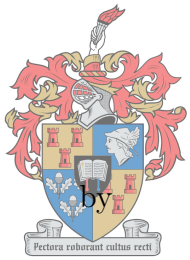
Development and demonstration of a Customer Super-Profiling tool utilising data analytics for alternative targeting in marketing campaigns



by

Marisa Walters

Thesis presented in fulfilment of the requirements for the degree of Master of Engineering (Industrial Engineering) in the Faculty of Engineering at Stellenbosch University

Supervisor: Prof JF Bekker

December 2018

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: December 2018

Acknowledgements

I would like to express my sincere gratitude to the following people and organisations for their contribution to this thesis:

- Professor James Bekker, my study leader, for your exceptional guidance, sharing your knowledge, wisdom and time. Thank you for giving me the opportunity to pursue my master's degree under your supervision.
- The completion of thesis would not have been realised without the love and support of my family and friends. Thank you for always believing in me, when I could not do so myself.
- The USMA members for the encouragement, memories and friendships that made the completion of this thesis possible and pleasant.
- The generous financial support by Bytes Universal Systems.
- Anne Erikson, for proofreading the document and making helpful suggestions.

Deo gratias

Abstract

Being part of a competitive generation demands that a business has good marketing policies to attract new customers as well as to retain existing ones. Marketing managers can develop long-term and healthy relationships with customers, if they can detect and predict changes in their customers' purchasing behaviour. With the growth of information systems and technology, businesses have an increasing capability to accumulate huge quantities of customer data in large databases. However, much of these potentially useful marketing insights into customer characteristics and their purchasing patterns often remains hidden and untapped. Therefore, businesses can achieve competitive advantages by studying customer behaviour through *data mining* tools (*i.e.* supervised and unsupervised learning) and techniques (*i.e.* classification, regression and clustering).

The goal of this research project was to develop a *Customer Super-Profiling* (CSP) tool that has the ability to analyse large (non-aggregate) customer datasets, considering both demographic and behavioural features. The data analytics was done by utilising more than one data mining tool, which generates customer super-profiles. These profiles are used to attract and classify new customers as well as to retain existing customers, providing the user with the ability to predict each customer's specific needs.

This research project outlines a general methodology for segmentation of customers by using the model of Recency, Frequency and Monetary (RFM), together with k-means clustering (unsupervised learning) to identify the various types of customers within the dataset. Customer profiles are then generated, in the form of decision rules (supervised learning) to identify each type of customer as well as classifying them into the various clusters created. These predictions are performed based on the customers' demographic and

behavioural features. The CSP tool was applied and demonstrated on large customer datasets from four different domains and useful results were found.

Opsomming

Om deel te wees van 'n mededingende generasie vereis dat 'n besigheid oor 'n goeie bemarkingsbeleid beskik om nuwe kliënte te werf asook om bestaande kliënte te behou. Bemarkingsbestuurders kan langtermyn en gesonde verhoudings met kliënte ontwikkel, as hulle veranderinge in hul kliënte se koopgedrag kan opspoor en voorspel. Met die groei van inligtingstelsels en tegnologie het besighede 'n toenemende vermoë om groot hoeveelhede kliëntedata in groot databasisse op te bou. Baie van hierdie potensieël nuttige bemarkingsinligting oor kliënteienskappe en hul kooppatrone bly egter steeds weggesteek en onbenut. Daarom kan besighede mededingende voordele behaal deur kliëntgedrag met data-ontginningsgereedskap (d.w.s. begeleiding en onbegeleide leer) en tegnieke (d.w.s. klassifikasie, regressie en groepering) te bestudeer.

Die doel van hierdie navorsingsprojek was om 'n *Kliënt-superprofiel* (KSP) instrument te ontwikkel wat die vermoë het om groot (nie-saamgestelde) kliëntdatastelle te analiseer, met inagneming van beide demografiese en gedragseienskappe. Die data-analise is gedoen deur gebruik te maak van meer as een data-ontginningsgereedskap, wat kliënte se superprofiele genereer. Hierdie profiele word gebruik om nuwe kliënte te lok en te klassifiseer, sowel as om bestaande kliënte te behou, wat die gebruiker die vermoë bied om elke kliënt se spesifieke behoeftes te voorspel.

Hierdie navorsingsprojek beskryf 'n algemene metodologie vir segmentasie van kliënte deur gebruik te maak van die model van Onlangs, Frekwensie en Monetêre waarde (OFM), tesame met k-mediane (onbegeleide leer) om die verskillende tipes kliënte binne die datastel te identifiseer. Kliëntprofiele word dan gegenereer in die vorm van besluitreëls (begeleide leer) om elke tipe kliënt te identifiseer asook om hulle in die verskillende groepe wat geskep word, te klassifiseer. Hierdie voorspellings word uitgevoer op grond van die

demografiese en gedragseienskappe van die kliënte. Die KSP-instrument is toegepas en gedemonstreer op groot kliëntdatastelle van vier verskillende domeine en nuttige resultate was gevind.

Contents

List of Figures

List of Tables

Nomenclature

Acronyms

kNN	k-nearest neighbour
AHP	Analytical hierarchical process
ANN	Artificial neural networks
BI	Business intelligence
CA	Correspondence analysis
cRFM	Combined Recency Frequency Monetary
CRISP	Cross-Industry Standard Process
CRM	Customer relationship management
CSP	Customer Super-Profiling
EERD	Extended entity relationship diagram
ERD	Entity relationship diagram
GA	Genetic Algorithm
GUI	Graphical user interface
KDD	Knowledge Discovery in Databases
MFA	Multiple factor analysis
MS	Microsoft

ODBC	Open Database Connectivity
OPD	Object–Process Diagram
OPL	Object–Process Language
OPM	Object–Process Methodology
PC	Principal components
PCA	Principal component analysis
RFC	Recency Frequency Cost
RFM	Recency Frequency Monetary
RFMTC	Recency Frequency Monetary Time Churn probability
RMSE	Root mean squared error
RSS	Residual sum of squares
SaaS	Software-as-a-Service
SEMMA	Sample, Explore, Modify, Model and Access
SL	Supervised learning
SOM	Self-organising maps
SQL	Structured Query Language
SSE	Sum of squared errors
SSMS	SQL Server Management Studio
STP	Segmenting Targeting Positioning
SVM	Support vector machine
UL	Unsupervised learning
WRFM	Weighted Recency Frequency Monetary

Chapter 1

Introduction

This chapter serves as introduction to the research presented in this thesis. The foundation of the research is explained by providing background information on the research question, followed by the formal research assignment, research scope, objectives and methodology. Finally, the structure of the document is explained.

1.1 Research background

In recent years, information technology has transformed the way marketing is done and how companies manage information about their customers (Shaw et al., 2001). According to Apeh et al. (2014), in the past, researchers used to apply statistical surveys in order to study customer behaviour. In today's fast-moving world of marketing from product-orientation to customer-orientation, the management of customer treatment can be seen as a key to achieving revenue growth and profitability (Hosseini and Shabani, 2015). Marketing managers can develop long-term and pleasant relationships with customers if they can detect changes in their purchasing behaviour. To gain more insights into customer behaviour, customer profiles should be constructed. Customer profiles are not the same as demographic information. Demographics usually provide the key dimensions that advertisers seek (age, gender, *etc.*), whereas profiling groups these dimensions along with other elements (behaviour) in order to create the ideal customer profile (Brown, 2016).

Marketing intelligence, which emphasises the marketing-related aspects of business intelligence, has traditionally relied on market surveys to understand the consumer's

behaviour and improve product design. For example, companies use consumer satisfaction surveys to study customer attitudes. Lately, data analytics technology can monitor key factors for strategic marketing decisions, for instance the customers' opinions on different aspects, namely, a product, service, or company, by using data mining tools (Fan et al., 2015). According to Fan et al. (2015), integrating mixed information from different sources provides a complete view of the area of interest and generates more accurate marketing intelligence, while analysis models developed on a single data source may only provide limited insights, leading to potentially biased business decisions.

This research offers an approach to building customer profiles through *data mining* tools (*i.e.* supervised or unsupervised learning) and techniques (*i.e.* classification, regression, clustering *etc.*), when having a customer dataset with typical monetary transactional data, demographic data and extra value adding customer attributes; which include mobile phone type, medical aid *etc.* Data mining is a technique used to extract knowledge from information (Chen and Chen, 2010). The goal of data mining differs from one area to another. When applying data mining to analyse data and create customer profiles, it will help to discover hidden knowledge in datasets to better understand customer behaviour and needs (Shaw et al., 2001). Thus, data mining can be defined, with respect to customer profiling, as being the technology that allows the building of customer profiles (among other functions), where each profile describes the specific habits, needs and behaviour of a customer group. Therefore, developing customer profiles is an important step for targeted marketing campaigns, for it not only classifies new customers, but also provides information on current customers.

Figure 1.1 shows the purpose of the proposed *Customer Super-Profiling* (CSP) tool. It functions as a super-profiling analytics tool that receives various customer attributes as input to create customer super-profiles (Walters and Bekker, 2017). The customer attributes include demographic information (age, gender, ethnicity, *etc.*), transactional data, as well as extra value-adding attributes (transportation type, mobile phone, *etc.*). Businesses in search of campaign ideas appoint advertising companies to assist them with marketing campaigns. Conversely, advertising companies may be in search of companies/developers that possess a profiling tool to provide them with reliable customer profiles for targeted marketing campaigns. These advertising companies are the value-creation partners: when they collaborate with the business partner they provide a revenue stream. The value that the advertising companies receive is knowledge about

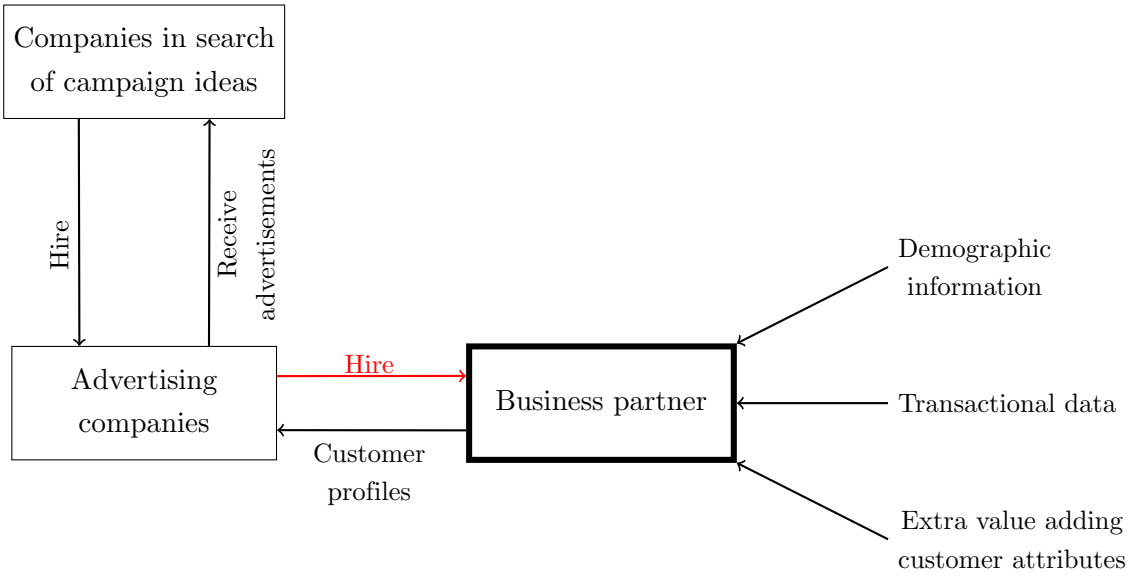


Figure 1.1: Illustrating the use of customer super-profiling (Walters and Bekker, 2017).

current and/or potential customers: who they are, what their behaviour and interests are and where to find them. This information provides the companies with insights in order to target suitable customers. The customers' information (personally identifiable information) is not sold to the advertising companies, they only receive the customer profiles generated after data analysis has been performed by the business partner. After the '*Facebook–Cambridge Analytica data scandal*' it is important to assure customers that their data and information are utilised and analysed *via* a safe and legal process (Lee, 2018; Solon, 2018).

Instead of stating the traditional research hypothesis, it is appropriate in this study to rather state a research assignment, as presented next.

1.2 Research assignment

Every business, large or small, needs a competitive advantage to distinguish itself from the competition. One of the strategic tasks performed to achieve this advantage is to understand the customers. The work to be done in this research project was inspired by a need experienced by an industry partner of the Industrial Engineering department of Stellenbosch University. The nature of the work is to determine whether a CSP tool can

be built using large datasets and machine learning. Hence a research assignment can be formulated, as follows:

> Develop a CSP tool which considers both demographic and behavioural features of customers utilising more than one data mining tool to generate customer profiles superior to the traditional profiles.

1.3 Research Scope

The scope of the research gives an indication of what the project entails. The data that will be used to develop the customer profiles will be fictitious data, created by a data simulator. The research will focus on at least two data domains which will be identified later. The number of transactions per '*customer*' in each data domain will be unspecified, yet finite. The amounts/tariffs for each product or service should be realistic and must usually correlate. Different types of customers will be simulated, those spending high, average and low amounts when participating in purchasing activities, as well as those participating in activities frequently, average and less frequently.

The research will include an assessment of different data analytics tools and techniques, which will guide the researcher to select the appropriate tools and techniques for the type of data that will be utilised. The objectives pursued to support the research assignment are discussed next.

1.4 Research objectives

The following objectives were stipulated in order to complete this research:

1. *Determine* if it is possible to develop a *Customer Super-Profiling* (CSP) tool that has the ability to analyse a large dataset.

2. The CSP tool should be able to *utilise* various Big Data Analytics tools and techniques and *generate* customer super-profiles which have more value than the typical demographic data (see also Figure 1.1).

1.5 Research problem-solving methodology

In order to achieve the objectives of this research, the following approach and methodology are proposed:

1. *Conduct* a comprehensive literature review by getting accustomed to relevant aspects in the domain of:
   (a) Segmentation
   (b) Customer profiling
   (c) Marketing strategies
   (d) Big Data Analytics
      i. Data Analytics processes
      ii. Data cleaning
      iii. Data transformation
      iv. Data mining and machine learning
2. *Develop* a solution architecture. This architecture will be the symbolic representation of the structural relations between objects in the system and its processes, demonstrating the planned abilities of the CSP tool.
3. *Determine* the customers' demographic and behaviour features that will be simulated for this research.
4. *Develop* a data simulator that is able to create customer datasets, which is necessary to provide the super-profiling tool with data. To conduct this research, big datasets are necessary. The data simulator will determine the structure and content of the data. Eventually, a different user of this super-profiling analytics tool could provide their own data, as long as the data have the same format and structure.
5. *Validate* the data simulator together with the datasets created (step 4). This is done to ensure that the customer data is reliable for use and analysis. Performing validation on the data simulator will also enable the researcher to create other datasets following the same principle.

6. *Design* and *develop* the CSP tool that contains a suite of data analytics techniques. The suite should contain more than one data mining tool (*i.e.* unsupervised and supervised learning).

7. *Demonstrate* the CSP tool by utilising the simulated South African demographic customer dataset. The *validation* of the CSP tool should be performed on another dataset(s), which will build confidence in the profiling capability of the tool.

8. *Draw* conclusions based on the results received from both the demonstration and the validation. These results should be able to provide a point of departure for future research.

Steps 2, 3 and 4 strive to fulfil Objective 1. The architecture that will be developed in step 2 will provide a broad overview of the tool, indicating the abilities and functionalities of the tool. The customers' demographic and behavioural data (steps 3 and 4) will function as the big datasets to be analysed.

Steps 1, and 4 to 8 will be performed in pursuit of Objective 2. Step 1 will provide the researcher with understanding of the various data analytics tools and techniques, as well as background on transitional customer profiles. Step 4 will also be used to achieve this objective, because the type of data that is available determines the data analytics techniques that can be utilised. Steps 5 to 8 will be performed to test, validate and document the results received from the CSP tool.

1.6 Deliverables envisaged

It is envisaged that this research will provide *another* point of departure in the domain of Big Data Analytics. The researcher will be developing a CSP tool containing a suite of Big Data Analytics tools and techniques that will allow for customer super-profiling. Guidance, to operate the CSP tool, will be provided from the architecture, while the aim of the tool is to build reliable customer models for targeted marketing campaigns.

1.7 Structure of the document

This chapter contains a description of using data analytics to develop customer profiles. This led to the formulation of the research assignment, objectives and methodology.

In **Chapter 2**, a literature review on segmentation and customer profiling is presented. This includes various segmentation models as well as the evolution of capturing customer information, and lastly marketing strategies.

Chapter 3 provides a comprehensive literature review regarding Big Data Analytics. The chapter includes an introduction to Big Data, Big Data Analytics and data mining, with application areas and references to various data mining tools and techniques as well as machine learning algorithms.

The theoretical background and literature reviews conducted in the previous chapters provide knowledge to develop a solution architecture for the proposed data simulator and CSP tool, which is described in **Chapter 4**. This chapter includes a toy problem as well as a big dataset problem to illustrate the solution architecture.

The demonstration and validation of the CSP tool is presented within **Chapter 5**. The demonstration is performed by utilising the customer datasets, whereas the validation is performed on different datasets. This chapter also includes several business case scenarios. The summary and general conclusions of the research are presented in **Chapter 6**.

Appendix A presents the process of creating domain-specific datasets, as well as the validation of the datasets. The datasets are created in Matlab® and imported into Microsoft® SQL Server®. Key descriptors of the information system are included in **Appendix B**, and pseudocode for the data simulator is shown in **Appendix C**.

This concludes **Chapter 1**. A literature review on segmentation and customer profiling is presented next.

Chapter 2

Segmentation and customer profiling

The previous chapter served as an introduction to this research. It stated the background, the problem and clearly defined the scope. The objectives were introduced and the research methodology was developed. The structure of the document was also presented in Chapter 1. Chapter 2 contains a literature review in order to fulfil Objective 2. The literature will focus on *segmentation* and *customer profiling*. The chapter will commence by outlining segmentation and the drawbacks involved in performing segmentation. This is followed by a brief development history of customer profiling and then defining customer profiling. Thereafter, two customer profile types will be reviewed.

Segmentation is often used in conjunction with profiling. However, segmentation is a term used to describe the process of dividing customers into homogeneous groups on the basis of shared or common characteristics/attributes, *e.g.* habits, tastes, *etc.*, while customer profiling is describing customers based on their personal attributes, such as *age, gender, income and lifestyles*. Having these two components, marketers can describe which marketing actions to take for each segment and then allocate resources to the segments in order to meet specific business objectives. Therefore, literature regarding a marketing strategy concludes Chapter 2. However, as this chapter expresses the viewpoint of an industrial engineering researcher, it is not as comprehensive as would be expected of a marketing student.

A synthesis at the end of the chapter will include the researcher's view and interpretation of what was observed while performing this review.

2.1 Segmentation

Segmentation is performed on an *unordered* customer dataset and is the process of *separating* markets into groups of potential customers with similar needs and/or characteristics, who are likely to exhibit similar purchasing behaviour (Weinstein, 2013). Segmentation is also seen as a way to have more targeted communication with customers (Jansen, 2007). According to Jansen (2007), the process of segmentation describes the characteristics of the customer groups, called the segments, or clusters, within the data.

Literature does not provide a plausible difference between market segmentation and customer segmentation, therefore the researcher took the view that market segmentation is generally used for high-level strategy, whereas customer segmentation provides a more detailed view. Market segmentation is a well-known and popular marketing technique and its benefits are emphasised in numerous marketing research textbooks (Müller and Hamm, 2014).

2.1.1 Market segmentation

Market segmentation was first defined by Smith (1956), as being based upon developments on the demand side of the market and representing a rational and more precise adjustment of product and marketing effort to customer requirements, thus providing a conceptual view of an inherently heterogeneous market. Marketing segmentation involves viewing a heterogeneous market (external heterogeneity) as a number of smaller homogeneous markets (internal homogeneity), in response to differing preferences, regarded as being caused by the desires of customers for more precise satisfactions of their varying wants (Wedel, 2002). In other words, searching for a process that minimises differences between members of a segment and maximises differences between segments. The segmentation approach must yield segments that are meaningful and applicable to specific marketing problems and then tailor a marketing mix for the targeted segments, rather than offering the same marketing mix to a huge heterogeneous group (Liu et al., 2012; Smith, 1956).

A fundamental task of market segmentation is to group the customers based on similarities in their needs, characteristics and preferences (Liu et al., 2012; Müller and Hamm, 2014). According to Weinstein (2013), the objective of segmentation research is to *analyse* markets, *find* niche opportunities and *capitalise* on superior competitive

positions. This process is begun by selecting one or more groups of users as targets for marketing activity and developing unique marketing programmes to reach these prime prospects (market segments). Similarly, Ungerer (2015) defined the purpose of market segmentation to be to meaningfully leverage scarce resources and target specific needs of different customer groups. This implies that different customer segments have different needs, and therefore it is vital to tailor specific offerings to them.

2.1.1.1 Dimensions for conducting market segmentation

According to Weinstein (2013), it is recognised that segmentation is both a science and an art. There are many alternative methods for segmenting business markets and several of these approaches are derived from the consumer behaviour field (Weinstein, 2013). Decision-making is impacted by both rational and emotional factors (*e.g.,* demographics, geographic, benefits, motivations, needs, purchasing habits, *etc.*). Customer demographic and socio-economic measures (age, gender, income, *etc.*) can be studied, and product consumption can be evaluated. In addition to these measures, credit card utilisation, brand loyalty and price sensitivity issues may also be insightful for segmenting the market. These variables provide evidence that the options are many; therefore, further research is necessary to determine the best approach(es), as indicated by Weinstein (2013).

The output of segmentation depends on the data that are available for use. However, there are some relatively standard segmentation proposals that fit with the most needs-based (Goyat, 2011) or value-based segmentation initiatives (Nguyen, 2016). Segmentation usually utilises customer attributes (Jansen, 2007). The following short examples and definitions, as referred to by Weinstein (2013) and Liu et al. (2012), illustrate six common business segmentation dimensions in action:

1. *Geographic location of a customer*: Collecting and analysing information according to the physical location of the customer is often used in marketing, for companies selling products and services would like to know where their products are being sold in order to increase advertising and sales efforts at that location. When referring to geographic features, it includes *region, size of city or town, density and climate* (Jooste et al., 2012).

2. *Business demographics*: A graphic supplies distributor can easily target advertising agencies by using business demographic variables or firmographics. Using *Advertising Age* and *Adweek* references, the company is able to find information about anticipated size, media specialisation, services offered, major accounts, key personnel, *etc.*

3. *Adopter categories:* Classifying customers on the basis of their degree of *readiness* to try a new product. Market researchers have identified five categories that exist in every market segment. Customers with the highest readiness are *innovators*, venturesome customers and risk takers who are the first users; these customers comprise 2.5% of the target market. *Early adopters* are status-oriented opinion leaders and represent about 13.5% of the target market. *Early majority* from the leading segment of the mass market; about 34% of the target market. *Late majority* are followers of the early majority and are also about 36% of the target market, and finally, *laggards* the conventional, price-conscious segment; making up about 14% of the target market.

   This segmentation dimension can be most informative for new product concepts using exploratory studies and qualitative procedures.

4. *Benefits:* A form of market segmentation, based on the differences in specific benefits that different groups of customers or companies look for in a product. It may be price, service, special features, and/or reputation of the seller (Xerox or Brand X). A benefit to one customer (enhanced features) may be a drawback to another (higher price).

5. *Product usage:* An approach often used by marketers is based on product or brand usage by customers. Product usage segmentation can take a number of directions, for example, the marketer may want to identify various segments of users for a particular product category or users of the company's brand. Marketers may also want to segment customers into those who buy frequently versus those who buy occasionally, or into those who usually purchase just one brand versus those who switch from brand to brand. In addition, the 'best' customer can be identified by several criteria: number of orders, revenues, unit sales, profitability, share of customer volume, *etc.*

6. *Purchasing approaches:* Characteristics of traditional and strategic purchasing approaches differ, they may be reactive or proactive, short term or long term, non-integrative or integrative, *etc.* For example, Dell's strategy of seeking sophisticated buyers and large accounts; not requiring much 'hand-holding', is accepted as accurate target marketing.

According to Ungerer (2015), customer groups represent separate segments if:

1. Their needs require and justify a distinct offer.
2. They are reached through different distribution channels.
3. They require different types of relationships.
4. They have substantially different profitabilities.
5. They are willing to pay for different aspects of the offer.

This could involve segmentation by means of geographic location, social standing, common needs, common behaviour and other attributes (Ungerer, 2015).

Segmentation is essential to cope with today's dynamically fragmenting consumer marketplace. By utilising segmentation, marketers are more effective in channelling resources and discovering opportunities. The overall objective of using a market segmentation strategy is to improve customer satisfaction and the competitive position of a business as well as better serve the needs of the customers. This is accomplished by tailoring a unique marketing mix (strategy) for targeted segments and in the process achieving maximum efficiency (Liu et al., 2012; Weinstein, 2013).

2.1.1.2 Methods for conducting marketing segmentation

A common tool used for grouping customers with similarities is called *clustering*. According to Hosseini and Mohammadzadeh (2016), the process of collecting a set of physical or abstract objects into groups of similar objects is called clustering. Both the academic researcher and the marketing applications researcher rely on the technique for developing empirical groupings of persons, products, or occasions which may serve as the basis for further analysis. The aim of applying the clustering technique is to maximise

within-segment homogeneity (Punj and Stewart, 1983). Each segment is a group of homogeneous customers that marketers can *identify*, *target* and *communicate* with (Liu et al., 2012).

In early market segmentation research, clustering was synonymous with market segmentation (Liu et al., 2012). However, as the spectrum of market segmentation expanded to studies concerning customer interaction with marketing mix, the market segmentation techniques evolved to simultaneously considering multiple sets of variables (more than one segmentation base). According to Liu et al. (2012), market segmentation is constantly under investigation by researchers. There is an abundance of segmentation methods available, including k-means clustering, hierarchical clustering, automation detection, classification and regression trees, neural networks, *etc.* The overall objective for utilising a market segmentation strategy is to improve a business's position and better serve the needs of the customers. This is accomplished by discovering and characterising customer groups and attaining profitable customer segments (Weinstein, 2013).

2.1.2 Customer segmentation

As mentioned, the difference between market segmentation and customer segmentation is still unclear and the researcher takes the view that customer segmentation provides a much more detailed view of the customers. This section will discuss customer segmentation and the different approaches.

According to Chan (2008), most marketers experience difficulties in identifying the right customers to engage in successful campaigns. Thus far, customer segmentation is a popular method that is used for selecting appropriate customers for a campaign. Bose and Chen (2009) mentioned that for product advertising and promotions, there are mainly two approaches that are used in practice; *mass marketing* and *direct marketing*. Mass marketing targets large groups of customers; it does not distinguish between customers within a cluster/group and the information delivered to customers is uniform, whereas direct marketing targets individuals or households. Different customers are subjected to different marketing information.

Direct marketing is defined as "the delivery of a marketing message or proposition to a target customer or potential customer, in a customer favourable format, put to the customer for the seller or the seller's agent without an intermediary person or indirect

media involved", while customer segmentation is the *classification* of the different customers that exist in a market, based on similar needs, product or service requirements, or some other characteristics (Ungerer, 2015). It can be concluded that direct marketing classifies customers so that *personalised advertising* and *promotional activities* can be targeted to specific classes of customers (Bose and Chen, 2009), and therefore, customer segmentation can be linked to direct marketing.

Customer segmentation can be performed by utilising various models (Berger and Nasr, 1998; Jain and Singh, 2002; Khajvand et al., 2011; Kim et al., 2006; Sarvari et al., 2016). However, the *Recency, Frequency and Monetary* (RFM) model, which represents *customer behaviour characteristics*, may be the most powerful and simplest technique for generating knowledge from customer relationship management (CRM) data (Dursun and Caber, 2016). This model will be discussed next in more detail.

To achieve business success, engaging in effective campaigns is a key task for marketers. Traditionally, marketers first segment the market, and then target profitable customers. However, this process brings forth problems, for the correlation between customer segments and a campaign is neglected. Therefore, as stated by Jonker et al. (2004), it is necessary to consider significant campaign-dependant variables of customer targeting in *customer segmentation*. An approach to combine customer segmentation and customer targeting for campaign strategies was defined by Chan (2008). The investigation identifies *customer behaviour*, using the well-known *Recency*, *Frequency* and *Monetary* (RFM) analytical model. The '*R*' refers to the duration of time between the last purchase time and the time of the 'survey'. The desired state is shorter in duration, so that R is bigger. The '*F*' indicates the total number of purchases during a specific period, thus the desired state is a bigger F-value, which means that there has been a high repetition of purchases, and the '*M*' indicates the monetary value spent during one specific period, where the desired state is for there to be much money, so M is bigger (Sarvari et al., 2016).

After using the RFM model to represent the customers' behaviour, the data is coded (encoded) into five categories. This is seen as one of the traditional applications of the RFM model, and is called '*the customer quintile method*'. By coding, each customer is compared with all the others, depending on the variables used (Chan, 2008; Dursun and Caber, 2016). If the value lies between 100% and 80%, the categorical value is set to 5, between 80% and 60%, the value is set to 4, *etc.* In this way, the database is divided

into 125 (5×5×5) equal clusters. The customers who obtain the highest RFM scores are generally the company's most profitable customers.

The purpose behind utilising the widely used behavioural-based data mining method, RFM, is to analyse the customer's behaviour and then make predictions based on the behaviour in the database (Wei et al., 2010). This model is used in various research areas, which defines valuable customers as those simultaneously having high *recency, frequency* and *monetary* values. According to Hosseini and Shabani (2015), one of the most effective customer segmentation models, based on customer value, is the RFM model. By adopting the RFM model, decision-makers have the ability to identify valuable customers and then develop effective marketing strategies (Wei et al., 2010). Previous studies show that "the bigger the values of R and F are, the more likely the customers are going to produce a new trade with the company; the bigger the M is, the more likely the customers are going to buy more services or products from the company" (Cheng and Chen, 2009).

The application of the RFM model is as follows. According to Wei et al. (2010) and Sarvari et al. (2016), the RFM model measures when customers have purchased lately (recency), how often (frequency) and how much (monetary) they spent. Customers' past purchases can effectively predict their future purchase behaviour. Companies can identify which customers are entitled to be contacted according to their past purchase behaviour based on the RFM model approach, which is extensively applied in database marketing and is a common tool to develop marketing strategies. When applying the RFM model, the customer's name and address need to be assigned by a unique key (account number) and order; the sales information also needs to be stored, with the unique key included in each transaction record (Kahan, 1998). Utilising the analysis of the RFM model, Thompson (1999) applied the RFM model to classify customers into (1) uncertain customers, (2) spenders, (3) frequent customers and (4) the best customers.

The RFM model has been widely applied in many practical areas (Wei et al., 2010) and its indicators are adaptable to measure customer value and to segment customers in different service areas (Dursun and Caber, 2016), as seen in Table 2.1. In addition to the areas listed in Table 2.1, Wei et al. (2010) mentioned that the RFM model could be used to segment customers, calculate customer value and customer lifetime value, observe customer behaviour, estimate the response probability for each offer type and evaluate online reviewers.

Table 2.1: Practical areas where the RFM model has been applied

Areas	Sources
Banking and insurance industries	Hsieh (2004), Sohrabi and Khanlari (2007)
Government agencies	King (2007)
Online industries	Li et al. (2010)
Telecommunication industries	Li et al. (2008)
Travel industries	Ha and Park (1998), Lumsden et al. (2008)
Marketing industries	Spring et al. (1999), Jonker et al. (2006)
(Global pizza) Restaurant chain	Sarvari et al. (2016)

There are, however, both advantages and disadvantages when utilising the RFM model (Dursun and Caber, 2016; Wei et al., 2010). The advantages include:

1. RFM is a powerful tool for assessing customer lifetime value, which is also combined with frequency pattern mining techniques.

2. RFM is cost-effective in acquiring important customer behaviour analysis and can easily quantify customer behaviour, where customers and transactional data can be stored in an accessible electronic form. Therefore, decision-makers can easily understand the application of the RFM model.

3. RFM is beneficial in predicting response and has the ability to boots a company's profits in the short term.

4. It is very effective to model using RFM variables as purchasing behaviour can be summarised by using a very small number of variables.

5. RFM variables are gathered *via* an internal database that contains customer-specific information regarding the transaction history and are not obtained through the aggregate level information in the demographic databases. Thus, RFM is more meaningful for targeting particular customers.

6. RFM is a well-known method used to measure the strength of the customer relationship as it can effectively identify valuable customers.

Although the RFM model is seen as a crucial tool for businesses to develop marketing strategies, it also has disadvantages, which include:

1. Given that the RFM model aims at identifying valuable customers in companies, it only focuses on 'the best customers'. It provides little meaningful scoring on recency, frequency and monetary when most customers do not buy often, spend little and have not purchased lately. This is particularly true for most company sales, and is referred to as the *Pareto Principle* –the 80/20 Rule. The Pareto rule states that 80% of the results come from 20% of the causes, similarly; 20% of the customers contribute to 80% of the company sales.

   It can be said that the model ignores the analysis of new companies setting up in a short period and customers that only purchased once and placed small orders. These customers are referred to as type 1-1-1 customers and it is stated that they are the biggest customer segment and may have the greatest untapped potential.

2. The RFM model can only use a limited number of selection variables. The simplicity of the RFM model has been overemphasised, while its ability to differentiate has little to be considered.

3. There are usually high correlations between the Frequency and Monetary values.

4. RFM focuses on a company's current customers and cannot be applied to scouting for new customers as a marketer does not have transactions for prospects.

5. RFM estimates a single response model for all the customers in the database, and therefore assumes the database is homogeneous, which is contradictory to the real situation, for customers often have considerable heterogeneity.

6. The RFM model is not introduced as a precise quantitative model and the importance of each RFM measure is different between industries and applications.

However, the weaknesses in some areas of the RFM model led to the disadvantages being discussed. These disadvantages introduce some minor modifications or extensions on the RFM model, for instance, it was suggested by Miglautsch (2002) that sub-segmentation can help to identify 1-1-1 customers, involving three classes of variables, (1) internal purchase information, (2) geo-demographic information connected to postal code, and (3) customer variables.

There are, however, numerous studies in favour of combining two types of data. The most popular combination is that of the customer's demographic data plus their purchasing history. The purchasing history is important for direct marketing, for marketers can predict the choices of customers more effectively than by using only demographic data (Bose and Chen, 2009).

Behavioural data has gained the favour of most researchers, and they have been using this data in various situations, industries and under various conditions, which include adding extra parameters, leading to the *extended* RFM analysis method (Khajvand et al., 2011). These models are also referred to as *expanded* or *adapted* RFM models. Due to the RFM model's disadvantages, researchers have attempted to improve the predictability of RFM models through adding additional variables to predict customer behaviour or develop new models to test whether they perform better than the traditional RFM model (Wei et al., 2010). Table 2.2 provides a review of researchers who adapted the traditional RFM model.

Table 2.2: Previous research regarding the adapted RFM model

Adapted RFM model consists of:	**Purpose:**	**Findings:**	**Sources:**
Two additional parameter: • Past purchase behaviour. • Additional customer variables.	Predict partial churn behaviour.	Past purchase behaviour, particularly RFM variables are the best predictors of partial customer defection.	Buckinx and Van den Poel (2005)
One additional parameter: • Period of product activity.	Classify customer product loyalty under business-to-business concept.	The developed methodology (adapted RFM model) produces better results than other commonly used models.	Hosseini et al. (2010)

Continued on next page

Adapted RFM model consists of:	**Purpose:**	**Findings:**	**Sources:**
Two additional parameters (RFMTC): • Time since first purchase. • Churn probability.	Selecting targets for direct marketing from a database together with estimating the probability that a customer will purchase a next time and the expected value of the total number of times that the customer will purchase in the future.	The proposed RFMTC model provides more predictive accuracy than the RFM model.	Yeh et al. (2009)
Change RFM to Recency, Frequency and Cost (RFC) model.	Targeting the highest-scoring citizens (*e.g.* drug users, vandals, noisy neighbours) in order to improve or reduce their use of services.	English authorities are planning to use this technology to assist them to understand their citizens better, for example, who is entitled to more benefits, or who is due for a visit from a social worker.	King (2007)
One additional parameter: • Count item.	Performing segmentation on customers.	Adding the extra parameter makes no difference to the clustering results.	Khajvand et al. (2011)
One additional parameter: • Weighted RFM (WRFM).	Determine loyalty degree of product to achieve an excellent CRM.	Massive improvements in classifying accuracy of loyal customers.	Hosseini et al. (2010)
One additional parameter: • Weighted RFM (WRFM).	Measure customer loyalty and estimating the customer loyalty rate.	The results indicated high precision when using the WRFM.	Zalaghi and Varzi (2014)

Continued on next page

Adapted RFM model consists of:	**Purpose:**	**Findings:**	**Sources:**
One additional parameter: • Types of customers.	Propose a loyalty measurement model.	The proposed segmentation model makes customers feel good, increases the sales for a company and helps to reach more targeted customers.	Bunnak et al. (2015)
Combination of WRFM model and demographic attributes.	Determines the best approach to customer segmentation and extrapolate associated rules for this, based on the RFM considerations, as well as demographic factors.	Results showed that having an appropriate segmentation approach is vital if there are strong association rules. Weight of RFM attributes affects rule association performance positively; moreover, to capture more accurate customer segments, a combination of RFM and demographic attributes is recommended for clustering.	Sarvari et al. (2016)
Use 3, 5 and 7 categories/classes while performing RFM.	Determine customer loyalty of the different category sizes, enhance classification accuracy.	Difficult to determine the best situation, it is a trade-off. Smaller classes (3) increase the accuracy rate as well as the number of target customers, while bigger class sizes (5,7) have a lower accuracy rate, but a smaller customer group to target.	Cheng and Chen (2009)

As seen in Table 2.2, the weighted RFM is used by various researchers, and can be employed as follows. First the R, F and M-values are specified for each customer and then their weights, which represent their relative importance, are determined using the analytical hierarchical process (AHP) method and each customer's value is calculated

based on the WRFM-value (Zalaghi and Varzi, 2014). It can be concluded that the weights differ for each research area where the RFM model is applied, as the importance of each variable (R,F and M) differs for each application area.

After performing customer segmentation, a customer segmentation and segment analysis should take place. There are various methods for analysing segments, these include statistical and machine-learning methods. After the analysis, the result enables marketers to plan and set up marketing strategies for each segment. Next, the drawbacks concerning segmentation will be discussed.

2.1.3 Segmentation drawbacks

Construction of user segmentation is not an easy task. The drawbacks associated with segmentation, mentioned by Jansen (2007), are:

1. *Relevance and quality of data*: These aspects are essential to develop meaningful segments. If the customer data is insufficient, the meaning of customer segmentation is unreliable and can almost be seen as worthless. On the other hand, too much data can lead to complex and time-consuming analysis. Poorly organised data (different formats, different source systems) makes it difficult to extract applicable information. The use of too many segmentation variables can be confusing and results in segments that are unfit for decision-making. As a result, effective variables may not be identified. Many of these problems are due to an inadequate customer database.

2. *Continuous process*: Segmentation demands continuous development and updating as new customer data is acquired. In addition, effective segmentation strategies will influence the behaviour of the customers affected by them. Therefore reclassification of customers will need to be kept in mind.

3. *Over-segmentation*: A segment can become too small to be treated as a separate segment.

This concludes the discussion of segmentation. Next, customer profiling is presented.

2.2 History of customer profiling

Customer profiling is a way to create a portrait of a customer to help make design decisions concerning a company's services. Customers are normally divided into groups of customers that share similar goals and characteristics. One of the goals of creating customer profiles is to ensure a better relationship with customers. The better the relationship, the easier it is to conduct business and generate revenue, this is usually referred to as CRFM. According to Soltani and Navimipour (2016), CRM is a management philosophy and strategy which enables a company to optimise revenue and increase customer value and service quality through understanding and satisfying the individual customer's needs. This section will make use of a time-line, as seen in Table 2.3, to uncover the history behind sales and the profiling platform that has transformed the business over the past few decades, with the help of various researchers referred to in the literature (Lyle, 2015; Reni, 2017; Soltani and Navimipour, 2016).

Table 2.3: Evolution of capturing customer information

Year	Method
1894:	*The telephone switchboard*: The telephone switchboard was the first major milestone for the customer, for it allowed the customer to easily resolve problems and receive product information without having to travel. The switchboard operator assists callers by answering the call and connecting the caller to the correct person or department. This invention planted itself firmly as an integrated part of customer relations for the next hundred years.
1950s:	*The ledger*: Businesses used pen and paper to track basic sales and information.
1960s:	*The call centres*: These centres were centralised offices used for receiving or transmitting large volumes of requests by telephone. These centres and the contact centre solutions that were developed were the beginnings of what are now customer service departments.
Early 1980s:	*The Rolodex*: The Rolodex offered companies the ability to spin through paper records, adding new customers while updating existing customer information, details and more.
Late 1980s:	*Database marketing*: This new process allowed companies to collect and analyse customer information; this enabled businesses to create customised communications in order to promote a product or service for marketing purposes (also known as direct marketing).

Continued on next page

Year	Method
Early 1990s:	*Contact management software*: This software enabled businesses to easily collect, store, find and organise customer contact information into what was effectively digital Rolodex.
1995:	*Sales force automation*: This is an integrated application of customised relationship management tools that effectively automate sales inventory, leads, forecasting, performance and analysis.
Late 1990s:	The acronym CRM is created.
2000:	*Mobile and SaaS*: CRM continues to evolve and by the end of the century, the first mobile CRM solution is introduced, as well as the first Software-as-a-Service (SaaS) CRM product.
Currently:	*Innovations*: Companies begin to see CRM as a way to manage all business relationships *via* a single platform. Key approaches include: 1. Increasing CRM's operability with legacy software. 2. Offering the platform *via* cloud. 3. Dramatically increasing the power of mobile and subsequently social CRM.

2.3 Customer profiling

Behaviour-based relationship marketing begins with customer profiling (Yankelovich and Meer, 2006). The term *customer profiling* involves a wide range of marketing and service approaches. Customer profiling provides a basis for marketers to interact with existing customers in order to offer them better services and retaining them. A customer profile is a snapshot of who a customer is, how to reach them and why they buy. In short, a customer profile is a collection of information that describes the customer. Customer profiling is the *process* of developing a profile using relevant and available information to describe the characteristics of an individual customer and to be able to identify discriminators from other customers and drives for their purchasing decisions (Ntawanga et al., 2010).

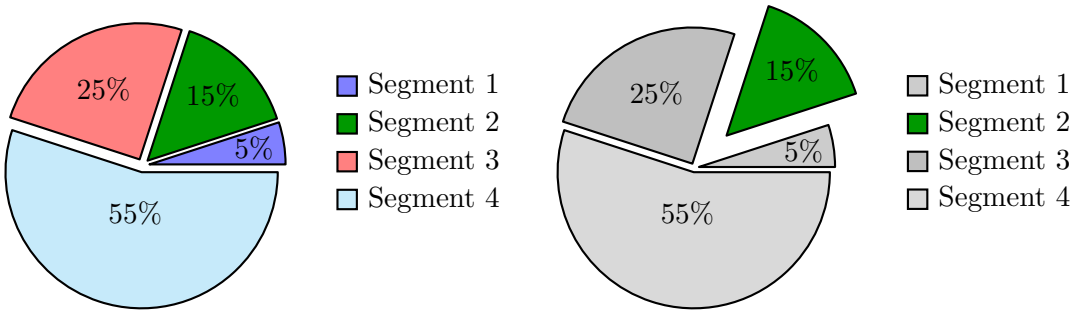
Profiling can be as simple as retaining credit card information at an e-commerce site or as complex as correlating a customer's demographic information with the relevant market segment statistics. Customer profiling is done by building a customer's behaviour model and estimating its parameters (Jansen, 2007). Customer profiling is

a way of applying external data to a population of possible customers. According to Yankelovich and Meer (2006), the purpose of profiling is to *identify* potential customers who could generate the maximum profit and *determine* the best possible ways of marketing to those customers, be it individuals or business enterprises. Thus, the goal remains to be that of building reliable customer models for targeted marketing campaigns, and consequently, better profitability (Romdhane et al., 2010). However, the effective utilisation of customer profiling has been a challenge for business enterprises (Yankelovich and Meer, 2006).

Customer data analysis enables the identification of customer profiles and customer preferences for specific products and services, as well as indicating the most appropriate channels to reach the customer and assess the profitability and life-time value of every individual (Zopounidis, 2012). The simplest way suggested for determining the manner in which to target new customers, is to profile existing ones (Yankelovich and Meer, 2006). While profiling, a firm identifies the characteristics of its best customers and then targets the non-customers with similar characteristics. These 'non-customers' may be first-time purchasers from the firm, or even individuals who have purchased from other divisions of the firm. According to Jansen (2007), there are various ways to use customer profiling, it all depends on the data available. It can be used to prospect for new customers or even to recognise existing bad customers.

A simple customer profile is a file that contains at least age and gender (Jansen, 2007). If one needs profiles for specific products, the file would contain product information and/or volume of money spent (Jansen, 2007). According to Yankelovich and Meer (2006), profiling is key for behaviour-based marketing.

Figure 2.1 offers a very basic understanding of the segmentation and profiling process. Segmentation was discussed earlier in this chapter, and can be explained as follows. As seen in Figure 2.1a, a customer population can be divided into different segments according to dimensions or characteristics. For example, suppose Figure 2.1a represents a population that was separated according to the region where the customers lived. It was seen that four segments of different sizes appeared. These different sizes illustrate that each region has a different number of customers. Thus, segmentation is done on the customer population. Customer profiling is a process that extracts each segment and evaluates the customers in that segment. Figure 2.1b illustrates the extraction of the 'green' segment, and then creates customer profiles for the customers present in



(a) Customer Segments (b) Customer Segment Exploded

Figure 2.1: Illustrating the segmentation and profiling process

that segment. These profiles contain demographic and other characteristics. There are broadly two kinds of profiling, *demographic* and *behavioural* profiling (Raines, 2009). These are discussed next.

2.3.1 Demographic customer profiles

Customer demographical profiling is a classic traditional marketing approach to follow, and contains a set of characteristics. The popular demographic categories are indicated in Table 2.4. Furthermore, customers at different stages of life have different needs and therefore purchase differently (Ungerer, 2015).

Table 2.4: Demographic categories for customer profiles

1.	Gender	7.	Children	13.	Medical cover
2.	Age	8.	Type(s) of cars	14.	Religion
3.	Income	9.	Savings	15.	Occupation
4.	Disabilities	10.	Race (Ethnicity)	16.	Geographical location
5.	Education	11.	Family size	17.	Marital status
6.	Employed	12.	Home ownership	18.	Political party affiliation

Consider the following customer profile:

Profile 1: Customer is married, has children, lives in an upmarket neighbourhood, and reads Economic Times newspaper.

Profile 1 involves a set of characteristics that are *demographic*. For someone in the advertising department, or when deciding the content for a website, a customer demographic profile is usually more important than a customer's behaviour profile, because it defines the market for advertisement sales and provides clues to editorial direction (Yankelovich and Meer, 2006). It is stated by Yankelovich and Meer (2006) that the *demographic profile provides vital help in attracting the customer and generating revenue in the early stages of an online project.*

2.3.2 Behavioural customer profiles

Customer behaviour is the process whereby individuals decide whether, what, when, where, how and from whom to purchase goods and services (Walters, 1974). However, Mowen and Minor (1998) provide a different definition by explaining customer behaviour as the study of the buying units and the exchange processes involved in acquiring, consuming, and disposing of goods, services, experiences as well as ideas. This definition focuses on buying units in an attempt to include not only the individual but also groups that purchase products or services (Mowen and Minor, 1998).

Behavioural profiling is based on customers' attitudes towards, use of, or response to a product. According to Larsen (2010), marketers believe that the behavioural variables that are the best starting points for constructing behavioural-based profiling include:

- *Occasions*: Customers are being profiled according to the time at which they get the idea to buy, make their purchase or use the purchased item. A company may choose one kind of marketing strategy around Christmas and another before Valentine's day.
- *Benefits*: The benefit profiling is a process that divides customers according to the different benefits they may look for in a product. Benefit profiling seeks to find:
  1. The benefits customers look for in a certain product.
  2. The type of customers who look for each benefit and the brands that deliver that benefit.

  Additionally, the benefit profiling process has the ability to identify customer profiles by making use of causal factors, rather than descriptive factors such as demographics.

- *User status*: Profiling according to non-users, ex-users, potential users, first-time users and regular users of a product, a company can customise and personalise its marketing for each group.
- *Usage rate*: Usage rate profiling separates the customers according to how much they use a product. They are divided into groups of non-users, light, medium and heavy product users.
- *Buyer-readiness stage*: This refers to customers' awareness and interest in the product.
- *Loyalty status*: A customer can also be profiled according to their loyalty. *Hard-core loyals* are customers that buy the same product many times, *split loyals* are customers that are loyal to two or three brands and buy these on a random basis, *shift loyals* are customers who shift from one brand to another and stay with that brand for a period until they shift to another brand, and lastly, *switchers* are customers who do not show loyalty or preference towards one particular brand, but rather buy a product or brand that is on sale or available at the time of purchase.
- *Attitude*: Customers can be separated based on whether they have an enthusiastic, positive, indifferent, negative or hostile attitude towards a product. By considering the customers' attitudes towards a brand or product the company will get a wide-ranging view of the market and its customers.

Now, consider the following customer profile:

Profile 2: A customer visited an enterprise website every day for two months, but has not visited the website at all in the past two weeks.

Profile 2 involves the real actions of a customer which are *behavioural-based*, and are concerned with customer action and behaviour. Utilising customer behaviour information to predict the future relationship with a customer is more prominent than utilising the demographic information about a customer (Profile 1). The database of customer behaviour provides a better criterion for business enterprises while forecasting their sales/transactions. Therefore, *customer behaviour profiling is critical to a company*

interested in retaining its customers and increasing their value. However, the combination of both the demographic and behavioural characteristics could serve as a powerful database in deciding the future profitability of a customer to a business enterprise (Yankelovich and Meer, 2006).

This concludes the literature review regarding customer profiling. Next, a literature review concerning a marketing strategy used for dealing with segmentation and customer profiling is presented.

2.4 Marketing strategy

To conclude Chapter 2, the marketing strategy associated with segmentation and customer profiling will be reviewed. For management to achieve its marketing objectives and support the core or marketing strategy, a marketing strategy or strategies must be developed (Jooste et al., 2012). According to Jooste et al. (2012), the formulation of the marketing strategy involves the following:

- *Target market/s selection*: Marketing is not about chasing any customer at any price. A decision must be made regarding which groups of customers (segments) are attractive to the organisation and match its supply capabilities.

- *Positioning the marketing offering*: The chosen marketing strategy includes a decision on the position within the market that the marketing offering is to occupy. Positioning is the process of designing an image and value proposition so that the customer within the selected target market can understand what the organisation or brand stands for in relation to its competitors. Positioning is the battle for the customer's mind, as the customer's perception of the company or its brands will determine success or failure. Positioning is therefore a fundamental element of the marketing strategy, since any decision on positioning has a direct effect on the determination of the marketing value mix.

- *Marketing value mix decisions*: The marketing value mix has been defined as the marketing management elements an organisation can coordinate and control when adopting a position in the selected target markets. Marketing managers have four broad tools they can use when matching their offerings to what customers require. These decisions consist of judgements about price levels, the blend of promotional

techniques, the distribution channels and service levels to use, and the types of products to manufacture. Normally it is not feasible to conquer the competition in every way. The elements of the marketing mix have been categorised by Jooste et al. (2012) under the headings of the modern approach to marketing as shown in Table 2.5.

Table 2.5: Elements of marketing value mix

Provide the value	**Communicate the value**	**Deliver the value**
Product planning	Advertising	Distribution channels
Branding	Personal selling	Physical handling
Packing	Direct marketing	Servicing
Pricing	Publicity	Promotions

According to Lynn (2012), almost any textbook states that the key to marketing success can be summed up by the STP (Segmenting, Targeting, Positioning) strategy, with segmentation being seen as the starting point to this success (Weinstein, 2013).

2.5 Synthesis: Chapter 2

The literature review for this research was initiated by discussing the segmentation process. This process consists of market segmentation as well as customer segmentation. It was clear from the literature that segmentation is constantly under investigation by researchers. The focus of this research is to create customer profiles from customer segments, for various markets. The markets will be simulated datasets that contains customer information. Thus, market segmentation as well as customer segmentation will not be applied as traditionally defined, but rather in a customised manner.

Next, customer profiling was explored. There is, however, a lack in the literature of another view on profiling. As mentioned in the literature, profiling has broadly two categories, namely *demographic* and *behavioural profiling*, and the attributes of both were discussed in this chapter. It was observed that there is not sufficient literature regarding customer *super* profiling available, which inspired the researcher to investigate the merits of combining both categories. This combination leads to the concept of super profiling.

This literature review also provides a solid foundation and understanding of basic marketing concepts, and presents areas for an industrial engineering researcher to explore. Next, a summary will be presented to conclude this chapter.

2.6 Summary: Chapter 2

In this chapter a literature review regarding segmentation, customer profiling and a marketing strategy was presented in order to pursue Objective 2.

Through the completion of this chapter it was established that there are millions of unique customers worldwide, covering many potential marketing segments, and marketing to every individual will be very difficult and probably even impossible. The best approach is to identify the sizeable groups within the customer base with shared characteristics/attributes (habits, tastes), referred to as segments, and then perform customer profiling (age, gender, income) on the customers in these different segments. According to literature, marketing success can be summarised by performing the STP strategy (Lynn, 2012), with segmentation being seen as the starting point (Weinstein, 2013).

Having discussed segmentation and customer profiling, the next topic to discuss is Big Data Analytics, which enables these concepts on massive datasets.

Chapter 3

Big Data Analytics

In the previous chapter an overview of topics related to segmentation and customer profiling was provided. The research in this chapter is based on *Big Data Analytics*. The first section of this chapter will be devoted to *data*, followed by *Big Data*, to get a clear understanding of what the two terms entail. After that, Big Data Analytics will be systematically reviewed. The review process will be initiated by providing a brief overview of what Big Data Analytics entails and how it differs from traditional data analytics. The section that follows will provide understanding and insights into various data analysis *processes*. This chapter will conclude by discussing *data mining* and the appropriate tools and techniques associated with it. These sections will provide the necessary background for the modelling approaches that will be adopted later in this research. Together, Chapters 2 and 3 fulfil Objective 2.

3.1 Data

According to Tien (2013), it will be helpful to first define the term *data*. It can be defined as "values of qualitative or quantitative variables, belonging to a set of items." The Oxford Dictionary defines data as "facts and statistics collected together for reference or analysis" (Oxford University Press, 2017). Data are the things assumed as facts which form the basis of reasoning or calculations. Data that have not yet been processed for use, are called *raw data*. There is often a distinction made between data and information, for information is the end product of data processing (Rouse, 2009).

According to Jiang et al. (1999), there are four kinds of data forms:

1. Textual data forms: This data form is used to represent texts, documents or social media posts. It can be seen as a collection of infinite characters that often do not follow a strict structure or format.

2. Temporal data forms: Time-series data, which represent data that varies with time (such as historical data), are stored in temporal data forms.

3. Transactional data forms: This data form is commonly represented by a list of items that were purchased in past market transactions. Transactional data forms describe an event, and therefore they can also contain a time dimension and refer to one or more objects.

4. Relational data forms: This data form is the most widely used data form, which can store different kinds of data. The data in relational data forms are presented in organised tables (relations). Each row of the table represents a record and each column an attribute or property of that record. Each attribute can also assume a different data type.

These four types of data forms are all within the context of data analytics or data mining.

Data types may be grouped into two main categories according to the properties of the underlying variables, namely qualitative (categorical) data and quantitative data. Quantitative data are described as data that are located on a numerical scale, such as the speed of a car or the academic averages of a student. While all quantitative data are numeric, not all numeric data are quantitative. For example, the national identification number of an individual is numeric, but not quantitative. On the other hand, qualitative data may be classified into categories based on the inherent characteristics of the objects described by the data. Examples would be a five-star hotel rating system or the gender of a person. Qualitative and quantitative data may be further divided into the categories shown in Figure 3.1.

Qualitative data may be broken down into *nominal* data and *ordinal* data. According to Hastie et al. (2009), nominal data contains two or more categories which may or may not be arranged in a meaningful sequence, but which cannot be quantified or ranked. If two mutually exclusive categories exist within the data, they may be referred to as *binary* or *dichotomous* nominal data. For example, the gender of a person can be either

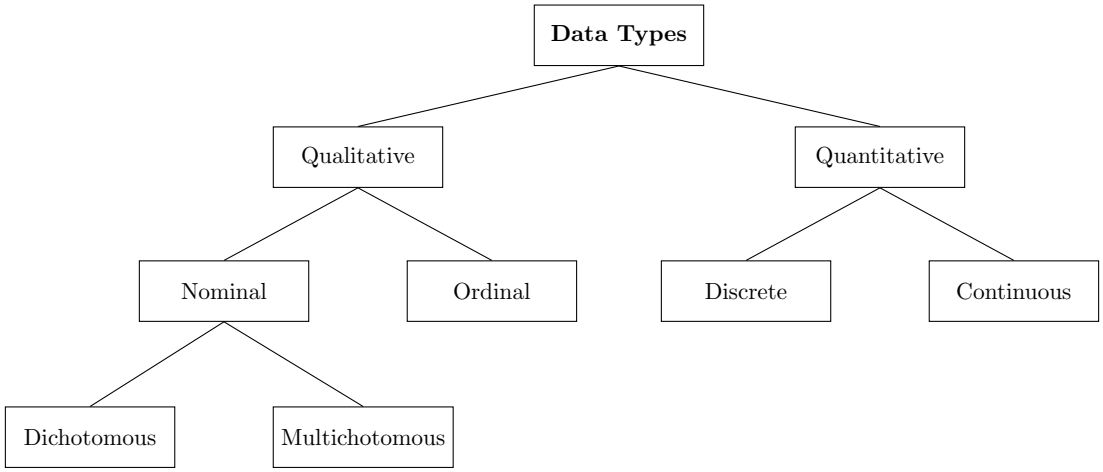


Figure 3.1: Taxonomy of data types (Steynberg, 2016).

male or female. Any other nominal data that exist with more than two categories are described as *multichotomous* data. An example of multichotomous data is an attribute describing a province, for it contains multiple categories (Western Cape, Eastern Cape, *etc.*) which cannot be ranked or quantified. On the other hand, an attribute describing customer satisfaction as *very satisfied*, *satisfied*, *unsatisfied etc.* consists of ordinal data, since these categories contain an intrinsic rank order.

There are also two types of quantitative data that may be further broken down into *discrete* and *continuous* data. Discrete data can assume any countable number of values, or values which are isolated and separated by gaps (such as the number that may be rolled on a die: 1, 2, 3, 4, 5 or 6). Furthermore, continuous data are measured along a continuum and may therefore take on an uncountable number of values (such as the temperature of a specific location).

The classification provided in Figure 3.1 will be employed throughout this research. This section introduced the concept of data and described the various categories into which data may be classified. The section that follows will provide an overview of what is meant when talking about massive volumes and huge variety of data, also known as *Big Data*.

3.2 Big Data

Two of the most famous Big Data pioneers are Billy Beane and Nate Silver. Beane, who was an American professional baseball player, popularised the idea of correlating various statistics with underestimated player traits in order to field an Oakland A's baseball team cheaply so that they could compete with teams like the Yankees.

Meanwhile, the effect that statistician Nate Silver had on forecasting Major League Baseball player performances was so strong, that people who did not believe his predictions created all sorts of analysis-free zones, such as 'Unskewed Polls'. Many think Silver is only a polling expert, but he is also a master at Big Data analysis (Harvey, 2017).

According to Prasad (2016), Big Data can be seen as any *voluminous* amount of structured, semi-structured and unstructured data that has the potential to be mined for information where the individual records stop mattering and only aggregates (collection of records) matter. It can be said that data becomes Big Data when it is difficult to process using traditional techniques. As stated by Bühlmann et al. (2016), conceptual confusion seems inevitable when referring to Big Data.

Similar to the previous description of Big Data, Gupta and George (2016) stated that the term Big Data is often used to describe massive, complex and real-time streaming data requiring sophisticated management, analytical and processing techniques to extract insights. Although there is no consensus on the definition and characteristics of Big Data, the term "Big Data" was initially formulated to reflect the *bigness* or *voluminous* size of data generated as a result of using new forms of technology (*e.g.*, social media, smart phones, sensors and radio-frequency identification (RFID) tags) (Gupta and George, 2016). The definition of Big Data was then extended to include *variety* (structured or unstructured data formats) and *velocity* (the speed at which data is created).

There are several dimensions or characteristics of Big Data, these dimensions are referred to as the V's. Erevelles et al. (2016) and Wang et al. (2016) refer to the 3V's, which include *volume*, *velocity* and *variety*. Furthermore, Wang et al. (2016) mention that there are other V's, such as *value*, *veracity*, *variability* and *virtual*, that appear in literature to serve as complementary features of Big Data. In like manner Gupta and George (2016) also indicated that over the years Big Data was further dimensionalised into *veracity* (messiness of data) and *value* (previously unknown insight).

To keep things straightforward, Zikopoulos et al. (2012) typically define Big Data by using four V's, namely, volume, variety, velocity and veracity. The veracity characteristic was recently added in response to the quality and source issues clients began facing with Big Data initiatives. Embracing Big Data leads to the adoption of new technologies to complement the traditional approach to data management and allow the utilisation of data:

- in different formats (*variety*).
- of high *volume*.
- entering system(s) at high rates (*velocity*).
- that users still try to clean, manage and maintain quality in (*veracity*).

Some analysts include other V-based descriptors, such as variability and visibility, that will not be included in the discussion to follow.

Volume is the obvious Big Data trait (Erevelles et al., 2016; Prasad, 2016; Rajaraman, 2016; Russom, 2011; Tien, 2013; Wang et al., 2016; Zikopoulos et al., 2012). Big Data implies enormous volumes of data generated by sensors, machines combined with internet explosion, social media, e-commerce, GPS devices, *etc.* (Prasad, 2016). According to Erevelles et al. (2016), the volume of Big Data is currently measured in petabytes, exabytes or zettabytes. One petabyte is equal to twenty million traditional filing cabinets of text. Walmart is estimated to create 2.5 petabytes of customer data every hour (Erevelles et al., 2016).

The *variety* characteristic of Big Data aims to capture all of the data that can have an effect on the decision-making process (Zikopoulos et al., 2012). According to Prasad (2016), variety implies to the type of formats and these formats are classified into three types:

- Structured: MySQL, Legacy files *e.g.* Microsoft Excel, Microsoft Access.
- Semi-structured: Emails, Tweets, User reviews.
- Unstructured: Photos, Videos, Audio files.

Many sources of Big Data provide a diverse richness that is superior to traditional data from the past (Erevelles et al., 2016). A major difference between contemporary Big Data and traditional data is the shift from structured transactional data to unstructured behavioural data.

Velocity is one of the favourite Big Data characteristics of Zikopoulos et al. (2012), but it is the least understood. It is defined as the rate at which data arrives at the enterprise and is processed or well understood. Velocity refers to the rate at which data is pouring in; for example Facebook users generate three million 'likes' per day (Prasad, 2016). According to Erevelles et al. (2016), marketing executives with access to rich, insightful, current data are able to make better decisions based on the evidence at a given time, rather than making decisions on intuition or laboratory-based consumer research.

Veracity is a term that is recently being used more frequently to describe Big Data. It refers to the quality and trustworthiness of the data (Zikopoulos et al., 2012). According to Prasad (2016), veracity refers to the biases, noise and abnormality in data. If meaningful data is needed, the initial step would be to cleanse the data. Veracity highlights the importance of being aware of data quality (Erevelles et al., 2016). The veracity of Big Data can be a major issue at times where the volume, velocity and variety of data are constantly increasing (Erevelles et al., 2016).

Big Data appears to be a notoriously difficult concept, with several dimensions and connotations, and it is therefore functionally vague (Bühlmann et al., 2016). Typically the term Big Data refers to more, or too much data than what is traditionally known as data, or which can be managed, accessed, analysed, interpreted and validated by traditional means, as a basis for useful information or reliable knowledge (Bühlmann et al., 2016).

At the beginning of the twenty-first century, the growing volumes of data presented seemingly inexplicable problems and challenges, and storage and CPU technologies were overwhelmed by the terabytes of data being generated (Tien, 2013). Figure 3.2 illustrates that the available data grows in all dimensions, and this availability of data has overloaded the capability to analyse data, as well as the capability to use the analysis; either to run or store analysis (computing and storage capability) (Kalickỳ, 2013).

A *knowledge gap*, as shown in Figure 3.2, expresses the inability to analyse data due to the limited analytical techniques. These techniques include data mining algorithms, natural language processing, *etc.* An *execution gap* expresses the inability to utilise analysis

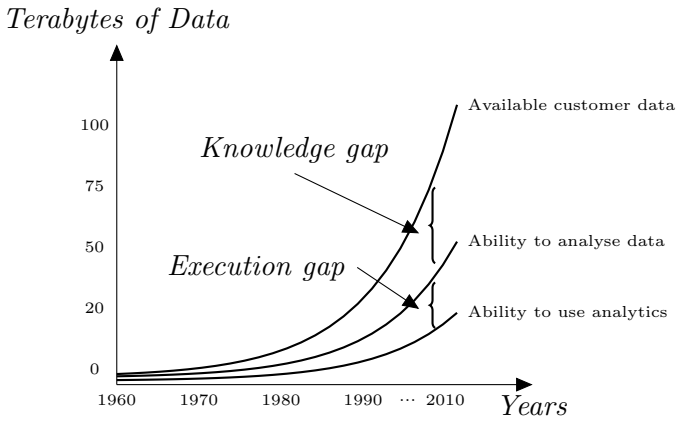


Figure 3.2: Data volume challenge (Kalickỳ, 2013).

due to the limited availability of resources. These resources include processing units and data storage. These gaps are getting smaller and smaller as technology evolves (execution gap) and mature analytics (knowledge gap) are applied to the datasets. Nowadays, Big Data is not a technical or storage problem, but has become a competitive advantage (Kalickỳ, 2013).

This section provided a brief introduction and discussion of Big Data and assisted in identifying databases that are referred to as Big Data. The terms volume, variety, velocity and veracity are used to provide ways to understand and classify a Big Data opportunity. Big Data can help to make the right decisions at the right time. The success of an organisation depends not only on how well the business is doing, but also on how well the organisation can *analyse* their data and derive insights about their company, their competitors, *etc.* Tools that assist the transformation of raw voluminous data into Big Data with trustworthy insights and to discard the noise is called *Big Data Analytics*. Next, an introduction and understanding regarding Big Data Analytics will be provided, as well as a framework regarding Big Data Analytics.

3.3 Big Data Analytics

Analytics is not a physical tool or technology, it is rather a way of thinking and acting (Prasad, 2016). Data analytics is the *process* of examining datasets in order to draw conclusions about the information they contain (Rouse, 2009). Therefore, Big Data Analytics, which is a relatively new term, describes the data analysis of Big Data. As

mentioned earlier, Big Data is any dataset which cannot be analysed with conventional tools (Prasad, 2016).

USMA (2017) defined Big Data Analytics to be the entire methodology that is utilised for the analysis of Big Datasets, in order to create value for an enterprise. This definition is in line with that of Russom (2011), which stated that Big Data Analytics is a methodology that is followed when advanced analytic techniques operate on large datasets. Big Data Analytics contains two elements, namely Big Data and Analytics, plus how the two terms have merged to create one of the most profound trends in business intelligence.

Analytics can be applied to various problems and in different industries. It is thus important for organisations to take time to understand the scope of analytics in their business. According to Prasad (2016), analytics can be classified into three broad groups:

1. Based on the industry.
2. Based on the business functions.
3. Based on the kind of insight offered.

Industries where analytics usage is very common would be industries which create a huge amount of data, for example, credit card companies and consumer goods vendors. These companies were among the first to adopt analytics (Prasad, 2016). Analytics can also be classified on the basis of the business functions it is used in. For example, marketing analytics, sales and HR analytics and supply chain analytics. However, the most popular way to classify analytics is on the basis of what it allows us to do.

Data analytics is concerned with the extraction of actionable knowledge and insights from Big Data (Rajaraman, 2016). To do so, a hypothesis needs to be formulated that is often based on speculation gathered from experience and discovering correlations among variables. Rajaraman (2016) stated that there are four types of data analytics, which include:

1. *Descriptive analytics*: This type of analytics essentially stated what happened in the past and presents it in a easily understandable form. Data gathered is organised as bar charts, graphs, pie charts, maps, scatter diagrams, *etc.* This is done to aid visualisation, which gives insight into what the data implies. This form of data presentation is often referred to as a *dashboard*. Examples of descriptive analytics

include presentations of population census data which classifies the population across a country by age, gender, education, income, population density and similar parameters.

2. *Predictive analytics*: The aim of predictive analysis is to be able to inform what is expected to happen in the near future by judging available data. Tools used for making these judgements include time series analyses using statistical methods, neural networks and machine-learning algorithms. One major use of predictive analytics is in marketing, by anticipating customers' needs and preferences. Another use includes the managing of election campaigns, by collecting a variety of data such as the composition of the electorate in various locations, and the perception of their requirements such as infrastructure and local issues.

3. *Diagnostic analytics*: This analytics type aims to determine the cause of a phenomenon that occurred in the past by using questions that focus on the reason behind the event (Erl et al., 2015). Diagnostic analysis is utilised to determine what information is related to the phenomenon in order to enable answering questions that seek to determine why something occurred, therefore some researchers refer to this type of analytics as exploratory or discovery analytics. The collection of data, from a variety of sources and analysis of the data provides additional opportunities for insights and unforeseen discoveries. Therefore, companies utilise customer feedback, tweets, blogs and sales trends, in order to discover patterns within their customers' behaviour. Based on the customers' behaviour, it may be possible for companies to forecast their actions, which then becomes exploration/discovery analytics. These actions include renewing a magazine subscription, changing a mobile phone service provider and cancelling a hotel reservation. A company then has the opportunity to formulate an attractive offer to try to change the customer's anticipated action.

4. *Prescriptive analytics*: This analytics type has the ability to identify opportunities to optimise solutions to existing problems, based on data that was gathered. Thus, the analysis has the ability to determine and notify a business how to achieve a goal. A common use is in airlines' pricing of seats in order to maximise the profit. It is based on historical data of travel patterns, popular origins and destinations, major events, holidays, *etc.*

To summarise the findings regarding data analytics, Figure 3.3 provides perspective on what the most frequently used analytics are, as well as what aspect within data they address.

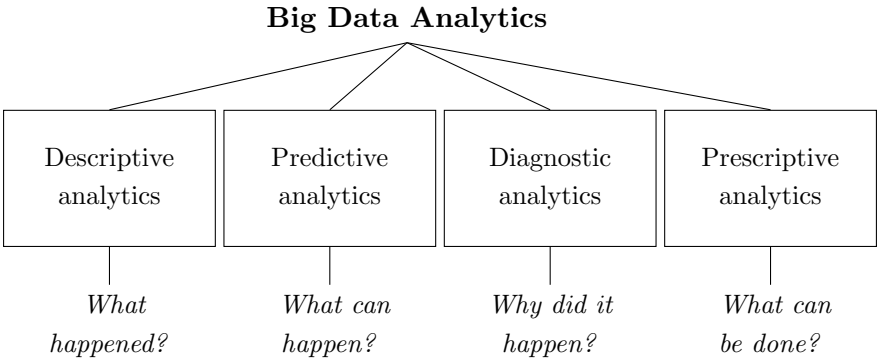


Figure 3.3: The four types of data analytics (Corcoran, 2015).

According to Minelli et al. (2012), market-leading companies are using Big Data Analytics to improve sales revenue, increase profits and better serve customers. The enterprises that have become skilled in Big Data Analytics will be able to simultaneously minimise operational costs while driving top-line revenues to net substantial profit margins for their enterprise (Minelli et al., 2012).

Big Data Analytics utilises a wide variety of advanced analytics, as presented in Figure 3.4, to provide:

1. *Deeper insights*: Instead of looking at segments, classifications, regions, groups or other summary levels, deeper insight provides knowledge of the individuals, products, parts, events, transactions, *etc.*

2. *Broader insights*: Operating a business in a global, connected economy is very complex, given constantly evolving and changing conditions. Big Data Analytics takes into account all the data, including new data sources, to understand the complex, evolving, and interrelated conditions to produce more accurate insights.

3. *Frictionless actions*: Increased reliability and accuracy that will allow the deeper and broader insights (mentioned above) to be automated into systematic actions.

To conclude this section, it can be said that unlike traditional analytics, Big Data Analytics is not constrained by predefined sets of questions or queries. According to Minelli

SQL analytics	Descriptive analytics	Data mining	Predictive analytics	Simulation	Optimisation
Count Mean OLAP	Univariate distribution Central tendency Dispersion	Association rules Clustering Feature extraction	Classification Regression Forecasting Spatial Machine learning Text analytics	Monte-Carlo Agent-based modelling Discrete event modelling	Linear optimisation Non-linear optimisation

Figure 3.4: Illustration of the wide variety of the analytics spectrum (Minelli et al., 2012).

et al. (2012), gathering data is often easier than figuring out how to use it. Therefore, with Big Data Analytics, organisations can receive the answers to their questions faster.

USMA (2017) created a framework as seen in Figure 3.5. This framework was constructed while attending a workshop, and provides an understanding of the term Big Data Analytics. The framework schematically indicates that Big Data Analytics contains various processes. Each process consists of numerous steps or phases. Each of the processes is initiated by data preparation steps or phases, seen in the first row of Figure 3.5. Following this step, each process also contains a data mining step or phase, as seen in the second row of Figure 3.5. The data preparation phase involves two main steps, namely *data cleaning* and *data transformation*. Exploring the data mining phase, USMA (2017) stated that data mining consists of various tools and techniques/tasks, which are collectively known as *machine learning*. Analytical techniques typically utilise large datasets. These datasets require techniques that can easily scale the dataset and provide results to increase company revenue (Bell and Mgbemena, 2018). The sections that follow within this chapter will be structured according to this framework diagram (Figure 3.5). Next, three of the Big Data Analytics processes will be discussed. To initiate this discussion, the definition of a *Data Analytics process*, as interpreted by USMA (2017), will be provided. This will be followed by the discussion of the first process namely, *Knowledge Discovery in Databases*.

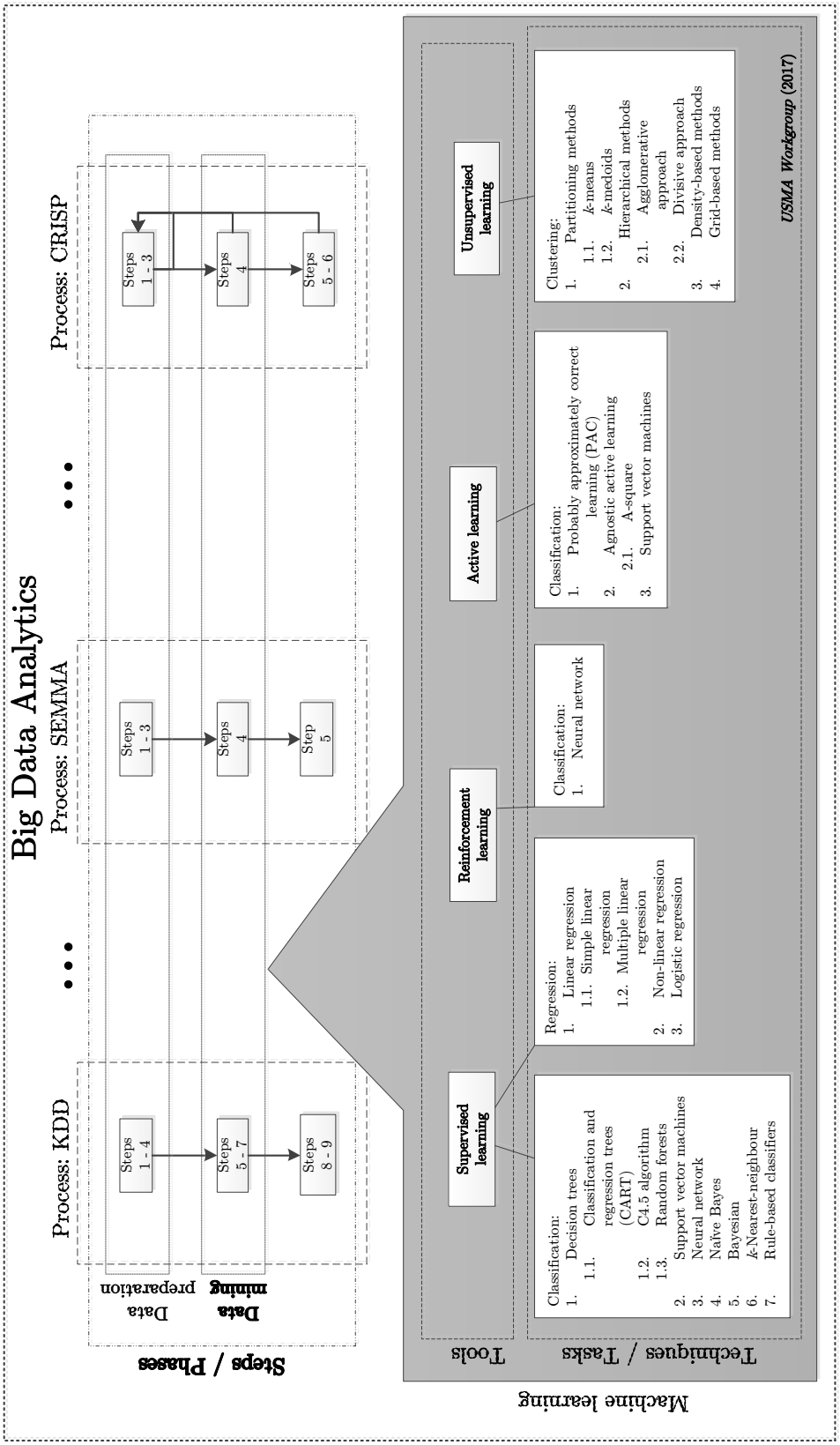


Figure 3.5: Illustrating Big Data Analytics (USMA, 2017).

3.4 Data Analytics processes

As seen in Figure 3.5, Big Data Analytics include various processes. For the purpose of this research, Knowledge Discovery in Databases (KDD), Sample, Explore, Modify, Model and Access (SEMMA) and Cross-Industry Standard Process (CRISP) have been selected to discuss, as they are considered to be the most popular. A process implies that the KDD, SEMMA and CRISP comprises many phases/steps, all repeated in multiple iterations (Mariscal et al., 2010). USMA (2017) defined a Big Data Analytics process, illustrated in Figure 3.5, as being a series of phases/steps that are followed in order to perform Big Data Analytics.

3.4.1 Process: Knowledge Discovery in Databases

Feyyad (1996) uses the term *Knowledge Discovery in Databases* (KDD) to denote the *overall* process used to extract high-level knowledge out of low-level data. This process is the first process shown in Figure 3.5. According to Fayyad et al. (1996), the term KDD was formulated at the first KDD workshop in 1989, to emphasise that knowledge is the end product of a data driven discovery. It is synonymous with large databases. A simple definition for KDD is as follows: *Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data* (Feyyad, 1996). Major KDD application areas include marketing, fraud detection, manufacturing and telecommunications.

Hamilton (2012) provides an outline of the steps of the KDD process as seen in Figure 3.6. The overall process of finding and interpreting patterns from data involves the repeated application of the steps seen in Figure 3.6. According to Feyyad (1996) and Hamilton (2012) the steps are as follows:

1. *Developing* an *understanding* of the application domain, the relevant prior knowledge, and the goals of the end user.

2. *Creating* a target set, selecting a dataset, or focusing on a subset of variables or data samples, on which discovery is to be performed.

3. Data *cleaning* and *preprocessing*.

4. Data *reduction*, *transformation* and *projection*.

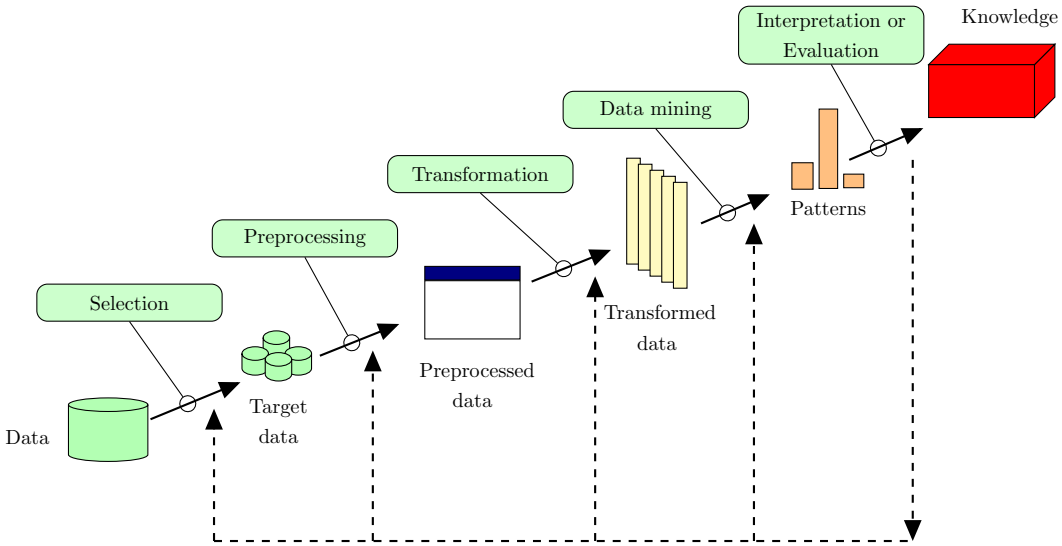


Figure 3.6: An overview of the KDD process (Hamilton, 2012).

5. *Selecting* the data mining technique/tasks.

6. *Selecting* the data mining algorithm(s), by matching the overall criteria of the KDD process (step 1) to a particular data mining task.

7. *Perform* data mining by searching for patterns of interest in a particular representational form or a set of such representations as classification, decision trees, *etc.*

8. *Evaluating* step 7 and interpreting mined patterns.

9. *Consolidating* discovered knowledge, incorporating this knowledge into the performance, or simply documenting the knowledge and reporting to users.

The data preparation phase is evident in the KDD process as steps 1–4, with step 3 as the data cleaning phase and step 4 as the data transformation phase. The KDD steps 5–7 form the data mining phase as indicated in the second row of Figure 3.5. In general, the terms KDD and data mining are used interchangeably by many (Feyyad, 1996; Mariscal et al., 2010). However, Fayyad et al. (1996) confirm that at the core of the KDD process, presented by Figure 3.6, is the application of data mining methods for pattern discovery. As stated by Fayyad et al. (1996), KDD is viewed as the overall process of discovering

useful knowledge from data, while data mining refers to a particular step in this process. Data mining will subsequently be discussed.

3.4.2 Process: Sample, Explore, Modify, Model, Assess process methodology

This is the second process shown in Figure 3.5. The *Sample, Explore, Modify, Model, Assess* process (SEMMA) was developed by the SAS Institute (Azevedo, 2008), which is seen as the leading company in business intelligence (BI) and it has the most comprehensive BI platform in the industry as well as the most advanced analysis capabilities (Mariscal et al., 2010).

The SAS Institute defines SEMMA as a logical organisation of the functional toolset of SAS Enterprise Miner for carrying out the core tasks of data mining (Mariscal et al., 2010). This process focuses on the model development aspects of data mining. Figure 3.7 illustrates the SEMMA process, and the SAS Institute considers a cycle with five phases for the process. These stages include:

Figure 3.7: An overview of the SEMMA process (Adapted from: Azevedo (2008); Mariscal et al. (2010).)

1. *Sample*: This stage consists of sampling the data by extracting a portion of a large dataset big enough to contain the significant information, yet small enough to manipulate quickly. This stage of the process is optional.

2. *Explore*: This stage consists of the exploration of the data by searching for unanticipated trends and anomalies, this is done to gain understanding and ideas.

3. *Modify*: The modification of the data by creating, selecting and transforming the variables in preparation for data modelling. Step 3 includes both the data cleaning and data transformation phases, whereas step 3 in the KDD process was observed to be the data cleaning phase and step 4 as the data transformation phase.

4. *Model*: This stage entails the data modelling by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.

This step, step 4, is seen as the data mining phase within this process, whereas steps 5–7 in the KDD process were observed to be the data mining phase. This can be seen in Figure 3.5.

5. *Assess*: This stage consists of assessing the results, which is done by evaluating the usefulness and reliability of the findings from the data mining process and estimating how well it performs.

Even though the SEMMA process is seen as independent from a chosen data mining tool, it is linked to the SAS Enterprise Miner software and pretends to guide the user on the implementations of data mining applications (Azevedo, 2008). According to Shafique and Qaiser (2014), the SEMMA phases assist in the solving of business problems as well as helping businesses to reach their goals.

3.4.3 Process: Cross-Industry Standard Process

This is the third and final process that will be discussed and is shown in Figure 3.5. *Cross-Industry Standard Process* (CRISP) was first suggested in the 1990s by a European consortium of companies as a standard process model for data mining (Azevedo, 2008). According to Mariscal et al. (2010), CRISP is the most commonly used methodology when developing data mining projects, and provides a framework for carrying out data mining activities (Giraud-Carrier and Povel, 2003; Tomar and Agarwal, 2013). However, its use is not becoming any more widespread due to rivalry with other in-house processes developed by work teams and the SEMMA process. This decrease in the use of CRISP is due to the fact that it just defines *what to do*, and not *how to do it* (Mariscal et al., 2010). The CRISP process clarifies what must be done and contributes to the speed, reliability and efficiency of projects. The CRISP process has six phases as seen in Figure 3.8.

The steps of the CRISP is as follows:

1. *Business understanding*: The initial phase focuses on understanding the project objectives and requirements from a business perspective; thereafter converting the knowledge into a data mining problem definition and a preliminary plan to achieve the objectives.

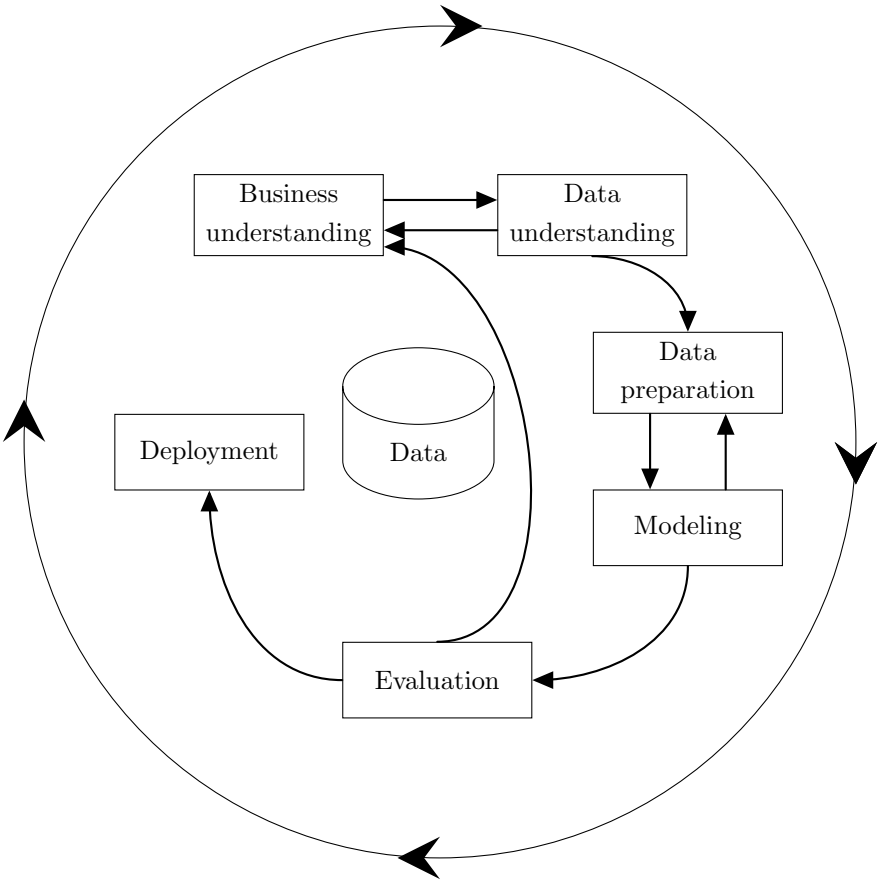


Figure 3.8: An overview of the CRISP (Azevedo, 2008; Mariscal et al., 2010).

2. *Data understanding*: Collect all data, explore data using graphics or basic statistics and determine what relationships exist in the data.

3. *Data preparation*: This phase covers all activities to construct the final dataset from the initial raw data. Take note that this phase is very time-consuming. Similar to the SEMMA process, step 3 of the CRISP includes both the data cleaning and transformation phases, as opposed to the KDD process where the data cleaning phase is evident as step 3 and the data transformation as step 4.

4. *Modelling*: Various modelling techniques are selected and applied and their parameters are calibrated to optimal values. This step (step 4) is seen as the data mining phase within this process, as opposed to steps 5–7 of the KDD process and step 4 of the SEMMA process, as seen in Figure 3.5.

5. *Evaluation*: Review the model(s) and determine which model or collection of models best satisfies or answers the business goals and objectives.

6. *Deployment*: The creation of the model is generally not the end of the project, even if the purpose of the model is to increase knowledge of data. The knowledge gained will need to be organised and presented in a way that customers can use it.

This concludes the discussion of three Big Data Analytics processes (as seen in Figure 3.5). The researcher acknowledges that there are more processes, but found that these three are the most popular processes discussed in literature. A comparative study of the three processes will follow. This is performed to provide more clarity and to establish a parallel between the three processes.

3.5 Comparative study of the analytics processes

This section will compare the KDD process against the SEMMA process and CRISP respectively. When performing a comparative study on the KDD and SEMMA processes, it could, on a first approach, be confirmed that these processes are equivalent to one another at the following stages:

Table 3.1: Summary of the correspondences between the KDD and SEMMA processes (Azevedo, 2008).

KDD	**SEMMA**
Pre KDD	———
Selection	Sample
Preprocessing	Explore
Transformation	Modify
Data mining	Model
Interpretation or Evaluation	Assessment
Post KDD	———

Table 3.1 provides a summary of the correspondence between the KDD and SEMMA processes. After thorough examination, it may be affirmed that the five phases of the SEMMA process can be seen as a practical implementation of the five corresponding

phases of the KDD process, since they are directly linked to the SAS Enterprise Miner software (Azevedo, 2008).

Table 3.1 states the following findings:

1. The Selection phase of the KDD process can be identified with the Sample phase of the SEMMA process.
2. The Preprocessing phase of the KDD process can be identified with the Explore phase of the SEMMA process.
3. The Transformation phase of the KDD process can be identified with the Modify phase of the SEMMA process.
4. The Data mining phase of the KDD process can be identified with the Model phase of the SEMMA process.
5. The Interpretation or Evaluation phase of the KDD process can be identified with the Assessment phase of the SEMMA process.

The comparison of the KDD phases with the CRISP phases is not as straightforward as in the SEMMA process situation. However, a similar table can be provided for this comparison, as seen in Table 3.2.

Table 3.2: Summary of correspondences between KDD and CRISP processes (Azevedo, 2008).

KDD	**CRISP**
Pre KDD	Business understanding
Selection	Data understanding
Preprocessing	
Transformation	Data preparation
Data mining	Modelling
Interpretation or Evaluation	Evaluation
Post KDD	Deployment

Table 3.2 states the following findings:

1. The KDD process that includes the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user can be associated with the business understanding phase of CRISP.

2. The post KDD phase can be associated with the deployment phase of CRISP.

These findings only entail the first and the last phases, therefore, the following can be stated of the remaining phases:

1. The combination of Selection and Preprocessing phases of the KDD process can be identified as the Data understanding phase of CRISP.

2. The Transformation phase of the KDD process can be identified with the Data preparation phase of CRISP.

3. The Data mining phase of the KDD process can be identified with the Modelling phase of CRISP.

4. The Interpretation or Evaluation phase of the KDD process can be identified with the Evaluation phase of CRISP.

Lastly, the comparison of the SEMMA process and CRISP can be seen in Table 3.3. The SEMMA process contains fewer steps than CRISP, therefore the Business understanding and Deployment phases of CRISP do not align with a SEMMA phase.

Table 3.3: Summary of correspondences between SEMMA and CRISP processes (Azevedo, 2008).

SEMMA	**CRISP**
———	Business understanding
Sample	Data understanding
Explore	
Modify	Data preparation
Model	Modelling
Assessment	Evaluation
———	Deployment

Table 3.3 stated the following findings:

1. Both the Sample and Explore phases of the SEMMA process can be associated with the Data understanding phase of CRISP.
2. The Modify phase of the SEMMA process can be associated with the Data preparation phase of CRISP.
3. The Model phase of the SEMMA process can be associated with the Modelling phase of CRISP.
4. The Assessment phase of the SEMMA process can be associated with the Evaluation phase of CRISP.

Taking into consideration the presented comparison analysis, it can be concluded that the SEMMA process and CRISP can be viewed as an implementation of the KDD process described by Feyyad (1996) and Azevedo (2008). At first impression it may seem that CRISP is more complete than the SEMMA process. However, analysing the SEMMA process in more detail leads to the integration of the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end user. This integration takes place in the sample phase (phase 1) of the SEMMA process, because the data cannot be sampled unless there is an understanding of all the presented aspects (Azevedo, 2008). With respect to the consolidation, by incorporating this knowledge into the system, it can be assumed that the knowledge will be present.

Thus, standards have been achieved concerning the overall process, *the SEMMA process and CRISP function as guidance on how to apply data mining in practice or in real systems* (Azevedo, 2008).

Next to be discussed is data cleaning, which forms part of the data preparation phase as seen in Figure 3.5. The data cleaning phase is evident in the KDD process as step 3, in the SEMMA process as step 3, and lastly in the CRISP as step 3.

3.6 Data cleaning

Data cleaning is an important phase of the *data preparation* process and is used to manipulate data into a form suitable for analysis (Salkind, 2010). The objective of data cleaning is to improve the quality of the data prior to the analysis by detecting and removing errors and inconsistencies. According to Nisbet et al. (2017) this entails

three primary activities: imputation (filling of black entries), handling error values and treating outliers.

The presence of '*dirty data*' reduces the reliability and validity of the associated data analysis. If responses or entries are missing or erroneous, they will not be reliable over time. Reliability sets the upper bound for validity, as unreliable entries reduce the validity Salkind (2010). Table 3.4 contains typical examples of such entries. The first column names the variable or attribute for which data was entered, the second column shows the '*true data*', or the way the entry should have looked, while the last column contains the data as it actually exists in the system. Data are *missing* in the attributes *ethnicity* and *annual income*, whereas the *date* and *place of birth* entries are *erroneous*, since they are incomplete or not entered as per the required format. Lastly, although the data entry for the *gender* of this record is in the correct format, it does not represent the correct value which is associated with this record. Erroneous data entries may therefore be grouped into two sets, namely those which are *valid* and those which are *invalid* (Bramer, 2007). An example of *outliers*, the third type of 'dirty data' entries, is not contained in the table. Consider the *annual income* attribute to illustrate this concept. If the typical range for this value is between \$10 000 and \$80 000, data entries of \$10 or \$6 000 000 would typically constitute outliers.

Table 3.4: Example of data errors and missing data (Salkind, 2010).

Variable	***'True data'***	***Incomplete, incorrect, or missing data***
Name	Maria Margaret Smith	Maria Smith
Date of birth	2/19/1981	1981
Gender	F	M
Ethnicity	Hispanic and Caucasian	
Education	B.A., Economics	College
Place of birth	Nogales, Sonora, Mexico	Nogales
Annual income	\$50 000	

Various approaches to data cleaning have been documented in literature. The most common methodologies applicable to each of the three data cleaning activities will subsequently be discussed.

3.6.1 Missing values

Nisbet et al. (2017) accentuated the fact that many statistical algorithms used for prediction or classification can only be used if all attributes in a data record contain entries, therefore it is necessary to address missing values in a dataset.

The best approach is to fill the missing values with the correct information. During the data gathering process, procedures should be established to detect missing values and ask a person with relevant knowledge to fill in the missing data as soon as possible (Salkind, 2010). However, this is not always a practical solution. The data may no longer be retrievable or, in the case of an automated data cleaning process, a manual entry is typically not desired.

A second approach entails deleting the data record or row that contains a missing value (Bramer, 2007). In this case, only the remaining, complete data records are used in the analysis. If the proportion of records containing missing entries for a particular column is small, this would be a logical approach. Contrarily, when the proportion of records containing missing entries is large, this approach may result in a loss of data integrity, contributing to the exclusion of several records from the analysis, thereby constituting poor practice. A threshold may be set for the proportion of deleted records, below which the implementation of this method is acceptable. If the proportion of records containing missing entries for a specific attribute or column is significantly high, the column may be removed entirely as its contribution to the data analysis is deemed to not be of value.

A final approach is to use *imputation* techniques. Imputation refers to the process of replacing a missing value with a reasonable estimation. Several imputation methods exist, varying from *mean imputation* (replacing a missing data entry with the average or most commonly occurring value in the entire dataset) to *hot deck imputation* (making estimates based on a similar, but complete dataset), *single imputation* if the proportion of missing values is small, and *multiple imputation* if the proportion is large or the data are not missing at random. All methods of imputation are considered preferable to case deletion, which can result in a biased sample (Salkind, 2010).

3.6.2 Erroneous values

According to Hellerstein (2008), error values may result from one of four main activities. The first activity is *data entry* or human activities, which may result in typographic error or errors due to misinterpretation of raw data. The second activity is that of *measurement*, where errors may occur when measuring devices are incorrectly used or the entire measuring process is approached incorrectly. The third activity is *distillation*, during which errors may occur during the preprocessing and summarising of data before they are added to the dataset. The final activity leads to errors that may occur when data from various sources are integrated into a single database. This activity is referred to as *integration*.

These four activities should be considered carefully when the data capture or data gathering process is designed. For example, if entries are made by humans into a computer, *data validation* techniques can be implemented in order to reduce errors made during the first activity. If a character is entered to describe a numerical variable, or dots are used to separate day, month and year of a date entry when the required format differs (dashed or forward slashes), this can be detected by the computer and the user may be prompted to re-enter the data.

In the case of invalid erroneous data entries, such as the '*Date of birth*' entry in Table 3.4, these errors should be corrected if the true data is known, otherwise the entry should be deleted and treated as a missing data value. Valid erroneous data entries, however, such as the '*Gender*' entry in Table 3.4, cannot be detected by the system. In the case where this entry differs greatly from the other values contained in the dataset, for example if *1 000 kg* is entered as the mass of a person, the data values may be detected and treated as outliers.

3.6.3 Outliers

The third and final data cleaning process described by Nisbet et al. (2017) is the treatment of outliers. An outlier may be defined as "an important kind of deviation" and "an individual value that falls outside the overall pattern" (Moore et al., 2009). If a data point's inclusion in or removal from a base model has a considerable impact on the model outcome, it is called *influential observation*; this may also be a cause for concern (Steynberg et al., 2017).

Aguinis et al. (2013) divided the treatment of outliers present in a dataset into three steps, namely *definition*, *identification*, and *handling* of outliers or influential observations. Definition and identification of outliers may be achieved using a wide range of outlier detection algorithms, many of which are built into existing statistical software packages (Nisbet et al., 2017). These detection methods can be described as *univariate* (considering an extreme value for one variable) or *multivariate* (for unusual values concerning at least two variables), as well as *parametric* (statistical) or *non-parametric* methods that are model-free (Williams et al., 2002). Parametric methods are typically based on statistical estimates or assumptions derived from an underlying distribution of the data and flag values as outliers when they deviate from model assumptions. On the other hand, non-parametric methods are suitable for multi-dimensional databases of which there is no knowledge of underlying distributions. These methods are usually based on local distance measures of data points, or on clustering techniques in which clusters of small sizes are considered clustered outliers (Knorr et al., 2001).

Although the process of outlier detection is remarkably well documented in literature, no standardised procedure currently exists pertaining to their handling in data-related applications (Aguinis et al., 2013). Generally, outliers can be processed in one of two ways; they can either be kept or deleted. In studies which focus on the detection of unusual activity, such as fraudulent bank transactions or equipment failures in factories, outliers provide essential information. On the other hand, if the goal of the data analysis is to identify a pattern or typical behaviour of a system, outliers may cause confusion. Therefore, the treatment of outliers should be chosen with due consideration for the application of the specific task at hand (Nisbet et al., 2017).

Next, the final step of the data preparation phase (Figure 3.5) will be discussed, namely dimensionality reduction. This technique forms part of the data transformation step and is evident in the KDD process as step 4, while in the SEMMA process as well as in the CRISP, the data transformation stage is evident alongside the data cleaning phase as step 3.

3.7 Data transformation

After data cleaning has occurred, there may be cases where the data are not ready for mining. When this happens the data need to be *transformed* into forms appropriate

for mining. The data analysis process responsible for this occurrence is called *data transformation*.

3.7.1 Dimensionality reduction

Over the part few years there have been advancements in data collection. This has resulted in data being bigger than before and therefore the term Big Data was devised. Not only has the amount of data objects increased, so have their dimensions (Brázdil, 2016). However, as stated by Tang et al. (2014), having bigger data usually comes with higher noise. Also, having more measured features or variables does not always guarantee that all of them are important. Big dimensionality of data can cause problems, and is therefore referred to as '*the curse of dimensionality*'. This expression was coined by Richard E. Bellman and refers to various phenomena that arise when analysing and organising data in high-dimensional datasets (*i.e.* with number of dimensions more than 10) that do not occur in low-dimensional settings such as the three-dimensional space of everyday experience. This is where *dimensionality reduction* techniques provide assistance, and is usually performed prior to applying a data mining tool and techniques in order to avoid the effect of 'the curse of dimensionality' (Beyer et al., 1999; Brázdil, 2016).

Dimensionality reduction is one of the most popular techniques used to remove noisy (*i.e.* irrelevant) and redundant features (Tang et al., 2014). Reducing the number of data features (variables) can help to improve the learning performance, create better generalisable models, lower computational complexity, decrease required storage and help visualise the data. The goal of dimensionality reduction is to introduce high-dimensional data in a lower-dimensional subspace, while essential features of the original data are kept as far as possible (Kadhim et al., 2014).

Burges et al. (2010) and Napoleon and Pavalakodi (2011) defined dimensionality reduction as the mapping or transformation of high dimensional data into a lower, yet meaningful representation of reduced dimensionality that corresponds to the intrinsic dimensionality of the data.

USMA (2017) constructed Table 3.5 to provide an overview of all the dimensionality reduction techniques that were schematically represented in Figure 3.5. The dimensionality reduction techniques are indicated in column two of Table 3.5 with the applicable sources for each technique in column three, while lastly column four provides insight

into what each technique is capable of. Next, the best-known dimensionality reduction technique will be discussed, namely principal component analysis.

Table 3.5: Summary of dimensionality reduction techniques (USMA, 2017).

	Dimensionality reduction technique:	*Source*:	*Known for/Applications*:
1	**Dimensionality reduction**	Brázdil (2016) Burges et al. (2010) Carreira-Perpinán (1997) Fodor (2002) Silipo (2015) Tang et al. (2014)	Process of reducing the number of random variables under consideration, through obtaining a set of principal variables.
1.1	Principal component analysis (PCA)	Abdi and Williams (2010) Carreira-Perpinán (1997) Ding and He (2004) Fodor (2002) Jolliffe (2002) Napoleon and Pavalakodi (2011) Silipo (2015) Udell and Boyd (2014) Yeung and Ruzzo (2001)	It is a statistical procedure that seeks to reduce the dimensions of the data by finding a few orthogonal linear combinations (called the principal components) of the original variables, possibly correlated, with the largest variance. PCA can be generalised as correspondence analysis (CA) in order to handle qualitative variables and as multiple factor analysis (MFA) in order to handle heterogeneous sets of variables.
1.2	Factor analysis	Jolliffe (2002) Kim and Mueller (1978) Statistics Solutions (2017)	A technique used to reduce a large number of similar variables into a smaller number of factors (dimensions). This process is also called *identifying latent variables*. Since factor analysis is an explorative analysis, it does not distinguish between independent and dependent variables.

Continued on next page

	Dimensionality reduction technique:	*Source:*	*Known for/Applications:*
1.3	Self-organising maps (SOM)	Cho et al. (2005) Ghnemat and Jaser (2015) Ha et al. (2002) Kohonen (1998) Lee et al. (2004) Min and Han (2005) Ngai et al. (2009) Pratiwi (2012)	Implements an orderly mapping of a high-dimensional distribution onto a regular low-dimensional grid. SOMs accomplish two things, they reduce dimensions as well as display similarities.
1.4	Projection pursuit	Friedman (1987) Friedman and Tukey (1974) Hand (1998) Huber (1985)	The basic idea of projection pursuit is to assign a numerical index to every (one or two dimensional) projection that characterised the amount of the structure present (data density variation) in the projection. This index is then maximised (*via* numerical optimisation) with respect to the parameters defining the projections.

3.7.1.1 Principal component analysis

Principal component analysis (PCA) dates back to Karl Pearson in 1901 and is considered as one of the oldest, yet most widely used technique in data analysis (Pearson, 2010; Sorzano et al., 2014; Udell and Boyd, 2014). PCA is shown in Table 3.5, indexed as 1.1, as a dimensionality reduction technique.

The key idea of PCA is to find a new coordinate system in which the input data can be expressed with less variables without a significant error. The PCA algorithm is based on the search of orthogonal directions explaining as much variance of the data as possible (Sorzano et al., 2014).

In the clustering literature, PCA can also be applied to reduce the dimensionality of the dataset prior to clustering. When PCA is performed prior to clustering, it is anticipated that the principal components (PCs) may 'extract' the cluster structure in the dataset. Since the PCs are uncorrelated and ordered, the first few PCs, which contain most of the variations in the data, are usually used when clustering is performed. There are some common rules of thumb to choose how many of the first PCs to keep for clustering, however, most of these rules are informal and ad hoc. On the other hand, there are also theoretical results that indicated that the first few PCs, in some cases, do not contain appropriate cluster information. Yeung and Ruzzo (2001) showed that the first few PCs may contain less cluster structure information than other PCs.

As mentioned, the PCs are uncorrelated and ordered such that the k^{th} PC has the k^{th} largest variance among all PCs. The k^{th} PC can be interpreted as the direction that maximises the variation of the projections of the data points such that it is orthogonal to the first $k-1$ PCs (Yeung and Ruzzo, 2001).

For better understanding of PCA, the researcher decided to perform a built-in Matlab®[1] example. The dataset that will be used can be accessed in Matlab by loading '*hald*'. This dataset contains a 13-by-4 table, called ingredients. Table 3.6 shows the data that are contained by the ingredients table.

Looking at Table 3.6, it would be difficult to graphically visualise this data, for it contains four features/variables. It is desired to obtain at most three features, and then construct a three-dimensional graph of those three features. Therefore, PCA will be performed on this dataset to reduce the dimensions; yet keeping as much variance

[1]The registered trademark for Matlab® will from now on be omitted.

Table 3.6: Ingredients table of the "hald" dataset in Matlab

7	26	6	60
1	29	15	52
11	56	8	20
11	31	8	47
7	52	6	33
11	55	9	22
3	71	17	6
1	31	22	44
2	54	18	22
21	47	4	26
1	40	23	34
11	66	9	12
10	68	8	12

as possible. After performing PCA, the output for the PC coefficients, also known as loadings, can be seen in Table 3.7. The rows of the ingredients dataset correspond to the observations, and the columns correspond to the variables; therefore the coefficient matrix is only a four-by-four matrix (Table 3.7). In other words, row two, column two, (value = -0.0678) in Table 3.7 represents all the data points in column one of Table 3.6. Table 3.7 indicates both ingredient 1 and 2 have a negative projection, while ingredient 3 and 4 have a positive projection. This means that ingredient 1 and 2 have a negative correlation with the other two ingredients. To check this interpretation, it is useful to use a tool called *biplot*, in Matlab, which plots the data, along with the projections of the original features. Before plotting this data, it is important to check if the coefficient matrix is orthonormal, as well as determining the expected variance of each PC. The statistical implication of the orthonormal property is that the last few PCs are not simply unstructured leftovers after removing the important PCs. The last PCs have variances as small as possible and therefore they are useful in their own right. They can help to detect unsuspected near-constant linear relationships between the data matrix elements, as well as being useful in regression and in outlier detection (Jolliffe, 2013).

The coefficient matrix for this problem is orthonormal. This can be checked by multiplying the coefficient matrix with its own inverse, and the answer should be an

identity matrix of the same size.

Table 3.7: PC coefficients or loadings of the original data

	PC 1	*PC 2*	*PC 3*	*PC 4*
Ingredient 1	-0.0678	-0.6460	0.5673	0.5062
Ingredient 2	-0.6785	-0.0200	-0.5440	0.4933
Ingredient 3	0.0290	0.7553	0.4036	0.5156
Ingredient 4	0.7309	-0.1085	-0.4684	0.4844

Next, the percentage of variance explained by the corresponding PC can be determined. Table 3.8 indicates that 89.6 percent of the variance lies in PC 1, and 11.3 percent in PC 2. PC 1 and 2 account for 95.3 percent of the variance, therefore a two-dimensional graph, containing feature one and two would represent the majority (97.9 percent) of the variance of the ingredients dataset.

Table 3.8: Variance explained for each PC

86.5974
11.2882
2.0747
0.0397

When performing PCA, it also returns the PC *scores*. These scores are the representation of the original dataset (ingredients) in the PC space. The rows of scores correspond to the observations, and the columns to the components, the same as the original data. The scores are the data formed by transforming the original data (ingredients) into the space of the principal components.

To illustrate how to calculate the scores, the researcher will calculate $score_{1,1}$ (score value = 36.8218). Firstly, the original data (Table 3.6) need to be centred, by subtracting the column means for each value. For the first data values of all the ingredients it would

be,

$$\begin{aligned}
\text{Centred}_{1,1} &= 7 - \frac{\sum(7+1+11+11+7+11+3+1+2+21+1+11+10)}{13} \\
&= -0.46154 \\
\text{Centred}_{1,2} &= 26 - \frac{\sum(26+29+56+31+52+55+71+31+54+47+40+66+68)}{13} \\
&= -22.15385 \\
\text{Centred}_{1,3} &= 6 - \frac{\sum(6+15+8+8+6+9+17+22+18+4+23+9+8)}{13} \\
&= -5.76923 \\
\text{Centred}_{1,4} &= 60 - \frac{\sum(60+52+20+47+33+22+6+44+22+26+34+12+12)}{13} \\
&= 30
\end{aligned}$$

After centring all of the data points, the PCs are multiplied by their corresponding centred data points to get the score values.

$$\begin{aligned}
\text{Score}_{1,1} = &(-0.0678 \times -0.46154) + (-0.6785 \times -22.15385) + \\
&(0.0290 \times -5.76923) + (0.7309 \times 30) \\
= &\ 36.8218
\end{aligned}$$

This value is the same as the first value in Table 3.9. When using Matlab to perform PCA, there is no need to calculate the centred matrix and perform the multiplications, for PCA has its built-in algorithm in Matlab.

Figure 3.9 represents the *biplot* (previously mentioned), which illustrates the orthonormal principal component coefficients for all four of the variables (X1, X2, X3 and X4) by a vector, with the direction and length of the vector indicating how each variable contributes to the two principal components in the plot. This figure also illustrates the PC scores (Table 3.9, only columns one and two) for each observation (red dots) in a single plot. A two-dimensional plot was selected to illustrate the data, for the first two PCs account for the majority of the variance. The scores represent how much each data point relates to the component.

Interpreting the four variables (Figure 3.9), it shows that the first PC, which is on the horizontal axis, has positive coefficients for the third and fourth variable. Therefore, vectors X3 and X4 are directed into the right half of the plot. The largest coefficient in the first principal component is the fourth, labelled as variable X4. The second principal

Table 3.9: The PC scores

36.8218	-6.8709	-4.5909	0.3967
29.6073	4.6109	-2.2476	-0.3958
-12.9818	-4.2049	0.9022	-1.1261
23.7147	-6.6341	1.8547	-0.3786
-0.5532	-4.4617	-6.0874	0.1424
-10.8125	-3.6466	0.9130	-0.1350
-32.5882	8.9798	-1.6063	0.0818
22.6064	10.7259	3.2365	0.3243
-9.2626	8.9854	-0.0169	-0.5437
-3.2840	-14.1573	7.0465	0.3405
9.2200	12.3861	3.4283	0.4352
-25.5849	-2.7817	-0.3867	0.4468
-26.9032	-2.9310	-2.4455	0.4116

component, which is on the vertical axis, has negative coefficients for the variables X1, X2 and X4, and a positive coefficient for the variable X3.

In turn, when interpreting the score values (red dots on Figure 3.9) it indicates that the points near the left edge of the plot have the lowest scores for the first principal component, for they have negative values. This means that those data points do not relate that much to the first PC. The points are scaled with respect to the maximum score value and maximum coefficient length, so only their relative locations can be determined from the plot.

This concludes the discussion regarding the data transformation stage that includes dimensionality reduction techniques. After *cleaning* and *transforming* of the raw data it ought to be suitable for initiating the analysis.

Next to be discussed is data mining, and as seen in section 3.4 (as well as in Figure 3.5) the data mining phase is evident in the KDD process as step 5-7, in the SEMMA process as step 4, and lastly in the CRISP as step 4.

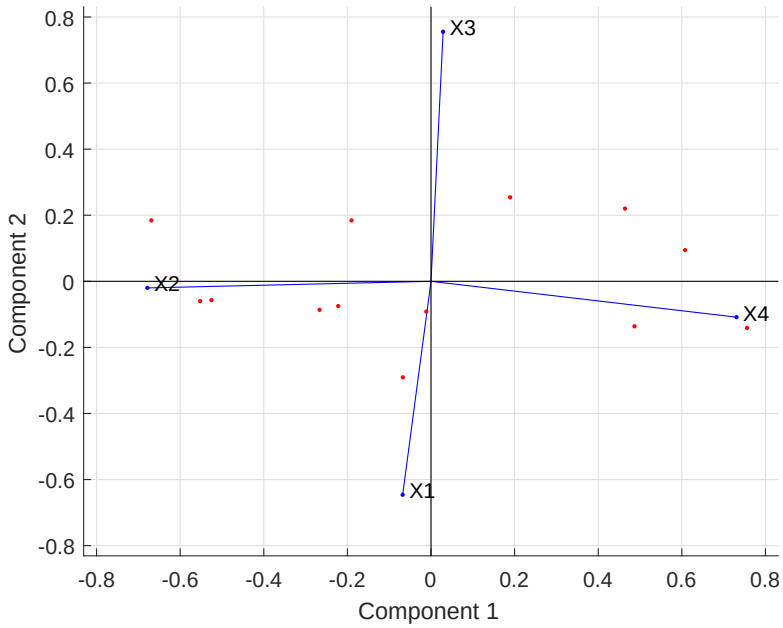


Figure 3.9: PCA plot for the example "hald"

3.8 Data mining

This section covers the data mining phase as shown in the second row of the schematics in Figure 3.5. Data mining consists of various tools and techniques/tasks, which are collectively known as machine learning. After defining what is meant by data mining, these tools and techniques will subsequently be discussed.

Shaw et al. (2001) referred to data mining as being the *process of searching and analysing data in order to find implicit (unspoken), but potentially useful, information.* This process involves *selecting*, *exploring* and *modelling* large amounts of data to uncover previously unknown patterns, and ultimately comprehensible information, from large databases (Shaw et al., 2001). As stated by Bell and Mgbemena (2018), data mining simply means extracting hidden knowledge from data, and is a popular method to utilise for understanding customer behaviour from raw data.

Similarly, Sharma (2014) stated that data mining is the process of analysing data from different perspectives and summarising the patterns, associations, or relationships among all the data to provide useful information.

According to Hand et al. (2001), data mining is the analysis of, often large, *observa-*

tional datasets to find unsuspected relationships and to summarise the data in unique ways that are understandable, as well as useful to the data owner. The relationships and summaries derived through data mining processes are referred to as *models* or *patterns* (Hand et al., 2001).

The definition above referred to *observational* data, as opposed to *experimental* data. This is because data mining typically deals with data that have already been collected for some purpose other than the data mining analysis (Hand et al., 2001). Therefore, it can be said that the objectives of data mining processes play no role in the data collection strategy. This is one way in which data mining differs from statistics, where data is often collected by using efficient strategies to answer specific questions. For this reason, data mining is often referred to as "secondary" data analysis (Hand et al., 2001).

The definition also mentions that the datasets that are examined in data mining are often large. Datasets that are referred to as large or complex, are datasets where traditional data processing application software is inadequate to deal with them. If only small datasets were involved it would merely be a discussion of classical exploratory data analysis as practised by statisticians. When large bodies of data are used, new problems arise (Hand et al., 2001).

According to Sharma (2014), data mining is an interdisciplinary field bringing together tools and techniques from machine learning, pattern recognition, statistics and visualisation to address the issue of information extraction from large databases.

As stated by Provost and Fawcett (2013), the data mining consists of tools and techniques, these tools include various learning techniques. Data mining uses a broad family of computational techniques that include classification, regression and clustering (Shaw et al., 2001). Although data mining techniques have been available for years, the advances in computers and software, in particular exploring techniques like visualisation and neural networks, have made data mining more attractive and practical to utilise (Shaw et al., 2001).

According to Sharma (2014), data mining is utilised for two main reasons:

1. There is too much data and too little information.

2. There is a growing need to extract useful information from data and to interpret the data.

Despite the large number of data mining algorithms developed over the years, there are only a few fundamentally different types of tasks these algorithms address (Provost and Fawcett, 2013). Data mining tasks are used to extract *patterns* from large datasets. Pattern extraction is an important component of any data mining activity and deals with relationships between subsets of data (Shaw et al., 2001). As stated by Shaw et al. (2001) a pattern can be formally defined as: *A statement S in L that describes relationships among a subset of facts F_s, of given set of facts F, with some certainty C, such that S is simpler than the enumeration of all facts in F_s.*

3.8.1 Supervised learning

As indicated in Figure 3.5, data mining entails various tools and techniques. These tools and techniques are collectively known as machine learning. The sections and subsections to follow will provide a better understanding of machine learning as schematically portrayed by USMA (2017) in Figure 3.5. Initiating the investigation of machine learning will be performed by discussing the first data mining tool, as shown in Figure 3.5, namely supervised learning (SL). According to Murphy (2012), SL is the form of machine learning most widely utilised in practice. SL is seen as a machine learning tool that is given a specific goal for grouping the data, for example to predict the target (Provost and Fawcett, 2013; USMA, 2017). In machine learning communities, prediction methods are referred to as SL (Rokach and Maimon, 2014). Therefore, when a question like *"Can we find groups of customers who have particularly high likelihoods of ending their subscription soon after their contracts expire?"* is asked, it defines a specific target of whether a customer will leave when their contract expires, and is referred to as a supervised data mining problem (Provost and Fawcett, 2013). Thus, any dataset used by machine learning algorithms is represented by using a set of features, if these features are given with known labels, the learning is called *supervised*.

According to Kotsiantis (2007), the aim of SL is to build a concise model of the distribution of class labels in terms of predictor features. In machine learning a dataset of observations called *instances* consists of a number of variables called *attributes*. SL is the modelling of these datasets containing *labelled* instances (Rechenthin, 2014). Each instance, in SL, can be represented as *(x,y)*, where x is a set of independent attributes (discrete or continuous) and y is the dependent target attribute. The target attribute y can be either *continuous* or *discrete*. However, when the category of modelling contains a

discrete target, it is a *classification* problem, but when the category contains a continuous target, it is a *regression* problem (Rechenthin, 2014).

For better understanding of a SL problem, the following table can be utilised for explanatory purposes. Table 3.10 demonstrates a dataset for SL with seven independent attributes, x_1,x_2,...,x_7, and one dependent target attribute, y. To be more specific, the attributes are defined as follows, x_1,$x_2 \in \{b,n\}$ and x_3,...,$x_7 \in \mathbb{R}$ and the target attribute $y \in$ {up,unchanged,down}. The attribute, *time*, is used to identify an instance and is not used in the model. The training and test datasets are represented in the same manner. However, where the training set contains a set of vectors of known label (y) values, the labels for the test sets are unknown.

Table 3.10: A supervised learning dataset (Rechenthin, 2014).

Time	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
09:30	b	n	-0.06	-116.9	-21.7	28.6	0.209	up
09:31	b	b	0.06	-85.2	-61	-21.7	0.261	unchanged
09:32	b	b	0.26	-4.4	-114.7	-61	0.17	down
09:33	n	b	0.11	-112.7	132.5	-114.7	0.089	unchanged
09:34	n	n	0.08	-128.5	-101.3	-132.5	0.328	down

The subsections to follow will provide brief discussions to illustrate the data mining techniques associated with this data mining tool called SL. These techniques are shown in Figure 3.5.

3.8.1.1 Classification

The first data mining technique to be discussed is *classification*, and can be seen in Figure 3.5 under the appropriate data mining tool, which is supervised learning. Classification is a well-known data mining technique that assigns items to discrete, previously learned classes and predicts the class to which a new item (data instance) will belong to (Erl et al., 2015; Gera and Goel, 2015; Taylor, 2013). Erl et al. (2015) stated that classification broadly consists of two steps:

1. The system is fed labelled or categorised training data, to develop an understanding of different categories.

2. The system is fed unknown, yet similar data for classification and based on the understanding it developed from the training data, the algorithm will classify the unlabelled data.

Common applications include, spam filtering, bank load applications, fraud detection, target marketing, *etc.* In classification problems the output of instances admits only discrete, unordered values (Kotsiantis, 2007). Figure 3.10 illustrates a simplified classification process, a machine is fed labelled data during training, that builds its understanding of the classification. When unlabelled data is fed to the machine, it classifies the data itself.

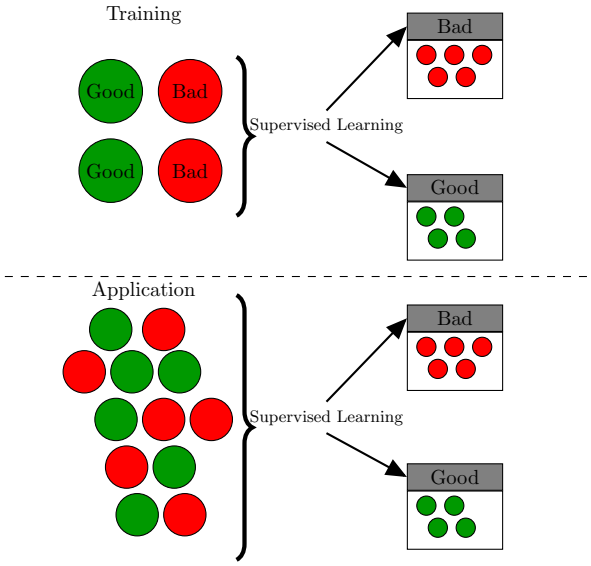


Figure 3.10: Classification used to automatically classify datasets (Erl et al., 2015).

USMA (2017) constructed Table 3.11 to provide an overview of techniques which fall under classification, with sources allocated to each. This table serves as a guideline when performing SL. All of the classification techniques presented in Figure 3.5 are indicated in column two of Table 3.11, with the numbers in column one referring to the numbers in Figure 3.4. Column three provides applicable sources for further reading regarding the specific classification technique, while column four indicates what each technique is known for, or for which situations it can be applied. In the subsections that follow classification, the majority of the classification techniques will be discussed to provide more insight.

Table 3.11: Summary of classification techniques (USMA, 2017).

	Classification technique:	*Source:*	*Known for/Application:*
1	**Decision trees**	Apté and Weiss (1997) Bell and Mgbemena (2018) Kim et al. (2006) Kotsiantis (2007) Larose and Larose (2014) Paramasivam et al. (2014) Rokach and Maimon (2014)	Segmentation using a target variable Profiling clusters Group clusters with common preferences Decision support tool Customer segmentation Customer identification Fraudulent behaviour Fault diagnosis
1.1	Classification and regression trees (CART)	Breiman et al. (1984) Larose and Larose (2014) Rokach and Maimon (2014) Steynberg (2016)	Financial analysis Building predictive models Spam filtering
1.2	C4.5 algorithm	Hssina et al. (2014) Kotsiantis (2007) Larose and Larose (2014) Quinlan (2014)	Generate decision tree Text processing Fault diagnosis

Continued on next page

	Classification technique:	*Source:*	*Known for/Application:*
1.3	Random forest	Breiman (2001) Murphy (2012)	Fraudulent behaviour Automatic medical diagnosis Identifying stock behaviour E-commerce: Recommendation system Outlier detection
2	**Support vector machines (SVM)**	Coussement and Van den Poel (2008) Huang et al. (2007) Jansen (2007) Kotsiantis (2007) Rechenthin (2014) Tomar and Agarwal (2013) Vapnik (1999)	Text and hypertext categorisation Pattern recognition Customer segmentation Image segmentation Bioinformatics
3	**Neural networks**	Bloom (2004) Chan (2005) Hastie et al. (2009) Izenman (2008) Jiawei et al. (2011) Kuo et al. (2006) Linoff and Berry (2011) Paliwal and Kumar (2009b) Petroulakis and Miaoudakis (2007)	Decision-making Pattern recognition Sequence recognition Face identification Automatic medical diagnosis Spam filtering Market segmentation Customer identification

Continued on next page

	Classification technique:	*Source:*	*Known for/Application:*
4	**Naïve Bayes network**	Jiawei et al. (2011) Rechenthin (2014) Li (2015)	Text categorisation Pattern recognition Automatic medical diagnosis Spam filtering
5	***k*-nearest neighbour (*k*NN)**	Kotsiantis (2007) Larose and Larose (2014) Li (2015) Rechenthin (2014) Salkind (2007)	Classifying new observations from predictions Concept search Recommendation system Outlier detection
6	**Rule-based classifiers**	Cakir and Aras (2012) Ishibuchi and Yamamoto (2005) Lawrence and Wright (2001)	Automatic medical diagnosis Concept search Recommendation systems Outlier detection Loyalty programs

3.8.1.1.1 Decision trees

Decision trees, shown as the first classification technique in Figure 3.5, form part of predictive and exploratory analysis (as defined in section 3.3), which can be utilised to represent both classification and regression models. In operational research, decision trees refer to a hierarchical model of decisions and their consequences. When a decision tree is used for classification, it is commonly referred to as a classification tree, while it is called a regression tree when it is used for regression (Rokach and Maimon, 2014). However, when the researcher refers to a decision tree within the classification section, it can be assumed that it is equivalent to a classification tree.

Decision trees are trees that classify objects or instances into a predefined set of classes (risky/non-risky), by sorting them based on feature or attribute values (age, gender, *etc.*) (Kotsiantis, 2007; Rokach and Maimon, 2014). The predefined set of classes which contain the feature and attribute values are discrete subcategories, and the selection of these attributes in the dataset is based on its predictability to a certain subcategory (Paramasivam et al., 2014).

Each node in a decision tree represents a feature of an instance to be classified and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values (Kotsiantis, 2007). A decision tree, as displayed in Figure 3.11, consists of decision nodes, which are connected by branches extending from the root node, which is usually at the top of the diagram, towards the terminating leaf nodes. Variables are tested from the root node at each decision node, with the possible outcome being represented by a branch, which again leads to another decision node, or terminating leaf node. When the tree cannot split further, no new nodes appear (Larose and Larose, 2014).

Figure 3.11 illustrates a simple decision tree, where the target variable is *credit risk*, with potential customers being classified as either good or bad credit risks. The predictor variables are *savings* (low, medium, high), *assets* (low or not low) and *income* ($\leq$\$30 000 or $>$\$30 000). In this example, the root node represents a decision node, testing whether each record has a low, medium or high savings level. The records with low savings are sent *via* the leftmost branch (savings = low) to another decision node, while records with high savings are sent *via* the rightmost branch to a different decision node. The records with medium savings are sent *via* the middle branch directly to a leaf node, indicating the termination of this branch. This occurs when all the instances for medium savings

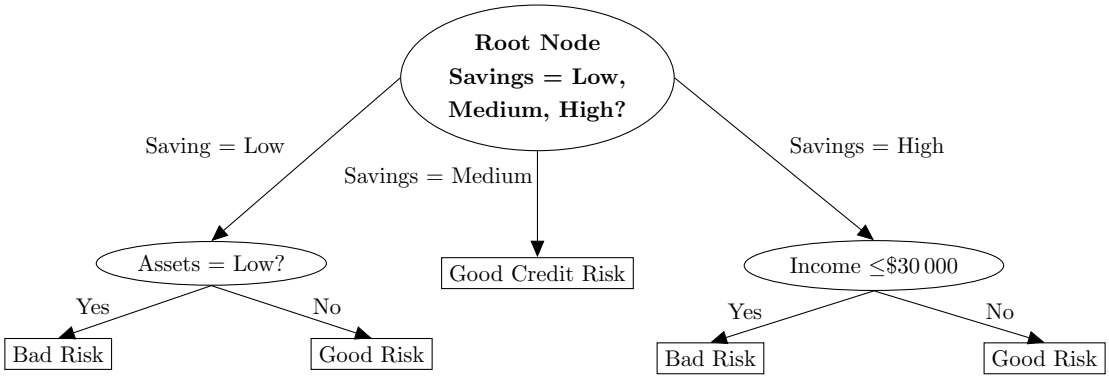


Figure 3.11: A simple decision tree (Larose and Larose, 2014).

are classified to be good credit risks, resulting in 100% accuracy, pure node, and no further splitting options are available.

The next decision node test is whether a customer with low savings has low assets, those with low savings are classified as *Bad Risk*; the remaining low savings customers are classified as *Good Risk*. Moreover, customers with high savings are tested at the next decision node whether they have an income of at most $30 000, if they have $30 000, or less, they are classified as *Bad Risk*, while the remaining high savings customers are classified as *Good Risk*. Thus, no further splits can be made and therefore the decision tree stops growing new nodes.

As mentioned earlier, when the savings are equal to medium, all the instances are classified to be good credit risk, where the target variable is unary (single option – *Good Credit Risk*) for the records in that node. This is not always the situation, therefore there are various methods for measuring leaf node purity and deciding on a cut-off value. The two leading algorithms for constructing decision trees include:

1. Classification and regression trees (CART) algorithms
2. C4.5 algorithm

There are certain requirements that must be met before applying the decision tree technique (Larose and Larose, 2014). The decision tree algorithms represent SL, thus requiring preclassified target variables. A rich and varied training dataset must be supplied, which will provide the algorithm with the values of the target variable, as well as a healthy cross section of the types of records for which classification may be needed

in the future. Decision trees learn by example, therefore classification and prediction will be problematic or impossible when the examples are systematically lacking a definable subset of records. The target attribute classes must be discrete. Tree analysis cannot be applied to a continuous target variable. Thus, the target variables must take on values that are clearly distinguishable as either belonging or not belonging to a particular class. This is why decision trees could be utilised for segmenting customers, profiling segments, as well as predicting responses to data.

Next, the second classification technique, as seen in Table 3.11, will be discussed, namely support vector machines.

3.8.1.1.2 Support Vector Machines

A support vector machine (SVM) is an algorithm that learns by example to assign labels to objects (Jansen, 2007). The algorithm works by classifying instances based on a linear function of the features (Rechenthin, 2014). As with the previous classification technique, SVMs may be utilised to predict a binary outcome of observations. SVMs can be utilised in customer segmentation by recognising the segment of a customer by examining thousands of customer data features (customer profiles) of each segment. As reported by Vapnik (1999) and Kotsiantis (2007), SVMs are the newest supervised learning technique.

The SVM was pioneered by Vapnik (1999) to address the problem of pattern classification and non-linear regression by minimising the structural risk. The SVM was initially developed for binary classification and will be explained accordingly in this section.

According to Huang et al. (2007), based on the theory of SVMs, they are proposed to cluster datasets, nowadays it could effectively be extended for multi-class problems (Rechenthin, 2014; Tomar and Agarwal, 2013). These areas, as mentioned by Huang et al. (2007), include various classification and curve fitting problems such as pattern recognition, text categorisation, bioinformatics, *etc.*

The SVM classifier creates a *hyperplane*, or multiple hyperplanes within the high dimensional space that is useful for classification, regression and other efficient tasks (Huang et al., 2007). In a binary classification context, SVMs try to find a linear optimal hyperplane so that the margin of separation between positive and negative examples is maximised (Coussement and Van den Poel, 2008). It can be concluded that

the objective of SVMs is to use a single linear surface, known as the hyperplane, to separate the observations in the training data belonging to two different classes, with the largest margin of separation possible. The outcome of new observations may be predicted based on which side of the hyperplane they lie (Vapnik, 1999).

In order to explain the linear separability of observations in two dimensions, consider Figure 3.12. The observations, denoted as filled circles, are the outcome of class 1, while the observations denoted as filled squares are the outcome of class 2. SVMs project data points onto a higher dimension, while determining the best hyperplane to separate the data. Additionally, kernel functions, for example Gaussian or polynomial, are used for non-linear mapping of the training sample to the higher dimensional space. Figure 3.12 shows a linear separable situation, therefore by utilising SVMs, the aim is to draw a single line (hyperplane) denoted by H, to separate the two types of observations with the largest possible margin of separation, denoted by M. Then, the observations that lie on the dashed margin of separation lines, B_1 and B_2, are called the *support vectors*.

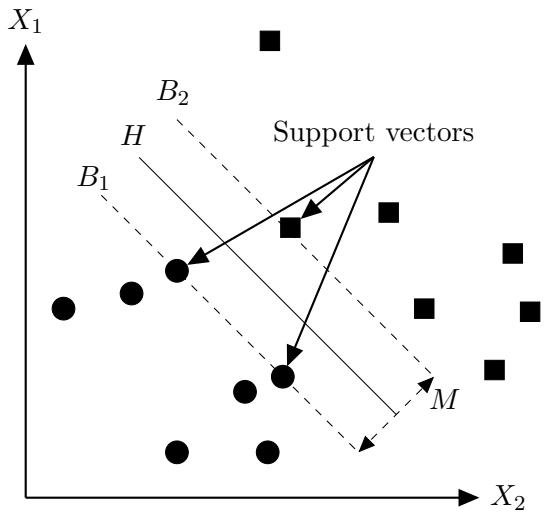


Figure 3.12: Fundamentals of SVM (Steynberg, 2016).

In Figure 3.13 it is shown that other lines may be drawn to separate the observations (M_1 and M_2). Such lines achieve smaller margins of separation. In Figure 3.13, M_2 is the larger of the two margins of separation and thus its corresponding hyperplane is favoured. As stated by Rechenthin (2014), the classifier is fed with pre-labelled instances, and by selecting points as support vectors the SVM searches for hyperplanes that maximise

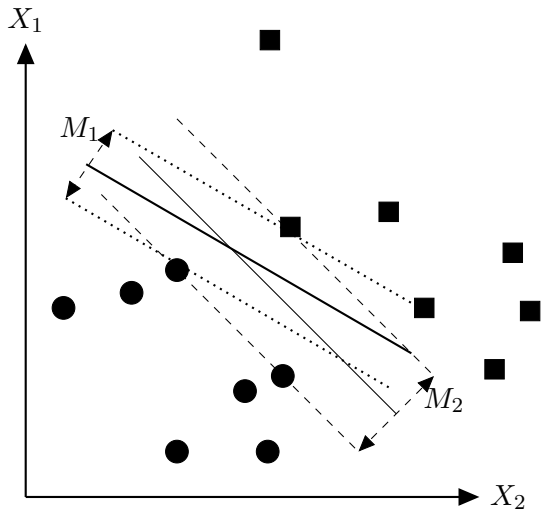


Figure 3.13: SVM margin of separation (Steynberg, 2016).

the margin. More information regarding the SVM decision boundary and mathematics behind the large margin classification can be found in Steynberg (2016), Vapnik (1999) and Kotsiantis (2007).

SVMs are a popular data mining technique, applied to various real-world problems namely, text and hypertext categorisation, classification of images, image segmentation, handwriting recognition, *etc.* These, and more applications that SVMs are known for, are shown in Table 3.11. Next, the third classification technique number, as shown in Table 3.11, namely neural networks, will be discussed.

3.8.1.1.3 Neural networks

As mentioned, USMA (2017) developed Table 3.11 to provide more insights into all the classification techniques that are schematically illustrated in Figure 3.5. The third classification technique mentioned by USMA (2017) is the neural network. This technique defines a wide field of study, therefore it was decided not to conduct an in-depth literature study. This section continues providing an overview of the technique.

Neural networks were developed in both statistical and artificial intelligence and are also known as artificial neural networks (ANN). According to Hastie et al. (2009), a neural network is just a non-linear statistical model, and consists of a two-stage regression or classification model, typically represented by a network diagram. According to Haykin

(2011), an ANN is a machine that is designed to model the way in which the human brain performs a task or function of interest.

It is known that the human brain consists of a huge number of neurons (nerve cells), which connect to each other to form neural networks. Each neuron can be seen as an information-processing unit which makes a simple decision so that a human can react to the environment. A simple example of a human response is that of the knee-jerk reflex that occurs when someone taps the tendon below the knee with a hammer (Khan Academy, 2013). This is known as the "all-or-none" character of nervous activity. When it reacts, an electrochemical pulse is generated and spreads to thousands of neurons that are connected to the reacting neuron. Each neuron that accepts this electrochemical signal in turn makes its own decision about reacting, based on the signal it received from the aforementioned neuron as well as signals from thousands of other neurons it is also connected to (Yoon, 2015).

McCulloch and Pitts (1943) attempted to model the functioning of neural networks mathematically, which led to the development of ANNs. They presented a mathematical model of a neuron that has three basic elements (Marsland, 2015):

1. A set of weights denoted by w_i for the i^{th} input of the neuron.
2. An adder to sum the input signals.
3. An activation function that determines whether the neuron reacts/fires for the current inputs.

Figure 3.14 illustrates the mathematical model of a neuron. The neuron has m inputs (x_1,x_2,...,x_m), a bias input x_0 (which always has a value of 1), and an output (y). The adder calculates the sum of the weighted inputs to form the net input $h = \sum_m^{i=1} w_i x_i + b$. The activation function accepts this value as its input and determines the output $y = \phi(h)$. A typical activation function that best represents the neuron's all-or-none character is shown in Figure 3.15a. If the sum of the weighted inputs is greater than or equal to zero, it reacts/fires. Otherwise it does not react/fire. This is called the *threshold activation function*. The mathematical form of the threshold activation function is as follows:

$$\phi(h) = \begin{cases} 1, & \text{if } h \geq 0; \\ 0, & \text{if } h < 0. \end{cases}$$

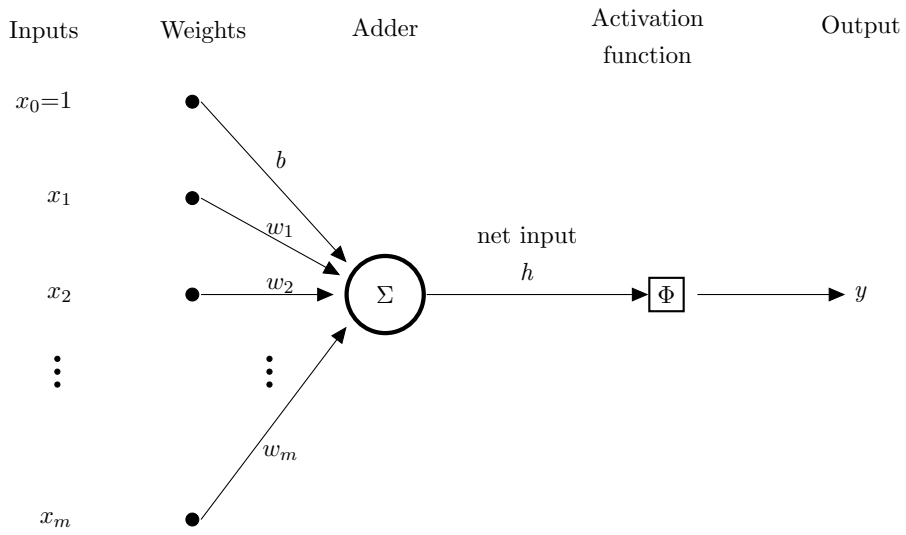
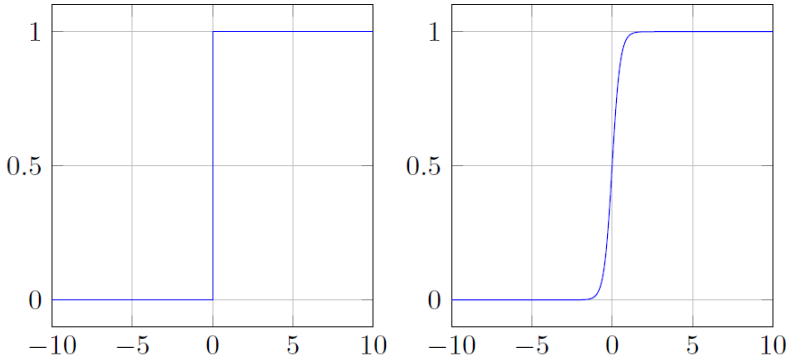


Figure 3.14: A neuron model indicating the three basic elements (Marsland, 2015).



(a) Threshold activation function (b) Sigmoid activation function

Figure 3.15: Types of activation function (Yoon, 2015).

Another form of activation function is called the *sigmoid function*. Figure 3.15b shows an example of a sigmoid function, of which the graph is S-shaped. It appears similar to that of the threshold activation function (Figure 3.15a), yet it increases smoothly. A common example of the sigmoid function is the logistic function. The mathematical form is

$$\phi(h) = \frac{1}{1 + \exp(-ch)},$$

where c is a positive parameter to indicate how quickly the function transitions from

low values to high values. The bigger the parameter, the more the shape of the sigmoid function resembles that of the threshold function (Figure 3.15a).

The primary uses of neural networks include pattern classification and prediction (*e.g.* problems entailing the recognition of speech, faces and characters as well as robotics). These problems entail large sample sizes and high-dimensionality (Izenman, 2008). More application areas are listed in Table 3.11.

Next, the fourth classification technique to be discussed is called the naïve Bayes network.

3.8.1.1.4 Naïve Bayes network

The naïve Bayes classifier is an efficient probabilistic model based on the Bayes theorem, that examines the likelihood of features appearing in the predicted classes (Rechenthin, 2014). Naïve Bayes is not a family of classification algorithms that share the assumption that every feature of the data that is being classified, within a certain class, is independent of the other features in that class (*e.g.* customer's age and location) (Li, 2015). It is called *naïve*, because of the assumption that all features of a dataset are independent.

Given a set of attributes $X = \{x_1,x_2,...,x_n\}$, the objective is to construct the posterior probability (probability that a hypothesis is true, calculated in the light of relevant observations) for the event C_k among a set of possible class outcomes $C = \{c_1,c_2,...,c_k\}$. Let H be some hypothesis such as that the dataset X belongs to a specific class C. For a classification problem $P(H|X)$ needs to be determined, the probability that H holds, given the observed dataset X. $P(H|X)$ is the *posterior probability* of H conditioned on X (Jiawei et al., 2011).

Suppose that the dataset X is confined to describe the attributes of customers, such as age and income respectively, and that X is a 35-year-old customer with an income of \$40 000. Again, suppose that H is the hypothesis that a customer will buy a computer. Then $P(H|X)$ reflects the probability that customer X will buy a computer given that the customer's age and income are known.

$P(H)$ is called the *prior probability* of H. For this example, it is the probability that any given customer will buy a computer, regardless of age, income or any other information. The *posterior probability*, $P(H|X)$, is based on more customer information than the *prior probability*, $P(H)$, which is independent of X.

Similarly, $P(X|H)$ is the *posterior probability* of X conditioned on H. It is the probability that customer X, who is 35 years old and earns \$40 000, given that the customer is known, will buy a computer. $P(X)$ is the *prior probability* of X. Therefore $P(X)$ is the probability that a person from the known set of customers is 35 years old and earns \$40 000.

In layman's terms, $P(H|X)$ can be interpreted as the probability that a computer will be bought by a 35–year-old man (customer), while $P(X|H)$ can be interpreted as the probability that a 35–year-old man (customer) buys a computer.

To conclude this example, $P(H)$, $P(X|H)$ and $P(X)$ may be estimated from the provided data. Bayes' theorem is useful in that it provides a way of calculating the *posterior probability*, $P(H|X)$, from $P(H)$, $P(X|H)$ and $P(X)$. Thus, the Bayes' theorem, for this example, can be written as

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}.$$

Naïve Bayes involves simple arithmetic, it relies on calculating up counts, multiplying and dividing. Once the frequency tables are calculated, classifying an unknown customer just involves calculating the probabilities for all the classes, and then selecting the highest probability. Despite its simplicity, naïve Bayes can be accurate, and it is commonly used for spam filtering. More areas of application for naïve Bayes can be seen in Table 3.11.

Next to be discussed is the fifth classification technique listed in Table 3.11, namely k-Nearest Neighbour.

3.8.1.1.5 *k*-nearest neighbour

k-nearest neighbour (kNN) is a classification technique that forms part of the data mining tool called SL since it is provided a labelled dataset (Li, 2015). Therefore, this technique is most often used to classify data points into any number of categories, although it can also be used for estimation and prediction (Salkind, 2007). The kNN technique is one of the simplest machine learning techniques and is often referred to as a *lazy* learner, because learning is not implemented until actual classification or prediction is required (Rechenthin, 2014). Lazy learners require less computation time during the training phase than eager-learning algorithms (*i.e.* decision trees, neural networks, Bayes), but more computation time during the classification process (Kotsiantis, 2007).

kNN is an example of *instance-based learning*, in which the training dataset is stored, when a new unclassified record is added, it is classified by simply comparing it to the most similar record in the training set. To begin, when new unlabelled data enters, the kNN operates in these three basic steps:

1. The *distance* between the new data point that is to be classified and each of the data points in the training dataset is computed.
2. The data points in the training dataset are then stored, in descending order, according to their distance to the new data point.
3. The predicted category is the most common category of the k data points that are the nearest to the new data point.

According to Larose and Larose (2014), the distance between the labelled and unlabelled data is determined with a function that accounts for three aspects:

1. The distance is always non-negative, and zero when the coordinates are the same.
2. The distance between two points, for example point x and point y, is the same as the distance from point y to point x, this is called *commutativity*.
3. Introducing a third point (point z), between point x and point y, will never shorten the distance between two other points, this is called *triangle inequality*.

The main principle outlining the kNN algorithm can be concluded to be that it can be used to estimate the probability that an observation belongs to a specific class by comparing it to its neighbouring data points and observing which they belong to. In order to decide which points counts as the closest neighbours, similarity between the data points must be defined. A plausible way to determine similarity when the observation features are quantitative is to make use of a distance function (Dizdarevic, 2017).

The most general distance function is *Euclidean distance*, which represents the usual manner in which humans think of distance in the real world. When measuring distance, certain attributes that contain large values (income) can overwhelm the influence of other attributes which are measured on a smaller scale (age). To avoid this, the attribute values should be *normalised*. The normalisation formulas can be seen in Larose and

Larose (2014). However, the Euclidean distance metric is not appropriate for categorical variables, for which a function has to be defined.

After establishing a method to determine which records are most similar to the new, unclassified record, a classification decision has to be made. This can be achieved either by unweighted or weighted voting. Unweighted voting is the most simple voting method, which entails *deciding* the value of k, that is the number of records that will have an influence on the classification decision, *comparing* the new record to the kNN (minimum distance from the new record) and then *classifying* the record based on one vote from each nearest neighbour. Weighted voting follows the same procedure, except that neighbours are weighted in inverse proportion to the distance from the new point. This results in closer neighbours having a greater weighting, which means that they have a larger influence in the classification decision, in contrast to the more distant neighbours. The weighted voting also minimises the likelihood of ties (Larose and Larose, 2014).

kNN is a popular method because of the understanding and implementation being simpler than other methods, and depending on the distance metric, quite accurate. However, kNN can get computationally expensive when determining the nearest neighbour on a large dataset, kNN is not robust regarding noisy data, and in the case where some attributes have a large range and other smaller ranges, it must be scaled (Li, 2015).

The classification technique that is indicated to follow kNN, is rule-based classifiers (indexed as 6). However, this technique will not be discussed in detail. Table 3.11 provides sources for further reading as well as application areas for rule-based classifiers.

Next, the second SL tool, namely regression, will be investigated, as well as regression techniques.

3.8.1.2 Regression

Regression and classification are the two SL techniques that will be discussed in this research project. Regression analysis is one of the most widely used of all statistical methods (Paliwal and Kumar, 2009a) and forms the basis for many other statistical techniques (*e.g.* variance analysis, covariance analysis, t-test, Pearson product-moment correlation and Spearman (ρ) correlation) which are all specially designed versions of regression (Salkind, 2007).

Regression is similar to classification; however, the difference lies with the response variable being continuous (Murphy, 2012), thus, regression is a data mining technique

that is used to predict a continuous and numerical value or target (Gera and Goel, 2015; Salkind, 2007; Taylor, 2013). Applying regression helps determine how the value of the dependent variable changes in relation to changes in the value of the independent variables (Erl et al., 2015). Regression techniques can be used to predict profit, sales, mortgage rates, house values, temperature or distance. For example, a regression model could be utilised to predict the value of a house based on location, number of rooms, plot size as well as other factors (Taylor, 2013).

According to Mohri et al. (2012); Salkind (2007) and Gera and Goel (2015), regression is based on a training process that consists of using dataset values already known, to predict, as closely as possible, the real-valued labels of the points or items considered. Regression is a common technique in machine learning with a variety of applications.

Similar to classification, regression consists of various methods, namely linear regression, non-linear regression and logistic regression. Similar to Table 3.11, Table 3.12 provides an overview of regression techniques, with applicable sources in column three and application areas in column four.

In the subsections that follow, each of the main regression techniques will be discussed to provide more insight. These techniques are schematically represented in Figure 3.5 and also indicated and summarised in Table 3.12.

Table 3.12: Summary of regression techniques (USMA, 2017).

	Regression technique:	*Source*:	*Known for/Application*:
1	**Linear regression**	Salkind (2007) Gera and Goel (2015) Paliwal and Kumar (2009a) Salkind (2007)Yang et al. (2017)	A model that can show relationships between two variables and how the independent variable impacts the dependent variable.
1.1	Simple linear regression	Bishop (2013) Paliwal and Kumar (2009a) Salkind (2007)	Evaluate trends Forecasting Analyse marketing effectiveness Assess finance/insurance risks
1.2	Multiple linear regression	Bishop (2013) Paliwal and Kumar (2009a) Salkind (2007)	Same as Simple Linear Regression, more variables for example in management – Is customer loyalty influenced by customer satisfaction, brand perception and price perception?
2	**Non-linear regression**	Bates and Watts (2007) Chatterjee and Hadi (2013) Gallant (1975) Gera and Goel (2015) Riffenburgh (2011) Ruckstuhl (2010) Tellis (2006) Tellis and Ambler (2007)	Assesses the effects of wearin and wearout on campaigns. This can be done by determining the effectiveness of advertisements on different age groups.

Continued on next page

	Regression technique:	*Source:*	*Known for/Application:*
3	**Logistic regression**	Chatterjee and Hadi (2013) Hosmer et al. (2013) Karp (1998) Montgomery et al. (2015) Riffenburgh (2011) Salkind (2007)	Assesses likelihood of remaining a customer Sales – purchase/re-purchase vs. No purchase Marketing – Respond vs. Not respond Fraud identification Health care – Cure vs. No cure

3.8.1.2.1 Linear regression

This subsection introduces the first regression technique to be discussed. Linear regression is shown in Table 3.12, and consists of two sub-techniques, namely simple linear regression and multiple linear regression.

Linear regression is a simple, yet powerful technique and well-known statistical learning method used to predict the value of quantitative variables. It forms the basis of many other learning methods. The prediction of quantitative variables is therefore referred to as a *regression* problem. The underlying assumption of this method is that there is an approximate linear relationship between the input and output variable. Therefore, linear regression is utilised when the relation between the target and the predictor can be represented as a straight line (Salkind, 2007).

The two forms of linear regression can be seen in Table 3.13, and provide a brief overview of the purpose of both single and multiple linear regression, along with their equations. Firstly, simple linear regression can be considered as an approach for predicting a quantitative response Y on the basis of a single predictor variable X. The assumed linear relationship can be seen in row three, column two of Table 3.13. In this equation β_0 and β_1 are the intercept and slope, respectively. Collectively, they are referred to as the model coefficients. When applying simple linear regression in practice, these coefficients are unknown and should be determined using the training data provided. The data can be written as n observation pairs $(x_i,y_i),...,(x_n,y_n)$, each consisting of a measurement of X and a measurement of Y. The goal is to fit the model to the data in such a way that each prediction, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$, is as close to the *true* value y_i as possible (Salkind, 2007).

Table 3.13: Simple linear regression and multiple linear regression

	Simple linear regression:	**Multiple linear regression:**
Purpose:	Examines the relationship between two variables. This include the *predictor* variable (X) and the *criterion* variable (Y).	Examines the relationship between at least two predictor variables (X_1,X_2,...,X_k) and the *criterion* variable (Y).
Equation:	$Y = \beta_0 + \beta_1 X$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$

There are several methods available to ensure this *closeness*. The best known is *least squares regression*. In this method the *residual* for each observation is defined as

$e_i = y_i - \hat{y}_i$ (Hastie et al., 2009). The residual for the least squares fit may be graphically represented by the distance from the red point to the fitted blue line in Figure 3.16.

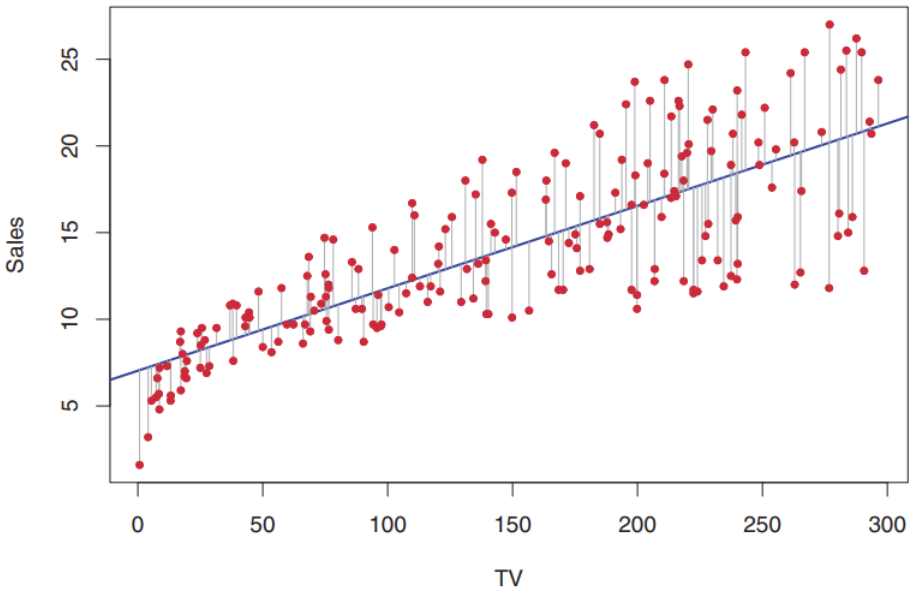


Figure 3.16: Simple linear regression (James et al., 2013).

The *residual sum of squares* (RSS) is defined as

$$\text{RSS} = {e_1}^2 + {e_2}^2 + ... + {e_n}^2$$

or, equivalently, as

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + ... + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2 \tag{3.1}$$

The minimum value of RSS is determined by calculating the partial derivatives of (3.1) with respect to β_0 and β_1 and setting them to zero. Solving for the coefficients then yields the following *least squares coefficient estimates*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad \text{and}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x},$$

where $\overline{x}$ and $\overline{y}$ denote the sample mean of the input (independent) variable and output (dependent) variable, respectively (James et al., 2013).

However, in most problems more than one input variable exists that dictates the model outcome (Hastie et al., 2009). To accommodate this, the simple linear regression model can be extended to form a multiple linear regression model with k predictors, as seen in Table 3.13, column three. In the equation, the X_i represents the i^{th} predictor and β_i quantifies the association between the variable and the response. For each predictor, β_i may be interpreted as the average effect on Y for a unit increase in X_i if all other predictors are held constant.

The coefficients are estimated in a similar fashion as in the simple linear regression case. Visually, the least squares for a three-dimensional model may be thought of as a minimisation of the vertical distance of each observation to the fitted plane, shown in Figure 3.17.

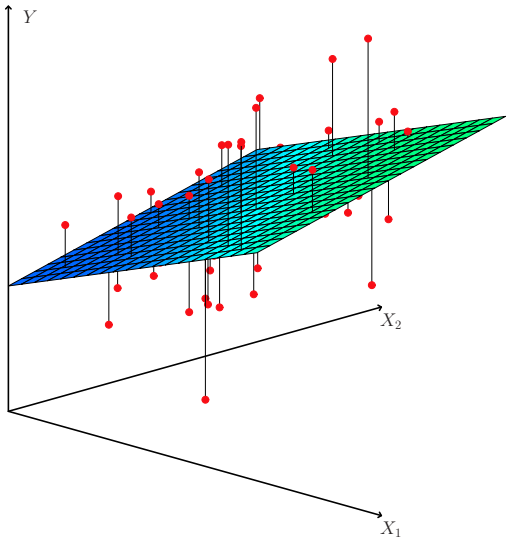


Figure 3.17: Multiple linear regression for a three-dimensional problem (James et al., 2013).

There are various metrics available to evaluate the accuracy of the model. One commonly used metric is the *coefficient of determination*, R^2 (Salkind, 2007). This provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. This coefficient

may be calculated using the formula

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}},$$

where $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the *residual sum of squares* and $TSS = \sum_{i=1}^{n}(y_i - \overline{y_i})^2$ is the *total sum of squares*. Another common measure of model accuracy is the *root mean squared error* (RMSE). This represents a measure of the difference between outcomes predicted by the model and actual observed outcomes by computing the mean of the square of the residuals and then taking the square root of this value as per the equation

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

By first squaring the error terms, larger errors are amplified. Subsequently, taking the square root of the sum of these values, yields a value of the error that bears the same units as the output variable that is to be predicted, making it easier to interpret.

The next regression technique to be discussed is called non-linear regression.

3.8.1.2.2 Non-linear regression

This subsection discusses the second regression technique as seen in Table 3.12, called non-linear regression. When the relationship between the response and some of the predictors is non-linear or some of the parameters appear non-linearly, the regression technique that should be applied is called non-linear regression. A non-linear function can be written as

$$Y_n = f(x_n, \theta) + e_n,$$

where f is the expectation/response function and x_n is a vector of independent variables or unknown parameters for the n^{th} case, and e_n represents unobservable observational or experimental errors (Gallant, 1975).

For non-linear models, at least one of the derivatives of the expected function, with respect to the parameters, depends on at least one of the parameters (Bates and Watts, 2007).

A non-linear regression equation can take on multiple forms, therefore it is worth emphasising the intuitive definition that states that when the relationship is, $Y = \beta_0 +$

$\beta_1 X$, it is linear, if not, it is non-linear. However, there are a few cases where a non-linear equation can be transformed to mimic a linear equation.

Non-linear functions that can be transformed into linear functions are called *linearisable* functions and those that are not able to transform are called *intrinsically* non-linear functions (Chatterjee and Hadi, 2013; Ruckstuhl, 2010). General linear models are not restricted to linear algebraic models. They may include non-linear forms (*e.g.* log, exponential and Gompertz, as well as a combination of algebraic and/or non-linear forms) (Riffenburgh, 2011).

The linearisable regression function transforms the variable of interest and the explanatory variables. For example, a power function such as:

$$h(x;\theta) = \theta_1 x^{\theta_2}$$

can be transformed to a linear (in the parameters) function

$$\ln(h(x;\theta)) = \ln(\theta_1) + \theta_2 \ln(x) = \alpha + \beta\tilde{x}$$

where $\alpha = \ln(\theta_1)$, $\beta = \theta_2$ and $\tilde{x} = \ln(x)$ (Ruckstuhl, 2010).

Non-linear regression is utilised when:

1. a transformation is necessary in order to obtain variance homogeneity, but the transformation destroys linearity.
2. linearity does not fit, and the transformation seems to destroy other parts of the model assumptions, for example, the assumption of variance homogeneity.
3. theoretical knowledge indicates that the proper relation is intrinsically non-linear.
4. interest is in the functions of the parameters, which do not enter linearly in the model.

One of the applications of non-linear regression in marketing was to assess the effects of wearin (increasing response to an advertisement, with increasing repetition of exposure of the advertisement, occurs in the early stages of a campaign) and wearout (decreasing response to an advertisement, with increasing repetition of exposure of the advertisement, occurs in the latter stages of a campaign) on campaigns. This can be done by determining the effectiveness of advertisements on different age groups (Tellis and Ambler, 2007; Tellis, 2006). Next, the last regression technique specified by USMA (2017) will be discussed.

3.8.1.2.3 Logistic regression

Logistic regression is the last regression technique and the final SL technique, to be discussed. Logistic regression represents the case where the response variable is qualitative (Chatterjee and Hadi, 2013) and where there are only two possible outcomes, generally called success and failure and denoted as 0 and 1 (Montgomery et al., 2015).

Logistic regression is a flexible technique when one variable is identified as the response variable and it is categorical (Salkind, 2007). A logistic regression model can be distinguished from a linear regression model by observing the outcome variable. The linear regression methods described above are suitable for predicting quantitative variables, while in the case of logistic regression the outcome variable is *binary* or *dichotomous* (qualitative) (Hosmer et al., 2013). Assigning a new observation to a category based on its known characteristics is referred to as *classification*.

The two main categories of data, qualitative and quantitative data, are very important in the context of classifying the output of a process. According to Hastie et al. (2009), attempts at predicting outputs of a quantitative nature are collectively known as *regression*, while attempts at predicting outputs of a qualitative nature is referred to as *classification*. Various authors agree with this taxonomy (Gelman and Hill, 2006; Luna, 2000; Shalizi, 2017). James et al. (2013) stated that the response variables, not the predictor variables, should be the main consideration when selecting between a regression or classification approach to tackle a specific statistical learning problem.

Numerous problems will arise when using ordinary linear regression analysis when the response variable is categorical, for linear regression is not appropriate for classification. Therefore a transformation is needed on the categorical response variable so that it can be predicted by a linear relationship with the explanatory variables.

Lekdee and Ingsrisawang (2010) express, for the case of m independent variables, the *logistic function* as

$$P(Y_i = 1|\mathbf{X} = [X_{1,i}, X_{2,i}, ..., X_{m,i}]) = \frac{e^{\beta_0+\beta_1 X_{1,i}+\beta_2 X_{2,i}...+\beta_m X_{m,i}}}{1 + e^{\beta_0+\beta_1 X_{1,i}+\beta_2 X_{2,i}...+\beta_m X_{m,i}}}, \quad i \in \mathcal{N}.$$

For ease of explanation, the single independent variable case of

$$P(Y_i|X_i) = \frac{e^{\beta_0+\beta_1 X_i}}{1 + e^{\beta_0+\beta_1 X_i}}, \quad i \in \mathcal{N},$$

or, in rearranged form,

$$\frac{P(Y_i|X_i)}{1-P(Y_i|X_i)} = e^{\beta_0+\beta_1 X_i}, \quad i \in \mathcal{N}, \tag{3.2}$$

is considered here. Gelman and Hill (2006) and Salkind (2007) call the left side of (3.2) the *odds*, which may assume any value in the interval $[0,\infty)$. The odds takes the form $\frac{P}{1-P}$ for any event which has the probability P of occurring and a probability $1-P$ of not occurring

Consider the following event as an example. One in five people watches sport, this implies that the odds is $\frac{1}{4}$ that someone watches sport. The calculation behind this is that 1 in 5 means that $p(X) = \frac{1}{5} = 0.2$, and when substituting into the odds function, the result is $\frac{0.2}{1\text{-}0.2} = \frac{1}{4}$ (James et al., 2013).

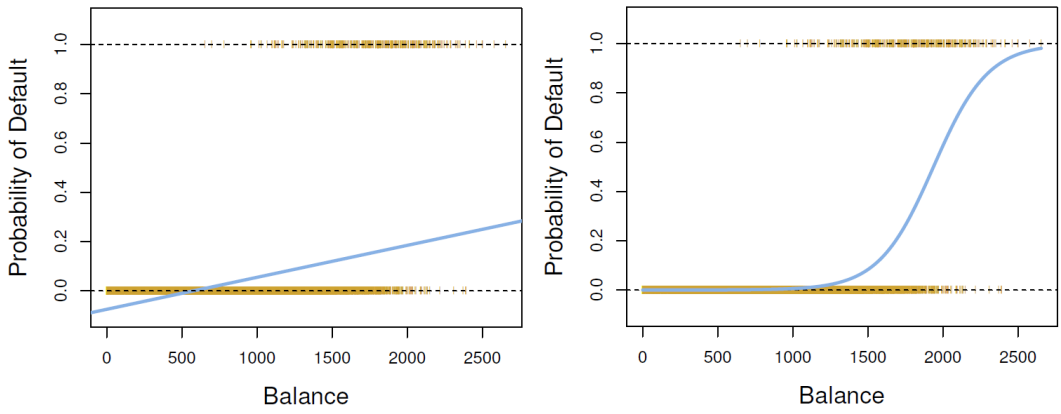
After the natural log-algorithm of the odds occurred, a linear relationship between the transformed variable and the explanatory variable can be established, called the logistic transformation or *logit*, expressed as

$$\text{Logit}[P] = \ln\left(\frac{P}{1-P}\right).$$

Since both the odds and logit values can be calculated, a general regression model may follow after the dependent variable is transformed to the log-odds ratio, calculated by

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_i.$$

Consider a case of *binary classification*, in which there are two categories. James et al. (2013) provided the following example for better understanding of logistic regression. One may wish to predict whether or not a customer will default on their loan, based on their credit card debt. The prediction may be interpreted as a probability of belonging to one of a number of categories. When using linear regression, it leads to an undesired result, as seen in Figure 3.18a. Some of the probability values are negative, and are difficult to interpret. In order to solve this problem, the probability $p(X)$ must be modelled using a function that yields output values between 0 and 1 for all X. A function that fits this criterion is the *logistic function*, also known as the sigmoid function, given by $\phi(z) = \frac{1}{1+e^{-1}}$. The sigmoid function is used to transform values in the range of $(\text{-}\infty,\infty)$ to values that fall in the range of (0,1), and is the inverse of the previously mentioned logit function. The linear model, previously introduced, can be transformed



(a) Linear regression model

(b) Logistic regression model

Figure 3.18: Binary classification using both a linear and logistic regression model to estimate whether or not a customer will *default* on a loan based on the average balance remaining on their credit card after their monthly payment James et al. (2013).

to the desired form simply by inserting it into the sigmoid function. The probability $p(x)$ of belonging to a category is then given by

$$p(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}}.$$

The coefficients could be estimated by utilising a non-linear form of least squares fit described earlier; however, it is more general to use the *maximum likelihood* method. The logic behind this method is to determine coefficients $\beta_0, ..., \beta_p$, such that $p(X)$ is close to 1 for customers who have defaulted and close to 0 for customers who did not (Riffenburgh, 2011; Salkind, 2007). This entails maximising the *likelihood function*

$$l(\beta_0 + \beta_1 X_1 + ... + \beta_p X_p) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=1} (1 - p(x_i')).$$

The resulting fit is shown in Figure 3.18b. The model now generates probabilities between 0 and 1, as desired. In order to predict the value of a category based on this model, a cut-off value is usually set as $p(X) = 0.5$. By raising or lowering the cut-off point, it is to some extent possible to increase or decrease the number of correct positive predictions and correct negative predictions.

For example, an observation with input variables $X_1, ..., X_p$ is classified into one category if $p(X)$ is larger than 0.5, or a second category otherwise. In the example illustrated in Figure 3.18, the probability of a customer defaulting on a loan versus the average balance X remaining on their credit card after their monthly payments is shown. The customer is predicted to default if $p(X) > 0.5$ (score = 1), otherwise, they are predicted to successfully repay the loan (score = 0) (James et al., 2013).

Table 3.14: Confusion matrix for comparing predictive outcomes *versus* actual outcomes

		Actual outcome:	
		Positive	Negative
Predicted outcome:	Positive	True positive	False positive
	Negative	False negative	True negative

The final part of a logistic regression model is to determine how well the model can predict dependent variable outcomes based on a set of given independent variables. This is called the *model evaluation*. To assess the predicting ability of quantitative variables, a *confusion matrix* for the predictions of the values on the *test data* can be constructed, as shown in Table 3.14.

On the diagonal (Table 3.14), the number of correctly classified observations are shown as *true positives* (TP) and *true negatives* (TN). Off the diagonal, the *false positives* (FP) and *false negatives* (FN) are shown (James et al., 2013; Powers, 2011).

The model accuracy can then be computed as the proportion of correctly classified observations, as stated by Powers (2011) to be

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}.$$

The applications of logistic regression, according to Karp (1998), generally include assessing the chance that a customer will repurchase a product, remain a customer or respond to direct mail or other marketing incentives. It can be concluded that logistic regression models yield powerful insights as to why some customers leave, while others stay. These insights can be employed to modify the implementation of retention strategies.

This concludes the discussion regarding regression techniques. Next, unsupervised learning will be discussed. Unsupervised learning is seen as a data mining tool, and is shown in Figure 3.5 as the fourth data mining tool.

3.8.2 Unsupervised learning

In SL, as previously discussed, the goal was to learn a mapping from input to an output, where the correct values are provided by a supervisor, whereas, in *unsupervised learning* (UL), there is no such supervisor and only input data is provided. UL is a field much less developed in literature as opposed to SL.

UL does not have a dependent variable and the methods are mainly *descriptive*, searching for unknown patterns or relationships (Bramer, 2007). According to James et al. (2013), UL is useful for better understanding of the relationship between the variables and/or observations. The goal is to seek regularities within the input (Alpaydin, 2009). Therefore, when a question like *"Do our customers naturally fall into different groups?"* is asked, no specific purpose or *target* has been specified for grouping. When no target is identified, the data mining problem is referred to as *unsupervised* (Provost and Fawcett, 2013). Therefore, Bramer (2007) stated that data mining of unlabelled data is known as UL.

There is a wide array of methods available for UL, yet only some of the most popular method will be introduced in this section. Next, the UL learning technique as indicated in Figure 3.5 and Table 3.16, called clustering, will be discussed.

3.8.2.1 Clustering

This section will be investigating an UL learning techniques called clustering. The term *data clustering* first appeared in the title of a 1954 article that dealt with anthropological data (Demšar and Zupan, 2013). Nowadays, clustering is the best-known UL method used for analysing multivariate statistical procedures that starts with a dataset containing information about a sample of entities and attempts to reorganise these entities into relatively homogeneous groups (Aldenderfer and Blashfield, 1984; Izenman, 2008; Madhulatha, 2011).

A similar definition of cluster analysis, according to Chiu and Tavella (2008), stated that cluster analysis is used to uncover interdependence between members of a sample. Clustering is a form of machine learning. The machine is the computer and the learning refers to an algorithm that is repeated until a set of predetermined conditions is met (Pierson and Porway, 2017). Learning algorithms are normally run until the point where the final analysis results are unchanged.

For a final definition of what is meant by clustering, Erl et al. (2015) stated that clustering is an UL technique by which data is divided into different groups so that the data in each group has similar properties. However, there is no prior learning of categories required, as opposed to SL. Instead, categories are implicitly generated based on data groupings.

These three definitions provide a clear understanding of what is meant by clustering. However, the manner in which the data is grouped depends on the type of clustering algorithm used. Each algorithm uses a different technique to identify clusters. These algorithms are derived from mathematics, statistics and numerical analysis.

The algorithm categorises the data into two or more groups (clusters) with the main goal of maximising the similarities between members in the clusters (Jacob and Ramani, 2012; Jiawei et al., 2011). In the marketing context, clustering can be applied to the categorisation of unknown data or documents and to personalised marketing campaigns by grouping together customers with similar behaviour (Erl et al., 2015). According to Izenman (2008), marketers utilise demographics and consumer profiles in an attempt to segment the marketplace into small, homogeneous groups, in order for promotional campaigns to be carried out more efficiently.

Clustering is a common technique used in data mining to get an understanding of the properties of a given dataset. After developing this understanding, classification can be used to make better predictions about similar, new or unseen data. Figure 3.19 represents a scatter graph that provides a visual representation of clusters.

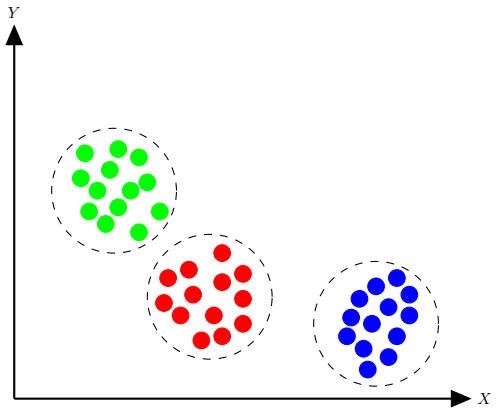


Figure 3.19: A scatter plot summarising the results of clustering (Erl et al., 2015).

Clustering resembles classification methods; however, these two data mining techniques possess rational differences (Izenman, 2008). Table 3.15 indicates three major differences between classification and clustering.

Table 3.15: Differences between clustering and classification (Izenman, 2008).

	Classification:	*Clustering:*
Technique:	Classify new items into classes based on rules learnt from the *learning set*.	No prior information regarding the class structure is available, which leads to defining clustering as an *exploratory data analysis*.
Size:	The number of *classes* (groups) in the dataset is *known*.	The number of *classes* (groups) in the dataset is *unknown*.
Item:	Classify *observations*.	Can be individually or simultaneously applied to group *observations* and/or *variables*, depending on the context.

USMA (2017) created Table 3.16 to provide an overview of all the clustering techniques that were schematically represented in Figure 3.5. The clustering techniques are indicated in column two of Table 3.16 with the applicable sources for each technique in column three, while column four provides insight into what each technique is capable of. Next, k-means clustering will be discussed, indexed as technique 1.1 in Table 3.16.

Table 3.16: Summary of clustering techniques (USMA, 2017).

	Clustering technique:	*Source:*	*Known for/Applications:*
1	**Clustering**	Aggarwal and Reddy (2016) Aldenderfer and Blashfield (1984) Chiu and Tavella (2008) Demšar and Zupan (2013) Izenman (2008) Jacob and Ramani (2012) Jiawei et al. (2011) Kuo et al. (2006) Madhulatha (2011) Pierson and Porway (2017) Rechenthin (2014)	Market segmentation Product positioning New product development Selecting test markets Grouping of items Object recognition Recommendation system
1.1	Partitioning (Non-hierarchical) methods k-means k-medoids	Dean (2014) Jansen (2007) Jiawei et al. (2011) Jiawei et al. (2011) Lanjewar and Yadav (2013) MathWorks (2018b) Napoleon and Pavalakodi (2011) Rajarajeswari and Ravindran (2015) Ross (2005) Salazar et al. (2007) Tsiptsis and Chorianopoulos (2011)	Algorithms create single sets of clusters, most effective for small/medium datasets.

Continued on next page

	Clustering technique:	*Source:*	*Known for/Applications:*
1.2	Hierarchical methods Agglomerative (bottom-up) approach Divisive (top-down) approach	Chiu and Tavella (2008) Halkidi et al. (2001) Izenman (2008) Madhulatha (2011) Pierson and Porway (2017)	Algorithms create separate sets of clusters, each in their own hierarchal level (multiple levels).
1.3	Density-based methods DBSCAN DENCLUE	Jiawei et al. (2011)	The key idea is to group neighbouring objects of a dataset into clusters based on density conditions. It grows clusters either according to the density of neighbourhood objects (*e.g.*, DBSCAN) or according to a density function (*e.g.*, DENCLUE).
1.4	Grid-based methods STING CLIQUE	Bounsaythip and Rinta-Runsala (2001) Jiawei et al. (2011)	These algorithms are mainly proposed for spatial data mining. Their main characteristic is that they estimate the space into a finite number of cells and then they do all operations on the quantised space.

3.8.2.1.1 Clustering: *k*-means

Within this subsection, the first clustering technique indicated in Table 3.16 will be discussed. This technique is called k-means, which forms part of the partitioning (non-hierarchical) method. The researcher decided to only discuss k-means, for it is the most used clustering technique, while it provides a good foundation for understanding clustering and is a simple and elegant approach for partitioning a dataset into k distinct clusters.

It is important to understand the broad picture of how k-means perform clustering. It starts by choosing k representative points as the initial centroids. Each data point is then assigned to the closest centroid based on a particular proximity (distance) measure chosen. Once the clusters are formed, the centroids for each cluster are updated. Because k-means is an *iterative algorithm*, these two steps repeat until the centroids do not change or another pre-specified criterion is met. In practice, the iterative procedure must be continued until one percent of the points change their cluster memberships (Aggarwal and Reddy, 2016).

The explanation of k-means will follow an example by making use of the *Fisher's iris dataset*. This dataset is freely available and is a built-in dataset in Matlab.

The example to follow was performed on Matlab. Firstly, the Fisher's iris dataset needs to be loaded into Matlab. The dataset contains the natural groupings among iris specimens based on their sepal and petal measurements. When applying k-means, the number of clusters need to be prespecified.

To illustrate the data visually before applying the clustering technique k-means, Figure 3.20 represents the scatter graph of the sepal length, sepal width and petal length of the data.

From Figure 3.20, it is clear that there are groups in the data. The larger group or cluster appears to be split into a lower variance region and a higher variance region. This might indicate that the larger cluster is actually two, overlapping clusters. Without applying any techniques, it would be difficult to cluster this dataset into more than two clusters for it is difficult to decide where to separate the large group.

Therefore, the first step is to set the desired number of clusters to two, and make use of the squared *Euclidean distance*, for this metric is the most popular choice. Applying the distance metric provides an idea of how well-separated the resulting clusters are. To visually represent this, a *silhouette plot* is constructed. The silhouette plot displays a

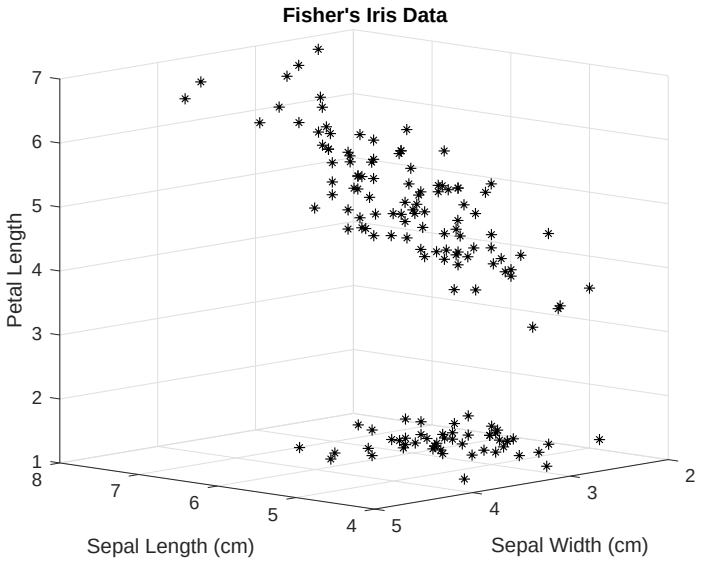


Figure 3.20: Scatter graph of Fisher's Iris Dataset before clustering is applied.

measure of how close each point in one cluster is to points in the neighbouring clusters. Figure 3.21 represents the silhouette plot when two clusters are present.

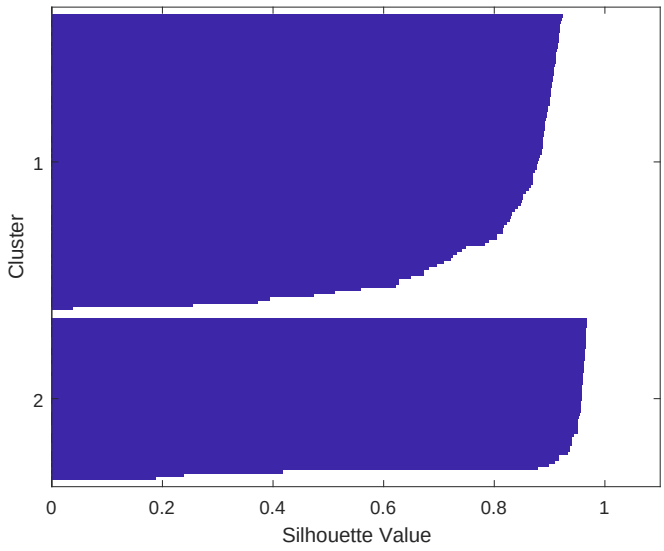


Figure 3.21: Silhouette plot for $k = 2$ applied to the Fisher's Iris Dataset

The silhouette plot indicates that most data points in both clusters have a large silhouette value. A silhouette value greater than 0.8 indicates that those data points are

well-separated from the neighbouring clusters. However, each cluster also contains a few data points with low silhouette values. This, on the other hand, indicates that they are close to data points from other clusters.

The *objective function* that is employed by k-means is called the *sum of squared errors* (SSE) or *residual sum of squares* (RSS). The mathematical formulation for SSE/RSS is

$$\text{SSE(C)} = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - c_k\|^2,$$

where c_k is the centroid of cluster C_k, and is denoted as

$$c_k = \frac{\sum_{x_i \in C_k} x_i}{|C_k|}.$$

As mentioned, the objective is to find a clustering that *minimises* the SSE score. This is where the iterative assignment and update step of the k-means algorithm aim to minimise the SSE score for the given set of centroids.

It turns out that the fourth measurement in this dataset, the petal width (not shown in Figure 3.20), is highly correlated with the third measurement, the petal length. For this reason, a 3-D plot of the first three measurements (Figure 3.20) gives a good representation of the data, without resorting to four dimensions.

Figure 3.22 shows the scatter graph after k-means was performed, with $k = 2$. The data points which have small silhouette values can be identified as those points that lie closer to the neighbouring cluster. The two clusters are distinguishable as the one cluster is represented as blue squares and the other as red triangles. In order to get the total sum of distances as small as possible, in other words minimise the SSE, Matlab had to perform three iterations.

The centroids of each cluster are represented by an '**x**', and can be seen in Figure 3.22. Three of the red triangle data points (cluster 2) seem to be very close to the lower data points of cluster 1. This occurs because cluster 1 is so spread out, those three points are closer to the centroid of cluster 2 than to that of cluster 1, even though they are separated from the bulk of the points in their own cluster by a gap. The k-means clustering technique only considers distances, and not densities, so this kind of result can occur.

By increasing the number of clusters, the researcher can test whether k-means can find further grouping structures in the data. Figure 3.23 represents the silhouette plot for $k = 3$.

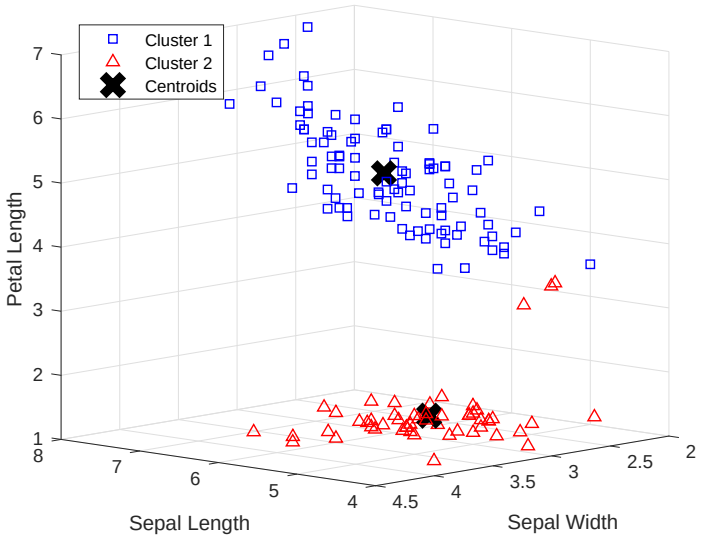


Figure 3.22: Scatter graph after clustering applied on the Fisher's Iris Dataset with $k = 2$

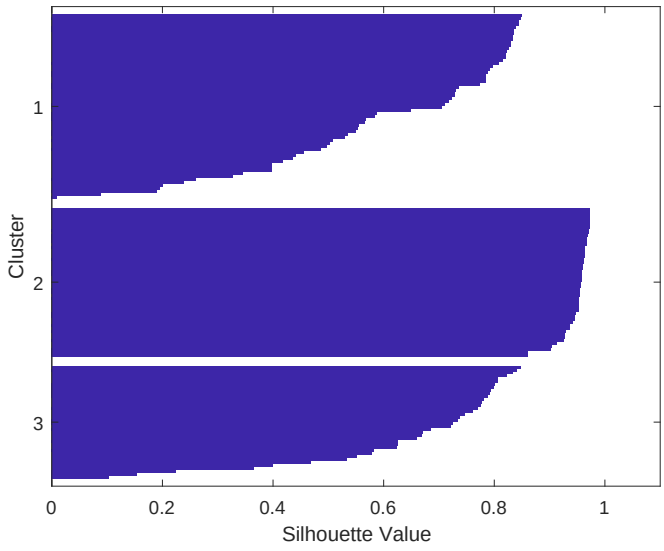


Figure 3.23: Silhouette plot for $k = 3$ applied to the Fisher's Iris Dataset

The silhouette plot for the three-cluster solution, as shown in Figure 3.23, indicates that there is one cluster that is well-separated (middle cluster), but that the other two clusters are not very distinct. Again, by plotting the data as a scatter graph, for $k = 3$,

the assignment of the data points can be visualised.

Figure 3.24 can be interpreted that k-means has split the upper cluster from the two-cluster solution (cluster 1 in Figure 3.22), into two clusters (pink diamonds and blue squares). Depending on what the indented purpose is of this dataset after clustering, this three-cluster solution may be more, or less useful than the previous, two-cluster, solution.

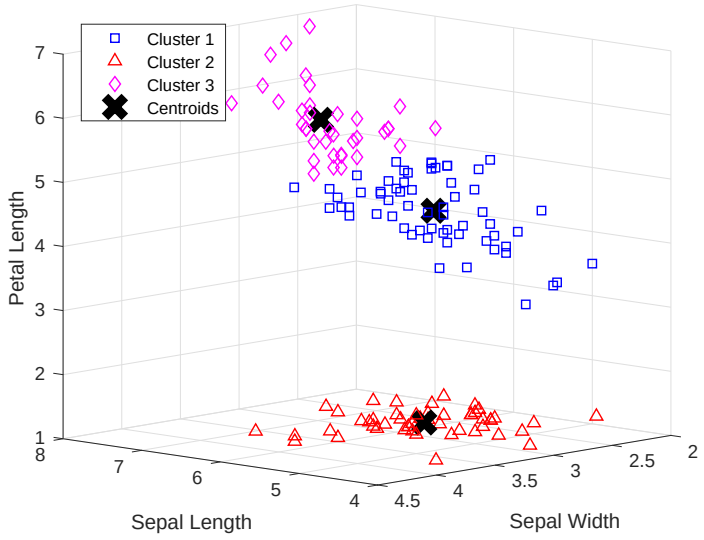


Figure 3.24: Scatter graph after clustering applied on the Fisher's Iris Dataset with $k = 3$

To conclude the discussion regarding k-means, the researcher will briefly discuss two major factors that can impact the performance of the k-means algorithm. The first factor is *choosing the initial centroids*. When using Matlab to perform k-means, the default for selecting the initial centroid is *via* the k-means++ algorithm. The algorithm follows a simple probability-based approach where initially the first centroid is selected at random. The centroid to follow is the one which is farthest from the current selected centroid. This selection is decided based on a weighted probability score. The selection is continued until there are k centroids and then k-means clustering is done using those centroids (Aggarwal and Reddy, 2016). There are other methods for selecting the initial centroid which include *Hartigan and Wong*, *Milligan*, and *Bradley and Fayyad* (Kaufman and Rousseeuw, 2009; Nimbalkar and Shah, 2013).

The second factor impacting the performance of k-means is *estimating the number of clusters* k. Several researchers have proposed new methods for addressing this challenge in the literature. The researcher will briefly discuss some of the most prominent methods. The first method is the *silhouette coefficient*, this is the method that was used for the example above, the Fisher's Iris dataset. The silhouette coefficient is formulated by considering both the intra- and inter-cluster distances.

For a given point x_i, first the average of the distances to all the data points in the same cluster is calculated. This value is set equal to a_i. Then for each cluster that does not contain x_i, the average distance of x_i to all the data points in each cluster is computed. This value is set equal to b_i. Using these two values the silhouette coefficient of a point is estimated. The average of all the silhouettes in the dataset is called the average silhouette width for all the points in the dataset. To evaluate the quality of clustering, the average silhouette coefficient is calculated for all the data points (Kaufman and Rousseeuw, 2009). This calculation is done using

$$S = \frac{\sum_{i=1}^{N} \frac{b_i - a_i}{max(a_i, b_i)}}{N}, \quad N = \text{number of data points.}$$

Looking at Figure 3.21 and 3.23, the first output argument from the silhouette plot contains the silhouette values for each data point. These values can be used to compare the two solutions quantitatively. It can be concluded that the average silhouette value was larger for the two-cluster solution (Figure 3.21). This indicated that it is a better answer purely from the point of view of creating distinct clusters.

The second method to estimate the number of clusters k is the *Calinski–Harabasz Index*. The Calinski–Harabasz index is defined by

$$\mathrm{CH}(K) = \frac{\frac{B(K)}{(K-1)}}{\frac{W(K)}{N-K}}.$$

The number of clusters is then chosen by maximising the Calinski–Harabasz index function. The B(K) and W(K) are the between and within cluster sum of squares, respectively (Caliński and Harabasz, 1974).

The third method to estimate the number of clusters k is the *gap statistic*. When using this method, B different datasets, each with the same range of values as the original data, are produced. The within cluster sum of squares is calculated for each of them

with a different number of clusters. $W_b^*(K)$ is the within cluster sum of squares for the b^{th} uniform dataset. The gap statistic equation is

$$\text{Gap}(K) = \frac{1}{B} \times \sum_b \log(W_b^*(K)) - \log(W(K)).$$

The number of clusters chosen is the smallest value of K that satisfies the gap statistic equation.

There are a few more methods to estimate the number of clusters k namely, *Akaike Information Criterion* (Yeung et al., 2001), *Bayesian Information Criterion* (Mojena, 1977), *Duda and Hart* (Duda and Hart, 1973), *Newman and Girvan* (Newman and Girvan, 2004) and *ISODATA* (Ball and Hall, 1965)

This brings the discussion regarding clustering to an end. The k-means technique provides insight into the main objective of clustering. For further reading on this or other clustering techniques, Table 3.16 provides applicable references.

Table 3.11, 3.12 and 3.16 provided a brief overview of data mining techniques. Not all of them will be utilised within this research; therefore the researcher did not describe them in detail. The review concerning data mining tools and techniques is now drawn to a close. Next, a synthesis of this chapter will be presented.

3.9 Synthesis: Literature review

The literature review performed in Chapter 3 was initiated by defining the term *data* and exploring the four types of data forms. It was found that the most commonly occurring data form is the relational dataset, with each attribute having a different data type, broadly defined as either quantitative or categorical. For the purpose of this study it was necessary to define, not only the term data, but also what is meant when referring to *Big Data*. Data becomes Big Data when it is too difficult to process datasets using traditional techniques. Therefore, Big Data can be seen as any voluminous amount of data of various formats that has the potential to be mined for information. It was concluded that Big Datasets consist of numerous dimensions, referred to as the V's, with volume being only one of the dimensions. The tools that assist the transformation of raw voluminous data into Big Data with trustworthy insights and to discard the noise is called *Big Data Analytics*.

Big Data Analytics is a relative new term to describe the analysis of Big Data. Within this chapter Big Data Analytics was defined as the entire methodology that is utilised for the analysis of big datasets, in order to create value for an enterprise. As one can anticipate, there are various types of data analytics that can be applied to a dataset. When trying to establish what happened previously, one would perform descriptive analytics; in trying to predict what is going to happen, predictive analytics may be performed; in demanding to understand why an incident occurred, diagnostic analytics can be applied; and lastly when faced with what can be done next, prescriptive analytics may supply the answer.

A framework was created while running a workshop to get a clear understanding of what Big Data Analytics entails. The body of the literature presented in this chapter followed the framework that was constructed and schematically presented in Figure 3.5. It was established that Big Data Analytics contains various processes, such as the KDD process, CRISP and the SEMMA process. These processes consists of numerous steps or phases. Each process contains a *data preparation* phase, which in this context means to manipulate the data into a form suitable for analysis. The data preparation phase includes two main steps, namely *data cleaning* and *data transformation*.

Data cleaning is evident in all three analytic processes as step 3. Data cleaning aims to improve the quality of data prior to the analysis by detecting and removing missing values, erroneous data entries and outliers. After completing the data cleaning phase, the data may still not be ready for mining.

This provided reasonable cause for highlighting the next step, namely *data transformation*. The transformation step is evident in the KDD process as step 4, while the CRISP and SEMMA processes perform data transformation in step 3. Transformation of data is necessary when a dataset contains many features or attributes, increasing the dimensions of the dataset. This is where dimensionality reduction techniques, such as PCA, provide assistance, with the goal being to introduce high-dimensional data in a lower dimensional subspace, while the essential features of the original data are kept as far as possible.

The literature regarding data preparation indicated that there is no unique procedure and the only criterion is to clean and transform the data for convenience of use. After the completion of the data preparation phase, the dataset should be 'clean' as well as 'transformed'. Following the data preparation phase is the *data mining* phase. Exploring

the data mining phase, it was recognised that data mining consists of various tools and techniques/tasks, which are collectively known as machine learning.

The two main data mining tools are called *supervised* and *unsupervised learning*. Supervised learning is the form of machine learning most widely utilised in practice. Supervised learning is a machine learning tool that is given a specific goal for grouping the data, for example to predict the target. This tool can be divided into two techniques, namely *classification* and *regression*.

Classification is a well-known data mining technique that assigns items to discrete, previously learned classes and automatically predicts the class to which a new item will belong to. The popular technique called *decision trees* initiated the discussion regarding classification techniques. When using a binary decision tree, a tree induction model with a "Yes–No" format can be built to split the data into different classes, according to its attributes. However, the classification obtained from tree induction may not produce an optimal solution where prediction power is limited. In such cases, building a *neural network* model could have more advantages. The attributes become input layers in the neural network, while the classes associated with the data are the output layer. Between the input and output layer, there is a larger number of hidden layers processing the accuracy of the classification. Although neural networks yield better results in many cases, the network involves complex non-linear relationships, and implementing this technique on large sets of attributes is a very difficult task. This provided a brief overview of two techniques discussed in Chapter 3. There are, however, numerous techniques discussed, along with their application areas.

Regression, together with classification, forms the two supervised learning techniques. Regression is one of the most widely used statistical methods and forms the basis for many other statistical techniques. The key idea of regression is to discover the relationship between the dependent and independent variables. For example, if sales is an independent variable, the profit may be a dependent variable. Making use of the historical data for both sales and profit, either *linear* or *non-linear regression* techniques may be performed, for both will be able to produce a fitted regression curve for profit prediction in the future. Similar to classification, various regression techniques were discussed in this chapter, along with their application areas.

Unlike supervised learning, unsupervised learning does not have a dependent variable, nor does it require a learning set. The methods are mainly descriptive, searching for

unknown patterns or relationships. The unsupervised learning tool discussed is clustering. The aim of clustering is to take ungrouped data and utilise automatic techniques to place the data into groups with similar properties. The k-means clustering technique was discussed and performed in this chapter to provide a good foundation for understanding clustering. Different clustering techniques were mentioned, as well as their application areas and sources for further reading.

After completing the literature review, the researcher noticed that very little has been done to apply unsupervised learning and supervised learning in conjunction, at least in the Industrial Engineering domain. This raises concerns, for this research focuses on applying both data mining tools, in sequence, to the same customer dataset. Table 3.17 provides a brief summary of the number of articles, in various journals, that address the key phrase, *Big Data Analytics*. A few of the articles that are used in the keyword count in Table 3.17 utilise Big Data Analytics in a different field and for a different purpose than what it is intended for in this research. This offers an opportunity to investigate the integrated use of both data analytics tools. This concludes the review of Big Data Analytics. Next, a chapter summary is provided.

Table 3.17: Summary of journal articles containing the keyword: 'Big Data Analytics'

Journal:	*Keyword count:*
Computers & Industrial Engineering (CIE Journal, 2018).	9
The South African Journal of Industrial Engineering (SAJIE, 2018).	2
Journal of Industrial and Production Engineering (JIPE, 2018).	3
IIE Transactions (IIE Transactions, 2018).	1

3.10 Summary: Chapter 3

This chapter reviewed important concepts that form part of the theoretical basis for this study. A brief definition of *data* was followed by a discussion of the various forms and types within the context of *data analysis*. This was followed by examining what is meant by the term *Big Data*, as well as defining the different dimensions that characterise this term.

An introduction to *Big Data Analytics* was then provided, following a framework that illustrates various data analytic processes that consist of numerous steps. The

different processes were defined and comparisons were drawn between them. This was followed by a description of the data cleaning process, including means of identifying and treating missing values, erroneous data entries and outliers. Following the cleaning stage was the data transformation stage. The transformation stage was initiated by discussing dimensionality reduction together with the well-known technique associated with it, namely PCA. These two processes form the *data preparation* phase, which is followed by the *data mining* phase.

Data mining was explored by outlining the two main tools, *supervised* and *unsupervised learning*, subsequent to which important techniques of both of them were discussed in more detail. Tables 3.11, 3.12 and 3.16 were constructed to summarise the findings regarding both supervised and unsupervised learning techniques.

Finally, the chapter concludes by providing a synthesis of the review performed. In the next chapter, the architectural development will commence, with support of Chapters 2 and 3.

Chapter 4

Architectural development

The previous chapter provided a comprehensive review of *Big Data Analytics*, as well as highlighting major phases, namely data cleaning, data transformation and data mining. The discussion of data mining led to defining machine learning tools and techniques, introducing the concept '*learning from data*'. This chapter presents the introduction to developing an analytics tool that has the ability to perform data analytics techniques on a real-world large database, and to yield reliable customer profiles that are acceptable to marketers.

First, the proposed solution architecture for the Customer Super-Profiling (CSP) tool is presented. Next, a *toy problem* is used to provide an overview of the proposed architecture. Then a second problem is studied based on a large dataset, making use of the methodology of the toy problem to illustrate the segmentation and profiling processes. Finally, a summary containing the focus areas will conclude this chapter.

4.1 Development of a solution architecture for super-profiling

The aim of the thesis is to develop a CSP tool containing a suite of Big Data Analytics tools and techniques which will allow for super-profiling. The need for a simulator that creates big datasets was identified. The datasets will be used by the CSP tool to illustrate the concept of super-profiling.

The solution architecture of the simulator and demonstration tool will be developed by following the *Object–Process Methodology* (OPM). This is an ISO19450 standard and it includes a clear and concise set of symbols that form a language, enabling the

expression of the system's building blocks and how they relate to each other. The OPM is a symbolic representation of the structural relations between objects in a system and its processes. *Objects* are what a system or a product is, while *processes* are what a system does. The OPM represents the system simultaneously in a graphic representation and in a natural language (semantic). The two are completely interchangeable, and they represent the same information. The OPM not only represents both objects and processes, but it also clearly shows the connections between them (Dori, 2011).

The concept of developing systems in a unified frame of reference is not new; on the contrary, as early as 1981 researchers noted that systematic development of basic concepts leads to methods that cover the system's entire life cycle. However, the OPM has a novel approach when modelling complex systems that include humans, physical objects and information.

The OPM does not make assumptions regarding the nature of the system being examined, and it can be applied in any domain of human study or endeavour. Both natural and artificial systems exhibit three major aspects:

1. *Function*: What these systems do.

2. *Structure*: How they are constructed.

3. *Behaviour*: How they change over time.

The OPM combines formal yet simple graphics, with natural language sentences to express the function, structure and behaviour of systems in an integrated, single model. The OPM is a mature tool/instrument for performing tasks that are involved in system development, and it does so in a direct and obvious manner. The initial solution architecture was developed and presented by Walters and Bekker (2017), to provide a broad overview; however, for the purpose of this thesis the original/initial OPM will be altered slightly to provide more detail. The *three* main components which constitute the simulator and demonstration tool proposed for predicting the customer profile for a targeted marketing campaign are 1) the simulating process, 2) the segmenting processes and 3) the profiling process. Every artificial system is designed to execute a certain *function*. A combination of objects and processes, defined as the system's *architecture*, enables the execution of this function. Figure 4.1 illustrates the top-level Object Process Diagram (OPD), as well as the interaction and flow of information between the components. The

description that follows, regarding the system diagram as illustrated in Figure 4.1, is a top-level OPD of the CSP tool, designed to move towards the aim, which is to develop customer super-profiles to enable efficient targeting in marketing campaigns. The simulation process consumes raw input data and in return yields customer data. This customer data is in a specific format and structure. The segmenting process consumes the customer data and requires the appropriate machine learning algorithms, in order to yield customer groups/clusters. These customer groups/clusters are consumed by the customer super-profiling process. This process also requires the appropriate machine learning algorithms as well as data descriptor analysis (together with the consumed objects) in order to yield customer profiles for campaign(s). Lastly, the produced customer profiles relate to the campaign success.

When a new customer 'enters' the system, it is possible to classify them without starting from the initial point of departure (segmenting). The classifying new customer(s) process consumes the new customer(s), and requires the customer groups created by the segmenting process as well as machine learning algorithms. The new customer(s) is then classified into the appropriate customer group, and the classifying new customer process yields new customer groups containing the new customer(s). The profiling process can then be followed, as mentioned earlier.

Various sets of classification rules can be determined by consuming the customer groups. The process of determining these classification rules will require machine learning algorithms and will yield a predictive model.

Object Process Language (OPL) is the counterpart of the graphic OPM system specifications. The OPL is automatically generated as a textual description of the system in a subset of natural English. The OPL is extracted from the diagrammatic description in the OPD set. Following the OPM guidelines, the OPL for Figure 4.1 is:

Customer Profiles for Campaign(s) relates to Campaign Success.
Customer Super-Profiling requires Data Descriptor Analysis and Machine Learning Algorithms.
Customer Super-Profiling consumes Customer Groups/Clusters.
Customer Super-Profiling yields Customer Profiles for Campaign(s).
Segmenting requires Machine Learning Algorithms.
Segmenting consumes Customer Data.
Segmenting yields Customer Groups/Clusters.
Simulating Process consumes Input Data.

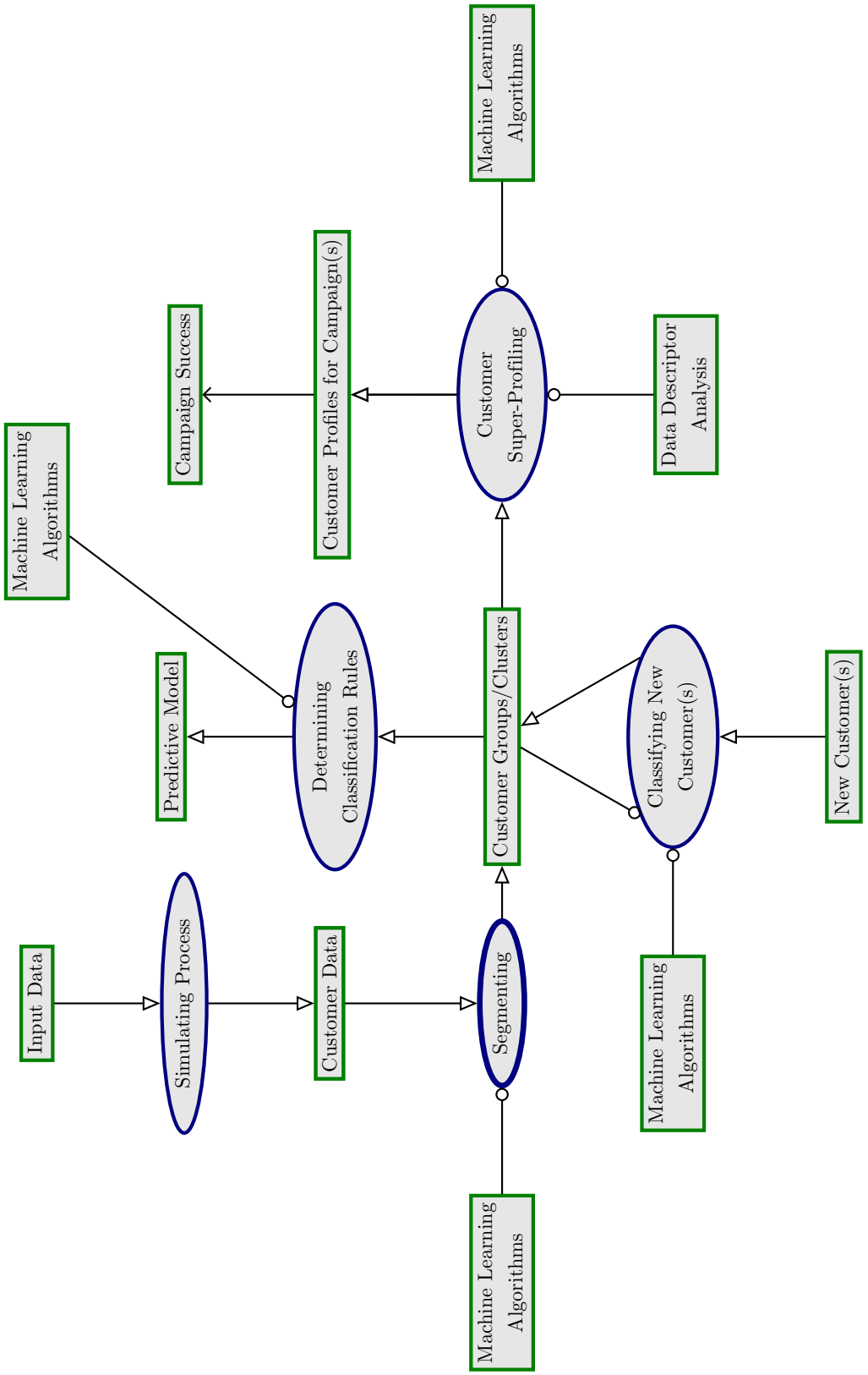


Figure 4.1: System diagram representing the top-level OPD of the proposed simulator and CSP tool

Simulating Process yields Customer Data.
Classifying New Customers requires Customer Groups/Clusters and Machine Learning Algorithms.
Classifying New Customers consumes New Customer(s).
Classifying New Customers yields Customer Groups/Clusters.
Determining Classification Rules requires Machine Learning Algorithms.
Determining Classification Rules consumes Customer Groups/Clusters.
Determining Classification Rules yields Predictive model.

Adding more detail to the top-level OPD (Figure 4.1) causes confusion. Therefore, refinement is necessary. The *'Segmenting'* process of the system diagram (Figure 4.1) is *zoomed-in*, as seen in Figure 4.2. Focusing on the result of the zooming-in operation, a few of the objects outside the zoomed-in process have temporarily been omitted.

The description to follow, regarding the zoomed-in diagram as illustrated in Figure 4.2, provides more detail to the segmenting process. The segmenting process consists of two processes; preparing customer data and customer segmenting. The preparing customer data process is similar to the data preparation phase of Big Data Analytics. This process consumes the customer data and yields prepared customer data. Customer segmenting consumes the prepared customer data and requires machine learning algorithms in order to yield customer groups/clusters.

To summarise this architecture; it is established that customer data in a specific format and structure are utilised, customer segmentation is then performed on this dataset which leads to the profiling of customers. The simulated customer dataset provides the means not to perform market segmentation, but only customer segmentation, for the entire dataset is seen as the *'market'*.

The OPL for Figure 4.2 is as follows:

Simulating Process consumes Input Data.
Simulating Process yields Customer Data.
Segmenting exhibits Prepared Customer Data.
Segmenting consist of Preparing Customer Data and Customer Segmenting.
Segmenting zooms into Customer Segmenting and Preparing Customer Data, as well as Prepared Customer Data.
Customer Segmenting requires Machine Learning Algorithms.
Customer Segmenting consumes Prepared Customer Data.
Customer Segmenting yields Customer Groups/Clusters.
Preparing Customer Data consumes Customer Data.
Market Segmenting yields Prepared Customer Data.

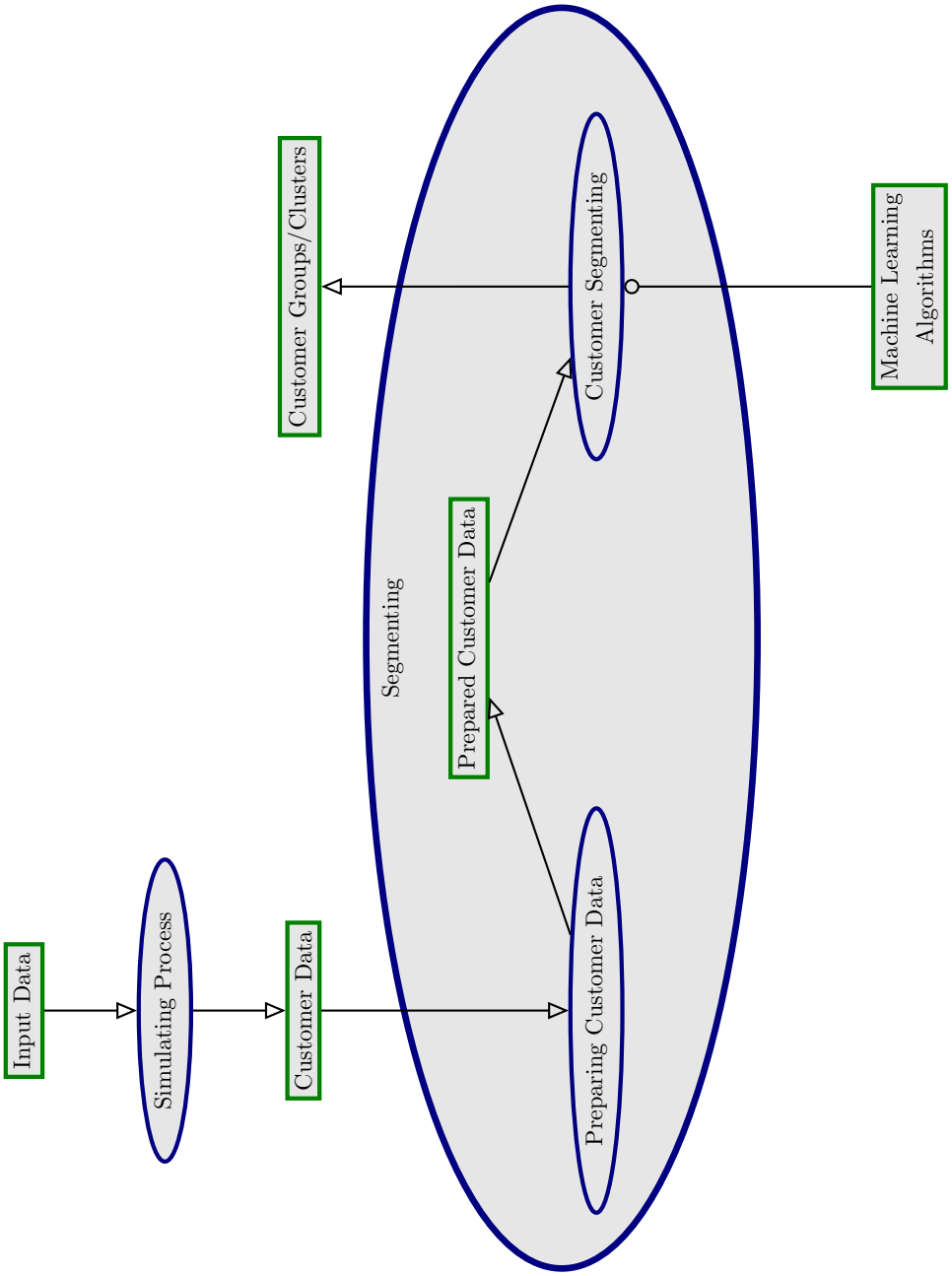


Figure 4.2: The zoomed-in segmenting process

OPM is powerful since it presents a system architecture in visual and textual format.

4.2 Toy problem and large dataset problem

This section provides an overview of the proposed architecture described in the previous section, by using a *toy problem* (Walters and Bekker, 2017). However, the processes: '*determining classification rules*' and '*classifying new customer(s)*' will not be performed by either the toy problem or the large dataset problem. A toy problem is a problem that is not of immediate scientific interest, yet it is a useful tool to create a simplified version of a complex problem, that is used to demonstrate the concept of a proposition. The toy problem will deliberately oversimplify the *customer segmentation* and *profiling* process. The segmentation process will be performed by using Matlab and its built-in clustering function.

Market segmentation was performed on a very small dataset that contains N=100 customers with many attributes; segmentation divided the customers into different geographical regions. After market segmentation took place, suppose a grocery store in region X (one of the market segments) desired to obtain more information about their customers; this included *demographic* and *behavioural* information, in order to promote targeted campaigns. To obtain this information, the customers of the segmented group (region X) will be 'filtered' to consist of only the customers that purchased at the specific local grocery store, and then *customer segmentation* will be performed on that group (less than the original 100 customers). The customer segmentation process will follow the Recency, Frequency and Monetary (RFM) model approach, determining the R, F and M parameter value for each customer. The ease of use and quick implementation are the reasons that researchers and marketers continue to employ RFM models. They are also easily understood by managers and decision-makers (Sarvari et al., 2016). After using the RFM model to represent *customer behaviour*, the data will be encoded by dividing the values of *recency*, *frequency* and *monetary* into five categories. If the value lies between 100 percent and 80 percent, the category value is 5, between 80 percent and 60 percent, the value is 4; *etc.* There is no need to perform data preparation on this dataset, for it is 'clean' and the RFM model approach utilises its own categories for normalisation.

For this toy problem, the well-known k-means clustering algorithm will be used for segmentation, as the clustering quality and runtime of the algorithm is reliable, and it is one of Matlab's built-in functions. To determine the optimal number of clusters, as the k-means algorithm largely depends on selecting the proper number of clusters, *silhouette plots* were generated as seen in Figure 4.3. The mean value of each silhouette plot is indicated in Table 4.1. To determine the optimal number of clusters from the silhouette plots (Figure 4.3), the mean value should be as close to one as possible; ideally the plot of each cluster should be above the mean value, and lastly the width of the plot should be as uniform as possible. According to Martinez et al. (2010), an average silhouette width greater than 0.5 indicates a reasonable partition of the data, while a value less than 0.2 would indicate that the data does not exhibit cluster structure. It can be observed from Table 4.1 that four clusters would be the optimal for this problem.

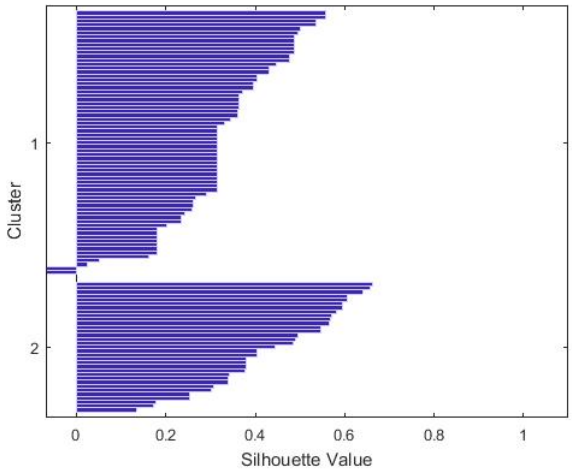
Table 4.1: Silhouette mean values of the toy problem dataset

$k =$	2	3	**4**	5	6
Mean value	0.3629	0.5582	**0.6030**	0.5136	0.5423

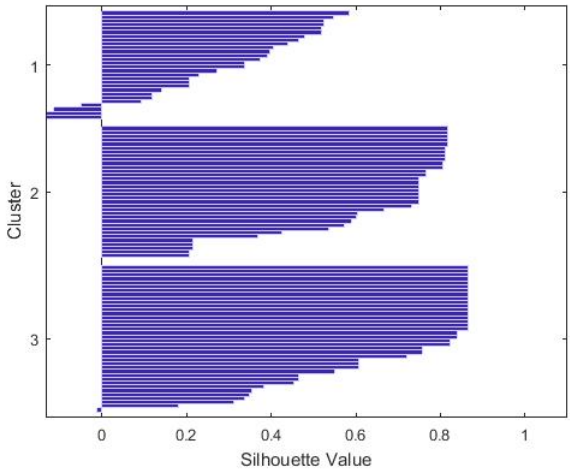
After investigation of Figure 4.3c, it is observed that there is a negative value present, which means that a customer was misclassified in cluster 3; and would be a better fit in a neighbouring cluster. Before determining the neighbouring cluster, the k-means cluster analysis was performed on the categorical dataset (containing the misclassified customer), with an input of $k = 4$. The four clusters can be seen in Figure 4.4. After the k-means analysis is completed, and each cluster's customers are known, it is easier to determine which cluster the misclassified customer belongs to. Table 4.2 indicates the customer that has been misclassified, its RFM-value and which cluster would be a better fit. Originally cluster 3 had 14 customers and cluster 4 had 34 customers; now cluster 3 has 13 customers, and cluster 4 has 35 customers. Figure 4.5 represents the four segments and the segmented population sizes.

Table 4.2: Misclassified customer(s) of the toy problem dataset

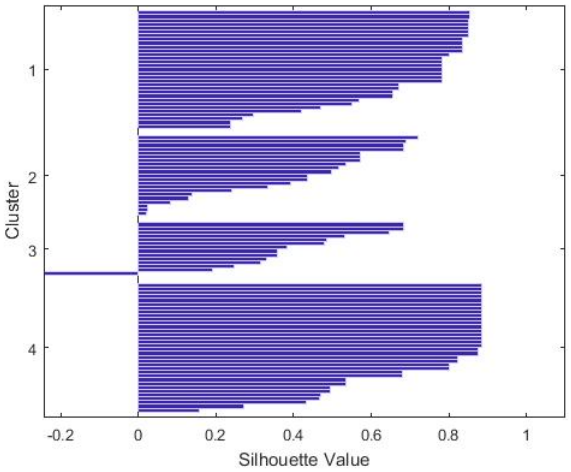
Customer Index	***Cluster***	***Silhouette Value***	***RFM-Value***	***Neighbour Cluster***
78	3	-0.2440	2-3-2	4



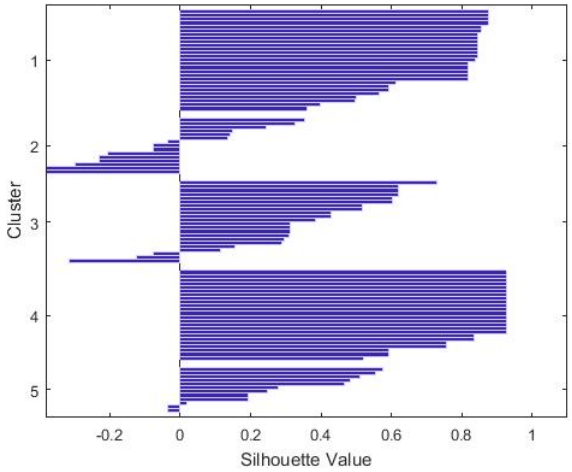
(a) Silhouette plot for $k = 2$



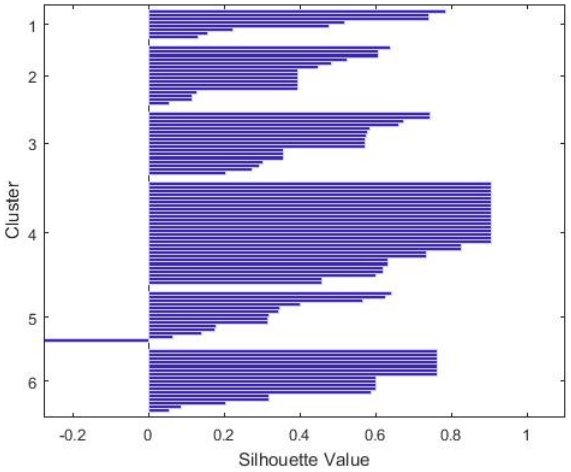
(b) Silhouette plot for $k = 3$



(c) Silhouette plot for $k = 4$



(d) Silhouette plot for $k = 5$



(e) Silhouette plot for $k = 6$

Figure 4.3: Silhouette plots for $k = 2$, 3, 4, 5, 6 of the toy problem dataset

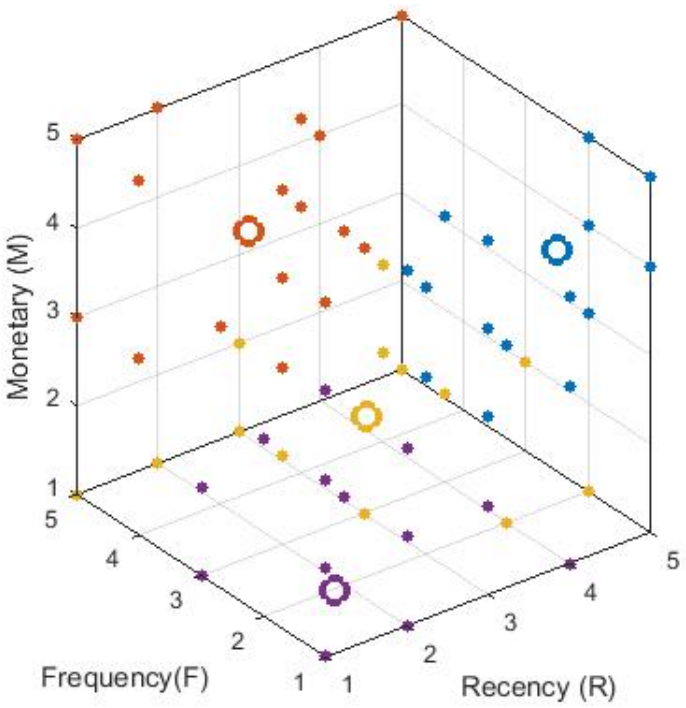


Figure 4.4: Scatter plot representing the four clusters of the toy problem dataset

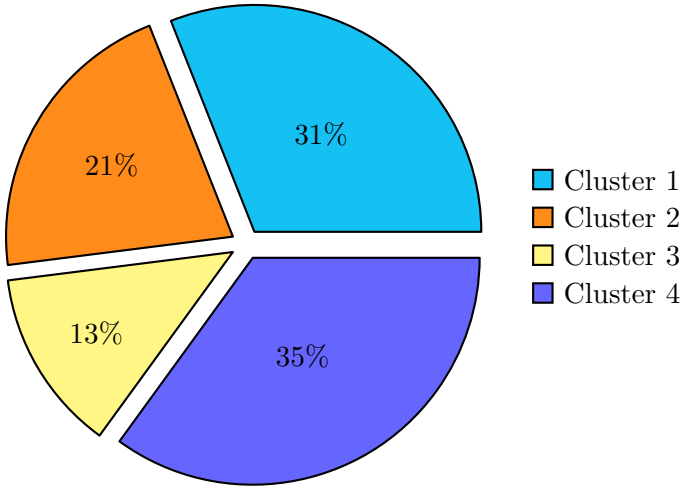


Figure 4.5: Pie chart indicating the four cluster sizes of the toy problem dataset

After this, additional information regarding each cluster and the customers that are grouped together within that cluster could be extracted. The grocery store could specify

what information they are interested in regarding the customers and further analysis could take place to create customer profiles containing the specified information. For this problem, demographic information is required. Therefore, each customer within a cluster is profiled according to their *gender*, *age*, *occupation*, *annual income*, *marital status* and whether or not they have *children*. This information is important for the grocery store, for it affects the customers' purchasing behaviour, and the store then has the ability to identify marketing strategies for their campaign programmes.

Next, individual cluster analysis will be performed, starting with cluster 1. Figure 4.6 illustrates the RFM ratio of cluster 1. The RFM parameter frequencies vary from one to five, with five being the highest '*score*'. It can be interpreted that most of the customers in cluster 1 (48 percent of the customers as seen in Figure 4.6) have a high recent purchase value (R=5), while 45 percent of the customers also have a high monetary value (M=5), but 58 percent of the customers have a low frequency value (F=1).

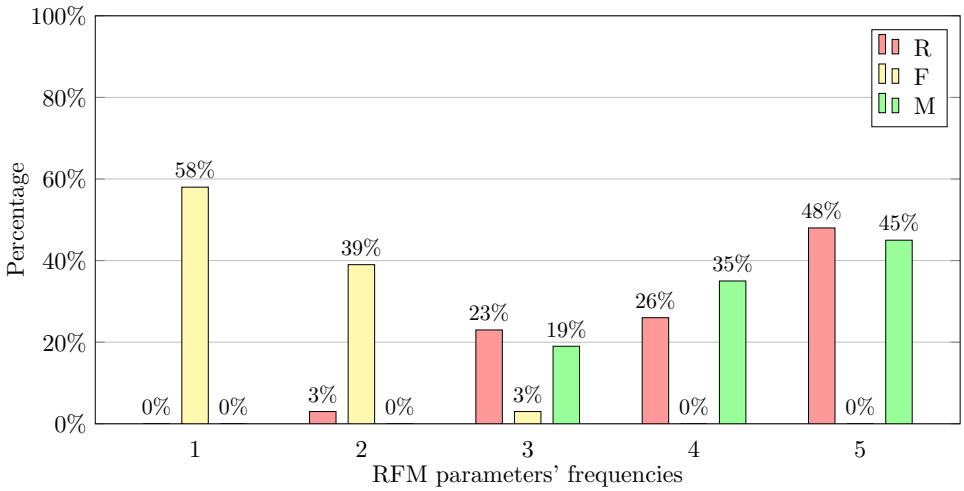


Figure 4.6: Cluster 1 of toy problem – RFM ratio

A summary of the customers' demographic information, providing the gender, age, occupation, annual income and children status are shown in Table 4.3. This summary may lead to understanding the low F-value and high R and M-values. The customers' marital and children statuses are intentionally not added, for they will not add value to this customer group.

Figure 4.7 illustrates the RFM ratio of cluster 2. As mentioned, the RFM parameter frequencies vary from one to five, which is the highest. As seen in Figure 4.7, most of the

Table 4.3: Cluster 1 of toy problem – Customer demographics

Gender	Male	59%	**Occupation**	Student	47%	**Annual**	**Income**	Low	47%	**Age**	16-32	73%
	Female	41%		Employed	21%						48-62	6%
				Retired	32%			Medium	53%		62-80	32%

customers in cluster 2 (48 percent of the customers) have a high monetary value (M=5). This occurs when a customer spends an excessive amount over a specified period, 29 percent of the customers have a medium frequency value (F=3), another 29 percent of customers have a good frequency value (F=4), and another 29 percent have an excellent frequency value (F=5). This indicates that the customers in cluster 2 frequently purchase at the grocery shop; however, the majority of the customers in cluster 2 (43 percent of the customers) have a low recent purchase value (R=2). This occurs when the time duration between the last purchase and the time of the survey is long, indicating that the majority of the customers in cluster 2 did not purchase recently.

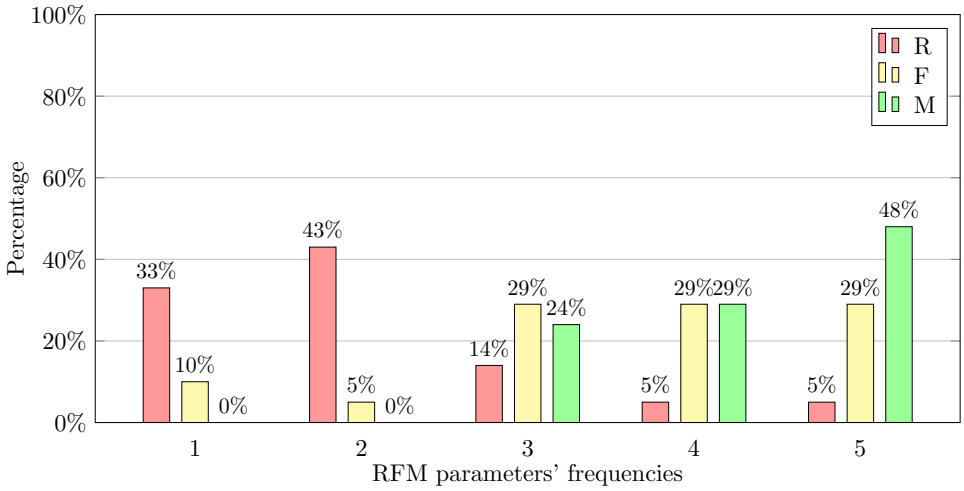


Figure 4.7: Cluster 2 of toy problem – RFM ratio

Table 4.4 provides a summary of the customers' demographic information, providing the gender, age, occupation and annual income of the customers, which may lead to explaining the low R-values and the high M-values. The customer's marital status is intentionally not added, for it will not add value to this specific customer cluster.

Table 4.4: Cluster 2 of toy problem – Customer demographics

Gender	Male	12%	**Occupation**	Employed	75%	**Annual**	**Income**	Medium	75%	**Age**	16-32	25%	**Children**	Yes	75%
	Female	88%		Home-maker	25%			High	25%		32-48	50%		No	25%
											48-62	25%			

Figure 4.8 represents cluster 3, the smallest cluster. Most of the customers in cluster 3 (47 percent) have a low monetary value (M=1), yet 40 percent of the customers have a high frequency value (F=5), while the recency value is spread from low to high values, with most customers falling in the R=4 class. Examining the demographic information of cluster 3 may provide more insight into their RFM scores. Table 4.5 provides a summary of the customers' demographic information, providing their gender, age, occupation, marital and children statuses.

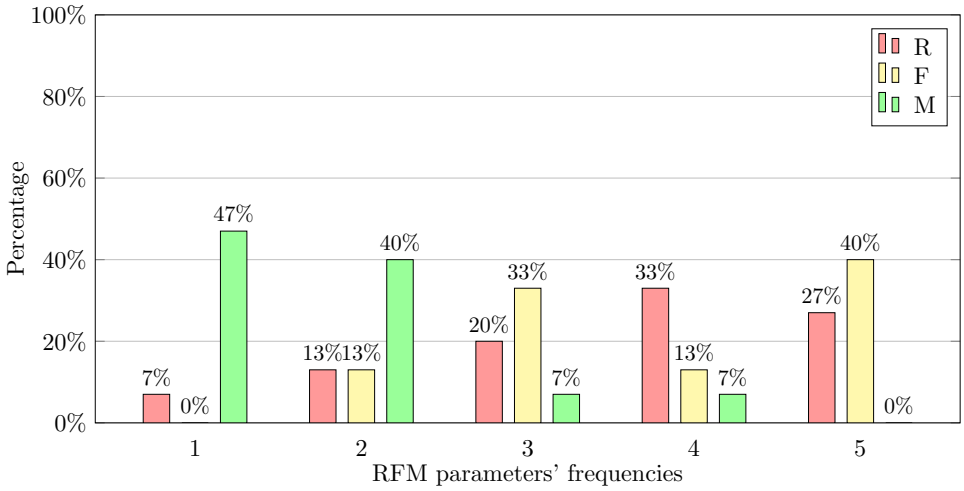


Figure 4.8: Cluster 3 of toy problem – RFM ratio

Cluster 3 displays fairly similar behaviour to cluster 2, regarding the gender and children statuses. However, when comparing cluster 3 to cluster 1, cluster 3 shows a major difference. Customers in cluster 3 differ from cluster 1 customers regarding the age ranges, gender and occupation, therefore the marital and children statuses were added, because cluster 3 customers are in different life stages from customers in clusters 1 and 2, and will require a different marketing/campaign strategy.

In contrast to the previous clusters, cluster 4 (the biggest cluster) displays entirely

Table 4.5: Cluster 3 of toy problem – Customer demographics

Gender	Male	20%	**Marital**	**Status**	Yes	67%	**Occupation**	Employed	60%	**Age**	32-48	73%	**Children**	Yes	87%
	Female	80%			No	33%		Home-maker	40%		48-62	27%		No	13%

different customer behaviour, as seen in Figure 4.9. The majority of the customers in cluster 4 have low R, F and M-values. Therefore, it is necessary to analyse the customer demographics of this cluster. Table 4.6 provides a summary of the customers' demographic information, providing the gender, occupation, annual income and age of the customers, which may lead to explaining the low R, F and M-values. Table 4.6 is similar to Table 4.3 (cluster 1), yet Figure 4.9 is very different from Figure 4.6. The researcher has shown here what customer profiling is by means of collected and analysed customer demographic information.

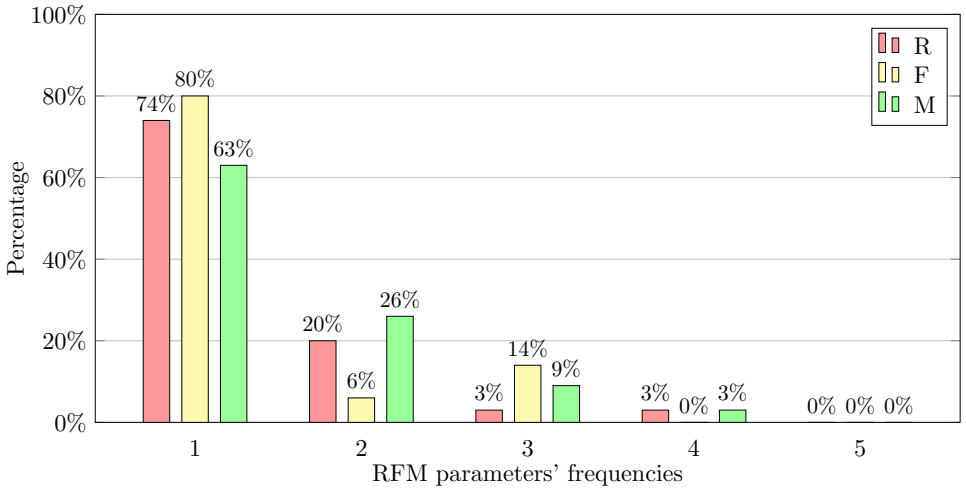


Figure 4.9: Cluster 4 of toy problem – RFM ratio

Table 4.6: Cluster 4 of toy problem – Customer demographics

Gender	Male	45%	**Occupation**	Student	46%	**Annual**	**Income**	Low	70%	**Age**	16-32	73%
	Female	55%		Employed	39%			Medium	30%		48-62	12%
				Retired	15%						62-80	15%

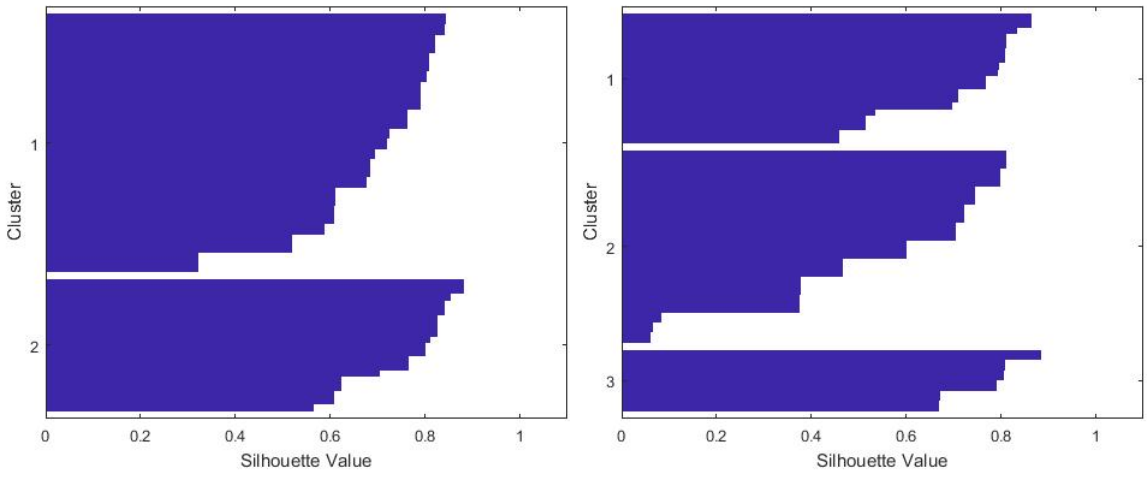
As demonstrated by the toy problem, each cluster is unique, just as the individual customers are unique. Each cluster has its own properties; therefore, the marketing strategy for each cluster will differ. At the conclusion of this toy problem, this information is presented (sold) to the local grocery store, and its marketing team can use the results to prepare and run targeted campaigns.

The researcher studied a second problem based on a large dataset, making use of the methodology of the toy problem to illustrate the segmentation and profiling processes on a different, less conventional, domain. The new problem will not focus on the purchasing behaviour of customers (as in the toy problem), but rather on the *participation* behaviour of *campers* (as opposed to *purchasing* behaviour). The researcher used the American Camping Report of 2014 as a platform on which to base this problem (American Outdoor Foundation, 2014). This is done to make the RFM model and analysis more realistic.

The RFM model is also used for this problem, with the survey period being a year (January to December). The R indicates how recently the individuals participated in camping, the F indicates the number of camping trips within the survey period, and the M indicated the amount of money spent while camping (all the trips added together). The length of the trip is not taken into consideration, for the information received from the profiling process is utilised by the outdoor domain for marketing. The frequency (number of times camped) of campers would be more useful to them than the number of days spent camping. For instance, when the frequency of campers is known, the outdoor domain could target high-frequency individuals with advertisements, more so than those with a low frequency. The dataset used for this problem consists of 100 000 records of individuals.

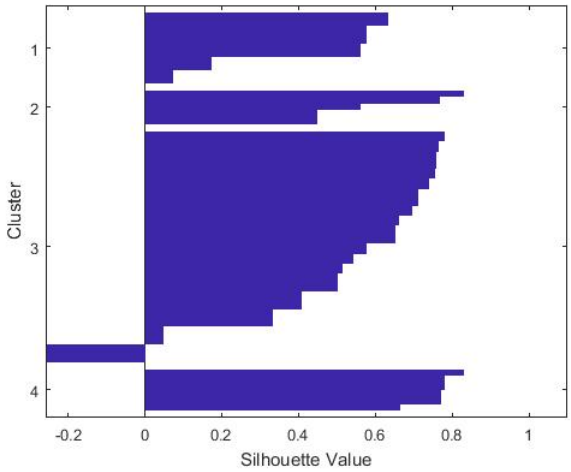
The first step is to determine the number of clusters to use in the analysis, by generating silhouette plots. Table 4.7 represents the mean value for each silhouette plot, and it is observed that less misclassification of data points occurs when the dataset is bigger. The mean value for $k = 2$ is closer to one than $k = 3$; however, the cluster sizes of $k = 3$ are closer to each other than that of the $k = 2$ cluster sizes (Figure 4.10). Therefore, the researcher decided to select three as the best number of clusters, for Figure 4.10b and Table 4.7 indicate that three clusters are still seen as a reasonable structure (Martinez et al., 2010).

The k-means cluster analysis was performed on the dataset, with $k = 3$, as seen in Figure 4.11. After further analysis, Figure 4.12 represents the three segments and

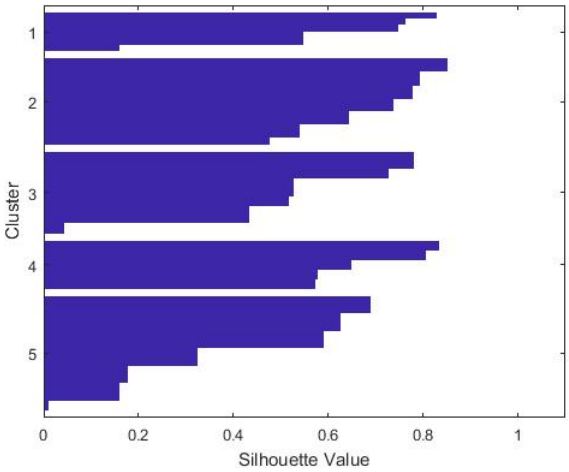


(a) Silhouette plot for $k = 2$

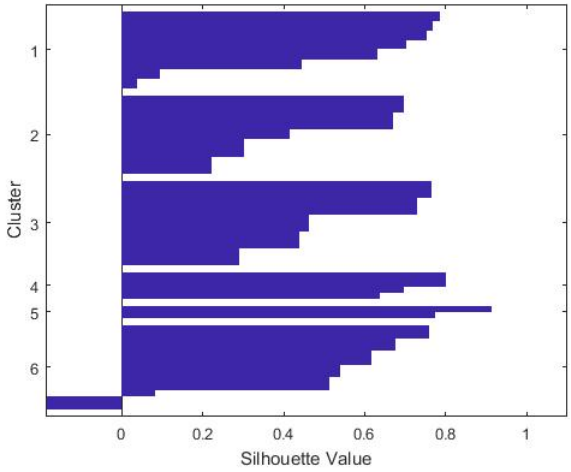
(b) Silhouette plot for $k = 3$



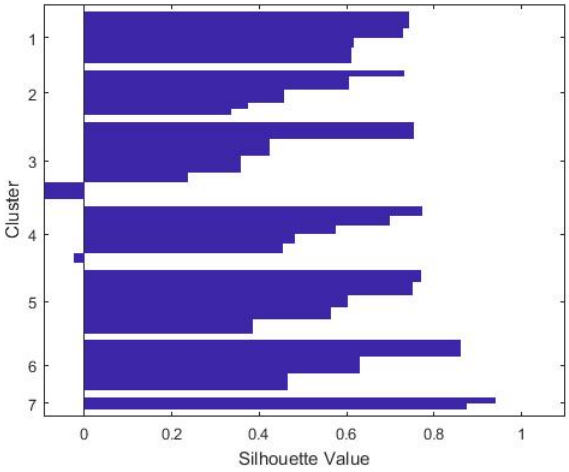
(c) Silhouette plot for $k = 4$



(d) Silhouette plot for $k = 5$



(e) Silhouette plot for $k = 6$



(f) Silhouette plot for k = 7

Figure 4.10: Silhouette plots for $k = 2, 3, 4, 5, 6, 7$ of the camping dataset

Table 4.7: Silhouette mean values of camping dataset

$k =$	2	**3**	4	5	6	7
Mean value	0.7135	**0.6323**	0.5239	0.5541	0.5242	0.5471

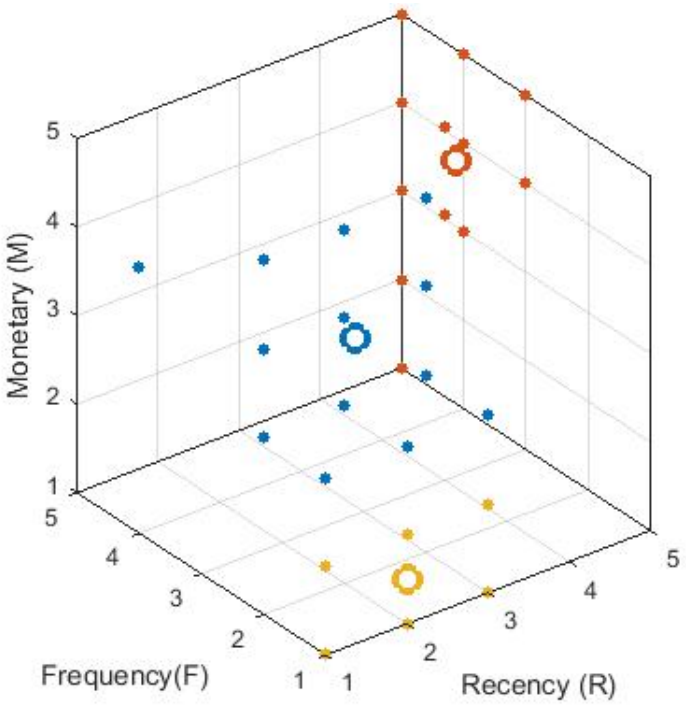


Figure 4.11: Scatter plot representing the three clusters of the camping dataset

their sizes. Additional information regarding each cluster and the individuals within that cluster could be extracted in the same manner as for the toy problem. The following variables were used: gender, age, annual income, marital and children statuses, occupation, campsite reservations and type of shelter.

The RFM ratios of each cluster are determined and illustrated in Figures 4.13 – 4.15. Figure 4.13 represents the RFM ratios of cluster 1 and can be interpreted as follows. The majority of participants (79 percent) have a high R-value (R=5). This occurs when the individual's latest camping trip took place very recently ($<$ 2 months from end of survey period). Most of the participants (42 percent) have a medium frequency value (F=3), meaning that they camp more than five times a year, but less than or equal to seven times. The M-values for this cluster are also in the higher categories, with 37 percent of

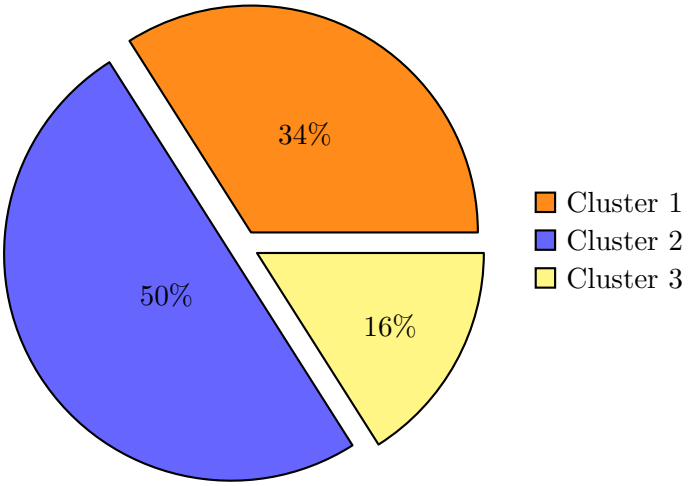


Figure 4.12: Pie chart indicating the three cluster sizes of the camping dataset

participants having an M-value equal to five, another 37 percent of participants having an M-value equal to four, and 16 percent of participants having an M-value equal to three. Only 5 percent of the participants have an M-value equal to one or two. Table 4.8 provides insights in order to understand the high R, F and M-values.

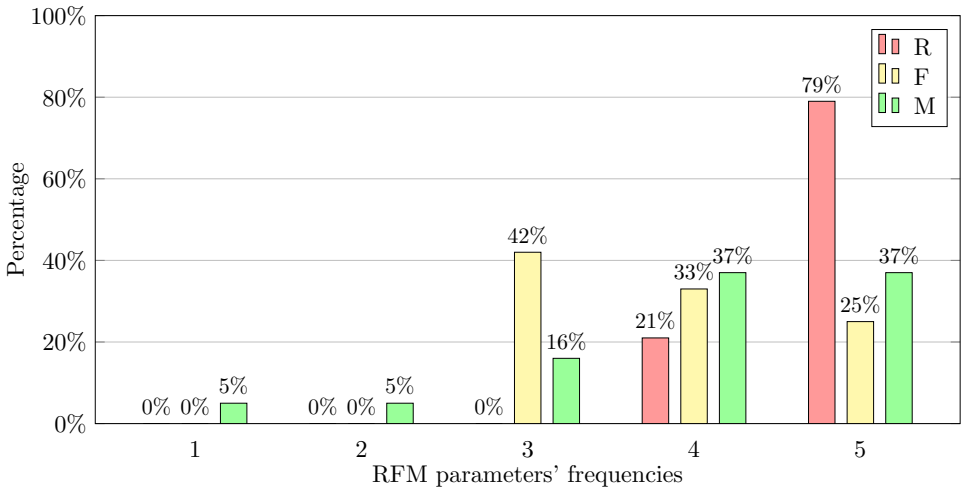


Figure 4.13: Cluster 1 of camping problem – RFM ratio

Figure 4.14 illustrates the RFM ratio of cluster 2 (largest cluster). It can be interpreted that most of the individuals (84 percent) have an F value equal to two, while 16 percent of the individuals' F-value is equal to one. This means that 84 percent (F=2) of

Table 4.8: Cluster 1 of camping problem – Customer demographics

Gender	Male	43%	**Annual Income**	Medium	63%	**Occupation**	Student	9%	**Age**	16-32	7%
	Female	57%		High	37%		Employed	18%		32-48	11%
							Retired	73%		48-62	79%
										62-80	3%
Marital Status	Yes	73%	**Children Status**	Yes	55%	**Reservations**	4-7 Days	9%	**Type of shelter**	Tent	45%
	No	27%		No	45%		2-4 Weeks	11%		RV	55%
							Month +	38%			
							3 Months +	42%			

the individuals participate in camping more than twice a year, but less than or equal to five times a year; and 16 percent (F=1) of the campers participate in camping twice or less a year. The recency value for cluster 2 varies from one (33 percent of the campers) to three (34 percent of the campers). When R=1, the individuals' latest camping trip occurred the previous year ($\geq$ 12 months from end of survey period), when R=2 the individuals camped within the survey year, but more than or equal to six months ago; and when R=3, the individuals camped about six months ago ($4 \leq$ months < 6 from the end of the survey period). It can be concluded, when investigating the R and F-values, that cluster 2 contains infrequent campers. Looking at the M-values, the majority (44 percent) have a medium M-value (M=3), while 28 percent of the campers subsequently fall in the M=4 and M=5 category. Table 4.9 is presented to provide insight into cluster 2.

Table 4.9: Cluster 2 of camping problem – Customer demographics

Gender	Male	55%	**Annual Income**	Medium	59%	**Occupation**	Home-maker	6%	**Age**	16-32	2%
	Female	45%		High	41%		Employed	86%		32-48	68%
							Retired	8%		48-62	6%
										62-80	24%
Marital Status	Yes	89%	**Children Status**	Yes	93%	**Reservations**	4-7 Days	20%	**Type of shelter**	Tent	74%
	No	11%		No	7%		2-4 Weeks	47%		RV	24%
							Month +	25%			
							6 Months +	8%			

In contrast to cluster 2, cluster 3 is the smallest. Cluster 3 displays similar behaviour regarding the R and F-values, however, there is a major difference regarding the M-

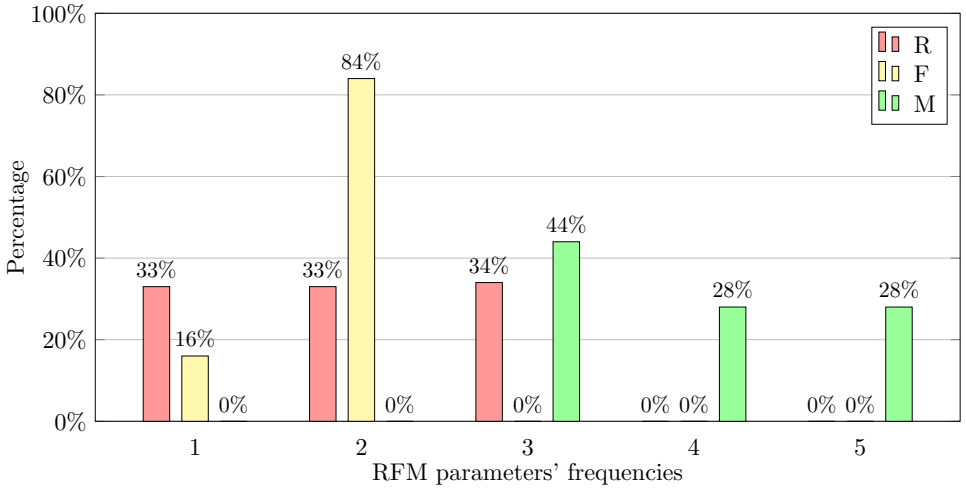


Figure 4.14: Cluster 2 of camping problem – RFM ratio

values as seen in Figure 4.15. Cluster 3 has the same R-values as cluster 2, however 100 percent of the participants have a low F-value (F=1). Looking at the M-values, 50% of the participants belong to category one and the other 50 percent to category two. Thus, it can be concluded that cluster 3 individuals have low R, F and M-values. Table 4.10 is constructed to provide an understanding of the low R, F and M-values.

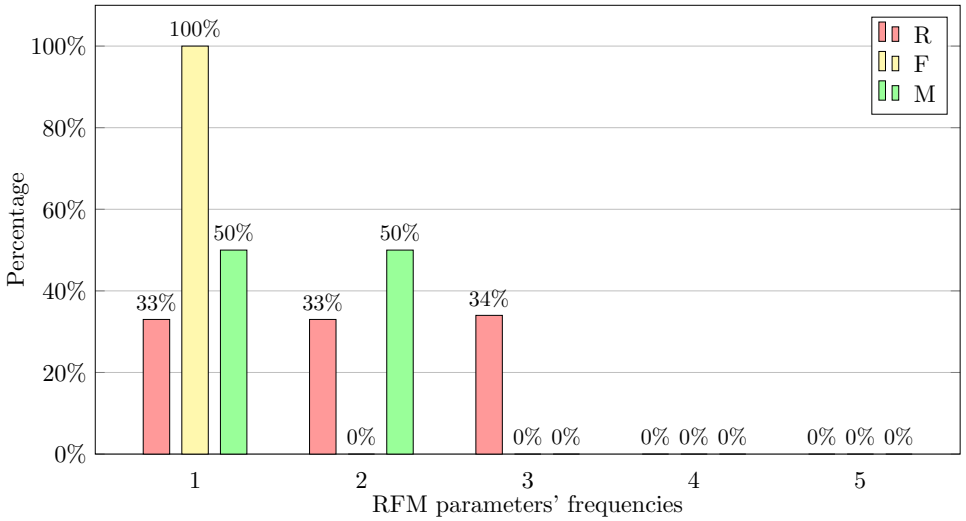


Figure 4.15: Cluster 3 of camping problem – RFM ratio

If further analysis is necessary for these clusters, the following variables could be

Table 4.10: Cluster 3 of camping problem – Customer demographics

Gender	Male	58%	**Annual Income**	Low	67%	**Occupation**	Student	6%	**Age**	16-32	36%
	Female	42%		Medium	33%		Home-maker	3%		32-48	42%
							Employed	89%		48-62	22%
							Retired	2%			
Marital Status	Yes	33%	**Children Status**	Yes	42%	**Reservations**	No advance	34%	**Type of shelter**	Tent	81%
							1-3 Days	16%		RV	13%
	No	67%		No	58%		4-7 Days	27%		Bivy/No shelter	6%
							2-4 Weeks	23%			

considered: region, school holiday, seasonality, ethnicity and medical. The correlation between these variables and camping might provide even deeper insight for the outdoor domain.

4.3 Summary: Chapter 4

After completing the literature review (Chapters 2 and 3), the next phase included the design and development of the CSP tool, containing a suite of Big Data Analytics tools and techniques which will allow for customer super-profiling. The solution architecture of this tool was developed by following the OPM. This methodology was selected because it represents the system simultaneously in a graphic representation and in a natural language. The three main components which constitute the simulator and demonstration tool proposed in this chapter are the simulating, segmentation and profiling processes. The customer super-profiling tool also has the ability to classify new customers into a customer group that already exists, as well as being able to build a predictive model from the customer data groups.

Following the development of the solution architecture, two problems were created to demonstrate the outcomes of the analysis. The first problem was a small dataset problem, referred to as a toy problem. This toy problem focused on customers with many attributes. According to the solution architecture, segmentation is implemented first. The toy problem dataset was not for a specific grocery store, therefore market segmentation was first implemented to divide the customers into different geographic regions. A grocery store in region X (one of the market segments) that wanted to obtain more demographic and behavioural information regarding their customers. To

obtain this information, Region X's customers are 'filtered' to include only the specific grocery store's customers. Customer segmentation was then performed on that group, by following the RFM model approach, and then clustering the customers by applying the k-means clustering algorithm. After performing the customer segmentation, four clusters were identified, the customers' behaviour was known, and each cluster's customers were profiled. These profiles could provide more information for the grocery store regarding its customers.

A second problem was studied, following the same methodology, but in a different domain. This problem focused on the participation behaviour of campers, as opposed to the purchasing behaviour of customers (toy problem). The RFM model approach was followed, as well as the k-means clustering algorithm, to cluster the campers. Three clusters were identified, and each cluster was profiled.

Next, Chapter 5 will present the development and implementation of the CSP tool. Specific Big Data Analytics techniques will be selected and utilised in conjunction, to perform the necessary processes in order to reach the goal of *super-profiling*.

Chapter 5

Customer Super-Profiling tool

The previous chapter presented the proposed solutions architecture. The architecture offers assistance in order to reach the goal of this research, which is to develop a Customer Super-Profiling (CSP) tool that contains a suite of Big Data Analytics tools and techniques which will allow for customer super-profiling. The need for a simulator that creates datasets with *specific properties* was identified. The elements needed for data creation and analysis are shown in Appendix A, as well as the customer datasets that were created with simulation. These datasets contain customer *information* (demographic and extra value adding features) and *typical transactional information*. The data simulator containing the customer data was validated according to various data distributions.

Chapter 6 will present the application and demonstration of the *Customer Super-Profiling* (CSP) *tool*. This CSP tool will have the ability to analyse a large dataset (Objective 1) by utilising various Big Data Analytics tools and techniques (Objective 2), in *conjunction*. Several business case scenarios will be illustrated to communicate the business value of this tool. The researcher will also revisit and apply the CSP tool on the "*large dataset problem*" discussed in a previous chapter, as well as validate the CSP tool. A summary at the end of the chapter includes the researcher's views and interpretation of the results obtained by the CSP tool.

5.1 Customer Super-Profiling tool road map

The need for a CSP tool was identified and documented as the overall objective for this research, as seen in Section 1.4. To conduct this research, big datasets were necessary. In order to determine the structure and content of the data, a data simulator was developed to provide the super-profiling tool with customer data. Eventually, a different user of this super-profiling analytics tool could provide their own data, as long as the data have the same format and structure.

The CSP tool will make use of various data analytics techniques in order to analyse customer data. The CSP tool requires a user proficient in data science. Previous chapters (see Chapter 2 and Chapter 3) provided an in-depth literature study which acquainted the researcher with the domain of data analytics. This knowledge led to selecting specific data analytics tools and techniques, to be used in conjunction, when developing the CSP tool.

In order to communicate the purpose of the CSP tool and keep consistency when developing and utilising the tool, an outline containing various steps is developed. Figure 5.1 illustrates this broad outline of the CSP tool which will be followed throughout this chapter. The main steps within this outline include:

1. *Select* data.
2. *Do* RFM analysis.
3. *Do* clustering.
4. *Develop* a predictive model.

Figure 5.1 indicates that specific data analytics techniques will be utilised, such as RFM analysis, clustering (k-means) and decision trees, which form part of the predictive model. This selection was made based on the combination of the (customer) data and the end goal, which is to generate *customer super-profiles*. The researcher refers to '*customers*' (see Figure 5.1) generically to refer to any person participating in purchasing activities (any domain). The subsections to follow will visit these steps individually, explain what each of them means and what needs to be done, as well as document the results retrieved during the steps.

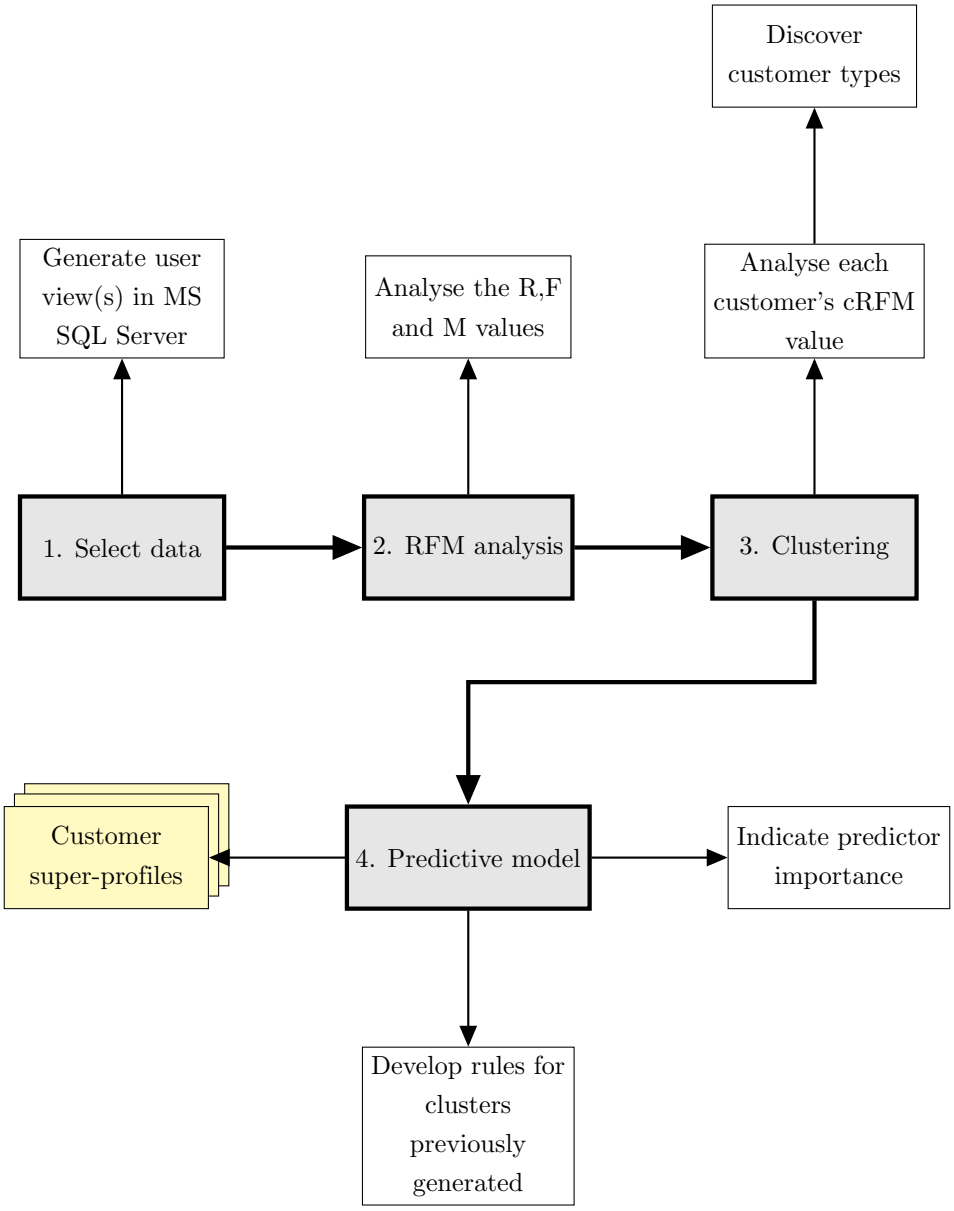


Figure 5.1: Schematic representing the CSP tool

5.1.1 Select data: Simulated South African demographic customer dataset

The first step (1) in Figure 5.1 indicates that data need to be selected. Large quantities of information, if used correctly, can help generate important patterns and trends. These patterns provide useful insights into customer purchasing behaviour, and when

used in combination with customer demographic information, even more insights can be generated, and powerful customer profiles created.

The simulated customer data is stored in Microsoft®[1] SQL Server ®[2], which functions as the database, and can be accessed in Matlab by using a '*selectquery*' command. Relevant datasets are created along various dimensions and imported into Matlab for analysis. The first dataset created contains the *customer transactional history*, which include the customers' primary keys (unique identification numbers), retail shop name, transaction date and amount spent. Data must first be 'cleaned', if necessary. The entire dataset used for this research project is simulated, therefore the likelihood of missing and incorrect values to occur is very slim. However, when using practical or industry data, missing values, erroneous values and/or outliers should be '*cleaned*'. Next, the RFM analysis will be performed.

5.1.2 RFM analysis: Simulated South African demographic customer dataset

After selecting the appropriate customer information, the second step includes the *Recency, Frequency and Monetary* (RFM) analysis. The RFM model is one of the best-known customer value analysis methods, which extracts characteristics of customers using fewer criteria as clustering attributes to reduce the complexity of the model, and provides a simple framework for quantifying customer behaviour. Note that the R, F and M parameters are dimensionless. A literature review regarding the RFM model was discussed in Chapter 2. This section will present the application of the RFM model, with regards to the simulated customer data. The RFM analysis that will be performed on the dataset created in the previous step (1); this dataset includes all *retail shops* each customer visited, thus will not focus on only one shop. The values that will be obtained from this analysis will provide insights into the individual R, F and M values of each customer taking into consideration all of the retail shop visits of each customer, and will be used to fulfil the next step (3), namely *clustering*.

The transaction data, which forms part of the behavioural feature called *activities* (see Table A.2) were simulated to vary from 01/01/2015 to 31/12/2016 (two years). The RFM method was implemented as follows:

[1]The registered trademark for Microsoft® will from now on be omitted.
[2]The registered trademark for SQL Server Management Studio® will from now on be omitted.

- Recency (R): It represents the interval between the customer's latest active date and the date selected as the last date (31/12/2016). The older the active date, the lower the recency category of that customer. This recency date does not consider the type of retail shop; each customer's latest active date could be for a different retail shop.

- Frequency (F): It represents the number of times a customer was active during the specified period for this study. The higher the number of transactions in an interval, the higher the frequency category. Again, the frequency of each customer is calculated with regards to all the retail shops they visited, therefore a customer with more retail shop excursions will most likely have a higher frequency than one with less.

- Monetary (M): It represents the monetary value of the purchases in the specified period for this study. The higher the amount spent by a customer, the higher the monetary category. The average amount spent by each customer is calculated by adding all the money spent by a customer and dividing that amount by their frequency value. This average amount spent is used to allocate a monetary value to each customer.

The minimum and maximum values for the R, F and M parameters of the dataset under study can be seen in Table 5.1. These results are utilised in order to provide each

Table 5.1: Minimum and maximum R, F and M parameter values

	Recency	Frequency	Monetary
Minimum	01/01/2015	3	R99.79
Maximum	31/12/2016	2 502	R826.50

customer with an individual R, F and M category value. In order to do so, the first step is to calculate the interval value of each parameter as follows:

$$\text{Interval} = \frac{\text{Maximum} - \text{Minimum}}{5}. \tag{5.1}$$

When calculating the interval value, the difference between the maximum and minimum value is divided by 5, because each RFM parameter has five categories, with 1 being the

lowest category a customer can be assigned and 5 the highest category a customer can be assigned.

Next, the R, F and M category range values can be determined. Each parameter has five categories. These range values, which limit each category, are calculated by taking the minimum value of each parameter and adding the interval value. This will provide the R, F and M of category value one, then the interval value is added to the R, F and M of category one to give the R, F and M values of category two *etc.* Figure 5.2 schematically represents the R, F and M category range values. For example, a customer will be assigned a recency category of 1 if his last visit to the retail shop(s) was between 01/01/2015 and 27/05/2015, a frequency category value of 3 indicates he visited the shop(s) more than 1 002.6 and less then 1 502.4 times within the survey period. Also, a monetary category value of 2 is assigned if he spent between R245.13 and R390.48 on all their transactions in the given period.

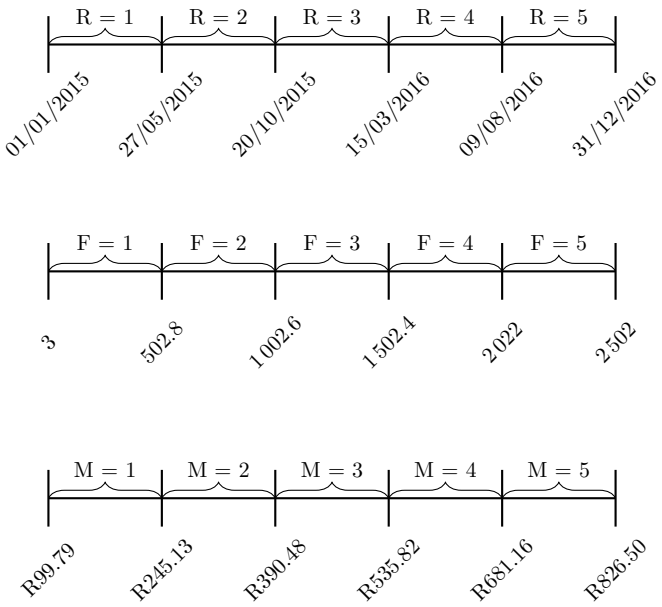


Figure 5.2: Schematic representing the R, F and M category range values

Algorithm 1 is executed in Matlab to assign an R category value to each customer. The input table defined as `DetermineRecency`, as shown in Algorithm 1, contains each customer's number alongside their most recent transaction date. Each case represents a different outcome, when a customer satisfies one of the cases, the R category value

(*i.e.* 1, 2,. . . , 5) associated with that case is assigned to the customer. Similarly, each customer is assigned an F and M category value, following Algorithm 2 (with input table defined as `DetermineFrequency`) and Algorithm 3 (with input table defined as `DetermineMonetary`), respectively. Algorithms 1 – 3 were executed in Matlab to assign an R, F and M category value to each customer, and stored in table `RFM`. The top 10 rows of table `RFM` can be seen in Table 5.2, after the R, F and M category values were assigned to each customer.

Algorithm 1 Determine the R category value

Begin
Input `DetermineRecency` table
For each row in table `DetermineRecency`
  Case 1: `DetermineRecency.Row` $<$ R2
    R category value $\leftarrow$ 1
  Case 2: `DetermineRecency.Row` $<$ R3
    R category value $\leftarrow$ 2
  Case 3: `DetermineRecency.Row` $<$ R4
    R category value $\leftarrow$ 3
  Case 4: `DetermineRecency.Row` $<$ R5
    R category value $\leftarrow$ 4
  Case 5: `DetermineRecency.Row` $<$ RMax
    R category value $\leftarrow$ 5
Next row
End

The second column in Table 5.2 indicates the *Customer_ID*s of those customers that visited one or more retail shop(s). Columns 3 to 5 indicate the R, F and M category values of each customer, respectively. Table 5.3 summarises the intermediate results by indicating the number of times, together with the percentage, each category value (*e.g.* R = 1, R = 2, *etc.*) occurred within the dataset.

Table 5.3 indicates that more than 80 percent of the customers have a high recency value, meaning that most of the customers' latest purchase date was between 09/08/2016 and 31/12/2016. The customers' frequency values vary from low to high, with most customers having a frequency value equal to 3. This occurrence means that the number of times that most customers were active lies between 1 002.6 and 1 502.4 (Figure 5.2).

Algorithm 2 Determine the F category value

Begin
Input `DetermineFrequency` table
For each row in table `DetermineFrequency`
  Case 1: `DetermineFrequency.Row` < F2
    F category value ← 1
  Case 2: `DetermineFrequency.Row` < F3
    F category value ← 2
  Case 3: `DetermineFrequency.Row` < F4
    F category value ← 3
  Case 4: `DetermineFrequency.Row` < F5
    F category value ← 4
  Case 5: `DetermineFrequency.Row` < FMax
    F category value ← 5
Next row
End

Algorithm 3 Determine the M category value

Begin
Input `DetermineMonetary` table
For each row in table `DetermineMonetary`
  Case 1: `DetermineMonetary.Row` < M2
    M category value ← 1
  Case 2: `DetermineMonetary.Row` < M3
    M category value ← 2
  Case 3: `DetermineMonetary.Row` < M4
    M category value ← 3
  Case 4: `DetermineMonetary.Row` < M5
    M category value ← 4
  Case 5: `DetermineMonetary.Row` < MMax
    M category value ← 5
Next row
End

Lastly, more than 80 percent of the customers have a low monetary value. These customers spent, on average, between R99.79 and R245.13. The data is now ready for the

Table 5.2: Top 10 rows of table `RFM`

	Customer_ID	R	F	M
1	2	5	3	1
2	3	5	4	1
3	4	5	3	1
4	5	5	3	1
5	7	5	4	1
6	9	5	2	1
7	11	5	4	1
8	13	5	4	1
9	15	5	3	1
10	16	5	3	1

next step (3), which is clustering.

Table 5.3: Summary of R, F and M category value occurrences in `RFM`

	R		F		M	
1	0	0%	506	1.51%	29 246	87.28%
2	2	0.01%	11 333	33.82%	1 596	4.76%
3	6	0.02%	15 337	45.77%	444	1.32%
4	4 055	12.10%	6 150	18.35%	1 956	5.84%
5	29 447	87.88%	184	0.55%	259	0.77%
Total customers	**33 510**					

5.1.3 Clustering: Simulated South African demographic customer dataset

The goal of the outline, schematically represented by Figure 5.1, is to develop a tool that has the ability to discover customer profiles. Thus, in order to achieve better understanding of the customers' behaviour and to improve the quality future predictions, this step (3) will perform segmentation (also known as clustering) on table `RFM` created in the previous step (2), in order to account for similarities or differences in broad segment needs.

The unsupervised learning technique that will be used in this step is the k-means clustering technique. This technique consist of grouping individuals based on their response variables (R, F and M values from step 2). Individuals are grouped according to how close their responses are to those of other customers, using the *Euclidean* distance metric as the selection criteria (Salazar et al., 2007). One of the challenges of using the k-means clustering algorithm is deciding on an appropriate or best number of clusters.

A Matlab function is utilised in order to determine the appropriate number of clusters. One of the input arguments of this function is to specify the '*clustering evaluation criterion*', which is selected as '*silhouette*' for this dataset. This criterion creates a cluster evaluation object containing silhouette values. The number of clusters associated with the highest silhouette value is seen as the best number of clusters. However, as indicated by literature, deciding on the most appropriate cluster solution is subjective, therefore two criteria were used.

The first involved an inspection of the silhouette values. The silhouette values provide a reasonable indication of the cluster structure; the higher the silhouette value, the better the cluster structure. It can be said that an average silhouette value greater than 0.5 indicates a reasonable grouping of the data, while a value less than 0.2 would indicate that the data does not exhibit a cluster structure (Martinez et al., 2010). Figure 5.3 indicates the silhouette criterion value for each number of clusters tested. Figure 5.3 and Table 5.4 indicate that the most appropriate solutions seemed to vary between two clusters and five to 10 clusters, according to the silhouette values. The best number of clusters suggested by the function is 10, with a silhouette value of 0.95.

Secondly, the sizes of the clusters were taken into account. Aspects that are related to cluster sizes were also considered, such as:

- Marketing cost; the more clusters, the more marketing efforts need to be funded.
- Smaller clusters lead to more analysis possibilities.

Under this criterion, the *two cluster solution* resulted in groups of reasonable sizes (silhouette value of 0.84). Taking both the cluster size and the silhouette values into account it was decided to proceed with a *two cluster solution*. In total 50 000 customers were simulated (primarily indicating their demographic information), and only 33 510 randomly selected customers participated in the RFM analysis. The remaining 16 490

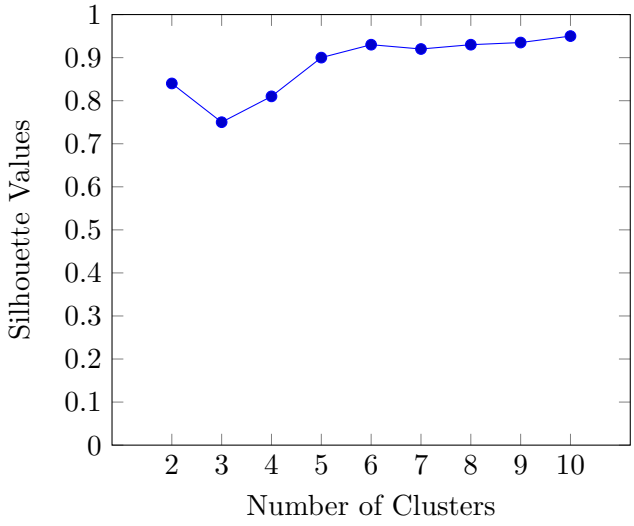


Figure 5.3: Plot of the silhouette criterion values for each number of clusters tested for the customer dataset

Table 5.4: Cluster solutions for customer dataset

	2 clusters	**5 clusters**	**6 clusters**	**7 clusters**	**8 clusters**	**9 clusters**	**10 clusters**
Size of cluster:	1 ⇒ 21 671	1 ⇒ 10 917	1 ⇒ 6 922	1 ⇒ 20 281	1 ⇒ 6 640	1 ⇒ 3 995	1 ⇒ 13 637
	2 ⇒ 11 839	2 ⇒ 13 641	2 ⇒ 13 641	2 ⇒ 6 273	2 ⇒ 6 276	2 ⇒ 13 641	2 ⇒ 10 429
		3 ⇒ 6 276	3 ⇒ 3 995	3 ⇒ 2 356	3 ⇒ 3 857	3 ⇒ 6	3 ⇒ 6 276
		4 ⇒ 2 668	4 ⇒ 6 276	4 ⇒ 4 285	4 ⇒ 13 581	4 ⇒ 3	4 ⇒ 488
		5 ⇒ 8	5 ⇒ 8	5 ⇒ 53	5 ⇒ 488	5 ⇒ 2 404	5 ⇒ 2 510
			6 ⇒ 2 668	6 ⇒ 254	6 ⇒ 2 356	6 ⇒ 254	6 ⇒ 8
				7 ⇒ 8	7 ⇒ 58	7 ⇒ 6 922	7 ⇒ 153
					8 ⇒ 254	8 ⇒ 10	8 ⇒ 3
						9 ⇒ 6 275	9 ⇒ 3
							10 ⇒ 3
Silhouette value	0.8402	0.9004	0.9290	0.9210	0.9306	0.9345	0.9535

customers did not visit the simulated retail shop(s) and will therefore by default form their own cluster.

Next, the silhouette plot for the *two clusters* was constructed and shown in Figure 5.4. The silhouette plot provides silhouette values for each customer. If a customer has a high silhouette value, it indicates that the customer is well-matched to its own cluster, and poorly-matched to the neighbouring cluster. If most customers have a high silhouette value, then the clustering solution is appropriate; however, if many customers

have a low or negative silhouette value, then the clustering solution may have either too many or too few clusters. Figure 5.4 indicates that a two cluster solution is appropriate because:

- All of the customers have positive silhouette values.
- The majority of customers also have silhouette values higher than 0.5.

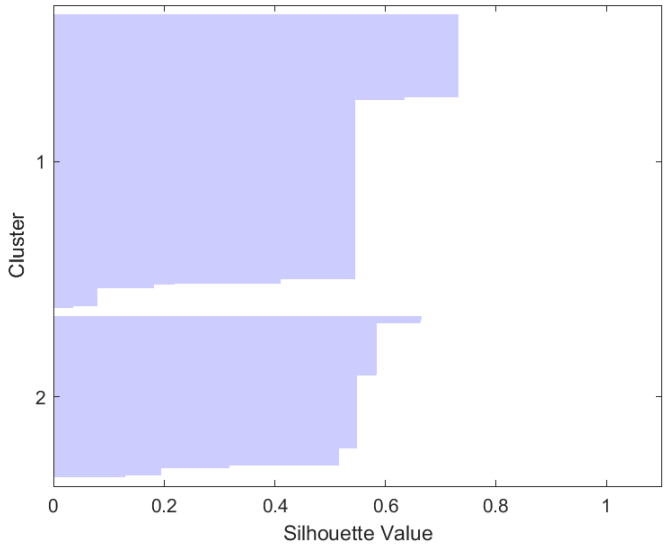


Figure 5.4: Plot of the silhouette values from clustered data of customer dataset

Figure 5.5 presents a scatter plot of the two clusters that were formed when applying the k-means clustering method. The red dots represent cluster 1 and the blue dots represent cluster 2. This figure indicates that if only the recency and frequency values (coordinates) were considered, the centroids would lie very close to each other (almost overlap). Therefore, the third dimension, the monetary values, contributes to creating a suitable cluster structure for this dataset, dividing the two centroids (groups) along the vertical axis.

Table 5.5 provides a summary of the clusters, indicating the percentage of customers in both clusters belonging to each RFM category. The results shown in Table 5.5 are consistent with Figure 5.5. Cluster 1 only has R values in categories 4 and 5 (indicated by the red dots), whereas cluster 2 has R values in categories 2, 3, 4 and 5 (indicated by the blue dots). The F values of cluster 1 are in category 3 and higher, while the F values

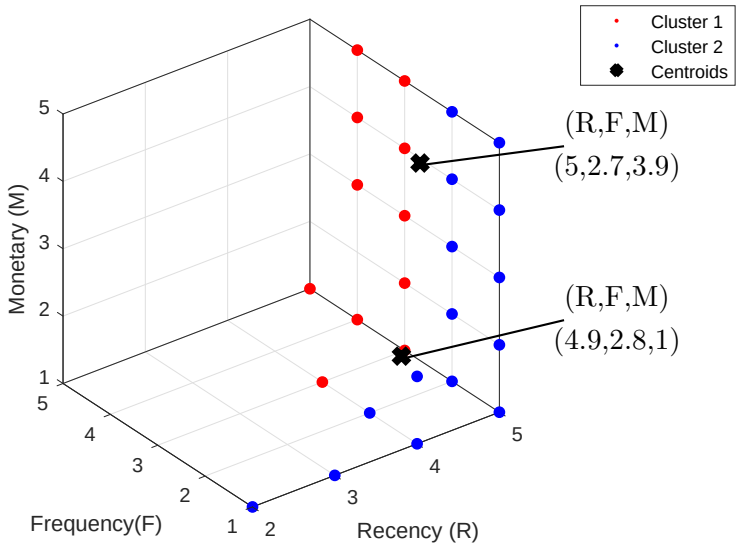


Figure 5.5: Scatter plot representing the two clusters for the customer dataset

of cluster 2 are in category 1 and 2, as indicated by the red and blue dots, respectively. The M values for both cluster 1 and 2 vary, with most of both clusters having low M values.

Table 5.5: Summary of the customers in each RFM category per cluster

	Cluster 1			**Cluster 2**			**Cluster 3**
	R	**F**	**M**	**R**	**F**	**M**	
1	0%	0%	90.46%	0%	4.27%	81.44%	Not applicable
2	0%	0%	1.44%	0.02%	95.73%	10.84%	
3	0%	70.77%	1.18%	0.05%	0%	1.59%	
4	0.28%	28.38%	6.40%	33.74%	0%	4.87%	
5	99.72%	0.85%	0.51%	66.19%	0%	1.26%	
Total customers	**21 671**			**11 839**			**16 490**

As indicated by Table 5.5, cluster 1 has the majority of observations, followed by cluster 3; which is the non-shoppers, followed by cluster 2. The biggest difference between cluster 1 and cluster 2 lies in the frequency parameter. The customers that belong to

cluster 1 have higher frequency values than the customers that belong to cluster 2. This means that cluster 1 customers are more frequent customers and could be seen as more loyal customers, as opposed to cluster 2 customers. ("Loyal" in this context is towards the retail stores involved in this research.)

RFM analysis is utilised in many ways by researchers; therefore, the RFM analysis can mean different things to different researchers. The classic RFM implementation is to assign an RFM category to each customer, as performed in step 2. This results in customer segments that are 'neatly' ordered from most valuable (category 5) to least valuable (category 1) customers. After each customer is assigned an R, F and M category, literature indicates that it is also feasible to create a concatenated RFM *score* for each customer (Hosseini et al., 2010; McCarty and Hastak, 2007). The RFM score is assigned to customers based on their past behaviour. Using the quintile method (mentioned in Chapter 2), at most 125 different scores (5×5×5) can be assigned. A customer's score can range from 555 being the highest, to 111 being the lowest. The best customers are in quintile 5 for each factor (555) that have purchased most recently, most frequently and have spent the most money (Birant, 2011).

The researcher adjusted this score value and called it the *combined* RFM (cRFM) value. This value is determined by adding each customer's R, F and M category value and then dividing the total by 3, as follows:

$$\text{cRFM/customer} = \frac{\text{R} + \text{F} + \text{M}}{3}. \tag{5.2}$$

Each customer has their own cRFM category value, and knowing this value provides a different perspective of the customers. The customers can easily be compared with each other when assigning a combined (cRFM) value to them. With a big customer dataset it is necessary to be able to compare customers and draw conclusions based on the comparisons. It is, however, still possible to interpret each customer's RFM category values separately, if needed. Many decision-making domains use some form of scoring with a single value to distinguish between alternatives/candidates. Two examples can be mentioned:

1. A *credit score*, which is a numerical expression based on a level analysis of a person's credit files, to represent the creditworthiness of an individual. A credit score can range between 330 and 850. The higher the score, the less risk the person under investigation is to creditors (NDA National Debit Advisors, 2016).

2. The Engineering faculty of Stellenbosch University utilises a scoring technique when calculating the *selection score* of applicants (Stellenbosch University, 2018). This score is calculated, based on the matric marks, as follows: *Selection score* = *Mathematics mark* + *Physical Sciences mark* + 6 × *Matric average*; with a maximum score of 800.

Thus, the researcher believes that the cRFM value will provide deeper insights into the customers' value and purchasing behaviour.

Again, the interval value of the cRFM values are calculated in order to have five category values for these customers. The same interval and category range values will be used to assign each cluster's customers a cRFM category value. Algorithm 4 is executed to provide each customer within a cluster with a cRFM value as well as assigning a cRFM *category value* to each customer.

Algorithm 4 Determine the cRFM category value for each customer

```
Input table RFM
For each row in table RFM
    cRFM ← (R + F + M) / 3
Next row
Determine the minimum and maximum cRFM values
Calculate the cRFM interval values
For each row in table cRFM
    Case 1: cRFM.Row < Interval 2
        cRFM category value ← 1
    Case 2: cRFM.Row < Interval 3
        cRFM category value ← 2
    Case 3: cRFM.Row < Interval 4
        cRFM category value ← 3
    Case 4: cRFM.Row < Interval 5
        cRFM category value ← 4
    Case 5: cRFM.Row < cRFMMax
        cRFM category value ← 5
Next row
End
```

Table 5.6 represents a summary of the percentage customers in each cRFM category,

for cluster 1 and 2. As anticipated from the RFM category values per cluster (as shown in Table 5.5), cluster 1 and cluster 2 have different *types* of customers, loyal and less loyal customers. Considering Table 5.6, the majority of customers in cluster 1 have high cRFM values, whereas the majority of customers in cluster 2 have lower, more distributed cRFM values. These two clusters sufficiently separate the two types of customers present in the dataset. Cluster 1, which contains the more *'loyal'* customers is the bigger cluster, whereas cluster 2 which contains the less *'loyal'* customers is smaller.

Table 5.6: Percentage customers in each cRFM category for clusters 1 and 2

	Percentage:	
	Cluster 1 (21 671)	**Cluster 2 (11 839)**
cRFM = 1	0%	0.068%
cRFM = 2	0%	35.73%
cRFM = 3	61.50%	56.51%
cRFM = 4	31.56%	6.48%
cRFM = 5	6.94%	1.22%

Now, the researcher will stray from the classical application of the RFM analysis to achieve marketing intelligence. With the help of the two clusters and the customers' cRFM values, it is possible to discover various *types of customers* present in the dataset. Identifying the customer types assists in targeting only a specific/type of audience, and not all the customers within a database. There are five types identified within this customer dataset, each associated with different cRFM category values. When utilising the CSP tool, this step (identifying the type of customer) is done by using human discretion, each dataset and their types will differ, because when the data are different (different customers, retail shops, purchasing history *etc.*) so are the types. The decision-maker should make intuitive judgements about the various customer types present within the dataset as well as which customers are associated to which type. By grouping (clustering) and 'ranking (scoring)' (cRFM category values) customers, the decision-maker can differentiate between types of customers and target them based on predefined and justified values instead of blindly reaching out to every customer. The researcher considered both the cRFM category values as well as the individual RFM category values to determine the association between the various types of customers and cRFM category values. It is possible to consider the individual RFM category values

when working with the cRFM category values, for they consist of the RFM category values. Shih and Liu (2003) conducted their study by making use of a '*RFM pattern*' to indicate how the RFM category values of each customer segment differ from the 'original' RFM category values in a dataset. This assists in explaining and allocating customers (cRFM category values) to each customer type.

The researcher decided to adapt this RFM pattern technique by applying it to the cRFM category values, illustrating how each cRFM category (consisting of R, F and M parameters) differs from the original RFM category values in each cluster. The *original* RFM category values refer to the values obtained in steps 2 and 3 (as seen in Table 5.5). This technique is initiated as follows. Table 5.7 indicates the *total average* R, F and M category values, which are determined by calculating the average category value for each RFM parameter in Table 5.5, for both clusters. The total average RFM category values

Table 5.7: Total average RFM category values per cluster

	Cluster 1	**Cluster 2**
R	4.9972	4.6610
F	3.3008	1.9573
M	1.2505	1.3367

in both clusters are compared with the average RFM category values which constitute the cRFM category values. If the average R (F, M) category value present within each cRFM category (*e.g.* the RFM category values which are combined to form the cRFM categories equal to 1, 2, ..., 5) exceeds the total average R (F, M), then an upward arrow (↑) is shown; otherwise, a downward arrow (↓) is shown. This is referred to as the RFM pattern (Shih and Liu, 2003). For example, cluster 1 customers belonging to cRFM category 5, have average recency and monetary category values (5 and 4.0699) exceeding the total average recency (as indicated in Table 5.7); therefore an upward arrow (↑) can be assigned to the recency and monetary categories, while a downward arrow (↓) is assigned to the frequency category (3.0386) of cRFM equal to 5 (cluster 1). Table 5.8, which indicates the various customer types and the appropriate cRFM category values of each type, is constructed by considering the following aspects:

1. Various customer types.

2. The customers' cRFM category values.

3. The RFM patterns.

4. The researcher's judgement.

The five customer types that were decided on are listed in Table 5.8, together with various customer characteristics' explanations, adapted from research conducted by Evaldas (2017), the customers' identification traits (cRFM category values) and the RFM patterns of each cRFM category. The table shows that deeper understanding of the customers can be obtained in the given dataset, for example a (new) low spender is characterised as a customer who made a significantly low purchase on their (first) buying experience(s) and obtains downward arrows for all RFM parameters within both clusters, as opposed to a big spender who made significantly high purchases on buying experiences and therefore has an upward arrow at the monetary parameter in both clusters.

Table 5.8: Types of customers in dataset

Type of customer	**Customer characteristics**	**Cluster 1**	**Clusters 2**
(New) low spenders	These customers have made significant low purchases on their (first) buying experience.	cRFM = 3 R ↓ F ↓ M ↓	cRFM = 2 R ↓ F ↓ M ↓
(New) big spenders	These customers, as opposed to the (new) low spenders, have made significant high purchases on their (first) buying experience. These customers are wealthy and will spend their money over a lifetime of their relationship with (a) retail group(s). They usually content themselves with a few big purchases, or a few small ones.	cRFM = 5 R ↑ F ↓ M ↑	cRFM = 4 R ↑ F ↑ M ↑ cRFM = 5 R ↑ F ↑ M ↑
Low loyal customers	These customers buy often but are not able to spend more than they can afford or more than they think something should cost. These customers make purchases carefully but trust the retail group(s) they support.	cRFM = 4 R ↑ F ↑ M ↓	cRFM = 3 R ↑ F ↑ M ↓
Churned cheap customers	These customers spend as little as possible, buy very few goods and their purchase history is from a long time ago. It is extremely unlikely that these customers are a source of repeat purchases. Marketers believe that these customers are not worth time and trouble.		cRFM = 1 R ↓ F ↓ M ↓
Prospects	No transactions are registered in the database; only customer information is available.	Only cluster 3	

Differentiating between the customer types and the associated cRFM category values (RFM patterns), as seen in Table 5.8, is a crucial step to complete when utilising the CSP

tool. The various types need to be known before a predictive model can be developed (step 4). While constructing Table 5.8, the researcher decided to differentiate between the customer types and their cRFM category values as follows:

1. *Churned cheap customers*: These customers' cRFM category value is equal to 1. If the cluster does not have a cRFM category that is equal to 1, the cluster will not have churned cheap customers, as seen in cluster 1.

2. *(New) low spenders*: This customer type represents new customers who have recently visited a retail store, still lower than the average recency of the dataset. To be identified as a *new* low spender their monetary contribution (low spender) and purchase frequency (new customer) are low. Thus combining these values (RFM) will provide a low cRFM category value (usually second lowest, after churned cheap customers), yet not higher than 3. There are no cRFM category values equal to 1 or 2 in cluster 1, therefore (new) low spenders in cluster 1 have a cRFM category value equal to 3.

3. *Low loyal customers*: This type represents customers with a good relationship with the retail shop(s), for they are identified as loyal customers. To be identified as a low loyal customer, they recently visited a retail store, purchase frequently and contribute little monetary value with each transaction. Combining these values will provide a higher cRFM category value than the previous types, yet not equal to cRFM category 5. Low loyal customers in cluster 1 have a different cRFM category value than those in cluster 2. Cluster 1 customers' cRFM category value is equal to 4, whereas the customers that belong to cluster 2, and identified as low loyal spenders, have a cRFM category of 3.

4. *(New) big spenders*: Big spenders are the biggest contributors to the retail store(s)' profitability. To be identified as a new big spender, the customer contributes a lot of monetary value (each transaction), and recently visited a retail store; however, their purchase frequency varies from medium to high. Combining these values will provide a high cRFM category, generally this customer type has the highest cRFM category value within their cluster. The new big spenders in cluster 1 have a cRFM category value of 5, with all the RFM parameters higher than the total average RFM category values, except the frequency parameter (new customers

have lower frequency values). New big spenders that belong to cluster 2 have a cRFM category value of 4 and 5, with all the RFM parameters higher than the total average RFM category values.

This process is to be executed each time an analysis is done to associate the cRFM category values with types. Next, a predictive model that includes the customer demographic and extra value adding features will be developed, according to the various types of customers.

5.1.4 Predictive model: Simulated South African demographic customer dataset

A customer segment, or cluster, is not sufficient to identify and then ultimately predict a customer's behaviour. Many researchers believe that the RFM values of customers are generally associated with customer profiling (Nimbalkar and Shah, 2013). Integrating the RFM analysis with both clustering (step 3) and classification provides useful information for current and new customers and more behavioural knowledge of the customers is attained; as opposed to other independent clustering and classification techniques.

Classification techniques (see Chapter 3) are used to derive rules from the clustered results, obtained from the previous step (3). These classification or decision rules are useful for identifying each and every customer from their purchasing patterns (RFM information) (Apté and Weiss, 1997). There are various techniques for classification, *e.g.* decision trees and neural networks, which were documented, whilst performing the literature review in Chapter 3.

After a thorough literature review regarding classification techniques, the researcher, together with the guidance of the study leader, decided to utilise decision trees in order to build the predictive model. A large quantity of research has been conducted on decision trees and their various application areas. They are known for being a decision support tool and are therefore considered as a technique for predicting customer behaviour and profiling customers. According to Trewartha (2006), using a decision tree in conjunction with other data mining techniques, such as unsupervised learning (k-means) which determines whether concept structures exist within the dataset, would provide a good, if not complete implementation of a data mining process. A decision tree consists of various response (or input) variables in order to make a prediction. These response variables

generate decision rules that need to be followed, while also identifying the important predictors within the dataset. Decision tree and decision rule solutions offer a level of interpretability that is unique to symbolic models. This makes these solutions easily understandable for non-technical end users and makes this technique very appealing in decision support-related data mining activities where insight and explanations are of critical importance. This approach is technically viable because most modern symbolic modelling methodologies succeed in formulating solutions that are also competitive in predictive accuracy, compared to non-intuitive or quantitative techniques, such as neural networks. This is an important reason for making use of decision rule modelling techniques to generate rules directly from data (Apté and Weiss, 1997).

As listed in Table 5.8, there are several types of customers within the dataset. Decision rules were discovered using the customers' features (age, gender, province, mobile phone *etc.* as seen in Table A.2) to identify the profiles of such customers. The cRFM values allocated to each type of customer refine the customer profiles, forming the *super-profiles*. Decision rules are extracted from generated decision trees. This is called an indirect method of creating decision rules. The decision rules may not be mutually exclusive, meaning more than one rule may cover the same instance.

Table 5.9 shows the decision rules that are utilised when predicting a customer's type (*i.e.* (new) low spender, low loyal customer, (new) big spender, churned cheap customer or prospect). This set of rules can provide (1) customer super-profiles for each type of customer and (2) classify new/future customers. The decision rules are developed by determining the most *distinguishing* customer feature within the dataset, for example, this feature would be 'province' for the decision rules shown in Table 5.9; then a rule is *formulated* to 'divide' the dataset into various groups (in this case provinces). The customers are classified as either belonging to 'Eastern Cape or Free State' (Rules 1 to 7) or to 'Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West or Western Cape' (Rules 8 to 18). Those customers that belong to the first group of provinces have a different second customer feature that are used to distinguish them further, namely age, as opposed to the other group of customers (employment status). This process is repeated until the rules contain all the customer features, no distinguishing customer features are present, or it is preferred that the rules only contain certain customer features.

Figure 5.6 revealed that the customer features with the *most* influence on the decision rules shown in Table 5.9 include: province, employment status, age, mobile phone type, household size and housing type.

Table 5.9: Decision rules to identify the type of customer

Rule 1:	**if** *Province* = Eastern Cape or Free State and *Age* = 20-24, 30-34, 35-39, 40-45, 50-54, 65-69 or 80+ and *HouseholdSize* = 3, 5, 8 or 9 **then** Low loyal customer.
Rule 2:	**if** *Province* = Eastern Cape or Free State and *Age* = 20-24, 30-34, 35-39, 40-45, 50-54, 65-69 or 80+ and *HouseholdSize* = 1, 2, 4, 6, 7 or 10+ and *Education* = Less than Gr.12 and with diploma or certificate or Degree or post graduate degree and *MobilePhoneType* = Apple, Nokia, Blackberry, HTC or Siemens **then** Low loyal customer.
Rule 3:	**if** *Province* = Eastern Cape or Free State and *Age* = 20-24, 30-34, 35-39, 40-45, 50-54, 65-69 or 80+ and *HouseholdSize* = 1, 2, 4, 6, 7 or 10+ and *Education* = Less than Gr.12 and with diploma or certificate or Degree or post graduate degree and *MobilePhoneType* = Samsung, Other, Huawei, Sony, LG or Motorola **then** Prospect.
Rule 4:	**if** *Province* = Eastern Cape or Free State and *Age* = 20-24, 30-34, 35-39, 40-45, 50-54, 65-69 or 80+ and *HouseholdSize* = 1, 2, 4, 6, 7 or 10+ and *Education* = Less than Gr.12 and no other qualification, Gr.12, Gr.12 with diploma or certificate or Honours degree or higher and *MobilePhoneType* = Samsung, Other, Huawei, Sony, LG or Motorola **then** Low loyal customer.
Rule 5:	**if** *Province* = Eastern Cape or Free State and *Age* = 15-19, 25-29, 45-49, 55-59, 60-64, 70-74 or 75-79 and *HousingType* = Flat or apartment in flat block, Informal – Shack not backyard, Townhouse, Traditional dwelling – hut or Overcrowding **then** Low loyal customer.
Rule 6:	**if** *Province* = Eastern Cape or Free State and *Age* = 15-19, 25-29, 45-49, 55-59, 60-64, 70-74 or 75-79 and *HousingType* = Cluster house in complex, House or brick structure on yard or strand, House, flat or room in backyard, Informal - Shack in backyard, Other, Room, granny flat or large dwelling or Semi-detached house and *Transportation* = Bus, Taxi, Car or Walk/cycle **then** Low loyal customer.
Rule 7:	**if** *Province* = Eastern Cape or Free State and *Age* = 15-19, 25-29, 45-49, 55-59, 60-64, 70-74 or 75-79 and *HousingType* = Cluster house in complex, House or brick structure on yard or strand, House, flat or room in backyard, Informal - Shack in backyard, Other, Room, granny flat or large dwelling or Semi-detached house and *Transportation* = Train or Other **then** Prospect.

Continued on next page

Rule 8: **if** *Province* = Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West or Western Cape and *EmploymentStatus* = Employed or Unemployed and *HouseholdSize* = 1, 2, 3, 4, 5, 6 or 9 and *HousingType* = Townhouse or Traditional dwelling – hut and *MobilePhoneType* = Samsung, Other, Huawei, Nokia, Blackberry, LG or HTC **then** (New) low spender.

Rule 9: **if** *Province* = Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West or Western Cape and *EmploymentStatus* = Employed or Unemployed and *HouseholdSize* = 1, 2, 3, 4, 5, 6 or 9 and *HousingType* = Cluster house in complex, Flat or apartment in flat block, House or brick structure on yard or strand, House, flat or room in backyard, Informal – Shack in backyard, Informal – Shack not backyard, Other, Room, granny flat or large dwelling, Semi-detached house or Overcrowding **then** (New) low spender.

Rule 10: **if** *Province* = Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West or Western Cape and *EmploymentStatus* = Employed or Unemployed and *HouseholdSize* = 1, 2, 3, 4, 5, 6 or 9 and *HousingType* = Townhouse or Traditional dwelling – hut and *MobilePhoneType* = Apple, Sony, Motorola or Siemens **then** Low loyal customer.

Rule 11: **if** *Province* = Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West or Western Cape and *EmploymentStatus* = Employed or Unemployed and *HouseholdSize* = 7, 8 or 10+ and *Province* = Gauteng, KwaZulu-Natal or North West and *Transportation* = Train or Car **then** Prospect.

Rule 12: **if** *Province* = Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West or Western Cape and *EmploymentStatus* = Employed or Unemployed and *HouseholdSize* = 7, 8 or 10+ and *Province* = Gauteng, KwaZulu-Natal or North West and *Transportation* = Bus, Taxi, Walk/cycle or Other **then** (New) low spender.

Rule 13: **if** *Province* = Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West or Western Cape and *EmploymentStatus* = Employed or Unemployed and *HouseholdSize* = 7, 8 or 10+ and *Province* = Limpopo, Mpumalanga, Northern Cape or Western Cape and *Transportation* = Bus, Taxi, Walk/cycle or Other **then** (New) low spender.

Rule 14: **if** *Province* = Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West or Western Cape and *EmploymentStatus* = Not economically active and *Age* = 30-34, 35-39 or 40-44 and *HouseholdSize* = 2, 7, 8 or 9 and *MobilePhoneType* = Motorola **then** (New) big spender.

Continued on next page

Rule 15: **if** *Province* = Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West or Western Cape and *EmploymentStatus* = Not economically active and *Age* = 30-34, 35-39 or 40-44 and *HouseholdSize* = 2, 7, 8 or 9 and *MobilePhoneType* = Samsung, Other, Apple, Huawei, Nokia, Blackberry, Sony, LG, HTC or Siemens **then** Prospect.

Rule 16: **if** *Province* = Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West or Western Cape and *EmploymentStatus* = Not economically active and *Age* = 30-34, 35-39 or 40-44 and *HouseholdSize* = 1, 3, 4, 5, 6 or 10+ **then** (New) big spender.

Rule 17: **if** *Province* = Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West or Western Cape and *EmploymentStatus* = Not economically active and *Age* = 15-19, 20-24, 25-29, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79 or 80+ and *Education* = Less than Gr.12 and no other qualification, Less than Gr.12 and with diploma or certificate, Gr.12, Gr.12 with diploma or certificate or Degree or post graduate degree **then** Prospect.

Rule 18: **if** *Province* = Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West or Western Cape and *EmploymentStatus* = Not economically active and *Age* = 15-19, 20-24, 25-29, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79 or 80+ and *Education* = Honours degree or higher **then** (New) low spender.

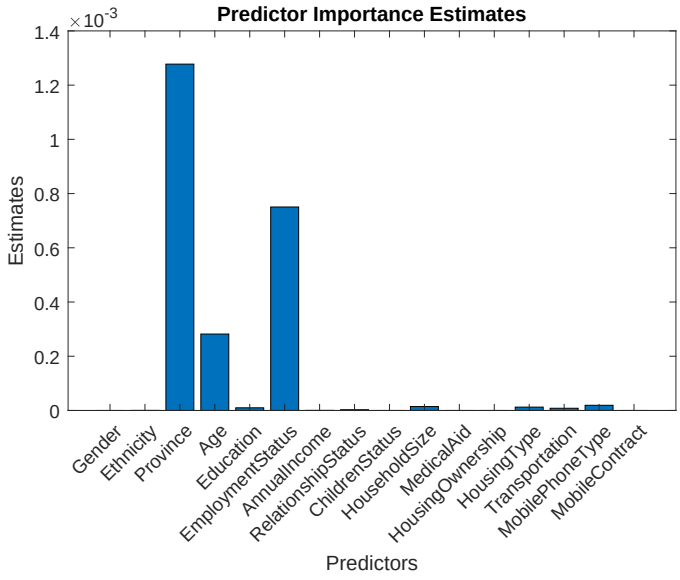


Figure 5.6: Predictor importance for determining the type of customer

The various customer profiles received from Table 5.9 can be presented in a more 'user friendly' manner. Figure 5.7 illustrates two customer profiles of (new) big spenders that

were discovered by the decision rules (rule 14 and rule 16). The difference between the

Employment status:

	Employed
	Unemployed
v	Not economically active

Age:

	15-19		50-54
	20-24		55-59
	25-29		60-64
v	30-34		65-69
v	35-39		70-74
v	40-44		75-79
	45-49		80+

Province:

	Eastern Cape
	Free State
v	Gauteng
v	KwaZulu-Natal
v	Limpopo
v	Mpumalanga
v	Northern Cape
v	North West
v	Western Cape

Household size:

	1		6
v	2	v	7
	3		8
	4	v	9
	5		10+

Mobile phone type:

	Samsung		Siemens
	Other		Blackberry
	Apple		Sony
	Huawei		LG
v	Motorola		HTC
	Nokia		

Province:

	Eastern Cape
	Free State
v	Gauteng
v	KwaZulu-Natal
v	Limpopo
v	Mpumalanga
v	Northern Cape
v	North West
v	Western Cape

Household size:

v	1	v	6
	2		7
v	3	v	8
v	4		9
v	5	v	10+

Age:

	15-19		50-54
	20-24		55-59
	25-29		60-64
v	30-34		65-69
v	35-39		70-74
v	40-44		75-79
	45-49		80+

Employment status:

	Employed
	Unemployed
v	Not economically active

Figure 5.7: Illustrating two customer profiles of (new) big spenders, following rule 14 and rule 16

two profiles is also more clear when schematically presented. For example, the top profile (in Figure 5.7) indicates the mobile phone type of the customer, whereas the bottom profile does not indicate the mobile phone type. The major difference between the two customer profiles of the same customer type is the household size. Having various customer profiles for the same type of customer provides more clarity and assurance when classifying a new customer. For example, a new member to the database might

not provide all their demographic and extra value adding information, and yet it will still be possible to find a suitable customer type for such a customer.

When comparing this customer profile ((new) big spender shown in Figure 5.7) to another customer profile, such as a low loyal customer, different customer features are used to profile these customers. Figure 5.8 illustrates two customer profiles of low loyal spenders, discovered by the decision rules (Table 5.9: rule 2 and rule 5). According

Education:

	Less than Gr.12 and no other qualification		Gr.12 with diploma or certificate
v	Less than Gr.12 and with diploma or certificate	v	Degree or postgraduate degree
	Gr.12		Honours degree or higher

Age:

	15-19	v	50-54
v	20-24		55-59
	25-29		60-64
v	30-34	v	65-69
v	35-39		70-74
v	40-44		75-79
	45-49	v	80+

Province:

v	Eastern Cape
v	Free State
	Gauteng
	KwaZulu-Natal
	Limpopo
	Mpumalanga
	Northern Cape
	North West
	Western Cape

Household size:

v	1	v	6
v	2	v	7
	3		8
v	4		9
	5	v	10+

Mobile phone type:

	Samsung	v	Siemens
	Other	v	Blackberry
v	Apple		Sony
	Huawei		LG
	Motorola	v	HTC
v	Nokia		

Province:

v	Eastern Cape
v	Free State
	Gauteng
	KwaZulu-Natal
	Limpopo
	Mpumalanga
	Northern Cape
	North West
	Western Cape

Age:

v	15-19		50-54
	20-24	v	55-59
v	25-29	v	60-64
	30-34		65-69
	35-39	v	70-74
	40-44	v	75-79
v	45-49		80+

Housing type:

	Cluster house in complex		Other
v	Flat or apartment in flat block		Room or granny flat or large dwelling
	House or brick structure on yard or strand		Semi-detached house
	House, flat or room in backyard	v	Townhouse
	Informal - Shack in backyard	v	Traditional dwelling – hut
v	Informal - Shack not backyard	v	Overcrowding

Figure 5.8: Illustrating two customer profiles of low loyal customers, following rule 2 and rule 5

to the customer profiles presented in Figures 5.7 and 5.8, the difference between being classified as a (new) big spender or a low loyal customer is dependent on the customer's age and province. Other customer features also influence the classification, as seen in Table 5.9 and Figure 5.6.

Not all types of customers are restricted to one cluster. Therefore, after the customer type is known (Table 5.9), the cluster to which that customer belongs can be determined. Table 5.8 indicated that churned cheap customers only belong to cluster 2, whereas prospective customers belong only to cluster 3. A set of decision rules for each type of customer is constructed. These rules can be used to (1) predict to which cluster a specific type of customer belongs *via* a customer super-profile as well as (2) provide customer super-profiles for targeted marketing campaigns, when the type of customer (*e.g.* low loyal customer) is known. As noted before, the cRFM values allocated to each type of customer refines the customer profiles, forming the *super-profiles*.

For (new) low spenders, only customers that belong to the specific cRFM 'groups' indicated in Table 5.8, columns three and four were selected. The rules are shown in Table 5.10. When a marketer wants to target a customer who is classified as a (new) low spender, they can follow the rules in Table 5.10 and identify the profile of such a customer. For example, when the identified customer is allocated to cluster 1, rule 1, rule 2, rule 3 and/or rule 6 customers, it is known that their cRFM value would be equal to 3. As apparent, this list indicates only some of the decision rules contained in the decision tree. The rules that are reported have a misclassification rate of 9.3 percent. This means that, for instance, at least 90.7 percent of customers following rule 1 are in cluster 1. Figure 5.9 revealed that the customer features with the *most* influence on the decision rules shown in Table 5.10 are: employment status, age, mobile phone type, relationship status and province.

Table 5.10: Decision rules to identify (new) low spenders

Rule 1:	**if** *EmploymentStatus* = Employed or Unemployed **then** Cluster 1.
Rule 2:	**if** *EmploymentStatus* = Not economically active and *Province* = Eastern Cape or Free State and *Age* = 15-19, 20-24, 25-29, 40-44, 45-49, 50-54, 55-59, 60-64, 70-74 or 80+ **then** Cluster 1.

Continued on next page

Rule 3: **if** *EmploymentStatus* = Not economically active and *Province* = Eastern Cape or Free State and *Age* = 30-34, 35-39, 65-69 or 75-79 and *MobilePhoneType* = Samsung, Other, Huawei, Nokia, Blackberry, Sony, LG or Motorola and *RelationshipStatus* = Married/domestic partner, Never married/single or Widowed **then** Cluster 1.

Rule 4: **if** *EmploymentStatus* = Not economically active and *Province* = Eastern Cape or Free State and *Age* = 30-34, 35-39, 65-69 or 75-79 and *MobilePhoneType* = Samsung, Other, Huawei, Nokia, Blackberry, Sony, LG or Motorola and *RelationshipStatus* = Divorced **then** Cluster 2.

Rule 5: **if** *EmploymentStatus* = Not economically active and *Province* = Eastern Cape or Free State and *Age* = 30-34, 35-39, 65-69 or 75-79 and *MobilePhoneType* = Apple **then** Cluster 2.

Rule 6: **if** *Province* =Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, North West, Northern Cape or Western Cape **then** Cluster 1.

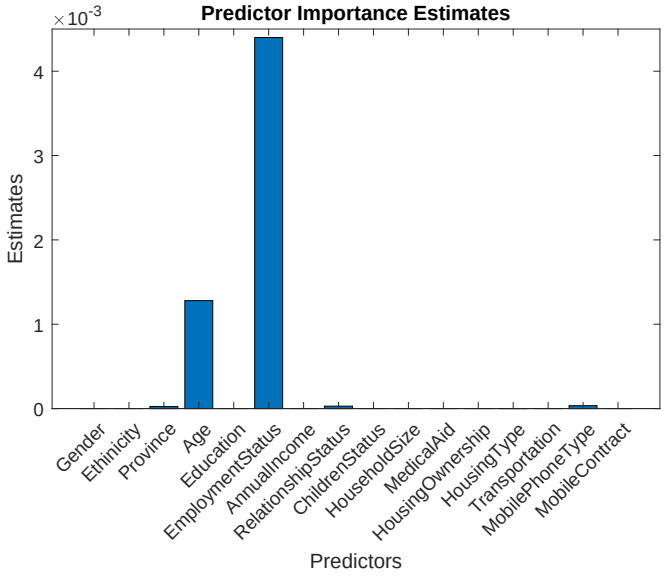


Figure 5.9: Predictor importance for (new) low spenders

Next, the decision rules for (new) big spenders are generated and shown in Table 5.11. The misclassification error is 10.9 percent. Figure 5.10 revealed that the customer features with the *most* influence on the decision rules shown in Table 5.11 are: age, housing type, education, mobile phone type, province, household size, housing ownership, ethnicity and gender. The number of predictors that are considered when predicting the cluster to which (new) big spenders belong is noticeably more than when predicting the cluster to which (new) low spenders belong. This indicates that different customer

features are important in order to be classified as a big spender.

When predicting (new) big spenders, the age category is significantly more important than any other feature. Customers in the age range of 30-44 years, and education to honours degree or higher belong to cluster 1, with a cRFM value of 5 (pre-specified), as shown by rule 12. The same age group, but with a different educational background, belongs to cluster 2, and will have a cRFM value equal to 4 or 5, as seen by rule 10. These rules provide customer profiles for (new) big spenders.

Table 5.11: Decision rules to identify (new) big spenders

Rule 1:	**if** *Age* = 15-19, 20-24, 25-29, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79 or 80+ and *Gender* = Male and *ChildrenStatus* = Yes and *Age* = 20-24, 45-49, 65-69, 75-79 or 80+ **then** Cluster 1.
Rule 2:	**if** *Age* = 15-19, 20-24, 25-29, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79 or 80+ and *Gender* = Male and *ChildrenStatus* = Yes and *Age* = 15-19, 25-29, 50-54, 55-59, 60-64 or 70-74 and *HousingType* = Flat or apartment in flat block, House or brick structure on yard or strand, Informal - Shack not backyard, Semi-detached house, Townhouse, Traditional dwelling - hut or Overcrowding **then** Cluster 1.
Rule 3:	**if** *Age* = 15-19, 20-24, 25-29, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79 or 80+ and *Gender* = Male and *ChildrenStatus* = Yes and *Age* = 15-19, 25-29, 50-54, 55-59, 60-64 or 70-74 and *HousingType* = House, flat or room in backyard, Informal - Shack in backyard or Room, granny flat or large dwelling **then** Cluster 2.
Rule 4:	**if** *Age* = 15-19, 20-24, 25-29, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79 or 80+ and *Gender* = Male and *ChildrenStatus* = No and *HousingOwnership* = Rented and *MobilePhoneType* = Samsung, Other, Apple, Nokia, Blackberry or LG **then** Cluster 1.
Rule 5:	**if** *Age* = 15-19, 20-24, 25-29, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79 or 80+ and *Gender* = Male and *ChildrenStatus* = No and *HousingOwnership* = Rented and *MobilePhoneType* = Huawei, Sony or Siemens **then** Cluster 2.
Rule 6:	**if** *Age* = 15-19, 20-24, 25-29, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79 or 80+ and *Gender* = Male and *ChildrenStatus* = No and *HousingOwnership* = Owned (not fully), Owned (fully), Occupied rent free or Other **then** Cluster 1.
Rule 7:	**if** *Age* = 15-19, 20-24, 25-29, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79 or 80+ and *Gender* = Female **then** Cluster 1.

Continued on next page

Rule 8: **if** *Age* = 30-34, 35-39 or 40-44 and *Education* = Less than Gr.12 and no other qualification, Less than Gr.12 and with diploma or certificate, Gr.12 or Gr.12 with diploma or certificate and *Age* = 30-34 **then** Cluster 2.

Rule 9: **if** *Age* = 30-34, 35-39 or 40-44 and *Education* = Less than Gr.12 and no other qualification, Less than Gr.12 and with diploma or certificate, Gr.12 or Gr.12 with diploma or certificate and *Age* = 35-39 or 40-44 and *HouseholdSize* = 1, 2, 5, 6 or 9 **then** Cluster 2.

Rule 10: **if** *Age* = 30-34, 35-39 or 40-44 and *Education* = Less than Gr.12 and no other qualification, Less than Gr.12 and with diploma or certificate, Gr.12 or Gr.12 with diploma or certificate and *Age* = 35-39 or 40-44 and *HouseholdSize* = 3, 4, 7, 8 or 10+ and *Province* = Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape or North West **then** Cluster 2.

Rule 11: **if** *Age* = 30-34, 35-39 or 40-44 and *Education* = Less than Gr.12 and no other qualification, Less than Gr.12 and with diploma or certificate, Gr.12 or Gr.12 with diploma or certificate and *Age* = 35-39 or 40-44 and *HouseholdSize* = 3, 4, 7, 8 or 10+ and *Province* = Western Cape **then** Cluster 1.

Rule 12: **if** *Age* = 30-34, 35-39 or 40-44 and *Education* = Honours degree or higher **then** Cluster 1.

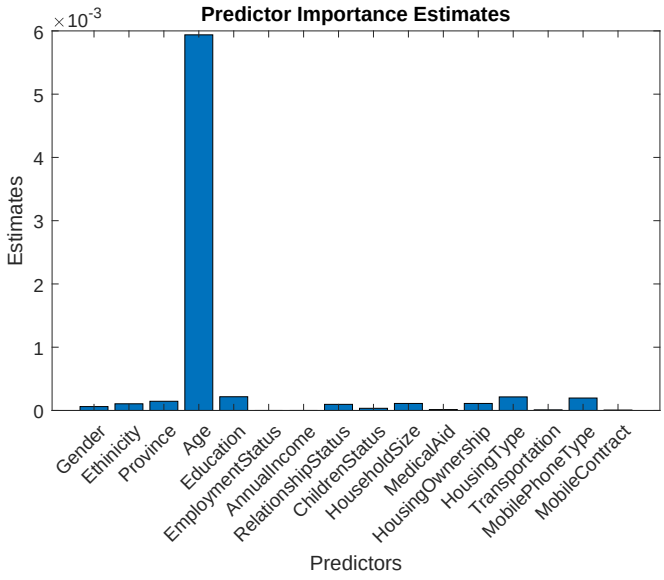


Figure 5.10: Predictor importance for (new) big spenders

Next, the decision rules for low loyal customers are shown in Table 5.12. The misclassification rate for this decision tree is 9.8 percent. When the customer belongs to

cluster 1, it is known that their cRFM value will be equal to 4, as opposed to when a customer belongs to cluster 2, when their cRFM value will be equal to 3. Figure 5.11 revealed that the customer features with the *most* influence on the decision rules shown in Table 5.12 are: province, age, employment status, mobile phone type and household size.

Table 5.12: Decision rules to identify low loyal customers

Rule 1:	**if** *Province* = Eastern Cape or Free State and *AnnualIncome* = R12 001 – R54 000 or R54 001 – R192 000 and *HousingOwnership* = Owned (not fully) and *HousingType* = House, flat or room in backyard **then** Cluster 1.
Rule 2:	**if** *Province* = Eastern Cape or Free State and *AnnualIncome* = R12 001 – R54 000 or R54 001 – R192 000 and *HousingOwnership* = Owned (not fully) and *HousingType* = Cluster house in complex, Flat or apartment in flat block, House or brick structure on yard or strand, Informal - Shack in backyard, Informal - Shack not backyard, Other, Room, granny flat or large dwelling, Semi-detached house, Townhouse, Traditional dwelling – hut or Overcrowding **then** Cluster 2.
Rule 3:	**if** *Province* = Eastern Cape or Free State and *AnnualIncome* = R12 001 – R54 000 or R54 001 – R192 000 and *HousingOwnership* = Rented, Owned (fully), Occupied rent free or Other and *RelationshipStatus* = Married/domestic partner or Never married/single **then** Cluster 2.
Rule 4:	**if** *Province* = Eastern Cape or Free State and *AnnualIncome* = R12 001 – R54 000 or R54 001 – R192 000 and *HousingOwnership* = Rented, Owned (fully), Occupied rent free or Other and *RelationshipStatus* = Widowed or Divorced and *Age* = 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54 or 55-59 **then** Cluster 2.
Rule 5:	**if** *Province* = Eastern Cape or Free State and *AnnualIncome* = R12 001 – R54 000 or R54 001 – R192 000 and *HousingOwnership* = Rented, Owned (fully), Occupied rent free or Other and *RelationshipStatus* = Widowed or Divorced and *Age* = 60-64 or 75-79 **then** Cluster 1.
Rule 6:	**if** *Province* = Eastern Cape or Free State and *AnnualIncome* = R0 – R12 000, R192 001 – R360 000 or More than R360 000 and *HousingType* = House or brick structure on yard or strand, House, flat or room in backyard, Informal – Shack in backyard, Informal – Shack not backyard, Other, Townhouse, Traditional dwelling – hut or Overcrowding and *HousingOwnership* = Rented, Owned (not fully), Owned (fully) or Occupied rent free **then** Cluster 2.

Continued on next page

Rule 7: **if** *Province* = Eastern Cape or Free State and *AnnualIncome* = R0 – R12 000, R192 001 – R360 000 or More than R360 000 and *HousingType* = House or brick structure on yard or strand, House, flat or room in backyard, Informal – Shack in backyard, Informal – Shack not backyard, Other, Townhouse, Traditional dwelling – hut or Overcrowding and *HousingOwnership* = Other and *Age* = 35-39, 55-59 or 80+ **then** Cluster 1.

Rule 8: **if** *Province* = Eastern Cape or Free State and *AnnualIncome* = R0 – R12 000, R192 001 – R360 000 or More than R360 000 and *HousingType* = House or brick structure on yard or strand, House, flat or room in backyard, Informal – Shack in backyard, Informal – Shack not backyard, Other, Townhouse, Traditional dwelling – hut or Overcrowding and *HousingOwnership* = Other and *Age* = 15-19, 20-24, 25-29, 30-34, 40-44, 50-54, 60-64, 65-69 or 70-74 **then** Cluster 2.

Rule 9: **if** *Province* = Eastern Cape or Free State and *AnnualIncome* = R0 – R12 000, R192 001 – R360 000 or More than R360 000 and *HousingType* = Cluster house in complex, Flat or apartment in flat block, Room, granny flat or large dwelling or Semi-detached house **then** Cluster 2.

Rule 10: **if** *Province* = Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, North West, Northern Cape or Western Cape and *EmploymentStatus* = Employed or Unemployed **then** Cluster 1.

Rule 11: **if** *Province* = Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, North West, Northern Cape or Western Cape and *EmploymentStatus* = Not economically active and *Age* = 30-34, 35-39 or 40-45 **then** Cluster 1.

Rule 12: **if** *Province* = Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, North West, Northern Cape or Western Cape and *EmploymentStatus* = Not economically active and *Age* = 15-19, 20-24, 25-29, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79 or 80+ and *HouseholdSize* = 1, 2, 4 or 6 **then** Cluster 1.

Rule 13: **if** *Province* =Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, North West, Northern Cape or Western Cape and *EmploymentStatus* = Not economically active and *Age* = 15-19, 20-24, 25-29, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79 or 80+ and *HouseholdSize* = 3, 5, 7, 8, 9 or 10+ and *MobilePhoneType* = Apple or Blackberry **then** Cluster 1.

Rule 14: **if** *Province* =Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, North West, Northern Cape or Western Cape and *EmploymentStatus* = Not economically active and *Age* = 15-19, 20-24, 25-29, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79 or 80+ and *HouseholdSize* = 3, 5, 7, 8, 9 or 10+ and *MobilePhoneType* = Samsung, Other, Huawei, Nokia, Sony, LG, HTC, Motorola or Siemens **then** Cluster 2.

As indicated in Table 5.8, the churned cheap customers are not of interest to marketers. However, in this customer dataset, churned cheap customers only belong to

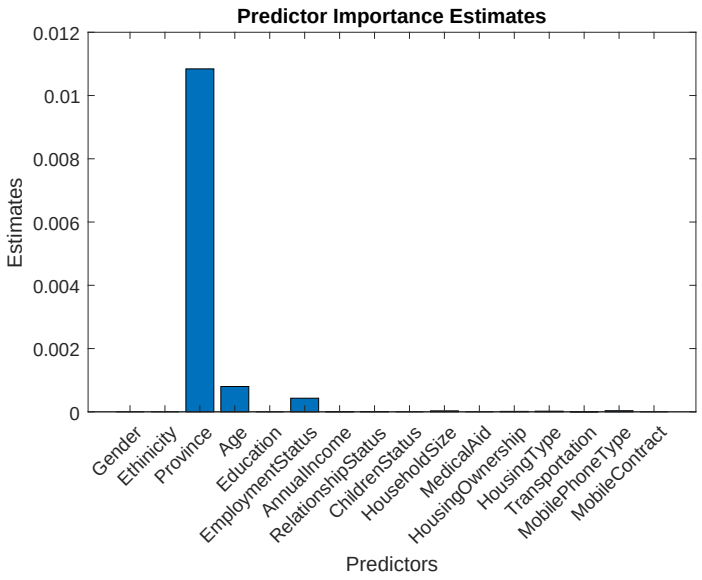


Figure 5.11: Predictor importance for low loyal customers

cluster 2, have a cRFM value of 1, and only make up 0.068 percent of the customers within cluster 2. No decision rules are necessary to profile this type of customer. The last type of customer is the *prospects*. These customers only belong to cluster 3, for no transactional information is registered in the database for such customers, only demographic and extra value adding features. Therefore no additional decision rules are necessary.

Comparing the individual decision rules for the type of customers (Tables 5.10 – 5.12), it is found that the rules used to identify (new) low spenders (Table 5.10) have the lowest misclassification rate. This can be as a result of the (new) low spenders being the biggest customer dataset, thus having more training and testing data.

This concludes the discussion regarding the development of the predictive model. The purpose of this research, as indicated in Section 1.2, was to develop a tool that can generate customer super-profiles given a specific dataset, through performing the steps illustrated in Figure 5.1. The predictive model for this research, which forms part of the CSP tool, includes various sets of decision rules, all leading to creating customer super-profiles for the five customer types present within the dataset (Table 5.8). Marketers using this tool will have more knowledge of their customers especially when they know which type of customer they are interested in. This model is able to perform classification

and prediction for existing and/or future customers. The ability of the model will be presented through business case scenarios, providing clear understanding and the context in which the CSP tool can be utilised.

5.2 Business case

The previous sections indicated how to develop a customer super-profile to enable efficient targeting in marketing campaigns. The business case to illustrate the value added by this research is as follows.

Being able to identify *who* to target, as well as *where* and *how* to advertise marketing campaigns is an important task. The proposed CSP tool has the ability to run a deterministic audience discovery to reveal customer profiles for the marketers. These profiles contain demographic information, typical transactional data as well as customer preferences. The tool can be used when marketing a product for a certain target group is necessary. This target group can be 'found' by the tool and will provide all relevant 'customer' information regarding that group, *e.g.* demographic information, transportation type, mobile phone ownership and activities as well as RFM values. This type of information will provide more insight into the customers and will decrease the frustration experienced by marketers when performing 'tossing a coin' type of targeting.

The business value of the CSP tool will be demonstrated by various scenarios. All of the scenarios will be regarding the simulated customer datasets.

5.2.1 Business case scenario 1: Targeting customers

The first scenario assumes that a marketing team desires to target a low loyal *existing* customer, with an age of 25 years or older, for a given marketing campaign. The decision rules in Table 5.9 provide the suitable customer super-profiles for such customers. The only required customer feature is the age, which focuses the profiles on a more specific group. The customer super-profiles are provided by rules 5 and 6 and are as follows.

The customer(s) can be found in the Eastern Cape or the Free State. Their type of housing is specified as a flat or apartment in flat block, informal – shack not backyard, townhouse, traditional dwelling – hut or overcrowding. When their housing type is specified as a cluster house in complex, house or brick structure on yard or strand, house, flat or room in backyard, informal – shack in backyard, other, room or granny

flat or large dwelling or semi-detached house; their type of transportation is specified as either a bus, taxi, car or walk/cycle.

This customer profile provides a broader perspective of how a low loyal customer of 25 years and/or older *looks*. The marketers can decide how, where and with what to target this group of customers after receiving the profile.

Assume that a marketing company wants to target less conventional customers, such as affluent young customers. Using the CSP tool it is possible to find suitable customer profiles. For example, rule 2 in Table 5.9 indicates a customer profile suitable for a affluent young person, identifying them as a low loyal customer. For a more specific customer profile, the decision rules in Table 5.12 are investigated. Rule 6 and rule 8, together with rule 2 (from Table 5.9) lead to customer profiles that the marketers would be interested in. It is identified that the customer falls in the age range of 20-24 years old, lives in Eastern Cape or Free State, has an Apple (iPhone) mobile phone, has an educational background of a degree or post graduate degree, earns an annual income of more than R360 000 as well as owning a house. The marketer can realise through the customer profile, that even though it is an affluent young customer, they are still classified as a *low* loyal customer. This information can be used to target the customer in various ways, such as targeting with small purchases, yet regularly, or try to get the customer to spend more.

5.2.2 Business case scenario 2: New members

The second scenario illustrates how the decision rules would be applied if a *new* member enters the database. No *default* type is assigned to the new member, only that they are *new*, which means that they can be a *new* low spender, *new* big spender, *new* loyal customer *etc.* Table 5.13 presents an example of a new member's features. Firstly, the decision rules in Table 5.9 are used to assign a customer type to the new member. Considering the member's features, rule 8 is applicable; identifying the new member as a *new low spender*.

Next, Table 5.10 is utilised to assign the customer to a cluster. The customer complies with both rule 1 and rule 6, which predict that the customer belongs to cluster 1. A new low spender that is assigned to cluster 1 has a cRFM value of 3. All of this customer information and predictions will allow for easier and less time-consuming marketing efforts.

Table 5.13: First new member information

Customer feature			
Gender	Male	Children status	Yes
Ethnicity	Coloured	Household size	3
Province	Western Cape	Medical aid	–
Age	33	Housing ownership	–
Education	Gr.12 with diploma	Housing type	Townhouse
Employment status	Employed	Transportation	Car and bus
Annual income	–	Mobile phone type	Huawei
Relationship status	Married	Mobile contract	Contract

To illustrate that different information leads to different predictions, Table 5.14 presents the information of a second new member. The decision rules in Table 5.9 are used to assign a customer type to the second new member. Considering the member's features, rule 6 is most applicable; identifying the new member as a *new low loyal customer*. Even though Table 5.8 does not list *new* low loyal customer as a type of customer, adding *new* to the type only indicates that it is a *new customer* and provides a probation period for the customer to change type (reclassify) when actually building a purchasing pattern, different than that predicted.

Table 5.14: Second new member information

Customer feature			
Gender	Female	Children status	No
Ethnicity	White	Household size	1
Province	Free State	Medical aid	–
Age	19	Housing ownership	Rent
Education	Gr.12	Housing type	Flat
Employment status	Employed	Transportation	Car
Annual income	R108 000	Mobile phone type	Apple
Relationship status	Single	Mobile contract	–

Allocating this new low loyal customer to a cluster, the decision rules in Table 5.12 are followed. Considering the new customer's information, rule 3 is most applicable, allocating this new customer to cluster 2.

5.2.3 Business case scenario 3: Change the customer type

Scenario 3 illustrates the possibility for a prospective customer (cluster 3 customers) to change into a different type of customer, considering their features. Before having to find new customers, existing prospective customers, with their features known, can be 'transformed' into possible customers participating in retail shop activities. This scenario is modelled for those customers that are already in the database, yet not exploited. Table 5.15 indicates the type of customer that was predicted for 10 randomly selected prospective customers, by making use of a trained decision tree.

Table 5.15: Predicting the possible purchasing behaviour of the existing prospective customers

Prospect	
Customer	***Predict***
301	New low loyal customer
3812	New big spender
1210	New low spender
1355	New low spender
2216	New low spender
787	New low loyal customer
3156	New low loyal customer
4052	New low spender
7731	New low spender
5082	New low spender

5.3 Decision tree analysis: Camping problem

The researcher presented a '*large dataset problem*' regarding the *participation behaviour of campers* in Section 4.2. After performing analysis on the camping dataset, the results illustrated customer (camper) profiles for each cluster. These profiles provide insight into each cluster, and show how the campers in each cluster differ. However, the researcher believes that by revisiting this problem and applying the CSP tool's methodology (Figure 5.1) deeper insight can be reached.

The first two steps of the CSP tool were already performed, as the dataset was selected (1) and RFM analysis was performed (2). The third step of the CSP tool is to perform clustering on the RFM dataset, this was also already performed, three clusters

were identified. However, step 3 also instructs to calculate the cRFM, which will be done subsequently.

5.3.1 Clustering: Camping dataset

As mentioned, the camping dataset contains *three clusters*. The cRFM category values were calculated for each cluster, according to (5.2). Table 5.16 represents a summary of the percentage of campers in each cRFM category, for clusters 1, 2 and 3. The majority of campers in cluster 1 have high cRFM values, whereas most campers in cluster 2 have lower cRFM values, and the campers that belong to cluster 3 have the lowest cRFM index.

Table 5.16: Percentage campers in each cRFM category for clusters 1, 2 and 3

	Percentage:		
	Cluster 1 (33 791)	**Cluster 2 (50 352)**	**Cluster 3 (15 857)**
cRFM = 1	0%	5.38%	82.95%
cRFM = 2	0%	38.58%	17.05%
cRFM = 3	0%	56.04%	0%
cRFM = 4	52.42%	0%	0%
cRFM = 5	47.58%	0%	0%

As performed earlier in this chapter, the various camper types are identified with the help of the three clusters and the campers' cRFM category values (RFM pattern). This information makes it possible to discover various types of campers present in this dataset. These types of campers are identified and listed in Table 5.17 together with their identification traits which are the cRFM category values and RFM patterns. The explanation of the type of customer (camper) can be seen in Table 5.8.

Next, a predictive model, including the campers' features (as seen in Table 5.18) will be developed, according to the various types of campers. It is anticipated that this model will provide deeper insight into the campers' participation behaviour, by generating super-profiles.

5.3.2 Predictive model: Camping dataset

The profiles that were presented in Chapter 4 gave an intermediate summary of the campers' demographic information per cluster. This section will present the super-

Table 5.17: Types of campers in dataset

Type of camper	**Cluster 1**	**Cluster 2**	**Cluster 3**
(New) low spenders			cRFM = 2 R ↑ F↓ M ↓
(New) big spenders	cRFM = 5 R ↑ F↑ M ↑	cRFM = 3 R ↑ F↑ M ↑	
Loyal big spenders	cRFM = 4 R ↓ F↑ M ↑	cRFM = 2 R ↓ F↑ M ↑	
Churned cheap campers		cRFM = 1 R ↓ F↓ M ↓	cRFM = 1 R ↓ F↓ M ↓

Table 5.18: Camper's features

Variable name	**Explanation**	**Scaling**
Gender	Male or Female	Categorical (Dichotomous, Figure 3.1)
Age	6-32, 32-48, 48-62 or 62-80	Categorical (Multichotomous, Figure 3.1)
Occupation	Student, Employed, Home-maker or Retired	Categorical (Multichotomous, Figure 3.1)
Annual income	Low, Medium or High	Categorical (Multichotomous, Figure 3.1)
Children status	Yes or No	Categorical (Dichotomous, Figure 3.1)
Marital status	Yes or No	Categorical (Dichotomous, Figure 3.1)
Reservations	1-3 days, 4-7 days, 2-4 weeks, 1 month+, 6 months+	Categorical (Multichotomous, Figure 3.1)
Type of shelter	Tent, RV, Bivy/No shelter	Categorical (Multichotomous, Figure 3.1)

profiles of the campers when utilising the CSP tool.

The previous results (in Chapter 4) indicated that 9 percent of the campers in cluster 1 were students, 18 percent employed and 73 percent were retired. From that group of campers, 63 percent receive a medium annual income and 37 percent a high annual

income. The rest of cluster 1 campers' features can be seen in Table 4.8. Having access to this information provides an overall perspective of the cluster, yet it does not state whether the student, employed camper or retired camper earn the medium or high annual income, or the type of camper they are.

The decision rules shown in Table 5.19 are used to identify the type of campers as listed in Table 5.17. When a campers' information is known, the rules shown in Table 5.19 can determine the camper's type. For example, rule 8 and rule 9 represent similar camper profiles: the camper is a 62-80 year old student or is retired, who makes reservations more than 4 days in advance. However, rule 8 indicates that the camper has a child (or children), while rule 9 indicates that the camper does not have a child. This one feature results in either classifying the camper as a churned cheap camper (rule 8: camper has child) or as a loyal big spender (rule 9: camper does not have a child). It can be concluded from these profiles that campers at a certain age (62-80) with children are less frequent campers than similar campers without children. If a campaign makes use of this information as a selling point, a more specific group will be targeted, which can lead to campaign success.

Table 5.19: Decision rules to identify the type of camper

Rule 1:	**if** *Occupation* = Student or Retired and *Age* = 16-32 or 48-62 and *AnnualIncome* = Low **then** (New) low spender.
Rule 2:	**if** *Occupation* = Student or Retired and *Age* = 16-32 or 48-62 and *AnnualIncome* = Medium or High and *TypeOfShelter* = Tent or RV and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ **then** (New) big spender.
Rule 3:	**if** *Occupation* = Student or Retired and *Age* = 16-32 or 48-62 and *AnnualIncome* = Medium or High and *TypeOfShelter* = Tent or RV and *Reservations* = 1-3 days **then** (New) low spender.
Rule 4:	**if** *Occupation* = Student or Retired and *Age* = 16-32 or 48-62 and *AnnualIncome* = Medium or High and *TypeOfShelter* = Bivy/No shelter **then** (New) low spender.
Rule 5:	**if** *Occupation* = Student or Retired and *Age* = 32-48 or 62-80 and *Reservations* = 1 month+ or 6 months+ and *AnnualIncome* = Low **then** (New) low spender.
Rule 6:	**if** *Occupation* = Student or Retired and *Age* = 32-48 or 62-80 and *Reservations* = 1 month+ or 6 months+ and *AnnualIncome* = Medium or High **then** Loyal big spender.

Continued on next page

Rule 7: **if** *Occupation* = Student or Retired and *Age* = 32-48 or 62-80 and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *Age* = 32-48 **then** (New) low spender.

Rule 8: **if** *Occupation* = Student or Retired and *Age* = 32-48 or 62-80 and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *Age* = 62-80 and *ChildrenStatus* = Yes **then** Churned cheap camper.

Rule 9: **if** *Occupation* = Student or Retired and *Age* = 32-48 or 62-80 and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *Age* = 62-80 and *ChildrenStatus* = No **then** Loyal big spender.

Rule 10: **if** *Occupation* = Employed or Home-maker and *AnnualIncome* = Low **then** (New) low spender.

Rule 11: **if** *Occupation* = Employed or Home-maker and *AnnualIncome* = Medium or High and *Reservations* = 6 months+ and *AnnualIncome* = Medium and *TypeOfShelter* = RV **then** Loyal big spender.

Rule 12: **if** *Occupation* = Employed or Home-maker and *AnnualIncome* = Medium or High and *Reservations* = 6 months+ and *AnnualIncome* = Medium and *TypeOfShelter* = Tent or Bivy/No shelter **then** (New) low spender.

Rule 13: **if** *Occupation* = Employed or Home-maker and *AnnualIncome* = Medium or High and *Reservations* = 6 months+ and *AnnualIncome* = High and *ChildrenStatus* = Yes **then** (New) big spender.

Rule 14: **if** *Occupation* = Employed or Home-maker and *AnnualIncome* = Medium or High and *Reservations* = 6 months+ and *AnnualIncome* = High and *ChildrenStatus* = No **then** Loyal big spender.

Rule 15: **if** *Occupation* = Employed or Home-maker and *AnnualIncome* = Medium or High and *Reservations* = 1-3 days, 4-7 days, 2-4 weeks or 1 month+ and *ChildrenStatus* = Yes **then** Churned cheap camper.

Rule 16: **if** *Occupation* = Employed or Home-maker and *AnnualIncome* = Medium or High and *Reservations* = 1-3 days, 4-7 days, 2-4 weeks or 1 month+ and *ChildrenStatus* = No and *AnnualIncome* = Medium and *Reservations* = 1 month+ **then** Loyal big spender.

Rule 17: **if** *Occupation* = Employed or Home-maker and *AnnualIncome* = Medium or High and *Reservations* = 1-3 days, 4-7 days, 2-4 weeks or 1 month+ and *ChildrenStatus* = No and *AnnualIncome* = Medium and *Reservations* = 1-3 days, 4-7 days or 2-4 weeks **then** (New) low spender.

Rule 18: **if** *Occupation* = Employed or Home-maker and *AnnualIncome* = Medium or High and *Reservations* = 1-3 days, 4-7 days, 2-4 weeks or 1 month+ and *ChildrenStatus* = No and *AnnualIncome* = High **then** (New) big spender.

Figure 5.12 revealed that the campers' features with the most influence on the decision rules shown in Table 5.19 are: age, occupation, reservations, annual income, type

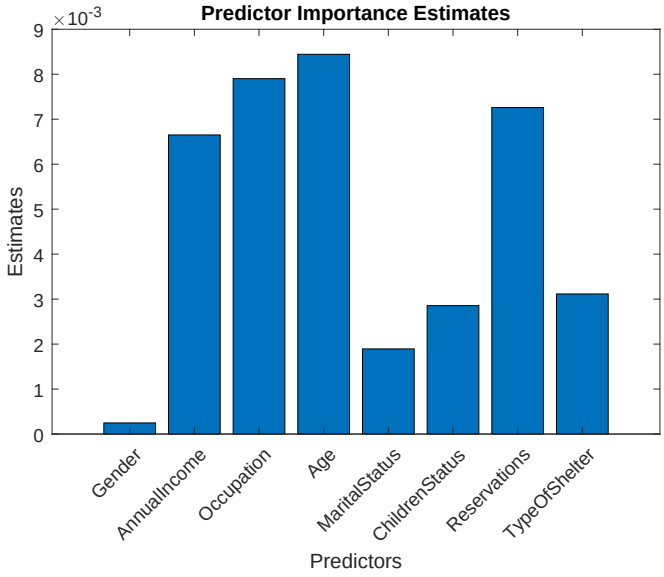


Figure 5.12: Predictor importance for identifying the type of camper

of shelter, children status, marital status and gender. The importance of the predictors is not known when initially constructing a decision tree. Thus, in this case, it is also initially not known which predictors are correct (gender is least important, while annual income is most important).

When the type of camper is known, the cluster to which the camper belongs is still unknown. Therefore, the following decision rules were constructed to allocate the camper to the appropriate cluster. When a camper within this dataset is identified as a *(new) low spender* they belong to cluster 3, as indicated in Table 5.17, and no additional decision rules will be generated for this type of camper. Table 5.20 shows the decision rules to be followed for allocating a *(new) big spender* to a cluster. For example, assume that a (new) big spender was identified by making use of the decision rules in Table 5.19, with the profile described in rule 13. The rules in Table 5.20 are used to identity the cluster to which this camper belongs. Rule 11 (Table 5.20) indicates a suitable camper profile given the information that are known, which indicates that the camper belongs to cluster 2.

Table 5.20: Decision rules to identify the cluster of a camper identified as a (new) big spender

Continued on next page

Rule 1:	**if** *Occupation* = Student or Retired and *Age* = 16-32, 32-48 or 48-62 and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *Reservations* = 4-7 days or 2-4 weeks and *ChildrenStatus* = Yes and *Reservations* = 4-7 days **then** Cluster 2.
Rule 2:	**if** *Occupation* = Student or Retired and *Age* = 16-32, 32-48 or 48-62 and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *Reservations* = 4-7 days or 2-4 weeks and *ChildrenStatus* = Yes and *Reservations* = 2-4 weeks **then** Cluster 1.
Rule 3:	**if** *Occupation* = Student or Retired and *Age* = 16-32, 32-48 or 48-62 and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *Reservations* = 4-7 days or 2-4 weeks and *ChildrenStatus* = No **then** Cluster 1.
Rule 4:	**if** *Occupation* = Student or Retired and *Age* = 16-32, 32-48 or 48-62 and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *Reservations* = 1 month+ or 6 months+ **then** Cluster 1.
Rule 5:	**if** *Occupation* = Student or Retired and *Age* = 16-32, 32-48 or 48-62 and *Reservations* = 6 months+ **then** Cluster 2.
Rule 6:	**if** *Occupation* = Student or Retired and *Age* = 62-80 and *Reservations* = 6 months+ **then** Cluster 1.
Rule 7:	**if** *Occupation* = Student or Retired and *Age* = 62-80 and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *ChildrenStatus* = Yes **then** Cluster 2.
Rule 8:	**if** *Occupation* = Student or Retired and *Age* = 62-80 and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *ChildrenStatus* = No and *Reservations* = 2-4 weeks or 1 month+ **then** Cluster 1.
Rule 9:	**if** *Occupation* = Student or Retired and *Age* = 62-80 and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *ChildrenStatus* = No and *Reservations* = 4-7 days or 6 months+ **then** Cluster 2.
Rule 10:	**if** *Occupation* = Employed or Home-maker and *Reservations* = 6 months+ **then** Cluster 1.
Rule 11:	**if** *Occupation* = Employed or Home-maker and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *ChildrenStatus* = Yes **then** Cluster 2.
Rule 12:	**if** *Occupation* = Employed or Home-maker and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *ChildrenStatus* = No and *Reservations* = 2-4 weeks or 1 month+ and *TypeOfShelter* = Tent **then** Cluster 2.
Rule 13:	**if** *Occupation* = Employed or Home-maker and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *ChildrenStatus* = No and *Reservations* = 2-4 weeks or 1 month+ and *TypeOfShelter* = RV **then** Cluster 1.

Continued on next page

Rule 14: **if** *Occupation* = Employed or Home-maker and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *ChildrenStatus* = No and *Reservations* = 4-7 days or 6 months+ **then** Cluster 2.

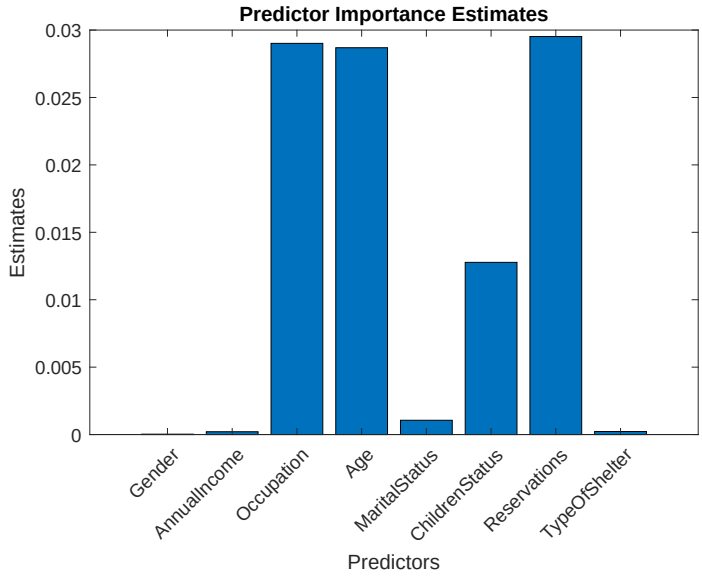


Figure 5.13: Predictor importance for (new) big spenders

The decision rules to identify the cluster number of loyal big spenders are shown in Table 5.21. The misclassification error for this set of rules is 4.9 percent. The decision rules followed for identifying the cluster allocated to both a (new) big spenders (Table 5.20) and a loyal big spender (Table 5.21) are the same. However, the misclassification error and predictor importance differ.

Table 5.21: Decision rules to identify the cluster of a camper identified as a loyal big spender

Rule 1: **if** *Occupation* = Student or Retired and *Age* = 16-32, 32-48 or 48-62 and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *Reservations* = 4-7 days or 2-4 weeks and *ChildrenStatus* = Yes and *Reservations* = 4-7 days **then** Cluster 2.

Rule 2: **if** *Occupation* = Student or Retired and *Age* = 16-32, 32-48 or 48-62 and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *Reservations* = 4-7 days or 2-4 weeks and *ChildrenStatus* = Yes and *Reservations* = 2-4 weeks **then** Cluster 1.

Continued on next page

Rule 3: **if** *Occupation* = Student or Retired and *Age* = 16-32, 32-48 or 48-62 and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *Reservations* = 4-7 days or 2-4 weeks and *ChildrenStatus* = No **then** Cluster 1.

Rule 4: **if** *Occupation* = Student or Retired and *Age* = 16-32, 32-48 or 48-62 and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *Reservations* = 1 month+ or 6 months+ **then** Cluster 1.

Rule 5: **if** *Occupation* = Student or Retired and *Age* = 16-32, 32-48 or 48-62 and *Reservations* = 6 months+ **then** Cluster 2.

Rule 6: **if** *Occupation* = Student or Retired and *Age* = 62-80 and *Reservations* = 6 months+ **then** Cluster 1.

Rule 7: **if** *Occupation* = Student or Retired and *Age* = 62-80 and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *ChildrenStatus* = Yes **then** Cluster 2.

Rule 8: **if** *Occupation* = Student or Retired and *Age* = 62-80 and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *ChildrenStatus* = No and *Reservations* = 2-4 weeks or 1 month+ **then** Cluster 1.

Rule 9: **if** *Occupation* = Student or Retired and *Age* = 62-80 and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *ChildrenStatus* = No and *Reservations* = 4-7 days or 6 months+ **then** Cluster 2.

Rule 10: **if** *Occupation* = Employed or Home-maker and *Reservations* = 6 months+ **then** Cluster 1.

Rule 11: **if** *Occupation* = Employed or Home-maker and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *ChildrenStatus* = Yes **then** Cluster 2.

Rule 12: **if** *Occupation* = Employed or Home-maker and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *ChildrenStatus* = No and *Reservations* = 2-4 weeks or 1 month+ and *TypeOfShelter* = Tent **then** Cluster 2.

Rule 13: **if** *Occupation* = Employed or Home-maker and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *ChildrenStatus* = No and *Reservations* = 2-4 weeks or 1 month+ and *TypeOfShelter* = RV **then** Cluster 1.

Rule 14: **if** *Occupation* = Employed or Home-maker and *Reservations* = 4-7 days, 2-4 weeks, 1 month+ or 6 months+ and *ChildrenStatus* = No and *Reservations* =4-7 days or 6 months+ **then** Cluster 2.

To receive more insight into the misclassification errors of Table 5.20 and Table 5.21, the *posterior probabilities* of both these set of rules could be investigated. Posterior probability is the probability that a hypothesis or prediction is true, calculated in the light of relevant observations (discussed in Chapter 3: 3.8.1.1.4). The posterior probabilities for (new) big spenders is shown in Table 5.22. The posterior probability table

Table 5.22: Posterior probability for (new) big spenders

Camper allocated to:	**Posterior probability**	
	Cluster 1	**Cluster 2**
Cluster 2	0.0320	0.9680
Cluster 1	0.9906	0.0094
Cluster 1	0.9955	0.0094
Cluster 1	0.9955	0.0094
Cluster 2	0.0320	0.9680

represents a sample of five randomly selected campers' probability of being misclassified. For example, the first entry in Table 5.22 indicates that when a camper identified as a (new) big spender is allocated to cluster 2, the probability that the allocation is correct is 96.8 percent, leaving 3.2 percent probability that the camper should rather be allocated to cluster 1. The posterior probability for identifying the cluster number for a loyal big spender is shown in Table 5.23. Figure 5.14 revealed that the campers' features with the

Table 5.23: Posterior probability for loyal big spenders

Camper allocated to:	**Posterior probability**	
	Cluster 1	**Cluster 2**
Cluster 1	0.9968	0.0032
Cluster 1	0.9965	0.0035
Cluster 2	0.0475	0.9525
Cluster 2	0.0475	0.9525
Cluster 1	0.9284	0.0716

most influence on the decision rules shown in Table 5.21 are: occupation, reservations, age, children status, type of shelter, marital status, annual income and gender.

When a camper has a profile as seen in Table 5.19 rule 9, Table 5.21 can be followed to allocate a cluster to her. The occupation of the camper is student or retired, thus only rules 1 to 9 are applicable, the age and children status of the camper refine the search, for the camper falls in the age range of 62-80 and indicated that does not have a child. No other information is known, therefore the most suitable rule to be applied to this camper is rule 8, which allocates the camper to cluster 1.

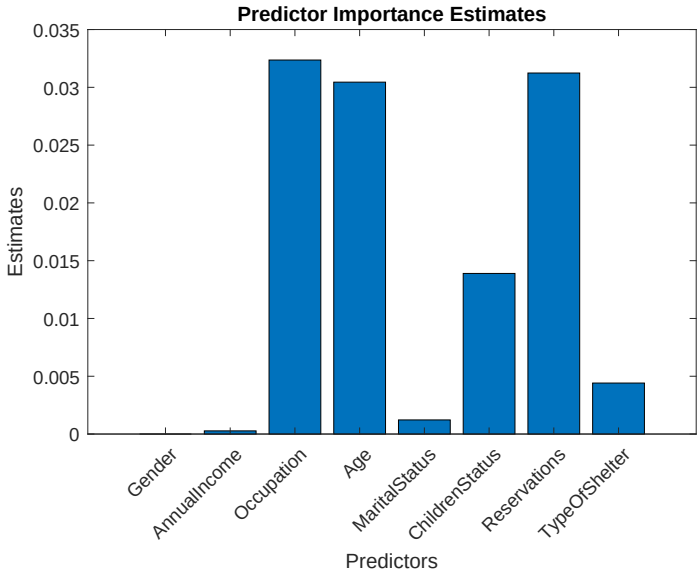


Figure 5.14: Predictor importance for loyal big spenders

Table 5.24 shows the decision rules for identifying the cluster number of the churned cheap camper. An example to allocate a churned cheap camper to a cluster is as follows. Rule 15, as shown in Table 5.19, provides a profile for such a camper. According to rule 2 in Table 5.24, cluster 3 is identified as the suitable cluster. The misclassification rate for these rules is 17.1 percent, which is higher than the previously mentioned rules. The posterior probability is shown in Table 5.25, which indicates that a churned cheap camper has a bigger probability to be allocated to cluster 3 than to cluster 2.

Table 5.24: Decision rules to identify the cluster of a camper identified as a churned cheap camper

Rule 1:	**if** *AnnualIncome* = Low **then** Cluster 3.
Rule 2:	**if** *AnnualIncome* = Medium or High and *ChildrenStatus* = Yes **then** Cluster 3.
Rule 3:	**if** *AnnualIncome* = Medium or High and *ChildrenStatus* = No and *AnnualIncome* = Medium **then** Cluster 3.
Rule 4:	**if** *AnnualIncome* = Medium or High and *ChildrenStatus* = No and *AnnualIncome* = High and *Age* = 16-32, 32-48 or 48-62 **then** Cluster 3.
Rule 5:	**if** *AnnualIncome* = Medium or High and *ChildrenStatus* = No and *AnnualIncome* = High and *Age* = 62-80 **then** Cluster 2.

Table 5.25: Posterior probability for churned cheap camper

Camper allocated to:	**Posterior probability**	
	Cluster 2	**Cluster 3**
Cluster 3	0.0000	1.0000
Cluster 3	0.3618	0.6382
Cluster 3	0.3618	0.6382
Cluster 3	0.0000	1.0000
Cluster 3	0.3618	0.6382

Figure 5.15 revealed that the campers' features with the most influence on the decision rules shown in Table 5.24 are: annual income, children status, age, reservations, marital status, type of shelter, occupation and gender.

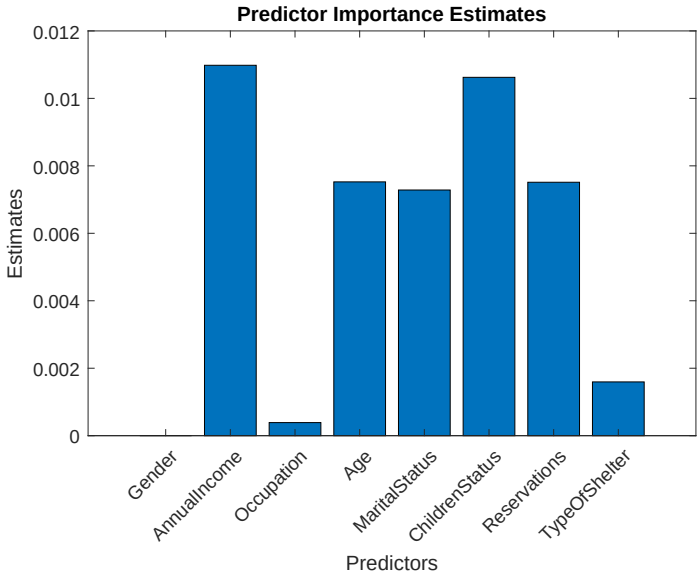


Figure 5.15: Predictor importance for churned cheap campers

This concludes the decision tree analysis performed on the camping dataset. The camper profiles that were discovered within this section provide more insight than the previously generated profiles (Chapter 4). This also validates the CSP tool, illustrating that it can generate super-profiles for a different data domain.

5.4 Validation of data simulator and CSP tool: Golfers

As a confidence building test, the researcher decided on performing another validation of the simulator as well as the CSP tool. A different domain and dataset size will be utilised. The domain that is used for validation is the *golf domain*. The golfing data is simulated according to golf participation in the United States (National Golf Foundation, 2013). The CSP tool outline, as seen in Figure 5.1, is followed. Each step will subsequently be discussed, and the results retrieved will be presented. Previously, when the CSP tool was applied on the South African demographic customer dataset, each step contained an in-depth explanation; for this section only the results of each step will be presented with their explanation.

5.4.1 Select data: Golf dataset

A dataset containing 500 000 golfers' demographic and golfing participation information was created according to the statistics provided by National Golf Foundation (2013). The dataset that will be used for the next step (2) contains the *golfer participation information*, which includes the golfers' primary key, dates participated in golfing as well as amount spent on golfing equipment. Next, the RFM analysis will be performed.

5.4.2 RFM analysis: Golf dataset

The RFM analysis will be performed on the dataset created and selected in the previous step (1). The RFM method was implemented as follows:

- Recency (R): It represents a month in the year 2013 for each golfer, in which they participated in their last game of that year.
- Frequency (F): It represents the number of times a golfer participated in a golfing game in the specified period (January 2013 – December 2013).
- Monetary (M): It represents the monetary value of purchases in the specified period for this study. The purchases only include golfing equipment, not club fees *etc.*

The RFM interval values are calculated as seen in (5.1), with the minimum and maximum values shown in Table 5.26. Algorithms 1, 2 and 3 were used to determine to which R, F and M category each golfer belongs, respectively. Table 5.27 indicates the

number, together with the percentage of golfers located in each RFM category. It can be concluded for this table that 46.04 percent of golfers have high recency value (recency category 5), the majority of the golfers (38.15 percent) belong to frequency category 2, while 58.99 percent of the golfers belong to monetary category 3. The data is now ready for the next step (3): clustering.

Table 5.26: Minimum and maximum R, F and M parameter values for the golf dataset

	Recency	Frequency	Monetary
Minimum	January 2013	1	$100
Maximum	December 2013	24	More than $15 000

Table 5.27: Summary of R, F and M parameter occurrences in `RFM` (golfers)

	R		**F**		**M**	
1	17 462	3.49%	113 056	22.61%	49 840	9.97%
2	47 829	9.57%	190 767	38.15%	49 474	9.89%
3	80 104	16.02%	94 847	18.97%	294 959	58.99%
4	124 404	24.88%	42 330	8.47%	49 970	9.99%
5	230 201	46.04%	59 000	11.80%	55 757	11.15%
Total golfers	**500 000**					

5.4.3 Clustering: Golf dataset

When performing clustering the first step is to determine the number of clusters, which is done by evaluating the silhouette values. Figure 5.16 represents the silhouette values calculated for a range of values, considering the given dataset. The figure indicates that a four-cluster solution is the most appropriate for the given dataset. Next, the k-means clustering algorithm is applied to the golf dataset, with $k = 4$. Figure 5.17 represents the scatter plot of the four clusters that were formed and Figure 5.18 indicates the different sizes of the four clusters.

Table 5.28 provides a summary of the clusters including their sizes as well as their RFM parameter information. Each cluster has M values in all the M categories (1, 2, ..., 5), however not all of the clusters have R and F values in each category. Therefore,

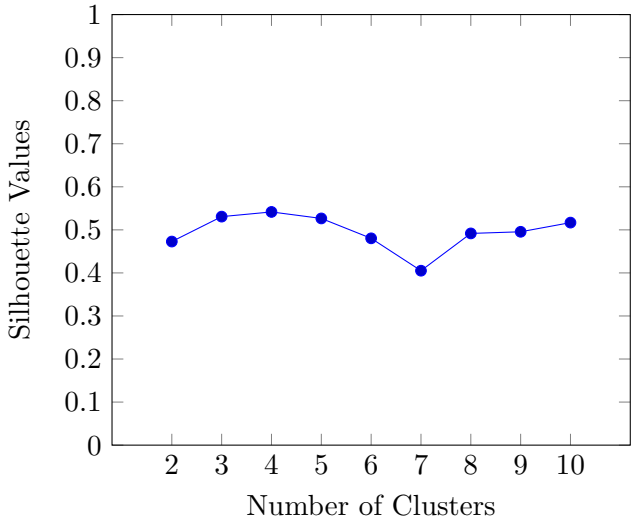


Figure 5.16: Plot of the silhouette criterion values for each number of clusters tested for the golf dataset

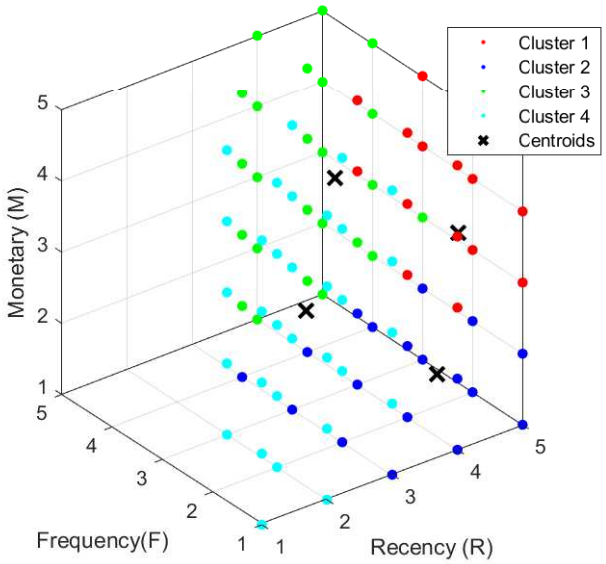


Figure 5.17: Scatter plot representing the four clusters of the golf dataset

it can be concluded that the various clusters are separated according to their R and F values; while the M values contribute to the golfers' purchasing behaviour when having different R and F values.

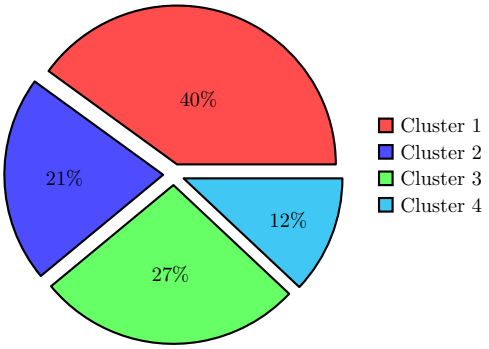


Figure 5.18: Pie chart indicating the four cluster sizes of the golf dataset

Table 5.28: Summary of the golfers and their RFM category values per cluster

	Cluster 1			Cluster 2			Cluster 3			Cluster 4		
	R	**F**	**M**	**R**	**F**	**M**	**R**	**F**	**M**	**R**	**F**	**M**
1	0%	32.73%	9.93%	0%	0%	10.05%	12.95%	34.81	10.01%	0%	0%	9.84%
2	0%	67.27%	13.81%	0%	0%	4.93%	35.47%	40.70%	10.00%	0%	0%	5.02%
3	0%	0%	58.13%	10.13%	59.36%	59.96%	51.58%	24.49%	58.96%	0%	0%	60.28%
4	36.11%	0%	10.06%	35.28%	40.64%	9.96%	0%	0%	9.92%	24.95%	0%	9.99%
5	63.89%	0%	8.07%	54.58%	0%	15.10%	0%	0%	11.10%	75.05%	100%	14.86%
	202 001			104 158			134 841			59 000		

As previously mentioned, the researcher determines each golfer's cRFM value for deeper insights into the golfers' participation behaviour. When determining the cRFM values of the golfers, it leads to discovering *types of golfers* within the dataset that are not restricted to a cluster. Clustering an RFM dataset already provides a certain structure, dividing the golfers into '*groups*' based on their golfing participation (with regards to the RFM parameters). Determining their cRFM values leads to further discovering of *golfer types*. Therefore, it can be stated that one cluster could contain more than one type of golfer, for one cluster contains more than one cRFM category value. For example, cluster 1 golfers have R values in category 4 and 5, F values in category 1 and 2 and M values in each category with the majority in category 3. This means that the golfers within this cluster participated in golfing activities very recently, however they are not frequent golfers and spent an average amount on their golfing equipment. When analysing only this cluster information, one could believe that this cluster only contains a certain type of golfer, such as disloyal golfers. However, the cRFM values provide more insight into the cluster and its golfers, for 80 percent of the golfers belong to cRFM category 3 which

is not a category disloyal golfers would belong to. There are, however, 7.99 percent of the golfers that belong to cRFM category 2, which might be classified as disloyal golfers. This indicates that each cluster could contain more than one type of golfer, even though the overall cluster has a certain group structure. Table 5.29 presents a summary of the percentage of golfers in each cRFM category.

Table 5.29: Percentage golfers in each cRFM category

	Percentage:			
	Cluster 1 (202001)	**Cluster 2 (104158)**	**Cluster 3 (134841)**	**Cluster 4 (59000)**
cRFM = 1	0%	0%	17.21%	0%
cRFM = 2	7.99%	0%	36.02%	0%
cRFM = 3	80.02%	34.69%	44.03%	2.49%
cRFM = 4	11.98%	53.43%	2.74%	27.35%
cRFM = 5	0%	11.88%	0%	70.16%

With the help of the four clusters, the cRFM category values and the RFM patterns, it is possible to discover various types of golfers present in the dataset. These types of golfers are identified and listed in Table 5.30. The table shows that an occasional golfer is characterised by visiting the golf course not too frequently, as well as not spending a lot of money on golfing equipment. The RFM patterns of the occasional golfers' cRFM categories indicate that all parameters are below the total average RFM values (derived from Table 5.28).

The most important variables that are considered when discovering the different types of golfers are the recency and frequency. It is evident from the RFM parameters that the R and F patterns differ from one type of golfer to another. The monetary values separate the clusters from one another, but are not the deciding factor for allocating a type to each golfer.

Next, the predictive model for this dataset can be developed, by using the golfers' features as response variables (as seen in Table 5.31) in order to predict the type of golfer they identify as, and identifying their cluster number.

Table 5.30: Types of golfers in dataset

Type of golfer	Golfer characteristics	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Occasional golfers	These golfers are similar to 'churned cheap customers'. These type of golfers usually belong to the low R, F and M categories. These golfers visit the golf course not too frequently and do not spend a lot of money on golfing equipment. Cluster 4 does not have occasional golfers for they do not have golfers that belong to the cRFM category values of 1 and/or 2.	cRFM = 2 R ↓ F ↓ M ↓	cRFM = 3 R ↓ F ↓ M ↓	cRFM = 1 R ↓ F ↓ M ↓ cRFM = 2 R ↓ F ↓ M ↓	
Social golfers	These golfers belong to the medium R, F and M categories. They visited the golf course more recently and more frequently than occasional golfers and spend more money on equipment. Again, clusters 2 and 3 do not have golfers that are identified as social golfers, for their frequency category values are high. The customers that belong to cluster 4 have an average frequency value equal to the total average frequency, therefore the pattern is indicated by an up and down arrow (↕).	cRFM = 3 R ↑ F ↓ M ↓			cRFM = 3 R ↓ F ↕ M ↓ cRFM = 4 R ↓ F ↕ M ↓
Core golfers	These golfers usually belong to the high R, F and M categories, for their cRFM values are generally the highest in the cluster. These golfers are very serious about the game. They participate in golfing activities the most frequently and more recently of all the golfers, as well as spending above average amounts on golfing equipment. As indicated, all of the average RFM values are higher than the total RFM values (cluster 4 customers' have an average frequency value equal to the total average frequency value).	cRFM = 4 R ↑ F ↑ M ↑	cRFM = 4 R ↑ F ↑ M ↑ cRFM = 5 R ↑ F ↑ M ↑	cRFM = 3 R ↑ F ↑ M ↑ cRFM = 4 R ↑ F ↑ M ↑	cRFM = 5 R ↑ F ↕ M ↑

Table 5.31: Golfer's features

Variable name	Explanation	Scaling
Gender	Male or Female	Categorical (Dichotomous, Figure 3.1)
Age	6-17, 18-29, 30-39, 40-49, 50-59, 60-69 or 70+	Categorical (Multichotomous, Figure 3.1)
Annual household income	Under $30 000, $30 000-$49 999, $50 000-$74 999, $75 000-$99 999, $100 000-$124 999 or More than $125 000	Categorical (Multichotomous, Figure 3.1)
Education	Non high school graduate, High school graduate, Some college or College graduate	Categorical (Multichotomous, Figure 3.1)

5.4.4 Predictive model: Golf dataset

After identifying the various types of golfers, the next step (4) is to develop the predictive model. Table 5.32 shows the decision rules that are utilised when predicting a golfer's type (*i.e.* occasional golfers, social golfers or core golfers). Figure 5.19 revealed the golfer's age is the feature with the most influence on the decision rules shown in Table 5.32.

Table 5.32: Decision rules to identify the type of golfer

Rule 1:	**if** *Age* = 6-17, 18-29, 30-39 or 40-49 **then** Occasional golfer.
Rule 2:	**if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 and *Education* = Some college and *AnnualHouseholdIncome* = Under $30 000 and *Gender* = Male **then** Core golfer.
Rule 3:	**if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 and *Education* = Some college and *AnnualHouseholdIncome* = Under $30 000 and *Gender* = Female **then** Social golfer.
Rule 4:	**if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 and *Education* = Some college and *AnnualHouseholdIncome* = $30 000-$49 999, $50 000-$74 999, $75 000-$99 999, $100 000-$124 999 or More than $125 000 **then** Core golfer.
Rule 5:	**if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 and *Education* = Non high school graduate, High school graduate or College graduate and *Gender* = Male and *AnnualHouseholdIncome* = Under $30 000, $30 000-$49 999, $50 000-$74 999, $75 000-$99 999 or $100 000-$124 999 **then** Occasional golfer.
Rule 6:	**if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 and *Education* = Non high school graduate, High school graduate or College graduate and *Gender* = Male and *AnnualHouseholdIncome* = More than $125 000 **then** Core golfer.
Rule 7:	**if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 and *Education* = Non high school graduate, High school graduate or College graduate and *Gender* = Female and *AnnualHouseholdIncome* = $50 000-$74 999 **then** Social golfers.
Rule 8:	**if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 and *Education* = Non high school graduate, High school graduate or College graduate and *Gender* = Female and *AnnualHouseholdIncome* = Under $30 000, $30 000-$49 999, $75 000-$99 999, $100 000-$124 999 or More than $125 000 **then** Occasional golfers.
Rule 9:	**if** *Age* = 50-59, 60-69 or 70+ and *Age* = 60-69 or 70+ and *Age* = 60-69 and *Gender* = Male **then** Core golfer.

Continued on next page

Rule 10:	**if** *Age* = 50-59, 60-69 or 70+ and *Age* = 60-69 or 70+ and *Age* = 60-69 and *Gender* = Female and *AnnualHouseholdIncome* = \$100 000-\$124 999 **then** Occasional golfer.
Rule 11:	**if** *Age* = 50-59, 60-69 or 70+ and *Age* = 60-69 or 70+ and *Age* = 60-69 and *Gender* = Female and *AnnualHouseholdIncome* = Under \$30 000, \$30 000-\$49 999, \$50 000-\$74 999, \$75 000-\$99 999 or More than \$125 000 **then** Core golfer.
Rule 12:	**if** *Age* = 50-59, 60-69 or 70+ and *Age* = 60-69 or 70+ and *Age* = 70+ **then** Core golfer.

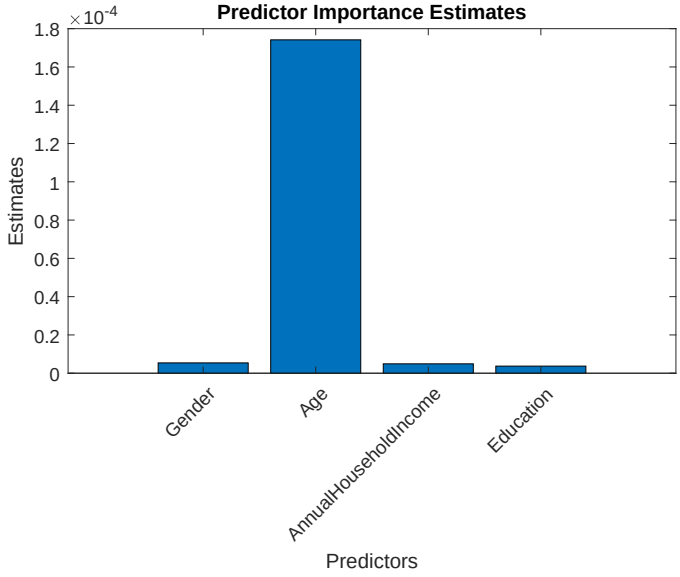


Figure 5.19: Predictor importance for identifying the type of golfer

The decision rules to identify the cluster number of an occasional golfer are shown in Table 5.33. The misclassification error for this set of rules is 18.42 percent. Figure 5.20 revealed the golfers' feature with the most influence on the decision rules shown in Table 5.33 is the golfers' age; the other features only slightly influence the predictions compared to the age factor.

Table 5.33: Decision rules to identify the cluster of a golfer identified as an occasional golfer

Rule 1:	**if** *Age* = 6-17, 18-29, 30-39 or 40-49 and *Age* = 6-17, 18-29 or 30-39 and *Gender* = Male **then** Cluster 1.

Continued on next page

Rule 2: **if** *Age* = 6-17, 18-29, 30-39 or 40-49 and *Age* = 6-17, 18-29 or 30-39 and *Gender* = Female and *Age* = 6-17 or 18-29 and *AnnualHouseholdIncome* = Under \$30 000, \$50 000-\$74 999 or \$100 000-\$124 999 **then** Cluster 1.

Rule 3: **if** *Age* = 6-17, 18-29, 30-39 or 40-49 and *Age* = 6-17, 18-29 or 30-39 and *Gender* = Female and *Age* = 6-17 or 18-29 and *AnnualHouseholdIncome* = \$30 000-\$49 999, \$75 000-\$99 999 or More than \$125 000 **then** Cluster 3.

Rule 4: **if** *Age* = 6-17, 18-29, 30-39 or 40-49 and *Age* = 6-17, 18-29 or 30-39 and *Gender* = Female and *Age* = 40-49 **then** Cluster 1.

Rule 5: **if** *Age* = 6-17, 18-29, 30-39 or 40-49 and *Age* = 40-49 and *Education* = Some college and *AnnualHouseholdIncome* = \$30 000-\$49 999, \$50 000-\$74 999 or \$100 000-\$124 999 and *Gender* = Male **then** Cluster 2.

Rule 6: **if** *Age* = 6-17, 18-29, 30-39 or 40-49 and *Age* = 40-49 and *Education* = Some college and *AnnualHouseholdIncome* = \$30 000-\$49 999, \$50 000-\$74 999 or \$100 000-\$124 999 and *Gender* = Female **then** Cluster 1.

Rule 7: **if** *Age* = 6-17, 18-29, 30-39 or 40-49 and *Age* = 40-49 and *Education* = Some college and *AnnualHouseholdIncome* = Under \$30 000, \$75 000-\$99 999 or More than \$125 000 **then** Cluster 2.

Rule 8: **if** *Age* = 6-17, 18-29, 30-39 or 40-49 and *Age* = 40-49 and *Education* = Non high school graduate, High school graduate or College graduate and *Education* = Non high school graduate **then** Cluster 1.

Rule 9: **if** *Age* = 6-17, 18-29, 30-39 or 40-49 and *Age* = 40-49 and *Education* = Non high school graduate, High school graduate or College graduate and *Education* = High school graduate or College graduate and *Gender* = Male **then** Cluster 3.

Rule 10: **if** *Age* = 6-17, 18-29, 30-39 or 40-49 and *Age* = 40-49 and *Education* = Non high school graduate, High school graduate or College graduate and *Education* = High school graduate or College graduate and *Gender* = Female **then** Cluster 1.

Rule 11: **if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 and *Gender* = Male and *Education* = Some college **then** Cluster 2.

Rule 12: **if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 and *Gender* = Male and *Education* = Non high school graduate, High school graduate or College graduate and *AnnualHouseholdIncome* = \$50 000-\$74 999 or \$100 000-\$124 999 **then** Cluster 3.

Rule 13: **if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 and *Gender* = Male and *Education* = Non high school graduate, High school graduate or College graduate and *AnnualHouseholdIncome* = Under \$30 000, \$30 000-\$49 999, \$75 000-\$99 999 or More than \$125 000 **then** Cluster 2.

Rule 14: **if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 and *Gender* = Female and *Education* = High school graduate or Some college **then** Cluster 2.

Continued on next page

Rule 15:	**if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 and *Gender* = Female and *Education* = Non high school graduate or College graduate **then** Cluster 3.
Rule 16:	**if** *Age* = 50-59, 60-69 or 70+ and *Age* = 60-69 or 70+ **then** Cluster 2.

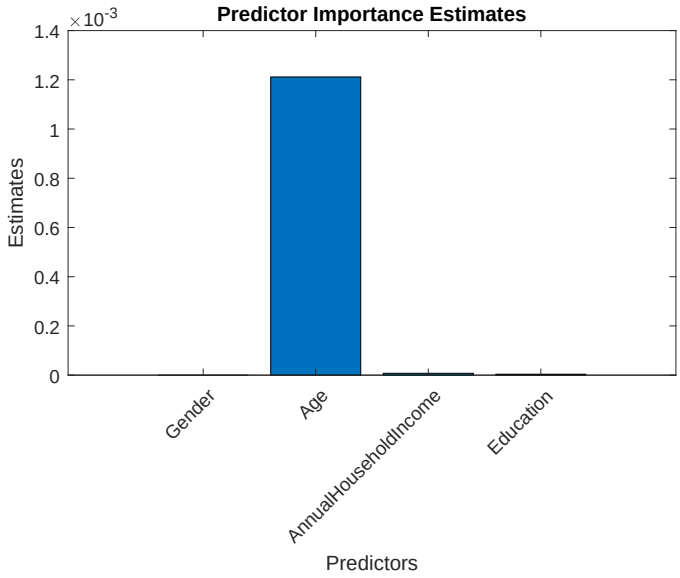


Figure 5.20: Predictor importance for occasional golfers

The decision rules to identify the cluster number of a social golfer are shown in Table 5.34. The misclassification error for this set of rules is 41.10 percent. Figure 5.21 revealed the golfers' feature with the most influence on the decision rules shown in Table 5.34 is the golfer's age. The other features (*i.e.* education, annual household income and gender) only slightly influence the predictions compared to the age factor.

Table 5.34: Decision rules to identify the cluster of a golfer identified as a social golfer

Rule 1:	**if** *Age* = 6-17, 18-29, 30-39 or 40-49 **then** Cluster 1.
Rule 2:	**if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 and *AnnualHouseholdIncome* = Under $30 000, $30 000-$49 999 or $50 000-$74 999 and *AnnualHouseholdIncome* = Under $30 000 or $30 000-$49 999 **then** Cluster 1.
Rule 3:	**if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 and *AnnualHouseholdIncome* = Under $30 000, $30 000-$49 999 or $50 000-$74 999 and *AnnualHouseholdIncome* = $50 000-$74 999 and *Education* = Non high school graduate, High school graduate or College graduate **then** Cluster 1.

Continued on next page

Rule 4: **if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 and *AnnualHouseholdIncome* = Under $30 000, $30 000-$49 999 or $50 000-$74 999 and *AnnualHouseholdIncome* = $50 000-$74 999 and *Education* = Some college **then** Cluster 4.

Rule 5: **if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 and *AnnualHouseholdIncome* = $75 000-$99 999, $100 000-$124 999 or More than $125 000 and *Gender* = Male and *Education* = Non high school graduate **then** Cluster 1.

Rule 6: **if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 and *AnnualHouseholdIncome* = $75 000-$99 999, $100 000-$124 999 or More than $125 000 and *Gender* = Male and *Education* = High school graduate, Some college or College graduate **then** Cluster 4.

Rule 7: **if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 and *AnnualHouseholdIncome* = $75 000-$99 999, $100 000-$124 999 or More than $125 000 and *Gender* = Female and *Education* = Non high school graduate, High school graduate or College graduate **then** Cluster 1.

Rule 8: **if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 and *AnnualHouseholdIncome* = $75 000-$99 999, $100 000-$124 999 or More than $125 000 and *Gender* = Female and *Education* = Some college **then** Cluster 4.

Rule 9: **if** *Age* = 60-69 or 70+ **then** Cluster 4.

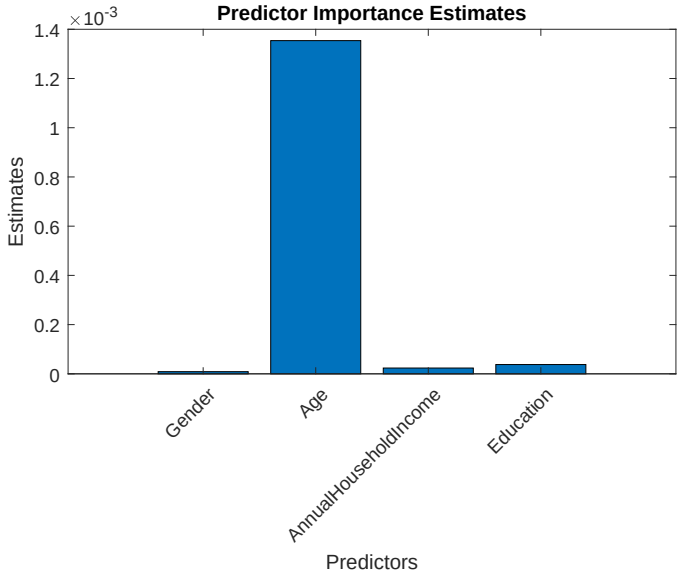


Figure 5.21: Predictor importance for social golfers

The decision rules to identify the cluster number of a core golfer are shown in Table 5.35. The misclassification error for this set of rules is 37.70 percent. Figure 5.22 revealed

the golfers' feature with the most influence on the decision rules shown in Table 5.35 is the golfer's age, followed by the annual household income, education and lastly the gender.

Table 5.35: Decision rules to identify the cluster of a golfer identified as a core golfer

Rule 1:	**if** *Age* = 6-17, 18-29, 30-39 or 40-49 and *Age* = 6-17, 18-29 or 30-39 **then** Cluster 1.
Rule 2:	**if** *Age* = 6-17, 18-29, 30-39 or 40-49 and *Age* = 40-49 and *Education* = Non high school graduate or High school graduate and *AnnualHouseholdIncome* = $50 000-$74 999 **then** Cluster 2.
Rule 3:	**if** *Age* = 6-17, 18-29, 30-39 or 40-49 and *Age* = 40-49 and *Education* = Non high school graduate or High school graduate and *AnnualHouseholdIncome* = Under $30 000, $30 000-$49 999, $75 000-$99 999, $100 000-$124 999 or More than $125 000 **then** Cluster 1.
Rule 4:	**if** *Age* = 6-17, 18-29, 30-39 or 40-49 and *Age* = 40-49 and *Education* = Some college or College graduate and *AnnualHouseholdIncome* = $75 000-$99 999 **then** Cluster 4.
Rule 5:	**if** *Age* = 6-17, 18-29, 30-39 or 40-49 and *Age* = 40-49 and *Education* = Some college or College graduate and *AnnualHouseholdIncome* = Under $30 000, $30 000-$49 999, $50 000-$74 999, $100 000-$124 999 or More than $125 000 and *Gender* = Male **then** Cluster 3.
Rule 6:	**if** *Age* = 6-17, 18-29, 30-39 or 40-49 and *Age* = 40-49 and *Education* = Some college or College graduate and *AnnualHouseholdIncome* = Under $30 000, $30 000-$49 999, $50 000-$74 999, $100 000-$124 999 or More than $125 000 and *Gender* = Female **then** Cluster 1.
Rule 7:	**if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 or 60-69 and *AnnualHouseholdIncome* = $75 000-$99 999 **then** Cluster 2.
Rule 8:	**if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 or 60-69 and *AnnualHouseholdIncome* = Under $30 000, $30 000-$49 999, $50 000-$74 999, $100 000-$124 999 or More than $125 000 and *Age* = 50-59 and *Education* = Non high school graduate **then** Cluster 3.
Rule 9:	**if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 or 60-69 and *AnnualHouseholdIncome* = Under $30 000, $30 000-$49 999, $50 000-$74 999, $100 000-$124 999 or More than $125 000 and *Age* = 50-59 and *Education* = High school graduate, Some college or College graduate **then** Cluster 4.
Rule 10:	**if** *Age* = 50-59, 60-69 or 70+ and *Age* = 50-59 or 60-69 and *AnnualHouseholdIncome* = Under $30 000, $30 000-$49 999, $50 000-$74 999, $100 000-$124 999 or More than $125 000 and *Age* = 60-69 **then** Cluster 2.

Continued on next page

Rule 11: **if** *Age* = 50-59, 60-69 or 70+ and *Age* = 70+ and *Education* = High school graduate or Some college and *AnnualHouseholdIncome* = $50 000-$74 999 **then** Cluster 4.

Rule 12: **if** *Age* = 50-59, 60-69 or 70+ and *Age* = 70+ and *Education* = High school graduate or Some college and *AnnualHouseholdIncome* = Under $30 000, $30 000-$49 999, $75 000-$99 999, $100 000-$124 999 or More than $125 000 **then** Cluster 2.

Rule 13: **if** *Age* = 50-59, 60-69 or 70+ and *Age* = 70+ and *Education* = Non high school graduate or College graduate and *AnnualHouseholdIncome* = Under $30 000, $50 000-$74 999 or $100 000-$124 999 **then** Cluster 2.

Rule 14: **if** *Age* = 50-59, 60-69 or 70+ and *Age* = 70+ and *Education* = Non high school graduate or College graduate and *AnnualHouseholdIncome* = $30 000-$49 999, $75 000-$99 999 or More than $125 000 **then** Cluster 4.

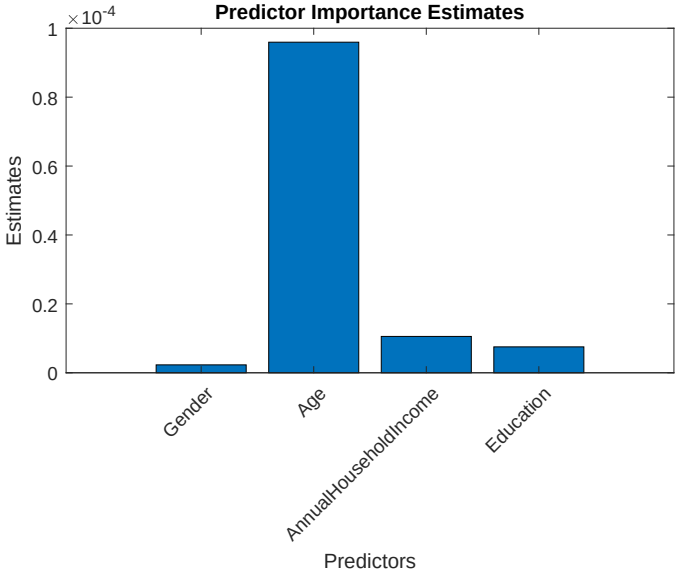


Figure 5.22: Predictor importance for core golfers

After completing this validation, the researcher made some conclusions. When building a predictive model, it is not only the size of the dataset that is important, but also the number of variables (features). When the researcher constructed the predictive model for the customer transactional (South African demographics) dataset, 50 000 customers containing 16 customer features were used. The accuracy/misclassification error of the predictive model was less than when validating the CSP tool with 500 000 golfers containing only four features. Another factor contributing to more accurate predictions, is

better defined clusters. A cluster is well defined when the silhouette value is high (higher than 0.5). This leads to easily discovering different *types* in the dataset, because each cluster and type contains specific demographic and extra value adding features; resulting in high prediction accuracy.

5.5 Validation of data simulator and CSP tool: Magazine readers

The researcher decided to perform another confidence building validation of the CSP tool with a dataset containing demographic and behavioural information regarding *magazine readers*. According to magazine readers and their reliable third-party sources, advertisements in magazines deliver a higher return on advertising spend than other media. A total of 150 000 magazine readers were simulated according to the statistics provided by MPA – The Association of Magazine Media (2017). For the purpose of this validation the following assumptions were made:

1. The simulated dataset contains information regarding the readers of three similar magazines (all containing the same magazine categories); however, each magazine has a different price.
2. The magazines contain various categories, and each reader is only assigned their favourite category.
3. The simulated data only considers readers who are not subscribed to the magazine(s).

Next, the various steps, as seen in Figure 5.1, will be performed on the magazine readers dataset.

5.5.1 Select data: Magazine dataset

The first step to be performed is to select the appropriate dataset. This dataset contains the magazine readers purchasing pattern information, which includes the readers' primary key, dates they last purchased the magazine(s), the number of magazines they have purchased within the survey period as well as the amount of money spent on the magazine(s) throughout the same period. Next, the RFM analysis will be performed on this dataset.

5.5.2 RFM analysis: Magazine dataset

The RFM method was implemented on the dataset selected in the previous step (1), as follows:

- Recency (R): It represents the most recent date on which the reader purchased the magazine(s).
- Frequency (F): It represent the number of magazines purchased by the reader within the survey period.
- Monetary (M): It represent the average amount of money spent, by the reader, on the magazine(s), within the survey period.

After calculating the RFM interval values, as seen in (5.1) and determining to which R, F and M category each reader belongs (Algorithms 1, 2 and 3); Table 5.36 indicates the number of readers located in each RFM category. Next, the well-known k-means clustering algorithm will be applied to this dimensionless RFM dataset.

Table 5.36: Summary of R, F and M parameter occurrences in `RFM` (magazine readers)

	R		**F**		**M**	
1	35 037	23.36%	21 035	14.02%	19 929	13.29%
2	24 667	16.44%	19 755	13.17%	17 565	11.71%
3	24 984	16.66%	40 006	26.67%	42 108	28.07%
4	24 704	16.47%	29 872	19.91%	31 245	20.83%
5	40 608	27.07%	39 332	26.22%	39 153	26.10%
Total magazine readers	**150 000**					

5.5.3 Clustering: Magazine dataset

As mentioned previously when performing clustering, the best cluster solution can be determined by calculating and examining a range of silhouette values. Figure 5.23 indicates that three clusters are the best solution for this dataset. Figures 5.24 and 5.25 represent the scatter plot of the three-cluster solution and the different cluster sizes for this dataset, respectively.

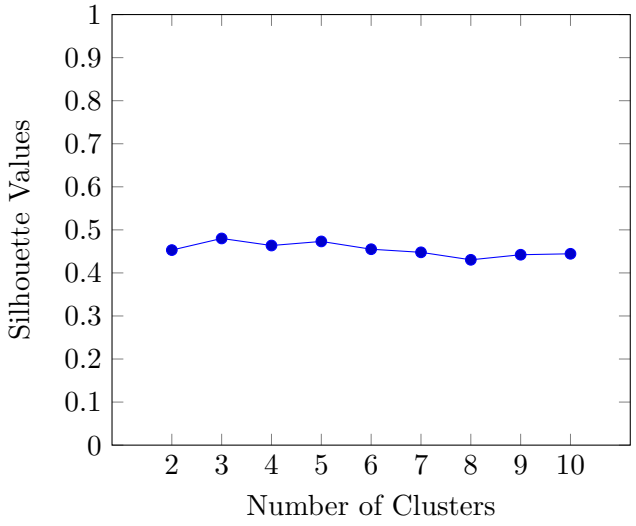


Figure 5.23: Plot of the silhouette criterion values for each number of clusters tested for the magazine dataset

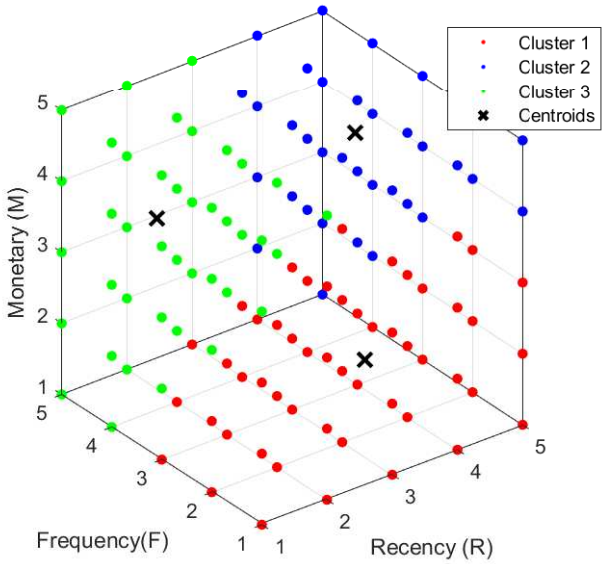


Figure 5.24: Scatter plot representing the three clusters of the magazine readers' dataset

Table 5.37 provides a summary of the clusters indicating their sizes as well as their RFM parameter information. It is evident from the table that cluster 1 contains magazine readers with different purchasing priorities, whereas it is clear that cluster 2 readers

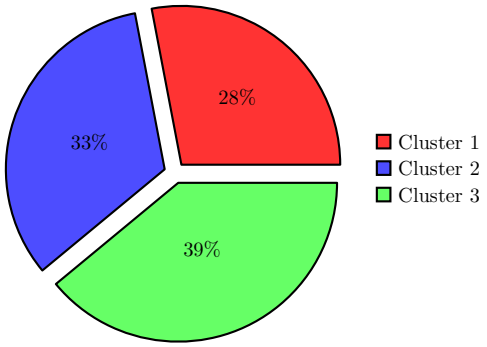


Figure 5.25: Pie chart indicating the three cluster sizes of the magazine readers' dataset

purchased the magazine(s) more recently than cluster 3 readers.

Table 5.37: Summary of the magazine readers and their RFM category values per cluster

	Cluster 1			**Cluster 2**			**Cluster 3**		
	R	**F**	**M**	**R**	**F**	**M**	**R**	**F**	**M**
1	19.03%	37.29%	41.19%	0%	8.11%	3.29%	45.79%	2.08%	1.37%
2	14.69%	26.85%	28.45%	0%	8.81%	7.32%	31.31%	6.88%	3.23%
3	27.07%	31.36%	20.41%	0%	23.84%	29.33%	22.89%	25.62%	32.57%
4	20.48%	3.46%	8.98%	32.88%	21.63%	25.77%	0%	30.40%	25.32%
5	18.73%	1.05%	0.97%	67.12%	37.60%	34.29	0%	35.02%	37.51%
	42 544			48 630			58 826		

After the RFM values have been calculated and categorised, each magazine reader's cRFM value and category can be calculated by applying Algorithm 4. Table 5.38 provides a summary of the percentage magazine readers in each cRFM category. This table,

Table 5.38: Summary of the magazine readers and their cRFM category values per cluster

	Percentage:		
	Cluster 1 (42544)	**Cluster 2 (48630)**	**Cluster 3 (58826)**
cRFM = 1	16.42%	0%	0%
cRFM = 2	39.57%	0%	11.46%
cRFM = 3	43.71%	18.62%	57.09%
cRFM = 4	0%	37.50%	27.17%
cRFM = 5	0%	43.87%	4.28%

together with the RFM patterns (as indicated in the cluster columns in Table 5.39), were used to discover the various types of magazine readers as seen in Table 5.39. For

Table 5.39: Types of magazine readers in the dataset

Type of magazine readers	**Magazine reader characteristics**	**Cluster 1**	**Clusters 2**	**Clusters 3**
Up-to-date readers	These readers belong to the high R, F and M categories, therefore their cRFM category value will be high. They are informed and avid magazine readers, who buy regularly, frequently and are likely to buy all three of the magazines. The magazine readers that belong to cluster 1 do not qualify to be identified as an up-to-date reader, for they do not have readers belonging to the cRFM category values of 4 and/or 5.		cRFM = 4 R ↑ F ↓ M ↑ cRFM = 5 R ↑ F ↑ M ↑	cRFM = 5 R ↑ F ↑ M ↑
Routine readers	These readers belong to the high F categories and medium to high R categories. The recency will either be high or low, because they recently purchased or are about to purchase, due to the high F category values. These readers are loyal to one or two magazines. They purchase the magazine(s) weekly or monthly. The magazine readers than belong to cluster 2 cannot be identified as routine readers for they do not belong to the cRFM category values of 1 and/or 2.	cRFM = 1 R ↑ F ↓ M ↓ cRFM = 2 R ↓ F ↑ M ↓		cRFM = 2 R ↓ F ↑ M ↓ cRFM = 4 R ↑ F ↑ M ↓
Cover buyers	These readers are unpredictable and a mixture, they most likely belong to the low F categories and medium to high R and M categories. These readers buy magazines when they spot something that attracts them, most likely on the cover page, or by word of mouth. They are characterised by an unstable purchasing pattern.	cRFM = 3 R ↑ F ↓ M ↑	cRFM = 3 R ↑ F ↓ M ↓	cRFM = 3 R ↑ F ↓ M ↓

example, an *up-to-date reader* is characterised by being an informed and avid magazine reader, buying the magazine(s) regularly, frequently as well as buying, at most, all three of the magazines. These readers belong to clusters 2 and 3 and their RFM patterns indicate that almost all parameter values exceed the total average RFM values. A *routine reader* is a more loyal magazine reader, and would buy at most two of the magazines

weekly or monthly, and can be found in cluster 1 and cluster 3. These customers' RFM patterns are more diverse, indicating that there is a relationship between the frequency and recency parameters. The last type of magazine reader is the *cover buyer*. These readers purchase a magazine based on the attractiveness of the cover page and little to no purchasing patterns are visible. Cover buyers are present in all three clusters.

Next, the predictive model for this dataset can be developed, by using the magazine readers' features (as seen in Table 5.40) as response variables.

Table 5.40: Magazine readers' features

Variable name	**Explanation**	**Scaling**
Gender	Male or Female	Categorical (Dichotomous, Figure 3.1)
Age	Under 18, 18-24, 25-34, 35-44, 45-54 or 55+	Categorical (Multichotomous, Figure 3.1)
Race	African American/Black, Hispanic/Latino, White, Asian, American Indian/Alaska Native, Native Hawaiian or other Pacific Islander or Other	Categorical (Multichotomous, Figure 3.1)
Annual household income	Less than \$75 000, \$75 000-\$149 000 or More than \$150 000	Categorical (Multichotomous, Figure 3.1)
Hobbies	Fishing, Restaurants Sports, Books, Travel or Pets	Categorical (Multichotomous, Figure 3.1)
Favourite magazine categories	Health and fitness, Business, Travel, Science and technology or Lifestyle	Categorical (Multichotomous, Figure 3.1)
Favourite holiday destination	Hawaii, Europe, South America, Caribbean, Florida or Mexico	Categorical (Multichotomous, Figure 3.1)

5.5.4 Predictive model: Magazine dataset

The predictive model for this dataset follows the same methodology as the previous models. Decision rules are discovered to generate customer super-profiles for each type of magazine reader. After the type is known, another set of rules is generated for each type, to identify the cluster to which the magazine reader belongs. Table 5.41 shows the generated customer super-profiles in the form of *rules*, to identify the type of magazine

reader. For example, rule 1 indicates the customer super-profiles of an *up-to-date reader*. The customer profiles indicate that the reader is in the age range of 45-54, their favourite holiday destination is Mexico, their race is either African American/Black or American Indian/Alaska Native as well as having an annual household income of $75 000-$149 000.

Table 5.41: Decision rules to identify the type of magazine reader

Rule 1:	**if** *Age* = 45-54 and *FavouriteHolidayDestination* = Mexico and *Race* = African American/Black or American Indian/Alaska Native and *AnnualHouseholdIncome* = $75 000-$149 000 **then** Up-to-date readers.
Rule 2:	**if** *Age* = 45-54 and *FavouriteHolidayDestination* = Mexico and *Race* = African American/Black or American Indian/Alaska Native and *AnnualHouseholdIncome* = Less than $75 000 or More than $150 000 **then** Cover buyers.
Rule 3:	**if** *Age* = 45-54 and *FavouriteHolidayDestination* = Mexico and *Race* = Asian, Hispanic/Latino, Native Hawaiian or other Pacific Islander, Other or White **then** Cover buyers.
Rule 4:	**if** *Age* = 45-54 and *FavouriteHolidayDestination* = Caribbean, Europe, Florida, Hawaii or South America and *FavouriteHolidayDestination* = Caribbean and *Race* = African American/Black, Asian or Hispanic/Latino and *Hobbies* = Fishing, Pets, Restaurants or Sports **then** Cover buyers.
Rule 5:	**if** *Age* = 45-54 and *FavouriteHolidayDestination* = Caribbean, Europe, Florida, Hawaii or South America and *FavouriteHolidayDestination* = Caribbean and *Race* = African American/Black, Asian or Hispanic/Latino and *Hobbies* = Books or Travel **then** Routine readers.
Rule 6:	**if** *Age* = 45-54 and *FavouriteHolidayDestination* = Caribbean, Europe, Florida, Hawaii or South America and *FavouriteHolidayDestination* = Caribbean and *Race* = American Indian/Alaska Native, Native Hawaiian or other Pacific Islander, Other or White and *Hobbies* = Books or Travel **then** Cover buyers.
Rule 7:	**if** *Age* = 45-54 and *FavouriteHolidayDestination* = Caribbean, Europe, Florida, Hawaii or South America and *FavouriteHolidayDestination* = Europe, Florida, Hawaii or South America **then** Cover buyers.
Rule 8:	**if** *Age* = Under 18, 18-24, 25-34, 35-44 or 55+ **then** Cover buyers.

Figure 5.26 indicates that the predictors with the most influence on this set of rules are the reader's race, followed by the reader's age, favourite holiday destination, hobbies, annual household income, favourite magazine category and lastly their gender.

Table 5.42 shows the decision rules utilised to identify the cluster number of an up-to-date reader. The misclassification error for this set of rules is 21.10 percent. Figure

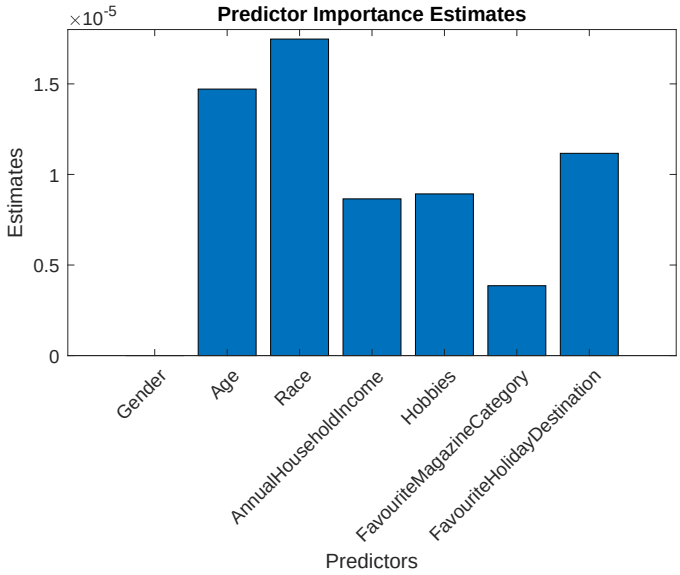


Figure 5.26: Predictor importance for type of magazine readers

5.27 revealed that the reader's features with the most influence on the decision rules, shown in Table 5.42, are the reader's hobbies, age, race, favourite holiday destination, favourite magazine category, annual household income and lastly the gender.

Table 5.42: Decision rules to identify the cluster of a magazine reader identified as an up-to-date reader

Rule 1:	**if** *Age* = 18-24 or 45-54 and *AnnualHouseholdIncome* = Less than \$75 000 and *Race* = African American/Black or American Indian/Alaska Native, Asian, Hispanic/Latino or White **then** Cluster 2.
Rule 2:	**if** *Age* = 18-24 or 45-54 and *AnnualHouseholdIncome* = Less than \$75 000 and *Race* = Other **then** Cluster 3.
Rule 3:	**if** *Age* = 18-24 or 45-54 and *AnnualHouseholdIncome* = \$75 000-\$149 000 or More than \$150 000 and *Race* = Hispanic/Latino, Native Hawaiian or other Pacific Islander or Other and *Hobbies* = Books, Pets, Restaurants, Sports or Travel and *FavouriteHolidayDestination* = Caribbean, Europe, Florida, Mexico or South America **then** Cluster 2.
Rule 4:	**if** *Age* = 18-24 or 45-54 and *AnnualHouseholdIncome* = \$75 000-\$149 000 or More than \$150 000 and *Race* = Hispanic/Latino, Native Hawaiian or other Pacific Islander or Other and *Hobbies* = Books, Pets, Restaurants, Sports or Travel and *FavouriteHolidayDestination* = Hawaii **then** Cluster 3.

Continued on next page

Rule 5: **if** *Age* = 18-24 or 45-54 and *AnnualHouseholdIncome* = $75 000-$149 000 or More than $150 000 and *Race* = Hispanic/Latino, Native Hawaiian or other Pacific Islander or Other and *Hobbies* = Fishing **then** Cluster 3.

Rule 6: **if** *Age* = 18-24 or 45-54 and *AnnualHouseholdIncome* = $75 000-$149 000 or More than $150 000 and *Race* = African American/Black, American Indian/Alaska Native, Asian or White and *Hobbies* = Fishing, Restaurants, Sports or Travel **then** Cluster 2.

Rule 7: **if** *Age* = 18-24 or 45-54 and *AnnualHouseholdIncome* = $75 000-$149 000 or More than $150 000 and *Race* = African American/Black, American Indian/Alaska Native, Asian or White and *Hobbies* = Books or Pets **then** Cluster 3.

Rule 8: **if** *Age* = Under 18, 25-34, 35-44 or 55+ and *FavouriteHolidayDestination* = Mexico and *Age* = Under 18, 25-34 or 35-44 and *Race* = African American/Black, Asian or Other and *FavouriteMagazineCategory* = Business Health and fitness, Lifestyle or Travel **then** Cluster 2.

Rule 9: **if** *Age* = Under 18, 25-34, 35-44 or 55+ and *FavouriteHolidayDestination* = Mexico and *Age* = Under 18, 25-34 or 35-44 and *Race* = African American/Black, Asian or Other and *FavouriteMagazineCategory* = Science and technology **then** Cluster 3.

Rule 10: **if** *Age* = Under 18, 25-34, 35-44 or 55+ and *FavouriteHolidayDestination* = Mexico and *Age* = Under 18, 25-34 or 35-44 and *Race* = Hispanic/Latino or White and *Hobbies* = Books, Pets or Sports **then** Cluster 2.

Rule 11: **if** *Age* = Under 18, 25-34, 35-44 or 55+ and *FavouriteHolidayDestination* = Mexico and *Age* = Under 18, 25-34 or 35-44 and *Race* = Hispanic/Latino or White and *Hobbies* = Fishing, Restaurants or Travel **then** Cluster 3.

Rule 12: **if** *Age* = Under 18, 25-34, 35-44 or 55+ and *FavouriteHolidayDestination* = Mexico and *Age* = 55+ and *Hobbies* = Fishing **then** Cluster 2.

Rule 13: **if** *Age* = Under 18, 25-34, 35-44 or 55+ and *FavouriteHolidayDestination* = Mexico and *Age* = 55+ and *Hobbies* = Books, Restaurants, Sports or Travel **then** Cluster 3.

Rule 14: **if** *Age* = Under 18, 25-34, 35-44 or 55+ and *FavouriteHolidayDestination* = Caribbean, Europe, Florida, Hawaii or South America and *Hobbies* = Pets, Restaurants or Travel and *FavouriteMagazineCategory* = Health and fitness or Science and technology and *Race* = Hispanic/Latino, Other or White **then** Cluster 2.

Continued on next page

Rule 15: **if** *Age* = Under 18, 25-34, 35-44 or 55+ and *FavouriteHolidayDestination* = Caribbean, Europe, Florida, Hawaii or South America and *Hobbies* = Pets, Restaurants or Travel and *FavouriteMagazineCategory* = Health and fitness or Science and technology and *Race* = African American/Black, Asian or Native Hawaiian or other Pacific Islander **then** Cluster 3.

Rule 16: **if** *Age* = Under 18, 25-34, 35-44 or 55+ and *FavouriteHolidayDestination* = Caribbean, Europe, Florida, Hawaii or South America and *Hobbies* = Pets, Restaurants or Travel and *FavouriteMagazineCategory* = Business, Lifestyle or Travel and *Age* = Under 18 or 25-34 **then** Cluster 2.

Rule 17: **if** *Age* = Under 18, 25-34, 35-44 or 55+ and *FavouriteHolidayDestination* = Caribbean, Europe, Florida, Hawaii or South America and *Hobbies* = Pets, Restaurants or Travel and *FavouriteMagazineCategory* = Business, Lifestyle or Travel and *Age* = 35-44 or 55+ **then** Cluster 3.

Rule 18: **if** *Age* = Under 18, 25-34, 35-44 or 55+ and *FavouriteHolidayDestination* = Caribbean, Europe, Florida, Hawaii or South America and *Hobbies* = Books, Fishing or Sports and *FavouriteMagazineCategory* = Business, Lifestyle or Travel and *Age* = 35-44 or 55+ *Race* = American Indian/Alaska Native **then** Cluster 2.

Rule 19: **if** *Age* = Under 18, 25-34, 35-44 or 55+ and *FavouriteHolidayDestination* = Caribbean, Europe, Florida, Hawaii or South America and *Hobbies* = Books, Fishing or Sports and *FavouriteMagazineCategory* = Business, Lifestyle or Travel and *Age* = 35-44 or 55+ *Race* = African American/Black, Asian, Hispanic/Latino, Other or White **then** Cluster 3.

The decision rules to identify the cluster number of a routine reader are indicated in Table 5.43. The misclassification error for this set of rules is 34.66 percent. Figure 5.28 revealed the reader's features with the most influence on decision rules, shown in Table 5.43, are as follows: the reader's age, favourite holiday destination, hobbies, race, gender, favourite magazine category and lastly their annual household income.

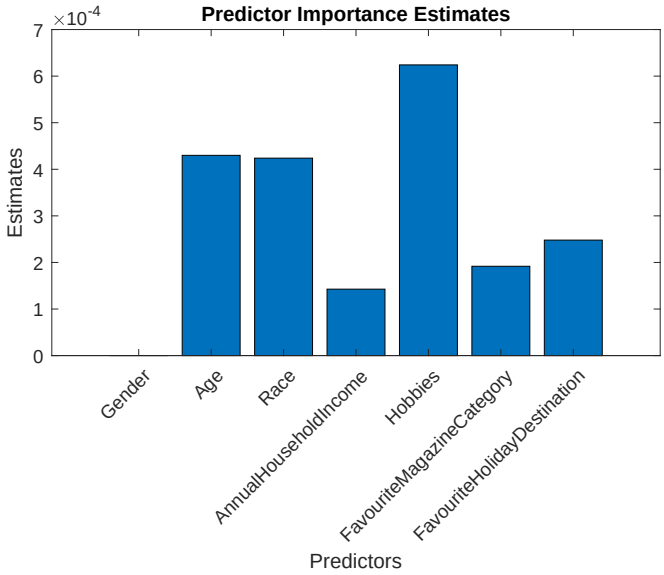


Figure 5.27: Predictor importance for up-to-date readers

Table 5.43: Decision rules to identify the cluster of a magazine reader identified as a routine reader

Rule 1:	**if** *Race* = African American/Black, American Indian/Alaska Native, Asian or Hispanic/Latino and *FavouriteMagazineCategory* = Business, Lifestyle, Science and technology or Travel and *Race* = American Indian/Alaska Native and *Hobbies* = Books, Restaurants, Sports or Travel **then** Cluster 1.
Rule 2:	**if** *Race* = African American/Black, American Indian/Alaska Native, Asian or Hispanic/Latino and *FavouriteMagazineCategory* = Business, Lifestyle, Science and technology or Travel and *Race* = American Indian/Alaska Native and *Hobbies* = Fishing or Pets and *Age* = Under 18, 25-34, 35-44 or 55+ **then** Cluster 1.
Rule 3:	**if** *Race* = African American/Black, American Indian/Alaska Native, Asian or Hispanic/Latino and *FavouriteMagazineCategory* = Business, Lifestyle, Science and technology or Travel and *Race* = American Indian/Alaska Native and *Hobbies* = Fishing or Pets and *Age* = 18-24 or 45-54 **then** Cluster 3.
Rule 4:	**if** *Race* = African American/Black, American Indian/Alaska Native, Asian or Hispanic/Latino and *FavouriteMagazineCategory* = Business, Lifestyle, Science and technology or Travel and *Race* = African American/Black, Asian or Hispanic/Latino and *Age* = 18-24, 25-34, 35-44 or 45-54 **then** Cluster 1.

Continued on next page

Rule 5: **if** *Race* = African American/Black, American Indian/Alaska Native, Asian or Hispanic/Latino and *FavouriteMagazineCategory* = Business, Lifestyle, Science and technology or Travel and *Race* = African American/Black, Asian or Hispanic/Latino and *Age* = 55+ and *FavouriteHolidayDestination* = Caribbean, Europe, Florida, Hawaii or Mexico **then** Cluster 1.

Rule 6: **if** *Race* = African American/Black, American Indian/Alaska Native, Asian or Hispanic/Latino and *FavouriteMagazineCategory* = Business, Lifestyle, Science and technology or Travel and *Race* = African American/Black, Asian or Hispanic/Latino and *Age* = 55+ and *FavouriteHolidayDestination* = South America **then** Cluster 3.

Rule 7: **if** *Race* = African American/Black, American Indian/Alaska Native, Asian or Hispanic/Latino and *FavouriteMagazineCategory* = Health and fitness and *Age* = 18-24, 45-54 or 55+ and *FavouriteHolidayDestination* = Europe or Florida **then** Cluster 1.

Rule 8: **if** *Race* = African American/Black, American Indian/Alaska Native, Asian or Hispanic/Latino and *FavouriteMagazineCategory* = Health and fitness and *Age* = 18-24, 45-54 or 55+ and *FavouriteHolidayDestination* = Caribbean, Hawaii, Mexico or South America and *Race* = African American/Black, Asian or Hispanic/Latino **then** Cluster 1.

Rule 9: **if** *Race* = African American/Black, American Indian/Alaska Native, Asian or Hispanic/Latino and *FavouriteMagazineCategory* = Health and fitness and *Age* = 18-24, 45-54 or 55+ and *FavouriteHolidayDestination* = Caribbean, Hawaii, Mexico or South America and *Race* = American Indian/Alaska Native **then** Cluster 3.

Rule 10: **if** *Race* = African American/Black, American Indian/Alaska Native, Asian or Hispanic/Latino and *FavouriteMagazineCategory* = Health and fitness and *Age* = Under 18, 25-34 or 35-44 and *FavouriteHolidayDestination* = Hawaii **then** Cluster 1.

Rule 11: **if** *Race* = African American/Black, American Indian/Alaska Native, Asian or Hispanic/Latino and *FavouriteMagazineCategory* = Health and fitness and *Age* = Under 18, 25-34 or 35-44 and *FavouriteHolidayDestination* = Caribbean, Europe, Florida, Mexico or South America **then** Cluster 3.

Rule 12: **if** *Race* = Native Hawaiian or other Pacific Islander, Other or White and *Hobbies* = Books, Fishing, Restaurants, Sports or Travel and *Gender* = Female and *Age* = 18-24, 25-34, 35-44, 45-54 or 55+ and *Hobbies* = Books, Fishing or Restaurants **then** Cluster 1.

Rule 13: **if** *Race* = Native Hawaiian or other Pacific Islander, Other or White and *Hobbies* = Books, Fishing, Restaurants, Sports or Travel and *Gender* = Female and *Age* = 18-24, 25-34, 35-44, 45-54 or 55+ and *Hobbies* = Sports or Travel **then** Cluster 3.

Continued on next page

Rule 14: **if** *Race* = Native Hawaiian or other Pacific Islander, Other or White and *Hobbies* = Books, Fishing, Restaurants, Sports or Travel and *Gender* = Female and *Age* = Under 18 **then** Cluster 1.

Rule 15: **if** *Race* = Native Hawaiian or other Pacific Islander, Other or White and *Hobbies* = Books, Fishing, Restaurants, Sports or Travel and *Gender* = Male and *Age* = 18-24, 25-34 or 35-44 **then** Cluster 1.

Rule 16: **if** *Race* = Native Hawaiian or other Pacific Islander, Other or White and *Hobbies* = Books, Fishing, Restaurants, Sports or Travel and *Gender* = Male and *Age* = Under 18, 45-54 or 55+ and *FavouriteHolidayDestination* = Caribbean, Europe, Florida or Mexico **then** Cluster 1.

Rule 17: **if** *Race* = Native Hawaiian or other Pacific Islander, Other or White and *Hobbies* = Books, Fishing, Restaurants, Sports or Travel and *Gender* = Male and *Age* = Under 18, 45-54 or 55+ and *FavouriteHolidayDestination* = Hawaii or South America **then** Cluster 3.

Rule 18: **if** *Race* = Native Hawaiian or other Pacific Islander, Other or White and *Hobbies* = Pets and *AnnualHouseholdIncome* = Less than $75 000 and *Age* = 18-24 or 35-44 **then** Cluster 1.

Rule 19: **if** *Race* = Native Hawaiian or other Pacific Islander, Other or White and *Hobbies* = Pets and *AnnualHouseholdIncome* = Less than $75 000 and *Age* = Under 18, 25-34, 45-54 or 55+ **then** Cluster 3.

Rule 20: **if** *Race* = Native Hawaiian or other Pacific Islander, Other or White and *Hobbies* = Pets and *AnnualHouseholdIncome* = $75 000-$149 000 More than $150 000 **then** Cluster 3.

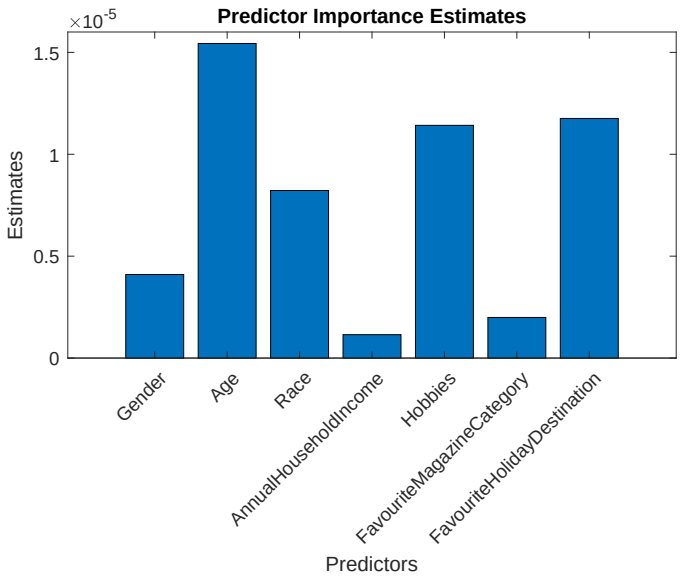


Figure 5.28: Predictor importance for routine readers

Table 5.44 displays the decision rules generated to identify the cluster number of a magazine reader identified as a cover buyer. The misclassification error is 39.71 percent. Figure 5.29 revealed that readers' features with the most influence on the decision rules shown in Table 5.44 are as follows: favourite holiday destination, age, annual household income, race, hobbies, favourite magazine category and lastly gender.

Table 5.44: Decision rules to identify the cluster of a magazine reader identified as cover buyers

Rule 1:	**if** *Race* = Other and *Age* = 25-34, 35-44, 45-54 or 55+ and *AnnualHouseholdIncome* = Less than $75 000 and *Hobbies* = Books, Fishing, Pets, Restaurants or Sports **then** Cluster 3.
Rule 2:	**if** *Race* = Other and *Age* = 25-34, 35-44, 45-54 or 55+ and *AnnualHouseholdIncome* = Less than $75 000 and *Hobbies* = Travel and *FavouriteHolidayDestination* = Caribbean or Hawaii **then** Cluster 3.
Rule 3:	**if** *Race* = Other and *Age* = 25-34, 35-44, 45-54 or 55+ and *AnnualHouseholdIncome* = Less than $75 000 and *Hobbies* = Travel and *FavouriteHolidayDestination* = Europe, Florida, Mexico or South America **then** Cluster 1.
Rule 4:	**if** *Race* = Other and *Age* = 25-34, 35-44, 45-54 or 55+ and *AnnualHouseholdIncome* = $75 000-$149 000 or More than $150 000 **then** Cluster 3.
Rule 5:	**if** *Race* = Other and *Age* = Under 18 or 18-24 and *FavouriteHolidayDestination* = Caribbean, Florida, Hawaii or Mexico **then** Cluster 3.
Rule 6:	**if** *Race* = Other and *Age* = Under 18 or 18-24 and *FavouriteHolidayDestination* = Europe or South America and *Gender* = Female **then** Cluster 3.
Rule 7:	**if** *Race* = Other and *Age* = Under 18 or 18-24 and *FavouriteHolidayDestination* = Europe or South America and *Gender* = Male and *FavouriteMagazineCategory* = Health and fitness **then** Cluster 2.
Rule 8:	**if** *Race* = Other and *Age* = Under 18 or 18-24 and *FavouriteHolidayDestination* = Europe or South America and *Gender* = Male and *FavouriteMagazineCategory* = Business, Lifestyle, Science and technology or Travel **then** Cluster 3.
Rule 9:	**if** *Race* = African American/Black, American Indian/Alaska Native, Asian, Hispanic/Latino, Native Hawaiian or other Pacific Islander or White and *Age* = 18-24, 35-44, 45-54 or 55+ **then** Cluster 3.
Rule 10:	**if** *Race* = African American/Black, American Indian/Alaska Native, Asian, Hispanic/Latino, Native Hawaiian or other Pacific Islander or White and *Age* = Under 18 or 25-34 **then** Cluster 3.

After generating these decision rules, customer super-profiles for magazine readers

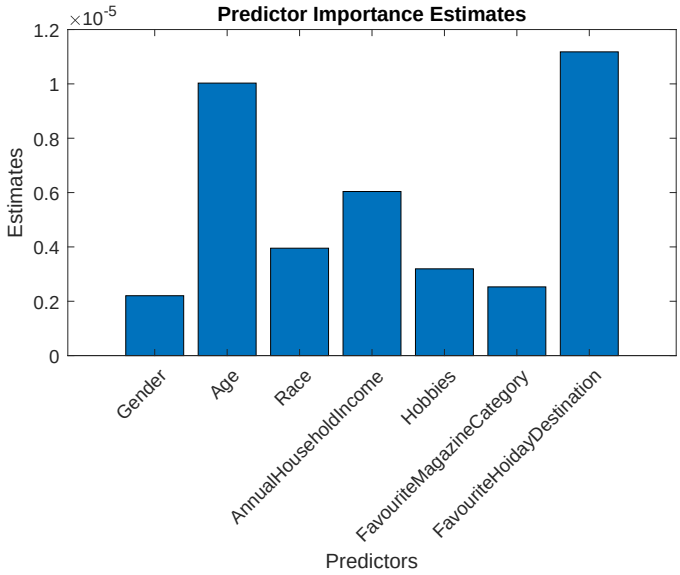


Figure 5.29: Predictor importance for cover buyers

can be discovered. The type of information utilised to generate the profiles provides a different view of the readers, and could be helpful for marketers when planning targeted marketing campaigns. For example, assume a magazine owner only wants to target a certain type of reader to subscribe to their magazine, *e.g.* routine readers because they regularly buy the magazine, they know who and how to target these readers based on the super-profiles generated. The marketers are also informed that the predictor with the most influence, with regards to routine readers, is the reader's age, and the predictor that is least important is the annual household income. Considering another example, assume a magazine runs a competition and the 'winner' gets a holiday to Mexico, they only have to target cover buyers and up-to-date readers (routine readers will not be interested in Mexico), according to the rules generated in Table 5.41. By utilising this information, regarding the readers, targeted magazine advertisements could be placed more selectively for the specific audience.

This concludes the validation of the CSP tool. Ideally, the CSP tool should be applied to many types of datasets (small, big, various domains *etc.*), however; this is not practical considering the time and other limitations of this research. Next, the researcher will briefly discuss overall findings of the developed tool, followed by a chapter summary.

5.6 Findings

After developing, testing and validating the CSP tool, several observations were made regarding input datasets and computational work. Some of the observations were anticipated, yet others were rather unexpected. Table 5.45 indicates the relative computational time of the four datasets. "Short computational time" means the time was the shortest for the given row, while "Long computational time" means the time to compute *e.g.* a dataset in Matlab was the longest relative to other data sets.

Table 5.45: Computational findings in the CSP steps for each dataset

Dataset:	**Customers**	**Camping**	**Golf**	**Magazine readers**
Size:	50 000 customers Over 36 million transactions	100 000 campers	500 000 golfers	150 000 readers
Creating dataset in Matlab	Long computational time	Short computational time	Short computational time	Short computational time
Importing from Matlab to SQL and vice versa	Long computational time	Short computational time	Short computational time	Short computational time
Performing RFM analysis	Medium computational time	Short computational time	Short computational time	Short computational time
Performing clustering (silhouette values)	Short computational time	Short computational time	Long computational time	Medium computational time
Developing predictive model (decision trees)	Short computational time	Short computational time	Short computational time	Short computational time

Another finding to be stated is that when a type of customer, camper, golfer or magazine reader is to be determined (using the cRFM values), human discretion is used. As it can be seen there are five types of customers, four types of campers and three type of golfers and magazine readers. These types depend on the datasets (domain) the cRFM values, of the dataset and the user operating the tool.

The business and marketing value of the CSP is summarised as follows:

1. Demographic and extra value adding features that traditionally distinguish groups/segments (*i.e.* gender) are not necessarily the most important predictor variable. For example, when predicting a *type of customer, camper* and *magazine reader*, gender and ethnicity (applicable only to type of customer) are among the least important predictor variables, as seen in Figures 5.6 and 5.12.

2. If the analyser doubts the results received from the CSP tool (*e.g.* the decision rules), they can still reinvestigate the results.

3. The decision rules generated in this chapter are considered as *descriptive* and not prescriptive. This means that the decision-maker can still choose to market to certain segments or types if, for example, experience suggests that. If the decision-maker has past experience that a camper's gender is important for classifying them into types or identifying a type of camper, they are still 'allowed' to create campaigns to target the campers based on their gender.

This chapter presented the incorporation and integration of RFM analysis into data mining techniques. This brought attention to the importance and advantages of using the classic RFM model in data analytics. Next, a chapter summary is presented.

5.7 Summary: Chapter 6

This chapter was initiated by developing a CSP tool, and schematically presented the broad overview of the tool in Figure 5.1. The tool contains several steps that need to be followed in order to receive the desired customer super-profiles. The CSP tool was developed and implemented on a South African demographic customer dataset containing typical monetary transactions. The dataset contained 50 000 customers and over 36 million customer transactions. After following steps 1 and 2 of the CSP tool, the transactional data were transformed into a dataset containing only the recency, frequency and monetary values of each customer. This RFM dataset was then clustered (step 3) by applying the k-means clustering method. After the clustering process and interpretation of the cluster were completed, the cRFM values were calculated. These values, together with their RFM patterns, were utilised to discover the various *types* of customers within the dataset. This concluded the unsupervised learning part of the CSP tool. Next the supervised learning was initiated. In order to develop the predictive model (indicated as step 4 in Figure 5.1), decision trees and rules were generated in Matlab. The first set of rules that emerged generated customer super-profiles for each *type of customer* present within the dataset. The rules to follow provided profiles to *identify the cluster number*, after the customer type was known. After completing this step, the CSP tool provided customer super-profiles to the user.

Various business case scenarios were created in order to illustrate the business value of the CSP tool. Next, the researcher decided to revisit the big dataset problem (camping), presented in Chapter 4, and apply the CSP tool on the camping dataset. The results received, when applying the CSP tool, provided much more depth to the customer super-profiles. After this, the researcher decided to perform two validations for building confidence. The datasets utilised for these validations were from two very different domains, different size datasets as well as having different numbers of demographic and extra value adding features. The first validation was in the domain of golfing, whereas the second validation was that of magazine readers.

After completing the validations, the researcher was confident that the CSP tool could be used in different domains. However, the researcher acknowledges that ideally the CSP tool should be tested on more domains, datasets with different sizes, datasets with various quantities of features, as well as industry data; which would need to be cleaned and/or have dimensionality reduction performed on it. Next, Chapter 7 will present the final conclusions of this research.

Chapter 6

Research summary and conclusions

The research is summarised in this chapter, and research conclusions are presented. Suggestions for future research are listed, appraisal of the research work is performed and concluding remarks are given to end this chapter.

6.1 Project summary and conclusion

The research assignment was stated in Section 1.2. The research aimed to develop a Customer Super-Profiling (CSP) tool that has the ability to identify types of customers and then predict customer profiles while using non-aggregate customer data. The researcher together with the study leader decided it would be best to simulate all the customer data that were utilised within this research to avoid ethical clearance delays (which could cause significant project delays). In order to simulate customer datasets to imitate real-life customers, the researcher had to build a data simulator that created datasets with *specific properties*. The CSP tool then utilised these datasets as '*input data*' to perform a sequence of data analytics techniques; to eventually yield customer super-profiles.

Chapter 2 initiated the research by providing a literature review regarding *segmentation and customer profiling*. This review expressed the viewpoint of an industrial engineering research project; it is not as comprehensive as would be expected of a marketing student. Segmentation is often used in conjunction with customer profiling. However,

segmentation is a term used to describe the process of dividing customers into homogeneous groups on the basis of shared or common attributes, *e.g.* habits, tastes, *etc.*, while customer profiling is describing customers by their attributes, such as *age, gender, income and lifestyles*.

Segmentation is performed on an unordered customer dataset and is the process of separating markets into groups of potential customers with similar needs and or characteristics, who are likely to exhibit similar purchasing behaviour. The researcher stated that there are two variations of segmentation: market segmentation and customer segmentation. Literature does not provide a plausible difference between market and customer segmentation. Therefore, the researcher took the view that market segmentation is generally used for a high-level strategy, whereas customer segmentation provides a more detailed view. It was attempted to indicate the different application areas of customer segmentation, such as the Recency, Frequency and Monetary (RFM) model, and to provide references to the most applicable and recent literature studies of this model, where possible. Several segmentation drawbacks were also discussed.

When discovering customer profiles for segmented datasets, it creates a portrait of the customers to help with decision-making concerning a company's products or services. Section 2.2 indicated the evolution of capturing customer information from as early as 1894. The literature indicated that there are broadly two kinds of customer profiling: demographic and behavioural profiling. Both these kinds of profiles with their distinctive variables, were discussed. In the conclusion of the literature review presented in **Chapter 2**, the researcher provided a foundation and understanding of basic marketing concepts, and presented areas for an industrial engineering researcher to explore.

Chapter 3 contained an in-depth literature review regarding *Big Data Analytics*. The chapter was initiated by first defining *data* and then *Big Data*. USMA (2017) defined and created a framework (Section 3.3) to provide an understanding of the term *Big Data Analytics*. The rest of the chapter was outlined as indicated in the framework (Figure 3.5). The data preparation phase was discussed with several data-cleaning options, this was followed by the data transformation phase, where dimensionality reduction techniques were listed with applicable references and application areas. Next, the data mining phase was discussed. USMA (2017) defined data mining as a data analytics phase containing various tools and techniques/tasks, which are collectively known as

machine learning. The researcher discussed two of the tools, namely *supervised* and *unsupervised leaning*. Supervised learning consists of two techniques/tasks: *classification* and *regression*. Both these techniques were discussed together with summaries containing most classification and regression techniques, their application areas as well as references to the most applicable research studies. Unsupervised learning together with most of the techniques were summarised in the same manner as the supervised learning techniques.

The researcher developed a high-level solution architecture (Section 4.1) for superprofiling to provide an understanding of the proposed CSP tool. An article reporting on this work, together with a toy problem and big dataset problem to illustrate the concept of discussion, was published (Walters and Bekker, 2017).

Datasets for analysis were created using a simulator implemented in Matlab, the detail design is described in **Appendix A**. An information system was created, containing all the simulated South African demographic customer information. Each customer has 16 features (demographic and extra value adding features), as well as behavioural features which indicate the retail shop(s) they visit, the date and amount spent at the retail shop(s). The data simulator was validated by creating the customer datasets containing 1 000 000 customers, populating the tables and confirming the output.

After developing the data simulator, the CSP tool was designed and built. The demonstration and validation of the CSP tool followed a deliberate path: after the development, the outline of the CSP tool (Section 5.1), which includes four steps, was followed. The demonstration of the CSP tool was performed on the customer dataset. The first step was to *select* the data that was necessary to initiate the CSP tool, which was the transactional history of the customers. This dataset was ready to continue to step 2, which is to *perform* RFM analysis on the selected dataset. After assigning each customer an RFM category value, the dataset was ready for the next step (3), which was *performing* k-means clustering. Next, the *combined* RFM (cRFM) values were deducted from the RFM category values, and served as a scoring technique in order to distinguish between customers. Each cRFM category has their own RFM pattern, which is determined when comparing the average R, F and M value (of each cRFM category per cluster) against the total average R, F and M value of each cluster. The customers' cRFM category values, together with the RFM patterns, contributed to defining the five types of customers in this dataset: (new) low spenders, (new) big spenders, low loyal

spenders, churned cheap customer and prospects. The final step to be performed when using the tool is *developing* a predictive model. The predictive model generated decision rules, by using the customer features to predict customer super-profiles for the various customer types. The profiles are referred to as *super-profiles*, because the dataset was first analysed using the RFM model, clustered into various groups, divided into types according to the cRFM values and RFM patterns and then their demographic and extra value adding features were trained and tested by decision trees to yield decision rules for predicting customer types. These rules provide the *customer super-profiles*. After getting to know the type of customer, another set of rules was used to allocate that customer (*e.g.* low loyal spender) to the most appropriate cluster. Each set of rules has their own predictor (feature) which has the most influence on those rules. These predictors were also indicated.

A few business case scenarios were designed to illustrate the business value added by utilising the CSP tool (Section 5.2). For example, not only customers within the dataset are classified into types and then targeted. New members entering the '*system*' are classified as a type of customer according to the trained decision tree and then targeted accordingly, before participating in any purchasing activities. Targeting the new customer with offers, those customers with the same or similar features are interested in, are seen as targeted marketing.

The researcher decided to revisit the big dataset problem (camping dataset), previously presented in **Chapter 4**, and perform decision tree analysis on the dataset, with the goal of generating super-profiles for the campers. Thus, the researcher only developed a predictive model for the camping dataset, because the RFM analysis and clustering had already been performed. It was concluded that the dataset contains three clusters. In order to develop the predictive model, the cRFM values of the campers had to be determined to be able to discover the type of campers. Four types of campers were discovered, *i.e.* (new) low spenders, (new) big spenders, loyal big spenders and churned cheap campers. The decision rules were generated to predict the super-profiles for each type of camper, as well as decision rules to identify the cluster number of such a camper.

Next, the researcher decided on validating the CSP tool with two other data domains: golfing and magazine readers. These additional validations were performed as confidence building tests. As indicated in Section 1.3, various data domains were utilised to test and validate the CSP tool. The validations showed that when the input data is of a good

quality and standard (demographic and extra value adding features), good predictions can be made. However, human discretion was still needed when determining the type of customers, campers, golfers and magazine readers. After the demonstration and the validations of the CSP tool the researcher is confident that the tool can be applied to various domains when utilised by an expert (knowledgeable) user. Ideally, the CSP tool should be validated and applied to many more data domains, big and/or small datasets and industry data (cleaned and transformed); however, this was not possible due to time limitations.

To summarise, the research aim and objectives set out in **Chapter 1** were achieved because:

1. The CSP tool was designed (solution architecture) to analyse input data, of various sizes, as presented in **Chapter 4**.
   (a) A data simulator, to *create big datasets*, was developed and validated, as shown in **Appendix A**.
   (b) The CSP tool has the ability to *analyse* big datasets, as demonstrated in **Appendix A** and **Chapter 5**.
2. The CSP tool, discussed in **Chapter 5**, was developed to utilise various data analytics tools and techniques, introduced in **Chapters 2** and **3**.
3. The developed CSP tool generated reliable customer super-profiles for different data domains, as demonstrated in **Chapter 5**.

Objective 1 was fulfilled as indicated by step 1 (a and b), whereas steps 2 and 3 were performed in pursuit of **Objective 2**. The summary and conclusions lead to the following suggestions for further research.

6.2 Future research

The research presented in this thesis is not complete and a few suggestions for further research are as follows:

1. Apply the CSP tool to other domains and dataset sizes. Utilising industry data should also be considered, for it will bring forth other challenges, such as data

cleaning (error and missing values) and data transformation (dimensionality reduction). Commercial users of the CSP will have to consider the *Protection of Personal Information* (PoPI) Act and privacy issues.

2. Altering the CSP tool to contain other data analytic techniques, such as k-nearest neighbour, neural networks, *etc.*

3. Creating a graphical user interface (GUI) for the CSP tool. The GUI should allow for, after the user has selected the input dataset, performing RFM and clustering automatically and output the results. This will guide the user to decide how many *types* would be appropriate for the dataset and then customer super-profiles will be generated.

4. Compare the results from this research project against results received when utilising other (machine learning) software, such as Microsoft Azure Machine Learning. The future researcher should be aware that these types of software packages usually have a cost associated with them.

5. Use or simulate customer transactional data which include retail stores (as seen in this research) as well as the *products purchased* by the customers at the retail store(s) that they visit. Apply data analytics to the dataset generating customer and *product focused* profiles, which provide *other* marketing advantages, such as cross-selling and upselling, promotional opportunities, *etc.*

6.3 Appraisal of research work

After conducting the research regarding developing a CSP tool, the researcher established the principles of Big Data Analytics. The researcher feels confident in having achieved the objective that the tool can be used in different data domains and applied to datasets of different sizes (number of rows and features). However, the researcher is aware that the CSP tool has not been used with industry data, and that the demonstration, as well as the validation, of the tool did not include data cleaning and data transformation, because the datasets were simulated.

Utilising the CSP tool requires a specific format for data input and steps to be performed in sequence, which a user should adhere to. However, the CSP tool can be

adapted and applied to various datasets when operated by a knowledgeable user. After completing this study, the researcher would recommend that when performing a study of the same nature one should consider collecting industry data and applying for ethical clearance.

The work has not yet been commercialised, although an industry partner of the researcher's home department is interested in the CSP tool. The issue of customer data privacy remains paramount, and if used for commercial purposes, it should be done with circumspection.

The researcher applied information system principles throughout this entire research, which included the interrelatedness of various aspects. Industrial engineers, with their understanding of systems and system integration as well as analytical knowledge should find these challenges that are presented exciting and relevant in our modern world.

6.4 Concluding remarks

In this final section of the project, the researcher wishes to share some reflections. The research that was conducted and documented introduced the industrial engineer to Big Data Analytics, machine learning and data science. It is evident that machine learning is used in various domains and is gaining popularity within the sports domain. Data scientists' world-wide participated in performing predictive analysis on which country would win the '*FIFA World Cup 2018*'. Various techniques, such as the ubiquitous decision trees, random forest, Bayesian interface method *etc.* were used, as well as several variables: which players are in each team, their recent performance, who they have played against and general public sentiment towards the team.

The extent to these predictions led to some data science teams to run over 2 000 000 scenarios, based on team data and individual player attributes to project-specific match scores and simulate over 1 000 000 variations of the tournament draw to calculate the *probable* winner. Other features that were used to predict the winner include: FIFA rankings, each country's population and their Gross Domestic Product (GDP), book-makers' odds, how many of the nations' team players played together in a club, the players' average age and how many Champions Leagues they have won. Considering all the data that were collected and mined to make prediction, *Goldman Sachs Group, Inc.*, learnt that past data do not always predict the future. This investment banking

group initially had Brazil, France, Germany and Portugal in the semi-finals. According to them, Brazil was supposed to win against Germany in the final (Bershidsky, 2018).

On the other hand, a South African data analytics and machine learning firm, *Principa*, who saw the FIFA World Cup 2018 as an opportunity to sharpen their skills and compare human and machine predictions, correctly predicted the winner of the World Cup, with their models out-predicting 99.96 percent of human-made predictions (Rangongo, 2018). It is *predicted* that data science enthusiasts will expand their knowledge and skill set to the Olympic Games Tokyo 2020.

Data is a resource that must be managed, since it costs money to acquire, secure and retrieve. Nowadays, data are used more and more to generate revenue and have competitive advantage, and the industrial engineer is an ideal candidate to be involved in this new drive which requires systems thinking, interfacing and analysis.

References

H. Abdi and L. J. Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010. DOI: http://dx.doi.org/10.1002/wics.101. 58

C. C. Aggarwal and C. K. Reddy. *Data clustering: Algorithms and Applications*. CRC press, 2016. ISBN 9781498785778. https://books.google.co.za/books?id=p8b1CwAAQBAJ. 99, 101, 105

H. Aguinis, R. K. Gottfredson, and H. Joo. Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, 16(2):270–301, 2013. DOI: https://doi.org/10.1177/1094428112470848. 54, 55

M. Aldenderfer and R. Blashfield. *Cluster Analysis*. SAGE Publications, Inc., 1984. ISBN 9780803923768. https://books.google.co.za/books?id=3XIqtQEACAAJ. 96, 99

E. Alpaydin. *Introduction to Machine Learning*. MIT Press, 2009. ISBN 9780262303262. https://books.google.co.za/books?id=TtrxCwAAQBAJ. 96

American Outdoor Foundation. 2014 American Camper Report. https://outdoorindustry.org/participation/outdoor-foundation-research/reports/, 2014. [Online, Accessed: 2017-09-14]. 126

E. Apeh, B. Gabrys, and A. Schierz. Customer profile classification: To adapt classifiers or to relabel customer profiles? *Neurocomputing*, 132:3–13, 2014. DOI: http://dx.doi.org/10.1016/j.neucom.2013.07.048. 1

C. Apté and S. Weiss. Data mining with decision trees and decision rules. *Future generation computer systems*, 13(2-3):197–210, 1997. DOI: https://doi.org/10.1016/S0167-739X(97)00021-6. 70, 153, 154

A. I. R. L. Azevedo. KDD, SEMMA and CRISP-DM: a parallel overview. In *IADS-DM European Conf. Data Mining*, volume 8, pages 182–185, 2008. http://hdl.handle.net/10400.22/136. 45, 46, 47, 48, 49, 50, 51

G. H. Ball and D. J. Hall. ISODATA, a novel method of data analysis and pattern classification. Technical report, Stanford research inst Menlo Park CA, 1965. http://www.dtic.mil/dtic/tr/fulltext/u2/699616.pdf. 107

D. Bates and D. Watts. *Nonlinear Regression Analysis and Its Applications*. Wiley Series in Probability and Statistics. Wiley, 2007. ISBN 9780470139004. https://books.google.co.za/books?id=m5I_AQAAIAAJ. 85, 90

D. Bell and C. Mgbemena. Data-driven agent-based exploration of customer behavior. *Simulation*, 94(3):195–212, 2018. DOI: https://doi.org/10.1177/0037549717743106. 41, 65, 70

P. D. Berger and N. I. Nasr. Customer lifetime value: Marketing models and applications. *Journal of interactive marketing*, 12(1):17–30, 1998. DOI: https://doi.org/10.1002/(SICI)1520-6653(199824)12:1<17::AID-DIR3>3.0.CO;2-K. 14

L. Bershidsky. How Goldman Sachs Lost the World Cup. https://www.bloomberg.com/view/articles/2018-07-14/world-cup-goldman-sachs-gs-model-got-it-all-wrong, July 2018. [Online, Accessed: 2018-08-07]. 220

K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "Nearest Neighbor" Meaningful? In *International conference on database theory*, pages 217–235. Springer, 1999. DOI: https://doi.org/10.1007/3-540-49257-7_15. 56

D. Birant. Data mining using RFM analysis. In *Knowledge-oriented applications in data mining*. InTech, 2011. https://cdn.intechopen.com/pdfs/13162.pdf. 147

C. Bishop. *Pattern Recognition and Machine Learning*. Information science and statistics. Springer, 2013. ISBN 9788132209065. https://books.google.co.za/books?id=HL4HrgEACAAJ. 85

J. Z. Bloom. Tourist market segmentation with linear and non-linear techniques. *Tourism Management*, 25(6):723–733, 2004. DOI: http://dx.doi.org/10.1016/j.tourman.2003.07.004. 71

I. Bose and X. Chen. Quantitative models for direct marketing: A review from systems perspective. *European Journal of Operational Research*, 195(1):1–16, 2009. DOI: https://doi.org/10.1016/j.ejor.2008.04.006. 13, 14, 18

C. Bounsaythip and E. Rinta-Runsala. Overview of Data Mining for Customer Behavior Modeling, June 2001. https://www.inf.utfsm.cl/~mcriff/Tesistas/lista-papers/customerprofiling.pdf. 100

M. Bramer. *Principles of Data Mining*. Undergraduate Topics in Computer Science. Springer London, 2007. ISBN 9781846287664. https://books.google.co.za/books?id=xVW7NslHNHsC. 52, 53, 96

B. J. Brázdil. Dimensionality reduction methods for vector spaces. Master's thesis, Masaryk University Faculty of Informatics, 2016. https://is.muni.cz/th/v9xlg/thesis.pdf?so=nx. 56, 58

L. Breiman. Random Forests. *Machine learning*, 45(1):5–32, 2001. DOI: https://doi.org/10.1023/A:1010933404324. 71

L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984. ISBN 9780412048418. https://books.google.co.za/books?id=JwQx-WOmSyQC. 70

O. Brown. The Importance of Customer Profiling. http://www.hellostarling.com/the-importance-of-customer-profiling/, March 2016. [Online, Accessed: 2018-05-21]. 1

W. Buckinx and D. Van den Poel. Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research*, 164(1):252–268, 2005. DOI: http://dx.doi.org/10.1016/j.ejor.2003.12.010. 18

P. Bühlmann, P. Drineas, M. Kane, and M. van Der Laan. *Handbook of Big Data*. A Chapman & Hall GroupHandbooks of Modern Statistical Methods. Taylor & Francis, 2016. ISBN 9781482249071. https://books.google.co.za/books?id=IbI6rgEACAAJ. 34, 36

P. Bunnak, S. Thammaboosadee, and S. Kiattisin. Applying Data Mining Techniques and Extended RFM Model in Customer Loyalty Measurement. *Journal of Advances in Information Technology Vol*, 6(4), 2015. DOI: http://dx.doi.org/10.12720/jait.6.4.238-242. 20

C. J. Burges et al. Dimension reduction: A guided tour. *Foundations and Trends® in Machine Learning*, 2(4):275–365, 2010. DOI: http://dx.doi.org/10.1561/2200000002. 56, 58

BusinessTech. These are the most and least popular loyalty reward programmes in South Africa. https://businesstech.co.za/news/business/204770/these-are-the-most-and-least-popular-loyalty-reward-programmes-in-south-africa-2/, 2017. [Online, Accessed: 2017-10-01]. 261

O. Cakir and M. E. Aras. A Recommendation Engine by Using Association Rules. *Procedia - Social and Behavioral Sciences*, 62:452–456, 2012. DOI: https://doi.org/10.1016/j.sbspro.2012.09.074. 72

T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974. DOI: https://doi.org/10.1080/03610927408827101. 106

M. A. Carreira-Perpinán. A review of dimension reduction technique. Technical report, Department of Computer Science, University of Sheffield, January 1997. http://www.pca.narod.ru/DimensionReductionBrifReview.pdf. 58

C. C. H. Chan. Online Auction Customer Segmentation Using a Neural Network Model. *International Journal of Applied Science and Engineering*, 3(2):101–109, 2005. https://www.researchgate.net/profile/Chu-chai_Chan/publication/228939097_Online_auction_customer_segmentation_using_a_neural_network_model/links/5444ef960cf2a76a3ccdc4a0/Online-auction-customer-segmentation-using-a-neural-network-model.pdf. 71

C. C. H. Chan. Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. *Expert systems with applications*, 34 (4):2754–2762, 2008. DOI: http://dx.doi.org/10.1016/j.eswa.2007.05.043. 13, 14

S. Chatterjee and A. Hadi. *Regression Analysis by Example*. Wiley Series in Probability and Statistics. Wiley, Fifth edition, May 2013. ISBN 9781118456248. https://books.google.co.za/books?id=86MCAZaY1noC. 85, 86, 91, 92

T.-H. Chen and C.-W. Chen. Application of data mining to the spatial heterogeneity of foreclosed mortgages. *Expert Systems with Applications*, 37(2):993–997, 2010. DOI: https://doi.org/10.1016/j.eswa.2009.05.076. 2

C.-H. Cheng and Y.-S. Chen. Classifying the segmentation of customer value via RFM model and RS theory. *Expert systems with applications*, 36(3):4176–4184, 2009. DOI: http://dx.doi.org/10.1016/j.eswa.2008.04.003. 15, 20

S. Chiu and D. Tavella. Chapter 7 - Introduction to Data Mining. In S. Chiu and D. Tavella, editors, *Data Mining and Market Intelligence for Optimal Marketing Returns*, pages 137–192. Butterworth-Heinemann, Boston, 2008. DOI: https://doi.org/10.1016/B978-0-7506-8234-3.00007-1. 96, 99, 100

Y. B. Cho, Y. H. Cho, and S. H. Kim. Mining changes in customer buying behavior for collaborative recommendations. *Expert Systems with Applications*, 28(2):359–369, 2005. DOI: https://doi.org/10.1016/j.eswa.2004.10.015. 59

CIE Journal. CIE journal with keywords: Big data analytics. https://www.sciencedirect.com/, 2018. [Online, Accessed: 2018-02-20]. 110

M. Corcoran. The Five Types of Analytics. http://www.informationbuilders.es/sites/www.informationbuilders.com/files/intl/co.uk/presentations/four_types_of_analytics.pdf?redir=true, 2015. [Online, Accessed: 2017-11-11]. 40

K. Coussement and D. Van den Poel. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications*, 34(1):313–327, 2008. DOI: http://dx.doi.org/10.1016/j.eswa.2006.09.038. 71, 75

J. Dean. *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners*. Wiley and SAS Business Series. John Wiley & Sons, First edition, 2014. ISBN 9781118920701. 99

J. Demšar and B. Zupan. Orange: Data Mining Fruitful and Fun - A Historical Perspective. *Informatica*, 37(1), 2013. http://www.informatica.si/ojs-2.4.3/index.php/informatica/article/viewFile/434/438. 96, 99

C. Ding and X. He. K-means Clustering via Principal Component Analysis. In *Proceedings of the Twenty-first International Conference on Machine Learning*, pages 29–36, New York, NY, USA, July 2004. ACM. DOI: https://doi.org/10.1145/1015330.1015408. 58

G. Dizdarevic. Data Fusion for Consumer Behaviour. Master's thesis, KTH Royal Institute of Technology School of Engineering Sciences, Stockholm, Sweden, 2017. http://www.diva-portal.org/smash/get/diva2:1111499/FULLTEXT01.pdf. 82

D. Dori. *Object-Process Methodology: A Holistic Systems Paradigm*. Springer Berlin Heidelberg, 2011. ISBN 9783642562099. https://books.google.co.za/books?id=rmirCAAAQBAJ. 113

R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*, volume 344. John Wiley & Sons, February 1973. ISBN 9780471223610. https://books.google.co.za/books?id=POMGRAAACAAJ. 107

A. Dursun and M. Caber. Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. *Tourism Management Perspectives*, 18: 153–160, 2016. DOI: http://dx.doi.org/10.1016/j.tmp.2016.03.001. 14, 15, 16

Effective Measure. South Arica Mobile Report 2017. http://hello.effectivemeasure.com/za-mobile-2017, 2015. [Online, Accessed: 2017-10-01]. 261

S. Erevelles, N. Fukawa, and L. Swayne. Big Data consumer analytics and the transformation of marketing. *Journal of Business Research*, 69(2):897–904, 2016. DOI: http://dx.doi.org/10.1016/j.jbusres.2015.07.001. 34, 35, 36

T. Erl, W. Khattak, and P. Buhler. *Big Data Fundamentals: Concepts, Drivers & Techniques*. The Prentice Hall Service Technology Series from Thomas Erl. Pearson Education, 2015. ISBN 9780134291208. https://books.google.co.za/books?id=tZtNCwAAQBAJ. 39, 68, 69, 84, 97

M. Evaldas. Practical use of RFM customer segmentation. https://stacktome.com/blog/rfm-customer-segmentation, December 2017. [Online, Accessed: 2018-05-14]. 151

S. Fan, R. Y. Lau, and J. L. Zhao. Demystifying Big Data Analytics for Business Intelligence Through the Llens of Marketing Mix. *Big Data Research*, 2(1):28–32, 2015. DOI: http://dx.doi.org/10.1016/j.bdr.2015.02.006. 2

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge Discovery and Data Mining: Towards a Unifying Framework site. *AAAI*, 17(3):82–88, 1996. http://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf. 43, 44

U. M. Feyyad. Data mining and knowledge discovery: making sense out of data. *IEEE Expert*, 11(5):20–25, 1996. DOI: http://dx.doi.org/10.1109/64.539013. 43, 44, 51

I. K. Fodor. A Survey of Dimension Reduction Techniques. Technical report, U.S. Department of Energy by the University of California, Lawrence Livennore National Laboratory, May 2002. https://e-reports-ext.llnl.gov/pdf/240921.pdf. 58

J. H. Friedman. Exploratory Projection Pursuit. *Journal of the American statistical association*, 82(397):249–266, December 1987. DOI: https://doi.org/10.1080/01621459.1987.10478427. 59

J. H. Friedman and J. W. Tukey. A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on computers*, 100(9):881–890, 1974. DOI: https://doi.org/10.1109/T-C.1974.224051. 59

A. R. Gallant. Nonlinear Regression. *The American Statistician*, 29(2):73–81, 1975. DOI: http://dx.doi.org/10.2307/2683268. 85, 90

A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*, volume 1. Cambridge University Press New York, NY, USA, 2006. ISBN 9781139460934. https://books.google.co.za/books?id=c9xLKzZWoZ4C. 92, 93

M. Gera and S. Goel. Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity. *International Journal of Computer Applications*, 113(18):22–29, March 2015. https://pdfs.semanticscholar.org/1c1f/ff550066be72e67f1a1efba709a4ece0489b.pdf. 68, 84, 85

R. Ghnemat and E. Jaser. Classification of Mobile Customers Behavior and Usage Patterns using Self-Organizing Neural Networks. *International Journal of Interactive Mobile Technologies*, 9(4), 2015. DOI: https://doi.org/10.3991/ijim.v9i4.4392. 59

C. Giraud-Carrier and O. Povel. Characterising Data Mining software. *Intelligent Data Analysis*, 7:181–192, 2003. http://sci2s.ugr.es/keel/pdf/specific/articulo/LloraGarrell03.pdf. 46

S. Goyat. The basis of market segmentation: a critical review of literature. *European Journal of Business and Management*, 3(9):45–54, 2011. http://www.iiste.org/Journals/index.php/EJBM/article/view/647. 10

M. Gupta and J. F. George. Toward the development of a big data analytics capability. *Information & Management*, 53(8):1049–1064, 2016. DOI: http://dx.doi.org/10.1016/j.im.2016.07.004. 34

S. H. Ha and S. C. Park. Application of data mining tools to hotel data mart on the Intranet for database marketing. *Expert Systems with Applications*, 15(1):1–31, 1998. DOI: http://dx.doi.org/10.1016/S0957-4174(98)00008-6,. 16

S. H. Ha, S. M. Bae, and S. C. Park. Customer's time-variant purchase behavior and corresponding marketing strategies: an online retailer's case. *Computers & Industrial Engineering*, 43(4):801–820, 2002. DOI: https://doi.org/10.1016/S0360-8352(02)00141-9. 59

M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of intelligent information systems*, 17(2):107–145, December 2001. DOI https://doi.org/10.1023/A:1012801612483. 100

H. Hamilton. Overview of the KDD process. http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html, June 2012. [Online, Accessed: 2017-04-15]. 43, 44

D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. A Bradford book. CogNet, 2001. ISBN 9780262082907. https://books.google.co.za/books?id=SdZ-bhVhZGYC. 65, 66

D. J. Hand. Data mining: Statistics and more? *The American Statistician*, 52(2): 112–118, 1998. DOI: https://doi.org/10.1080/00031305.1998.10480549. 59

C. Harvey. Big Data Analytics. http://www.datamation.com/big-data/big-data-analytics.html?utm_medium=email&utm_campaign=DM_NL_ITMD_20170731_STR3L2&dni=420577398&rni=8160815, 2017. [Online, Accessed: 2017-08-28]. 34

T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. *Elements*, 1:337–387, 2009. DOI: http://dx.doi.org/10.1007/b94608. 32, 71, 77, 88, 89, 92

S. Haykin. *Neural Networks and Learning Machines*. Pearson Education, 2011. ISBN 9780133002553. https://books.google.co.za/books?id=faouAAAAQBAJ. 77

J. M. Hellerstein. Quantitative Data Cleaning for Large Databases. *United Nations Economic Commission for Europe (UNECE)*, February 2008. http://db.cs.berkeley.edu/jmh/papers/cleaning-unece.pdf. 54

D. Hosmer, S. Lemeshow, and R. Sturdivant. *Applied Logistic Regression*. Wiley Series in Probability and Statistics. Wiley, 2013. ISBN 9780470582473. https://books.google.co.za/books?id=64JYAwAAQBAJ. 86, 92

M. Hosseini and M. Shabani. New approach to customer segmentation based on changes in customer value. *Journal of Marketing Analytics*, 3(3):110–121, 2015. DOI: http://dx.doi.org/10.1057/jma.2015.10. 1, 15

S. M. S. Hosseini, A. Maleki, and M. R. Gholamian. Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*, 37(7):5259–5264, 2010. DOI: https://doi.org/10.1016/j.eswa.2009.12.070. 18, 19, 147

Z. Z. Hosseini and M. Mohammadzadeh. Knowledge discovery from patients' behavior via clustering-classification algorithms based on weighted eRFM and CLV model: An empirical study in public health care services. *Iranian journal of pharmaceutical research: IJPR*, 15(1):355–367, 2016. http://ijpr.sbmu.ac.ir/pdf_1827_f7c24808a63baec4e386600af0f6dc38.html. 12

N.-C. Hsieh. An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert systems with applications*, 27(4):623–633, 2004. DOI: https://doi.org/10.1016/j.eswa.2004.06.007. 16

B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali. A comparative study of decision tree ID3 and C4.5. *International Journal of Advanced Computer Science and Applications*, (2):13–19, 2014. DOI: http://dx.doi.org/10.14569/SpecialIssue.2014.040203. 70

J.-J. Huang, G.-H. Tzeng, and C.-S. Ong. Marketing segmentation using support vector clustering. *Expert systems with applications*, 32(2):313–317, 2007. DOI: https://doi.org/10.1016/j.eswa.2005.11.028. 71, 75

P. J. Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–475, June 1985. https://www.jstor.org/stable/2241175?seq=1#page_scan_tab_contents. 59

IBM. *IBM Dictionary of Computing*. McGraw-Hill, Inc., New York, NY, USA, Tenth edition, 1993. https://dl.acm.org/citation.cfm?id=541721. 253

IIE Transactions. IIE transactions journal with keywords: Big data analytics. http://www.tandfonline.com/action/doSearch?AllField=%22Big+Data+Analytics%22&SeriesKey=uiie20, 2018. [Online, Accessed: 2018-02-20]. 110

H. Ishibuchi and T. Yamamoto. Rule weight specification in fuzzy rule-based classification systems. *IEEE transactions on fuzzy systems*, 13(4):428–435, 2005. DOI: https://doi.org/10.1109/TFUZZ.2004.841738. 72

A. J. Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. 2008. DOI: http://dx.doi.org/10.1007/978-0-387-78189-1. 71, 80, 96, 97, 98, 99, 100

S. G. Jacob and R. G. Ramani. Data Mining in Clinical Data Sets: A Review. *International Journal of Applied Information Systems (IJAIS)*, 4 (6):15–26, December 2012. https://pdfs.semanticscholar.org/2339/51bd19c3959b3a054f31596c76db68618f65.pdf. 97, 99

D. Jain and S. S. Singh. Customer lifetime value research in marketing: A review and future directions. *Journal of interactive marketing*, 16(2):34–46, 2002. DOI: https://doi.org/10.1002/dir.10032. 14

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Iintroduction to Statistical Learning: with Applications in R*, volume 103 of *Springer Texts in Statistics*. Springer Science Business Media, 2013. DOI: http://doi.org/10.1007/978-1-4614-7138-7. 88, 89, 92, 93, 94, 95, 96

S. Jansen. Customer Segmentation and Customer Profiling for a Mobile Telecommunications Company Based on Usage Behavior: A Vodafone Case Study. Master's thesis, University of Maastricht, July 2007. https://pdfs.semanticscholar.org/7a3a/688783e0424bd89f7413138bbfc24deeef8f.pdf. 9, 10, 21, 23, 24, 71, 75, 99

M.-F. Jiang, S.-S. Tseng, and S.-Y. Liao. Data Types Generalization for Data Mining Algorithms. In *IEEE SMC'99 Conference Proceedings*, volume 3, pages 928–933. IEEE, 1999. DOI: https://doi.org/10.1109/ICSMC.1999.823352. 31

H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Elsevier, third edition, 2011. ISBN 9780123814807. https://books.google.co.za/books?id=pQws07tdpjoC&dq=data+mining+concepts+and+techniques&source=gbs_navlinks_s. 71, 72, 80, 97, 99, 100

JIPE. Journal of Industrial and Production Engineering with keywords: Big data analytics. http://www.tandfonline.com/action/doSearch?AllField=%22Big+Data+Analytics%22&SeriesKey=tjci21, 2018. [Online, Accessed: 2018-02-20]. 110

I. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer New York, 2013. ISBN 9781475719048. https://books.google.co.za/books?id=-ongBwAAQBAJ. 61

I. T. Jolliffe. *Principal Component Analysis and Factor Analysis*. Springer Series in Statistics. Springer, Second edition, 2002. https://www.google.co.za/search?q=Jolliffe+I.+Principal+Component+Analysis+(2ed.,+Springer,+2002)(518s)_MVsa_.pdf&source=lnms&tbm=bks&sa=X&ved=0ahUKEwjH18-G2t3cAhVdFMAKHQucBJMQ_AUIECgB&biw=1922&bih=976. 58

J.-J. Jonker, N. Piersma, and D. Van den Poel. Joint optimization of customer segmentation and marketing policy to maximize long-term profitability. *Expert Systems with Applications*, 27(2):159–168, August 2004. DOI: https://doi.org/10.1016/j.eswa.2004.01.010. 14

J.-J. Jonker, N. Piersma, and R. Potharst. A decision support system for direct mailing decisions. *Decision support systems*, 42(2):915–925, November 2006. DOI: https://doi.org/10.1016/j.dss.2005.08.006. 16

C. Jooste, J. Strydom, A. Berndt, and P. du Plessis. *Applied Strategic Marketing*. PEARSON, Fourth edition, 2012. ISBN 9781775781271. 10, 28, 29

A. I. Kadhim, Y.-N. Cheah, and N. H. Ahamed. Text Document Preprocessing and Dimension Reduction Techniques for Text Ddocument Clustering. In *Fourth International Conference on Artificial Intelligence with Applications in Engineering and Technology*, pages 69–73. IEEE, 2014. DOI: http://doi.org/10.1109/ICAIET.2014.21. 56

R. Kahan. Using database marketing techniques to enhance your one-to-one marketing initiatives. *Journal of Consumer Marketing*, 15(5):491–493, 1998. DOI: https://doi.org/10.1108/07363769810235965. 15

A. Kalickỳ. High Performance Analytics. Master's thesis, Charles University in Prague: Department of Software Engineering, September 2013. https://is.cuni.cz/webapps/zzp/detail/139442/?lang=en. 36, 37

A. H. Karp. Using Logistic Regression to Predict Customer Retention. In *Proceedings of the Eleventh Northeast SAS Users Group Conference*, 1998. https://lexjansen.com/nesug/nesug98/solu/p095.pdf. 86, 95

L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*, volume 344 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, 2009. ISBN 9780470317488. https://books.google.co.za/books?id=YeFQHiikNo0C&dq=Finding+groups+in+data:+an+introduction+to+cluster+analysis&source=gbs_navlinks_s. 105, 106

K. Kendall and J. Kendall. *Systems Analysis and Design, Global Edition*. Pearson Education Limited, 2013. ISBN 9780273788515. https://books.google.co.za/books?id=rvyoBwAAQBAJ. 253, 255

M. Khajvand, K. Zolfaghar, S. Ashoori, and S. Alizadeh. Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science*, 3:57–63, 2011. DOI: https://doi.org/10.1016/j.procs.2010.12.011. 14, 18, 19

Khan Academy. Overview of neuron structure and function. https://www.khanacademy.org/science/biology/human-biology/neuron-nervous-system/a/overview-of-neuron-structure-and-function, September 2013. [Online, Accessed: 2018-01-19]. 78

J. Kim and C. Mueller. *Factor Analysis: Statistical Methods and Practical Issues*. Number no. 14 in A Sage university paper. SAGE Publications, 1978. ISBN 9780803911666. https://books.google.co.za/books?id=raQzQnbET9QC. 58

S.-Y. Kim, T.-S. Jung, E.-H. Suh, and H.-S. Hwang. Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert systems with applications*, 31(1):101–107, July 2006. DOI: https://doi.org/10.1016/j.eswa.2005.09.004. 14, 70

S. F. King. Citizens as customers: Exploring the future of CRM in UK local government. *Government Information Quarterly*, 24(1):47–63, January 2007. DOI: https://doi.org/10.1016/j.giq.2006.02.012. 16, 19

E. M. Knorr, R. T. Ng, and R. H. Zamar. Robust Space Transformations for Distance-based Operations. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135, 2001. DOI: http://doi.org/10.1145/502512.502532. 55

T. Kohonen. The self-organizing map. *Neurocomputing*, 21(1):1–6, November 1998. DOI: http://doi.org/10.1016/S0925-2312(98)00030-7. 59

S. B. Kotsiantis. Supervised Machine Learning: A Review of Classification Techniques. In *Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, pages 3–24, Amsterdam, The Netherlands, The Netherlands, 2007. IOS Press. http://dl.acm.org/citation.cfm?id=1566770.1566773. 67, 69, 70, 71, 72, 73, 75, 77, 81

D. Kroenke. *MIS Essentials*. Pearson Education, Fourth edition, 2014. ISBN 9780133546811. https://books.google.co.za/books?id=1SOvAgAAQBAJ. 251

R. Kuo, Y. An, H. Wang, and W. Chung. Integration of self-organizing feature maps neural network and genetic K-means algorithm for market segmentation. *Expert Systems with Applications*, 30(2):313–324, February 2006. DOI: http://dx.doi.org/10.1016/j.eswa.2005.07.036. 71, 99

A. Kurniawan. *Database Programming Using Matlab:*. PE Press, First edition, 2013. https://books.google.co.za/books?id=rWiVAgAAQBAJ. 261

R. Lanjewar and O. P. Yadav. Understanding of Customer Profiling and Segmentation Using K-Means Clustering Method for Raipur Sahkari Dugdh Sangh Milk Products. *International Journal of Research in Computer and Communication Technology (IJRCCT)*, 2(3):103–107, 2013. http://www.ijrcct.org/index.php/ojs/article/view/189. 99

D. Larose and C. Larose. *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley Series on Methods and Applications in Data Mining. Wiley, Second edition, 2014. ISBN 9781118873588. https://books.google.co.za/books?id=nZWtAwAAQBAJ&source=gbs_similarbooks. 70, 72, 73, 74, 82, 83

N. Larsen. Market Segmentation - A Framework for Determining the Right Target Customers. Master's thesis, Aarhus School of Business, May 2010. http://pure.au.dk/portal/files/11462/ba.pdf. 26

R. L. Lawrence and A. Wright. Rule-Based Classification Systems Using Classification and Regression Tree (CART) Analysis. *Photogrammetric engineering and remote sensing*, 67(10):1137–1142, 2001. https://pdfs.semanticscholar.org/0a3c/ec5d1c1c1ba8ee2ce44dcae1b2bcc93c5519.pdf. 72

D. Lee. Facebook's security boss to leave firm. https://www.bbc.com/news/technology-45040289?intlink_from_url=https://www.bbc.com/news/topics/c81zyn0888lt/facebook-cambridge-analytica-data-scandal&link_location=live-reporting-correspondent, August 2018. [Online, Accessed: 2018-08-04]. 3

S. C. Lee, Y. H. Suh, J. K. Kim, and K. J. Lee. A cross-national market segmentation of online game industry using SOM. *Expert systems with applications*, 27(4):559–570, November 2004. DOI: https://doi.org/10.1016/j.eswa.2004.06.001. 59

P. Lehohla. South African Satistics, 2012. http://www.statssa.gov.za/publications/SAStatistics/SAStatistics2012.pdf, 2016. [Online, Accessed: 2017-10-01]. 260

K. Lekdee and L. Ingsrisawang. The Empirical Distribution of Wald, Score, Likelihood Ratio, Hosmer–Lemeshow (HL), and Deviance for a Small Sample Logistic Regression Model. In *Proceedings of the International MultiConference of Engineering and Computer Scientists*, volume 3, March 2010. http://www.iaeng.org/publication/IMECS2010/IMECS2010_pp2062-2065.pdf. 92

R. Li. Top 10 Data Mining Algorithms, Explained. http://www.kdnuggets.com/2015/05/top-10-data-mining-algorithms-explained.html, May 2015. [Online, Accessed: 2017-05-16]. 72, 80, 81, 83

S.-T. Li, L.-Y. Shue, and S.-F. Lee. Business intelligence approach to supporting strategy-making of ISP service management. *Expert Systems with Applications*, 35 (3):739–754, 2008. DOI: https://doi.org/10.1016/j.eswa.2007.07.049. 16

Y.-M. Li, C.-H. Lin, and C.-Y. Lai. Identifying influential reviewers for word-of-mouth marketing. *Electronic Commerce Research and Applications*, 9(4):294–304, 2010. DOI: https://doi.org/10.1016/j.elerap.2010.02.004. 16

G. Linoff and M. Berry. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. IT Pro. Wiley, Third edition, 2011. ISBN 9781118087459. https://books.google.co.za/books?id=AyQfVTDJypUC. 71

Y. Liu, M. Kiang, and M. Brusco. A unified framework for market segmentation and its applications. *Expert Systems with Applications*, 39(11):10292–10302, 2012. DOI: https://doi.org/10.1016/j.eswa.2012.02.161. 9, 10, 12, 13

M. Lombard, B. Cameron, M. Mokonyama, and A. Shaw. Report on Trends in Passenger Transport in South Africa. https://www.dbsa.org/, 2017. [Online, Accessed: 2017-10-01]. 261

S.-A. Lumsden, S. Beldona, and A. M. Morrison. Customer Value in an All-Inclusive Travel Vacation Club: An Application of the RFM Framework. *Journal of Hospitality & Leisure Marketing*, 16(3):270–285, 2008. DOI: https://doi.org/10.1080/10507050801946858. 16

J. Luna. Predicting student retention and academic success at new mexico tech. Master's thesis, New Mexico Institute of Mining and Technology, August 2000. http://euler.nmt.edu/~brian/students/julie.pdf. 92

B. Lyle. A Brief History of Customer Relationship Management. http://www.business2community.com/business-innovation/brief-history-customer-relationship-management-01245936#pzJtoZjIrxcHAcHo.97, 2015. 22

M. Lynn. *Segmenting and Targeting Your Market: Strategies and Limitations*. John Wiley & Sons, Inc., 2012. DOI: http://dx.doi.org/10.1002/9781119200901.ch23. 29, 30

T. S. Madhulatha. Comparison between K-Means and K-Medoids Clustering Algorithms. *Advances in Computing and Information Technology*, pages 472–481, 2011. DOI: https://doi.org/10.1007/978-3-642-22555-0_48. 96, 99, 100

G. Mariscal, Ó. Marbán, and C. Fernández. A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, 25

(2):137–166, Month 2010. DOI: http://dx.doi.org/10.1017/S0269888910000032. 43, 44, 45, 46, 47

S. Marsland. *Machine Learning: An Algorithmic Perspective*. Machine Learning & Pattern Recognition Series. CRC press, Second edition, 2015. ISBN 9781498759786. https://books.google.co.za/books?id=y_oYCwAAQBAJ&dq=Machine+learning:+an+algorithmic+perspective&source=gbs_navlinks_s. 78, 79

W. Martinez, A. Martinez, A. Martinez, and J. Solka. *Exploratory Data Analysis with MATLAB, Second Edition*. Chapman & Hall/CRC Computer Science & Data Analysis. CRC Press, Second edition, 2010. ISBN 9781439812211. https://books.google.co.za/books?id=_J3MBQAAQBAJ. 119, 126, 143

MathWorks. Microsoft SQL Server ODBC for Windows. https://www.mathworks.com/help/database/ug/microsoft-sql-server-odbc-windows.html, 2018a. [Online, Accessed: 2018-03-12]. 261

MathWorks. Exchange data with relational and nonrelational databases. https://www.mathworks.com/products/database.html, 2018b. [Online, Accessed: 2018-03-06]. 99, 261

J. A. McCarty and M. Hastak. Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research*, 60(6):656–662, June 2007. DOI: https://doi.org/10.1016/j.jbusres.2006.06.015. 147

W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943. DOI: https://doi.org/10.1007/BF02478259. 78

J. Miglautsch. Application of RFM principles: What to do with 1–1–1 customers? *Journal of Database Marketing & Customer Strategy Management*, 9(4):319–324, 2002. DOI: https://doi.org/10.1057/palgrave.jdm.3240080. 17

S.-H. Min and I. Han. Detection of the customer time-variant pattern for improving recommender systems. *Expert Systems with Applications*, 28(2):189–199, 2005. DOI: https://doi.org/10.1016/j.eswa.2004.10.001. 59

M. Minelli, M. Chambers, and A. Dhiraj. *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*. Wiley CIO. Wiley, 2012. ISBN 9781118239155. https://books.google.co.za/books?id=Mg3WvT8uHV4C. 40, 41

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning series. The MIT Press, 2012. ISBN 9780262018258. https://books.google.co.za/books?id=maz6AQAAQBAJ&dq=Foundations+of+Machine+Learning&source=gbs_navlinks_s. 84

R. Mojena. Hierarchical grouping methods and stopping rules: An evaluation. *The Computer Journal*, 20(4):359–363, January 1977. DOI: https://doi.org/10.1093/comjnl/20.4.359. 107

D. Montgomery, E. Peck, and G. Vining. *Introduction to Linear Regression Analysis*. Wiley Series in Probability and Statistics. John Wiley Sons, 2015. ISBN 9781119180173. https://books.google.co.za/books?id=27kOCgAAQBAJ. 86, 92

D. Moore, G. MacCabe, and B. Craig. *Introduction to the Practice of Statistics*. Introduction to the Practice of Statistics. W.H. Freeman and Company, Sixth edition, 2009. ISBN 9781429216227. https://books.google.co.za/books?id=FiZ7SgAACAAJ. 54

J. Mowen and M. Minor. *Consumer Behavior*. Prentice-Hall, Fifth edition, 1998. ISBN 9780137371150. https://books.google.co.za/books?id=_SLZPfe0oDkC. 26

MPA – The Association of Magazine Media. Magazine Media Factbook 2016/17. http://www.magazine.org/sites/default/files/MPA-FACTbook201617-ff.pdf, 2017. [Online, Accessed: 2018-07-03]. 195

H. Müller and U. Hamm. Stability of market segmentation with cluster analysis–A methodological approach. *Food Quality and Preference*, 34:70–78, June 2014. DOI: https://doi.org/10.1016/j.foodqual.2013.12.004. 9

K. Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning. MIT Press, 2012. ISBN 9780262018029. https://books.google.co.za/books?id=NZP6AQAAQBAJ. 67, 71, 83

D. Napoleon and S. Pavalakodi. A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Set. *International Journal of Computer Applications*, 13(7):41–46, January 2011. https://pdfs.semanticscholar.org/f555/69a0a7484b6086185f8c4119b8246e5da4da.pdf. 56, 58, 99

National Golf Foundation. Golf Participation in the United States. http://leisurepropertiesgroup.com/wp-content/uploads/2014/08/NGF-Golf-Participation-in-the-US-2013-Edition.pdf, 2013. [Online, Accessed: 2018-06-26]. 182

NDA National Debit Advisors. What is a good credit score in South Africa? https://nationaldebtadvisors.co.za/what-is-a-good-credit-score-in-south-africa/, October 2016. [Online, Accessed: 2018-08-04]. 147

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, February 2004. DOI: http://doi.org/10.1103/PhysRevE.69.026113. 107

E. Ngai, L. Xiu, and D. Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36(2, Part 2):2592–2602, 2009. DOI: http://doi.org/10.1016/j.eswa.2008.02.021. 59

T. A. Nguyen. CUSTOMER SEGMENTATION: A GUIDE TO THE BEST B2B PRACTICES. http://labs.openviewpartners.com/customer-segmentation/#.WMe5nm-GNpg, September 2016. [Online, Accessed: 2017-03-14]. 10

D. D. Nimbalkar and P. Shah. Data mining using RFM analysis. *International Journal of Scientific & Engineering Research (IJSRE)*, 4(12):940–943, December 2013. https://www.ijser.org/researchpaper/Data-mining-using-RFM-Analysis.pdf. 105, 153

R. Nisbet, G. Miner, and K. Yale. *Handbook of Statistical Analysis and Data Mining Applications*. Elsevier Science, Second edition, 2017. ISBN 9780124166455. https://books.google.co.za/books?id=QVgXAwAAQBAJ. 51, 53, 54, 55

F. Ntawanga, A. Calitz, and L. Barnard. A Customer Profile Model Using A Service-Oriented Architecture. 2010. http://www.satnac.org.za/proceedings/2009/papers/software/Paper50.pdf. 23

Oxford University Press. *Definition of data in English*. 2017. [Online, Accessed: 2017-11-20]. 31

M. Palamuleni, I. Kalule-Sabiti, and M. Makiwane. Fertility and childbearing in South Africa. http://www.hsrc.ac.za/uploads/pageContent/1655/FertilityandchildbearinginSouthAfrica.pdf, 2007. [Online, Accessed: 2017-10-01]. 260

M. Paliwal and U. A. Kumar. A study of academic performance of business school graduates using neural network and statistical techniques. *Expert Systems with Applications*, 36(4):7865–7872, May 2009a. DOI: http://dx.doi.org/10.1016/j.eswa.2008.11.003. 83, 85

M. Paliwal and U. A. Kumar. Neural networks and statistical techniques: A review of applications. *Expert Systems with Applications*, 36(1):2–17, January 2009b. DOI: http://dx.doi.org/10.1016/j.eswa.2007.10.005. 71

V. Paramasivam, T. S. Yee, S. K. Dhillon, and A. S. Sidhu. A methodological review of data mining techniques in predictive medicine: An application in hemodynamic prediction for abdominal aortic aneurysm disease. *Biocybernetics and Biomedical Engineering*, 34(3):139–145, 2014. DOI: http://dx.doi.org/10.1016/j.bbe.2014.03.003. 70, 73

K. Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, June 2010. https://www.tandfonline.com/doi/abs/10.1080/14786440109462720?journalCode=tphm17. 60

N. Petroulakis and A. Miaoudakis. An Application of Neural Networks in Market Segmentation. *In Proc of 2nd Pan-Hellenic Conference in New Technologies and Marketing (NTM2007)*, 2007. https://pdfs.semanticscholar.org/b187/cb3e519377e548e38d91c29256e792e96d9d.pdf. 71

L. Pierson and J. Porway. *Data Science For Dummies*. Wiley, 2017. ISBN 9781119327653. https://books.google.co.za/books?id=o2ovDgAAQBAJ. 96, 99, 100

D. M. Powers. Evaluation: from precision, recall and F-measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies*, 2:37–63, 2011. http://www.bioinfo.in/contents.php?id=51. 95

Y. Prasad. *Big Data Analytics Made Easy*. Notion Press, 2016. ISBN 9781946390721. https://books.google.co.za/books?id=43q2DQAAQBAJ. 34, 35, 36, 37, 38

D. Pratiwi. The Use of Self Organizing Map Method and Feature Selection in Image Database Classification System. *IJCSI International Journal of Computer Science Issues*, 9(2), May 2012. https://arxiv.org/ftp/arxiv/papers/1206/1206.0104.pdf. 59

F. Provost and T. Fawcett. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media, 2013. ISBN 9781449374280. https://books.google.co.za/books?id=4ZctAAAAQBAJ. 66, 67, 96

G. Punj and D. W. Stewart. Cluster Analysis in Marketing Research: Review and Suggestions for Application. *Journal of marketing research*, pages 134–148, May 1983. DOI: http://dx.doi.org/10.2307/3151680. 13

J. Quinlan. *C4.5: Programs for Machine Learning*. Ebrary online. Elsevier Science, 2014. ISBN 9780080500584. https://books.google.co.za/books?id=b3ujBQAAQBAJ. 70

K. Raines. Factsheet : Customer Profiling and Segmentation Tools. http://asimetrica.org/wp-content/uploads/2014/08/FACTSHEET-CUSTOMER-PROFILING-AND-SEGMENTATION-TOOLS.pdf, 2009. [Online, Accessed: 2017-05-12]. 25

A. Rajarajeswari and R. M. Ravindran. A Comparative Study of K-means K-Medoid And Enhanced K-Medoid Algorithms. *International Journal of Advance Foundation and Research in Computer (IJAFRC)*, 2(8):7–10, 2015. https://pdfs.semanticscholar.org/6854/e0d6554fefaa69d561e4133dc7149d33606d.pdf. 99

V. Rajaraman. Big data analytics. *Resonance*, 21(8):695–716, August 2016. DOI: http://dx.doi.org/10.1007/s12045-016-0376-7. 35, 38

T. Rangongo. These SA data scientists have had great success with predicting World Cup scores - and they believe they know who will win the tournament. https://www.businessinsider.co.za/these-sa-data-scientists-say-france-will-win-the-2018-fifa-world-cup-2-0-against-croatia-based-on-this-method-2018-7, July 2018. [Online, Accessed: 2018-07-20]. 220

M. D. Rechenthin. *Machine-learning classification techniques for the analysis and prediction of high-frequency stock direction*. Doctor of philosophy, University of Iowa, May 2014. https://ir.uiowa.edu/cgi/viewcontent.cgi?article=5248&context=etd. 67, 68, 71, 72, 75, 76, 80, 81, 99

C. Reni. Customer Success: A Brief History. https://customergauge.com/news/a-brief-history-of-customer-success/, 2017. [Online, Accessed: 2017-03-13]. 22

R. Riffenburgh. *Statistics in Medicine*. Elsevier Science, 2011. ISBN 9780080541747. https://books.google.co.za/books?id=zoipeXzsA7IC. 85, 86, 91, 94

L. Rokach and O. Maimon. *Data Mining with Decision Trees: Theory and Applications*. World Scientific Publishing Company, 2014. ISBN 9789814590099. https://books.google.co.za/books?id=0VYCCwAAQBAJ. 67, 70, 73

L. Romdhane, N. Fadhel, and B. Ayeb. An efficient approach for building customer profiles from business data. *Expert Systems with Applications*, 37(2):1573–1585, March 2010. DOI: http://dx.doi.org/10.1016/j.eswa.2009.06.050. 24

D. F. Ross. E-CRM from a supply chain management perspective. *Information Systems Management*, 22(1):37–44, 2005. DOI: https://doi.org/10.1201/1078/44912.22.1.20051201/85737.5. 99

S. M. Ross. *Simulation*. Elsevier, 2013. ISBN 9780124158252. 259

M. Rouse. Definition: raw data (source data or atomic data). http://searchdatamanagement.techtarget.com/definition/raw-data, February 2009. [Online, Accessed: 2017-11-20]. 31, 37

A. Ruckstuhl. Introduction to nonlinear regression. *IDP Institut fur Datenanalyse und Prozessdesign, Zurcher Hochschule fur Angewandte Wissenschaften*, October 2010. https://pdfs.semanticscholar.org/8fa1/3fead47cc6ecf3d27de9e682dcef36c77502.pdf. 85, 91

P. Russom. Big data analytics. Technical report, TDWI Research, 2011. https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf. 35, 38

SAJIE. SAJIE with keywords: Big data analytics. http://sajie.journals.ac.za/pub/search/search?simpleQuery=Big+Data+Analytics&searchField=query, 2018. [Online, Accessed: 2018-02-20]. 110

M. T. Salazar, T. Harrison, and J. Ansell. An approach for the identification of cross-sell and up-sell opportunities using a financial services customer database. *Journal of Financial Services Marketing*, 12(2):115–131, November 2007. https://doi.org/10.1057/palgrave.fsm.4760066. 99, 143

N. Salkind. *Encyclopedia of Measurement and Statistics*. A Sage reference publication. SAGE Publications, 2007. ISBN 9781412916110. https://books.google.co.za/books?id=dqc5DQAAQBAJ. 72, 81, 83, 84, 85, 86, 87, 89, 92, 93, 94

N. Salkind. *Encyclopedia of Research Design*. SAGE Publications, 2010. ISBN 9781412961271. https://books.google.co.za/books?id=pvo1SauGirsC. 51, 52, 53

P. A. Sarvari, A. Ustundag, and H. Takci. Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. *Kybernetes*, 45 (7):1129–1157, 2016. DOI: https://doi-org.ez.sun.ac.za/10.1108/K-07-2015-0180. 14, 15, 16, 20, 118

U. Shafique and H. Qaiser. A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, 12(1):217–222, November 2014. http://www.ijisr.issr-journals.org/abstract.php?article=IJISR-14-281-04. 46

C. Shalizi. Advanced data analysis from an elementary point of view, January 2017. http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/ADAfaEPoV.pdf. 92

M. Sharma. Data mining: A literature survey. *International Journal of Emerging Research in Management & Technology*, 3(2), 2014. ISSN: 2278-9359. 65, 66

M. J. Shaw, C. Subramaniam, G. W. Tan, and M. E. Welge. Knowledge management and data mining for marketing. *Decision support systems*, 31(1):127–137, May 2001. DOI: https://doi.org/10.1016/S0167-9236(00)00123-8. 1, 2, 65, 66, 67

Y.-Y. Shih and C.-Y. Liu. A method for customer lifetime value ranking – Combining the analytic hierarchy process and clustering analysis. *Journal of Database Marketing & Customer Strategy Management*, 11(2):159–172, December 2003. DOI: https://doi.org/10.1057/palgrave.dbm.3240216. 150

R. Silipo. Seven Techniques for Data Dimensionality Reduction. https://www.kdnuggets.com/2015/05/7-methods-data-dimensionality-reduction.html, May 2015. [Online, Accessed: 2017-10-31]. 58

W. R. Smith. Product differentiation and market segmentation as alternative marketing strategies. *The Journal of Marketing*, 21(1):3–8, July 1956. DOI: http://doi.org/10.2307/1247695. 9

B. Sohrabi and A. Khanlari. Customer lifetime value (CLV) measurement based on RFM model. *Iranian Accounting & Auditing Review*, 14(47):7–20, 2007. https://acctgrev.ut.ac.ir/article_18552_3c49bb9a53ce8058c5e7d811b5515f2d.pdf. 16

O. Solon. Cambridge Analytica closing after Facebook data harvesting scandal. https://www.theguardian.com/uk-news/2018/may/02/cambridge-analytica-closing-down-after-facebook-row-reports-say, May 2018. [Online, Accessed: 2018-07-20]. 3

Z. Soltani and N. J. Navimipour. Customer relationship management mechanisms: A systematic review of the state of the art literature and recommendations for future research. *Computers in Human Behavior*, 61:667–688, August 2016. DOI: http://dx.doi.org/10.1016/j.chb.2016.03.008. 22

C. O. S. Sorzano, J. Vargas, and A. P. Montano. A survey of dimensionality reduction techniques. *CiteSeer*, March 2014. https://arxiv.org/abs/1403.2877. 60

P. Spring, P. S. Leeflang, and T. Wansbeek. The Combination Strategy to Optimal Target Selection and Offer Segmentation in Direct Mail. *Journal of Market-Focused Management*, 4(3):187–203, October 1999. DOI: http://dx.doi.org/10.1023/A:1009899802421. 16

Statistics Solutions. Conduct and Interpret a Factor Analysis. http://www.statisticssolutions.com/factor-analysis-2/, 2017. [Online, Accessed: 2017-10-31]. 58

Statistics South Africa. Census 2011 Fertility in South Africa. http://www.statssa.gov.za/publications/Report-03-01-63/Report-03-01-632011.pdf, 2011. [Online, Accessed: 2017-10-01]. 260

Statistics South Africa. Gender statistics in South Africa, 2011. http://www.statssa.gov.za/publications/Report-03-10-05/Report-03-10-052011.pdf, 2013a. [Online, Accessed: 2017-10-01]. 260

Statistics South Africa. Overview of the South African Retail Market. http://www.tv.camcom.gov.it/docs/Corsi/Atti/2013_11_07/OverviewOFTHESOUTHAfrica.pdf, 2013b. [Online, Accessed: 2017-10-01]. 261

Statistics South Africa. General Household Survey. https://www.statssa.gov.za/publications/P0318/P03182015.pdf, 2015a. [Online, Accessed: 2017-10-01]. 260, 261

Statistics South Africa. Living Conditions of Households in South Africa. http://www.statssa.gov.za/publications/P0310/P03102014.pdf, 2015b. [Online, Accessed: 2017-10-01]. 260, 261

Statistics South Africa. Education Series Volume III: Education Enrolment and Achievement, 2016. http://www.statssa.gov.za/, 2016. [Online, Accessed: 2017-10-01]. 260

Statistics South Africa. Mid-year population estimates 2017. http://www.statssa.gov.za/publications/P0302/P03022017.pdf, 2017. [Online, Accessed: 2017-10-01]. 257, 260

S. Stein, J. Hubbard, J. Vance, and C. Guyer. SQL Server Management Studio (SSMS). https://docs.microsoft.com/en-us/sql/ssms/sql-server-management-studio-ssms, 2017. [Online, Accessed: 2018-02-20]. 252

Stellenbosch University. Faculty of Engineering Academic Programmes and Faculty Information. https://www.sun.ac.za/english/Documents/Yearbooks/Current/Engineering.pdf, May 2018. [Online, Accessed: 2018-08-03]. 148

R. Steynberg. A framework for identifying the most likely successful underprivileged tertiary bursary applicants. Master's thesis, Stellenbosch: Stellenbosch University, December 2016. http://hdl.handle.net/10019.1/100336. 33, 70, 76, 77

R. Steynberg, D. Lötter, and J. H. Van Vuuren. Framework for identifying the most likely successful underprivileged tertiary study bursary applicants. *South African Journal of Industrial Engineering*, 28(2):59–77, 2017. DOI: http://dx.doi.org/10.7166/28-2-1695. 54

Study.com. What Are Information Systems? - Definition & Types. https://study.com/academy/lesson/what-are-information-systems-definition-types-quiz.html, 2018. [Online, Accessed: 2018-02-20]. 251

J. Tang, S. Alelyani, and H. Liu. *Feature selection for classification: A review.* CRC Press, July 2014. https://www.crcpress.com/Data-Classification-Algorithms-and-Applications/Aggarwal/p/book/9781466586741. 56, 58

K. L. Taylor. *Oracle® Data Mining.* Oracle, June 2013. 11g Release 2 (11.2). https://docs.oracle.com/cd/E11882_01/datamine.112/e16808.pdf. 68, 84

Techopedia. What does Entity-Relationship Diagram (ERD) mean? https://www.techopedia.com/definition/1200/entity-relationship-diagram-erd, 2018. [Online, Accessed: 2018-02-27]. 253

G. Tellis and T. Ambler. *The SAGE Handbook of Advertising.* SAGE Publications, 2007. ISBN 9781473971561. https://books.google.co.za/books?id=ovNcCwAAQBAJ. 85, 91

G. J. Tellis. Modeling Marketing Mix. *Handbook of marketing research*, pages 506–522, August 2006. http://www-bcf.usc.edu/~tellis/mix.pdf. 85, 91

The Housing Development Agency (HDA). South Africa: Informal settlements Status. http://www.thehda.co.za/uploads/files/HDA_South_Africa_Report_lr.pdf, 2013. [Online, Accessed: 2017-10-01]. 260

H. Thompson. *The Customer-Centered Enterprise: How IBM and Other World-Class Companies Achieve Extraordinary Results by Putting Customers First*. McGraw-Hill, New York, 1999. ISBN 9780071371407. https://books.google.co.za/books?id=bSTWzV1P384C&dq=The+Customer-centered+enterprise:+How+IBM+and+other+world-class+companies+achieve+extraordinary+results+by+putting+customers+first&source=gbs_navlinks_s. 15

J. M. Tien. Big Data: Unleashing information. *Journal of Systems Science and Systems Engineering*, 22(2):127–151, June 2013. DOI: https://doi-org.ez.sun.ac.za/10.1007/s11518-013-5219-4. 31, 35, 36

D. Tomar and S. Agarwal. A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology*, 5(5):241–266, 2013. DOI: http://dx.doi.org/10.14257/ijbsbt.2013.5.5.25. 46, 71, 75

D. Trewartha. Investigating data mining in MATLAB. Master's thesis, Department of Science, Rhodes University, Grahamstown, 2006. http://pppj2012.ru.ac.za/g03t2052/CSHnsThesis.pdf. 153

Truth. South African Loyalty Landscape 2017. http://truth.co.za/wp-content/uploads/Truth-Whitepaper-October-2017.pdf, 2017. [Online, Accessed: 2018-10-01]. 261

K. K. Tsiptsis and A. Chorianopoulos. *Data Mining Techniques in CRM: Inside Customer Segmentation*. John Wiley & Sons, 2011. ISBN 9781119965459. https://books.google.co.za/books?id=t4ZIKY7sMRsC&dq=Data+mining+techniques+in+CRM:+inside+customer+segmentation&source=gbs_navlinks_s. 99

M. Udell and S. Boyd. PCA on a Data Frame. 2014. https://people.orie.cornell.edu/mru8/doc/udell15_pca_dataframe.pdf. 58, 60

G. D. Ungerer. *A Competitive Strategy Framework for E-Business Start-Ups*. PhD thesis, Stellenbosch University, 2015. http://hdl.handle.net/10019.1/97930. 10, 12, 14, 25

USMA. USMA Working Group, Dept. of Industrial Engineering, Stellenbosch University. Unit for Systems Modelling and Analysis, 2017. 38, 41, 42, 43, 56, 58, 67, 69, 70, 77, 85, 91, 98, 99, 214

V. Vapnik. *The Nature of Statistical Learning Theory*. Information Science and Statistics. Springer Science & Business Media, 1999. ISBN 9780387987804. https://books.google.co.za/books?id=sna9BaxVbj8C. 71, 75, 76, 77

C. Walters. *Consumer Behaviour: Theory and Practice*. Richard D. Irwin, Inc, Third edition, 1974. ISBN 9780256015973. https://books.google.co.za/books?id=aVqJcDXS8QEC. 26

M. Walters and J. Bekker. Customer Super-Profiling Demonstrator to Enable Efficient Targeting in Marketing Campaigns. *South African Journal of Industrial Engineering*, 28(3):113–127, 2017. DOI: http://dx.doi.org/10.7166/28-3-1846. 2, 3, 113, 118, 215

H. Wang, Z. Xu, H. Fujita, and S. Liu. Towards felicitous decision making: An overview on challenges and trends of Big Data. *Information Sciences*, 367–368:747–765, 2016. DOI: https://doi.org/10.1016/j.ins.2016.07.007. 34, 35

W. A. Wedel, Michel Kamakura. Introduction to the Special Issue on Market Segmentation. *Intern. J. of Research in Marketing*, 19:181–183, 2002. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2395277. 9

J.-T. Wei, S.-Y. Lin, and H.-H. Wu. A review of the application of RFM model. *African Journal of Business Management*, 4(19):4199–4206, 2010. http://www.academicjournals.org/journal/AJBM/article-full-text-pdf/EB3418D18198. 15, 16, 18

A. Weinstein. *Handbook of Market Segmentation: Strategic Targeting for Business and Technology Firms, Third Edition*. Taylor & Francis, 2013. ISBN 9781135185664. https://books.google.co.za/books?id=sQXfAQAAQBAJ. 9, 10, 12, 13, 29, 30

G. Williams, R. Baxter, H. He, S. Hawkins, and L. Gu. A Comparative Study of RNN for Outlier Detection in Data Mining. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 709–712. IEEE, 2002. DOI: https://doi.org/10.1109/ICDM.2002.1184035. 55

L. Yang, S. Liu, S. Tsoka, and L. G. Papageorgiou. A regression tree approach using mathematical programming. *Expert Systems with Applications*, 78:347–357, July 2017. DOI: https://doi.org/10.1016/j.eswa.2017.02.013. 85

D. Yankelovich and D. Meer. Rediscovering Market Segmentation. *Harvard Business Review*, 84(2):122, 2006. http://viewpointlearning.com/wp-content/uploads/2011/04/segmentation_0206.pdf. 23, 24, 26, 28

I.-C. Yeh, K.-J. Yang, and T.-M. Ting. Knowledge discovery on RFM model using Bernoulli sequence. *Expert Systems with Applications*, 36(3):5866–5871, April 2009. DOI: https://doi.org/10.1016/j.eswa.2008.07.018. 19

K. Y. Yeung and W. L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, September 2001. DOI: https://doi.org/10.1093/bioinformatics/17.9.763. 58, 60

K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10): 977–987, October 2001. DOI: https://doi.org/10.1093/bioinformatics/17.10.977. 107

M. Yoon. Developing basic soccer skills using reinforcement learning for the robocup small size league. Master's thesis, Stellenbosch: Stellenbosch University, 2015. http://hdl.handle.net/10019.1/96823. 78, 79

Z. Zalaghi and Y. Varzi. Measuring customer loyalty using an extended RFM and clustering technique. *Management Science Letters*, 4(5):905–912, 2014. DOI: http://dx.doi.org/10.5267/j.msl.2014.3.026. 19, 21

P. Zikopoulos, D. deRoos, K. Parasuraman, T. Deutsch, J. Giles, and D. Corrigan. *Harness the Power of Big Data The IBM Big Data Platform*. McGraw-Hill Education, 2012. ISBN 9780071808187. https://books.google.co.za/books?id=HhSONOxOCQOC. 35, 36

C. Zopounidis. *New Trends in Banking Management*. Contributions to Management Science. Physica-Verlag HD, 2012. ISBN 9783642574788. https://books.google.co.za/books?id=wqOMBwAAQBAJ. 24

Appendix A

Data simulation

The development of a simulator to create datasets for analysis is presented here. This includes building an information system, creating and storing domain-specific datasets, as well as validating these datasets. The datasets will be used by the CSP tool to illustrate the concept of customer super-profiling.

A.1 Domain identification

The aim of this research is to create a CSP tool, containing datasets with specific properties. The datasets will be used by the CSP tool to illustrate the concept of customer super-profiling. The demonstrator is designed to contain the specific simulated datasets as input data, and if an enterprise wants to utilise the tool, customer data needs to be provided in the same format.

Before creating datasets, an information system needs to developed. This is to manage the creation and growth of the (customer) records. Next, the information system for this research will be created.

A.2 Creating the information system

An *information system* can be defined as the software that helps with collection, organisation, storage and communication of data. The purpose of an information system is to turn raw data into useful information that can help with decision-making within an organisation (Study.com, 2018). Kroenke (2014) stated: *"An information system is*

a group of components that interact to produce information. It focuses on the internal rather than the external."

The software utilised to host the information system for this research is Microsoft SQL Server Management Studio (SSMS) 2014. SSMS is an integrated environment for managing any SQL infrastructure. SSMS is utilised to access, configure, manage, administer, and develop all components of SQL Server (Stein et al., 2017). One of the main reasons for selecting SSMS as the database tool is that the researcher had already worked with this software in an undergraduate module, and has knowledge regarding this software.

The next step regarding the simulated datasets includes determining the customer attributes/properties. Once this is determined the SQL database can be created which consists of various tables. The subsection to follow will discuss these tables in more detail.

A.2.1 Domain properties

This section will provide more insight into the database that will be constructed for this research. Firstly, the various customer features that provide more knowledge into customer behaviour will be mentioned. These attributes will form the tables which constitute a database. Following this, the data dictionary of this database will be illustrated.

The researcher decided to utilise 16 customer features (demographic and extra value adding features) as well as behavioural features which include typical monetary transactional history. Table A.1 lists all the customer tables that need to be created. The table is divided into two 'groups'. The first 'group' consists of the first 16 customer features that are used to create the customer database. These first 16 tables include all the characteristics that customers can possess only one of, for example each customer is only assigned one gender, one ethnicity, one province, *etc.*. The second 'group' in Table A.1 contains the customer attributes that customers can possess zero or more of (many-to-many relationships). For example, each customer could visit up to 20 retail shops: if so, a date and the amount spent at the retail shop will be documented; however, a customer can also be present in the database and not participate in any retail shopping activities.

Next, the data dictionary of the information system will be illustrated, indicating the name of the table, followed by an explanation for each table as well as indicating the

Table A.1: Illustrating the customer tables

Customer Tables : One-to-many relationships			
1	Gender	9	Children Status
2	Ethnicity	10	Household Size
3	Province	11	Medical Aid
4	Age	12	Housing Ownership
5	Education	13	Housing Type
6	Employment	14	Transportation
7	Annual Income	15	Mobile Phone
8	Relationship Status	16	Mobile Contract
Customer Tables: Many-to-many relationships			
1	Retail Shop (intersection table)	1.1	Retail Shop Name
		1.2	Activities/Transactions

data type. A *data dictionary* or a metadata repository, as defined by IBM (1993), is a *"centralised repository of information about data such as meaning, relationships to other data, origin, usage and format."* Kendall and Kendall (2013) defined data dictionaries as a collective work of data about data (metadata), compiled by systems analysts to guide them through analysis and design. The data dictionary can also be used to control access to and manipulation of the database.

Tables A.2 and A.3 contain the data dictionary of the information system created. Table A.2 represents the first 'group' indicated in Table A.1. Table A.3 indicates all the retail shops which customers visit: only 20 anonymised retail shops were considered. This concludes the discussion regarding the data dictionary, next an *extended entity relationship diagram* will be constructed to illustrate the relationships between all of these entities.

A.2.2 Extended entity relationship diagram

An *entity relationship diagram* (ERD) is a data modelling technique that graphically illustrates data relationships in an information system (Techopedia, 2018). An ERD contains many entities, many different types of relations, and numerous attributes (Kendall and Kendall, 2013). An ERD is a conceptual and representational model of data used to represent the entity framework infrastructure (Techopedia, 2018). In this subsection, the

Table A.2: Customer features used in the study

Variable name	**Explanation**	**Scaling**
Gender	Male or Female	Categorical (Dichotomous, Figure 3.1)
Ethnicity	Black African, Coloured, Indian/Asian or White	Categorical (Multichotomous, Figure 3.1)
Province	Eastern Cape, Free State, Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West or Western Cape	Categorical (Multichotomous, Figure 3.1)
Age	15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79 or 80+	Categorical (Multichotomous, Figure 3.1)
Education	Less than Gr.12 and no other qualification, Less than Gr.12 and with diploma or certificate, Gr.12, Gr.12 with diploma or certificate, Degree or post graduate degree or Honours degree or higher	Categorical (Multichotomous, Figure 3.1)
Employment status	Employed, Unemployed or Not economically active	Categorical (Multichotomous, Figure 3.1)
Annual income	R0–R12 000, R12 001–R54 000, R54 001–R192 000, R192 001–R360 000 or More than R360 001	Categorical (Multichotomous, Figure 3.1)
Relationship status	Married or domestic partner, Never married or single, Widowed or Divorced	Categorical (Multichotomous, Figure 3.1)
Children status	Yes or No	Categorical (Dichotomous, Figure 3.1)
Household size	1, 2, 3..., 10+	Categorical (Multichotomous, Figure 3.1)
Medical aid	Yes or No	Categorical (Dichotomous, Figure 3.1)
Housing ownership	Rented, Owned (not fully), Owned (fully), Occupied rent free or Other	Categorical (Multichotomous, Figure 3.1)
Housing type	Cluster house in complex, Flat or apartment in flat block, House or brick structure on yard or stand, House, flat or room in backyard, Informal – shack in backyard, Informal – shack not backyard, Other, Room, granny flat or large dwelling, Semi-detached house, Townhouse, Traditional dwelling – hut or Overcrowding	Categorical (Multichotomous, Figure 3.1)
Transportation	Train, Bus, Taxi, Car, Walk/Cycle or Other	Categorical (Multichotomous, Figure 3.1)
Mobile phone	Samsung, Other, Apple, Huawei, Nokia, Blackberry, Sony, LG, HTC, Motorola or Siemens	Categorical (Multichotomous, Figure 3.1)
Mobile contract	Prepaid or Contract	Categorical (Dichotomous, Figure 3.1)

extended entity relationship diagram (EERD) will be created. The difference between the ERD and the EERD is that the EERD has *intersection entities* where necessary,

Table A.3: Customer purchasing (transactional) behaviour features

Variable name	**Explanation**	**Scaling**
Retail shop names (anonymised)	ShopWrong, Select&Debt, RetailA, Nylonworths, WePay, Kliks, ThisKem, RetailB, JetPlane, Cokcor, VosGroup, MrsFee, RetailC, Inspectets, WoolOn, RetailD, Poems, Kara, MarkHim, Retail E	Categorical (Multichotomous, Figure 3.1)
Activities/ Transactions	Retail shop	Categorical (Multichotomous, Figure 3.1)
	Transaction date	Date
	Amount spent	Numeric

as well as *optionalities*. The optionalities are modelled using 0's and 1's and follow the revised crow's foot notation (Kendall and Kendall, 2013).

Figure A.1 illustrates the EERD created, indicating the various relationships between all the entities. The table *tbl_Customers*, which all the entities are connected to, contains all the customer records. A record is seen as the collection of all the attributes (entities), and has to have a unique primary key (identification key/number) for each record. All the customer variables in Table A.2 have one-to-many relationships with the customer table, while the *retail shop name* entity in Table A.3 has a many-to-many relationship with the customer table. Many-to-many relationships do not get modelled on an EERD, therefore an *intersection entity* is necessary. The *retail shop* entity is denoted as an intersection entity, as seen in Figure A.1. An intersection entity functions only as an index set connecting the other two entities, namely *retail shop name* and *activities*. Thus, the *retail shop* entity indicates which customer(s) visit(s) what retail shop, providing a unique key for each 'combination'.

Figure A.2 represents five customers and their demographic and extra value adding features, as seen in the database table *tbl_Customers*. Each customer has 16 features (as indicated by the columns) which are represented by their primary key values. For example the second column indicates the gender of the customer, either by displaying '1' or '2'. Figure A.3 displays the user view *i.e.* the primary key values of Figure A.2 are 'hidden'. Each customer feature has its own amount of features that can be selected (*e.g.* gender has two values, age has 14 category values *etc.*), as indicated in Table B.1.

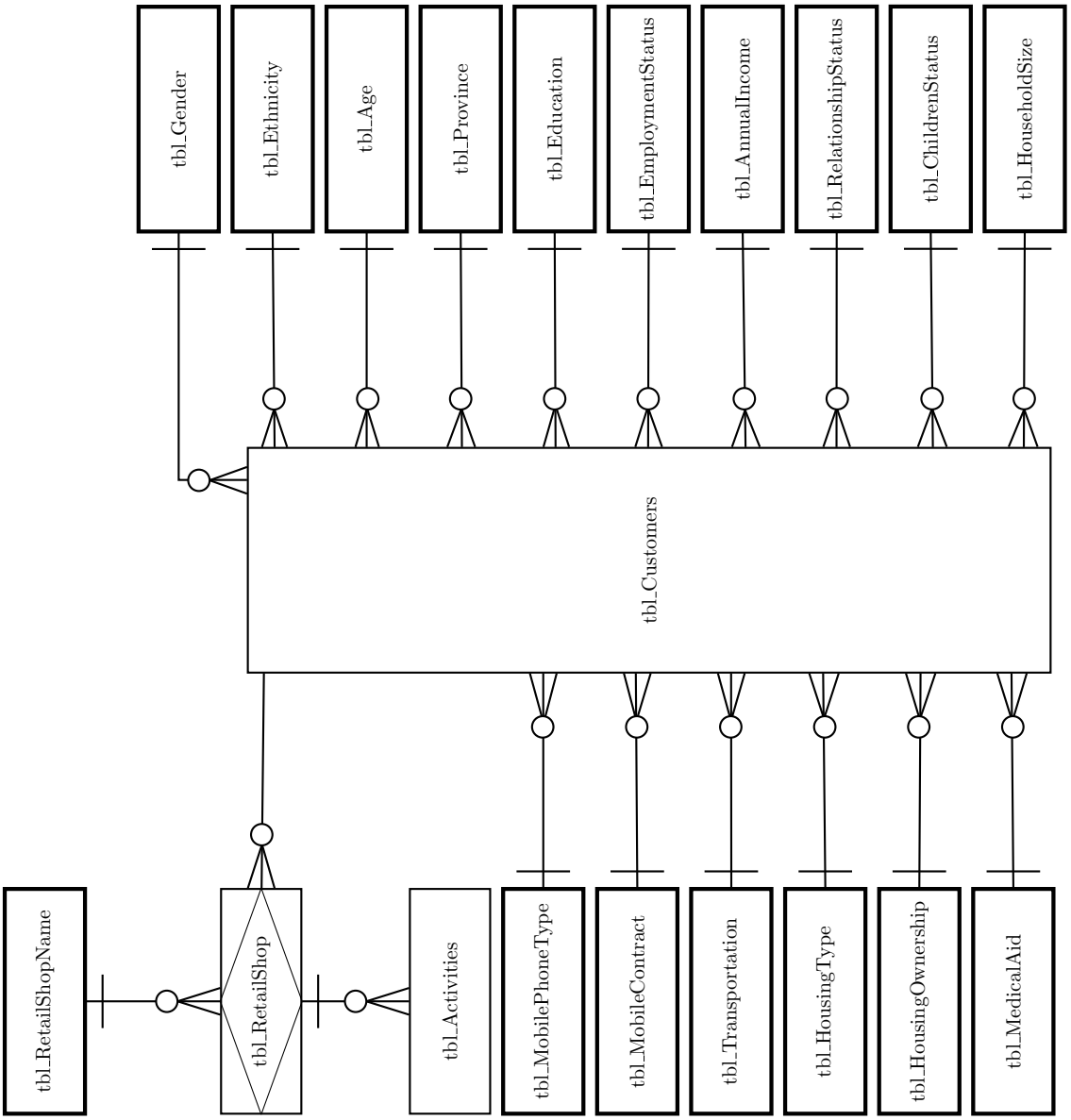


Figure A.1: The EERD supporting the CSP tool

Customer_ID	Gender_IDFK	Ethinicity_IDFK	Province_IDFK	Age_IDFK	Education_IDFK	...	Transportation...	MobilePhone_...	MobileContra...
1	2	1	4	4	1	...	5	1	1
2	1	1	4	4	1	...	5	1	1
3	1	4	9	5	2	...	3	8	2
4	1	1	4	2	1	...	5	2	2
5	2	1	5	10	1	...	5	1	1

Figure A.2: Illustrating the top five customers in table *tbl_Customers*

Customer_ID	Gender	Ethinicity	Province	Age	Education	...	Transportation	MobilePhoneType	MobileContract
1	Female	Black African	KwaZulu-Natal	30-34	Less than Gr.12 and no other qualification	...	Walk/cycle	Samsung	PrePaid
2	Male	Black African	KwaZulu-Natal	30-34	Less than Gr.12 and no other qualification	...	Walk/cycle	Samsung	PrePaid
3	Male	White	Western Cape	35-39	Less than Gr.12 and with diploma or certificate	...	Taxi	LG	Contract
4	Male	Black African	KwaZulu-Natal	20-24	Less than Gr.12 and no other qualification	...	Walk/cycle	Other	Contract
5	Female	Black African	Limpopo	60-64	Less than Gr.12 and no other qualification	...	Walk/cycle	Samsung	PrePaid

Figure A.3: User view of top five customers in table *tbl_Customers*

Similar to Figures A.2 and A.3, the top five customers' retail shops are displayed in Figure A.4. *Customer 1* does not visit a retail shop, and is therefore not present in Figure A.4, while customers 2 to 5 appear numerously in Figure A.4. Table B.2 indicates the retail shops linked to each primary key value (*RetailShopType_IDFK*).

A.3 Data simulator and Customer Super-Profiling tool logic

Figure A.5 schematically presents the elements of data creation and analysis necessary to reach the goal of this research. The subsections to follow will discuss the contents of Figure A.5 in more detail.

A.3.1 Data simulator

The first element to be discussed is the simulator, as indicated in Figure A.5. The researcher, with guidance from the study leader, decided on simulating datasets according to South African demographics (Statistics South Africa, 2017). The datasets contained in different tables have different distributions so the data are random and more realistic. These data distributions were determined by the researcher and the study leader as

RetailShop_ID	Customer_IDFK	RetailShopType_IDFK
2	2	1
3	2	2
4	2	6
5	2	7
6	2	5
7	2	3
8	3	11
9	3	2
10	3	1
11	3	5
12	3	6
13	3	3
14	3	8
15	3	7
16	4	1
17	4	3
18	4	10
19	4	5
20	4	6
21	4	2
22	5	5
23	5	16
24	5	1
25	5	4
26	5	7
27	5	2

RetailShopType
ShopWrong
Select&Debt
Kliks
ThisKem
WePay
RetailA
VosGroup
Select&Debt
ShopWrong
WePay
Kliks
RetailA
RetailB
ThisKem
ShopWrong
RetailA
Cokcor
WePay
Kliks
Select&Debt
WePay
RetailD
ShopWrong
Nylonworths
ThisKem
Select&Debt

Figure A.4: Illustrating the top five customers in table *tbl_RetailShop* together with the user view

suitable solutions. The CSP tool will be developed to contain these simulated datasets as input data.

The researcher started the data simulation with a set of assumptions derived from the real world (deductive), and produced simulation-based data that can be analysed (inductive). The set of assumptions includes various South African statistics that were utilised in order to create values for the datasets seen in Table A.2. These assumptions include: (1) customers that are still in school will have an employment status of not economically active and (2) customers below the age range of 25-29 will not be able to

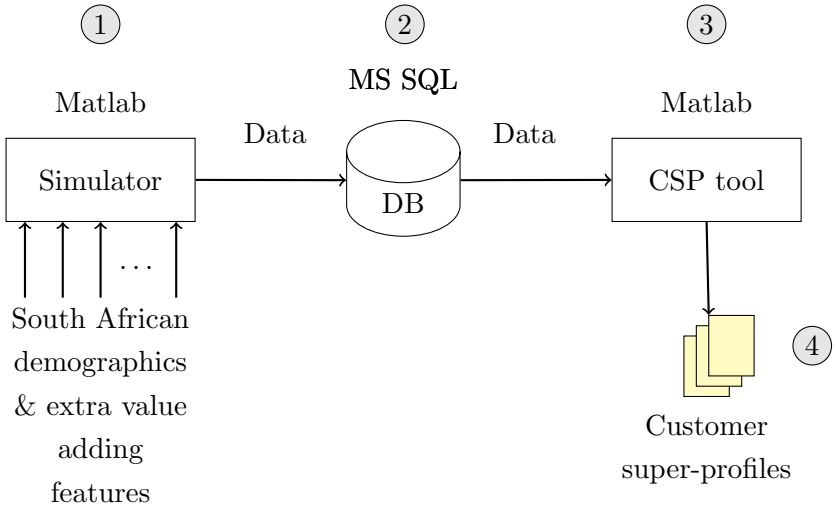


Figure A.5: Conceptual illustration of the elements of data creation and analysis

have an educational level higher than a first degree or diploma qualification. Appendix C contains the pseudocode of the data simulator utilised to create the 50 000 customer dataset with the 16 customer features. The data of the intersection entity (retail shop) was created by using the *acceptance/rejection sampling method* (Ross, 2013). The distribution utilised by this method was derived from data regarding the popularity of retail shops, as seen in Table A.4. Table A.5 indicates all the sources that were utilised in order to create datasets reflecting South African statistics (indicated in Figure A.5 under phase 1). One of the entries in Table A.3 indicates the *'transaction date'*. The researcher assumed an integer number of days between transactions. Since these intervals must be finite, an offset truncated Poisson distribution was used. The *'amount spent'* entry in Table A.3, is distributed by following the Beta distribution, with various alpha (α) and beta (β) values.

All of the data values are created according to each distribution (Matlab) and need to be stored within a database (Microsoft SQL Server). These are fairly expensive products for commercial users, but the proposed solution is independent of software products. Open-source software can be used instead. The next section will briefly discuss the relationship between these two software packages.

Table A.4: Retail shop distributions

Retail shop	Percentage	Retail shop	Percentage
ShopWrong	67%	VosGroup	14%
Select&Debt	66%	MrsFee	14%
RetailA	44%	RetailC	12%
Nylonworths	43%	Inspectets	11%
WePay	43%	WoolOn	11%
Kliks	38%	RetailD	11%
ThisKem	32%	Poems	9%
RetailB	23%	Kara	8%
JetPlane	15%	MarkHim	5%
Cokcor	15%	RetailE	5%

Table A.5: Sources used to distribute data accordingly

Table	Source
Gender	Statistics South Africa (2017)
Ethnicity	Statistics South Africa (2017)
Province	Statistics South Africa (2017)
Age	Statistics South Africa (2017)
Education	Statistics South Africa (2013a) Statistics South Africa (2016)
Employment	Lehohla (2016) Statistics South Africa (2016)
Annual Income	Statistics South Africa (2015b)
Relationship Status	Lehohla (2016) Statistics South Africa (2013a) Statistics South Africa (2016)
Children Status	Palamuleni et al. (2007) Statistics South Africa (2011)
Household Size	Statistics South Africa (2015a) Statistics South Africa (2015b)
Medical Aid	Statistics South Africa (2013a)
Housing Ownership	Statistics South Africa (2015a)
Housing Type	Statistics South Africa (2015a) Statistics South Africa (2015b) The Housing Development Agency (HDA) (2013)

Continued on next page

Table	Source
Transportation	Lombard et al. (2017) Statistics South Africa (2015a) Statistics South Africa (2015b)
Mobile Phone	Effective Measure (2015)
Mobile Contract	Effective Measure (2015)
Retail Shop Type	BusinessTech (2017) Statistics South Africa (2013b) Truth (2017)

A.3.2 Matlab and Microsoft SQL Server

After completing the simulation of the datasets (element 1) it needs to be stored within a database (element 2). Matlab is a high-level language and interactive environment for numerical computation, visualisation and programming. Matlab can be used to analyse data, develop algorithms and to create models and applications. Matlab has the ability to access a database server and then perform data manipulation (Kurniawan, 2013).

Matlab possesses the *Database Toolbox*™ [1] that provides functions with relational databases. Data from relational databases can be accessed when making use of SQL commands or the Database Explorer app, without using Microsoft SQL Server (MathWorks, 2018b).

To access this data from Matlab, a data source and connection to the Microsoft SQL Server database is necessary. The Database Explorer app accesses the Microsoft *Open Database Connectivity* (ODBC) Data Source Administrator automatically when configuring an ODBC data source. Figure A.6 conceptualises the connection between Matlab and Microsoft SQL Server, also indicating that an ODBC connection is created (MathWorks, 2018a). The ODBC connection serves as the data flow (Figure A.5) from the simulator (Matlab) to the database (Microsoft SQL), and then again from the database to the CSP tool (Matlab).

A.4 Validation of simulator

The validation of the simulator is necessary because it will be used to create customer datasets. The first 16 data tables were created by following the same method and

[1] The trademark for Database Toolbox™ will from now on be omitted.

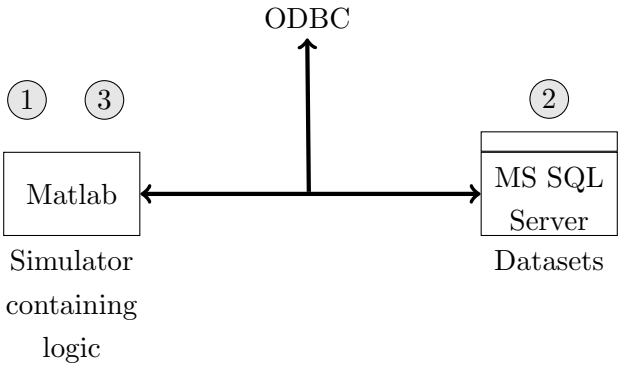


Figure A.6: Conceptual illustration of the connection between Matlab and Microsoft SQL Server

distribution; however, each table has its own preliminary conditions that need to be satisfied.

Table A.6 indicates the preliminary conditions of the datasets mentioned in Table A.2. The preliminary conditions are the customer features that need to be considered when distributing the contents of a specific table, *e.g.* when distributing the ethnicity of the customers, the researcher considered whether the customers were male or female and then assigned a race accordingly (sources for distributions are indicated in Table A.5). There are several tables which do not have preliminary conditions. In such cases only the statistics were utilised for the creation of values in the tables.

Table A.6: Indicating the preliminary conditions of the one-to-many tables created in Matlab

Table	**Preliminary Condition(s)**
Gender	–
Ethnicity	Gender
Province	–
Age	Gender Ethnicity
Education	Age Ethnicity
Employment	Age Gender Education

Continued on next page

Table	Preliminary Condition(s)
Annual Income	Employment(1 = Employed) Gender Education
Relationship Status	Ethnicity Age Gender
Children Status	Ethnicity Relationship Status Age
Household Size	Ethnicity Relationship Status Age
Medical Aid	Ethnicity
Housing Ownership	Ethnicity Gender
Housing Type	Annual Income
Transportation	Annual Income Employment (2 = Unemployed, 3 = Not economically active) Education Ethnicity
Mobile Phone	–
Mobile Contract	–

The researcher validated the simulator by creating the datasets containing 1 000 000 customers, populating the tables and confirming the output. Tables A.7, A.8 and A.9 represent, as examples, the validation of the customers' ethnicity, age and housing ownership tables, respectively.

Table A.7: Validation of table 'Ethnicity'

Male – 490 000			
Ethnicity	**Theoretical percentage**	**Simulated percentage**	**Absolute percentage deviation**
Black African	80.80%	80.83%	0.06%
Coloured	8.70%	8.70%	0.03%
Indian/Asian	2.60%	2.62%	0.54%
White	7.90%	7.85%	0.78%
Female – 510 000			

Continued on next page

Ethnicity	Theoretical percentage	Simulated percentage	Absolute percentage deviation
Black African	80.80%	80.68%	0.12%
Coloured	8.86%	8.99%	1.49%
Indian/Asian	2.39%	2.43%	1.93%
White	7.98%	7.90%	1.00%

Table A.8: Validation of table 'Age'

Male: Black African – 396 060							
Age	**Theoretical percentage**	**Simulated percentage**	**Absolute percentage deviation**	**Age**	**Theoretical percentage**	**Simulated percentage**	**Absolute percentage deviation**
15-19	12.62%	12.68%	0.50%	50-54	5.03%	5.02%	0.16%
20-24	13.88%	13.82%	0.41%	55-59	4.06%	4.01%	1.23%
25-29	15.36%	15.37%	0.04%	60-64	3.06%	3.11%	1.37%
30-34	14.59%	14.62%	0.26%	65-69	2.03%	1.98%	2.40%
35-39	11.62%	11.61%	0.10%	70-74	1.17%	1.19%	1.95%
40-44	8.92%	9.02%	1.04%	75-79	0.66%	0.66%	0.73%
45-49	6.54%	6.47%	1.13%	80+	0.45%	0.44%	2.69%
Female: Black African – 411 449							
Age	**Theoretical Percentage**	**Simulated percentage**	**Absolute percentage deviation**	**Age**	**Theoretical percentage**	**Simulated percentage**	**Absolute percentage deviation**
15-19	11.95%	11.96%	0.01%	50-54	5.78%	5.74%	0.67%
20-24	13.15%	13.16%	0.02%	55-59	4.77%	4.82%	1.04%
25-29	14.52%	14.56%	0.22%	60-64	3.85%	3.85%	0.19%
30-34	13.61%	13.50%	0.77%	65-69	2.80%	2.80%	0.03%
35-39	10.65%	10.66%	0.14%	70-74	1.87%	1.86%	0.52%
40-44	7.98%	7.97%	0.04%	75-79	1.24%	1.23%	0.93%
45-49	6.48%	6.54%	0.84%	80+	1.35%	1.35%	0.43%
Male: Coloured – 42 626							
Age	**Theoretical percentage**	**Simulated percentage**	**Absolute percentage deviation**	**Age**	**Theoretical percentage**	**Simulated percentage**	**Absolute percentage deviation**
15-19	12.01%	11.82%	1.62%	50-54	7.45%	7.54%	1.20%
20-24	12.58%	12.54%	0.34%	55-59	6.24%	6.17%	0.98%
25-29	12.69%	12.78%	0.69%	60-64	4.39%	4.31%	1.96%
30-34	11.61%	11.62%	0.08%	65-69	2.98%	3.12%	4.60%
35-39	9.61%	9.49%	1.25%	70-74	1.72%	1.67%	2.87%
40-44	8.89%	9.04%	1.70%	75-79	0.88%	0.92%	5.33%
45-49	8.36%	8.34%	0.21%	80+	0.59%	0.64%	8.32%
Female: Coloured – 45 839							
Age	**Theoretical Percentage**	**Simulated percentage**	**Absolute percentage deviation**	**Age**	**Theoretical percentage**	**Simulated percentage**	**Absolute percentage deviation**
15-19	10.86%	10.77%	0.79%	50-54	8.08%	8.19%	1.35%
20-24	11.42%	11.52%	0.81%	55-59	6.74%	6.91%	2.51%
25-29	11.60%	11.60%	0.07%	60-64	5.27%	5.35%	1.51%
30-34	10.72%	10.51%	1.91%	65-69	3.87%	3.87%	0.03%
35-39	9.14%	9.07%	0.76%	70-74	2.42%	2.32%	4.00%
40-44	8.35%	8.47%	1.40%	75-79	1.57%	1.55%	1.45%
45-49	8.61%	8.56%	0.54%	80+	1.36%	1.31%	3.69%
Male: Indian/Asian – 12 829							
Age	**Theoretical percentage**	**Simulated percentage**	**Absolute percentage deviation**	**Age**	**Theoretical percentage**	**Simulated percentage**	**Absolute percentage deviation**

Continued on next page

15-19	7.79%	7.55%	3.10%	50-54	7.57%	7.19%	5.02%
20-24	9.48%	9.43%	0.51%	55-59	6.26%	6.20%	0.93%
25-29	11.48%	11.13%	3.00%	60-64	4.99%	4.79%	3.85%
30-34	12.91%	13.46%	4.25%	65-69	3.63%	3.35%	7.55%
35-39	12.06%	12.16%	0.81%	70-74	2.26%	2.19%	3.13%
40-44	10.59%	11.02%	4.07%	75-79	1.23%	1.23%	0.29%
45-49	9.01%	9.53%	5.76%	80+	0.76%	0.77%	1.26%
Female: Indian/Asian – 12 407							
Age	**Theoretical Percentage**	**Simulated percentage**	**Absolute percentage deviation**	**Age**	**Theoretical percentage**	**Simulated percentage**	**Absolute percentage deviation**
15-19	7.71%	7.64%	0.92%	50-54	8.01%	8.66%	8.23%
20-24	9.19%	9.36%	1.81%	55-59	7.02%	7.02%	0.02%
25-29	10.46%	10.91%	4.30%	60-64	5.98%	5.75%	3.95%
30-34	11.21%	10.97%	2.17%	65-69	4.85%	4.93%	1.72%
35-39	10.18%	9.77%	3.99%	70-74	3.44%	3.46%	0.57%
40-44	9.32%	8.89%	4.63%	75-79	2.22%	2.20%	0.81%
45-49	8.53%	8.19%	3.99%	80+	1.88%	2.25%	19.53%
Male: White – 38 485							
Age	**Theoretical percentage**	**Simulated percentage**	**Absolute percentage deviation**	**Age**	**Theoretical percentage**	**Simulated percentage**	**Absolute percentage deviation**
15-19	6.98%	6.87%	1.57%	50-54	8.42%	8.29%	1.47%
20-24	7.38%	7.31%	1.02%	55-59	8.18%	8.19%	0.05%
25-29	7.82%	7.73%	1.23%	60-64	7.45%	7.50%	0.68%
30-34	8.49%	8.38%	1.26%	65-69	6.63%	6.64%	0.10%
35-39	8.28%	8.59%	3.72%	70-74	5.27%	5.34%	1.42%
40-44	8.88%	8.88%	0.05%	75-79	3.59%	3.73%	3.69%
45-49	9.39%	9.27%	1.20%	80+	3.23%	3.28%	1.34%
Female: White – 40 305							
Age	**Theoretical Percentage**	**Simulated percentage**	**Absolute percentage deviation**	**Age**	**Theoretical percentage**	**Simulated percentage**	**Absolute percentage deviation**
15-19	6.33%	6.24%	1.44%	50-54	8.25%	8.33%	0.90%
20-24	6.83%	6.93%	1.58%	55-59	8.38%	8.65%	3.19%
25-29	7.24%	7.33%	1.29%	60-64	7.59%	7.60%	0.21%
30-34	7.88%	7.65%	2.94%	65-69	7.11%	7.13%	0.24%
35-39	7.69%	7.64%	0.68%	70-74	5.76%	5.90%	2.55%
40-44	8.63%	8.58%	0.55%	75-79	4.33%	4.28%	1.26%
45-49	8.94%	8.73%	2.39%	80+	5.03%	5.00%	0.61%

Table A.9: Validation of table 'Housing ownership'

Male: Black African – 396 060			
Housing ownership	**Theoretical percentage**	**Simulated percentage**	**Absolute percentage deviation**
Rented	24.49%	24.63%	0.58%
Owned (not fully)	6.04%	5.98%	0.97%
Owned (full)	50.54%	50.00%	1.07%
Occupied rent free	16.20%	16.65%	2.80%
Other	2.73%	2.74%	0.20%

Continued on next page

Female: Black African – 411 449			
Housing Ownership	**Theoretical Percentage**	**Simulated percentage**	**Absolute percentage deviation**
Rented	12.58%	12.57%	0.08%
Owned (not fully)	3.77%	3.74%	0.69%
Owned (full)	68.52%	68.55%	0.04%
Occupied rent free	12.94%	12.96%	0.17%
Other	2.20%	2.18%	0.67%
Male: Coloured – 42 626			
Housing ownership	**Theoretical percentage**	**Simulated percentage**	**Absolute percentage deviation**
Rented	19.91%	20.19%	1.43%
Owned (not fully)	22.46%	22.33%	0.57%
Owned (full)	40.87%	40.81%	0.13%
Occupied rent free	12.57%	12.39%	1.50%
Other	4.19%	4.28%	2.09%
Female: Coloured – 45 839			
Housing Ownership	**Theoretical Percentage**	**Simulated percentage**	**Absolute percentage deviation**
Rented	19.86%	20.24%	1.92%
Owned (not fully)	11.45%	11.37%	0.65%
Owned (full)	51.64%	51.18%	0.88%
Occupied rent free	10.51%	10.50%	0.14%
Other	6.54%	6.70%	2.47%
Male: Indian/Asian – 12 829			

Continued on next page

Housing ownership	**Theoretical percentage**	**Simulated percentage**	**Absolute percentage deviation**
Rented	27.03%	27.45%	1.58%
Owned (not fully)	24.32%	24.03%	1.20%
Owned (full)	41.70%	41.69%	0.01%
Occupied rent free	3.09%	2.74%	11.42%
Other	3.86%	4.08%	5.79%
Female: Indian/Asian – 12 407			
Housing Ownership	**Theoretical Percentage**	**Simulated percentage**	**Absolute percentage deviation**
Rented	26.60%	26.09%	1.90%
Owned (not fully)	14.89%	15.11%	1.47%
Owned (full)	56.38%	56.62%	0.42%
Occupied rent free	0.00%	0.00%	n/a
Other	2.13%	2.18%	2.28%
Male: White – 38 485			
Housing ownership	**Theoretical percentage**	**Simulated percentage**	**Absolute percentage deviation**
Rented	22.77%	22.45%	1.45%
Owned (not fully)	34.55%	34.86%	0.88%
Owned (full)	38.13%	38.26%	0.35%
Occupied rent free	2.27%	2.18%	3.79%
Other	2.27%	2.25%	0.93%
Female: White – 40 305			

Continued on next page

Housing Ownership	**Theoretical Percentage**	**Simulated percentage**	**Absolute percentage deviation**
Rented	30.51%	31.02%	1.67%
Owned (not fully)	19.70%	19.44%	1.34%
Owned (full)	42.58%	42.30%	0.66%
Occupied rent free	3.81%	3.79%	0.59%
Other	3.39%	3.45%	1.74%

The output values are not exactly equal to the theoretical values, but the order sizes are acceptable. The absolute deviations are not significant, except in the case of very small fractions of age (*i.e.* 80+ year old Indian/Asian female). The simulated values will converge when simulating large datasets. This validation technique was performed on all the tables that were populated, to ensure that valid datasets will be created. This concludes the section regarding the validation of the simulator.

Appendix B

Key descriptors of the information system

Table B.1: Look up values for the customer features

Customer features	**Description**
Gender	1 = Male 2 = Female
Ethnicity	1 = Black African 2 = Coloured 3 = Indian/Asian 4 = White
Province	1 = Eastern Cape 2 = Free State 3 = Gauteng 4 = KwaZulu-Natal 5 = Limpopo 6 = Mpumalanga 7 = Northern Cape 8 = North West 9 = Western Cape

Continued on next page

Customer features	**Description**
Age	1 = 15–19 2 = 20–24 3 = 25–29 4 = 30-34 5 = 35–39 6 = 40–44 7 = 45–49 8 = 50–54 9 = 55–59 10 = 60–64 11 = 65–69 12 = 70–74 13 = 75–80 14 = 80+
Education	1 = Less than Gr.12 and no other qualification 2 = Less than Gr.12 and with diploma or certificate 3 = Gr.12 4 = Gr.12 with diploma or certificate 5 = Degree or post graduate degree 6 = Honours degree or higher
Employment status	1 = Employed 2 = Unemployed 3 = Not economically active
Annual income	1 = R0–R12 000 2 = R12 001–R54 000 3 = R54 001–R192 000 4 = R192 001–R360 000 5 = More than R360 001
Relationship status	1 = Married or domestic partner 2 = Never married or single 3 = Widowed 4 = Divorced
Children status	1 = Yes 2 = No

Continued on next page

Customer features	**Description**
Household size	1 = 1 2 = 2 3 = 3 4 = 4 5 = 5 6 = 6 7 = 7 8 = 8 9 = 9 10 = 10+
Medical aid	1 = Yes 2 = No
Housing ownership	1 = Rented 2 = Owned (not fully) 3 = Owned (fully) 4 = Occupied rent free 5 = Other
Housing type	1 = Cluster house in complex 2 = Flat or apartment in flat block 3 = House or brick structure on yard or stand 4 = House, flat or room in backyard 5 = Informal – shack in backyard 6 = Informal – shack not backyard 7 = Other 8 = Room, granny flat or large dwelling 9 = Semi-detached house 10 = Townhouse 11 = Traditional dwelling – hut 12 = Overcrowding
Transportation	1 = Train 2 = Bus 3 = Taxi 4 = Car 5 = Walk/cycle 6 = Other

Continued on next page

Customer features	**Description**
Mobile phone type	1 = Samsung 2 = Other 3 = Apple 4 = Huawei 5 = Nokia 6 = Blackberry 7 = Sony 8 = LG 9 = HTC 10 = Motorola 11 = Siemens
Mobile contract	1 = Prepaid 2 = Contract

Table B.2: Look up values for the customer purchasing (transactional) behaviour features

Customer features	**Description**
Retail shop type (anonymised)	1 = ShopWrong 2 = Select&Debt 3 = RetailA 4 = Nylonworths 5 = WePay 6 = Kliks 7 = ThisKem 8 = RetailB 9 = JetPlane 10 = Cokcor 11 = VosGroup 12 = MrsFee 13 = RetailC 14 = Inspectets 15 = WoolOn 16 = RetailD 17 = Poems 18 = Kara 19 = MarkHim 20 = RetailE

Appendix C

Pseudocode: Data simulator

C.1 Matlab data simulator

Algorithm 5 Data simulator for South African demographic customer dataset

Begin
Generate 50 000 customers
For each row in table `Customers`
  Assign **Gender** [Male, Female]
  Assign **Ethnicity** [Black, Coloured, Indian/Asian, White]
  Assign **Province** [Eastern Cape, Free State, Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West, Western Cape]
  Assign **Age** [15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80+]
  Assign **Education** [Less than Gr.12 and no other qualification, Less than Gr.12 and with diploma or certificate, Gr.12, Gr.12 with diploma or certificate, Degree or postgraduate degree, Honours degree or higher]
  Assign **Employment** [Employed, Unemployed, Not economically active]
  Assign **Annual income** [R0-R12 000, R12 001-R54 000, R54 001-R192 000, R192 001-R360 000, More than R360 000]
  Assign **Relationship status** [Married/domestic partner, Never married/single, Widowed, Divorced]
  Assign **Children status** [Yes, No]
  Assign **Household size** [1, 2, 3, 4, 5, 6, 7, 8, 9, 10+]
  Assign **Medical aid** [Yes, No]
  Assign **Housing ownership** [Rented, Owned (not fully), Owned (fully), Occupied rent free, Other]
  Assign **Housing type** [Cluster house in complex, Flat or apartment in flat block, House or brick structure on yard or stand, House, flat or room in backyard, Informal – shack in backyard, Informal – shack not backyard, Other, Room, granny flat or large dwelling, Semi-detached house, Townhouse, Traditional dwelling – hut, Overcrowding]
  Assign **Transportation type** [Train, Bus, Taxi, Car, Walk/Cycle, Other]
  Assign **Mobile phone type** [Samsung, Other, Apple, Huawei, Nokia, Blackberry, Sony, LG, HTC, Motorola, Siemens]
  Assign **Mobile contract type** [Contract, Prepaid]
Next row
End