SAIIE29 Proceedings, 24th – 26th of October 2018, Spier, Stellenbosch, South Africa © 2018 SAIIE

DEVELOPMENT AND DEMONSTRATION OF A CUSTOMER SUPER-PROFILING TOOL TO ENABLE EFFICIENT TARGETING IN MARKETING CAMPAIGNS

M. Walters[1*] & J. Bekker[2]

[1]Department of Industrial Engineering
Stellenbosch University, South Africa
17618142@sun.ac.za

[2]Department of Industrial Engineering
Stellenbosch University, South Africa
jb2@sun.ac.za

ABSTRACT

Being part of a competitive generation demands having good marketing policies to attract new customers as well as to retain existing customers. This research outlines a general methodology for segmentation of customers by using the model of Recency, Frequency and Monetary (RFM) to identify types of customers, and then predict their customer profiles, based on demographic and behavioural features. A few previous studies dealt with the question using non-aggregate customer data. We, however, also address the problem by using decision trees, something which has rarely been done before. We applied and demonstrated this tool on a large customer dataset and found useful results.

[1] The author was enrolled for an M Eng (Industrial) degree in the Department of Industrial Engineering, Stellenbosch University, South Africa
*Corresponding author

1. INTRODUCTION

The research work reported on in this paper is the continuation of previous work published in [1] in the domain of data analytics and deals with specific aspects of customer profiling. 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 [2]. 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 creating the ideal customer profile [3].

This paper offers an approach to build customer profiles through data mining tools (*i.e.* supervised and 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 referred to as a technique used to extract knowledge from information [5]. 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 [4]. Thus, one can define data mining, 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.

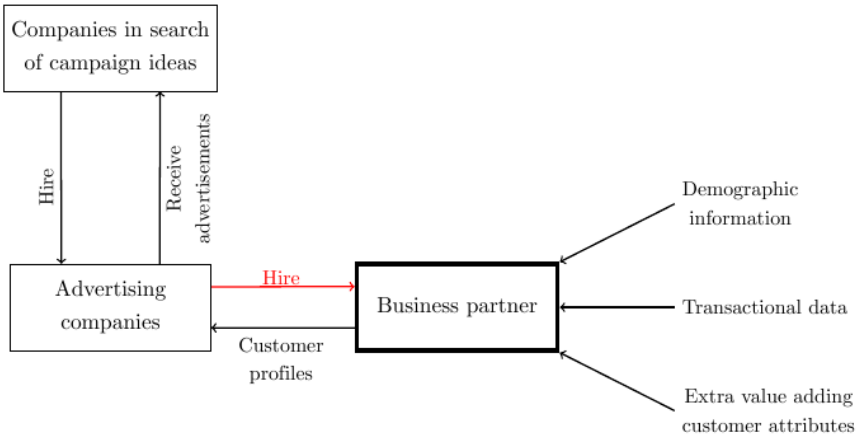

Figure 1: Illustrating the use of customer super-profiling

Figure 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 [1]. The customer attributes include demographic information (age, gender, ethnicity, *etc.*), transactional data, as well as extra value-adding attributes (transportation type, mobile phone, *etc.*). Companies 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 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.

It is the authors' purpose to develop a CSP *tool* that has the ability to analyse a large dataset by utilising various big data analytics tools and techniques. 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 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.

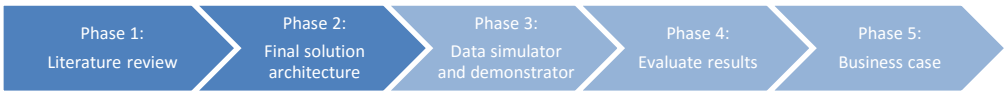

Figure 2: Project methodology

Figure 2 indicates the structure of this paper. Phase 1 of this project refers to the literature review that will be presented in *section 2*, which is brief, as the literature necessary to perform this work was presented previously in [1]. *Section 3* corresponds with Phase 2 in the figure, and presents the final solution architecture for the CSP tool. Phases 3 to 5 form the 'new' work, indicated as future work in [1]. *Section 4* contains the development of the data simulator and demonstrator, as well as evaluating the results. *Section 5* documents the business case applicable to this demonstration tool, while the last section gives the concluding remarks.

2. LITERATURE REVIEW

This section will provide a literature summary of specific data mining tools and techniques utilised to create the proposed customer super-profiling demonstrator. To initialise the literature review, a brief summary regarding segmentation will be provided.

2.1 Segmentation

Segmentation is often used in conjunction with profiling. However, segmentation is a term used to describe the process of dividing customers into homogeneous groups based on shared or common attributes, *e.g.* habits, tastes, *etc.*, while customer profiling describes the customers by using demographic information, *e.g.* age, gender, *etc.* as well as the behaviour of their homogeneous group.

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 [6]. Segmentation of a customer dataset provides a more targeted communication with the customers, because the characteristics of a certain group of customers, or a cluster, are known [7].

Literature does not provide a distinguishable difference between market segmentation and customer segmentation, therefore the authors adopted the view that market segmentation is generally used for high-level strategy, whereas customer segmentation provides a more detailed view. Next to be discussed is the well-known heuristic approach called Recency, Frequency and Monetary (RFM) analysis.

2.2 RFM analysis

According to [8], 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. Subsequently, [9] 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 individual customers within a cluster/group and the information delivered to customers is uniform, whereas direct marketing targets individuals or households. Different customers receive different marketing information.

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 produces problems, as the correlation between customer segments and a campaign is neglected. Therefore, as stated by [10], it is necessary to consider significant campaign-dependent variables of customer targeting in customer segmentation. An approach to combine customer segmentation and customer targeting for campaign strategies was defined by [11]. The investigation identified customer behaviour, using the well-known *Recency*, *Frequency* and *Monetary* (RFM) analytical model ([2], [11], [12]). The detailed definition for each RFM parameter is as follows [13]:

- Recency (R): Represents the duration of time between the last purchase date/time and the date/time the 'survey' took place. The shorter the interval, the higher the recency value of a customer.
- Frequency (F): Represents the total number of purchases during the specific period (survey). The higher the number of purchases in an interval, the higher the frequency value of a customer.
- Monetary (M): Represents the monetary value of the purchases in the time interval considered. The higher the amount spent in an interval, the higher the monetary value of a customer.

After applying 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

[11], [14]. If the value lies between 100% and 80%, the categorical value is set to 5; if between 80% and 60%, the value is set to 4, *etc*. In this way, the database is divided into 125 ($5 \times 5 \times 5$) 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 method, RFM, is to analyse customers' behaviour and then make predictions based on the behaviour in the database [15]. The model is used in various research areas, which defines valuable customers as those simultaneously having high *recency*, *frequency* and *monetary* values. According to [2], one of the most effective customer segmentation models, based on customer value, is the RFM model. By adopting the RFM model, decision-makers can identify valuable customers and then develop effective marketing strategies [15].

The RFM model has been widely applied in many practical areas and its indicators are adaptable to measure customer value and to segment customers in different service areas ([14], [15], [16]). Next, a clustering technique (unsupervised learning), that can be performed on values obtained from the RFM analysis, will be discussed. This approach is explored because many researchers have considered the RFM variables when developing clustering models.

2.3 Clustering — *k*-means

The clustering technique that will be discussed is called, *k*-means, which forms part of the partitioning (non-hierarchical) method. The authors decided on using *k*-means as it is the most frequently used clustering technique, whilst it provides a good foundation for understanding clustering and is a simple and elegant approach to portioning a dataset into *k* distinct clusters.

The motive for using a clustering technique within this demonstrator is mainly to group together customers with similar purchasing or transactional patterns. When performing RFM analysis prior to the clustering process, it is possible to cluster customers based on the values obtained from the RFM analysis. The resulting clusters should have minimum dissimilarity within the cluster and maximum dissimilarity with other clusters [17].

It is important to understand the broad picture of how *k*-means performs clustering. After all the customer transactional data are transformed into RFM categories, the data can be grouped by using the *k*-means clustering technique. As mentioned, to be able to group the data into several clusters, certain steps need to be performed ([18], [19], [20]):

1. **Determine the desired number of clusters:** To find the optimal/desired number of clusters in a dataset, the *k*-means clustering algorithm needs to be performed for a range of *k* values and the results compared. In general, there is no method for determining the exact value for *k*, yet an accurate estimate can be obtained by using various techniques [19]. These techniques include: elbow plot, cross-validation, information criteria, the information theoretic jump method, the silhouette method, and the G-means algorithm. The *silhouette method* together with the squared *Euclidean distance* as distance metric will be used in this paper. Silhouette plot refers to a method of interpretation and validation of consistency within clusters of data. This technique provides a concise graphical representation of how well each object lies within its cluster (testing over the predefined range of *k* values). The silhouette method provides silhouette values for each data point by using the distance metric. The silhouette value is a measure of how similar an object is to its own cluster compared to the neighbouring cluster. The silhouette values range from -1 to +1 where a high value indicates that the object is well matched to its own cluster and poorly matched to the neighbouring cluster [20], [24].

2. **Determine the centroids:** The specialised algorithm called *k*-means++ algorithm uses a heuristic to find centroid seeds for *k*-means clustering. The *k*-means++ algorithm was proposed as a specific way of selecting centroids for the *k*-means clustering algorithm, instead of generating centroids randomly. It determines the initial centre points by calculating their squared distance from the closest centre already chosen.

3. **Allocate data points to each cluster:** The data points are assigned to its closest centroid. This is done by calculating the distance of each data point with regards to the centroid's position. 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 formula for SSE/RSS is

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

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

$$c_k = \frac{\sum_{x_i \epsilon c_k} x_i}{|c_k|}.$$

4. **Recalculate the new cluster centroids**: After all the data points are assigned to the closest cluster, compute the average of the data points in each cluster to obtain *k* new centroid locations.

5. **Repeat steps 2—4:** After obtaining the new centroids for each cluster, repeat steps 2 through 4 until cluster assignments do not change, or the maximum number of iterations is reached, because *k*-means is an *iterative algorithm*

This concludes the discussion regarding *k*-means for the purpose of this paper. A supervised learning technique that can be applied on clustered data, discovered by a clustering algorithm, will be discussed next.

2.4 Predictive model(s)

Classification algorithms are used to derive rules from the clustered results, obtained from the previous sections' results. These classification or decision rules are useful for identifying each and every customer from their purchasing patterns (RFM information) [17], [21]. There are various techniques for classification, *e.g.* decision trees, neural networks, *etc.*

The classification technique that will be utilised in this paper is decision trees. Once a decision tree or decision rule solution is generated from data, it can be used for estimating or predicting the response or class variable for a new case. The application of a decision tree is a straightforward top-down decision process, controlled by evaluating the tests and taking the appropriate branch, beginning at the root and terminating when a leaf node is reached. 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 [21].

Estimating the true accuracy of a decision tree or rule model is an important aspect of the modelling process. A solution generated from a set of training examples will almost always be highly accurate on the same dataset, but far less accurate on new data. A two-fold strategy will be applied in this paper, where the first step involves generating the model from training data, and the second step involves testing the proposed solution on independent cases as part of future work [21], [24].

After completing the literature review, the authors noticed that very little has been done to apply unsupervised learning and supervised learning in conjunction, at least in the Industrial Engineering domain. This offers an opportunity to investigate the integrated use of both data analytics tools. This concludes the literature review for the paper. Next to be discussed is the final architecture for the super-profiling tool.

3. FINAL ARCHITECTURE FOR THE CUSTOMER SUPER-PROFILING TOOL

The proposed data simulator and super-profiling tool were developed in precursory work of [1], using the standard Object Process Methodology (OPM) approach [22]. Revisions have been made and the final solution architecture can be seen in Figure 3.

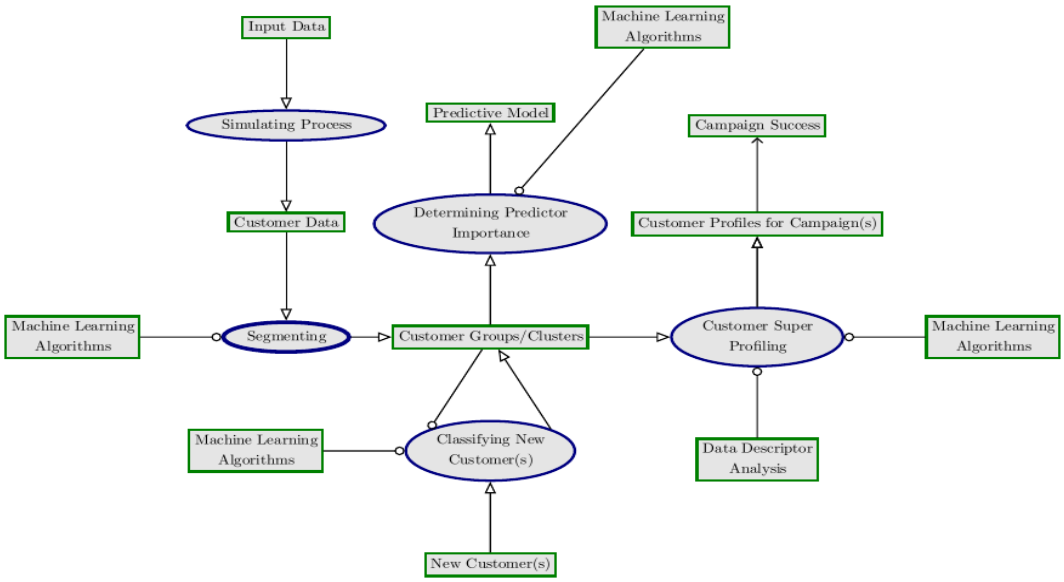

Figure 3: System diagram (SD) representing the working of the final proposed data simulator and CSP tool (Revised from [1]).

Object Process Language (OPL) is the semantic counterpart of the graphic OPM system specifications. The OPL is automatically generated as a textual description of the system in a subset natural English. Following the OPM guidelines, the OPL for Figure 3 is:

Customer Profiles for Campaign(s) relates to Campaign Success.
Customer Super Profiling requires Statistical Analysis and Machine Leaning Algorithms.
Customer Super Profiling consumes Customer Groups/Clusters.
Customer Super Profiling yields Customer Profiles for Campaign(s).
Segmenting requires Machine Leaning Algorithms.
Segmenting consumes Customer Data.
Segmenting yields Customer Groups/Clusters.
Simulating Process consumes Input Data.
Classifying New Customers requires Customer Groups/Clusters and Machine Leaning Algorithms.
Classifying New Customers consumes New Customer(s).
Classifying New Customers yields Customer Groups/Clusters.
Determining Predictor Importance requires Machine Leaning Algorithms.
Determining Predictor Importance consumes Customer Groups/Clusters.
Determining Predictor Importance yields Predictive model.

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

4. DATA SIMULATION AND CUSTOMER SUPER-PROFILING TOOL

The previous section proposed the final system's architecture. This architecture aids in reaching the goal of the research, which is to develop a tool that contains a suite of Big Data Analytics tools and techniques which will allow for customer super-profiling.

The need for a data simulator that creates datasets with *specific properties* was identified in the previous sections. This section presents the process followed to create the specific datasets. The datasets will be used by the 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 demonstrator, customer data with the same format and structure would need to be extracted.

4.1 Customer data

This section will provide more insight into the database that was constructed for this research. Firstly, the various customer attributes that provide more knowledge about a customer will be mentioned. These attributes will form the tables which constitute a database. The authors decided on utilising various customer

characteristics (*e.g.* demographic and behavioural attributes). A few of the attributes are basic demographic information, while other attributes are more specific to customer behaviour (*e.g.* mobile phone type). Table 1 and Table 2 describe the available customer demographic data and customer behavioural features — those related to customer purchasing behaviour.

Table 1: Customer features used in the study

Variable name	Explanation	Scaling
Gender	Male or Female	Categorical
Ethnicity	Black African, Coloured, Indian/Asian or White	Categorical
Province	Eastern Cape, Free State, Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West or Western Cape	Categorical
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
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
Employment	Employed, Unemployed or Not economically active	Categorical
Annual income	R0-R12 000, R12 001-R54 000, R54 001-R192 000, R192 001-R360 000 or More than R360 001	Categorical
Relationship status	Married or domestic partner, Never married or single, Widowed or Divorced	Categorical
Children status	Yes or No	Categorical
Household size	1, 2, 3…,10+	Categorical
Medical aid	Yes or No	Categorical
Housing ownership	Rented, Owned (not fully), Owned (fully), Occupied rent free or Other	Categorical
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 or granny flat or large dwelling, Semi-detached house, Townhouse, Traditional dwelling — hut or Overcrowding	Categorical
Transportation	Train, Bus, Taxi, Car, Walk/Cycle or Other	Categorical
Mobile phone	Samsung, Other, Apple, Huawei, Nokia, Blackberry, Sony, LG, HTC, Motorola or Siemens	Categorical
Mobile contract	Prepaid or Contract	Categorical

Table 2: 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
Activities/ Transactions	Retail shop	Categorical
	Transaction date	Date
	Amount spent	Numeric

The authors decided on simulating the datasets according to South African demographics [23]. The datasets contained in the different tables have different distributions so the data are random and more realistic. The authors 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). 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 have an educational level higher than a first degree or diploma qualification.

Two software packages were used to support this project, namely Matlab® and Microsoft® (MS) SQL Server®. Matlab is a high-level language and interactive environment for numerical computation, visualisation and programming and has the ability to access a database server and then perform data manipulation. To access data from Matlab a data source and connection to MS SQL Server database is necessary. The Database Explorer app (in Matlab) accesses the Microsoft ODBC Data Source Administrator automatically when configuring an ODBC data source. Figure 4 conceptualises the connection between Matlab and MS SQL Server, also indicating that an ODBC connection is created.

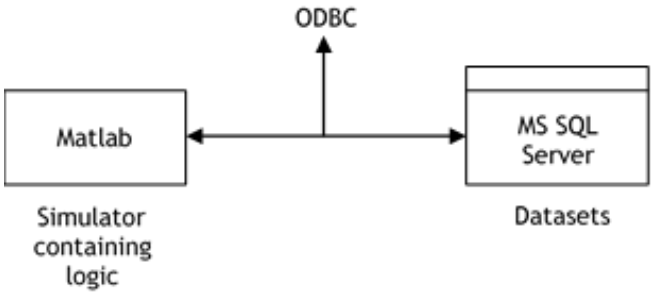

Figure 4: Conceptual illustration of the connection between Matlab and MS SQL Server

4.2 Application of proposed tool

The goal of this research is to develop a customer super-profiling demonstrator. In order to create this demonstrator, a simulator was developed which creates various datasets. These datasets contain demographic, typical transactional and personal preference information, of 50 000 customers. The demonstrator will make use of various techniques to analyse the data in its control. The techniques as well as the results of each technique will subsequently be discussed.

Figure 5 illustrates the broad outline of the CSP tool which will be developed throughout this section. The main steps within this outline include:

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

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 relevant results retrieved during the steps.

4.2.1 Select data

The first step (1) in Figure 5 indicates that data needs to be selected. Large amounts 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 MS SQL Server, 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 data must first be 'cleaned'. Next, the RFM analysis will be performed.

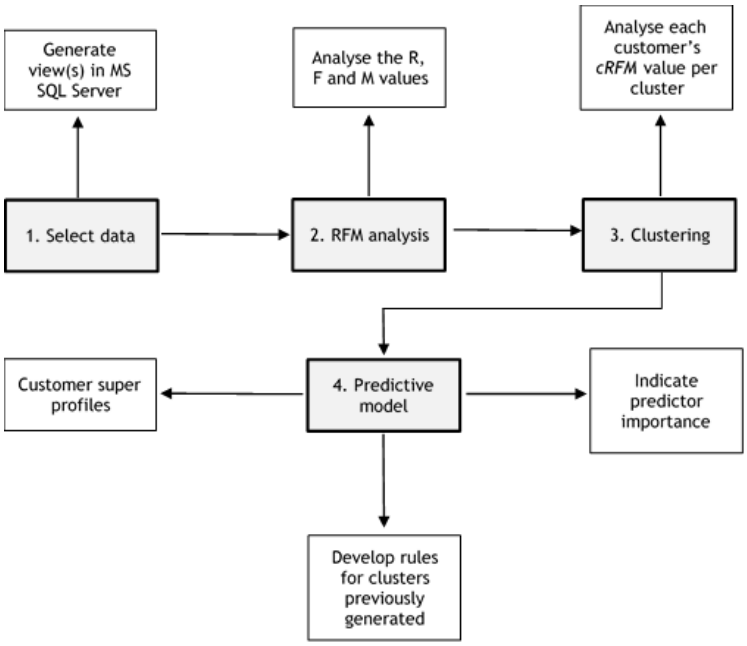

Figure 5: Schematic representing the CSP tool

4.2.2 RFM analysis

After selecting the appropriate customer information, the second step includes the *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. The RFM analysis will be performed on the dataset created in the previous step (1): this dataset includes all *retail shops* each customer visited, will not focus on only one shop or products purchased. 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 the retail shop visits of each customer, and will be used to fulfil the next step, namely *clustering*.

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

- *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.
- *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
- *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.

Figure 6 schematically represents the R, F and M category range values.

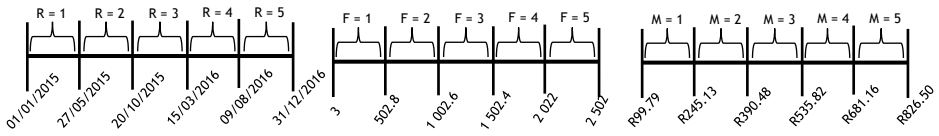

Figure 6: Schematic representing the R, F and M category range values

Table 3 indicates the transformation of the data after each customer has been assigned their R, F and M category value of the transactional data. After assigning each customer their own R, F and M value, it can be stated that more than 80% of the customers have a high R-value; the majority of customer have an F-value equal to 3, and

lastly more than 80% of the customers have a low M-value. This data is now ready for the next step, which is clustering.

Table 3: Top five rows of customer RFM matrix

Customer_ID	*R*	*F*	*M*
2	5	3	1
3	5	4	1
4	5	3	1
5	5	3	1
7	5	4	1

4.2.3 Clustering

Clustering is the process of grouping similar objects. Clustering is performed on the RFM dataset created in the previous step (2). The *k*-means clustering method will be applied to this dataset. The best number of clusters is determined using Matlab to yield silhouette plots. The number of clusters associated with the highest silhouette value is seen as the best number of clusters. However, as indicated by literature, the selection of the best number of clusters is subjective. The silhouette values provide a reasonable indication of the cluster structure; the higher the silhouette value, the better the cluster structure. Figure 7 indicates the silhouette criterion value for each number of clusters tested. The best number of clusters suggested by this function is 10, with a silhouette value of 0.957. As mentioned, this is only a suggestion, and the authors decided to select two clusters for this dataset (silhouette value of 0.84). Various aspects were considered for not selecting the best number of clusters, but rather two clusters, such as marketing cost, the more clusters the more marketing efforts need to be funded and smaller clusters lead to more analysis possibilities. The number of customers that need to be clustered is 33 510, and with ten clusters, each cluster would possess a small group of customers. Therefore, two clusters is a reasonable number to perform analysis on, and the clusters will be divided more equally.

In total 50 000 customers were simulated, and only 33 510 customers participated in the RFM analysis. The remaining 16 490 customers did not visit a retail shop (those involved in this study) and will form by default their own cluster. Figure 8 illustrates a scatter plot of the two clusters that were formed when applying the *k*-means clustering method. This figure indicates that if only the recency and frequency values were considered, the centroids would lie very close to each other (almost overlap), therefore a third dimension, the monetary values, contributes to creating a suitable cluster structure for this dataset, dividing the two centroids along the vertical axis. The results received from Table 4 (percentage customers in each RFM category per cluster) are consistent with the structure and positioning shown in Figure 8.

As indicated by Table 4, cluster 1 has the majority observations, followed by cluster 3; which is the non-shoppers, followed by cluster 2. The biggest difference between (the customers of) cluster 1 and cluster 2 lies in the frequency parameter. This means that the customers allocated to cluster 1 are more frequent customers and could be seen as more loyal customers, as opposed to the customers allocated to cluster 2. ("Loyal" in this context is towards the retail stores involved in this study.)

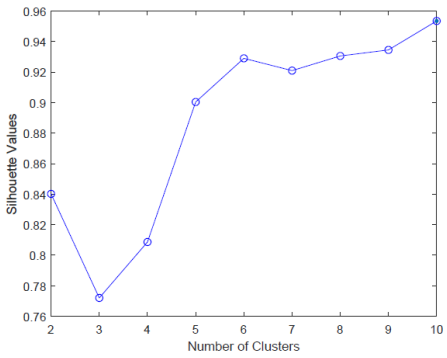

Figure 7: Plot of the silhouette criterion values for each number of clusters tested

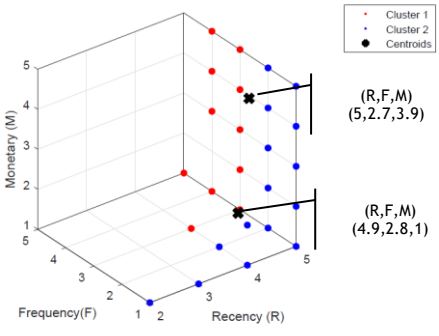

Figure 8: Scatter plot representing the two clusters

Table 4: 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%	
	21 671			11 839			16 490

For deeper insights into the customers' value and purchasing behaviour, we defined a *combined* RFM for each customer, calling it cRFM. This value is determined by adding each customer's R, F and M category value and then dividing the total by 3, as follows:

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

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.

Table 5 represents a summary of the percentage customers in each cRFM category, for clusters 1 and 2. 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: Percentage customers in each cRFM category for clusters 1 and 2

	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 present *types of customers* in the dataset.

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. A study conducted by [25] made use of a *'RFM pattern'* to indicate how the RFM category values of each customer segment differ from the 'original' RFM category values (Table 4) in the dataset.

The authors 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 total average RFM category values 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.

The five customer types that were decided on are listed in Table 6, together with various customer characteristics' explanations, adapted from research conducted by [26], the customers' identification traits (cRFM category values) and the RFM patterns of each cRFM category.

Table 6: Types of customers in dataset

Type of customer	Customer characteristics:	Cluster 1	Cluster 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 brand(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 groups that 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 demographic information is available.	Only cluster 3.	

Next, a prediction model that includes the customer demographic information will be developed, according to the various types of customers.

4.2.4 Predictive model

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 [17]. Integrating the RFM analysis with both clustering (step 3) and classification provides useful information for current and new customers and more behavioural knowledge of customers is attained; as opposed to other independent clustering and classification techniques.

According to [27] 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.

For these reasons, decision rules were discovered using the customers' demographic and extra value adding features (age, gender, province, medical aid, mobile phone type, *etc.*), as seen in Table 1, to generate customer super-profiles for the various types of customers (Table 6). 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.

In total 18 decision rules were generated for this study to identify a type of customer. Table 7 shows a selection of four decision rules that are utilised when predicting a customer's type (*i.e.* (new) low spenders, low loyal customer, (new) big spender, or prospect). This set of rules can provide (1) customer super-profiles for each type of customer and (2) classify new/future customers.

Table 7: Selection of decision rules to identify the type of 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 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 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.

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 7; 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 2 and 3) or to 'Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West or Western Cape' (Rules 8 and 14). Those customers that belong to the first group of provinces have a different second customer feature that are used to distinguish then 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.

Not all types of customers are restricted to one cluster. Therefore, after the customer type is known (Table 7), the cluster to which that customer belongs can be determined. 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 groups indicated in Table 6, columns three and four were selected. Again, only a selection of four rules are shown, as seen in Table 8. 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. The most distinguishing customer feature for this set of rules is the *employment status*.

Table 8: Selection of decision rules to identify (new) low spenders

Rule 1:	**if** *EmploymentStatus*= Employed or Unemployed **then** Cluster 1.
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.

Table 9 shows four selected decision rules generated to identify (new) big spenders and the cluster to which they belong. The misclassification error for this set of rules is 10.9 percent, with the most distinguishing customer feature as the *age* category.

Table 9: Selection of 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 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 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 10:	**if** *Age*= 1 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.

Next, the decision rules for low loyal customers are shown in Table 10. The misclassification rate for this decision tree is 9.8 percent. The distinguishable customer feature is the customer's *province*.

Table 10: Selection of decision rules to identify low loyal customers

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

As indicated in Table 6, the churned cheap customers are not of interest to marketers. However, in this customer dataset, churned cheap customers only belong to 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 (Table 8—Table 10), it is found that the rules used to identify (new) low spenders (Table 8) 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 article was to develop a tool that demonstrates the generating of customer super-profiles given a specific dataset, through performing the steps illustrated in Figure 5. The predictive model for this study, 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. Marketers using this tool will have more knowledge of their customers especially when they know which type of customer they are interested in.

5. 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 project 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 demonstrator 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 demonstrator can be used when marketing a product for a certain target group is necessary. This target group can be 'found' by the demonstrator and will provide all relevant 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.

6. CONCLUSIONS

Customer super-profiling consists of a large set of analysis models that could be used for predicting the behaviour and characteristics of current and/or new customers, which enables efficient targeting in marketing campaigns. This paper presents a proposed customer profiling demonstrator by incorporating the RFM analysis into data mining techniques to provide marketing intelligence. Initialising the study by performing the RFM analysis draws attention to the importance and advantages of this analysis. In order to evaluate the proposed demonstrator and show the benefit of using it in direct marketing, a customer dataset was simulated containing 50 000 customers, with purchasing behaviour over a two-year period. The results received consider several parameters together, such as the customer clusters, the *cRFM* values of the customers, as well as potential future customer behaviour. Industrial engineers, with their understanding of systems and system integration as well as analytical

knowledge, should find these challenges exciting and relevant in our modern world. With the fourth industrial revolution upon us, who better to lead it than the adaptable and flexible industrial engineers.

REFERENCES

[1] **Walters, M. and Bekker, J.** 2017. Customer super-profiling demonstrator to enable efficient targeting in marketing campaigns, *South African Journal of Industrial Engineering*, 28(3), pp. 113-127.

[2] **Hosseini, M. and Shabani, M.** 2015. New approach to customer segmentation based on changes in customer value, *Journal of Marketing Analytics*, 3(3), pp. 110-121.

[3] **Brown, O**. 2016. The Importance of Customer Profiling, http://www.hellostarling.com/the-importance-of-customer-profiling/. Accessed 21 May 2018.

[4] **Shaw, M., Subramaniam,W., Tan, W. and Welge, M**. 2001. Knowledge management and data mining for marketing, *Decision Support System*, 31(1), pp.127-137.

[5] **Chen, T. and Chen, C.** 2010. Application of data mining to the spatial heterogeneity of foreclosed mortgages, *Expert Systems with Applications*, 37(2), pp.993-997.

[6] **Weinstein, A.** 2013. *Handbook of Market Segmentation: Strategic Targeting for Business and Technology Firms*, 3rd edition, USA: Taylor & Francis.

[7] **Jansen, S.** 2007. Customer Segmentations and Customer Profiling for a Mobile Telecommunications Company Based on Usage Behaviour: A Vodafone Case Study. Masters thesis, Maastricht: University of Maastricht.

[8] **Chan, C.C.H.** 2008. Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer, *Expert systems with applications*, 34(4), pp.2754-2762.

[9] **Bose, I. and Chen, X.** 2009. Quantitative models for direct marketing: A review from systems perspective, *European Journal of Operational Research*, 195(1), pp.1-16.

[10] **Jonker, J.-J., Piersma, N. and Van den Poel, D.** 2004. Joint optimization of customer segmentation and marketing policy to maximize long-term profitability, *Expert Systems with Applications*, 27(2), pp.159-168.

[11] **Chan, C.** 2008. Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer, *Expert System Application*, 24(4), pp. 2754-2762.

[12] **Kim, S., Jung, T., Suh, E. and Hwang, H.** 2006. Customer segmentation and strategy development based on customer lifetime value: A case study, *Expert System Application*, 31(1), pp. 101-107.

[13] **Sarvari, P. A., Ustundag, A. and Takci, H.** 2016. Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis, *Kybernetes*, 45(7), pp. 1129-1157.

[14] **Dursun, A. and Caber, M.** 2016. Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis, *Tourism Management Perspectives*, 18(1), pp.153-160.

[15] **Wei, J.T., Lin, S.Y. and Wu, H.H.** 2010. A review of the application of RFM model, *African Journal of Business Management*, 4(19), pp.4199-4206.

[16] **Hsieh, N.C.,** 2004. An integrated data mining and behavioral scoring model for analyzing bank customers, *Expert systems with applications*, 27(4), pp.623-633.

[17] **Nimbalkar, D. D. and Shah, P.** 2013. Data mining using RFM Analysis, *International Journal of Scientific & Engineering Research*, 4(12), pp.940-943.

[18] **MathWorks**.2018. Kmeans, https://www.mathworks.com/help/stats/kmeans.html#bueftl4-1. The MathWorks, Inc. Accessed 24 January 2018.

[19] **Trevino, A**. 2016. Introduction to K-means Clustering. https://www.datascience.com/blog/k-means-clustering. Accessed 23 May 2018.

[20] **Kaufman, L**. and Rousseeuw, P.J. 2009. *Finding groups in data: an introduction to cluster analysis*, USA: John Wiley & Sons.

[21] **Aptè, C. and Weiss, S.**1997. Data mining with decision trees and decision rules, *Future Generation Computer Systems*, 13(219), pp. 197-210.

[22] **Dori, D.** 2002. *Object-process methodology: A holistic systems paradigm*. Germany: Springer Science & Business Media.

[23] **Statistics South Africa,** 2017. *Mid-year population estimates 2017*, http://www.statssa.gov.za/publications/P0302/P03022017.pdf . Accessed 1 October 2017.

[24] **Aggarwal,C.C. and Reddy, C.K**. 2014. *Data clustering: Algorithms and Applications*. USA: CRC press.

[25] **Shih, Y.-Y. and Liu, C.-Y.** 2003. A method for customer lifetime value ranking – Combining the analytic hierarchy process and clustering analysis, *Journal of Database Marketing & Customer Strategy Management*, 11(2), pp.159-172.

[26] **Evaldas, M.** 2017. *Practical use of RFM customer segmentation*, https://stacktome.com/blog/rfm-customer-segmentation . Accessed 14 May 2018.

[27] **Trewartha, D.** 2006. Investigating data mining in MATALAB. Masters thesis, Department of Science, Rhodes University, Grahamstown. http://pppj2012.ru.ac.za/g03t2052/CSHnsThesis.pdf