Identity theft risk quantification for social media users

Nicola Michau

Department of Industrial Engineering

University of Stellenbosch

Study leader: James Bekker

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

M. Eng Industrial

March 2017

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: March 2017

Acknowledgements

"If the only prayer you ever say in your entire life is thank you, it will be enough." – Meister Eckhart

Magriet Treurnicht,
Jerall Toi,
South African Fraud Prevention Services,
Anne Erikson,
Professor Martin Kidd,
Joubert Maarschalk,
PW & Barbara Michau,
my family,
my friends

and especially

Professor James Bekker,

Thank you.

Abstract

The information era has made it difficult to protect and secure one's personal information. One such struggle is that of identity theft, a crime that has caused great suffering to its victims. Offenders guilty of the crime use the identities of their victims for the purpose of entertainment or fraud. Social media has extended the capability of people to interact and share information, but without the appropriate guidelines to protect individuals from becoming victims of identity theft. There is a lack of studies on identity theft and its determinants. The purpose of the research is therefore to assist with the prevention of identity theft by determining the effect that information-sharing on social media has on the risk of individuals becoming identity theft victims. The details of reported identity theft victims were collected from the South African Fraud Prevention Services. Data on individuals' information-sharing habits on social media networks, like Facebook and LinkedIn, was collected via surveys that were sent to a relevant group at the Stellenbosch University. It was found that the two variables, `Age` and `Gender`, were the greatest predictors of identity theft victims. A prediction model was developed that serves as a tool to score individuals as high-risk or low-risk victims according to their attributes and social media information-sharing habits. The findings benefit research on the prevention of identity theft, by raising awareness of the potential risks the sharing of sensitive data on social media has.

Opsomming

Die tegnologiese era het dit moeilik gemaak vir individue om hulle persoonlike inligting te beskerm. Identiteitsdiefstal is ’n voorbeeld hiervan en veroorsaak lyding onder slagoffers. Oortreders, skuldig aan hierdie misdaad, gebruik die identiteite van hulle slagoffers bloot vir vermaak of bedrog. Die vooruitgaan van tegnologie en die totstandkoming van sosiale media het dit vir die mens vergemaklik, om persoonlike inligting te deel sonder die gepaste voorsorgmaatreëls.

Daar is ’n tekort aan inligting rakende studies oor identiteitsdiefstal en die bepalers daarvan. Die doel van hierdie navorsing is om by te dra tot die voorkoming van identiteitsdiefstal, deur die tendense te bepaal in die persoonlike inligting wat sosiale media gebruikers op die netwerke verskaf, vir beide die wat al slagoffers was of nie. Inligting van verklaarde identiteitsdiefstal slagoffers is verkry vanaf die South African Fraud Prevention Services. Steekproefopnames is uitgestuur na relevante groepe in die Stellenbosch Universiteit rekenaar netwerk. Die inligting rakende individue se gewoontes om persoonlike inligting op sosiale media netwerke, soos Facebook en LinkedIn te deel, is verkry van die bogenoemde steekproefopnames. `Ouderdom` en `Geslag` is gevind as die kernbepalers van identiteitsdiefstal slagoffers. ’n Model is ontwikkel wat gedien het as ’n instrument, om individue as hoë- of lae-risiko slagoffers te bepunt, volgens hulle kenmerke en die persoonlike inligting wat hul op sosiale media deel. Die bevindinge dra by tot die navorsing rakende die voorkoming van identiteitsdiefstal deur die bewusmaking van die potensiële risikos, wat gepaard gaan daarmee om sensitiewe inligting op sosiale media te deel.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter serves as an introduction that gives background information on the research question, it states the research objectives and discusses the research strategy.

1.1 Background to the research hypothesis

The information era has introduced great struggles into the lives of individuals. This era has made it difficult to protect and secure one's personal information. One such struggle is that of identity theft, a crime that has caused great suffering to its victims. Offenders guilty of the crime use the identities of their victims to steal money, obtain loans and generally violate the law (Saunders & Zucker, 1999).

In South Africa, as in most countries, the problem arises because most people are naïve when opening online accounts and are markedly careless with their personal information. Social media has extended the capability of people to interact and share information, but without the appropriate guidelines. Details such as identity numbers, contact details and physical addresses are freely available on social media. Through various channels criminals can create an identity book and proof of address, thus having enough information to open a bank account, which will be billed to the original identity document's owner. In some cases, fraudulent accounts are even opened with the data of deceased individuals. South Africa is one of the top three countries internationally with the highest rates of fraud using recycled deceased identities (Alfreds, 2015c). Social media has immensely simplified the ability of obtaining an individual's personal information. Information is exploding on the internet and individuals have lost control over it.

The increase in the volume, velocity and variety of data online has steered society to the age of big data (He *et al.*, 2014). Social media is an important contributor to the big data regime. This enormous amount of data, big data, requires new technologies and architectures to process data and acquire results (Katal *et al.*, 2013). Technologies like Hadoop together with MapReduce and the Hadoop Distributed File System (HDFS) are both options to query and analyse extremely large data sets. Metadata infrastructures like the Resource Description Framework (RDF) provide a general framework which can be used to graphically represent insights emerging from big data.

Big data is a developing field that has the potential to generate results that are extra reliable. The use of big data can lead to more precise, constant and dependable measurements. Better predictions can be made and experiments can be conducted based on data rather than gut feel or intuition.

1.2 Rationale of research

Identity theft is a crucial problem worldwide and an accelerating problem in South Africa. According to a study by tech firm IBM, one billion parts of personal information were lost in South Africa in 2014, costing the country R432 256 000. Unfortunately, cyber thieves increased their deeds in 2015, with costs running up to R465 412 000 in July. The challenge for society is to minimise the chance of becoming part of these statistics (Alfreds, 2015b).

According to Reyns & Henson (2015) there is a major lack of studies on identity theft and its determinants. Reyns & Henson (2015) state, 'Considering the possible link between online activities and identity theft, research is needed to identify risk factors for online victimization'.

The purpose of the research is therefore to assist the country and individuals with the prevention of the cybercrime, identity theft, and to raise awareness of the potential risks the sharing of certain sensitive data on social media might have.

If the research results in a definite conclusion and social media vulnerability can be connected to identity theft, the benefits expected from the research to science or society will be to raise awareness of the potential risks the sharing of certain sensitive data on social media might have. Recommendations for the managing of social media accounts can then be made.

1.3 Research hypothesis

People are not cautious when sharing personal information on social media platforms. Personal details such as identity numbers, contact details and physical addresses are posted without second thought. Individuals are thus unaware of the effect their social media interaction has on the risk of them becoming identity theft victims.

It is anticipated that there should be a recognisable difference between the amount of data that is shared by identity theft victims compared to that of people who have not been offended with the crime.

Furthermore it is hypothesised that the attributes commonly found in historic identity theft victim cases are the attributes that will serve as important predictor variables in a model that classifies individuals as high-risk or low-risk victims of identity theft.

1.4 Aim and objectives

In order to address the research hypothesis effectively, the following objectives must be met:

1. Determine the attributes that have noteworthy correlations with victims of identity theft.

2. Develop a method to estimate vulnerability scores for individuals based on the data they have revealed on social media.

3. Build a prediction model that best classifies identity theft victims.

4. Determine the variables that best predict identity theft victims.
5. Use the prediction model to score new data as either at high risk of identity theft or at low risk according to their social media information-sharing habits.

1.5 Proposed research methodology

Data on actual identity theft incidents will be collected. The data must contain personal attributes that describe the victims. Data on individuals' sharing of information activities on social media networks, like Facebook and LinkedIn, will then be collected via the method of web crawling. Data will be ingested onto the Hadoop Distributed File System (HDFS) and then processed and cleaned with MapReduce. Hadoop is an open source framework that provides a shared storage and analysis system. The storage is provided by HDFS and the analysis by the programming paradigm, MapReduce. The previously offended victims' attribute outcomes will be used to determine predictor variables that will serve as reference variables throughout the processing of the data. To group and visually graph this information, a technique called the Resource Description Framework (RDF) will be used.

The RDF is a general purpose language for representing information on the web. One of the main applications of the RDF is the integration of data. Data is structured in graphs with vertices and edges. The format of the RDF data model enables the model to be easily reconstructed compared to the complex reconstruction of a relational model. A query language such as SPARQL can then be used to manipulate and retrieve data stored in RDF format.

The data on individual's social media sharing habits will be grouped with RDFs and used to estimate the vulnerability of these individuals based on the amount and type of personal data they have shared on social media. The Hadoop project, Jena, will then be used to compare the data for previous identity theft victims and the data for non-victims to determine if there is a significant difference in the type of data shared by the two groups.

The data on individual's social media sharing habits will then be used to build prediction models that classify individuals as either high- or low-risk identity theft victims and to determine the variables that best predict identity theft victims. SPARQL will be used to mine the RDF models and to build the prediction models. The data analysis results found with the models will be used to accept or reject the hypothesis.

1.6 Research proposal summary

In conclusion it is determined that there is a lack of studies done on identity theft and its determinants. Social media is a big contributor to the amount of individuals' personal information available online, which assists perpetrators with the crime of identity theft. It is therefore planned to conduct a study on identity theft and social media literature. Data will be collected on historic identity theft victims, and the social media sharing habits of individuals. The data will be used to build a prediction model and the results will assist society with the identification of identity theft determinants and be used to classify individuals as high- or low-risk victims.

Chapter 2

Literature Study

This chapter gives an overview of the literature relevant to the study. Data was collected from the sources described in Figure 2.1.

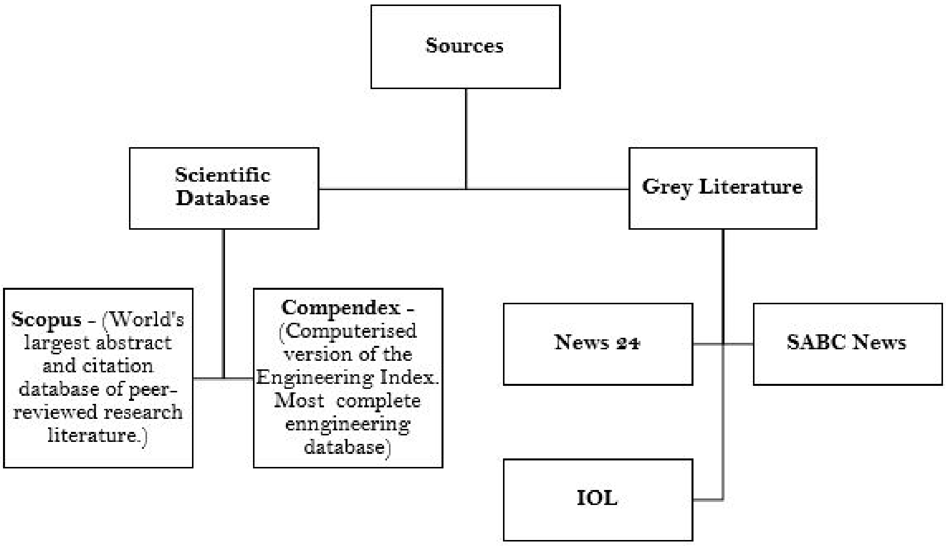

Sources
Scientific Database
Grey Literature
Scopus - (World's largest abstract and citation database of peer-reviewed research literature.)
Compendex - (Computerised version of the Engineering Index. Most complete ennginneering database)
News 24
SABC News
IOL

Figure 2.1: Literature Review Sources.

2.1 Identity theft

Identity theft became a national crime in 1998 in the United States. Today identity theft is an increasing problem worldwide (Reyns, 2013). In South Africa, residents are seen as sitting ducks when discussing the crime. Government, banking and other corporate databases are leaked and are currently being spread by criminals around the world (News24, 2015a).

2.1.1 Identity theft definition

According to Reyns (2013), identity theft is a term used to describe particular crimes, which include the use of an individual's personal information, without their approval or permission, to commit a crime. The Identity Theft Assumption and Deterrence Act agree with Reyns (2013) that identity theft happens when an individual's identity is knowingly used, without consent, to commit an illegal activity. Identity theft is often confused with identity fraud. Reyns (2013) clarifies that identity theft is a category of identity fraud. The ultimate differences between identity theft and fraud are first consent and secondly if the identity is owned by a person (Reyns, 2013). For the purpose of the study, identity theft is defined as the unlawful act when one's personal information is used, without consent, to either commit a crime such as credit theft or for the illegal activity of impersonation.

2.1.2 Crime types

Crimes committed with the personal information of victims include (Reyns, 2013):

- The illegal application for credit;
- Banking fraud like loans;
- Posing as a letting or estate agent to receive deposits (Maluleke & Pheko, 2015);
- Document fraud like a driver's licence; and
- The unlawful application for governmental benefits.

Offenders depend on the good reputation of their victims in order for them to use the benefits that victims qualify for (Dirk, 2015b).

2.1.3 How identities are acquired

According to Reyns (2013) the most common methods of acquiring the identities of victims are: personal data online; phishing; skimming; hacking; and theft of actual identification documents. This study focuses on the stealing of personal data that is shared online.

By knowing how offenders acquire the details of their victims, individuals who develop online spaces can decrease the crime of identity theft by designing these spaces in such a manner that makes these acquisition methods impossible (Reyns & Henson, 2015).

2.1.4 Identity theft cases

Cases of identity theft are progressing due to the information explosion on the internet. Table 2.1 lists identity theft incidents that actually happened in South Africa.

Table 2.1: Cases of Identity Theft.

Type	**Case**
Actual document theft	A resident from Pretoria wanted to open a First National Bank account, but to his surprise, an account had already been opened in his name and with his unique identity number. It was determined that the offender used his official identity document, which he had lost four years prior to the event. (Makhubu, 2015)
Actual document theft	The stolen identity of an individual was used to take out a life insurance policy worth one million rand. After a few months had passed, it was then claimed that the individual had died. A corpse was bought for R20 000 from the King Edward Hospital to obtain a death certificate. (Hlophe, 2015)
Personal data online	Perpetrators used stolen identities from dating sites to buy and ship illegal products. The fraud was discovered after a woman reported the receipt of an suspicious package from a person she met online. The package enclosed a request that she had to reship it to Pretoria. (Cronje, 2015)
Fake letting agent	A Cape Town local paid a deposit to rent a house and signed a lease, but on arrival found out that the property belonged to someone else and the letting agent, to whom he had paid the deposit and with whom he had confirmed the lease, was a duplicate of the official agent and not the real agent himself. (Maluleke & Pheko, 2015)
Fake estate agent	When a woman tried to sell her home, she was astonished to find out that someone, posing as her, had already sold her home. The criminal went to the conveyancer and signed documents under her name. It is not known how the offender acquired her details. (Barbeau, 2015)
Recruit-ment Scam	Criminals posing as the company Netcare 911 sent out messages to the public that Netcare 911 was offering training for paramedics. When individuals then applied to the advertisement, their personal details were stolen and they were required to make payments to the fraudulent account. (News24, 2015b)

2.1.5 The impact of identity theft on South Africa

It is of foremost importance to recognise that identity theft is not only a problem in South Africa, but an increasing issue globally. According to SABC (2015), the estimated global cost of identity theft is 450 billion dollars per year. It is found that in Canada out of 100 000 individuals, 11.5 % were identity theft victims in 2013 and of the entire United Kingdom in 2012, 8.8 %, costing 3.3 billion pounds, were found to have been victims. Statistics show that in 2012, 7 % of all families in the United States were already identity theft victims and this number has increased ever since (Reyns & Henson, 2015).

In South Africa identity theft has escalated by more than 200 percent over the past six years and prime victims are men from Gauteng and KwaZulu-Natal between the ages of 28 and 40 (Dirk, 2015b) and (Erasmus, 2015). Ngwenya (2015) and (Erasmus, 2015) declare that identity theft costs the South African economy over 1 billion rand a year, but Dirk (2015a); Mkheze (2015) reported only a few months later that the same figure stands at 2 billion rand a year. The conclusion however is that the South African economy, which is already suffering, is losing a substantial amount of money to identity theft yearly.

2.1.6 Discovering you are a victim

Many victims learn that their identities have been stolen when applying for credit. It is a clear warning that one is a victim when a debt collector calls in terms of an outstanding balance, which one has no record of (Dirk, 2015a). According to Carol McLoughlin, the South African Fraud Prevention Services (SAFPS) spokesperson, individuals only discover they are identity theft victims once they are notified by credit bureaus that they have been blacklisted for not paying accounts, which were probably opened without their consent (Dirk, 2015b). Recovering from such a crime is frequently worse than discovering it. Victims must prove their innocence by verifying that transactions were not done by them, they must be removed from being blacklisted and sometimes even change their identity numbers (Erasmus, 2015).

2.1.7 Identity theft prevention

The Protection of Personal Information Act gives legal right to confidentiality and individuals should be aware that unauthorised admission to information regarding a person's education, medical records, financial statements, criminal records, personal information or employment history is prohibited (SABC, 2015). It is therefore important that individuals know their right to privacy and make it their own responsibility to guard their personal information. Recommendations seen as good practice by Dirk (2015b) include the following: never give a password or personal identification number (PIN) telephonically, by email or via fax; do not transport redundant personal information in wallets or purses; avoid doing private banking by using internet cafes or insecure terminals; guard documents containing personal information and be sure to destroy them when these papers are not needed anymore; and regularly check accounts and credit records to notice when strange transactions are made.

There are systems and strategies that have been implemented by authorities to prevent crimes such as identity theft. The following systems and strategies are available in South Africa:

1. Home Affairs established that official identity documents have a coat-of-arms that is tactically positioned as an overlay to distinguish real documents from fake ones (Makhubu, 2015).

2. The Department of Home Affairs upgraded their application process for identity documents from manually to online (Dirk, 2015a).

3. Estate agents must be registered with the Estate Agency Affairs Board and have a valid fidelity fund certificate (Maluleke & Pheko, 2015).

4. Social media accounts like Facebook have optional security settings that limit the group of people who can view your profile (IOL, 2015).

5. Security officer certificates expire every 18 months and security business certificates expire every 12 months to ensure that information is legitimate (Mkhabela, 2015).

6. The Financial Intelligence Centre Act 28 of 2001 was introduced to safeguard bank customers (SAPA, 2015).

7. Certain bank branches, for example Capitec, have installed biometric technology to increase client security (Alfreds, 2015a).

2.2 Social media

Before the internet existed, personal information was far more difficult to collect and associate with individuals. Reyns & Henson (2015) determined that the posting of personal information online is a substantial predictor of identity theft. The following section therefore discusses the internet and social media, the benefits and risks thereof and online sharing of personal data.

2.2.1 The internet and social media

In the last decade, with the development of various electronic devices, internet usage has increased exponentially (Henson *et al.*, 2013). The internet has produced a completely new set of fraudulent practices, which include click fraud, email spam, phishing schemes and identity theft (Becker *et al.*, 2010). Internet usage only became universally popular in the 1990s and therefore literature on the fear of online crimes is scarce up until then (Henson *et al.*, 2013). Today, however, the fear of online crime is vital for investigation.

A big contributor to the data available on the internet is the expansion of social media. Social media began with bulletin board systems years ago, but is commonly known today with the growth of platforms such as Facebook, Twitter and LinkedIn (Gaff, 2014). According to Hamed Haddadi (2010), in 2010 statistical results showed that two out of three people in the United States and the United Kingdom were members of at least one social media platform. Today it has become the norm and almost everyone who has access to the internet joins one or more social media platforms.

2.2.2 Definition of social media

Social media is defined as internet-based platforms for people to meet in virtual communities and share information (Gaff, 2014). Social media includes platforms such as blogs, social networking sites (*e.g.*, Facebook), virtual social worlds (*e.g.*, Second Life), collaborative projects (*e.g.*, Wikipedia), content communities (*e.g.*, YouTube) and virtual game worlds (*e.g.*, World of Warcraft). For the purpose of the study social media is defined as an online platform where users upload data to a personal profile that serves as a description of themselves when communicating with other platform users.

2.2.3 Benefits and risks of social media

Social media platforms have the ability to share information widely and quickly. This can be very valuable in companies or among communities when rapid decisions or updates need to be broadcast. Social media platforms are used for the marketing of products and services as well as communication with clients. It is not limited to plain text files, but rather supports all types of information including videos, images, website links and audio files. The medium is therefore prodigious for advertising.

Social media platforms are an extraordinary contribution to the development of communication technology. News reports on important events, governmental decisions, public warnings, *etc.*, are spread globally in seconds. Today it is possible to get in contact or keep contact with family and friends worldwide. Social media even creates a potential to make new friends with platforms such as dating websites. Job-hunting is simplified with platforms like LinkedIn where individuals and companies load their professional information, successes and requests. The communication of information is endless and above all, at an affordable price.

The downside to the broad and speedy sharing of data, is that it is possible to spread harmful or incorrect information in the same manner. Negative publicity on social media can damage a company's reputation or brand. Destructive information about individuals or cyber bullying can lead to serious consequences like depression. It is important to acknowledge that the hacking of an account is a likelihood. Hacking not only leads to the spreading of phoney information, but also to more advanced complications such as identity theft.

2.2.4 Online sharing of personal data

Social media has made individuals more vulnerable to cybercrime. Facebook is one of the world's biggest social media networks. In March 2013 Facebook launched a new feature, namely the graph search. Graph search is based on semantics and enables users to search for questions written in natural language.

The technique has made it possible to acquire personal information effortlessly. By simply writing a query like "all single ladies, aged 22, who live in Cape Town", the graph search will crawl Facebook and return all individuals satisfying the question (Khan, 2013).

Companies and individuals should have strict policies as to what and how information is presented on social media. The sharing of personal data, on social media platforms, is facilitated due to the minimum bar set by the terms and conditions of these platforms (Gaff, 2014). A study done by Reyns & Henson (2015) concluded that the posting of personal data online is a major predictor of identity theft. The problem with social media platforms is that the designers of the platform take copyright over the information that has been uploaded (Hamed Haddadi, 2010). Once users quit the social network, their personal information is not necessarily deleted, but rather remains part of the platform's data.

Sophos, a security company, did a test on Facebook users to determine their naïvety. The company created a fake profile with the name of 'Freddi Staur' and sent out 200 friend requests to random people, of which 87 were accepted. Results showed that 82 out of the 87 accounts shared personal information like email addresses, physical addresses, dates of birth and employment information (Krishnamurthy & Wills, 2009).

It is important to recognise that users themselves determine the vulnerability of the their personal information on social media platforms. Security and privacy settings are available, but they do not guarantee protection from cybercrimes (Patsakis *et al.*, 2014).

2.3 Big data

It is known that social media is a great contributor to the phenomenon of big data. This section gives a background to what big data is. It describes the benefits and challenges of the field and provides examples of where it has been applied.

2.3.1 The analytics timeline

The idea to use data to guide decision-making is an early concept which was already in use in the 1950s. Due to technology innovation and the increase in data velocity, volume, veracity and variety, an era of 'business intelligence' developed. Information systems were built to organise data. Data was captured and business intelligence technologies were used to query and report it. Data sets were small enough that they were manageable in data warehouses. The processing of data was very time-consuming and the minimum time was spent on analysis. (Davenport, 2013)

In the mid-2000s, internet-based and social networks led to the era of big data (Davenport, 2013). Data became too great in volume, velocity and variety for it to be analysed on a single server. The rate at which data sets started growing became so complex to process, that traditional database management and processing applications were no longer able to handle the large amounts of data (Lee *et al.*, 2015). Therefore a need for more advanced tools and technologies developed. It was found that the solution to the problem was the use of a network that processes batch data across parallel servers. Hadoop is the most favoured distributed processing core technology and will be discussed in Section 2.4 (Lee *et al.*, 2015).

The increased speed of data processing and data analysis has made it possible to use information in decision-making, product creation and service development. Davenport (2013) stated, 'Today it is not just online and information firms that can create products and services from analyses of data. It's every firm in every industry'.

2.3.2 Big data definition

There are several definitions for big data (Schneider, 2012). Boyd & Crawford (2012) refer to big data as a 'poor term'. The reason for this blunt statement being that primarily the word 'big' in the term 'big data' accounts for the volume of the data, but fails to grasp the remaining properties of the term. According to Katal *et al.* (2013) big data implies the following properties:

- **Volume** – The 'big' in 'big data' should actually be improved to something like 'massive', 'gigantic' or 'colossal', because the existing data is measured in zettabytes. The extent to which the volume of big data stretches is far beyond the limit that traditional systems can handle (McAfee & Brynjolfsson, 2012).
- **Variety** – Big data comprises several types of data including raw, structured, semi–structured and even unstructured which is very complex to handle with traditional systems (McAfee & Brynjolfsson, 2012).
- **Velocity** – Big data is directly related to the speed of incoming data and the speed of data flow. Data is continuously moving and this specifically makes it impossible for traditional systems to keep up when it comes to the analysis of big data (McAfee & Brynjolfsson, 2012).
- **Veracity** – The quality of the data has improved immensely over the past years.
- **Variability** – Aspects such as social media cause peaks in data masses. The inconsistency of incoming data is an important feature of big data.
- **Complexity** – The cleaning, sorting, linking and analysis of big data requires a difficult set of skills and entails advanced techniques and technologies when handled.
- **Value** – Trends surface from filtered data and reliable information is captured through queries. The use of big data creates incredible results and thus adds value to businesses.

Big data is generated in real time or it is accumulated over time (He *et al.*, 2014). Companies gain an understanding of information that was not conceivable before this era. Knowledge gained by the analysis of big data is measured and directly interpreted into decision-making (McAfee & Brynjolfsson, 2012). Computerised decision-making is becoming a reality with the exponential growth of data and the capabilities of the tools and techniques available (Elmegreen & Sanchez, 2014). Computers will eventually take over the human component and make most decisions independently, based on data analysis. It is, however, not smart to erase the need for human insight completely. Before a decision is finalised, it should be well appraised by company leaders (Katal *et al.*, 2013).

2.3.3 The benefits and challenges of big data

The greatest benefit of big data is that a department, company or industry can make fact–based decisions on a daily basis (Ross *et al.*, 2013). Big data has the power to measure and manage data more effectively than ever before. Exceptionally large data sets are used in analysis, which clearly increases the reliability of results. Results are based on current data and not historical data, making an organisation much more agile and giving it a competitive advantage (McAfee & Brynjolfsson, 2012). Big data results in better predictions and decisions that are more accurate. A variety of big data projects have been done in many different fields. Table 2.2, extracted from Katal *et al.* (2013), lists a few of these projects under their specific domains.

Table 2.2: Previous Big Data Projects.

Domain	**Project Description**
Science	Large Hydron Collider (worlds biggest and highest-energy particle accelerator) – Data flow comprises 25 petabytes and extends to 200 petabytes after replication. Sloan Digital Sky Survey (multi-filter imaging and spectroscopic redshift survey) – Includes more than 140 terabytes of data and generates data at 200 gigabytes per night.
Government	Obama Administration Project – Involved 84 different big data programs. Community Comprehensive National Cyber Security (for the delivery of cyber security) – Data is stored in yottabytes.
Private Sector	Amazon.com – Possess the three largest Linux databases in the world. Capacities range from 7.8 terabytes to 24.7 terabytes. Walmart – One million customer transactions are processed hourly and more than 2.5 petabytes of data is stored. Falcon Credit Card Fraud Detection System – There are more than 2.5 billion active accounts.
International Development	Information and Communication Technologies for Development – Big data adds to international development by producing fact-based decisions.

There are however many challenges concerning big data. The most significant ones, according to Redman (2013) are:

- **Data Quality** – Up to 50% of employees' time is wasted due to poor quality data. A great amount of time is spent on data cleansing and specifically on searching, identifying and correcting data errors.
- **Data Credibility** – Unreliable data causes managers to lose faith in the data system and forces them to go back to their gut feelings and intuition rather than hard decisions based on information.
- **Privacy and Security** – Personal information regarding users is collected by invasion of their lives. Facts, desired to be kept secret, are collected from a person without their consent.

Katal *et al.* (2013) agrees when Redman (2013) includes privacy and security as a big data challenge, but argues that it is the most crucial issue. Data quality and credibility are included by Katal *et al.* (2013), under technical challenges, but are not the only challenges addressed. The list of core issues according to Katal *et al.* (2013) are:

- **Privacy and Security** – As Redman (2013) suggested, personal data, not necessarily meant for anyone else except the user, is revealed and users are uninformed about the fact that their data is being used to create insights.
- **Data Access and Sharing** – Due to organisations' striving to have a competitive advantage and culture of confidentiality, it is difficult to gain access to certain client data and databases or to have a company agree to sharing their data.
- **Storage and Processing Challenges** – The exceptionally large amount of incoming data, produced by various sources, is too much to store and moves so fast that it would be problematic to upload it in cloud, especially in real time (Barbierato *et al.*, 2014).
- **Analytical Challenges** – The fact that big data consists of raw, unstructured, structured and semi–structured data requires a need for advanced analytical skills.
- **Essential Skills** – Due to the commercial use of big data being fairly new, universities should offer programs to teach the wide range of skills needed to process and analyse such data. Skills include not only technical and analytical skills, but research, creativity and interpretive skills.
- **Technical Challenges** – Machines and software used for the processing of big data are not 100% reliable or fault-proof. Complex algorithms are vital for fault-tolerant computing. The scalability of big data makes it difficult to know when data is sufficient, relevant or accurate enough to extract conclusions from (Barbierato *et al.*, 2014).

2.3.4 The processing and analysis of big data

According to White (2009) the following tools and techniques are available for the management of data:

- **Hadoop and its Components** – Hadoop, together with HDFS and MapReduce, offer a trustworthy joint storage and analysis system. It entails various projects that contribute to its distributed computing ability.
- **High Performance Computing (HPC) and Grid Computing** – Data is distributed across a cluster with a combined file system hosted by a Storage Area Network. The framework relies on Application Program Interface (API) tools such as the Message Passing Interface to control data flow, which becomes very difficult to manage.
- **Volunteer Computing Technique** – Work is broken down into portions and shared among computers across the globe to be analysed and then returned. This technique is hardware-intensive.
- **Relational Database Management System (RDBMS)** – RDBMSs are different in structure and means of analysis when paralleled to MapReduce. Traditional databases only work with data sizes in range of gigabytes. Big data requires databases to deal with data sizes in ranges of petabytes (Lee *et al.*, 2015).

Katal *et al.* (2013) compared the available techniques and found that Hadoop is more user-friendly than HPC and Grid Computing because MapReduce automatically does the tasks users have to control when using APIs. The Volunteer Computing Technique is not as reliable as MapReduce due to the risk of computer hardware failure. RDBMSs lack the ability to manage the size of big data and are thus not suitable tools (Lee *et al.*, 2015). It is therefore concluded that Hadoop, combined with its components, is the best technique for the management of big data.

2.4 Hadoop

In the previous section, it was determined that Hadoop is the preferred technique to manage big data. The following section explains the phenomenon, Hadoop. It provides a brief introduction to Hadoop, what the framework consists of and how it works.

2.4.1 What is Hadoop?

The velocity, volume and variety of big data leads to continuously growing unstructured data files. No single record is predominantly valuable, but having every single record is beyond valuable. The great extent of data created can be of extreme worth, but the challenge remains that it must first be filtered, processed and analysed. It is possible to perform such functions with a framework like Hadoop. Hadoop has the ability to give meaning to an enormous amount of insignificant random data (Lee *et al.*, 2015).

Hadoop, founded by Doug Cutting and Michael Cafarella in 2005, is hosted by Apache Software Foundation (Katal *et al.*, 2013). Hadoop is an open-source framework that offers shared storage and large-scale processing of data. The storage is provided by the HDFS and the data processing by the programming paradigm, MapReduce. Hadoop can store various types of data and execute challenging data analysis (Lee *et al.*, 2015). The framework's ability to rapidly process big data is due to the fact that batch data is distributed among parallel servers as displayed in Figure 2.2 (Lee *et al.*, 2015).

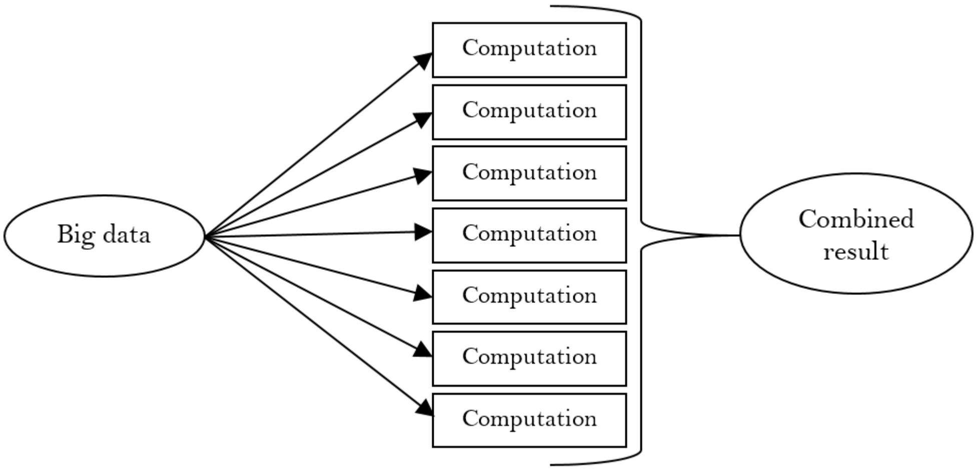


Figure 2.2: Distributed Data in Hadoop.

2.4.2 The Hadoop architecture

A typical Hadoop environment consists of a master node and slave nodes. It is common to find more than one master node in an environment; this is to reduce the risk of a single point of failure. The master node requires elements such as: JobTracker; TaskTracker; and NameNode.

A Hadoop deployment includes several slave nodes. Slave nodes entail a DataNode, which stores data in the HDFS and replicates data across clusters, as well as a TaskTracker. Slave nodes provide a large amount of processing power that has the ability to analyse hundreds of terabytes or even a petabyte of data. The JobTracker element distributes MapReduce tasks to numerous nodes within a cluster. The master node TaskTracker as well as the various slave node TaskTrackers have the ability to receive the MapReduce tasks. The NameNode element stores a directory tree of all the files in the HDFS and keeps an index of where file data is stored among the DataNodes in the cluster. Figure 2.3 is a graphical representation of the communication between the master node and the slave nodes (Schneider, 2012).

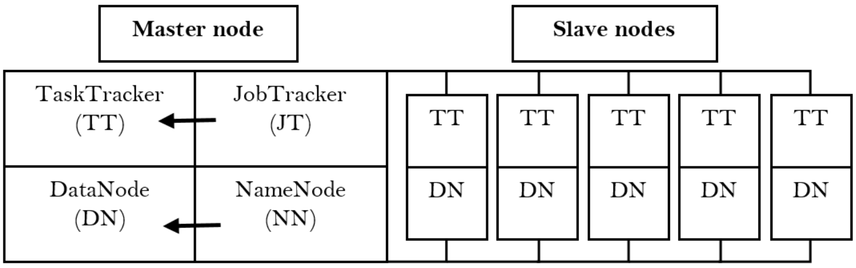


Figure 2.3: Hadoop Cluster.

The Hadoop distributed model is Linux-based and makes use of low-cost computers. Therefore Hadoop was constructed to keep hardware failures in mind. The framework saves three copies of each file by default and these files are spread among different computers (Schneider, 2012).

2.4.3 The Hadoop infrastructure

The Hadoop architecture includes a set of tools generally known as projects. The main subprojects are the HDFS and MapReduce, but the various other subprojects each contribute to a specific area by providing it with higher–level functionality and complementary services (White, 2009). Table 2.3 lists a few of the essential projects according to Wayner (2015).

Table 2.3: Hadoop Projects

Project Name	**Description**
AVRO	Sequential system that packs data with a diagram to make it comprehensible.
FLUME	Online analytical application that collects, aggregates and moves log data.
GIS	Geographic Information Systems (GIS) that handle geographic queries with the use of coordinates.
HIVE	Data warehouse infrastructure used for data summaries, queries and analysis.
HBASE	Open source Java based project, which runs on top of HDFS, that MapReduce jobs can run locally.
JENA	Project which supports the writing of applications that work on RDF data.
LUCENE	Tool that catalogues bulky blocks of unstructured text.
MAHOUT	Project that learns algorithms and provides recommendations established on user's taste.
NOSQL	Data store with particular devices to store data across nodes.
OOZIE	Project that schedules jobs in Hadoop system.
PIG	High-level command project that is responsible for actual computation.
SQOOP	Tool that transforms bulk data between Hadoop and structured data stores.
SQL	Traditional database query language adapted to support Hadoop in quick, ad hoc queries.
SPARK	Framework to support iterative algorithms.

The Hadoop framework is generally used to build recommendation systems, searching tools, for online advertising, market analysis and even to sensor data (Lee *et al.*, 2015). The Hadoop framework is a data lake from which data can be organised, queried, connected and made sense of. The Hadoop framework contributes to the development of the semantic web.

2.5 Semantic web

Hadoop serves as the perfect platform to make sense of an enormous amount of data. Semantic web tools are applied to the Hadoop framework to add further clarity to the data. The following section describes what the semantic web is and the significance of connected data. It discusses applicable tools and techniques and focuses on the Resource Description Framework.

2.5.1 What is the semantic web

The term 'semantic web' was invented by Tim Berners-Lee to describe the future of the World Wide Web (WWW). Information is given expressive meaning (semantics) in such a manner that computers can associate and understand the connection between data from various sources (Mika, 2004).

The web consists of an enormous jumble of data that is very difficult to differentiate if it is unknown how the data interconnects. WWW Consortium (W3C) therefore developed the semantic web infrastructure to create some consensus and eventually a dynamic web of data. The goal of the semantic web model is to integrate data in such a manner that it is consistent throughout the web. If data is connected and organised in a systematic way, more of the available smart web applications will be able to extract the applicable information needed for analysis, and consequently more value will be obtained from data (Allemang & Hendler, 2011).

2.5.2 Semantic web tools and techniques

According to Allemang & Hendler (2011) there are a few general methods to create integrated Web applications. Allemang & Hendler (2011) suggest the following two methods:

1. The first approach is to save data in a relational database. Then let queries run against the database to build websites. If changes or updates need then to be made to the data, they must be done in the database itself. Webpages will then be consistent with information due to them extracting their data from the same source, the mutual database.

2. The second method is to write program code in some general-purpose language like Java, Python or C. The code will connect data in different places and in doing so keep them up to date with changes.

The goal of these approaches is to create an environment where websites are not a collection of pages, but rather a collection of data. The interconnected information can then be queried and presented as it is needed and websites will have the ability to change dynamically as required (Allemang & Hendler, 2011).

2.5.3 Connected web of data

Currently the web offers a distributed network where web pages are universally connected with links called Uniform Resource Locators (URLs). Websites that are more refined use their own structure. They are backed up with a database or Extensible Markup Language (XML) that guarantees that information remains consistent. With the semantic web the desire is that the whole network follow the structure where data items are interconnected with Uniform Resource Identifiers (URIs) links. The connection of data items (URIs) replaces that of webpages (URLs). The semantic web uses the Resource Description Framework (RDF) to graph and present the connected web of the data (Allemang & Hendler, 2011).

2.5.4 Resource Description Framework

The Resource Description Framework is a data model that uses a general–purpose language to present data in the web. A RDF is constructed with a set of triplets that each contain a subject, predicate and object (Alkhateeb *et al.*, 2012). The subject and predicate of a triple entail URIs, and an object involves literal values or blank nodes. Data is organised in graphs with edges and vertices as seen in Figure 2.4.

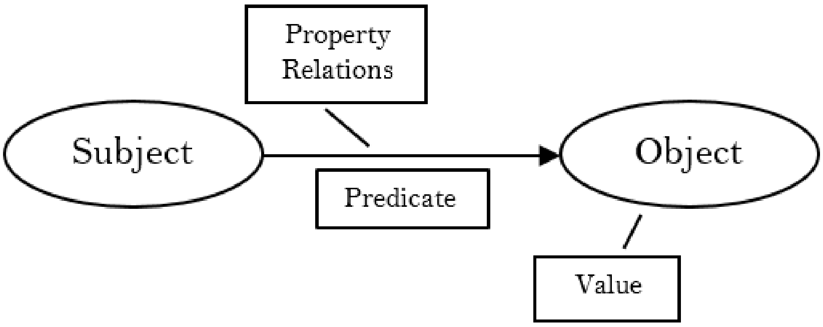


Figure 2.4: RDF Triple.

The prime use for RDFs is the integration of data. Figure 2.5 is a demonstration of how a relational database will look in RDF format. When RDFs are compared to relational databases, it is found that RDFs are much easier reconstructed to fit different types of queries. RDFs have been widely adopted by companies because of their simple model. Examples of companies that are available in RDF format are Wikipedia's RDF image called dbpedia and Facebook's format called Open Graph Protocol (Allemang & Hendler, 2011).

	Predicate (Property)	…
Subject	Object (Value)	
…		

Figure 2.5: Adaption of a Relational Model to RDF.

2.6 Data mining

Rohanizadeha & Moghadam (2009) reported that data mining is frequently explained as the process by which technologies such as pattern recognition technologies and statistical and mathematical techniques are used to find trends and relationships among variables in large amounts of data.

2.6.1 Data mining models

According to Martins *et al.* (2016); Shafique & Qaiser (2014b), the three most widely used data mining models are: Knowledge Discovery in Databases (KDD), Sample, Explore, Modify, Model and Access (SEMMA) Methodology and the Cross Industry Standard Process for Data Mining (CRISP). The three methods are therefore discussed in the following sections.

2.6.1.1 Knowledge Discovery in Databases Model

The Knowledge Discovery Database (KDD) model is the process by which the knowledge hidden in databases is obtained. The process is repetitious and interactive in nature. (Shafique & Qaiser, 2014a)

The nine phases of the model, presented in Figure 2.6 are described as follows:

1. **Developing and Understanding of the Application Domain** – customer requirements are determined and transformed into goals.

2. **Creating a Target Data Set** – the required data set is created and sampled.

3. **Data Cleaning and Pre-processing** – in this phase data is cleaned and pre-processed to remove noise and inconsistencies.

4. **Data Transformation** – data transformation techniques are used to transform data into a suitable format for model building.

5. **Choosing the Suitable Data Mining Task** – data mining tasks such as classification, regression, clustering, *etc.* are identified.

6. **Choosing the Suitable Data Mining Algorithm** – data mining algorithms that best suit the data set and objectives are selected.

7. **Employing the Data Mining Algorithm** – data mining algorithms are applied.

8. **Interpreting Mined Patterns** – data mining algorithms are evaluated and it is determined what they portray.

9. **Using Discovered Knowledge** – knowledge gained is shared and applied in practice.

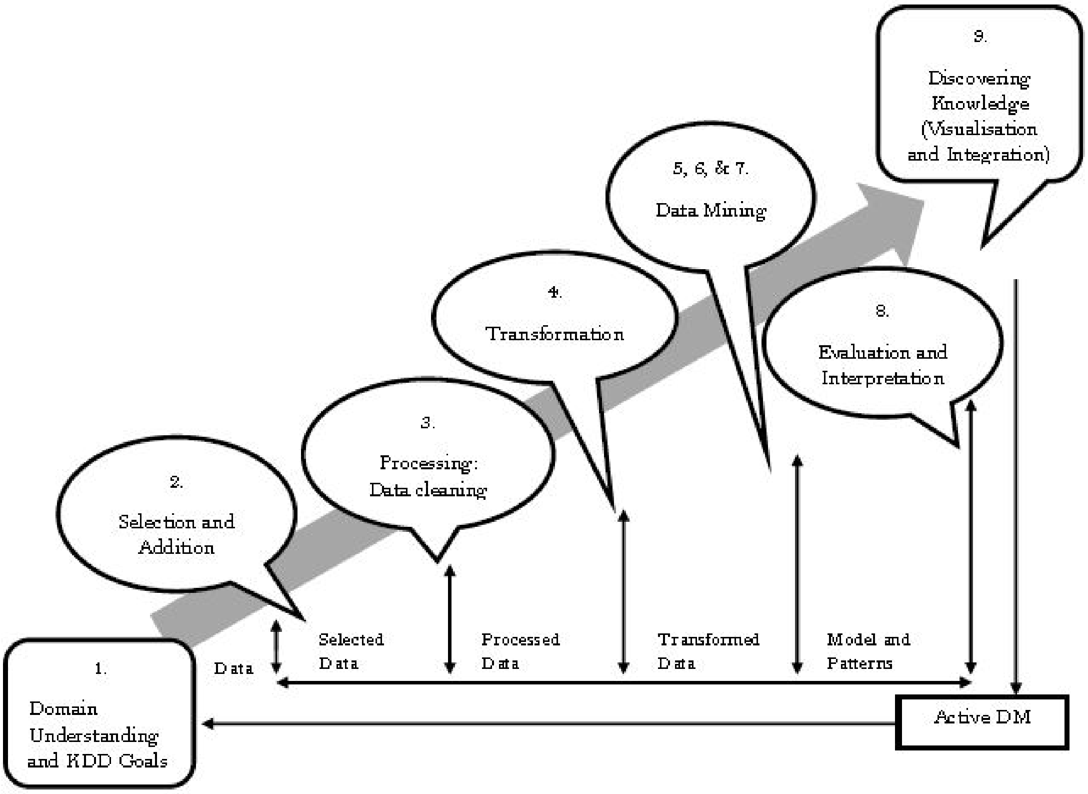


Figure 2.6: KDD Process Cycle (Shafique & Qaiser, 2014a).

2.6.1.2 SEMMA Methodology

SEMMA is a data mining method that was first proposed by the SAS institute, which is a leading company in the development of statistical software applications (Rohanizadeha & Moghadam, 2009). According to Shafique & Qaiser (2014a), the SEMMA cycle assists in the solving of business problems and the reaching of business goals.

The acronym SEMMA stands for Sample, Explore, Modify, Model and Access, which are the names of the five phases included in the cycle. The SEMMA Cycle's five phases are described as follows (Shafique & Qaiser, 2014a):

1. **Sample** – a sample of data is removed from a large data set. The sample must be large enough to produce significant information, but small enough for rapid manipulation. This phase is optional.

2. **Explore** – explore data to determine trends and anomalies that exist in the data.

3. **Modify** – determine outliers and screen data for variable removal. Modify the data in order to simplify the model selection process.

4. **Model** – in this phase the data is modelled. Modelling techniques are applied according to the data type and situation.

5. **Access** – evaluation of model performance and applicability of findings.

2.6.1.3 The Cross–Industry Standard Process Data Mining Model

The Cross–Industry Standard Process is used for data mining. CRISP was first suggested in the 1990s by a European consortium of companies as a standard process model for data mining. The model was non–proprietary at the time. CRISP defines the steps of data mining and is relevant to any industry. The model clarifies what must be done and contributes to the speed, reliability and efficiency of projects Electronic Version: StatSoft (2013).

The CRISP Cycle has six phases as seen in Figure 2.7:

1. **Business Understanding** – define goals of project, define what data indicates, construct possible questions, determine business objectives, create a project plan and finalise hypothesies.

2. **Data Understanding** – collect all data, explore data using graphs or basic statistics and determine what relationships exist in the data.

3. **Data Preparation** – clean data. Remember that this phase is very time-consuming.

4. **Modelling** – apply the several predictive data models that are available.

5. **Evaluation** – review models and determine which model or collection of models satisfies or answers the business goals and objectives best.

6. **Deployment** – score new data with models.

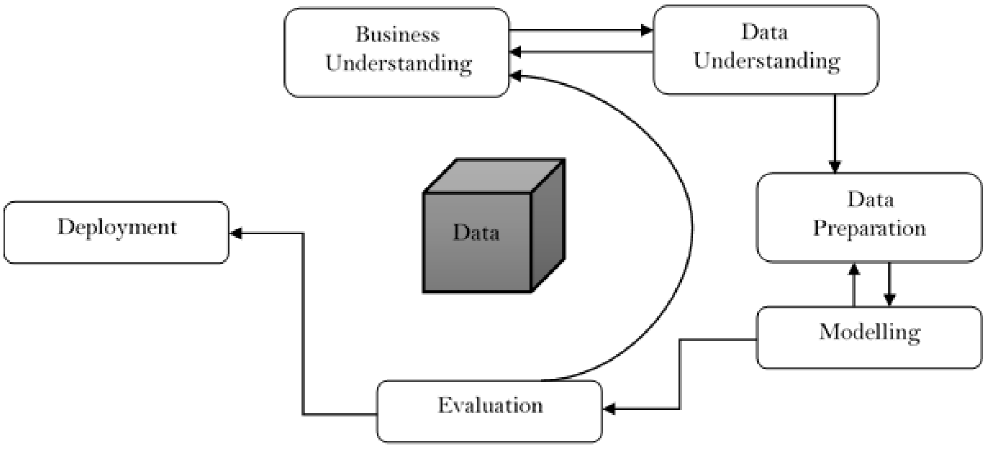


Figure 2.7: CRISP Cycle (Electronic Version: StatSoft, 2013).

2.6.2 Data mining techniques

There are several techniques available in data mining. It is however, important to note that only specific techniques are applicable to certain data types. Some of the available techniques were summarised by Rohanizadeha & Moghadam (2009) as:

1. **Traditional Statistics** – These techniques include cluster analysis, discriminant analysis, logistic regression and time series forecasting (Han & Kamber, 2001).

2. **Induction and Decision Trees** – Classification and Regression Trees (CART), Chi–squared Automatic Interaction Detector (CHAID), Exhaustive CHAID, Quick Unbiased Efficient Statistical Tree (QUEST), Random Forest Regression and Classification and Boosted Tree Classifiers and Regression are six of the different tree data mining methods investigated in Sut & Simsek (2011).

3. **Neural Networks** – According to West (2000), the multilayer perceptron network is the most frequently used in literature.

4. **Data Visualisation** – In Soukup & Davidson (2002) there is a chapter on data visualisation tools such as column and bar graphs, distribution and histogram graphs, box graphs, line graphs, scatter graphs, tree visualisations and map visualisations.

2.6.3 Data mining functions

According to Han & Kamber (2001), the following data mining functions are frequently required:

1. **Description and Summarisation** – is when the data is investigated to determine certain characteristics . Techniques commonly used for this are basic descriptive statistics and data visualisation tools (Rohanizadeha & Moghadam, 2009).

2. **Concept Description** – is the process taken to describe and understand data classes and to determine what the data represents. Techniques such as clustering and induction are frequently used for this function (Rohanizadeha & Moghadam, 2009).

3. **Segmentation** – is when data is sorted into groups according to similar characteristics. Techniques fitted for this function are clustering, neural networks and data visualisation (Rohanizadeha & Moghadam, 2009).

4. **Classification** – uses techniques such as discriminant analysis, induction and decision trees, neural networks and genetic algorithms to group data according to known classes (Rohanizadeha & Moghadam, 2009).

5. **Prediction models** – forecasts unexplored continuous data according to a certain determined class. Techniques used to build these models include neural networks, regression analysis, regression trees and genetic-algorithms (Rohanizadeha & Moghadam, 2009).

6. **Dependency Analysis** – determines the dependencies among data points (Rohanizadeha & Moghadam, 2009). According to Han & Kamber (2001) there are two essential dependency analysis techniques: association and sequential patterns.

2.7 Literature review summary

The Literature Review discussed the topic of identity theft, the impact thereof on South Africa, the different crime types, how offenders manage to acquire victims' data, and how identity theft can be prevented. The study focuses on online identity theft, specifically via social media platforms. Social media platforms were therefore discussed, the benefits and risks thereof and the observed sharing of personal information habits of individuals. Social media contributed to the era of big data, which produced the next topic discussed. Big data was discussed along with the popular big data analysis paradigm, Hadoop. The description of the large amounts of data online led to the topic of the sematic web and its preferred data model, the RDF. To make sense of all the data, data mining is required. Information on data mining models, techniques and functions was therefore collected and summarised as the final section of the Literature Review.

In the next chapter the required data will be determined and then collected. A preliminary basic statistical analysis will then be done on the data to establish if the data fits the study.

Chapter 3

Data Acquisition and Preliminary Investigation

If you do not know how to ask the right question, you discover nothing.
- W. Edwards Deming

This chapter stipulates what data was needed to execute the research study, the required data sources, the methods that were used to collect the information and a preliminary investigation of the data.

3.1 Data required for research study

The purpose of the study was to identify the attributes that were commonly observed in a historic data set of known identity theft victims, then to analyse the social media information-sharing habits of a large group of people and determine if there was a significant difference between the variables and habits of previous identity theft victims compared to non–victims. The final goal of the research study was to build a predictive model, which best classified identity theft victims in order to determine if an individual was at risk of being an identity theft victim. The following information was therefore required:

1. Information on recorded identity theft cases that included descriptive attributes of the victims; and

2. A data set that contained examples of personal data, including the related variables found in the first set of data, that is publicly shared on social media profiles.

The study was limited to South African identity theft cases and therefore the data collected was that of South African citizens only.

3.2 Identification of data sources

In this section the data sources that were needed to obtain the necessary data from, were identified.

3.2.1 Data source: historic identity theft cases

The first fundamental data collected was that of reported identity theft victims. In South Africa identity theft incidents are reported to the South African Police Service (SAPS) or the SAFPS). Hillcrest PI Rick Crouch said in Erasmus (2015) that the SAPS is inadequate to handle identity theft types of crimes. Due to the possibility of insufficient or incorrect information received from the SAPS, it was decided to rather source the necessary data from the SAFPS. The SAFPS is the top commercial reference in fraud prevention in South Africa (SAFPS, 2016).

3.2.2 Data source: social media sharing habits

The second set of information collected was on individuals' personal information-sharing activities on social media. The initial idea was to obtain this information by crawling the relevant social media sites. Many social media platforms have features where users can build applications, like Facebook's Graph API, discussed in Weaver & Tarjan (2013), to collect certain information from the site's database. The privacy policies of other social media platforms, such as LinkedIn, however prevents this data collection method as it publicly states that it is illegal to crawl the site. The privacy policies of the social media platforms discussed in this study are summarised and included in Appendix A. The method of crawling to collect the required information would therefore exclude certain social media platforms due to legislation. The study however focuses on social media platforms relevant in South Africa. It is therefore important that all of the most popular social media platforms in South Africa were incuded in the research.

MyBroadband (2015) discussed the results presented by We Are Social's 'Digital, Social, and Mobile in 2015' report on South Africa's most popular social media platforms in 2015. Table 3.1 lists the social media platforms according to the most popular, calculated by the percentage of the nation that is subscribed to the platform.

Table 3.1: 2015's Most Popular Social Media Sites in South Africa (MyBroadband, 2015).

Social Media Platform	**% of Nation Active**	**Millions of users**
WhatsApp	31%	16.74
Facebook	26%	14.04
Facebook Messenger	19%	10.26
Google+	15%	8.1
Twitter	13%	7.02
LinkedIn	12%	6.48
Skype	11%	5.94
Pinterest	9%	4.86
Instagram	8%	4.32
WeChat	7%	3.78

It can thus be seen that the method of crawling is not feasible for the problem. LinkedIn is one of the vital social media sites for investigation and would have to be excluded if information was gathered via crawling.

It was decided to rather gather the data by making use of surveys. A study conducted by Kaplowitz *et al.* (2004) concluded that if members of a population all have internet access, online surveys are suggested as it was found that when compared to hard copy questionnaires, online surveys achieve better response rates. In Bryman *et al.* (2014) two types of online surveys were discussed; email surveys and web-based surveys. Email surveys send a questionnaire directly to potential participants via email. Web-based surveys present respondents with a link that directs them to a website with a questionnaire. According to Ilieva *et al.* (2002), email surveys achieve a much better response rates than web-based surveys.

The data was therefore collected with a survey that was sent via email to the population of Stellenbosch University (SU). The SU population included all students and staff members. The reasons for selecting the SU population were the following:

- The more information gathered, the better the results. SU is a very large academic institution with more than 30,000 students enrolled for 2016. This yields a potential survey of 30,000 participants, excluding staff members.
- All individuals enrolled at the Stellenbosch University have access to computers and the internet. Therefore every individual has the opportunity to subscribe to social media platforms.
- The minimum age requirements for social media platforms vary from 13 to 18 years (Bennett, 2014). It was decided that the target group for the surveys will therefore be a group consisting of individuals with ages in the range of 18 to 80 years. The minimum age limit of 18 years was to ensure that no social media site was excluded and the maximum age of 80 years is due to the fact that most individuals are retired at that age. The identified population of Stellenbosch University caters for the determined target group.
- The research study is done in South Africa and is based on South African identity theft cases. The population is therefore applicable, because SU is located in South Africa.
- Tertiary educated people are assumed to be intelligent and are therefore expected to be cautious on social media. The population would therefore test the actions of wise people. The results on an uneducated population would be expected to be much less cautious.
- A drawback of the population is that the majority of the population are young people. The staff members, who are generally older, consist of a much

smaller proportion. It was however decided that the population is still valid, because the majority of social media users are younger people.

3.3 Data collection methods

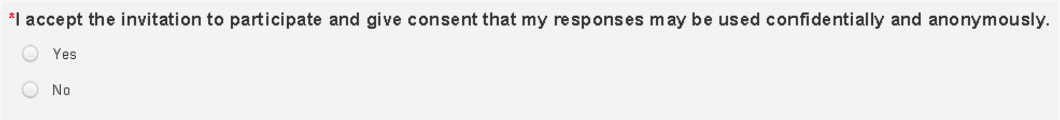
This section describes how permission was granted to acquire the required data and how the data collection process commenced.

3.3.1 Data collection: SAFPS

In order to obtain ethical clearances and receive information from the SAFPS a proposal was constructed. The proposal explained how the data will contribute to the research study and it specified how the data will be processed and presented. The SAFPS accepted the research study proposal and agreed to share the indicated information needed for the research. It was agreed that the information of the various identity theft victims were to be kept confidential. Therefore only general, depersonalized information that could not be tracked to the victims was received. The information included the age, gender, marital status and income of all identity theft incidents reported in the year 2015. No data was connected to victims' personal identification.

3.3.2 Data collection: email survey

The second set of data was collected via a survey that was directly emailed to individuals from the population group Stellenbosch University. The researcher had to follow the Stellenbosch University's ethics clearance protocol. The data acquisition only began after institutional permission was received and the Faculty Ethics Screening Committee classified the research study as a low-risk study. The introduction of the survey was of such nature that potential participants first needed to read the purpose of the study and then give consent to participate in the survey as illustrated in Figure 3.1.



*I accept the invitation to participate and give consent that my responses may be used confidentially and anonymously.

○ Yes

○ No

Figure 3.1: Survey Page One, Consent Question.

The first part of the survey asked participants to select the social media platforms that they were subscribed to. It was then required of them to provide the personal attributes which they publicly shared on these various sites. Before the survey could be created, it was essential to first determine the social media sites and attributes that were to be included in the survey.

It was decided to include the social media platforms that were most popular in the country at the time. The sites listed in Table 3.1 were the most recent statistics for social media platform subscriptions. For the purpose of the study, as recalled from the Literature Review Chapter, social media was defined as an online platform where users can upload data to a personal profile that serves as a description of themselves when communicating and sharing information with other platform users. Facebook, LinkedIn, Twitter, Instagram and Pinterest were therefore included in the survey. WhatsApp, Facebook Messenger, Google+, Skype and WeChat were according to this definition not seen as social media platforms as they are mainly communication mediums and were therefore excluded from the research.

It was decided to include Mxit, as it is a South African developed social media platform and YouTube, because Stellenbosch University students have free access to the platform via the SU's intranet. Dawid Jacobs, Independent Identity Verification expert, stated that recently, more people have become identity theft victims through dating websites (Alfreds, 2015b). Dating sites, which refers to all dating platforms as a whole, were therefore the final platform included in the survey.

The social media platforms that were included in the survey are presented in Table 3.2 along with the feedback categories the survey would yield.

Table 3.2: Survey Page Two, Part One, Platform Subscription Numbers.

Social Media Subscription Numbers		
Social media platforms:	Subscribed:	
Facebook	Yes = 1	No = 0
LinkedIn	Yes = 1	No = 0
Twitter	Yes = 1	No = 0
Instagram	Yes = 1	No = 0
YouTube	Yes = 1	No = 0
Pinterest	Yes = 1	No = 0
Mxit	Yes = 1	No = 0
Dating Sites	Yes = 1	No = 0
Other	Yes = ”Specify”	No = 0

The second part of the survey requested participants to specify what attributes from the list in Table 3.3 they have shared on which of the afore-mentioned social media sites. The 15 attributes, incorporated in the survey, were determined after investigating:

1. the privacy policies of the included social media sites that are summarised in Appendix A;
2. the personal profile options provided by the different social media sites; and
3. the minimum information these sites requested to create social media accounts.

Table 3.3: Survey Page Two, Part Two, Shared Attributes.

Attributes of participants that are revealed on social media platforms		Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10
		No Platform	Facebook	LinkedIn	Twitter	Instagram	YouTube	Pinterest	Mxit	Dating Sites	Other
X1	Name	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0
X2	Surname	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0
X3	Identity number (pictures of an ID doc, etc.)	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0
X4	Birthday (day, month and year)	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0
X5	Gender	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0
X6	Race	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0
X7	Physical address	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0
X8	Email address	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0
X9	Cell number	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0
X10	Relationship status	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0
X11	School details	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0
X12	University/College details	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0
X13	Job details	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0
X14	Income	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0
X15	Credit card details	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0	Yes = 1, No = 0

The information gathered in the third and final part of the survey is presented in Table 3.4. Participants had to provide their age, gender, relationship status and monthly income. These attributes were requested because they were the attributes that were captured and received from the SAFPS victim cases. The attributes were therefore necessary to compare the two data sets. Participants were then asked to state whether they had been a victim of identity theft or not. If they answered 'Yes' to the question they had to select which type of identity theft crime they were a victim of, where they could select one or multiple answers from a given list. After investigating several identity theft crime cases in the Literature Review, the most common crime types boiled down to the following options:

1. Existing Account Fraud (credit or debit card fraud);
2. New Account Creation (apply for credit cards, mortgages, phone service, rent an apartment, buy or lease a car, *etc.*);
3. Tax Identity Theft (file fraudulent tax returns);

4. Criminal Impersonation (offender using a person's identity to hide theirs); and

5. Other (type not mentioned above).

Table 3.4 shows the crime type options survey participants could choose from if they had been identity theft victims and how the responses were recorded.

Table 3.4: Survey Page Three, Personal Details.

Attributes	**Possible Survey Responses**					
Age	Numerical value between 18 - 100					
Gender options:	Male Yes = 1, No = 0	Female Yes = 1, No = 0				
Relationship status options:	Married Yes = 1, No = 0	Divorced Yes = 1, No = 0	Single Yes = 1, No = 0	In a relationship Yes = 1, No = 0	No status Yes = 1, No = 0	
Monthly income	Numerical value					
Identity theft victim options:	No = 0	Yes - Existing Account Fraud = 1	Yes - New Account Creation = 2	Yes - Tax Identity Theft = 3	Yes - Criminal Impersonation = 4	Yes - Other = 5

According to a study done by Wiley *et al.* (2009), there are three effective methods to encourage internet-based survey responses: survey follow-ups, pre-notification of the survey and incentives. In order to encourage individuals to participate in the survey an incentive of three prizes, worth R1 000 each, were therefore arranged to be given away to three participants via a lucky draw. Participants who wished to compete in the lucky draw, had to provide their email addresses. When the option 'Yes', to compete in the lucky draw, was selected in the survey, a secondary survey commenced that retrieved the email address of the respondent and included it in a database from which the lucky draw winners were chosen in a depersonalised manner. This was done to keep email addresses anonymous and to satisfy the privacy controls requested by the Research Ethical Council of SU.

The functioning of the survey was first tested with a pilot test that was sent to a group consisting of 11 family members and friends. After careful consideration and a few alterations to the design of the survey, the survey was sent out in May 2016 to 35,808 Stellenbosch University members falling into the groups: staff members, postgraduates, undergraduates except first and final years, final years, and first years. The surveys were sent out in batch groups to streamline the process and to avoid overloading the server. Participants were allowed two weeks to respond to the survey.

The survey had a 12% response rate. Out of the 4 336 respondents, 16 did not give consent for their information to be used for research purposes, and these were therefore removed from the data set. This meant that 4 320 viable responses remained. A total number of 4 124 participants competed in the lucky draw. The three prizes were arranged after the survey deadline and the winners had to collect them from the SU Industrial Engineering Department. All three of the winners cashed in on the opportunity within a week's time.

3.4 Revised research methodology

The original research methodology proposed the use of HDFS for the storing of the data and the Hadoop programming paradigm, MapReduce, together with RDFs for the analysis of the data. This approach would have been feasible if data shared on social media was collected via the method of crawling. Data would then be in URI format, which is required to build RDFs and the volume, variety and velocity of the data would then have qualified as big data, which makes Hadoop and its projects like Jena appropriate.

Owing to legal restrictions, the method of crawling caused the exclusion of certain vital social media platforms. It was therefore decided to collect data on the SU population's social media sharing habits via an email survey. The survey generated a total number of 4 320 responses. The survey retrieved 171 variables per response, which means that 738 720 data points had to be analysed. A data set of this volume can be processed by traditional data applications and does not qualify as big data. The original research approach was therefore not applicable as data was not in URI format and big data techniques were not necessary.

The revised research methodology is therefore to continue the study as initially planned, but to adapt the data analysis techniques. It is proposed that data is stored on a local secure drive rather than the HDFS and sorted with traditional data processing tools such as Excel and not with RDFs and MapReduce. Instead of SPARQL, traditional data mining with software such as Statistica will then commence.

3.5 Preliminary investigation of collected data

The data received from the SAFPS and the data collected via the email surveys are examined in this section to gain perspective.

3.5.1 Preliminary investigation: SAFPS data

The information received from the SAFPS was cleaned and sorted in Excel. The data was then imported into Statistica. The total number of feasible incidents was 2,039. Graphs were drawn to display the results on the different variables received per incident: Age, Gender, Relationship Status and Income. As shown in Figure 3.2 and Figure 3.4 the results received on the variables Age and Gender complied with that found by Erasmus (2015) and Dirk (2015b), which stated that males between the ages of 28 and 40 were prime targets for identity theft.

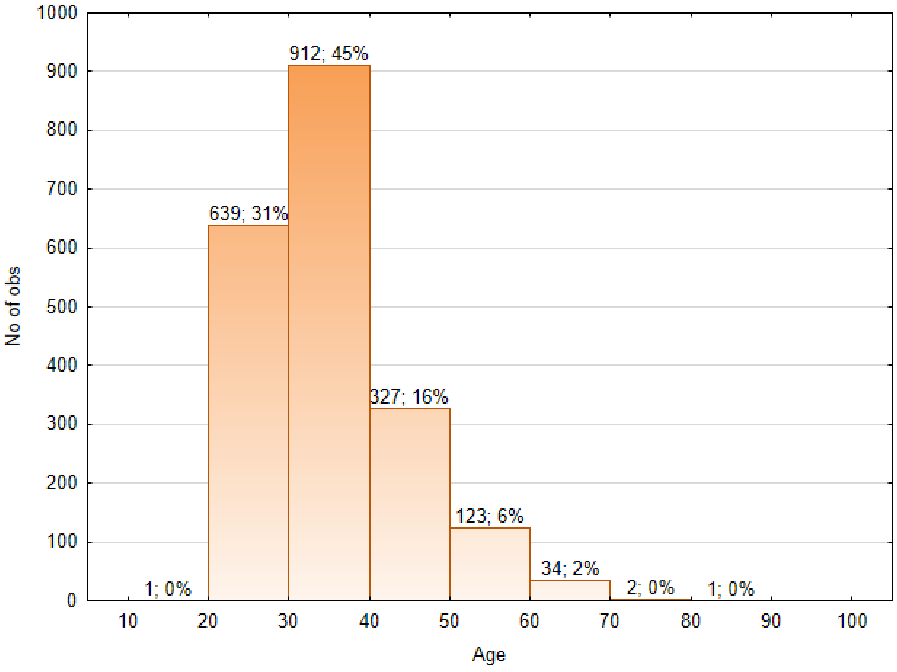


Figure 3.2: SAFPS Data: Histogram of Variable Age.

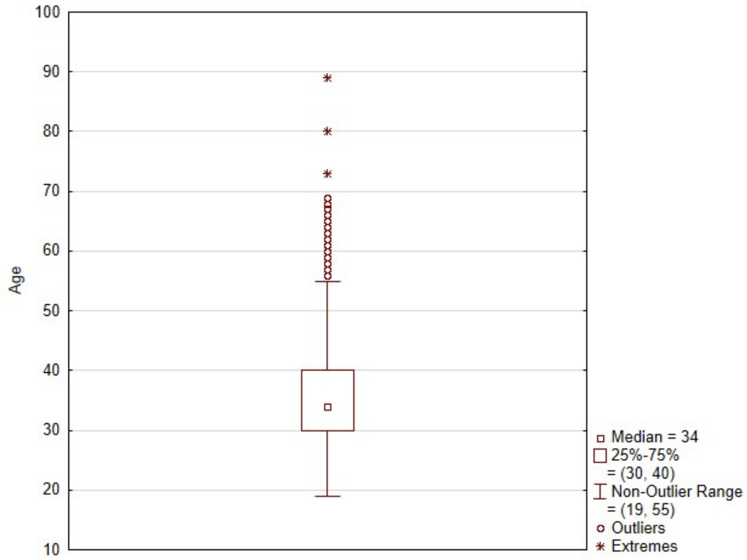


Figure 3.3: SAFPS Data: Box Plot of Variable `Age`.

The SAFPS data shows that 69% of the incidents reported were male victims compared to only 31% being female and 45.5% of the victims were between the ages of 30 and 40 years. The Box Plot in Figure 3.3 of the variable `Age` illustrates that the most frequent age among the victims was 34 years.

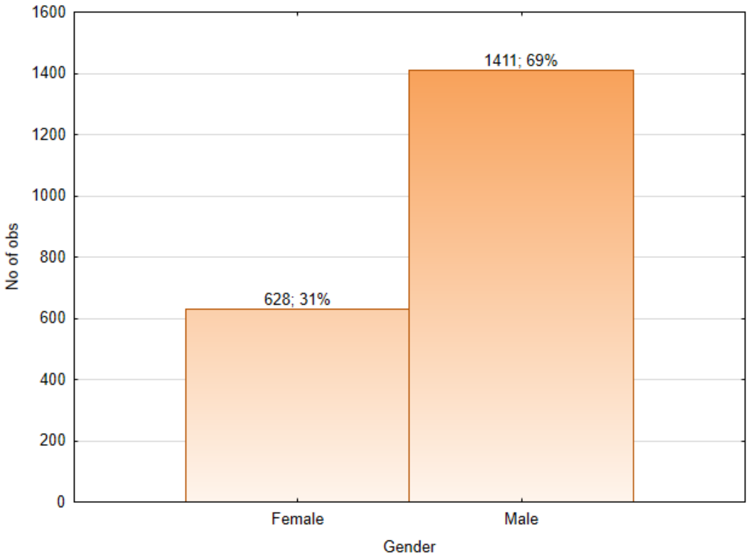


Figure 3.4: SAFPS Data: Histogram of Variable `Gender`.

The results on the variables `Relationship Status` and `Income` are presented in Figure 3.6 and Figure 3.5. Figure 3.5 shows that 96% of victims stated that they have no income and according to Figure 3.6, 77%, which was the majority of victims, did not reveal their relationship status. The results on both these variables are considered as poor answers, which can either be due to insufficient data, the fact that victims wanted to keep this data confidential or random chance.

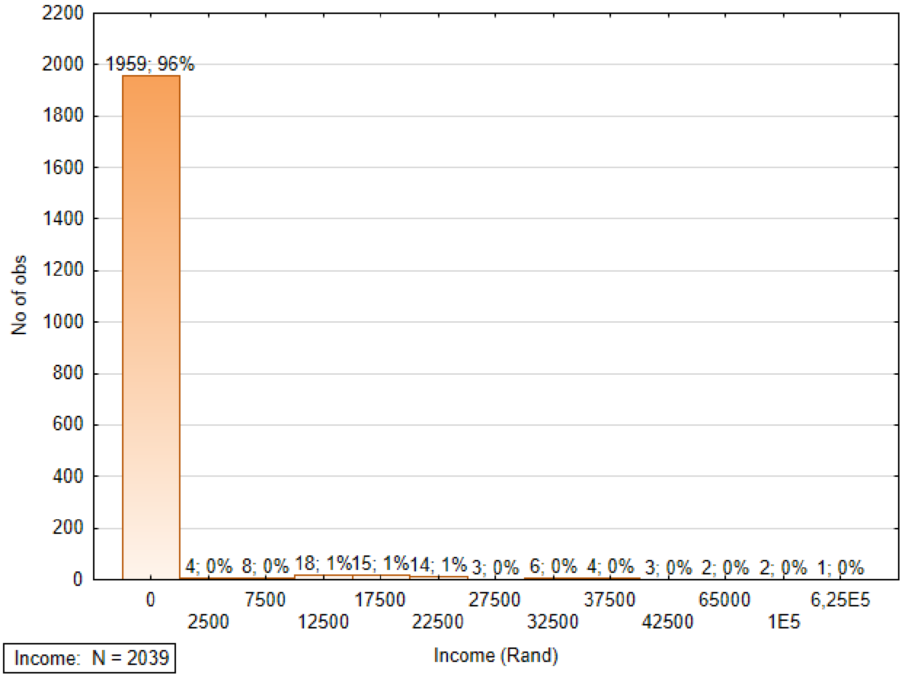


Figure 3.5: SAFPS Data: Histogram of Variable `Income`.

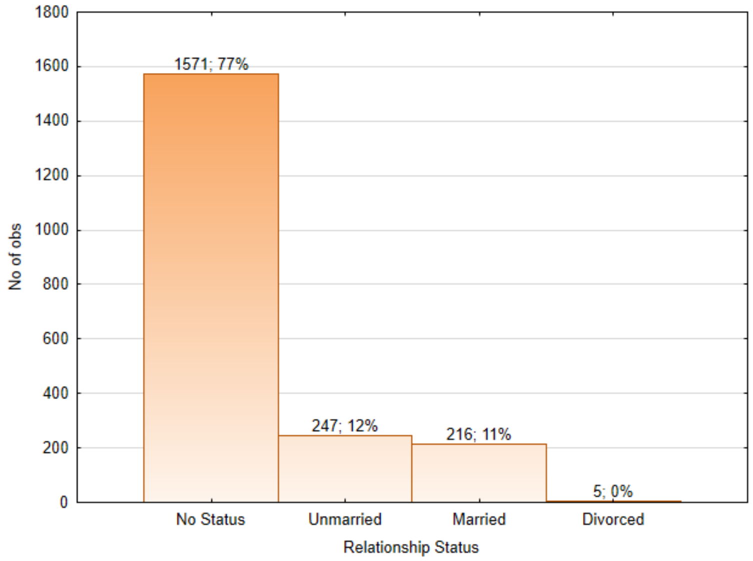


Figure 3.6: SAFPS Data: Histogram of Variable `Relationship Status`.

It can be concluded that the variables `Gender` and `Age` may possibly have a significant influence on the data set and can be expected to be predictor variables of identity theft in the predictive model built for the research study.

3.5.2 Preliminary investigation: email survey data

The survey data was exported from the SU server into Excel. The responses were in the survey designed format, mostly binary, as described in Tables 3.2, 3.3 and 3.4. The total number of feasible responses was 4 320 and the total number of variables per response was 165, which contained 10 platform subscription variables Yj, where $j = 1, 2, \ldots 10$, according to the platforms demonstrated in Table 3.5, 150 shared attributes among the various platforms variables as presented in Table 3.5 and finally the variables `Age`, `Gender`, `Income`, `Relationship Status` and `Victim`.

Table 3.5: Variables for Attributes shared on Social Media Platforms.

150 Variables		Platform Revealed On									
		Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10
Attribute Revealed		None	Facebook	LinkedIn	Twitter	Instagram	Youtube	Pinterest	Mxit	Dating Sites	Other
X1	Name	X1Y1	X1Y2	X1Y3	X1Y4	X1Y5	X1Y6	X1Y7	X1Y8	X1Y9	X1Y10
X2	Surname	X2Y1	X2Y2	X2Y3	X2Y4	X2Y5	X2Y6	X2Y7	X2Y8	X2Y9	X2Y10
X3	Identity number	X3Y1	X3Y2	X3Y3	X3Y4	X3Y5	X3Y6	X3Y7	X3Y8	X3Y9	X3Y10
X4	Birthday	X4Y1	X4Y2	X4Y3	X4Y4	X4Y5	X4Y6	X4Y7	X4Y8	X4Y9	X4Y10
X5	Gender	X5Y1	X5Y2	X5Y3	X5Y4	X5Y5	X5Y6	X5Y7	X5Y8	X5Y9	X5Y10
X6	Race	X6Y1	X6Y2	X6Y3	X6Y4	X6Y5	X6Y6	X6Y7	X6Y8	X6Y9	X6Y10
X7	Physical address	X7Y1	X7Y2	X7Y3	X7Y4	X7Y5	X7Y6	X7Y7	X7Y8	X7Y9	X7Y10
X8	Email address	X8Y1	X8Y2	X8Y3	X8Y4	X8Y5	X8Y6	X8Y7	X8Y8	X8Y9	X8Y10
X9	Cell number	X9Y1	X9Y2	X9Y3	X9Y4	X9Y5	X9Y6	X9Y7	X9Y8	X9Y9	X9Y10
X10	Relationship status	X10Y1	X10Y2	X10Y3	X10Y4	X10Y5	X10Y6	X10Y7	X10Y8	X10Y9	X10Y10
X11	School details	X11Y1	X11Y2	X11Y3	X11Y4	X11Y5	X11Y6	X11Y7	X11Y8	X11Y9	X11Y10
X12	University/College details	X12Y1	X12Y2	X12Y3	X12Y4	X12Y5	X12Y6	X12Y7	X12Y8	X12Y9	X12Y10
X13	Job details	X13Y1	X13Y2	X13Y3	X13Y4	X13Y5	X13Y6	X13Y7	X13Y8	X13Y9	X13Y10
X14	Income	X14Y1	X14Y2	X14Y3	X14Y4	X14Y5	X14Y6	X14Y7	X14Y8	X14Y9	X14Y10
X15	Credit card details	X15Y1	X15Y2	X15Y3	X15Y4	X15Y5	X15Y6	X15Y7	X15Y8	X15Y9	X15Y10

The goal of this section was to study the survey results in order to see what the nature of the data was and to determine if the effort of building predictive models with the data would be worth it. In order to gain some perspective on the data, it was decided that an average vulnerability score would be calculated per respondent according to what attributes they have shared on how many social media platforms.

Table 3.6: Attribute Risk Factors for Vulnerability Score Calculation.

Risk Factor $\mathcal{R}$	Degree of Sensitivity	Attributes
5	Offenders can commit various types of identity theft fraud with this information.	Identity number (X3), Physical address (X7)
4	This is highly sensitive information, but crimes are limited and easy to stop.	Credit card details (X15)
3	Personal details that reveal an individuals' reputation and therefore makes them targets.	Birthday (X4), Gender (X5), Relationship Status (X10), Income (X14)
2	Data that is often known publicly and used in combination with other details for crimes such as impersonation.	Name (X1), Surname (X2), Race (X6), School Details (X11), University/College Details (X12), Job Details (X13)
1	Information that can contribute to identity theft crimes, but is not enough solely to commit a crime.	Email Physical Address (X8), Cell Number (X9)

In furtherance of the vulnerability score calculation, it was decided to assign a risk factor to each attribute shared, according to its degree of sensitivity. Table 3.6 lists the attributes with their degree of sensitivity and respective risk factors. Risk factors were not assigned to the specific social media platforms the attributes were shared on. The purpose of the research study is not to discriminate between social media platforms' privacy settings. All of the included social media platforms were therefore assigned equal risks of one.

The vulnerability factor per attribute was therefore calculated with:

$$V(Xi) = 0, \text{ when } j = 1,$$

and

$$V(Xi) = R(Xi) \sum_{j=2}^{10} XiYj$$

where

$$i = 1, 2, \ldots 15.$$

The average vulnerability V_A score could then be calculated per response case with

$$V_A = \frac{\sum_{i=1}^{15} VXi}{15}.$$

The maximum average vulnerability score possible is therefore 24 and would be received if an individual happened to share all 15 attributes on all of the nine social media platforms and the highest possible vulnerability score per single social media platform is 2.67. Figure 3.7 displays the number of responses and the average vulnerability scores per age groups: below 20 years, 20–30 years, 30–40 years, 40–50 years, 50–60 years, 60–70 years, 70–80 years, 80–90 years and 90–100 years. It is seen that the age group 30–40 years revealed the highest average vulnerability score of 2.85.

It must be noted that the average vulnerability score gradually increased from the age group below 20 years to the group 30–40 years and then gradually decreased to the group 50–60 years. The lower scores for the younger generation could perhaps be explained by the fact that they grew up with technology and social media and are therefore more streetwise, as well as the fact that they are still in the initial stages of their careers and not financially established yet. The peak score generation of 30–40 years probably use social media as society expects it of them to adapt to the new technology era and the majority of them are financially established. Then finally the decrease in average vulnerability for the age groups thereafter is probably due to the cautious behaviour of the older generation as they were not brought up with technology and are in general either not comfortable with social media or not even using it.

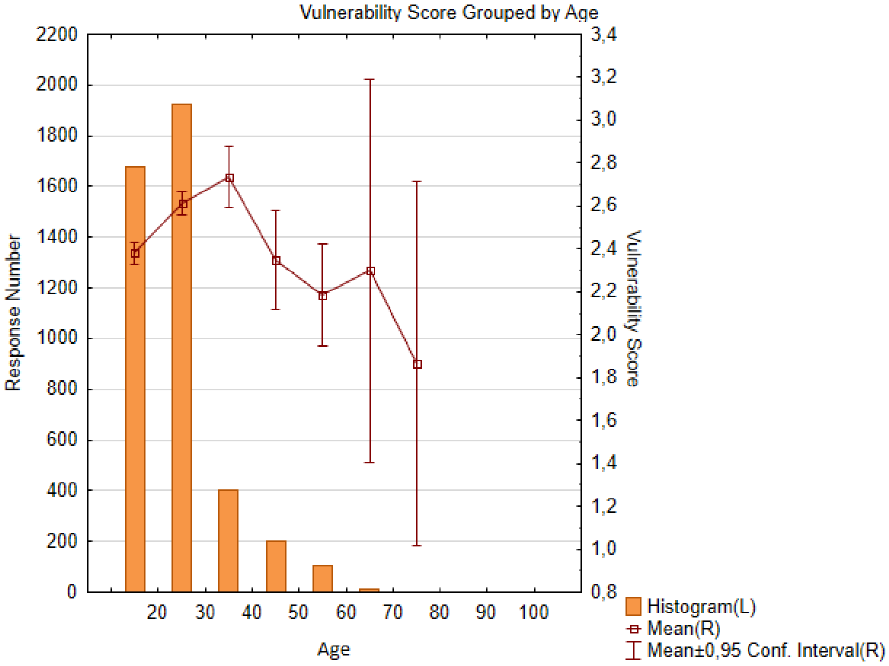


Figure 3.7: Email Survey Data: Histogram of Variable `Age`.

Figure 3.8 shows the number of responses and the average vulnerability scores per gender groups: male and female. Male victims presented a higher average vulnerability score of 2.68 compared to the 2.39 score among females.

The results for the variables `Age` and `Gender` come with no surprise as they once again agree with the most popular age of 28–40 years among identity theft victims and the much higher percentage of male victims compared to female victims as reported by Erasmus (2015) and Dirk (2015b).

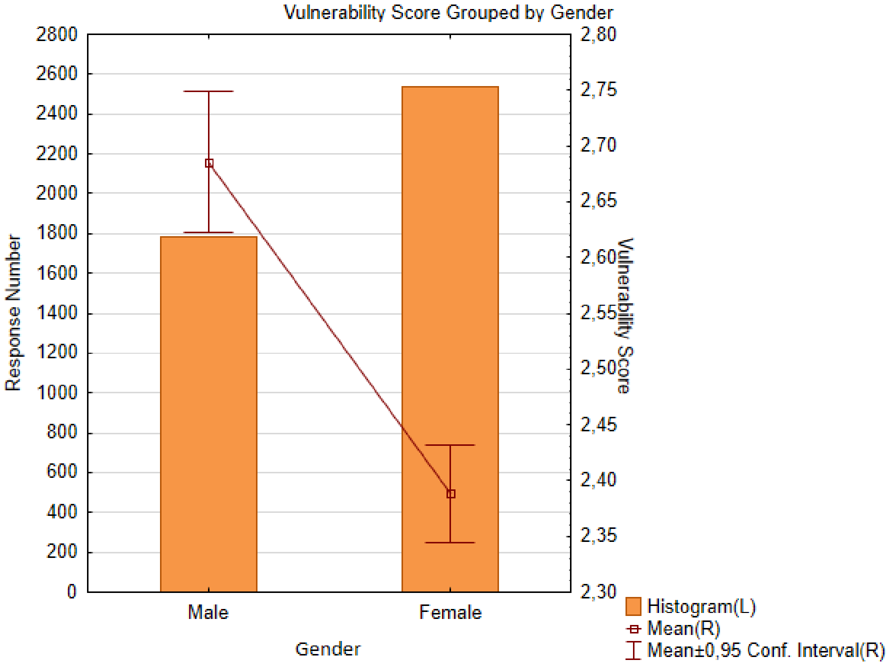


Figure 3.8: Email Survey Data: Histogram of Variable `Gender`.

In Figure 3.9 the total number of responses and average vulnerability scores are displayed according to the five relationship status options: ‘Single’, ‘Married’, ‘In a Relationship’, ‘Divorced’, and ‘No Status’. The status with the highest average vulnerability score was ‘Married’ with a score of 2.62 and the lowest score was of the variable ‘No Status’ with a score of 2.12.

Respondents could optionally provide their monthly income. A total number of 1 247 income responses were retrieved. The average vulnerability scores were calculated for these cases and are graphed in Figure 3.10 according to income subsets of the received results. The results peaked at two income values, R350 000 and R650 000 per year with average vulnerability scores of 5 and 4 respectively. The variable `Income` contained many missing values and therefore these results are skewed and were not considered as noteworthy.

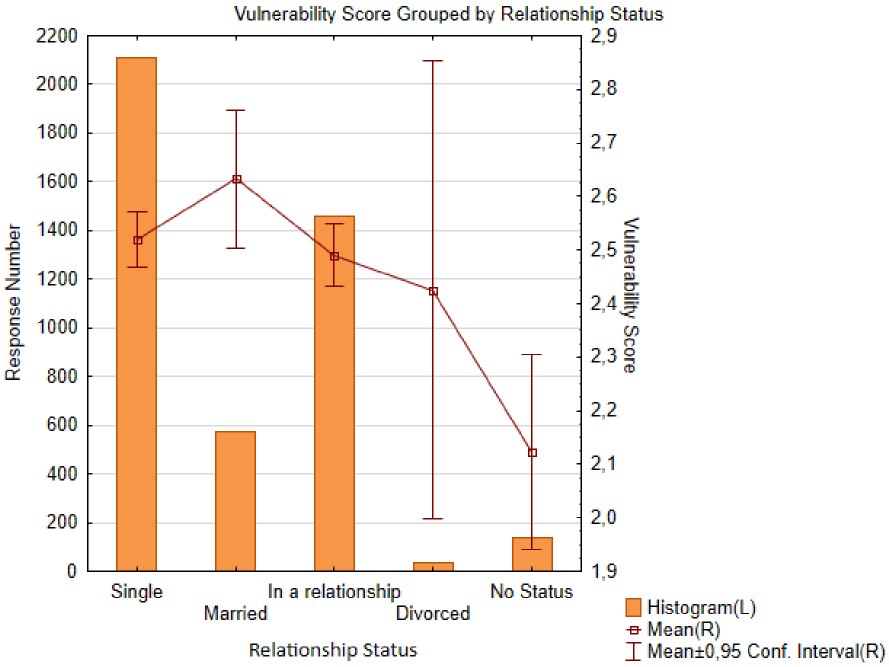


Figure 3.9: Email Survey Data: Histogram of Variable `Relationship Status`.

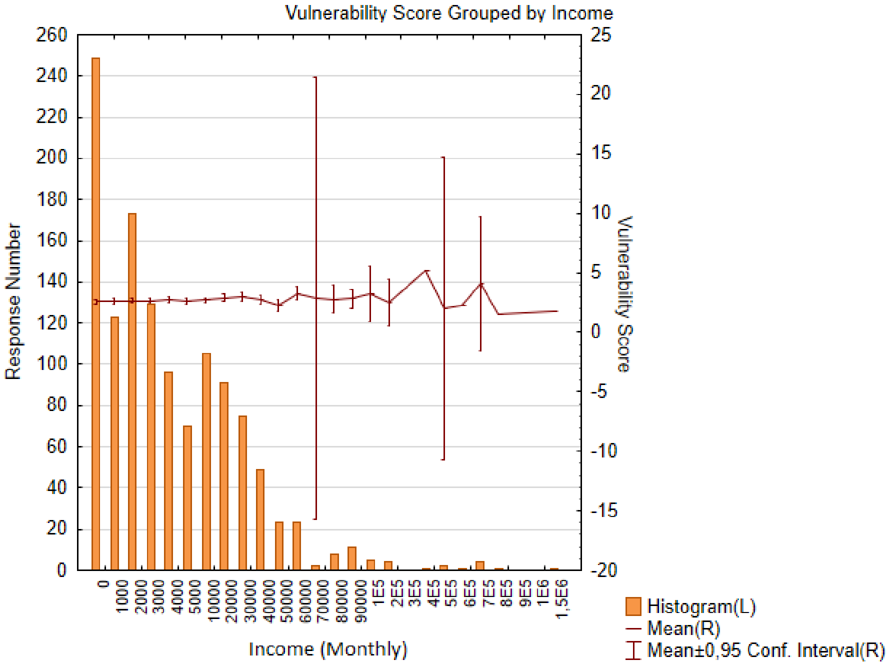


Figure 3.10: Email Survey Data: Histogram of Variable `Income`.

The final variable investigated was `Victim`. The results on this variable were crucial, because the research study assumed that identity theft victims are less cautious with the sharing of personal information on social media platforms compared to non-victims. If the results therefore prove the opposite, the research hypothesis will have to be rejected. When respondents selected that they have been identity theft victims, they were given five types of identity theft to choose from as described in Table 3.4. Due to a lack of significant responses it was decided to combine the different types to a single 'Yes' answer for the question as to whether they had been an identity theft victim. The derived average vulnerability scores for the dependent variable `Victim`, graphed in Figure 3.11, however presented the expected results. The average vulnerability score for previous identity theft victims was reasonably higher than non-victims with a score of 2.72 compared to 2.48 for non-victims.

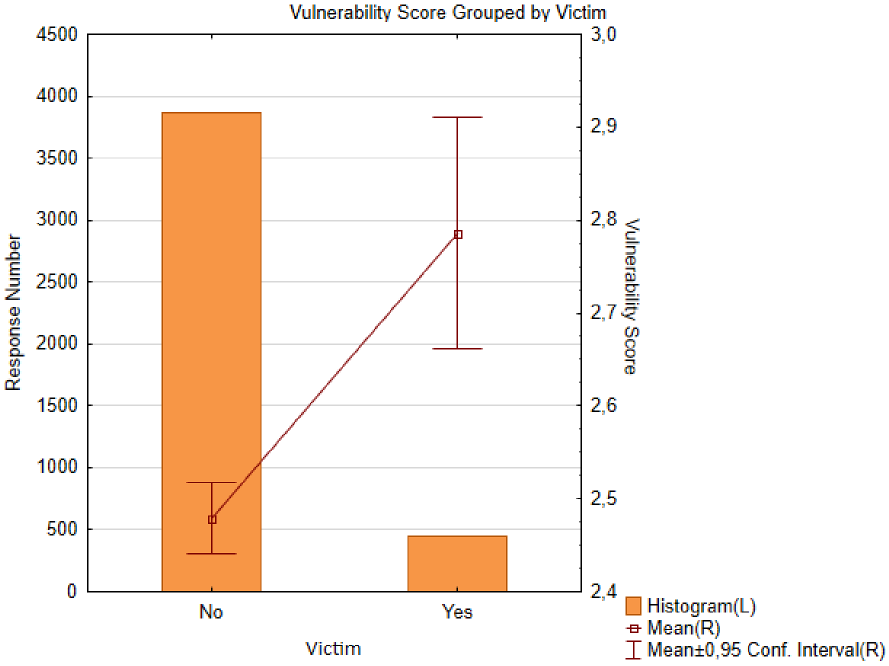


Figure 3.11: Email Survey Data: Histogram of Variable `Victim`.

It can be concluded that the collected survey data displayed the expected results and is therefore sufficient to develop a prediction model. The prediction model was therefore a realistic objective for the research study. The variables `Gender` and `Age` consistently presented significant results and were therefore strongly expected to be predictor variables of identity theft in the prediction model developed in the next chapter.

Chapter 4

Data Mining

In this chapter, the data mining of the survey data commences. The content of the chapter includes a brief introduction to Statistica and an overview and application of the Cross-Industry Standard Process (CRISP), which was the chosen model for the data mining process.

4.1 Introduction to Statistica

Before data mining commenced the researcher went for assistance at the Centre for Statistical Consultation at Stellenbosch University. It was recommended to use the software program Statistica.

The Statistica suite of analytics software products and solutions is the top product provided by StatSoft, which is now part of Dell Software. According to Dell (2016), Statistica offers the most inclusive collection of data analysis, management, visualisation, and mining procedures in one software platform. Its techniques are the most comprehensive range and include methods like predictive modelling, clustering, classification and exploratory techniques. The program has been in use for over two decades and has a good record with various accomplishments. Its global user base of over one million users testifies that it is a known and tested analytics platform (Dell, 2016).

An article by Thompson (2015) stated that Statistica received the 2015 Technology Innovation Award from Dresner Advisory Services and Dell Statistica was recognised as a leader in Advanced and Predictive Analytics for 2015.

It was therefore decided that Statistica was the program of choice. StatSoft, Inc. presents a series of 35 tutorial videos that cover essential concepts, processes and data mining techniques that are included in Statistica (Electronic Version: StatSoft, 2013). The concepts, processes and data mining techniques recommended in the series of videos, were used as guidelines for the data mining process as they are built into the software and therefore most relevant.

4.2 Data mining overview

In this section the different types of data were investigated to assist with the selection of the data mining model to use as guideline for the data mining process.

4.2.1 Data mining types

It was fundamental to first identify the types of data collected by the survey, before the application of the data mining model and the determination of the necessary data mining techniques could kick off. There are three main types of data mining applications (Electronic Version: StatSoft, 2013):

1. **Classification type problems** – The variable of interest is categorical in nature. The goals of classification problems are to find variables that are strongly related to the variable of interest and to develop a predictive model where a set of variables is used to classify the variable of interest.

2. **Regression type problems** – The variable of interest is continuous in nature. The regression type problem is therefore not applicable to the data used in this study and will not be elaborated on.

3. **Clustering type problems** – There is no traditional variable of interest and data is sorted into clusters. The clustering type problem is therefore again not relevant to the data used in this study and will not be discussed in any further detail.

The email survey asked members of the SU population to give consent that their information could be used for research purposes. The responses of individuals who answered 'Yes' to giving consent were recorded and those who answered 'No' could not complete the survey. The second part of the survey was to determine the number of users per social media platform for all the various platforms included in the survey and then to learn what attributes these users share on which and how many platforms. These questions were constructed as checkboxes that allowed participants to tick multiple boxes, but was restricted to a minimum of one box per question. The final part of the survey requested the participant's age, which had a lower limit of 18 years and a higher limit of 100. The participants were then requested to select one of the radio buttons concerning their gender and their relationship status. Monthly income was an optional question. The survey was designed in such a manner that only the variables `Age` and `Income` were continuous variables. The question as to whether the survey participants had been identity theft victims or not had six possible checkboxes they could select. Participants were allowed to tick multiple boxes, but were once again restricted to one answer and limited to five. The variable of interest in the data set was whether the individual had been an identity theft victim or not. The dependent variable resulted in a binary response.

It can be concluded that all the variables, except for `Age` and `Income` were categorical in nature. The purpose of the research study was to determine the predictor variables of the dependent variable `Victim` and to build a prediction model to classify new cases according to the dependent variable. Due to the binary dependent variable `Victim` and the purpose of the research, it was therefore clear that the research problem was a classification type problem.

4.2.2 Identification of the data mining model

The following data mining models were discussed in the Literature Review: Knowledge Discovery in Databases (KDD), SEMMA Methodology and the Cross Industry Standard Process (CRISP). Shafique & Qaiser (2014b) did a comparative study of the three different models and the comparison results that were found, are summarised in Table 4.1. It is seen that the SEMMA model does not provide for the initial phase where objectives are determined or the final phase that scores new data with the final model. It was therefore decided that SEMMA is not a good fit for the research, because it does not satisfy the objective to classify new data as high-risk or low-risk identity theft victims. The two remaining models, KDD and CRISP, are very similar. After the consideration of both models, the CRISP model was preferred. It covers the same information range as KDD, but in fewer steps. The CRISP model was developed in 1996 and has been thoroughly researched. The CRISP 1.0 version is published, complete and documented and the model is well structured and illustrated (Shafique & Qaiser, 2014b).

Table 4.1: Comparison of the Data Mining Models: KDD, CRISP-DM and SEMMA (Shafique & Qaiser, 2014b).

Data Mining Process Models	**KDD**	**CRISP-DM**	**SEMMA**
Number of Steps	9	6	5
Name of Steps	Developing and Understanding of the Application	Business Understanding	*
	Creating a Target Data Set	Data Understanding	Sample
	Data Cleaning and Pre-processing		Explore
	Data Transformation	Data Preparation	Modify
	Choosing the suitable Data Mining Task	Modelling	Model
	Choosing the suitable Data Mining Algorithm		
	Employing Data Mining Algorithm		
	Interpreting Mined Patterns	Evaluation	Assessment
	Using Discovered Knowledge	Deployment	*

4.3 Application of CRISP

The CRISP data mining model provides a systematic framework that guides the data mining process with the model's six phases. The phases were discussed in

Section 2.6.1.3 in the Literature Review. These phases will now be applied in this study and are subsequently described.

4.3.1 CRISP phase 1: Business understanding

The business understanding phase is to determine the goals of the project and to clarify what outcomes are expected from the project.

4.3.1.1 CRISP objectives

The following objectives are essential for the data mining phase:

1. Determine the variables that best predict identity theft victims;
2. Find a predictive model that best classifies victims; and
3. Deploy the model to make decisions on whether or not a social media user is at risk of becoming an identity theft victim or not.

The handling and execution of the above goals will determine the success of the research study.

4.3.1.2 CRISP expected outcomes

The historic data that was received from the SAFPS was delegated to serve as reference variables. The data was cleaned, sorted and analysed during the preliminary investigation that is discussed in Section 3.5.1. The two variables that significantly stood out were `Age` and `Gender`. It is therefore expected that these two variables will be identified by prediction models as predictor variables of identity theft victims.

4.3.2 CRISP phase 2: Data understanding

The data collected in Chapter 3 is explored graphically in order to screen variables and determine possible predictor variables.

4.3.2.1 Histograms of survey variable occurrences

Histograms were graphed for the number of occurrences of each of the 171 survey variables in order to determine the variables that presented very low responses and that could therefore be neglected.

The first variable, `Consent`, was obtained by the compulsory question in the survey that is presented in Figure 3.1. Only 16 participants did not give consent and were therefore removed from the data set, the remaining 4320 participants were accepted as viable responses.

Figure 4.1 to Figure 4.9 illustrate the percentage of survey participants that were subscribers and non-subscribers to the various social media platforms. Figure 4.1, Figure 4.4 and Figure 4.5 indicate that Facebook, Instagram and Youtube are the most popular social media sites. Figure 4.7, which presents Mxit subscribers, and Figure 4.8, which presents Dating Site subscribers, show low user percentages.

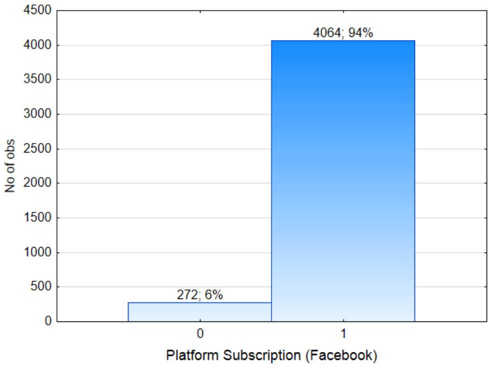


Figure 4.1: Facebook Subscription.

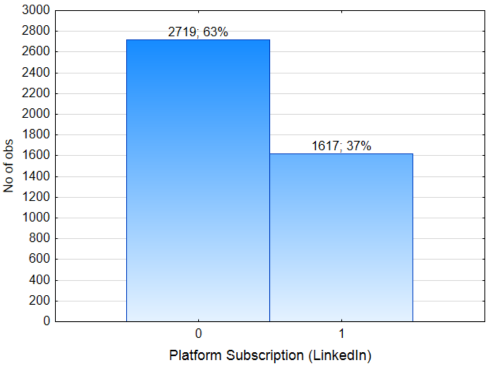


Figure 4.2: LinkedIn Subscription.

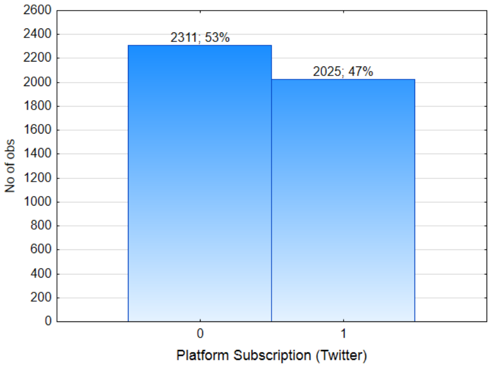


Figure 4.3: Twitter Subscription.

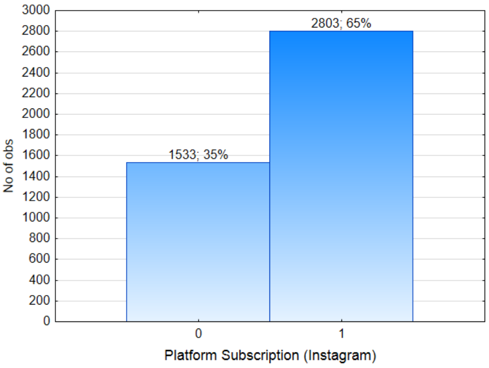


Figure 4.4: Instagram Subscription.

The category with the lowest percentage response rate was for the platform Mxit with a 3% response rate for the category subscribers, which was a total of 137 responses. Mxit responses are therefore expected to be insufficient. For now none of the platform variables should be considered to be excluded for further analysis, but Mxit must be investigated more.

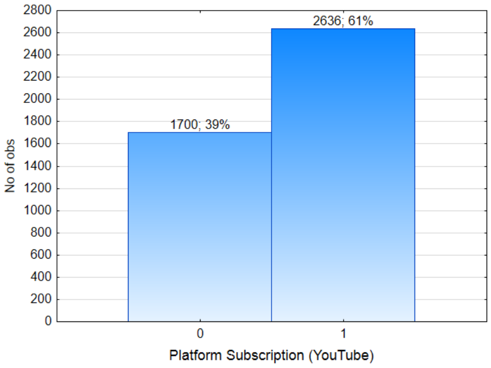


Figure 4.5: YouTube Subscription.

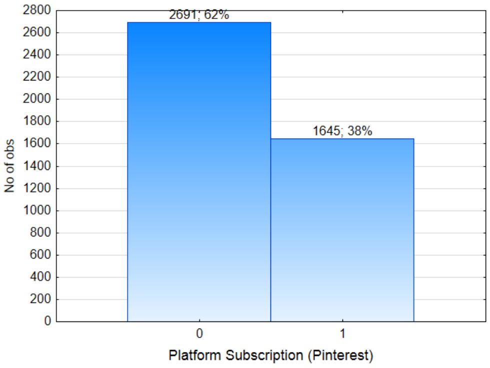


Figure 4.6: Pinterest Subscription.

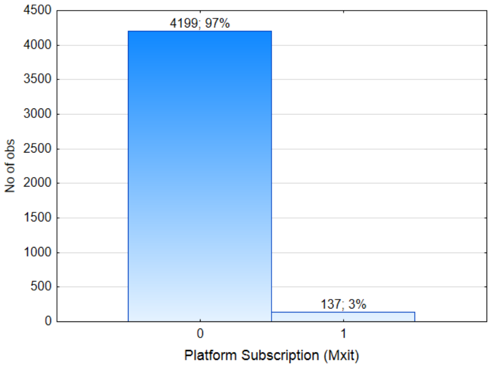


Figure 4.7: Mxit Subscription.

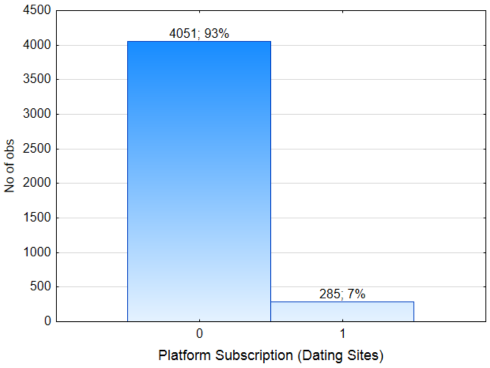


Figure 4.8: Dating Sites Subscription.

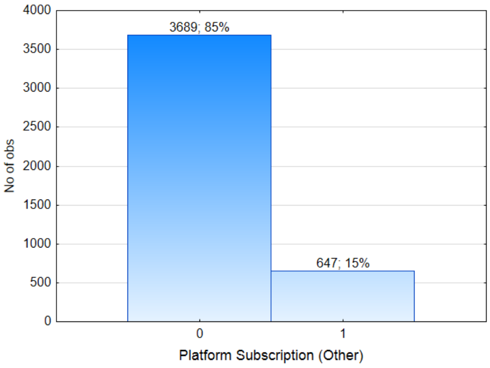


Figure 4.9: Other Subscription.

The recording of the information sharing habits of social media users resulted in 150 variables. The variables are listed in Table 3.5 and discussed in Section 3.5.2. The histograms of these 150 variables are included in Appendix B, section B.1, and only the variables with noteworthy results are summarised in tables and discussed in this chapter.

The five attributes, revealed on social media platforms, that stood out to be quite meaningless due to very low responses were `Identity Number`, `Physical Address`, `Job Details`, `Income` and `Credit Card Details`. As presented in Table 4.2, 93% of the population specified that they have not revealed their identity numbers on any social media platforms, 83% of the population indicated that they have not disclosed their physical address on any social media platforms, 98% did not share their income and 98% never revealed their credit card details on any platforms. Except for Facebook, where the physical address of 14% and job details of 41% of the population were revealed and LinkedIn, where 27% of the population displayed their job details, no other social media platform displayed significant results concerning any of these attributes.

Table 4.2: Noteworthy Attributes.

% of population that revealed Xi on Yj		Platform Revealed On									
		Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10
Attribute Revealed		None	Facebook	LinkedIn	Twitter	Instagram	Youtube	Pinterest	Mxit	Dating Sites	Other
X3	Identity number	93%	4%	3%	1%	2% (73)	1%	0%	0%	0%	1%
X7	Physical address	83%	14%	4%	1%	1%	1%	0%	0%	0%	1%
X13	Job details	47%	41%	27%	2% (74)	1%	0%	0%	0%	1%	1%
X14	Income	98%	1%	1%	0%	0%	0%	0%	0%	0%	0%
X15	Credit card details	98%	0%	0%	0%	0%	0%	0%	0%	0%	1%

Table 4.3 shows the attribute sharing responses for the social media platforms 'Twitter', 'YouTube', 'Pinterest', 'Mxit', 'Dating Sites', the option of other social media platforms, 'Other' and the option of no social media platform, 'None', that were investigated. These were the social media platforms that displayed attribute responses that were noteworthy. 'YouTube', 'Pinterest', 'Dating Sites' and 'Mxit' were the only platforms that indicated low responses to the sharing of cellular phone numbers. Mxit was the only platform that presented low percentages for the sharing of every single attribute. This can be due to the major drop, of 6.3 million users, in the Mxit user base between 2013 and 2015 (MyBroadband, 2016).

The platforms 'Twitter', 'YouTube', 'Pinterest', 'Dating Sites' and 'Other' presented very low response rates for the variables: `Relationship Status`, `School Details`, `Job Details`, `Income` and `Credit Card Details`. The low responses can be explained by the fact that these platforms, except for 'Dating Sites' that normally require users' relationship status, do not explicitly require these details. The same reason can be given for the low rate of responses to `University Details` shared on 'YouTube', 'Pinterest', 'Mxit', 'Dating Sites' and 'Other' platforms. All the platforms included in Table 4.3, except for the 'None' option, show that users do not generally share their identity numbers and physical addresses on these sites, which is very positive for South African online fraud prevention.

The option to select 'None', which means that the attribute is not revealed on any platform displayed the following results:

- Almost the entire population has their name, surname, birthday and gender revealed on some or other social media platform;
- An extremely small percentage of the population has shared their identity number, physical address, income and credit card details on any social media platforms;
- The majority of the population has their email address, relationship status, school details and university/college details stated somewhere on social media; and
- Approximately half of the population present their race, cellphone number and job details on social media platforms.

Table 4.3: Noteworthy Platforms.

% of population that revealed Xi on Yj		Platform Revealed On						
		Y1	Y4	Y6	Y7	Y8	Y9	Y10
Attribute Revealed		None	Twitter	Youtube	Pinterest	Mxit	Dating Sites	Other
X1	Name	1%	39%	38%	29%	2% (94)	6%	10%
X2	Surname	2% (108)	34%	32%	24%	1%	2% (105)	7%
X3	Identity number	93%	1%	1%	0%	0%	0%	1%
X4	Birthday	9%	13%	10%	4%	1%	2% (93)	3%
X5	Gender	4%	24%	16%	11%	2% (80)	6%	5%
X6	Race	56%	8%	5%	3%	0%	3%	2% (103)
X7	Physical address	83%	1%	1%	0%	0%	0%	1%
X8	Email address	26%	16%	21%	10%	0%	1%	3%
X9	Cell number	50%	4%	2% (98)	1%	1%	1%	5%
X10	Relationship status	34%	2% (70)	1%	0%	1%	3%	1%
X11	School details	15%	2% (89)	1%	0%	0%	1%	1%
X12	University/college details	11%	3%	1%	0%	0%	2% (69)	1%
X13	Job details	47%	2% (74)	0%	0%	0%	1%	1%
X14	Income	98%	0%	0%	0%	0%	0%	0%
X15	Credit card details	98%	0%	0%	0%	0%	0%	1%

The response numbers for the continuous variable `Age` are presented in Figure 4.10. According to Figure 4.10 the minimum age of respondents was 18 years, which was due to the lower age limit of the survey and the maximum age was 77 years that is well within the target group of 18 to 80 years as described in Chapter 3. The average age of respondents was 25 years and the age that occurred most frequently was 21 years. The great number of respondents between the ages of 18 and 25 years is explanatory to the chosen population of university students.

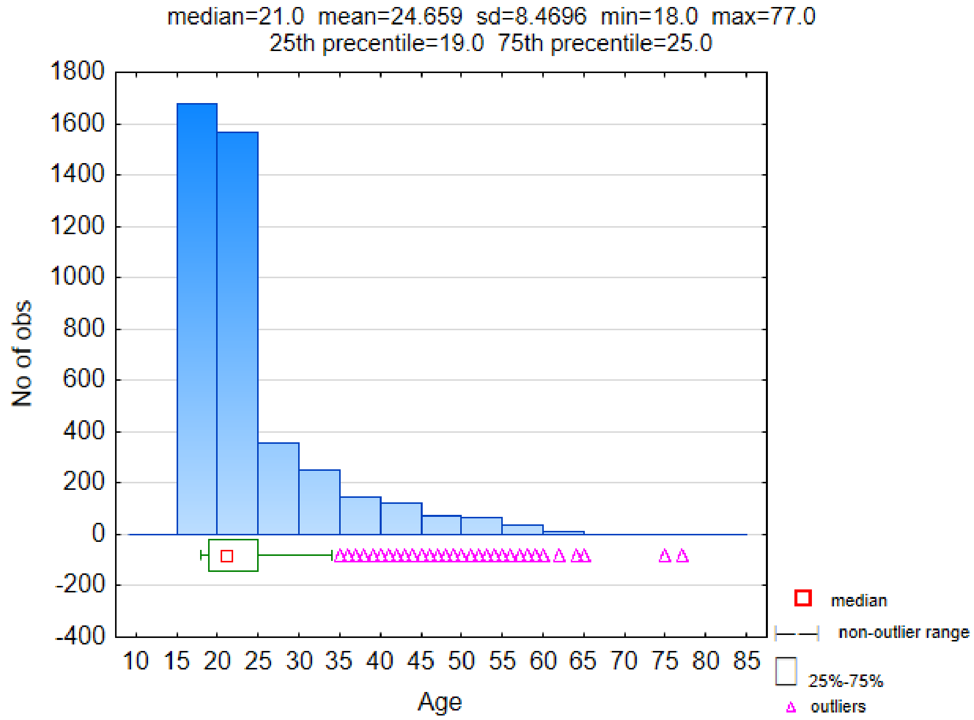


Figure 4.10: Variation in Age of Respondents.

Figure 4.11 presents the percentages of male and female responses of the variable Gender. The female responses outweighed the male responses with a percentage of 59% compared to 41%.

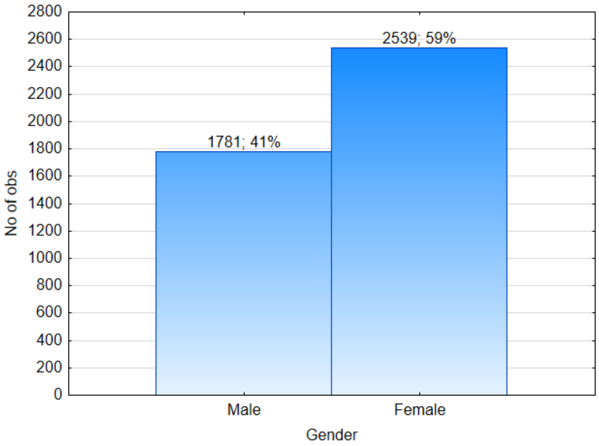


Figure 4.11: Number of Male and Female Respondents.

The `Relationship Status` variable occurrences among respondents are presented in Figure 4.12. The low percentages of divorced and married individuals are probably related to the age range of the population illustrated in Figure 4.10.

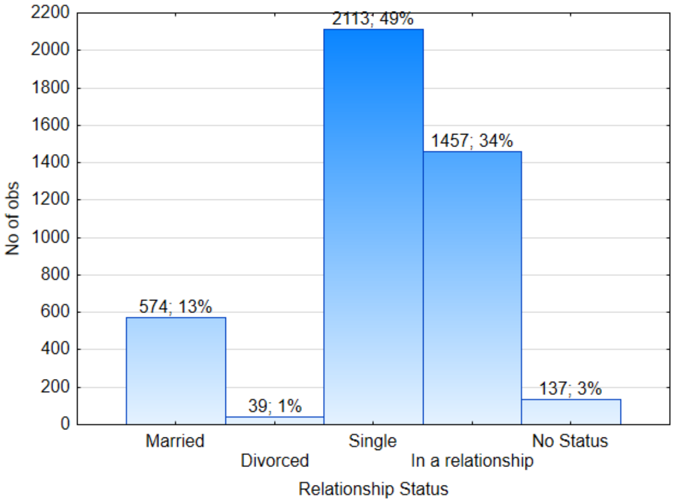


Figure 4.12: `Relationship Status` Occurrences among Respondents.

The variable `Income` is graphed in Figure 4.13. Due to the lack of responses, as observed by the graph that displays no significant results, the variable was removed from the data set and no further analysis was done on the variable.

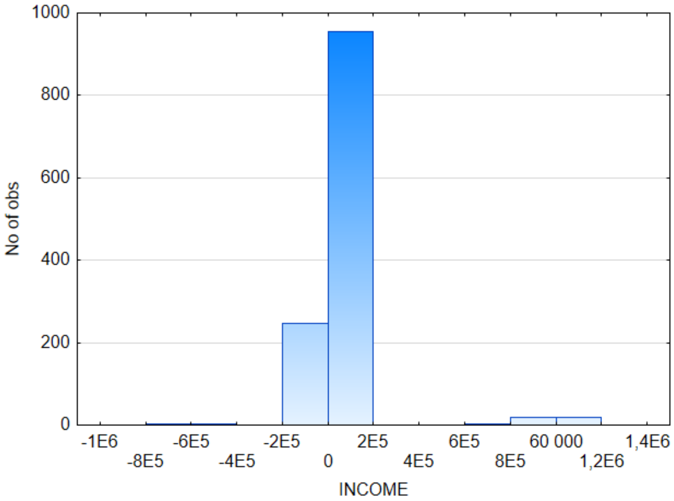


Figure 4.13: `Income` of Participants.

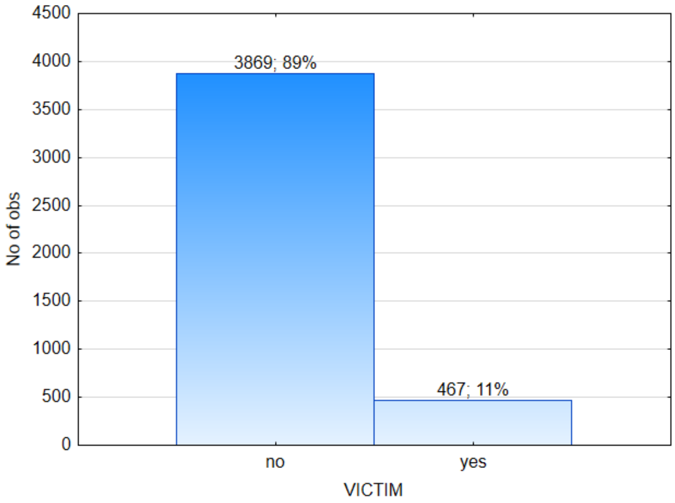


Figure 4.14: Number of Victims and Non-Victims.

The dependable variable of interest is graphed in Figure 4.14. The survey asked participants who indicated that they have been identity theft victims to specify the type of identity theft they have been victims of. Due to no significant results to a particular type, as seen in Appendix B, Figures B.151, B.152, B.153, B.154 and B.155, it was decided to merge the different types of fraud to a single answer, 'Yes, I have been a victim'.

Of all respondents, 11% had previously been identity theft victims. The resulting number of 467 victims out of the total population of 4 320 respondents were enough occurrences to do analysis on. This was a crucial breaking point in the research study.

4.3.2.2 Variable screening

In this section the variables that were graphically inspected with histograms in the previous section, were screened.

According to Electronic Version: StatSoft (2013), the effect of too many variables is:

- The curse of Dimensionality: The more predictor variables included in a prediction model, the more data points needed; and

- Deployment (the scoring of new data with a trained prediction model) Complexity: All variables included in the model are required when new data is gathered for deployment. Deployment is therefore made less complex with less input variables.

It is therefore important that all variables that are not related to the variable of interest and which add no or little information to the outcome of the dependent variable should be eliminated before modelling commences. It was decided that variables with response numbers of less than 100, were to be excluded from the data set.

The variable `Consent` was excluded, because it was insignificant to the analysis. The platform subscription variables presented in Figure 4.1 to Figure 4.9 displayed sufficient responses and all remained in the data set. Table 4.4 is a summary table of all the attribute sharing variables' response results.

The cut-off point for variable exclusion was 100, which is 2.31%. The percentages in the Tables are rounded values, which is why the number of responses are indicated in brackets next to the variables with responses of 2%.

The variables with unsatisfactory response rates were discussed in the previous section. The remaining attribute sharing variables, after all variables with response numbers less than 100, were removed, are listed in Table 4.5.

The variables `Age` and `Gender`, in Figure 4.10 and Figure 4.11, portrayed significant responses and were included in the data set for analysis. The relationship status 'Divorced', graphed in Figure 4.12, had only 39 responses and was therefore removed from the data set. The variable `Income` displayed inconsistent results, as shown in Figure 4.13, because of a lack of responses and was excluded from the data set. The different types of identity theft for the dependent variable, `Victim`, were merged in the previous section, which resulted in a single variable instead of six variables.

Table 4.4: Summary of Variable Occurrences according to the Histograms' Results.

% of population that revealed Xi on Yj		Platform Revealed On									
		Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10
	Attribute Revealed	None	Facebook	LinkedIn	Twitter	Instagram	YouTube	Pinterest	Mxit	Dating Sites	Other
X1	Name	1%	93%	37%	39%	58%	38%	29%	2% (94)	6%	10%
X2	Surname	2% (108)	91%	36%	34%	51%	32%	24%	1%	2% (105)	7%
X3	Identity number	93%	4%	3%	1%	2% (73)	1%	0%	0%	0%	1%
X4	Birthday	9%	87%	21%	13%	12%	10%	4%	1%	2% (93)	3%
X5	Gender	400%	92%	30%	24%	32%	16%	11%	2% (80)	6%	5%
X6	Race	56%	39%	12%	8%	14%	5%	3%	0%	3%	2% (103)
X7	Physical address	83%	14%	4%	1%	1%	1%	0%	0%	0%	1%
X8	Email address	26%	62%	27%	16%	16%	21%	10%	0%	1%	3%
X9	Cell number	50%	44%	10%	4%	4%	2% (98)	1%	1%	1%	5%
X10	Relationship status	34%	65%	3%	2% (70)	3%	1%	0%	1%	3%	1%
X11	School details	1500%	81%	24%	2% (89)	3%	1%	0%	0%	1%	1%
X12	University/college details	1100%	82%	28%	3%	7%	1%	0%	0%	2% (69)	1%
X13	Job details	47%	41%	27%	2% (74)	1%	0%	0%	0%	1%	1%
X14	Income	98%	1%	1%	0%	0%	0%	0%	0%	0%	0%
X15	Credit card details	98%	0%	0%	0%	0%	0%	0%	0%	0%	1%

Table 4.5: Variable Screening.

Screened Variables		Platform Revealed On									
		Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10
	Attribute Revealed	None	Facebook	LinkedIn	Twitter	Instagram	Youtube	Pinterest	Mxit	Dating Sites	Other
X1	Name	*	X1Y2	X1Y3	X1Y4	X1Y5	X1Y6	X1Y7	*	X1Y9	X1Y10
X2	Surname	X2Y1	X2Y2	X2Y3	X2Y4	X2Y5	X2Y6	X2Y7	*	X2Y9	X2Y10
X3	Identity number	X3Y1	X3Y2	X3Y3	*	*	*	*	*	*	*
X4	Birthday	X4Y1	X4Y2	X4Y3	X4Y4	X4Y5	X4Y6	X4Y7	*	*	X4Y10
X5	Gender	X5Y1	X5Y2	X5Y3	X5Y4	X5Y5	X5Y6	X5Y7	*	X5Y9	X5Y10
X6	Race	X6Y1	X6Y2	X6Y3	X6Y4	X6Y5	X6Y6	X6Y7	*	X6Y9	X6Y10
X7	Physical address	X7Y1	X7Y2	X7Y3	*	*	*	*	*	*	*
X8	Email address	X8Y1	X8Y2	X8Y3	X8Y4	X8Y5	X8Y6	X8Y7	*	*	X8Y10
X9	Cell number	X9Y1	X9Y2	X9Y3	X9Y4	X9Y5	*	*	*	*	X9Y10
X10	Relationship status	X10Y1	X10Y2	X10Y3	*	X10Y5	*	*	*	X10Y9	*
X11	School details	X11Y1	X11Y2	X11Y3	*	X11Y5	*	*	*	*	*
X12	University/College details	X12Y1	X12Y2	X12Y3	X12Y4	X12Y5	*	*	*	*	*
X13	Job details	X13Y1	X13Y2	X13Y3	*	*	*	*	*	*	*
X14	Income	*	*	*	*	*	*	*	*	*	*
X15	Credit card details	*	*	*	*	*	*	*	*	*	*

The final sum of variables consisted of the nine platform subscription variables, 80 out of the 150 attribute sharing variables, the variables `Age`, `Gender` and `Victim` and the four remaining relationship status options variables. The final number of variables that remained in the data set after the screening process was therefore 96.

4.3.2.3 Relationship histograms of survey variables

Histograms viewing the relationship between previous identity theft victims versus non-victims of the 96 remaining variables were constructed to identify possible identity theft predictor variables.

In Figure 4.15 to Figure 4.23 the percentage of individuals who have been identity theft victims compared to non-victims are graphed according to platform subscribers and non-users for every social media platform included in the survey.

It is notable that LinkedIn in Figure 4.16, Mxit in Figure 4.21 and Dating Sites in Figure 4.22 are the only social media platforms where subscribers have a higher percentage of victims compared to non-users. These sites are therefore possible predictor variables. It is noteworthy that the graph drawn for other social media platforms in Figure 4.23 has exactly the same percentage of victims for users and non-users and will possibly not have a significant influence on predictions.

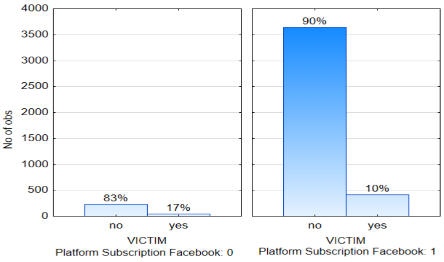


Figure 4.15: Facebook ID Victims.

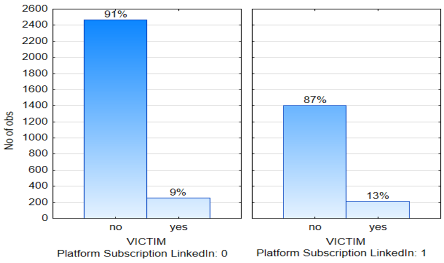


Figure 4.16: LinkedIn ID Victims.

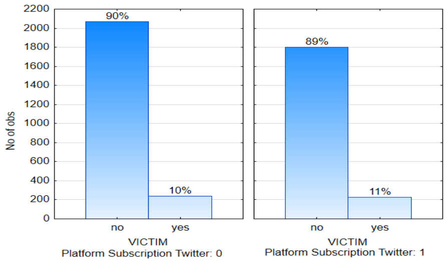


Figure 4.17: Twitter ID Victims.

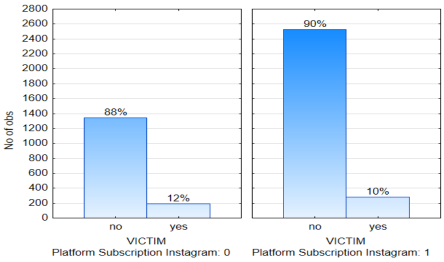


Figure 4.18: Instagram ID Victims.

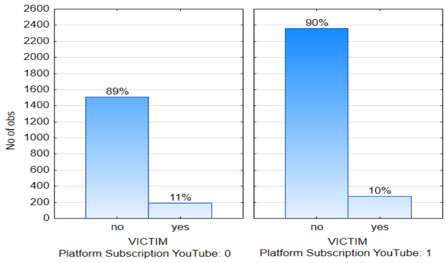


Figure 4.19: YouTube ID Victims.

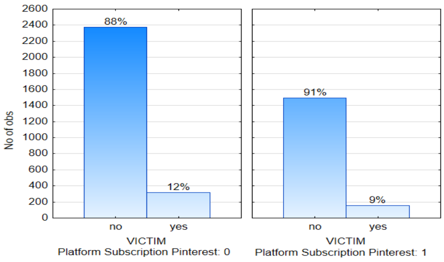


Figure 4.20: Pinterest ID Victims.

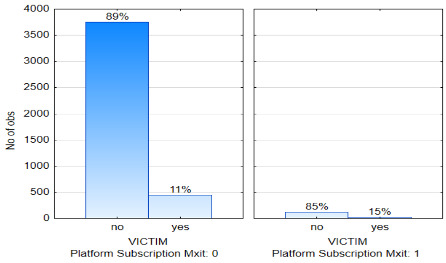


Figure 4.21: Mxit ID Victims.

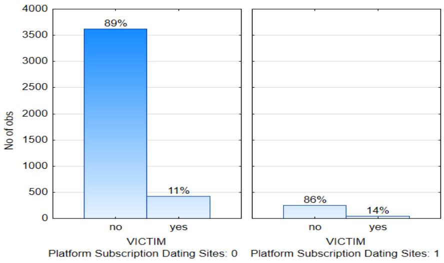


Figure 4.22: Dating Sites ID Victims.

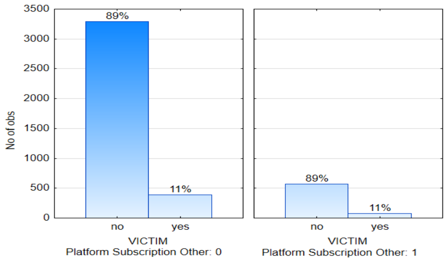


Figure 4.23: Other ID Victims.

Table 4.6 shows a summary of the relationship histogram results of the attributes `Identity Number`, `Physical Address`, `School Details` and `Job Details`. For each of these attributes the percentage victims of non-users and users are listed as observed for the social media platforms 'None', 'Facebook' and 'LinkedIn'. These were the attributes that displayed significant results, because for the majority of social media platforms that they were shared on, these attributes presented a much higher victim percentage among users, when compared to non-users. The only incidence that differed from the recurrent trend was: `School Details` on Facebook. The platform option 'None' points to the fact that for the attributes listed in Table 4.6, the victim percentage was much higher when shared on some social media sites, compared to when not shared on any social media platform. The variables in Table 4.6, except for `School Details` on Facebook, are therefore possible predictor variables.

Table 4.6: Noteworthy Victim Relationships for Certain Attributes.

% victims (non-users) — % victims (users)		Platform Revealed On		
		Y1	Y2	Y3
Attribute Revealed		None	Facebook	LinkedIn
X3	Identity number	22%—10%	10%—19%	11%—18%
X7	Physical address	16%—10%	10%—14%	10%—19%
X11	School details	11%—10%	12%—10%	10%—14%
X13	Job details	13%—8%	10%—12%	9%—14%

In Table 4.7, the social media platforms that demonstrated notable results, are listed along with the percentage of non-user victims versus user victims per attribute shared on them. The no platform option, 'None', is a meaningful variable, because the percentage of non-user victims represents the attributes shared on some or other social media platform and the user victims percentage stands for the attributes not shared on any social media platform. Keeping this in mind it is conspicuous that the victim percentage for every single attribute, except `Birthday` and `Gender`, is higher when shared on some type of social media platform. The exceptions `Birthday`, which had equal percentages for non-user and user victims, and `Gender` that shows a slightly higher victim percentage when not shared on social media, are both unusual occurrences. The SAFPS information, used to determine the expected variables in the CRISP business understanding phase, clearly foresees Age and Gender to be predictor variables. The only variables on 'Facebook' with higher user victim percentages than non-user victim percentages were Identity Number, Physical Address and Job Details. Furthermore, it seems that 'Facebook' users have lower victim percentages compared to non-users. 'LinkedIn' and 'Dating Sites', however, testify to higher user victim percentages than non-user victim percentages for every single attribute. Other social media platforms not included in the survey could be added by participants under the name 'Other'. These platforms revealed that the percentage victims for the users of these platforms are higher than the percentage for non-users for all attributes except Name, which had equal percentages. The possible predictor variables resulting from Table 4.7 were therefore: Identity Number, Physical Address and Job Details shared on 'Facebook'; all the relevant attributes shared on 'LinkedIn' and 'Dating Sites'; all the attributes not shared on any platform, except for Birthday and Gender and finally all the relevant attributes shared on 'Other' platforms, excluding the attributes Name and Gender.

Table 4.7: Noteworthy Victim Relationships for Certain Platforms.

% victims (non-users) — % victims (users)		Platform Revealed On				
		Y1	Y2	Y3	Y9	Y10
	Attribute Revealed	None	Facebook	LinkedIn	Dating Sites	Other
X1	Name	*	17%—10%	9%—13%	11%—13%	11%—11%
X2	Surname	11%—7%	15%—10%	9%—13%	11%—12%	11%—13%
X3	Identity number	22%—10%	10%—19%	11%—18%	*	*
X4	Birthday	11%—11%	14%—10%	10%—14%	*	11%—14%
X5	Gender	11%—13%	16%—10%	10%—13%	11%—13%	11%—11%
X6	Race	11%—10%	11%—10%	10%—13%	11%—16%	11%—15%
X7	Physical address	16%—10%	10%—14%	10%—19%	*	*
X8	Email address	11%—10%	11%—10%	10%—13%	*	11%—16%
X9	Cell number	12%—10%	11%—11%	11%—12%	*	11%—13%
X10	Relationship status	11%—10%	11%—11%	11%—15%	11%—16%	*
X11	School details	11%—10%	13%—10%	10%—14%	*	*
X12	University/college details	11%—9%	14%—10%	10%—14%	*	*
X13	Job details	13%—8%	10%—12%	9%—14%	*	*

Figure 4.24 displays the average age of victims and the average age of non-victims. According to the expected variable information in the business understanding phase it is anticipated that the average age of victims is between 26 and 38 years. The results in Figure 4.24 support this expectation with the average age of non-victims being 24.25 years and that of victims 27.85 years, which is much closer to the anticipated age. It is strongly expected that Age is the greatest predictor variable.

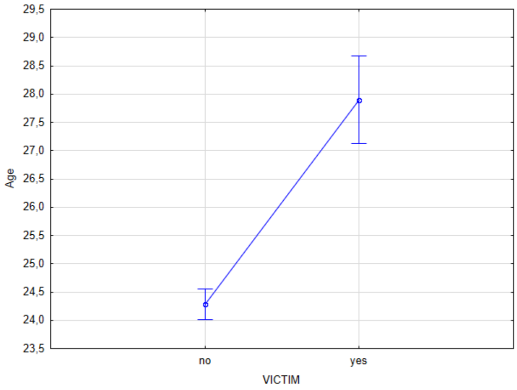


Figure 4.24: Variation in Age of Victims and Non-Victims.

The results received from the SAFPS in the business understanding phase revealed that 69% of all the victims recorded were males. In Figure 4.25 the percentage of male victims is 3% higher than the percentage of female victims. The variable Gender is therefore firmly believed to be a predictor variable.

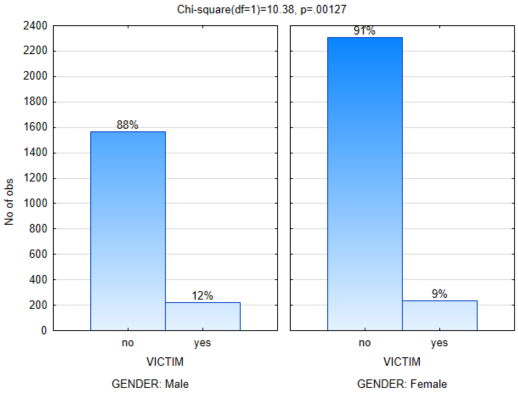


Figure 4.25: Number of Male and Female Victims and Non-Victims.

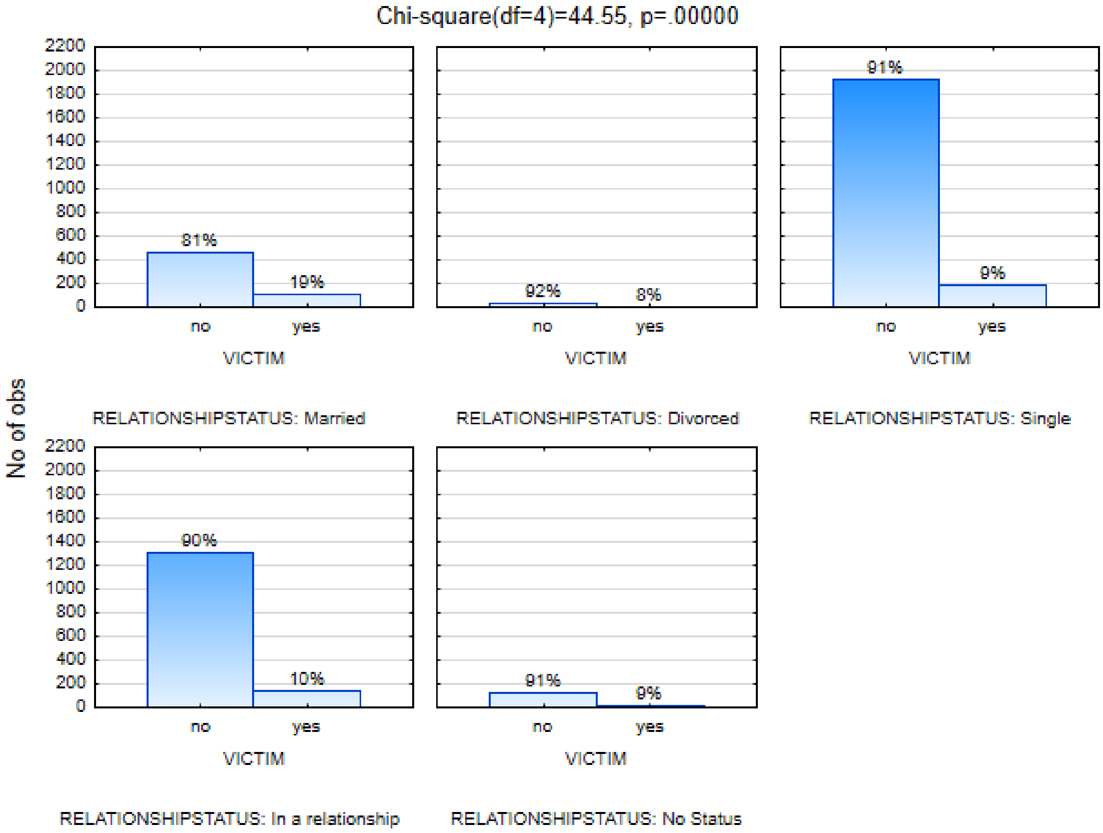


Figure 4.26: `Relationship Status` Occurrences for Victims and Non-Victims.

Figure 4.26 displays victims and non-victims according to their relationship status. The relationship status, 'Married', is by far the variable with the highest victim percentage and is thus expected to be a predictor variable. This observation can possibly be connected to age, as `Age` is expected to be a predictor variable and married people generally represent the older part of the population. The remaining relationship statuses show more or less the same results.

4.3.2.4 Scatterplot of sensitivity against specificity

Considering all the variables above, `Age` stood out to be the predictor variable that is expected to be the most influential and it was therefore decided to construct a scatterplot of *sensitivity* against *specificity* of the variables `Age` and `Victim`.

Sensitivity is the probability to correctly identify an individual as a victim and *specificity* the probability to correctly identify the person as a non-victim. *Sensitivity* and *specificity* are inversely proportional and the crux of the operation is to determine the point that takes both *sensitivity* and *specificity* into account and best benefits both.

In Figure 4.27 it is seen that the point in age that best fits *sensitivity*, at 0.39, and *specificity*, at 0.55, is 23.5 years. It is therefore more likely for individuals to be identified as victims when they are older than 23.5 years and non-victims when younger.

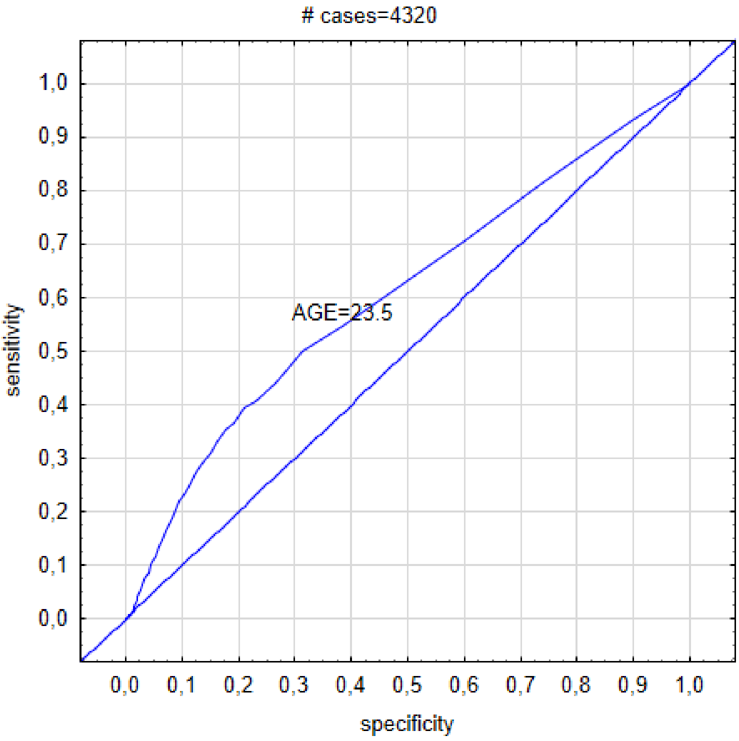


Figure 4.27: Scatterplot for Variables `Age` and `Victim`.

4.3.3 CRISP phase 3: Data preparation

In this phase the final preparation of the data set is done by the handling of outliers and missing data.

4.3.3.1 Outlier identification

Outliers can be identified by either graphically viewing the data or by making use of statistical tests. Box plots are typically used to present the outliers in continuous data and histograms in categorical data.

The only continuous variable in the data set was `Age`. A box plot was constructed for this variable. As illustrated by Figure 4.28, the non-outlier range of the variable `Age` is 18 to 34 years. The outliers and extremes are all greater than 34 years and are due to the small percentage of staff members in the population that are generally older than the students. It was therefore decided to not remove the outliers and extremes as they are not random.

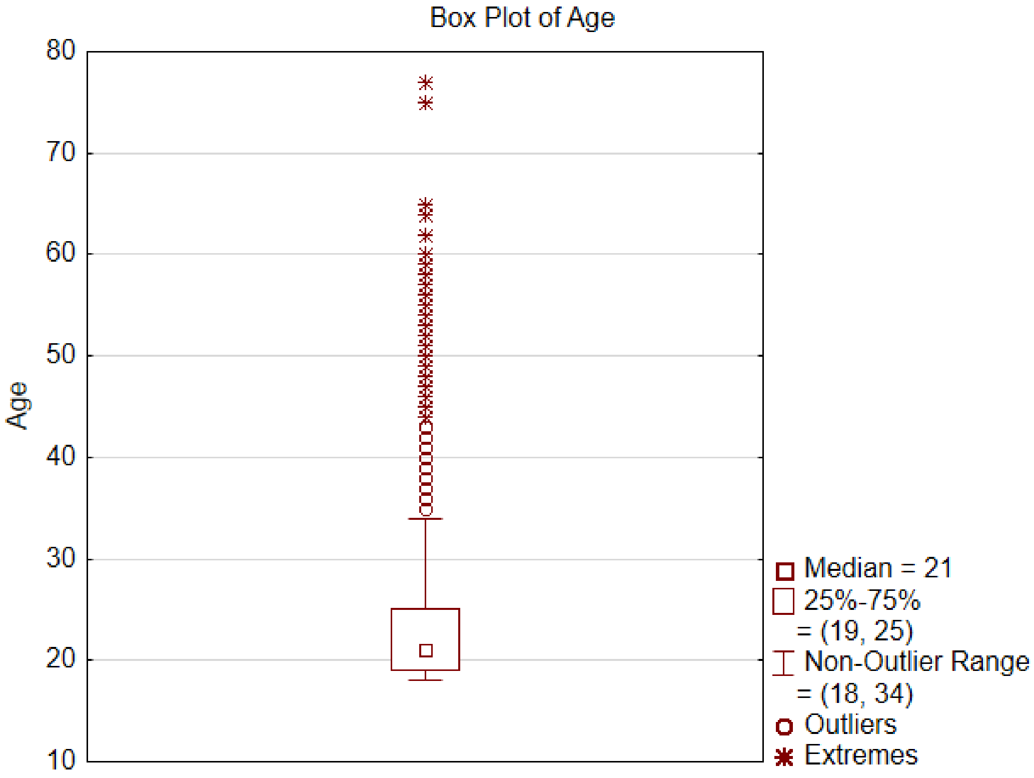


Figure 4.28: Box Plot of `Age`.

The histograms presented in the data understanding phase were used to identify the outliers of the categorical variables. The categorical outliers were then dealt with in Section 4.3.2.2. Except for `Relationship Status` that could take on one of five possible fixed values, all other variables were binary. The variables that remained after the screening process were presented in Table 4.5.

4.3.3.2 Missing data

Most data mining algorithms do not consider the missing entries and simply ignore them. This is problematic, because missing data frequently includes valuable data on other variables. If the missing data points are simply ignored, it can cause the loss of data and perhaps biased results. Statistica offers tools that substitutes missing data. Substitution methods include the mean, median or specific value method. The problem with substituting data is that there is a possibility that these methods can falsely decrease variance, which can affect correlation. Another approach is to use the k-nearest neighbours method where the value, k, is selected and then the k cases most similar to the missing data are used to replace the blank spaces (Electronic Version: StatSoft, 2013).

The survey that was used to collect the data for the research project was constructed in such a manner that participants could not proceed with the survey if they did not complete the compulsory questions. This function made provision for thorough answers and eliminated the risk of having missing data. No missing data tools were therefore needed for any of the compulsory questions.

4.3.3.3 Other data problems

According to Electronic Version: StatSoft (2013) attention should be given to the following additional problems:

- **Sparse Data:** Variables that consist of too many missing values. These variables are not valuable to a model due to the lack of observations. It does not add tangible information and should therefore be removed.

  In the survey sent to SU members, participants were optionally requested to provide their monthly income. Due to the lack in responses the variable `Income` was classified as sparse data and removed from the data set.

- **Invariant Data:** Variables with little or no variability contribute nothing to a model and are therefore not valuable to model building. Invariant variables should be removed from data.

  All participants were asked to give consent and without this permission their responses were not viable and therefore not recorded. The remaining responses all had identical values for the variable `Consent`. The variable consent was thus identified as invariant data and removed from the data set.

- **Duplicate Records:** Duplicate records will skew analysis. The influence multiple entries of the same record will have on a data set is not equal to the actual weight of the record. Only one of these identical records should remain in the data set; the rest must be removed.

  The survey was constructed in such a manner that duplicate records were not possible. Every participant could only submit the survey once.

4.3.4 CRISP phase 4: Modelling

In this section the modelling phase of the CRISP cycle took place. Various models for classification type problems were applied to the research data.

4.3.4.1 Determination of models for classification of identity theft victims and non-victims

A classification type problem is when a certain case must be assigned to a preordained group based on the observed attributes of the case that are related to the group. It was therefore decided to investigate the different techniques applicable to classification type problems, as the goal of this phase was to classify individuals as identity theft victims or non-victims. In the Literature Review in Chapter 2, it was found that techniques such as the discriminant analysis, induction and decision trees, neural networks and genetic algorithms were used to group data according to known classes (Rohanizadeha & Moghadam, 2009). In order to determine which would be the best fit for the research study the following techniques were therefore explored:

1. **Discriminant Analysis** – studies the differences between two or more groups, such as identity theft victims and non-victims in the case of this study, with respect to numerous variables concurrently. The basic assumptions of this technique are that all data cases must belong to only one of the two or more groups; there must be at least two cases per group; the number of discriminant variables must not exceed the total number of cases minus two; discriminant variables are not allowed to be linear combinations of other discriminating variables and should be measured at the interval level; covariance matrices for the various groups must be more or less equal and finally all groups must be from a population with a normal distribution on the discriminating variables. (Klecka, 1980)

2. **Induction and Decision Trees** – Decision trees are used to classify cases according to certain classes. Decision trees have nodes that represent either a class name or an attribute test that branches cases according to possible test outcomes and then the divided subsets are solved further by each test outcome with subtrees (Utgoff, 1989). Classification and regression tree properties include the ability to process many different data types such as numerical, categorical, censored, multivariate data and dissimilarity matrices; complex problems are modelled easily; missing values are handled without losing too much data; and trees are constant to monotonic transformations of predictor variables (De'Ath, 2007).

3. **Neural Networks** – are according to Electronic Version: StatSoft (2013) a non-parametric modelling tool that represents the function of biological neurons to learn from data. The model uses weights and neurons to detect complex relationships among variables and performs well with rough data. Neural Networks are applicable to classification, regression, time series and clustering tasks (Electronic Version: StatSoft, 2013). Elements or nodes used in Neural Networks are typically nonlinear and analog and they do not assume the shape of the underlying distributions as well as traditional statistical classifiers (Lippmann, 1987).

4. **Genetic Algorithms** – are part of the metaheuristic methods that model natural evolution processes. It is used to solve complex engineering optimisation problems, which contain a large search space. Genetic Algorithms evolve a population in a systematic manner in order to find the best solution by making use of computational processes. (Bagchi, 1999)

It was seen that the data mining techniques that were applicable to the research were the discriminant analysis; induction and decision trees; and neural networks. Genetic algorithms are typically used for optimisation problems and were therefore not relevant.

A study conducted on credit scoring was used to determine which of the three remaining techniques was most applicable to the study. The study by Lin & McClean (2001) used the following techniques to score 8 000 customers as either good or bad credit customers: Discriminant Analysis, Logistic Regression, Neural Networks, Classification and Regression Trees, and Multivariate Adaptive Regression Splines (MARS). The dependent variable for the study was the credit status of customers, which could either be good or bad, and the data set further contained nine variables namely gender, age, marriage status, educational level, occupation, job position, annual income, residential status and credit limits (Lin & McClean, 2001). The study by Lin & McClean (2001) was similar to this research as it consisted of the same objective to build a predictive model that classified cases according to a categorical dependent variable. The methods used in the credit scoring study were therefore applicable in this study. Lin & McClean (2001) concluded that the commonly used techniques like discriminant analysis and logistic regression were frequently dismissed due to their strict model assumptions and neural networks due to its lengthy training process, interpretative difficulties and the general struggle to determine the importance of input variables. The analytical results showed that CART and MARS presented better credit scoring accuracies and lower Type II errors than Discriminant Analysis, Logistic Regression, Neural Networks and Support Vector Machines (Lin & McClean, 2001).

It was decided that the data mining technique best suited for the research study was therefore decision trees as it included the CART method. The results founded by Lin & McClean (2001) eliminated Neural Networks and the Discriminant Analysis. Decision trees will subsequently be discussed.

4.3.4.2 Different types of decision trees

Decision trees are predictive models used in data mining. Classification trees are used for classification type problems and regression trees for regression type problems. In this study the research problem is a classification task. Classification trees classify a case (such as a survey response) into a predefined class (such as an identity theft victim or non-victim) based on attributes (such as attributes shared on social media) (Rokach & Maimon, 2015). Classification trees are widely used in various fields like finance, marketing, health, engineering, education, *etc.*

The advantages of decision trees are that they are self-explanatory; they accompany both nominal and numeric input variables; they have the ability to represent any discrete value; data sets that contain errors or missing values are handled; and finally they are non-parametric models, which means there are no assumptions on the space distribution or classifier structure. Disadvantages however include that the majority of models require discrete values for the target variable; they perform best if a few very relevant variables exist opposed to many complex ones; the over-sensitivity of the training set can cause instability; fragmentation problems may arise; and the handling of missing values is sometimes a battle (Rokach & Maimon, 2015).

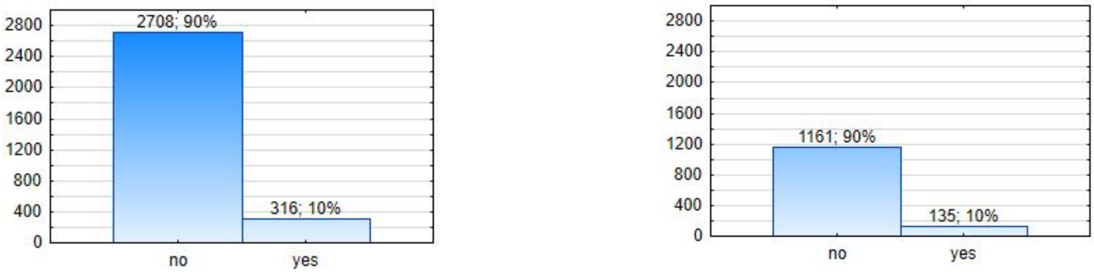
According to Rokach & Maimon (2015) some of the popular decision trees' induction algorithms are ID3, C4.5, CART, CHAID and QUEST. CHAID performs well with survey data, because of the data's categorical nature (Electronic Version: StatSoft, 2013). In an ecological analysis study done by De'Ath (2007) it was found that boosted trees were more interpretable than neural networks. The technique demonstrated the correlations between predictors very well and was found to be an excellent prediction model (De'Ath, 2007). The results from a study conducted by Rodriguez-Galiano *et al.* (2012) on land-cover classification concluded that the Random Forest model outperformed simple decision tree models as it amplified variation between the diverse categories of the research area.

It was therefore decided that the models; CART, CHAID, Random Forest and Boosted Trees were most applicable to the research study's data type and testified to good performance in previous studies. These trees were ideal as they are included in the Statistica tree building tools (Electronic Version: StatSoft, 2013). These trees are now applied and described in their own subsections.

4.3.4.3 Survey data sampling for model building

The data that was prepared in the data preparation phase of the CRISP cycle, was sampled before it was used for model building. Data was randomly divided into two samples; a 70% training sample and a 30% testing sample. Data sampling is necessary for the validation phase of the CRISP cycle and it is a safeguard against overfitting (Electronic Version: StatSoft, 2013). The training sample of data is used to construct the models and to identify patterns and the testing sample of data measures the performance of the model relative to the training data.

It was decided to use stratified random sampling, because the proportion of survey respondents who were identity theft victims is quite small. Stratified random sampling selects even outcomes from both groups of the strata variable, `Victim`, to ensure that the model building will give enough attention to the critical event, which is the prediction of identity theft victims. As seen in Figure 4.29a and Figure 4.29b, both the train and test samples contain even proportions (10%) of the strata variable.



(a) Train Sample.

(b) Test Sample.

Figure 4.29: Stratified Random Sampling

4.3.4.4 Classification and regression trees

Classification and Regression Trees (CART) is a nonparametric data mining algorithm popularised by Breiman *et al.* (1984) and used to generate either classification or regression trees. Classification trees are applied to data with a categorical dependent variable and regression trees to data with a numeric dependent variable. In this study the variable of interest, `Victim`, is categorical in nature and therefore the exploration of the classification trees technique follows.

The CART data analysis determines the correlation between the variable of interest and a large set of possible predictor variables (Sut & Simsek, 2011). Tree branches or splits are made by variables that best predict the variable of interest. Every tree node can be split into two child nodes and all variables are seen as independents when splits are calculated. The stopping rules of CART then determine the size of the tree (Electronic Version: StatSoft, 2013).

Misclassification of the variable of interest is inevitable since no model is perfect. Some misclassifications are worse than others. For example, it is worse misclassifying an individual as not an identity theft victim when the individual actually is one than when misclassifying the person as a victim when not a victim. Statistica accounts for misclassification in the CART model tool by assigning misclassification costs to the variable of interest (Electronic Version: StatSoft, 2013).

It was noted that misclassification costs were not needed for the CART model in the research study. This was tested by assigning a misclassification cost of two to the misclassifying of victims as non-victims, the result of this action was that the model classified 100% of the cases as victims, which was not feasible, because as demonstrated in Figure 4.14, 89% of the cases were actually non-victims. The misclassification cost for the model was therefore set to equal 1 for both of the dependent variable outcomes as seen in Table 4.8.

Table 4.8: CART Misclassification Costs.

Misclassification Costs.		
	Observed no	Observed yes
Predicted no		1
Predicted yes	1	

It is recommended by Breiman *et al.* (1984) to use the Gini index to decrease the impurity when splitting for classification. According to Sut & Simsek (2011), the Gini index becomes zero when a single class is present at a node, meaning that all the cases at a node are from the same class if the Gini index is zero. The Gini index was therefore selected as the goodness of fit parameter. The Centre for Statistical Consultation at Stellenbosch University advised that Statistica performs better when the option for equal prior probabilities is selected as opposed to when prior probabilities are estimated. The researcher did a preliminary investigation by comparing the results of the prospective model for both options and found that estimated priors presented better results.

According to Ripley (1996), when misclassification costs are equal and priors estimated, decision trees are pruned with the minimal cost-complexity cross validation pruning method. The stopping rule was therefore the pruning on misclassification error along with the stopping parameters ‘minimum number of cases’ and ‘maximum number of nodes’ that were set to the Statistica default values. Table 4.9 presents a summary of all the model input parameters.

Table 4.9: CART Parameter Tuning Values.

CART Parameters	
Parameter	Values
Goodness of fit	Gini
Prior Probabilities	Equal
Min n of Cases	400
Max n of nodes	1000

The test sample estimate was selected for the validation of the model. The total number of cases were divided into two samples. A train sample containing

70% of the cases that was used to construct the predictor variables and a test sample of the remaining cases to test the performance of the trained model.

The chain of optimal trees, theory being explained in Breiman *et al.* (1984), that was found for the CART model according to their *cross validation (CV)* and resubstitution costs or learning costs are exhibited in Figure 4.30. It illustrates that resubstitution costs decrease as the tree sizes increase. This is due to the misclassification rate that improves with the increase of the number of terminal nodes. *V-fold cross validation* is performed, when using the *cost-complexity pruning* method, as each split is added to the tree to determine the CV costs. It is important to notice that the absolute cost value is not what is being observed, the trend is the aspect of importance.

Tree number four, which had three terminal nodes, was the best tree size option, because it was located at the point where CV cost started to level out and where the resubstitution cost has not dropped to a point where predictive accuracy could start to be jeopardised by overfitting.

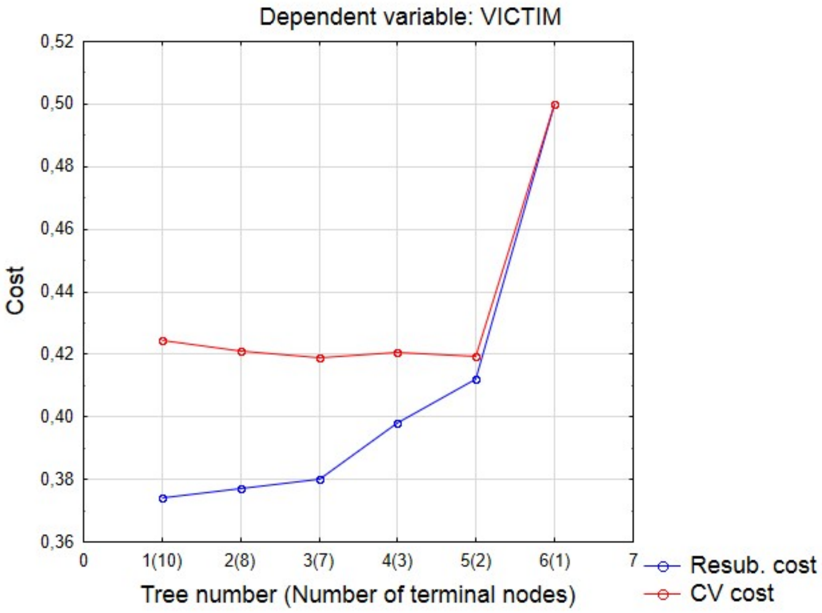


Figure 4.30: CART Cost Sequence.

Figure 4.31 presents the chosen tree with its three terminal nodes. Splits were made for the variable `Age` and the variable `Physical Address` on platform option

'None'. If respondents were older than 26.5 years, or if they were younger than 26.5 years and their physical address was shared on some social media platform, they were classified as identity theft victims.

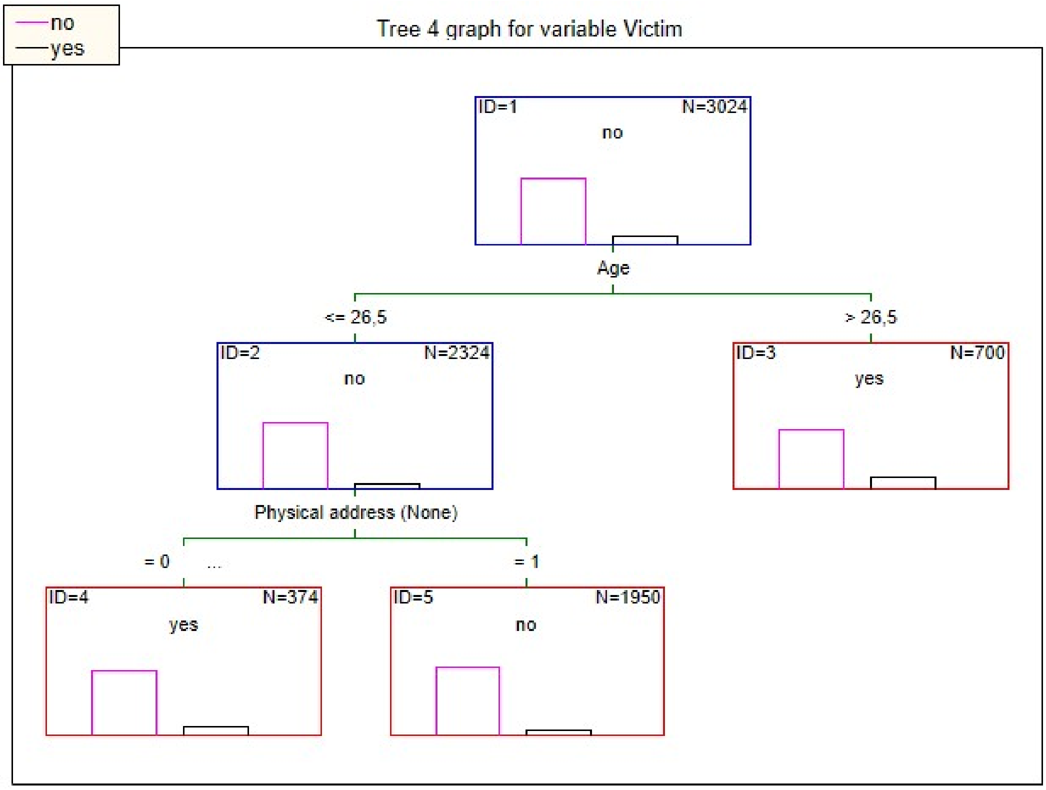


Figure 4.31: Selected CART Tree with Three Terminal Nodes.

Table 4.10 lists the top 20 predictor variables according to a 0–100 scale as determined by the calculation described in Breiman *et al.* (1984). It is observed that the predictor variable with the highest ranking score was `Age`, which was expected in the CRISP data understanding phase.

Table 4.10: CART Variable Importance.

Variable Importance		
n	Variable Name	Rank
1	Age	100
2	Relationship Status (Married)	91
3	ID Details (Facebook)	60
4	Address (None)	57
5	ID Details (None)	56
6	Job Details (None)	53
7	Address (Facebook)	46
8	Relationship Status (Single)	43
9	Platform Subscription (Mxit)	41
10	Job Details (LinkedIn)	35
11	Name (Pinterest)	33
12	Name (LinkedIn)	31
13	Race (Pinterest)	29
14	University Details (LinkedIn)	29
15	Surname (LinkedIn)	26
16	School Details (LinkedIn)	26
17	Surname (Pinterest)	26
18	Gender	25
19	Platform Subscription (LinkedIn)	25
20	University Details (Instagram)	25

The prediction results for the train sample response variable, `Victim`, is presented in Table 4.11. More than half of the observed victims were classified correctly and almost 70% of the non-victims were classified correctly. The model detected victims quite well, when considering the small number (10%) of victims in the actual data sample.

Table 4.11: CART Train Sample Prediction Results.

Train Prediction Results		
Variable: Victim	Predicted no	Predicted yes
Observed no	1804 (66.62%)	146 (33.38%)
Observed yes	904 (46.20%)	170 (53.80%)

The testing sample displayed very similar prediction results to the training sample as observed in Table 4.12.

Table 4.12: CART Test Sample Prediction Results.

Test Prediction Results		
Variable: Victim	Predicted no	Predicted yes
Observed no	769 (66.24%)	62 (33.76%)
Observed yes	392 (45.93%)	73 (54.07%)

Train and test sample cross-tabulation was used to determine if the right size tree was found based on how well the tree performed when classifying the test data (Electronic Version: StatSoft, 2013). The cross-tabulation results for the train sample and test samples are presented in Figure 4.32 and Figure 4.33 respectively. The two samples indicated very similar results and it was therefore concluded that the right tree size was chosen.

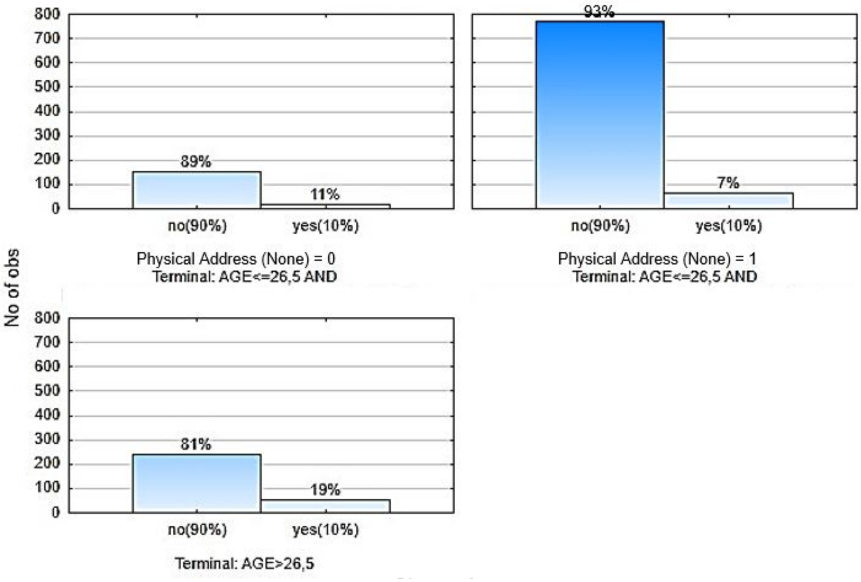


Figure 4.32: Cross-tabulation Results for Training Sample.

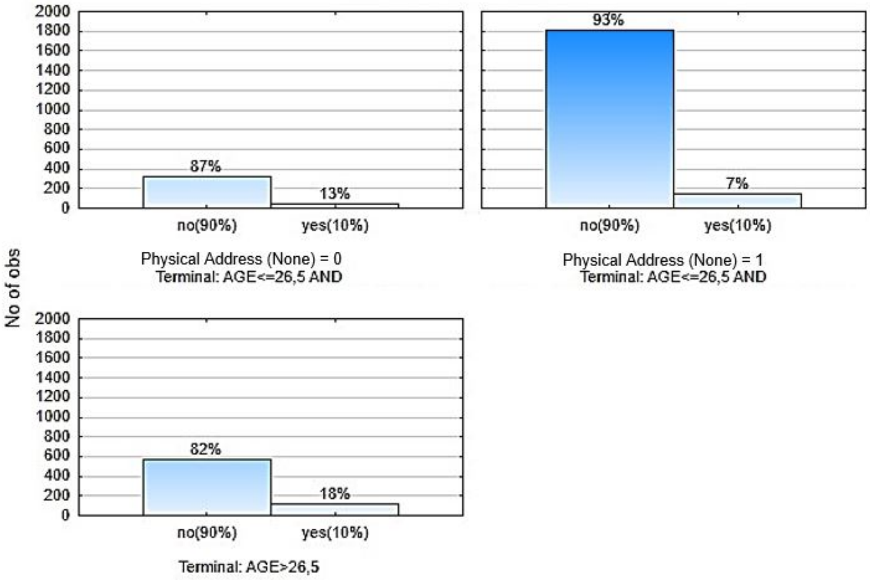


Figure 4.33: Cross-tabulation Results for Testing Sample.

4.3.4.5 Boosted trees for classification

Boosted trees compare a sequence of simple binary trees according to their misclassifications. For every step the boosted trees algorithm calculates the optimal partitioning of the data as well as the deviations of the observed values from the residuals for each partition. The next tree is then fitted to the preceding tree's residuals, to identify another partition that will reduce the residual variance for the data. By voting the best tree from these simple trees, a final classification is made. This technique has developed into one of the top methods for predictive data mining over the last few years (Electronic Version: StatSoft, 2013).

The default tree options to build a boosted tree model in Statistica are presented in Table 4.13.

Table 4.13: Default Boosted Tree Options.

Default Boosted Tree Options			
Learning Rate	Number of additive terms	Subsample proportion	Test Data Proportion
0.1	200	1	0.3

Stopping parameters control the complexity of the tree. The default stopping parameters for boosted trees in Statistica are given in Table 4.14.

Table 4.14: Default Boosted Tree Stopping Parameters.

Default Stopping Parameters			
Minimum n of cases	Minimum n in child node	Maximum n of levels	Maximum n of nodes
15	1	10	3

The goal of the model is to best predict identity theft victims as victims and non-victims as non-victims. It is however much more costly to incorrectly classify a victim as a non-victim than it is to mistake a non-victim for a victim. Misclassification costs account for situations like these. The first parameter to determine was therefore misclassification cost. Models were built for the misclassification costs listed in Table 4.15. The default boosted tree options in Table

4.13 and default stopping parameters in Table 4.14 were used in the initial model building.

By comparing standard errors and prediction results, the misclassification cost of 70 for misclassifying 'Yes' as 'No', yielded the best results. The training set resulted in a standard error of 0.001079, 78.69% of non-victims classified correctly and 38.92% of victims classified correctly. The testing set presented results that differed only a little with a standard error of 0.001694, 79.24% of non-victims classified correctly and 40.74% of victims classified correctly. These were the results obtained before parameter tuning commenced. The train and test samples indicated similar results, which indicated good model performance, but further investigation was still required in terms of the parameters.

Table 4.15: Boosted Tree Misclassification Cost Options.

Victim	
Observed no - Predicted yes	Observed yes - Predicted no
1	1; 10; 20; 30; 40; 50; 60; 70; 80; 90; 100

In Electronic Version: StatSoft (2013) learning rates of 0.1 and smaller are recommended as they tend to produce the top prediction models. Subsample proportion indicates the proportion of data points randomly selected at each iteration without replacement (Elith *et al.*, 2008). The Centre for Statistical Consultation at SU advised the use of all the data points, meaning a subsample of one, for smaller data sets like the one for this study, but Elith *et al.* (2008) reckon that subsample proportions improve model performance and therefore used a subsample of 0.5 in their study. The parameter values listed in Table 4.16 were therefore chosen to be tested for model building.

Cross-validation, as recommended by Elith *et al.* (2008) for small data sets, was used to determine the best combination of parameters. After numerous models were run, the parameter values that yielded the best prediction of identity theft victims were determined. The two parameters first resolved were *learning rate* and *subsample proportion*. It was observed that a learning rate of 0.05 in combination with a subsample proportion of one yielded the best set of train and test sample prediction results for the variable `Victim`. The parameters 'maximum

n of levels', 'minimum n of cases' and 'minimum n of nodes' all have an affect on tree complexity (Elith *et al.*, 2008).

Table 4.16: Boosted Tree Parameter Tuning Values.

Parameter Tuning	
Parameter	Values
Learning Rate	0.1, 0.05
Subsample proportion	0.5, 1
Minimum n of cases	15, 30
Maximum n of levels	3, 4, 5, 6, 10, 15, 20
Maximum n of nodes	3, 6, 9, 10, 11, 12, 13, 14, 15
Number of trees	100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1500

The trial and error process was used to test the different values for the parameters by comparing the standard errors calculated by the various models. The parameter combination with the lowest standard error was for the model with parameter values: 'maximum n of levels' = 5, 'minimum n of cases' = 15 and 'minimum n of nodes' = 12. The model yielded a standard error of 0.0011 for the training sample and a standard error of 0.0015 for the test sample.

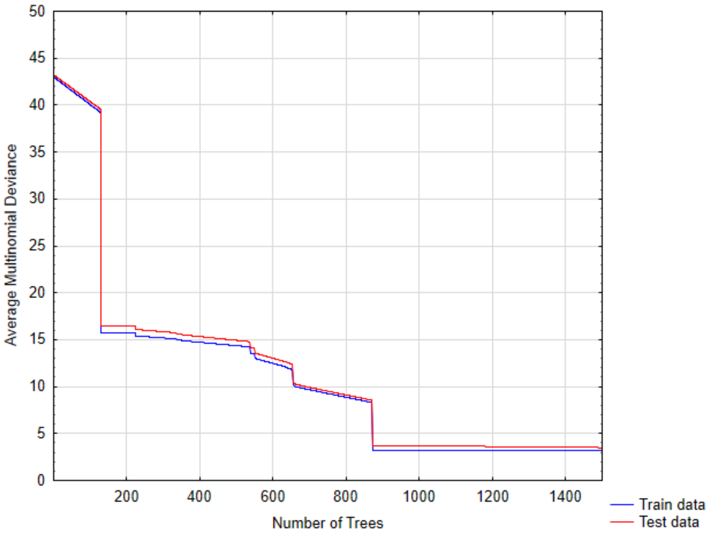


Figure 4.34: Boosted Optimal Trees.

The final parameter to determine was the number of trees. The optimal number of trees was calculated as 900 for the model constructed with the above mentioned parameter values. As illustrated in Figure 4.34, the average multinomial deviance did not improve much after a number of 900 trees. Table 4.17, Table 4.18 and Table 4.19 display a summary of the final parameter values that were used for final model building.

Table 4.17: Boosted Tree Options.

Boosted Tree Options			
Learning Rate	Number of additive terms	Subsample proportion	Test Data Proportion
0.05	200	1	0.3

Table 4.18: Boosted Stopping Parameters.

Stopping Parameters				
Minimum n of cases	Minimum n in child node	Maximum n of levels	Maximum n of nodes	Number of trees
15	1	5	12	900

Table 4.19: Boosted Tree Misclassification Costs.

Misclassification Costs		
	Observed no	Observed yes
Predicted no		70
Predicted yes	1	

Table 4.20 lists the top 20 predictor variables according to the 0–100 scale determined by the calculation described in Breiman *et al.* (1984) as previously mentioned during the CART model construction.

Table 4.20: Boosted Tree Variable Importance.

Variable Importance		
n	Variable	Rank
1	Age	100
2	University Details (Instagram)	59
3	Relationship Status (Married)	38
4	ID Details (Facebook)	25
5	Platform Subscription (Mxit)	22
6	School Details (Instagram)	21
7	ID Details (None)	18
8	Platform Subscription (LinkedIn)	14
9	Job Details (LinkedIn)	14
10	Name (LinkedIn)	13
11	Surname (LinkedIn)	13
12	Gender (LinkedIn)	11
13	Name (Instagram)	11
14	University Details (LinkedIn)	10
15	Platform Subscription (Instagram)	10
16	Email (LinkedIn)	9
17	Job Details (None)	9
18	Relationship Status (In a relationship)	9
19	Platform Subscription (Youtube)	8
20	School Details (LinkedIn)	8

The training sample's prediction results are presented in Table 4.21. The model predicts observed non-victims as non-victims much more accurately than observed victims as victims. This is due to the small proportion of only 11% of victims in the original data set.

Table 4.21: Boosted Tree Train Sample Prediction Results.

Train Prediction Results		
Variable: Victim	Predicted no	Predicted yes
Observed no	1923 (71.01%)	785 (28.99%)
Observed yes	160 (50.63%)	156 (49.37%)

In Table 4.22 the prediction results of the testing sample is given. The prediction results of the testing sample are slightly less accurate compared to those of the training sample, but still very close. The model performed well and the parameters chosen for the Boosted Tree model were therefore deemed feasible.

Table 4.22: Boosted Tree Test Sample Prediction Results.

Test Prediction Results		
Variable: Victim	Predicted no	Predicted yes
Observed no	808 (69.60%)	353 (30.40%)
Observed yes	71 (52.59%)	64 (47.41%)

4.3.4.6 Chi Square Automatic Interaction Detection

The Chi Square Automatic Interaction Detection (CHAID) model is one of the oldest classification methods. The model is a type of decision tree that uses the Bonferoni p–value testing. CHAID implements multi-level splits instead of the binary splits performed in CART. Like the CART model, the CHAID model only builds one tree. This method is often used for direct marketing and it requires large data sets for reliable analysis.

The default values for the CHAID model in Statistica are presented in Table 4.23. All explanatory variables in the CHAID model must be categorical (Electronic Version: StatSoft, 2013). The only continuous variable `Age` was therefore categorised into nine age groups: 18–20, 21–30, 31–40, 41–50, 51–60, 61–70, 71–80, 81–90 and 91–100.

Table 4.23: Default CHAID Stopping Parameters.

Default Stopping Parameters			
Minimum n	Maximum n of nodes	Probability for splitting	Probability for merging
432	1000	.05	.05

Table 4.24: CHAID Misclassification Cost Options.

Victim	
Observed no - Predicted yes	Observed yes - Predicted no
1	1; 10; 20; 30; 40; 50; 60; 70; 80; 90; 100

The misclassification costs listed in Table 4.24 were tested with the default parameters as input parameters to the models. The best classification cost was then determined by the comparison of the various model's prediction results. It turned out that misclassification costs had no effect on the prediction outcomes of the model and it was thus decided that the misclassification costs for the final model were set to equal for the response variable `Victim` as seen in Table 4.25.

Table 4.25: CHAID Misclassification Costs.

Misclassification Costs		
	Observed no	Observed yes
Predicted no		1
Predicted yes	1	

The remaining parameters that needed tuning and the considered values are listed in Table 4.26. It was seen that if the probability for splitting was set to zero, all cases were classified as non-victims, which is not viable. All the options with a splitting and merging probability bigger than or equal to 0.15 presented identical prediction results, but as the splitting and merging probabilities decreased to less than 0.15, the model gradually performed better when the train sample prediction results were compared to the test sample results. It was determined that the best combination for the parameters were the default values, which were a 0.05 probability for splitting and a 0.05 probability for merging.

Table 4.26: CHAID Parameter Tuning Values.

Parameter Tuning	
Parameter	Values
Minimum n	1, 15, 30, 45, 60, 75, 100, 200, 300, 400, 500
Maximum n of nodes	500, 1000, 2000
(Probability for splitting; Probability for merging)	(1;0), (0;1), (0.7;0.3), (0.3;0.7), (0.5;0.5), (0;0), (1;1), (0.05;0.05), (0.1;0.1), (0.15;0.15), (0.2;0.2)

The various investigated values for the parameter ‘maximum number of nodes’ did not have any effect on the prediction results, risk estimates or standard errors of the train and test samples. The default value of 1 000 was therefore again chosen for the final model. The final parameter that needed tuning was ‘minimum n’. It was noticed that the model performance, measured by comparing the train and test sample prediction results, increased as the value for ‘minimum n’ increased until it reached 60. After that the model performance weakened drastically. The final value chosen for parameter 'minimum n' was therefore 60. Table 4.27 shows a summary of the final stopping parameters.

Table 4.27: CHAID Stopping Parameters.

Stopping Parameters			
Minimum n	Maximum n of nodes	Probability for splitting	Probability for merging
60	1000	0.05	0.05

The final model, constructed with the parameters in Table 4.27, delivered the risk estimates and standard errors listed in Table 4.28.

Table 4.28: CHAID Risk Estimates.

	Risk Estimate	Standard error
Train	0.102513	0.005515
Test	0.104938	0.008513

The predictor variables presented in Table 4.29 were ranked according to the Bonferoni adjusted p–value of the corresponding chi-square test calculated for the first node of the tree. The final tree consisted of 43 nodes.

Table 4.29: CHAID Variable Importance.

Variable Importance			
n	Variable	Chi–square	p–value
1	Age	49.36	0.00000
2	Relationship Status (Married)	44.23	0.00000
3	Job Details (None)	18.28	0.00002
4	ID Details (Facebook)	16.53	0.00005
5	Job Details (LinkedIn)	16.15	0.00006
6	ID Details (None)	15.16	0.00010
7	Relationship Status (Single)	13.65	0.00022
8	University Details (LinkedIn)	13.08	0.00030
9	Name (LinkedIn)	13.00	0.00031
10	School Details (LinkedIn)	11.99	0.00053
11	Surname (LinkedIn)	11.45	0.00072
12	Physical Address (None)	11.36	0.00075
13	Platform Subscription (LinkedIn)	10.93	0.00095
14	Birthday (LinkedIn)	10.81	0.00101
15	Physical Address (LinkedIn)	10.50	0.00119
16	Physical Address (Facebook)	9.91	0.00164
17	Gender	8.98	0.00273
18	Platform Subscription (Mxit)	8.73	0.00313
19	Name (Pinterest)	8.24	0.00410
20	Surname (Pinterest)	6.90	0.00861

The final prediction results for the response variable `Victim` for the train and test samples are displayed in Table 4.30 and Table 4.31 respectively.

Table 4.30: CHAID Train Sample Prediction Results.

Train Prediction Results		
Variable: Victim	Predicted no	Predicted yes
Observed no	2697 (99.59%)	11 (0.41%)
Observed yes	268 (84.81%)	48 (15.19%)

Table 4.31: CHAID Test Sample Prediction Results.

Test Prediction Results		
Variable: Victim	Predicted no	Predicted yes
Observed no	1140 (98.19%)	21 (1.81%)
Observed yes	128 (94.81%)	7 (5.19%)

Table 4.30 and Table 4.31 illustrate that the model classified 15.19% of the observed victims correctly in the training sample and only 5.19% in the test sample. The model is therefore not a very good predictor for victims. This can be explained by the small proportion of victims in the data set and the fact that CHAID works best for large data sets and the data set for this study only consisted of 4 320 cases. It is however a very good predictor for non-victims.

4.3.4.7 Random forest

The random forest model is a non-parametric model that consists of a series of simple trees. For every single tree a different training sample is randomly selected from the data set, with replacement, and then used to predict a classification (Touw *et al.*, 2013). The random forest then predicts the classification that was predicted by the majority of the trees, the forest (Electronic Version: StatSoft, 2013).

The default parameter options for the random forest model in Statistica are shown in Table 4.32 and the default stopping parameters in Table 4.33. The Gini index is used as impurity measure in the Random Forest model (Touw *et al.*, 2013).

Table 4.32: Default Random Forest Options.

Default Random Forest Options		
Number of predictors	Number of trees	Subsample proportion
7	100	1

Table 4.33: Default Random Forest Stopping Parameters.

Default Random Forest Stopping Parameters			
Minimum n of cases	Minimum n in child node	Maximum n of levels	Maximum n of nodes
432	1	10	100

According to Svetnik *et al.* (2003), the only parameter that really needs tuning in random forest is ‘number of predictors’. The number of predictors controls how many independent predictors are considered at each node (Electronic Version: StatSoft, 2013). The “number of trees’ parameter determines the model complexity and should be considered to avoid overfitting. The two parameters that were therefore tuned were ‘the number of trees’ and ‘the number of predictors’. The values that were tested are displayed in Table 4.34.

Table 4.34: Random Forest Parameter Tuning Values.

Parameter Tuning	
Parameter	Values
Number of trees	100, 500, 1000
Number of predictors	3, 7, 14, 28, 42, 56, 70, 84, 95

All the values for the parameter 'number of predictors' were tested. It was observed that the value that performed the best when prediction results for the response variable `Victim` were compared for the train and test samples, was 42. The value for the 'number of trees' parameter that presented the best performing model was 100. It is therefore safe to say that overfitting occurred for the values 500 and 1 000. The default values for misclassification were equal. The effect of assigning a misclassification cost of two for misclassifying victims as non-victims, was tested, and it resulted in a model that performed much weaker.

After considering all of the above the decided parameters for the final Random Forest model were the misclassification costs displayed in Table 4.35 and the random forest options as listed in Table 4.36.

Table 4.35: Random Forest Misclassification Costs.

Misclassification Costs		
	Observed no	Observed yes
Predicted no		1
Predicted yes	1	

Table 4.36: Random Forest Options.

Random Forest Options		
Number of predictors	Number of trees	Subsample proportion
42	100	1

Figure 4.35 illustrates that the misclassification rate of the train data sample

is less than the misclassification rate of the test data sample for all of the 100 trees.

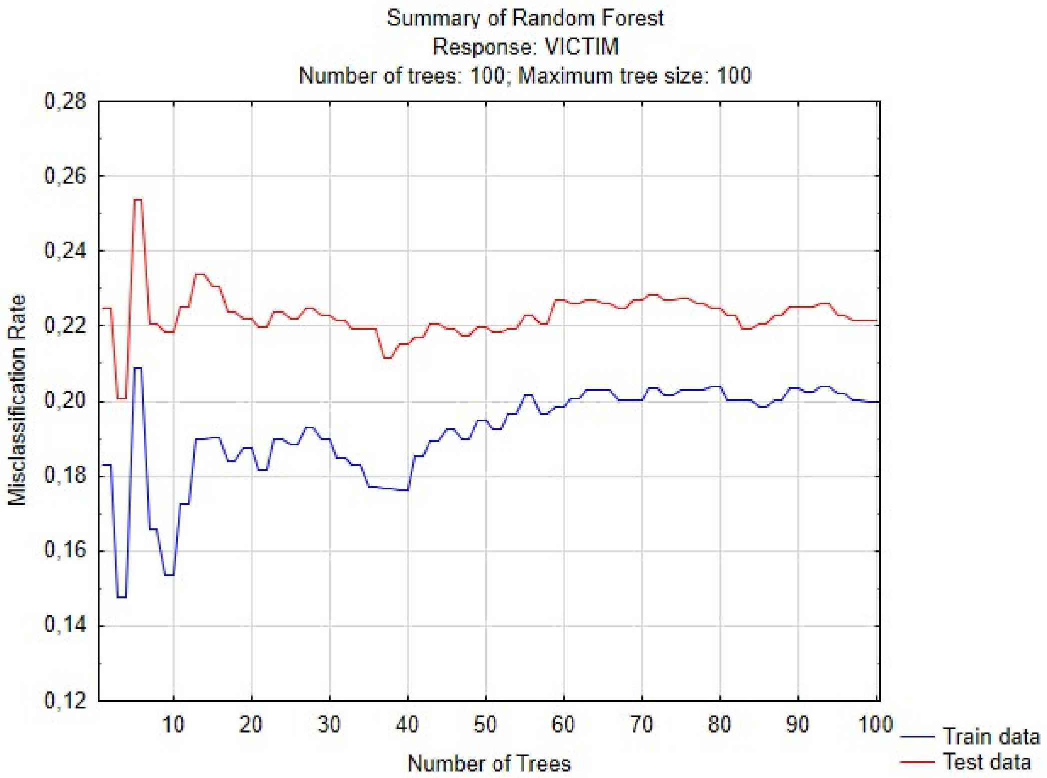


Figure 4.35: Random Forest Misclassification Rate.

As calculated by the constructed Random Forest model, the 20 variables that were seen as the best predictors for the response variable `Victim` were the variables listed in Table 4.37.

Table 4.37: Random Forest Variable Importance.

Variable Importance		
n	Variable	Rank
1	Age	100
2	Gender	40
3	Race (Facebook)	39
4	Cell Number (None)	39
5	Race (None)	39
6	Email (Facebook)	38
7	Cell Number (Facebook)	38
8	Relationship Status (None)	38
9	Surname (Youtube)	38
10	Relationship Status (Facebook)	38
11	Platform Subscribtion (Youtube)	37
12	Job Details (Facebook)	37
13	Name (Youtube)	37
14	Platform Subscribtion (Pinterest)	36
15	Surname (Twitter)	36
16	Email (Youtube)	36
17	Gender (LinkedIn)	36
18	Birthday Details (LinkedIn)	35
19	Gender (Instagram)	35
20	Platform Subscription (Twitter)	35

The prediction results for the train and test samples are seen in Table 4.38 and Table 4.39. The model predicted non-victims much better than victims, but considering the small number of responders that were actually observed as victims, the model predicted victims quite well. The training and test samples presented similar predictions for the classification of non-victims and the test sample performed slightly weaker than the train sample with the classification of victims. It can therefore be concluded that the model performed well in general and that the input parameters were therefore feasible.

Table 4.38: Random Forest Train Sample Prediction Results.

Train Prediction Results		
Variable: Victim	Predicted no	Predicted yes
Observed no	2267 (83.71%)	441 (16.29%)
Observed yes	164 (51.90%)	152 (48.10%)

Table 4.39: Random Forest Test Sample Prediction Results.

Test Prediction Results		
Variable: Victim	Predicted no	Predicted yes
Observed no	958 (82.52%)	203 (17.48%)
Observed yes	83 (61.48%)	52 (38.52%)

4.3.5 CRISP phase 5: Evaluation

In this phase all models were compared and it was determined which model performed the best, relative to the other models, with the classification of victims and non-victims of identity theft.

In the credit scoring study done by Lee *et al.* (2006), the decision tree models were evaluated by the comparison of the training and test sample prediction results. Table 4.40 and Table 4.41 are summaries of the prediction results of the train and test samples for all of the models tested in this study; Boosted Tree, CART, CHAID and Random Forest. When models were built using the training data sample, the CHAID model performed the best with the correct classification of non-victims with a 99.59% success rate and the CART model with the correct classification of victims with a 53.80% success rate as illustrated in Table 4.40. The model with the best overall classification was the Random Forest model with an average correct classification of 65.91%.

Table 4.40: Summary of Prediction Results for Training Set.

Summary of Prediction Results for training set				
Correctly Classified Victim	Boosted Tree	CART	CHAID	Random Forest
Category: No	71.01%	66.62%	99.59%	83.71%
Category: Yes	49.37%	53.80%	15.19%	48.10%
Average Correct Classification %	60.19%	60.21%	57.39%	65.91%

In Table 4.41 it is seen that the models that correctly classified the dependent variable best when the test data was used in the models, were once again CHAID with a 98.19% success rate for non-victims and CART with a 54.07% success rate for victims. The CART and Random Forest model indicated very similar average correct classifications, but once again the Random Forest model performed the best with 0.36%, which is a total number of 16 cases more correctly classified.

Table 4.41: Summary of Prediction Results for Testing Set.

Summary of Prediction Results for testing set				
Correctly Classified Victim	Boosted Tree	CART	CHAID	Random Forest
Category: No	69.60%	66.24%	98.19%	82.52%
Category: Yes	47.41%	54.07%	5.19%	38.52%
Average Correct Classification %	58.51%	60.16%	51.69%	60.52%

In general Type I errors, when ID non-victims are wrongly classified as victims, have lower misclassification costs than Type II errors, when ID-victims are misclassified as non-victims (Lee *et al.*, 2006). In this study the misclassification costs for the models CART, CHAID and Random Forest were set to equal as it did not affect the prediction results. Boosted Trees, however performed best with a misclassification cost of 70 for Type II errors to 1 for Type I errors. It is therefore needed to review the Type I and Type II error results presented in

Table 4.42. The model with the lowest Type II error percentage was the CART model. The misclassification costs assigned for the Type II errors in the Boosted Tree model helped as it presented the second lowest Type II error percentages. The CHAID model displayed the lowest Type I error percentages, followed by the Random Forest model with the second lowest Type I error percentages. It is however important that Type I and Type II errors are taken into account and therefore the model with the lowest combination of errors was the Random Forest model.

Table 4.42: Type I and Type II Errors of the Four Models.

	Train Sample Results		**Test Sample Results**	
	Type I error	**Type II error**	**Type I error**	**Type II error**
CART	33.38%	46.2%	33.76%	45.93%
Boosted Trees	28.99%	50.63%	30.4%	52.59%
CHAID	0.41%	84.81%	1.81%	94.81%
Random Forest	16.29%	51.9%	17.48%	61.48%

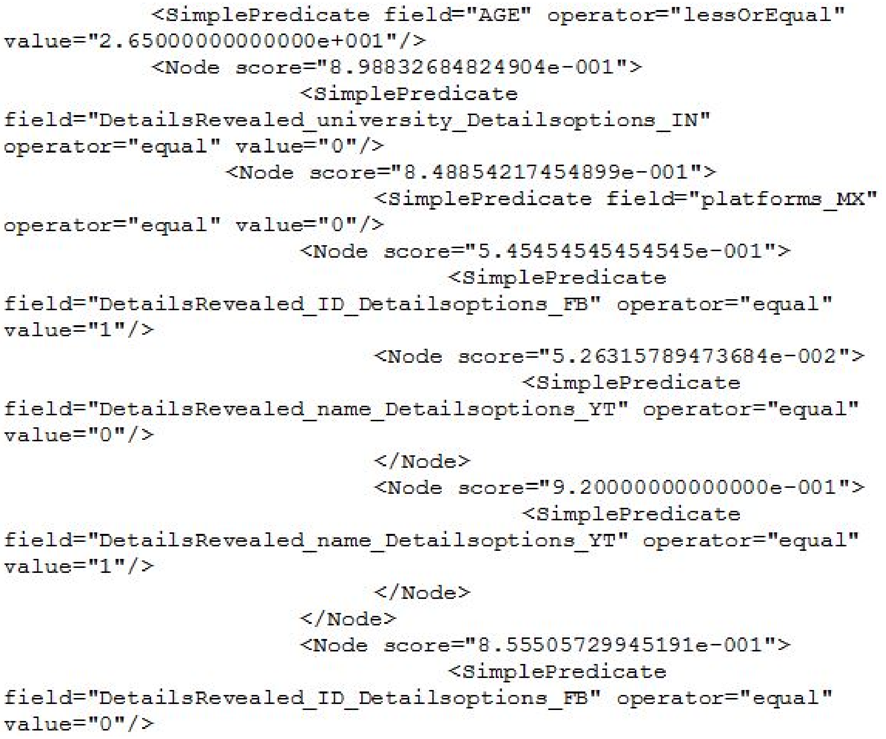
These observations were recorded during the building phase of the models and were however not the best reflection of how well the models were actually performing. In order to compare the models more effectively and to score new data with the models, the Rapid Deployment Tool in Statistica, described in Nisbet *et al.* (2009), was used.

4.3.5.1 Application of the rapid deployment tool

The program Statistica has the capability of generating deployment code for the evaluation of prediction models and the scoring of new data. The four deployment code options are: Statistica Visual Basic for dataminer workspaces, C or C++ language for custom deployment tools, Predictive Model Markup Language (PMML) script for rapid deployment or code for deployment to the Statistica enterprise (Electronic Version: StatSoft, 2013).

All top commercial and open source data mining tools produce and support PMML code (Guazzelli, 2015). This function enables users to rapidly implement prediction models in practice without a hassle (Guazzelli, 2015). It was therefore decided to use PMML code in the rapid deployment tool that Statistica offers. The tool is able to load the PMML code of multiple models, compare the models with lift and gains charts and make predictions with new data (Electronic Version: StatSoft, 2013).

The PMML code for the CART, CHAID, Boosted Trees and Random Forest models were extracted and uploaded to the rapid deployment tool in Statistica. Refer to the enclosed CD for the PMML code files of the various models. Figure 4.36 is a snapshot of the PMML code for the Boosted Tree model. The tool ran the various models on the full data set, which included the training and test sample data.



```
           <SimplePredicate field="AGE" operator="lessOrEqual"
value="2.65000000000000e+001"/>
           <Node score="8.98832684824904e-001">
                      <SimplePredicate
field="DetailsRevealed_university_Detailsoptions_IN"
operator="equal" value="0"/>
                <Node score="8.48854217454899e-001">
                           <SimplePredicate field="platforms_MX"
operator="equal" value="0"/>
                      <Node score="5.45454545454545e-001">
                                 <SimplePredicate
field="DetailsRevealed_ID_Detailsoptions_FB" operator="equal"
value="1"/>
                           <Node score="5.26315789473684e-002">
                                      <SimplePredicate
field="DetailsRevealed_name_Detailsoptions_YT" operator="equal"
value="0"/>
                           </Node>
                           <Node score="9.20000000000000e-001">
                                      <SimplePredicate
field="DetailsRevealed_name_Detailsoptions_YT" operator="equal"
value="1"/>
                           </Node>
                      </Node>
                      <Node score="8.55505729945191e-001">
                                 <SimplePredicate
field="DetailsRevealed_ID_Detailsoptions_FB" operator="equal"
value="0"/>
```

Figure 4.36: Snapshot of Boosted Trees PMML Code as Example.

Gains charts were used to display the percentage of observations that were correctly classified for the two dependent variable categories (Electronic Version: StatSoft, 2013). The gain percentage is equal to the number of accurately predicted responses relative to the total number of responses in that percentile.

The goal is to maximize the curve from the baseline. In Figure 4.37 the gains chart for the category 'Yes' of the variable `Victim` is displayed. The baseline, which represents no-model, is indicated with a blue line. The results for the four models were very close and almost in line with the baseline. The Boosted Tree model was on the baseline, the CART model performed a little better than the Boosted Tree model, the CHAID model slightly better than the CART model and the Random Forest model closely outperformed CHAID to be the best classifier of victims. The small gain for the four models is subjected to the small proportion, 11% of the total cases, that were observed as victims in the data set. In conclusion it is seen that the prediction of victims with models presents almost the same results as if there was no model. This problem can be solved if more data points that were actual observed victims are included in the data set.

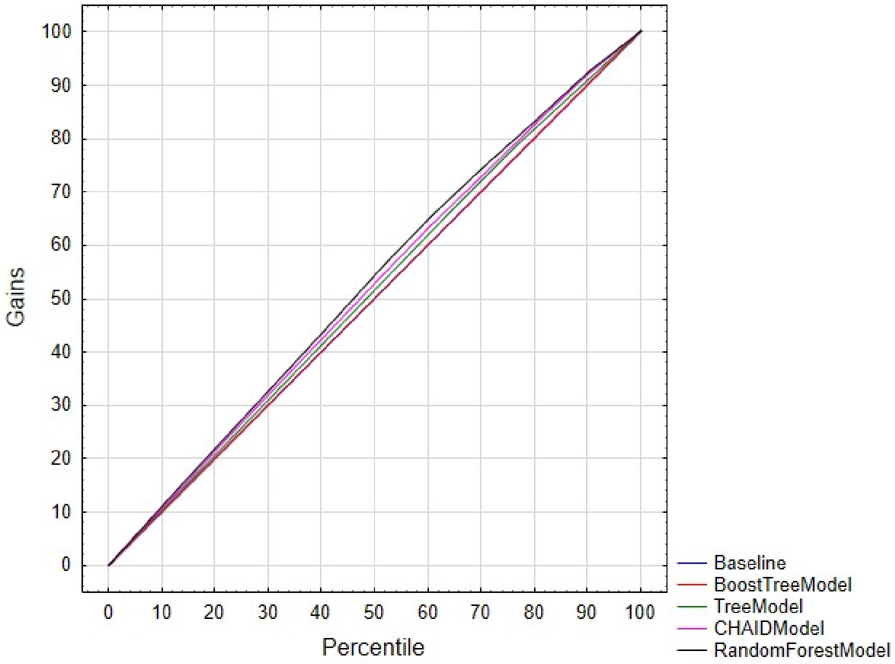


Figure 4.37: Gains Chart for Victim Category: 'Yes'.

In Figure 4.38 the gains chart for the category 'No' of the variable `Victim` is presented. It can be observed that the Random Forest model presented the biggest gain from the baseline. This therefore indicates that the Random Forest model contributes the most accuracy to the correct classification of non-victims

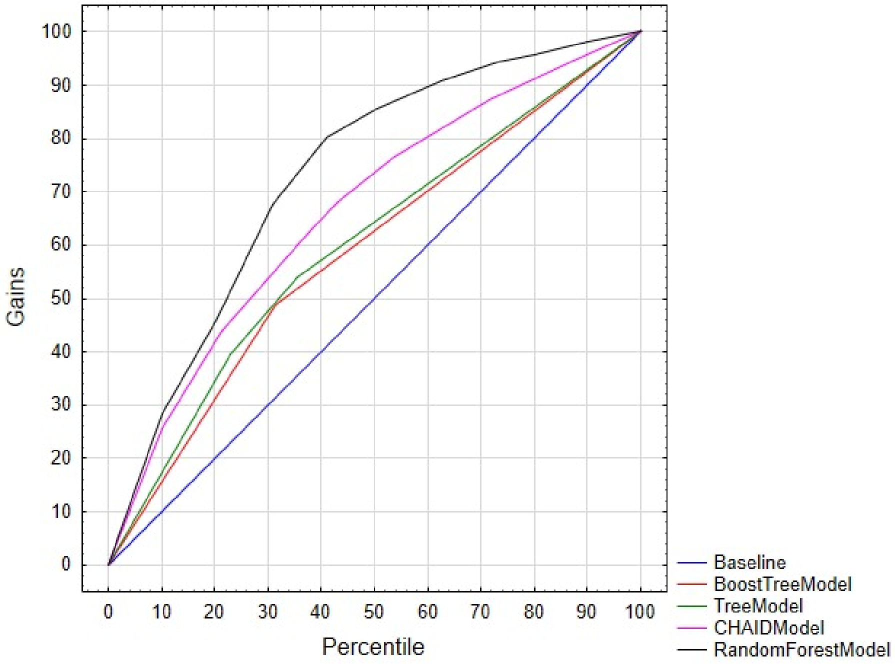


Figure 4.38: Gains Chart for Victim Category: 'No'.

when compared to the other models or the baseline, which represents no model. The model that performed second best was again the CHAID model, but results were not as close as in Figure 4.37 with the classification of victims.

Lift charts are used to present the effectiveness of a model relative to no model (Electronic Version: StatSoft, 2013). The baseline is the percentage of actual observed responses, for the dependent variable category investigated, in the data set. The actual observed values were 89.6% for non-victims and 10.4% for victims. The lift value is then calculated by dividing the expected response when using a predictive model with the expected response when no model is used, the baseline percentage, per cumulative percentile.

It is illustrated in Figure 4.39 that the Random Forest model was the most effective with the prediction of non-victims as it had the highest percentage of expected classifications for every percentile. Figure 4.40 indicates with the lift values that once again the Random Forest model was the most effective with the prediction of victims compared to the CART, CHAID and Boosted Tree models. Once again it is important to note that the scale of the figures are proportional to

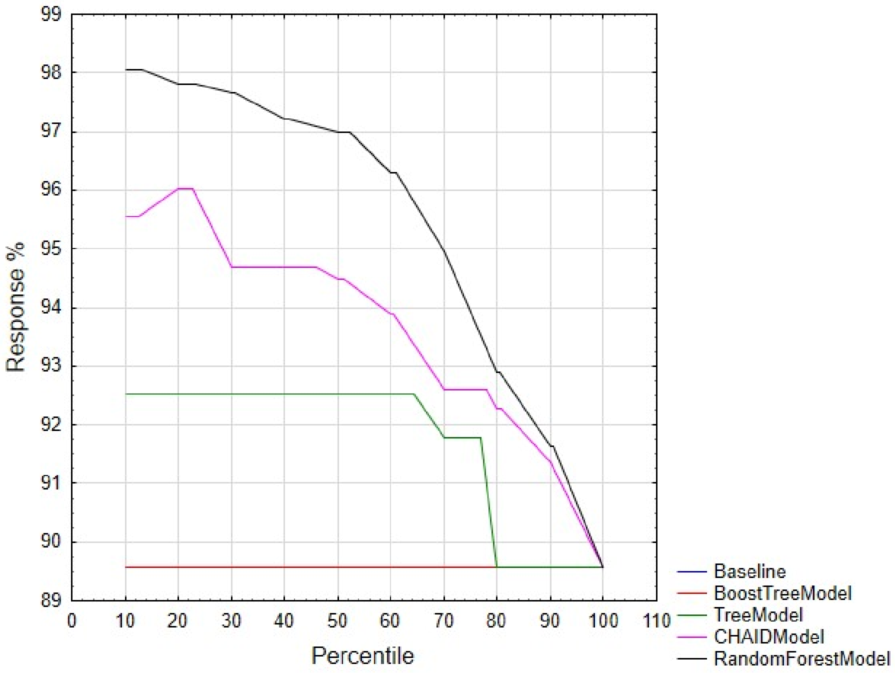


Figure 4.39: Lift Chart for Victim Category: 'No'.

the amount of actual observed values and the figures serve to present the trends in the predictions rather that the absolute prediction values.

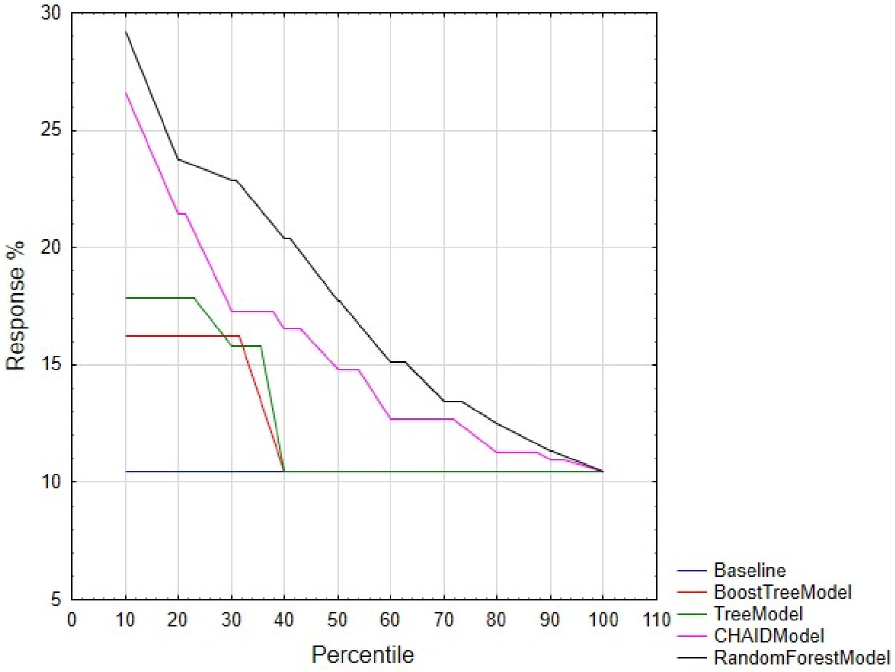


Figure 4.40: Lift Chart for Victim Category :‘Yes’.

The gains charts’ model accuracy presentation and lift charts’ model effectiveness presentation both suggest that the Random Forest model performs better than the CART, CHAID and Boosted Trees models. With these results in combination with the prediction results in Table 4.40 and Table 4.41 it is concluded that the Random Forest model is the prediction model that best classified the dependent variable, `Victim`.

4.3.6 CRISP phase 6: Deployment

In this phase new data is scored with the best performing prediction model, the Random Forest model, determined in the Evaluation phase.

The cut-off date for the survey, discussed in Chapter 3, was the end of May 2016. There were 41 survey responses received after this date, which were not included in the original data set that was used in the data analysis. The 41 late survey responses were retrieved in August and used as the new cases for the deployment of the model. Figure 4.41 shows a histogram of the responses for the variable `Victim`, which is a good representation of the original data set, because the victim percentage of 10% is a good reflection of the common observed proportion.

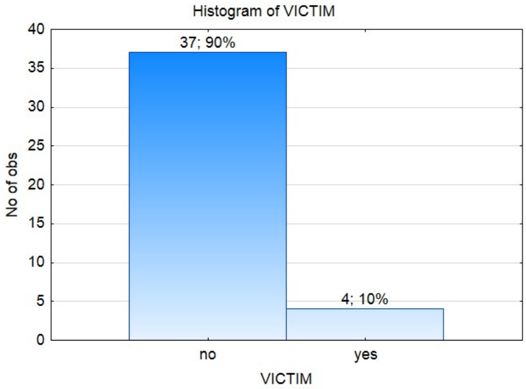


Figure 4.41: Number of Responses for the Variable: `Victim`.

The rapid deployment tool in Statistica was used to apply the PMML code, generated for the Random Forest Model, on the deployment data set. The resulting error rate for the deployment data was 0.146341, showing that the model adapted very well to the new data. The prediction results for the new data are presented in Table 4.43. There were only four victim cases, three of them, 75%, were classified correctly and 86.49% of the non-victims were classified correctly resulting in an overall accuracy of 85.37%. The model performed very well, considering that the average prediction results for the training and test sample data in the Evaluation phase were 43.31% for the correct classification of victims and 83.12% for the correct classification of non-victims.

Table 4.43: Summary of Deployment Prediction Results for the Random Forest Model.

Random Forest Deployment Prediction Results		
	Predicted No	Predicted Yes
Observed No	86.49%	13.51%
Observed Yes	25.00%	75.00%

For a final validation of the model it was decided to apply the PMML code of the three remaining models on the deployment data set. Table 4.44, Table

4.45 and Table 4.46 present the prediction results for the models Boosted Trees, CART and CHAID respectively.

Table 4.44: Summary of Deployment Prediction Results for the Boosted Trees Model.

Boosted Tree Deployment Prediction Results		
	Predicted No	Predicted Yes
Observed No	0%	100%
Observed Yes	0%	100%

Table 4.45: Summary of Deployment Prediction Results for the CART Model.

CART Deployment Prediction Results		
	Predicted No	Predicted Yes
Observed No	51.35%	48.65%
Observed Yes	25%	75%

Table 4.46: Summary of Deployment Prediction Results for the CHAID Model.

CHAID Deployment Prediction Results		
	Predicted No	Predicted Yes
Observed No	97.30%	2.7%
Observed Yes	100%	0%

The Boosted Tree model misclassified all non-victims and the CHAID model misclassified all victims. These two models did therefore not present feasible results. The CART model, like the Random Forest model, correctly classified 75% of the victims, but only classified 51.35% of the non-victims correctly compared to the 86.49% correctly classified non-victims of the Random Forest model. It is thus safe to conclude that the Random Forest model is the prediction model that best satisfies the project objectives listed in the CRISP Business Understanding phase. The associated variables that best predict identity theft victims are therefore the variables presented in Table 4.37 in the CRISP Modelling phase.

4.4 Data mining chapter summary

In this chapter the CRISP data mining model was used as guideline during the processing of the email survey data collected in Chapter 3. Data was screened and prepared for model building. It was decided to use decision trees as prediction models and the four identified models that were applied in the study were CART, Boosted Trees, CHAID and Random Forest. Data was sampled into a training and test set. The models were evaluated by cross-validation of the two sets. The PMML code of each model was generated and imported into the Statistica Rapid Deployment Tool. The Rapid Deploment Tool was used to construct gains and lift charts to compare model's prediction accuracy and effectiveness respectively. The Random Forest model was observed as the model that performed the best. The Random Forest model was therefore used to score 41 new cases in the CRISP Deployment phase. To validate the model, the PMML code of the remaining three models were applied to the deployment data set. It was concluded that the Random Forest model outperformed the CART, CHAID and Boosted Trees models.

Chapter 5

Results

In this chapter the research results are summarised for all of the research steps and an overall conclusion of the results is then given.

5.1 Introduction to results

The data that was required for the research study was discussed in Section 3.1. After a comprehensive literature review and reading of the social media privacy policies included in Appendix A, it was found that the initial research methodology was not practical and a revised research methodology was suggested in Section 3.4. The revised research methodology was followed for the execution of the remaining study.

5.2 Summary of data included in the research

In Chapter 3, Section 3.3 information on identity theft victims that was general and in a depersonalised format was collected from the SAFPS. The information included the age, gender, marital status and income of all identity theft incidents reported in the year 2015. A total number of 2 039 feasible incidents were collected.

Information was then collected on the social media sharing habits and the attributes age, gender, income, relationship status and identity theft victim status via an email survey. The email survey was sent to the SU population and had a 12% response rate. A total number of 4 320 responses were viable.

5.3 Results from the preliminary investigation

Chapter 3, Section 3.5 presented a preliminary investigation that was conducted on the SAFPS and email survey data. This section therefore summarises the results obtained from the investigation.

5.3.1 Preliminary investigation results: SAFPS

The information received from the SAFPS corresponded with that found in literature. The results received on the variables `Age` and `Gender` complied with that found by Erasmus (2015) and Dirk (2015b), which stated that males between the ages of 28 and 40 were prime targets for identity theft. The SAFPS data showed that 69% of the incidents reported were male victims compared to only 31% being female and 45.5% of the victims were between the ages of 30 and 40 years. The data on the variables `Relationship Status` and `Income` did not present significant results due to insufficient data.

It was concluded that `Gender` and `Age` were possible identity theft predictor variables. The data received from the SAFPS would serve as reference data for the email survey data set.

5.3.2 Preliminary investigation results: email survey

Vulnerability scores were calculated for all the survey responses according to the attributes that they have shared and the number of social media platforms the attributes have been shared on. The average vulnerability scores were then grouped for the variables `Age`, `Gender`, `Relationship Status`, `Income` and `Victim`. The results on the variables `Age` and `Gender` for the survey data came as no surprise as they once again agreed with literature and with the SAFPS results that the ages which presented the highest average vulnerability scores were between 28 to 40 years and the average vulnerability percentage was again much higher for male victims when compared to female victims. The relationship status with the highest average vulnerability score was 'Married' and the lowest score was for the 'No Status' option. This could be that the relationship status of married is mostly related to older people and not to students. The variable `Income` contained many missing values and therefore these results were skewed and not considered as noteworthy.

The final variable inspected was `Victim`, which grouped the population according to individuals who have previously been offended by identity theft and people who have never been offended. The derived average vulnerability scores for the dependent variable `Victim` presented the expected results. The average vulnerability score for previous identity theft victims was higher than the score for non-victims.

It was concluded that the collected survey data displayed the expected results and was sufficient to develop a prediction model. The prediction model was therefore a realistic objective for the research study. The variables `Gender` and `Age` consistently presented significant results and were strongly expected to be predictor variables of identity theft in the prediction model.

5.4 Data mining results

In Chapter 4, Section 4.2.1, the email survey data task was determined as a classification type problem. As discussed in Section 4.2.2, the data mining model found most suited for the research study was the CRISP model and the recommended data technique for classification problems was determined as decision trees in Section 4.3.4.1.

The CRISP data mining framework was followed in order to reach the following goals presented in Section 4.3.1:

1. Determine the variables that best predict identity theft victims;
2. Find a predictive model that best classifies victims; and
3. Use the model to make decisions as to whether or not a social media user is at risk of becoming an identity theft victim or not.

The SAFPS data was used as reference variables and the email survey data was used in the data mining process to construct the models.

5.4.1 Data exploration results

The survey data variables were screened graphically with histograms in Section 4.3.2.1. Variables that presented very low responses were excluded from the data set in Section 4.3.2.2. A total number of 96 variables remained after the screening process.

Relationship histograms were then constructed for the remaining 96 variables in Section 4.3.2.3 to determine the expected predictor variables by observing the percentage difference of previous identity theft victims for all the variable options. The process revealed the expected predictor variables listed in Table 5.1. The shared attributes along with the platforms they were shared on, presented in brackets, are listed according to the percentage increase of identity theft victims when the attribute was shared on the social media platform compared to when it was not shared. This was done for social media platform subscription numbers too. The average age of previous identity theft victims was determined as 27.85 years and it was seen that 3% more males than females have been victims. The relationship status option of 'Married' showed an average of 9% more identity theft victims than all of the other relationship status options.

These were only the variables that could potentially be predictor variables. The final predictor variables were however determined by the prediction model that performed the best during the Modelling Phase of the CRISP data mining model.

Table 5.1: Possible Predictor Variables.

	Shared Attributes	% Difference		Shared Attributes	% Difference
1	*Identity number (None)*	12%	**23**	*Birthday (Other)*	3%
2	*Identity number (Facebook)*	9%	**24**	*Job details (Facebook)*	2%
3	*Physical address (LinkedIn)*	9%	**25**	*Cell number (LinkedIn)*	2%
4	*Identity number (LinkedIn)*	7%	**26**	*Name (Dating Sites)*	2%
5	*Physical address (None)*	6%	**27**	*Gender (Dating Sites)*	2%
6	*Job details (LinkedIn)*	5%	**28**	*Cell number (None)*	2%
7	*Job details (None)*	5%	**29**	*University/College details (None)*	2%
8	*Email address (Other)*	5%	**30**	*Surname (Other)*	2%
9	*Race (Dating Sites)*	5%	**31**	*Cell number (Other)*	2%
10	*Relationship status (Dating Sites)*	5%	**32**	*School details (None)*	1%
11	*Relationship status (LinkedIn)*	5%	**33**	*Surname (Dating Sites)*	1%
12	*Physical address (Facebook)*	4%	**34**	*Race (None)*	1%
13	*Name (LinkedIn)*	4%	**35**	*Email address (None)*	1%
14	*Surname (LinkedIn)*	4%	**36**	*Relationship status (None)*	1%
15	*Birthday (LinkedIn)*	4%		**Platform Subscriptions**	**% Difference**
16	*University/College details (LinkedIn)*	4%	**1**	*LinkedIn Subscription*	4%
17	*School details (LinkedIn)*	4%	**2**	*Mxit Subscription*	4%
18	*Surname (None)*	4%	**3**	*Dating Sites Subscription*	3%
19	*Race (Other)*	4%		**Description Attributes**	
20	*Gender (LinkedIn)*	3%	**1**	*Age*	Average 27.85
21	*Race (LinkedIn)*	3%	**2**	*Gender*	Male (3%)
22	*Email address (LinkedIn)*	3%	**3**	*Relationship status (Married)*	9% lead

5.4.2 Prediction model results

After the careful consideration of all the different prediction techniques presented in Section 4.3.4.1 it was decided to process the data with decision trees. The following popular decision tree models that were discussed in Section 4.3.4.2 were included in the research study: CART, CHAID, Random Forest and Boosted Trees. The models were applied to the survey data in Section 4.3.4. The data set was divided into a training sample, which consisted of 70% of the population cases and a testing sample, which consisted of the remaining 30% of the population cases. This was done to enable cross-validation between samples for model evaluation. The results of the four prediction models are discussed in the sections to follow.

5.4.2.1 CART model results

The optimal CART tree, determined by resubstitution and CV costs as seen in Figure 4.31, had three terminal nodes where splits were made for the variable `Age` and the variable `Physical Address` on the platform option 'None'. If individuals were older than 26.5 years, or if they were younger than 26.5 years and their physical address was shared on any social media platform, they were classified as a potential identity theft victim.

The CART model performance was evaluated by the comparison of the train and test sample prediction results. Table 5.2 displays the prediction results for both samples. The train and the test samples illustrated similar results. The model therefore performed quite well.

Table 5.2: CART Model Prediction Results for Train and Test Sample.

	CART Model			
	Train Sample Results		Test Sample Results	
	Predicted no	**Predicted yes**	**Predicted no**	**Predicted yes**
Observed no	66.62%	33.38%	66.24%	33.76%
Observed yes	46.20%	53.80%	45.93%	54.07%

Table 4.10 presents the top 20 predictor variables for the CART model that were determined according to a 0–100 scale as calculated by the method described in (Breiman *et al.*, 1984).

5.4.2.2 Boosted trees model results

The optimal number of trees was calculated as 900 for the Boosted Trees model. In Figure 4.34 it is seen that the multinomial deviance did not improve much after a number of 900 trees. It was determined that the Boosted Trees model performed best when a misclassification cost of 70 was assigned to Type II errors. The same evaluation process completed for the CART model was used to determine model performance for the Boosted Trees model. The results for the training and test samples is presented in Table 5.3. Again the results for the two samples were complementary, which shows that the Boosted Trees model performed soundly.

Table 5.3: Boosted Trees Model Prediction Results for Train and Test Sample.

	Boosted Trees Model			
	Train Sample Results		Test Sample Results	
	Predicted no	Predicted yes	Predicted no	Predicted yes
Observed no	71.01%	28.99%	69.60%	30.40%
Observed yes	50.63%	49.37%	52.59%	47.41%

Table 4.20 presents the top 20 predictor variables calculated by the Boosted Trees model. The variables are again ranked according to the 0–100 scale mentioned during the CART model results discussion.

5.4.2.3 CHAID model results

The CHAID model requests that all variables must be categorical. The only continuous variable `Age` was therefore categorised into nine age groups. The final CHAID tree consisted of 43 nodes.

The CHAID model exposed the prediction results exhibited in Table 5.4. The comparison of the train and test sample data results disclosed the performance of the model. The model detected non-victims of identity theft very accurately for both samples, but the test sample performed weaker than the train sample with the classification of victims. The model was therefore not as stable as the CART and Boosted Trees models were. The poor results for the prediction of victims could be explained by the fact that CHAID works best for large data sets and the data set for this study consisted of 4 320 cases, of which only 475 cases were observed victims.

Table 5.4: CHAID Model Prediction Results for Train and Test Sample.

	CHAID Model			
	Train Sample Results		Test Sample Results	
	Predicted no	Predicted yes	Predicted no	Predicted yes
Observed no	99.59%	0.41%	98.19%	1.81%
Observed yes	84.81%	15.19%	94.81%	5.19%

The predictor variables listed in Table 4.29 were ranked according to the Bonferoni adjusted p–value of the corresponding chi-square tests that were calculated for the first node of the tree.

5.4.2.4 Random forest model results

A total number of 100 trees was built, each with a number of 42 predictors, for the Random Forest model. The performance of the Random Forest Model is demonstrated by the prediction results of the train and test samples on view in Table 5.5. The model performed well with the recognition of identity theft non-victims as the train and test sample results were very close, but with the identification of victims, performance decreased with the lower percentage correctly classified victims in the test sample data.

Table 5.5: Random Forest Model Prediction Results for Train and Test Sample.

	Random Forest Model			
	Train Sample Results		**Test Sample Results**	
	Predicted no	**Predicted yes**	**Predicted no**	**Predicted yes**
Observed no	83.71%	16.29%	82.52%	17.48%
Observed yes	51.90%	48.10%	61.48%	38.52%

As calculated by the constructed Random Forest model, the 20 variables that were seen as the best predictors for the response variable `Victim` were the variables listed in Table 4.37.

5.4.3 Prediction model evaluation results

In the credit-scoring study done by Lee *et al.* (2006), the decision tree models were evaluated by the comparison of the training and test sample prediction results. It was found that this method was suitable for the evaluation of the prediction models applied in this study. Table 5.2, 5.3, 5.4 and 5.5 display the prediction results for the training and test samples of the CART, Boosted Trees, CHAID and Random Forest models respectively. It was however important to not only compare Type I and Type II errors separately, but to compare an overall average prediction of the models. The average results for the training and test samples for all four of the models are therefore presented in Table 5.6. The CHAID model performed the best with the correct classification of non-victims with a 99.59% success rate and the CART model with the correct classification of victims with a 53.80% success rate. The model with the best overall classification was however the Random Forest model with an average correct classification of 65.91% for the training sample and 60.52% for the test sample.

Table 5.6: Average Prediction Results for the Four Decision Tree Models.

Average Classification %	**Training Sample**	**Test Sample**
CART	60.21%	60.16%
Boosted Trees	60.19%	58.51%
CHAID	57.39%	51.69%
Random Forest	65.91%	60.52%

In order to evaluate the models even further, the PMML codes of all the models were generated and applied to the full data set with the Rapid Deployment Tool in Statistica. Gains charts were created to display the prediction accuracy of the four models and lift charts to demonstrate the effectiveness of the models.

Figure 4.38 on page 113 displays the gains chart for the classification of victims. The results for the four models were very close. The small gain for the four models is subjected to the small proportion, 11% of the total cases, that were observed as victims in the data set. The model with the best accuracy in the classification of victims, even though it was by a very small gain, was the Random Forest model. The gains chart for non–victim classification can be seen in Figure 4.37 on page 112. The Random Forest clearly presented the biggest gain and was therefore the model that performed the best in this category. Overall it could be concluded that the Random Forest model was the most accurate in the classification of the dependent variable, `Victim`.

Lift charts were used to present the effectiveness of the models relative to no model. It was illustrated in Figure 4.39 on page 114 that the Random Forest model was the most effective with the prediction of non-victims as it had the highest percentage of expected classifications for every percentile. Figure 4.40 on page 115 indicates with the lift values that once again the Random Forest model was the most effective with the expected prediction of victims compared to the CART, CHAID and Boosted Tree models.

The gains charts' model accuracy presentation and lift charts' model effectiveness exhibition both suggested that the Random Forest model performed better than the CART, CHAID and Boosted Trees models. With these results in combination with the prediction results in Table 5.6 it was concluded that the Random Forest model was the prediction model that best classified the dependent variable, `Victim`.

5.4.4 Deployment and validation of the random forest model

In order to deploy the Random Forest model, new data that contained the same input variables was required. In Section 4.3.6 it is discussed how the data of the 41 respondents, who did not make the survey deadline and which was therefore not included in the data mining data set, was used to create a deployment data set. The histogram in Figure 4.41 presents the responses for the variable, `Victim`, which was found to be a good representation of the original data set, because the victim percentage of 10% is a good reflection of the victim percentage of 11% in the original data mining data set.

The Rapid Deployment Tool was used to apply the trained PMML code of the Random Forest model to the new deployment data set. The prediction error rate for the Random Forest model in the evaluation phase was 0.206250. The prediction error rate is calculated by adding the test sample error to the training sample error. The error rate for the deployment data was 0.146341, showing that the model adapted very well to the new data. The prediction results can be seen in Table 4.43. There were only four victim cases, three of them, 75%, were classified correctly and 86.49% of the non-victims were classified correctly resulting in an overall accuracy of 85.37%. The model performed very well, considering that the average prediction results for the training and test sample data in the Evaluation phase were 43.31% for the correct classification of victims and 83.12% for the correct classification of non-victims.

For a final validation of the model, it was decided to compare the model's ability to score new data by applying the PMML code of the CART, CHAID and Boosted Trees models and observe how they perform relative to the Random Forest Model. As seen in Table 5.7, the Boosted Tree model misclassified all non-victims and the CHAID model misclassified all victims. These two models did not therefore present feasible results. The CART model, like the Random Forest model, correctly classified 75% of the victims, but only classified 51.35% of the non-victims correctly compared to the 86.49% correctly classified non-victims of the Random Forest model. It can therefore be concluded that the Random Forest model was indeed the best prediction model for the classification of identity theft victims.

Table 5.7: Prediction Results on Deployment Data for Decision Trees

Prediction Results	**Correct classification of non-victims**	**Correct classification of victims**
CART	51.35%	75%
Boosted Trees	0%	100%
CHAID	97.3%	0%
Random Forest	86.49%	75%

5.4.5 Overall results discussion and conclusion

During the preliminary investigation of the SAFPS data it was determined that 69% of the identity theft victims reported were males compared to only 31% being female and 45.5% of the victims were between the ages of 28 and 40 years. These two variables were therefore set as reference variables as they represented real identity theft incidents. The email survey data was then investigated and it was found that the average vulnerability on social media sites of male users was higher than the score for female users. The average vulnerabilities for different age groups were highest for ages in the range of 30–40 years. These results therefore complied with the SAFPS expected results. Relationship diagrams were constructed during the data mining of the survey data for the 96 variables that remained after the screening process. The victim percentages for these variables were compared and it was again found that the average age for victims was 28 years and the victim percentage for males was 3% higher than the victim percentage for females. The two most important predictor variables were conistently expected to be `Age` and `Gender` and the Random Forest model confirms the expection to be true as it ranked `Age` as the most important predictor variable and `Gender` as second most important.

Table 5.8 lists the most important predictor variables according to the Random Forest model next to the expected variables that were determined during the data exploration. The actual predictor variables are listed on the left side of the table. The expected predictor variables are listed on the right side of the table and are cross-referenced in bold to indicate which expected variables the prediction model determined to actually be predictor variables. It is observed that 18 of the expected predictor variables turned out to be actual predictor variables in the Random Forest model. It can therefore be concluded that the predictor variables are not as obvious as one would think.

The variables listed in Table 5.8 can therefore be used as a guideline when one creates a social media profile. It is advised that social media users should not share these attributes on social media platforms to avoid the risk of becoming an identity theft victim. According to Alfreds (2015c), South Africa is one of the top three countries internationally with the highest rates of fraud using recycled deceased identities. It is therefore advised to remove these personal details of loved ones from their online social media profiles if something should happen to them.

The PMML code for the Random Forest model can be applied, via a tool such as the Rapid Deployment Tool in Statistica, to score any new case that completed a questionaire with the same input variables. It can then be determined if such an individual classifies as a high risk identity theft victim, when classified as a victim or a low risk identity theft victim, when classified as a non-victim. If an individual classifies as a high-risk victim, they should revise their profile on social media platforms as they have either already been a victim or could be at risk of becoming one.

As Arthur Conan Doyle said, ‘There is nothing more deceptive than an obvious fact.’ When it comes to identity theft, we are all possible victims. The only thing we have control over is our actions, which can either increase or decrease our vulnerability. Be wise. Be informed.

Table 5.8: Final Predictor Variables

ID	Important predictor variables	Rank	Cross	ID	Expected Predictor Variables	Refer to
	According to Random Forest Model	**(Scale 1–100)**	**Referenced**		**Determined by Relationship Diagrams**	**Table**
1	Age	100	**1**	1	Age	Average 27.85
2	Gender (Male)	40	**2**	2	Gender (Male)	3%
3	Race (Facebook)	39		3	Relationship status (Married)	9% lead
4	Cell Number (Any Platform)	39	**31**	4	Identity number (Any Platform)	12%
5	Race (Any Platform)	39	**37**	5	Identity number (Facebook)	9%
6	Email (Facebook)	38		6	Physical address (LinkedIn)	9%
7	Cell Number (Facebook)	38		7	Identity number (LinkedIn)	7%
8	Relationship Status (Any Platform)	38	**39**	8	Physical address (Any Platform)	6%
9	Surname (Youtube)	38		9	Job details (LinkedIn)	5%
10	Relationship Status (Facebook)	38		10	Job details (Any Platform)	5%
11	Platform Subscribtion (Youtube)	37		11	Email address (Other)	5%
12	Job Details (Facebook)	37	**27**	12	Race (Dating Sites)	5%
13	Name (Youtube)	37		13	Relationship status (Dating Sites)	5%
14	Platform Subscribtion (Pinterest)	36		14	Relationship status (LinkedIn)	5%
15	Surname (Twitter)	36		15	Physical address (Facebook)	4%
16	Email (Youtube)	36		16	Name (LinkedIn)	4%
17	Gender (LinkedIn)	36	**23**	17	Surname (LinkedIn)	4%
18	Birthday Details (LinkedIn)	35	**18**	18	Birthday (LinkedIn)	4%
19	Gender (Instagram)	35		19	University/College details (LinkedIn)	4%
20	Platform Subscription (Twitter)	35		20	School details (LinkedIn)	4%
21	Email (Twitter)	35		21	Surname (Any Platform)	4%
22	Gender (Twitter)	34		22	Race (Other)	4%
23	School Details (Facebook)	34		23	Gender (LinkedIn)	3%
24	Gender (YouTube)	34		24	Race (LinkedIn)	3%
25	Surname (Instagram)	33		25	Email (LinkedIn)	3%
26	Email (LinkedIn)	33	**25**	26	Birthday (Other)	3%
27	School Details (LinkedIn)	33	**20**	27	Job details (Facebook)	2%
28	Name (Twitter)	33		28	Cell number (LinkedIn)	2%
29	Job Details (Any Platform)	33	**10**	29	Name (Dating Sites)	2%
30	Birthday Details (Twitter)	32		30	Gender (Dating Sites)	2%
31	Name (Instagram)	32		31	Cell number (Any Platform)	2%
32	Name (Pinterest)	32		32	University/College details (Any Platform)	2%
33	Platform Subscription (Instagram)	31		33	Surname (Other)	2%
34	Patform Subscription (Other)	31		34	Cell number (Other)	2%
35	University Details (LinkedIn)	31	**19**	35	School details (Any Platform)	1%
36	Physical Address (Facebook)	31	**15**	36	Surname (Dating Sites)	1%
37	Email (Instagram)	31		37	Race (Any Platform)	1%
38	Relationship status (Married)	31	**3**	38	Email (Any Platform)	1%
39	Job details (LinkedIn)	31	**9**	39	Relationship status (Any Platform)	1%
40	Platform Subscription (LinkedIn)	30	**40**	40	Platform Subscription (LinkedIn)	4%
41	Email (Any Platform)	30	**38**	41	Platform Subscription (Mxit)	4%
42	Physical Address (Any Platform)	30	**8**	42	Platform Subscription (Dating Sites)	3%

Chapter 6

Conclusion and Recommendations

This chapter presents a short summary of what has been done in the research study; it gives a conclusion that demonstrates how the study objectives were satisfied and then ends with a discussion of recommendations for further research.

6.1 Research summary

A literature study was conducted on the topics: *identity theft*, *social media*, *big data*, *Hadoop*, the *semantic web* and *data mining*. Data was collected from the SAFPS on historic identity theft cases and data on the social media information-sharing habits of the SU population was collected via surveys sent by email. The SAFPS data was used to determine common attributes among identity theft victims. These attributes were then used as reference variables during the survey data analysis. The survey data respondents were each assigned a vulnerability score that was calculated according to the attributes they have shared on social media platforms and the number of platforms the associated attributes were shared on. Average vulnerability scores were then calculated for each of the SAFPS reference variables. The survey data was graphically explored and transformed into a data set that could be mined with prediction models. The CART, CHAID, Random Forest and Boosted Trees models were applied to the data set to classify individuals as high-risk or low-risk identity theft victims. The Random

Forest model predicted victims best, the trained PMML code of the model was therefore used to score new data. The model was found valid and delivered a list of high ranked predictor variables of identity theft that it was advised to not share on social media platforms.

6.2 Conclusion of the research study

The first objective of the study was to *determine the attributes that have noteworthy correlations with victims of identity theft.* The data collected from the SAFPS on identity theft incidents in South Africa was used to determine these attributes. It was concluded that the attributes `Gender` and `Age` were strongly related to identity theft victims.

The second objective was to *develop a method to estimate vulnerability scores for individuals based on the data they have revealed on social media.* Risk factors were assigned to each of the possible shared attributes and an average vulnerability score was then calculated by the summing of risk factors according to the number of times they were shared for all attributes divided by the total number of attributes. These scores were then used in a preliminary investigation to see if the data set was feasible for the study and it was concluded that the expected data characteristics were visible, which made the data set applicable.

The third objective was to *build a prediction model that best classified identity theft victims.* The following decision trees were explored: CART, CHAID, Random Forest and Boosted Trees. The Random Forest model performed the best in victim prediction. Gains charts were used to present model accuracy and lift charts to show model effectiveness. Both the gains and lift charts suggested that the Random Forest model performed better than the CART, CHAID and Boosted Trees models. With these results in combination with the prediction results it was concluded that the Random Forest model was the prediction model that best classified the dependent variable, `Victim`.

The fourth objective was to *determine the variables that best predicted identity theft victims.* All prediction models presented a list of important variables. The final predictor variables were however determined by the model that best classified the dependent variable. The top predictor variables for the Random Forest model were determined according to a 0–100 scale as calculated by the method described in Breiman *et al.* (1984). It was concluded that 18 of the expected predictor variables that were determined during data exploration turned out to be predictor variables in the Random Forest model.

The fifth objective was to *use the prediction model to score new data as either at high risk of identity theft or at low risk according to their social media information sharing habits.* The Rapid Deployment Tool in Statistica was used to apply the trained PMML code of the Random Forest model to a new deployment data set. The PMML code of the CART, CHAID and Boosted Trees models were applied to the deployment data set to observe how they performed relative to the Random Forest Model. The Boosted Tree model misclassified all non-victims and the CHAID model misclassified all victims. These two models did not therefore present feasible results. The CART model performed the same as the Random Forest model with the classification of victims, but the Random Forest model outperformed CART with the classification of non-victims. It was therefore concluded that the Random Forest model was indeed the best choice for the prediction model. The PMML code for the Random Forest model could therefore be used to score new cases as high-risk or low-risk identity theft victims.

It was hypothesised that *there is a recognisable difference between the amount of data that is shared by people who have been identity theft victims compared to people who have not been offended with the crime.* The average vulnerability score, which was calculated in the preliminary investigation of the survey data, for previous identity theft victims was reasonably higher than the score for non-victims, which meant that the attributes that victims shared on social media were attributes with higher risk factors than that shared by non-victims and the attributes were shared on more social media sites. The hypothesis was therefore accepted as the average vulnerability results on the dependent variable `Victim` satisfied the statement.

Furthermore it was hypothesised that *the attributes commonly found in historic identity theft victim cases are the attributes that will serve as important predictor variables in a model that classified individuals as high-risk or low-risk* *victims of identity theft.* The SAFPS data concluded that the identity theft victims that were reported consisted of more males compared to females and that most of the victims were between the ages of 30 and 40 years. These two variables were therefore set as reference variables as they represented real identity theft incidents. The email survey data concluded that the average vulnerability on social media sites of male users was higher than the score for female users. The average vulnerabilities for different age groups were highest for ages in the range of 30–40 years. These results therefore complied with the SAFPS expected results. The relationship diagrams that were constructed during the survey data mining concluded that the average age for victims was 28 years and the victim percentage for males was again higher than the victim percentage for females. The two most important predictor variables were cosistently found to be `Age` and `Gender` and the Random Forest model then confirmed it as `Age` was ranked as the most important predictor variable and `Gender` as the second most important. The hypothesis was therefore accepted.

Finally it is agreed with Patsakis *et al.* (2014) that the user determines the vulnerability of their own personal information on social media platforms. Security and privacy settings are available, but they do not guarantee protection from cybercrimes. If users avoid the sharing of predictor variables they will decrease their vulnerability on social media, which will decrease their risk of becoming an identity theft victim.

As mentioned before, South Africa is one of the top three countries internationally with the highest rates of fraud using recycled deceased identities (Alfreds, 2015c). It is therefore advised to remove these personal details of loved ones from their online social media profiles if something should happen to them.

6.3 Recommendations

After a comprehensive literature review and reading of the social media privacy policies included in the study, it was found that the initial research methodology was not practical and a revised research methodology was suggested. It is however recommended that for further research big data can be collected, in URI format, on the topic. The initial research methodology would then be applicable. More data would result in more accurate results and if it could be possible to find a way to legally crawl social media sites, the perfect population could be extracted.

References

ALFREDS, D. (2015a). Capitec partners with Home Affairs to beat crime. *News 24*. 11

ALFREDS, D. (2015b). Hackers rob SA of 1 billion data points - IBM. *News 24*. 2, 39

ALFREDS, D. (2015c). Here's how easy it is for crooks to steal your ID. *News 24*. 2, 130, 135

ALKHATEEB, F., MANASRAH, A.M. & BSOUL, A.A.R. (2012). Bank web sites phishing detection and notification system based on semantic web technologies. *International Journal of Security and its Applications*, **6**, 53–66. 26

ALLEMANG, D. & HENDLER, J. (2011). *Semantic Web for the Working Ontologist*. Morgan Kaufmann Publishers, United States, 2nd edn. 25, 26, 27

BAGCHI, T.P. (1999). What are genetic algorithms? In *Multiobjective Scheduling by Genetic Algorithms*, 19–54, Springer. 83

BARBEAU, N. (2015). Widow's home 'stolen' by estate agency. *IOL*. 9

BARBIERATO, E., GRIBAUDO, M. & IACONO, M. (2014). Performance evaluation of NoSQL big-data applications using multi-formalism models. *Future Generation Computer Systems*, **37**, 345–353. 19

BECKER, R.A., VOLINSKY, C. & WILKS, A.R. (2010). Fraud Detection in Telecommunications: History and Lessons Learned. *Technometrics*, **52**, 20–33. 12

Bennett, S. (2014). Minimum Age Requirements: Twitter, Facebook, Instagram, Snapchat, WhatsApp, Secret. *SocialTimes*. 37

Boyd, D. & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, **15**, 662–679. 15

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984). *Classification and Regression Trees (Wadsworth Statistics/Probability)*. Chapman and Hall/CRC, 1st edn. 86, 87, 88, 89, 96, 124, 134

Bryman, A., Bell, E., Hirschsohn, P., dos Santos, A., du Toit, J., Masenge, A., van Aardt, I. & Wagner, C. (2014). *Research Methodology Business and Management Contexts*. Oxford University Press Southern Africa (Pty) Ltd, Cape Town, 1st edn. 36

C-date (2016). Privacy Policy. 155

Cronje, J. (2015). SA link to global internet scams. *IOL*. 9

Davenport, T.H. (2013). Analytics 3.0. *Harvard Business Review*. 15

De'Ath, G. (2007). Boosted trees for ecological modeling and prediction. *Ecology*, **88**, 243–251. 82, 84

Dell (2016). Statistica product index. 56

Dirk, N. (2015a). Home Affairs on course to root out corruption. *IOL*. 10, 11

Dirk, N. (2015b). ID theft costs SA R1bn a year. *IOL*. 7, 10, 11, 44, 51, 120

Electronic Version: StatSoft, I. (2013). Electronic statistics textbook. x, 30, 31, 57, 69, 80, 82, 84, 85, 86, 91, 93, 94, 99, 104, 110, 111, 113

Elitesingles (2016). Privacy Policy. 156

Elith, J., Leathwick, J.R. & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802–813. 94, 95

Elmegreen, B.G. & Sanchez, S.M. (2014). The Future of Computerized Decision Making. In *Winter Simulation Conference*, 943–949. 16

Erasmus, J. (2015). Identity theft in SA booming. *News 24*. 10, 35, 44, 51, 120

Facebook (2016). Data Policy. 145

Gaff, B.M. (2014). Corporate Risks from Social Media. *Computer*, **47**, 13–15. 12, 14

Guazzelli, A. (2015). Predictive Analytics, Big data, Hadoop, PMML. 110

Hamed Haddadi, P.H. (2010). To add or not to add: Privacy and social honeypots. *2010 IEEE International Conference on Communications Workshops, ICC 2010*. 12, 14

Han, J. & Kamber, M. (2001). *Data Mining, Concepts and Techniques*. Morgan Kaufmann Publishers. 31, 32, 33

He, M., Hao, J., Qinhua, W. & Ren, C. (2014). Big data fueled process management of supply risks: sensing, prediction, evaluation and mitigation. In *Winter Simulation Conference*, vol. 1, 1005–1013. 2, 16

Henson, B., Reyns, B.W. & Fisher, B.S. (2013). Fear of crime online? Examining the effect of risk, previous victimization, and exposure on fear of online interpersonal victimization. *Journal of Contemporary Criminal Justice*. 12

Hlophe, N. (2015). Hawks nab five over alleged KZN corpse syndicate. *SABC News*. 9

Ilieva, J., Baron, S. & Healey, N. (2002). Online surveys in marketing research: pros and cons. *International Journal of Market Research.*, **44**, 361 – 376. 37

Instagram (2016). Privacy Policy. 151

IOL (2015). Facebook flaw leaves personal data vulnerable. *IOL*. 11

Kaplowitz, M.D., Hadlock, T.D. & Levine, R. (2004). A comparison of web and mail survey response rates. *Public Opinion Quarterly*, **68**, 94 – 101. 36

Katal, A., Wazid, M. & Goudar, R.H. (2013). Big data: Issues, challenges, tools and good practices. *2013 6th International Conference on Contemporary Computing, IC3 2013*, 404–409. 2, 15, 17, 18, 19, 20, 21

Khan, Z.C. (2013). An Analysis of Facebook ' s Graph Search. 14

Klecka, W.R. (1980). *Discriminant Analysis*. Sage Publications, Inc. 82

Krishnamurthy, B. & Wills, C.E. (2009). On the leakage of personally identifiable information via online social networks. 14

Lee, M., Jung, H. & Cho, M. (2015). On a Hadoop-based analytics service system. *International Journal of Advances in Soft Computing and its Applications*, **7**, 1–8. 15, 20, 21, 24

Lee, T.S., Chiu, C.C., Chou, Y.C. & Lu, C.J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics Data Analysis*, **50**, 1113 – 1130. 108, 109, 126

Lin, F.Y. & McClean, S. (2001). A data mining approach to the prediction of corporate failure. *Knowledge-Based Systems*, **14**, 189–195. 83, 84

LinkedIn (2016). Your Privacy Matters. 147

Lippmann, R. (1987). An introduction to computing with neural nets. *IEEE Assp magazine*, **4**, 4–22. 82

Makhubu, N. (2015). ID theft victim sees red at bank. *IOL*. 9, 11

Maluleke, T. & Pheko, L. (2015). Fake 'letting agents' scam would–be tenants. *IOL*. 7, 9, 11

Martins, S., Pesado, P. & García-Martínez, R. (2016). Information mining projects management process. 28

McAfee, A. & Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Business Review*, 60–68. 16, 17

MIKA, P. (2004). Social Networks and the Semantic Web. *IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)*, 285–291. 25

MKHABELA, N. (2015). Security guards irate over renewals. *IOL*. 11

MKHEZE, S. (2015). Identity theft costs SA R2 billion. *SABC News*. 10

MYBROADBAND (2015). Most popular social platforms in south africa. xiii, 36

MYBROADBAND (2016). Mxit is shutting down. 64

NEWS24 (2015a). Identity Theft. *News 24*. 7

NEWS24 (2015b). Netcare 911 warns of job scam. *News 24*. 9

NGWENYA, J.S. (2015). Identity thieves becoming more resourceful. *IOL*. 10

NISBET, R., MINER, G. & ELDER IV, J. (2009). *Handbook of statistical analysis and data mining applications*. Academic Press. 110

PATSAKIS, C., ZIGOMITROS, A., PAPAGEORGIOU, A. & GALVÓN-LÓPEZ, E. (2014). Distributing privacy policies over multimedia content across multiple online social networks. *Computer Networks*, **75**, 531–543. 14, 135

PINTEREST (2016). Privacy Policy. 152

REDMAN, T. (2013). Data s Credibility Problem. *Harvard Business Review*, 84–88. 18, 19

REYNS, B.W. (2013). Online Routines and Identity Theft Victimization: Further Expanding Routine Activity Theory beyond Direct-Contact Offenses. *Journal of Research in Crime and Delinquency*, **50**, 216–238. 6, 7, 8

REYNS, B.W. & HENSON, B. (2015). The Thief With a Thousand Faces and the Victim With None: Identifying Determinants for Online Identity Theft Victimization With Routine Activity Theory. *International Journal of Offender Therapy and Comparative Criminology*. 2, 8, 10, 12, 14

RIPLEY, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press. 87

Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M. & Rigol-Sanchez, J.P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, **67**, 93–104. 84

Rohanizadeha, S.S. & Moghadam, M.B. (2009). A proposed data mining methodology and its application to industrial procedures. *Journal of Industrial Engineering*, 37 – 50. 27, 29, 31, 32, 33, 81

Rokach, L. & Maimon, O. (2015). *Data Mining with Decision Trees: Theory and Applications*. World Scientific Publishing Co. Pte. Ltd., 2nd edn. 84

Ross, J.W., Beath, C.M. & Quaadgras, A. (2013). You May Not Need Big Data After All. *Harvard Business Review*. 17

SABC (2015). ID theft costs SA firms a fortune. *SABC News*. 9, 10

SAFPS (2016). The Southern African Fraud Prevention Service. 35

SAPA (2015). Submit FICA info, urges Sabric. 11

Saunders, K.M. & Zucker, B. (1999). Counteracting Identity Fraud in the Information Age: The Identity Theft and Assumption Deterrence Act. *Cornell Journal of Law and Public Policy*, **13**, 183–192. 1

Schneider, R.D. (2012). *Hadoop for dummies*. John Wiley & Sons Canada, Ltd., United States, special edn. 15, 22, 23

Shafique, U. & Qaiser, H. (2014a). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, **12**, 217–222. x, 28, 29

Shafique, U. & Qaiser, H. (2014b). A comparative study of data mining process models (kdd, crisp–dm and semma). *International Journal of Innovation and Scientific Research*, **12**, 217 – 222. xiii, 28, 58, 59

Soukup, T. & Davidson, I. (2002). *Visual Data Mining: Techniques and Tools for Data Visualization and Mining*. John Wiley Sons, Inc. 32

SUT, N. & SIMSEK, O. (2011). Comparison of regression tree data mining methods for prediction of mortality in head injury. *Expert Systems with Applications*, **38**, 15534 – 15539. 31, 86, 87

SVETNIK, V., LIAW, A., TONG, C., CULBERSON, J.C., SHERIDAN, R.P. & FEUSTON, B.P. (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences*, **43**, 1947–1958, pMID: 14632445. 104

THOMPSON, J. (2015). Dell Statistica Receives 2015 Technology Innovation Award from Dresner Advisory Services. 56

TOUW, W.G., BAYJANOV, J.R., OVERMARS, L., BACKUS, L., BOEKHORST, J., WELS, M. & VAN HIJUM, S.A.F.T. (2013). Data mining in the life sciences with random forest: a walk in the park or lost in the jungle?. *Briefings in Bioinformatics*, **14**, 315 – 326. 104

TWITTER (2016). Twitter Privacy Policy. 149

UTGOFF, P.E. (1989). Incremental induction of decision trees. *Machine Learning*, **4**, 161–186. 82

WAYNER, P. (2015). 18 Essential Hadoop Tools for Crunching Big Data. 23

WEAVER, J. & TARJAN, P. (2013). Facebook Linked Data via the Graph API. *Semantic Web*, **4**, 245 – 250. 35

WEST, D. (2000). Neural network credit scoring models. *Computers Operations Research*, **27**, 1131 – 1152. 32

WHITE, T. (2009). *Hadoop The Definitive Guide*. O'Reilly Media, Inc., United States, 1st edn. 20, 23

WILEY, J., HAN, V., ALBAUM, G. & THIRKELL, P. (2009). Selecting techniques for use in an internet survey. *Asia Pacific Journal of Marketing and Logistics*, **21**, 455 – 474. 42

YOUTUBE (2016). YouTube Privacy Guidelines. 152

Zoosk (2016). Privacy Policy. 155

Appendix A

Privacy Policy Summaries of Social Media Platforms

The privacy policies included in this study were extracted on 31 July 2016. It should therefore be noted that changes could have been made to these policies at the time of reading.

A.1 Facebook privacy policy

Table A.1 displays a short summary of the Facebook privacy policy that was obtained from the official online site. The complete privacy policy can be extracted from Facebook (2016).

A.1 Facebook privacy policy

Table A.1: Summary of Facebook Privacy Policy.

Facebook	
Kinds of information collected:	Things users do and information users provide. Things other users do and information they provide. Users' networks and connections. Information about payments. Device information. Information from websites and apps that use Facebook's services. Information from third-party partners.
How information is used:	To provide, improve and develop services. To provide shortcuts and suggestions to users. To communicate with users. To show and measure advertisements and services. To promote safety and security by making use of cookies and similar technologies.
How information is shared:	On Facebook services. To people users share and communicate with. Public information is available to anyone on or off Facebook services and can be seen or accessed through online search engines. It is shared with apps, websites and third-party integrations on or using Facebook services. Sharing happens within Facebook companies. New owners. Sharing with third-party partners and Customers.
Manage and delete information:	Users can manage their accounts through the activity log tool. Data is stored for as long as it is necessary to provide products and services to users and others. Information associated with users' accounts will be kept until their account is deleted.
How harm is prevented:	Facebook may access, preserve and share users' information in response to a legal request, if Facebook has good belief that the law requires them to do so. Facebook may access, preserve and share users' information when they believe it is necessary to: detect, prevent and address fraud and other illegal activity, to protect users, others and themselves (this includes investigations).
Operation of global services:	Complies with US-ES, US-Swiss and participates in the Safe Harbor program as set by the Department of Commerce (Resolves disputes through TRUSTe). Facebook may share information for purposes described in this policy.
Privacy policy changes:	Users will be notified if privacy policy changes and given opportunity to review and comment before continuing use of services.

A.2 LinkedIn privacy policy

Table A.2 and Table A.3 present a short summary of the LinkedIn privacy policy that was obtained from the official online site. The complete privacy policy can be extracted from LinkedIn (2016).

Table A.2: Summary of LinkedIn Privacy Policy (a).

LinkedIn	
Kinds of information collected:	Information users provide with registration. Profile information users fill out. Address book and other services that synchronize with LinkedIn. When users contact for customer support. When using LinkedIn sites, applications and advertisements. Using third-party services and visiting third-party sites. Cookies. Advertising technologies and web beacons. Log files, IP addresses and information about an user's computer and mobile device.
How information is used:	Users agree that information they provide on their profile can be seen and used by LinkedIn. LinkedIn communications (messages and emails). User communications (a recipient can see a user's name, email address and some network information). To conduct research and development and to customise users' experience. Third parties using LinkedIn platform services. Polls and surveys. Searches. Groups. Testimonials and advertisements placed through LinkedIn advertisements. Talent recruiting, marketing and sales solutions. Pages for companies, schools, influencers and other entities. Compliance with legal process and other disclosures. Disclosure to others as the result of a change in control or sale of LinkedIn Corporation. Service providers if LinkedIn needs assistance. Data processing outside the company users live.
How information is shared:	Information may be shared with user's consent or as required by law. Shared across LinkedIn's different services, among companies in the LinkedIn family. Information is shared with LinkedIn Affiliates. Information is shared with third-parties (a users' profile and any data they post can be found by others through search engines, if public).
Manage and delete information:	Users have rights to access, correct or delete information, and closing their account. If an account is deleted, LinkedIn will remove all the user's information within 24 hours. Data retention – LinkedIn keeps users' information for as long as their accounts is active or needed. Some information is kept even after an account is deleted for example when it is necessary to comply with LinkedIn obligations , to meet regulatory requirements, resolve disputes, prevent fraud and abuse, or enforce this agreement.

Table A.3: Summary of LinkedIn Privacy Policy (b).

LinkedIn	
How harm is prevented:	Personal information is protected by using industry-standard safeguards. LinkedIn takes privacy and security seriously and have enabled HTTPS access to their site, in addition to existing SSL access over mobile devices. The internet is not a secure environment – select strong passwords. LinkedIn do not permit the use of any third-party software. No ‘crawlers’, bots, browser plug-ins, or browser extensions are permitted to scrape or copy data.
Operation of global services:	Partner with TRUSTe and complies with US-EU and US-Swiss Safe Harbor programs. Californias Shine the Light Law. LinkedIn does not share any of users’ personal information with third parties for direct marketing.
Privacy policy changes:	LinkedIn only notifies users when they make changes to this privacy policy if it affects users’ personal information, else not.

A.3 Twitter privacy policy

Table A.4 displays a short summary of the Twitter privacy policy that was obtained from the official online site. The complete privacy policy can be extracted from Twitter (2016).

Table A.4: Summary of Twitter Privacy Policy.

Twitter	
Kinds of information collected:	Basic account information. Contact information. Additional information users provide. Tweets. Following, lists and other public information. Location information. Interaction with links. Cookies. Log data – such as IP addresses, browser types, operating systems, the referring web page, pages visited, location, users' mobile carriers, device information, search terms and cookie information. Widget data. Commerce services – to facilitate future purchases Twitter saves user payment information (excluding CVV code) and shipping address. Third parties and affiliates. Twitter receives users' information through their various websites, SMS, API's, email notifications, applications, buttons, widgets, advertisements, commerce services and other covered services, and from partners or other third parties.
How information is used:	Twitter engages with service providers to perform functions and provide services to users. Commerce transactions – to help with future purchases. As required by law. Users' information may be sold or transferred as part of a transaction.
How information is shared:	Users' information may be shared at their directing such as when they authorise a third-party web client or application to access their account. Service providers. Payment information is shared with the provider, commerce provider, marketplace, or charity. Government request – information is shared if it complies with the law. Business transfers, third parties and affiliates.
Manage and delete information:	Users can manage their data with the account settings. What users say on the Twitter Services may be viewed all around the world instantly. 'You are what you tweet'. Search engines and other third parties may still retain copies of users' public information even after users have deleted the information or deactivated their account.
How harm is prevented:	NA
Operation of global services:	NA
Privacy policy changes:	Users will be notified about changes via an @Twitter update or via email.

A.4 Instagram privacy policy

Table A.5 displays a short summary of the Instagram privacy policy that was obtained from the official online site. The complete privacy policy can be extracted from Instagram (2016).

Table A.5: Summary of Instagram Privacy Policy.

Instagram	
Kinds of information collected:	Information users provide Instagram directly – profile and contact details for example. Finding friends – contact details or third parties connection. Analytics information – example web pages users visited, add-ons, and other information that assists Instagram services. Cookies and similar technologies. Log file information. Device identifiers. Metadata – technical data that is associated with user content.
How information is used:	To help users efficiently access their information after they sign in. Remember information so that users do not have to re-enter it every time. Provide personalised content and information to users and others, can include online advertisements or other forms of marketing. Provide, improve, test and monitor the effectiveness of Instagram's services. Develop and test new products and features. Monitor metrics such as total number of visitors. Diagnose or fix technology problems. Automatically update the Instagram application.
How information is shared:	Information will not be rented or sold to third-parties outside of Instagram without users' consent. Information may be shared with affiliates. Third-party organisations within the Instagram family. Service providers. Third party advertisers. Users can choose to share information with users that use the API service. Instagram may transfer users' information in the event of a change of control. Responding to legal requests and preventing harm.
Manage and delete information:	After termination or deactivation of an account, Instagram and affiliates may keep users' information and user information for a commercially reasonable time.
How harm is prevented:	Commercially reasonable safeguards are used to keep information secure.
Operation of global services:	NA
Privacy policy changes:	It is the users own responsibility to review the privacy policy periodically.

A.5 YouTube privacy policy

Table A.6 displays a short summary of the YouTube privacy policy that was obtained from the official online site. The complete privacy policy can be extracted from YouTube (2016).

Table A.6: Summary of YouTube Privacy Policy.

YouTube	
Kinds of information collected:	Any information that is uploaded with videos.
How information is used:	Videos are viewed by users all over the globe.
How information is shared:	Privacy violation complaints are not accepted from third-parties.
Manage and delete information:	If a privacy complaint is filed, YouTube provides the uploader with an opportunity to remove the video within 48 hours, if the uploader does not respond – YouTube will take action.
How harm is prevented:	Content like videos are removed when an individual can be uniquely identifiable by image, voice, full name, national insurance number, bank account number, or contact information.
Operation of global services:	NA
Privacy policy changes:	NA

A.6 Pinterest privacy policy

Table A.7 displays a short summary of the Pinterest privacy policy that was obtained from the official online site. The complete privacy policy can be extracted from Pinterest (2016).

A.6 Pinterest privacy policy

Table A.7: Summary of Pinterest Privacy Policy.

Pinterest	
Kinds of information collected:	Information the user provides or give permission to obtain (example – name, surname, email, cell phone, location, *etc.*). Whenever the user uses websites, mobile applications or other internet services through Pinterest. Log data. Cookie data. Device information. Partners and advertisers may share information with Pinterest.
How information is used:	Pinterest uses information collected to provide better products to users, develop new products, and protect Pinterest and its users. To offer users customised content, including: suggesting pins or boards they might like and showing them advertisements they may be interested in. Send users updates, newsletters, marketing materials and other information that may be of interest to them. Help users' friends and contacts find them on Pinterest. Respond to users' questions or comments. The information Pinterest collect may be personally identifiable or non-personally identifiable.
How information is shared:	Anyone can see the public boards and pins users create, and the profile information they give Pinterest. Pinterest may make public information available to APIs. With the user's consent Pinterest may share information with other services such as Facebook. When users buy something on Pinterest using their credit card, Pinterest may share their credit card information, contact information, and other information about the transaction with the merchant they were buying from. (The merchants treat this information just as if the user had made a purchase from their website directly, which means their privacy policies and marketing policies apply to the data users share with them.) If users buy something on Pinterest via Apple Pay, their credit card number is not shared to merchants, but contact and transaction info is still. Pinterest allows third-party companies to audit the delivery and performance of advertisements on Pinterest. Pinterest may employ third-party companies or individuals to process personal information on their behalf based on their instructions and in compliance with this policy. If Pinterest believes that disclosure is reasonable necessary to comply with law, regulation or legal request, to protect the safety, rights, or property of the public, any person, or Pinterest, or to detect, prevent or otherwise address fraud, security or technical issues. Pinterest may engage in a merger, acquisition, bankruptcy, dissolution, reorganisation, or similar transaction or proceeding that involves the transfer of the information described in this policy. Pinterest may share aggregated or non-personally identifiable information with partners, advertisers or others.
Manage and delete information:	A user may access and change information on their profile page at any time, choose whether their profile page is available to search engines, or choose whether others can find their Pinterest account using their email address. Link or unlink their Pinterest account from an account on another service (like Facebook). Create or be added to a secret board that is only visible to them and other participants on the board. Choose whether Pinterest will be customised for them using information from off-Pinterest websites or apps. Choose whether their purchase on Pinterest will be used to customise recommendations and ads for them. Pinterest supports the do not track browser setting. Users may close their account at any time. When users close their account, Pinterest deactivates it and removes their pins and boards from Pinterest. The browser users use may provide them with the ability to control cookies or other types of local data storage. Users' mobile device may provide them with choices around how and whether location or other data is shared with Pinterest.
How harm is prevented:	NA
Operation of global services:	NA
Privacy policy changes:	If the privacy policy is changed and the user continues to use Pinterest after those changes are in effect, they agree to the revised policy. If the changes are significant, Pinterest may provide more prominent notice or get users' consent as required by law.

A.7 Mxit privacy policy

Table A.8 displays a short summary of the Mxit privacy policy that was obtained from the official online site. The complete privacy policy is no longer available as Mxit officially closed the service on 30 September 2016.

Table A.8: Summary of Mxit Privacy Policy.

Mxit	
Kinds of information collected:	The user can decide how much information they give Mxit, but users must provide Mxit at least the minimum information as required by the registration process. Mxit collects information directly from users. Mxit receives different types of information about users. Included in this is a user's personal information that is needed to create an user account: it includes the following – name, mobile number, ID, username and password. Other information such as date of birth, gender, interests, the device the user uses, their location and other information about them can be added to a user's account.
How information is used:	It is used to deliver services. To maintain systems. To improve service offerings to users. To improve user experience on Mxit. To make it possible for users to use third-party applications on the platform. To put together statistics about the use of the services.
How information is shared:	With developers and service providers. Third-party applications. Service providers such as advisors or information technology providers. To generate income – share info with developers, advertisers, and service providers. To process personal information on servers in other countries. When required by law to hand over information to the authorities.
Manage and delete information:	NA
How harm is prevented:	Users must continuously review the information security practices to make sure no one can gain unauthorised access to their personal information. The personal info users share with contacts, other Mxit users, in chat rooms or with developers whose applications users use is not covered by this privacy policy.
Operation of global services:	NA
Privacy policy changes:	Changes are announced 14 days before applied, if the user does not agree they may terminate their account.

A.8 Dating Sites privacy policy

In the study it is referred to 'Dating Sites' as a whole. It was therefore decided to look at three randomly selected dating sites that were popular at the time of research for some insight.

Table A.9 displays a short summary of the dating site Zoosk's privacy policy that was obtained from the official online site. The complete privacy policy can be extracted from Zoosk (2016).

Table A.9: Summary of Zoosk Privacy Policy.

Zoosk	
Kinds of information collected:	Information users provide – (like name, birth date, photographs email address, password, phone number, including mobile phone number, billing information, credit card information and other contact or demographic information provided, also extra info such as personal interests, background, gender, age, geographical location and physical characteristics). Information automatically collected – (like IP address, browser type, internet service provider, platform type, the site from which the user came and the site to which they are going when they leave Zoosk, date and time stamp and one or more cookies that may uniquely identify the user's browser or their account, UDID, mobile carrier, device manufacturer and phone number). Cookies and other technologies. Third party advertisers. Social networking sites services through a social networking site such as Facebook.
How information is used:	To operate, provide, maintain, develop, deliver, protect, and improve services, products, applications, content and advertising. To develop new products and services. To maintain and administer users' Zoosk account. To maintain and display users' Zoosk profile. To respond to users' comments and questions and provide customer service. To detect and prevent abusive, fraudulent, malicious, or potentially illegal activities, and to protect the rights, safety or property of Zoosk users. To enforce Zoosk's terms of use agreement. To keep users posted on Zoosk's latest product or service announcements, software updates, security alerts and any technical notices. To communicate with users. To provide users with confirmation of purchases, invoices, support and administrative messages or notice about changes to Zoosk's terms, conditions or policies. To administer any contests or promotions. To perform other functions as otherwise described to users at the time of collection. To link or combine information about users with other information Zoosk gets from third-parties to help understand users' needs and provide them with better service. Zoosk may store and process personal information in the United States and other countries.
How information is shared:	Users are allowed to share information about them with other individuals and other companies. Personal info is shared to third parties with users' consent, To satisfy applicable law, regulation, legal process or government request. For the investigation of potential violation of the terms of use agreement. To defend Zoosk against third party claims or allegations. For protection against harm. To detect, prevent or otherwise address fraud, security or technical issues. In connection with any merger, sale of company assets, reorganisation, financing, change of control or acquisition or in the event of bankruptcy. With third party vendors, consultants, and other service providers that perform services on Zoosk's behalf. Zoosk may also share information with others in an aggregated form that does not directly identify a user.
Manage and delete information:	Users can manage communications from Zoosk with settings. Remove or block cookies with settings. Opt out of advertising. Users can access or change their account with settings.
How harm is prevented:	Zoosk takes reasonable measures to help protect users' personal information in an effort to prevent loss, misuse, and unauthorised access, disclosure, alteration and destruction.
Operation of global services:	NA
Privacy policy changes:	Zoosk reserves the right to modify this policy from time to time. If Zoosk makes any changes to the policy, they will change the "last revision' date below and will post the updated policy on the policy page.

Table A.10 displays a short summary of the dating site C-date's privacy policy that was obtained from the official online site. The complete privacy policy can be extracted from C-date (2016).

Table A.10: Summary of C-date Privacy Policy.

C-date	
Kinds of information collected:	Personal data such as age, gender, zip code, telephone number, email address, evaluation results of questionnaires, personal statements and pictures posted. Cookies.
How information is used:	To provide services, contractors and other third parties. To a buyer or merger in the event of selling. For the purpose of advertising. Market research. For research and analysis. To subsidiaries and affiliates. To fulfil the purpose for which users provide it. With users' consent.
How information is shared:	C-date may disclose aggregated information about users and information that does not identify any individual, without restriction. Third-parties, affiliates or service providers. If the company should be sold, personal data will be transformed to third-party. To comply with any court order. If C-date believes disclosure is necessary to protect rights.
Manage and delete information:	Users can change account settings to not receive cookies.
How harm is prevented:	Interdate does not sell, trade, or otherwise disclose customer lists or personal data to unaffiliated third-parties without permission. C-date has implemented measures designed to secure users' personal information. Any payment transactions are encrypted using SSL technology.
Operation of global services:	Childrens Online Privacy Protection Act
Privacy policy changes:	Users will be notified by the sending of a notice to their mailboxes if material changes to how personal information is treated are made.

Table A.11 displays a short summary of the dating site Elitesingles' privacy policy that was obtained from the official online site. The complete privacy policy can be extracted from Elitesingles (2016).

A.8 Dating Sites privacy policy

Table A.11: Summary of Elitesingles Privacy Policy.

Elitesingles	
Kinds of information collected:	Info users provide with registration (gender, email, postal code, date of birth, marital status, education, occupation, income). For paid membership users must provide their name and surname, bank account details or credit card information and home address. Further info collected during site visits are: IP address, page name, date and time of access, referred URL, session cookies.
How information is used:	Elitesingles may record services users are interested in as well as user traffic patterns. For statistical purposes. Users profile info is used to suggest other members for match making. Other members can see when a user has viewed their profile and vice versa. Other members can connect with users. Email notifications and to receive information like specials and partners offers. Import info from Facebook. In order to find suitable partners.
How information is shared:	Other members can see users' profile information. Third-parties and service providers. Use of cookies. Use of analysis programs and remarketing. Use of google doubleclick. Use of Facebook, Google+, Twitter and social plugins. Checking and verification of personal information. When required by law.
Manage and delete information:	Users can change their message and profile settings. Right to any time receive information about personal information and settings Elitesingles store to a user's account. Users' right to erase personal information may be limited by legal retention.
How harm is prevented:	Written permission is obtained for collection, collation, processing or disclosure of information. Personal info is not electronically collected. Reason for personal data collection will be given in writing. Record will be kept of personal data during use and at least one year after, dates and third parties. Obsolete personal data will be destroyed. Elitesingles uses technological, organizational and physical protection measures. All info is encrypted with Secure Socket Layer. Users can decide on settings whether they want to continue to have their profile suggested to other members.
Operation of global services:	Consumer Protection Act 68 of 2008 Protection of Personal Information Act 4 of 2013 Electronic Communications and Transactions Act 25 of 2002
Privacy policy changes:	NA

Appendix B

CRISP: Data Understanding – Variable Description Graphs

B.1 Histograms of shared attributes

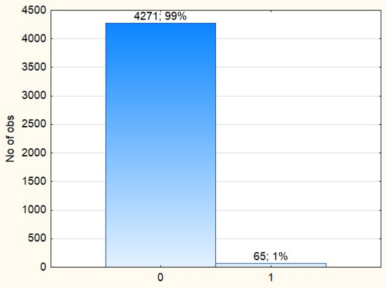


Figure B.1: `Name` shared on – (None).

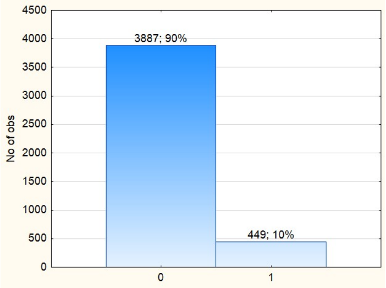


Figure B.2: `Name` shared on – (Other).

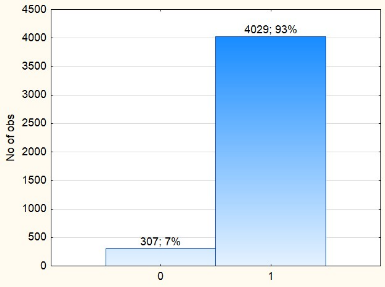


Figure B.3: `Name` shared on – (Facebook).

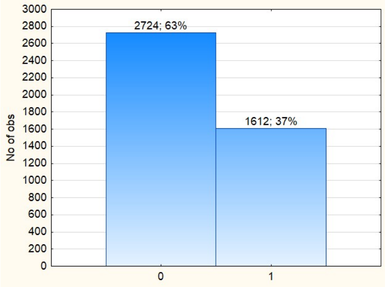


Figure B.4: `Name` shared on – (LinkedIn).

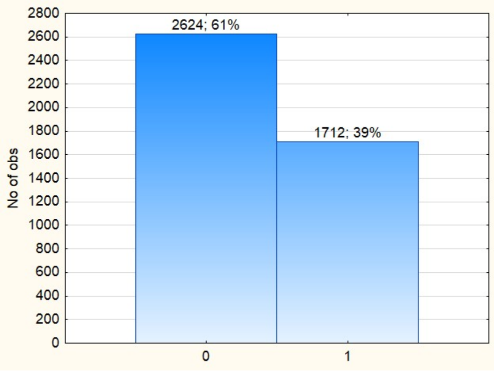


Figure B.5: `Name` shared on – (Twitter).

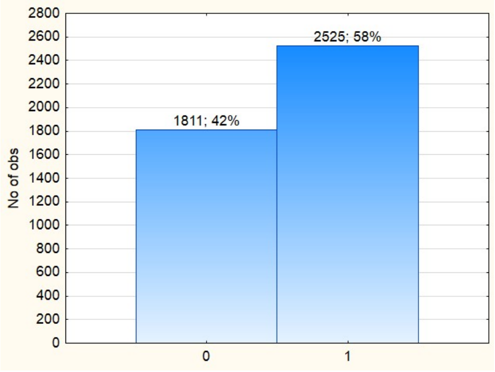


Figure B.6: `Name` shared on – (Instagram).

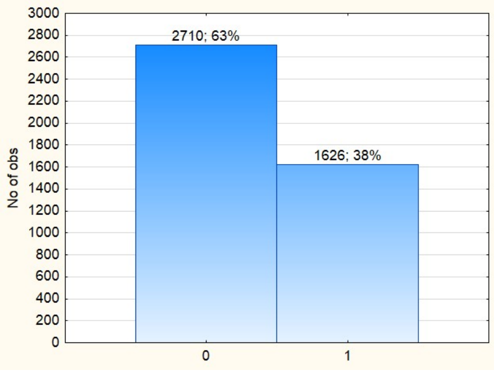


Figure B.7: `Name` shared on – (YouTube).

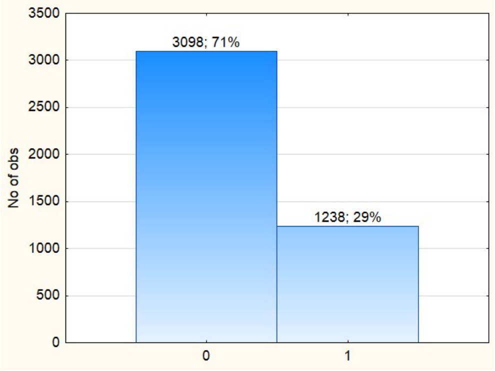


Figure B.8: `Name` shared on – (Pinterest).

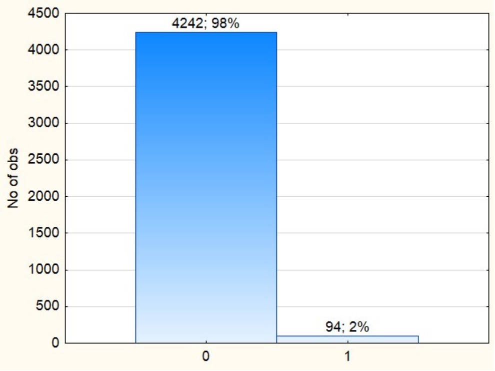


Figure B.9: `Name` shared on – (Mxit).

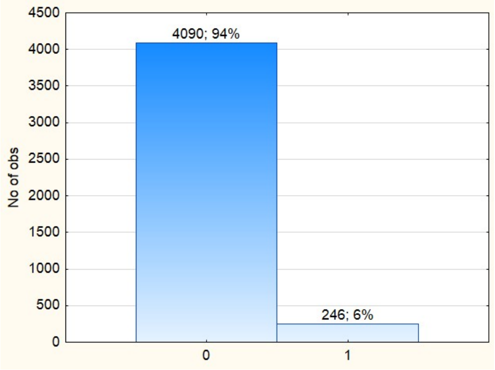


Figure B.10: `Name` shared on – (Dating Sites).

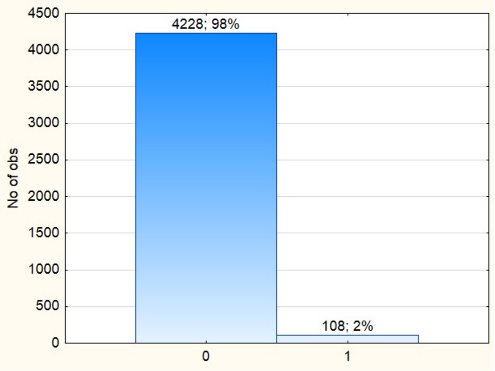


Figure B.11: Surname shared on – (None).

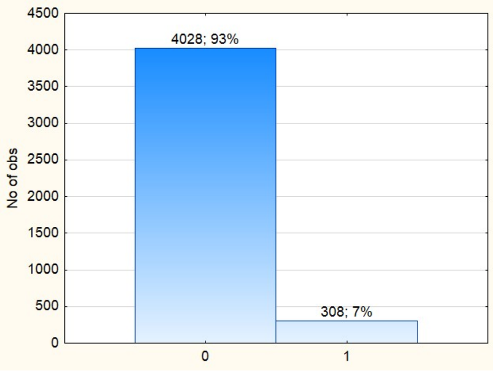


Figure B.12: Surname shared on – (Other).

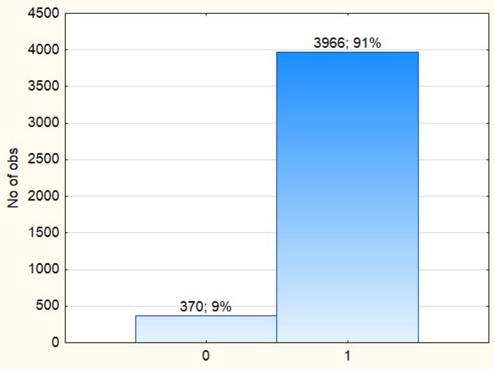


Figure B.13: Surname shared on – (Facebook).

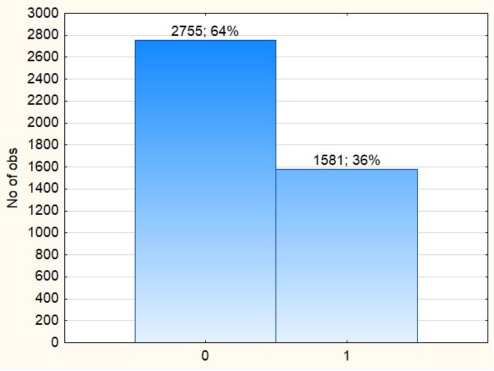


Figure B.14: Surname shared on – (LinkedIn).

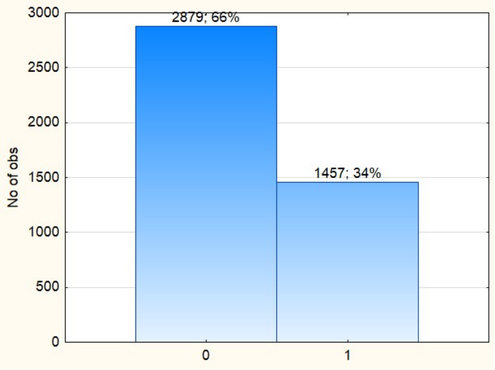


Figure B.15: Surname shared on – (Twitter).

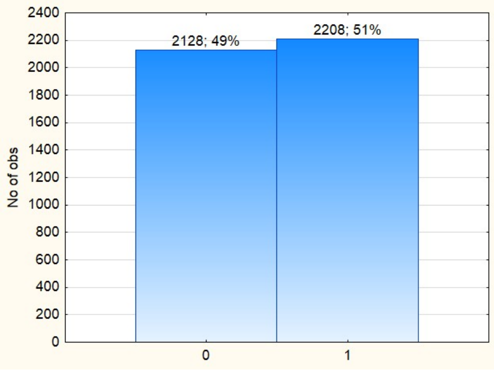


Figure B.16: Surname shared on – (Instagram).

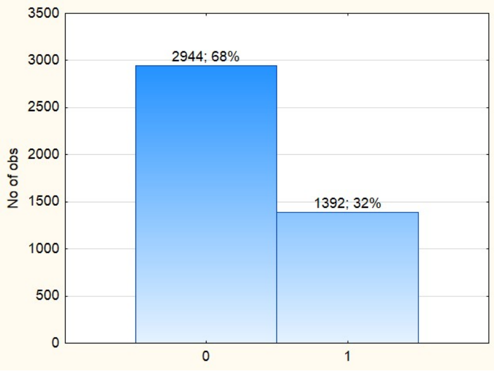


Figure B.17: `Surname` shared on – (YouTube).

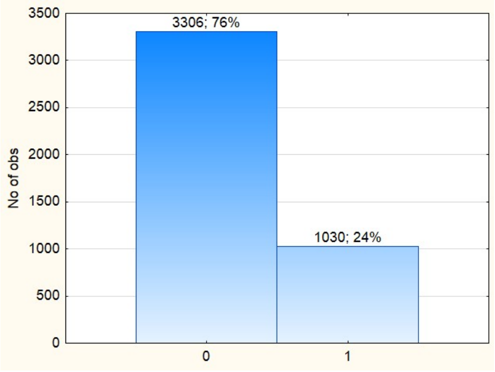


Figure B.18: `Surname` shared on – (Pinterest).

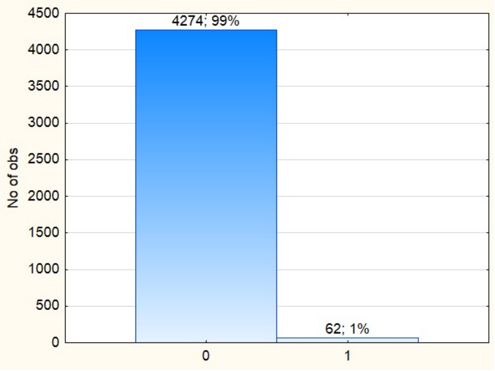


Figure B.19: `Surname` shared on – (Mxit).

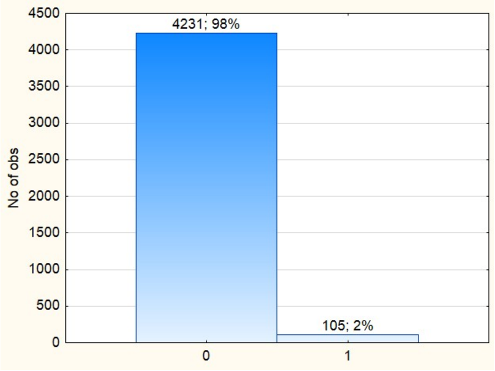


Figure B.20: `Surname` shared on – (Dating Sites).

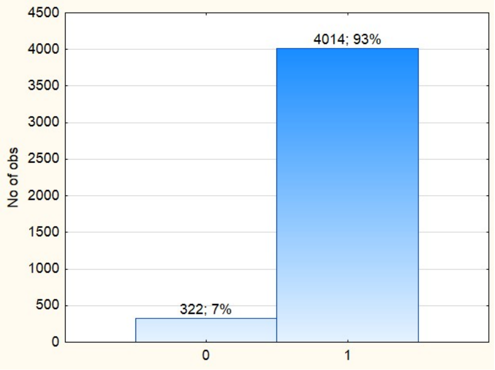


Figure B.21: `Identity Number` shared on – (None).

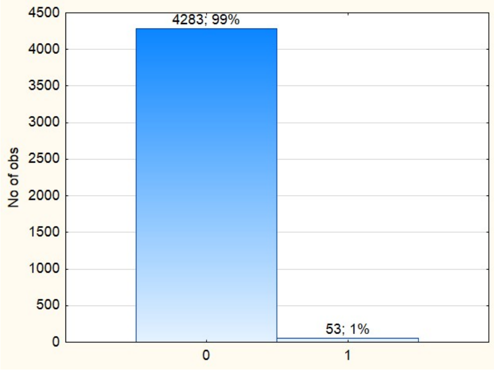


Figure B.22: `Identity Number` shared on – (Other).

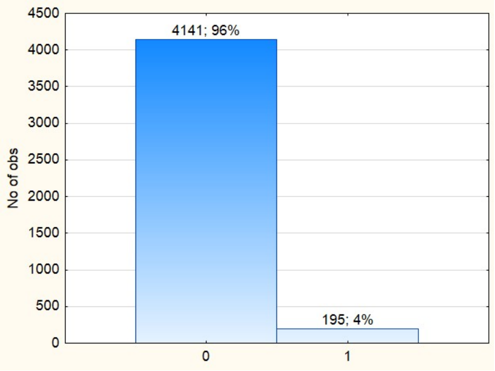


Figure B.23: `Identity Number` shared on – (Facebook).

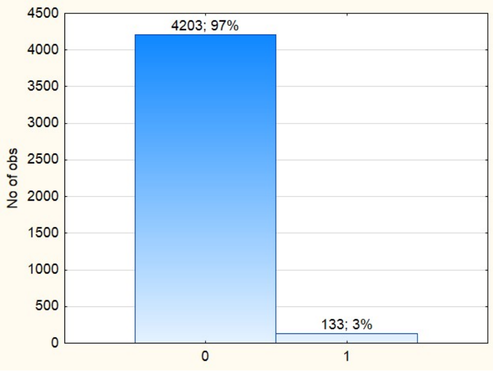


Figure B.24: `Identity Number` shared on – (LinkedIn).

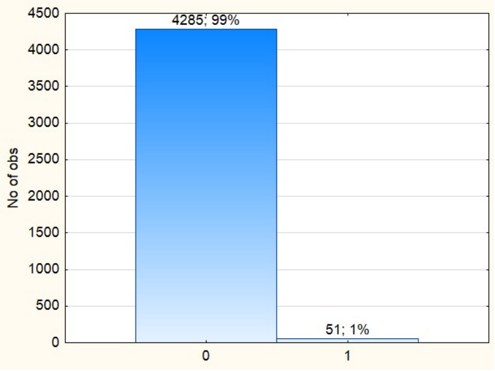


Figure B.25: `Identity Number` shared on – (Twitter).

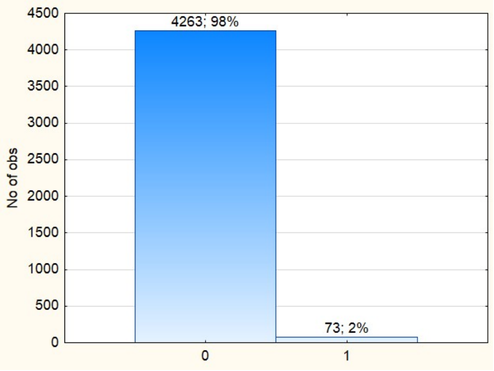


Figure B.26: `Identity Number` shared on – (Instagram).

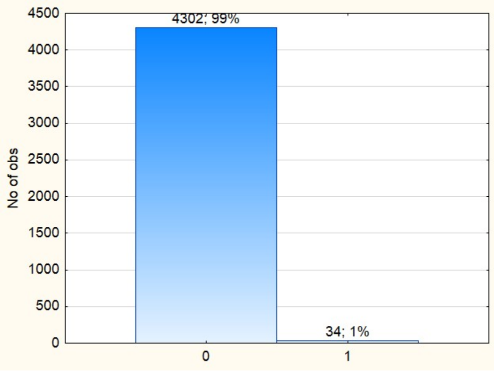


Figure B.27: `Identity Number` shared on – (YouTube).

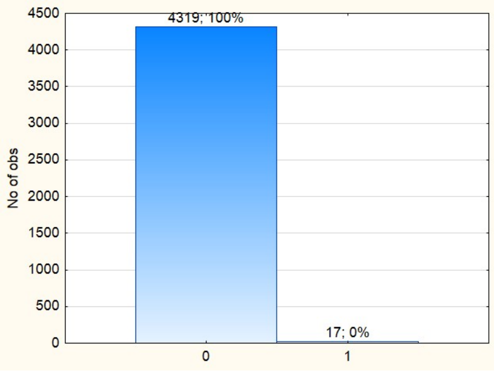


Figure B.28: `Identity Number` shared on – (Pinterest).

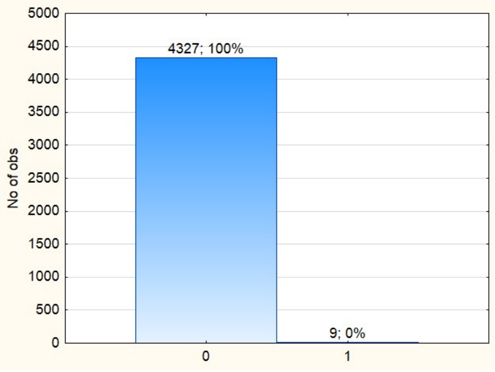


Figure B.29: `Identity Number` shared on – (Mxit).

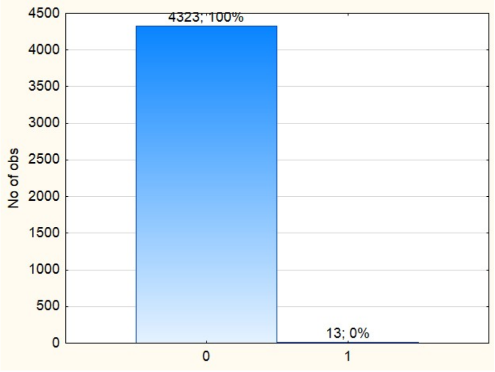


Figure B.30: `Identity Number` shared on – (Dating Sites).

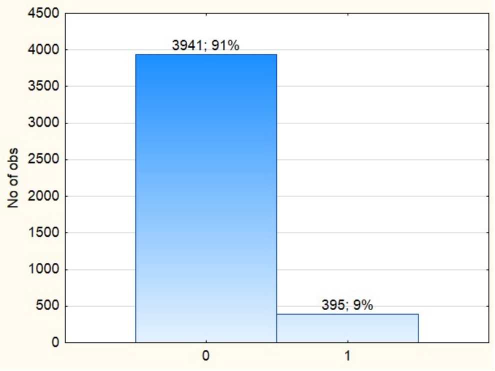


Figure B.31: `Birthday` shared on – (None).

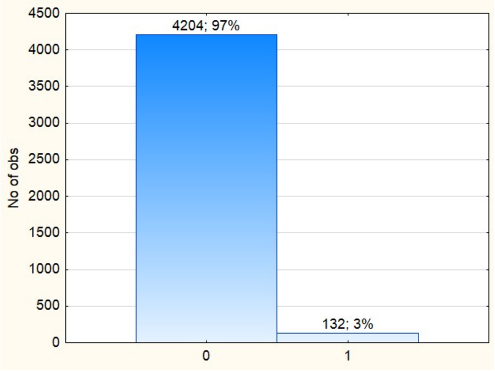


Figure B.32: `Birthday` shared on – (Other).

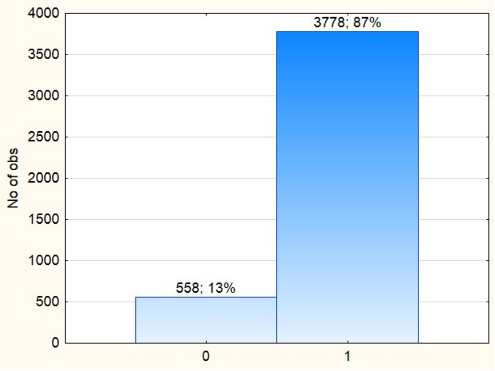


Figure B.33: `Birthday` shared on – (Facebook).

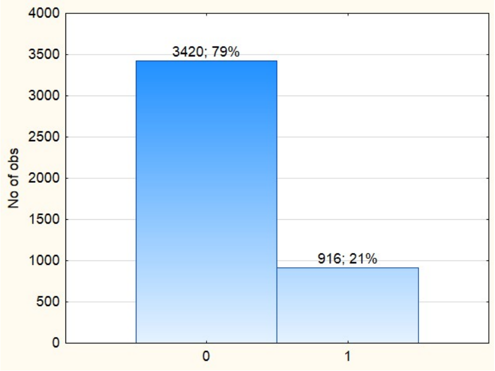


Figure B.34: `Birthday` shared on – (LinkedIn).

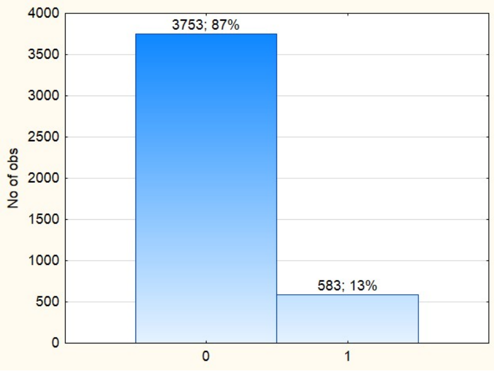


Figure B.35: Birthday shared on – (Twitter).

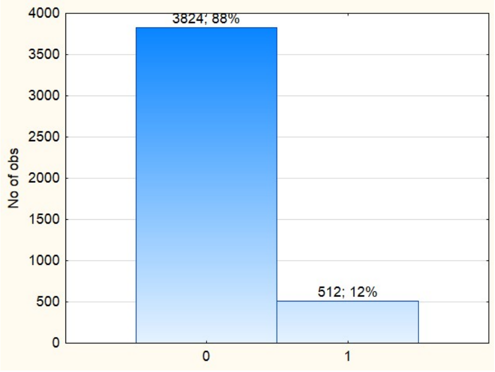


Figure B.36: Birthday shared on – (Instagram).

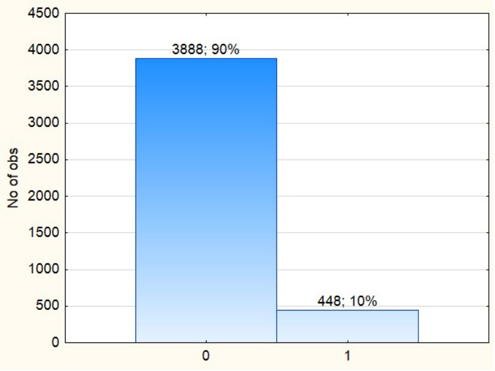


Figure B.37: Birthday shared on – (YouTube).

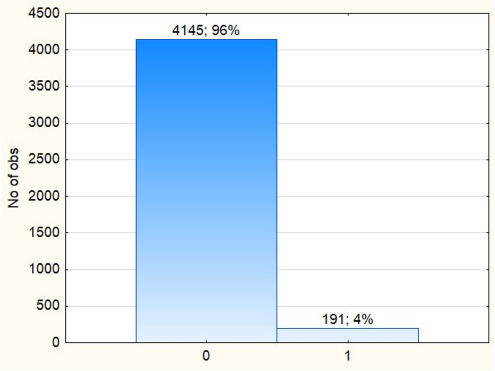


Figure B.38: Birthday shared on – (Pinterest).

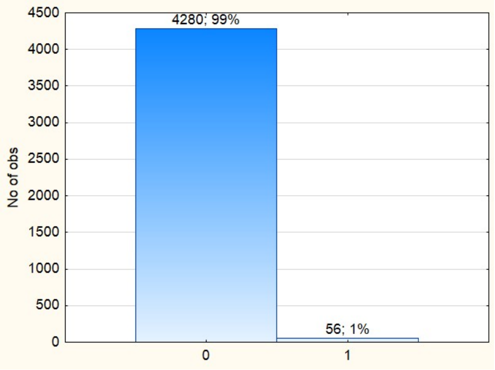


Figure B.39: Birthday shared on – (Mxit).

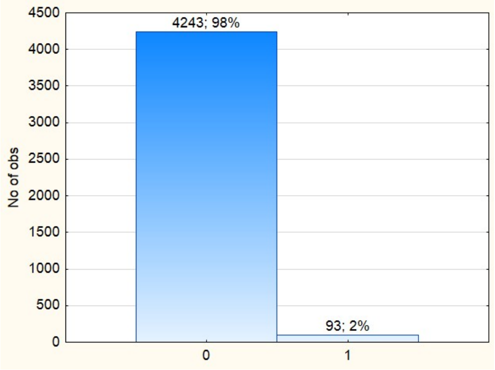


Figure B.40: Birthday shared on – (Dating Sites).

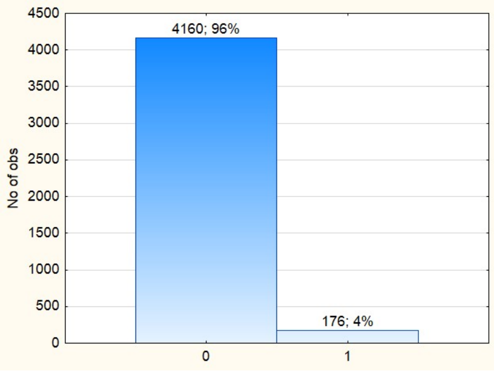


Figure B.41: Gender shared on – (None).

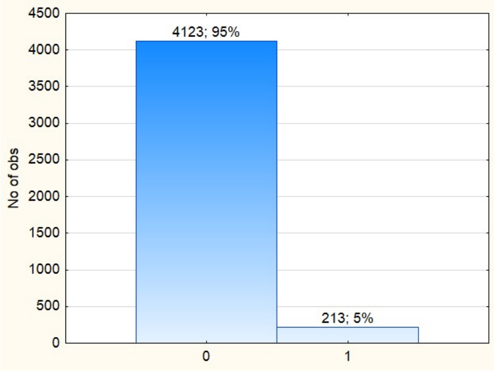


Figure B.42: Gender shared on – (Other).

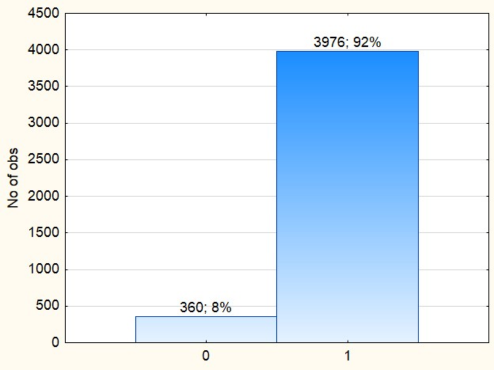


Figure B.43: Gender shared on – (Facebook).

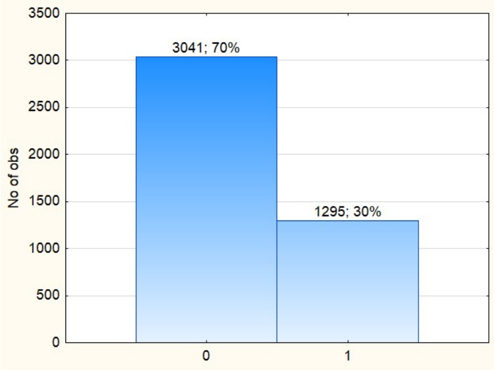


Figure B.44: Gender shared on – (LinkedIn).

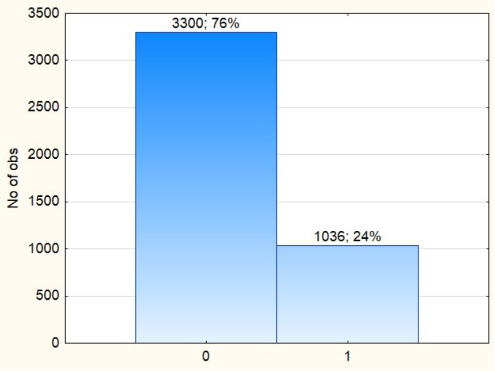


Figure B.45: Gender shared on – (Twitter).

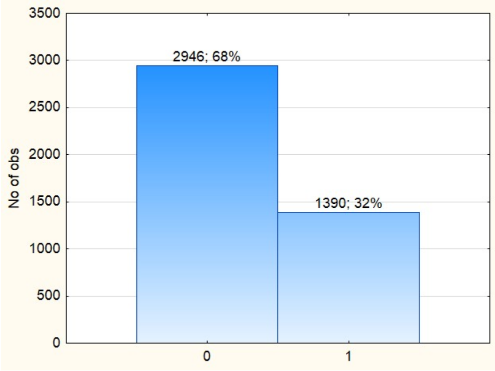


Figure B.46: Gender shared on – (Instagram).

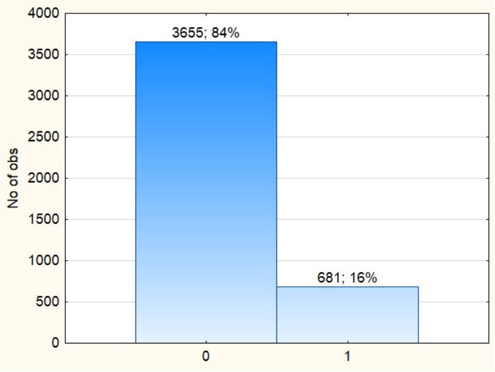


Figure B.47: `Gender` shared on – (YouTube).

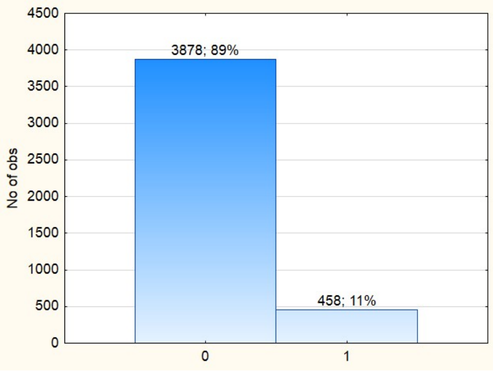


Figure B.48: `Gender` shared on – (Pinterest).

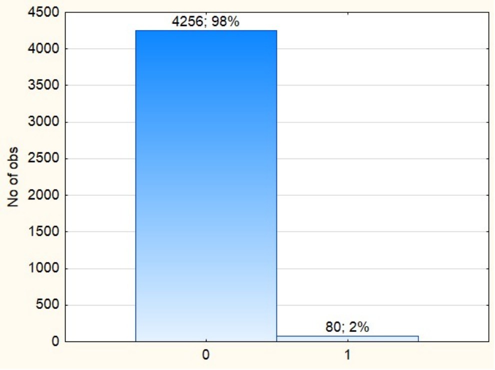


Figure B.49: `Gender` shared on – (Mxit).

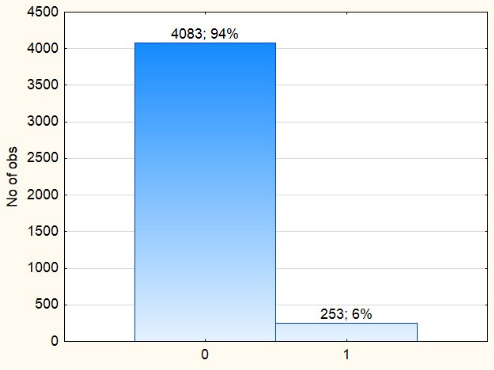


Figure B.50: `Gender` shared on – (Dating Sites).

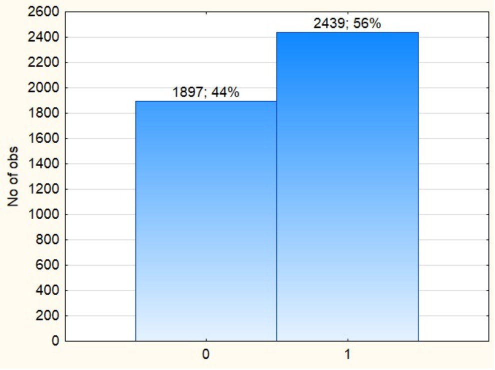


Figure B.51: `Race` shared on – (None).

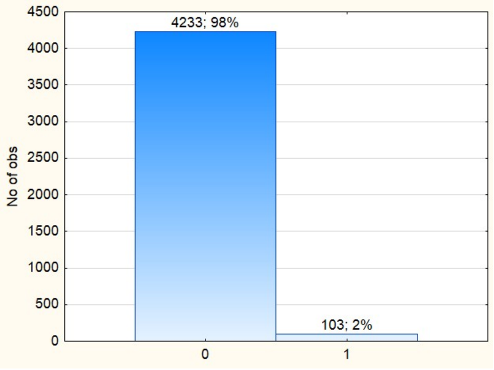


Figure B.52: `Race` shared on – (Other).

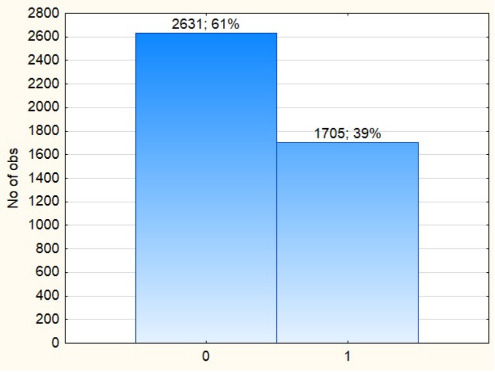


Figure B.53: `Race` shared on – (Facebook).

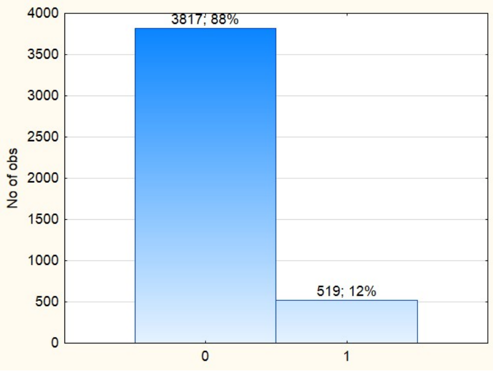


Figure B.54: `Race` shared on – (LinkedIn).

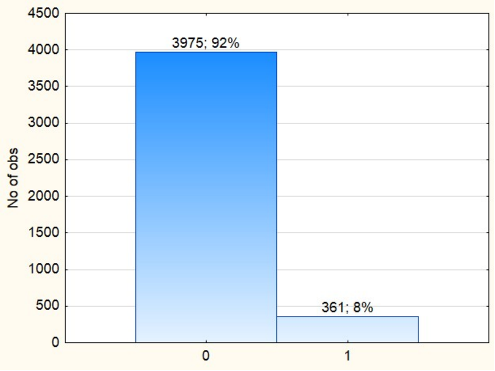


Figure B.55: `Race` shared on – (Twitter).

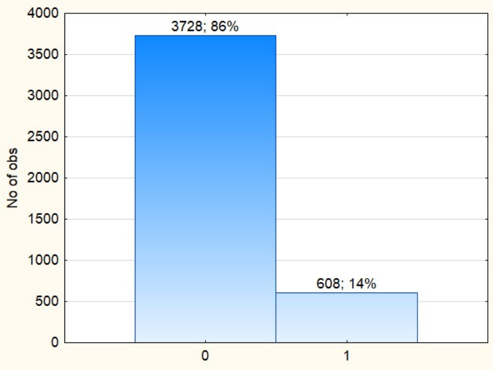


Figure B.56: `Race` shared on – (Instagram).

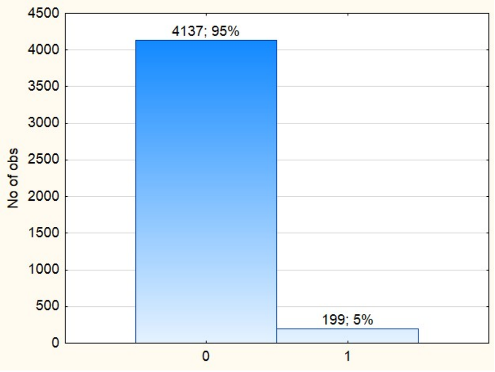


Figure B.57: `Race` shared on – (YouTube).

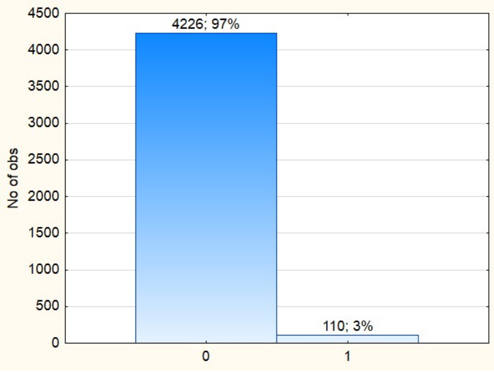


Figure B.58: `Race` shared on – (Pinterest).

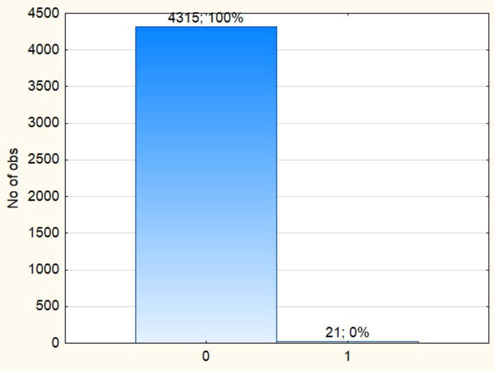


Figure B.59: Race shared on – (Mxit).

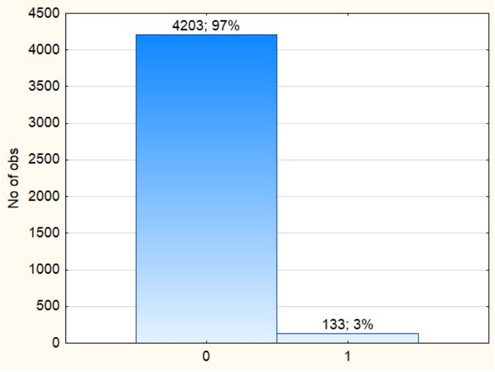


Figure B.60: Race shared on – (Dating Sites).

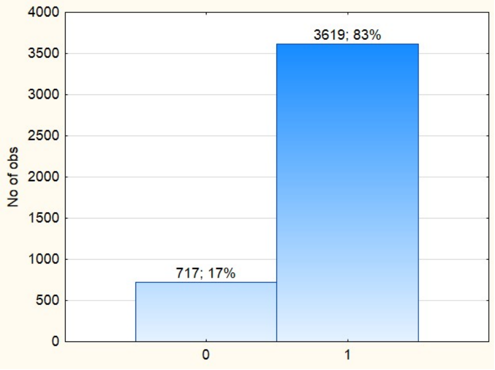


Figure B.61: Physical Address shared on – (None).

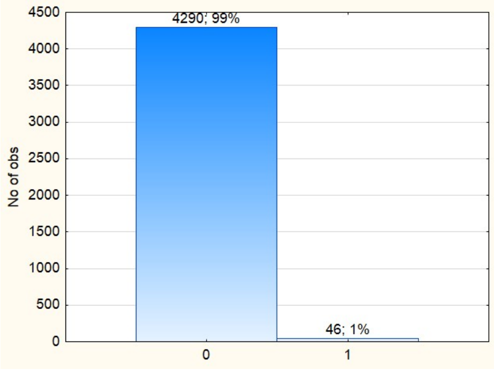


Figure B.62: Physical Address shared on – (Other).

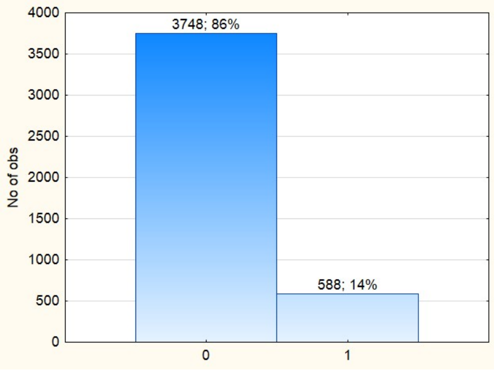


Figure B.63: Physical Address shared on – (Facebook).

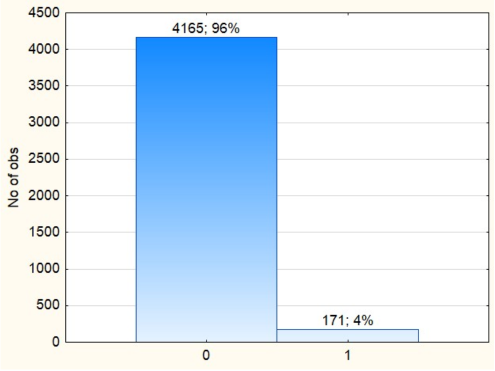


Figure B.64: Physical Address shared on – (LinkedIn).

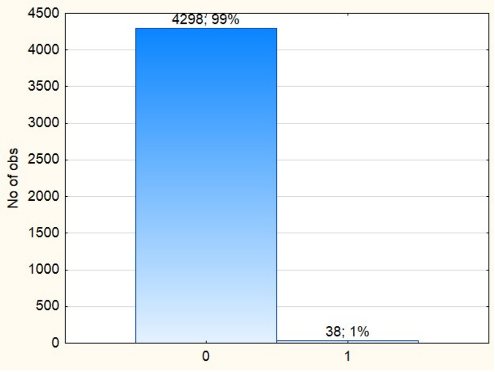


Figure B.65: `Physical Address` shared on – (Twitter).

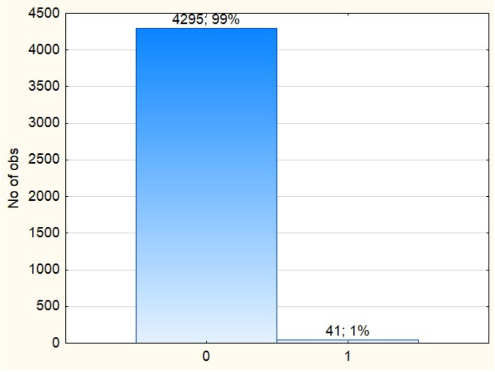


Figure B.66: `Physical Address` shared on – (Instagram).

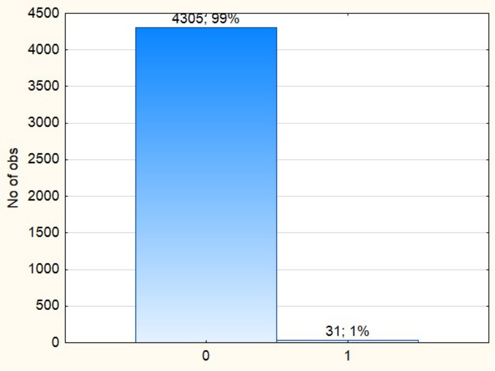


Figure B.67: `Physical Address` shared on – (YouTube).

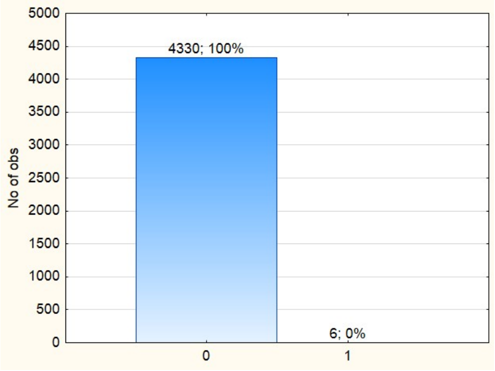


Figure B.68: `Physical Address` shared on – (Pinterest).

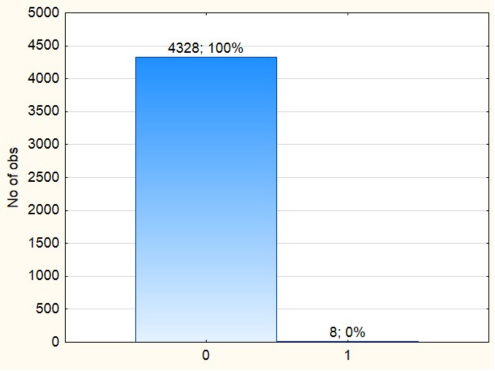


Figure B.69: `Physical Address` shared on – (Mxit).

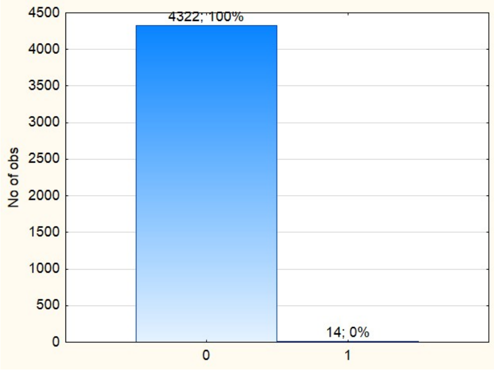


Figure B.70: `Physical Address` shared on – (Dating Sites).

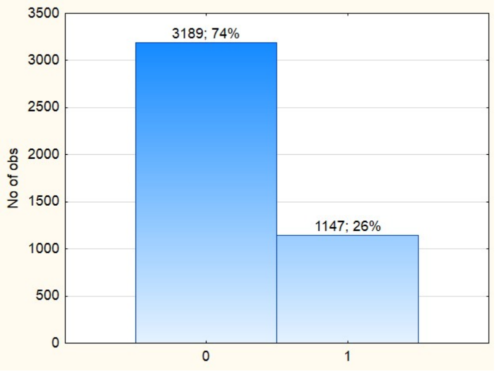


Figure B.71: **`Email`** shared on – (None).

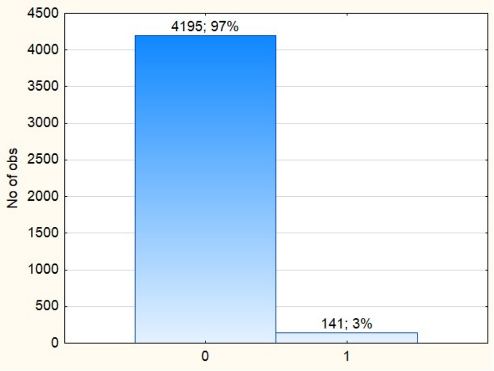


Figure B.72: **`Email`** shared on – (Other).

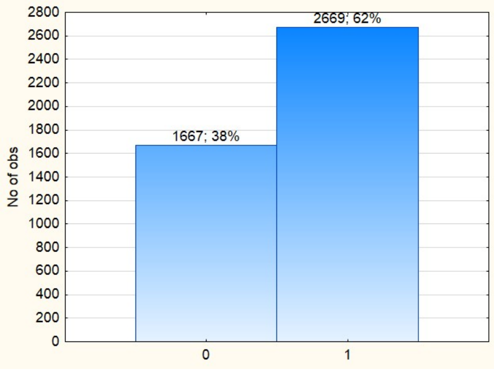


Figure B.73: **`Email`** shared on – (Facebook).

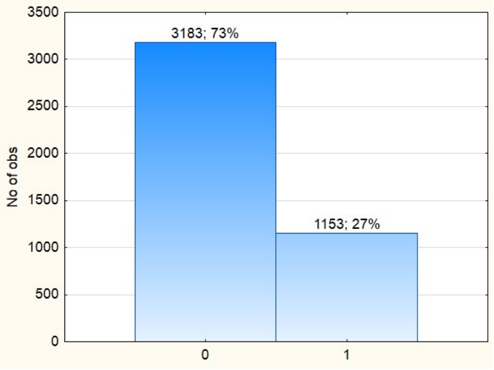


Figure B.74: **`Email`** shared on – (LinkedIn).

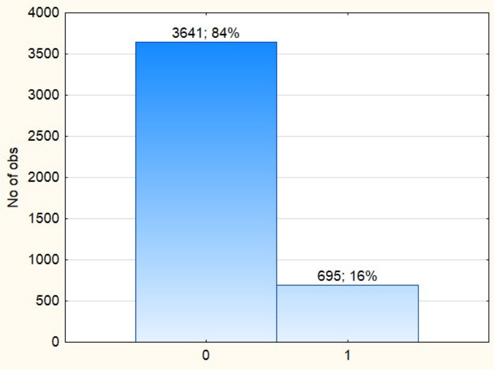


Figure B.75: **`Email`** shared on – (Twitter).

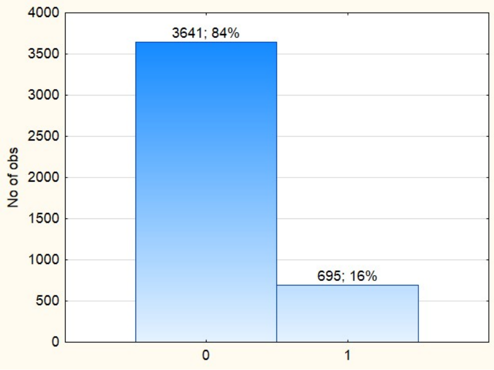


Figure B.76: **`Email`** shared on – (Instagram).

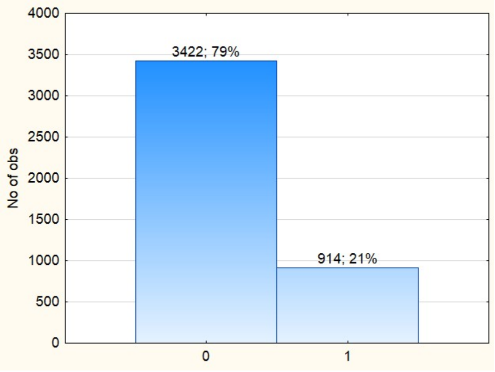


Figure B.77: `Email` shared on – (YouTube).

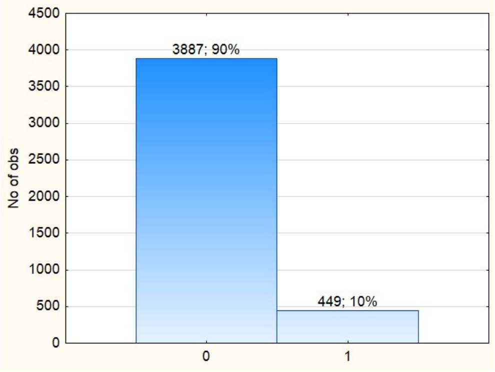


Figure B.78: `Email` shared on – (Pinterest).

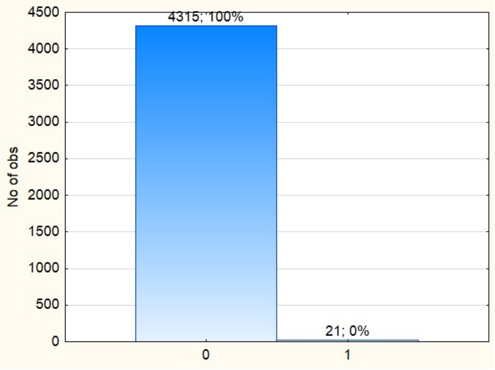


Figure B.79: `Email` shared on – (Mxit).

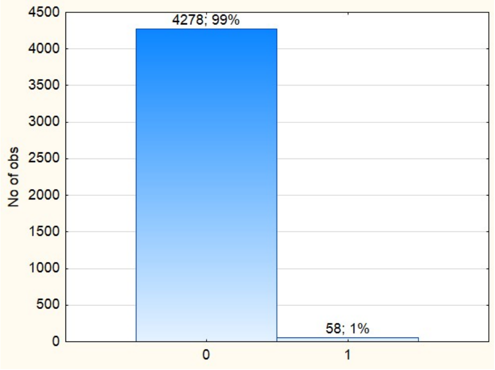


Figure B.80: `Email` shared on – (Dating Sites).

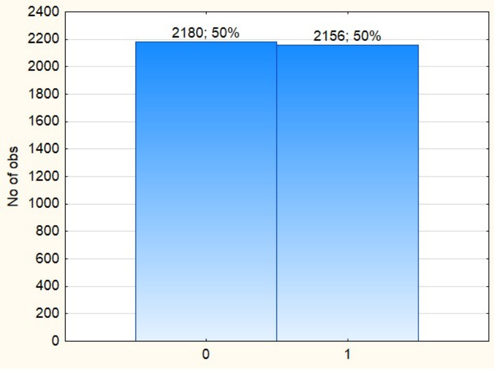


Figure B.81: `Cellphone Number` shared on – (None).

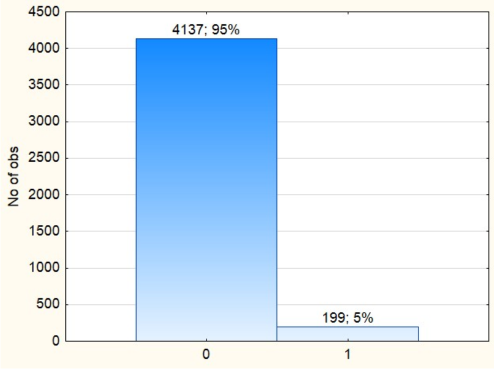


Figure B.82: `Cellphone Number` shared on – (Other).

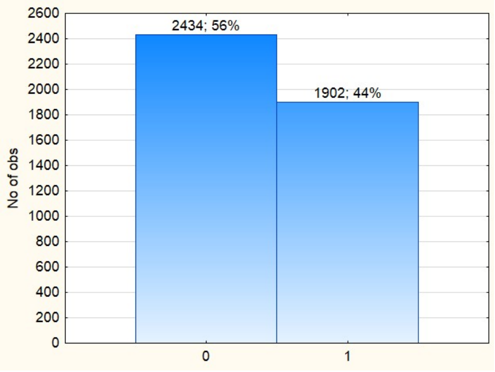


Figure B.83: `Cellphone Number` shared on – (Facebook).

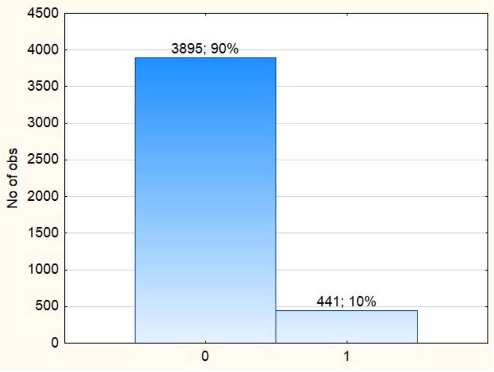


Figure B.84: `Cellphone Number` shared on – (LinkedIn).

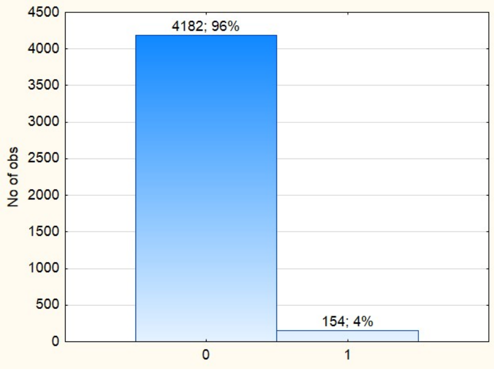


Figure B.85: `Cellphone Number` shared on – (Twitter).

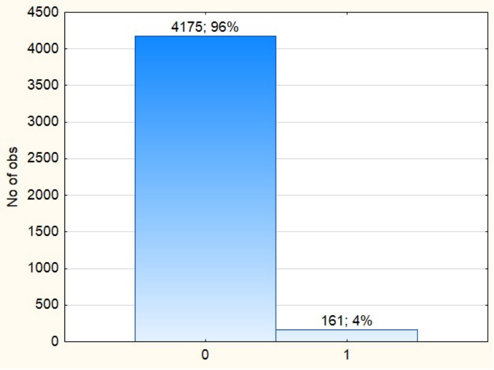


Figure B.86: `Cellphone Number` shared on – (Instagram).

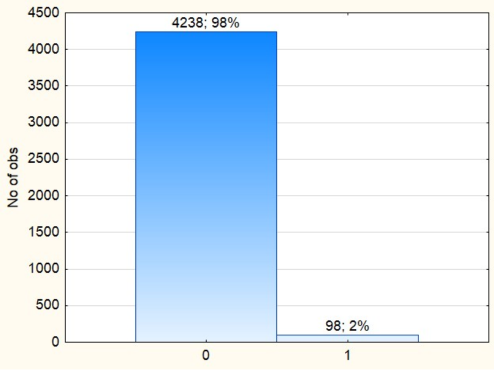


Figure B.87: `Cellphone Number` shared on – (YouTube).

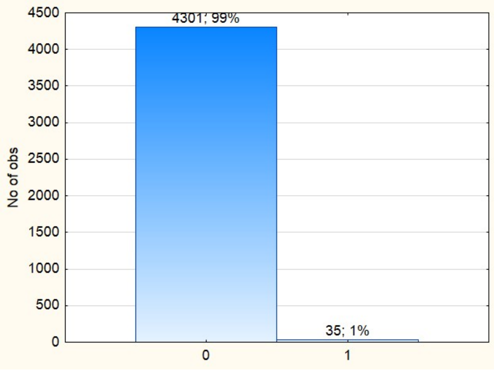


Figure B.88: `Cellphone Number` shared on – (Pinterest).

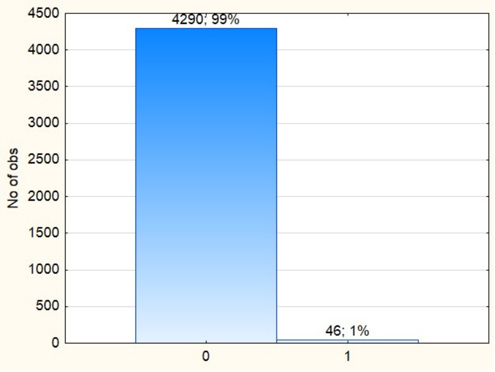


Figure B.89: `Cellphone Number` shared on – (Mxit).

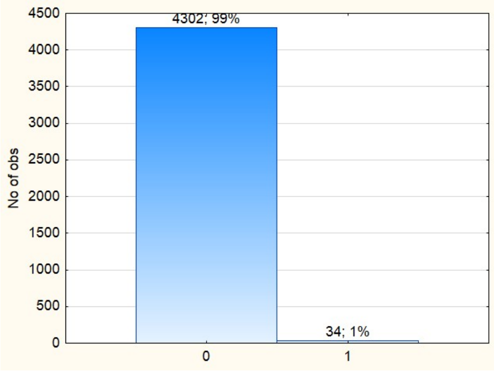


Figure B.90: `Cellphone Number` shared on – (Dating Sites).

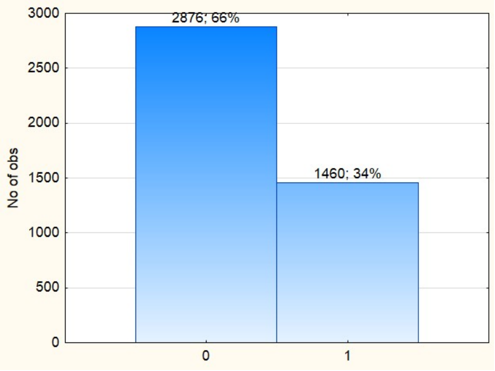


Figure B.91: `Relationship Status` shared on – (None).

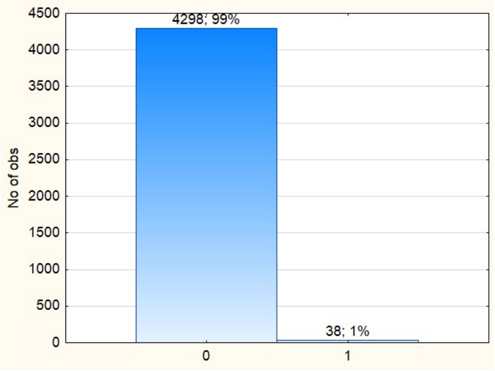


Figure B.92: `Relationship Status` shared on – (Other).

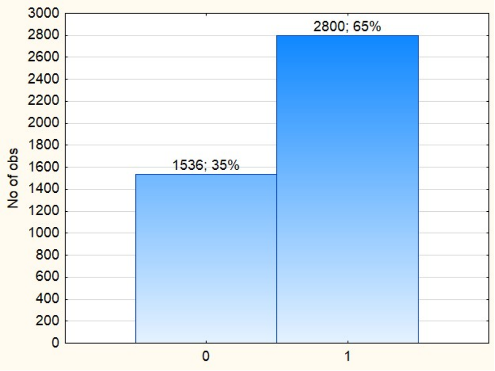


Figure B.93: `Relationship Status` shared on – (Facebook).

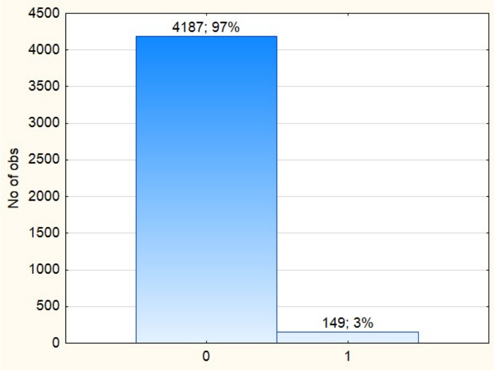


Figure B.94: `Relationship Status` shared on – (LinkedIn).

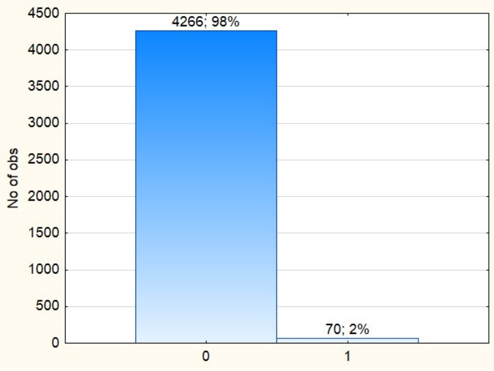


Figure B.95: `Relationship Status` shared on – (Twitter).

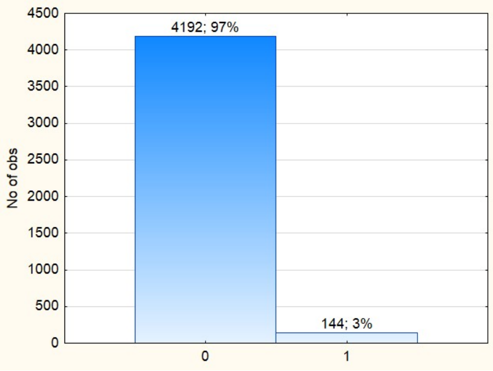


Figure B.96: `Relationship Status` shared on – (Instagram).

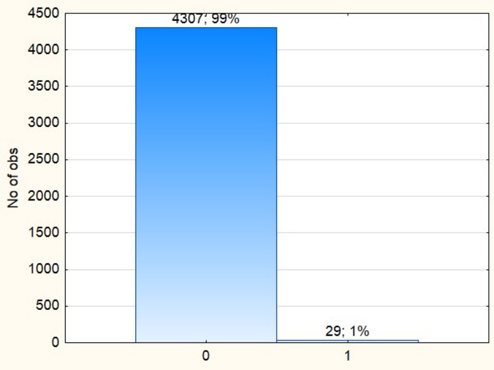


Figure B.97: `Relationship Status` shared on – (YouTube).

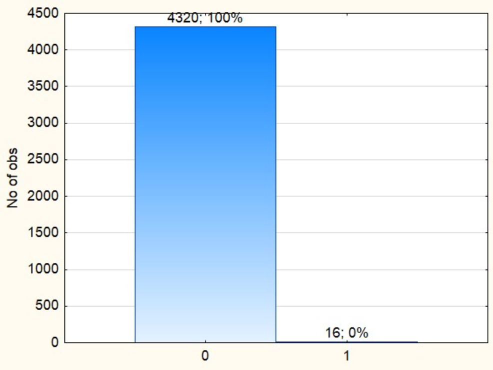


Figure B.98: `Relationship Status` shared on – (Pinterest).

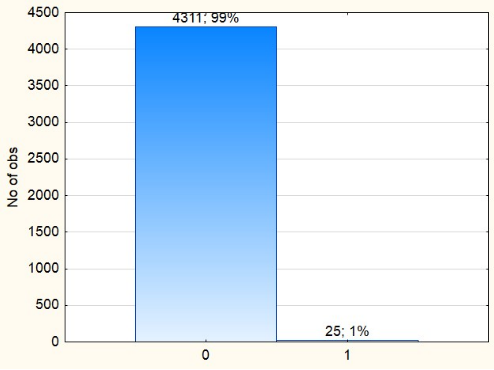


Figure B.99: `Relationship Status` shared on – (Mxit).

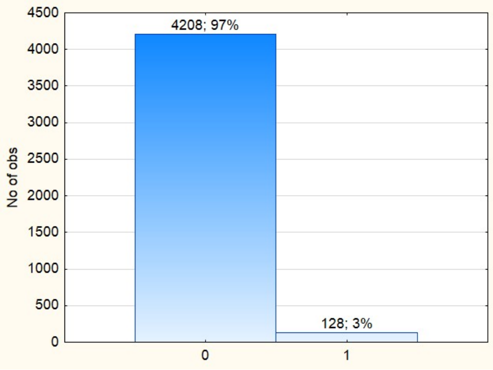


Figure B.100: `Relationship Status` shared on – (Dating Sites).

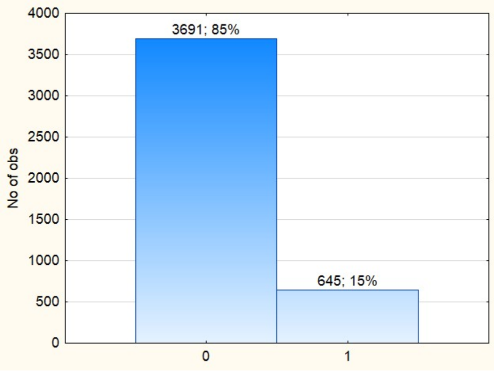


Figure B.101: `School Details` shared on – (None).

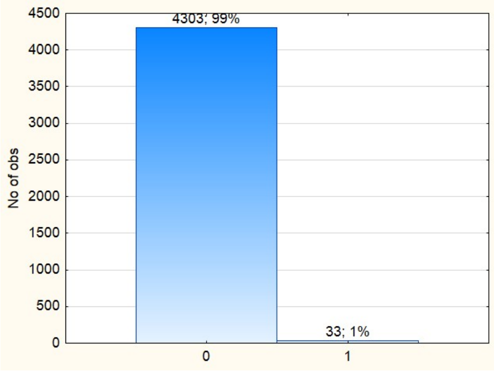


Figure B.102: `School Details` shared on – (Other).

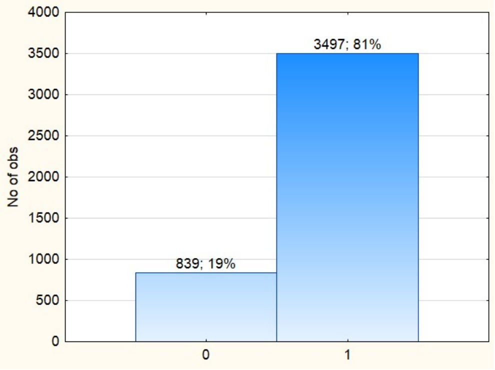


Figure B.103: `School Details` shared on – (Facebook).

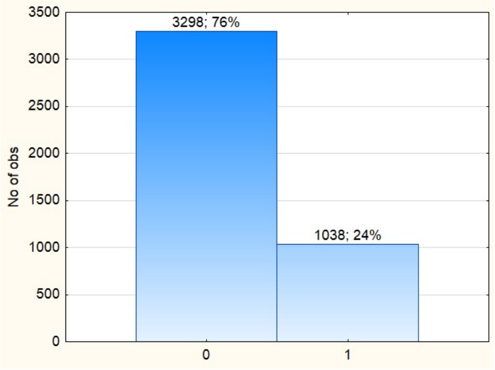


Figure B.104: `School Details` shared on – (LinkedIn).

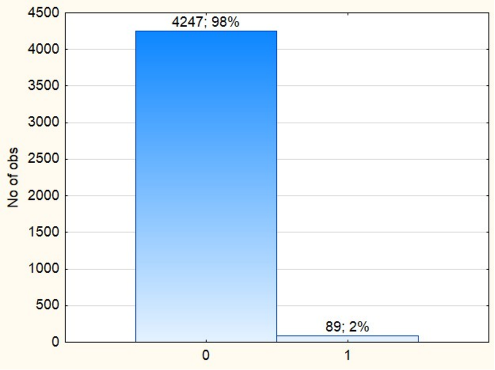


Figure B.105: `School Details` shared on – (Twitter).

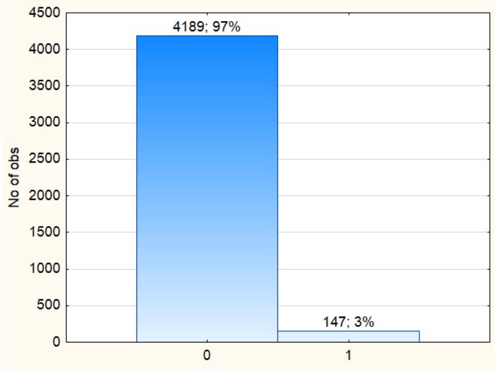


Figure B.106: `School Details` shared on – (Instagram).

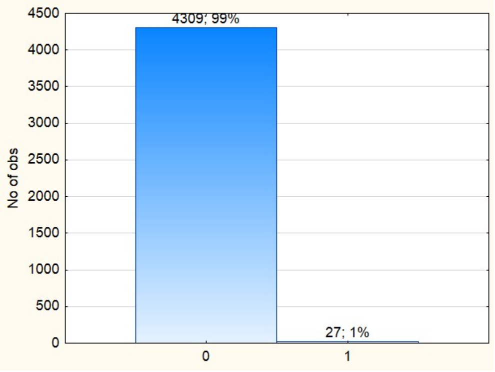


Figure B.107: `School Details` shared on – (YouTube).

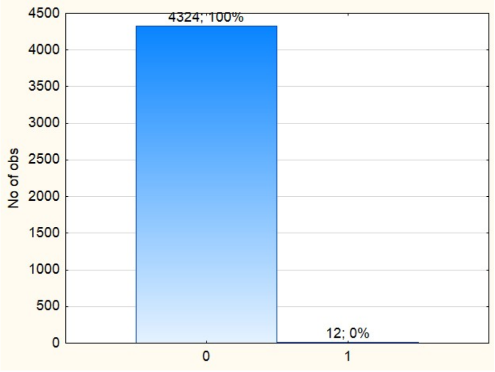


Figure B.108: `School Details` shared on – (Pinterest).

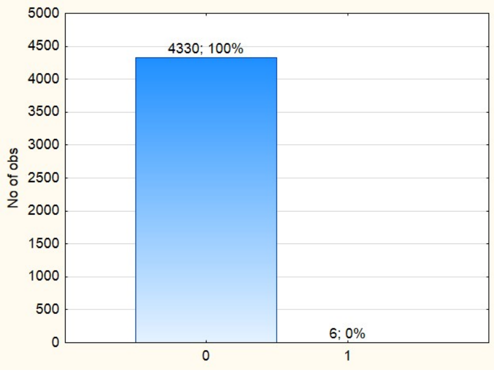


Figure B.109: `School Details` shared on – (Mxit).

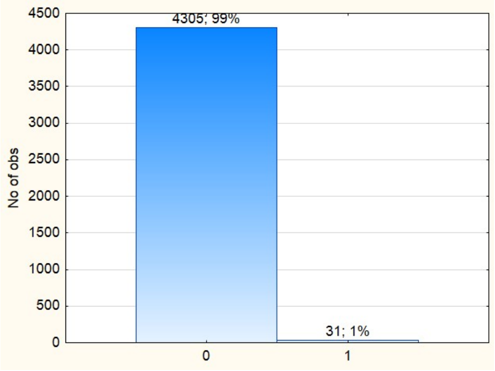


Figure B.110: `School Details` shared on – (Dating Sites).

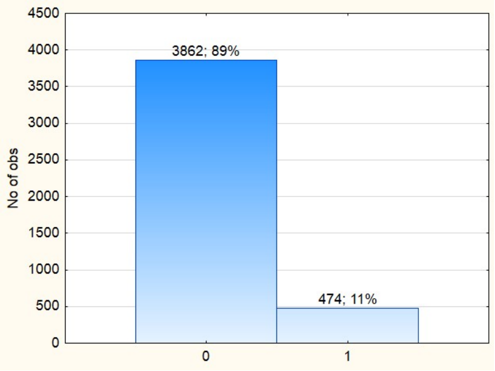


Figure B.111: `University Details` shared on – (None).

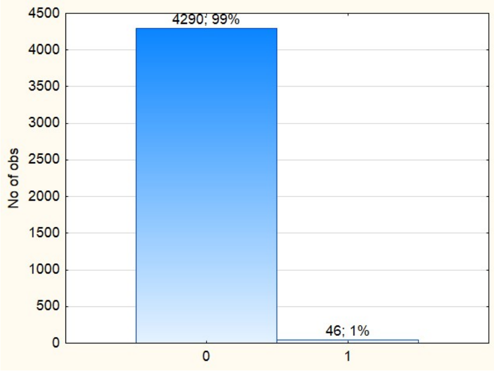


Figure B.112: `University Details` shared on – (Other).

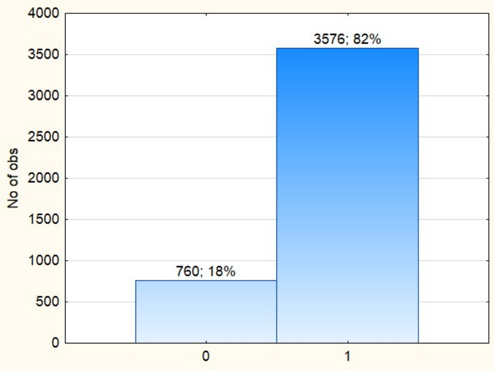


Figure B.113: `University Details` shared on – (Facebook).

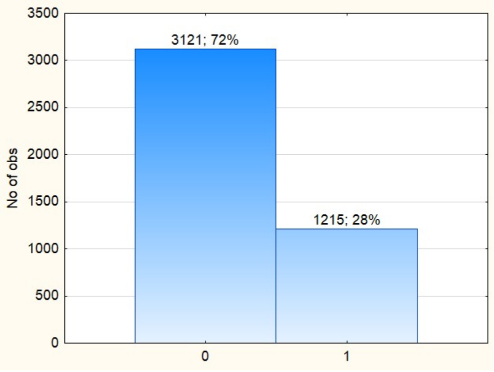


Figure B.114: `University Details` shared on – (LinkedIn).

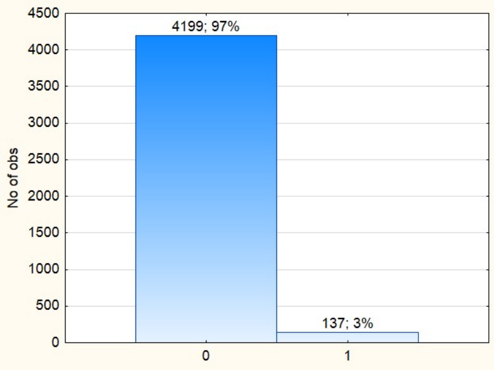


Figure B.115: `University Details` shared on – (Twitter).

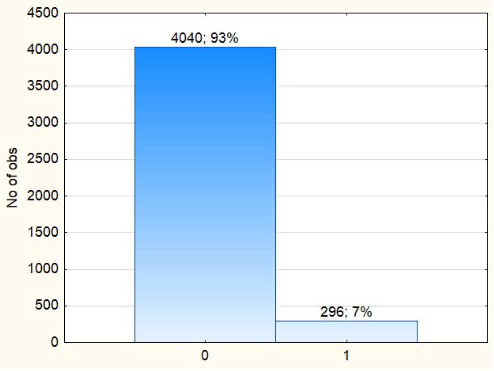


Figure B.116: `University Details` shared on – (Instagram).

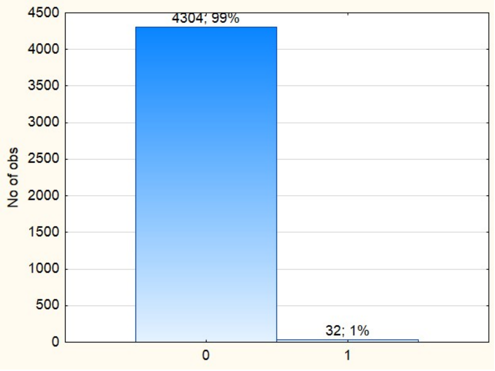


Figure B.117: `University Details` shared on – (YouTube).

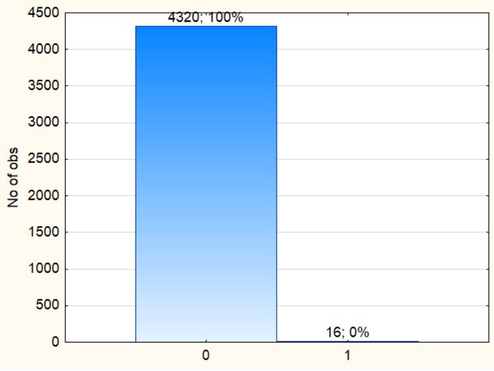


Figure B.118: `University Details` shared on – (Pinterest).

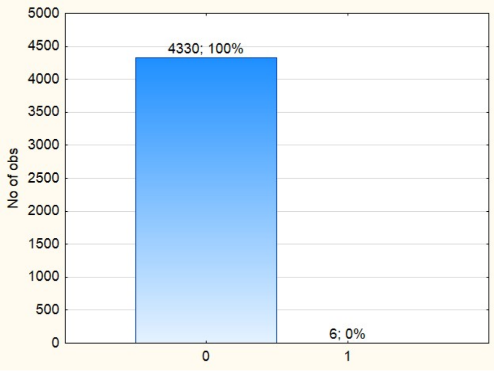


Figure B.119: `University Details` shared on – (Mxit).

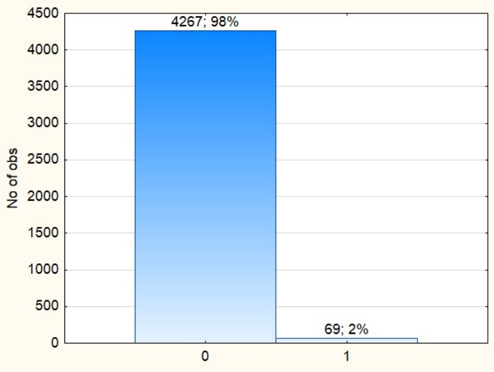


Figure B.120: `University Details` shared on – (Dating Sites).

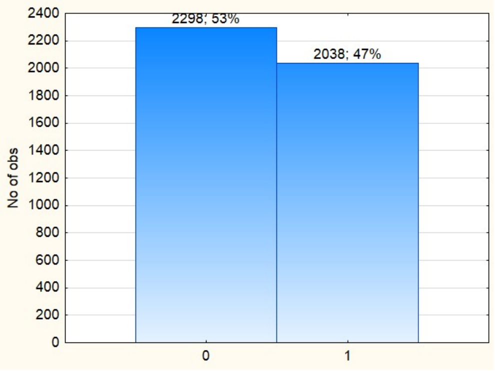


Figure B.121: `Job Details` shared on – (None).

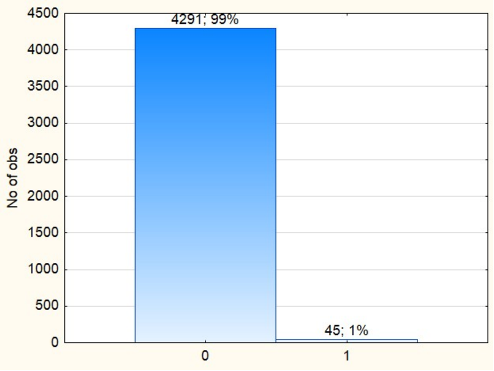


Figure B.122: `Job Details` shared on – (Other).

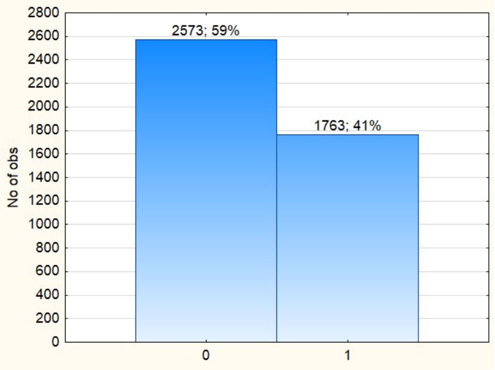


Figure B.123: `Job Details` shared on – (Facebook).

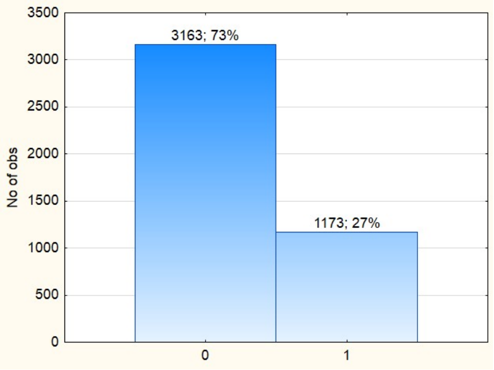


Figure B.124: `Job Details` shared on – (LinkedIn).

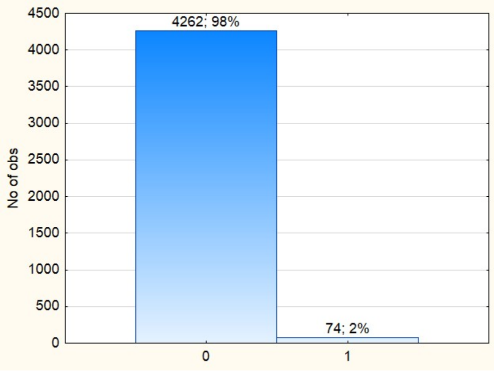


Figure B.125: `Job Details` shared on – (Twitter).

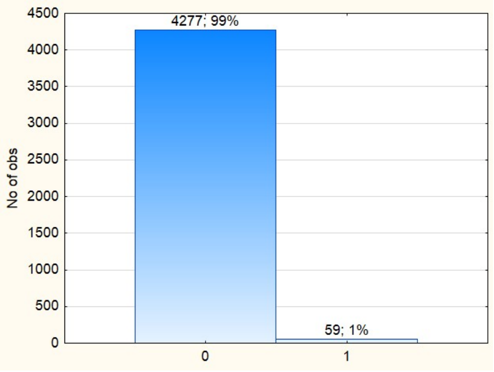


Figure B.126: `Job Details` shared on – (Instagram).

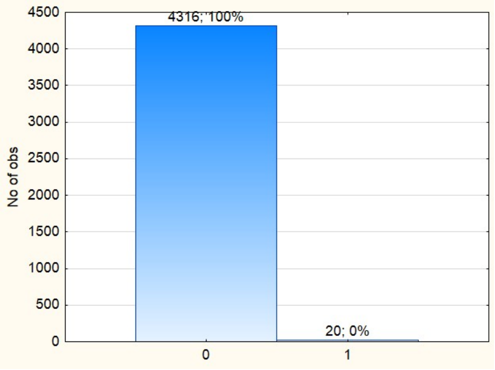


Figure B.127: `Job Details` shared on – (YouTube).

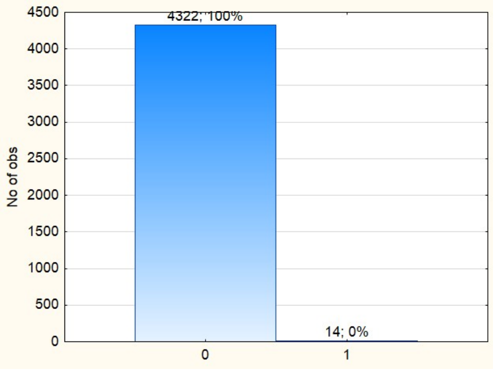


Figure B.128: `Job Details` shared on – (Pinterest).

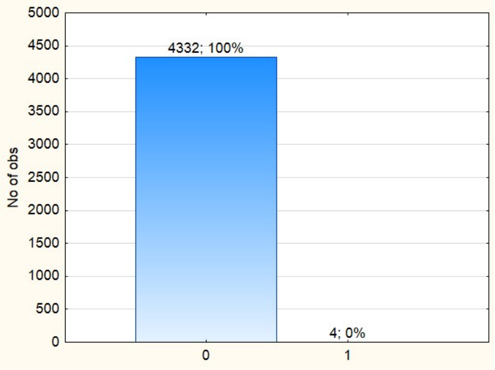


Figure B.129: `Job Details` shared on – (Mxit).

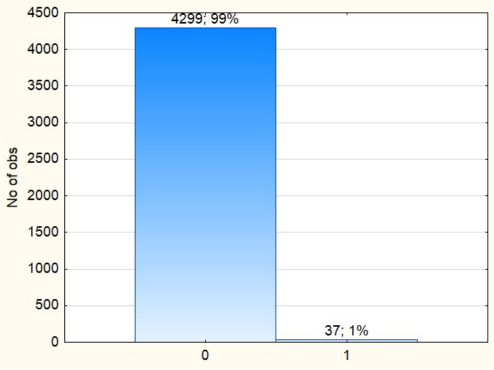


Figure B.130: `Job Details` shared on – (Dating Sites).

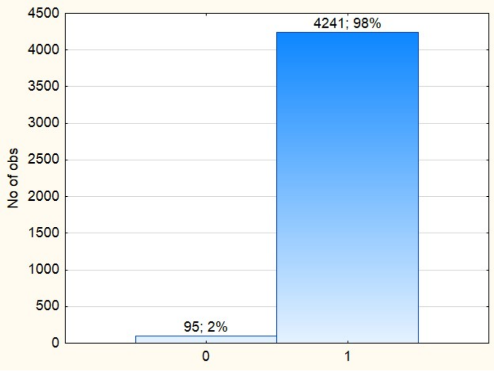


Figure B.131: `Income` shared on – (None).

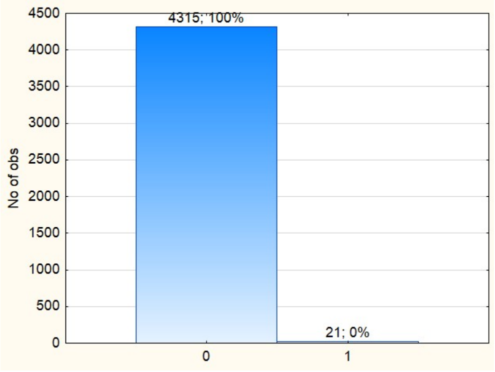


Figure B.132: `Income` shared on – (Other).

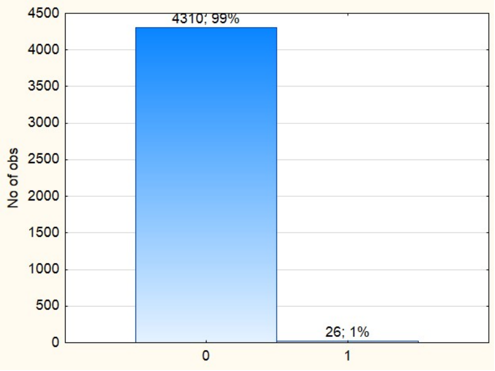


Figure B.133: `Income` shared on – (Facebook).

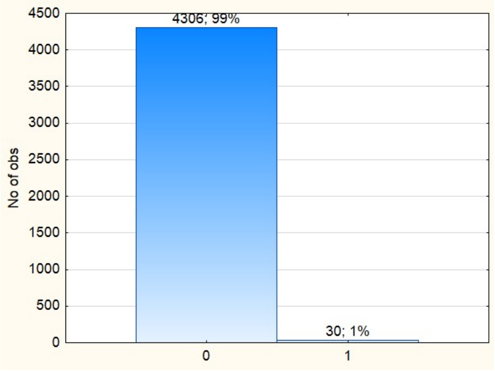


Figure B.134: `Income` shared on – (LinkedIn).

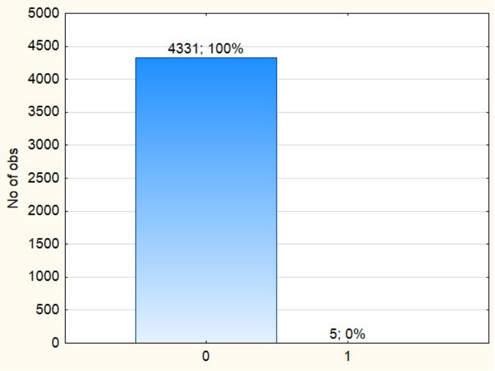


Figure B.135: `Income` shared on – (Twitter).

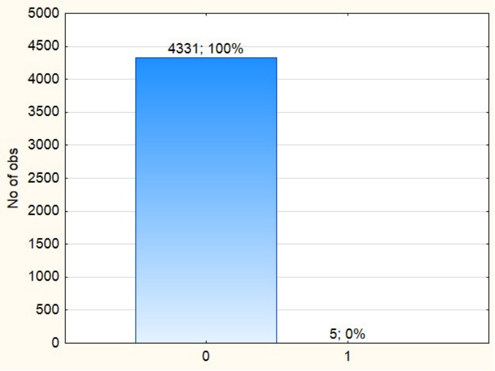


Figure B.136: `Income` shared on – (Instagram).

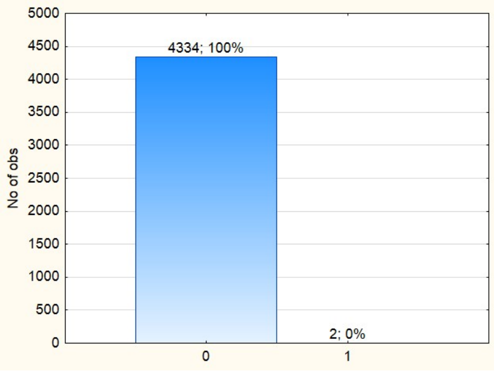


Figure B.137: `Income` shared on – (YouTube).

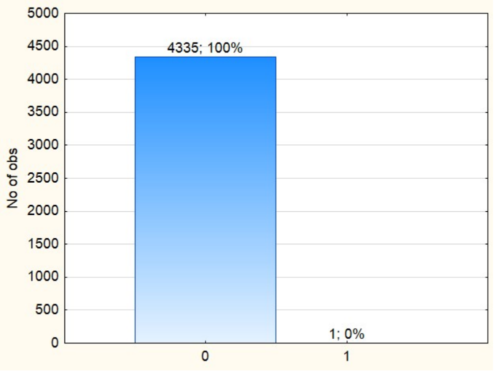


Figure B.138: `Income` shared on – (Pinterest).

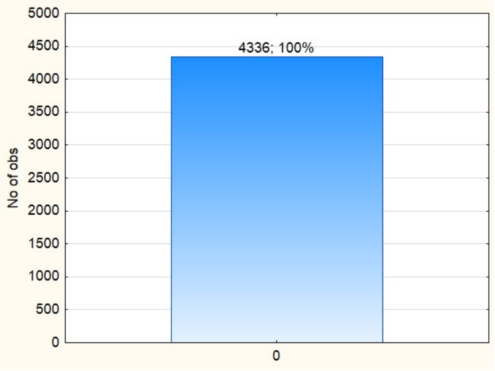


Figure B.139: `Income` shared on – (Mxit).

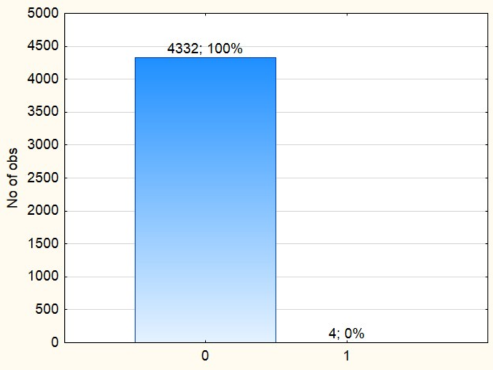


Figure B.140: `Income` shared on – (Dating Sites).

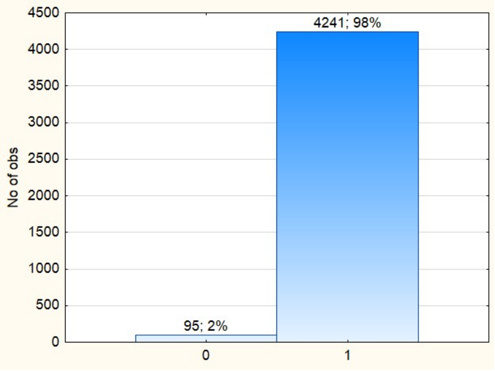


Figure B.141: `Credit Details` shared on – (None).

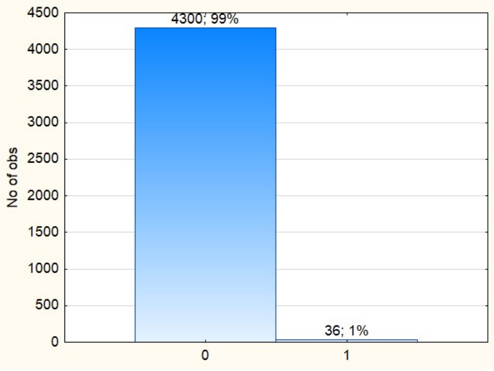


Figure B.142: `Credit Details` shared on – (Other).

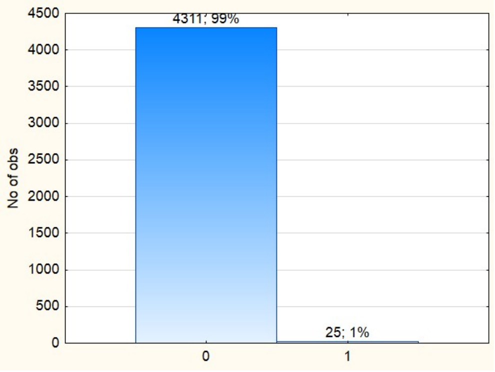


Figure B.143: `Credit Details` shared on – (Facebook).

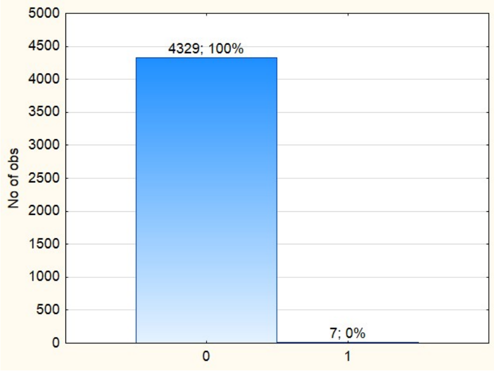


Figure B.144: `Credit Details` shared on – (LinkedIn).

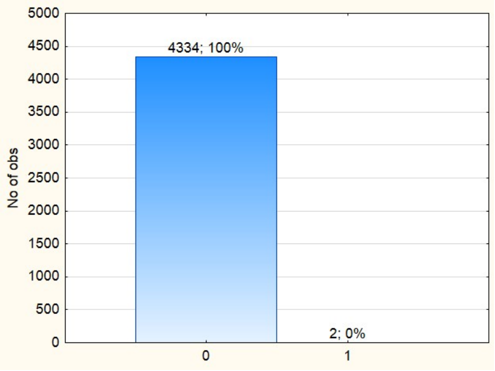


Figure B.145: `Credit Details` shared on – (Twitter).

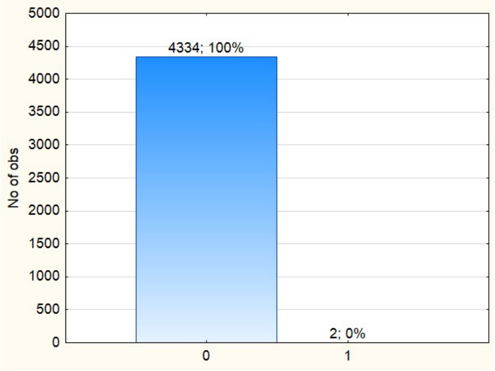


Figure B.146: `Credit Details` shared on – (Instagram).

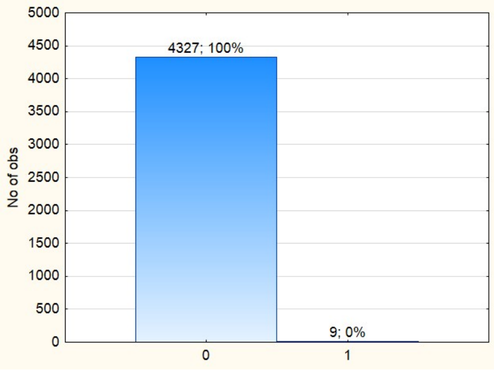


Figure B.147: `Credit Details` shared on – (YouTube).

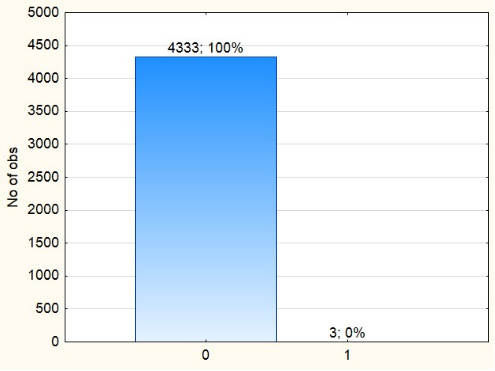


Figure B.148: `Credit Details` shared on – (Pinterest).

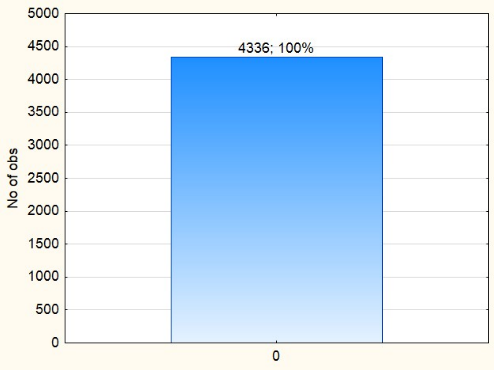


Figure B.149: `Credit Details` shared on – (Mxit).

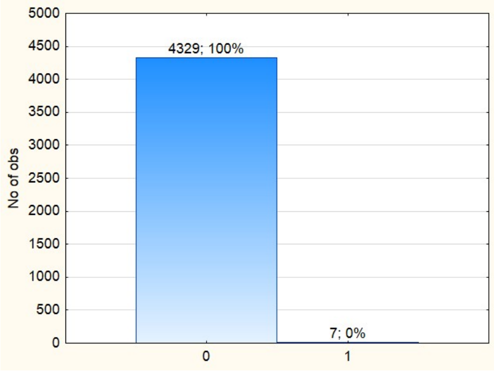


Figure B.150: `Credit Details` shared on – (Dating Sites).

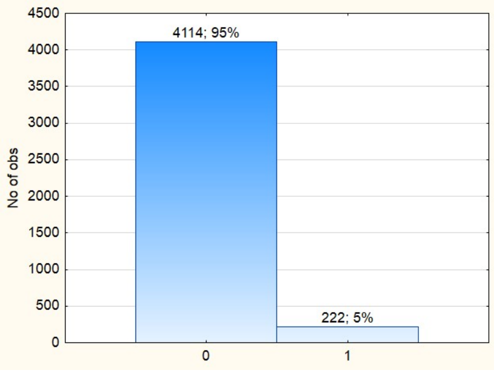


Figure B.151: `Victim` Type (1).

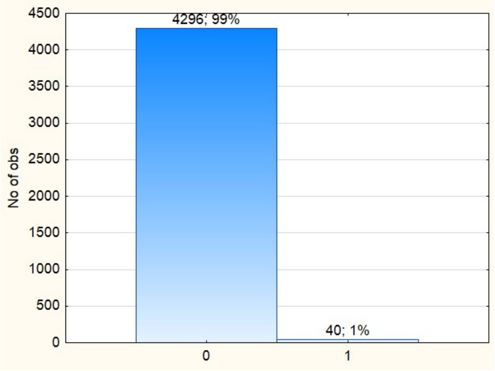


Figure B.152: `Victim` Type (2).

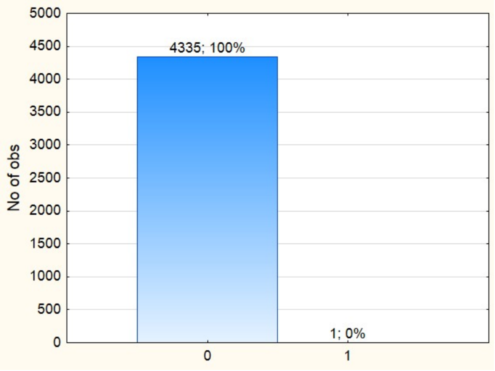


Figure B.153: `Victim` Type (3).

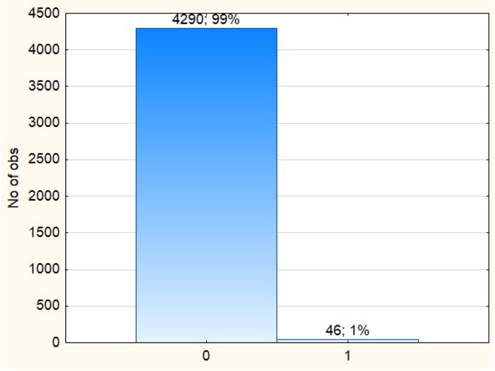


Figure B.154: `Victim` Type (4).

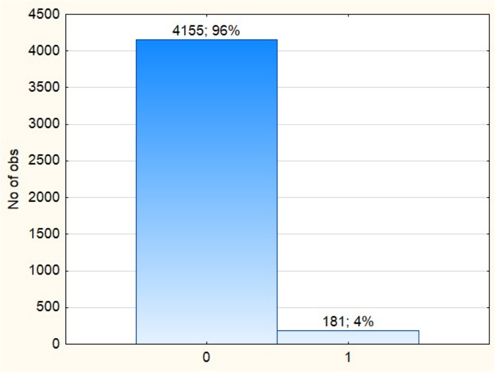


Figure B.155: `Victim` Type (5).

B.2 Relationship diagrams of shared attributes

The relationship diagrams included in this section present the identity theft victim percentages for the shared attributes according to the various social media sites's non-subscribers versus subscribers.

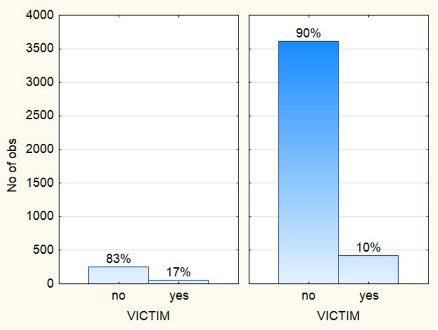


Figure B.156: Victim % for `Name` (Facebook Non-Subscribers vs. Subscribers).

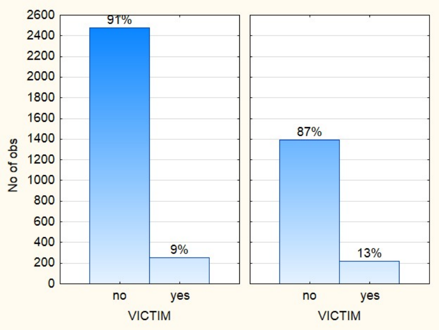


Figure B.157: Victim % for `Name` (LinkedIn Non-Subscribers vs. Subscribers).

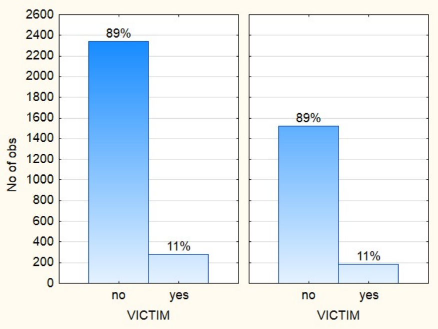


Figure B.158: Victim % for `Name` (Twitter Non-Subscribers vs. Subscribers).

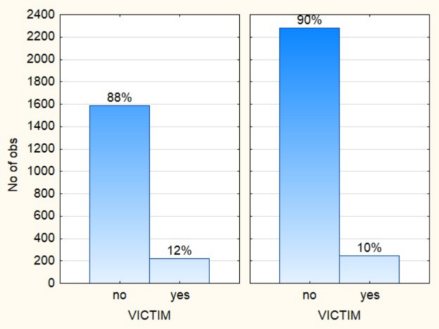


Figure B.159: Victim % for `Name` (Instagram Non-Subscribers vs. Subscribers).

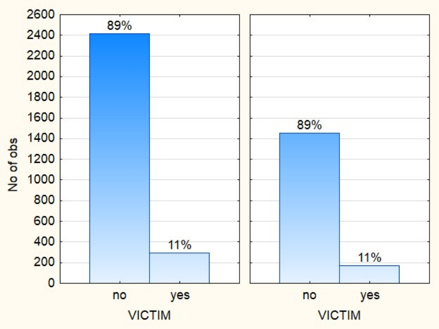


Figure B.160: Victim % for Name (YouTube Non-Subscribers vs. Subscribers).

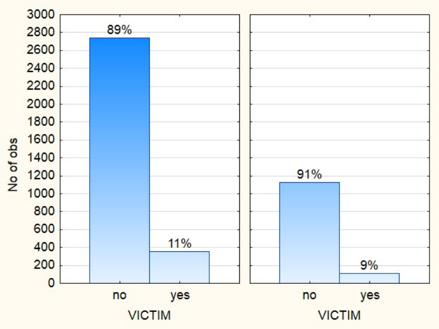


Figure B.161: Victim % for Name (Pinterest Non-Subscribers vs. Subscribers).

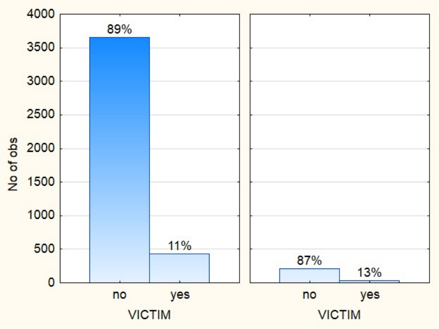


Figure B.162: Victim % for Name (Dating Sites Non-Subscribers vs. Subscribers).

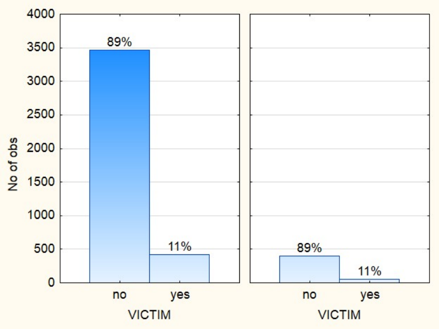


Figure B.163: Victim % for Name (Other Non-Subscribers vs. Subscribers).

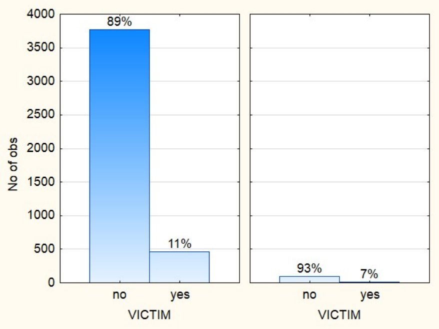


Figure B.164: Victim % for Surname (All vs. None).

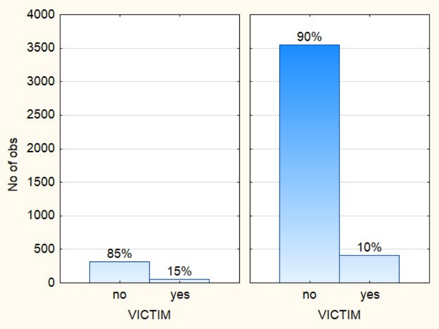


Figure B.165: Victim % for Surname (Facebook Non-Subscribers vs. Subscribers).

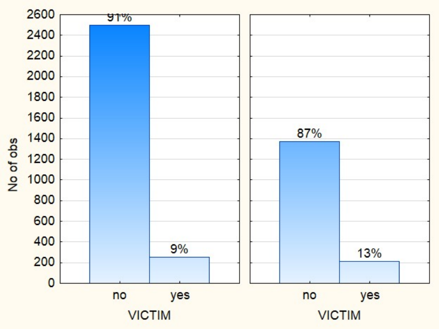


Figure B.166: Victim % for Surname (LinkedIn Non-Subscribers vs. Subscribers).

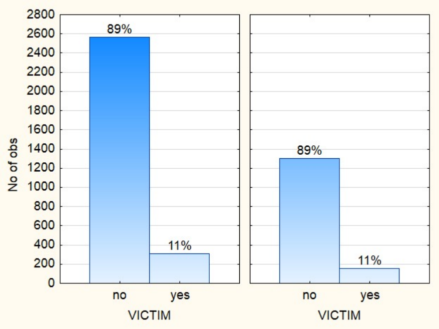


Figure B.167: Victim % for Surname (Twitter Non-Subscribers vs. Subscribers).

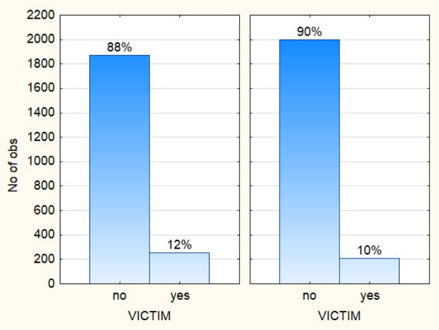


Figure B.168: Victim % for Surname (Instagram Non-Subscribers vs. Subscribers).

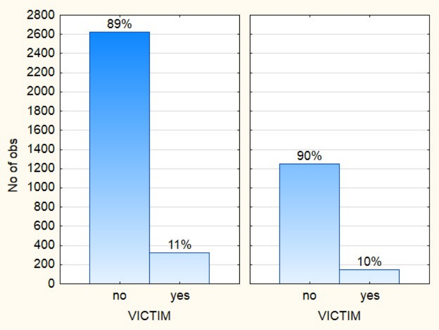


Figure B.169: Victim % for Surname (YouTube Non-Subscribers vs. Subscribers).

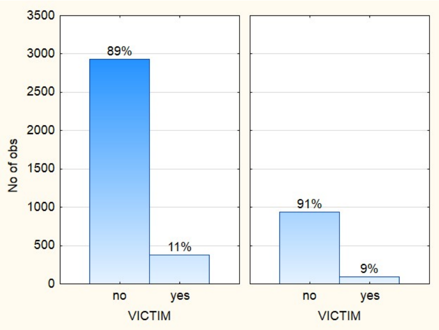


Figure B.170: Victim % for Surname (Pinterest Non-Subscribers vs. Subscribers).

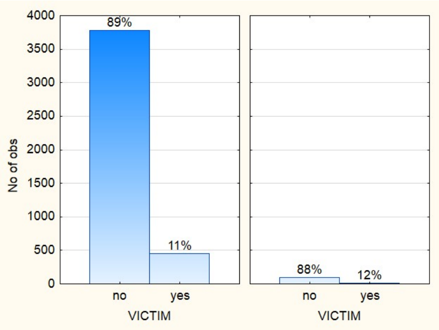


Figure B.171: Victim % for Surname (Dating Sites Non-Subscribers vs. Subscribers).

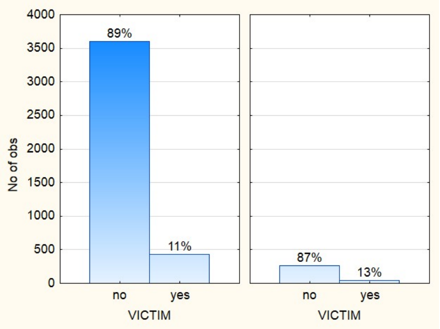


Figure B.172: Victim % for `Surname` (Other Non-Subscribers vs. Subscribers).

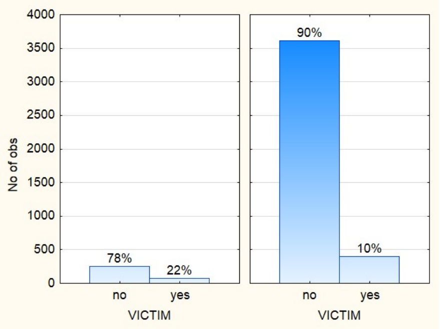


Figure B.173: Victim % for `Identity Number` (All vs. None).

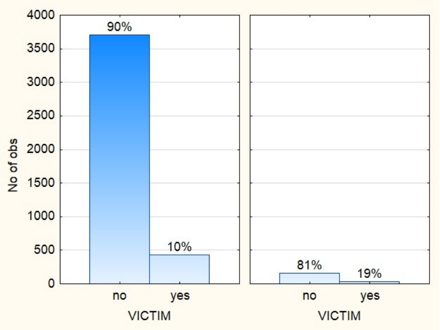


Figure B.174: Victim % for `Identity Number` (Facebook Non-Subscribers vs. Subscribers).

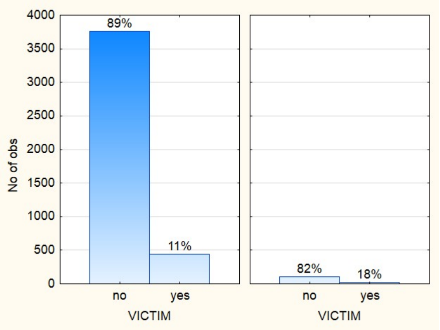


Figure B.175: Victim % for `Identity Number` (LinkedIn Non-Subscribers vs. Subscribers).

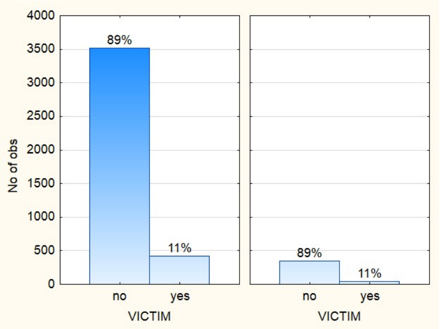


Figure B.176: Victim % for `Birthday` (All vs. None).

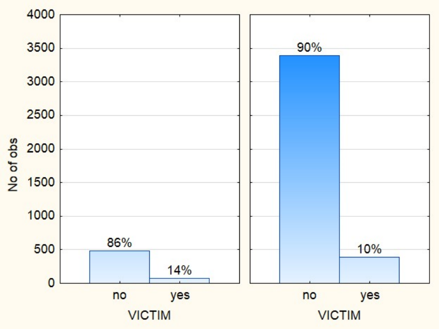


Figure B.177: Victim % for `Birthday` (Facebook Non-Subscribers vs. Subscribers).

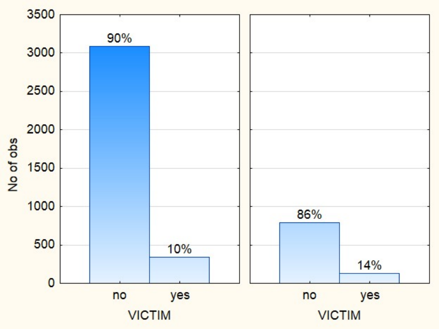


Figure B.178: Victim % for Birthday (LinkedIn Non-Subscribers vs. Subscribers).

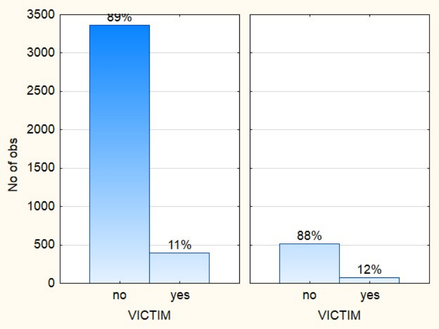


Figure B.179: Victim % for Birthday (Twitter Non-Subscribers vs. Subscribers).

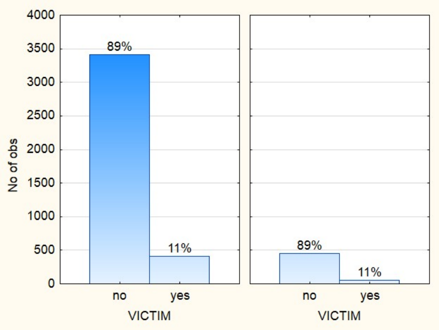


Figure B.180: Victim % for Birthday (Instagram Non-Subscribers vs. Subscribers).

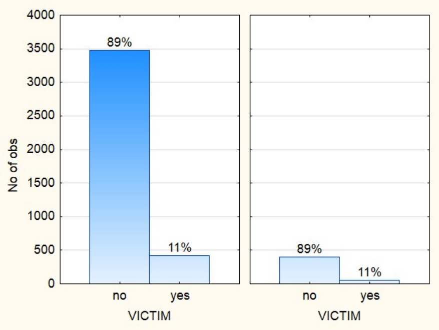


Figure B.181: Victim % for Birthday (YouTube Non-Subscribers vs. Subscribers).

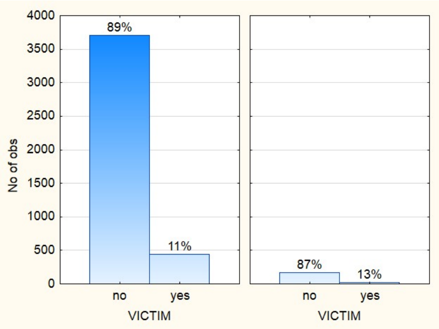


Figure B.182: Victim % for Birthday (Pinterest Non-Subscribers vs. Subscribers).

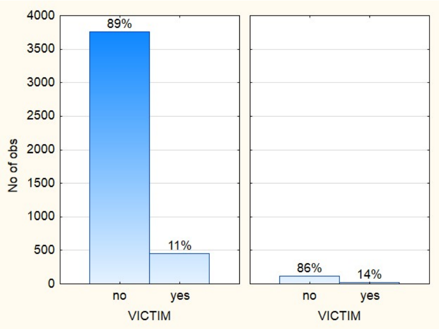


Figure B.183: Victim % for Birthday (Other Non-Subscribers vs. Subscribers).

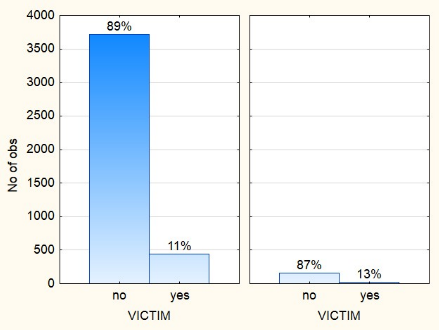


Figure B.184: Victim % for Gender (All vs. None).

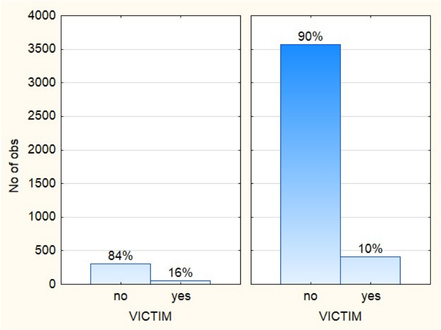


Figure B.185: Victim % for Gender (Facebook Non-Subscribers vs. Subscribers).

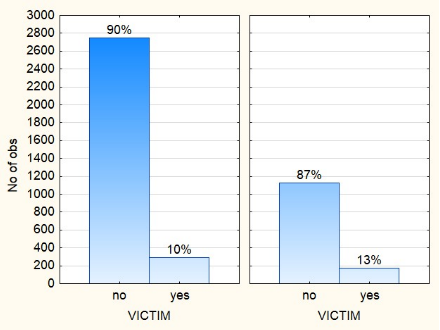


Figure B.186: Victim % for Gender (LinkedIn Non-Subscribers vs. Subscribers).

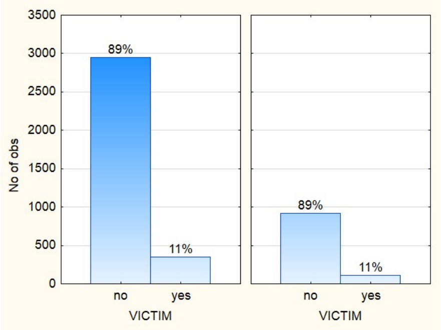


Figure B.187: Victim % for Gender (Twitter Non-Subscribers vs. Subscribers).

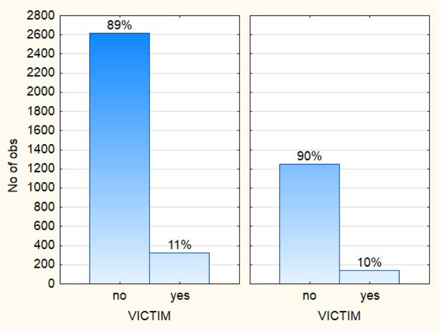


Figure B.188: Victim % for Gender (Instagram Non-Subscribers vs. Subscribers).

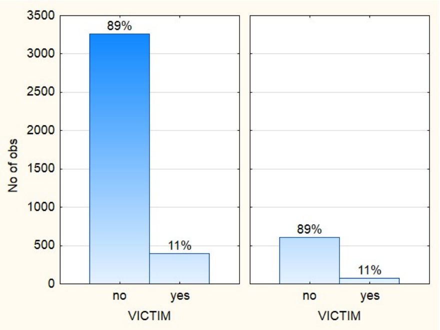


Figure B.189: Victim % for Gender (YouTube Non-Subscribers vs. Subscribers).

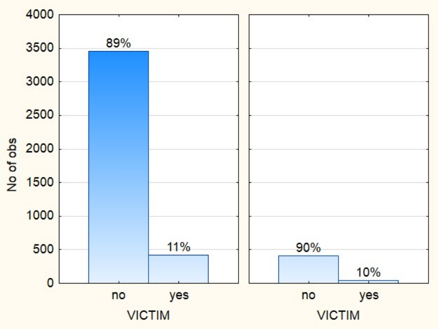


Figure B.190: Victim % for `Gender` (Pinterest Non-Subscribers vs. Subscribers).

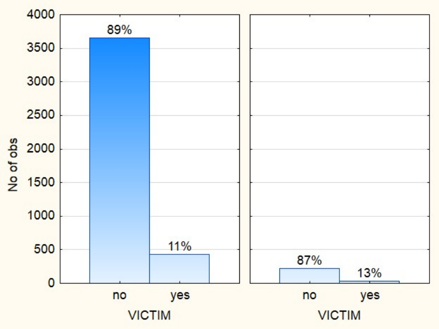


Figure B.191: Victim % for `Gender` (Dating Sites Non-Subscribers vs. Subscribers).

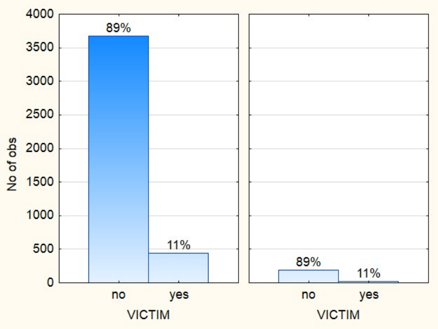


Figure B.192: Victim % for `Gender` (Other Non-Subscribers vs. Subscribers).

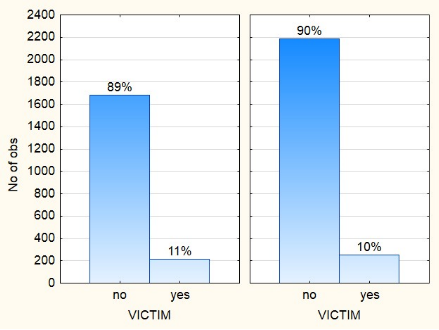


Figure B.193: Victim % for `Race` (All vs. None).

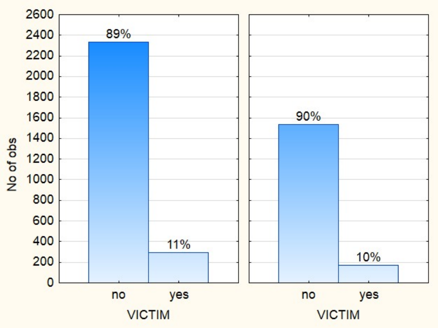


Figure B.194: Victim % for `Race` (Facebook Non-Subscribers vs. Subscribers).

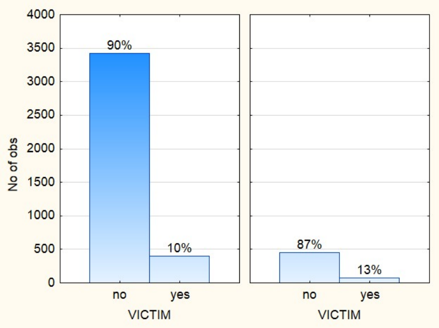


Figure B.195: Victim % for `Race` (LinkedIn Non-Subscribers vs. Subscribers).

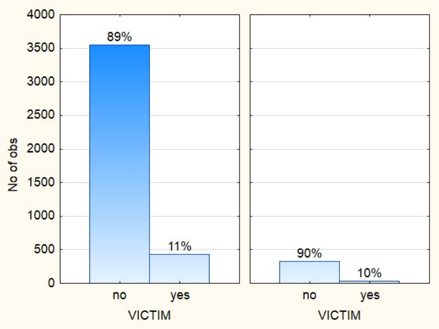


Figure B.196: Victim % for Race (Twitter Non-Subscribers vs. Subscribers).

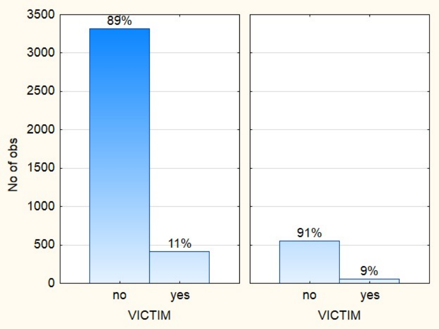


Figure B.197: Victim % for Race (Instagram Non-Subscribers vs. Subscribers).

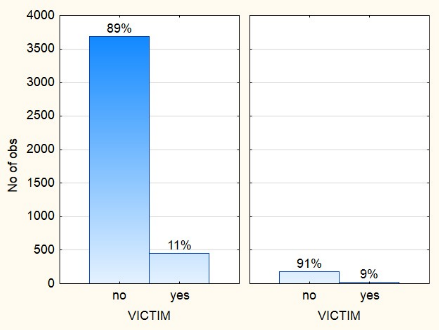


Figure B.198: Victim % for Race (YouTube Non-Subscribers vs. Subscribers).

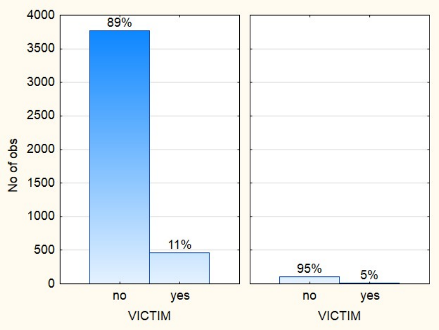


Figure B.199: Victim % for Race (Pinterest Non-Subscribers vs. Subscribers).

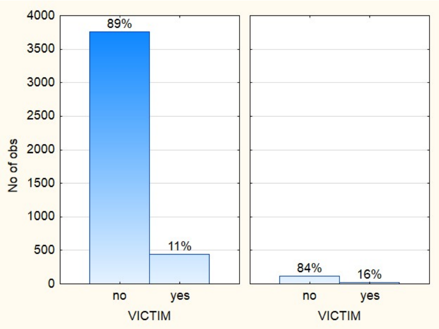


Figure B.200: Victim % for Race (Dating Sites Non-Subscribers vs. Subscribers).

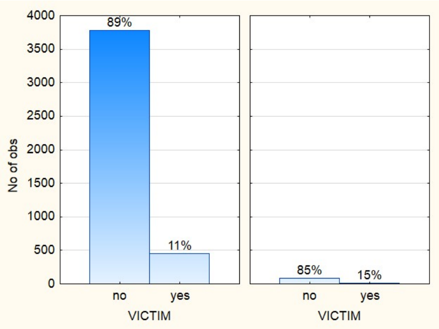


Figure B.201: Victim % for Race (Other Non-Subscribers vs. Subscribers).

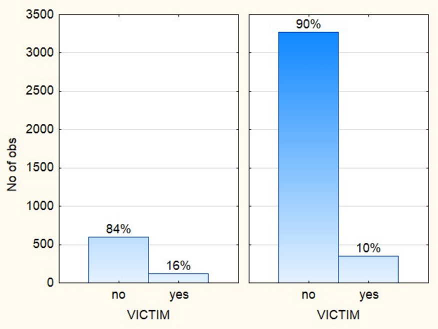


Figure B.202: Victim % for `Physical Address` (All vs. None).

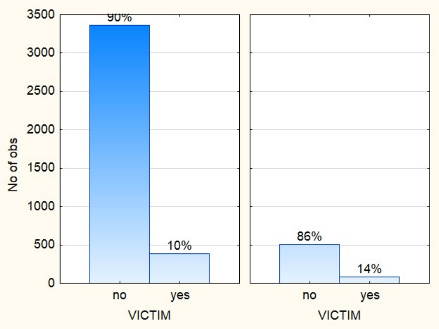


Figure B.203: Victim % for `Physical Address` (Facebook Non-Subscribers vs. Subscribers).

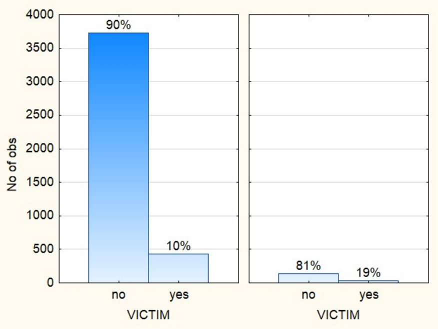


Figure B.204: Victim % for `Physical Address` (LinkedIn Non-Subscribers vs. Subscribers).

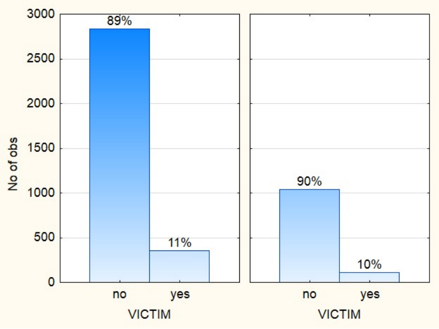


Figure B.205: Victim % for `Email` (All vs. None).

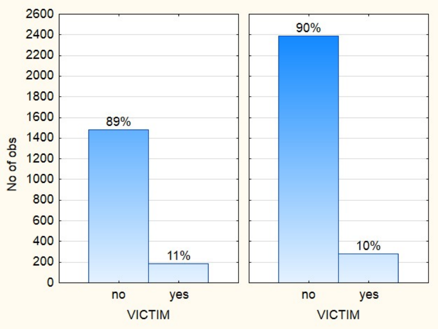


Figure B.206: Victim % for `Email` (Facebook Non-Subscribers vs. Subscribers).

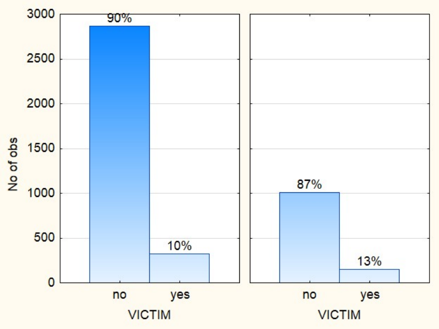


Figure B.207: Victim % for `Email` (LinkedIn Non-Subscribers vs. Subscribers).

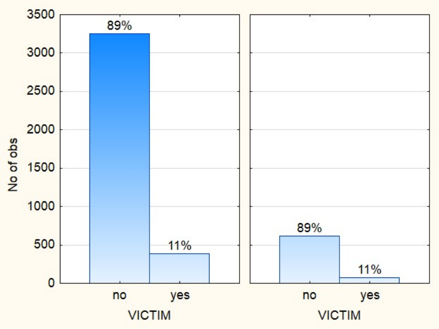


Figure B.208: Victim % for `Email` (Twitter Non-Subscribers vs. Subscribers).

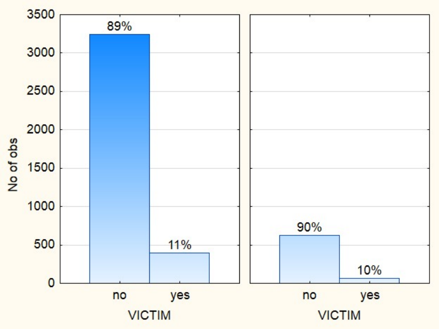


Figure B.209: Victim % for `Email` (Instagram Non-Subscribers vs. Subscribers).

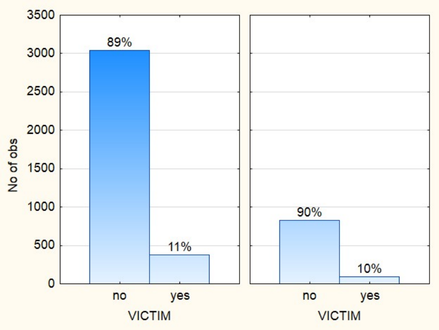


Figure B.210: Victim % for `Email` (YouTube Non-Subscribers vs. Subscribers).

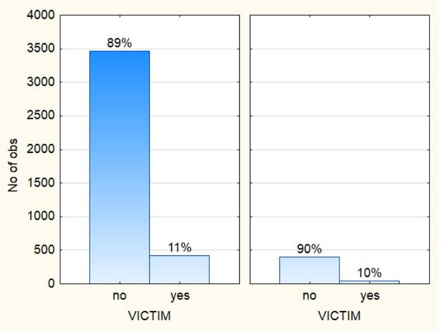


Figure B.211: Victim % for `Email` (Pinterest Non-Subscribers vs. Subscribers).

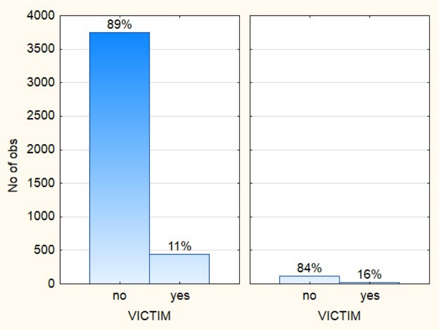


Figure B.212: Victim % for `Email` (Other Non-Subscribers vs. Subscribers).

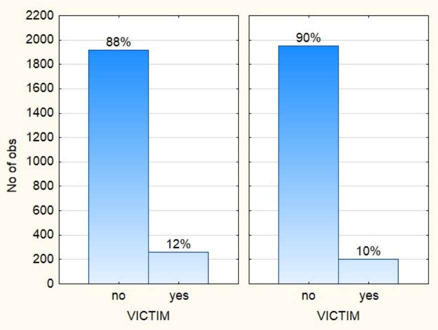


Figure B.213: Victim % for `Cellphone Number` (All vs. None).

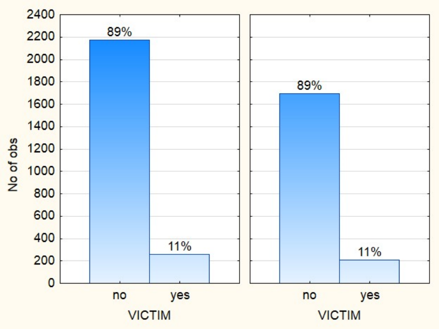


Figure B.214: Victim % for `Cellphone Number` (Facebook Non-Subscribers vs. Subscribers).

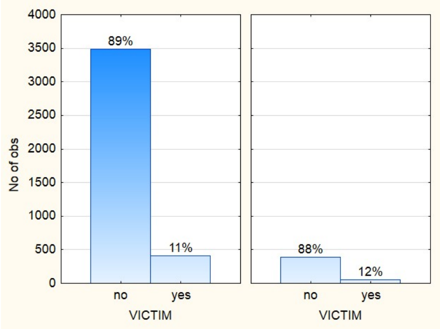


Figure B.215: Victim % for `Cellphone Number` (LinkedIn Non-Subscribers vs. Subscribers).

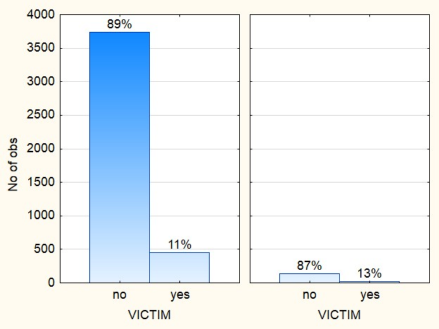


Figure B.216: Victim % for `Cellphone Number` (Twitter Non-Subscribers vs. Subscribers).

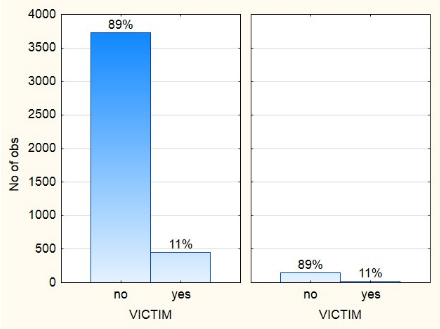


Figure B.217: Victim % for `Cellphone Number` (Instagram Non-Subscribers vs. Subscribers).

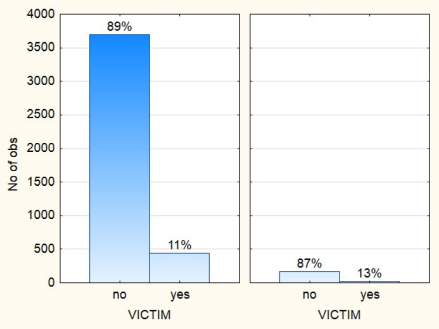


Figure B.218: Victim % for `Cellphone Number` (Other Non-Subscribers vs. Subscribers).

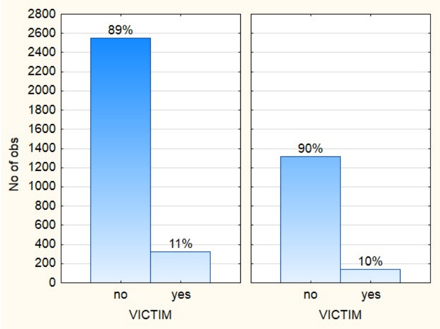


Figure B.219: Victim % for `Relationship Status` (All vs. None).

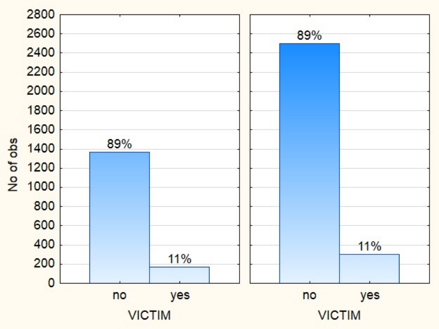


Figure B.220: Victim % for `Relationship Status` (Facebook Non-Subscribers vs. Subscribers).

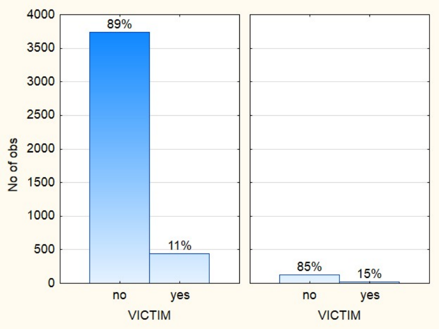


Figure B.221: Victim % for `Relationship Status` (LinkedIn Non-Subscribers vs. Subscribers).

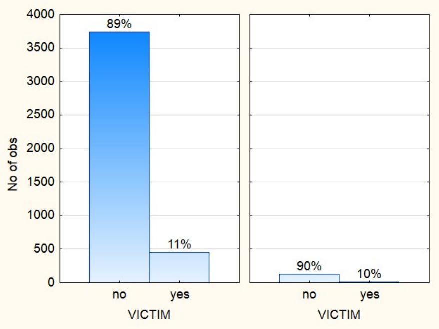


Figure B.222: Victim % for `Relationship Status` (Instagram Non-Subscribers vs. Subscribers).

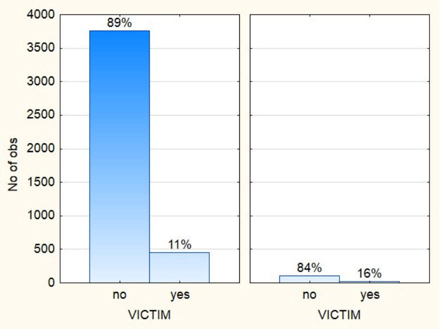


Figure B.223: Victim % for `Relationship Status` (Dating Sites Non-Subscribers vs. Subscribers).

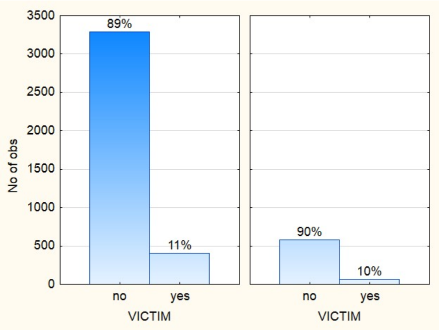


Figure B.224: Victim % for `School Details` (All vs. None).

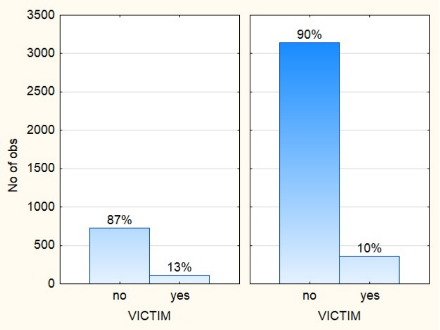


Figure B.225: Victim % for `School Details` (Facebook Non-Subscribers vs. Subscribers).

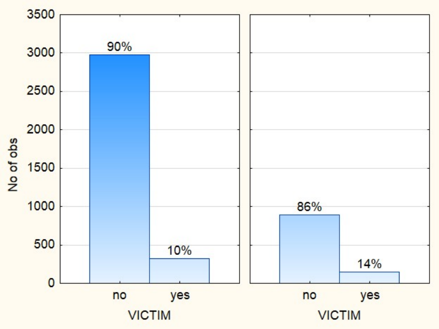


Figure B.226: Victim % for School Details (LinkedIn Non-Subscribers vs. Subscribers).

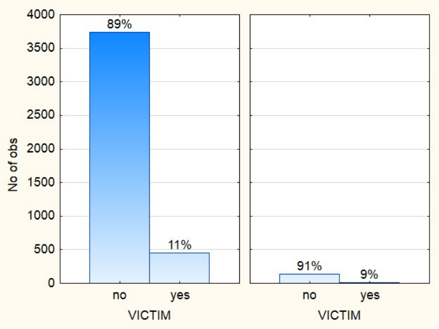


Figure B.227: Victim % for School Details (Instagram Non-Subscribers vs. Subscribers).

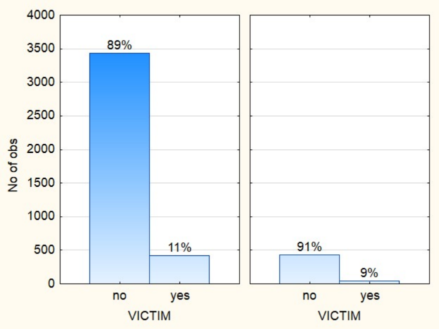


Figure B.228: Victim % for University Details (All vs. None).

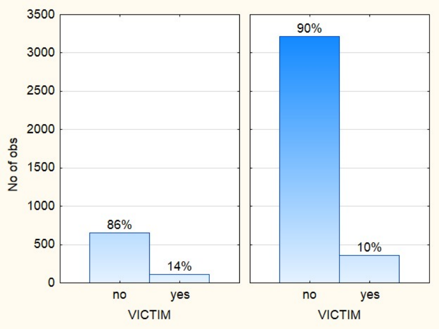


Figure B.229: Victim % for University Details (Facebook Non-Subscribers vs. Subscribers).

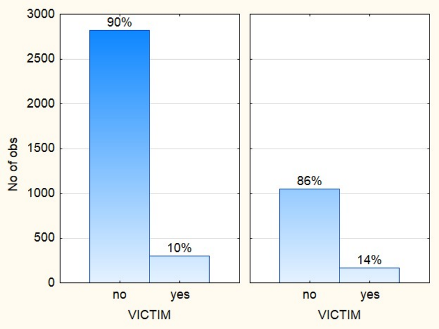


Figure B.230: Victim % for University Details (LinkedIn Non-Subscribers vs. Subscribers).

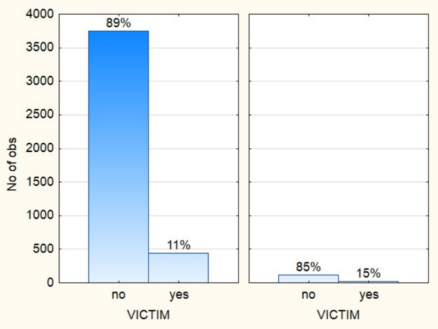


Figure B.231: Victim % for University Details (Twitter Non-Subscribers vs. Subscribers).

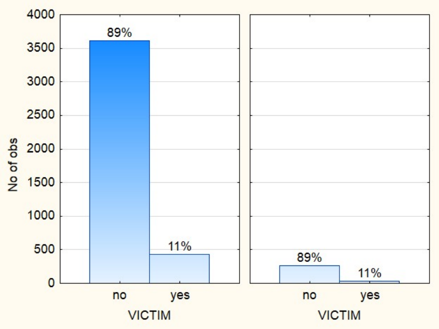


Figure B.232: Victim % for `University Details` (Instagram Non-Subscribers vs. Subscribers).

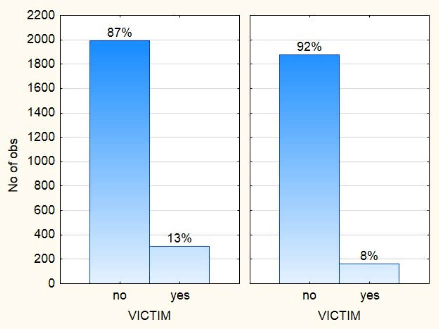


Figure B.233: Victim % for `Job Details` (All vs. None).

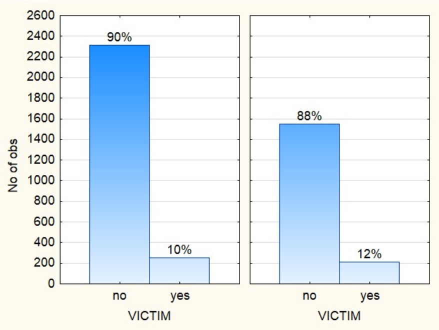


Figure B.234: Victim % for `Job Details` (Facebook Non-Subscribers vs. Subscribers).

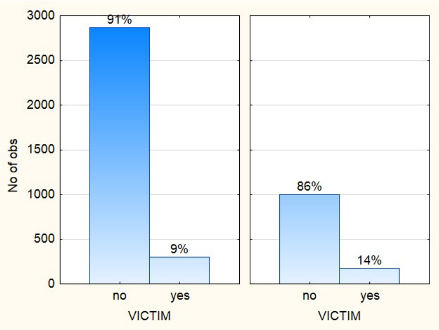


Figure B.235: Victim % for `Job Details` (LinkedIn Non-Subscribers vs. Subscribers).