MSc Thesis Report

A Study to Use a Data Mining Approach to Classify Customer Price Sensitivity Using a Retail Banking Foreign Exchange Historical Dataset

Prepared By

Ntombizodwa Maplanka

Supervisor

Professor David Lubinsky

A research report submitted to the Faculty of Science, University of the Witwatersrand, in partial fulfilment of the requirements for the degree of Master of Science in Mathematical Statistics with the School of Statistics and Actuarial Science, Johannesburg, South Africa

December 10, 2018

Dedication

To my three wonderful children Siphesihle(CJ), Zandisile and Christian Maplanka. Thank you for your patience, love and understanding. I love you endlessly and I promise to make up for the times I have missed out on your activities. It will be worth it in the end.

Acknowledgements

A special thanks to my supervisor Professor David Lubinsky for giving me an opportunity to complete this research under his supervision. His love and passion for R as the software of choice and his love for machine learning motivated me to ask him to be my supervisor; I also have a great interest in R. Many thanks to Professor Jacky Galpin for affording me an opportunity to be part of School of Actual Science and Mathematical Statistics at Wits University after many years of being absent from the formal education system. A big thank you to Dr Charles Chimedza for believing in me. I would also like to thank Sibusiso 'Leadership' Ndzukuma for his wisdom, genius and a significant role he played. I thank Siphiwo Maganda for believing in me and his major contribution. Hermann Donfack I thank you for your guidance. Above all, I thank God for without His grace none of this would have been possible.

Abstract

Data analysis combined with machine learning has become an essential part of the modern scientific methodology, offering automated procedures for the prediction of a phenomenon based on past observations, identifying underlying patterns in data and providing insights about the problem. This thesis seeks to demonstrate the use of data mining techniques to classify price sensitive customers using a historical dataset from a retail banking forex department. Two data sets were merged (customer data and deals data), and statistical models were fitted and compared; namely decision trees, random forests and neural networks. All models produced excellent results when fitted on the datasets; the random forests performed slightly better with marginal improvements over decision trees and neural networks. These models gave the area under the receiver operating characteristic curve of at least 0.90 and percentage correctly classified of least 0.95 for the datasets. Apart from making the most accurate predictions of the response variable random forests and decision trees were used to identify predictor variables that are most important to make these predictions. The study shows that in retail banking under the given setting, the foreign exchange division can price the clients appropriately and increase competitive edge by using data mining techniques to predict customers' price sensitivity to foreign exchange rates. The next step for the bank is to use these methods to retain the customers, increase revenue as well as make improvements in pricing where warranted.

Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence

AUC Area Under The ROC Curve

BP Backpropagation

CART Classification and Regression Trees

CRM Customer Relationship Management

DT Decision Trees

FMCG Fast Moving Customer Goods

FOREX Foreign Exchange

k-NN k-Nearest Neighbours

LR Logistic Regression

MLP Multilayer Perception

NN Neural Networks

PCC Percentage Correctly Classified

RF Random Forest

ROC Receiver Operating Characteristic

SVM Support Vector Machine

Chapter 1

Introduction

1.1 Customer Price Sensitivity

In today's world of intense competition and globalisation, banks are struggling to gain a competitive edge against each other. Customers are becoming more conscious of costs and shop around for cheaper rates. Price sensitivity of clients depends on market conditions and how savvy customers are, amongst other factors. This study considers variables such as volumes traded, customer segment, customer history with the bank, demographic information and other factors to ascertain customer price sensitivity. In this paper the client is regarded price sensitive if he declines the standard system exchange rate and accepts an exchange rate offered by the dealing desk or the client is on special pricing otherwise the customer is considered not price sensitive.

With the current unpredictable FOREX market and currency fluctuations, cus-

tomers are becoming more aware of FOREX market dynamics. The customers have developed an intense bargaining power to sustain their business strategies effectively by comparing exchange rates from various sources as they remain competitive. It is important for banks to understand how price sensitive their customers are so that they focus on strategies such as customer retention to try on customers that are least sensitive. Retaining these customers could result to opportunities such as cross selling. Gathering this knowledge can help banks price competitively and satisfy their clients. The client may try to reduce the margins by negotiating the price as a means of cutting costs. Setting correct strategies and analysing customer price sensitivity is one of the most significant challenges an organisation could face.

1.1.1 Markets' Influence

According to Reuters (2016) [29], the average daily turnover in the global FOREX market rose to USD 4.84 trillion. South African economic outlook paints a bleak picture for the 2016/2017 period as the country's sovereign debt has been downgraded to non-investment grade (Anderson, 2017) [2]. He cautions that the South African Reserve Bank could possibly increase interest rates and suggests that the downgrades may have short to medium term impacts such as a recession, weaker local currency, higher inflation and job losses. Customers are faced with tough economic challenges and therefore will try and save every cent and cut costs as much as possible. Banks are trying to survive in this unfavourable economic climate.

Forward-looking organisations seek to understand their customers individually

and use that understanding to make it easier for customers to do business with them rather than with the competition. It is important for banks to identify which customers are worth investing money and time in and which ones not to put so much effort into. This study focuses on using statistical data mining techniques to predict price sensitivity of customers using confidential information obtained from one of the largest banks in South Africa. The dataset consists of 3 474 123 individual FOREX transactions and 445 935 bank-wide customer demographic records.

1.1.2 Data Mining

The study uses data mining techniques to explore and analyse big data to discover meaningful patterns and classify customers. The goal is to classify price sensitive customers to understand and service customers better. Pulakkazhy and Balan (2013)[27] describe Data Mining and Knowledge Discovery as a concept that combines the fields of statistics, machine learning, database management, information science and visualization. According to the authors, the extracted information can then be used to predict or classify future behaviour, identify trends and associations, refine an existing model, or provide a summary of the database being mined.

Hormozi and Giles (2004) [16] state that data mining is a process that has been widely accepted and used by organisations to enhance performance and gain a competitive advantage. "Banks can use data mining to discover sequences, episodes and periodicity of the transaction behaviour of customers which may lead the bank to better segmenting, targeting, acquiring, retaining and maintaining a profitable cus-

tomer base" (Dass, 2006) [11]. The literature on bank customers' price sensitivity classification is scarce due to information privacy and confidentiality. Evidently, in the FOREX data used for the purpose of this research, there is confidential information that the bank would not share with external entities. However, literature does show other classification problems solved in the banking industry using data mining techniques:

- Customers likely to default on payments or not (Piramuthu, 2004) [26].
- Fraudulent transaction or not (Phua, Alahakoon and Lee, 2004) [25].

It can then be concluded that the highly competitive nature of the banking industry prevents public dissemination of the results of applications in customer pricing. Therefore, the available literature on classification applications using data mining techniques became the framework of this study.

1.2 Problem Statement

Pricing decisions are challenging and create uncertainty for decision makers in organisations. These decisions affect revenue and profitability since pricing influences the level of demand. This research seeks to understand customer price sensitivity by classifying individual FOREX transactions that are likely to be price sensitive using historical FOREX data. Price sensitive transactions are based on negotiated exchange rates given to customers while non price sensitive are not negotiated. The negotiated exchange rates are obtained from dealing desks (sensitive price) and non-

negotiated rates are default rates generated by the system (not sensitive price).

Customers that demand negotiated exchange rates are classified as sensitive while clients who take default rates are categorised as not sensitive. Different classification models were employed and a comparison of their levels of accuracy was made. Customers have many options regarding where they choose to do business. Therefore, it is important for a bank to understand its customers' price sensitivity in order to provide better services to them and be able to retain them. By revealing patterns of consumer behaviour using data mining techniques, the client's price sensitivity can be predicted.

1.3 Aim of Study

This research seeks to demonstrate the use of data mining techniques to classify price sensitive customers using a historical dataset. Successful marketing strategies can result in significant market growth, but unless the impact of this increase is understood, profitability is at risk. Data mining techniques are proving to be beneficial to business regarding identifying patterns in the data and predicting future behaviours of customers (Smith and Gupta, 2000) [31]. FOREX data, in combination with the bank-wide customer and product data were used to predict whether a customer is price sensitive or not price sensitive. Furthermore, the study aims to understand the type of statistical techniques that are most successful in predicting customer price sensitivity in this setting.

1.4 Objectives

The primary objective of this research is to convert FOREX data into meaningful information using data mining techniques. The aim is to develop a predictive model that is capable of classifying the price sensitivity of customers. The steps outlined below were undertaken:

1. Identifying attributes which are highly significant regarding price sensitivity classification.

2. Determining the underlying properties of the data by performing descriptive statistics and frequency analysis.

3. Identifying any anomalies with the data through data exploration.

4. Classifying price sensitivity using decision trees, random forests and neural networks.

5. Evaluating model performance and comparing outputs to determine the best fitting classification model.

1.5 Significance of the Study

The creation of a knowledge base and its utilisation for the benefit of the bank is becoming a strategic tool to compete, apart from the execution of business processes. The banking industry has gone through tremendous changes following the 2007-2009

financial crisis and requires new strategies such as data mining to compete and improve efficiency.

This study examines a dataset that consists of FOREX data in combination with the bank-wide customer data, including demographic and product data. Currently, the bank uses product data only to classify customer pricing using a regression model. Classifying price sensitive clients against those that are not will help banks price and service customers better and be in a better position to gain customer loyalty and retention.

Chapter 2

Literature Review

2.1 Introduction

Data mining techniques have become a key focus in the banking sector. However, due to confidentiality of customer data and the competitive nature of the industry, literature is limited. The first part of this review outlines research done in the data mining classification/prediction approach. The available articles focused on problems such as fraud detection, customer loyalty, credit scoring, churning and Customer Relationship Management (CRM). This study utilises available literature on data mining classification methodology as a framework for the research methodology.

In the existing literature, classification techniques such as DT, RF, NN have been used to classify dichotomous variables, similar to the proposed approach. Yamini

and Ramakrishna (2015) [36] describe classification as a well-known supervised data mining learning technique, which uses a set of pre-classified examples on a model that can easily classify the population of records and they also provide applications of these techniques. They also observed that the classification approach frequently employs DT or NN algorithms.

2.2 Data Mining Techniques Applications in Customer Relationship Management

Ngai et al.(2009) [24] reviewed 87 articles related to data mining techniques in CRM published in academic journals between 2000 and 2006. These reports were reviewed and classified according to four categories of CRM, namely: customer attraction, customer retention, customer development, customer identification/acquisition and seven classes of data mining models. The authors agreed that data mining tools can help organisations discover hidden knowledge in the data. The idea behind the application of data mining techniques in these CRM categories is to create a deeper understanding of customers to maximise customer value to the organisation in the long term.

They argue that elements of customer identification include target analysis which involves seeking profitable segments of clients through analysis of customers' underlying characteristics as well as customer segmentation. Furthermore, the authors employed NN and DT classification techniques aimed at building models to predict

future consumer behaviour through classifying database records into some predefined classes based on certain criteria and standard tools.

Selected articles were reviewed and classified according to the proposed classification framework by three independent researchers. Among the four CRM dimensions, the majority of the reviewed articles related to customer retention (62%; 54 out of 87 items); the most common aspect for which data mining was used to support decision making. Of these, 51% and 44% were related to one-on-one marketing and loyalty programs, respectively. They suggest that these articles could provide insight to organisation policy makers on the standard data mining practices used in retaining customers. Customer identification and customer development appeared on 13 of the 87 articles reviewed.

The authors argue that even though customer identification has fewer articles related to target customer analysis, it does not prevent the application of data mining in other CRM domains if they possess the same goal of analysing the characteristics of customers. The distribution of the data mining techniques showed that out of the 87 articles, NN were the most commonly used (35%), followed by DT (24%). Association rules, regression and genetic algorithm were among other data mining techniques which were applied. This review did not claim to be exhaustive, however, it does provide some insight.

2.3 Use of Data Mining in Banking

This section covers three papers that focus on the application of data mining classification techniques in banking. According to Moin and Ahmed (2012) [23], data mining techniques can help banks target and acquire new customers, detect fraud in real time, provide segment based products and analyse customer purchasing patterns over time for better retention and relationships. Similarly, Bhambri (2011) [4] and Hormozi and Giles (2004) [16] state that data mining can help banks and other financial institutions better target and acquire new customers, detect fraud, analyse customer purchase patterns, detect emerging trends to take a proactive approach in a highly competitive market, adding more value to existing products and services as well as launching of a new product and service bundle.

All three studies agreed that organisations that effectively use data mining techniques have the ability to develop or enhance a competitive edge. However, Bhambri (2011) [4] made a valid point that, even though banks have ever growing databases, the problem is that they fail to take full advantage of the true benefits that can be gained from this important information. All these papers focused on the increase in the ability to generate, capture and store data and the need for transforming such data into knowledge by using data mining techniques.

Moin and Ahmed (2012) [23] argue that data mining can help by finding patterns, associations and correlations which are hidden in the business information stored in

the database. They define data mining as an extraction of knowledge from big data. They further point out that by using data mining to analyse patterns and trends, bank executives can predict with increased accuracy, how customers are likely to accept new product offers, which customers will be high risk for defaulting on a loan and how to make customer relationships more profitable.

The authors note that banks have the richest and largest pool of customer information, covering variables such as customer demographics, transactional data, credit card usage patterns, income and the length of employment. They suggest that banks need to invest resources in understanding existing and prospective customers better, to offer "tailor-made" products and services. According to the authors, data mining can be used in a number of areas. These are discussed in the following sections:

2.3.1 Marketing

In order to determine customer behaviour concerning products, price and distribution channels, practitioners can analyse past trends, identify immediate demand and forecast client behaviour of various products and services to gain more business opportunities and anticipate behavioural patterns. Additionally, data mining techniques can help identify profitable from non-profitable customers. These techniques determine customers' reaction to adjustments in interest rates and the risk profile of a customer segment for defaulting on loans.

Clustering and classification methods are some of the data mining methods that

can be used to find the solution in marketing depending on the problem at hand (Bhambri, 2011) [4]. Clustering is a statistical method that helps target groups that share similar characteristics and the binary classification model distinguishes between "profitable" and "non-profitable".

2.3.2 Risk Management

Bankers want to know whether a customer is reliable or not. There are risky decisions that need to be taken by the banks, such as offering new credit cards, extending existing customer line of credit and approving loans. Data mining techniques help determine which customers are likely to default on loan repayment. Banks need to analyse reliability and behaviour of customers to distinguish prompt or delayed payers if credit cards are sold to clients. By knowing the chances of a customer default, the bank is in a better position to reduce risk. Bhambri (2011) [4] recommends classification techniques as they are most commonly used in data mining, which employs pre-classified examples to develop a model that can classify the population.

2.3.3 Fraud Detection

Data mining techniques will help banks focus on ways to analyse customer data to identify patterns that point to scams. Again, classification is a recommended method for detecting fraudulent transactions. DT and NN are mostly used. Both fraudulent and valid records are included in the learning (training data) and classification (test data).

Jayasree and Balan (2013) define fraud detection as the detection of criminal activities occurring in commercial organizations such as banks, credit card issuing organizations, insurance agencies, mobile companies and stock market. Pulakkazhy and Balan (2013) suggest that early detection of fraudulent transactions can help the banks limit the damages and they also point out that fraud detection is mostly crucial in the areas such as credit card products. According to Jayasree and Balan (2013) the regulators define certain rules for finding the fraud transaction patterns. The authors argue that patterns of customer's transactions can be discovered and alerts can be generated if any measurable deviations are found by employing DM techniques.

2.4 Data Mining Feature Selection for Credit Scoring

In the paper by Liu and Schumann (2005) [21], four data mining classification algorithms namely: model tree M5, k-Nearest Neighbours (k-NN), multi-layer perceptron NN and Logistic Regression (LR) were used to build credit scoring models. This study addresses the challenge of feature selection faced by researchers. Since real world data used might come from different sources and may be collected for a general task, it is likely that some of the features may be redundant due to their high inter-correlation, and some may be irrelevant to credit scoring (Liu and Schumann, 2005) [21].

The authors argue that due to redundant and irrelevant features most classification methods suffer from extensive computational time, a possible decrease in model accuracy and a decrease in scoring interpretation. They provide methods of feature selection mostly used by some scholars and practitioners:

- Univariate analysis to evaluate the effect of each independent feature on the class feature
- Statistical correlation analysis to determine relationships between independent features (multicollinearity).
- Sometimes, feature selection is incorporated into the classification algorithm.

Liu and Schumann (2005) [21] present four machine learning feature selection methods to show the performance in the real-world credit data. The data were obtained from a German credit insurance company. There is a risk that clients may make delayed payments to the supplier for goods and services received; therefore the supplier seeks the expertise of the insurance company to cover the payments of its clients. The granted protection was related to the creditworthiness of the providers' customers. These clients were regarded as risk partners of the insurance company. The dataset in the paper consists of 72 categorical variables. Examples of these variables are:

- Development: company development classified as expansive stagnant or positive or constant or stagnant or declining or strongly declining.
- Reputation: whether the company has a good reputation.
- Property: whether the company has real estate.

- The payment behaviour of debts.

The insurance company classified payment behaviours by ratings given by the current risk assessment system. These ratings are:

- Very good (A): above average good creditworthiness and very low default probability.
- Rather good (B): good or satisfactory creditworthiness and low default probability.
- Problematic (C): problematic and having a high risk of default.
- Very bad (D): very high credit risk and insurance applications must be refused.

Based on the credit experts' knowledge the ratings were assigned and classified into two classes, good (rating A/B) and bad (rating C/D). There were 38 283 cases, 34 562 were classified as 'good' and 3 721 were classified as 'bad'. The authors argued that this data with a low number of 'bad' examples led to inaccurate parameter estimations in the NN, DT and k-NN models. They suggested that a better result can be achieved by using a balanced selection of samples, where the training sample included the same number of 'good' and 'bad' examples by randomly selecting the same number of 'good' examples as 'bad' examples. All 'bad' examples were selected.

The chosen cases were randomly divided into training sample, test sample 1 and test sample 2 (validation sample). The performance of the models before (72 features) and after feature selection was compared regarding the number of used features, speed and accuracy.

- NN features were reduced to 40 features, DT and LR features were reduced to 26 and 33, respectively; while the k-NN models had a minimum of 13 features all selected by 'the consistency-based feature selection algorithm (CON) method.
- The improvements in the model speed were significant for DT, LR and k-NN when using every selection method. There was considerable increase in the speed of the NN model when using the CON method.
- The changes in error rates were small for the tree, LR and NN after feature selection, while a decrease in error in the k-NN was significant.

From the above literature it is shown that there is no improvement in model accuracy for DT, NN and LR after feature selection; maybe this result is owing to the nature of these algorithms. Tree models can select useful features during the building of the tree structure. NN are trained to assign small weights to irrelevant features by learning from the data. LR weighs the used features differently according to predictability. It may therefore not be necessary to reduce features when choosing these models in practice. However, feature reduction can help decrease training time and simplify the final model at a small sacrifice of accuracy.

2.5 Data Mining Feature Selection on Loan Default Data

Piramuthu (2004) [26] argues that recent advances in computing technology regarding speed, cost as well as access to improved computing power and the ability to

process large quantities of data in a reasonably short space of time has resulted in interest in the extraction of useful knowledge from the data using data mining techniques. The study aims to classify a set of firms into those that default on payments and those that would not. The data was extracted from the Index of Corporate Events in the 1973-1975 issues of Disclosure Journal. Sixteen defaulted firms were matched with sixteen non-default firms to obtain data for the study. The dataset was then combined with another sixteen non-default firms, resulting in data of 48 firms from which the subsets were formed randomly using tenfold cross-validation. The data consisted of 18 ratio variables.

The attributes were then reduced to 10 using feature selection methods. Results from using the DT classification technique were provided. The classification accuracy for the training sample and test examples before and after pruning, as well as tree size before and after pruning, were reported. The results showed that by selecting useful features from the dataset through pre-processing, the performance of induced DT could be improved. The author states that selecting useful features translates to a reduction in data gathering costs as well as storage and maintenance costs associated with features that are irrelevant for the decision problem of interest. This paper does not explore the use of other classification techniques such as NN to see how they would have performed in comparison to DT. It would have been interesting to see the impact of the feature reduction on other classification techniques.

2.6 Partial defection of behaviourally loyal clients in a not- contractual FMCG retail setting

This article was reviewed in the interest of observing the contribution of demographics where the predicting customer loyalty is concern. Buckinx and Van den Poel (2005) [6] build models using LR, NN and RF, to predict partial defection by behaviourally loyal clients. They use Percentage Correctly Classified (PCC) and Area Under The ROC Curve (AUC) to evaluate the classifier performance on the test sample. Partial defect aims to determine which behaviourally loyal clients may partially switch their purchases to another store at some point in time. As a result, binary classification models were employed to determine if a particular customer is loyal or a partial defector. The authors focused on partial attrition of high-frequency shoppers who exhibit a regular visit pattern hoping to overcome the challenge of not being able to identify total defection in non-contractual settings.

Data was provided by one of the largest retailers in the Fast Moving Customer Goods (FMCG) services. It consisted of 61 predictor variables made up of behavioural information at the level of the individual customer and customer demographics. The purchasing occasions could be traced using loyalty card, and these cards registered at least 85% purchases at this retailer. Personal records of 158 884 customers, from April 2000 to January 2001, were used as a random sample from the primary dataset. The first five months were used to define retailer's clients that were regarded as loyal. The number of customers selected was 32 371. The aver-

age spending of a customer was EUR 1 417 per annum, whereas behaviourally loyal customers spent almost twice as much. This group was randomly separated into a training set (16 079) and test set (16 292).

Models were fitted to determine whether the customers defected in the period from September to January. When comparing the different classification methods used, they all offered similar performance regarding AUC as well as PCC. RF had a PCC performance of 0.8040 on a test set and an AUC performance of 0.8310. NN and LR had a PCC performance of 0.8040 and 0.8013; an AUC performance of 0.8310 and 0.8220, respectively. It was further concluded that behavioural variables are better in separating behaviourally loyal customers from those that are likely to defect than customer demographics, which is inconsistent with some existing research that puts great emphasis on demographic variables. A model that excluded the demographics resulted in an AUC of 0.8224 as compared to 0.8319 for RF. NN and LR results were not reported.

This section suggests that decision trees, random forests and neural network methods are useful for a binary classification. These techniques use classification rule to classify the elements of a given dataset set into two groups. This study illustrates the use these methods in a classification problem encountered in the banking environment to classify whether the client is sensitive to exchange rates pricing or not.

Chapter 3

Statistical Theory

3.1 Modelling

Building models is fundamental to understanding the data mining world. This research employs classification methods to extract knowledge from the historical data and presents it in such a way that the resulting model can be applied to new cases. Firstly, data was split into a training set to build the model, validation set to fine tune the model and a test set to test for accuracy. Binary classification techniques such as decision trees, random forests were employed to categorise data into two classes.

3.1.1 Decision Trees

DT, also referred to as Classification and Regression Trees (CART), have been widely used in data mining due to the simplicity of the resulting model. DT are easy to

view, understand and explain when presenting output. The data are partitioned recursively; within each partition, a simple predictive model is fitted. CART is a very powerful analytic technique that helps determine variables with the most predictive power and can help researchers produce a powerful explanatory model.

These methods construct prediction models that offer a non-parametric framework for uncovering non-linear and interactive structures in the data. Classification trees allow for the possibility that different relationships may exist between predictors at various times under different cross-sectional conditions. DT can handle high dimensional data (Zekić-Sušac et al., 2014)[38]

Classification Tree Construction

A training set of labelled examples is given, and the goal is to build a model that can be used to predict unseen examples. The tree is split in such a way that the terminal nodes contain observations of the same class as much as possible. The construction of classification trees and the gini split rule section described below follows (Timofeev, 2004) [33] almost exactly.

Let n_l and n_r be left and right child nodes of parent node n_p, respectively. Consider the learning sample matrix X_{nm} with M variables and N observations. Let the class vector Y_{n1} consist of N observations with a total of T classes. A classification tree is built in accordance with a splitting rule - the rule that performs the splitting of the learning sample into smaller parts. Each time data have to be split into two sections with maximum homogeneity;

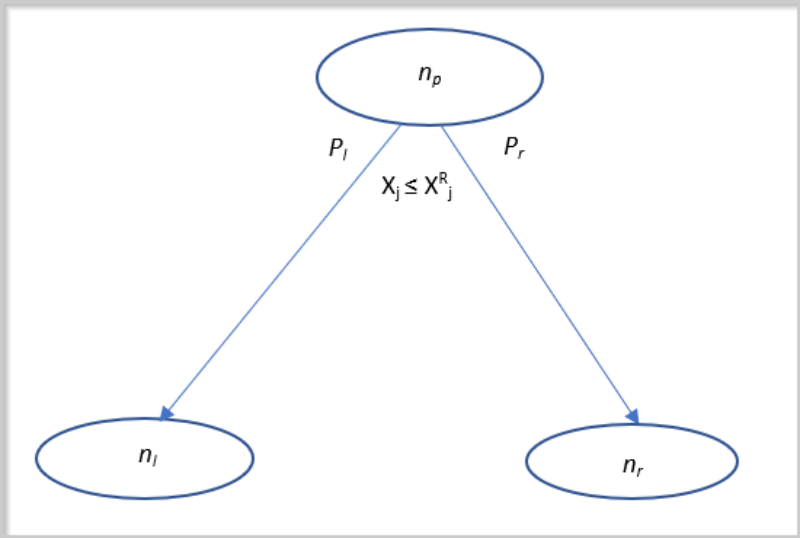

Figure 3.1: Splitting algorithm of CART

where x_j^R is the best splitting value of variable x_j. The minimum homogeneity of child nodes is defined by the so-called impurity function $i(n)$. Since the impurity of parent node n_p is constant for any of the possible splits $x_j \leq x_j^R, j = 1, \cdots, M$, the maximum homogeneity of left and right child nodes will be equivalent to the maximisation of change of impurity function $\Delta i(n)$:

$$\Delta i(n) = i(n_p) - E[i(n_c)] \tag{3.1}$$

n_c means left and right child nodes of the parent node n_p. Assuming that the P_l, P_r - are the probabilities of left and right nodes respectively, we get:

$$\Delta i(n) = i(n_p) - P_l i(n_l) - P_r i(n_r) \tag{3.2}$$

Therefore, at each node CART solves the following maximisation problem:

$$\underset{x_j \leq x_j^R, j=1,\cdots,M}{\arg\ max} \left[i(n_p) - P_l i(n_l) - P_r i(n_r) \right] \tag{3.3}$$

Equation 3.3 implies that CART will search through all possible values of all variables in matrix X for the best split question $x_j < x_j^R$ which will maximise the change of impurity measure. The approach examines multiple possible independent variable splits and then chooses the splits with the highest separation ratio between classes. Additionally, the analysis puts focus on the most commonly used function called Gini splitting rule. The rule is mainly focused on minimising the loss function related to the cost that tends to rise in situations where the split is deviating from a perfect split(this is where the best split results in the partitioning of data into the desired classification).

Gini Split rule

This rule is also known as the Gini index (Timofeev, 2004) [33]. It is the most commonly used amongst similar rules. The below function is applied in this rule and it seeks to minimise:

$$i(n) = \sum_{t \neq l} p(t|n)p(l|n) \tag{3.4}$$

here, $t, l\ 1, \cdots, T$ -represents the index of the class; $p(t|n)$ - the conditional probability of class t provided we are in node n. Applying the Gini impurity Equation 3.4 to maximisation problem Equation 3.3 we will get the following change of impurity measure:

$$\Delta i(n) = -\sum_{t=1}^{T} p^2(t|n_p) + P_l \sum_{t=1}^{T} p^2(t|n_l) + P_r \sum_{t=1}^{T} p^2(t|n_r) \tag{3.5}$$

The algorithm above seeks to address the problem below:

$$\underset{x_j \leq x_j^R, j=1,\cdots,M}{\arg\ max} \left[\sum_{t=}^{T} p^2(t|n_p) + P_l \sum_{t=1}^{T} p^2(t|n_l) + P_r \sum_{t=1}^{T} p^2(t|n_r) \right] \tag{3.6}$$

The algorithm will search in the learning sample for the largest class and isolate it from the rest of the data. This impurity function reaches a minimum when the terminal nodes contain only one of the two classes of observations; price sensitive or non price sensitive. The Gini split rule works well for noisy data.

CART steps

CART analysis consists of four basic steps (Lewis, 2000) [19]; this section follows the steps outlined by the author. The first step consists of tree building, during which a tree is built using recursive splitting of nodes. Each resulting node is assigned a predicted class, based on the distribution of classes in the learning dataset which would occur in that node and the decision cost matrix. The assignment of a predicted class to each node occurs whether or not that node is subsequently split into child nodes. The second step consists of stopping the tree building process. At this point a "maximal" tree has been produced, which may greatly overfits the information contained within the learning dataset.

The third step consists of tree "pruning", which results in the creation of a sequence of simpler and simpler trees, through the cutting off increasingly important nodes. The fourth step consists of optimal tree selection, during which the tree which fits the information in the learning dataset, but does not overfit the information, is selected from among the sequence of pruned trees. Termination does not imply that it is always possible to obtain entirely homogeneous final nodes. The best tree in terms of goodness of fit is achieved by minimising the misclassification rate while penalising larger trees with more nodes.

Advantages and disadvantages

Williams (2011) [35] highlights the benefits of using CART as follows:

1. Classification trees are non-parametric, meaning they make no assumption about the underlying distribution of values of the independent variables. Therefore, CART can handle numerical data that are skewed, or bimodal as well categorical predictors. This feature is very critical because it will save time as there will be no need to determine if variables are normally distributed and the need to transform data if not normally distributed.

2. CART can search all possible variable splitters, even in problems with hundreds of predictors.

3. Cases with missing predictor values are not dropped from the analysis; surrogate variables containing information similar to that provided in the primary splitter are used instead.

4. CART analysis is relatively automatic machine learning, unlike multivariate modelling techniques where a number of inputs are required.

5. CART trees are relatively easy to interpret.

6. The CART algorithm can identify variables that are most significant and eliminate non-significant ones, and it can handle outliers (Timofeev, 2004) [33].

The classification trees approach suffers from some shortcomings despite its simplicity. RF algorithm, an extension of CART, are used to overcome some of the weaknesses of the CART approach.Yohannes and Hoddinott (1999) [37] and Joy et al.(2015) [18] highlight the shortcomings of using CART as follows:

1. A critical weakness with CART is that it is not based on a probabilistic model. Therefore, confidence intervals for threshold values cannot be determined.

2. Since classification trees attach a single probability to all cases belonging to the same class, the marginal contribution of each variable to the probability of observing an episode of either price sensitive or not price sensitive cannot be determined.

3. Recursive partitioning procedure draws, with each step, less information from the full sample.

4. Trees can be prone to over-fitting if measures are not taken to avoid the problem.

5. Each subsequent split depends on the previous one, so an error in a higher split is filtered down.

3.1.2 Random Forest

Despite its many strengths, the classification tree approach can result in overfitting if employed without due care. That is to say this method can create overly complex trees that do not generalise well from the training data. The RF algorithm is used to avoid this situation. RF is used to address two main classes of problems: to construct a prediction rule in a supervised learning problem and to assess and rank variables on their predictive power of a response variable (Boulesteix et al., 2012) [5]. RF is a powerful statistical classifier that has very high classification accuracy, a method of determining variable importance, an ability to model complex interactions amongst predictor variables, flexibility to perform several types of statistical data analysis and an algorithm for imputing missing values (Cutler et al., 2007) [10].

Firstly, rather than run the classification tree algorithm on the entire space of potential candidate variables, it is run only on those variables found to be most important by an implementation of the RF algorithm (Breiman et al., 1984)[7]. This algorithm grows many trees instead of one tree and introduces some randomness so that each tree is most likely different from the others. RF do not overfit, even though each tree is probably overfitting since the components are all different. When errors are averaged they cancel each other. There is no need for cross-validation. Around two-thirds of the data is selected in each sub-sample. The remaining data (called out-of-bag data) is preserved to establish the variable importance and out-of-sample error rates. Each iteration issues a vote on variable importance and majority votes

are used to yield the final, pooled, variable-importance rankings. For classification, the trees are combined using majority voting with one vote per tree over all the trees in the forest.

Siroky (2009) [30] defines RF as a classifier consisting of a collection of tree-structured classifiers: $h(\boldsymbol{x}, \theta_k, k = 1, \cdots)$, where the θ_k are independent and identically distributed random vectors and each tree casts a unit vote for the most popular class at input $\boldsymbol{x}$. Hastie et al. (2001) [15] then present the algorithm as follows:

1. For $B_i, i = 1, \cdots, B$
   - Draw a bootstrap sample S of size $\hat{N}$ from the training data
   - Grow an unpruned RF tree, T_b using the bootstrapped data, until the minimum terminal node size, n_{min}, is obtained by recursively following the sub-algorithm.
2. Randomly select $\hat{p}$ variables from the total set of $\hat{P}$ variables. Select the optimal variable/split-point among the $\hat{p}$ variables. Split node into two daughter nodes.
3. Output ensemble of trees $T_{b_1}^B$
4. Predict new observations, or out-of-bag observations
   For regression: $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$
   For classification: Let $\hat{C}_b(x)$ be the class prediction of the b^{th} random forest tree $\rightarrow$ $\hat{C}_{rf}^B(x) = \text{majority vote}\left\{\hat{C}_b(x)\right\}_1^B$

Advantages and Disadvantages

RF offer two key advantages in this context:

1. insensitivity to outliers and

2. avoidance of overfitting.

Variable-importance rankings are therefore more robust than those issued by a single classification tree. Through construction, RF produce an unbiased estimate of the test set error internally by constructing many bootstrap samples from the original data and leaving about one-third of the cases out of the bootstrap sample and the construction of the n^{th} tree in the forest. Although RF can improve accuracy, this method can be very demanding on computer resources.

3.1.3 Neural Networks

Cheng and Titterington (1994) [9] define NN as mathematical models represented by a collection of simple computational units interlinked by a system of connections. This method of data mining classification is inspired by the human brain; the aim is to require the construction of a system that could compute, learn, remember and optimise. The processing tasks in the brains are distributed among nerve cells called neurons.

Each neuron is connected to many others and can be activated by inputs from elsewhere and can stimulate other neurons. NN are used for applications of pattern classifications and pattern recognitions. The reason behind using this approach is that it can fit the data where linear and other models have proved inadequate. Another excellent quality of NN is that they have an ability to learn and to make

adjustments to their structures over time.

Neural Networks Construction

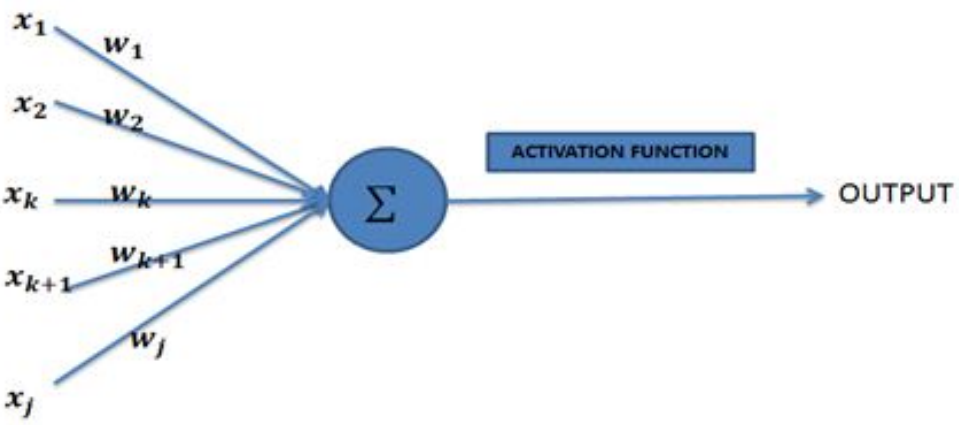

Figure 3.2: A typical single neuron

Figure 3.2 shows a typical single neuron where x_i for $i = 1, \cdots, j$ are input vectors, w_i are input vector weights, and $\sum = \sum_{i=1}^{j} w_i x_i$ is the sum of each weight times the input vector. Figure 3.2 is the simplest form of a NN called a perceptron, a model of a single neuron. It consists of a "feed forward" model, meaning inputs are sent into the neuron, are processed and result in an output. Each input that is sent into the neuron must be weighted. The output is generated by passing the sum of the weighted inputs through an activation function. The method called supervised learning is used to train the NN. NN is provided with inputs for which there is a known answer. Cheng and Titterington (1994) [9] generalise input and output relationship at a neuron as: $y = f(\phi(w, x))$ where f and ϕ are prescribed functional forms, x are inputs, w are the connecting weights associated with connections leading

into the unit, f is the activation function which is chosen from the small selection of functions, including the following:

- $f(v) = sgn(v) = f_h(v)$, the hard limiter is non-linear and yields dichotomous (+-1) output.
- $f(v) = (1 + e^{-v})^{-1} = f_s(v)$, the sigmoidal (logistic), produces output between zero and one.
- $f(v) = v_+$, produces non negative output.
- $f(v) = v$, is linear
- $tanh(k) = \frac{e^k - e^{-k}}{e^k + e^{-k}}$, hyperbolic tangent

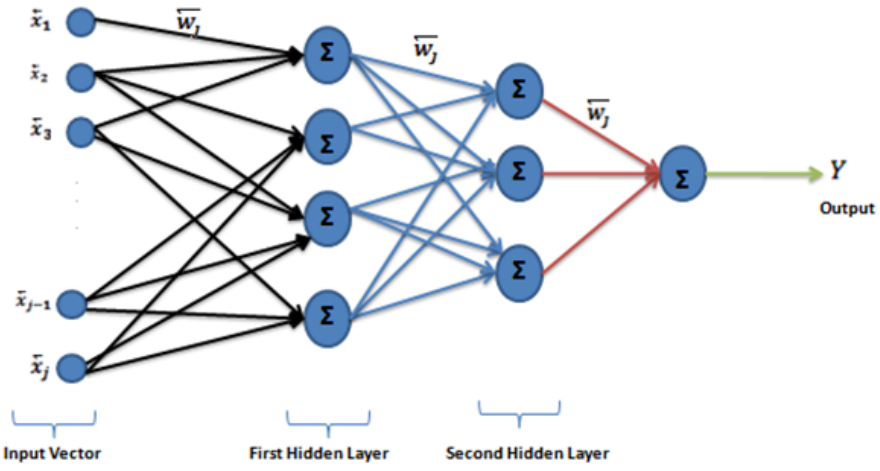

Figure 3.3: Feed forward network with two hidden layers

Figure 3.3 shows feed forward NN architecture, a two layered perceptron with many neurons. The $w's$ are the weights of each neuron; Y is the output, and the $x's$

are the input vectors. The back-propagation algorithm is used to optimise weights w_i. When fitting the NN model we try to find the unknown weights w_i by minimising the error of the output from the estimated weights. "Learning the class involves finding a specific set of weights, based on training examples, which yields a predictor that has good performance on future examples"(Livni et al.,2014) [22]. The starting values of the weights to be estimated and overfitting could pose a problem when fitting the NN. A zero value can be used as a starting point of estimating the weights and an early stopping rule can be applied to overcome overfitting.

Advantages and Disadvantages

Tu (1996) [34] outlines the benefits of using artificial NN as:

1. The ability to implicitly detect complex nonlinear relationships between dependent and independent variables.
2. Ability to detect all possible interactions between independent variables.
3. The availability of multiple training algorithms.

He then provides the following disadvantages:

1. "Black box" nature, that is to say, it is difficult for a system to provide a suitable explanation how it arrived at an answer.
2. NN are a greater computational burden.
3. They are also prone to overfitting.

Regularisation and early stopping are two methods for improving generalisation that are implemented in NN to avoid overfitting. According Livni et al. (2014) [22], with the sufficiently over-specified NN, the computational hardness of learning does decrease. They suggest that changing the activation function can also address the computational burden.

3.2 Model Evaluation

This study focuses on confusion matrix and the Receiver Operating Characteristic (ROC) because these two methods are commonly used. The model was used to predict the class for each observation. The predicted class was then compared with the actual class.

3.2.1 Confusion Matrix

A confusion matrix is appropriate when predicting a binary target. In this evaluation method, the predictions are compared to the actual outcomes; true and false positives and negatives are then calculated. One of the benefits of using this performance evaluation tool is that it is easy to detect if the model is confusing two classes. The matrix also shows the accuracy of the classifier as the percentage of correctly classified patterns in a given class divided by the total number of patterns in that class. However, when the number of samples in one class is significantly more than that in the other class, the accuracy evaluated of a classifier is not representative of the true performance of the classifier. It is, therefore necessary to use other evaluation

models apart from the confusion matrix to evaluate the model performance(Fawcett, 2006) [13].

	Predicted (P')	Predicted (N')	
Actual (P)	True Positive (TP)	False Negative(FN)	TP/(TP+FN)
Actual (N)	False Positive (FP)	True Negative (TN)	TN/(FP+TN)
	TP/(TP+FP)	TN/(FN+TN)	(TP+TN)/(TP+FN+FP+TN)

Table 3.1: Confusion matrix in general

Where,

- Accuracy = $\frac{TP+TN}{TP+FN+FP+TN}$: the proportion of the total number of predictions that were correct.
- Positive Predictive Value = $\frac{TP}{TP+FP}$ or Precision.
- Negative Predictive Value = $\frac{TN}{TN+FN}$
- Sensitivity = $\frac{TP}{TP+FN}$ or Recall. An equivalent of True Positive Rate.
- Specificity = $\frac{TN}{TN+FP}$.
- False Positive Rate = $\frac{FP}{FP+TN}$.

3.2.2 Receiver Operating Characteristic Curve

A ROC curve is a graphical method that is used for visualising, organising and selecting classifiers based on their performance (Fawcett, 2006) [13]. According to the

author ROC curves have properties that make them useful for domains with skewed class distribution and unequal classification error costs. The given dataset summaries indicate the presence of the unbalanced classes. Given that the objective of the study is to predict a class (price sensitive or not) and considering that the dataset was split into training and test sample, the model is built on the training set and evaluated on the test set. The ROC curve is a plot of the True Positive Rate (price sensitive cases that were correctly classified) on the y-axis and False Positive Rate (not price sensitive classified as price sensitive) on the x-axis. A ROC plot presents relative tradeoffs between (True Positives) correctly predicted cases as price sensitive and (False Positives) incorrectly predicted cases as price sensitive.

Figure 3.4 illustrates the ROC curve, where the 45-degree line (y = x which is labelled D) corresponds to random chance (not a good model), curve C show that the model is performing better, curve B is a better fitting test and curve A is the perfect model. The higher the area under the ROC curve the better is the performance of the model. As the accuracy improves, the ROC curve moves towards A and the AUC approaches 1. The AUC is a measure of classifier performance and is an element of the set [0;1] (Hanley and McNeil, 1982) [14].

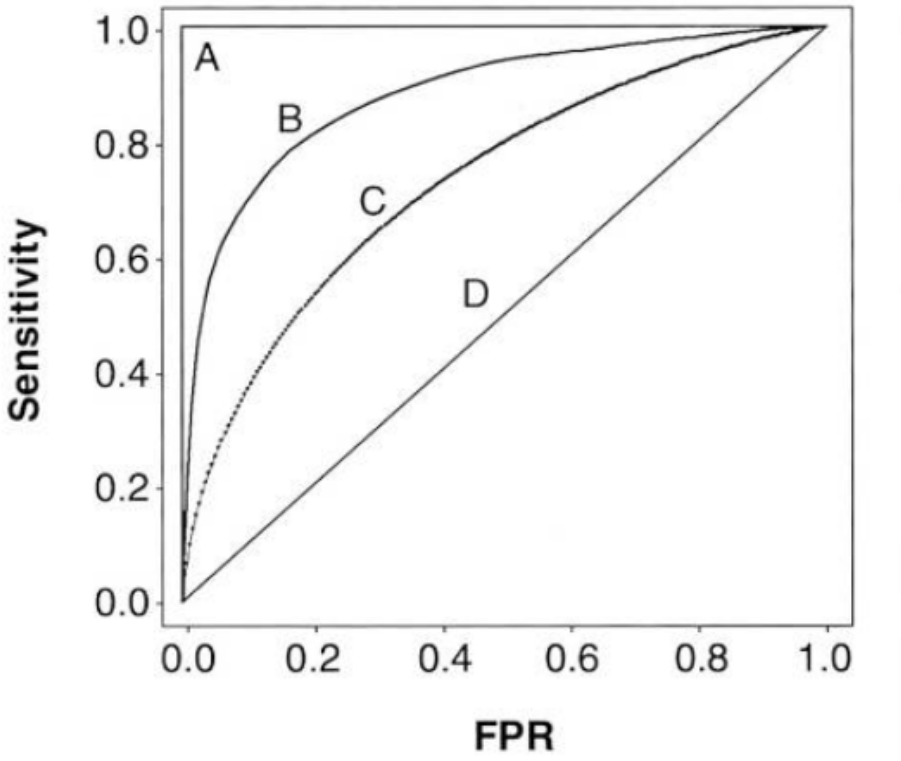

Figure 3.4: ROC Curve

Chapter 4

Methodology

4.1 Analysis Process

This chapter explains the processes, procedures and techniques that are involved in analysing the data. The data used in this study comprises of two datasets from a FOREX customer dealing department. One set consists of bank-wide customer information of 445 935 customers including demographics, and a deal dataset storing 1 048 575 (after removing unwanted data) individual dealing transaction records, recorded from 2011 to 2015. The two sets were merged together using a customer identifier to create a single set. The binary variable *sensitivity*, is a dependent variable considered sensitive if a customer receives a special pricing and not sensitive if a customer accepts standard system pricing.

4.1.1 Class variable

In FOREX dealing a customer is deemed price sensitive if he receives exchange rate from the dealing desks (coded as "Y" and "O"), where "Y" is a large dealing desk that handles clients of large business size and "O" is a small dealing desk that handles clients of small to medium business size. The client on specialised pricing (coded as "I") loaded in the system is also considered price sensitive and not price sensitive if he receives standard system (coded as "N") exchange rates. "Y", "O" and "I" were then assigned the number 1, while "N" was assigned the number 0. If a customer declines the standard system exchange rates, he then gets quoted from the dealing desks for R50 000 upwards. Given this scenario and the nature of retail banking, the client is always going to be given a quote that is rarely rejected.

4.2 Understanding the data sets

4.2.1 Deals data

The initial deals data set consists of 30 variables and had missing values on the variable product. Each record represents details of a foreign exchange transaction. The class variable *sensitivity* is a binary variable (0 and 1) and was converted to a nominal binary variable (not sensitive and sensitive). Some of the variables are not interesting for this paper: namely *agreement date*, *deal number*, *dealer code*, *date*, *year*, *month*, *zfx account number*, *customer sub-segment*, *dealer name* and *application*

number and were subsequently removed. Table A.1 and Table A.2 in the appendix show the summaries of variables contained in the deals data. The distributions of these variables are presented in section A.2.1 in the Appendix section.

The analysis was based on the data set that consists of dealt currencies USD, GBP, EUR as these base currencies contribute 85% to the data, most of the deal transactions were based on these three main currencies, see Table4.1 below.

No.	Dealt Currency	Frequency	Percentage (%)	Cumulative %
1	United States Dollar (USD)	585479	55.84	55.84
2	European Currency (EUR)	201502	19.22	75.06
3	Great British Pound (GBP)	108562	10.35	85.41
4	South African Rand (ZAR)	69347	6.61	92.02
5	Australian Dollar (AUD)	22872	2.18	94.2
6	Botswana Pula (BWP)	12584	1.2	95.4
7	Canadian Dollar (CAD)	8863	0.85	96.25
8	Lesotho Loti (LSL)	8538	0.81	97.06
9	Swiss Francs (CHF)	7899	0.75	97.81
10	New Zealand Dollar (NZD)	4178	0.4	98.21

Table 4.1: Frequency Analysis of Dealt Currency

Variable Description

Table A.3 in the Appendix describes the variables contained in the deals data. It shows the type and description of each variable and the variables are defined below:

No.	Variable	Definition
1	*customer number*	The unique number used to identify each customer. It was used to merge the deals and customer data.
2	*Sensitivity*	The dichotomous response variable with levels sensitive and not sensitive. A customer is described as sensitive if he receives the exchange rate from the dealing desk or is on special pricing and described as not sensitive if he receives a default system rates.
3	*Deal Type*	Describes whether the transaction is a purchase or a sale transaction.
4	*Dealt currency*	The base currency that is bought or sold. e.g. USDZAR (USD is a base)
5	*Dealt amount*	The value of the dealt currency. e.g. USD 10 000
6	*Counter currency*	The currency that is traded against the dealt currency.
7	*Counter amount*	The value of the counter currency
8	*Turnover*	The absolute value of the counter currency. This variable is a Rand equivalent of the dealt currency amount, regardless of the deal type.
9	*Agreement type*	A four-level categorical variable. It describes whether the agreement is a partial optional forward contract, a spot deal, a fixed forward contract or an optional forward contract.

10	*Spot exchange rate*	The foreign exchange rate agreed today for settlement in two business days, i.e. USDZAR, EURZAR, GBPZAR.
11	*effective exchange rate*	The rate at which the deal is concluded. Just as the spot exchange rate, the effective exchange rate is based on the USDZAR spot rate.
12	*Deal revenue*	A product of a turnover and margin; therefore it assumes a positive value.
13	*Customer segment*	a categorical variable consisting of consumer, wealth, business banking and commercial clients. The consumer and wealth segments are individual customers while business banking and commercial are organisations/companies.
14	*Dealing desk*	a categorical variable with four levels. Sensitive customers receive pricing from the large dealing desk (Y), small dealing desk (O) and individualised/special pricing (I). None sensitive customers receive pricing from the standard system pricing (N).
15	*Forward exchange contract* (FEC)	A contract with an agreed amount at an agreed exchange rate for an agreed future date.If the deal is not a forward agreement, then it is a spot contract.
16	*USD spot rate*	The prevailing USDZAR spot rate at the time the trade is executed.

17	*Range*	A categorical variable with fourteen levels. These levels group transaction amounts according to specific values. Various amounts such as deals that fall within the R0 – R9 999.99 range are said to be in range 1; contracts that fall within the R10 000 – R19 999.99 range are said to be in range 2; transactions that fall within the R20 000 – R49 999.99 range are said to be in range 3. All deals worth R4 000 000 and above are in range 14.
18	*Margin*	Represents a margin taken on a USD spot rate. If a deal is a purchase, then the margin is subtracted from the price, if the agreement is a sale the margin is added. An absolute value of the margin is considered as the negative sign only implies a purchase deal.
19	*Product*	A categorical variable that consists of different foreign exchange products offered by the bank,such as cash passport, notes, swifts.

Table 4.2: Definition of variables in the deals data

4.2.2 Customer data

The customer data set is made up of customer demographics and bank-wide customer information. It has 445937 records and 26 variables. Some records had missing values namely: customer *age*, the *highest level of education*, *Transactional (DDA) Account Balance*, *Investment Account (TDA) Balance* and *Credit Card Balance*. Table A.4 in the appendix shows the summaries of variables contained in the customer data. The

distributions of these variables are presented in section A.2.2 in the Appendix section.

Variable Description

Table A.5 in the Appendix presents the type and the description of each variable in the customer data and the explanation of each variable is given below:

No.	Variable Name	Definition
1	*customer type*	is a categorical variable that consists of private individuals, public companies, private companies as well as non-profit organisations.
2	*Branch Code*	variable is an identifier for the branch where customers do their banking.
3	*age*	represents individual customer age in years or the number of years the company is in existence.
4	*Highest education level*	shows different levels of customers' qualifications.
5	*Customer segment*	consists of individual customers and companies. This variable was excluded as it is in the deals data as well; these two sets were merged.
6	*Income*	is the variable that consists of income received by customers via the bank.

7	*Bank VIP*	is a two-level categorical variable that indicates the level of customer importance across retail banking.
8	*Years banking with the bank*	indicates the number of years the client has been banking with the bank.
9	*online indicator*	indicates the exchange rates levels online to clients.
10	*Salary with the bank*	indicates whether a customer receives their salary via the bank.
11	*Points incentive indicator*	is a reward scheme for using the bank's products.
12	*DDA account*	is the number of transactional accounts the client holds with the bank.
13	*DDA balance*	is the balance in the transactional account; this is the most commonly used account foreign exchange settlements.
14	*TDA account*	is the number of investment accounts a customer holds.
15	*TDA balance*	is the balance in the investment account. This account is seldom used in the foreign exchange settlements.
16	*Card account*	is a number of credit cards the customer holds with the bank.
17	*Card balance*	is a balance on the credit card account. This account is seldom used in foreign exchange settlements.
18	*Global account indicator*	indicates whether the client has a global account or not.
19	*Banking with the bank*	indicates whether the customer is banked with the bank the analysis is based on or not.

20	*Risk category code*	indicates the client's risk profile.
21	*FX VIP*	shows whether the customer is classified as a VIP or not in the foreign exchange dealing department. VIP clients get special rates.
22	*pre-dealing indicator*	shows if a client can book deals before submitting documents or not. Clients that have pre-dealing facility are most likely on special pricing.
23	*Deal per annum*	is an average number of deals done a by a client per annum.
24	*Turnover per annum*	is the total business or customer turnover per annum.

Table 4.3: Definition of variables in the customer data

4.2.3 Merged dataset

After cleaning, Deals and Customer data sets were merged using a universal identifier variable *customer number*. Cramer's V test was used to measure the association between two categorical variables. The test procedure is based on a Cramer's V contingency table analysis, which can test the strength of the association between the response variable and the explanatory variables. Variables with a Cramer's V value of about [0.3 to 0.5) are considered to have a strong association while those with values of 0.5 or higher are considered to have a stronger association. Correlation was used to measure the strength of the relationship between numerical variables. When the correlation is close to 1, this means that there is a strong relationship between two variables. Decision trees, random forests and neural networks were fitted to the

merged set.

Firstly, 50% of this data set was randomly sampled and was split into training and test set. The model was built on the training set and tested on the test set. The remaining 50% of the merged data was treated as a validation set. The split that was used when comparing the models was 70 percent training and 30 percent test set.

4.3 Machine Learning methods used

4.3.1 Decision Trees

CART model was fitted to the data set using the R rpart package; this package has an advantage in that it can use surrogate variables when it encounters an 'NA' which is a missing value. Rpart uses Gini impurity to select splits when performing classification. The decision tree models were grown on the merged data using all variables and without some variables to predict the response attribute. This approach was used to check whether excluding certain variables would have a significant impact on the model results.

4.3.2 Sampling and filling in missing data for RF and NN

RF and NN were also built on the merged data with all variables and without some variables to predict the dependent attribute. This approach resulted in four different data sets that were used for modelling. This method was done to check the impact

the exclusion of some variables would have on the model results.

- Dataset 1: consists of all variables selected by the decision trees to have the most predictive power.

- Dataset 2: excludes all variables with missing values,*tda balance*,*dda balance*,*card balance*, six records of variable *age* and *product*. These variables were excluded because the employed techniques cannot handle missing values and the information on them is not available. There were many missing values in the customer data for variables *tda balance* (80%),*card balance* (64%),*dda balance* (17%). *Age* variable had a fewer missing values in the records (0.0013%), *education level* (13%), these cases were also be excluded. Deals data consist of missing values in *product* variable (30%). Variables *branch code* and *education level* were left out as it does not seem they have much impact on predicting price sensitivity in the given setting.

- Dataset 3: excludes all variables with missing values,*tda balance*,*dda balance*,*card balance*, six records of*age* variable; excludes variables *branch code*, *education level* and includes imputed *product*. Missing values in the variable *product* were imputed using the missForest algorithm (see paragraph B.4.1 in the Appendix section). Tang and Ishwaran (2017)[32] suggest that RF imputation is robust with performance improving with increasing correlation. They discovered that the performance was good under moderate to high missingness,

and even (in some instances) when data was missing not at random. For each variable missForest fits a random forest on the observed part and then predicts the missing part. After three iterations using ten trees, missForest produced the proportion of falsely classified entries in the variable *product* of 0.0165. A good performance of missForest leads to a value close to 0 and bad performance to a value around 1; therefore this prediction is quite good. The imputed *product* consists of 316585 Notes, 238755 Inward swifts, 148597 Outward swifts, 75804 foreign cheques, 55349 CFC, 45846 Cash Passport and 12263 other products (see Appendix Table B.9). Variable *product* with missing values consists of 287079 Notes, 185790 Inward swift, 65659 Outward swift, 19931 foreign cheques, 269155 NAs, 36971 Cash Passport and 28640 other products.

- Dataset 4: excludes all variables with missing values, *branch code*,*education level*,*tda balance*,*dda balance*,*card balance*, six records of *age* variable and includes the imputed *product*.*Dda balance* variable was imputed using the decision tree; a decision tree is now used to fill in those missing values. The R function is.na () and it's reciprocal!is.na() are used (the bang symbol represents "not"). This subsets on whether a value is missing or not. The method="anova" version of the decision tree was used to predict a continuous variable. Grow a tree on the subset of the data with the *dda balance* values available; then substitute the missing data. After inspecting the summary there were no NA values, the average *dda balance* change from R120 609 to R121 100, the median before imputing was R7 541 and changed to R14 655 (see Appendix Table B.10)

Random Forests

Unlike the rpart package for a CART, the R package randomForest cannot handle missing values. RF also uses Gini Index based impurity measures for building decision trees. Decision trees were built on each of the four data sets (Dataset 1, Dataset 2, Dataset 3, Dataset 4) using the RF algorithm based learning. The error rates were plotted across decision trees.The Gini measures how pure the nodes are at the end of the tree.

Neural Networks

The two main kinds of ANN learning algorithms are supervised and unsupervised. In supervised learning, the correct results (target values, desired outputs) are known and are given to the ANN during training so that the ANN can adjust its weights to try matching its outputs to the target values. After training, the ANN is tested by giving it only input values, not target values, and seeing how close it comes to outputting the correct target values. This research utilises a supervised learning algorithm. The model was fitted in the four described training datasets (see paragraph 5.3) using R Neuralnet package. NN requires an all numeric input data frame; therefore, the variables were transformed. The NN use activation functions between -1 and +1; it is essential to scale the variables down so that the networks do not spend time training iterations doing the scaling.

The min-max method was used to scale the numeric variables, and the factor variables were converted into dummy variables using model matrix function. The model was built to train networks in the context of classification analysis. Resilient back-propagation with one hidden layer consisting of one node was used. Since the response variable is binary, the logistic activation function was chosen and the error function used is cross-entropy. Additionally, the item linear output was stated as false to ensure that the output is mapped by the activation function to the interval [0, 1]. The activation function transforms aggregated input signals into output, and the error function describes the deviation of the predicted outcomes from the observed ones. The same architecture model described was fitted in Dataset 1, Dataset 2, Dataset 3 and Dataset 4 to evaluate if the model performs better on the data set with and without some variables. This approach was taken to see if some of the variables contain the most information about the data than others.

Chapter 5

Analysis and Results

All training, validation and testing of the CART, RF and NN were conducted using R software. The input feature space was constructed by importing the input variables from a comma separated variable (CSV) formatted file into R environment.

5.1 Data cleaning and exploration

Descriptive Statistics for deals data

Table 5.1 gives summaries of the numerical variables and below the variables are described.

	Minimum	Median	Mean	Maximum
Dealt Amount	R -28,720,227.00	R 149.00	R 392.00	R 9,420,454.00
Counter Amount	R -120,000,000.00	R -1,298.00	R -2,274.00	R 300,000,000.00
Turnover	R 75,435.00	R 4,908.00	R 36,964.00	R 300,000,000.00
Spot Rate	-0.02	8.86	9.33	24.37
Effective Rate	0.17	8.78	9.26	24.37
Deal Revenue	R -1,585,989.00	R 102.50	R 435.30	R 1,563,200.00
USD Spot Rate	-0.02	8.09	8.13	15.92
Range	1.00	1.00	1.82	14.00
Margin	-1.08	0.02	0.02	1.00

Table 5.1: Summary statistics for numerical variables in the deals data

- The *dealt amount* consists of mixed currency values; e.g. USD10 000, GBP 10 000, EUR 10 000. This variable was excluded from the analysis, and the Rand equivalent turnover was taken into consideration.

- The *turnover* consists of a single negative value of R75 440 which was removed and it also has a low percentage of zeros (0.02%). The observations are concentrated at the same point. Transactions of R50 000 or less constitute 89% of this data. Deals greater than R50 000 and less than R5 000 0000 make up 11%. Fewer transactions with the turnover value greater than R5 million make up 0.04% of the set and are all price sensitive; therefore, the threshold is kept at R5 million. Half of the observations lie below the median value of R4 907,

and half lies above.

- According to Reuters, the USDZAR traded between R6.40 and R15.95 during the period the data was collected. A negative value was excluded from the analysis as this currency pair never traded at this rate. This paper uses this exchange rate as it is a base currency pair that is used to calculate the effective exchange rate and the *spot rate* and the three rates essentially correlate.

Frequency Analysis

Table 5.2 to Table 5.9 demonstrate the categorical variables and the explanation of each variable is given.

- *Sensitivity*

  In the below Table 5.2 it can be observed that 71.38% of the customers are not price sensitive, while 28.62% are recorded as sensitive. This proportion is probably because a majority of this data is made up of individual consumers who transact in small volumes and therefore receive a standard system pricing.

	Frequency	Percentage
Not sensitive	638619	71.38%
Sensitive	255997	28.62%
Total	894616	100.00%

Table 5.2: Distribution of Sensitivity

- *Deal Type*

  There are 66% purchase and 34% sale transactions.

	Frequency	Percentage
Purchase	592990	66.28%
Sell	301626	33.72%
Total	894616	100.00%

Table 5.3: Distribution of Deal Type

- *Dealt Currency*

  USD, EUR, ZAR make up most of the transactions traded and therefore the analysis was based on these currencies traded against the ZAR. USD trades constitute 65%, EUR and GBP deals contribute 23% and 12% respectively.

Dealt Currency	Frequency	Percentage
USD	585082	65.40%
EUR	201175	22.49%
GBP	108359	12.11%
Total	894616	100.00%

Table 5.4: Distribution of Dealt Currency

- *Agreement Type*

  The partially optional agreement has the highest frequency 470038 (52.54%)

of all agreement types, followed by spot 320988 (35.88%) and fixed type 10354 (11.54%). The fully optional agreement type has the lowest number of agreements, constituting only 336 (0.04%).

	Frequency	Percentage
Fixed	103254	11.54%
Fully optional	336	0.04%
Partially optional	470038	52.54%
Spot	320988	35.88%
Total	894616	100.00%

Table 5.5: Distribution of Agreement Type

- *Customer Segment*

  Individual consumers contribute 77.8% of the data, followed by business banking at 12.81%; wealth clients contribute 5.73% and commercial contribute 3.66%.

	Frequency	Percentage
Business banking	114644	12.81%
Commercial	32787	3.66%
Consumer	695711	77.77%
Wealth	51474	5.75%
Total	894616	100.00%

Table 5.6: Customer Segment

- *Dealing Desk*

  This variable was excluded from the analysis as it behaves as a surrogate variable; there is a perfect association between the response variable and dealing desk.

	Frequency	Percentage
Individualised/Special Pricing	226096	25.27%
Standard System Pricing	638619	71.38%
Small Dealing Desk	22469	2.51%
Large Dealing Desk	7432	0.83%
Total	894616	100.00%

Table 5.7: Distribution of Dealing Desk

- *Forward Exchange Contract (FEC)*

  Almost all of these deals are not FECs, only 2500 (0.28%) are forward contracts.

Therefore it will be excluded from the analysis.

	Frequency	Percentage
No	892116	99.72%
Yes	2500	0.28%
	894616	100.00%

Table 5.8: Distribution of Forward Exchange Contract

- *Product*

  Products vary from cash passport, notes to swifts. Notes are the most traded product (32.10%), followed by inward swift (20.81%) and then outward swifts (7.37%). Travelers' cheques and multiple currency accounts are the least favourite products, 30.14% values of this variable are missing.

	Frequency	Percentage
Notes	287158	32.10%
Inward swift	186158	20.81%
Outward swift	65958	7.37%
Cash Passport	36996	4.14%
Foreign checks	19977	2.23%
Other	28761	3.21%
NAs	269608	30.14%
Total	894616	100.00%

Table 5.9: Distribution of Product Type

Descriptive Statistics for Customer data

Table 5.10 below shows the summary statistics of each numerical variable in the customer data and further below the variables are explained.

	Minimum	Median	Mean	Maximum	NAs
Customer Age	0	39	40	116	6
Income	R -	R 256,700.00	R 3,661,000.00	R 1,000,000,000,000.00	
Years banking with the bank	0	12	14	116	
DDA Balance	R -91,319.00	R 7,541.00	R 120,609.00	R 197,000,000.00	75555
TDA Balance	R -	R 12,493.00	R 207,606.00	R 282,000,000.00	357236
Card Balance	R -4,050,000.00	R 8,069.00	R 18,363.00	R 651,818.00	287142
Average no. of deals per annum	1	1	3	79929	
Turnover per annum	R -	R 14,150.00	R 162,700.00	R 2,350,000,000.00	

Table 5.10: Summary statistics for numerical variables in the Customer Data

- Customer *age* shows that age zero appears 41826 times (9.38%), which indicates new business. The average age is 39 years, and the maximum age is 116 years. Ages between 0 and 17 indicate the year's companies have been in existence which is approximately 10% of the data. There are six records with missing values. See appendix Figure A.35.

- *Income*, the minimum value of R0.00 shows the income of customers that do not receive their salary via bank. Income less than R100 000 makes up approximately 29% of the variable, while income between R100 000 and R500 000 contribute the most at 43% and amounts greater than R500 000 and less than R1 000 000 make up 18%. Income higher than R1 000 000 contributes the least, close to 10%. Fifty percent of the income value lies below the median value R256 700, while the other fifty percent lies below. See appendix Figure A.25.

- *Years with bank*, the new customers at the bank show zero number of years, while fifty percent of the customers have been banking with the bank for less than twelve years and fifty percent for more than twelve years. See appendix Figure A.34.

- *DDA balance*, approximately 17% of the data is missing probably because of the sensitivity of the financial information. Minimum negative amount of R91 320 shows that some customers are in deficit and the maximum amount of R196 700 000 is a balance for a customer from the wealth segment. The median is R7 541; fifty percent of the balance falls on either side of this value. See appendix Figure A.20.

- *TDA balance* summaries show a large number is missing values (80%), not much financial information is contained in this variable as it was pointed out that this account is rarely used in the FOREX department. See appendix Figure A.32.

- In *card balance* approximately 64% of the data are missing, probably owing to the sensitivity of the financial information. Negative minimum value shows that the account has been overdrawn and the maximum balance is R651 800. See appendix Figure A.18.

- *Deal per annum* shows that fifty percent of the deals lie below the median value of 1.33 and the other fifty percent lie above. The maximum number of deals done on average is 79 930 per annum. See appendix Figure A.21.

- *Turnover per annum* fifty percent of the turnover lie below the median value of R14 150 000 and the other fifty percent lie above. The zero value indicates turnover of new clients; See appendix Figure A.33.

Frequency Analysis

Table 5.11 to Table 5.23 describes the categorical variables and further below the detail of each variable is given.

- *Customer Type*

  In it can be observed that private individuals are a majority in this data set, making up approximately 87% of the data and the rest are companies and non-profit organisations.

	Frequency	Percentage
Private Individual	387727	86.95%
Close Cooperation	18934	4.25%
Private Company	17567	3.94%
Sole Proprietorship	7941	1.78%
WB Individuals	6460	1.45%
Non-Profit Company	1659	0.37%
(Other)	5649	1.26%
Total	445937	100.00%

Table 5.11: Distribution of Customer Type

- *Highest Education Level*

  Highest education level was re-coded to address what appeared to be capturing errors. It consisted of meaningless characters like "*", "&" and "Y" which were removed. Other values such as "*10" and "*12" were assumed to represent Grade 10 and Grade 12 respectively and were thus re-coded accordingly. See Appendix Table A.8. "NAs" was assigned to the 13% of the data that were missing. After re-coding this variable consist of 64% Grade 0 to Grade 12 records and 23% post matric.

	Frequency	Percentage
G0 - G12	285557	64.04%
Post Matric	101843	22.84%
NAs	58537	13.12%
Total	445937	100.00%

Table 5.12: Distribution of Education Level

- *Customer segment*

  Customer segment was excluded as it is in the deals data as well, the two data sets were merged.

	Frequency	Percentage
Consumer	355002	79.61%
Wealth	42012	9.42%
Business Banking	41541	9.32%
Commercial	6577	1.47%
Commercial GTS	713	0.16%
Corporate	59	0.01%
Other	33	0.01%
Total	445937	100.00%

Table 5.13: Distribution of Customer Segment

- *Bank VIP*

  Almost all these clients are considered not VIP. This variable was not included in modelling.

Bank VIP	Frequency	Percentage
No	443548	99.46%
Yes	2389	0.54%
Total	445937	100.00%

Table 5.14: Distribution of Bank VIP

- *Online Indicator*

  The Online indicator; "0" was re-coded to "No" and 1 to "Yes". 71% of the

customers have access to online exchange rates indication, while 29% do not have access.

	Frequency	Percentage
No	128290	28.77%
Yes	317647	71.23%
Total	445937	100.00%

Table 5.15: Distribution of Online Indicator

- *Salary with the Bank*

  Salary with bank was re-coded from "0" to "No" and "1" to "Yes". Only 30% of the customers earn their salary via the given bank.

	Frequency	Percentage
No	310737	69.68%
Yes	135200	30.32%
Total	445937	100.00%

Table 5.16: Distribution of salary with the bank

- *Points Incentive Indicator*

  Points incentive indicator shows 74% of the customers participate in the reward scheme while 26% do not. This variable was re-coded from "0" to "No" and "1" to "Yes".

	Frequency	Percentage
No	116657	26.16%
Yes	329280	73.84%
Total	445937	100.00%

Table 5.17: Distribution of Points Incentive Indicator

- *TDA Account*

  TDA account is the investment account consisting of thirty levels (see Appendix Table A.10). This variable was re-coded to two levels. 80% of the customers hold zero TDA account while 20% hold at least one account.

TDA Account Recoded	Frequency	Percentage
One and above	88701	19.89%
Zero	357236	80.11%
	445937	100.00%

Table 5.18: Distribution of TDA account

- *Global Account Indicator*

  Global account indicator is a two-level categorical and was re-coded from "0" to "No" and from "1" to "Yes". 99% of the customers have no global accounts; this variable was excluded from the analysis.

	Frequency	Percentage
No	440433	98.77%
Yes	5504	1.23%
Total	445937	100.00%

Table 5.19: Distribution of Global Account Indicator

- *Banked with the given bank*

  This is a two-level variable re-coded from "0" to "No" and "1" to "Yes". Customers that are banked with this bank make up 83% of this variable.

Banked with the given bank	Frequency	Percentage
No	75555	16.94%
Yes	370382	83.06%
Total	445937	100.00%

Table 5.20: Distribution of banked with bank

- *FX VIP*

  FX VIP variable was re-coded from "0" to "No" and "1" to "Yes". VIP clients qualify for special rates. Almost all clients are not FX VIP, perhaps because the majority of the customers are consumers and they deal mainly in small amounts. This variable was excluded as almost 100% constitutes none VIP status.

	Frequency	Percentage
No	444421	99.66%
Yes	1516	0.34%
Total	445937	100.00%

Table 5.21: Distribution of FX VIP

- *Risk Category Code*

  In risk category a majority of clients (86%) have a normal risk profile, 8% are under investigation, 2% are regarded intensive, 2% are treated as special, 1% insolvent and only one client shows as fraudulent.

Risk Category Code	Frequency	Percentage
Fraud	1	0.00%
Insolvent	5010	1.12%
Intensive	10553	2.37%
Investigative	35957	8.06%
Normal	384566	86.24%
Special	9850	2.21%
Total	445937	100.00%

Table 5.22: Distribution of Risk Category Code

- *Pre Dealing Indicator*

  This variable has two levels "0" was re-coded to "No" and "1" to "Yes". 99.97%

of the customers do not have a pre dealing facility; this variable was excluded.

	Frequency	Percentage
No	445791	99.97%
Yes	146	0.03%
Total	445937	100.00%

Table 5.23: Distribution of Pre Dealing Indicator

5.2 Merged datasets

After preprocessing, Deals and Customer data sets were merged using a universal identifier variable *customer number*. The merged set consists of 893225 records and 31 variables. The objective is to classify customers that are price sensitive from those that are not. A set of summary statistics for selected attributes are shown in the Appendix from Table A.14 to Table A.33. After merging and cleaning the two datasets it can be seen that the turnover was capped at R5 million, the USD Spot Rate is within the correct trading ranges of between R6.40 and R15.92, and all the recorded variables are shown. The models were built based on this data set.

Correlation

Figure 5.1 suggests a strong positive correlation of 0.9 between *turnover per annum* and *deals per annum*, as the number of deals increases so does the turnover per annum. The *turnover* is correlated with *deals revenue* at 0.7.

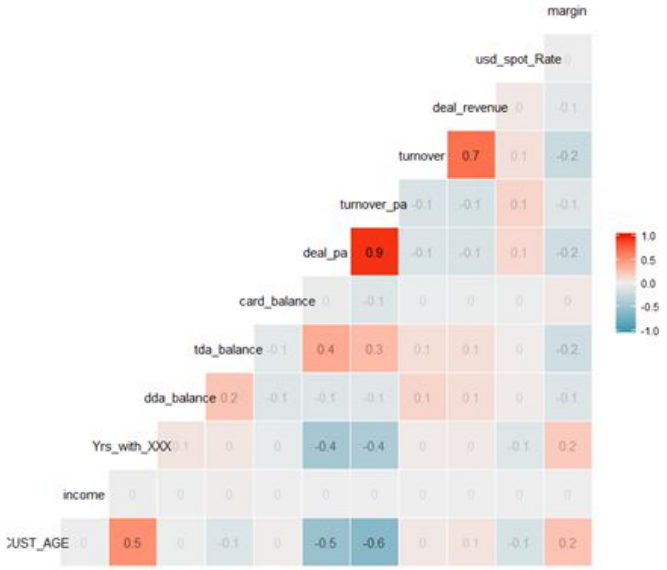

Figure 5.1: Correlations of numerical variables

Cramer's V

Cramer's V test results are shown in (Table 5.24).The range of Cramer's V is [0,1] for tables larger than 2X2; [-1,1] for 2X2 tables and cannot determine the direction of the association. A strong level of association was found between the predictor variables *customer type* (Cramer's V = 0.4), *online indicator* (Cramer's V = -0.3), *dda account* (Cramer's V = -0.5) and the response variable. There is a very strong association between explanatory variables *points incentive indicator* (Cramer's V = 0.6), *banked with bank* (Cramer's V = 0.7), *agreement type* (Cramer's V = 0.7), *product* (Cramer's V = 0.6) and *sensitivity. Dealing desk* has a perfect relationship with the response variable (Cramer's V = 1), if the independent variable is known, the dependent variable can be perfectly predicted. The rest of the variables seem to

have a weak relationship with the dependent variable.

	Customer Type	Education Level	Customer Segment	Bank VIP	Online Indicator	Salary with bank	Points incentive
Chi-Square	84941	289.3238	19254.0358	524.0174	51101.8642	14485.3289	233794
Likelihood Ratio Chi-Square	112823	276.8038	23086.4012	469.8153	60222.9628	19295.2952	243043
Continuity Adj. Chi-Square	84939	288.9463	10950.0305	523.111	51100.4115	14484.1059	233791
Mantel-Haenszel Chi-Square	84940	289.3223		524.0165	51101.7768	14485.3041	233794
Phi Coefficient	0.38118	0.039	0.1815	0.0299	-0.2957	-0.1574	0.6324
Contingency Coefficient	0.35618	0.0389	0.1786	0.0299	0.2835	0.1555	0.53449
Cramer's V	0.38118	0.039	0.1815	0.0299	-0.2957	-0.1574	0.6324
Comment	Strong	Very Weak	Weak	Very Weak	Strong	weak	Very strong
	DDA Account	TDA Account	Card Account	Global Account	Banked with bank	Risk category	FX VIP
Chi-Square	122153	10303.3796	20020.2247	15.7024	258348	9222.2938	1230.968
Likelihood Ratio Chi-Square	141333	13063.4163	23700.3541	15.3991	260280	12142.8661	1063.061
Continuity Adj. Chi-Square	122151	10302.2948	20019.1449	15.5752	258344	7322.3525	1229.552
Mantel-Haenszel Chi-Square	122153	10303.362	20020.1904	15.7023	258347		1230.9659
Phi Coefficient	-0.45712	-0.1328	-0.1851	0.0052	0.66478	0.1256	0.0459
Contingency Coefficient	0.41574	0.1316	0.182	0.0052	0.55361	0.1246	0.0458
Cramer's V	-0.45712	-0.1328	-0.1851	0.0052	0.66478	0.1256	0.0459
Comment	Strong	Weak	Weak	Very Weak	Very strong	Weak	Very Weak
	Pre dealing	Deal Type	Agreement Type	Customer Segment	Dealing Desk	FEC	Product
Chi-Square	678.0522	829.4523	291267	19091.8727	584586	1208.359	151619
Likelihood Ratio Chi-Square	585.4874	820.6892	296852	23239.8155	664558	1014.5741	168717
Continuity Adj. Chi-Square	676.6445	829.2659	323.74494	4326.1536	549021	1206.331	46453
Mantel-Haenszel Chi-Square	678.0511	829.4509				1208.357	
Phi Coefficient	0.0341	-0.0377	0.70586	0.1807	1	0.0455	0.64509
Contingency Coefficient	0.034	0.0376	0.57667	0.1778	0.70711	0.0454	0.54208
Cramer's V	0.0341	-0.0377	0.70586	0.1807	1	0.0455	0.64509
Comment	Very Weak	Very Weak	Very strong	Weak	Perfect	Very Weak	Very strong

Table 5.24: Statistics for categorical variables

5.3 Splitting the data

The training set consists of 625257 records, of these cases 71% are not price sensitive while 29% are price sensitive. The test set consists of 267968 records, 71% are not price sensitive while 29% are. The validation set also consists of 71% not price sensitive customers and 29% sensitive ones.

5.4 Decision Trees

The decision tree models grown on the merged data resulted in seven different trees. Tree 1 B.1 includes all variables, Tree 2 B.2 excludes the *branch code*, Tree 3 B.3 excludes the *branch code* and *margin*, Tree 4 B.4 replaces zero values in the *income* variable with missing values 'NA' and excludes *branch code*. The last two trees consider the highly correlated variables (correlation >= 0.9, which is close enough to 1) *turnover per annum* and *deal per annum*. Tree 5 excludes *turnover per annum* and *branch code*; Tree 6 B.6 excludes *deals per annum* and *branch code*, Tree 7 B.7 excludes *branch code* and includes imputed variables *product* and *dda balance*.

These variables sets were selected to see the impact they have on the model accuracy. The output of these trees analysis is shown in the appendix (Tree 1 B.1 – Tree 7 B.7). All seven trees model performances were shown and compared, however only the tree with the highest accuracy level is described below. The reason behind is that there are many trees produced in this section and the method to describe each

tree is the same. Trees were used to extract variables that would be used in RF and NN modelling.

This section describes and considers the CART procedure that generated (Tree 5 5.2) with the highest level of accuracy amongst the seven produced trees, this tree contains seven terminal nodes and six splits (Tree 5 B.5). The percentage of customers who are not price sensitive range from 98% to 100% in the four groups and those that are price sensitive range between 88% to 99% in the three groups. The first variable selected for splitting is *agreement type*. Among those who had foreign exchange transactions of *agreement type* (fixed or partial), the *margin* taken per transaction was greater or equal to 0.0012 (Group 1; 99% are not price sensitive).

A group that had foreign exchange transactions of *agreement type* (fixed or partial), the *margin* taken per transaction was less than 0.0012, *average number of deals per annum* was greater or equal to 3045 (Group 2; 99% are not price sensitive) or average *number of deals per annum* is less than 3045 (Group 3; 88% are price sensitive). Those who had foreign exchange transactions of *agreement type* (spot or fully optional contracts) were differentiated by average *number of deals per annum*, separating those with a *deals per annum* less than 258, *margin* greater of equal to 0.0011 (Group 4; 98% are not price sensitive) or *margin* less than 0.0011 (Group 5; 91% are price sensitive). Those who had foreign exchange transactions of *agreement type* (spot or fully optional contracts) were differentiated by average *number of deals per annum*, separating those with *deals per annum* greater or equal to 258, *years*

with bank less than 7.5 (Group 6; 100% are price not sensitive) or *years with bank* greater or equal to 7.5 (Group 7; 99% are price sensitive).

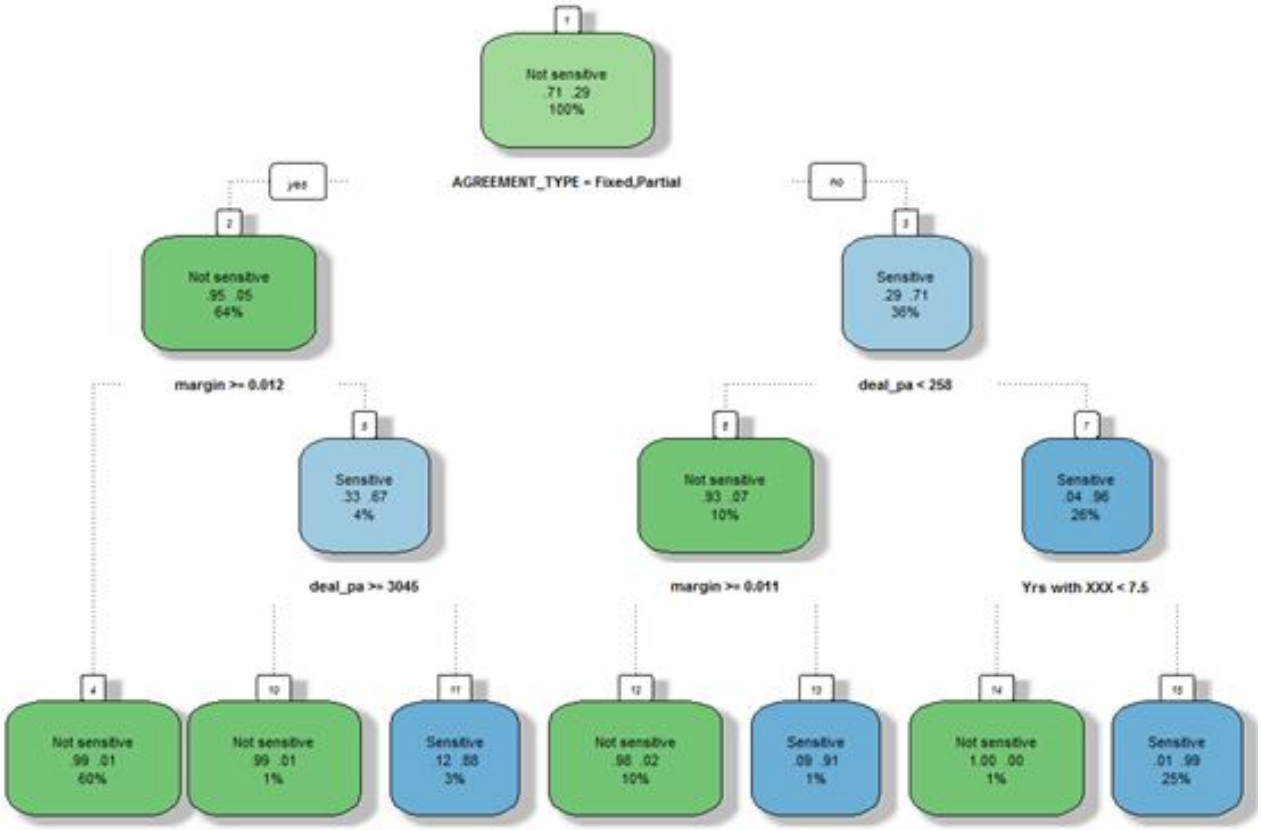

Figure 5.2: Decision Tree fitted without branch code and turnover per annum

5.4.1 Confusion Matrix

The confusion matrix is used to describe the performance of the classification tree. The prune function is also used to simplify the tree based on complexity parameter associated with the smallest cross-validated; however, the pruned trees did not make any improvement on the actual trees.Table 5.25 demonstrates a comparison of results obtained from the confusion matrix using CART fitted in data sets with and without some variables.

In Table5.26 it can be seen that Tree 3 B.3 has the highest error 3.29% amongst the seven trees, followed by Tree 1 B.1 with an error rate 1.97%, Tree 6B.6 has a misclassification rate of 1.64%, and Tree 2 B.2 and Tree 4 B.4 have the same error rate of 1.62%. The decision tree performed best on Tree 5 B.5 showing the lowest misclassification rate of 1.54% when fitted on the test set without the *branch code* and *turnover per annum*. This tree indicates that the most important predictors of deciding on customer price sensitivity are *agreement type*, *deals per annum*, *margin* and *years with bank*. Tree 7 B.7 fitted in the data with the imputed *product* and *dda balance* variables has the second lowest misclassification rate of 1.6%, and the selected predictors to have the most predictive power are *dda balance,margin,deals per annum,product* and *customer type*.

	Tree One		Tree Two		Tree Three	
	Not sensitive	Sensitive	Not sensitive	Sensitive	Not sensitive	Sensitive
Not sensitive	188922	2593	189772	1743	190757	758
Sensitive	2685	73768	2589	73864	8054	68399
	Tree Four		Tree Five		Tree Six	
	Not sensitive	Sensitive	Not sensitive	Sensitive	Not sensitive	Sensitive
Not sensitive	189772	1743	190014	1501	189776	1739
Sensitive	2589	73864	2637	73816	2643	73810

Table 5.25: Comparison of accuracy levels obtained from the classification trees

	Tree 1	Tree 2	Tree 3	Tree 4	Tree 5	Tree 6	Tree 7
Accuracy	98.03%	98.38%	96.71%	98.38%	98.45%	98.36%	98.04%

Table 5.26: Accuracy levels

5.5 Random Forests

After sampling and splitting the data, training, test and validation samples had similar target variable distribution. Twenty decision trees were built on each of the four data sets (Dataset 1, Dataset 2, Dataset 3, Dataset 4). The plots indicate that after 20 decision trees; there is no significant reduction in error rate for the model fitted in all four sets(See Appendix, Figure B.8 - Figure B.11). These plots show a

black solid line for overall OOB (out of the bag) error and a red and green lines, one for each class' error (not sensitive and sensitive). The Gini tests to see the result if each variable is taken out and a high score means the variable was important. Table B.1 to Table B.4 demonstrate the decreasing order of variable importance based on a measure of node impurity for each data set.

5.5.1 Variable Importance

The highest purity means that each node contains only elements of a single class. Table 5.27 shows the decreasing order of the top ten most important variables regarding "Mean Decreasing Gini" based on Dataset 1, Dataset 2, Dataset 3 and Dataset 4. The higher the value of "mean decrease Gini" score ,the higher the importance of the variable in the model. Essentially, it estimates the impact a variable has on the model by comparing prediction accuracy rates for models with and without the variable.

It can be seen from the table below that for Dataset 1 the *turnover per annum* is the most critical variable followed by the *deal agreement type*, the *average number of deals* the client dealt per annum is the third most crucial variable, the *years with the bank* and *margin*. In this set the *USD Spot Rate* is shown to have the least impact on the classification model. The customers with the highest turnover are most likely to be price sensitive.

The top ten important variables across all four data sets consist of those that were

selected to have the most predictive power by the decision tree model, except for the *USD Spot Rate* ranked at number twelve in Dataset 2. Dataset 2 shows variables which had the greatest impact in the classification model that was built based on the dataset that excludes all the variables with missing values and missing records. In this set variable *points incentive indicator* was chosen to be the most important variable, followed by *banked with the bank*. These variables indicate a customer's loyalty to the bank, for instance, a customer gets incentivised for using the bank's products; most customers are multi-banked therefore the latter variable determines if bank is the bank of choice. Demographic variable *age* and *customer type* are ranked in the top ten in the three data sets.

In Dataset 3 the missing values in a variable *product* were imputed. RF chose the imputed variable to be the seventh most important variable that has an impact on the model. RF was built on the Dataset 4 with the imputed *dda balance* included in the data set. The *dda balance* is the transactional account; therefore, it was of interest to see the model performance with this variable. According to the "Mean Decreasing Gini" in Table 5.27 *dda balance* is by far the most critical variable in this setting. This result is pleasing as this account type is actively used for settlements. The imputed *p*roduct is ranked the second most important variable.

The variables vary between the datasets because each set excludes some variables to see the impact their absence has on the prediction. Looking at the vari-

ables that feature at the top mostly in the given variable importance table,it can be seen that customer loyalty related variables such as *banked with the bank*,*years with the bank*,*Points Incentive Indicator*; financial variables such as *turnover per annum*, *turnover*, transactional account balance *DDA balance*; the type of *product* the customer buys or sells,*margin* taken from the deal, the contract *agreement type*; demographic variables *age* and *customer type* were ranked as the highest importance variables.

Variables such as *Card Account*,*TDA account*,*DDA account*,*dealt currency*,*deal type*,*online indicator*,*risk category code*, if the customer earns *salary* via the bank seem to be consistently ranked as the least important.

Data set 2			Data set 1		Data set 3		Data set 4	
Rank	Variables	Mean Decrease Gini	Variables	Mean Decrease Gini	Variables	Mean Decrease Gini	Variables	Mean Decrease Gini
1	Points Incentive Indicator	21461.13	Turnover per annum	35292.49	Banked with the Bank	22990.37	DDA balance	23111.35
2	Banked With The Bank	15688.69	Agreement Type	32721.23	Agreement Type	20507.11	Product	16994.94
3	Agreement Type	15169.21	Deals per annum	25684.57	Deals per annum	14767.25	Banked with the Bank	13373.76
4	Customer Type	13708.78	Years with the bank	16146.55	Turnover per annum	13914.38	Margin	12412.07
5	Deals per annum	13304.007	Margin	12891.51	Margin	10607.21	Turnover per annum	12187.95
6	Margin	10684.015	USD Spot Rate	4108.47	Customer Type	21461.13	Deals per annum	10554.61
7	Turnover per annum	7261.51			Product	9048.66	Agreement Type	10092.11
8	Years with the bank	7101.82			Years with the bank	7787.31	Years with the bank	7080.28
9	Customer Age	5545.48			Customer Age	2956.39	USD Spot Rate	2866.07
10	Income	3415.61			Points Incentive Indicator	2788.80	Customer Age	2787.90

Table 5.27: Variable importance for data sets

Accuracy measures of predictions

The models were built in an attempt to predict whether the customer is price sensitive or not based on the training data sets of Dataset 1, Dataset 2, Dataset 3 and Dataset 4 described above (Paragraph 5.3). Predictions were made on the unseen test set and validation set to see how well the models perform. The accuracy measures used were the overall percentage correctly classified (PCC), sensitivity (the percentage of presences correctly classified), specificity (the percentage of absences correctly classified) and the area under the receiver operating characteristic curve (Table 5.28).The models did a decent job showing accuracy levels and area under the curve of at least 0.99 across all fitted models(See B.2.3 in the Appendix section).

The specificity and sensitivities are all high and similar. After building models to the four datasets with and without some predictors, similar prediction accuracy was substantially retained. However, it is evident from the below table that datasets with imputed variables produced the best performing model. Dataset 3 gave a PCC of 0.9959 and 0.9961 on the test and validation set respectively and the AUC of 0.9991. Dataset 4 gave a PCC of 0.9958 and 0.9959 on the test and validation set respectively and the AUC of 0.999.

	Test 1	Validation 1	Test 2	Validation 2	Test 3	Validation 3	Test 4	Validation 4
PCC	0.9915	0.9918	0.9954	0.9955	0.9959	0.9961	0.9958	0.9959
Specificity	0.9949	0.9953	0.9983	0.9985	0.9983	0.9986	0.9985	0.9986
Sensitivity	0.9826	0.983	0.9882	0.9879	0.9899	0.9899	0.989	0.9893
AUC	0.9968		0.9989		0.9991		0.999	

Table 5.28: Model accuracy levels

5.6 Neural Networks

The topology of the NN was visualised (see Appendix from Figure B.16 - Figure B.19). In the topology the signal from each of the predictors is received by input node. The weights are displayed above each line. The circles indicate bias, corresponding to the intercept in the conventional regression model. There is one hidden layer consisting of one unit.This architecture was chosen because it is often faster than training with back propagation and it does not require the specification of any free parameter values(Ahmad et al, 2008).

The variable *sensitivity* is an output neuron. The weights dictate the relative influence of information that is processed in the network such that input variables that are not relevant in their correlation with a response variable are suppressed by the weights. The opposite effect is seen for weights assigned to explanatory variables that have strong, positive associations with a response variable. A significant variance of generalised weights for a covariate indicates non-linearity of its independent

effect. If generalized weights of a covariate are approximately zero, the covariate is considered to have no effect on the outcome. The same architecture was fitted on the four sets.

A summary of the results in Table 5.30 shows the total error of 53607 and 21299 steps were needed for iterations to converge for Dataset 1. In Dataset 2, the total error of 26178 and 57060 steps were needed. The architecture produced the total error of 22926 and the steps needed for iterations to converge were 31127 for Dataset 3. NN produced the total error of 22926 in Dataset 4 and the steps needed for iterations to converge were 31127. The intercepts of the hidden layer is 5.88, -2,49, -0.54 and -0.77 for the given set respectively and the threshold 0.01 for all data sets.

The full summary tables are in the appendix from Table B.5 to Table B.8.Table 5.29 shows the top five variables with the highest weights. NN fitted in the data sets that consist of the imputed variables produced the same estimated weights 2380.27, -1013.75, -230.47, 36.66 for the covariates *deal revenue*, *agreement type* (optional), *turnover*, *customer type* (government). The weights for *customer type* (registered friendly society) and *customer type* (stokvel) were similar. *Customer type*(government,registered friendly society, stokvel,cooperative) weights leading to the hidden neuron are estimated as -710.44, -708.80, 708,37 and 708,33, these weight denote how much this feature matters in the model.*Deal revenue* has the highest weight of -4718.97 in Dataset 2. The top five weights in this set are negative.

In Dataset 1 the four weights leading to the hidden neuron are estimated as 161.47, -45.11, -0.35, 3.29, -111,23, 127.54, 5.92 and -146.51 for the covariates *agreement type* (optional, partial optional, spot), *years with the bank*, *deals per annum*, *turnover per annum*, *USD spot rate* and *margin*, respectively.

Data set 1		Data set 2		Data set 3		Data set 4	
Variables	Weight	Variables	Weight	Variables	Weight	Variables	Weight
Agreement Type Optional	161.47	Deal revenue	-4718.97	Deal revenue	2380.27	Deal revenue	2380.27
Margin	-146.51	Customer Type government	-710.44	Agreement Type Optional	-1013.75	Agreement Type Optional	-1013.75
Turnover per annum	127.54	Customer Type reg. friendly society	-708.80	Turnover	-230.47	Turnover	-230.47
Deals per annum	-111.23	Customer Type stokvel	-708.37	Customer Type government	36.76	Customer Type government	36.66
Agreement Type Partial	-45.11	Customer Type Cooperative	-708.33	Customer Type reg. friendly society	36.30	Customer Type stokvel	35.03

Table 5.29: Summary of the NN results based on the four data sets

Data set 1		Data set 2		Data set 3		Data set 4	
error	53607.56	error	26178.70	error	22926.90	error	22909.43
Reached threshold	0.01	Reached threshold	0.01	Reached threshold	0.01	Reached threshold	0.01
Steps	21299.00	Steps	57060.00	Steps	31127.00	Steps	30719.00
Intercept	5.88	Intercept	-2.49	Intercept	-0.54	Intercept	-0.77

Table 5.30: The table shows the number of steps needed in the training process, the threshold for the partial derivatives of the error function and the intercept

Accuracy measures of predictions

The generalised weights help with the interpretation of NN model, allowing examination of the independent effect of the covariate; another purpose of NN is to

assist prediction of future observations. The predicted values were computed based on the trained network on the new samples. The accuracy measures used were again the PCC, sensitivity, specificity and the AUC (Table 5.31). The models did a good job showing accuracy levels and area under the curve of at least 0.9 across all fitted models. It is evident from the accuracy results that the model fitted in the set with variables selected by the decision trees gives the lowest accuracy regarding performance in comparison to other three datasets, while the model fitted in the data sets with imputed variables *product* and *dda balance* give the highest accuracy measures.

The accuracy of the model fitted in the test set 1 was found to be 0.9573, validation set 1 accuracy was 0.9566; test set 2 accuracy was found to be 0.9791, for validation set 2 accuracy was 0.9791; test set 3 was found to be 0.9809; validation set 3 accuracy was 0.9805; test set 4 was found to be 0.9809; validation set 4 accuracy was 0.9806. Furthermore, the evaluation of the model performance on the four data sets was done using ROC curve. The area under the ROC curve is higher than 0.9 for all four models, meaning that the models are performing very well.

The four charts in paragraph B.3.3 in the Appendix section are results from the ROC curves of the four samples. The area under the curve based on Dataset 1 is 0.9286, for Dataset 2 (0.9844), for Dataset 3 (0.9895), and Dataset 4 (0.9895). All four sets had similarly good results based on AUC. The results are however not that different which means that the model was stable.

	Test 1	Validation 1	Test 2	Validation 2	Test 3	Validation 3	Test 4	Validation 4
PCC	0.9573	0.9566	0.9791	0.9791	0.9809	0.9805	0.9809	0.9806
Specificity	0.9965	0.9961	0.9876	0.9884	0.989	0.9894	0.989	0.9888
Sensitivity	0.8588	0.8581	0.9579	0.9559	0.9607	0.9583	0.9607	0.9602
AUC	0.9286		0.9844		0.9895		0.9895	

Table 5.31: Model accuracy levels

Chapter 6

Comparison of Models

The prediction results produced by the models can be used to price the customers accordingly. If the prediction is *"sensitive"* which means the client is sensitive to exchange rate then the suggested decision is to price competitively from the dealing desk or the client be put under special pricing on the system. Whereas, if the prediction is *"not price sensitive"*, it means the client is not sensitive to exchange rate then the suggested pricing decision is to give standard pricing. Any wrong prediction can cost the bank a great deal of money. Hence, the model should be evaluated for its robustness.

The parameters that are used to evaluate the robustness of a binary classifier are accuracy, specificity, sensitivity and AUC. Accuracy measures the portion of all testing and validation samples classified correctly. Sensitivity measures the ability of a classifier to correctly identify *sensitive* labels while specificity measures the

classifier's ability to identify *not sensitive* labels correctly. These parameters were calculated for Dataset 1, Dataset 2, Dataset 3 and Dataset 4 prediction model using RF and NN. It can also be seen from the ROC curves that the RF model proves to be the most optimal; however, the NN results were also excellent. The area under the ROC curve is an essential parameter for evaluating the performance of a binary classifier.

An area of 1 represents an excellent classifier; an area of .5 represents a worthless classifier which produces random outputs. In other words, the area measures discrimination, that is, the ability of the classifier to correctly classify a positive shift and a negative shift in customer price sensitivity in case of this research. The area under the ROC curve was above 0.9 for both models using all four datasets. The results obtained from RF and NN analysis are comparable. In the former analysis, there seem to be slight improvements in most accuracy measures. For instance, all AUC, Sensitivity and PCC levels in the RF were slightly higher than in the NN. These models were complicated, and they both took a long time to execute. The results are provided in the table below(Table 6.1):

Results for test set 1					**Results for validation set 1**			
Model	PCC	Specificity	Sensitivity	AUC	Model	PCC	Specificity	Sensitivity
RF	0.9915	0.9949	0.9826	0.9968	RF	0.9918	0.9953	0.983
NN	0.9573	0.9965	0.8588	0.9286	NN	0.9566	0.9961	8.8581
Results for test set 2					**Results for validation set 2**			
Model	PCC	Specificity	Sensitivity	AUC	Model	PCC	Specificity	Sensitivity
RF	0.9954	0.9983	0.9882	0.9989	RF	0.9955	0.9985	0.9879
NN	0.9791	0.9876	0.9579	0.9844	NN	0.9791	0.9884	0.9559
Results for test set 3					**Results for validation set 3**			
Model	PCC	Specificity	Sensitivity	AUC	Model	PCC	Specificity	Sensitivity
RF	0.9959	0.9983	0.9899	0.9991	RF	0.9961	0.9986	0.9899
NN	0.9809	0.989	0.9607	0.9895	NN	0.9805	0.9894	0.9583
Results for test set 4					**Results for validation set 4**			
Model	PCC	Specificity	Sensitivity	AUC	Model	PCC	Specificity	Sensitivity
RF	0.9958	0.9985	0.989	0.999	RF	0.9959	0.9986	0.9893
NN	0.9809	0.989	0.9607	0.9895	NN	0.9806	0.9888	0.9602

Table 6.1: Test and Validation sets accuracy results based on four datasets

Chapter 7

Limitations of this Research Report

This research is dependent on the information that entails financial status of the customers received from a financial institution. Data of this nature is confidential, and institutions are very sceptical to divulge records. The study also depends on the accuracy of scoring of customer sensitivity assigned by the institution. Scoring customers as price sensitive has proven not be an easily determinable concept due to the competitive nature of the industry and possibly because there is no prior study on the subject matter. This work, therefore, assumes that NNs and RF can be used as an additional support system, in conjunction with the existing system, to improve the ability to predict customer price sensitivity.

The data consist of many missing values in the financials related variables; this information is critical and could add more value to the analysis. Demographic variables also consist of a considerable missing value level. The models employed in this

study are based on training with the use of past data on the premise that the past predicts the future. Changes in conditions such as attitudes and economic climate over time will affect the accuracy of the system; therefore, the models may have to be reviewed to adapt to changes. Memory is one of the most significant challenges in NN and RF modelling. The limited memory bandwidth of the device that was used by the system to store the vast amounts of data, weights and activations NN proved to be a challenge. Computer memory capacity limitation during the analysis stage was dealt with by sampling the data. The default resolution of some of R Studio images such as NN topology is low.

Chapter 8

Conclusion and Recommendation

The customer data that was used had a few demographic variables, and it had five variables with missing values; some of these variables had large levels of missingness. This set was much polluted, therefore it required much cleaning. The deals data also contained a variable with missing values. Descriptive and summary statistics were used to explore the data. Cramer's V tests were used to test for relationships between categorical input variables and customers' price sensitivity. Correlations amongst numerical variables were also considered. The belief that variables *product* and *dda balance* are influential resulted in the imputation of these two variables. The imputation methods used were MissForest and decision trees. The models that were fitted in the datasets that include these imputed variables resulted in a slightly improved accuracy level.

After merging customer and deals datasets, the CART, RF and NN models were

fitted to classify customer price sensitivity. Models fitted were RF using twenty decision trees and the Resilient backpropagation NN with one hidden layer consisting of one node. Both models produced good results however the RF performed slightly better than NN in the given setting. When RF and NN were fitted in Dataset 4 they performed better than when fitted in other three datasets. NN fitted in Dataset 4 obtained a PCC of 0.98 and an AUC of 0.99 when tested on the test dataset, while RF fitted in Dataset 4 obtained a PCC of 0.99 and an AUC of 0.99.

In the sets that exclude the imputed variables all models showed that *agreement type*, *turnover per annum*, *average number of deals per annum* and *margin* were ranked the highest on the significant effect in predicting customer price *s*ensitivity. RF and Tree 7 (see Appendix B.7) suggest that the imputed *dda balance* was the most significant variable when fitted in Dataset 4; however NN assigned a small weight to this variable relative to the rest and assigned a high weight to *deal revenue*. The variables that were assigned high values of weights by an NN that produced the highest accuracy level are *deal revenue*, *turnover*, *agreement type*, *customer type*, *margin*, range, *deals per annum*, *product*. The categorical variables chosen by the models to have a significant impact on the response variable were also shown through Cramer's V in Table 5.24 to have an association with *s*ensitivity.

Both RF and NN techniques performed well giving high model accuracy in all four datasets. RF and DT have advantages of being able to reveal what the variable is able to contribute to the classification and their relative importance while the NN

are trained to assign small weights to irrelevant features. Based on the findings and conclusions presented, it would be recommended to use the combination of neural networks, decision trees and random forests to predict customer price sensitivity. Early, automatic detection of price sensitivity which can be prioritised and analysed further by foreign exchange experts can result to potential savings by preventing incorrect pricing..

Chapter 9

Summary and Future Research

In retail banking, FOREX dealing division is very competitive, the likelihood of losing customers due to incorrect pricing is very high. Pricing is a crucial element in determining profitability; however, it is not always easy to get it right. Using statistical models can help a business price the customers accordingly and also retain some customers by predicting the ones that are price sensitive and put their profiles on "specialised pricing" for simple execution of deals. In this research, the deployment of neural networks and random forests for predicting customer price sensitivity has proved to be helpful; furthermore, the variable importance feature proved to be valuable.

The deals dataset consists of variables such as income, account balances with high levels of missing values, a complete set of this valuable information would have been preferred. However, due to confidentiality it could not be obtained. It will therefore

be interesting to pursue a further internal study that includes these variables together with the variables of importance obtained from the RF and CART. The importance of data management cannot be over-emphasised; it would be valuable to improve the accuracy of the information as several errors were identified during data cleaning. Perhaps an additional future study to ascertain how sensitive the customers are to the actual changes in the exchange rates by considering a given exchange rate at the time of transacting and record whether the deal was a hit or a miss would be useful.

For instance, if a deal was a miss, meaning that the customer did not accept the rate; this may well mean that the customer is taking a view on the market and is possibly price sensitive. In this exercise prediction models could be built using variables that are deemed important by the models in this paper. Using the chosen variables may reduce the training time instead of using global input.

Bibliography

[1] Ahmad, I., Ansari, M. A., & Mohsin, S. (2008, April). *Performance comparison between backpropagation algorithms applied to intrusion detection in computer network systems.* In Proceedings of the 7th WSEAS International Conference on Applied Computer and Applied Computational Science (pp. 47-52). World Scientific and Engineering Academy and Society (WSEAS).

[2] Anderson, N. (2017). *Implication of junk status for South African investors.* `https://www.thesouthafrican.com/read-implications-of-junk-status-for-south-african-investors` Retrieved: 16 April 2016

[3] Bengio, Y. (2009). *Learning deep architectures for AI.* Foundations and trends in Machine Learning, 2(1), 1-127.

[4] Bhambri, V. (2011). *Application of data mining in banking sector* IJCST,2(2).

[5] Boulesteix, A. L., Janitza, S., Kruppa, J., and König, I. R. (2012). *Overview of RF methodology and practical guidance with emphasis on computational biology and bioinformatics.*

Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(6), 493-507.

[6] Buckinx, W., & Van den Poel, D. (2005). *Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting* European Journal of Operational Research, 164(1), 252-268.

[7] Breiman L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees.* CRC press.

[8] Chen, C. P., and Zhang, C. Y. (2014). *Data-intensive applications, challenges, techniques and technologies: A survey on Big Data.* Information Sciences, 275, 314-347.

[9] Cheng, B., and Titterington, D. M. (1994). *Neural networks: A review from a statistical perspective.* Statistical science, 2-30.

[10] Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). *Random forests for classification in ecology.* Ecology, 88(11), 2783-2792.

[11] Dass, R. (2006). *Data mining in banking and finance: A note for bankers. Indian Institute of Management Ahmadabad*

[12] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *From data mining to knowledge discovery in databases* AI magazine, 17(3), 37.

[13] Fawcett, T. (2006). *An introduction to ROC analysis.* Pattern recognition letters, 27(8), 861-874.

[14] Hanley, J. A., and McNeil, B. J. (1982). *The meaning and use of the area under a receiver operating characteristic (ROC) curve.* Radiology, 143(1), 29-36.

[15] Hastie, T.; Tibshirani, R.; Friedman, J. (2001). *The elements of statistical learning. Data mining, inference, and prediction.* Springer Series in Statistics. Springer-Verlag, New York, 2001, p.588 . ISBN: 0-387-95284-5. Retrieved from:
`http://statweb.stanford.edu/ tibs/ElemStatLearn/printings/ESLII_print10.pdf`
Retrieved: 14 April 2017

[16] Hormozi, A. M., & Giles, S. (2004). *Data mining: A competitive weapon for banking and retail industries* Information systems management, 21(2), 62-71.

[17] Jayasree, V., & Balan, R. V. S. (2013). *A review on data mining in banking sector* American Journal of Applied Sciences, 10(10), 1160.

[18] Joy, M., Rusnák, M., Smidkova, K., & Vasicek, B. (2015). *Banking and currency crises: differential diagnostics for developed countries.*

[19] Lewis, R. J. (2000, May). *An introduction to classification and regression tree (CART) analysis* In Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California (pp. 1-14).

[20] Liaw, A. and Weiner M. (2002): *Classification and Regression by Random Forest.* R News Vol. 2/3 (December 2002), pp. 18-22.

[21] Liu, Y., & Schumann, M. (2005). *Data mining feature selection for credit scoring models* Journal of the Operational Research Society, 56(9), 1099-1108.

[22] Livni, R., Shalev-Shwartz, S., and Shamir, O. (2014). *On the computational efficiency of training neural networks.* In Advances in Neural Information Processing Systems (pp. 855-863).

[23] Moin, K. I., & Ahmed, D. Q. B. (2012). *Use of data mining in banking* International Journal of Engineering Research and Applications, 2(2,738-742).

[24] Ngai, E. W., Xiu, L., & Chau, D. C. (2009). *Application of data mining techniques in customer relationship management: A literature review and classification.* Expert systems with applications, 36(2), 2592-2602.

[25] Phua, C., Alahakoon, D., & Lee, V. (2004). *Minority report in fraud detection: classification of skewed data.* Acm sigkdd explorations newsletter, 6(1), 50-59.

[26] Piramuthu, S. (2004). *Evaluating feature selection methods for learning in data mining applications.* European journal of operational research, 156(2), 483-494.

[27] Pulakkazhy, S., & Balan, R. V. S. (2013). *Data mining in banking and its applications-a review.*

[28] Rahm, E., & Do, H. H. (2000). *Data cleaning: Problems and current approaches. IEEE Data Eng.* Bull., 23(4), 3-13.

[29] Reuters. (2016, February). Daily FX volumes rise to $4.84 trillion in Jan - CLS Retrieved from http://www.reuters.com/article/global-forex-volumes-idUSL8N15P2VA [30 July 2016]

[30] Siroky, D. S. (2009). *Navigating RFs and related advances in algorithmic modeling.* Statistics Surveys, 3, 147-163.

[31] Smith, K. A., and Gupta, J. N. (2000). *Neural networks in business: techniques and applications for the operations researcher.* Computers and Operations Research, 27(11), 1023-1044.

[32] Tang, F., & Ishwaran, H. (2017). *Random forest missing data algorithms.* Statistical Analysis and Data Mining: The ASA Data Science Journal.

[33] Timofeev, R. (2004). *Classification and regression trees (CART) theory and applications (Masters Thesis).* Humboldt University, Berlin

[34] Tu, J. V. (1996). *Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes.* Journal of clinical epidemiology, 49(11), 1225-1231.

[35] Williams, G. (2011). *Data mining with Rattle and R: The art of excavating data for knowledge discovery.* Springer Science & Business Media

[36] Yamini, O. and Prof. S. Ramakrishna (2015). *International Journal of Engineering Research & Technology.*

`http://www.ijert.org/view-pdf/14161/a-study-on-advantages-of-data-mining-classification-techniques` Retrieved: 17 June 2016

[37] Yohannes, Y., and Hoddinott, J. (1999). *Classification and regression trees: an introduction.* International Food Policy Research Institute, 2033

[38] Zekić-Sušac, M., Pfeifer, S., and Šarlija, N. (2014). *A Comparison of Machine Learning Methods in a High-Dimensional Classification Problem.* Business Systems Research Journal, 5(3), 82-96.

Appendix A

Deals and Customer data

A.1 Summary statistics

Customer No.		Sensitivity		Deal Type		Dealt Currency	
Min.	214	Not sensitive	767752	Purchase	706443	USD	585479
1st Qu.	9705574	Sensitive	280823	Sell	342132	EUR	201502
Median	11071233					GBP	108562
Mean	14576629					ZAR	69347
3rd Qu.	23695452					AUD	22872
Max.	45666072					BWP	12584
						(Other)	48229
Dealt Amount		Counter Currency		Counter Amount		Turnover	
Min.	-226000000	ZAR	978084	Min.	-328608735	Min.	75435
1st Qu.	-150	USD	18847	1st Qu.	-6803	1st Qu.	1420
Median	200	GBP	18399	Median	-1180	Median	5000
Mean	-961	EUR	9497	Mean	-3188	Mean	37080
3rd Qu.	1200	AUD	9240	3rd Qu.	1176	3rd Qu.	17402
Max.	40000000	INR	4101	Max.	343040084	Max.	300000000
		(Other)	10407				
Agreement Type		Spot Exchange Rate		Effective Exchange Rate		Deal Revenue	
Fixed	103791	Min.	-0.015	Min.	0.061	Min.	1585988.5
Optional	352	1st Qu.	7.162	1st Qu.	7.108	1st Qu.	29.0
Partial	581383	Median	8.639	Median	8.580	Median	108.1
Spot	363049	Mean	9.068	Mean	9.003	Mean	474.1
		3rd Qu.	10.295	3rd Qu.	10.278	3rd Qu.	354.7
		Max.	5000.000	Max.	4849.515	Max.	1563200.0

Table A.1: Deals data summary - 1

Customer Segment		Dealing Desk		FEC		USD Spot Rate		Range	
Bus banking	131778	I	247383	N	1045508	Min.	-0.015	Min.	1.000
Commercial	38917	N	767752	Y	3067	1st Qu.	7.080	1st Q	1.000
Consumer	817182	O	25256			Median	8.080	Media	1.000
Wealth	60698	Y	8184			Mean	8.119	Mean	1.834
						3rd Qu.	8.785	3rdQu	2.000
						Max.	15.920	Max.1	4.000

Margin		Product	
Min.	-1.07800	Notes	323114
1st Qu.	0.01700		300070
Median	0.02130	Inward swift	198380
Mean	0.02097	Outward swift	133406
3rd Qu.	0.02400	Cash Passport	41092
Max.	1.00000	Foreign cheques	22187
		(Other)	30326

Table A.2: Deals data summary before - 2

No.	Variable	No. of zeros	(%) zeros	No. of NAs	(%) NAs	Type	Unique
1	Customer Number	0	0	0	0	integer	151739
2	Sensitivity	0	0	0	0	factor	2
3	Deal Type	0	0	0	0	factor	2
4	Dealt Currency	0	0	0	0	factor	3
5	Dealt Amount	602	0.07	0	0	integer	42792
6	Counter Currency	0	0	0	0	factor	1
7	Counter Amount	174	0.02	0	0	integer	163473
8	Turnover	174	0.02	0	0	integer	117524
9	Agreement Type	0	0	0	0	factor	4
10	Spot Exchange Rate	0	0	0	0	numeric	87898
11	Effective Exchange Rate	0	0	0	0	numeric	91177
12	Deal Revenue	3957	0.44	0	0	numeric	149403
13	Customer Segment	0	0	0	0	factor	4
14	Dealing Desk	0	0	0	0	factor	4
15	Forward Exchange Contract	0	0	0	0	factor	2
16	USD Spot Rate	0	0	0	0	numeric	37369
17	Range	0	0	0	0	integer	14
18	Margin	5461	0.61	0	0	numeric	983
19	Product	0	0	0	0	factor	10

Table A.3: Description of variables in the deals data

Customer No.		Customer Type		Branch Code		Age		Pre dealing Indicator	
Min.	214	PI	387727	Min.	1.0	Min.	0.00	No	445791
1st Qu.	7769040	CC	18934	1st Qu.	135.0	1st Qu.	29.00	Yes	146
Median	15922665	Pty Ltd	17567	Median	448.0	Median	39.00		
Mean	16850725	SP	7941	Mean	733.6	Mean	39.92		
3rd Qu.	25203599	WBI	6460	3rd Qu.	855.0	3rd Qu.	53.00		
Max.	45921390	NPC	1659	Max.	9943.0	Max.	116.00		
		(Other)	5649	NA's	6				
Education Level		Customer Segment		Income		Bank VIP		Deals per annum	
	58537		380405	Min.	0	No	443548	Min.	1.00
G0 - G12	285557	Business banking	2295	1st Qu	70540	Yes	2389	1st Qu.	1.00
Post Matric	101843	Commercial	767	Median	256700			Median	1.33
		Commercial GTS	54	Mean	3661000			Mean	3.07
		Consumer	54100	3rd Qu.	547900			3rd Qu.	2.67
		Corporate	6	Max.	1000000000000			Max.	79928.60
		Wealth	8310						
Years with the bank		Online indicator		Salalary with the bank		Points Incentive Indicator		DDA Account	
Min.	0.00	No	128290	No	310737	No	116657	One and below	248696
1st Qu.	6.00	Yes	317647	Yes	135200	Yes	329280	Two and above	197241
Median	12.00								
Mean	13.96								
3rd Qu.	22.00								
Max.	116.00								
DDA Balance		TDA Account		TDA Balance		Card Account		Turnover per annum	
Min.	-91319	One and above	88701	Min.	0	Equal to zero	287142	Min.	0
1st Qu.	1066	Zero	357236	1st Qu	1172	One or more	158795	1st Qu.	3760
Median	7541			Median	12493			Median	14150
Mean	120609			Mean	207606			Mean	162700
3rd Qu.	40816			3rd Qu.	102512			3rd Qu.	59420
Max.	196736195			Max.	282479349			Max.	2346000000
NA's	75555			NA's	357236				
Card Balance		Global Account Indicator		XXX banked		Risk Cat. Code		FX VIP	
Min.	-4051916	No	440433	No	75555	Fraud	1	No	444421
1st Qu.	336	Yes	5504	Yes	370382	Insolv	5010	Yes	1516
Median	8069					Intensive	10553		
Mean	18363					Investi	35957		
3rd Qu.	26222					Normal	384566		
Max.	651818					Special	9850		
NA's	287142								

Table A.4: Customer data summary

No.	Variable	No. of zeros	Perc of zeros	No. of NAs	(%) NAs	type	unique
1	Customer Number	0	0	0	0	integer	445937
2	Customer Type	0	0	0	0	factor	23
3	Customer Branch Code	0	0	0	0	integer	986
4	Customer Age	41826	9.38	6	0	integer	117
5	Education Level Recoded	0	0	0	0	factor	3
6	Customer Segment	0	0	0	0	factor	7
7	Income	73669	16.52	0	0	numeric	272307
8	Bank VIP	0	0	0	0	factor	2
9	Years banking with the bank (bank XXX)	7574	1.7	0	0	integer	106
10	Online Indicator	0	0	0	0	factor	2
11	Salary with the bank	0	0	0	0	factor	2
12	Points incentive Indicator	0	0	0	0	factor	2
13	Transactional Account (DDA) Account	0	0	0	0	factor	2
14	Transactional Account (DDA) Balance	21002	4.71	75555	16.94	numeric	313404
15	Investment Account (TDA)	0	0	0	0	factor	2
16	Investment Account (TDA) Balance	177	0.04	357236	80.11	numeric	81568
17	Credit Card Account	0	0	0	0	factor	2
18	Credit Card Balance	3775	0.85	287142	64.39	numeric	147808
19	Global Account Indicator	0	0	0	0	factor	2
20	Banking with the bank (bank XXX banked)	0	0	0	0	factor	2
21	Risk Category Code	0	0	0	0	factor	6
22	Foreign Exchange VIP	0	0	0	0	factor	2
23	Pre-Dealing Indicator	0	0	0	0	factor	2
24	Average no. of deals per annum	0	0	0	0	numeric	724
25	Turnover per annum	0	0	0	0	numeric	398009

Table A.5: Description of variables in the Customer Data

No	Customer Type	Frequency	Percentage	Cumulative Percentage
1	1 = PRIVATE INDIVIDUAL	387727	86.95	86.95
2	4 = CLOSE CORPORATION - CC	18934	4.25	91.2
3	5= PRIVATE COMPANY - (PTY) LTD	17567	3.94	95.14
4	2 = SOLE PROPRIETORSHIP	7941	1.78	96.92
5	14 = XYZ BANK INDIVIDUAL	6460	1.45	98.37
6	7 = NON PROFIT COMPANIES - NPC	1659	0.37	98.74
7	19 = OTHER ASSOCIATIOND / CLUB / ETC	1656	0.37	99.11
8	9 = REGISTERED TRUST	1492	0.33	99.44
9	20 = JOINT HOME LOANS	798	0.18	99.62
10	3 = PARTNERSHIP	660	0.15	99.77
11	8 = PERSONAL LIABILITY COMPANY - INC	492	0.11	99.88
12	6 = PUBLIC COMPANY - LTD	149	0.03	99.91
13	12 = GOVERNMENT/QUASI GOVERNMENT	148	0.03	99.94
14	22 = FOREIGN COMPANY NOT REGISTERED IN SA	86	0.02	99.96
15	10 = FOUNDATION	40	0.01	99.97
16	31 = FNB ASSOCIATED	37	0.01	99.98
17	21 = FOREIGN COMPANY REGISTERED IN SA	36	0.01	99.99
18	11 = CO-OPERATIVE	28	0.01	100
19	18 = REGISTERED FREINDLY SOCIETY	15	0	100
20	13 = STOKVEL	9	0	100
21	0 = OFFICE ACCOUNTS	1	0	100
22	35 = FOREIGN BANKS	1	0	100
23	88 = RESERVED ACCOUNTS PROFILE	1	0	100

Table A.6: Customer Type frequency table

No	Branch Code	Frequency	Percentage	Cumulative Percentage
1	5276	11816	2.65	2.65
2	132	6414	1.44	4.09
3	620	6384	1.43	5.52
4	144	6239	1.4	6.92
5	635	5890	1.32	8.24
6	2861	5743	1.29	9.53
7	923	5056	1.13	10.66
8	838	4871	1.09	11.75
9	824	4646	1.04	12.79
10	26	4554	1.02	13.81
.....	.....	.....	.....	.....
977	8310	1	0	99.75
978	8380	1	0	99.75
979	8626	1	0	99.75
980	8647	1	0	99.75
981	8675	1	0	99.75
982	9214	1	0	99.75
983	9258	1	0	99.75
984	9381	1	0	99.75
985	9400	1	0	99.75
986	9901	1	0	100

Table A.7: Branch Code frequency table

No	**High Education Level**	**Frequency**	**Percentage**	**Cumulative Percentage**
1	G12	259187	58.12	58.12
2		57307	12.85	70.97
3	BCH	33526	7.52	78.49
4	DIP	31883	7.15	85.64
5	G10	23191	5.2	90.84
6	HNR	14590	3.27	94.11
7	MST	14136	3.17	97.28

8	DOC	4645	1.04	98.32
9	L10	3124	0.7	99.02
10	GRD	3062	0.69	99.71
11	INC	765	0.17	99.88
12	*	450	0.1	99.98
13	G11	10	0	99.98
14	G8	10	0	99.98
15	G5	4	0	99.98
16	G9	4	0	99.98
17	G	3	0	99.98
18	G7	3	0	99.98
19	GR8	3	0	99.98
20	OTH	3	0	99.98
21	Y	3	0	99.98
22	G1	2	0	99.98
23	G3	2	0	99.98
24	G6	2	0	99.98
25	GR9	2	0	99.98
26		1	0	99.98
27	*10	1	0	99.98
28	*12	1	0	99.98
29	.	1	0	99.98
30	12	1	0	99.98
31	2	1	0	99.98
32	F10	1	0	99.98
33	F12	1	0	99.98
34	G01	1	0	99.98
35	G05	1	0	99.98
36	G08	1	0	99.98
37	G09	1	0	99.98
38	G4	1	0	99.98
39	GR4	1	0	99.98
40	GR7	1	0	99.98
41	GRA	1	0	99.98
42	L13	1	0	99.98
43	MI	1	0	99.98

44	P	1	0	99.98
45	P10	1	0	100

Table A.8: Highest Education Level frequency table

No	DDA Accounts	Frequency	Percentage	Cumulative Percentage
1	0	173141	38.83	38.83
2	1	146892	32.94	71.77
3	2	75555	16.94	88.71
4	3	35716	8.01	96.72
5	4	9953	2.23	98.95
6	5	2927	0.66	99.61
7	6	940	0.21	99.82
8	7	321	0.07	99.89
9	8	157	0.04	99.93
10	9	65	0.01	99.94
.....	.....	.....	.....	.....
55	77	1	0	99.97
56	81	1	0	99.97
57	85	1	0	99.97
58	86	1	0	99.97
59	91	1	0	99.97
60	111	1	0	99.97
61	148	1	0	99.97
62	152	1	0	99.97
63	207	1	0	99.97
64	258	1	0	99.97
65	577	1	0	100

Table A.9: DDA Account frequency table

No	TDA Accounts	Frequency	Percentage	Cumulative Percentage
1	0	357236	80.11	80.11
2	1	76000	17.04	97.15
3	2	9778	2.19	99.34
4	3	1902	0.43	99.77
5	4	568	0.13	99.9
6	5	186	0.04	99.94
7	6	108	0.02	99.96
8	7	56	0.01	99.97
9	8	26	0.01	99.98
10	9	14	0	99.98
.....	.....	.....	.....	.....
20	25	2	0	99.98
21	15	1	0	99.98
22	20	1	0	99.98
23	23	1	0	99.98
24	28	1	0	99.98
25	32	1	0	99.98
26	34	1	0	99.98
27	66	1	0	99.98
28	91	1	0	99.98
29	96	1	0	99.98
30	162	1	0	100

Table A.10: TDA Account frequency table

No	Card Accounts	Frequency	Percentage	Cumulative Percentage
1	0	287142	64.39	64.39
2	1	141109	31.64	96.03
3	2	13269	2.98	99.01
4	3	2359	0.53	99.54
5	4	718	0.16	99.7
6	5	546	0.12	99.82
7	6	235	0.05	99.87
8	7	174	0.04	99.91
9	8	109	0.02	99.93
10	9	73	0.02	99.95
.....	.....	.....	.....	.....
25	31	2	0	99.99
26	21	1	0	99.99
27	22	1	0	99.99
28	26	1	0	99.99
29	27	1	0	99.99
30	29	1	0	99.99
31	32	1	0	99.99
32	36	1	0	99.99
33	41	1	0	100

Table A.11: Card Account frequency table

A.2 Merged Data Summary

CUST YPE		Branch Code		Cust Age		Highest Ed Level	
LTD	398212	480	239378	Min.	0.00	G0 - G12	250115
PI	336445	5798	152032	1st Qu.	0.00	Post Matric	89825
CC	70332	5276	13149	Median	0.00	NA's	553285
Pty Ltd	43899	824	9549	Mean	20.62		
SP	14722	144	7740	3rd Qu.	42.00		
WBI	8651	132	7474	Max.	116.00		
(Other)	20964	(Other)	463903	NA's	26		
Income		bank VIP		Yrs with XXX		online ind	
Min.	0	No	886759	Min.	0.00	No	578498
1st Qu.	0	Yes	6466	1st Qu.	6.00	Yes	314727
Median	0			Median	14.00		
Mean	3700000			Mean	13.69		
3rd Qu.	240300			3rd Qu.	16.00		
Max.	1000000000000			Max.	116.00		
eXXX ind		DDA Acc		dda balance		TDA Acc RECODED	
No	375733	One and below	536489	Min.	-91319	One and above	90814
Yes	517492	Two and above	356736	1st Qu.	0	Zero	802411
				Median	4871		
				Mean	261442		
				3rd Qu.	62089		
				Max.	129733122		
				NA's	336810		
tda balance		Card Acc		card balance		XXX banked	
Min.	0	Equal to zero	697346	Min.	-3008416.9	No	336810
1st Qu.	3170	One or more	195879	1st Qu.	65.4	Yes	556415
Median	38532			Median	6659.0		
Mean	521471			Mean	16952.2		
3rd Qu.	261321			3rd Qu.	24229.2		
Max.	95164963			Max.	651818.4		
NA's	802411			NA's	697346		

Table A.12: Merged Data Summary - 1

risk cat cde		pre dealing ind		deal pa		turnover pa	
Fraud	1	No	890747	Min.	1.0	Min.	1
Insolv	3792	Yes	2478	1st Qu.	4.0	1st Qu.	73495
Intensive	13381			Median	23.2	Median	1567908
Investi	38386			Mean	19848.7	Mean	142495990
Normal	829274			3rd Qu.	29660.0	3rd Qu.	283670922
Special	8391			Max.	79928.6	Max.	548316704
Sensitivity		DEAL TYPE		DLT CURR		Turnover	
Not sensitive	637707	Purch	592251	EUR	200816	Min.	1
Sensitive	255518	Sell	300974	GBP	108221	1st Q	1372
				USD	584188	Mdn	4904
						Mean	29246
						3rd Q	17099
						Max	5000000
deal revenue		Cust Segment		usd spot Rate		range	
Min.	0.0	Business banking	114355	Min.	6.461	1	585968
1st Qu.	28.6	Commercial	32395	1st Qu.	7.085	3	107350
Median	102.6	Consumer	695313	Median	8.092	2	105671
Mean	423.4	Wealth	51162	Mean	8.132	4	48430
3rd Qu.	336.9			3rd Qu.	8.813	5	16357
Max.	496515.8			Max.	15.920	7	8178
						(Other)	21271
margin		product		salary with XXX		AGREE TYPE	
Min.	0.00000	Notes	287079	No	793026	Fixed	103194
1st Qu.	0.01680	Inward swift	185790	Yes	100199	Optional	322
Median	0.02110	Outward swift	65659			Partial	468971
Mean	0.02021	Cash Passport	36971			Spot	320738
3rd Qu.	0.02340	Foreign cheques	19931				
Max.	0.19950	(Other)	28640				
		NA's	269155				

Table A.13: Merged Data Summary - 2

	Minimum	First Quartile	Median	Mean	Third Quartile	Maximum	NA's
Age	0	0	0	20.62	42	116	26
Income	R -	R -	R -	R 3,700,000.00	R 240,000.00	R1,000,000,000,000.00	
Years with the bank	0	6	14	13.69	16	116	
DDA Balance	R -91,319.00	R -	R 4,871.00	R 261,442.00	R 62,089.00	R 129,733,122.00	336810
TDA Balance	0	3170	38532	521471	261321	95164963	802411
Card Balance	R -3,008,416.90	R 65.40	R 6,659.00	R 16,952.20	R 24,229.20	R 651,818.40	697346
Deal P/A	1	4	23.2	19848.7	29660	79928.6	
Turnover P/A	R 1.00	R73,495.00	R1,567,908.00	R142,000,000.00	R283,670,922.00	R 548,316,704.00	
Turnover	R 1.00		R 4,909.00	R 29,246.00	R 17,099.00	R 5,000,000.00	
Deal Revenue	R -	R 28.60	R 102.60	R 423.40	R 336.90	R 496,515.80	
USD Spot Rate	6.461	7.085	8.092	8.132	8.813	15.92	
Margin	0	0.0168	0.0211	0.02021	0.0234	0.1995	

Table A.14: Summary statistics for numerical variables in the Merged Data

	Frequency	Percentage	Cumulative Percentage
Private Individual	398212	44.58%	44.58%
Close Cooperation	336445	37.67%	82.25%
Private Company	70332	7.87%	90.12%
Sole Proprietorship	43899	4.91%	95.04%
WB Individuals	14722	1.65%	96.68%
Non-Profit Company	8651	0.97%	97.65%
(Other)	20964	2.35%	100.00%
Total	893225	100.00%	

Table A.15: Distribution of Customer Type

	Frequency	Percentage	Cumulative Percentage
G0 - G12	250115	28.0%	28.0%
Post Matric	89825	10.1%	38.1%
NA's	553285	61.9%	100.0%
	893225	100.0%	

Table A.16: Distribution of Education Level

	Frequency	Percentage	Cumulative Percentage
No	886759	99.28%	99.28%
Yes	6466	0.72%	100.00%
Total	893225	100.00%	

Table A.17: Distribution of bank VIP

	Frequency	Percentage	Cumulative Percentage
No	578498	64.77%	64.77%
Yes	314727	35.23%	100.00%
Total	893225	100.00%	

Table A.18: Distribution of Online Indicator

	Frequency	Percentage	Cumulative Percentage
No	793026	88.78%	88.78%
Yes	100199	11.22%	100.00%
Total	893225	100.00%	

Table A.19: Distribution of Salary With The Bank

	Frequency	Percentage	Cumulative Percentage
No	375733	42.06%	42.06%
Yes	517492	57.94%	100.00%
Total	893225	100.00%	

Table A.20: Distribution of Points Incentive Indicator

	Frequency	Percentage	Cumulative Percentage
One and below	536489	60.06%	60.06%
Two and above	356736	39.94%	100.00%
Total	893225	100.00%	

Table A.21: Distribution of DDA Account

	Frequency	Percentage	Cumulative Percentage
None	802411	89.83%	89.83%
One and above	90814	10.17%	100.00%
Total	893225	100.00%	

Table A.22: Distribution of TDA Account

	Frequency	Percentage	Cumulative Percentage
None	697346	78.07%	78.07%
One and above	195879	21.93%	100.00%
Total	893225	100.00%	

Table A.23: Distribution of Card Account

	Frequency	Percentage	Cumulative Percentage
No	336810	37.71%	37.71%
Yes	556415	62.29%	100.00%
Total	893225	100.00%	

Table A.24: Distribution of Banked With The Bank

	Frequency	Percentage	Cumulative Percentage
Fraud	1	0.00%	0.00%
Insolvent	3792	0.42%	0.42%
Intensive	13381	1.50%	1.92%
Investigation	38386	4.30%	6.22%
Normal	829274	92.84%	99.06%
Special	8391	0.94	100.00%
Total	893225	100.00%	

Table A.25: Distribution of Risk Category Code

	Frequency	Percentage	Cumulative Percentage
No	890747	99.72%	99.72%
Yes	2478	0.28%	100.00%
Total	893225	100.00%	

Table A.26: Distribution of Pre-Dealing Indicator

	Frequency	Percentage	Cumulative Percentage
Not sensitive	637707	71.39%	71.39%
Sensitive	255518	28.61%	100.00%
Total	893225	100.00%	

Table A.27: Distribution of Sensitivity

	Frequency	Percentage	Cumulative Percentage
Purchase	592251	66.30%	66.30%
Sell	300974	33.70%	100.00%
Total	893225	100.00%	

Table A.28: Distribution of Deal Type

	Frequency	Percentage	Cumulative Percentage
EUR	200816	22.48%	22.48%
GBP	108221	12.12%	34.60%
USD	584188	65.40%	100.00%
Total	893225	100.00%	

Table A.29: Distribution of Dealt Currency

	Frequency	Percentage	Cumulative Percentage
Fixed	103194	11.55%	11.55%
Optional	322	0.04%	11.59%
Partial	468971	52.50%	64.09%
Spot	320738	35.91%	100.00%
Total	893225	100.00%	

Table A.30: Distribution of Agreement Type

	Frequency	Percentage	Cumulative Percentage
Business Banking	114355	12.80%	12.80%
Commercial	32395	3.63%	16.43%
Consumer	695313	77.84%	94.27%
Wealth	51162	5.73%	100.00%
Total	893225	100.00%	

Table A.31: Distribution of Customer Segment

	Frequency	Percentage	Cumulative Percentage
1	585968	65.60%	65.60%
3	107350	12.02%	77.62%
2	105671	11.83%	89.45%
4	48430	5.42%	94.87%
5	16357	1.83%	96.70%
7	8178	0.92%	97.62%
Other	21271	2.38%	100.00%
Total	893225	100.00%	

Table A.32: Distribution of Range

Notes	287079	32.14%	32.14%
Inward swift	185790	20.80%	52.94%
Outward swift	65659	7.35%	60.29%
Cash passport	36971	4.14%	64.43%
Foreign cheques	19931	2.23%	66.66%
Other	28640	3.21%	69.87%
NAs	269155	30.13%	100.00%
	893225	100.00%	

Table A.33: Distribution of Product

A.2.1 Deals Data graphs

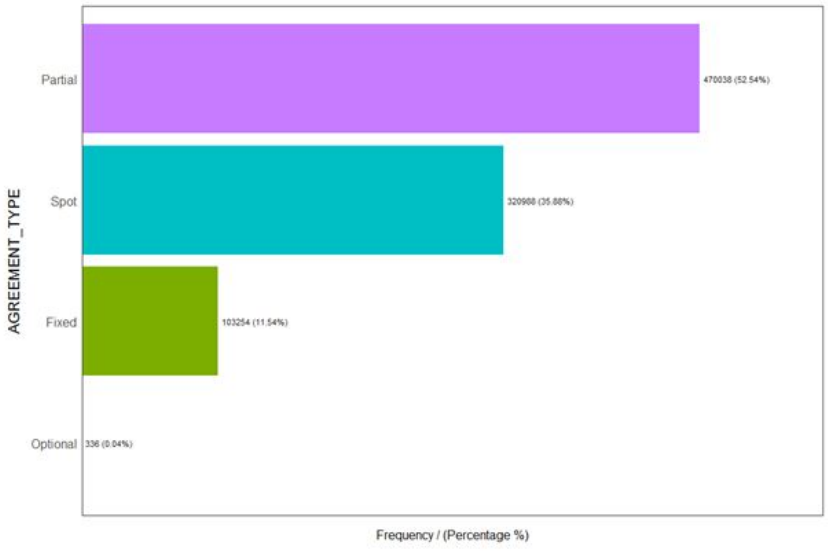

Figure A.1: Agreement Type distribution

Figure A.2: Distribution of Counter Currency

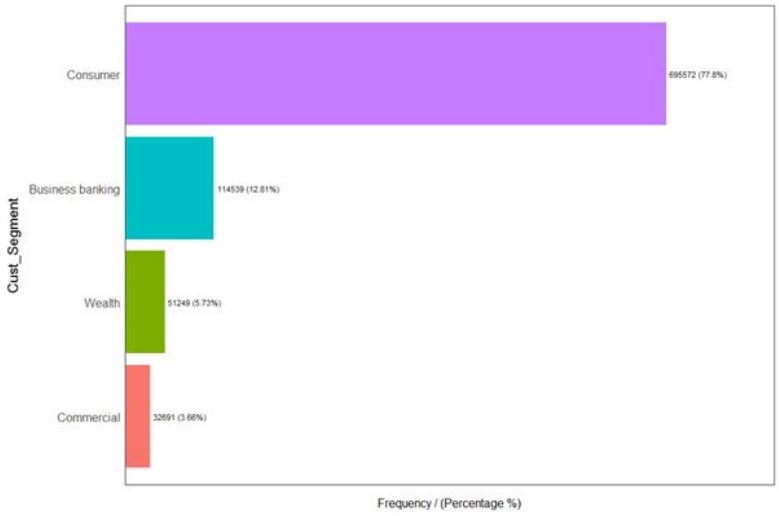

Figure A.3: Distribution of Customer Segment

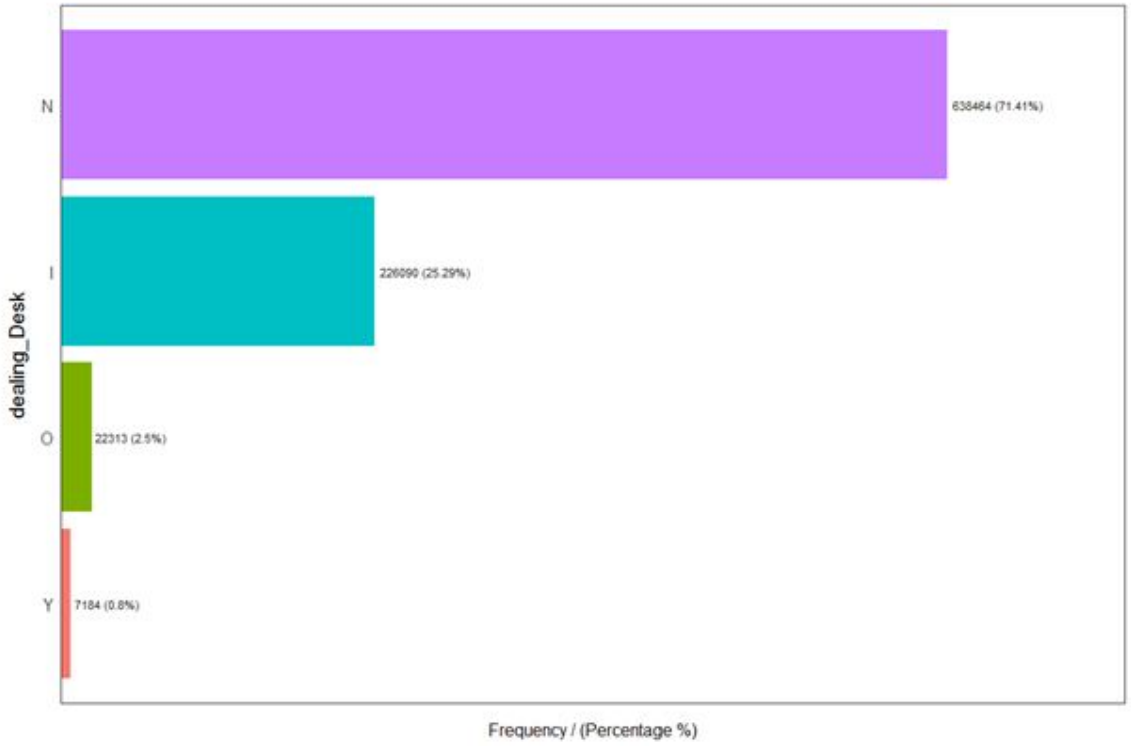

Figure A.4: Distribution of Dealing Desk

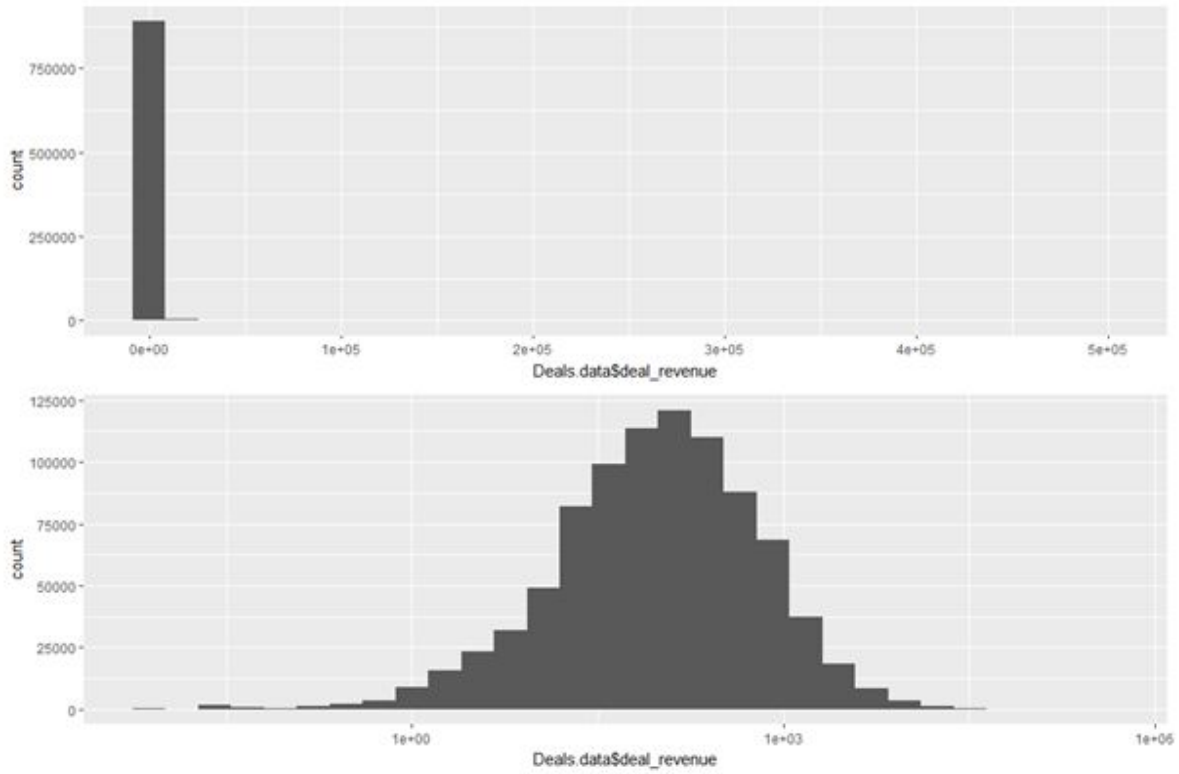

Figure A.5: Distribution of Deal Revenue

Figure A.6: Distribution of Dealt Currency

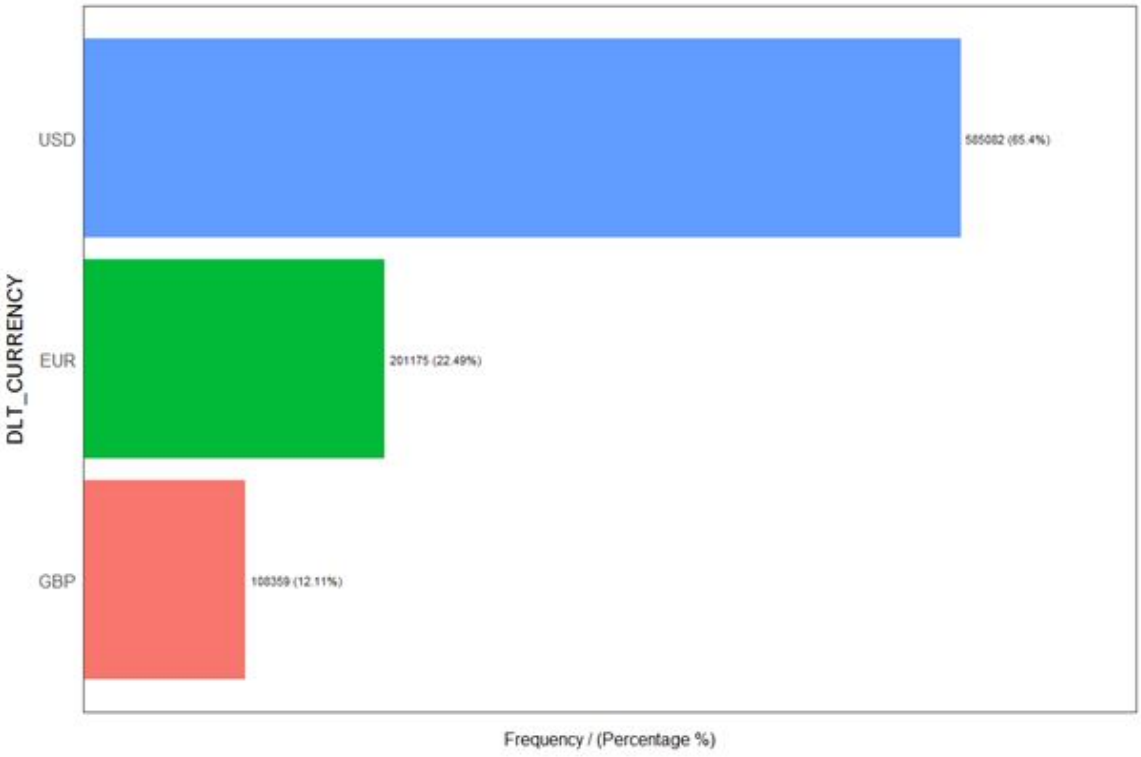

Figure A.7: Distribution of Cleaned Dealt Currency

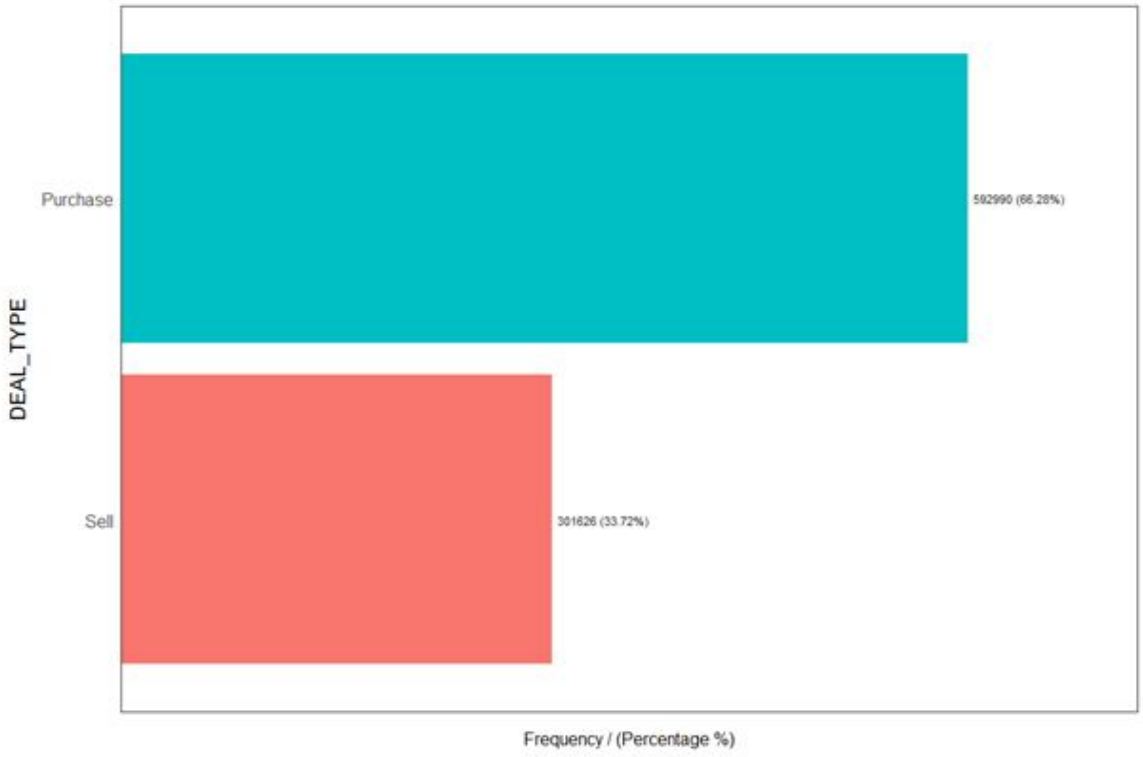

Figure A.8: Distribution of Deal Type

Figure A.9: Distribution of FEC

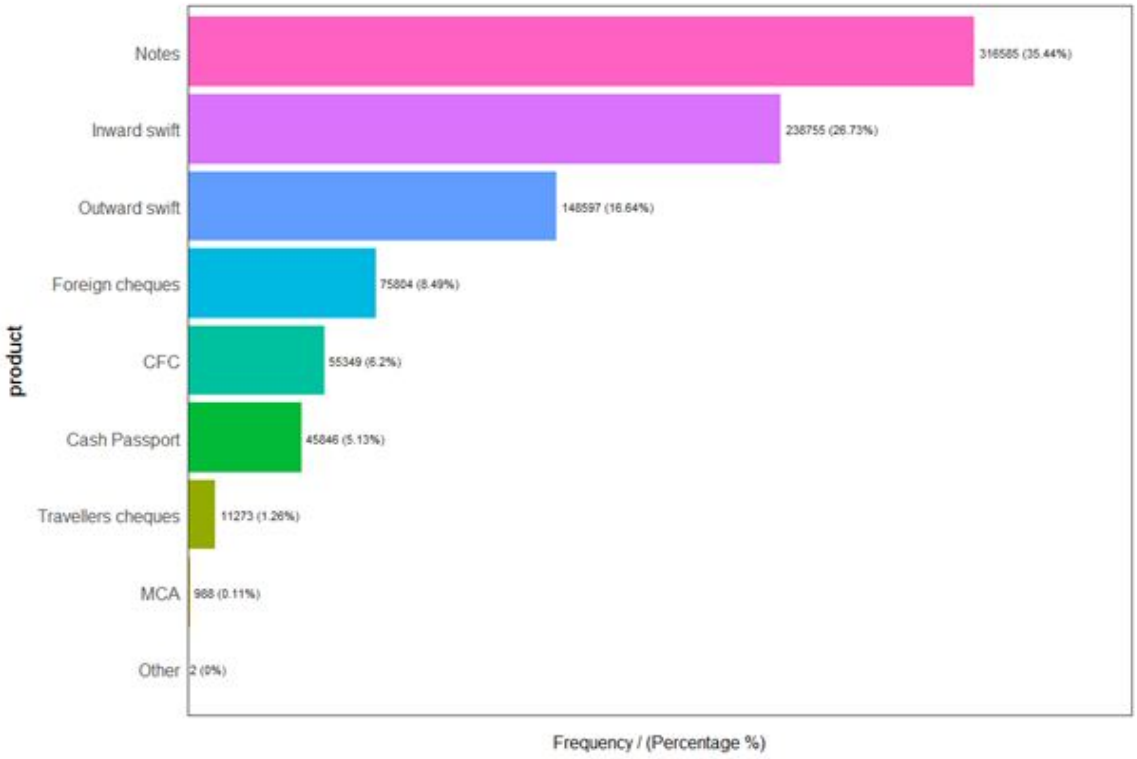

Figure A.10: Distribution of Imputed Product

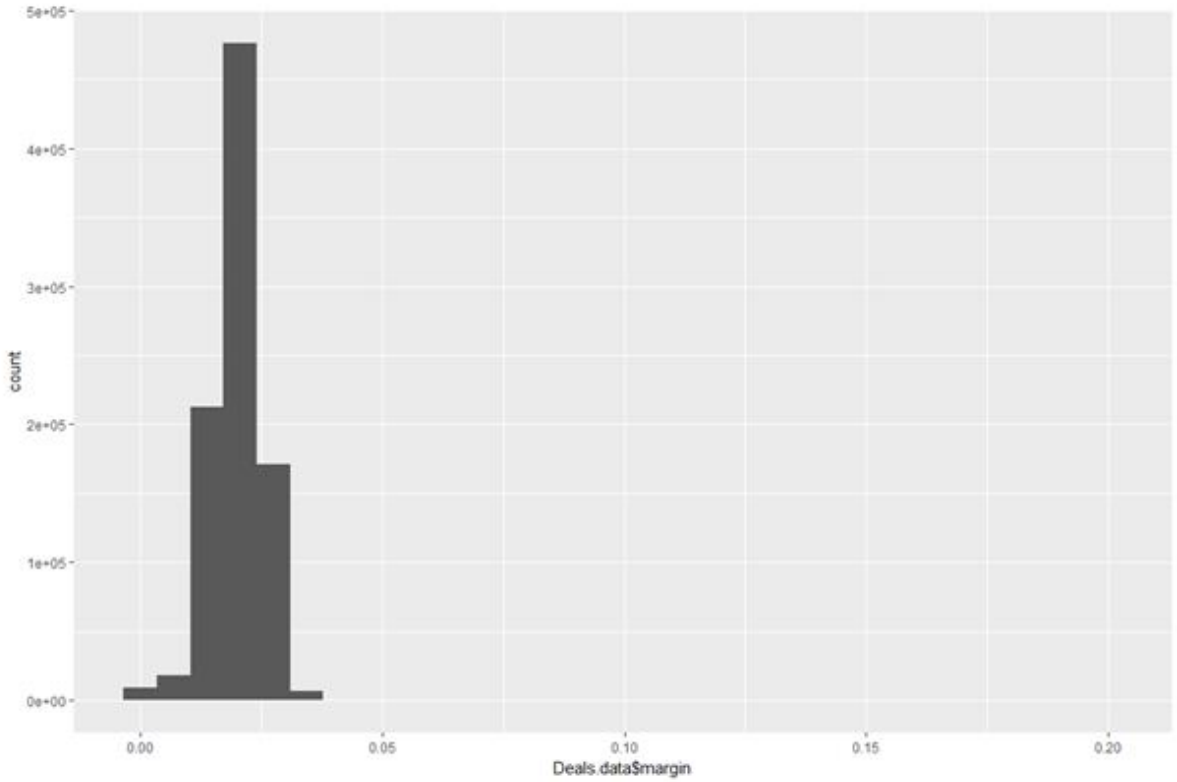

Figure A.11: Distribution of Margin

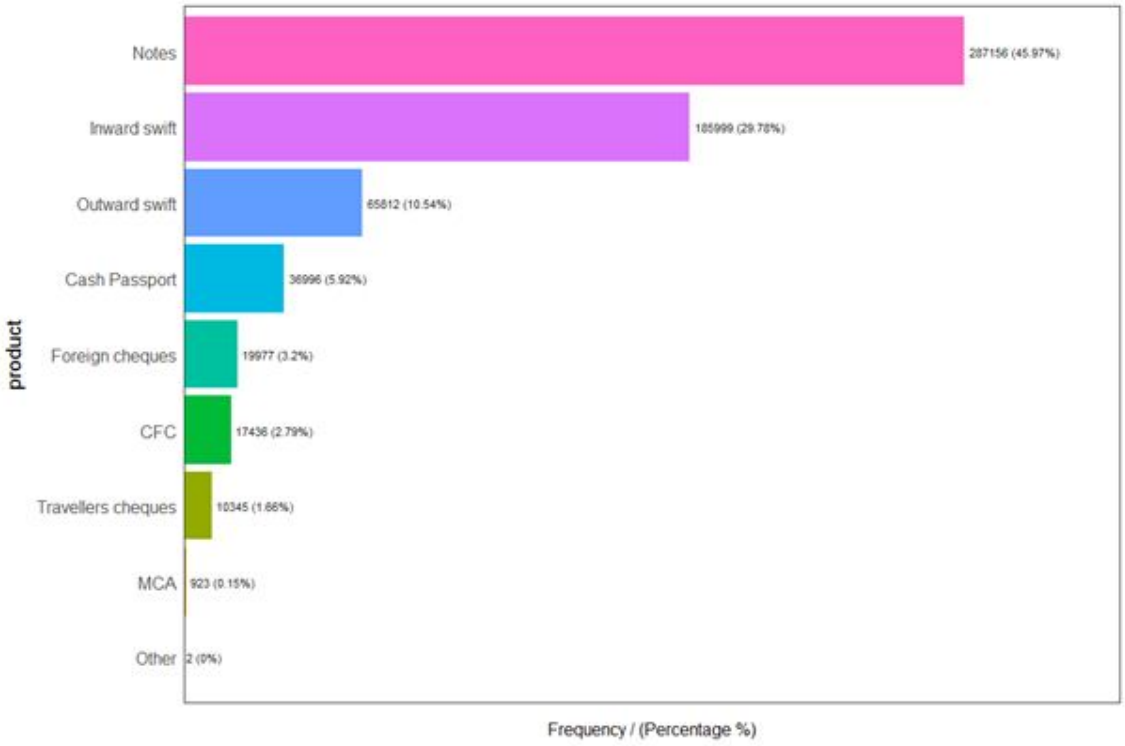

Figure A.12: Distribution of Product

Figure A.13: Distribution of Range

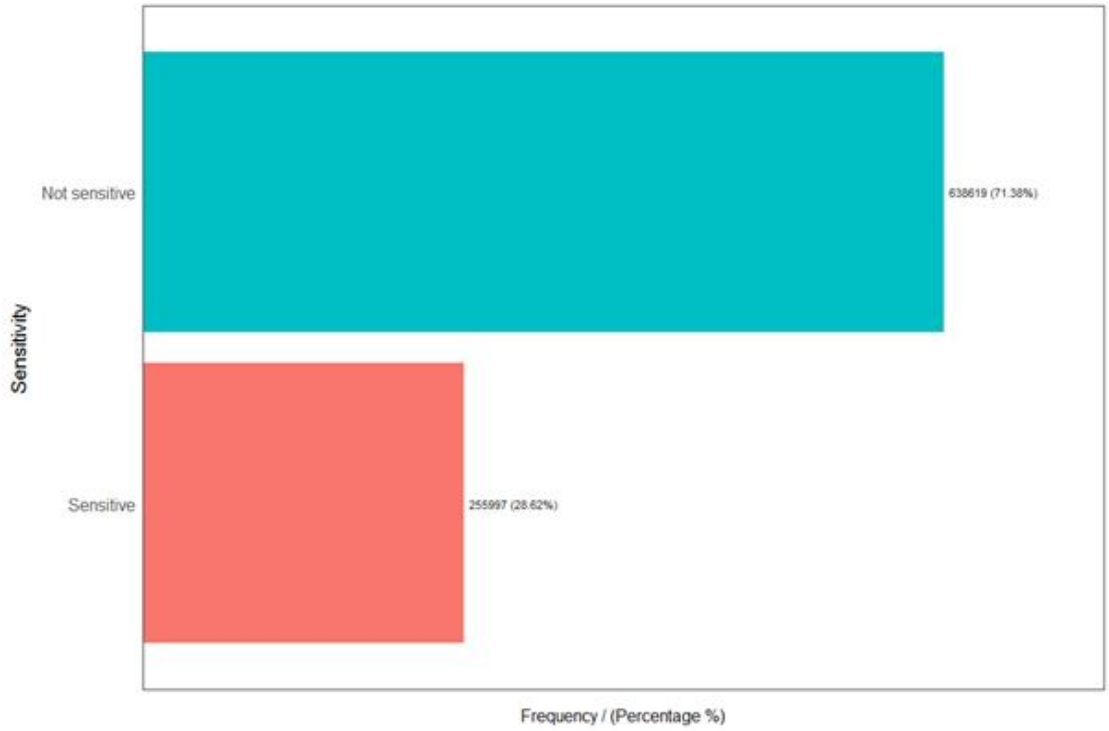

Figure A.14: Distribution of Sensitivity

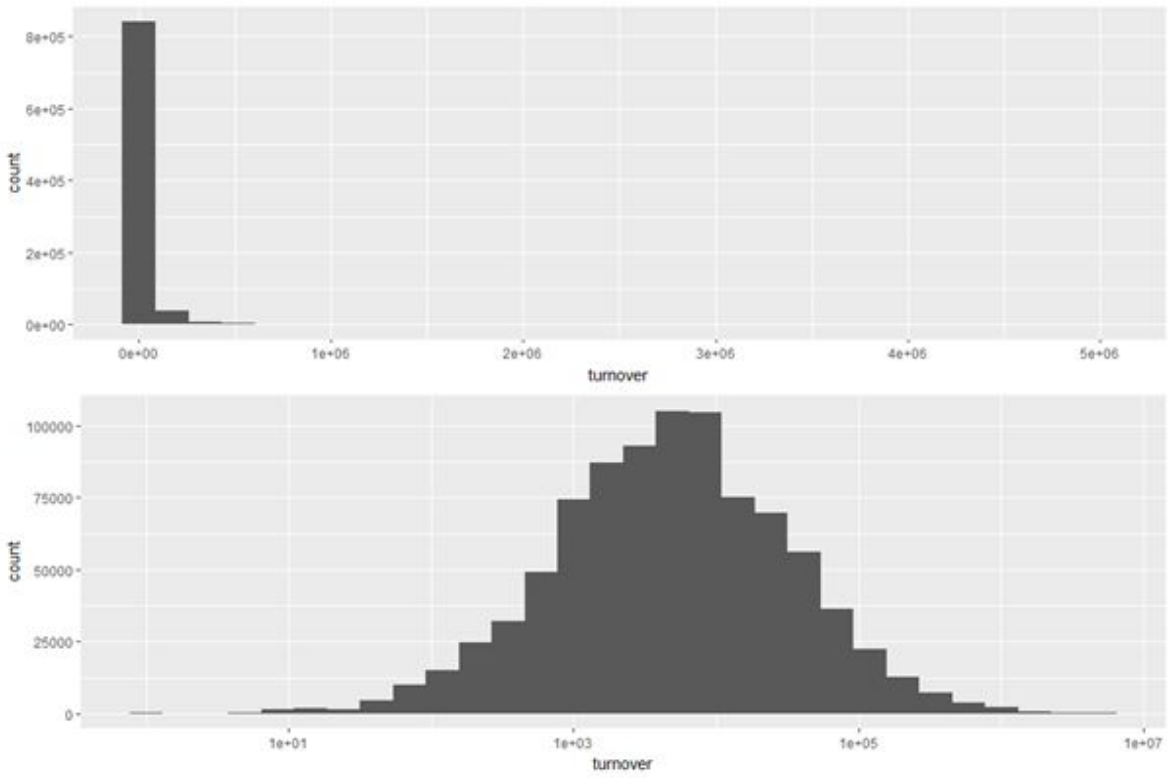

Figure A.15: Distribution of Turnover

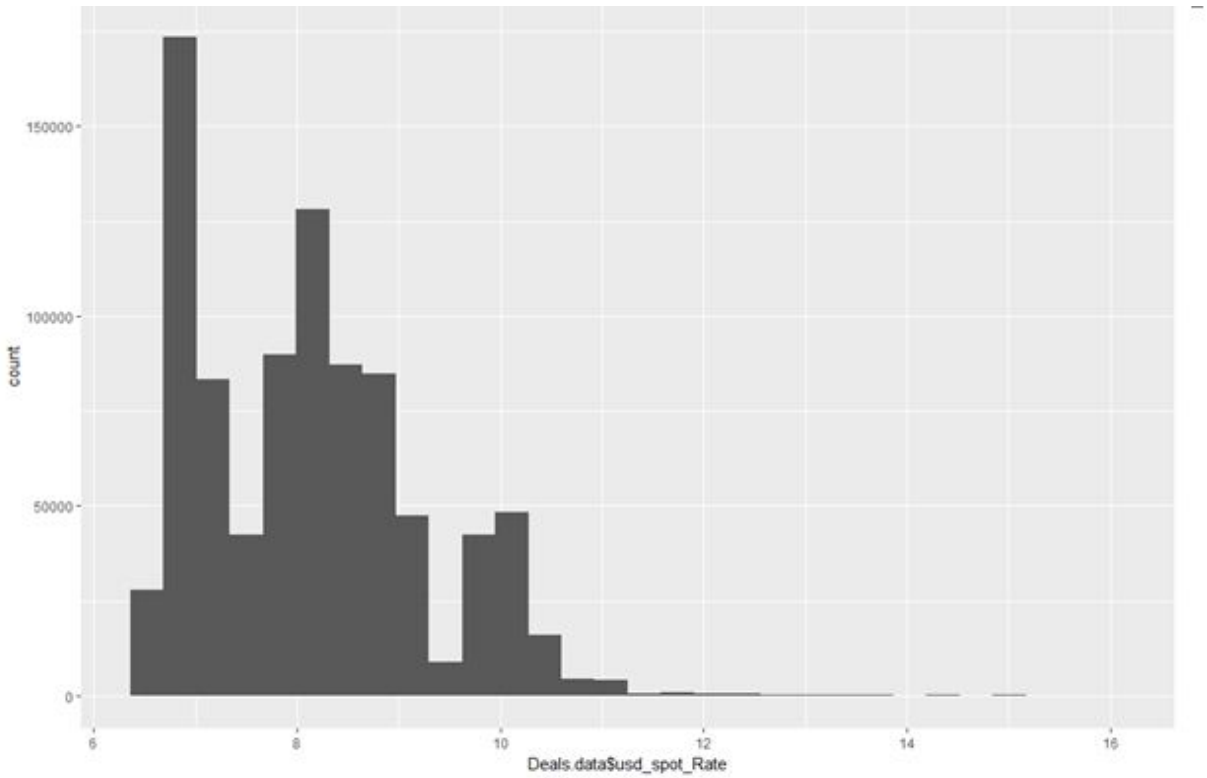

Figure A.16: Distribution of USD Spot Rate

A.2.2 Customer Data graphs

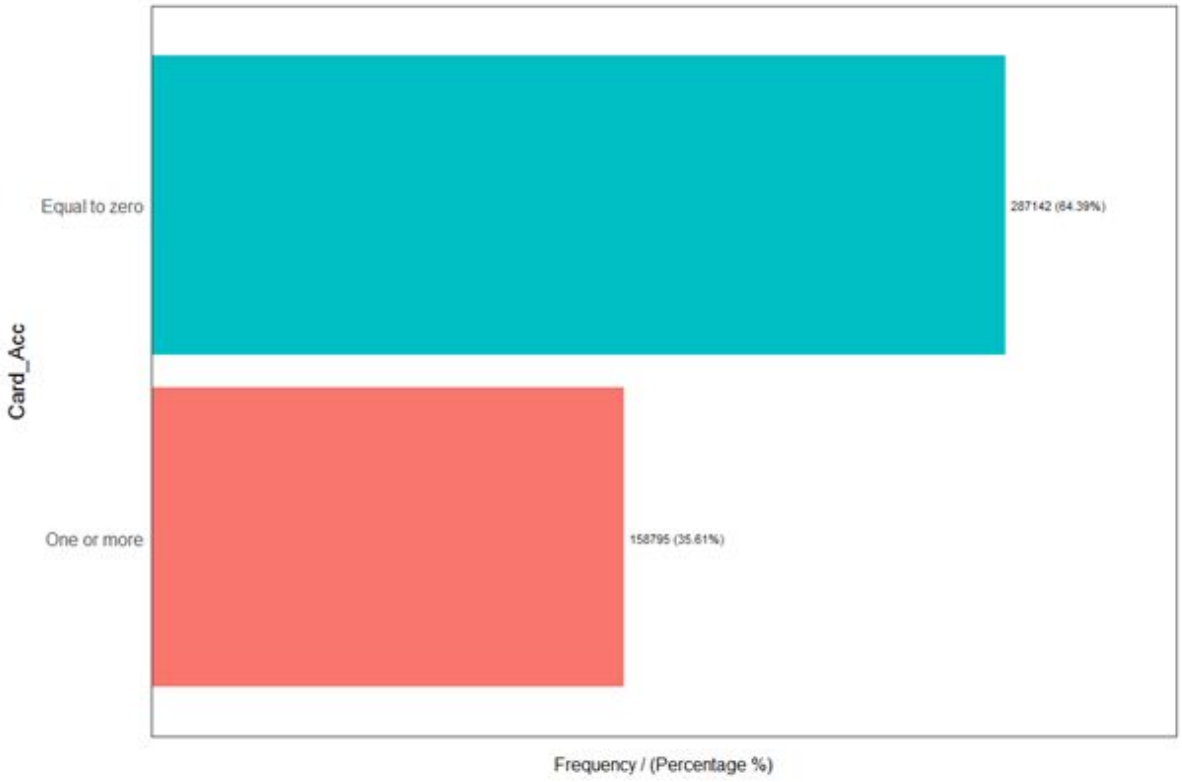

Figure A.17: Distribution of Card Account

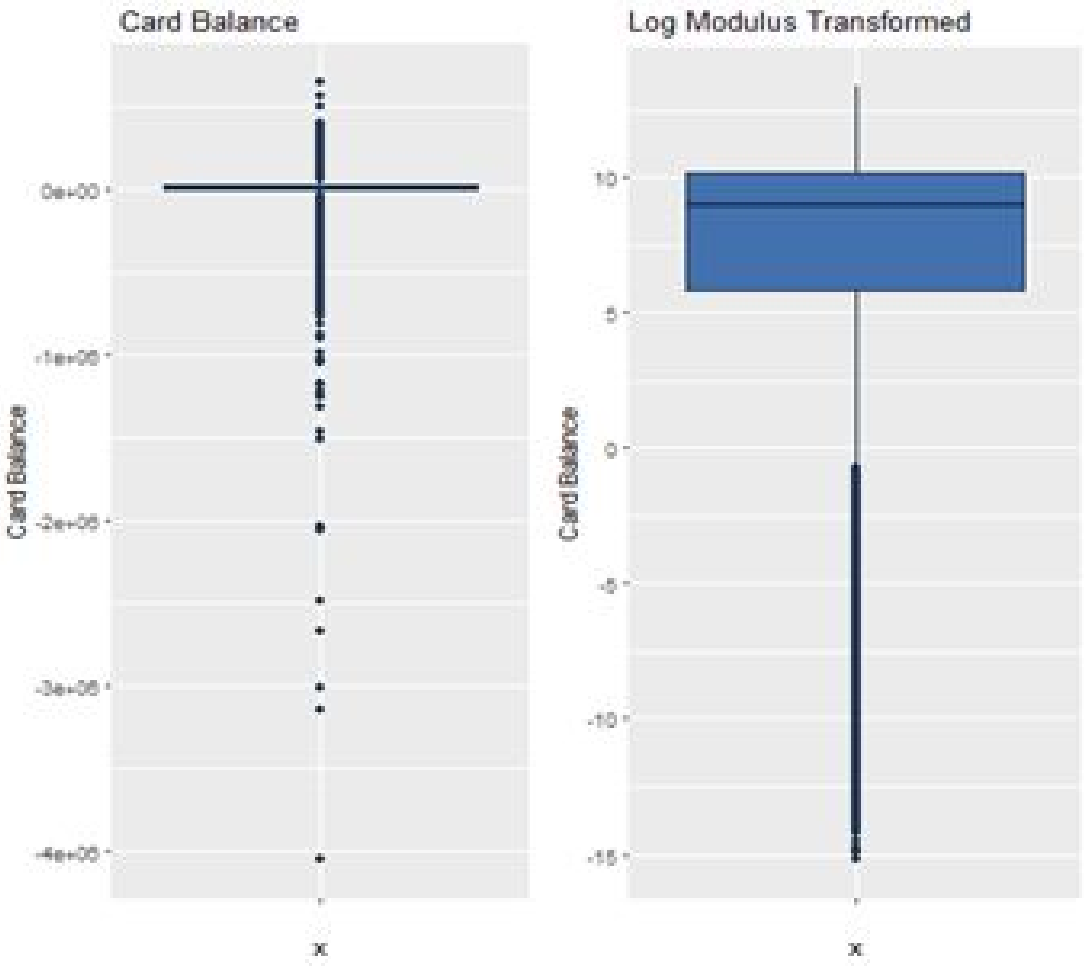

Figure A.18: Distribution of Card Balance

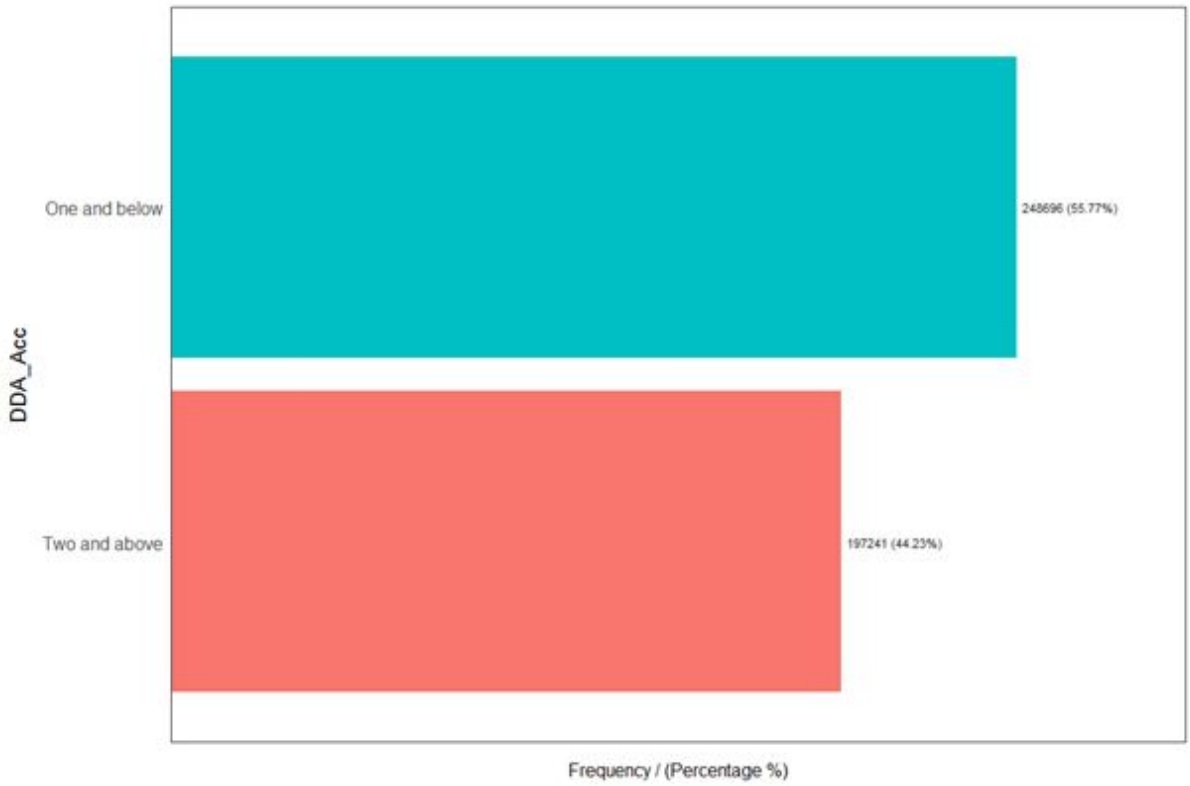

Figure A.19: Distribution of DDA Account

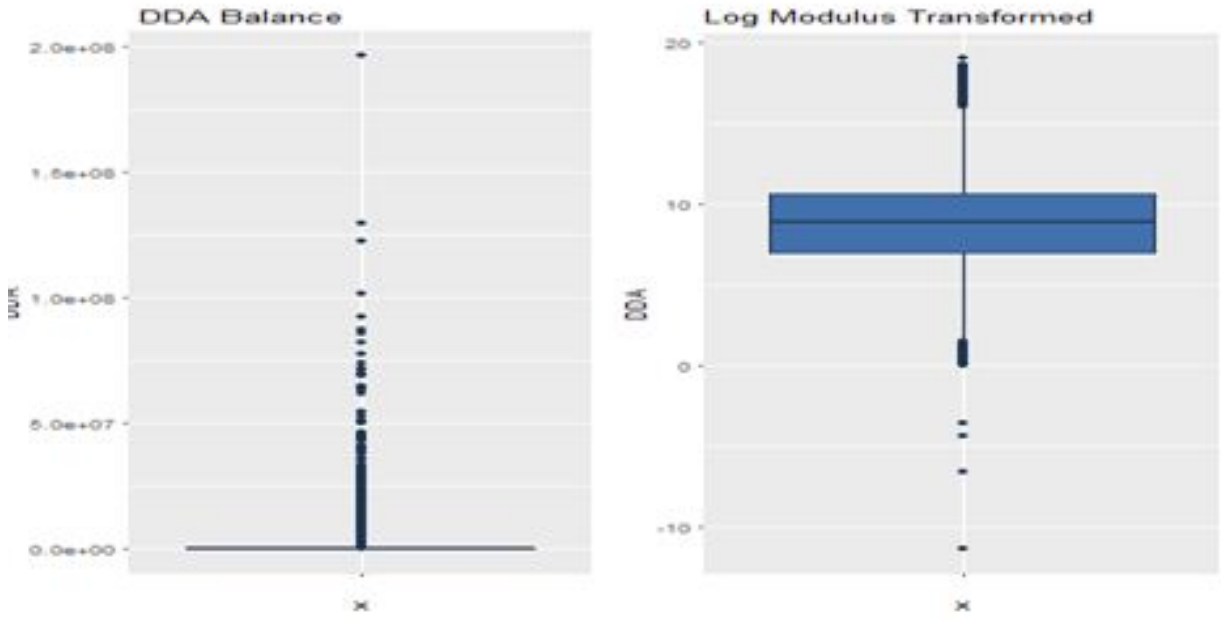

Figure A.20: Distribution of DDA Balance

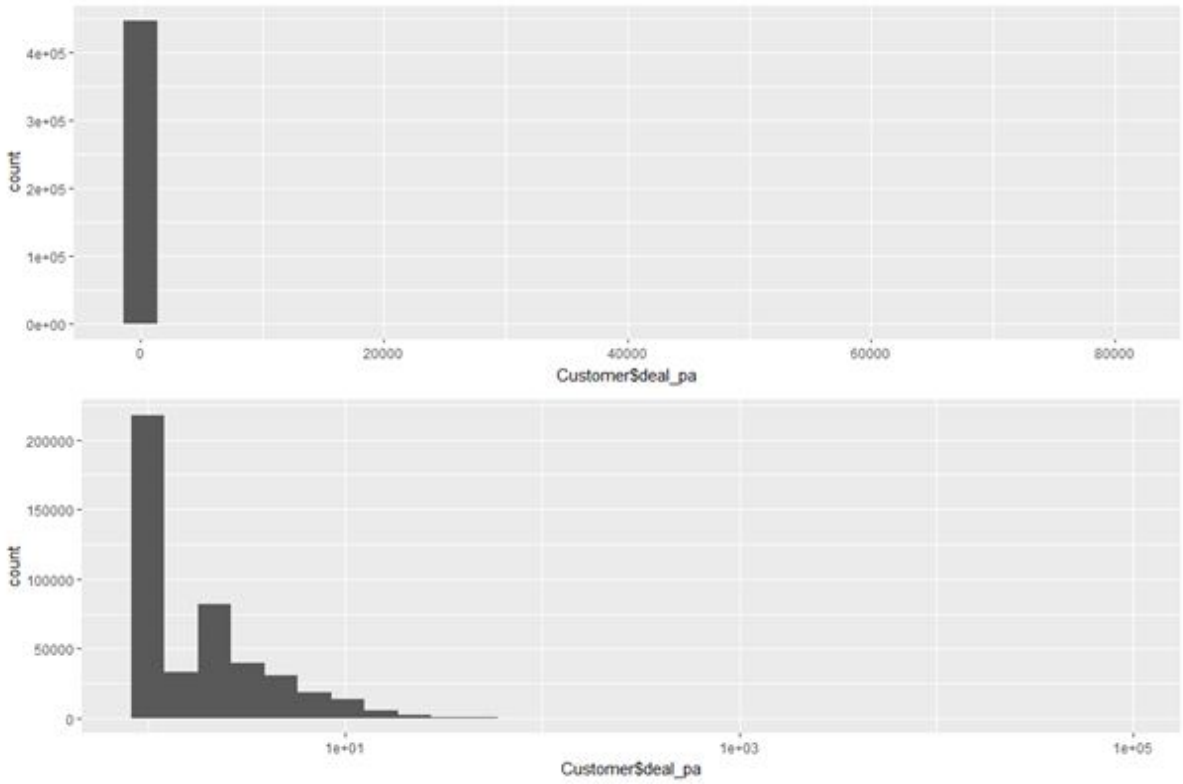

Figure A.21: Distribution of Deals Per Annum

Figure A.22: Distribution of FX VIP

Figure A.23: Distribution of Global Account

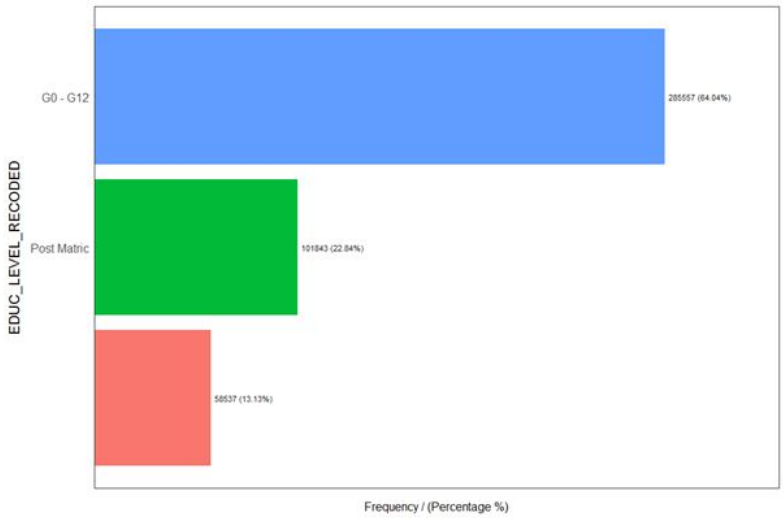

Figure A.24: Distribution of Highest Education Level

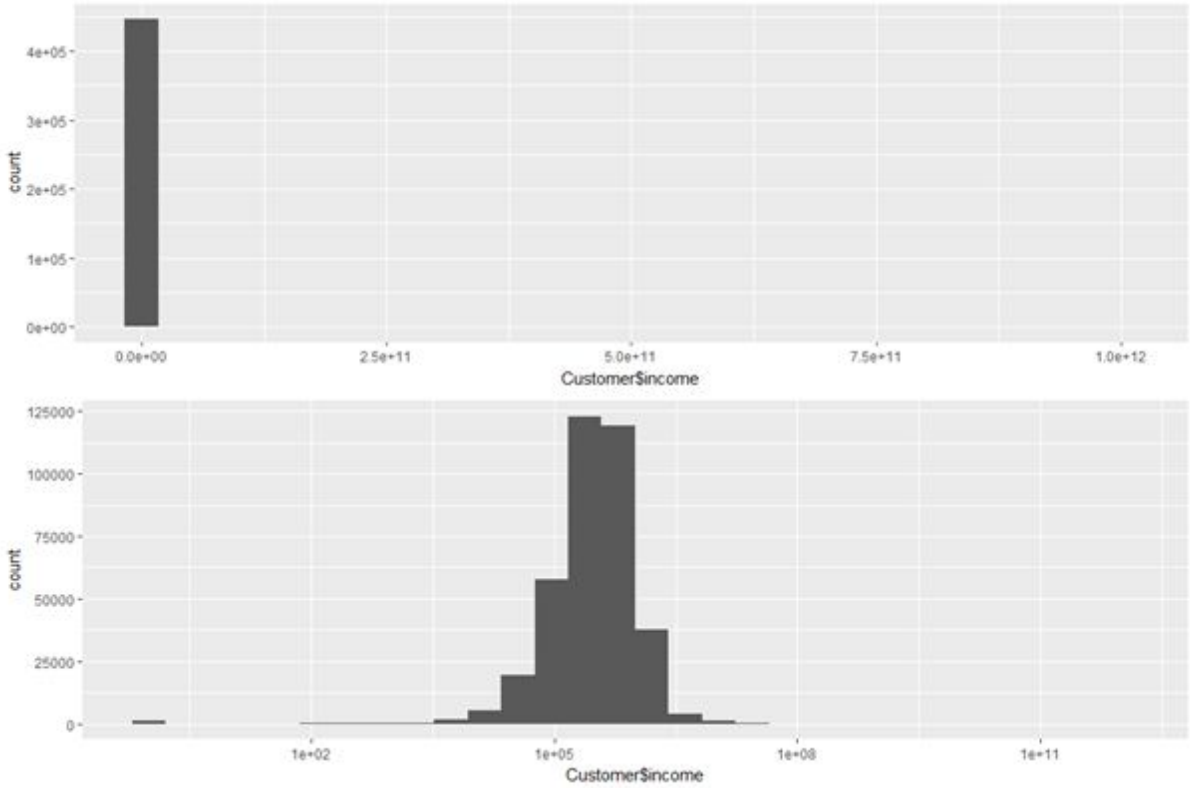

Figure A.25: Distribution of Income

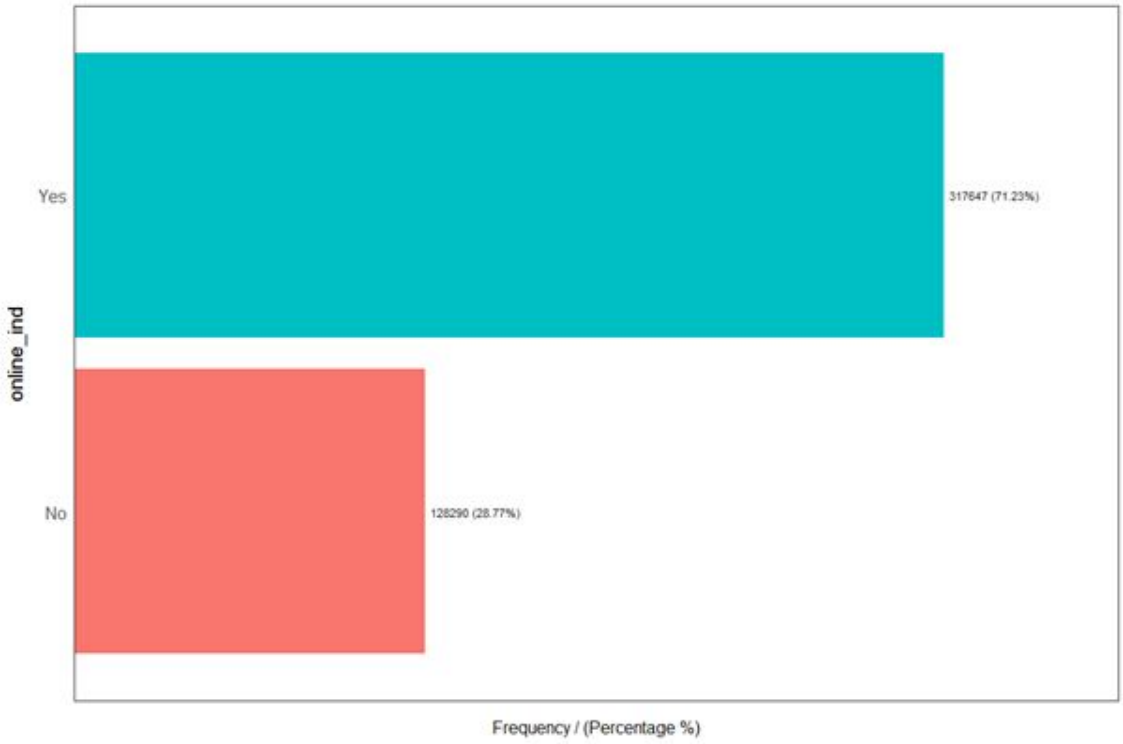

Figure A.26: Distribution of Online Indicator

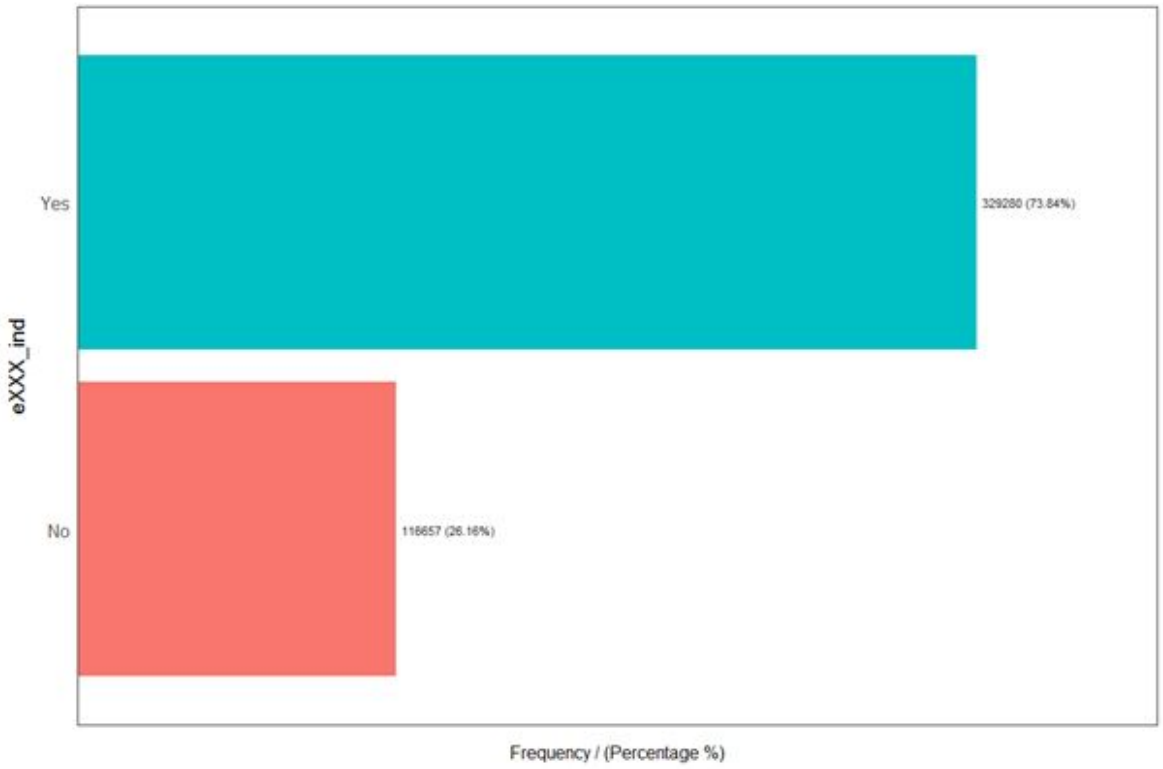

Figure A.27: Distribution of Points Incentive Indicator

Figure A.28: Distribution of Pre Dealing Indicator

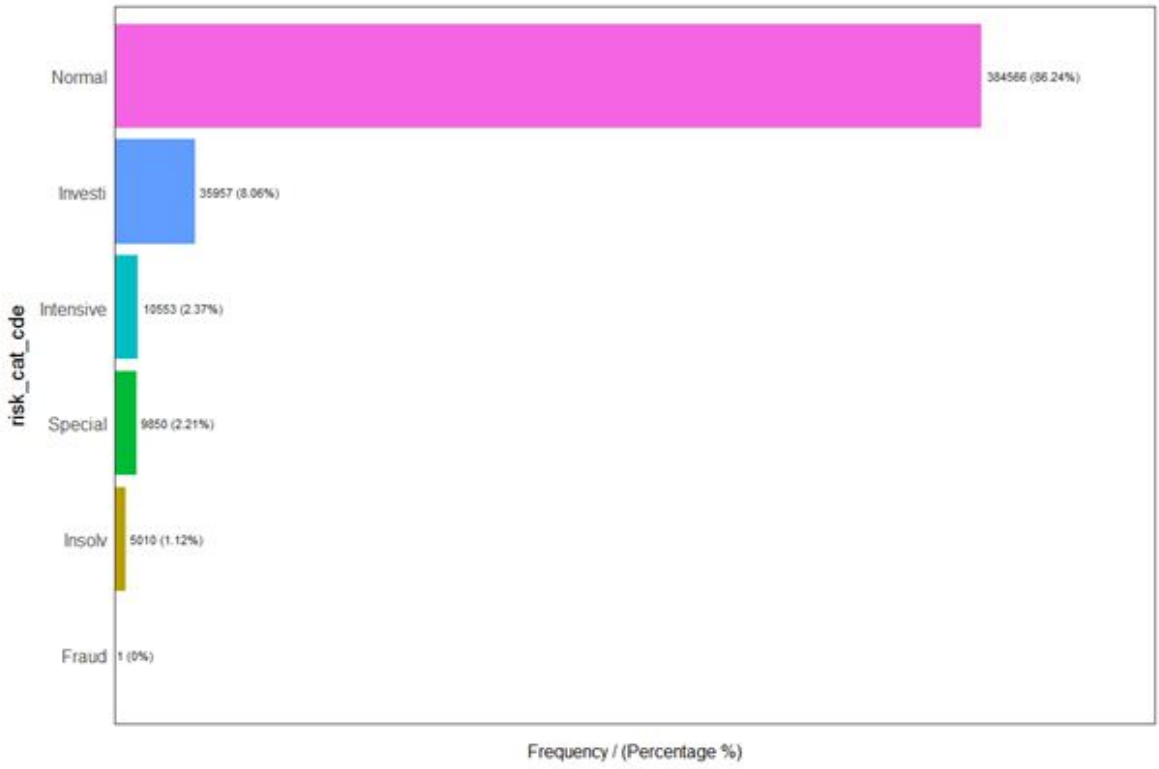

Figure A.29: Distribution of Risk Category Code

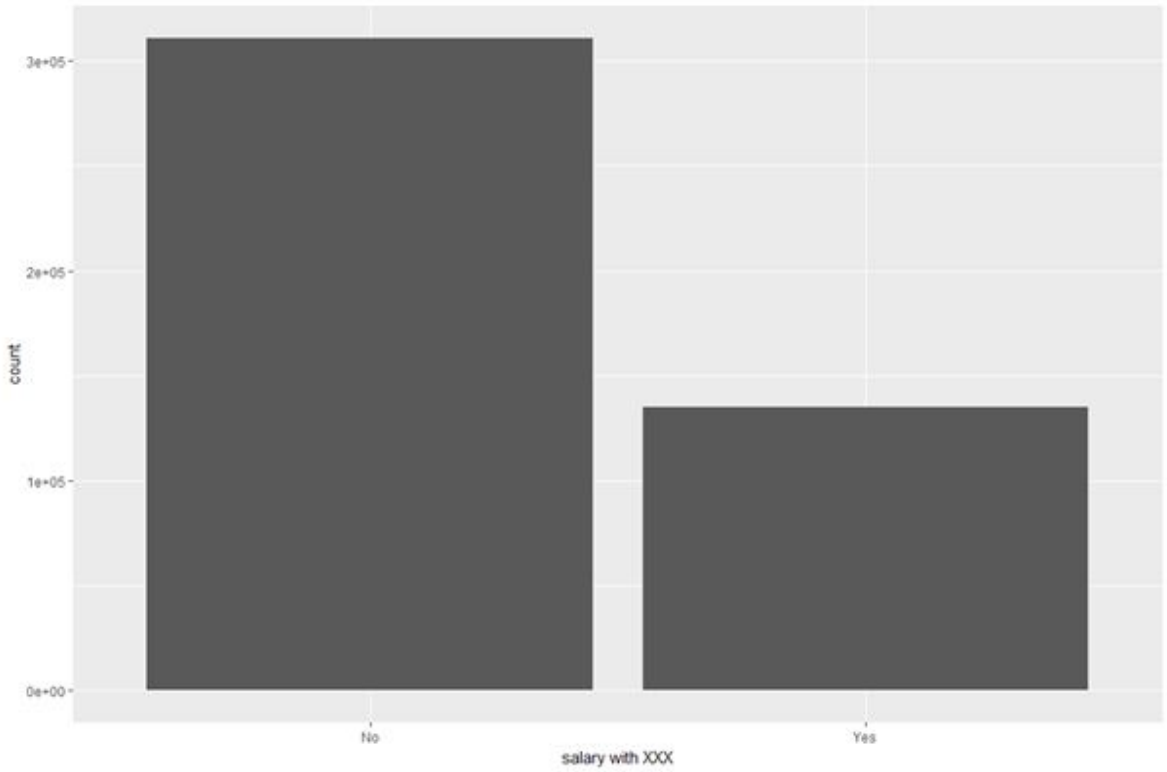

Figure A.30: Distribution of Salary With The Bank

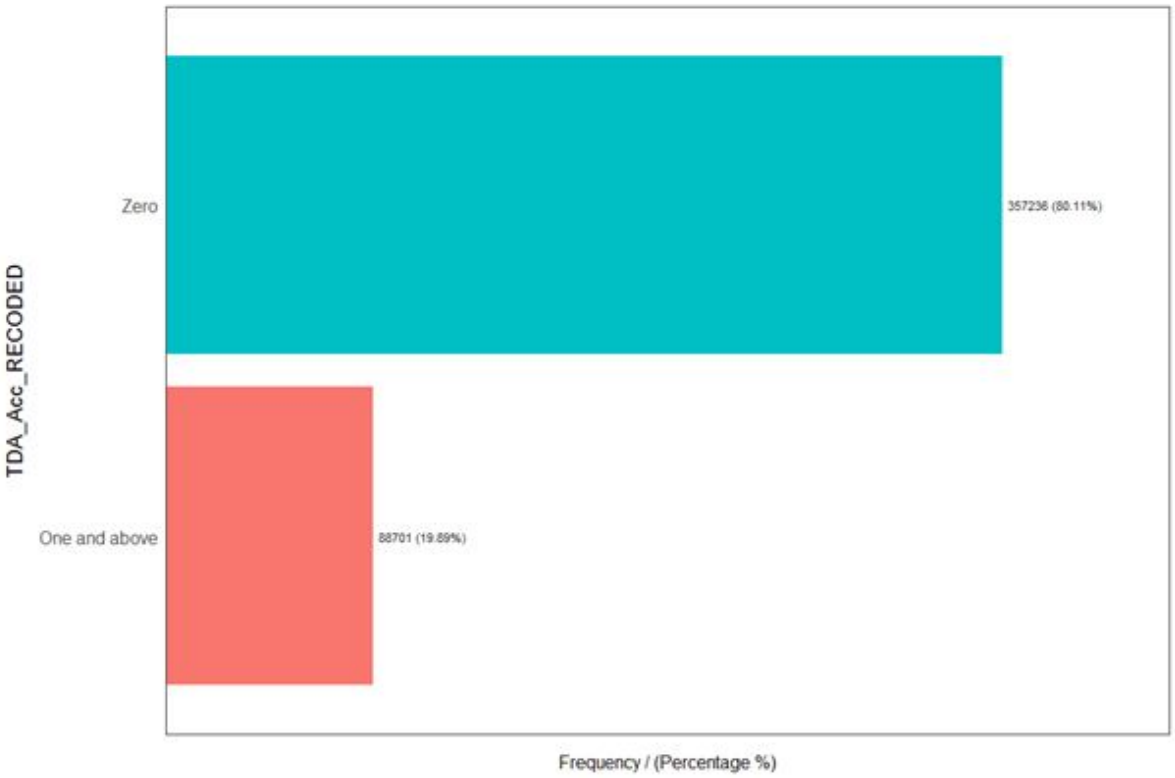

Figure A.31: Distribution of TDA Account

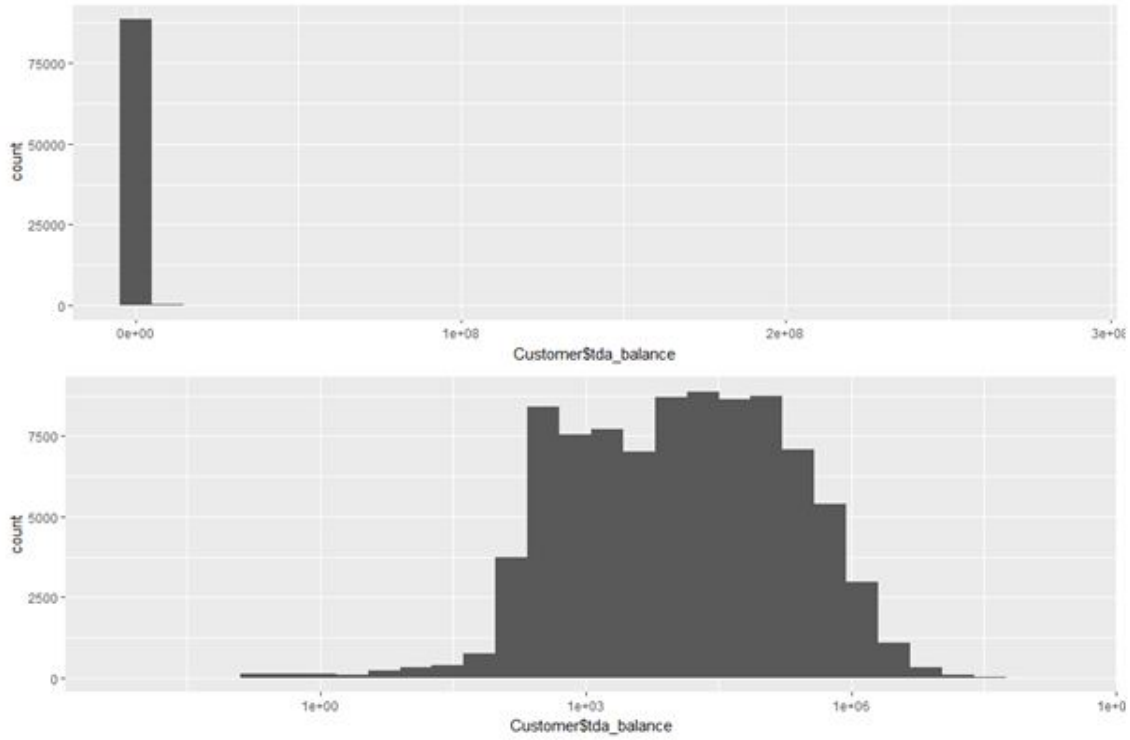

Figure A.32: Distribution of TDA Balance

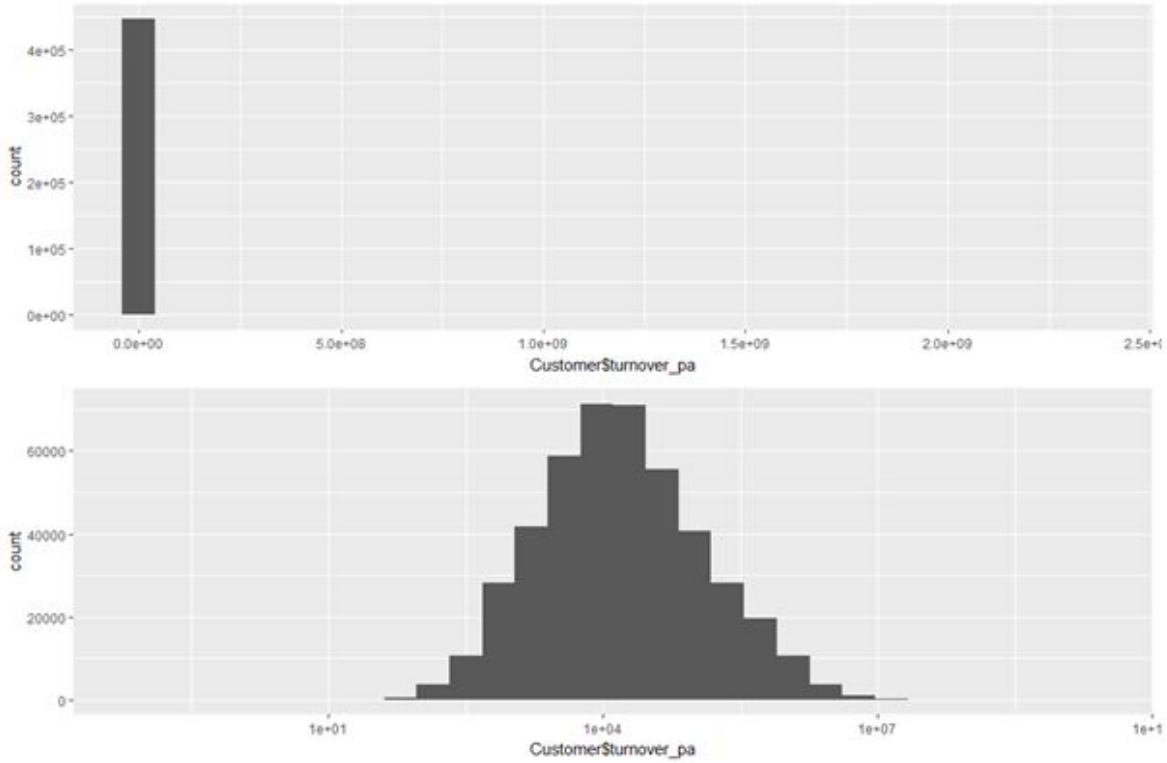

Figure A.33: Distribution of Turnover Per Annum

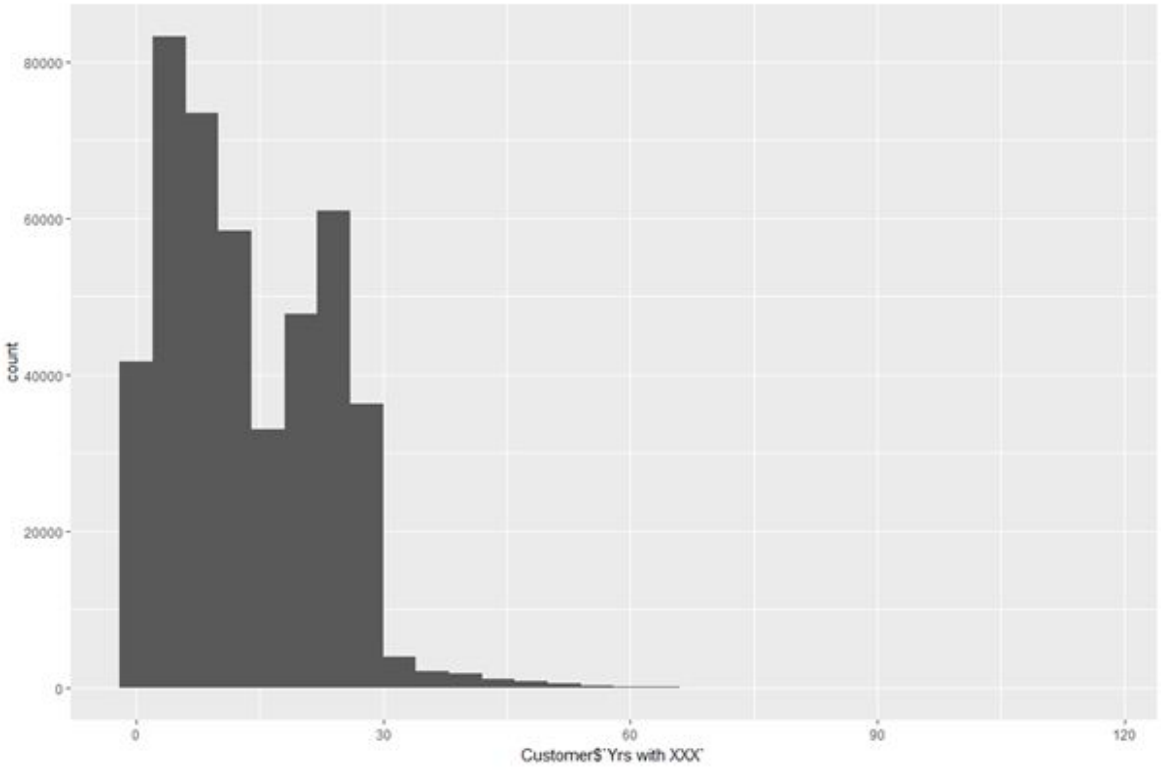

Figure A.34: Distribution of Years With The Bank

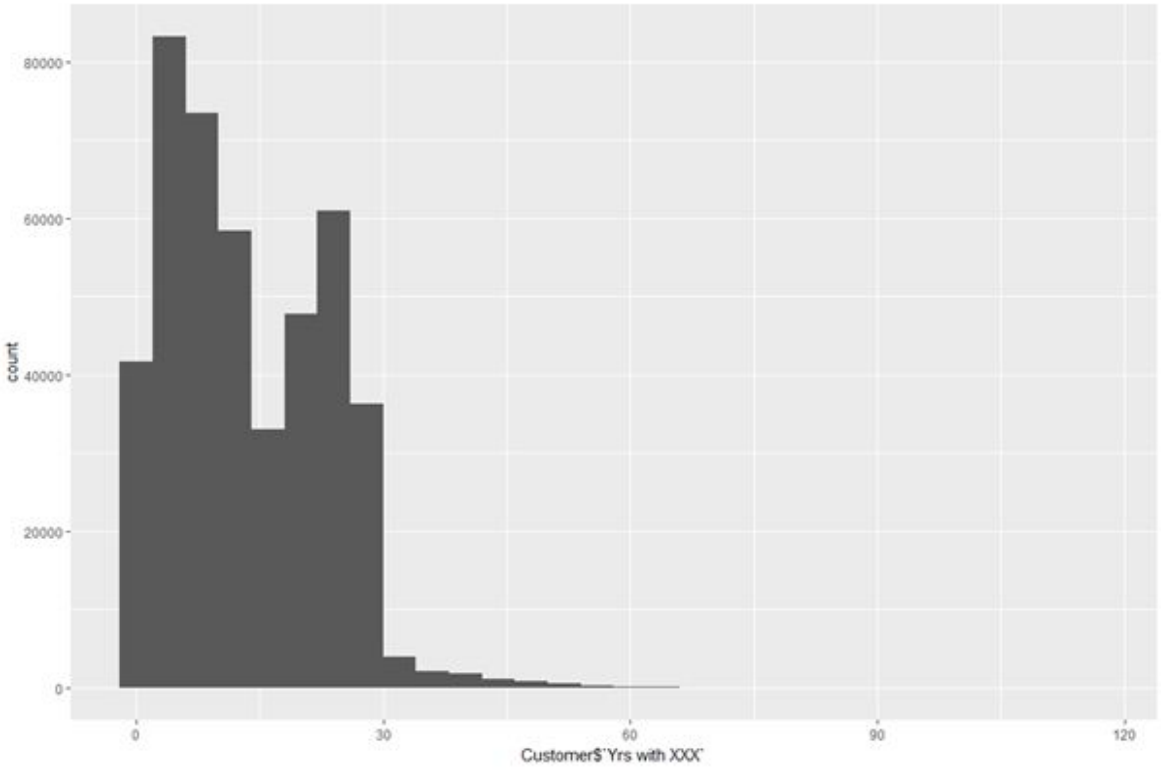

Figure A.35: Distribution of Age

Appendix B

Analysis Results

B.1 Decision Trees

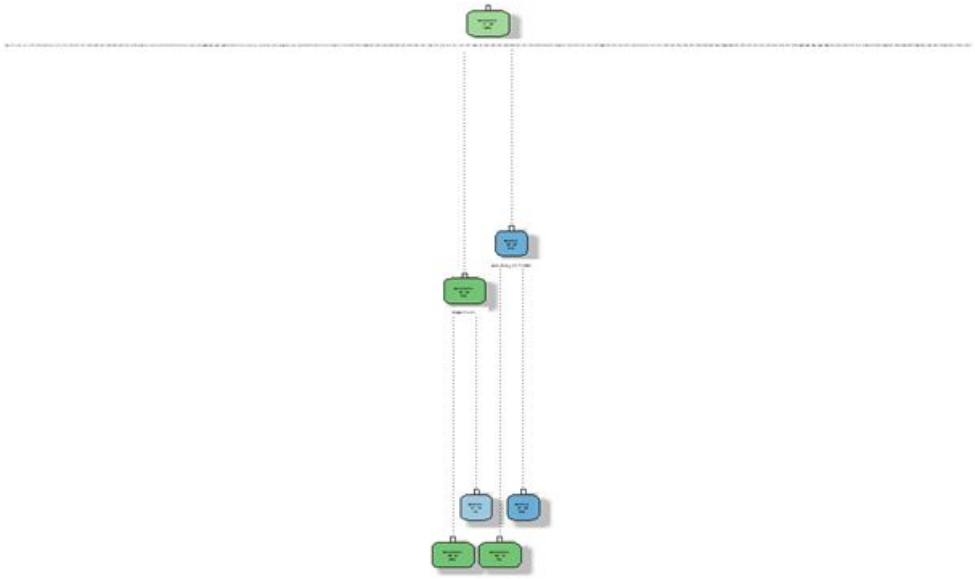

Figure B.1: Decision Tree fitted using all predictor variables

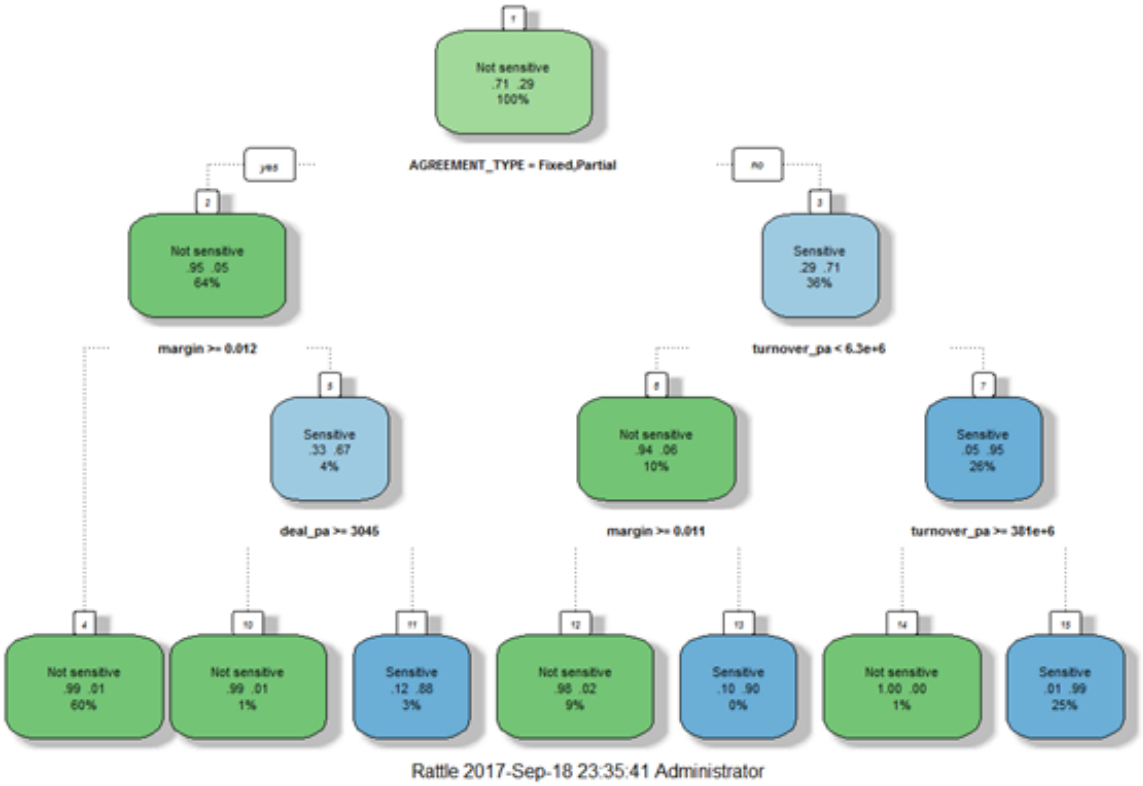

Figure B.2: Decision Tree fitted without branch code variable

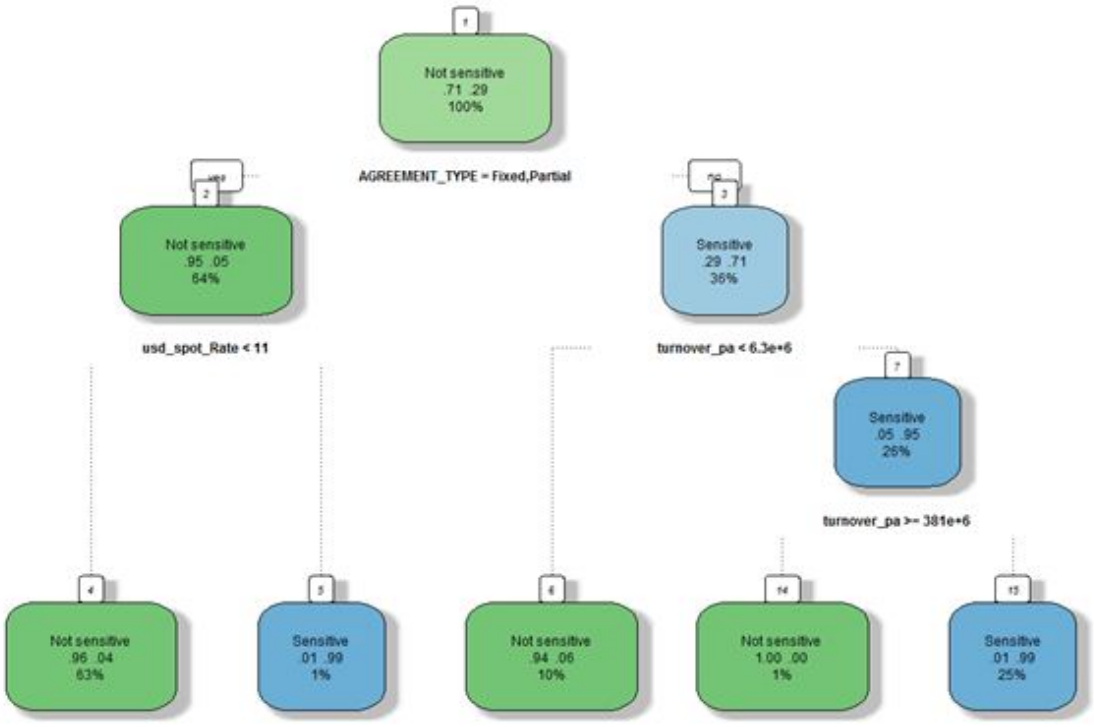

Figure B.3: Decision Tree fitted without branch code and margin

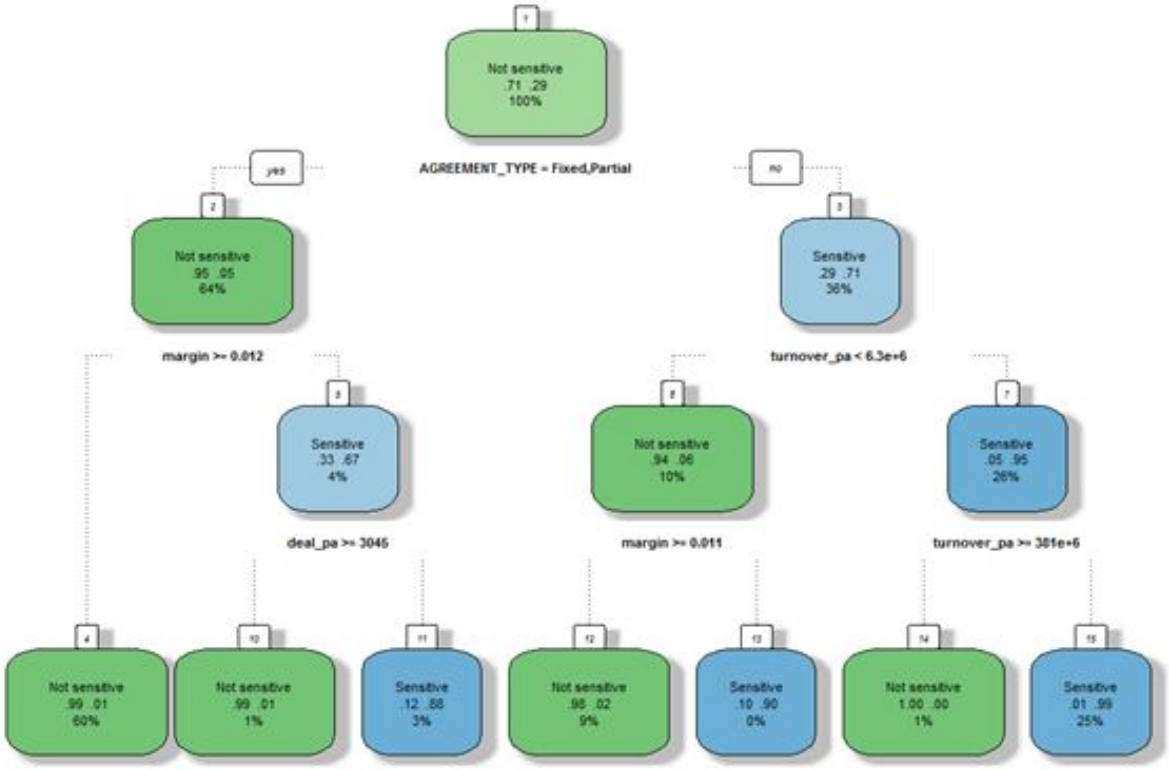

Figure B.4: Decision Tree fitted without branch code and NAs assigned to zero values in the Income

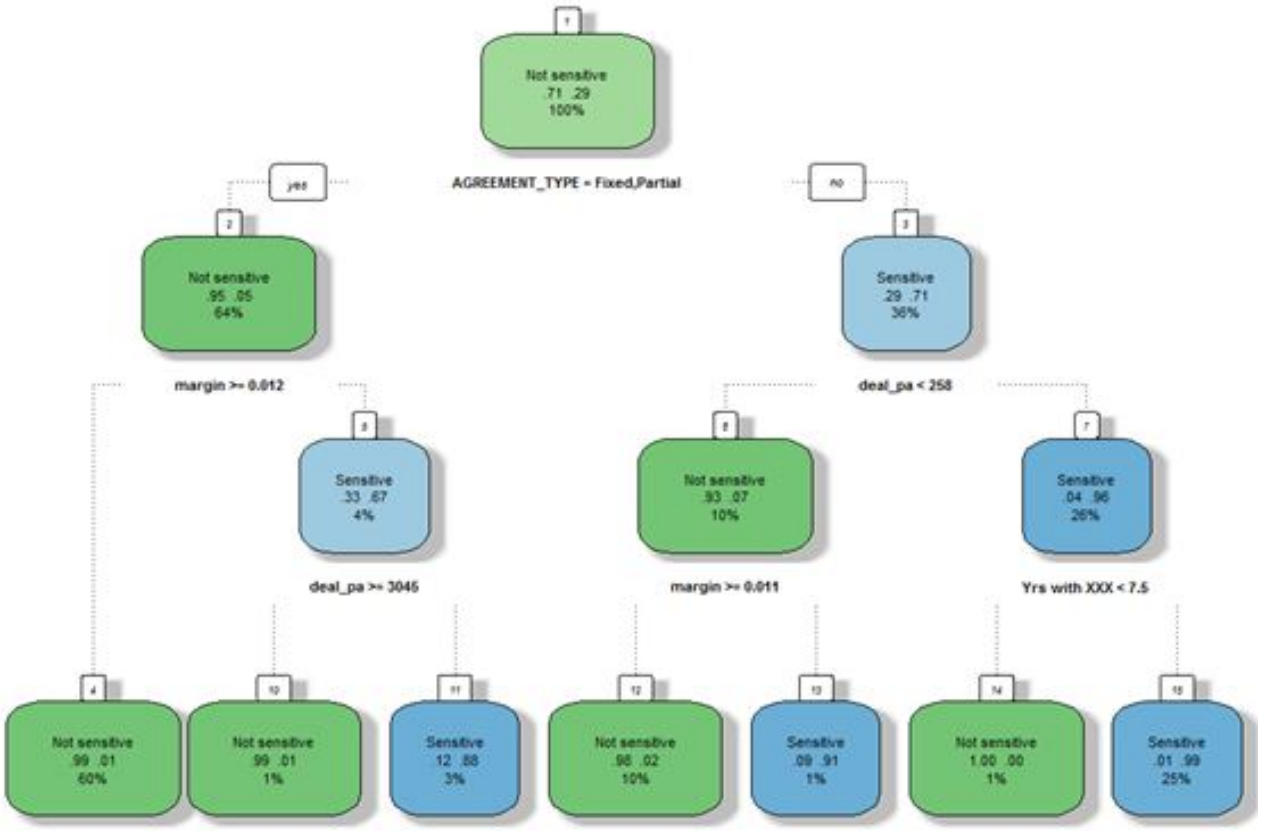

Figure B.5: Decision Tree fitted without branch code and turnover per annum

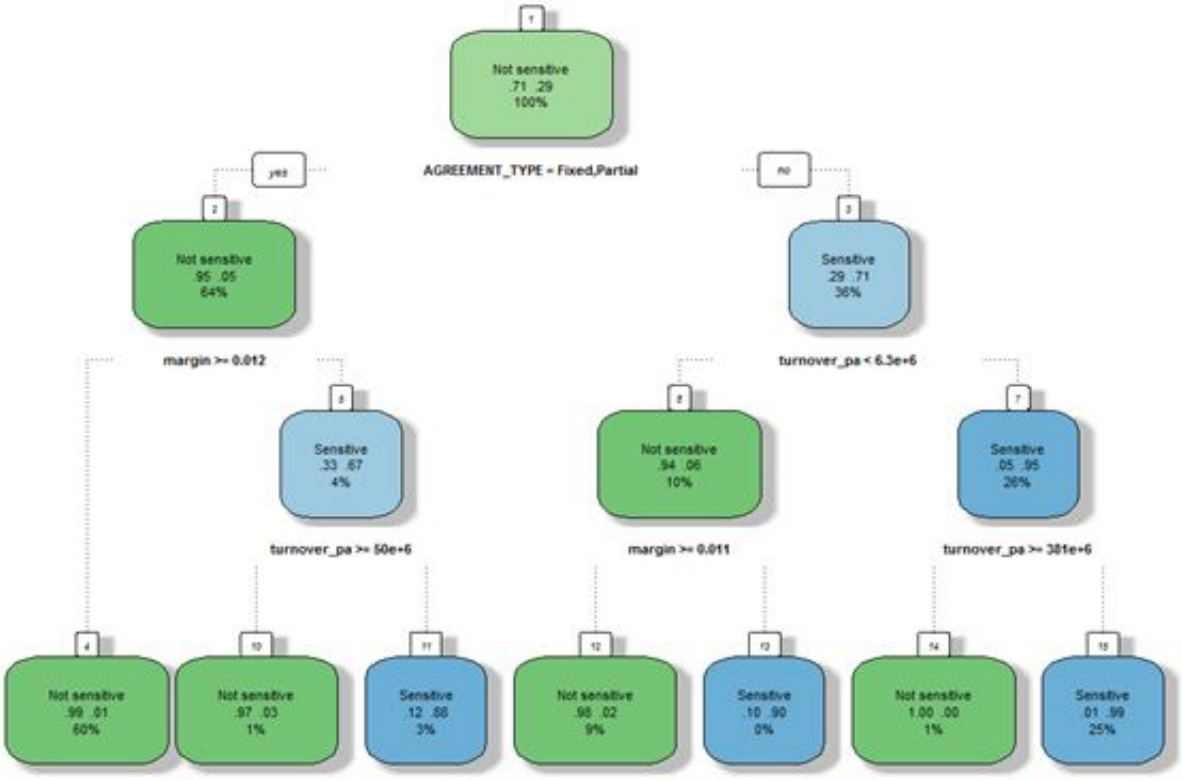

Figure B.6: Decision Tree fitted without branch code and average number of deals

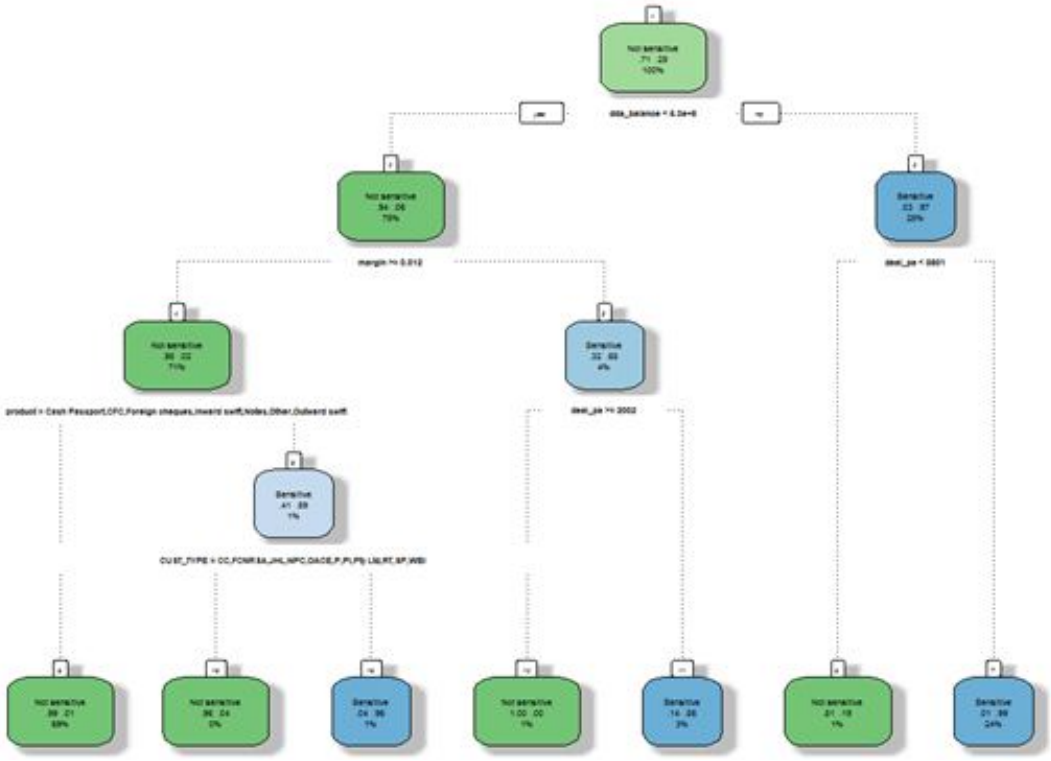

Figure B.7: Tree fitted without Branch Code and with Imputed Product and DDA Balance

B.2 Random Forests

B.2.1 Error Rates across Decision Trees

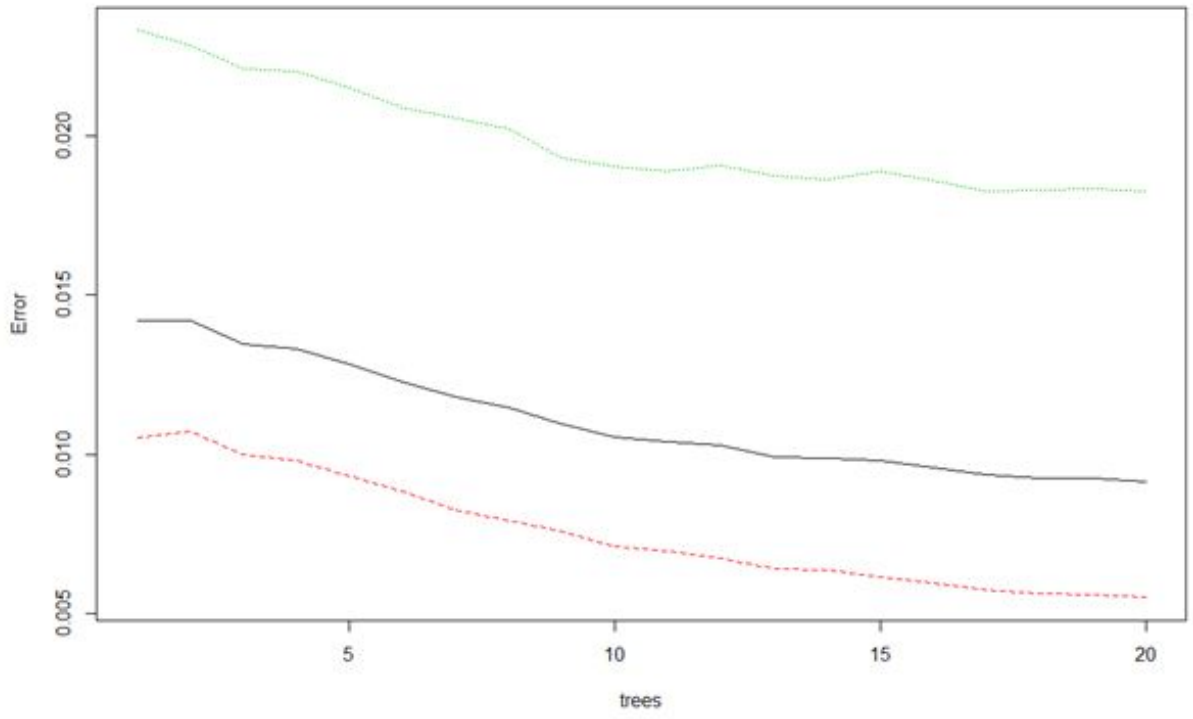

Figure B.8: Error Rate Across Decision Trees - data set 1

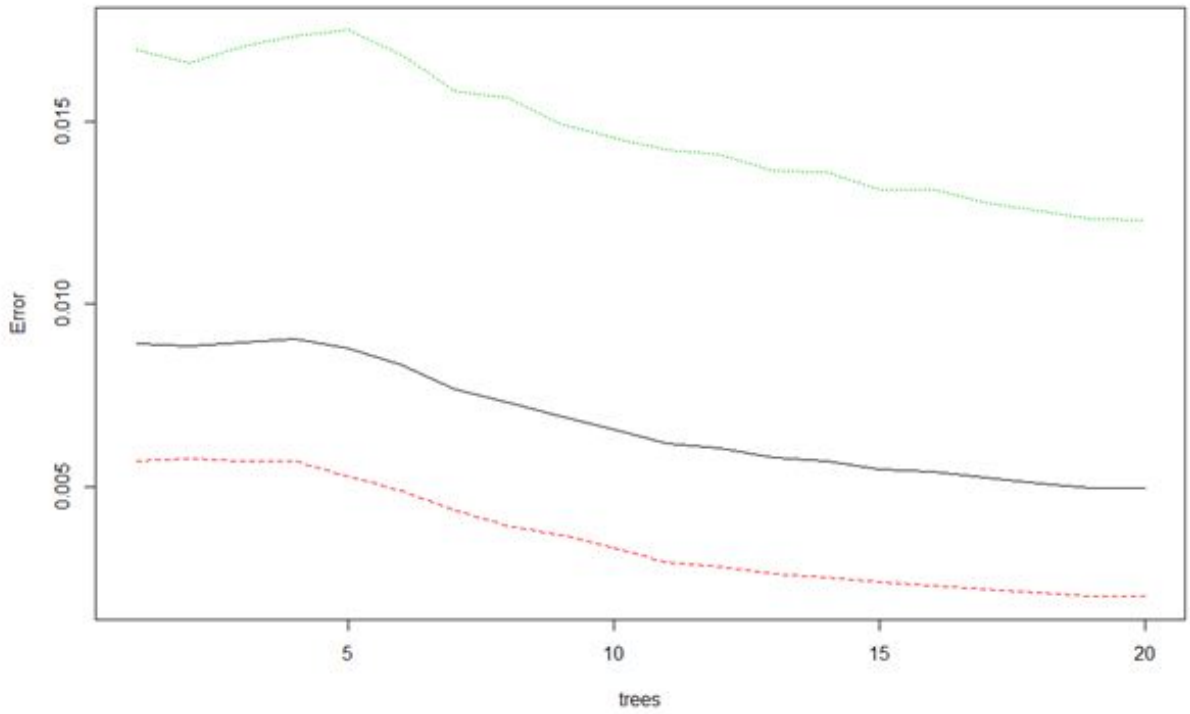

Figure B.9: Error Rate Across Decision Trees for - data set 2

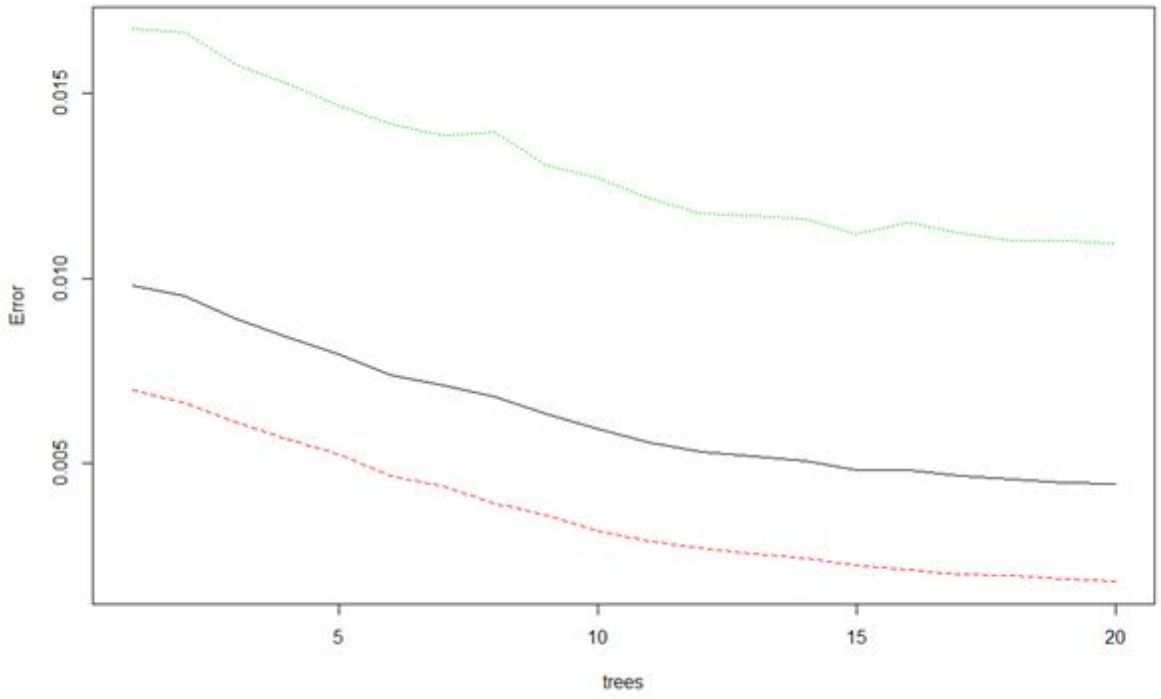

Figure B.10: Error Rate Across Decision Trees for - data set 3

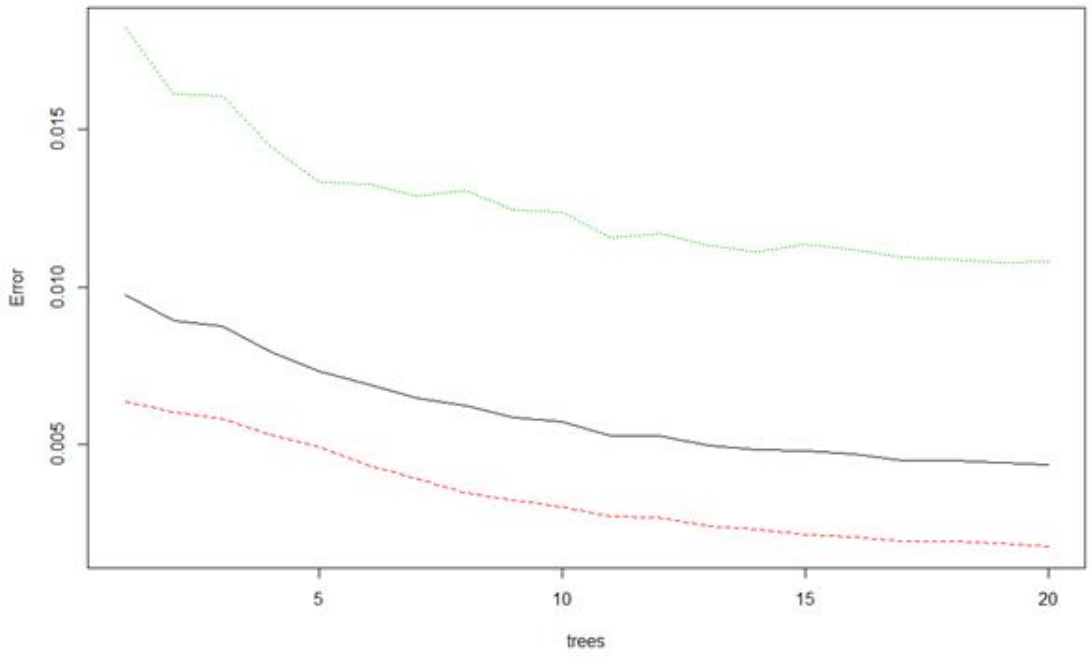

Figure B.11: Error Rate Across Decision Trees for - data set4

B.2.2 Variable importance

Variable importance		
Variables	Mean Decrease Gini	Rank
Turnover per annum	35292.49	1
Agreement Type	32721.23	2
Deals per annum	25684.57	3
Years with the bank	16146.55	4
Margin	12891.51	5
USD Spot Rate	4108.47	6

Table B.1: Variable importance for Dataset 1

Variable importance		
Variables	Mean Decrease Gini	Rank
Points Incentive Indicator	21461.13	1
Banked With The Bank	15688.69	2
Agreement Type	15169.21	3
Customer Type	13708.78	4
Deals per annum	13304.007	5
Margin	10684.015	6
Turnover per annum	7261.51	7
Years with the bank	7101.82	8
Customer Age	5545.48	9
Income	3415.61	10
DDA Account	3238.60	11
USD Spot Rate	2866.30	12
Turnover	2498.39	13
Deal Revenue	2081.99	14
Customer Segment	1946.59	15
Range	1208.60	16
Deal Type	301.27	17
Dealt Currency	189.67	18
Online Indicator	66.48	19
Risk Category Code	58.36	20
Card Account	57.85	21
TDA Account	39.21	22
Salary with the bank	25.69	23

Table B.2: Variable importance for Dataset 2

Variable importance		
Variables	Mean Decrease Gini	Rank
Banked with the Bank	22990.37	1
Agreement Type	20507.11	2
Deals per annum	14767.25	3
Turnover per annum	13914.38	4
Margin	10607.21	5
Customer Type	21461.13	6
Product	9048.66	7
Years with the bank	7787.31	8
Customer Age	2956.39	9
Points Incentive Indicator	2788.80	10
Turnover	2562.73	11
USD Spot Rate	2255.84	12
Branch code	2103.95	13
Deal Revenue	1511.25	14
Range	1402.18	15
Customer Segment	1205.42	16
Online Indicator	852.63	17
Dealt Currency	750.94	18
DDA Account	564.10	19
Income	300.85	20
Deal Type	200.47	21
Salary with the bank	113.46	22
Card Account	63.55	23
Risk Category Code	44.78	24
TDA Account	38.32	25

Table B.3: Variable importance for Dataset 3

Variable importance		
Variables	Mean Decrease Gini	Rank
DDA balance	23111.35	1
Product	16994.94	2
Banked with the Bank	13373.76	3
Margin	12412.07	4
Turnover per annum	12187.95	5
Deals per annum	10554.61	6
Agreement Type	10092.11	7
Years with the bank	7080.28	8
USD Spot Rate	2866.07	9
Customer Age	2787.90	10
Customer Type	2763.96	11
Income	2570.11	12
Range	1726.02	13
Turnover	1698.01	14
Branch code	1652.12	15
Points Incentive Indicator	1629.85	16
Deal Revenue	1366.59	17
DDA Account	1334.47	18
Customer Segment	806.31	19
Card Account	423.41	20
Dealt Currency	196.50	21
Deal Type	161.20	22
Online Indicator	106.65	23
Risk Category Code	39.49	24
TDA Account	36.66	25
Salary with the bank	22.81	26

B.2.3 Area Under the Area

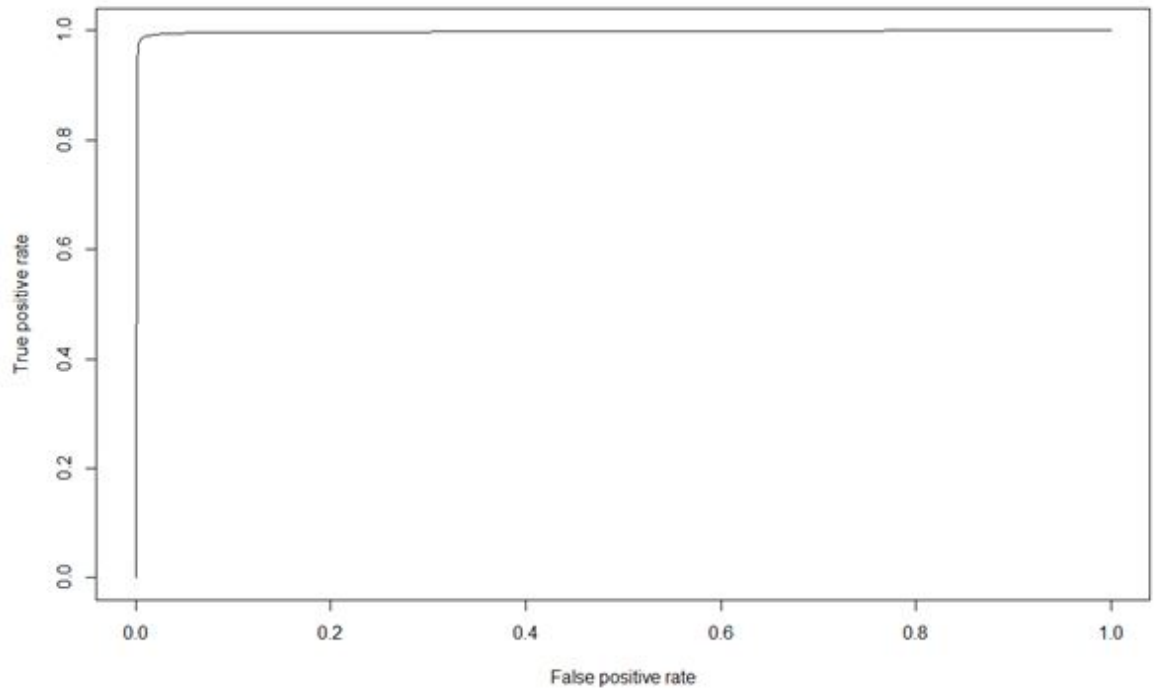

Figure B.12: Area Under the Curve - data set 1

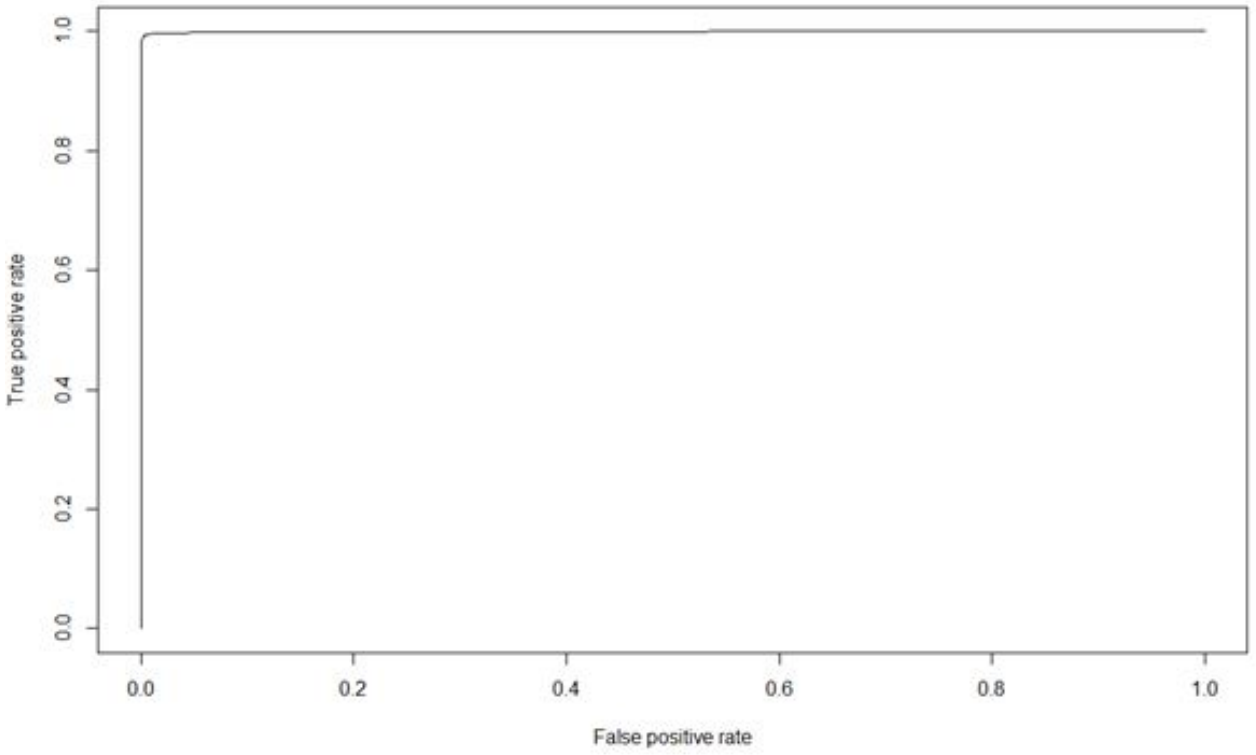

Figure B.13: Area Under the Curve - data set 2

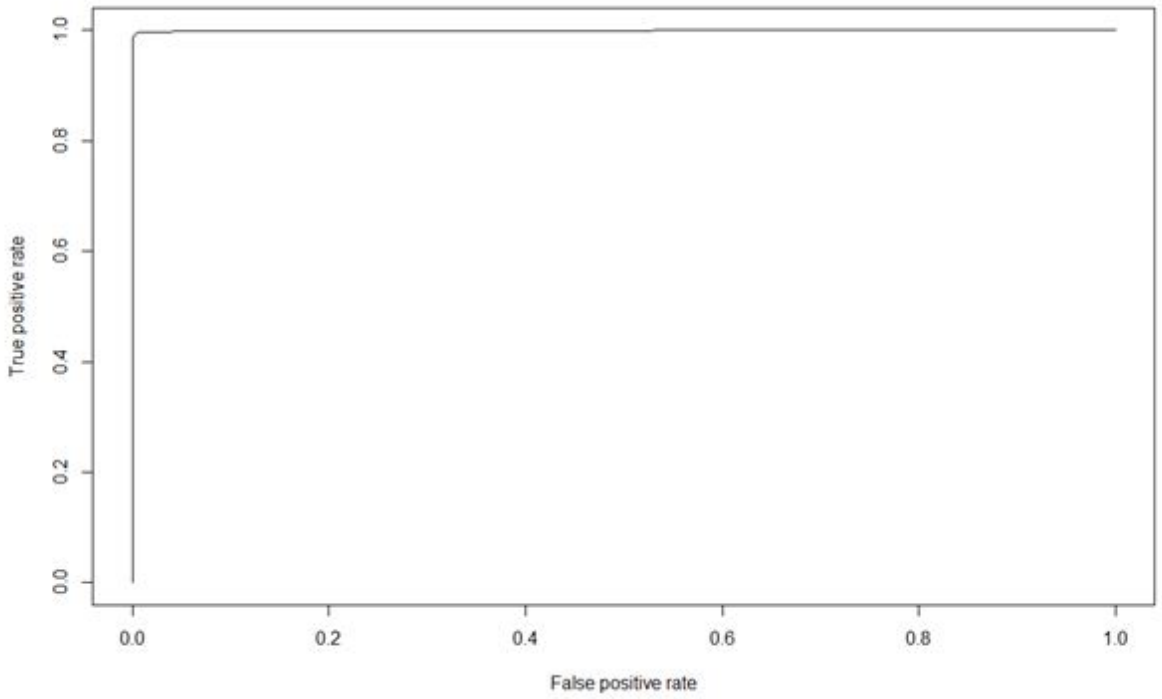

Figure B.14: Area Under the Curve - data set 3

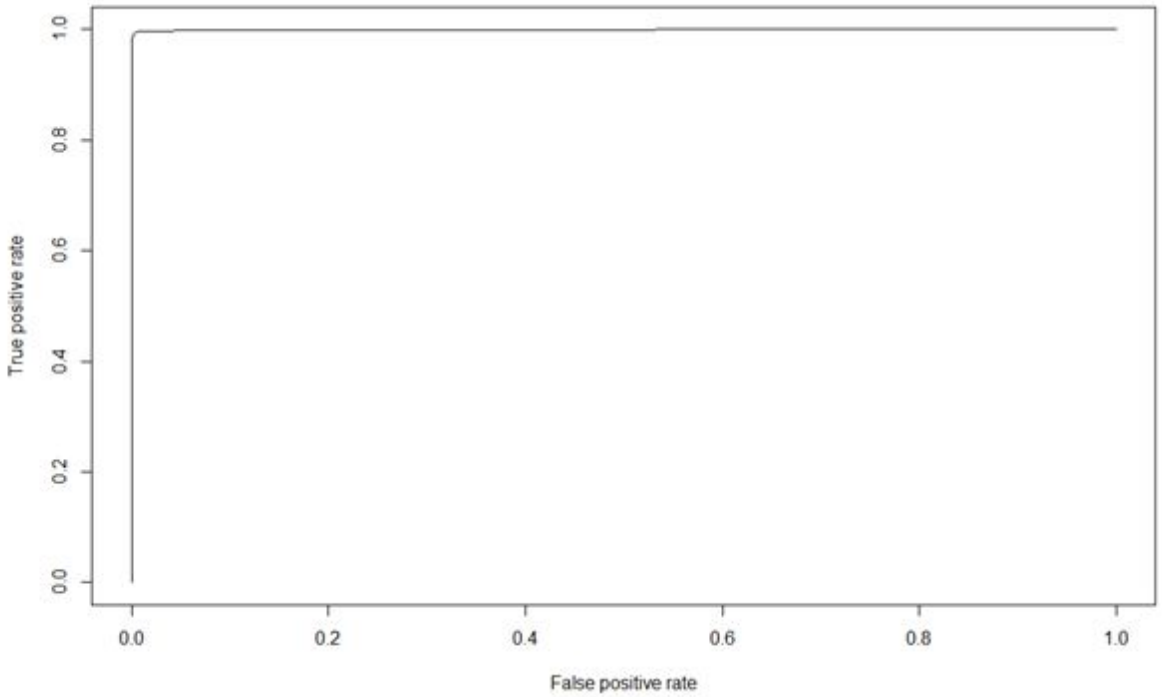

Figure B.15: Area Under the Curve - data set 4

B.3 Neural Networks

B.3.1 Summary of the NN results

Model results	
error	53607.56
Reached threshold	0.01
Steps	21299.00
Intercept	5.88
Agreement TypeOptional	161.47
Agreement TypePartial	-45.11
Agreement TypeSpot	-0.35
Years with the bank	3.29
Deals per annum	-111.23
Turnover per annum	127.54
USD Spot Rate	5.92
Margin	-146.51
Intercept.to.SensitivitySensitive	-2.88
1layhid.1.to.SensitivitySensitive	7.64

Table B.5: Summary of the NN results based on Dataset 1. The table shows the number of steps needed in the training process, the threshold for the partial derivatives of the error function and the estimated weights

Model results	
error	26178.70
Reached threshold	0.01
Steps	57060.00
Intercept	-2.49
Customer Type Cooperative	-708.33
Customer Type Foundation	-0.08
Customer Type Bank Associated	2.52
Customer Type Foreign co. Not reg. in SA	-0.09
Customer Type Foreign co. reg. in SA	0.12
Customer Type government	-710.44
Customer Type joint home loans	-1.05
Customer Type public company	0.62
Customer Type non profit company	-0.19
Customer Type other associations	-0.37
Customer Type partnership	-0.17
Customer Type private individual	-1.46
Customer Type personal liability company	-0.06
Customer Type private company	-0.32
Customer Type reserved accounts	-707.51
Customer Type reg. friendly society	-708.80
Customer Type reg. trust	-0.68

Customer Type sole proprietorship	-0.49
Customer Type stokvel	-708.37
Customer Type asset bank individual	-0.87
online indicator Yes	0.15
Salary with bank Yes	-0.14
Incentive Points Indicator Yes	0.13
DDA Account Two.and.above	0.06
TDA Account Zero	0.04
Card Account One.or.more	0.17
Banked with the bank Yes	-0.10
Risk Category Insolvency	-0.79
Risk Category Intensive	-0.66
Risk Category Investigation	-0.79
Risk Category Normal	-0.84
Risk Category Special	-1.07
Deal Type Sell	-2.13
Dealt Currency GBP	-0.03
Dealt Currency USD	-0.40
Agreement TypeOptional	587.95
Agreement TypePartial	-0.48
Agreement TypeSpot	1.40
Customer Segment Commercial	-0.48

Customer Segment Customer	2.73
Customer Segment Wealth	3.64
Customer Age	-0.26
Income	-3.25
Years with the bank	-1.73
Deals per annum	-25.13
Turnover per annum	25.73
Turnover	458.00
Deal revenue	-4718.97
USD Spot Rate	4.42
Range	20.70
Margin	-3.97
Intercept.to.SensitivitySensitive	-4.97
1layhid.1.to.SensitivitySensitive	9.08

Table B.6: Summary of the NN results based on Dataset 2. The table shows the number of steps needed in the training process, the threshold for the partial derivatives of the error function and the estimated weights.

Model results	
error	22926.90
Reached threshold	0.01
Steps	31127.00

Intercept	-0.54
Customer Type Cooperative	34.97
Customer Type Foundation	0.49
Customer Type Bank Associated	-1.76
Customer Type Foreign co. Not reg. in SA	-0.09
Customer Type Foreign co. reg. in SA	0.41
Customer Type government	36.76
Customer Type joint home loans	0.65
Customer Type public company	1.16
Customer Type non profit company	0.06
Customer Type other associations	0.22
Customer Type partnership	0.10
Customer Type private individual	1.03
Customer Type personal liability company	-0.03
Customer Type private company	0.17
Customer Type reserved accounts	34.29
Customer Type reg. friendly society	36.30
Customer Type reg. trust	0.54
Customer Type sole proprietorship	0.45
Customer Type stokvel	34.93
Customer Type asset bank individual	0.54
online indicator Yes.to1layhid1	-0.19

Salary with bank Yes	0.24
Incentive Points Indicator Yes	0.10
DDA Account Two.and.above	0.03
TDA Account Zero	0.10
Card Account One.or.more	-0.14
Banked with the bank Yes	0.28
Risk Category Insolvency	0.70
Risk Category Intensive	0.86
Risk Category Investigation	0.74
Risk Category Normal	0.81
Risk Category Special	1.08
Deal Type Sell	2.15
Dealt Currency GBP	-0.18
Dealt Currency USD	0.39
Agreement TypeOptional	-1013.75
Agreement TypePartial	-0.19
Agreement TypeSpot	-2.13
Customer Segment Commercial	0.46
Customer Segment Customer	-1.69
Customer Segment Wealth	-2.34
Product Foreign Currency Account	0.70
Product Foreign.cheques	2.94

Product Inward Swift	0.49
Product Multi Currency Account	-0.81
Product Notes	-1.38
product Other	32.53
Product Outward Swift	0.51
Product Travellers Cheques	-1.95
Customer Age	-0.88
Income	4.01
Years with the bank	2.77
Deals per annum	12.42
Turnover per annum	-15.09
Turnover	-230.47
Deal revenue	2380.27
USD Spot Rate	-4.41
Range	-11.99
Margin	30.36
Intercept.to.SensitivitySensitive	5.01
1layhid.1.to.SensitivitySensitive	-10.31

Table B.7: Summary of the NN results based on Dataset 3. The table shows the number of steps needed in the training process, the threshold for the partial derivatives of the error function and the estimated weights.

Model results	
error	22909.43
Reached threshold	0.01
Steps	30719.00
Intercept	-0.77
Customer Type Cooperative	34.97
Customer Type Foundation	0.48
Customer Type Bank Associated	-1.76
Customer Type Foreign co. Not reg. in SA	-0.08
Customer Type Foreign co. reg. in SA	0.42
Customer Type government	36.66
Customer Type joint home loans	0.62
Customer Type public company	1.15
Customer Type non profit company	0.09
Customer Type other associations	0.23
Customer Type partnership	0.13
Customer Type private individual	0.99
Customer Type personal liability company	0.08
Customer Type private company	0.18
Customer Type reserved accounts	34.29
Customer Type reg. friendly society	36.30
Customer Type reg. trust	0.66

Customer Type sole proprietorship	0.44
Customer Type stokvel	35.03
Customer Type asset bank individual	0.50
online indicator Yes	-0.19
Salary with bank Yes	0.24
Incentive Points Indicator Yes	0.10
DDA Account Two.and.above	0.03
DDA balance	-2.64
TDA Account Zero	0.10
Card Account One.or.more	-0.14
Banked with the bank Yes	0.28
Risk Category Insolvency	0.70
Risk Category Intensive	0.86
Risk Category Investigation	0.74
Risk Category Normal	0.81
Risk Category Special	1.08
Deal Type Sell	2.15
Dealt Currency GBP	-0.18
Dealt Currency USD	0.39
Agreement Type Optional	-1013.75
Agreement Type Partial	-0.19
Agreement Type Spot	-2.13

Customer Segment Commercial	0.46
Customer Segment Consumer	-1.69
Customer Segment Wealth	-2.34
Product Foreign Currency Account	0.70
Product Foreign.cheques	2.94
Product Inward Swift	0.49
Product Multi Currency Account	-0.81
Product Notes	-1.38
product Other	32.53
Product Outward Swift	0.51
Product Travellers Cheques	-1.95
Customer Age	-0.88
Income	4.01
Years with the bank	2.77
Deals per annum	12.42
Turnover per annum	-15.09
Turnover	-230.47
Deal revenue	2380.27
USD Spot Rate	-4.41
Range	-11.99
Margin	30.36
Intercept.to.SensitivitySensitive	5.01

1layhid.1.to.SensitivitySensitive	-10.31

Table B.8: Summary of the NN results based on Dataset 4. The table shows the number of steps needed in the training process, the threshold for the partial derivatives of the error function and the estimated weights.

B.3.2 Topology of the Neural Networks

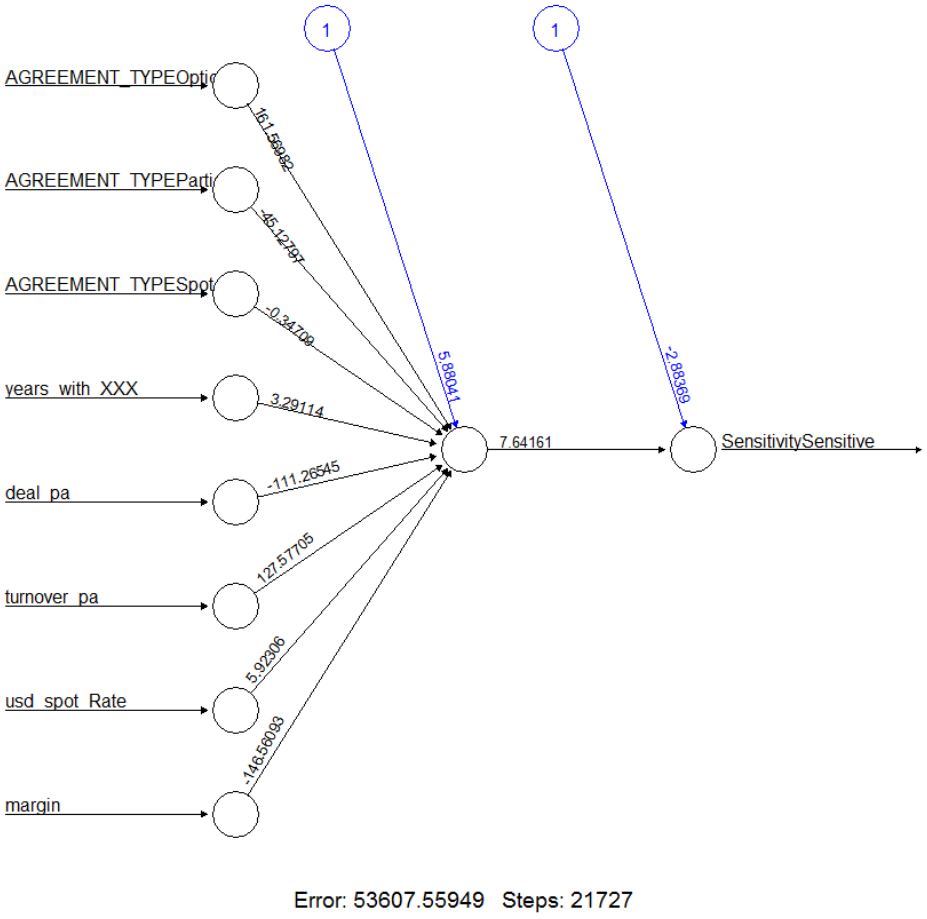

Figure B.16: Topology of the NN - data set 1

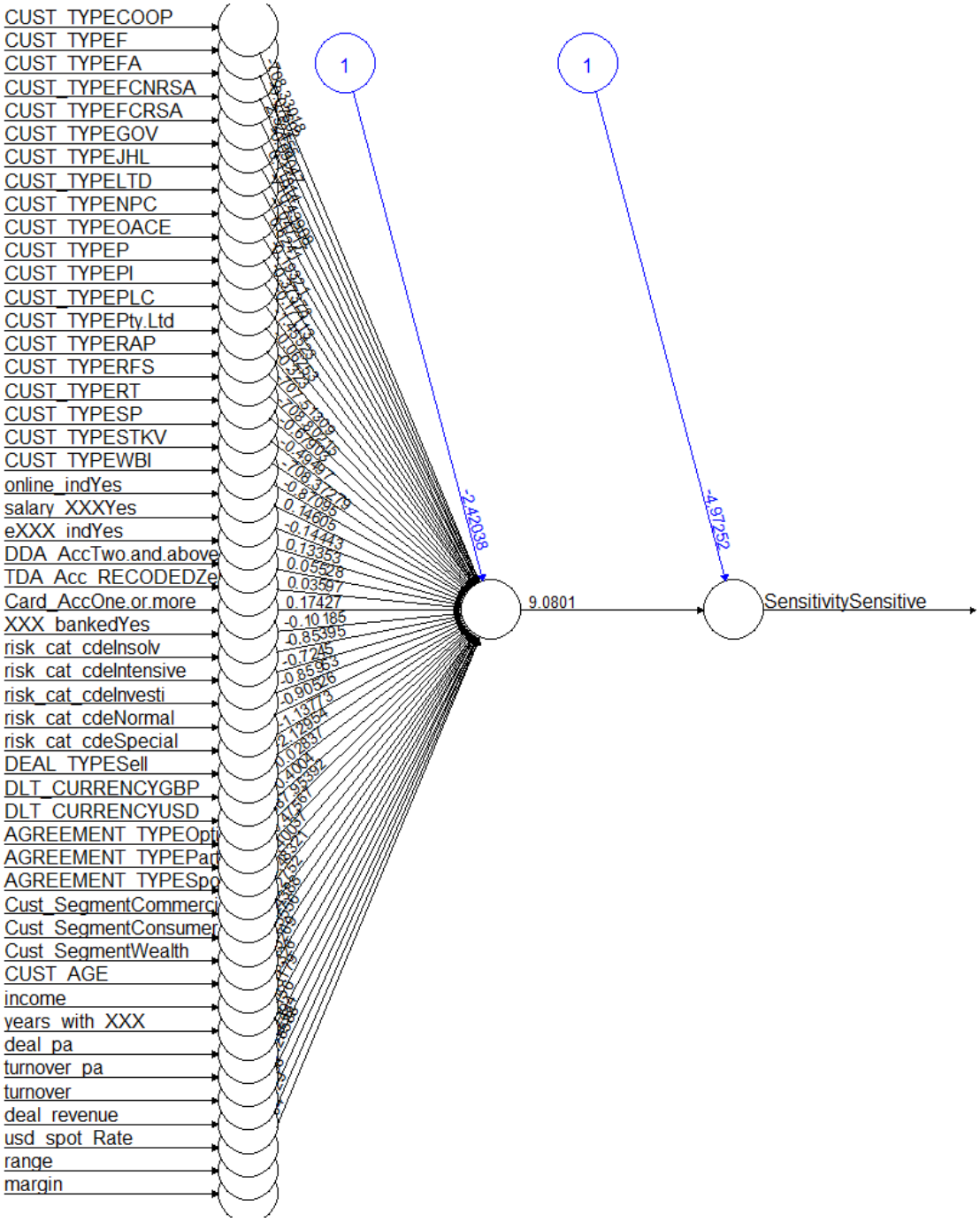

Figure B.17: Topology of the NN - data set 2

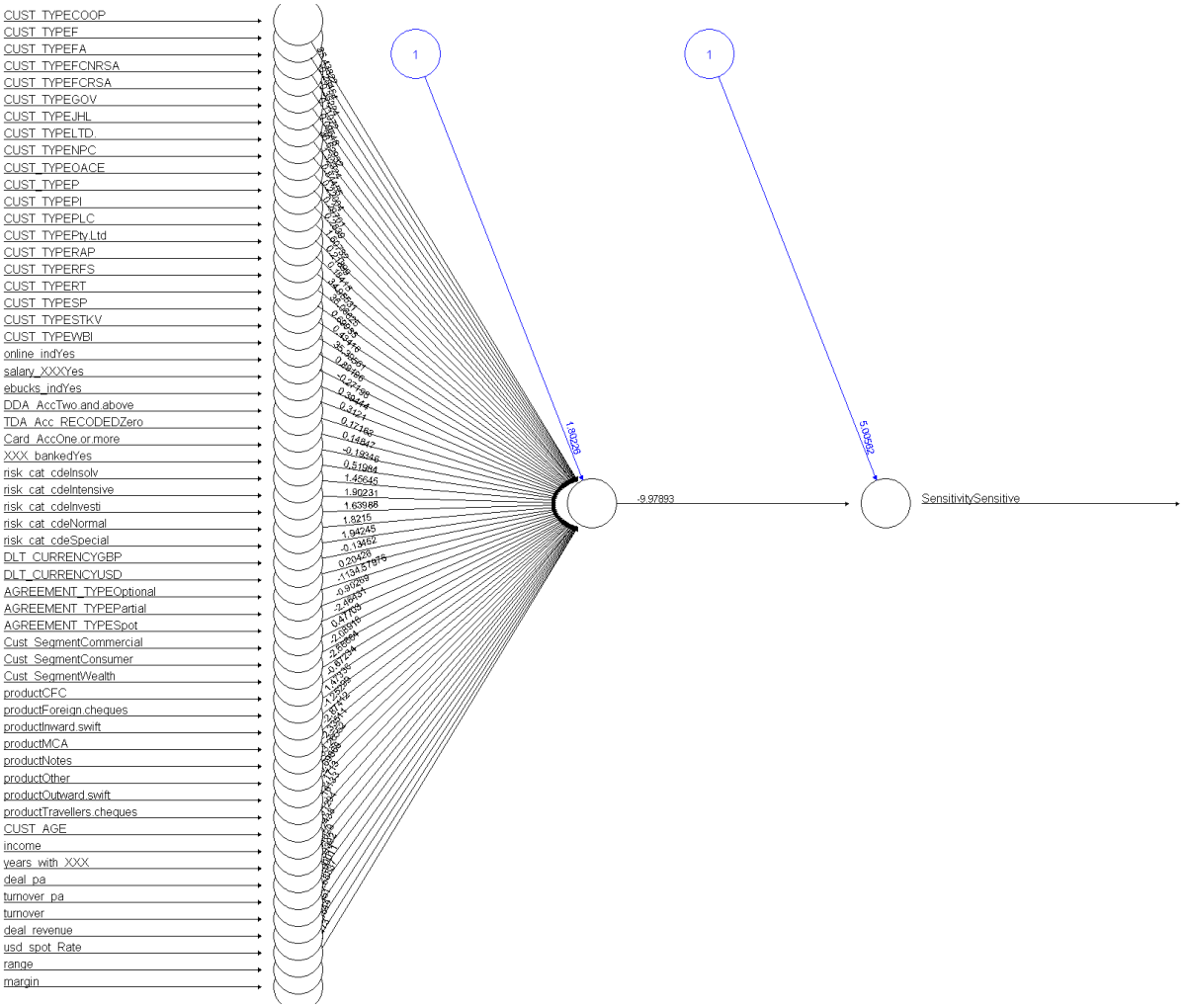

Figure B.18: Topology of the NN - data set 3

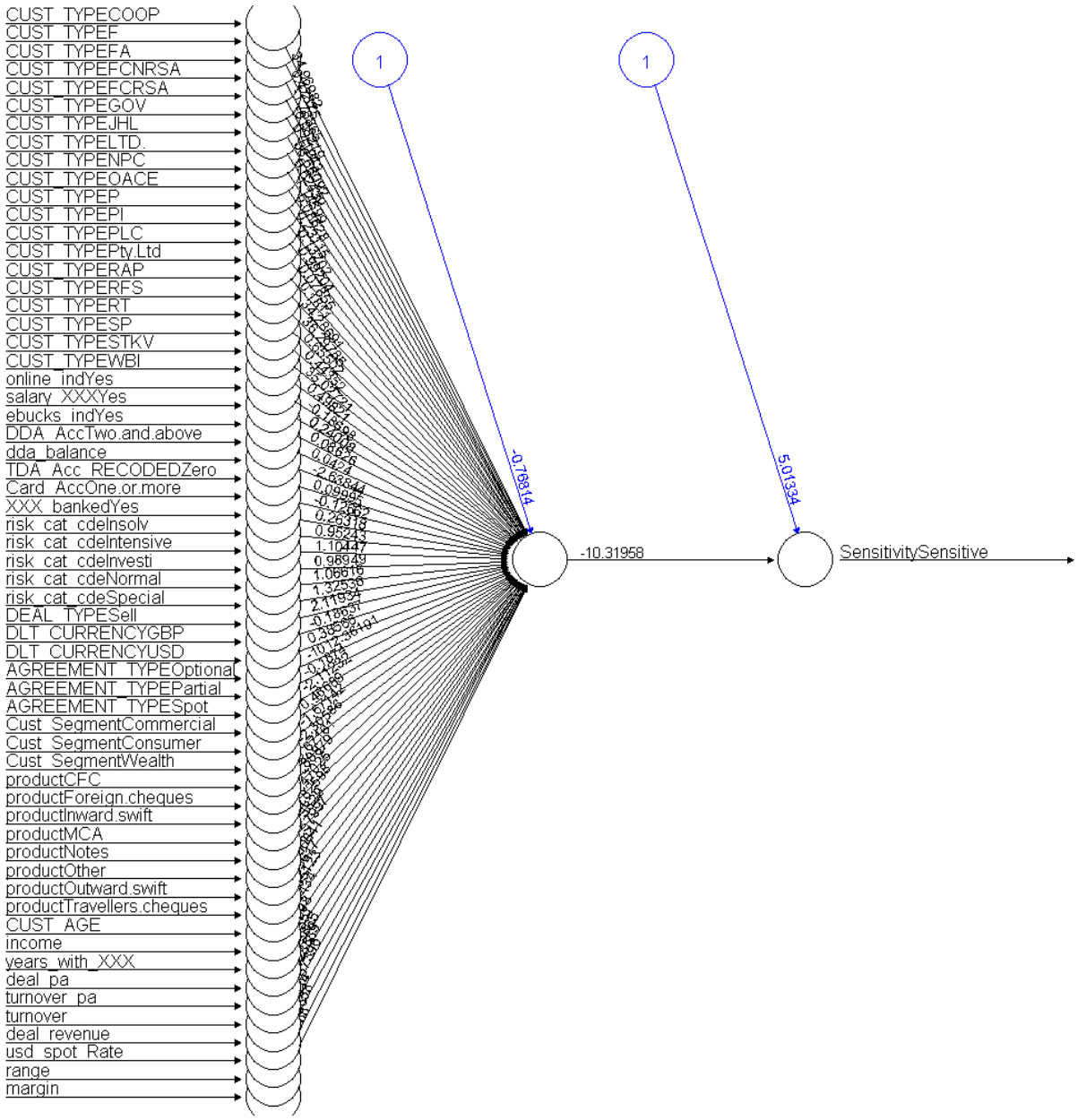

Figure B.19: Topology of the NN - data set 4

B.3.3 Area Under the Curve

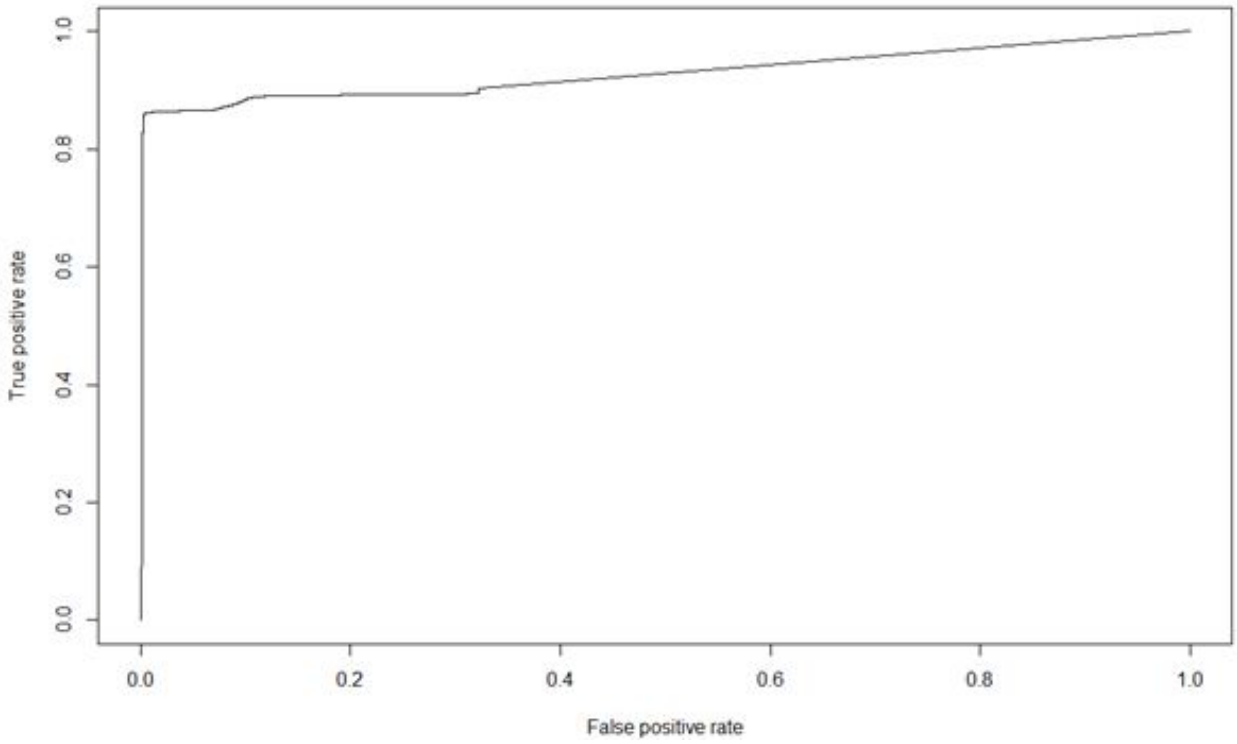

Figure B.20: Area Under the Curve - data set 1

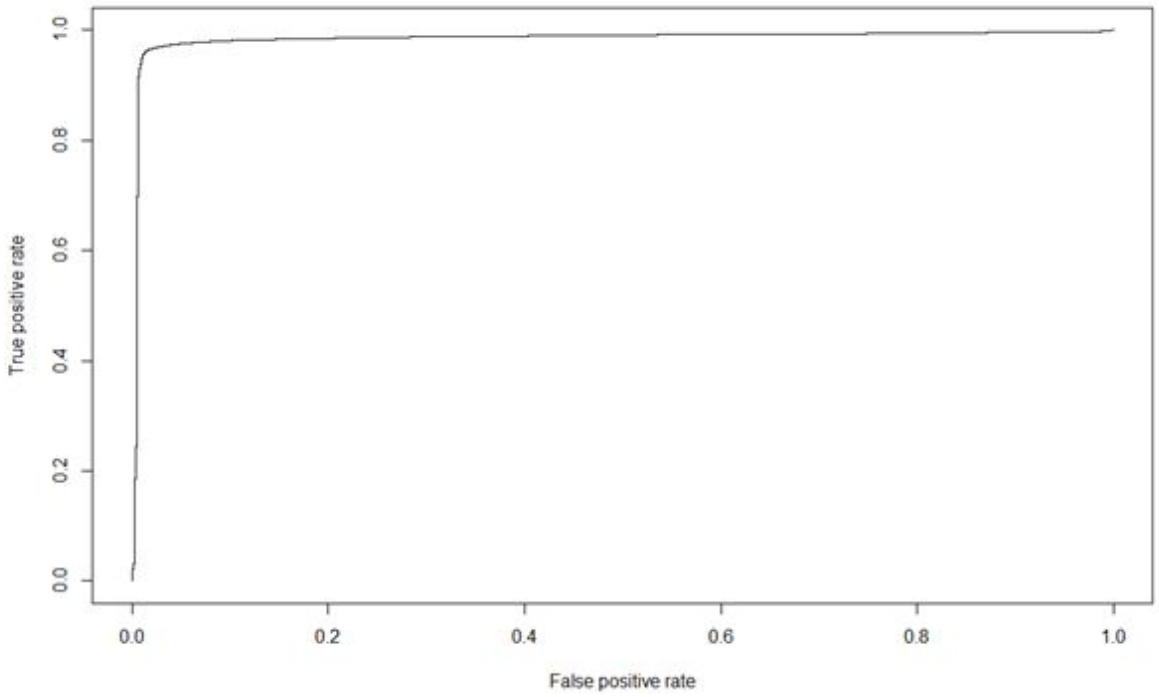

Figure B.21: Area Under the Curve - data set 2

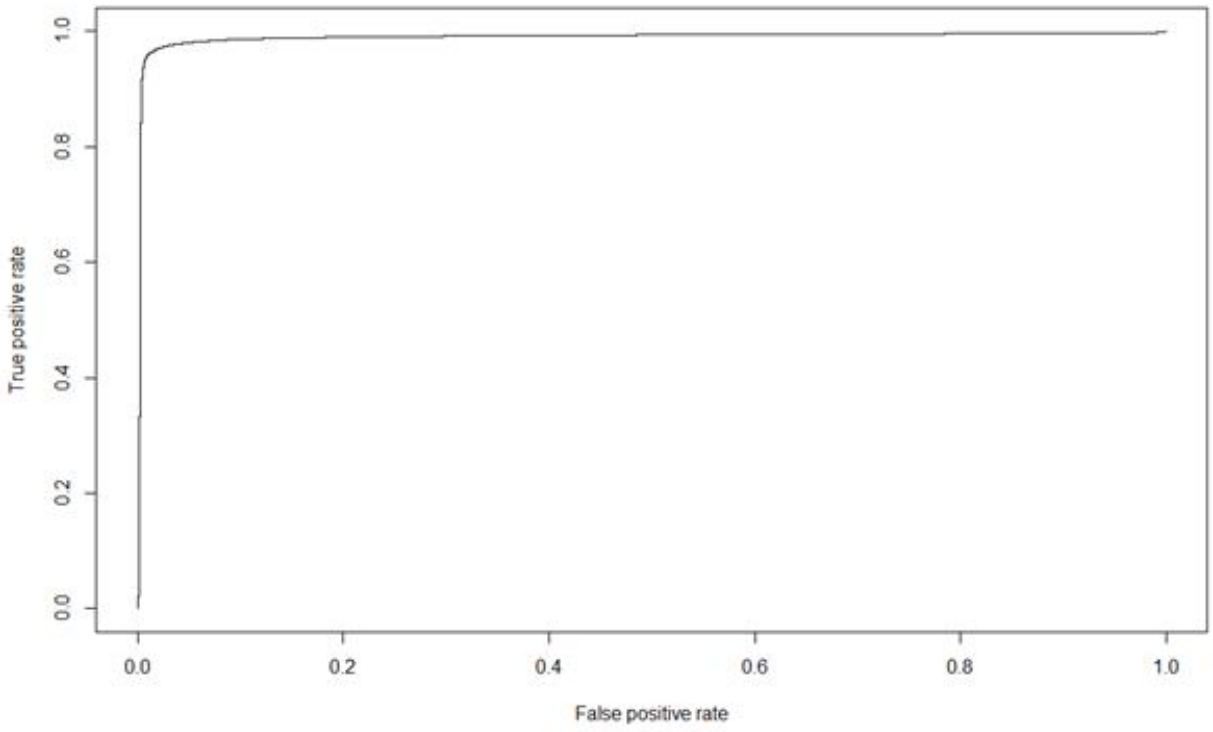

Figure B.22: Area Under the Curve - data set 3

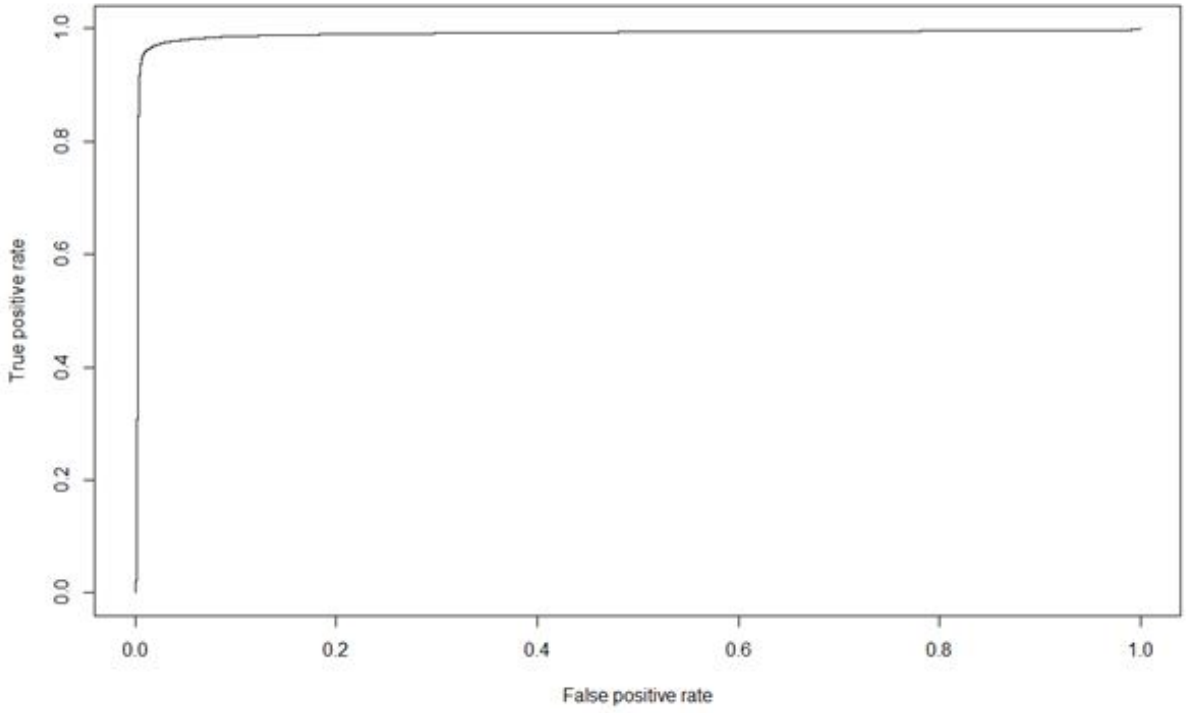

Figure B.23: Area Under the Curve - data set 4

B.4 Imputed Variables

B.4.1 Imputing variable Product using MissForest

Deals.data.imp <- missForest(Deals.data, verbose = TRUE, maxiter = 3 ,ntree= 10) missForest iteration 1 in progress...done! estimated error(s): 0 0.01647108405 difference(s): 0 0.03833909362 time: 43.16 seconds

missForest iteration 2 in progress...done! estimated error(s): 0 0.0165084787 difference(s): 0 0.008716415602 time: 42.89 seconds

missForest iteration 3 in progress...done! estimated error(s): 0 0.01653722551

difference(s): 0 0.009516182048 time: 41.47 seconds

Not sensitive Sensitive 0.7141131373 0.2858868627

Deals.data.imp$OOBerror NRMSE PFC 0.00000000000 0.01653722551

- *Imputed Product*

Notes	316585	37.79%	37.79%
Inward swift	238755	28.50%	66.28%
Outward swift	148597	17.74%	84.02%
Cash passport	45846	5.47%	89.49%
Foreign cheques	75804	9.05%	98.54%
Other	28640	1.46%	100.00%
	837850	100.00%	

Table B.9: Distribution of Imputed Product

- *Imputed DDA Balance*

Min.	R -91,319
1st Qu.	R 1,741
Median	R 14,655
Mean	R 121,081
3rd Qu.	R 106,021
Max.	R 196,736,195

Table B.10: Distribution of Imputed DDA Balance