An Approach to Improve Bank Churn Modelling Using Supervised Machine Learning Algorithms

Minna George Kaiprambadan

3008351

Submitted in partial fulfillment for the degree of

Master of Science in Big Data Management and Analytics

Griffith College Dublin
September, 2020

Under the supervision of

Osama Abushama

Disclaimer

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the Degree of Master of Science in Big Data Management and Analytics at Griffith College Dublin, is entirely my own work and has not been submitted for assessment for an academic purpose at this or any other academic institution other than in partial fulfilment of the requirements of that stated above.

Signed: Minna George **Date: 07/09/2020**

Acknowledgement

This project report is a testament to all the help and the support i have recieved from the people in my pursuits.

I would like to express my heartfelt gratitude to **Osama Abushama** for his help and intellectual guidence without which this project would not have been completed. I would like to thank him for spending his valuable time in mentoring me during the course of this project.

MINNA GEORGE KAIPRAMBADAN

3008351

Table of Contents

List of Figures

List of Tables

Abstract

Customer churn has become a major problem for financial institutions. With continued competition in the market place and the high cost of attracting new customers, companies are under pressure to focus on more effective customer retention strategies. Although the banking and financial sectors have lower churn rates than other sectors, the impact of lost customers on profitability is relatively large. Churn management plays an important role in improving long-term profitability. As a result, more research has been done to understand consumers' switching behaviour and to determine its determinants. In the banking and financial industries, churn management methods have been significantly developed to effectively shape consumerism. Therefore, in an effort to address this issue, we have focused on reviewing and analyzing business banking data from online collections, identifying factors that influence consumers to voluntarily or intentionally abandon them, and then construct various classification models such as random forest, logistic regression, support vector machine, etc.., you can get accurate results efficiently with respective models. Therefore, with the results obtained, we see that random forest outperforms other algorithms in terms of accuracy, outpacing the complexity of the market by providing potential suggestions and keeping the user with strategic action plans for companies and helps to make actionable decisions.

Chapter 1. Introduction

In this world, customer churn has become one of the significant concerns where managers of various associations have perceived the significance and have been striving hard in coming up with new methodologies. Customer churn fundamentally alludes in stopping his/her relationship in proceed the services with the organization. To deal this, each specialist co-op in the business are raising new customer relationship management system and various arrangements as the organizations comprehend that it is adaptable and simple enough in holding the current customers as opposed to securing new and the deserted customers and have been continually attempting to be comparable to the activities beating the rivals in the market. The decision of a CRM (Customer Relationship Management) system always improves the quality of the relationship by also increasing retention. [3]

The prediction of the CRM system and support helps banks to identify the customer. This gives a form of personalization in the service of offers and marketing, which will later help to develop loyalty. It also brings great prosperity and quality to customer interaction. In general, most companies tend to differentiate between voluntary abandonment and involuntary abandonment, where voluntary abandonment refers to the customer’s decision to switch to a service provider or switch to a different company and involuntary abandonment refers to the decision made by the customer. It means withdrawing services for certain unavoidable conditions such as relocation, long-term commitments, death, etc.[2] Furthermore, between the two, all companies ignore voluntary behaviour when analyzing and voluntarily follow the behaviour, as this occurs for a number of reasons, such as the company based on the services provided client relationship, Pre-sale and after-sale. Similarly, after identifying the type of arrogance, analysts within companies also differentiate between overall concessions and pure jealousy when measuring consumer turnover. Aggregate benefits are related to customer losses and tariffs on contracted goods and services that have a term and net nutrition as a whole atrium + net pollution, that is the additional cost that consumers receive and similar revenue is collected. In addition, most business analysts are developing analytics and forecasting models to measure customer satisfaction and better ways to attract a new customer base. Some sectors of the banking and financial sector include

1. Being Customer Centric
2. Fees and Interest Rates
3. Branch Locations
4. Customer Service and Wait Time Reduction
5. Product Offerings
6. Quality of Digital (E-Banking, Mobile banking, etc.)

Some of the above-mentioned areas also inclines with areas of various respective sectors like telecom industry, insurance companies, HRM, retail outlets, etc where the advancement in technologies helps each firm to inculcate the features within the system and its employees in getting new customers and also retaining the existing customers. Likewise, most of the analysts have also inclined employee attrition with customer churn as both of them strongly co-relates with each other with the major motivation of getting satisfied i.e. customer being satisfied from employees of the firm by getting better service and employees being satisfied from the firm by giving better promotions, salary, incentives, etc.[1] Similarly, it is also seen that over the years many researchers and analysts have been trying to predict and analyze customer churn to either reduce or increase the numbers (customers) enabling the firm to take necessary actions. Considering the previous studies as a benchmark laying a strong foundation in understanding the criticality, thus, the aim of this research is building a prediction models using machine learning algorithms by considering various factors and making it worthy enough for potential recommendations.

1.1 Motivation

Since banks are closely connected to our daily lives, such as drawing wages, paying taxes, buying houses, building up investments, and taking out loans, all includes transactions with banks. Any company relies on the banking system. Upon analysing the importance of banking system, personally I felt like upon improving the Customer

Relationship Management [CRM] by some techniques and thereby reduce the customer attrition.

1.2 Research project

In customer churn problems, supervised machine learning techniques been used with SVM poly. Random Forest, Decision trees are most common strategies used to prop up clients. By using SVM model, researchers solve total minimization, high dimensional and non linear problems. Future of the model depends upon the condition and structure of the data. Logistic Regression, Random Forest and Neural networks models are the most widely used algorithms for addressing churn. The literature on data mining research suggests that non-parametric data use the machine learning approach such as neural networks, as they often outperform conventional statistical methods such as quadratic and linear methods of analysis.

For classification problem, statistical classification model like logistic regression is primarily used. Combination of different variables is used to find customer churn accuracy. Random Forest is a good method to find regression problems, classification and uses the bagging technique to get results. To find customer churn four supervised machine learning techniques are used, Logistic Regression, Random Forest, Naive Bayes and SVM. [5]

Research Question: Can we eliminate the difficulty in predicting the potential for customer churn from banking companies and be a better forecast model for taking and identifying preventive measures to reduce customer churn?

To do this, we selected a dataset from an open data repository and proposed to implement a supervised machine learning algorithm analyzed using decision tree, random forest, neural network and performed analysis using models and e their assessment, in the hope of changing the currently accepted methods and strategies. Chapter 2 explains an overview of the literature and related work to help you understand how the previous task of reducing churn was accomplished. The chapter 3 details the methods and chapter 4 focuses on the system design and specification, chapter 5 implementation and chapter 6 testing and last conclusion. [6]

1.3 Research Objectives

The research objectives are as follows–

1) Collecting necessary data on banks’ outflow from bank for research.

2) Understand the dataset, identify different data problems, and then to implement different machine learning algorithms.

3) Prepare the data using cleaning, sampling and removing missing values.

4) Build the supervised machine learning algorithms like SVM, Logistic Regression, Random Forest, Naive Bayes, to see the result of the training dataset.

5) The validation and evaluation of the models in the authentication dataset is based on a matrix that identifies the best model to predict customer churn.

6) Test the best model among all the monitored algorithms in the test dataset and then evaluate the outcomes.

7) Identify limited areas and future research suggestions.

1.4 Dataset Description

Bank churn prediction dataset contains 10000 rows and 14 columns.

Feature	Type	Description and Values
RowNumber	Numeric	Row number with no missing values
CustomerId	Numeric	Distinct customer id with no missing values
Surname	Nominal	Surname of the customer
CreditScore	Numeric	Credit score contains 460 distinct values.
Geography	Nominal	Geography contains 3 distinct values such as France, Spain, Germany
Gender	Nominal	Gender of the customer
Age	Numeric	Age of the customer contains 70 distinct values
Tenure	Numeric	Tenure contains 11 distinct values
Balance	Numeric	Bank balance of the customer which contains 6382 distinct values
NumOfProducts	Numeric	Number of products contains 4 distinct values mentioned as 1 to 4
HasCrCard	Numeric	HasCrCard contains 2 distinct values 0 and 1
IsActiveMember	Numeric	This also contains 2 distinct values like 0 and 1
EstimatedSalary	Numeric	Estimated salary of the customer and it is distinct for everyone
Exited	Numeric	Exited explains customer exited or not such as 0 and 1

Table 1 Dataset Description

Chapter 2. Background

2.1 Literature Review

This chapter provides an overview of the available literature on methods for predicting bank churn, the different approaches to dealing with the problem, and the valuation matrix used to test the models. This chapter concludes with a gap in current research and sets out the purpose of the study.

2.1.1 Background

To improve the future revenue and understanding the customers, predicting customer churn rate is important in all business sectors. It develops good business when understand your customer lacking. For this, lot of work has to be done and there is lot of customer data from different industries for study and the output vary by industry. Customer service means that customer who transfers one sale to other. The attrition rate of customers forecast is used proactively find churn rates of potential customers before they move other company.[7]For retention policies this step help the company plan needed to attract likely churn rates, reducing the company's financial losses.

The loss of customers is a concern for many industries and is especially acute in highly competitive industries. Consumer losses result in financial losses due to falling revenue and a rising need for new clients. Retaining customers is critical number of industries because it is more costly to gain new customers than retention. Due to customer surprise, tingling whether or not a customer leaves the company is quite a daunting task. For financial institutions, determining the number of 10 customers is more complicated because of the small amount of data compared to other domain. This requires further investigation periods for migration forecasts.[9]

Customer churning is a major problem in banking, and banks have always tried to monitor clients and interactions in order to identify clients who are move the bank.

Customer attrition modelling mainly focuses on exiting customers and measures can be taken to prevent churn.

In the age of competition, more and more companies recognize that their most valuable assets are their existing customer base and their data. Examine churn predictors primarily in the context of Customer Relationship Management (CRM). Bank attrition management is an major task to keep your valuable customers.[8] Business associations such as banks, insurance companies and other service providers have established strategies to transform their employees into more customer and service oriented and secure customer loyalty. The best marketing strategy for the future is to keep existing customers and avoid them.

2.1.2 Data Exploration and Pre-processing

Data mining is essential for a better understanding of data and bank issues. The CRISP-DM method for data mining models is widely accepted. It is primarily for performing a data mining process, with a life span of six stages, as shown in the figure below.

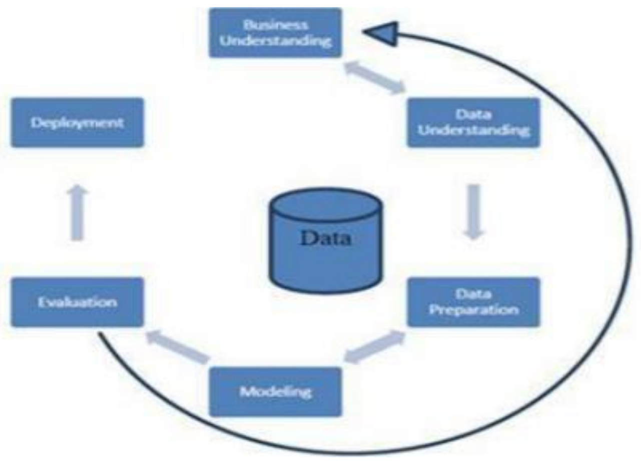


Figure 1 CRISP-DM phases

The most important step in data analysis is database planning. In general, it takes around 80% of the time to clean and pre-process data. Preparing data is a component which is more complicated and time consuming. Data from the real world can be chaotic, incomplete and incoherent. The process of the data preparation deals with:

incomplete data where certain adjective values were missing, where certain significant attributes were missing. During the data preparation phase, data observations and errors were also handled, including data discrepancies. Data creation produces a smaller data set than the original.[10] This work involves the selection of relevant data, the selection of attributes, the removal of discrepancies, and the removal of duplicate records. This stage is also related to replacing lost values, reducing ambiguity, and eliminating outsiders.

(1) The actual data is inaccurate.
(2) Standard data is required for high performance mining.
(3) High quality samples produce high quality data.

2.1.3 Machine Learning Techniques

Logistic Regression

Logistic regression has proved to be a strong algorithm and has become a widely used predictive model for clients. The formula in Figure 2 below is the logistical regression, pi is the likelihood and the independent variables predict the results. [11]

$$p_i = \frac{e^{x_i \beta}}{1 + \; e^{x_i \beta'}}$$

Figure 2 Logistic Regression

A study on predictions about the credit card market in China's banking industry modelled the Decision Tree and logistic regression. It was found that Logistical Regression performed better than the Decision Tree. The models were collaborative and in the study, it was found that the deduction tree algorithm shows less performance compared to logistic regression. [12] The researchers applied binary and general logistic regression models to predict customer brainstorming using SAS 9.2 using the logistic regression process and the Cox regression model.

Another study for predicting customer churn in telecommunications datasets found that logistic regression models outpaced the decision tree. Logistic regression makes the maximum use of logistic regression to convert the dependent variable into a logistic variable. The proposed method offered a statistical analysis tool for the estimation of consumer churns. The matrix of the uncertainty was used for assessment purposes.

Random Forest

Random Forest is a comprehensive learning approach to the problem of classification or regression. The decision tree is the building block of a random forest. Decision tree congestion creates random forestry and is a class tool in production classification and forecasting the recession problem. Random forests may not be overly compatible with a large number of trees. [13] It can deal with missing values and suitable for hierarchical variable as well. The research has been done previously on financial consumerism, random forestry classification techniques has been used.

Support Vector Machine

The software Support Vector Machine can be used for classification and regression problem since it is a machine learning software. SVMs are commonly used in classification problems, as they can distinguish two classes using hyperplanes. SVM's goal is to find a hyperplane which clearly classifies the data. Hyperplanes are boundaries of decision making which help to classify data points. The support vector is the nearest data point to the hyperplane, which influences the hyperplane 's direction which orientation. Some researchers used two methods to predict brainstorming among consumers. The first approach was a traditional method of classification using observational learning primarily for quantitative data and the second was an artificial method for non-linear, large-scale intelligence, large-magnitude, non-linear time series and time series data. It has used the SVM model as it can solve problems of non-linearity, high dimension and local reduction. The prediction of the model is based on the structure and position of the data.[14]

Neural Network

Neural networks are a set of algorithms that can be used for pattern recognition. Neurons are the fundamental building blocks of a neural network. The production depends on the activation mechanism of the neurons. For his research into predicting consumer turnover in banking, researcher uses a neural network model in the Alyuda NeuroInteligence software kit, as the neural network is well suited for pattern recognition , image analysis, optimization problems, etc. Another group of researchers suggested comparing common modeling techniques - multilayered perceptronal neural networks and decision trees - with innovative modeling techniques - SVM for predicting customer churn in the telecommunications industry. MLP and SVM were more efficient than decision trees. [17]

Model Evaluation

The problem of consumer intelligence forecasting is a classification problem and was calculated using the confusion matrix to evaluate the performance of the observed machine learning models. The AUC also have been used by many researchers in addition to the confusion matrix to assess the model. AUC is the area under the receiver rating operating curve (ROC) curve. ROC is the plot between false positive rate and true positive rate. The second evaluation matrix was the TDL (top-diesel lift), which is focused on the most churning customers. [15]

Prediction of Customer churns using Machine Learning

As technologically advanced era, begins the data is being expanded immensely and became a hectic challenge to data analysts to analyse huge volume of data. Churn rate of customers can be predicted by using machine learning and data mining techniques and this plays an important role. It follows a series of steps to predict consumer churn using machine learning models. The data is collected, and then the selected data is pre-processed and converted into an appropriate form for creating a model of machine learning. After modelling, tests were carried out and the model was finally deployed. Machine learning analyzed the data and identified the underlying data trends for evaluating the rate of customer churn. [16] Using machine learning, customer churn prediction was more accurate than conventional method.

2.2 Related Work

With the growing need to identify potential customers, many researchers and analysts have been instructed to develop strategic models and forecasting models. Most importantly, companies recognize that losing a customer is a low-cost opportunity for competitors in the market to gain a new customer. In addition, companies have developed state-of-the-art CRM systems for handling, storing and converting consumer data into valuable information-that is, to identify behaviours before the user is lost because the market the customer mantra is growing rapidly in every domain In recent years, some researchers have not only focused on implementing predictive models, but have also developed some strategies. One of them is targeting active retention, which means doing a thorough analysis to predict which customers will turn off the service, and then decide which marketing plan to retain on the customer. And focus on both customer traction predictions. And improve the marketing process. Some degree of theory and flow such as attribution dependence and accuracy of estimation in information systems, material and classification uncertainty, shortcomings, decision rules, modelling of buildings to make predictions about potential credit card users studied the use of network graph concepts. Appropriately, charismatic and compassionate focused on a conceptual model for identifying customer-linked spells, which clearly agrees with the data from which blood models are selected, and is illustrated below.

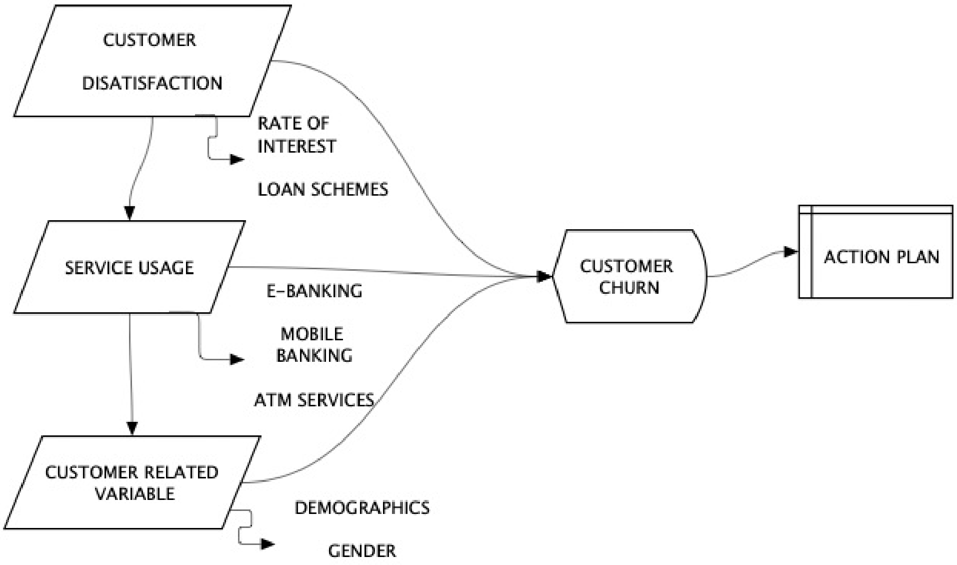


Figure 3 Customers behaviour on financial sector

Researchers have already reviewed and investigated patterns of behaviour on consumers in financial services. As mentioned above, churned customers are those who have closed their bank accounts. Churn research contributes to three levels: consumer behaviour, consumer demographics and the macro environment. Second, most studies to date have analysed churn by collecting churn data through questionnaires and observing the total population during this period. The third is to observe when the customer has left the account or the relationship with the company or the bank until the customer leaves or the end of the analysis. [18]

Research suggests that customer dissatisfaction is a major problem driving them to switch banks. Usually, the increase in fees is manifested as a way to mark the dissatisfaction of banks among customers and change banks to obtain more favorable prices. Market share and bank profits will suffer in the event of customer loss. also recognized some of the elements of customer dissatisfaction with banking, namely the interest rate compared to the competition, loan systems, ease of use, profitability, security, problem management, efficiency, etc. and states that customer dissatisfaction leads to desertion if the company does not comply with the above elements. Then, the level of use of the service would be like credit card, electronic banking, mobile banking, kiosks, etc., which also play a vital role in building relationships and trust with the customer. , then the variables related to the client such as client income, age, nationality, occupation, demographics, gender, etc. inadvertently, that could be one of the reasons the services were withdrawn.[19] Therefore, the figure shows that all three or one of the factors could influence a client to unsubscribe services, which raises the importance of having a biased prior knowledge of the client’s activity in the deployment of plans.

2.2.1 Critical Review of the Related Work

This part of the study helps in understanding various machine learning algorithms as well as forecasting models implemented by previous researchers to identify customer churns in previous years, thereby outperforming historical methods in this study and Banking helps to provide a better model for making potential proposals. Pioneered some research to detect churn, in which researchers looked at the Pareto principle of the 80:20 rule and applied the association rule method to identify similar patterns among abandoned customers when initiating regulatory action, And then used several advanced methods for addressing customer churn. Focused on previous literature on the use of a better-balanced random forest with its effectiveness and more to make accurate predictions by integrating it into sampling and cost-sensitive learning. Emphasized historical churn modeling between different mining techniques, namely the binary logit model, and applied the Generalized Additive Model (GAM), which they compared using two different matrices using logistic regression and curate with metrics like area under curve, Receiver Operating Characteristics (RC), root mean square error (MSE) and RMES, etc.[20] First, there is an opportunity to reveal nonlinear structures in the data that can often be ignored by relaxing general assumptions. Stated constructs may provide insight into the effect of the covariate.

Second, GAM data allows identifying more complex relationships between cooperative and dependent variables, making predictions more accurate and precise, helping companies add value to their business. Meanwhile, proposed to use a Bayesian validation network using a sample of variables and continuous variables using the CHAID algorithm, which helps to analyze the correlation between data and evaluate it using sensitivity analysis, discussed traditional classification models and proposed principles as part of a model selection strategy, starting with the good performance theory, which measures accuracy for overall accuracy, completeness, and model evaluation. Second, the classification time and model effectively evaluates how the customer departs and adheres to the principle of ranking pattern, taking into account the principle of efficiency by calculating the total time, and observing the pattern in customer behaviour. And to estimate the ease of feasibility distribution; And finally the principle of making rules by setting a set of assumptions and constraints on the data to implement and achieve high accuracy.

Meanwhile, reviewed the findings on customer churn and proceeded to implement classification algorithms, focusing on several areas such as root cause, retention, value models, and model expectations. I have identified what supports the machine. (SVM) outperforms the Bayesian hub classification algorithm and integrates it with the customer lifetime value model.[21] In addition, focus on the most commonly used analytical methods in the literature on churn, namely classification and logistic regression trees, and the time to find and evaluate models between clients. Implemented a frame strategy model using metric and also focused on classification methods, implemented SVMs, and evaluated using accuracy and recall measures on various kernels to determine the best model. Similarly, amines (2015) also implemented SVM in 2015 and became obsessed with scoring metrics according to the old method.

In addition, most researchers focused on comparing machine learning algorithms, one of which was comparison and optimization of algorithms using validation techniques through study, enhancing algorithms and evaluating most metrics and monitoring with machine learning algorithms will improve the performance of the model over the years. In addition, with recent advances in algebra, came up with a biased approach to the hybrid classification algorithm based on decisive tree and logistical regression and focuses more on the diagnostic matrix based on the search for powder literature.

In addition to using the most commonly used methods to predict customer churn rates, took an approach using a predictive system with Apache Spark and compared the results with the MLlib package to identify the best of the selected methods, and the

Mllib package with the RDD API Longer training times like the M1 with APIs based on data frames reduce test time and give the desired accuracy. Therefore, it is not possible to glimpse the previous literature, understanding how the processing of bank details and data of various clients becomes a mysterious task for analysts, and ultimately how to find a solution by presenting an action plan, always presented as the most important concern for banking companies.

Nowadays, picking from any industry perspective tends to confront the mindset of both the employee and the customer as the two become best friends together, who is why most of their research has been done by companies through employees and adopted strategies to deal with the use of mistrust. Overcoming consumer frustration because they both rely on the same approach means that finding an alternative will cost as much as paying back to those valued customers who will experience it with employees when someone the market becomes the need of any industry both of them. Furthermore, previous studies have shown that there are some factors for potential issues such as financial status for consumers, demographic sector, etc., which cannot be considered to feed the model of features. Making predictions and companies also believe that it does the same thing to frustrate employees because unnecessary reasons cannot be considered, which is beyond the scope of coping.[22] Therefore, the overall motivation for the research in the paper is the work done by the analysts and the most discussed topics in the analytical field are expected to help the business move forward with better plans and policies.

Chapter 3. Methodology

This section will describe the experimental methods and behaviour in conducting this research project. This project is based on the CRISP-DM approach where the data was first downloaded from the Kaggle repositories. This project is closely related to data pre-processing performed by exploratory data analysis to understand the variables that apply data transformation techniques that visualize the result and then evaluate the results.

3.1 Tools and Technologies

3.1.1 R Programming Language

R is a free software environment for programming language and statistical computing and graphics, supported by the R Foundation for Statistical Computing. The R language is widely used by statisticians and data miners to develop statistical software and data analysis. It is one of the most popular languages used by statisticians, data analysts, researchers, and marketers to extract, cleanse, analyze, visualize, and present data.

Open Source: R is a programming language which is open source. That means anyone can work with R without having to pay a license or fee. You may also help improve R by improving packages, designing new ones and solving problems.

Highly compatible: R is strongly compatible with many other programming languages including C, C++ , Java and Python. It can also be combined with Hadoop technologies and numerous other database management systems.

3.1.2 R Shiny

Shiny comes with a flexible programming library that will be used to build the logic of your application. Changing the input values will automatically cause the correct sections of the R code to be re-executed while using this library, which will in turn cause any changed output to be updated and a foundation for web application building using R. Shiny helps you turn your analytics into interactive web applications without the need for HTML, CSS, or JavaScript knowledge.

Chapter 4. System Design and Specifications

This section of the article focuses on the details of the methods adopted as part of the implementation process to produce successful predictive models, all of which begin with collecting data based on a well-thought-out approach, then performing exploratory data analysis needed to help build models that meet business needs. In terms of business perspective, the study clearly leans towards the CRISP-DM method (Cross Industry Standard Process for Data Mining) which grants a greater scope in the extraction of knowledge from the chosen data emphasizing the use of algorithms. machine learning to obtain prediction models. The following figure details the study approach in 5 different phases, integrating it into a CRISP-DM framework. [23]

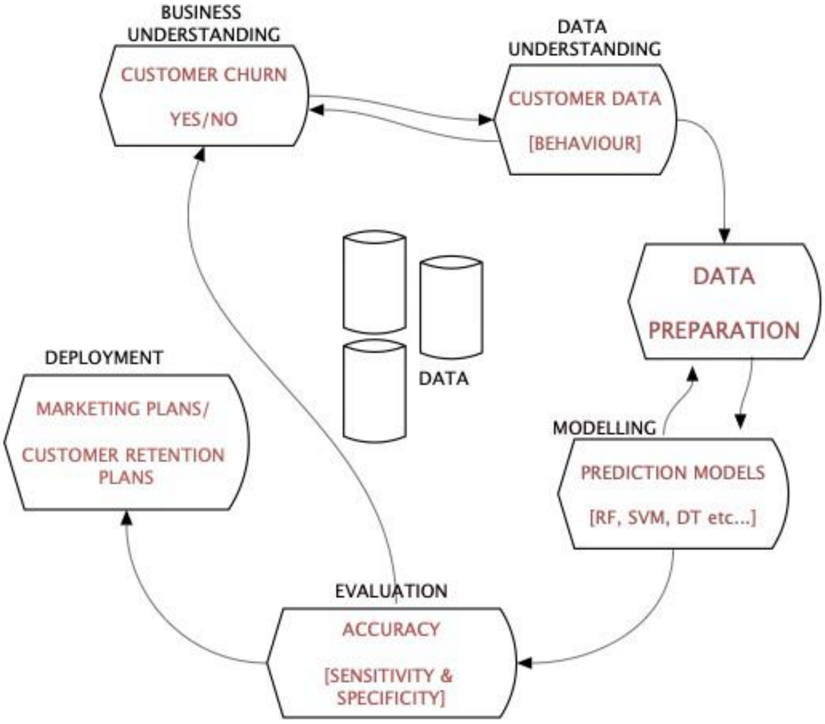


Figure 4 CRISP-DM Framework

As you can see in the figure, the business understanding phase begins with the availability of systems that meet the business needs, and we determine that the main requirement of banking companies is to find the client with or without a possible customer attitude, followed by the data understanding stage means. Analysis of customer behaviour in terms of actions, use of services, correlation factors contributing to the achievement of the target feature, its preparation and provision for building models contribute greatly. In the modelling phase, we create classification models such as random forest, Bayesian classifier, support vector machine, artificial neural network, etc. to compare when looking for algorithms that go beyond the desired precision, and the next step is building assessment. Models with metrics such as uc, rock, sensitivity, specificity, etc. to see how well the model predicts.[23] The final step in the approach is for companies to come up with strategies and deploy a model in terms of business standards with knowledge base to meet business requirements for timely action such as customer retention plans, marketing activities and advanced CRM systems.

4.1 Exploratory Data Analysis

This part of the study involves the selection of data based on the assumptions and speculations that have been made to meet the research objective. First, we selected data from the online data store Kaggle.com, where the next step was to understand the data and its properties in order to obtain a relationship with the key target variable. It also usually includes some basics about the customer's behaviour, why it leaves them and then follows a choice with the algorithm. Because the target in the data was variable, the classification technique was the most commonly chosen technique, and history is dominant with literature in this regard. Then, the implementation l. Processing steps have already been taken to detail and describe the assessment measurements required to determine the scope of the data and to evaluate it.[24]

Step 1: Data Collection

Step 2: Data Pre-processing phase

(i) Feature selection has been done corresponding to the target variable.

(ii) Factors conversion has been done into binary form 1 and 0 based on the need as machine works well with numbers.

(iii) Required libraries have been installed.

(iv)Checked correlation matrix and checked for the variables which are not statistically significant and discarded it.

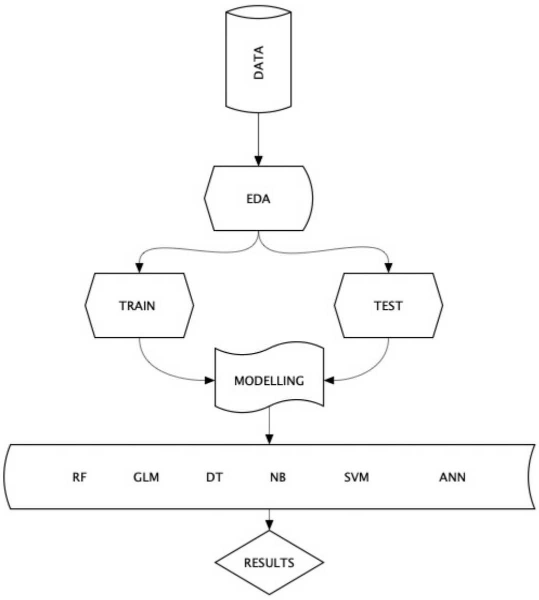


Figure 5 Exploratory Data Analysis

Step 3: Segregated the data into train and test sets in the ratio 80 percentage and 20 percentage respectively and started building classification models. Since, we had a Predefined target column as customer exited or no in binary format because of which modelling was obvious with classification technique when compared to regression.
Step 4: Build models respectively and evaluated them accordingly.

4.2 Prediction Models

Decision tree and random forest:- The random forest is considered as a combination of predictor trees in such a way that each of the trees depends on the values of the random vector, which is sampled independently also with the same distribution for all. forest trees According to the previous literature, the random forest and the decision tree as a joint learning method for regression, classification and other tasks that have been very dominant due to their efficiency generated in various machine learning applications in which many decision trees have been driven by a random forest that divides the data into partitions, so none of the partitioned trees include the full training data. Furthermore, the visual representation that explicitly represents the decision in the analysis makes the algorithm one of the best chosen methods in business practice to overcome prediction systems. In general, in these trees the leaves represent the class labels and the branches represent the characteristics that lead the class labels.

A random forest machine learning technique was chosen to predict the consumer churn rate. It is a combination of many decisive trees. This is a configuration learning method (group of decision trees) to use classification, regression problems, and bagging techniques to produce this result. In learning together, a group of weak learners together form a strong learner. Bagging, also known as bootstrap integration, is reduced to a decision tree. It is a combination of machine learning that is used to generate precise results by combining predictions of specific machining algorithms. Random Forest's default default hyperparameter works well and is excellent for avoiding overheating.[25] In the random forest, the most common output result of all decision trees was selected as the predictive class.

Navy Base Rating: The Navy Base Rating is used to understand and understand the probability of an event given a pre-existing event that has already occurred in Rest Eight. (2001) Many researchers are following the trend of bid-based classifier algorithms as a simple probabilistic model based solely on its strong empirical hypotheses based on the full application of Byers' theory. Statistics are independent of each other in any class .With this study, it is generally assumed that the presence or absence characteristic of the class (customer mantra) is unrelated or independent of the absence or presence of another class, when it is proved in a previous work, the result is significant.

Support Vector Machine: Support Vector Machine, developed by Vapnik, Boser, and Guyon, also known as Support Vector Networks, is supervised machine learning models that are widely used to analyze and recognize the pattern in data for various classification and regression studies. In SVM, minimizing structural risk acts as a key factor and improves performance through the use of various kernel functions.

The Support Vector Machine (SVM) is a supervised machine learning model with associated learning algorithms that analyze data for classification or regression problems. This algorithm works with the kernel function. [26] The data is converted based on the kernel function and a maximum limit is set between possible outputs and used SVM to predict churning in telecommunications customer data. SVMA solved the problems of non-linearity, high parameters and local minimization in predicting consumer attitudes. According to available studies, SVM can also work well with financial customer data sets.

Generalized Linear Model (GLM): The generalized linear model was designed by John and Robert to unify various statistical models, including some regression techniques. Statistically, GLM has the flexibility to normalize linear regression, taking responses with the diffusion model versus normal, that is, by allowing variance as a function of the predicted value for each measurement.

Artificial Neural Network (ANN): Over the years, ANN has been one of the most popular algorithms for processing complex data and models such as process modelling. This driven model can use a variety of topologies and learning algorithms. One of them is the Multiplayer Preceptor which is trained in various forms of backpass algorithm (BPN). In the case of this study, we see that ANN lags behind NBC and the linear model in achieving the desired results.

Logistic Regression: First, a logistic regression model was created. This is the most preferred algorithm for modeling binary dependent variables. This is a type of statistical probabilistic classification model used primarily for classification problems. This technique works well with different combinations of variables and helps to more accurately predict customer churn. You can calculate the predictive power of a variable. This is a statistical model with curves fitted to the dataset. This technique is useful when the target variable bisects. This is a predictive analytics algorithm based on the concept of probability. [27]

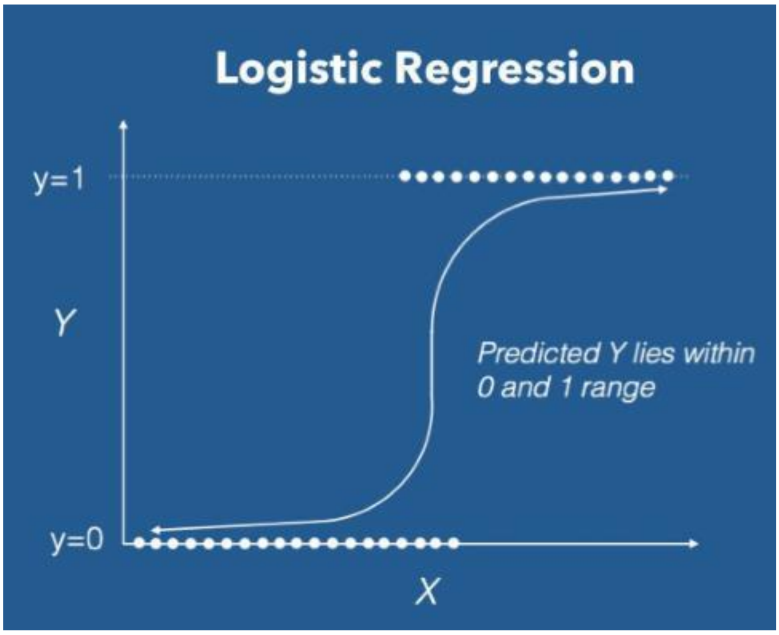


Figure 6 Logistic Regression

Logistic regression uses a complex cost function that is defined as a "sigmoid function" or also known as a "logistic function." The sigmoid function was used to compare predictions with probabilities. Restricts the output and returns a probability estimate of 0 to 1. Logistic regression was one of the popular algorithms for classification problems. [31]

4.3 Machine Learning

Machine learning is a method of data analysis that helps you to develop analytical models. This is the field of artificial intelligence (AI). Machine learning models identify common patterns in data to make decisions with minimal human intervention. Machine learning is mainly used when you have a complex problem or task that involves a large amount of data. This is a good option for more complex data and provides faster and more accurate results. [28] It helps organizations identify profitable opportunities or unknown threats.
Machine learning mainly uses two types of learning techniques:

1) Supervised Machine Learning

2) Unsupervised Machine Learning

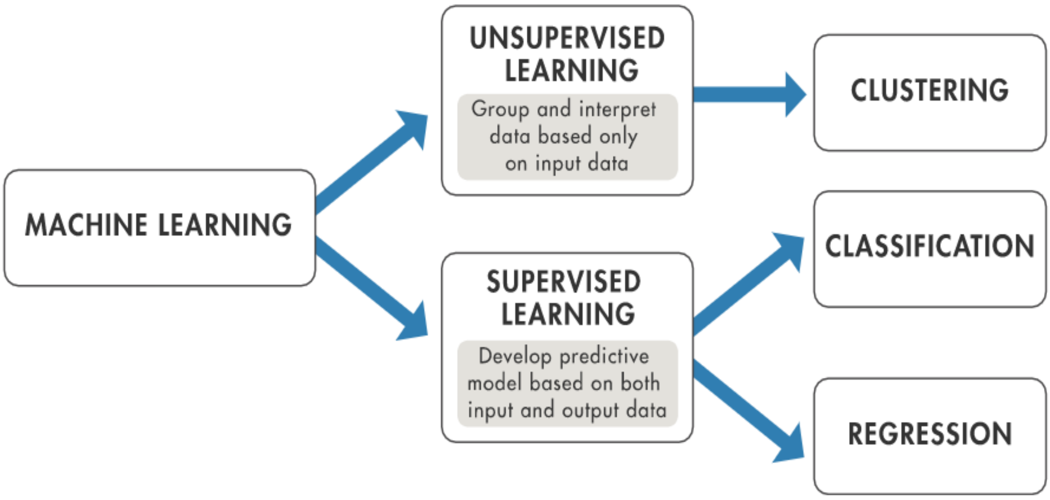

MACHINE LEARNING
UNSUPERVISED LEARNING
Group and interpret data based only on input data
SUPERVISED LEARNING
Develop predictive model based on both input and output data
CLUSTERING
CLASSIFICATION
REGRESSION

Figure 7 Types of Machine Learning

Supervised Machine Learning

Supervised machine learning is the computational task of learning the correlation between variables in a training dataset and then using this information to create a forecasting model to enable annotation of new data. In monitoring machine learning, we have an input variable (X) and a variable output variable (Y), and we use the algorithm to learn the mapping from input to output. [30]

Y = f(X)

The goal is to properly estimate the mapping function so that when new input data (x) is introduced, the output variable (y) for that data is predicted. If a well-known labeled example is provided, the study is called observational guidance. Features can be continuous, categorized, or binary.

The supervised learning problems can be grouped into regression and classification –

1) Classification - If the output variable is categorical, e.g. "red" or "blue" and "yes" or "no" this is considered a classification problem.

2) Regression - If the output variable is a real value, these problems are considered regression problems.

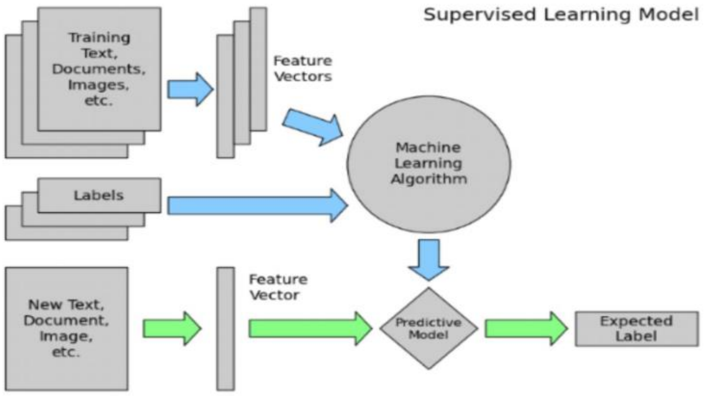

Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

Figure 8 Supervised Learning Model

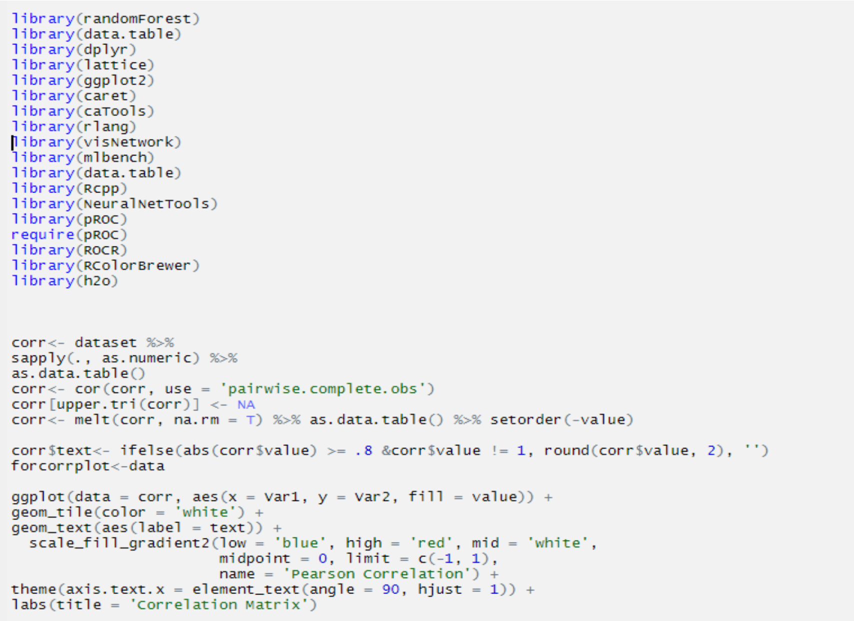
Chapter 5. Implementation

This section describes the implementation of your chosen algorithm and critically interprets the results achieved to suit your business needs. For implementation, I used R Studio to build a model for R and also used MS Excel to perform analyzes and fined relationships between variables. The figure below shows the details of the Pearson correlation matrix for understanding the variables, which are correlated and range from -1 to 1.

This chapter explains how the experiment was carried out in accordance with the measures outlined in Chapter 3. Includes all stages of pre-processing the data, how to construct a model of machine learning. This chapter also explains how each supervised model of machine learning is calculated. [29] The study was conducted using the Data Mining Cross Industry Standard Process (CRSP-DM) tool, providing the data miners with a standardized framework and guidance. This approach consists of six stages or phases: understanding the market, understanding data, preparing data, modelling, evaluating and deploying.

5.1 Correlation Analysis

Correlation Analysis of the data was carried out to determine the correlation between the independent and dependent variable and to establish multicollinearity between the independent characteristics. The correlation approach 'Spearman' has been used to define correlation, as the dissertation consists of both continuous and categorical variables. The heatmap of correlation was produced, as shown in the figure below.



```
library(randomForest)
library(data.table)
library(dplyr)
library(lattice)
library(ggplot2)
library(caret)
library(caTools)
library(rlang)
library(visNetwork)
library(mlbench)
library(data.table)
library(Rcpp)
library(NeuralNetTools)
library(pROC)
require(pROC)
library(ROCR)
library(RColorBrewer)
library(h2o)



corr<- dataset %>%
sapply(., as.numeric) %>%
as.data.table()
corr<- cor(corr, use = 'pairwise.complete.obs')
corr[upper.tri(corr)] <- NA
corr<- melt(corr, na.rm = T) %>% as.data.table() %>% setorder(-value)

corr$text<- ifelse(abs(corr$value) >= .8 &corr$value != 1, round(corr$value, 2), '')
forcorrplot<-data

ggplot(data = corr, aes(x = Var1, y = Var2, fill = value)) +
geom_tile(color = 'white') +
geom_text(aes(label = text)) +
  scale_fill_gradient2(low = 'blue', high = 'red', mid = 'white',
                       midpoint = 0, limit = c(-1, 1),
                       name = 'Pearson Correlation') +
theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
labs(title = 'Correlation Matrix')
```

Figure 9 Library Loading

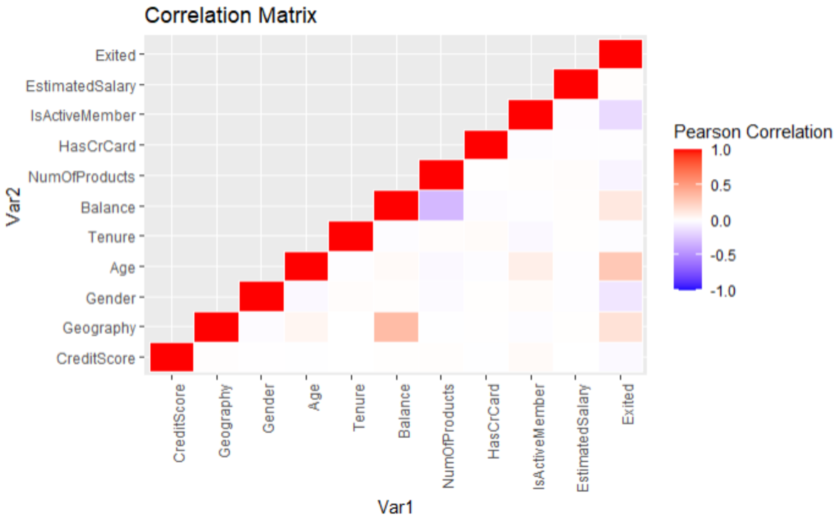
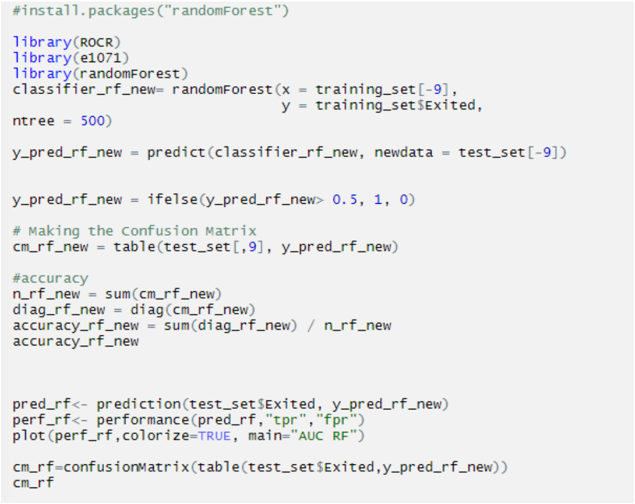


Figure 10 Correlation Matrix

The figure also depicts that the feature has credit card and estimated salary doesn’t make much impact on the target variable exited based on the values in the dataset.

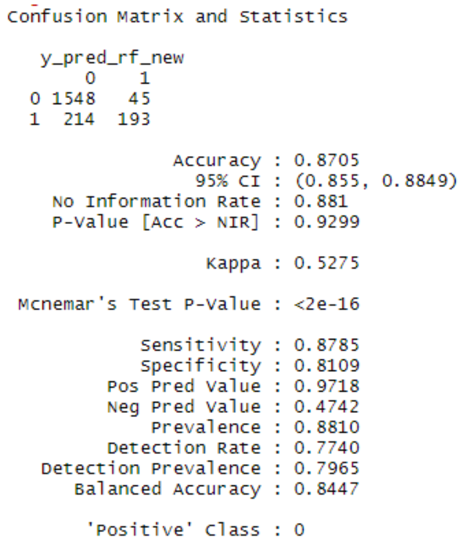
5.2 Random Forest Interpretation

Random forest classifier model has been built using random Forest package in R studio by setting up a tree value of 500, followed with building a confusion matrix to understand the number of true positive, false positive respectively. Therefore, an accuracy of 87.05% percentage was achieved by the model generating a kappa value of 0.52 which insights that the model performance is good/moderate reading all the features with test and train sets. Meanwhile, the model also achieved 87.85% of sensitivity(recall) and 81.09% of specificity which also portrays better model performance.



```
#install.packages("randomForest")

library(ROCR)
library(e1071)
library(randomForest)
classifier_rf_new= randomForest(x = training_set[-9],
                                y = training_set$Exited,
ntree = 500)

y_pred_rf_new = predict(classifier_rf_new, newdata = test_set[-9])


y_pred_rf_new = ifelse(y_pred_rf_new> 0.5, 1, 0)

# Making the Confusion Matrix
cm_rf_new = table(test_set[,9], y_pred_rf_new)

#accuracy
n_rf_new = sum(cm_rf_new)
diag_rf_new = diag(cm_rf_new)
accuracy_rf_new = sum(diag_rf_new) / n_rf_new
accuracy_rf_new



pred_rf<- prediction(test_set$Exited, y_pred_rf_new)
perf_rf<- performance(pred_rf,"tpr","fpr")
plot(perf_rf,colorize=TRUE, main="AUC RF")

cm_rf=confusionMatrix(table(test_set$Exited,y_pred_rf_new))
cm_rf
```

Figure 11 Random Forest Implementation



```
Confusion Matrix and Statistics

   y_pred_rf_new
       0    1
  0 1548   45
  1  214  193

               Accuracy : 0.8705
                 95% CI : (0.855, 0.8849)
    No Information Rate : 0.881
    P-Value [Acc > NIR] : 0.9299

                  Kappa : 0.5275

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8785
            Specificity : 0.8109
         Pos Pred Value : 0.9718
         Neg Pred Value : 0.4742
             Prevalence : 0.8810
         Detection Rate : 0.7740
   Detection Prevalence : 0.7965
      Balanced Accuracy : 0.8447

       'Positive' Class : 0
```

Figure 12 Random Forest : Accuracy

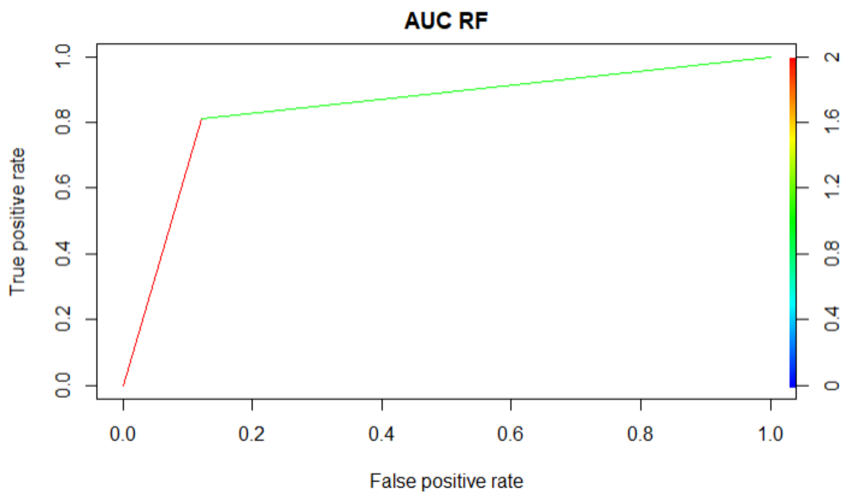
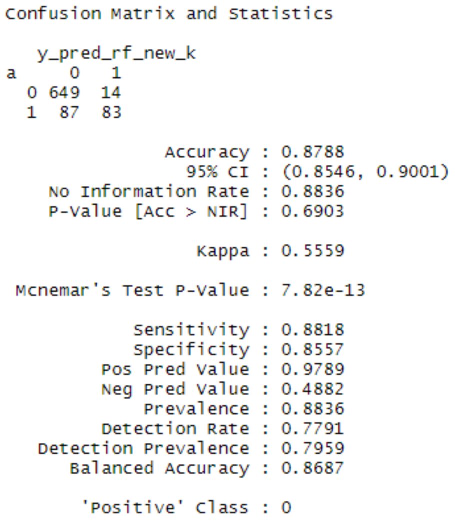


Figure 13 Random Forest Graph

To evaluate more accurately, we plotted the area under the curve (auc) to sight the skewness mapping it with the obtained accuracy with true positive and false positive rates. Further, we tried optimizing the algorithm using K-fold validation and could see that the accuracy was increased with some margin i.e. 87.78% and also with an increase in sensitivity and specificity values of 87.91% and 86.87%. Thus, we could see that validation techniques works better in churn modelling considering all the features enhancing the model performance.

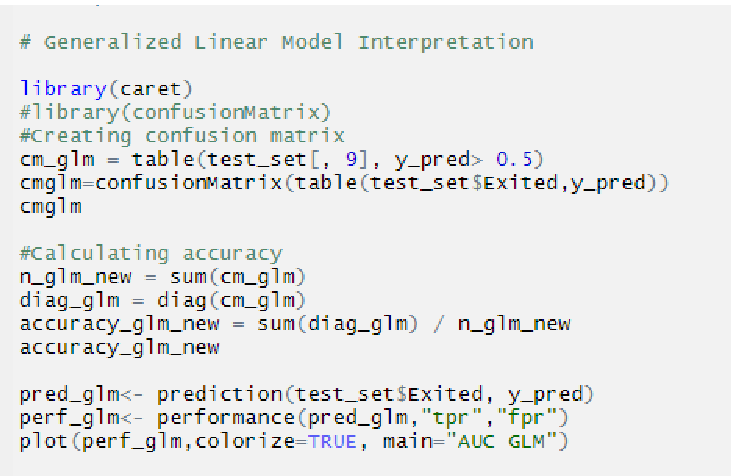


```
Confusion Matrix and Statistics

   y_pred_rf_new_k
a     0   1
  0 649  14
  1  87  83

               Accuracy : 0.8788
                 95% CI : (0.8546, 0.9001)
    No Information Rate : 0.8836
    P-Value [Acc > NIR] : 0.6903

                  Kappa : 0.5559

 Mcnemar's Test P-Value : 7.82e-13

            Sensitivity : 0.8818
            Specificity : 0.8557
         Pos Pred Value : 0.9789
         Neg Pred Value : 0.4882
             Prevalence : 0.8836
         Detection Rate : 0.7791
   Detection Prevalence : 0.7959
      Balanced Accuracy : 0.8687

       'Positive' Class : 0
```

Figure 14 Random Forest: After K fold

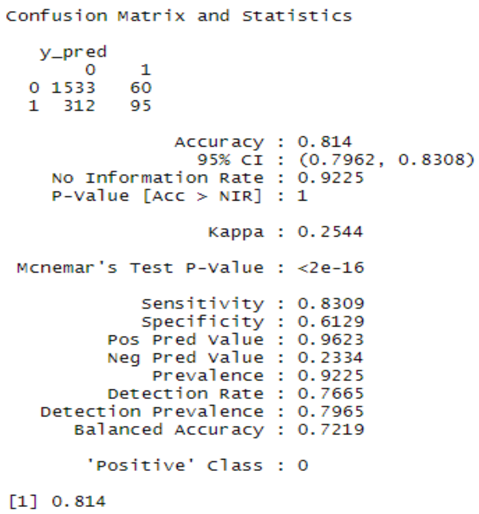
5.3 Generalized Linear Model Interpretation

The study implemented GLM specifying the family class as binomial and achieved an accuracy of 81.40% generating better accuracy but could see that the generated cohens kappa value is marginal with a value of 25.4% describing with an insight that all features were not completely considered. It has also achieved a better recall value of 83.09% and specificity of 61.2%.



```
# Generalized Linear Model Interpretation

library(caret)
#library(confusionMatrix)
#Creating confusion matrix
cm_glm = table(test_set[, 9], y_pred> 0.5)
cmglm=confusionMatrix(table(test_set$Exited,y_pred))
cmglm

#Calculating accuracy
n_glm_new = sum(cm_glm)
diag_glm = diag(cm_glm)
accuracy_glm_new = sum(diag_glm) / n_glm_new
accuracy_glm_new

pred_glm<- prediction(test_set$Exited, y_pred)
perf_glm<- performance(pred_glm,"tpr","fpr")
plot(perf_glm,colorize=TRUE, main="AUC GLM")
```

Figure 15 GLM code



```
Confusion Matrix and Statistics

   y_pred
       0    1
  0 1533   60
  1  312   95

               Accuracy : 0.814
                 95% CI : (0.7962, 0.8308)
    No Information Rate : 0.9225
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2544

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8309
            Specificity : 0.6129
         Pos Pred Value : 0.9623
         Neg Pred Value : 0.2334
             Prevalence : 0.9225
         Detection Rate : 0.7665
   Detection Prevalence : 0.7965
      Balanced Accuracy : 0.7219

       'Positive' Class : 0

[1] 0.814
```

Figure 16 GLM : Accuracy

We also checked area under the curve and achieved 72.18% with respective true positive and false positive rates and is plotted below.

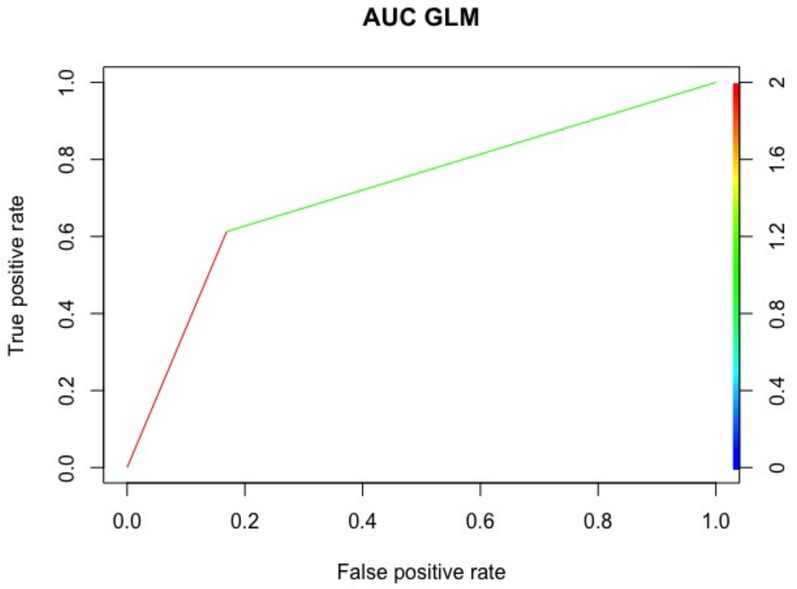
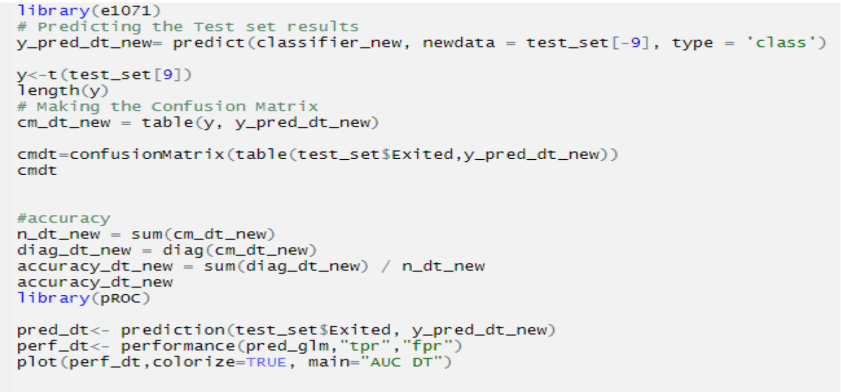


Figure 17 GLM : Graph

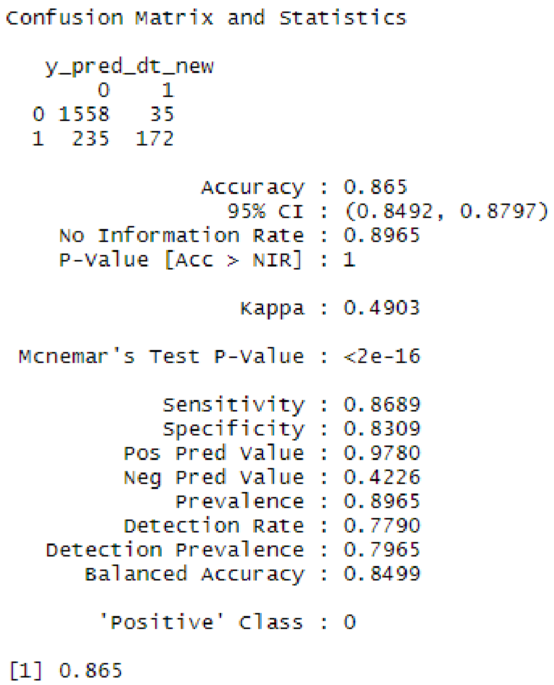
5.4 Decision Tree

Decision tree was implemented using Rpart library package and predicted an accuracy of 86.5% with respect to the confusion matrix generated using caret library function with false positives, true positives, false negatives and true negatives rates. Meanwhile, it also generated a kappa value of 49.03 which states that the model performance is efficient and have considered all the features in the dataset. It also achieved sensitivity value of 86.89% and specificity of 83.09% evaluating the models performance with an auc value of 82.47.



```
library(e1071)
# Predicting the Test set results
y_pred_dt_new= predict(classifier_new, newdata = test_set[-9], type = 'class')

y<-t(test_set[9])
length(y)
# Making the Confusion Matrix
cm_dt_new = table(y, y_pred_dt_new)

cmdt=confusionMatrix(table(test_set$Exited,y_pred_dt_new))
cmdt


#accuracy
n_dt_new = sum(cm_dt_new)
diag_dt_new = diag(cm_dt_new)
accuracy_dt_new = sum(diag_dt_new) / n_dt_new
accuracy_dt_new
library(pROC)

pred_dt<- prediction(test_set$Exited, y_pred_dt_new)
perf_dt<- performance(pred_glm,"tpr","fpr")
plot(perf_dt,colorize=TRUE, main="AUC DT")
```

Figure 18 Decision Tree: Code



```
Confusion Matrix and Statistics

   y_pred_dt_new
       0    1
  0 1558   35
  1  235  172

               Accuracy : 0.865
                 95% CI : (0.8492, 0.8797)
    No Information Rate : 0.8965
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4903

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8689
            Specificity : 0.8309
         Pos Pred Value : 0.9780
         Neg Pred Value : 0.4226
             Prevalence : 0.8965
         Detection Rate : 0.7790
   Detection Prevalence : 0.7965
      Balanced Accuracy : 0.8499

       'Positive' Class : 0

[1] 0.865
```

Figure 19 Decision Tree: Accuracy

We also plotted a decision tree in understanding how the algorithm works in partitioning the data into root and branches in obtaining the required results satisfying it in terms of binomial representation as 1 and 0.

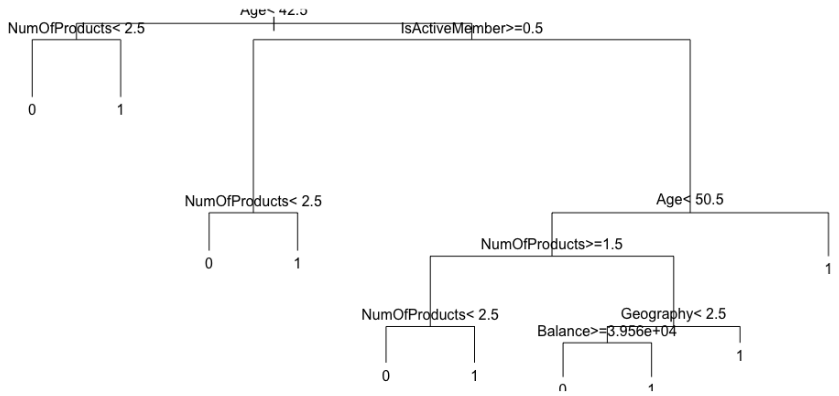


Figure 20 Decision Tree: Tree

We can see in the plot that the age is considered as the root and number of products and is an active member as its main related branches further setting up the values based on the information gain from the data and comparing it with the constraints and continues evaluating in the form of tree structure.

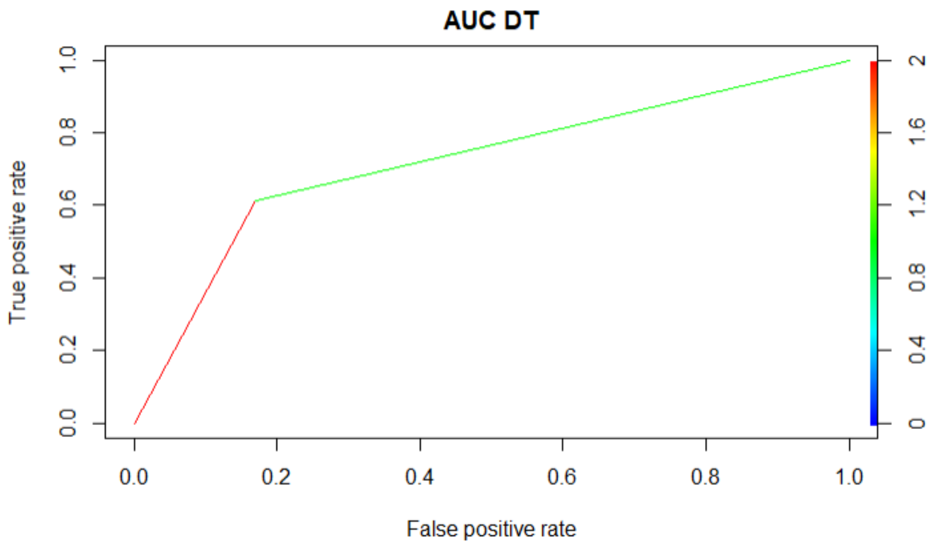
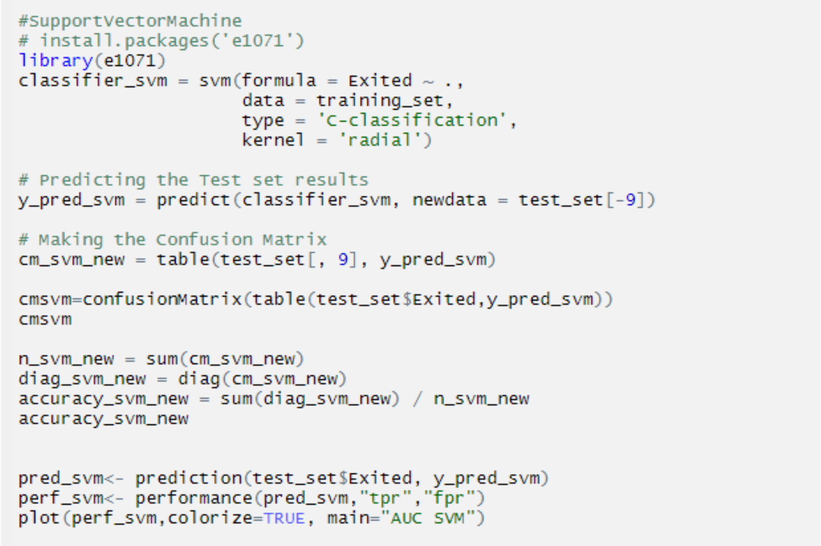


Figure 21 Decision Tree: Graph

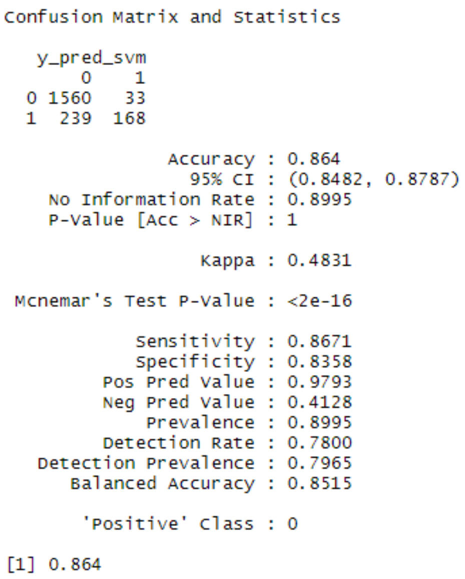
5.5 Support Vector Machine

Support Vector Machine has been implemented using e1071 library function and performed a C type classification with the kernel radial. In addition, various trial and error were preformed with different kernels such polynomial, linear, etc. to check which kernel function works the best in generating better results. Thus, SVM achieved an accuracy of 86.40% with a moderate kappa value of 48.30 along with the sensitivity and specificity values of 86.71% and 83.58% respectively. It also achieved an auc value of 83.91% with true positive and false positive rates.



```
#SupportVectorMachine
# install.packages('e1071')
library(e1071)
classifier_svm = svm(formula = Exited ~ .,
                     data = training_set,
                     type = 'C-classification',
                     kernel = 'radial')

# Predicting the Test set results
y_pred_svm = predict(classifier_svm, newdata = test_set[-9])

# Making the Confusion Matrix
cm_svm_new = table(test_set[, 9], y_pred_svm)

cmsvm=confusionMatrix(table(test_set$Exited,y_pred_svm))
cmsvm

n_svm_new = sum(cm_svm_new)
diag_svm_new = diag(cm_svm_new)
accuracy_svm_new = sum(diag_svm_new) / n_svm_new
accuracy_svm_new


pred_svm<- prediction(test_set$Exited, y_pred_svm)
perf_svm<- performance(pred_svm,"tpr","fpr")
plot(perf_svm,colorize=TRUE, main="AUC SVM")
```

Figure 22 SVM: Code



```
Confusion Matrix and Statistics

   y_pred_svm
       0    1
  0 1560   33
  1  239  168

               Accuracy : 0.864
                 95% CI : (0.8482, 0.8787)
    No Information Rate : 0.8995
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4831

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8671
            Specificity : 0.8358
         Pos Pred Value : 0.9793
         Neg Pred Value : 0.4128
             Prevalence : 0.8995
         Detection Rate : 0.7800
   Detection Prevalence : 0.7965
      Balanced Accuracy : 0.8515

       'Positive' Class : 0

[1] 0.864
```

Figure 23 SVM: Accuracy

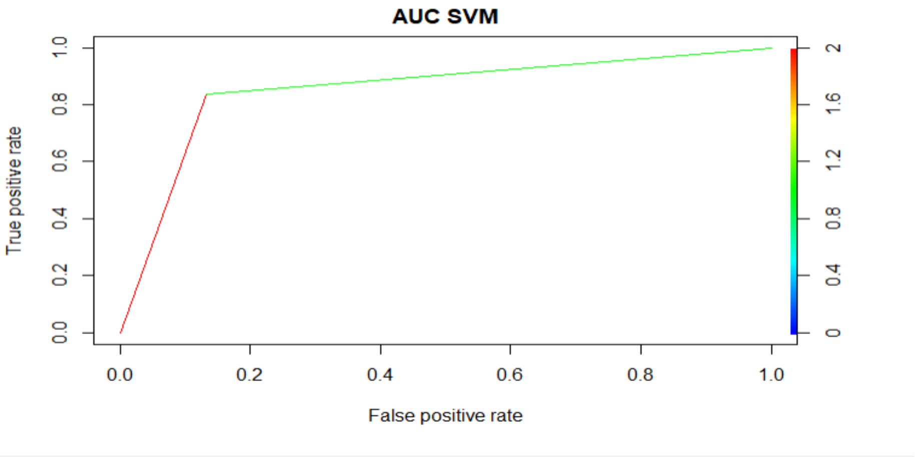
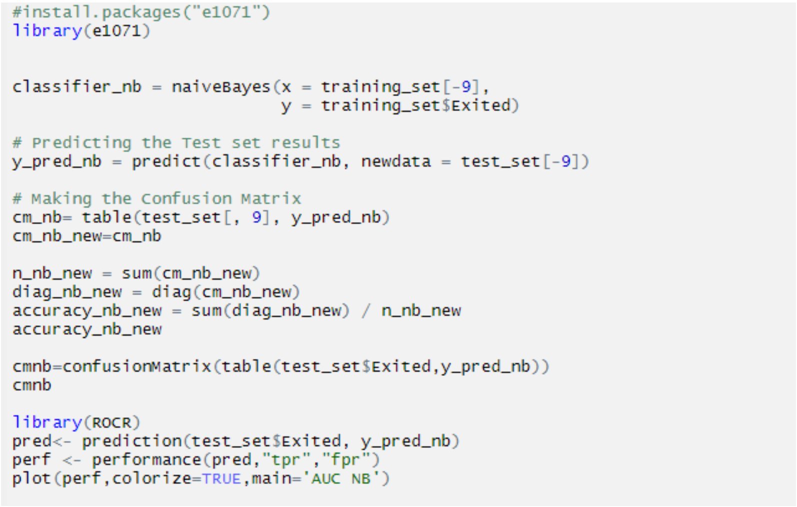


Figure 24 SVM: Graph

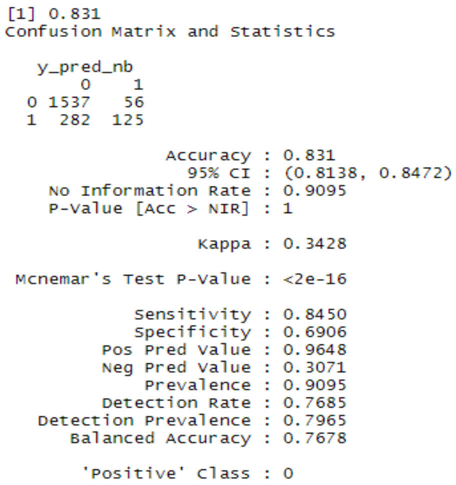
5.6 Nave Bayes Classifier

Library Function e1071 function has been used for the implementation and achieved an accuracy of 83.10% but with a marginal kappa value 0.34 and sensitivity value of 84.49% and 69.06% respectively and got an auc value of 79.01 in evaluating the model and portrayed below.



```
#install.packages("e1071")
library(e1071)

classifier_nb = naiveBayes(x = training_set[-9],
                           y = training_set$Exited)

# Predicting the Test set results
y_pred_nb = predict(classifier_nb, newdata = test_set[-9])

# Making the Confusion Matrix
cm_nb= table(test_set[, 9], y_pred_nb)
cm_nb_new=cm_nb

n_nb_new = sum(cm_nb_new)
diag_nb_new = diag(cm_nb_new)
accuracy_nb_new = sum(diag_nb_new) / n_nb_new
accuracy_nb_new

cmnb=confusionMatrix(table(test_set$Exited,y_pred_nb))
cmnb

library(ROCR)
pred<- prediction(test_set$Exited, y_pred_nb)
perf <- performance(pred,"tpr","fpr")
plot(perf,colorize=TRUE,main='AUC NB')
```

Figure 25 Naive Bayes: Code



```
[1] 0.831
Confusion Matrix and Statistics

   y_pred_nb
       0    1
  0 1537   56
  1  282  125

               Accuracy : 0.831
                 95% CI : (0.8138, 0.8472)
    No Information Rate : 0.9095
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3428

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8450
            Specificity : 0.6906
         Pos Pred Value : 0.9648
         Neg Pred Value : 0.3071
             Prevalence : 0.9095
         Detection Rate : 0.7685
   Detection Prevalence : 0.7965
      Balanced Accuracy : 0.7678

       'Positive' Class : 0
```

Figure 26 Naive Bayes: Accuracy

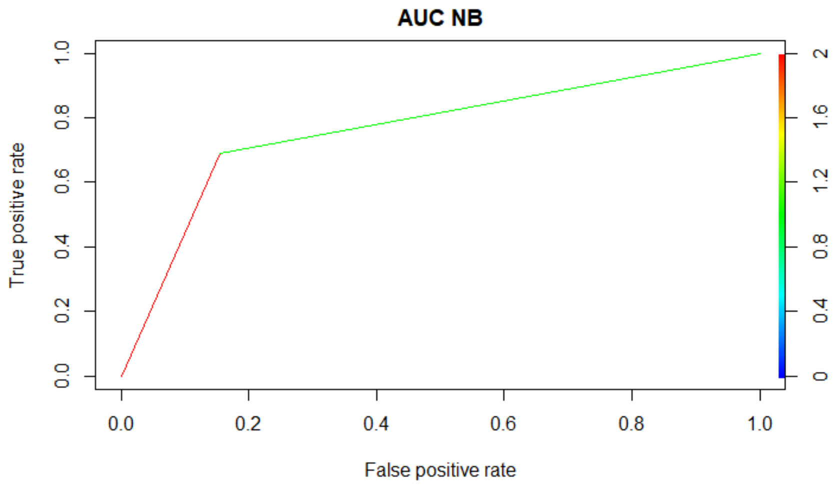


Figure 27 Naive Bayes: Graph

The below table is the summary of the results achieved by each algorithm making it easier in understanding and comparing the results the opt the best model.

ALGORITHMS	ACCURACY	PRECISION	RECALL
RANDOM FOREST	87.05 %	81.09 %	87.85%
LINEAR MODEL	81.04%	61.29%	83.09%
DECISION TREE	86.5%	83.09%	86.89%
NAIVE BAYES	83.01%	69.06%	84.50%
SVM-SUPPORT VECTOR MACHINE	86.04%	83.58%	86.71%

Table 2 Accuracy, Precision and Recall of Algorithms

As per the above findings of each of the algorithm accuracy and metrics, it varies along with as we change the set seed value. Below table is the output this was conducted to examine how the model performance is as a part of different test cases.

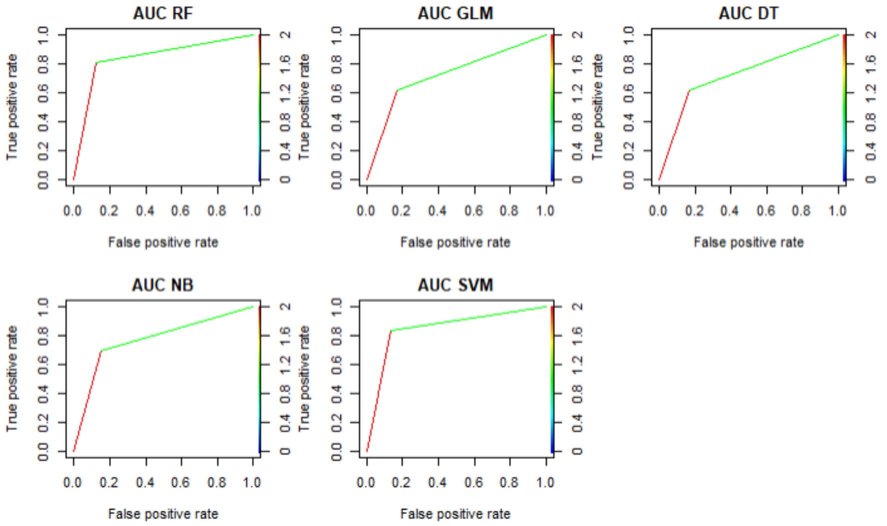
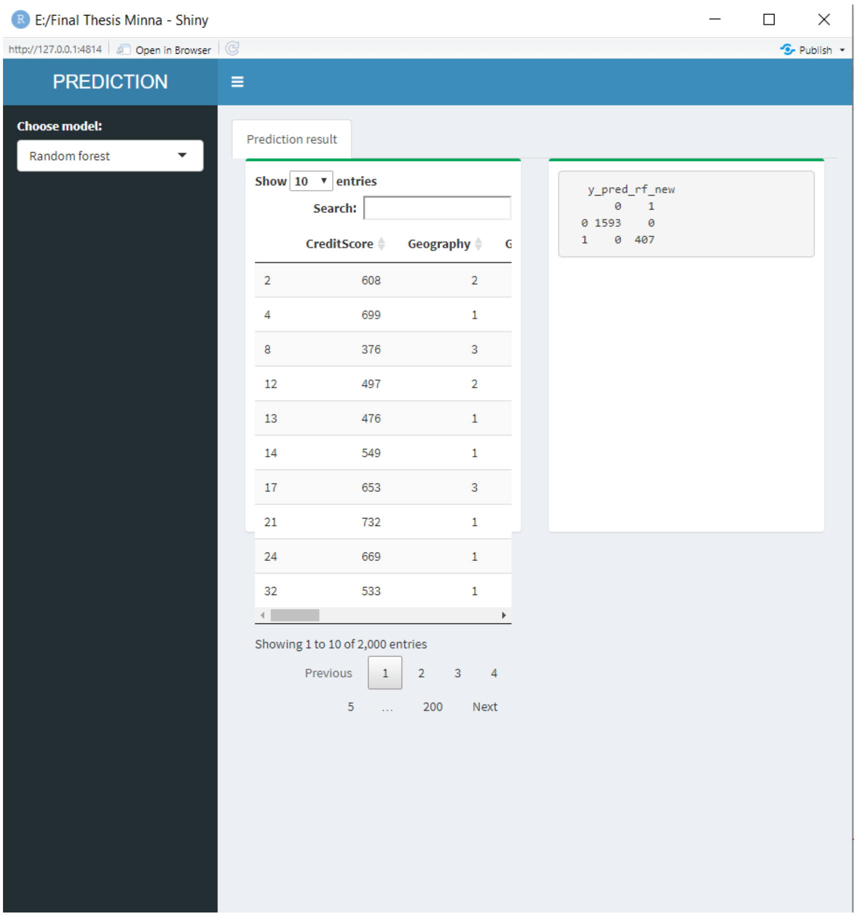


Figure 28 Graphs of all algorithms

Chapter 6. Testing and Evaluation

6.1 UI Testing

User interface is developed using shiny app using r programming language.



Choose model: Random forest

Show 10 entries

Search:

	CreditScore	Geography	G
2	608	2	
4	699	1	
8	376	3	
12	497	2	
13	476	1	
14	549	1	
17	653	3	
21	732	1	
24	669	1	
32	533	1	

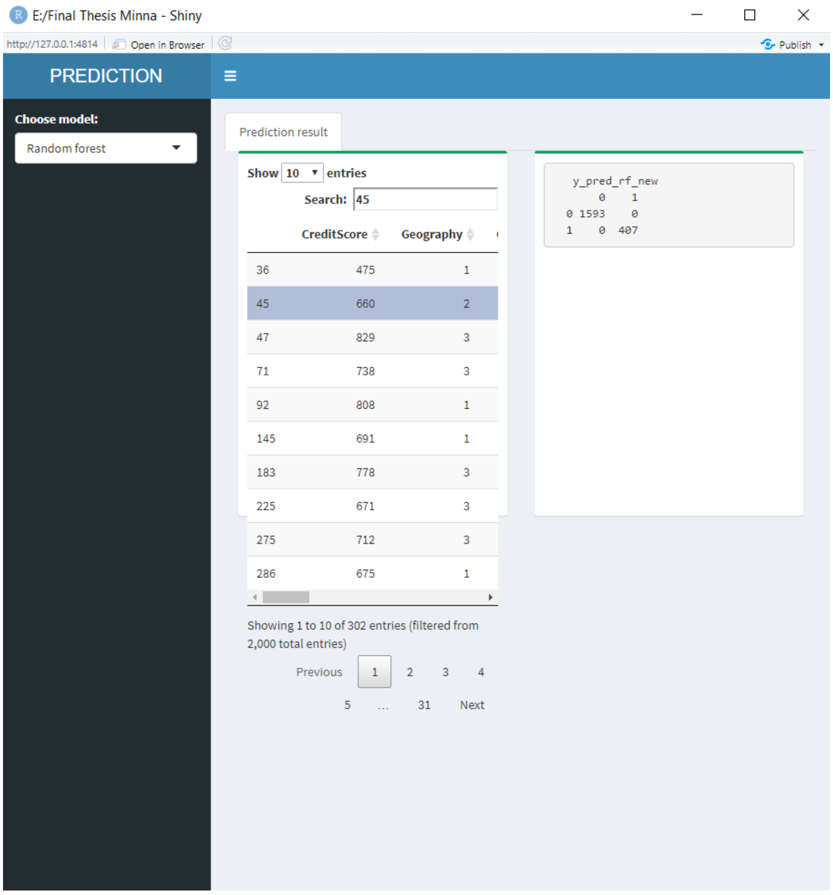
Showing 1 to 10 of 2,000 entries

Previous 1 2 3 4 5 … 200 Next

```
  y_pred_rf_new
       0    1
  0 1593    0
  1    0  407
```

Figure 29 UI Interface in Shiny web app

The prediction result of the bank churn modelling is showed using shiny web application. Random Forest has highest accuracy, results predicted using this algorithm.



Choose model: Random forest

Show 10 entries

Search: 45

	CreditScore	Geography
36	475	1
45	660	2
47	829	3
71	738	3
92	808	1
145	691	1
183	778	3
225	671	3
275	712	3
286	675	1

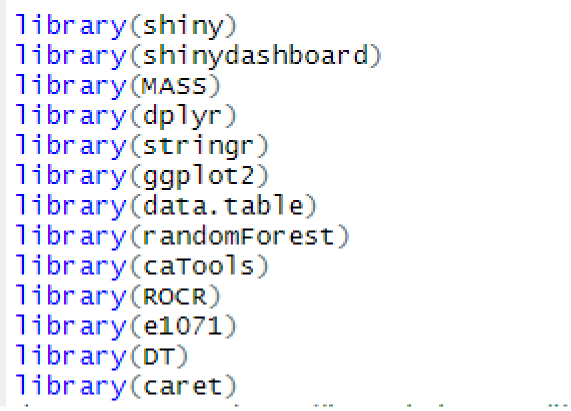
Showing 1 to 10 of 302 entries (filtered from 2,000 total entries)

Previous 1 2 3 4 5 … 31 Next

```
  y_pred_rf_new
      0    1
  0 1593    0
  1    0  407
```

Figure 30 Showing Prediction in shiny

If searching a particular customer using number or any field, particular entry will be displayed from the test dataset.



```
library(shiny)
library(shinydashboard)
library(MASS)
library(dplyr)
library(stringr)
library(ggplot2)
library(data.table)
library(randomForest)
library(caTools)
library(ROCR)
library(e1071)
library(DT)
library(caret)
```

Figure 31 Loading Library

When we type particular number to search then the bank end starts working to fetch the row which contains that number in any of the columns.

Chapter 7. Conclusions and Future Work

7.1 Conclusion

Any day, finding churn rate modelling with the banking industry always presents itself as a difficult task, which can be passed on to the client due to a lot of uncertainties and due to its demand in the market, technology at its peak. Other related findings and insights gained during this research were that the modelling churn rate in banking is still the most difficult to find compared to other sectors of the industry such as telecom, insurance. Retail, etc. Take the example of the telecommunications industry, the customer will have many reasons to leave services with a service provider (e.g. network coverage). So, this could be a reason why the customer said goodbye, but despite knowing the reason for the business, but they still can't retain the customer because coverage is a kind of service that can't. Fix overnight, but only if the cause was different. Then the customer may be more likely to respond. Therefore, in any industry, there are always gaps for customers who voluntarily leave the company. In other words, a company has developed sophisticated CRM systems and data mining tools to identify who will or will be retired and their needs. Understand that it is very costly to attract new customers, rather than retain old ones, and understand that you withhold them. With respect to banking, based on available data, the approach will be modified to make effective forecasts to reduce potential churn. As a further result, churn modelling could be predicted using other approaches based on data availability, such as using clustering techniques. That means clustering by splitting the data into smaller test cases, clustering K-to identify patterns of behaviour between running clients, and analyzing the likelihood that he/she will leave the company.

Thus, we can conclude that the analysis conducted to identify customer behaviour and patterns presents the model as it takes into account the relevant factors taking into account the forecasts, potential strategies, recommendations and business needs to help develop the recognition, as well as effective And effective implementation of goals. Thus, in the selected classifier models in terms of improving the accuracy of 87.5% and increasing the accuracy using optimization methods that reached 78%, random forests outperformed other algorithms, but each in data prediction and work. The model has its own meaning. So it is worth doing research. As such, we must point out that effective marketing strategies and CRM systems are a success factor for any relevant industry in the fight against consumer attitudes.

7.2 Future Work

Within the project some future work that can be carried out is identified. In this study, only one branch of the data set was explored and analyzed. Another branch of the bank dataset can be explored in the future. In this project, four machine learning methods were used for the bank dataset. You can explore other methods as well. You can explore various machine learning algorithms and analyze data. More research can be done to build a time series model to predict customer churn. Additionally, you can use unsupervised clustering machine learning techniques to explore data. With this method, you can determine the similarity of data or some patterns.

References

[1]G. XIA, "Customer churn prediction on kernel principal component analysis feature abstraction", *Journal of Computer Applications*, vol. 28, no. 1, pp. 149-151, 2008. Available: 10.3724/sp.j.1087.2008.00155.

[2]A. Amin, F. Rahim, M. Ramzan and S. Anwar, "A Prudent Based Approach for Customer Churn Prediction", *Beyond Databases, Architectures and Structures*, pp. 320-332, 2015. Available: 10.1007/978-3-319-18422-7_29 [Accessed 28 August 2020].

[3]M. Ballings and D. Van den Poel, "Customer event history for churn prediction: How long is long enough?", *Expert Systems with Applications*, vol. 39, no. 18, pp. 13517-13522, 2012. Available: 10.1016/j.eswa.2012.07.006.

[4]M. Fridrich, "Experimental Parameter Tuning of Artificial Neural Network in Customer Churn Prediction", *Trends Economics and Management*, vol. 11, no. 28, p. 9, 2017. Available: 10.13164/trends.2017.28.9.

[5]Y. Zhao, B. Li, X. Li, W. Liu and S. Ren, "Customer Churn Prediction Using Improved One-Class Support Vector Machine", *Advanced Data Mining and Applications*, pp. 300-306, 2005. Available: 10.1007/11527503_36 [Accessed 1 September 2020].

[6]K. Coussement, "Research – Kristof Coussement", *Kristofcoussement.com*, 2020. [Online]. Available: http://www.kristofcoussement.com/research/. [Accessed: 01-Sep- 2020].

[7]X. Wang, K. Nguyen and B. Nguyen, "Churn Prediction using Ensemble Learning", *Proceedings of the 4th International Conference on Machine Learning and Soft Computing*, 2020. Available: 10.1145/3380688.3380710 [Accessed 1 September 2020].

[8]A. Ahmad, A. Jafar and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform", *Journal of Big Data*, vol. 6, no. 1, 2019. Available: 10.1186/s40537-019-0191-6 [Accessed 2 September 2020].

[9]A. Bilal Zoric, "Predicting Customer Churn in Banking Industry using Neural Networks", *Interdisciplinary Description of Complex Systems*, vol. 14, no. 2, pp. 116-124, 2016. Available: 10.7906/indecs.14.2.1.

[10]Ö. Gür Ali and U. Arıtürk, "Dynamic churn prediction framework with more effective use of rare event data: The case of private banking", *Expert Systems with*

Applications, vol. 41, no. 17, pp. 7889-7903, 2014. Available: 10.1016/j.eswa.2014.06.018.

[11]"Bank Customer Churn Prediction Based on Support Vector Machine: Taking a Commercial Bank's VIP Customer Churn as the Example - IEEE Conference Publication", *Ieeexplore.ieee.org*, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/4680698. [Accessed: 02- Sep- 2020].

[12]"http://learn.aible.com/ai_for_customer_retention/", *Learn.aible.com*, 2020. [Online]. Available: https://learn.aible.com/ai_for_customer_retention/?utm_term=customer%20retention&utm_campaign=AI+for+Customer+Retention+-+US&utm_source=adwords&utm_medium=ppc&hsa_acc=1900714811&hsa_cam=10829873256&hsa_grp=106240494229&hsa_ad=456511443594&hsa_src=g&hsa_tgt=kwd-10359666&hsa_kw=customer%20retention&hsa_mt=b&hsa_net=adwords&hsa_ver=3&gclid=Cj0KCQjwy8f6BRC7ARIsAPIXOjiSRbzgxHcQQS7ePTlBjRafztuTD4nNNlM_gBLopId2hfK895OBspkaAk0mEALw_wcB. [Accessed: 02- Sep- 2020].

[13]A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo and S. Anwar, "Customer churn prediction in telecommunication industry using data certainty", *Journal of Business Research*, vol. 94, pp. 290-301, 2019. Available: 10.1016/j.jbusres.2018.03.003 [Accessed 2 September 2020].

[14]V. Subramanian, M. Hung and M. Hu, "An experimental evaluation of neural networks for classification", *Computers & Operations Research*, vol. 20, no. 7, pp. 769-782, 1993. Available: 10.1016/0305-0548(93)90063-o [Accessed 2 September 2020].

[15]L. Bin, S. Peiji and L. Juan, "Customer Churn Prediction Based on the Decision Tree in Personal Handyphone System Service", *2007 International Conference on Service Systems and Service Management*, 2007. Available: 10.1109/icsssm.2007.4280145 [Accessed 2 September 2020].

[16]2020. [Online]. Available: https://www.researchgate.net/publication/332528099_Early_Prediction_of_Employee_Attrition_using_Data_Mining_Techniques. [Accessed: 04- Sep- 2020].

[17]B. He, Y. Shi, Q. Wan and X. Zhao, "Prediction of Customer Attrition of Commercial Banks based on SVM Model", *Procedia Computer Science*, vol. 31, pp. 423-430, 2014. Available: 10.1016/j.procs.2014.05.286.

[18]A. Keramati, H. Ghaneei and S. Mirmohammadi, "Developing a prediction model for customer churn from electronic banking services using data mining", 2020.

[19]L. Khaidem, S. Saha and S. Dey, "Predicting the direction of stock market prices using random forest", *arXiv.org*, 2020. [Online]. Available: https://arxiv.org/abs/1605.00003. [Accessed: 02- Sep- 2020].

[20]P. Kisioglu and Y. Topcu, "Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey", 2020. .

[21]H. Academy and I. Tree, "Create your own social network with the best community website builder - NING", *NING*, 2020. [Online]. Available: http://api.ning.com/files/sb0MFra59lH4G*UJbMUFf3deK5IBa7EVN9v9glAVh4njDVmgyYJsNcXmFhZuufX1*N1fxlqkS*d4td*hp4W5*Yxa8MOy3IR2/Predictingcreditcardcustomerchurninbanks.pdf. [Accessed: 03- Sep- 2020].

[22]"Data mining | Guide books", *Dl.acm.org*, 2020. [Online]. Available: https://dl.acm.org/doi/book/10.5555/323651. [Accessed: 04- Sep- 2020].

[23]2020. [Online]. Available: https://www.researchgate.net/publication/308399969_PREDICT_CUSTOMER_CHURN_BY_USING_ROUGH_SET_THEORY_AND_NEURAL_NETWORK. [Accessed: 04- Sep- 2020].

[24]L. McDonald and S. Rundle-Thiele, "Corporate social responsibility and bank customer satisfaction", *International Journal of Bank Marketing*, vol. 26, no. 3, pp. 170-182, 2008. Available: 10.1108/02652320810864643.

[25]"Bayes and Empirical Bayes Methods for Data Analysis", *Technometrics*, vol. 43, no. 2, pp. 246-246, 2001. Available: 10.1198/tech.2001.s608.

[26]V. Saradhi and G. Palshikar, "Employee churn prediction", *Expert Systems with Applications*, vol. 38, no. 3, pp. 1999-2006, 2011. Available: 10.1016/j.eswa.2010.07.134.

[27]H. Sayed, M. A. and S. Kholief, "Predicting Potential Banking Customer Churn using Apache Spark ML and MLlib Packages: A Comparative Study", *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 11, 2018. Available: 10.14569/ijacsa.2018.091196.

[28]T. Vafeiadis, K. Diamantaras, G. Sarigiannidis and K. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction", *Simulation Modelling Practice and Theory*, vol. 55, pp. 1-9, 2015. Available: 10.1016/j.simpat.2015.03.003.

[29]D. Van den Poel and B. Larivière, "Customer attrition analysis for financial services using proportional hazard models", 2020. .

[30]2020. [Online]. Available: https://www.researchgate.net/publication/222929949_Customer_churn_prediction_using_improved_balanced_random_forests. [Accessed: 04- Sep- 2020].

[31]P. Dalvi, S. Khandge, A. Deomore, A. Bankar and V. Kanade, "Analysis of customer churn prediction in telecom industry using decision trees and logistic regression", *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, 2016. Available: 10.1109/cdan.2016.7570883 [Accessed 4 September 2020].

[32]B. Huang, T. Kechadi, B. Buckley, G. Kiernan, E. Keogh and T. Rashid, "A new feature set with new window techniques for customer churn prediction in land-line telecommunications", *Expert Systems with Applications*, vol. 37, no. 5, pp. 3657-3665, 2010. Available: 10.1016/j.eswa.2009.10.025 [Accessed 4 September 2020].