Open University of Cyprus

School of Economics and Management

Master's Degree: *Enterprise Risk Management*

Master Thesis

"Big Data Analytics in Banks: Comparison of Classification Models in predicting customers churn"

Chara A. Gavrielidou

Supervisor

Dr Pandelis Ipsilandis

May 2021

Open University of Cyprus

School of Economics and Management

Master's Degree: *Enterprise Risk Management*

Master Thesis

"Big Data Analytics in Banks: Comparison of Classification Models in predicting customers churn"

Chara A. Gavrielidou

Supervisor

Dr Pandelis Ipsilandis

"The present postgraduate Dissertation was submitted in partial fulfillment of the requirements for obtaining a Postgraduate Degree on Enterprise Risk Management from the School of Economics and Management of the Open University of Cyprus.

May 2021

ACKNOWLEDGEMENTS

I would first like to express my sincere thanks to my supervisor Dr Pandelis Ipsilandis, for providing me with his immense knowledge, continuous support, guidance and advice throughout the thesis. His kind guidance encouraged me a lot to do better in my thesis writing.

Finally, I would like to thank my family for their support and motivation during my studies. This accomplishment would not have been possible without them.

Table of Contents

Table of Figures

CHAPTER 1

Introduction

Competition in banking industry is growing and banks are trying to increase their market share by acquiring new customers and at the same time to use effective customer retention strategies. As a result, by improving the retention rate by up to 5 %, can increase a bank's profit up to 85 % (Nie et al., 2011).

Additionally, it costs more to any company attracting new customers rather than retaining the old ones who are more likely to produce profit for the company (Verbeke et al., 2011).

This process of movement from one bank to another is usually happening due to better services or due to different benefits that each competitor bank offers to any potential customer.

In order to maintain their competitive advantage and find valuable information from data, many banks are using predictive models and data analytical techniques to predict customer churn. Zoric, B et al. 2016).

Customer churn prediction is studied very commonly across different industries such as financial services, electronic commerce, telecommunications, retail markets and subscription management (Chen, Fan and Sun, 2012).

Many companies are taking different measures for improving intervention strategies and convince these customers to stay and to prevent the loss of their businesses (Zorn et al., 2010).

Future Predictions can be very valuable since they allow companies to adjust to the possible outcome (Roos and Gustafsson, 2007).

Customer churn prediction has become a field with much research and different models where used through the years. From our literature review models used before 2011 were Logistic Regression, Decision trees as well modern methods such as Artificial neural networks (ANN), Support Vector Machines (SVM) and Random Forests (RF). In 2015, Logistic Regression, Artificial Neural Networks and Decision Trees were on top of most used models for the telecommunication industry. (Mahajan, V., Misra, R. and Mahajan, R. 2015).

This dissertation has two parts:

The first part includes the description and analysis of Big data Analytics, its concepts, methods and analytical tools as well an extended analysis for Predictive Analytics, Classification Problem and literature review of classification Models predicting customer churn.

Second part includes the creation of two models which predict customer churn with classification especially Logistic Regression and Radial Basic Function and compare their performance.

The overall accuracy of the model, Receiver Operating Characteristic curve and Area Under the Receiver Operating Characteristic Curve is used as the evaluation metrics for this research to identify the best classifier.

CHAPTER 2

Big Data

2.1. History of big data

Big data origins go back in the 60s and 70s when first data centers were developed and the word of data has just started and relational databases had begun.

With the development of online services such as Facebook and You Tube around 2005, people started to understand how much data was generated. In the same year, Roger Mougalas from O'Reilly Media used the term Big Data for the first time:

It refers to a large set of data that is almost impossible to manage and process using traditional business intelligence tools.

Through the next years, new open-source frameworks were created such as Hadoop and Spank, in order to analyze and store big data sets.

2.2. Big Data Definition

Big data was also introduced by Gartner Analysts as:

Big data *is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.*

2.3 5Vs of Big Data

Big Data qualities known as the 5Vs are: Volume, Variety, Velocity, Veracity, and Value.

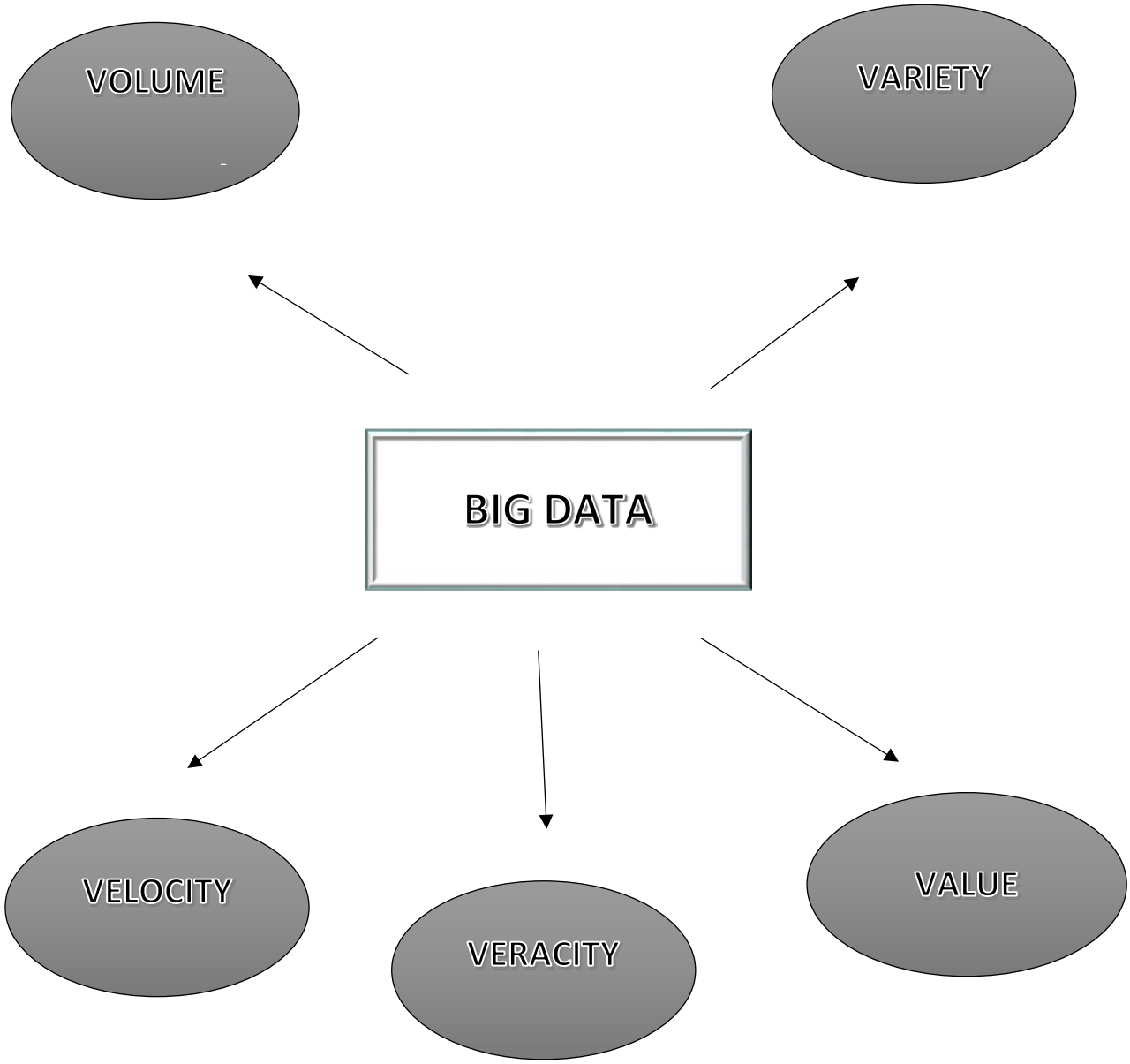


Figure 1: Author's illustration of 5Vs

Volume

One of the qualities of big data is volume which refers to the amount of data. For sure with big data, you have to process high volumes of data which is unstructured and of low-density and that matters. There is data which has unknown value such as clickstreams on a mobile application or Twitter/Facebook data feeds or on any webpage. This might be hundreds of terabytes for some organizations as for others this might be tens of petabytes.

Velocity

This quality refers to the rate at which data is received and how fast is generated. Companies need to have access to the data to the right time in order to make best managerial decisions. So, Velocity is important for the organizations because it gives competitive advantage if data is available as close to real time to be evaluated and take the best action.

Variety

This quality of big data refers to all types of big data which are available from many different sources. For example, from smartphone technology, from social networks or form in-house devices. These types of data can be structured as traditional data types who can fit neatly in a relational database to semi structured and to new unstructured types such videos, audio and text which require more complicated processing to be evaluated.

Veracity

This quality of big data refers to how accurate a data set can be and it gives a deeper understanding of data and this depends on if the data type source and processing is trustworthy. Veracity helps to understand and filter what is important or not in a timely manner. Data Veracity ensures through the processing method makes sense and ensures that the output is relevant and focused to the business needs in order to take action.

Value

This quality of big data refers to the ability to transform the data into business. This requires business professionals and analysts who can use the right tools, make informed assumptions and transform data to business decisions.

2.4. Benefits and used cases of big data and data analytics

Big data has changed the way that financial institutions are using data to predict their futures steps.

Big data gives the opportunity to gain complete answers because of the available information given. This gives more confidence and a completely different approach to deal with problems.

Big data can help financial institutions to address a lot of business activities from customer experience, product development to fraud detection and faster compliance reporting.

Customer experience

More than ever before, though big data and data analytics, a better picture of the consumer experience is now possible. In order to optimize the interaction experience and enhance the value delivered, big data helps you to collect data from social media, site visits, call logs, and other sources. Start to deliver customized deals, reduce consumer satisfaction, and proactively manage problems.

Netflix and Procter & Gamble are using big data to anticipate customer demand with predictive models and foe the creation of new products and services. These predictive models help them to plan produce and launch new products using big data from social media, focus groups and early store rollouts.

Operational efficiency

Big data has the most impact on Operational efficiency.

With big data, you can analyze and assess production, customer feedback and returns, and other factors to reduce outages and anticipate future demands. Big data can also be used to improve decision-making in line with current market demand.

Drive innovation

Big data will help organizations to innovate.

By researching interdependencies between individuals, organizations, agencies, and pro cesses and then identifying new ways to use those insights.

To enhance decisions on financial and planning considerations, use data insights.

Examine patterns and what new goods and services consumers want to offer. Enforce dy namic pricing. Big data gives you new insights that open up new opportunities and business models.

2.5. Big Data Analytical Types and Tools

2.5.1. Types of Big Data Analytics

1. Descriptive Analytics:

In order to gain guidance into the past and reply to the question: "What has happened?", it uses data aggregation and data mining.

Descriptive analytics do just as the name means, "describe" or summarize raw data and make it human-interpretable. What is happening now based on incoming data.

One good example of Descriptive analytics is the Google Analytics Tool. Through the tool will help the business to get results from the web server and understand what happened in the past.

2. Predictive analytics:

Predictive Analytics uses statistical models and methods of forecasting such as data mining, predictive modelling, and machine learning to understand the future and to answer: "What could happen?

Predictive Analytics provides organizations with measurable data-based insights. This introduces predictions about the possibility of a potential result. The aim is to move beyond understanding what has happened in order to have a better evaluation of what might happen in the future.

Predictive analytics is used in different fields such as marketing, business management, actuarial science, insurance, travel, healthcare pharmaceuticals and other fields.

Credit scoring, which is used in business management, is one of the best-known predictive applications. In order to rate customers by their possibility of making potential credit payments on time, scoring models process a customer's credit history, loan application, customer data.

3. Prescriptive Analytics:

"It utilizes algorithms for optimization and simulation to decide on possible results and responses on the question: "What should we do? It enables users to "prescribe" and direct them towards a response to a variety of different potential acts. This analytics is all about offering guidance and advice what action should be taken.

Prescriptive analytics software can help with both locating and producing hydrocarbons] by taking in seismic data, well log data, production data, and other related data sets to prescribe specific recipes for how and where to drill, complete, and produce wells in order to minimize cost, reduce environmental footprint and optimize recovery.

4. Diagnostic Analytics:

Diagnostic analytics is used to determine why something happened in the past. It is characterized by techniques such as drill-down, data discovery, data mining and correlations. Diagnostic analytics takes a deeper look at data to understand the root causes of the events.

A great example of Diagnostic analytics is Social Media marketing campaign which uses Diagnostic analytics to assess the number of followers, fans, page views, reviews, posts and analyses the success or failure rate of the campaign.

2.5.2 Basic Big Data Tools

The basic Big Data Tools that are used for Big Data Analytics are the following: Apache Pig, Hadoop, Apache HBase, Talend ,Splunk ,Apache Spark, Apache Hive and Kafka.

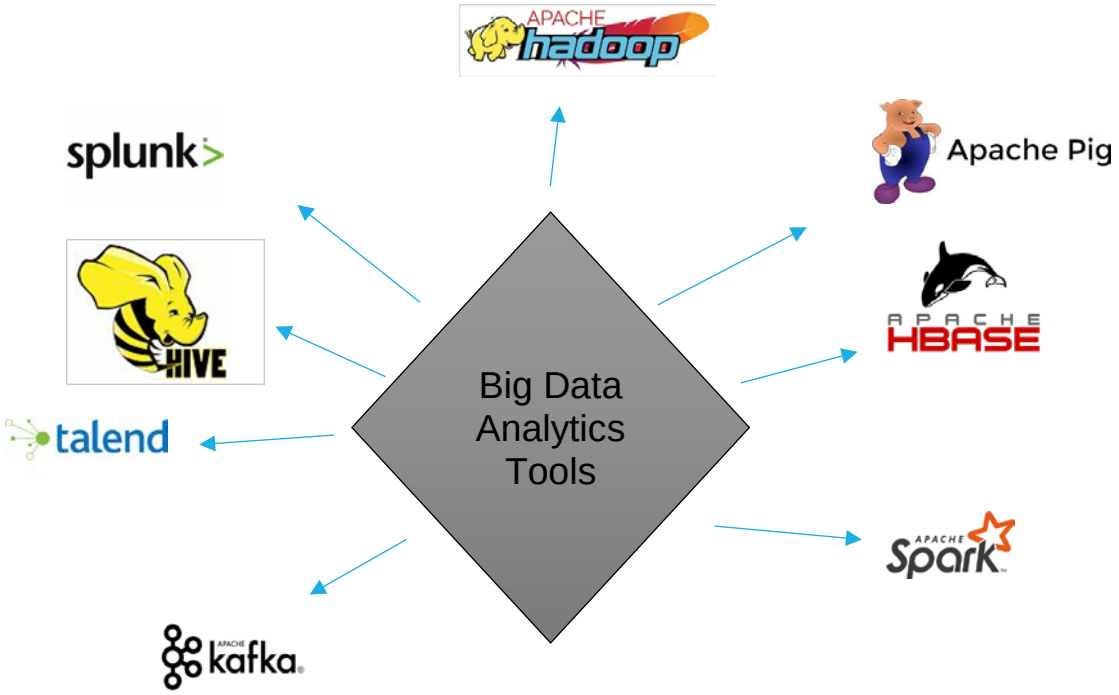


Figure 2: Authors Illustration of Big Data Tools

2.6 Big data and data analytics challenges

Big Data brings a lot of opportunities to financial institutions but one the other hand has to deal with a huge number of challenges. Although a lot of new technologies have been developed for data storage the size of data volume is doubled every two years.
For various researchers these challenges can be identified as management and processing issues storage and accuracy issues, data accessing, hardware or technical related challenges, privacy and security challenges and other analytical challenges.

Organizations struggle to keep pace with their data and find ways to effectively store it.
On the basis of literature review we have addressed the main challenges that need immediate attention to be evaluated and successfully eliminated if possible or at least to be minimized to the best outcome.

Lack of Skill Requirements:

The most recent Big Data Analytical Tools such as Hadoop, Apache Spark, Apache Pig, use complex technologies for the processing of big data. This requires professional with exceptional skills to handle them. According to Mckinesey & Company special report, US needs more than 150,000 data analysts or high skilled professionals for data analysis and more than one million managers to take result-based decisions.
Depending upon the role in the field of Big Data Analytics some of the skills which are required are the following:

I. **Basic programming:** Knowledge of at least some general programming languages such as Java and Python.
II. **Statistical and Quantitative Analysis:** It is important to know statistics and quantitative analysis.
III. **Data Warehousing:** Knowledge of SQL and NoSQL databases is needed.
IV. **Data Visualization:** It is essential to know how to visualize the data in order to be able to understand the insights and apply it in action.

V. **Specific Business Knowledge:** One must necessarily be aware of the business where they are applying analytics in order to optimize their operations.

VI. **Computational Frameworks:** Preferably one should know about at least one or two tools which are required for Big Data Analytics.

Management

Data distribution, access systems and governance are management challenges. The facts that new technologies are developing, companies are uncertain which technique is best to be used without any emerging risks or any potential problems.

Security and Privacy of Data

When business businesses figure out how to use Big Data, a wide variety of possibilities and opportunities are presented to them. It also includes, however, the possible threats associated with big data in terms of privacy and data protection. The Big Data program used for research and storage uses various sources of data. Eventually, this leads to a high risk of data exposure of the data, making it making it fragile. Thus, the huge amounts of data generate privacy and security issues. A Big Data corporate training program should be coordinated by company owners and administrators to solve these Big Data issues in enterprises and large organizations.

Synchronization between Sources of Data

Data loading involves bringing data into a single data repository from various heterogeneous data sources. For example, multiple data sources should be mapped into a single structural system, instruments and facilities should be available that cooperate with the size and speed of big data. The loading method suffers from numerous problems that involve the keen attention of researchers and practitioners and should transfer the data in a timely manner. Synchronization through various data sources is often considered as one of the crucial challenges, along with these loading problems.

Once the data is loaded from various data sources into a big data network, it is likely to get out of synchronization at different time intervals at different speeds. Data source synchronization refers to the process of ensuring data continuity over time between various sources of data and a shared repository. In other words, in terms of time and sequence, the data coming from different sources should balance each other. It is likely to become inaccurate or even invalid if the big data processing method seems unable to ensure synchronization.

This might lead to unfavorable and/or wrong outcome of mining. Thus, in order to help business organizations, avoid uncertainties in the analysis process and therefore draw reliable and reasonable conclusions, considerable attention should be given to the synchronization of data sources. The heterogeneous nature of data also makes it more difficult for companies to transform and clean before loading them for review in the warehouse. Hadoop and MapReduce are good examples to process unstructured data and effectively are used by several businesses.

Visualization

In order to enable more effective decision making, data visualization is the method of representing information in a standardized manner. As big data exponentially expands with unfettered velocity and enormous volume, due to the lack of scalable visualization software, it becomes very hard to extract the secret information. There is no denying that web markets to convert their huge, complex data sets into image formats, web markets use big data visualization tools such as Tableau and other tools to make all the data easily comprehensible. However, it is possible that these current visualization techniques will be of no use in the immediate future.

To address the current challenges and to be prepared to deal with future challenges, both in terms of hardware and software, a comprehensive effort by researchers is required. This can be done by creating knowledgeable processes that take advantage of great technology in data integration that addresses the changing threats and the reduction of the associated challenges and risks.

2.7 How Big Data Can Help you predict potential customer churn

The growth in volume, variety and velocity of data generated about the customers and their interactions across multiple channels has made it almost impossible to store, analyze and retrieve meaningful insights using traditional data management technologies.

Big Data can help you solve these challenges and allows you to leverage both structured and unstructured data from multiple channels such as bank visits, customer call logs, web interactions, transactional data such as credit card histories, and social media interactions.

The challenges that banks are facing can be solved by storing, analyzing and retrieving the massive volume and variety of structured and unstructured data economically on commodity hardware and scale elastically as the data grows.

It also allows banks to tap in to real-time customer interactions that are more likely to provide early warning signs before it is too late.

Additionally, sophisticated data matching capabilities allows banks to build a comprehensive holistic customer profile.

CHAPTER 3

Predictive Analytics

3.1. Definition of Predictive analytics

Predictive Analytics is the advance Analytics that is used in order to make prediction about unknown future outcomes. Predictive Analytics uses historical data combined with Data mining techniques, Big Data Machine Learning and Statistical Modeling in order to find patterns and identify potential opportunities and rising risks.

Predictive Analytics allows business to use statistics and modeling techniques to determine future performance and also is been used from business as a decision-making tool.

Predictive Analytics is used to boost revenue, mitigate risks and streamline operation for many kinds of businesses, including Retail, Public Sector, Manufacturing, Banking, Utilities, and Healthcare Sector.

3.2. Importance of Predictive analytics

Predictive Analytics is very important in our days because it helps organizations to solve difficult problems, minimize their risks and uncover new opportunities. Predictive Analytics helps organizations to the following:

1. Detection of fraud

Predictive Analytics combines multiple analytics methods to prevent criminal behavior and improve pattern detection. Through high-performance behavioral analytics and cybersecurity, all actions are examined on a network in real time to identify abnormalities that may indicate fraud and monitor advanced persistent threats.

2. Improve operations of organizations

Predictive models are used in many fields to manage resources, forecast inventory and maximize their market share. Predictive Analytics can help the companies interpret big data for their benefit and function more efficiently.
In order to increase their revenue and maximize occupancy, hotels use predictive models to predict the number of hotel quests for any given night. Airlines through predictive models, set ticket prices, predict the impact of specific maintenance operations on aircraft reliability, uptime, fuel use and availability. In manufacturing industry, predictive models are used to predict the location and rate of machine failures as well as in automotive industry by studying driver behavior to develop better driver assistance technologies.

3. Risk Reduction

Predictive models are used from many banks, insurance companies and other financial institutions, to assess their customers and predict credit score.
Predictive Analytics through predictive models can assess risks with a specific set of conditions and can capture relationships between different factors. One of the most used examples of predictive model is credit scoring. Credit scoring is a number which is generated by applying all customer data in the predictive models to assess the customer's creditworthiness.

4. Optimization of marketing campaigns

Predictive analytics are used by marketing campaigns to promote cross-selling opportunities and predict customer responses or purchases. Businesses use predictive models for decision making, to be proactive, retain existing customers, attract new and grow their profit.

3.3. Predictive Analytics process

Figure 3: Predictive Analytics Process

The above figure describes the steps that are followed in Predictive Analytics process:

1. **Definition of the Problem**

Determine what the problem's outcome, expected results, and business goals will be, and then begin collecting the data sets that will be used.

2. **Data Collection**

All data from different sources is grouped together for use. Data Collection presents an overview of the different customer experiences.

3. **Data Analysis and Preprocessing**

The data is inspected, cleansed, transformed, and modelled to determine whether it truly provides useful information and, ultimately, to come to a conclusion.

4. Statistical Analysis

This allows you to validate whether your findings, assumptions, and hypotheses are reasonable enough to proceed with and test using a statistical model.

5. Modeling

This allows for the creation of precise predictive models for the future. The best choice from the available options could be chosen as the appropriate solution using model evaluation.

6. Deployment

The predictive model deployment option allows the analytics results to be deployed into everyday decision making.

3.4. Predictive Analytical Techniques

Machine learning, data mining, and statistics are only a few of the data analysis methods used in predictive analytics. We'll concentrate on how we can use specific prediction-based methods within the machine learning area to gain greater insight into future events and patterns because machine learning is at the core of predictive analytics.

3.4.1. Classification Model

The majority of machine-learning algorithms are classified as either classification-based or regression-based. Classification algorithms are useful for sorting data into groups, and both forms have different predictive analytics applications.
They will assist businesses in determining whether a website user is a “purchaser” or a “browser,” or whether a subscriber is a “monthly” or “yearly” subscriber type of customer.
Organizations can use classification models to more effectively distribute resources, both human and nonhuman. Companies are ideally able to maintain inventory at acceptable levels and avoid overstaffing a store at certain hours, for example.

3.4.2. Regression Model

These kinds of regression algorithms find patterns that predict relationships between variables.
When a company needs to estimate a numerical value, such as how long a new client will take to return to an airline reservation before purchasing, or how much money someone will spend on car payments for a given period of time, a regression algorithm comes in handy. For instance, linear regression is a commonly used regression technique for determining whether two variables have a relationship.

3.4.3. Neural Networks

Neural networks are biologically inspired data processing methods that take in both historical and current data to predict future values. Their architecture allows them to identify complex correlations hidden in data in a way that mimics the pattern detection mechanisms of the human brain.

Widely used for applications, neural network modeling can predict events by simulating mechanisms of the human brain such as patient diagnosis, image recognition They are composed of several layers that take input (input layer), estimate predictions (hidden layer), and provide output (output layer) as a single prediction.

3.4.4. Decision Trees

Branching is used in decision trees to illustrate the possibilities that arise from each result or selection.

A decision tree is a visual map that looks like an upside-down tree: beginning at the "roots," one moves down through a narrowing set of choices, each of which represents a possible decision outcome. Although decision trees can solve a wide range of classification problems, when used in predictive analytics, they can also address far more complicated questions.

3.5. Training and Test data for Predictive analytics

Any machine learning algorithm needs to be trained before it can be used for prediction. To ensure that the trained model is correct, it also needs to be tested. These two procedures require that the data are split into training and testing data, because the model needs to be tested on unseen data (Burez et al.)

If a sufficiently large data set minimizes the problem of introducing bias, this cannot be generalized to the entire model building process, which requires the separation of training and test data. A model is created from the training data, which can then be used for classification on the test data (Burez et al.). Since the labels of the test data are known, the application to the generated model gives an indication of how well the predictions were made.

A common method for doing this is by the holdout evaluation, in which the data is randomly split into these two sets, which are usually sized so that 2/3 of the data is used for training and 1/3 for testing.

A smaller training set generally results in the model being built on less available data, while a smaller testing set can lead to a less accurate predictor of the achieved accuracy. When applied to a large dataset, holdout evaluation provides a reliable indicator of the model's results. Multiple holdout sampling or cross validation are other options that re-sample the data at random.

Other methods for re-sampling the data at random include multiple holdout sampling and cross validation.

Cross validation is mainly used for prediction, and we want to estimate the performance of a predictive model.

Multiple holdout sampling or cross validation, which re-sample the data one or more times at random, or bootstrapping sample datasets are other choices.

These are more time - consuming and are commonly only used on small datasets because they significantly increase the computation time on the entire process. (Kohavi et al, 1995)

In our case, one can over-sample the churned customers or under-sample the no churned customers.

Oversampling alone can increase noise or contribute to overfitting, while under sampling alone can minimize the amount of information available.

As Liu et al. suggested, combining the two approaches is a promising solution. “Over-sampling the minority class offers complementary information for the training data, and under-sampling alleviates the over-fitting problem,” (Liu et al. 2007)

To maintain the statistical properties of the original dataset and to ensure that the unbalanced dataset had a reasonable balance between the size and representation of the training and test sets, the data was divided into a training set (70 %) and a test set (30 %).

The decision to use a 70:30 dataset split ratio was affected in the first place by the fact that the number of data instances was sufficiently large. In general, having less training data resulted in higher variance in parameter estimates, while having less testing data resulted in higher variance in static results.

The aim was to ensure that dividing the data into training and test sets resulted in a low variance, which could be accomplished by using a 70:30 ratio.

3.6. Imbalanced data: consequences and treatments

Regarding the problem of class imbalance, Weiss (2004) identifies six types of issues that occur when mining unequal classes.

I. ***Incorrect assessment metrics***: In many cases, the wrong metrics are used to direct data mining algorithms and to analyze data mining performance.

II. ***Relative lack of data: relative rarity***: objects are not uncommon in the absolute sense, but are rare in comparison to other objects, which make greedy search heuristics difficult to use, and more global methods are not tractable in general.

III. ***Lack of data***: absolute rarity: the number of examples associated with the rare class is small in an absolute sense, which makes it difficult to detect regularities within the rare class.

IV. ***Inappropriate inductive bias***: Extra-evidentiary bias is required when generalizing from particular examples, or induction. Inductive leaps are impossible to make without such a bias, and learning is impossible. As a result, a data mining system's bias is crucial to its success. In order to promote generalization and prevent overfitting, often students use a general bias which can have a negative effect on one's ability to learn rare cases and classes.

V. ***Data fragmentation***: Many data mining algorithms, such as decision trees, use a divide-and-conquer technique, in which the original problem is broken down into smaller and smaller problems, resulting in the partitioning of the instance space into smaller and smaller bits. This is a problem since regularities can only be identified within each individual partition, which means there would be fewer data.

VI. ***Noise***: Any data mining system can be affected by noisy data, but it's important to note that noise has a greater effect on rare cases than on common cases. Weiss, 2004)

CHAPTER 4

Classification problem in predictive analytics: - classifying objects based on measurable and observed attributes

4.1. Definition of Classification Problem and Basic Classification Techniques

Variables can be characterized as qualitative or quantitative variables.

Quantitative variables take numerical values for example a person's age, income, price of a stock etc.

Qualitative are known also as categorical variables, take values with one or more categories. For example: a person's gender (female or male), whether a person has a loan (yes or no values) or a cancer diagnosis (Leukemia or No Leukemia).

Problems with a categorical variable are known as **Classification Problems.**

Classification Problems in Machine Learning and Statistics is a predictive modeling problem in which we want to assign a class label category to a new observation.

There are two types of Classification Problems:

a) Binary Classification and b) Multiclass Classification

Binary Classification is when we must categorize a given dataset into two different classes, for example to determine if a person has a disease or not, given data with certain health conditions.

Multiclass Classification is when the number of classes that we must categorize a given dataset, are more than two. For example, to determine in which specie a flower belongs to, given data about the different species of flowers.

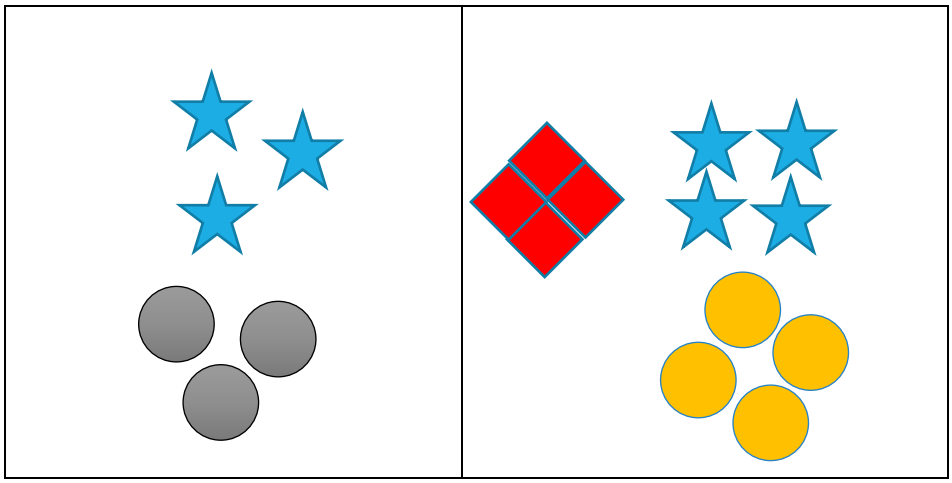


Figure 4: Binary Classification Vs Multiclass Classification

There are many possible classification techniques, or classifiers, that one might use to predict a categorical variable. In this thesis, we will apply two of the most widely-used classifiers for binary classification problem: **Logistic regression** and Artificial Neural Networks and specifically **Radial Basic Function.**

4.2. LOGISTIC REGRESSION

Logistic Regression is a type of probability statistical classification model mainly used for classification problems (Nie, Rowe, Zhang, Tian & Shi, 2011). The technique can work well with a different combination of variables and can help in predicting the customer churn with higher accuracy.

Logistic regression is used to predict a dichotomous categorical variable given a set of predictor variables.

Logistic Regression is often chosen if the predictor variables are not nicely distributed and if they are a mix of continuous and categorical variables.

Logistic regression is one of the most used predictive methods especially in medical research and in customer churn.

Logistic regression may be used to predict the risk of developing a certain disease, for example heart disease or diabetes, based on observed characteristics of the patient. These include body mass index, age, gender, blood results and other characteristics.

Logistic regression is a classification algorithm for predicting the probability of an event's success or failure.

When the dependent variable is binary in nature (0/1, True/False, Yes/No), it is applied. It aids in the classification of data into discrete classes by examining the relationship between a set of sample data. It first learns a linear relationship from the given dataset before introducing a non-linear relationship.

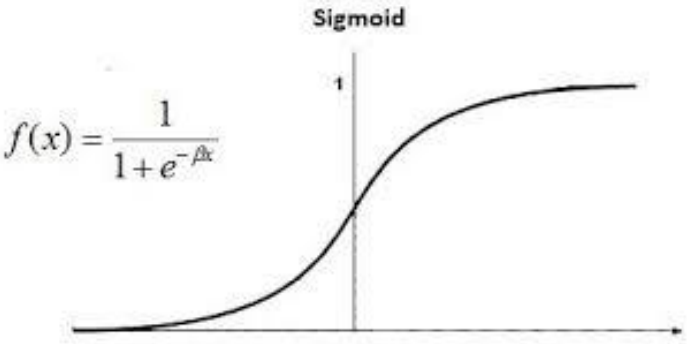


Figure 5: Logistic function

The goal of logistic regression, according to the given use case, is to compute the expected churn behavior given specific data. The churn variable is denoted as y, which has two values: A churned customer is denoted by Y = 1 and a customer who does not churn is denoted by Y = 0. The random variable for the event of churning is Y. This expected churn behavior is written as

$$\pi(x) = E(Y|x),$$

where x are the values of the attributes describing the customer.

Ultimately, one wants to attain a function in the form of

$$P(Y|x1, x2,..., xN) \sim b0 + b1x1 + b2x2 + ... + bdxd$$

The logistic function (logit function) is a link function which is used to transform the values to probabilities. The logistic function creates a sigmoid function which maps the corresponding prediction to each observation along the curve (James et al., 2017).

Logistic equation is:

$$P(x) = \frac{e^{b_0 + b_1 x_1 + \cdots + bdxd}}{1 + e^{b_0 + b_1 x_1 + \cdots + bdxd}}$$

Maximum likelihood

The estimation of b_0,b_1,...,b_d coefficients of the logistic function equation are unknown and have to be estimated using the Maximum Likelihood Estimation.

In logistic regression, we are predicting the probability of a customer falling into a target group [e.g., pr(Y=1, churned)] as a function of the predictors in the model.

If we attempted to model them as a function of the predictors using OLS regression, this would create serious statistical problems, since probabilities are necessarily bounded at 0 and 1, (Pampel, 2000).

In logistic regression, we are not modelling the pr(Y=1) directly as a linear function of the predictors. We use a mathematical transformation of probabilities into a new variable called a logit. This allows us to model pr(Y=1) as a linear function of the predictors. This linearization of the relationship between the predictors and the pr(Y=1) occurs via the use of the logit function (Heck et al., 2012).

$$\mathbf{logit}(\mathbf{Y}=\mathbf{1}) = \mathbf{ln}\left(\frac{\mathbf{p\,r}(\boldsymbol{Y}=\mathbf{1})}{\mathbf{1}-\mathbf{p\,r}(\boldsymbol{Y}=\mathbf{1})}\right) = \mathbf{ln}(\mathbf{odds}(\mathbf{Y}=\mathbf{1})) = \mathbf{b_0} + \mathbf{b_1}\mathbf{X_1} + \cdots \mathbf{b_1}\mathbf{X_k}$$

$$\boldsymbol{odds}(\boldsymbol{Y}=\mathbf{1}) = \boldsymbol{e}^{\boldsymbol{logit}} = \boldsymbol{e}^{\mathbf{b_0}+\mathbf{b_1}\mathbf{X_1}+\cdots\mathbf{b_1}\mathbf{X_k}}$$

$$\boldsymbol{pr}(\boldsymbol{Y}=\mathbf{1}) = \frac{\boldsymbol{odds}(\boldsymbol{Y}=\mathbf{1})}{\mathbf{1}+\boldsymbol{odds}(\boldsymbol{Y}=\mathbf{1})}$$

When implementing Logistic Regression, there are a few criteria that must be met in terms of data. In order to improve the model's performance and ensure that we have correct results. (Garson, G. David. 2009).

Firstly, the answer variable must be binary, which means it can only have two possible values.

Secondly, each predictor variable must have a linear relationship with the logit function outcome. Third, there should be no multicollinearity among the predictors, which means that there should be no high correlations between them.

Finally, in predictors with continuous values, no influential points are possible. Outliers in data are known as influential points.

4.2.1. Advantages and Disadvantages of Logistic Regression

Advantages:

1. It performs well when the dataset is linearly separable and has good accuracy for many simple data sets.
2. It classifies unknown data very quickly.
3. Logistic regression is more straightforward to apply, interpret, and train.
4. It's simple to be applied also to multiple classes (multinomial regression) and a probabilistic view of class predictions.
5. Model coefficients can be interpreted as measures of attribute importance.
6. Logistic Regression, not only informs you how reliable a predictor is (coefficient size), but it also tells you how strong the relationship is (negative or positive.).
7. Makes no assumptions about distributions of classes in feature space
8. Overfitting is less likely in logistic regression, but it can happen in high-dimensional datasets. To prevent over-fitting in these cases, regularization (L1 and L2) techniques could be used. The primary distinction between L1(Lasso)and L2 (Ridge)regularization is that L1 regularization attempts to estimate the data's median, while L2 regularization attempts to estimate the data's mean to avoid overfitting.

Disadvantages:

1. Logistic Regression has Linear decision boundary
2. Logistic Regression should not be used if the number of observations is less than the number of features; otherwise, it should result in overfitting.
3. Since logistic regression has a linear decision surface, it cannot solve non-linear problems. In real-world scenarios, linearly separable data is unusual.

4. The average or no multicollinearity between independent variables is needed for logistic regression.
5. Only discrete functions can be predicted with Logistic Regression. As a result, the dependent variable of Logistic Regression is bound to the discrete.
6. Complex relationships are difficult to obtain using logistic regression. On the other hand, Neural Networks, which are more efficient and compact than this algorithm, can significantly outperform it.
7. The independent and dependent variables are linearly related in Linear Regression. However, in order to use Logistic Regression, independent variables must be linearly related to the log odds (log (p/ (1-p)).

4.3. RADIAL BASIC FUNCTION

Radial Basic Function Network is a supervised learning network which is feedforward and comprises of three layers including an input layer, only one hidden layer with a nonlinear RBF activation function, based on the fact that the nervous system is comprised of a large number of neurons with locally tuned receptive fields, and a linear output layer.(Bishop, 2008). Radial Basic Function network can be used for classification and prediction problems. It takes a different approach than multi-layer perceptron (MLP) neural network since the nodes from the Hidden layer implement a set of radial basic functions.

The RBF layer transfer function is generally a Gaussian function which has the following equation:

$$\boldsymbol{\varphi(x) = exp(-\frac{\|\chi - \mu\kappa\|}{2\sigma\kappa^2}}$$

where $\|\boldsymbol{\chi - \mu k}\|$ is the Euclidean distance between the input data vector x and the corresponding center μκ and σκ is the width parameter, which can adjust the sensitivity of RBF neuron.

RBF networks have three layers:

1. **The Input Layer:**

Each predictor variable is represented by one neuron in the input layer. If there is more than one predictor variable, then the RBF function has as many dimensions as many are the variables. N-1 neurons are employed for categorical variables, where N is the number of categories.
By subtracting the median and dividing by the interquartile range, the input neurons (or processing before the input layer) normalize the range of values. The values are then fed to each of the neurons via the input neurons in the hidden layer.

2. The Hidden Layer:

The number of neurons in this layer varies (the optimal number is determined by the training process). Each neuron is mainly composed of a radial basis function centered on a point that has the same number of dimensions as the predictor variables. For each dimension, the RBF function's spread (radius) may be different.

The training process determines the centers and spreads. Given the x vector of input values from the input layer, a hidden neuron computes the Euclidean distance of the test case from the neuron's center point and then uses the spread values to apply the RBF kernel function to this distance. The summation receives the resulting value.

3. The Output Layer:

The value output by a hidden layer neuron is multiplied by a weight associated with the neuron (W_1, W_2,...,W_n in this case) and transferred to the summation, which adds up the weighted values and displays the sum as the network's output.

For classification problems, each target category has its own output (along with its own set of weights and summation unit). The probability that the case being evaluated has that category is the value output for that category.

The output layer has a weighted sum of outputs from the hidden layer to from the output networks.

The equation for the k_{th} hidden node is:

$$\boldsymbol{zk} = \mathbf{exp}(-\frac{\|\boldsymbol{\chi} - \boldsymbol{\mu\kappa}\|^2}{\mathbf{2}\boldsymbol{\sigma\kappa}^2}$$

The below figure shows the general form of Radial Basic Function networks, with N inputs, k hidden units and output layer:

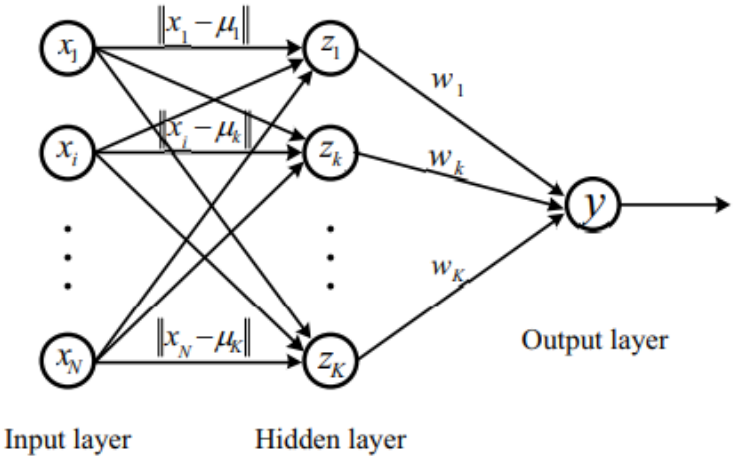


Figure 6: Architecture of an RBF network with Gaussian activation function (Bishop, 2006)

The process of finding the parameters w_k and μ_k from formula (1) is called Training. The RBF network is trained in two stages (Zhang et al., 1996):

1) It uses the two-step clustering algorithm to determine the μ_k centers and their width (Zhang et al., 1996).

2) Given the radial basis function, estimate the synaptic weights w_k.

This cluster algorithm can process both continuous and categorical variables, as well as very huge data sets.

There are two steps:

- Pre-cluster input data into small sub-clusters – scans the data and, if the conditions are met, merges the current input record with previously generated clusters or creates a new one depending on the distance criterion (Integral Solutions Limited, 2007).
- Cluster the sub-clusters into the desired number of clusters by using an agglomerative hierarchical clustering method (Zhang et al., 1996).

Each variable's mean and standard deviation are computed for each cluster.

The center μ_k of the k_{th} radial basis function is equal to the k_{th} cluster mean of the i_{th} input variable if the input variable is a continuous variable;

The center μ_k of the k_{th} radial basis function is equal to the proportion of the category of a categorical variable that the ith variable corresponds to.

The output Layer has the following equation:

$$Y = \sum_{k=1}^{N} w_k z_k$$

Due to the simplicity of these two stages, the training speed of the RBF network is much faster than that of the MLP.

4.3.1 Advantages of RADIAL BASIC FUNCTION

We can easily interpret what is the meaning / function of each node in hidden layer of Radial Basic Function Network.

Radial Basis Function Neural Network (RBFNN) has deep physiological basis, simple network structure, rapid learning ability, and excellent approximation performance, so it has been widely used in the fields of f signal processing, system identification, function approximation and pattern recognition, Today; it remains an important part of neural network research.

4.4. Evaluation and Comparison of Predictive Models

Evaluation criteria

The percentage of cases that are correctly or incorrectly categorized is associated with the accuracy and miscalculation of a classifier's result on a given test set, while precision and recall are considered as measures of precision and validity, respectively. As previously mentioned, choosing one of these steps is closely related to the case of classification.

General classification evaluation rules like prediction accuracy can no longer effectively measure the predictive power of the models. Thus, many scholars introduced several measure evaluation criteria such as sensitivity, ROC curve-measure, etc.

As has been mentioned numerous times in the literature, the cost of losing a loyal customer is usually much greater than the cost of taking preventive or retention measures.

We're interested in the accuracy and completeness of classification performance because important and significant customers are usually considered in churn classifications.

In this case study our goal is about predicting whether a customer will churn or not, in order to make a decision about taking proper retention measures, we claim that precision and recall are more eligible indicators.

Accuracy refers to how well the model classifies data points correctly. This metric is appropriate when the observations are evenly distrusted between the two classes.

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

Precision is the ability to find all relevant instances, the percentage of customers predicted to churn that actually churned.

$$\text{Precision} = \frac{TP}{TP+FP}$$

CHAPTER 5

Churn Prediction Analysis Using Basic Classification Techniques: Neural Networks and Logistic Regression

How to predict customer churn in a bank:

Given a randomly sampled population of 10000 customers from three European-based banks, this case study, intends to propose an efficient predictive model for customer churn in banking industry, by using different classification techniques.

Model performance, goodness of fit, feature selection, class imbalance, will be discussed in the following sections.

5.1. Basic research questions of the case study

1. Identify statistically which variables affect customer churn
2. Build a prediction model that will Classify if a customer is going to churn or not
3. Compare and Evaluate the performance of two predictive models.

5.2. Methodology

The data collection and experiments that were conducted in the study are described in this section.

Data analysis: The analysis was performed using SPSS.

For the predictive analytics, two supervised machine learning classifiers were applied: Logistic Regression and Radial Basic Function of Neural Network, in order to predict churned and no churned customers of thesis project. These classifiers were chosen because they are frequently used for predicting churn, and each has demonstrated good and comparable performance in predicting customer churn. The data set was split into training and testing set.

The training set consisted of 70% of the total data set, and the testing set the remaining 30%. After implementation of the predictive models, features were ranked using the information gain ratio. To measure the prediction performance of the different models, the area under the receiver operating characteristic curve (AUC) was obtained, along with measures for precision and recall. Below is the methodology followed:

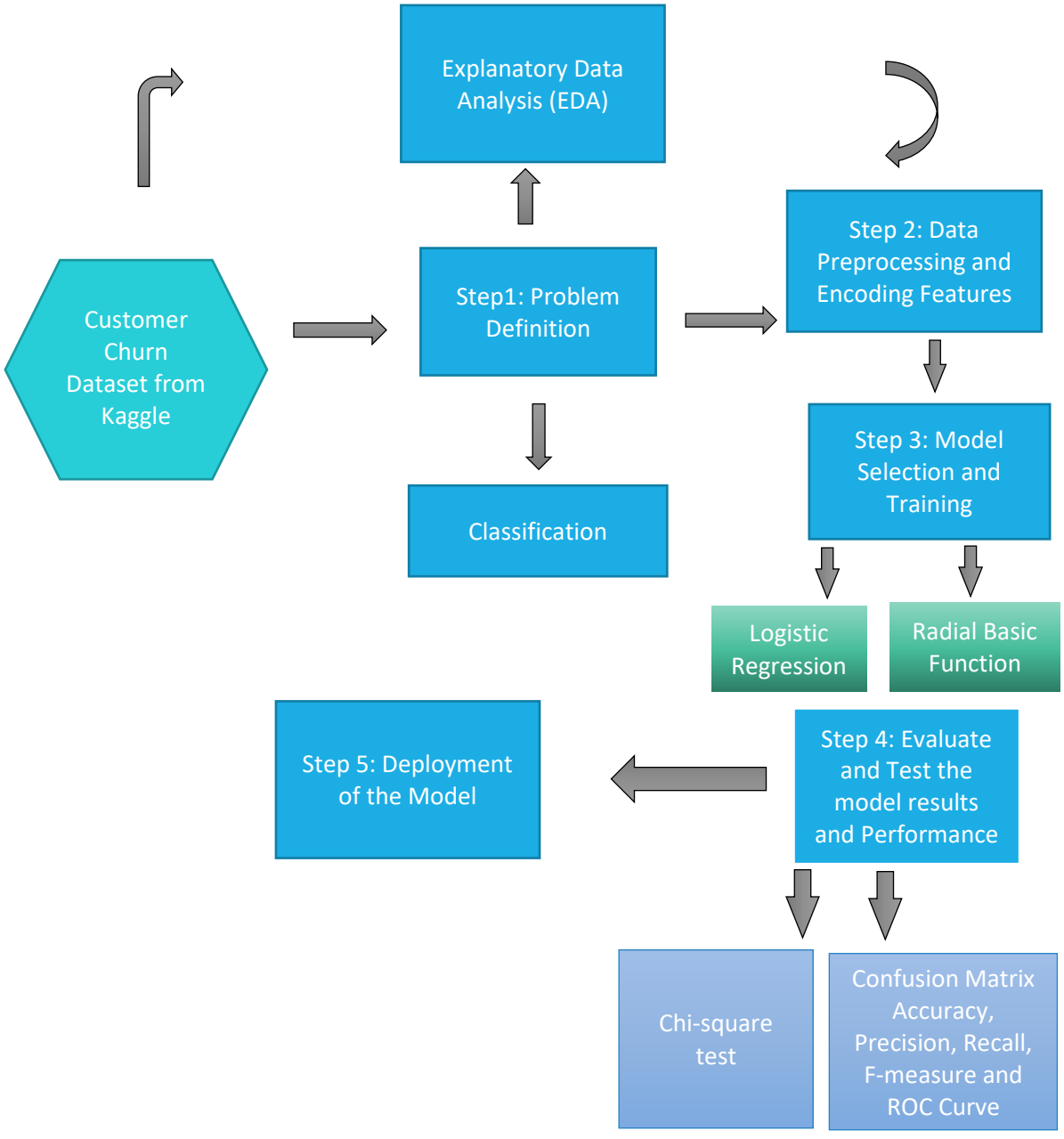


Figure 7: Methodology

5.3 DATA COLLECTION AND DESCRIPTION OF THE DATASET

It is common knowledge that banks do not provide customers' profile information and transaction, due to confidentiality and its sensitive nature. Consequently, the dataset of the case study was downloaded from Kaggle.com (https://www.kaggle.com/mathchi/churn-for-bank-customers/download)

Kaggle is a depository of big data databases used widely by data scientists and it is used to test models, share machine learning techniques and compete.

The dataset represents a collection of information from an imaginary bank.

We will use data entirely in the analysis and don't follow any sampling procedure because we need the training sample to be sufficiently large.

Table 1 shows a description of the 14 variables that are included in the dataset. This includes 13 independent variables and 1 dependent variable.

Out of 13 variables, Customerid and Surname need to be removed as they don't have any contribution to the classification purpose. We have also replaced binary value of the outcome variable (Exited= Churn) with Churned and No churned, in order to have a better representation of the output.

We do not follow any sampling procedure, because Training Sample must be sufficiently large.

VARIABLES		DESCRIPTION
Independent Variables		
1.	RowNumber	Row number of the customer in the csv file. We have 10,000 customers
2.	CustomerId	Each customer's identification number retrieved from bank's records
3.	Surname	Surname of the customer
4.	CreditScore	This score is assigned by the bank to each customer according to personal credit history in order to measure the customer's creditworthiness. The higher the credit score, the more creditworthy the customer is
5.	Geography	The country that the customer lives
6.	Gender	Customer's gender (male or female)
7.	Age	Customer's age
8.	Tenure	How long has been the customer with the bank calculated in years
9.	Balance	Present monetary value of a customer's account
	NumOfProducts	Number of different bank's Products that the customer is currently using (e.g., current accounts, housing loans, personal loans, Internet banking, currency or savings accounts, etc.)
10.	HasCrCard	Indicator whether a customer possesses a credit card or not
11.	IsActiveMember	Indicator of whether a customer has used any of the bank's products in the last 6 months
12.	EstimatedSalary	Estimation of customer's salary
Dependent Variable		
13. Churned		indicates whether the customer has left the bank after 6 months (1 for yes and 0 for no)

Figure 8: Description of the dataset parameters

*CustomerId and Surname will be excluded from the analysis

	Frequency	Percent	Valid Percent	Cumulative Percent
No churned	7963	79.6	79.6	79.6
Churned	2037	20.4	20.4	100.0
Total	10000	100.0	100.0	

Figure 9: Frequency and Percentage of Churn

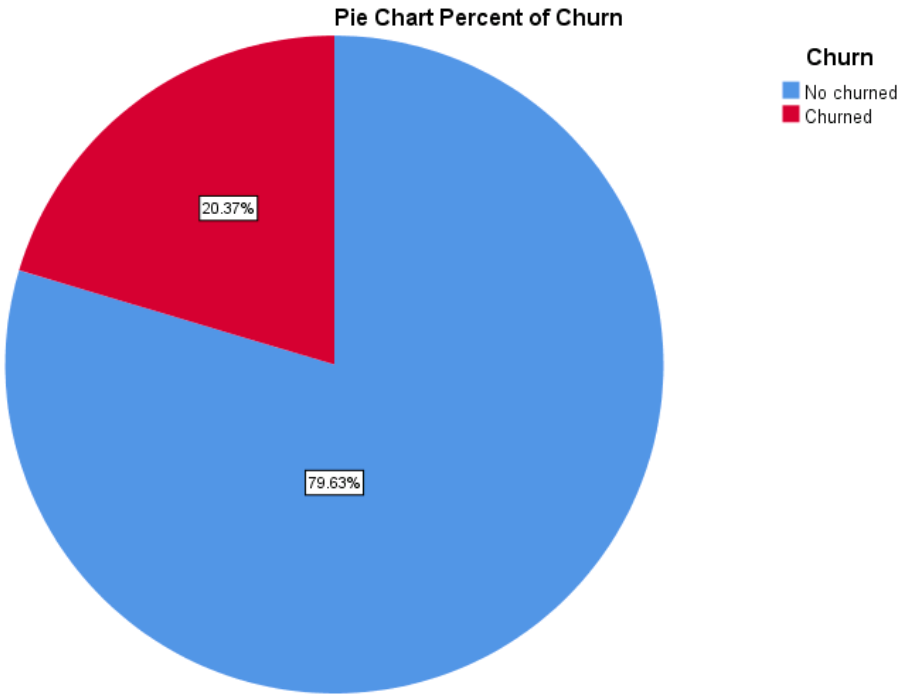


Figure10: Pie Chart Percent of Churn of the dataset

About 20% of the customer of this dataset churned.

Our aim is to build a model that based on customer characteristics could predict specific customer that will abandon their bank to take their business at another bank. We need to ensure that the chosen model does predict with great accuracy those customers as it is of interest to the bank to identify and keep this bunch as opposed to accurately predicting the customers that are retained.

Our dataset is **skewed/imbalanced** since the number of instances in the 'No churned' class outnumbers the number of instances in the 'Churned' class by a lot. Therefore, accuracy is probably not the best metric for model performance.

However, customers who stayed with the bank (7963 customers) are around four times the number of those who left (2037 customers). Therefore, data is imbalance with respect to the outcome variable and this concern needs to be addressed in the modelling section.

5.4. EXPLANADORY DATA ANALYSIS OF DATASET (EDA)

5.4.1 Categorical Variables in the dataset

1. ***GEOGRAPHY:***

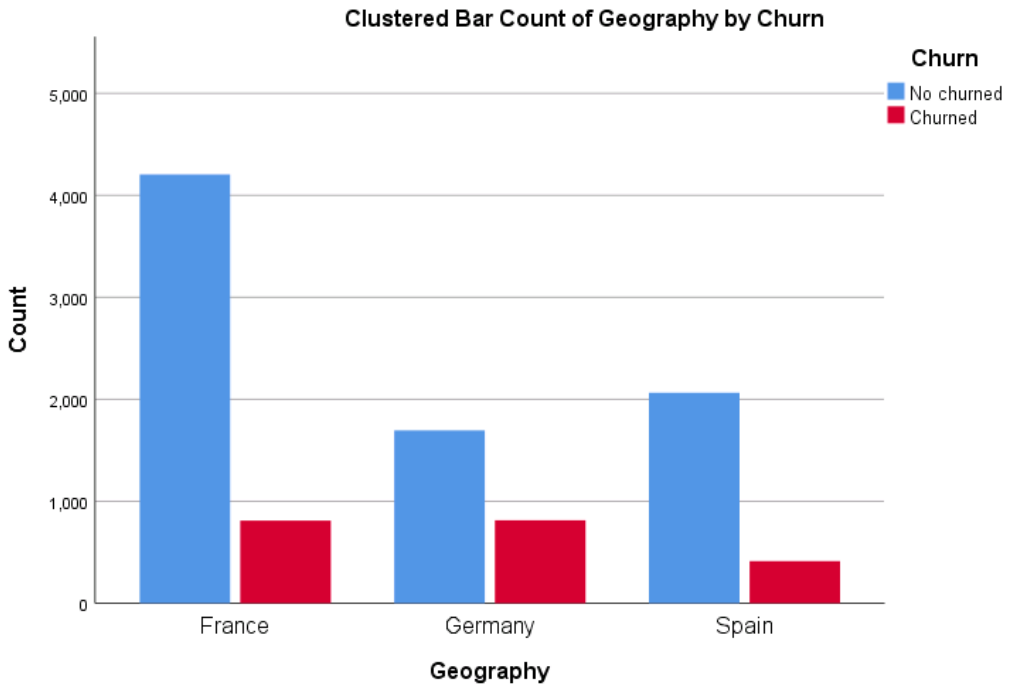


Figure 11: Geography by Churn Bar

The bank has customers in three countries (France, Spain, and Germany).

Most customers are in France.

We can observe that the churn rate for German customers is almost double compared to customers from Spain and France. One could see a division between “Southern” and “Northern” Europe customers with the first group i.e., French customers, to be more loyal to their bank. Many reasons could explain this finding such as higher competition or different preferences for German customers.

Independence of Categorical Variables:

Pearson Chi-square Test Pearson Chi-square test is used to evaluate the variables which are associated with the decision of churn that can be used in the predictive model. Pearson and likelihood ratio chi-square tests are conducted using SPSS Statistics.

The test produced significant results (p–value is less than α level of 0.05) to indicate that some of the variables have an association with the decision to churn.

Our Null hypothesis is: ***whether the two categorical variables are associated with each other – that is, are they dependent or independent.***

Crosstabs

In order two describe the relationship between two categorical variables we use a cross-tabulation (or "crosstab")

The categories of one variable determine the rows of the table, and the categories of the other variable determine the columns.

If we compute the row percentages or column percentages, we can calculate the proportion of the row or column that fell within a particular category, in this case churned or not churned customers.

			Churn		
			No churned	Churned	Total
Geography	France	Count	4204	810	5014
		% within Geography	83.8%	16.2%	100.0%
	Germany	Count	1695	814	2509
		% within Geography	67.6%	32.4%	100.0%
	Spain	Count	2064	413	2477
		% within Geography	83.3%	16.7%	100.0%
Total		Count	7963	2037	10000
		% within Geography	79.6%	20.4%	100.0%

Figure 12: Geography * Churn Cross tabulation

As we can see from the above table 83,8%, 67,6% and 83,3% no churned customers and 16,2%, 32,4% 16,7% churned customers, are from France, Germany and Spain, respectively.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	301.255[a]	2	.000
Likelihood Ratio	280.341	2	.000
N of Valid Cases	10000		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 504.56.

Figure13: Chi-Square Tests

From the above table, the value of the chi square statistic is 301,255. The p-value appears in the same row in the "Asymptotic Significance (2-sided)" column (.000). The result is significant if this value is equal to or less than the designated alpha level (a=0.05).

Since Pearson Chi Square p-value =.000 is less than our chosen significance level α = 0.05, we can reject the null hypothesis that asserts the two variables are independent of each other. The result is significant – the data suggests that the variables Churn and Geography are associated with each other i.e., they are dependent to each other. This indicates that geography affects churn rate.

So, we can conclude that the variable Geography **must be included** in our predictive model.

2. ***GENDER:***

			Churn		
			No churned	Churned	Total
Gender	Female	Count	3404	1139	4543
		% within Gender	74.9%	25.1%	100.0%
	Male	Count	4559	898	5457
		% within Gender	83.5%	16.5%	100.0%
Total		Count	7963	2037	10000
		% within Gender	79.6%	20.4%	100.0%

Figure 14: Gender * Churn Cross tabulation

As we can see from the above table 74,9% and 83,5% no churned customers and 25,1% and 16,5% churned customers, are female and male customers, respectively.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	113.449[a]	1	.000		
Continuity Correction[b]	112.919	1	.000		
Likelihood Ratio	113.044	1	.000		
Fisher's Exact Test				.000	.000
N of Valid Cases	10000				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 925.41.

b. Computed only for a 2x2 table

Figure 15: Chi-Square Tests

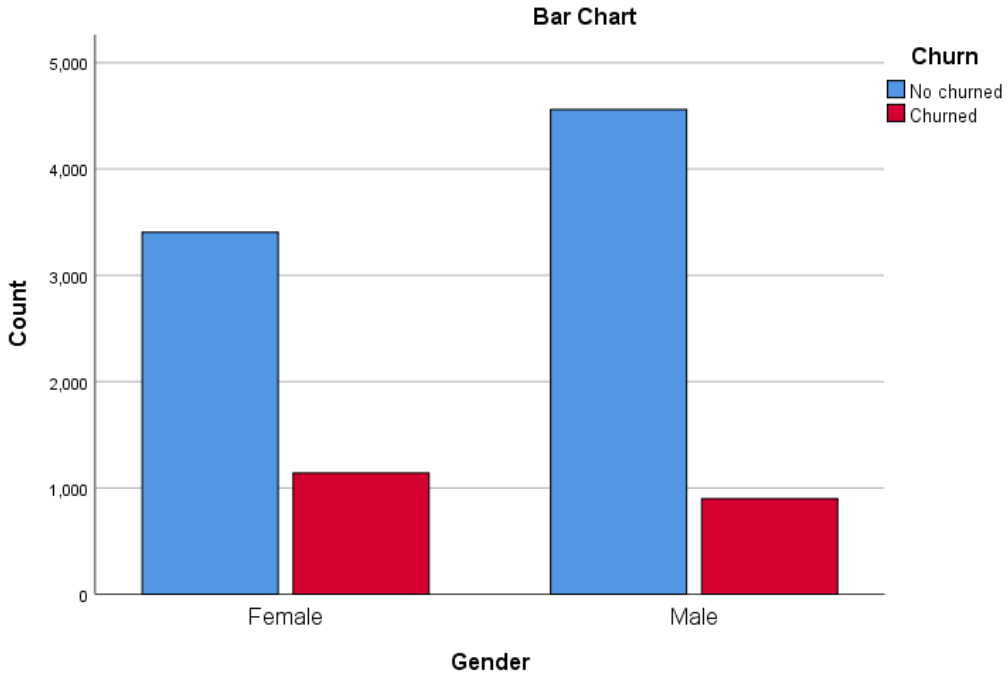


Figure 16: Bar Chart for Gender

There are more male customers than females. Female customers are more likely to churn compared to male customers.

From the above table, the value of the chi square statistic is 113.449. The p-value appears in the same row in the "Asymptotic Significance (2-sided)" column (.000). The result is significant if this value is equal to or less than the designated alpha level (a=0.05).

Since Pearson Chi Square p-value p =.000 is less than our chosen significance level α = 0.05, we can reject the null hypothesis that asserts the two variables are independent of each other. The result is significant – the data suggests that the variables Churn and Gender are associated with each other i.e., they are dependent to each other.

So, we can conclude that the variable Gender **must be included** in our predictive model.

3. ***ISACTIVEMEMBER:***

			Churn		
			No churned	Churned	Total
IsActiveMember	0	Count	3547	1302	4849
		% within IsActiveMember	73.1%	26.9%	100.0%
	1	Count	4416	735	5151
		% within IsActiveMember	85.7%	14.3%	100.0%
Total		Count	7963	2037	10000
		% within IsActiveMember	79.6%	20.4%	100.0%

Figure 17: IsActiveMember * Churn Cross tabulation

As we can see from the above table 73,1% and 85,7% no churned customers and 26,9% and 14,3% churned customers, are no active members and active members of the bank, respectively.

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	243.760[a]	1	.000		
Continuity Correction[b]	242.985	1	.000		
Likelihood Ratio	245.830	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	243.736	1	.000		
N of Valid Cases	10000				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 987.74.

b. Computed only for a 2x2 table

Figure 18: Chi-Square Tests

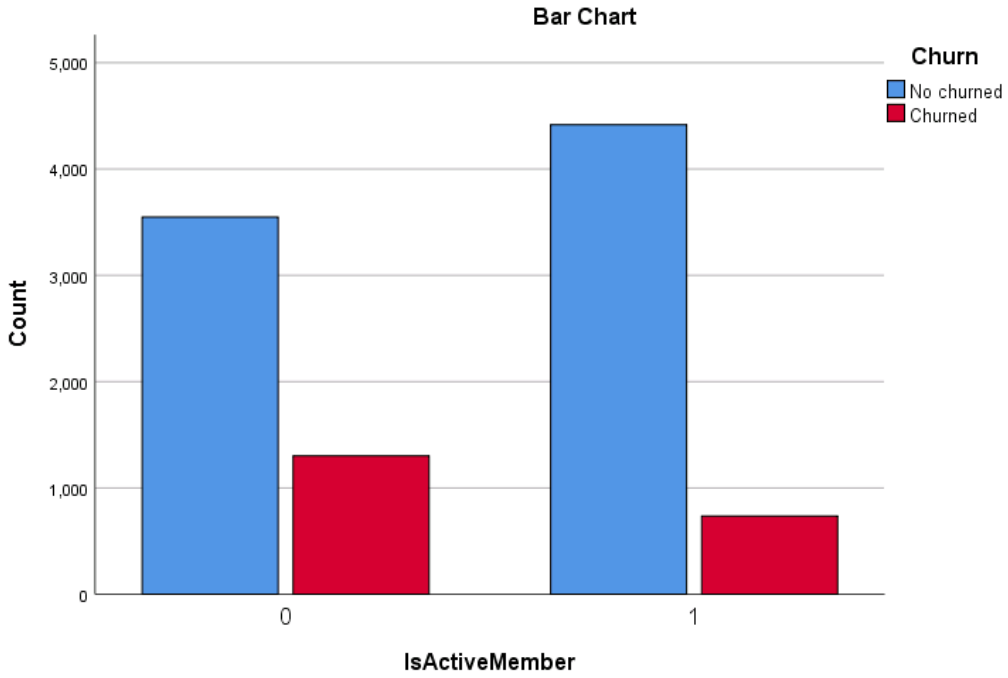


Figure 19: Is Active Member by Churn Bar Chart

We can see that inactive customers are more likely to churn. A big portion of customer is inactive; therefore, the bank will definitely benefit from changing its policy so that more customers become active.

From the above table, the value of the chi square statistic is 243.760. The p-value appears in the same row in the "Asymptotic Significance (2-sided)" column (.000).
The result is significant if this value is equal to or less than the designated alpha level (a=0.05).
Since Pearson Chi Square p-value p =.000 is less than our chosen significance level α = 0.05, we can reject the null hypothesis that asserts the two variables are independent of each other. The result is significant – the data suggests that the variables Churn and IsActiveMember are associated with each other i.e., they are dependent to each other.

So, we can conclude that the variable IsActiveMember **must be included** in our predictive model.

4. ***HASCRCARD:***

			Churn		
			No churned	Churned	Total
HasCrCard	0	Count	2332	613	2945
		% within HasCrCard	79.2%	20.8%	100.0%
	1	Count	5631	1424	7055
		% within HasCrCard	79.8%	20.2%	100.0%
Total		Count	7963	2037	10000
		% within HasCrCard	79.6%	20.4%	100.0%

Figure 20: HasCrCard * Churn Cross tabulation

	Value	Df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.509[a]	1	.475		
Continuity Correction[b]	.471	1	.492		
Likelihood Ratio	.508	1	.476		
Fisher's Exact Test				.479	.246
Linear-by-Linear Association	.509	1	.475		
N of Valid Cases	10000				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 599.90.

b. Computed only for a 2x2 table

Figure 21: Chi-Square Tests

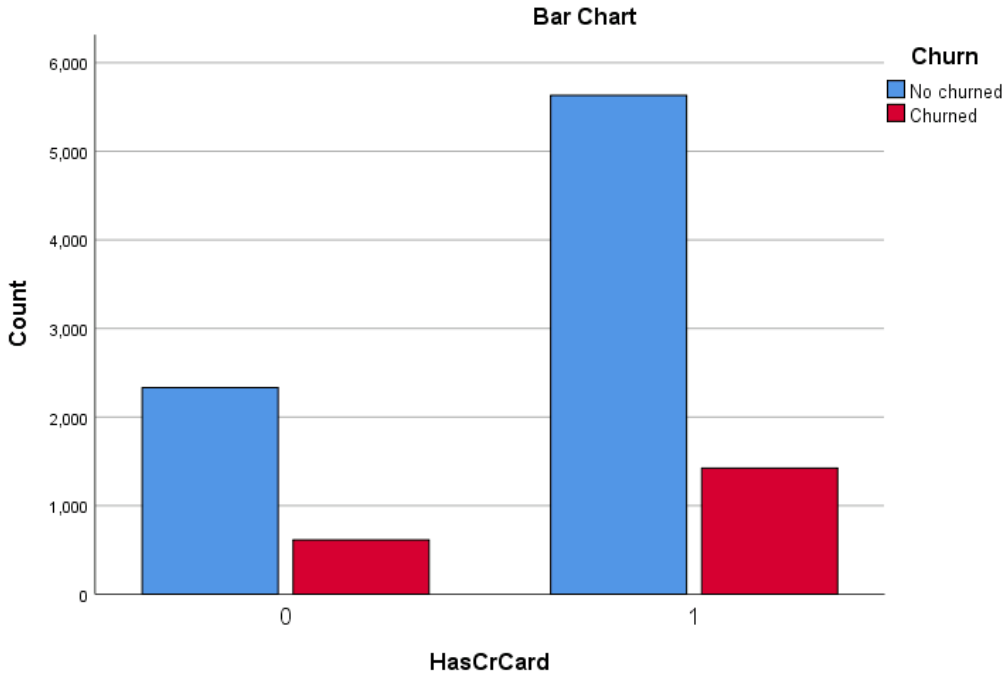


Figure 22: Has CrCard Bar Chart

HasCrCard may not be a useful feature as we cannot really tell if a customer has credit card will churn or not.

From the above table, the value of the chi square statistic is .509. The p-value appears in the same row in the "Asymptotic Significance (2-sided)" column (.475). The result is significant if this value is equal to or less than the designated alpha level (a=0.05).

Since Pearson Chi Square p-value p =.475 is more than our chosen significance level α = 0.05, we cannot reject the null hypothesis that asserts the two variables are independent of each other. The result is not significant – the data suggests that the variables Churn and HasCrCard are not associated with each other i.e., they are independent to each other.

So, we can conclude that the variable HasCreditCard **must be removed** from our predictive model.

5.4.2 Continuous numeric variables in the dataset

We have plotted histograms and box plots for each of the four continuous numerical variables which are: Age, Balance, Estimated Salary and Credit Score comparing the two categories of the dependent variable (churned and no churned):

1. AGE:

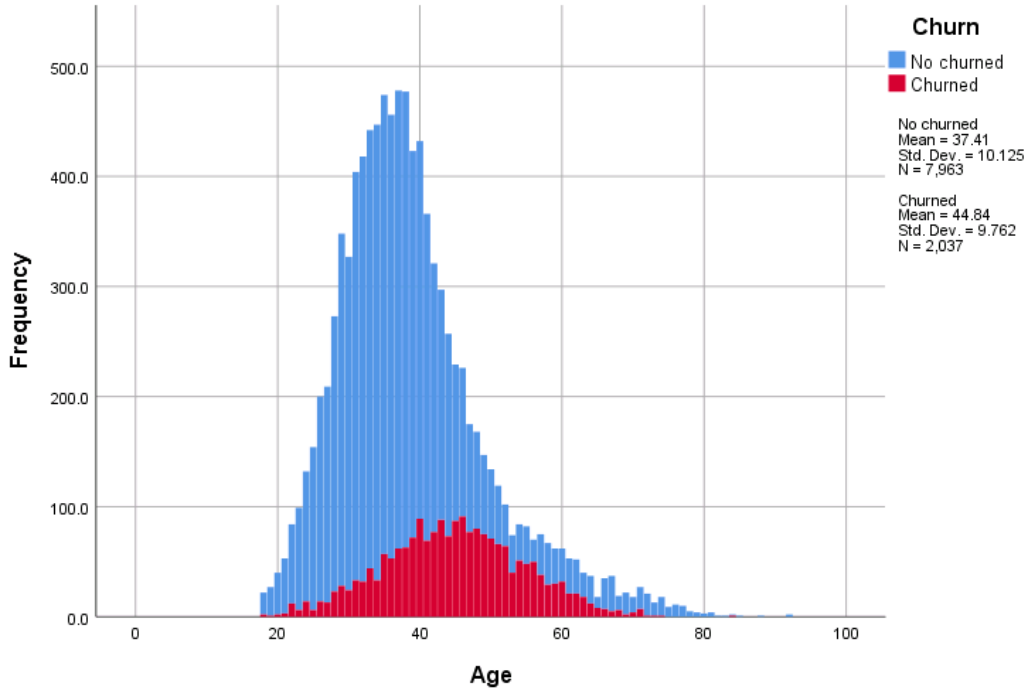


Figure 23: Age Histogram by Churn

Age' is slightly tail-heavy, i.e., it extends more further to the right of the median than to the left.

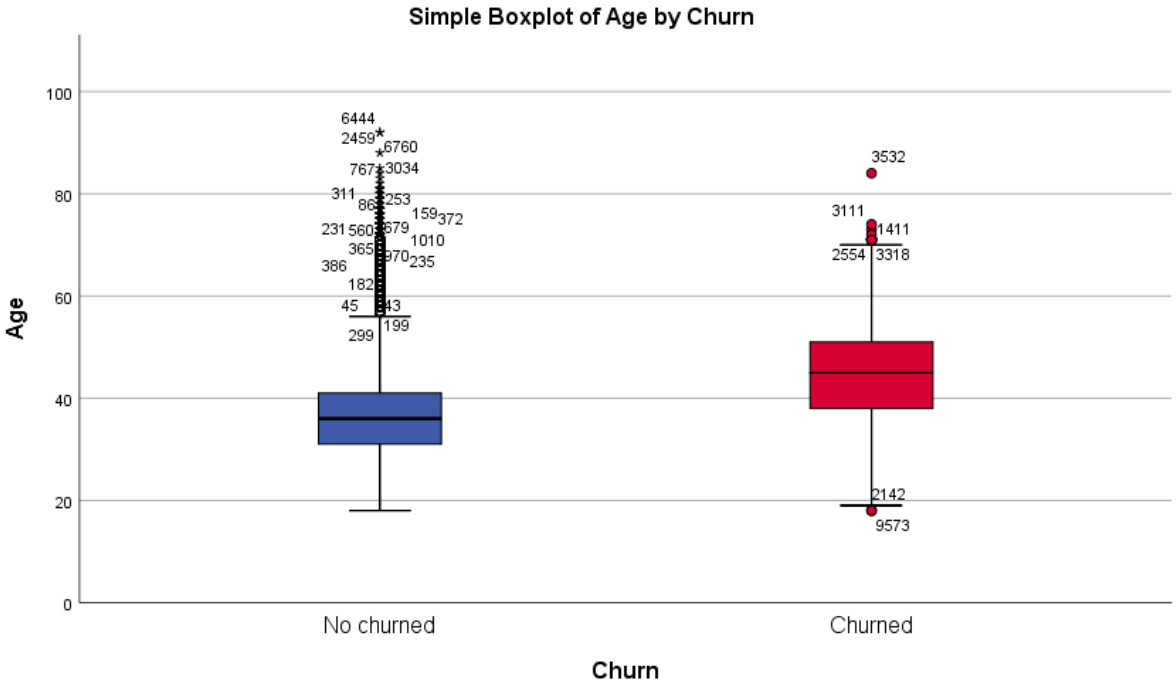


Figure 24: Age Boxplot by Churn

Interestingly, there is a clear difference between age groups since older customers are more likely to churn. This could indicate that preferences of the customers change with age, and the bank hasn't modified its strategy to meet the requirements of older customers.

2. ***CREDIT SCORE:***

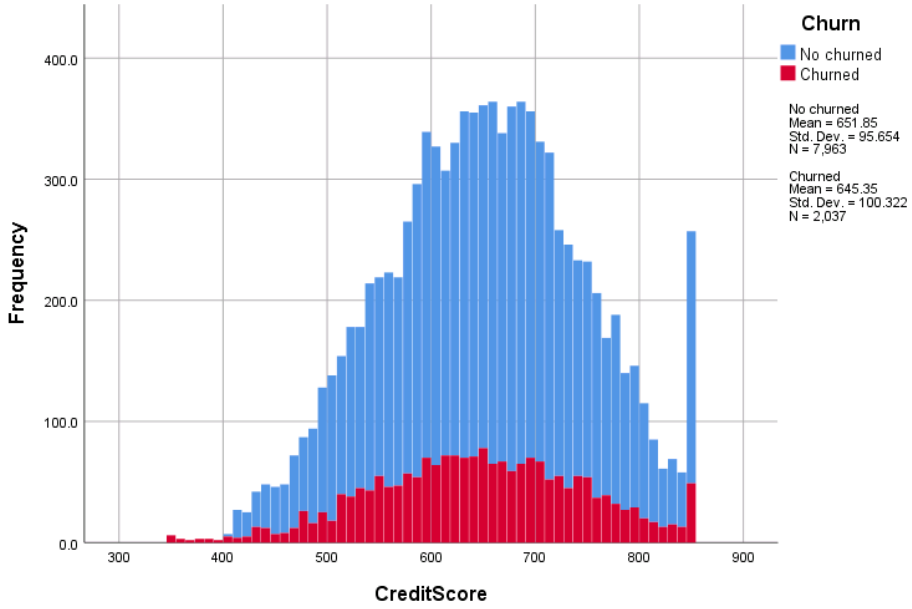


Figure 25: CreditScore Histogram by Churn

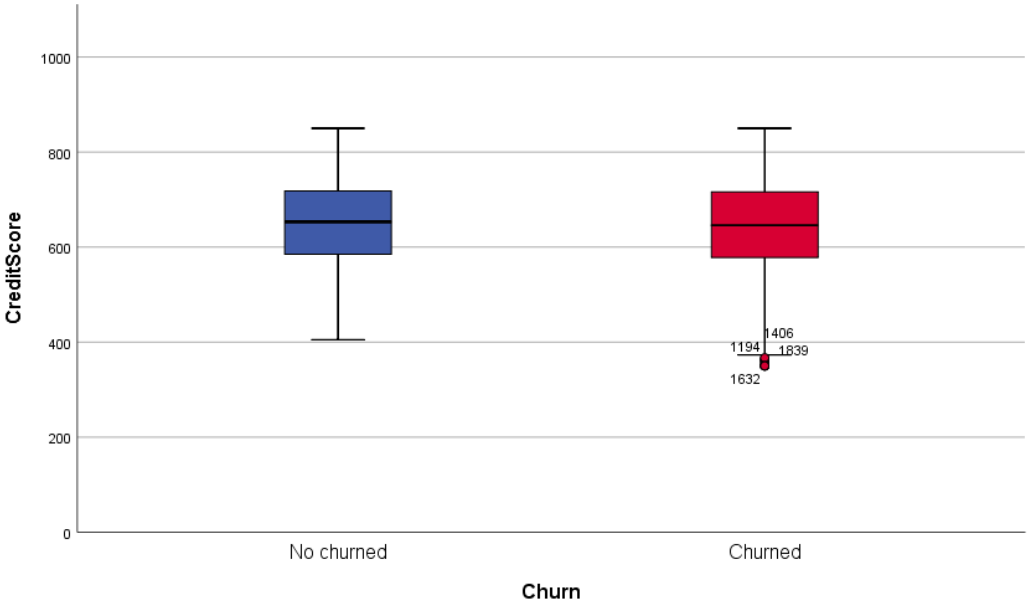


Figure 26: CreditScore Boxplot by Churn

Most values for Credit Score are above 600.

3. ***BALANCE:***

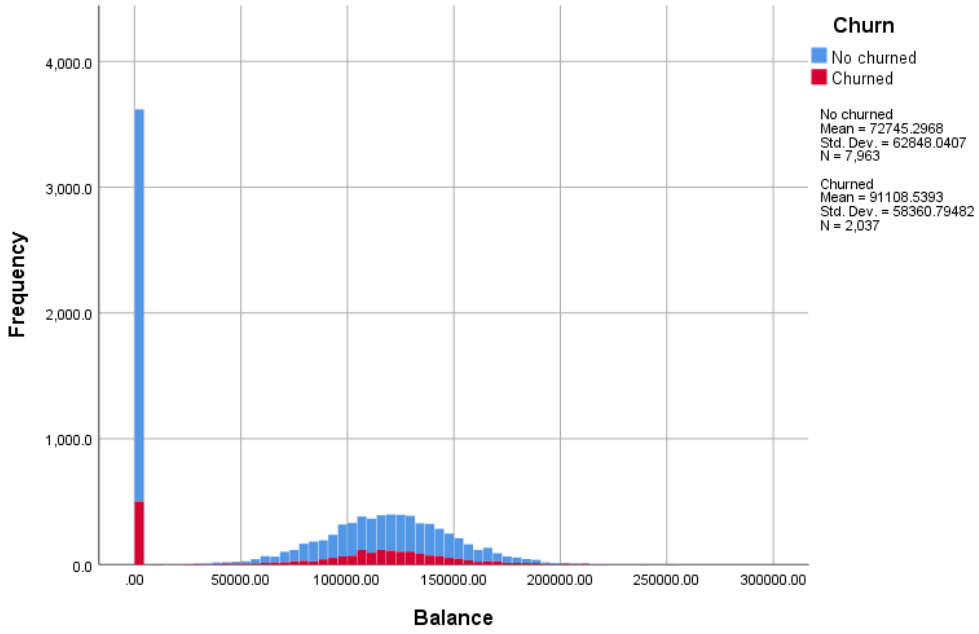


Figure 27: Balance Histogram by Churn

If we ignore the first bin, Balance follows a fairly normal distribution

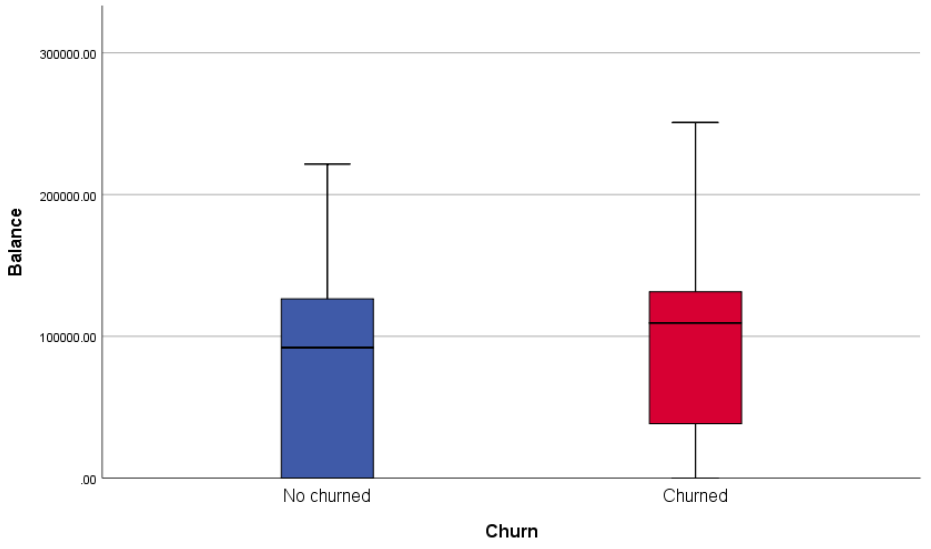


Figure 28: Balance Boxplot by Churn

We can see that the two distributions are quite similar. There is a big percentage of non-churned customers with a low account balance.

4. ***EstimatedSalary***

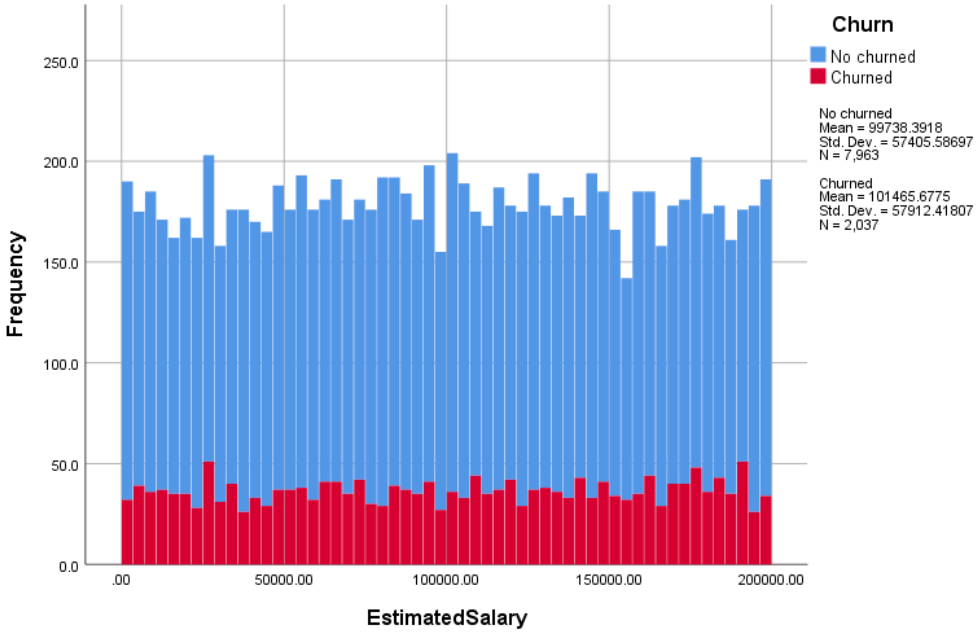


Figure 29: EstimatedSalary Histogram by Churn

The distribution of EstimatedSalary is more or less uniform; Consequently, it provides little information.

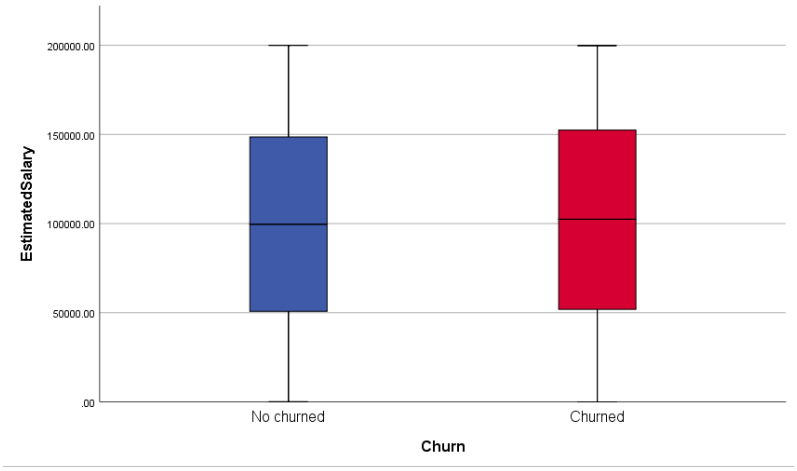


Figure 30: EstimatedSalary Boxplot by Churn

Both churned and no churned customers display a similar uniform distribution for their salary. Consequently, we can conclude that salary doesn't have a significant effect on the likelihood to churn.

5.4.3. Comparing the 2 groups

We proceeded with an independent-samples t-test in order to compare the means between two unrelated groups on the same continuous, dependent variable. We want to understand whether all the below independent variables differed based on churn (which is our dependent variable).Churn has two groups: "churned" and "no churned".

Our **null hypothesis** for an independent samples t-test is that:
“Two populations (churned and not churned in this case) have equal means on some metric variables”.

Group Statistics						
	Churn	N	Mean	Std. Deviation	Std. Error Mean	p-value
NumOfProducts	No churned	7963	1.54	.510	.006	p-value = 0.000 **Significant**
	Churned	2037	1.48	.802	.018	
CreditScore	No churned	7963	651.85	95.654	1.072	p-value = 0.008 **Significant**
	Churned	2037	645.35	100.322	2.223	
Age	No churned	7963	37.41	10.125	.113	p-value = 0.000
	Churned	2037	44.84	9.762	.216	**Significant**
Tenure	No churned	7963	5.03	2.881	.032	p-value = 0.162
	Churned	2037	4.93	2.936	.065	**No Significant**
Balance	No churned	7963	72745.2968	62848.04070	704.29302	p-value = 0.000
	Churned	2037	91108.5393	58360.79482	1293.08086	**Significant**
HasCrCard	No churned	7963	.71	.455	.005	p-value = 0.475
	Churned	2037	.70	.459	.010	**No Significant**
IsActiveMember	No churned	7963	.55	.497	.006	p-value = 0.000
	Churned	2037	.36	.480	.011	**Significant**
EstimatedSalary	No churned	7963	99738.3918	57405.58697	643.30334	p-value = 0.226
	Churned	2037	101465.6775	57912.41807	1283.14632	**No Significant**

Figure 31: independent-samples t-test

Conclusions and Results

❖ NumOfProducts:

The p-value of Levene's test is .000, so we reject the null of Levene's test and conclude that the variance in Number of Products of churned customers is significantly different than that of not churned customers. Then we should look at the "Equal variances not assumed" row for the t test (and corresponding confidence interval) results.

Since p =.000 is less than our chosen significance level α = 0.05, we can reject the null hypothesis, and conclude that the mean Number of products for churned and not churned customers is significantly different.

Based on the above results, we can state the following:

There was ***a significant difference*** in the mean NumofProducts between churned and no churned customers.

❖ CreditScore:

The p-value of Levene's test is .019, so we reject the null of Levene's test and conclude that the variance in Credit Score of churned customers is significantly different than that of no churned customers. Then we should look at the "Equal variances not assumed" row for the t test (and corresponding confidence interval) results.

Since p =.008 is less than our chosen significance level α = 0.05, we can reject the null hypothesis, and conclude that that the mean Credit Score for churned and not churned customers is significantly different.

Based on the above results, we can state the following:

There was a ***significant difference*** in the mean Credit Score between churned and not churned customers.

❖ Age:

The p-value of Levene's test is .003, so we reject the null of Levene's test and conclude that the variance in Age of churned customers is significantly different than that of not churned customers. Then we should look at the "Equal variances not assumed" row for the t test (and corresponding confidence interval) results.

Since p =.000 is less than our chosen significance level α = 0.05, we can reject the null hypothesis, and conclude that the mean Age for churned and not churned customers is significantly different.

Based on the above results, we can state the following:

There was a ***significant difference*** in the mean Age between churned and not churned customers

❖ Tenure:

The p-value of Levene's test is .201, so we cannot reject the null of Levene's test and conclude that the variance in Tenure of churned customers is not significantly different than that of not churned customers. Then we should look at the "Equal variances assumed" row for the t test (and corresponding confidence interval) results.

Since p =.162 is more than our chosen significance level α = 0.05, we cannot reject the null hypothesis, and conclude that the mean Tenure for churned and not churned customers is not significantly different.

Based on the above results, we can state the following:

Our **population means Tenure are equal** between churned and not churned customers.

❖ Balance:

The p-value of Levene's test is .000, so we reject the null of Levene's test and conclude that the variance in Balance of churned customers is significantly different than that of not churned customers. Then we should look at the "Equal variances not assumed" row for the t test (and corresponding confidence interval) results.

Since p =.000 is less than our chosen significance level α = 0.05, we can reject the null hypothesis, and conclude that the mean Balance for churned and not churned customers is significantly different.

Based on the above results, we can state the following:

There was a ***significant difference*** in the mean Balance between churned and not churned customers.

- **HasCrCard:**

The p-value of Levene's test is .158, so we cannot reject the null of Levene's test and conclude that the variance in HasCrCard of churned customers is not significantly different than that of not churned customers. Then we should look at the "Equal variances assumed" row for the t test (and corresponding confidence interval) results.

Since p =.475 is more than our chosen significance level α = 0.05, we cannot reject the null hypothesis, and conclude that the mean HasCrCard for churned and not churned customers is not significantly different.

Based on the above results, we can state the following:

Our **population means HasCrCard are equal** between churned and not churned customers.

- **IsActiveMember:**

The p-value of Levene's test is .000, so we reject the null of Levene's test and conclude that the variance in IsActiveMember of churned customers is significantly different than that of not churned customers. Then we should look at the "Equal variances not assumed" row for the t test (and corresponding confidence interval) results.

Since p =.000 is less than our chosen significance level α = 0.05, we can reject the null hypothesis, and conclude that the mean IsActiveMember for churned and not churned customers is significantly different.

Based on the above results, we can state the following:

There was a ***significant difference*** in the mean IsActiveMember between churned and not churned customers.

❖ EstimatedSalary:

The p-value of Levene's test is .335, so we cannot reject the null of Levene's test and conclude that the variance in EstimatedSalary of churned customers is not significantly different than that of not churned customers. Then we should look at the "Equal variances assumed" row for the t test (and corresponding confidence interval) results.

Since p =.226 is more than our chosen significance level α = 0.05, we cannot reject the null hypothesis, and conclude that the mean EstimatedSalary for churned and not churned customers is not significantly different.

Based on the above results, we can state the following:

Our **population means EstimatedSalary are equal** between churned and not churned customers.

EDA revealed that the variables that can be dropped from our predictive model as they do not provide any value in predicting our target variable are:

1. Customer row
2. Customer id
3. Customer surname
4. Has Credit Card
5. Tenure
6. Estimate Salary
7. Credit Score
8. Balance

Keeping variables that are not statistically significant can reduce the model's precision.

The variables that should be included to our predictive model are the following:

1. Age
2. Gender
3. Number of products
4. Is Active Member
5. Geography

5.5. Encoding Categorical Variables

Machine learning algorithms usually require that all input (and output) features are numeric. Consequently, categorical variables need to be converted (encoded) to numbers, or to be regrouped before using them for building models.

Our dataset contains three variables that require recoding.

- For Gender', we have recoded Male and Female to two numbers (Male = 1 and Female=0).
- For Geography, we have recoded the three Countries to two subgroups:

First Group is German Customers second group are all other customers from France and Spain which were recoded to FranceSpain.

We have chosen this method since the churn rate for customers in the other two countries is almost equal and considerably lower than in Germany. Therefore, it makes sense to encode this variable so that it differentiates between German and non-German customers.

- For Numberofproducts we have recoded 4 groups to two groups:

First group are the customers we have 1 product and second group are the customers which have 2 and more products with the bank.

5.6. Methodology

Machine Learning Application methodology is that the dataset is split into two data samples: A Training Data Sample and a Test Data Sample.

Then the Machine Learning model is applied to the training data and the model learns from the data. Then the model is applied to the Test Sample which is yet unseen data from which the model predicts the predicted values which are compared to the real values. Evaluation metrics are calculated from the differences between the predicted values and real values with conclusions of how good the predictive model is. (Louridas and Ebert, 2016).

5.7. Research question

Evaluate and compare the performance of the two basic classification techniques Logistic Regression and Neural Network-specifically Radial Basic Function.
Our methodology includes applying both techniques to the same Kaggle dataset.

5.7.1. Radial Basis Function

The first classification predictive technique that we have computed is Neural Network and specifically Radial Basic Function. Radial Basic Function predictive technique produces a predictive model for one or more dependent variables, based on values of predictor variables

Case Processing Summary

		N	Percent
Sample	Training	7026	70.3%
	Testing	2974	29.7%
Valid		10000	100.0%
Excluded		0	
Total		10000	

Figure 32: Training and Test Samples Summary

We have divided the dataset in training and test sample data using the variables we found significant in EDA: i.e. Geography_regrouped, NumofProducts_regrouped, Gender, IsActiveMember and Age. SPSS has automatically chosen 70% of the data as the Training Sample and 30% of the data as the Testing Sample.

For this case Study we have 70% of all cases (I.e., 7026 cases) were assigned to the Training Sample and 30% of all cases (i.e., 2974 cases) were assigned to the Testing Sample.

No cases were excluded from the analysis.

Network Information: (appendix page):

Displays information about the neural network. This is useful for ensuring that the specifications are correct.

In the INPUT LAYER we have the number of units which include the number of covariates plus the number of factors. In this case:

Factors are:

- Geography_regrouped
- NumberofProducts_regrouped
- Gender
- Is Active Member

Covariates are:

- Age

A separate unit is created for each category of factors and covariates, none of them are considered redundant units. In this case we have 9 units.

In the OUTPUT LAYER a separate output unit is created for each category of the dependent variable “churn”. In this case we have two units (i.e., for churned/not churned)

Covariates are rescaled using the adjusted normalized method.

In the HIDDEN LAYER automatic architecture selection has chosen 4 units.

All other information in the Network is default for Radial Basic Function.

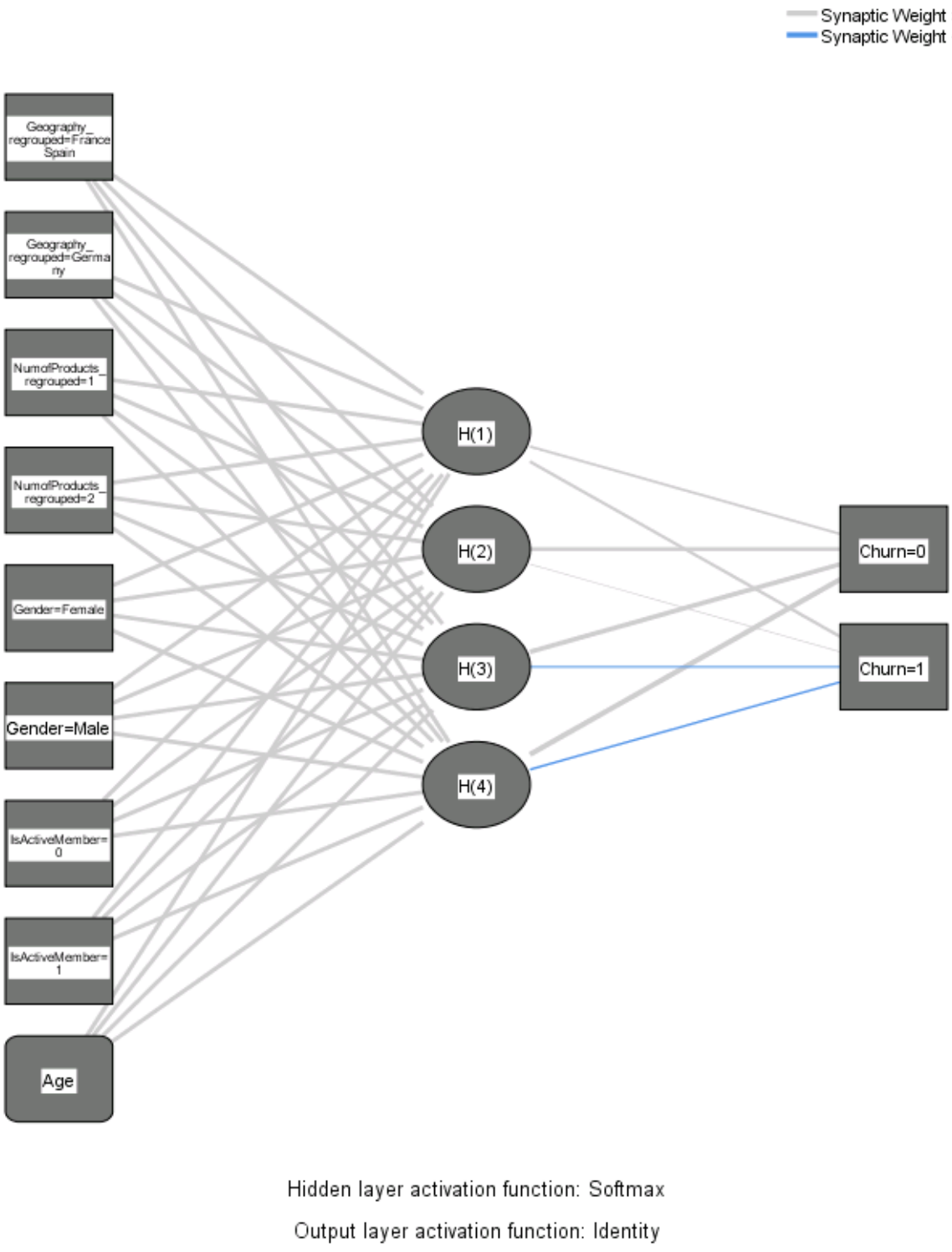


Figure 33: Radial Basic Function Model: Input Layer; Hidden Layer and Output Layer

Training	Sum of Squares Error	1022.079
	Percent Incorrect Predictions	20.2%
	Training Time	0:00:00.39
Testing	Sum of Squares Error	435.914[a]
	Percent Incorrect Predictions	20.8%

Dependent Variable: Churn

a. The number of hidden units is determined by the testing data criterion: The "best" number of hidden units is the one that yields the smallest error in the testing data.

Figure 34: Model Summary

MODEL SUMMARY

Displays information about the results of Training and applying the final network to the Testing Sample. The percentage of incorrect predictions is ***20.8%.***

The percentage of incorrect predictions in the Training Sample is 20.8%, which means that we have achieved ***79.8% accuracy in our Training Sample.***

Then we can see that when our model is applied to our Testing Sample in order to be confirmed, we see that the percentage of incorrect prediction in the Testing Sample us 20,8% which means that we have achieved ***79,2% of accuracy in our Testing Sample.***

		Predicted		
Sample	Observed	No churned	Churned	Percent Correct
Training	No churned	5609	0	100.0%
	Churned	1416	1	0.1%
	Overall Percent	100.0%	0.0%	79.8%
Testing	No churned	2354	0	100.0%
	Churned	620	0	0.0%
	Overall Percent	100.0%	0.0%	79.2%

Dependent Variable: Churn

Figure 35: Classification Table

CLASSIFICATION TABLE

The Classification Table displays all number of cases classified correctly and incorrectly for each dependent variable category. It also includes the percentage of total cases which were correctly classified by the model.

For each case the predicted response is the category with the highest predicted pseudo-probability.

Cells on diagonal are correct predictions and cell off the diagonal are incorrect predictions.

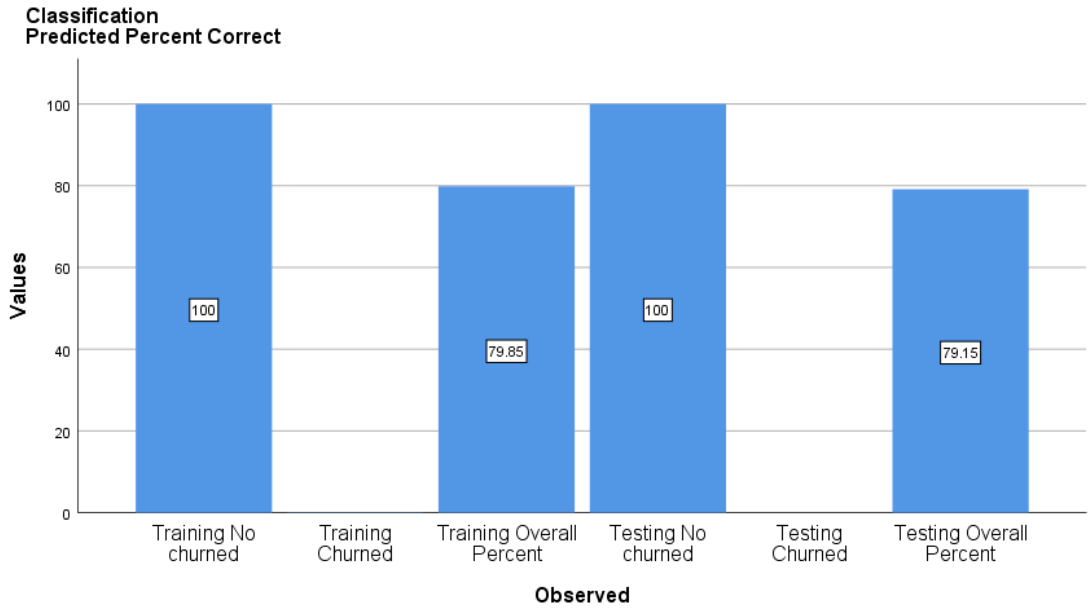


Figure 36: Classification Predicted Percent Correct Bar Chart

ANALYSIS OF THE RESULTS

TRAINING SAMPLE:

5609 No churned customers were also correctly classified as no churned customers, so we have 100% accuracy for no churned customers.

1 churned customer was also correctly classified as churned customer and 1416 churned customers were incorrectly predicted as churned customers. We have 0.1% accuracy for churned customers.

Overall accuracy is 79,8% for the Training Sample.

TESTING SAMPLE:

2354 No churned customers were also correctly classified as no churned customers, so we have 100% accuracy for no churned customers.

620 churned customers were incorrectly predicted as churned customers. We have 0% accuracy for churned customers.

Overall percentage accuracy is 79, 2% for the Testing Sample

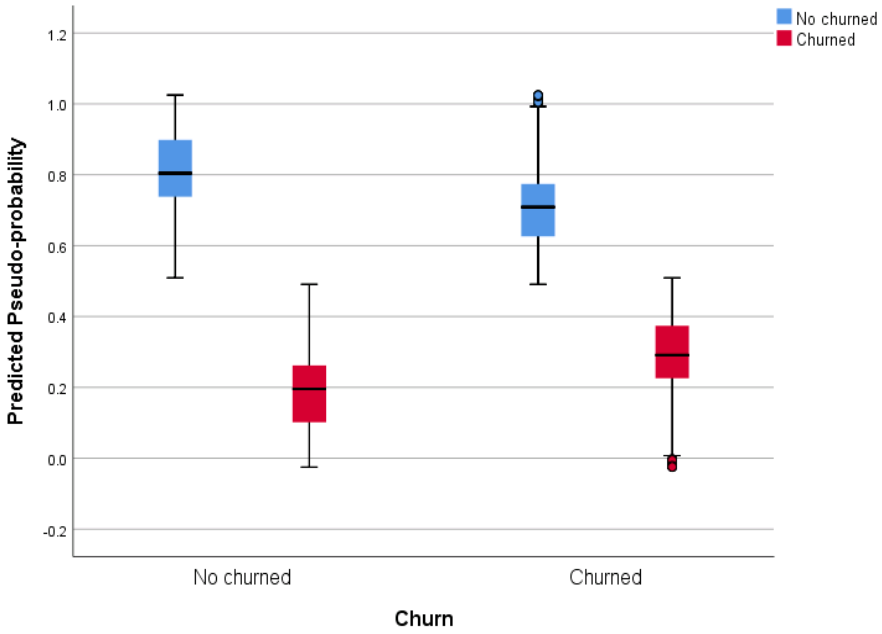


Figure 37: Predicted by observed chart

Predicted by observed chart

It displays a predicted-by-observed-value chart for the dependent variable churn. For our categorical dependent variable, this displays clustered boxplots of predicted pseudo-probabilities for the combined training and testing samples. The *x* axis corresponds to the observed response categories (no churned and churned customers), and the y- axis corresponds to predicted pseudo-probabilities.

- The first boxplot shows, for cases that have observed category no churned customers, the predicted pseudo-probability of category no churned customers.

- The second boxplot shows, for cases that have observed category no churned customers, the predicted pseudo-probability of category churned customers
- The third boxplot shows, for cases that have observed category churned customers, the predicted pseudo-probability of category no churned customers.
- The fourth boxplot shows, for cases that have observed category churned customers, the predicted pseudo-probability of category churned customers

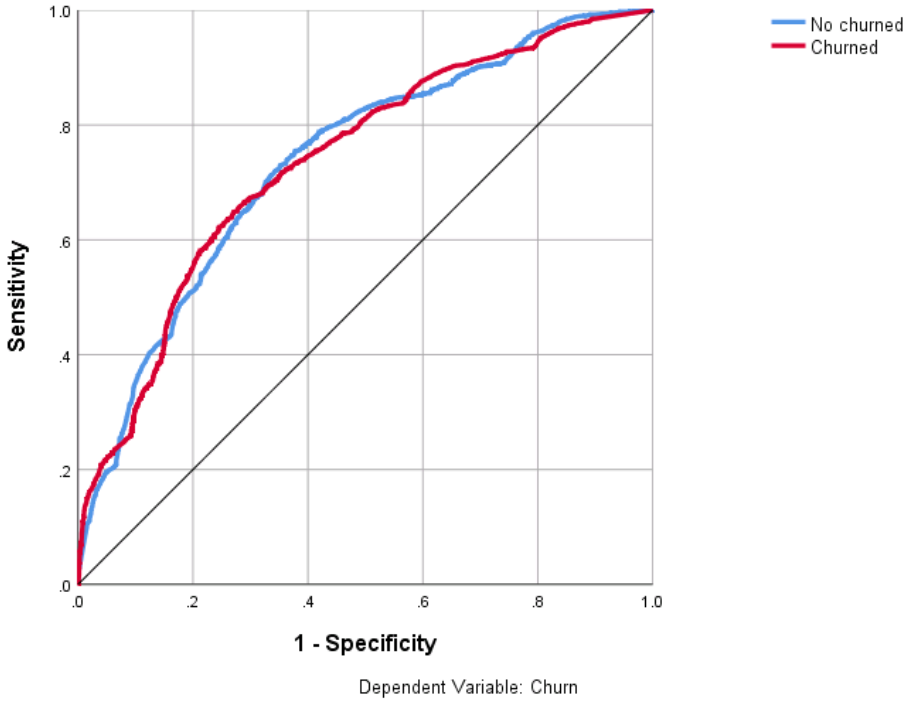


Figure 38: Receiver Operating Characteristic Curve

ROC CURVE (Receiver Operating Characteristic Curve)

It displays the Sensitivity by Specificity for all classification cutoffs. This chart shown here displays two curved, one of each category of the dependent variable churn (Churned/No churned Customers)

It is based on the combined Training and Testing Samples. Commonly used Predictive variables for the performance of a model are the Sensitivity and Specificity.

Sensitivity and **specificity** are statistical measures of the performance of a binary classification test that are widely used:

- **Sensitivity** (True Positive rate) measures the proportion of positives that are correctly identified (i.e., the proportion of those who have some condition (affected) who are correctly identified as having the condition).
- **Specificity** (True Negative rate) measures the proportion of negatives that are correctly identified (i.e., the proportion of those who do not have the condition (unaffected) who are correctly identified as not having the condition).

ROC Curves of a classifier shows its performance as a tradeoff between Sensitivity and Specificity. The equations of these are given by:

$$\text{SENSITIVITY} = \frac{NUMBER\ OF\ TRUE\ POSTITIVES}{NUMBER\ OF\ TRUE\ POSITIVES + NUMBER\ OF\ FALSE\ NEGATIVES} = \frac{TP}{TP+FN}$$

$$\text{SPECIFICITY} = \frac{NUMBER\ OF\ TRUE\ NEGATIVES}{NUMBER\ OF\ TRUE\ NEGATIVES + NUMBER\ OF\ FALSE\ POSITIVES} = \frac{TN}{TN+FP}$$

Where:

TN = True negatives

TP = True Positives

FN = False negatives

FP = False positives

In this case, true positives TP are the no churned customers which are correctly predicted as no churned customers; true negatives TN are the churned customers which are correctly predicted as churned customers; False positives FP are the churned customers which are incorrectly predicted as no churned customers; False Negatives FN are the no churned customers which are incorrectly predicted as no churned customers.

		Area
Churn	No churned	.736
	Churned	.736

Figure 39: Area Under the Curve

AREA UNDER THE CURVE (AUC CURVE)

It is a numerical summary of the ROC Curve and the values in the table represent the probability of each category that the predicted pseudo-probability of being in that category us higher for a randomly chosen case in that category than for a randomly chosen case not in that category. In this case, for a randomly selected customer in no churned customers category and a randomly selected customer in churned category there is a **0.736 (73,6%) probability** that the model -predicted pseudo-probability of default will be higher for the customer in no churned category.

Area under the curve is an evaluation criterion used and a measure for the performance of the model. AUC of a good classifier should be preferably close to 1, as a value of 1 represents a perfect classifier (Wouter Verbeke, Karel Dejaeger, David Martens, Joon Hur, and Bart Baesens, 2012)

Figure present ROC curve for both values of the dependent variable. All points fall on the two jagged curves above the diagonal, indicating good classification.

The Area under the curves based on both Training and Testing Sample was 0.736 indicating a quite high classification accuracy rate, since exceeds 70% for both samples.

Cumulative gains chart

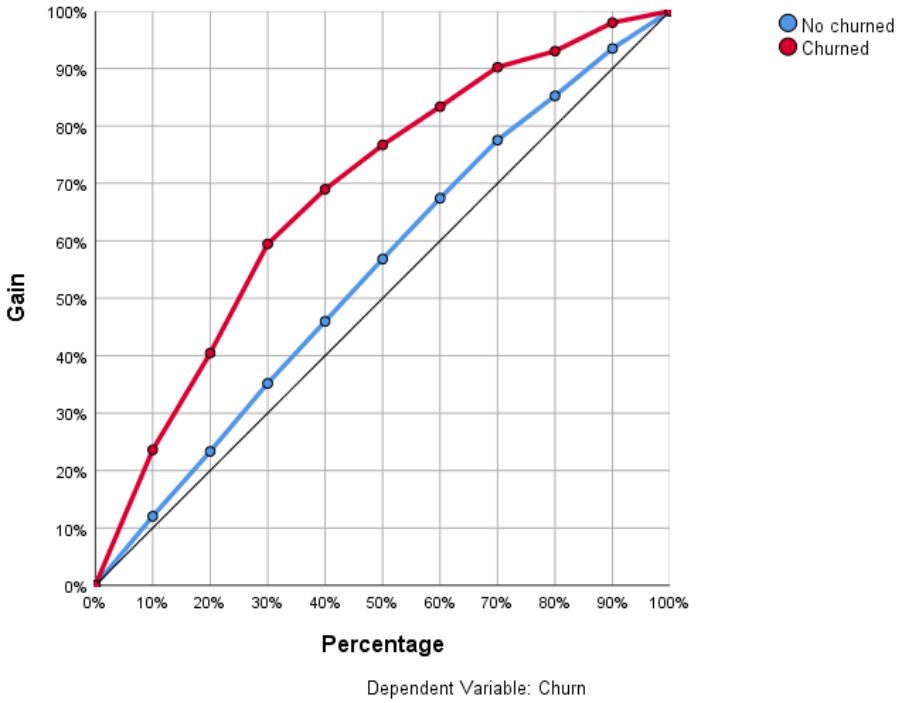


Figure 40: Cumulative gains chart

The cumulative gains chart shows the percentage of the overall number of cases in a given category "gained" by targeting a percentage of the total number of cases and it is based on the combined training and testing samples.

The diagonal line is the "baseline" curve and the farther above the baseline a curve lies, the greater the gain.

As we can see from the above chart, the first point on the curve for example of the *churned customers* category is approximately at (10%, 25%), meaning that if you score a dataset with the network and sort all of the cases by predicted pseudo-probability of *churned customers*, you would expect the top 10% to contain approximately 25% of all of the cases that actually take the category *churned customers* Likewise, the top 20% would contain approximately 40%, the top 30% of cases, would contain approximately 60%, the top 40% of cases, would contain approximately 80%, the top 70% of cases, would contain approximately 90%of all of the cases that actually take the category churned customers.

5.7.2 Logistic Regression

Allows us to predict categorical outcomes based on predictor variables.

Logistic regression estimates the probability of an event (in this case, predict customer churn) occurring. If the estimated probability of the event occurring is greater than or equal to 0.5 (better than even chance), SPSS Statistics classifies the event as occurring (e.g., customer churn being present). If the probability is less than 0.5, SPSS Statistics classifies the event as not occurring. (e.g., no customer churn). It is very common to use binomial logistic regression to predict whether cases can be correctly classified (i.e., predicted) from the independent variables. Therefore, it becomes necessary to have a method to assess the effectiveness of the predicted classification against the actual classification.

In order to run Logistic regression, we have divided dataset in training and test sample and we have selected same size samples as we did for Radial Basic Function. We have randomly selected 70% of the cases for the Training Sample and 30% of the cases for the Testing Sample.7026 out of 10000 cases were selected for Training Sample and 2974 cases for the Testing Sample.

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	7026	70.3
	Missing Cases	0	.0
	Total	7026	70.3
Unselected Cases		2974	29.7
Total		10000	100.0

a. If weight is in effect, see classification table for the total number of cases.

Figure 41: Training and Test Sample

Recoding of Variables

Machine learning algorithms usually require that all input variables are numeric. Consequently, variables need to be converted (encoded) to numbers before using them for building models.

Our dataset contains variables that require recoding.

- For Dependent variable, we have recoded churn to two numbers (No churned=0 and Churned=1)
- For Gender, we have recoded Male and Female to two numbers (Male = 1 and Female=0).
- For NumofProducts_regrouped we have recoded to two numbers (1 Product = 0 and 2 or more Products =1)
- For IsActiveMember we have recoded to two numbers (No active customer =0 and Active Customer=1)
- For Geography_regrouped, we have recoded the two subgroups to two numbers (FranceSpain = 0 and Germany =1)

Original Value	Internal Value
No churned	0
Churned	1

Figure 42: Dependent Variable Encoding

		Frequency	Parameter coding (1)
Gender	Female	3205	.000
	Male	3821	1.000
NumofProducts_regrouped	1	3581	.000
	2 or more	3445	1.000
IsActiveMember	0	3411	.000
	1	3615	1.000
Geography_regrouped	FranceSp	5238	.000
	Germany	1788	1.000

Figure 43: Categorical Variables Codings

This part of the output tells you about the cases that were included and excluded from the analysis, the coding of the dependent variable, and coding of any categorical variables listed on the categorical subcommand.

Classification Table[a,b]								
			Predicted					
			Selected Cases[c]			Unselected Cases[d]		
			Churn			Churn		
	Observed		No churned	Churned	Percentage Correct	No churned	Churned	Percentage Correct
Step 0	Churn	No churned	5576	0	100.0	2387	0	100.0
		Churned	1450	0	.0	587	0	.0
	Overall Percentage				79.4			80.3

a. Constant is included in the model.

b. The cut value is .500

c. Selected cases 7026 from the first 10000 cases (SAMPLE) EQ 1

d. Unselected cases 7026 from the first 10000 cases (SAMPLE) NE 1

Figure 44a: Null Model

Null Model

This part of the output describes a "null model", which is model with no predictors and just the intercept (which SPSS calls the constant). This is why you will see all of the variables that you put into the model in the table titled "Variables not in the Equation".

- For the Training Sample:

5576/7026 No churned customers will not churn (79.36%) and 1450/7026 churned customers wrongly predicted as no churned (20, 63%) and none will churn (0%).

The best strategy is to predict that customers will churn, for every case, with no other information that you would be correct 79.4% of the time.

- For the Testing Sample:

2387/2974 No churned customers will not churn (80, 26%) and 587/2974 churned customers wrongly predicted as no churned (19, 74%).

The best strategy is to predict, for every case that customers will no churned, with no other information that you would be correct 80, 3% of the time.

Block 0: Beginning Block

Variables in the Equation

		B	S.E.	Wald	Df	Sig.	Exp(B)
Step 0	Constant	-1.347	.029	2087.654	1	.000	.260

Figure 44b: Null Model

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Geography_regrouped (1)	236.020	1	.000
		NumofProducts_regrouped (1)	247.830	1	.000
		IsActiveMember (1)	212.332	1	.000
		Age	573.767	1	.000
		Gender(1)	90.308	1	.000
	Overall Statistics		1271.360	5	.000

Figure 44c: Null Model

Under Variables in the Equation, you see that the intercept-only model is Ln (odds) = -1.347. If we exponentiate both sides of this expression we find that our predicted odds [Exp (B)] = .260.

That is, the predicted odds of customer churn are .260. Since 5576 of our customers will not churned and 1450 wrongly predicted as churned, our observed odds are 1450/5576 = .260.

Our Model

Block 1: Method = Enter

		Chi-square	df	Sig.
Step 1	Step	1334.674	5	.000
	Block	1334.674	5	.000
	Model	1334.674	5	.000

Figure 45: Omnibus Tests of Model Coefficients

The section contains what is frequently the most interesting part of the output: the overall test of the model (in the "Omnibus Tests of Model Coefficients" table) and the coefficients and odds ratios (in the "Variables in the Equation" table).

The overall model is statistically significant, since χ^2 (5) = 1334.674, p= .000 < 0.05.This is a test of the null hypothesis that adding the variables to the model has not significantly increased our ability to predict customer churn.

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	5819.422[a]	.173	.271

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Figure 46: Model Summary

Under Model Summary we see that the -2 Log Likelihood statistic is 5819.422.

This statistic measures how poorly the model predicts customer churn -the smaller the statistic the better the model.

This table also contains the Cox & Snell R Square and Nagelkerke R Square values, which are both methods of calculating the explained variation. These values will have lower values than in multiple regression. Therefore, the explained variation in the dependent variable based on our model ranges from 17.30% to 27.10%, depending on whether you reference the Cox & Snell R^2 or Nagelkerke R^2 methods, respectively.

Step	Chi-square	df	Sig.
1	3.499	8	.899

Figure 47: Hosmer and Lemeshow Test

		Churn = No churned		Churn = Churned		
		Observed	Expected	Observed	Expected	Total
Step 1	1	674	677.491	21	17.509	695
	2	676	670.920	29	34.080	705
	3	652	657.004	56	50.996	708
	4	639	633.480	62	67.520	701
	5	618	617.619	89	89.381	707
	6	594	590.194	113	116.806	707
	7	553	549.214	148	151.786	701
	8	491	497.417	210	203.583	701
	9	411	418.347	287	279.653	698
	10	268	264.316	435	438.684	703

Figure 48: Contingency Table for Hosmer and Lemeshow Test

The Hosmer-Lemeshow tests the null hypothesis that predictions made by the model fit perfectly with observed group memberships. A chi-square statistic is computed comparing the observed frequencies with those expected under the linear model. A nonsignificant chi-square indicates that the data fit the model well. So, in our case since p= 0.899> 0.05 no significant, data ***fits the model well***.

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			Churn			Churn		
	Observed		No churned	Churned	Percentage Correct	No churned	Churned	Percentage Correct
Step 1	Churn	No churned	5323	253	95.5	2282	105	95.6
		Churned	1036	414	28.6	442	145	24.7
	Overall Percentage				81.7			81.6

a. The cut value is .500

b. Selected cases 7026 from the first 10000 cases (SAMPLE) EQ 1

c. Unselected cases 7026 from the first 10000 cases (SAMPLE) NE 1

Figure 49: Classification Table[a]

With the independent variables added, the overall percentage correct indicates the percentage of cases with an observed outcome that were correctly predicted by the model ***i.e., Percentage accuracy in classification.***

Classification Table

Training Sample:

- In this output, the overall percentage is 81.7% computed as: 100%[(5323 +414) / (5323+253+1036+414)] = 100%[5737 / 7026] = 81.7%. ***Percentage accuracy in classification for the Training Sample is 81, 7%.***

Sensitivity refers to percentage of cases observed to fall in the target group (Y=1; e.g., observed as churned customers) who were correctly predicted by the model to fall into that group (e.g., predicted as churned customers).

- The sensitivity for the model is calculated as: 100%[414/(414+1036)] = 100%[414/1450] = 28,6% of participants who churned, were also predicted by the model to churn for the Training Sample. ***Sensitivity for the Training Sample is 28, 6%.***

Specificity refers to percentage of cases observed to fall into the non-target category (e.g., observed as no churned customers) who were correctly predicted by the model to fall into that group (e.g., predicted as no churned customers).

- The specificity for this model is calculated as: 100%[5323/(5323+253)] = 100%[5323/5576] = 95.5%. of participants who did not leave, were correctly predicted by the model not to leave. ***Specificity for the Training Sample is 95, 5%.***

Overall, for the Training Sample, the ***accuracy rate was reasonably high*** at 81,7%. The model exhibits ***poor sensitivity*** since among those customers who churned, only 28,6% were correctly predicted as churned by the model. The model exhibits ***very high specificity*** since among those customers who didn't churn 95,5% were correctly predicted not to churn.

Testing Sample:

- In this output, the overall percentage is 81.6% computed as: 100 %[(2282 +145) / (2282+145+442+145)] = 100%[2727 / 2974] = 81.6%. ***Percentage accuracy in classification for the Testing Sample is 81,6% .***
- The sensitivity for the model is calculated as: 100%[145/(145+442)] = 100%[145/587] = 24,70% of participants who churned, were also predicted by the model to churn for the Testing Sample. ***Sensitivity for the Testing Sample is 24, 70%.***
- The specificity for this model is calculated as: 100%[2282/(2282+105)] = 100%[2282/2387] = 95.5%. of participants who did not leave, were correctly predicted by the model not to leave. ***Specificity for the Testing Sample is 95,60%.***

Overall, for the Testing Sample, the accuracy rate was ***reasonably high*** at 81,6%. The model exhibits ***very poor*** sensitivity since among those customers who churned, only 24, 70% were correctly predicted as churned by the model. The model exhibits ***very***

high specificity since among those customers who didn't churn 95, 6% were correctly predicted not to churn.

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Geography_regrouped (1)	.936	.070	179.500	1	.000	2.550	2.224	2.925
	NumofProducts_regrouped (1)	-.972	.069	200.039	1	.000	.378	.331	.433
	IsActiveMember (1)	-1.134	.070	263.403	1	.000	.322	.281	.369
	Age	.072	.003	541.829	1	.000	1.075	1.068	1.081
	Gender (1)	-.568	.066	73.709	1	.000	.567	.498	.645
	Constant	-3.392	.139	596.917	1	.000	.034		

a. Variable(s) entered on step 1: Geography_regrouped, NumofProducts_regrouped, IsActiveMember, Age, and Gender.

Figure 50: Variables in the Equation

This output contains the unstandardized regression slopes and associated significance tests, odds ratios, and confidence intervals (for the odds ratios)

The Wald test ("Wald" column) is used to determine statistical significance for each of the independent variables. The statistical significance of the test is found in the "Sig." column.

From these results, you can see that all added variables as predictors, are significant to the model/prediction. This is a test of the null hypothesis that adding the variables to the model has not significantly increased our ability to predict customer churn.

You can use the information in the "Variables in the Equation" table to predict the probability of an event occurring based on a one-unit change in an independent variable when all other independent variables are kept constant.

In logistic regression, we are predicting the probability of a customer falling into a target group [e.g., pr(Y=1, churned)] as a function of the predictors in the model.
If we attempted to model them as a function of the predictors using OLS regression, this would create serious statistical problems, since probabilities are necessarily bounded at 0 and 1, (Pampel, 2000).
In logistic regression, we are not modeling the pr(Y=1) directly as a linear function of the predictors. We use a mathematical transformation of probabilities into a new variable called a logit. This allows us to model pr(Y=1) as a linear function of the predictors. This linearization of the relationship between the predictors and the pr(Y=1) occurs via the use of the logit function (Heck et al., 2012).

$$\mathbf{logit}(\mathbf{Y}=\mathbf{1}) = \mathbf{ln}\left(\frac{\mathbf{pr}(\boldsymbol{Y}=\mathbf{1})}{\mathbf{1}-\mathbf{pr}(\boldsymbol{Y}=\mathbf{1})}\right) = \mathbf{ln}(\mathbf{odds}(\mathbf{Y}=\mathbf{1})) = \mathbf{b_0} + \mathbf{b_1}\mathbf{X_1} + \cdots \mathbf{b_1}\mathbf{X_k}$$

$$\boldsymbol{odds}(\boldsymbol{Y}=\mathbf{1}) = \boldsymbol{e}^{\boldsymbol{logit}} = \boldsymbol{e}^{\mathbf{b_0}+\mathbf{b_1}\mathbf{X_1}+\cdots\mathbf{b_1}\mathbf{X_k}}$$

$$\boldsymbol{pr}(\boldsymbol{Y}=\mathbf{1}) = \frac{\boldsymbol{odds}(\boldsymbol{Y}=\mathbf{1})}{\mathbf{1}+\boldsymbol{odds}(\boldsymbol{Y}=\mathbf{1})}$$

The Variables in the Equation output shows us that the regression equation is

Ln (ODDS) =-3.392 + 0.936* Geography_regrouped - .972*NumofProducts_regrouped -1.134*IsActiveMember+ 0.072*Age - 0.568*Gender.

A positive coefficient indicates that as the value of the independent variable increases, the mean of the dependent variable also tends to increase. A negative coefficient suggests that as the independent variable increases, the dependent variable tends to decrease.

We can now use this model to **predict the odds** that a customer with a specific characteristic will churn or not. The Variables in the Equation output also gives us the Exp (B). This is better known as the odds ratio predicted by the model. This odds ratio can be computed by raising the base of the natural log to the b^{th} power, where b is the slope from our logistic regression equation. An odd ratio is what the odds of an event is happening.

The odds ratio is the increase (or decrease if the ratio is less than one) in odds of being in one outcome category when the value of the predictor increases by one unit (Tabachnick and Fidell, 2001)

- If an odds ratio is 1, then it is indicating that there is no change in odds per unit increase on the predictor.
- If an odds ratio is > 1, then it is indicating that the odds associated with target group membership are increasing with increases on the predictor.
- If an odds ratio is < 1, then it is indicating that the odds of target group membership are decreasing with increases on the predictor.

The last two columns contain the 95% confidence interval for the odds ratios.

ODDS RATIO = $e^{coefficent}$

a)ODDS for German Customers

ODDS for German Customers=$e^{coefficent}$ = $e^{0.936}$=2.55. That tells us that the model predicts that the odds to churn are 2.55 times higher for German customers than for rest customers, if all other factors being Constant.

b) ODDS for Customer which have 2 or more Products with the bank

ODDS for Customer which have 2 or more Products with the bank =$e^{coefficent}$ = $e^{-0.972}$=0.378, meaning that for the odds of customer to churn (Y=1) change by a factor of .378 with every unit increase on number of products.

Since we are multiplying odds by .378 per unit increase on the predictor, this must mean our odds are decreasing with each increase on the predictor.

Since odds<1, we can simply compute the inverse of the odds ratio and this will provide us with an odds ratio where Y=0 (when interpreting the predictor):

$$OR(Y=0) = \frac{1}{OR(Y=1)}$$

If we take its multiplicative inverse, we obtain OR=2.64[1/.378 = 2.64].

We interpret this to mean that for each one unit increase on number of products, the predicted odds of customers not to churn changes by a factor of 2.64.

This means that customers which maintained 2 or more products with the bank seem to be more loyal and will not leave the bank than those which maintain only 1 product.

c) ODDS for Active Member

ODDS for Active Member=$e^{coefficent}$ = $e^{-1.134}$=0.322. meaning that for the odds of customer to churn (Y=1) change by a factor of .322 with every unit increase on Active Member.

Since we are multiplying odds by .322 per unit increase on the predictor, this must mean our odds are decreasing with each increase on the predictor

If we take its multiplicative inverse, we obtain OR= [1/.322= 3.10].

We interpret this to mean that for each one unit increase on active member, the predicted odds of customers not to churn changes by a factor of 3.10.

This means that the odds for Active members is 3.10 times that of the odds for customers which are active members of the bank seem to be more loyal and will not leave the bank than those which are not active members.

d) ODDS for Age

ODDS for Age =$e^{coefficent}$ = $e^{0.072}$=1.075, meaning that for the odds of customer to churn (Y=1) change by a factor of 1.075 with every unit increase on Age, if all other factors being constant.

This means that customers which are older seem to be more loyal and will not leave the bank than the younger customers.

From the perspective of single factor analysis of age characteristics, the churn rate of low age customers is higher.

Banks should find ways to attract young customers by using online banking, and at the same time, retain older customers with more traditional ways.

e) ODDS for Men Customers

ODDS for Men Customers $=e^{coefficent} = e^{-0.568}=0.567$, meaning that for the odds of customer to churn (Y=1) change by a factor of 0.567 with every unit increase on gender , if all other factors being constant.

Since we are multiplying odds by .567 per unit increase on the predictor, this must mean our odds are decreasing with each increase on the predictor.

If we take its multiplicative inverse, we obtain OR= [1/.567= 1.76].

We interpret this to mean that for each one unit increase on gender, the predicted odds of customers not to churn changes by a factor of 1.76.

This means that female customers seem to be more loyal and will not leave the bank than male customers.

After we got the results of the logistic regression, a new variable is created in the dataset which is called PGR_1. This new variable is the prediction of which customers of our dataset will churn or not.

Interpretation of Results of Logistic Regression Model

In order to get the percentage of correct and incorrect cases, we ran a cross tabulation between our dependent variable churn and predicted variable PGR_1. , for entire dataset.

Cross tabulation for the entire Sample

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Churn * Predicted group	10000	100.0%	0	0.0%	10000	100.0%

			Predicted group		
			No churned	Churned	Total
Churn	No churned	Count	7605	358	7963
		% within Predicted group	83.7%	39.0%	79.6%
	Churned	Count	1478	559	2037
		% within Predicted group	16.3%	61.0%	20.4%
Total		Count	9083	917	10000
		% within Predicted group	100.0%	100.0%	100.0%

Figure 51: Churn * Predicted group Cross tabulation

As we can see from the above table the proportions of no churned customers who were correctly predicted as no churned customers is 83.7% and 39,0% of no churned customers were incorrectly predicted as churned customers by our model.

Also 16.3% of churned customers were correctly predicted as churned customers and 61,0% were incorrectly predicted as churned customers.

Overall performance of our model is 79,6 % correctly predicted for no churned customers and 20,4% for churned customers for Entire Sample.

The Pearson's Chi-Square test will assess whether the two cross-tabulated variables are independent or unrelated. A significant chi-square means that the two variables are related.

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1025.423[a]	1	.000		
Continuity Correction[b]	1022.670	1	.000		
Likelihood Ratio	814.441	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	1025.320	1	.000		
N of Valid Cases	10000				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 186.79.

b. Computed only for a 2x2 table

Figure 52: Chi-Square Tests

As we can see from the above table, Pearson's Chi-Square p=.000 < 0.05 which means that two variables are related, meaning that **our classification model is significant.**

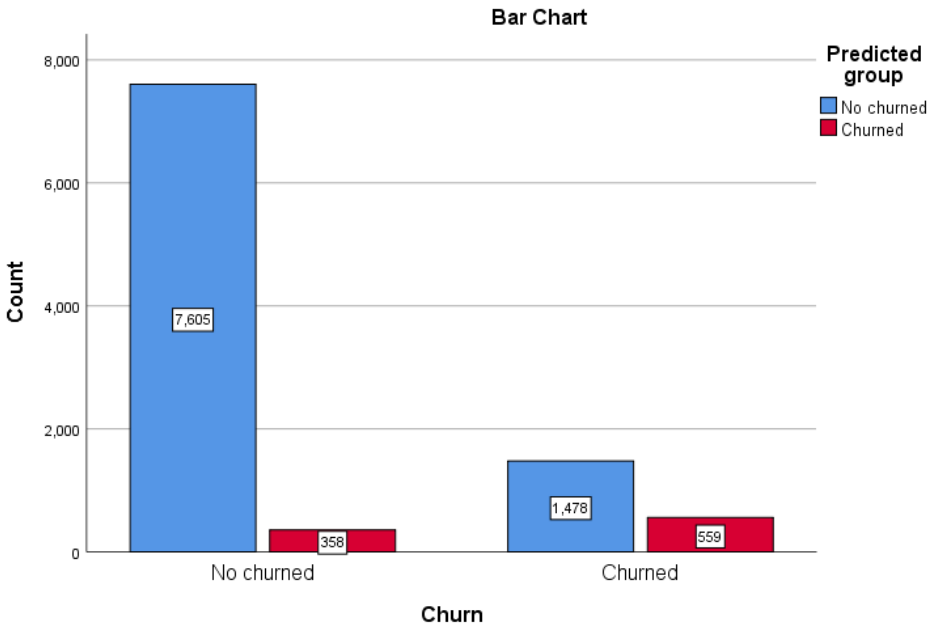


Figure 53: Predicted group Bar Chart

From the above results the model has very good performance on predicting no churned customers but poor performance predicting churned customer for the entire Sample.

Cross tabulation for the Test Sample

1) LOGISTIC REGRESSION:

In order to evaluate the performance of our prediction model, we have selected only the cases that are included in the Test Sample.

We ran a cross tabulation between our dependent variable churn and predicted variable PGR_1. , including only the Test Sample.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Churn * Predicted group	2974	100.0%	0	0.0%	2974	100.0%

			Predicted group		
			No churned	Churned	Total
Churn	No churned	Count	2282	105	2387
		% within Predicted group	83.8%	42.0%	80.3%
	Churned	Count	442	145	587
		% within Predicted group	16.2%	58.0%	19.7%
Total		Count	2724	250	2974
		% within Predicted group	100.0%	100.0%	100.0%

Figure 54: Churn * Predicted group Cross tabulation

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	252.236[a]	1	.000		
Continuity Correction[b]	249.606	1	.000		
Likelihood Ratio	198.846	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	252.151	1	.000		
N of Valid Cases	2974				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 49.34.

Computed only for a 2x2 table

Figure 55: Chi-Square Tests

As we can see from the above table, Pearson's Chi-Square p=.000 < 0.05 which means that two variables are related, meaning that **our classification model is significant.**

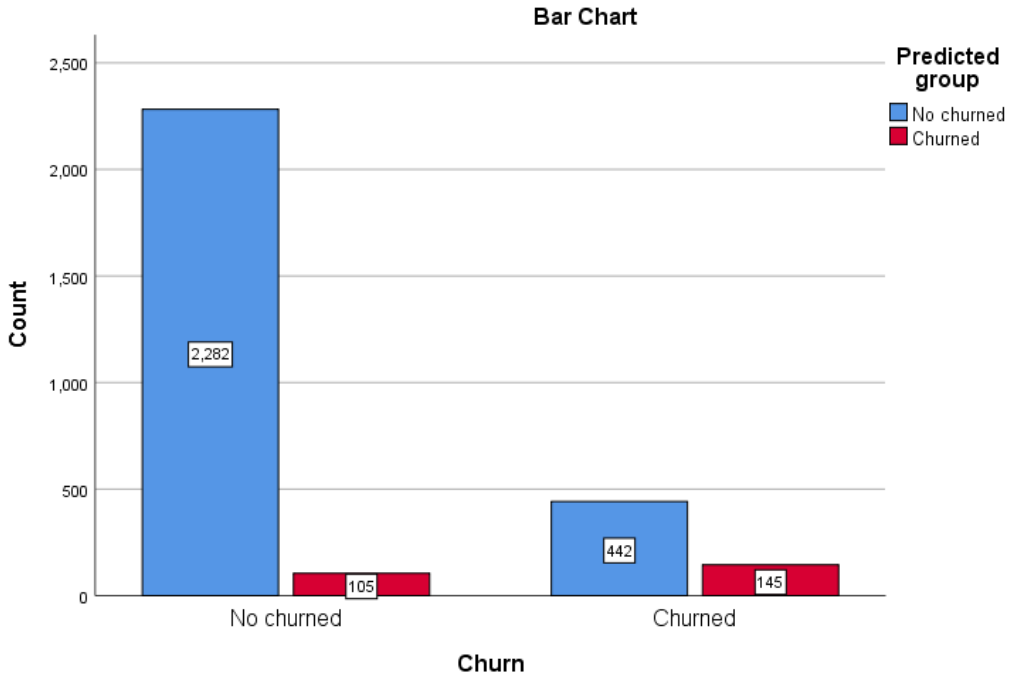


Figure 56: Predicted group Bar Chart

Our prediction model has **correctly predicted 80.3% as no churned customers** and **19, 7% as churned customers.**

From the above results the model has very good performance on predicting no churned customers and poor performance predicting churned customer for the Testing Sample.

2) RADIAL BASIC FUNCTION:

In order to evaluate the performance of our prediction model, we have selected only the cases that are included in the Test Sample.

We ran a cross tabulation between our dependent variable churn and predicted variable RBF_PredictedValue for Radial Basic Function, including only the Test Sample.

			Predicted Value for Churn		
			No churned	Churned	Total
Churn	No churned	Count	2248	138	2386
		% within Predicted Value for Churn	84.9%	42.5%	80.2%
	Churned	Count	401	187	588
		% within Predicted Value for Churn	15.1%	57.5%	19.8%
Total		Count	2649	325	2974
		% within Predicted Value for Churn	100.0%	100.0%	100.0%

Figure 57: Churn * Predicted Value for Churn Cross tabulation

As we can see from the above table the proportions of no churned customers who were correctly predicted as no churned customers is 84.9% and 42,5% of no churned customers were incorrectly predicted as churned customers by our model.

Also 15,1% of churned customers were correctly predicted as churned customers and 57,5% were incorrectly predicted as churned customers.

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	328.098[a]	1	.000		
Continuity Correction[b]	325.431	1	.000		
Likelihood Ratio	262.177	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	327.988	1	.000		
N of Valid Cases	2974				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 64.26.

b. Computed only for a 2x2 table

Figure 58: Chi-Square Tests

The Pearson's Chi-Square test will assess whether the two cross-tabulated variables are independent or unrelated. A significant chi-square means that the two variables are related.

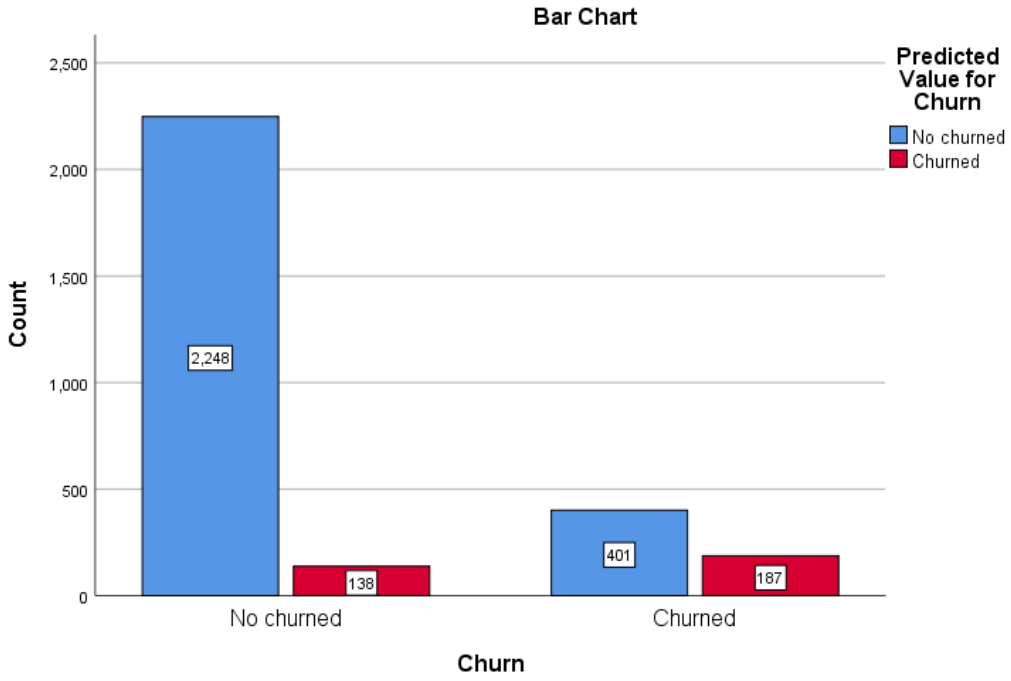


Figure 59: Predicted Value for Churn Bar Chart

Overall performance of our model is 80, 2 % correctly predicted for no churned customers and 19, 8% for churned customers for Testing Sample.

From the above results the model has very good performance on predicting no churned customers and poor performance predicting churned customer for the Testing Sample.

CHAPTER 6

COMPARISON OF THE PERFORMANCE OF THE TWO PREDICTIVE MODELS

LOGISTIC REGRESSION PREDICTION MODEL						RADIAL BASIC FUNCTION PREDICTION MODEL		
Churn * Predicted group Cross tabulation						**Churn * Predicted Value for Churn Cross tabulation**		
			Predicted group			Predicted Value for Churn		Total
			No churned	Churned	Total	No churned	Churned	
Churn	No churned	Count	2282	105	2387	2248	138	2386
		% within Predicted group	83.80%	42.00%	80.30%	84.90%	42.50%	80.20%
	Churned	Count	442	145	587	401	187	588
		% within Predicted group	16.20%	58.00%	19.70%	15.10%	57.50%	19.80%
Total		Count	2724	250	2974	2649	325	2974
		% within Predicted group	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Figure 60: Cross tabulations:

-

Churn * Predicted group Cross tabulation and Churn * Predicted Value for Churn Cross tabulation

6.1. Evaluation Criteria

Accuracy refers to how well the model classifies data points correctly. This metric is appropriate when the observations are evenly distrusted between the two classes.

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

Precision is the ability to find all relevant instances, the percentage of customers predicted to churn that actually churned.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall is also known as sensitivity and as the True Positive Rate because it refers to how well the model classifies true values. It is also described as the percentage of the customers who will churn and those that the model is able to predict as churned.

$$\text{Recall} = \frac{TP}{FN+TP}$$

$$\text{F -- Measure} = \frac{2 \times \textbf{Recall} \times \textbf{Precision}}{\textbf{Recall} + \textbf{Precision}}$$

Model	Accuracy (%)	Precision (%)	Recall (%)	F – Measure (%)
Logistic Regression model	81,60%	58,00%	24,70%	34,65%
Radial Basic Function (Neural Network)	81,87%	57,54%	31,80%	40,96%

Figure 61: Results of evaluation of Predictive Models

The above table results, performance measures for all developed models after splitting the data into training and testing Samples using a 70/30 split rule.

The two models have almost the same percentages.

Radial Basic Function has 81.87% Accuracy ***slightly higher than the percentage of*** Logistic Regression, which has 81.60%.

Radial Basic Function has 57.54% Precision ***slightly lower than the percentage of*** Logistic Regression, which has 58% Precision.

Radial Basic Function has 31.80% Recall which is ***higher than the percentage of*** Logistic Regression, which has 24.70%.

Radial Basic Function has 40.96 % F -measure which is ***higher than the percentage of*** Logistic Regression, which has 34.65%.

Radial Basic Function Model has been selected as the best, because it has shown better performance results for capturing customer churn than Logistic Regression.

The most difficult class of customers to predict and the most important are the churners. This means that the most accurate method for the author's research is the neural network.

6.2. Comparison of the Area under the curve for the two models

Area under the curve is an evaluation criterion used and a measure for the performance of the model. We have calculated an ROC Analysis to compare the two models.

Case Processing Summary

Churn	Valid N (listwise)
Positive[a]	2037
Negative	7963
Missing	0
Total	10000

Larger values of the test result variable(s) indicate stronger evidence for a positive actual state.

a. The positive actual state is Churned.

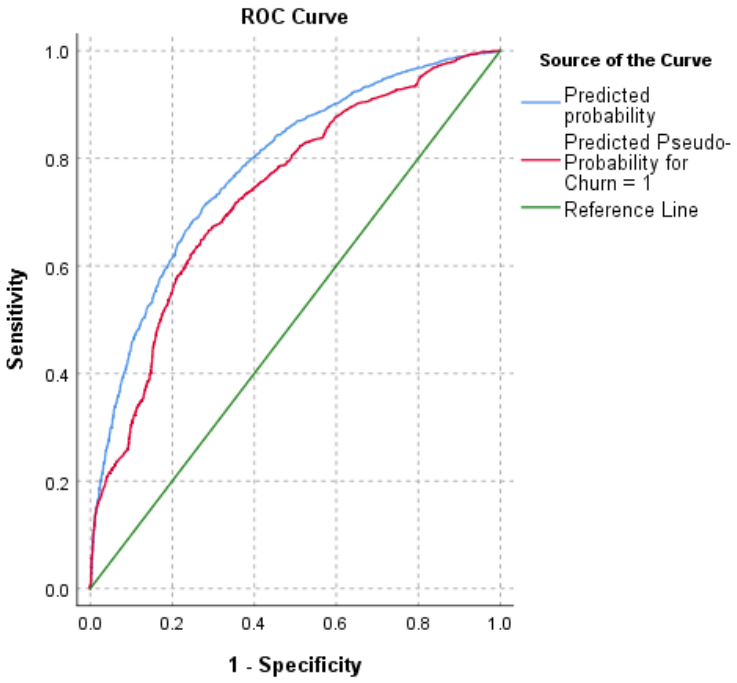


Figure 62: ROC Curve

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
Predicted probability	.783	.006	.000	.772	.795
Predicted Pseudo-Probability for Churn = 1	.737	.006	.000	.725	.749

The test result variable(s): Predicted probability, Predicted Pseudo-Probability for Churn = 1 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

Null hypothesis: true area = 0.5

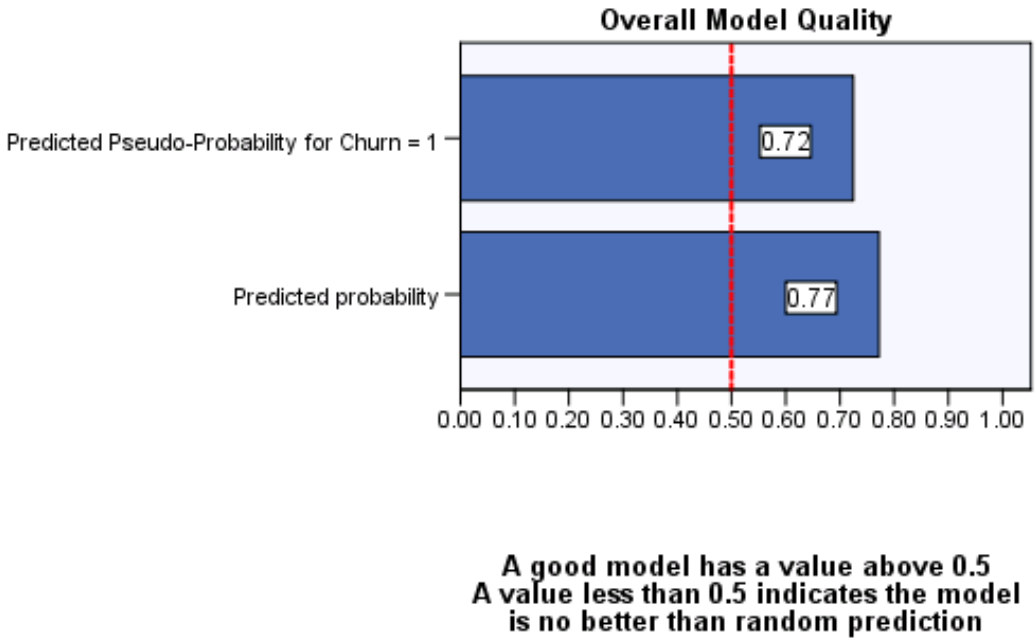
Figure 63: Area Under the ROC Curve

Test Result Pair(s)	z	Asymptotic Sig. (2-tail)[a]	AUC Difference	Std. Error Difference[b]	Asymptotic 95% Confidence Interval	
					Lower Bound	Upper Bound
PRE_1 - RBF_PseudoProbability_2	14.432	.000	.047	.109	.040	.053

a. Null hypothesis: true area difference = 0

b. Under the nonparametric assumption

Figure 64: Paired-Sample Area Difference Under the ROC Curves



Note: Use caution in interpreting this chart since it only reflects a general measure of overall model quality. The model quality can be considered "good" even if the correct prediction rate for positive responses does not meet the specified minimum probability. Use the classification table to examine correct prediction rates.

Figure 65: Overall Model quality

It is a numerical summary of the ROC Curve and the values in the table represent the probability of each category that the predicted probability of being in that category us higher for a randomly chosen case in that category than for a randomly chosen case not in that category.

Figure present ROC curves for the two models, for both values of the dependent variable. All points fall on the two jagged curves above the diagonal, indicating good classification.

For **Logistic Regression** the Area under the curve based on both Training and Testing Sample was **0.783** indicating a quite high classification accuracy rate, since exceeds 70% for both samples.

For **Radial Basic Function**, The Area under the curves based on both Training and Testing Sample was **0.737** indicating a quite high classification accuracy rate, since exceeds 70% for both samples.

Comparing the two AUCs we can see that Logistic Regression has a higher classification accuracy rate than Radial Basic Function.

CHAPTER 7

Summary and Conclusions

This thesis applied the two basic classification techniques namely artificial neural networks – specific Radial Basic Function and Logistic Regression, in order to predict which bank customers will churn or not.

The overall predictive accuracy was between 73, 7% and 78, 3 %.

The companies would be more interested in the prediction accuracies of those who are likely to leave the company. Radial Basic Function appears to higher predictive accuracy at 8 percent in comparison with logistic regression.

The identification of the ***four variables*** Age of the customer, from where the customer comes from, i.e., Geography, Gender of the customer, Number of Products and if the customer is an Active customer of the bank and has not left the last 6 months to churn is significant from a research perspective. The role of age as a variable is particularly significant.

While these predictive accuracies are specific to the data used in the analysis, the study has shown that it is possible to predict the customer churn, and identify those who are likely to leave the bank even before they had made their final decision to leave.

Such predictive abilities could help the bank to take proactive measures to minimize the attrition.

It is important for the bank to try out different models and techniques and identify important variables before finalizing on a specific technique or model.

Finally, understanding your customers' needs, preferences, sentiments, behavior and propensity to switch are the keys to use for predicting their loyalty.

Banks have a distinct competitive edge with the ability to anticipate and avoid churn, drive cross-selling, and generate customer loyalty thanks to customer intelligence

management focused on deep business process expertise, as well as the use of Big Data with Predictive Analytics and advanced machine learning.

Research Limitations and future work

A detailed study of state-of-the-art approaches in predicting customer churn has been performed for this research but the study had some limitations. The first limitation of the study was that the study dataset was only a fictitious dataset from a public data repository site, which may have been collected from only one bank within a short time period. In this case, the dataset may not apply to other banks, so generalizing the results to other banks should be done with extreme caution. In the future, more longitudinal studies are needed to test the reproducibility of the results, with more data samples collected over a long time from different banks, in order to use the findings to the banking industry in general.

Second limitation is that the study dataset was unbalanced in distribution between the two classes of churn. (Churned customers = 2037, no churned= 7963).

Although the stratified cross-validation method was used to ensure a representation of each category, this could have affected the prediction accuracy of the machine learning classifiers. It is, however, worthy of note that within the context of these limitations, the study achieved its goals.

Finally, this thesis focused solely on developing a neural network model specifically Radial Basic Function and on the other hand a Logistic Regression model approach to estimate churn.

Hence overlooking the potential benefit of considering and comparing alternative approaches for models such as Support Vector Machines or Decision Trees.

REFERENCES

1. Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert systems with applications, 38(3), 2354-2364
2. H. V. Jagadish University of Michigan, Jagadish, H., Michigan, U., University, J., Gehrke, J., Authors: H. V. Jagadish University of Michigan. (2014, July 01). Big data and its technical challenges. Retrieved September 17, 2020, from https://dl.acm.org/doi/fullHtml/10.1145/2611567
3. Grazia Dicuonzo & Graziana Galeone & Erika Zappimbulso & Vittorio Dell'Atti, 2019. "Risk Management 4.0: The Role of Big Data Analytics in the Bank Sector," International Journal of Economics and Financial Issues, Econjournals, vol. 9(6), pages 40-47.
4. Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. Expert Systems with Applications, 38(12), 15273-15285
5. Chen, Z.-Y., Fan, Z.-P. and Sun, M. (2012) 'Decision support A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data', European Journal of Operational Research, 223, pp. 461–472. https://doi.org/10.1016/j.ejor.2012.06.040
6. Zorn, S., Jarvis, W. and Bellman, S. (2010), "Attitudinal perspectives for predicting churn", Journal of Research in Interactive Marketing, Vol. 4 No. 2, pp. 157-169. https://doi.org/10.1108/17505931011051687
7. Coussement, K., Lessmann, S. and Verstraeten, G. (2017) 'A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry', Decision Support Systems, 95, pp. 27–36. doi: 10.1016/j.dss.2016.11.007.
8. Coussement, K. and Van den Poel, D. (2008) 'Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques', Expert Systems with Applications, 34(1), pp. 313–327. doi: 10.1016/j.eswa.2006.09.038.

9. Louridas, P. and Ebert, C. (2016) 'Machine Learning', IEEE Software, 33(5), pp. 110–115. doi: 68 10.1109/MS.2016.114.
10. Roos, I. and Gustafsson, A. (2007) 'Understanding frequent switching patterns', Journal of Service Research, 10(1), pp. 93–107. doi: 10.1177/1094670507303232.
11. Mahajan, V., Misra, R. and Mahajan, R. (2015) 'Review of data mining techniques for churn prediction in telecom', Journal of Information and Organizational Sciences, 39(2), pp. 183– 197.
12. Zoric, B. Predicting customer churn in banking industry using neural networks. Interdiscip. Descr. Complex Syst. 2016, 14, 116–124. DOI 10.7906/indecs.14.2.1
13. James G., Witten D., Hastie T., Tibshirani R. (2013) Classification. In: An Introduction to Statistical Learning. Springer Texts in Statistics, vol 103. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-7138-7_4
14. Freedman, D. (2009). Statistical Models: Theory and Practice (2nd ed.). Cambridge: Cambridge University Press https://doi.org/10.1017/CBO9780511815867
15. Guangli Nie, Wei Rowe, Lingling Zhang, Yingjie Tian, Yong Shi, Credit card churn forecasting by logistic regression and decision tree, Expert Systems with Applications, Volume 38, Issue 12,2011,Pages 15273-15285, https://doi.org/10.1016/j.eswa.2011.06.028.
16. Wouter Verbeke, Karel Dejaeger, David Martens, Joon Hur, and Bart Baesens. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, 218(1):211–229, 2012.
17. Jonathan Burez and Dirk Van den Poel. Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3):4626–4636, 2009.
18. Pampel, F.C. (2000). Logistic regression: A Primer. Thousand Oaks, CA: Sage.
19. Heck, R.H., Thomas, S.L., & Tabata, L.N. (2012). *Multilevel modeling of categorical outcomes using IBM SPSS*. New York: Routledge.
20. R. Kohavi. A study of cross validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th international joint conference on Artificial Intelligence, 2:1137–1143, 1995.
21. Metz, C. E.: Basic Principles of ROC Analysis. Sem Nuc Med, (1978) 283-298

22. Farid Shirazi, Mahbobeh Mohammadi,A big data analytics model for customer churn prediction in the retiree segment, International Journal of Information Management, Volume 48,2019,Pages 238-253,ISSN 0268-4012, https://doi.org/10.1016/j.ijinfomgt.2018.10.005.
23. Christopher M. Bishop. 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.
24. Zhang, T., Ramakrishnon, R., and Livny, M. (1996). BIRCH: An efficient data clustering method for very large databases. Proceedings of the ACM SIGMOD Conference on Management of Data, pp. 103–114. Montreal, Canada
25. Integral Solutions Limited (2007). Clementine® 12.0, Algorithms Guide. Chapter 10.
26. James, Gareth. Witten, Daniela. Hastie, Trevor. Tibshirani, Robert. 2017. An Introduction to Statistical Learning. Springer. New York.
27. https://www.financemagnates.com/thought-leadership/banks-are-leveraging-big-data-and-ai-toadvance-the-industry-heres-how/
28. https://www.icarvision.com/en/how-does-big-data-help-with-financial-risk-management- OUC 62 DOCUMENT 3
29. http://cab-inc.com/wp-content/uploads/2019/11/A-Case-Study-on-Big-Banks-Integrating-BigData-For-Success-Julian-Sevillano-Promontory-an-IBM-Company.pdf
30. https://www.upgrad.com/blog/data-science-use-cases-finance-industry/
31. Hasan, M.M., Popp, J. & Oláh, J. Current landscape and influence of big data on finance. J Big Data 8, 21 (2020). https://doi.org/10.1186/s40537-020-00291-z
32. Liu, A., Ghosh, J., and Martin, C. E. (2007). Generative oversampling for mining imbalanced datasets. In DMIN, pages 66–72.
33. https://www.datameer.com/wp-content/uploads/2018/02/big-data-use-case-financialservices.pdf
34. https://www.finextra.com/blogposting/17847/big-data-in-the-financial-services-industry---fromdata-to-insights
35. https://link.springer.com/chapter/10.1007/978-3-319-21569-3_12
36. Technavio. (2013). Global big data market in the financial services sector 2012-2016. Retrieved from http://www.technavio.com.
37. Horcher, K. A. (2005). *Essentials of financial risk management*. Hoboken, NJ: Wiley.

38. Hale G, Lopez JA. Monitoring banking system connectedness with big data. J Econ. 2019;212(1):203–20. https://doi.org/10.1016/j.jeconom.2019.04.027.
39. link.springer.com › article Current landscape and influence of big data on finance | SpringerLink
40. https://www.google.com/url?sa=t&source=web&rct=j&url=https://link.springer.com/article/10.1186/s40537-020-00291-z&ved=2ahUKEwiEl_uJsLXuAhUd8uAKHRcwDtkQFjAEegQIDRAC&usg=AOvVaw1U7mBWtT3kd4KdwPiXDxMq
41. https://www.weforum.org/agenda/2015/02/a-brief-history-of-big-data-everyone-should-read
42. https://www.firmex.com/resources/blog/big-data-3-open-source-tools-to-know/
43. https://www.firmex.com/resources/blog/7-big-data-techniques-that-create-business-value/
44. https://www.firmex.com/resources/uncategorized/does-big-data-make-sense-for-your-business/
45. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0206-3
46. Cimaglobal. Using big data to reduce uncertainty in decision making. 2015. http://www.cimaglobal.com/Pages-that-we-will-need-to-bring-back/velocity-archive/Student-e-magazine/Velocity-December-2015/P2-using-big-data-to-reduce-uncertainty-in-decision-making/.
47. IBM big data and analytics hub. Extracting Business Value from the 4 V's of Big Data. 2016. http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data
48. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0206-3
49. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0030-3
50. https://doi.org/10.1016/j.jbusres.2016.08.001
51. Big data stream analysis: a systematic literature review
52. https://thesai.org/Downloads/Volume7No2/Paper_67_A_Survey_on_Big_Data_Analytics_Challenges.pdf

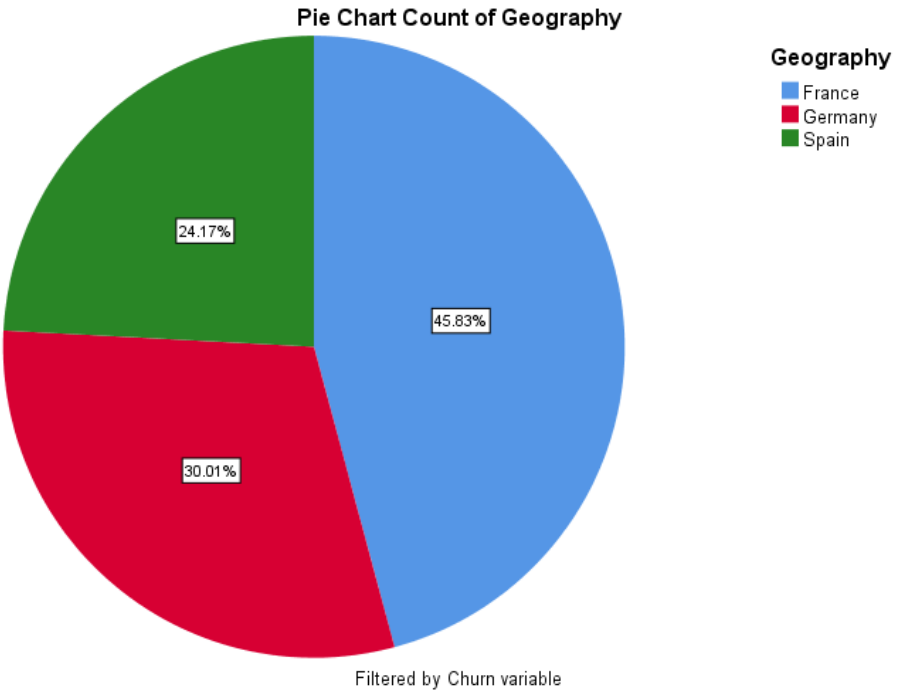
53. https://thesai.org/
54. https://www.gartner.com/en/information-technology/glossary/big-data
55. Basu, Atanu (December 2013). "How Data Analytics Can Help Frackers Find Oil". Datanami
56. "Predictive Analytics". *IBM. 2018. Retrieved 21 January* 2021.
57. 43 Examples of Analytical Skills for Greater Success". *Mindmonia*. 2019-03-01. Retrieved *29 January 2021*.
58. Jianqing Fan, Fang Han, Han Liu, Challenges of Big Data analysis, ***National Science Review***, Volume 1, Issue 2, June 2014, Pages 293–314, https://doi.org/10.1093/nsr/nwt032
59. https://www.guru99.com/big-data-tools.html
60. Beyer, M.A., Laney, D.: The importance of big data: A definition. Gartner (2012).
61. https://www.progress.com/docs/default-source/default-document-library/Progress/Documents/Papers/Addressing-Five-Emerging-Challenges-of-Big-Data.pdf
62. https://www.stat.purdue.edu/~doerge/BIOINFORM.D/SPRING16/KatalWazidGoudar_2013.pdf
63. J. Marvis, "Agencies rally to tackle big data," Science, vol. 336, no. 6077, p. 22, 2012.
64. Intel, "Big Data Analytics," 2012, http://www.intel.com/content/dam/www/public/us/en/documents/reports/data-insights-peer-research-report.pdf.
65. P. Russom, "Big data analytics," TDWI Best Practices Report, Fourth Quarter, 2011.https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf
66. Hale G, Lopez JA. Monitoring banking system connectedness with big data. J Econ. 2019;212(1):203–20. https://doi.org/10.1016/j.jeconom.2019.04.027.
67. link.springer.com › article Current landscape and influence of big data on finance | SpringerLink
68. https://www.google.com/url?sa=t&source=web&rct=j&url=https://link.springer.com/article/10.1186/s40537-020-00291-z&ved=2ahUKEwiEl_uJsLXuAhUd8uAKHRcwDtkQFjAEegQIDRAC&usg=AOvVaw1U7mBWtT3kd4KdwPiXDxMq

69. https://www.weforum.org/agenda/2015/02/a-brief-history-of-big-data-everyone-should-read/
70. https://www.firmex.com/resources/blog/big-data-3-open-source-tools-to-know/
71. https://www.firmex.com/resources/blog/7-big-data-techniques-that-create-business-value/
72. https://www.firmex.com/resources/uncategorized/does-big-data-make-sense-for-your-business/
73. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0206-3
74. Cimaglobal. Using big data to reduce uncertainty in decision making. 2015. http://www.cimaglobal.com/Pages-that-we-will-need-to-bring-back/velocity-archive/Student-e-magazine/Velocity-December-2015/P2-using-big-data-to-reduce-uncertainty-in-decision-making/.
75. IBM big data and analytics hub. Extracting Business Value from the 4 V's of Big Data. 2016. http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data
76. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0206-3
77. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0030-3
78. https://doi.org/10.1016/j.jbusres.2016.08.001
79. Big data stream analysis: a systematic literature review
80. https://thesai.org/Downloads/Volume7No2/Paper_67-A_Survey_on_Big_Data_Analytics_Challenges.pdf

Appendix

Geography

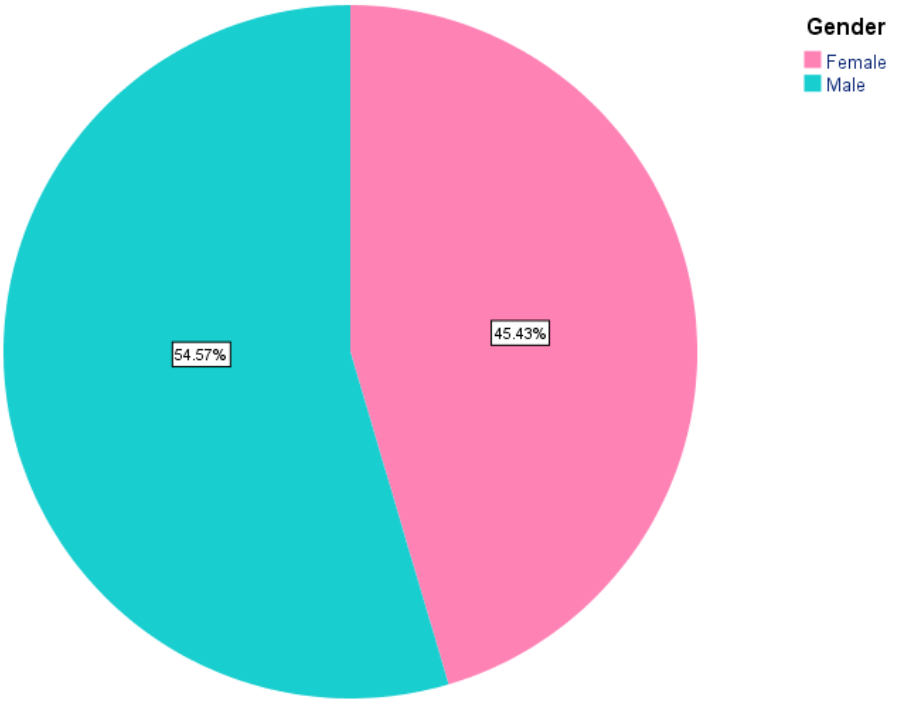
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	France	5014	50.1	50.1	50.1
	Germany	2509	25.1	25.1	75.2
	Spain	2477	24.8	24.8	100.0
	Total	10000	100.0	100.0	



Pie chart for Geography

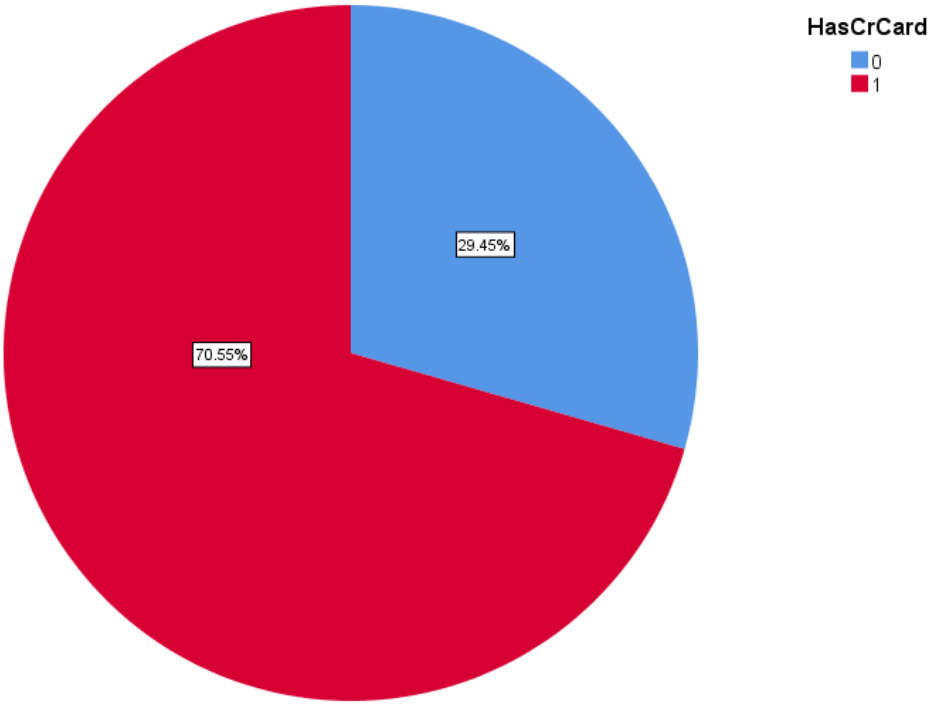
Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	4543	45.4	45.4	45.4
	Male	5457	54.6	54.6	100.0
	Total	10000	100.0	100.0	



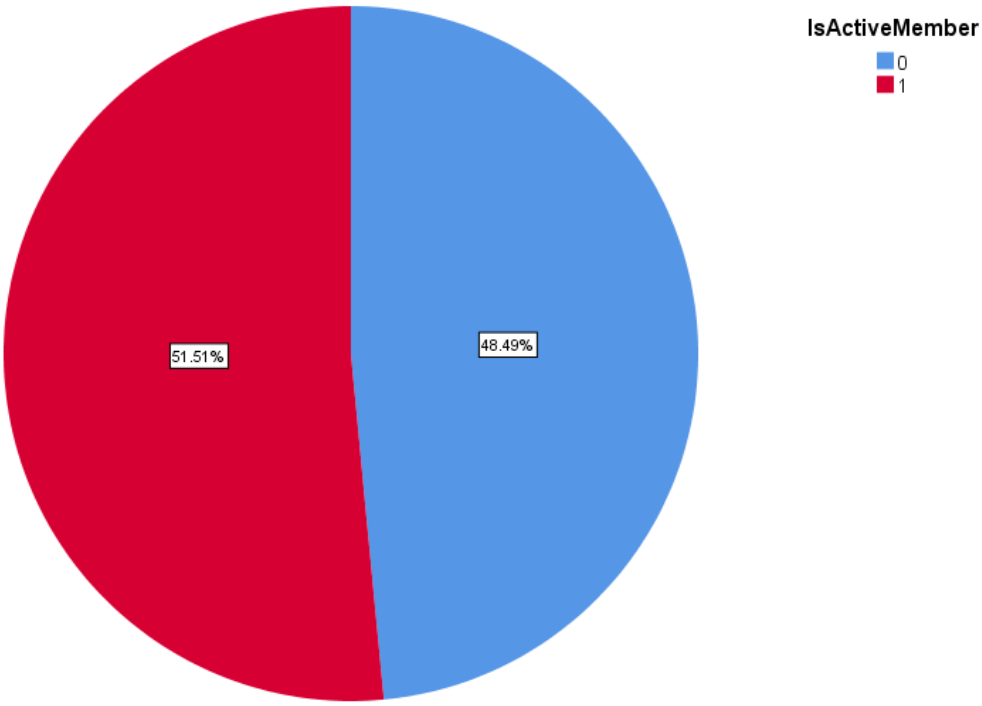
HasCrCard

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	2945	29.5	29.5	29.5
	1	7055	70.6	70.6	100.0
	Total	10000	100.0	100.0	



IsActiveMember

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	4849	48.5	48.5	48.5
	1	5151	51.5	51.5	100.0
	Total	10000	100.0	100.0	



Radial Basis Function

*Radial Basis Function Network.

RBF Churn (MLEVEL=N) BY Geography_regrouped NumofProducts_regrouped Gender IsActiveMember

WITH Age

/RESCALE COVARIATE=ADJNORMALIZED

/PARTITION TRAINING=7 TESTING=3 HOLDOUT=0

/ARCHITECTURE MINUNITS=AUTO MAXUNITS=AUTO HIDDENFUNCTION=NRBF

/CRITERIA OVERLAP=AUTO

/PRINT CPS NETWORKINFO SUMMARY CLASSIFICATION

/PLOT NETWORK ROC GAIN LIFT PREDICTED

/SAVE PREDVAL PSEUDOPROB

/MISSING USERMISSING=EXCLUDE

Radial Basis Function

Notes

Output Created		10-MAY-2021 18:01:01
Comments		
Input	Data	C:\Users\000100001759\Desktop\datasetfinal.sav
	Active Dataset	DataSet1
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	10000
Missing Value Handling	Definition of Missing	User- and system-missing values are treated as missing.
	Cases Used	Statistics are based on cases with valid data for all variables used by the procedure.
Weight Handling		not applicable

Syntax		RBF Churn (MLEVEL=N) BY Geography_regrouped NumofProducts_regrouped Gender IsActiveMember WITH Age /RESCALE COVARIATE=ADJNORMALIZED /PARTITION TRAINING=7 TESTING=3 HOLDOUT=0 /ARCHITECTURE MINUNITS=AUTO MAXUNITS=AUTO HIDDENFUNCTION=NRBF /CRITERIA OVERLAP=AUTO /PRINT CPS NETWORKINFO SUMMARY CLASSIFICATION /PLOT NETWORK ROC GAIN LIFT PREDICTED /SAVE PREDVAL PSEUDOPROB /MISSING USERMISSING=EXCLUDE.
Resources	Processor Time	00:00:02.03
	Elapsed Time	00:00:01.90
Variables Created or Modified	Predicted Value	RBF_PredictedValue
	Predicted Pseudo-Probability	RBF_PseudoProbability_1

	RBF_PseudoProbability_2

Network Information

Input Layer	Factors	1	Geography_regrouped
		2	NumofProducts_regrouped
		3	Gender
		4	IsActiveMember
	Covariates	1	Age
	Number of Units		9
	Rescaling Method for Covariates		Adjusted normalized
Hidden Layer	Number of Units		4[a]
	Activation Function		Softmax
Output Layer	Dependent Variables	1	Churn
	Number of Units		2
	Activation Function		Identity
	Error Function		Sum of Squares

a. Determined by the testing data criterion: The "best" number of hidden units is the one that yields the smallest error in the testing data.

Logistic Regression

```
USE ALL.
do if $casenum=1.
compute #s_$_1=7026.
compute #s_$_2=10000.
end if.
do if  #s_$_2 > 0.
compute filter_$=uniform(1)* #s_$_2 < #s_$_1.
compute #s_$_1=#s_$_1 - filter_$.
compute #s_$_2=#s_$_2 - 1.
else.
compute filter_$=0.
end if.
VARIABLE LABELS filter_$ '7026 from the first 10000 cases (SAMPLE)'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
FILTER OFF.
USE ALL.
EXECUTE.
LOGISTIC REGRESSION VARIABLES Churn
  /SELECT=filter_$ EQ 1
  /METHOD=ENTER Geography_regrouped NumofProducts_regrouped IsActiveMember Age Gender
  /CONTRAST (Geography_regrouped) =Indicator (1)
  /CONTRAST (Gender)=Indicator (1)
  /CONTRAST (NumofProducts_regrouped) =Indicator (1)
  /CONTRAST (IsActiveMember)=Indicator (1)
```

/SAVE=PRED PGROUP

/CLASSPLOT

/PRINT=GOODFIT CI(95)

/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).

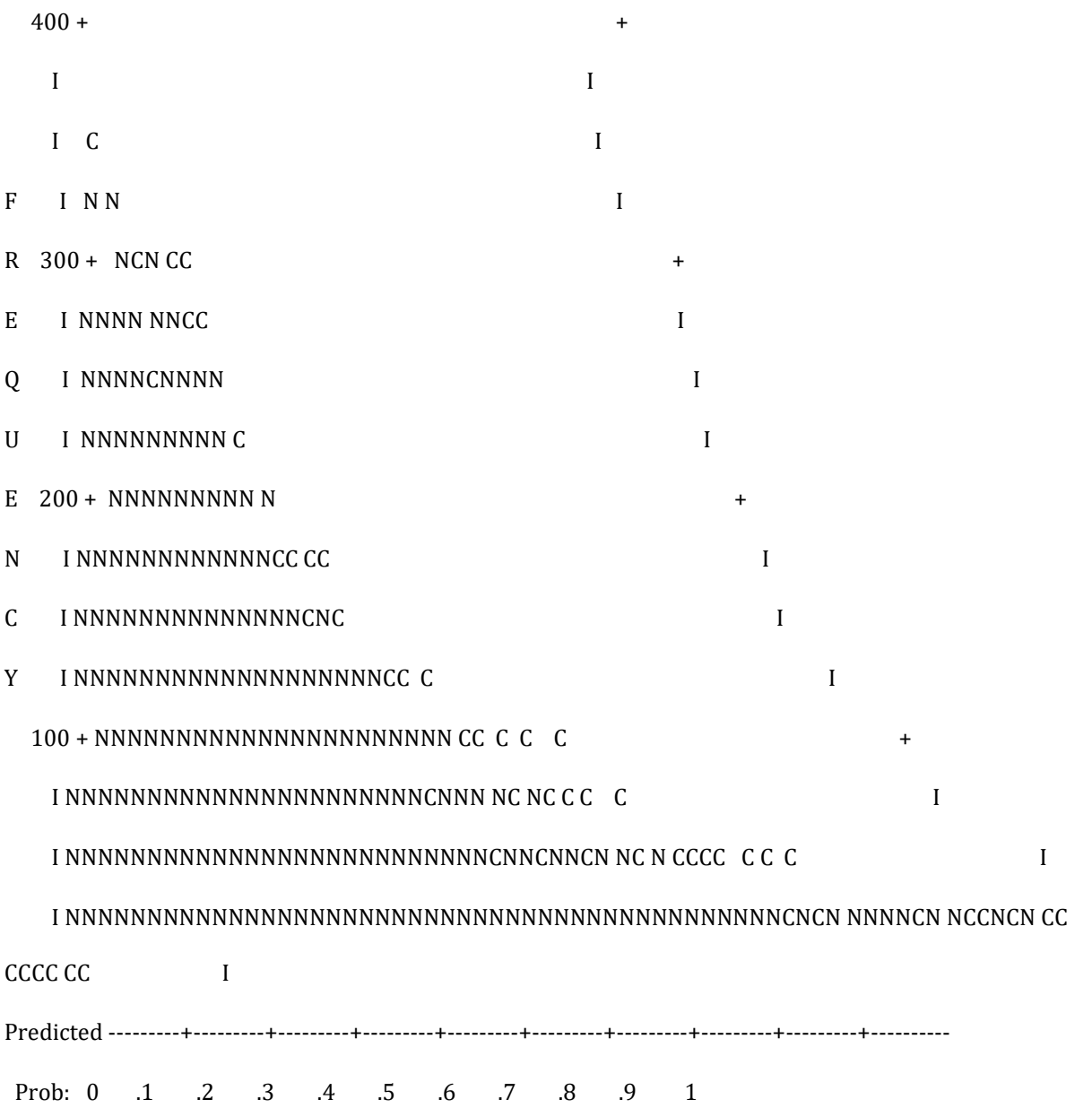
Notes

Output Created		10-MAY-2021 18:06:17
Comments		
Input	Data	C:\Users\000100001759\Desktop\datasetfinal.sav
	Active Dataset	DataSet1
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	10000
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing

Syntax		LOGISTIC REGRESSION VARIABLES Churn /SELECT=filter_$ EQ 1 /METHOD=ENTER Geography_regrouped NumofProducts_regrouped IsActiveMember Age Gender /CONTRAST (Geography_regrouped) =Indicator (1) /CONTRAST (Gender)=Indicator (1) /CONTRAST (NumofProducts_regrouped) =Indicator (1) /CONTRAST (IsActiveMember)=Indicator (1) /SAVE=PRED PGROUP /CLASSPLOT /PRINT=GOODFIT CI(95) /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
Resources	Processor Time	00:00:00.09
	Elapsed Time	00:00:00.09
Variables Created or Modified	PRE_1	Predicted probability
	PGR_1	Predicted group

Step number: 1

Observed Groups and Predicted Probabilities



Predicted Probability is of Membership for Churned

The Cut Value is .50

Symbols: N - No churned C – Churned

Each Symbol Represents 25 Cases.

Crosstabs

```
CROSSTABS
  /TABLES=Churn BY PGR_1
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT COLUMN
  /COUNT ROUND CELL
  /BARCHART.
```

Crosstabs details

Notes

Output Created		10-MAY-2021 18:11:43
Comments		
Input	Data	C:\Users\000100001759\Desktop\datasetfinal.sav
	Active Dataset	DataSet1
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	10000
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics for each table are based on all the cases with valid data in the specified range(s) for all variables in each table.
Syntax		CROSSTABS /TABLES=Churn BY PGR_1 /FORMAT=AVALUE TABLES /STATISTICS=CHISQ /CELLS=COUNT COLUMN /COUNT ROUND CELL /BARCHART.
Resources	Processor Time	00:00:00.42
	Elapsed Time	00:00:00.30
	Dimensions Requested	2
	Cells Available	524245

```
USE ALL.
COMPUTE filter_$=(filter_$=0).
VARIABLE LABELS filter_$ 'filter_$=0 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
CROSSTABS
  /TABLES=Churn BY PGR_1
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT COLUMN
  /COUNT ROUND CELL
  /BARCHART.
```

Crosstabs

Notes

Output Created		10-MAY-2021 18:14:11
Comments		
Input	Data	C:\Users\000100001759\Desktop\datasetfinal.sav
	Active Dataset	DataSet1
	Filter	filter_$=0 (FILTER)
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	2974
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics for each table are based on all the cases with valid data in the specified range(s) for all variables in each table.
Syntax		CROSSTABS /TABLES=Churn BY PGR_1 /FORMAT=AVALUE TABLES /STATISTICS=CHISQ /CELLS=COUNT COLUMN /COUNT ROUND CELL /BARCHART.
Resources	Processor Time	00:00:00.25
	Elapsed Time	00:00:00.23
	Dimensions Requested	2
	Cells Available	524245

ROC Analysis

Paired-sample design compares two ROC curves in a paired-sample scenario

Notes

Output Created		10-MAY-2021 21:36:02
Comments		
Input	Data	C:\Users\000100001759\Desktop\datasetfinal.sav
	Active Dataset	DataSet1
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	10000
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Only cases with valid data for all analysis variables are used in computing any statistics.
Weight Handling		not applicable

Syntax		ROC ANALYSIS PRE_1 RBF_PseudoProbability_2 BY Churn (1) /MISSING USERMISSING=EXCLUDE /CRITERIA CUTOFF=INCLUDE TESTPOS=LARGE CI=95 /DESIGN PAIR=TRUE /PLOT CURVE=ROC(REFERENCE) MODELQUALITY=TRUE /PRINT SE=TRUE COORDINATES=ROC.
Resources	Processor Time	00:00:01.26
	Elapsed Time	00:00:00.99

Chara A. Gavrielidou

"Big Data Analytics in Banks: Comparison of Classification Models in predicting customers churn"

May 2021