

Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of Statistics



Master's Thesis

**Data mining methods and their application in solving
credit card customer churn situations**

Bc. YuFeng Gao

© 2024 CZU Prague

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

DIPLOMA THESIS ASSIGNMENT

Bc. YuFeng Gao

Economics and Management

Thesis title

Data mining methods and their application in solving credit card customer churn situations

Objectives of thesis

The main objective of this thesis is to perform in-depth analysis and information mining of big data to determine the best model for predicting customer churn. Also, this thesis has partial objectives, such as

- Clarifying the statistical challenges of data mining for predicting customer churn in the context of big data.

- Identify the determinants that influence credit card customer churn behavior.
- Identify the main factors influencing customer churn based on different variables.
- Identify possible suitable models to predict credit card customer churn, obtain the best model and validate it with test set data.

Methodology

To achieve these objectives, the theoretical part of this thesis involves a literature review of Big Data, its characteristics, data mining methods and customer churn. A part of the literature review will also describe data mining methods and statistical methods and techniques that will be applied to the data, such as linear regression, binary logistic regression and decision tree models.

Practical work will include the creation of appropriate classification algorithms and the use of analytical tools to identify the factors that influence credit card customer churn. Based on these factors, binary logistic regression models and decision tree models will be used to construct predictive models to predict customer churn behavior. Finally, the analytical tools and the application of test set data are used to validate the predictive power of the model.

The end result is a predictive model with high classification accuracy that can provide a reliable reference for predicting credit card customer churn.

The proposed extent of the thesis

60 – 80 pages

Keywords

Big data, statistical analysis, predictive models, binary logistic regression, decision trees, customer churn.

Recommended information sources

- ABBOTT, D. Applied Predictive Analytics : Principles and Techniques for the Professional Data Analyst. Praha: John Wiley & Sons, Incorporated, 2014. ISBN 9781118727935.
- AGRESTI, A. Categorical data analysis. Hoboken: John Wiley & Sons, 2013. ISBN 978-0-470-46363-5.
- AU W.-H., Chan C.C., Yao X. A novel evolutionary data mining algorithm with applications to churn prediction. IEEE Transactions on Evolutionary Computation, 2003. ISSN:1089-778X.
- AZZALINI, A., SCARPA, B. Data Analysis and Data Mining : An Introduction. Oxford University Press, USA, March 2012. ISBN 9780199909285
- BERRY, M. J., & LINOFF, G. S. Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons, 2004. ISBN: 978-0-471-47064-9
- HUANG B, KECHADI M T, BUCKLEY B. Customer churn prediction in telecommunications. Expert Systems with Applications, 2012. 39(1): 1414-1425.
- LEMESHOW, S. et al. Applied Logistic Regression. Hoboken: John Wiley & Sons, 2013. ISBN 978-0-470-58247-3.
- OTT, Lyman; LONGNECKER, Michael. *An introduction to statistical methods & data analysis*. Australia: Cengage Learning, 2016. ISBN 9781305269477.
- SIEGEL, E. Predictive Analytics. Hoboken: John Wiley & Sons, 2013. ISBN 978-1-118-35685-2
- TUFFERY, S. Data Mining and Statistics for Decision Making. UK, West Sussex: Wiley, 2011. ISBN 978-0-470-68829-8.
-

Expected date of thesis defence

2023/24 SS – PEF

The Diploma Thesis Supervisor

Ing. Tomáš Hlavsa, Ph.D.

Supervising department

Department of Statistics

Electronic approval: 17. 5. 2023

Ing. Tomáš Hlavsa, Ph.D.

Head of department

Electronic approval: 3. 11. 2023

doc. Ing. Tomáš Šubrt, Ph.D.

Dean

Prague on 14. 03. 2024

Declaration

I declare that I have worked on my master's thesis titled " Data mining methods and their application in solving credit card customer churn situations" by myself and I have used only the sources mentioned at the end of the thesis. As the author of the master's thesis, I declare that the thesis does not break any copyrights.

In Prague on 08.03.2024

高城峰

Acknowledgement

I would like to thank Ing. Tomáš Hlavsa, Ph.D. for his guidance during my work on this thesis. I would also like to thank my parents for their support and trust in me all the time.

Data mining methods and their application in solving credit card customer churn situations

Abstract

The theoretical part of the paper first we start with the definition of big data, we introduce the basic concepts and principles of data mining algorithmic models in big data, then we focus on the development of statistical data processing and the basic methods of statistical and predictive analysis. After that we introduce three different data mining models used to process big data: logistic regression model, decision tree CHAID model and decision tree C&D model. they are after that we will use to predict and determine credit card customer churn. At the end of the theoretical part, we briefly introduce the development and use of credit cards and the general definition of customer churn behavior, and we provide a preliminary analysis of the comprehensive factors influencing customer churn behavior. In the practical part of the work, we deal with a large dataset of 10,127 objects, a collection containing information related to credit card customers, which is placed on the Kaggle online platform for educational purposes.

The practical work involved creating appropriate classification algorithms and using analytical tools to identify the factors that influence credit card customer churn. Based on these factors, logistic regression models and two different decision tree models were constructed to predict customer churning behavior. Finally, data analysis tools and testing of test datasets were used to validate the predictive power of the models.

Three data prediction models were built using logistic regression and decision tree algorithms encapsulated in IBM SPSS to build data models based on a dataset of credit card customer churn. Their prediction quality was checked by using confusion metrics and the area under the receiver operating characteristic (ROC) curve (AUC).

By comparing the three prediction models, a data model with 91% prediction accuracy is established, which can provide a reliable reference for banks' credit card customer attrition prediction.

Keywords: Big data, statistical analysis, predictive models, binary logistic regression, decision trees, customer churn.

Metody dolování dat a jejich využití při řešení situace odlivu zákazníků kreditních karet

Abstrakt

V teoretické části článku nejprve začneme definicí velkých dat, představíme základní pojmy a principy algoritmičtých modelů dolování dat ve velkých datech, poté se zaměříme na vývoj statistického zpracování dat a základní metody statistické a prediktivní analýzy. Poté představíme tři různé data miningové modely používané ke zpracování velkých dat: logistický regresní model, model rozhodovacího stromu CHAID a model rozhodovacího stromu C&D. Ty poté použijeme k predikci a určení odchodu zákazníků kreditních karet. V závěru teoretické části stručně představíme vývoj a používání kreditních karet a obecnou definici chování zákazníků při odchodu a uvedeme předběžnou analýzu komplexních faktorů ovlivňujících chování zákazníků při odchodu. V praktické části práce se zabýváme rozsáhlým datovým souborem 10 127 objektů, souborem obsahujícím informace týkající se zákazníků kreditních karet, který je umístěn na online platformě Kaggle pro vzdělávací účely.

Praktická část práce zahrnovala vytvoření vhodných klasifikačních algoritmů a využití analytických nástrojů k identifikaci faktorů, které ovlivňují odchod zákazníků kreditních karet. Na základě těchto faktorů byly sestaveny logistické regresní modely a dva různé modely rozhodovacích stromů, které předpovídají chování zákazníků při odchodu z trhu. Nakonec byly použity nástroje pro analýzu dat a testování testovacích datových sad k ověření prediktivní síly modelů.

Byly sestaveny tři modely predikce dat s využitím algoritmů logistické regrese a rozhodovacího stromu zapouzdřených v programu IBM SPSS, aby bylo možné sestavit datové modely na základě souboru dat o odchodu zákazníků z kreditních karet. Jejich predikční kvalita byla ověřena pomocí metrik záměny a plochy pod křivkou ROC (receiver operating characteristic) (AUC).

Porovnáním tří predikčních modelů byl vytvořen datový model s 91% přesností predikce, který může poskytnout spolehlivou referenci pro predikci úbytku zákazníků kreditních karet bank.

Klíčová slova: Big data, statistická analýza, prediktivní modely, binární logistická regrese, rozhodovací stromy, odliv zákazníků.

Table of content

1. Introduction.....	1
2. Objectives and Methodology	3
2.1 Objectives	3
2.2 Methodology.....	4
3. Literature Review	8
3.1 CRM and Credit card services.....	8
3.1.1 Definition of CRM and how it works	8
3.1.2 Customer churn behavior and churn prediction.....	9
3.1.3 Credit card services.....	10
3.2 Factors affecting the churn prediction	11
3.2.1 Socio-demographic characteristics	12
3.2.2 Customer income level.....	13
3.2.3 Marital and family status.....	14
3.2.4 Service Quality & Customer Care	14
3.2.5 Credit card limits and types	15
3.2.6 Client trading volume and recent trading changes.....	16
3.3 Big data.....	16
3.3.1 Definition	17
3.3.2 Characteristics of Big data	17
3.4 Data mining and algorithmic models.....	18
3.4.1 Logistic regression	18
3.4.2 Linear regression.....	20
3.4.3 Decision Tree (CART and CHAID algorithm).....	22
3.4.4 Evaluation of model.....	25
4. Practical Part	28
4.1 Overview of the Case Study	28
4.2 Data pre-processing	29
4.2.1 Measurement types, labels	30
4.2.2 Missing values analysis.....	30
4.3 Explanatory Data Analysis	30
4.3.1 Descriptive statistics – Numerical variables	31
4.3.2 Univariate statistical analysis – Category variables.....	38
4.4 Data preparation.....	44
4.4.1 Categorization - reclassification.....	44
4.4.2 Partition.....	44
4.5 Modeling.....	45

4.5.1 Binary logistic regression model.....	45
4.5.2 Model Evaluation – Binary logistic regression	50
4.5.3 Decision trees models - CHAID tree models	53
4.5.4 CHAID tree model evaluation.....	56
4.5.5 Decision trees models - C&R tree model.....	57
4.5.6 C&R tree model evaluation.....	59
5. Model comparison and final evaluation- Testing.....	61
5.1 Confusion Matrix	61
5.1.1 Logistic regression testing.....	61
5.1.2 CHAID tree model testing.....	62
5.1.3 C&R tree model testing.....	62
5.2 AUC, Gini and ROC curve	63
6. Result and Discussion	66
7. Conclusion	69
8. References.....	71
9. List of Figure and tables.....	73
8.1 List of Figure.....	73
8.2 List of tables.....	74
Appendix.....	75

1. Introduction

In recent years, based on the rapid development of science and technology, some new applications of information technology have provided a highway for the rapid development of people's consumption methods and the development of the financial industry, and for the financial and monetary aspects, the emergence and use of credit cards can be said to have changed people's daily consumption methods and provided great convenience for the rapid development of the financial industry, so credit cards are accepted and widely used.

Of course with the development of the information age, new data is being generated and stored every moment. There is a geometric explosion in the generation of data and its storage. This growth also includes sources of data collection, such as credit card payments and repayments, the spending preferences of different customers and the storage of their personal information. Big Data is the term used to describe the vast amount of data that can impact our daily activities. With the explosion of data, the application of data mining and its algorithmic models is involved in how to get the quality information we need from this massive amount of data. It can be said that data mining offers us many possibilities, and its applications and uses are expanding over time. In the business world, many companies have been able to obtain a large amount of important information hidden in big data, which puts the company far ahead of the business competition and at a competitive advantage.

The focus of this paper is to use data mining algorithmic models in big data analytics to predict the occurrence of customer churn and to identify the key factors that influence customer churn and ultimately determine which of the predictive models used is the best model to predict credit card customer churn. To estimate the potential credit card customers that will be churned in advance for the company and to recover the potential lost customer base in time.

The main purpose of this paper is not to provide a new data mining algorithm, but to focus on the application of customer churn prediction, providing a framework for understanding the hidden pattern knowledge of cardholders using big data and data mining. The goal is the application of customer churn prediction from data preparation to useful knowledge.

In this paper, we discuss the application of data mining, including logistic regression and decision trees, in predicting credit card user churn. Banks can take appropriate measures to retain customers based on the model's recommendations.

The structure of this paper is as follows. Section 3 presents the definition of customer churn behavior and the various factors that contribute to credit card customer churn through a literature review. It also explains the basic understanding of the definitions of big data, data mining and data analytics and illustrates the challenges faced in data mining. Section 4 introduces EDA analysis, correlation analysis and influence factor importance analysis to filter out the key factors influencing credit card customer churn and enumerates the various algorithms introduced for data mining and data analytics. Section 5 details the data analysis algorithms used for predictive modeling and their performance on the dataset. Finally, Section 6 concludes the paper and discusses future work.

2. Objectives and Methodology

2.1 Objectives

The main objective of this thesis is to conduct an in-depth research analysis and information mining on big data of credit card user information to determine the best model for customer churn prediction. In addition, this thesis has partial objectives such as

- To clarify the statistical challenges of data mining for predicting customer churn in the context of big data.
- To identify the determinants that influence credit card customer churn behavior.
- Identify the main factors influencing customer churn based on different variables.
- Identify potentially suitable models to predict credit card customer churn, obtain the best model and validate it using test set data.

The focus of this paper is to use data mining algorithmic models in big data analytics to predict the occurrence of customer churn and to identify the key factors influencing churn and ultimately determine which of the predictive models used is the best model to predict credit card customer churn. To anticipate potential credit card customers that will be churned in advance for the company and to recover the potentially lost customer base in time.

The main objective of this paper is not to provide a new data mining algorithm, but to focus on the application of customer churn prediction by providing a framework to understand the hidden pattern knowledge of cardholders using big data and data mining. The goal is the application of customer churn prediction. The resulting predictive model has high classification accuracy and provides a reliable reference for predicting credit card churn. It can predict potential credit card customer churn in advance for banks and recover potential customer base loss in time. Predicting credit card user churn. Banks can take appropriate measures to retain customers based on the model's recommendations.

Hypothesis 1: Big data analytics has a positive impact in CRM.

Hypothesis 2: The logsitc regression model has high accuracy in predicting credit card customer churn behavior, comparing the other two decision tree models.

2.2 Methodology

The methodology of the study is to first conduct a preliminary analysis of the factors influencing credit card customer churn, and to filter out the influencing variables through analysis and literature review. Then data preprocessing is performed on the credit card customer churn prediction data obtained from Kaggle data mining website, where checking for missing data values and eliminating their impact on subsequent modelling is very important. The analytical tools used were: SPSS Modeler and SPSS Statistics. After data preprocessing, univariate statistical descriptive analyses were performed on each variable in the dataset to gain more insight into the proportional relationship between each variable and churned customers. The next step is to perform predictive analysis on the preprocessed dataset, in this case logistic regression model, decision tree CHAID model and decision tree C&R model will be used for data modelling and predictive analysis. The dataset for this example consists of 10127 observations with 1 target variable and 22 input variables.

- Literature Review on Influencing Factors

The first stage begins with a preliminary analysis of the influencing factors leading to credit card customer churn through a literature review to screen out the variables that have a greater impact on customer churn. The theoretical part of this paper also covers the literature review of big data and its characteristics, data mining methods and customer churn. The literature review section will also present data mining methods and statistical analysis techniques applied to customer data, including linear regression, binary logistic regression and decision tree modeling.

- Data preprocessing

The second stage was to perform initial data organisation on the original dataset and then obtain ordered, neat and high quality data. Running datasets and data processing with SPSS modeler and SPSS Statistic. The descriptive information of the original dataset is shown in Table 1 below, we checked the duplicates, missing data and had a basic statistical understanding of the data through EDA, after which some categorical variables with too large a gap in distribution were recoded to achieve data integration in order to improve the fit of the data model.

Nominal includes (2): “Gender”, “Marital_Status”.

Ordinal (3): “Education_Level”, “Income_Category”, “Card_Category”.

Continuous (14): “Age”, “Dependent_count”, “Months_on_book”,
 “Total_Relationship_Count”, “Months_Inactive_12_mon”, “Contacts_Count_12_mon”,
 “Credit_Limit”, “Total_Revolving_Bal”, “Avg_Open_To_Buy”,
 “Total_Amt_Chng_Q4_Q1”, “Total_Trans_Amt”, “Total_Trans_Ct”,
 “Total_Ct_Chng_Q4_Q1”, “Avg_Utilization_Ratio”.

Attribute Information	Data type	Description
Attrition_Flag	category	Flag indicating whether or not the customer has churned out.
Customer_Age	numerical	Age of customer.
Gender	category	Gender of customer.
Dependent_count	numerical	Number of dependents that customer has.
Education_Level	category	Education level of customer.
Marital_Status	category	Marital status of customer.
Income_Category	category	Income category of customer.
Card_Category	category	Type of card held by customer.
Months_on_book	numerical	How long customer has been on the books.
Total_Relationship_Count	numerical	Total number of relationships customer has with the credit card provider.
Months_Inactive_12_mon	numerical	Number of months customer has been inactive in the last twelve months.
Contacts_Count_12_mon	numerical	Number of contacts customer has had in the last twelve months.
Credit_Limit	numerical	Credit limit of customer.
Total_Revolving_Bal	numerical	Total revolving balance of customer.
Avg_Open_To_Buy	numerical	Average open to buy ratio of customer.
Total_Amt_Chng_Q4_Q1	numerical	Total amount changed from quarter 4 to quarter 1.
Total_Trans_Amt	numerical	Total transaction amount.
Total_Trans_Ct	numerical	Total transaction count.
Total_Ct_Chng_Q4_Q1	numerical	Total count changed from quarter 4 to quarter 1.
Avg_Utilization_Ratio	numerical	Average utilization ratio of customer.
Note: * see appendix for variable code value.		

Table 1 Category data, data type, Description
 Source: author’s own work

- Data analysis

In the third stage of data analysis, interpretive data analysis is used to analyze each of the integrated variables independently through univariate statistical analysis. Through the second stage, we further screened and integrated the data by eliminating unqualified variables.

- Preparation of modeling data – partition

In the fourth stage, to prepare the data for modelling, data partition is performed on the preprocessed dataset (70% training dataset and 30% test dataset). The eligible variables are selected to prepare the dataset for modelling.

- Data mining, modeling

The next phase is the modeling phase where the actual work will include creating appropriate classification algorithms and using analytical tools to identify the factors that influence credit card customer churn. Based on these factors, predictive models will be constructed using a binary logistic regression model and two different decision tree models to predict customer churn behavior. Predictive models are developed based on previously organized datasets. According to the literature review, three data mining analysis methods, logistic regression, decision tree CHAID and C&R models, meet the needs of this study, so we decided to use these three models to model the data.

Logistic regression is a classification algorithm that can be used to deal with binary and multivariate classification problems. It is a classical classification model commonly used in machine learning, data mining and data prediction. The coefficients derived from the logistic regression model are easy to understand and interpret. By using SPSS statistic to detect multicollinearity for all the variables in the 70% training data set, we found that none of the variables had a $VIF > 5$ by comparing the VIF values, thus eliminating the effect of multicollinearity on the modeling of the logistic regression model.

For logistic regression modeling the following data variables were used: Dependent_count, Total_Relationship_Count (Keaveney, 1995), Months_Inactive_12_mon (Anil Kumar & Ravi, 2008a), Contacts_Count_12_mon (Mahajan et al., 2017), Credit_Limit (Cronin Jr et al., 2000) and (Anil Kumar & Ravi, 2008a), Total_Revolving_Bal (Anil Kumar & Ravi, 2008a), Total_Trans_Amt (Reichheld, 1993), Total_Trans_Ct, Total_Ct_Chng_Q4_Q1 (Anil Kumar & Ravi, 2008a), Gender (Athanasopoulos, 2000), Income_Category (Cronin Jr et al., 2000) and Education_level_Re (Athanasopoulos, 2000). And we use significant level $\alpha=0.05$.

Decision tree model is a classical classification and regression algorithm in data mining and machine learning. We built decision tree CHAID and C&R models. The decision tree C&R method builds decision tree models of binary trees, this is because the CART algorithm is applicable to the problem of yes or no sample characteristics.

- Evaluation

The last phase is the evaluation, which consists of validating the predictive ability of the models using analytical tools and test set data, and evaluating the previously developed predictive models to determine if their predictive ability meets the research objectives. During the evaluation phase, feedback will be given as to whether the developed model meets the criteria and whether the predictive accuracy meets the criteria. To determine the best forecasting model, the following characteristics will be examined: receiver operating characteristic curve (ROC) and AUC curve.

The resulting predictive model has high classification accuracy and provides a reliable reference for predicting credit card churn. It can predict potential credit card customer churn in advance for banks and recover potential customer base loss in time. Predicting credit card customer churn. Banks can take appropriate measures to retain customers based on the model's recommendations.

3. Literature Review

Big data analysis and Data mining (DM) methodology has a tremendous contribution for researchers to extract the hidden knowledge and information which have been inherited in the data used by researchers.

3.1 CRM and Credit card services

3.1.1 Definition of CRM and how it works

Customer Relationship Management (CRM) technology solutions emerged in their infancy in the 1970s, beginning as a solution aimed explicitly and exclusively at automating sales people (Buttle & Maklan, 2019).

Since the early 1980s, the concept of customer relationship management (CRM) in marketing, which consists of the four dimensions of customer identification, customer attraction, customer retention and customer development, has gained prominence. It is difficult to find a fully recognized definition of CRM. we can describe it as the integrated strategy and process of acquiring, retaining and working with selective customers to create superior value for the company and its customers (Parvatiyar & Sheth, 2001)

Customer Relationship Management (CRM) is a combination of people, processes and technology designed to understand a company's customers. CRM is an integrated approach to relationship management that focuses on maintaining customer relationships and relationship development (Chen & Popovich, 2003).

Customer Relationship Management (CRM) has developed on the basis of advances in information technology and changes in the organization of customer-centric processes. Customer relationship management solutions initially consisted of three pillar modules: sales, marketing and service. These are also the three classic axes of global customer business management. Customer relationship management can lead to harvesting customer loyalty and long-term profitability for a company (Chen & Popovich, 2003)

CRM is a comprehensive business and marketing strategy that integrates skills, processes and all business activities around the customer (Anton, 1996).

CRM is "the key competitive strategy you need to stay focused on the needs of your customers and to integrate a customer-facing approach throughout your organisation" (Brown & Coopers, 1999).

With today's cost cutting and intense competitive pressures, more and more companies are focusing on Customer Relationship Management (CRM). The unknown future behavior of customers is very important to CRM (Nie et al., 2011).

Management and marketing services are trying to cope with the increasing competition in the industry by focusing their efforts on strong customer relationship management (CRM). In particular, customer retention is generating interest because it is clear that retained customers can help a company greatly by spreading positive word-of-mouth, a behavior that can subsequently reduce the marketing costs of new customer acquisition (Bolton & Bronkhorst, 1995). The cost of acquiring new customers can be much higher than the cost of retaining existing customers (Geiler et al., 2022). Moreover, in saturated markets, the cost of attracting new customers is 5-10 times higher than the cost of retaining existing customers (Jamalian & Foukerdi, 2018).

As a result, companies focus their marketing efforts on customer retention rather than customer acquisition. Building an effective and trustworthy customer churn prediction model is necessary for customer retention.

3.1.2 Customer churn behavior and churn prediction

Customer churn is the number of current clients who are potentially leaving the service provider within a given period of time. These customers may be referred to as attrition customers. The main objective of predicting churn analysis is to forecast potential churn as early as possible and to identify the root causes that lead to churn (Wagh et al., 2024)

(Berson & Thearling, 1999) identified customer churn in the telecommunications industry as the movement of existing customers from one service provider to another.

(Gürsoy, 2010) point out that if an existing customer ends his contract with one service provider and turns into a customer of another service supplier, the subscriber is said to be "churned" or "attrition". Churn can be active or on purpose, passive or involuntary, and rotational or casual.

(Kentrias, 2001) says that customer churn management in the telecommunication industry is an important process for retaining the company's regular customers. He also emphasized the importance of evaluating each customer's attitude in response to different specific offers and assessing which customers will be positively affected. (Berson & Thearling, 1999) explained that churn management is a process that involves constructing

churn prediction models using past churn data and identifying key factors that influence churn.

The main objective of Predictive Churn Analysis is to predict the potential customer churn as early as possible and to identify the root causes leading to churn. This will help companies to provide better services to their customers, satisfy their needs and change their customer loyalty so that they will continue to use the services.

Credit card customers are increasingly focused on the quality of service provided by their bank (Lin et al., 2011). When a customer has no particular loyalty to his or her bank, the customer moves from one bank to another when he or she perceives that other banks offer better quality of service or better interest rates. This results in a potential and objective loss of customers.

Many studies have applied customer churn analysis to a variety of areas, such as the credit card industry (Kim et al., 2005).

While the negative effects of customer churn are easily observed - lack of revenue or supplemental costs to attract new customers - the causes of customer churn continue to be researched, as they often vary by economic sector and customer group.

3.1.3 Credit card services

Before the advent of credit card services, banks only offered face-to-face services to their customers, which largely limited the upper limit of people's spending (Tudeal, 2022). After the emergence of electronic banking, new digital currency platforms were quickly added to almost all bank payment systems in order to maintain the competitive advantage of banks, and credit cards were accepted and widely used through the development and support of Internet technology.

Customers typically carry one or two bank credit cards, as well as other types of credit cards, so managing customer churn is a top priority for most banks in the credit card industry. Credit card services offered by commercial banks include account acquisition and activation, receivables financing, card authorization, private label credit card issuance, statement generation, remittance processing, customer service functions, and marketing services. Intense competition among banks to provide these services has led to a significant increase in service reliability and quality (Anil Kumar & Ravi, 2008a).

3.2 Factors affecting the churn prediction

(Keaveney, 1995) identifies eight major causal variables for customer churn, namely price, inconvenience, core service failure, service encounter failure, competitive issues, ethical issues, and involuntary factors.

Predictors of customer churn include comprehensive demographic details (such as age, gender, marital status and income category). as well as insights into each customer's relationship with their credit card provider (e.g., card type, months of card ownership, and period of inactivity.) It also includes maintaining key data on customer spending behaviour that more closely approximates the churn decision, (e.g. total revolving balance, credit limit, average open purchase rate, and analyzable metrics) (Anil Kumar & Ravi, 2008b).

The customer loyalty is described by the relationship between the customer and the enterprise, which in the case of the banking and credit card industry is reflected in the following variables: the type of card held by the customer, the duration of the customer's stay on the books, the total number of relationships the customer has with the credit card provider (number of credit cards held), the frequency of purchases made by the customer with the credit card per year, the frequency of contact between the customer and the credit card institution per year, the number of months that the customer has been active in a year, the credit limit and total amount of purchases made by the customer. These are important factors in predicting credit card churn.

(Reichheld, 1993) indicated that the number of customers is critical to business performance. Therefore, retaining and creating loyal customers should be part of a firm's strategy. The authors noted that if firms focus on retaining repeat customers, they can save on the cost of acquiring new customers, and repeat customers can sometimes lead to new customers. Studies have also found that long-term repeat customers are less price sensitive, resulting in higher profits for the company. By reducing attrition by as little as 5 percent, a company's profits could be doubled. The economic benefits of high customer loyalty are significant for companies. Customer loyalty can be built through the value a company provides through its services or products, and the feedback that customers receive from the personal feeling of using a service is much more convincing than advertisements, telemarketing calls, and other marketing campaigns. This is what makes the difference between companies.

(Mahajan et al., 2017) concluded that service quality is positively correlated with customer loyalty. The quality of service is reflected in how many times a business contacts its customers. Good service quality will effectively contribute to the number of times customers purchase and the total amount of money spent. The frequency of customer usage can also reflect the service quality of the company side by side.

3.2.1 Socio-demographic characteristics

A study by (Athanasopoulos, 2000) identified that customer retention is dependent on customer satisfaction and the socio-demographic characteristics of the customer (age, gender, marital status, and education level).

The impact of customers' socio-demographic characteristics (e.g., age, gender, marital status, and education level) on credit card customer churn is a complex issue that touches on a variety of aspects such as consumer behavior, financial service preferences, and customer loyalty.

- **Age:** Younger customers may be more likely to churn because they may be looking for more innovative or technologically advanced banking products. Older customers, on the other hand, may show greater loyalty because they have developed a long-term relationship with the bank.
- **Gender:** Research suggests that gender may influence an individual's financial management habits and product preferences, but its direct impact on customer churn is likely to be small. However, gender differences may affect customer service preferences, which in turn may indirectly affect customer loyalty.
- **Marital status:** Married customers may have more stable financial needs, such as home purchase and education loans, which may make them more inclined to maintain a long-term relationship with the bank. Single or divorced customers may change their banking services more frequently in search of better deals or services.
- **Education level:** Customers with higher levels of education are likely to have a better understanding of financial products and therefore may be more inclined to look for products and services that best suit their needs. This means that they may be more likely to churn if their current bank does not offer services that meet their expectations.

3.2.2 Customer income level

(Cronin Jr et al., 2000) identified a significant positive relationship between service quality and customer loyalty. Customer income level, good service quality and brand image make customers brand conscious. Customers who continue to use their existing service providers are mainly concerned with switching costs.

A customer's income has a significant impact on credit card churn, primarily because income levels are directly linked to a customer's ability to pay, credit risk assessment, and demand for financial products and services.

- Higher income levels generally imply greater ability to pay, which makes customers more likely to pay their bills on time, thereby reducing the risk of churn due to late payments or defaults. Conversely, customers with lower incomes may have more difficulty managing their finances, increasing the likelihood of churn.
- Banks and credit card companies often use income as a key indicator in assessing credit risk. Higher-income customers may be perceived as low risk and therefore more likely to receive more favorable credit card terms, such as lower interest rates or higher credit limits, which can increase customer satisfaction and loyalty and reduce churn. Customers with lower incomes, on the other hand, may be offered less favorable terms, increasing the likelihood that they will switch to other credit card providers.
- Customers' income levels affect their demand for various financial products and services. For example, customers with higher incomes may be more interested in premium credit cards, which offer additional rewards, benefits and services. If a credit card company fails to meet the needs of these higher-income customers, they may lose out to competitors who offer these premium services.
- Changes in the economic environment may affect different groups of customers at different income levels differently. In times of recession or instability, lower-income customers may be more likely to experience financial difficulties, increasing the risk of churn. At the same time, changes in income, whether increasing or decreasing, may affect customers' credit card usage habits and loyalty.

3.2.3 Marital and family status

(Misra, 2012) found that customers preferred the same businesses that provided services to their family and friends. Impact of marital status and number of family members on customer churn: The number of family members may affect the spending habits, preferences, and disposable income of the rest of the family, which in turn affects customer loyalty to the credit card service provider.

- A larger number of household members may mean that the overall spending power of the household is higher, but the per capita spending power may be relatively low. This may result in such households being more price sensitive and more likely to change brands or service providers for financial reasons.
- Offering customized services or products that can meet the needs of different family sizes for customers who are married with children may affect customer loyalty. Credit card ancillary card services may be more attractive to families with children, thus reducing churn among these families.
- Households with a larger number of family members may place more value on family connections and shared experiences, which may affect their loyalty to the service. For example, a service that offers a shared experience for families may reduce churn for these families more than a service that is only for individuals. This indicates that family size status and marital status have a significant impact on customer churning behavior.

(Sathish et al., 2011) concluded that value added services and customer care are important factors influencing customer churn. The authors also found that family influence is the most important factor with about 41% of the subscribers switching providers due to family influence. Thus marital status and family status are important influences on customer churn.

3.2.4 Service Quality & Customer Care

(Ramzi & Mohamed, 2010) found that service quality is directly affected by customer loyalty, which in turn helps to retain customers and increase the company's market share. The amount of customer loyalty is the most important factor that affects customer churn because the business in the credit card industry relies on long-term relationships, that is, the longer a customer uses a credit card, the more loyal the customer will be. The quality of the

bank's services and the number of times customers contact the bank have a direct and indirect impact on customer churn.

- Bank service quality affects customer satisfaction and loyalty: high quality service increases customer satisfaction, which in turn increases customer loyalty and reduces churn. Service quality includes the quality of credit card services provided by the bank, the responsiveness of customer service and the ability to solve problems. It directly affects the average utilization rate of credit cards by customers
- Good service quality can attract new customers through word-of-mouth effect and also increase the retention rate of existing customers. Satisfied customers are more likely to recommend the bank's services to those around them, and this positive feedback can reduce customer churn.

In the banking industry, service quality is an important differentiator. A high standard of service can differentiate a bank from the competition and reduce customer churn due to competition.

Enhancement of customer relationship: frequent interactions can enhance the relationship between the customer and the bank and make the customer feel more valued, thus increasing customer satisfaction and loyalty.

- By interacting with the customer, the bank can gather more information about the customer and in turn provide a more personalized service. Personalized services enhance customer experience and reduce churn.
- Frequent contact helps banks to identify and resolve problems that customers may encounter in a timely manner, preventing problems from accumulating and leading to customer dissatisfaction and churn.

(Paulrajan & Rajkumar, 2011) argued that the important factors influencing customers' choice of service provider are service attitude and price followed by response to customer complaints.

3.2.5 Credit card limits and types

Bank card limit: A higher credit card limit usually means that the bank has a high level of trust in the customer and the customer has a good credit score. However, if a customer feels that the limit is too high for them to use, they may consider cancelling the card to simplify their financial management. Conversely, if the limit is too low to meet the

customer's spending needs, the customer may churn and look for a credit card product with a higher limit. Higher income customers may have a greater need for a higher limit.

Type of credit card: The type of credit card (e.g., regular, gold, platinum, etc.) and its associated services and benefits, such as points rewards, airport lounge access, travel insurance, etc., can have a significant impact on customer retention. If customers find that the services and benefits offered by the card no longer meet their needs, or if they find that other card brands offer more favorable services, this may lead to customer churn. Higher income customer segments are more favorable to more premium credit card tiers.

3.2.6 Client trading volume and recent trading changes

The total amount of transactions made by customers using a credit card reflects the degree of reliance on the card. A high frequency and high value of transactions means that the customer is more dependent on the card, reducing the likelihood of churn. However, if a card's frequency of use and transaction value continue to decline, it may indicate that the customer is less interested in the card, increasing the risk of churn.

Changes in a customer's transaction behavior, particularly within the last six months, can provide early signals of possible churn. For example, if there is a sudden decrease in the frequency or amount of a customer's transactions, it may indicate that the customer is beginning to move to other credit card products. Conversely, if the frequency and amount of transactions increase, it may indicate that the customer is more satisfied with their current card and is at a lower risk of churn.

3.3 Big data

Since the end of the 19th century and the beginning of the 21st century, the development of science and technology has grown at an astonishingly rapid rate. The constant updating of the Internet, electronic communication technologies and storage devices has resulted in the continuous generation of data at an exponential rate. The rapid growth of digital data eruption, which also led to the term "big data" also came into being in the development of information and Internet technology.

The boundaries of big data are very fuzzy, and in statistics, data that can no longer be processed by ordinary means are generally categorized as big data. But the definition of Big Data is not so simple, but much more complex. Even some digital data or files that do not

require much storage space are considered big data due to their complexity. At the same time, not all digital data or files that require big data storage capacity are complex enough to be considered big data.

Big data can help organizations better understand their operations, customer needs, market trends, etc. so that they can make more informed decisions. It can also be used to predict future trends and improve products or services.

3.3.1 Definition

Generally big data is data that consists of large sets of data, usually including structured data (e.g., tabular data in databases), semi-structured data (e.g., XML files), and unstructured data (e.g., text, image, and audio files). Big data is usually characterized by high rates of generation, diversity and large scale.

In the 1980s, the concept of "Big Data" appeared in Toffler's (1990) *The Third Wave*, which he called "the third wave of wonderful music" (Wang & Wang, 2021).

"John Mashey verbally used the term Big Data in a lunchtime chat at Silicon Graphics Incorporated (SGI) in the 1990s" (Diebold, 2012).

Whenever we say "what is big data", the first idea that comes to mind must be the sheer volume of data. However (Laney, 2001) proposed the three main challenges of data management called the "three V's": volume, variety and velocity.

Then some scholars started to use the concept of "V" to define "Big Data". (Gartner, 2020) to define "Big Data" in terms of V: Volume that massive amounts of data consuming large amounts of storage space; Velocity that The speed of data generation and frequency of data transmission; Variety which emphasises that data are generated from a variety of sources and formats and include multidimensional data fields such as structured and unstructured data.

The triple-V critical point is the turning point at which a dataset is transformed into big data.

3.3.2 Characteristics of Big data

There is no uniform definition of Big Data, however, it is commonly accepted that Big Data has 5V characteristics, namely Volume, Variety, Velocity, Value, and Veracity (Tang et al., 2022).

(Laney, 2001) first identified three primary characteristics of Big Data, the "three V's": volume, variety, and velocity. Volume refers to the size of the data. The size of big data is measured in multiple terabytes and petabytes. 1024 TB is equivalent to 1 PB. Variety refers to the structural heterogeneity in a data set: structured data, such as relational databases and tabular data in Excel; and unstructured data, such as text, images, audio, and video, which have no machine-analysable structure. Velocity refers to the speed at which data is transferred and the efficiency with which it is generated per minute and second.

Value: Unprocessed raw data is usually worthless, but by analysing the data, its raw value is enhanced and a higher value is realised. The feature of adding value to big data is used to emphasise the importance of deriving economic benefits from existing big data (Sivarajah et al., 2017).

Veracity implies the presence of unreliable elements in the data source (Gandomi & Haider, 2015). They also propose veracity in big data to emphasise the importance of data quality and trust in agnostic sources.

Literature suggests that Big Data has 3-5 characteristics, three of which are Volume, Velocity and Variety: Volume, Velocity and Variety are the most commonly identified. There is no agreement on the veracity and value of the other characteristics. Volume, Velocity and Variety can be referred to as the three V's criticality.

3.4 Data mining and algorithmic models

Data mining is generally considered to be the process of discovering meaningful new associations, patterns and trends by analysing stored big data using statistical and mathematical techniques. Data mining as: the extraction of previously unknown, implicitly useful information from potentially valuable data.

(Fayyad et al., 1996) showed that data mining is the process of applying different algorithms to find valuable patterns or relationships in a data set.

(Linoff & Berry, 2011) suggest that data mining is the process of exploring and analysing large amounts of data to discover meaningful patterns and rules.

3.4.1 Logistic regression

Logistic regression is a generalised linear regression. It represents a multiple regression model where the outcomes are categorical (usually binary variables) and the

predictors can be continuous or categorical. The dependent variable in logistic regression can be either a binary or a multicategorical variable, but binary variables are more commonly used. Therefore, logistic regression models are often used when the predictors are binary variables.

Logistic regression is one of the most important models for categorical response data. This logistic regression algorithm can be used to discover relationships or trends between large data sets and binary target variables. For example, the probability that a subject is trustworthy is modelled in business in the form of a credit score.

This logistic regression algorithm can be obtained by using the relationship between the large data set and the binary target variable. In logistic regression, the Odds ratio is used to determine the incidence rate or trigger probability. We generally think of it as the proportion of events that occur.

$$\text{odds ratio} = \frac{P(1)}{1 - P(1)} = \frac{P(1)}{P(0)} \quad (1)$$

We can determine the parameters of the logistic regression by the maximum likelihood function.

The degree of matching between the statistical model and the data sample is shown to be expressed by the likelihood function, and the maximum likelihood estimation is applied to find the parameters with the best match by the gradient descent method. When the parameters of the regression are determined, the odds ratio of logistic regression can be presented as follow:

$$\text{Log odds ratio} = \beta_0 + \beta_1 \cdot x_1 + \dots + \beta_n \cdot x_n \quad (2)$$

The probability of an event occurring can be represented as a logistic curve to show. The shape of the function visible on Figure. Unlike linear regression, the shape of the logistic regression function is similar to the inverse "sin function". The logistic curve takes values in the interval [0;1], and generally we interpret the value of 0.5 as the decision boundary between the two classes. The probability of an event is represented by the following function:

$$P(\text{target} = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot x_1 + \dots + \beta_n \cdot x_n)}} \quad (3)$$

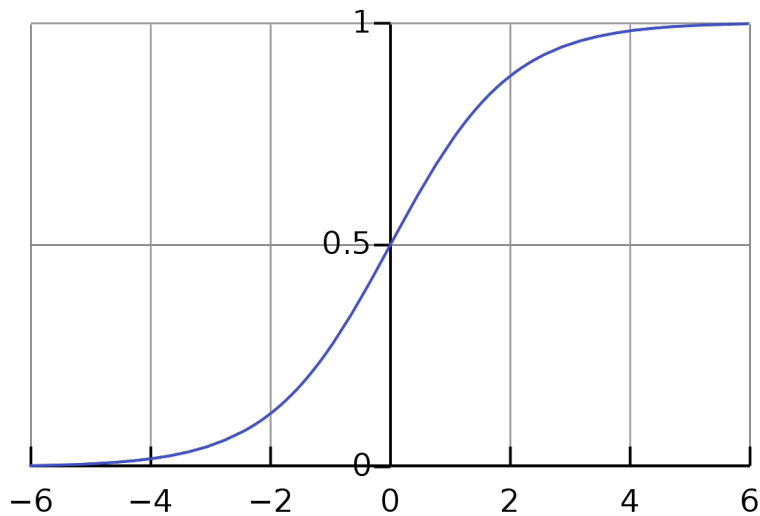


Figure 1 Logistic regression curve (Qef, 2008)

Since logistic regression is a number-based algorithm, all categorical variables need to be converted into a numerical form during data preparation.

3.4.2 Linear regression

Regression analysis is a statistical technique which is usually used for the estimation of the relationship among different variables. In other words, we will use regression model to analyze the correlation among many variables having cause-effect relationships and based on our model to make a prediction for our data.

In the very beginning of the book named “Applied linear regression” by (Weisberg, 2005), regression is discussed with the fundamental graphical tool, a two-dimensional scatterplot. Based on the definition, the scatterplot of the response will be presented in the graph showing the relationship between dependent variable and independent variable.

The independent variables are also called as “predictors”, and is shown as “ X_i ” in the graph; while the dependent variable is shown as “ Y ” in the scatterplot. Our aim is to understand how the value of Y change as X is varied over its range of possible values. Before we really get into the details of our sample, we can always first graphically look at our data to get a general picture of how our variables look like.

The following figure shows the very basic scatterplot of the linear regression model. As we can see in the figure that, the scatterplot is composed by a series of individual blue plots, and each plot represents corresponding value of “x” and “y”. If we want to know the prediction of a corresponding value of “y” based on a specific value of “x”, we need to understand the trend of the relationship between “x” and “y”, which is represented as a red straight line in the following graph. It is very clear to say that, the independent variable “x” has a positive relation with dependent variable “y” as the increase of x will cause the increase of y at the same time.

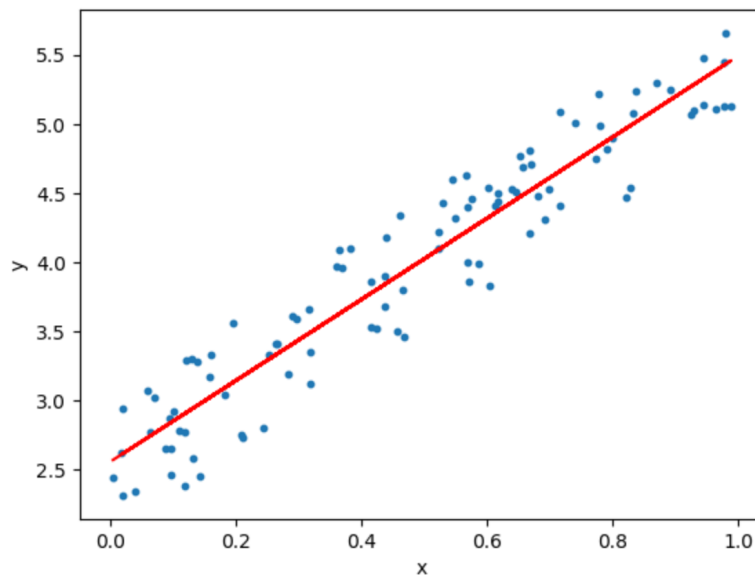


Figure 2 Linear regression (Agarwal, 2018).

The regression can be divided into two kinds, first one is using only one single independent variable, which is called “uni-variate regression”; and the other one is using more than one independent variable, which is called “multi-variate regression” (Tabachnick et al., 2013). For practical reason, we commonly apply the later approach. However, we will introduce both the simple linear regression and the multivariate regression .

Simple linear regression

The function of simple linear regression is expressed as follow:

$$E(Y|x) = \beta_0 + \beta_1 * x \quad (4)$$

Where E(), means the expected value, and we aim to find the possible value of Y when x is limited to some specific value. Beta zero β_0 is the intercept parameter, and beta one β_1 is the slope parameter. We usually call the parameter that can estimated the value in the model as “coefficient”.

The hypotheses of the simple linear regression is given as:

$$H_0: \beta_1 = 0 \quad (5)$$

$$H_1: \beta_1 \neq 0 \quad (6)$$

H_0 is the null hypothesis, and if the null hypothesis is true, then we can say that based on our model, x has no effect on Y at all.

Multivariate regression

The logic behind the multivariate regression is just like the simple linear regression one, We expect to find the best fit line of our model, but with more than one independent variable in the model (Abbott, 2014).

$$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_i X_i \quad (7)$$

Where Y is the dependent variable, and we have the X_i and different coefficient to explain Y . This function is actually quite similar to the simple linear regression, only adding more dependent variables in the model (Statistics Solutions, 2022).

Because we have more than one dependent variables in the model, the assumptions of multivariate regression is as follow:

- Linearity
- Normality
- No multi-collinearity
- No auto-correlation
- Homoscedasticity

3.4.3 Decision Tree (CART and CHAID algorithm)

Decision tree learning is widely used in data mining, statistics and machine learning. Decision tree learning is a decision model built using a tree structure based on the attributes of the input data to predict the value of a target attribute, and the output is a classification model. Each internal node in these trees represents an input variable respectively, the number of branches induced by the node is equal to the number of all possible values of that input variable, and the leaf nodes represent the final judgement outcome which is the value of the attribute. Decision tree learning can handle both continuous and categorical variables so it requires less effort in data preparation than other methods and is used here to predict the probability of a credit card customer churn.

According to the definition on The Economic Times, decision tree analysis is used to break down complex problems. Usually it involves making a tree-shaped diagram to chart

out a course of action or a statistical probability analysis. We need to note here that the decision tree is not only used in finance and economic issues, but also in philosophy and machine learning problems. It is very useful to draw a final conclusion for the problems with many branches (The Economic Times, 2022).

There are five steps of decision tree analysis generally:

Define the problem area for which decision making is necessary.

Draw a decision tree with all possible solutions and their consequences.

Input relevant variables with their respective probability values.

Determine and allocate payoffs for each possible outcome.

Calculate the Expected Monetary Value for every chance node in order to determine which solution is expected to provide the most value. Circles represent chance nodes in a tree diagram (Team Asana, 2021).

Decision Tree Hugging

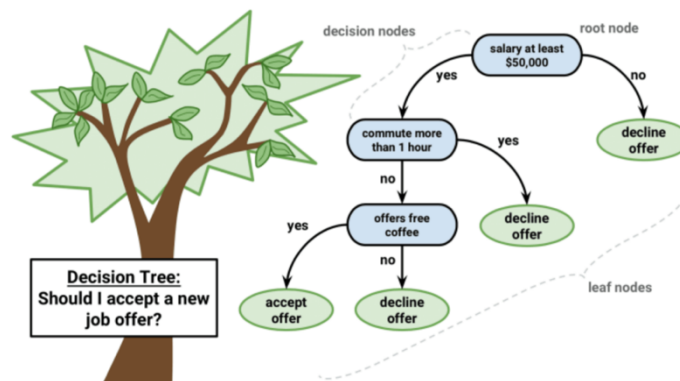


Figure 3 Decision Tree (Gray, 2017)

For each of the decision tree algorithms described, the algorithmic steps are as follows (Abbott, 2014).

1. for each candidate input variable, evaluate the best way to split the data into two or more. Select the best way to split the data into subgroups defined by the splitting method.
2. Select one of the subgroups and repeat step 1 (this is the recursive part of the algorithm). Repeat for each subgroup.
3. Continue the segmentation until all records in the segmentation belong to the same target variable value, or apply another stop condition. The stopping condition can be a complex test of statistical significance or a simple minimum number of records.

CART analysis is a statistical method used to identify the explanatory variables that influence the target variable (Takeshi , et al., 2021). In the CART algorithm binary trees are constructed. The impurity measure is in the form of Gini index (Leszek , Maciej, Lena, & Piotr, 2014).

The CART algorithm, explains how to predict the value of an outcome variable based on other values. the output of CART is a decision tree where each fork is a fork of the predictor variable and each end node contains a prediction of the outcome variable.

We will now briefly describe the CART algorithm. The CART algorithm starts with a single node - the root. In the learning process, a specific subset S_q of the training data set S is processed in each created node. If all elements of the set S_q belong to the same class, the node is marked as a leaf and no segmentation is performed. Otherwise, the best attribute among the available attributes of the considered nodes is selected for segmentation according to the segmentation measurement function (Leszek , Maciej, Lena, & Piotr, 2014).

Any method has its advantages and its shortcomings, the decision tree model is not an exception as well. First we can talk about its strengths, it is very easy and direct to interpret the process of the whole decision-making. It is also quite valuable without requiring large amounts of hard data. Third, it really helps decision makers ascertain best, worst, and expected results for many kinds of scenarios. Last but not least, it can be combined with various decision models (Team Asana, 2021).

Decision tree learning algorithms are very efficient and scale well as the number of records or fields in the modeling data increase (Abbott, 2014).

Any extreme values are separated in small nodes and do not affect the classification problem (Tufféry, 2011).

(Singh et al., 2024) and (Wagh et al., 2024) indicated that the decision tree model had a poor performance on customer churn determination.

On the other hand, there are some weaknesses of the decision trees model. First one we note here is that, even minor data changes could lead to major structure changes of our decision-making. Second that we mention is, information gain in decision trees can be biased. Third is uncertain values can lead to complex calculations and uncertain outcomes. Last but not least is, decision trees can often be relatively inaccurate (Team Asana, 2021).

There are two types of decision tree models that are used to construct them, namely CART and CHAID. CART stands for classification and regression trees where as CHAID represents Chi-Square automatic interaction detector.

Although CHAID can mine as much information as possible in the learning of the training sample set, its generated decision trees are relatively large in branching and size. The dichotomy of CART algorithm can simplify the size of decision trees and improve the efficiency of generating decision trees. Its operation is simpler. Of course, its performance is also very close to that of the entropy model.

- CHAID is a multinomial tree with slow operation speed, CART is a binary tree with fast operation speed.
- CART uses the Gini coefficient as an impurity measure for variables, reducing a large number of logarithmic operations.

CHAID is most frequently used for descriptive analysis whereas CART is frequently used in predictive analysis.

The CART algorithm uses a dichotomous recursive partitioning method. The generation of a decision tree is the process of recursively constructing a dichotomous decision tree. The algorithm always divides the current sample set into two sub-sample sets so that the resulting decision tree has only two branches per non-leaf node. The decision tree generated by the CART algorithm is therefore a binary tree with a simple structure. Therefore, the CART algorithm is suitable for problems where the sample features are yes or no.

3.4.4 Evaluation of model

When it comes to the evaluation of the logistic regression model, We often use VIF, Likelihood ratio, (receiver operating characteristic)ROC curve and AUC value to evaluate logistic regression models.

The Receiver Operator Characteristic (ROC) curve is evaluation metric for binary classification problems

- it is a probability curve that plots the **TPR** against **FPR** at various threshold values
- it shows the performance of a classification model at all classification thresholds

TPR= true positive rate (Sensitivity)

FPR=false positive rate (1- Specificity)

In the following paragraph, we will introduce some commonly used evaluation to check the performance of logistic regression model.

Confusion matrix and some relevant terms:

We will introduce the confusion matrix first in this part, which is quite useful for the further evaluation of the logistic regression model. The confusion matrix is a simple table that can be used to evaluate the performance of a classification model. It summarizes the count combinations of every predicted and actual class.

In the table, We have “positive/negative” which means the class is predicted as “positive/negative”, while “true/false” means whether our model detect or classify the data sample as “correct/incorrect” groups.

We need to stress that predicted class and actual class are two different concepts. For example, False Positive indicates that actual classification should be Negative, even though it is classified as positive by model. likewise, False Negative means actual classification should be Positive.

As shown in the following figure, we can see clearly there are four combinations of the results:

- True positive
- False negative
- True negative
- False positive

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

Figure 4 Combination of results (An, 2020)

Based on the values of the above four situation, we can now calculate precision, sensitivity, specificity, accuracy and so on. For instance, the accuracy is the proportion of the total number of all the correct predictions.

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (8)$$

And we define the precision as the ratio of the total number of correctly classified positive classification and the total number of predicted positive classification. That's why it is also known as positive predictive value.

$$Precision = \frac{TP}{(TP + FP)} \quad (9)$$

Sensitivity is also called as “recall” or we just say “true positive rate”, is the proportion of the total number of actual positives that were identified correctly. Actual positives include situations of true positives and false negatives.

$$Sensitivity = \frac{TP}{(TP + FN)} \quad (10)$$

As we probably all know that this is vital for the COVID-19 test. To be specific, if the COVID-19 test is 95% of sensitivity, this indicates that 95 out of 100 infected patients will be correctly diagnosed, and only 5 infected patients are off the radar based on test results.

Specificity should also be mentioned as the proportion of the total number of the actual negative that were identified correctly. The actual negative include situations of true negatives and false positives. It is also known as the true negative rate (TNR).

$$Specificity = \frac{TN}{(TN + FP)} \quad (11)$$

Nevertheless, we should be aware that there is a trade-off between recall and precision. It shows how the recall vs precision relationship changes as we vary the threshold for identifying a positive in our model. When we increase the recall rate by adjusting the classification threshold of a model, the precision rate is decreased.

After we introduce the confusion matrix and all relevant terms, we can get to know ROC/AUC. Generally, ROC means a curve plots the true positive rate on the y-axis versus the false positive rate on the x-axis. It shows the true positive and false positive rate for every probability threshold of a binary classifier.

In other words, the higher the result of ROC, the better is our model. And AUC is just “area under the curve”, so it represents the similar meanings, with high value suggesting better performance of the model.

4. Practical Part

Introduction

This chapter introduces the research analysis as specified set out in the research methodology. The analysis of the data set is consistent with the objectives. The results of the analysis are presented in graphical form.

4.1 Overview of the Case Study

The dataset was obtained from Kaggle, which contains information on 10,127 customers collected from a portfolio of consumer credit cards. Kaggle is the world's largest data science community with powerful tools and resources. After obtaining the credit card churn prediction dataset from this website, data exploration and cleaning activities were conducted. Data cleaning was performed using EXCEL and IBM SPSS Modeler. only 20 variables were selected from the given data, any duplicates were removed, and variable values were defined before any analysis was performed. The purpose of analyzing this dataset was for academic purposes only and the data will not be used for any other purpose. The dataset's information on the 10,127 credit card spending users includes comprehensive demographic data such as age, gender, marital status, and income category, as well as each customer's relationship with their credit card provider, such as type of credit card, number of months of billing, and period of inactivity. It also includes key data on customer spending behaviour, which is closer to the decision to churn customers, such as total revolving balance, credit limit, average open purchase rate and analyzable metrics, such as total change amount from Q4 to Q1, and average usage rate. As a result, churn forecasting of customers through predictive modelling when managing portfolios or servicing individual customers improves service offerings and provides a higher likelihood of retaining existing users.

With this complex dataset spanning multiple variables, data mining and big data analytics will be used to extract and capture valuable information from it, and it can be used to build credit card customer churn prediction models. Determining the long-term stability of an account can also determine if an account is about to leave.

Before analyzing the data and its variables, it is important to understand the benefits of this study. Customer acquisition and retention are two major components for the banking and credit card industry. Where retaining customers costs 30% of the total expenditure for the business and this sunk cost expenditure is very important for the business. Therefore

understanding and analyzing the major factors of credit card customer churn behavior and building predictive models to help banks and credit card organizations to predict credit card customer churn from their own customer databases is crucial for both businesses and organizations.

The variables in this dataset will be analyzed in detail in the next sections of this study. The tool used to analyze the dataset was IBM SPSS Modeler. IBM SPSS Modeler a data mining and predictive analytics software that uses a visual interface to create predictive models by invoking statistical and data mining algorithms such as logistic regression and decision trees, and to generate the results of the predictive models.

4.2 Data pre-processing

Field	Measurement	Values	Missing	Check	Role
CLIENTNUM	Typeless			None	Record ID
Attrition_Flag	Flag	1,0	0	None	Target
Age	Continuous	[26,0,73,0]		None	Input
Gender	Nominal	0,0,1,0		None	Input
Dependent_count	Continuous	[0,0,5,0]		None	Input
Education_Level	Ordinal	0,0,1,0,2,0,3,0,4,0...		None	Input
Marital_Status	Nominal	0,0,1,0,2,0,3,0		None	Input
Income_Category	Ordinal	0,0,1,0,2,0,3,0,4,0...		None	Input
Card_Category	Ordinal	1,0,2,0,3,0,4,0		None	Input
Months_on_book	Continuous	[13,0,56,0]		None	Input
Total_Relationship_C...	Continuous	[1,0,6,0]		None	Input
Months_Inactive_12...	Continuous	[0,0,6,0]		None	Input
Contacts_Count_12...	Continuous	[0,0,6,0]		None	Input
Credit_Limit	Continuous	[1438,3,34516,0]		None	Input
Total_Revolving_Bal	Continuous	[0,0,2517,0]		None	Input
Avg_Open_To_Buy	Continuous	[3,0,34516,0]		None	Input
Total_Amt_Chng_Q4...	Continuous	[0,0,3,397]		None	Input
Total_Trans_Amt	Continuous	[510,0,18484,0]		None	Input
Total_Trans_Ct	Continuous	[10,0,139,0]		None	Input
Total_Ct_Chng_Q4_Q1	Continuous	[0,0,3,714]		None	Input
Avg_Utilization_Ratio	Continuous	[0,0,0,999]		None	Input
Naive_Bayes_Classi0...	Continuous	[7,66E-6,0,99958]		None	None
Naive_Bayes_Classi0...	Continuous	[4,1998E-4,0,99999]		None	None

Figure 5 Dataset overview

Source: Author's own work

The most important thing to do before starting EDA is to check for duplicates and missing data. If the check result is "Yes", then we need to create some indexes to remove duplicate data. As shown in Figure 5, there is no duplicate data in the whole dataset because each credit card customer's data is unique and cannot be duplicated.

Secondly, defining the type of variable is a key part of getting started so that understanding the variable and its effect on the dependent variable is quite vital.

As shown in Figure 5, the original dataset contains 23 variables, and the target variable to "Attrition_Flag" will be set in the role. The role of variable "CLIENTNUM" was changed to Record ID. The role of variables "Naive_Bayes_Classi0..." were changed to none, as it is not relevant for our data analysis. All other variables are inputs.

After preprocessing, there are 20 variables, the target variable "Attrition_Flag" and 19 other Input variables.

4.2.1 Measurement types, labels

As shown in Figure 5 above, the measurement of data types has been changed where,

- Nominal includes (2): "Gender", "Marital_Status".
- Ordinal (3): "Education_Level", "Income_Category", "Card_Category"
- Continuous (14): "Age", "Dependent_count", "Months_on_book", "Total_Relationship_Count", "Months_Inactive_12_mon", "Contacts_Count_12_mon", "Credit_Limit", "Total_Revolving_Bal", "Avg_Open_To_Buy", "Total_Amt_Chng_Q4_Q1", "Total_Trans_Amt", "Total_Trans_Ct", "Total_Ct_Chng_Q4_Q1", "Avg_Utilization_Ratio".

4.2.2 Missing values analysis

This step is a process that can be demonstrated through quality in Data Audit if there are missing values in the dataset.

As shown in Figure 6 below, all 20 variables are 100% complete and there are no missing values. And all samples shown in Figure are 10127.

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String
Attrition_Flag	Flag	--	--		Never	Fixed	100	10127	0	0
Age	Continuous	1	0 None		Never	Fixed	100	10127	0	0
Gender	Nominal	--	--		Never	Fixed	100	10127	0	0
Dependent...	Continuous	0	0 None		Never	Fixed	100	10127	0	0
Education_L...	Ordinal	--	--		Never	Fixed	100	10127	0	0
Marital_Sta...	Nominal	--	--		Never	Fixed	100	10127	0	0
Income_Cat...	Ordinal	--	--		Never	Fixed	100	10127	0	0
Card_Catego...	Ordinal	--	--		Never	Fixed	100	10127	0	0
Months_on_...	Continuous	0	0 None		Never	Fixed	100	10127	0	0
Total_Relat...	Continuous	0	0 None		Never	Fixed	100	10127	0	0
Months_Ina...	Continuous	124	0 None		Never	Fixed	100	10127	0	0
Contacts_Co...	Continuous	54	0 None		Never	Fixed	100	10127	0	0
Credit_Limit	Continuous	0	0 None		Never	Fixed	100	10127	0	0
Total_Revol...	Continuous	0	0 None		Never	Fixed	100	10127	0	0
Avg_Open_...	Continuous	0	0 None		Never	Fixed	100	10127	0	0
Total_Amt_...	Continuous	135	28 None		Never	Fixed	100	10127	0	0
Total_Trans...	Continuous	391	0 None		Never	Fixed	100	10127	0	0
Total_Trans...	Continuous	2	0 None		Never	Fixed	100	10127	0	0
Total_Ct_Ch...	Continuous	76	37 None		Never	Fixed	100	10127	0	0
Avg_Utilizat...	Continuous	0	0 None		Never	Fixed	100	10127	0	0

Figure 6 Missing value - check
Source: author's own work

4.3 Explanatory Data Analysis

The first phase of the study should provide an understanding of the dataset. The purpose of this phase is to provide as much information about the dataset as possible.

Explanatory data analysis is an activity that includes classification and numerical activity. Therefore we need to have a comprehensive analysis and understanding of the variables and the effect of each variable on the dependent variable through EDA.

4.3.1 Descriptive statistics – Numerical variables

As shown in Figure 7 and Figure 8 below. The directly obtain statistical descriptions of the data and data exploratory analysis (EDA) through SPSS Modeler’s Data Audit will be shown below.

Field	Sample Graph	Measurement	Min	Max	Sum	Mean	Std. Dev	Variance	Skewness	Unique	Valid
Attrition_Flag		Flag	0.000	1.000	--	--	--	--	--	2	10127
Age		Continuous	26.000	73.000	469143.000	46.326	8.017	64.269	-0.034	--	10127
Gender		Nominal	0.000	1.000	--	--	--	--	--	2	10127
Dependent_count		Continuous	0.000	5.000	23760.000	2.346	1.299	1.687	-0.021	--	10127
Education_Level		Ordinal	0.000	6.000	--	--	--	--	--	7	10127
Marital_Status		Nominal	0.000	3.000	--	--	--	--	--	4	10127
Income_Category		Ordinal	0.000	5.000	--	--	--	--	--	6	10127
Card_Category		Ordinal	1.000	4.000	--	--	--	--	--	4	10127
Months_on_book		Continuous	13.000	56.000	363847.000	35.928	7.986	63.783	-0.107	--	10127
Total_Relationship_Count		Continuous	1.000	6.000	38610.000	3.813	1.554	2.416	-0.162	--	10127
Months_Inactive_12_mon		Continuous	0.000	6.000	23709.000	2.341	1.011	1.021	0.633	--	10127

Figure 7 Data Audit-1

Source: author's own work

Contacts_Count_12_mon		Continuous	0.000	6.000	24865.000	2.455	1.106	1.224	0.011	--	10127
Credit_Limit		Continuous	1438.300	34516.000	87415795.100	8631.954	9088.777	82605860.998	1.667	--	10127
Total_Revolving_Bal		Continuous	0.000	2517.000	11775818.000	1162.814	814.987	664204.357	-0.149	--	10127
Avg_Open_To_Buy		Continuous	3.000	34516.000	75639977.100	7469.140	9090.685	82640559.654	1.662	--	10127
Total_Amt_Chng_Q4_Q1		Continuous	0.000	3.397	7695.919	0.760	0.219	0.048	1.732	--	10127
Total_Trans_Amt		Continuous	510.000	18484.000	44600182.000	4404.086	3397.129	11540487.165	2.041	--	10127
Total_Trans_Ct		Continuous	10.000	139.000	656824.000	64.859	23.473	550.962	0.154	--	10127
Total_Ct_Chng_Q4_Q1		Continuous	0.000	3.714	7212.676	0.712	0.238	0.057	2.064	--	10127
Avg_Utilization_Ratio		Continuous	0.000	0.999	2783.847	0.275	0.276	0.076	0.718	--	10127

¹ Indicates a multimode result ² Indicates a sampled result

Figure 8 Data Audit-2

Source: author's own work

The histogram is to see the distribution, set the Attrition_Flag to be normalized by color, and the results are shown in Figure below.

For the continuous variable "Age" we can obtain that the minimum value is 26, the maximum value is 73 and the average value is 46.326. This means that the minimum age in

the sample is 26 years old already adult, the maximum age is 73 years old and the average age is about 46 and a half years old.

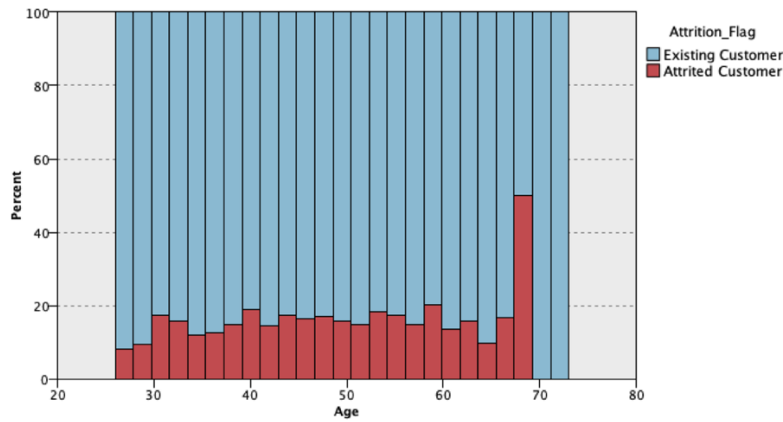


Figure 9 Age distribution
Source: author's own work

For the continuous variable "Dependent_count" we can come up with a minimum value of 0, a maximum value of 5 and a mean value of 2.346. This means that Number of dependents that customer has in the sample has a minimum of 0 and a maximum of 5, with an average number of about 2 and a half.

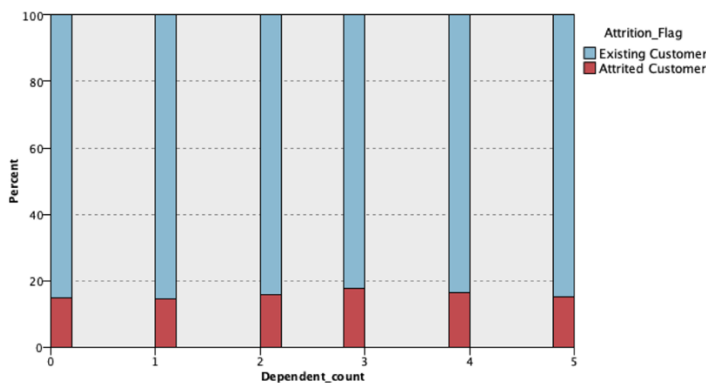


Figure 10 Dependent_count distribution
Source: author's own work

For the continuous variable "Months_on_book", we can come up with a minimum value of 13, a maximum value of 56, and a mean value of 35.928. This means that the minimum time that customer has been on the books in the sample is 13 months i.e. one year, the maximum is 56 months i.e. 4 and a half years and the average is 36 months i.e. 3 years.

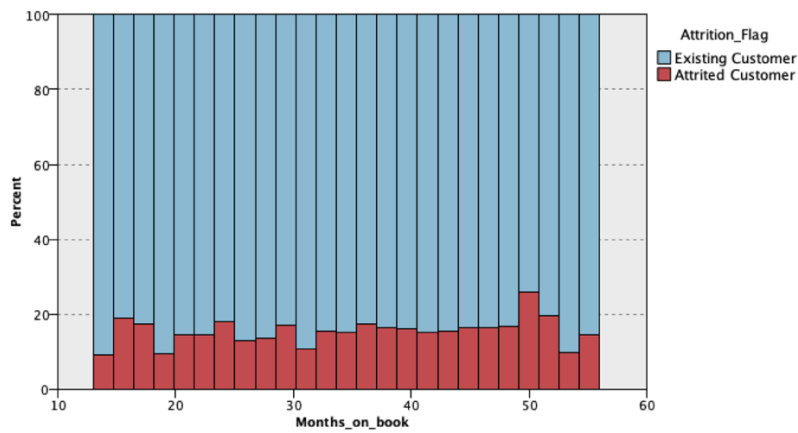


Figure 11 Months_on_book distribution
Source: author's own work

For the continuous variable "Total_Relationship_Count", we can conclude that the minimum value is 1, the maximum value is 6 and the average value is 3.813. This means that the Total number of relationships customer has with the credit card provider in the sample has a minimum of 1 and a maximum of 6, with an average of about 4 credit cards.

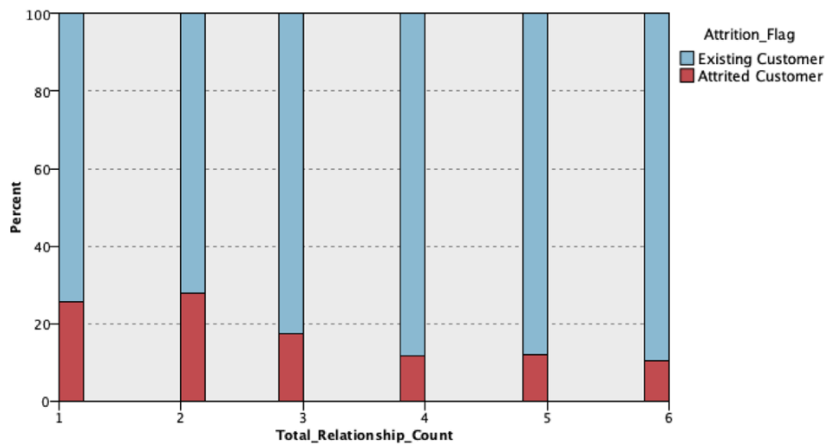


Figure 12 Total_Relationship_Count distribution
Source: author's own work

For the continuous variable "Months_Inactive_12_mon", we can come up with a minimum value of 0, a maximum value of 6, and a mean value of 2.341. This means that the sample Number of months customer has been inactive in the last twelve months has a minimum of 0, a maximum of 6 months, and an average of about 2 months of inactivity.

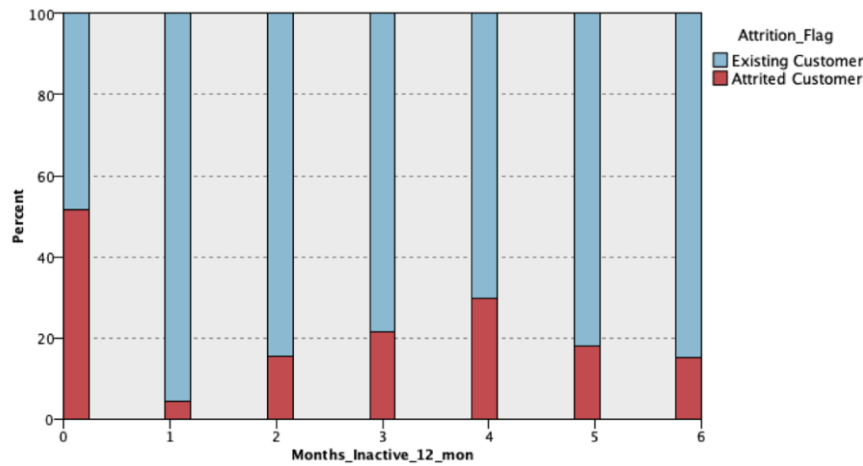


Figure 13 Months_Inactive_12_mon distribution

Source: author's own work

For the continuous variable "Contacts_Count_12_mon", we can come up with a minimum value of 0, a maximum value of 6, and a mean value of 2.455. This means that the sample Number of contacts customer has had in the last twelve months The minimum is 0, the maximum is 6, and the average is about 3.

The higher number of contacts customer has had, the higher the churn rate.

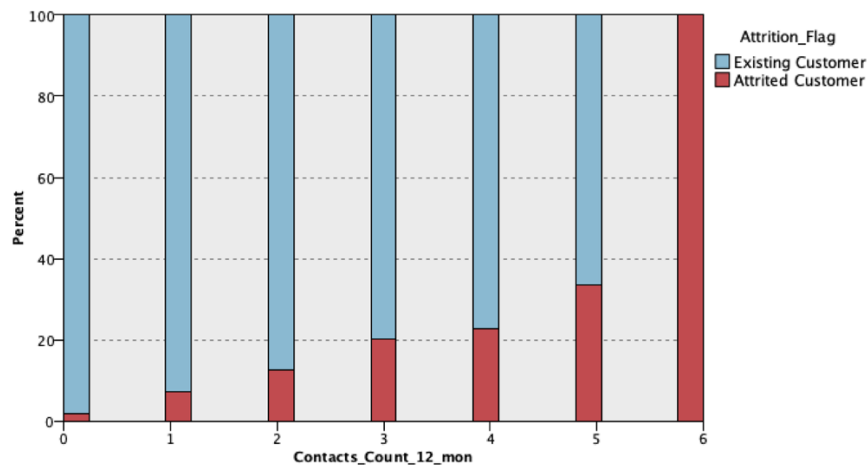


Figure 14 Contacts_Count_12_mon distribution

Source: author's own work

For the continuous variable "Credit_Limit", we can come up with a minimum value of 1438.3, a maximum value of 34516, and an average value of 8631.954. This means that the minimum Credit limit of customer in the sample is 1438.3 , the maximum is 34516 and the average is approximately 8632.

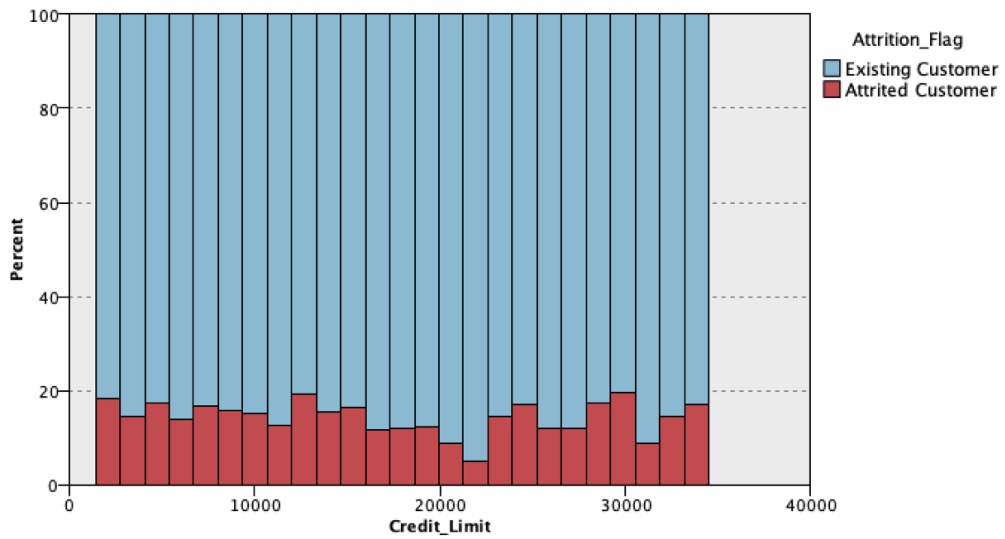


Figure 15 Credit_Limit distribution

Source: author's own work

For the continuous variable "Total_Revolving_Bal", we can come up with a minimum value of 0, a maximum value of 2517 and an average value of 1162.814. This means that the minimum Total revolving balance of customer in the sample is 0 , the maximum is 2517 and the average is about 1163.

The lower Total revolving balance of customer, the higher the customer churn rate.

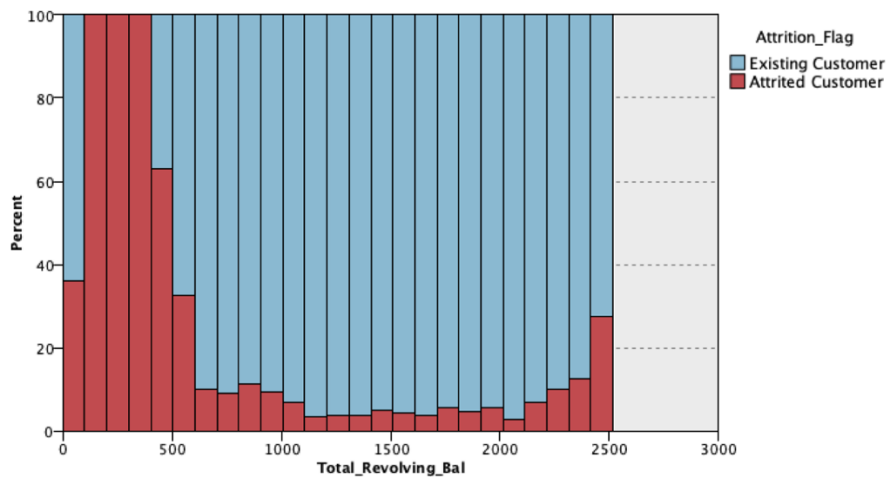


Figure 16 Total_Revolving_Bal distribution

Source: author's own work

For the continuous variable "Avg_Open_To_Buy", we can come up with a minimum value of 3, a maximum value of 34516, and an average value of 7469.14. This means that the Average open to buy ratio of customer in the sample has a minimum of 3 , a maximum of 34516 and an average of about 7469.

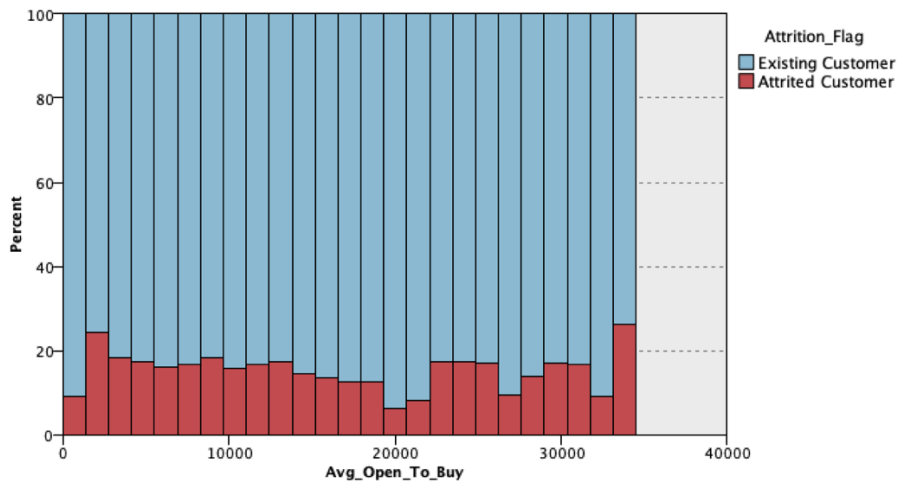


Figure 17 Avg_Open_To_Buy distribution
 Source: author's own work

For the continuous variable "Total_Amt_Chng_Q4_Q1", we can come up with a minimum value of 0, a maximum value of 3.397, and a mean value of 0.76. This means that Total amount changed from quarter 4 to quarter 1 in the sample has a minimum of 0, a maximum of 3, and an average of about 1.

The smaller the Total amount changed from quarter 4 to quarter 1, the higher the probability of customer churn. Customer churn is mainly concentrated at less than 1.

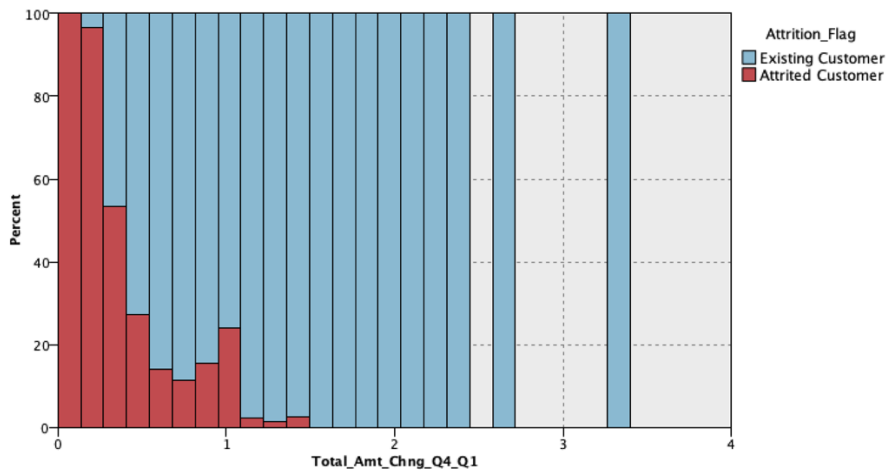


Figure 18 Total_Amt_Chng_Q4_Q1 distribution
 Source: author's own work

For the continuous variable "Total_Trans_Amt", we can come up with a minimum value of 510, a maximum value of 18,484, and an average value of 4404.086. This means that the minimum Total transaction amount in the sample is 510, the maximum is 18484 and the average is about 4404.

When Total transaction amount is between 0 and 11000, the larger the amount, the higher the probability of customer churn. Above 11,000 there is almost no credit card churn.

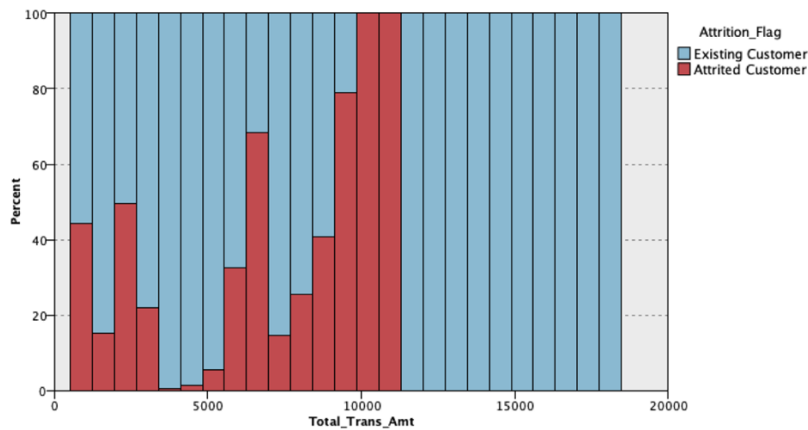


Figure 19 Total_Trans_Amt distribution

Source: author's own work

For the continuous variable "Total_Trans_Ct" , the conclusion is that the minimum value is 10, the maximum value is 139 and the average value is 64.859. This means that the minimum Total transaction count in the sample is 10, the maximum is 139 and the average is about 65.

The smaller the Total transaction count, the higher the probability of churn. Above 100 there is almost no credit card churn.

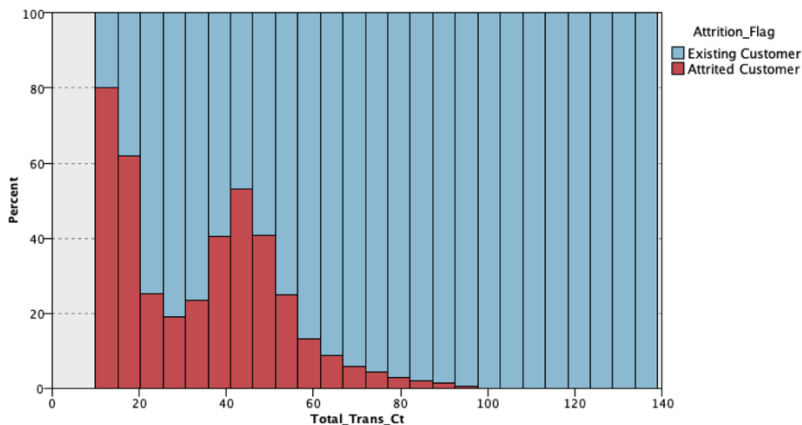


Figure 20 Total_Trans_Ct distribution

Source: author's own work

For the continuous variable "Total_Ct_Chng_Q4_Q1", we can conclude that the minimum value is 0, the maximum value is 3.714 and the average value is 0.712. This means that Total count changed from quarter 4 to quarter 1 in the sample has a minimum of 0, a maximum of 4 and an average of about 1.

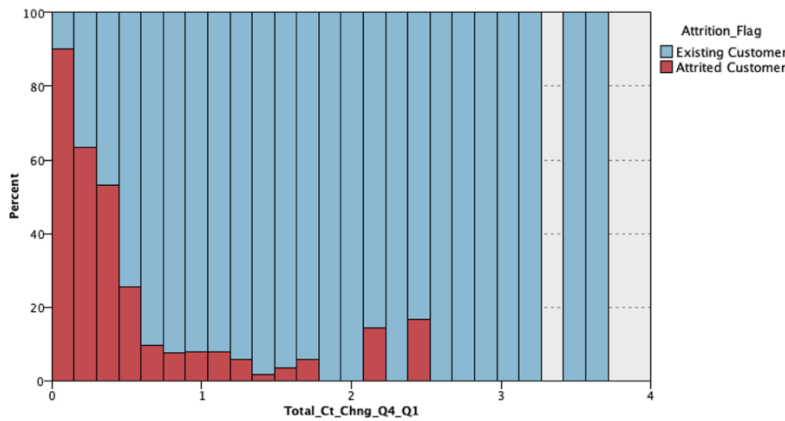


Figure 21 Total_Ct_Chng_Q4_Q1 distribution
 Source: author's own work

For the continuous variable "Avg_Utilisation_Ratio", we can conclude that the minimum value is 0, the maximum value is 0.999 and the average value is 0.275. This means that the Average utilisation ratio of customer in the sample has a minimum of 0 and a maximum of 1, with an average of about 0.275.

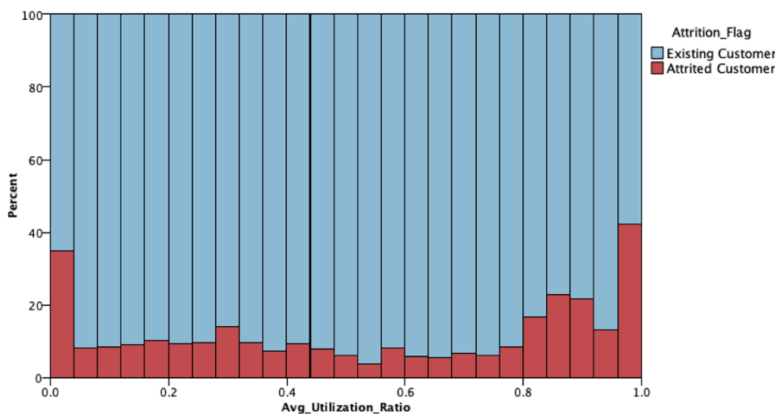


Figure 22 Avg_Utilization_Ratio distribution
 Source: author's own work

4.3.2 Univariate statistical analysis – Category variables

Firstly, the univariate statistical analysis will be used to assess and analyze each variable separately and present the distribution in graphs.

The first variable to be described is the dependent variable "y", which is the target variable for the current dataset, Attrition_Flag. It contains information about whether credit card customers are churning or not.

Target variable: Attrition_Flag

Figures 23 and 24 show the data distribution of target variable "Attrition_Flag". The value 0 represents "Customer attrition No" and 1 represents "Customer attrition Yes". It can

be seen that the number of un-churned customers is 8500 and the number of credit card customers that have been churned is 1627, which is 16.07 per cent of the total observations.

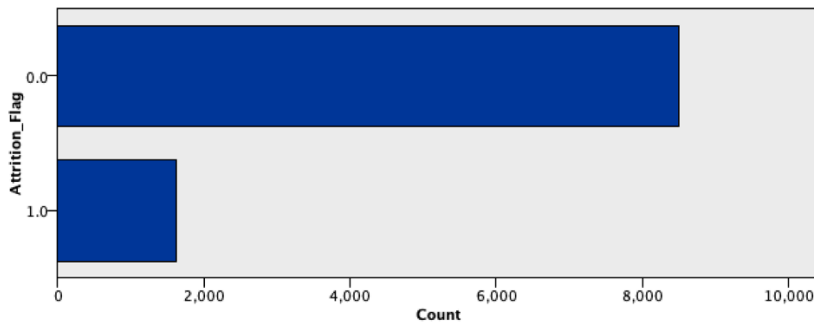


Figure 23 Attrition_Flag-1
Source: author's own work

Table **Graph** **Annotations**

Value #	Proportion	%	Count
0.000		83.93	8500
1.000		16.07	1627

Figure 24 Attrition_Flag_2
Source: author's own work

The target variable's visualized distribution is shown in Figure 24, where we can observe the number of churned customers vs. the number of un-churned customers as a percentage of the total number of observations. The dataset is very unbalanced. It shows that 83.93% of the customers are not churned while the probability of churning a credit card customer is 16.07%. From the comparison in the graph, we can conclude that in general the probability of churning a credit card customer may be around 15%.

We can see the distribution, set the Attrition_Flag to be normalized by color, and the results are shown in Figure below.

Input variable: Gender

Value #	Proportion	%	Count
0.000		52.91	5358
1.000		47.09	4769

Attrition_Flag

0.0

1.0

Figure 25 Gender-1 Distribution
Source: author's own work

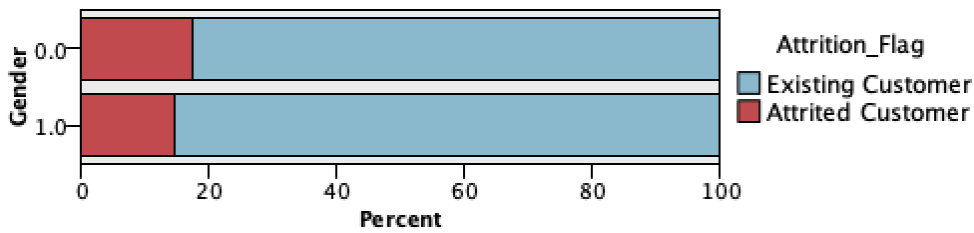


Figure 26 Gender-2 Distribution

Source: author's own work

Figures 25 and 26 show the distribution of data for the input variable "gender". The value 0 represents "female" and 1 represents "male". It can be seen that the number of females is 5358 or 52.91 per cent of the total number of observations and the number of males is 4769 or 47.09 per cent of the total number of observations. The number of males is 4769, representing 47.09 per cent of the total number of observations.

The visual distribution of the input variables is shown in Figure 25, where we can observe the number of females and males as a percentage of the total number of observations. It can be seen that this dataset is relatively balanced. It shows 52.91% female customers and 47.09% male customers. The credit card customer churn rate is higher among female customers than male.

Input variable: Education_Level

Value /	Proportion	%	Count
0.000		15.0	1519
1.000		14.68	1487
2.000		19.88	2013
3.000		10.0	1013
4.000		30.89	3128
5.000		5.1	516
6.000		4.45	451

Attrition_Flag



Figure 27 Education_Level-1 Distribution

Source: author's own work

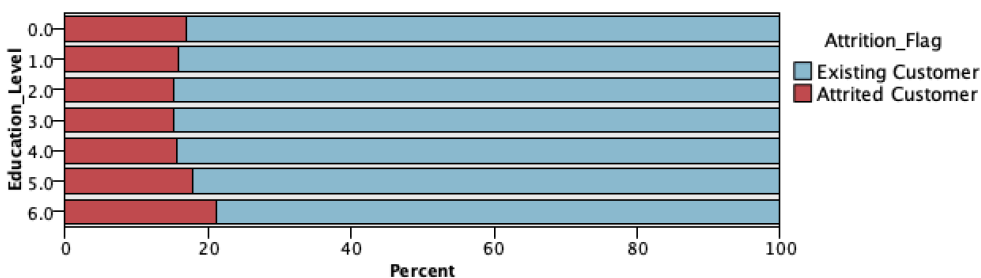


Figure 28 Education_Level-2 Distribution

Source: author's own work

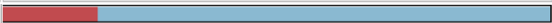
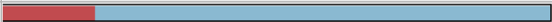
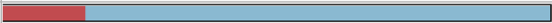
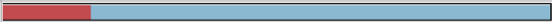
Education_Level	Frequency	Percentage of total
Unknown (0)	1519	15%
Uneducated (1)	1487	14.68%
High School (2)	2013	19.88%
College (3)	1013	10%
Graduate (4)	3089	31.28%
Post-Graduate (5)	516	5.1%
Doctorate (6)	451	4.45%

Table 2 Education_Level Distribution

Source: author's own work

The input variable is expressed in numerical rather than actual values, with each value indicating the category of the client's educational attainment, and there are seven categorical values. As can be seen in Figure 27 above, the largest percentage of clients is Graduate a total of 3089 (31.82%). However from Figure 28, we can get that the credit card customer churn percentage is slightly higher in Post-Graduate and Doctorate than the other classifications.

Input variable: Marital_Status

Value /	Proportion	%	Count
0.000		7.4	749
1.000		38.94	3943
2.000		46.28	4687
3.000		7.39	748

Attrition_Flag

 0.0

 1.0

Figure 29 Marital_Status-1 Distribution

Source: author's own work

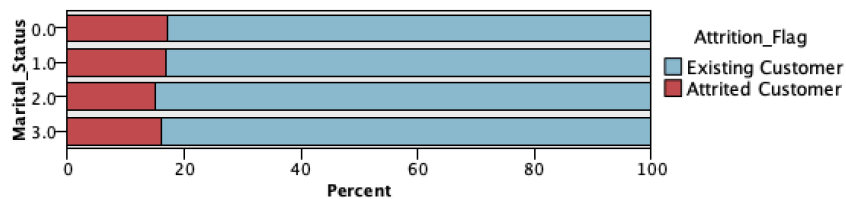


Figure 30 Marital_Status-2 Distribution

Source: author's own work

Marital_Status	Frequency	Percentage of total
Unknown (0)	749	7.4%
Single (1)	3943	38.94%







Married (2)	4687	46.28%
Divorced (3)	748	7.39%

Table 3 Marital_Status Distribution

Source: author's own work

The input variable is expressed in numerical rather than actual values, with each value indicating the category of the client's marital status, and there are four categorical values. From Figure 29 above, we see that the largest percentage of customers are married with a total of 4687 (46.28%), followed by single with a total of 3943 (38.94%). However from figure 30, we get that the lowest percentage of credit card customer churn is the married customers.

Input variable: Income_Category

Value #	Proportion	%	Count
0.000		10.98	1112
1.000		35.16	3561
2.000		17.68	1790
3.000		13.84	1402
4.000		15.16	1535
5.000		7.18	727

Attrition_Flag

 0.0  1.0

Figure 31 Income_Category-1 Distribution

Source: author's own work

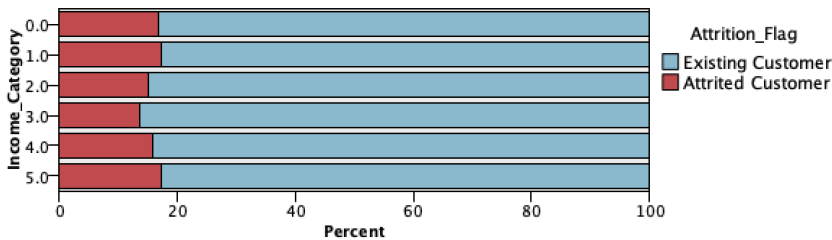


Figure 32 Income_Category-2 Distribution

Source: author's own work

Income_Category	Frequency	Percentage of total
Unknown (0)	1112	10.98%
Less than \$40K(1)	3561	35.16%
\$40K - \$60K (2)	1790	17.68%
\$60K - \$80K (3)	1402	13.84%
\$80K - \$120K (4)	1535	15.16%
\$120K + (5)	727	7.18%

Table 4 Income_Category Distribution

Source: author's own work

The input variable is expressed in numerical values rather than actual values, and each value represents the category of income status of the client, with six categorical values. From

Figure 31 above, the largest percentage of customers are Less than \$40K with a total of 3,561 (35.16%) and the least is \$120K + with a total of 727 7.18%. However, from Figure 32, the lowest percentage of credit card customer churn are customers with income of \$40K - \$60K and \$60K - \$80K.

Input variable: Card_Category

Value #	Proportion	%	Count
1.000		93.18	9436
2.000		5.48	555
3.000		1.15	116
4.000		0.2	20



Figure 33 Card_Category-1 Distribution
Source: author's own work

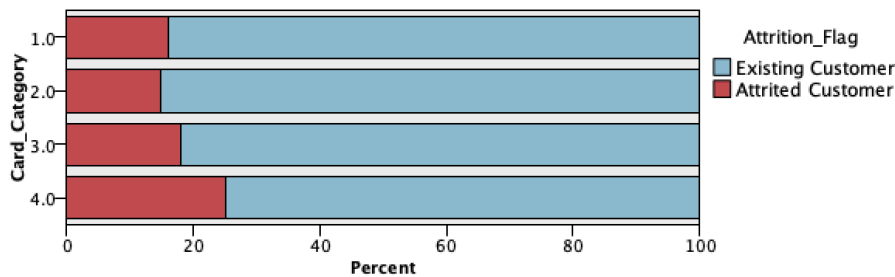


Figure 34 Card_Category-2 Distribution
Source: author's own work

Card_Category	Frequency	Percentage of total
Blue (1)	9436	93.18%
Silver (2)	555	5.48%
Gold (3)	116	1.15%
Platinum (4)	20	0.2%

Table 5 Card_Category Distribution
Source: author's own work

The input variable is expressed as numerical values rather than actual values and each value represents the customer credit card category with 4 categorical values. From Figure 33 above, it shows that the largest percentage of customers are Blue cards with a total of 9,436 (93.18%) and the smallest are Platinum cards with a total of 20 (0.2%).

However, from Figure 34, we observe that the lowest percentage of credit card customer churn is Silver card customers and the highest percentage of customer churn is Platinum card users.

4.4 Data preparation

4.4.1 Categorization - reclassification

Reclassify the input variable "Education_Level" based on similar proportions.

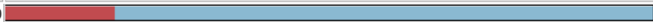












Value /	Proportion	%	Count
0.000		15.0	1519
1.000		14.68	1487
2.000		19.88	2013
3.000		10.0	1013
4.000		30.89	3128
5.000		5.1	516
6.000		4.45	451

Figure 35 Reclassify Education_Level-1

Source: author's own work

As shown in Figure 35 above, the Count of group 5 and group 6 are not more than 1000, which is far lower than the five groups of the Encounter, and their proportions are relatively similar, so we combined group 5 and group 6, and named the merged group as group 5. The distribution of the combined variable "Education_Level_Re" is shown in Figure 36 below.

Value /	Proportion	%	Count
0.000		15.0	1519
1.000		14.68	1487
2.000		19.88	2013
3.000		10.0	1013
4.000		30.89	3128
5.000		9.55	967

Attrition_Flag

 0.0

 1.0

Figure 36 Reclassify Education_Level-2

Source: author's own work

4.4.2 Partition

Before starting modeling the data, we need to partition the dataset into two parts: a training set and a testing set. The training portion of the dataset will be used to build the predictive model and train the regression tree model, while the testing portion will be used to validate the accuracy of the built predictive model and its various measurement data. This is important to validate the model on data that has not been used for modeling.

The best way for me to divide the dataset is to randomly select 30% of the dataset as the test set and the remaining 70% of the dataset as the training set. The commands used to generate the training/test dataset in IBM SPSS Modeler.

4.5 Modeling

4.5.1 Binary logistic regression model

We examined multicollinearity for all variables in the 70% training dataset using SPSS statistical software, which excludes the qualitative/categorical variables “Gender”, “Marital_Status”, “Education_Level”, “Income_Category” and “Card_Category”. By comparing the statistical VIF values in Figure 38, we found that none of the variables had a VIF value greater than 5, thus eliminating the effect of multicollinearity on the modelling of the logistic regression model.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	.940	.034		27.530	<.001		
	Age	-.001	.001	-.016	-1.035	.301	.380	2.630
	Dependent_count	.014	.003	.049	5.121	<.001	.974	1.027
	Months_on_book	.001	.001	.013	.853	.394	.383	2.611
	Total_Relationship_Count	-.043	.002	-.182	-17.844	<.001	.867	1.153
	Months_Inactive_12_mon	.047	.003	.130	13.665	<.001	.990	1.010
	Contacts_Count_12_mon	.041	.003	.123	12.751	<.001	.966	1.036
	Credit_Limit	-1.487E-6	.000	-.037	-2.931	.003	.569	1.759
	Total_Revolving_Bal	-9.804E-5	.000	-.218	-15.413	<.001	.451	2.217
	Total_Amt_Chng_Q4_Q1	-.045	.017	-.027	-2.571	.010	.843	1.187
	Total_Trans_Amt	3.358E-5	.000	.314	18.278	<.001	.306	3.272
	Total_Trans_Ct	-.010	.000	-.609	-36.899	<.001	.330	3.029
	Total_Ct_Chng_Q4_Q1	-.302	.016	-.197	-18.991	<.001	.834	1.200
	Avg_Utilization_Ratio	.007	.022	.005	.328	.743	.347	2.883

a. Dependent Variable: Attrition_Flag

Figure 37 Logistic Multicollinearity-VIF

Source: author's own work

Then we select 70% training dataset in SPSS modeler and build a binary logistic regression model named “Logistic Model 1” using the previously selected dataset variables.

Valid	7059	100.0%
Missing	0	
Total	7059	
Subpopulation	7059 ^a	

a. The dependent variable has only one value observed in 7059 (100.0%) subpopulations.

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	6222.361			
Final	3241.271	2981.089	30	<.001

Pseudo R-Square

Cox and Snell	.344
Nagelkerke	.588
McFadden	.479

Figure 38 Logistic The Omnibus Tests-model 1

Source: author's own work

The result of the binary logistic regression model 1 is shown in Figure 38, the total number of training data sets 7059 without missing values.

The Omnibus Tests:

From the results of the “model fit information”, we conclude that the likelihood ratio test is used to verify that the coefficients of all variables are zero at the same time. It shows whether $\beta_1=\beta_2=\dots=\beta_{19}=0$ is valid or not.

$$H_0: \beta_1=\beta_2=\dots=\beta_{19}=0$$

H_1 : at least one β not equal to 0

Results of the likelihood ratio test: Chi-Square = 2981.089 df = 30, p-value < .001. alpha = 0.05 > p-value, then H_0 will be rejected. There is at least one parameters statistically significant in model 1.

Nagelkerke's R-Square = 0.588 in Pseudo R-Square, which indicates that 58.8% of the target variable “Attrition_Flag” is explained by the input explanatory variable in regression model 1.

		Parameter Estimates					95% Confidence Interval for Exp(B)		
Attrition_Flag ^a		B	Std. Error	Wald	df	Sig.	Exp(B)	Lower Bound	Upper Bound
Attrited Customer	Intercept	6.824	.928	54.087	1	<.001			
	Age	-.013	.009	1.971	1	.160	.987	.969	1.005
	Dependent_count	.176	.037	23.197	1	<.001	1.192	1.110	1.281
	Months_on_book	.010	.009	1.059	1	.304	1.010	.991	1.028
	Total_Relationship_Count	-.459	.033	191.911	1	<.001	.632	.592	.674
	Months_Inactive_12_mon	.576	.046	155.295	1	<.001	1.779	1.625	1.948
	Contacts_Count_12_mon	.488	.044	124.334	1	<.001	1.629	1.495	1.775
	Credit_Limit	.000	.000	4.405	1	.036	1.000	1.000	1.000
	Total_Revolving_Bal	-.001	.000	130.144	1	<.001	.999	.999	.999
	Avg_Open_To_Buy	0 ^b	.	.	0
	Total_Amt_Chng_Q4_Q1	-.184	.224	.675	1	.411	.832	.536	1.290
	Total_Trans_Amt	.000	.000	294.079	1	<.001	1.000	1.000	1.001
	Total_Trans_Ct	-.119	.005	691.399	1	<.001	.887	.880	.895
	Total_Ct_Chng_Q4_Q1	-2.987	.230	168.350	1	<.001	.050	.032	.079
	Avg_Utilization_Ratio	-.035	.297	.014	1	.905	.965	.540	1.726
	[Gender=.000]	.744	.175	18.004	1	<.001	2.104	1.492	2.967
	[Gender=1.000]	0 ^b	.	.	0
	[Marital_Status=.000]	.192	.230	.694	1	.405	1.211	.772	1.902
	[Marital_Status=1.000]	.077	.184	.174	1	.677	1.080	.753	1.549
	[Marital_Status=2.000]	-.521	.184	8.060	1	.005	.594	.414	.851
	[Marital_Status=3.000]	0 ^b	.	.	0
	[Card_Category=1.000]	-.805	.729	1.218	1	.270	.447	.107	1.867
	[Card_Category=2.000]	-.505	.737	.469	1	.493	.604	.142	2.558
	[Card_Category=3.000]	.556	.811	.469	1	.493	1.743	.356	8.545
	[Card_Category=4.000]	0 ^b	.	.	0

Figure 39 Result-binary logistic regression model 1-1

Source: author's own work

[Income_Category=.000]	-.720	.277	6.763	1	.009	.487	.283	.838
[Income_Category=1.000]	-.747	.260	8.262	1	.004	.474	.285	.788
[Income_Category=2.000]	-.977	.240	16.610	1	<.001	.376	.235	.602
[Income_Category=3.000]	-.740	.212	12.126	1	<.001	.477	.315	.724
[Income_Category=4.000]	-.476	.197	5.816	1	.016	.622	.422	.915
[Income_Category=5.000]	0 ^b	.	.	0
[Education_level_Re=.000]	-.375	.184	4.185	1	.041	.687	.479	.984
[Education_level_Re=1.000]	-.289	.184	2.459	1	.117	.749	.522	1.075
[Education_level_Re=2.000]	-.325	.174	3.503	1	.061	.722	.514	1.015
[Education_level_Re=3.000]	-.500	.207	5.818	1	.016	.607	.404	.911
[Education_level_Re=4.000]	-.406	.163	6.199	1	.013	.667	.484	.917
[Education_level_Re=5.000]	0 ^b	.	.	0

a. The reference category is: Existing Customer.

b. This parameter is set to zero because it is redundant.

Figure 40 Result-binary logistic regression model 1-2

Source: author's own work

From Figures 39 and 40, which show the results of model 1, we find that some variables have Sig. values greater than $\alpha = 0.05$, which indicates that these variables are not statistically significant.

As shown in Figures, we therefore screened out variables in Model 1 with Sig. values less than $\alpha=0.05$. The results are as follows:

- Dependent_count
- Total_Relationship_Count
- Months_Inactive_12_mon
- Contacts_Count_12_mon
- Credit_Limit
- Total_Revolving_Bal
- Total_Trans_Amt
- Total_Trans_Ct
- Total_Ct_Chng_Q4_Q1
- Gender
- Income_Category
- Education_level_Re

Then, the 12th data will be introduced to set mentioned above and reconstructed a new binary logistic regression model named Model 2.

Valid	7059	100.0%
Missing	0	
Total	7059	
Subpopulation	7059 ^a	

a. The dependent variable has only one value observed in 7059 (100.0%) subpopulations.

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	6222.361			
Final	3297.353	2925.007	20	<.001

Pseudo R-Square

Cox and Snell	.339
Nagelkerke	.579
McFadden	.470

Figure 41 Result logistic model 2 - The Omnibus Tests
Source: author's own work

The result of The Omnibus Tests for the binary logistic regression model 2 is shown in Figure 41, where we can obtain the total number of training data sets without missing values as 7059.

The Omnibus Tests: From the results of "model fit information", we can get that the likelihood ratio test is used to verify that the coefficients of all variables are zero at the same time. It shows whether formula (12) is valid or not.

$$\beta_1 = \beta_2 = \beta_3 = \dots = \beta_{12} = 0 \tag{12}$$

$$H_0 = \beta_1 = \beta_2 = \beta_3 = \dots = \beta_{12} = 0 \tag{13}$$

H1: At least one β is not equal to 0

Results of the likelihood ratio test: Chi-Square = 2925.007 df = 20, p-value < .001 . alpha = 0.05 > p-value , then we reject H0. At least one of the models is statistically significant.

Nagelkerke's R-Square = 0.579 in Pseudo R-Square , which indicates that 57.9% of the target variable "Attrition_Flag" is explained by the 12 input explanatory variables in Logistic regression model 2.

		Parameter Estimates					95% Confidence Interval for Exp(B)		
Attrition_Flag ^a		B	Std. Error	Wald	df	Sig.	Exp(B)	Lower Bound	Upper Bound
Attrited Customer	Intercept	5.079	.378	180.881	1	<.001			
	Dependent_count	.178	.036	24.729	1	<.001	1.194	1.114	1.281
	Total_Relationship_Count	-.465	.033	200.828	1	<.001	.628	.589	.670
	Months_Inactive_12_mon	.577	.045	161.156	1	<.001	1.781	1.629	1.947
	Contacts_Count_12_mon	.482	.043	124.183	1	<.001	1.619	1.488	1.763
	Credit_Limit	.000	.000	.444	1	.505	1.000	1.000	1.000
	Total_Revolving_Bal	-.001	.000	332.227	1	<.001	.999	.999	.999
	Total_Trans_Amt	.000	.000	294.632	1	<.001	1.000	1.000	1.001
	Total_Trans_Ct	-.115	.004	680.765	1	<.001	.892	.884	.899
	Total_Ct_Chng_Q4_Q1	-3.072	.216	202.593	1	<.001	.046	.030	.071
	[Income_Category=.000]	-.540	.271	3.984	1	.046	.583	.343	.990
	[Income_Category=1.000]	-.505	.251	4.065	1	.044	.603	.369	.986
	[Income_Category=2.000]	-.756	.231	10.709	1	.001	.469	.298	.738
	[Income_Category=3.000]	-.578	.207	7.786	1	.005	.561	.374	.842
	[Income_Category=4.000]	-.401	.195	4.257	1	.039	.669	.457	.980
	[Income_Category=5.000]	0 ^b	.	.	0
	[Education_level_Re=.000]	-.329	.182	3.276	1	.070	.720	.504	1.028
	[Education_level_Re=1.000]	-.267	.183	2.139	1	.144	.766	.535	1.095
	[Education_level_Re=2.000]	-.328	.172	3.619	1	.057	.720	.514	1.010

Figure 42 Result-Logistic Model 2-1
Source: author's own work

[Education_level_Re=3.000]	-.487	.205	5.622	1	.018	.615	.411	.919
[Education_level_Re=4.000]	-.382	.161	5.613	1	.018	.682	.497	.936
[Education_level_Re=5.000]	0 ^b	.	.	0
[Gender=.000]	.709	.173	16.747	1	<.001	2.032	1.447	2.855
[Gender=1.000]	0 ^b	.	.	0

a. The reference category is: Existing Customer.
b. This parameter is set to zero because it is redundant.

Figure 43 Result-Logistic Model 2-2
Source: author's own work

Estimate "Attrition_Flag=1"

The value of the variable Credit_Limit in Figure 44 is -0.000004139097777 and the value of the variable Total_Trans_Amt is 0.000464157015285

According to Figure 44 and 45, we can obtain the logistic regression of model 2 as:

$$\ln\left(\frac{P}{1-P}\right) = 5.079 + 0.178 * Dependent_Count - 0.465$$

$$* Total_Relationship_Count + 0.577 * Months_Inactive_12_mon$$

$$+ 0.482 * Contacts_Count_12_mon - 0.000004 * Credit_Limit$$

$$- 0.001 * Total_Revolving_Bal + 0.00046 * Total_Trans_Amt$$

$$- 0.0115 * Total_Trans_Ct - 3.072 * Total_Ct_Chng_Q4_Q1 - 0.54$$

$$* Income_category - 0.329 * Education_level_Re + 0.709$$

$$* Gender \tag{14}$$

In the results we see that Sig., basically all the Sig. values are less than alpha = 0.05, which means that almost all of them are statistically significant, for Credit_Limit Sig. = 0.505 but the result can be accepted.

For input variable the value of Exp(B) if it is greater than 1, it means that the probability of risk leading to Attrition_Flag = 1 increases; If it is less than 1, it means that the probability of causing the risk Attrition_Flag=1 decreases.

For variable [Income_category=1.000] and [Education_level_Re=.000], When we look at result "B", positive shows increase risk of Attrition_Flag=Attrition(P=1), negative shows decrease risk of Attrition_Flag=Attrition(P=1). Following this we can say:

The influencing factors that are positively related to attrition are:

Dependent_count(1.194 times), Contacts_Count_12_mon (1.619 times), Total_Trans_Amt(0.000464 times).

The influencing factors that are negatively correlated with customer churn are:

Total_Relationship_Count (0.628 times), Months_Inactive_12_mon (1.781 times), Credit_Limit (0.00000413 times.), Total_Revolving_Bal (0.999 times), Total_Trans_Ct(0.892 times.), Total_Ct_Chng_Q4_Q1(0.046 times).

Evidence found that the variables Months_Inactive_12_mon and Contacts_Count_12_mon are very significant on credit card customer churn, and also coincide with the explanatory variable factors mentioned in the literature review, where the less active a customer is the more likely they are to churn. Banks need to be efficient when communicating with their customers, reducing the number of ineffective communication with customers and telephone advertising harassment has a positive impact on customer churn. The higher Months_Inactive_12_mon, the higher chance to be Attrition_Flag

=Attrition (P=1). Decreasing Exp(B)= 1.781 times. The higher Contacts_Count_12_mon, the higher chance to be Attrition_Flag =Attrition (P=1). Increasing Exp(B)=1.619 times.

Interestingly we also found that Credit_Limit has a very small effect on credit card customer churning, The higher Credit_Limit, the lower chance to be Attrition_Flag=Attrition(P=1). Decreasing Exp(B)=0.00000413 times.

When all other variables are held constant, for the variable "Income_Category". Evidence shows that the larger the Income_category, the higher the probability of Attrition. This shows that Income_category has a positive effect on the probability of customer churn and an increase in Income_category leads to an increase in the probability of attrition. This is in line with the analysis in the literature review.

With all other variables held constant, for the variable "Education_Level We found that a higher level of education leads to a higher probability of customer churn. This shows that the Education_Level has a positive effect on the probability of churn, and increasing the Education_Level leads to an increase in the probability of Attrition. This is consistent with the analyses in the literature review.

4.5.2 Model Evaluation – Binary logistic regression

The importance of each variable in the forecast is shown in Figure 46 below, although all 12 variables are statistically significant. The variable “Total_Trans_Ct” is more important than the other 11 variables for predicting "Attrition_Flag=1".

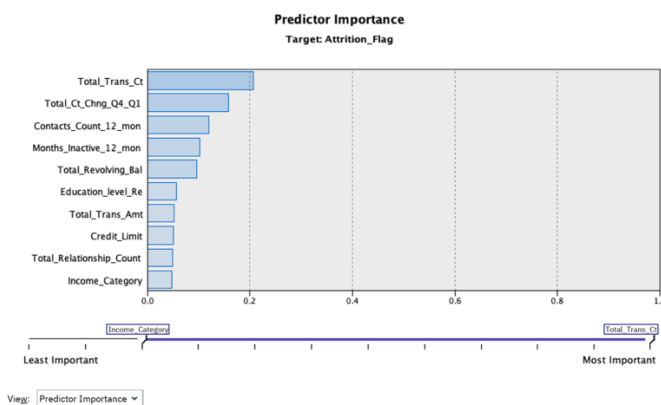


Figure 44 Predictor Importance- logistic model 2
Source: author's own work

	Attrition_Flag	\$L-Attrition_Flag	\$LP-Attrition_Flag	\$LRP-Attrition_Flag
1	0.000	0.000	0.997	0.003
2	0.000	0.000	1.000	0.000
3	0.000	0.000	0.994	0.006
4	0.000	0.000	0.999	0.001
5	0.000	0.000	0.844	0.156
6	0.000	0.000	0.948	0.052
7	0.000	0.000	0.857	0.143
8	0.000	0.000	0.992	0.008
9	0.000	0.000	0.919	0.081
10	0.000	0.000	0.858	0.142
11	0.000	0.000	0.986	0.014
12	0.000	0.000	0.999	0.001
13	0.000	0.000	0.924	0.076
14	0.000	0.000	0.923	0.077
15	0.000	0.000	0.955	0.045
16	0.000	0.000	0.871	0.129
17	0.000	0.000	0.978	0.022
18	0.000	0.000	0.946	0.054
19	0.000	1.000	0.897	0.897
20	0.000	1.000	0.767	0.767
21	0.000	0.000	0.895	0.105
22	0.000	0.000	0.758	0.242
23	0.000	0.000	0.998	0.002
24	0.000	0.000	0.998	0.002

Figure 45 Table of logistic model 2

Source: author's own work

We run the results of logistic Model 2 with Table and get three new data columns, as shown in Figure 45:

\$L : predicted (target variable=1)

\$LP confidence

\$LRP: propensity score , if >0.5 belong to P(Attrition_Flag)= 1 , <0.5 belong to P=0

For observation 19, predicted its 1, propensity score its 0.897>0.5, then its automatic filter to Attrition Customer Yes. Confidence=0.897, there is 89.7% confidence that predicted result belong to Attrition Customer Yes “P=1”

		\$L-Attrition_Flag	
Attrition_Flag		0.0	1.0
0.0	Count	5720	205
	Row %	96.540	3.460
1.0	Count	474	660
	Row %	41.799	58.201

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 2,652.681, df = 1, probability = 0

Figure 46 Confusion Matrix – logistic Model 2

Source: author's own work

As shown in Figure 46, we create the Confusion matrix using the actual Attrition_Flags and the \$L predicted Attrition_Flags, with the graph showing TN=5720; FP=205; FN=474; TP=660, we can obtain:

TPR-Sensitivity is related to “P=1”, $\frac{TP}{FN+TP}=58.201\%$

TNR-Specificity is related to “P=0”, $\frac{TN}{TN+FP}=96.54\%$

$$Accuracy = \frac{TP + TN}{FN + TP + TN + FP} = 90.38\% \quad (15)$$

Results for output field Attrition_Flag

Individual Models

Comparing \$L-Attrition_Flag with Attrition_Flag

'Partition'	1_Training	
Correct	6,380	90.38%
Wrong	679	9.62%
Total	7,059	

Evaluation Metrics

'Partition'	1_Training	
Model	AUC	Gini
\$L-Attrition_Flag	0.926	0.853

Figure 47 Analysis model 2 result
Source: author's own work

We run the results of logistic model 2 with Analysis. The results are shown in Figure 47, AUC [0,1], The closer the AUC is to 1.0, the higher the authenticity of the test method.

AUC=0.926 is greater than 0.9 which is pretty high. This indicates that trained prediction model has 92.6% prediction truth

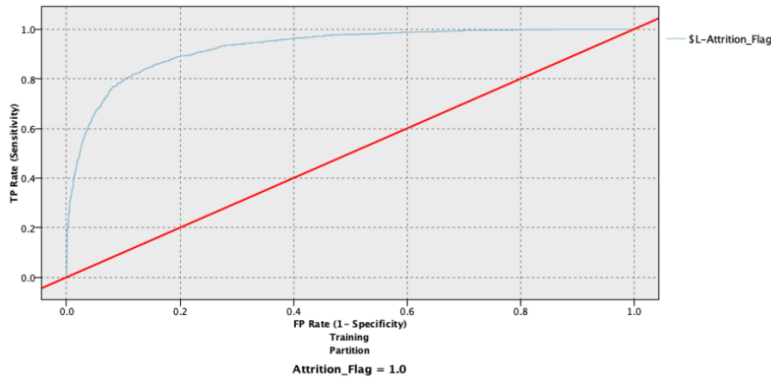


Figure 48 ROC curve for logistic model 2
Source: author's own work

We run the results of logistic model 2 in Evaluation to get the ROC curve as shown in Figure 48.

4.5.3 Decision trees models - CHAID tree models

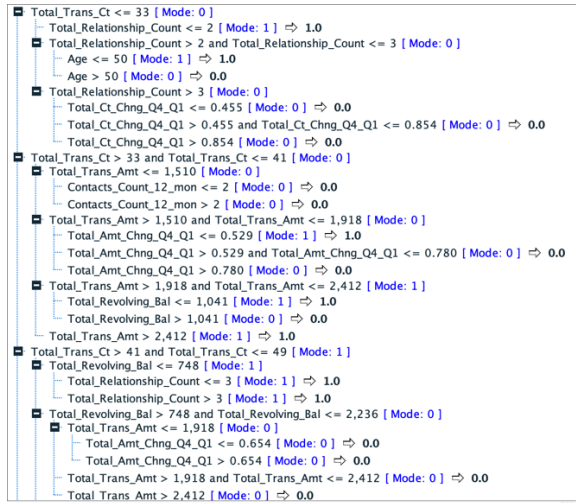


Figure 49 CHAID-2

Source: author's own work

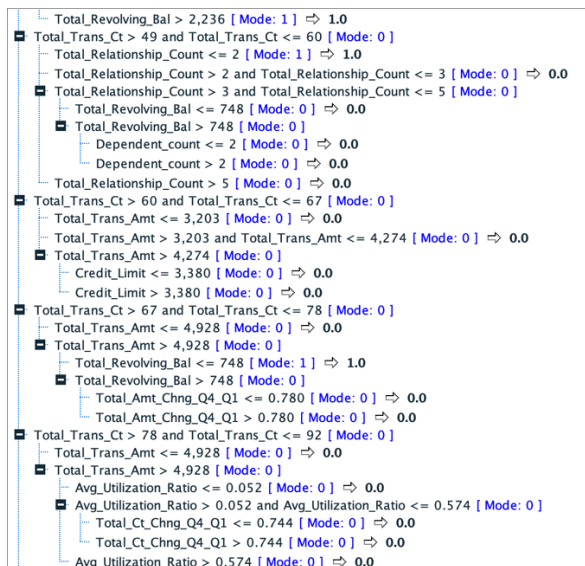


Figure 50 CHAID-3

Source: author's own work



Figure 51 CHAID-3

Source: author's own work

Observations with a lower Total_Revolving_Bal will have a greater probability of customer attrition. Observations with a lower age will have a greater probability of customer attrition. Observations with a lower Total_Relationship_Count will have a better chance of customer churn.

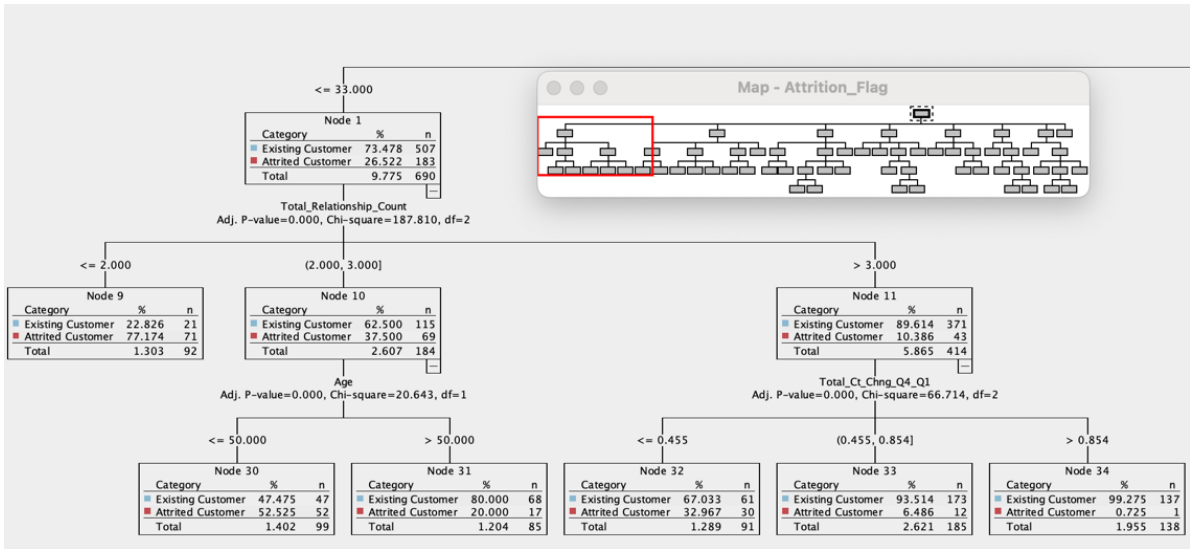


Figure 52 CHAID tree-left 1
Source: author's own work

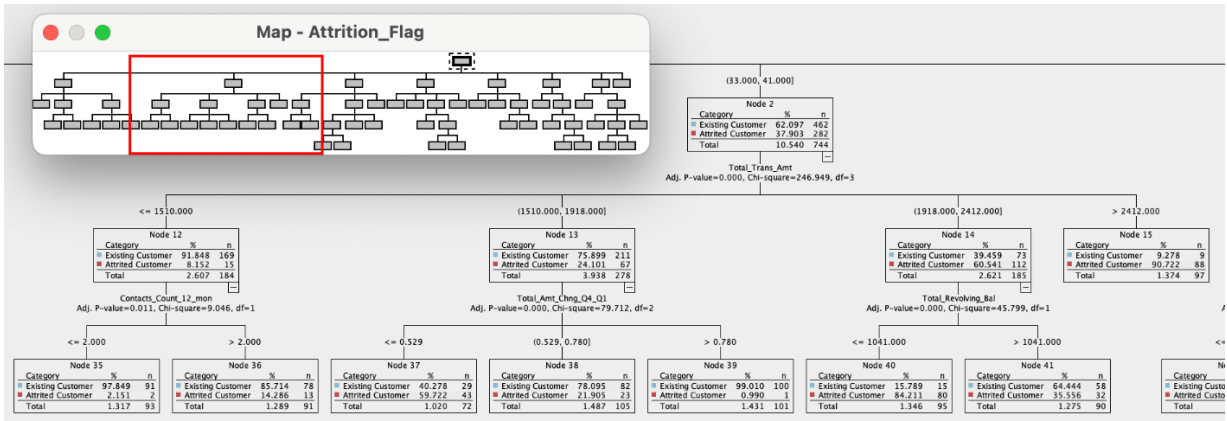


Figure 53 CHAID tree-left 2
Source: author's own work

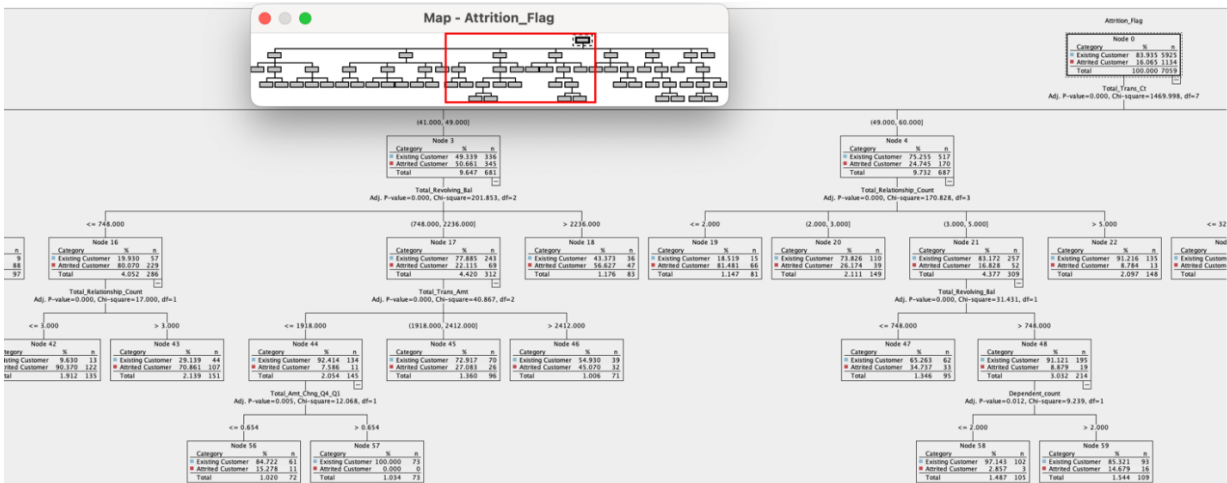


Figure 54 CHAID tree-right 1
Source: author's own work

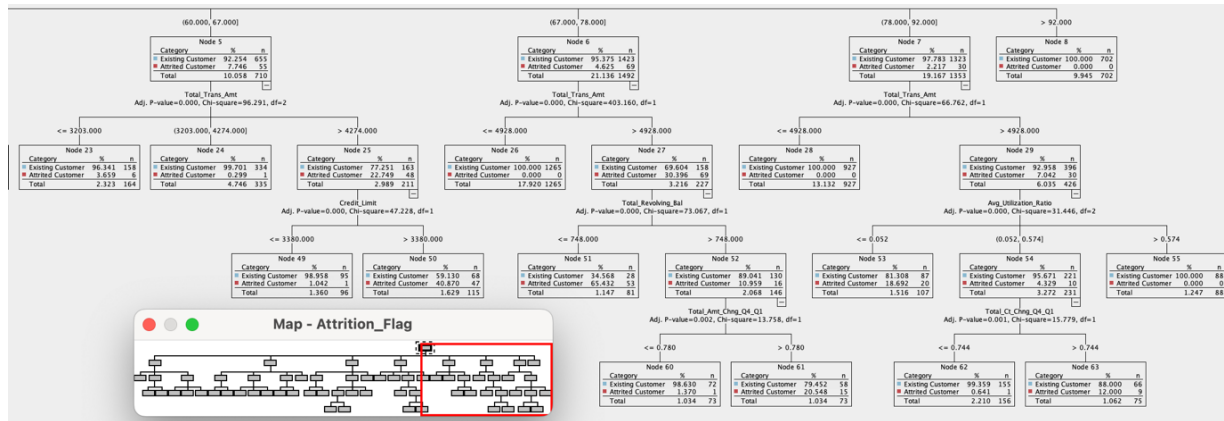


Figure 55 CHAID tree-right 2

Source: author's own work

Total 7059 base on 5925 observations with “Existing Customer=0” and 1134 observations with “Attrited Customer=1”

As shown in figures 52-55 above, The highest chi-square value = 1469.998 determines that the first node branches with the variable Total_Trans_Ct. At the next level, Node 1 the chi-square value = 187.81 determines branching with the variable Total_Relationship_Count. Node 2 the chi-square value = 246.949 determines branching with the variable Total_Trans_Amt. Node 3 the chi-square value = 201.853 determines branching with the variable Total_Revolving_Bal. Node 4 branches to the variable Total_Relationship_Count, and nodes 5, 6, and 7 branch to the variable Total_Trans_Amt. CHAID tree modeling results show: Total_trans_Ct is the most important categorical variable as preferred. Total_Relationship_Count, Total_Trans_Amt and Total_Revolving_Bal are the second most important branches of the CHAID tree model after Total_trans_Ct.

4.5.4 CHAID tree model evaluation

Table Annotations

	Attrition_Flag	\$R-Attrition_Flag	\$RC-Attrition_Flag	\$RRP-Attrition_Flag
1	0.000	0.000	0.987	0.013
2	0.000	0.000	0.986	0.014
3	0.000	0.000	0.986	0.014
4	0.000	0.000	0.986	0.014
5	0.000	1.000	0.525	0.525
6	0.000	0.000	0.930	0.070
7	0.000	0.000	0.981	0.019
8	0.000	0.000	0.986	0.014
9	0.000	0.000	0.986	0.014
10	0.000	0.000	0.987	0.013
11	0.000	0.000	0.986	0.014
12	0.000	0.000	0.986	0.014
13	0.000	0.000	0.986	0.014
14	0.000	0.000	0.986	0.014
15	0.000	1.000	0.766	0.766
16	0.000	0.000	0.986	0.014
17	0.000	0.000	0.986	0.014
18	0.000	0.000	0.986	0.014
19	0.000	1.000	0.525	0.525
20	0.000	0.000	0.930	0.070
21	0.000	0.000	0.930	0.070
22	0.000	0.000	0.986	0.014
23	0.000	0.000	0.968	0.032

Figure 56 Table CHAID Tree Training
Source: author's own work

\$R-Attrition_Flag

Attrition_Flag		0.0	1.0
0.0	Count	5668	257
	Row %	95.662	4.338
1.0	Count	405	729
	Row %	35.714	64.286

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 2,846.532, df = 1, probability = 0

Figure 57 CHAID tree Matrix Training
Source: author's own work

As shown in Figure 57, the Confusion matrix using the actual Attrition Flags and the \$R1 predicted Attrition Flags, with the graph showing TN=5668; FP=257; FN=405; TP=729, we can obtain:

TPR-Sensitivity is related to “P=1”, $\frac{TP}{FN+TP}=64.286\%$

TNR-Specificity is related to “P=0”, $\frac{TN}{TN+FP}=95.662\%$

$$Accuracy = \frac{TP + TN}{FN + TP + TN + FP} = 90.62\% \quad (16)$$

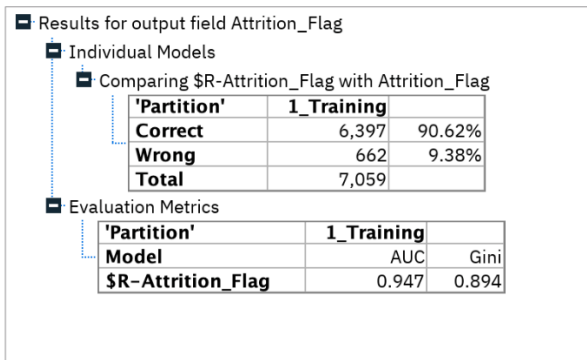


Figure 58 CHAID tree Accuracy Analysis

Source: author's own work

The AUC of the CHAID model was 0.947, indicating that this was a good fit and that the model predicted with high accuracy(>0.90).

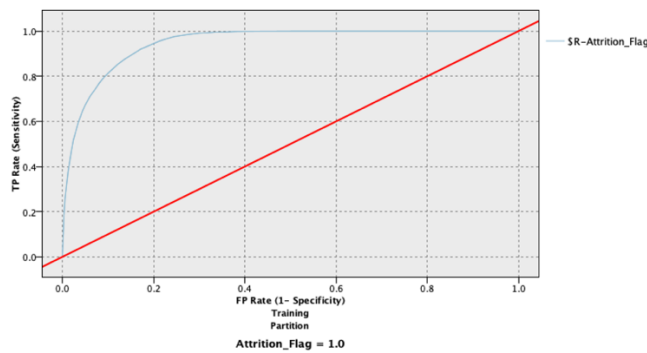


Figure 59 CHAID AUC curve

Source: author's own work

4.5.5 Decision trees models - C&R tree model

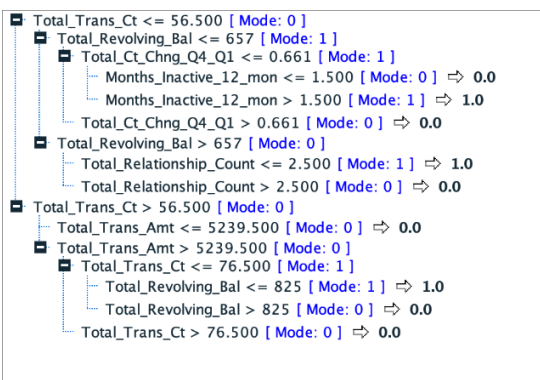


Figure 60 C&R tree-1

Source: author's own work

Observations with a higher Months_Inactive_12_mon have a greater chance of customer churn. Observations with lower Total_Relationship_Count have a greater chance of attrition. Observations with lower Total_Revolving_Bal have a better chance of Customer Attrition.

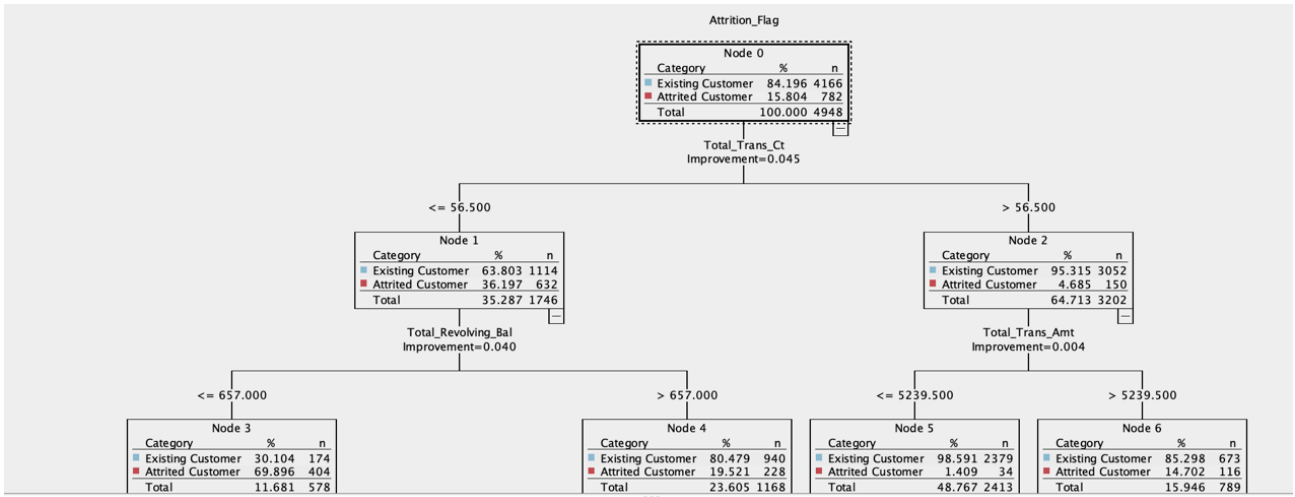


Figure 61 C&R tree-top
Source: author's own work

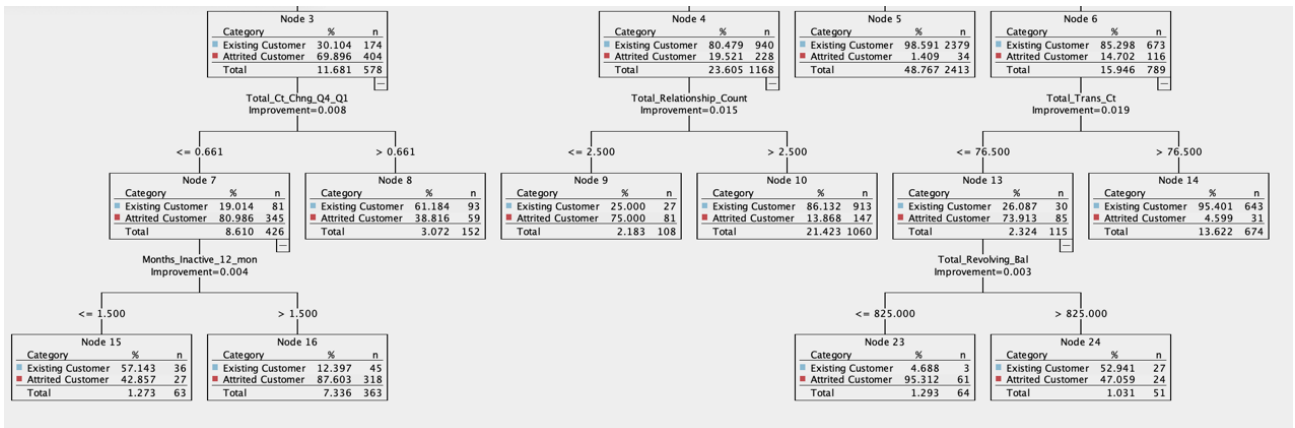


Figure 62 C&R tree-down
Source: author's own work

As shown in Figures 61 and 62 above, the highest value of Improvement = 0.045 determines that the first node branches with the variable Total_Trans_Ct. At the next level, node 1, the value of Improvement = 0.04 determines branching with the variable Total_Revolving_Bal. Node 2, the value of Improvement = 0.004 determines branching with the variable Total_Trans_Amt. In the third level of nodes, the value of root node 3 Improvement = 0.008 determines branching with the variable Total_Ct_Chng_Q4_Q1. The value of root node 4 Improvement = 0.0015 determines branching with the variable Total_Relationship_Count. The value of root node 6 Improvement = 0.0019 determines branching with the variable Total_Trans_Ct.

In the fourth level of nodes, the value of root node 7 Improvement = 0.004 determines branching with the variable Months_Inactive_12_mon. The value of root node 13 Improvement = 0.003 determines branching with the variable Total_Revolving_Bal.

C&R tree modeling results show: Total_trans_Ct is the most important categorical variable as preferred. Total_Trans_Amt and Total_Revolving_Bal are the second most important branching variables in the C&R tree model.

For both CHAID tree model and C&R tree model Total_trans_Ct is the most important categorical variable as preferred.

Total_Relationship_Count, Total_Trans_Amt and Total_Revolving_Bal are the second most important branches of the CHAID tree model after Total_trans_Ct.

Total_Trans_Amt and Total_Revolving_Bal are the second most important branching variables in the C&R tree model.

4.5.6 C&R tree model evaluation

Table Annotations

	Attrition_Flag	\$R-Attrition_Flag	\$RC-Attrition_Flag	\$RRP-Attrition_Flag
1	0.000	0.000	0.861	0.139
2	0.000	0.000	0.861	0.139
3	0.000	0.000	0.610	0.390
4	0.000	0.000	0.610	0.390
5	0.000	0.000	0.861	0.139
6	0.000	0.000	0.861	0.139
7	0.000	1.000	0.745	0.745
8	0.000	0.000	0.861	0.139
9	0.000	0.000	0.861	0.139
10	0.000	0.000	0.861	0.139
11	0.000	0.000	0.861	0.139
12	0.000	0.000	0.861	0.139
13	0.000	0.000	0.861	0.139
14	0.000	0.000	0.861	0.139
15	0.000	1.000	0.745	0.745
16	0.000	0.000	0.861	0.139
17	0.000	0.000	0.861	0.139
18	0.000	0.000	0.861	0.139
19	0.000	0.000	0.861	0.139
20	0.000	0.000	0.610	0.390
21	0.000	0.000	0.861	0.139

Figure 63 C&R tree table Training

Source: author's own work

Terminal node 10 shows 1st observation , P(Attrition)= 0.861%, \$RC= 0.861, \$RRP-Attrition_Flag=0.139 which showing in table above. This is probability with “branch”, Total_Trans_Ct<=56.5, Total_Revolving_Bal>0.657, Total_Relationship_Count>2.5.

		\$R-Attrition_Flag	
Attrition_Flag		0.0	1.0
0.0	Count	5812	113
	Row %	98.093	1.907
1.0	Count	452	682
	Row %	39.859	60.141

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 3,229.816, df = 1, probability = 0

Figure 64 C&R tree matrix Training

Source: author's own work

As shown in Figure 64, we create the Confusion matrix using the actual Attrition Flags and the \$R1 predicted Attrition Flags, with the graph showing TN=5812; FP=113; FN=452; TP=682, we can see that:

$$\text{TPR-Sensitivity is related to "P=1"}, \frac{TP}{FN+TP} = 60.141\%$$

$$\text{TNR-Specificity is related to "P=0"}, \frac{TN}{TN+FP} = 98.093\%$$

$$\text{Accuracy} = \frac{TP + TN}{FN + TP + TN + FP} = 92\% \quad (17)$$

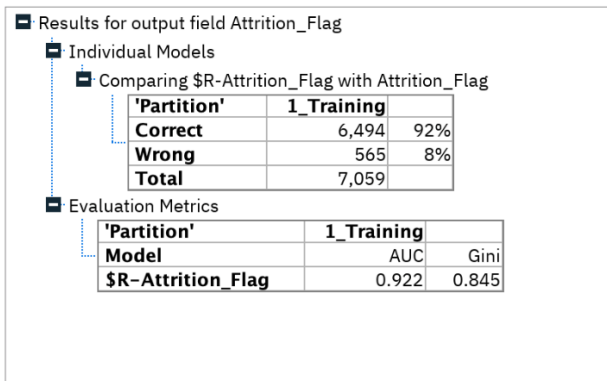


Figure 65 C&R tree Accuracy Analysis Training
Source: author's own work

The AUC of the C&R model was 0.922, indicating that this was a good fit and that the model predicted with high accuracy(>0.9).

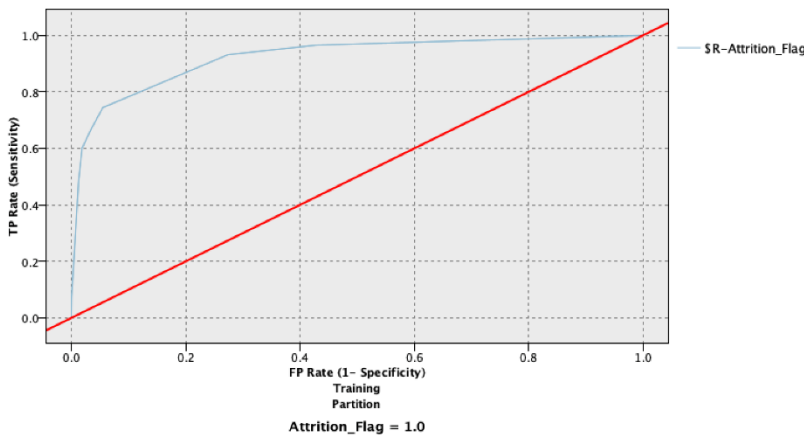


Figure 66 C&R tree AUC curve Training
Source: author's own work

5. Model comparison and final evaluation- Testing

To evaluate the models, we first created a "Test" selection in Partition Generation, then connected the three model outputs created with the training dataset to the "Test" selection.

5.1 Confusion Matrix

We used Matrix to create confusion matrices for the target variable Attrition_Flags and the three model outputs for the prediction outcome \$L, \$R and \$R1.

5.1.1 Logistic regression testing

Attrition_Flag		0.0	1.0
0.0	Count	2469	106
	Row %	95.883	4.117
1.0	Count	195	298
	Row %	39.554	60.446

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 1,148.258, df = 1, probability = 0

Figure 67 Confusion Matrix- Testing logistic regression model 2

Source: author's own work

As shown in Figure 67, the Confusion matrix using the actual Attrition Flags and the \$L predicted Attrition Flags, with the graph showing TN=2469; FP=106; FN=195; TP=298, we can obtain:

TPR-Sensitivity is related to "P=1", $\frac{TP}{FN+TP}=60.446\%$

TNR-Specificity is related to "P=0", $\frac{TN}{TN+FP}=95.883\%$

$$Accuracy = \frac{TP + TN}{FN + TP + TN + FP} = 90.19\% \quad (18)$$

5.1.2 CHAID tree model testing

Attrition_Flag		0.0	1.0
0.0	Count	2460	115
	Row %	95.534	4.466
1.0	Count	203	290
	Row %	41.176	58.824

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 1,067.02, df = 1, probability = 0

Figure 68 Confusion Matrix-Testing CHAID model

Source: author's own work

As shown in Figure 68, the Confusion matrix using the actual Attrition Flags and the \$R predicted Attrition Flags, with the graph showing TN=2460; FP=115; FN=203; TP=290, we can obtain:

TPR-Sensitivity is related to “P=1”, $\frac{TP}{FN+TP}=58.824\%$

TNR-Specificity is related to “P=0”, $\frac{TN}{TN+FP}=95.534\%$

$$Accuracy = \frac{TP + TN}{FN + TP + TN + FP} = 89.63\% \quad (19)$$

5.1.3 C&R tree model testing

Attrition_Flag		0.0	1.0
0.0	Count	2525	50
	Row %	98.058	1.942
1.0	Count	208	285
	Row %	42.191	57.809

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 1,327.745, df = 1, probability = 0

Figure 69 Confusion Matrix- Testing C&R model

Source: author's own work

As shown in Figure 69, the Confusion matrix using the actual Attrition Flags and the \$R1 predicted Attrition Flags, with the graph showing TN=2525; FP=50; FN=208; TP=285, we can obtain:

TPR-Sensitivity is related to “P=1”, $\frac{TP}{FN+TP}=57.809\%$

TNR-Specificity is related to “P=0”, $\frac{TN}{TN+FP}=98.058\%$

$$Accuracy = \frac{TP + TN}{FN + TP + TN + FP} = 91.59\% \quad (20)$$

5.2 AUC, Gini and ROC curve

Results for output field Attrition_Flag

Individual Models

Comparing \$L-Attrition_Flag with Attrition_Flag

'Partition'	1_Training	
Correct	6,380	90.38%
Wrong	679	9.62%
Total	7,059	

Comparing \$R-Attrition_Flag with Attrition_Flag

'Partition'	1_Training	
Correct	6,397	90.62%
Wrong	662	9.38%
Total	7,059	

Comparing \$R1-Attrition_Flag with Attrition_Flag

'Partition'	1_Training	
Correct	6,494	92%
Wrong	565	8%
Total	7,059	

Agreement between \$L-Attrition_Flag \$R-Attrition_Flag \$R1-Attrition_Flag

'Partition'	1_Training	
Agree	6,128	86.81%
Disagree	931	13.19%
Total	7,059	

Comparing Agreement with Attrition_Flag

'Partition'	1_Training	
Correct	5,899	96.26%
Wrong	229	3.74%
Total	6,128	

Evaluation Metrics

'Partition'	1_Training	AUC	Gini
Model			
\$L-Attrition_Flag		0.926	0.853
\$R-Attrition_Flag		0.947	0.894
\$R1-Attrition_Flag		0.922	0.845

Figure 70 Comparing training model results: AUC and Gini
Source: author's own work

Results for output field Attrition_Flag

Individual Models

Comparing \$L-Attrition_Flag with Attrition_Flag

'Partition'	2_Testing	
Correct	2,767	90.19%
Wrong	301	9.81%
Total	3,068	

Comparing \$R-Attrition_Flag with Attrition_Flag

'Partition'	2_Testing	
Correct	2,750	89.63%
Wrong	318	10.37%
Total	3,068	

Comparing \$R1-Attrition_Flag with Attrition_Flag

'Partition'	2_Testing	
Correct	2,810	91.59%
Wrong	258	8.41%
Total	3,068	

Agreement between \$L-Attrition_Flag \$R-Attrition_Flag \$R1-Attrition_Flag

'Partition'	2_Testing	
Agree	2,648	86.31%
Disagree	420	13.69%
Total	3,068	

Comparing Agreement with Attrition_Flag

'Partition'	2_Testing	
Correct	2,542	96%
Wrong	106	4%
Total	2,648	

Evaluation Metrics

'Partition'	2_Testing	AUC	Gini
Model			
\$L-Attrition_Flag		0.914	0.827
\$R-Attrition_Flag		0.928	0.856
\$R1-Attrition_Flag		0.918	0.837

Figure 71 Comparing testing model results: AUC and Gini
Source: author's own work

AUC provides an aggregate measure of performance across all possible classification thresholds, AUC ranges in value from 0 to 1

- Model whose predictions are 100% wrong → AUC of 0.0
- Model whose predictions are 100% correct → AUC of 1.0

The closer the AUC is to 1 the higher the accuracy of the model prediction

For the training dataset as shown in Figure 70 above, AUC for , $\$L=0.926$, $\$R=0.947$, $\$R1=0.922$,The prediction results of the three models are very close to each other in the range of 0.92-0.94 This means that the results of all three models are very good and exceed 0.9. But the 2nd model CHAID tree model has the highest prediction result AUC=0.947.

For the testing dataset as shown in Figure 71 above, AUC for , $\$L=0.914$, $\$R=0.928$, $\$R1=0.918$,The prediction results of the three models are very close to each other in the range of 0.91-0.93 This means that the results of all three models are very good and exceed 0.9. But the 2nd model CHAID tree model has the highest prediction result AUC=0.928.

We found the AUC of CHAID Tree model The AUC result is the highest in both the training model and the test model.

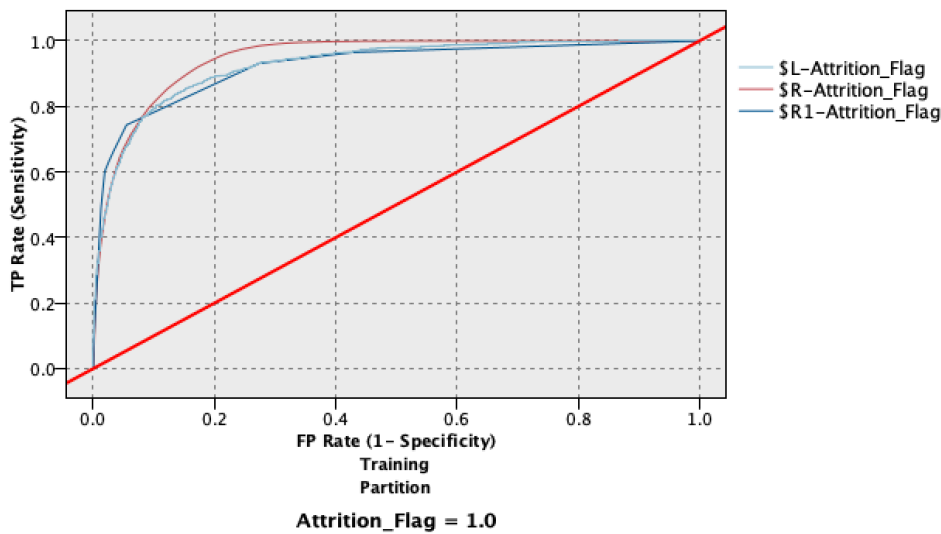


Figure 72 All ROC curve-Training

Source: author's own work

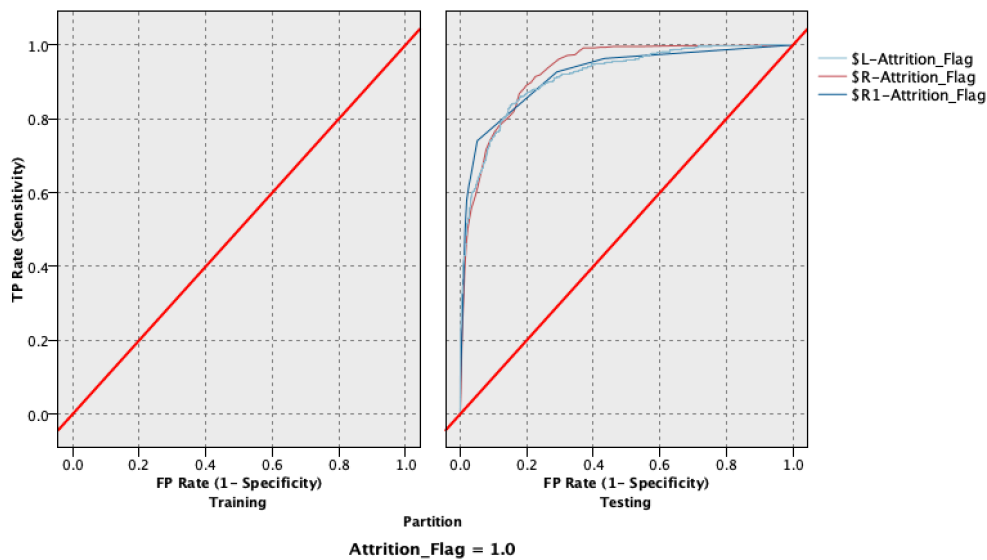


Figure 73 All ROC curve-Testing

Source: author's own work

The Receiver Operator Characteristic (ROC) curve is evaluation metric for binary classification problems

- it is a probability curve that plots the **TPR** against **FPR** at various threshold values
- it shows the performance of a classification model at all classification thresholds

TPR= true positive rate (Sensitivity)

FPR=false positive rate (1- Specificity)

The above graph shows that the ROC curves of the three models are very close to each other. But the C&D model is the largest.

As shown in Figures 72 and 73 above, comparing the modelling ROC results for both the training dataset and the test dataset shows that the ROC curve for the CHAID tree model is higher than the other two models (logistic regression and the C&R tree model). Comparing the results of the two models on the training dataset and the test dataset we can conclude: the CHAID tree model is slightly better than the other two models.

6. Result and Discussion

For both CHAID tree model and C&R tree model Total_trans_Ct is the most important categorical variable as preferred.

Total_Relationship_Count, Total_Trans_Amt and Total_Revolving_Bal are the second most important branches of the CHAID tree model after Total_trans_Ct.

Total_Trans_Amt and Total_Revolving_Bal are the second most important branching variables in the C&R tree model.

Comparing in the logistic regression model, the area of AUC under the curve is equal to 0.914, which is slightly lower than the area of AUC in the decision tree CHAID model of 0.928. for the prediction of big data, the difference in accuracy is 2% lower, Comparison with the findings of (Wagh et al., 2024) in the previous literature review. so we obtain that the decision tree model is slightly more accurate than the logistic regression model (1.4%).

Considering the overall specification and purpose of the study, in its current form, the logistic regression results are superior to the results of the decision tree CHAID and C&R models for warning banks and credit card organizations by predicting credit card customer churn, even though their overall accuracy values are slightly lower than those of the decision tree CHAID model. The decision tree CHAID model is characterized by more accurate predictions, while logistic regression provides a more detailed and accurate statistical analysis of risk factors, despite slightly lower predictions (1.4%). Considering that credit card customer churn is delayed and reversible in banking enterprises, the logistic regression model seems to be more appropriate here, because predicting the probability of customer churn occurring in advance through big data and data mining, human intervention in communicating with churn-prone customers and providing personalized and customized services, and timely allocation and strategic adjustments based on the feedback from churn-prone customers can reduce the probability of churn occurring by increasing customer satisfaction while decreasing customer The probability of churn is more favorable to market competition and market share capture and protection than other banks and credit card organizations that do not make predictions, reducing potential economic losses and brand impact losses for banking firms.

By observing the coefficients of the various influencing factors in the three models, we are able to explain the impact of each factor on the final results. Therefore, we can compare the results obtained from the models with the characteristics of each factor mentioned in the theoretical section.

- Total_trans_Ct
- Total_Ct_Chng_Q4_Q1
- Contacts_Count_12_mon
- Months_Inactive_12_mon
- Total_Revolving_Bal
- Eduaction_lvl_Re
- Total_Trans_Amt
- Credit_Limit
- Total_Relationship_Count
- Income_Category

From the regression model we can see that Total transaction count is a very important factor and the most important factor leading to customer churn as mentioned earlier. Therefore, as Total transaction count grows, the chances of customer churn decreases. Comparison with the findings of (Lin et al., 2011), (Cronin Jr et al., 2000) and (Ramzi & Mohamed, 2010) in the previous literature review showed consistent results. Increased customer satisfaction also stimulates higher frequency of credit card usage, which in turn reduces the probability of customer churn. This is consistent with the results in this case. In this case, the effects of the variables in the model are as expected.

Total count changed from quarter 4 to quarter 1 describes the change from quarter 4 to quarter 1, which is inversely related to churn in the model; the larger the change, the lower the risk of churn, but it also implies that churn can be reduced by offering more benefits in the fourth quarter to increase customer spending. Comparison with the findings of (Lin et al., 2011) and (Mahajan et al., 2017), in the previous literature review. The quality of the service will increase customer satisfaction, which will increase customer loyalty, and changes in the frequency of using the service will tend to stabilize, which in turn will reduce customer churn. This is consistent with the conclusions in this case.

Number of contacts customer has had in the last twelve months is a measure of customer satisfaction with banks and credit card organizations. It is directly proportional to the chances of churn in the model, which means that the more contacts with the bank and credit card organization, the higher the risk of churn. Customers who are more satisfied with their credit card services tend to contact their banks less because there is nothing to give feedback or complain about.

Number of months customer has been inactive in the last twelve months describes the change in customer activity. It is proportional to churn in the model, which means that the more active a credit card customer is, the lower the risk of churn. Customers who are more satisfied with their credit card services tend to use them more often because nowadays we spend money all the time and credit cards are more portable and easier to use compared to cash.

Total revolving balance of customer is inversely proportional to the churn rate in the model; the higher the balance, the lower the risk of churn, which means that the churn rate can be reduced by increasing the revolving balance of the customer to increase customer satisfaction and promote spending. Comparison with the findings of (Ramzi & Mohamed, 2010) in the previous literature review. Increase repeat customer satisfaction and customer loyalty by increasing the revolving credit limit, which in turn reduces the chances of churn. Consistent with the findings in this case. The effects of the variables in the model are as expected.

Education level of customer is directly proportional to the churn rate in the model, which means that banks may need to pay more attention to the feedback of highly educated credit card users and provide them with personalized services to improve their sense of belonging and satisfaction.

Income category of customer is directly proportional to the churn rate in the model, which means that the higher the income of credit card customers, the higher the risk of churn. This means that for high income customers, the bank needs to contact them more often, assign personalized service, create a sense of belonging to the customer group, and increase customer satisfaction to reduce the risk of churn. Comparison with the findings of (Cronin Jr et al., 2000) in the previous literature review. The higher the income category, the more customers will demand credit card services. When the services provided by banks do not meet the needs of higher income groups, this creates a risk of customer attrition. This is consistent with the findings in this case.

7. Conclusion

This study has helped banks to successfully predict the churn of credit card customers. However, there are limitations and areas for improvement. Access to large datasets is limited due to the sensitivity of bank data. Access to more datasets would enhance the generalizability of this prediction.

Credit card customer churn has always been one of the most immediate challenges facing banks and credit card organizations, and the act of customer churn often leads to a chain reaction: the long-term profitability and sustainability of the organization is affected, the brand and company reputation is negatively impacted, and market share is lost. This prompted me to choose Predicting Credit Card Customer Churn as my dissertation topic. In turn, big data analytics and data mining provide opportunities and challenges for predictive modeling of credit card customer churn. For micro and small businesses and new banking organizations, it is simply not possible to form and generate predictive models without the support of a large amount of data. The data used in this thesis is from the Kaggle data exploration website and is used for academic research purposes only.

This thesis begins with a theoretical understanding of the terminology of credit card user churn and its influencing factors through a literature review, in addition, the literature study provides a general discussion of the concept of big data, listing the structure and challenges of big data. It then illustrates the definition and application of data mining and lists the data analysis models and their methods applied in this thesis.

In the practical part, this study compares the IBM SPSS Modeler based model with the algorithms of Logistic Regression, Decision Tree CHAID and C&R, and the test results show that the Decision Tree CHAID slightly outperforms the Logistic Regression model in terms of accuracy. The study showed that these algorithms predicted credit card customer churn with an accuracy of 91%-93%. All three models showed that

Total transaction count is the largest influence on credit card customer churn, followed by Total revolving balance of customer and Total transaction amount on customer churn. Number of months customer has been inactive in the last twelve months, Total number of relationships customer has with the credit card provider, Age, Number of dependents that customer has and Number of contacts customer has had in the last twelve months, These factors also have an impact on customer attrition.

This study helps provide useful information and assistance to banks and credit card companies in predicting customer churn for manual intervention and early warning of churn-prone customer segments. Personalized services are provided to churn-prone customers after early warning to understand their needs, improve customer satisfaction and retain customers at the root. Through churn predictive analytics, banks and credit card organizations are able to identify common patterns or triggers of churn and understand which customers are at risk of churn after which they can improve the overall customer experience and reduce the likelihood of churn by enhancing products, services and customer support.

Feedback from churn-tending customers can provide valuable perspectives on the development of new products or improvements to existing credit card services, which can facilitate more informed strategic decisions by management, enabling the bank to allocate resources more efficiently, priorities measures to improve credit card customer retention and ultimately increase profitability.

This paper also confirms that 1. Data mining and big data analytics have a positive impact on customer relationship management. 2. The logistics regression model is slightly less accurate than the Decision Tree CHAID model in predicting credit card churn, but the analysis and statistical description of the influencing factors are more detailed and accurate than the other two Decision Tree Models.

There are some limitations in this thesis, some other factors (region, geographic culture, technology) may have an impact on the probability of credit card customer churn, and appropriate data is needed for a more comprehensive analysis and comparison. The results of this study may not be applicable to clients in all countries due to cultural factors.

The accuracy of the logistic regression used in this paper needs to be improved and the comparison of models needs to be expanded. Therefore, future research could explore further comparisons and predictions using more advanced algorithms (e.g., random forest algorithms and neural network algorithms, which help to identify non-linear and complex relationships between data variables).

8. References

- Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst*. John Wiley & Sons.
- Anil Kumar, D., & Ravi, V. (2008a). Predicting credit card customer churn in banks using data mining. *International Journal of Data Analysis Techniques and Strategies*, 1(1), 4–28.
- Anil Kumar, D., & Ravi, V. (2008b). Predicting credit card customer churn in banks using data mining. *International Journal of Data Analysis Techniques and Strategies*, 1(1), 4–28. <https://doi.org/10.1504/IJDATS.2008.02002>
- Anton, J. (1996). Customer relationship management: Making hard decisions with soft numbers. (*No Title*).
- Athanassopoulos, A. D. (2000). Customer satisfaction cues to support market segmentation and explain switching behavior. *Journal of Business Research*, 47(3), 191–207.
- Berson, A., & Thearling, K. (1999). *Building data mining applications for CRM*. McGraw-Hill, Inc.
- Bolton, R. N., & Bronkhorst, T. M. (1995). The relationship between customer complaints to the firm and subsequent exit behavior. *ACR North American Advances*.
- Brown, S. A., & Coopers, P. W. (1999). *Customer relationship management: A strategic imperative in the world of e-business*. John Wiley & Sons, Inc.
- Buttle, F., & Maklan, S. (2019). *Customer relationship management: Concepts and technologies*. Routledge.
- Chen, I. J., & Popovich, K. (2003). Understanding customer relationship management (CRM): People, process and technology. *Business Process Management Journal*, 9(5), 672–688.
- Cronin Jr, J. J., Brady, M. K., & Hult, G. T. M. (2000). Assessing the effects of quality, value, and customer satisfaction on consumer behavioral intentions in service environments. *Journal of Retailing*, 76(2), 193–218.
- Diebold, F. X. (2012). *On the Origin (s) and Development of the Term 'Big Data'*.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37–37.
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144.
- Gartner, I. (2020). *Glossary, nd Retrieved from <http://www.gartner.com/it-glossary/big-data>*.
- Geiler, L., Affeldt, S., & Nadif, M. (2022). An effective strategy for churn prediction and customer profiling. *Data & Knowledge Engineering*, 142, 102100.
- Gürsoy, U. Ş. (2010). Customer churn analysis in telecommunication sector. *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, 39(1), 35–49.
- Jamalian, E., & Foukerdi, R. (2018). A hybrid data mining method for customer churn prediction. *Engineering, Technology & Applied Science Research*, 8(3), 2991–2997.
- Keaveney, S. M. (1995). Customer switching behavior in service industries: An exploratory study. *Journal of Marketing*, 59(2), 71–82.
- Kentrias, S. (2001). Customer relationship management: The SAS perspective. *Retriev. Mar*, 24, 2011.
- Kim, S., Shin, K., & Park, K. (2005). *An application of support vector machines for customer churn analysis: Credit card case*. 636–647.
- Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6(70), 1.

- Lin, C.-S., Tzeng, G.-H., & Chin, Y.-C. (2011). Combined rough set theory and flow network graph to predict customer churn in credit card accounts. *Expert Systems with Applications*, 38(1), 8–15.
- Linoff, G. S., & Berry, M. J. (2011). *Data mining techniques: For marketing, sales, and customer relationship management*. John Wiley & Sons.
- Mahajan, V., Misra, R., & Mahajan, R. (2017). Review on factors affecting customer churn in telecom sector. *International Journal of Data Analysis Techniques and Strategies*, 9(2), 122–144.
- Misra, R. (2012). An empirical study on the preference and satisfaction for the pre-paid and post-paid cellular subscribers. *Abhigyan*, 30(3), 23.
- 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.
- Parvatiyar, A., & Sheth, J. (2001). Conceptual framework of customer relationship management. *New Delhi*.
- Paulrajan, R., & Rajkumar, H. (2011). Service quality and customers preference of cellular mobile service providers. *Journal of Technology Management & Innovation*, 6(1), 38–45.
- Ramzi, M., & Mohamed, B. (2010). Customer loyalty and the impacts of service quality: The case of five star hotels in Jordan. *International Journal of Economics and Management Engineering*, 4(7), 1702–1708.
- Reichheld, F. F. (1993). Loyalty-based management. *Harvard Business Review*, 71(2), 64–73.
- Sathish, M., Naveen, K., & Jeevanantham, V. (2011). A study on consumer switching behaviour in cellular service provider. *Far East Journal of Psychology and Business*, 5(2), 13–22.
- Singh, P. P., Anik, F. I., Senapati, R., Sinha, A., Sakib, N., & Hossain, E. (2024). Investigating customer churn in banking: A machine learning approach and visualization app for data science and management. *Data Science and Management*, 7(1), 7–16. <https://doi.org/10.1016/j.dsm.2023.09.002>
- Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, 70, 263–286.
- Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2013). *Using multivariate statistics* (Vol. 6). Pearson Boston, MA.
- Tang, L., Li, J., Du, H., Li, L., Wu, J., & Wang, S. (2022). Big data in forecasting research: A literature review. *Big Data Research*, 27, 100289.
- Tudeal, K. Y. D. (2022). Internet Banking Development as A Means of Providing Efficient Financial Services in South Sudan. *Specialusis Ugdymas*, 2(43), Article 43.
- Tufféry, S. (2011). *Data mining and statistics for decision making*. John Wiley & Sons.
- Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. (2024). Customer churn prediction in telecom sector using machine learning techniques. *Results in Control and Optimization*, 14, 100342. <https://doi.org/10.1016/j.rico.2023.100342>
- Wang, B., & Wang, Y. (2021). Big data in safety management: An overview. *Safety Science*, 143, 105414. <https://doi.org/10.1016/j.ssci.2021.105414>
- Weisberg, S. (2005). *Applied linear regression* (Vol. 528). John Wiley & Sons.

9. List of Figure and tables

8.1 List of Figure

Figure 1 Logistic regression curve (Qef, 2008)	20
Figure 2 Linear regression (Agarwal, 2018).	21
Figure 3 Decision Tree (Gray, 2017)	23
Figure 4 Combination of results (An, 2020)	26
Figure 5 Dataset overview.....	29
Figure 6 Missing value - check	30
Figure 7 Data Audit-1.....	31
Figure 8 Data Audit-2.....	31
Figure 9 Age distribution	32
Figure 10 Dependent_count distribution	32
Figure 11 Months_on_book distribution	33
Figure 12 Total_Relationship_Count distribution.....	33
Figure 13 Months_Inactive_12_mon distribution.....	34
Figure 14 Contacts_Count_12_mon distribution	34
Figure 15 Credit_Limit distribution	35
Figure 16 Total_Revolving_Bal distribution	35
Figure 17 Avg_Open_To_Buy distribution	36
Figure 18 Total_Amt_Chng_Q4_Q1 distribution	36
Figure 19 Total_Trans_Amt distribution	37
Figure 20 Total_Trans_Ct distribution.....	37
Figure 21 Total_Ct_Chng_Q4_Q1 distribution	38
Figure 22 Avg_Utilization_Ratio distribution	38
Figure 23 Attrition_Flag-1	39
Figure 24 Attrition_Flag_2.....	39
Figure 25 Gender-1 Distribution	39
Figure 26 Gender-2 Distribution	40
Figure 27 Education_Level-1 Distribution.....	40
Figure 28 Education_Level-2 Distribution.....	40
Figure 29 Marital_Status-1 Distribution	41
Figure 30 Marital_Status-2 Distribution	41
Figure 31 Income_Category-1 Distribution	42
Figure 32 Income_Category-2 Distribution	42
Figure 33 Card_Category-1 Distribution	43
Figure 34 Card_Category-2 Distribution	43
Figure 35 Reclassify Education_Level-1	44
Figure 36 Reclassify Education_Level-2	44
Figure 37 Logistic Multicollinearity-VIF.....	45
Figure 38 Logistic The Omnibus Tests-model 1	45
Figure 39 Result-binary logistic regression model 1-1	46
Figure 40 Result-binary logistic regression model 1-2	46
Figure 41 Result logistic model 2 - The Omnibus Tests	47
Figure 42 Result-Logistic Model 2-1	48
Figure 43 Result-Logistic Model 2-2	48
Figure 44 Predictor Importance- logistic model 2.....	50
Figure 45 Table of logistic model 2	51

Figure 46 Confusion Matrix – logistic Model 2	51
Figure 47 Analysis model 2 result	52
Figure 48 ROC curve for logistic model 2	52
Figure 49 CHAID-2	53
Figure 50 CHAID-3	53
Figure 51 CHAID-3	53
Figure 52 CHAID tree-left 1	54
Figure 53 CHAID tree-left 2	54
Figure 54 CHAID tree-right 1	54
Figure 55 CHAID tree-right 2	55
Figure 56 Table CHAID Tree Training	56
Figure 57 CHAID tree Matrix Training	56
Figure 58 CHAID tree Accuracy Analysis	57
Figure 59 CHAID AUC curve	57
Figure 60 C&R tree-1	57
Figure 61 C&R tree-top	58
Figure 62 C&R tree-down	58
Figure 63 C&R tree table Training	59
Figure 64 C&R tree matrix Training	59
Figure 65 C&R tree Accuracy Analysis Training	60
Figure 66 C&R tree AUC curve Training	60
Figure 67 Confusion Matrix- Testing logistic regression model 2	61
Figure 68 Confusion Matrix-Testing CHAID model	62
Figure 69 Confusion Matrix- Testing C&R model	62
Figure 70 Comparing training model results: AUC and Gini	63
Figure 71 Comparing testing model results: AUC and Gini	63
Figure 72 All ROC curve-Training	64
Figure 73 All ROC curve-Testing	65
Figure 74 variable code value	75

8.2 List of tables

Table 1 Category data, data type, Description	5
Table 2 Education_Level Distribution	41
Table 3 Marital_Status Distribution	42
Table 4 Income_Category Distribution	42
Table 5 Card_Category Distribution	43

Appendix

	Attrition_Flag	Gender	Education_Level	Marital_Status	Income_Category	Card_Category
0	Existing Customer	F	Unknown	Unknown	Unknown	
1	Attrited Customer	M	Uneducated	Single	Less than \$40K	Blue
2			High School	Married	\$40K - \$60K	Silver
3			College	Divorced	\$60K - \$80K	Gold
4			Graduate		\$80K - \$120K	Platinum
5			Post-Graduate		\$120K +	
6			Doctorate			

Figure 74 variable code value
Source: author's own work