CREDIT CARD CUSTOMERS PREDICTION

Dissertation

Hrishikesh Rane

2228534

June 2024

Swansea University
Prifysgol Abertawe

Project Dissertation submitted to Swansea University in Partial Fulfilment for the Degree of Master of Science

Department of Computer Science

Swansea University

SUMMARY

In this research study, the intricate world of credit card company customer attrition is examined. Customer churn—the term for when customers cease using or cancel credit card services—poses a significant issue for businesses, particularly those in the finance sector. Various facets of credit card client churn, including causes and preventive strategies, have been examined in this study. The issue of customer attrition is a major concern for credit card firms nowadays. Identifying if the client would choose to stop and, more importantly, devising workable strategies to do so could be challenging. Customers vary in what they like to use for payment methods. If a firm can't keep consumers, no matter how amazing its products and services are, it won't succeed; this could be because of a lack of understanding about how credit cards are used. It is essential to make an effort to enhance the user experience if you want to make sure that customers can simply receive what they need. By stressing these crucial elements, businesses can lower customer turnover and develop a devoted clientele. To forecast churn, we'll use six algorithms: Random Forest, AdaBoost, Decision Tree, Naive Bayes, Neural Network and Logistic Regression. Random Forest performs the best in predicting churning customers.

Declaration

This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed: Hrishikesh Rane

Date: June 2, 2024

Statement 1

This dissertation is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by giving explicit references. A bibliography is appended.

Signed: Hrishikesh Rane

Date: June 2, 2024

Statement 2

I hereby give my consent for my thesis, if accepted, to be made available for photocopying and inter-library loan, and for the title and summary to be made available to outside organizations.

Signed: Hrishikesh Rane

Date: June 2, 2024

ACKNOWLEDGEMENT

My sincere thanks go out to my supervisor, Dr. Muneeb Imtiaz Ahmad, and Dr. Benjamin Mora for their important advice, criticism, and mentoring during the dissertation process. My work has improved as a result of your insights. Also deserving of recognition are the Computer Science Department faculty members who, through their research and teaching, enhanced my postgraduate studies. I developed my research questions thanks to our conversations, and I was motivated to keep continuing with your support. Without the help and participation of everyone listed above, this dissertation would not have been feasible. I am grateful that you joined me on this journey and assisted me in finishing this work.

Contents

Figures

Tables

CHAPTER 1: INTRODUCTION

1.1 Understanding Customer Churn

Customer attrition, also known as customer churn, or the loss of clients to rival enterprises, is one of the more significant issues faced by organizations in highly competitive markets with little switching costs. It is difficult to prevent customer churn in industries such as banking, insurance, telecommunications, and subscription services because it is easy for customers to switch from one provider to the next. Therefore, it is critical for both industry and academia to prioritize the science of predicting customer churn properly to combat the churn [1]. The loss of these customers can have serious impacts on company revenue, future growth potential, and long-term viability.

Customers may leave for any number of reasons, insufficient product/service quality, bad customer service experience, perceived lack of value, changing customer preferences/needs or a monetary issue. Moreover, customer attrition can be influenced by external factors such as industry trends, regulatory shifts and economic environments. While the root causes of such churn can vary, its adverse effects are consistent: diminished revenue streams, heightened customer acquisition costs and harm to brand reputation.

There is nothing wrong with customer churn beyond the obvious financial ramifications. When they go, companies miss out on the chance to build long-lasting relationships and upsell or upsell more stuff and to get good old-fashioned word-of-mouth endorsements. Even more, significant cost is linked to customer churn i.e., internal resources are focused on acquiring instead of servicing the customers.

Understanding and acting on customer attrition has become a key strategic imperative, for any company seeking to maintain (or expand) a loyal customer base, decrease the costs of customer acquisition, and ensure sustainable growth. Proactive identification and retention of churn customers may help minimize the direct costs of churn damage while also giving companies an edge of learning through the entire customer maintenance exercise, with the target of ultimately increasing shopper pleasure and bonding.

Even slight decreases in customer churn rates can result in considerable financial gains and competitive advantages in businesses with low switching barriers. As a result, organisations are increasingly recognising the value of establishing advanced analytical capabilities to better understand the complex processes of customer attrition and inform focused retention initiatives.

1.2 Causes of Customer Churn

It is important for banks to have precise churn prediction in order to take proactive steps to hold onto key clients. Various sectors have different causes for experiencing consumer loss. Customers leaving a business due to a variety of reasons, such as bad customer service, competitor offerings, perceived value being lower, shifts in the demands or preferences of the client, and price concerns, are some of the common causes of churn. Customer behaviour can also be influenced by external variables that can cause churn, including industry trends, legislative changes, and economic situations. For companies to create sustainable client bases and retention strategies, it is imperative that these elements are recognised and addressed [4].

Yet, the cause of customer churn can be complex and can be inter-related. This could, say, start with some disappointment at the quality of the customer service, which then becomes a feeling that the product or service was not worth as much as the customer initially perceived. This, combined with compelling offers from competitors might eventually push the customer to switch vendors. Hence, it is critical to identify the myriads of drivers of customer attrition in order to create specialised and successful retention strategies.

1.3 Applying Machine Learning for the Prediction and Prevention of Customer Churn

In the past, companies have used a range of tactics to reduce customer attrition, such as raising the calibre of their products, providing better customer service, introducing retention campaigns, and providing incentives and loyalty programmes. Nevertheless, these methods frequently depend on reactionary actions and might not adequately recognise and tackle the fundamental reasons for customer or segment churn. Moreover, these approaches might fall short in capturing the complex relationships among different elements affecting consumer behaviour, producing less than ideal outcomes.

In recent years, the adoption of data-driven strategies and machine learning in the workplace has transformed how companies approach customer attrition. Machine learning algorithms have the capability to uncover hidden patterns and insights that human analysts might not immediately recognize by leveraging vast amounts of customer data, including demographics, usage patterns, purchase history, and interactions with the organization. Once predictive models have been developed based on these insights, proactive retention tactics and targeted interventions become possible by identifying high-risk customers.

The effective application of machine learning techniques has benefited customer churn analysis across various industries. For instance, in the telecommunications industry, churn prediction models have been utilized to identify customers likely to switch providers, enabling businesses to retain these customers by offering personalized incentives or improving service quality. Similarly, in the financial services sector, churn prediction models have been employed by banks and credit card companies to pinpoint customers who may close their accounts or defect to a competitor, allowing for more focused retention efforts.

Machine learning (ML) can be an effective tool in predicting customer attrition by leveraging behavioural patterns and customer data analysis. This predictive capability enables proactive measures to be taken to keep customers who are likely to leave satisfied. Once ML models have been trained on historical customer churn data and relevant customer data has been collected, their effectiveness is commonly evaluated using metrics such as AUC-ROC. The highest-performing model is then selected, and its scores are applied to current customers to identify those most at risk of churning. Strategies, such as personalized services, can then be implemented to retain these valuable customers.

Moreover, machine learning models can not only predict customer attrition but also reveal deeper patterns that explain the phenomenon. By examining the characteristics and trends of these customers, businesses can gain a more comprehensive understanding of the factors driving customer attrition and take proactive measures to address these issues. This approach not only improves customer retention but also optimizes marketing efforts, resource allocation, and the overall customer experience.

To enhance the accuracy and interpretability of churn prediction models, this thesis will investigate advanced machine learning techniques, including regression, probabilistic, supervised, deep learning, and ensemble methods. The focus will be on understanding the key drivers of customer attrition and generating actionable insights to inform effective retention strategies tailored to different customer segments.

With the help of this project, organisations will be able to anticipate customer attrition and take targeted action to reduce it and increase enduring customer loyalty. The goal of customer churn analysis is to give a comprehensive and useful framework. This thesis seeks to add to the body of knowledge in this field by providing a thorough analysis of customer churn and the application of state-of-the-art machine learning techniques. It also offers insightful observations and useful suggestions for companies looking to improve customer retention, lower churn rates, and eventually boost long-term growth and profitability.

CHAPTER 2: BACKGROUND

2.1 Factors Impacting Churning Behaviour

There are several reasons why customers leave typically with the sectors of banking, telecoms. They include service quality, price, customer satisfaction, technology availability, and consumer need shifting. Factors that affect churn include both banking-specific and demographic variables including credit limits, account balances, and transaction patterns. Customer behaviour data enables us to identify churn risks. Models to predict churn and effective retention tactics rely on an understanding of these elements [3]. Such areas are collectively accepted as covenanted inputs in machine learning algorithms for predicting churn likelihood in banking customers. These include consumer demographics, usage habits, service, experience and context among others [4].

2.1.1 Demographic Factors:

Changing customer concerns are analysed on various topics. Demographic indicators include age, gender, education level, and marital status. The pre-Bank customers, having been there so long, are less likely to bank away from us. It may also depend on the gender. Churn rates may also be affected by the level of education that people have, and churn levels can be directly related to both married and unmarried populations [4].

2.1.2 Customer-Bank Relationship:

Customer relationships with the bank with longer tenure are also generally correlated with less turnover. Less customer churn is a known phenomenon; a customer who has multiple bank products with the bank is less likely to churn. In an inconsistent way, too much inactivity leads to a higher likelihood of client churn, but with regular contacts, it helps prevent attrition [4].

2.1.3 Credit Card Usage:

A bank that carries higher revolving balances but has higher credit limits may result in lower attrition but also higher churn probabilities. This ratio, and how low it may be, can indicate your customer is poor and churning. Churn is also dependent on the overall transaction value and count, and an unexpected change in transaction behaviour might be an indication of a churn risk [4].

2.1.4 Card Category:

Finally, the churn rate of different card categories may vary; this is also a way a card belongs to (Blue, Silver, Gold, Platinum). A comprehensive examination, this work provides a nuanced picture of client attrition—its antecedents and results [4].

Although these components give some understanding of customer turnover behaviour, they need to be further looked into and modelled statistically for a complete picture. On the basis of these findings, financial institutions should modify their strategies to reduce customer churn.

2.2 Existing ML Methods

The study by Ünlü proposes a machine learning method with a hybrid solution to predict client attrition within the credit card business. The author uses Support Vector Machines (SVM) using different kernel functions and performs hyperparameter optimisation of SVMs using Bayesian optimisation. The study was performed on a Kaggle dataset for credit card users. Once the SVM models were trained and optimised for the training set, they were evaluated on the test set. According to the results, the linear kernel SVM model works better than others, with an accuracy of 91%. The above may provide a more accurate solution on how to solve the bank client churn forecast problem deeply addressed in the study [1].

Table 1: Metrics of Assessment for the Linear Kernel [1]

Label	Precision	Recall	F_1-score	Support
0	0.78	0.60	0.68	319
1	0.93	0.97	0.95	1707
Macro Average	0.85	0.79	0.81	2026
Weighted Average	0.91	0.91	0.91	2026

Table 2: Assessment Measures for the Polynomial Kernel [1]

Label	Precision	Recall	F_1-score	Support
0	0.61	0.63	0.62	319
1	0.93	0.92	0.93	1707
Macro Average	0.77	0.78	0.77	2026
Weighted Average	0.88	0.88	0.88	2026

Table 3: Assessment Measures for the Radial Kernel [1]

Label	Precision	Recall	F_1-score	Support
0	0.62	0.31	0.42	319
1	0.88	0.96	0.92	1707
Macro Average	0.75	0.64	0.67	2026
Weighted Average	0.84	0.86	0.84	2026

Table 4: Assessment Measures for the Sigmoid Kernel [1]

Label	Precision	Recall	F_1-score	Support
0	0.20	0.00	0.01	319
1	0.84	1.00	0.91	1707
Macro Average	0.52	0.50	0.46	2026
Weighted Average	0.74	0.84	0.77	2026

Azzopardi et al.'s study examined processes in virtual credit card transactions in relation to customer attrition prediction that could be achieved by a machine learning approach. Next, the researchers used two models: the Gradient Boosting Model and Artificial Neural Networks (GBM and ANN, respectively) to examine their hypotheses on demographics and the transaction information obtained from the transaction data. The GBM model outperformed the ANN model with a higher AUROC of 0.69 and was able to forecast churn more accurately. The study also found that 1 month of observation was sufficient for accurate predictions, and using transactional data together with demographic data drastically improved predictions for new consumers. This has significant implications for the financial services industry. The study shows that machine learning models, in the form of GBM, can be employed to enhance overall customer pool retention strategies and that dynamic transaction variables can effectively predict client attrition. Especially when combined with initial transaction data for new consumers and a small observation window [2].

Table 5: Evaluation findings for the GBM prediction model using the Confusion Matrix as a basis [2].

Metric	Score
Sensitivity	0.6989
Specificity	0.6865
False Positive Rate	0.3135
False Negative Rate	0.3011
Precision	0.6581

Here, Al-Najjar et al. used a multitude of independent variables and machine learning models to forecast customer churn. We found that the C5 tree model struck the best deal. Important variables were the revolving balance total, the total number of transactions, and the change in transaction count. 10,127 credit card owners were selected for the sample, with a subset of 1,627 churn users used in the final analysis. Authors employed five diverse machine learning techniques, which were used in conjunction with three variable selection models. To evaluate the performance, accuracy, precision, recall, false omission rate (FOR), and F1 score of the recommendations were assessed. This study makes it easier to forecast client turnover [3].

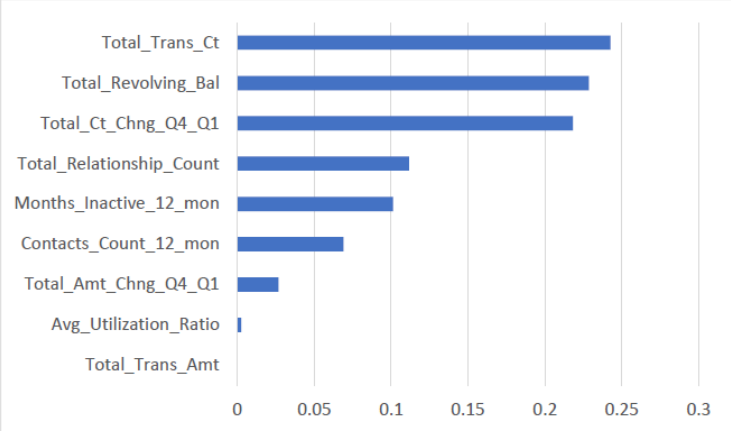

Figure 1: Significant factors following the use of the C5 tree customer churn model [3].

The banking sector was selected to predict customer attrition utilising machine-learning algorithms in a study by Kumara N V et al. The processing of the data and how and where the data were used for training and testing sets was done using customer features like: Age, Location, and Balance Naive Bayes and logistic regression models were trained using the following input combinations on which the evaluation was done: The Naive Bayes algorithm outperformed logistic regression at 91.95% versus 84.75%. So, according to this, in the banking industry, Naive Bayes will work better to predict client churn. For other datasets, the results can be less clear [4].

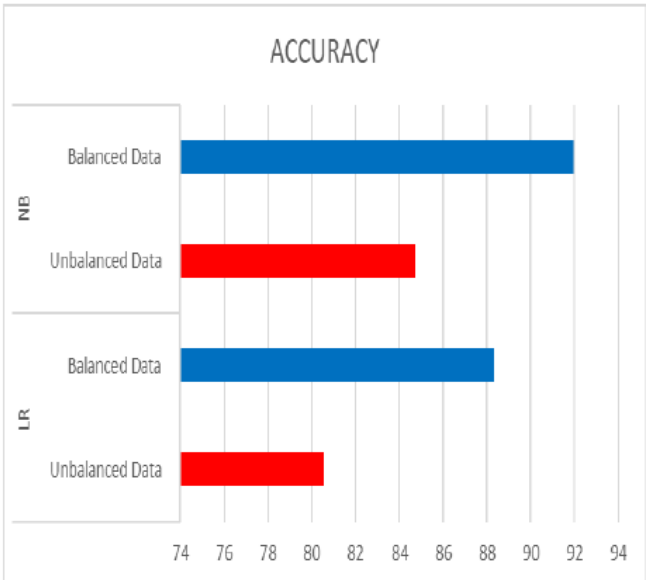

Figure 2: Demonstrates the superior accuracy of Naive Bayes [4].

In the Chang et al. study, credit card data was utilised as the input in order to model customer attrition using machine learning. The authors worked on the pre-processed and balanced data with the help of the ADASYN and SMOTE methods and evaluated six various algorithms. Authors also solved the problem of class imbalance due to the dataset. Class imbalance is a term in machine learning when one class outnumbers the other significantly in machine learning datasets. In this work, in order to alleviate this issue, the authors utilised the Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic (ADASYN) over-sampling opportunities. This improvement makes machine learning models more reliable because these strategies establish a balanced dataset. Then they trained and assessed six machine learning algorithms, which will be: K-Nearest Neighbours, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and LightGBM. The best-performing algorithm was then clearly the LightGBM algorithm, with accuracy, recall, and AUC scores of 92.5%, 91.3%, and 0.97. These findings have demonstrated how machine learning can be used to address real-world problems and help businesses develop strategies to establish client retention [5].

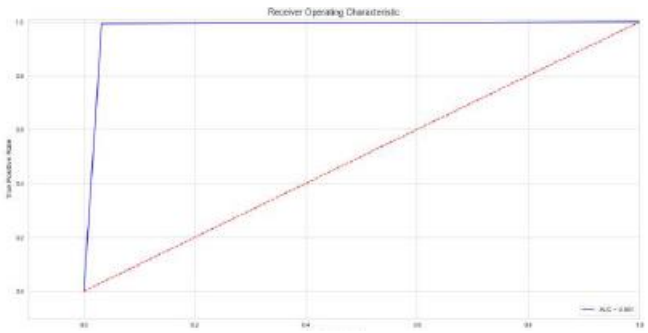

Figure 3: SMOTE-based ROC curve for the LightGBM classifier [5].

This study by Panduro-Ramirez et al. examines a serious banking issue: banking client turnover, the real beast that reduces your revenue. The number of categorical variables in the dataset was converted to numerical values by the authors before use. In order to cope with class imbalance, the SMOTE method was applied. Random forest, K-Nearest Neighbours, XGBoost, and CatBoost are a few of the machine learning methods applied and compared. Hyperparameter

tuning was conducted with grid search and the elbow curve method to get the most out of each model. Models were evaluated using a variety of metrics, including F1 score, accuracy, precision, and recall. Since it was robust with categorical variables and imbalanced data with outliers, CatBoost showed the best results across all models in customer churn prediction, with an accuracy of 97.85%. This has consequences for banks and other businesses. [6]

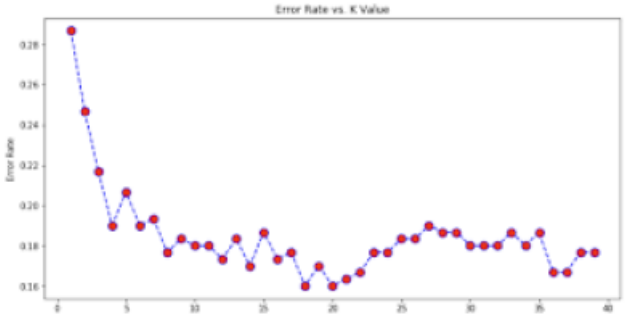

Figure 4: Elbow curve [6].

Prabadevi et al. used machine learning processes to study and seek to predict client churn. The authors conducted their work on the Kaggle dataset, comparing Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbours (KNN), and Stochastic Gradient Boosting (SGB). After the data processing, the algorithms were fit and tuned by hyperparameter tweaking for optimal performance. The results showed the superiority of SGB as having the highest AUC of 0.84 and accuracy of 83.9% among the compared algorithms. Except for KNN, which had the worst performance, LR and RF worked fine. To provide insights that can help businesses forecast customer churn and initiate responsive strategies to reduce customer attrition, an analysis was conducted. The authors recommended that the SGB model be integrated into customer-centric strategies in order to monitor and predict churn rates [7].

Figure 5: A comparison of LR, RF, SGB, and KNN [7].

Arram et al.'s work sought to lower the risks and costs for banks by using machine learning algorithms for predicting credit card defaults. Pre-processing steps like replacing missing entries, outlier detection, and over-sampling for an imbalanced class distribution were applied to a new dataset from an American bank. The models tested included Logistic Regression, Decision Trees, Random Forests, XGBoost, LightGBM, and MLP Neural Networks. With the lowest possible false negative rate comes the overall percent of correctly identified defaults, an indicator measured in recall, which the MLP did perform better than the others, exceeding 80% in this case as well. However, by suggesting that it may help banks identify potential defaulters in time and thereby reduce risks and losses related to defaults, the study concludes that the MLP model is the most suitable to perform this task. This research provides valuable information for future work in credit risk modelling and prediction [8].

Table 6: The effectiveness of the suggested machine learning models [8].

Model	Accuracy (%)	AUC (%)	Precision (%)	Recall (%)
LogisticRegression	85.42	78.95	48.28	70
LightGBM	93.75	83.79	82.35	70
LightGBM	94.44	84.19	87.5	70
Decision Tree	89.58	81.37	60.87	70
Random Forest	93.06	85.48	75	75
MLP Classifier	91.67	86.77	66.67	80

Credit card turnover is a problem for the banks because of the costs that come with it. Wang et al. show the application of machine learning and deep learning, respectively, in the study. This paper discussed various deep learning and machine learning-based churn prediction techniques and models, including inputs in models obtained by using consumer data. The bottom line in churn prediction is that, in the main results, the best deep learning methods (sequence models and autoencoders) outperform the best conventional approaches (like Logistic Regression). Nonetheless, problems such as concept drift, complex nonlinear dynamics, data fusion, and classification imbalance still prevail. This review really helps to emphasise how important it is to perform feature selection and a more rigorous evaluation of performance. In conclusion, it is a comprehensive assessment of where churn prediction with credit cards is today or tomorrow and gives comments for future research in this area. [10]

The work by Zhu uses Random Forest and Logistic Regression strategies to predict customer churn in the banking sector. After preprocessing and splitting in training and test samples, they utilised the Kaggle dataset regarding the bank customer churn. The two algorithms were implemented, and feature importance and K-fold cross-validation performance evaluations were done. The model was good in terms of AUROC of 0.852; however, the Random Forest model performed better in terms of precision of 76.9% and accuracy of 85.9% than the Logistic Regression model. Age was discovered to be the most crucial factor affecting customer attrition. Around 70% of respondents said machine learning has great potential to boost customer relationship management and customer churn prediction, the report found. It called for a more nuanced approach to the jobs to be done and the factors contributing to churn, which posed serious questions about the simple jobs to be done to build such interventions to reduce churn. The banking sector utilises long-term success plans, which are aided by this effort [10].

Table 7: The Results of Cross-Validated Logistic Regression and Random Forest Models [10].

	Random Forest	Logistic Regression
Cross-validation Scores	0.863 0.869 0.86 0.864 0.853	0.795 0.788 0.793 0.792 0.796
Mean Score	0.862	0.793
Standard Deviation	0.00522	0.00266

Kumar et al.'s report examines the trigonometric portfolio and evaluates credit card industry customer attrition. Using the predictive model they created, companies are able to design specific retention tactics to help save customers at greater risk of leaving. The researchers used a Kaggle dataset and applied many machine learning models like XGBoost, Decision Tree, K-Nearest Neighbours, and Logistic Regression. They also proposed Logistic Regression with K-Nearest Neighbours and Logistic Regression with Decision Tree hybrid models. The better forecast accuracy of the hybrid models relative to separate models had important consequences for the credit card industry. Hybrid models can provide a solution for hard prediction problems [11].

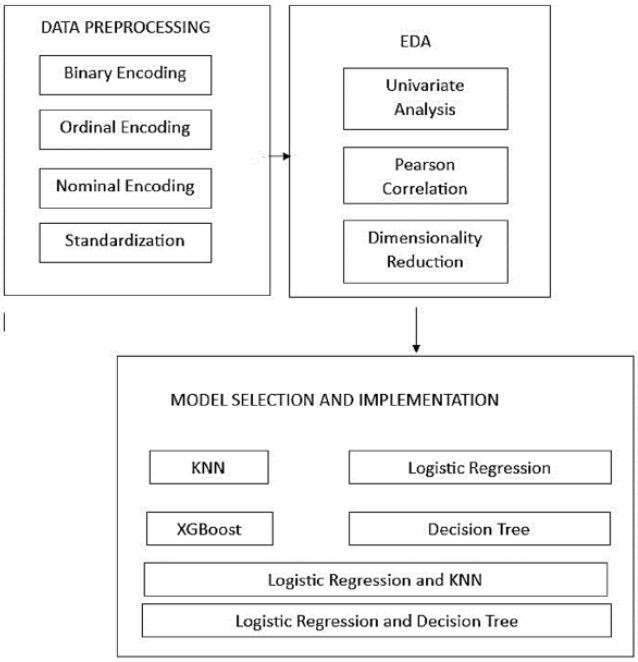

Figure 6: Flow of the suggested technique [11]

In Agarwal et al.'s work, the authors are going to try to predict client churn in banking using machine learning. With the objective of creating models that could help in identifying clients that were more likely to leave and, hence, take proactive steps to retain them. Demographic and account information for the customers were collected and pre-processed using this method. Subsequently, the researchers utilised this dataset to carry out Logistic Regression (LR) and Naive Bayes (NB) models as customer churn prediction approaches. The performance of the models was measured on the basis of accuracy, and the models were evaluated on balanced as well as imbalanced datasets. The accuracy in the case of LR is 90.8%, and with NB it is 91.95% (compare to performing LR). NB is recommended for predicting client churnbuffers for client attrition prediction. This is important because banks would know in advance where and how to rectify consumer discontent through an accurate churn forecast. This seems to provide a narrow pathway for further research and application in customer churn prediction to optimise banking strategies to retain customers [12].

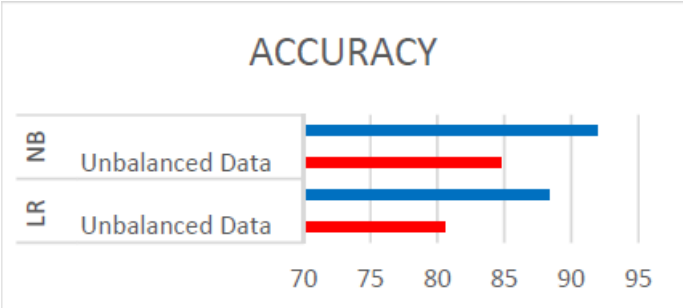

Figure 7: Demonstrates the superior Accuracy of NB [12]

Table 8: Author(s) and their best-performing algorithm.

Author(s)	Algorithms Used	Best Performing Algorithm
Ünlü [1]	Support Vector Machines (SVM) with various kernel functions	Linear kernel SVM
Azzopardi et al. [2]	Gradient Boosting Model, Artificial Neural Networks	Gradient Boosting Model
Al-Najjar et al. [3]	Bayesian network, C5 Decision Tree, CHAID decision tree, Classification and Regression tree, Neural network	C5 tree model
Kumara N V et al. [4]	Naive Bayes, Logistic Regression	Naive Bayes
Chang et al. [5]	K-Nearest Neighbours, Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM	LightGBM
Panduro-Ramirez et al. [6]	Random Forest, KNN, XGBoost, CatBoost	CatBoost
Prabadevi et al. [7]	Logistic Regression, Random Forest, K-Nearest Neighbours, Stochastic Gradient Boosting	Stochastic Gradient Boosting
Arram et al. [8]	Logistic regression, decision trees, random forests, XGBoost, LightGBM, MLP neural networks	MLP model
Wang et al. [9]	Various Deep Learning Techniques, Logistic Regression	Deep learning techniques (specifically, sequence models and autoencoders)
Zhu [10]	Random Forest, Logistic Regression	Random Forest
Kumar et al. [11]	XGBoost, Decision Tree, K-Nearest Neighbours, Logistic Regression, Logistic Regression with K-Nearest Neighbours and Logistic Regression with Decision Tree (Hybrid models)	Logistic Regression with K-Nearest Neighbours and Logistic Regression with Decision Tree (Hybrid models)
Agarwal et al. [12]	Logistic Regression, Naive Bayes	Naive Bayes

In the literature review carried out for this thesis research, a variety of machine learning algorithms have been used for forecasting customer churn in the banking and credit card sectors. The research was done using the following tested classification algorithms: Decision Trees,

XGBoost, LightGBM, CatBoost, K-Nearest Neighbours, GBM, Artificial Neural Networks, Support Vector Machines, Naive Bayes, and Logistic Regression However, each method has its own merits, and each has different performance with the dataset properties. On the other hand, Support Vector Machines with a linear kernel function achieved up to 91% accuracy, while GBM performed slightly better than the Artificial Neural Networks, which exhibited a lower AUROC score. With an accuracy of 92.5% for LightGBM and 97.85% for CatBoost, both have stunning performance.

Through an extensive review, I have concluded my dissertation is to implement AdaBoost, Random Forest, Decision Tree, Neural Network, Naive Bayes and Logistic Regression as the models. They were utilized to show how one can apply different types of methods to demonstrate different types of patterns in the data.

AdaBoost: It is a type of boosting method and works very effectively on tougher data sets, as it has the capacity to boost the performance of weak learners.

Random Forest: It is an ensemble method like decision trees, can handle all data types, and is immune to overfitting because it includes several decision trees predicting simultaneously.

Decision Tree: It is a transparent and simple method and less restrictive than many other classification methods because it can be highly non-linear and handles the relations of the features to the dependent variable well despite its simplicity.

Neural Network: With enough data and computational power, neural networks can be a useful feature detection technique for discovering nuanced patterns and relationships within high-dimensional data.

Naive Bayes: If the dimensions of the data set are large, then Naive Bayes (probabilistic classifiers) are best suited, so they work best in higher-dimensional space; therefore, this kind of algorithm is very easy and fast to implement.

Logistic Regression: It is one of the basic methods to address binary classification problems, which is a logistic regression. It works well even on your tiny test data and is super easy to understand.

With the use of these six algorithms, I would be able to use the positives of each individual algorithm and, at the same time, reduce the weaknesses of each individual algorithm so that all my predictions come from reliable models and will be robust enough. This feature will help highlight the data and the topic of customer attrition in the banking sector. The performance of my system is better than testing to know how well they will do on my very task. Since the majority of information is already there with them, the banking industry would readily be able to use this as a strategy in their customer retention campaign. In the following chapters of this thesis, I embrace the results of my execution.

CHAPTER 3: METHODOLOGY

The methodology chapter is a vital part of the research study. It helps in preparing a notion of the method and the ways that are executed to accomplish the aims of the research paper. This chapter attempts to provide insights into the approach followed for analysing the data and developing the churn prediction model for the financial services sector. The primary concern of this study is to predict the possibility of attrition among credit card users by applying several machine learning methods and techniques to better understand consumer behaviour. The overall system of research work, with its basic steps and approach, can be explained through a block diagram.

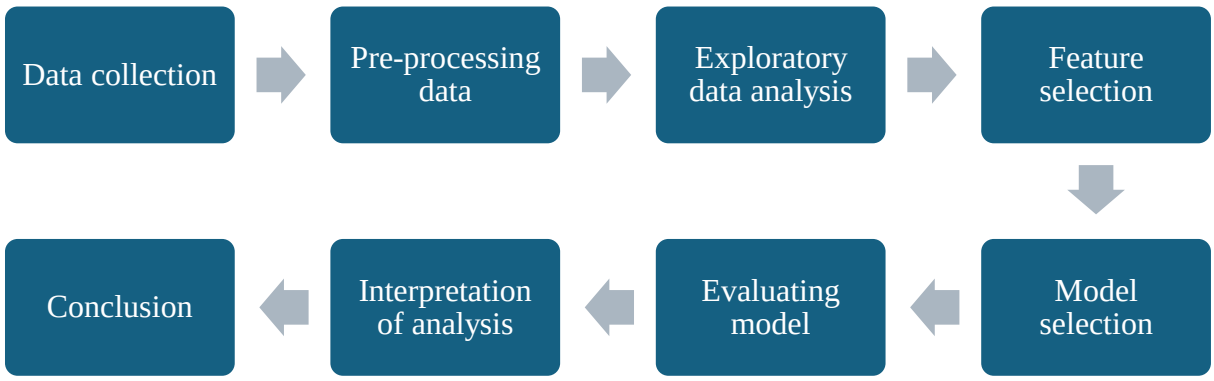

Figure 8: Block Diagram

The block diagram gives an idea of the steps that are followed to perform the research work and make the analysis appropriate. The first task is to collect data, pre-process the data to clean the raw data, perform exploratory data analysis, select the appropriate features, choose a good machine learning model, evaluate the model, make an interpretation of the analysis done, and finally draw conclusions from the study performed.

3.1 Research Philosophy

The research philosophy “interpretivism” is highly appropriate for the investigation of the prediction of credit card customers. This philosophy recognises that individual settings, subjective judgements, and personal experiences all have an impact on human behaviour. In this case, the use of credit cards and financial decision-making. By embracing interpretivism, we acknowledge that the decisions credit card users make are highly influenced by their socio-cultural origins, personal beliefs and motivations in addition to quantitative information [17]. In order to explore the complex motivations behind credit card usage, qualitative research techniques like in-depth interviews and surveys, as well as the use of models, are encouraged by an interpretive approach. These techniques offer a broader insight into consumer behaviour than quantitative data alone. We can use different models, such as Random Forest, AdaBoost, Neural Networks, Naive Bayes, Logistic Regression and Decision Trees.

3.2 Research Approach

The purpose of an inductive research approach is to draw assumptions from specific facts and observations about credit card behaviour among consumers. Data on different customer characteristics, including credit history, income, purchasing habits, and demographic data, is initially gathered. Patterns and trends in this data are found using statistical analysis and machine learning methods. Using these patterns, predictive models can be created to predict consumer behaviour and elements like card usage, repayment tendencies, and churn rates. Inductive research attempts to develop predictive models by extrapolating from the observed data which can help credit card companies make wise judgements and successfully tailor their services to customer's demands [17].

3.3 Data Collection

This dataset of "Credit Card Customers Prediction" is collected from Kaggle, which is open source [16]. This dataset contains information on users of a credit card firm and their behaviour. This dataset is frequently used for machine learning and predictive analytics activities, notably for forecasting customer attrition or churn. The different columns in the dataset are as follows:

Table 9: Data Dictionary [16]

Column	Description
CLIENTNUM	It is an unique identifier for each customer
Attrition_Flag	It indicates whether the customer is an "Existing Customer" or an "Attrited Customer." This is typically the target variable in predictive modeling tasks
Customer_Age	It shows the age of the customer
Gender	This variable shows the gender of the customer (e.g., "Male" or "Female")
Dependent_count	The number of dependents associated with the customer
Education_Level	The customer's educational level, which could include categories like "High School," "Graduate," etc.
Marital_Status	The marital status of the customer (e.g., "Single," "Married")
Income_Category	The income category of the customer (e.g., "Less than $40K," "$40K - $60K")
Card_Category	The type of card the customer holds, such as "Blue," "Silver," "Gold," or "Platinum"
Months_on_book	The number of months the customer has been with the company
Total_Relationship_Count	The total number of products held by the customer
Months_Inactive_12_mon	The number of months the customer has been inactive in the last 12 months
Contacts_Count_12_mon	The number of contacts the customer had with the company in the last 12 months
Credit_Limit	The credit limit on the customer's credit card
Total_Revolving_Bal	The total revolving balance on the credit card
Avg_Open_To_Buy	The average amount available for the customer to spend on the card
Total_Amt_Chng_Q4_Q1	The total change in transaction amount from Q1 to Q4
Total_Trans_Amt	The total transaction amount in the last 12 months
Total_Trans_Ct	The total transaction count in the last 12 months

Total_Ct_Chng_Q4_Q1	The total change in transaction count from Q1 to Q4
Avg_Utilization_Ratio	Ratio of Average Card Use

This dataset is mostly used for tasks that involve client retention and churn prediction. Such information is used by businesses, particularly credit card firms, to evaluate consumer behaviour and pinpoint the elements that lead to client attrition. The dataset is appropriate for a variety of data analysis and machine learning approaches since it contains both category and numerical variables. This dataset and the information in it would help in creating machine learning models to forecast client attrition. This also helps us determine the attributes that increase a customer's likelihood of leaving and take preventive measures by looking at their behaviour and customer information. Based on the dataset's analysis, it may be possible to divide customers into groups according to their transactional and demographic features. Targeted advertising can employ these divisions. It helps in understanding the customer's behaviour and also assess the credit risk.

3.4 Libraries

Code for importing libraries:

```
import pandas as pd

from sklearn.model_selection import train_test_split,
cross_val_score, StratifiedKFold

from sklearn.preprocessing import LabelEncoder, StandardScaler

from sklearn.metrics import classification_report,
confusion_matrix, accuracy_score

from sklearn.ensemble import RandomForestClassifier,
AdaBoostClassifier

from sklearn.naive_bayes import GaussianNB

from sklearn.linear_model import LogisticRegression

from sklearn.tree import DecisionTreeClassifier

from keras.models import Sequential

from keras.layers import Dense

from sklearn.neural_network import MLPClassifier

import seaborn as sns

import matplotlib.pyplot as plt

from imblearn.over_sampling import SMOTE

from scipy import stats
```

In this study, we benefited from many Python libraries and modules for data preprocessing, training the model, evaluation of the model, and visualization. Data manipulation and analysis

were supported by Pandas, a powerful data manipulation library. The Scikit-learn package had all sorts of model selection and preprocessing methods, scoring, and cross-validation; clearly, machine learning in Python did not. We Used Scikit-learn Techniques: Random Forest, AdaBoost, Naive Bayes, Logistic Regression, Decision Trees, and Neural Networks. Along with the Keras package, which really simplified the process of creating and training complete deep learning models. The tools Seaborn and Matplotlib were used for visualisation, which helped us plot the data and generate relevant and eye-catching plots and figures, thereby helping us understand and interpret our findings better. In order to correct for any class imbalance that may exist within the dataset, we rebalanced the dataset by oversampling the minority class using the SMOTE (Synthetic Minority Oversampling Technique) algorithm with the Imbalanced-Learn library. We also employed the statistical capabilities of the SciPy library for comprehensive and stable experiments. We utilised this whole suite of libraries and modules that facilitated data preprocessing, model training, evaluation, and visualisation, allowing us to draw educated inferences from our research.

3.5 Data pre-processing

Pre-processing the data is a very important task in machine learning projects. There are many different steps carried out in data pre-processing for the analysis and modelling of the data. The basic steps include handling the missing values, dealing with the categorical columns, encoding the target variables, removing the unwanted columns, etc. This dataset contains 23 columns and 10,127 rows.

So, the first step is handling the missing values. For this, we are checking if there are any null values present in the dataset or not. It is important to ensure that the dataset is comprehensive and that the missing values in the data won't compromise the accuracy of the analysis or the machine learning model. Depending on the dataset, missing values can either be replaced with suitable values such as the mean, median, or mode, or they can be removed just by dropping the rows or columns. With the help of the function `df.isnull().sum()` we have checked if there are any null values present in the dataset. It has been found that there are no null values in the dataset, so the dataset is good to proceed further. I am removing the last two columns named "Naive_Bayes_Classifier_attribution" and "Naive_Bayes_Classifier_attribution", as they have the same name but different values in their respective columns. I am also going to drop "CLIENTNUM," as we really don't need it.

The next step is to check the data types of the columns. It is found that some of the columns are integers, while some columns are categorical in data type. So, it's important to convert the categorical columns into numerical so that the machine learning models can accept the data in numerical values. Categorical variables must be encoded since many machine learning techniques demand numerical inputs [18]. Label encoding is a popular technique for this, where each category is given a different number. This is done to transform categorical data that is not numerical into a format that the machine learning models can understand.

In this project, there are some categorical columns that are important to be used in the analysis and need to be converted into numerical values. The label encoding technique is used for converting the categorical columns into numerical data types. Using the label encoding approach, category variables may be transformed into a numerical representation that machine learning algorithms can use. When working with categorical data, this approach is straightforward and often used, especially when the categorical variable has an ordinal

connection, which indicates that the categories have a meaningful order. Label encoding effectively creates a mapping from category to number by assigning a distinct integer to each category in a categorical variable. Usually, the integer values are assigned according to the categories' order or ranking in ascending order. The label encoding technique is popular due to its simplicity and ease of understanding. It helps in maintaining the ordinality between the categories in the categorical variables. As it doesn't build new columns for each category, it can be more memory-efficient than one-hot encoding for high-cardinality categorical data.
Code for label encoding:

```
categorical_columns = ['Gender', 'Education_Level',
'Marital_Status', 'Income_Category', 'Card_Category']

# Create a copy of the DataFrame to preserve the original data
encoded_df = df.copy()

# Create a LabelEncoder object
label_encoder = LabelEncoder()

# Iterate through the categorical columns and apply label
encoding
for col in categorical_columns:
    encoded_df[col] =
label_encoder.fit_transform(encoded_df[col])

encoded_df = encoded_df.drop(categorical_columns, axis=1)

# Display the updated DataFrame
encoded_df.head()
```

	Attrition_Flag	Customer_Age	Dependent_count	Months_on_book	Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Count_12_mon	Credit_Limit
0	Existing Customer	45	3	39	5	1	3	12691.0
1	Existing Customer	49	5	44	6	1	2	8256.0
2	Existing Customer	51	3	36	4	1	0	3418.0
3	Existing Customer	40	4	34	3	4	1	3313.0
4	Existing Customer	40	3	21	5	1	0	4716.0

Figure 9: Label Encoding

The attribute ‘Attrition_Flag’ is converted into a numerical value by using the mapping technique. This column has two categories: one is ‘Existing customer’, and another is ‘Attrited customer’. So, by using the mapping technique, the existing customers are assigned a value of 1, while the Attrited customers are assigned a value of 0. A dictionary called mapping describes the relationship between categories and numbers. In this instance, the values “Existing Customer” and “Attrited Customer” are mapped to 1 and 0, respectively. The mapping is applied to the ‘Attrition_Flag’ column using the map function. Based on the dictionary mapping, it substitutes the category values with the equivalent numerical values. In the ‘Attrition_Flag’ column of the Data Frame, ‘Existing Customer’ is changed to 1 and ‘Attrited Customer’ to 0 as a consequence.

Code for mapping:

```
# Define the mapping
```

```
mapping = {'Existing Customer': 1, 'Attrited Customer': 0}

# Apply the mapping to the 'Attrition_Flag' column

encoded_df['Attrition_Flag'] =
encoded_df['Attrition_Flag'].map(mapping)

# Display the updated DataFrame

encoded_df.head()
```

	Attrition_Flag	Customer_Age	Dependent_count	Months_on_book	Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Count_12_mon	Credit_Limit
0	1	45	3	39	5	1	3	12691.0
1	1	49	5	44	6	1	2	8256.0
2	1	51	3	36	4	1	0	3418.0
3	1	40	4	34	3	4	1	3313.0
4	1	40	3	21	5	1	0	4716.0

Figure 10: Converting "Attrition_Flag" Column into Numeric

3.6 Exploratory Data Analysis

A crucial phase of the data analysis process is exploratory data analysis (EDA). It entails a careful examination of the dataset to obtain knowledge, comprehend the data's structure, spot trends, and look for abnormalities. Before creating predictive models or making data-driven judgements, EDA is frequently carried out. In EDA, basically the data is reviewed, checked for the columns and rows, and the data types of the columns are checked. The statistical summary, such as the mean, median, etc., is examined, and the frequency of each of the categorical variables is examined and plotted. With the help of histograms, box plots, or density plots, we have tried to depict the distribution of numerical data and then analyse the data. This aids in comprehending the data's distribution and central tendency. We have created bar plots, or pie charts, to depict the distribution of categories for categorical variables. All this helps in understanding the patterns and trends in the data.

3.7 Model Selection

A few years ago, the credit card companies were anticipating customer churn, mostly using different manual processes. This manual process was basically used for performing different analyses and taking critical decisions against them. But with the development of a new era where everything is changing to be automated by making use of data science and machine learning, we could get accurate and precise results.

The word 'churn' basically means the customers are withdrawing or cancelling their credit card services, and ultimately, this leads to substantial difficulty for credit card providers. Churn is a major issue since it results in revenue loss and a drop in consumer loyalty. The causes of client attrition in the credit card business, methods for predicting when it could happen, and in-depth analyses required to comprehend credit card customer churn will all be covered in this article [19].

Through this project, there has been an attempt to build a model that would determine why many customers would stop using credit cards. Thus, there are six machine learning models proposed in this research work to understand the churn rate.

3.7.1 Random Forest

One of the most often used ensemble techniques, Random Forest aggregates the output of several decision trees to produce a single outcome. It produces more reliable and accurate outcomes. Overfitting is a potential concern when employing a single decision tree, however this random forest regression approach addresses the issue when the prediction outputs of several decision trees are added together.

The ensemble method means that it is going to combine the results of multiple models to give a single prediction. In ensemble method there are two types of techniques one is Bagging, and another is boosting.

Bagging method is used to lower a model's error by averaging its forecasts from multiple models. It works by resampling with replacement to divide the training data into several subsets which is different bags. Then decision tree is trained using each bag separately. Predictions are often made through majority voting for classification task and for regression, there is averaging of the results from each individual model's predictions [18]. A popular ensemble approach that uses the bagging process is Random Forest. Using random feature subsets and bootstrapped samples from the training data, it constructs numerous decision trees. In categorization tasks, the outcome is determined by the majority vote of the trees.

Depending on whether a customer is an "Existing Customer" or an "Attrited Customer," Random Forest might be utilised in the project to anticipate client attrition. This dataset is appropriate for Random Forest since it can handle a combination of numerical and category characteristics. Additionally, it can offer feature significance scores that can be used to determine which customer characteristics are most effective at forecasting attrition [18].

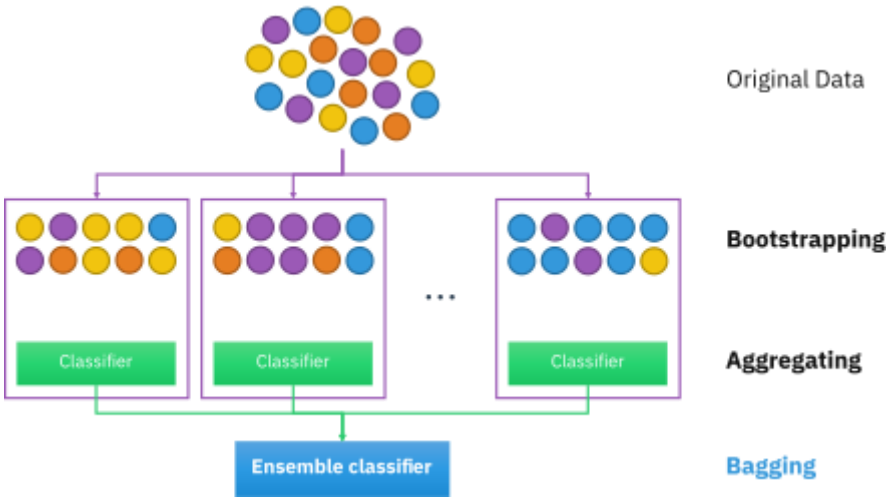

Figure 11: Random Forest Method (Bagging) [20]

3.7.2 AdaBoost

Another ensemble technique is AdaBoost, which combines a number of weak learners which is usually decision trees so as to produce a strong learner. To enhance classification performance, it gives each data point a distinct weight and concentrates on the ones that are incorrectly identified [21]. AdaBoost adds more models to the training set and increases the weight of misclassified data items. The models are then blended, giving more weight to the models that perform better. AdaBoost has the potential to improve the model's classification performance

in the project. If the dataset is uneven or the ensemble consists of weak learners, then it's a smart choice. In this context, it might increase the predictability of client attrition.

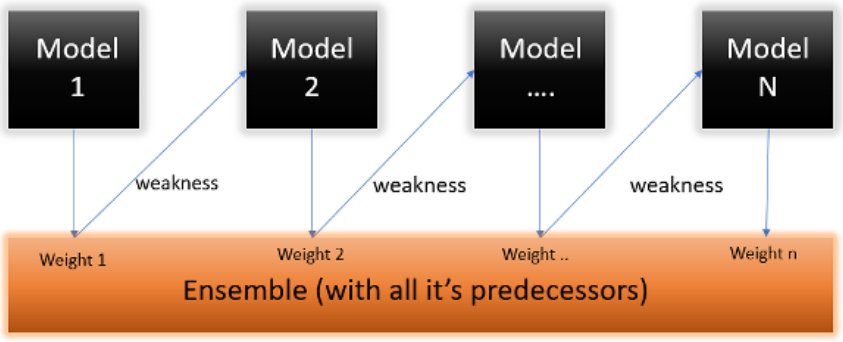

Figure 12: AdaBoost Algorithm [22]

3.7.3 Neural Networks

In machine learning, a neural network is a computational model that matches the human brain's decision-making process. One or more hidden layers, an input layer, and an output layer are among the layers of nodes, or artificial neurons, that make up this structure. Nodes have weights and thresholds of their own and are interconnected. A node is triggered and starts transferring data to the following layer of the network if its output exceeds a predetermined threshold. Otherwise, no information is shared. In order to learn and become more accurate over time, neural networks are trained using data. They can quickly classify and cluster data, making them effective tools in computer science and artificial intelligence if they are refined. With input data, weights, a bias (or threshold), and an output, every node in a neural network can be compared to a linear regression model. After that, an activation function is applied to the output to determine its final value. Data is passed to the following layer of the network when the node is activated if the output surpasses a predetermined threshold. This neural network's feedforward function is defined by the way data is passed from one layer to the next. Neural networks are the foundation of deep learning models and an essential part of machine learning. They go by the names simulated neural networks (SNNs) or artificial neural networks (ANNs) occasionally [14].

Formula: $\sum w_i x_i + \text{bias} = w_1 x_1 + w_2 x_2 + w_3 x_3 + \text{bias}$

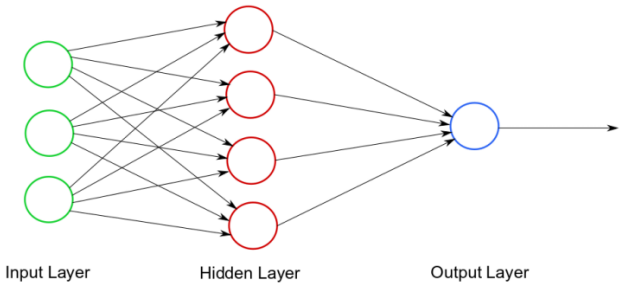

Figure 13: Working of NN [14]

3.7.4 Naive Bayes

Naive Bayes is a popular probabilistic machine learning model that is extensively employed for solving classification problems. The Bayes Theorem algorithm is based on the 'naïve'

assumption that each characteristic in a dataset is independent of each other. This means that the presence of any feature does not influence the presence of any other feature, and vice versa. This model does this by taking the features of a set of features of an input, assigning a likelihood of that input belonging to each class, and returning the class with the highest likelihood as the output. Naive Bayes is a good choice for live prediction because of its effectiveness and preference for high-dimensional data [13].
When there are many X variables (features), the formula for n numbers of X is as follows [13]:

$$P(Y = k|X1, X2. \ldots Xn) = \frac{P(Y)\prod_{i=1}^{n} P(Xi|Y)}{P(X1) * P(X2) \ldots . * P(Xn)}$$

3.7.5 Logistic Regression

Logistic regression (supervised machine learning) can be used for jobs that fall under binary classification. Its main purpose is to predict the probability that an instance belongs to a given class. For example, it can be used to find out if an email is spam or not or to determine the medical diagnosis of a test result from a patient. This illustrates that it's used to forecast the result of a categorical dependent variable, where the result should be a discrete or categorical value but not a continuous value, i.e., true or false, 0 or 1, yes or no, due to its binary values. It does not tell you the exact numbers between 0 and 1, it tells you a range of probabilistic values. In logistic regression, we map each real-valued number to a particular value ranging from 0 to 1 using the sigmoid functions. This y depends on the notion of the threshold value: the probability of 0, and the probability of 1. Values above the cutoff point tend to one, values below to zero. According to their respective criteria, binomial, multinomial, and ordinal regression fall into three different categories that are listed above. A prominent difference from linear regression is that, instead of predicting a continuous value, the output is a binary value (0 or 1) [15].

Formula: $y = \frac{e^{(b0+b1*x)}}{1+e^{(b0+b1*x)}}$

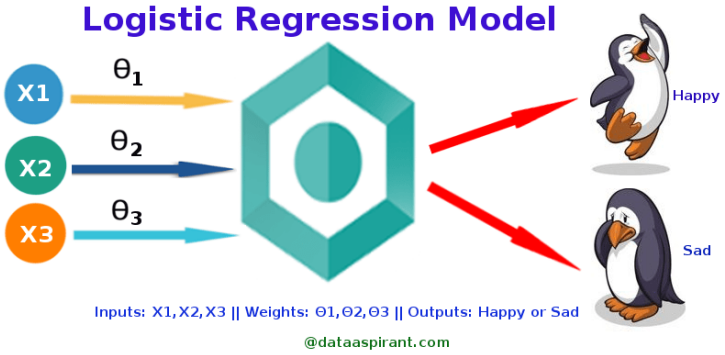

Figure 14: Working of Logistic Regression [15]

3.7.6 Decision Tree

A powerful tool in the field of supervised learning algorithms, decision trees are used for tasks involving both regression and classification. With the help of the data's properties, it builds a decision-tree form model. The test results are represented by branches in the tree, internal nodes that each represent a test on an attribute, and leaf nodes that carry a class label. A stopping requirement, such as the maximum tree depth or the least number of samples needed to split a node, is met by recursively partitioning the training data into subsets based on attribute values in order to create the decision tree. Based on a measure of the degree of impurity or randomness

in the subsets, such as entropy or Gini impurity, the algorithm determines which characteristic is ideal for data splitting. Finding the characteristic that maximises either the reduction in impurity after splitting or the information gain is the goal. Decision trees are a popular choice for novices in machine learning because of its simplicity, interpretability, and ease of implementation. Additionally, they serve as the foundation for one of the most reliable machine learning algorithms, the Random Forest method, which trains on various subsets of training data.

We can utilise these techniques to develop user attrition prediction models for this study project. After preparing the data and doing exploratory data analysis (EDA), the dataset may be divided into training and testing sets. After being trained on training data using testing data, models such as Random Forest, AdaBoost, Neural Networks, Naive Bayes, Logistic Regression, and Decision Tree are then evaluated for anticipated accuracy. Then, several assessment measures are applied to these models.

CHAPTER 4: RESULTS AND DISCUSSION

This chapter covers the various analytical findings as well as the exploratory analysis carried out for the study project. It aids in our comprehension of the various elements and criteria that determine whether to keep or lose clients by using credit cards.

For the credit card companies, it is important to understand the actual value of the customer retention for the credit card companies so that they can understand the customer's base and make proper decisions. Credit card churn is basically the withdrawal of credit card customers which could be either due to cancellation of the credit card or expiry of the credit card which is not renewed on time. At such an instance the customers are then considered as a churned customer. A customer can also be considered a churned customer if they don't use some particular card for a longer duration and switch on to some other credit card company. Thus, it is very important to understand why a customer would stop using this particular credit card so that we can find out the key factor in predicting and reducing the churn. There could be many different reasons and different factors that could be affecting the customers due to which they don't feel like continuing with this particular credit card [23]. Different factors could be lack of awareness, very little information related to credit cards, lack of marketing from the credit card company's side, age factor, income of the customer, etc. It could happen that if the customer has no knowledge and doesn't know about the benefits of using the credit card, they won't use it, and ultimately, this would lead to an increase in the churn percentage [24].

Thus, it is very important to understand the different factors, parameters, and conditions so that it would be useful for the companies to take productive actions against such factors and try to retain their customers. Suppose the credit card companies don't understand why the customers are discontinuing the use of their credit cards. Then the company is at risk of losing their customers and is also not able to get new customers, leading to a great financial loss for the company. Thus, through this analysis and different exploratory data analyses, we are trying to understand these factors and make proper insights that would be useful for the companies in changing their marketing strategy, giving more offers to attract customers, and also trying to retain their customers by providing proper knowledge to the customers regarding the use of credit cards.

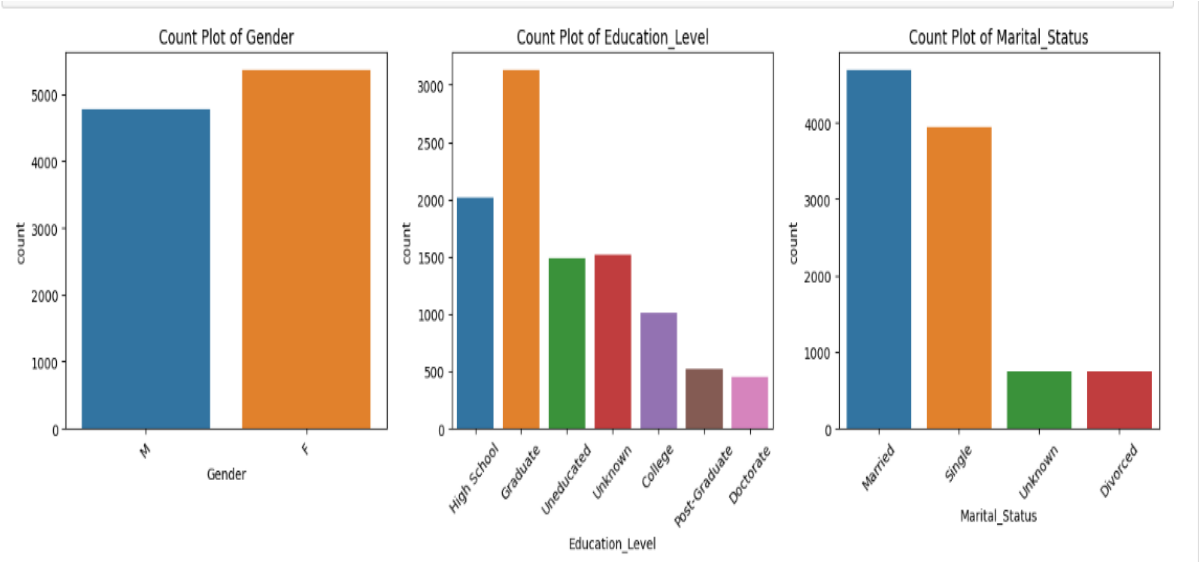

Figure 15: Gender, Education Level and Marital Status of Customers

First, we begin by exploring the univariate analysis, where the frequency count of each of the variables is examined and analysis is performed. From the frequency count of the gender, it was found that most of the customers using the credit card are female. The percentage of male members using this credit card is lower than that of female users.

The next plot shows the education level of the credit card users. The analysis could predict that most of the people using the card are graduates, which is then followed by the people who have studied until high school who use this credit card. It is observed that very few doctoral and postgraduate people make use of this credit card.

The marital status analysis shows that most of the credit card users are married, followed by unmarried users. According to the data, there is a sizable gender gap among credit card users, with more females than males. This knowledge could be useful for marketing plans. It can imply that the business should modify its marketing initiatives to appeal to a primarily female user base. The kinds of financial goods and services available might vary depending on the education level of customers. This realisation could influence the creation of particular financial solutions tailored to certain educational groupings [23].

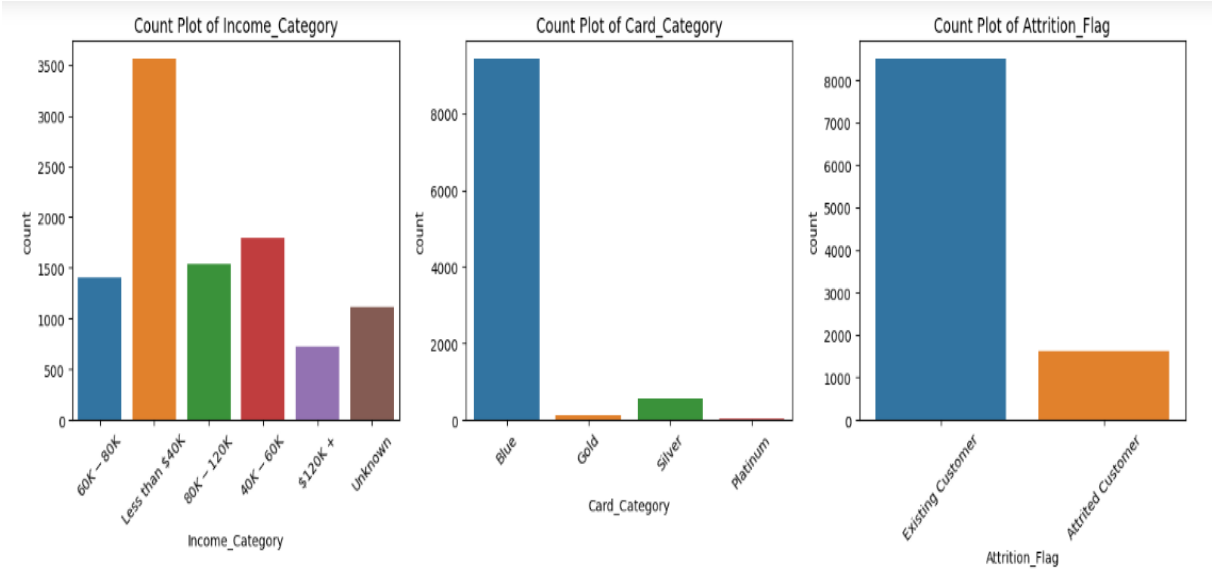

Figure 16: Income, Card Type and Attrition Flag Plot

The next analysis shows that people who have an income less than 40K are using this credit card, followed by people who have an income in the range of 40K to 60K. People are mostly using the blue card, and very few people have access to the silver and gold cards. The users' income level is a key consideration when designing financial goods and services. It can direct the creation of credit card features, interest rates, and incentive schemes that are tailored to various income levels. Understanding how prevalent various card kinds are might help with portfolio management. It can imply the necessity for marketing campaigns to highlight premium cards or a review of the functionality provided by each card type.

From the frequency plot of the attrition flag, it is found that most of the people who are using this credit card are existing customers of this credit card, while very few of the customers tend to churn from the use of this credit card. Even when the majority of customers aren't leaving, it's still crucial to focus on client retention. The business may wish to look into the factors contributing to the tiny number of customers that churn and put plans in place to further lower churn. Positive indications include the finding that more consumers are currently transacting business with the company than those who have stopped. It suggests a strong rate of client retention. Even if the retention percentage is good, the business should keep putting retention and possible new client attraction techniques into practice. Understanding the factors that contribute to client turnover might enhance retention efforts.

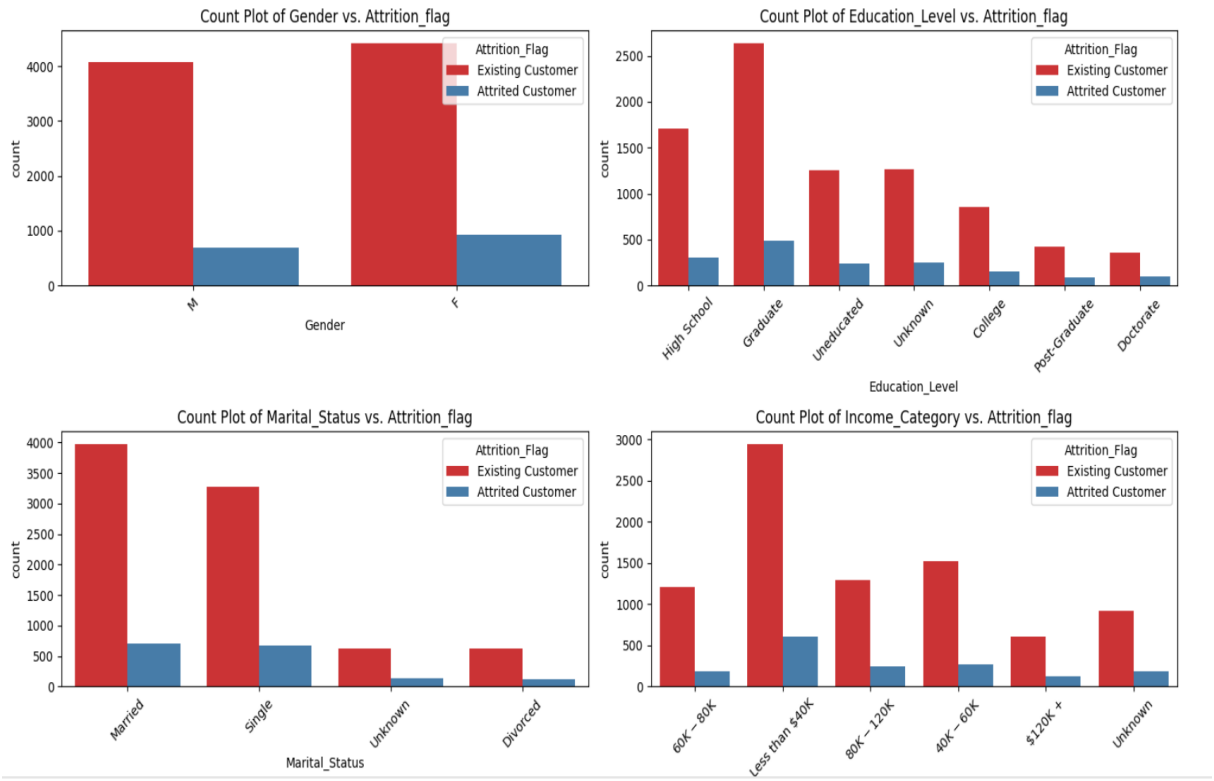

Figure 17: Bivariate Analysis

The next analysis shows the bivariate analysis of the different variables. It is observed that female members tend to churn or tend to stop using the credit card compared to male customers.

From the education level, it is found that most of the people who are graduates tend to stop using their credit cards. The marital status does not have much impact on the churning of the customers from the use of this credit card. The fact that there are more women than men in both business status groups (current customers and attired customers) indicates that the majority of the company's clients are women. This knowledge may direct marketing initiatives targeted towards the female market, ensuring that the company's goods and services meet the tastes and requirements of this market group.

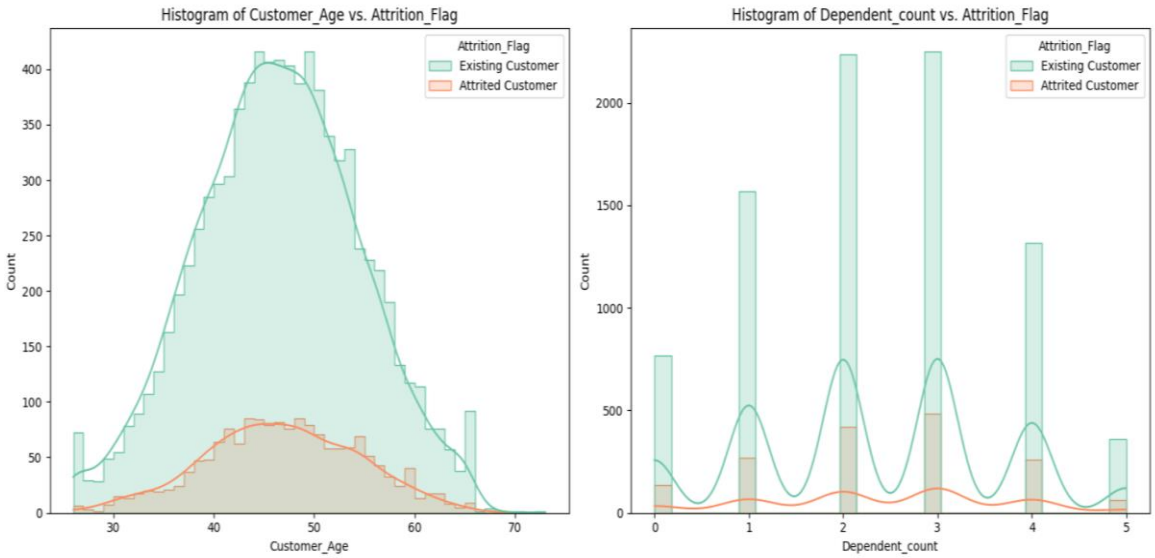

Figure 18: Histogram of Age and Dependent Count with the Target Variable

The finding that customer attrition begins to rise at the age of 29 suggests that there may be an age-related component to customer turnover. The business could take into account providing unique promotions, incentives, or items designed specifically for clients in this age range to solve the issue of attrition around age 29. The finding of a rise in the number of dependents after the age of 30, perhaps as a result of marriage and family expansion, suggests that these consumers are more financially responsible. For evaluating credit risk, this data is useful. Customers with more dependents could have different financial habits and needs, which could have an impact on loan decisions.

The assessments shed light on the characteristics and habits of customers. However, it's crucial to investigate the underlying reasons impacting customer attrition, dependent count, and educational choices in order to make more practical and data-informed judgements. The association between these characteristics and other facets of consumer behaviour, including spending patterns or credit card usage patterns, should be examined in further investigation.

In addition, taking into account the pattern of age-related churn, the business may wish to perform customer surveys or gather feedback to better understand the particular causes of attrition among clients around the age of 29. Developing focused retention efforts may be aided by this.

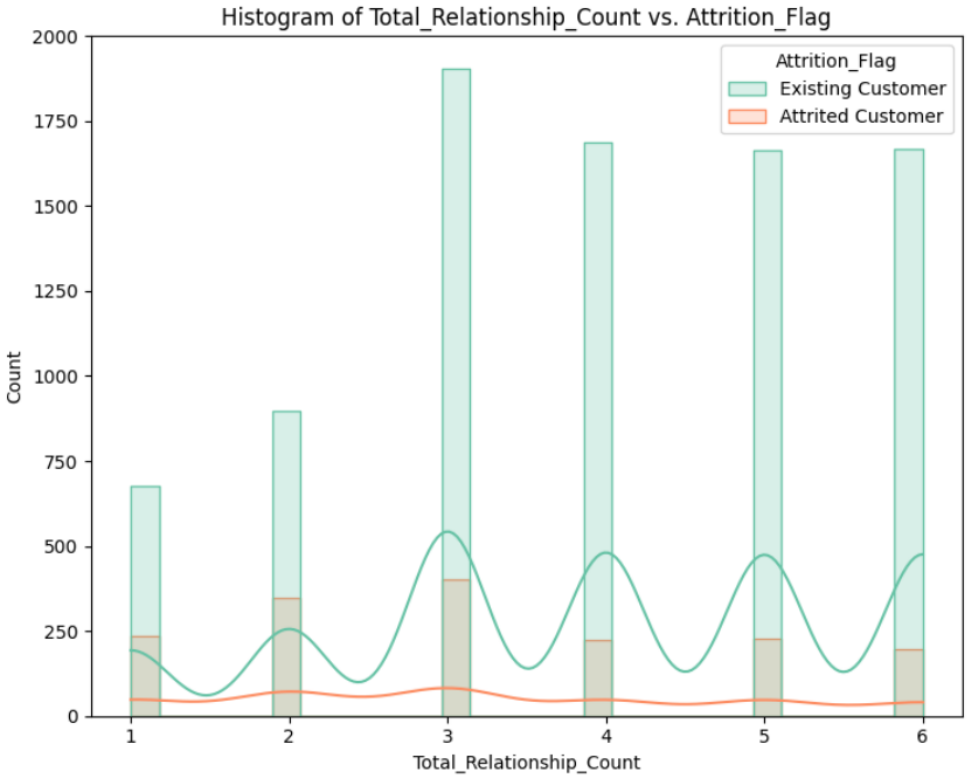

Figure 19: Total Relationship Count Verses Attrition Flag

From the relationship count, it is observed that most of the customers are giving a relationship count in the range of 3 to 6. It shows that the customers are having a good relationship status with the credit card providers, and the customer relationship is healthy, which is making people retain and continue to use the credit card.

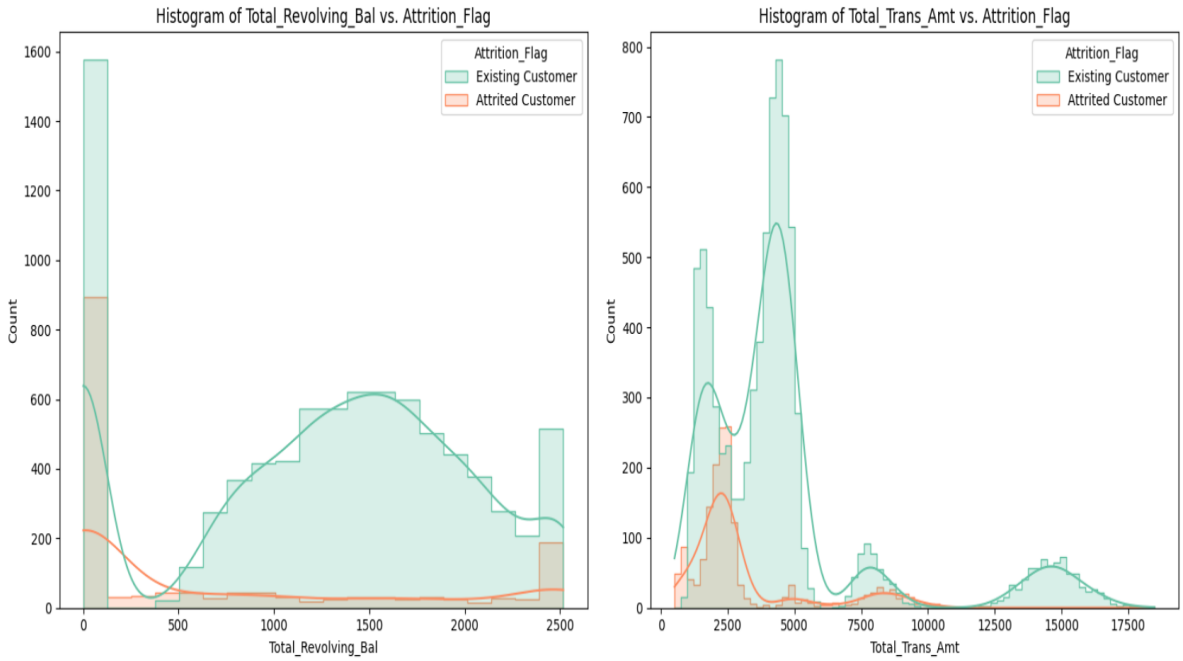

Figure 20: Total Transaction Amount and Revolving Balance

The next analysis shows that most people are revolving their balances in the range of 500 to 2500 and most of the customers are still using this credit card. The maximum transaction amount also lies in the range of $2500 to $5000. This indicates that customers are engaged in transactions with values ranging from modest to high. The corporation may build reward programmes or incentives that correspond with consumer behaviour by taking into account these spending trends. The majority of clients with balances between $500 and $2500 are kept, which is consistent with the notion that consumers in this range appreciate the card. It's conceivable that their purchasing habits have a role in this retention.

Analysis of the distribution of balance and transaction amounts across credit card users offers important insights on consumer behaviour and usage trends. Customers that keep balances between $500 and $2500 are often kept, showing that they respect their credit card and use it regularly. To maintain this group's happiness and loyalty, retention methods should be customised for them.

Benefits like cashback rebates, reduced interest rates, or raised credit limits might be very alluring to this clientele. The usual range of the highest transaction amount is between $2500 and $5000, indicating that the majority of clients do transactions with values ranging from low to high. The business can create incentives or rewards programmes that fit with these transactional patterns. Offering cashback or extra points for purchases inside this range, for instance, might promote card use and client loyalty.

Customers that have balances between $500 and $2500 are more likely to be kept, which shows they are happy with the card. It’s probable that these clients' purchasing patterns contribute to their retention. The business can better comprehend and meet their needs by looking at their individual purchasing patterns. This can entail determining popular expenditure categories or tying marketing to customary buying patterns. The organisation is able to create financial goods and services that are in line with client behaviour by comprehending balance and transaction trends. The organisation can more successfully attract and keep consumers by creating items

that are tailored to their requirements and purchasing preferences. In order to do this, new card kinds, features, or services that appeal to client preferences may be introduced.

In addition to being educational, the study of balance and transaction amounts offers valuable information. This information may be used by the business to establish customer-focused retention, product development, and marketing strategies. The firm is more likely to satisfy the demands and expectations of its credit card customers by matching its services with consumer behaviour, which eventually results in higher customer satisfaction and loyalty.

Moving ahead with the analysis of different factors affecting the credit card, the next histogram plot is of the total transaction count in a period of 12 months for the customers. This is shown in the figure 21 plot. It shows that most of the customers who are not continuing to use the credit card have made a maximum of 40 transactions in the last 12 months. This shows that these customers might not have liked the customer service or the fees of the card, and the interest rate in repaying the dues is higher, which is making it difficult for the customers to not use the credit card further. While the people who have proper knowledge regarding the use of credit cards and are aware of the benefits of using credit cards are making more transactions in a year. The maximum number of transactions made by existing customers in a period of 12 months is more than 80.

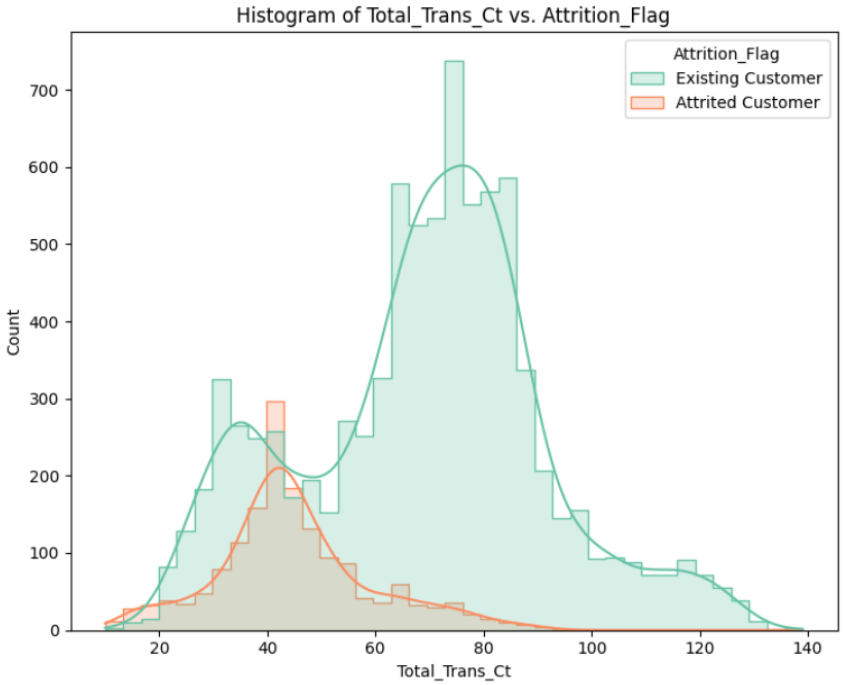

Figure 21: Histogram Plot for Total Transaction Count in Last 12 Months

The credit card companies also face the dilemma of pricing. Each and every individual in this world is facing a crisis for their livelihood. People tend to make strict decisions for any kind of spending that is not properly justified.

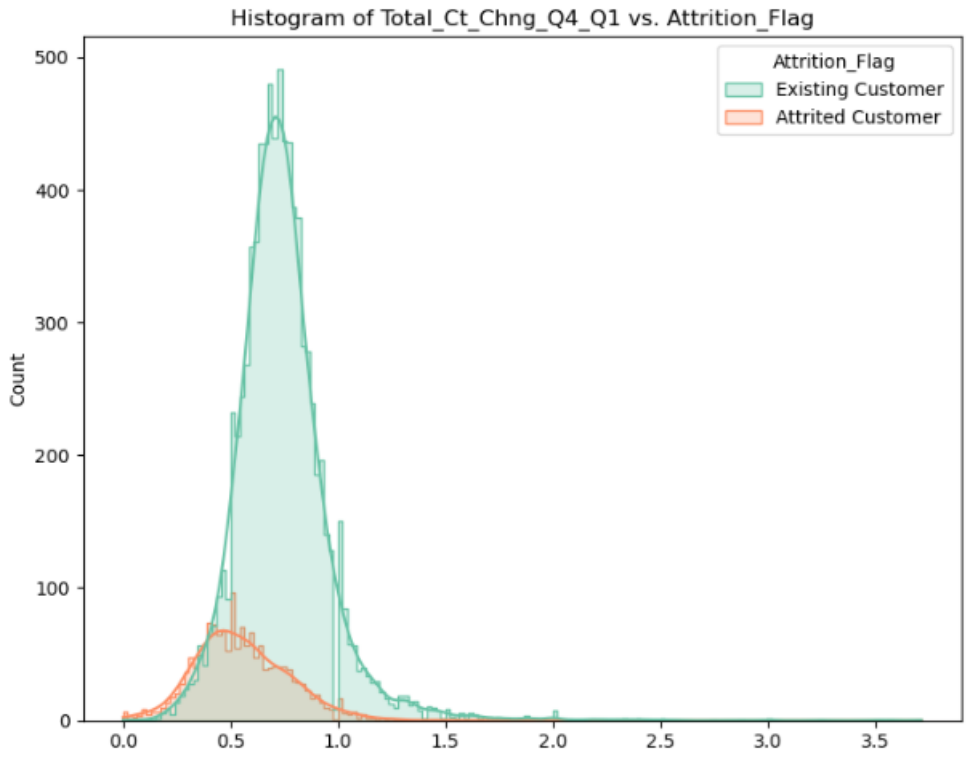

Figure 22: Histogram for Total change in transaction count from Q1 to Q4.

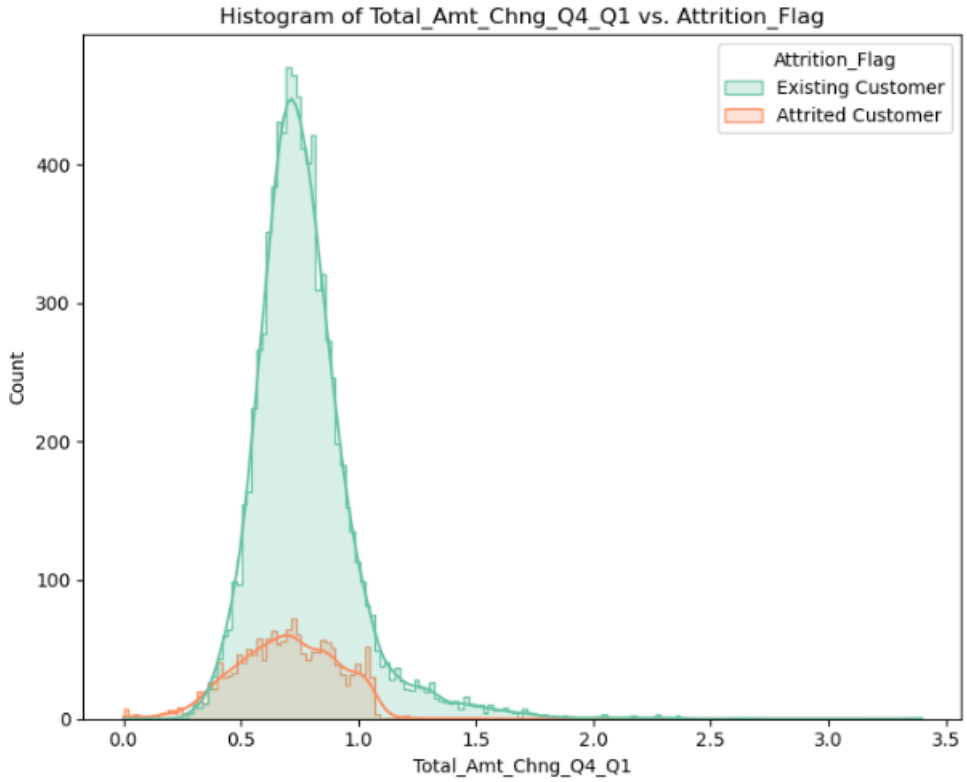

Figure 23: The total change in transaction amount from Q1 to Q4

In the histograms of "Total_Ct_Chng_Q4_Q1" and "Total_Amt_Chng_Q4_Q1," it shows data for the first two quarters, which are Q1 and Q2. We can see that the change in transaction count for attrited customers peaked in the middle of quarter 1, and after that, it dropped to the end of quarter 1 and passed a little bit to the first few days of quarter 2. The change in transaction

amount for attrited customers peaked in the 3rd month of quarter 1 and started dropping. We can see that these two attributes are correlated.

Some of the credit cards may be expensive and have very high yearly fees, which could be hundreds of dollars, making them unworkable for those with little resources. Some credit card companies say that they provide credit cards that are "free," but they still include additional fees. High interest rates may add up rapidly, and clients could find it difficult to take full advantage of reward schemes. Customers frequently max out their credit cards and struggle to pay off their bills, locking them in a vicious cycle of debt that is difficult to escape.

Another crucial issue in the credit card sector is the standard of customer care. Many consumers must endure excruciatingly long wait periods, and they frequently get vague or inadequate answers to their questions. Customers who get poor customer service may feel frustrated and helpless, which may lead them to look for better service elsewhere.

Next, we have box plots to check if there are outliers in the dataset. It was found that the dataset has a lot of outliers. This is shown in figure 14. These outliers are removed so that it doesn't affect the accuracy of the model. Z-scores serve as a gauge for how much a data point deviates from the mean. Z-scores are computed in the code for each chosen column in the training data. The SciPy library's stats. zscore function is used for this. The variable threshold is used to reflect the threshold value that is used for outlier identification. The threshold is set to 3 in this instance. This means that outliers will be defined as data points with a Z-score of more than 3 or less than -3. This threshold is fairly arbitrary and can be changed in accordance with your unique requirements. With the help of this technique, we can detect and understand the outliers and also remove them from the data.

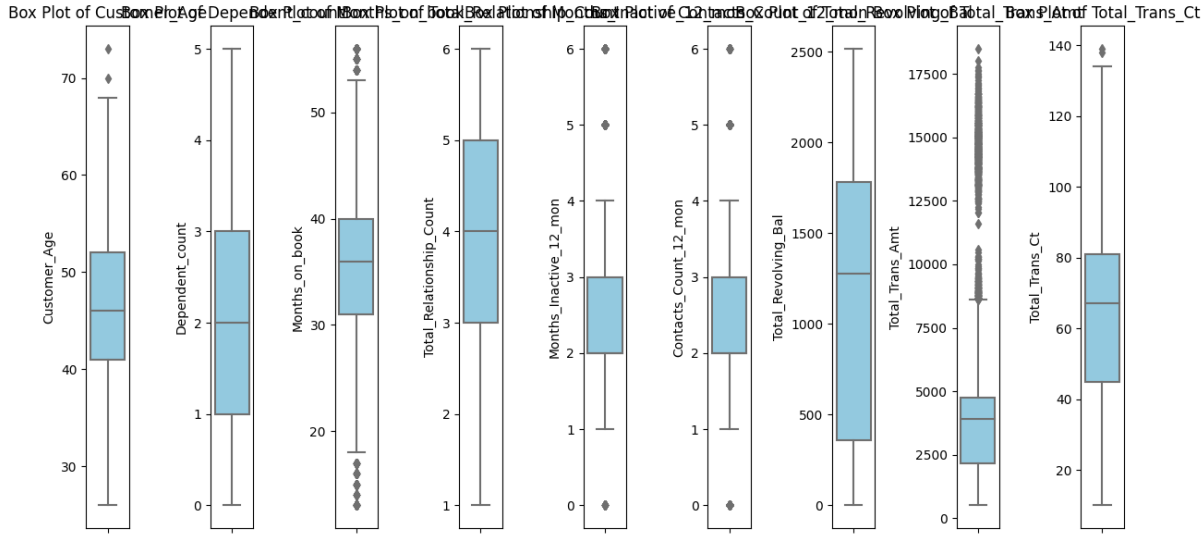

Figure 24: Checking Outliers in the Data

The dataset is then split into a training dataset and a testing dataset. 80% of the data is used for training the model, while 20% of the data is used for testing the model. The 'Attrition_flag' is considered the target variable that is to be predicted. Then the six machine learning models are built and trained on the data, and finally, the evaluation of the models is performed.

Code for splitting dataset:

```
X = df.drop('Attrition_Flag', axis=1)
```

```
y = df['Attrition_Flag']

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)
```

4.1 Results:

The table below sorts every model according to its overall accuracy for the class “Attrited Customer” in descending order.

Table 10: Overall Accuracy for “Attrited Customer”

Model	Precision	Recall	F1-score	Overall Accuracy
AdaBoost	0.88	0.81	0.84	0.9516288252714709
Random Forest	0.91	0.75	0.83	0.9486673247778875
Decision Tree	0.77	0.77	0.77	0.9259624876604146
Neural Networks	0.67	0.56	0.61	0.8840078973346496
Logistic Regression	0.57	0.79	0.66	0.8706811451135242
Naive Bayes	0.38	0.78	0.51	0.7620927936821322

AdaBoost, Random Forest, and Decision Tree models were able to predict attrited customers as well as give excellent overall accuracy regardless of class imbalance in the dataset. The AdaBoost model performed with an accuracy of about 95.16%, while the Random Forest model's accuracy is about 94.86%, which is slightly lower as compared to the AdaBoost model. The Decision Tree model also performed with impressive 92.59% accuracy.

Neural Networks, Logistic Regression, and Naive Bayes models failed to predict attrited customers at initail findings because of class imbalance. Only 16.07% of customers stopped using their credit card, according to the dataset [16]. Dealing with imbalanced datasets is a typical issue in machine learning. This imbalance can cause a bias in the model's predictions towards the majority class, hurting its predictive performance for the minority class. The Synthetic Minority Oversampling Technique (SMOTE) is used in this study to solve this issue.

The SMOTE algorithm starts with a sampling method of 0.7 and a random state of 0. The sampling strategy parameter determines the target ratio between the number of samples in the minority class and the number of samples in the majority class following the sampling process. In this situation, the minority class will be oversampled to produce a class distribution of 70% majority. This method improves the models' performance with the minority class.

Code for SMOTE:

```
#Using SMOTE to handle class imbalances for algorithms like
NN, NB, and LR

smote = SMOTE(sampling_strategy=0.7, random_state=0)

X_resampled, y_resampled = smote.fit_resample(X_train,
y_train)
```

We implemented SMOTE on Neural Networks, Naive Bayes, and Logistic Regression models. The Neural Networks model gave an impressive accuracy of about 88.40% but a poor F1-score of 0.61. Same with the Naive Bayes model; it has an overall accuracy of 76.20% but a poor F1-score of 0.51.

But even after implementing SMOTE on the Logistic Regression model, it again failed to predict attrited customers. To fix this issue, we had to scale the test data and resampled data. "StandardScaler" normalises a feature by removing its mean and scaling to unit variance. To calculate the unit variance, divide all values by the standard deviation. This operation was performed on both the resampled data and the test data.

Code for Standard Scaler:

```
# Scale the data for LR

scaler = StandardScaler()

X_resampled_scaled = scaler.fit_transform(X_resampled)

X_test_scaled = scaler.transform(X_test)
```

After scaling, the Logistic Regression model has an impressive overall accuracy of about 87.06%. Even after scaling, the F1-score is still poor, at 0.66.

4.2 Confusion matrix for each model

A confusion matrix is a performance metric that summarises a classification model's performance by displaying the amount of true positives, true negatives, false positives, and false negatives in each class. It enables for the display of a machine learning classification model's performance [24].

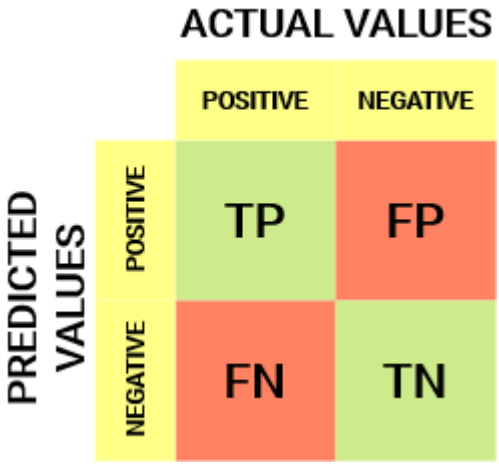

Figure 25: 2x2 Confusion Matrix [25]

Key Terms:

True Positives (TP): Actual positive cases that are correctly forecasted as positive.

True Negatives (TN): Actual cases that were correctly forecasted as negative.

False Positives (FP): Actual negative cases wrongly forecasted as positive.

False Negatives (FN): Actual positive cases that are wrongly forecasted as negative.

Let's take a look at the confusion matrix of all six models for class 0, as it represents “Attrited Customers” also called churning customers and class 1, as it represents “Existing Customers”.

4.2.1 The Random Forest Model

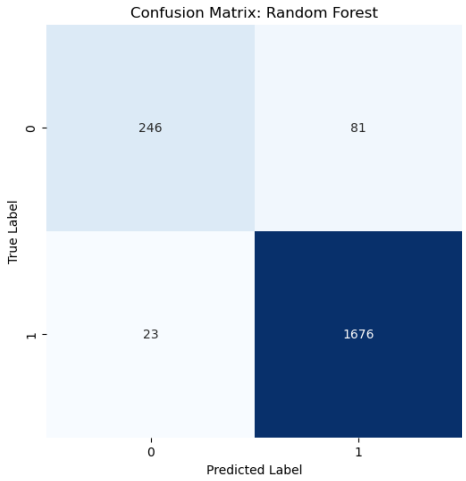

Figure 26: Confusion Matrix for the Random Forest model

True Positives: The model correctly predicted 246 samples as churning customers.

False Negatives: The model incorrectly predicted 23 samples which are churning customers as existing customers.

True Negatives: The model correctly predicted 1676 samples as existing customers.

False Positives: The model incorrectly predicted 81 samples which are existing customers as churning customers.

4.2.2 The AdaBoost Model

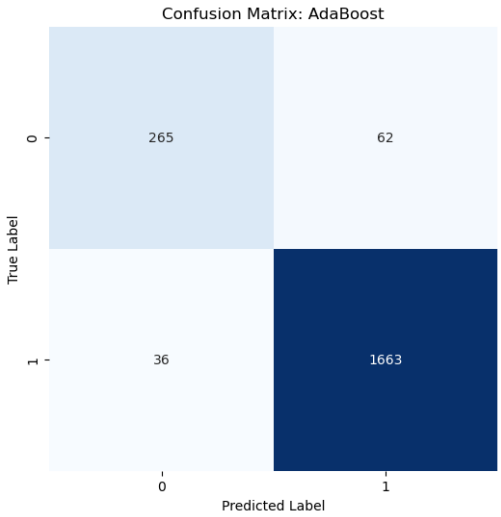

Figure 27: Confusion Matrix for the AdaBoost model

True Positives: The model correctly predicted 265 samples as churning customers.

False Negatives: The model incorrectly predicted 36 samples which are churning customers as existing customers.

True Negatives: The model correctly predicted 1663 samples as existing customers.

False Positives: The model incorrectly predicted 62 samples which are existing customers as churning customers.

4.2.3 The Neural Networks Model

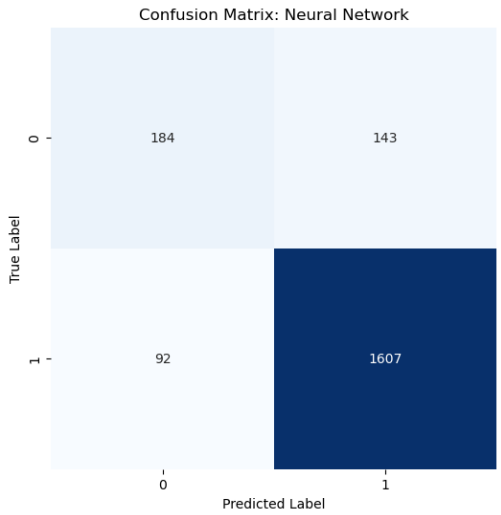

Figure 28: Confusion Matrix for the Neural Networks model

True Positives: The model correctly predicted 184 samples as churning customers.

False Negatives: The model incorrectly predicted 92 samples which are churning customers as existing customers.

True Negatives: The model correctly predicted 1607 samples as existing customers.

False Positives: The model incorrectly predicted 143 samples which are existing customers as churning customers.

4.2.4 The Naive Bayes Model

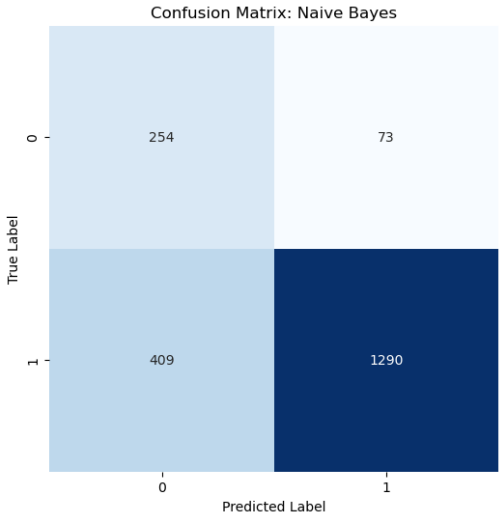

Figure 29: Confusion Matrix for the Naive Bayes model

True Positives: The model correctly predicted 254 samples as churning customers.

False Negatives: The model incorrectly predicted 409 samples which are churning customers as existing customers.

True Negatives: The model correctly predicted 1290 samples as existing customers.

False Positives: The model incorrectly predicted 73 samples which are existing customers as churning customers.

4.2.5 The Logistic Regression Model

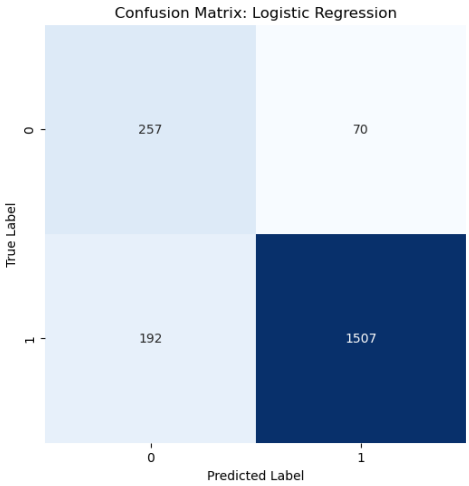

Figure 30: Confusion Matrix for the Logistic Regression model

True Positives: The model correctly predicted 257 samples as churning customers.

False Negatives: The model incorrectly predicted 192 samples which are churning customers as existing customers.

True Negatives: The model correctly predicted 1507 samples as existing customers.

False Positives: The model incorrectly predicted 70 samples which are existing customers as churning customers.

4.2.6 The Decision Tree Model

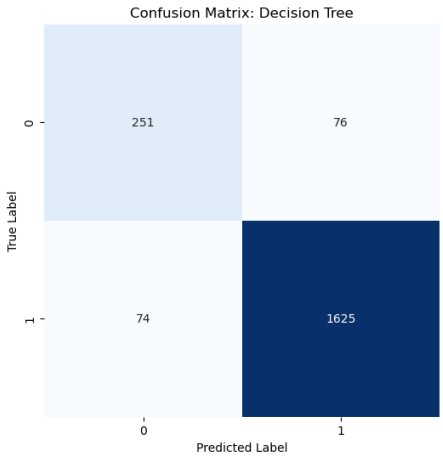

Figure 31: Confusion Matrix for the Decision Tree model

True Positives: The model correctly predicted 251 samples as churning customers.

False Negatives: The model incorrectly predicted 74 samples which are churning customers as existing customers.

True Negatives: The model correctly predicted 1625 samples as existing customers.

False Positives: The model incorrectly predicted 76 samples which are existing customers as churning customers.

4.3 Cross Validation

We will use a Stratified K-fold with 10 folds. As we have class imbalance in the dataset, SKFold will perform better. The table below sorts each model according to its average in descending order.

Table 11: Cross-validation Scores

Model	Mean
Random Forest	0.9404536480301534
AdaBoost	0.93709552497376
Decision Tree	0.9071723718146313
Naive Bayes	0.8793343647210767
Logistic Regression	0.8666889722149606
Neural Network	0.8429978461814592

If we compare cross-validation scores and F1-scores, we can see that Random Forest and AdaBoost models are the best for predicting attrited customers. The Random Forest model with a cross-validation mean of around 0.94 and an F1 score of 0.83 and the AdaBoost model with a cross-validation mean of around 0.93 and an F1 score of 0.84 show that the Random Forest model will perform better on unseen data as compared to the AdaBoost model, as seen on the cross-validation score. Even though there is just a little difference between both models, we can see that the Random Forest model has a precision of 0.91 and the AdaBoost model has a precision of 0.88. This shows that the Random Forest model is better at predicting attrited customers as compared to the AdaBoost model.

4.4 Feature Importance

4.4.1 The Random Forest Model

We will use the Gini Importance of Random Forest algorithm to measure the importance of each feature for the prediction of customer churning factors.

4.4.1.1 Gini importance (MDI):

Gini importance, also called mean decrease in impurity (MDI), is a commonly used measure of variable importance in random forest models for classification. It quantifies the importance or impact of a specific feature/ predictor variable to predict the outcome or to take decisions in the random forest model [26].

The Gini importance is calculated for each predictor variable as follows:

While building a tree more and more leaf nodes are created, which are then used to calculate the output based on the sum/average of the input data,and it uses the Gini impurity criterion to select the splits of leafnode which in turn used during the tree formation and there by resulting in Random forest's number of decision trees. The weighted impurity (weighted Gini impurity) at each split the variable is used to decide on is computed as a decrease comparing the weighted impurity from the parent node to the weighted impurity of the child nodes. It aggregates the weighted impurity among how variables of the independent from those nodes are decreasing for all trees in the forest. It then computes the summed absolute value of all of the changes in impurity over the estimation steps for that predictor, and normalizes it by the number of trees—the result is the Gini importance score. The higher the score of the features, the most significant the feature is to accurately classify the samples or to make good predictions. But, the Gini importance is biased and will always select variables that have high cardinality (i.e. many unique values or categories) as important. A higher potential for node organization variance, ie more room for pure subset creation, entails a greater probability of impurity decrease just by randomness alone. To reduce the bias, alternative importance measures like permutation importance are frequently used instead [26].

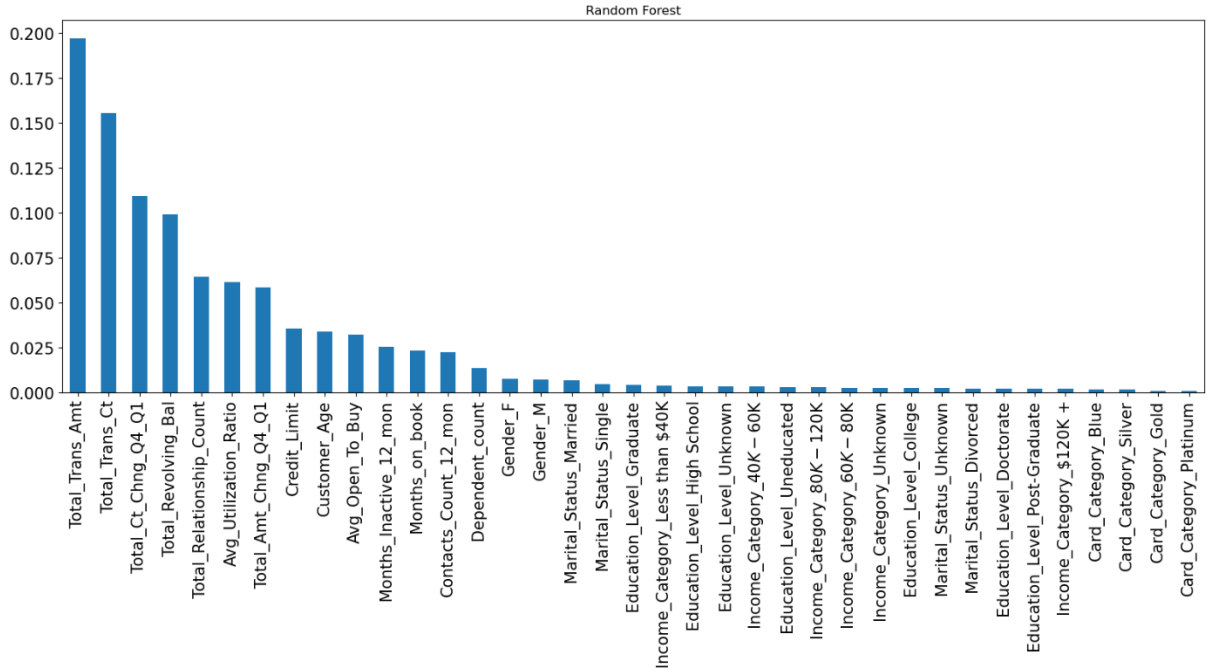

Figure 32: Important features for the Random Forest Model

Table 12: Score of each Feature for the Random Forest Model

Feature	Importance
Total_Trans_Amt	0.197028
Total_Trans_Ct	0.155504
Total_Ct_Chng_Q4_Q1	0.109037
Total_Revolving_Bal	0.098930

We will consider the top four features that the Random Forest model used to predict churning customers. As we can see, the "Total_Trans_Amt" feature has an importance of 19.70%, which is the most important feature as per the model, followed by "Total_Trans_Ct" with 15.55%, "Total_Ct_Chng_Q4_Q1" with 10.90%, and "Total_Revolving_Bal" with 9.89%.

4.4.2 The AdaBoost Model

Let's see the importance of features for the AdaBoost model. In AdaBoost algorithm, feature importance is calculated by the amount that each feature helps to reduce the weighted error.

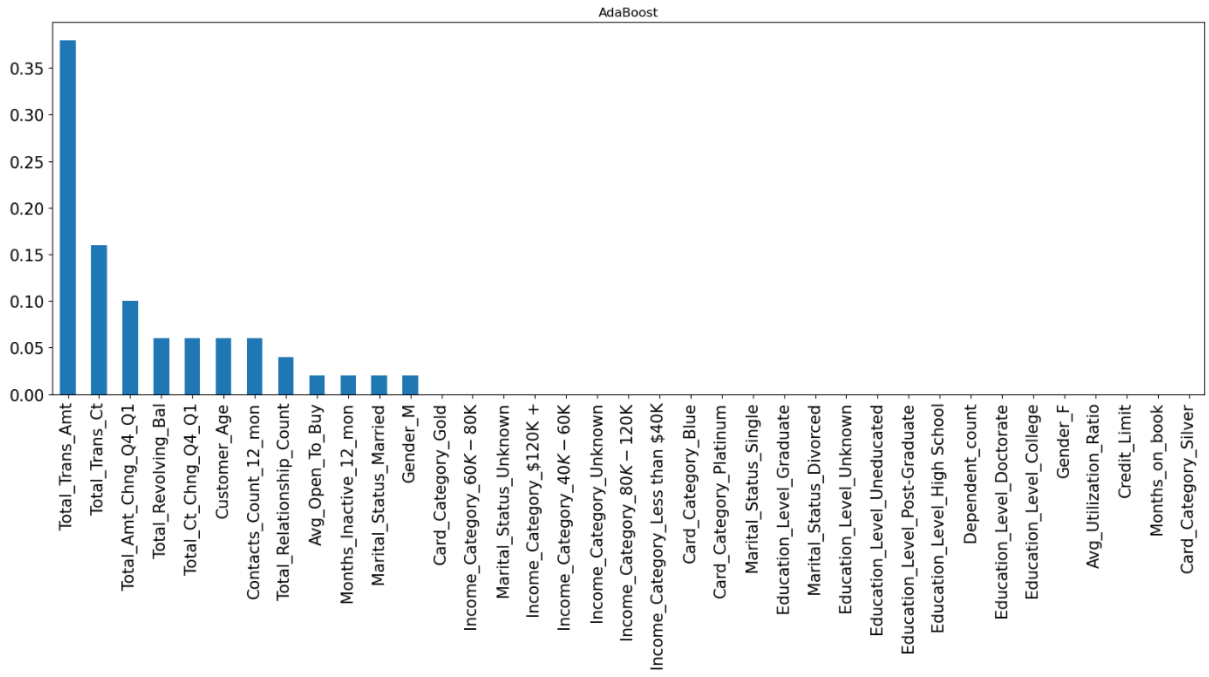

Figure 33: Important Features for the AdaBoost Model

Table 13: Score of each feature for the AdaBoost Model

Feature	Importance
Total_Trans_Amt	0.38
Total_Trans_Ct	0.16
Total_Amt_Chng_Q4_Q1	0.10
Total_Revolving_Bal	0.06

Also, this time we will consider the top four features that the AdaBoost model used to predict churning customers. As we can see, the "Total_Trans_Amt" feature is the most important feature for the model, with an importance of 38%. After this, "Total_Trans_Ct" has an importance of 16%, "Total_Amt_Chng_Q4_Q1" has an importance of 10%, and "Total_Revolving_Bal" has an importance of 6%.

CHAPTER 5: CONCLUSION

The study ultimately provides a detailed examination of credit card customer churn, where customer churn is predicted using numerous machine learning algorithms. Main determinants affecting customer turnover: gender, education level, marital status, income, card type, transaction behaviour. Based on the data, churned clients are dominated by women, graduates and people earning less than $40K. An additional relationship between age and customer churn was noted, where retention began to decline by 29 years old.

This study tested several machine learning models, including AdaBoost, Decision Tree, Random Forest, Logistic Regression, Neural Network, and Naive Bayes. The Random Forest and AdaBoost models had the best F1-scores with superior cross-validation metrics and were the highest performing. While examining the importance of all features, we obtained "Total_Trans_Amt" as the best predictor for both models because of their top scoring feature, followed by "Total_Trans_Ct" and "Total_Ct_Chng_Q4_Q1" and "Total_Amt_Chng_Q4_Q1" and "Total_Revolving_Bal". We may give the Random Forest model preference as it is the best as compared to the AdaBoost model. We should keep in mind that both models have very little difference in terms of scores.

Knowing how customers behave and why is what will give you the right compass for your retention strategy. Loyalty programmes, discounts, and personalised offers will be the ways businesses try to get more transactions through and engage more customers these next few months. This shift to visibility into transaction and volume changes from quarter to quarter gives companies a heads-up on customers in trouble so they can be proactive instead of reactive. Businesses may use the revolving balance of their customers to determine the state of their financial health and offer advice or solutions, such as a grace period, to indebted consumers in a tight financial position where they are struggling to pay their dues on time. Ultimately, you need to find and train high-value customers that cost more to acquire, which also brings in more subscribers. If properly designed and executed, these moves would result in happier customers, longer retention, and, therefore, more business. The thesis has contributed a comprehensive model to conduct a holistic approach towards understanding customer engagement and retention and, hence, can be of huge benefit to the pool of businesses fighting to survive in this era of cutthroat competition.

Further the study made use of Stratified K-fold cross-validation to address class imbalance in the dataset which added more robustness to the study. Thus, this research contributes by enhancing our understanding of customer attrition, and it reminds credit card corporations of important choices they can make in order to improve their customer retention strategies. Future studies could explore other factors that may affect customer churn and the research on how the different intervention strategies work.

References

[1] K. D. Ünlü, "Predicting credit card customer churn using support vector machine based on Bayesian optimization," *Communications Faculty of Sciences University of Ankara Series A1: Mathematics and Statistics*, vol. 70, no. 2, pp. 827–836, Dec. 2021, doi: 10.31801/cfsuasmas.899206.

[2] A. Azzopardi and J. Azzopardi, "Predicting Customer Behavioural Patterns using a Virtual Credit Card Transactions Dataset," *OAR@UM (University of Malta)*, Jan. 2022, doi: https://doi.org/10.5220/0011342300003280.

[3] D. AL-Najjar, N. Al-Rousan, and H. AL-Najjar, "Machine Learning to Develop Credit Card Customer Churn Prediction," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 17, no. 4, pp. 1529–1542, Nov. 2022, doi: https://doi.org/10.3390/jtaer17040077.

[4] M. Kumara, B. Kumar, and A. Mudhol, "Machine Learning based Prediction of Customer Churning in Banking Sector," *IEEE Conference Publication | IEEE Xplore*, Nov. 24, 2022. https://ieeexplore.ieee.org/abstract/document/10011126

[5] V. Chang, X. Gao, K. Hall, and E. Uchenna, "Machine learning techniques for predicting customer churn in a credit card company," *IEEE Conference Publication | IEEE Xplore*, Sep. 01, 2022. https://ieeexplore.ieee.org/document/10077145

[6] J. Panduro-Ramirez, S. V. Akram, Ch. Srinivasa. Reddy, J. M. Ruiz-Salazar, B. Kanwer, and R. Singh, "Implementation of Machine Learning Techniques for predicting Credit Card Customer action," *IEEE Xplore*, Jul. 01, 2022. https://ieeexplore.ieee.org/document/9914238 (accessed Jun. 26, 2023).

[7] B. Prabadevi, R. Shalini, and B. R. Kavitha, "Customer churning analysis using machine learning algorithms," *International Journal of Intelligent Networks*, vol. 4, pp. 145–154, Jan. 2023, doi: 10.1016/j.ijin.2023.05.005

[8] A. Arram, M. Ayob, M. A. A. Albadr, A. Sulaiman, and D. Albashish, "Credit card score prediction using machine learning models: A new dataset," *arXiv.org*, Oct. 15, 2023. https://arxiv.org/abs/2310.02956 (accessed May 15, 2024).

[9] S. Wang and B. Chen, "Credit card attrition: an overview of machine learning and deep learning techniques," *Informatika. Èkonomika. Upravlenie*, vol. 2, no. 4, pp. 0134–0144, Nov. 2023, doi: 10.47813/2782-5280-2023-2-4-0134-0144

[10] H. Zhu, "Bank Customer Churn Prediction with Machine Learning Methods," *Advances in Economics, Management and Political Sciences*, vol. 69, no. 1, pp. 23–29, Jan. 2024, doi: 10.54254/2754-1169/69/20230773.

[11] R. P. R. Kumar, B. Sahithi, K. Neeharika, M. Shivaleela, D. Singh, and K. R. K. Reddy, "Automation of Credit Card Customer Churn Analysis using Hybrid Machine Learning Models," *E3S Web of Conferences*, vol. 430, p. 01034, Jan. 2023, doi: 10.1051/e3sconf/202343001034.

[12] V. Agarwal, S. Taware, S. A. Yadav, D. Gangodkar, A. Rao, and V. K. Srivastav, "Customer - Churn Prediction Using Machine Learning | IEEE Conference Publication | IEEE Xplore," *ieeexplore.ieee.org*, Oct. 10, 2022. https://ieeexplore.ieee.org/document/9988187

[13] A. Saini, "Naive Bayes Algorithm: A Complete guide for Data Science Enthusiasts," *Analytics Vidhya*, Sep. 16, 2021. https://www.analyticsvidhya.com/blog/2021/09/naive-bayes-algorithm-a-complete-guide-for-data-science-enthusiasts/

[14] V. ANAND, "Introduction to Neural Networks," *Analytics Vidhya*, Jan. 31, 2022. https://www.analyticsvidhya.com/blog/2022/01/introduction-to-neural-networks/

[15] A. Saini, "Logistic Regression | What is Logistic Regression and Why do we need it?," *Analytics Vidhya*, Aug. 03, 2021. https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/

[16] A. Chauhan, "Credit Card Customers Prediction," *www.kaggle.com*. https://www.kaggle.com/datasets/whenamancodes/credit-card-customers-prediction/data

[17] T. C. Tran and T. K. Dang, "Machine Learning for Prediction of Imbalanced Data: Credit Fraud Detection," *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, Jan. 2021, doi: https://doi.org/10.1109/imcom51814.2021.9377352.

[18] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *The Stata Journal: Promoting Communications on Statistics and Stata*, vol. 20, no. 1, pp. 3–29, Mar. 2020, doi: https://doi.org/10.1177/1536867x20909688.

[19] P. M. Saanchay and K. T. Thomas, "An Approach for Credit Card Churn Prediction Using Gradient Descent," pp. 689–697, Jan. 2022, doi: https://doi.org/10.1007/978-981-16-3945-6_68.

[20] S. E R, "Random Forest | Introduction to Random Forest Algorithm," *Analytics Vidhya*, Jun. 17, 2021. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[21] M. Habibpour *et al.*, "Uncertainty-aware credit card fraud detection using deep learning," *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106248, Aug. 2023, doi: https://doi.org/10.1016/j.engappai.2023.106248.

[22] A. Saini, "AdaBoost Algorithm - A Complete Guide for Beginners," *Analytics Vidhya*, Sep. 15, 2021. https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/

[23] S. Yang and H. Zhang, "Comparison of Several Data Mining Methods in Credit Card Default Prediction," *Intelligent Information Management*, vol. 10, no. 05, pp. 115–122, 2018, doi: https://doi.org/10.4236/iim.2018.105010.

[24] E. Beauxis-Aussalet and L. Hardman, "Visualization of Confusion Matrix for Non-Expert Users," 2020. Available: https://vis.cs.ucdavis.edu/vis2014papers/VIS_Conference/infovis/posters/beauxis-aussalet.pdf

[25] A. Bhandari, "Confusion matrix for machine learning," *Analytics Vidhya*, Apr. 17, 2020. https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/

[26] S. Nembrini, I. R. König, and M. N. Wright, “The revival of the Gini importance?,” *Bioinformatics*, vol. 34, no. 21, pp. 3711–3718, May 2018, doi: https://doi.org/10.1093/bioinformatics/bty373.