Blue text on a black background

Description automatically generated

**WQD7005 DATA MINING**

**PROJECT REPORT**

**SEMESTER 2, SESSION 2023/2024**

**DATA MINING APPROACH IN ANALYZING CHURN OF CREDIT CARD CUSTOMER**

**OCC 2 | Group G3**

**Teacher:**

**Tutut Herawan**

|  |  |
| --- | --- |
| **Name** | **Matric Number** |
| SAGAL MOHAMED YUSUF | 22093010 |
| TARSVINI A/P RAVINTHER | 17193844 |
| RABITA BHUIYA TANAYA | 22111473 |
| ZHANGXUAN | 22091244 |

**Abstract:**

In today's digitally driven financial landscape, credit card usage is ubiquitous, providing unparalleled convenience and flexibility for both consumers and businesses. Credit card firms face a major threat to their profit margins and brand loyalty from customer churn, which occurs when customers stop their engagement with a company. This research initiative explores a data mining approach to analyze credit card customer behavior and predict churn. The study aims to (1) identify usage patterns and spending behaviors that correlate with customer churn, (2) develop a predictive model using machine learning techniques to estimate the likelihood of churn, and (3) evaluate the model's performance in predicting actual customer churn. Our findings indicate that the Random Forest algorithm provided the highest accuracy score (95%) while Gradient Boosting delivered the best recall score (94%) on an unbalanced dataset. Upon applying the Synthetic Minority Over-sampling Technique (SMOTE), Random Forest's accuracy slightly decreased by 1%, yet the recall score of Gradient Boosting remained consistent at 94%. Given the comparable accuracy scores and based on the importance of recall in unbalanced datasets, we selected Gradient Boosting as the optimal model for predicting credit card customer churn. By leveraging advanced analytics and predictive modeling, credit card companies can better understand the factors driving customer churn. This understanding enables the development of tailored retention strategies that enhance customer loyalty and support sustainable business growth. Ongoing research and adaptation to evolving market trends are crucial for maintaining a competitive edge in the dynamic landscape of credit card services.

**Keywords:** predict churn, customer churn, usage pattern, customer loyalty

1. **Introduction**

Credit card use has become ubiquitous in today's rapidly evolving financial landscape, transcending geographical and socioeconomic boundaries. These plastic cards serve as highly practical tools, facilitating a wide range of monetary transactions, from small purchases to large ones. Offering unparalleled convenience and flexibility to both consumers and businesses, credit cards have become indispensable for economic transactions in the era of digitization. However, alongside their widespread adoption, credit card companies face the formidable challenge of customer attrition.

Customer churn, the phenomenon where customers cease their relationship with a company, poses a significant problem in the face of expanding marketplaces and heightened competition among enterprises. This concern permeates various sectors, including banking and telecommunications, where businesses strive to retain their customer base. In the context of credit cards, churn occurs when customers choose to cancel their credit card accounts, often migrating to different card issuers or discontinuing credit card usage altogether. [6] Guliyev and Tatoğlu As large companies endeavor to retain both existing and new clients, customer churn analysis emerges as a crucial strategy for identifying and retaining valuable customers, introducing new products, and making informed customer retention decisions.

The impact of customer churn extends beyond immediate financial considerations, affecting credit card companies' bottom lines and influencing spending habits across the banking industry. With customers leaving the company, credit card companies experience a decline in profit margins, exacerbating the financial burden of attracting new customers. Moreover, churn undermines consumer trust and brand loyalty, posing strategic challenges for businesses striving for sustainable growth and market resilience.

Against this backdrop, our research initiative aims to explore and analyze credit card customer behavior to develop predictive models for churn identification. Drawing upon insights from previous studies, including machine learning approaches by [1], Muneer, *et al*. and AL-Najjar, *et al*., [2], and experimental analyses of hyperparameters by [3] Domingos, *et al*., our study seeks to leverage advanced analytics and predictive modeling techniques to delineate the intricate web of factors contributing to churn.

Furthermore, by evaluating customer behavior with temporal centrality metrics [9] Calzada-Infante, et al. and integrated churn prediction frameworks [10] Wu et al, our research aims to enhance the predictive accuracy and interpretability of churn prediction models. By delineating the complex interplay of factors driving churn, credit card companies can develop tailored retention strategies that foster long-term customer loyalty and sustainable growth.

Our research initiative is motivated by the dual imperatives of consumer interest protection and market competitiveness enhancement for credit card companies. Through a data-driven approach, we seek to unravel the intricacies of credit card user behavior and make accurate churn predictions, thereby facilitating informed decision-making and fostering mutually beneficial relationships between credit card companies and their customers. As market trends and consumer preferences continue to evolve, ongoing research efforts are essential to stay abreast of the latest developments and maintain a competitive edge in the dynamic landscape of credit card services.

Analyzing credit card customer behavior for churn prediction is incredibly significant because customers represent the cornerstone of any business [21]. It costs six times less to maintain the loyalty of an existing customer than to acquire a new one [22][12]. The ability to forecast customer attrition enables companies to fine-tune their retention strategies, reducing losses and improving marketing strategies [23].

This project embarks the following objectives:

1. Identify how frequently a credit card is used, how much it costs on average, and what kinds of purchases are made, and how these factors relate to customer churn rates over time.
2. Create a predictive model using machine learning techniques to estimate the likelihood that credit card users will churn. The model will be based on transactional data from the past, consumer interaction patterns, and demographic data.
3. Evaluate the performance of the model in predicting customer churn by comparing the predictions with real instances of customer churn throughout a designated timeframe.

The scope of the project should encompass the following aspects:

1. Integrated Data Analysis: To find out what causes customers to leave, this study will examine a lot of data, including demographics, consumer interaction patterns, and transactional data.
2. Client Category: To better target certain demographics, the study will divide customers into several groups and analyze the impact of various retention strategies and characteristics on each.
3. Evaluation of Retention Techniques: This evaluation determines which customer retention strategies work best in lowering churn rates across different demographics.
4. Predictive Modelling: By analyzing past data, we may create a model that uses machine learning to estimate the probability of client attrition.
5. Ongoing research: Due to the ever-changing nature of consumer preferences and market trends, this project must be continuously studied to include new insights and trends.

Even though there are several aspects related to customer churn, this project may not include them as the main focus:

1. Macroeconomic conditions: Recession periods, inflation rates, and employment levels are factors that undoubtedly influence customer behavior, but they are outside the typical scope of individual customer data analysis.
2. Regulatory changes: Changes in regulations and data usage policies affecting financial institutions may have an indirect impact on customer churn.
3. Competitive landscape analysis: This requires a different set of data and analytical frameworks that focus on market analysis rather than individual customer behavior.

At the same time, the project data also has certain limitations.

1. The first is that the data is based on data collected during a specific period. Within this time frame, seasonal changes or special events (e.g., economic crisis, pandemic) may distort results.
2. Secondly, the collection of data is limited to a specific financial institution, which is affected by differences in consumer behaviour, economic and cultural backgrounds in different regions. The availability of the model may also limit the geographical coverage of the research.
3. **Literature Review**

Customer churn, the departure of customers from a business, is a critical concern for banks. Understanding and predicting churn is essential for retaining customers and informing strategic decisions. Recent advancements in machine learning (ML) have transformed churn analysis, with a focus on explainable ML models for transparency. Traditional methods like logistic regression and decision trees have been supplanted by ML algorithms like support vector machines and gradient boosting. These algorithms handle large, complex datasets better and offer more accurate predictions. However, concerns about model interpretability have led to the development of explainable ML approaches.

In [1], Muneer, *et al*., applied machine learning techniques such as AdaBoost, Random Forest (RF), and Support Vector Machine (SVM) to forecast customer churn within the banking sector. They utilized a dataset comprising credit card customer churn data and employed the synthetic minority oversampling technique (SMOTE) to address the issue of imbalanced datasets which involves employing a combination of under sampling and oversampling techniques. Based on the results, RF has predictive of 88.7% and an F1 score of 0.91. It outperformed the other two models.

In another study by AL-Najjar, *et al*., [2], feature selection approach alongside five distinct machine learning models. including Bayesian Network, chi-square automatic interaction detection (CHAID) tree, classification and regression (CR) tree, C5 tree and a neural network were utilized to predict credit card customer churn. Based on the results, C5 tree machine learning model performed the best. The findings of the study highlighted three key variables essential for developing a model for credit card customer churn prediction: total transaction count, total evolving balance on the credit card, and the change in transaction count. Additionally, merging multiple categorical variables into a single variable was shown to enhance the model's performance.

In [3], Domingos, *et al*., used the deep neural network (DNN) model to forecast customer churn within the banking sector. Analysis of the effect of different hyperparameters when using DNN model is presented in this paper. The findings indicate that the Deep Neural Network (DNN) model surpassed the Multilayer Perceptron (MLP) when employing a rectifier function for activation in the hidden layer and a sigmoid function in the output layer. Additionally, the DNN showed improved performance with smaller batch sizes compared to the size of the test dataset. Among the tested training algorithms, RemsProp demonstrated superior accuracy.

In [4], Ahmet and Kozłowska conducted a prediction on Telco customer churn by using various data mining techniques and classification algorithms such as Decision Tree, Random Forest, logistic regression, *K-*nearest neighbor, and Support Vector Classifier. Logistic Regression provided the best accuracy score (81%) and Decision Tree provided the best recall score (57%) on unbalanced data. Upon detecting the imbalance in the data, the researchers employed the SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset. Then the logistic regression model was tested again, and the accuracy score improved by 3% and recall score improved by 30%. Although accuracy metric is important, recall was considered as the more important metric for model performance evaluation due to unbalanced data in target class. The study suggested to choose model with a better recall score if there is no significant difference in accuracy scores. Hence, decision tree was recommended as a model when using unbalanced data.

In [5], de Lima Lemos, *et al*., has done horserace between many supervised machine learning models including decision tree, elastic net, *k-*nearest neighbors, random forests, logistic regression, SVMs, and ensemble models consisting of above models. The study utilized the same cross-validation and evaluation setup across all experiments to ensure fair comparison between algorithms. Random forests emerged as the top-performing model in terms of predictive accuracy. Additionally, the study identified key attributes strongly associated with potential customer churn, including the frequency of financial service usage (measured by the number of transactions), the volume of credit extended, and the range of banking products possessed by customers.

According to [6], Guliyev and Tatoğlu studied that explainable ML models, like the XgBoost model, outperform others in identifying churn customers. By incorporating SHAP values, these models offer insights into the factors driving churn, aiding in targeted retention efforts. Challenges remain, including data privacy and model monitoring, but continued research is vital for improving churn prediction and customer retention in banking.

Customer churn prediction in e-commerce is crucial for maintaining business sustainability and growth. In this study [7] Xiahou and Harada focuses on the use of AdaBoost classifier and BP neural network techniques in churn prediction models. *K-*means clustering is commonly employed to segment customers based on behavior and characteristics, enhancing the accuracy of churn prediction models. AdaBoost, an ensemble learning method, is effective in identifying potential churners and outperforms BP neural networks in predictive accuracy. Integrating *k-*means clustering with AdaBoost further improves performance. While BP neural networks can capture complex patterns in customer behavior, they may require extensive parameter tuning. However, AdaBoost exhibits superior performance in churn prediction tasks. The integration of advanced AI techniques like AdaBoost and BP neural networks enables proactive customer retention strategies in e-commerce. Continued research is needed to address challenges and optimize the implementation of churn prediction solutions for sustained business success.

Data mining techniques are widely applied in financial transactions analysis, particularly for fraud detection and customer behavior understanding. Machine learning algorithms like artificial neural networks (ANN) and gradient boosting (GBM) are commonly employed for predictive analytics in this domain. Customer churn prediction, a vital aspect in various industries, including finance, relies heavily on data mining. Machine learning algorithms such as ANN, decision trees, logistic regression, and ensemble methods like GBM are commonly used for churn prediction tasks. In this study [8], Azzopardi and Azzopardi emphasize the importance of integrating diverse datasets, including demographic information and transactional data, to enhance the accuracy of churn prediction models. Challenges such as the "cold start" problem and determining minimum activity thresholds for predicting churn are also addressed in the literature. leveraging data mining techniques in financial transactions analysis facilitates actionable insights for businesses, especially in predicting customer churn. The study conducted on a virtual credit card transactions dataset in Malta contributes by evaluating machine learning techniques for churn prediction and addressing pertinent challenges in the field.

In the competitive telecommunications sector, retaining customers is vital for profitability. Predicting customer churn, often through binary classification methods, is crucial. Recent research emphasizes the role of social network analysis (SNA) in understanding churn, considering social relationships' influence. According to this study [9] Calzada-Infante, et al. proposes a novel approach integrating SNA techniques with similarity forests, using call detail records to construct temporal graphs. Centrality metrics are pivotal in quantifying individual customer importance within the network. The comparison of centrality metrics on two types of temporal graphs—Time-Order Graph and Aggregated Static Graph—reveals insights into churn prediction effectiveness. Previous studies also explore temporal data and time series analysis to capture evolving customer behavior. Overall, research underscores the value of combining individual behaviors and social interactions to improve churn prediction accuracy and inform retention strategies. This article's methodology contributes significantly to this field, offering nuanced insights into customer behavior dynamics for churn prediction.

The telecommunications industry prioritizes customer retention due to its cost-effectiveness over customer acquisition. In this research [10], Wu et al discussed “Integrated Churn Prediction and Customer Segmentation Framework for Telco Business" proposes a comprehensive approach to churn management by combining churn prediction and customer segmentation. Churn prediction, crucial for identifying potential churners, employs six machine learning classifiers, including AdaBoost and Random Forest. Synthetic Minority Oversampling Technique (SMOTE) addresses imbalanced datasets, enhancing predictive accuracy. Experimental results show varying classifier performance across datasets, with AdaBoost, Random Forest, and Multi-layer Perceptron exhibiting notable performance in different scenarios. Bayesian Logistic Regression conducts factor analysis to identify significant churn-related features, guiding customer segmentation using *K-*means clustering. This integration enables tailored retention strategies for specific customer segments, enhancing overall churn management effectiveness. The framework offers telco operators actionable insights to mitigate churn and improve customer loyalty, contributing to optimized retention efforts in the telco sector.

Customer churn prediction is crucial for B2C e-commerce, facilitating effective retention strategies and targeted marketing. In this study [11], Xiahou and Harada combine *k-*means clustering with SVM prediction to enhance churn forecasting. Segmentation techniques like *k-*means clustering have been pivotal in improving churn prediction models by identifying distinct customer groups. Additionally, SVM's superiority over traditional methods like logistic regression in predictive performance has been noted. Previous research underscores the significance of factors like product satisfaction and pricing strategies in churn prediction. Challenges include data quality and model interpretability. Future research should focus on overcoming these challenges to refine churn prediction models further. In summary, the proposed hybrid model offers valuable insights for customer relationship management in e-commerce enterprises, emphasizing the importance of continued research in churn prediction methodologies.

The study "Ensemble Methods in Customer Churn Prediction: A Comparative Analysis of the State-of-the-Art" addresses the growing interest in using ensemble methods for predicting customer churn in CRM. In this study [12], Bogaert & Delaere highlights the underutilization of heterogeneous ensembles despite their potential superiority over homogeneous ensembles and single classifiers. Prior research primarily focused on individual classifiers or homogeneous ensembles, neglecting the potential benefits of heterogeneous ensembles. However, this study fills the gap by conducting a comprehensive benchmarking of 33 classifiers across 11 datasets, incorporating novel methods in performance evaluation and model comparison. The findings demonstrate the superiority of heterogeneous ensembles, particularly those optimized by simulated annealing classifier selection and non-negative binomial likelihood, in terms of predictive accuracy and generalization. By integrating diverse classifiers and evaluation metrics, the study contributes to a deeper understanding of churn prediction and informs decision-making in real-world applications.

Customer churn prediction is crucial for businesses, yet traditional models struggle due to imbalanced data. In this article [13], Li et al., introduces the Focal Loss hard example mining technique to enhance identification of potential churners. By augmenting LightGBM with class weights and a focus parameter, the proposed approach aims to improve accuracy and stability. Empirical results demonstrate its superiority over existing methods. Despite challenges, such as class imbalance, this novel approach represents a promising advancement in churn prediction research. Further exploration of techniques to enhance model accuracy is warranted.

In article [14], ÜNLÜ, K. D. introduces a hybrid machine learning algorithm combining support vector machine (SVM) and Bayesian optimization (BO) to predict the churn of bank credit card customers. And use four different evaluation indicators (accuracy, precision, recall and F1 score) to compare the prediction performance of different models. The results show that although SVM has potential when dealing with complex data sets, the choice of kernel function and super Optimization of parameters is also critical to the success of the model. In addition, this paper also highlights that although Bayesian optimization provides a flexible and effective hyperparameter tuning method, choosing the appropriate model and parameter settings is still the key to ensuring predictive performance. At the same time, it is said that future research can consider comparing different hyperparameter optimization tools in predicting customer churn.

In article [15], Rahman, M., & Kumar, V. proposes a method to predict bank customer churn using machine learning technology, which is a branch of artificial intelligence. This study facilitates the exploration of customer churn possibilities by analyzing customer behavior. KNN, SVM, decision tree and random forest classifiers were used in the study. Furthermore, each model was evaluated by the accuracy obtained after 10-fold cross-validation. And a random confusion matrix was also generated for each model. The results show that the RF classifier works better when used with oversampling. Feature selection methods are independent of tree classifiers (decision trees and RF).

In the article [16] by Wu, C., & Wang, L., credit card is used as the background, the necessary preprocessing is performed on the data, and various modern churn models (including logistic regression, random forest and XGBoost models) are established to achieve the dual goals of accuracy and interpretability. Preprocessing includes resampling of imbalanced data (synthetic minority oversampling technique) and feature selection. and provides an overview of the most important churn drivers. Then the baseline model, random forest and XGBoost were compared. Results show that XGBoost and RandomizedSearchCV provide the best predictive capabilities on a selected set of performance metrics

Besides, in [17], Fujo, et al., has implemented a deep learning model called deep backpropagation artificial neural network (Deep-BP-ANN) in this study to predict customer churn in telecommunication industry. This study has used 2 datasets related to telecom (IBM Telco and Cell2cell). Random Oversampling method was used to balance the datasets. The study mentions that the model outperformed Machine Learning Techniques such as Logistic Regression, Naïve Bayes, XG\_Boost, and KNN in forecasting churn in customer. When tested with holdout or 10 Cross Validation method with the same datasets, this model’s accuracy is better than existing deep learning models.

In the article [18], Dorokhov et al., illustrates the process of constructing predictive models that classify users based on their probability of churning, using a cellular service provider as an example. Customer analytics is responsible for understanding and analyzing the commercial activity of customers. The main factors contributing to client attrition are presented, together with the specific details of attrition for enterprises associated with subscription and transactional business models, which involve regular customer payments. Special attention is placed on the analysis of forecasting strategies based on classification techniques. The article focused on forecasting models that utilize the Bayesian network and decision tree methodologies. The C5.0 algorithm served as the fundamental basis for the decision tree methodology. The Bayesian model is constructed for a Naive and Markov structure. Customer service has been identified as a crucial factor in client attrition across all three models. The models were subjected to a comparative analysis using the AUC and Gini metrics. The decision tree model exhibited the most optimal outcomes. Furthermore, the decision tree model offers insights for tailoring a personalized strategy for each customer and elucidates the factors that may lead a consumer to discontinue their engagement with the firm. SPSS Modeler was a software program used for generating models.

Electronic commerce is one of the many fields that sees daily innovation in data. Therefore, data mining techniques are necessary in this field. Tehran is the top online meal delivery service in Iran, and this article [19] by S. Raeisi & H. Sajedi has used its dataset to explain it. In addition to retaining customers, data analysis can reveal why customers were leaving. Businesses that are growing must prepare ahead of time to retain customers since client attrition is a significant performance metric. The purpose of this article is to provide a framework for predicting customer churn based on online activity and property attributes. The results of different data mining approaches have been compared through several tests. The results show that Gradient Boosted Trees outperform the other approaches with an accuracy of 86.90%.

Based on article [20], Kim & Lee aimed to predict the rate of client attrition in influencer commerce. After endorsing something on social media, influencers will often link to the product's website, allowing their followers to buy the product straight from their account. One form of online trade is influencer commerce. Advertising products and services on social media sites like Twitter, Instagram, and Facebook is what influencers do for a living. This study's customer churn estimate is based on the premise that influencers have passionate support from their followers. The Korean influencer marketing agency tracked all purchases, from August 2018 to October 2020, including the total money paid, the specific item purchased, and the details of the customer. They used the Decision Trees (DT) method in the computer application RapidMiner to predict which clients will be leaving. The investigation showed that the greatest prediction accuracy is 90% based on the F-measure. Predicting customer turnover from an influencer's point of view is made easier by this study.

**Table 1.** Related Works

|  |  |  |  |
| --- | --- | --- | --- |
| **Authors & Year** | **Method** | **Advantages** | **Disadvantages** |
| Muneer, et al., [1] | 1. Random Forest (RF) 2. AdaBoost 3. Support vector machine (SVM) 4. SMOTE Technique. | This study made comparison between the 3 proposed models with related literature contributions (KNN, XG Boost, Naïve Bayes, Decision Trees, Random Forest, ANN) in aspect of Recall, F1 Score and Accuracy %. | Increased Computational Complexity:  Up sampling techniques like SMOTE increase the size of the dataset by generating synthetic instances, which can significantly increase computational overhead during model training. |
| AL-Najjar, et al., [2] | 1. Feature-selection method. 2. Bayesian Network. 3. Chi-square automatic interaction detection (CHAID) tree. 4. The classification and regression (CR) tree. 5. The C5 tree. 6. Neural network. | Model Flexibility:  The models employed in the study include:  Model 1: Utilizing all independent variables.  Model 2: Employing all continuous variables along with cluster values obtained through two-step clustering and *k-*nearest neighbor.  Model 3: Incorporating selected variables identified through the feature-selection method. | Potential Information Loss:  While feature selection techniques like those used in Model 3 aim to identify the most important variables, there is a risk of information loss by discarding marginal or unimportant features. |
| Domingos, et al.,[3] | 1. Deep neural network (DNN) model. | Enhanced Understanding of Hyperparameters:  The study enriches theoretical understanding by investigating the impact of diverse hyperparameter configurations on the efficacy of deep neural networks (DNNs) for predicting churn in the banking sector. | Dataset was unbalanced  (2000 churners, 8000 non-churners) |
| Çalış and Kozłowska, [4] | 1. Logistic regression. 2. *K-*nearest neighbor. 3. Decision Tree. 4. Random Forest. 5. Support Vector Classifier. 6. SMOTE technique | Better Utilization of Unbalanced Data:  By prioritizing recall over accuracy, the study acknowledges the importance of correctly identifying churn instances, especially in scenarios where the class distribution is highly imbalanced. | Limited Scope:  While the dataset provides valuable insights into customer behavior within the Telco communication company, its focus on customers in California during a specific quarter may limit the generalizability of the study's findings |
| de Lima Lemos, et al., (2022) [5] | 1. Decision trees. 2. Logistic regression. 3. *K-*nearest neighbors. 4. Elastic net. 5. SVMs 6. Random forests. | Rich dataset from a large Brazilian bank - 500,000 records. | Resource Intensive:  Random forests are computationally intensive models, especially when dealing with large datasets or many features. |
| Guliyev and Tatoğlu,[6] | Explainable Machine Learning (ML) models, specifically XgBoost model  Incorporating SHAP (SHapley Additive exPlanations) values | The XgBoost model, along with SHAP values, offers accurate predictions of churn customers, helping banks identify at-risk customers more effectively. | Analyzing customer data to predict churn raises privacy concerns, as sensitive information may be used without explicit consent. Ensuring compliance with data protection regulations and maintaining customer trust is crucial. |
| Xiahou and Harada, [7] | AdaBoost classifier  BP neural network  *K-*means clustering | AdaBoost classifier:  Effectively identifies potential churners.  BP neural network:  Capable of capturing complex patterns in customer behavior.  *K-*means clustering:  Customers segment based on behavior and characteristics. | AdaBoost classifier:  May require careful parameter tuning for optimal performance.  BP neural network:  Requires extensive parameter tuning for optimal performance.  *K-*means clustering:  Sensitive to the choice of initial cluster centroids. |
| Azzopardi and Azzopardi (2022) [8] | Machine learning algorithms like artificial neural networks (ANN)  Gradient boosting (GBM)  Decision trees  Logistic regression  Ensemble methods like GBM | Ability to handle large and complex datasets.  Capability to identify complex patterns and relationships in the data.  Flexibility to incorporate diverse types of data, including demographic information and transactional data | Some machine learning algorithms, such as ANN and GBM, can be complex and challenging to interpret, requiring specialized expertise for implementation and tuning. |
| Calzada-Infante, et al. (2020) [9] | Social Network Analysis (SNA) techniques with similar forests. | Integration of social network analysis with similarity forests provides a holistic understanding of customer churn dynamics.  Temporal graphs allow for the capture of evolving customer behavior over time, enabling more accurate churn prediction. | Implementing the proposed approach may require a high level of technical expertise and computational resources. |
| Wu et al. (2021) [10] | AdaBoost  Random Forest  Multi-layer Perceptron | The framework combines churn prediction and customer segmentation, providing a holistic solution to churn management.  By employing multiple classifiers, the model can capture different aspects of churn prediction, potentially improving overall performance. | The experimental results show varying performance across different classifiers and datasets, indicating that certain classifiers may be more effective in specific scenarios. This variability may introduce uncertainty in model selection and deployment. |
| Xiahou and Harada(2022a) [11] | *K-*means clustering combined with SVM prediction for churn forecasting in B2C e-commerce. | Combining *k-*means clustering with SVM prediction can improve the accuracy and effectiveness of churn forecasting models by identifying distinct customer segments and leveraging SVM's predictive performance. | One of the challenges mentioned is data quality, which can affect the performance and reliability of the churn prediction model. Poor-quality data, such as missing values or inaccuracies, may lead to biased or erroneous predictions. |
| Bogaert & Delaere (2023) [12] | It investigates the use of heterogeneous ensembles, which combine diverse classifiers, for improved predictive accuracy and generalization. | The study demonstrates that heterogeneous ensembles, particularly those optimized by simulated annealing classifier selection and non-negative binomial likelihood, outperform other models in terms of predictive accuracy. This implies that combining diverse classifiers can lead to better churn prediction outcomes. | Heterogeneous ensembles can be more complex to implement and interpret compared to individual classifiers or homogeneous ensembles. Managing and optimizing a diverse set of classifiers may require more computational resources and expertise. |
| Li et al., (2023) [13] | Focal Loss/ LightGBM with class weights and a focus parameter | By augmenting LightGBM with class weights and a focus parameter, the proposed approach aims to improve both the accuracy and stability of the churn prediction model, leading to more reliable predictions. | Implementing the Focal Loss with hard example mining technique and augmenting LightGBM with class weights and a focus parameter may introduce additional complexity and implementation overhead, requiring expertise and resources for deployment. |
| ÜNLÜ, K. D. (2021) [14] | Model: Support Vector Machine (SVM)  Optimization method: Bayesian optimization (BO)  Evaluation metrics: accuracy, precision, recall and F1 score | SVM can effectively handle linearly inseparable and nonlinear problems by using different kernel functions. And Bayesian optimization can effectively optimize hyperparameters, reduce the need for manual adjustments, and is more efficient than traditional grid search and random search. | Different kernel functions have a significant impact on model performance, which can lead to difficulties in practical applications,  In addition, SVM requires high computing resources on large-scale data sets and is weak in the interpretation of nonlinear models. |
| Rahman, M., & Kumar, V. (2020, November) [15] | Models: KNN, SVM, decision tree and random forest classifiers. | The random forest model showed higher accuracy after Jin carried out sampling processing, and the use of the decision tree model can effectively improve the decision-making path of the Chiang Kai-shek model, making the whole process more interpretable. In addition, the method of using multiple models is more generalizable and can be applied to multiple data sets. | Random forests and support vector machines may require higher computing resources when processing large data sets, increasing computing costs. At the same time, parameter tuning of SVM and random forests may be complex, requiring precise methods to select appropriate parameters. SVM is more sensitive to data imbalance and may require additional data preprocessing steps to deal with this problem. |
| Wu, C., & Wang, L. (2022) [16] | Models: logistic regression, random forest, XGBoost: Parameter optimization using RandomizedSearchCV | The article uses SMOTE technology to improve the model's prediction ability on imbalanced data sets and uses RandomizedSearchCV for automatic parameter adjustment. The XGBoost model performs better on performance evaluation indicators and has higher prediction accuracy. | The article uses SMOTE technology to improve the model's prediction ability on imbalanced data sets and uses RandomizedSearchCV for automatic parameter adjustment. The XGBoost model performs better on performance evaluation indicators and has higher prediction accuracy. |
| Fujo, et al., [17] | 1. Deep-BP-ANN model  2. Feature selection methods such as  Variance Thresholding  & Lasso Regression  3. XG\_Boost  4. Logistic\_Regression  5. Naïve Bayes  6. KNN  7. Early stopping technique | Comparison with Other Studies: The study claims that the Deep-BP-ANN model not only outperforms traditional machine learning techniques but also surpasses other studies that have used deep learning techniques for similar tasks. This indicates that the implemented model may have certain architectural or methodological advantages over existing deep learning approaches. | Computational Resources: Training deep learning models can be computationally intensive, requiring significant amounts of computational resources, including high-performance GPUs or TPUs. This can be a disadvantage for organizations with limited computational infrastructure or budget constraints. |
| Dorokhov et al., [18] | 1. Decision Tree  2. Bayesian Network (Naive & Markov)  Platform:  SPSS Modeler | The decision tree model yields superior outcomes. Hence, it is more advantageous to employ the decision tree model for forecasting customer churn in a mobile service provider. | Main disadvantage is the use of only two model. Customer churn predictive modeling first used the Bayesian model, but the decision tree model exceeded it in accuracy and effectiveness.  As there are only two models to compare, it doesn't conclude to a concrete result. |
| S. Raeisi & H. Sajedi, [19] | 1. GBT  2. KNN  3. Naive Bayes  4. Decision Tree  5. Random Forest  6. Rule Induction Method | Using Gradient Boosted Trees, the highest accuracy achieved was 86.90%. | The current system's performance or capability is limited by the amount or depth of data available. |
| Kim & Lee, [20] | Model Used:  Decision Tree  Software Used:  RapidMiner  Evaluation Matrics:  Recall, Precision, Accuracy, F-measure | The Decision Trees (DT) model has a maximum prediction accuracy of 90% in terms of f-measure. | More research is required to use other algorithms to improve their practicality. |

1. **Dataset and Proposed Method**

This section described the dataset used for this project.

* 1. **Dataset**

The 'Bank Churners' dataset is sourced from Kaggle and provides a rich compilation of data from over 10,000 bank customers.

Dataset Overview:

* Total Entries: 10,127 rows
* Total Columns: 23 attributes

### Attributes:

1. CLIENTNUM: Unique identifier for the client.
2. Attrition\_Flag: Indicates if the customer left the bank or not (e.g., "Attrited Customer" or "Existing Customer").
3. Customer\_Age: Age of the customer.
4. Gender: Gender of the customer (e.g., "M" or "F").
5. Dependent\_count: Number of dependents the customer has.
6. Education\_Level: Educational background of the customer.
7. Marital\_Status: Marital status (e.g., "Married", "Single").
8. Income\_Category: Income category for the customer (e.g., "Less than $40K", "$40K - $60K").
9. Card\_Category: Type of card the customer holds (e.g., "Blue", "Silver").
10. Months\_on\_book: Duration (in months) that the customer has been with the bank.
11. Total\_Relationship\_Count: Total number of products the customer has with the bank.
12. Months\_Inactive\_12\_mon: Number of months the customer has been inactive in the last 12 months.
13. Contacts\_Count\_12\_mon: Number of contacts between the customer and bank in the last 12 months.
14. Credit\_Limit: Credit limit on the customer’s cards.
15. Total\_Revolving\_Bal: Total revolving balance on the customer’s cards.
16. Avg\_Open\_To\_Buy: Average open to buy credit line (total credit limit minus the revolving balance).
17. Total\_Amt\_Chng\_Q4\_Q1: Change in transaction amount (Q4 over Q1).
18. Total\_Trans\_Amt: Total transaction amount in the last 12 months.
19. Total\_Trans\_Ct: Total transaction count in the last 12 months.
20. Total\_Ct\_Chng\_Q4\_Q1: Change in transaction count (Q4 over Q1).
21. Avg\_Utilization\_Ratio: Average card utilization ratio.
22. Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_1: Naive Bayes Classifier output for attrition.
23. Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_2: Another Naive Bayes Classifier output for attrition.
    1. **Proposed Method**

For methodology, we are adhering to the structured approach SEMMA, developed by SAS Institute. SEMMA comprises five stages: Sample, Explore, Modify, Model, and Assess. Table 2 outlines how each step is applied in our project following the SEMMA methodology.

**Table 2:** SEMMA

|  |  |  |
| --- | --- | --- |
| **No** | **Step** | **Description** |
| 1 | Sample | We will collect a subset of the credit card customers data. This will be a comprehensive data set from credit card users. This data includes customer personal information, credit card limit, credit card balance, number of transactions, transaction amount, etc. |
| 2 | Explore | We will conduct data visualization and correlation analysis to better capture the nuances of customer behavior that lead to churn. |
| 3 | Modify | We will Identify and handle missing values, outliers, and duplicates, and transform raw data to make it more suitable for the model. |
| 4 | Model | We will divide training sets and test sets, build and train multiple models (Random Forest, Logistic Regression, Gradient Boosting Machines (XGBoost/LightGBM), and Neural Networks) to ensure that the models can generalize well to unseen data. |
| 5 | Assess | We will assess the performance of the models using evaluation metrics such as accuracy, precision, recall, and AUC. |

We have chosen python as data mining tool for our project because Python has extensive libraries for data manipulation, analysis, and visualization. For data manipulation and analysis, Pandas can be used as they have intuitive data structures. Besides, we can also do visualization tasks using matplotlib and seaborn as it provides high-quality plotting capabilities. Moreover, python also has a scikit-learn library for modelling and stats models for statistical analysis. We will use Google Collaboratory (Google Collab) to do this project. Google Collab is a cloud-based platform which provides seamless accessibility and sharing on Jupyter notebook. Team members can share their code in the same notebook.

Figure 1 below shows the proposed methodology.

**Figure 1**: Proposed Methodology

A diagram of a model

Description automatically generated

We are expecting some challenges in the project development. The first challenge is imbalanced data. Credit card customers churning data usually tends to be imbalanced, with a significantly higher number of non-churned customers compared to churned ones. This can lead to biased models that perform poorly in predicting the minority class. The second challenge is feature selection. Choosing the right features that are most predictive of churn can be challenging. It requires domain knowledge and experimentation to identify the most suitable and relevant features. The third challenge is related to data quality. Ensuring the data is consistent and without any missing values or faulty values is crucial as it can impact the model results.

* 1. **Expected Outcome**

The expected outcomes are:

1. Early identification of at-risk customers: By predicting churn, businesses can identify at-risk customers early and proactively engage with them through personalized offers, enhanced services, or loyalty programs to improve retention.
2. Understand the drivers of churn: Analysing the factors that lead to customer churn helps companies understand the factors that influence customer satisfaction. This knowledge can guide product, customer service, and overall user experience improvements.
3. Improve corporate profitability: By reducing customer churn, companies can significantly increase the lifetime value of their customers, thereby improving long-term profitability.
4. **Results and Discussions**
   1. **Data Cleaning**

After collecting the dataset, we uploaded it to GitHub in raw format. In this project, we accessed the dataset via the GitHub URL. The dataset comprises 10,127 rows and 23 columns, with three columns deemed irrelevant and subsequently dropped.

Columns below were removed:

1. CLIENTNUM
2. Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_1
3. Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_2'

Following this, we conducted a check for missing or null values to ensure the dataset was devoid of such issues and ready for further analysis.

Since no missing or null values were found, we proceeded to conduct Exploratory Data Analysis (EDA).

* 1. **Exploratory Data Analysis**

The goal of exploratory data analysis (EDA) is to identify and describe the key features of datasets. To fully understand the data's structure and quality, it is necessary to examine data types, missing values, and fundamental statistics. Discovering trends, correlations, and anomalies is made easier with the use of visualizations like histograms and pair plots. EDA is useful for hypothesis testing, data cleaning, and directing the analysis that follows. Obtaining insights and making sure data is ready is an important first step in any machine learning or data analysis effort.

**Univariate Analysis:**

Data analysis places a premium on carefully exploring and understanding the unique features of any given data set. Univariate analysis is a way to examine the characteristics and distribution of one independent variable. The central tendency and variability of a particular quantity can be understood by analyzing statistical markers, such as the mean, standard deviation, and quartiles.

Data Mining approach in analyzing churn of credit card customer project might use univariate analysis to look at the features in the dataset and their properties. Histograms, box plots, and summary statistics are some of the statistical tools that can be used to find missing values, outliers, and information about the data's distribution and range.

We can find out which features significantly affect the dependent variable by using the univariate analysis, which provides helpful information about the distribution of all features. In addition, it makes it easier to spot and fix issues with data quality, like missing or out-of-range numbers.

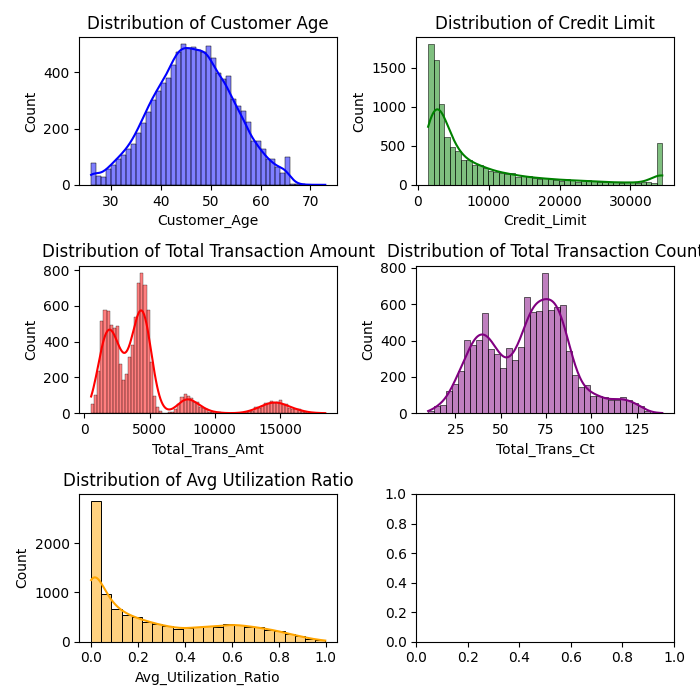
A methodical strategy for investigating a single variable can be put into action by employing the SEMMA methodology. It is feasible to improve the dataset for future research and model building by thoroughly examining all characteristics. Using this approach, one can learn all there is to know about the data and find the important features, which leads to a better model in the end.

In conclusion, understanding the unique properties of a dataset requires performing univariate analysis. This procedure makes it easier to learn about the characteristics of the data and identify problems that may need fixing. For a more effective prediction model, it is recommended to use a systematic technique like SEMMA to ensure a thorough and accurate univariate analysis.

**Histogram Analysis:**

One of the fundamentals of exploratory data analysis (EDA) is histogram analysis, which plots numerical variables according to their distribution. Knowing the data's shape, central tendency, spread, and frequency distribution is helpful.

**Figure 2**: Histogram Analysis of Numerical Variables



**1. Customer Age Distribution:** This histogram shows a normal distribution centred around 45 years of age. The smooth bell curve shows the typical population distribution,

**2. Credit Line allocation:** The credit distribution is skewed to the right, which means that most customers have low credit limits, and a few customers have very high credit limits. A high value on the far right suggests that a small number of customers qualify for higher credit amounts.

**3. Distribution of Total Transaction Volume:** A right-skewed distribution is shown, indicating that most transactions have lower amounts, with fewer transactions representing significantly higher amounts. The smaller of these transactions happen more frequently.

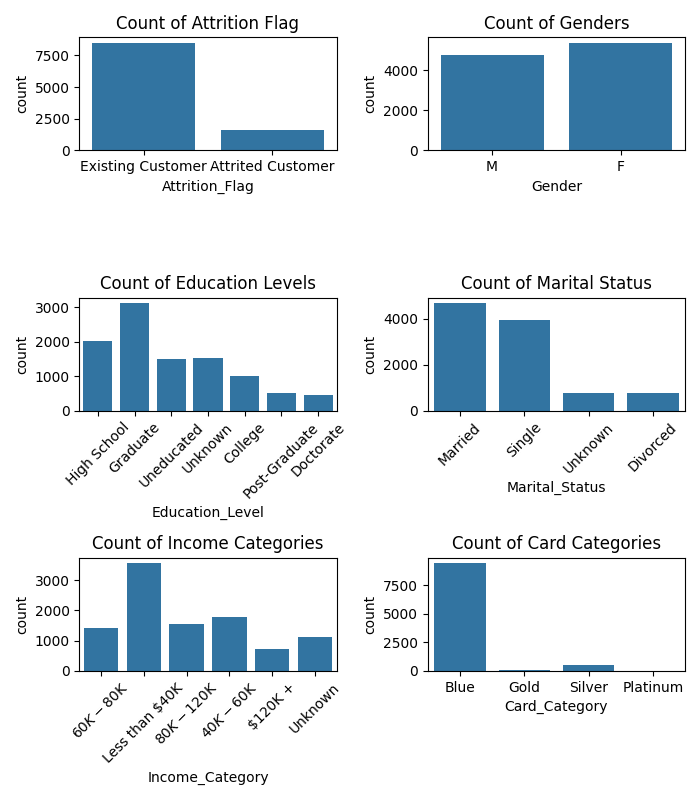
**4. Distribution of Transaction Times:** The distribution of the number of transactions is bimodal, indicating that customers have two common behaviours in the frequency of transactions.

**5. Average Utilization Distribution:** The histogram of average utilization is heavily skewed to the left, with most values clustered near zero. This suggests that most customers are only using a fraction of their available credit.

**Bar plot Analysis:**

One way to see the connection between numerical values and categorical variables is with bar plot analysis. You can use it to better grasp the categories' distribution, frequency, and comparison.

**Figure 3**: Bar Plot Analysis of Categorical Variables



1. The Count of Attrition Flag shows that the number of existing customers is significantly higher than the number of lost customers, indicating a high overall customer retention rate.

2. The distribution between male (M) and female (F) customers is balanced, suggesting that the dataset or customer base is not heavily skewed towards one gender or the other.

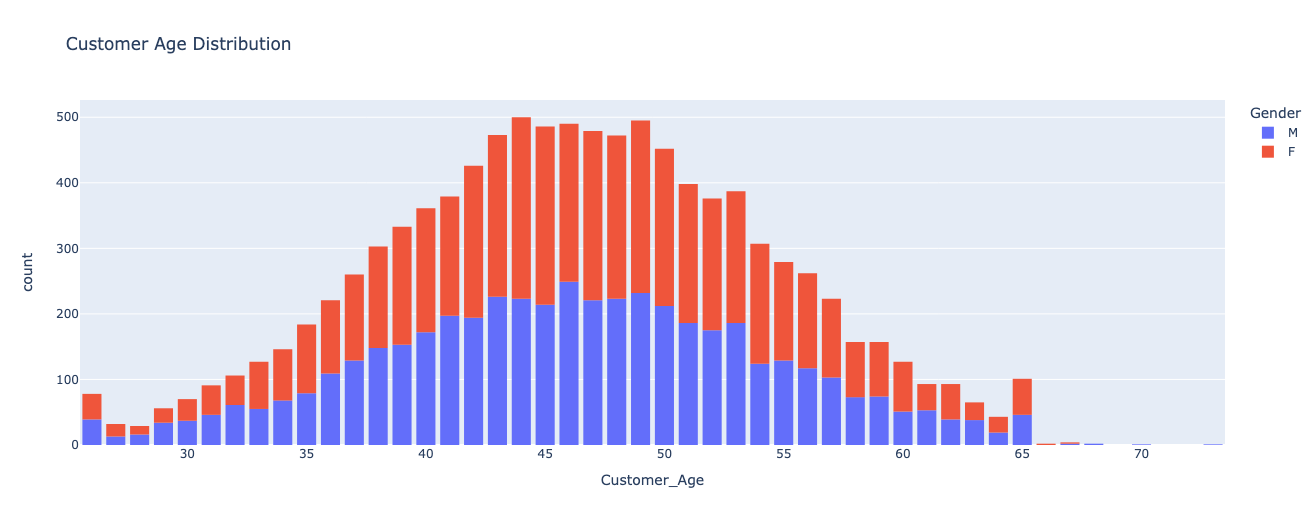
3. The graph shows different levels of education from high school to PhD. Most clients have at least a college degree, with "graduate" being the most common level.

4. Most customers are married, followed by single, then divorced, with a small percentage classified as unknown. This can affect product and marketing strategies, as marital status often influences financial needs and decisions.

5. The bar chart of revenue categories shows a wide range of customers' income, with significant numbers in brackets for both "less than $40K" and "$40K - $60K." Higher income brackets, such as "$120,000 +," are less common.

6. Most customers have a "blue" card, which may be a standard or entry card, while the "gold", "silver" and "platinum" card holders are becoming fewer and fewer. This suggests that premium cards, which may offer more benefits, are less common among customers.

**Figure 4**: Customer Age Distribution Analysis



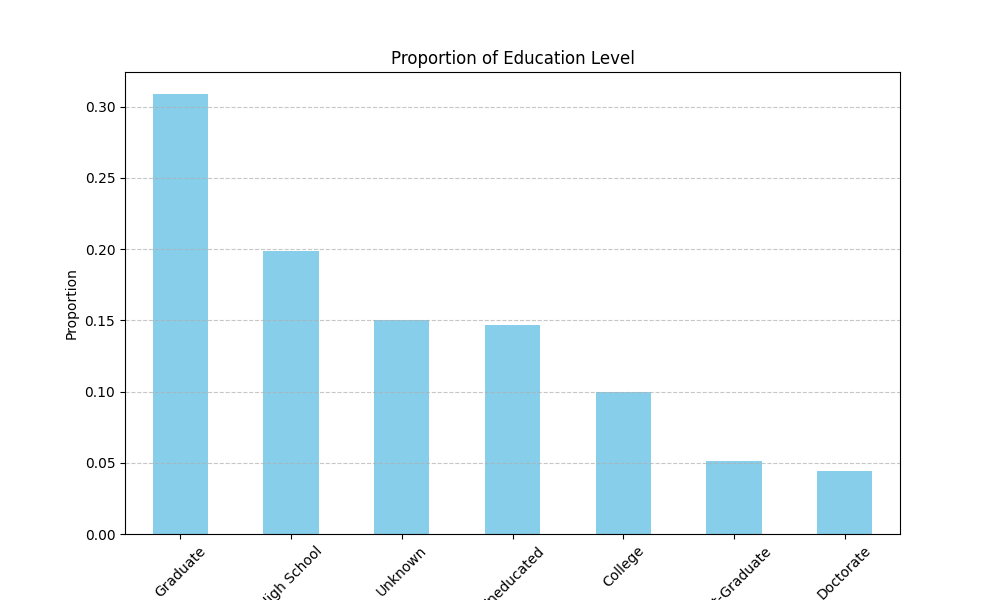
**Customer Age Distribution:**

This graph compares the age distribution between men and women at different ages. It can be seen more intuitively that the ratio of men to women is the same at different ages.

**Proportion:**

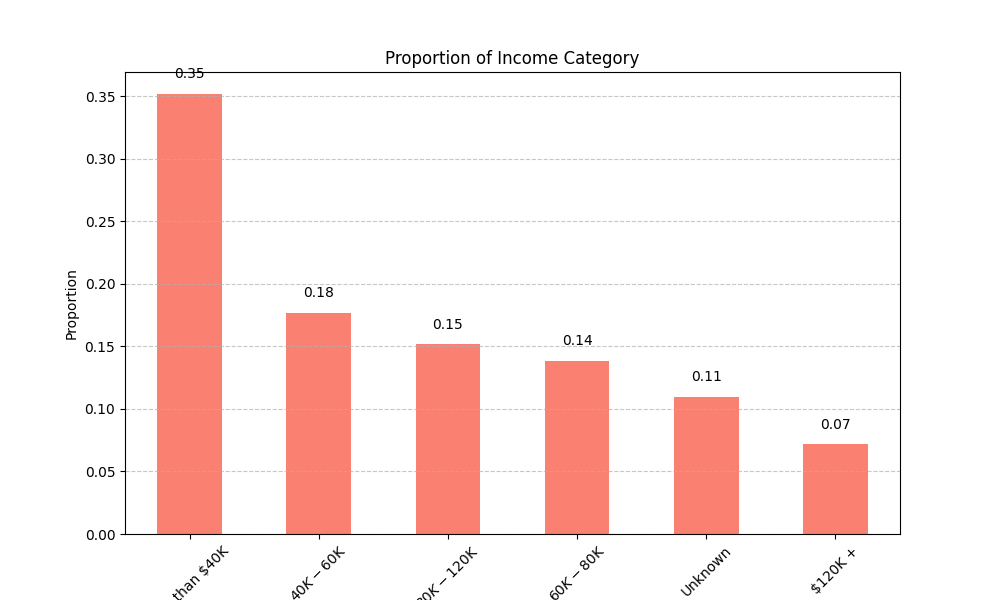
The distribution of categorical variables is typically described using proportions in univariate analysis. Percentages show how many observations fall into each group in relation to the overall count.

**Figure 5**: Proportion of Education Level



**Proportion of Education Level:**

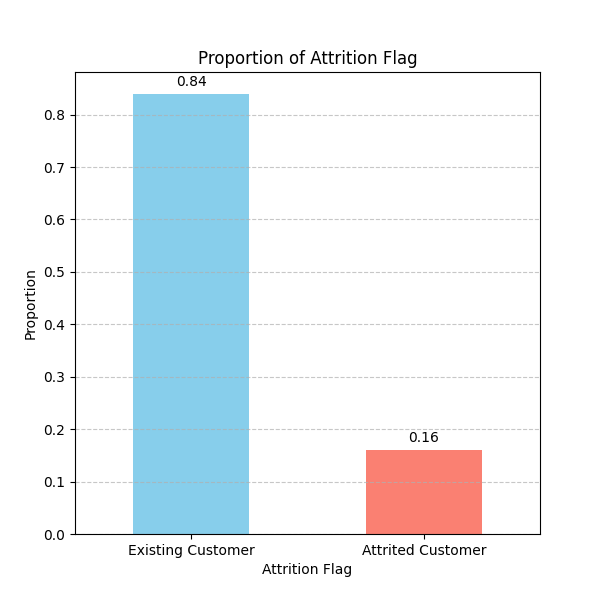
* Above 30% clients are Graduate, followed by 20% from High School, 10% from college, nearly 10% from post-graduate and Doctorate in this dataset.
* Approximately 30% Clients are listed as Unknown and Uneducated.

**Figure 6**: Proportion of Income Category

**Proportion of Income Category:**

* We can see 35% of Clients have income less than 40k(USD) per year.
* 18% Clients have income from 40k-60k(USD) per year, followed by 15% from 80k-120K(USD) per year and 14% from 60k-80k(USD) per year.
* We have 11% Clients in our dataset whose income is Unknown.
* 7% Clients have income more than 120K(USD) per year which is very low.

**Figure 7:** Proportion of Attrition Flag



**Proportion of Attrition Flag:**

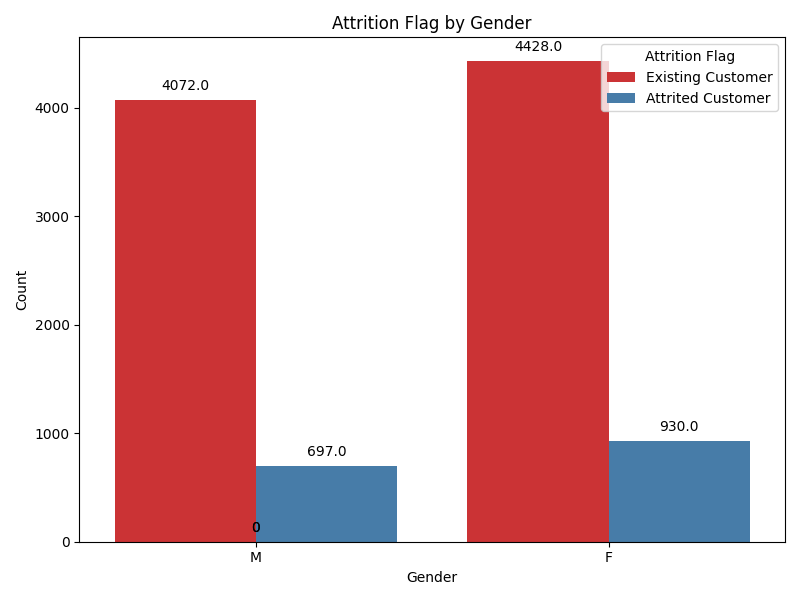
* 84% of the dataset are Existing Clients and 16% are Attired Clients.
* No missing values.

**Bivariate Analysis:**

Through plotting techniques such as box plots, scatter plots, and correlation coefficients, bivariate analysis investigates the relationship between two variables. The objective of this analysis is to gain an understanding of the strength, direction, and type of the link between the variables. This analysis helps to facilitate insights into patterns, trends, and predictive modeling.

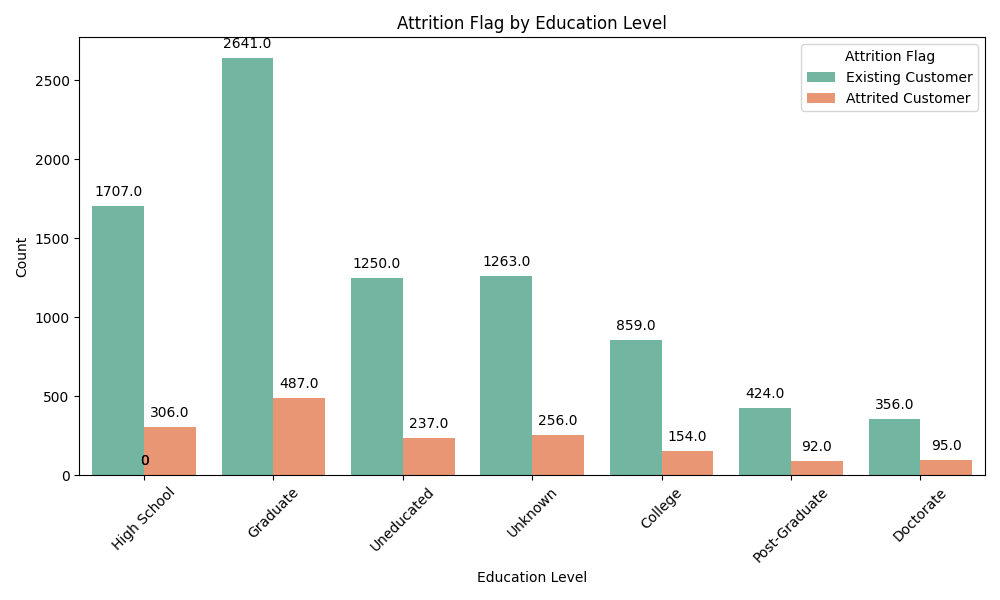
**Bar Plot Analysis:**

**Figure 8**: Distribution of Attrition Flag by Gender

**Attrition Flag by Gender:**

* We can observe from the bar plot that there are more Female Clients (4428) than Male (4072) in this dataset.
* Attired Clients are more from Female segment which is 930. There are 697 attired Male Clients in this dataset.
* So, the proportion of attired customer is almost similar in Male and Female Clients.

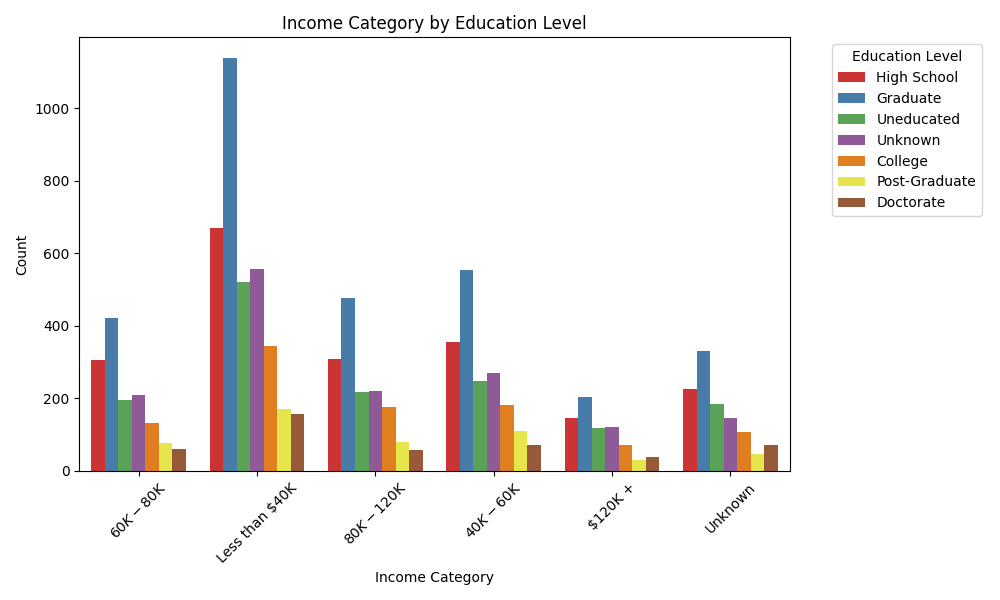
**Figure 9**: Distribution of Attrition Flag by Education Level



**Attrition Flag by Education Level:**

* There is no correlation between customer education level and customer attrition, since both the attrition rate and the retention rate of existing customers are behaving similarly.
* Most of the Existing Clients are from Graduate level and the least no of Clients are from Doctorate and Post-Graduate Level.

**Figure 10**: Distribution of Income Category by Education Level

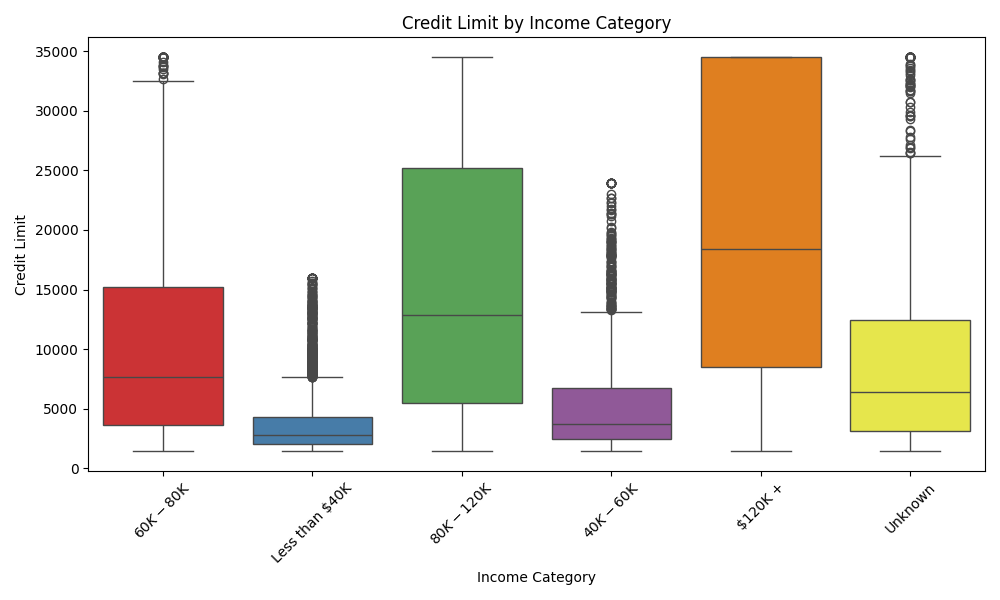


**Income Category by Education Level:**

* Regardless of one's level of education, there is a consistent trend with respect to income.
* Most of the Clients are grom Graduate level and their Income Category is less than 40k(USD)

**Boxplot Analysis:**

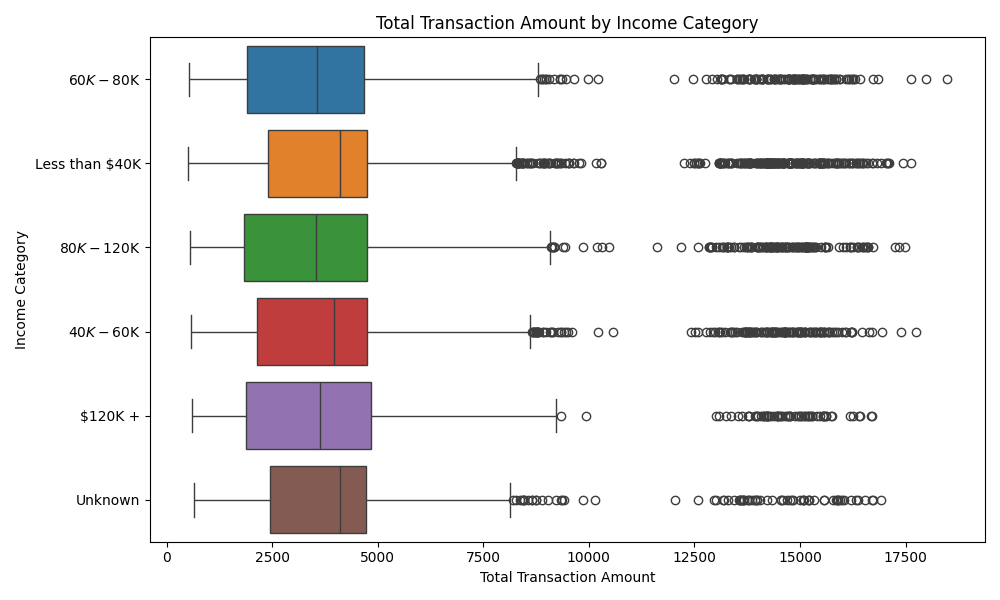
**Figure 11:** Distribution of Credit Limit by Income Category



**Credit Limit by Income Category:**

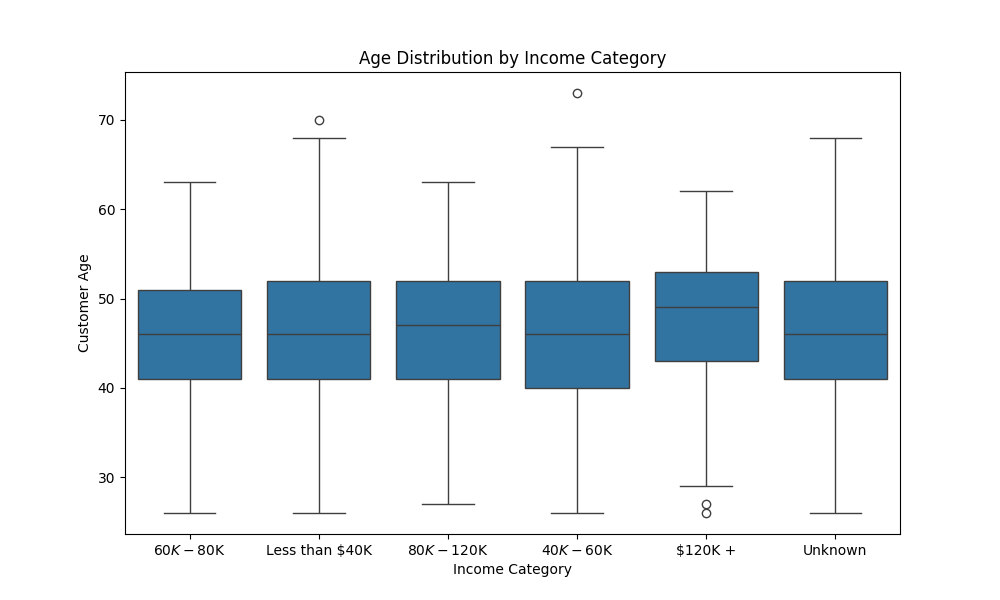
* All income categories have the same credit limit, which is to be expected given that those with lesser incomes would have a smaller credit limit.
* There are outliers in every category.

**Figure 12**: Distribution of Total Transaction Amount by Income Category

**Total Transaction Amount by Income Category:**

Transaction Amount is almost similar in all Income Categories.

**Figure 13**: Box plot Analysis of Age Distribution by Income Category

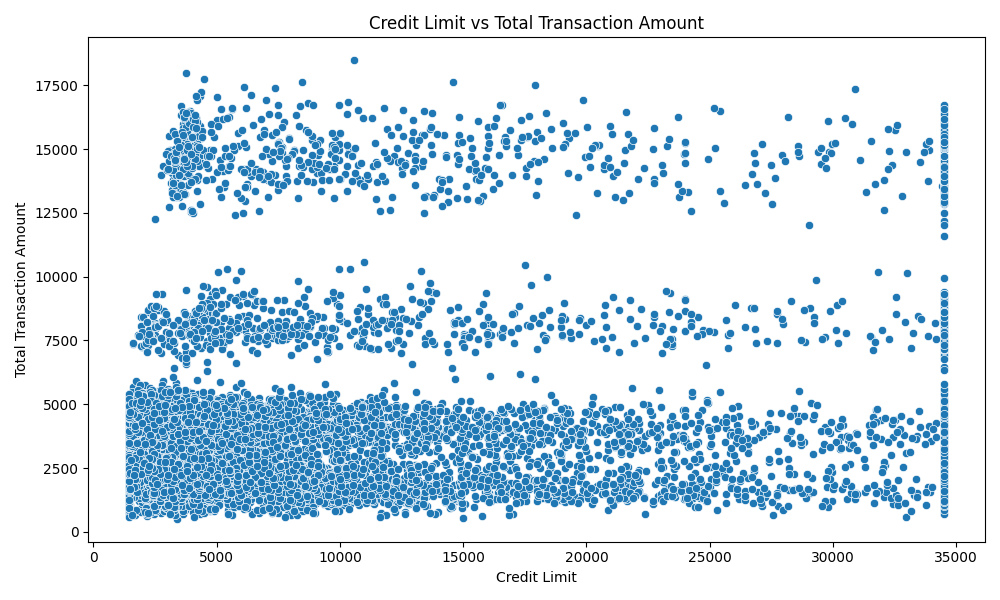


**Age Distribution by Income Category:**

* As can be seen from the box diagram, the age distribution of the income groups of $60K-$80K and $80K-$120K is concentrated, indicating that the age of customers at these income levels is consistent.
* The wide age distribution of the $40K-$60K income group indicates that this group has a large age span.
* The highest income category of $120K+ has a narrower age distribution, but there are young outliers, suggesting that high earners are concentrated in specific age groups, but there are also a few younger high earners.

**Scatter Plot Analysis:**

**Figure 14**: Scatter Plot Analysis of Credit Limit Vs Total Transaction Amount



**Credit Limit by Total Transaction Amount:**

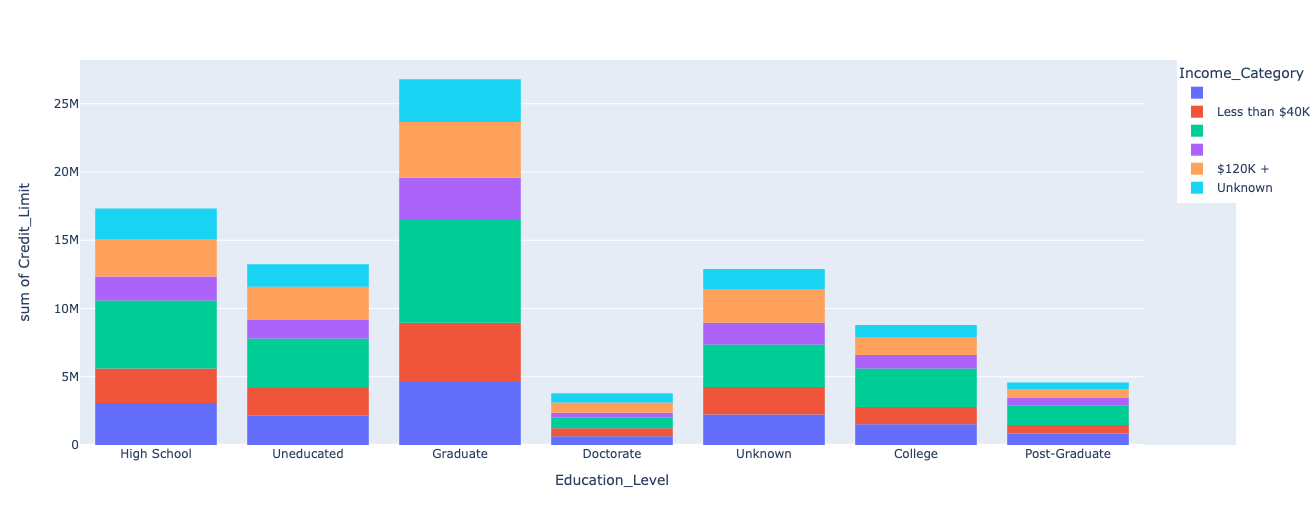
* + It can be observed that most clients have credit limits lower than they were previously.
  + When compared to greater credit limit ranges, the total transaction amount is significantly higher when the credit limit range is significantly lower.

**Multivariate Analysis:**

Statistical methods used to examine data sets with several variables are referred to as multivariate analysis. To find patterns, trends, and underlying structures in the data, it investigates the connections and interactions between various variables. It covers a wide range of techniques, including principal component analysis, discriminant analysis, cluster analysis, regression analysis, and more, allowing for a deeper comprehension of complex data and assisting in decision-making across a range of industries, from social sciences and healthcare to finance and marketing.

**Bar Plot Analysis:**

**Figure 15**: Bar Plot Analysis of Education Level by Credit Limit by Income Category



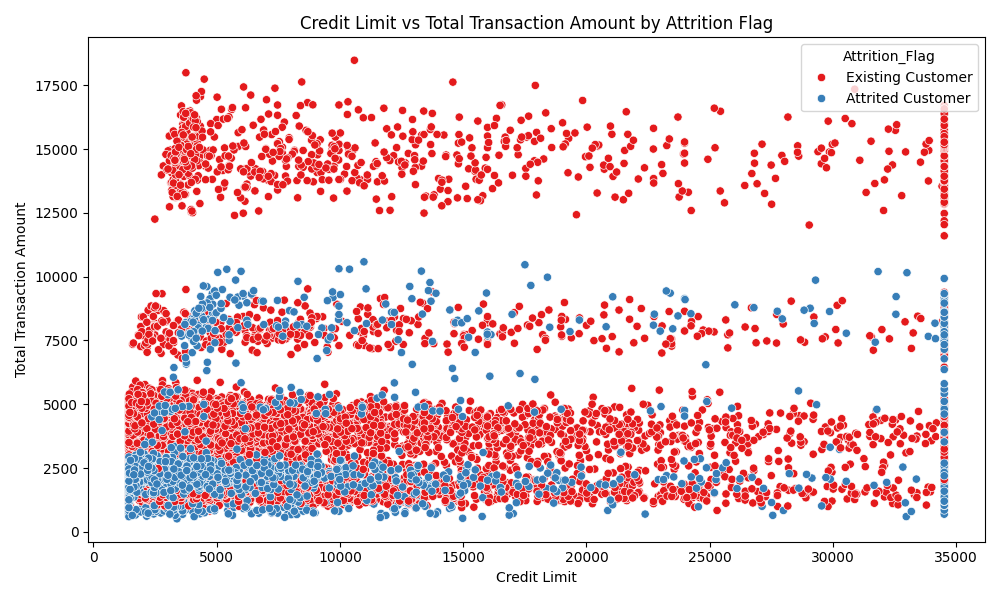
**Education Level by Credit Limit by Income Category:**

This graph shows the total credit lines of customers with different levels of education, stratified by income category.

* Uneducated customers have lower credit limits, especially with little representation in higher income categories.
* College graduate customers have higher credit limits in all income categories, especially in the $120K+ income category.
* Doctoral degree holders have the highest credit limits in the high-income category, showing the potential advantages of higher education in obtaining high credit limits.

**Scatter Plot Analysis:**

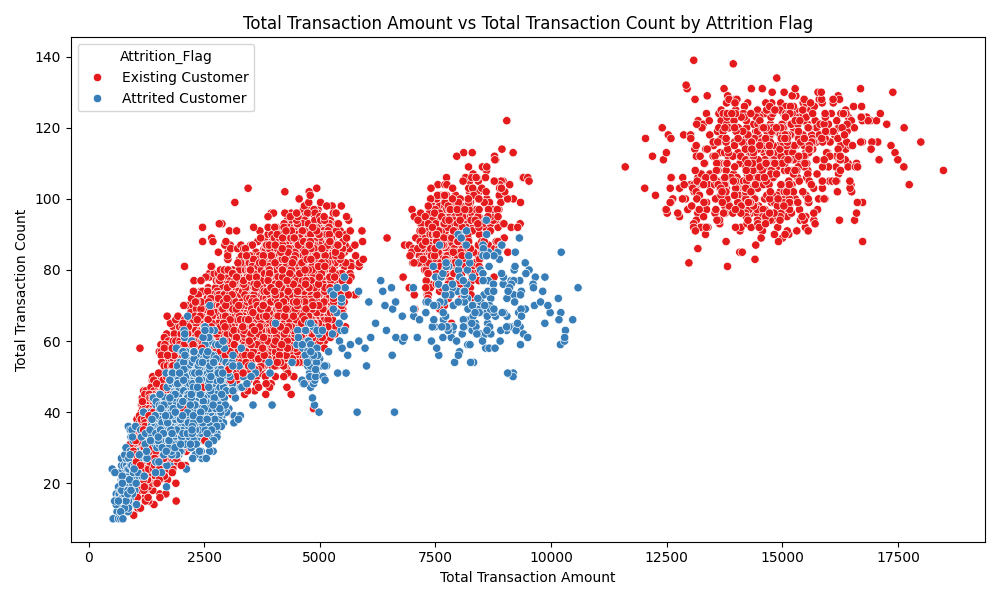
**Figure 16**: Scatter Plot Analysis of Credit Limit Vs Total Transaction Amount vs Attrition Flag



**Credit Limit Vs Total Transaction Amount by Attrition Flag:**

* While the chart displays the credit limit in comparison to the total quantity of transactions, the attrition flag is used to differentiate between the various hues.
* It has been discovered that the majority of the consumers who have left the company are from the lower range of both their credit limit and the total amount of their transactions.

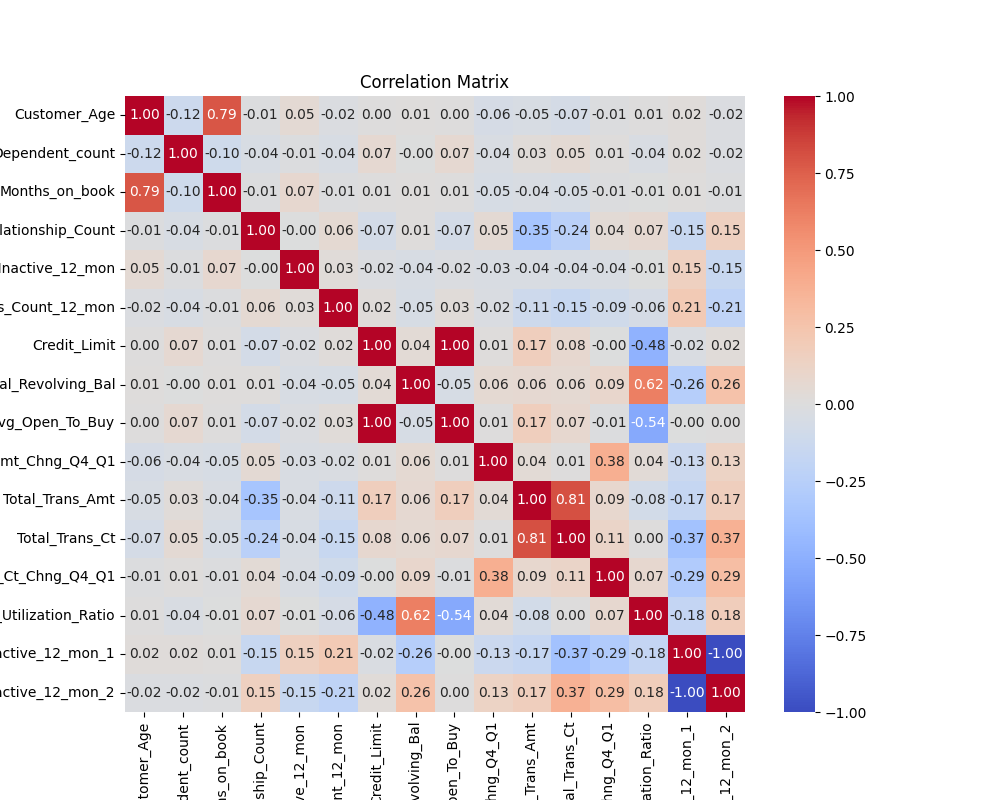
**Figure 17**: Scatter Plot Analysis of Total Transaction Amount vs Total Transaction Count by Attrition Flag



**Total Transaction Amount Vs Total Transaction Count by Attrition Flag:**

* The chart displays the overall transaction value in comparison to the total transaction count, with the attrition flag using a variety of colors to differentiate between the two levels.
* It is possible to see that there is a noticeable pattern in which the client who has left the company has a smaller number of transactions to count and a lower total amount than the customers who are still with the company.

**Figure 18**: Correlation Matrix



**Correlation Matrix:**

Correlation matrix heat maps provide a visual representation of the strength of the relationship between various variables.

* Number of months on book and customer age (0.79): This indicates a strong relationship, with older customers tending to have a longer relationship with the bank, which may indicate loyalty and a lower likelihood of churn.
* Number of contacts and inactive months (1.00): A perfect correlation shows that the more inactive months, the more contacts with the bank, which may be a response to inactivity alerts or retention efforts, a key signal in predicting customer churn.
* Total transaction amount and total number of transactions (0.81): indicates that customers who trade more frequently also tend to have higher total transaction amounts, which is a sign of active account use and may reduce the risk of churn.
* Customer age and total number of transactions (-0.24): indicates that younger customers are likely to transact more frequently, allowing for targeted marketing strategies.
  1. **Data Transformation**

After completing the exploratory data analysis (EDA), we continued with the data transformation process. This step is crucial for preparing the data for modelling and analysis. Here are the key transformations we performed:

1. Imputation
2. Encoding Categorical Variables
3. Feature Selection

**Imputation**

During the exploratory data analysis, we identified that the columns Marital\_Status, Education\_Level, and Income\_Category contain unknown values. Specifically, 15% of the values in Education\_Level, 7.4% in Marital\_Status, and 11% in Income\_Category are marked as 'Unknown'. As these columns are categorical variables, we used mode imputation to replace the 'Unknown' values with the most frequently occurring values in their respective columns.

We used SimpleImputer from scikit-learn to replace values in our dataset.

The mode in Education\_Level is ‘Graduate’. So, after mode imputation, the ‘Unknown’ value was replaced by ‘Graduate’. ‘Graduate’ count increased from 3128 to 4647 after mode imputation.

In Marital\_Status column, ‘Married’ is the highest occurring value. So, after mode imputation, the ‘Unknown’ value was replaced by ‘Married’ increasing count of ‘Married’ from 4687 to 5436.The mode in Income\_Category Column is ‘Less than $40k’ value. The ‘Unknown’ was replaced  
by the mode causing the count of ‘Less than $40k’ increase from 3561 to 4673.

**Encoding Categorical Variables**

Then, we performed encoding for categorical variables. Encoding was required because categorical variables must be transformed into numerical variables necessary for the machine learning algorithms because most of the algorithms are designed to work with numerical data. The categorical variable to encode into numerical is as below:

1. Attrition Flag (Target Variable)
2. Gender
3. Education\_Level
4. Marital\_Status
5. Income\_Category
6. Card\_Type

We used label encoding, each unique category in the variable is assigned a unique

integer number. Figure below shows the value of category column after encoding.

In Attrition\_Flag, we encoded ‘Existing Customer’ as 0 and ‘Attrited Customer’ as 1.

In Gender, we encoded ‘F’ as 0 and ‘M’ as 1.

In Education\_Level, we encoded ‘Uneducated’ as 0, ‘High School’ as 1, ‘College’ as 2, ‘Graduate’ as 3, ‘Post-Graduate’ as 4, ‘Doctorate’ as 5.

In Marital\_Status we encoded ‘Single’ as 0, ‘Married’ as 1, ‘Divorced’ as 2.

In Income\_Category, we encoded 'Less than $40K' as 0, '$40K - $60K' as 1, '$60K - $80K' as 2, '$80K - $120K' as 3, '$120K +' as 4. Figure below shows the encoded categorical variables.

**Figure 19:** Categorical Variable after encoding

A group of blue bars

Description automatically generated

**Feature Selection**

Feature selection is a crucial step in the data preprocessing. It involves identifying the most relevant features that contribute to the predictive performance of a model. By reducing the number of features, we can simplify the model, reduce overfitting, and improve computational efficiency. We extracted the important features for all the 4 models by using steps below:

**Logistic Regression:**

We extracted the coefficients of the logistic regression model as the coefficients can indicate the importance of each feature. Features with larger absolute coefficient values are considered more important.

**Random Forest:**

Feature importance in a random forest is determined based on how much each feature reduces the impurity (e.g., Gini impurity) across all the trees in the forest. We used feature\_importances\_ function in RandomForestClassifier to retrieve the feature and their important scores.

**Gradient Boosting:**

Similar to Random Forest, feature importance in Gradient Boosting is derived from how much each feature contributes to reducing the loss function. We also used feature\_importances\_ function in GradientBoostingClassifier to extract the features and their scores.

**Multi-layer perceptron (MLP):**

MLPs don't provide direct feature importances like tree-based models. So, we used permutation feature importance. Permutation feature importance involves shuffling the values of each feature and measuring how the performance of the model decreases.

Then we created plot to visualize feature importance in descending order for all 4 models using as shown in below figure. The vertical red dashed line at the 0.1 threshold helps to quickly identify which features are considered important by each model.

**Figure 20:** Feature importance of all 4 models

A screenshot of a graph

Description automatically generated

Based on the plot, it is evident that the features with importance scores higher than 0.1 differ across the various models. Therefore, we have decided to employ a threshold aggregation method for feature selection. This method involves selecting features that exceed 0.1 score in all the models.

Below features were selected:

1. Avg\_Open\_To\_Buy
2. Contacts\_Count\_12\_mon
3. Credit\_Limit
4. Dependent\_count
5. Months\_Inactive\_12\_mon
6. Total\_Ct\_Chng\_Q4\_Q1
7. Total\_Revolving\_Bal
8. Total\_Trans\_Amt
9. Total\_Trans\_Ct
   1. **Model Building**

To ensure robust model performance, we undertook a systematic approach to model building, which included data splitting, hyperparameter tuning, and selection of the best parameters for each model. We began by splitting the dataset into training and testing sets. This step is crucial to evaluate the model's performance on unseen data.

The dataset was split according to below ratio:

Training Set: 80% of the data (8101 records)

Testing Set: 20% of the data (2026 records)

Next, we continued with hyperparameter tuning. Hyperparameter tuning involves searching for the best combination of parameters that optimize the model's performance. We used different search strategies for different models:

1. Logistic Regression: Grid Search CV
2. Random Forest, Multi-Layer Perceptron (MLP) , and Gradient Boosting: Randomized Search CV

Logistic regression typically has fewer hyperparameters to tune compared to ensemble methods like Random Forest or Gradient Boosting. Since the hyperparameter space is relatively small, and it's feasible to try out all possible combinations, GridSearchCV was used. Logistic regression hyperparameters (like regularization strength) often require an exhaustive search over a specified range of values to find the optimal combination. On the other hand, Random Forest, Gradient Boosting and MLP uses RandomizedSearchCV as these models has more hyperparameters to tune. Performing an exhaustive grid search over a large hyperparameter space can be computationally expensive and time-consuming.

Besides, RandomizedSearchCV samples a fixed number of parameter settings from the specified distributions, making it more efficient and practical when dealing with many hyperparameters and a large search space. We have referred and selected this method from [16] as the study used RandomizedSearchCV for automatic parameter adjustment for their model.

Once we have completed hyperparameter tuning for all four models (logistic regression, random forest, gradient boosting, and MLP), we trained each model using the best hyperparameters found during tuning and then evaluate their performance on the testing dataset.

**4.4.1 Model Evaluation**

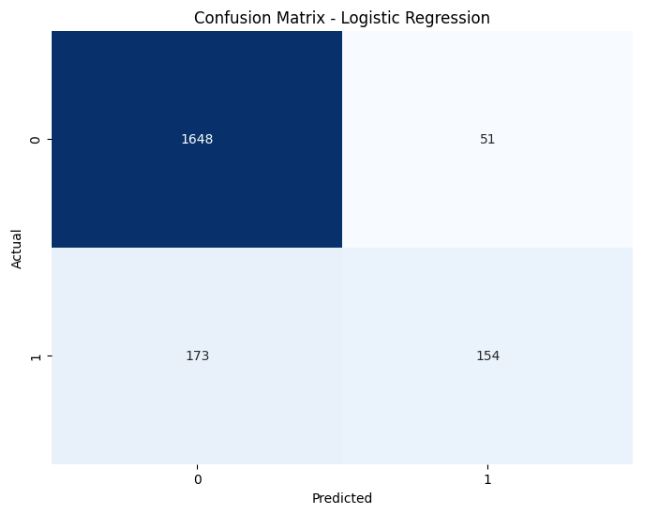
Below figures show the logistic regression model performance’s in predicting credit card customer churn.

Figure 21: Classification Report of Logistic Regression

A screenshot of a computer

Description automatically generated

**Figure 22:** Confusion Matrix of Logistic Regression



Below is the interpretation for Confusion Matrix of Logistic Regression:

True Negatives (TN): 1648

False Positives (FP): 51

False Negatives (FN): 173

True Positives (TP): 154

We have summarized the performance of logistic regression as below:

**Accuracy:**

The model has an accuracy of 0.89, meaning it correctly classifies 89% of the instances.

**Precision:**

For the negative class (0), precision is 0.90, indicating that when the model predicts the negative class, it is correct 90% of the time.

For the positive class (1), precision is 0.75, indicating that when the model predicts the positive class, it is correct 75% of the time.

**Recall:**

For the negative class (0), recall is 0.97, meaning the model correctly identifies 97% of the true negative cases.

For the positive class (1), recall is 0.47, meaning the model correctly identifies 47% of the true positive cases.

**F1-Score:**

For the negative class (0), the F1-score is 0.94, indicating a good balance between precision and recall.

For the positive class (1), the F1-score is 0.58, indicating a moderate balance between precision and recall.

The model performs very well in predicting the negative class with high precision, recall, and F1-score. However, its performance in predicting the positive class is weaker, with a lower recall and F1-score. The overall accuracy and weighted average metrics suggest a robust performance, but the lower recall for the positive class indicates that the model misses a significant number of true positive cases.

Below figures show the random forest model performance’s in predicting credit card customer

churn.

**Figure 23:** Classification Report of Random Forest

**A screenshot of a computer screen

Description automatically generated**

**Figure 24:** Confusion Matrix of Random Forest

**A blue and white graph

Description automatically generated**

Below is interpretation for Confusion Matrix of Random Forest:

True Negatives (TN): 1671

False Positives (FP): 28

False Negatives (FN): 83

True Positives (TP): 244

We have summarized the performance of random forest as below:

**Accuracy:**

The model has an accuracy of 0.95, indicating that it correctly classifies 95% of the instances.

**Precision:**

For the negative class (0), precision is 0.95, meaning the model is correct 95% of the time when it predicts the negative class. For the positive class (1), precision is 0.90, indicating the model is correct 90% of the time when it predicts the positive class.

**Recall:**

For the negative class (0), recall is 0.98, meaning the model correctly identifies 98% of the true negative cases. For the positive class (1), recall is 0.75, meaning the model correctly identifies 75% of the true positive cases.

**F1-Score:**

For the negative class (0), the F1-score is 0.97, indicating a high balance between precision and recall.

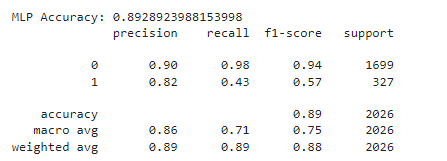
For the positive class (1), the F1-score is 0.81, indicating a good balance between precision and recall.

The random forest model outperforms the logistic regression model across most metrics. It shows a significant improvement in recall for the positive class, which addresses the weakness observed in the logistic regression model. The random forest model achieves higher overall accuracy and provides a better balance between precision and recall, making it a more robust choice for this classification task.

Below figures show the MLP performance’s in predicting credit card customer

churn.

**Figure 25:** Classification Report of MLP



**Figure 26:** Confusion Matrix of MLP

**A blue squares with white text

Description automatically generated**

Below is interpretation for Confusion Matrix of MLP:

True Negatives (TN): 1668

False Positives (FP): 31

False Negatives (FN): 186

True Positives (TP): 141

We have summarized the performance of MLP as below:

**Accuracy:**

The model has an accuracy of 0.89, indicating that it correctly classifies 89% of the instances.

**Precision:**

For the negative class (0), precision is 0.90, meaning the model is correct 90% of the time when it predicts the negative class. For the positive class (1), precision is 0.82, indicating the model is correct 82% of the time when it predicts the positive class.

**Recall:**

For the negative class (0), recall is 0.98, meaning the model correctly identifies 98% of the true negative cases. For the positive class (1), recall is 0.43, meaning the model correctly identifies 43% of the true positive cases.

**F1-Score:** For the negative class (0), the F1-score is 0.94, indicating the model is very good at identifying negative instances correctly, balancing both high precision and high recall for this class.

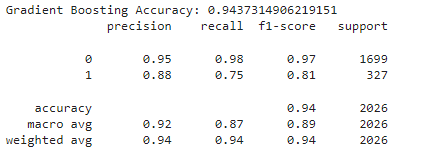
For the positive class (1), the F1-score is 0.58, indicating the model is not as effective at identifying positive instances.

The MLP model demonstrates strong performance in identifying negative instances, with high precision, recall, and F1-score for the negative class. However, the model's performance in identifying positive instances is considerably weaker, as indicated by lower recall and F1-score for the positive class.

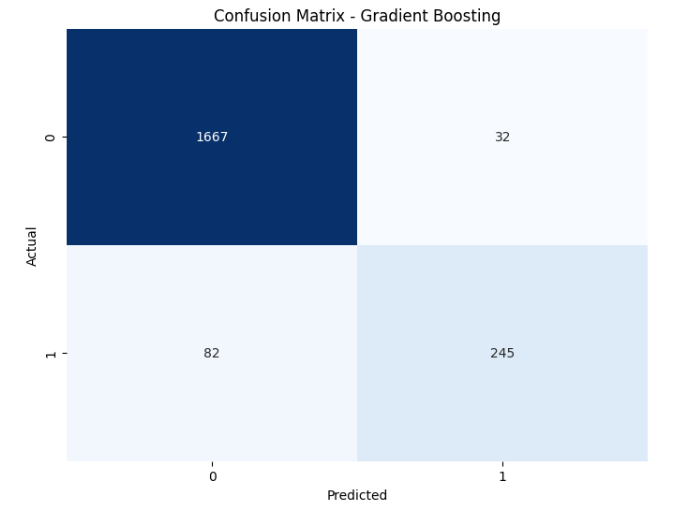
Below figures show the Gradient Boosting model’s performance in predicting credit card customer

churn.

**Figure 27:** Classification Report of Gradient Boosting



**Figure 28:** Confusion Matrix of Gradient Boosting



Below is interpretation for Confusion Matrix of Gradient Boosting:

True Negatives (TN): 1667

False Positives (FP): 32

False Negatives (FN): 245

True Positives (TP): 82

We have summarized the performance of Gradient Boosting as below:

**Accuracy:**

The model has an accuracy of 0.94, indicating that it correctly classifies 94% of the instances.

**Precision:** For the negative class (0), precision is 0.95, meaning the model is correct 95% of the time when it predicts the negative class. For the positive class (1), precision is 0.88, indicating the model is correct 88% of the time when it predicts the positive class.

**Recall:**

For the negative class (0), recall is 0.98, meaning the model correctly identifies 98% of the true negative cases. For the positive class (1), recall is 0.75, meaning the model correctly identifies 75% of the true positive cases.

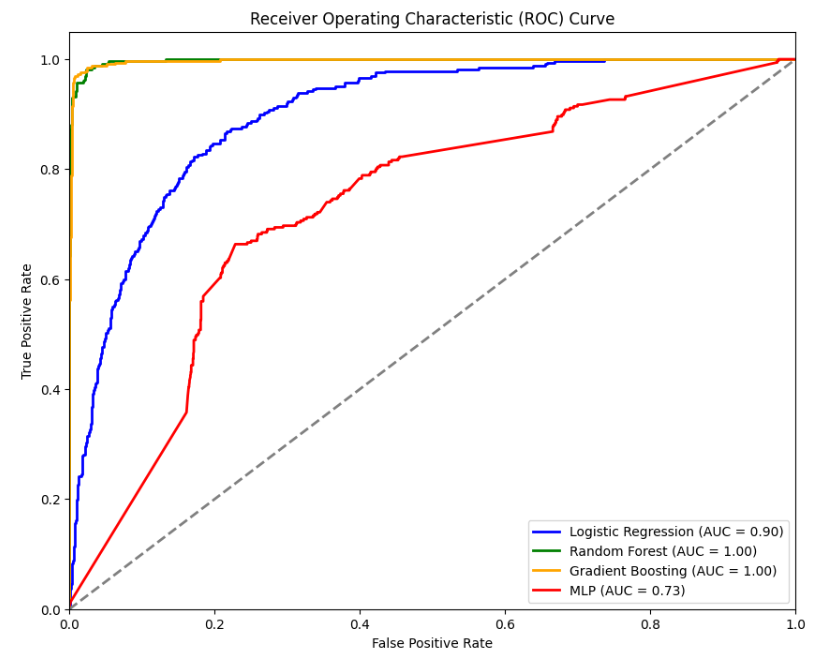
**F1-Score:**

For the negative class (0), the F1-score is 0.97, indicating a high balance between precision and recall. For the positive class (1), the F1-score is 0.81, indicating a good balance between precision and recall

Gradient Boosting demonstrates strong performance across various evaluation metrics. It achieves high accuracy, precision, recall, and F1-scores for both classes, indicating its effectiveness in classification tasks. Specifically, the model excels in correctly identifying negative instances, with very high precision and recall. While its performance on positive instances is slightly lower, it still achieves a good balance between precision and recall, as indicated by the F1-score.

We also plotted ROC curves for all the 4 models. The ROC curve is a plot of the true positive rate (TPR) against the false positive rate (FPR) for different threshold values. The closer the ROC curve is to the top-left corner, the better the model's performance. This indicates a high true positive rate and a low false positive rate.

**Figure 29:** Comparing ROC Curve between 4 models



The ROC curve graph shows the performance of four models: Logistic Regression, Random Forest, Gradient Boosting, and MLP (Neural Network). The x-axis represents the False Positive Rate (FPR), and the y-axis represents the True Positive Rate (TPR). The diagonal line indicates the performance of a random classifier.

Based on the ROC Curve above, Random Forest and Gradient Boosting have AUC of 1.00 respectively. Logistic Regression has AUC of 0.90 whereas MLP has the lowest AUC,0.73.

Random Forest and Gradient Boosting achieved the highest AUC scores, indicating excellent discriminatory power. Logistic Regression follows closely with a slightly lower AUC but still performs well. MLP, while having the lowest AUC.

**4.4.2 Performance comparison between balanced dataset and imbalanced dataset**

According to the model performance, we noticed that the models have higher scores for Precision, F1-Score and Recall in identifying negative classes. This could be caused by the imbalanced data in the dataset. So, we decided to use Synthetic Minority Over-sampling Technique (SMOTE) to balance the dataset, train and test the models again.

**Table 3: Comparison of Accuracy and Recall before and after applying SMOTE**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy before SMOTE** | **Accuracy after SMOTE** | **Recall before**  **SMOTE** | **Recall after SMOTE** |
| **Logistic Regression** | 88.90% | 82.4% | 89.00% | 82.00% |
| **Random Forest** | 95.00% | 93.00% | 93.00% | 93.00% |
| **MLP** | 89.29% | 74.43% | 89.00% | 74.00% |
| **Gradient Boosting** | 94.37% | 93.68% | 94.00% | 94.00% |

Based on the table above, Random Forest provided the highest accuracy score (95%) and Gradient Boosting provided the best recall score (94%) on unbalanced dataset. After applying SMOTE, the accuracy of Random Forest reduced by 1% only and the recall score of Gradient Boosting remained the same (94%). There is no significant difference in the accuracy scores between Random Forest and Gradient Boosting. However, in [4], study conducted by Ahmet and Kozłowska suggested to choose model with a better recall score if there is no significant difference in accuracy scores. According [4], compared to accuracy, recall is considered the most important model evaluation metric in unbalanced dataset. Hence, we select Gradient Boosting as the best model for predicting credit card customer churn.

**5.0 Conclusion**

In conclusion, credit card usage has become essential in the modern financial landscape, but companies face significant challenges with customer churn, where clients discontinue their credit card services. In order to create effective retention tactics, businesses must study and predict customer behavior, as this churn has a negative impact on profit margins, trust, and brand loyalty. Our research uses transaction, customer, and demographic data to create predictive models with machine learning approaches that uncover churn-causing elements. Through a structured approach (SEMMA) and overcoming challenges such as imbalanced data and feature selection, we evaluated several models including Logistic Regression, Random Forest, Gradient Boosting, and MLP. Among these, Gradient Boosting emerged as the most effective, achieving high accuracy and recall, making it the best model for predicting customer churn. The findings advocate for the use of data-driven strategies to foster long-term customer loyalty, highlighting that while the immediate financial impacts of churn are significant, the broader implications on consumer trust and market resilience are equally critical. Consequently, continuous research and adaptation are imperative for credit card companies to maintain a competitive edge and achieve sustainable growth in a dynamic market environment. This study underscores the importance of continuous research to adapt to evolving market trends and consumer preferences, ensuring credit card companies can retain valuable customers and maintain a competitive edge.

**References:**

[1] Muneer, Amgad, Rao Faizan Ali, Amal Alghamdi, Shakirah Mohd Taib, Ahmed Almaghthawi, and EA Abdullah Ghaleb. "Predicting customers churning in banking industry: A machine learning approach." Indonesian Journal of Electrical Engineering and Computer Science 26, no. 1 (2022): 539-549.

[2] Al-Najjar, Dana, Nadia Al-Rousan, and Hazem Al-Najjar. "Machine learning to develop credit card customer churn prediction." Journal of Theoretical and applied electronic commerce research 17, no. 4 (2022): 1529-1542.

[3] Domingos, Edvaldo, Blessing Ojeme, and Olawande Daramola. "Experimental analysis of hyperparameters for deep learning-based churn prediction in the banking sector." Computation 9, no. 3 (2021): 34.

[4] Çalış, Ahmet, and Justyna Kozłowska. "Customer churn prediction with popular machine learning algorithms." (2021).

[5] de Lima Lemos, Renato Alexandre, Thiago Christiano Silva, and Benjamin Miranda Tabak. "Propension to customer churn in a financial institution: A machine learning approach." Neural Computing and Applications 34, no. 14 (2022): 11751-11768.

[6] Guliyev, Hasraddin, and Ferda Yerdelen Tatoğlu. "Customer churn analysis in banking sector: Evidence from explainable machine learning model." Journal of Applied Microeconometrics 1, no. 2 (2021).

[7] Xiahou, Xiancheng, and Yoshio Harada. "Customer churn prediction using AdaBoost classifier and BP neural network techniques in the E-commerce industry." American Journal of Industrial and Business Management 12, no. 3 (2022): 277-293.

[8] Azzopardi, A. S., & Azzopardi, J. (2022). Predicting Customer Behavioural Patterns using Virtual Credit Card Transactions Dataset. ICSBT International Conference on Smart Business Technologies, 2022-July(Icsbt), 160–167. https://doi.org/10.5220/0011342300003280

[9] Calzada-Infante, Laura, María Óskarsdóttir, and Bart Baesens. "Evaluation of customer behavior with temporal centrality metrics for churn prediction of prepaid contracts." Expert Systems with Applications 160 (2020): 113553.

[10] Wu, Shuli, Wei-Chuen Yau, Thian-Song Ong, and Siew-Chin Chong. "Integrated churn prediction and customer segmentation framework for telco business." Ieee Access 9 (2021): 62118-62136.

[11] Xiahou, Xiancheng, and Yoshio Harada. "B2C E-commerce customer churn prediction based on *K-*means and SVM." Journal of Theoretical and Applied Electronic Commerce Research 17, no. 2 (2022): 458-475.

[12] Bogaert, Matthias, and Lex Delaere. "Ensemble methods in customer churn prediction: A comparative analysis of the state-of-the-art." Mathematics 11, no. 5 (2023): 1137.

[13] Li, Jianfeng, Xue Bai, Qian Xu, and Dexiang Yang. "Identification of Customer Churn Considering Difficult Case Mining." *Systems* 11, no. 7 (2023): 325.

[14] ÜNLÜ, Kamil Demirberk. "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* 70, no. 2 (2021): 827-836.

[15] Rahman, M., & Kumar, V. (2020, November). Machine learning based customer churn prediction in banking. In 2020 4th international conference on electronics, communication and aerospace technology (ICECA) (pp. 1196-1201). IEEE.

[16] Wu, Chongqi, and Lan Wang. "A Comparative Analysis of Churn Prediction Models: A Case Study in Bank Credit Card." *Journal of Supply Chain and Operations Management* 20, no. 2 (2022): 120.

[17] Fujo, S. Wael, Suresh Subramanian, and M. Ahmad Khder. "Customer churn prediction in telecommunication industry using deep learning." *Information Sciences Letters* 11, no. 1 (2022): 24.

[18] Dorokhov, O. V., L. P. Dorokhova, L. M. Malyarets, and Iryna Ushakova. "Customer churn predictive modeling by classification methods." (2020).

[19] IEEE Conference Publication | IEEE Xplore. “E-Commerce Customer Churn Prediction By Gradient Boosted Trees,” October 29, 2020. https://ieeexplore.ieee.org/document/9303661.

[20] Kim, Sulim, and Heeseok Lee. "Customer churn prediction in influencer commerce: An application of decision trees." *Procedia Computer Science* 199 (2022): 1332-1339.

[21] Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., and Anwar, S. 2019a. "Customer Churn Prediction in Telecommunication Industry Using Data Certainty," Journal of Business Research (94), pp. 290-301.

[22] Amin, A., Shah, B., Khattak, A. M., Lopes Moreira, F. J., Ali, G., Rocha, A., and Anwar, S. 2019b. "CrossCompany Customer Churn Prediction in Telecommunication: A Comparison of Data Transformation Methods," International Journal of Information Management (46), pp. 304-319.

[23] Coussement, K., Lessmann, S., and Verstraeten, G. 2017. "A Comparative Analysis of Data Preparation Algorithms for Customer Churn Prediction: A Case Study in the Telecommunication Industry," Decision Support Systems (95), pp. 27-36.