Statistical Analysis of Customer Credit Card Attrition

Victoria Agboola, Ilayda Bekircan

ABSTRACT

This project studies customer attrition in a bank's credit card portfolio, employing four statistical methods for predictive modeling: Binary Logit, CART (Classification and Regression Trees) algorithm, Random Forest, and XGBoost. This study aims to identify the most effective model in predicting customer churn using statistical key metrics such as accuracy, sensitivity, specificity, F1 score and balanced accuracy. Upon comprehensive analysis, the CART algorithm emerged as the optimal choice. CART demonstrated competitive accuracy, excelled in sensitivity, and achieved a high specificity, showcasing a balanced performance in capturing both positive and negative instances. Its strong precision and balanced accuracy further supported its selection as the preferred model for predicting credit card customer attrition. The findings from this study not only contribute to the understanding of customer churn in the banking sector but also provide actionable insights for the implementation of targeted retention strategies. The success of CART underscores its utility in enhancing predictive analytics for customer behavior, thereby aiding financial institutions in proactive decision-making to mitigate customer attrition and foster long-term customer relationships.

INTRODUCTION

In the fast-paced environment of finance and banking industries, customer attrition threatens business managers and companies seeking to maintain a healthy business relationship with their customers. As financial institutions aims to follow market trends and consumer behaviors, to understand and predict credit card attrition becomes increasingly important. Recently, population of newly credit card customers has shrunk in an important scale. Therefore, banks are forced to have different and brilliant strategies to attract new customers from other financial institutions to improve their portfolio. To attract more new customers, banks strategize lower interest rates in credit cards for a time period at the beginning. However, when this advantageous period ends, customers tend to transfer their balance to an another bank before the interest rate pivots to the default value. Consequently, rather than maintaining existing

customers, attracting new customers becomes more and more costly for the institutions. Banks set a course for managing their existing customers not to leave their credit card products since it cost less to invest on them. Accordingly, portfolio of existing customers are prone to increase their spendings on their credit cards more [1].

The significance of anticipating which clients, particularly those with a high return on investment, are prone to leaving has grown in importance for banks. Focusing on this foresight, financial institutions can implement targeted marketing initiatives that have demonstrated efficiency in customer retention. Given the elevated importance and widespread interest among financial entities, this project introduces machine learning methodologies for classifying credit card attrition [2].

An increasing rate of customer attrition causes unsettling problems for the company with a consumer credit card portfolio. Managers and data analytics teams must answer important questions while clients stop using the bank's credit card services: What causes this attrition and what motivates it? Are there any trends that causes customers stop using their credit cards? How can this information be properly used to forecast and stop attrition? It is essential to minimize financial losses to build client loyalty and improve business sustainability by comprehending the analysis of credit card attrition. The bank can implement strategies to retain valuable customers and improve customer engagement procedures by resolving root causes of customer churn. In order to provide business managers with useful insights, this project uses extensive statistical methods to comb through the complexity of customer attrition in the banking industry and identify the underlying causes.

The primary goal of this project is to conduct a comprehensive statistical analysis of credit card attrition of customers in a bank's portfolio. The project aims to reveal trends and root causes that originates customer attrition by developing predictive models including binary logit, CART (Classification and Regression Trees), Random Forest, and XGBoost to forecast customer churn with enhanced accuracy. By leveraging machine learning algorithms, we seek to understand the reasons behind driving customer attrition in banks. This predictive capability will equip the business managers with valuable insights, enabling them to implement targeted retention strategies and mitigate potential financial losses.

DATA DESCRIPTION

Attrition_Flag: This variable is the binary response variable which refers if a customer's account is closed or not. If a credit card is not in use anymore, this variable is flagged as "Attrited Customer" and if it is still in use, it is flagged as "Existing Customer". This variable will be converted to 0 for existing customers and 1 for attrited customers for easy interpretation and simple usage in machine learning models. The data set consists of 84% of existing customers and 16% of attrited customers.

The following parameters will be used as independent variables (predictors) in the machine learning models you will see in the following parts of the project.

Customer_Age: This variable represents age of account holders in the bank's portfolio. The unit of this variable is years. The minimum and maximum age in this data set are 26 and 73, respectively. The frequency of occurence is at high point between ages 40 and 50.

Gender: This is the sex of the account holders which gives us information about the demographic structure of the population in the data set. 53% of the observations are constructed by females while 47% of them are by males. Since the percentages are close to each other, we can assume gender distribution is balanced.

Dependent_count: This variable shows how many other individuals are considered as dependent for the corresponding customer. If a customer has no dependent individual in their financial account, this parameter is labeled as 0. The maximum number of dependent individuals in this data set is 5.

Education_Level: This parameter represents seven different education level status for the customers in the portfolio. The distinct values are Uneducated, High School, College, Graduate, Post-Graduate, Doctorate and Unknown.

Marital_Status: This predictor involves four unique marital status: Divorced, Marries, Single and Unknown.

Income_Category: This is an independet variable that gives information about the account holders' annual incomes in categories: Less than $40K, $40K - $60K, $60K - $80K, $80K - $120K, $120K+ and Unknown.

Card_Category: The bank has different credit card types that each one has different benefits. There are four unique products Blue, Gold, Platinum, Silver and 93.18% of the observations have Blue card in the data set.

Months_on_book: This one represents the period of relationship of the client to their credit cards in months. The minimum months on book is 13 while the maximum amount in the data set is 56.

Total_Relationship_Count: Total relationship count shows the total number of products that the corresponding customer holds in the bank. Since all customers have at least one product in the data set, the minimum amount is always 1 which can increase up to 6.

Months_Inactive_12_mon: If a customer has any inactivity for a whole month in the last 12 months, this parameter gives the total number of inactive month number. If there is no inactivity for a whole year, it gives 0.

Contacts_Count_12_mon: This predictor is the contact number between the customer and the bank in the last 12 months. This number can differ from 0 to 6.

Credit_Limit: This gives the credit card limit of the client in dollars ($). The range is between $1438 and $34,516.

Total_Revolving_Bal: Reveolving balance is the balance carries over from one month to the next and this predictor gives the information in dollars ($). If a customer pays their checks in time, the total revolving balance is 0.

Avg_Open_To_Buy: Open to buy means the amount left in the customer's credit card that they can use montly. This parameter (Average open to buy) gives the average of open to buy values of last 12 months in dollars ($).

Total_Trans_Amt: Total transaction amount is the sum of all the transactions happening in the customer's account in the last 12 months in dollars ($).

Total_Trans_Ct: Total transaction count is the count of all the transactions happening in the customer's account in the last 12 months.

Total_Ct_Chng_Q4_Q1: This variable gives information about the ratio of the total transaction count in 4th quarter to the total transaction count in 1st quarter.

Total_Amt_Chng_Q4_Q1: This is the ratio of the total transaction amount in 4th quarter to the total transaction amount in 1st quarter.

Avg_Utilization_Ratio: Average utilization ratio represents how much of the available credit the customer spent by their credit cards.

GOAL

This project aims to construct a predictive model that serves as a practical tool for informing business strategies within the banking sector. Our objective is to accurately predict customer attrition by analyzing a dataset of 10,000 customers from Kaggle, which includes variables such as age, salary, marital status, credit card limit, and category, among 18 total features. With 16.07% of the customers in the dataset having churned, our focus is on identifying the underlying patterns and indicators that lead to customer turnover.

Employing advanced statistical techniques like logistic regression, CART, Random Forest, and Gradient Boosting, the project seeks to delve into the complexities of customer behavior and its impact on attrition. The target audience for this project includes stakeholders within a bank, as the insights derived aim to guide and improve customer retention strategies. The ultimate goal is to provide actionable intelligence that can reverse the trend of customer turnover, thereby ensuring the bank's competitive advantage through enhanced service delivery and customer satisfaction.

STATISTICAL METHODS

In this project, four different methods (Binary logistic regression, CART algorithm, Random forest and XGBoosting) are used to analyze credit card attrition of customers with the given data set. There are several reasons behind using not only one method but comparing four models with each other. By comparing multiple models, one can assess their performance metrics, such as accuracy, precision, recall and F1 score. This allows us to identify which model provides the most accurate and reliable predictions for credit card customer attrition. Different models may handle the data differently and may be more or less robust to variations in the dataset. Comparing multiple models helps evaluating their generalization capabilities and choosing the one that is likely to perform well on a new data.

Logistic regression models are used to examine how predictor variables influence categorical outcomes. Typically, the outcome for logistic regression is binary indicating the presence or absence of an event. If the logistic regression model involves just one predictor variable, it is called simple logistic regression. However, if there are multiple predictors, encompassing both categorical and continuous variables, the model is called multiple or multivariable logistic regression. The logit link function serves to transform a linear combination of covariate values, which can range from negative to positive infinity, into a probability scale ranging between 0 and 1. Binary logistic regression is a type of generalized linear models (GLIM) with the link function in Eq1 where π_i is the ith mean response and η_i is the systematic component [3].

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i. \tag{1}$$

The systematic component η_i for the logit model is stated in Eq2 as

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \tag{2}$$

where p is the number of predictors, β is the regression coefficient for the corresponding covariate and x is the value for the respective covariate [4].

By using both Eq1 and Eq2, we can state the binary logit model function as in Eq3 [4].

$$\pi_i = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j})}. \tag{3}$$

The second statistical method that is used in this project is CART (Classification and Regression Trees) algorithm. CART algorithm offers several advantages because of its tree-structure. To create a binary tree, a series of node splits is created where each internal node undergoes a test on one of the features. Depending on the outcome of the test, the algorithm branches to the left or right. This splitting process is determined by the predictor that maximizes the

decrease in heterogeneity of the binary response variable, credit attrition status in our consideration. The tree continues to grow until a leaf node (terminal node) is reached, at which point further growth stops, and a binary prediction for the response is finalized [4].

Unlike certain other nonparametric methods for classification and regression, the resulting tree-structured predictors can be relatively straightforward functions of the input variables. For analysts seeking accurate results without time and expertise needed for traditional methods, CART algorithm can be advantageous. Even in situations where conventional methods are preferred, trees remain helpful, especially when dealing with numerous variables, as they can assist in identifying significant variables and interactions [5]. CART can effectively model non-linear relationships making it versatile in capturing complex patterns. This algorithm automatically identifies the most relevant predictors for splitting nodes in the decision tree focusing on the most informative ones which can be valuable in handling datasets with a large number of variables. In other words, it is a greedy algorithm that chooses the best discriminatory feature at each step in the modeling process [4].

The rpart package in R that is used in this project to model CART algorithm uses Gini index as an impurity index. The main focus behind the model is to minimize cost while having satisfactory amount of leaf nodes [4]. Equation 4 gives the formula of the cost of this tree with error indicating the ratio of misclassified cases and N indicating the number of leaf nodes in the model.

$$Cost_{CP}(Tree) = Error(Tree) + C_p \cdot N(Tree) \tag{4}$$

As the third model used in this project, random forest is a member of decision tree models in statistics which is developed as a randomized version of the tree algorithms [6]. This model creates a large number of decision trees using randomization. Also, while creating different forms of decision trees, predictors in these trees are chosen randomly from the data set. The data sets for each decision tree are sampled with replacement which means the same observations may occur in different decision trees since all samples will be drawn from the original data set and this process is named bagging or bootstrap aggregating [4]. The ensemble approach of Random Forest enables the model to prioritize the most informative variables while automatically reducing the importance of those with lower predictive effectiveness. In the context of overfitting, random forest is more advantageous than CART algorithm since CART has only one decision tree. Random forest algorithm constructs an ensemble of several trees which provides more generalized results on new data rather than fitting closely of the training set.

As the final statistical method that is used in the project, XGBoost (eXtreme Gradient Boosting) is applied on the credit card customer attrition dataset. The main concept behind XGBoost is to create a better learner with good hypothesis from weak learners with poor hypothesis by building predictive models [7]. Gradient Boosting Algorithms create a series of shallow trees by providing correcting the errors for each tree from its prior tree, then combine together to have the final version. The most precisive characteristics of XGBoosting when comparing to

other boosting algorithms are its fast and accurate structure [4]. Also, adding a regularization term (L1 and L2) to XGBoost helps it become better at making predictions on new data. This is important because decision trees can sometimes memorize the training data too much, and the regularization term helps preventing that [8].

Data Preprocessing

Before initializing the statistical processes by creating models with different methods, the raw data set is studied in some different perspectives. Some statistical models typically cannot handle missing data so that we check if there is any missing values, which concludes there is no null values in the data set. On the other hand, to ensure consistent distribution between train and test sets for binary outcomes, a seed that gives a the same proportions for the whole data, train and test sets is set with 16.1% attrited and 83.9% existing customers. Also, to understand the distributions of some predictors and their relations, we employed a data visualization study which can be found under the supplementary file with all the extended details. Upon cleaning the data, we then proceeded to calculate a correlation matrix to explore the relationships between all numeric variables. Notably, we identified a high correlation value between 'Credit_Limit' and 'Avg_Open_To_Buy'. The presence of a high correlation coefficient of 0.996 between 'Credit_Limit' and 'Avg_Open_To_Buy' indicates a strong multicollinearity, which is problematic for the statistical models due to the risk of unstable estimates and model overfitting. To mitigate this issue, a prudent step in our analysis is to exclude the 'Avg_Open_To_Buy' variable from the dataset, thereby simplifying the model and enhancing its interpretability and reliability.

Model 1: Logistic Regression

A null model with no predictors and a full model with all the predictors are performed on the clean data set. The full model demonstrates strong predictive power with 90.52% accuracy, significantly surpassing the baseline rate of random chance predictions. It excels in identifying those who will continue as customers in the bank (specificity of 96.99%), but less so in pinpointing those likely to leave the bank (sensitivity of 57.14%). The model is correct 78.66% of the time when predicting customer attrition (precision), and the F1 score of 0.66197 indicates a balanced model considering both precision and recall. The Area Under the Curve (AUC) of 0.9332 from the ROC analysis indicates a very good ability of the model to distinguish between customers who will stay and those who will attrite. On the other hand, the null model predicts all customers as existing (negative class), failing to identify any attrited customers (positive class). Despite an overall accuracy of 83.75%, which matches the proportion of existing customers, the model lacks true predictive ability, as indicated by a sensitivity of 0 and a Kappa statistic of 0. Essentially, it didn't correctly predict any customer attrition. The high specificity of 1 is misleading since it merely reflects the model's bias towards predicting the majority class.

After assessing our logistic regression model's ability to predict customer attrition, we check for multicollinearity to ensure the predictors' independence. The output indicates that all predictor variables in the logistic regression model have Generalized Variance Inflation Factor (GVIF) values well below the common threshold of 10, even after adjustment for degrees of freedom. This suggests that multicollinearity is not a concern for this model, indicating that each variable contributes independently to the prediction of customer attrition. This is a positive sign for the model's validity, as it implies that the predictors can be reliably used to assess their individual impact on the likelihood of customer attrition.

We employ a stepwise regression function, which iteratively adds or removes variables based on specified criteria, such as the Akaike information criterion (AIC) or Bayesian information criterion (BIC). This technique aims to identify a simpler model that retains predictive power with fewer variables, thereby enhancing the interpretability and generalizability of the model. The model demonstrates high accuracy (90.81%) in predicting customer attrition, significantly better than random chance, as indicated by the p-value being less than 2.2e-16. Sensitivity is moderate at 58.66%, while specificity is very high at 97.05%, showing the model is particularly effective at identifying customers who will not attrite. The positive predictive value (79.42%) and negative predictive value (92.36%) are both strong. The balanced accuracy of 77.85% and an F1 score of 0.67483 reflect a reasonable balance between precision and recall. Overall, the model is quite adept at discerning the likelihood of customers leaving the bank, which is critical for informed decision-making in customer retention strategies. The stepwise model exhibits a slight improvement in sensitivity compared to the full model. This improvement indicates that the stepwise model is marginally better at correctly identifying customers who will attrite. Even a small increase in sensitivity is valuable in a banking context, as it means fewer customers who are likely to leave the bank go unnoticed. The AUC (Area Under the Curve) for the stepwise logistic regression model is 0.9329, which is an excellent score. An AUC close to 1 indicates that the model has a high ability to correctly classify those customers who will attrite (case = 1) and those who will not (control = 0). This high AUC score signifies that the stepwise model provides strong discriminative power in identifying the likelihood of customer attrition, which is critical for effective customer retention strategies in the banking industry.

Outliers can significantly impact the performance of a model so we flag data points where residuals from the logistic regression model exceed three times the standard deviation from the mean residual. This suggests the presence of substantial atypical values which could affect the model's accuracy. Such a significant number of outliers merits further investigation to understand their nature and consider appropriate adjustments to the model or data. Then, it is checked if there are high leverage points in the logistic regression model. Leverage points can unduly influence the model, but the output "named integer(0)" indicates no such points are found in the data, suggesting a stable model with no single data point having an excessive impact on the fit. Observations with Cook's distance greater than $4/(n-p-1)$ are typically considered to be influential. Given that numerous influential points were found, it suggests that the model may be sensitive to specific data points. This could indicate either actual influential observations that merit closer inspection or potential data issues. The presence of many

influential points might necessitate a deeper investigation into the data or a reconsideration of the model's specifications.

After removing observations identified as influential based on high Cook's distance from the training data set, a refined logistic regression model is generated. This process aims at enhancing the model's robustness by excluding data points that could unduly affect the model's estimates. The refined logistic regression model shows a reduction in both residual deviance and AIC values compared to the previous full model. The residual deviance and AIC of the previous full model are 3804 and 3858 while this models parameters are 2233 and 2297, respectively. This improvement indicates a better fit to the data with less unexplained variance, while maintaining model parsimony, meaning the model is achieving more with less.

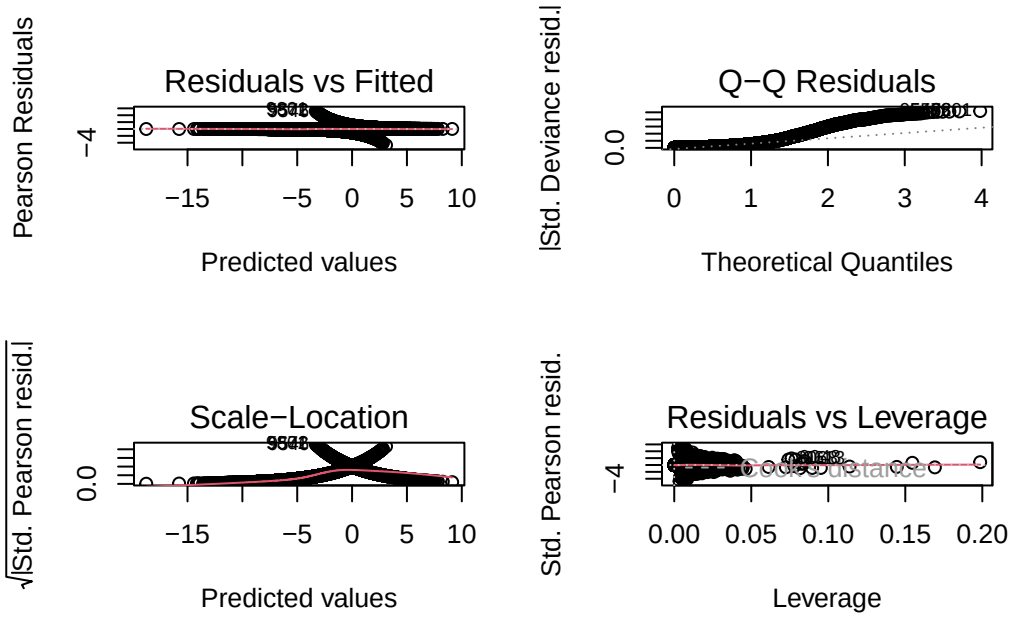

Figure 1: Residual plots for the final logit model

The diagnostic plots for the refined model in Figure 1 reveal that some issues persist (You can see the plots for the previous model in Supplementary Files). The "Residuals vs Fitted" plot shows a curve, hinting at potential non-linearity in the residuals, similar to the full model. The "Q-Q" plot indicates a departure from normality, with more pronounced deviations from the expected line than in the full model, suggesting that the distribution of residuals may be skewed or have heavy tails. The "Scale-Location" plot suggests possible heteroscedasticity, with residuals spreading as fitted values increase, an issue that also appeared in the full model. However, the "Residuals vs Leverage" plot demonstrates an improvement, with data points well-contained within Cook's distance, suggesting effective mitigation of influential points. In summary, while the refined model shows fewer influential points, indicating some improvement, concerns regarding non-linearity and non-normality of residuals remain when compared to the full model's residual plots.

The refined model achieves an accuracy of 90.77%, with sensitivity at 58.05% and specificity at 97.11%. These metrics indicate a reliable model, but when compared to the full and stepwise models, the improvement in performance is not significant. The precision and F1 score, standing at 79.58% and 67.13% respectively, also do not show marked improvement. The balanced accuracy of 77.58% underscores this point. Despite refining the model by removing influential points, the gains in predictive performance are marginal, suggesting that the initial models were already capturing the essential patterns in the data effectively.

In summarizing the application of logistic regression to the banking dataset, it is evident that while the models achieved commendable accuracy and specificity, the sensitivity scores remained modest. The full logistic regression model, the stepwise adjusted model, and the refined model post-outlier adjustment consistently showed high accuracy, exceeding 90% in some cases, and specificity scores were similarly robust, often surpassing 97%. These results affirm the models' proficiency in correctly identifying customers who will not attrite. However, the sensitivity scores—reflecting the models' ability to correctly identify those customers who will attrite—were consistently lower, hovering around 58%. This suggests a potential limitation in the models' capacity to capture the true positive rate effectively, which is critical for the bank's objective of accurately predicting attrition. Despite attempts to enhance model performance by excluding influential data points, improvements in sensitivity were not substantial, indicating the original models already encapsulated the primary predictive patterns within the data. The takeaway from this method is a recognition of logistic regression as a valuable predictive tool, yet one that may require supplementary techniques or data enrichment to improve sensitivity and thus more effectively meet the bank's objective of identifying at-risk customers.

Model 2: Decision Trees

A decision tree with rpart package is generated to classify customers as likely to attrite, using all available predictors. It is set to grow a detailed tree, allowing splits even with single observations. The intent here is to display the complexity parameter table for the decision tree model to aid in selecting the optimal tree size.

The provided CP table from the decision tree output highlights the variables used in predicting customer attrition and the model's performance at various complexity levels. A lower 'rel error' suggests better fit to the training data, but a minimized 'xerror' indicates robustness in cross-validation, which is essential for model generalization. The decision on pruning the tree will balance these errors, aiming for a model that captures critical patterns without overfitting, crucial for effective attrition prediction in banking.

```
(cp= fit.allpred$cptable[which.min(fit.allpred$cptable[, "xerror"]), "CP"])
```

```
[1] 0.002118644
```

```
(xerr = fit.allpred$cptable[which.min(fit.allpred$cptable[, "xerror"]), "xerror"])
```

```
[1] 0.3243451
```

Above, there are the chosen CP value and its corresponding minimum cross-validation error. This minimal error suggests that at this particular CP value, the decision tree model strikes an optimal balance between complexity and predictive performance, potentially reducing the likelihood of overfitting while maintaining a robust predictive capability for unseen data. This CP value will be used to prune the tree, simplifying the model to this level of complexity to improve its generalization to new data.

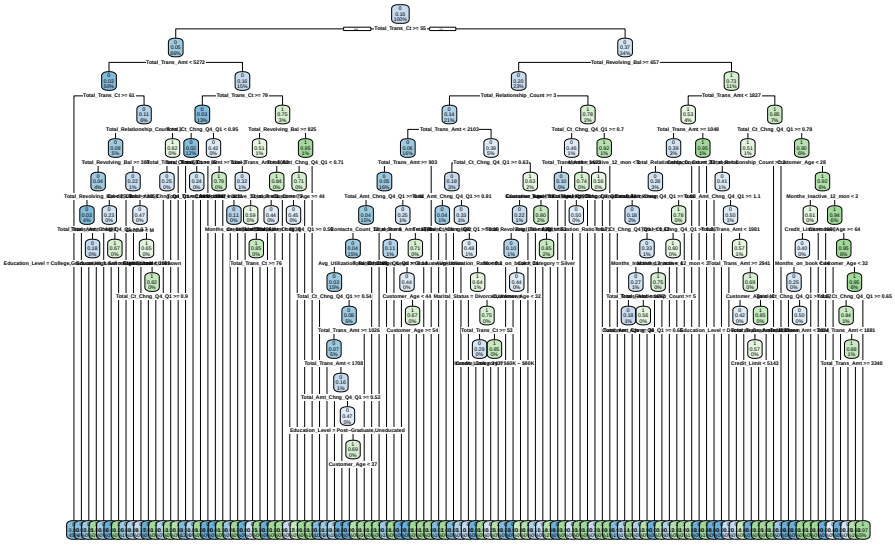

Figure 2: CART decision tree

In Figure 2, a decision tree has been generated to model customer attrition. The tree is complex, with numerous nodes representing various decisions based on customer data attributes. Each node in the tree makes a decision, leading to either another decision node or a terminal node that predicts the customer's likelihood of attrition. The warning indicates that due to the tree's complexity and the large number of nodes, the labels may overlap in the visual representation, making it challenging to discern individual nodes and paths clearly. This complexity suggests a highly detailed model that may capture subtle nuances in the data but also raises concerns about overfitting and model interpretability.

The decision tree model showcases strong performance with an accuracy of 94.07%, indicating reliable predictive capabilities. Notably, the model's sensitivity or true positive rate is 80.24%, demonstrating its effectiveness in identifying customers likely to attrite. This, paired with a high specificity of 96.76%, means the model is accurate in predicting both attrition and

retention. The balanced accuracy of 88.50% suggests a well-tuned model that is equally adept at identifying both classes. This robust performance suggests that the decision tree could be a valuable tool for targeting customer retention efforts in a banking context.

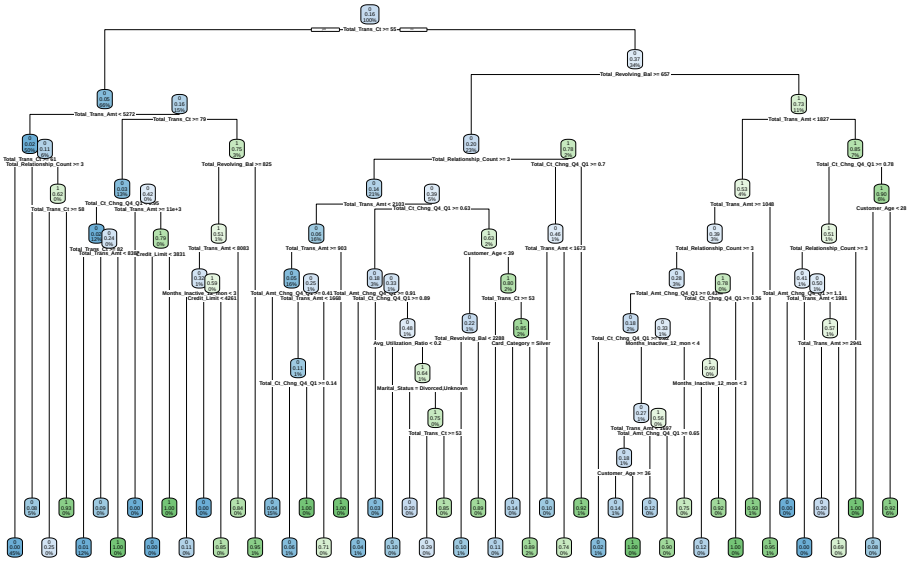

Figure 3: CART pruned decision tree

The decision tree in Figure 2 is pruned to reduce complexity and prevent overfitting so the decision tree in Figure 3 is formed. Pruning is done at the complexity parameter (cp) that minimizes cross-validation error, which is likely to yield a model that generalizes better to unseen data. This visual represents a pruned decision tree, which has been simplified to focus on the most predictive variables for customer attrition. The pruning process removes branches that have little impact on the classification outcome, aiming to improve the model's generalization to new data. This results in a more interpretable and efficient model, potentially enhancing its performance.

The computed resubstitution error rate is 0.0333251. This suggests that after pruning the tree to avoid overfitting, approximately 3.33% of the training data would still be misclassified by the model, indicating the tree's performance on the data it was trained on. The computed cross-validation error rate is 0.05196248. This represents the expected proportion of misclassifications if the model were applied to new, unseen data, based on the pruning parameter (cp) chosen from cross-validation. The pruned decision tree model exhibits a modest enhancement in classification performance over its unpruned counterpart. A slight uptick in accuracy from 0.9407 to 0.9437 and an improved Kappa statistic from 0.7796 to 0.7906 both suggest a more reliable prediction after pruning. Sensitivity shows a marginal increase, indicating a better detection of true positives. The specificity and predictive values also see slight improvements, reinforcing the model's precision and reliability in classifying both positive and negative cases. The balanced accuracy rate, a critical measure of overall performance, has risen from 0.8850

to 0.8904, highlighting a more balanced approach in predicting outcomes across classes. The misclassification error rate is 0.0562963. This value indicates that approximately 5.63% of the predictions made by the pruned decision tree model were incorrect. This low error rate suggests that the model is performing well, accurately predicting the correct class for a high percentage of the instances. It is a direct measure of the model's error and is a complement to the accuracy rate, which in this case is approximately 94.37%.

Model 3: Random Forest

Random forest model targets the prediction of the 'Attrition_Flag' outcome based on a configuration of 500 trees and a maximum of three variables tried at each split (mtry = 3), the model emphasizes variable importance based on impurity measures. The reported out-of-bag (OOB) prediction error is 3.85%, which is a direct assessment of prediction accuracy on the training set itself, using each tree to predict the data not included in its bootstrap sample. This relatively low OOB error rate hints at a robust model with potentially good predictive performance. However, validation on an independent test set is crucial to evaluate the model's ability to generalize beyond the training data.

The top two variables by importance are Total_Trans_Amt and Total_Trans_Ct, indicating they are the most significant predictors in the model. Variables related to transaction amount, count, and changes over time are highlighted as key factors, with other customer-related metrics like revolving balance, utilization ratio, and age following in importance. The data suggests that transactional behavior is more predictive of the model's target variable than customer's demographic data.

The performance metrics for the random forest model on bank customer data indicate strong predictive power. With an accuracy of 95.46% and a Kappa statistic of 0.8208, the model is highly reliable and shows substantial agreement beyond chance. High specificity (98.94%) indicates excellent identification of existing customers (class 0), while good sensitivity (77.51%) means the model is also reasonably effective at identifying customers likely to attrite (class 1). The positive predictive value of 93.41% and the negative predictive value of 95.78% suggest the model is robust in its predictions across both classes. The balanced accuracy of 88.22% demonstrates the model generalizes well for both classes despite any imbalance in the dataset. Overall, the model appears to be highly effective for this analysis. The recall is 0.775, indicating the model correctly identifies 77.5% of the positive class. The precision is high at 0.958, showing that 95.8% of the instances predicted as positive are indeed positive. The F1 score, which balances recall and precision, is also high at 0.973. These metrics collectively suggest that the model has a strong performance in identifying the positive class while maintaining a high level of accuracy in its predictions.

Model 4: Gradient Boosting

For bank customer attrition data, Gradient Boosting is particularly effective due to its ability to capture complex non-linear patterns and interactions between features. It's adept at handling varied types of data, which is often the case with the multitude of variables involved in customer attrition, such as demographics, transaction behavior, and product usage. Its iterative refinement helps in accurately pinpointing customers at high risk of churn.

For our model implementation, the evaluation metric is set to 'AUC' to assess the model's ability to rank predictions rather than output hard probabilities. We train the XGBoost model, iterating through seven rounds, and monitor the Area Under the Curve (AUC) metric for both the training and test datasets. This metric provides insight into the model's ability to distinguish between the classes across each training round.

The XGBoost model demonstrates commendable performance on the test data, as evidenced by the provided metrics. It achieves an accuracy of 93.53%, with a 95% confidence interval between 92.37% and 94.56%, underscoring the model's reliability in making predictions. The model's Kappa statistic is 0.7484, indicating a substantial agreement beyond chance. Sensitivity, or the true positive rate, is 73.25%, suggesting a robust capability to identify the positive class. Specificity is at 97.46%, showing excellent recognition of the negative class. The model's output shows a recall of 0.732, precision of 0.849, and an F1 score of 0.787, indicating a balanced performance in terms of sensitivity and precision. However, the metrics for XGBoost is not sufficient enough comparing to the other statistical models we used.

Table 1: Results for Logit, CART, Random Forest and XGBoost Models

Metrics	Logit	CART	Random Forest	XGBoost
Accuracy	90.77	94.37	95.46	93.53
Sensitivity	58.05	81.16	77.51	73.25
Specificity	97.11	96.93	98.94	97.46
F1	67.13	96.65	97.33	78.63
Precision	79.58	96.36	95.78	84.86
Balanced Accuracy	77.58	89.04	88.22	85.36

Table 1 shows the metrics of four models that we have implemented. As a result, it is seen that both CART and Random Forest shows a better performance than the others. CART has a competitive accuracy (94.37%), slightly trailing behind Random Forest (95.46%). However, the difference is marginal, and CART still demonstrates a strong overall correct classification rate. This model outperforms all other models in sensitivity with the highest value (81.16%). This suggests that CART is particularly effective in correctly identifying positive instances, which is crucial in credit card attrition scenario where capturing all positive cases is a priority. While Random Forest has a slightly higher specificity, CART's specificity (96.93%) is still robust. It excels in correctly identifying negative instances, which is essential for avoiding false

positives. Also, CART leads in balanced accuracy (89.04%) among all models, showcasing a superior balance between correctly identifying positive and negative instances. This well-rounded performance is a key strength of CART. In summary, while Random Forest has a slightly higher accuracy, CART distinguishes itself by achieving the highest sensitivity, competitive accuracy, robust specificity, and a strong balance between precision and recall. The decision to choose CART as the best model is supported by its consistently strong performance across multiple key metrics, making it a reliable choice for credit card attrition analysis.

SUMMARY AND CONCLUSION

In the comprehensive analysis undertaken to predict customer attrition, the Decision Tree (CART) model stood out with an impressive balance across key performance indicators. It achieved an accuracy of 94.37%, with a particularly high sensitivity of 81.16%, which is paramount in detecting potential customer churn. Specificity also ranked high at 96.93%, indicating a robust ability to identify loyal customers. The Random Forest model followed closely with an accuracy of 95.46% and a sensitivity of 77.51%. While the XGBoost model displayed a solid accuracy of 93.53%, its sensitivity at 73.25% was slightly lower than desired for this specific objective.

The logistic regression model, despite high overall accuracy (90.77%) and specificity (97.11%), lagged in sensitivity at just 58.05%. This metric is critical in the context of banking, where the ability to predict customer attrition is valued. In essence, the logistic model's capacity to identify churn was not as robust as the other models explored.

In terms of variable importance, transaction amount ('Total_Trans_Amt') and transaction count ('Total_Trans_Ct') emerged as top indicators for predicting customer departure, underscoring the significance of transaction frequency and volume as markers of customer engagement and satisfaction.

It is also pertinent to note that a high correlation (0.9959) was detected between 'Credit Limit' and 'Average Open To buy', prompting the removal of the latter to prevent multicollinearity in the models, thus refining the predictive accuracy.

Conclusively, the CART model's superior performance, especially in terms of sensitivity, makes it the recommended approach for the bank to adopt in predicting customer churn. The bank can leverage this model to enhance its customer retention strategies by closely monitoring transaction behaviors and adjusting credit limits accordingly.

Moving forward, banks are advised to implement these findings into their operational frameworks, potentially integrating real-time predictive scoring into their customer relationship systems. Further research could benefit from exploring deeper customer data insights, refining model features, and perhaps considering ensemble models that combine the predictive

strengths of various algorithms. Continuous refinement and updating of these models are essential to adapt to the evolving customer profiles and market dynamics, ensuring that banks remain proactive in their customer retention efforts.

REFERENCES

[1] Rico-Poveda, C.A., & Galpin, I. (2020). Forecasting Credit Card Attrition using Machine Learning Models. ICAI Workshops.

[2] Mourtas, S. D., Katsikis, V. N., & Sahas, R. (2023). Credit Card Attrition Classification Through Neuronets. In P. Stanimorovic, A. A. Stupina, E. Semenkin, & I. V. Kovalev (Eds.), Hybrid Methods of Modeling and Optimization in Complex Systems, vol 1. European Proceedings of Computers and Technology (pp. 86-93). European Publisher.

[3] Nick, T.G., Campbell, K.M. (2007). Logistic Regression. In: Ambrosius, W.T. (eds) Topics in Biostatistics. Methods in Molecular Biology™, vol 404. Humana Press.

[4] Bar, H., Ravishanker N., Asha, G. Statistical Practice for Data Science: with Hands-on Illustrations using R (Draft Version)

[5] Sutton, Clifton D. (2005). Handbook of Statistics: Chapter 11 - Classification and Regression Trees, Bagging, and Boosting.

[6] Rigatti, Steven J. (2017) Random Forest. J Insur Med. 47 (1): 31–39.

[7] Ramraj, S., Uzir N., Sunil R., Banerjee, S. (2016). Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets. International Journal of Control Theory and Applications. Vol 9. Number 40.

[8] Chen M., Liu Q., Chen S., Liu Y., Zhang C. H., Liu, R. (2019). XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System. IEEE Access, vol. 7, pp. 13149-13158.