Essentials of Data Analytics

CSE3506 - Winter 2022 2023

Final Report

Project Title:

Credit Card Churn Predicition - A Machine Learning Approach

By

20BCE1420 Karthik Raj R

20BCE1441 Kartik Deepu

20BCE1452 Aryan Vigyat

20BCE1479 Manvik Sreedath

CONTENT

ABSTRACT

Predicting credit card churn is a crucial task for credit card businesses to keep consumers and boost profits. The phenomenon known as credit card churn occurs when a client discontinues using or cancels their credit card. Due to the decline in revenue and rise in marketing expenses, this might cause credit card firms to suffer large losses.

Many consumer behavior patterns, including credit card usage, payment history, and demographics, are examined in order to predict credit card churn. To create predictive models for credit card churn, machine learning algorithms such as logistic regression, decision trees, and neural networks are used. Credit card firms can use these algorithms to identify at-risk clients and take preventative action to keep them. These actions could take the form of better customer service, targeted marketing initiatives, and exclusive deals.

Credit card firms may boost client loyalty, boost revenue, and boost their bottom line by lowering credit card churn. Credit card firms can make educated judgments about client retention efforts thanks to the useful insights into consumer behavior that predictive models for credit card churn offer. By anticipating client requirements and preferences, these models can also be utilized to find fresh chances for revenue development. Overall, predicting credit card churn is a critical challenge for credit card firms, and machine learning algorithms can assist them in minimizing churn and boosting customer happiness to achieve considerable commercial gains.
.

Keywords: credit card churn, consumer behavior, predictive models, machine learning algorithms, logistic regression, decision trees, neural networks, client retention, customer service, targeted marketing

INTRODUCTION

An important topic of financial industry research is credit card churn prediction. When a consumer stops using their credit card or cancels it, this is referred to as credit card churn. For credit card issuers, this practice creates a serious problem because it has a negative effect on both their customer base and revenue. In order to identify consumers who are at risk of leaving and take proactive steps to keep them, credit card firms have been actively building predictive algorithms.

Data collection, pre-processing, feature selection, model construction, evaluation, and deployment are just a few of the processes that go into the credit card churn prediction process. Data collection includes gathering pertinent client information, including credit card usage, payment history, demographics, and client comments. Pre-processing is the process of preparing data for analysis by removing outliers, handling missing numbers, and normalizing and transforming the data. The process of feature selection entails determining which customer characteristics are most important for churn prediction. Model development is choosing the proper methods and algorithms to create predictive models that can precisely pinpoint clients who are at danger. A predictive model's performance is evaluated using a variety of metrics, including accuracy, precision, recall, and F1-score. In order to identify at-risk customers in real time, the predictive models must be deployed and integrated with the operational systems of the credit card firm.

Effective credit card churn prediction faces a number of difficulties that must be overcome. The dataset's asymmetry is one of the main problems. It might be difficult to create precise predictive models because the proportion of customers who churn is often a small portion of the total customer base. Additionally, because customer behavior patterns are dynamic and subject to change over time, it is challenging to create models that can take these changes into account. Keeping customer data private and secure is another difficulty. When creating predictive

models, credit card firms must ensure that client data is private and secure by adhering to data protection laws like GDPR and CCPA.

Credit card businesses are looking into a number of methods and strategies for enhancing credit card churn prediction in order to deal with these issues. To determine which client categories are most likely to churn, one strategy is to employ customer segmentation. This strategy entails categorizing clients according to a range of factors like age, income, credit history, and consumption trends. Credit card firms can create customized retention strategies that cater to the particular requirements and preferences of these clients by identifying at-risk customer segments.

Another strategy is to ask for consumer input to learn more about the factors that contribute to credit card churn. Surveys, social media, and encounters with customer service are just a few of the ways that businesses can get feedback from their customers. Using this feedback, predictive models can be improved, and retention methods can be made more potent.

In order to create prediction models for credit card churn, credit card issuers are also investigating the use of machine learning algorithms including decision trees, logistic regression, and neural networks. These algorithms can examine massive amounts of client data and find patterns and trends that manual research might miss. To guarantee the precision and efficacy of the prediction models, however, the right algorithms must be chosen and model parameters must be optimized.

In predicting credit card churn, interpretability is also a crucial factor. Credit card churn prediction models should be interpretable, which means that the variables that influence churn prediction should be simple to comprehend. Credit card firms can use this information to better understand the causes of churn and create tailored retention tactics that take these causes into account. Interpretability can also increase consumer confidence in the predictive models and retention tactics used by the credit card firm.

Predictive models must be regularly evaluated to ensure that they continue to be accurate in identifying clients who are at risk because customer behavior patterns

might change over time. Also, in order to adapt to changing market conditions and consumer behavior patterns, credit card businesses must constantly change their models and strategies.

Finding at-risk consumers who are likely to churn and taking proactive steps to keep them is the main goal of credit card churn prediction. This can be accomplished by looking at several client behavior patterns, including demographics, payment history, and credit card usage. To create predictive models for credit card churn, machine learning algorithms such as logistic regression, decision trees, and neural networks might be used.

LITERATURE SURVEY

[2] Al-Najjar et al. (2022) propose a study in which a prediction model for credit card user attrition is created using machine learning techniques. The study used a dataset of credit card transactions and customer information obtained from a major Middle Eastern bank in order to address the issue of customer churn in the credit card industry by creating a model that can identify at-risk customers and assist credit card companies in taking proactive measures to retain them. In order to prepare the data for analysis, the study did data pre-processing, including data cleaning, transformation, and normalization. The dataset contained characteristics such as customer demographics, credit card usage, payment history, and customer feedback. To determine the most important customer variables that affect churn prediction, a correlation-based feature selection technique was used for feature selection. The study came to the conclusion that machine learning algorithms, particularly decision trees, can be used to develop precise predictive models for credit card customer churn. The selected features were then used to develop various machine learning algorithms, including decision trees, support vector machines, and logistic regression. The results of the study can aid credit card firms

in improving client retention and revenue by helping them create more efficient retention tactics.

[10] Joana Dias et.al. focus on the use of predictive models to identify customers who are likely to leave the bank and the development of retention strategies to improve customer retention. The study uses a dataset containing customer transactional and demographic data from a Portuguese bank. The authors compare the performance of several machine learning algorithms, including logistic regression, decision trees, random forests, and support vector machines. The results show that random forest and support vector machines outperform other algorithms in terms of accuracy. The paper also identifies the key factors that influence customer churn, such as the frequency of transactions, account balance, and customer age.

[1] Nie et al. (2019) offer a method for forecasting credit card churn using decision trees and logistic regression, two well-liked machine learning algorithms. By creating a predictive model that can precisely identify at-risk clients and assist credit card issuers in taking preventative actions to retain them, the study aims to solve the issue of customer churn in the credit card business.
To get the data ready for analysis, the study pre-processed it by cleaning, transforming, and normalizing the data. To determine the most important customer variables that affect churn prediction, a correlation-based feature selection technique was used for feature selection. The study assessed the effectiveness of the predictive models using a variety of metrics, including accuracy, precision, recall, and F1-score. The chosen features were then utilized to create logistic regression and decision tree models for churn prediction. The outcomes demonstrated that, in terms of accuracy and F1-score, the logistic regression model performed better than the decision tree model. The study also conducted a sensitivity analysis to pinpoint the key elements that influence churn prediction. The study came to the conclusion that logistic regression is a promising algorithm for credit card churn prediction and can be used by credit card companies to identify at-risk customers and develop targeted retention strategies. The results showed that variables such as the number of credit card transactions, credit card balance, and payment amount were the most critical factors in predicting credit card churn. The study by Nie et al. (2011) offers important insights into credit card

churn prediction using machine learning algorithms and highlights the potential of logistic regression for this application. It also highlighted the importance of feature selection in developing accurate predictive models and the need for continuous monitoring and updating of models to ensure their effectiveness. The results of the study can aid credit card firms in improving client retention and revenue by helping them create more efficient retention tactics.

[3]The application of machine learning algorithms for predicting customer turnover in the telecom sector is covered in the article. Lan Wang et. al. put a lot of effort into creating a predictive model to determine which customers are most likely to leave and creating retention measures to increase client retention. The study makes use of a dataset from a Turkish telecoms firm that includes demographic and user usage information. The performance of several machine learning techniques, such as artificial neural networks, decision trees, and logistic regression, is compared by the authors. The outcomes demonstrate that artificial neural networks perform better in terms of accuracy than other methods. The report also analyses the crucial variables that affect customer churn, including call volume, call volume, and customer tenure. The study sheds light on how machine learning algorithms might help the telecommunications industry estimate customer turnover and increase client retention.

The study by [4] V. Ravi presents a comprehensive investigation into the credit card churn prediction problem in banks, using data mining techniques. The authors propose an ensemble system with majority voting that combines multiple machine learning algorithms, including MLP, LR, decision tree (J48), RF, RBF network, and SVM. The results show that the best performance is achieved when the unbalanced original data is SMOTED, or with the combination of undersampling and oversampling. The tenfold cross-validation method on SMOTED data yielded excellent results with 92.37% sensitivity, 91.40% specificity, and 91.90% overall accuracy. The authors also generated a set of ‘if-then’ rules using a decision tree J48, which could serve as an early warning system for churn modeling, prediction, and management. Overall, the paper highlights the potential of data mining and machine learning techniques in predicting credit card churn and improving customer retention for banks.

[5] Sundarkumar, G., Ganesh, G., Vadlamani, R., and Siddeshwar, V. explains a technique for enhancing the efficiency of one-class support vector machines (SVMs) in the context of unbalanced data sets, particularly for churn prediction and insurance fraud detection.The suggested approach entails selecting a subset of the majority class' instances that are nearer to the one-class SVM's decision border, hence undersampling the majority class. To do this, a one-class SVM is trained on the full data set first, and the instances that have a distance measure (i.e. decision function value) closer to zero are then chosen.The second one-class SVM, which is used for prediction on fresh, unused data, is trained using the undersampled data set. The authors use a number of data sets, including a churn prediction data set and an insurance fraud detection data set, to show the efficacy of their methodology. According to experimental findings, the proposed method works better than other techniques for handling unbalanced data sets, such as random undersampling and oversampling. The authors also demonstrate how the method may be utilized to strike a compromise between specificity and sensitivity, which is crucial in fraud detection applications where false positives and false negatives have different costs.Overall, the research offers a strategy that shows promise for enhancing one-class SVM performance in the presence of unbalanced data sets, with potential applications in a variety of industries, including banking, healthcare, and cybersecurity.

The flow network graph is used by [6] Lin, Chiun-Sin, Gwo-Hshiung Tzeng, and Yang-Chieh Chin to forecast customer attrition. The nodes in the flow network graph are given weights based on the PageRank algorithm, which represents the relative significance of each attribute in forecasting customer turnover. They produce a churn probability score for each customer using these weights, which is used to determine whether or not they are likely to leave.

The performance of the authors' suggested approach is assessed using data on credit card customer attrition, and its results are contrasted with those of a number of different machine learning techniques, such as logistic regression, support vector machines, and decision trees. According to experimental findings, their approach surpasses these other algorithms in terms of F1 score, accuracy, precision, recall, and other metrics.

Ultimately, by fusing rough set theory and flow network graph, the research proposes a novel method for anticipating consumer turnover in credit card accounts. Outside credit cards, the proposed approach has potential applications in numerous other industries and yields encouraging results.

[7] Castro and Tsuzuki describe a churn prediction model for online games that uses login data from users. The goal of the study was to create a model that can reliably pinpoint players who are at danger and assist game producers in taking preventative actions to keep them.The study made use of a dataset of login data gathered from a well-known online game. The dataset contained elements like the quantity of logins, the length of play, and the interval between logins.In order to identify the pertinent elements that contribute to churn prediction, the study used data pre-processing and feature engineering. A frequency analysis-based churn prediction model was then created using the chosen features.

The study by [8] Jamal and Bucklin (2006) suggested a heterogeneous hazard modeling strategy to enhance customer churn diagnosis and prediction. The goal of the study was to create a model that can precisely predict when and how likely churn will occur for certain customer segments.A sizable dataset of demographic and customer transaction data from a telecommunications business was employed in the study. In order to identify the pertinent elements that contribute to churn prediction, the study used data pre-processing and feature engineering. A heterogeneous hazard model for churn prediction was created using the chosen features.The c-statistic and root mean squared error, among other measures, were used in the study to assess how well the prediction model performed. The findings demonstrated that the heterogeneous hazard model performed better in forecasting churn than other models.The study also conducted a segmentation analysis to determine which customer segments had varying turnover risks. The findings demonstrated that customers may be divided into groups according to their tenure, consumption patterns, and demographic traits.According to the study's findings, segmenting and predicting churn can be accomplished effectively using the heterogeneous hazard modelling approach. The results of the study can aid

businesses in creating retention tactics that are more profitable and boost client retention.

[9] The study utilizes customer data including transaction information, credit card limits, feature usage, credit bureau information, and demographic information. The authors compare the performance of different machine learning algorithms to identify the best model with the highest accuracy. The results suggest that random forest is the best model to predict customer churn in the bank's credit card business. The authors also identify the key factors that influence customer churn, which can be used by the bank to develop effective retention strategies. Overall, the study provides insights into the importance of customer retention in the credit card business and the potential of machine learning algorithms to predict customer churn and improve customer retention strategies.

[11] delves into the application of artificial neural networks (ANN) in the banking sector in order to forecast the amount of customers who will stop banking with a particular institution. The authors explain that the rate at which clients transfer to other banks or financial institutions can have a major impact on a bank's profitability. This rate is referred to as "customer churn." As a result, forecasting and avoiding the loss of customers is an essential challenge for financial institutions like banks. In this article, a model for predicting customer turnover that makes use of various machine learning approaches, in particular ANN, is presented. The authors explain that in order to train and test the ANN model, they utilised a dataset that included customer transactions as well as demographic information. This dataset was utilised both to train and test the model. They also discuss the several stages that are required in preparing the data for analysis, such as cleaning the data, normalising the data, and selecting the features to be analysed. After that, the authors proceed to discuss the structure of the ANN model that they applied, which included an input layer, one or more hidden layers, and an output layer. In addition to this, they discuss the activation functions that are utilised in each layer, as well as the training algorithm and performance measures that are used to evaluate the performance of the model. In the latter part of the essay, the findings of the study are discussed. These results demonstrate that the ANN model was successful in achieving a high level of accuracy when forecasting customer turnover. The authors argue that the model might be used by financial institutions

to identify consumers who are at risk of leaving the institution and then take preventative steps to keep those customers.

[12] outlines a revolutionary strategy for predicting the amount of customers who will cancel their credit card accounts. In order to enhance the precision of churn prediction, the strategy that has been developed makes use of a number of different techniques, including unsupervised learning and rough clustering. In the beginning of the article, the author explains the notion of "credit card churn" and the importance of this phenomenon for credit card firms. It then moves on to a discussion of the limits of standard churn prediction models and emphasises the necessity of a churn prediction model that is more accurate and efficient. The methodology being offered is comprised of three primary stages: the preprocessing of the data, the approximate clustering of the data, and the supervised learning stage. During the procedure known as "data preprocessing," the data are checked for errors, converted, and normalised. During the process known as "rough clustering," the data is first divided up into a number of distinct clusters using the rough set theory. During the process of supervised learning, a classification model that can predict churn is trained on each cluster individually. The study analyses and assesses the proposed method by applying it to a real-world credit card dataset and contrasting it with a number of other conventional churn prediction models. The findings demonstrate that the strategy that was proposed achieves higher levels of accuracy, precision, and recall than any of the other models. In conclusion, the article illustrates that the suggested method can greatly increase the accuracy of credit card churn prediction by utilising rough clustering and supervised learning approaches.

[13] evaluates the efficacy of Apache Spark's ML and MLlib machine learning libraries by comparing and contrasting their results when it comes to forecasting the number of customers who may leave a banking institution. The rate at which consumers transfer from one company or brand to another is referred to as customer churn, and it can be extremely important for organisations to accurately estimate customer churn in order to keep their existing customers.
In this study, many different machine learning models are trained and tested with the help of a real dataset obtained from a bank. This dataset includes client demographics, transactional history, and account information. Using the ML and

MLlib libraries, the authors analyse the effectiveness of a variety of methods, such as logistic regression, decision trees, random forests, and gradient-boosted trees. The findings indicate that both ML and MLlib libraries are capable of accurately predicting the likelihood of possible customer churn; however, ML displays an overall performance that is marginally superior. In instance, ML displayed a higher degree of accuracy when predicting users in the test set who will leave the service by utilising logistic regression and random forest techniques. Nevertheless, both libraries accomplished a high level of accuracy, as evidenced by F1 ratings that ranged from 0.77 to 0.87. The authors suggest that additional studies should be conducted to investigate the application of different machine learning algorithms and methods for the purpose of improving the accuracy of churn prediction in the banking industry. In addition to this, they recommend utilising a more extensive dataset in conjunction with feature engineering techniques in order to increase the performance of the model.
In conclusion, the study highlights the potential of machine learning in the banking sector to anticipate client attrition and retention rates. The ML and MLlib libraries can be used to construct successful models for forecasting churn customers; however, ML displays significantly superior performance when used in this particular scenario. This study offers businesses valuable data that can be used to construct machine learning models that can accurately forecast customer churn and help them keep their existing clients.

[14] investigates how machine learning and automated machine learning (AutoML) techniques can be applied in the context of a credit card firm in order to forecast the attrition rate of customers. The loss of customers, which can have a negative impact on a company's income and profitability, is a serious issue that must be addressed. The purpose of this research is to construct prediction models that can reliably identify factors that drive customer turnover and provide insights into how to enhance customer retention rates.
In order to accomplish this goal, the study conducts an analysis of customer data, looking for patterns and trends connected to customer churn by examining factors such as demographics, transaction history, and credit score. When developing predictive models, a number of different algorithms, including Random Forest, XGBoost, and AutoML tools such as H2O.ai, are utilised. In this study, the performance of each model is assessed by contrasting its accuracy, precision,

recall, and F1 scores. According to the findings of the study, predictive models that were constructed with the help of machine learning and autoML technologies are capable of effectively predicting client turnover. The models offer insights into which elements are most influential in customer retention, enabling organisations to make the required changes to improve customer experience and retention rates. The research demonstrates the potential of machine learning and other AI methods in predicting client turnover for financial institutions such as credit card firms. Businesses have the ability to improve their profitability and customer retention rates by developing strategies based on the precise prediction of customer churn. The research also shows how AutoML tools may make the process of generating predictive models much simpler, which makes it much simpler for organisations to apply machine learning solutions. In conclusion, the study sheds light on the ways in which machine learning and autoML tools can be utilised to make accurate predictions regarding the amount of customers leaving a credit card firm. Because the utilisation of these technologies can assist businesses in maintaining their customer base and increasing their profitability, financial institutions can benefit from using this approach.

The issue of churning customers is brought to light in [15], which focuses on the banking business. The practise of clients moving their banking relationships to other institutions, which can lead to a reduction in revenue for the institution in question, is referred to as "customer churn." Using a dataset that includes demographic and transactional data, the authors suggest a supervised learning strategy as a method for predicting the rate at which customers may churn out of a business. The authors address issues with missing values, outliers, and class imbalance as part of the preparation process for the dataset. After that, they evaluate the efficacy of a number of classification algorithms, including logistic regression, decision trees, random forests, and gradient boosting, among others. According to the authors' findings, gradient boosting achieves the highest level of accuracy compared to the other algorithms.
The authors also determine the most essential characteristics for churn prediction, which include the number of transactions, the age of the consumer, and the kind of account. These findings can assist financial institutions in prioritising their client retention efforts for consumers who are at risk. The strategy that has been suggested can be useful for banks since it enables the institutions to recognise

consumers whose accounts are at risk and to take preventative steps to keep such customers. Banks are able to offer individualised incentives to retain clients, such as lower fees, greater interest rates, or customised services, if they can predict churn and know when consumers will leave.
In general, the paper offers insightful information about how the banking industry may make use of machine learning to combat the issue of customer churn and retain existing customers. The strategy can assist financial institutions in retaining more customers, which can result in higher income and higher levels of customer satisfaction.

Proposed Methodology

The following steps are involved in credit card churn prediction:

1. Data gathering: Gathering information on consumer behavior patterns, credit card usage, payment history, and demographics is the initial step in credit card churn prediction.

2. Data pre-processing involves removing any duplicate or missing values from the obtained data and transforming it into a format that machine learning algorithms can use.

3. Feature selection is the process of choosing pertinent features that will significantly affect the prediction of credit card churn.

4. Model Development: Predictive models for credit card churn are created using machine learning methods including logistic regression, decision trees, and neural networks.

5. Model Evaluation: Metrics like accuracy, precision, recall, and F1-score are used to assess the constructed model.

6. Deploying the predictive model into a production environment is the last phase, which will be used to spot customers who are at risk of leaving and take preventative action to keep them.

Dataset Description:

0	CLIENTNUM	8101 non-null	int64
1	Customer_Age	8101 non-null	int64
2	Gender	8101 non-null	object
3	Dependent_count	8101 non-null	int64
4	Education_Level	8101 non-null	object
5	Marital_Status	8101 non-null	object
6	Income_Category	8101 non-null	object
7	Card_Category	8101 non-null	object
8	Months_on_book	8101 non-null	int64
9	Total_Relationship_Count	8101 non-null	int64
10	Months_Inactive_12_mon	8101 non-null	int64
11	Contacts_Count_12_mon	8101 non-null	int64
12	Credit_Limit	8101 non-null	float64
13	Total_Revolving_Bal	8101 non-null	int64
14	Avg_Open_To_Buy	8101 non-null	float64
15	Total_Amt_Chng_Q4_Q1	8101 non-null	float64
16	Total_Trans_Amt	8101 non-null	int64
17	Total_Trans_Ct	8101 non-null	int64
18	Total_Ct_Chng_Q4_Q1	8101 non-null	float64
19	Avg_Utilization_Ratio	8101 non-null	float64
20	Attrition_Flag	8101 non-null	object

Machine Learning Algorithms used are:

1. Extreme Gradient Boosting, also known as XGBoost, is a potent machine learning technique that boosts model accuracy using gradients. It operates by gradually including decision trees into the model while attempting to minimize a loss function. An increasingly accurate model is produced as

each new tree is taught to repair the flaws of the prior trees. XGBoost is ideal for a variety of applications, including classification, regression, and ranking because it is very versatile and can handle complex, non-linear connections between variables.

2. Another boosting approach that iteratively trains weak models to increase overall accuracy is called AdaBoost, short for Adaptive Boosting. It can be used for regression tasks but is most helpful for binary classification challenges. AdaBoost functions by giving samples that were incorrectly classified heavier weights and then retraining the model using the new weights. Until the desired accuracy is attained, this process is repeated.

3. Decision trees are a straightforward yet powerful machine learning approach that may be applied to both regression and classification applications. They operate by repeatedly separating data depending on the variables that provide the most information until a prediction is made. Decision trees can be utilised to understand the underlying facts because they are very interpretable. With complicated datasets, they can, however, overfit and perform poorly.

4. An ensemble technique called random forests mixes various decision trees to increase model accuracy and decrease overfitting. They operate by combining the predictions of various trees that have been trained on various random subsets of the data. With its ensemble approach, random forests perform classification tasks better than individual decision trees and are less prone to overfitting.

5. A common statistical technique for estimating the likelihood of an event based on a number of predictor factors is multiple logistic regression. It is frequently employed in social science and healthcare research, and it can be expanded to handle numerous outcome factors. In order to determine the probability of the outcome variable, multiple logistic regression first estimates the coefficients of the predictor variables. It is a highly interpretable technique that may be utilized to learn more about the connections between the predictors and the results.

RESULTS

Various models have been used in order to generate the most accurate result, i.e. , whether a current customer would switch over to another competitor. Models that have been implemented in our project are as follows:

- Logistic regression - A linear model used for classification that predicts the probability of an event occurring. It uses the concept of the sigmoid function to get the probability.

```
#Model Accuracy
table(test$Attrition_Flag,predicted)
```

```
##    predicted
##        0    1
##   0  227  170
##   1   73 1942
```

```
classerr=mean(predicted!=test$Attrition_Flag)
paste("Accuracy is ",1-classerr)
```

```
## [1] "Accuracy is  0.899253731343284"
```

- Decision trees - A tree-like model used for classification or regression that recursively partitions data based on the features.

```
# Predicting on test data'
y_pred <- predict(classifier_cl, newdata = test)

# Confusion Matrix
cm <- table(test$Attrition_Flag, y_pred)
cm
```

```
##    y_pred
##        0    1
##   0  296  101
##   1   56 1959
```

```
# Model Evaluation
#confusionMatrix(cm)

#Accuracy
accuracy = mean(y_pred!=test$Attrition_Flag)
accuracy = 1 - accuracy
print(paste('Accuracy of the model = ',accuracy))
```

```
## [1] "Accuracy of the model =  0.934908789386401"
```

- Random forest - An ensemble model of decision trees that combines multiple decision trees to improve accuracy and reduce overfitting.

```
# Predicting the Test set results
y_pred = predict(model_rf, newdata = test[-c(1,2)])

# Confusion Matrix
confusion_mtx = table(test[, 2], y_pred)
#confusion_mtx

# Compute accuracy
accuracy = mean(y_pred == test$Attrition_Flag)
print(paste("Accuracy: ", round(accuracy * 100, 2), "%"))
```

```
## [1] "Accuracy:  96.52 %"
```

- XGBoost Classifier - A gradient boosting algorithm that builds a sequence of decision trees to improve the predictive accuracy of a model.

```
# Convert data to DMatrix format and using most important features
dtrain <- xgb.DMatrix(data = as.matrix(train[,c(11,15,17:20)]), label = train$Attrition_Flag)
dtest <- xgb.DMatrix(data = as.matrix(test[,c(11,15,17:20)]), label = test$Attrition_Flag)

# Define parameters for XGBoost model
params <- list(
  objective = "binary:logistic",
  eval_metric = "auc",
  min_child_weight = 1,
  max_depth = 6,
  eta = 0.1,
  subsample = 0.8,
  colsample_bytree = 0.7
)

# Train XGBoost model
model <- xgb.train(
  params = params,
  data = dtrain,
  nrounds = 1000,
  verbose = FALSE
)

predictions <- predict(model, dtest)
binary.predictions <- ifelse(predictions > 0.5, 1, 0)
accuracy <- sum(binary.predictions == test$Attrition_Flag) /  length(test$Attrition_Flag)
print(paste("Accuracy = ",accuracy))
```

```
## [1] "Accuracy =  0.96558872305141"
```

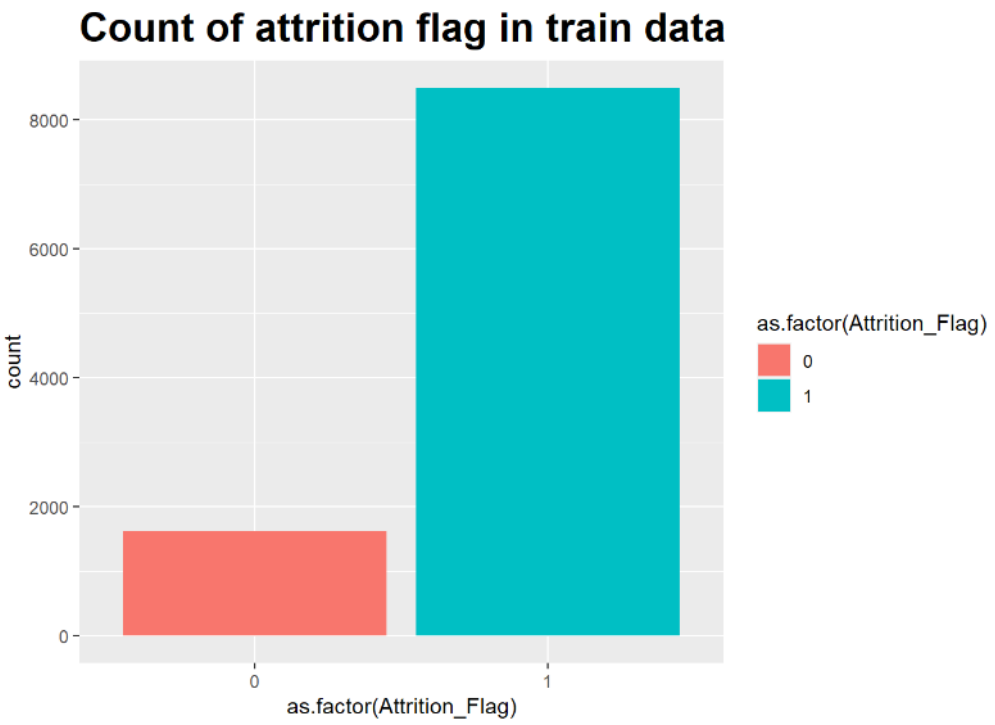
- ADABoost Classifier - An ensemble model that combines multiple weak classifiers into a strong classifier by adjusting the weights of incorrectly classified samples.

```
train$Attrition_Flag=as.factor(train$Attrition_Flag)
model_adaboost <- boosting(Attrition_Flag~., data=train, boos=TRUE, mfinal=50)
summary(model_adaboost)
```

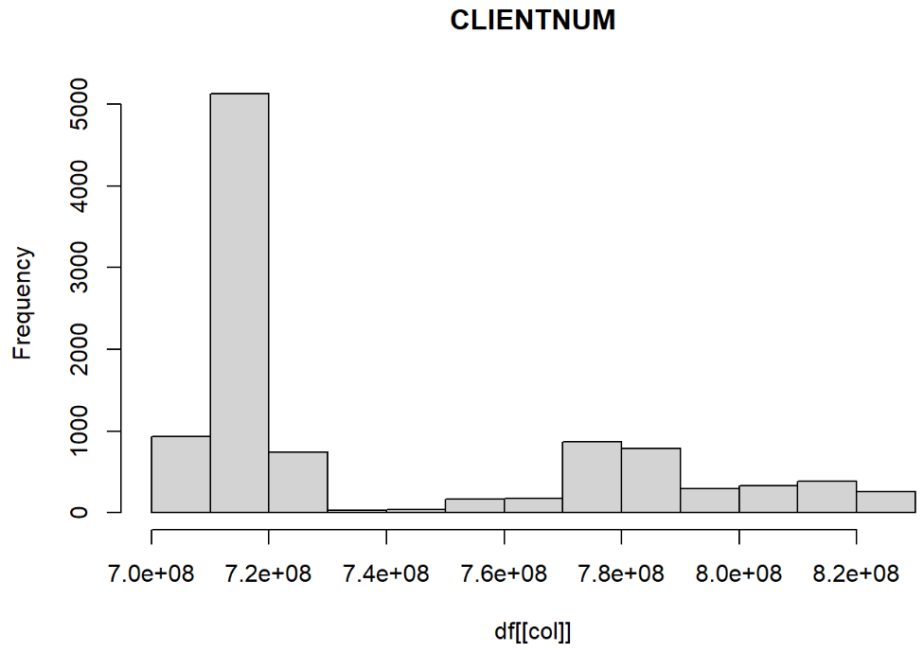
```
##             Length Class   Mode
## formula          3 formula call
## trees           50 -none-  list
## weights         50 -none-  numeric
## votes        15430 -none-  numeric
## prob         15430 -none-  numeric
## class         7715 -none-  character
## importance      20 -none-  numeric
## terms            3 terms   call
## call             5 -none-  call
```

```
#Make Predictions
pred_test=predict(model_adaboost,test)
cm <- confusionMatrix(as.factor(pred_test$class), as.factor(test$Attrition_Flag))
accuracy <- cm$overall[1]
print(paste("Accuracy: ", round(accuracy * 100, 2), "%"))
```

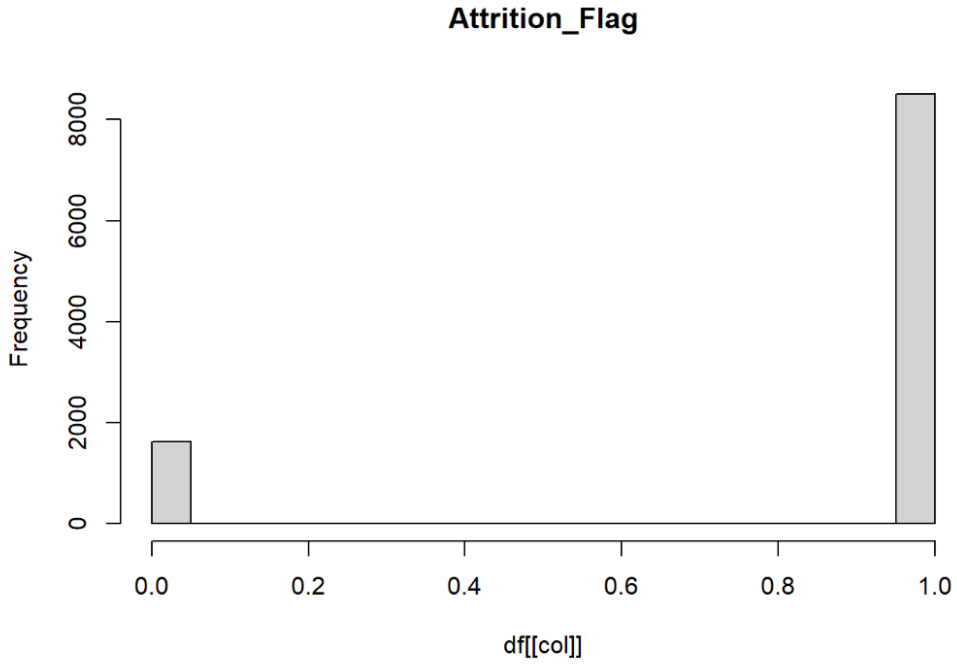
```
## [1] "Accuracy:  97.18 %"
```



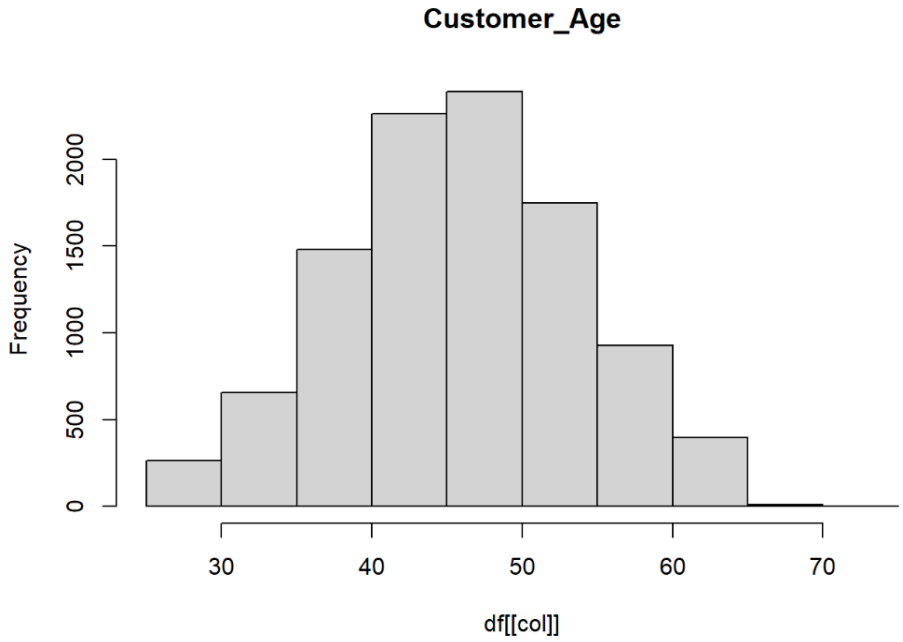
Frequency of clients:



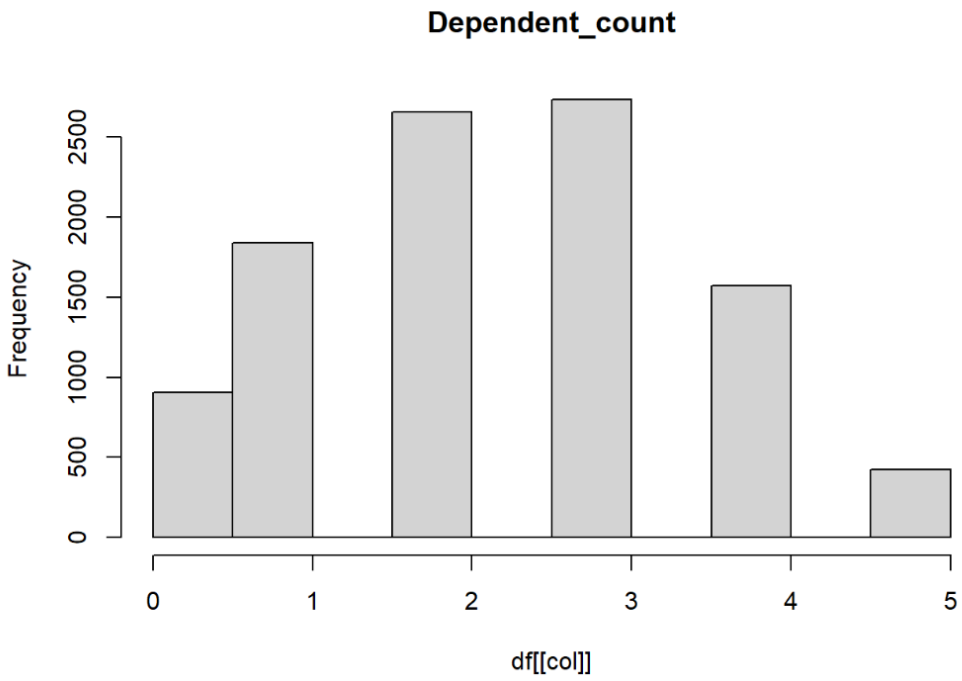
Churn frequency:



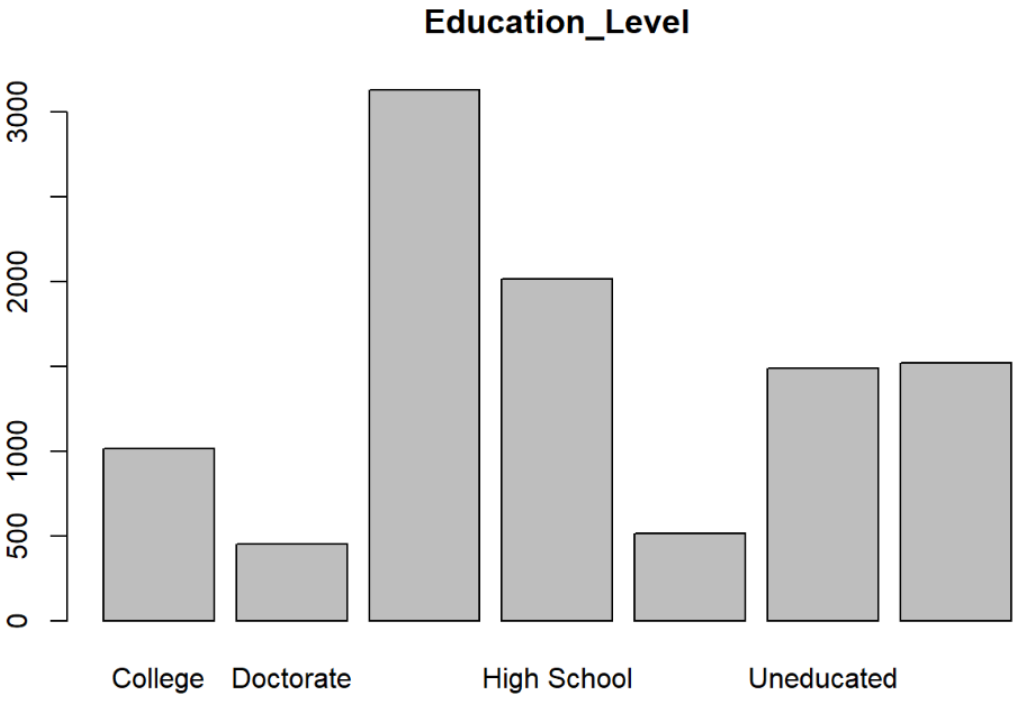
Histogram for customer age:



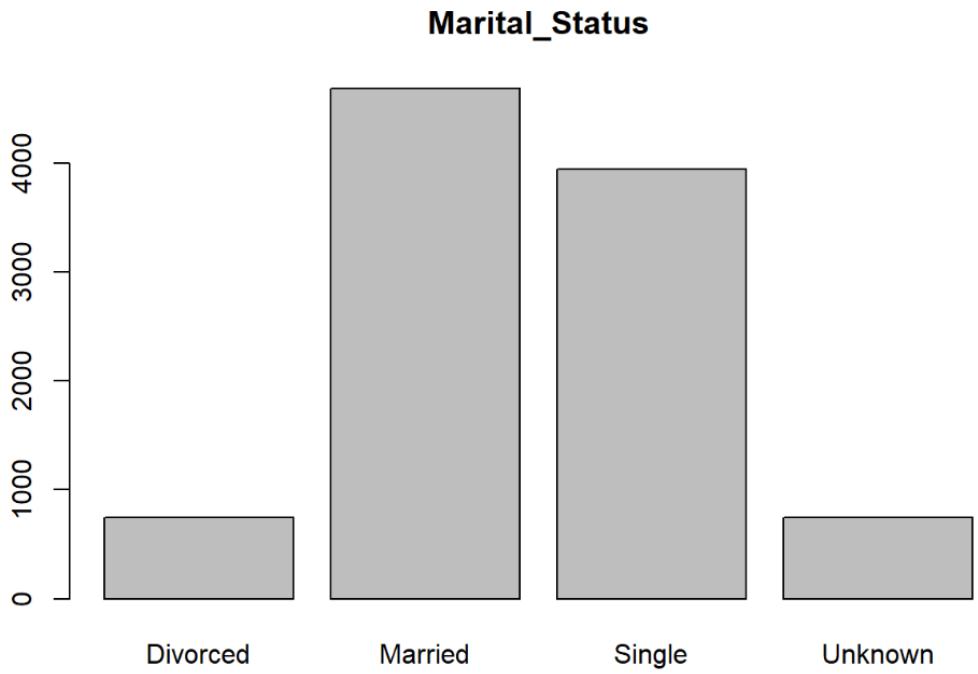
Gender frequency:



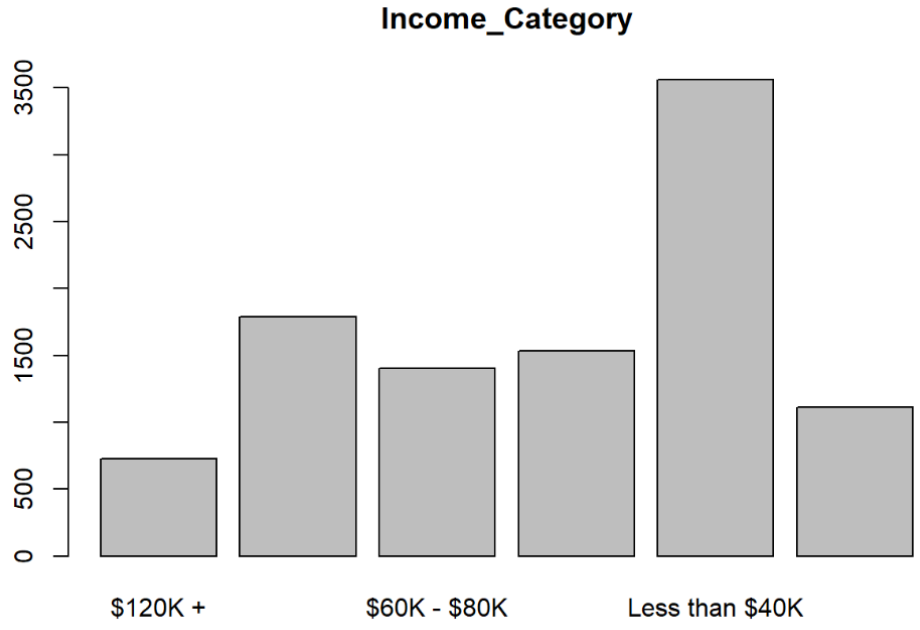
Education level of card holder:



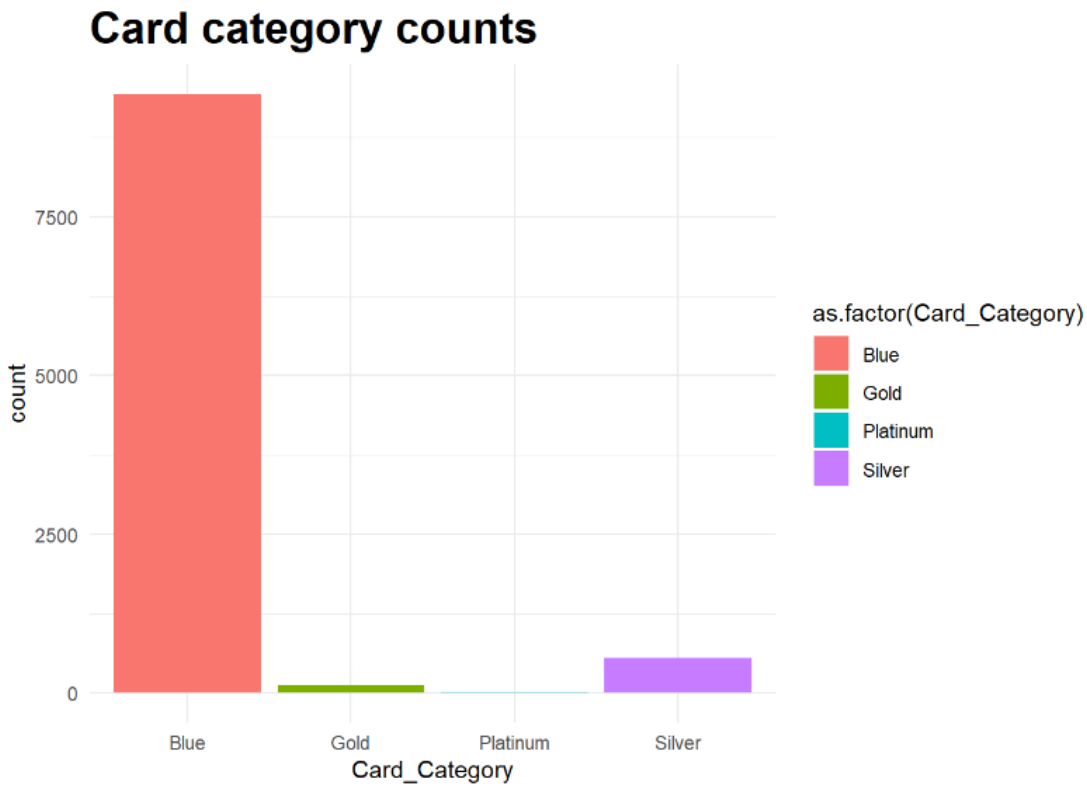
Marital status of card holders:



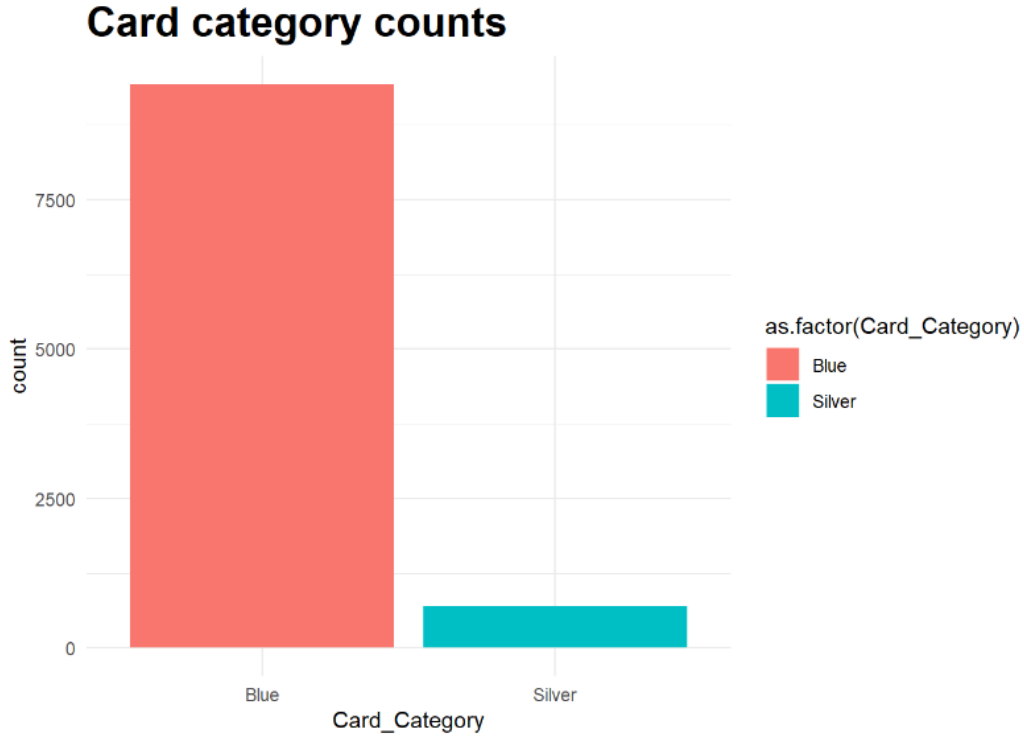
Income category of card holders:

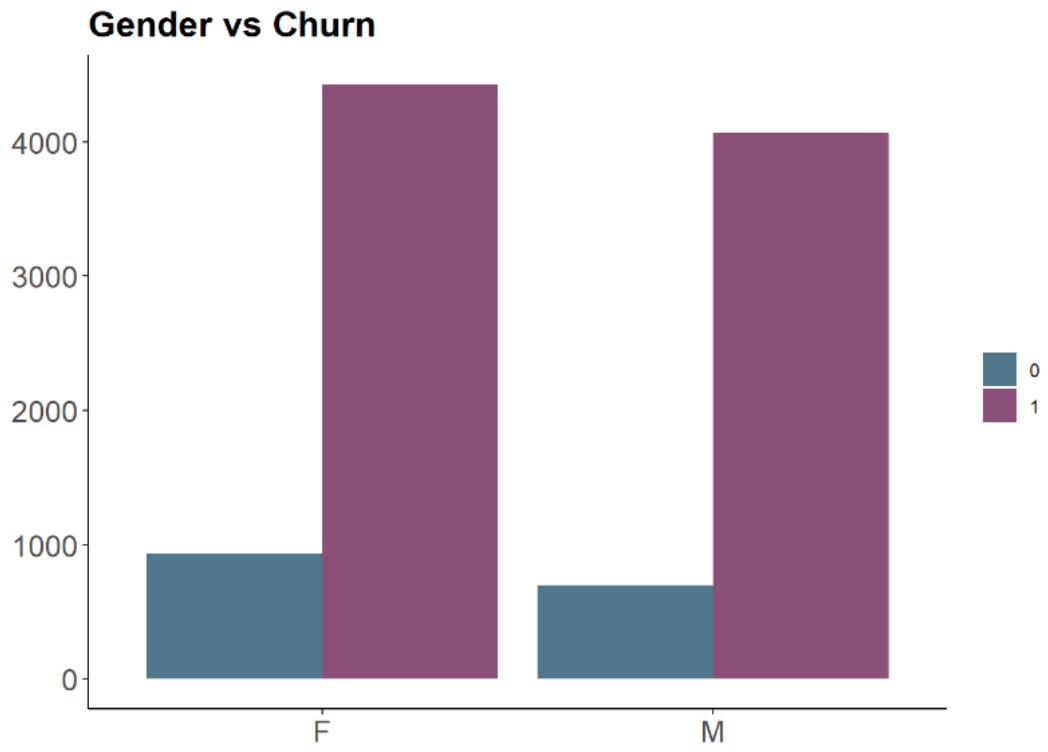


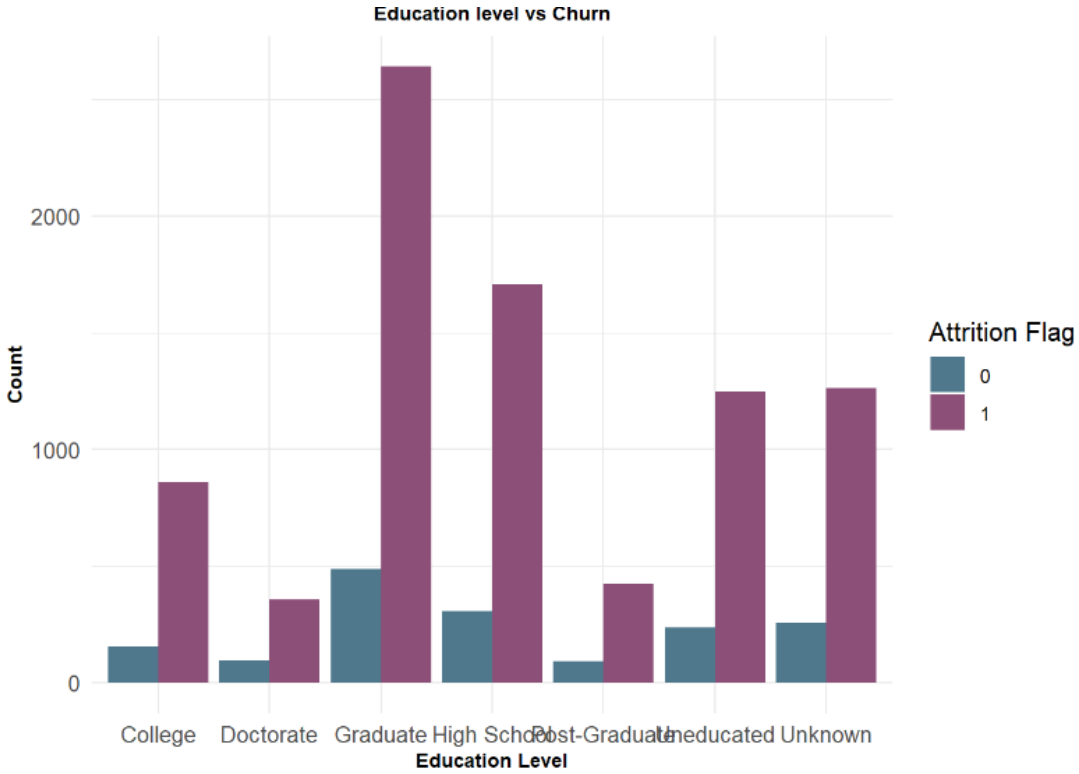
Card categories offered by bank:

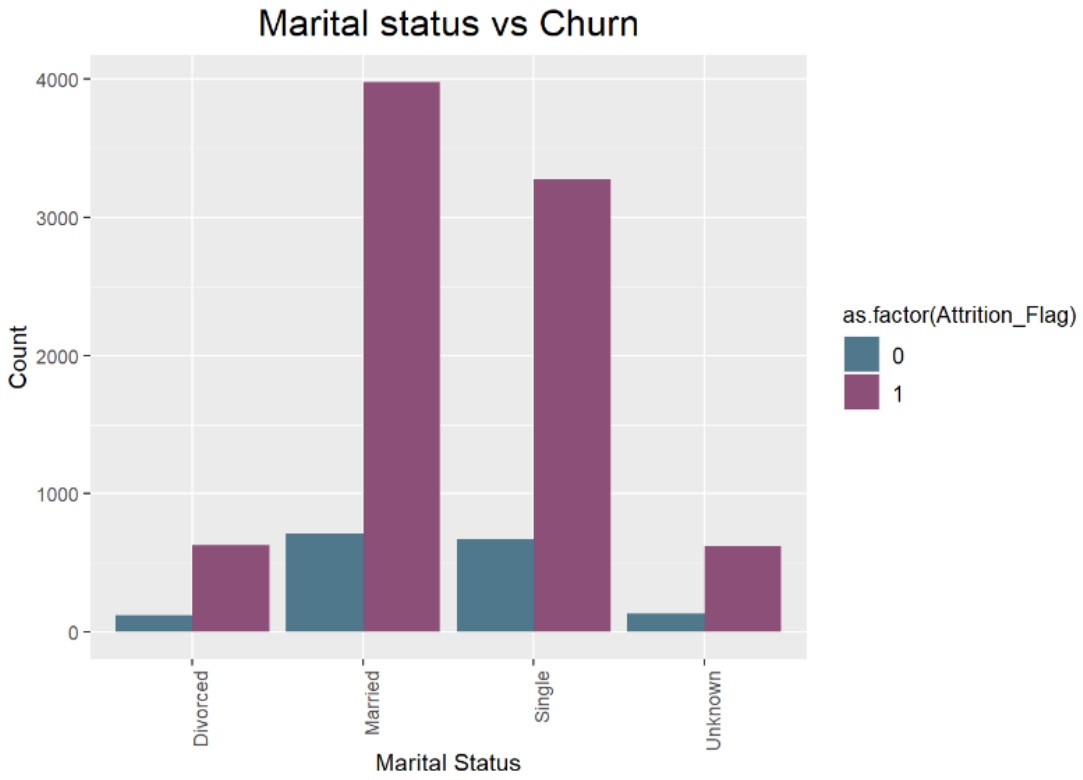


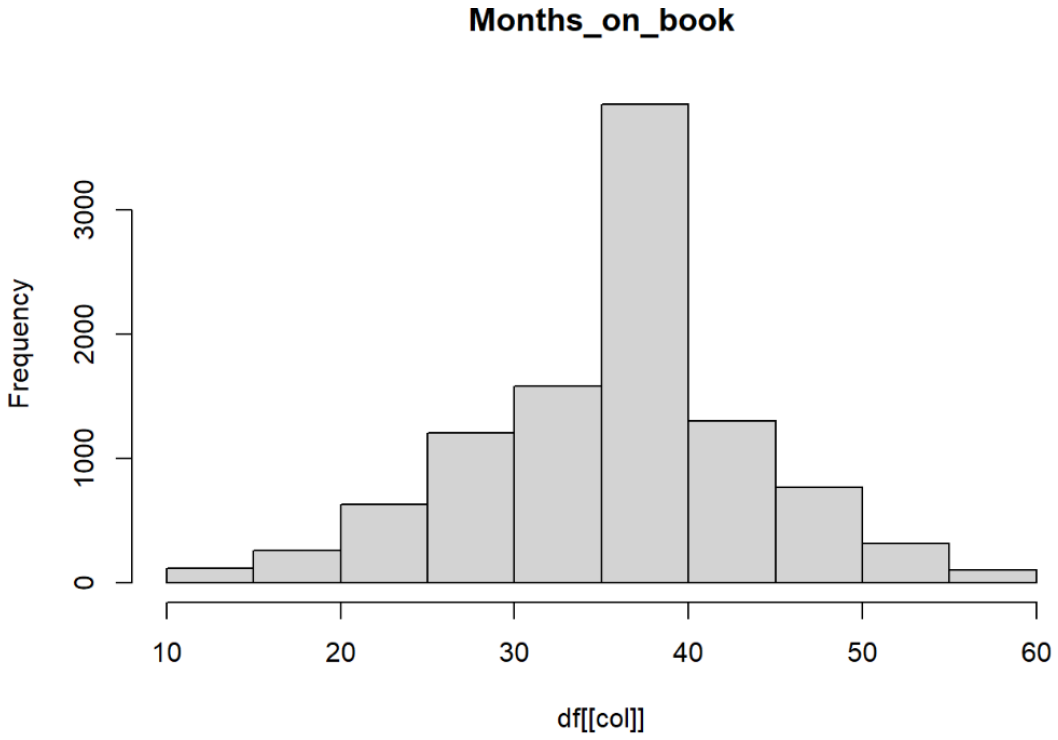
Computing Gold and platinum cards (outliers) with silver

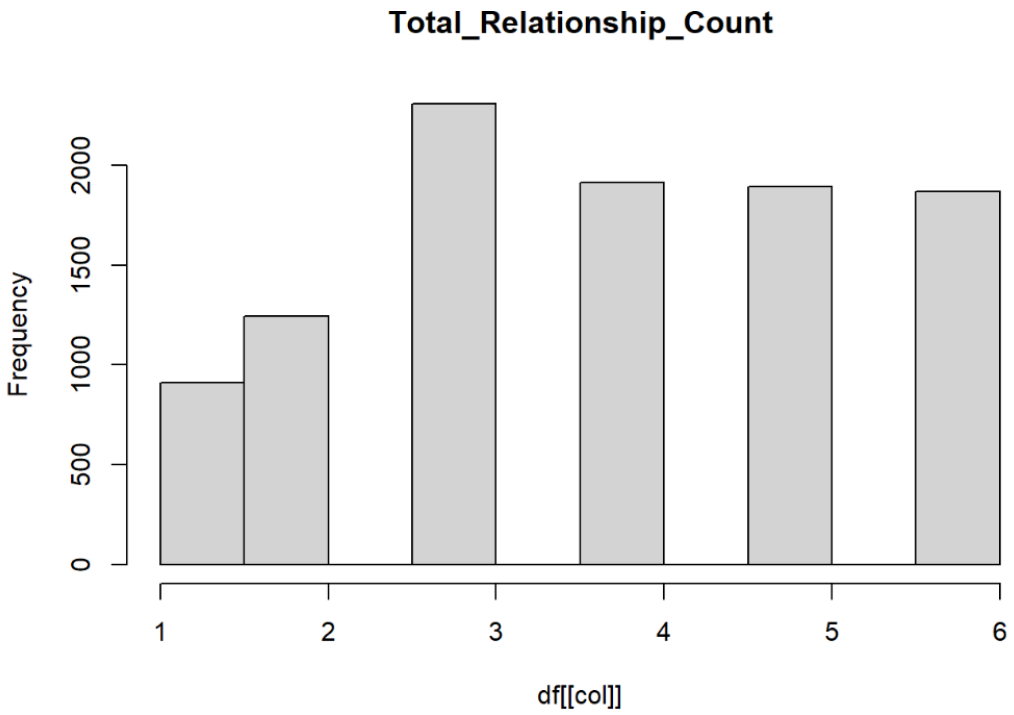



Gender vs Churn
4000
3000
2000
1000
0
F
M
0
1

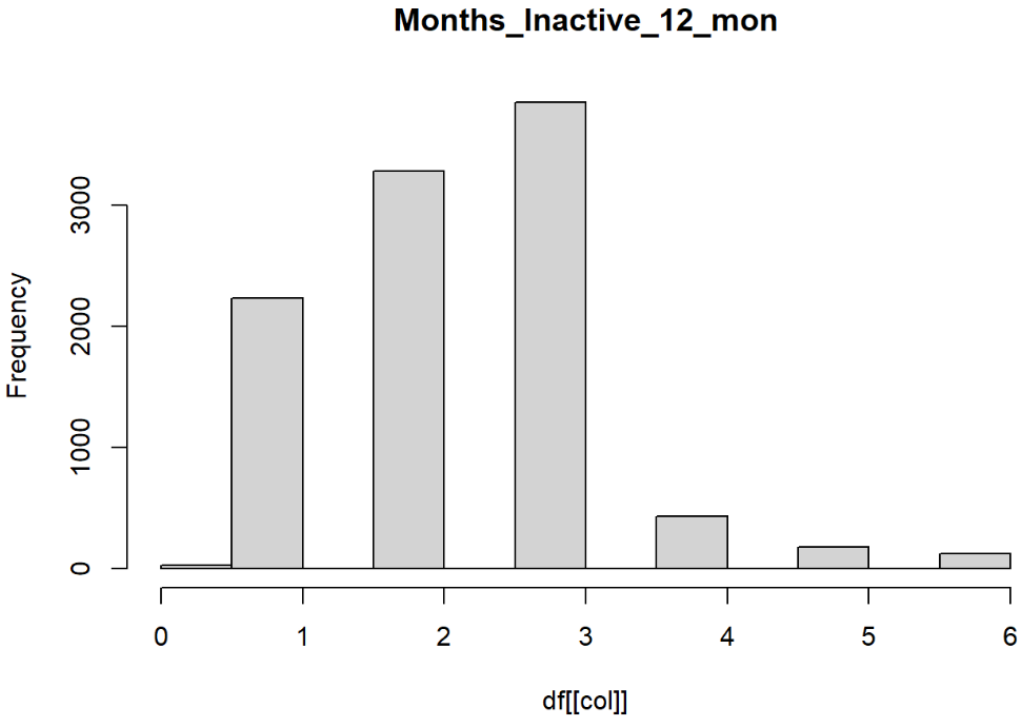

Education level vs Churn
Count
2000
1000
0
College
Doctorate
Graduate
High School
Post-Graduate
Uneducated
Unknown
Education Level
Attrition Flag
0
1

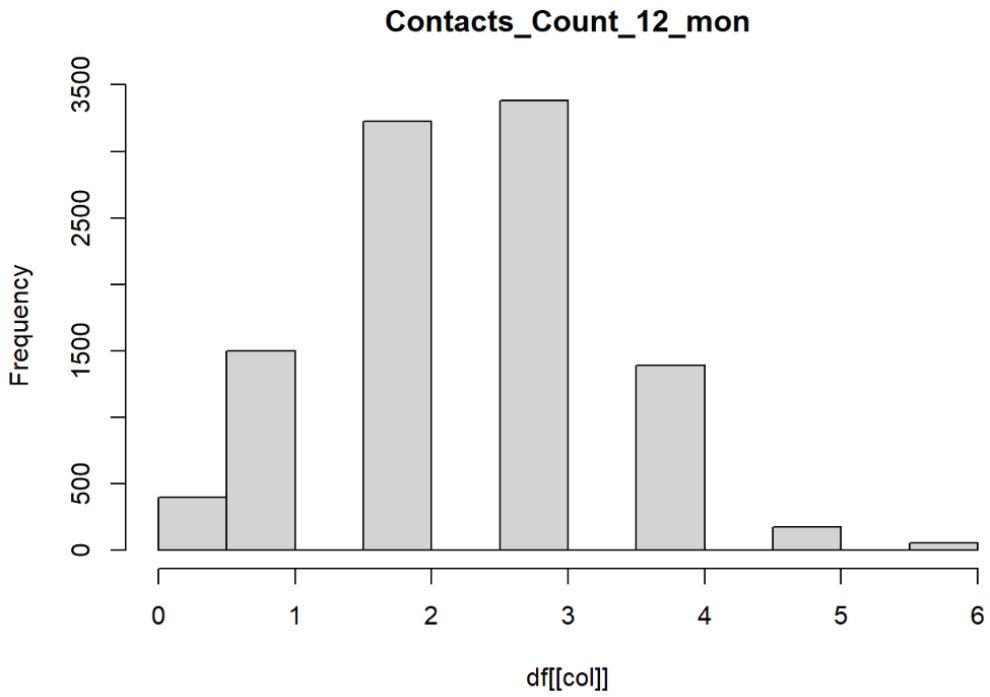

Marital status vs Churn
4000
3000
2000
1000
0
Count
Divorced
Married
Single
Unknown
Marital Status
as.factor(Attrition_Flag)
0
1


Months_on_book
Frequency
0
1000
2000
3000
10
20
30
40
50
60
df[[col]]

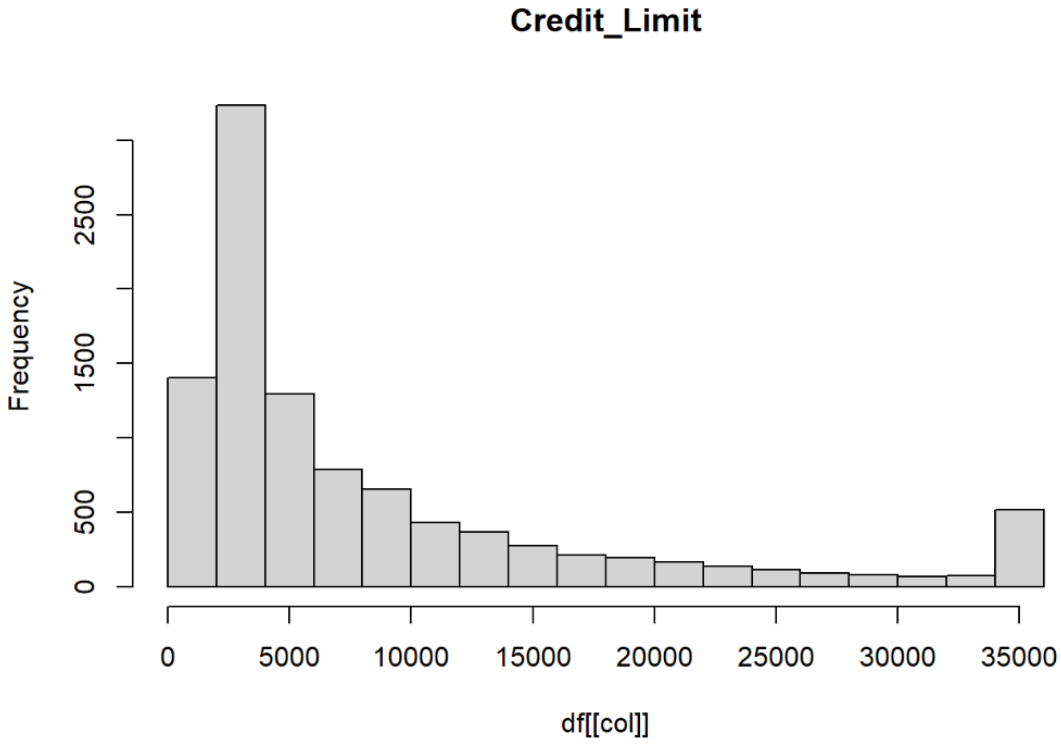


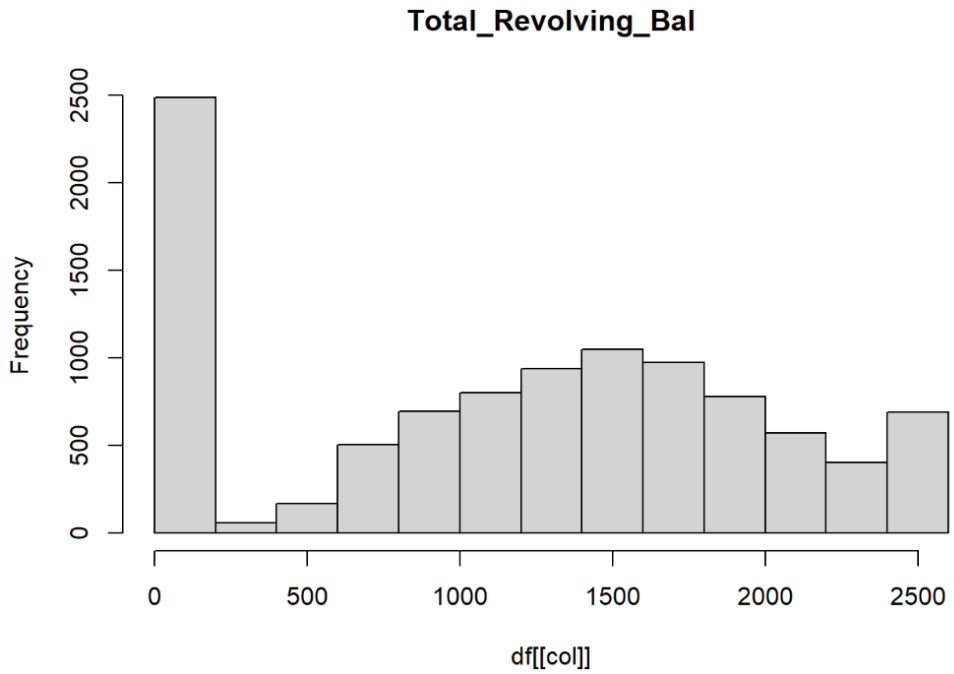
Inactivity duration of card holders:

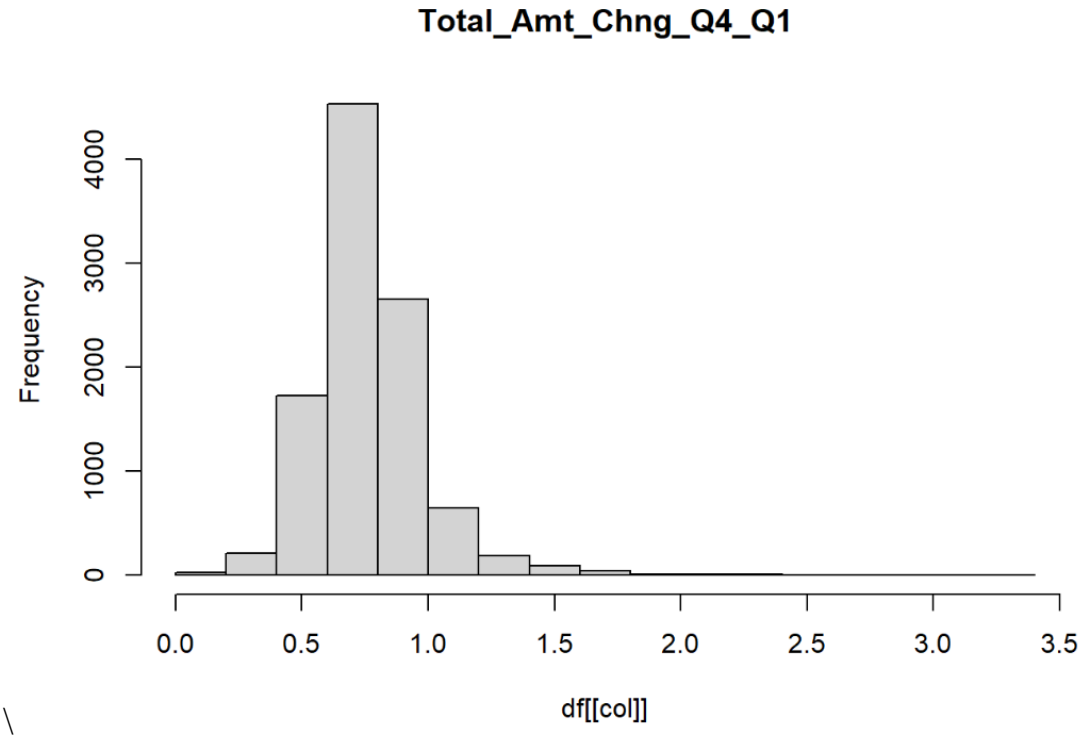




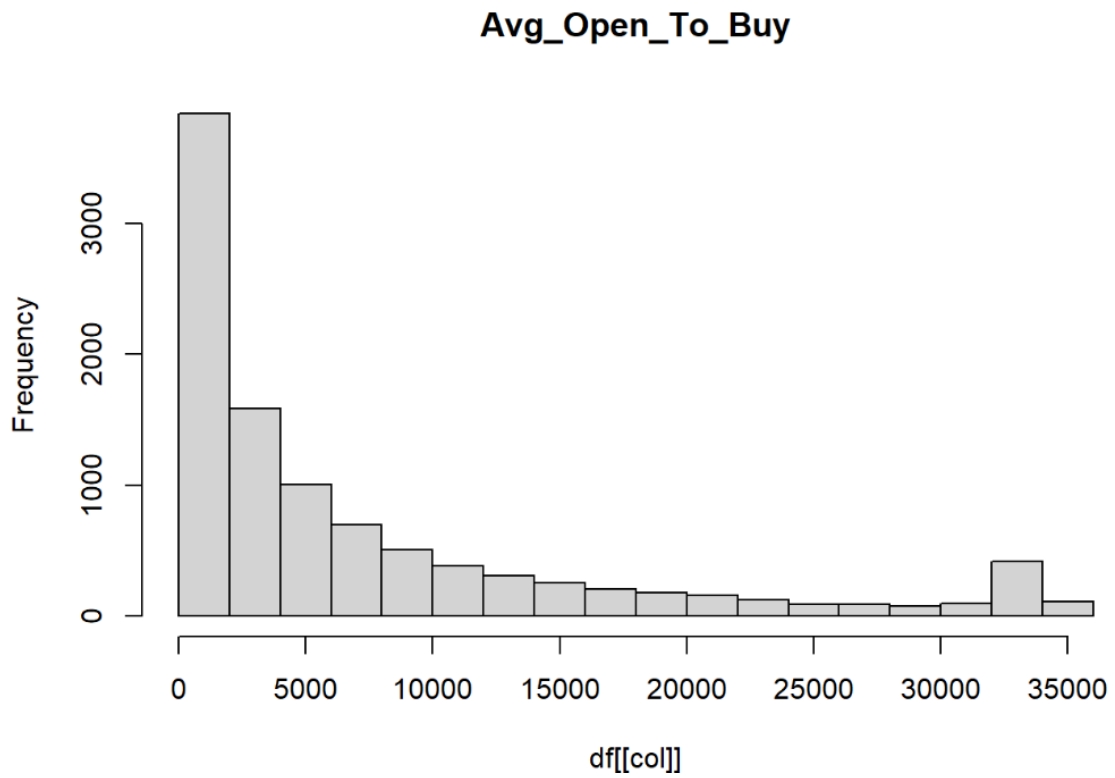
Frequency of credit card limit set by users:

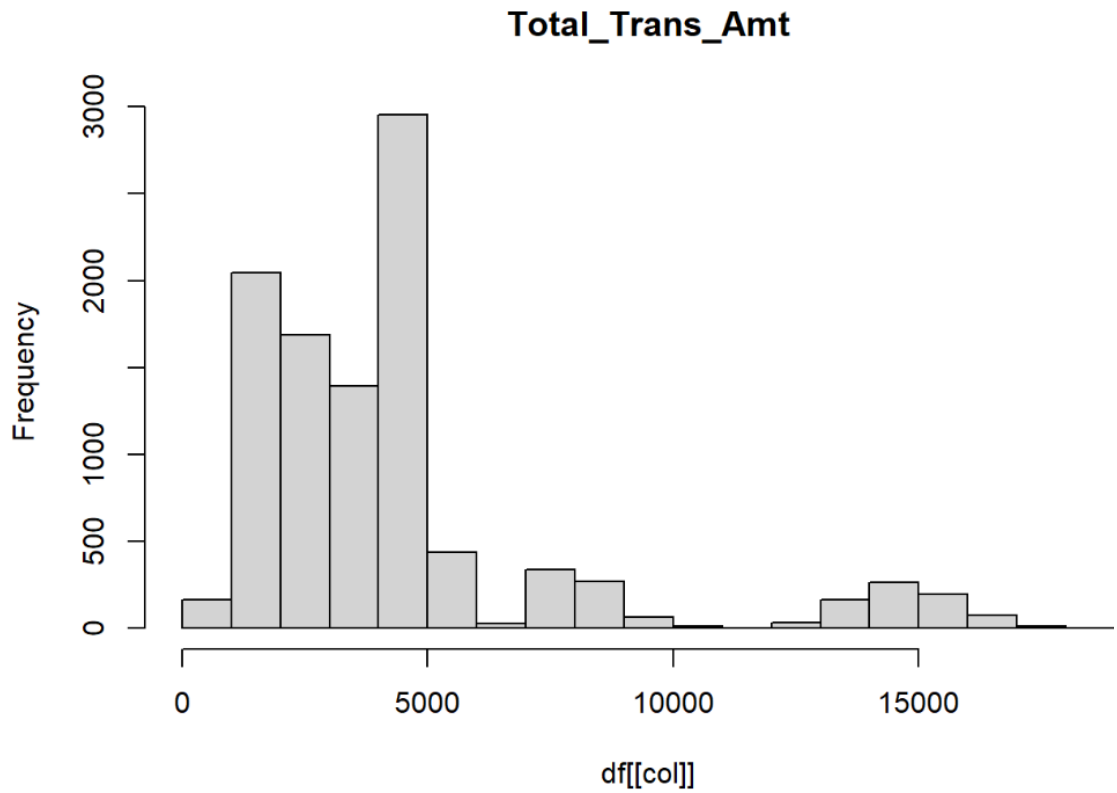


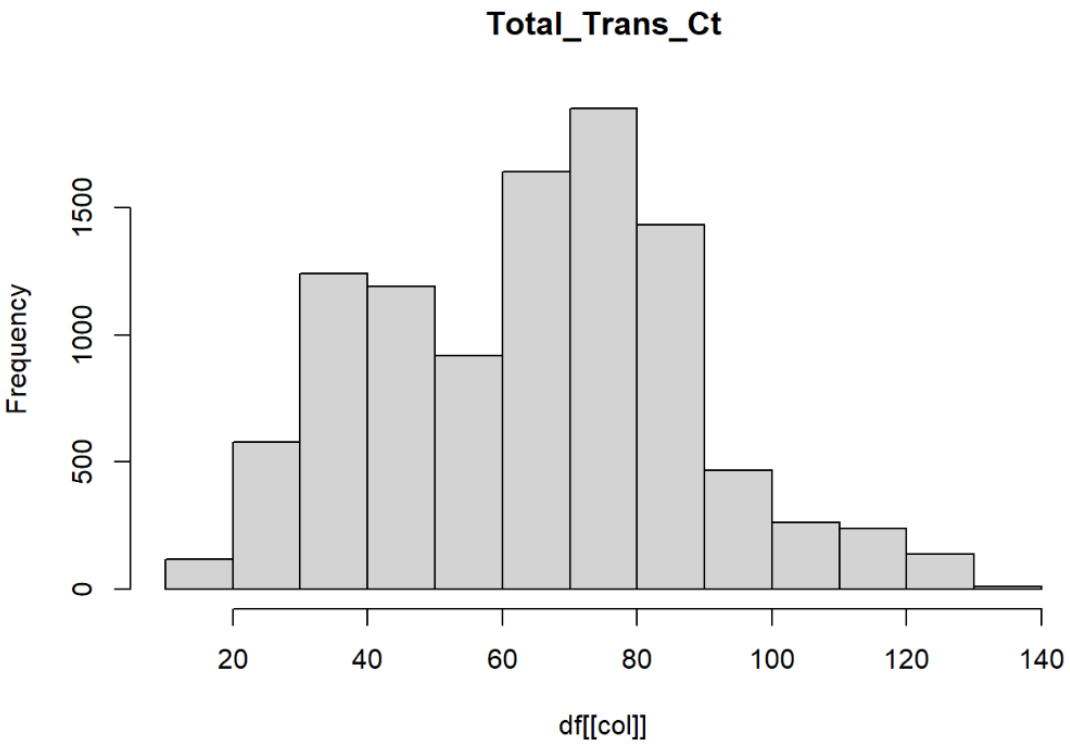

Total_Revolving_Bal
Frequency
0
500
1000
1500
2000
2500
df[[col]]

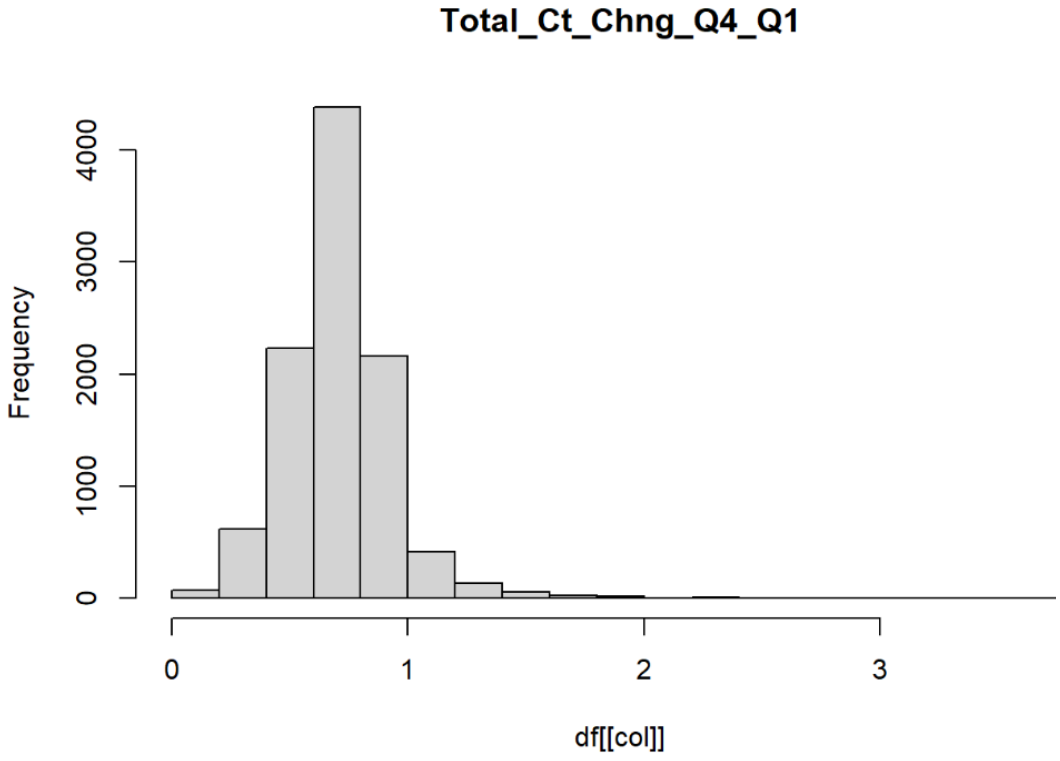

Total_Amt_Chng_Q4_Q1
Frequency
0
1000
2000
3000
4000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
df[[col]]

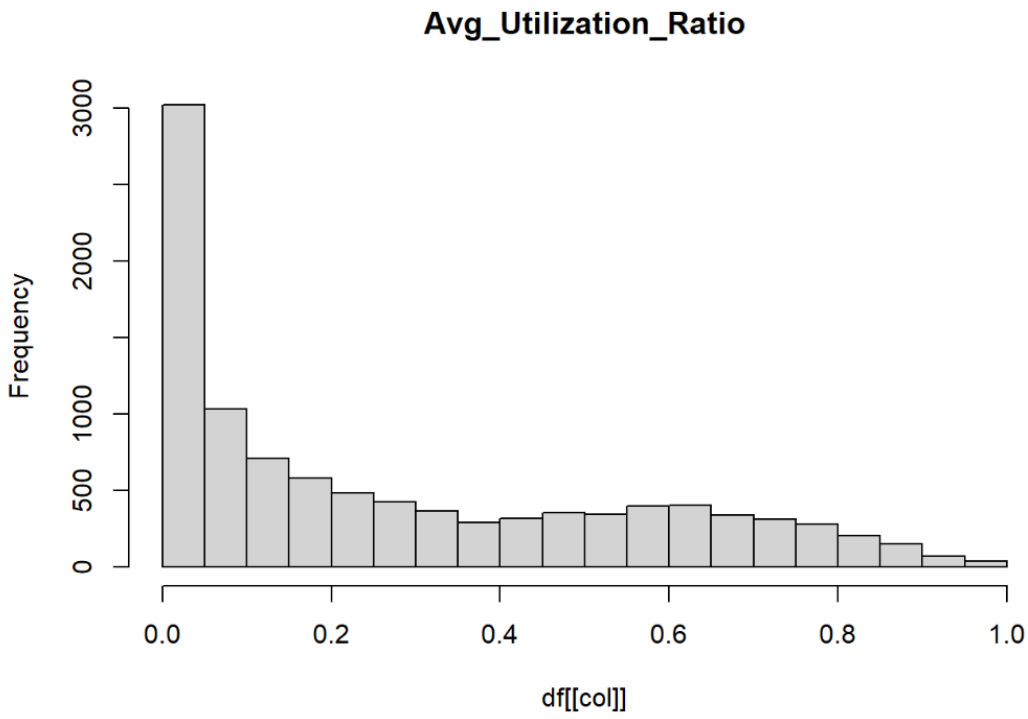
\

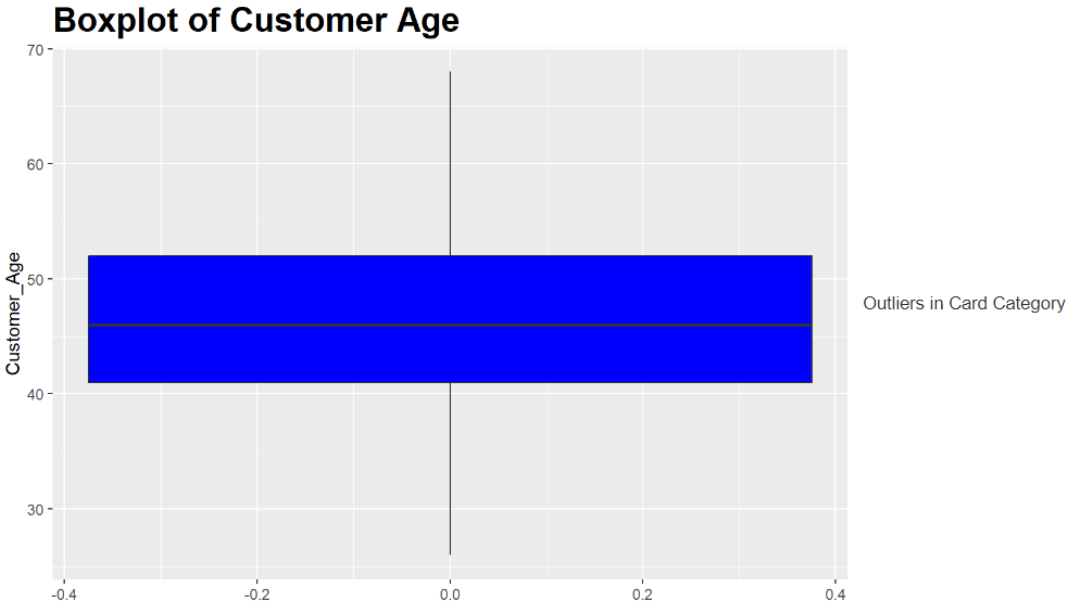

Avg_Open_To_Buy
Frequency
0
1000
2000
3000
0
5000
10000
15000
20000
25000
30000
35000
df[[col]]


Total_Trans_Amt
Frequency
0
500
1000
2000
3000
0
5000
10000
15000
df[[col]]

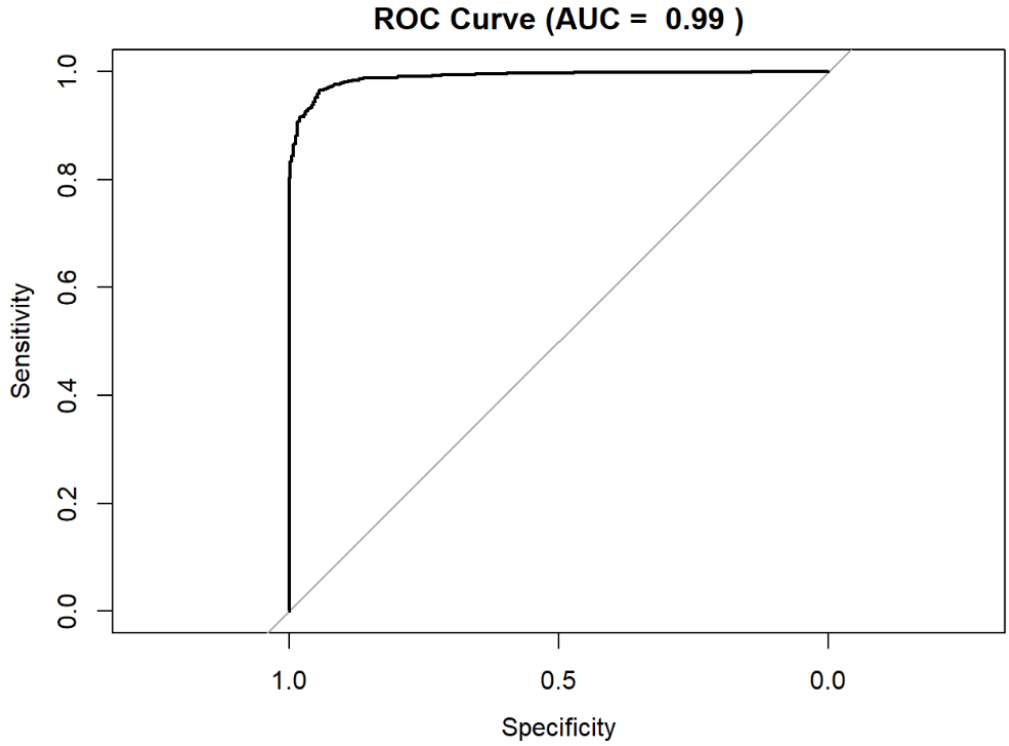

Total_Trans_Ct
Frequency
0
500
1000
1500
20
40
60
80
100
120
140
df[[col]]


Total_Ct_Chng_Q4_Q1
Frequency
0
1000
2000
3000
4000
0
1
2
3
df[[col]]

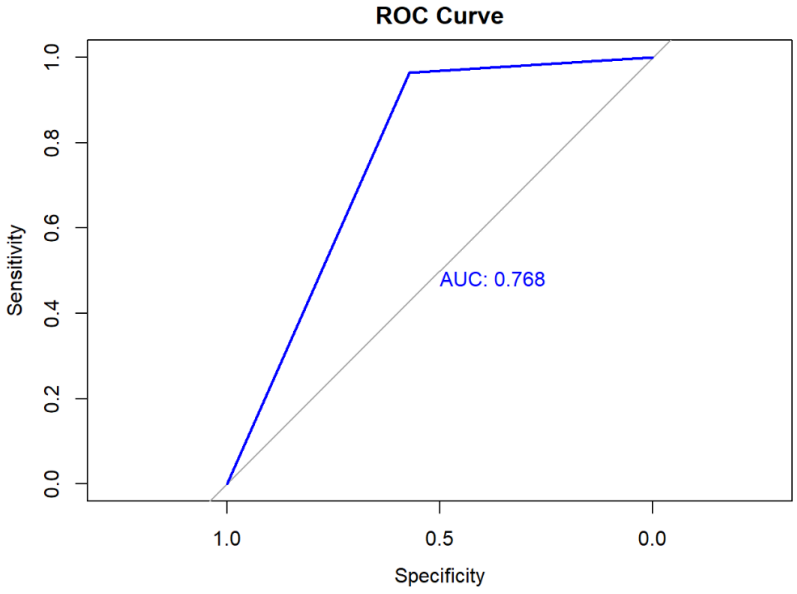

Avg_Utilization_Ratio
Frequency
0
500
1000
2000
3000
0.0
0.2
0.4
0.6
0.8
1.0
df[[col]]


Boxplot of Customer Age
Customer_Age
70
60
50
40
30
-0.4
-0.2
0.0
0.2
0.4
Outliers in Card Category

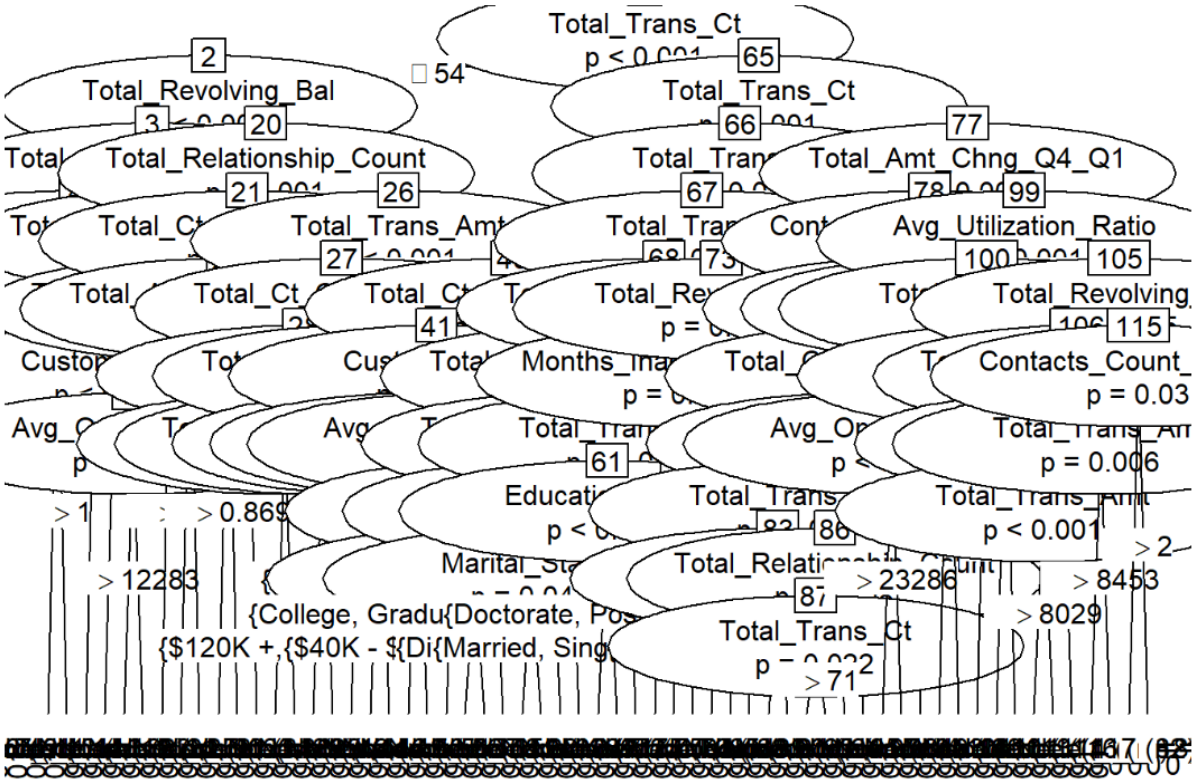
ROC curve for XGBOOST



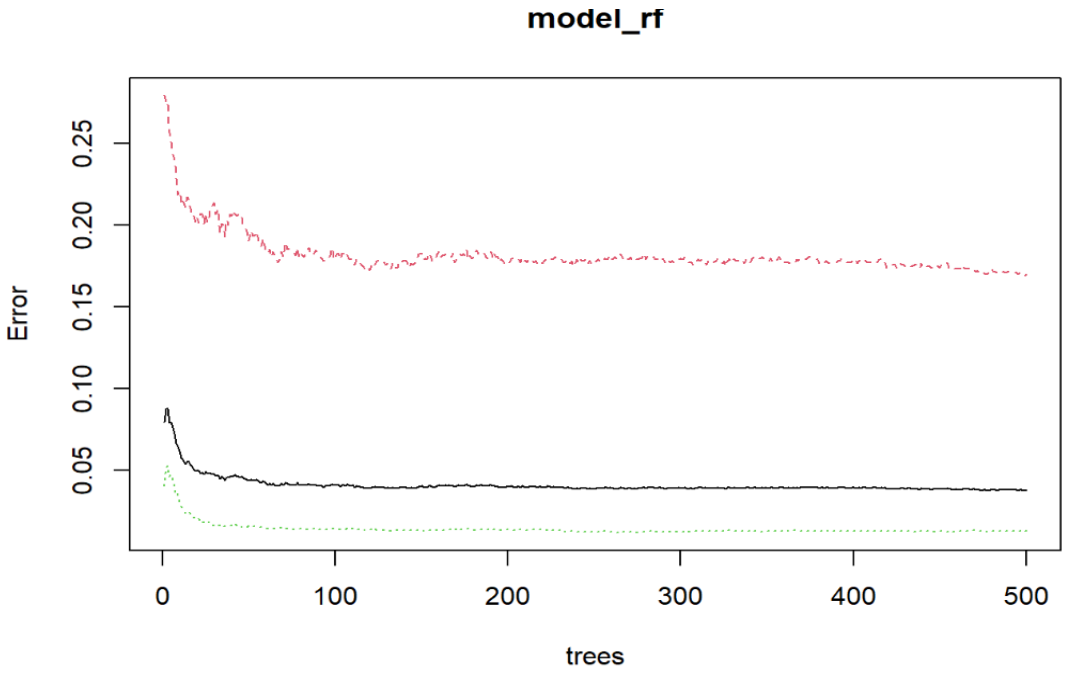
ROC curve for logistic regression



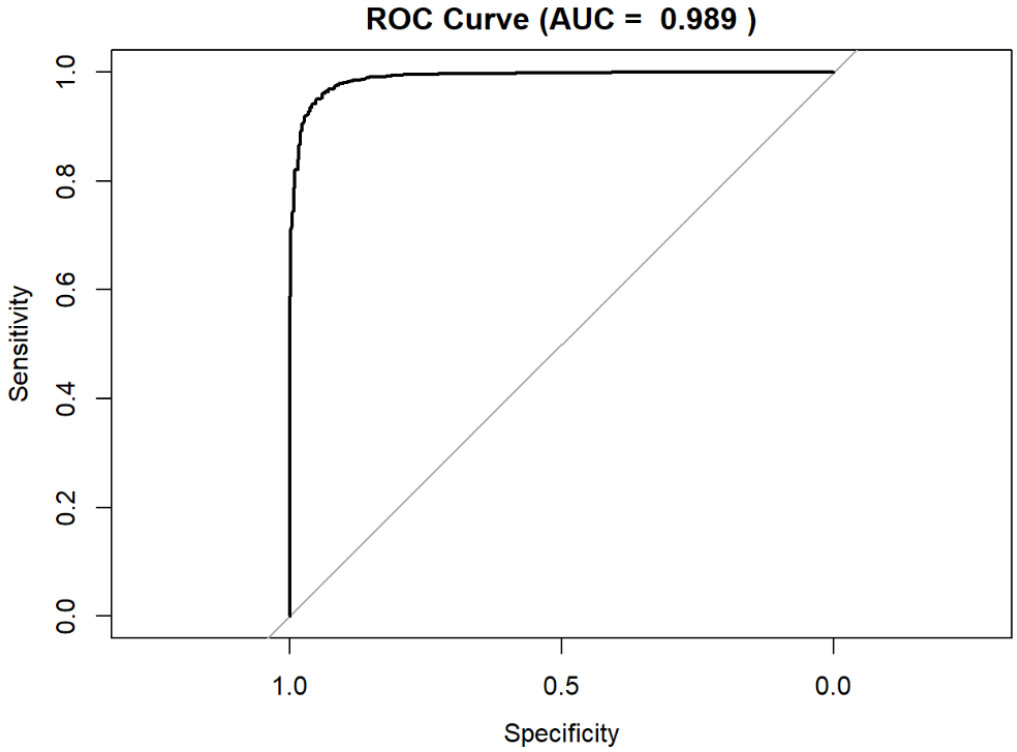
Decision tree:



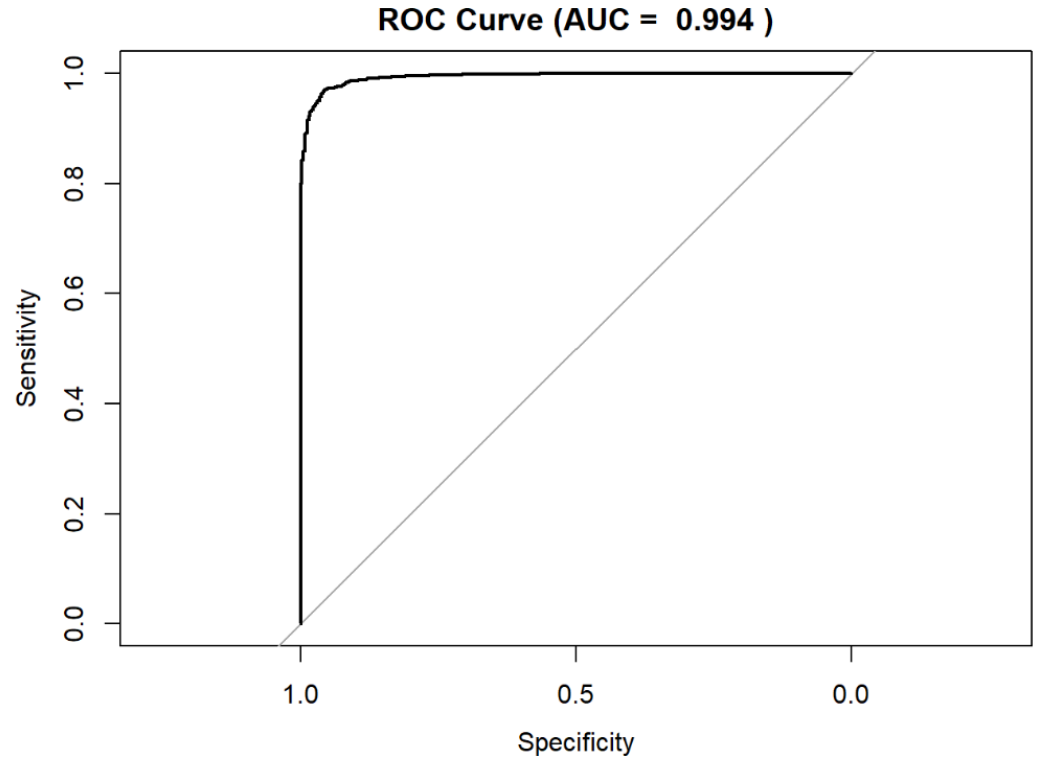
Random forest with number of trees vs. error



ROC curve random forest



ADABoost ROC curve



CONCLUSION AND FUTURE SCOPE

For credit card firms to increase their retention rates and profitability, credit card churn prediction is a crucial duty. In this project, a number of machine learning methods have been used to predict credit card churn, like Adaboost, XGBoost, Random Forest, and Linear Regression.

In summary, all four models have produced positive predictions for credit card churn. However, depending on the volume and complexity of the data as well as the characteristics of the features, their performance may differ. When compared to Adaboost and Linear Regression, XGBoost and Random Forest have shown to perform better.

There are various opportunities to enhance credit card churn prediction in terms of future scope. First, to further improve the precision of churn prediction models, researchers should look at the usage of more sophisticated ensemble methods, like Stacking. Second, adding more intricate features and variables, such credit score and payment history, might strengthen the models' robustness. Finally, when working with complicated, non-linear data, the use of deep learning techniques like Recurrent Neural Networks and Neural Networks can be used to increase the accuracy of churn prediction models.

In conclusion, although Adaboost, XGBoost, Random Forest, and Linear Regression have all demonstrated promising results in predicting credit card churn, there is still a great deal of room for future research to develop more precise and robust models to aid credit card companies in lowering customer churn rates and enhancing profitability.

REFERENCES

1. Nie, Guangli, et al. "Credit card churn forecasting by logistic regression and decision tree." Expert Systems with Applications 38.12 (2019): 15273-15285
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 17.4 (2022): 1529-1542
3. Wu, C., & Wang, L. (2022). A Comparative Analysis of Churn Prediction Models: A Case Study in Bank Credit Card. *Journal of Supply Chain and Operations Management*, *20*(2), 120.
4. Anil Kumar, Dudyala, and Vadlamani Ravi. "Predicting credit card customer churn in banks using data mining." International Journal of Data Analysis Techniques and Strategies 1.1 (2018): 4-28.
5. Sundarkumar, G. Ganesh, Vadlamani Ravi, and V. Siddeshwar. "One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection." 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). IEEE, 2015
6. Lin, Chiun-Sin, Gwo-Hshiung Tzeng, and Yang-Chieh Chin. "Combined rough set theory and flow network graph to predict customer churn in credit card accounts." Expert Systems with Applications 38.1 (2011): 8-15
7. Castro, Emiliano G., and Marcos SG Tsuzuki. "Churn prediction in online games using players' login records: A frequency analysis approach." IEEE Transactions on Computational Intelligence and AI in Games 7.3 (2015): 255-265
8. Jamal, Zainab, and Randolph E. Bucklin. "Improving the diagnosis and prediction of customer churn: A heterogeneous hazard modeling approach." Journal of Interactive Marketing 20.3‐4 (2018): 16-29
9. Konuksal, S. (2018). Credit card churn prediction with machine learning algorithms.

10. Dias, J., Godinho, P., & Torres, P. (2020). Machine learning for customer churn prediction in retail banking. In *Computational Science and Its Applications–ICCSA 2020: 20th International Conference, Cagliari, Italy, July 1–4, 2020, Proceedings, Part III 20* (pp. 576-589). Springer International Publishing.
11. Y. Kavyarshitha, V. Sandhya and M. Deepika, "Churn Prediction in Banking using ML with ANN," 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2022, pp. 1191-1198, doi: 10.1109/ICICCS53718.2022.9788456.
12. Rajamohamed, R., and J. Manokaran. "Improved credit card churn prediction based on rough clustering and supervised learning techniques." *Cluster Computing* 21.1 (2018): 65-77.
13. Sayed, Hend, Manal A. Abdel-Fattah, and Sherif Kholief. "Predicting potential banking customer churn using apache spark ML and MLlib packages: a comparative study." *International Journal of Advanced Computer Science and Applications* 9.11 (2018).
14. Doshi, Amrita. "Predicting Customer Churn in a Credit card company by applying ML and AutoML tools."
15. Dalmia, Hemlata, Ch VSS Nikil, and Sandeep Kumar. "Churning of bank customers using supervised learning." *Innovations in Electronics and Communication Engineering: Proceedings of the 8th ICIECE 2019*. Springer Singapore, 2020.