Credit Card Approval Prediction Machine Learning Models Report

Prepared by:

Adrian Vasques

Johnny Zhang

12/11/2023

ISDS 577 – Master of Science Capstone Seminar

Fall 2023

Prof. Daniel Soper

**Table Of Contents**

[**1. Executive Summary 2**](#_tmam9o4ryvfn)

[2.1 Data Source: 3](#_2t4nlcivh12d)

[2.2 Structure of Demographic Data: 3](#_hloa0j2gevws)

[2.3 Data Cleaning and Preparation: 5](#_iag5vl1jbb9m)

[2.3.1 Data Cleaning 5](#_nmpyoc73jxqj)

[2.3.2 Data Preparation 7](#_u0ngcfufexfb)

[2.4 Creating a Balanced Dataset 9](#_d391uph7urc4)

[2.5 Combining Categorical Variables 9](#_l5y3le4au7q7)

[2.6 Final Dataset 10](#_l5aiz7fwfs4t)

[**3. Analytical Techniques Used 13**](#_cj2juk5yk1st)

[3.1 Predicting Customer Delinquency: 13](#_7mxcf7widw4)

[3.1.1 Logistic Regression: 13](#_afda4hnv3609)

[3.1.2 Classification Trees: 13](#_9yzxzh9qyoki)

[3.2 Predicting Late Fee Generation: 14](#_bqgh54krhgp3)

[3.2.1 Logistic Regression: 14](#_3yqjl43hekv)

[3.2.2 Classification Trees: 14](#_e1yv0el68jgs)

[3.3 Choosing the Right Algorithm: 14](#_baj7hfc268ob)

[**4. Research Questions Answered 15**](#_7od45dj0pjtv)

[4.1 How do we decide what classifies as a bad customer based on credit history? 15](#_xlanu3ibaaix)

[4.2 How can we predict if customers will default on their loan/credit? 15](#_nm1k5knrkl3n)

[4.3.1 Logistic Regression: 17](#_kkm8bfneqk24)

[4.3.2 Classification Trees: 17](#_qqgp6yerfqbv)

[**5. How will answering this research question inform managerial decision-making? 23**](#_4tcsmx8wzzq2)

[**6. Tables and Charts 25**](#_l09djfl5m6qw)

[6.1 XG Boost Variable importance 25](#_8yhpev8c333y)

[6.2 Models and Charts for Predicting a Delinquent Customer 26](#_ryonu8wyvinr)

[**7. References 31**](#_qq3ukihyxgzu)

# **1. Executive Summary**

The main objective of this project is to help management identify customers who are likely to pay back their loans and to determine the customer groups that generate the most profits based on their credit history and demographic information. The analysis will start with descriptive statistics to gain a comprehensive overview of the data. After that, the data will be cleaned and prepared to ensure its accuracy and suitability for analysis. The project will turn the data into actionable insights using supervised machine learning models, allowing management to fine-tune their strategies and increase profitability.

The following are the research questions we hope to answer in this report:

1. How can we predict if customers will default on their loan/credit?
2. How will we determine what classifies as a bad customer based on credit history?
3. What models are best at predicting if a customer will default？
4. What variables are most significant in predicting whether a consumer will default?
5. What variables will predict if a customer is past due for more than 180 days and will eventually be charged off?
6. How can our analysis increase profitability?

**2. Descriptive Statistics and Data Preparation**

## 2.1 Data Source:

We have a collection of 2 CSV files combined into a single dataset. The first dataset has information about customer applications, including age, gender, car and home ownership, number of children, income, marital status, housing, arrangement, education level, and occupation. The second dataset contains credit histories and the application status of each customer.

Both datasets are identified using customer IDs so we can merge the data vertically. However, we face challenges, such as creating a binary decision variable and feature engineering, since there is no decision variable.

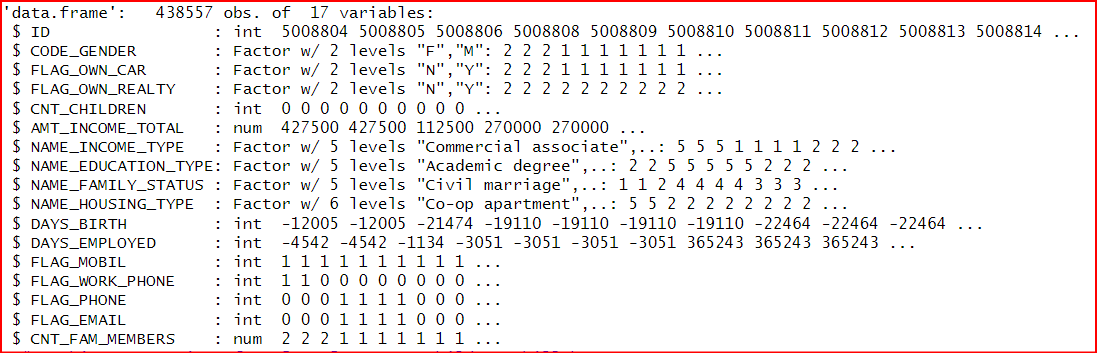
We attained the data from the Kaggle link below:

<https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction?resource=download>

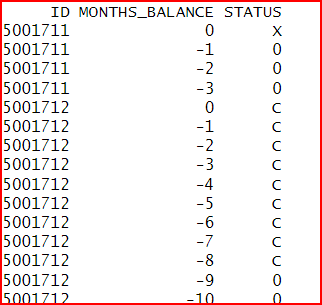
## 2.2 Structure of Demographic Data:

Below is the Structure of the Demographic Data, which consists of 438,557 observations of 17 variables. The first structure contains the application data, and the next has the credit card holder's records.

Demographic Data:



Sample Credit History Data



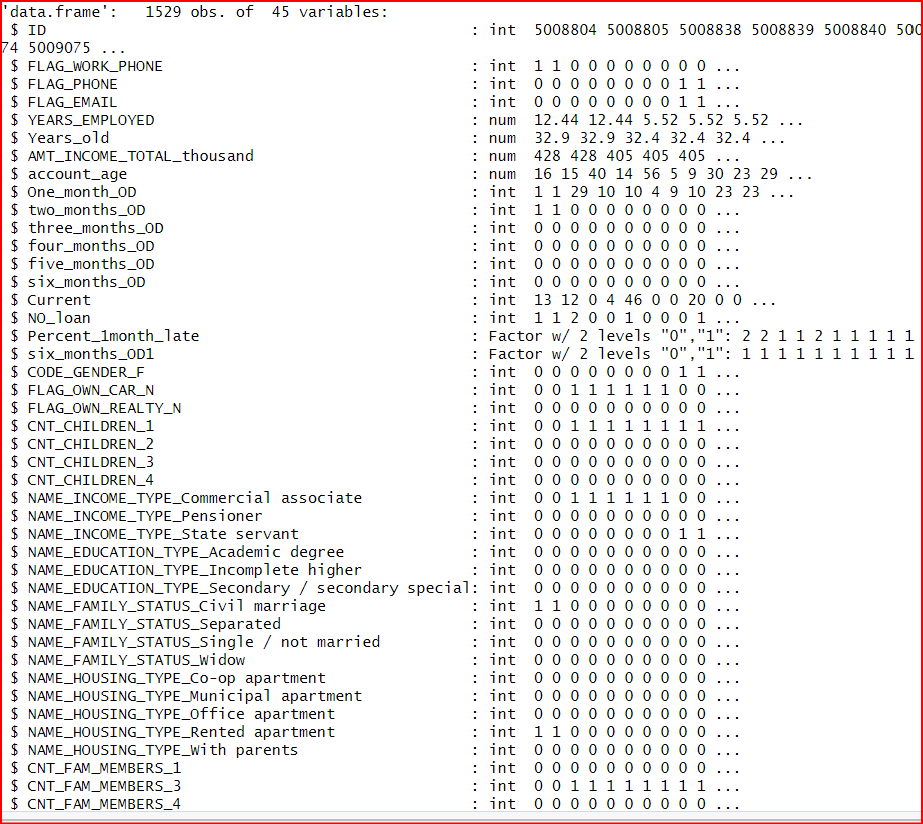
## 2.3 Data Cleaning and Preparation:

Our dataset comprises multiple files, namely 2 CSV, amalgamated into a single comprehensive dataset. The primary dataset entails customer applications and provides an array of information, including age, gender, car, home ownership, number of children, income, marital status, housing arrangement, education level, and occupation. The secondary dataset, on the other hand, encompasses each customer's credit history and application status. Since both datasets utilize customer IDs, we can vertically merge the data. However, we are met with certain challenges, which include devising a binary decision variable and feature engineering due to the absence of such a variable.

### 2.3.1 Data Cleaning

Our cleaning tasks included:

* **Handling missing values:** There was no missing data**.**
* **Checking for duplicates:** Our Data did not have traditional duplicate data but had varying records between demographic and credit history data. Refer to preparation for more information.
* **Transforming categorical variables into numerical or dummy variables:** Out of 19 original dimensions, we had 4 continuous variables, so 15 variables needed to be transformed into dummy variables. After transformation, we had 45 dimensions.
* **Removed Attributes**: We removed the mobile phones attribute because all records reported having a mobile phone, so it offers no added value to the analysis. Also, “Occupation\_type '' had over 30% missing variables, so we removed that attribute altogether.
* **Create Dummy Variables:** Since most machine learning models can analyze using words, we had to turn string data into binary classifiers using 1s and 0s.

****

* **Removing outliers:** The only variable that had Outliers was “amount\_of\_Income\_in thousands.” using a boxplot and the statistical definition of an outlier, we decided to remove any values greater than 380 thousand (using q3 + 1.5 IQR= 380250)

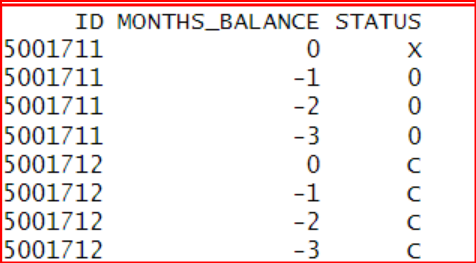
### 2.3.2 Data Preparation

Feature Engineering: We did Feature Engineering for 2 features to accomplish our goals and answer our research questions.

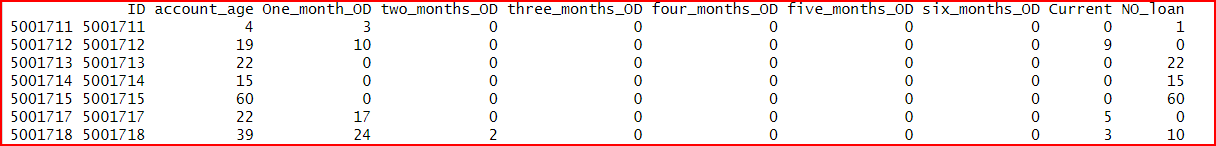
**Feature 1:** The data we received needed a decision variable, which posed a challenge in using most of the supervised classification methods available. However, we identified the issue with the data: the inability to merge it without losing important information or creating unnecessary duplicates. For instance, the customer credit history data contained multiple records for each customer ID, representing one month of credit history. We pinpointed this issue and avoided merging the credit history and demographic data, which would have resulted in missing and duplicate values. Additionally, we needed to create a variable that could effectively distinguish between good and bad customers. According to Business Insider, after conducting thorough research, customers are considered to have defaulted on their credit if they have been overdue by six months or more. To tackle this, we pivoted the data wider by creating additional columns and aggregating the number of months each customer was delinquent, from one to six months. Overall, we were confident in our approach to handling the data and were able to overcome the challenges it posed.

Below is a before-and-after picture of our data.

Before Feature Engineering:



After Feature Engineering



**Feature 2:** To develop a reliable model for customer classification, it is important to have a variable that identifies the most profitable customers. However, we had limited customer payment history information in our case, making it challenging to create a balanced dataset. We devised a variable that predicts future profitability based on the available data to address this issue. This involved analyzing customer payment history and identifying customers who frequently incur late fees but are not delinquent.

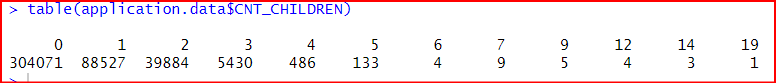
To achieve this, we created a dummy variable representing customers who were late paying their bills by 1-3 months but never became delinquent. We used customer credit history to calculate the proportion of missed payments to total history length and defined a function to identify customers meeting the late payment criteria. Finally, we classified customers as "1" (potential for generating late fees) or "0" (no late fees) based on this variable.

## **2.4 Creating a Balanced Dataset**

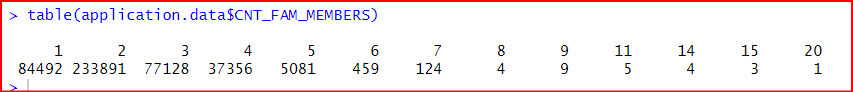
Our team faced a significant challenge with our data set, which had a severe imbalance. Out of 36,456 records, only 180 were identified as delinquent. This data imbalance can cause issues with machine learning models, leading to biased predictions, inaccurate evaluations, and limited generalizability. We had to balance our data before deploying our models to address this issue. We used two approaches to do this: undersampling and oversampling. For undersampling, we randomly selected a sample of 180 records from the majority class, leaving us with a data set of 360 records. For oversampling, we randomly selected multiple samples from the minority class until we had the same number of records as the majority class. In the end, we had 36277 records classified as delinquent and 36277 classified as good customers.

## 

## **2.5 Combining Categorical Variables**

Lastly, we combined categories with too many levels with few observations—the variable “CNT\_children'' which recorded the number of children per customer. We chose to combine all records that reported over 6 children because they had few records for each class. The table below supports this claim.

In Addition, the attribute “CNT\_Family\_Members' ' also had 20 levels. However, only a few records had more than 7 family members, so we combined all records that reported having more than 7 family members into a single level. We chose to do this because most records had 7 or fewer family members, and we believed the higher number of family members would skew the data.



## **2.6 Final Dataset**

Our final data set has 36457 obs. of 47 variables:

|  |  |
| --- | --- |
| Filed Name | Data Type |
| $ FLAG\_WORK\_PHONE | int |
| $ FLAG\_PHONE | int |
| $ FLAG\_EMAIL | int |
| $ YEARS\_EMPLOYED | num |
| $ Years\_old | num |
| $ AMT\_INCOME\_TOTAL\_thousand | num |
| $ account\_age | num |
| $ One\_month\_OD | int |
| $ two\_months\_OD | int |
| $ three\_months\_OD | int |
| $ four\_months\_OD | int |
| $ five\_months\_OD | int |
| $ six\_months\_OD | int |
| $ Current | int |
| $ NO\_loan | int |
| $ Percent\_1month\_late | Factor w/ 2 levels "0","1" |
| $ six\_months\_OD1 | Factor w/ 2 levels "0","1" |
| $ CODE\_GENDER\_M | int |
| $ FLAG\_OWN\_CAR\_Y | int |
| $ FLAG\_OWN\_REALTY\_N | int |
| $ CNT\_CHILDREN\_1 | int |
| $ CNT\_CHILDREN\_2 | int |
| $ CNT\_CHILDREN\_3 | int |
| $ CNT\_CHILDREN\_4 | int |
| $ CNT\_CHILDREN\_fiveOrMore | int |
| $ NAME\_INCOME\_TYPE\_Commercial associate | int |
| $ NAME\_INCOME\_TYPE\_Pensioner | int |
| $ NAME\_INCOME\_TYPE\_State servant | int |
| $ NAME\_INCOME\_TYPE\_Student | int |
| $ NAME\_EDUCATION\_TYPE\_Academic degree | int |
| $ NAME\_EDUCATION\_TYPE\_Higher education | int |
| $ NAME\_EDUCATION\_TYPE\_Incomplete higher | int |
| $ NAME\_EDUCATION\_TYPE\_Lower secondary | int |
| $ NAME\_FAMILY\_STATUS\_Civil marriage | int |
| $ NAME\_FAMILY\_STATUS\_Separated | int |
| $ NAME\_FAMILY\_STATUS\_Single / not married | int |
| $ NAME\_FAMILY\_STATUS\_Widow | int |
| $ NAME\_HOUSING\_TYPE\_Co-op apartment | int |
| $ NAME\_HOUSING\_TYPE\_Municipal apartment | int |
| $ NAME\_HOUSING\_TYPE\_Office apartment | int |
| $ NAME\_HOUSING\_TYPE\_Rented apartment | int |
| $ NAME\_HOUSING\_TYPE\_With parents | int |
| $ CNT\_FAM\_MEMBERS\_1 | int |
| $ CNT\_FAM\_MEMBERS\_3 | int |
| $ CNT\_FAM\_MEMBERS\_4 | int |
| $ CNT\_FAM\_MEMBERS\_5 | int |
| $ CNT\_FAM\_MEMBERS\_sixOrMore | int |
| $ CNT\_FAM\_MEMBERS\_sixOrMore | int |

# **3. Analytical Techniques Used**

To answer our research question, we use Logistic regression and classification trees to predict if a customer is delinquent and if our customers generate late fees.

## **3.1 Predicting Customer Delinquency:**

### **3.1.1 Logistic Regression:**

Logistic regression is a statistical method that estimates the probability of a binary outcome, making it well-suited for predicting customer delinquency. It analyzes the relationship between various input features, such as credit score, payment history, and debt-to-income ratio, and the target variable, which in this case is whether a customer will become delinquent.

### **3.1.2 Classification Trees:**

Classification trees are decision-tree models that categorize data based on rules and splits. They can effectively predict customer delinquency by constructing a tree structure that progressively divides the data into smaller subgroups based on relevant credit history and demographic factors.

## **3.2 Predicting Late Fee Generation:**

### **3.2.1 Logistic Regression:**

Similar to predicting delinquency, logistic regression can estimate the probability of a customer generating late fees. It analyzes the relationship between factors such as payment history, credit utilization, and income stability with the target variable, whether a customer will incur late fees.

### **3.2.2 Classification Trees:**

Classification trees can also predict late fee generation by constructing a tree structure that categorizes customers based on their likelihood of generating late fees. The tree's branches represent decision rules based on payment behavior, creditworthiness, and financial habits.

## **3.3 Choosing the Right Algorithm:**

The choice between logistic regression and classification trees depends on specific requirements and data characteristics. Logistic regression is often preferred when interpretability is crucial, while classification trees may be preferred when accuracy is the primary concern. However, it is advisable to experiment with both algorithms to determine the best fit for the specific data and application.

# **4. Research Questions Answered**

## **4.1 How do we decide what classifies as a bad customer based on credit history?**

The data we received needed to indicate which customers would likely default on their payments. Therefore, we had to devise a solution and create a new feature to identify customers who were in default based on the available data. We followed the industry standard of considering customers delinquent if they were late for six consecutive months. Using this criterion, we created a binary attribute assigned a value of "1" for delinquent customers and "0" for those who were not.

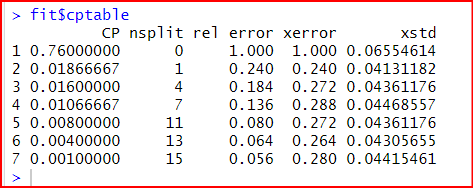
## **4.2 How can we predict if customers will default on their loan/credit?**

Binary classification is a common task in machine learning, and various algorithms can be used to achieve this goal. Some of the most popular algorithms for binary classification include:

* Logistic regression: This is a simple and effective algorithm that works well for problems with a linear relationship between the features and the target variable.
* Support vector machines (SVMs): SVMs are a more complex algorithm that can handle non-linear relationships between the features and the target variable.
* Decision trees: Decision trees are easy to interpret and can handle linear and non-linear relationships between the features and the target variable.
* Random forests: Random forests are an ensemble of decision trees, meaning they combine the predictions of multiple decision trees to make a more accurate prediction.
* Neural networks: Neural networks are a powerful machine learning model that can be used to solve various problems, including binary classification.

The best algorithm for a particular problem will depend on the specific characteristics of the data and the accuracy and error of the various models.

We answered our research question using supervised machine learning models such as Decision trees, Logistic regression, and Random Forest. From there, we will measure the accuracy based on the percentage of accurate and sensitive records, specificity and special specificity, and false negative rate for delinquent customers. We want a low false negative rate because it is important to find all instances where a customer will default on a loan. Also, having a low false negative rate will help the corporation not to lose money because it will enable them to not classify someone as a good customer who is a customer that will default.



**4.3 What variables are most significant in predicting whether a consumer will default 180 days?**

**Methods Used to Identifying Important Variables:**

Both logistic regression and classification trees provide insights into the relative importance of input features in predicting delinquency and late fee generation. Below are the techniques and measures we will use to evaluate var importance.

### **4.3.1 Logistic Regression:**

1. **Coefficient Magnitude:** The magnitudes of the logistic regression coefficients indicate the relative strength of each feature's impact on the probability of delinquency or late fee generation. Larger coefficients represent a stronger influence.
2. **Odd Ratio:** The larger the odds ratio, the more it contributes to whether a customer will default.
3. **P-Value**: If the P value is smaller than the alpha of .05, it is statistically significant in predicting a particular outcome.

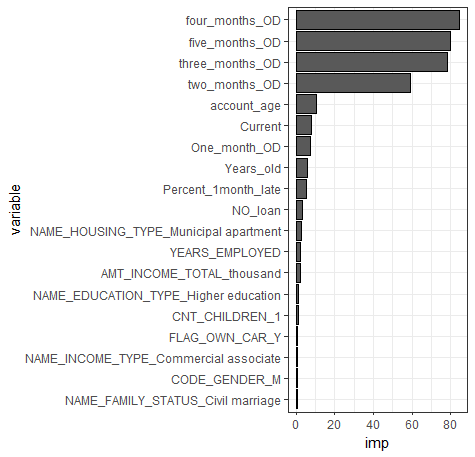
### **4.3.2 Classification Trees:**

1. **Split Frequency:** The frequency with which a feature is used to split the data in the classification tree reflects its importance in predicting delinquency or late fee generation. Features used in higher-level splits are considered more influential.
2. **Mean Decrease in Impurity:** MDI measures the average impurity reduction across all splits where a given variable is used. Impurity is a measure of the diversity of the classes within a node. A lower impurity value indicates that the node is pure and that the variable used to split the node is more important. The bigger the MDI, the more important the variable.

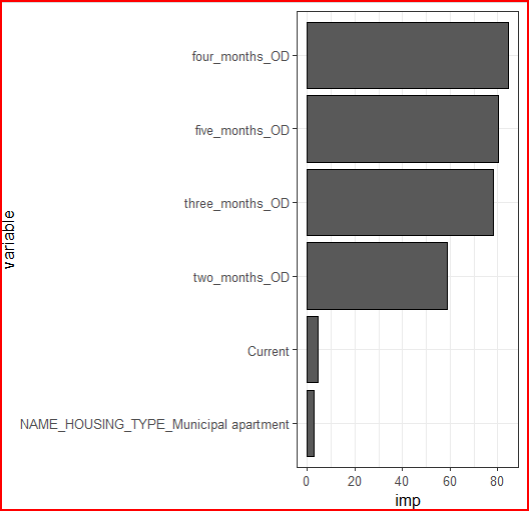
By analyzing these measures, you can identify the most critical variables contributing to customer delinquency and late fee generation. This information can be valuable for targeted interventions and risk mitigation strategies.

Below are charts that display what variables are most important based on MDI. The Higher the MDI, the better the variable. Please refer to the charts below for more information:

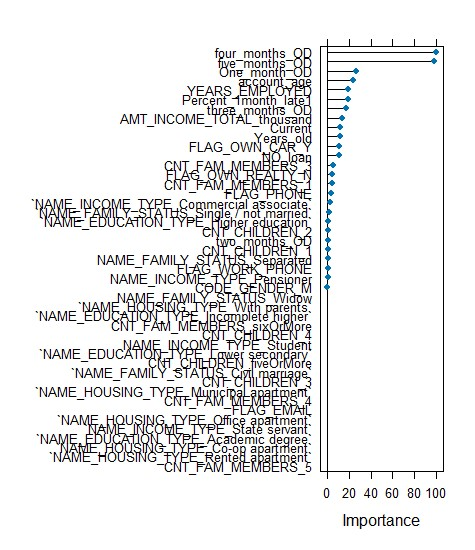
**Best Pruned Tree**



**Min Error**

****

**XG Boost**

****

Overall, each model had mentioned Four\_Months\_od, Five\_Months\_od, and One\_Months\_od as important variables. Also, each model had different demographic data, so we used the XG Boost results to tell us which variables were most important. According to XG Boost, the 5 most important variables were:

**Four\_Months\_od:** This is the most important variable that reduces the most error and tells us that if a customer is 4 months over, they will likely become delinquent for 6 months.

**Five\_Months\_od:** This important variable also reduces error compared to the other variables. It tells us that if a customer is 5 months overdue, they are very likely to become delinquent for 6 months.

**One\_Months\_od:** This variable tells us that if you miss a payment, you have a higher chance of becoming delinquent, But compared to the first 2 variables, it does not have such a substantial impact as four and five months overdue.

**Account Ages:** According to our chart, account age is in several splits of XG Boost Tree and reduces accuracy significantly. This makes sense because a customer with a long credit history is reliable, especially if one must make payments to avoid defaulting and having bad credit. When that happens, no credit card company will offer them credit. Thus, customers with a small account history are usually you and do not have a history, or they have had bad credit and were denied credit.

**Years Employed:** Records that had much work experience seemed to pay off their bill and did not default. However, customers who did not work or had a long employment history were more likely to default. This makes sense because if you cannot hold a job, you will probably not have the money or responsibility to maintain your credit card payments.

**4.4 Which model is best at predicting a default customer**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model: (delinquent) | Accuracy | Sensitivity | specificity | False Negative |
| Logistic regression  forward | .8222 | 1 | .709 | 0 |
| XG Boost | .90 | .94 | .87 | .06 |
| Min error Decision Tree | .8222 | 1 | .709 | 0 |
| Best Pruned | .8333 | .9429 | .7636 | .0571 |

We conducted a study where we trained three different models, namely Decision trees, Logistic regression, and XG BOOST Decision trees. We used a training dataset to determine which model performed the best and then evaluated the model’s predictive ability using a separate test data set. Our evaluation criteria included accuracy, specificity, false negative rate, and error rate. We were particularly interested in identifying false negative rates since we wanted to identify all customers who would default on their loans.

Our analysis found that the XGBOOST model performed the best with an accuracy rate of 90. However, despite our efforts, none of the models we created could beat McNemar's Test, suggesting that the models were ineffective. Our model failed to

predict the major class, which makes it unsuitable for prediction purposes. Nonetheless, we were able to gather valuable insights on variables that were important, and we plan to investigate them further.

# **5. How will answering this research question inform managerial decision-making?**

**Delinquency Prediction:**

Our analysis reveals a 90% accuracy rate in predicting customer delinquency, though it does not outperform the 99.5% "no information" baseline. However, valuable insights can still be gleaned.

**Targeted Monitoring and Communication:**

Management can implement targeted interventions by identifying customers with a high delinquency risk, especially those who are late for at least three consecutive months. This could involve increased monitoring, proactive outreach, or personalized payment plans.

**Refined Customer Acquisition Strategies**:

Analyzing demographic data associated with delinquency risk can help identify customer profiles prone to default. This information can then be utilized to refine acquisition strategies and adjust loan terms for specific demographics.

**Payment History: The Key Predictor:**

All models consistently identified payment history as the most critical variable in predicting delinquency. This emphasizes the importance of closely monitoring payment patterns and promptly addressing delays.

**Flag for Potential Defaults:**

Identifying the 66.7% delinquency rate for customers late for four consecutive months provides a valuable benchmark. This data can be used to flag potentially high-risk customers for immediate intervention and prevent future defaults.

**Continuous Improvement:**

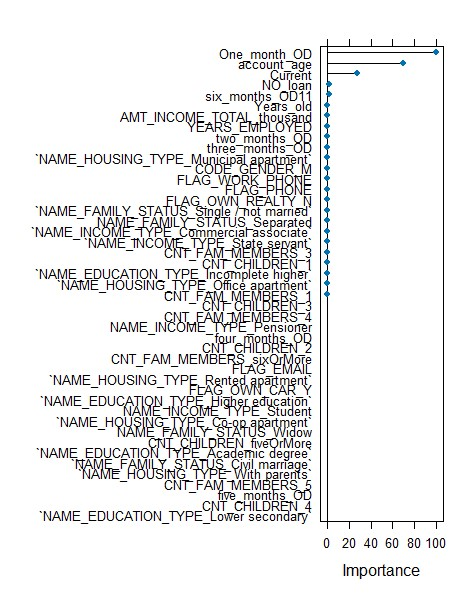
While the current model provides valuable insights, exploring additional data sources beyond payment history can improve prediction accuracy. This could include factors like employment status, income fluctuations, or credit history.

Even though the prediction accuracy is imperfect, this analysis offers valuable information to help management proactively manage customer delinquency risk and improve financial performance. Organizations can mitigate potential losses and ensure long-term financial stability by implementing the suggested strategies, such as targeted monitoring, refined acquisition practices, and close attention to payment history.

# **6. Tables and Charts**

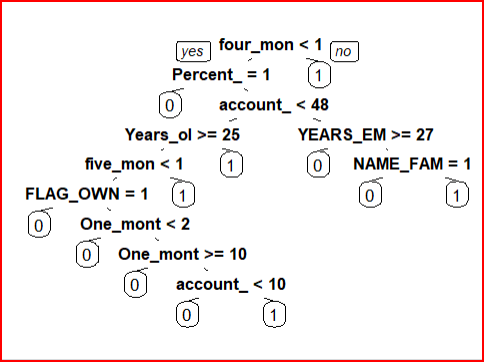
## **6.1 XG Boost Variable importance**

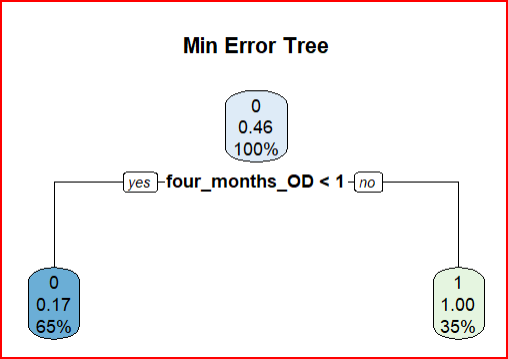
|  |  |  |  |
| --- | --- | --- | --- |
| **Late Fee Models** | Accuracy | Sensitivity | Specificity |
| Cart full | .9182 | .9311 | .8969 |
| Xg boost | .8889 | .9730 | .5 |
| Best pruned | .9182 | .9311 | .8969 |
| Min error | .8575 | .8567 | .8588 |
| Log forward | .6417 | .9203 | .1825 |

****

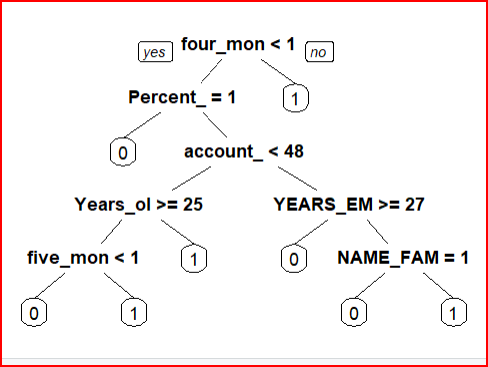
## **6.2 Models and Charts for Predicting a Delinquent Customer**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model: delinquent | Accuracy | Sensitivity | specificity | False Negative |
| Logistic regression  forward | .8222 | 1 | .709 | 0 |
| XG Boost | .90 | .94 | .87 | .06 |
| Min error Decision Tree | .8222 | 1 | .709 | 0 |
| Best Pruned | .8333 | .9429 | .7636 | .0571 |

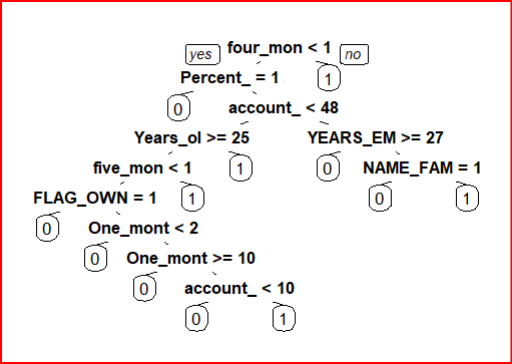


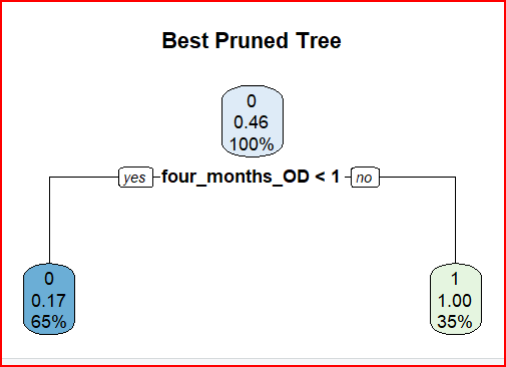


**Best Pruned Tree**



**Full tree**

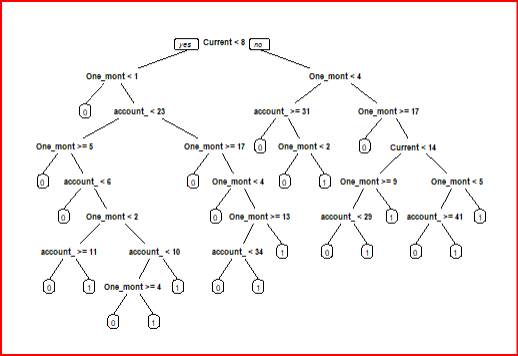




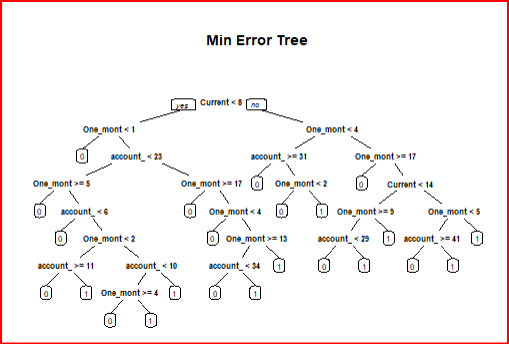
**6.3 Models that Predict Late Fees**

|  |  |  |  |
| --- | --- | --- | --- |
| **Late Fee Models** | Accuracy | Sensitivity | Specificity |
| Cart full | .9182 | .9311 | .8969 |
| Xg boost | .8889 | .9730 | .5 |
| Best pruned | .9217 | .9370 | .8968 |
| Min error | .8575 | .8567 | .8588 |
| Log forward | .6417 | .9203 | .1825 |

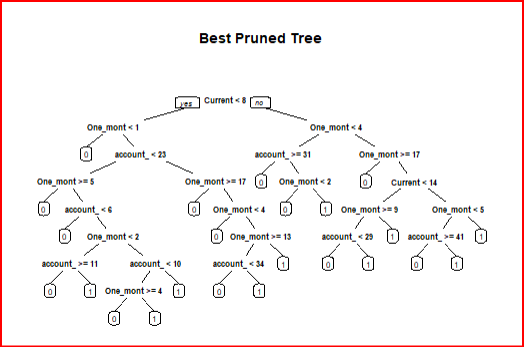
**Full Tree**

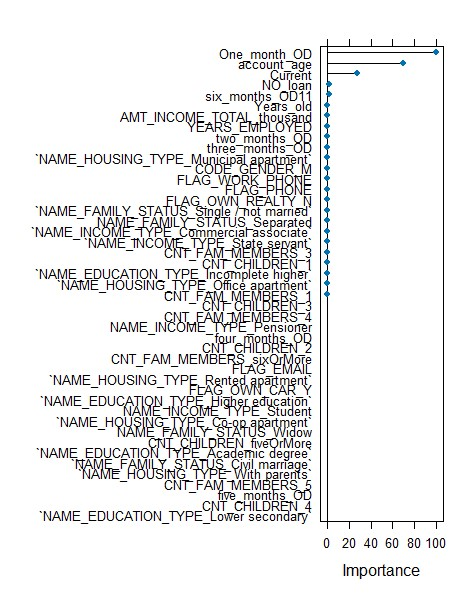


**Min Error Tree**



**Best Pruned Tree**



**XG BOOST Variable Importance Late Fees**.****

# 

# **7. References**

Bruce, P. C., Shmueli, G., Gedeck, P., & Others. (2018). *Machine Learning for Business Analytics: Concepts, Techniques, and Applications in R 2nd Edition*. Wiley.

# 

# 

# 