
# Credit Card Approval Prediction

Group 3:
Adrian Vasques
Johnny Zhang
Akhil Nair

# Problem Statement

Credit score cards are a common risk control method in the financial industry

It uses personal information and data submitted by credit card applicants to predict the probability of future defaults and credit card borrowings.

Build a machine learning model to predict if an applicant is 'good' or 'bad' client

# Research Questions

- How can we predict if customers will default on their loan/credit?
- How to predict if a customer will be profitable based on?
- How will we determine what classifies as a bad customer based on credit history?
- What variables are most significant in predicting whether a consumer will default?
- What variables will predict if a customer is past due for more than 180 days and will eventually be charged off?

# Data Source

- Kaggle
- application_record.csv contains appliers personal information
- credit_record.csv records users' behaviors of credit card.
- Connected by Customer ID

# Definition of 'good' or 'bad' Customer

**Percent_Late:** create a variable a binary attribute that helps us identify what customers were most profitable by means of late free.

**Method** We  aggregated the total credit history in months, and divided it by the total amount of months the record was late for 1-2 months

**Criteria**: record that had  lat no more than 40 percent of the time but were never late for for more than 3 months in a row

# Descriptive Analysis

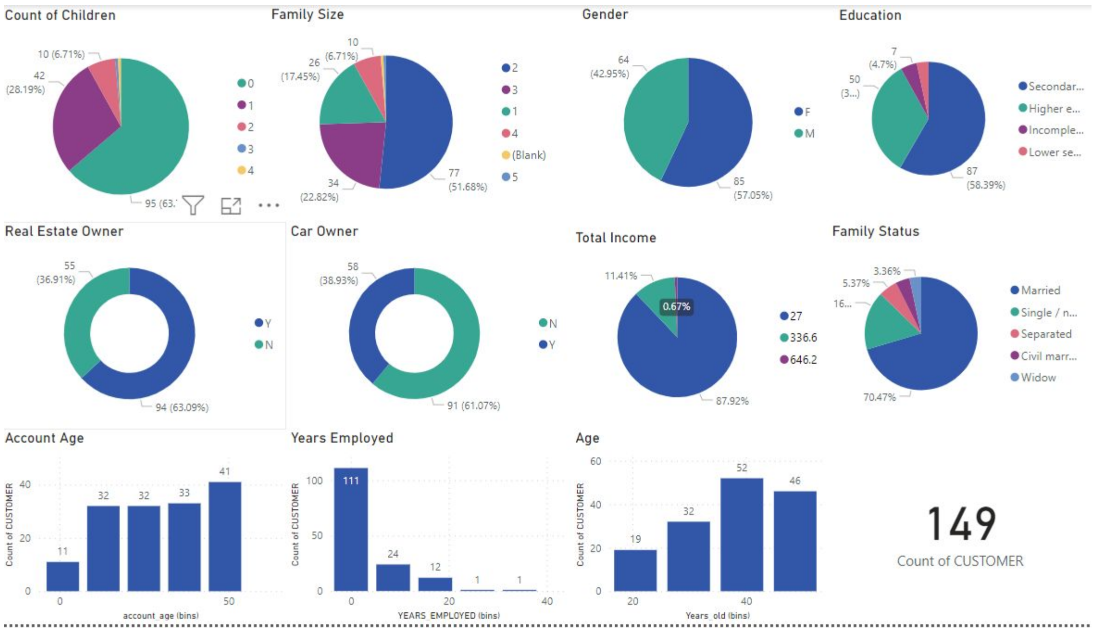
Created by PowerBi

# Good Customer


- Count of Children : 0
- Family Size : 2 - 3
- Gender : Female
- Education :  Secondary
- Property Owner : Yes
- Car Owner : No
- Total Income: under 27k
- Family Status: Married
- Years Employed: 0-5 years

# Bad Customer

- Count of Children : 0
- Family Size : 2 - 3
- Gender : Female
- Education :  Secondary
- Property Owner : Yes
- Car Owner : No
- Total Income: under 27k
- Family Status: Married
- Years Employed: 0-5 years



# Data Cleaning and Preparation

- Identify missing data: No Null Value or Missing Value
- Identifying any outliers: Income
- Removing duplicates: “Month_balance” and “payment status”
- Create dummy variables: CODE_GENDER","FLAG_OWN_CAR","FLAG_OWN_REALTY","CNT_CHILDREN","NAME_INCOME_TYPE","NAME_EDUCATION_TYPE", ”NAME_FAMILY_STATUS", "NAME_HOUSING_TYPE", "CNT_FAM_MEMBERS"
- Removing Columns: Cell phone
- Merge categorical variable with many levels: Amt Children, Amt_family_member

# Feature Engineering

**Create a  binary feature** which identifies if a customer is delinquent based on industry standard of 6 month overdue

# Before Feature Engineering:

| ID | MONTHS_BALANCE | STATUS |
|---|---|---|
| 5001711 | 0 | X |
| 5001711 | -1 | 0 |
| 5001711 | -2 | 0 |
| 5001711 | -3 | 0 |
| 5001712 | 0 | C |
| 5001712 | -1 | C |
| 5001712 | -2 | C |
| 5001712 | -3 | C |

# After Feature Engineering

| | ID | account_age | One_month_OD | two_months_OD | three_months_OD | four_months_OD | five_months_OD | six_months_OD | Current | NO_loan |
|---|---|---|---|---|---|---|---|---|---|---|
| 5001711 | 5001711 | 4 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 5001712 | 5001712 | 19 | 10 | 0 | 0 | 0 | 0 | 0 | 9 | 0 |
| 5001713 | 5001713 | 22 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 22 |
| 5001714 | 5001714 | 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 15 |
| 5001715 | 5001715 | 60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 60 |
| 5001717 | 5001717 | 22 | 17 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 5001718 | 5001718 | 39 | 24 | 2 | 0 | 0 | 0 | 0 | 3 | 10 |

# Unbalanced Data Set

Majority Class:    36277 = 99.5%

Minority class: 180 = .5%

Undersample: 180 majority and minority

Oversample: 36277 majority and minority

# Engineered Feature 2

**Percent_Late:** Created a binary variable that helps us identify what customers were profitable by means of late free generation.

**Feature Engineering:** We aggregated the total credit history in months, and divided it by the total amount of months the record was late for 1-3 months

**Criteria**: record that had  lat no more than 40 percent of the time but were never late for for more than 3 months in a row

# Merge Demographic data and New Data with feature engineering

```
'data.frame':	36457 obs. of  27 variables:
 $ ID                        : int  5008804 5008805 5008806 5008808 5008809 5008810 5008811 5008812 5008813 5008814 ...
 $ CODE_GENDER               : Factor w/ 2 levels "F","M": 2 2 2 1 1 1 1 1 1 1 ...
 $ FLAG_OWN_CAR              : Factor w/ 2 levels "N","Y": 2 2 2 1 1 1 1 1 1 1 ...
 $ FLAG_OWN_REALTY           : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
 $ CNT_CHILDREN              : chr  "0" "0" "0" "0" ...
 $ NAME_INCOME_TYPE          : Factor w/ 5 levels "Commercial associate",..: 5 5 5 1 1 1 1 2 2 2 ...
 $ NAME_EDUCATION_TYPE       : Factor w/ 5 levels "Academic degree",..: 2 2 5 5 5 5 5 2 2 2 ...
 $ NAME_FAMILY_STATUS        : Factor w/ 5 levels "Civil marriage",..: 1 1 2 4 4 4 4 3 3 3 ...
 $ NAME_HOUSING_TYPE         : Factor w/ 6 levels "Co-op apartment",..: 5 5 2 2 2 2 2 2 2 2 ...
 $ FLAG_WORK_PHONE           : int  1 1 0 0 0 0 0 0 0 0 ...
 $ FLAG_PHONE                : int  0 0 0 1 1 1 1 0 0 0 ...
 $ FLAG_EMAIL                : int  0 0 0 1 1 1 1 0 0 0 ...
 $ CNT_FAM_MEMBERS           : chr  "2" "2" "2" "1" ...
 $ YEARS_EMPLOYED            : num  12.44 12.44 3.11 8.36 8.36 ...
 $ Years_old                 : num  32.9 32.9 58.8 52.4 52.4 ...
 $ AMT_INCOME_TOTAL_thousand: num  428 428 112 270 270 ...
 $ account_age               : num  16 15 30 5 27 27 39 21 17 18 ...
 $ One_month_OD              : int  1 1 7 2 0 6 6 14 14 14 ...
 $ two_months_OD             : int  1 1 0 0 0 0 0 0 0 0 ...
 $ three_months_OD           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ four_months_OD            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ five_months_OD            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ six_months_OD             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Current                   : int  13 12 7 0 0 15 27 0 0 0 ...
 $ NO_loan                   : int  1 1 16 3 5 6 6 3 3 3 ...
 $ Percent_1month_late       : num  1 1 1 1 0 1 1 0 0 0 ...
 $ six_months_OD1            : num  0 0 0 0 0 0 0 0 0 0 ...
```

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

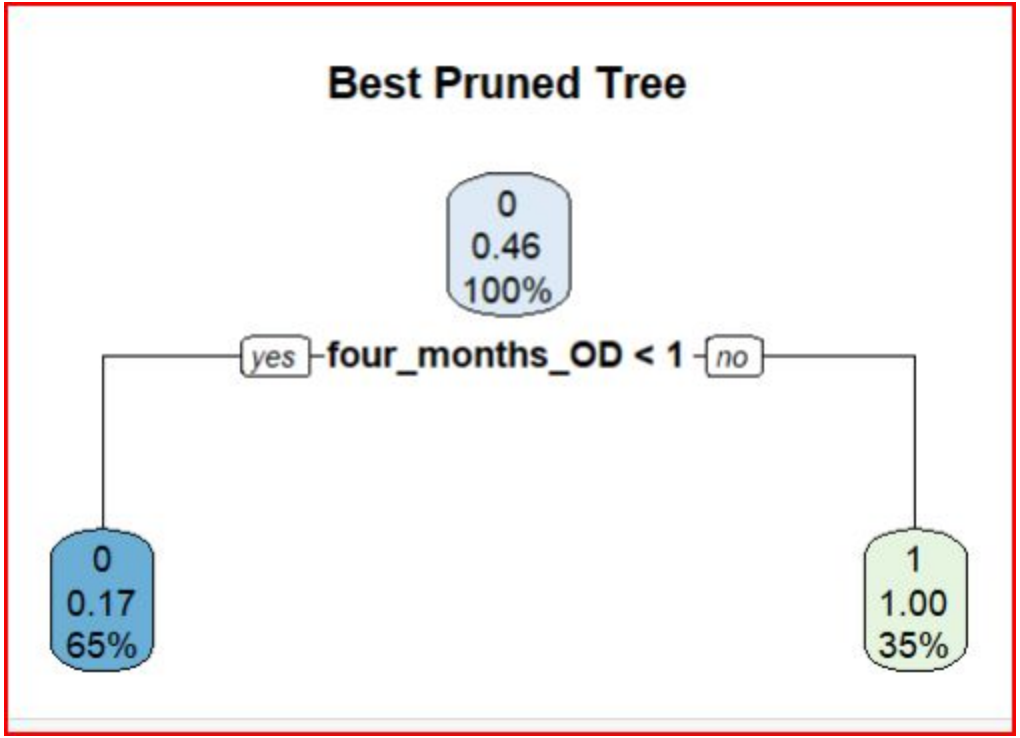
1. Perform Feature engineering to create dependent variable to be used classification analysis.
2. Create a Balanced Data set to be used in classification analysis.
3. Build  various supervised learning algoriths to predict if a customer will delinquint.
4. Evaluate the model ,and choose the best model and gain insight on how it can help management

# Cost Parameter

```
Classification tree:
rpart(formula = six_months_OD1 ~ ., data = under_train, method = "class",
    minsplit = 7)

Variables actually used in tree construction:
[1] AMT_INCOME_TOTAL_thousand    FLAG_OWN_CAR_Y             four_months_OD             NAME_FAMILY_STATUS_Separated
[5] One_month_OD                 Percent_1month_late        Years_old

Root node error: 125/270 = 0.46296

n= 270

      CP nsplit rel error xerror     xstd
1 0.7600      0     1.000  1.000 0.065546
2 0.0144      1     0.240  0.240 0.041312
3 0.0120      7     0.152  0.320 0.046698
4 0.0100      9     0.128  0.304 0.045713
```

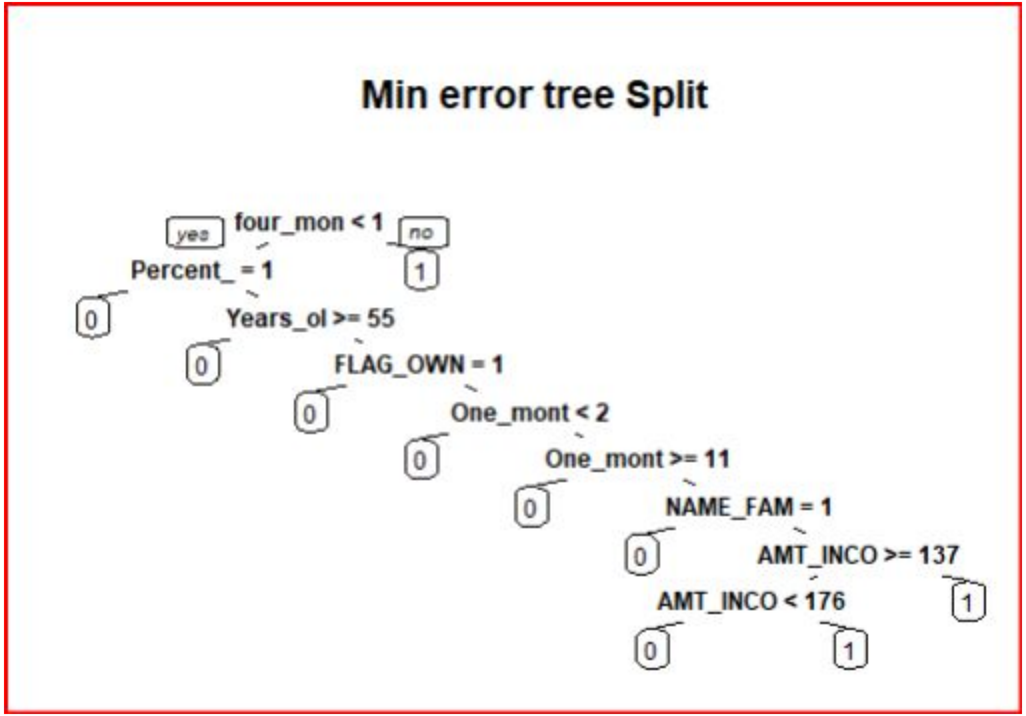
# Best pruned tree

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 34 13
         1  1 42

               Accuracy : 0.8444
                 95% CI : (0.7528, 0.9
    No Information Rate : 0.6111
    P-Value [Acc > NIR] : 1.258e-06

                  Kappa : 0.6919

 Mcnemar's Test P-Value : 0.003283

            Sensitivity : 0.9714
            Specificity : 0.7636
         Pos Pred Value : 0.7234
         Neg Pred Value : 0.9767
```

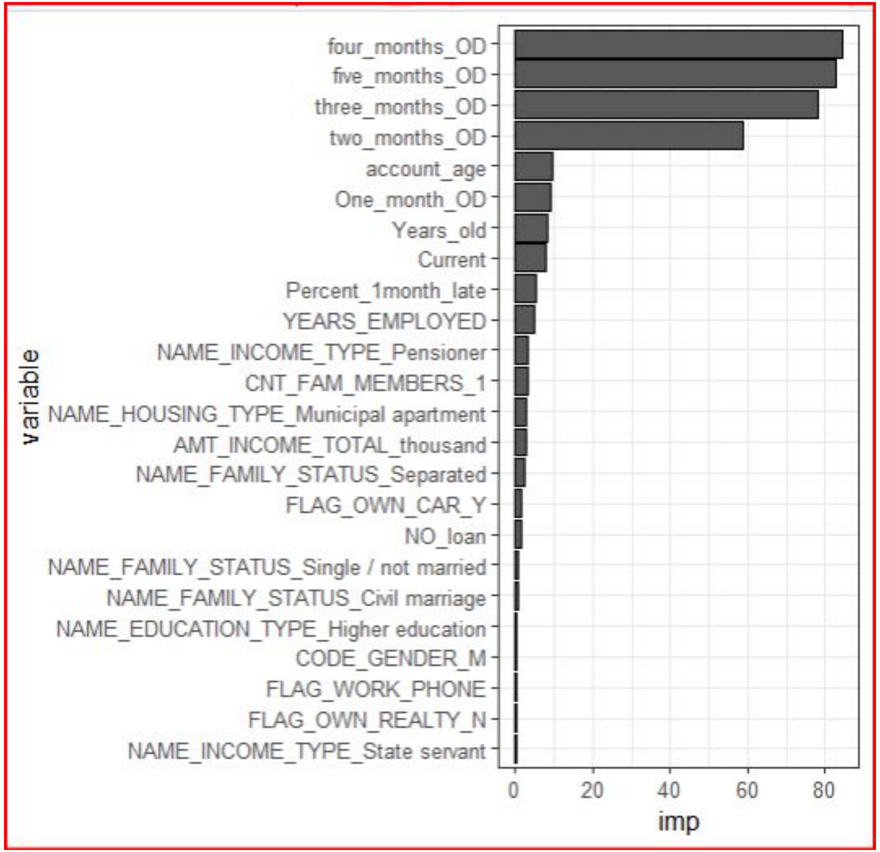


# Full Tree

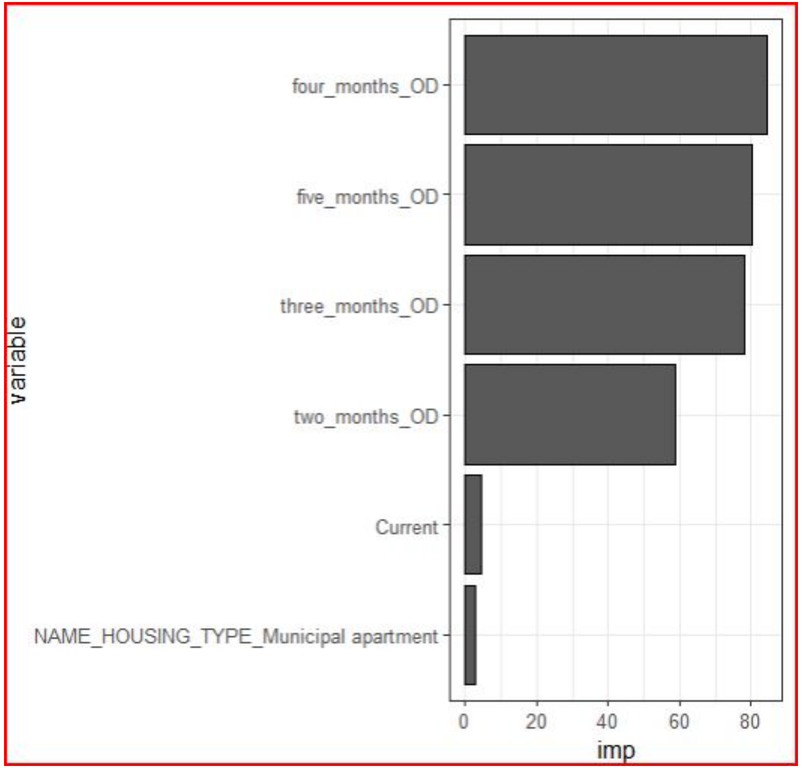
```
> confusionMatrix(predict_under_train_full
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 33 13
         1  2 42

               Accuracy : 0.8333
                 95% CI : (0.74, 0.9036)
    No Information Rate : 0.6111
    P-Value [Acc > NIR] : 4.19e-06

                  Kappa : 0.6683

 Mcnemar's Test P-Value : 0.009823

            Sensitivity : 0.9429
            Specificity : 0.7636
```



# Best Pruned

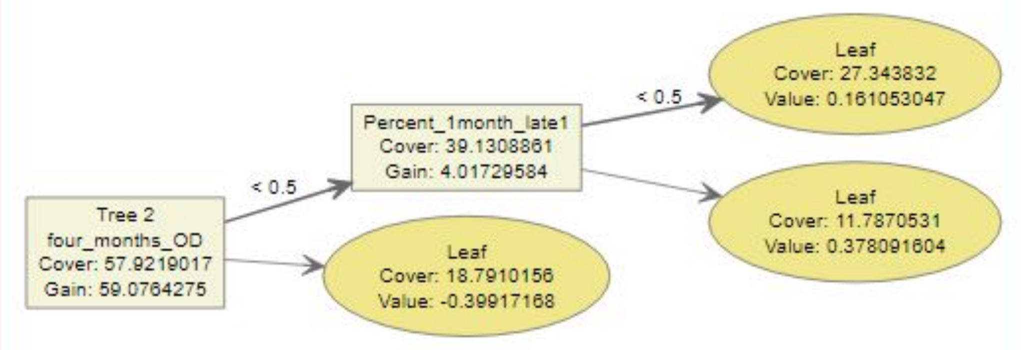



# Min error

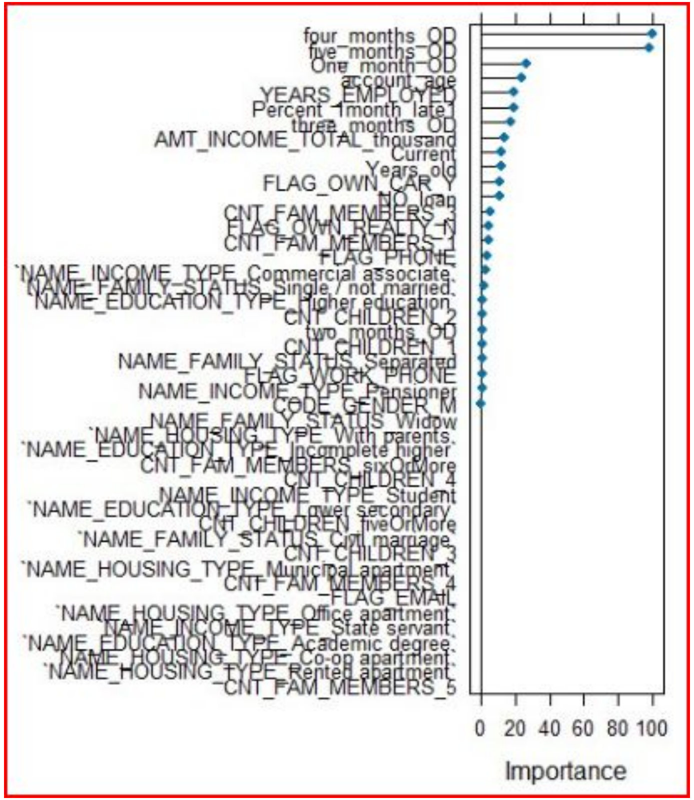


# XG Boost

```
          Reference
Prediction  0  1
         0 32  7
         1  3 48

               Accuracy : 0.8889
                 95% CI : (0.8051, 0.9454)
    No Information Rate : 0.6111
    P-Value [Acc > NIR] : 4.329e-09

                  Kappa : 0.771

 Mcnemar's Test P-Value : 0.3428

            Sensitivity : 0.9143
            Specificity : 0.8727
         Pos Pred Value : 0.8205
         Neg Pred Value : 0.9412
```



# XG Boost to predict If a customer will be delinquint




Five Most Important Variables

1. Four Month Overdue
2. Five Months Overdue
3. One Month Overdue
4. Percent Late
5. Account age

# Best method to predict is a customer will default on there loan

Methods used: After careful analysis we learned that no model deemed fit to predict whether the customer will default on there loan.

Assigning them them all to  the majority class will produce better accuracy than any of the models we built.

Mcnemar test P -Value .11  Not significant in predicting better than the majoriety class

# Late Fee Prediction

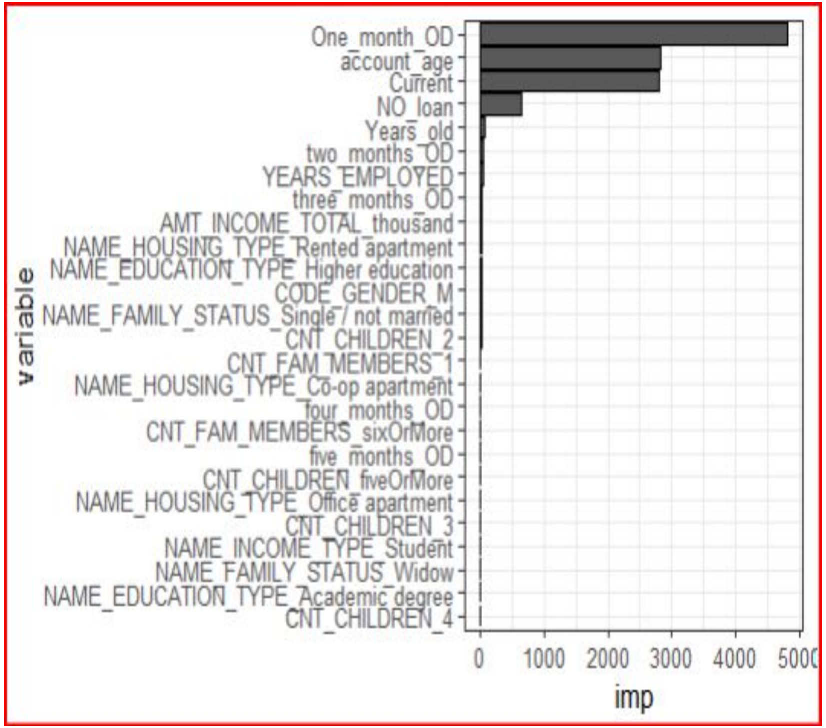
How can We predict if a customer will be profitable based on our data.

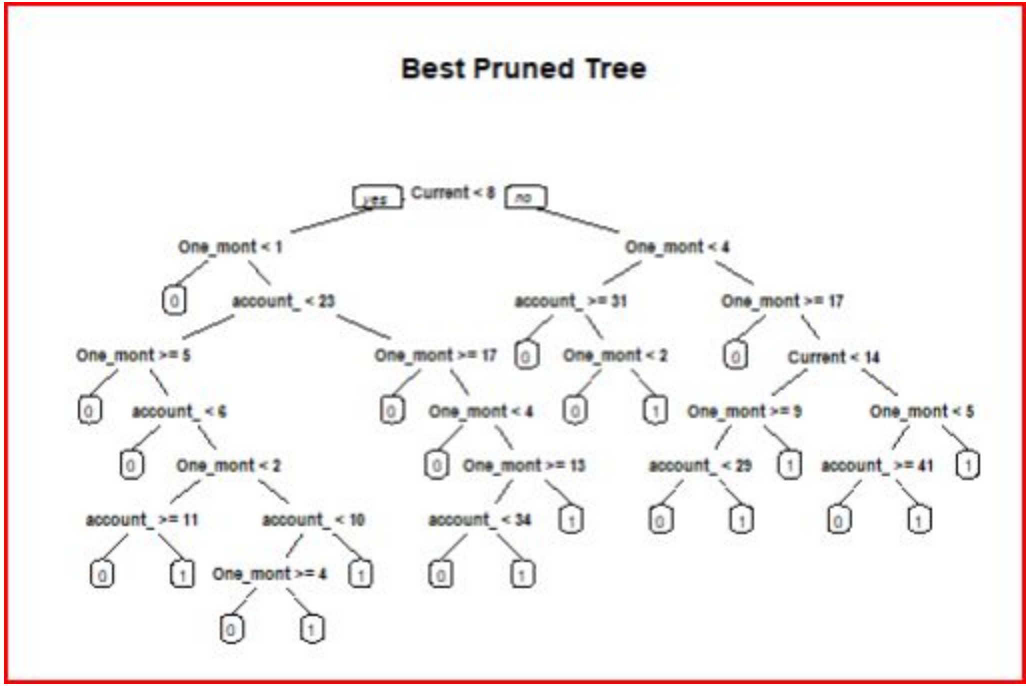
1. Perform Feature engineering to create dependent variable to be used classification analysis.
2. Build  various Decision Trees and logistics regression model to predict if a customer will be late.
3. Evaluate the model ,and choose the best model and gain insight on how it can help management

# Late Fee prediction

| Late Fee Models | Accuracy | Sensitivity | specificity |
|---|---|---|---|
| Cart full | .9182 | .9311 | .8969 |
| Xgboost | .8889 | .9730 | .5 |
| Best pruned | .9182 | .9311 | .8969 |
| Min error | .8575 | .8567 | .8588 |
| Log forward | .6417 | .9203 | .1825 |
| Backward | 0.6421 | 0.1836 | 0.9203 |

# Best Pruned Tree



# Conclusion

**Optimize Credit Limit Management:**

Identify a threshold for credit limits that minimizes the risk of customers incurring late fees while still providing sufficient credit access.

**Targeted Communication and Education:**

Develop targeted communication strategies to educate customers about the implications of outstanding debt and late fees.

**Customized Credit Limit Assignments:**

Explore personalized credit limit assignments based on individual customer profiles and financial behaviors.

**Early Warning Systems:**

Implement early warning systems or alerts for customers approaching their credit limits or exhibiting patterns associated with late fees.

**Reward Programs for Responsible Behavior:**

Introduce or enhance reward programs that incentivize responsible credit card usage, timely payments, and maintaining lower debt levels.

**Continuous Monitoring and Adaptation:**

Establish a framework for continuous monitoring of credit-related metrics and adapt strategies based on evolving customer behavior and economic conditions.