LOAN DEFAULT PREDICTION

Introduction

Problem Statement Vehicle Loan Default Prediction Financial institutions incur significant losses due to the default of vehicle loans. This has led to the tightening up of vehicle loan underwriting and increased vehicle loan rejection rates. The need for a better credit risk scoring model is also raised by these institutions. This warrants a study to estimate the determinants of vehicle loan default. A financial institution has hired us to accurately predict the probability of loanee/borrower defaulting on a vehicle loan in the first EMI (Equated Monthly Instalments) on the due date.

Data

disbursed_amount: Amount of Loan disbursed

asset_cost: Cost of the Asset

Ltv: Loan to Value of the asset

Date.of.Birth: Date of birth of the customer

Employment.Type: Employment Type of the customer (Salaried/Self Employed)

DisbursalDate: Date of disbursement

Passport_flag: if passport was shared by the customer then flagged as 1

Driving_flag: if DL was shared by the customer then flagged as 1

PRI.NO.OF.ACCTS: count of total loans taken by the customer at the time of disbursement

PRI.ACTIVE.ACCTS: count of active loans taken by the customer at the time of disbursement

CREDIT.HISTORY.LENGTH: Time since first loan

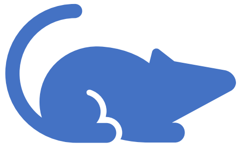
loan_default: Payment default in the first EMI on due date

Formatting

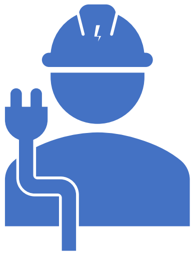
- Merging
- Checking the presence of NAs
- Changing data types
- Variable transformation
- Feature creation

Missing value treatment

Prediction using rpart

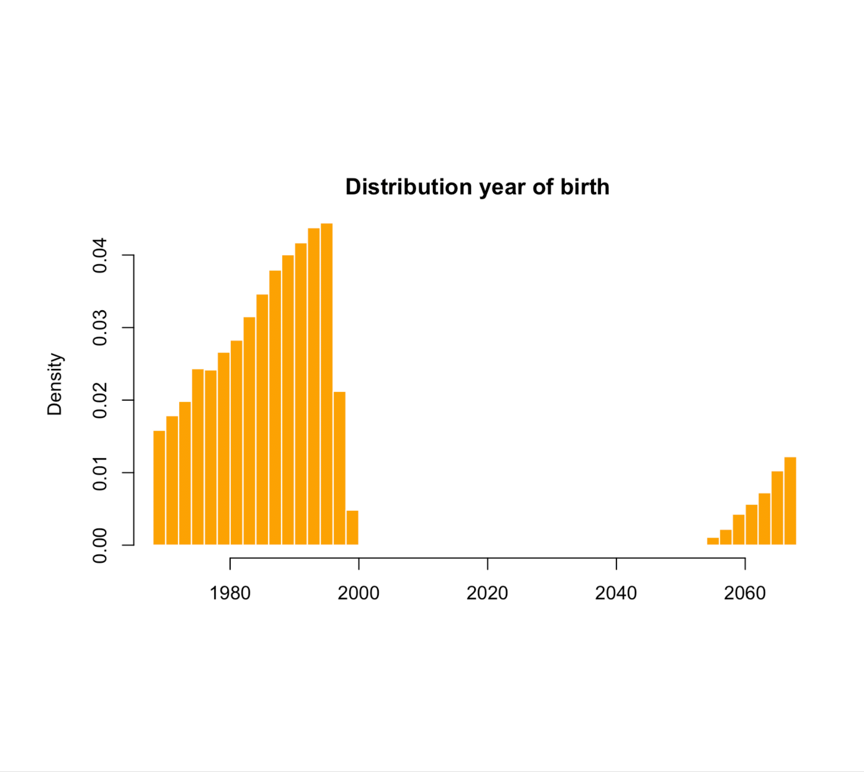


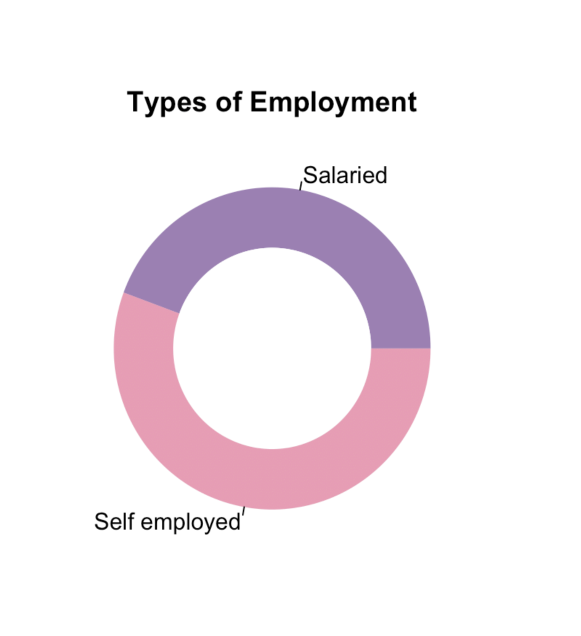
Prediction using mice

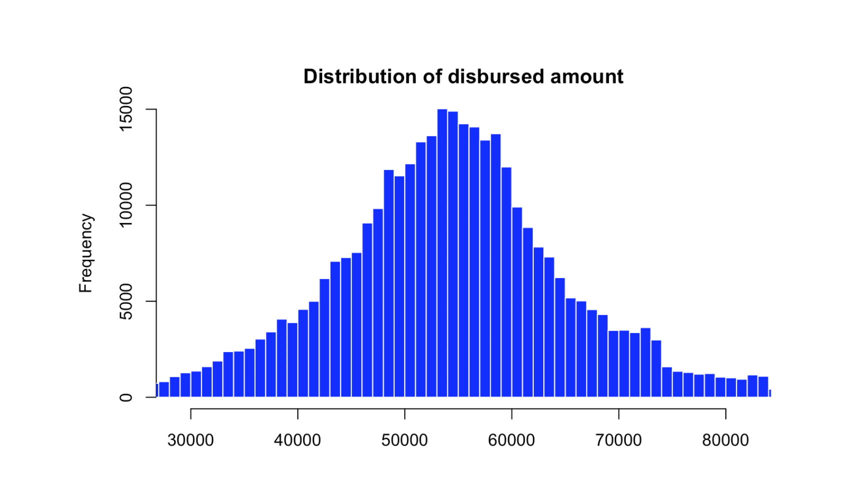


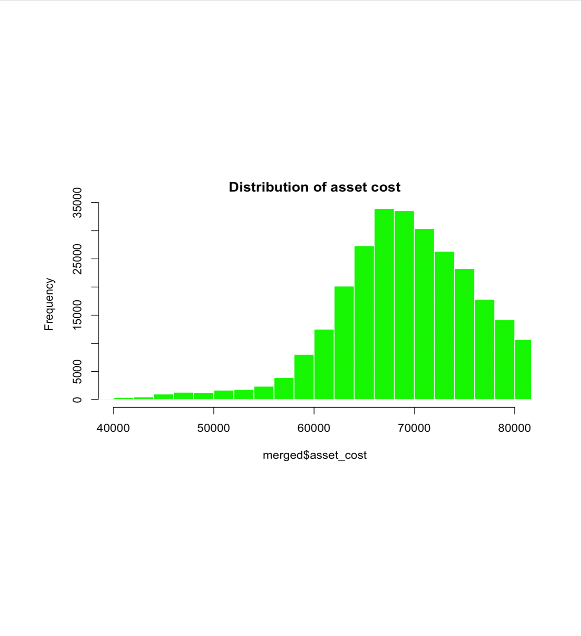
Replacing with median

Visualization









Variable selection

- Caret
- Random forest

Linear Discriminant Analysis (LDA)

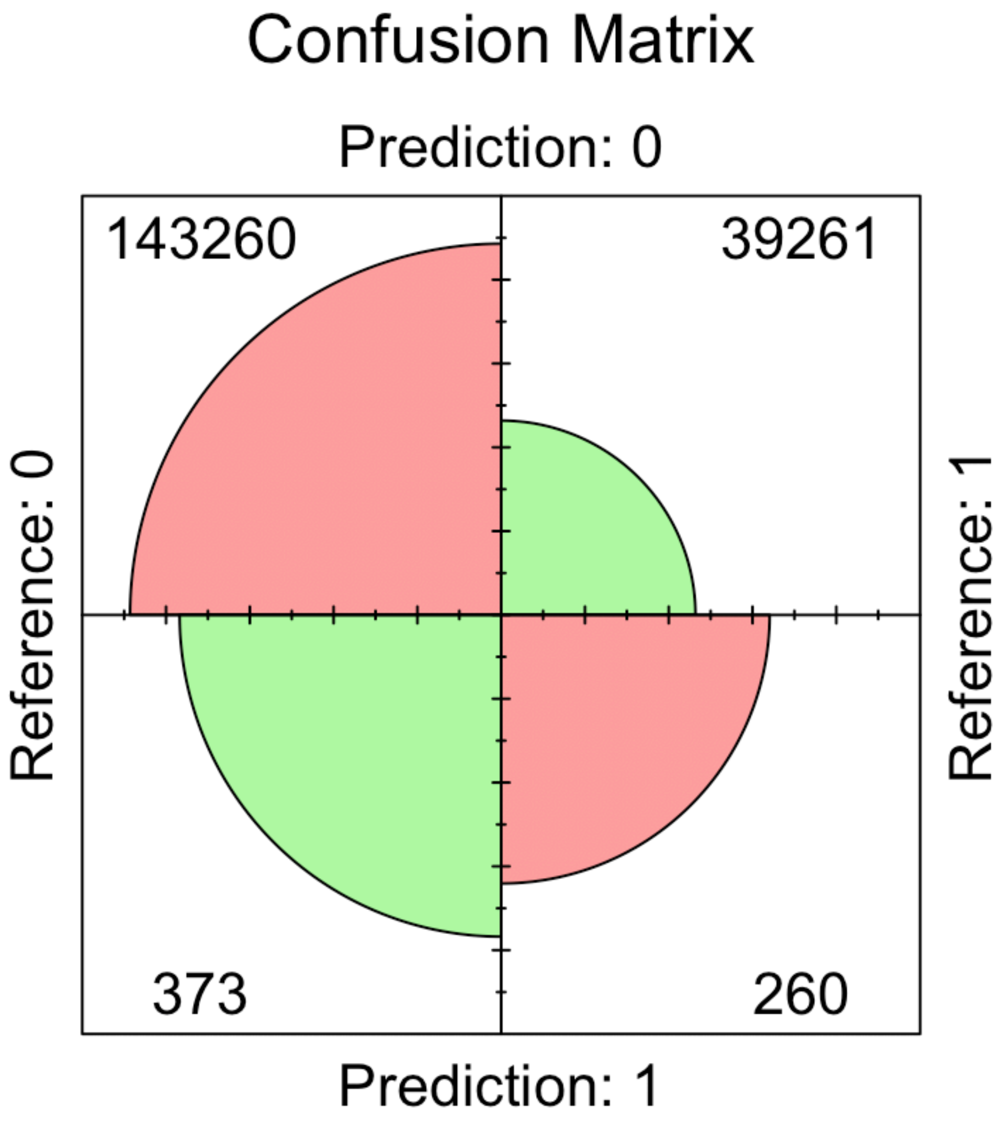
Logistic regression

Modeling

Linear Discriminant Analysis

```
          Reference
Prediction      0      1
         0 143051  39608
         1    294    201
                                          
               Accuracy : 0.7821          
                 95% CI : (0.7802, 0.784) 
    No Information Rate : 0.7826          
    P-Value [Acc > NIR] : 0.702           
                                          
                  Kappa : 0.0047          
 Mcnemar's Test P-Value : <2e-16          
                                          
            Sensitivity : 0.997949        
            Specificity : 0.005049        
         Pos Pred Value : 0.783159        
         Neg Pred Value : 0.406061        
             Prevalence : 0.782647        
         Detection Rate : 0.781042        
   Detection Prevalence : 0.997297        
      Balanced Accuracy : 0.501499        
                                          
       'Positive' Class : 0               
```

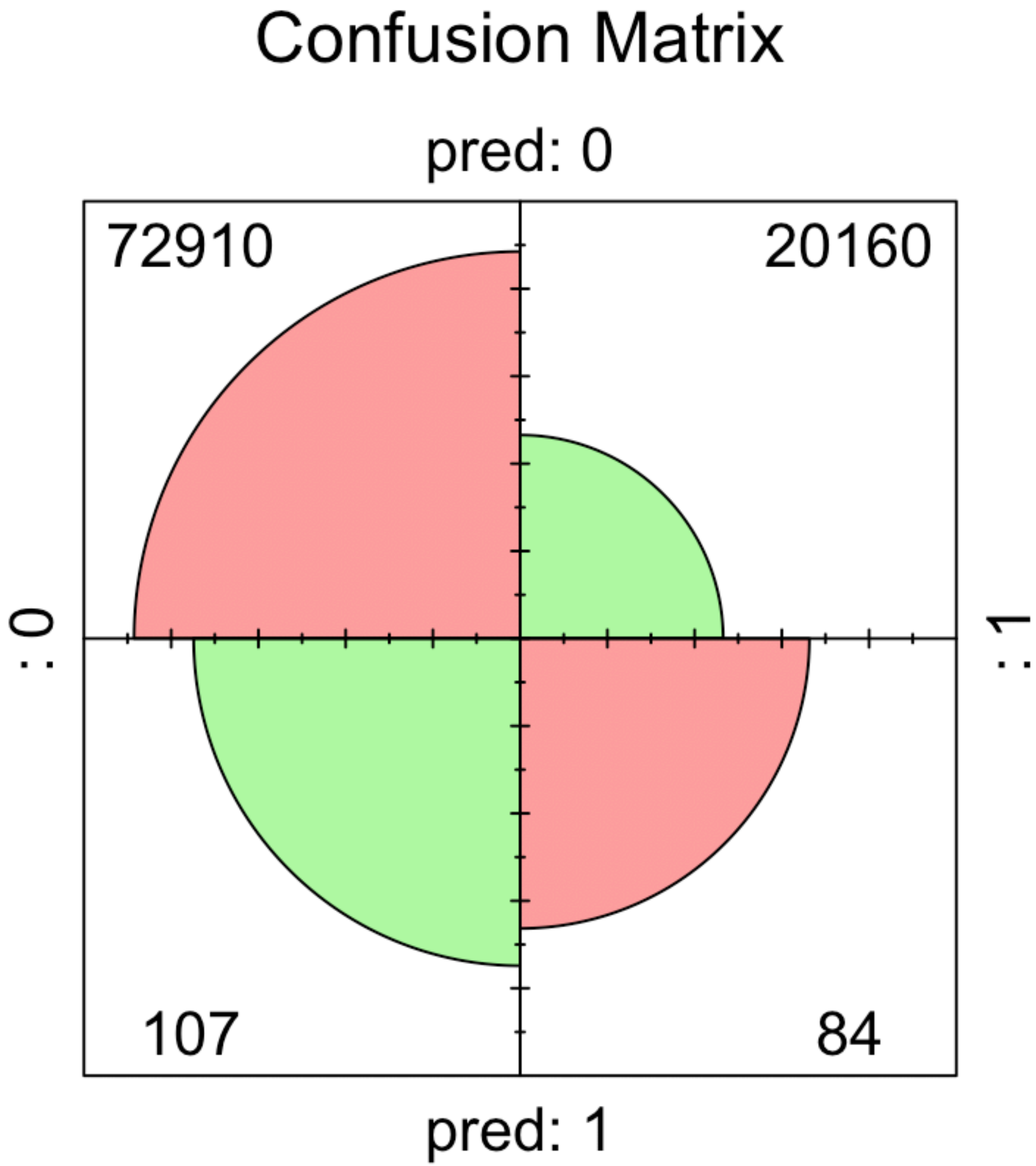
LDA. Visualization of results



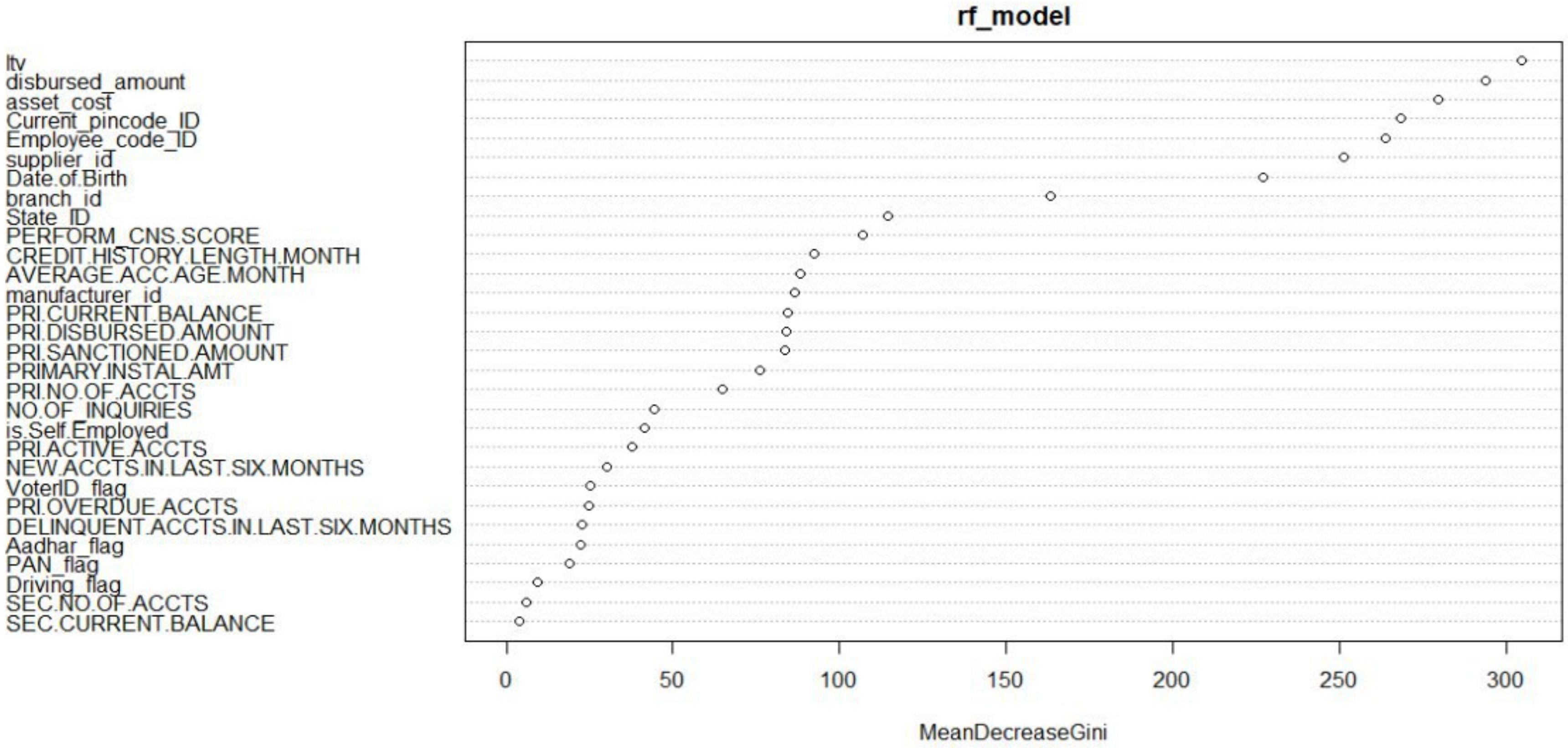
Logistic regression

```
Confusion Matrix and Statistics

pred     0     1
   0 72910 20160
   1   107    84

               Accuracy : 0.7827
                 95% CI : (0.78, 0.7853)
    No Information Rate : 0.7829
    P-Value [Acc > NIR] : 0.5743

                  Kappa : 0.0042
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.998535
            Specificity : 0.004149
         Pos Pred Value : 0.783389
         Neg Pred Value : 0.439791
             Prevalence : 0.782932
         Detection Rate : 0.781784
   Detection Prevalence : 0.997952
      Balanced Accuracy : 0.501342

       'Positive' Class : 0
```

GLM. Visualization of results



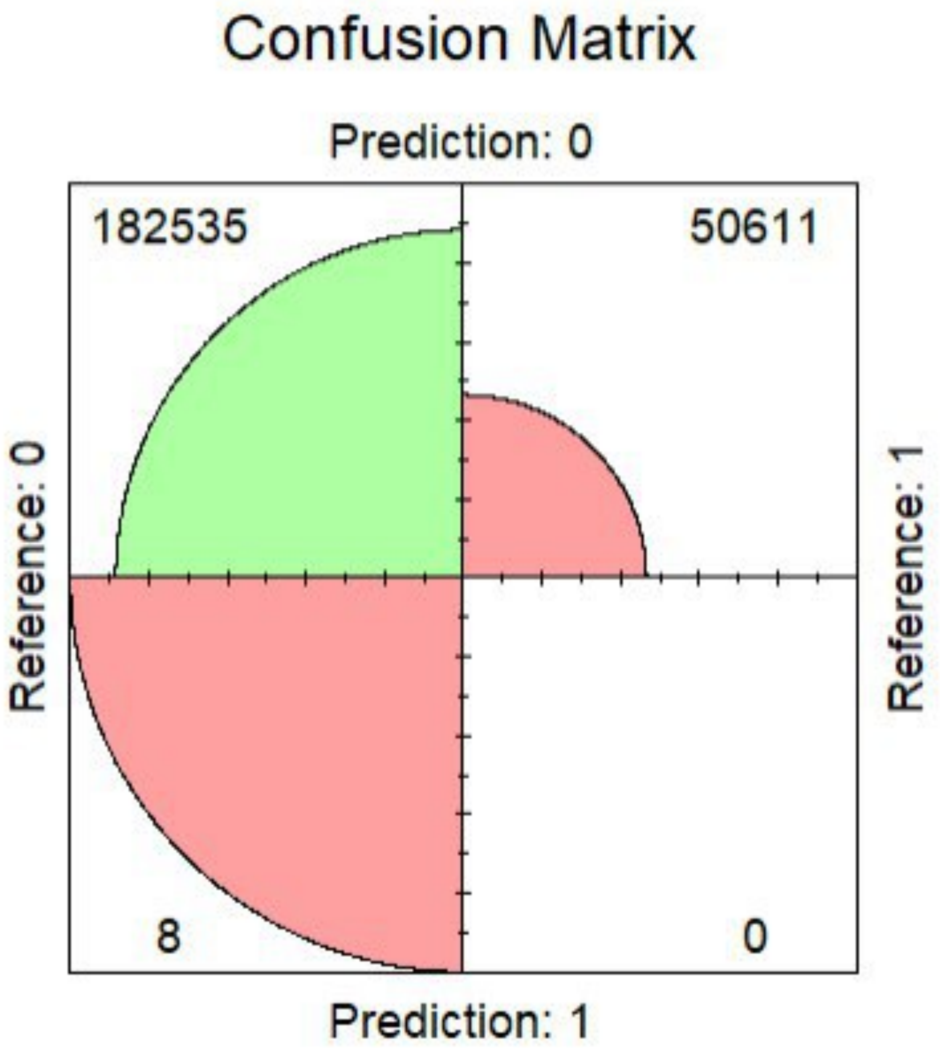
Random forest model. Variable importance



Model for these variables (10)

```
Call:
glm(formula = loan_default ~ disbursed_amount + asset_cost +
    ltv + branch_id + supplier_id + Current_pincode_ID + Date.of.Birth +
    State_ID + Employee_code_ID + PERFORM_CNS.SCORE, family = binomial(link = logit),
    data = data.model1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9722  -0.7435  -0.6528  -0.4716   2.6677

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.441e+00  4.630e-01  -7.433 1.06e-13 ***
disbursed_amount   -1.021e-05  2.225e-06  -4.590 4.42e-06 ***
asset_cost          1.224e-05  1.501e-06   8.156 3.46e-16 ***
ltv                 3.887e-02  1.783e-03  21.807  < 2e-16 ***
branch_id           4.580e-04  7.450e-05   6.148 7.86e-10 ***
supplier_id         1.081e-05  1.549e-06   6.980 2.96e-12 ***
Current_pincode_ID  5.029e-05  2.485e-06  20.241  < 2e-16 ***
Date.of.Birth      -8.335e-04  2.229e-04  -3.740 0.000184 ***
State_ID            2.104e-02  1.147e-03  18.344  < 2e-16 ***
Employee_code_ID    3.506e-05  5.259e-06   6.667 2.61e-11 ***
PERFORM_CNS.SCORE  -4.321e-04  1.559e-05 -27.711  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 243961  on 233153  degrees of freedom
Residual deviance: 238818  on 233143  degrees of freedom
AIC: 238840
```

Random Forest. Visualization of results



Thanks for your attention!

Resources:
https://github.com/SofiyaHevorhyan/LoanDefaultAnalysis