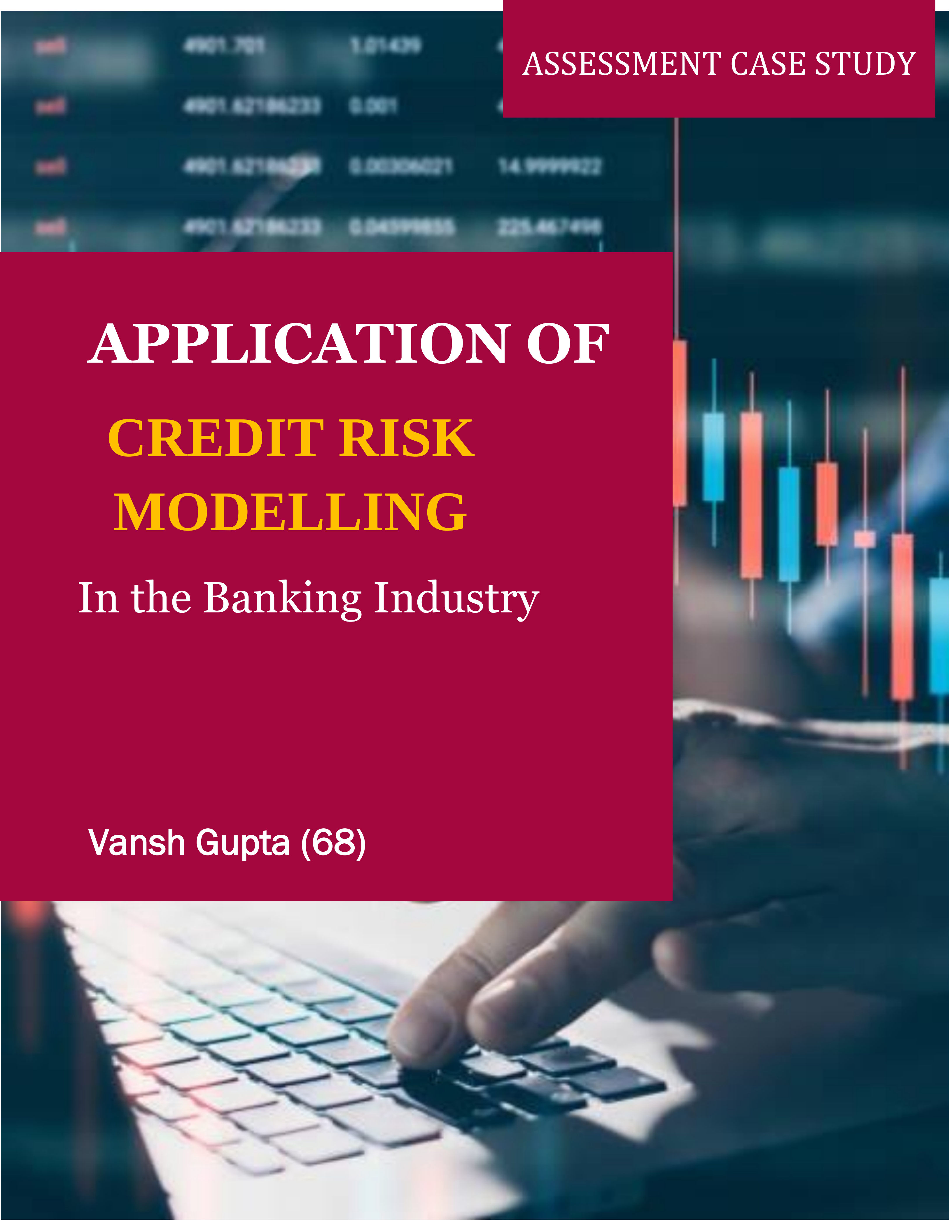
ASSESSMENT CASE STUDY

APPLICATION OF CREDIT RISK MODELLING

In the Banking Industry

Vansh Gupta (68)



INDEX

ADDITIONAL INFORMATION

BACKGROUND

Non-Performing Assets (NPAs) have expanded significantly in the sector, creating a riskier climate that has resulted in significant losses for banks and businesses. These losses have been exacerbated by the absence of a strong framework for credit evaluation. Our work attempts to design efficient models to meet this difficulty and reduce hazards. In light of the most recent regulatory framework, the Internal Model Approach is our choice.

To reduce losses in the long run, the goal is to develop models that may greatly enhance credit evaluation procedures. It is in compliance with the most recent regulatory recommendations to use the Internal Model Approach. Using historical data to estimate the chance of default, the established credit risk model facilitates the assessment of possible obligors' creditworthiness.

This approach allows us to evaluate borrowers' creditworthiness and predict defaults, providing a proactive approach to loan disbursements. The company's total risk of loss in subsequent loan deals is intended to be reduced by this proactive strategy.

STEPS IN MODEL BUILDING PROCESS

1 **Data Collection and Cleaning:**
   - Collaborated with the bank's team to collect raw data.
   - Conducted a thorough cleaning process to identify and rectify errors in the raw data.

2 **Exploratory Data Analysis:**
   - Performed exploratory data analysis to gain insights into variable types and relationships.
   - Utilized graphical representations to visualize important relationships within the data

3 **Preparing the Data for the Model:**
   - Segregated the data into training and test datasets to facilitate model building and result testing.
   - Employed data splitting whenever necessary for the modelling process.

4 **Building the Model:**
   - Developed models using the relevant data, engaging in multiple iterations to identify the best-fitting model

5. **Model Testing and Evaluation**:
   - Conducted various tests on the selected model as needed.
   - Performed in-sample analysis and out-sample analysis to assess underfitting and overfitting.

6. **Model Selection:**
   - Evaluated models based on metrics and their predictive accuracy.
   - Selected the final model based on its superior accuracy and speed among all alternatives.

DATA CLEANING

1) Loan Status Variable:

- Converted the "loan_status" variable to a factor type, designating it as the dependent variable for model building.

2) Member ID and Funded Amount:

- Dropped the "member_ID" variable as it was redundant.
- Removed the "funded amount" variable due to its similarity to the "loan_amount."
- Eliminated the "acc_now_relinq" column due to its overwhelming similarity across the dataset.

3) Number of Instalments Variable:

- Removed NA values in the "number_of_instalments" variable.
- Converted "60 months" to 60 and "36 months" to 36 for consistency in preparation for model building.

4) Home Ownership Variable:

- Addressed irregularities in the "home ownership" variable.
- Classified values as OWN and RENT based on the 3rd quartile of annual income.
- If income is above the 3rd quartile, both OTHER and NONE were classified as OWN; otherwise, they were classified as RENT.

5) Handling NA Values in Total Current Balance:

- Retained NA values in the "total_current_bal" variable.
- Replaced NA values with the median of the remaining data.

6) Length of Employment Variable:

- Observed irregularities in the "length_of_employment" variable.
- Attempted coarse classification but opted for converting "10>" to 10 and "<1" to 1.
- Filled missing values with the median values for the same variable.

7) Payment Received Variable:

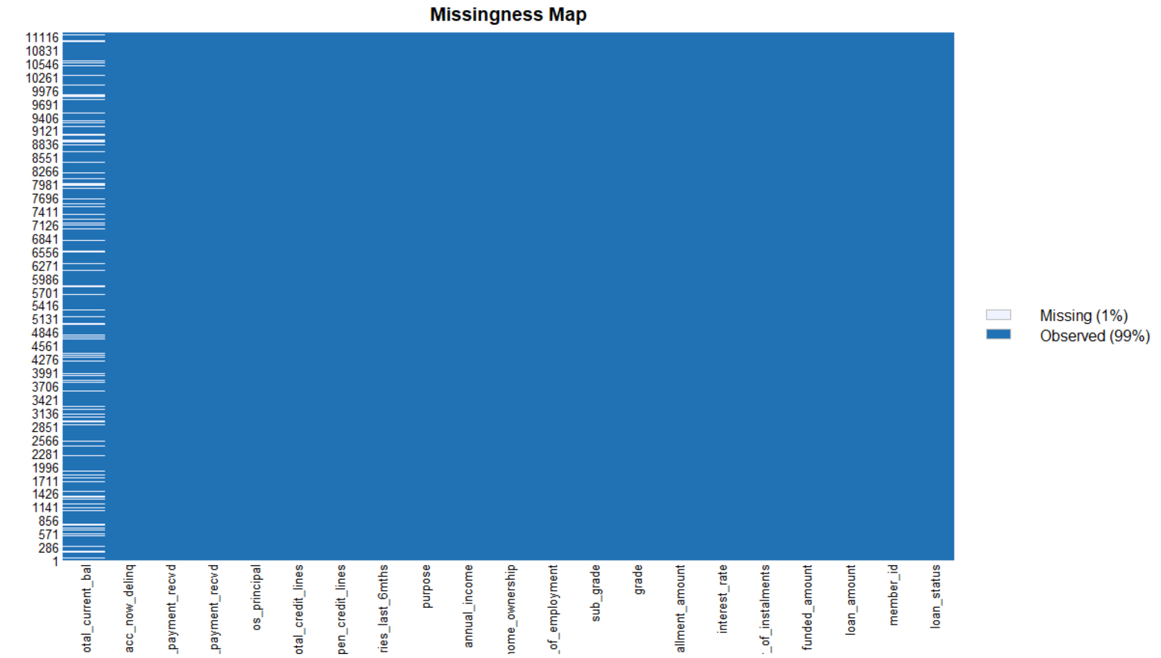
- Checked for NA values and negative values in the "payment_recvd" variable, resulting in a negative outcome.

8) Grade and Sub Grade Variables:

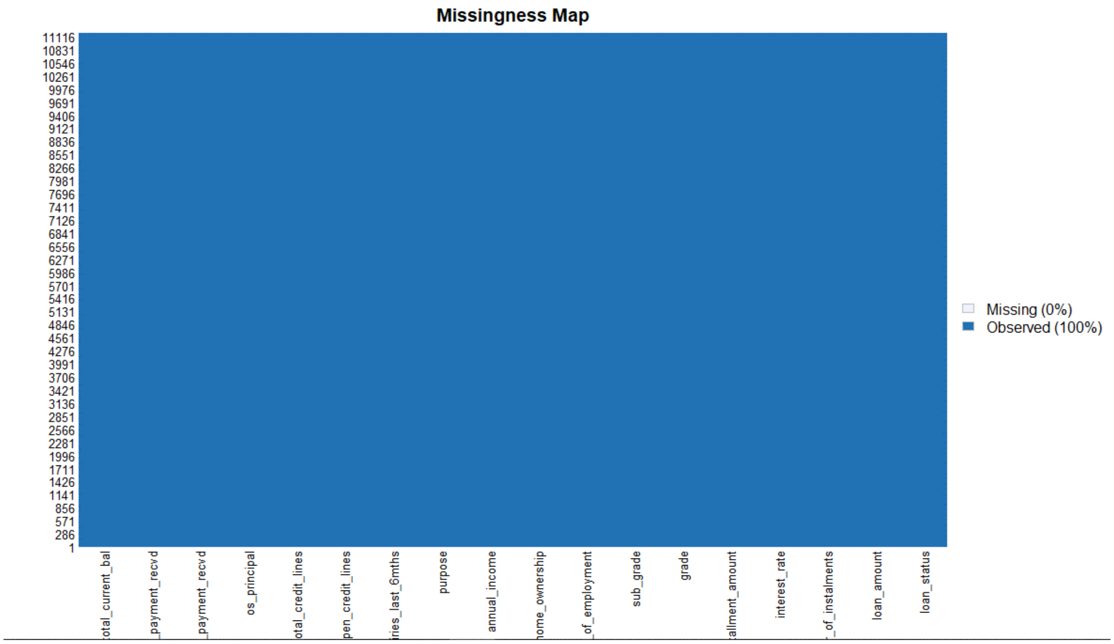
- Examined "grade" and "sub-grade" variables for unique values.

DATA CLEANING GRAPH

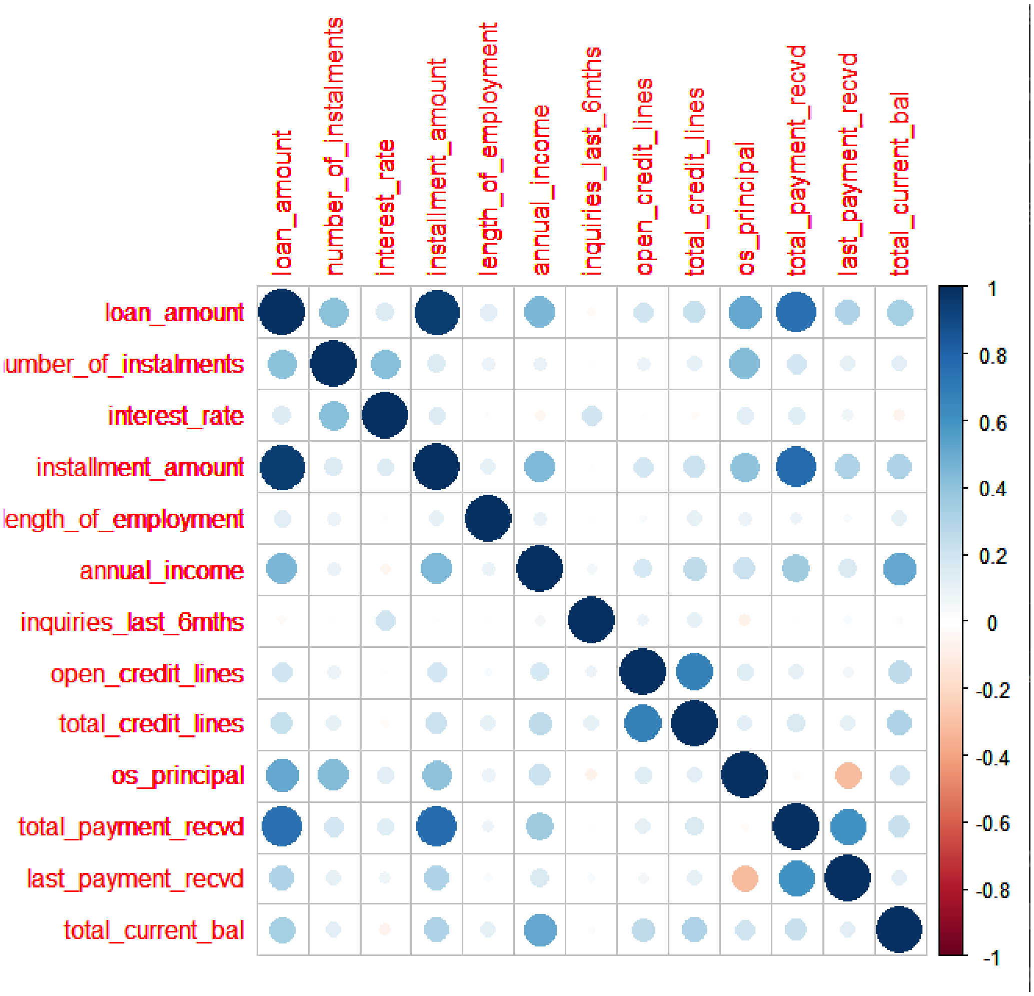
Before Data Cleaning



After Data Cleaning



IDENTIFYING MULTICOLLINEARITY



Total Payment Received and Instalment amount seem to have a high positive correlation.

Multicollinearity might be present.

EXPLORATORY DATA ANALYSIS (EDA)

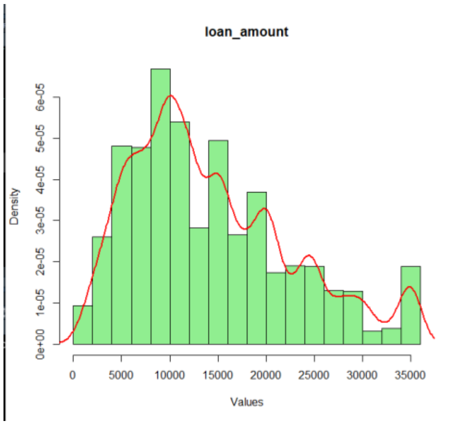
IDENTIFYING THE EXPLANATORY VARIABLE:

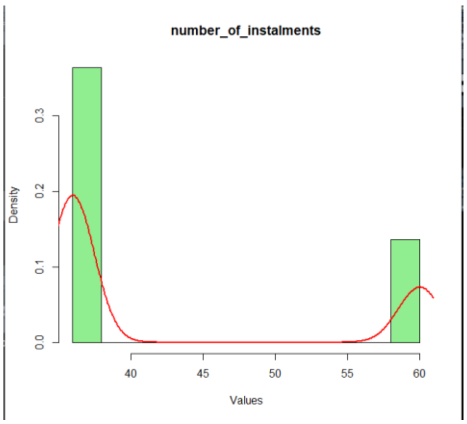
Following data cleaning, a crucial step involves conducting univariate analysis to identify the dependent variable for the model. In this context, "loan status" was selected as the response variable, while the remaining 17 variables were designated as explanatory variables.

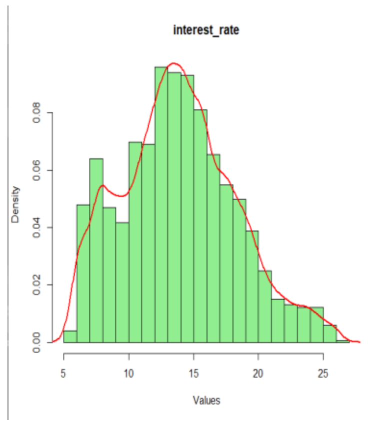
ABOUT THE DATA:

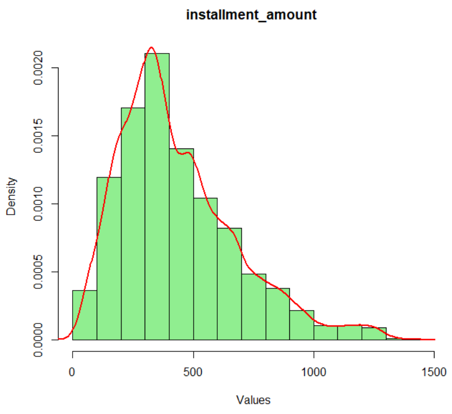
- Post data cleaning, the dataset comprises approximately 11,191 observations spanning 18 variables.
- Numeric and factor conversions were applied to specific variables to facilitate modelling.
- Various graphs were generated for modelling purposes, and exploratory data analysis (EDA) was extensively conducted on most variables to detect outliers and understand data distributions.
- Subjective exploration was performed on the data to assess potential removal of variables and intuitively identify significant factors.

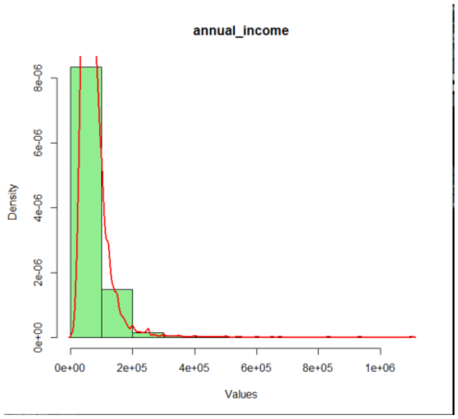
NUMERICAL VARIABLE GRAPH

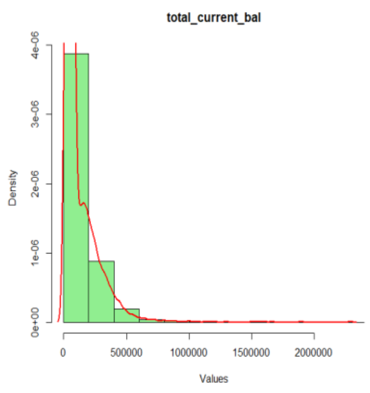


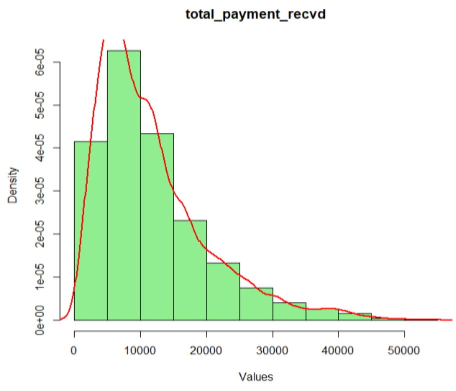


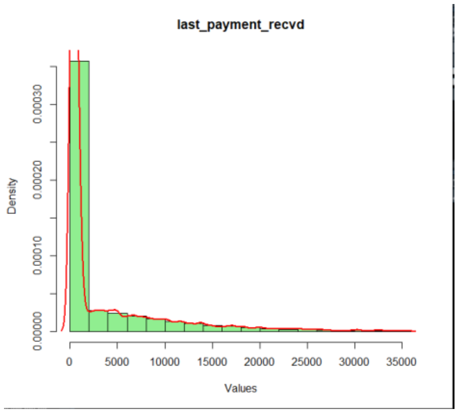








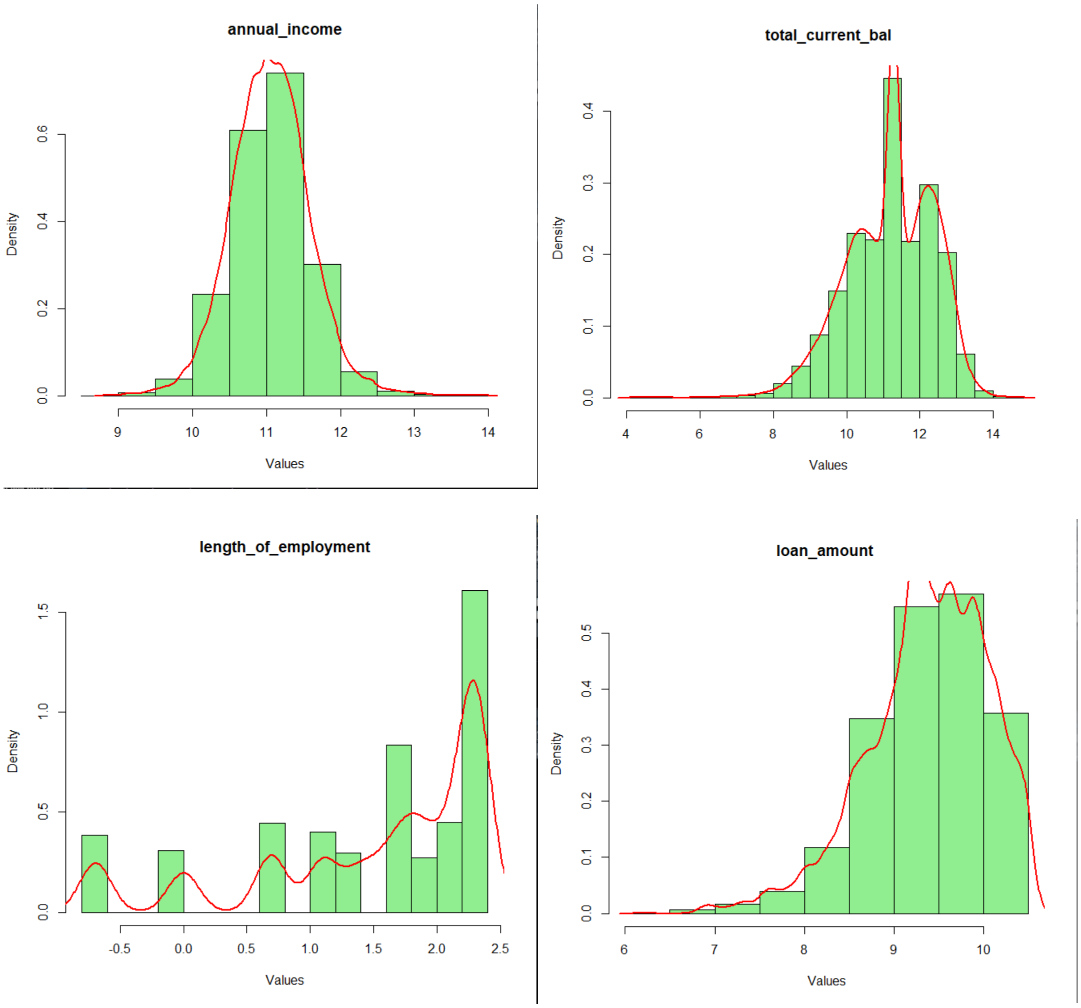




Here we can see that a lot of data is particularly peaked and skewed towards a specific point. Performing log transformations will smoothen the curve. It is required to be transformed in order for better modelling and no skewed results.

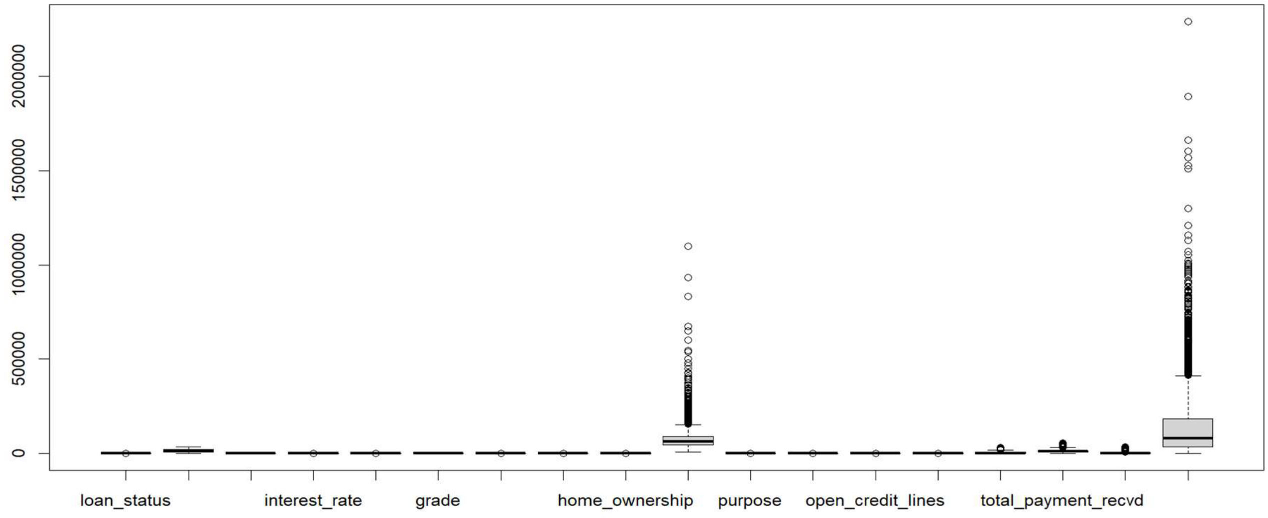
GRAPHS FOR

LOG TRANSFORMED NUMERIC VARIABLE



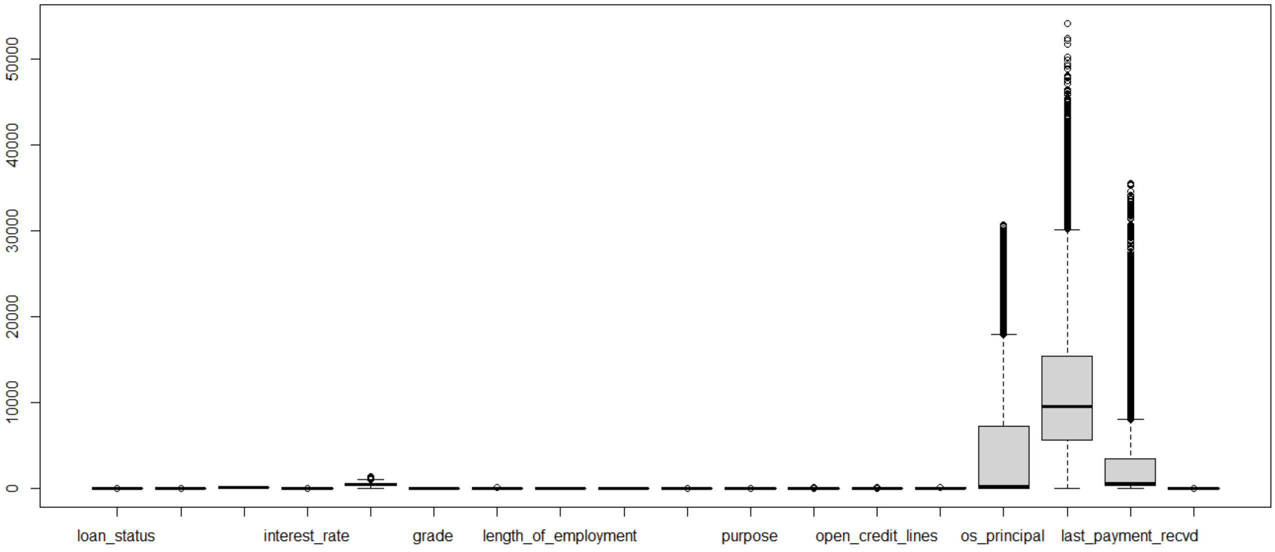
- The log transformed variables have their data residing towards the centre showing a better approximation overall.
- The character variables were difficult to plot and made lesser sense.
- Data becomes more symmetric.

BOXPLOT FOR ALL NUMERICAL VARIABLES



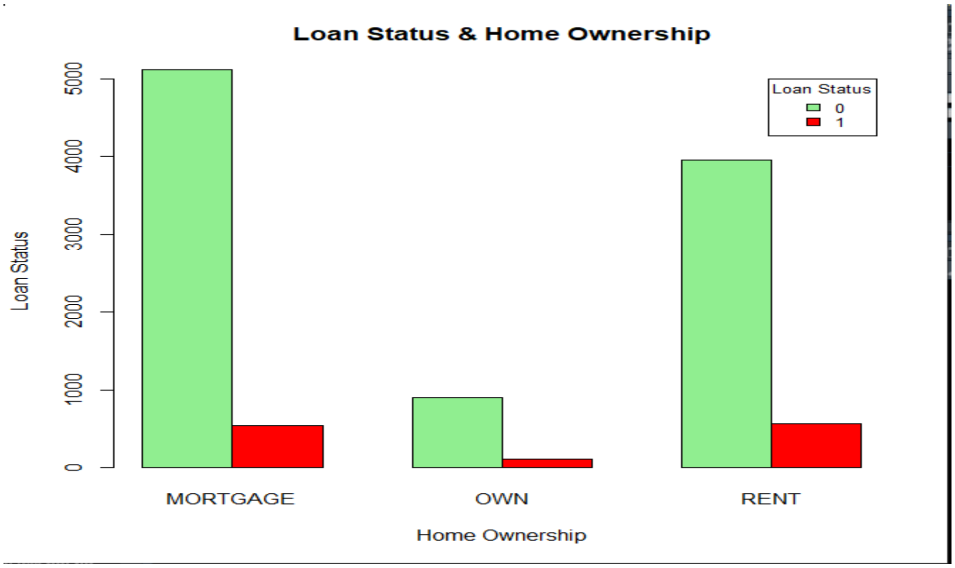
BOXPLOT FOR

LOG TRANSFERED NUMERICAL VARIABLES

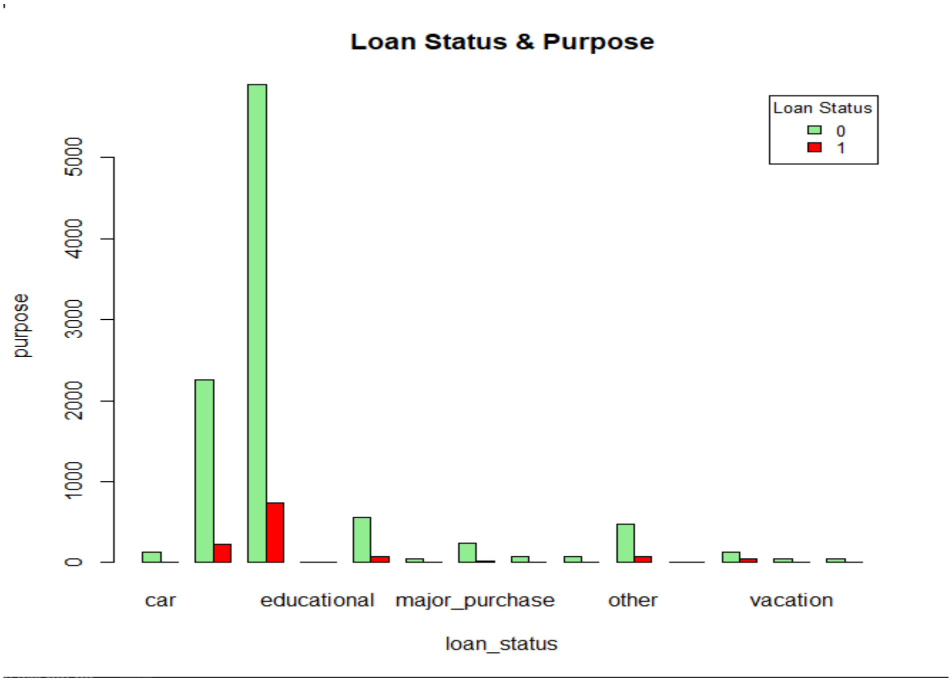


HISTOGRAM FOR CATEGORICAL VARIABLE

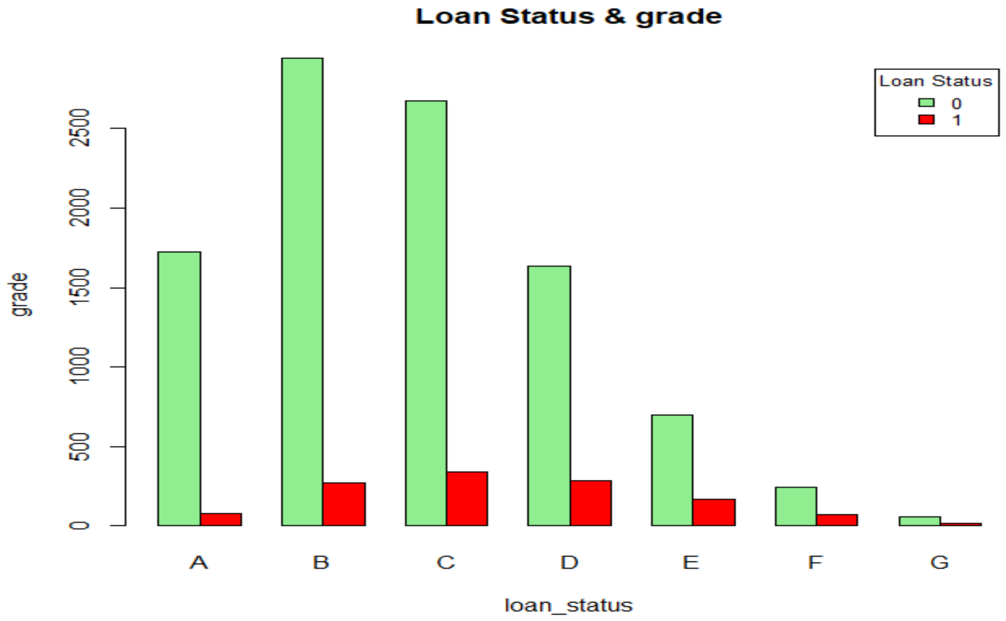
Loan Status and Home Ownership



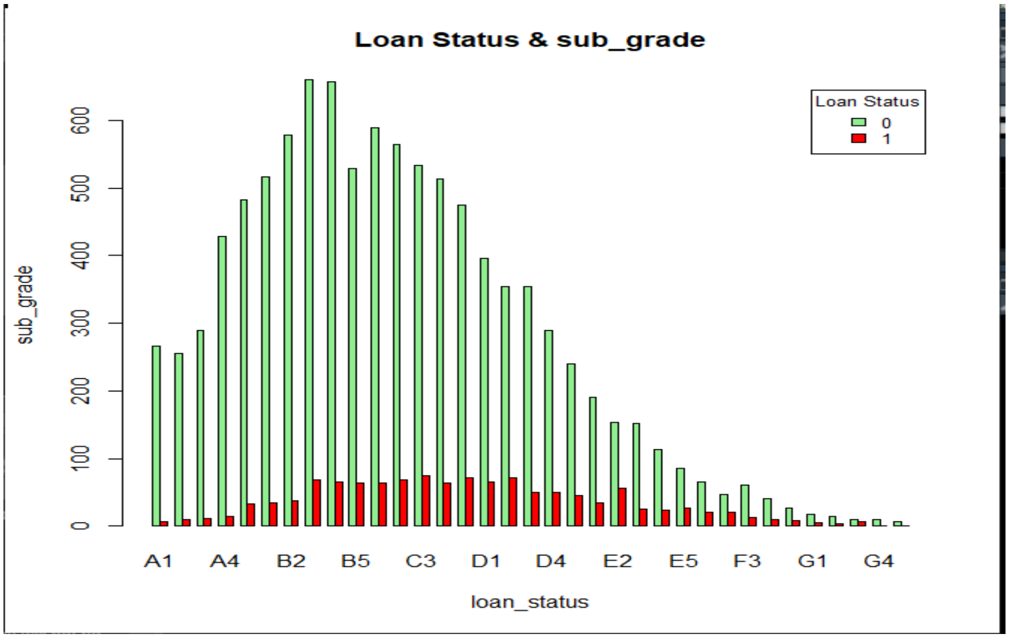
Loan Status and Purpose



Loan Status and Grade



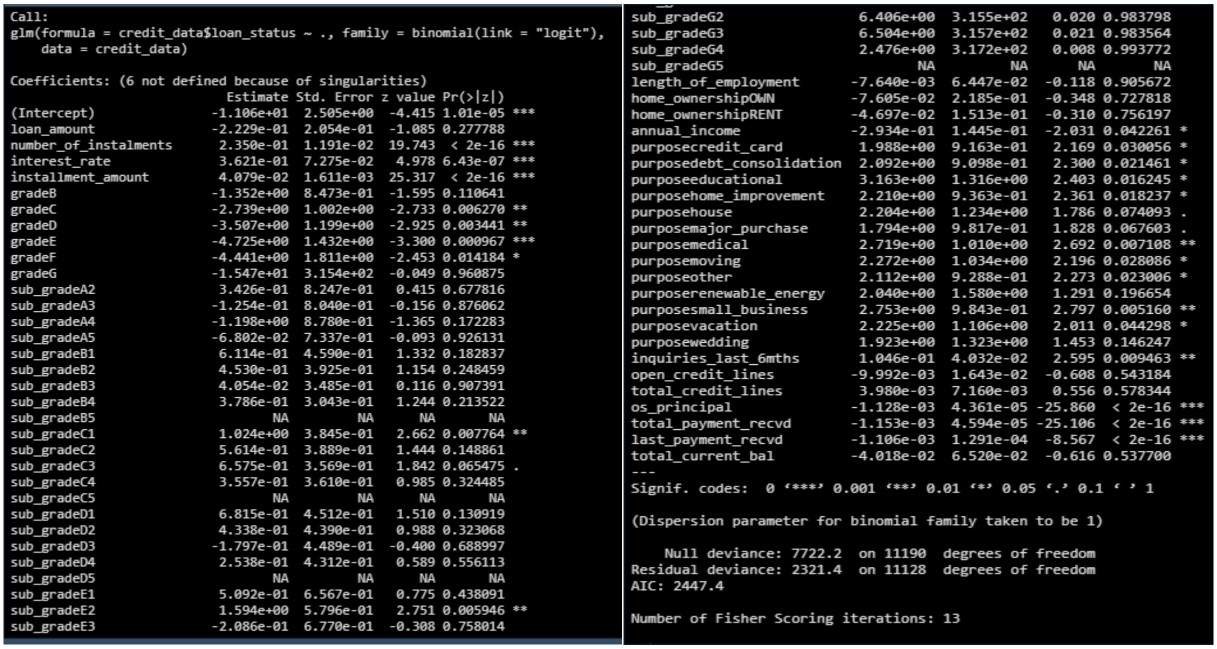
Loan Status and Sub-Grade



BUILDING THE MODEL

Logistic regression models are generally built using a 5% level of significance and insignificant variables will be discarded in the subsequent iterations.

Summary for the Initial Model

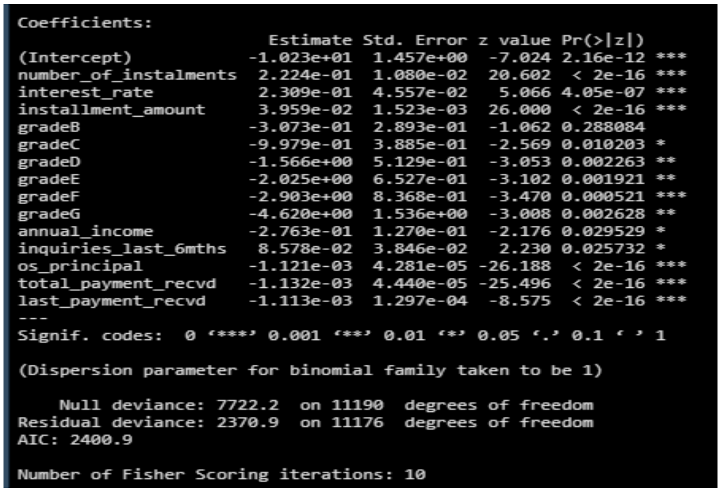


```
Call:
glm(formula = credit_data$loan_status ~ ., family = binomial(link = "logit"),
    data = credit_data)

Coefficients: (6 not defined because of singularities)
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.106e+01  2.505e+00  -4.415 1.01e-05 ***
loan_amount              -2.229e-01  2.054e-01  -1.085 0.277788
number_of_instalments     2.350e-01  1.191e-02  19.743  < 2e-16 ***
interest_rate             3.621e-01  7.275e-02   4.978 6.43e-07 ***
installment_amount        4.079e-02  1.611e-03  25.317  < 2e-16 ***
gradeB                   -1.352e+00  8.473e-01  -1.595 0.110641
gradeC                   -2.739e+00  1.002e+00  -2.733 0.006270 **
gradeD                   -3.507e+00  1.199e+00  -2.925 0.003441 **
gradeE                   -4.725e+00  1.432e+00  -3.300 0.000967 ***
gradeF                   -4.441e+00  1.811e+00  -2.453 0.014184 *
gradeG                   -1.547e+01  3.154e+02  -0.049 0.960875
sub_gradeA2               3.426e-01  8.247e-01   0.415 0.677816
sub_gradeA3              -1.254e-01  8.040e-01  -0.156 0.876062
sub_gradeA4              -1.198e+00  8.780e-01  -1.365 0.172283
sub_gradeA5              -6.802e-02  7.337e-01  -0.093 0.926131
sub_gradeB1               6.114e-01  4.590e-01   1.332 0.182837
sub_gradeB2               4.530e-01  3.925e-01   1.154 0.248459
sub_gradeB3               4.054e-02  3.485e-01   0.116 0.907391
sub_gradeB4               3.786e-01  3.043e-01   1.244 0.213522
sub_gradeB5                      NA         NA      NA       NA
sub_gradeC1               1.024e+00  3.845e-01   2.662 0.007764 **
sub_gradeC2               5.614e-01  3.889e-01   1.444 0.148861
sub_gradeC3               6.575e-01  3.569e-01   1.842 0.065475 .
sub_gradeC4               3.557e-01  3.610e-01   0.985 0.324485
sub_gradeC5                      NA         NA      NA       NA
sub_gradeD1               6.815e-01  4.512e-01   1.510 0.130919
sub_gradeD2               4.338e-01  4.390e-01   0.988 0.323068
sub_gradeD3              -1.797e-01  4.489e-01  -0.400 0.688997
sub_gradeD4               2.538e-01  4.312e-01   0.589 0.556113
sub_gradeD5                      NA         NA      NA       NA
sub_gradeE1               5.092e-01  6.567e-01   0.775 0.438091
sub_gradeE2               1.594e+00  5.796e-01   2.751 0.005946 **
sub_gradeE3              -2.086e-01  6.770e-01  -0.308 0.758014
```

```
sub_gradeG2               6.406e+00  3.155e+02   0.020 0.983798
sub_gradeG3               6.504e+00  3.157e+02   0.021 0.983564
sub_gradeG4               2.476e+00  3.172e+02   0.008 0.993772
sub_gradeG5                      NA         NA      NA       NA
length_of_employment     -7.640e-03  6.447e-02  -0.118 0.905672
home_ownershipOWN        -7.605e-02  2.185e-01  -0.348 0.727818
home_ownershipRENT       -4.697e-02  1.513e-01  -0.310 0.756197
annual_income            -2.934e-01  1.445e-01  -2.031 0.042261 *
purposecredit_card        1.988e+00  9.163e-01   2.169 0.030056 *
purposedebt_consolidation 2.092e+00  9.098e-01   2.300 0.021461 *
purposeeducational        3.163e+00  1.316e+00   2.403 0.016245 *
purposehome_improvement   2.210e+00  9.363e-01   2.361 0.018237 *
purposehouse              2.204e+00  1.234e+00   1.786 0.074093 .
purposemajor_purchase     1.794e+00  9.817e-01   1.828 0.067603 .
purposemedical            2.719e+00  1.010e+00   2.692 0.007108 **
purposemoving             2.272e+00  1.034e+00   2.196 0.028086 *
purposeother              2.112e+00  9.288e-01   2.273 0.023006 *
purposerenewable_energy   2.040e+00  1.580e+00   1.291 0.196654
purposesmall_business     2.753e+00  9.843e-01   2.797 0.005160 **
purposevacation           2.225e+00  1.106e+00   2.011 0.044298 *
purposewedding            1.923e+00  1.323e+00   1.453 0.146247
inquiries_last_6mths      1.046e-01  4.032e-02   2.595 0.009463 **
open_credit_lines        -9.992e-03  1.643e-02  -0.608 0.543184
total_credit_lines        3.980e-03  7.160e-03   0.556 0.578344
os_principal             -1.128e-03  4.361e-05 -25.860  < 2e-16 ***
total_payment_recvd      -1.153e-03  4.594e-05 -25.106  < 2e-16 ***
last_payment_recvd       -1.106e-03  1.291e-04  -8.567  < 2e-16 ***
total_current_bal        -4.018e-02  6.520e-02  -0.616 0.537700
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7722.2  on 11190  degrees of freedom
Residual deviance: 2321.4  on 11128  degrees of freedom
AIC: 2447.4

Number of Fisher Scoring iterations: 13
```

The subsequent models were refined by excluding variables with a significance level greater than 5%. The variables removed in this process included loan amount, sub grade, length of employment, home ownership, purpose, open credit lines, total credit lines, and total current balance.

After removing columns

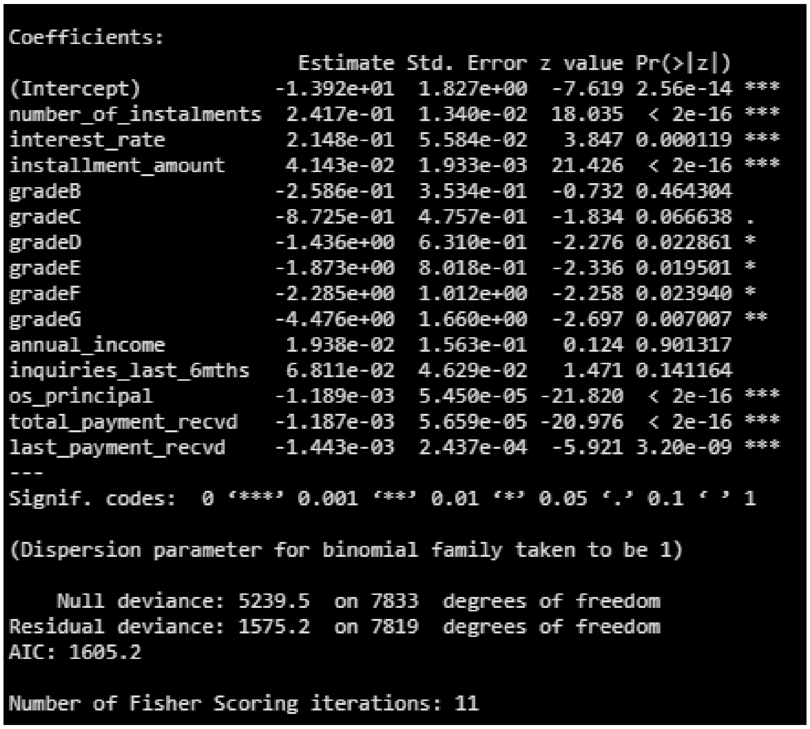


```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.023e+01  1.457e+00  -7.024 2.16e-12 ***
number_of_instalments  2.224e-01  1.080e-02  20.602  < 2e-16 ***
interest_rate          2.309e-01  4.557e-02   5.066 4.05e-07 ***
installment_amount     3.959e-02  1.523e-03  26.000  < 2e-16 ***
gradeB                -3.073e-01  2.893e-01  -1.062 0.288084
gradeC                -9.979e-01  3.885e-01  -2.569 0.010203 *
gradeD                -1.566e+00  5.129e-01  -3.053 0.002263 **
gradeE                -2.025e+00  6.527e-01  -3.102 0.001921 **
gradeF                -2.903e+00  8.368e-01  -3.470 0.000521 ***
gradeG                -4.620e+00  1.536e+00  -3.008 0.002628 **
annual_income         -2.763e-01  1.270e-01  -2.176 0.029529 *
inquiries_last_6mths   8.578e-02  3.846e-02   2.230 0.025732 *
os_principal          -1.121e-03  4.281e-05 -26.188  < 2e-16 ***
total_payment_recvd   -1.132e-03  4.440e-05 -25.496  < 2e-16 ***
last_payment_recvd    -1.113e-03  1.297e-04  -8.575  < 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: 7722.2  on 11190  degrees of freedom
Residual deviance: 2370.9  on 11176  degrees of freedom
AIC: 2400.9

Number of Fisher Scoring iterations: 10
```

Data Preparation

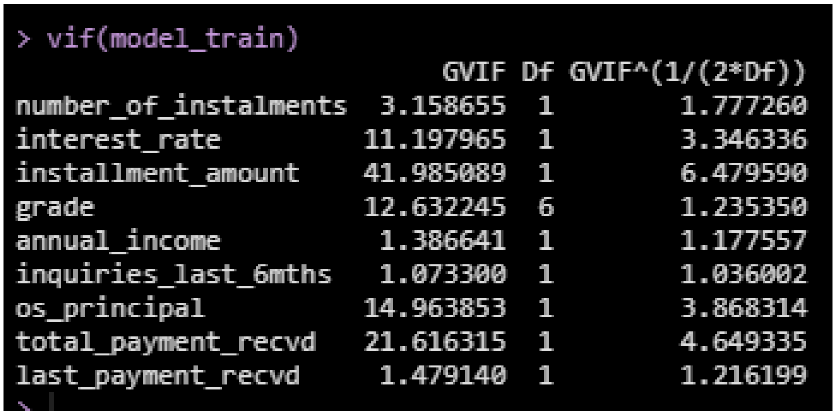
Further moving ahead with the model building process, the data was split into train and test sets by stratified sampling. It is divided in 70:30 ratio where 70% observations are contained by the training set whereas 30% are contained in the testing set.

Summary for Model built on Train Set



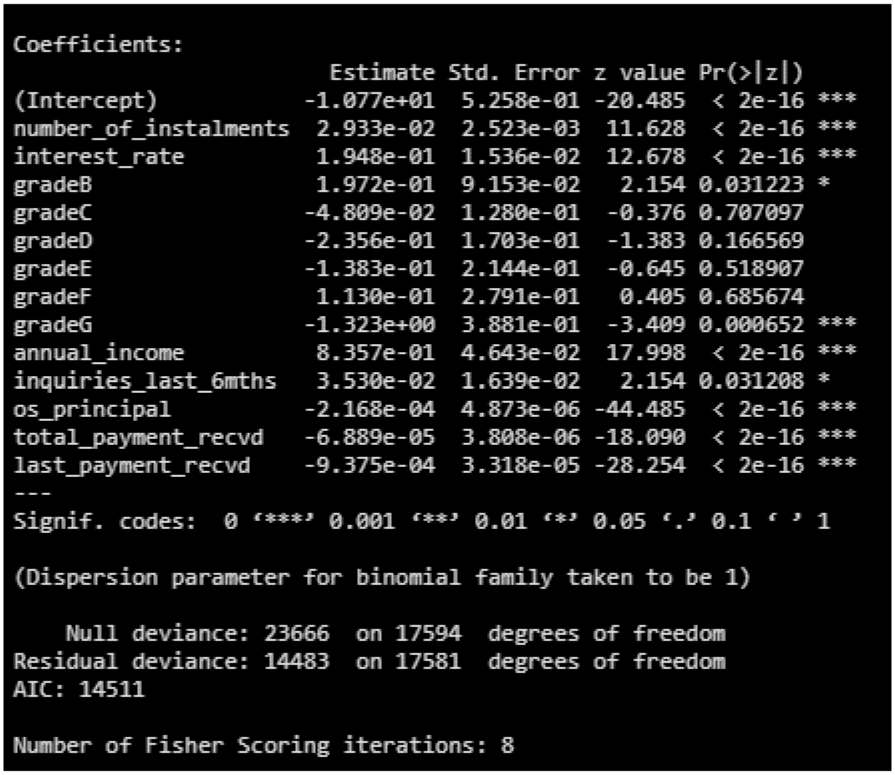
```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.392e+01  1.827e+00  -7.619 2.56e-14 ***
number_of_instalments  2.417e-01  1.340e-02  18.035  < 2e-16 ***
interest_rate          2.148e-01  5.584e-02   3.847 0.000119 ***
installment_amount     4.143e-02  1.933e-03  21.426  < 2e-16 ***
gradeB                -2.586e-01  3.534e-01  -0.732 0.464304
gradeC                -8.725e-01  4.757e-01  -1.834 0.066638 .
gradeD                -1.436e+00  6.310e-01  -2.276 0.022861 *
gradeE                -1.873e+00  8.018e-01  -2.336 0.019501 *
gradeF                -2.285e+00  1.012e+00  -2.258 0.023940 *
gradeG                -4.476e+00  1.660e+00  -2.697 0.007007 **
annual_income          1.938e-02  1.563e-01   0.124 0.901317
inquiries_last_6mths   6.811e-02  4.629e-02   1.471 0.141164
os_principal          -1.189e-03  5.450e-05 -21.820  < 2e-16 ***
total_payment_recvd   -1.187e-03  5.659e-05 -20.976  < 2e-16 ***
last_payment_recvd    -1.443e-03  2.437e-04  -5.921 3.20e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5239.5  on 7833  degrees of freedom
Residual deviance: 1575.2  on 7819  degrees of freedom
AIC: 1605.2

Number of Fisher Scoring iterations: 11
```

Checking Multicollinearity



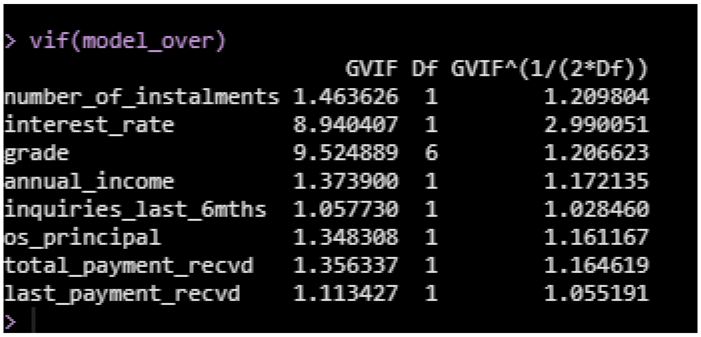
```
> vif(model_train)
                           GVIF Df GVIF^(1/(2*Df))
number_of_instalments  3.158655  1        1.777260
interest_rate         11.197965  1        3.346336
installment_amount    41.985089  1        6.479590
grade                 12.632245  6        1.235350
annual_income          1.386641  1        1.177557
inquiries_last_6mths   1.073300  1        1.036002
os_principal          14.963853  1        3.868314
total_payment_recvd   21.616315  1        4.649335
last_payment_recvd     1.479140  1        1.216199
>
```

Summary for Model built on Over-Sampled Data



```
Coefficients:
                         Estimate Std. Error z value Pr(>|z|)    
(Intercept)            -1.077e+01  5.258e-01 -20.485  < 2e-16 ***
number_of_instalments   2.933e-02  2.523e-03  11.628  < 2e-16 ***
interest_rate           1.948e-01  1.536e-02  12.678  < 2e-16 ***
gradeB                  1.972e-01  9.153e-02   2.154 0.031223 *  
gradeC                 -4.809e-02  1.280e-01  -0.376 0.707097    
gradeD                 -2.356e-01  1.703e-01  -1.383 0.166569    
gradeE                 -1.383e-01  2.144e-01  -0.645 0.518907    
gradeF                  1.130e-01  2.791e-01   0.405 0.685674    
gradeG                 -1.323e+00  3.881e-01  -3.409 0.000652 ***
annual_income           8.357e-01  4.643e-02  17.998  < 2e-16 ***
inquiries_last_6mths    3.530e-02  1.639e-02   2.154 0.031208 *  
os_principal           -2.168e-04  4.873e-06 -44.485  < 2e-16 ***
total_payment_recvd    -6.889e-05  3.808e-06 -18.090  < 2e-16 ***
last_payment_recvd     -9.375e-04  3.318e-05 -28.254  < 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: 23666  on 17594  degrees of freedom
Residual deviance: 14483  on 17581  degrees of freedom
AIC: 14511

Number of Fisher Scoring iterations: 8
```

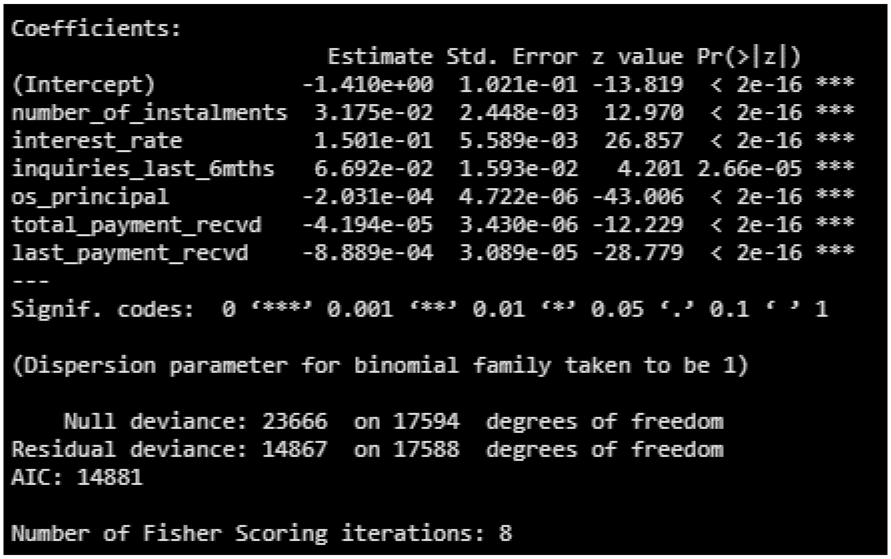
Checking Multicollinearity



```
> vif(model_over)
                          GVIF Df GVIF^(1/(2*Df))
number_of_instalments 1.463626  1        1.209804
interest_rate         8.940407  1        2.990051
grade                 9.524889  6        1.206623
annual_income         1.373900  1        1.172135
inquiries_last_6mths  1.057730  1        1.028460
os_principal          1.348308  1        1.161167
total_payment_recvd   1.356337  1        1.164619
last_payment_recvd    1.113427  1        1.055191
> 
```

- The model was checked for insignificant variables above 5 % level of significance were removed from the model.
- We will Purpose and Grade variable for the next iterations as they proved to be insignificant at 5% level.
- No Multicollinearity present after removing these variables.

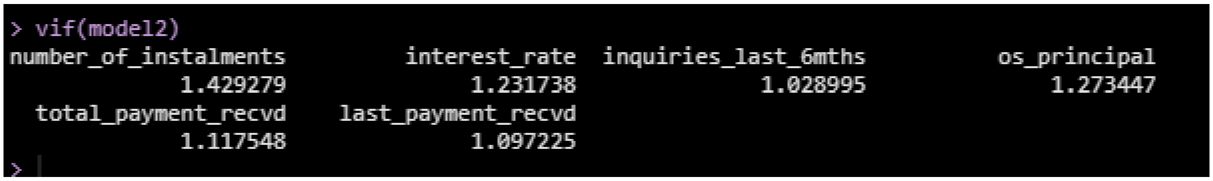
Summary for Model built without Grade and Annual Income



```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -1.410e+00  1.021e-01 -13.819  < 2e-16 ***
number_of_instalments  3.175e-02  2.448e-03  12.970  < 2e-16 ***
interest_rate          1.501e-01  5.589e-03  26.857  < 2e-16 ***
inquiries_last_6mths   6.692e-02  1.593e-02   4.201 2.66e-05 ***
os_principal          -2.031e-04  4.722e-06 -43.006  < 2e-16 ***
total_payment_recvd   -4.194e-05  3.430e-06 -12.229  < 2e-16 ***
last_payment_recvd    -8.889e-04  3.089e-05 -28.779  < 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: 23666  on 17594  degrees of freedom
Residual deviance: 14867  on 17588  degrees of freedom
AIC: 14881

Number of Fisher Scoring iterations: 8
```

The new model shows a higher AIC value although, is much more significant as compared to the previous one as this has fewer insignificant variables statistically.

Checking Multicollinearity

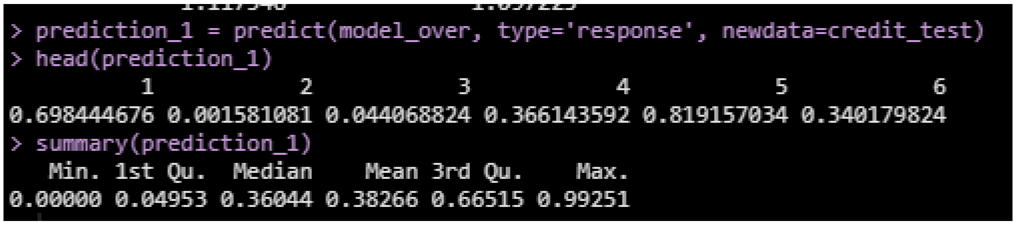


```
> vif(model2)
number_of_instalments         interest_rate  inquiries_last_6mths          os_principal 
             1.429279              1.231738              1.028995              1.273447 
  total_payment_recvd    last_payment_recvd 
             1.117548              1.097225 
> 
```

The model demonstrates outstanding results, indicating the absence of multicollinearity, especially for variables with VIF values exceeding 2.

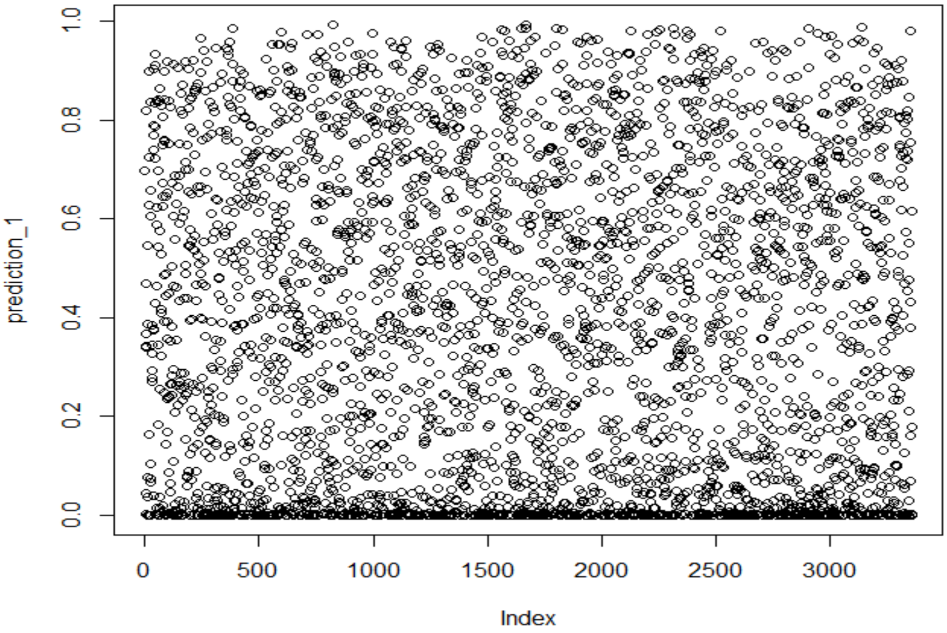
FITTING THE MODEL

The model demonstrates outstanding results, indicating the absence of multicollinearity, especially for variables with VIF values exceeding 2.



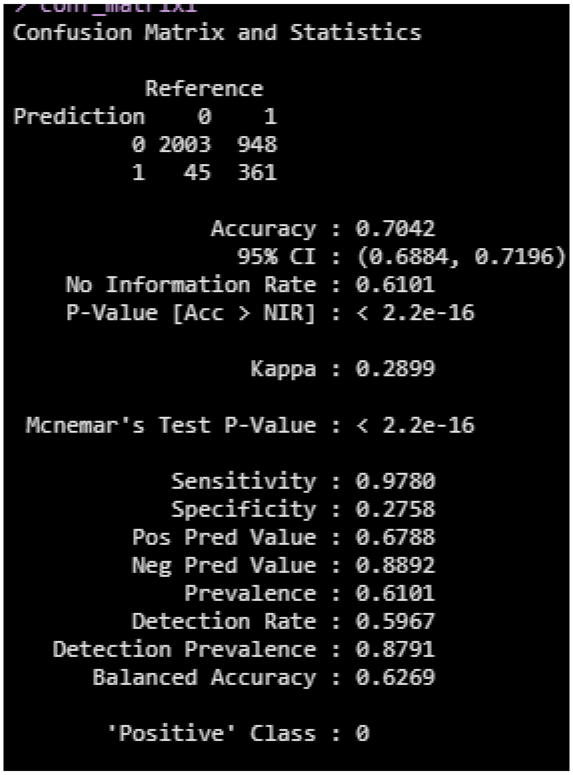
```
> prediction_1 = predict(model_over, type='response', newdata=credit_test)
> head(prediction_1)
          1           2           3           4           5           6 
0.698444676 0.001581081 0.044068824 0.366143592 0.819157034 0.340179824 
> summary(prediction_1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.00000 0.04953 0.36044 0.38266 0.66515 0.99251 
```

Scatter plot for the predicted values



BUILDING THE CONFUSION MATRIX

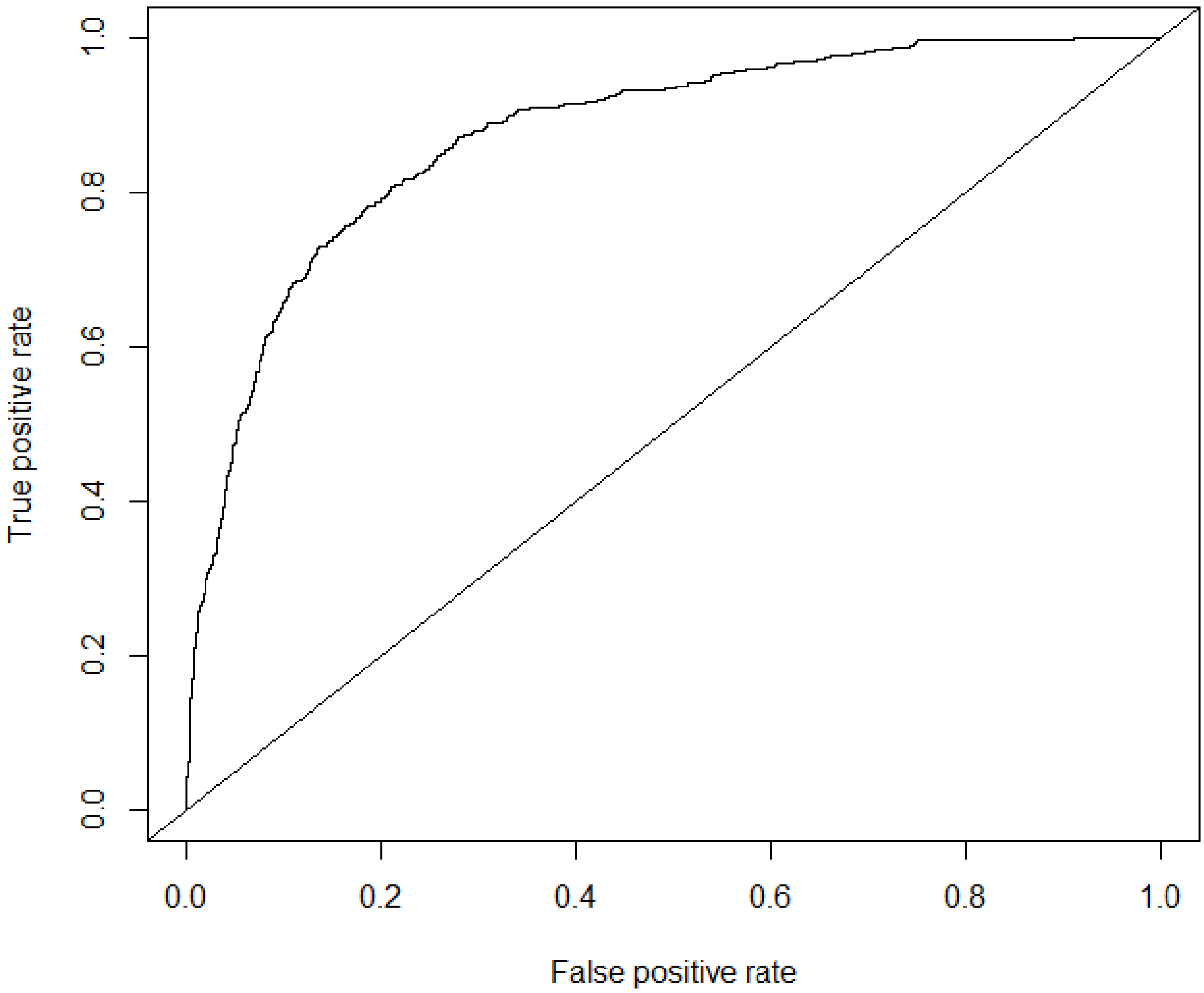
ON THE TESTING SET (OUT-SAMPLE)



```
> conf_matrix1
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2003  948
         1   45  361

               Accuracy : 0.7042
                 95% CI : (0.6884, 0.7196)
    No Information Rate : 0.6101
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2899

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9780
            Specificity : 0.2758
         Pos Pred Value : 0.6788
         Neg Pred Value : 0.8892
             Prevalence : 0.6101
         Detection Rate : 0.5967
   Detection Prevalence : 0.8791
      Balanced Accuracy : 0.6269

       'Positive' Class : 0
```

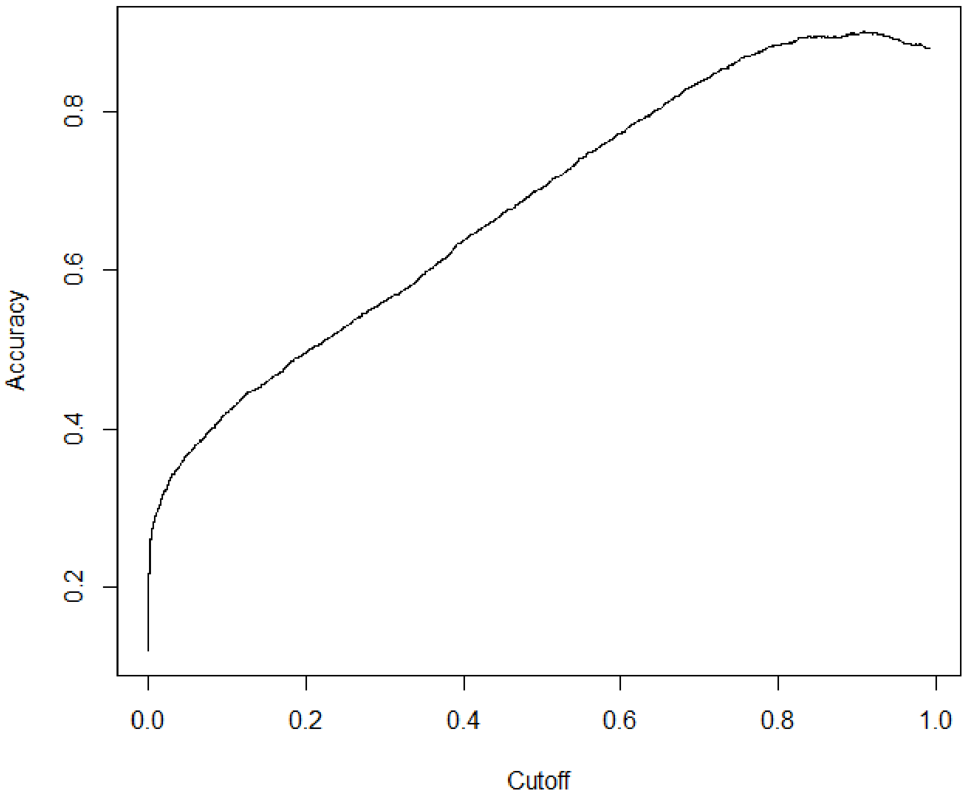
CONSTRUCTING THE ROC CURVE

(OUT SAMPLE)



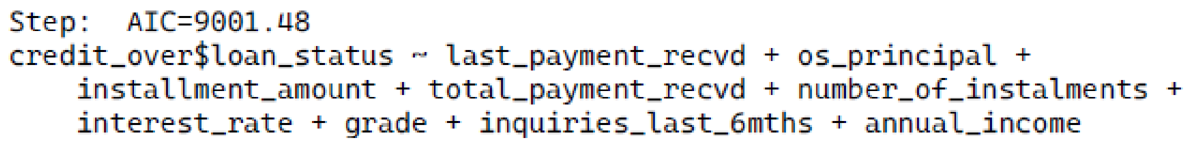
- The area under the ROC curve approximately comes out to be 0.9024116 which clearly is high.
- The higher the ROC the better is the model.

GRAPH FOR ACCURACY OF PREDICTED VALUE (OUT SAMPLE)



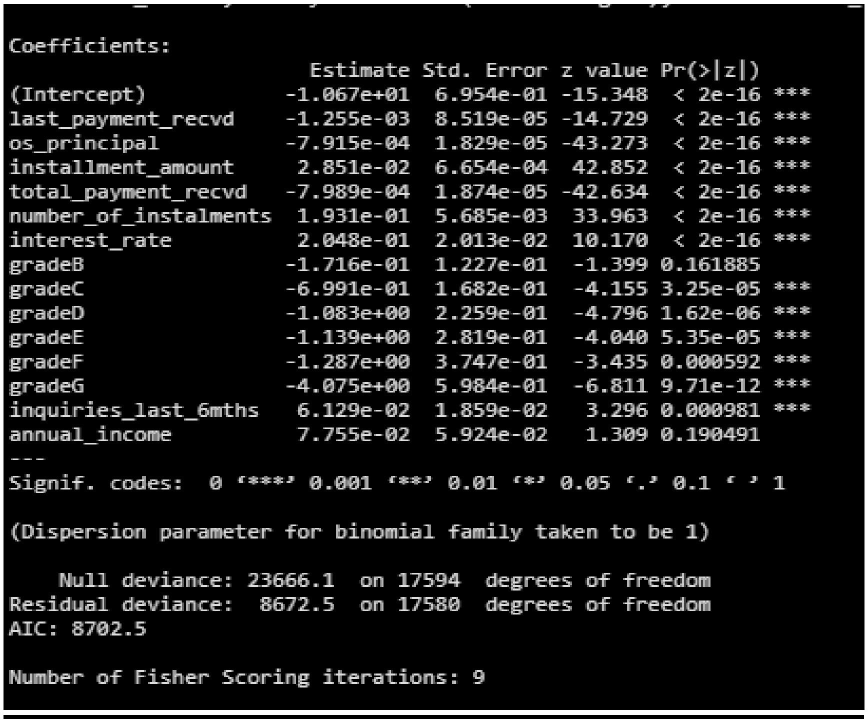
The graph is left skewed which means the model demonstrates relatively high accuracy, correctly predicting class 0 in 1913 instances and class 1 in 374 instances. The high number of true negatives (1913) suggests that the model is proficient in identifying instances of class 0. However, the presence of false positives (1043) indicates instances where the model incorrectly predicted class 1. The false negatives (27) suggest instances wh

SELECTING THE BEST MODEL USING STEPAIC



```
Step:  AIC=9001.48
credit_over$loan_status ~ last_payment_recvd + os_principal +
    installment_amount + total_payment_recvd + number_of_instalments +
    interest_rate + grade + inquiries_last_6mths + annual_income
```

SUMMARY FOR THE FINAL MODEL

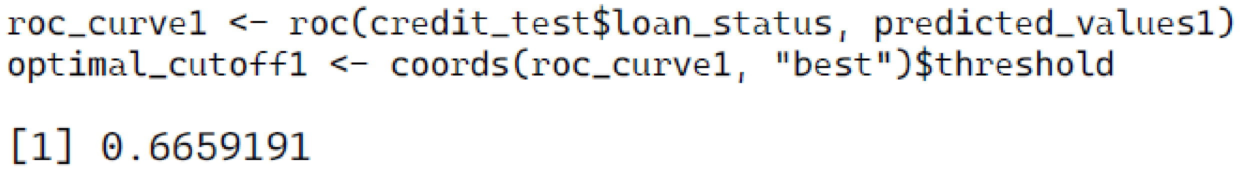


```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.067e+01  6.954e-01 -15.348  < 2e-16 ***
last_payment_recvd    -1.255e-03  8.519e-05 -14.729  < 2e-16 ***
os_principal          -7.915e-04  1.829e-05 -43.273  < 2e-16 ***
installment_amount     2.851e-02  6.654e-04  42.852  < 2e-16 ***
total_payment_recvd   -7.989e-04  1.874e-05 -42.634  < 2e-16 ***
number_of_instalments  1.931e-01  5.685e-03  33.963  < 2e-16 ***
interest_rate          2.048e-01  2.013e-02  10.170  < 2e-16 ***
gradeB                -1.716e-01  1.227e-01  -1.399 0.161885
gradeC                -6.991e-01  1.682e-01  -4.155 3.25e-05 ***
gradeD                -1.083e+00  2.259e-01  -4.796 1.62e-06 ***
gradeE                -1.139e+00  2.819e-01  -4.040 5.35e-05 ***
gradeF                -1.287e+00  3.747e-01  -3.435 0.000592 ***
gradeG                -4.075e+00  5.984e-01  -6.811 9.71e-12 ***
inquiries_last_6mths   6.129e-02  1.859e-02   3.296 0.000981 ***
annual_income          7.755e-02  5.924e-02   1.309 0.190491
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23666.1  on 17594  degrees of freedom
Residual deviance:  8672.5  on 17580  degrees of freedom
AIC: 8702.5

Number of Fisher Scoring iterations: 9
```

- The model exhibits no variables that are deemed insignificant at the 5% significance level.
- The observed relationship where the residual deviance is lower than the null deviance suggests a favourable fit for the model.
- Despite the increase in AIC caused by oversampling, it effectively addresses data imbalance, resulting in more meaningful models for practical applications.

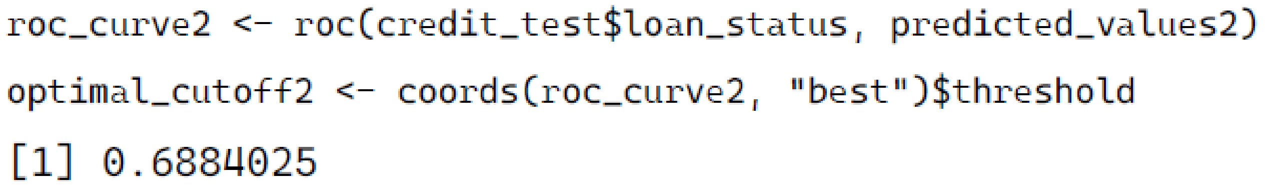
CHOOSING AN OPTIMAL THRESHOLD FOR THE MODEL

Out Sample



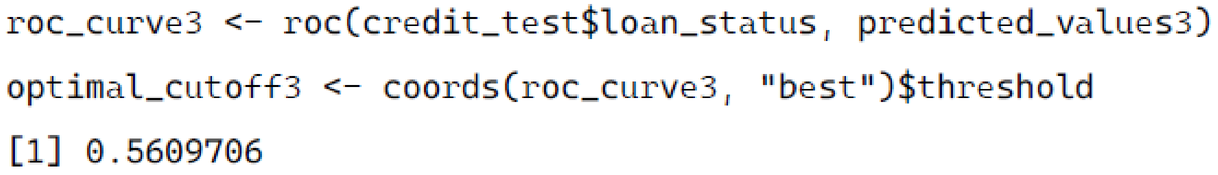
```
roc_curve1 <- roc(credit_test$loan_status, predicted_values1)
optimal_cutoff1 <- coords(roc_curve1, "best")$threshold

[1] 0.6659191
```

In Sample



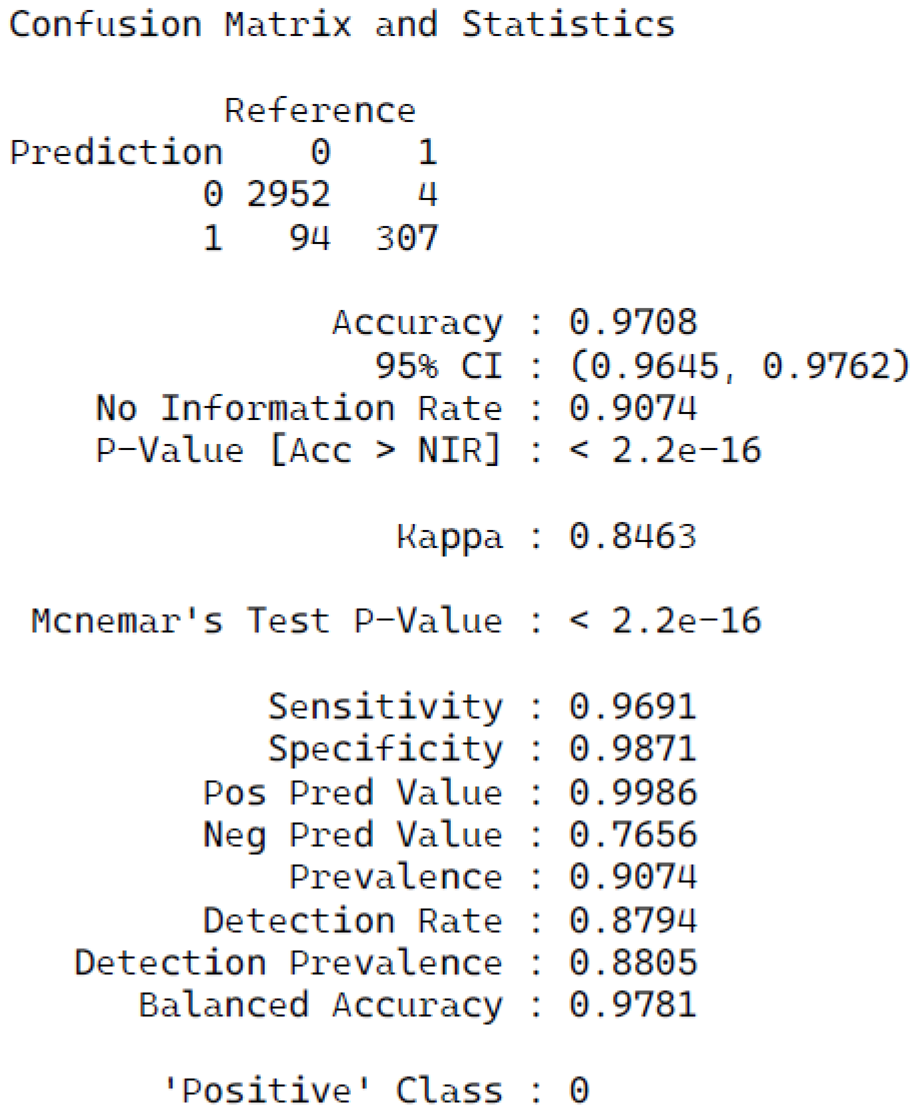
```
roc_curve2 <- roc(credit_test$loan_status, predicted_values2)

optimal_cutoff2 <- coords(roc_curve2, "best")$threshold

[1] 0.6884025
```

Final



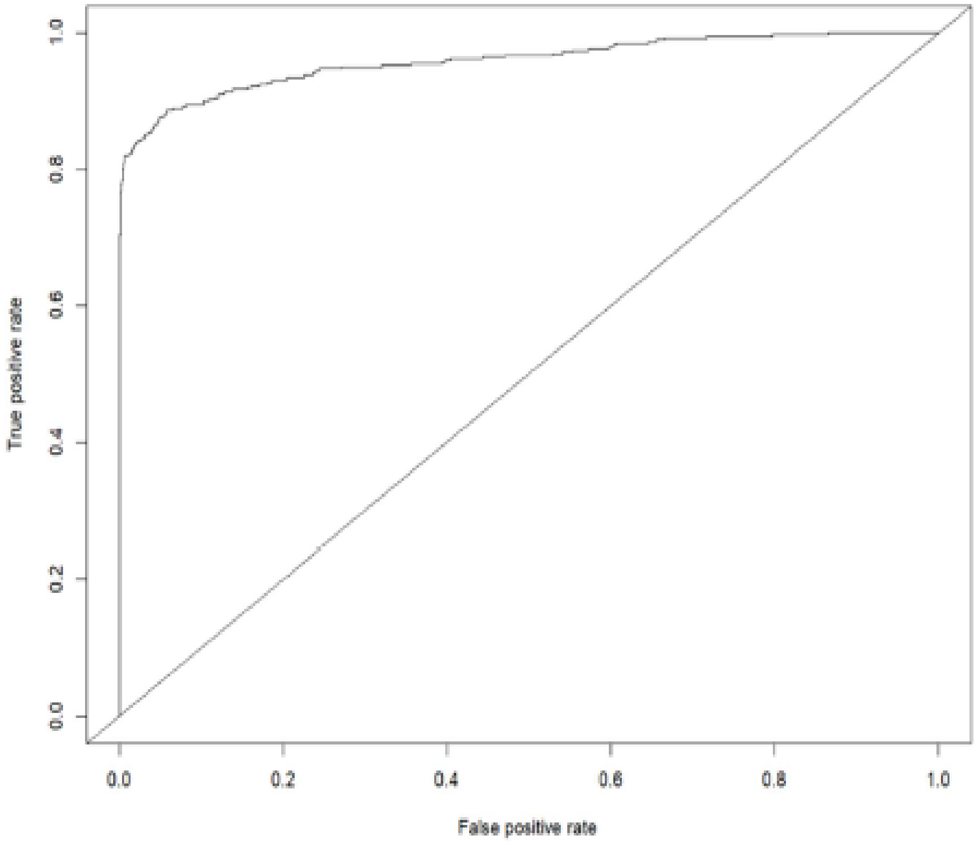
```
roc_curve3 <- roc(credit_test$loan_status, predicted_values3)

optimal_cutoff3 <- coords(roc_curve3, "best")$threshold

[1] 0.5609706
```

BUILDING THE CONFUSION MATRIX ON THE TESTING SET (IN-SAMPLE)



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2952    4
         1   94  307

               Accuracy : 0.9708
                 95% CI : (0.9645, 0.9762)
    No Information Rate : 0.9074
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8463

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9691
            Specificity : 0.9871
         Pos Pred Value : 0.9986
         Neg Pred Value : 0.7656
             Prevalence : 0.9074
         Detection Rate : 0.8794
   Detection Prevalence : 0.8805
      Balanced Accuracy : 0.9781

       'Positive' Class : 0
```

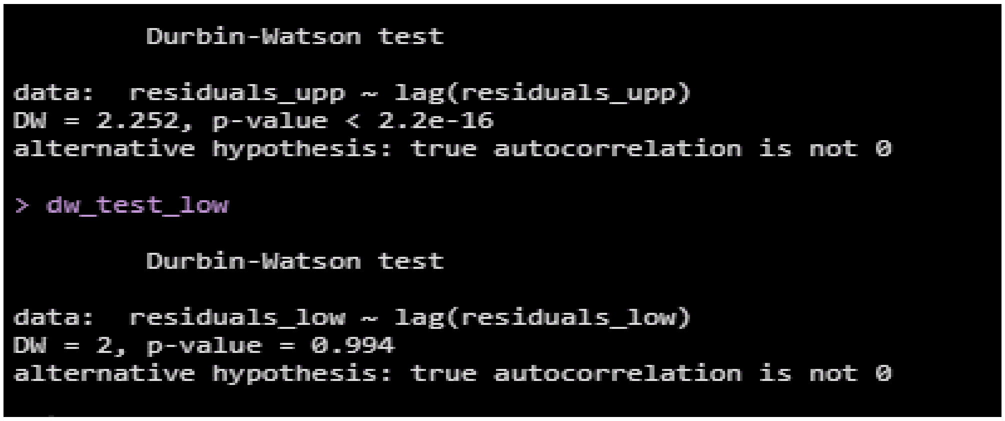
CONSTRUCTING ROC CURVE (IN-SAMPLE)



- The area under the ROC curve approximately comes out to be 0.959591 which clearly is high.
- The higher the ROC the better is the model.

ASSUMPTION TESTING

Checking Autocorrelation



```
        Durbin-Watson test

data:  residuals_upp ~ lag(residuals_upp)
DW = 2.252, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is not 0

> dw_test_low

        Durbin-Watson test

data:  residuals_low ~ lag(residuals_low)
DW = 2, p-value = 0.994
alternative hypothesis: true autocorrelation is not 0
```

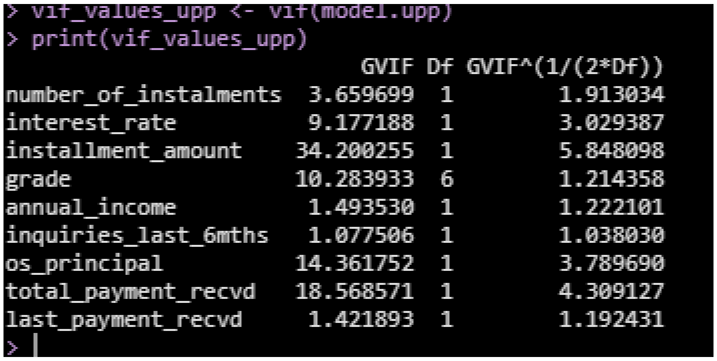
For upper model

the Durbin-Watson test on the residuals of the upper model yielded a statistic of 2.252 with a highly significant p-value (< 2.2e-16). The results strongly indicate the presence of positive autocorrelation in the residuals, suggesting that consecutive residuals are positively correlated.

For lower model

The Durbin-Watson test results for the lower model show a DW statistic of 2 with a p-value of 0.994. With such a high p-value, we do not have enough evidence to reject the null hypothesis. Therefore, we conclude that there is no significant autocorrelation in the residuals of the lower model, indicating that the residuals exhibit independence.

Checking



```
> vif_values_upp <- vif(model.upp)
> print(vif_values_upp)
                          GVIF Df GVIF^(1/(2*Df))
number_of_instalments  3.659699  1        1.913034
interest_rate          9.177188  1        3.029387
installment_amount    34.200255  1        5.848098
grade                 10.283933  6        1.214358
annual_income          1.493530  1        1.222101
inquiries_last_6mths   1.077506  1        1.038030
os_principal          14.361752  1        3.789690
total_payment_recvd   18.568571  1        4.309127
last_payment_recvd     1.421893  1        1.192431
> |
```

For upper model

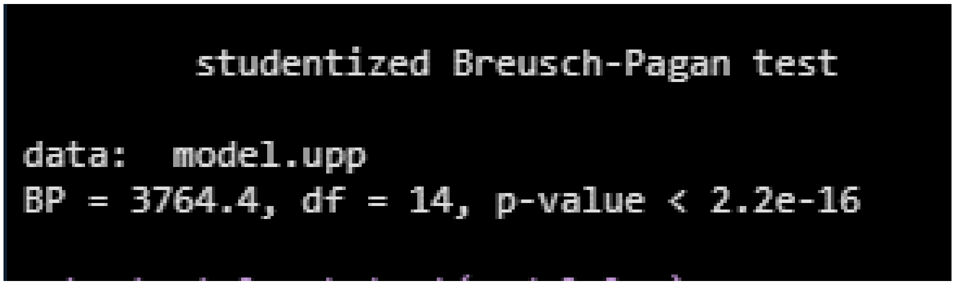
The variance inflation factor (GVIF) results indicate the following regarding multicollinearity:

→ Moderate multicollinearity for "Number of Installments" and "Grade."
→ High multicollinearity for "Interest Rate."

→ Severe multicollinearity for "Installment Amount," "Outstanding Principal," and "Total Payment Received."
→ Low multicollinearity for "Annual Income," "Inquiries Last 6 Months," and "Last Payment Received."

Addressing severe multicollinearity in certain predictors, especially "Interest Rate," "Installment Amount," "Outstanding Principal," and "Total Payment Received," may be crucial for model stability and interpretation. Consider techniques such as variable selection or regularization.

Checking Homocedascity



```
        studentized Breusch-Pagan test

data:  model.upp
BP = 3764.4, df = 14, p-value < 2.2e-16
```

For upper model

The studentized Breusch-Pagan test on the upper model yielded a highly significant p-value (< 2.2e-16), indicating the presence of heteroscedasticity. The assumption of homoscedasticity is violated, suggesting that the variance of the residuals is not constant across all levels of the predictor variables. Further analysis or model adjustments, such as heteroscedasticity-robust standard errors, may be require

CONCLUSION

The model exhibits satisfactory performance, as evidenced by the conducted statistical tests and the accuracy derived from its predictions. The assumptions made during the modeling align well with the outcomes, with both precision and sensitivity on the testing data demonstrating commendable results.

Despite facing challenges such as imbalances in the dataset, the model showcases resilience in handling complexities. While the logistic regression model performs well, there is an opportunity to explore alternative models, such as Decision Trees and Random Forest, which may provide more practical solutions.

In conclusion, the model's performance could be further enhanced with a larger dataset and fewer missing values. The need to make assumptions in the absence of complete information is acknowledged. Improved significance of variables, especially factors like loan amount, could contribute to a more robust modelling process.

END-TO-END PROCESS TO DEVELOP A MODEL
THE INTERNAL CREDIT SCORING MODEL
FOR THE RETAIL PORTFOLIO

1. Data Collection and Preparation:

- Collect relevant data for the retail portfolio, including variables such as loan amount, interest rate,
- annual income, payment history, and other relevant features.
- Cleanse and preprocess the data to handle missing values, outliers, and ensure data quality.

2. Exploratory Data Analysis (EDA):

- Conduct exploratory data analysis to understand the distribution of variables, identify patterns, and
- assess relationships between features.
- Visualize key insights using statistical plots and summary statistics.

3. Variable Selection:

- Choose variables that are likely to have a significant impact on credit risk, considering factors such as
- historical performance, economic indicators, and industry trends.
- Use statistical methods or domain knowledge to prioritize variables.

4. Data Splitting:

- Split the dataset into training and testing sets. The training set is used to train the model, while the testing set evaluates its performance on unseen data.

5. Model Selection:

- Choose an appropriate modelling technique for credit scoring. Common choices include logistic
- regression, decision trees, random forests, or gradient boosting.
- Consider the interpretability of the model and its ability to handle imbalanced data.

6. Model Training:

- Train the selected model on the training dataset, using historical credit performance as the target
- variable.
- Fine-tune hyperparameters to optimize the model's predictive performance.

7. Model Evaluation:

- Evaluate the model's performance on the testing dataset using metrics such as accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic (ROC) curve.
- Assess the model's ability to discriminate between good and bad credit risks.

8. Model Interpretation:

- Interpret the coefficients (for linear models) or feature importance (for tree-based models) to understand the impact of each variable on the credit score.
- Ensure that the model aligns with business goals and regulatory requirements.

9. Validation and Back testing:

- Validate the model's performance using out-of-sample data to ensure robustness.
- Conduct back testing by applying the model to historical data to assess how well it would have performed in the past.

10. Deployment:

- Deploy the model to the production environment, integrating it with the credit decision-making process.
- Implement monitoring mechanisms to track the model's performance over time.

11. Documentation and Reporting:

- Document the model development process, including data sources, variable definitions, and modelling methodologies.
- Generate regular reports on model performance and present findings to stakeholders.

12. Continuous Monitoring and Updating:

- Implement continuous monitoring of the model's performance in real-world conditions.
- Regularly update the model to incorporate new data and improve predictive accuracy over time.
- This end-to-end process ensures a comprehensive and systematic approach to developing, validating, and deploying an internal credit scoring model for

IMPORTANCE OF DATA CLEANSING, EXPANATORY ANALYSIS AND TRANSFORMATION OF VARIABLES

1. Data Cleaning:

- Handle Missing Values:
- Identify and analyse missing values in each variable.
- Impute missing values using appropriate techniques (e.g., mean imputation, median imputation, or
- advanced imputation methods).
- Outlier Detection and Treatment:
- Identify outliers using statistical methods or visualization tools.
- Decide whether to remove outliers or transform them based on the impact on the model.
- Duplicate Removal:
- Check for and remove duplicate records, if any, to avoid bias in the model.

2. Exploratory Data Analysis (EDA):

- Univariate Analysis:
- Examine the distribution of each variable to identify patterns and outliers.
- Generate summary statistics (mean, median, range) for continuous variables.
- Bivariate Analysis:
- Explore relationships between pairs of variables using scatter plots, correlation matrices, or other
- relevant visualizations.
- Assess the correlation between independent variables to identify multicollinearity.
- Target Variable Analysis:
- Examine the distribution of the target variable (e.g., loan default) to understand the class balance.
- Visualize the relationship between the target variable and key features.

3. Data Transformation:

- Variable Encoding:
- Convert categorical variables into numerical format using methods like one-hot encoding or label
- encoding.
- Handling Skewed Variables:
- Identify and transform skewed continuous variables using techniques like log transformation or Box-
- Cox transformation.
- Feature Engineering:
- Create new features that may capture relevant information for credit scoring.

- Generate interaction terms or polynomial features to capture complex relationships.
- Handling Time-Related Variables:
- If applicable, process and extract relevant information from time-related variables (e.g., length of
- employment) to make them suitable for modelling.

4. Data Standardization and Scaling:

- Standardize or normalize numerical features to ensure that variables with different scales have a
- similar impact on the model.
- Scaling methods include Min-Max scaling or Z-score normalization.

5. Handling Categorical Variables:

- Choose an appropriate method for handling categorical variables based on the modelling technique e.g., one-hot encoding, label encoding, or target encoding).
- Ensure that the encoding method aligns with the chosen modelling approach.

6. Data Splitting:

- Split the dataset into training and testing sets to facilitate model evaluation.
- Consider using techniques like stratified sampling to maintain the distribution of the target variable in both sets.

7. Documentation:

- Document the decisions made during data cleaning and transformation for transparency and
- reproducibility.
- Keep a record of any assumptions or imputations made during the process.

Effective data cleansing, exploratory analysis, and transformation lay the foundation for a reliable credit scoring model by ensuring that the model is trained on high-quality, representative data.

TYPES OF MODELLING METHODOLOGIES AND FRAMEWORKS THAT CAN BE USED FOR CREDIT RISK MODELLING

Various modeling methodologies and frameworks are employed in credit risk modeling, each with its own strengths and characteristics. Here are some common types:

1. Logistic Regression:
Explanation: A statistical method for binary classification, modeling the probability of an event (e.g., default) based on predictor variables.
Advantages: Simple, interpretable, and suitable for probability estimation.

2.Decision Trees:
Explanation: Divides the dataset based on input features, useful for both classification and regression tasks in credit scoring.
Advantages: Intuitive, easy to understand, and capable of capturing non-linear relationships.

3. Random Forest:
Explanation: Ensemble learning method that combines predictions from multiple decision trees, offering improved accuracy and robustness.
Advantages: Handles overfitting, provides variable importance measures, and is resistant to outliers.

4. Gradient Boosting Machines (GBM):
Explanation: Builds an ensemble of weak learners sequentially, correcting errors made by previous learners.
Advantages: High predictive accuracy, accommodates complex relationships and variable interactions.

5. Support Vector Machines (SVM):
Explanation: Finds a hyperplane that best separates data into different classes, applicable to linear and non-linear classification problems.
Advantages: Effective in high-dimensional spaces, versatile with different kernel functions.

6. Neural Networks:
Explanation: Deep learning models that capture complex patterns and relationships in data through interconnected layers of nodes.
Advantages: Ability to model intricate relationships, suitable for large and complex datasets

7. Scorecard Development (Traditional Credit Scoring):
Explanation: Involves assigning scores to different features based on their contribution to credit risk, widely accepted in the credit industry.
Advantages: Transparent, easy to understand.

8.Ensemble Models:
Explanation: Combine predictions from multiple models to improve overall performance. Examples include model stacking, blending, and bagging.
Advantages: Improved generalization, reduced overfitting, increased model robustness.

9. Bayesian Models:
Explanation: Incorporates prior knowledge and updates it based on observed data to make predictions. Can be applied to various model types.
Advantages: Incorporates uncertainty, useful for handling limited data.

10. Hybrid Approaches:
Explanation: Combine multiple modelling techniques to leverage the strengths of each approach. For example, combining logistic regression with decision trees.
Advantages: Improved robustness, better handling of diverse data characteristics.

The choice of modelling methodology depends on factors such as the nature of the data, business context, and the specific goals of the credit risk model. It's common to experiment with multiple approaches and choose the one that best suits the requirements and constraints of the credit risk modelling task.

CONTROL-CHECKS FOR OPTIMAL EXPLANATORY VARIABLES SELECTION

In the context of explanatory variable selection, commonly known as feature selection, several methods can be employed to identify the most relevant variables for your analysis. Here are some control checks and considerations:

Correlation Analysis:
Check for multicollinearity by examining correlations between explanatory variables. High correlations may indicate redundancy.

Variable Importance Measures:
Use methods like Random Forest or Gradient Boosting to assess the importance of each variable in predicting the outcome.

Stepwise Regression:
Perform stepwise regression to add or remove variables based on statistical criteria like AIC or BIC.

VIF (Variance Inflation Factor):
Calculate the VIF for each variable to identify high multicollinearity. High VIF values suggest that a variable is highly correlated with others.

Recursive Feature Elimination (RFE):
Use RFE algorithms to iteratively remove the least important variables based on model performance.

LASSO Regression:
Apply LASSO (Least Absolute Shrinkage and Selection Operator) regression to shrink some coefficients to zero, effectively performing variable selection.

Information Criteria:
Use information criteria such as AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion) to guide variable selection.

Domain Knowledge:
Leverage subject-matter expertise to identify and include variables that are theoretically relevant to the problem.

Cross-Validation:
Employ cross-validation techniques to assess how well your model generalizes to new data, helping you avoid overfitting.

Remember, the optimal method may depend on the specific characteristics of your data and the goals of your analysis. It's often a good practice to combine multiple approaches and validate your choices using appropriate performance metrics.

OTHER APPLICATIONS OF CREDIT RISK MODELS

Credit risk models, initially developed for assessing the likelihood of default on loans, have found applications beyond traditional credit scoring. Here are some other applications of credit risk models:

Fraud Detection:
Credit risk models can be adapted to identify potentially fraudulent activities. Unusual patterns in transactions or application data may signal fraudulent behavior.

Insurance Underwriting:
Similar risk assessment principles can be applied in insurance underwriting. Models help evaluate the risk associated with insuring individuals or entities, impacting premium rates and coverage decisions.

Supply Chain Finance:
Credit risk models are used to assess the creditworthiness of suppliers or buyers in supply chain finance. This helps optimize working capital and manage financial risk along the supply chain.

Trade Finance:
Assessing the credit risk of parties involved in international trade transactions can facilitate more informed decisions in trade finance, such as letters of credit and trade credit insurance.

Corporate Bond Rating:
Credit risk models are employed to evaluate the creditworthiness of corporate entities issuing bonds. This aids investors in making decisions regarding bond purchases and portfolio management.

Peer-to-Peer Lending:
In the context of peer-to-peer lending platforms, credit risk models assist in evaluating the risk associated with lending to individual borrowers, enabling more accurate interest rate determination.

Mortgage Underwriting:
Credit risk models are utilized in mortgage underwriting to assess the risk of default on mortgage loans. This helps mortgage lenders make informed decisions on loan approvals and interest rates.

Debt Collection Strategies:
Credit risk models inform debt collection strategies by identifying accounts most likely to default. This allows for the prioritization of collection efforts and the optimization of resource allocation.

Credit Limit Management:
For credit card issuers, credit risk models aid in setting appropriate credit limits for cardholders based on their creditworthiness, spending patterns, and payment history.

Regulatory Compliance:
Credit risk models play a crucial role in regulatory compliance, helping financial institutions meet requirements related to capital adequacy and risk management.

Microfinance:
Credit risk models are applied in microfinance to assess the creditworthiness of individuals and small businesses, facilitating responsible lending in underserved markets.

Investment Decision-Making:
Investors use credit risk models to evaluate the credit risk associated with various financial instruments, including bonds, asset-backed securities, and other debt instruments.
Rating Agencies:

Credit rating agencies use sophisticated credit risk models to assign credit ratings to various entities, providing investors with assessments of credit risk.
The versatility of credit risk models demonstrates their adaptability across diverse financial domains, helping organizations make informed decisions in risk management and financial planning.