BFS CAPSTONE

SUBMISSION

Group Name:

1. Surya Prakash Tripathi
2. Varun Ahalawat
3. Mukul Tamta
4. Bharath BS

CredX– Credit Card Provider

CredX is a leading credit card provider that gets thousands of credit card applicants every year. But in the past few years, it has experienced an increase in credit loss. The CEO believes that the best strategy to mitigate credit risk is to ‘acquire the right customers’.

In this project, we will help CredX in identifying the right customers using predictive models. Using past data of the bank’s applicants, we need to determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of the project.

Objectives:

1. Using past data of the bank’s applicants, determine the factors affecting credit risk, create strategies to mitigate the acquisition risk
2. Assessing the financial benefit of the project
3. Acquiring the right customers
4. Predicting the likelihood of default for the rejected candidates
5. Application scorecard development and deciding cutoff of the score (below which CredX would not grant credit cards to applicants.)

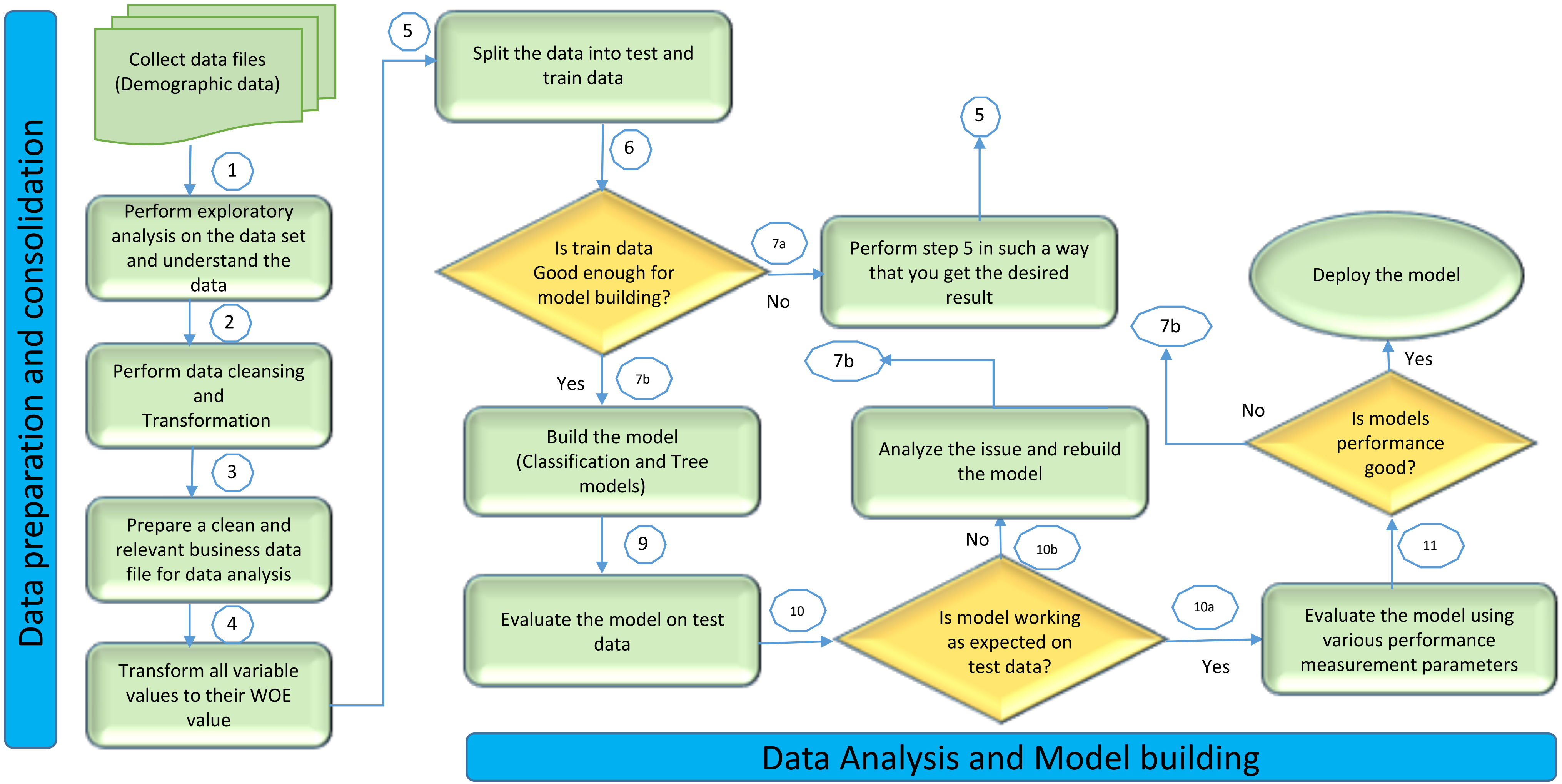
Data Understanding

There are two data sets in this project —

- Demographic and
- Credit bureau data

- **Demographic/application data:** This is obtained from the information provided by the applicants at the time of credit card application. It contains customer-level information on age, gender, income, marital status, etc.
- **Credit bureau:** This is taken from the credit bureau and contains variables such as 'number of times 30 DPD or worse in last 3/6/12 months', 'outstanding balance', 'number of trades', etc. Both files contain a performance tag which represents whether the applicant has gone 90 days past due or worse in the past 12-months (i.e. defaulted) after getting a credit card. In some cases, all the variables in the credit bureau data are zero and credit card utilization is missing. These represent cases in which there is a no-hit in the credit bureau. There are some cases wherein credit card utilization is missing. These are the cases in which the applicant does not have any other credit card.

Process flow of Demographic Data



Data preparation and consolidation

Collect data files (Demographic data)

1

Perform exploratory analysis on the data set and understand the data

2

Perform data cleansing and Transformation

3

Prepare a clean and relevant business data file for data analysis

4

Transform all variable values to their WOE value

5

Split the data into test and train data

6

Is train data Good enough for model building?

7a

No

Perform step 5 in such a way that you get the desired result

5

Yes

7b

Build the model (Classification and Tree models)

9

Evaluate the model on test data

10

Is model working as expected on test data?

No

10b

Analyze the issue and rebuild the model

7b

10a

Yes

Evaluate the model using various performance measurement parameters

11

Is models performance good?

No

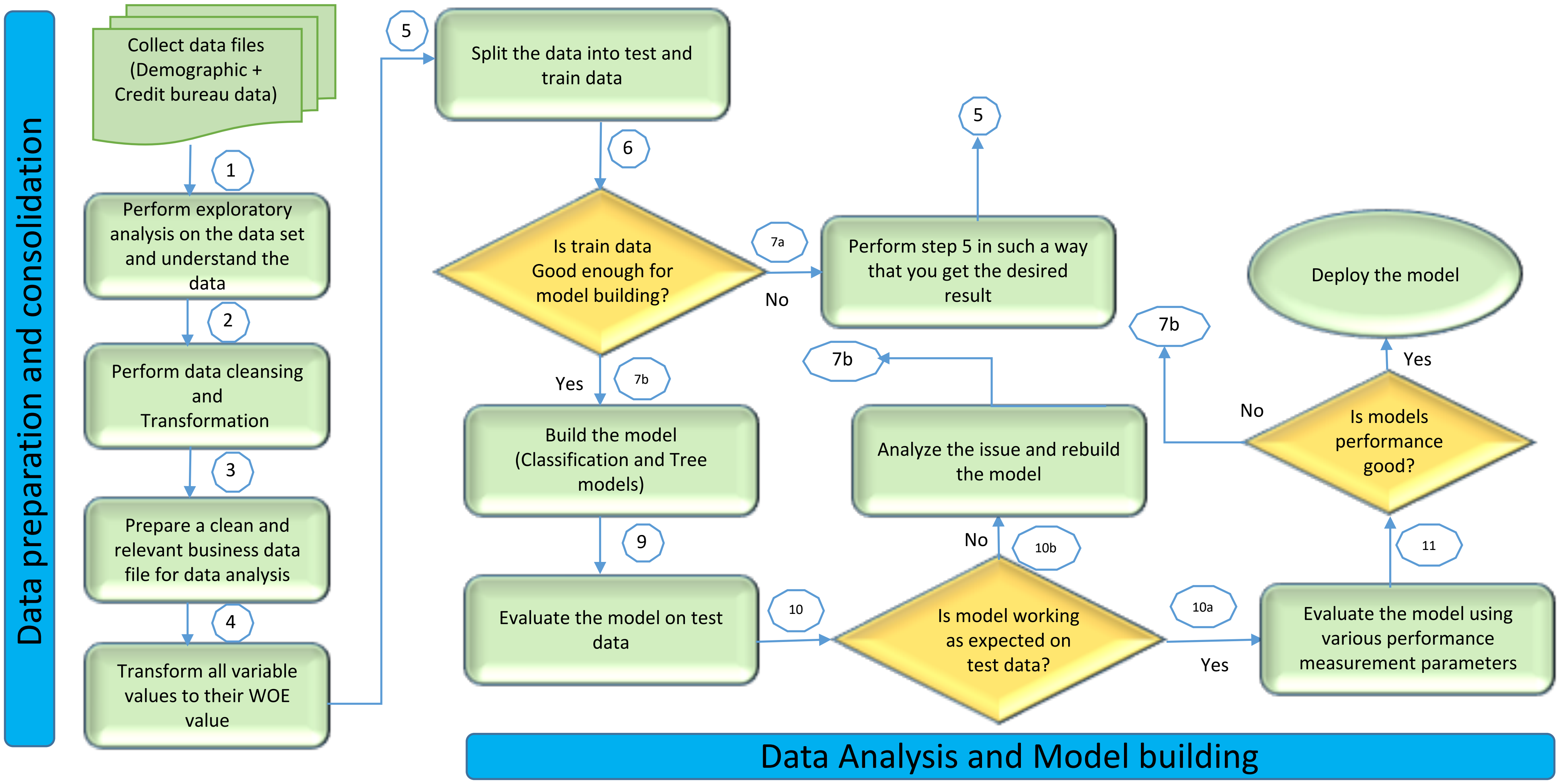
7b

Yes

Deploy the model

Data Analysis and Model building

Process flow of Demographic and Credit Bureau Data



Data preparation and consolidation
Collect data files (Demographic + Credit bureau data)
1
Perform exploratory analysis on the data set and understand the data
2
Perform data cleansing and Transformation
3
Prepare a clean and relevant business data file for data analysis
4
Transform all variable values to their WOE value
5
Split the data into test and train data
6
Is train data Good enough for model building?
7a
No
Perform step 5 in such a way that you get the desired result
5
Yes
7b
Build the model (Classification and Tree models)
9
Evaluate the model on test data
10
Is model working as expected on test data?
No
10b
Analyze the issue and rebuild the model
7b
Yes
10a
Evaluate the model using various performance measurement parameters
11
Is models performance good?
No
7b
Yes
Deploy the model
Data Analysis and Model building

Assumptions

1. Model is built using WOE values as WOE automatically impute missing values and treat outliers hence it’s not covered as part of EDA however we have analyzed this in EDA.
2. Rejected application data is only used for scorecard verification. NOT for EDA/ modelling.
3. Expected credit loss is calculated using following formula : **Expected credit loss = PD*EAD*LGD**

   We assumed **recovery as 25%** hence LGD = 1-.25 = .75 is used in above formula
4. We have removed the records for which age has negative values.
5. We have replaced records for which age’s value was less than 18 to 18.
6. As income cant be negative hence we have removed negative values of income from the dataset.
7. We have created a separate dataset for records which has NA values in performance tag(Target variable) and treated this dataset as rejected population.
8. Score card developed manually using following formula:

   factor = pdo / log(2)

   offset = points0 - factor * log(odds0)

   Score = offset - factor * logit

Data Understanding and Analysis

Data Understanding

- We have 2 different Dataset – Demographic data and Credit bureau data
- Demographic data contains customers demographic details like Age, Gender, Marital Status, Income, No of dependents, Education, Profession etc
- Credit bureau data contains customers transaction details related to CC or loan products like DPD details, Average CC utilisation, Outstanding Balance, No of inquiries in certain time frame etc.
- Distribution of default data is around 4.13% in the both the dataset.
- There were no strong predictor observed using EDA hence feature selection is done based on IV value and StepAIC function.

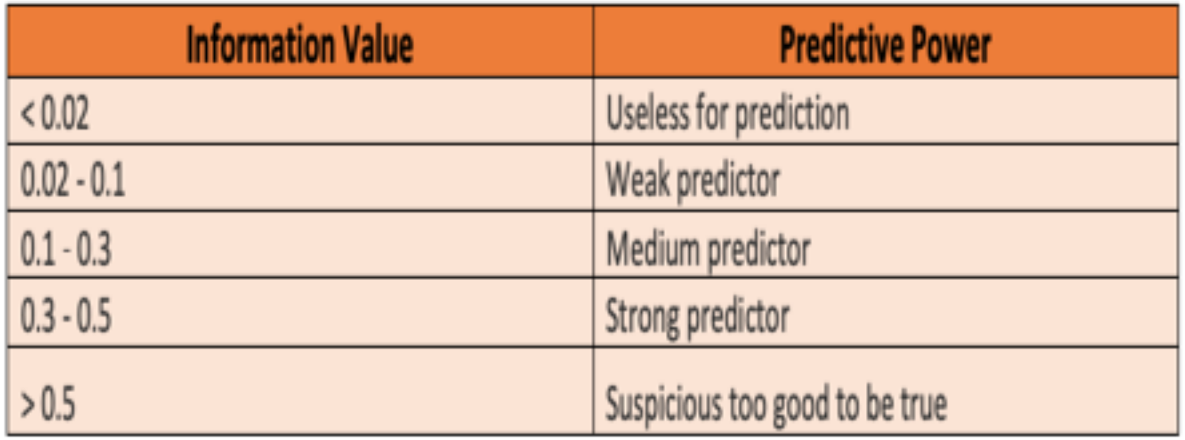
Data cleansing and transformation

Following are the rules which were applied while calculating the result and same will be applicable if any incremented data processed through the designed workflow:

- For performing any aggregation operation, we had excluded the rows which contains NA values or blank in amount column. These data were not removed from the master data frame.
- Demographic dataset and credit bureau data set is converted into their WOE values and model is built on the WOE dataset.
- Merging of the following two data sets : Demographic data set and credit bureau dataset into Master data set.
- Negative values of age and income is removed from the dataset and values of age which were less than 18 is replaced with value 18.
- NA values of target variable are removed from the final merged dataset and demographic dataset and these removed records were collated and used as separate dataset named as rejected population dataset.
- Outliers and NA values treated using WOE values.
- We have created bins of woe by checking monotonicity of the bins. If bins plot is not monotonic then bin size were reduced or increased to get the desired result.

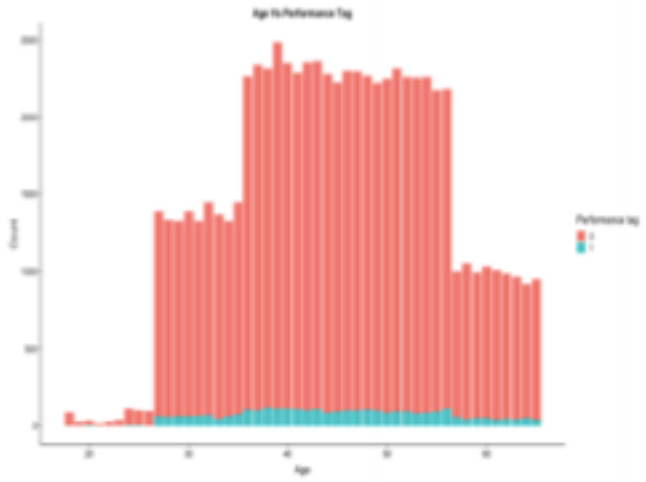
Variable importance using IV value

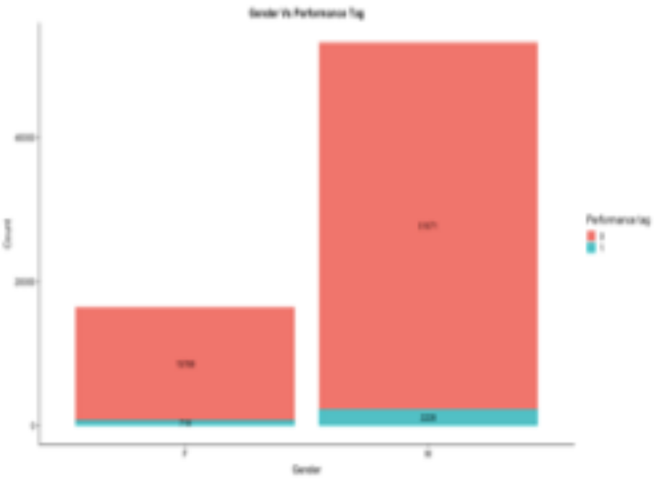
Variable	IV
Avgas.CC.Utilization.in.last.12.months	**0.31017091**
No.of.trades.opened.in.last.12.months	**0.298410295**
No.of.PL.trades.opened.in.last.12.months	**0.295992511**
No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.	**0.295662565**
Outstanding.Balance	**0.245301444**
No.of.times.30.DPD.or.worse.in.last.6.months	**0.241511513**
Total.No.of.Trades	**0.236972999**
No.of.PL.trades.opened.in.last.6.months	**0.219649332**
No.of.times.90.DPD.or.worse.in.last.12.months	**0.213808841**
No.of.times.60.DPD.or.worse.in.last.6.months	**0.205772033**
No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.	**0.205100412**
No.of.times.30.DPD.or.worse.in.last.12.months	0.198164749
No.of.trades.opened.in.last.6.months	0.186131091
No.of.times.60.DPD.or.worse.in.last.12.months	0.185410021
No.of.times.90.DPD.or.worse.in.last.6.months	0.160046408
No.of.months.in.current.residence	0.079052037
Income	0.04283854
No.of.months.in.current.company	0.021616865
Presence.of.open.home.loan	0.017351979
Age	0.003433062
No.of.dependents	0.002652335
Profession	0.00216856
Presence.of.open.auto.loan	0.001628672
Type.of.residence	0.000925666
Education	0.000787917
Gender	0.000340095
Marital.Status..at.the.time.of.application.	0.000097

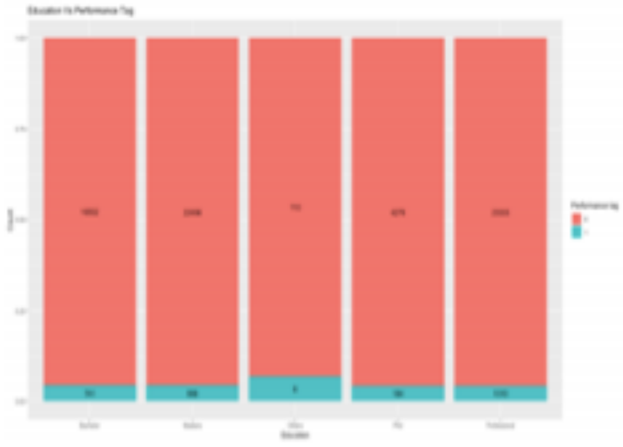


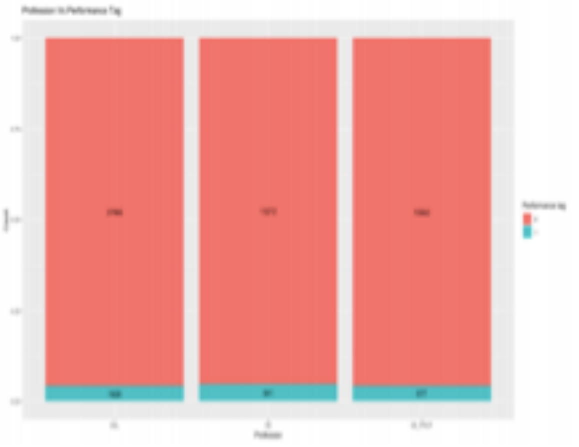
Information Value	Predictive Power
< 0.02	Useless for prediction
0.02 - 0.1	Weak predictor
0.1 - 0.3	Medium predictor
0.3 - 0.5	Strong predictor
> 0.5	Suspicious too good to be true

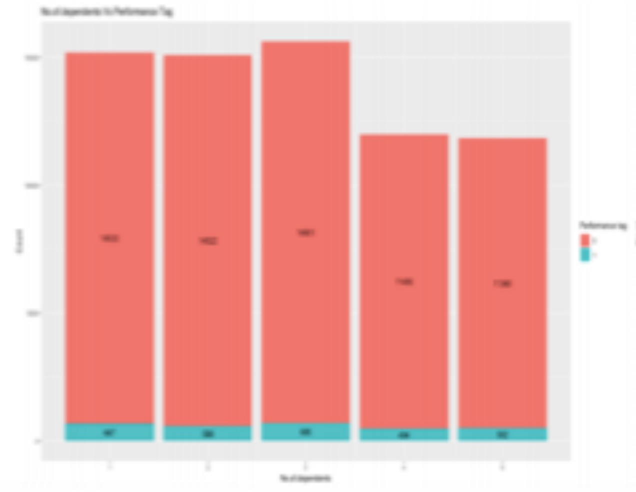
EDA-Demographic data

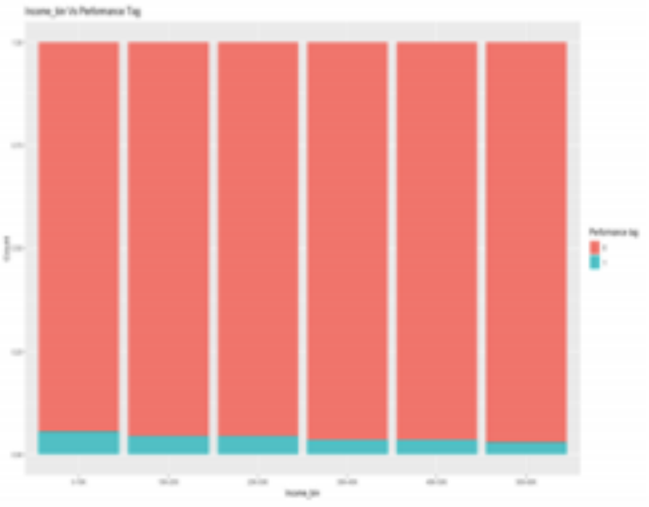


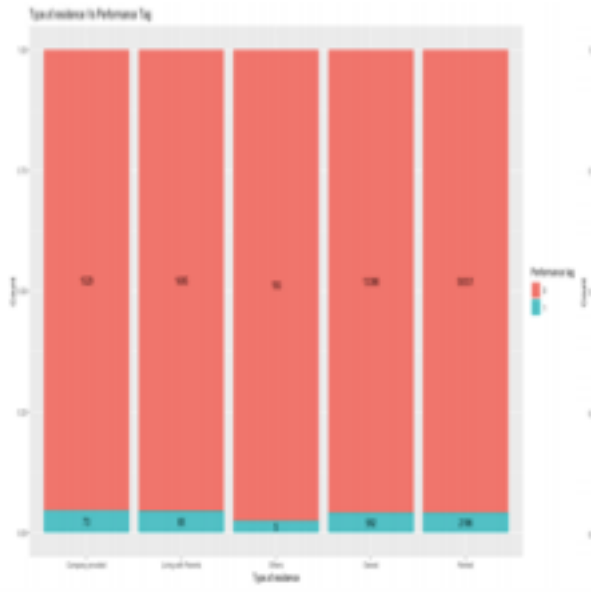


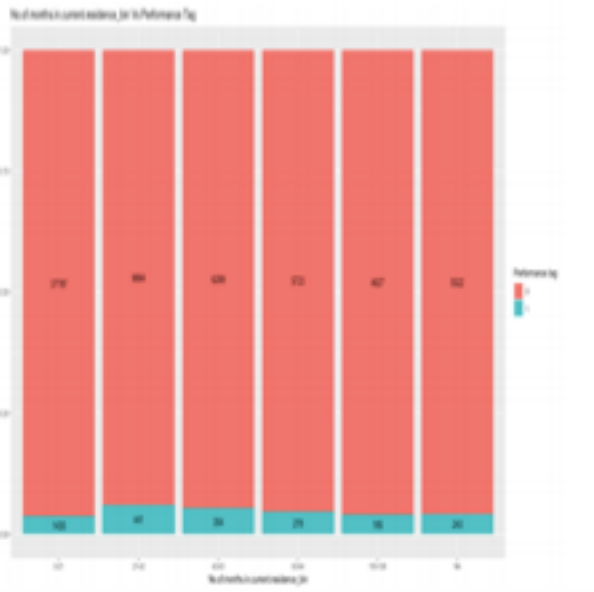




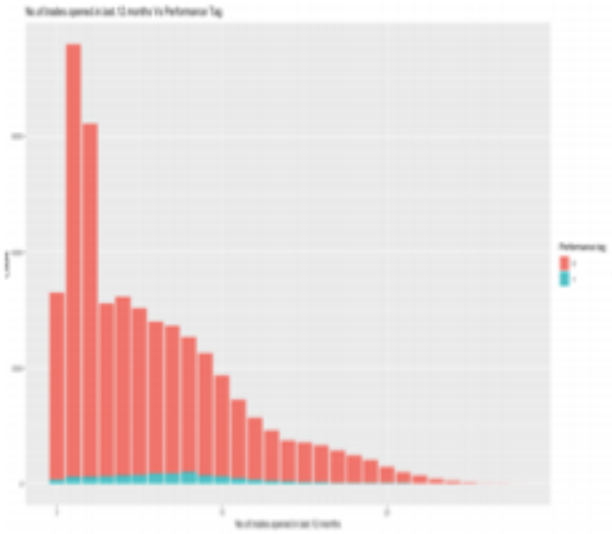


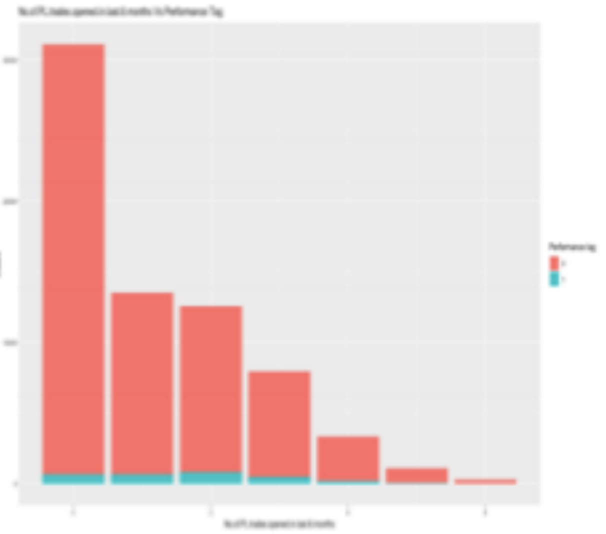


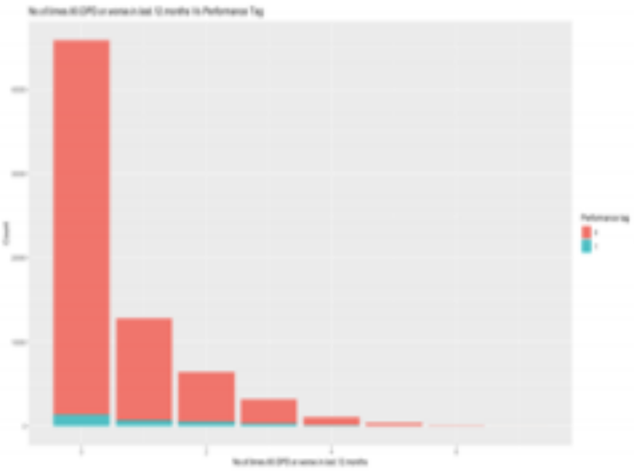


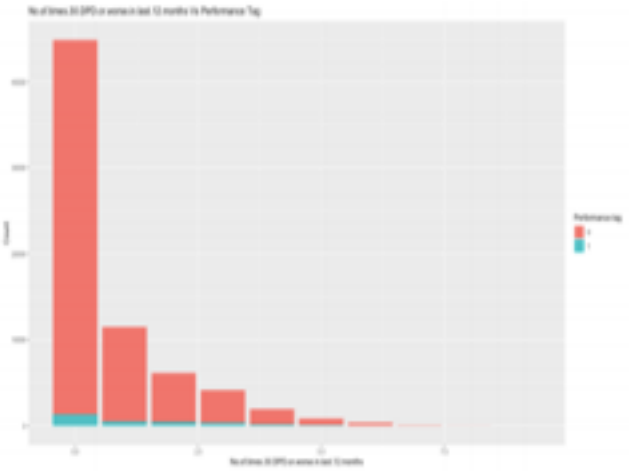


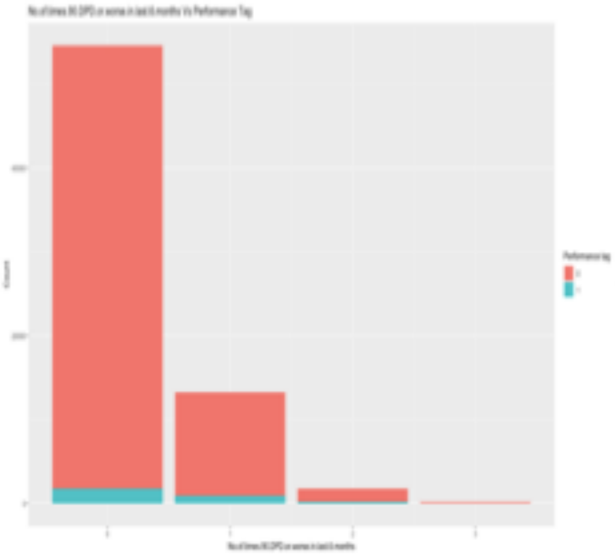
EDA-Credit bureau data

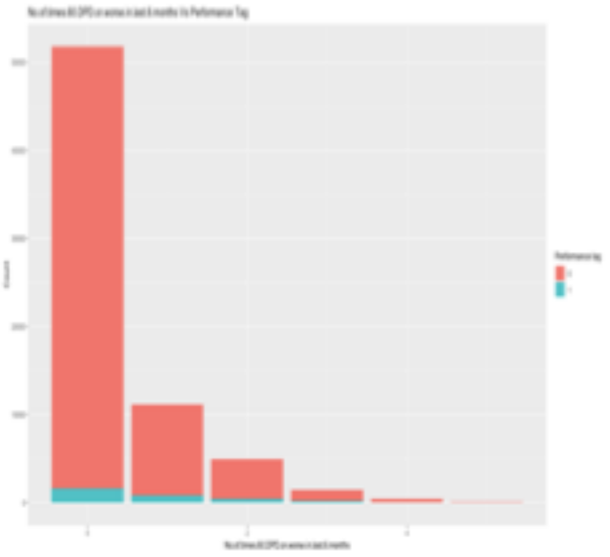


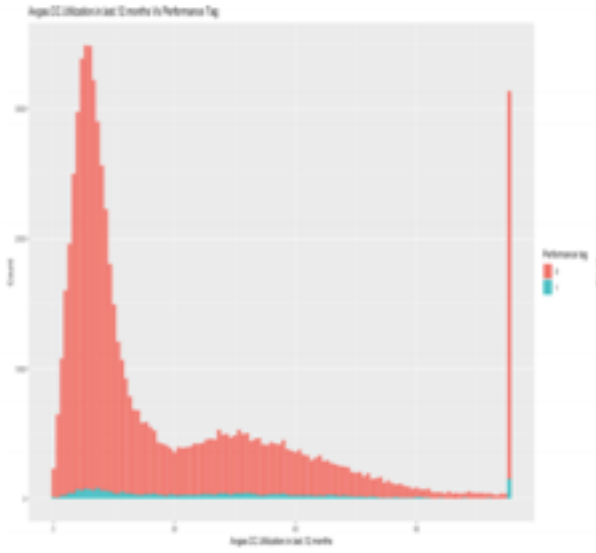


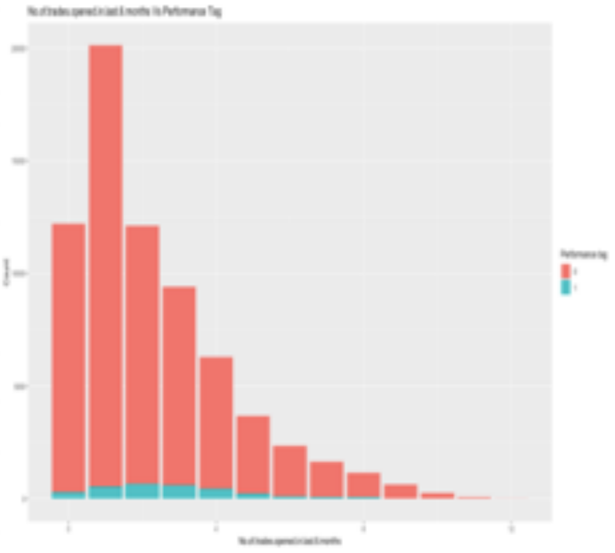




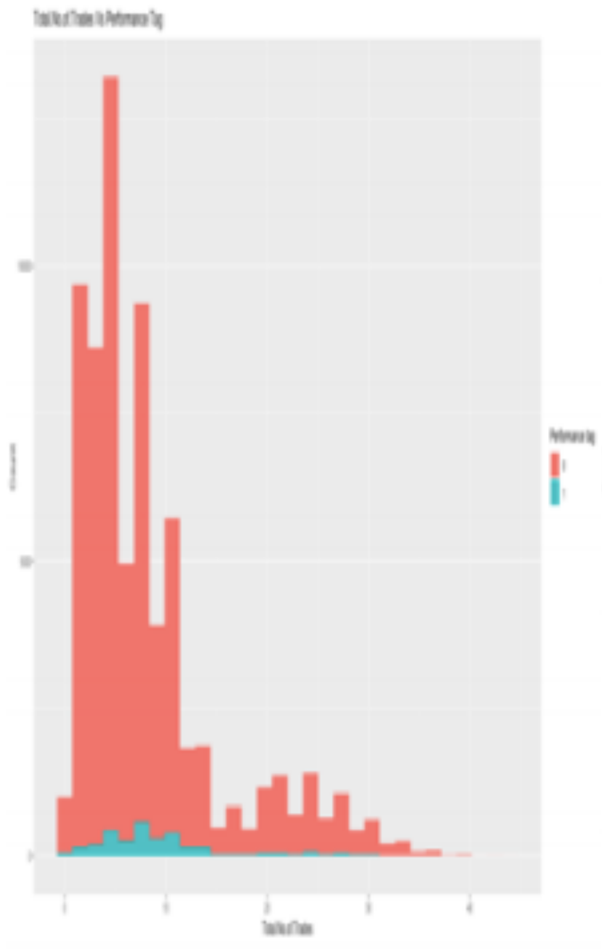


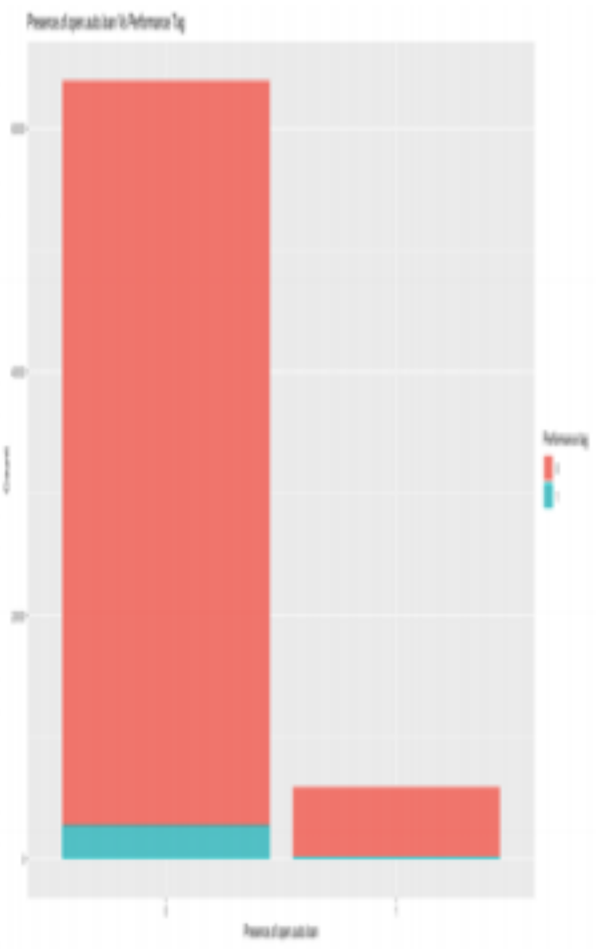


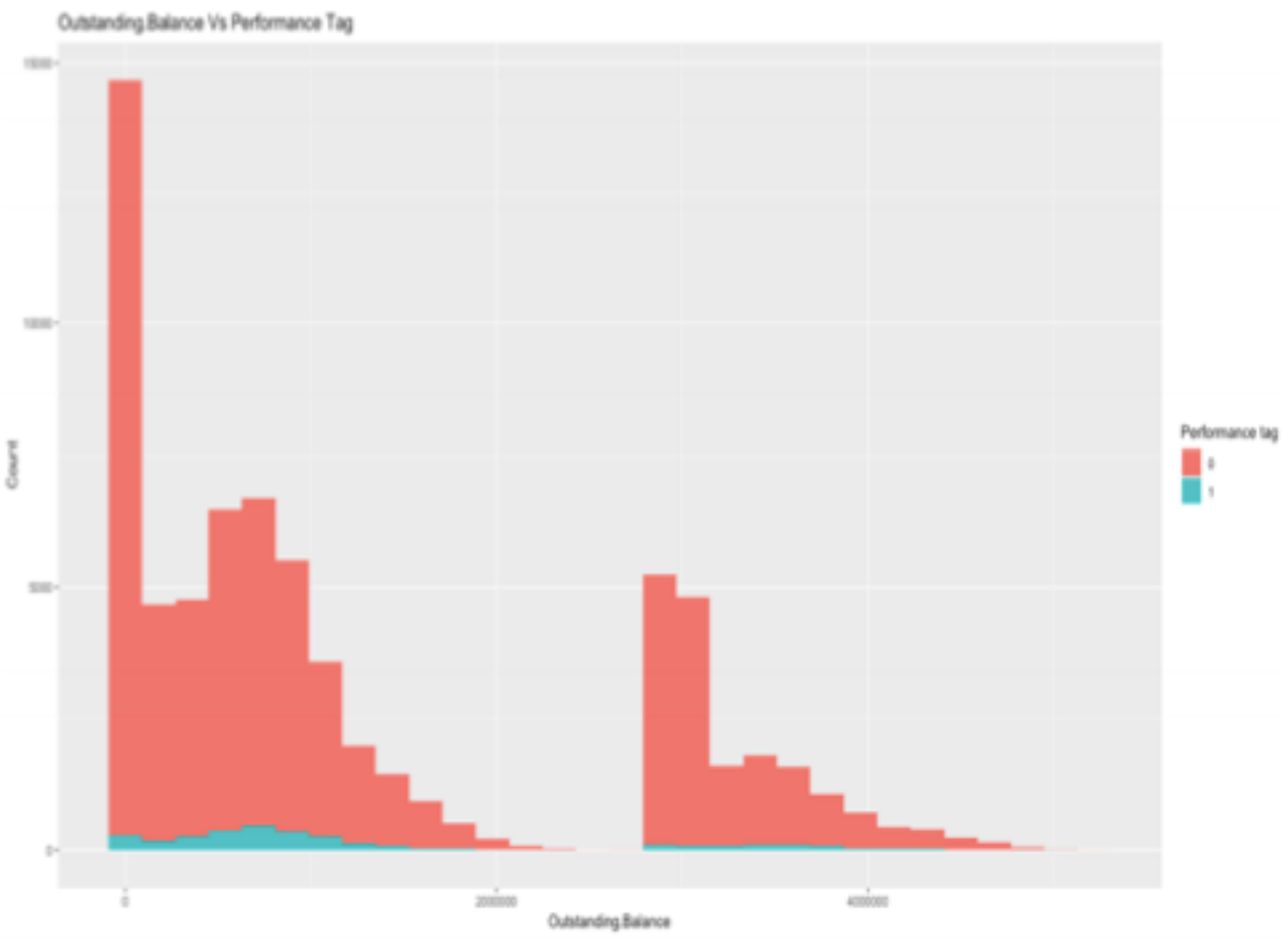




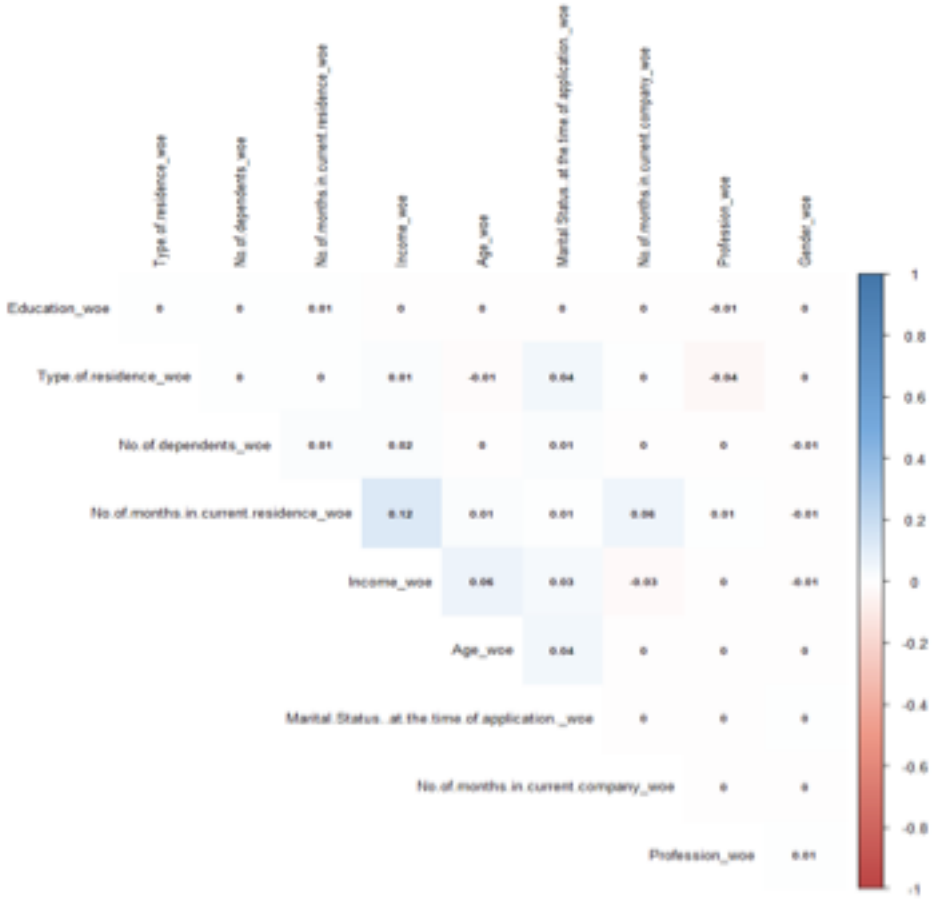
EDA- Credit bureau data

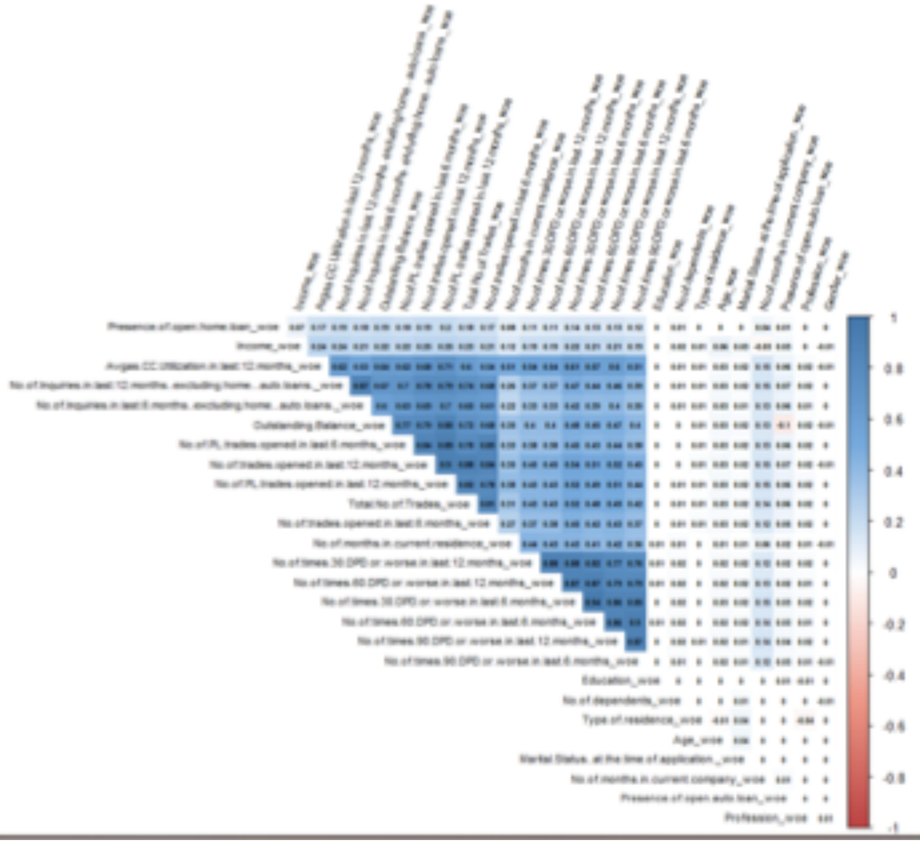




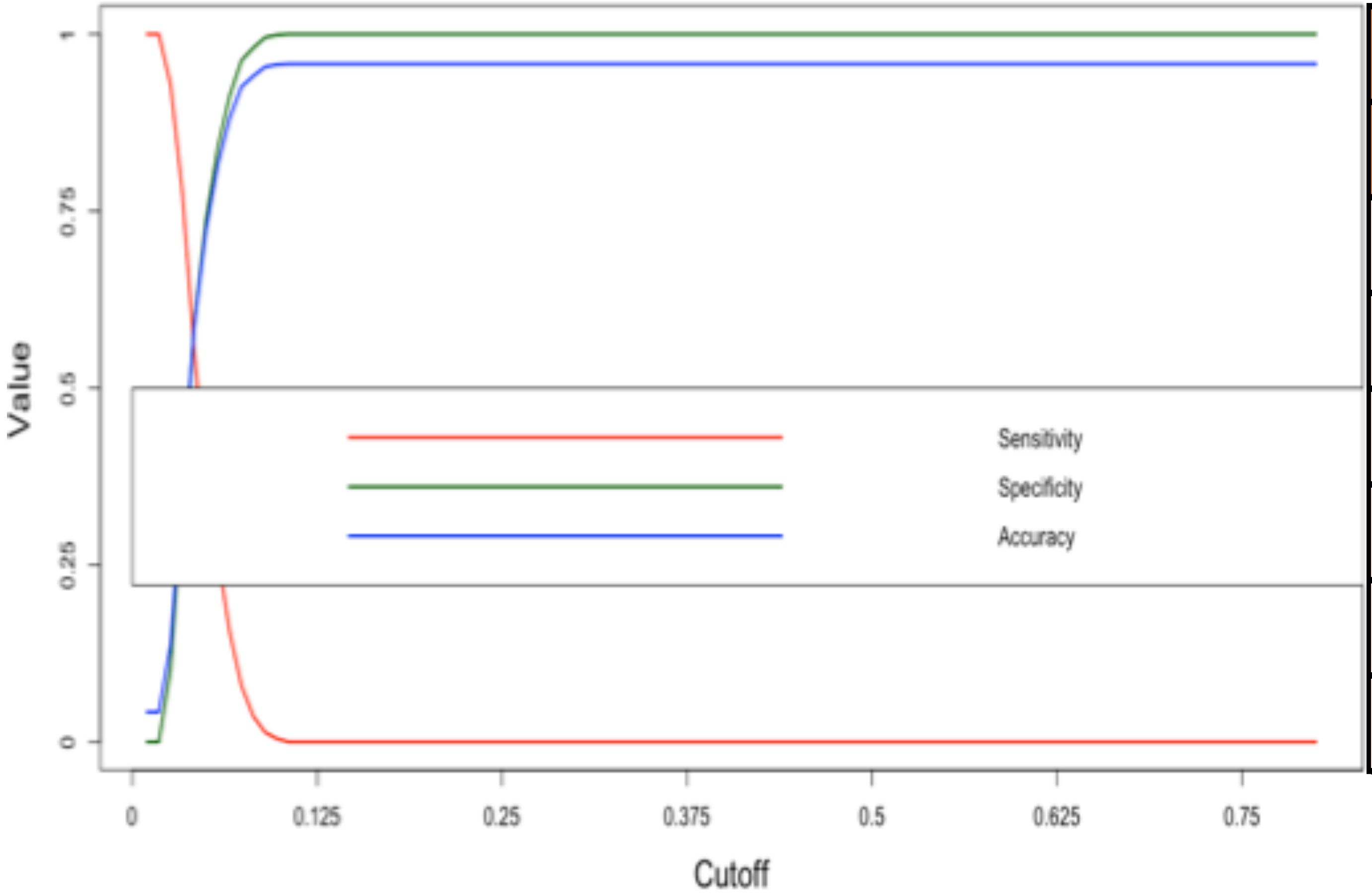


Correlation Analysis of independent features





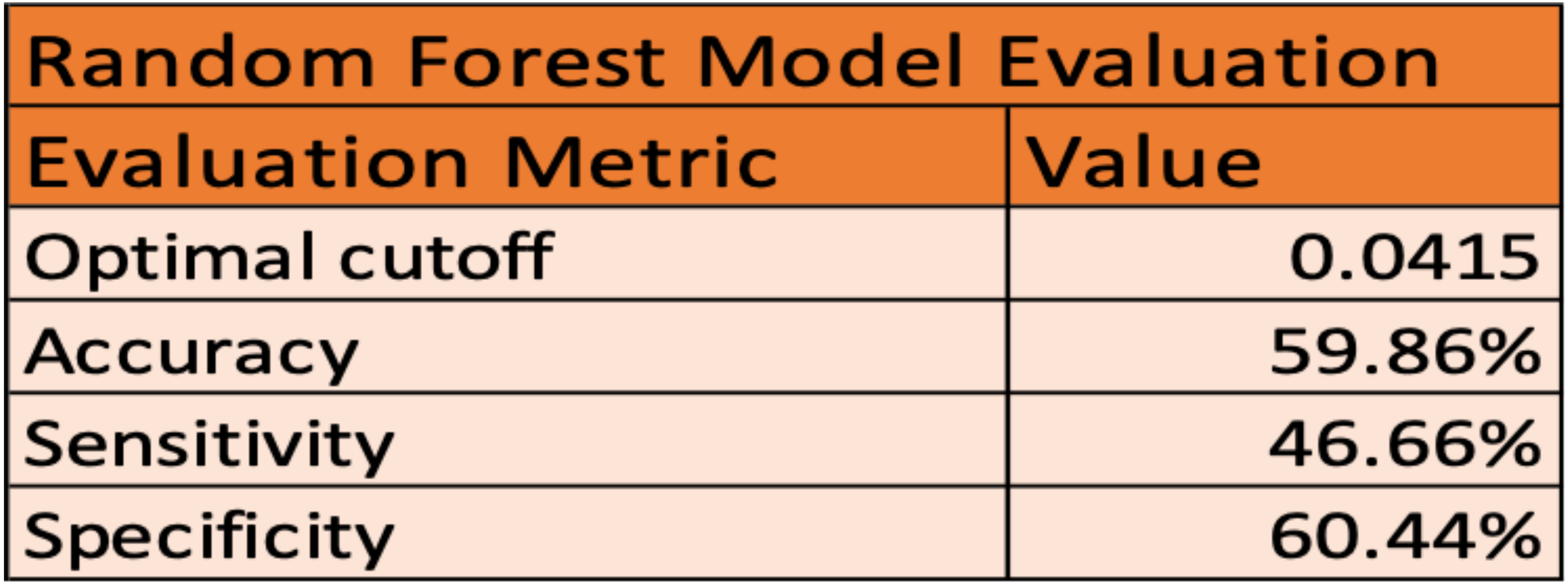
Logistic Model and Result of Demographic Data



Logistic Regression Model Evaluation	
Evaluation Metric	Value
Optimal cutoff	0.0415
Accuracy	57.15%
Sensitivity	57.42%
Specificity	57.14%
KS Statistics	14.55%
AUC	60.20%

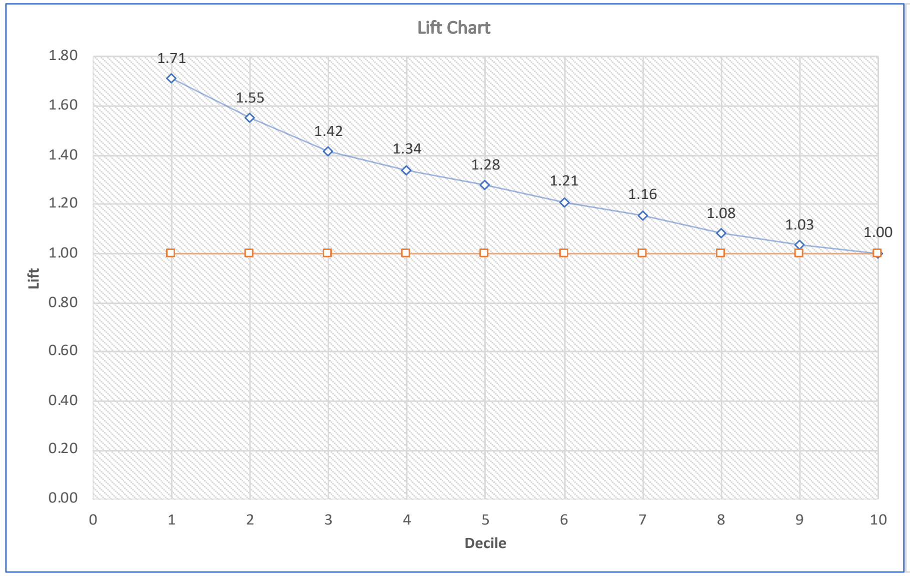
Note : We have developed random forest and logistic model but we preferred logistic regression due to better evaluation parameter values

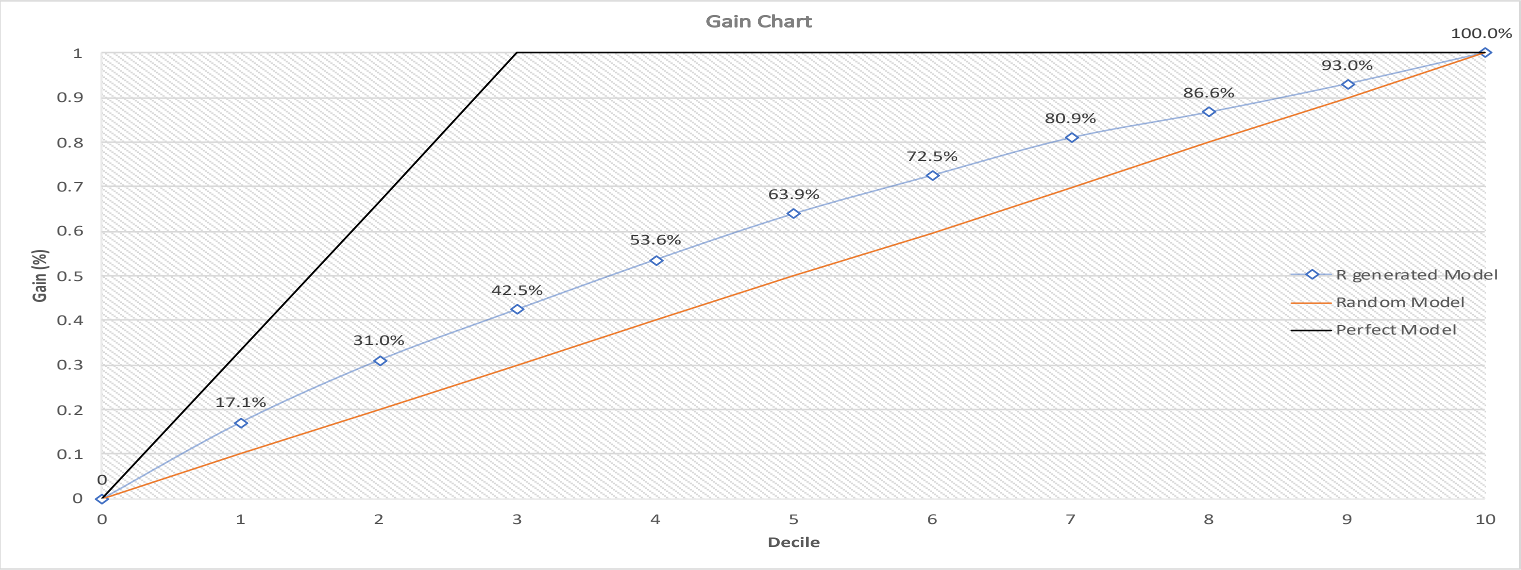
RF Model Result of Demographic Data

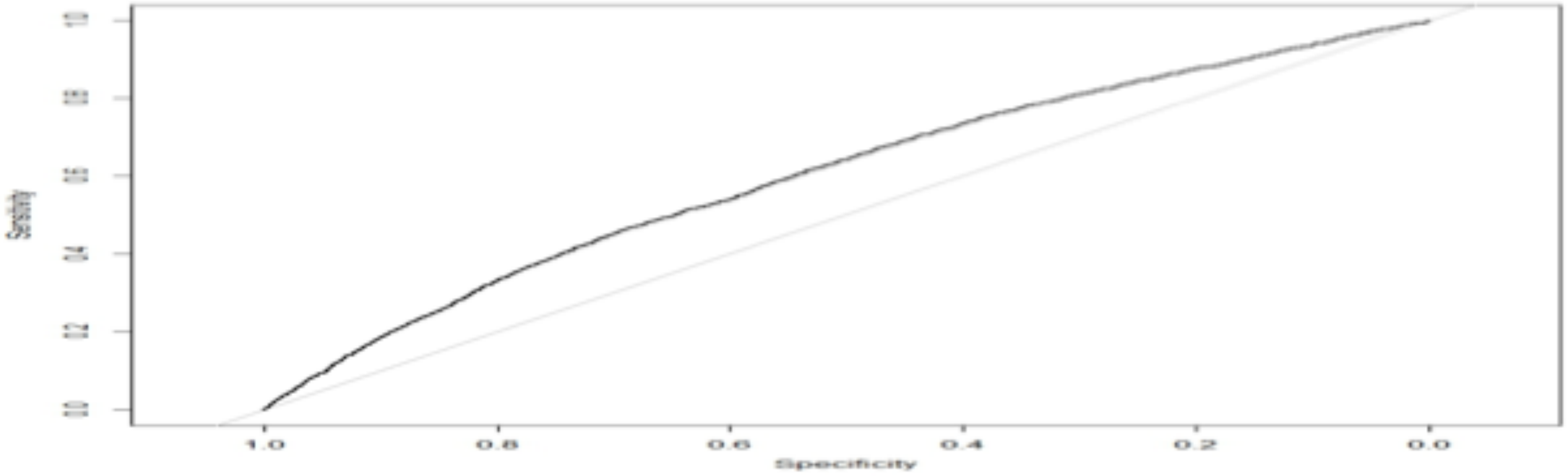


Random Forest Model Evaluation	
Evaluation Metric	Value
Optimal cutoff	0.0415
Accuracy	59.86%
Sensitivity	46.66%
Specificity	60.44%

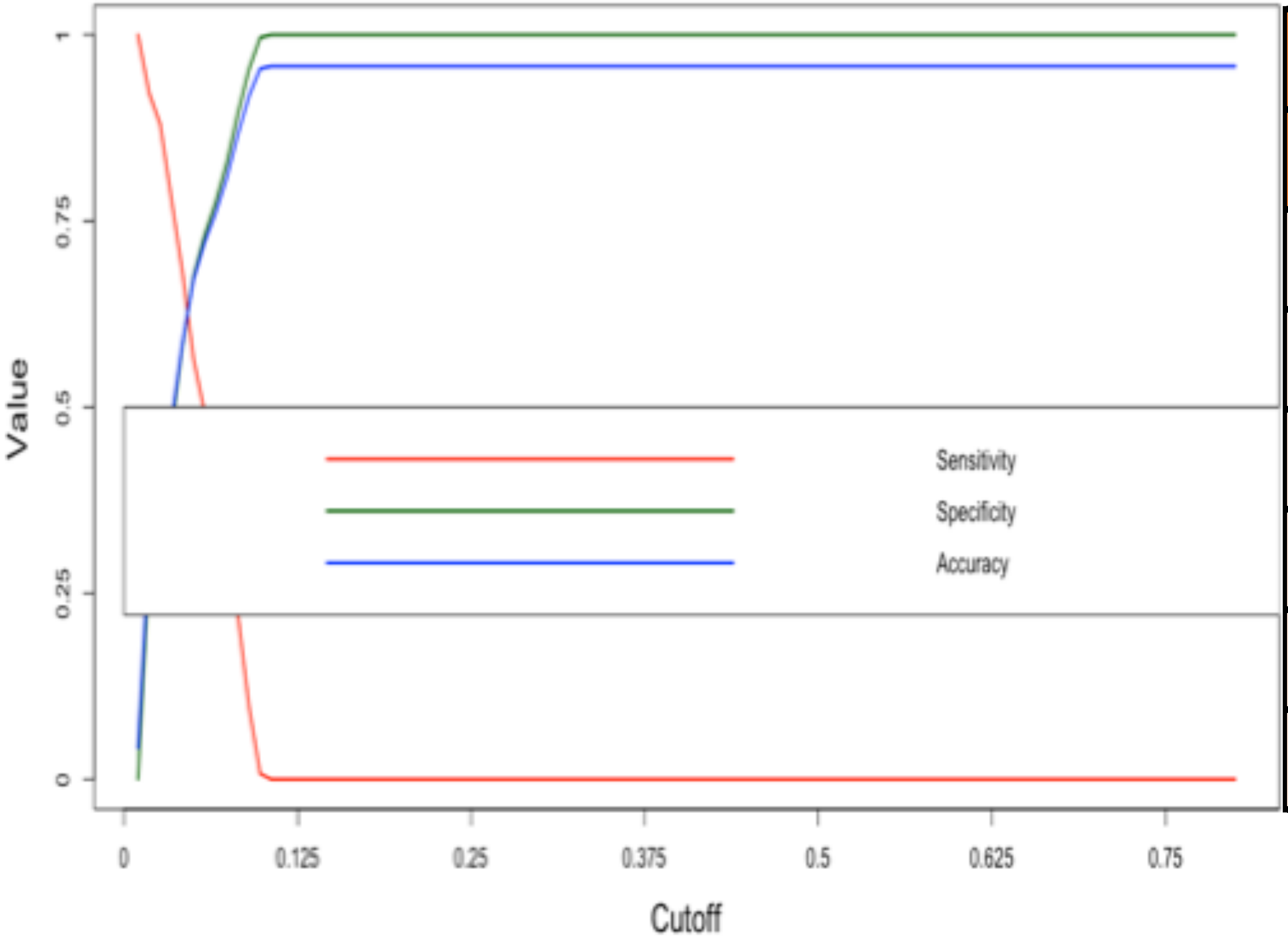
Model Evaluation (Logistic Regression) of Demographic Data







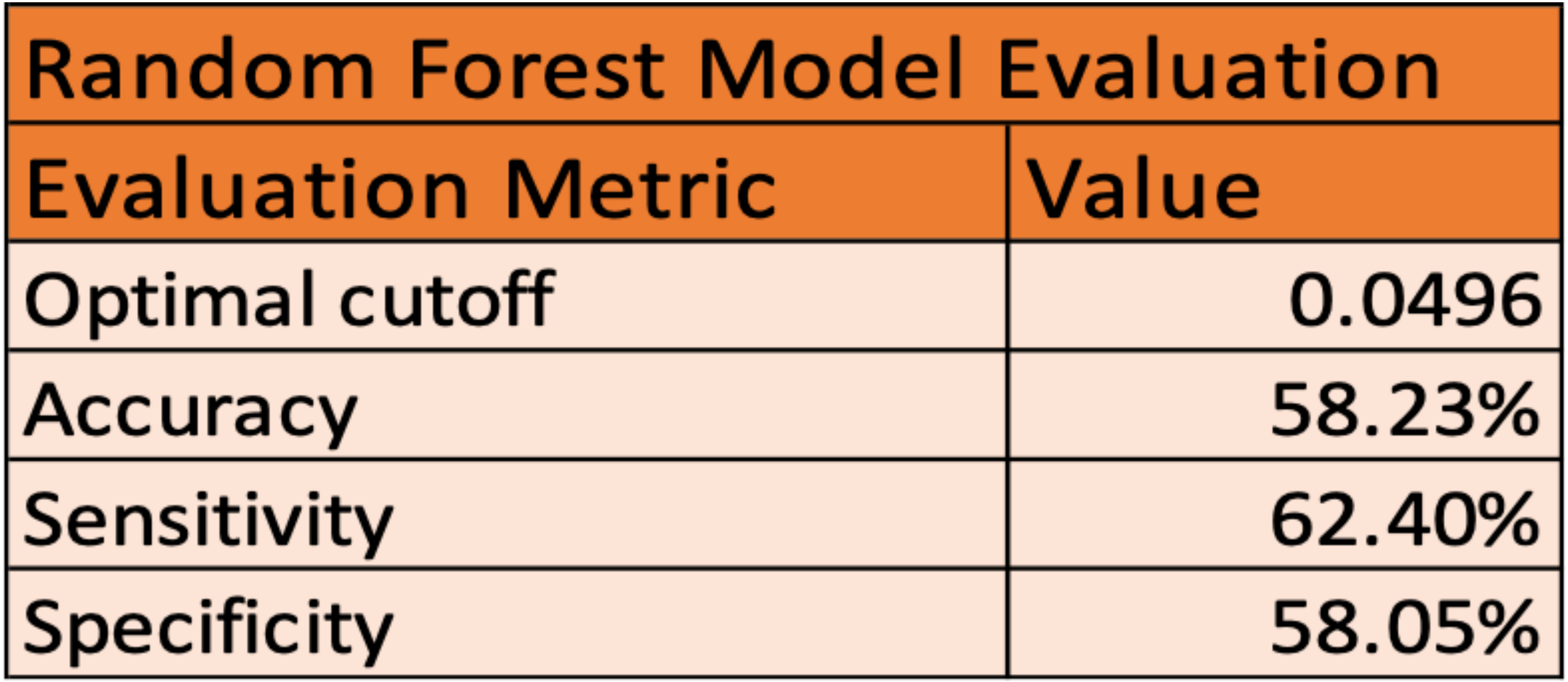
Logistic Model and Result of Demographic and Credit bureau Data



Logistic Regression Model Evaluation	
Evaluation Metric	Value
Optimal cutoff	0.047
Accuracy	64.35%
Sensitivity	60.36%
Specificity	64.52%
KS Statistics	24.88%
AUC	67.50%

Note : We have developed random forest and logistic model but we preferred logistic regression due to better evaluation parameter values

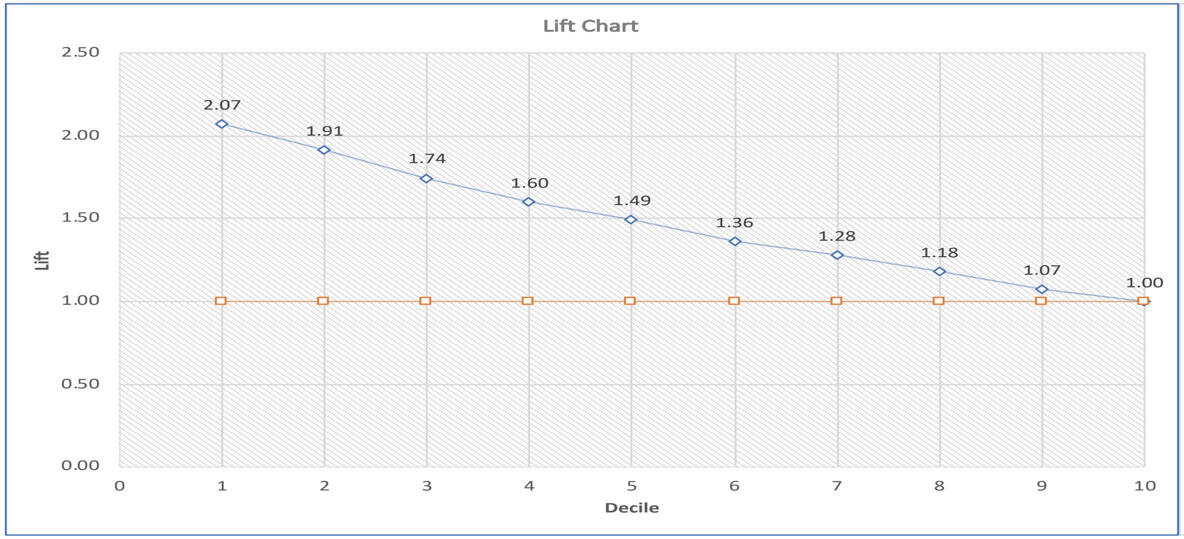
RF Model Result of Demographic and Credit bureau Data

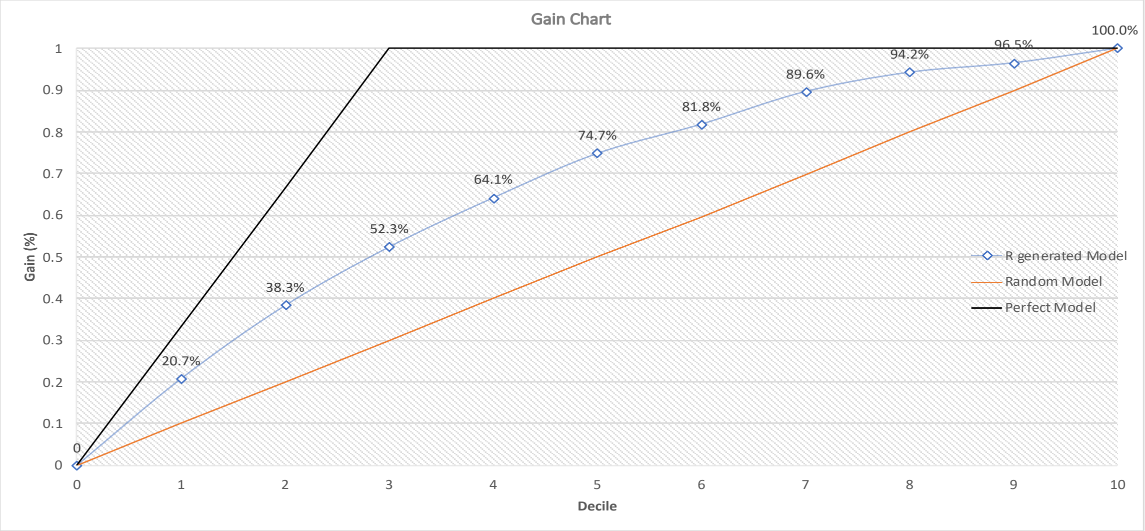


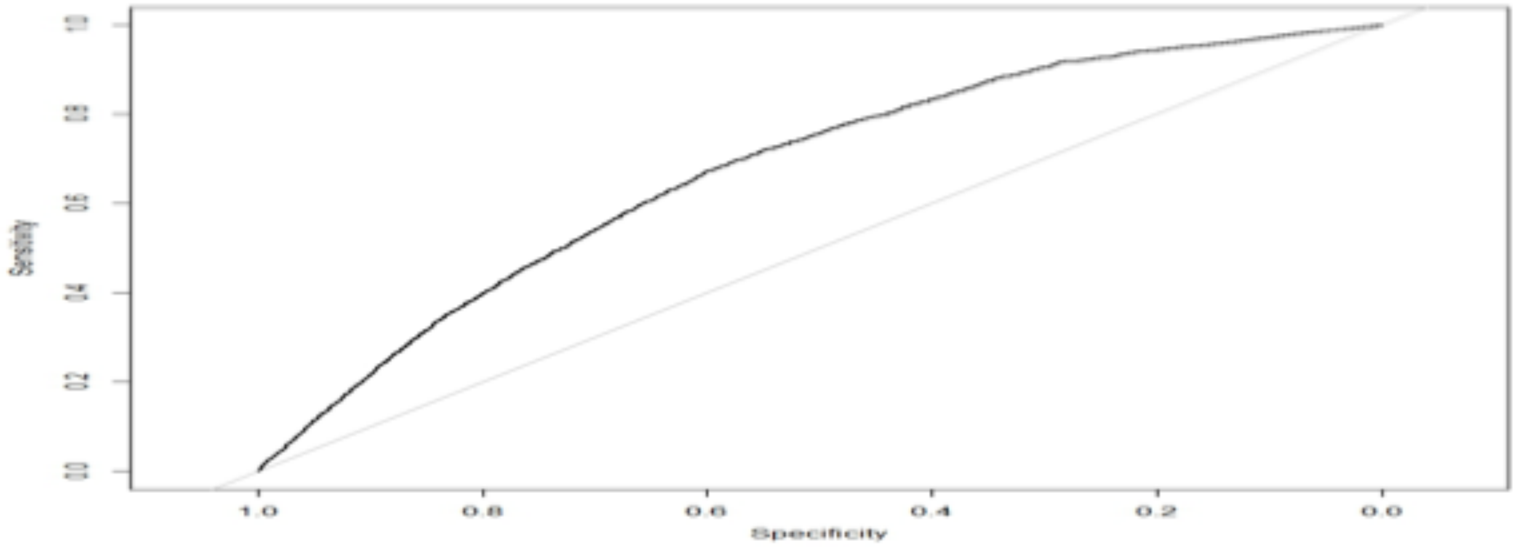
Random Forest Model Evaluation	
Evaluation Metric	Value
Optimal cutoff	0.0496
Accuracy	58.23%
Sensitivity	62.40%
Specificity	58.05%

Note : We have developed random forest and logistic model but we preferred logistic regression due to better evaluation parameter values

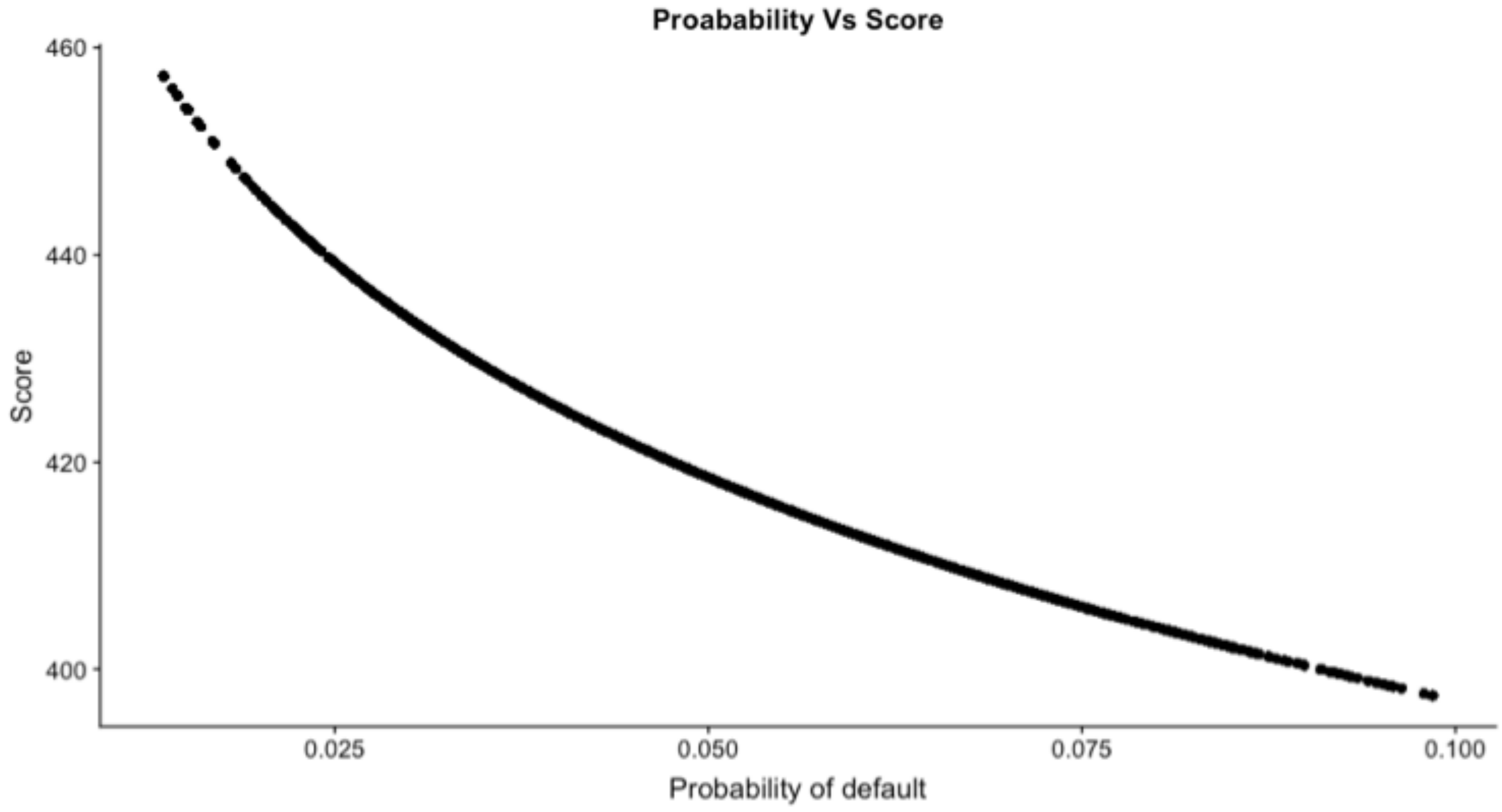
Model evaluation(Logistic Regression) of Demographic and Credit bureau Data





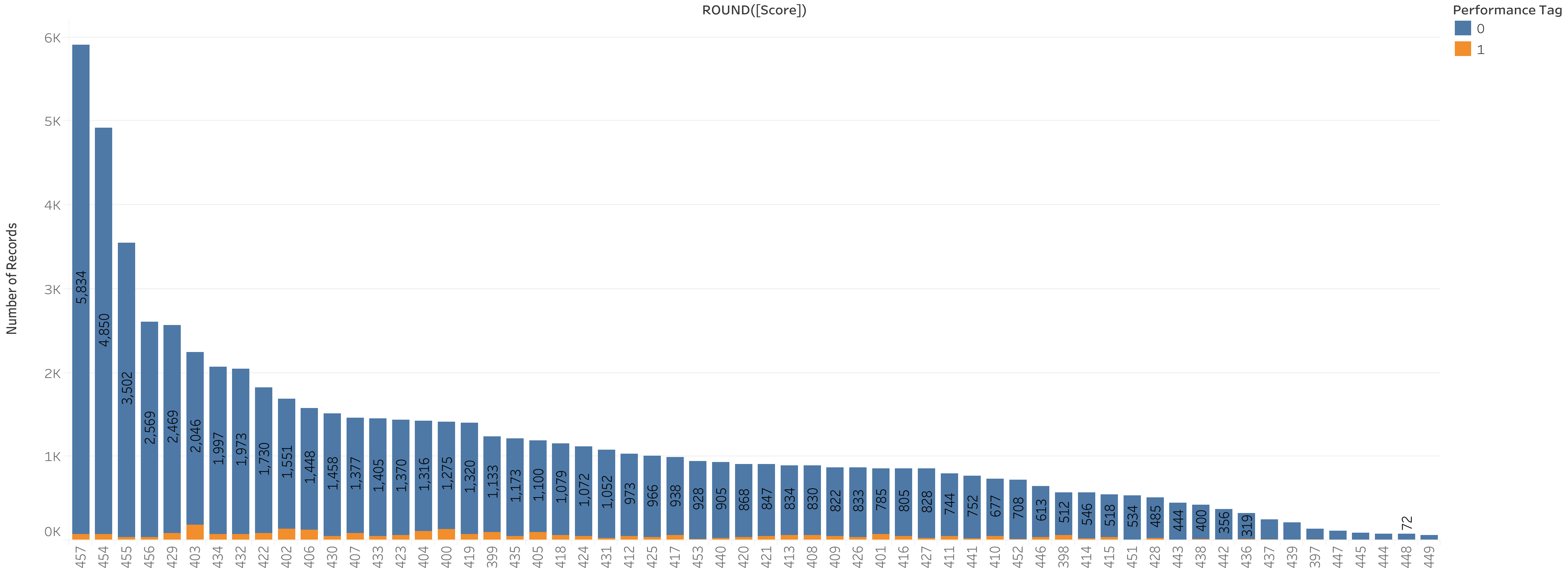


Trend of Score Vs Probability of default



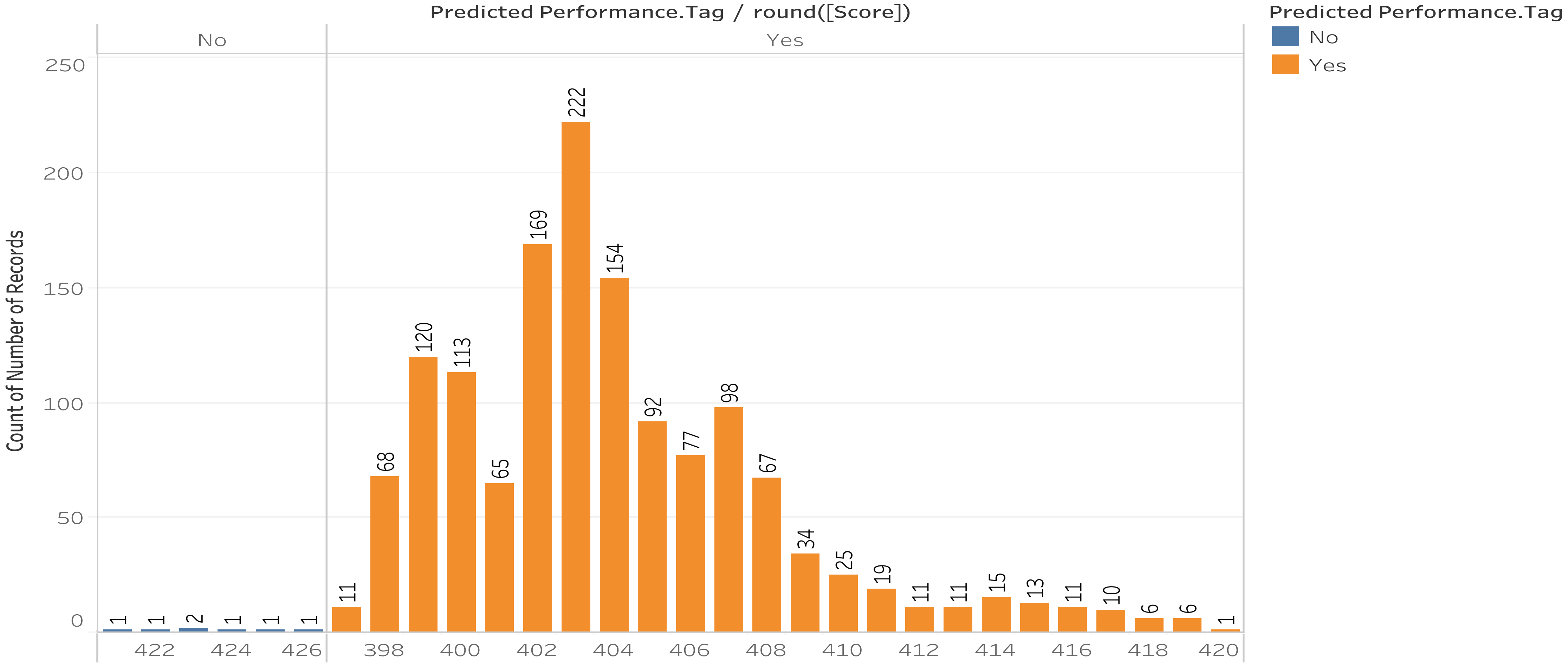
Score card distribution

Score Vs Performance tag

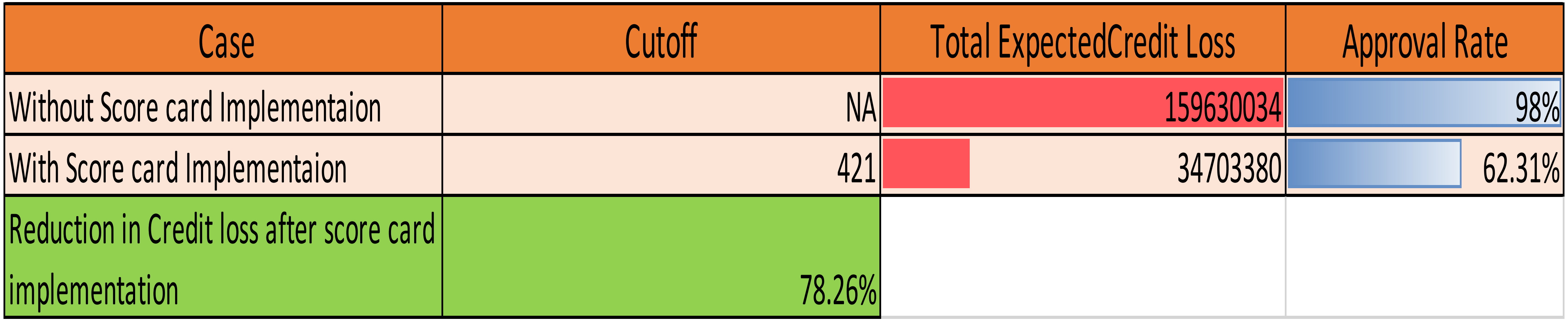


Score card distribution of Rejected population

Score Vs Predicted performance tag on Rejected Population



Financial Analysis



Case	Cutoff	Total ExpectedCredit Loss	Approval Rate
Without Score card Implementaion	NA	159630034	98%
With Score card Implementaion	421	34703380	62.31%
Reduction in Credit loss after score card implementation	78.26%		

Conclusion

Summary of the inferences made from the analysis are as listed below –

- Approval rate without score card implementation was 98% which was the main driving factor in net credit loss.
- With suggested optimal cut off score of 421, avg. 62.31% of applicants would be auto approved.

  Hence 37.69% applicants would be auto rejected.
- From P & L point of view, we have implemented scorecard which in turns reduces net credit loss by 78.26%
- With scorecard implementation we have reduced revenue loss to 4673367 Unit*.

* Unit here refers currency value which can be INR,USD etc.

Thank You!!!!!!!