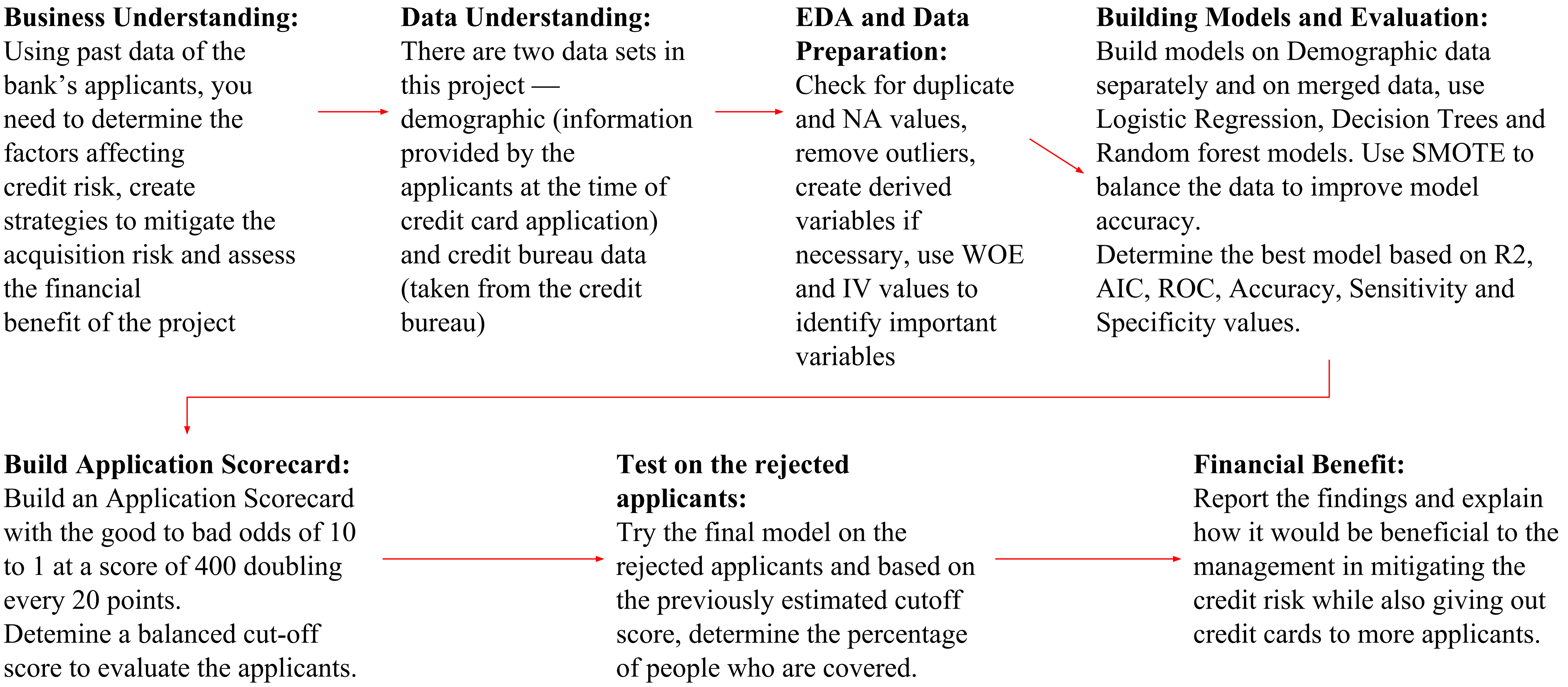
BFS CAPSTONE PROJECT

FINAL SUBMISSION

Group Members:

1. Anuj Arya
2. Asim Pattnaik
3. Mohammed Suhail Y
4. Rakesh Bosu

Problem Solving Methodology



Business Understanding:
Using past data of the bank’s applicants, you need to determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of the project

Data Understanding:
There are two data sets in this project — demographic (information provided by the applicants at the time of credit card application) and credit bureau data (taken from the credit bureau)

EDA and Data Preparation:
Check for duplicate and NA values, remove outliers, create derived variables if necessary, use WOE and IV values to identify important variables

Building Models and Evaluation:
Build models on Demographic data separately and on merged data, use Logistic Regression, Decision Trees and Random forest models. Use SMOTE to balance the data to improve model accuracy.
Determine the best model based on R2, AIC, ROC, Accuracy, Sensitivity and Specificity values.

Build Application Scorecard:
Build an Application Scorecard with the good to bad odds of 10 to 1 at a score of 400 doubling every 20 points.
Detemine a balanced cut-off score to evaluate the applicants.

Test on the rejected applicants:
Try the final model on the rejected applicants and based on the previously estimated cutoff score, determine the percentage of people who are covered.

Financial Benefit:
Report the findings and explain how it would be beneficial to the management in mitigating the credit risk while also giving out credit cards to more applicants.

Business Understanding

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 identify the right customers using predictive models. Using past data of the bank’s applicants, you need to determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of the project.

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 datasets have Performance Tag column which signifies whether the person has defaulted after getting a credit card

Data Understanding

- Demographic Data
  - Rows - 71295
  - Columns- 12

- Credit Bureau Data
  - Rows - 71295
  - Columns- 19

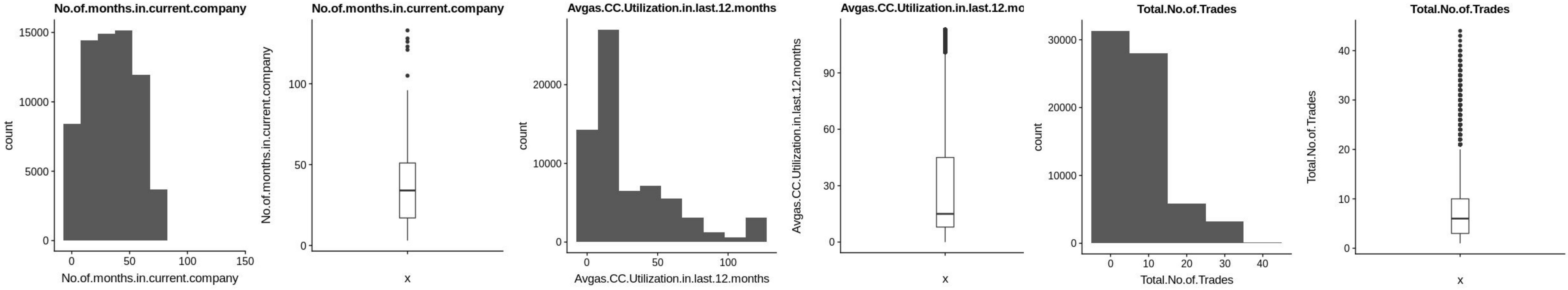
- 71292 Unique values are found in both datasets.
- There are 3 Duplicate records in both.

Data cleaning and preparation

- Remove the duplicate rows in both the datasets.
- Merge both the datasets into one.
- Remove all the records with Age less than 18 as only people above age 18 can
- apply for credit cards.
- Remove all the records where Income less than or equal to 0 as it is not possible to have negative income. This constitutes only a negligible size of the entire dataset, hence it can be removed.
- Check for NA Values
- Performance tag has many NA values which indicate they were rejected at the onset. Separate these records as a individual dataset and use it later to test the model and remove other records with NA values.
- Balanced the dataset using SMOTE Package in R, since the number of records corresponding to non default customers is very less and this will improve the model accuracy.

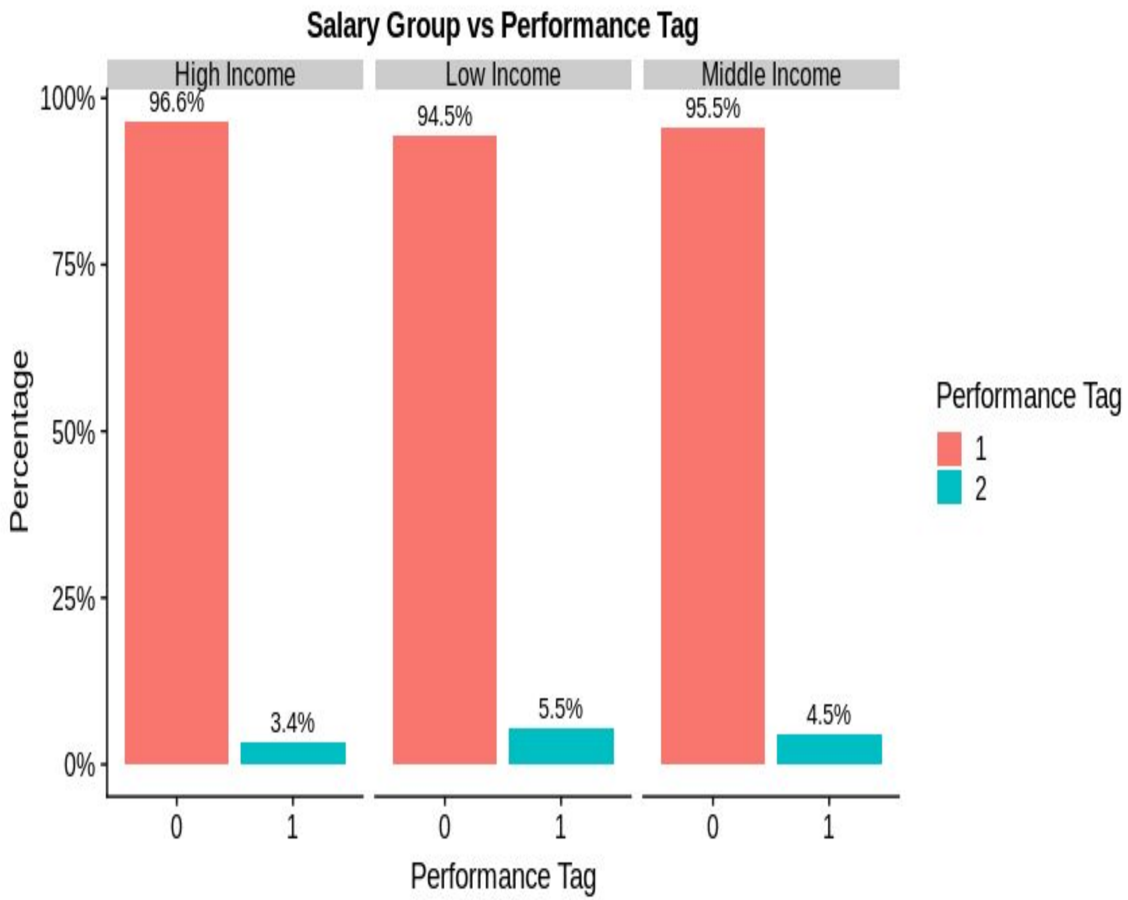
Find Outliers

Outliers were found in the following fields and were treated using the quantile method.

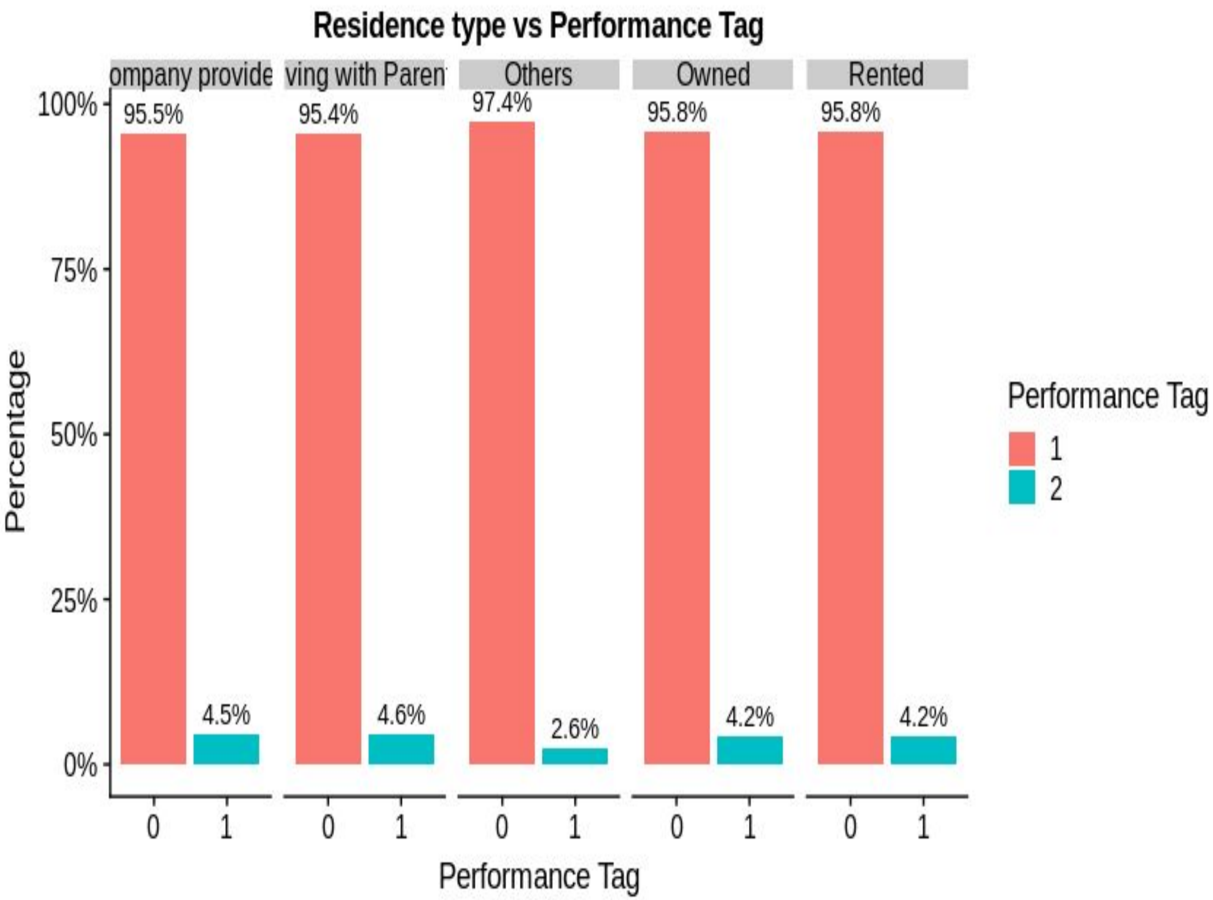


- No.of.months.in.current.company- Sudden jump from 99% to 100%. Cap values to 74
- Avgas.CC.Utilization.in.last.12.months- Sudden jump from 94% to 95%. Cap values to 91
- Total.No.of.Trades- Sudden jump from 99% to 100%. Cap values to 31

Insights from EDA

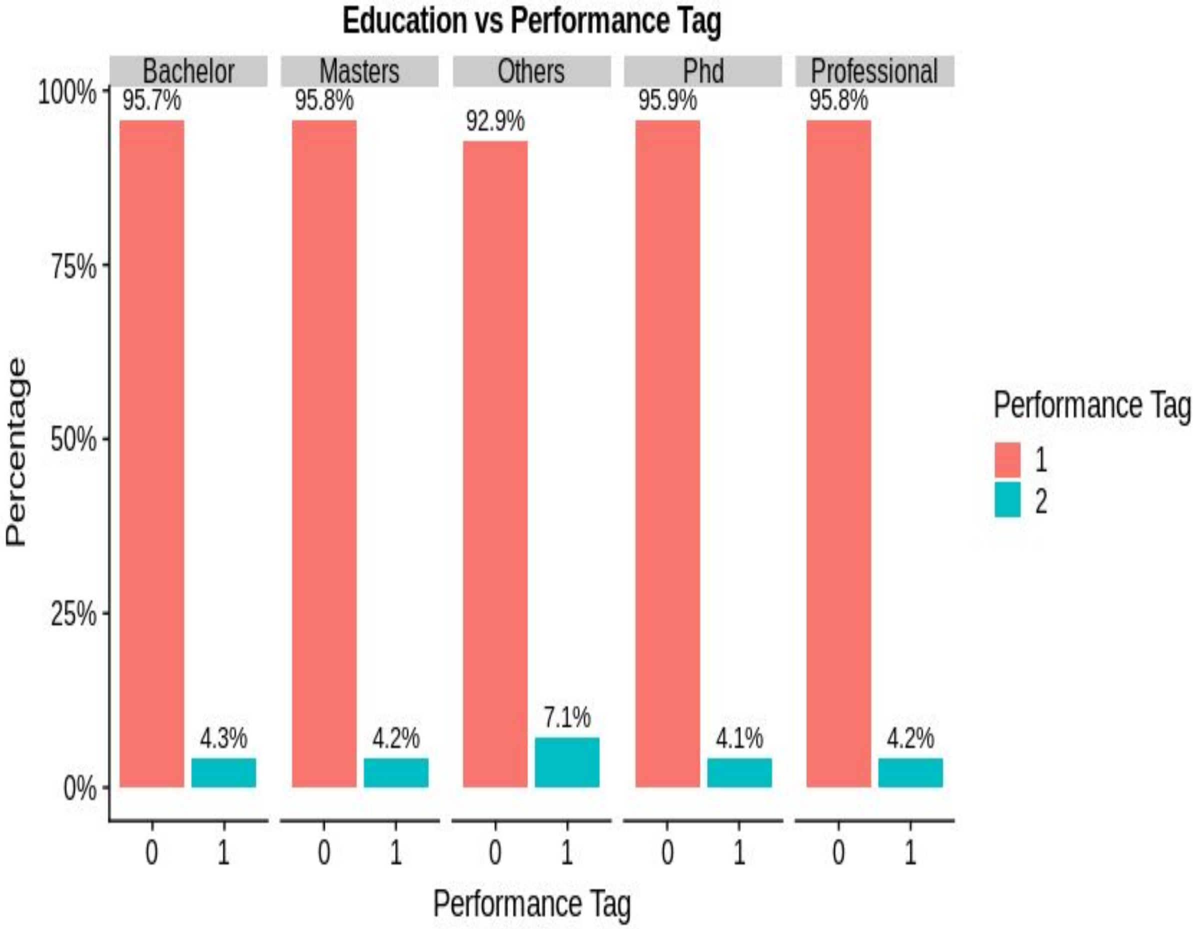


Low Income group people tends to default slightly more compared to high or middle income group

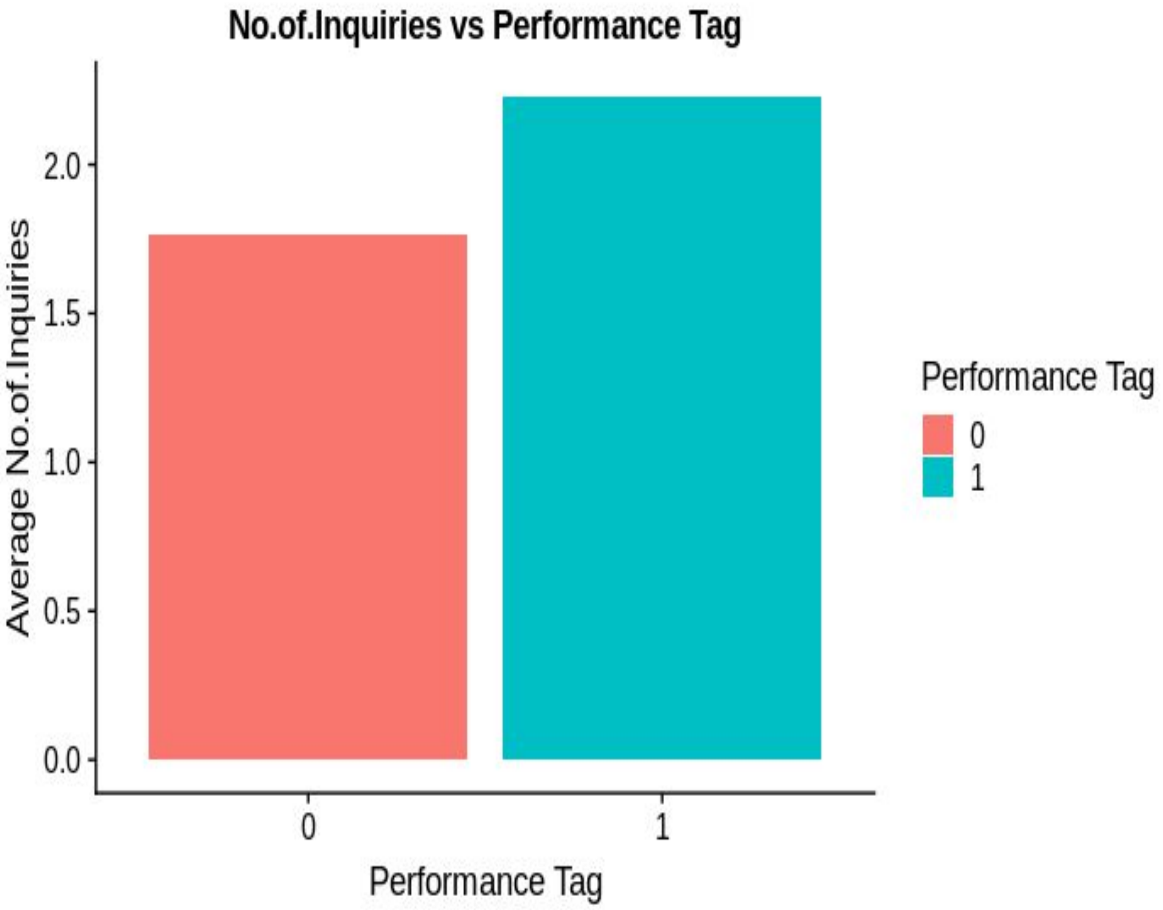


People whose residence type is not disclosed seem to default less.

Insights from EDA

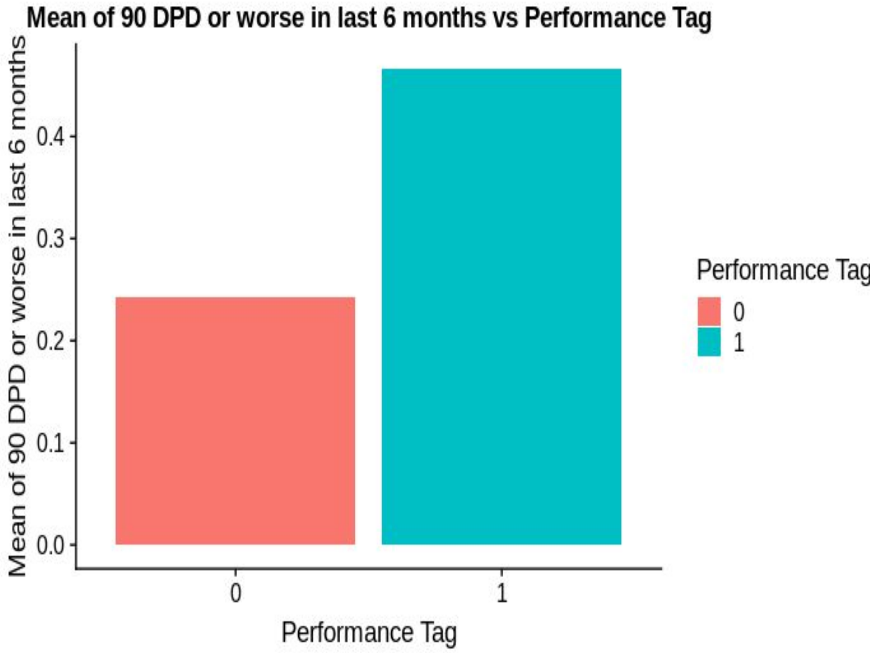


People with Education as others or undisclosed education details tend to default credit cards more compared to other groups.

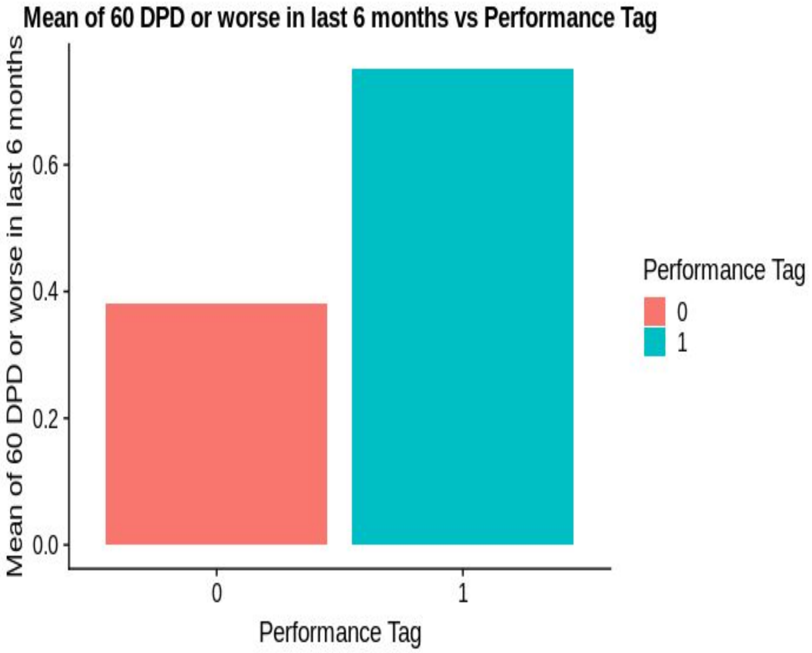


People who tend to default also tend to have a higher number of Inquiries in the past 6 months.

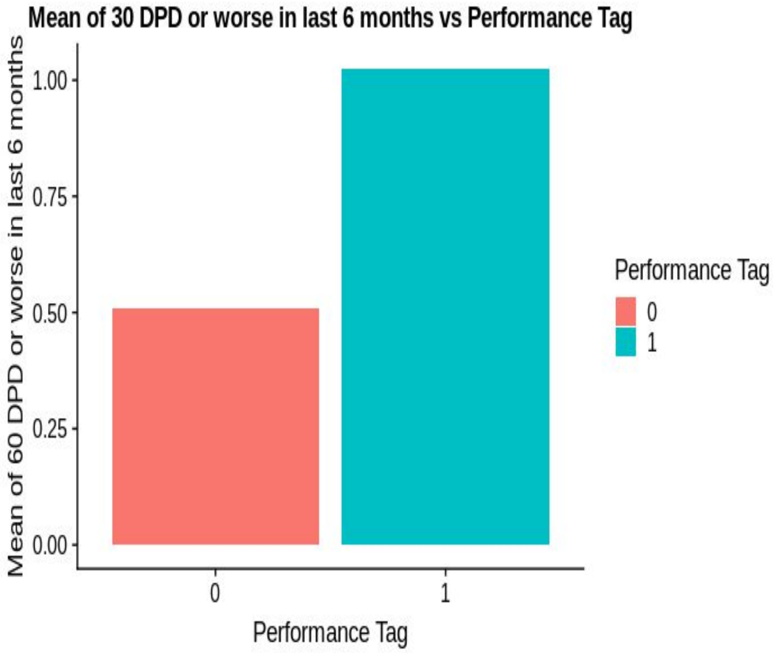
Insights from EDA



Mean of 90 DPD or worse in last 6 months vs Performance Tag



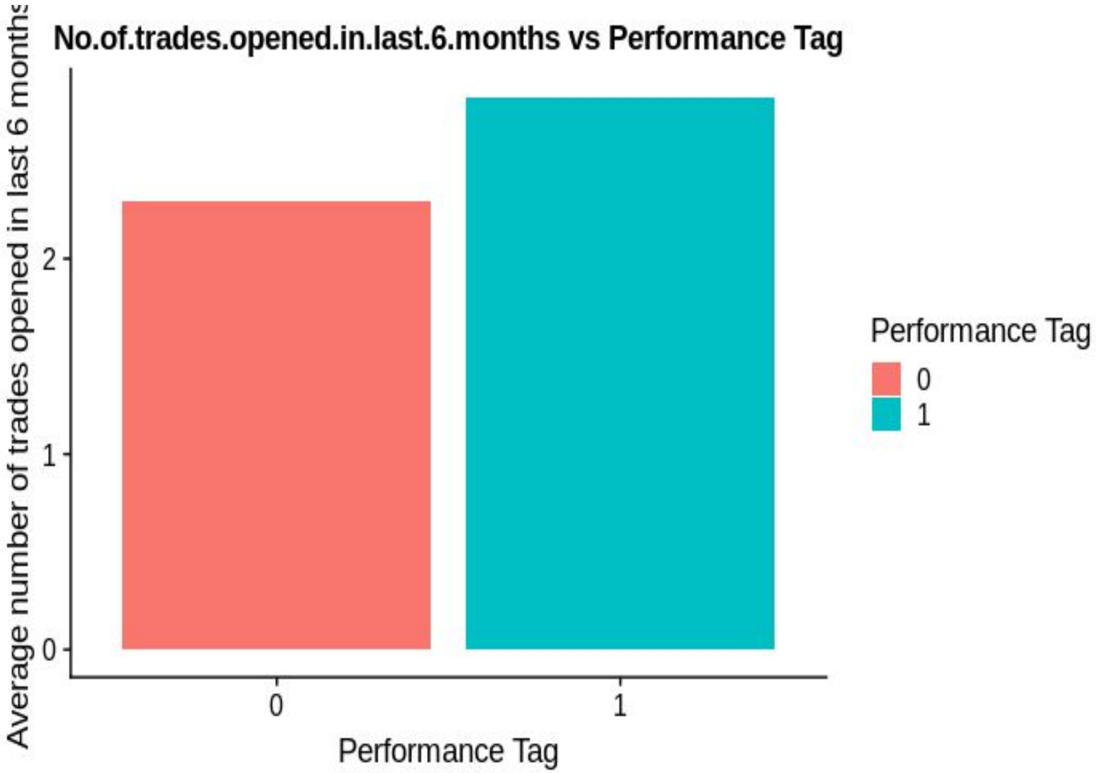
Mean of 60 DPD or worse in last 6 months vs Performance Tag

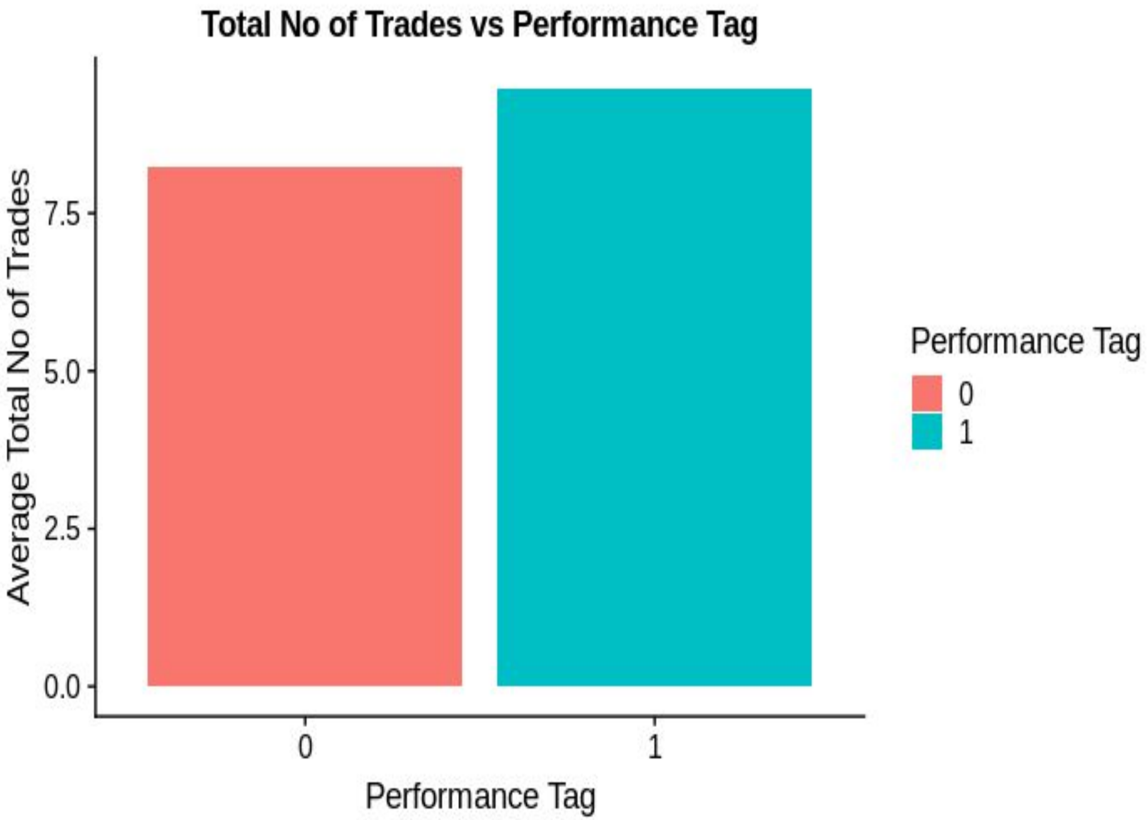


Mean of 30 DPD or worse in last 6 months vs Performance Tag

On Average, people who haven’t paid their dues since 90, 60 and 30 days in the past 6 months are drastically more likely to default in their credit card bills.

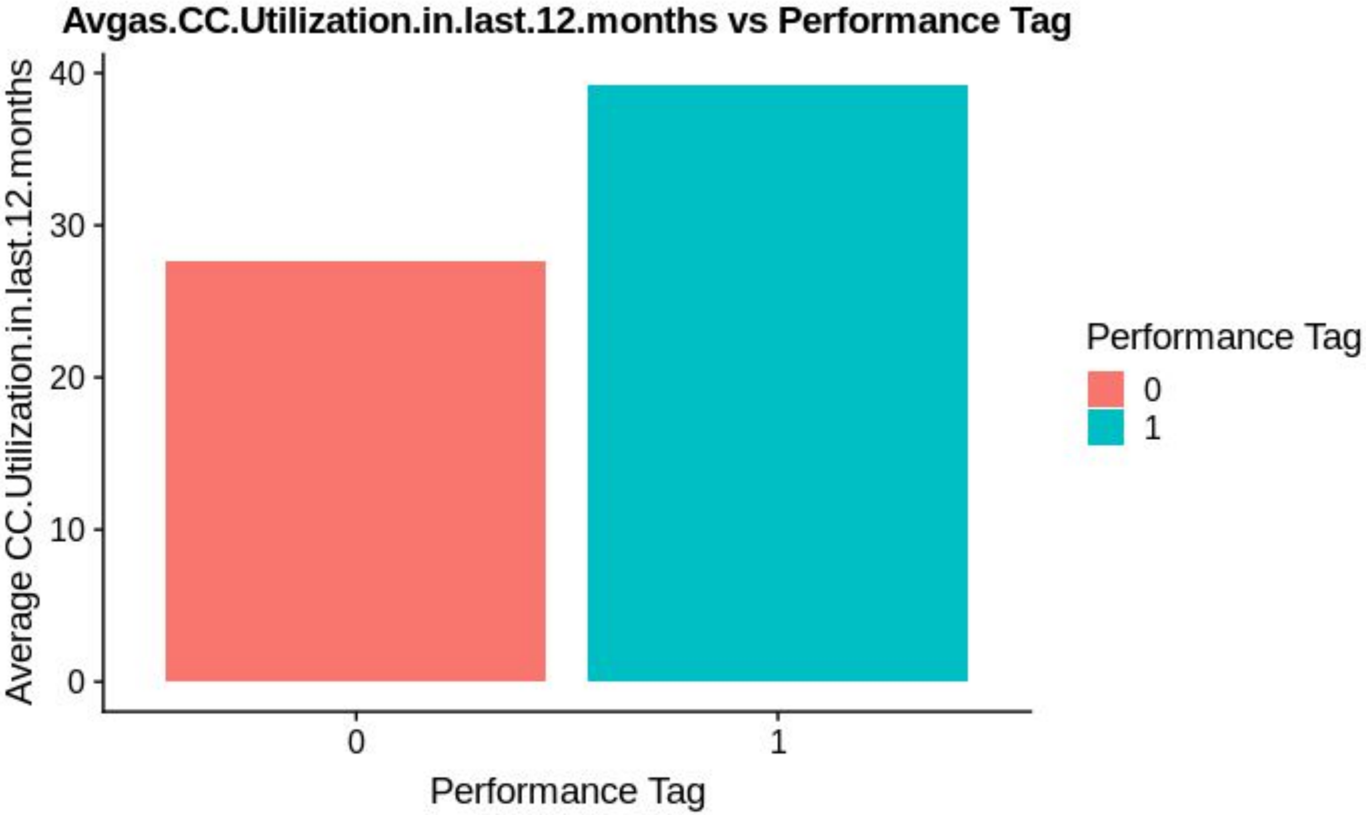
Insights from EDA





On Average, people who have more number of trades or have opened more number of trades in the last 6 months have a marginally higher chance of being defaulters in the credit card payment.

Insights from EDA



On Average, people who have utilized their credit card more in the last 12 months have a considerably higher chance of being defaulters in the credit card payment.

Insights from EDA

From EDA, we have identified some variables as strong predictors

- No of times 90 DPD or worse in last 6 months
- No of times 60 DPD or worse in last 6 months
- Avg CC Utilization
- Outstanding Balance
- Education
- No Of Inquiries
- No of Trades
- Gender
- Marital Status
- Salary
- Age
- No. of Dependents

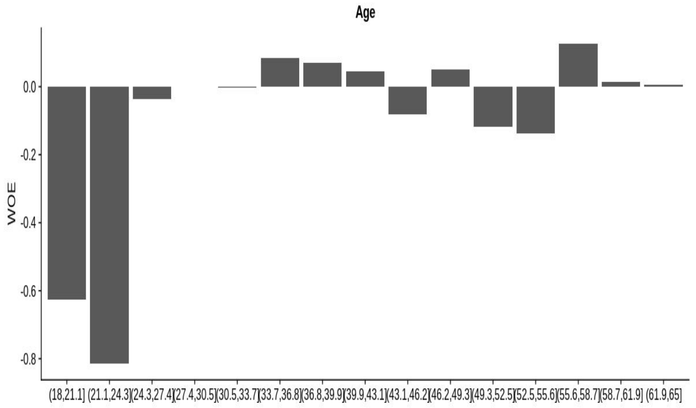
WOE Transformation and Information Values

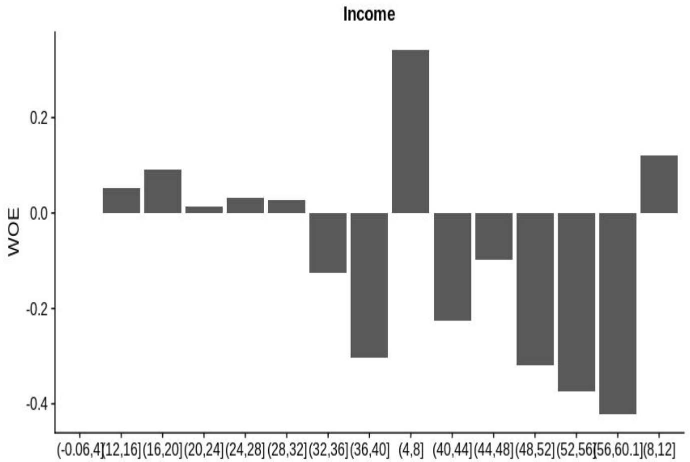
Variable	Information Value
Avgas.CC.Utilization.in.last.12.months	0.32
No.of.trades.opened.in.last.12.months	0.31
No.of.PL.trades.opened.in.last.12.months	0.27
No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.	0.27
Outstanding.Balance	0.26
Total.No.of.Trades	0.25
No.of.times.30.DPD.or.worse.in.last.6.months	0.25
No.of.PL.trades.opened.in.last.6.months	0.23
No.of.times.30.DPD.or.worse.in.last.12.months	0.22
No.of.times.90.DPD.or.worse.in.last.12.months	0.22
No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.	0.22
No.of.times.60.DPD.or.worse.in.last.6.months	0.22

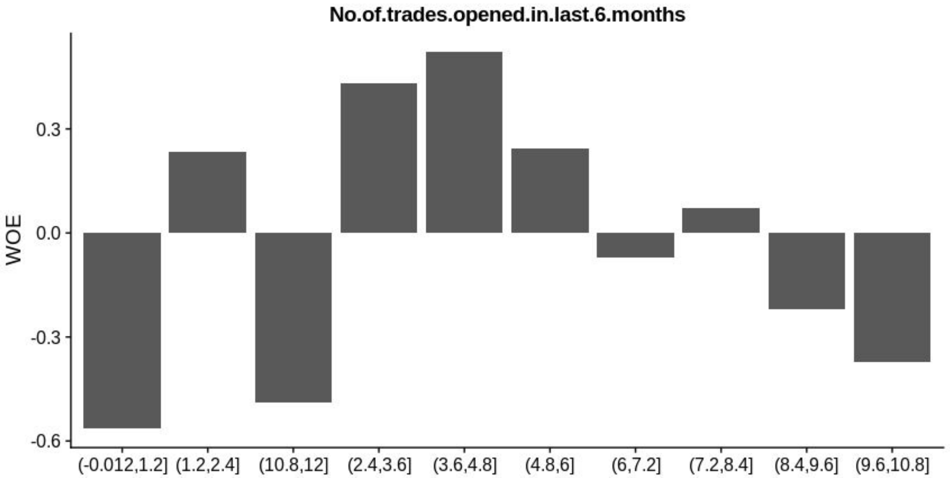
- WOE Transformation is carried out in all the variables and the Corresponding Information values are calculated to find the important variables.

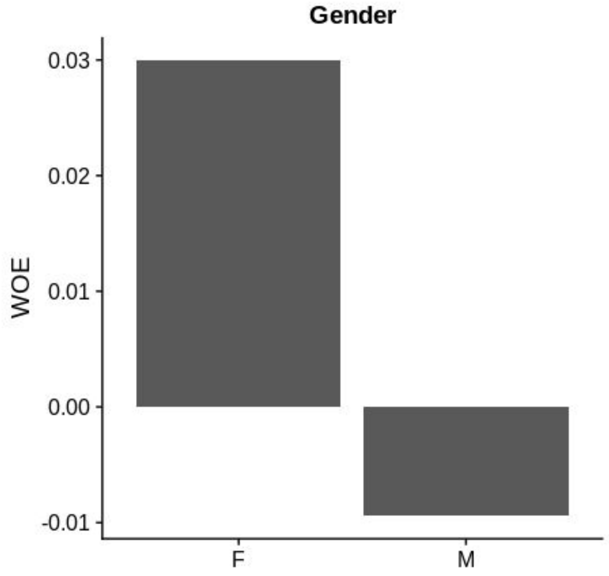
- The following are the identified important variables based on the Information Value. Considering IV values greater than 0.20 as significant.

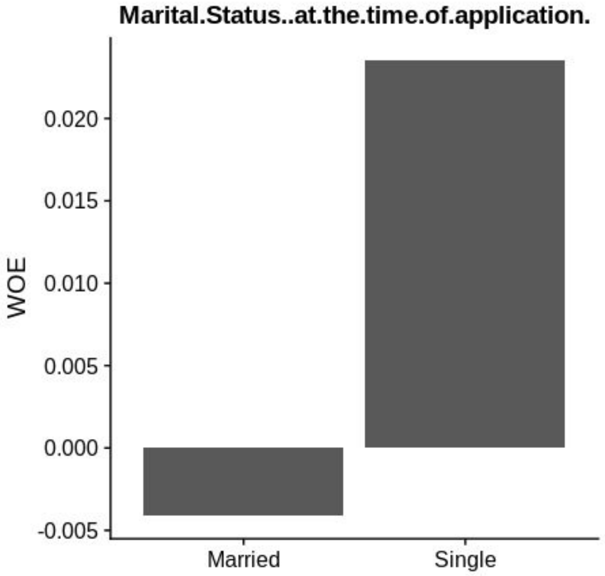
WOE Plots

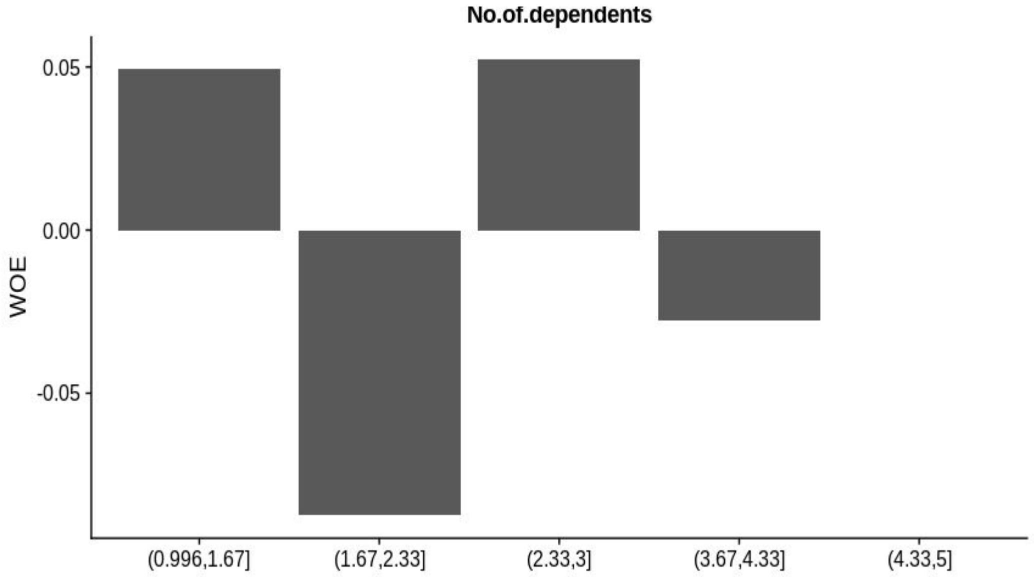

Age
WOE
0.0
-0.2
-0.4
-0.6
-0.8
(18,21.1]
(21.1,24.3]
(24.3,27.4]
(27.4,30.5]
(30.5,33.7]
(33.7,36.8]
(36.8,39.9]
(39.9,43.1]
(43.1,46.2]
(46.2,49.3]
(49.3,52.5]
(52.5,55.6]
(55.6,58.7]
(58.7,61.9]
(61.9,65]


Income
WOE
0.2
0.0
-0.2
-0.4
(-0.06,4]
(12,16]
(16,20]
(20,24]
(24,28]
(28,32]
(32,36]
(36,40]
(4,8]
(40,44]
(44,48]
(48,52]
(52,56]
(56,60.1]
(8,12]

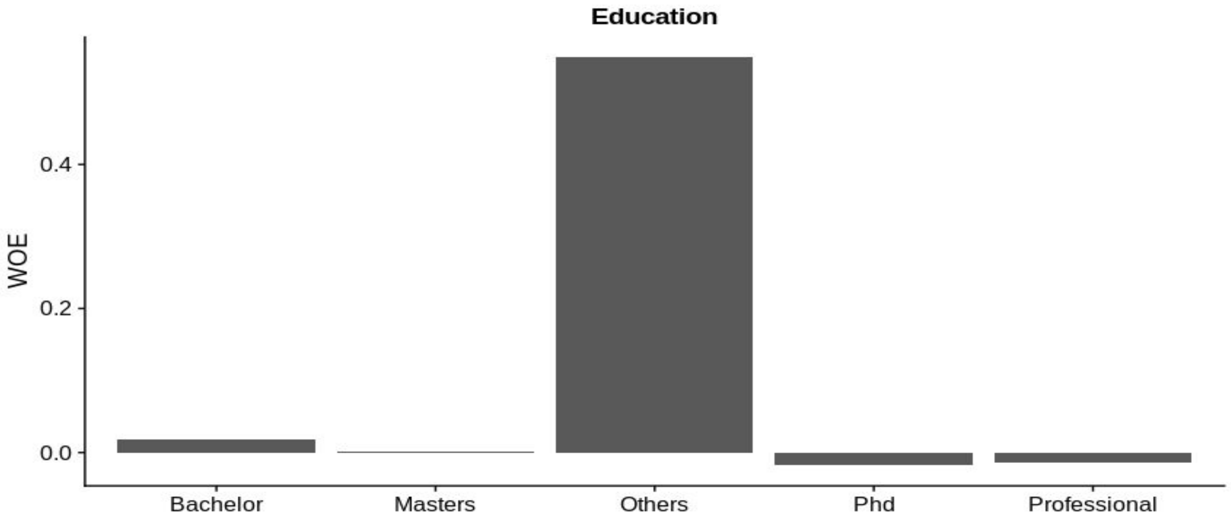

No.of.trades.opened.in.last.6.months
WOE
0.3
0.0
-0.3
-0.6
(-0.012,1.2]
(1.2,2.4]
(10.8,12]
(2.4,3.6]
(3.6,4.8]
(4.8,6]
(6,7.2]
(7.2,8.4]
(8.4,9.6]
(9.6,10.8]

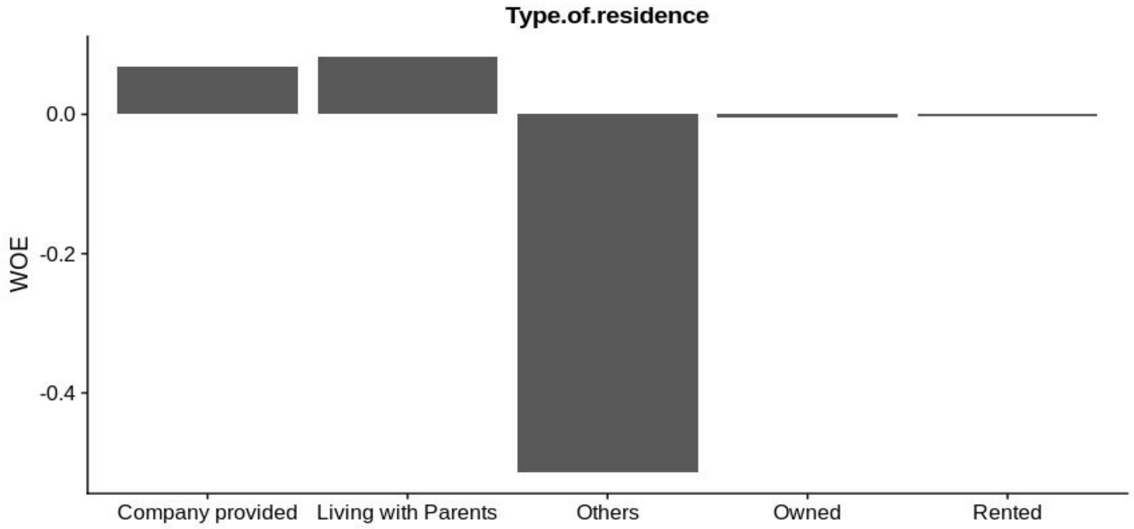

Gender
WOE
0.03
0.02
0.01
0.00
-0.01
F
M

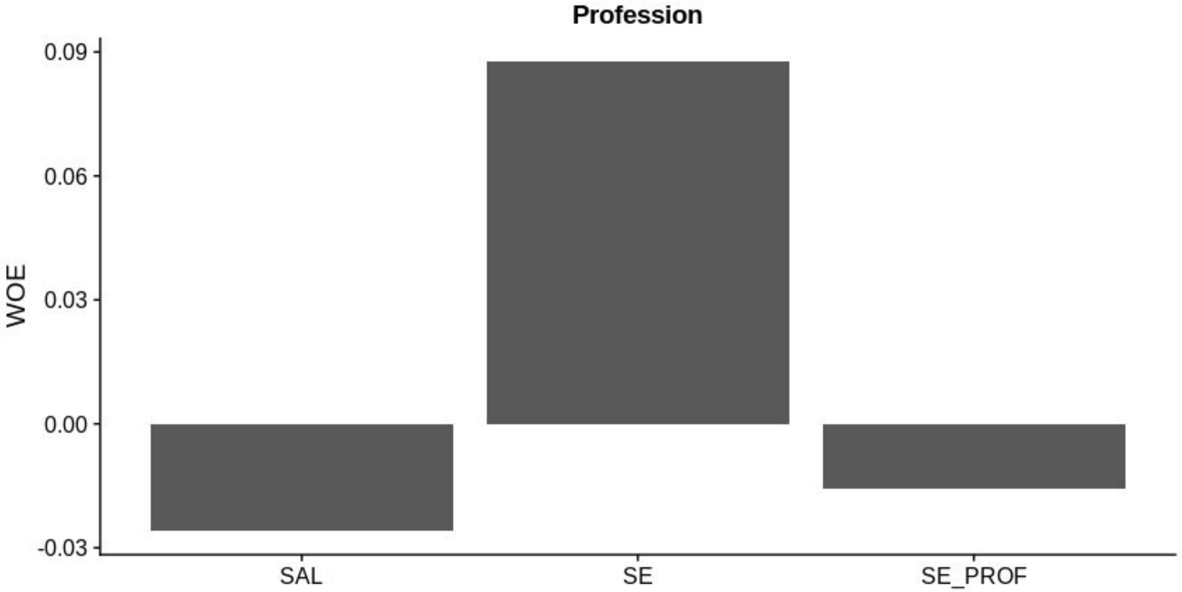

Marital.Status..at.the.time.of.application.
WOE
0.020
0.015
0.010
0.005
0.000
-0.005
Married
Single

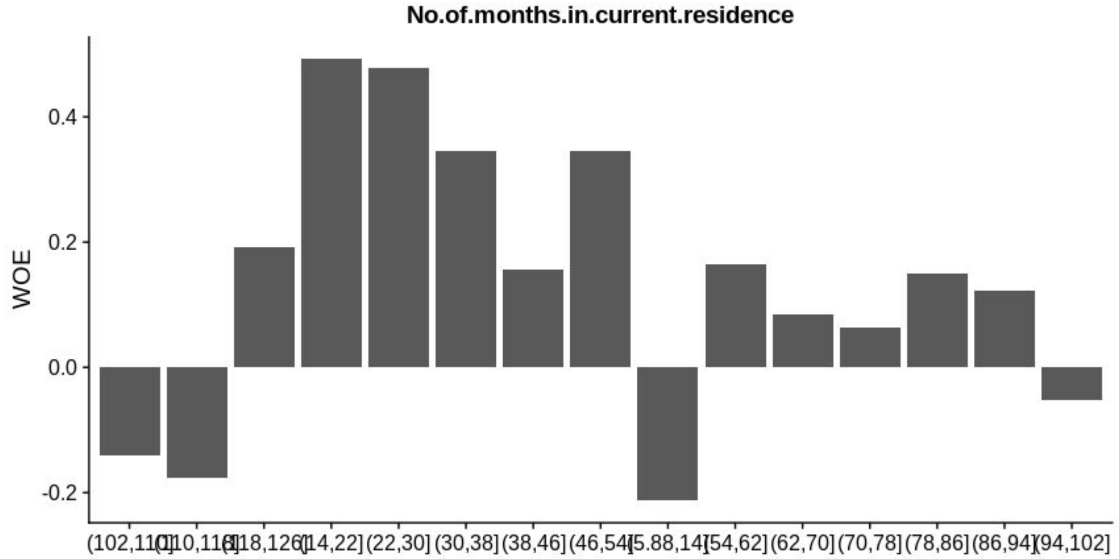

No.of.dependents
WOE
0.05
0.00
-0.05
(0.996,1.67]
(1.67,2.33]
(2.33,3]
(3.67,4.33]
(4.33,5]

WOE Plots

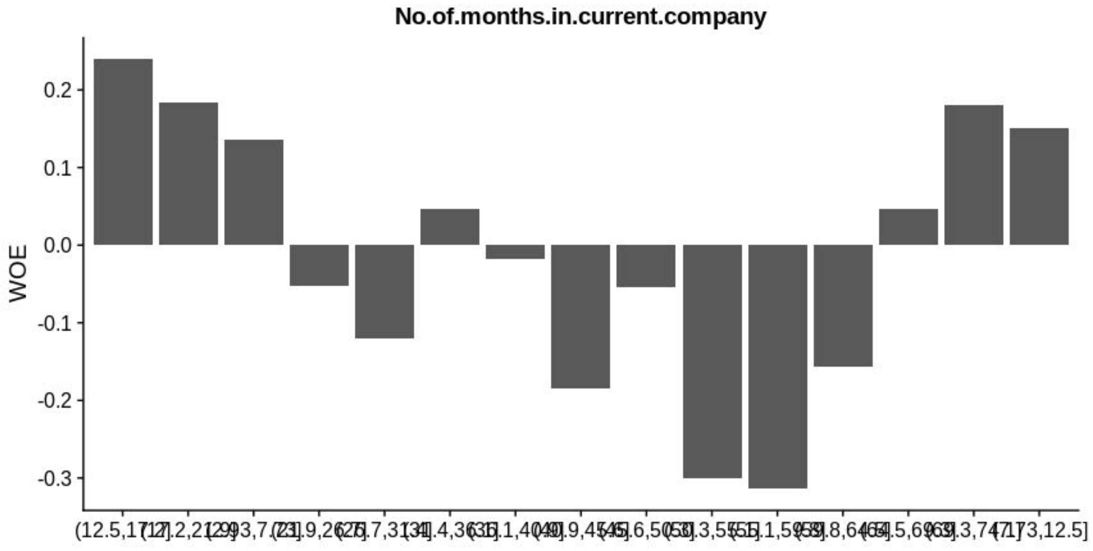

Education

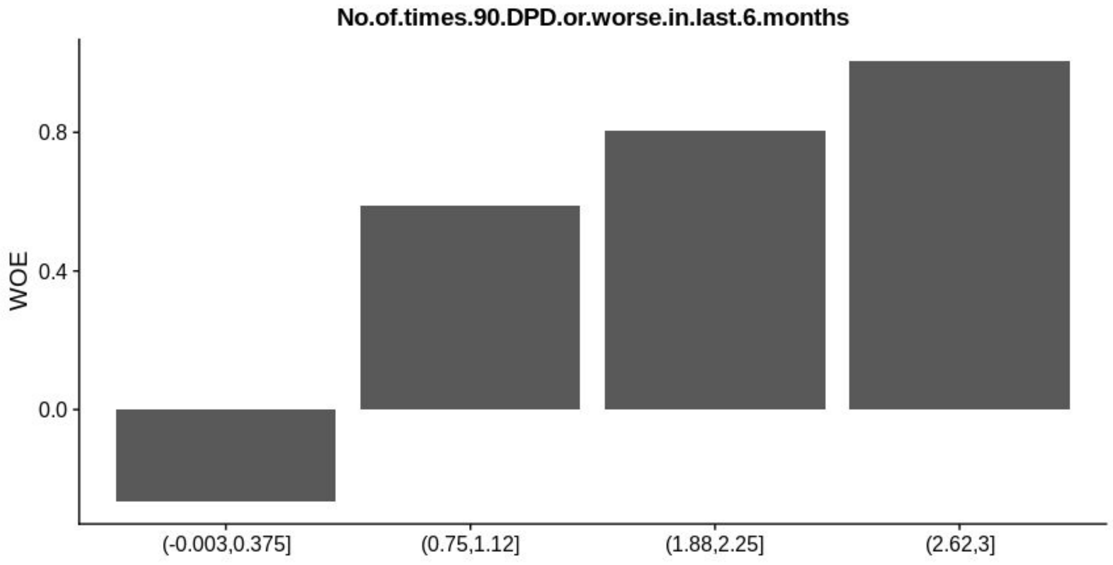

Type.of.residence

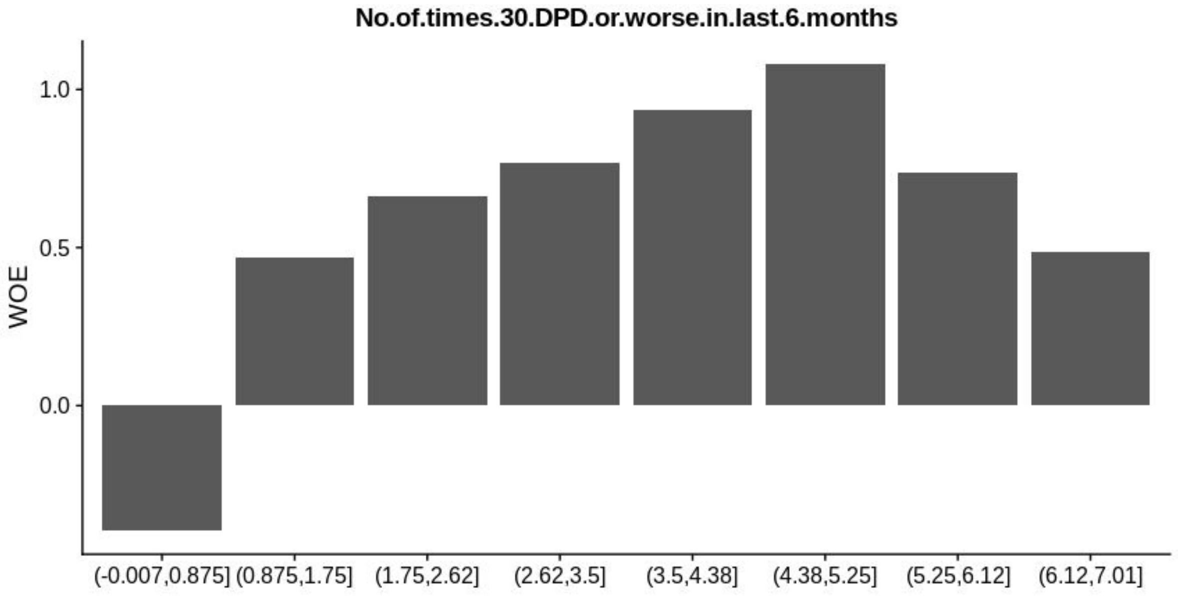

Profession

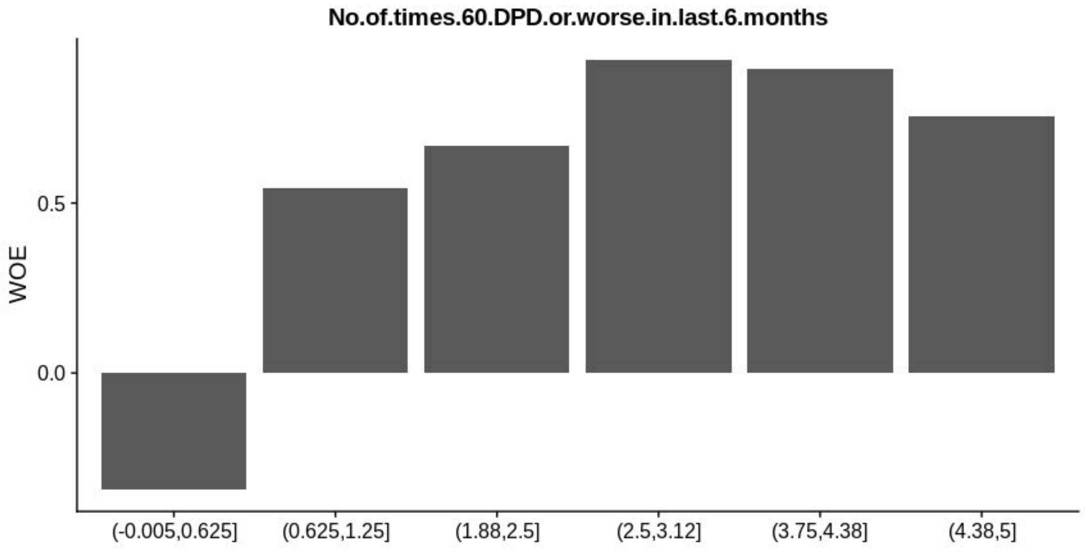

No.of.months.in.current.residence

WOE Plots

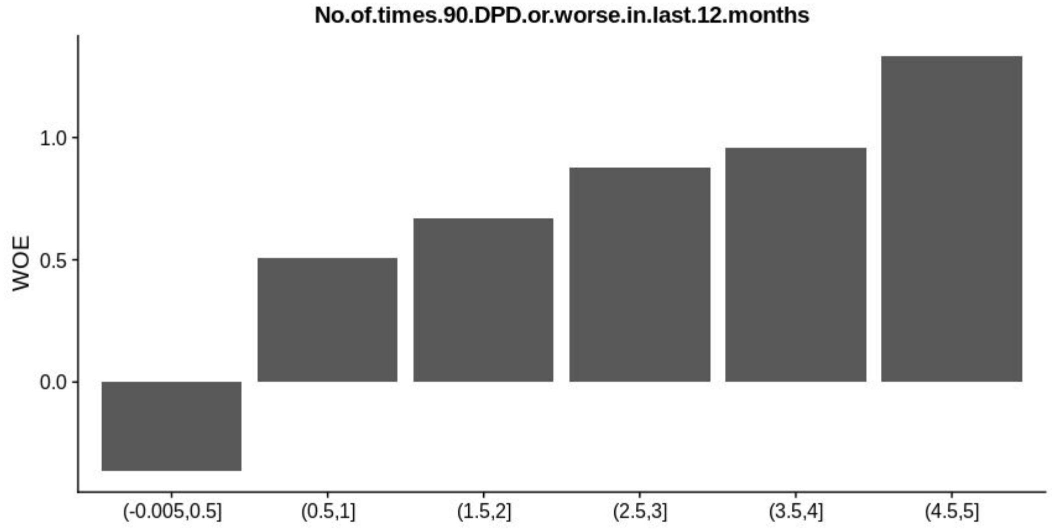

No.of.months.in.current.company

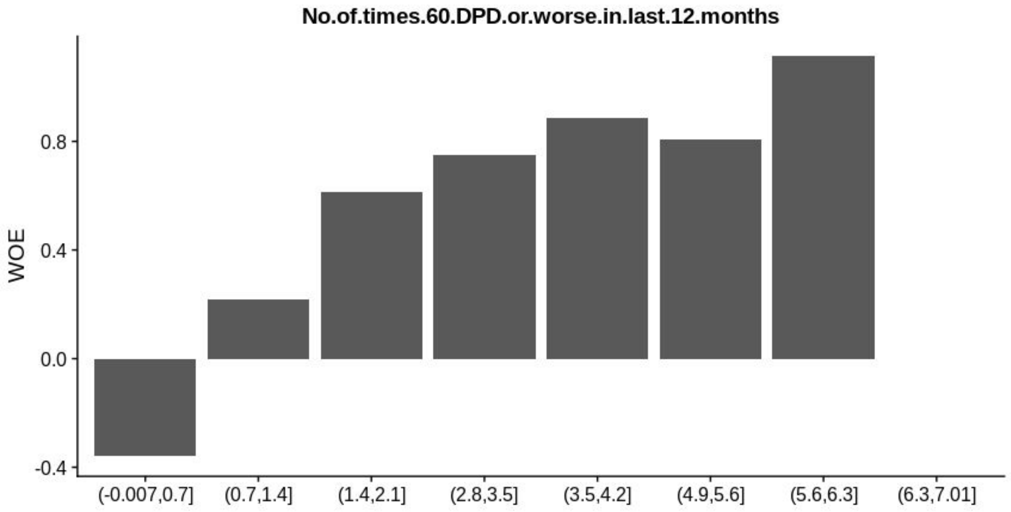

No.of.times.90.DPD.or.worse.in.last.6.months

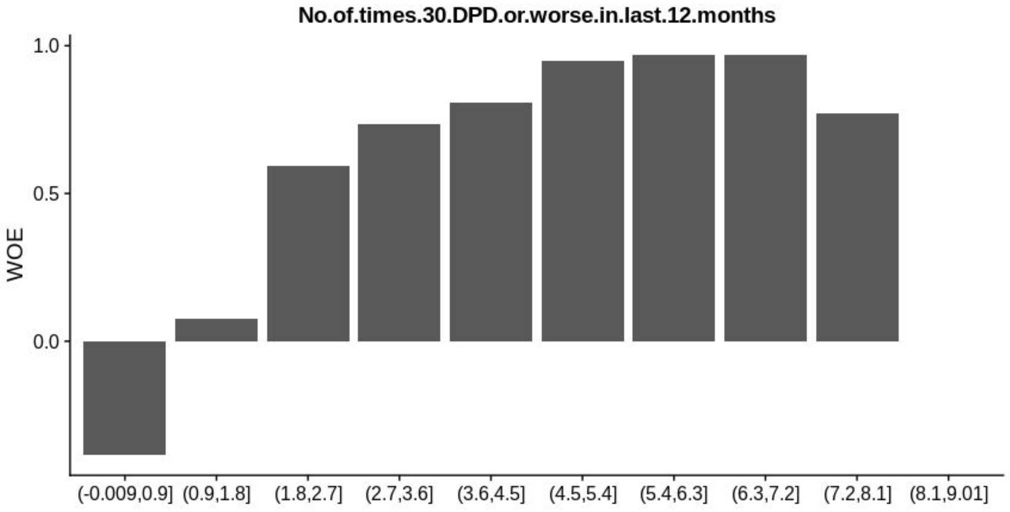

No.of.times.30.DPD.or.worse.in.last.6.months

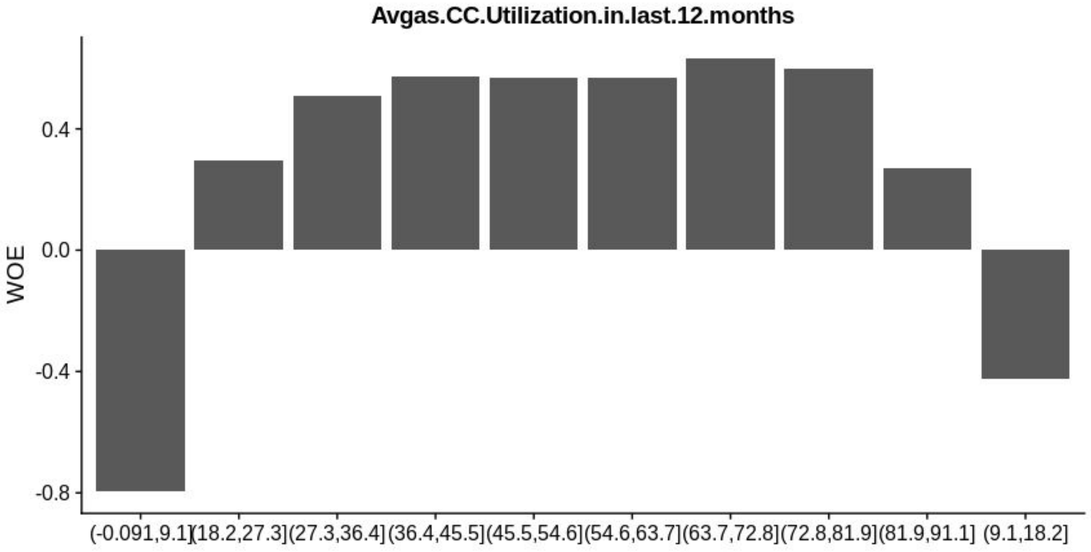

No.of.times.60.DPD.or.worse.in.last.6.months

WOE Plots


No.of.times.90.DPD.or.worse.in.last.12.months


No.of.times.60.DPD.or.worse.in.last.12.months


No.of.times.30.DPD.or.worse.in.last.12.months


Avgas.CC.Utilization.in.last.12.months

Model Building Procedure

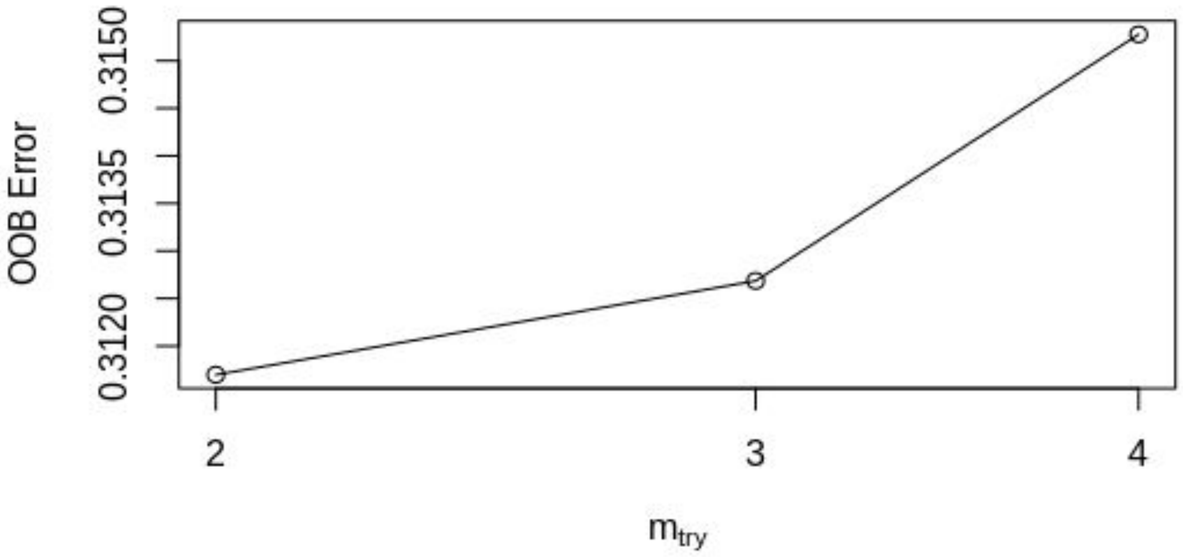
- For building the model we use both the Demographic Dataset as well as the Merged dataset of demographic and credit bureau data.
- First build the model using the cleaned dataset and check the accuracy.
- After that, apply the SMOTE method of synthetic data generation in order to handle the problem of class imbalance.
- Once the data is synthetically treated, use three different algorithms (Logistic Regression, Decision Tree and Random Forest) to build the models and determine the cutoff values to find the best values of accuracy, sensitivity and specificity for each model.
- Determine the best model based on R2, AIC, ROC, Accuracy, Sensitivity and Specificity values.

Final Model for Demographic Data- Random Forest

- Out of the Logistic Regression Model, Decision Tree and Random Forest Model, the Random Forest Model worked best on the Demographic Data and gave reasonable accuracy.

```
            Reference
Prediction    0      1
         0  15103   582
         1  4601    283
```

- Accuracy : 73.29%
- Sensitivity: 74.91%
- Specificity: 36.30%



```
 mtry  OOBError
2.OOB    2 0.3116979
3.OOB    3 0.3126851
4.OOB    4 0.3152764
```

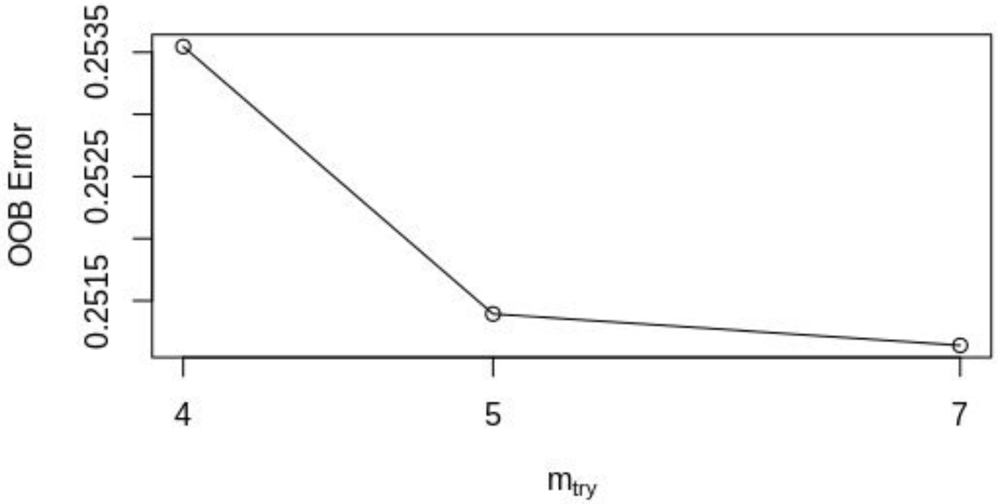
Optime mtry: 2

Final Model for Merged Data- Random Forest

- Out of the Logistic Regression Model, Decision Tree and Random Forest Model, the Random Forest Model worked best on the Demographic Data and gave reasonable accuracy.

```
              Reference
Prediction     0        1
         0   15471   576
         1   4181    341
```

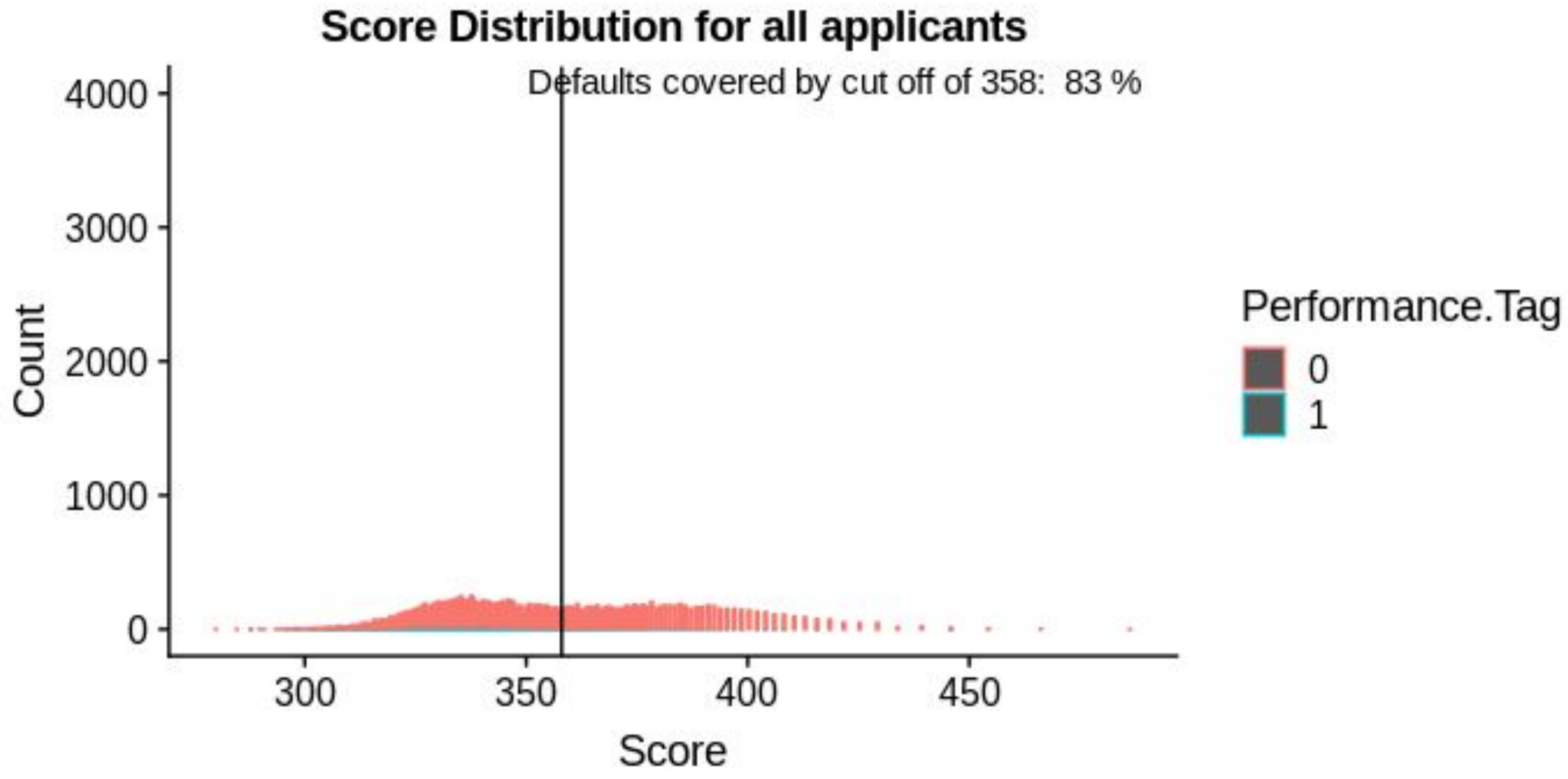
- Accuracy : 76.87%
- Sensitivity: 78.72%
- Specificity: 37.18%



```
 mtry  OOBError
4.OOB    4 0.2535461
5.OOB    5 0.2513931
7.OOB    7 0.2511398
```

Optime mtry: 7

Application Scorecard

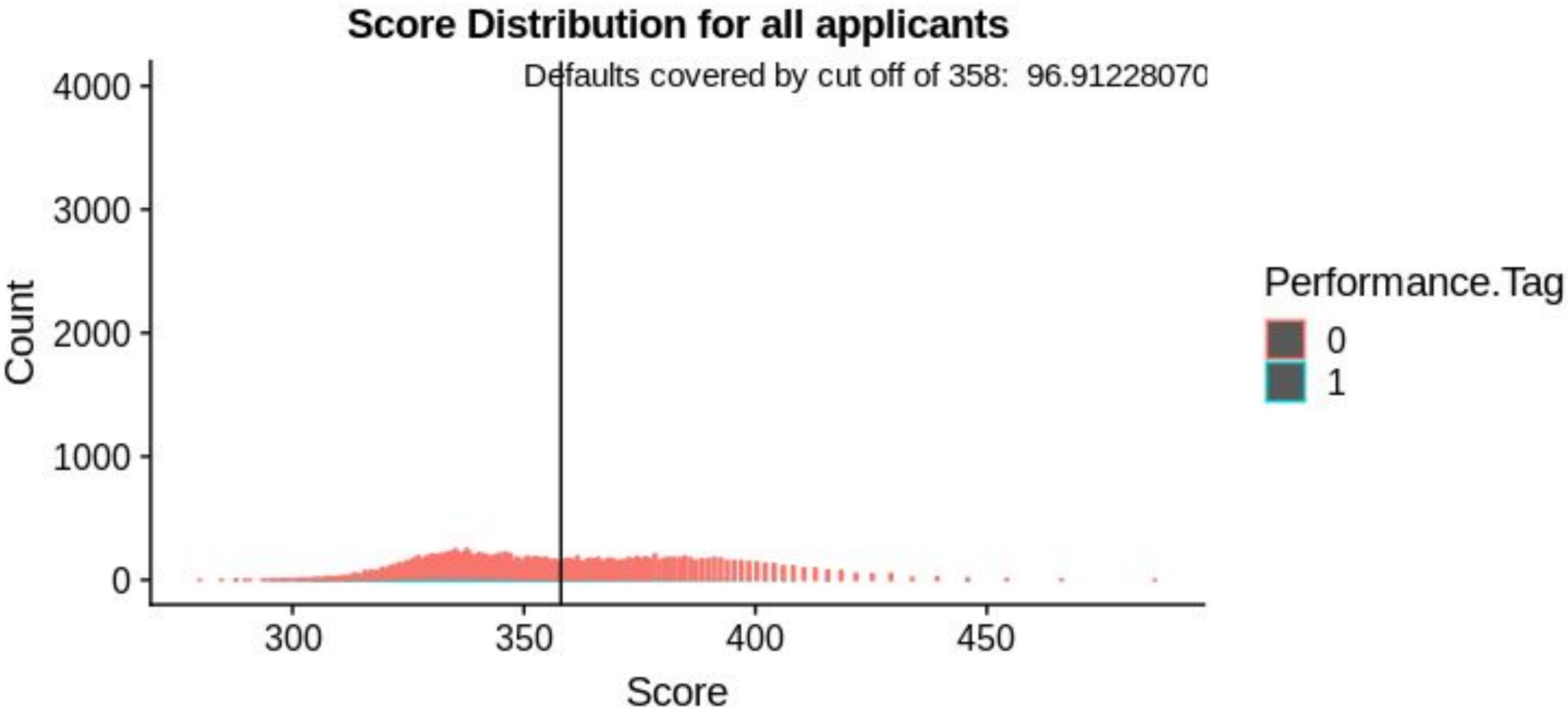


- Based on the analysis, the cutoff score is set to 358 which will help determine which applicants can be given credit card and which can't.

- More than 80% of the applicants are covered with the cutoff at 358

Percent:	0%	20%	40%	60%	80%	100%
Score:	180.8289	309.1136	327.7713	339.3516	358.0093	486.2939

Application Scorecard for Rejected Population



- Around 97% of the applicants are covered based on the cutoff value of 358
- This means we can provide credit cards safely to around 3% of the rejected population as well.

Financial Analysis and Implications of using the model

- The main aim of the Credit card provider is to mitigate the credit loss.
- The following model is built on the most important factors which are affecting the credit risk.
- If the following model is followed, using the balanced cutoff score (which helps find the right tradeoff between risk level and approval rate) which is determined to be 358, the company can attract as much customers as they can while avoiding most of the people who might default, thus reducing the credit risk.
- This cutoff score can be used as an important criteria to determine if an applicant should get the credit card issued or not.
- The management can re-evaluate all the applicants based on this cutoff score and, with this cutoff, they can ideally approve more than 80% of the applicants without worrying about the credit risk