CredX Acquisition Analytics

Application Scorecard

Team Details

- Vijay Choudhary
- Aman Kumar
- Anand Vaishnao
- Sarthak Gupta

Strategy and Business Objectives

Problem Statement

- To help CredX identify 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 your project.

Business Objective

- 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'.

Business Impact

- To mitigate financial losses caused due to default of a customer.
- Additionally, identify the factors which determine the riskiness of a customer (quantify the risk while issuing a credit card to a customer) .
- Creating Strategies which can help in reducing this acquisition risk and further assess the financial benefit of the same.

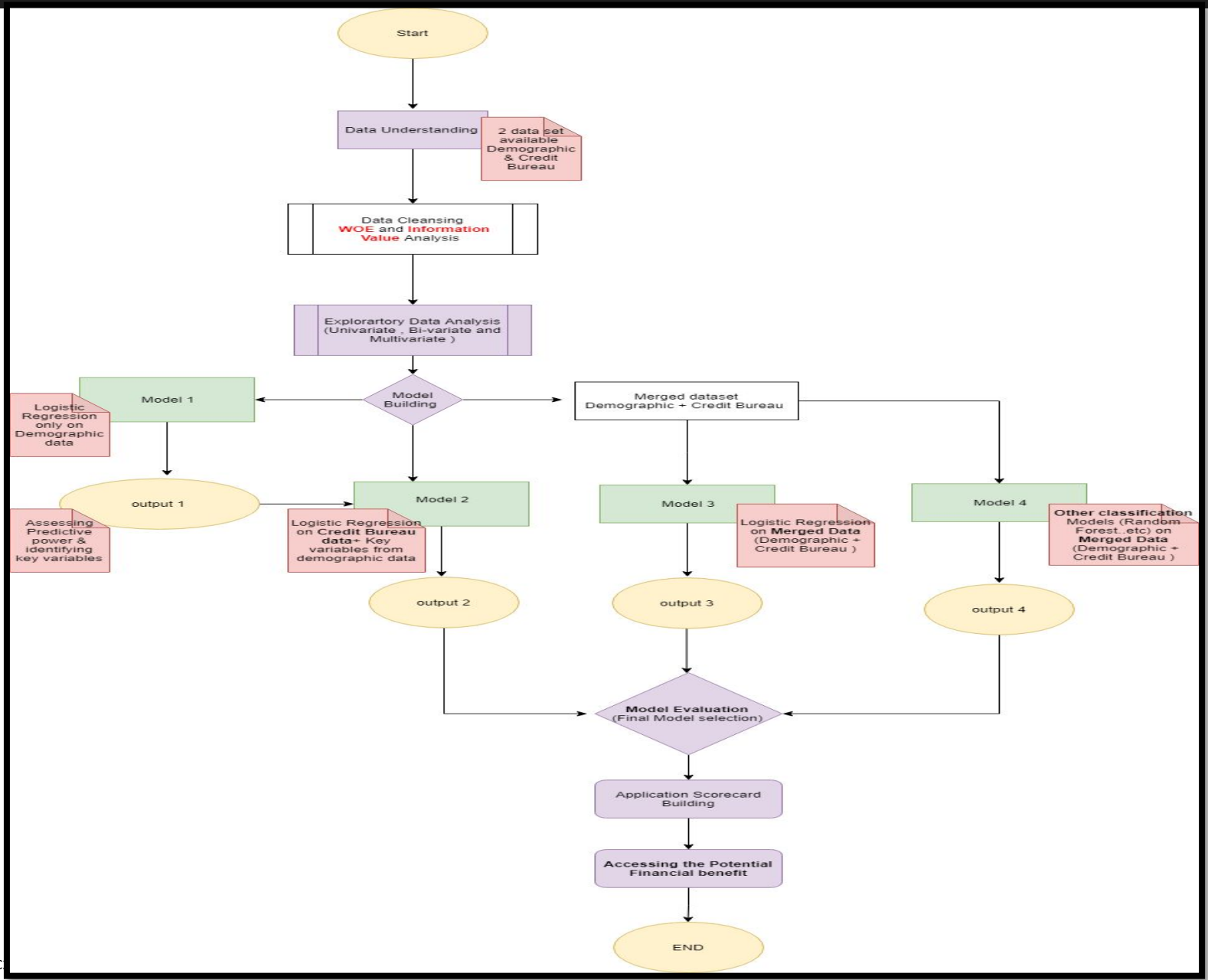
Data Understanding - 71292 unique customers records

Data

- 71292 unique customer data consist of 2 different datasets (**Demographic** and **Credit Bureau**)
- Both files contain a **performance tag** (target variable) 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.
- Total 1425 records (around **2%**) are having 'NA' values in 'performance tag' column.
- Total 66922 records (around 94%) are having '0' (non defaulted) in 'performance tag' column.
- Total 2948 records (around **4%)** are having '1' (**defaulted**) in 'performance tag' column.

Dataset	Records	Columns/Variables
Demographic data.csv	71292	12
Credit Bureau data.csv	71292	19

Approach – Future Roadmap



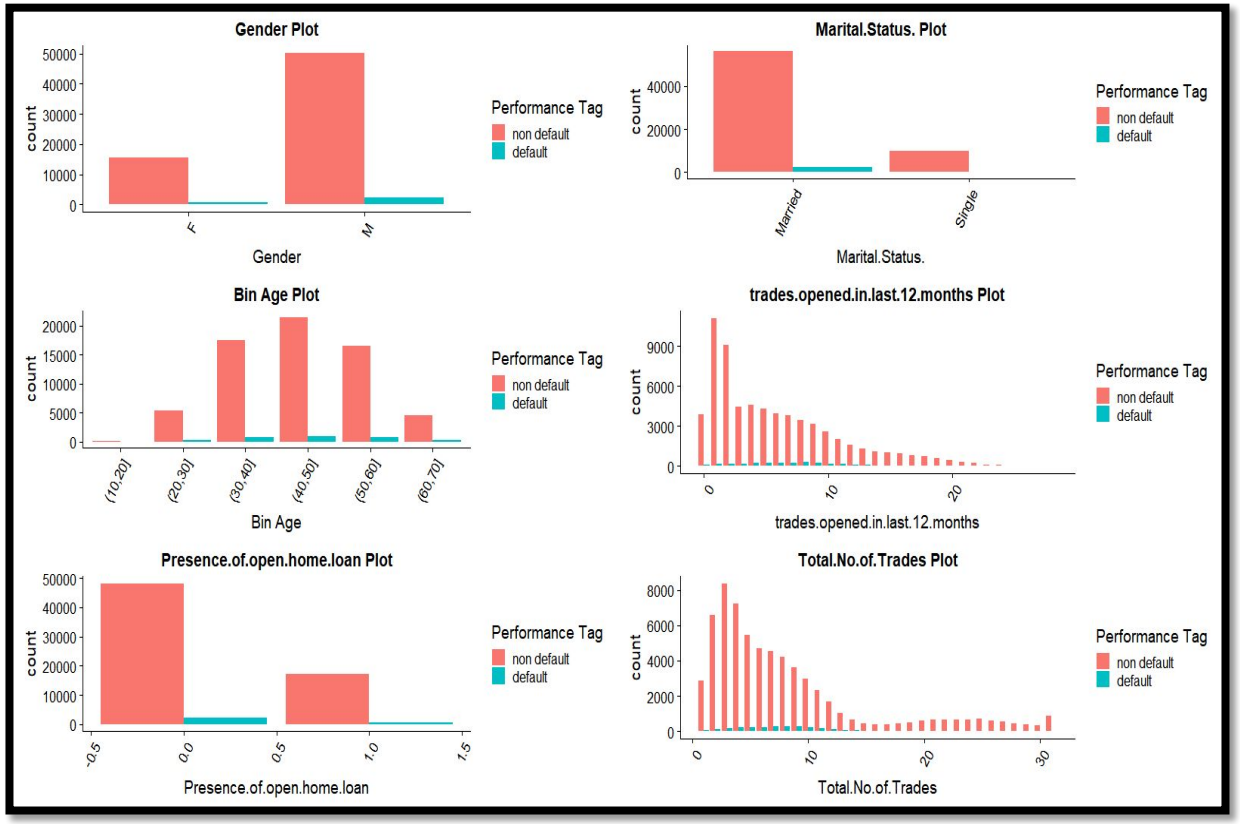
Highlights

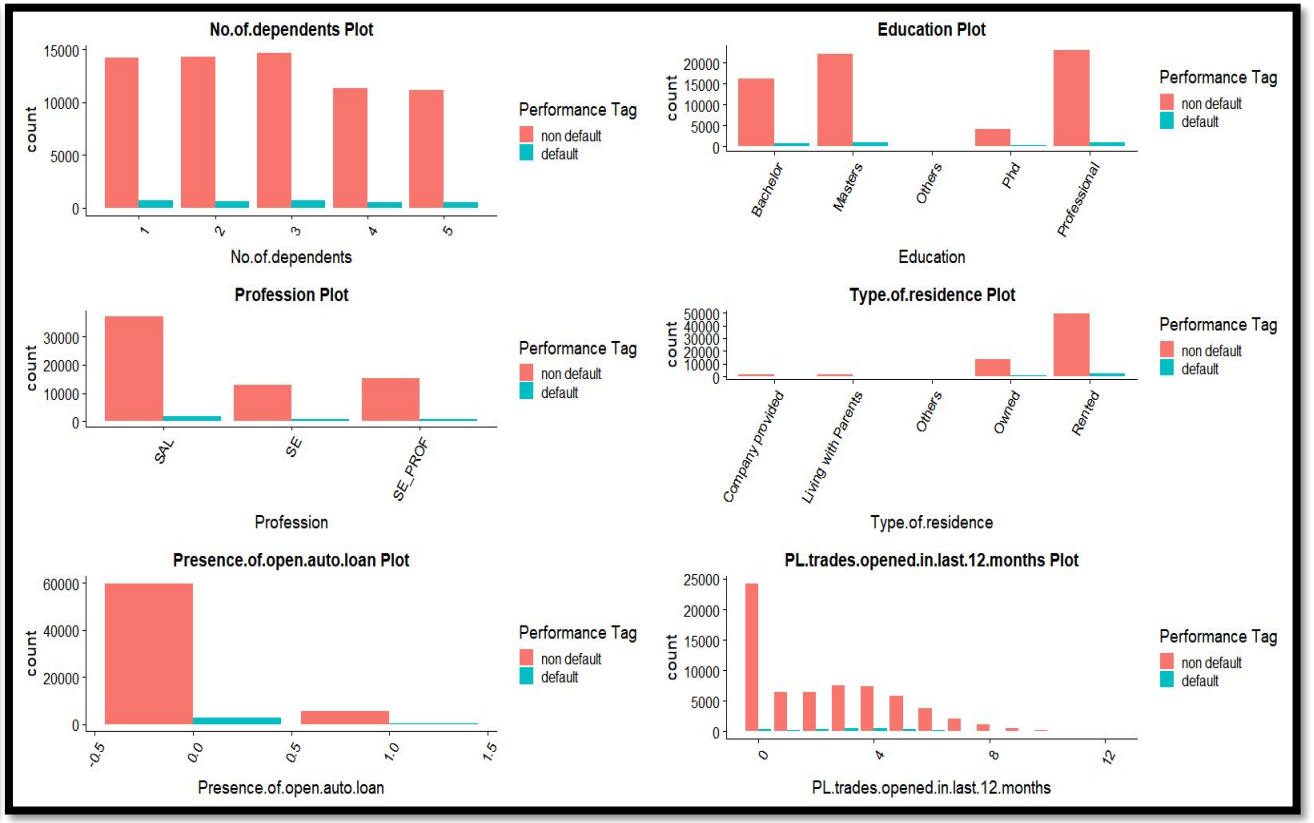
- **SMOTE:** Super sampling approach to handle the imbalanced classification problems
- **EDA:** Univariate , Bi-Variate, Correlation, WOE for all variables and IV plots.
- **Model**: Classification models like Logistic, Random Forest with their stats.
- **Model Evaluation:**
  Specificity, Sensitivity & Accuracy metrics
  KS Statistics, Lift & Gain plot
- **Application Scorecard**
- **Financial Assessment**

Data Cleaning and Assumptions

- Duplicate application id treatment before merging the 2 dfs (3 duplicates were removed)
- Outliers are present in the following variables : Age,No. of Months in current company,No. of Times 30/60/90 PDP or worse in last 6/12 months, No. of trades open in last 6/12 months, Total No. of trades , Total inquiry in last 6/12 months.
- Assumptions: age below 18yrs is minors hence removing them from data if application % is below 1 (65 applications)
- Income <= 0 removal if less than 1% applications present (0.15%, 106 applications).
- Load the data with blanks and “NA” included in the na.strings.
- NA treatment: ~1.9% in the dependent variable, removed.
- NA treatment: ~1% in column: “Avgas.CC.Utilization.in.last.12.months” which will be removed as well.
- As missing values are less than 2% hence not imputing WOE for the missing values.

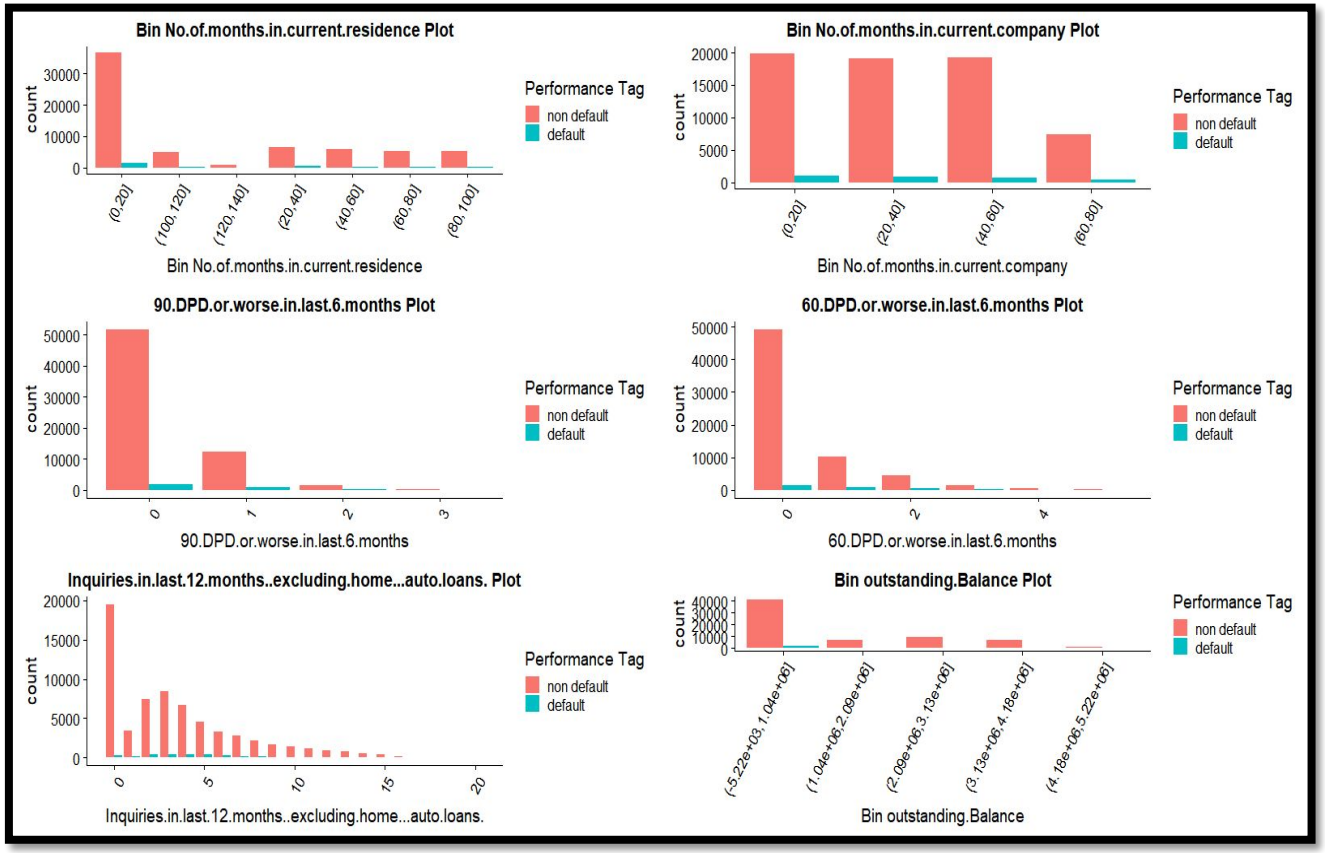
Exploratory Data Analysis- Variable influence on Target Variable

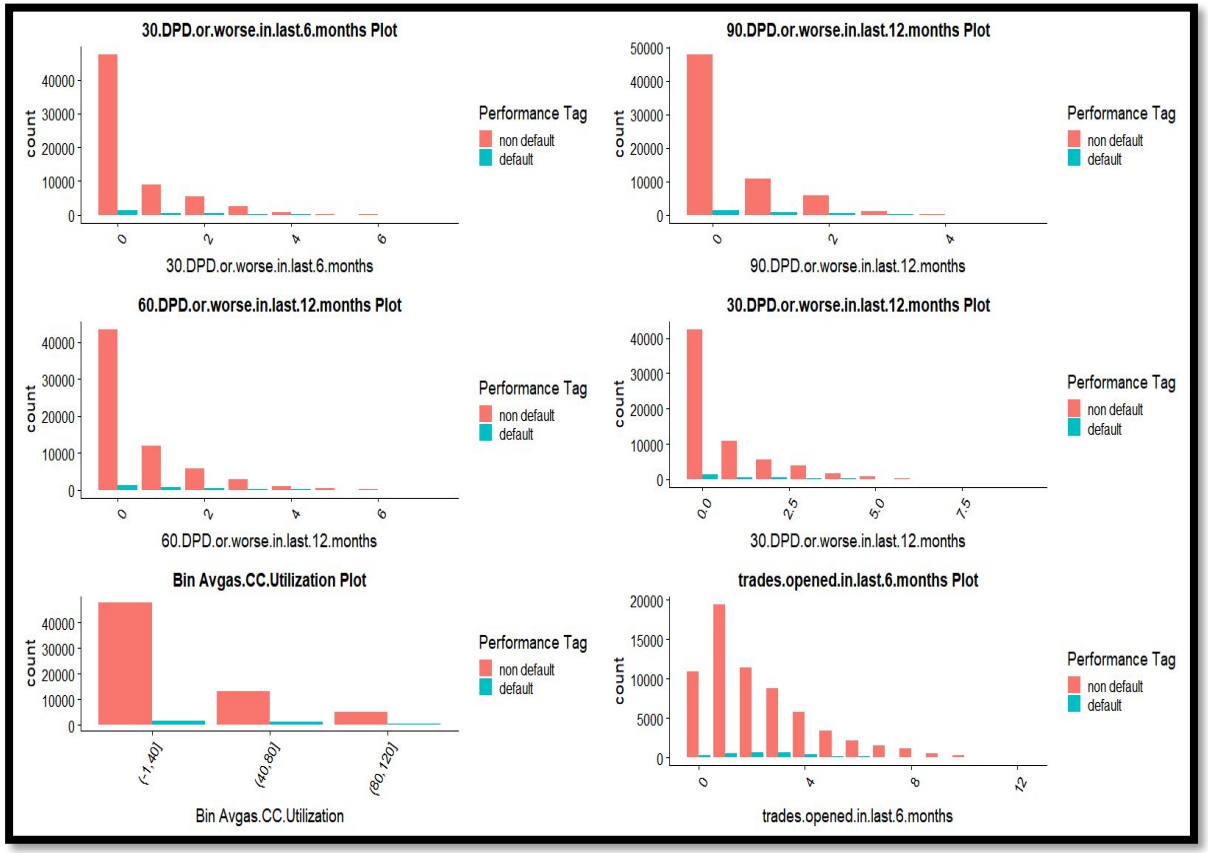




- All the variables are showing almost the same trend for Performance Tag

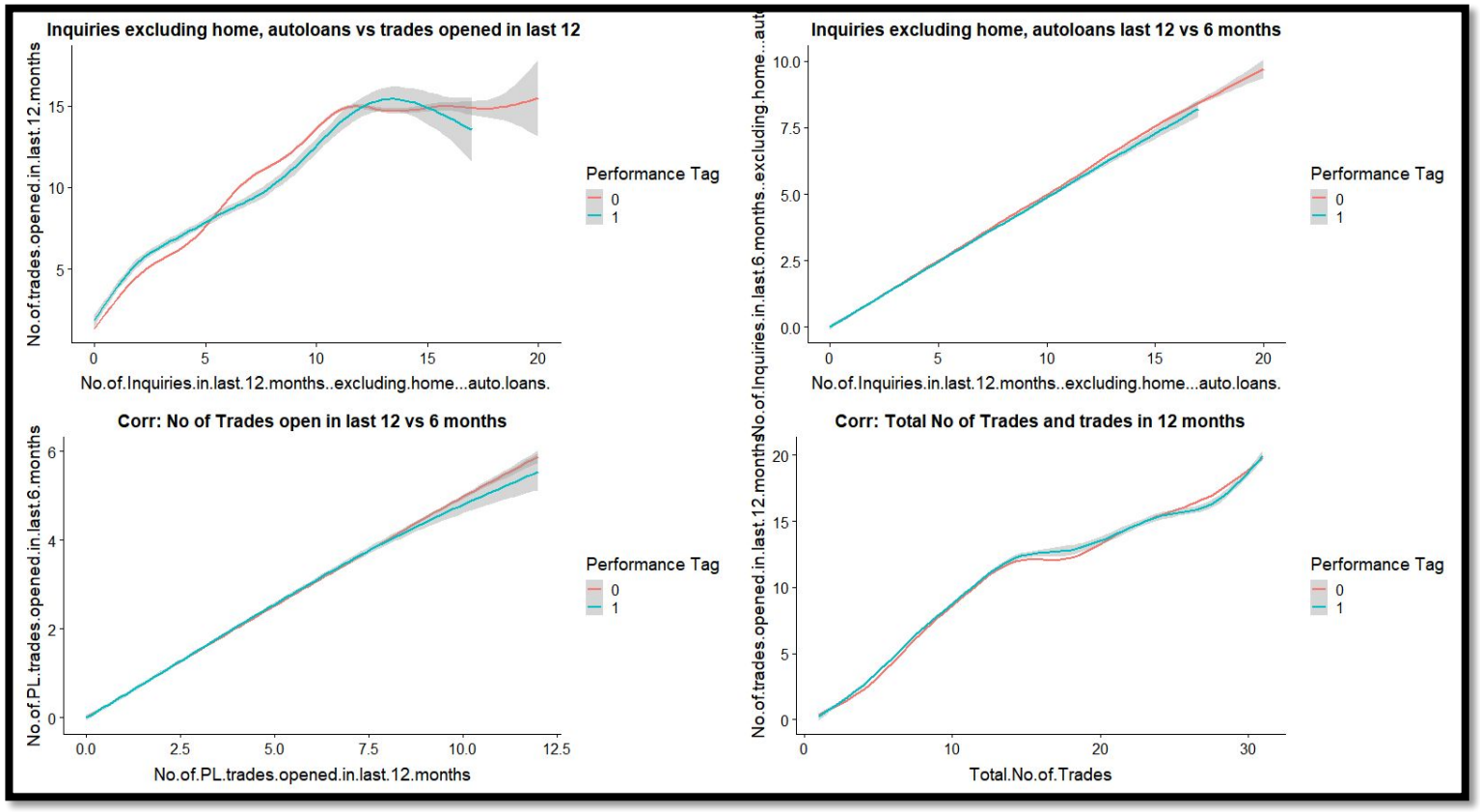
Exploratory Data Analysis- Univariate analysis

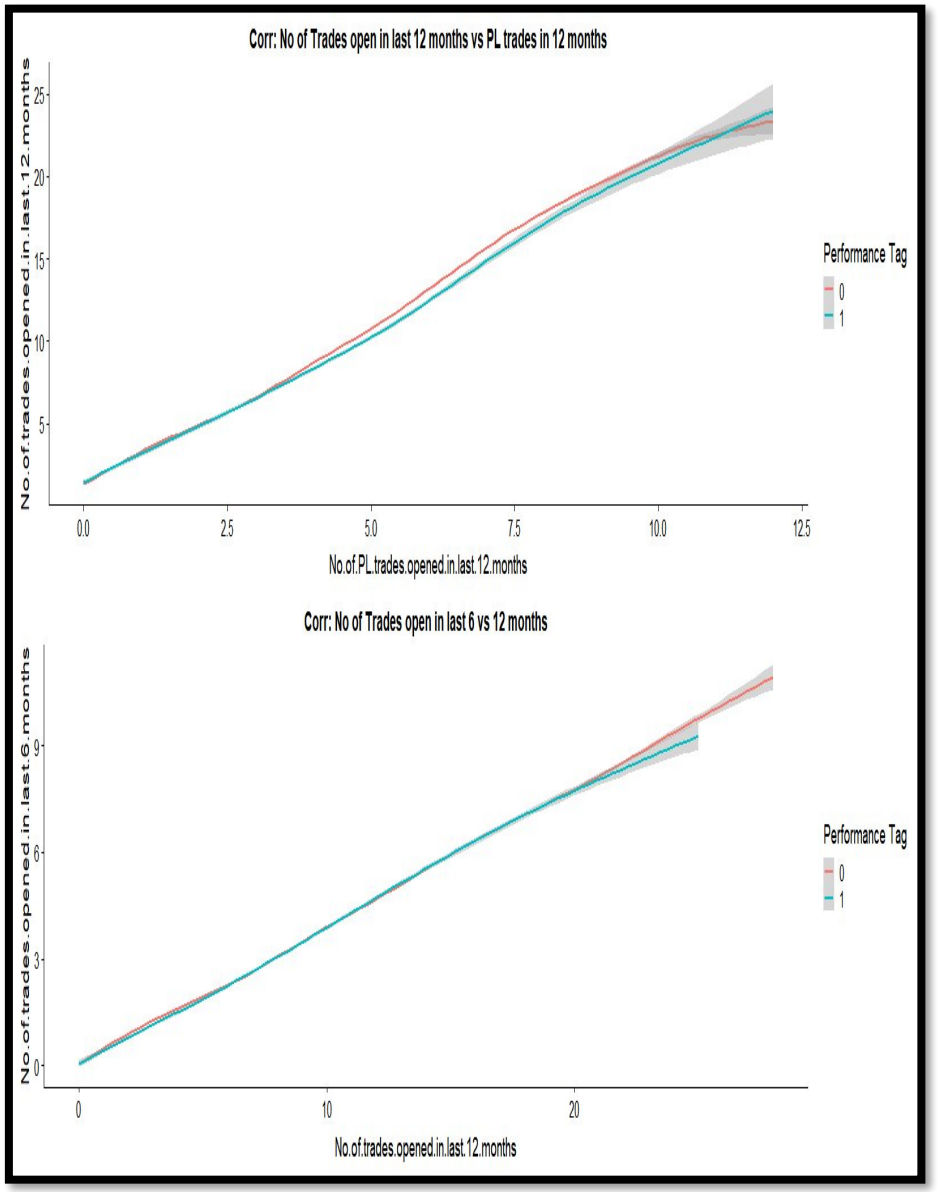




- All the variables are showing almost the same trend for Performance Tag

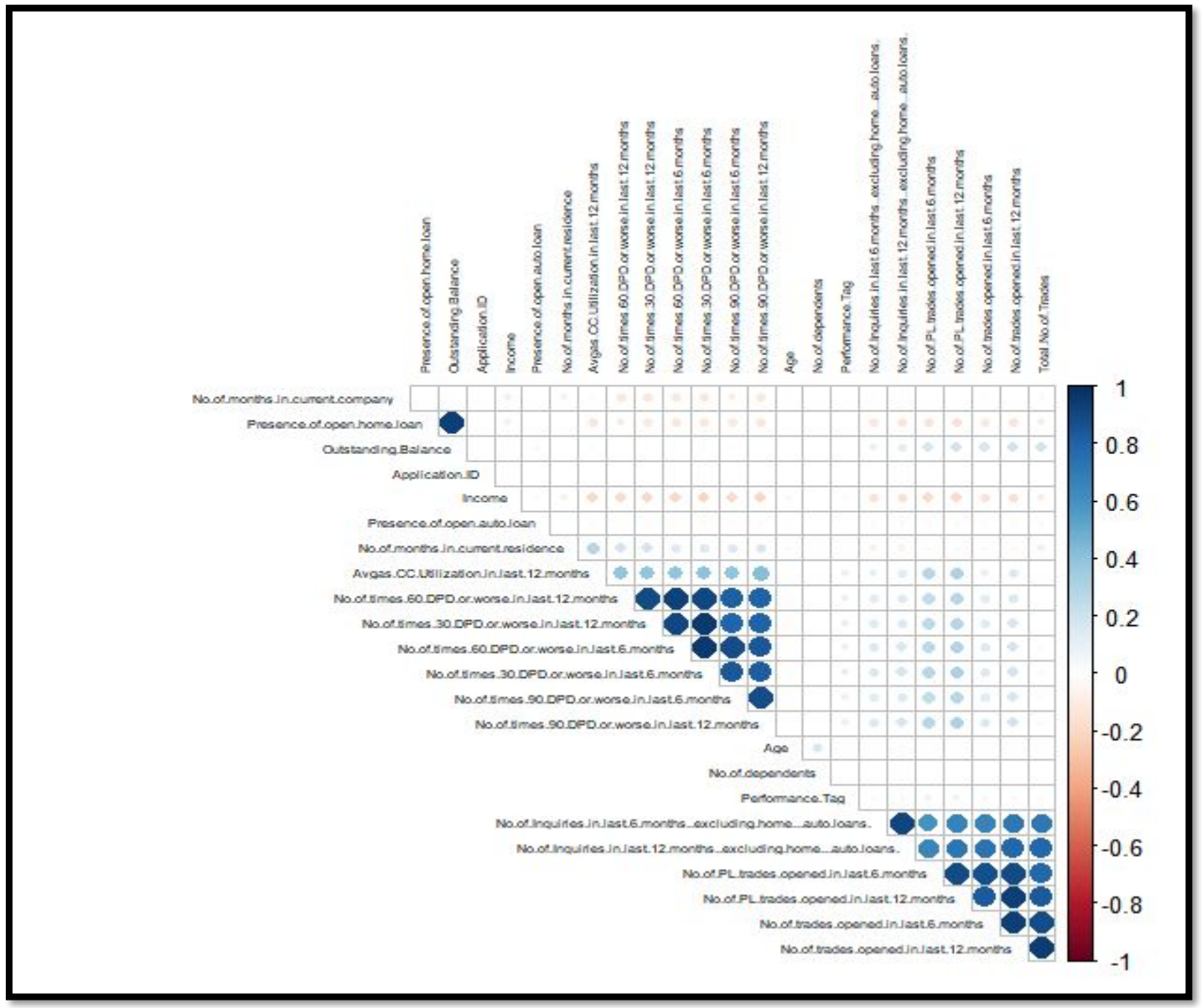
Exploratory Data Analysis- Bivariate Analysis





- Trend shows variables are highly correlated in the provided data set

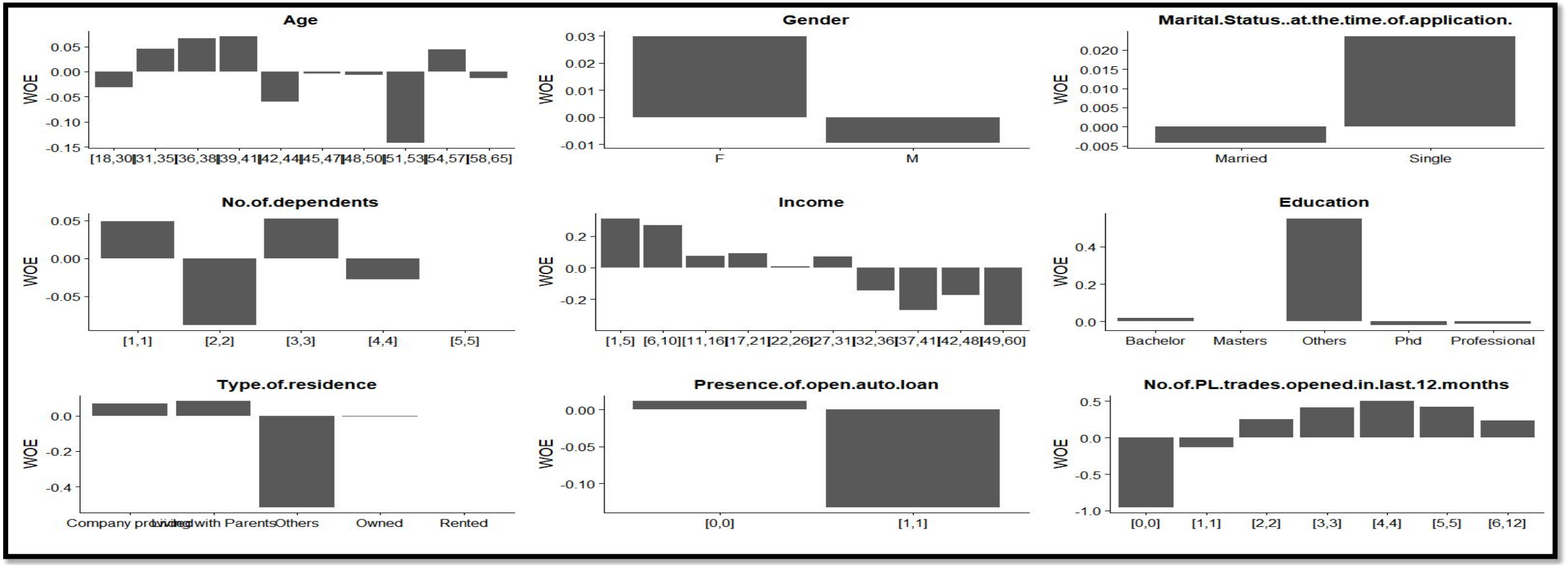
Data Preparation : Variable Correlations



Important High Correlation Variables Observed

- ✔ Outstanding Balance & Presence of Open home loan.
- ✔ All the variables of "No. of time DPD" are highly correlated with one another.
- ✔ Total No. of trades is correlated with all No. of inquiry ,No. of PL trade and No. of trade opened in last 6,12 months variables.

Exploratory Data Analysis- WOE & Information Value Analysis



- Most of the parameter have impact on the Performance tag except Education, so these parameters can be useful for model building

Final Model – Logistic with WOE + Smote Technique

Observation In Final Model

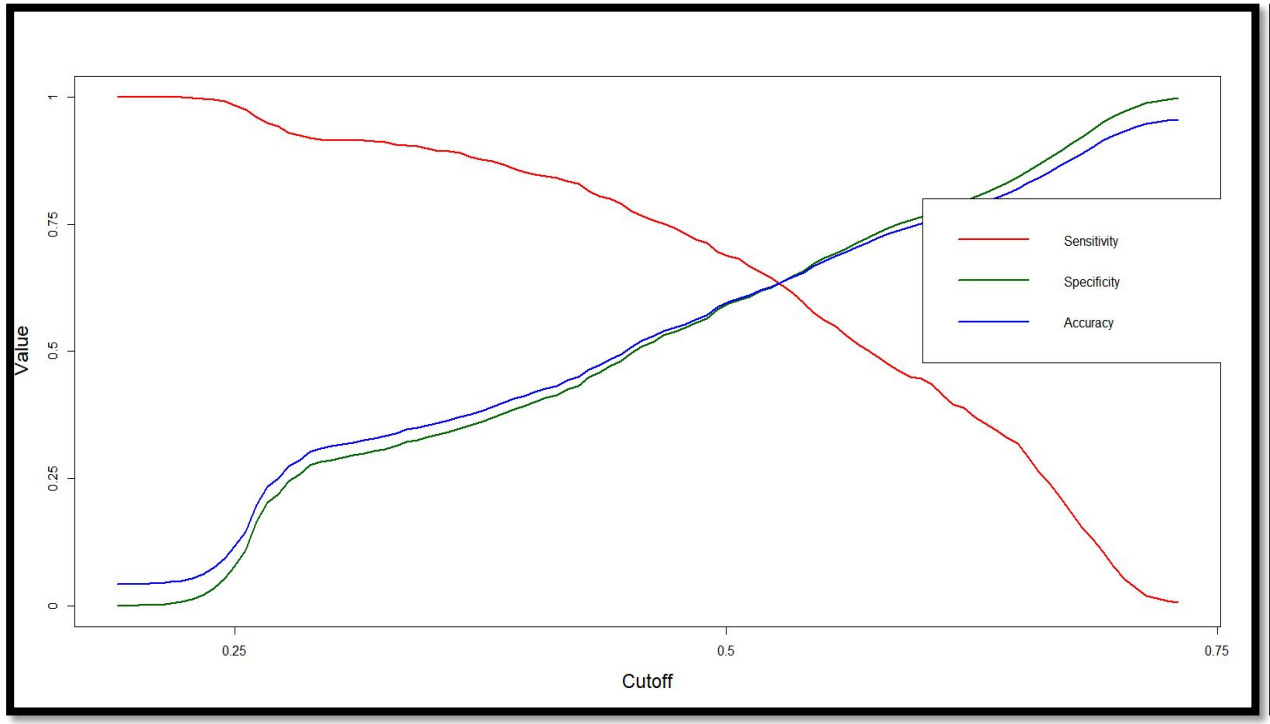
- **More chance of Default if**
  - No. of month in current residence is less.
  - No. of times 30 DPD or worse in last 6 months is more.
  - Avg CC Utilization in last 12 months is more.
  - No. of trades opened in last 12 months is more
  - No of Inquiries in last 6 months (excluding home & auto loans) is more.
  - Outstanding Balance is more.
  - No. of Presence of Open Auto loan is more

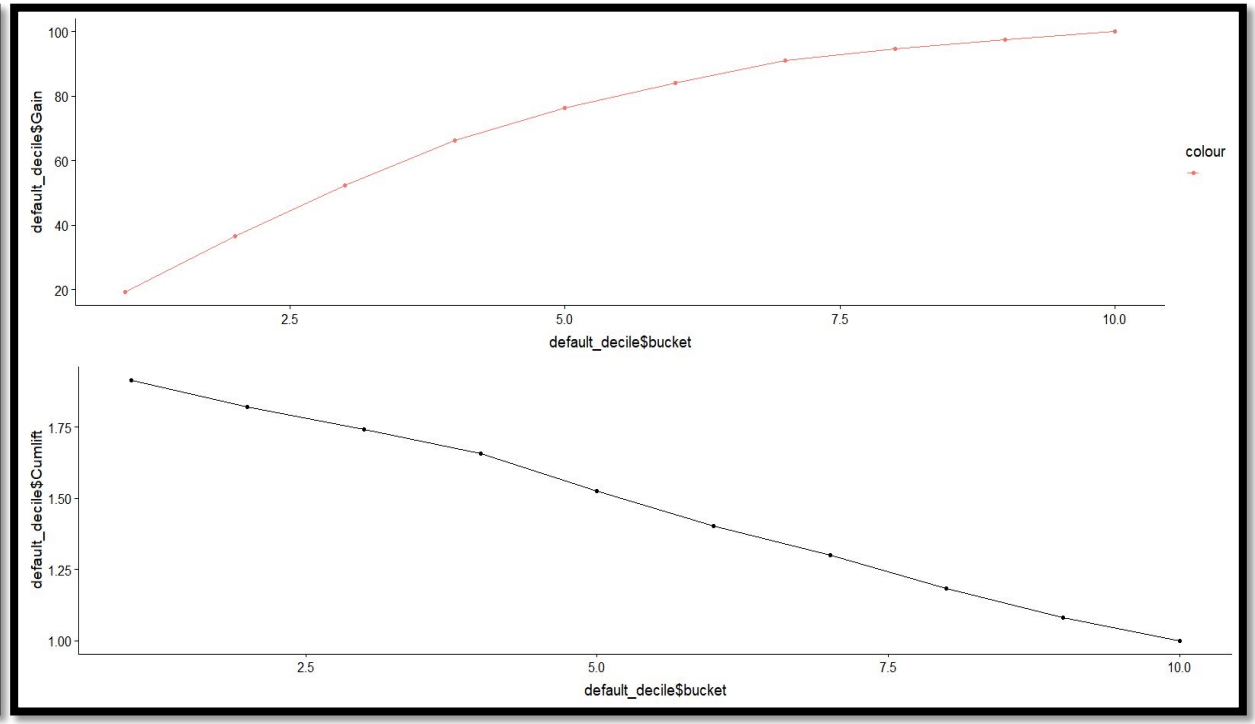
```
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-1.63478  -1.10215   0.07284   1.05785   1.75750

Coefficients:
                                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                                     -0.007136   0.023391  -0.305  0.76030
No.of.months.in.current.residence               -0.249395   0.099104  -2.516  0.01185
No.of.times.30.DPD.or.worse.in.last.6.months     0.452984   0.059593   7.601 2.93e-14
Avgas.CC.Utilization.in.last.12.months           0.364138   0.064730   5.625 1.85e-08
No.of.trades.opened.in.last.12.months            0.223775   0.077672   2.881  0.00396
No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.  0.342770   0.066068   5.188 2.12e-07
Outstanding.Balance                              0.216075   0.077713   2.780  0.00543
Presence.of.open.auto.loan                       1.670374   0.619663   2.696  0.00703
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11223  on 8095  degrees of freedom
Residual deviance: 10438  on 8088  degrees of freedom
AIC: 10454

Number of Fisher Scoring iterations: 4
```

Model Building & Evaluation

We have successfully build a logistic regression model to predict the probability of defaulters

- Final model was achieved in 9 iterations at cutoff value is 0.53
- There are total 7 key variables.

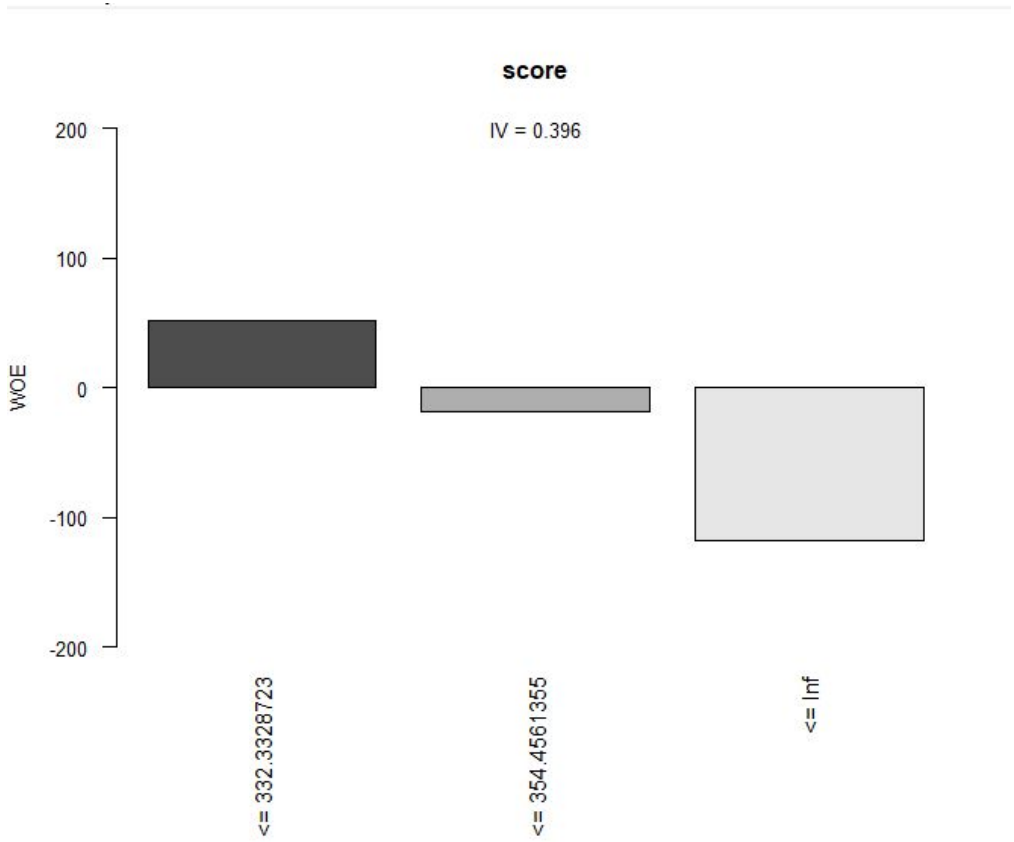




- ❑ Accuracy: 64 %
- ❑ Sensitivity(True Positive Rate): 63%
- ❑ Specificity(True Negative Rate): 64%

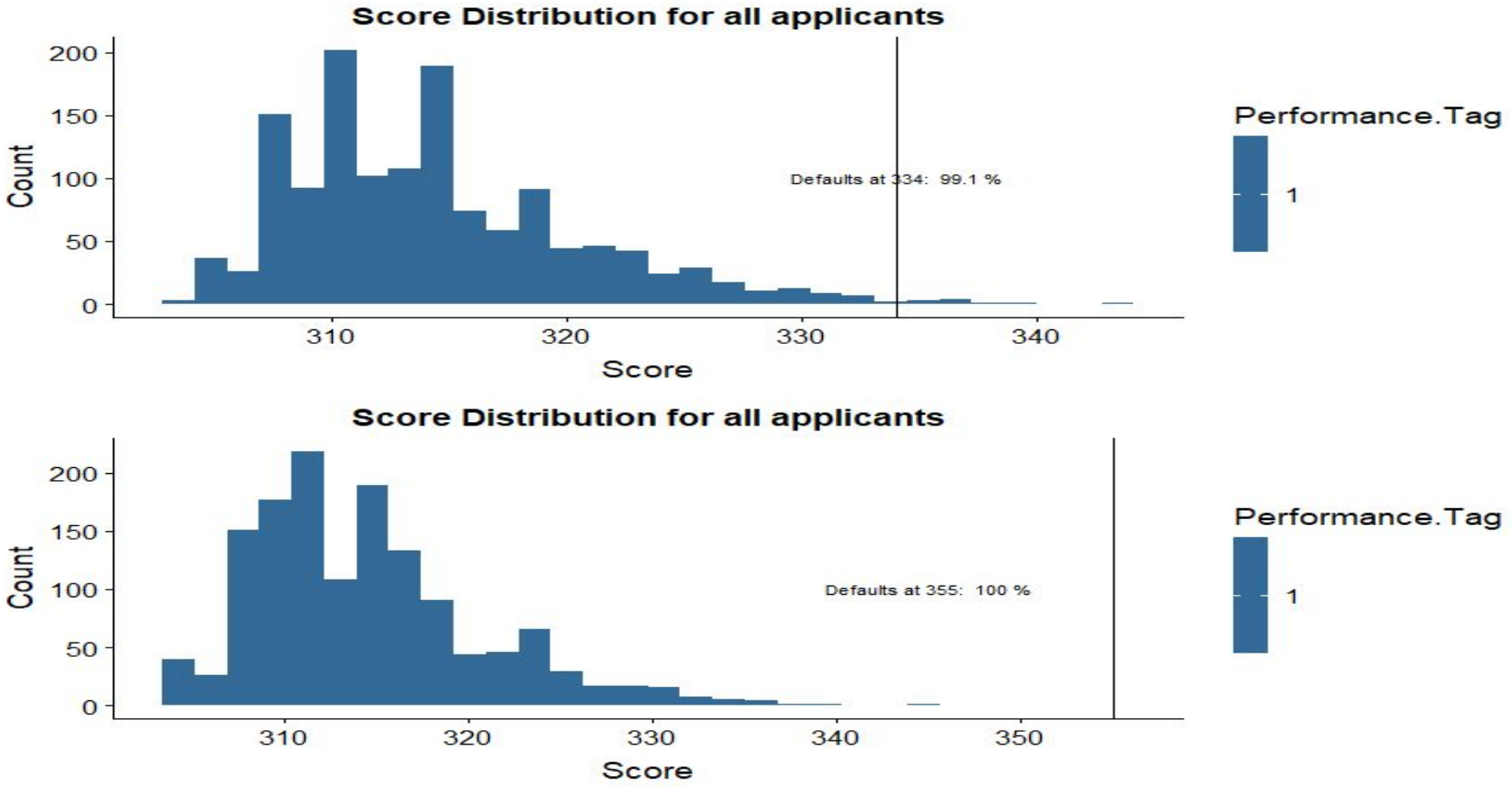
- ❑ Gain at 4th Decile: 66.3%
- ❑ Lift at 4th Decile: 1.66%

Application Scorecard Building



- Formulae used for Scorecard Calculation is : **333.5614 + (28.8539 * score_card_data$odds)**
- Got the Scores between 303 to 376 and got cut off score as 335 and 355.
- Cutoff 355 : 91% of applicants had score more than the cutoff score.
- Cutoff 334 : 68% of applicants had score more than the cutoff score.

Scorecard Evaluation for rejected applicants



- Ran the model on the Rejected applicant data set and none of them got score more then cutoff score.
- Score ranged from 304 to 344 for this data set of ~1400 applicants
- for score 355: : 100% applicants who defaulted were captured
- for score 335: 99.1% defaulted applicants were captured

Financial Benefit Assessment

- Assumptions to choose cutoff value for scorecard:
  - Bank losses 1,000 INR per credit card per customer for a rejected applicant who did not default
  - Per credit card default, bank losses 15 times what it makes as revenue (Reason: default generally happens when amount is high) ##i.e 15,000 INR per default
- Metrics used to decide cutoff:
  - **Revenue loss** = No of candidates rejected by the model who didn't default / Total No of candidates who didn't default
  - **Financial benefit by the model** = correctly_defaulted_customer_prediction*default_loss - non_default_rejections*good_customer_loss
- For cut_off: 355:
  - **Revenue loss =** 45615/45615+20030 = **~70%**
  - **Financial benefit by the mode** = 2623*15000 -45615*1000 = **-6,270,000 INR**
- For cut_off: 334
  - **Revenue loss =** 27102/(27102+38543) = **~41%**
  - **Financial benefit by the mode =** 1954*15000 -27102*1000 = **+2,208,000 INR**
- We can clearly see here that based on the assumptions made, the cut off of 334 is optimum for bank to be profitable from the model made.

Thank You