BFS Capstone Project

BFSI

Group Members:

1. **Raju Kumar (Group Facilitator)**
2. **Sangram Jethy**
3. **Manish Mani**
4. **Sidharth Mahapatra**

Content

Overall Approach

- Overview/Approach
- Data Preparation
- Data Transformation
- Model Building
- Model Evaluation
- Model Selection

Results & Observations

- Information Value
- EDA
- Lift and Gain
- Score Card Calculation
- Financial Benefit Analysis

Overall Approach

Overview

Project Background

- Our bank receives thousands of credit card applications from different types of customers every year.
- In the past few years we have experienced an increase in credit loss due to acquisition of more number of bad customers.

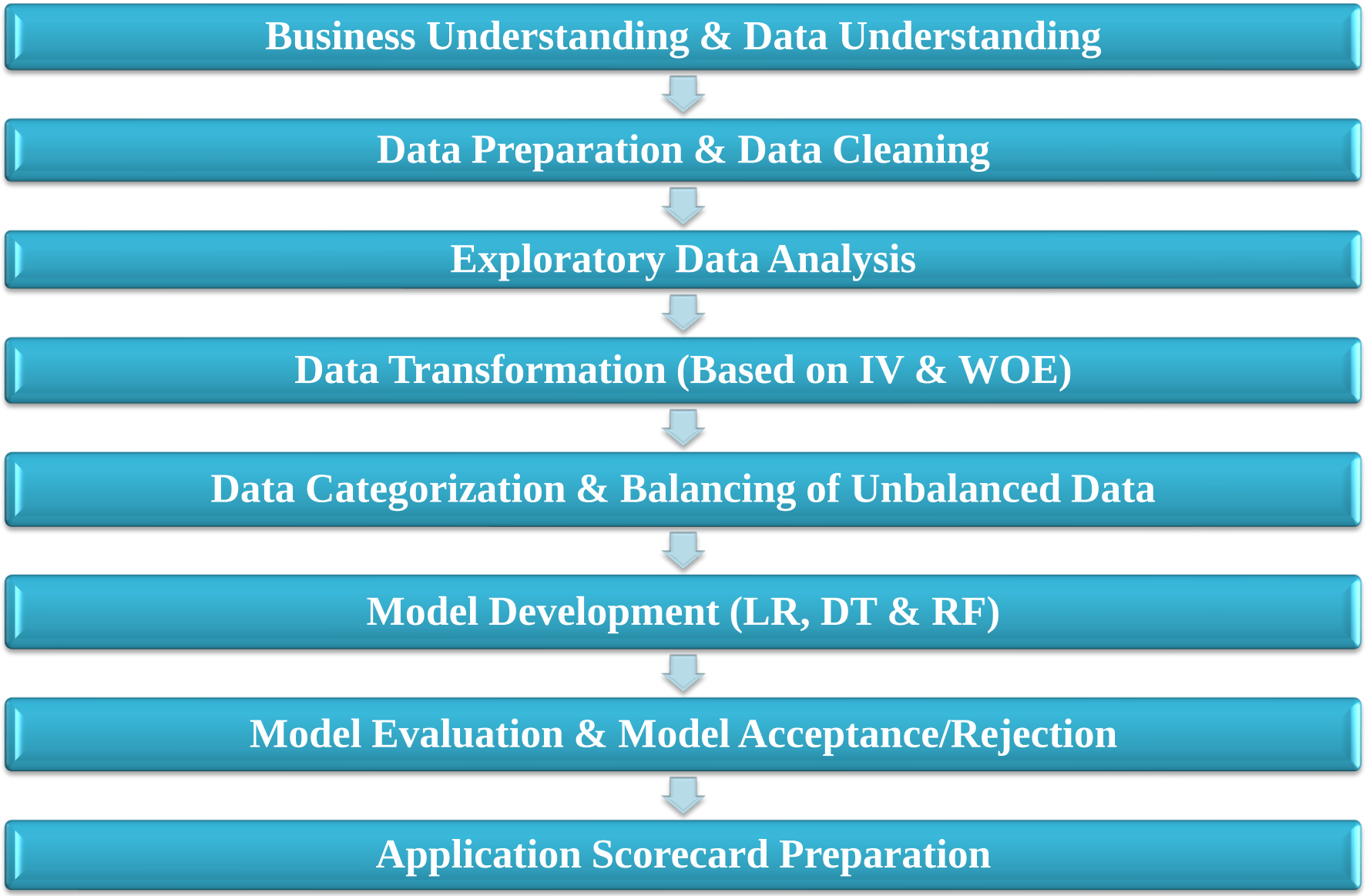
Business Objective

- The objective here is to identifying the right customers using predictive models to reduce credit loss. And the best strategy to mitigate credit risk is to ‘acquire the right customers’.

Proposed Solution

- We have developed a robust credit risk predictive model by analyzing bank’s past credit card applicants’ demographic and credit bureau data.
- Application score card has been derived by using the result of predictive model to find the credit worthiness of each applicant.

Model Building Approach



Data Preparation

Data Sets Used

- Demographic Data
- Credit Bureau Data
- Both the data sets are merged based on Application ID

Data Cleaning

- Duplicate Values, Incorrect Values are treated
- Outlier & missing value treatment was done
- Binning is done

Data Categorization

- Approved Population (Performance Tag value is Available)
- Rejected population (Performance Tag value is Missing)

Data Transformation

WOE Transformation

- The WOE stands for Weight of Evidence. It tells the predictive power of an independent variable in relation to the dependent variable i.e. Performance Tag (The default indicator) in our case.
- The master dataset is transformed with WOE converted dataset on which predictive model is built.

Balancing the Unbalanced Data

- We found that the approved population dataset obtained is unbalanced.
- The unbalanced data was converted into balanced data by using Synthetic Minority Oversampling Technique (SMOTE). So now we have both balanced and unbalance data set for model building.

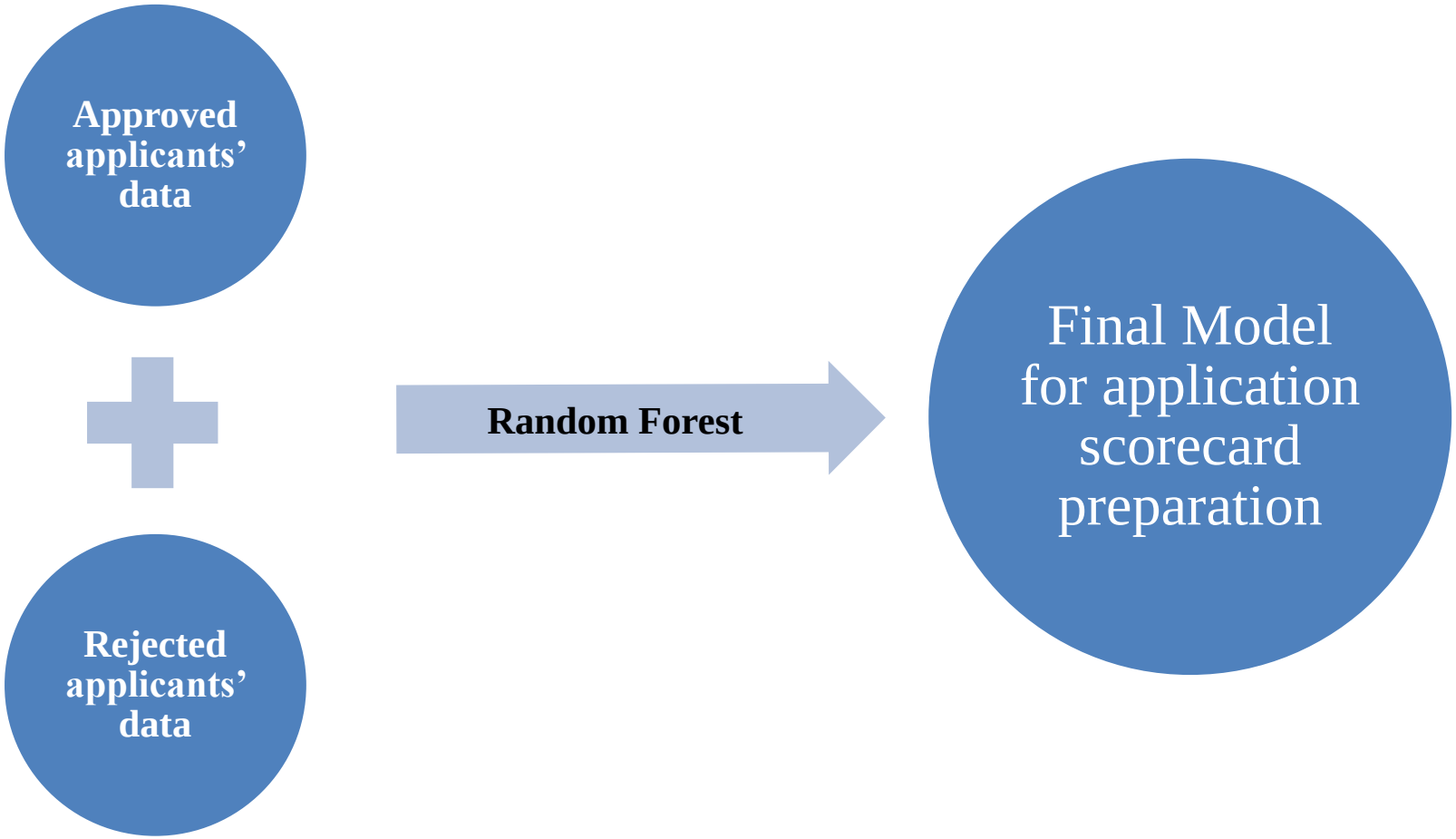
Model Building

The following models are built and best model (pink highlighted) is chosen based on accuracy obtained.

DATA	BALANCED/UNBALANCED	MODEL	METRICS
Demographic Data	Unbalanced	Logistic Regression	Accuracy : 0.555 Sensitivity : 0.60068 Specificity : 0.05589
Master Data	Unbalanced	Logistic Regression	Accuracy : 0.6294 Sensitivity : 0.59389 Specificity : 0.63097
Master Data	Unbalanced	Decision Tree	Accuracy : 0.9573 Sensitivity : 0.002262 Specificity : 0.9994
Master Data	Unbalanced	Random Forest	Accuracy : 0.5864916 Sensitivity : 0.6346154 Specificity : 0.5852686
Master+ Rejected Data	Unbalanced	Random Forest	Accuracy : 0.6852259 Sensitivity : 0.6733902 Specificity : 0.6859862 KS Statistics : 0.359
Master+ Rejected Data	Balanced	Random Forest	Accuracy : 0.660102 Sensitivity : 0.6857361 Specificity : 0.6584272

Model Selection

The best model chosen is built using both approved and rejected applicants' unbalanced data.



Model Selection

The following variables are used for final model selection

- **No of times 30 DPD or worse in last 12 months**
- **Avg CC Util 12 Months**
- **Outstanding Balance**
- **Total No Of Trades**

Model Evaluation

Evaluation Metric

- **Accuracy:** Accuracy, Specificity and Sensitivity are calculated
- **Discriminatory Power** : KS statistics is calculated
- **Stability** : Model is evaluated on the rejected population

Application Scorecard

- Application scorecard is calculated for all the trained data set.
- Cut-off score is figured out.

Financial Benefit

- Model is applied on both approved and rejected data
  - Rejected Population: To find out potential loss of revenue
  - Approved Population: To find out Credit loss avoided

Results & Analysis

Important Predictors

- Total 18 variables have been selected based on information value for model building.

- Top 8 variables have been listed below for which exploratory data analysis has been presented.

    1. Average Credit Card Utilization in last 12 Months
    2. No of trades opened in last 12 months
    3. Total No Of Trades
    4. No of PL trades opened in last 12 months
    5. No of times 90 DPD or worse in last 6 months
    6. Outstanding Balance
    7. No of times 30 DPD or worse in last 6 months
    8. No of times 60 DPD or worse in last 12 months

- All the above variables have **medium predictive power** except “Average Credit Card Utilization in last 12 Months” which have **strong predictive powers** for model development.

Information Value

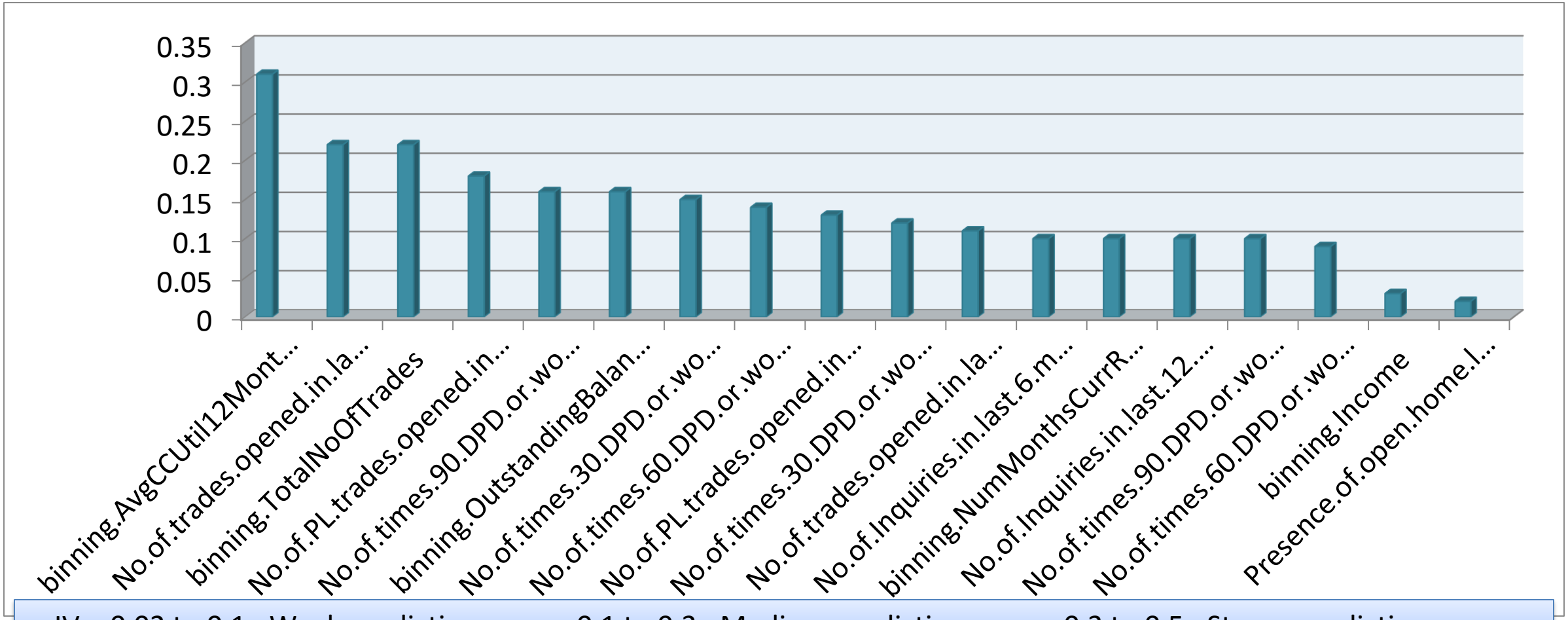
▪ Information Value (IV) is one of the most useful technique to select important variables in predictive modeling. It helps to rank variables on the basis of their predictive power.

Sl	Variables	IV	Sl	Variables	IV
1	Average Credit Card Utilization in last 12 Months	0.31	10	No of times 30 DPD or worse in last 12 months	0.12
2	No of trades opened in last 12 months	0.22	11	No of trades opened in last 6 months	0.11
3	Total No Of Trades	0.22	12	No of Inq in last 6 mnth excl home auto loans	0.1
4	No of PL trades opened in last 12 months	0.18	13	Num of Months in Current Residence	0.1
5	No of times 90 DPD or worse in last 6 months	0.16	14	No of Inq in last 12 mnth excl home auto loans	0.1
6	Outstanding Balance	0.16	15	No of times 90 DPD or worse in last 12 months	0.1
7	No of times 30 DPD or worse in last 6 months	0.15	16	No of times 60 DPD or worse in last 6 months	0.09
8	No of times 60 DPD or worse in last 12 months	0.14	17	Income	0.03
9	No of PL trades opened in last 6 months	0.13	18	Presence of open home loan	0.02

IV = 0.02 to 0.1 : Weak predictive power, 0.1 to 0.3 : Medium predictive power, 0.3 to 0.5 : Strong predictive power

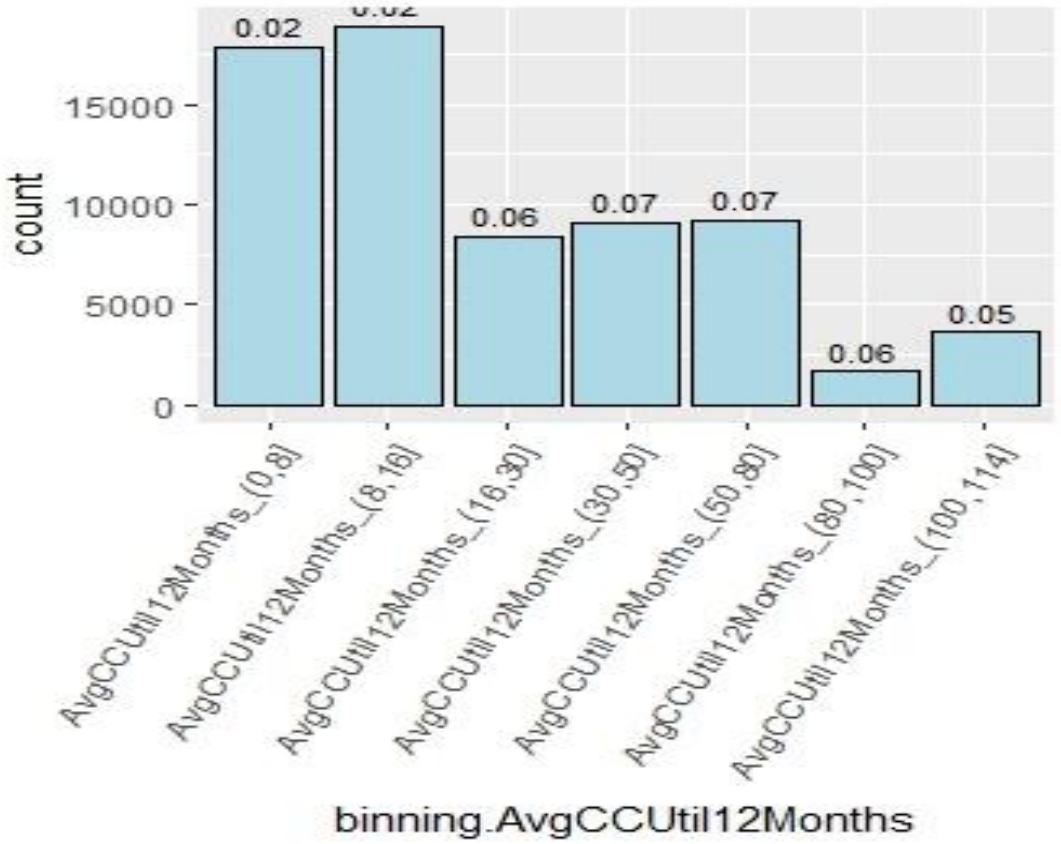
Information Value

For model building we have considered only those variables whose information value is greater than or equal to 0.02.

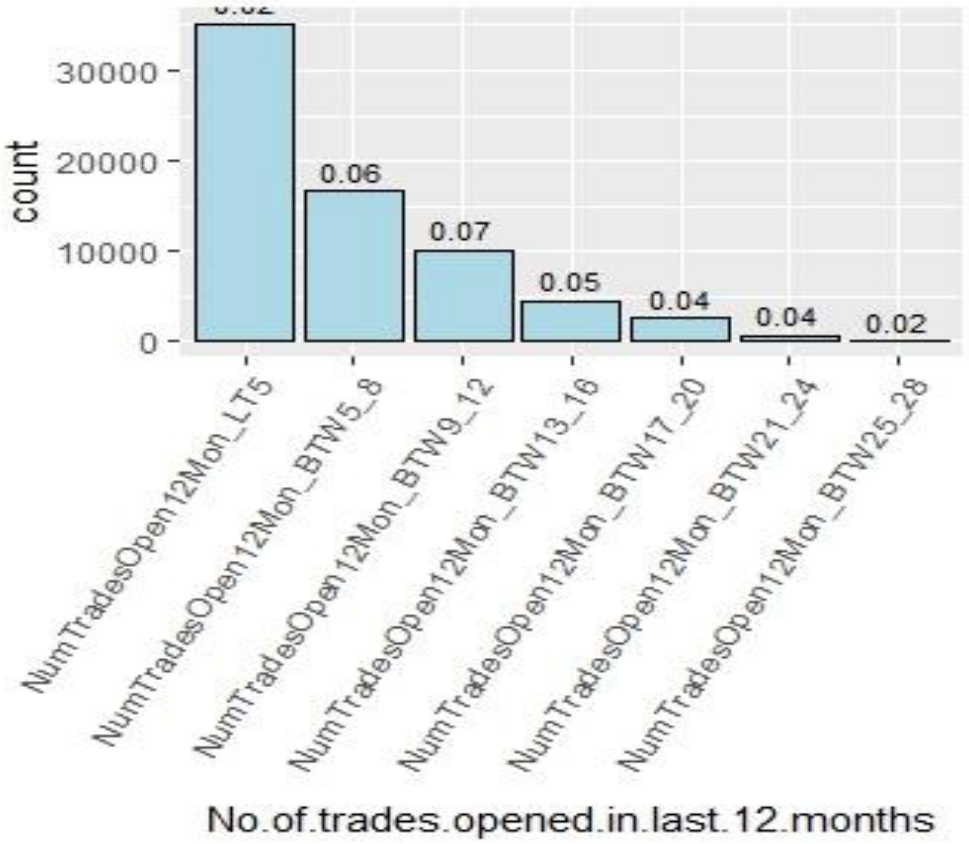


IV = 0.02 to 0.1 : Weak predictive power, 0.1 to 0.3 : Medium predictive power, 0.3 to 0.5 : Strong predictive power

Exploratory Data Analysis

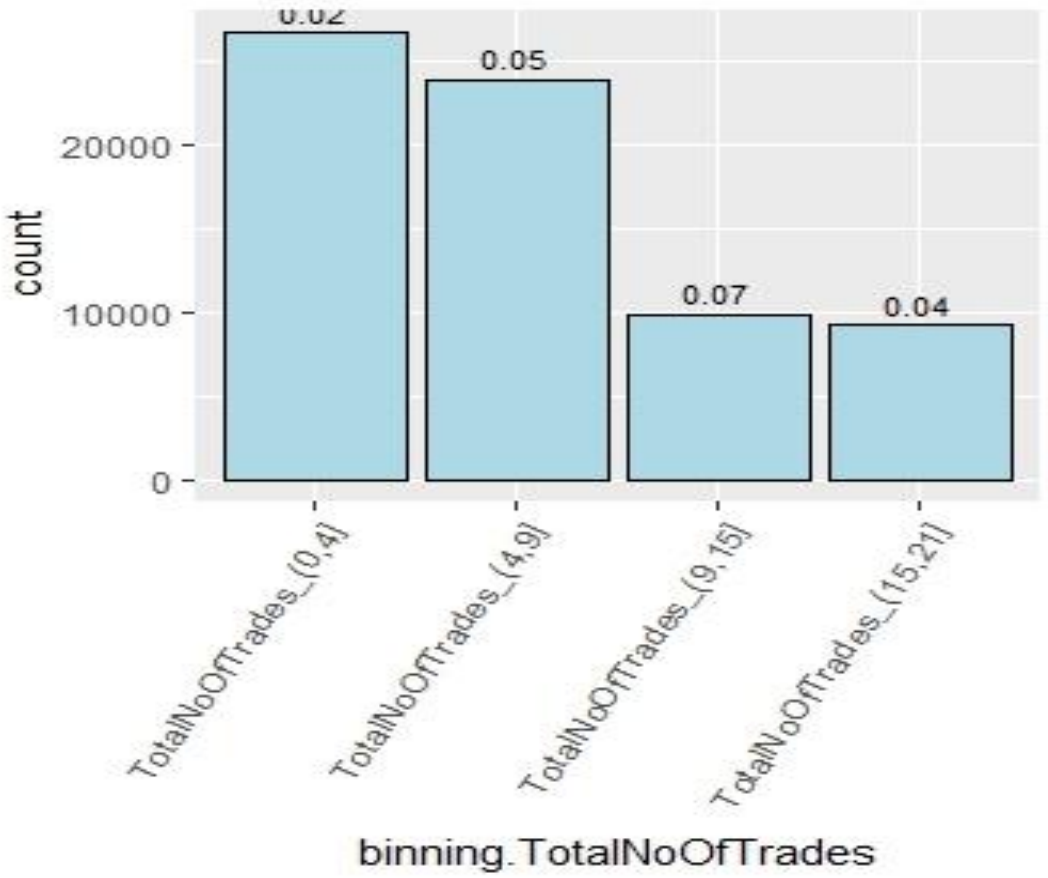


Default rate is comparatively more for credit card utilization [30,50] & [50,80]

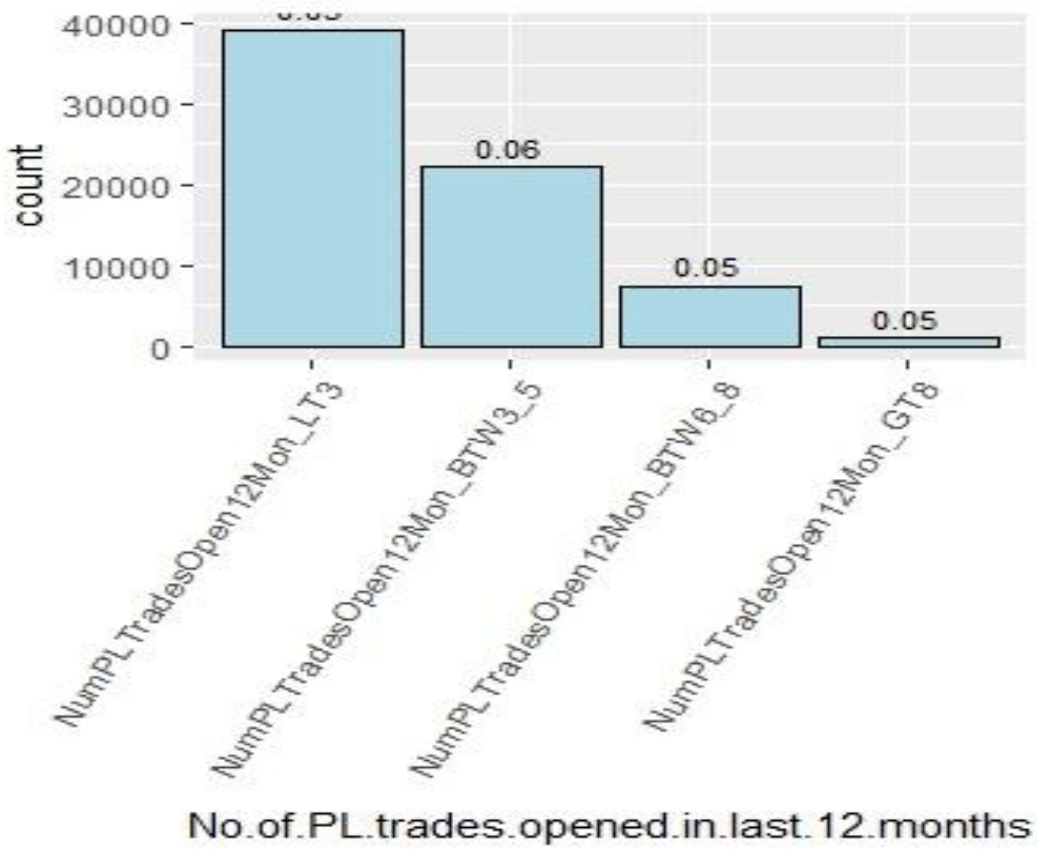


Default rate is comparatively more No of trades opened in last 12 months [5,8] & [9,12]

Exploratory Data Analysis

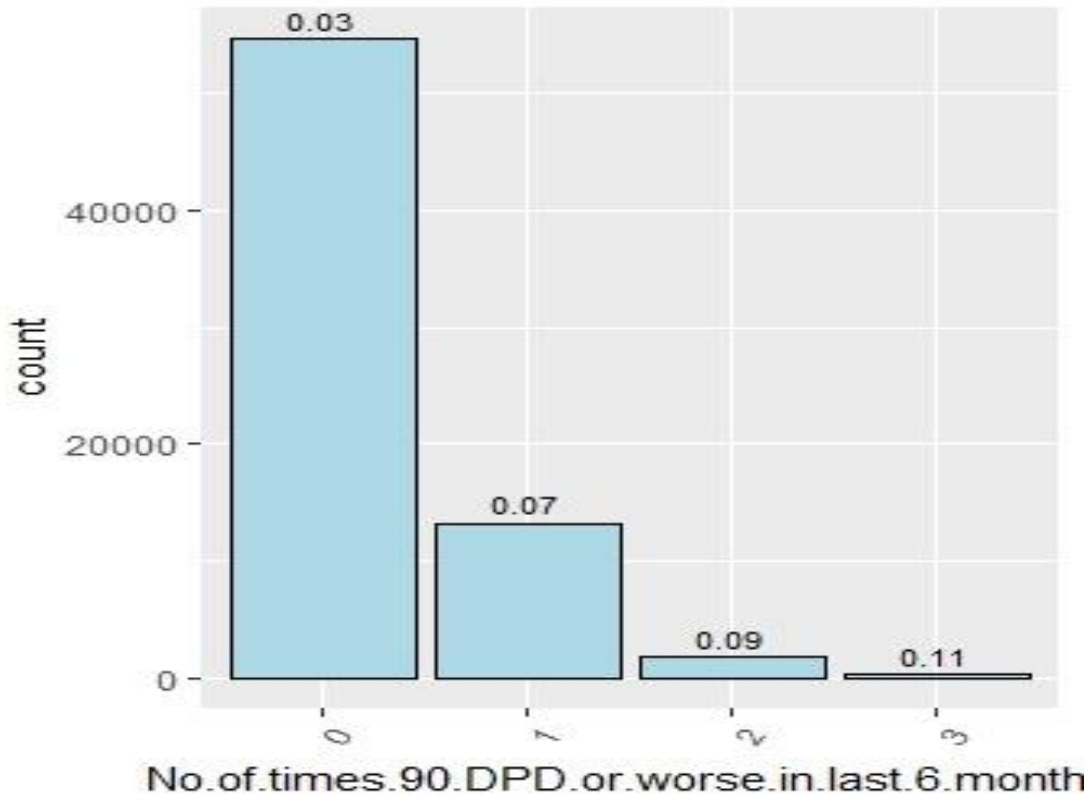


Default rate is comparatively more for total number of trades [9 , 15]

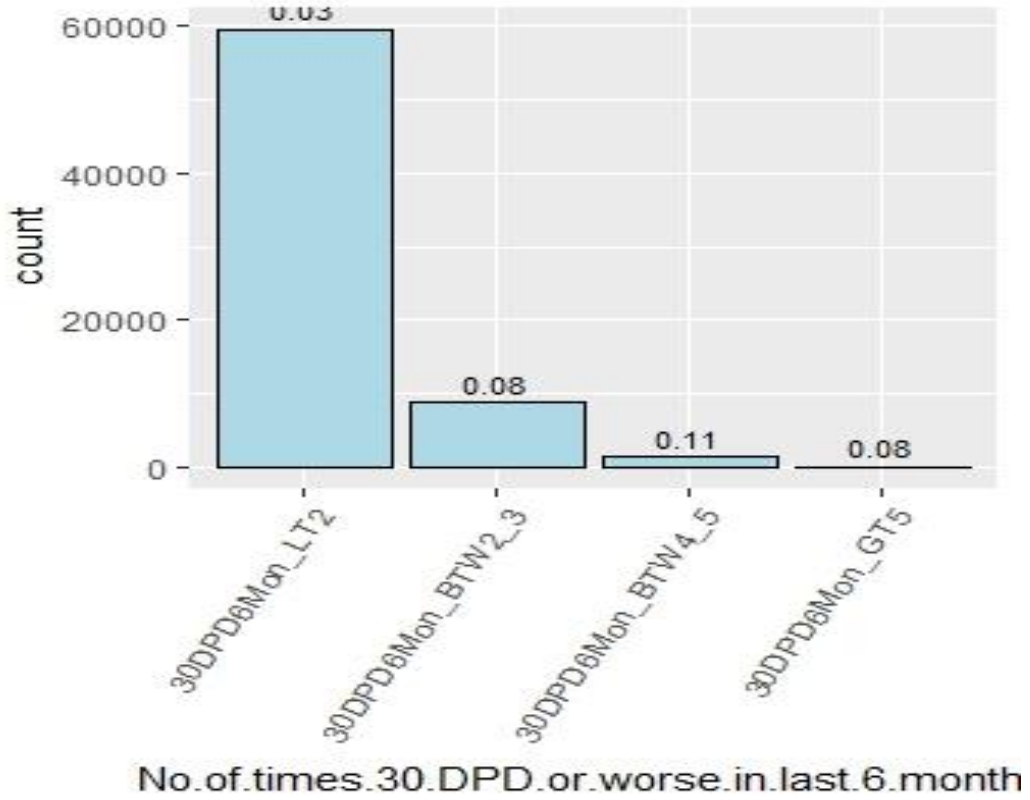


Default rate is comparatively more for no of pl trades opened in last 12 months [3, 5]

Exploratory Data Analysis

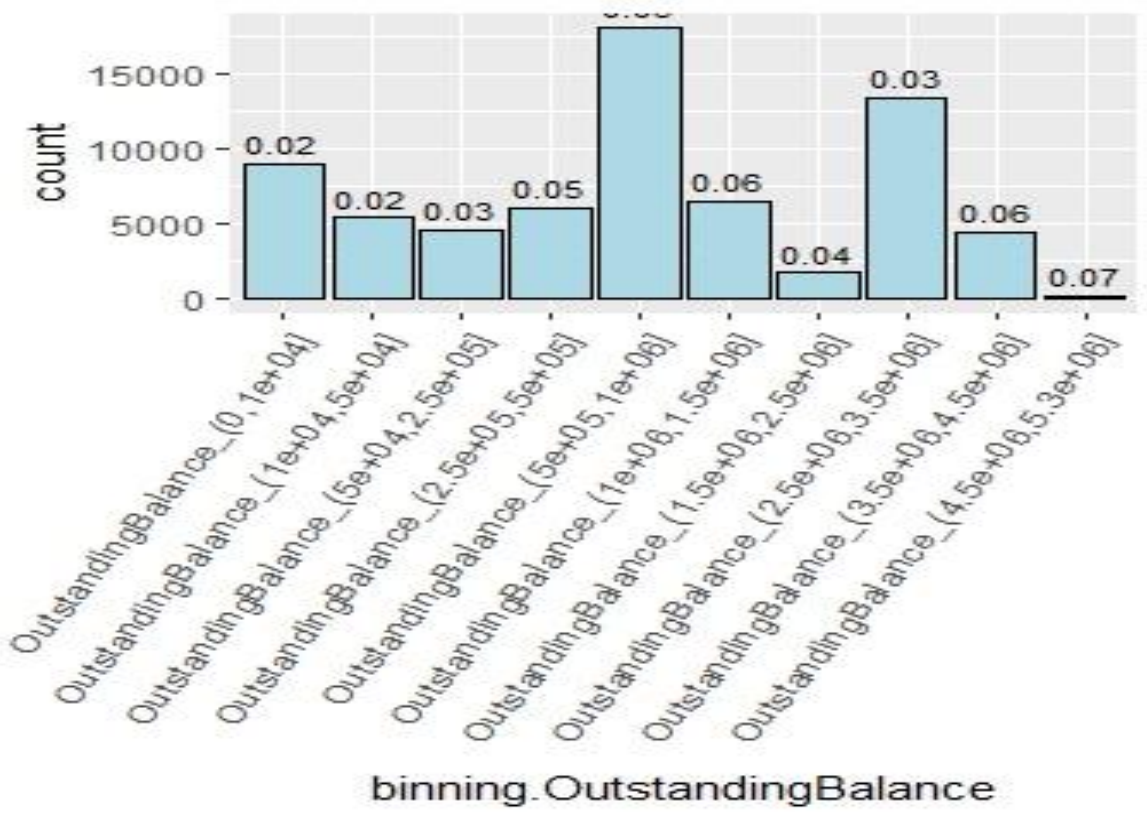


Default rate is comparatively more for those applicants who have 3 times 90 DPD in last 6 months

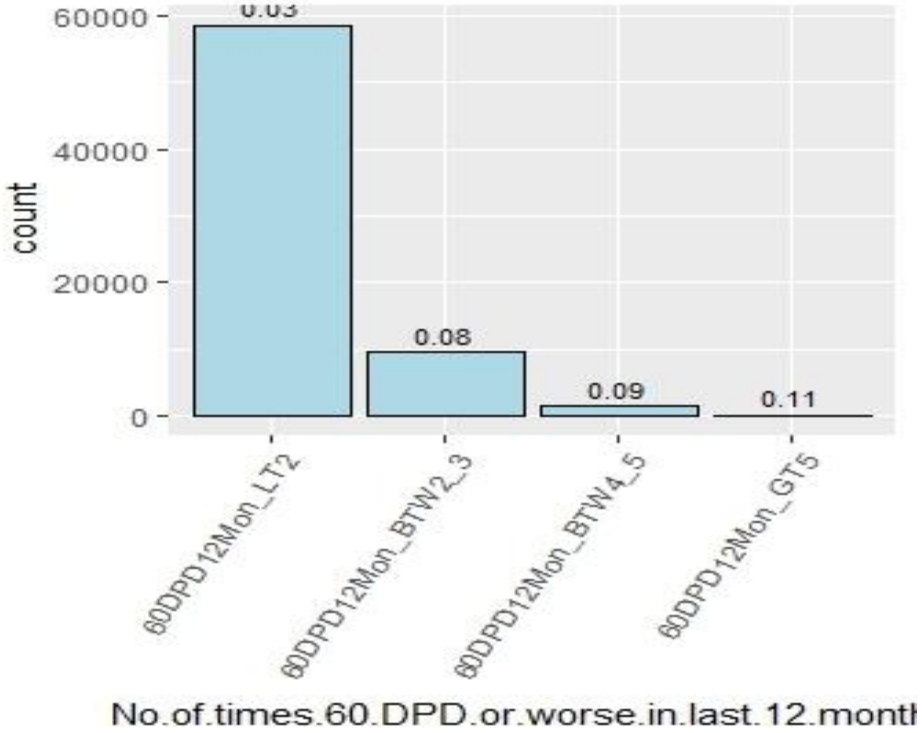


Default rate is comparatively more for those applicants who have 4 to 5 times 30 DPD in last 6 months

Exploratory Data Analysis



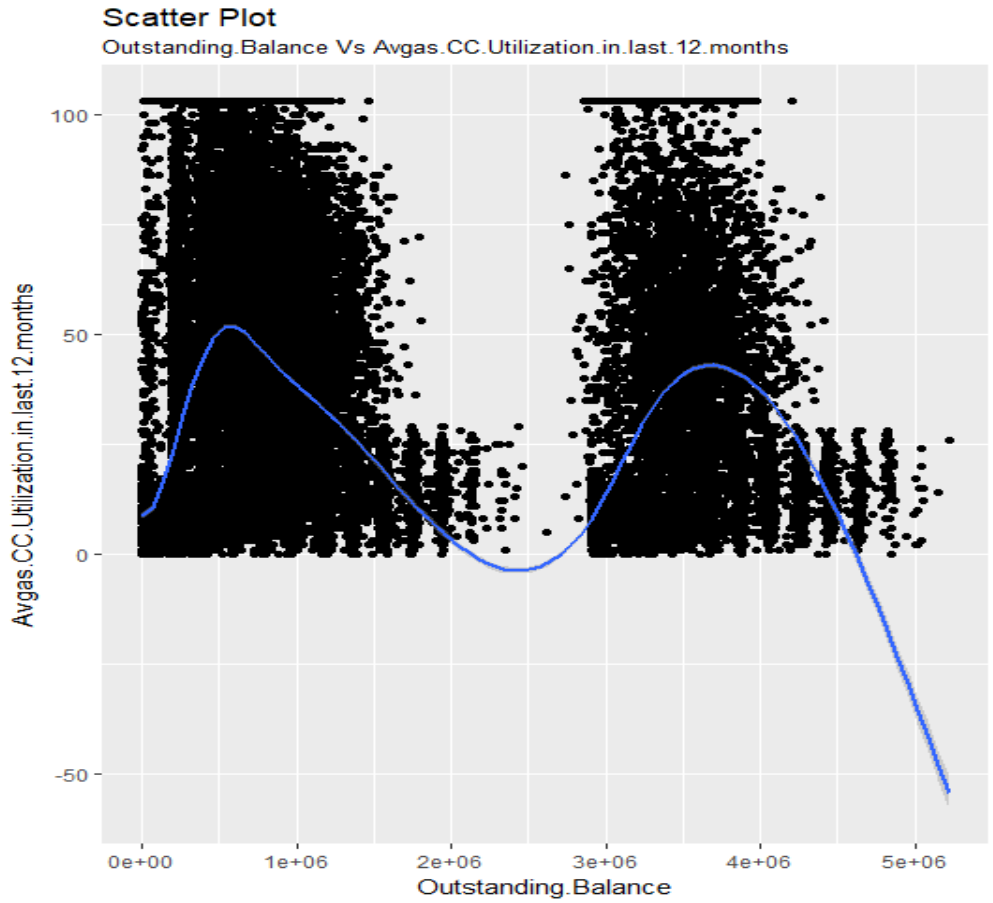
Default rate is comparatively more for those applicants who have maximum outstanding balance [4500000,5300000]



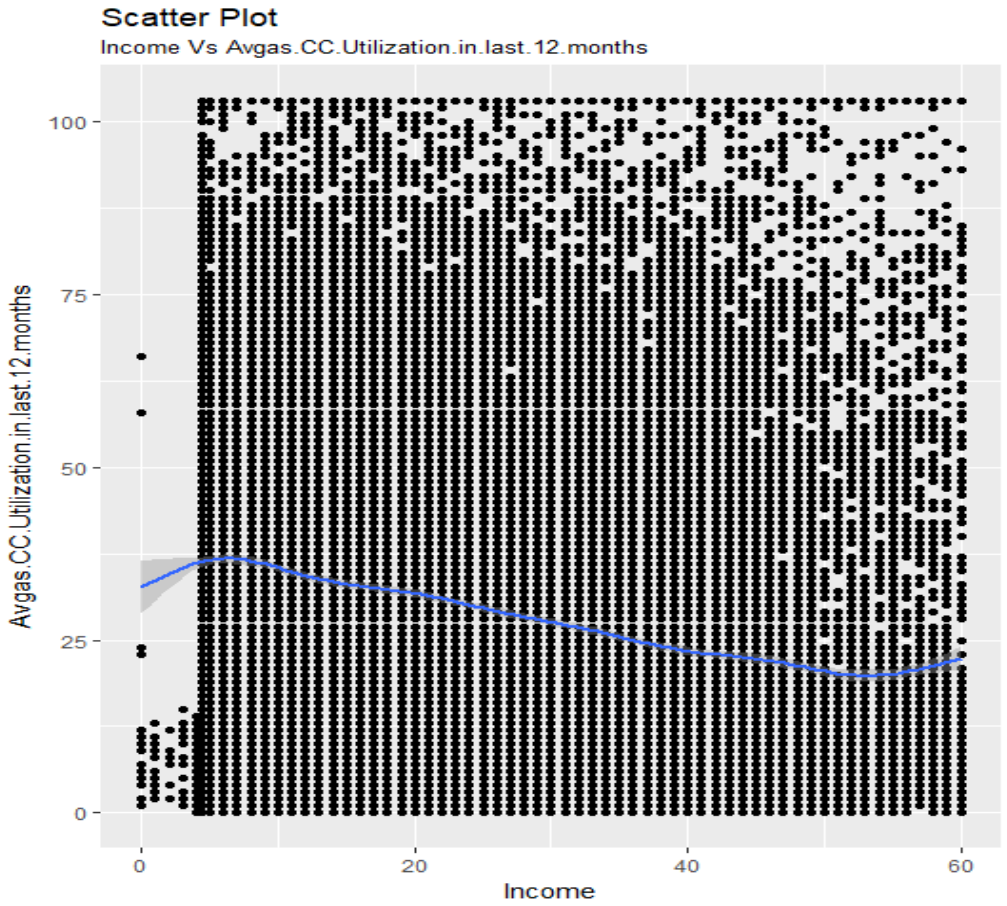
Default rate is comparatively more for those applicants who have 5 or more than 5 times 60 DPD in last 12 months

Exploratory Data Analysis

Multivariate Analysis



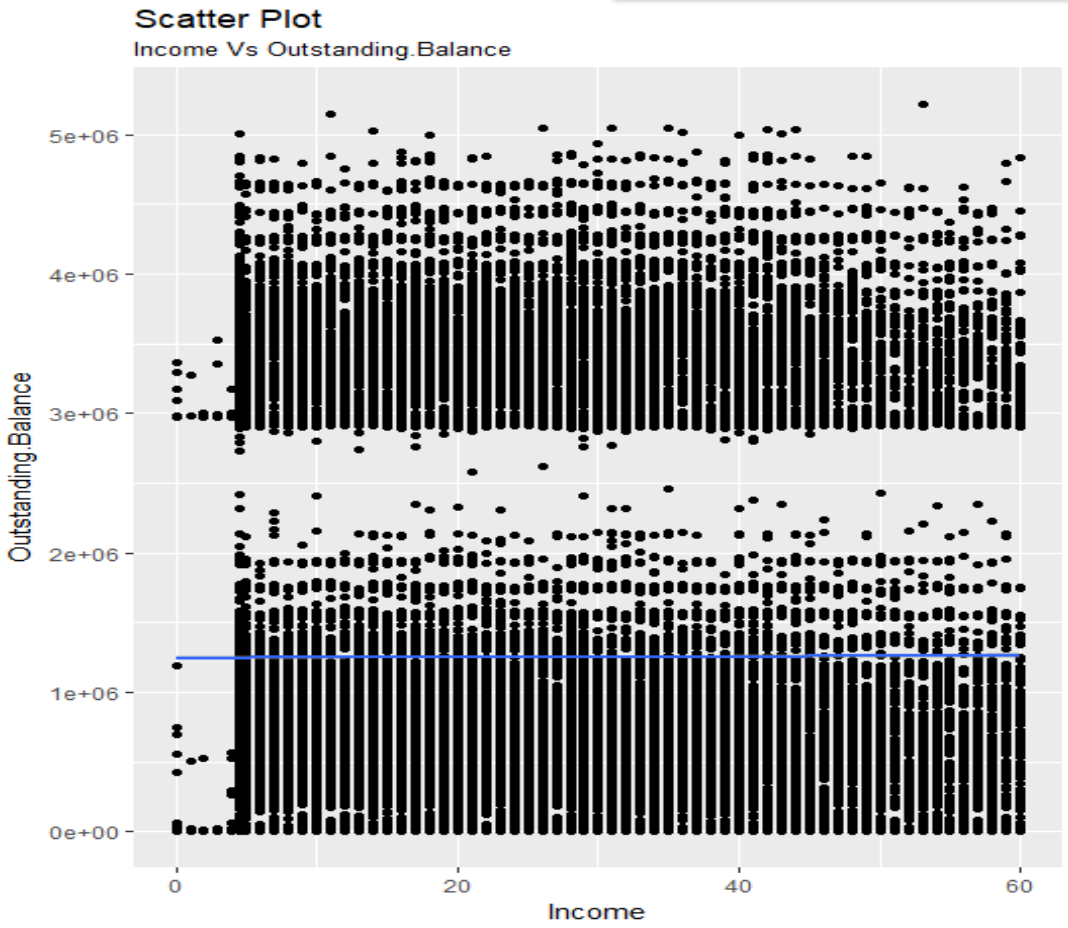
Their is a decreasing trend of Avgas.CC.Utilization.in.last.12.months when Outstanding Balance increases.



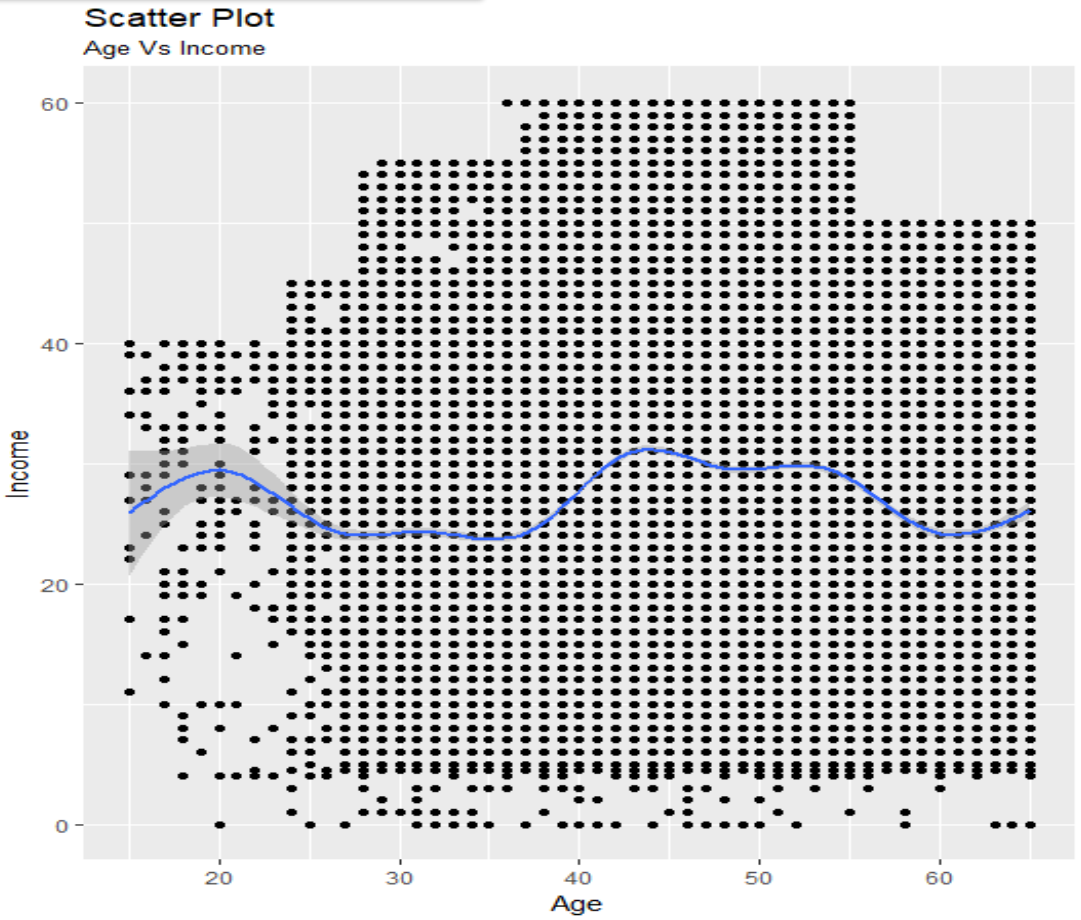
Avg CC utilization slightly decreases with increase in income. It has linear relationship with -ve coefficient.

Exploratory Data Analysis

Multivariate Analysis

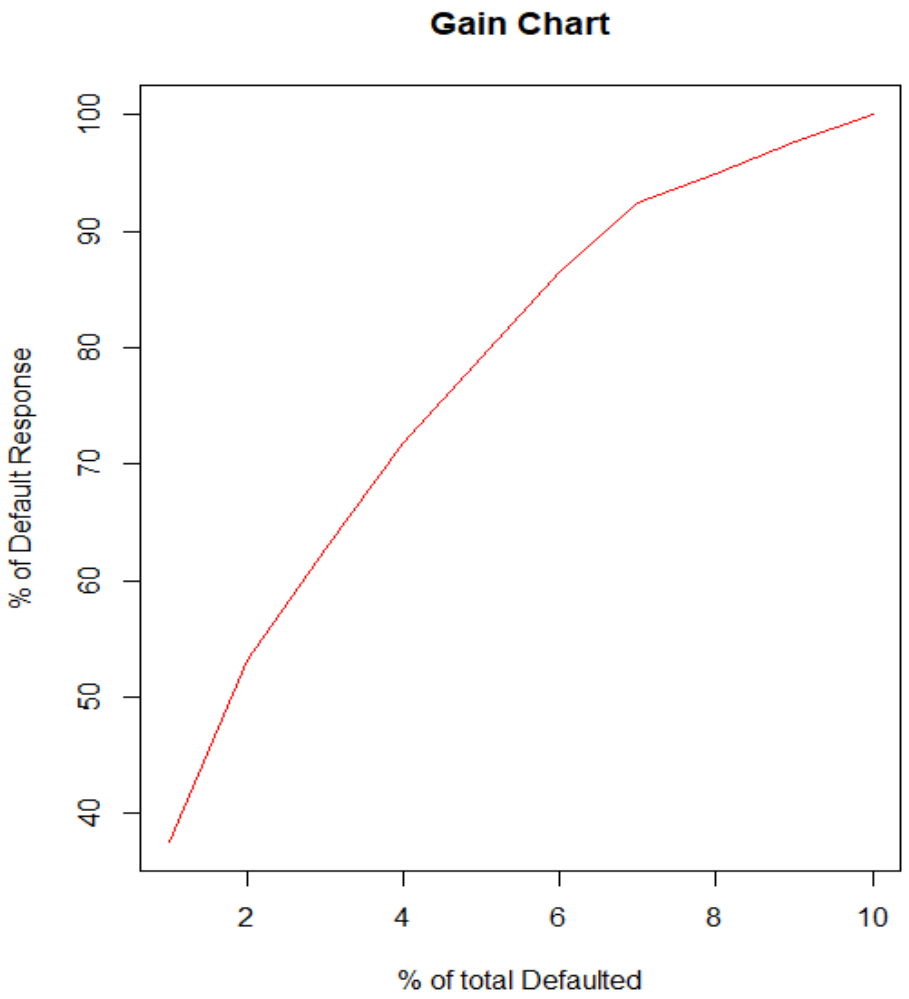


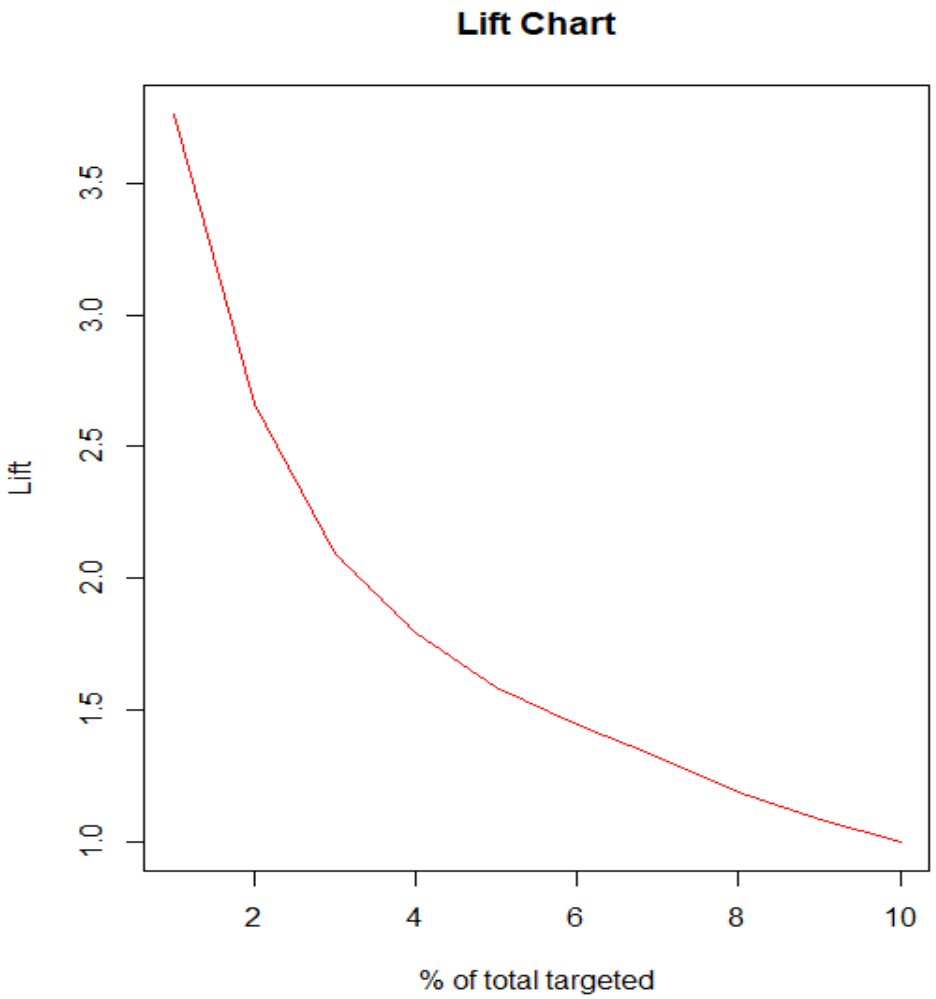
Income and Outstanding Balance have a perfect linear relationship.



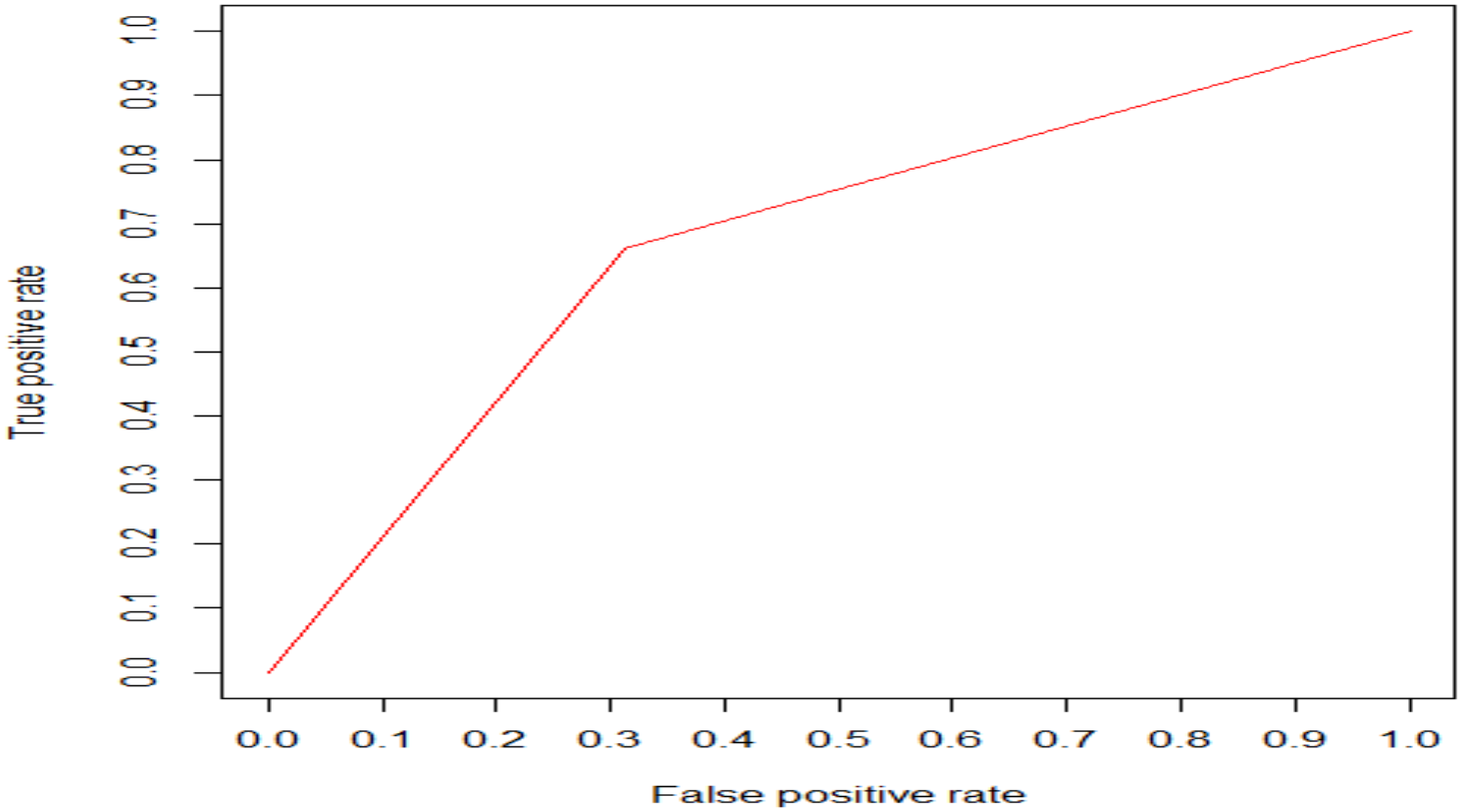
Age and Income does not have any increasing or decreasing trend .Its almost varies up and down with in a small range.

Gain / Lift Charts

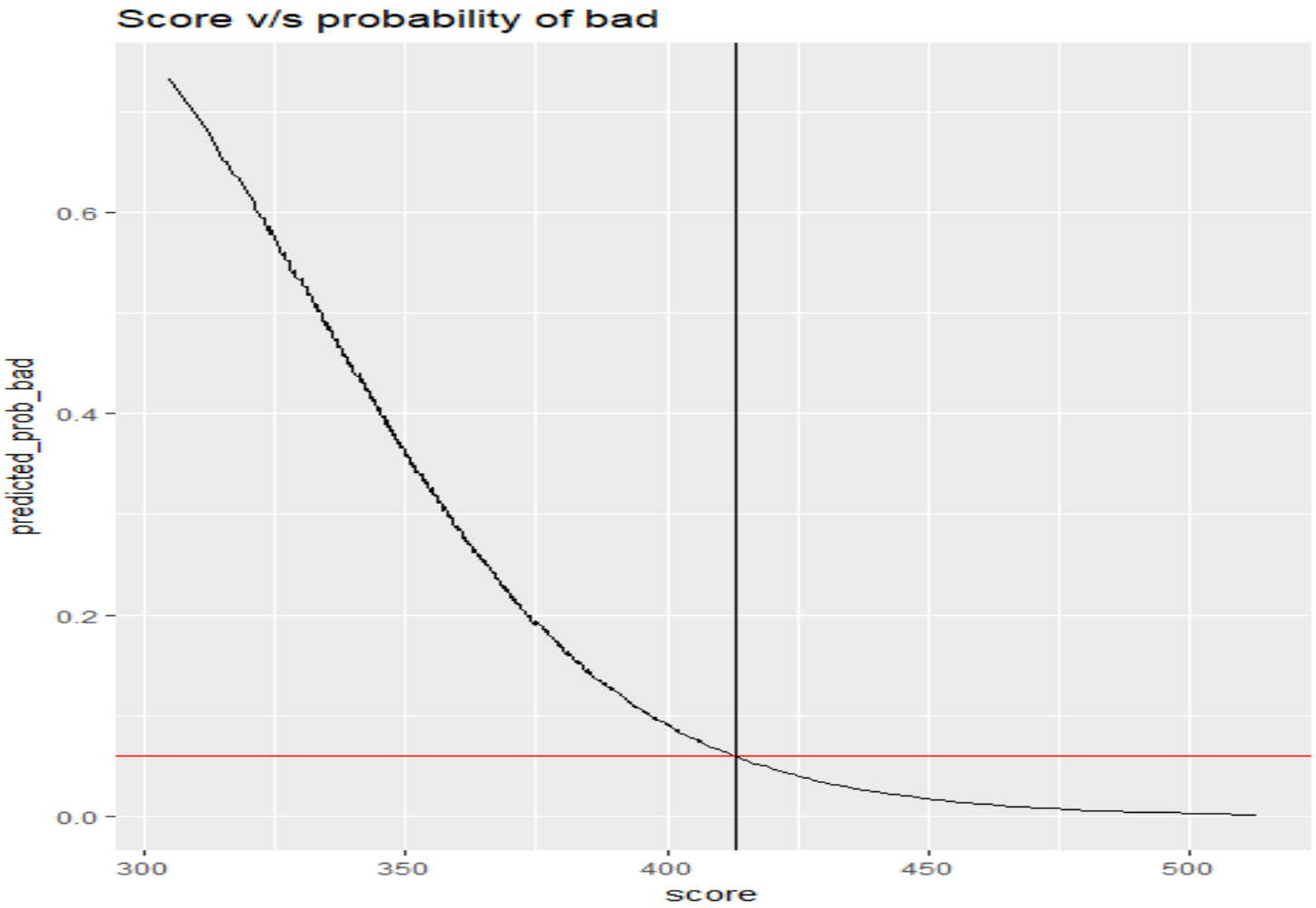




KS Statistics



Score Card Plot



Score Card Calculation

- The score of each the applicant is calculated based on the below formula.

Score = (400- (20/ln2)) + (20/ln2) * ln(Odds)
where Odd = P/(1-P) and P = Probability of default

- The optimal score is calculated to be **413.**

- Applicants having score more than 413 are good customers whose predicted output is “NO” and those having score less than 413 are bad customers whose predicted output is “YES”.

- The score is calculated for both the rejected and approved population and compared against the optimum score of 413 to predict they are good or bad customer.

- Min Application Score 319
  Max Application Score 513

Financial Benefit Analysis

EVALUATION OF MODEL ON REJECTED APPLICANTS

- The final model is applied on the rejected applicants.

NO	YES
72	1352

- So out of total 1424 rejected applicants, 1352 applicants for whom our model predict yes, .the bank has done a right thing by not giving them credit card.

- As per our model, 1352 i.e. 95% of applicants would have defaulted. Hence we have avoided annual credit loss by denying them credit card.

- The rest 72 i.e. 5% of applicants, would not have defaulted had they given credit card facility. Hence bank has incurred a potential **loss of revenue**.

Financial Benefit Analysis

- Total number of rejected customers by the bank without using model : 1424

As per our model ,

Predicted non default customer : 1352
Predicted default customer : 72

- Potential revenue loss by declining credit card to 72 predicted good applicants = 12,608,520 which is calculated by summing the total outstanding balance.

Financial Benefit Analysis

EVALUATION OF MODEL ON APPROVED APPLICANTS

- Our model is applied on the approved population and the records are sorted on decreasing order of probability.
- It is observed that top 30% of applicants gives us 63% of total applicants who defaulted as per prediction.
- The credit loss for the top 30 % applicant is 825,7714,779 as calculated using outstanding balance.
- They are those applicants who have actually defaulted and our model also predicted correctly.
- Hence though our model we could have avoided the above credit risk.

Thank You!