# BFSI Project - Acquisition Analytics

# Introduction

## Problem Statement

CredX, leading credit card provider, has experienced increased credit loss in recent years. To mitigate this risk, CredX  wants to acquire the right customers.

## Key Objectives

- Help CredX identify the right customers using appropriate predictive models. This involves
    - Using historical data to determine factors affecting credit risk
    - Creating strategies to mitigate acquisition risk
    - Assessing financial benefits of the project

## Data Description

- Demographic Data : 71295 observations of 12 variables
    - Contains data on customers age, income , gender, marital status etc.
- Credit Data : 71295 observations of 19 variables
    - Data obtained from Credit Bureau, contains information on loans, outstanding balance, trades, DPD, etc.

## Assumptions

- There are cases where all the variables in the credit bureau data are zero and credit card utilization is missing-this is missing data from Credit Bureau
- Cases wherein only credit card utilization is missing are customers without credit cards
- The dependent variable "Performance Tag" is missing in a few cases(1425) -these are treated as applicants who have been rejected by CredX. We are treating this as rejected data and will use this for testing the cut-off for credit score.

# Approach

**Data Preparation**

- Checking for duplicates in Application ID
- Checking that Application IDs across datasets are same
- Checking for missing values-
  - Replaced missing values with median
- Outlier detection and treatment
- Creating appropriate derived variables
- Formatting and creating dummies for categorical data and scaling numerical data

**EDA**

- Univariate, bivariate, multivariate and correlation analysis of variables to determine which factors are likely to have more influence on credit default
- WOE and Information Value Analysis to determine predictive value of variable

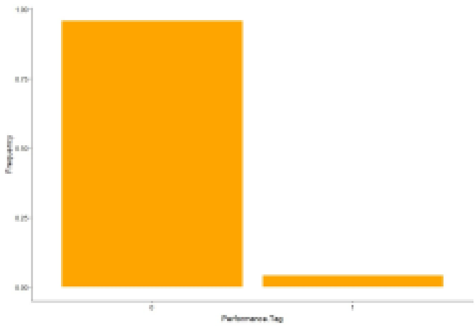
**Model building, selection and testing**

- Divide data into Training and Test datasets in 70:30 ratio
- Since data is unbalanced, oversampling is done using SMOTE(Synthetic Minority Oversampling Technique) using ROSE package
- Iterative model building on training dataset using
  - Logistic regression
  - Decision Tree
  - Random Forest
  - SVM
  - GBM
- Model selection based on specific parameters
- Using model to Predict test data
- Evaluating Model accuracy, sensitivity and specificity
- Plotting the AUC - ROC curve
- Evaluating KS statistic
- Plotting Gain and Lift Charts

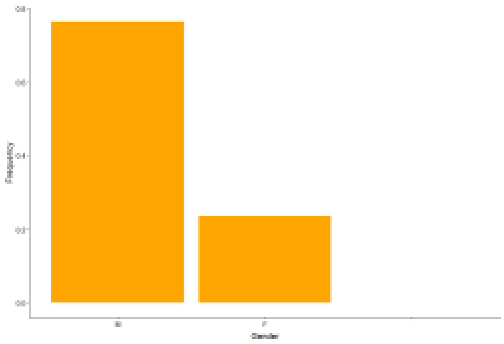
**Application Scorecard and Financial Benefit**

- Chosen model is used to build an Application Scorecard
- A cut-off is chosen below which applicants will not be granted credit card
- The financial benefit of the project is assessed in terms of the credit loss minimized as well as in terms of the revenue maximized by acquiring right customers , vis-à-vis a no model approach

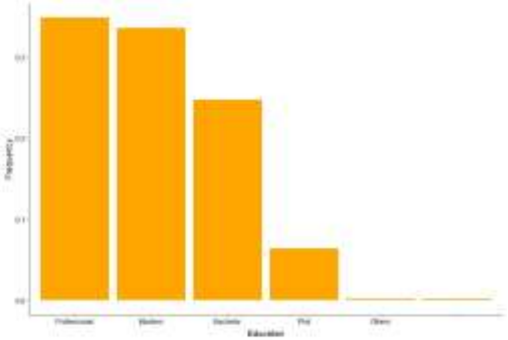
# EDA –Overall Approved Applicant Characteristics



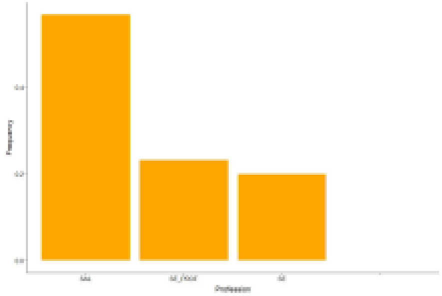
Distribution by default status- only 4.2% of the applicants are defaulters



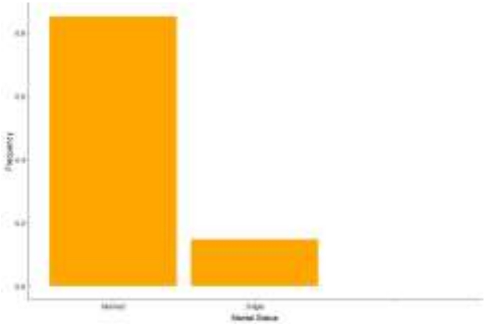
Distribution by gender-4 times more male applicants than female applicants



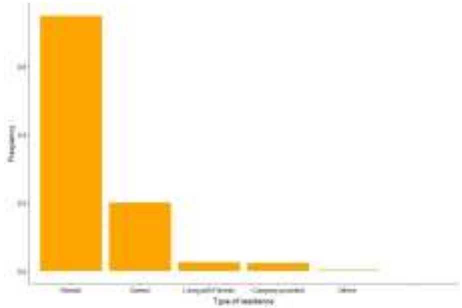
Distribution by education-Professional and masters degree holders are the largest in number



Distribution by job type-3 times more salaried professionals than other job holders



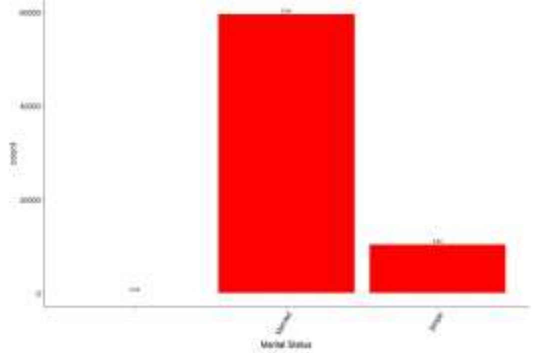
Distribution by marital status-5 times more married than single applicants



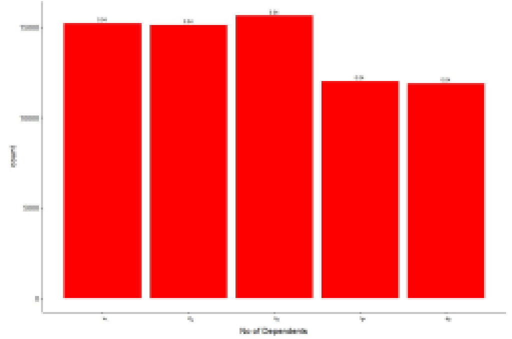
Distribution by residence type-applicants living in rented housing are 5 times more than those living with in own houses

- Overall, most approved applicants are male, with professional or masters degree, in a salaried profession, married and living in rented housing
- Also, most of them are non defaulters, indicating that basic approval guidelines are valid

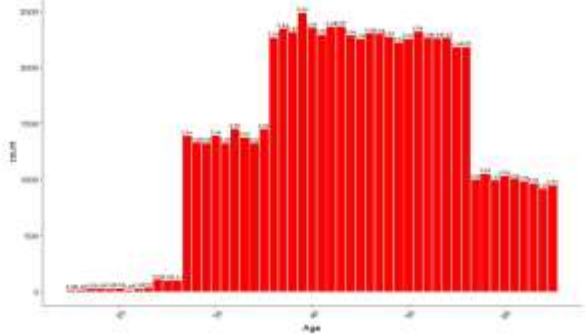
# EDA –Default % across categories



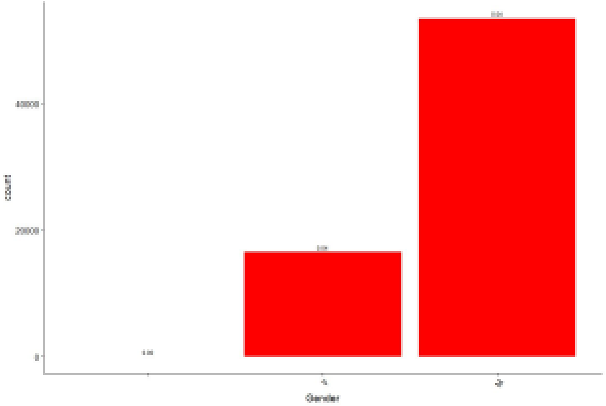
Defaulters % by Marital Status



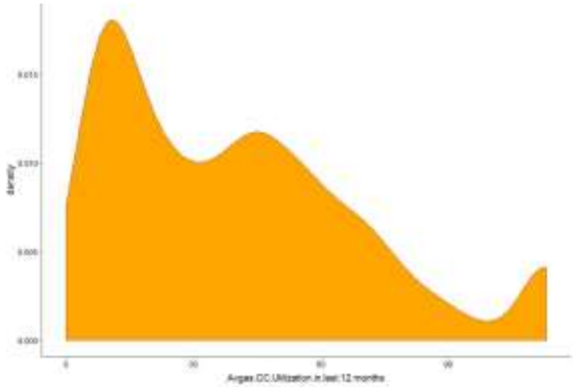
Defaulters % by Number of Dependents



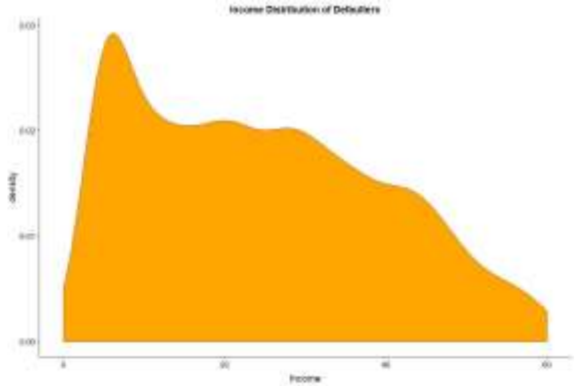
Defaulters % by Age



Defaulters % by Number of Dependents



Average Credit card Utilization by Defaulters

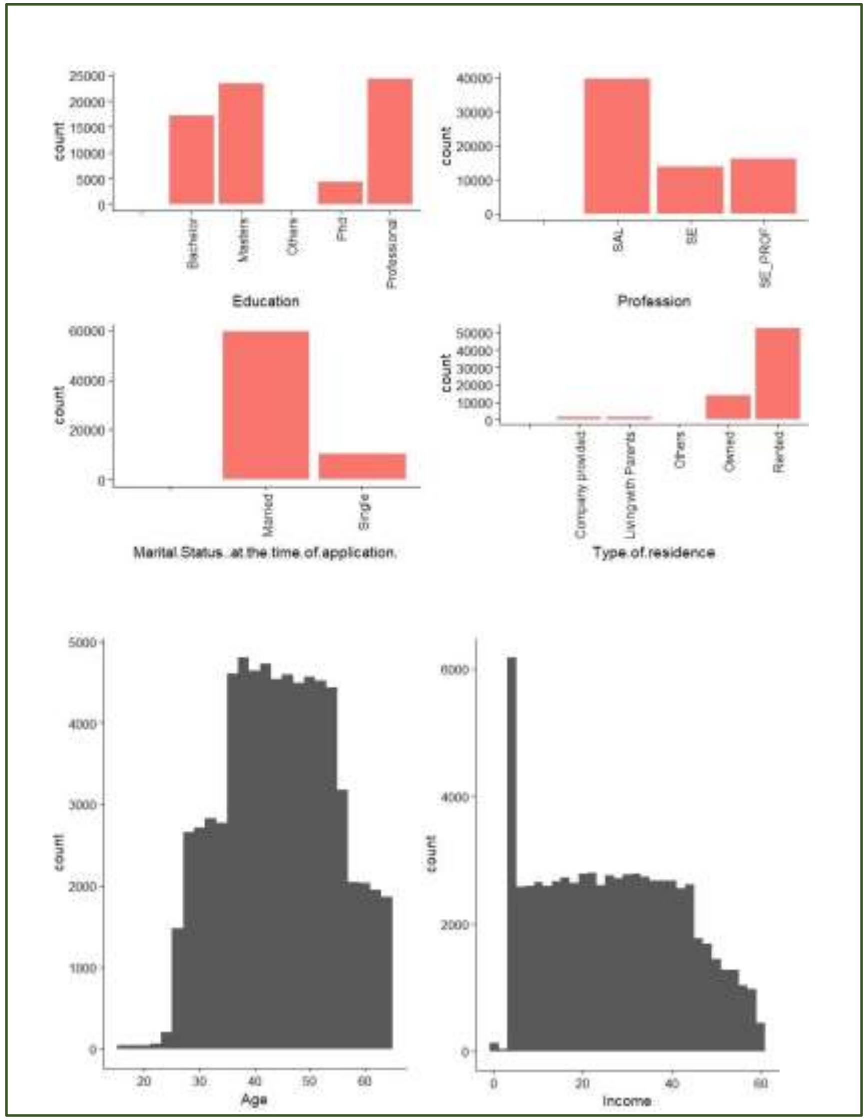


Income density plot of defaulters

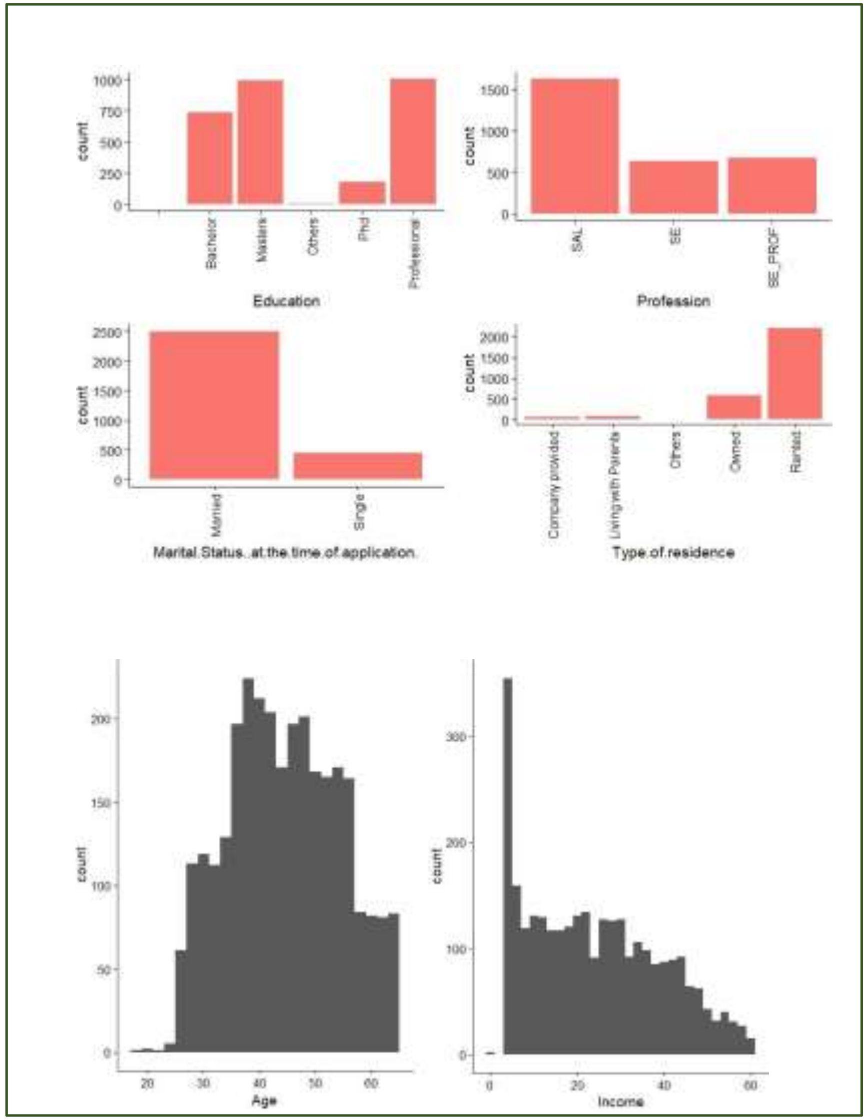
- Percentage of defaulters is approximately the same across gender, marital status and number of dependents
- Though it is slightly higher i.e. 5% for some ages, however there is no visible pattern
- Both Average credit card utilization and Income of defaulters is right skewed

# Factors affecting Credit Risk-Demographic Data

All



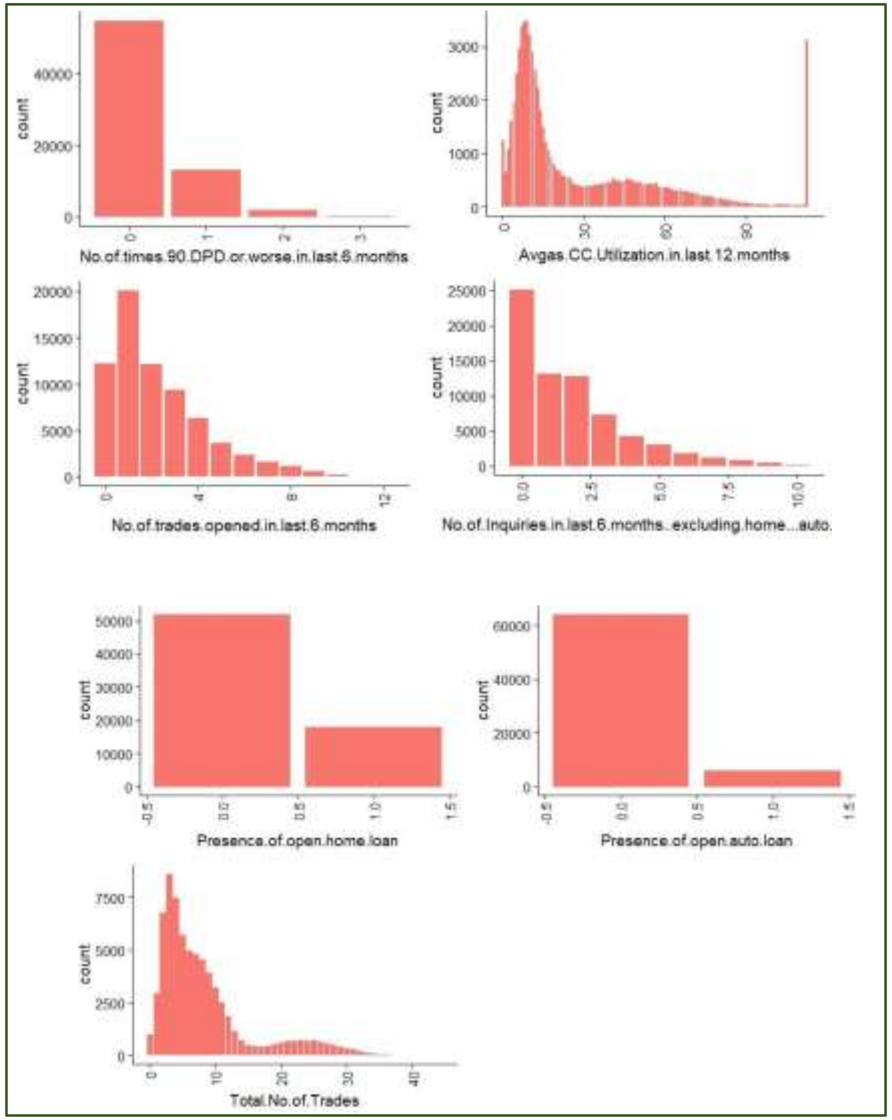

Defaulters



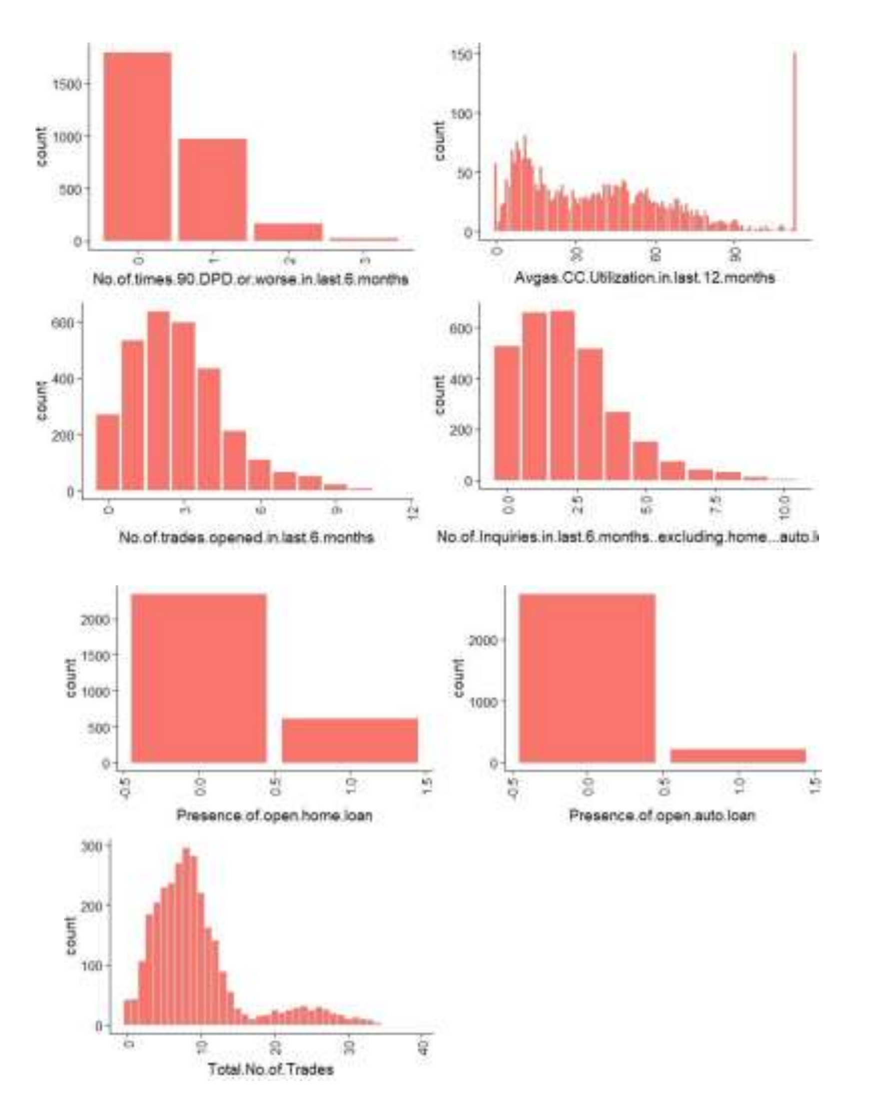

- Except for marginal differences in age and income distribution among defaulters vs the entire data set of applicants, other demographic variables show no difference in pattern, indicating that they may be poor predictors of credit risk

# Factors affecting Credit Risk- Credit History

**All**

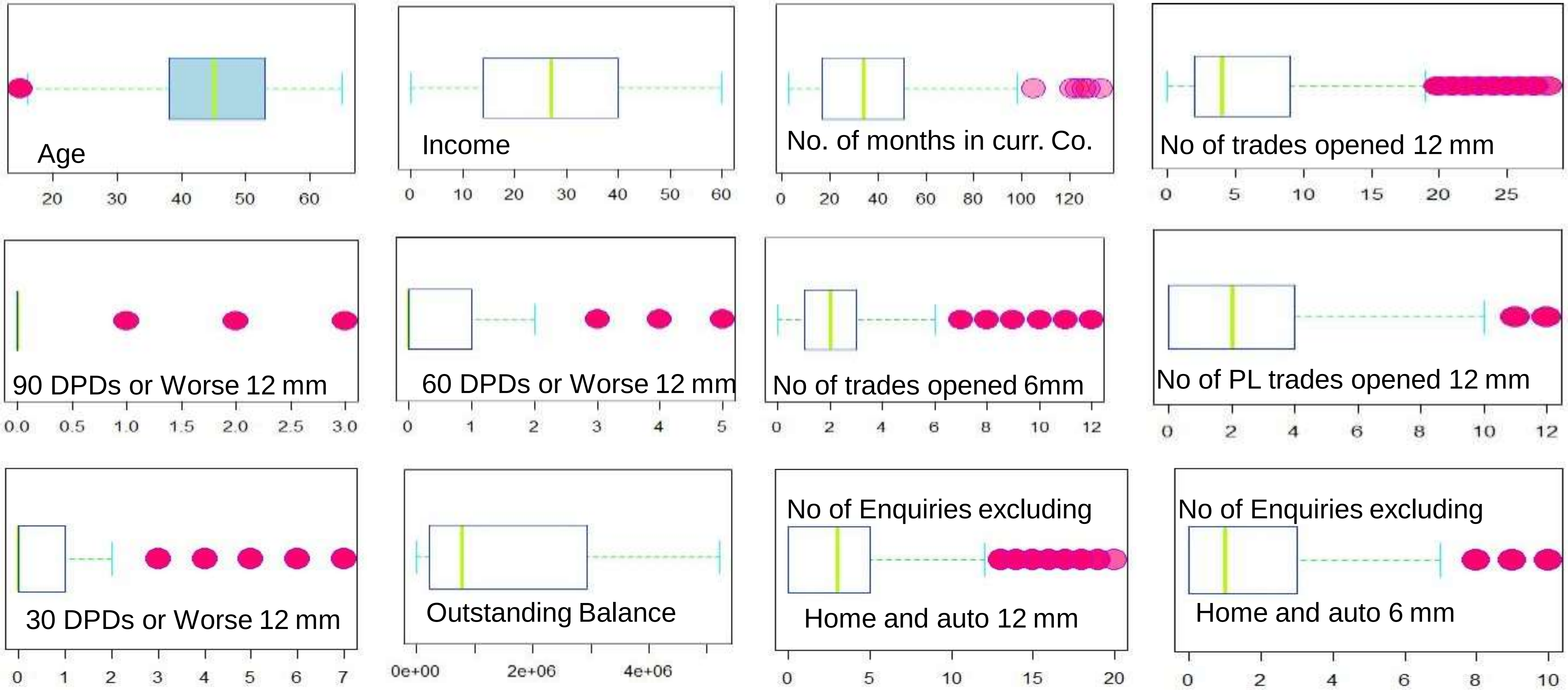



**Defaulters**



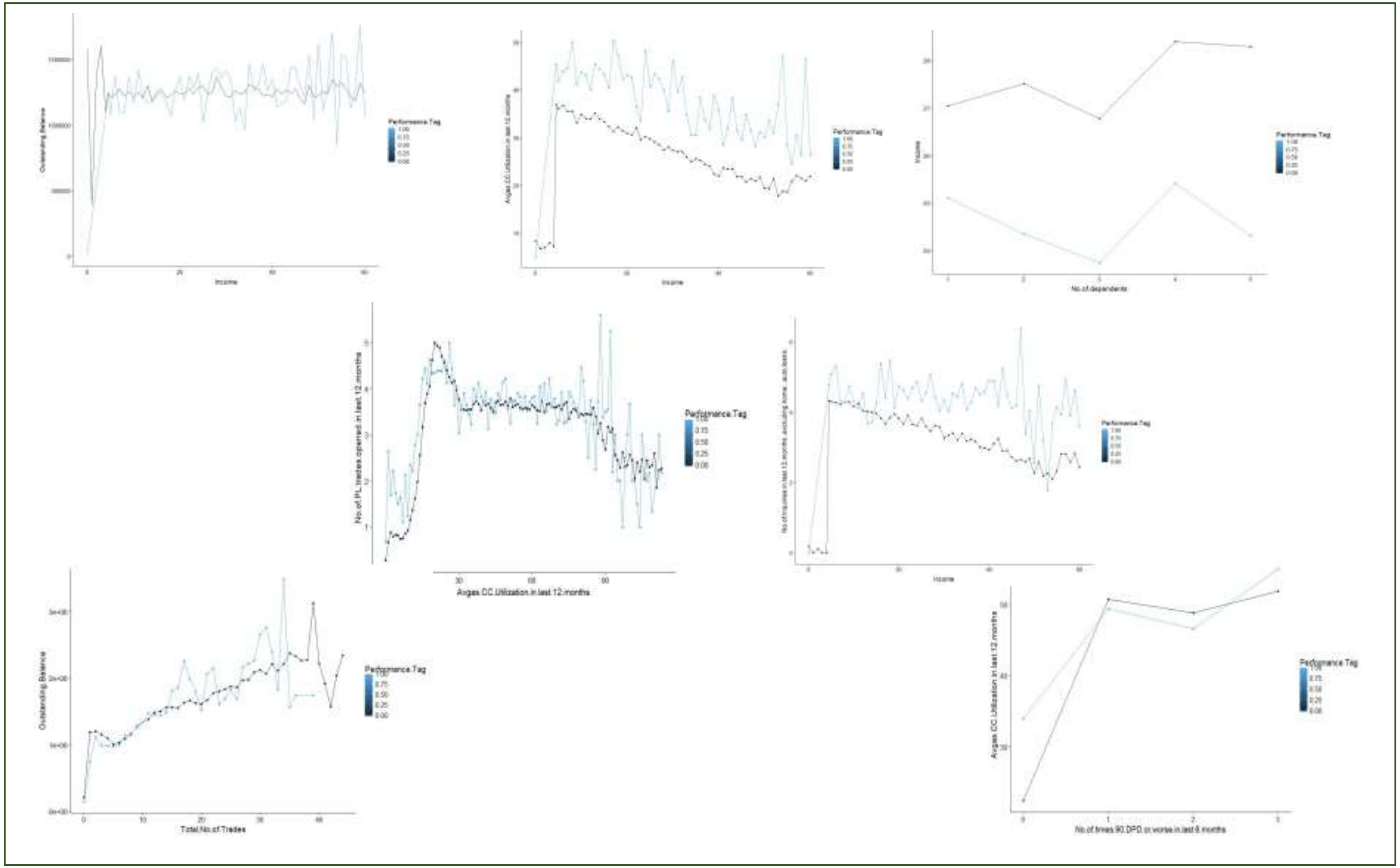

- Both number of trades opened and number of loan inquiries are less right skewed for defaulters, indicating average number of inquiries and trades is likely to be higher for this set

# Boxplots for Outlier Identification




- Outliers are in red
- While demographic variables do not have too many outliers, financial variables like loan enquiries and number of trades have substantial outliers

# Factors affecting Credit Risk- Credit History



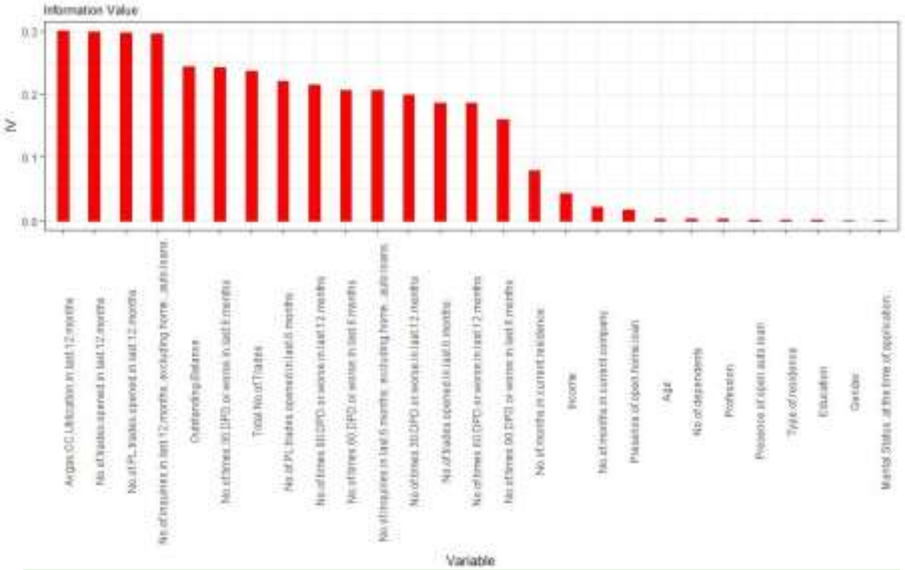

- Outstanding balance remains constant with increase in come, although outstanding is higher for defaulters
- Average credit card utilization is higher for defaulters and in general falls with rise in income
- Income per dependent is lower for defaulters as compared to non-defaulters
- For same level of credit card utilization, Personal Loan trades opened is higher for defaulters
- Number of inquiries falls with rise in income for non-defaulters while for defaulters it is largely constant. Also defaulters have a much higher number of inquiries compared to non-defaulters
- Credit card utilization is also higher for defaulters for minimum and maximum number of times of 90 days DPD, indicating that at extreme values, defaulters may exhibit less control over spending habits
- For the same number of trades, outstanding balance is higher for defaulters, especially as the number of trades increases

# Data Manipulation

- **Missing Value Imputation**
  - With Median : No.of.dependents, No.of.trades.opened.in.last.6.months,Presence.of.open.home.loan,Outstanding.Balance
  - With 0 : NA values in
    - No.of.dependents
    - Presence.of.open.home.loan
    - Outstanding.Balance
    - Avgas.CC.Utilization.in.last.12.months indicate no utilization of the credit card by the user
  - Using Mode :
    - Gender : M
    - Marital.Status..at.the.time.of.application. : Married
    - Education : Professional
    - Profession : SAL
    - Type.of.residence : Rented
- **Binning**
  - Age : Values of less than 10 are imputed with median values
  - Income : Negative values have been imputed with median values
- **Scaling**
  - All numeric columns were scaled : Age, Income,No.of.months.in.current.residence, No.of.months.in.current.company,Total.No.of.Trades,Outstanding.Balance,Avgas.CC.Utilization.in.last.12.months,No.of.times.90.DPD.or.worse.in.last.6.months,No.of.times.60.DPD.or.worse.in.last.6.months,No.of.times.30.DPD.or.worse.in.last.6.months, 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, No.of.trades.opened.in.last.6.months,No.of.trades.opened.in.last.12.months, No.of.PL.trades.opened.in.last.6.months,No.of.PL.trades.opened.in.last.6.months, No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.,No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.,No.of.PL.trades.opened.in.last.12.months,Presence.of.open.home.loan,Presence.of.open.auto.loan
- **Outlier Treatment**
  - The outlier values were treated within the range of 20% - 80% for No.of.months.in.current.company,Avgas.CC.Utilization.in.last.12.months,No.of.trades.opened.in.last.6.months, No.of.trades.opened.in.last.12.months,No.of.PL.trades.opened.in.last.6.months,No.of.PL.trades.opened.in.last.12.months,No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.,No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.,Total.No.of.Trades
- **Dummy Variables**
  - Dummy Variables were created for all categorical variables:
    - Gender, Marital Status,Education,Profession, Type of Residence
  - ApplicationId was not considered for model building

# Factors affecting Credit Risk- Information Value Analysis

## Information Value Analysis




$$IV = \sum(DistributionGood_i - DistributionBad_i) \times WOE_i$$

$$WeightofEvidence = ln(\frac{DistributionGood_i}{DistributionBad_i})$$

- We observe from the plot that IV values lie in between 0.3 and 0.00009
- If the IV statistics is less than 0.01, then the predictor is not useful for modeling (separating the Goods from the Bads)
- If the IV statistics is 0.2 to 0.3, then the predictor has a medium strength relationship to the Goods/Bads odds ratio
- If the IV statistics is 0.3 to 0.5, then the predictor has a strong relationship to the Goods/Bads odds ratio
- None of the demographic variables are important predictors.Some significant predictors with IV>0.2 are credit card utilization, number of inquiries in last 12 months, number of PL trades opened in last 12 months, number of times 30 DPD or worse in last 12 months

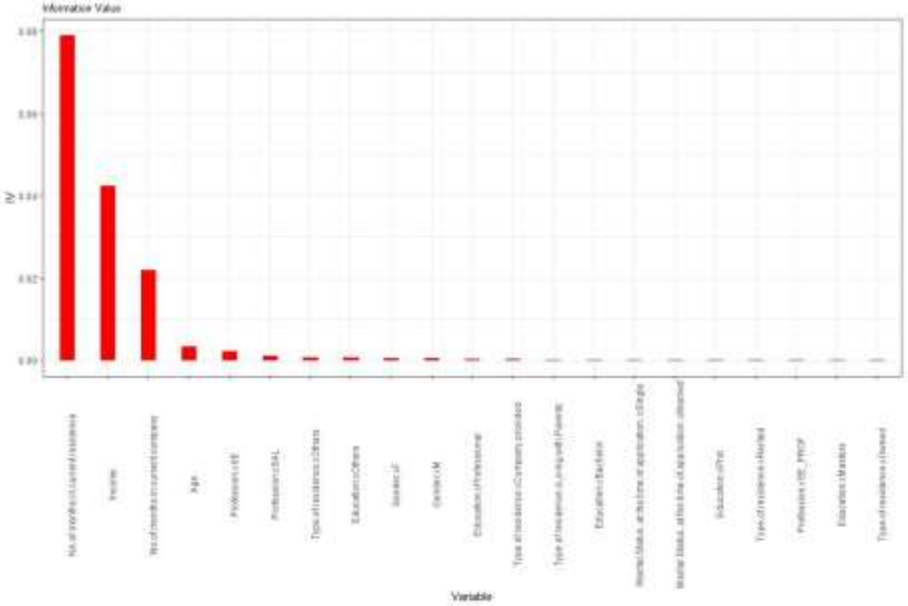
| Variable | IV |
|---|---|
| Avgas.CC.Utilization.in.last.12.months | 0.299347483 |
| No.of.trades.opened.in.last.12.months | 0.297952306 |
| No.of.PL.trades.opened.in.last.12.months | 0.295897393 |
| No.of.Inquiries.in.last.12.months..excluding.home...auto.loans. | 0.295391062 |
| Outstanding.Balance | 0.242788956 |
| No.of.times.30.DPD.or.worse.in.last.6.months | 0.241530717 |
| Total.No.of.Trades | 0.236609276 |
| No.of.PL.trades.opened.in.last.6.months | 0.219734357 |
| No.of.times.90.DPD.or.worse.in.last.12.months | 0.213879846 |
| No.of.times.60.DPD.or.worse.in.last.6.months | 0.205810019 |
| No.of.Inquiries.in.last.6.months..excluding.home...auto.loans. | 0.205160567 |
| No.of.times.30.DPD.or.worse.in.last.12.months | 0.198218247 |
| No.of.trades.opened.in.last.6.months | 0.186015041 |
| No.of.times.60.DPD.or.worse.in.last.12.months | 0.185470651 |
| No.of.times.90.DPD.or.worse.in.last.6.months | 0.160118393 |
| No.of.months.in.current.residence | 0.078962672 |
| Income | 0.042404833 |
| No.of.months.in.current.company | 0.021761038 |
| Presence.of.open.home.loan | 0.016958143 |
| Age | 0.003329512 |

## Top 6 variables

```
> knitr::kable(head(IV_Value$Summary))


|   |Variable                                                         |
|:--|:----------------------------------------------------------------|
|7  |Avgas.CC.Utilization.in.last.12.months                           |
|9  |No.of.trades.opened.in.last.12.months                            |
|11 |No.of.PL.trades.opened.in.last.12.months                         |
|13 |No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.  |
|14 |Outstanding.Balance                                              |
|3  |No.of.times.30.DPD.or.worse.in.last.6.months                     |
```

# Factors affecting Credit Risk- Information Value Analysis –Demographic Data

## Information Value Analysis



| Variable | IV |
|---|---|
| No.of.months.in.current.residence | 7.895394e-02 |
| Income | 4.241078e-02 |
| No.of.months.in.current.company | 2.176071e-02 |
| Age | 3.329732e-03 |
| Profession.xSE | 2.193400e-03 |
| Profession.xSAL | 1.005126e-03 |
| Type.of.residence.xOthers | 6.313788e-04 |
| Education.xOthers | 5.212903e-04 |
| Gender.xF | 3.264734e-04 |
| Gender.xM | 3.239346e-04 |
| Education.xProfessional | 1.702266e-04 |
| Type.of.residence.xCompany.provided | 1.566350e-04 |
| Type.of.residence.xLiving.with.Parents | 1.229363e-04 |
| Education.xBachelor | 9.971957e-05 |
| Marital.Status..at.the.time.of.application..xSingle | 9.546226e-05 |
| Marital.Status..at.the.time.of.application..xMarried | 9.058620e-05 |
| Education.xPhd | 5.858175e-05 |
| Type.of.residence.xRented | 5.437432e-05 |
| Profession.xSE_PROF | 5.319518e-05 |
| Education.xMasters | 3.215927e-05 |
| Type.of.residence.xOwned | 4.324374e-06 |

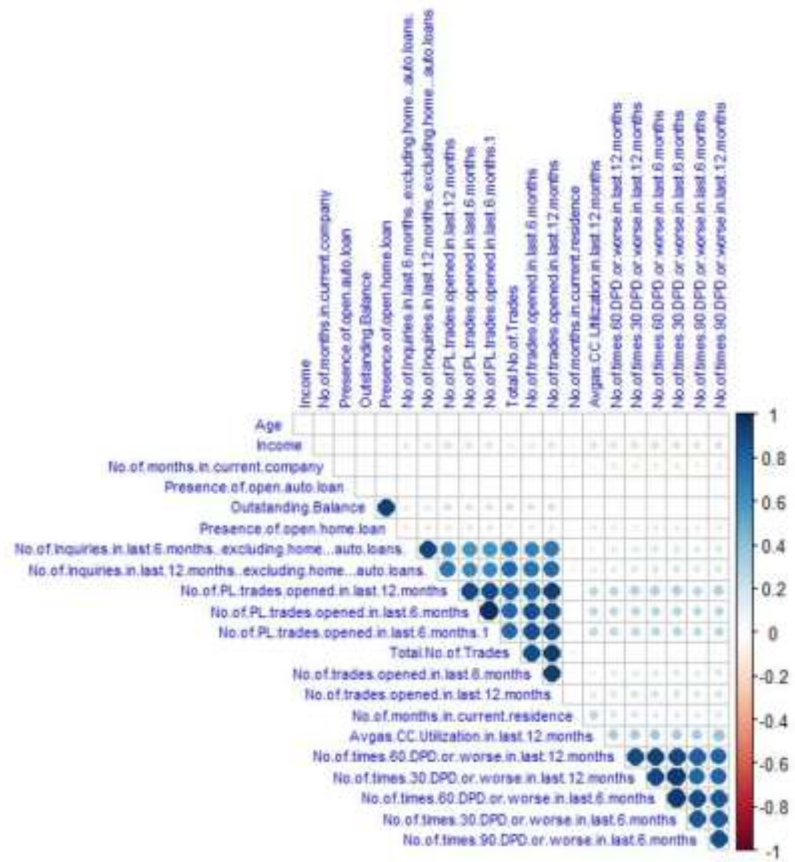
- We observe from the plot that IV values lie in between .08 to almost 0
- Among the demographic variables, number of months in current residence and income are the only variables with IV>=0.04
- As discussed before, demographic variables are weak predictors in this case

## Top 6 variables

| | Variable | IV |
|:--|:--|--:|
| 3 | No.of.months.in.current.residence | 0.0789539 |
| 2 | Income | 0.0424108 |
| 4 | No.of.months.in.current.company | 0.0217607 |
| 1 | Age | 0.0033297 |
| 15 | Profession.xSE | 0.0021934 |
| 14 | Profession.xSAL | 0.0010051 |

# Factors affecting Credit Risk-Correlation Analysis

**Correlation Analysis**




- Income is  Negatively correlated with all credit related attributes, indicating that people with higher income are likely to have better credit history. However the correlation is very minor hence not conclusive
- Number of trades and outstanding balance are positively correlated , as are outstanding balance and average credit card utilization
- Trades, Inquiries and DPD values over different periods of time are correlated, which is expected
- Most of the correlation between credit history variables is expected, however, it indicates need for iterative variable selection using VIF during model building to avoid multicollinearity related issues

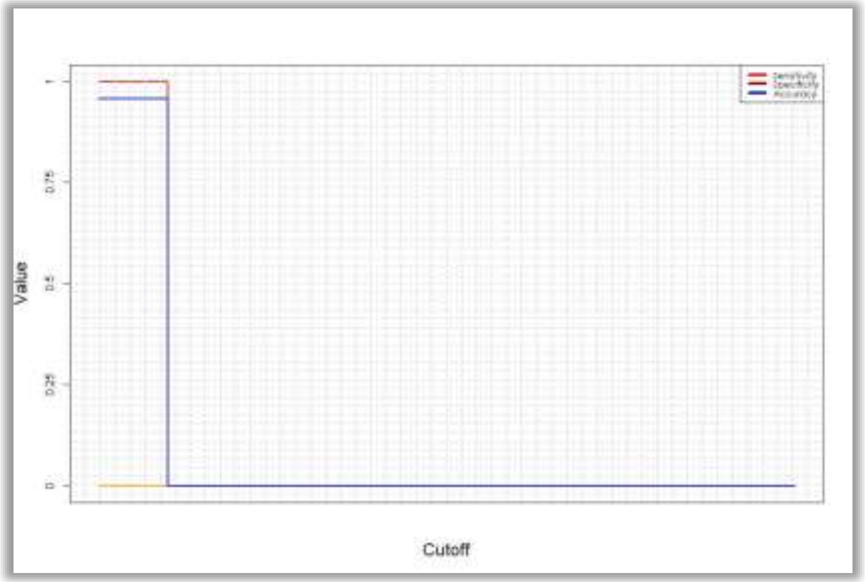
# Sampling, Model Building, Outcomes and Selection

## Logistic Regression based on Demographic data

- As seen before, demographic data is not expected to yield good predictive results
- However, we apply the model to the demographic data to understand which among these might be useful for prediction
- In the logistic model applied iteratively using Step AIC, the final model has only two predictors i.e. Income and Number of months in current company
- These were also among the top 3 in the information value analysis
- As we can see, model is completely biased, classifying most observation in the majority class
- Even after choosing optimal cutoff , sensitivity is 1.3%
- We therefore reject this model and consider other models based on a combination of demographic and credit bureau data

***Final model summary- Logistic Regression-Demographic Data***

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 21171   928
         1   117    13

               Accuracy : 0.953
                 95% CI : (0.9501, 0.9557)
    No Information Rate : 0.9577
    P-Value [Acc > NIR] : 0.9997

                  Kappa : 0.0141
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.0138151
            Specificity : 0.9945039
         Pos Pred Value : 0.1000000
         Neg Pred Value : 0.9580071
             Prevalence : 0.0423321
         Detection Rate : 0.0005848
   Detection Prevalence : 0.0058482
      Balanced Accuracy : 0.5041595

       'Positive' Class : 1
```



# Sampling, Model Building, Outcomes and Selection

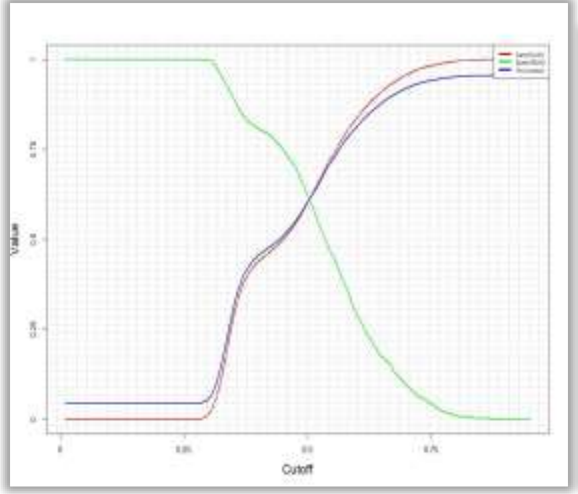
- Only 4% of the data comprises defaulters, hence data is highly unbalanced
- To correct this, we use SMOTE algorithm for oversampling of minority cases in the training data using ROSE package-tis creates a 0-1 distribution i.e. roughly equal
- This mitigates the risk of the model producing biased outcome because of the rare occurrence of the event

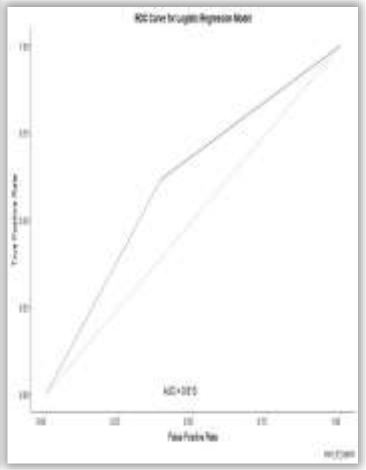
## Logistic Regression

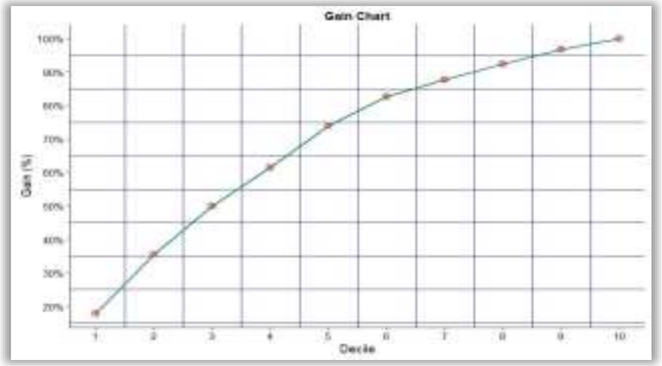
- Logistic regression is performed to predict the log odds of default depending upon categorical and numerical variables
- First model is created using glm() on all variables, next we used stepAIC() to remove insignificant variables
- After several iterations, final model is selected using p-values and VI.
- Variables that have a negative impact on log odds of default are
    - Income
    - Average months in current company
    - Average credit card utilization
- Variables that have a positive impact on log odds of Attrition are
    - 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 90 DPD or worse in last 12 months
    - No of times 60 DPD or worse in last 12 months
    - No of times 90 DPD or worse in last 6 months
    - No of times 30 DPD or worse in last 6 months
    - No of PL trades opened in last 12 months
    - No of inquiries in last 12 months excluding home and auto loans
    - No of PL trades opened in last 6 months
- **Overall Accuracy at optimal cutoff is 61%, Sensitivity-62%,Specificity-61%**
- **KS Statistic is 22.6%, Area under the Curve is 61%**

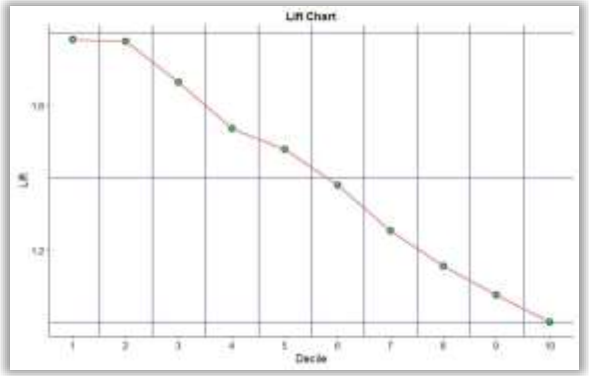
***Final model summary- Logistic Regression***

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 12203   352
         1  7835   568

               Accuracy : 0.6094
                 95% CI : (0.6027, 0.616)
    No Information Rate : 0.9561
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0464
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.61739
            Specificity : 0.60899
         Pos Pred Value : 0.06759
         Neg Pred Value : 0.97196
             Prevalence : 0.04390
         Detection Rate : 0.02710
   Detection Prevalence : 0.40094
      Balanced Accuracy : 0.61319

       'Positive' Class : 1
```





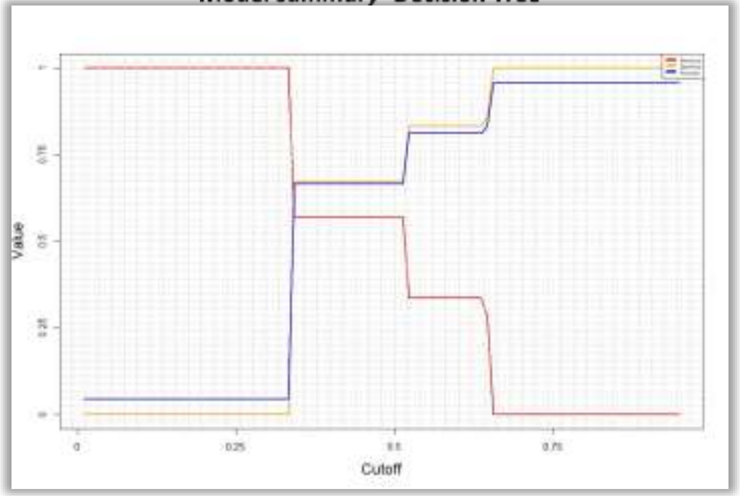
# Sampling, Model Building, Outcomes and Selection

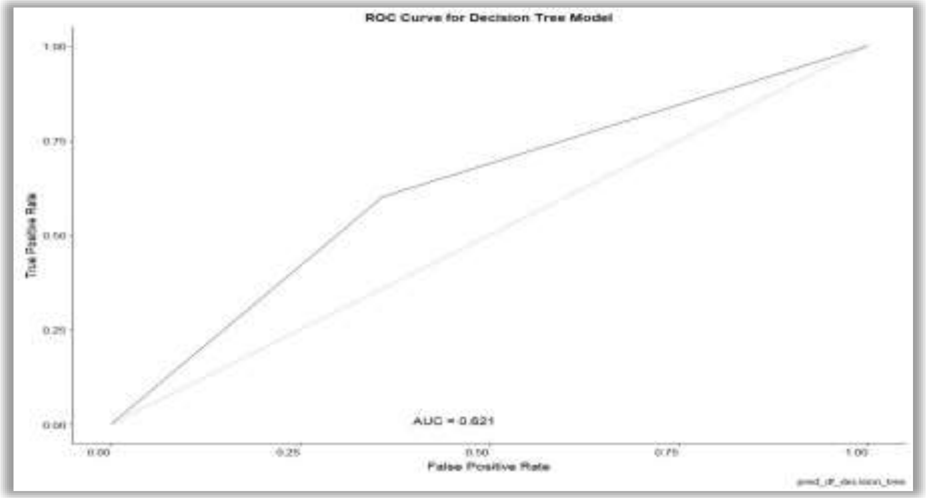
**Decision Tree**

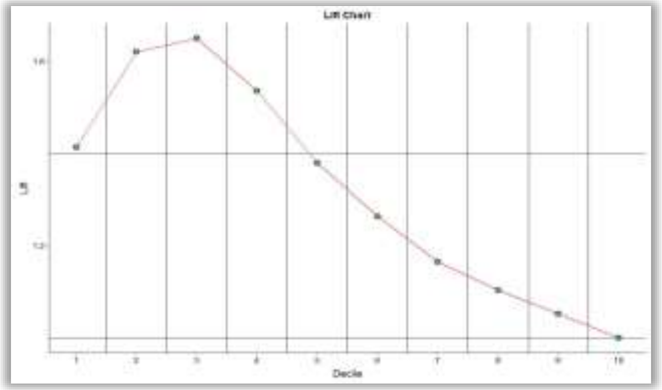
- Model Accuracy-67%, Sensitivity-57%, Specificity-67%
- KS Statistic-25%
- Area under curve-62.1%

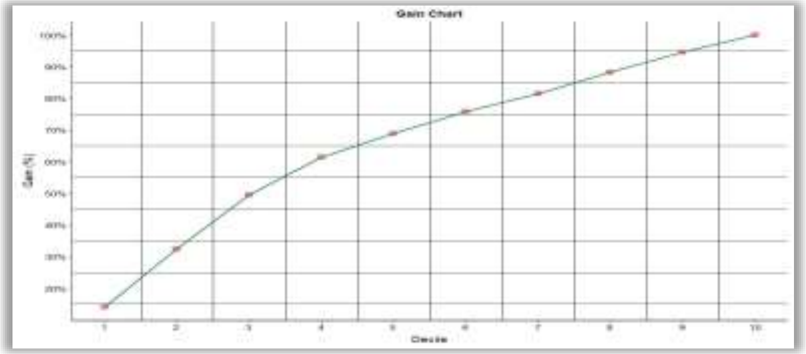
***Model summary- Decision Tree***

```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  12836   366
       yes  7202   554

               Accuracy : 0.6389
                 95% CI : (0.6324, 0.6454)
    No Information Rate : 0.9561
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0534
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.60217
            Specificity : 0.64058
         Pos Pred Value : 0.07143
         Neg Pred Value : 0.97228
             Prevalence : 0.04390
         Detection Rate : 0.02643
   Detection Prevalence : 0.37007
      Balanced Accuracy : 0.62138

       'Positive' Class : yes
```



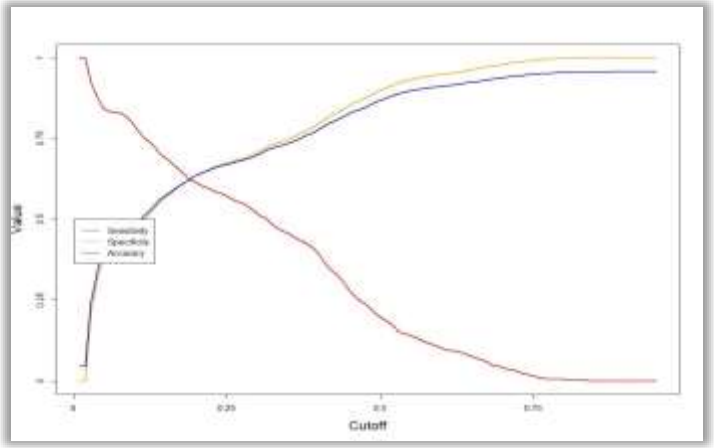
# Sampling, Model Building, Outcomes and Selection

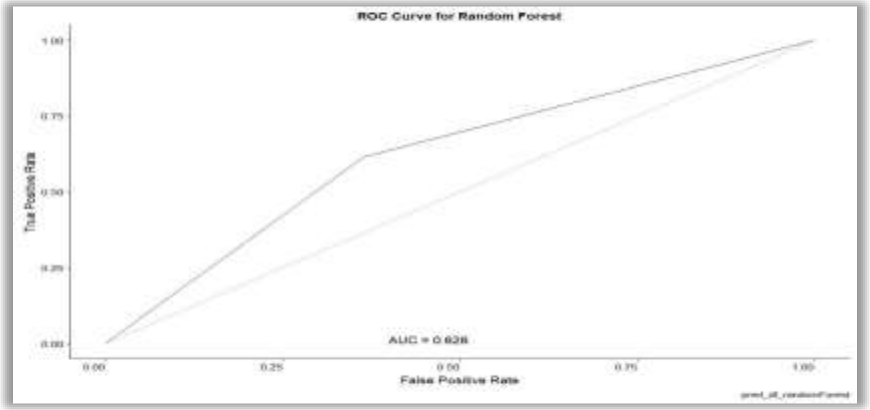
**Random Forest**

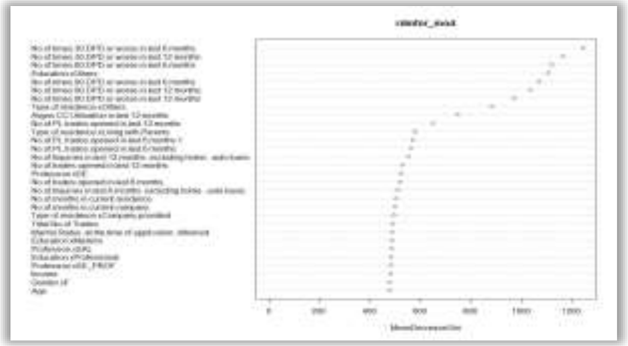
- Model Accuracy-64%, Sensitivity-54%, Specificity-69%
- KS Statistic-24%
- Area under curve-61.7%

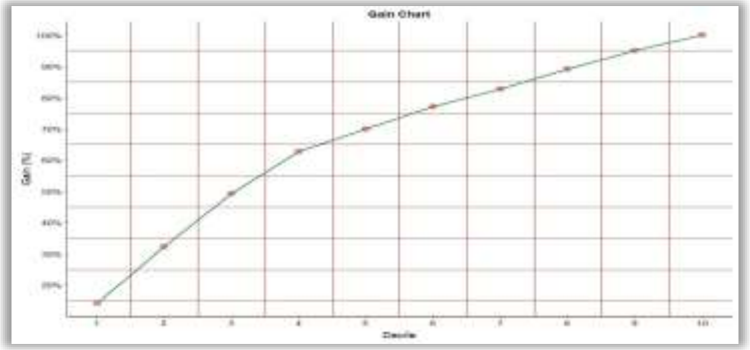
*Model summary- Random Forest*

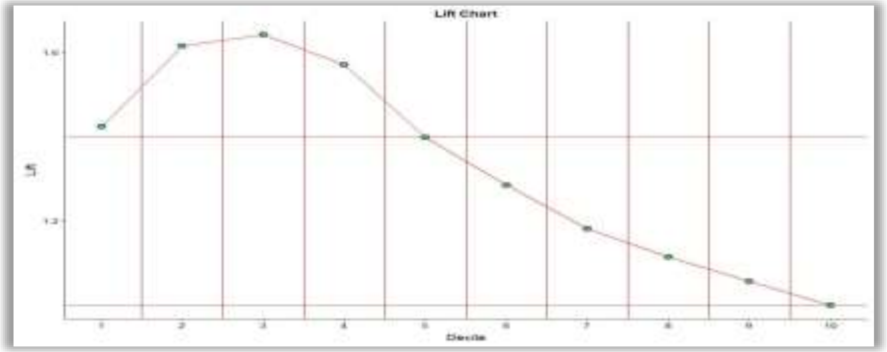
```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  12742   354
       yes  7296   566

               Accuracy : 0.635
                 95% CI : (0.6284, 0.6415)
    No Information Rate : 0.9561
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0546
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.61522
            Specificity : 0.63589
         Pos Pred Value : 0.07199
         Neg Pred Value : 0.97297
             Prevalence : 0.04390
         Detection Rate : 0.02701
   Detection Prevalence : 0.37513
      Balanced Accuracy : 0.62555

       'Positive' Class : yes
```





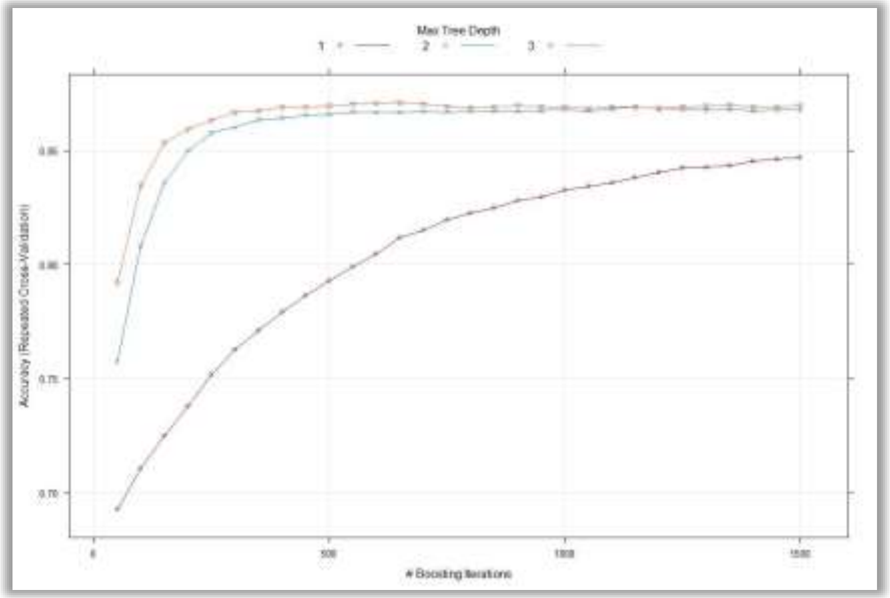
# Sampling, Model Building, Outcomes and Selection

**Gradient Boosting Model**

- Model Accuracy-93%, Sensitivity-6%, Specificity-97%
- Area under curve-51.2%

***Model summary- GBM***

| var | rel.inf |
|---|---|
| No.of.Inquiries.in.last.6.months..excluding.home...auto.loans. | 16.58784400 |
| No.of.PL.trades.opened.in.last.6.months | 15.51838553 |
| No.of.Inquiries.in.last.12.months..excluding.home...auto.loans. | 9.48422862 |
| No.of.times.30.DPD.or.worse.in.last.6.months | 8.60027429 |
| No.of.times.90.DPD.or.worse.in.last.12.months | 8.45394811 |
| No.of.trades.opened.in.last.6.months | 6.51784012 |
| No.of.months.in.current.residence | 4.20034021 |
| Age | 3.58859431 |
| Avgas.CC.Utilization.in.last.12.months | 3.50436474 |
| No.of.times.30.DPD.or.worse.in.last.12.months | 3.34039239 |
| No.of.months.in.current.company | 2.91204561 |
| No.of.PL.trades.opened.in.last.12.months | 2.85470240 |
| Income | 2.41909935 |
| Outstanding.Balance | 2.34350330 |
| No.of.times.60.DPD.or.worse.in.last.12.months | 1.99186636 |
| No.of.trades.opened.in.last.12.months | 1.89996478 |
| Total.No.of.Trades | 1.70854362 |
| No.of.times.60.DPD.or.worse.in.last.6.months | 1.09544022 |
| Presence.of.open.auto.loan | 0.81909822 |
| No.of.times.90.DPD.or.worse.in.last.6.months | 0.44087568 |
| Presence.of.open.home.loan | 0.32460225 |
| Education.xProfessional | 0.23764831 |
| Education.xMasters | 0.17505297 |
| Type.of.residence.xLiving.with.Parents | 0.16552612 |
| Profession.xSAL | 0.10504634 |
| Profession.xSE_PROF | 0.10190310 |
| Education.xPhd | 0.09795234 |
| Type.of.residence.xOwned | 0.08935930 |
| Type.of.residence.xCompany.provided | 0.08754926 |
| Gender.xF | 0.07966643 |
| Education.xBachelor | 0.07674253 |
| Profession.xSE | 0.05281619 |
| Type.of.residence.xRented | 0.04550581 |
| Gender.xM | 0.04052440 |
| Marital.Status..at.the.time.of.application..xMarried | 0.02018120 |
| Marital.Status..at.the.time.of.application..xSingle | 0.01857157 |
| No.of.PL.trades.opened.in.last.6.months.1 | 0.00000000 |
| Education.xOthers | 0.00000000 |
| Type.of.residence.xOthers | 0.00000000 |




```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 13560   613
         1   462    38

               Accuracy : 0.9267
                 95% CI : (0.9224, 0.9309)
    No Information Rate : 0.9556
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0286
 Mcnemar's Test P-Value : 0.000004763

            Sensitivity : 0.05837
            Specificity : 0.96705
         Pos Pred Value : 0.07600
         Neg Pred Value : 0.95675
             Prevalence : 0.04437
         Detection Rate : 0.00259
   Detection Prevalence : 0.03408
      Balanced Accuracy : 0.51271

       'Positive' Class : 1
```

# Sampling, Model Building, Outcomes and Selection

**Support Vector Machine Model**

- SVM models using linear and RBF kernels were run
- However they proved to be too time intensive without corresponding increase in accuracy
- Hence these models are rejected
- Note-for these models, positive class=0, meaning sensitivity refers to true positive rate for majority class i.e. “No” and specificity to true negative for minority class i.e. “Yes”

***Model summary- SVM***

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 11231   300
         1  8807   620

               Accuracy : 0.5655
                 95% CI : (0.5587, 0.5722)
    No Information Rate : 0.9561
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0433
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.56049
            Specificity : 0.67391
         Pos Pred Value : 0.97398
         Neg Pred Value : 0.06577
             Prevalence : 0.95610
         Detection Rate : 0.53588
   Detection Prevalence : 0.55020
      Balanced Accuracy : 0.61720

       'Positive' Class : 0
```

Linear

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 19646   879
         1   392    41

               Accuracy : 0.9394
                 95% CI : (0.936, 0.9425)
    No Information Rate : 0.9561
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0334
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.98044
            Specificity : 0.04457
         Pos Pred Value : 0.95717
         Neg Pred Value : 0.09469
             Prevalence : 0.95610
         Detection Rate : 0.93740
   Detection Prevalence : 0.97934
      Balanced Accuracy : 0.51250

       'Positive' Class : 0
```

RBF

# Model Selection

**Comparative table of metrics across models**

| | Logistic Regression | Decision Tree | Random Forest | GBM |
|---|---|---|---|---|
| Accuracy | 60.90% | 63.89% | 63.50% | 93.54% |
| Sensitivity | 61.70% | 60.22% | 61.52% | 5.22% |
| Specificity | 60.70% | 64.06% | 63.58% | 97.63% |
| KS | 22.63% | 24.27% | 25.11 | |
| ROC | 61.31% | 62.13% | 62.55 | 51.27% |
| Gini | 0.2263 | 0.2427 | 0.2511 | |

- From the table, logistic , decision tree and random forest are all yielding comparable results
- Both decision tree and random forest have higher accuracy and specificity than logistic
- However, random forest has the best measures across , in terms of KS statistics and area under the curve
- Also considering that our aim is to predict the minority class of defaulters, sensitivity is the most important
- Hence we select the random forest model for scorecard creation
- While GBM has high accuracy, it is highly biased with very low sensitivity hence we do not consider this model
- Our choice of model is also validated by the rejected data-while logistic model predicts 36% of the rejected applicants as defaulters, for decision tree and random forest the percentages are 83% and 95% respectively
- Since we expect that rejected applicants have low credit worthiness, using random forest which predicts maximum % of these applicants as defaulters is logical

**Random Forest**

| bucket | total | totalresp | Cumresp | Gain | Cumlift |
|---|---|---|---|---|---|
| 1 | 2096 | 131 | 131 | 14.23913043 | 1.423913043 |
| 2 | 2096 | 166 | 297 | 32.2826087 | 1.614130435 |
| 3 | 2096 | 156 | 453 | 49.23913043 | 1.641304348 |
| 4 | 2096 | 125 | 578 | 62.82608696 | 1.570652174 |
| 5 | 2095 | 65 | 643 | 69.89130435 | 1.397826087 |
| 6 | 2096 | 66 | 709 | 77.06521739 | 1.28442029 |
| 7 | 2096 | 52 | 761 | 82.7173913 | 1.181677019 |
| 8 | 2096 | 59 | 820 | 89.13043478 | 1.114130435 |
| 9 | 2096 | 55 | 875 | 95.10869565 | 1.056763285 |
| 10 | 2095 | 45 | 920 | 100 | 1 |

**Logistic Regression**

| bucket | total | totalresp | Cumresp | Gain | Cumlift |
|---|---|---|---|---|---|
| 1 | 2096 | 166 | 166 | 18.04347826 | 1.804347826 |
| 2 | 2096 | 157 | 323 | 35.10869565 | 1.755434783 |
| 3 | 2096 | 127 | 450 | 48.91304348 | 1.630434783 |
| 4 | 2096 | 117 | 567 | 61.63043478 | 1.54076087 |
| 5 | 2095 | 112 | 679 | 73.80434783 | 1.476086957 |
| 6 | 2096 | 80 | 759 | 82.5 | 1.375 |
| 7 | 2096 | 46 | 805 | 87.5 | 1.25 |
| 8 | 2096 | 45 | 850 | 92.39130435 | 1.154891304 |
| 9 | 2096 | 33 | 883 | 95.97826087 | 1.066425121 |
| 10 | 2095 | 37 | 920 | 100 | 1 |

**Decision Tree**

| bucket | total | totalresp | Cumresp | Gain | Cumlift |
|---|---|---|---|---|---|
| 1 | 2096 | 130 | 130 | 14.13043478 | 1.413043478 |
| 2 | 2096 | 168 | 298 | 32.39130435 | 1.619565217 |
| 3 | 2096 | 157 | 455 | 49.45652174 | 1.648550725 |
| 4 | 2096 | 110 | 565 | 61.41304348 | 1.535326087 |
| 5 | 2095 | 69 | 634 | 68.91304348 | 1.37826087 |
| 6 | 2096 | 63 | 697 | 75.76086957 | 1.262681159 |
| 7 | 2096 | 53 | 750 | 81.52173913 | 1.164596273 |
| 8 | 2096 | 62 | 812 | 88.26086957 | 1.10326087 |
| 9 | 2096 | 59 | 871 | 94.67391304 | 1.051932367 |
| 10 | 2095 | 49 | 920 | 100 | 1 |

# Scorecard

## Calculating Scores

- Post model selection, a scorecard is created using the formula
  - 'Points to double the odds' (pdo = 20)
  - Factor = pdo / ln(2)
  - Offset = Score—{Factor * ln(Odds)}

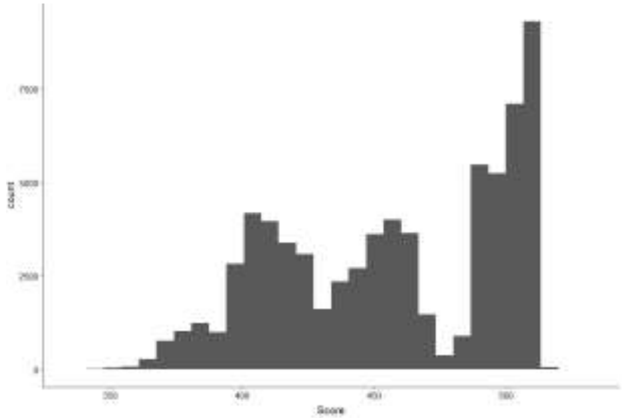
$$\text{Score} = \sum_{i=1}^{n}\left(-\left(woe_i * \beta + \frac{a}{n}\right) * factor + \frac{offset}{n}\right)$$

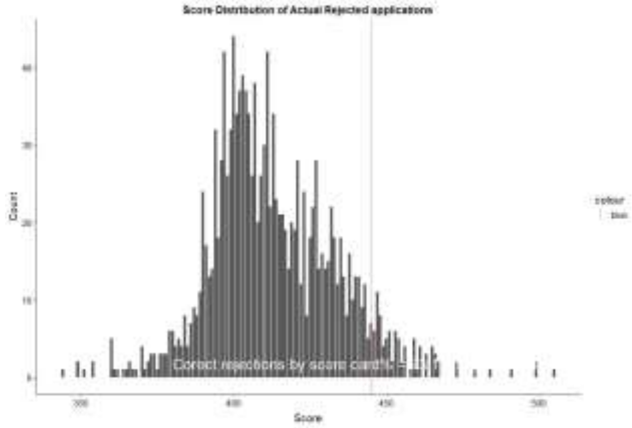
### Quantile distribution of Scores

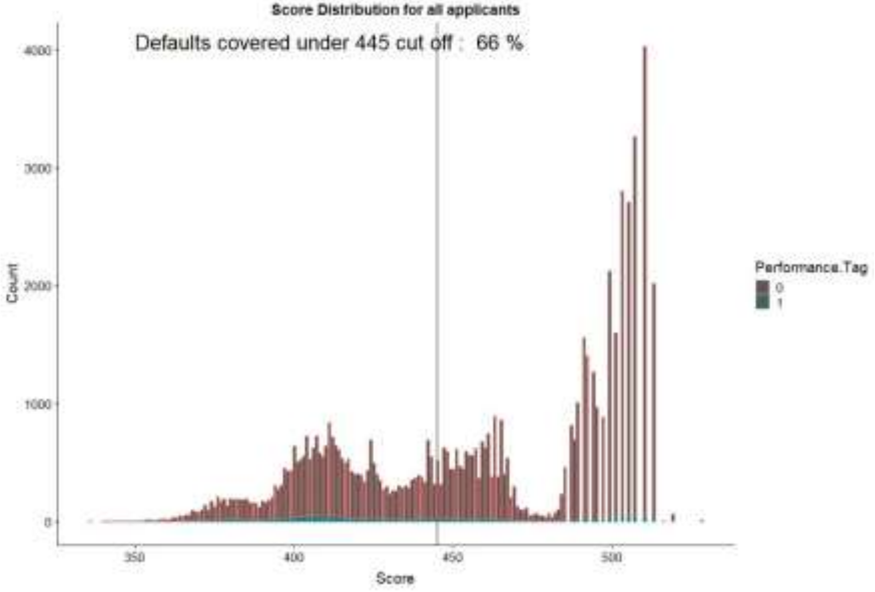
| 0% | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|---|---|---|---|---|
| 336 | 400 | 411 | 425 | 445 | 458 | 483 | 494 | 503 | 507 | 528 |

- Cutoff selection-a cut off of 445 is selected based on the distribution of scores, because very high cutoff will impact approval rate hugely
- Using a cutoff of 445 to reject applicants, we apply the scores to Rejected Data (separated earlier for validation)-93% of these rejections are correctly classified by the scorecard
- Overall 66% defaulters were filtered out based on cutoff=445

### Quantile distribution of Scores



# Financial Benefit Analysis

## Objective

- From the financial perspective, we try to optimize the percentage of defaulters, while also ensuring that we do not compromise revenue by drastically reducing approved applicant percentage
- Also, the ultimate aim is to reduce credit loss for CredX , in terms of the outstanding balance of defaulters

## Benefits

- Originally 4.2% of approved applicants were defaulters, with the model the number of defaulters has come down to 2.4%
- Total credit loss= defaulters outstanding= 3.7B
- Filtering out defaulters with score<445, the new credit loss=1.3B, i.e. almost 3 times less

## Potential loss and Recommendations

- Original Approval Rate- % of applicants granted credit (Total applicants-Rejected Applicants)-98% while new approval rate is 60%
- This implies that there is some potential loss in terms of rejecting credit worthy applicants-34% applicants who were non-defaulters (98%-4.2% defaulters -60% approved) will be rejected using this approach
- However the company can decide to approve applicants in the low-medium score category by imposing a higher rate of interest on these applicants. In this way it can neutralize any losses expected from potential defaulters

## Recommendations for better Customer Selection

- As seen demographic data is not a good predictor of potential default
- In addition, CredX should consider credit history variables, particularly
  - ✔ Average credit card utilization in last 12 months.
  - ✔ No of trades opened in last 12 months.
  - ✔ No of enquiries in last year.
  - ✔ No of time 30 dpd or worse.