Case Study: BFS Capstone Project – Mid Submission

GROUP DETAILS :

1. SHARATH ATHREY
2. SUMIT KAUSHIK
3. VANDHANA SHRI

BUSINESS UNDERSTANDING

PROBLEM STATEMENT :

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.

OBJECTIVE :

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

DATA UNDERSTANDING

DATASETS :

- **Demographics Data** : Applicants information collected during the credit card application. It contains 12 variables listed below.

APP ID	AGE	GENDER	MARITAL STATUS	DEPENDENTS	INCOME	EDUCATION	PROFESSION	TYPE OF RESIDENCE

NO OF MONTHS IN CURRENT RESIDENCE	NO OF MONTHS IN CURRENT COMPANY	PERFORMANCE TAG

- **Credit Bureau Data** : The information is obtained directly from credit bureau. It contains 19 variables listed below.

App ID	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	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)	Presence of open home loan	Outstanding Balance
Total No of Trades	Presence of open auto loan	Performance Tag	

Data Observations :

- 71295 observations (rows) in both Demographics and Credit Bureau Data
- Both datasets can be merged using the common variable "Application ID"
- "Performance tag" variable is the target variable

Data Preparation / Cleaning :

Data Quality (Both Demographics and Credit Bureau):

1. **Unwanted Data**
   - **3** Duplicate Application ID observations found – excluded
   - **65** records found with age < 18 (-3, 0, 15,16, 17) – excluded
   - **107** records found with income <=0 – excluded

2. **Missing Data**
   - **1425** observations found to be missing "Performance tag"
   - **2** observations found to be missing "Gender"
   - **6** observations found to be missing "Marital Status"
   - **3** observations found to be missing "No of Dependents"
   - **119** observations found to be missing "Education"
   - **14** observations found to be missing "Profession"
   - **8** observations found to be missing "Type of Residence"
   - **1058** observations found to be missing "Avgas CC Utilization in last 12 months"
   - **1** observations found to be missing "No of trades opened in last 6 months"
   - **272** observations found to be missing "Presence of Open Home Loan"
   - **272** observations found to be missing "Outstanding Balance"

Detailed Data Analysis for Demographic Data Variables:

1. **Age**

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
18	37	45	45	53	65

Capping the lower values of age with 18.

Binning the age into various buckets and checking the performance rate:

	age	performance	count_prospects	No.of_prospect
1	(18,20]	0.038	2	53
2	(20,30]	0.041	238	5805
3	(30,40]	0.044	830	18688
4	(40,50]	0.042	958	22872
5	(50,60]	0.041	718	17533
6	(60,70]	0.041	200	4825

This shows that people between age 30 to 50 are slightly more likely to default.

2. **Gender**

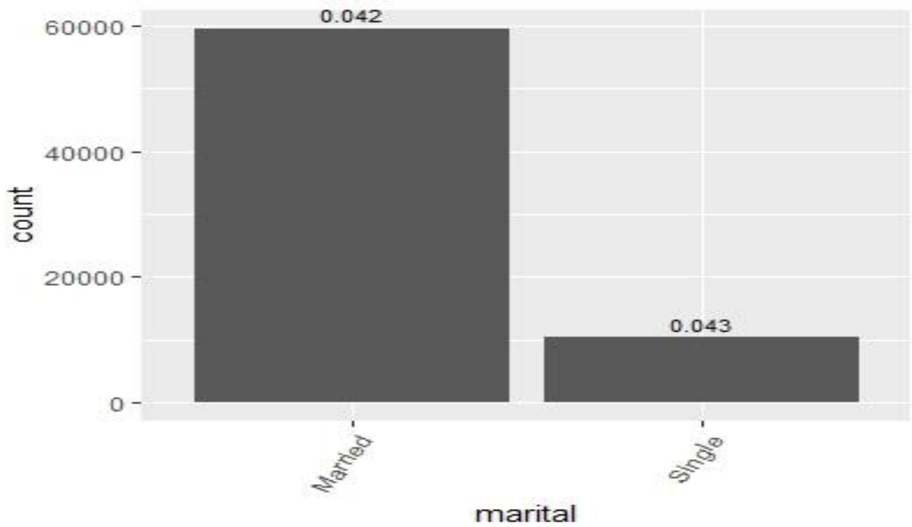
There are three levels:

"" "F" "M"
We are changing spaces level to M.
On plotting there is no significant difference between the performance of males and females

3. **Marital Status**

	Married	Single
6	59542	10316

Replace Unknown level to married

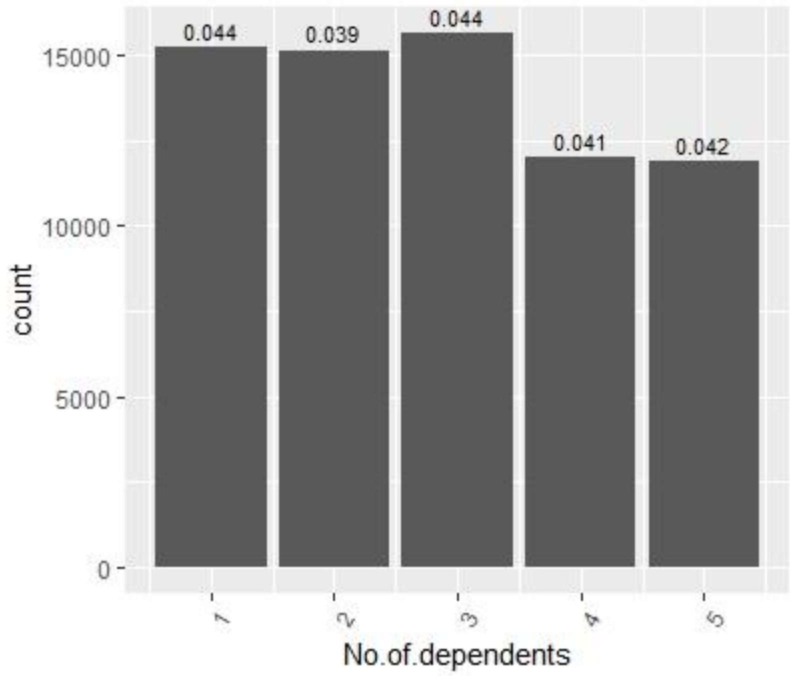


The analysis shows single people are more likely to default.

4. **No. of dependents**

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  1.000   2.000   3.000   2.859   4.000   5.000       3 
```

The 3 NA values will be replaced while WOE analysis.

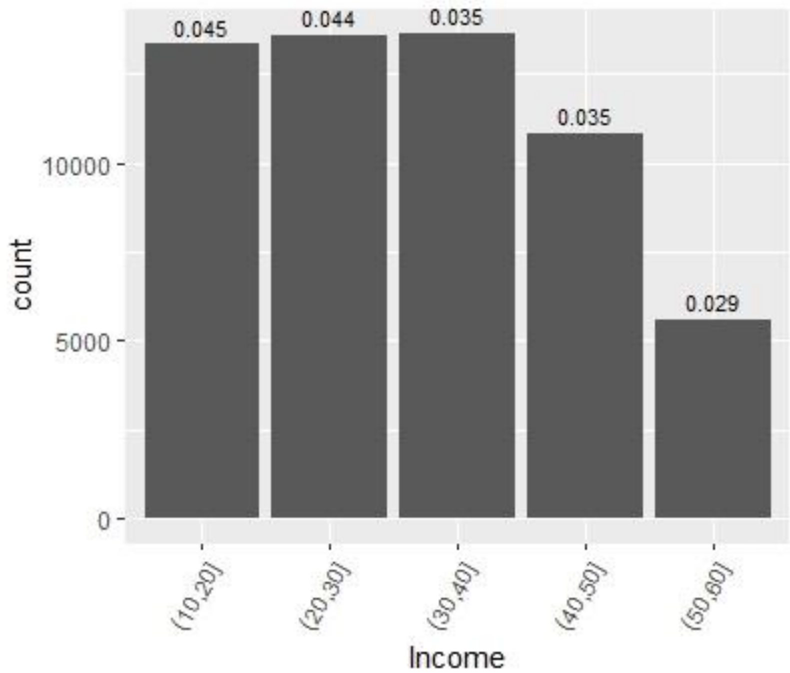


No significant trend on no. of dependents.

5. **Income**

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-0.50   14.00   27.00   27.41   40.00   60.00 
```

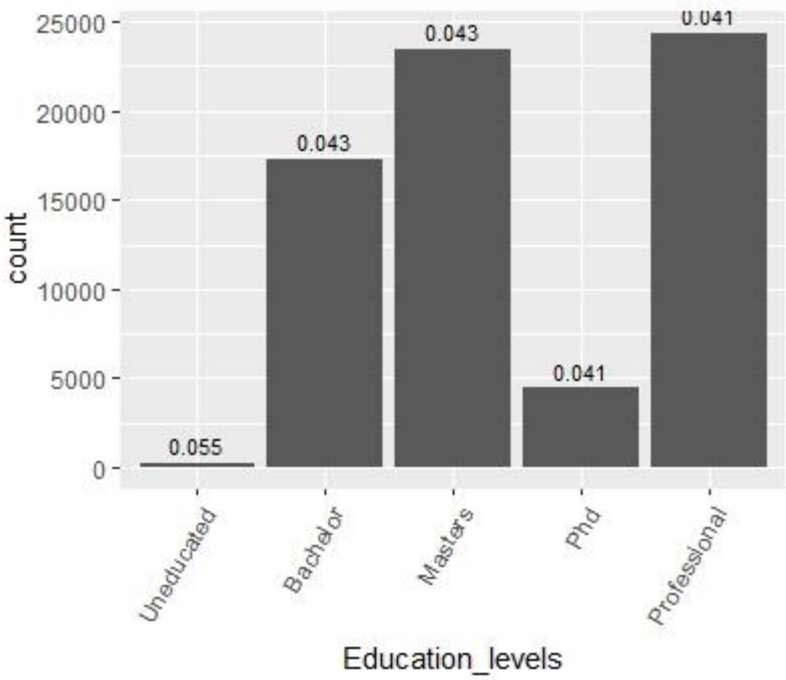
107 records found with income <=0 – excluded



Analysis with performance shows for income > 30 people are less likely to default.

6. Education

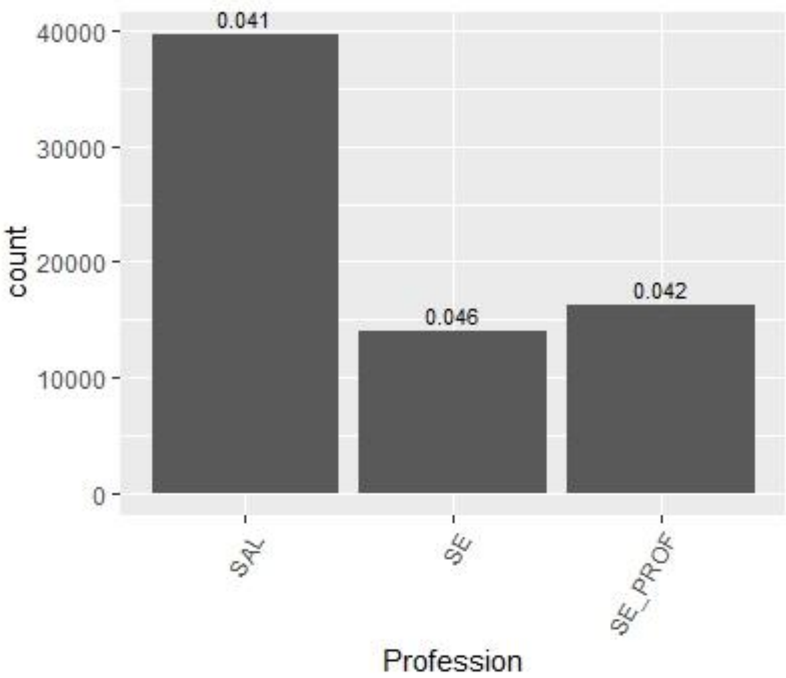
Reducing the levels of education variable, change spaces and others to uneducated.



People with higher education levels are less likely to default.

7. Profession

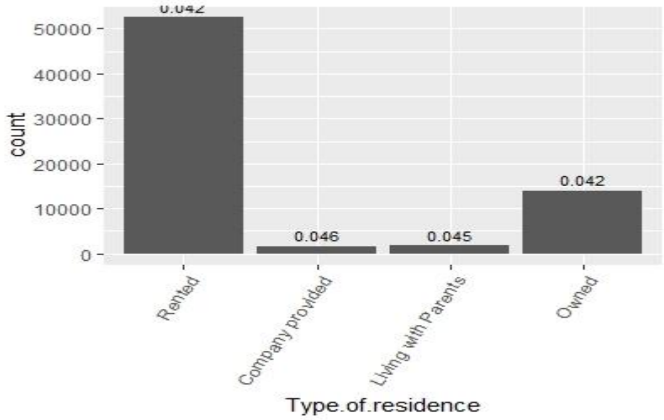
Change unknown level to "SAL"



The 'SAL' people are less likely to default.

8. **Type of residence**

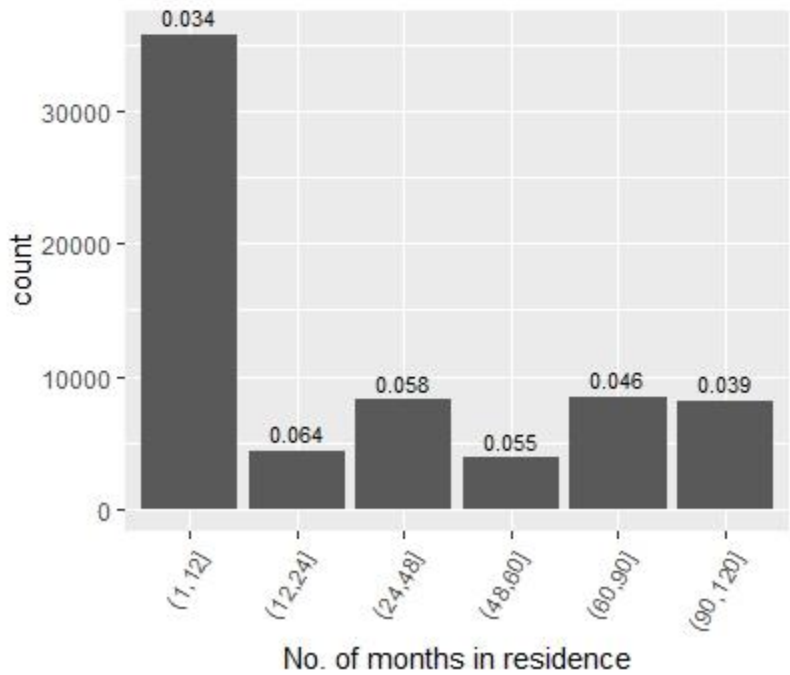
Change spaces and Others level to Rented.



People with rented and owned accommodation are less likely to default.

9. **No of months in current residence**

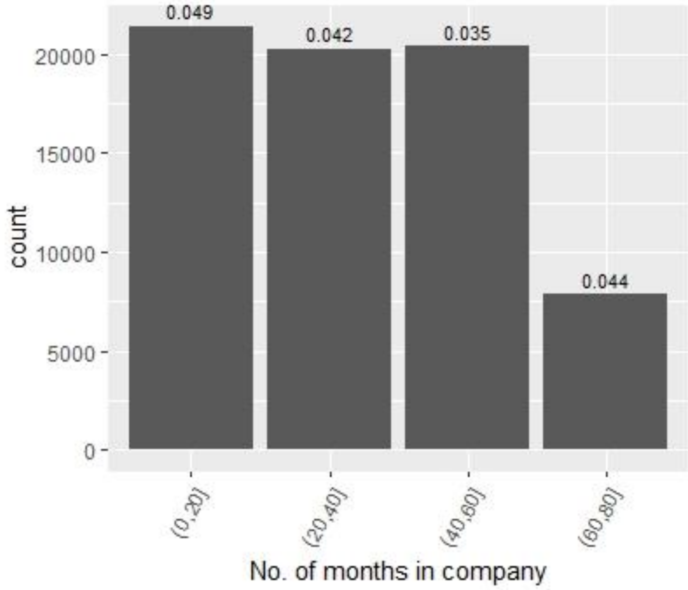
Outliers were treated.



People with more than 90 months and less that 12 months are less likely to be default.

10. No of months in current company

There are outliers shown in box plot that were treated. Binning was done for analysis.

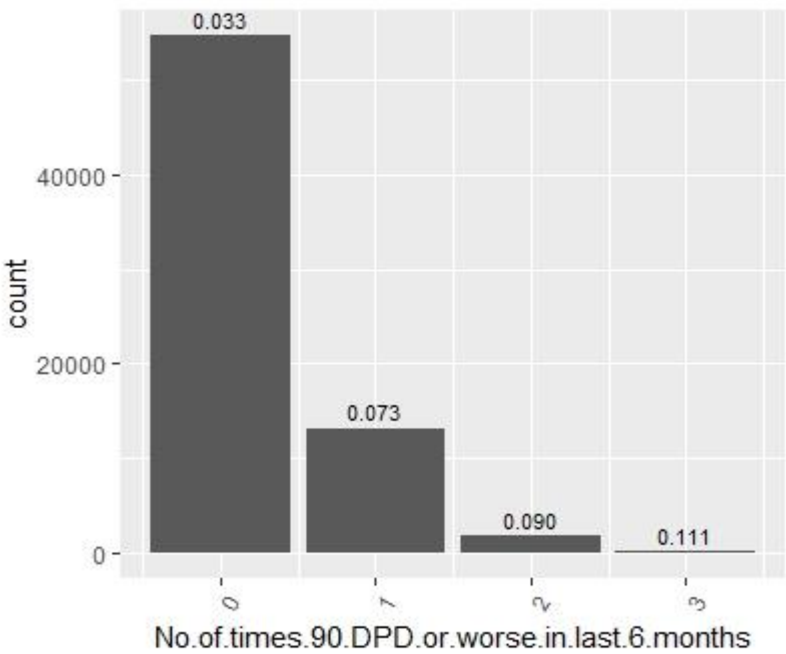


People between 20 and 60 months in current company are less likely to default.

Credit Bureau Data

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

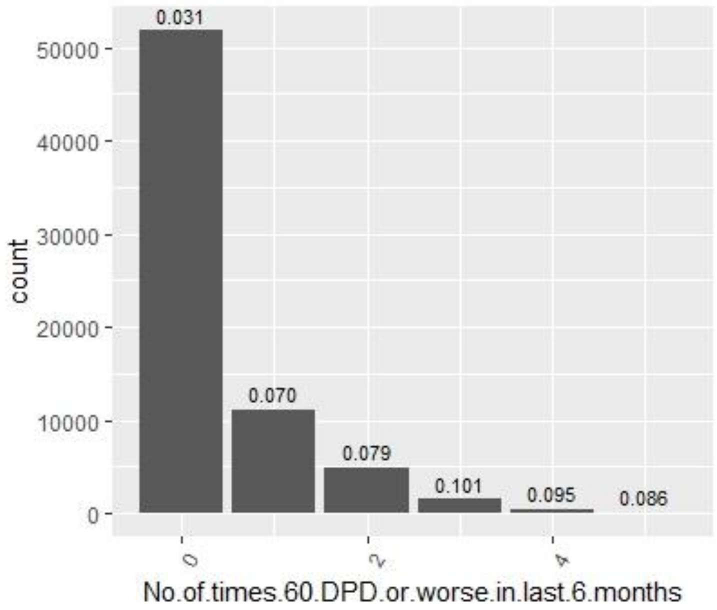
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   0.249   0.000   3.000
```



The more no. of times means person more likely to default.

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

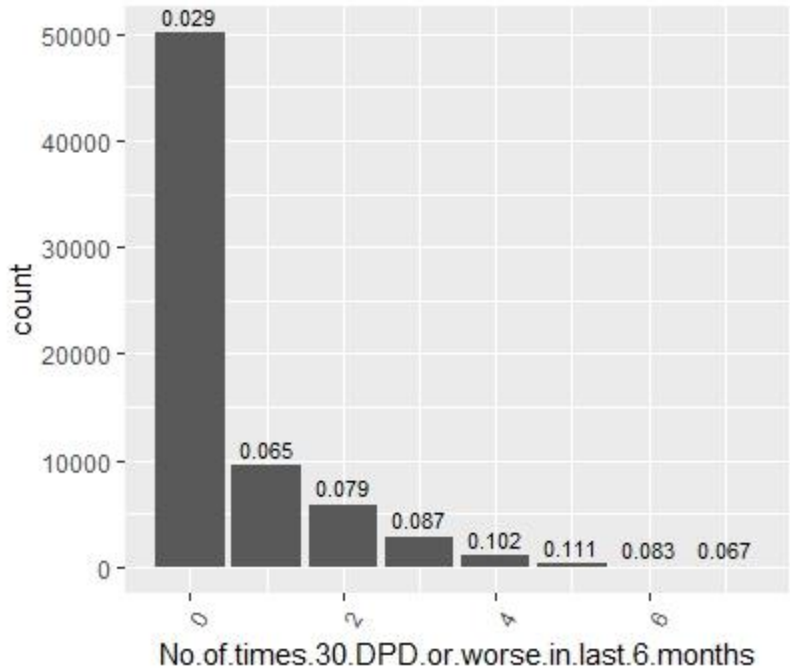
```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0000  0.0000  0.0000  0.3917  1.0000  5.0000
```



The more no. of times means person more likely to default.

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

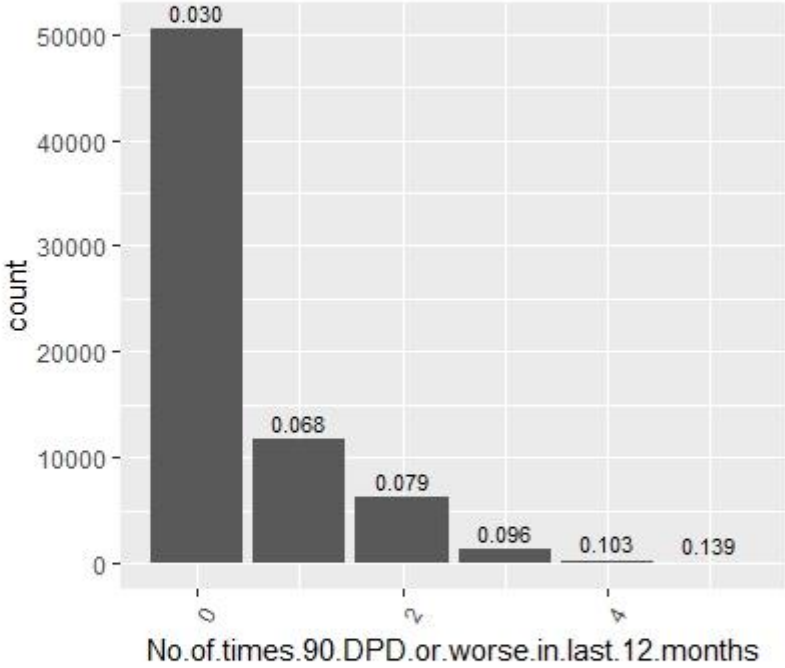
```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.5235  1.0000  7.0000
```



The more no. of times means person more likely to default.

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

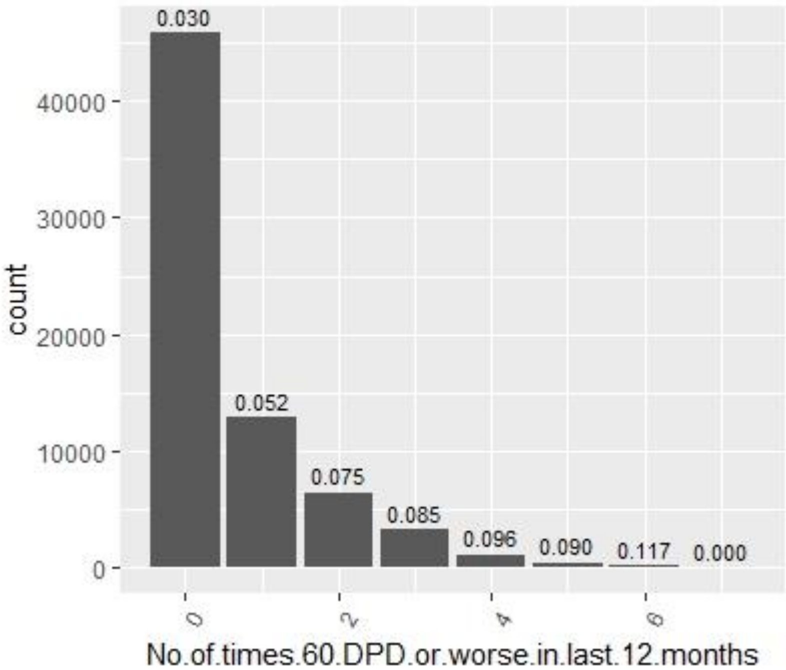
```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0000  0.0000  0.0000  0.4148  1.0000  5.0000
```



The more no. of times means person more likely to default.

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

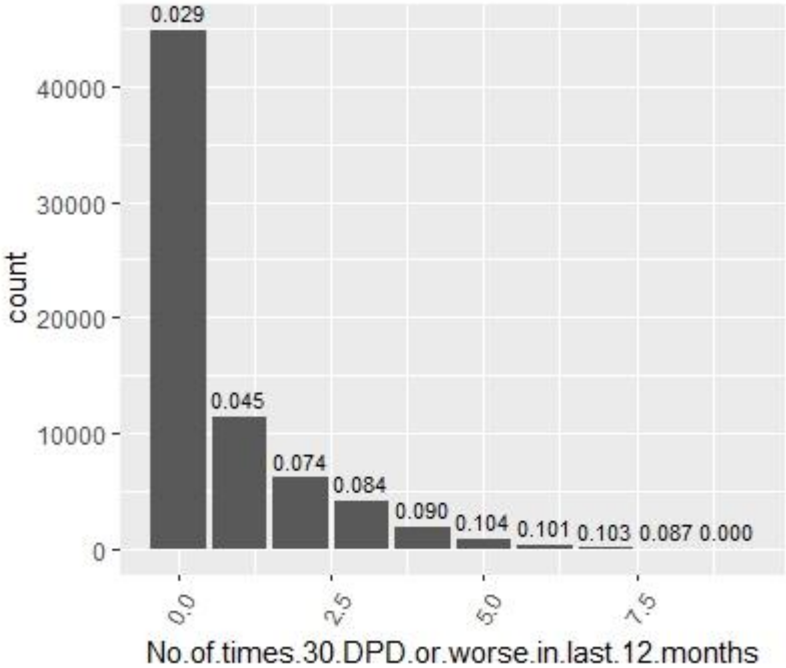
```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0    0.0000  0.0000  0.6034  1.0000  7.0000
```



The more no. of times means person more likely to default.

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

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0     0.0000  0.0000  0.7339  1.0000  9.0000
```



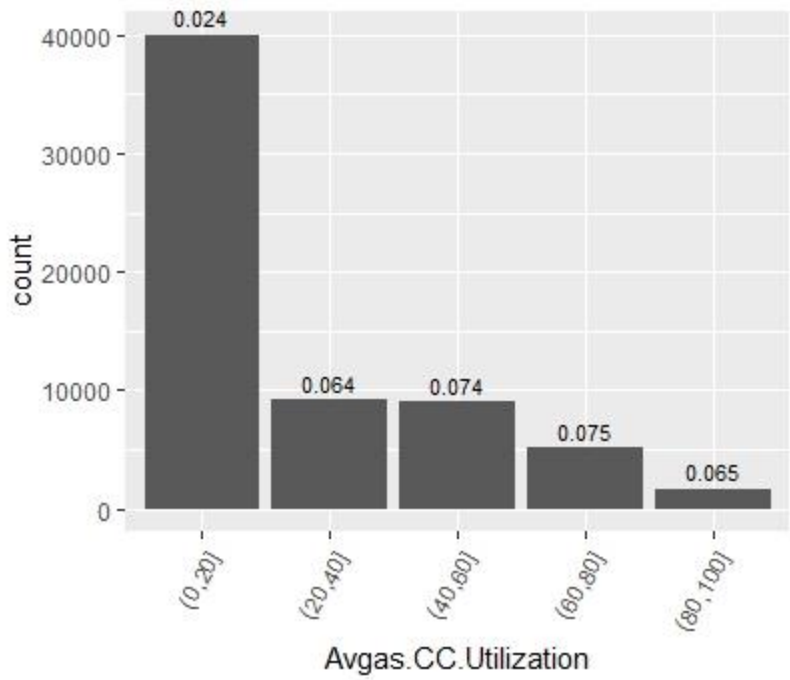
It clearly shows the more number of times means more likely a person is going to default.

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

```
1023 NA values to be handled while WOE analysis.

Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00    8.00   15.00   29.26   45.00  113.00    1023
```

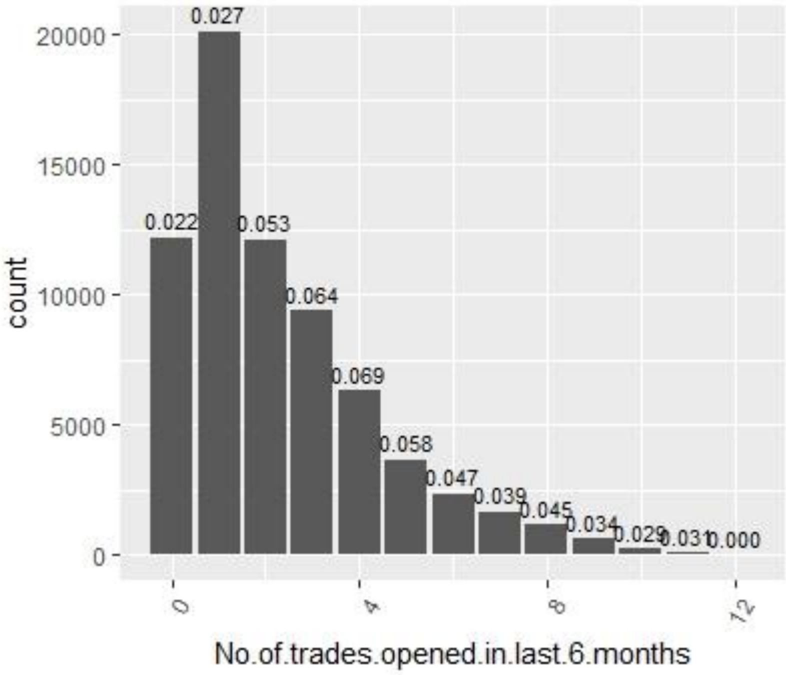
Box plot showed outliers. The outliers treatment was done.
Binning was done for analysis:



People with 0-20 values are less likely to default.

8. No of trades opened in last 6 months

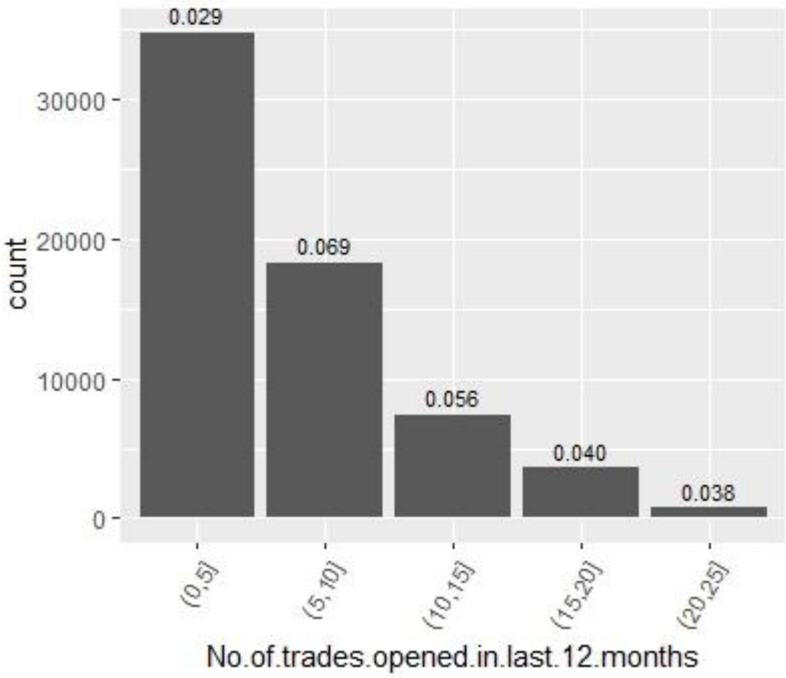
```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
 0.000   1.000   2.000   2.285   3.000  12.000       1 
```



9. No.of.trades.opened.in.last.12.months

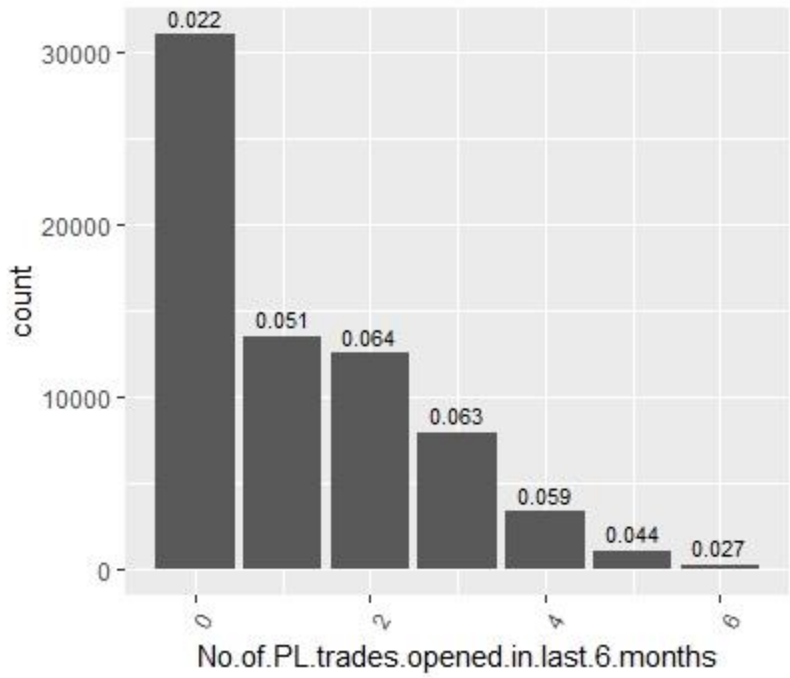
```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   2.000   4.000   5.785   9.000  28.000 
```

Binning the No.of.trades.opened.in.last.12.months



10. No.of.PL.trades.opened.in.last.6.months

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00    1.00    1.19    2.00    6.00
```



People with 0 trades are less likely to default.

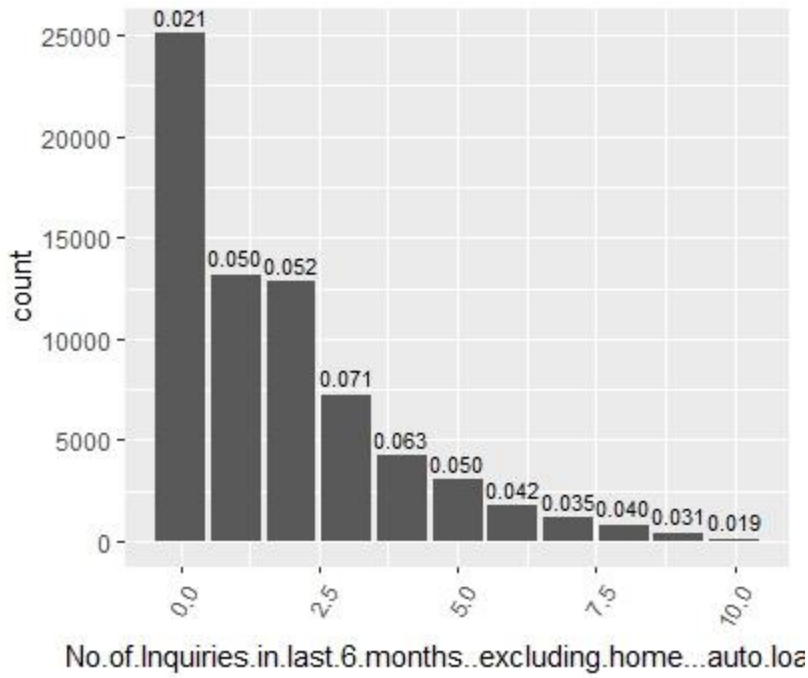
11. No.of.PL.trades.opened.in.last.12.months

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   2.000   2.363   4.000  12.000
```

People with 0 PL trades are less likely to default.

12. No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.

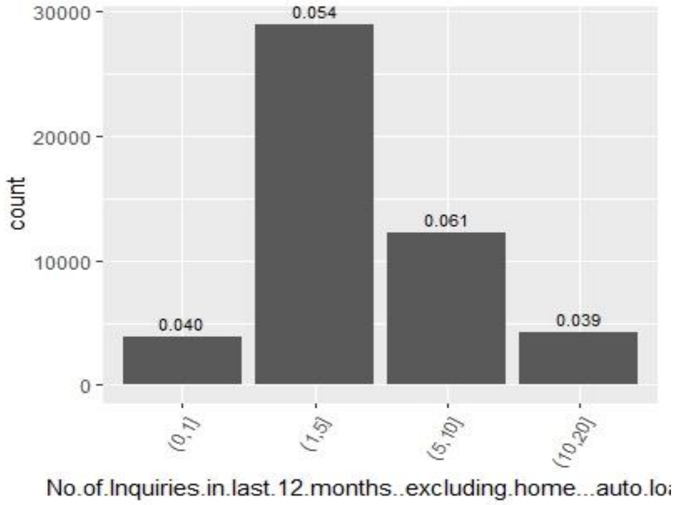
```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   1.000   1.758   3.000  10.000
```



People with 0 inquiries are least likely to default.

13. No.of.Inquiries.in.last.12.months..excluding.home...auto.loans

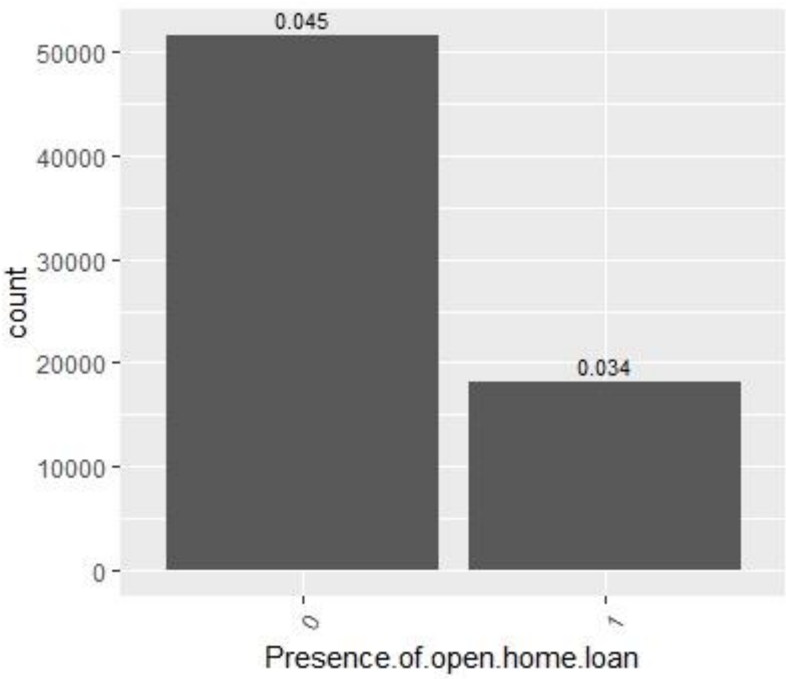
```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   3.000   3.525   5.000  20.000
```



People with less than 1 or more than 10 inquiries are less likely to default.

14. Presence.of.open.home.loan

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 0.0000  0.0000  0.0000  0.2597  1.0000  1.0000     272
```



People with home loan are less likely to default.

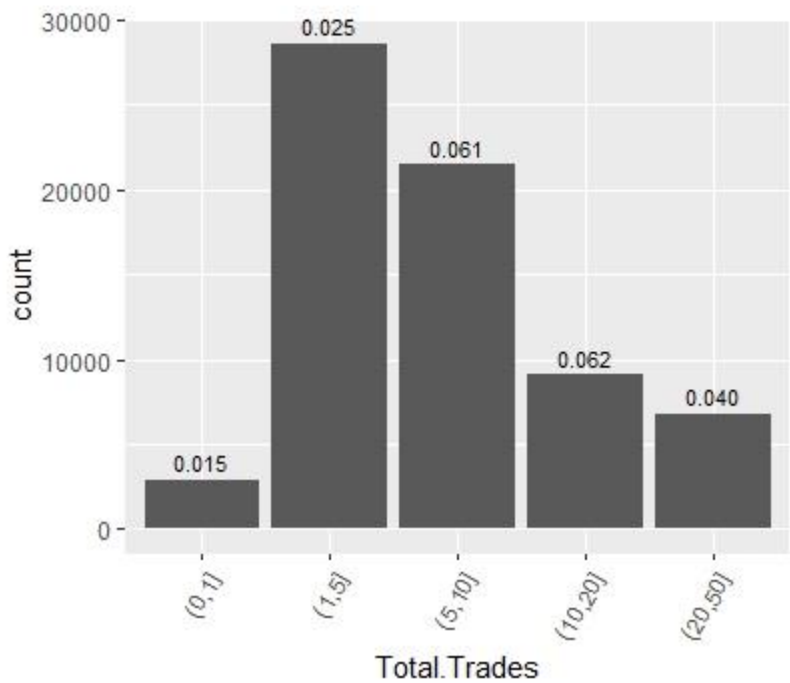
15. Outstanding.Balance

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 0  208400  774242 1253410 2926250 5218801     272
```

Binning was done for analysis with performance.
The analysis shows people with less outstanding balance are less likely to default.

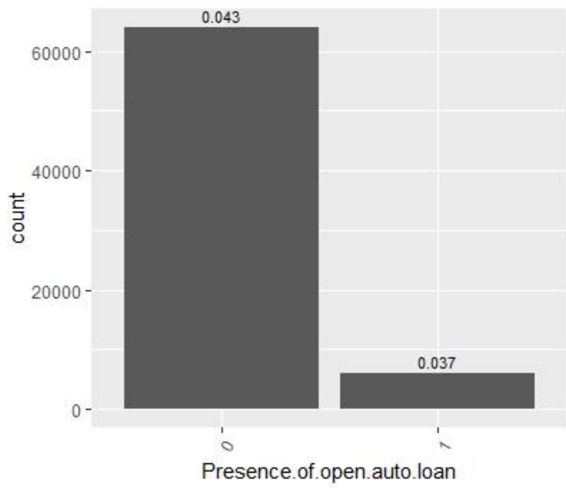
16. Total.No.of.Trades

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.000   3.000   6.000   8.175  10.000  44.000
```



People with less total trades are less likely to default.

17. **Presence.of.open.auto.loan**



People with auto loan are less likely to default.

Weight Of Evidence (WOE), IV Analysis :

- All varibales are analysed using the for WOE using the woe.binning and information packages.
- WOE values are used to replace the missing values in the variables
- From the analysis, it is evident that the variables in the Credit Bureau are significant as compared to the demographic data.
- There are no variables found with predictive power which is strong
- We have identified a total of 12 variables that has relatively less strong (medium) predictive power based on their IV values that are listed below. (Marked yellow)

For demographic Data:

```
                                          Variable           IV
4                     No.of.months.in.current.residence 7.894353e-02
3                                                Income 4.241780e-02
5                       No.of.months.in.current.company 2.175441e-02
1                                                   Age 3.349157e-03
7                                woe.Profession.binned 2.228309e-03
9                                    woe.Gender.binned 3.255737e-04
8                                 woe.Education.binned 2.694278e-04
10                        woe.Type.of.residence.binned 1.093045e-04
11 woe.Marital.Status..at.the.time.of.application..binned 9.592186e-05
2                                      No.of.dependents 5.556324e-05
6                                       Performance.Tag 0.000000e+00
```

For Credit Bureau data:

#All variables except the following 6 variales are monotonically changing across bins:
#"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"
So we will have to make coarse bins for these 6 variables

Overall Table:

Significant Variables (Medium Pred Power)	IV Values
No.of.Inquiries.in.last.12.months	0.2715
Avgas.CC.Utilization.in.last.12.months	0.2607
No.of.times.30.DPD.or.worse.in.last.6.months	0.2415
No.of.times.90.DPD.or.worse.in.last.12.months	0.2138
No.of.times.60.DPD.or.worse.in.last.6.months	0.2058
No.of.times.30.DPD.or.worse.in.last.12.months	0.1982
No.of.trades.opened.in.last.12.months	0.1943
No.of.times.60.DPD.or.worse.in.last.12.months	0.1854
Total.No.of.Trades	0.1822
No.of.PL.trades.opened.in.last.12.months	0.1766
No.of.trades.opened.in.last.6.months	0.1697
No.of.times.90.DPD.or.worse.in.last.6.months	0.1601
Other Variables	**IV Values**

No.of.PL.trades.opened.in.last.6.months	0.12474369
No.of.Inquiries.in.last.6.months	0.09293914
No.of.months.in.current.residence	0.07894352
Income	0.07894352
No.of.months.in.current.company	0.02175441
Presence.of.open.home.loan	0.01762652
Outstanding.Balance	0.0142395
Age	0.0033491
woe.Profession.binned	0.0021820
Presence.of.open.auto.loan	0.001654
woe.Gender.binned	0.00032497
woe.Type.of.residence.binned	0.00028927
woe.Education.binned	0.0002694
woe.Marital.Status..at.the.time.of.application..binned	9.52E-05
No.of.dependents	5.56E-05
Performance.Tag	0

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

Next Steps :

- We are going to use all variables with Information value > 0.02 for the model building.
- We will start with logistic regression model.
- We will then proceed with random tree or SVM models.
- Compare the results of all models and finalize the best model.
- Calculate scorecard using r scorecard package.

Model Building & Evalution :

- **Logistic Regression / Random Forest** estimators will be used
- We plan to build 2 models
    - **Demographic Model**
    - **Merged Data Model**
- Removal of Insignificant variables and model evaluation will be based on Sensitivity, Specificity and Accuracy.
- Application scorecard will be built on the final model leading to cut-off score

WOE and IV Formulas Usage :

$$\text{WoE:} \quad [\, ln\left(\frac{\text{Distr Good}}{\text{Distr Bad}}\right)] \times 100.$$

$$\text{IV:} \quad \sum_{i=1}^{n} (\text{Distr Good}_i - \text{Distr Bad}_i) * ln\left(\frac{\text{Distr Good}_i}{\text{Distr Bad}_i}\right)$$

Application Score Formula = (β×WoE+ α/n)×Factor + Offset/n

Where:
β—logistic regression coefficient for characteristics that contains the given attribute
α—logistic regression intercept
WoE—Weight of Evidence value for the given attribute
n—number of characteristics included in the model
Factor, Offset—scaling parameter