BFS CAPSTONE PROJECT

Submitted by:
1. Venkata Ramana Gudi
2. Munish Bansal
3. Anjali Nair
4. Rathan Gopi

BUSINESS CASE

Problem Statement:
CredX is experiencing an increase in credit loss for the past few years.

Objective:
The objective of this case study is to acquire the right customers to mitigate the credit risk.

Goal:
Determine the factors affecting credit risk, create right strategies & models to mitigate the acquisition risk and assess the financial benefit .

Approach

Data Understanding

- Understand data w.r.t . Data dictionary and data files provided

Data cleanup and preparation

- Import data files into R and merge datasets
- Check for missing values, outliers, treat them & rename variables
- Convert variables into categorical & continuous values and derive new matrices

Exploratory data analysis

- Perform the Uni-variate and Bi-variate Analysis with graphs
- Perform WOE & IV Analysis for influential factors and use WOE values
- Prepare cleaned dataset for modeling

Model Building

- Separate data into Train & Test datasets
- Try different classification models e.g. Logistic and Random Forest
- Perform Data Balancing as well and do the modelling

Model Evaluation

- Check the final model against Test dataset
- Performance metrics using confusion matrix,ROC,AUC KS

Model Validation

- Validate model using K-fold method
- Compare Performance metrics across logistic & RF models

Conclusion

- Choose the best model
- Build the application score-card with good & bad odds
- Calculate the Score on the Rejected Applications to identify potential customers
- Come up with the Financial Benefits

Persona

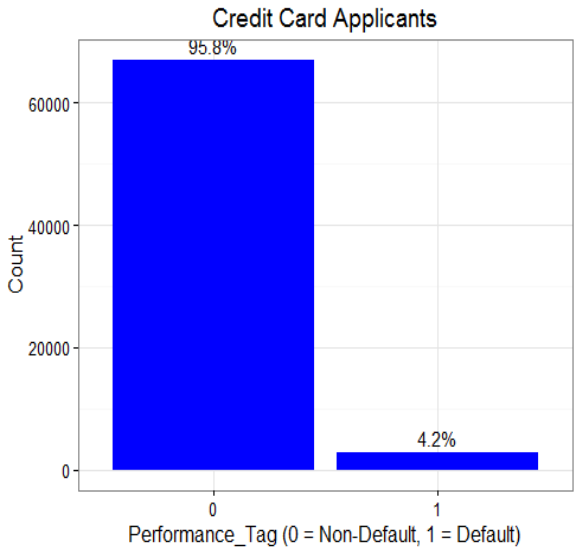
High level data findings:

Data Understanding:

- The credx company has provided us two datasets, demographic/application data and credit bureau data.
- The demographic data is obtained from the information provided by the applicants at the time of credit card application, which includes customer level information on age, gender, income, marital status, Performance tag etc.
- The credit bureau data contains variables such as number of time 30daypassdue or worse in last 3/6/12months, outstandingbalance, numberoftrades, presenceofhomeloan, presenceofautoloan, performancetag etc.
- The demographic data consists of **71295** observations with **12** variables including **1577** NA's and 3 duplicates applicationid, the credit bureau data consists of **71295** observations with **19** variables including **3028** NA's 3 duplicates applicationid.

- Master Data obtained by merging two data sets after removing the duplicated application id's Data has **71292** observations and **30** variables.

Performance tag trend

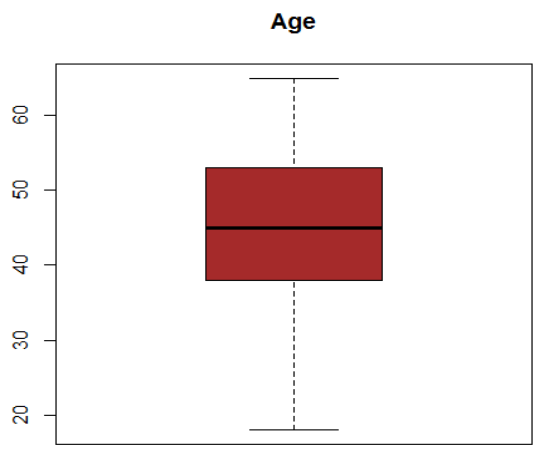


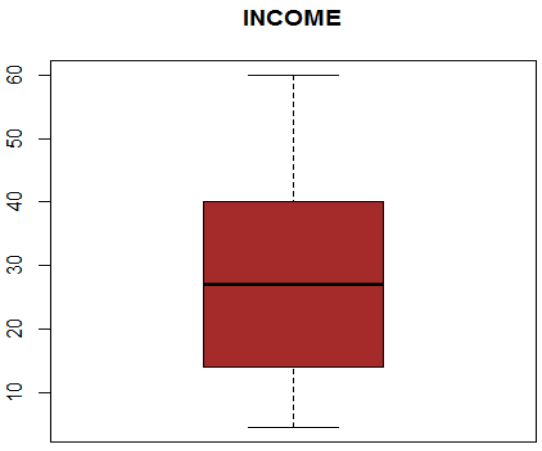
Inference & Actions:

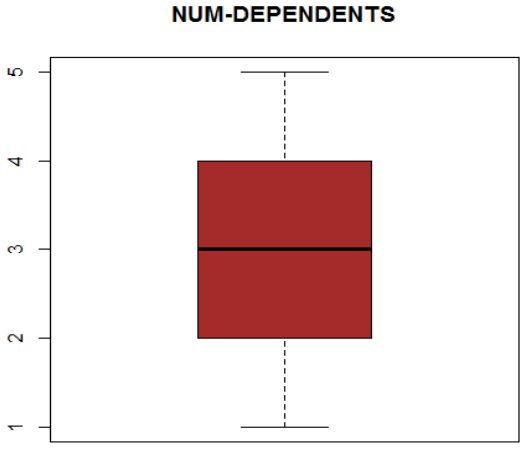
- There are 1425 records with missing PerformanceTag values which can be removed as this signifies customers with applications rejected.
- Records with variables having NA values as well as outliers treated accordingly e.g. average credit card utilization in 12months, outstanding balance etc.
- Some variables have faulty data , treated as well e.g. Age with negative values removed
- Continuous variables are converted into **categorical and binned** them appropriately. e.g. CC utilization, outstanding balance, number of trades etc.
- The data is very much **imbalanced** as ,defaults are 4.2% compared to non-defaults which are 96% hence sampling techniques are applied

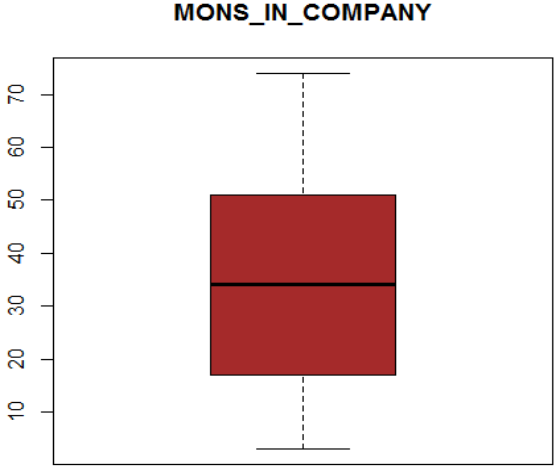
- Application Rejected Rate ~2%

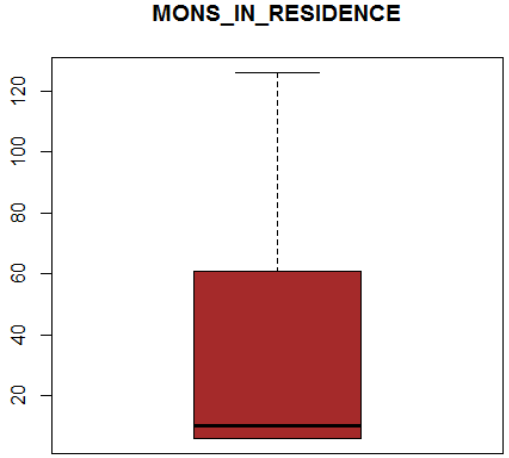
Box plots after Outlier Treatment

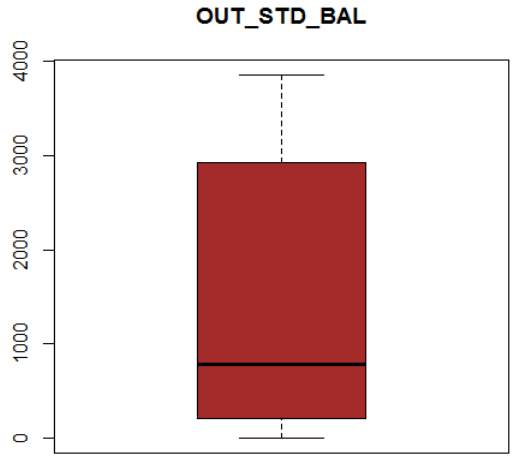






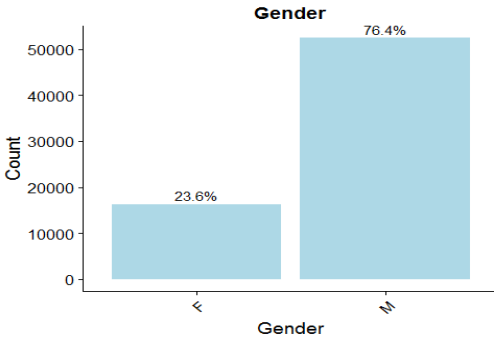


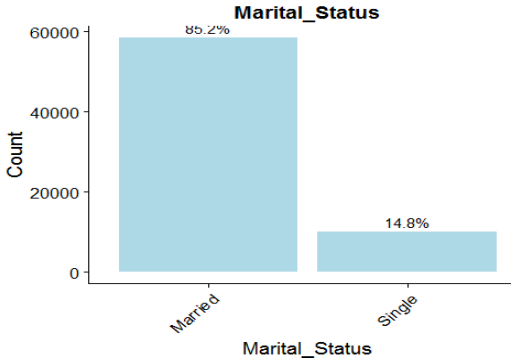


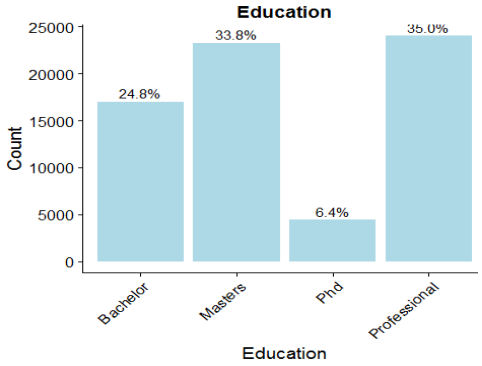


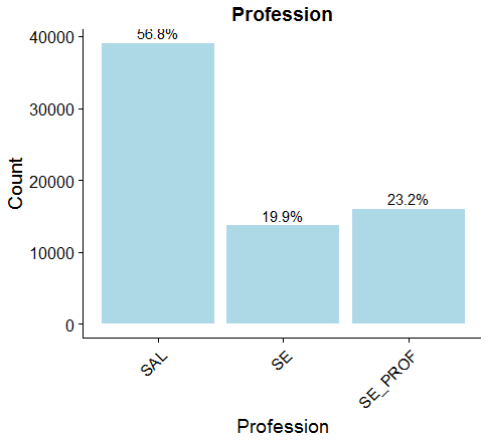
Personal

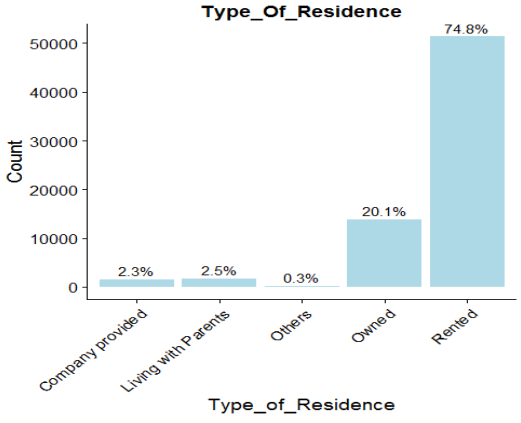
Demographic Data Distribution

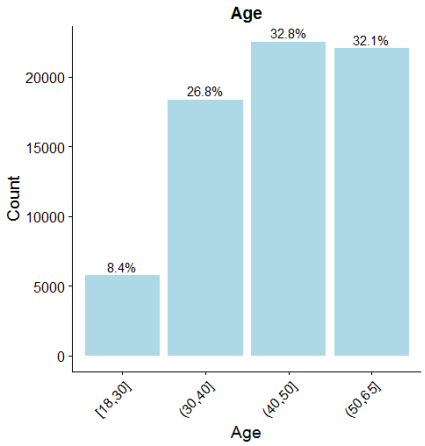






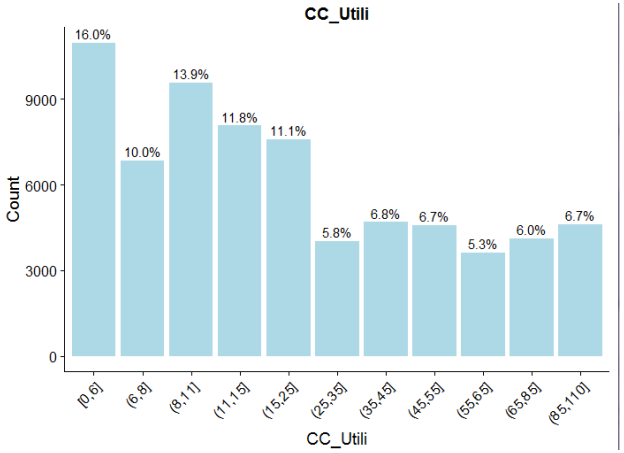


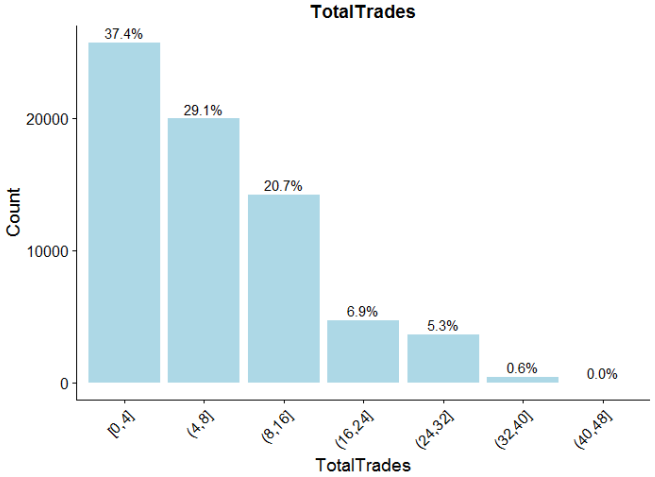


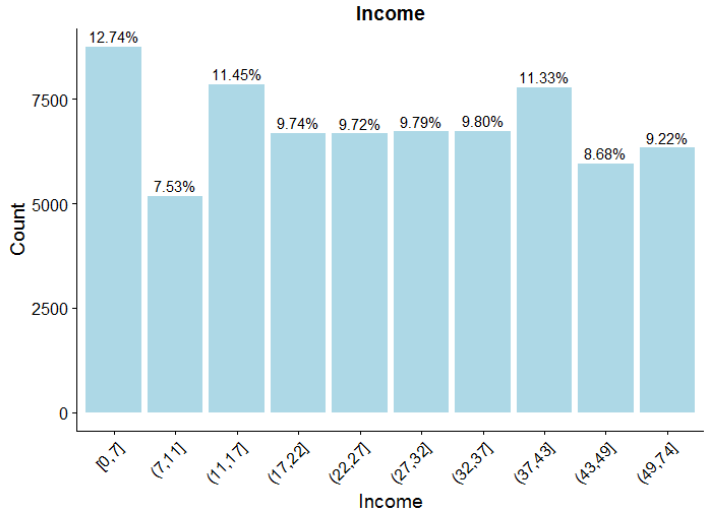


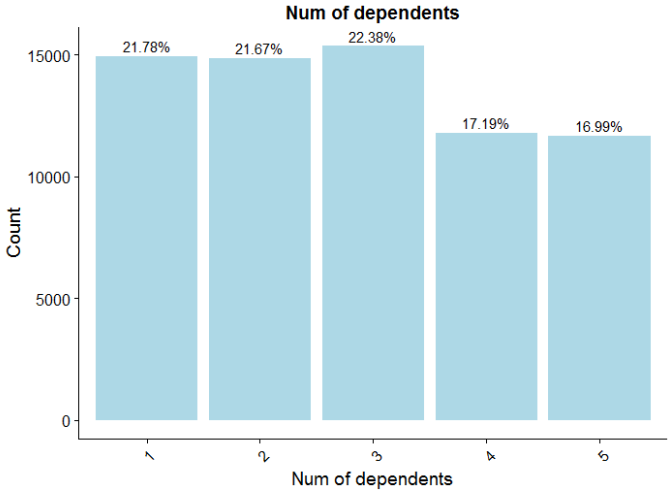
- There are more male applicants than female
- Married people having credit card much more than singles
- Salaried people are using credit card more than other professionals
- People with rented accommodation having highest credit cards
- Age group of 40-50 has maximum people carrying credit card.

Credit Bureau Data Distribution



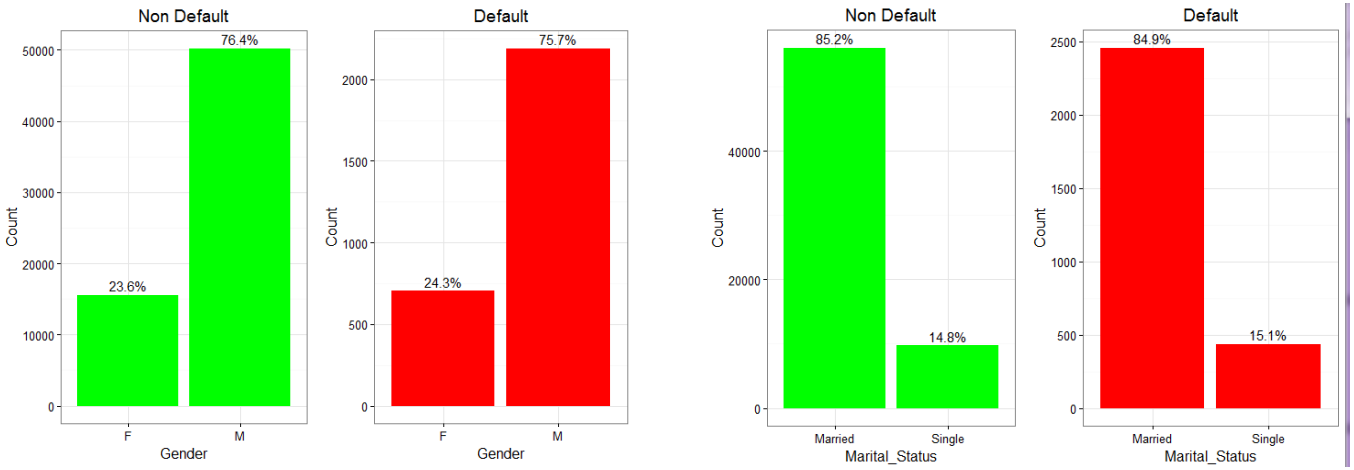


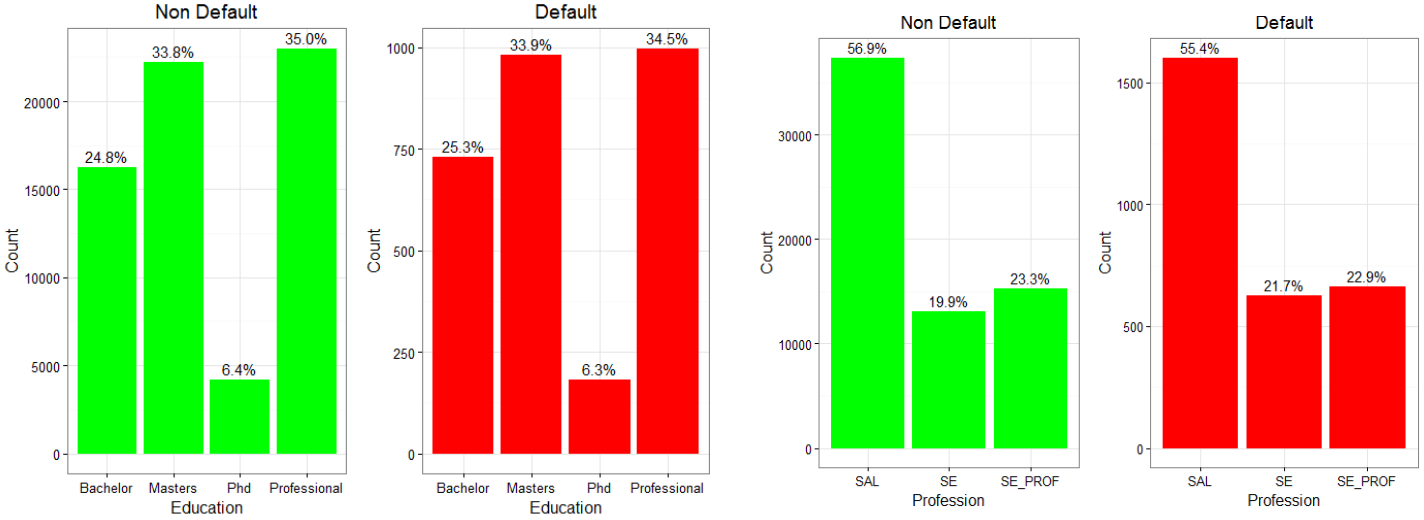




- There are more people with relatively less avg CC utilization
- There are large number of customers with lesser trades done
- Income portfolio seems to be reasonably equally distributed
- No of dependents showing no significant inclination towards having credit card.

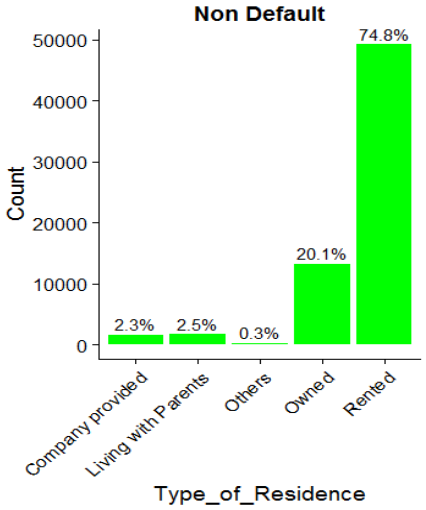
Demographic Data Vs Defaulters

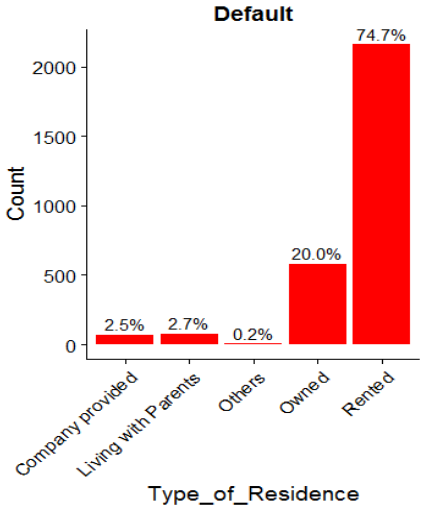


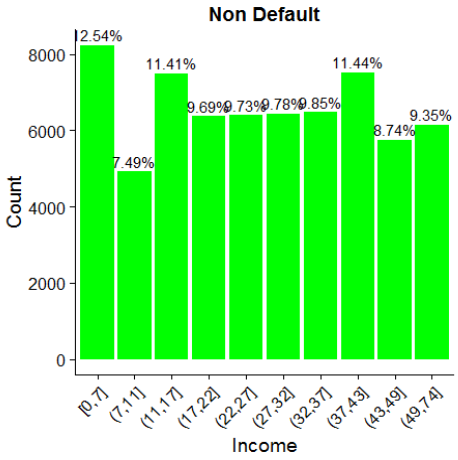


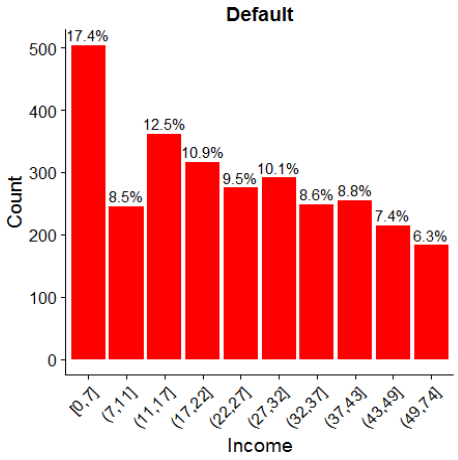
- More or less equal distribution of Gender across default & non-default. Hence no major influence of gender on performance.
- Similarly Marital Status has no significant impact.

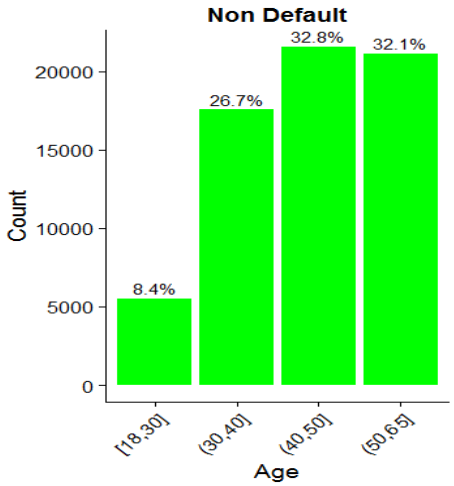
Demographic Data Vs Defaulters

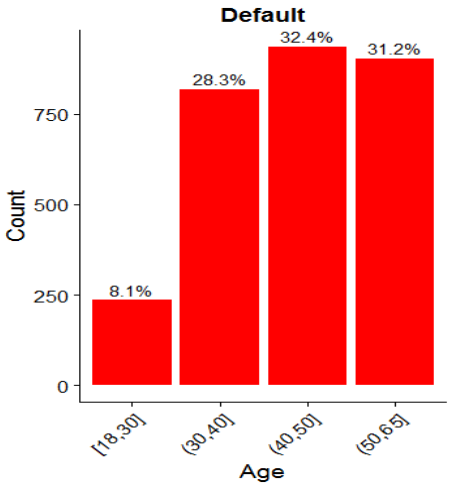






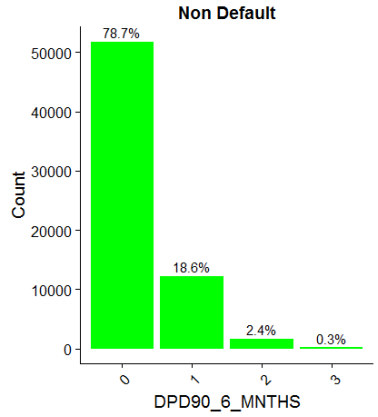


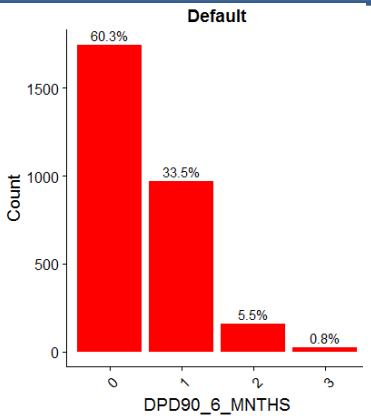


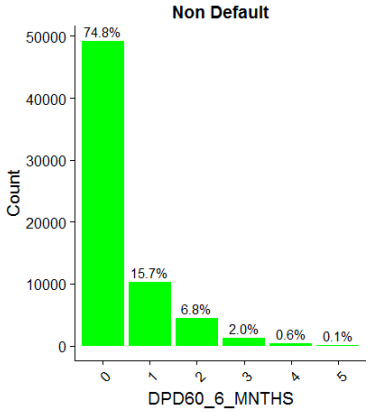


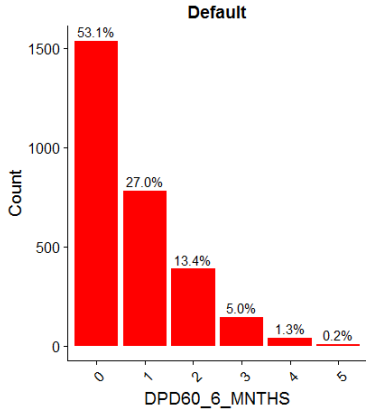
- **Low income group** are more likely to default
- There is **less impact** of income buckets on non-defaulters as compared to defaulters

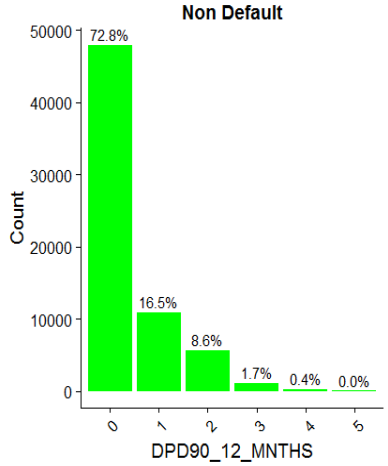
Credit Bureau Data Vs Defaulters

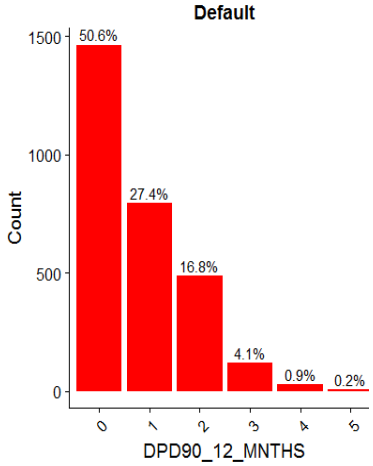


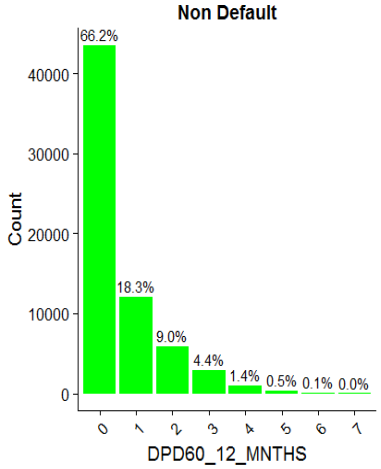


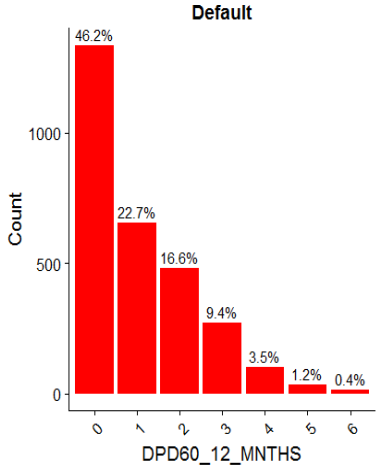






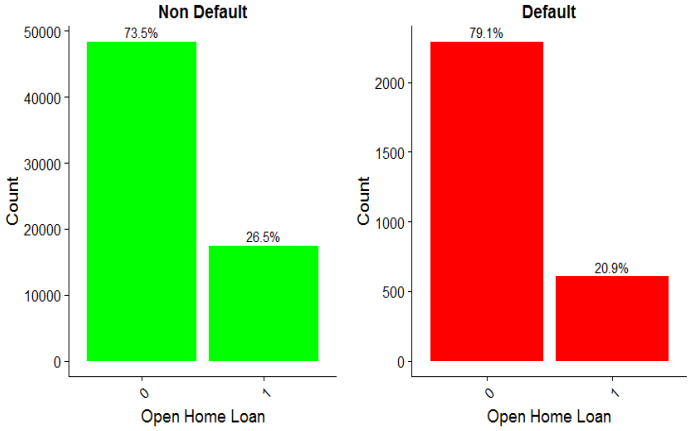


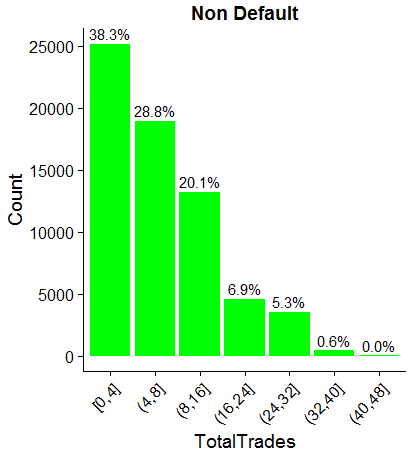


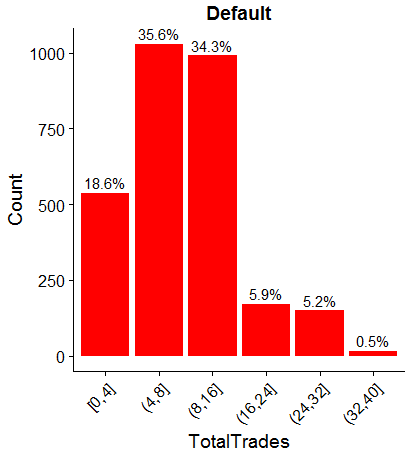


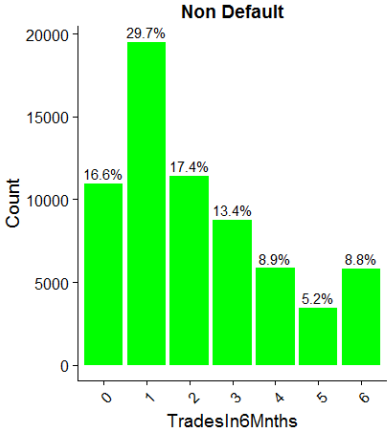
- More number of DPD90/60 in last 6/12 months resulted into more defaults

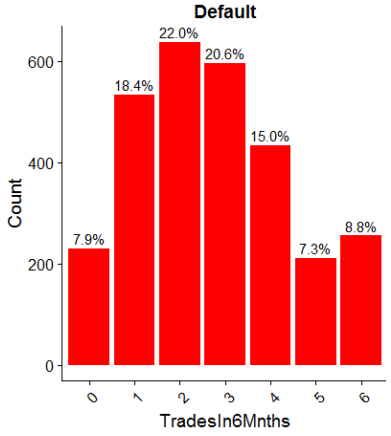
Credit Bureau Data Vs Defaulters

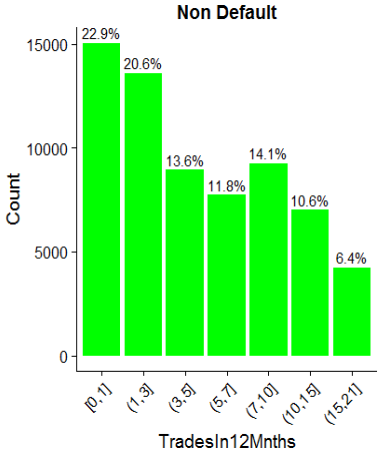


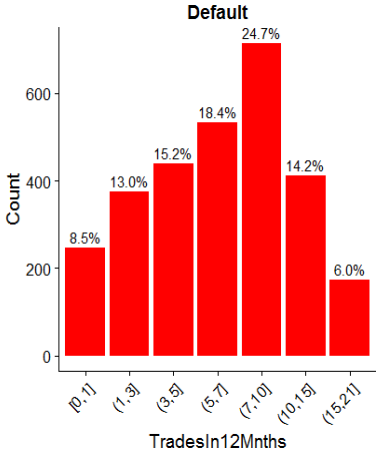






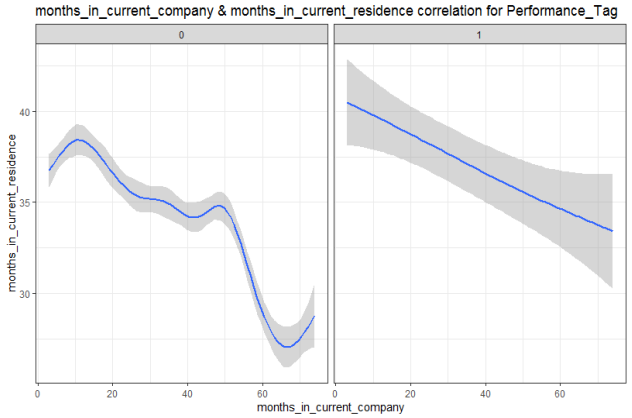


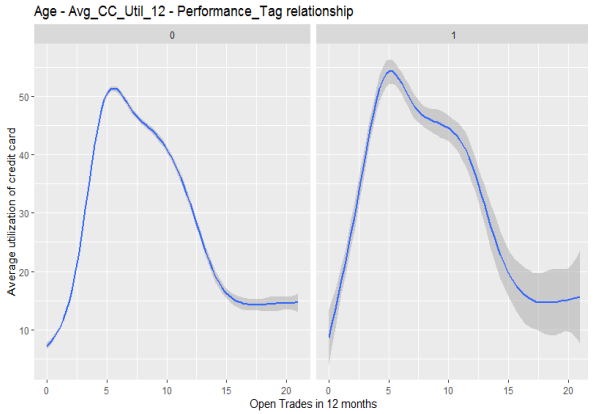


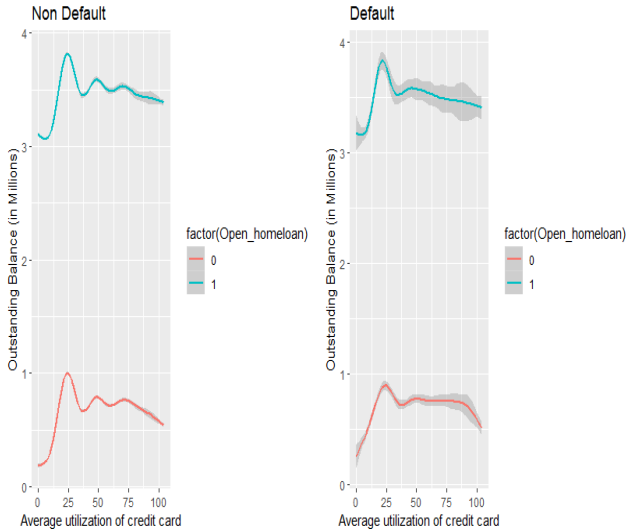


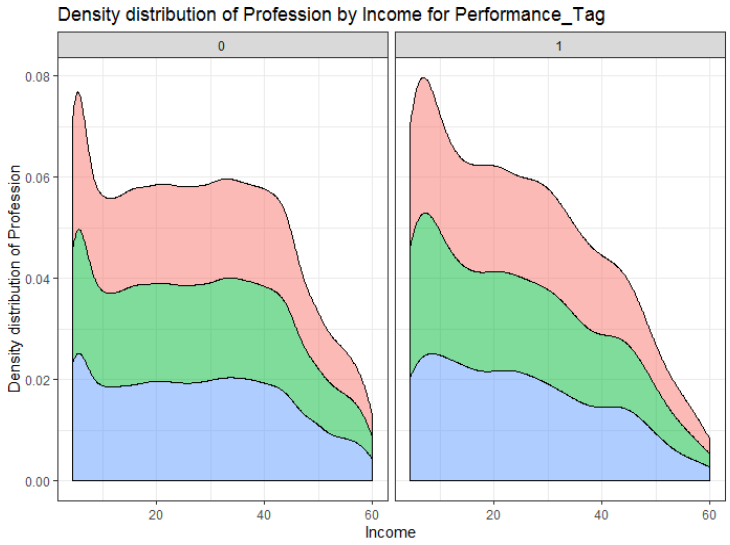
•There is an interesting finding w.r.t. TotalTrades which shows people done more trades in recent time are more likely to default comparatively

Multivariate Analysis against Performance Tag



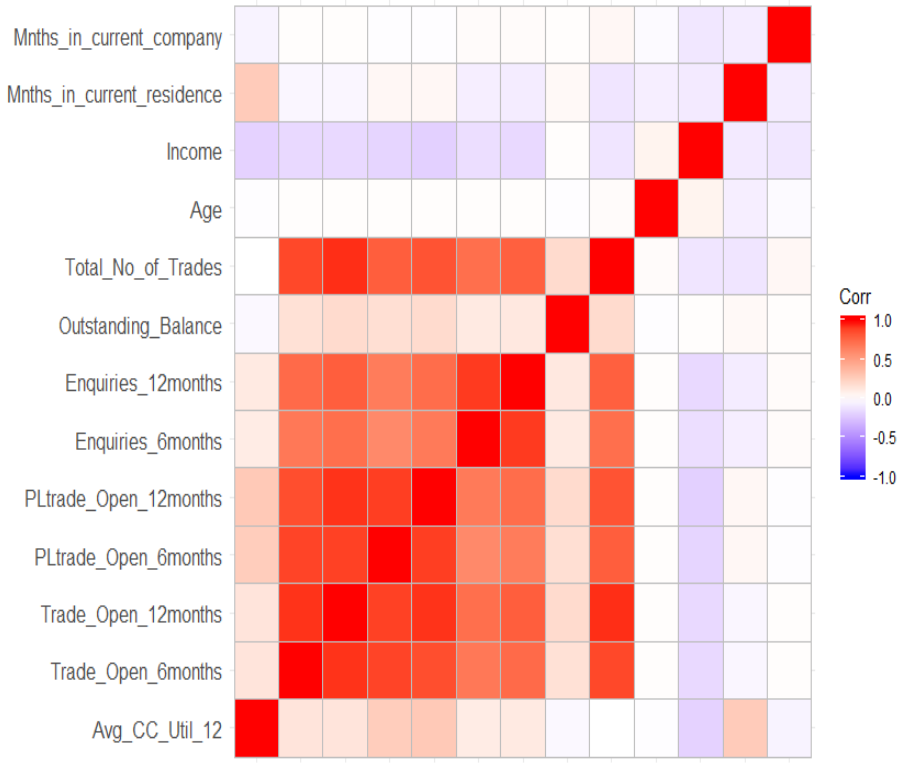




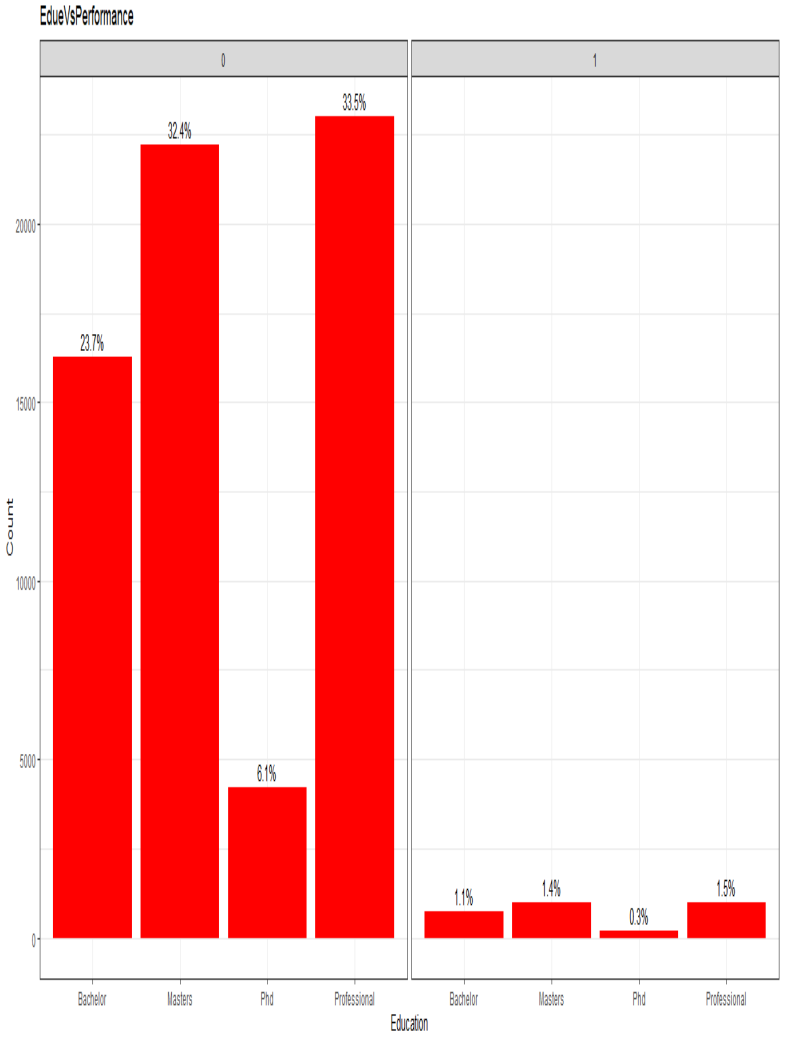


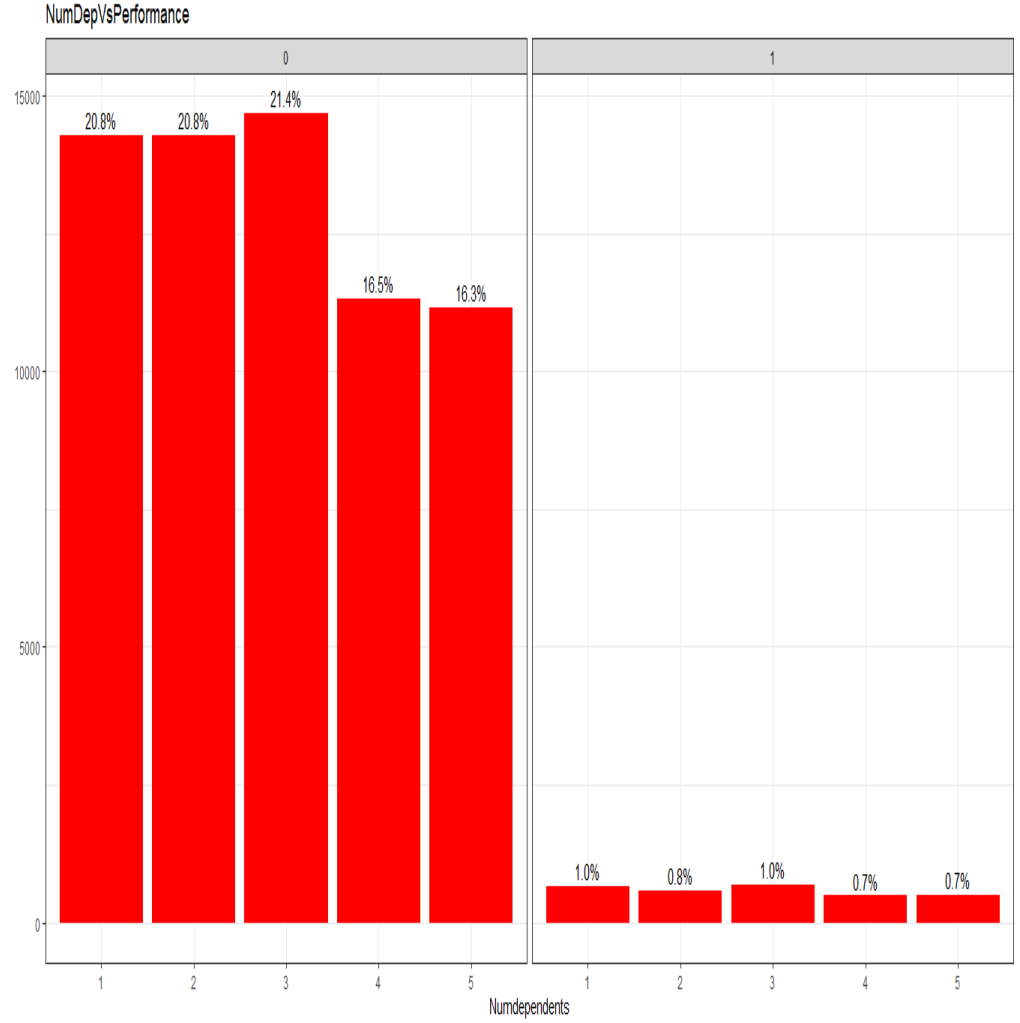
Personal

Correlation Among Continuous Variables

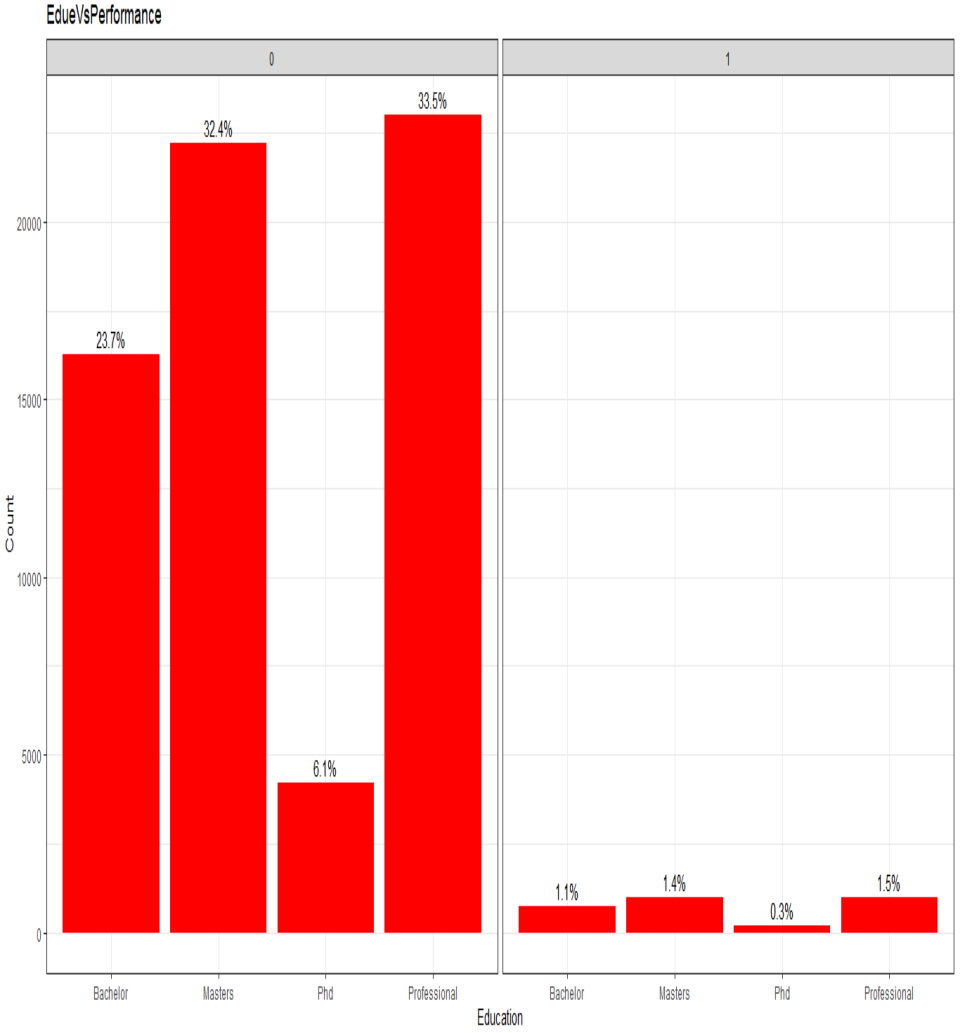


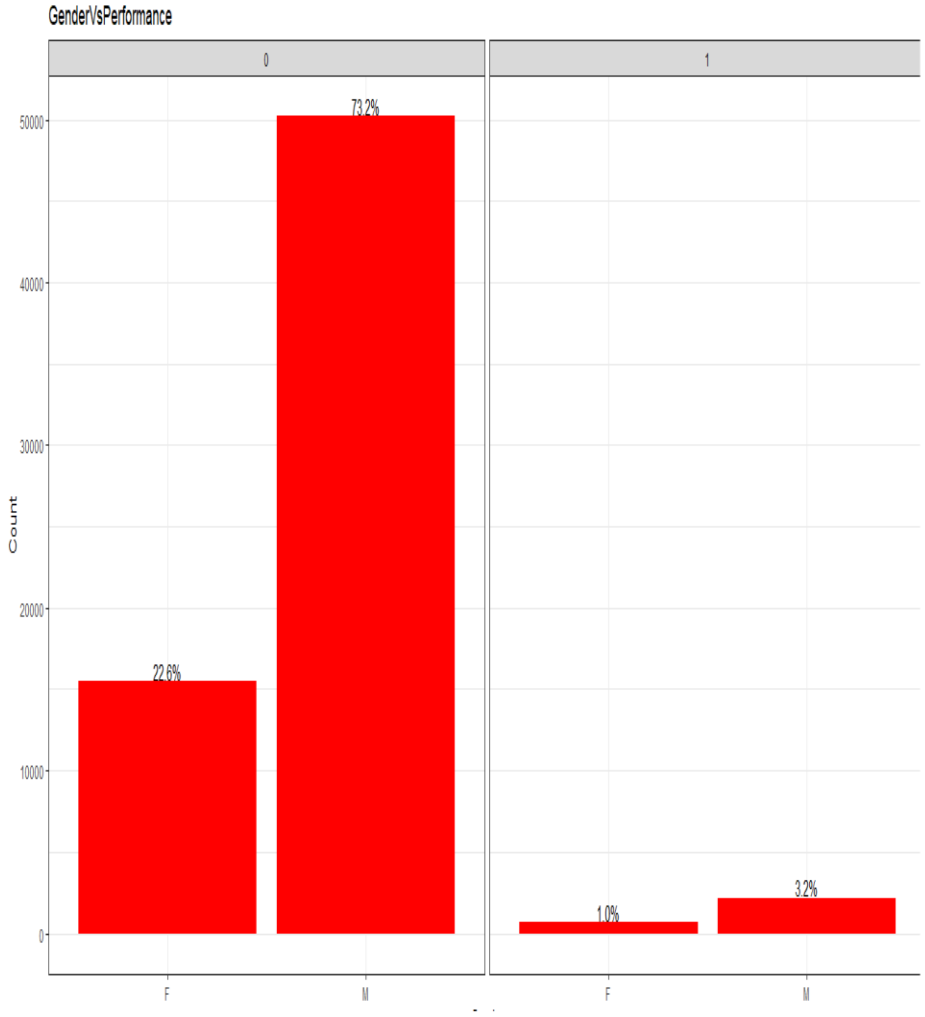
Uni-variate Analysis



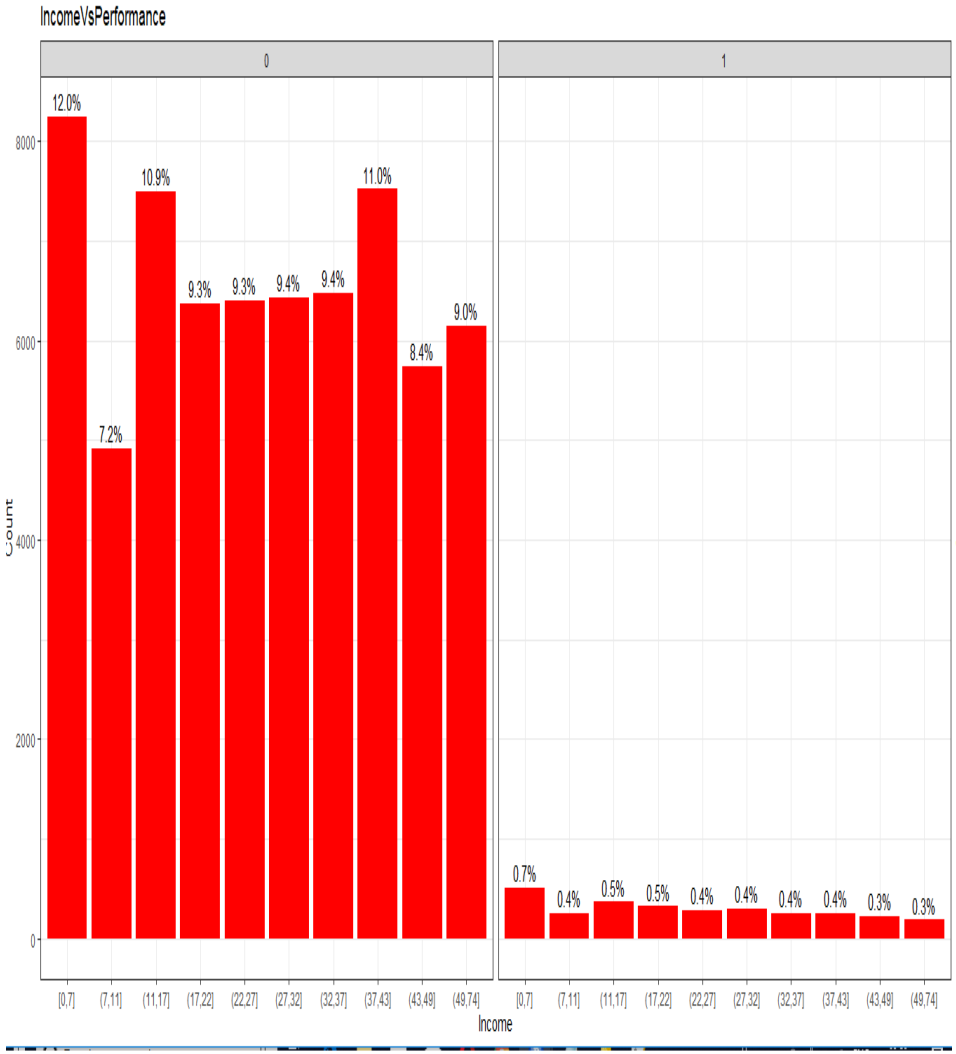


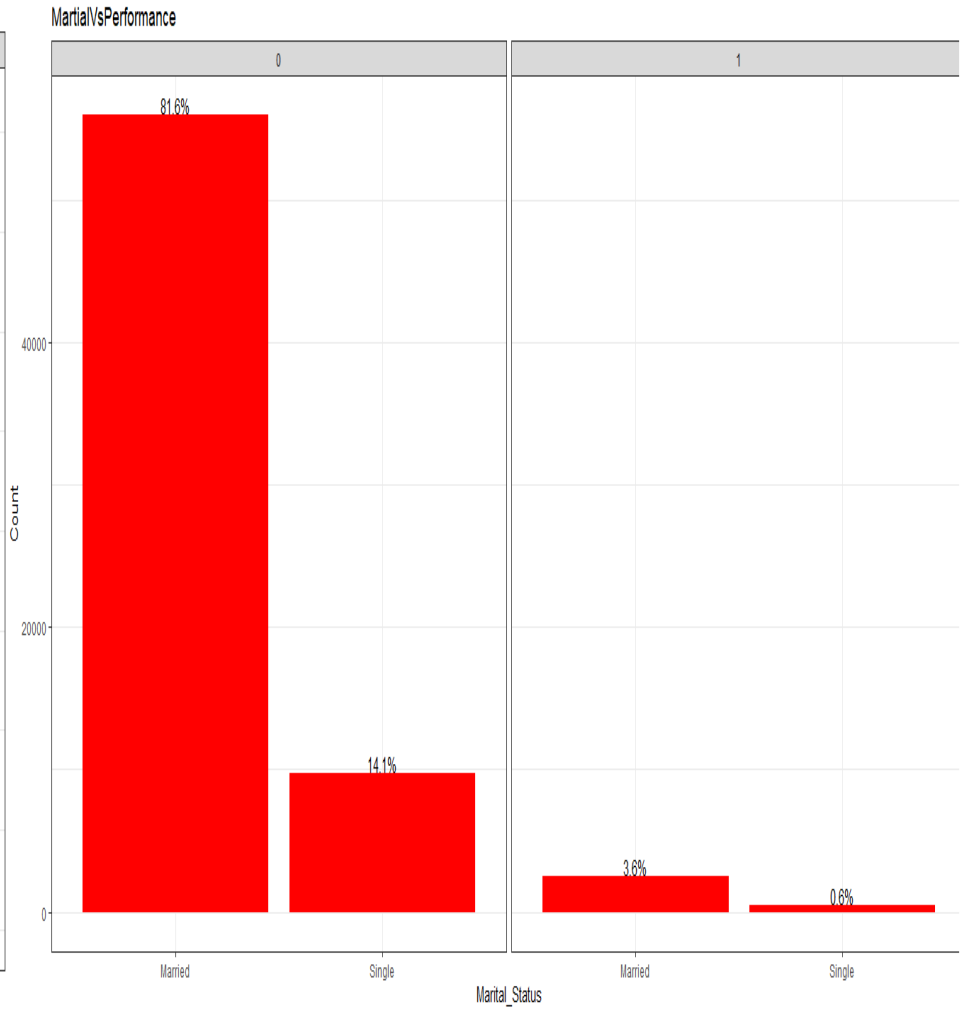
Uni-variate Analysis



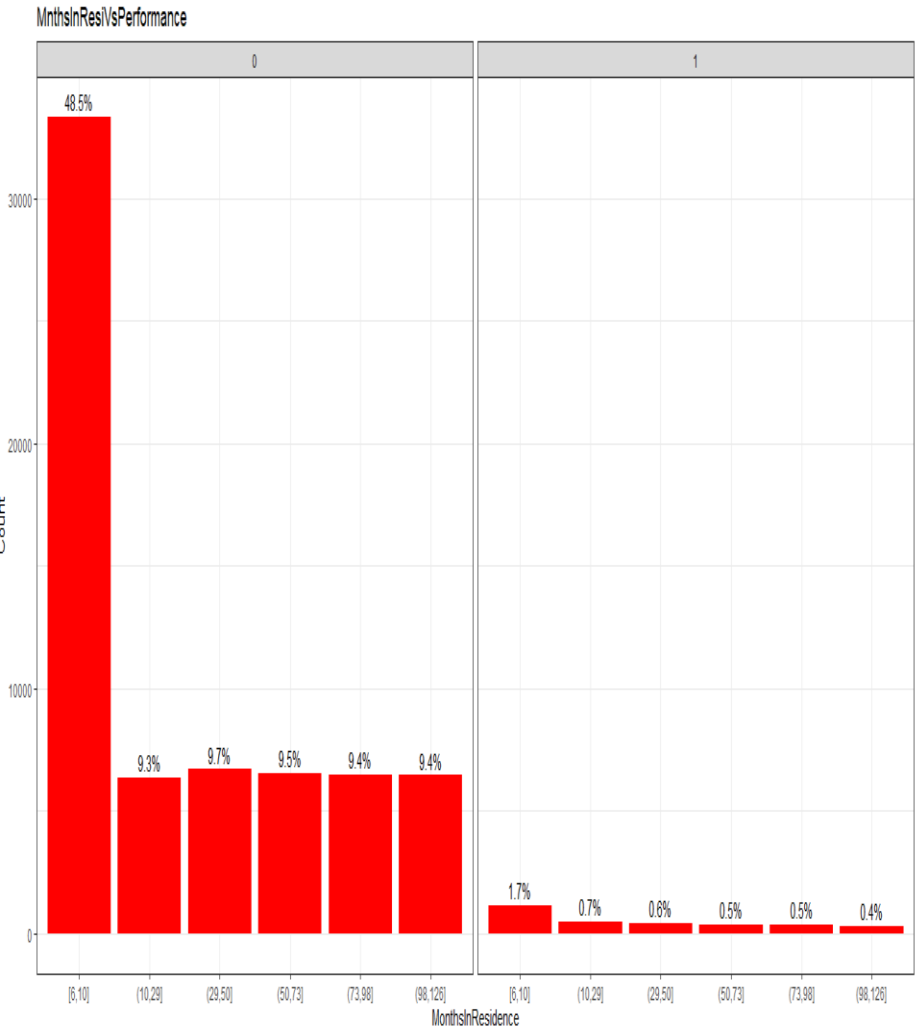


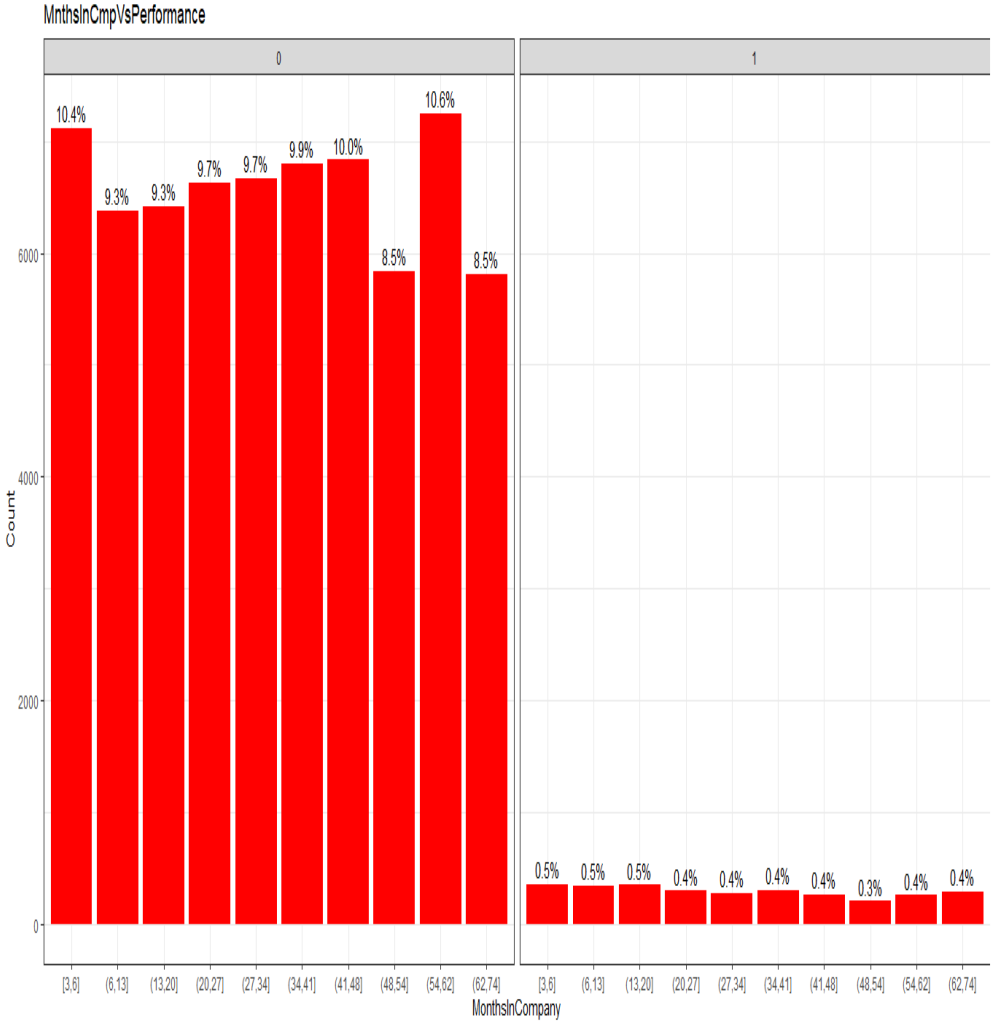
Uni-variate Analysis



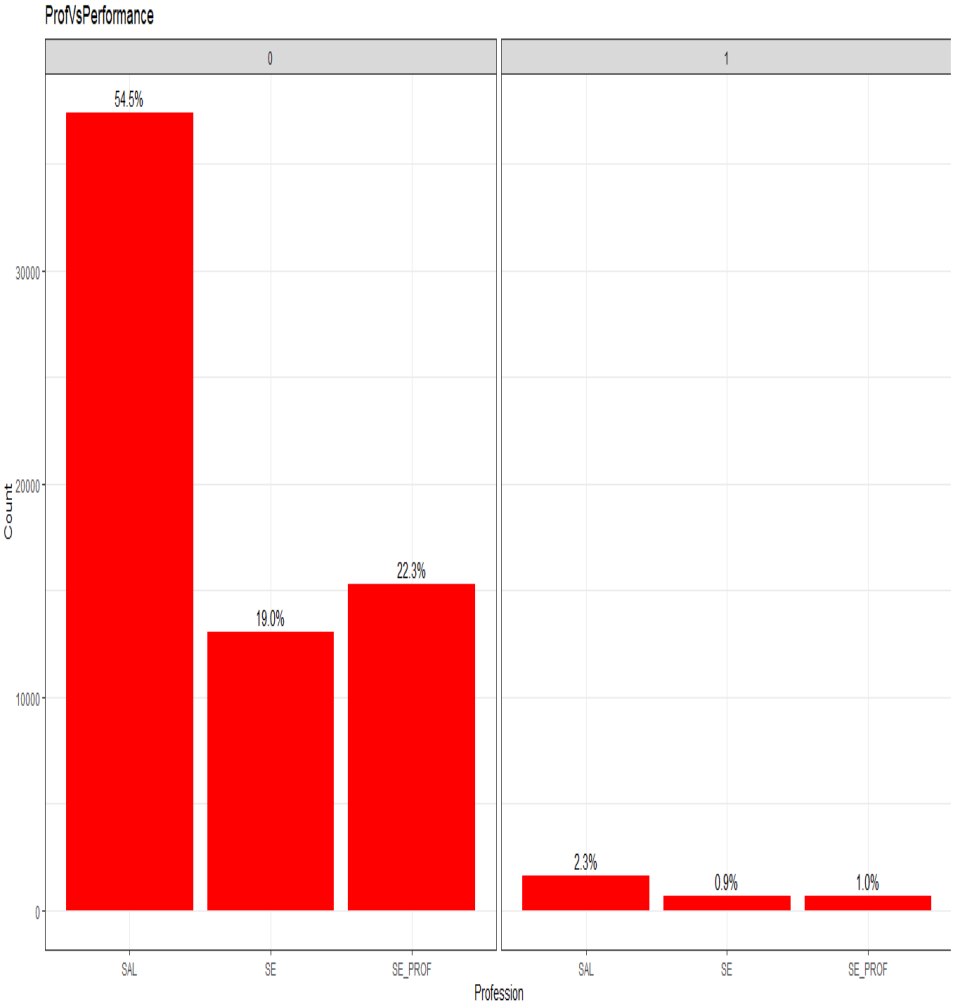


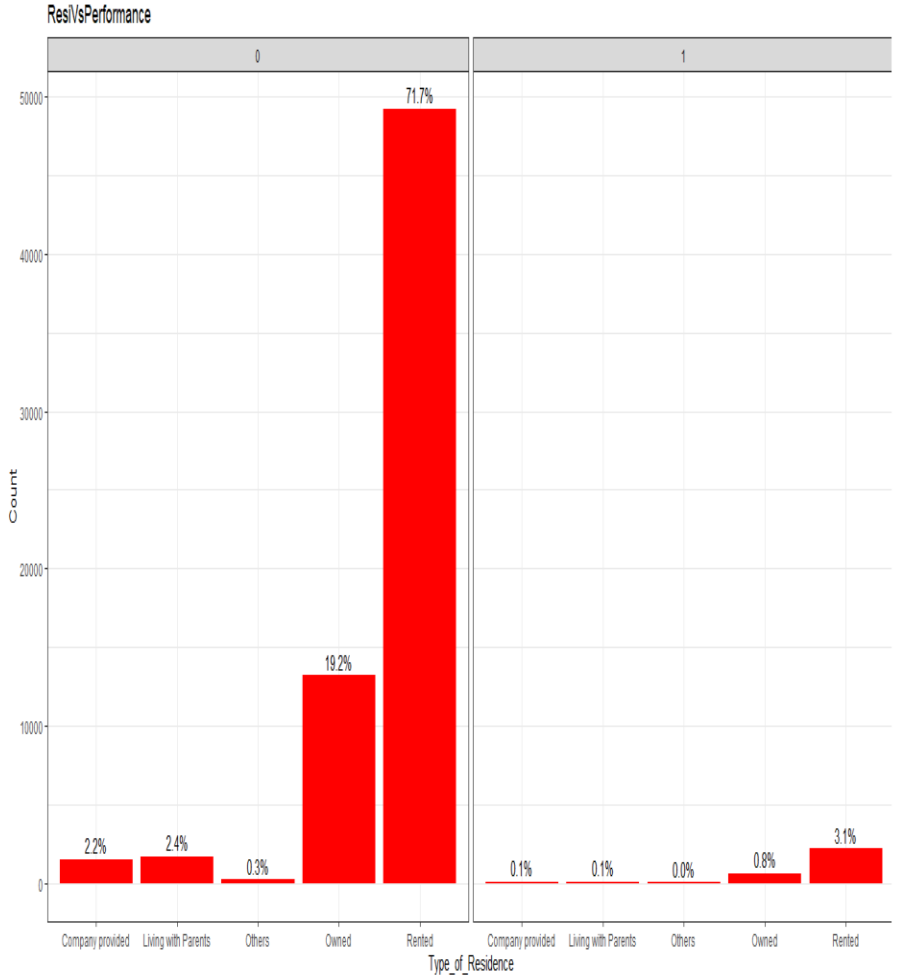
Uni-variate Analysis



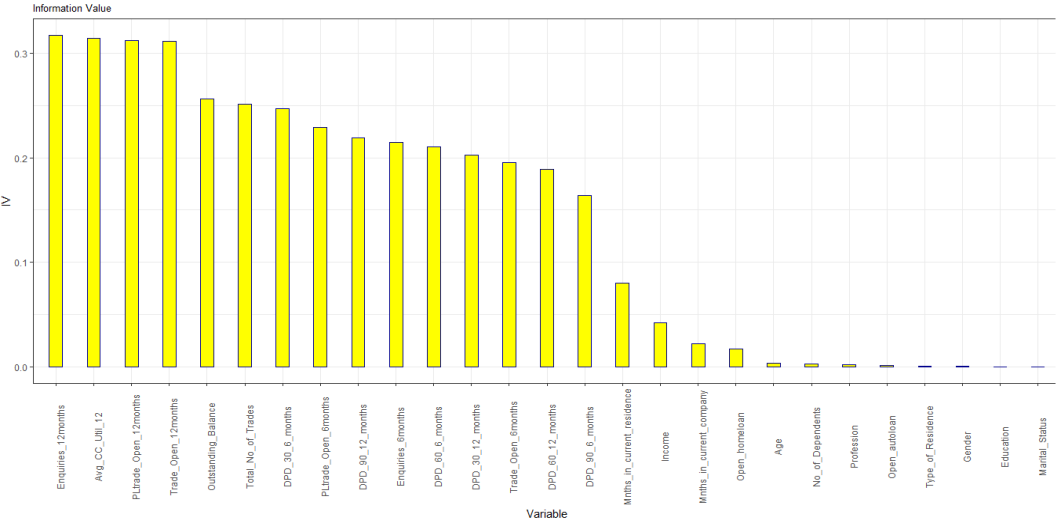


Uni-variate Analysis





Feature Selection Using WOE & IV




```
# using the IV criteria:

# < 0.02          Not useful for prediction
# 0.02 to 0.1     Weak predictive Power
# 0.1 to 0.3      Medium predictive Power
# 0.3 to 0.5      Strong predictive Power
# > 0.5           Suspicious Predictive Power
```

```
knitr::kable(head(IV_Tables$Summary))
#   |    |Variable                |        IV|
#   |:---|:-----------------------|---------:|
#   |13  |Enquiries_12months      | 0.3170205|
#   |7   |Avg_CC_Util_12          | 0.3141667|
#   |11  |PLtrade_Open_12months   | 0.3124460|
#   |9   |Trade_Open_12months     | 0.3117646|
#   |14  |Outstanding_Balance     | 0.2561862|
#   |15  |Total_No_of_Trades      | 0.2511140|
```

Variable	IV	feedback
Enquiries_12months	0.317020521	Strong
Avg_CC_Util_12	0.314166665	Strong
PLtrade_Open_12months	0.312445981	Strong
Trade_Open_12months	0.311764592	Strong
Outstanding_Balance	0.256186192	Medium
Total_No_of_Trades	0.251113955	Medium
DPD_30_6_months	0.247140305	Medium
PLtrade_Open_6months	0.228913872	Medium
DPD_90_12_months	0.218716562	Medium
Enquiries_6months	0.214752243	Medium
DPD_60_6_months	0.210422551	Medium
DPD_30_12_months	0.202860262	Medium
Trade_Open_6months	0.195074909	Medium
DPD_60_12_months	0.189306036	Medium
DPD_90_6_months	0.163911992	Medium
Mnths_in_current_residence	0.079897684	Weak
Income	0.042098313	Weak
Mnths_in_current_company	0.02210909	Weak
Open_homeloan	0.017331704	Useless
Age	0.003629963	Useless
No_of_Dependents	0.002885618	Useless
Profession	0.002075011	Useless
Open_autoloan	0.001581178	Useless
Type_of_Residence	0.000893177	Useless
Gender	0.000267128	Useless
Education	0.000196672	Useless
Marital_Status	9.45E-05	Useless

Model Building, Evaluation and Validation

1. Logistic Regression Model
   - Train & Test datasets prepared
   - Firstly logistic regression model (glm) is used to identify the important variables
   - StepAIC is used to form the initial model with important variables
   - Final model is prepared by eliminating non-significant variables by means of VIF & p-values
   - This gives a good view as a baseline model of important variables in acquiring new customers in terms of default performance. Important variables found like Avg_CC_Util_12, Trade_Open_12months, Outstanding_Balance, DPD_30_6_months etc.

   Model Evaluation:
   - Predictions performed on test dataset
   - Confusion matrix calculated to check Accuracy, Sensitivity & Specificity
   - ROC Curve AUC and KS Statistic – calculated to further verify the model

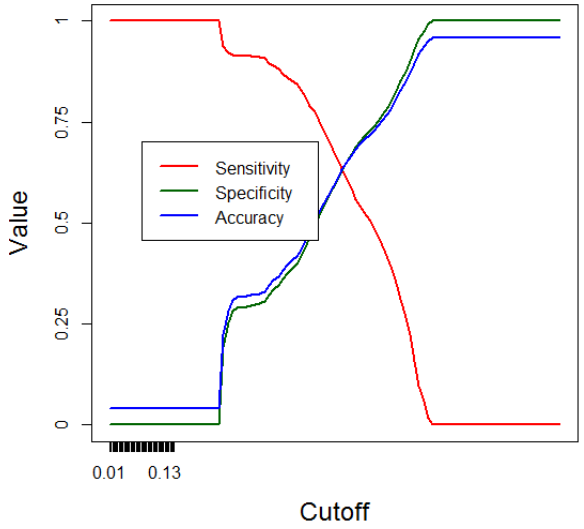
2. Next we tried **Random Forest** and compared their performance with the logistic regression model. There is little significant improvement in their performance.

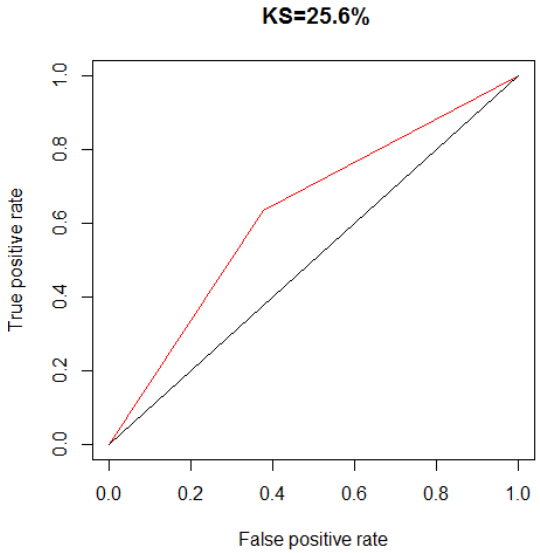
3. Random Forest model & logistic models then applied onto balanced dataset and opted to further check the important variables affecting the performance of credit card holders

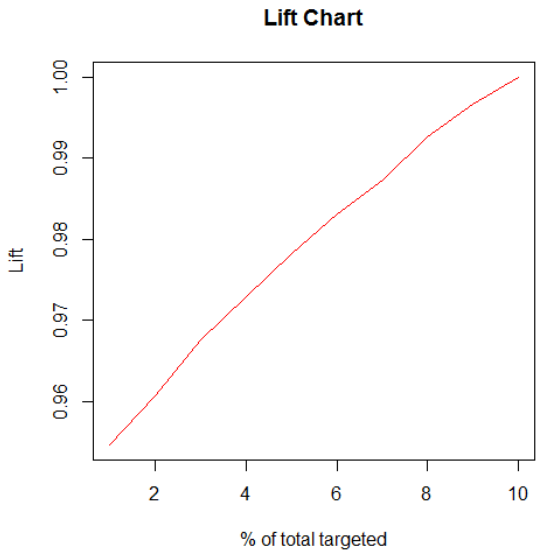
4. At the end, a final model selected based upon comparison of all the models in terms of the following - Accuracy, Sensitivity, Specificity, ROC AUC, Cross Validation, Classification Error, Confusion matrix,F1 Score, KS-Statistic, and Computational cost and Real time deployments. And choose the model that gives a right balance of achieving the goals and simplicity in nature.

Model Building, Evaluation and Validation

Logistic model







Cutoff value of 0.04275 is selected with below results:

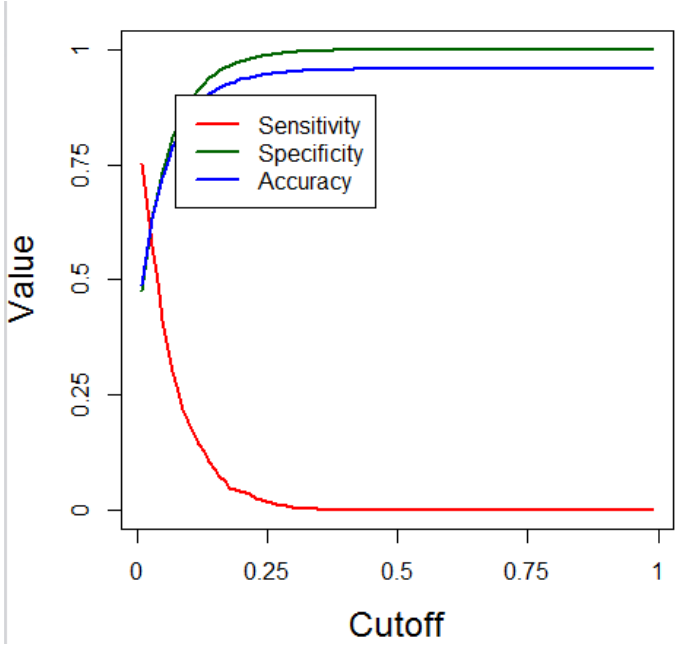
Accuracy : 0.6146375
Sensitivity : 0.64286
Specificity : 0.61340

#KS Statistics for the Test data
KS : 0.2562527 AUC : 0.6281263

```
#-----------------------------------GLM MODEL PARAMETERS FOR  IMBALANCED DATA-------------
#1. Accuracy                                                                    63.3
#2. Sensitivity (True positive rate)                                            61.8
#3. Specificity (True negative rate)                                            63.363
#4. Precision/Specificity: how many selected instances are relevant.            0.042
#5. Recall/Sensitivity: how many relevant instances are selected.               1.000
#6. F1 score: harmonic mean of precision and recall.                           0.040
#7. AUC: relation between true-positive rate and false positive rate.          0.6261447
#8. KS-Static Paramaer                                                         0.2522893
```

Model Building & Evaluation

Random Forest Model




```
#-------------------------------------RF MODEL PARAMETERS FOR  IMBALANCED DATA----------------

#1. Accuracy                                                          61.68
#2. Sensitivity (True positive rate)                                  59.908
#3. Specificity (True negative rate)                                  61.760
#4. Precision/Specificity: how many selected instances are relevant.   0.042
#5. Recall/Sensitivity: how many relevant instances are selected.      1.000
#6. F1 score: harmonic mean of precision and recall.                   0.040
#7. AUC: relation between true-positive rate and false positive rate.  0.608
```

Model Building & Evaluation– Balanced Data

Since given data is unbalanced so model building to be tried on balanced dataset.
The ROSE (Random Over Sampling Examples) package is used to generate artificial data based on sampling methods and smoothed bootstrap approach.

Oversampling
This method over instructs the algorithm to perform oversampling. As the original dataset had 2947 good observations, this method is used to oversample minority class until it reaches 69864. The dataset has a total of 69864 records. This can be attained using method = "over".

Under Sampling
This method functions similar to the oversampling method and is done without replacement. In this method, good transactions are equal to fraud transactions. Hence, no significant information can be obtained from this sample. This can be attained using method = "under".

Both Sampling
This method is a combination of both oversampling and under-sampling methods. Using this method, the majority class is under-sampled without replacement and the minority class is oversampled with replacement. This can be attained using method = "both".

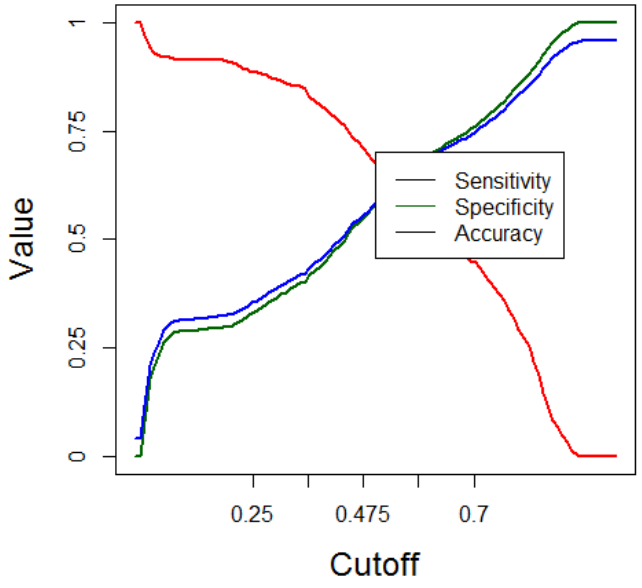
ROSE Sampling
ROSE sampling method generates data synthetically and provides a better estimate of original data.

Synthetic Minority Over-Sampling Technique (SMOTE) Sampling
This method is used to avoid over fitting when adding exact replicas of minority instances to the main dataset.
Handling Unbalanced data

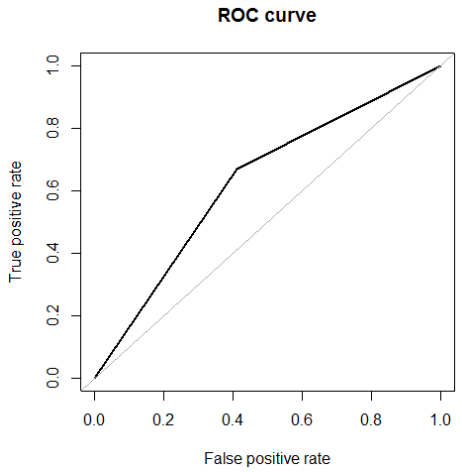
Model Building & Evaluation– Balanced Data

Logistic Regression – ROSE method

```
glm(formula = Performance_Tag ~ DPD_30_6_months + DPD_90_12_months +
    DPD_30_12_months + Avg_CC_Util_12 + Trade_Open_12months +
    PLtrade_Open_6months + PLtrade_Open_12months + Enquiries_6months +
    Enquiries_12months, family = "binomial", data = rose_sampling_data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.795  -1.108  -0.645   1.047   1.886
```



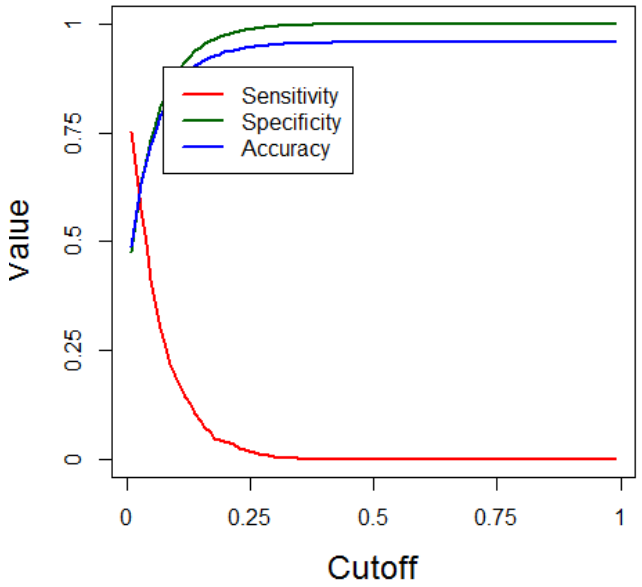

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 12276   318
         1  7447   550

               Accuracy : 0.6229
                 95% CI : (0.6162, 0.6295)
    No Information Rate : 0.9578
    P-Value [Acc > NIR] : 1

                  Kappa : 0.052
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.63364
            Specificity : 0.62242
         Pos Pred Value : 0.06878
         Neg Pred Value : 0.97475
             Prevalence : 0.04215
         Detection Rate : 0.02671
   Detection Prevalence : 0.38837
      Balanced Accuracy : 0.62803

       'Positive' Class : 1
```

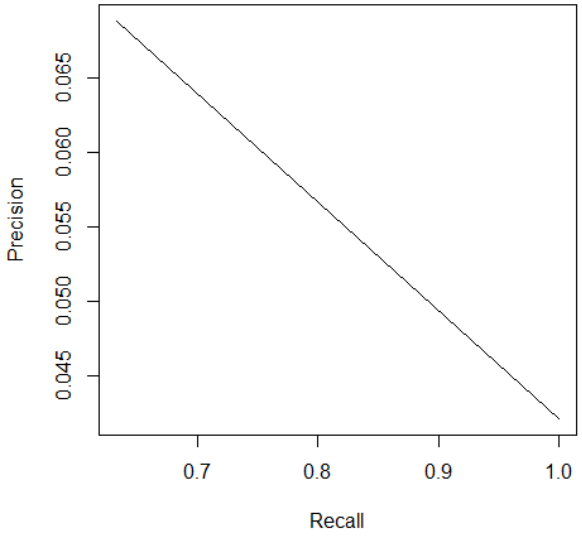


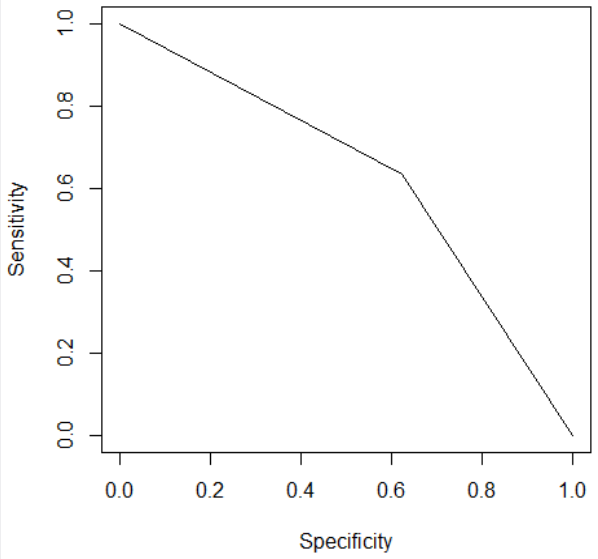
#----------GLM MODEL PARAMETERS FOR ROSE BALANCED DATA---------	
#1. Accuracy	62.29
#2. Sensitivity (True positive rate)	63.364
#3. Specificity (True negative rate)	62.242
#4. Precision/Specificity: how many selected instances are relevant.	0.042
#5. Recall/Sensitivity: how many relevant instances are selected.	1.000
#6. F1 score: harmonic mean of precision and recall.	0.040
#7. AUC: relation between true-positive rate and false positive rate.	0.628

Model Building & Evaluation – Balanced Data

Random Forest – ROSE method








```
#-------------------------RF MODEL PARAMETERS FOR ROSE BALANCED DATA----------------------
#1. Accuracy                                                    0.7178
#2. Sensitivity (True positive rate)                            0.7205
#3. Specificity (True negative rate)                            0.7205
#4. Precision/Specificity: how many selected instances are relevant.    0.496
#5. Recall/Sensitivity: how many relevant instances are selected.       1.000
#6. F1 score: harmonic mean of precision and recall.            0.332
#7. AUC: relation between true-positive rate and false positive rate.   0.718
#8. KS- Static                                                  0.4356921

# CONCLUSION : RF Model with ROSE BALANCED DATA Has given better Model Evaluation Parameter
```

Application Scorecard/ Financial Benefits (Predicted)

Credit scoring is a **key risk assessment technique** to analyse and quantify a potential obligor's credit risk. Essentially, credit scoring aims **at quantifying the likelihood** that an obligor will repay the debt. The outcome of the credit scoring exercise is a score reflecting the **creditworthiness** of the obligor

Application Score Card :-

- ❑An Application scorecard is built with the good to bad odds of 10 to 1 at a score of 400 doubling every 20 points. 10 will be highest and 1 will be lowest.

- ❑Application score is associated with application Id and bank can decide good and bad applicants considering their scorecard.

- ❑The Application Score Cutoff is identified which will act as a threshold, below which the bank should not grant of Credit Cards to the applicants

Financial benefit :-

- ❑ Bank can decide good and bad customer using Application score card and find the Potential Customers/ reject the Customers who might Default.

- ❑ This exercise will provide significant variables and their importance to bank which will help to work on certain area for avoiding credit loss and Increase the financial benefit

Conclusion

- As part of EDA, after removing NA values and outliers, we saw valuable relationship of the variables with performance_tag – as showed & explained in above slides. E.g. more number of male, married, salaried or people with high income are using credit cards and that too more in 40-50yrs of age group. Similarly, people who have done more open trades in last given months, are more likely to default.

- By means of using WOE and deriving IV for all these variables, we could figure out strong to medium to week predictors of default/non-default performance.

- The weak/not useful predictors are mostly the demographic variables e.g. Age, Income, Profession, Gender, Marital_status, Education, No. of months in current company, Types of residence etc. are dropped in the final model as they don't play a good role in making predictions for defaulters.

- After performing all these steps, we have a cleaned data set/file having all the good predictor fields with no missing values and having their actual values converted to WOE values, which will help us to build our model(s) successfully.

- After **Logistic, ensemble techniques** are applied on the variables arrived by logistic regression, the models with **imbalanced and balanced** data set are evaluated compared and final model is decided. **Random forest with ROSE Sampling data** gave better Performance Metrics.

- Finally we built the **application score card** with application Id so that bank can decide good and bad applicants considering their scorecard. It could find out a application score cut off of 332 with approx. 63% of accuracy, specificity & sensitivity and could find out 50% of applications below cut off in the balanced data for applicants. Similarly, could suggest 98% of the default in the rejected applications.