Approach Document - BFS Capstone Project

Yogesh Kulkarni, Maxim Rohit

PGDDS -March 2018

Contents

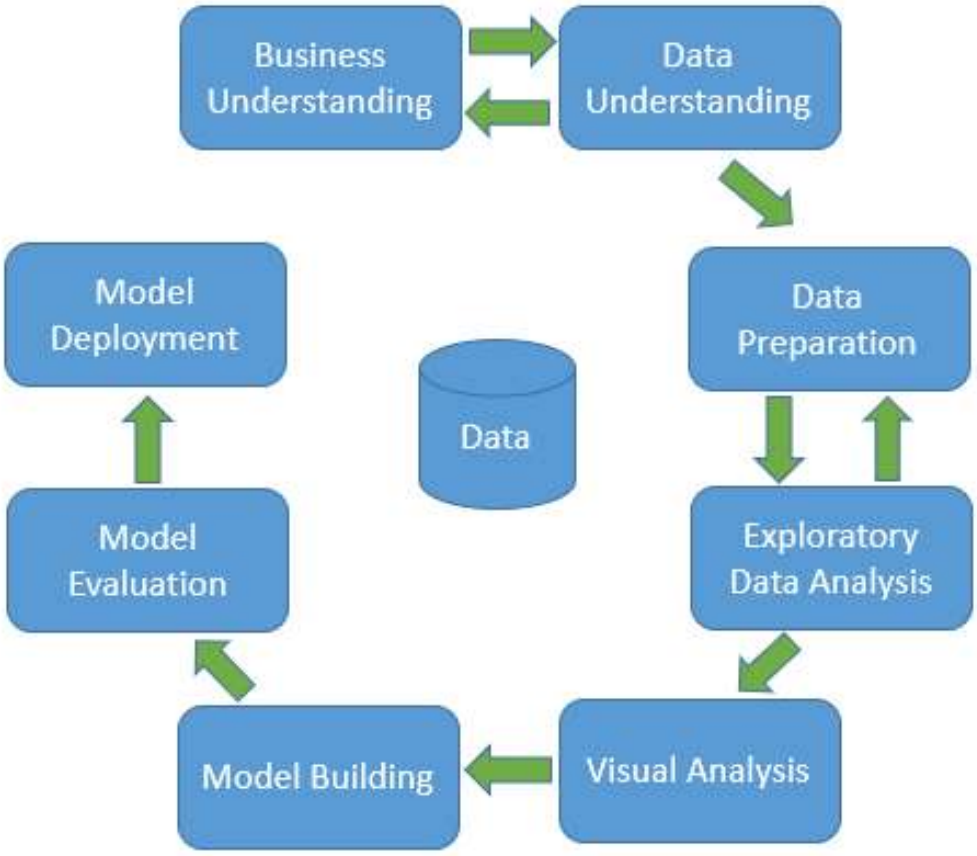
Problem Statement

CredX is a leading credit card provider that gets thousands of credit card applicants every year. However, it has experienced an increase in credit loss in the past few years.

The best strategy to mitigate credit risk is to '***acquire the right customers***'. We will help CredX identify the right customers using predictive models to minimize credit losses.

As part of this project, we are required to determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and highlight the financial benefit of this project.

Problem Solving Methodology



Business Understanding

CredX is a leading credit card provider has experienced an increase in credit loss due to customer defaults in the past few years. The best strategy to mitigate credit risk is to acquire right customers. We will help CredX identify the right customers using predictive models to minimize credit loss. Using past data of the bank's applicants, we will determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and provide assessment of the financial benefits.

For this, we are required to develop a model that can be used for approvals and minimize credit loss by rejecting customers that would default, maximize approvals of customer who will not default thus improving revenue for CrediX and minimizing loss. We are required to build application scorecard that can provide a cutoff below which credit cards should not be approved.

Data Understanding

- **Demographic Data**: This data is obtained from the information provided by the applicants at the time of credit card application. It contains customer-level information on age, gender, income, marital status, etc. Some of this information could be sensitive and incomplete.
- **Credit Bureau Data**: This is taken from the credit bureau and contains variables such as 'number of times 30 DPD or worse in last 3/6/12 months', 'outstanding balance', 'number of trades', etc. Such data obtained from credit bureau is usually paid and can indicate credit discipline as well as credit worthiness of individuals.

Data Preparation

Data loading

We loaded the data separately to observe both data sets and perform analysis prior to merge. As part of data load, we replaced empty values with NA. Following observations were noted and actions were taken.

Data Observations and prep

- Changed data frame field names to intuitive and short names.
- **Demographic data** contained 71295 records and 12 variables
- **Credit Bureau data** contained 71295 observations and 19 variables
- Both datasets contain same application id and performance – there's no conflict. We merged both datasets to get a complete view of applicant data. We created isolated data frames from this in further EDA & model building process as needed.
- It is observed that there're ~4.13% subjects in the data who have defaulted. Since this number is low from model building perspective, it was noted that we will probably need to balance/boost the data for better accuracy.
- 'Performance' has ~ 2%(1425) blank values i.e. applicants rejected for credit card. We saved such data in another data frame for later use – to check our model against and excluded it from model building process.

- Since we need to predict customers to whom credit should be provided, we **reversed performance indicator** for automated model deployment. i.e: 1 will be good customer and 0 will be defaulter
- There were 3 duplicate records. These duplicates were analysed and removed
- Credit bureau data contains 565 subjects who do not have any information other than performance. Of these, 30 customers have defaulted, and rest other have not defaulted. Since this can be valid scenario, i.e. good customers may not have any open loan credit lines, trades or DPD we are assuming that we can keep this data as is.
- **Missing values observed in 'Gender',' Education',' Profession' 'Marital Status', No of dependents', 'Type of residence' and 'Outstanding Balance' in Demographic Data**
- **Missing values observed in 'Avg CC Utilization', 'No of trades opened in last 6 months', 'Presence of open home loan' and 'Outstanding Balance' in Credit Bureau Data**
- Summary also showed that age and income had negative/invalid values which were imputed as explained in EDA.
- As per info given in problem statement, cases where Avg CC utilization is missing are the cases in which the applicant does not have any other credit card. We imputed such empty values first with 0 and then with imputation methods (to evaluate impact on model).
- Other missing & Invalid values were analysed & corrected as part of EDA for each variable
- **Note that for building logistic regression model, separate frames with original columns replaced by WoE values were used. For building other models MICE imputations were performed. Sample code snippet:**

WoE column imputation for regression model:

```
#WOE imputations
cat_bucket<-WOETable(X=as.factor(demo_df_woe$Education), Y=demo_df_woe$performance)
cat_bucket
names(cat_bucket)
demo_df_woe<-merge(demo_df_woe,cat_bucket[,c("CAT","WOE")],by.x="Education",by.y="CAT")
length(names(demo_df_woe))
names(demo_df_woe)[length(names(demo_df_woe))]<-"Education_WOE"

#Since we have WOE column, we can drop original column
demo_df_woe$"Education"<-NULL
```

MICE Imputation for other models

```
mice_demo<- mice(demo_df, m=1, maxit = 50, method = 'pmm', seed = 500)
summary(demo_df)
mice_demo_df <- complete(mice_demo)

sapply(mice_demo_df, function(x) sum(is.na(x)))
```

Exploratory Data Analysis - Approach

We first performed IV/WoE analysis to understand important variables that could be further analysed in detail. Later on, we performed analysis of each variable in demographic data and credit bureau data. While performing EDA, separate data frames with WoE values for columns with missing values was created which were further used for building logistic regression models.

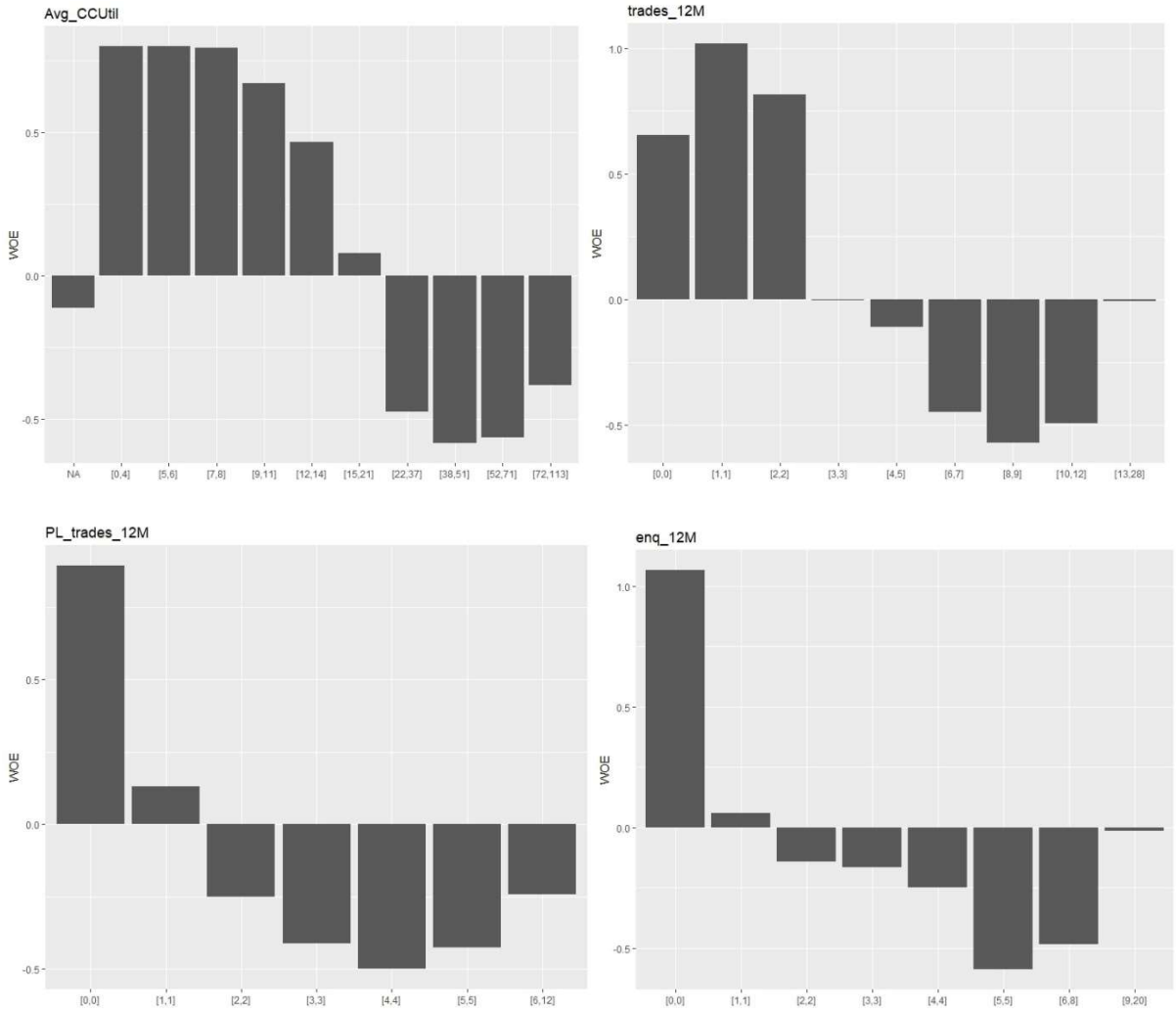
Exploratory Data Analysis – IV/WoE Analysis

- As part of EDA, First IV/WoE analysis was performed. IV analysis shows following variables as important/relevant.

```
#Based on IV following are some important variables in decreasing order of IV
#There may be multicollinarity that will need to be checked for amongst these vars as part of model building
#It's clear that demographic variables are weak predictors and hence building model only on demographic data may not yield
#good model

# Variable           IV

# 0.3 to 0.5  Strong predictor

# 17        Avg_CCUtil 3.099364e-01
# 19       trades_12M 2.979571e-01
# 21    PL_trades_12M 2.958955e-01
# 23           enq_12M 2.954243e-01

# 0.1 to 0.3  Medium predictor

# 25         outst_bal 2.462692e-01
# 13          DPD30_6M 2.415627e-01
# 26      total_trades 2.366049e-01
# 20      PL_trades_6M 2.197050e-01
# 14         DPD90_12M 2.138748e-01
# 12          DPD60_6M 2.058339e-01
# 22            enq_6M 2.051870e-01
# 16         DPD30_12M 1.982549e-01
# 18         trades_6M 1.860089e-01
# 15         DPD60_12M 1.854989e-01
# 11          DPD90_6M 1.601169e-01

# 0.02 to 0.1 Weak predictor

# 9    current.res.dur 7.894353e-02
# 5             Income 4.241780e-02
# 10  current.empl.dur 2.175441e-0
```

- WoE plots indicated below observations for above important variables. Since we reversed performance indicator, negative WoE values indicate higher likelihood of default:
- WoE was observed to be -ve for Avg CC Util beyond 22% , 4 or more trades in 12M, 2 or more PL Trades in 12M, 2 or more enquiries in 12M, outstanding balances above 3.86 Lakh, 1 or more 30DPD in 6 Months, total trades above 6, PL Trades above 1 in 6M, 1 or more 90DPD in 12M, 1 or more 60DPD in 12M, 1 or more enquiries in 6M, 1 or more 30DPD in 12M, 2 or more trades in 6M, 1 or more 60DPD in 12M, 1 or more 90DPD in 6M

Sample WoE plots for strong predictors are shown below:



We further performed EDA on each variable in demographic data & credit bureau data separately to understand inferences.

Exploratory Data Analysis – Demographic Data

Summary Statistics:

```
> summary(demographic_data)
 Application.ID         Age        Gender     No.of.dependents     Income            Education
 Min.   :1.004e+05   Min.   :-3.00    :    2   Min.   :1.000    Min.   :-0.5              :  119
 1st Qu.:2.484e+08   1st Qu.:37.00   F:16837   1st Qu.:2.000    1st Qu.:14.0   Bachelor   :17697
 Median :4.976e+08   Median :45.00   M:54456   Median :3.000    Median :27.0   Masters    :23970
 Mean   :4.990e+08   Mean   :44.94             Mean   :2.865    Mean   :27.2   Others     :  121
 3rd Qu.:7.496e+08   3rd Qu.:53.00             3rd Qu.:4.000    3rd Qu.:40.0   Phd        : 4549
 Max.   :1.000e+09   Max.   :65.00             Max.   :5.000    Max.   :60.0   Professional:24839
                                               NA's   :3
   Profession             Type.of.residence No.of.months.in.current.residence No.of.months.in.current.company
         :   14                     :    8   Min.   :  6.00                    Min.   :  3.00
 SAL     :40439   Company provided   : 1630   1st Qu.:  6.00                    1st Qu.: 16.00
 SE      :14307   Living with Parents: 1818   Median : 11.00                    Median : 34.00
 SE_PROF:16535   Others             :  199   Mean   : 34.56                    Mean   : 33.96
                 Owned              :14243   3rd Qu.: 60.00                    3rd Qu.: 51.00
                 Rented             :53397   Max.   :126.00                    Max.   :133.00

 Performance.Tag  Marital.Status   performance
 Min.   :0.0000          :    6   Min.   :0.0000
 1st Qu.:0.0000   Married:60730   1st Qu.:0.0000
 Median :0.0000   Single :10559   Median :0.0000
 Mean   :0.0422                   Mean   :0.0422
 3rd Qu.:0.0000                   3rd Qu.:0.0000
 Max.   :1.0000                   Max.   :1.0000
 NA's   :1425                     NA's   :1425
```

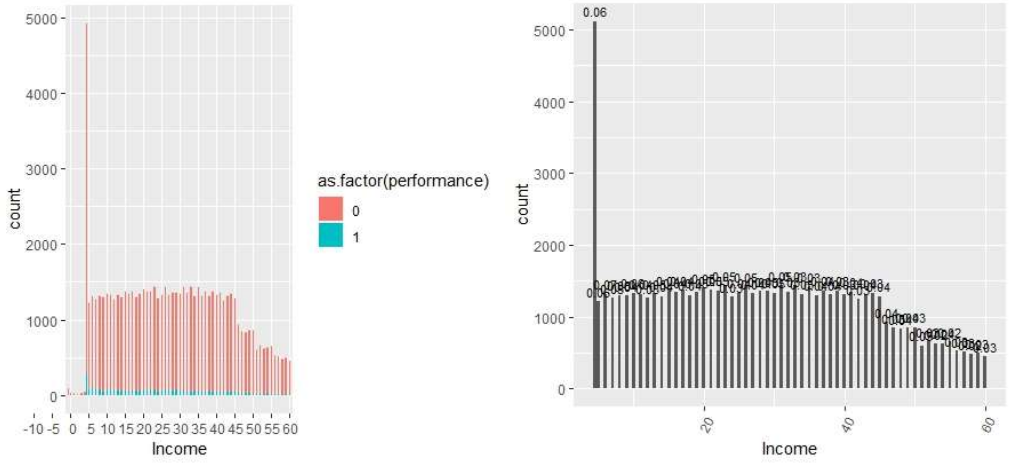
We've listed details of important variables as per IV/WoE analysis below. EDA of remaining variables is documented in addendum for reference.

Income

Income has negative/invalid values. Lower income groups have higher defaults.

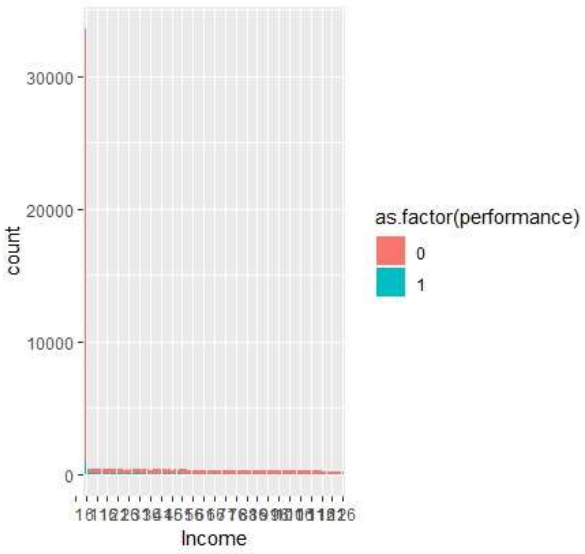
WOE analysis shows that up to 4.5, there are smaller number of records. WOE & IV values for these are also very high and there is a sudden change in post this category. Hence, we set the lower limit of income to 4.5 and imputed the lower values with 4.5.

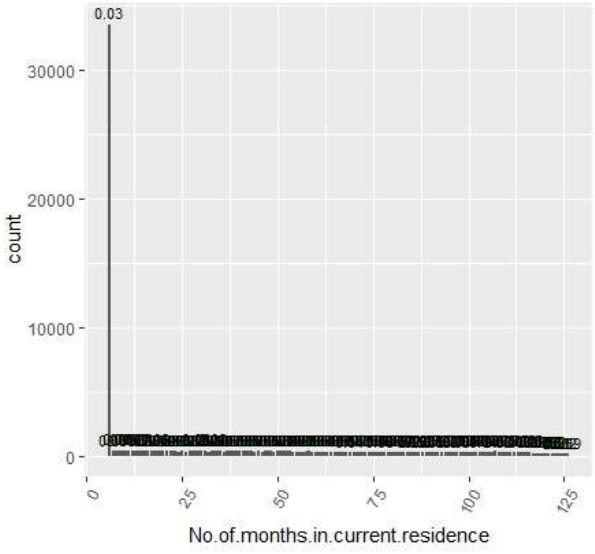
From plots, Low income group - 4.5LPA seems to have highest defaults



No. of months in current residence

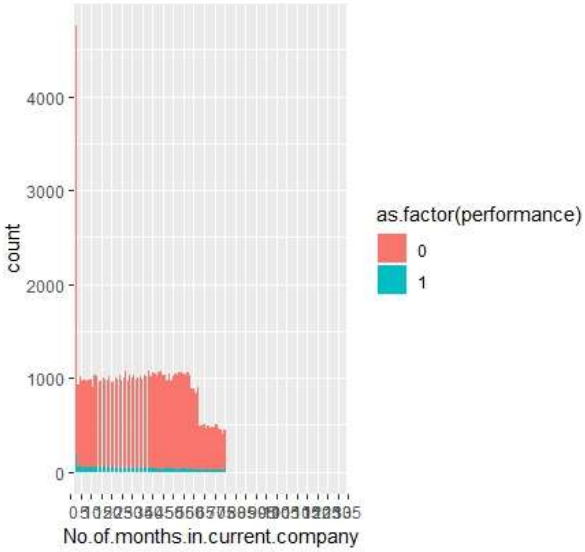
48% of subjects have stayed in current residence for 6 months. Observed a spike at value 6. Fluctuating default ratio across values. This variable was observed to be a weakly significant variable in IV/WoE analysis. However since almost half of the data has single value, it may not be able to a strong predictor.

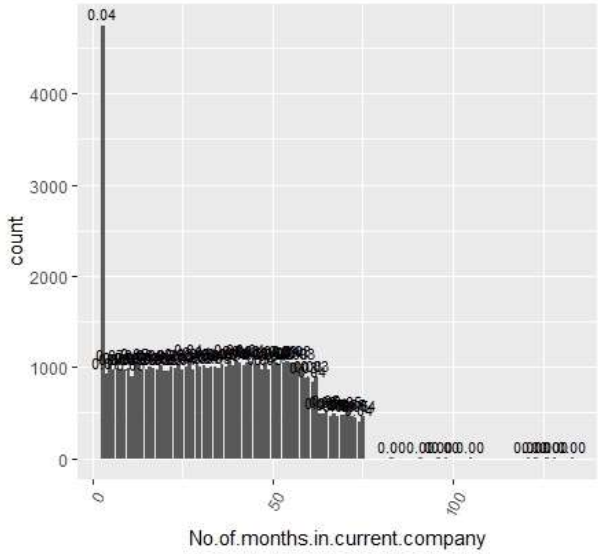




No. of months in current company

Defaults are high in lower no. of months. There's spike of applications & default in applicants with 3 month duration with current employer. This variable was found to have weak prediction strength in IV Analysis.





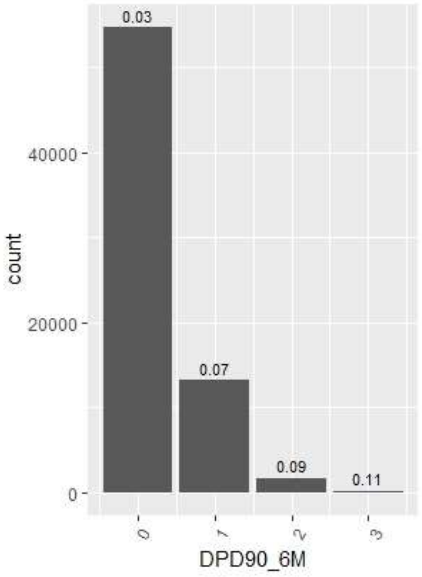
Exploratory Data Analysis – Credit Bureau Data

During IV analysis, it was observed that most strong to medium predictor variables were from credit bureau data. It is important to analyse these variables carefully to get insights that can help our model building process.

```
> summary(credit_bureau_data)
     appid              DPD90_6M         DPD60_6M         DPD30_6M         DPD90_12M        DPD60_12M
 Min.   :1.004e+05   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:2.484e+08   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :4.976e+08   Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :4.990e+08   Mean   :0.2703   Mean   :0.4305   Mean   :0.5772   Mean   :0.4503   Mean   :0.6555
 3rd Qu.:7.496e+08   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :1.000e+09   Max.   :3.0000   Max.   :5.0000   Max.   :7.0000   Max.   :5.0000   Max.   :7.0000

   DPD30_12M        Avg_CCUtil      trades_6M        trades_12M      PL_trades_6M   PL_trades_12M        enq_6M
 Min.   :0.0000   Min.   :  0.0   Min.   : 0.000   Min.   : 0.000   Min.   :0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.:0.0000   1st Qu.:  8.0   1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.:0.000   1st Qu.: 0.000   1st Qu.: 0.000
 Median :0.0000   Median : 15.0   Median : 2.000   Median : 5.000   Median :1.000   Median : 2.000   Median : 1.000
 Mean   :0.8009   Mean   : 29.7   Mean   : 2.298   Mean   : 5.827   Mean   :1.207   Mean   : 2.397   Mean   : 1.764
 3rd Qu.:1.0000   3rd Qu.: 46.0   3rd Qu.: 3.000   3rd Qu.: 9.000   3rd Qu.:2.000   3rd Qu.: 4.000   3rd Qu.: 3.000
 Max.   :9.0000   Max.   :113.0   Max.   :12.000   Max.   :28.000   Max.   :6.000   Max.   :12.000   Max.   :10.000
                  NA's   :1058    NA's   :1
    enq_12M        open_hsgloan      outst_bal        total_trades    open_autoloan      performance
 Min.   : 0.000   Min.   :0.0000   Min.   :      0   Min.   : 0.000   Min.   :0.00000   Min.   :0.0000
 1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.: 211532   1st Qu.: 3.000   1st Qu.:0.00000   1st Qu.:0.0000
 Median : 3.000   Median :0.0000   Median : 774992   Median : 6.000   Median :0.00000   Median :0.0000
 Mean   : 3.535   Mean   :0.2564   Mean   :1249163   Mean   : 8.187   Mean   :0.08462   Mean   :0.0422
 3rd Qu.: 5.000   3rd Qu.:1.0000   3rd Qu.:2920796   3rd Qu.:10.000   3rd Qu.:0.00000   3rd Qu.:0.0000
 Max.   :20.000   Max.   :1.0000   Max.   :5218801   Max.   :44.000   Max.   :1.00000   Max.   :1.0000
                  NA's   :272      NA's   :272                                          NA's   :1425
```

DPD90_6M

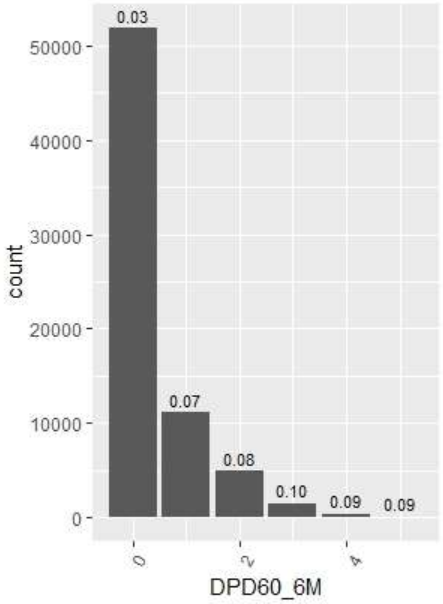
Nearly 78% of subjects are not 90 Day Past Due in past 6M which can help narrow our focus on 90DPDs. High number of people-970 who are 90 DPD in past 6 month have defaulted Higher 90DPDs in past 6M seem to show more default ratio as below plot shows.



About 7.59 % default rate is observed for those who are 90DPD for 1 or more time while default rate is about 3.28% in those who have not been 90DPD in 6M.This is one of the medium strong predictors as per IV/WoE analysis.

DPD60_6M

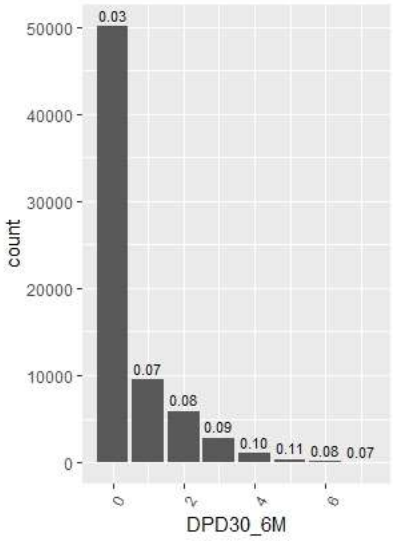
Nearly 74% of subjects are not 60DPD in past 6M which can help narrow our focus when screening applications. High number of people who are 1 or 2 times 60DPD show defaults. Higher number of 60 DPDs seem to show more default ratio as shown in below .



Default rate is 7.59 % for those who are 60DPD for 1 or more time vs 3.05% default rate in those who have not been 60DPD in 6M.This variable is an medium important predictor as per IV/WoE analysis as well.

DPD30_6M

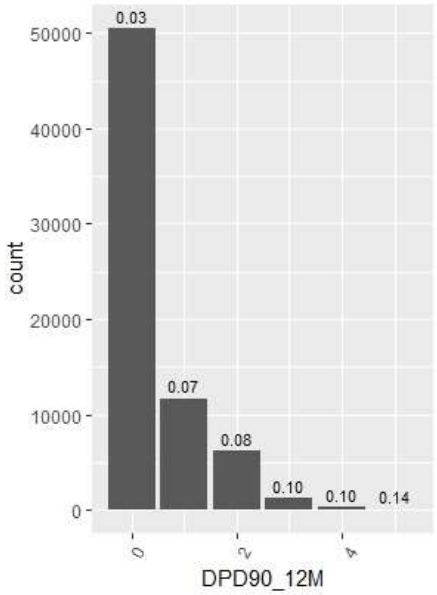
Nearly 71% of subjects are not 30DPD in past 6M which can help narrow our focus in analysing applications. High number of people who are 1 or 2 times 30DPD have. Also, it is observed that 1st time or second time 30DPD have higher default rate compared to 60DPD, 90DPD 1st timers. Which means that we can focus on those who have been 30DPD 1-3 times in past 6 months. Higher DPDs seem to show more default ratio.



Default rate is 7.55 % for those who are 30DPD for 1 or more time. 2.91% default rate in those who have not been 30DPD in 6M. This variable is an medium important predictor as per IV/WoE analysis as well.

DPD90_12M

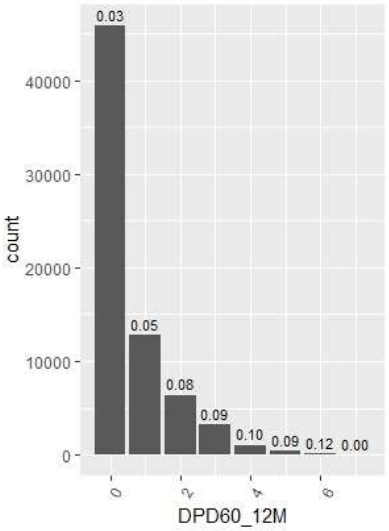
Nearly 72% of subjects are not 90DPD in past 12M which can help narrow our focus in application approvals. People who have been 90DPD in 12M 1-3 times show high defaults. Higher 90DPD results in higher default ratio as seen from plot below. This is one of the medium important variables as per IV analysis.



Default rate is 7.42 % for those who are 90DPD for 1 or more time. 2.99% default rate in those who have not been 90DPD in 12M.

DPD60_12M

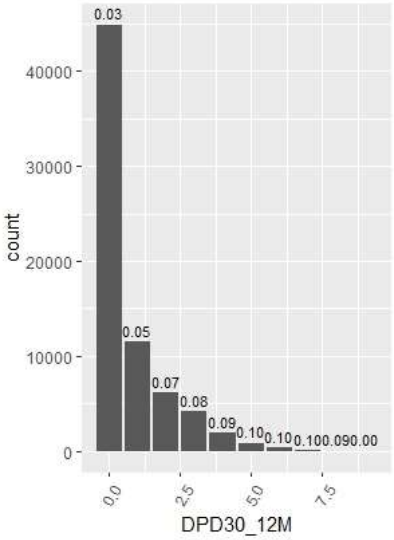
Nearly 65% of subjects are not 60DPD in past 12M which can help narrow our focus in application process. People who have been 60DPD 1-4 times seem to be defaulting more. Default rate increases with number of 60DPD in 12M.



Default rate is 6.54 % for those who are 60DPD for 1 or more time. 3% default rate in those who have not been 60DPD in 12M. This variable is a medium important predictor as per IV/WoE analysis as well.

DPD30_12M

Nearly 64% of subjects are not 30DPD in past 12M which can help narrow our focus in application process. Subjects who have been 30DPD in past 12 M show higher numbers Highest default rates at 5-7 times DPD.



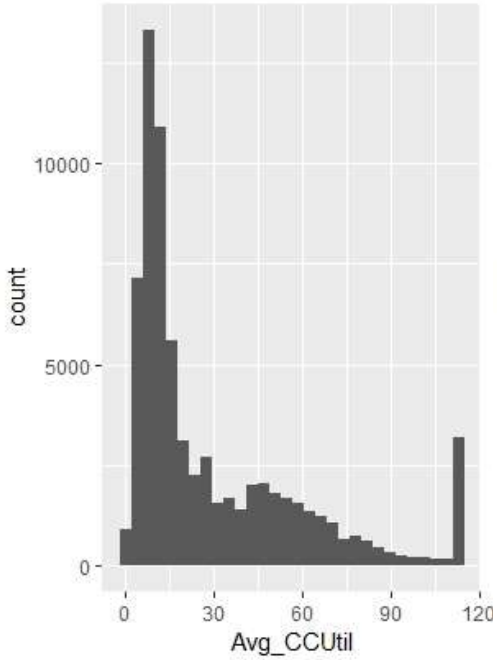
6.53 % default rate for those who are 60DPD for 1 or more time. 2.93% default rate in those who have not been 60DPD in 12M. This variable is a medium important predictor as per IV/WoE analysis as well.

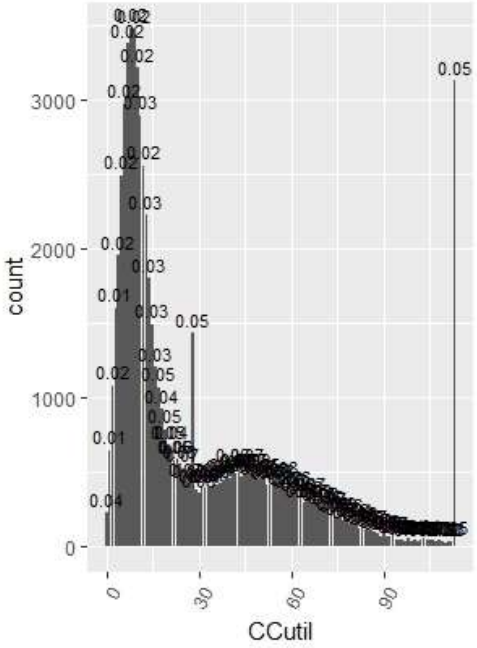
In general, it can be concluded that DPD is a good indicator to probed as a variable in our model since it shows ability to discriminate between defaulters and those who will not default. DPD variables were found to have moderate importance based on IV value.
Although large number of subjects who defaulted but were never DPD, such subjects will be part of the group that never defaulted and was never DPD. Hence having watch on DPD seems to be a good predictor.
Secondly, DPD categories seem to have strong correlation with each other.This could be due the fact that someone who is 60DPD or 90DPD is 30DPD in past 6M/12M as well as there could be trend that someone who's 30DPD is likely to be 60DPD hence scrutinizing subjects which show consistent default tendency will help contain default rate.

Avg. CC Utilization

75% people have used about 45% of their cc limit. Plot shows high number of people who are using more than 100% of their limit and are defaulting at higher rate. 3666 subjects have fully used/exceeded their credit limit-this is more than 5% of the data. 80% of subjects have used nearly 50% of their credit card limit. 6.567534 - High default rate in people who have used more than 50% of their credit limit.

We imputed empty values first with 0 and then with imputation methods(to evaluate impact on model).



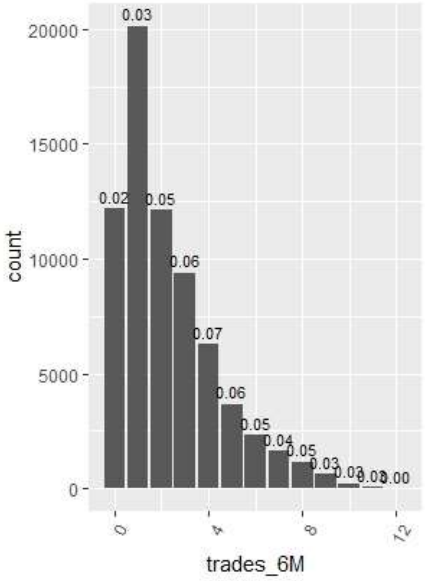


Based on IV/WoE analysis, this variable was found to be strongest predictor.

trades_6M

High number of people who have 1-4 trades have defaulted. Only 17% population does not have any trades in 6M. 22% population has 4 or higher trades. people with 3-5 trades have higher default rate. As per WoE/IV analysis, people with 2 or more trades have -ve WoE.

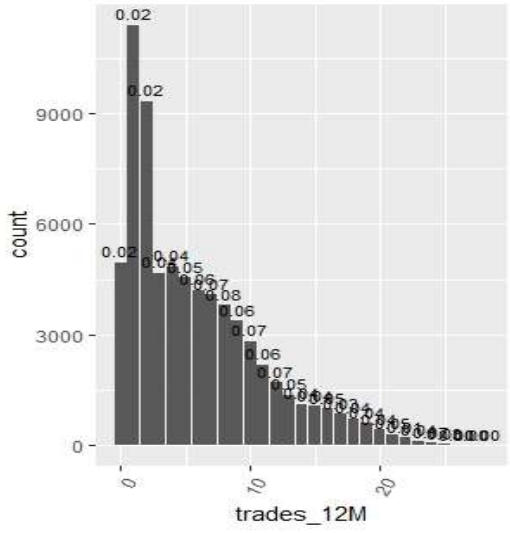
5.7% default rate for 3 or more trades. While those below 3 were found to default only 3.25. This column was seen to have medium importance as per IV/WoE analysis.



trades_12M

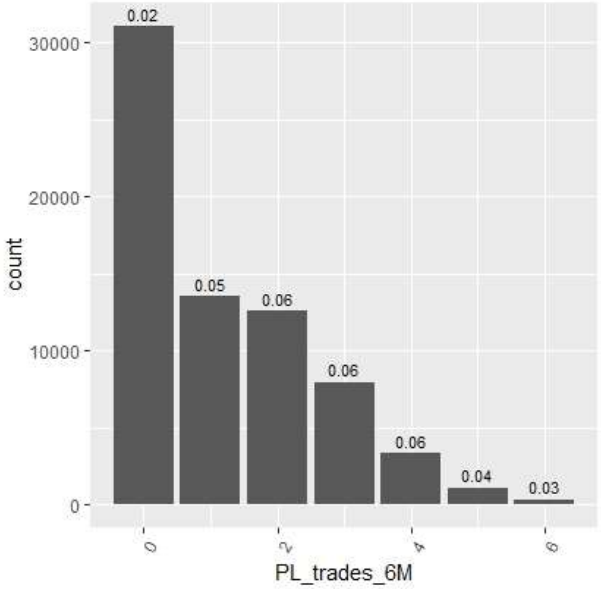
This variable is of strong importance from IV/WoE analysis. Defaults seem to gradually grow and peak at 8 trades and then taper down. Only 7% population does not have any trades in 12M. People with 8 trades have highest default rate as well as actual defaults.

High default rate of 5.78 for those who have 4 or more trades in a year vs those who have less than 4 trades at 2.19



PL_trades_6M

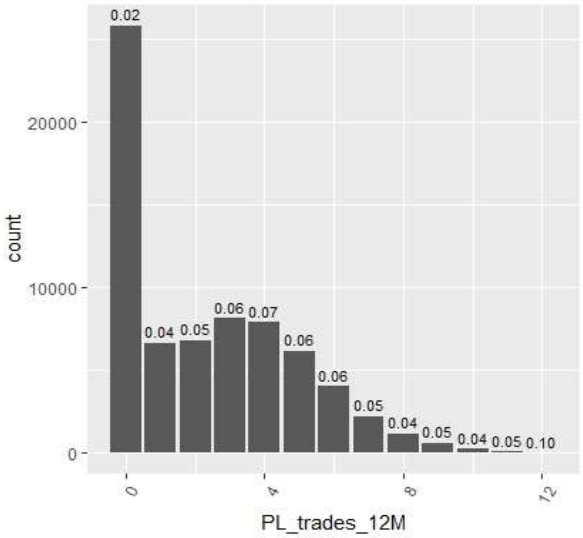
44% subjects do not have even 1 PL trade in past 6M. High number of defaults at 1-4 PL trades in 6M. Subjects with 2-4 trades have highest default rate (34% of population). 5.8% default rate in population with 1 or more pl trade in 6M. vs 2.24 in those with no PL trade.



This column was seen to have medium importance as per IV/WoE analysis.

PL_trades_12M

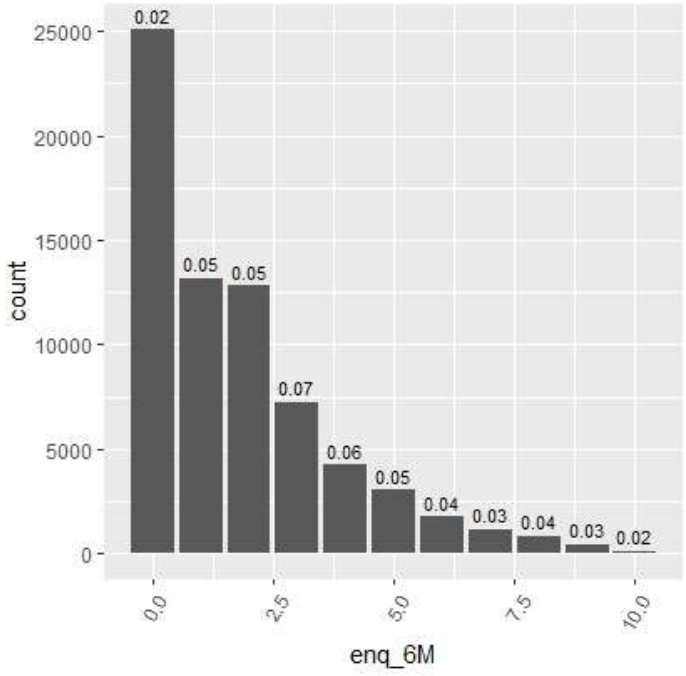
As per IV/WoE analysis, this is strong importance predictor. 36% subjects do not have single PL trade in past 12M. High number of defaults at 1-6 PL trades in 12M. People with 3-5 trades have high rate of defaults.



High default rate of 5.65 for people with one or more PL Trade in past 12M vs 1.77 in those with no PL Trade in 12 M

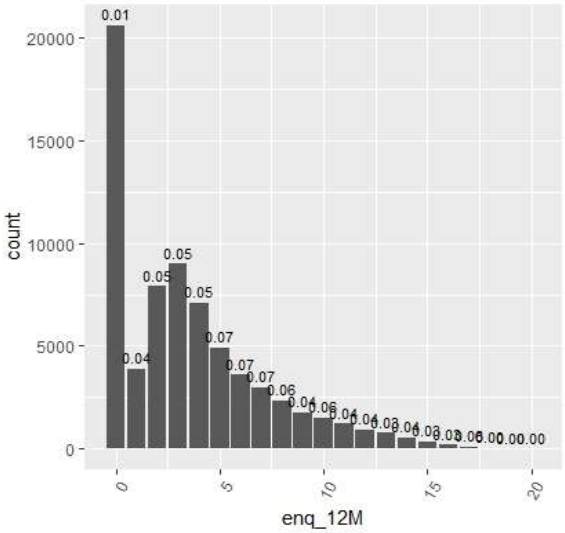
enq_6M

35% did not have any enquiry in past 6M. People with 1-3 enquiries in past 6M have higher defaults. 1-5 enquiries in 6M shows more default ratio. 5.4% default rate when there is 1 or more enq in past 6M VS 2.1% with no enq. This column was seen to have medium importance as per IV/WoE analysis.



enq_12M

This is one of the strong predictors as per IV/WoE analysis. 29% population did not have any enquiry in past 12M.



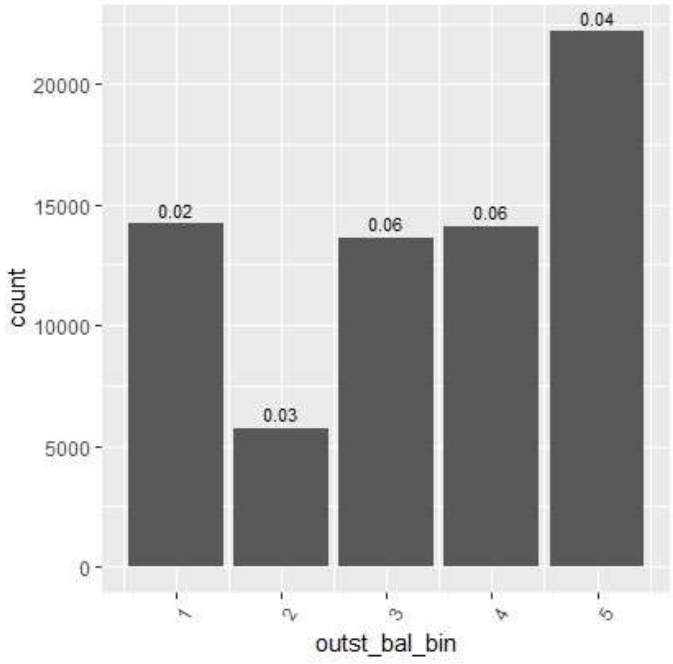
3-6 enquiries in past 12M show higher number of defaults. 2-7 enquiries in 6M has more default ratio. 5.36% default rate when there is 1 or more enq in past 12M VS 1.49% with no enq.

outst_bal

Higher loan amounts seem to have higher defaults- bad sign for the company. 51% population is in 4th & 5th bin-high amount loans, which makes it important to scrutinize high amount loans. Those who have taken higher amounts of loan seem to default more - which is cause of concern.

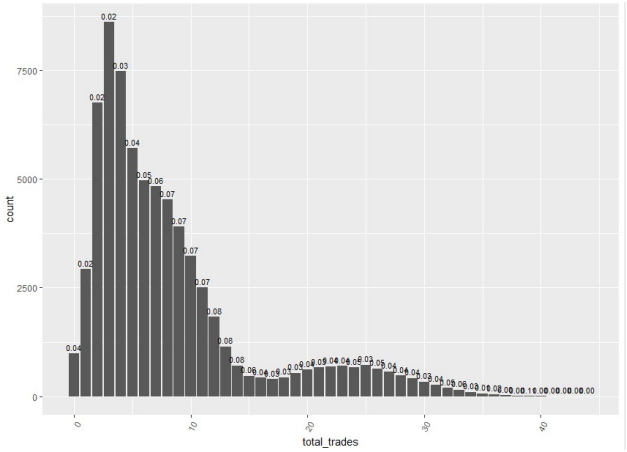
This column has NA values.Column with WoE values was used to replace original values in a separate data frame for building regression model.

Outstanding balance is binned to find default rates. **This is one of the medium importance predictor as per IV/WoE analysis.**



Total trades

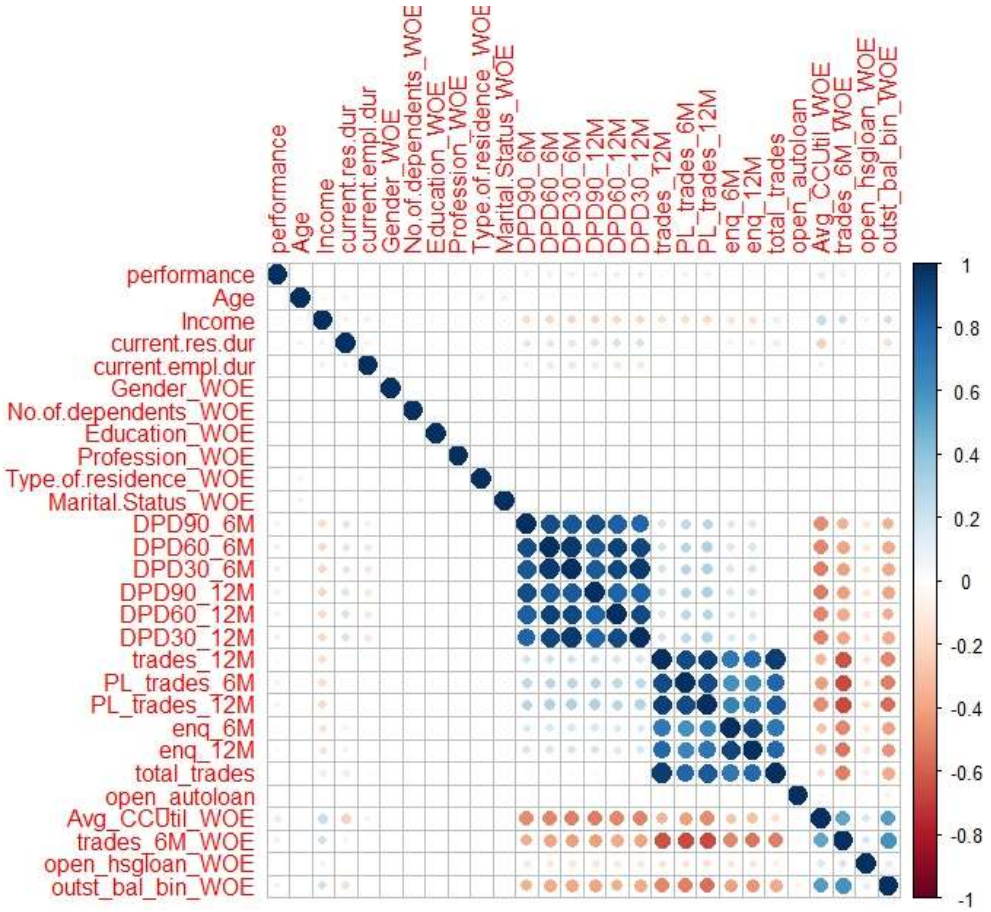
Only 1% population does not have trades. 59% population is between 4-10 trades. 7-14 range of trades shows higher defaults.



5.863464 default rate for subjects with 7 trades or above.It is observed that default rate increased beyond 5% above 4 trades or more have very strong correlation which also true in case of trades and enquiries. **This is one of the medium strong variables as per IV/WoE analysis.**

Correlation check

Correlation plot/table is able to show correlation of performance with important variables identified in IV/WoE analysis. Average CC Utilization, Trades_6M, DPD's, PL Trades, trades and enquiries have correlation with performance.
It is also observed that strong correlation exists amongst DPD columns - which indicates that people may have tendency to be repeat DPS in payments
Average CC Util shows strong correlation with DPD, open loans, trades and moderate correlation with income and current residence duration.



Open housing loan has correlation with outstanding balance – which is obvious.

Model Building

- **Models based only on Demographic Data**

Logistic Regression:

We have started model building process. As a first step, we built a logistic regression model to predict the likelihood of default using only the demographic data.

- For building logistic regression, we used a separate data frame that had columns containing WoE values instead of original columns in which values were missing.
- Inverted performance tag since we want to predict approvals via an automated model.
- We created some derived variables and scaled the data.
- Split the data into train & test in 70:30 ratio.
- Performed stepwise selection using stepAIC post a generic first model.
- Variables are selected basis of lower p-value and to lower VIF keeping close check on AIC.
- Final model is selected with significant variables with least correlation and having lowest set of variables. **Income** and derived variable based on **number of months in current company** (which were found to have weak strength as per IV analysis) were part of final model.
- Found the optimal probability cutoff to keep balanced performance. Creating cutoff values from 0.01 to 0.99 for plotting and initializing a matrix of 1000 X 4. Plotted cutoff values vs confusion matrix values. Tried to further fine tuning the range of 0.07. The model with only demographic data is giving
    - Accuracy: 51-58%
    - Sensitivity: 51-58%
    - Specificity: 51-57%,

SMOTE-The overall data contains only 4% are defaulted customers of the total customers. We tried smote on train data to even the data. Accuracy, Sensitivity, Specificity are improved to some extent.

We also tried **Gradient Boosting/Generalized Boosted Regression** which did give similar results to logistic regression.

Model Validation using KS Statistic and Lift & Gain- We validated model using KS Statistics & Lift & Gain charts.

Our model is producing very low accuracy mostly due to uneven spread of performance values and low predictive power of demographic variables.

Next we tried **Random Forest & SVM** models on demographic data. Data was balanced using SMOT and imputed using MICE.

Random Forest Modelling

For RF, we cleaned up demographic data frame with NA values imputed using MICE. The data frame also included derived variables. Observed that higher number of trees yield slight improvements in accuracy. However since the improvement is not significant, we'll select moderate number(100) of trees & variables(4-5) for sampling.

SVM

So far we have tried building SVM using Linear Kernel, Polydot Kernel and RBF Kernel using default hyperparameters. We tried to tune RBF kernel with different cost and sigma values using grid search and are evaluating results. So far all the SVM models built are giving high accuracy and sensitivity and low specificity.

Overall it is observed that, with only demographic data, model may not provide required accuracy and the data needs to be supplemented with Credit Bureau data since as per IV analysis, credit bureau data contained strong and medium strong predictors of performance.

Future Roadmap(Model Building-contd../Evaluation & Deployment)

Since only demographic data is not enough to build model that can give predictions, we will be building model on combined data set that will contain both Demographic and Credit Bureau Data. For logistic regression we used separate data frame that will include WoE columns for columns with missing values.

For other model types, we will impute data using MICE and perform SMOT to balance data.

Model Building-contd

- **Model using both Demographic and Credit Bureau Data**: we will then build a model to predict default using both the data sets. We will start with a logistic regression model. Further, we would try and build other models such as random forest, SVM.
- For logistic regression, we will select variables based on P-Value, VIF and AIC impact.
- So far, with logistic regression on data without SMOTE, we were able to get accuracy, sensitivity, specificity values of 64%. We will continue build other models further to compare improvements.
- For Random Forest, we will try to find cut off that can give us balanced values of accuracy, sensitivity and specificity.
- SVM model hyperparameters-cost and sigma will be tuned by grid search.
- We'll try to observe that the resultant variables are relevant based on IV/WoE analysis.
- We shall try four different models (logistic regression, random forest, SVM, xg boost) and select the best one based on key performance metrics & stability.

Model Evaluation

- We will evaluate the models using tests and focus on achieving best of Discriminatory power, Accuracy/Calibration and Stability.
- Foe evaluation, we will use measures like confusion matrix, KS statistics, Lift-Gain chart etc. and ensure stability and generalisability of the model.
- As part of evaluation, we will predict the likelihood of default for the rejected candidates and assess whether the results correspond to actuals.

- We will also see how this model performs on defaulted candidates.

Building Application Scorecard

- Through model building and evaluation process, we will pick an optimum model and based on the final model, we will predict odds of being good for each applicant. Once we have the odds, we will sort the applicants from high to low odds or in simple terms most good to most bad. The good to bad odds needed as per requirement is 10 to 1 at a score of 400 doubling every 20 points. We'll use this scale in the scorecard and decide cut-off or threshold score above which CreditX can approve credit cards. Based on the scorecard, we will recommend the cut-off score below which CreditX could reject credit cards applications. The application scorecard will be evaluated on rejected population to compare the results with approved population.

This Application Scorecard will help CreditX to:

- Automate the credit card application decision processes which will help reduce cost of manual application underwriting.
- Help businesses to make accurate, consistent, data-based decisions.
- Help justify approvals in cases which otherwise would have been rejected w/o model this improving revenue.
- It will help optimize credit cards based on risk score, help identify customers for up/cross selling and manage bad debts closely.

Model Deployment & Financial Benefit Analysis

As last part of modelling process-deployment, we will provide CreditX with analysis of financial benefits in using the model. Important aspect of building models for approval of credit decisions is financial benefits over long term in terms of P&L. This information will be provided in a presentation that will be shared with CreditX and will include all assumptions made in the modelling process.

We will identify the metrics to be optimised and explain how the analysis and the model would work and share the results of the model as part of our presentation.

We will explain implications of using the model for auto approval or rejection which will speed up application process. We will also highlight potential credit loss avoided with the help of the model/scorecard for those candidates which otherwise would have been rejected based on current process.

Overall, our presentation will include

- Important factors/variables for credit decision with explanation of importance
- Assumptions
- The model building process & final model
- Assessment details of the model
- Benefits of the model to CreditX including financial as well as process optimization
- Maintenance of the model

Addendum-EDA for remaining variables

We performed EDA for variables that were not significant as per IV/WoE analysis. The results are detailed below.

Demographic Data

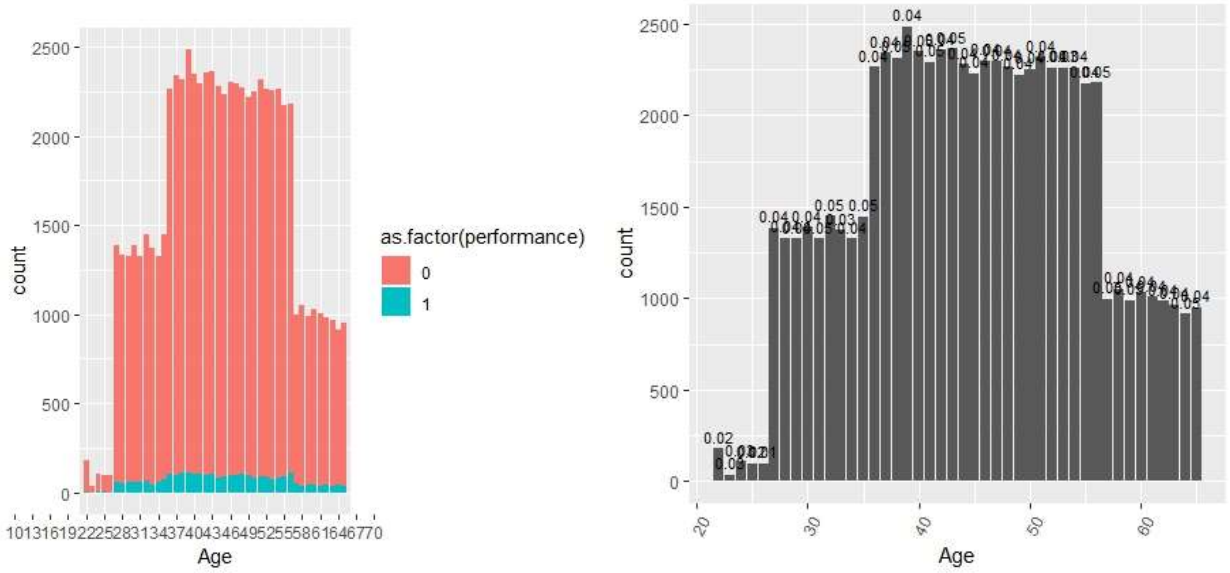
Age

First quartile is 37. Mean 45. 0th quartile contains -ve value which needed correction.

We made assumption that for getting credit cards, one needs to be 18 or above and imputed all 65 such values below 18 with the same. It was observed that this imputation did not alter the mean.

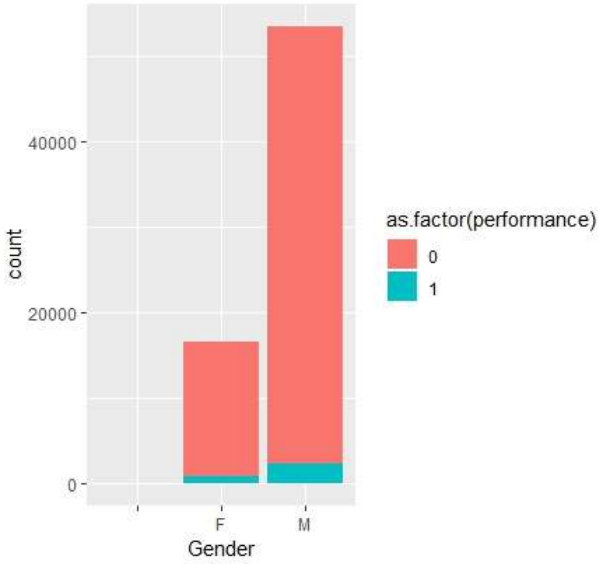
35-56 age range shows defaults above 70 and highest number of defaults at 38-41.

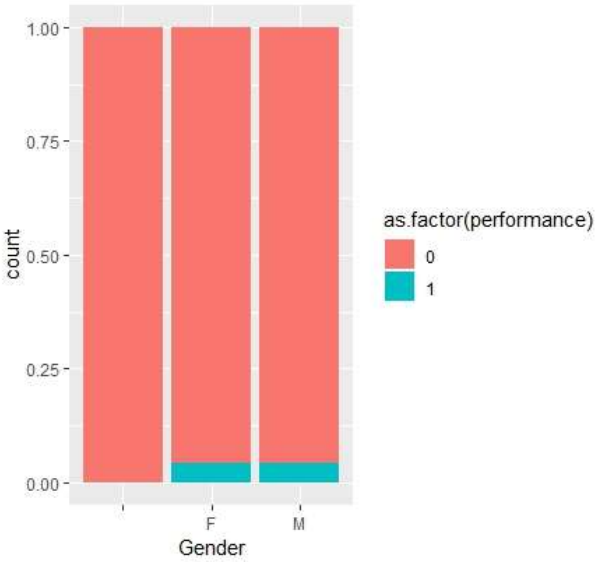
35-56 shows almost same default rate and high numbers of applicants are cause of more default numbers Mid age people seem to very slightly default more however as per IV analysis, age may not an significant predictor. Plots for analysis of age are shown below:



Gender

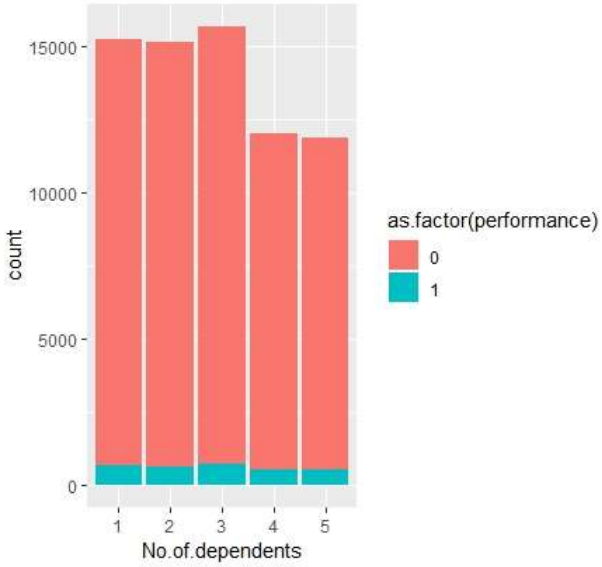
Proportion of male applicants is high. Default rate is the same for male and female. Default rate is same.
We <u>will not use</u> 'Gender' in our model to adhere to fair lending norms.

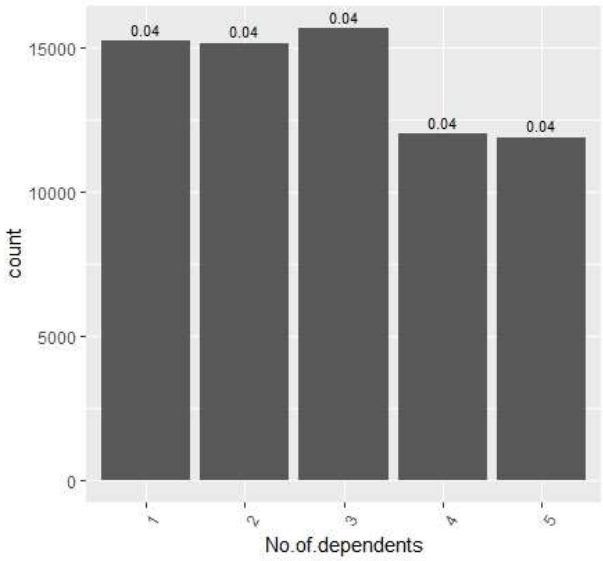




No of Dependents

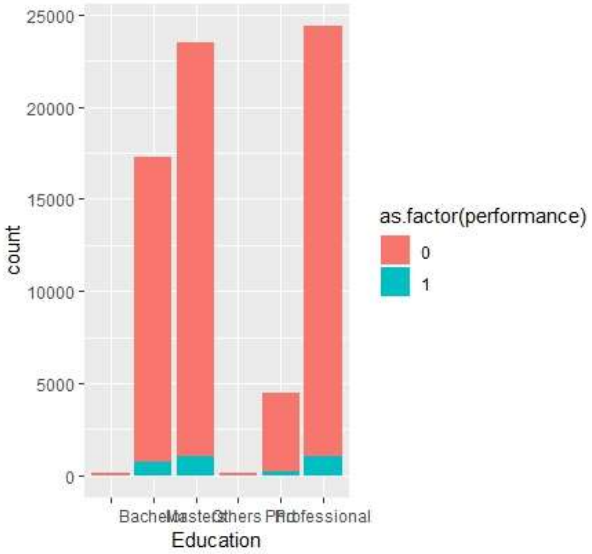
Applicants with 1 & 3 dependents have higher defaults. However, all the levels have same rate of default.

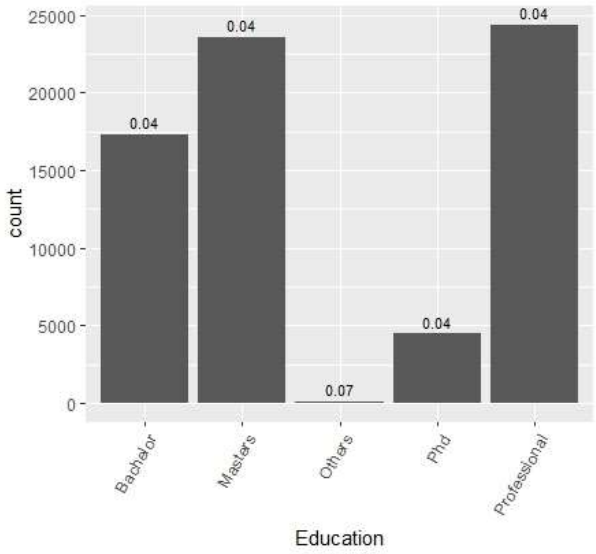




Education

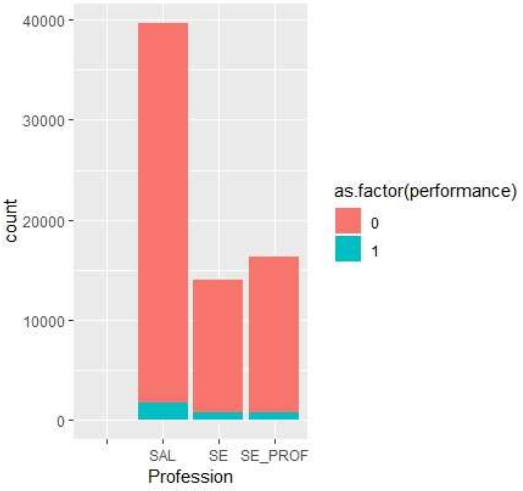
118 records do not have education mentioned. Masters category has high number of application as well as defaults.

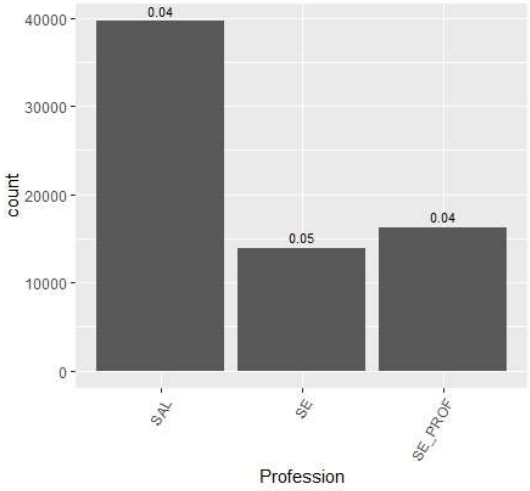




Profession

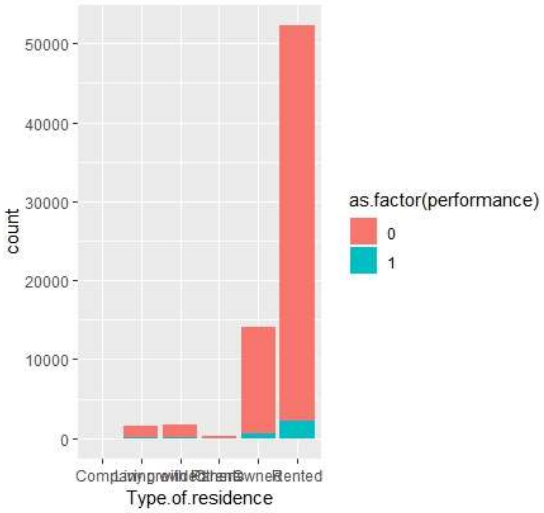
Salaried subjects have highest requests as well as defaults. Default rate is almost similar across categories.

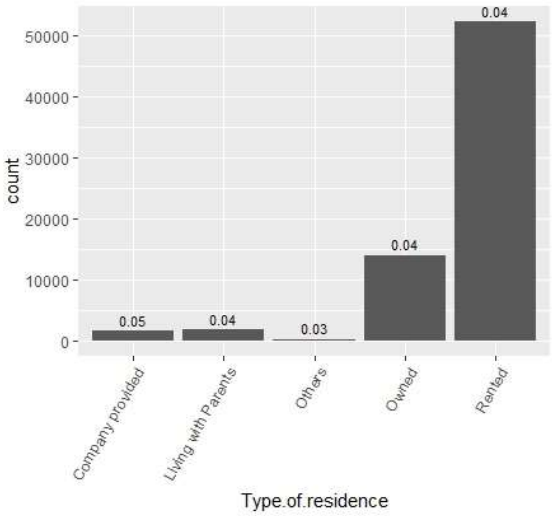




Type of Residence

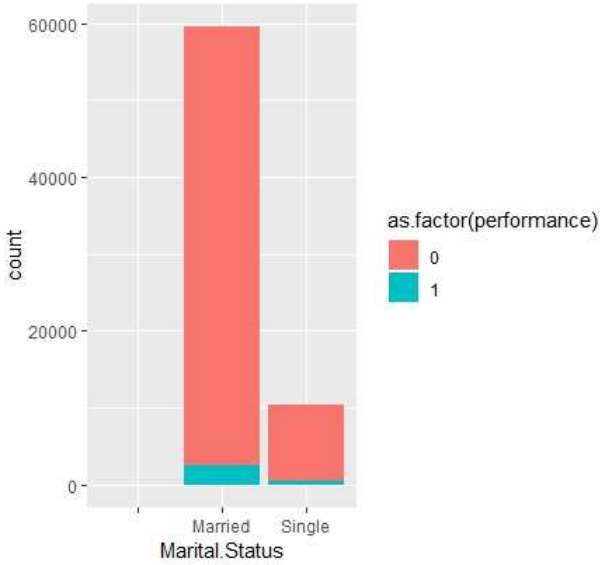
Rented and owned cover maximum number of applicants as well as defaults. Not much difference in default rates.

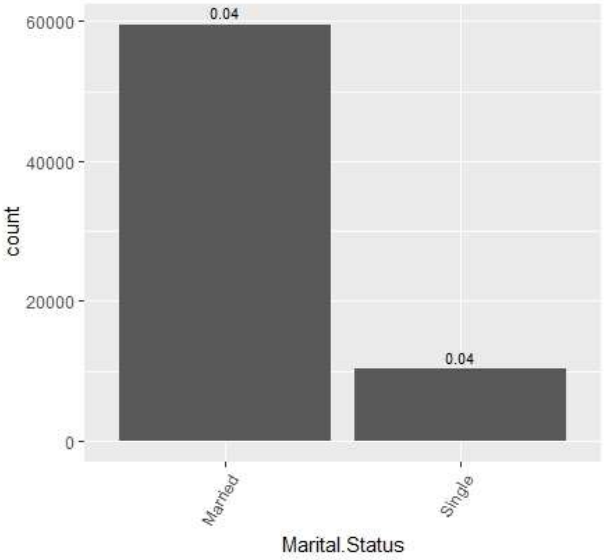




Marital Status

More than 85% of applicants are married. There're 5 NA values in data. Married applicants have higher default numbers however default rate is the same for both categories.

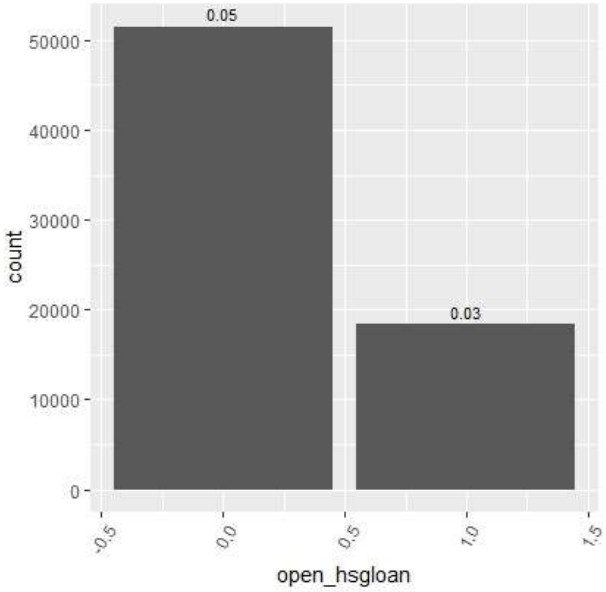




Credit Bureau Data

open_hsgloan

73% of people do not have open hsg loan. 615 Subjects who have open hsg loan defaulted. People that do not have housing loan seem to have more default rate. Housing loan may be a weaker indicator of default compared to other variables. It does not show as important variable in IV/WoE analysis.



open_autoloan

91% people do not have auto loan. Small number of people with auto loan have defaulted. Those who have taken auto loan or have not taken it show equal default rate. Auto loan may not be a significant variable in our model.