CredX Risk Analytics Approach Document

Batch : Jun-2018

Aditya Kumar
Raghuram Krishnamurthy
Rohit Saini
Veenu Bhanot

Business Objective

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.

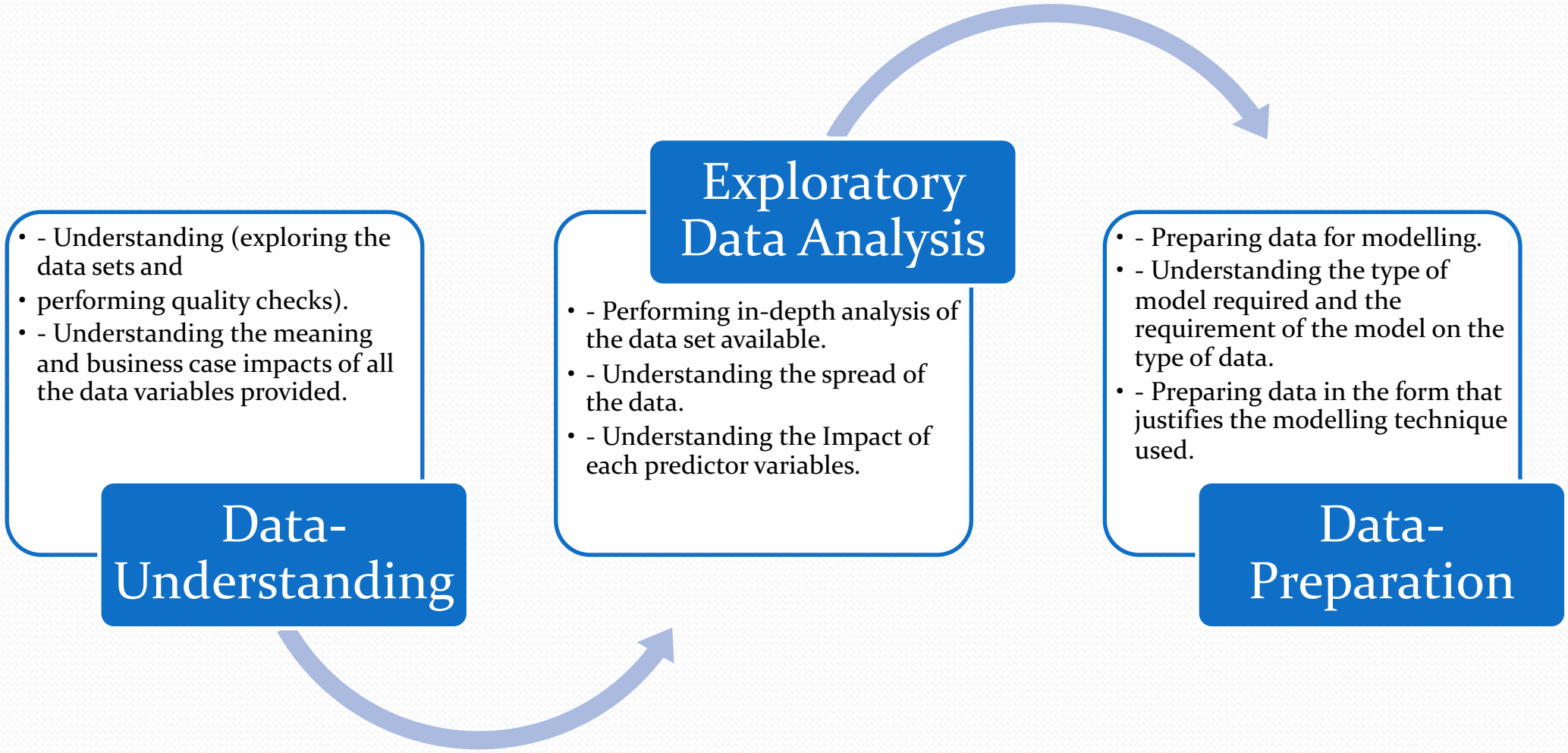
In this project, we will help CredX:

- Identify the right customers using predictive models.
- Using past data of the bank's applicants, we need to determine the factors affecting credit risk, create strategies to mitigate the acquisition risk.
- Assess the financial benefit of the project.

The datasets available to us are:

- *Demographic/application data*: This is obtained from the information provided by the applicants at the time of credit card application.
- *Credit bureau*: This is taken from the credit bureau.

Analysis Steps Followed

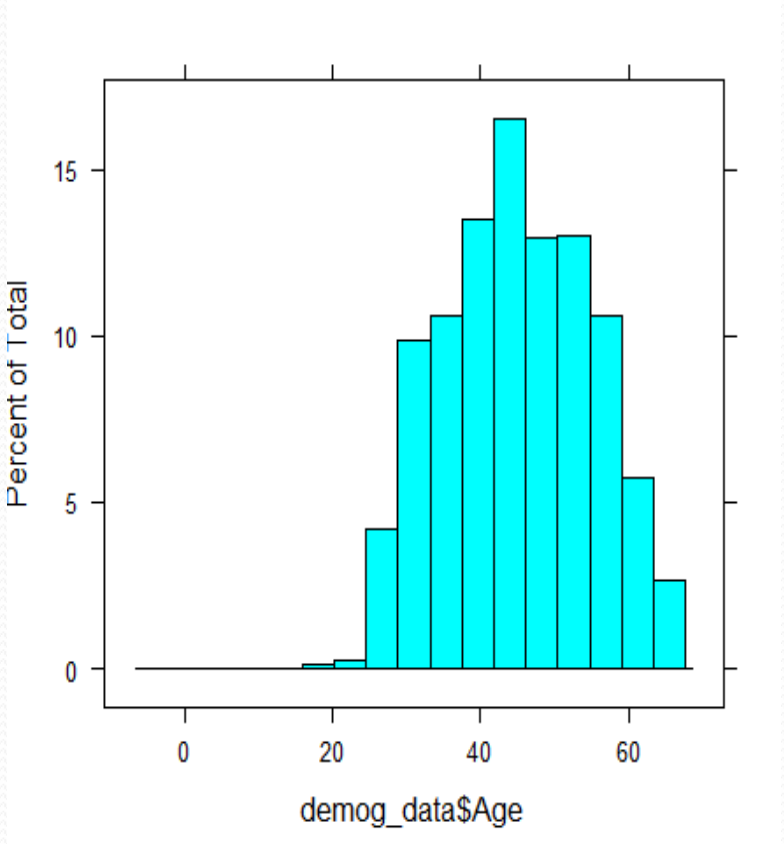


Data Understanding and EDA Summary

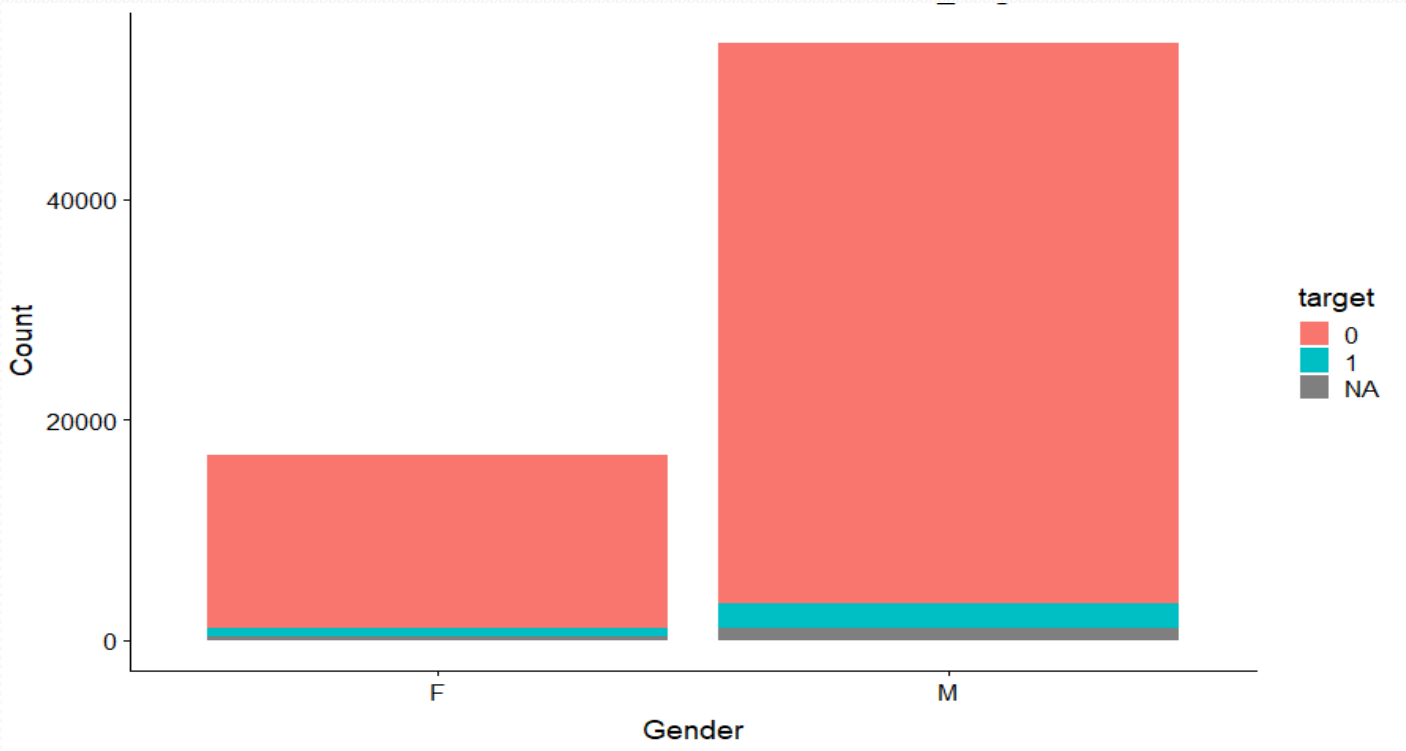
- Data is available in a structured format(csv), in two files for demographic and credit-bureau data.
- The Grain of data is at an individual customer level.
- There are duplicate Application IDs with different features in both demographic data and credit data.
- There are NA values in 'No. of dependents' and 'Performance Tag' in demographic data set.
- There are NA values in 'CC Utilization in 12 months', 'Trades opened in last 12 months', 'Open home loan', 'Outstanding balance' and 'Performance tag' in Credit bureau data set which has been imputed with WOE values.
- There are Age values in demographic data set that are less than 18, and has been dropped as erroneous data.
- There are negative values in Gender, MaritalStatus, No_of_dependents, Income, Education, Profession, Residence_type, No_of_mons_curr_residence, No_of_mons_curr_company and Performance_Tag which are dropped as erroneous data.
- The distribution of data in Performance tag is very skewed, the case of default is only around 4% of the data-set. The data is imbalanced.
- The inner join set of demographic and credit data is used for further data-preparation.

Univariate Analysis :

Understanding data-distribution



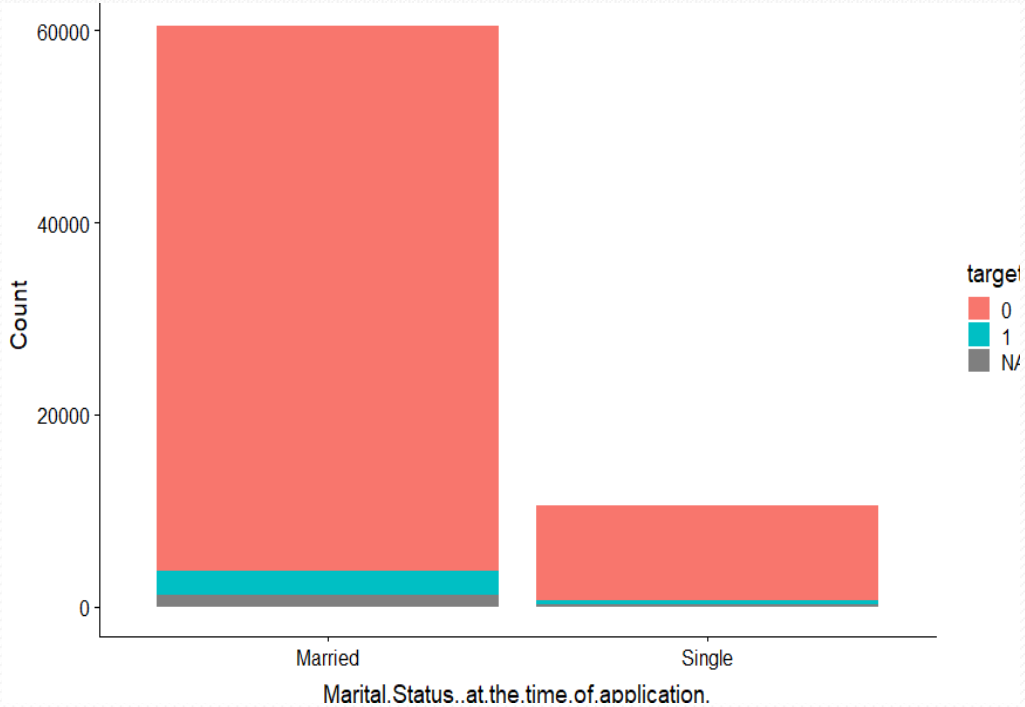
Observation : Age group between 40-55 tend to default the most



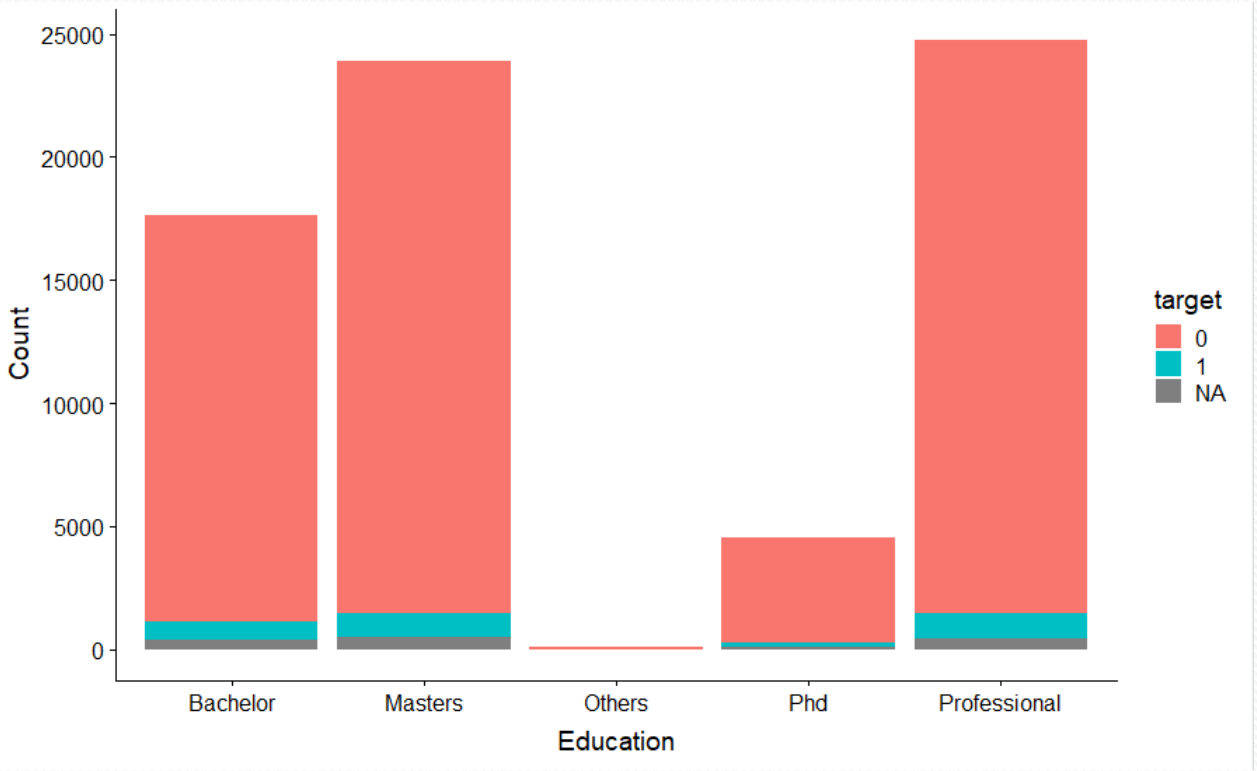
Observation : Males seems to default more than females.

Education Qualification has higher risk of defaulting.

Univariate Analysis :

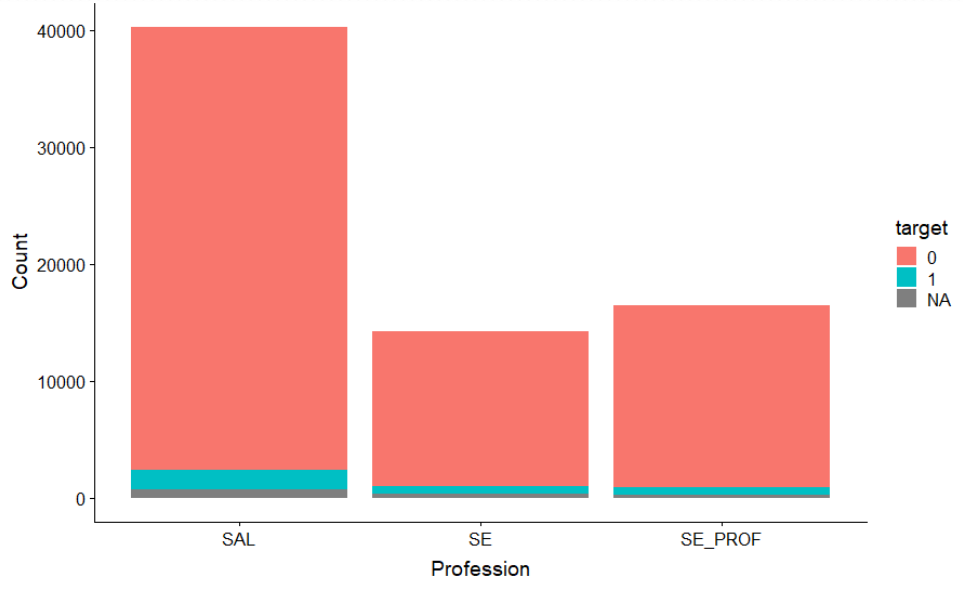


Observation : Applicants having marital status married has high risk of defaulting

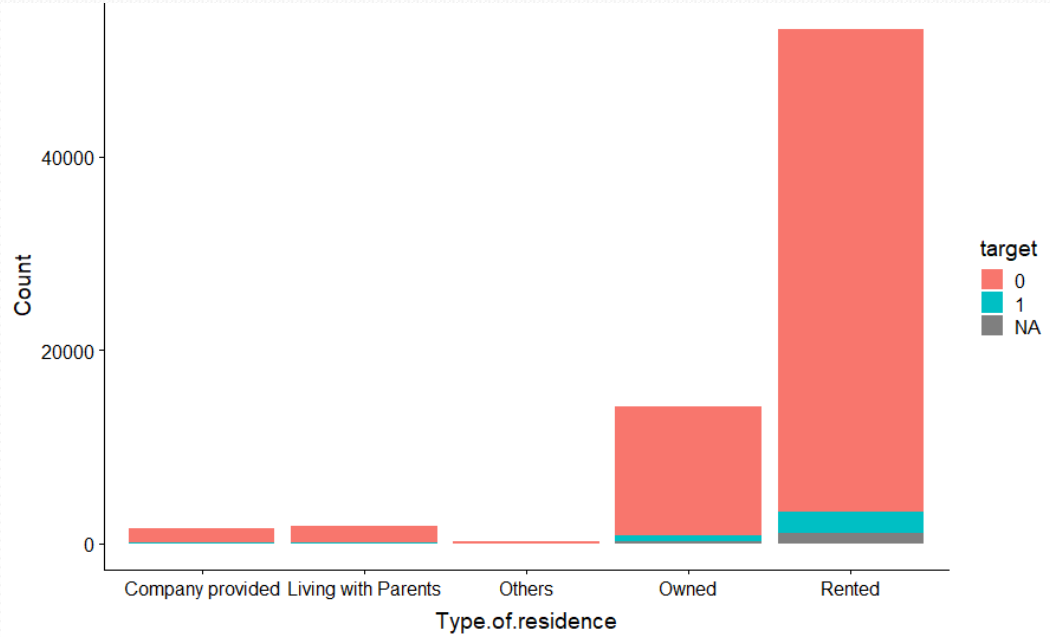


Observation : Applicants with Masters or Professional Educational qualification has higher risk of defaulting

Univariate Analysis :

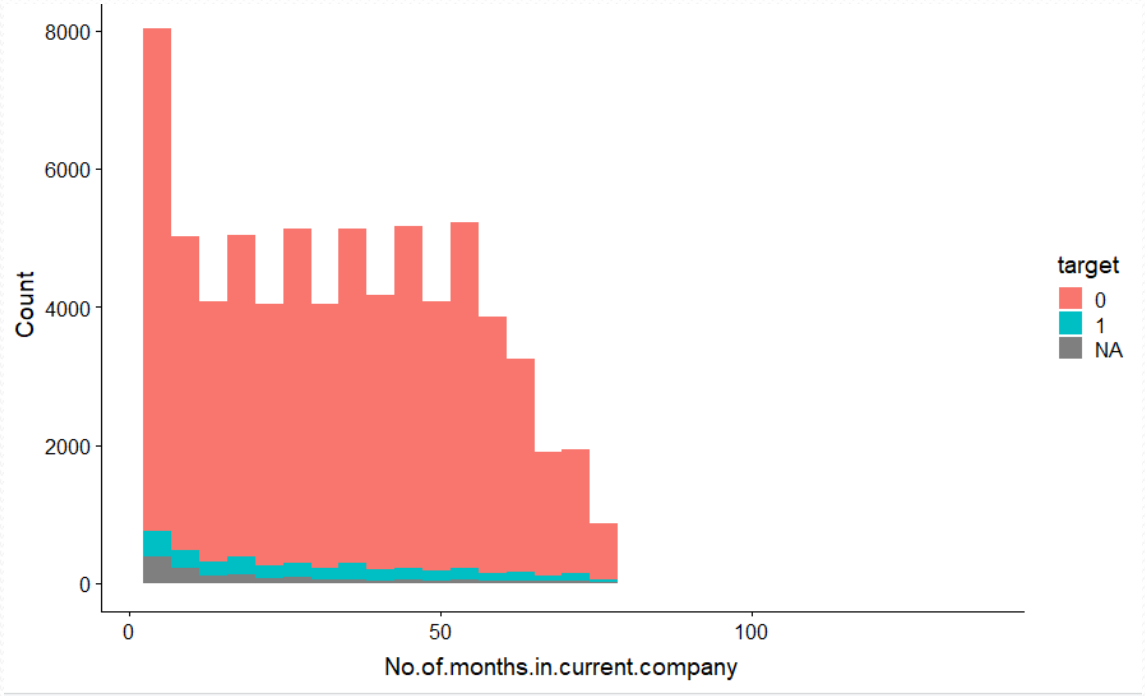


Observation : Salaried Applicants are the ones who default the most.

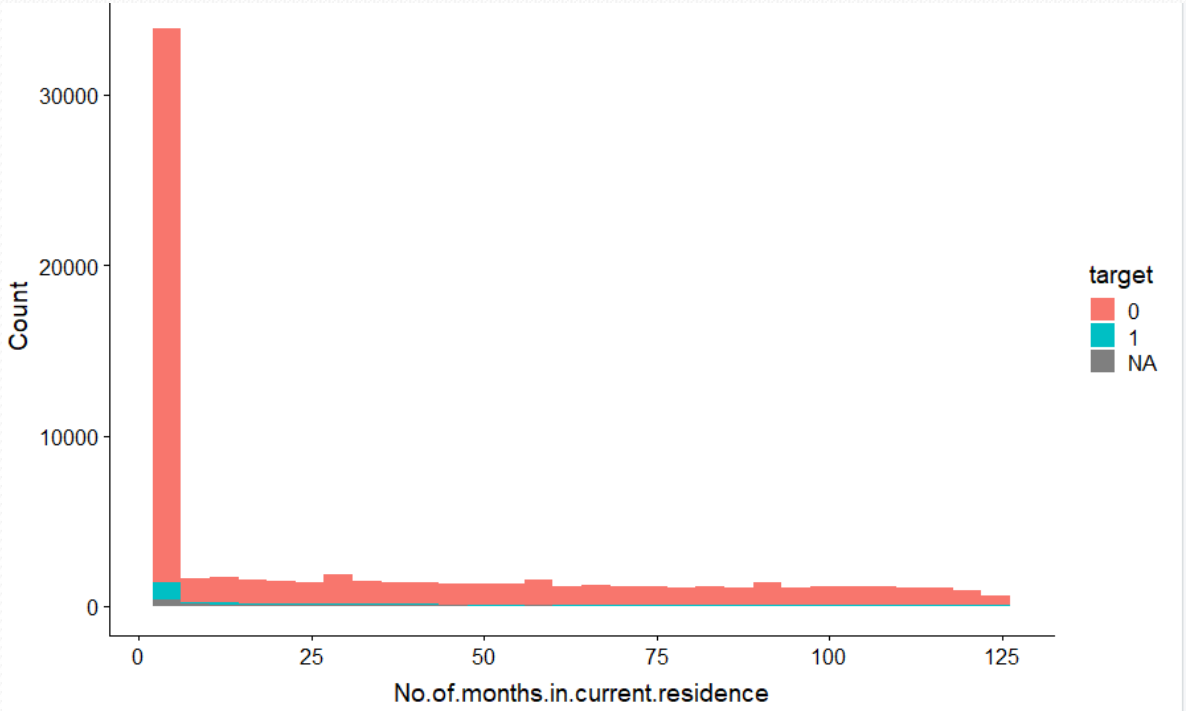


Observation : Rented ones are having high default chances .

Univariate Analysis :

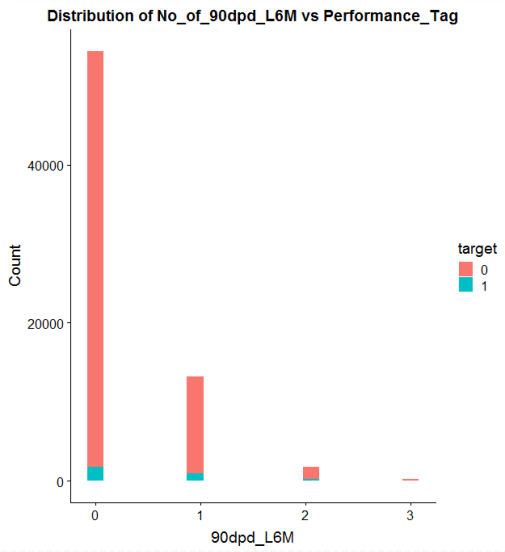


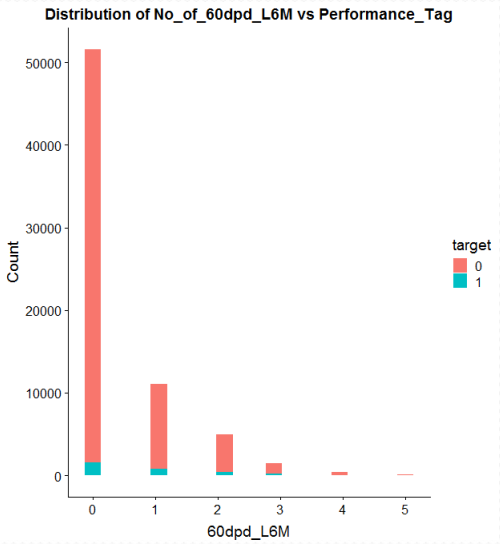
Observation : Less no of months in current company seems to be more defaulted case.

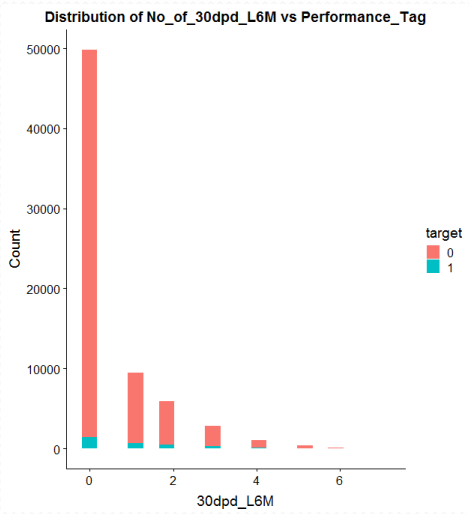


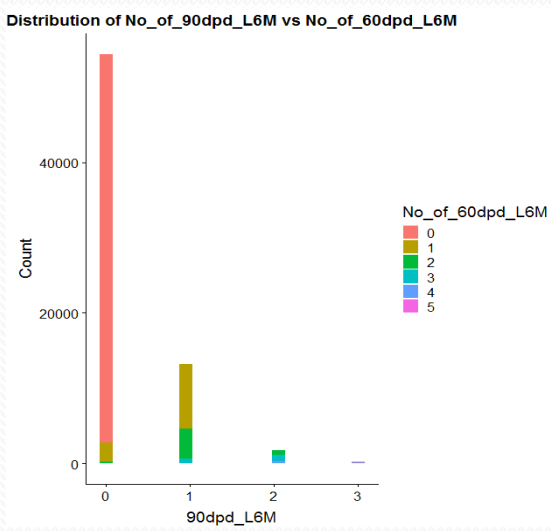
Observation : Less no of months in current residence seems to be more defaulted case.

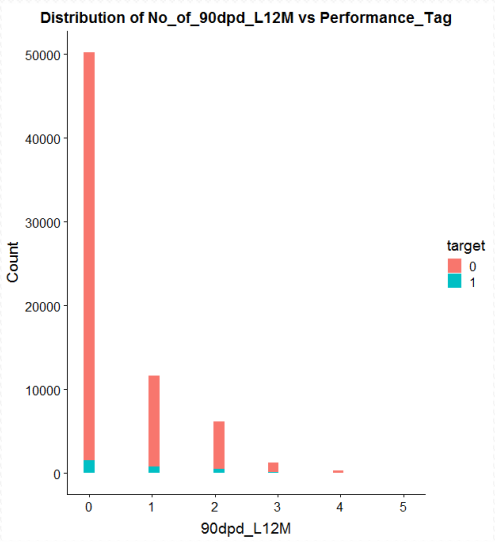
Bivariate Analysis

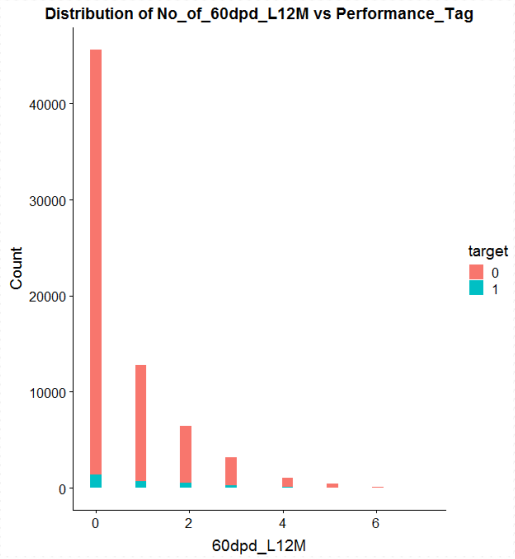


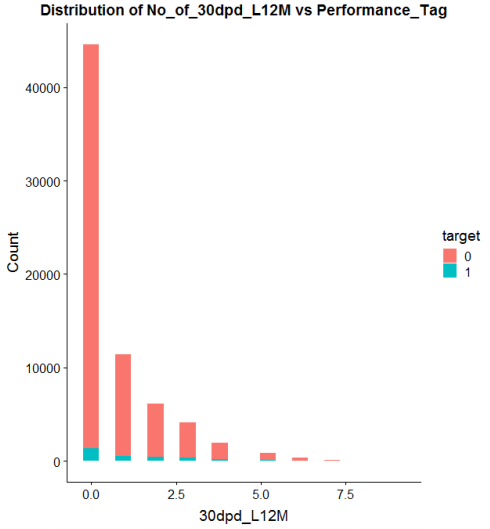


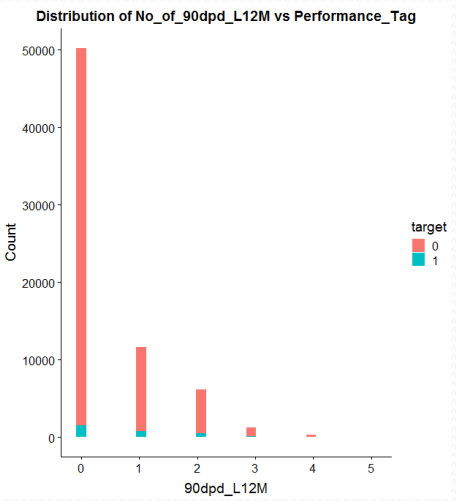




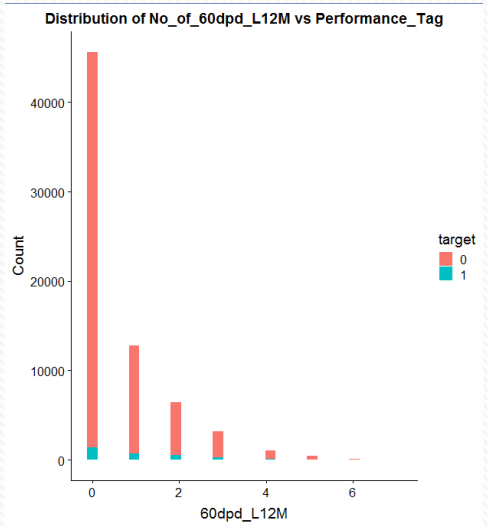


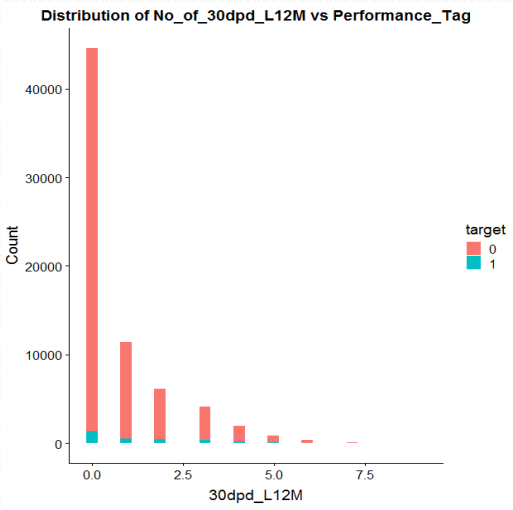


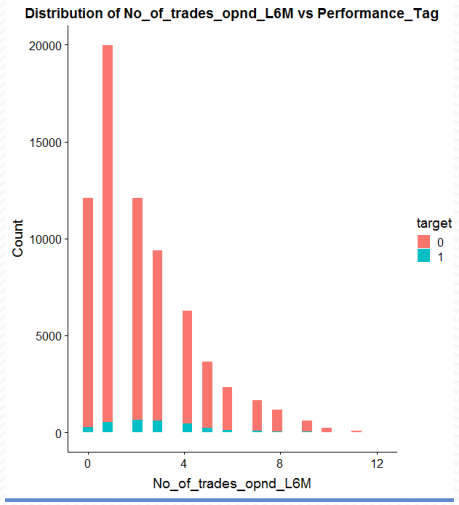


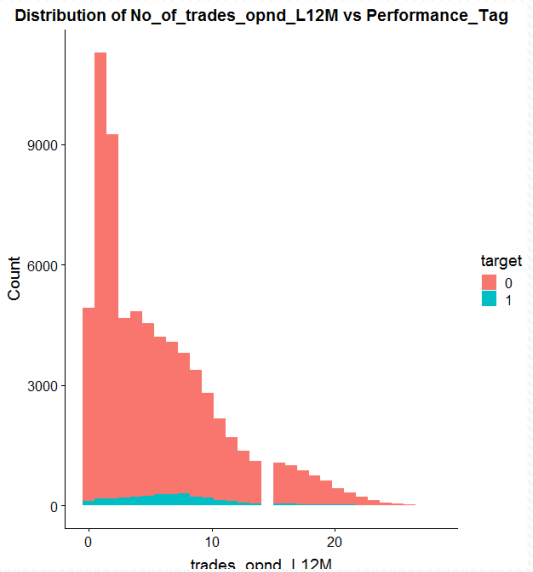


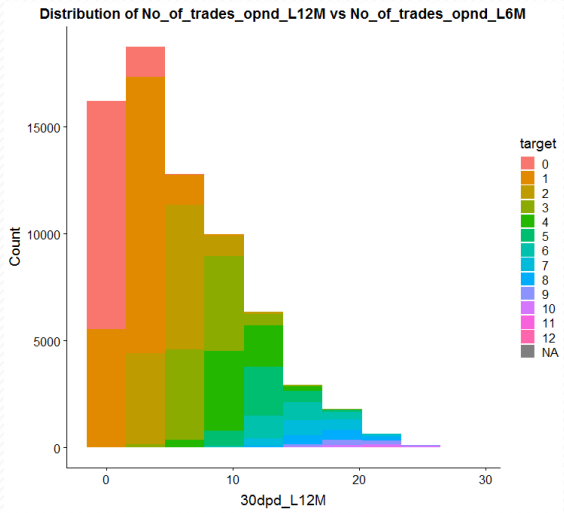
Bivariate Analysis

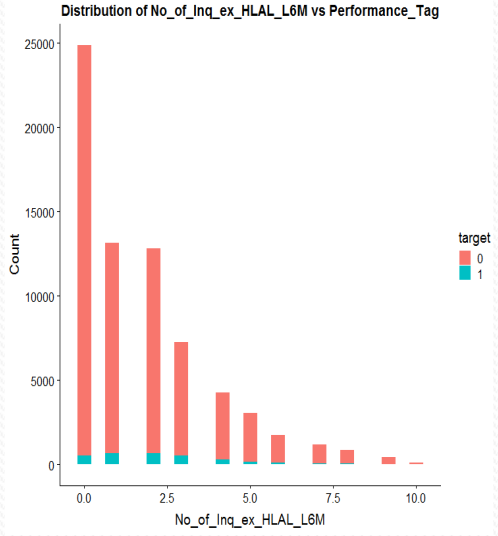


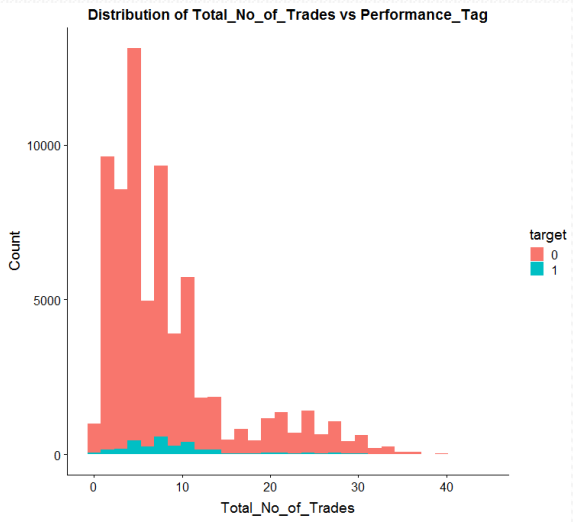


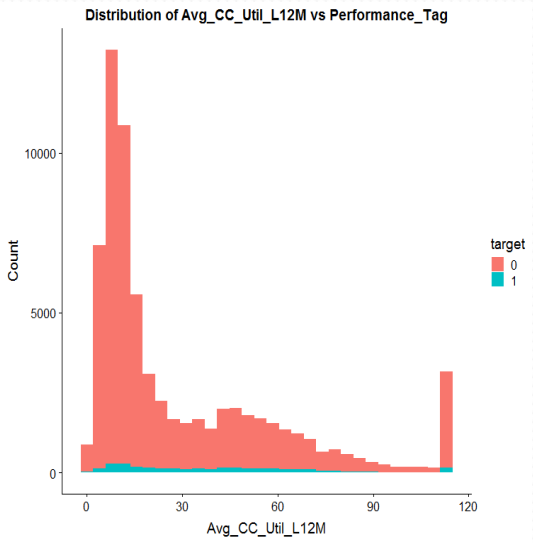












Important Predictors

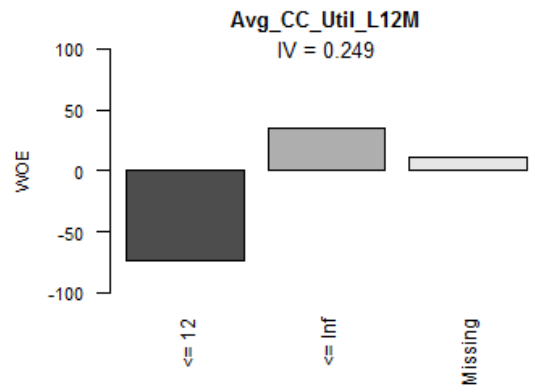
- The Information Value associated with the predictors are as follows:

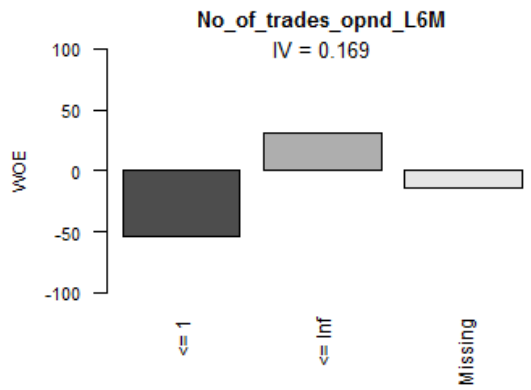
"No_of_PL_trades_opnd_L12M"	0.2685437	"Income"	0.03625863
"No_of_Inq_ex_HLAL_L12M"	0.2618248	"No.of.months.in.current.company"	0.02648455
"No_of_trades_opnd_L12M"	0.2532209	"woe.Presence_of_opn_HL.binned"	0.01716079
"woe.Avg_CC_Util_L12M.binned"	0.2478046	"woe.Outstanding_Bal.binned"	0.0146238
"No_of_30dpd_L6M"	0.234797	"Age"	0.003083984
"No_of_30dpd_L12M"	0.2142864	"No.of.dependents"	0.002494763
"No_of_PL_trades_opnd_L6M"	0.2122808	"ProfessionSE"	0.002280984
"No_of_90dpd_L12M"	0.2102182	"Presence_of_open_AL"	0.001561284
"No_of_60dpd_L6M"	0.2063916	"Type.of.residenceOthers"	0.0006247534
"Total_No_of_Trades"	0.1865789	"gender_dummy"	0.0003349801
"No_of_60dpd_L12M"	0.1855611	"Type.of.residenceLiving.with.Parents"	0.0001625963
"No_of_Inq_ex_HLAL_L6M"	0.1829944	"marital_status_dummy"	0.0001066424
"woe.No_of_trades_opnd_L6M.binned"	0.1686347	"ProfessionSE_PROF"	7.825915e-05
"No_of_90dpd_L6M"	0.1604187	"Type.of.residenceRented"	4.557127e-05
"No.of.months.in.current.residence"	0.09278225	"EducationMasters"	3.465606e-05
"Type.of.residenceOwned"	1.275249e-06		

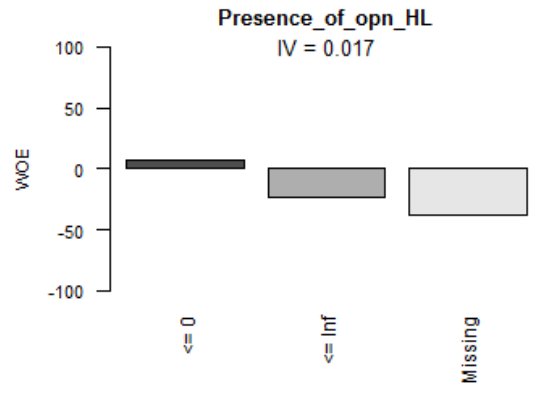
Data Preparation

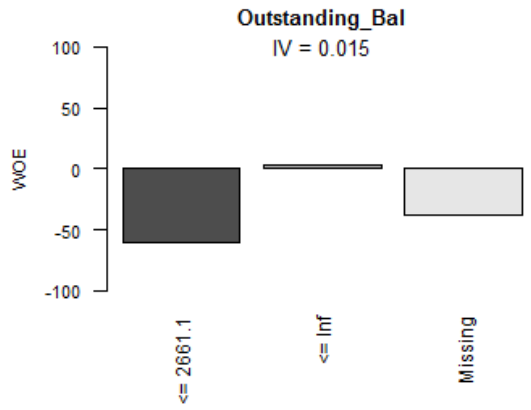
Missing value treatment in Demographic Data	- The data points with missing values in the demographic data were dropped from the data set as they were only around 1% of the data set.
Erroneous Value Treatment	- The erroneous data points like, the negative value on income, age less than 18 , duplicate application IDs were removed from the data set as they were around 1% of the data.
Missing value treatment in Credit-bureau Data	- The data points with missing values in the credit-bureau data were replaced by WOE values.

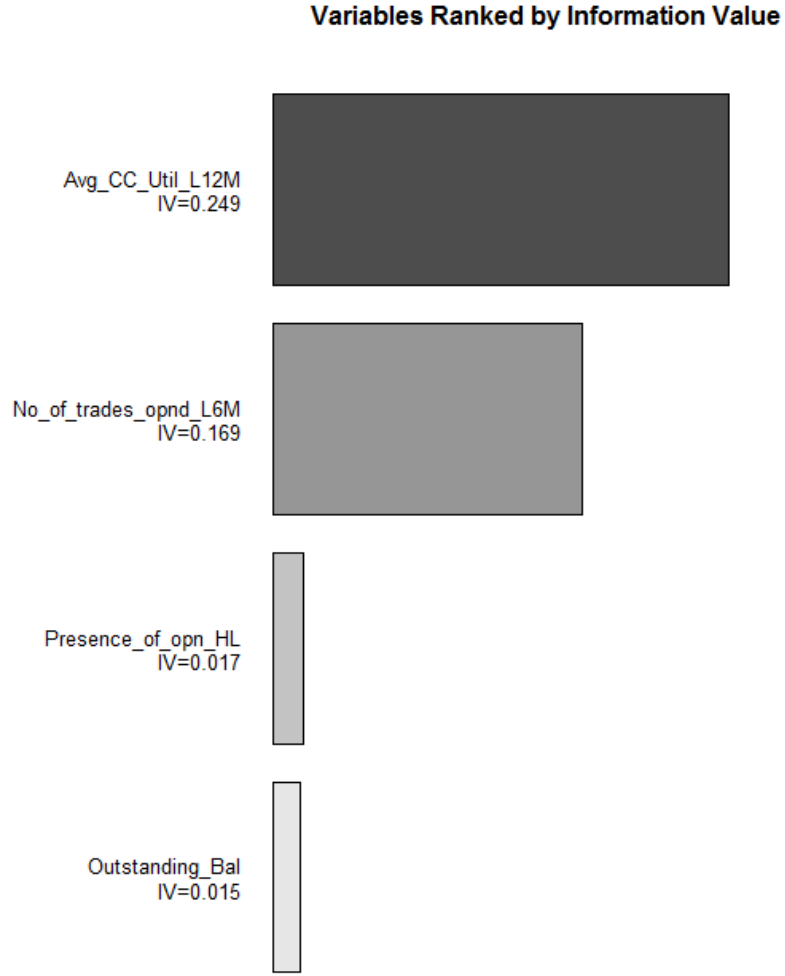
Woe Binning for variables with missing values











Imbalance in the Data-set

- The data set has an inherent imbalance in the Performance Tag variable (3:66), the two categories available(Defaulters and non-defaulters) are to be predicted
- Due to defaulting being a rare case event, there is so much skewness in the dataset.
- This imbalance would impact the training of a model as the entropy would not change much when splitting on a variable that is a predictor of the defaulters case.
- Accuracy can not be used to evaluate the model as majority of the data is biased towards non-defaulters case.
- The combination of Sensitivity and Specificity or F1 score can be used to evaluate a model.

Overcoming Imbalance

The imbalance in the data-set is overcome using the following strategy:

- The data chunk of the non-defaulters is separated from the defaulters case.
- The non-defaulters data set is clustered into ‘n’ clusters.
- A random sample of datapoints is grabbed from each cluster formed, grabbing as many data-points as the defaulters case contained, from each cluster, hence down scaling the non-defaulters data-set.
- The defaulters datapoints are then replicated ‘n’ number of times to match the count of data points for non-defaulters case, hence upscaling.
- The distribution of defaulters and non-defaulters are now 1:1

Future Roadmap

Binary classification algorithms needs to be used:

- Logistic regression
- Random Forest

The model evaluation metric used has to be sensitivity/specificity or F1 score, accuracy cannot be used to judge the performance of a model due to skewness of the data.

A proper application score needs to evaluated based on the probability value predicted by the binary classification algorithms.