CredX Capstone Project Presentation

Batch : Sep-2018

Gaurav Jawrani
Gopinath B
Haritha Appam
Tamilvanan Rajendran

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.

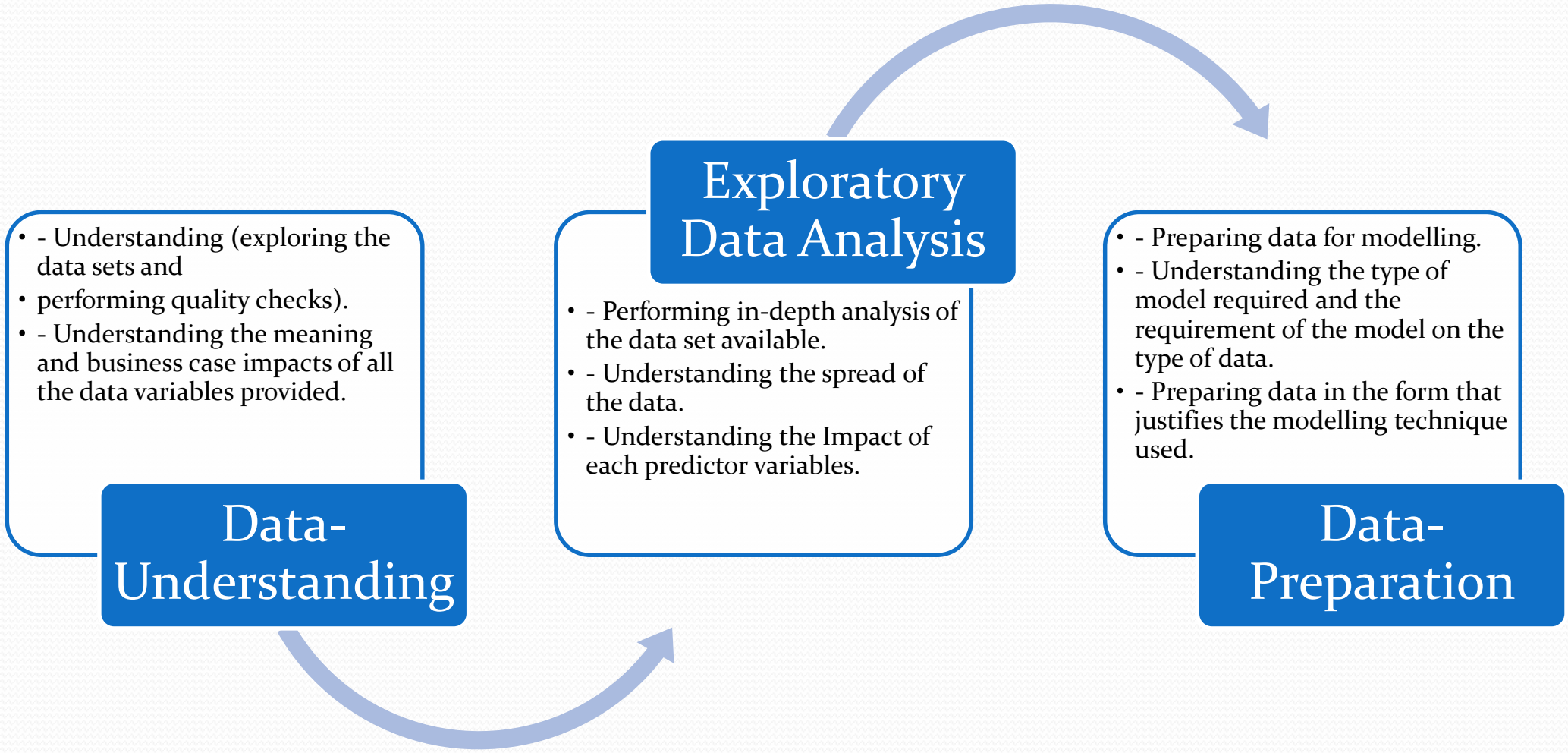
In this project, we will help CredX:

- 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

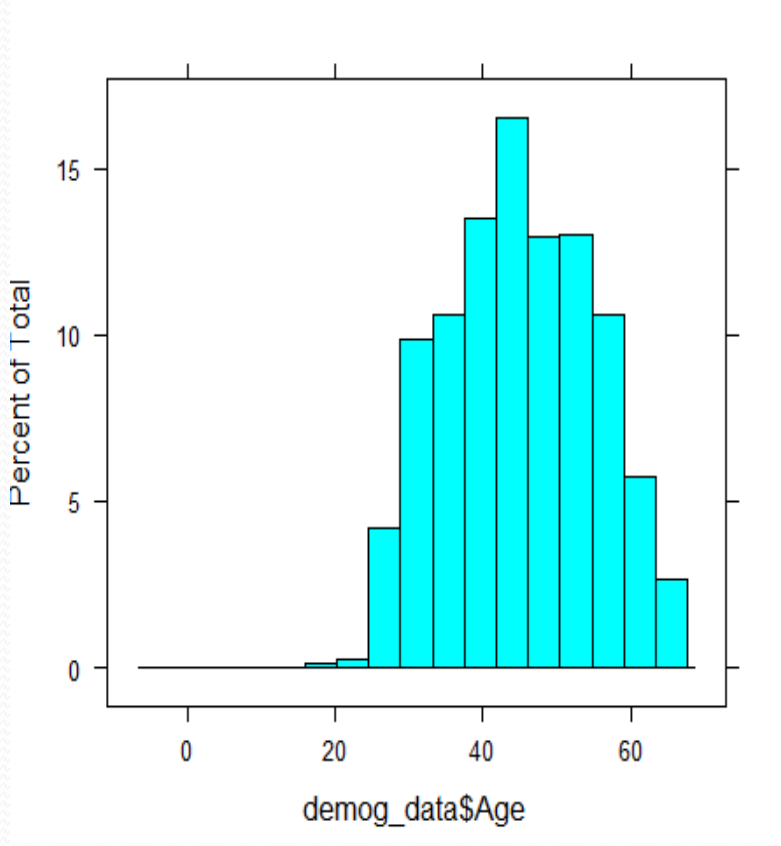


Data Understanding and EDA Summary

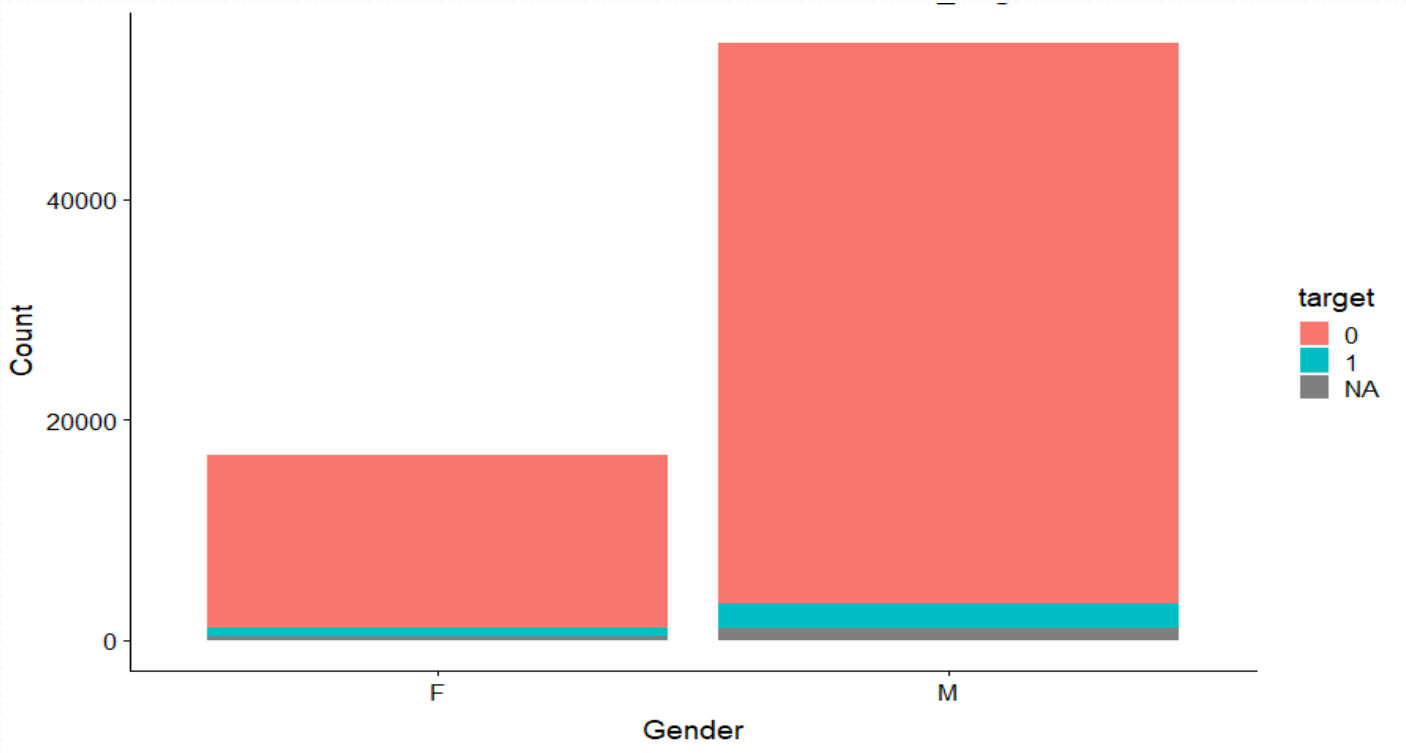
- Data is available in a structured format(csv), in two files for demographic and credit-bureau data.
- The Grain of data is at an individual customer level.
- There are duplicate Application IDs with different features in both demographic data and credit data.
- There are NA values in 'No. of dependents' 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

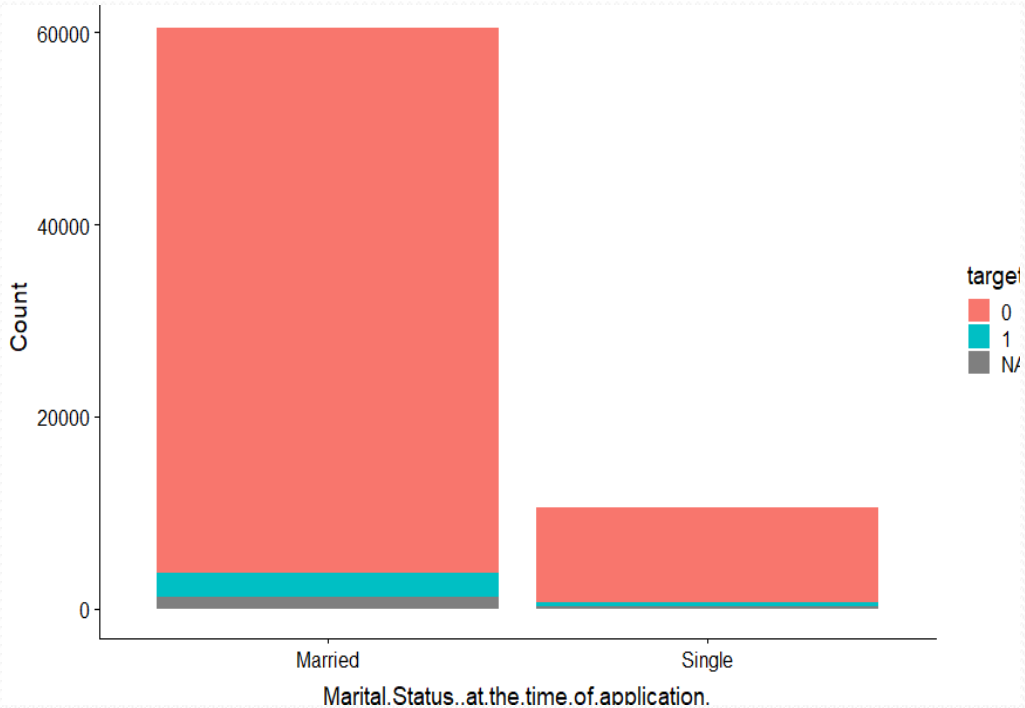


Observation : *Age group between 40-55 tend to default the most*

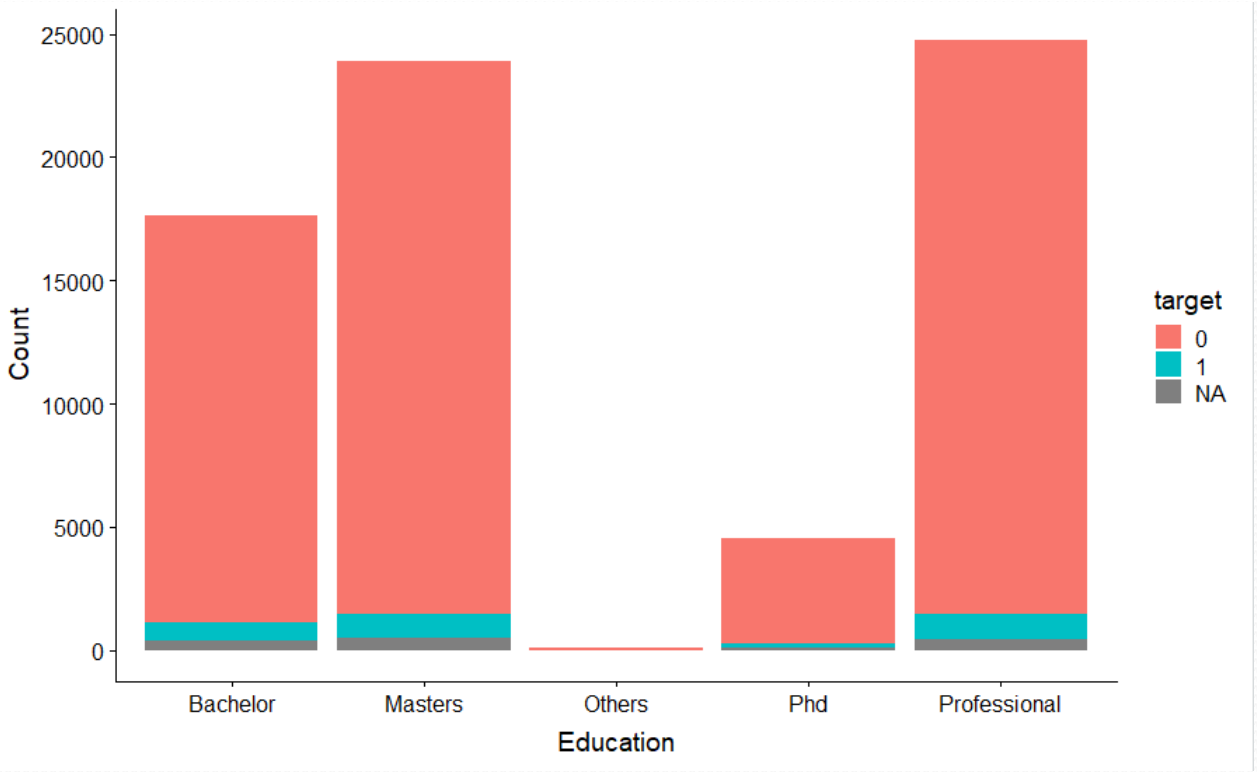


Observation : *Males seems to default more than females.*

Univariate Analysis :

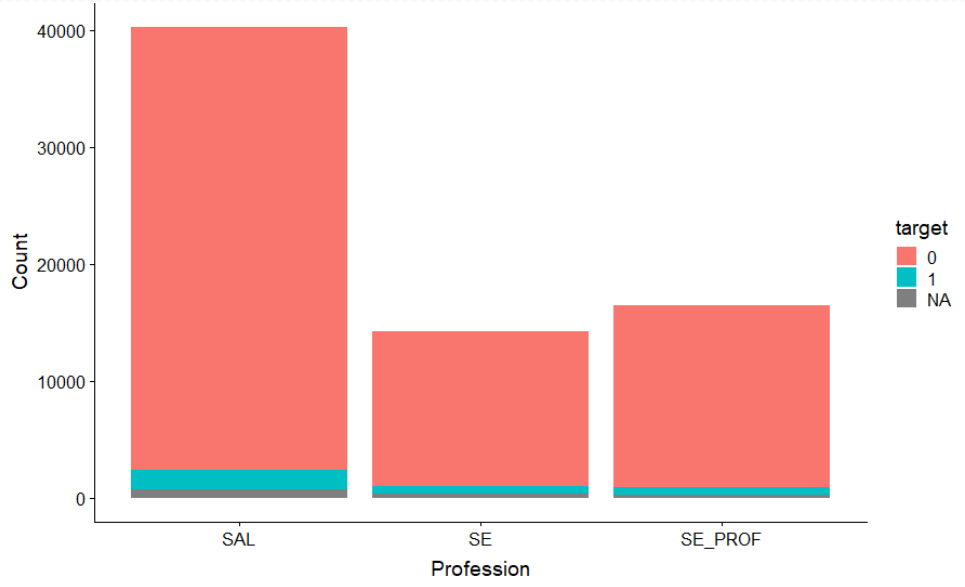


Observation : *Applicants having marital status married has high risk of defaulting*

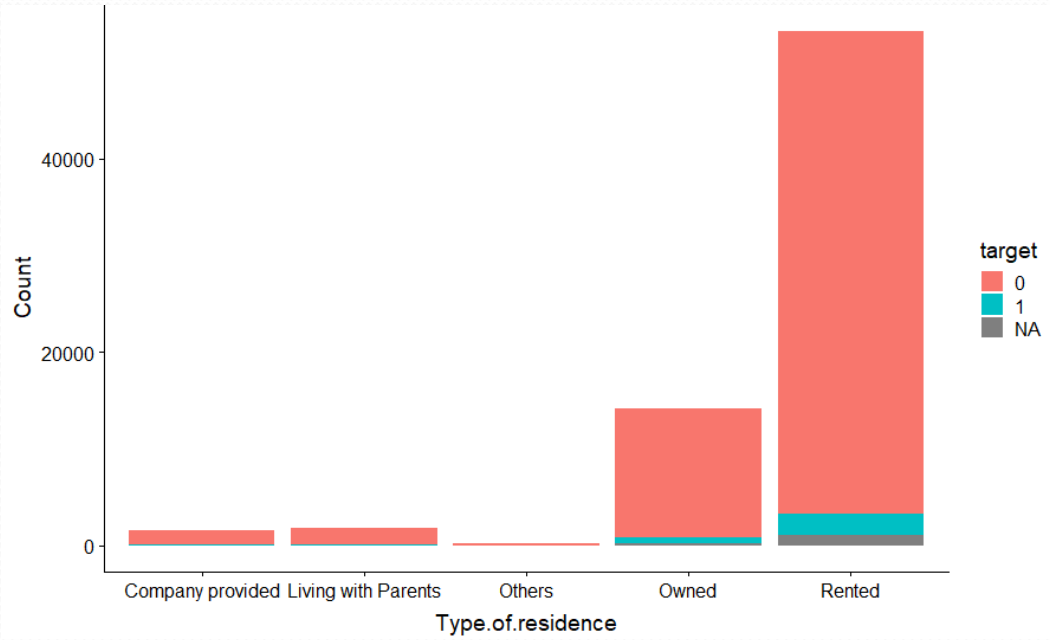


Observation : *Applicants with Masters or Professional Educational qualification has higher risk of defaulting*

Univariate Analysis :

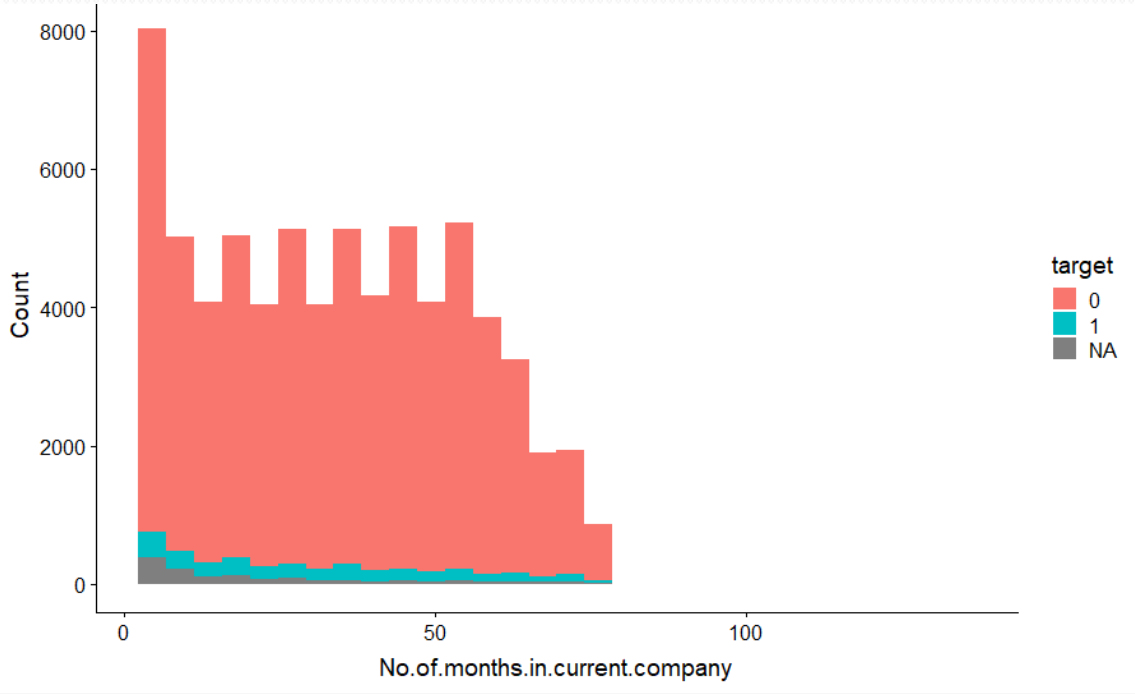


Observation : *Salaried Applicants are the ones who default the most.*

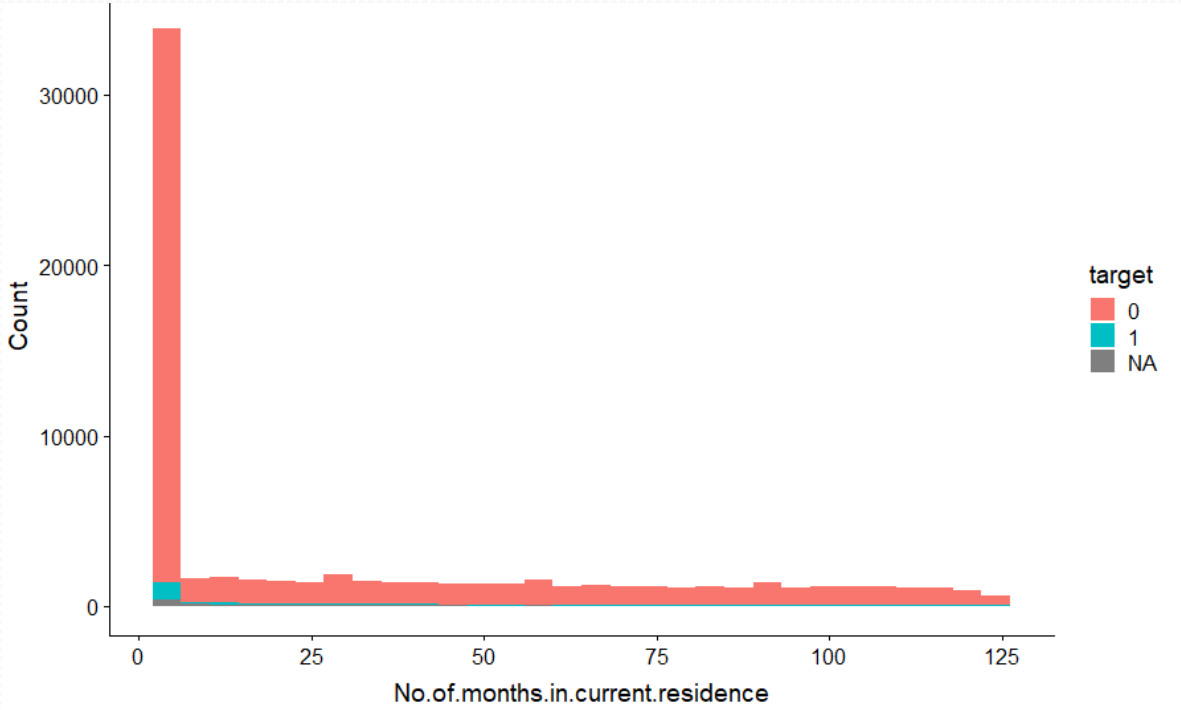


Observation : *Rented ones are having high default chances .*

Univariate Analysis :

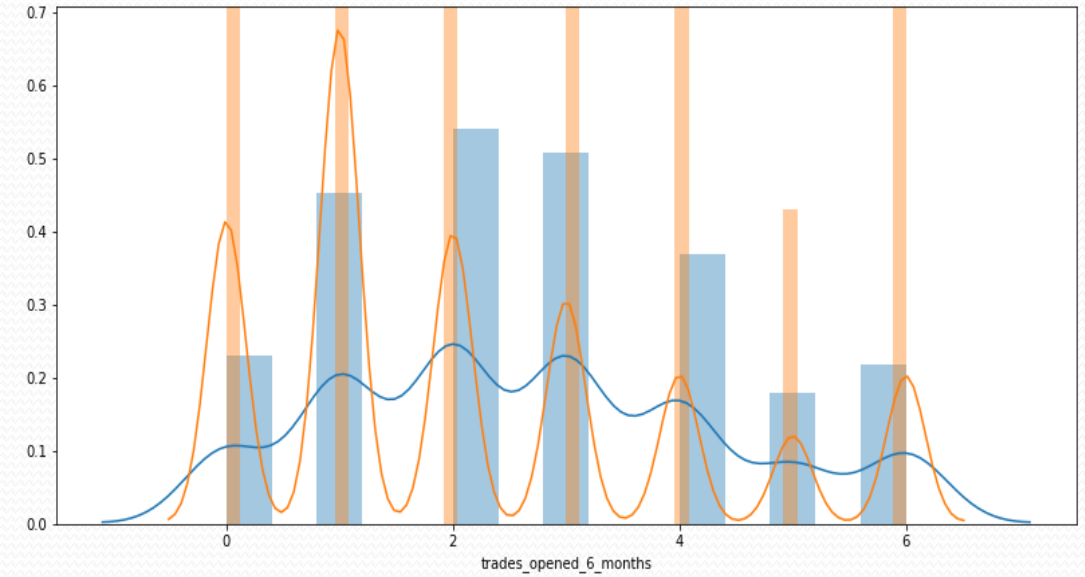


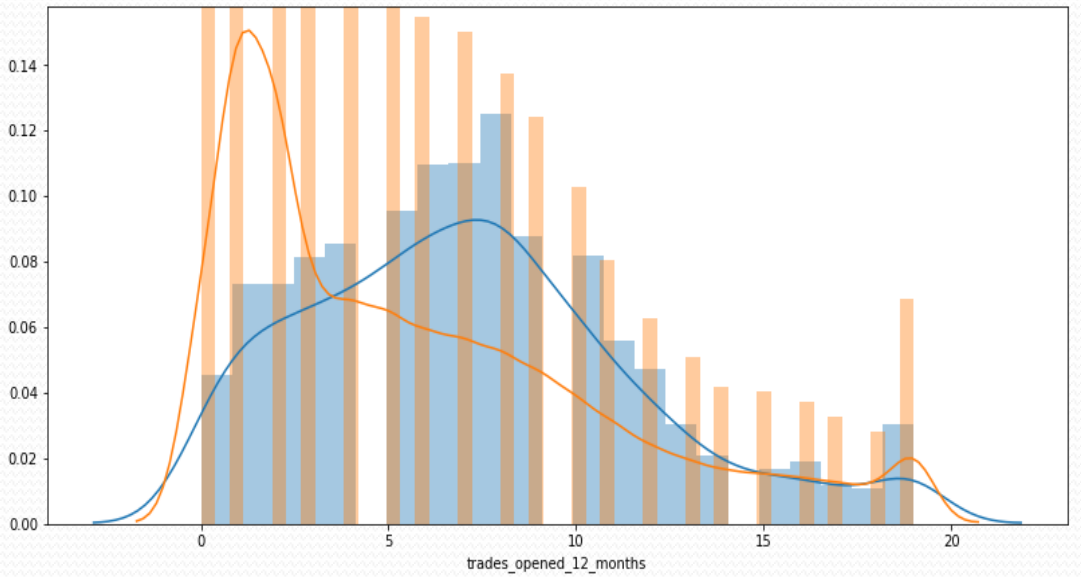
Observation : *Less no of months in current company seems to be more defaulted case.*

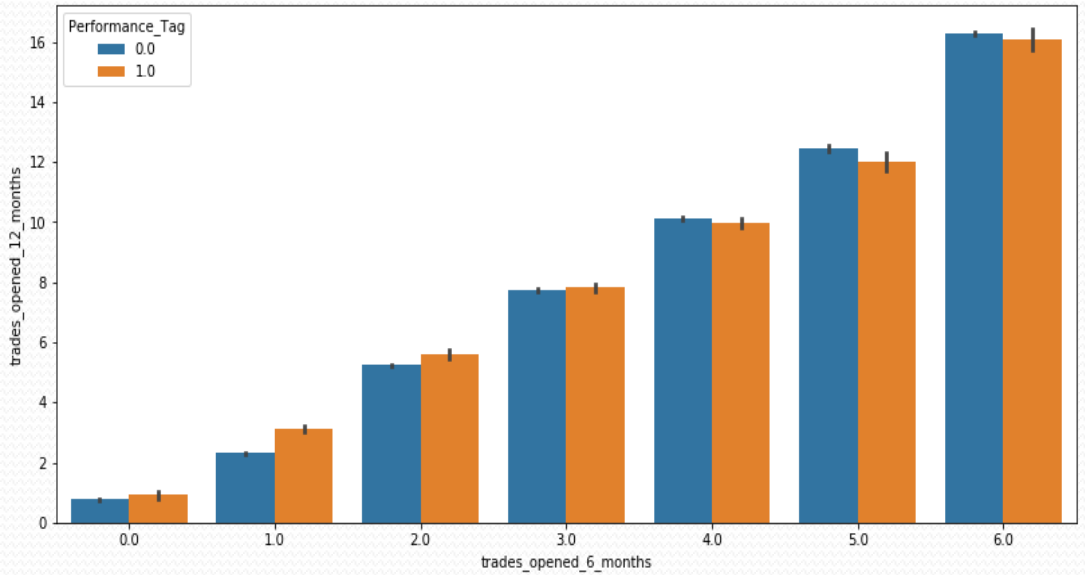


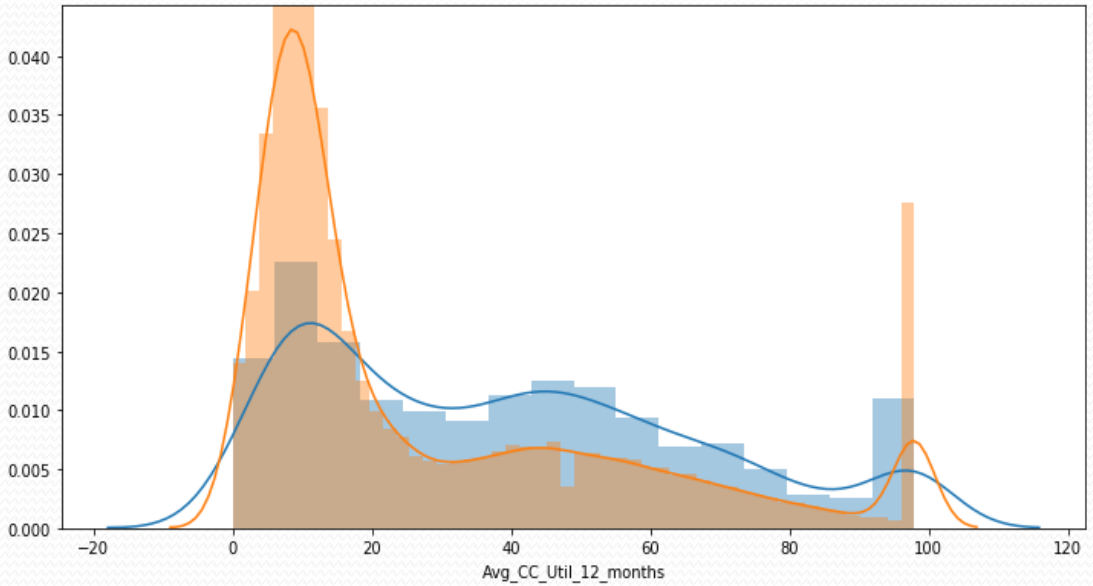
Observation : *Less no of months in current residence seems to be more defaulted case.*

Bivariate Analysis

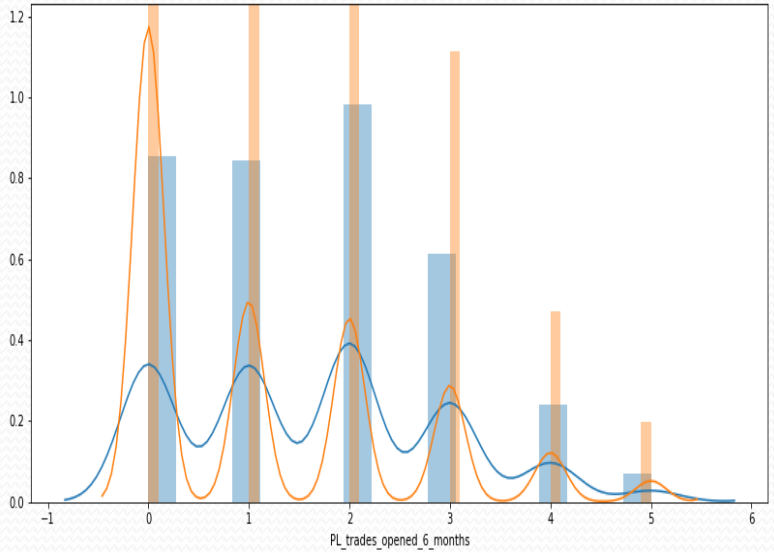


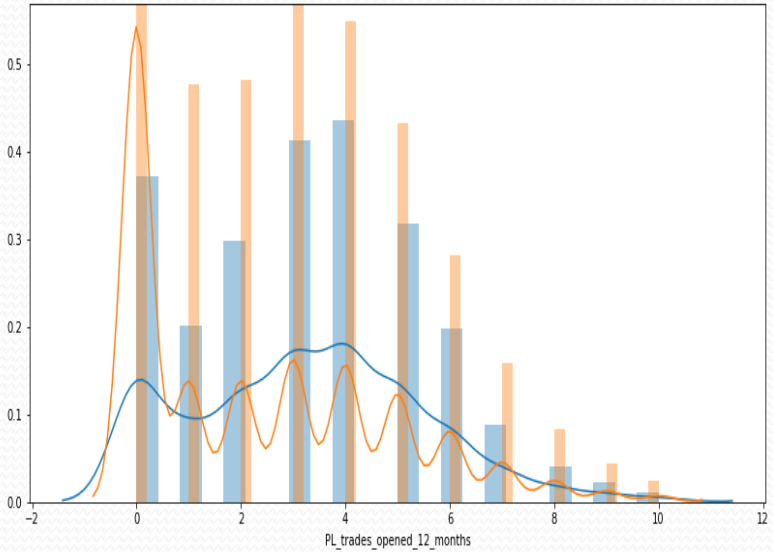


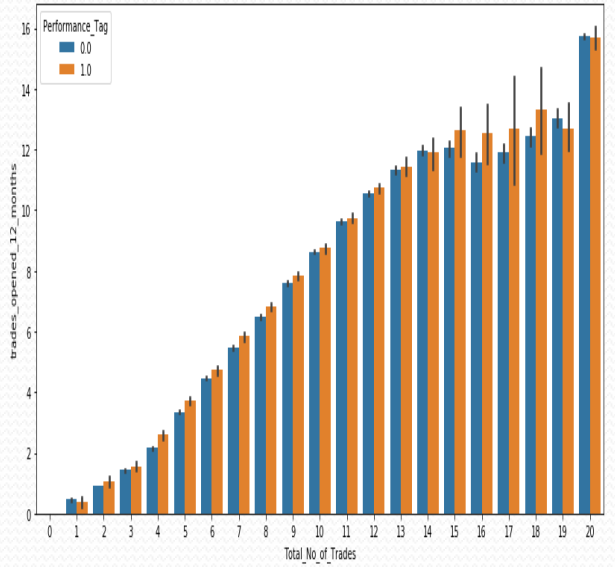


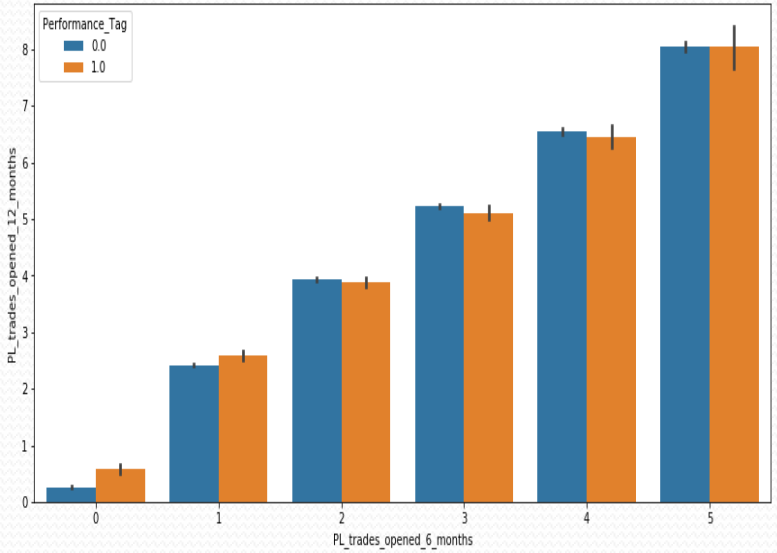


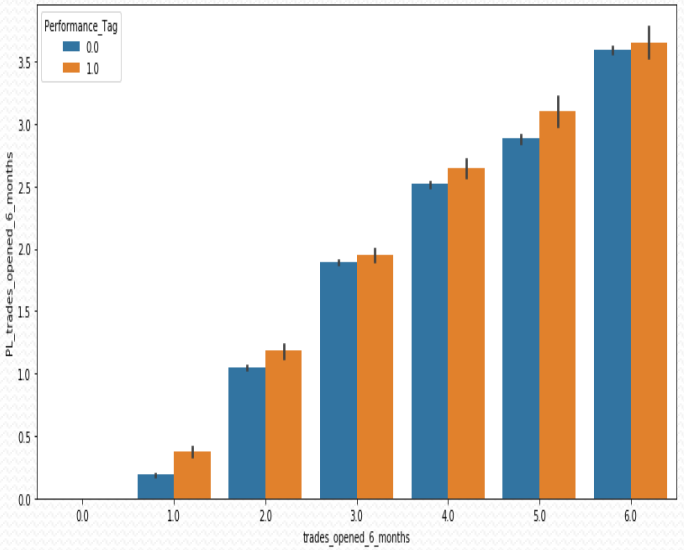
Bivariate Analysis

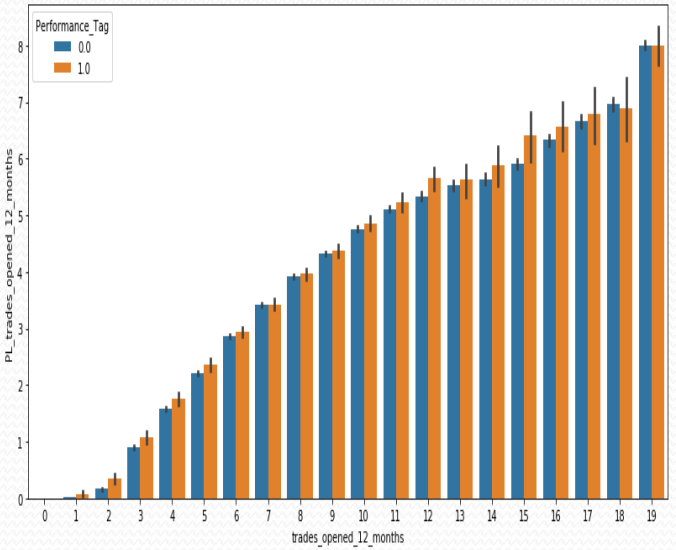




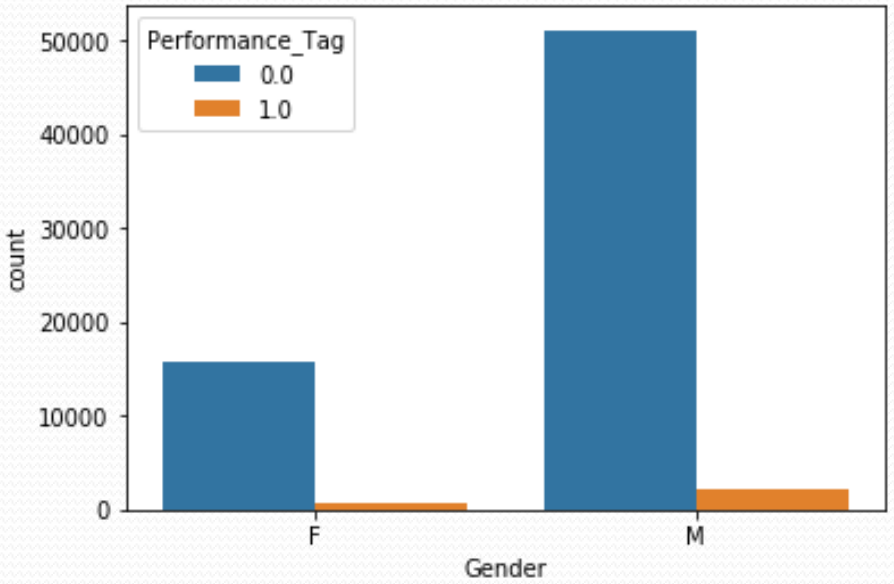


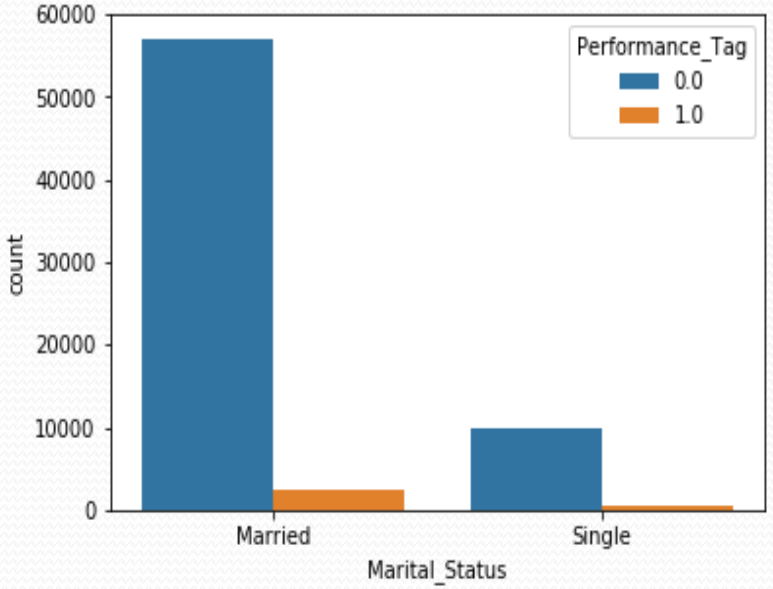


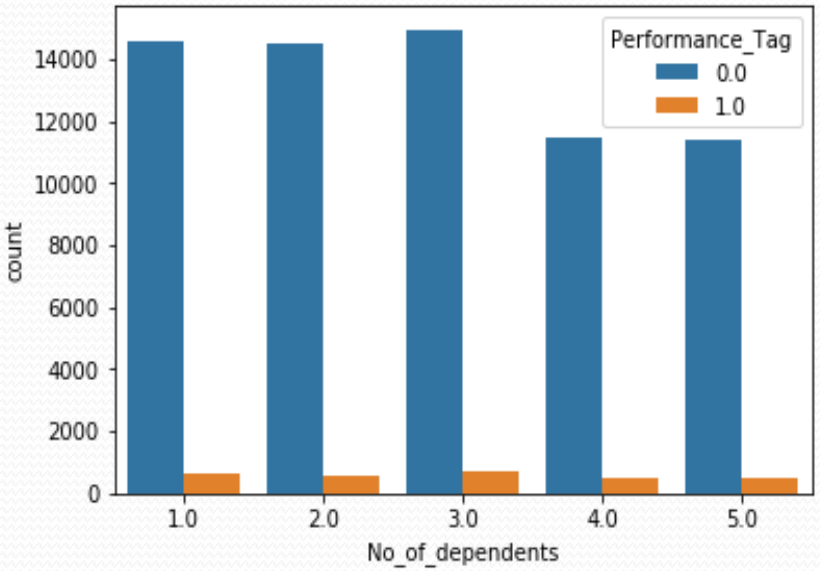


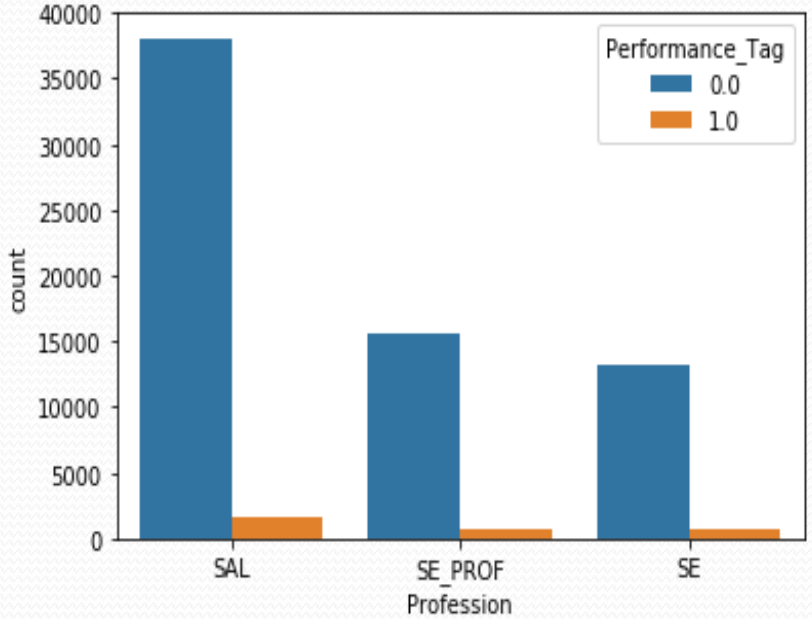


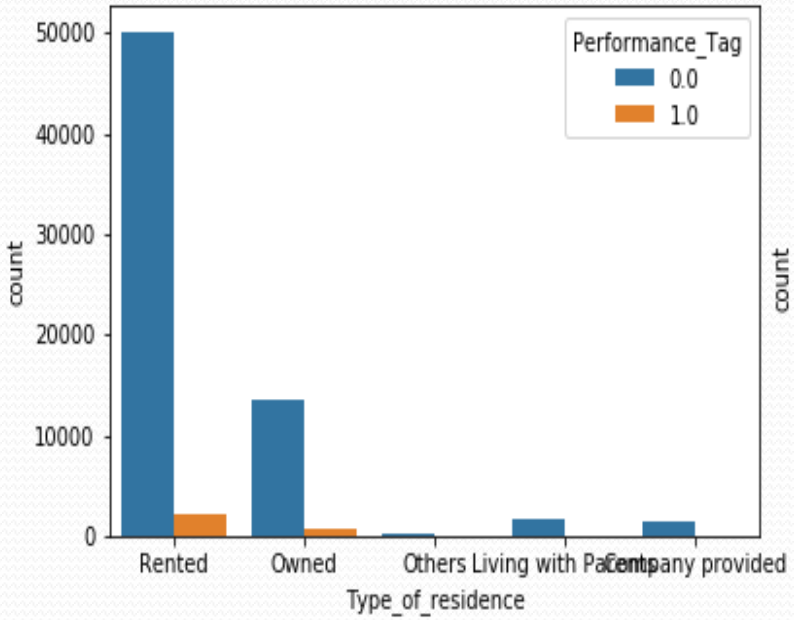
Bivariate Analysis

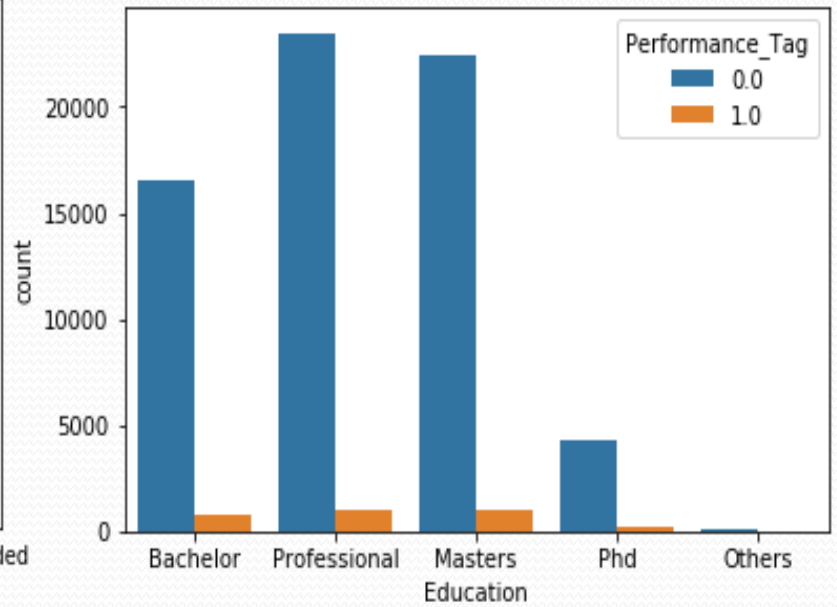












Important Predictors

- The Information Value associated with the predictors are as follows:
- Avg_CC_Util_12_months 0.27101025
- trades_opened_12_months 0.25740911
- Inquiries_12_months 0.22919523
- Total_No_of_Trades 0.189891030
- DPD_12_months 0.18805318
- PL_trades_opened_12_months 0.176626130
- DPD_6_months 0.145723260
- DPD_12_months 0.13768819
- PL_trades_opened_6_months 0.124712490
- DPD_12_months 0.09573226
- trades_opened_6_months 0.09530912
- Inquiries_6_months 0.092925360
- DPD_6_months 0.08959416
- No_of_months_in_current_residence 0.05208810
- Income 0.037622590
- DPD_6_months 0.03074021
- Presence_of_open_home_loan 0.01697915
- No_of_months_in_current_company 0.01275117
- Outstanding_Balance 0.00897122
- Profession 0.00222620
- Presence_of_open_auto_loan 0.00165624
- Type_of_residence 0.0009258
- Education 0.0007826
- Age 0.0006499
- Gender 0.00032713
- Marital_Status 0.00009614
- No_of_dependents 0.000056

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.

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

Regression Analysis

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.

MODEL BUILDING APPROACH

- **OUTLIER TREATMENT:** Outlier detection is done using boxplot on continuous variables and quantiles function and the variables with outliers has been corrected by capping the outliers to the nearest non-outlier values.

- **DATA SCALING:** Scaling is performed for all variables except Application ID and performance tag to standardize the data into common scale.

- **DATA SPLIT:** The final dataset is split into Train and Test in 70:30 ratio for model building.
  - All models are trained on training datasets and regularization was done by tuning of hyper parameters with cross validation on validation datasets.
  - All the models are tested on test datasets that were kept separate from training and validation datasets.

- **DATA SAMPLING:** The given data is highly imbalanced. We have sampled data using ROSE package for balancing the training data sets.

- The cutoff value for the probability of default was chosen such that model evaluation metrics like accuracy ,sensitivity and specificity were almost equal to each other.

- Logistic Regression was built by iteratively removing using these two algorithms

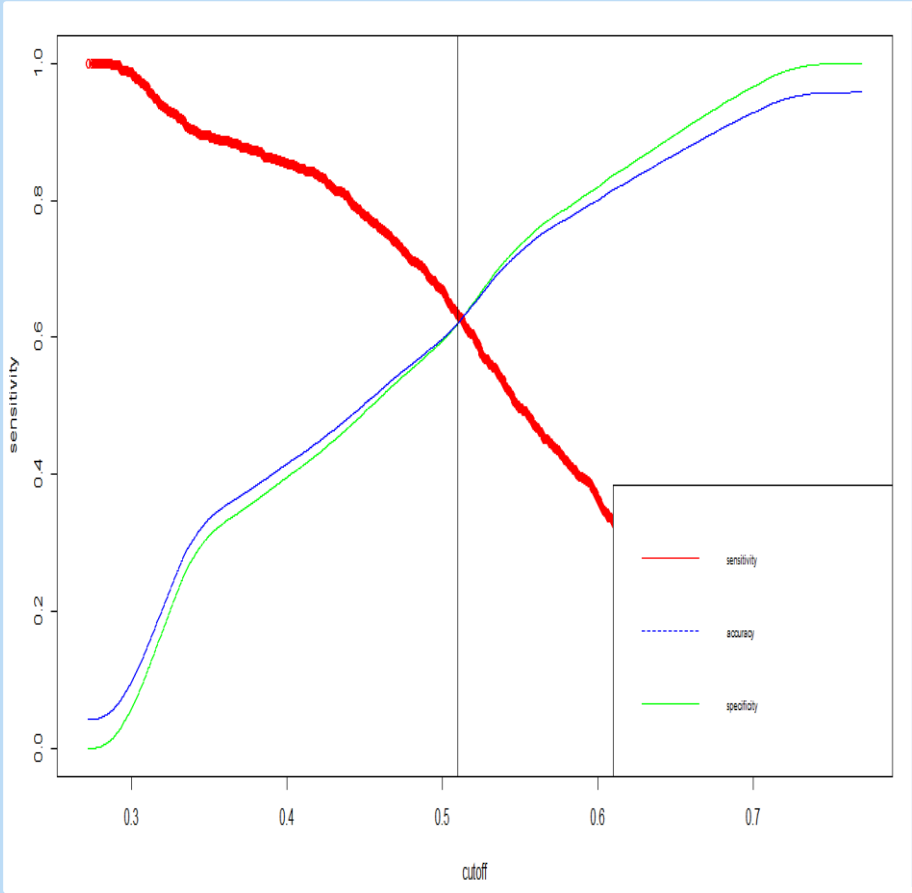
1. Stepwise variable selection based on AIC[using stepAIC()]
2. Backward variable selection based on VIF and p value

LOGISTIC REGRESSION MODEL ON MERGED CREDIT BUREAU AND DEMOGRAPHIC DATASET WITHOUT REJECTED 1425 RECORDS

Predictors in logistic regression model trained on a part of merged credit bureau and demographic dataset (merged on the application id column) without rejected 1425 records which does not have performance tags are as follows :

- INCOME
- NO.OF.MONTHS.IN.CURRENT.RESIDENCE
- NO.OF.MONTHS.IN.CURRENT.COMPANY
- WOE.EDUCATION.BINNED
- AVGAS.CC.UTILIZATION.IN.LAST.12.MONTHS
- NO.OF.TRADES.OPENED.IN.LAST.6.MONTHS
- NO.OF.PL.TRADES.OPENED.IN.LAST.6.MONTHS
- NO.OF.PL.TRADES.OPENED.IN.LAST.12.MONTHS
- NO.OF.TIMES.90.DPD.OR.WORSE.IN.LAST.6.MONTHS
- NO.OF.TIMES.60.DPD.OR.WORSE.IN.LAST.6.MONTHS
- NO.OF.TIMES.30.DPD.OR.WORSE.IN.LAST.6.MONTHS
- NO.OF.TIMES.90.DPD.OR.WORSE.IN.LAST.12.MONTHS
- NO.OF.TIMES.30.DPD.OR.WORSE.IN.LAST.12.MONTHS
- NO.OF.INQUIRIES.IN.LAST.12.MONTHS..EXCLUDING.HOME...AUTO.LOANS.
- PRESENCE.OF.OPEN.HOME.LOAN
- OUTSTANDING.BALANCE
- TOTAL.NO.OF.TRADES

Statistics	Values
Cut-off	0.51
Accuracy	57%
Sensitivity	63%
Specificity	63%



ALL VARIABLES HAVE EXTREMELY LOW P VALUES AND VIF LESS THAN OR ALMOST EQUAL TO 2, HENCE KEEPING ALL VARIABLES ON THAT CRITERIA

LOGISTIC REGRESSION MODEL ON MERGED DATASET WITHOUT REJECTED 1425 RECORDS

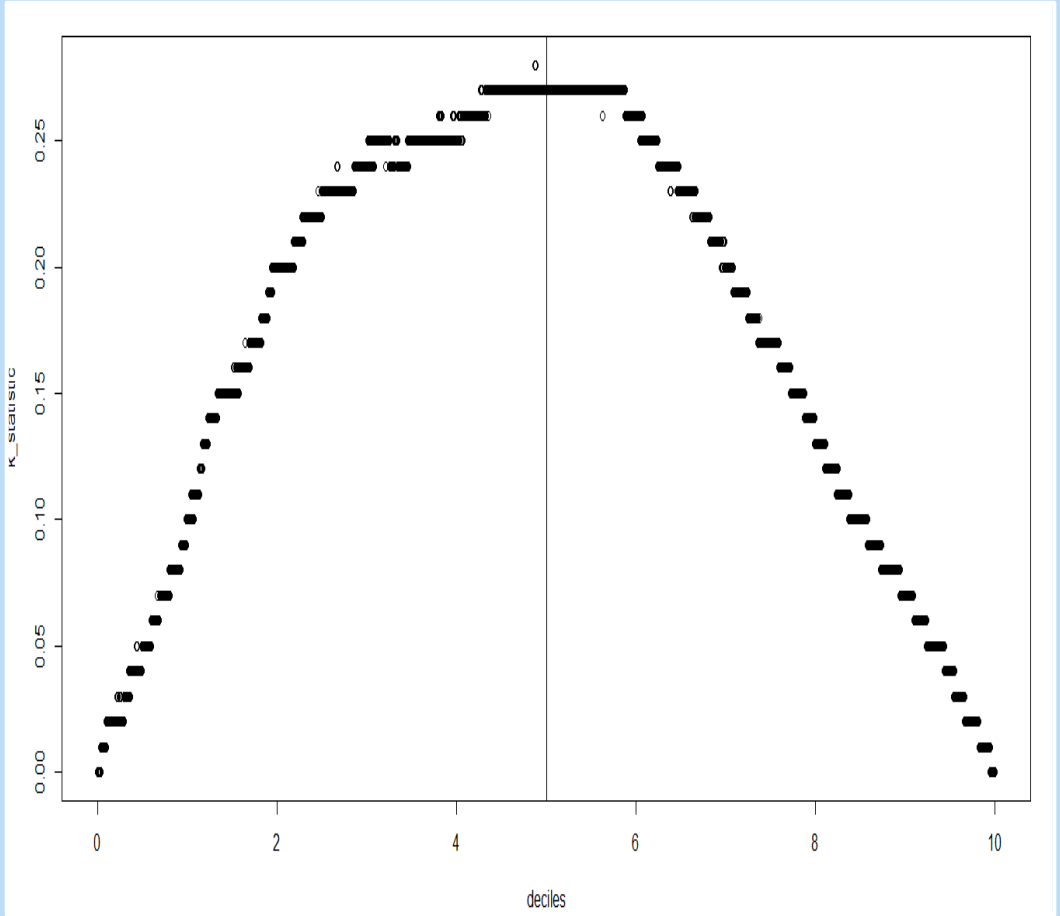
CONFUSION MATRIX AND KS CHART:

Prediction	0	1
0	11405	299
1	8698	556

Statistics	Values
Accuracy	0.5707
Sensitivity	0.62006
Specificity	0.65029

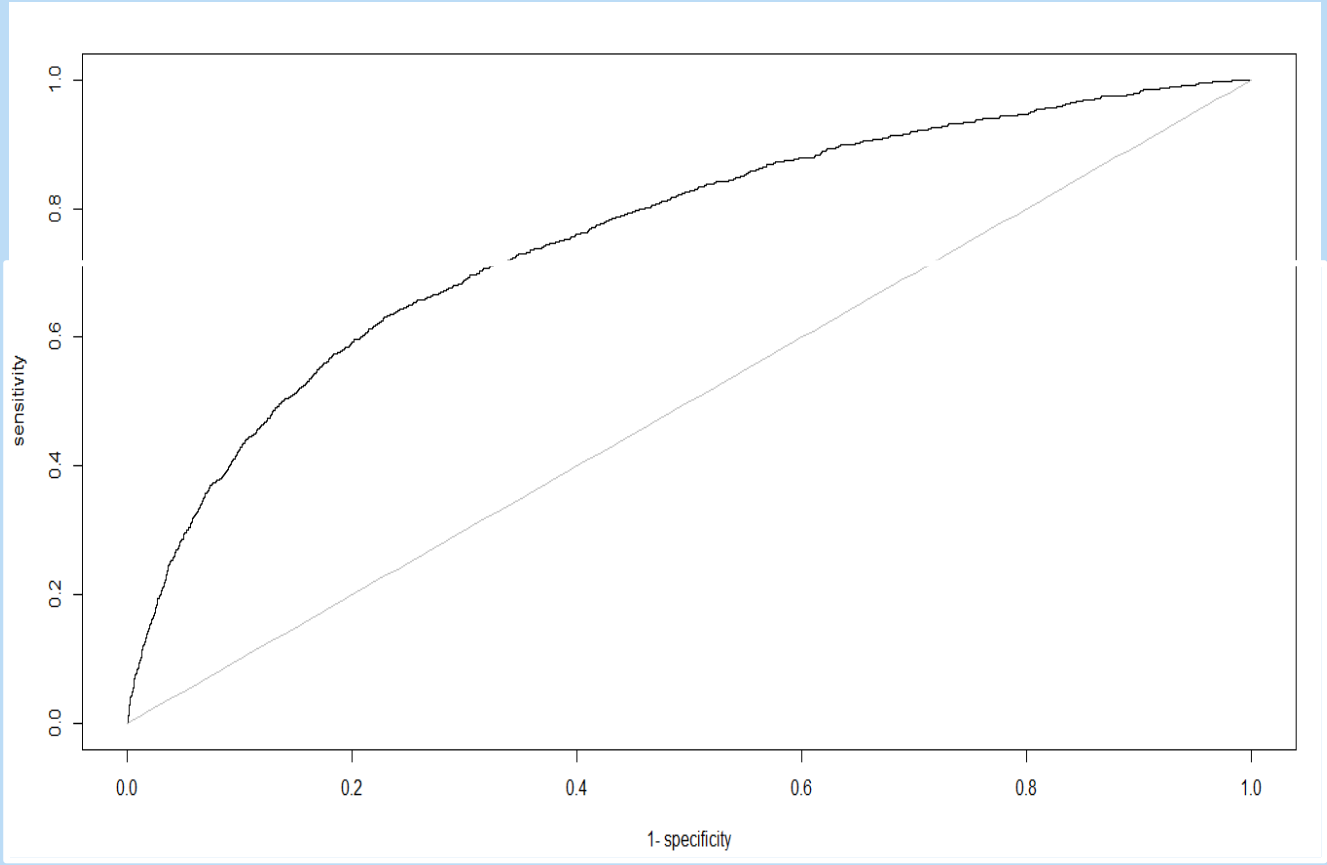
KS STATISTIC FOR THIS MODEL IS 0.27 AND LIES WITHIN IN FIRST 5 DECILES

KS CHART



LOGISTIC REGRESSION MODEL ON MERGED DATASET WITH REJECTED 1425 RECORDS

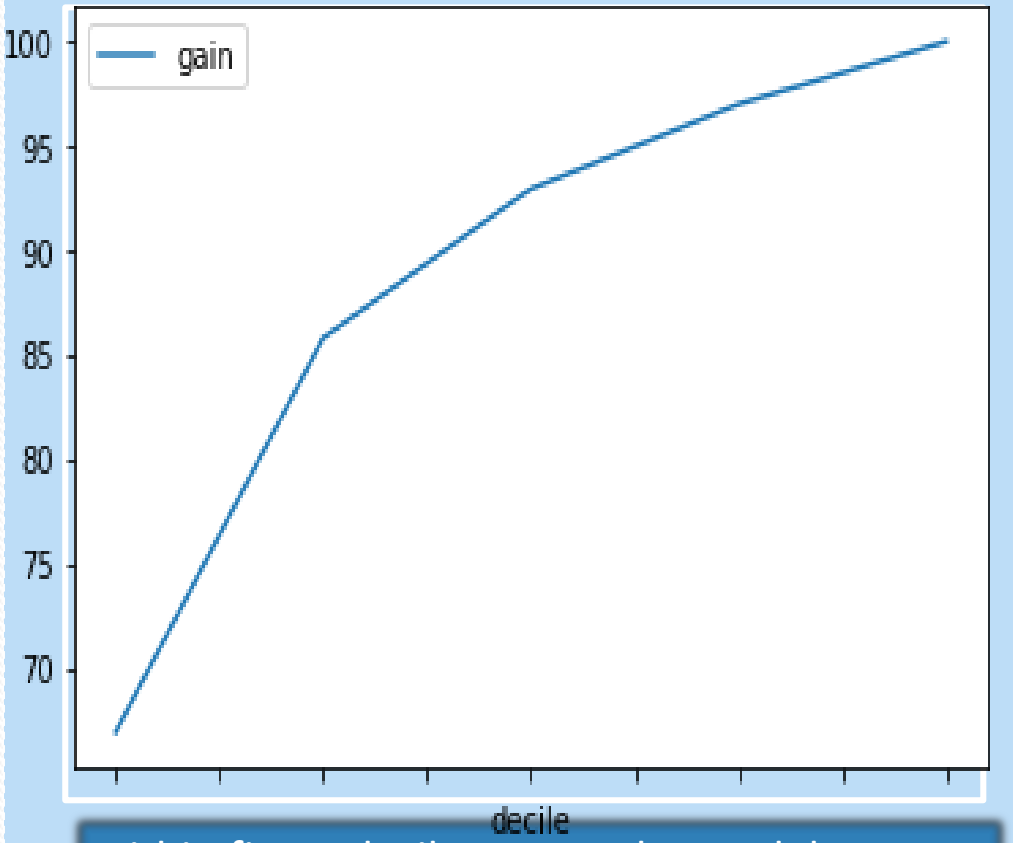
AREA UNDER THE CURVE :



AREA UNDER ROC CURVE = 0.92

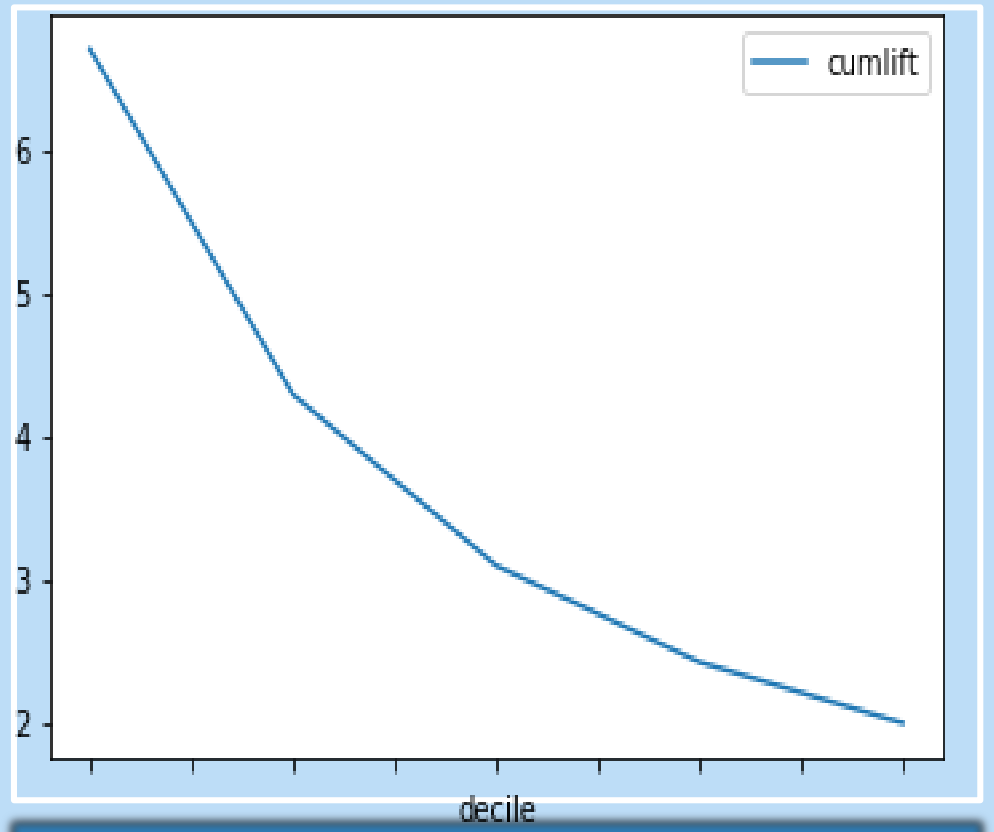
DECISION TREE MODEL ON MERGED DATASET WITH REJECTED 1425 RECORDS

Gain Chart



Within first 4 deciles as per the model we are able to predict 97% of defaulters correctly.

Lift Chart



A lift of 2.8 times is achieved with the model within first 2 deciles compared to randommodel;

DIFFERENT MODEL'S ACCURACY, SENSITIVITY & SPECIFICITY ON MERGED DATASET WITH REJECTED 1425 RECORDS

DIFFERENT MODEL'S ACCURACY, SENSITIVITY & SPECIFICITY ON MERGED DATASET WITH REJECTED 1425 RECORDS WHICH HAVE PERFORMANCE TAGS

Models	Accuracy	Sensitivity	Specificity
Logistic Regression on the Final Merged Dataset (with rejected 1425 records)	57.07%	69.80%	69.14%
Decision Tree on the Final Merged Dataset(with rejected 1425 records)	94.46%	69.63%	69.44%

CONCLUSION: FOR MERGED DATA WITH PERFORMANCE TAG MISSING RECORDS, DECISION TREE MODEL IS PERFORMING BETTER COMPARED TO RANDOM FOREST. HENCE CONSIDERING DECISION TREE MODEL AS FINAL MODEL FOR APPLICATION SCORECARD.

Application Scorecard

Final application scorecard was made using the Decision Tree model on the entire dataset which also contained predictions for missing values in "Performance Tag" in 1425 records.

The Decision Tree model was chosen since its evaluation metrics were comparable to other models as well it's an easily interpretable simple model.

The scorecard was made using the following steps:

1. Application score card was made with odds of 10 to 1 being a score of 400. Score increases by 20 points for doubling odds.
2. Probability of default for all applicants were calculated
3. Odds for good was calculated. Since the probability computed is for rejection (bad customers), Odd(good) = (1-P(bad))/P(bad)
4. ln(odd(good)) was calculated
5. Used the following formula for computing application score card:
   400 + slope * (ln(odd(good)) - ln(10)) where slope is 20/(ln(20)-ln(10))
   Where, slope=20/(log(20)-log(10))

Summary of application_score_card values:

- Scores range from 248 to 333 for applicants with median score being 309.
- Higher scores indicate less risk for defaulting

Application Scorecard

CUTOFF SCORE FOR ACCEPTING OR REJECTING AN APPLICATION

- Cutoff selected for probability of default for Decision Tree model was 0.55
- CUTOFF_SCORE= 400 + (slope * (log((1-0.55)/0.55) - log(10)))
- CUTOFF SCORE is equal to **340**
- No.of applicants above score 340 and thus their credit card application will be accepted as per our model is 358
- No.of applicants below score 340 and thus their credit card application will not be accepted as per our model is 20600

Financial Benefits of the Model

The Confusion Matrix for calculating the Financial gain using our model was made on the dataset without missing Performance tag records, since we need to evaluate how much gain was achieved using our model for applicants who were provided with credit card compared to when no model was used.

	Reference	
Prediction	**0**	**1**
0	19771	332
1	829	26

Profit calculations – with model Vs without model

•We have considered an average profit of Rs.5000 from each non defaulters and
• an average loss of Rs.1,00,000 when each accepted applicant defaults

•Net Profit without model = Rs 3.9665 crores

•Profit using model will be total profit due to each true positive and each true negative minus loss from each false positive and each false negative prediction
•Profit with model = Rs15.6865 crores

•Net financial gain with using our model = **Rs. 11.72 crores**
•Percentage financial gain = **295.47%**

Revenue loss and Potential Credit loss saved

Revenue Loss : Occurs when good customers are identified as bad and credit card application is rejected.

- No of candidates rejected by the model who didn't default – 20600.
- Total No of candidates who didn't default – 66853
- % of good candidates rejected by our model – 31.38%

- About 31.38% of the non defaulting customers are rejected which resulted in revenue loss.

Credit Loss Saved : The candidates who have been selected by the bank and have defaulted are responsible for the credit loss to the bank.

- % of candidates approved and then defaulted when model was not used = 4.2%
- % of candidates approved and then defaulted when model was used = 1311/69799 = 1.8%

- Credit loss saved => 4.2 – 1.8 = 2.4%

Conclusion

- Decision Tree model is chosen as the final Model with 94% of Accuracy.
- Optimal score cut-off value of 340 is derived to approve and reject the applications.
- By this we found out that credit loss % was decreased when we used this model. Hence it is accurate in rejecting the candidate who may default in future.
- There is Net Financial gain of 295.47% after using the model.