credX

Capstone BFSI Case Study

Final Submission

Date: 16th September 2019

Submitted by - ***Srinivasan. G, Naveed J, Denny J and Abhishek C***

*Mentored by – **Nikhil Mehrotra***

Problem Statement & Objective

- A Credit card provider company CredX who issue credit card to its customers based on the number of applications they receive every year. However, the company experienced in recent years, there were increase in credit loss as it perceived that it is due to acquiring 'bad' customers.
- The objective of this project, to mitigate the credit risk, by predicting the right customer during acquisition of new customers by using predictive models, based on the past data of the applicants who are seeking for credit cards.
- Also, the objective to identify the factors which are influencing the credit risk and create strategies to mitigate the risk and assess financial benefit.

Methodology & Solution Approach

Data Understanding
- Observing the data and its anomalies
- Renaming long column names

Data Cleansing
- Dropping Null Values and Duplicate values rows, Treating Outliers
- Dropped Null Values are been kept aside later for scorecard calculation
- Merging the Credit Bureau and Demographic data

Data Analysis
- Univariate and Bi-Variate Analysis
- Key plots Density, Histogram, Box and Violin plots to understand the data distribution of each feature

WOE and IV
- Bin for independent Numerical and Categorical variables and calculate WOE and IV
- Plot for WOE and IV and Identify the top predictive power variables

Develop Model
- Develop Standard Logistics Regression Model for Demo and Demo + Credit Bureau Model
- Develop other Logistic Regression models using Under SMOTE, SMOTE-TOMEK, Random Forests, Decision Tree, SVM
- Develop Logistic Regression model on WOE data

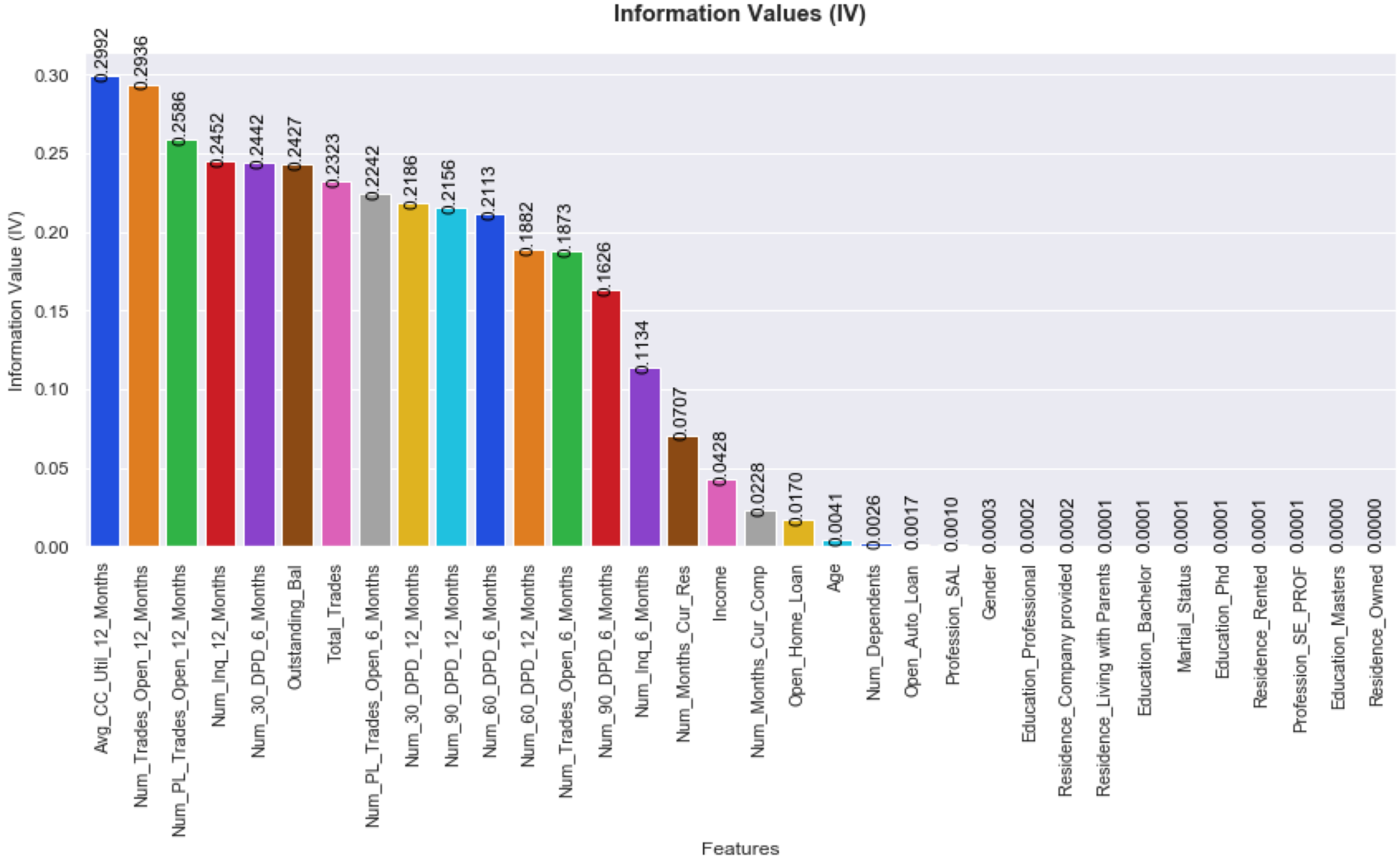
Scores
- Generate Credit card scores and find the cut-off score
- Evaluate for Rejected Population and Original Population

FBA
- Use the cut-off scores to assess the financials from the P&L perspective
- Give recommendations based on the overall analysis and model assessment

Constraints and Assumptions

- Constraints
  - The data is highly imbalanced. Only 4.2% of data is about defaulters. We would use sampling techniques such as SMOTE and SMOTE Tomek while generating the models
- Assumptions
  - The missing data/outliers/invalid data treatment has been done either by replacing median values/limit values.
  - Rejected application data is only used for scorecard verification. NOT for EDA/modelling.
  - It is assumed that all the outstanding balance for the defaulted users would be lost and attributed as credit loss (Written off) for final financial strategy.

Important Variables (IV) Analysis

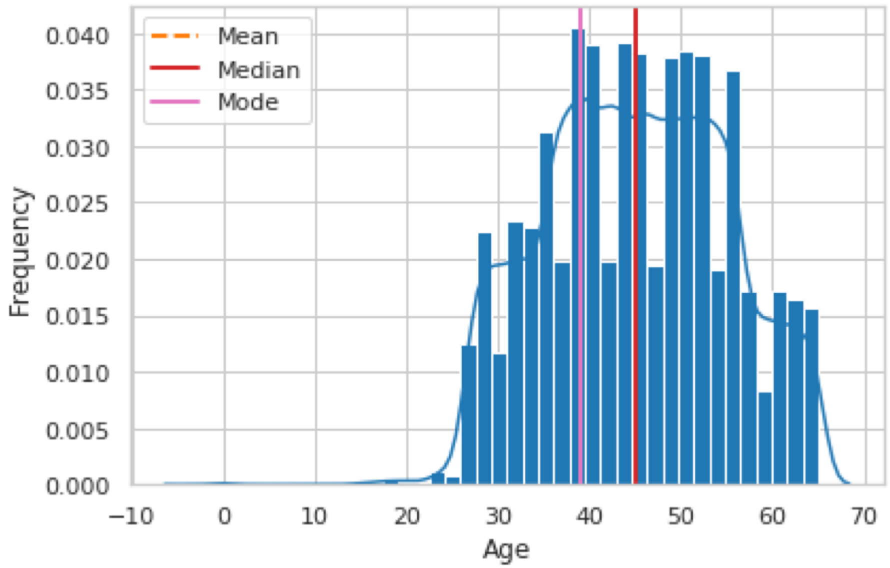


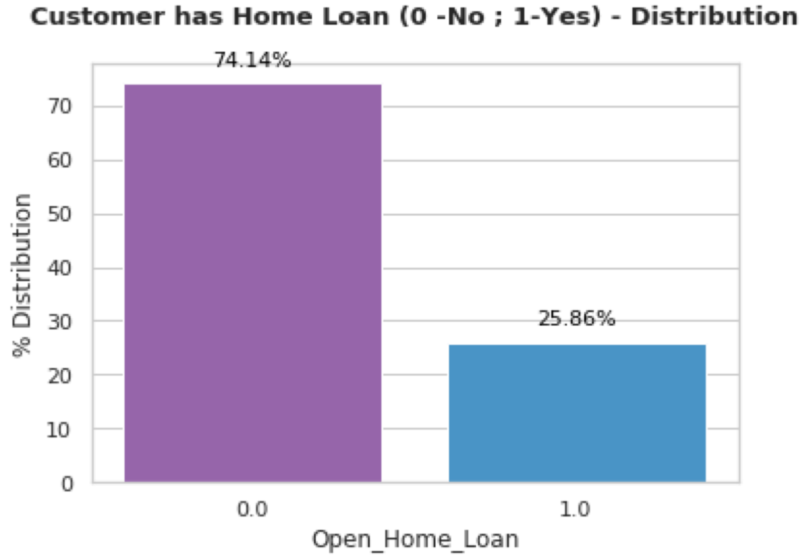
Information Value	Predictive Power
<0.02	Useless for prediction
0.02 to 0.1	Weak predictor
0.1 to 0.3	Medium predictor
0.3 to 0.5	Strong predictor

S.NO	FEATURE
1	Avg_CC_Util_12_Months
2	Num_Trades_Open_12_Months
3	Num_PL_Trades_Open_12_Months
4	Num_30_DPD_6_Months
5	Num_Inq_12_Months
6	Outstanding_Bal
7	Total_Trades
8	Num_PL_Trades_Open_6_Months
9	Num_30_DPD_12_Months
10	Num_90_DPD_12_Months

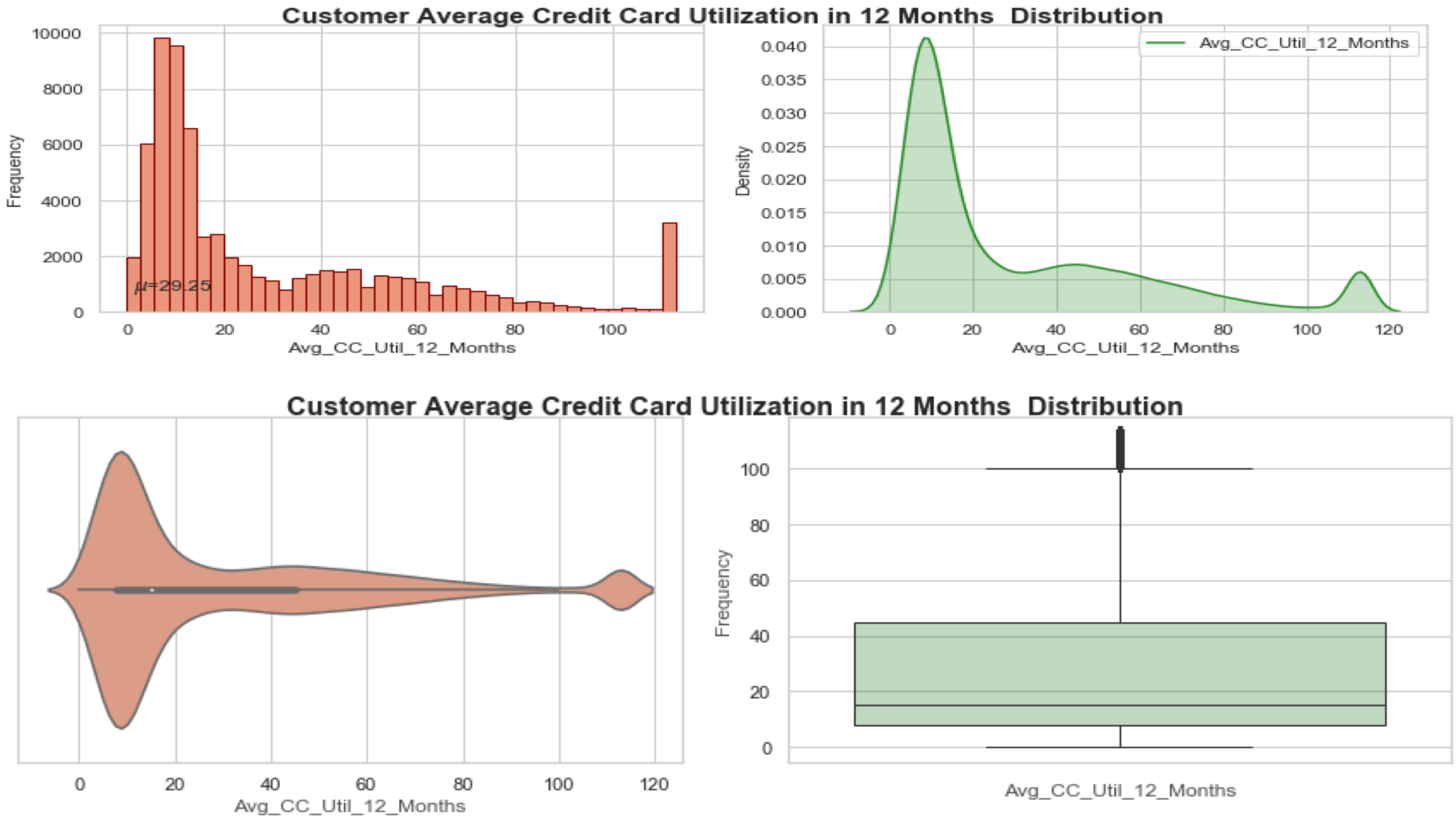
EDA Insights

- Univariate
  - Shows that the cards were issued to individuals in the right age group, and income brackets, and employed by a company for more than 3 months
  - Also, that nearly 26% of the defaulters opened a home loan, whereas 9% had opened an auto loan

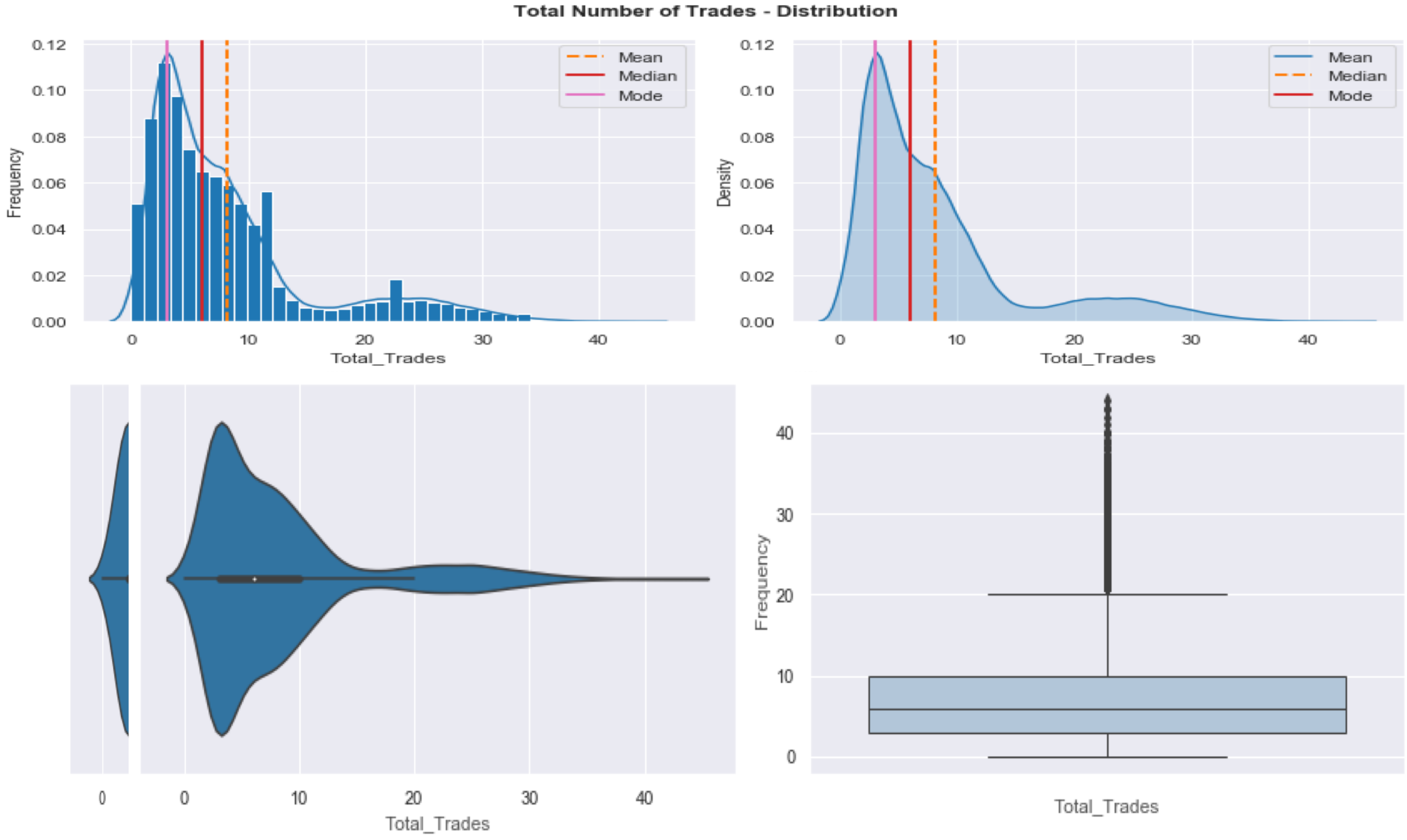




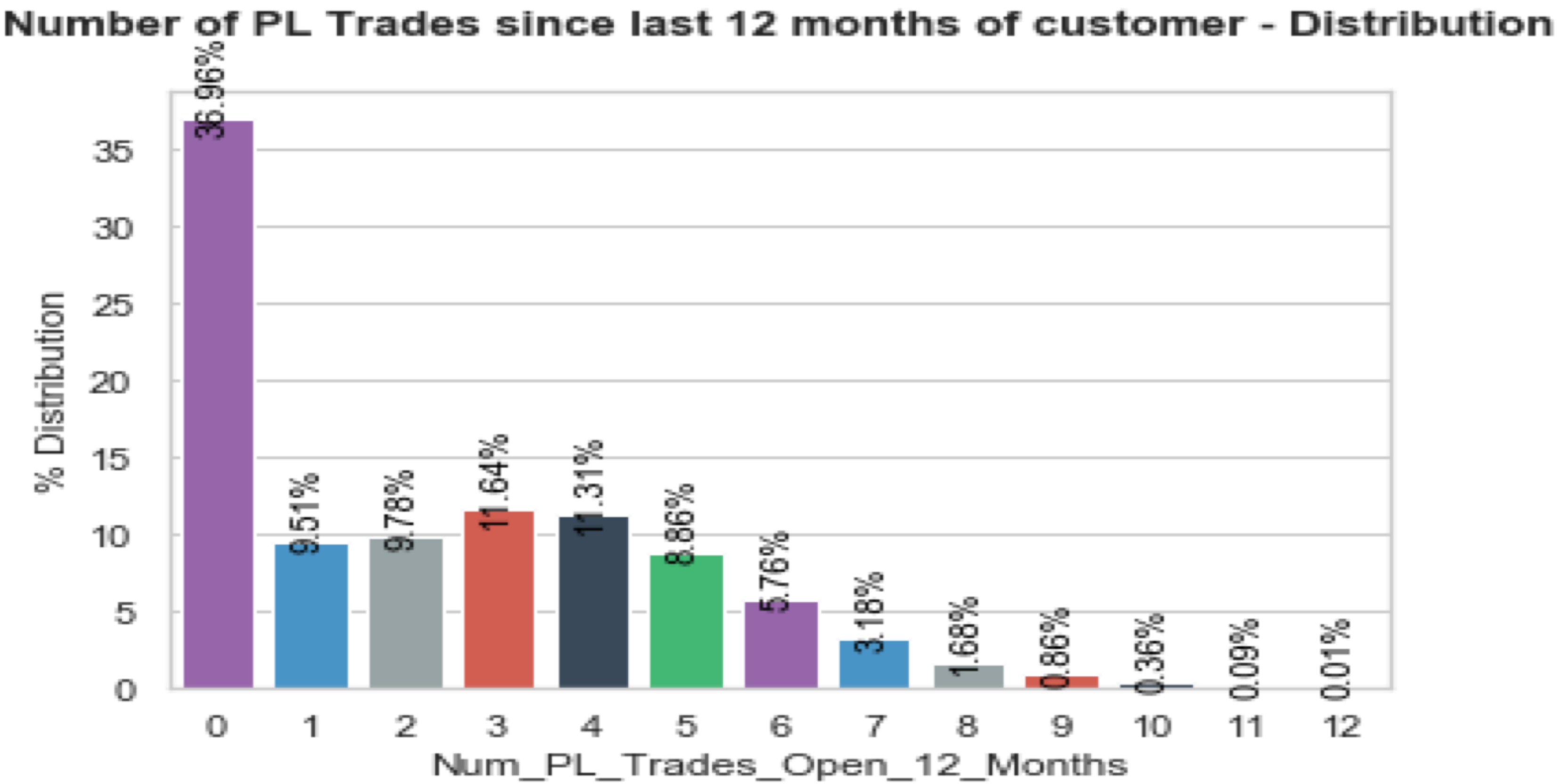
Univariate -Average Credit Card Utilization in 12 months



Univariate - Total Number of Trades Distribution



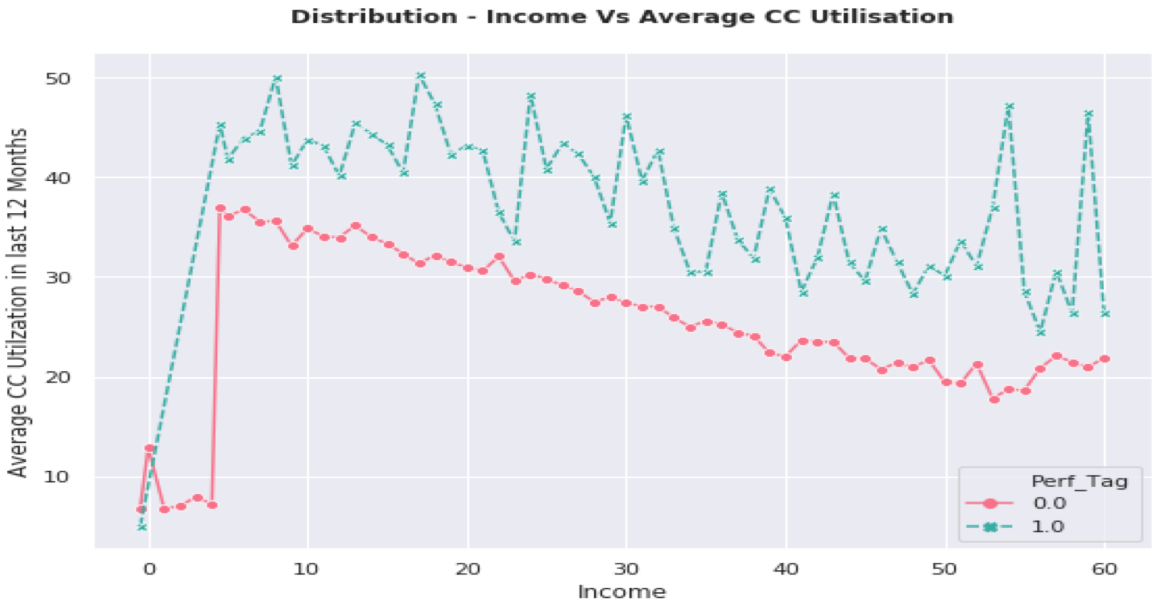
Univariate - Number of PL Trades in 12 months Distribution



EDA Insights

Bivariate

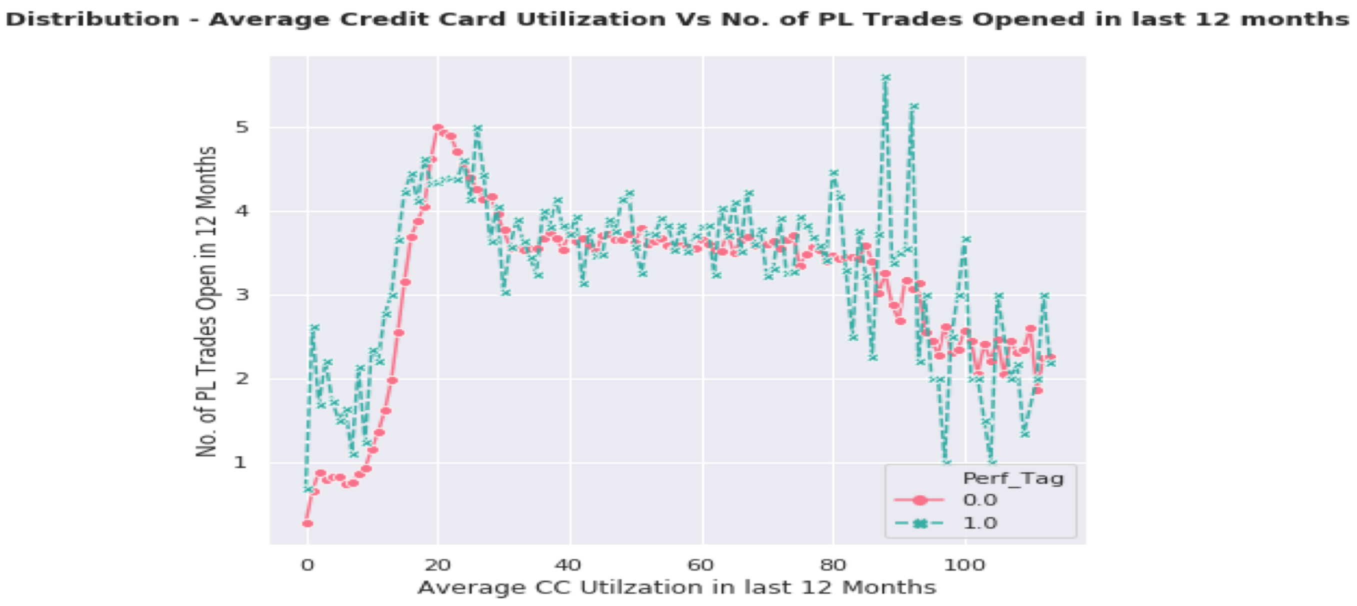
- With Increase in Income the Average Credit Card Utilisation is decreasing with population.
- The average CC usage is greater than 40 for a low income, greater than 30 for middle income and greater than 25 for higher income.
- The defaulters have higher outstanding balance

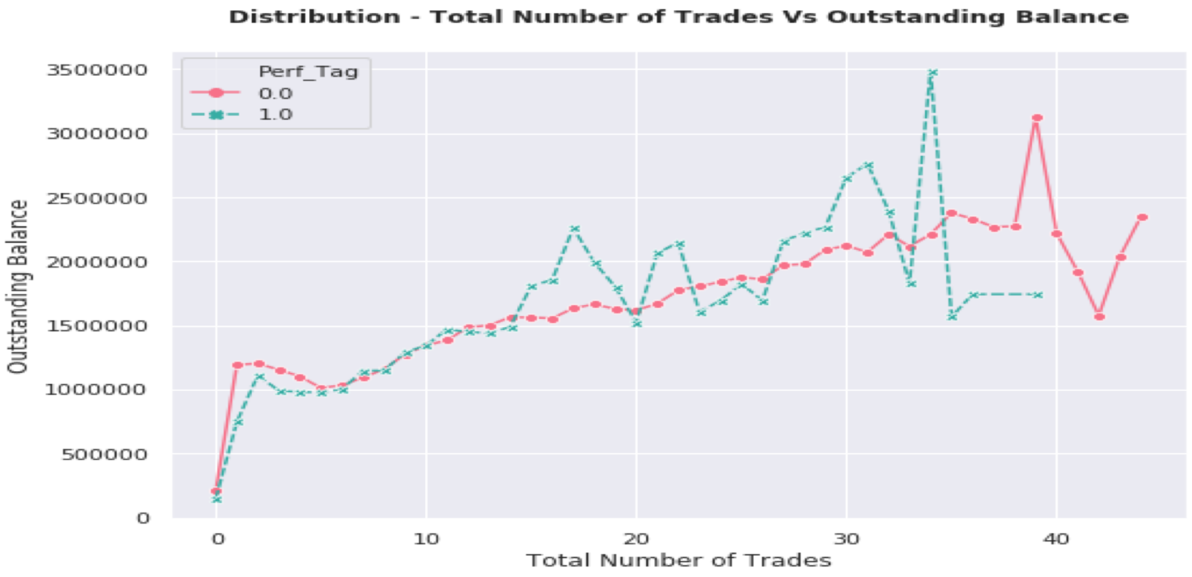


EDA Insights

Bivariate

- The number of PL trades opened in last 12 months is higher for the defaulting customers.
- Overall the total number of trades are higher for the default users. Correspondingly, the outstanding balance also increases gradually increase with number of trades.

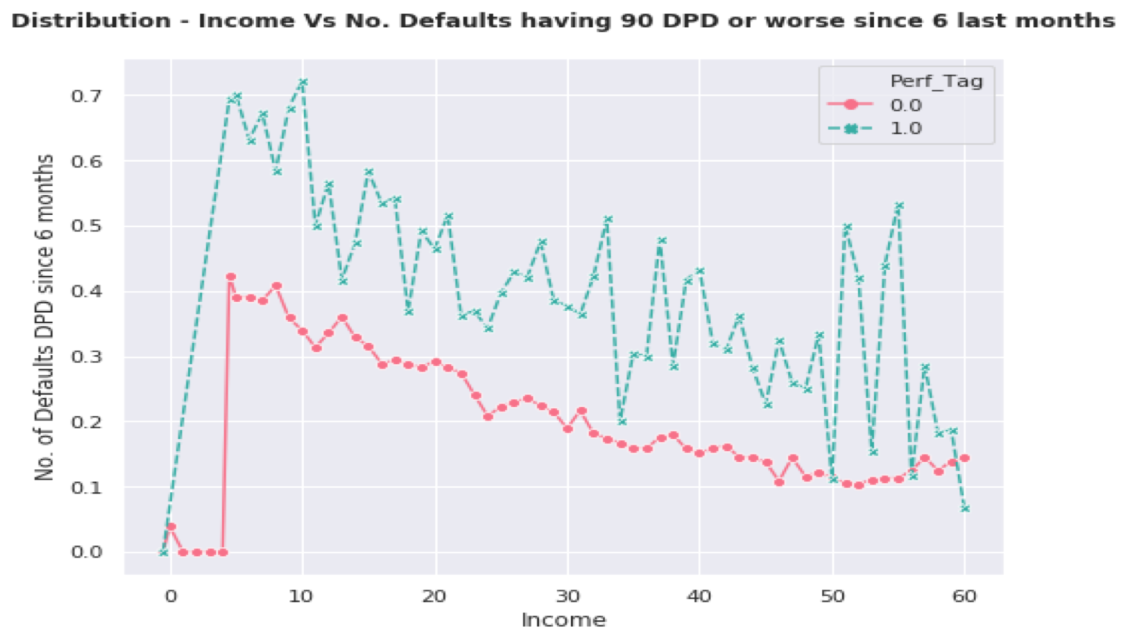


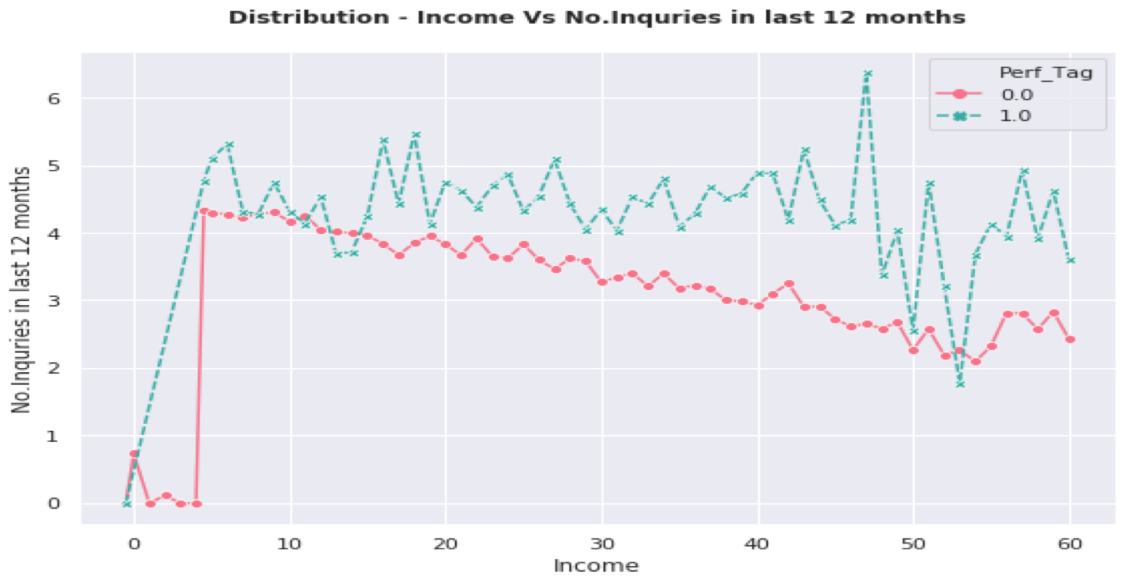


EDA Insights

Bivariate

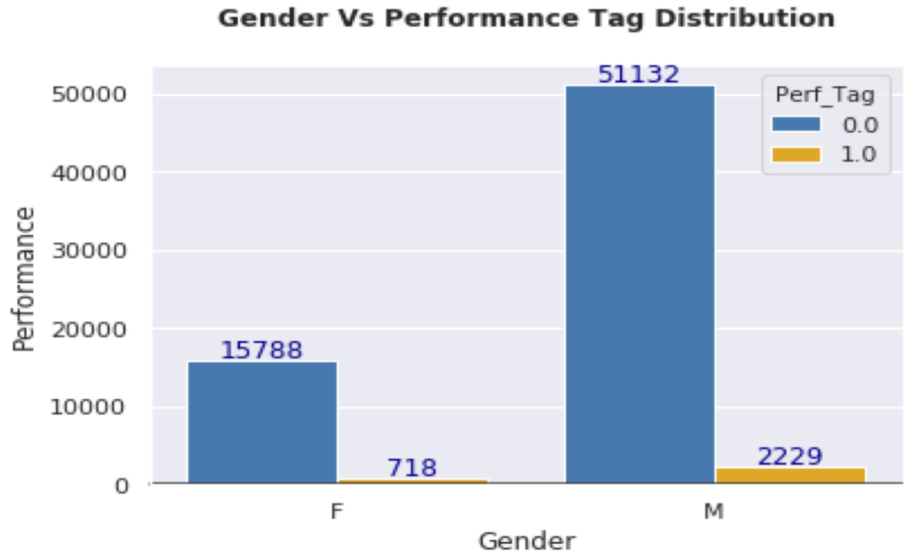
- With increasing Income, DPD numbers are decreasing.
- Also for defaulting users DPD numbers are way higher and higher the numbers of defaulters are in lower to medium income range.
- With increase in income number of inquiries are decreasing for non defaulters. Similarly, the with increase in income, the number inquiries are also in increase for default customers.

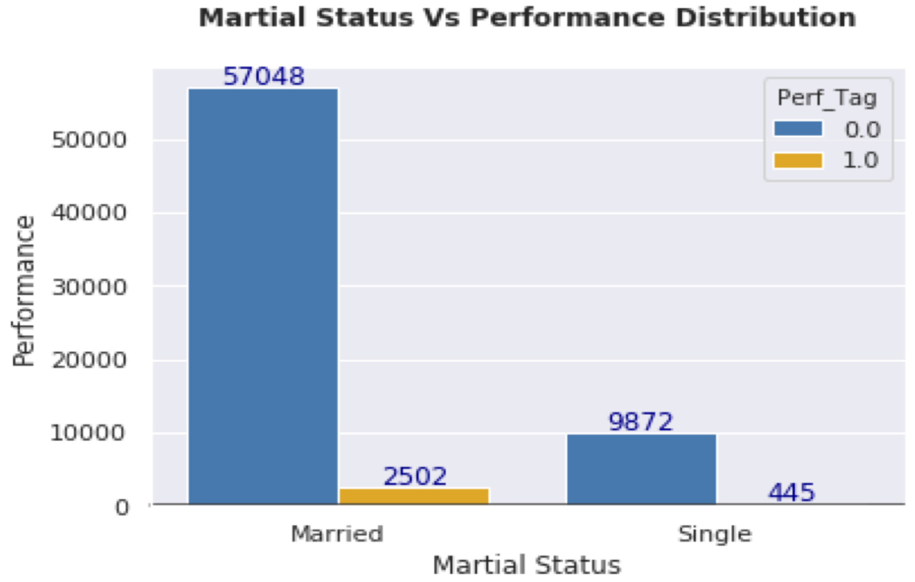




EDA Insights

- Bivariate
  - Male customers 3 times more likely to default
  - Married customers are 6 times more likely to default, and highest for customers with 3 dependents
  - Relative percentage wise, the education background did not affect defaulting nature
  - Relative percentage wise, the mode of salary did not affect defaulting nature





![iiit-b logo]

EDA Insights – Correlation

UpGrad

We see two groups of variables being correlated with variables within the group.

Group-1

- Num_90_DPD_6_Months
- Num_60_DPD_6_Months
- Num_30_DPD_6_Months
- Num_90_DPD_12_Months
- Num_60_DPD_12_Months
- Num_30_DPD_12_Months
- Avg_CC_Util_12_Months

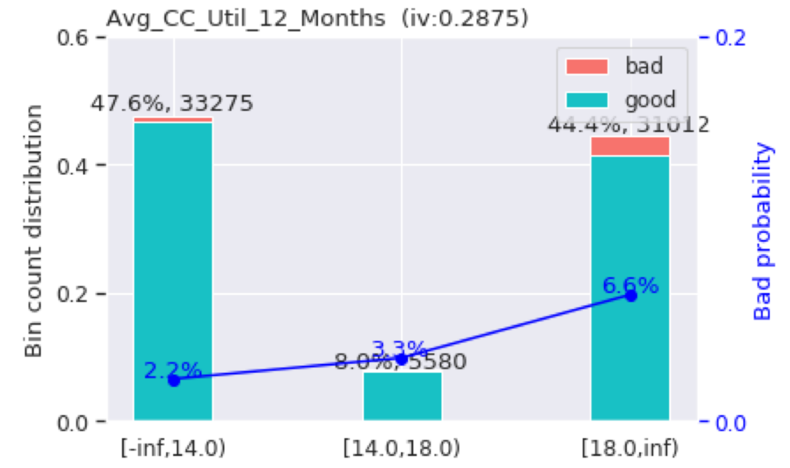
Group-2

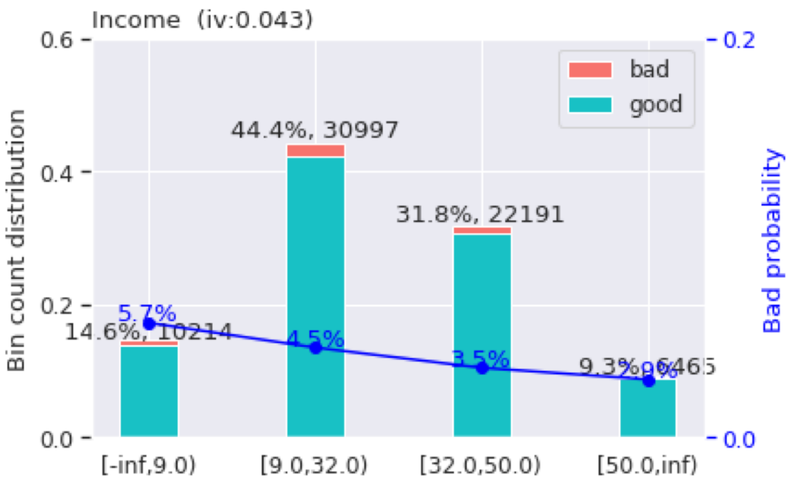
- Num_Trades_Open_6_Months
- Num_Trades_Open_12_Months
- Num_PL_Trades_open_6_Months
  Num_PL_Trades_Open_12_Months
- Num_Inq_6_Months
- Num_Inq_12_Months
- Outstanding_Bal
- Total_Trades

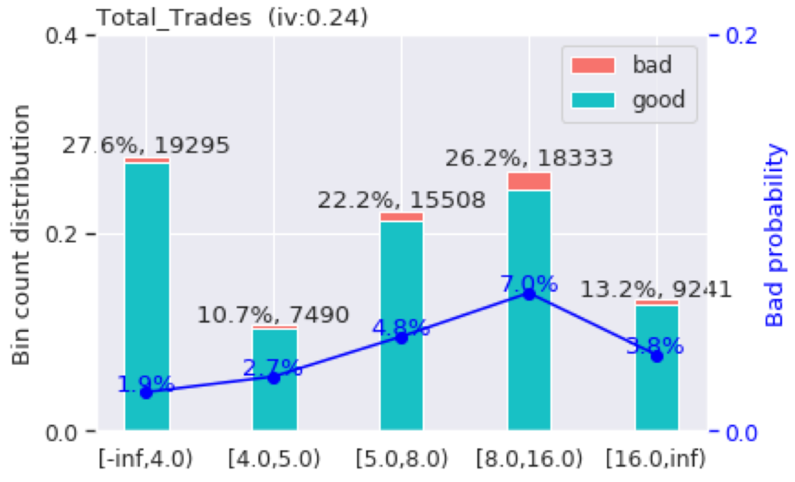
	Age	Num_Dependents	Income	Num_Months_Cur_Res	Num_Months_Cur_Comp	Perf_Tag	Num_90_DPD_6_Months	Num_60_DPD_6_Months	Num_30_DPD_6_Months	Num_90_DPD_12_Months	Num_60_DPD_12_Months	Num_30_DPD_12_Months	Avg_CC_Util_12_Months	Num_Trades_Open_6_Months	Num_Trades_Open_12_Months	Num_PL_Trades_Open_6_Months	Num_PL_Trades_Open_12_Months	Num_Inq_6_Months	Num_Inq_12_Months	Open_Home_Loan	Outstanding_Bal	Total_Trades	Open_Auto_Loan
Age	1.00	0.17	0.06	-0.07	-0.02	-0.00	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.00	0.02	0.02	0.01	0.01	0.01	0.02	-0.01	-0.01	0.02	-0.01
Num_Dependents	0.17	1.00	0.03	-0.01	-0.01	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	-0.00	-0.01	-0.00	-0.01	-0.00	-0.01	-0.01	-0.01	0.00	0.00	-0.00	-0.00
Income	0.06	0.03	1.00	-0.09	-0.10	-0.04	-0.19	-0.20	-0.21	-0.21	-0.19	-0.20	-0.18	-0.14	-0.16	-0.18	-0.19	-0.14	-0.15	0.07	0.00	-0.11	0.03
Num_Months_Cur_Res	-0.07	-0.01	-0.09	1.00	-0.08	0.02	0.15	0.15	0.15	0.16	0.18	0.18	0.26	-0.06	-0.04	0.04	0.04	-0.07	-0.08	0.00	0.02	-0.11	0.01
Num_Months_Cur_Comp	-0.02	-0.01	-0.10	-0.08	1.00	-0.02	-0.10	-0.12	-0.12	-0.12	-0.12	-0.13	-0.04	0.02	0.01	-0.01	-0.01	0.02	0.02	0.01	0.01	0.05	0.00
Perf_Tag	-0.00	-0.00	-0.04	0.02	-0.02	1.00	0.09	0.10	0.10	0.10	0.09	0.10	0.08	0.05	0.06	0.07	0.08	0.05	0.06	-0.03	0.00	0.03	-0.01
Num_90_DPD_6_Months	-0.02	-0.00	-0.19	0.15	-0.10	0.09	1.00	0.89	0.84	0.89	0.82	0.80	0.36	0.14	0.18	0.25	0.28	0.14	0.16	-0.11	-0.02	0.04	-0.03
Num_60_DPD_6_Months	-0.02	-0.00	-0.20	0.15	-0.12	0.10	0.89	1.00	0.95	0.84	0.92	0.90	0.35	0.16	0.20	0.27	0.30	0.16	0.17	-0.12	-0.02	0.04	-0.03
Num_30_DPD_6_Months	-0.02	-0.00	-0.21	0.15	-0.12	0.10	0.84	0.95	1.00	0.83	0.90	0.95	0.35	0.17	0.21	0.28	0.32	0.17	0.19	-0.12	-0.02	0.05	-0.03
Num_90_DPD_12_Months	-0.02	-0.00	-0.21	0.16	-0.12	0.10	0.89	0.84	0.83	1.00	0.81	0.80	0.39	0.16	0.20	0.28	0.31	0.16	0.18	-0.12	-0.02	0.04	-0.03
Num_60_DPD_12_Months	-0.02	0.00	-0.19	0.18	-0.12	0.09	0.82	0.92	0.90	0.81	1.00	0.90	0.34	0.13	0.17	0.25	0.28	0.13	0.15	-0.11	-0.01	0.01	-0.02
Num_30_DPD_12_Months	-0.02	-0.00	-0.20	0.18	-0.13	0.10	0.80	0.90	0.95	0.80	0.90	1.00	0.35	0.14	0.18	0.27	0.29	0.15	0.16	-0.11	-0.01	0.02	-0.02
Avg_CC_Util_12_Months	-0.00	-0.01	-0.18	0.26	-0.04	0.08	0.36	0.35	0.35	0.39	0.34	0.35	1.00	0.10	0.15	0.25	0.28	0.10	0.11	-0.12	-0.02	0.01	-0.04
Num_Trades_Open_6_Months	0.02	-0.00	-0.14	-0.06	0.02	0.05	0.14	0.16	0.17	0.16	0.13	0.14	0.10	1.00	0.94	0.88	0.84	0.67	0.74	-0.11	0.18	0.89	-0.04
Num_Trades_Open_12_Months	0.02	-0.01	-0.16	-0.04	0.01	0.06	0.18	0.20	0.21	0.20	0.17	0.18	0.15	0.94	1.00	0.89	0.93	0.72	0.79	-0.12	0.20	0.94	-0.04
Num_PL_Trades_Open_6_Months	0.01	-0.00	-0.18	0.04	-0.01	0.07	0.25	0.27	0.28	0.28	0.25	0.27	0.25	0.88	0.89	1.00	0.90	0.60	0.66	-0.13	0.18	0.79	-0.04
Num_PL_Trades_Open_12_Months	0.01	-0.01	-0.19	0.04	-0.01	0.08	0.28	0.30	0.32	0.31	0.28	0.29	0.28	0.84	0.93	0.90	1.00	0.66	0.73	-0.15	0.20	0.84	-0.05
Num_Inq_6_Months	0.01	-0.01	-0.14	-0.07	0.02	0.05	0.14	0.16	0.17	0.16	0.13	0.15	0.10	0.67	0.72	0.60	0.66	1.00	0.91	-0.11	0.12	0.72	-0.04
Num_Inq_12_Months	0.02	-0.01	-0.15	-0.08	0.02	0.06	0.16	0.17	0.19	0.18	0.15	0.16	0.11	0.74	0.79	0.66	0.73	0.91	1.00	-0.13	0.13	0.78	-0.05
Open_Home_Loan	-0.01	0.00	0.07	0.00	0.01	-0.03	-0.11	-0.12	-0.12	-0.12	-0.11	-0.11	-0.12	-0.11	-0.12	-0.13	-0.15	-0.11	-0.13	1.00	0.94	-0.09	0.02
Outstanding_Bal	-0.01	0.00	0.00	0.02	0.01	0.00	-0.02	-0.02	-0.02	-0.02	-0.01	-0.01	-0.02	0.18	0.20	0.18	0.20	0.12	0.13	0.94	1.00	0.20	0.05
Total_Trades	0.02	-0.00	-0.11	-0.11	0.05	0.03	0.04	0.04	0.05	0.04	0.01	0.02	0.01	0.89	0.94	0.79	0.84	0.72	0.78	-0.09	0.20	1.00	-0.03
Open_Auto_Loan	-0.01	-0.00	0.03	0.01	0.00	-0.01	-0.03	-0.03	-0.03	-0.03	-0.02	-0.02	-0.04	-0.04	-0.04	-0.04	-0.05	-0.04	-0.05	0.02	0.05	-0.03	1.00

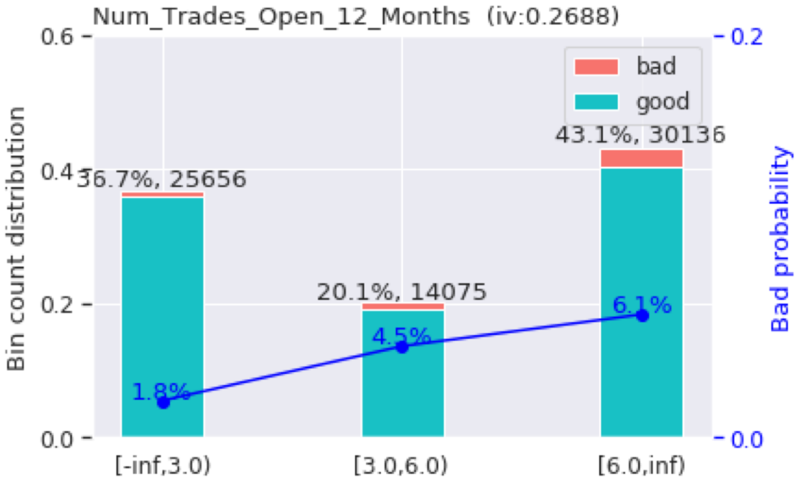
1.00 0.75 0.50 0.25 0.00

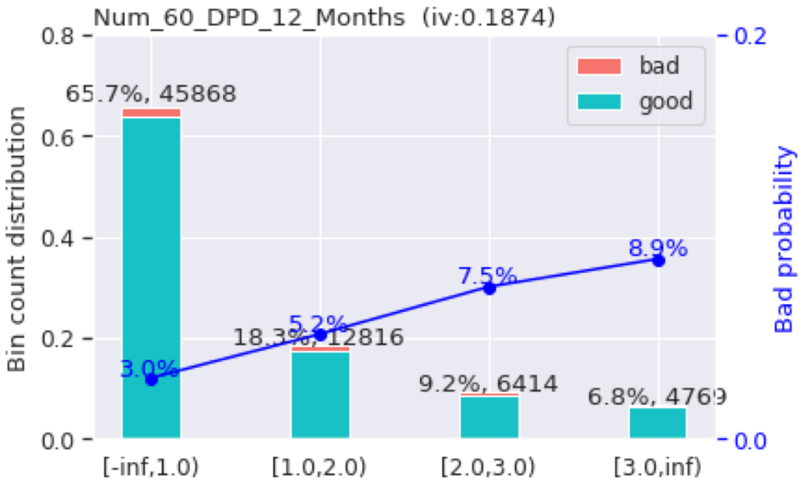
WOE bins of few features

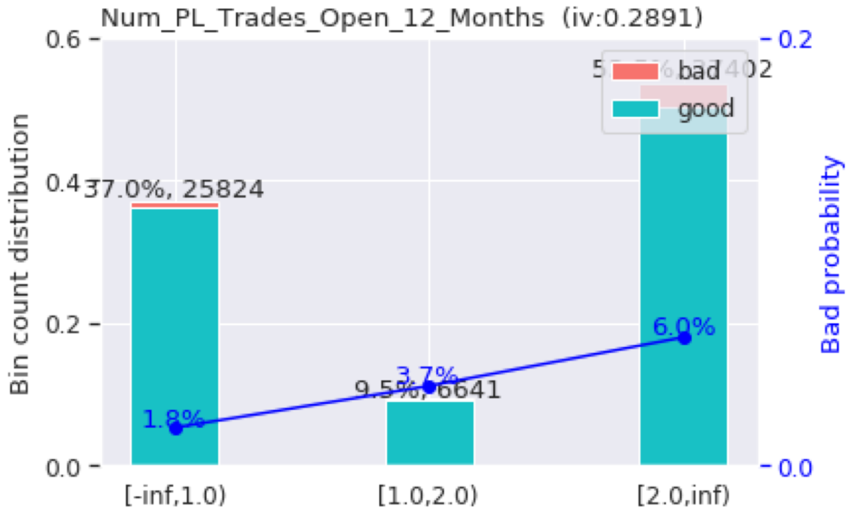






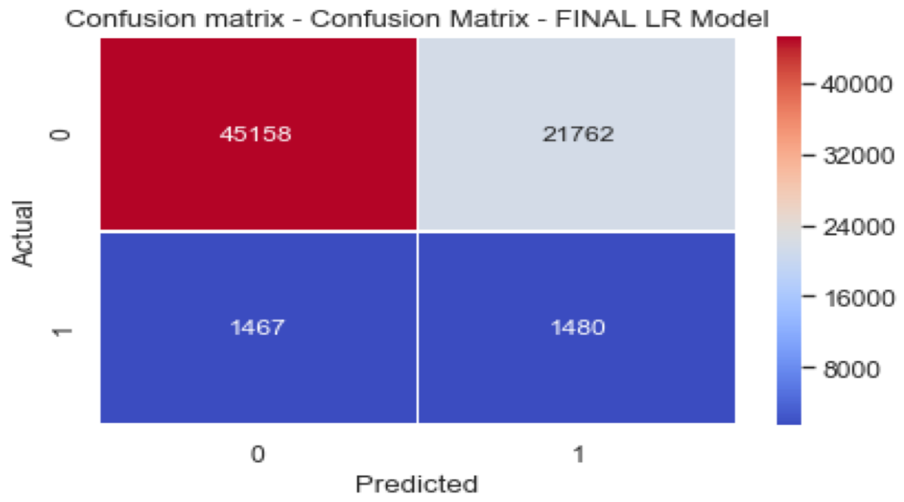






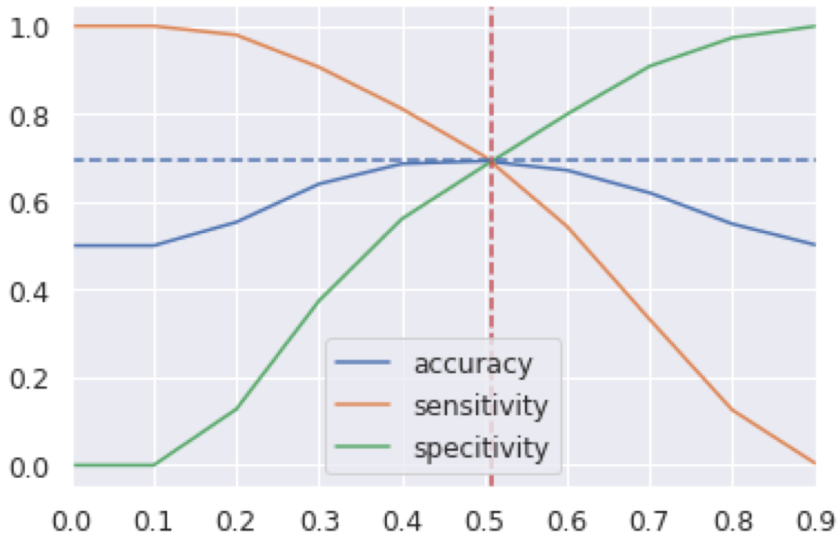
Models Developed & Evaluated

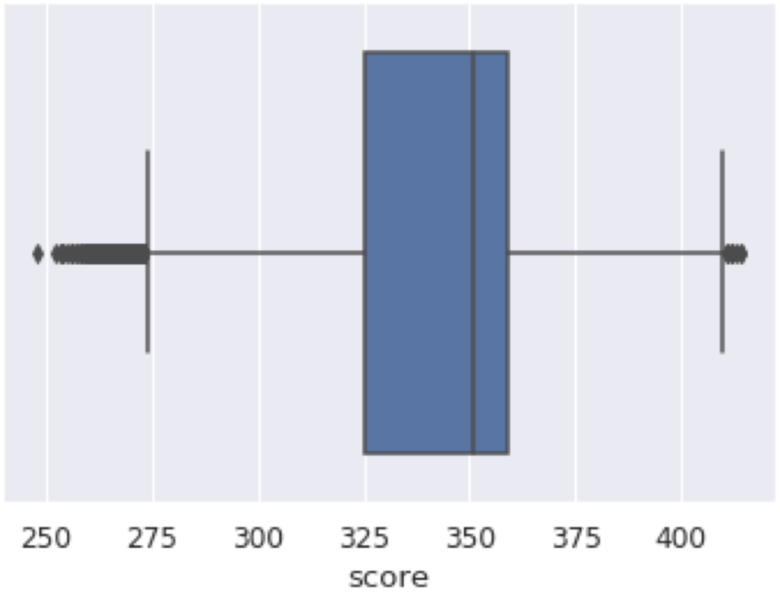
- A series of models such as Logistic Regression, Random Forest, Decision Tree, SVM & Logistic regression for WOE data were developed and the metrics were evaluation
- Logistics Regression with SMOTE for WOE data was found to be the best model with Accuracy of 68%
- This model was chosen to evaluate for the Rejected Population (1425 records) and the Original population (69867 records)



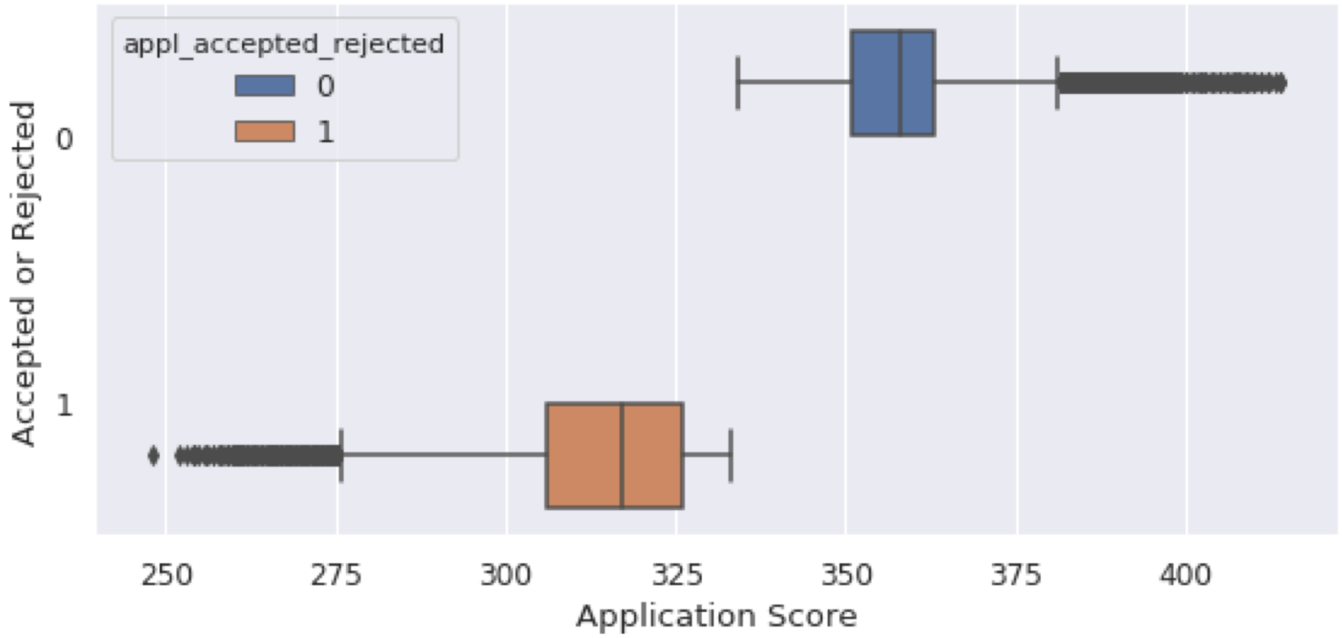
Cut-off Score

- The cut-off score is calculated based on the optimum cut-off probability (0.503)
- The optimum probability was found at the merging of accuracy, sensitivity, and specificity, i.e. ~ 0.53
- It was found to be approximately **333**





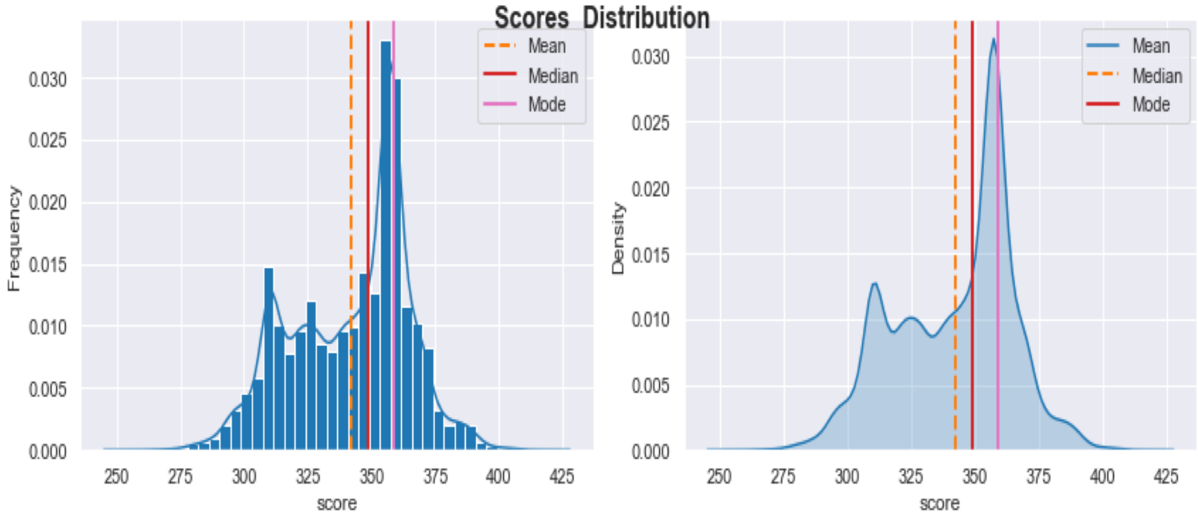
Application Scorecard



- The best model i.e. LR on sampled (SMOTE) WOE data is used to generate the scores for all the customers
- The good to bad odds are 10 to 1 at a score of 400 doubling every 20 points
- The scores are calculated to be in the range of 248 to 414

Optimum Model Results Summary

- The conservative cut-off score for the Optimum model was found to be 333
- The model was evaluated for the Rejected population and the defaulter's rate was found to be 43%
- The Defaulter's rate for the Original population was found to be 36%
- At cut-off score 333, the below results were obtained
  - Conservative Net Approval Rate : 65.00%
  - Total Outstanding Balance of Accepted Applicants/Profit @cut off 333 : 56027836200.00
  - Total Outstanding Balance of Rejected Applicants/Credit Loss @cut off 333 : 31411073333.00



Recommendations

- Applying the Bar of 80-20% Acceptance rate, it was found based on the 20% quantile of scores, the revised balanced cut-off score is 318
- At 80% approval rate, the outstanding balance is found to be
  - Total Outstanding Balance of Accepted Applicants/Net profit - 71027114302.00
  - Total Outstanding Balance of Rejected Applicants/Credit Loss - 16411795231.00
- Comparing with the Conservative approval rate of 65%, the net profit is
  - Final Net Profit after revised balanced cut off : 14999278102.00
  - Final Net Credit Loss after revised balanced cut off : 14999278102.00

Scenario	Cut-off Score	Net Credit Loss	Profit %
Approval rate 96%	NA	3711178158	0
Approval rate 65%	333	31411073333	33%
Approval rate 80%	318	16411795231	79%

- Clearly based on the final P&L numbers, its obvious that there is no profit and loss based on cut-off score of 318
- Adopting no Profit & no loss policy, we suggest to use the developed model at cut-off score of 318
- Based on the risk the Credit card company management wants to take, the model can be used at a cut-off score less than 318 which would increase the final Net Credit loss and reduce the final Net Profit.