BFS CAPSTONE PROJECT

CREDX ANALYTICS

Final Submission

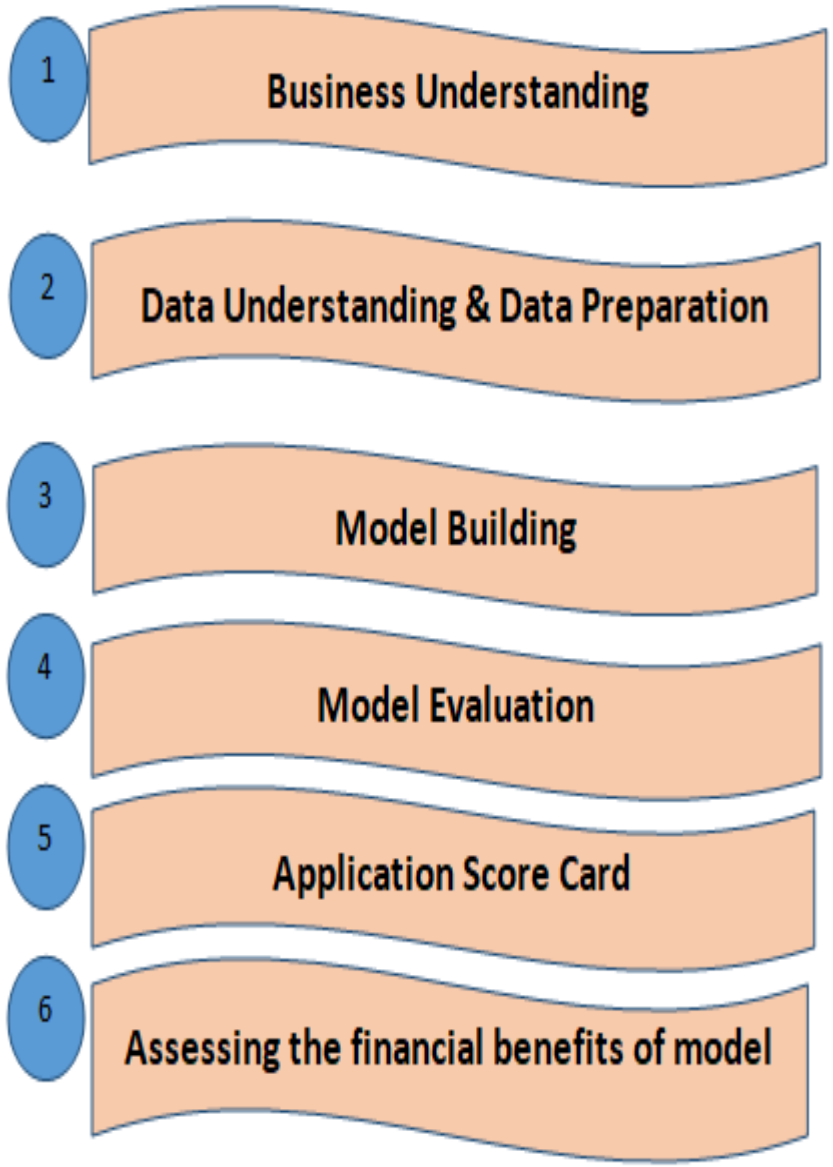
-Presented By

1. Anuradha Cherukupalli
2. Monika Iyer M

PROBLEM STATEMENT

BUSINESS UNDERSTANDING

- **Objective**: Help CredX identify the right customers using predictive models by determining the determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of the project.
- **Problem Statement**: 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 due to increase in defaults. Create strategies to mitigate acquisition risk and assess the financial benefit to Credx.
- **Solution Approach**: Build binary classification models such as Logistic Regression, Random Forest, Decision Trees etc. to identify the customers who are at a risk of defaulting.


1
Business Understanding
2
Data Understanding & Data Preparation
3
Model Building
4
Model Evaluation
5
Application Score Card
6
Assessing the financial benefits of model

DATA UNDERSTANDING

1. **Demographic/Application Data** : This data contains information provided by the applicantsat the time of Credit card application and contains variables such as Age, Marital Status, Gender,Income and Type of Residence etc.

   This dataset consists of 71295 observations and 12 Variables.

2. **Credit Bureau Data** : This information is taken from credit bureau and contains variables such as Number of time 90DPD, Number of times 60DPD, Number of times 30 dpd in last 6 months and 12 months, Avg CC utilization, Presence of Open auto loan or open home loan, etc.

   This dataset consists of 71295 observations and 19 variables

Performance Tag is the target variable with 0 (non-default) and 1(default). It also has null values which means that the applications were rejected manually due to various reasons.

Application ID is the common column for merging both data sets.

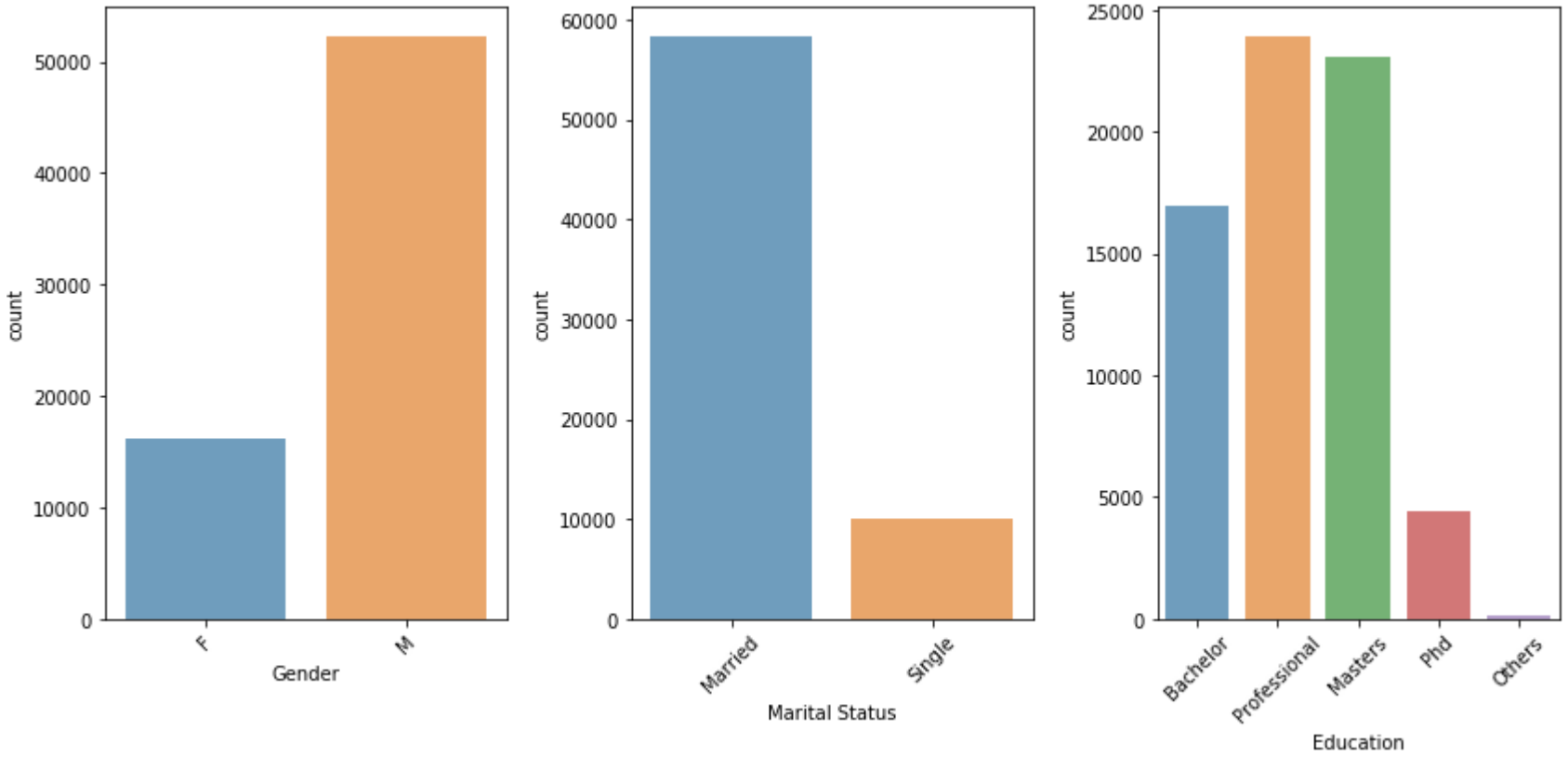
DATA CLEANING AND PREPARATION

- As a first step, unique values were found on Application Id column in Credit Bureau dataset and the Demography dataset. Three duplicate Application ids were found in both the datasets respectively. These were removed.
- The data of demography and credit bureau were merged on the Application id column.
- There are 1425 values where performance tag is null. Hence we have separately kept this dataset as rejected dataset for later predictions in modeling phase.
- The original dataset had a space in the column Profession, which was removed.
- We also dropped other null values in the dataset as it was less than 1 percent of the whole dataset.
- As people who are less than 18 are considered minors, we have treated this as an outlier treatment criteria for Age column and removed all applicants less than 18 years of age.
- Similarly, people who have income less than 0 have also been removed

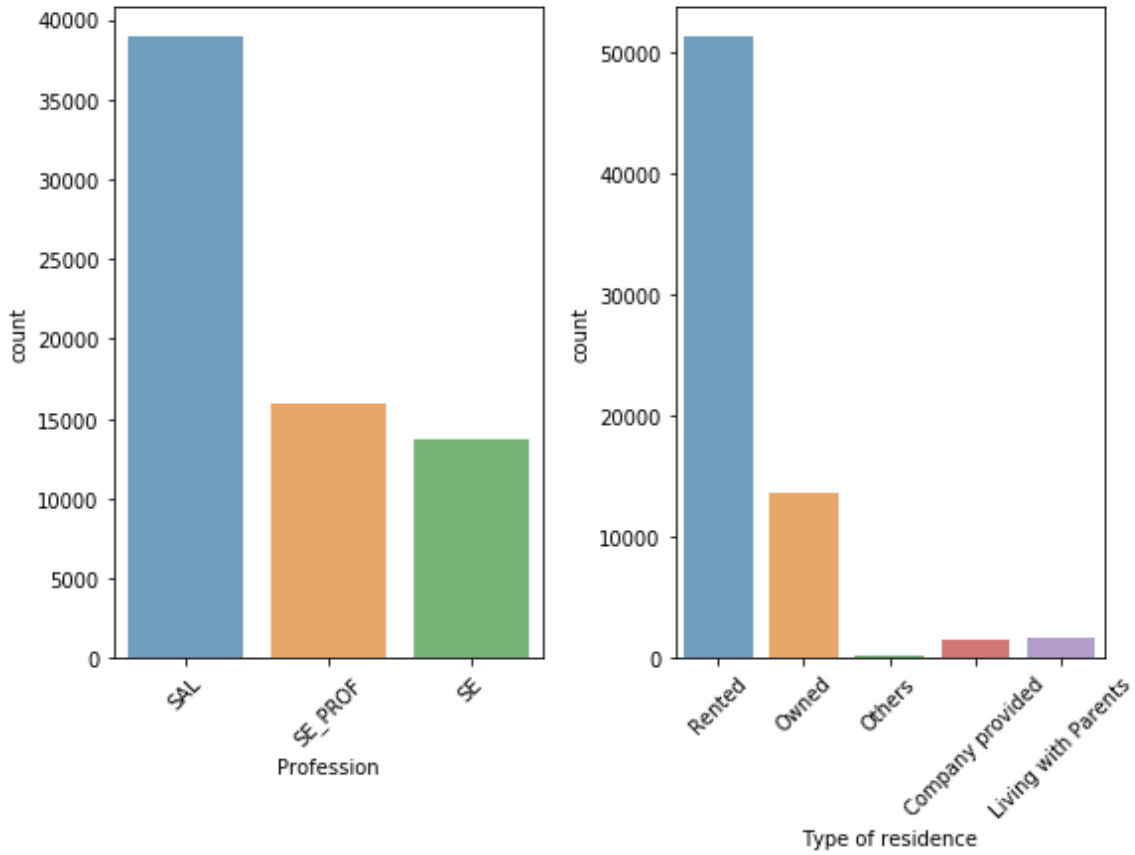
EDA

- We then proceeded to do EDA on the master dataset.
- There were few variables like Income, Age, Outstanding balance etc, which were continuous variables. These variables were binned for analysis purposes.
- The EDA plots are presented as following.

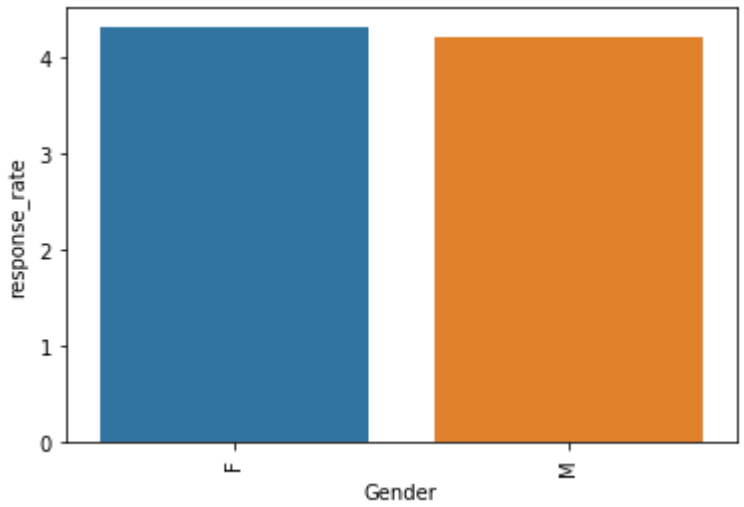
Univariate Analysis on categorical

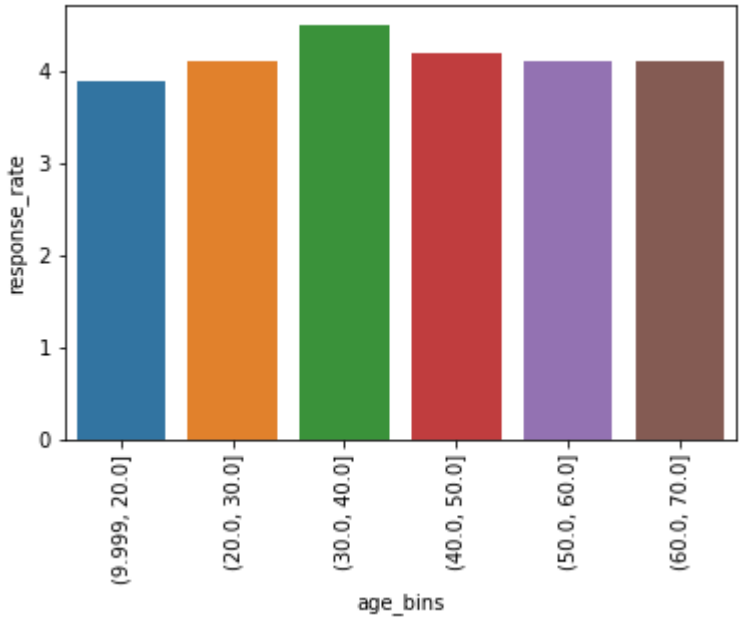


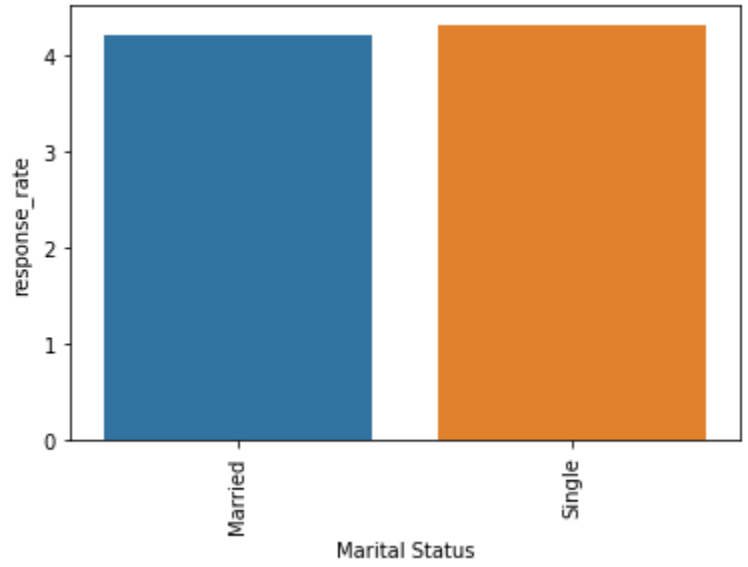
Univariate Analysis on Categorical

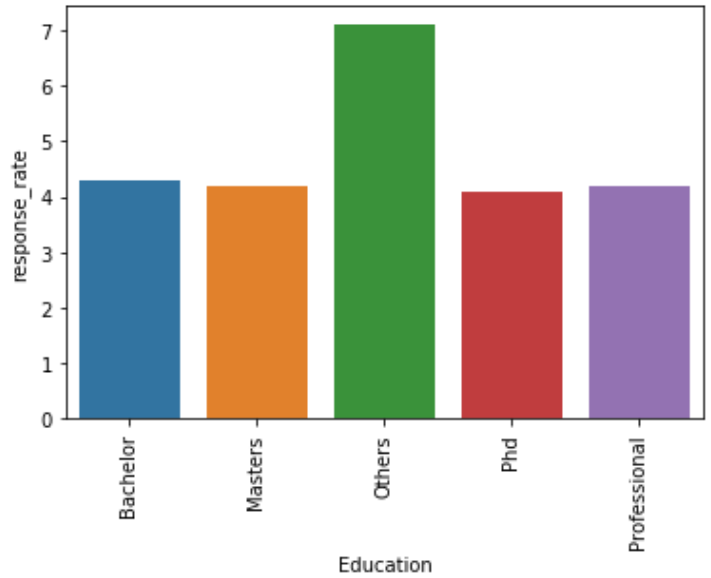


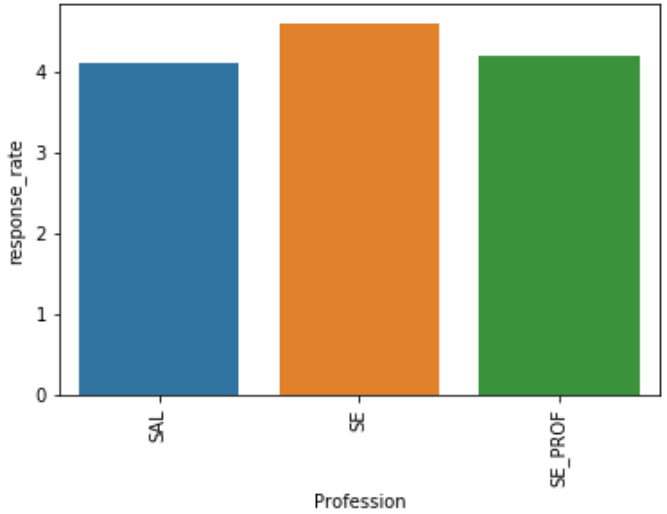
Bivariate Analysis

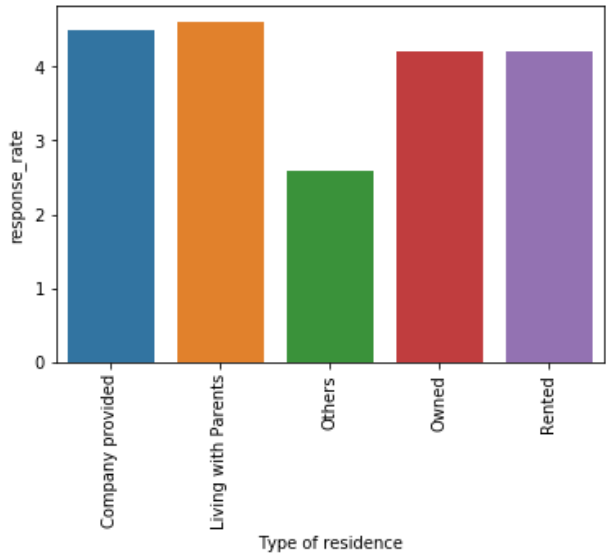


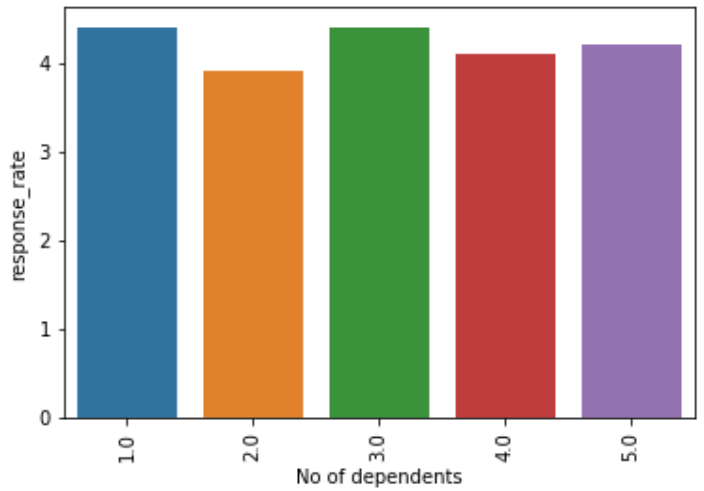


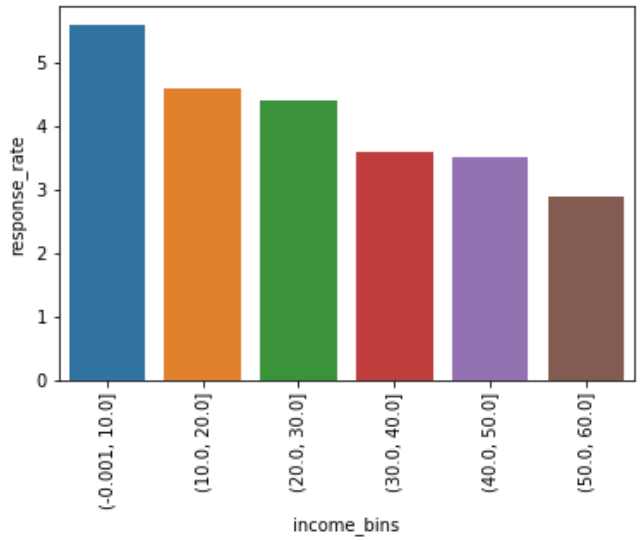


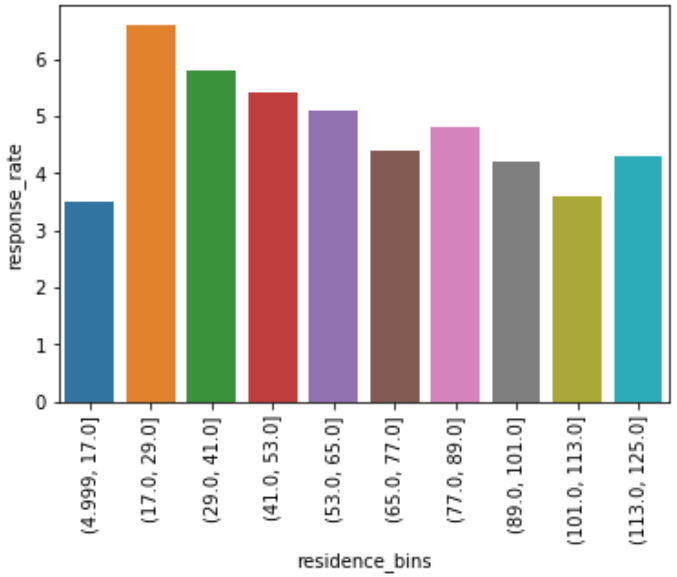

response_rate
0
1
2
3
4
5
6
7
Bachelor
Masters
Others
Phd
Professional
Education

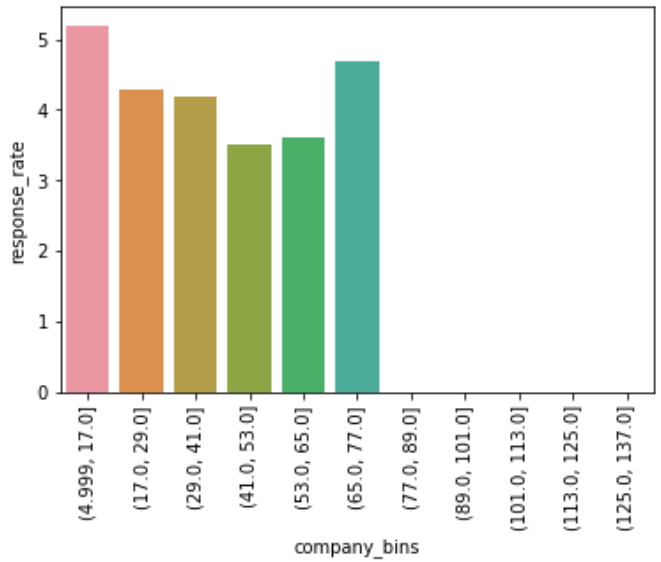

response_rate
0
1
2
3
4
SAL
SE
SE_PROF
Profession

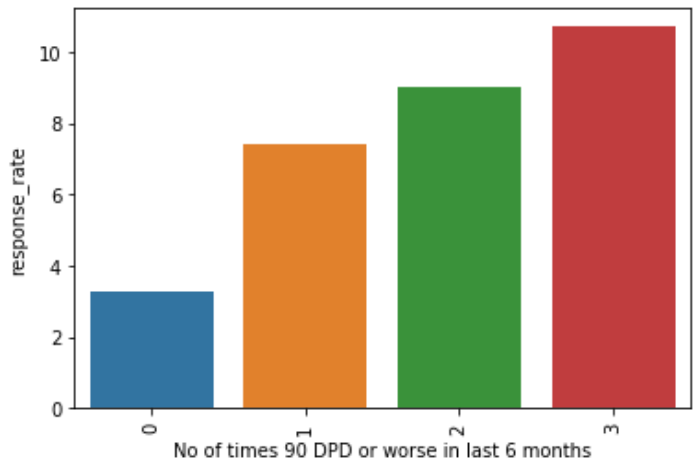

response_rate
0
1
2
3
4
Company provided
Living with Parents
Others
Owned
Rented
Type of residence

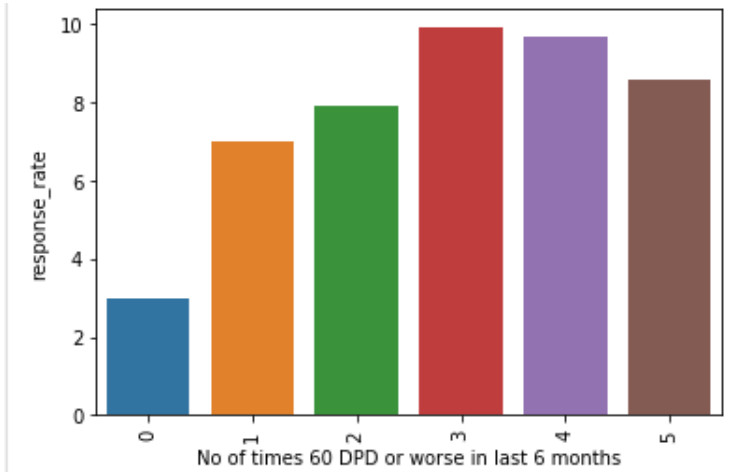

response_rate
0
1
2
3
4
1.0
2.0
3.0
4.0
5.0
No of dependents

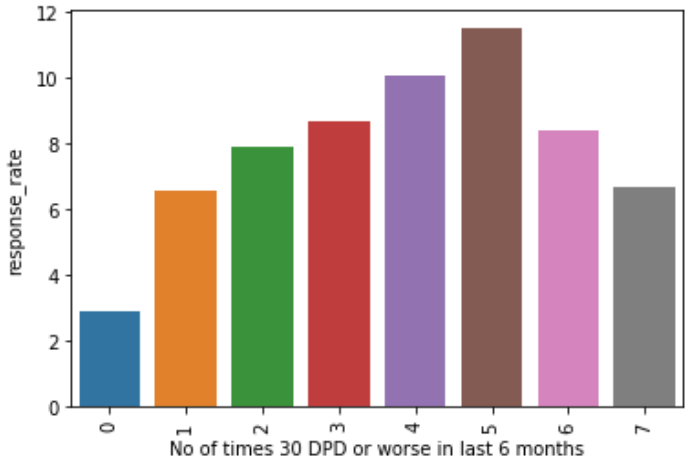

response_rate
0
1
2
3
4
5
(-0.001, 10.0]
(10.0, 20.0]
(20.0, 30.0]
(30.0, 40.0]
(40.0, 50.0]
(50.0, 60.0]
income_bins

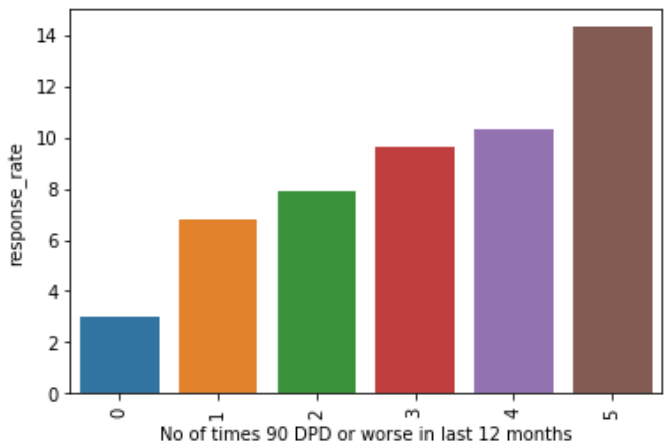

response_rate
0
1
2
3
4
5
6
(4.999, 17.0]
(17.0, 29.0]
(29.0, 41.0]
(41.0, 53.0]
(53.0, 65.0]
(65.0, 77.0]
(77.0, 89.0]
(89.0, 101.0]
(101.0, 113.0]
(113.0, 125.0]
residence_bins

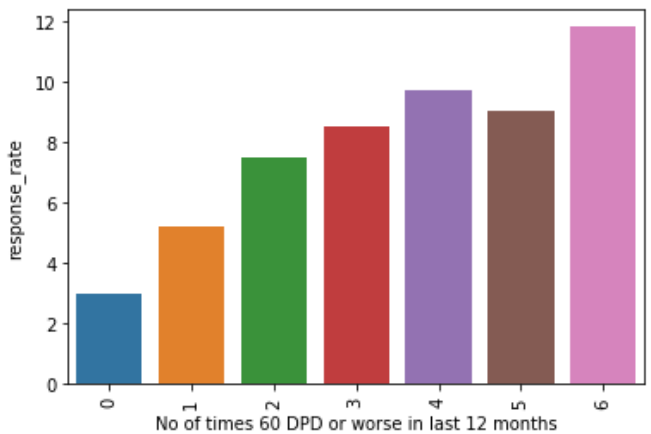

response_rate
0
1
2
3
4
5
(4.999, 17.0]
(17.0, 29.0]
(29.0, 41.0]
(41.0, 53.0]
(53.0, 65.0]
(65.0, 77.0]
(77.0, 89.0]
(89.0, 101.0]
(101.0, 113.0]
(113.0, 125.0]
(125.0, 137.0]
company_bins

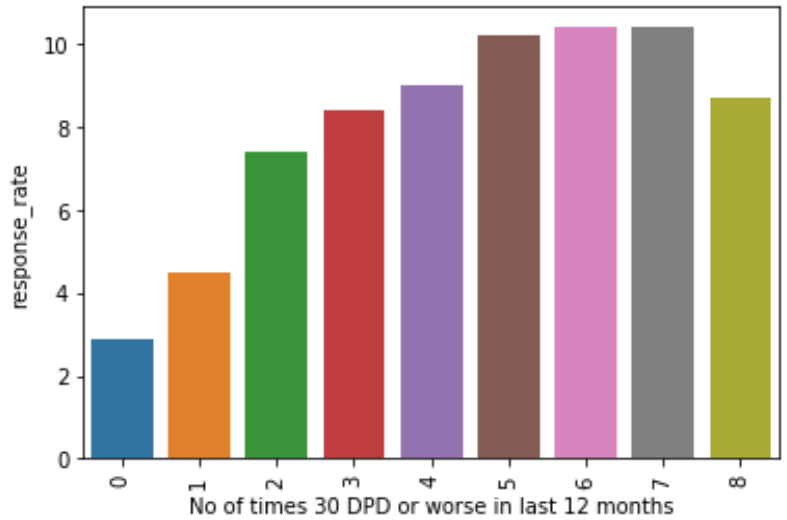

response_rate
0
2
4
6
8
10
0
1
2
3
No of times 90 DPD or worse in last 6 months

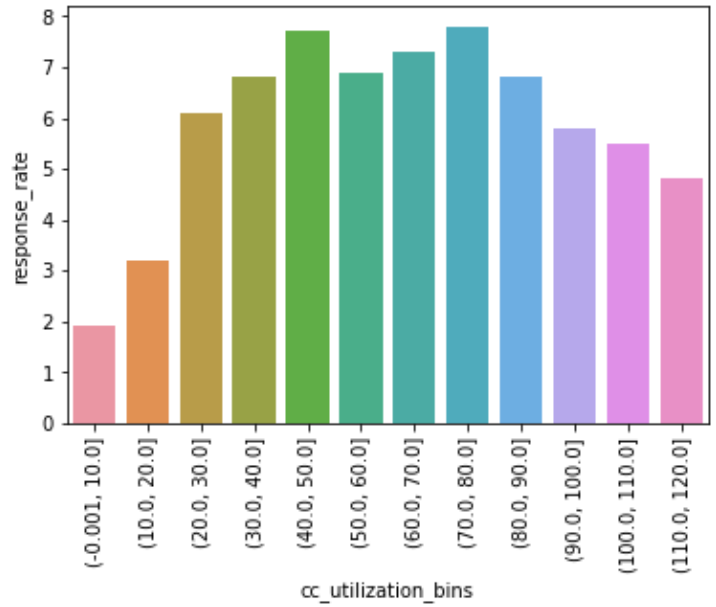

response_rate
No of times 60 DPD or worse in last 6 months

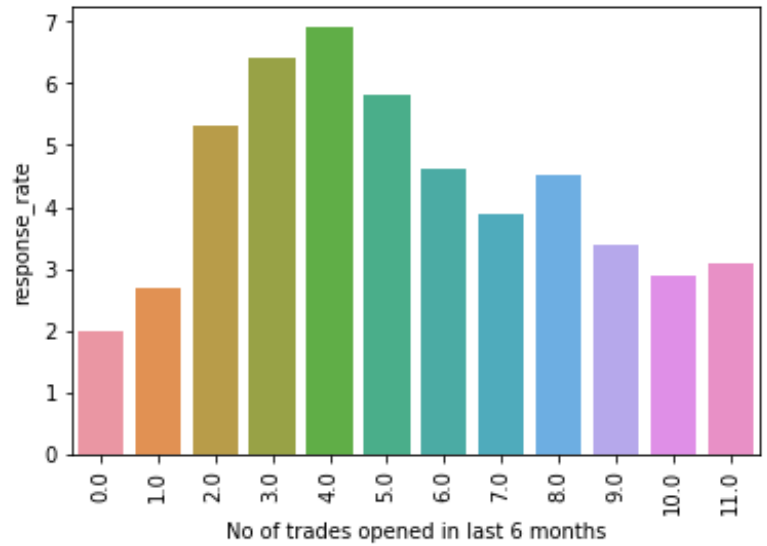

response_rate
No of times 30 DPD or worse in last 6 months

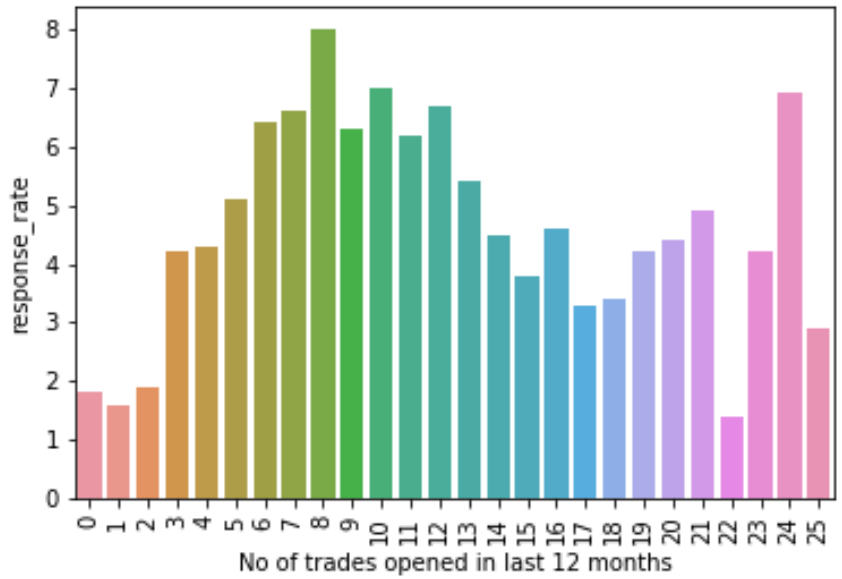

response_rate
No of times 90 DPD or worse in last 12 months

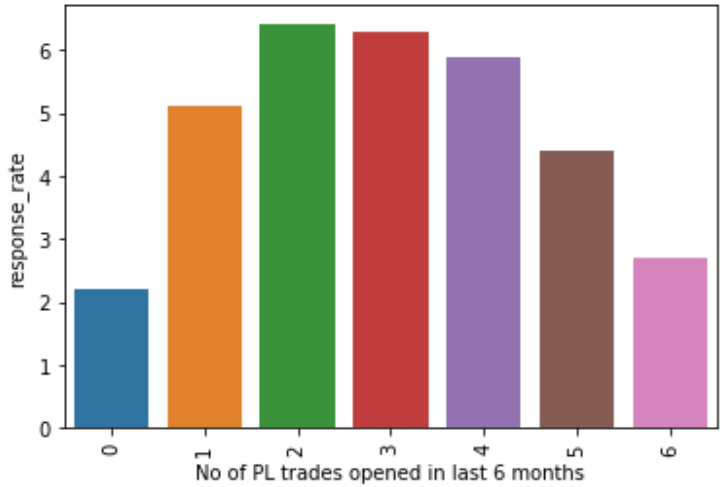

response_rate
No of times 60 DPD or worse in last 12 months

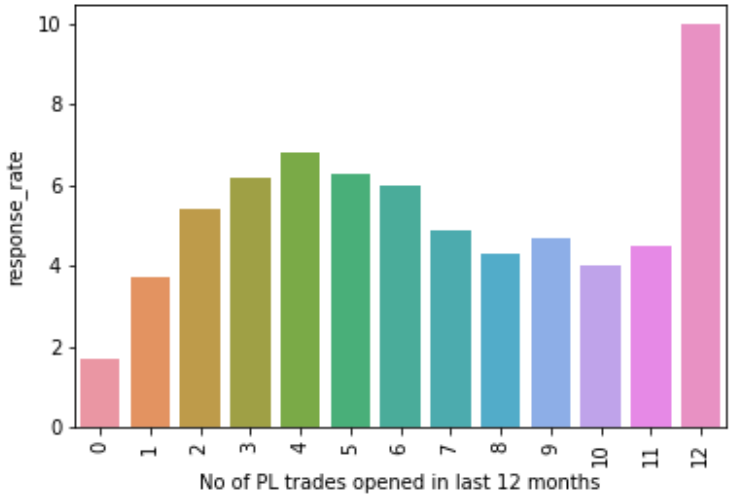

response_rate
No of times 30 DPD or worse in last 12 months

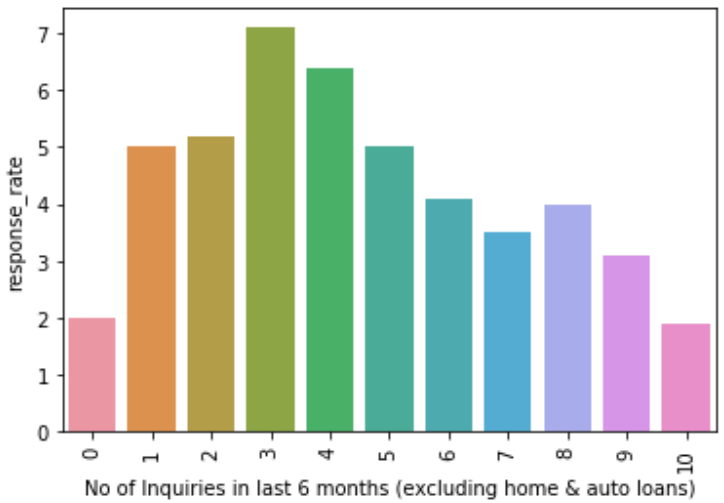

response_rate
(-0.001, 10.0]
(10.0, 20.0]
(20.0, 30.0]
(30.0, 40.0]
(40.0, 50.0]
(50.0, 60.0]
(60.0, 70.0]
(70.0, 80.0]
(80.0, 90.0]
(90.0, 100.0]
(100.0, 110.0]
(110.0, 120.0]
cc_utilization_bins

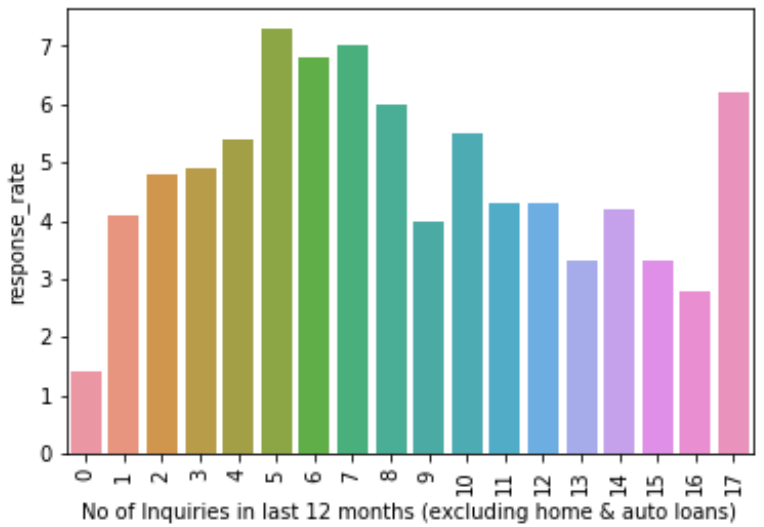

response_rate
No of trades opened in last 6 months

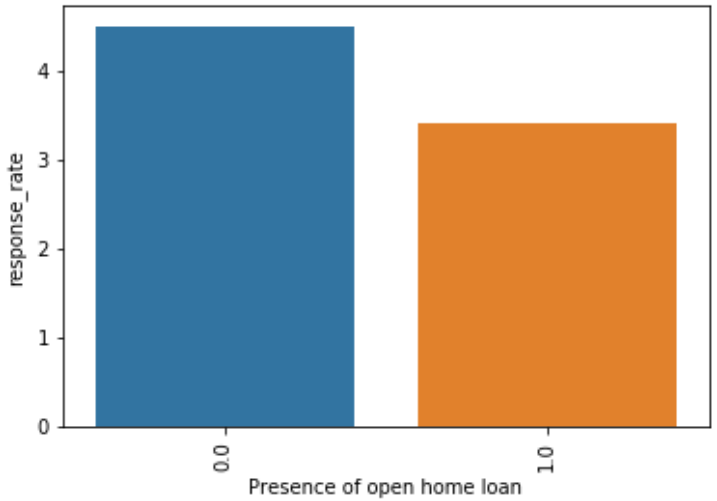

response_rate
No of trades opened in last 12 months

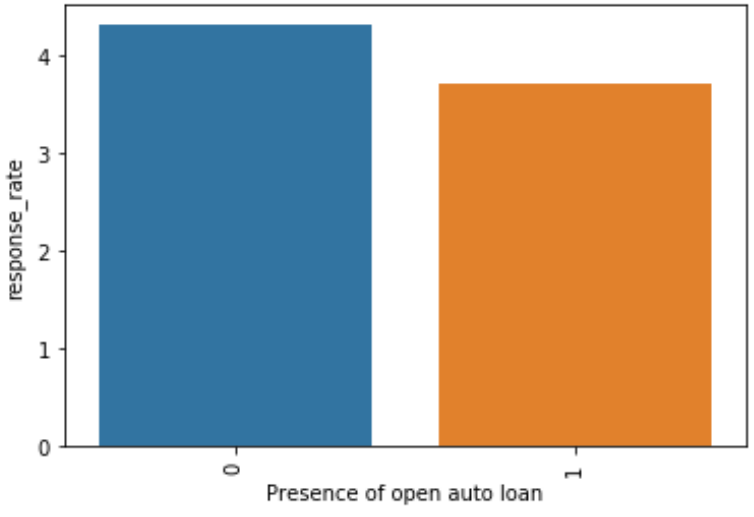

response_rate
No of PL trades opened in last 6 months

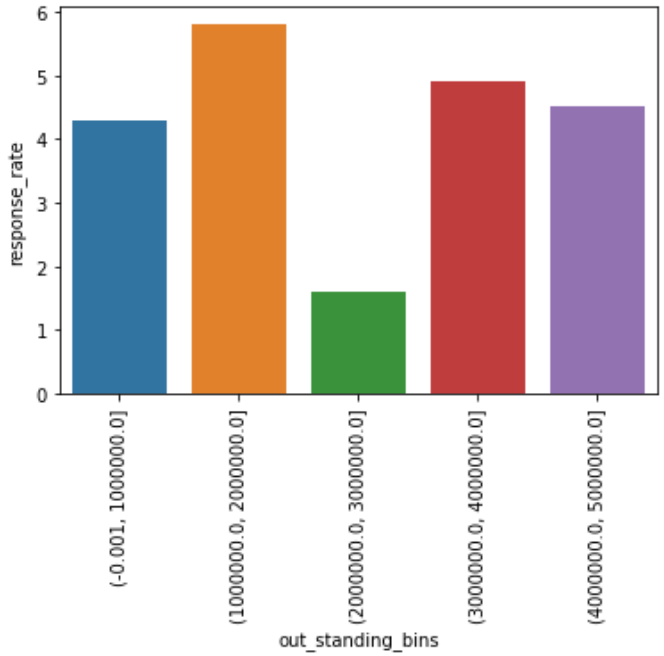

response_rate
No of PL trades opened in last 12 months

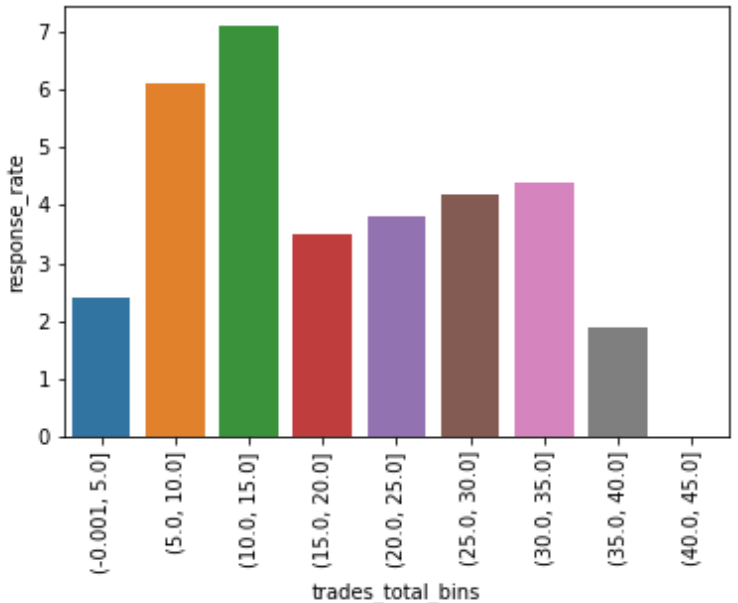

response_rate
No of Inquiries in last 6 months (excluding home & auto loans)


response_rate
No of Inquiries in last 12 months (excluding home & auto loans)


response_rate
Presence of open home loan
0.0
1.0


response_rate
Presence of open auto loan
0
1


response_rate
out_standing_bins
(-0.001, 1000000.0]
(1000000.0, 2000000.0]
(2000000.0, 3000000.0]
(3000000.0, 4000000.0]
(4000000.0, 5000000.0]


response_rate
trades_total_bins
(-0.001, 5.0]
(5.0, 10.0]
(10.0, 15.0]
(15.0, 20.0]
(20.0, 25.0]
(25.0, 30.0]
(30.0, 35.0]
(35.0, 40.0]
(40.0, 45.0]

WOE AND IV ANALYSIS

- As the next step, we did WOE and IV analysis.
- Weight of Evidence(WOE) and Information Value(IV) are often used in credit risk analytics to identify the important variables.
- A function was written to calculate the woe and IV values.
- The actual dataset was then replaced by WOE calculated values and stored in a separate dataset.
- IV values were then used to know the important variables. All the variables that have IV values were considered as weak predictors. Any variable with a IV value between 0.1and 0.3 was considered as medium predictor and variables with values more than 0.3 and less than 0.5 were considered as strong predictors.

DATA PREPARATION

- Two datasets are used for modeling.
- One is the demographic woe dataset which contains the woe replaced values of demographic data.
- Another one is the master dataset that contains credit bureau data and demographic data but with woe replaced values.
- Another model is considered only with IV values greater than 0.1

DATA TRANSFORMATION

- Dummy variables were created for categorical variables on the demographic woe dataset as well as on the merged dataset of credit bureau and demographic woe dataset.
- First we start with demographic woe dataset.
- The dataset was split into train and test data in the ratio 70:30.
- Scaling has been performed on all variables except for the variables Application ID and Performance Tag.
- Logistic Regression, Decision Tree and Random Forest models are created for Demographic and the master dataset.

MODEL BUILDING AND EVALUATION

- Logistic Regression for merged dataset performed better and hence was chosen as the final model. Let us now check the metrics of each of the models. Decision Trees is not used, as we have already used an ensemble(Random Forest).
- SVM usually requires higher computing power due to large data.

Logistic Regresssion-Demographic	
Accuracy	60.1
Sensitivity	54
Specificity	56
AUC	57

Random Forest- Demographic	
Accuracy	93.4
Sensitivity	8
Specificity	94
AUC	52

MODEL BUILDING AND EVALUATION

Logistic Regresssion- Master	
Accuracy	66.4
Sensitivity	69
Specificity	56
AUC	66

Random Forest- Master	
Accuracy	99
Sensitivity	0
Specificity	100
AUC	55

Logistic Regression- IV-Master	
Accuracy	62
Sensitivity	62
Specificity	63
AUC	

APPLICATION SCORE CARD

- An application scorecard with the good to bad odds of 10 to 1 at a score of 400 doubling every 20 points was built.
- Application score was calculated for rejected population and accepted population.
- The optimal cut-off score below which credit card should not be granted was found to be 337.1.
- We could have given credit card to 1791 out of 2891 applicants.
- The approval rate is 80.99%.

FINANCIAL BENEFIT

- Keeping cut off score off 337.1, we see that there is a difference of 4.91 percent between the model accepted rate and the rejected population.
- Increase in approval rate will increase the credit loss. There has been a significant saving when the approvals have been done using the model compared to not using a model.
- If we assume that we lose 1000 rupees for every defaulter, then the profit earned by using this model is approximately 6 lakhs.