CredX Acquisition Risk Analytics

Group members:

- Geetakrishnasai Gunapati
- Pramodini V Nayak

Background:

- 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.
- The CEO believes that the best strategy to mitigate credit risk is to acquire "the right customers".

Objective:

- To identify the right customers using predictive models. We need to determine the factors affecting credit risk and create strategies to mitigate the acquisition.
- Build an application scorecard and identify the cut-off score below which we would not grant credit cards to applicants.
- We need to assess and explain the potential financial benefit of our project and identify the metrics we are trying to optimize.

Problem Solving Methodology:

This is a binary supervised classification problem. We have followed CRISP–DM framework. It involves the following series of steps:

- **Business understanding**
- **Data understanding and preparation :**
  - ✓ Reading demographic and credit bureau data.
  - ✓ Doing the basic data check to understand the data shape, column names and types.
  - ✓ Removing duplicate rows.
  - ✓ Merging demographic and credit bureau data on ‘Applicant ID’.
  - ✓ Checking NAs.
- **Exploratory data analysis:**
  - ✓ Exploring each and every column by Univariate and Multivariate Analysis.
  - ✓ Checking correlation among variables.
  - ✓ Handling data quality issues.

Problem Solving Methodology:

- **WOE and IV analysis:**
  - ✓ Computing WOE and Information value.
  - ✓ Identifying the important variables using Information value.
  - ✓ Replacing null values with WOE values.
  - ✓ Populating features with their WOE values.
- **Model Building:**
  - ✓ Splitting data into train and test
  - ✓ Building the models only demographic data to get an idea about the predictive power of the application data.
  - ✓ Building the models on the merged data using Logistic Regression, Decision Tree ,Random Forest and SVM algorithms.
- **Model Evaluation**
  - ✓ Evaluating the model on AUC, accuracy, sensitivity and specificity.
- **Identifying the factors affecting credit risk using the chosen model.**
- **Building Application Score Card using the chosen model.**
- **Assessing financial benefits of the chosen model.**

Business Understanding

With the large potential profit in credit card banking also comes the risk of customers not paying off their credit card balance.
As CredX is dealing with increased credit loss, it is important that they exercise good risk control. Because while they try to expand their customer base they may also acquire certain risky customers.

- Customers who regularly pay within due date are 'low risk - low revenue' customers.
- Customers who are habitual defaulters from DPD but eventually pay - are 'medium risk- high revenue' customers.
- Customers who do not pay at all - are 'high risk- high credit loss' customers.

Customers of category 'medium risk', are the right customers to be acquired. 'Low risk' are good to increase the customer base but they are not high revenue customers. However, CredX in the past few years, has experienced an increase in credit loss due to 'high risk' customers. Therefore, the best strategy to mitigate the credit risk would be to acquire 'the right customers'.

Data Understanding

Basic structure of the data:

- To begin with – two datasets are provided. Demographic data and credit bureau data.
  - **Demographic/application data:** This dataset contains 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.
  - **Credit bureau data:** This information 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.
- In both the datasets we noted 6 records in which 3 were duplicate records and we dropped all 6 records.
- Merged both the data sets on ‘Application ID’.

Data Understanding

Basic structure of the data:

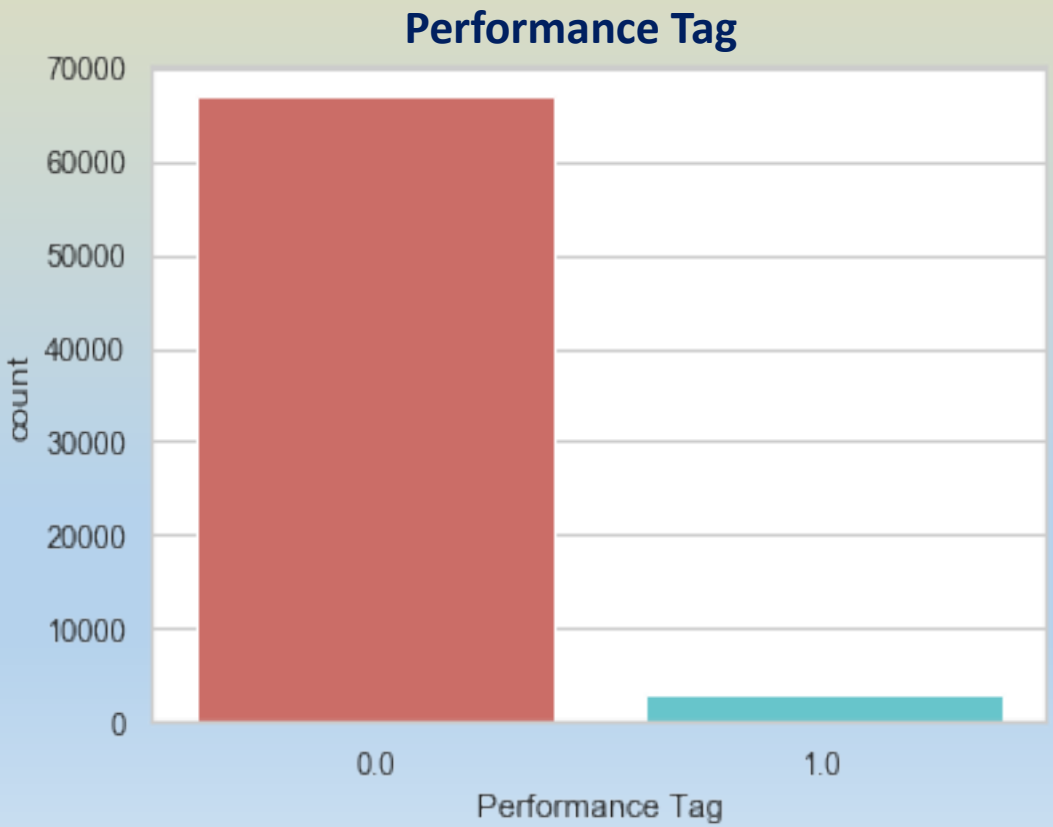
- The variables such as - gender, marital status , no. of dependents , education , profession, type of residence, avg. credit card utilization in last 12 months, no. of trades opened in last 6 months, outstanding balance, presence of open home loan and performance tag have null values. The null % in all the these variables are less than 3%.

- Performance Tag indicates whether the applicant defaulted or not. Hence we have assumed NA in performance tag(1425 rows) as 'rejected credit card applicants'. We later predicted performance tag for the rejected population and calculated the application scores for assessing the model results.

- The data is highly unbalanced. The ratio of non-defaulters to defaulters instances is 96:4.

Exploratory Data Analysis

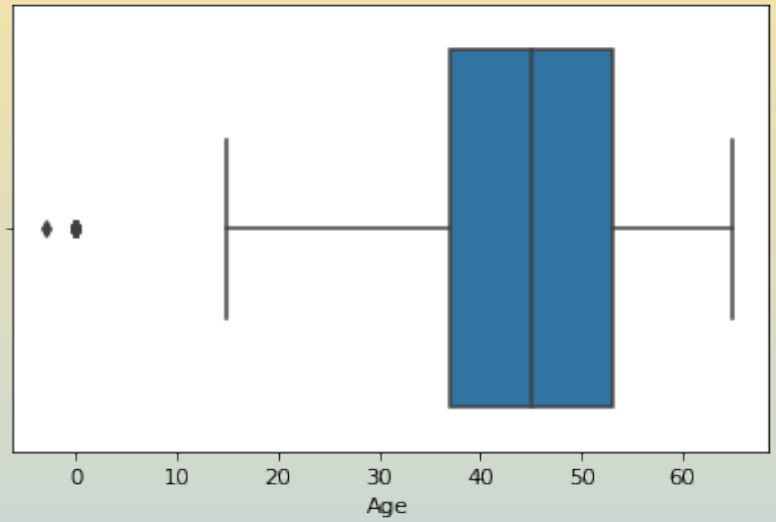
We have done univariate and bivariate analysis on each and every column. Column with data quality issues and their response towards target variable will be explained here. Let us start with our target variable - Performance Tag.

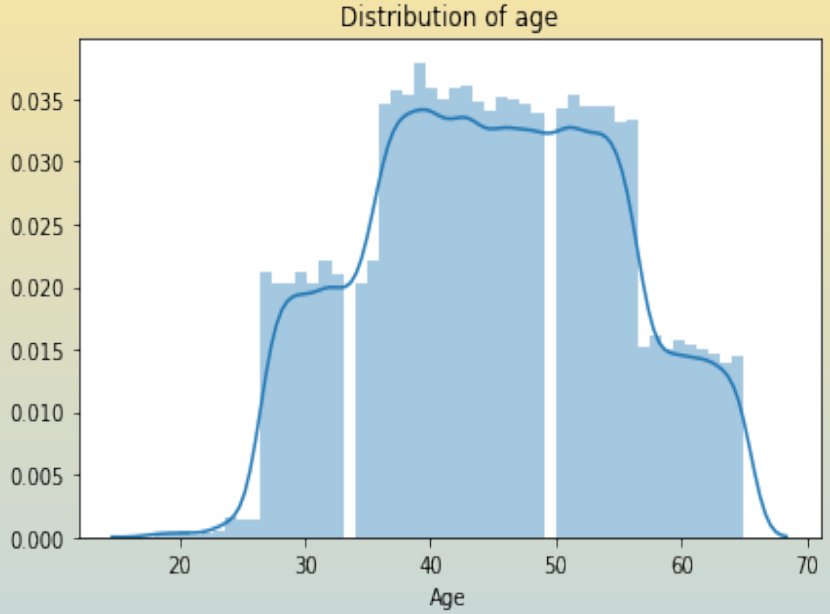


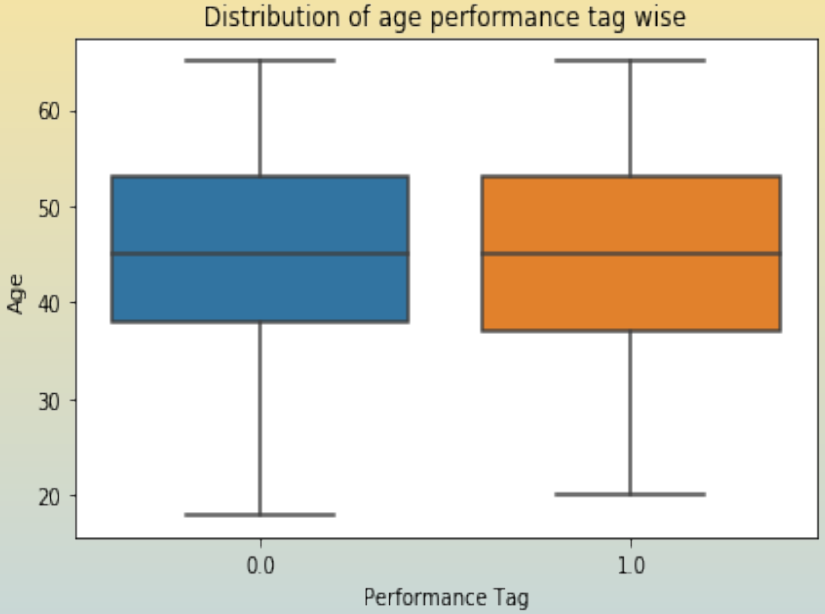
- Performance Tag column had NA values. As performance tag indicates whether the applicant has gone 90 days past due (DPD) or worse in the past 12 months (i.e. defaulted) after getting a credit card. So we have assumed NA in performance tag(1425 rows) as 'rejected credit card applicants'.

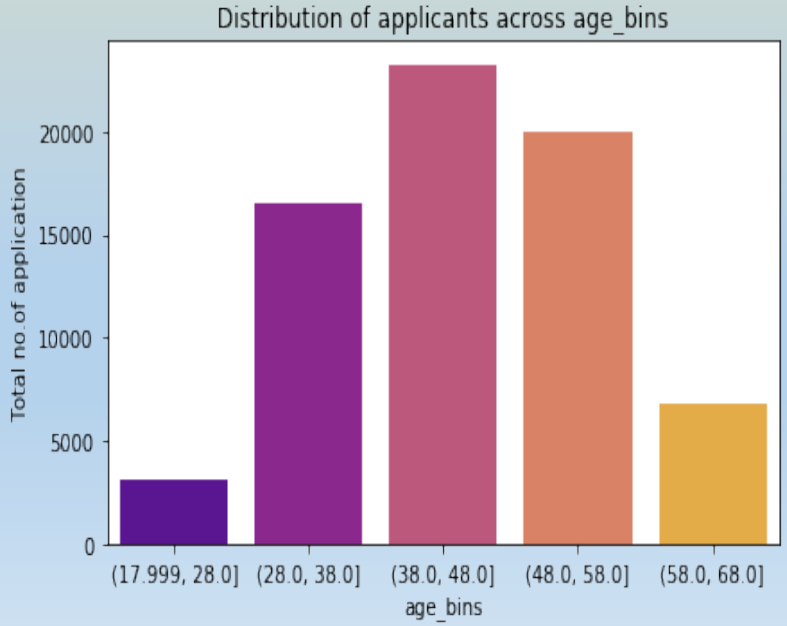
- We have 66917 non-defaults records(0.0) and 2947 defaulted records(1.0). So it is highly unbalanced dataset with non-defaulters to defaulters ratio to be 96:4.

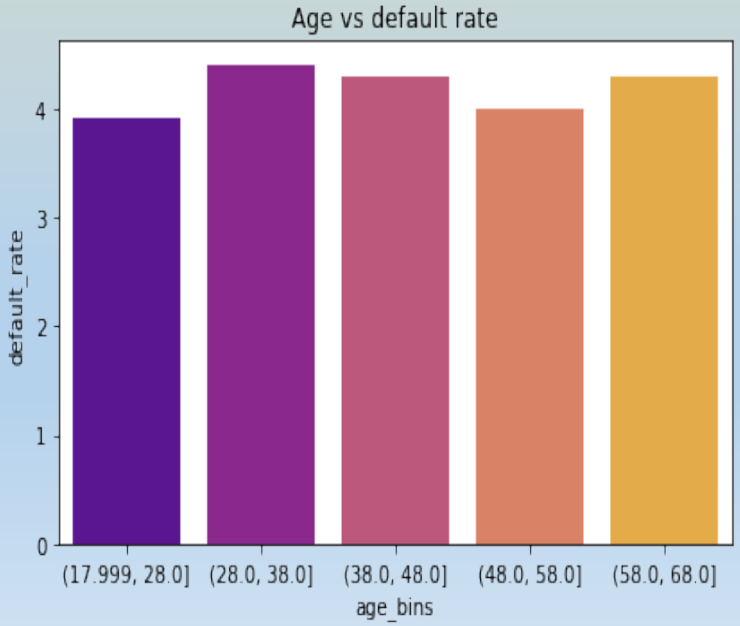
Demographic Data Column: Age





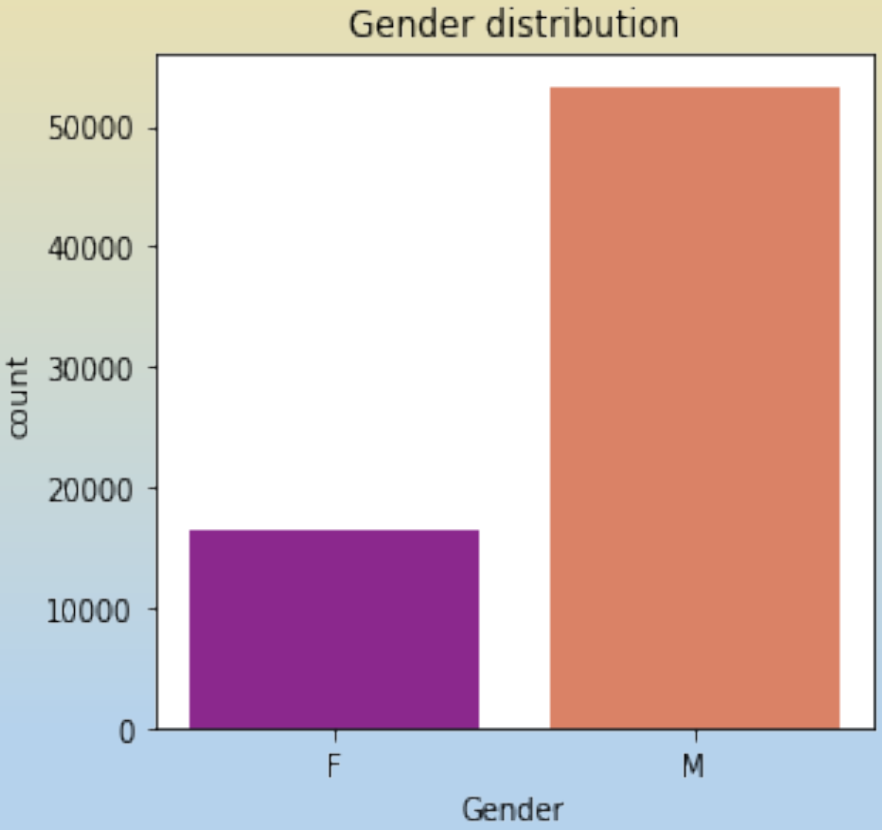


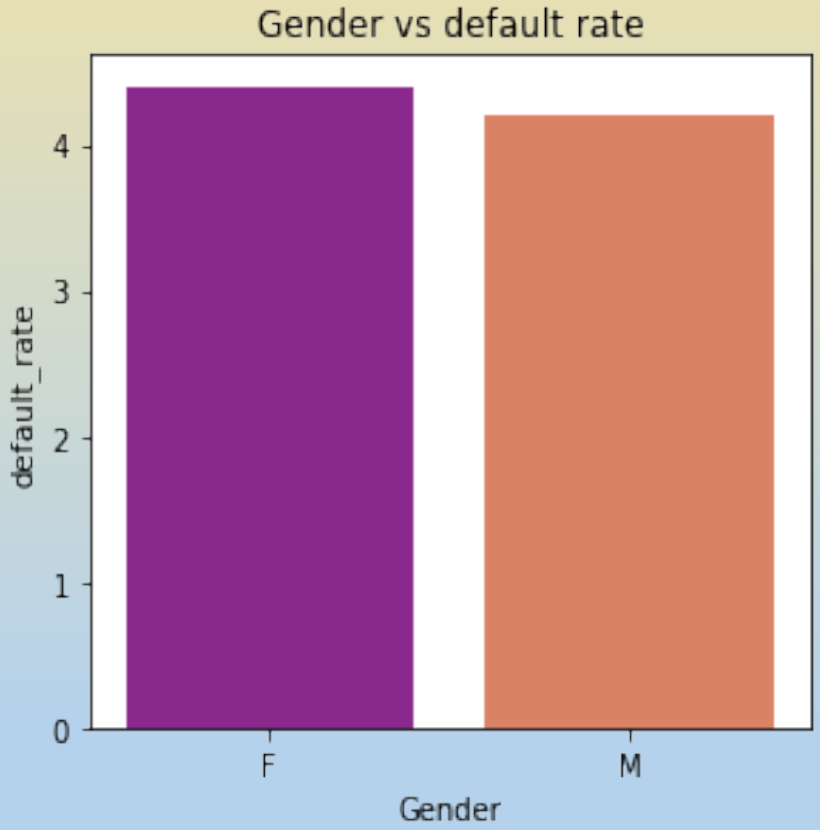




- We note a data quality issue in 'Age' column.
- Few data entries have negative age. We have assumed the minimum age for credit card has to be 18 years.
- We see maximum credit holders are of age group 30+ to 58 years old.
- The average age of default and non defaulters is similar as seen from the box plot.
- We don't see any extreme default rate difference between the age groups. Almost all the age group has the same default rate.

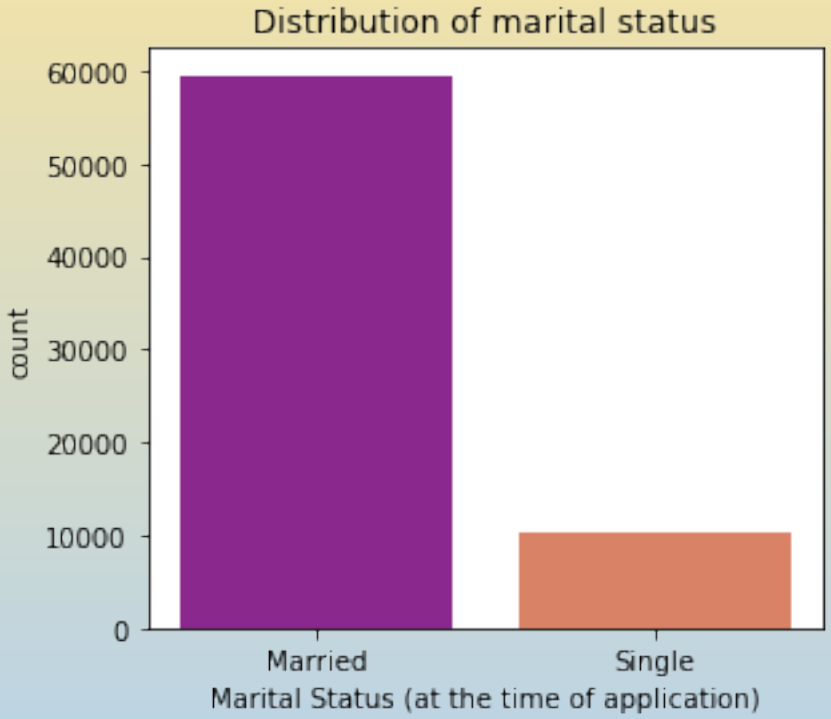
Demographic Data Column: Gender

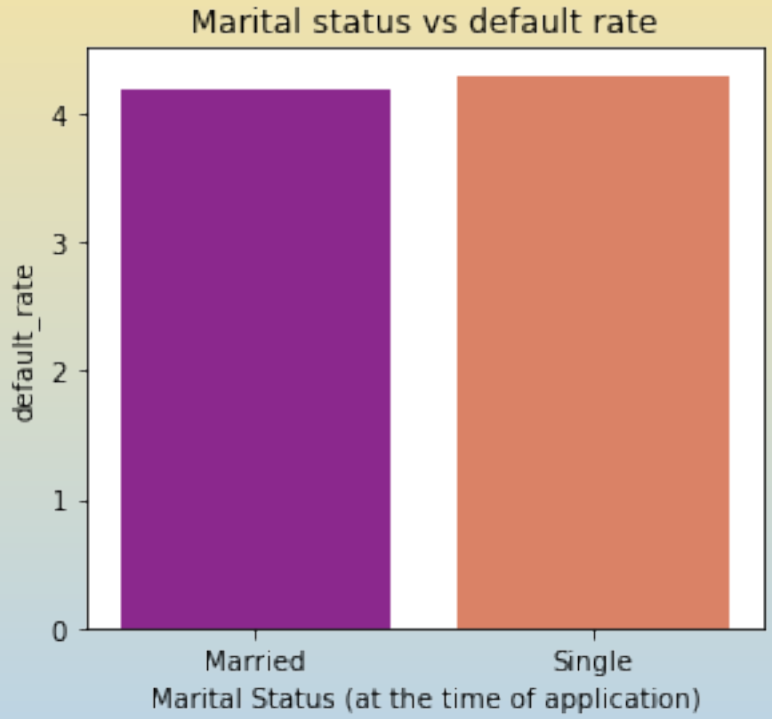


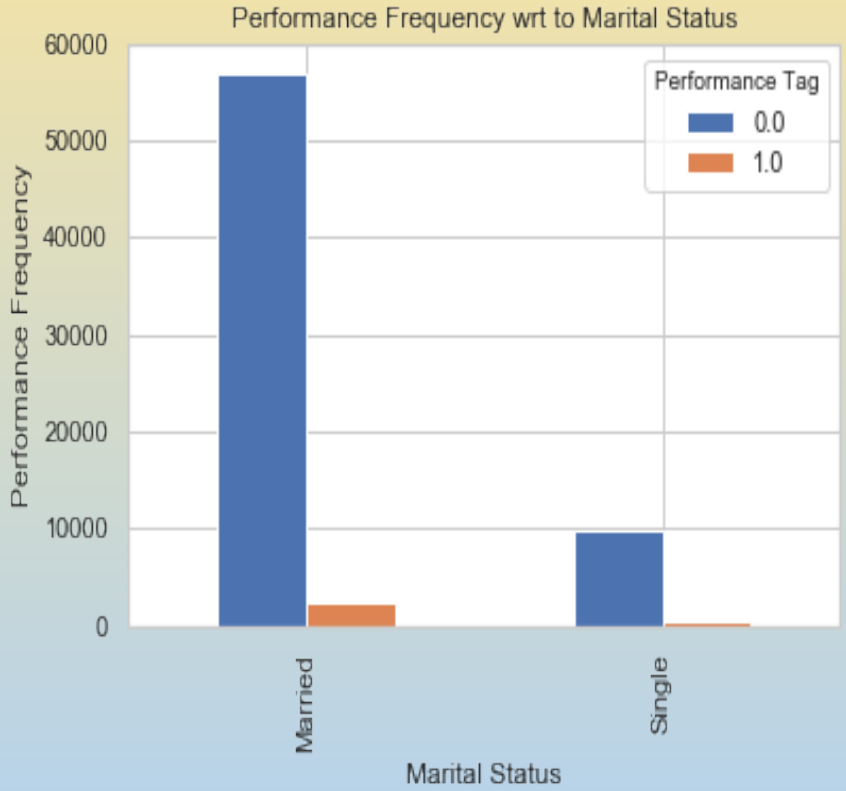


- Maximum credit card holders are Male.
- There seems to be no correlation between a gender and default rate.

Demographic Data Column: Marital Status

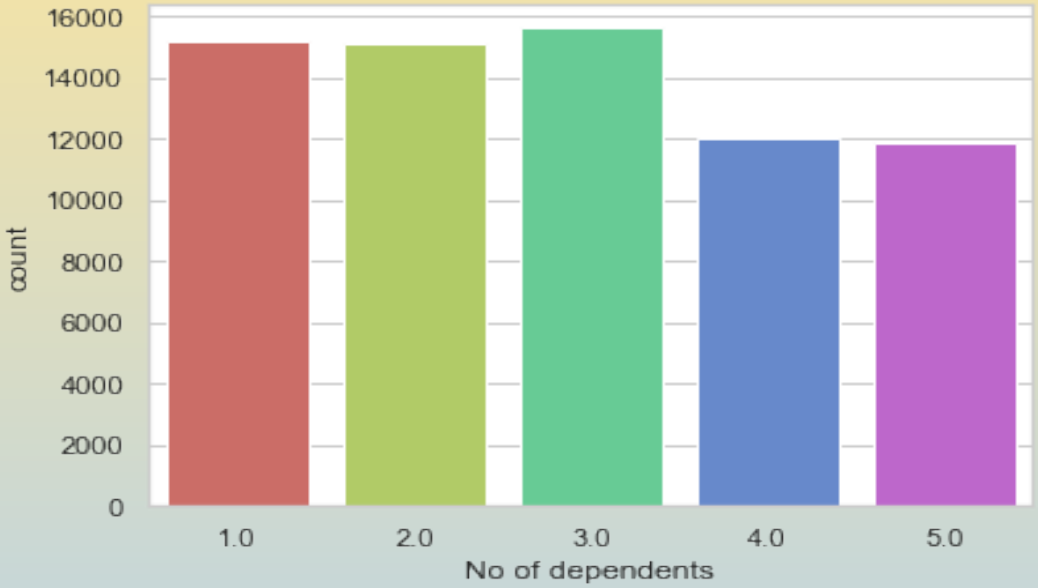


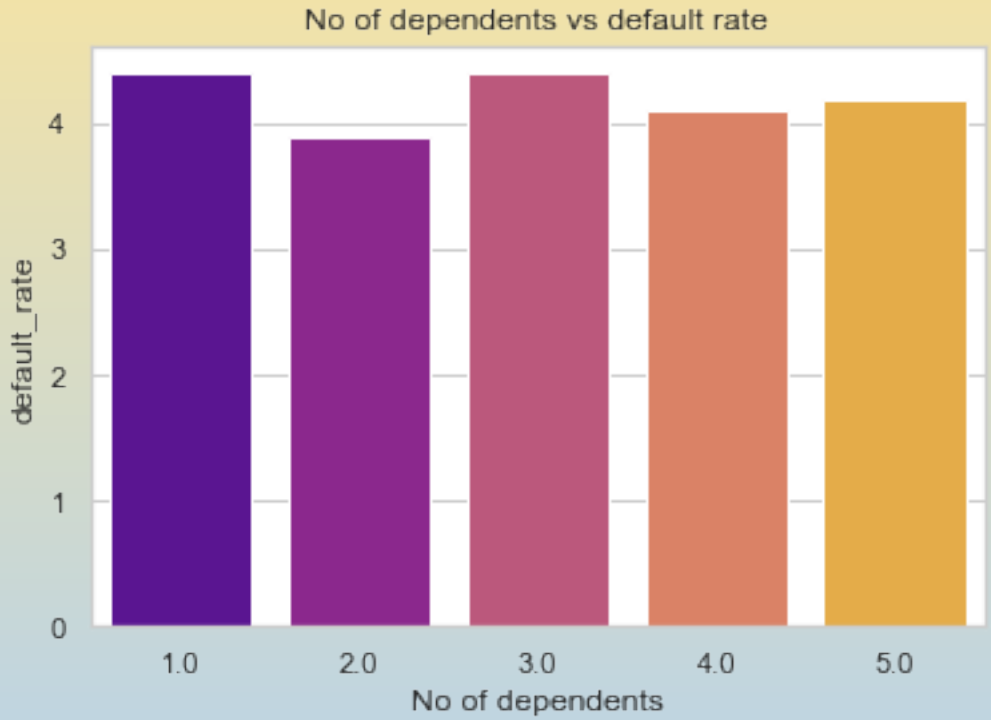


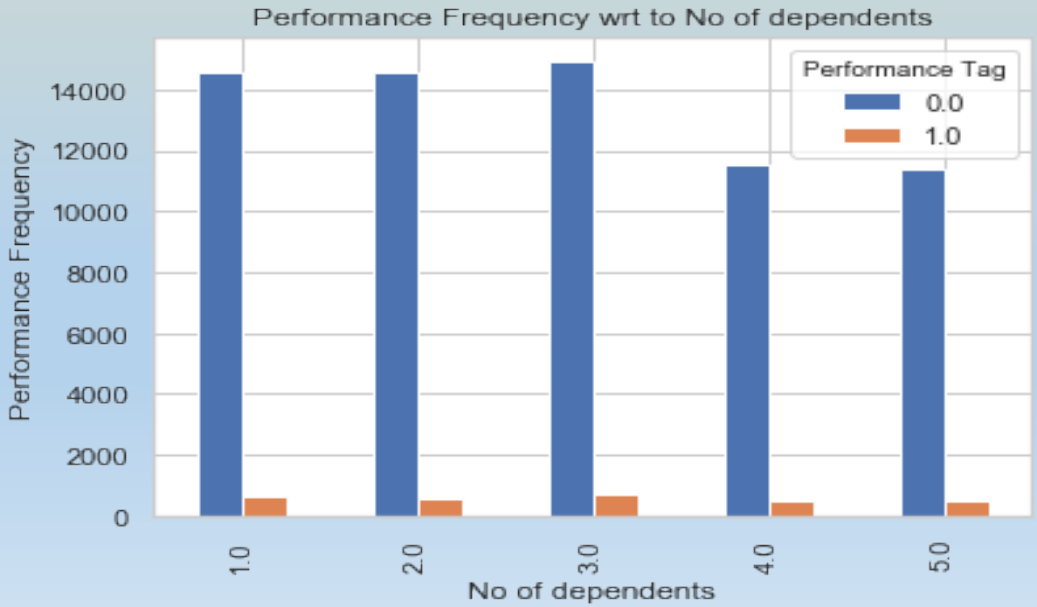


- Married people were offered the majority of credit cards (about 85%).
- The default rate is similar across both the categories

Demographic Data Column: No. of dependents

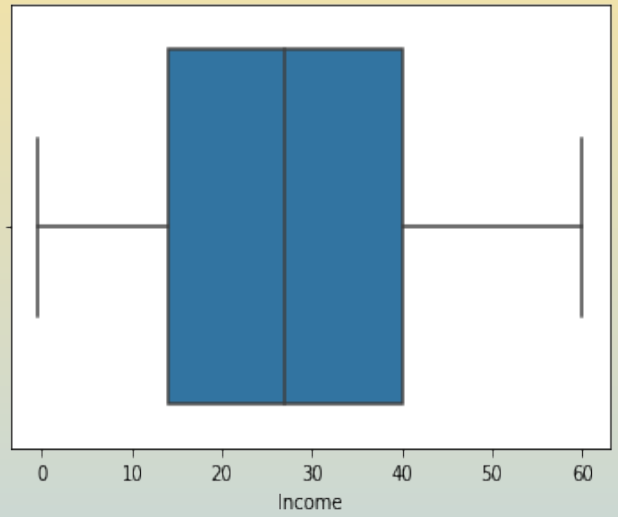


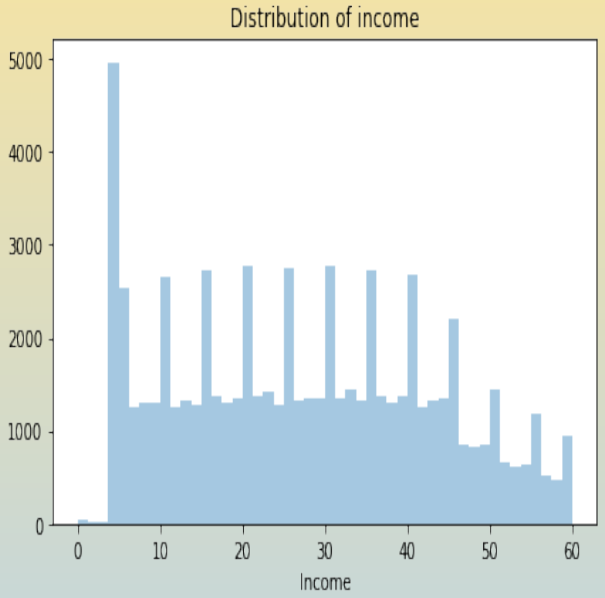


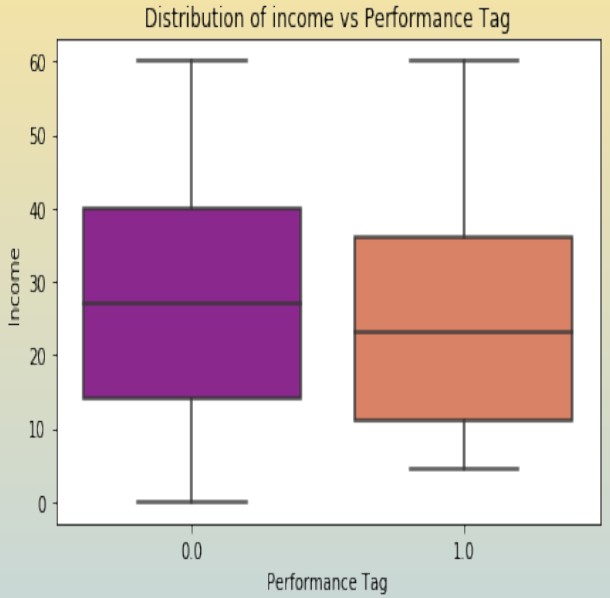


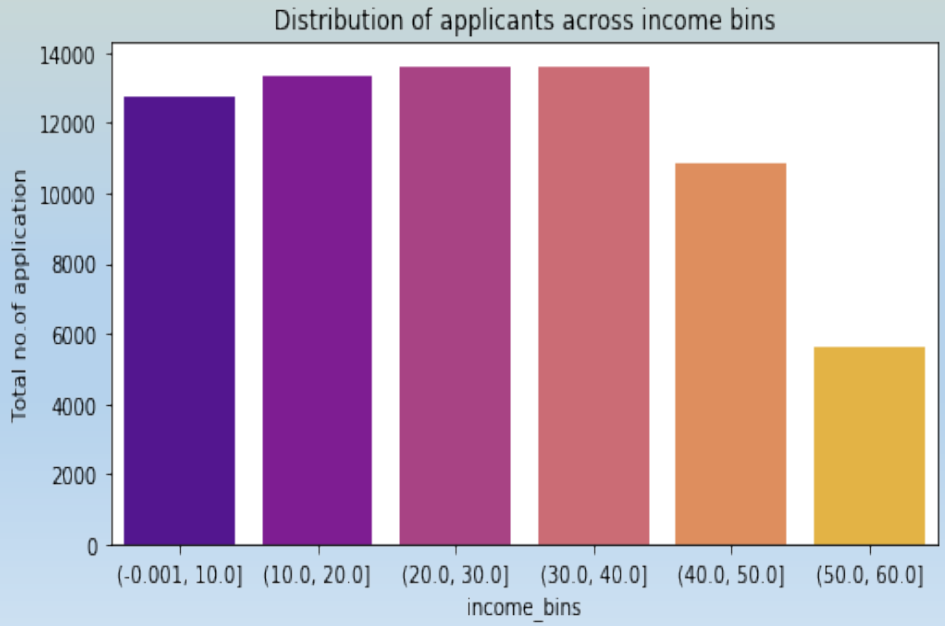
- Maximum credit card holders have atleast 1, 2 or 3 depending on them.
- We note a slight uptick in default rate from people having 1 and 3 dependents on them when compared to others.

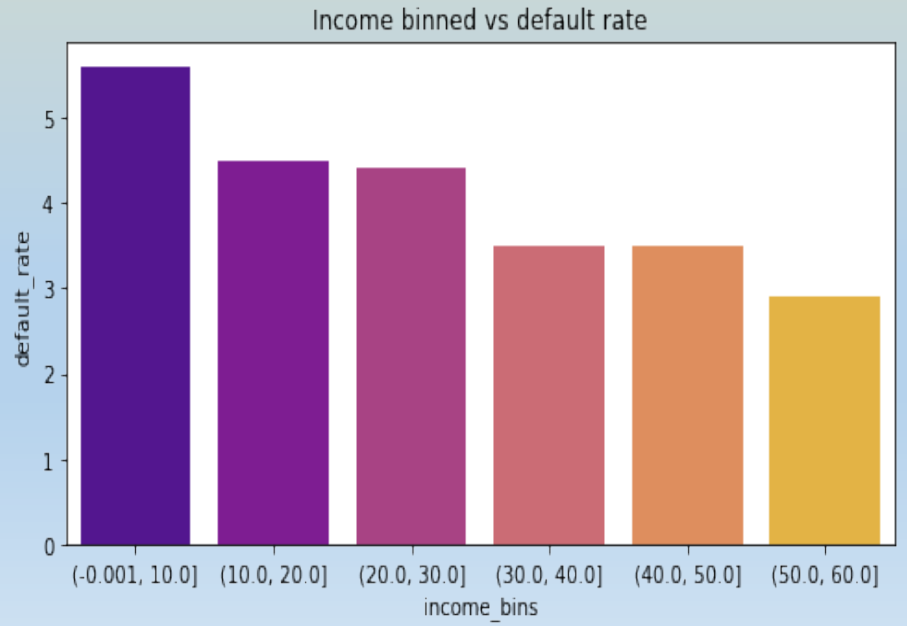
Demographic Data Column: Income





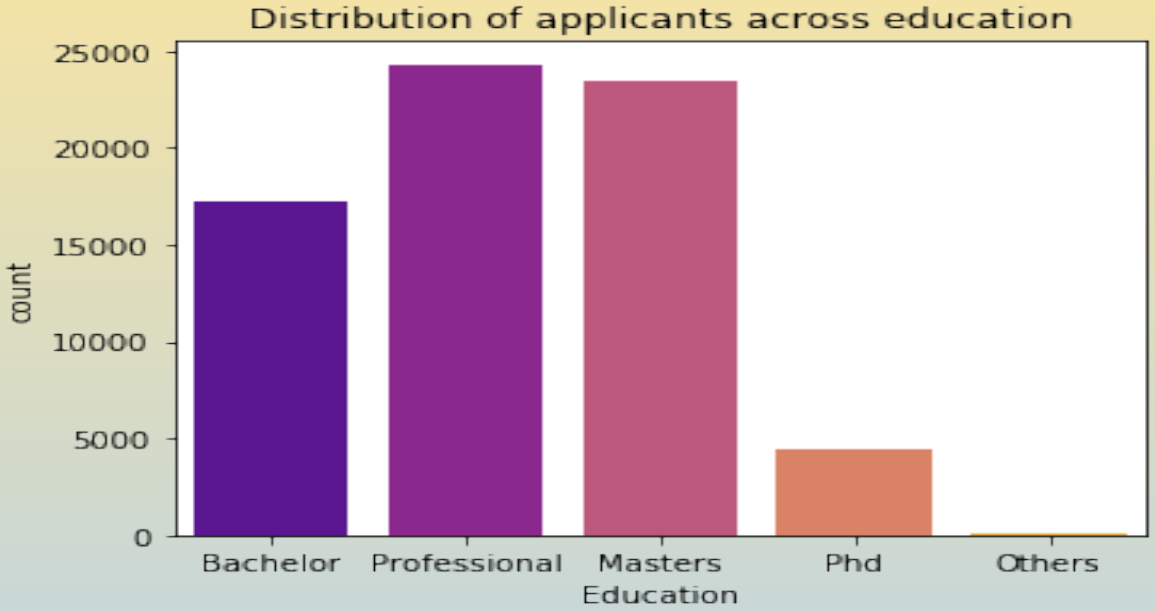


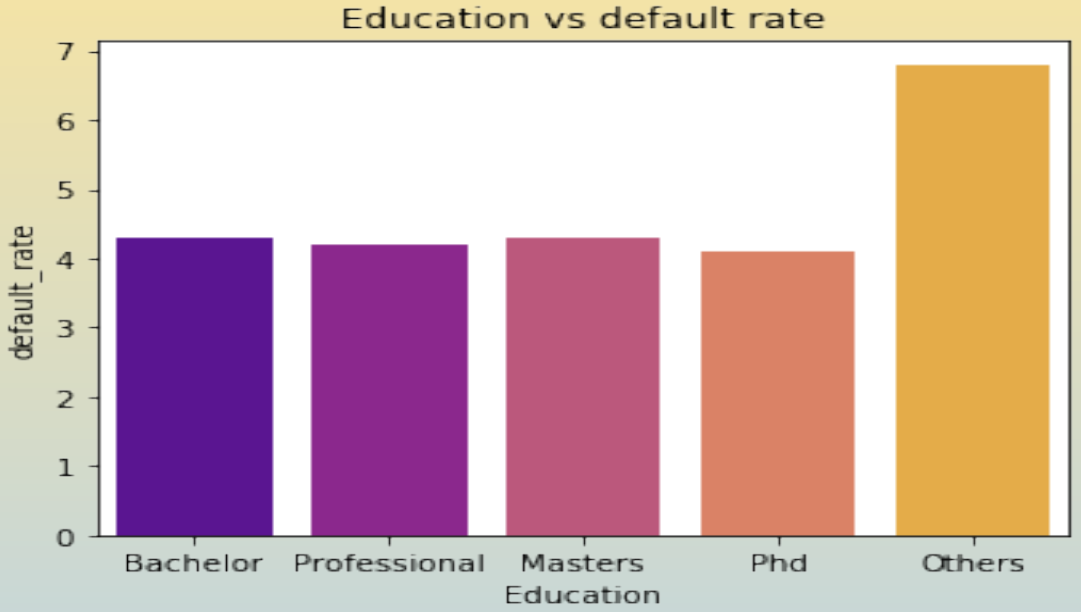


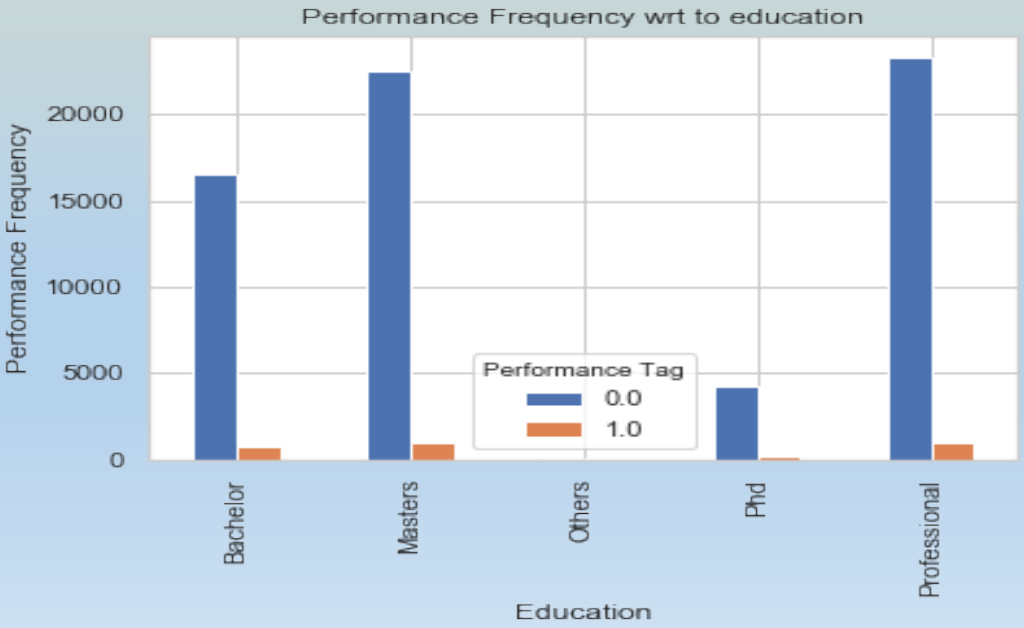


- We note a data quality issue in 'Income' column.
- Few data entries had negative income. We have assumed the minimum income to be as 0.0.
- A large number of credit cards were issued to people with income less than 5Lakh(assuming income is in INR lakhs).
- The median income for defaulters is lower than the median income for defaulters.
- Lesser the income, higher is the defaulted rate.

Demographic Data Column: Education

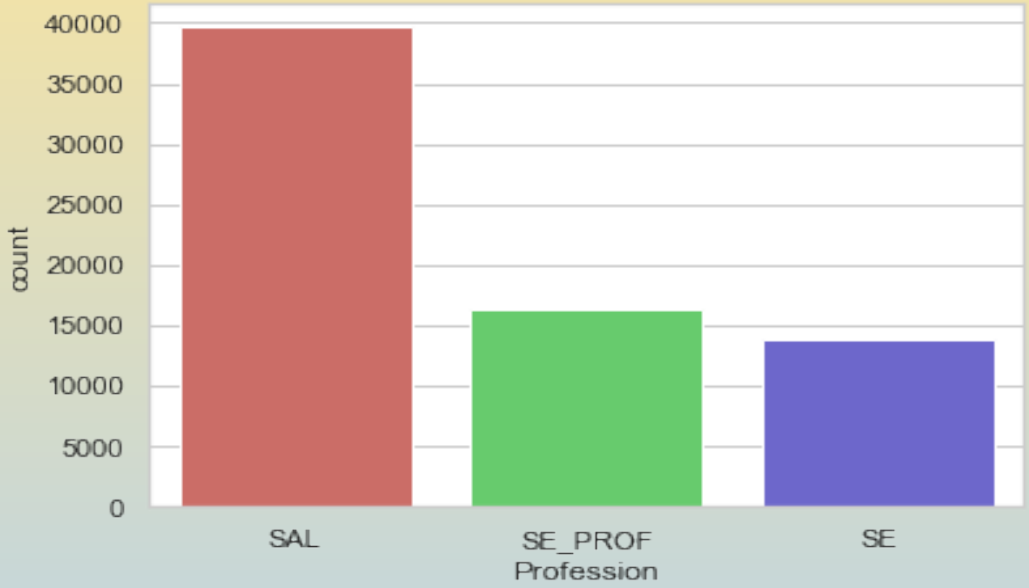


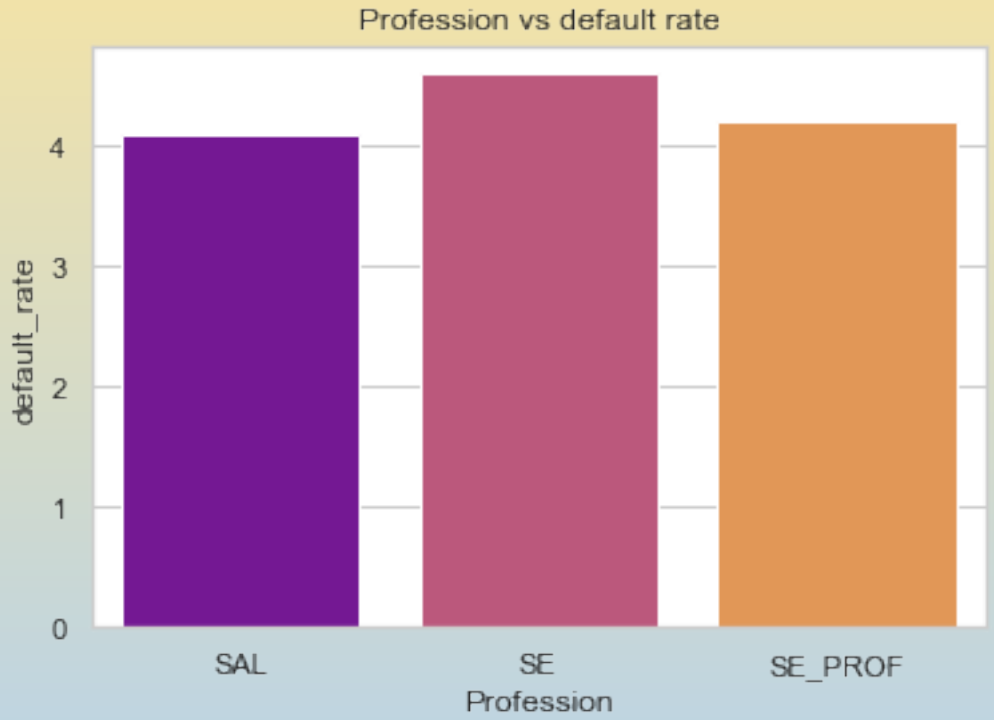


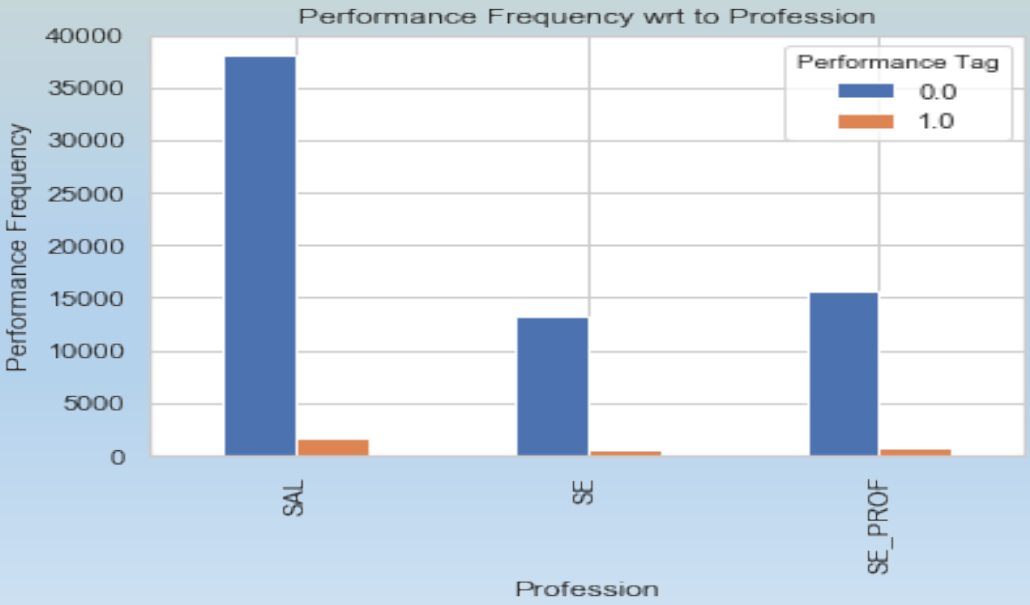


- Maximum credit card holders either are professionals or master degree holders.
- The default rate is similar across all the categories, except for others which is very high because it has negligible proportion of credit card holders when compared with other categories.

Demographic Data Column: Profession

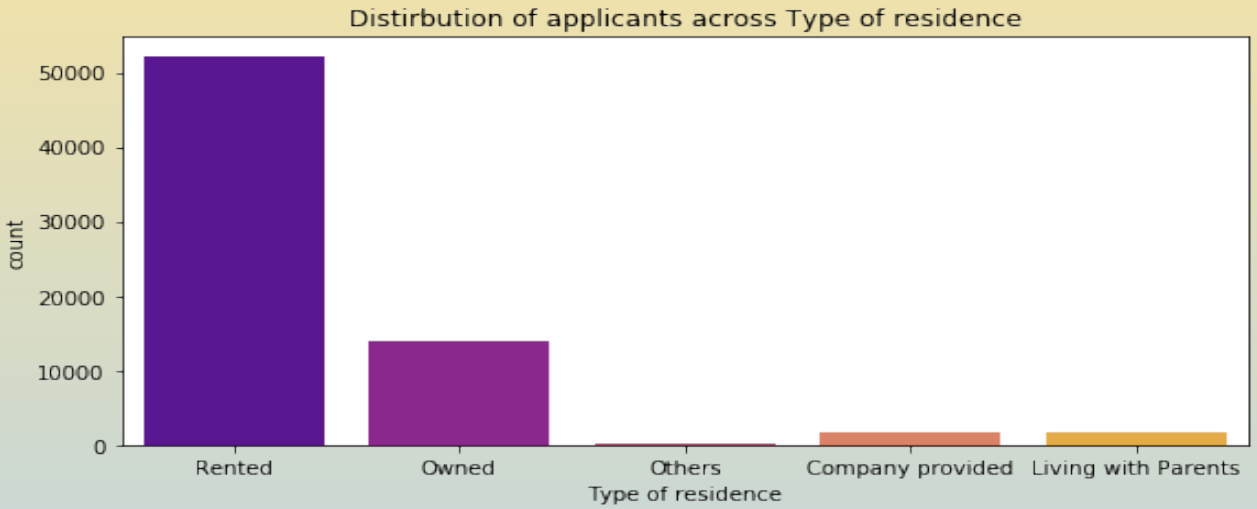


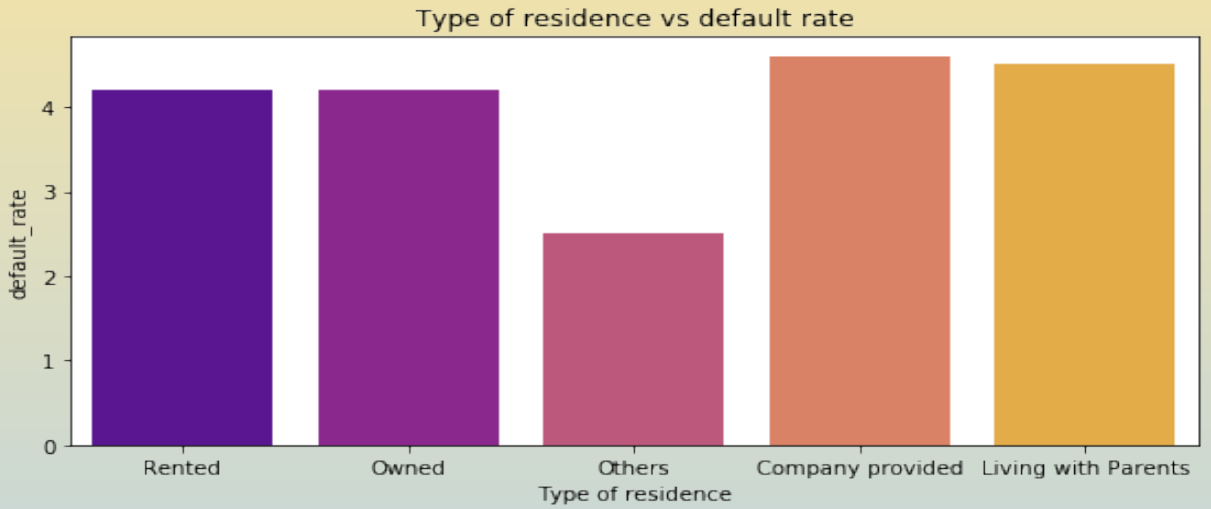


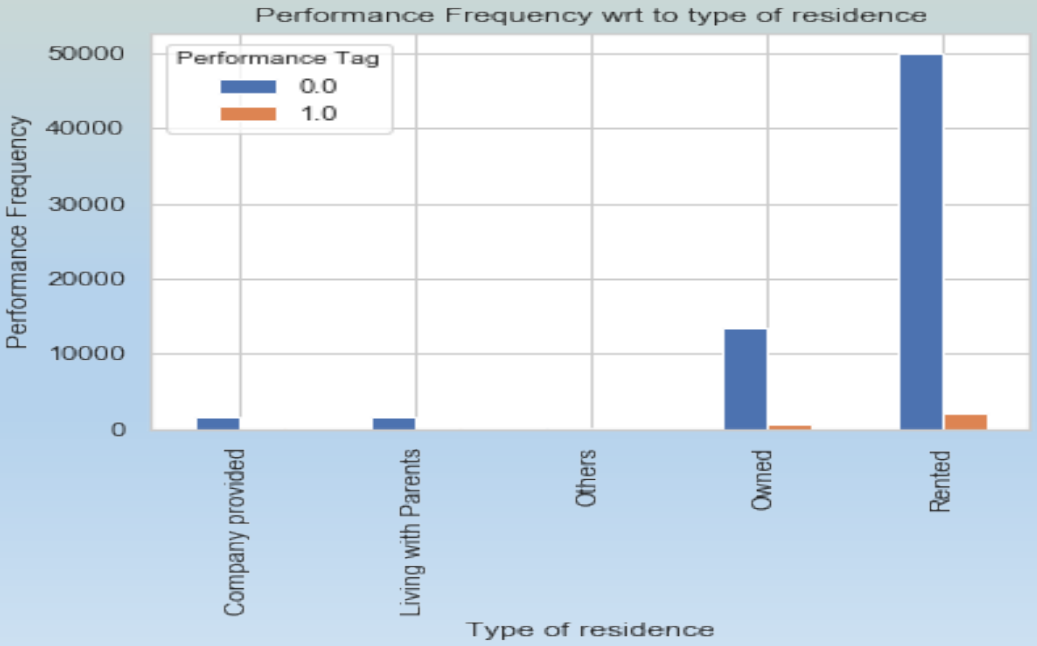


- Maximum credit card holders are ‘salaried’ professionals.
- Self employed people seem to do slightly bad than the salaried people.

Demographic Data Column: Type of residence

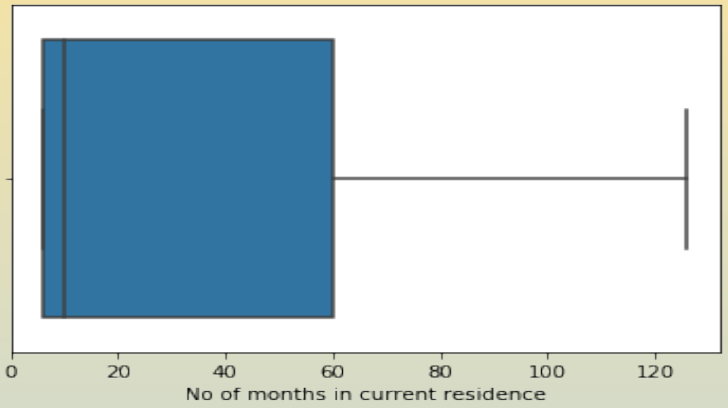


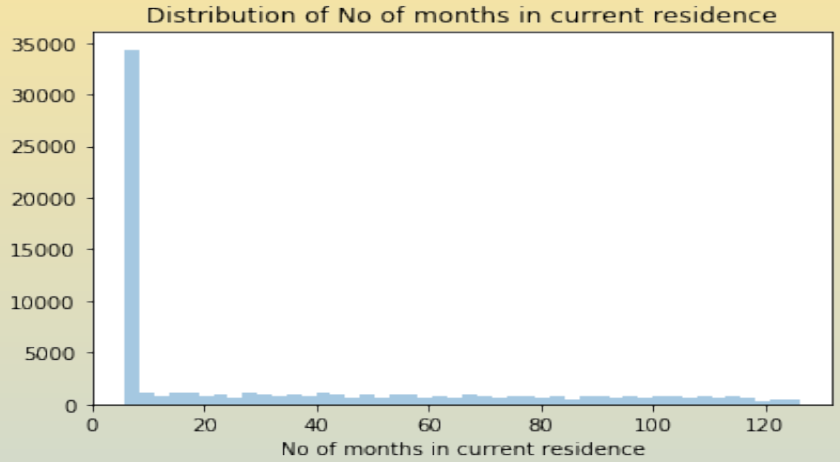


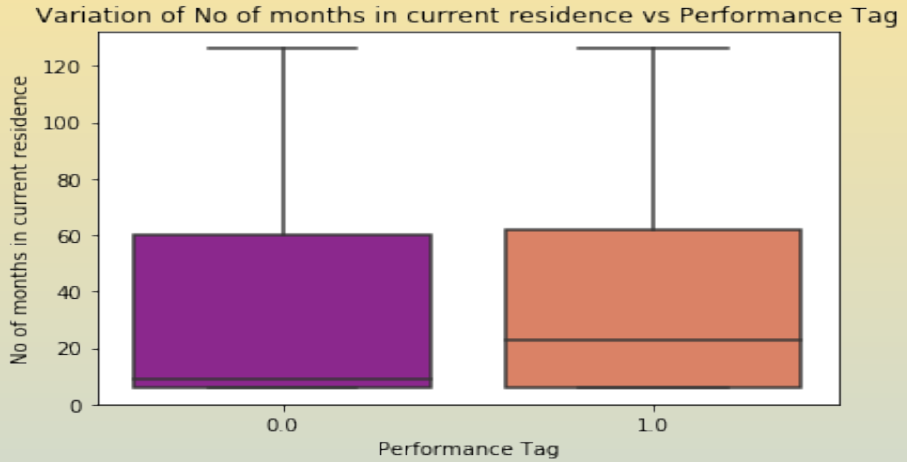


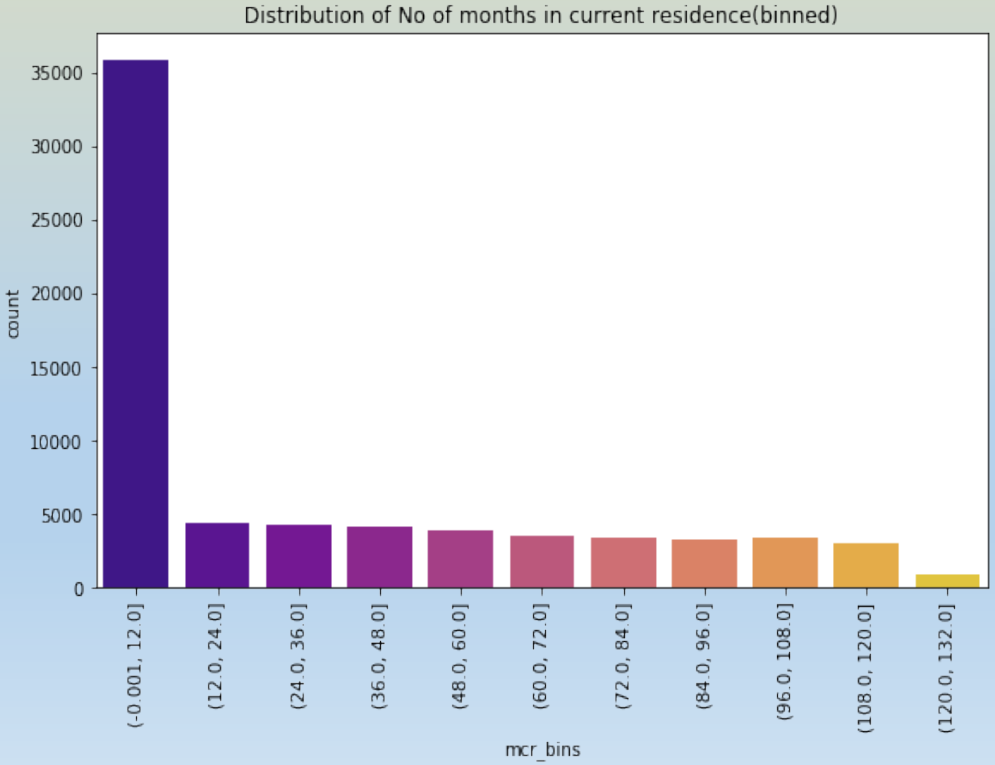
- Maximum credit card holders stay in ‘Rented’ residence.
- We note a slight higher default rate for type of residence- company provided and living with parents when compared to other types. But it is important to note that though they have fewer data entries as well.

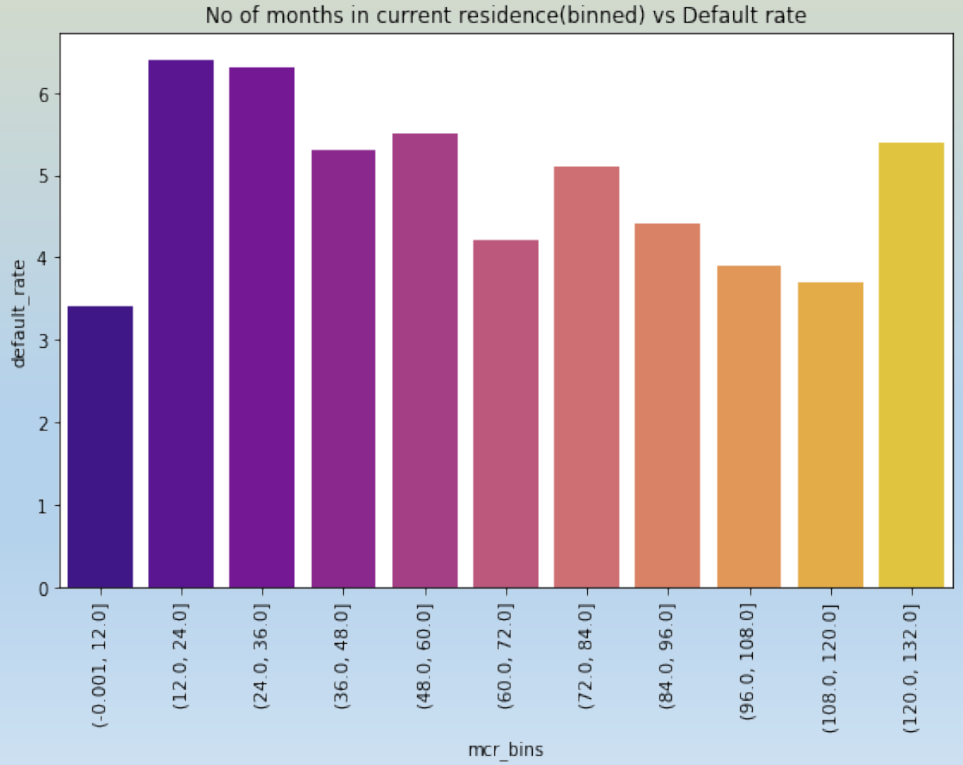
Demographic Data Column: No. of months in current residence





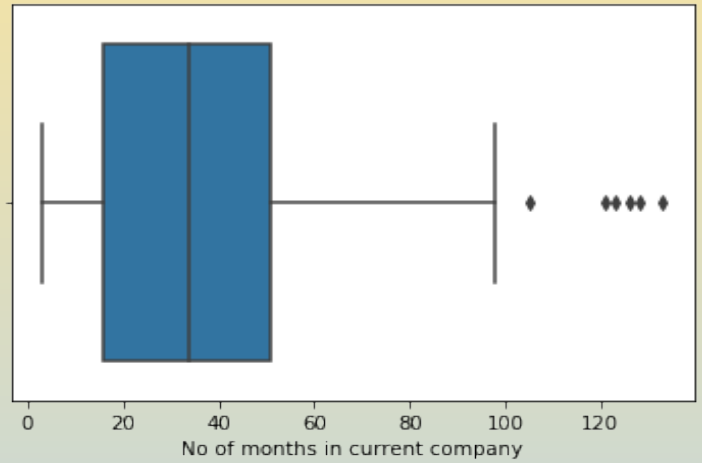


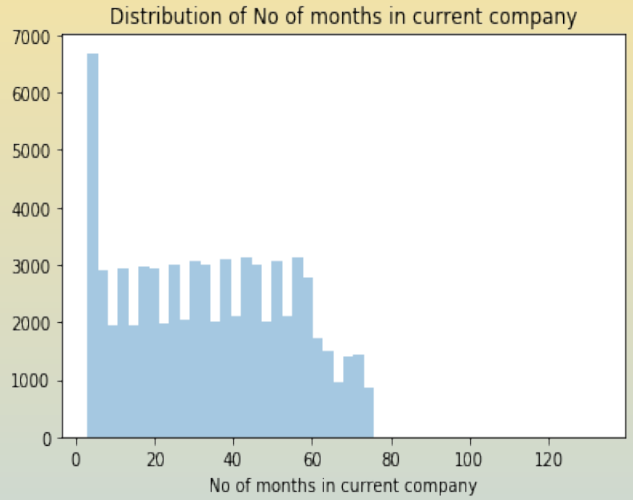


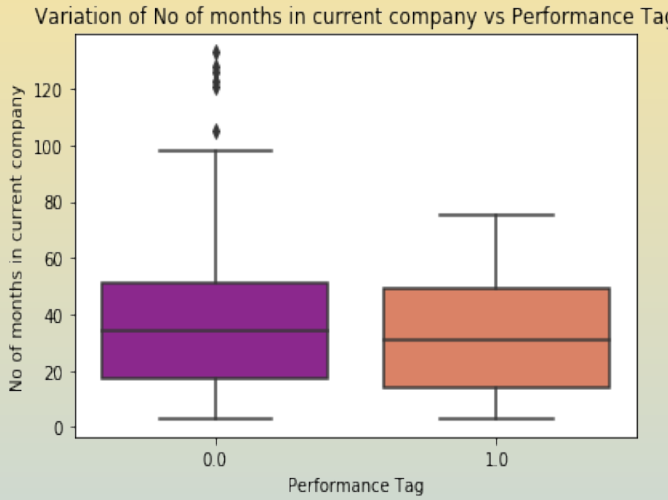


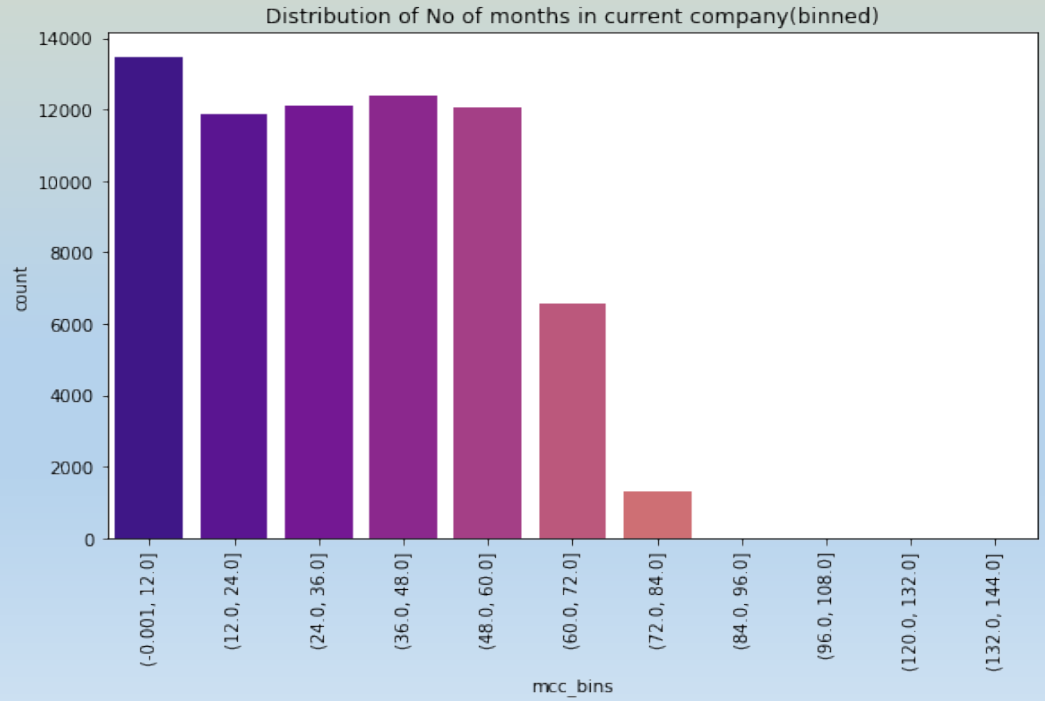
- Maximum credit card was taken in first few months into current residence.
- Leaving the category of less than a year, we can see a general trend of decrease in default rate with increase in number of months in current residence.

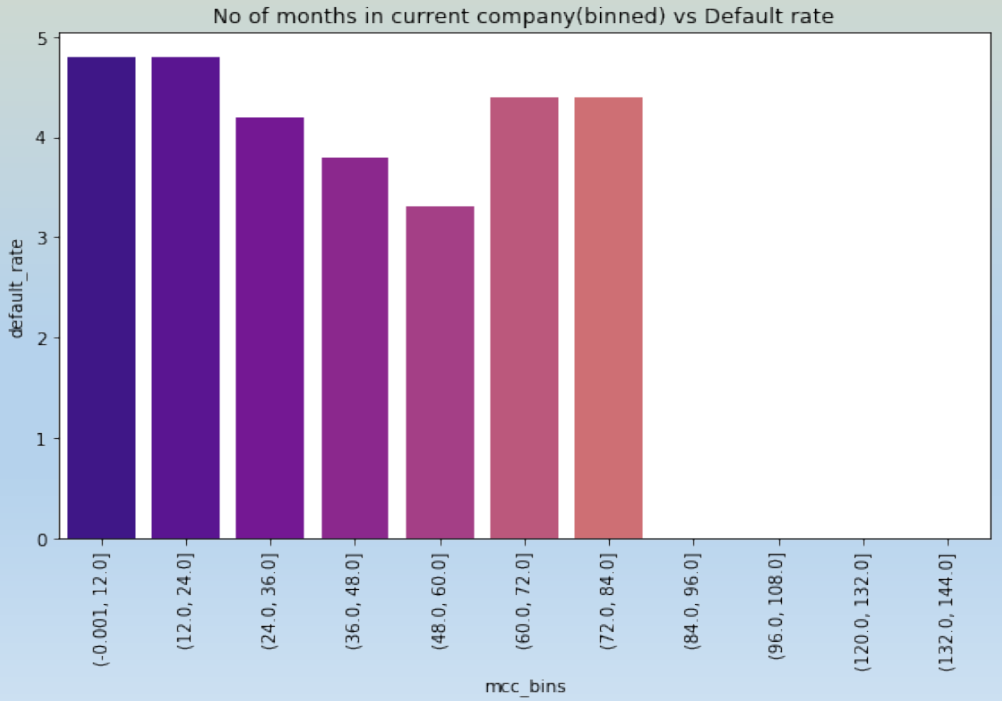
Demographic Data Column: No. of months in current company





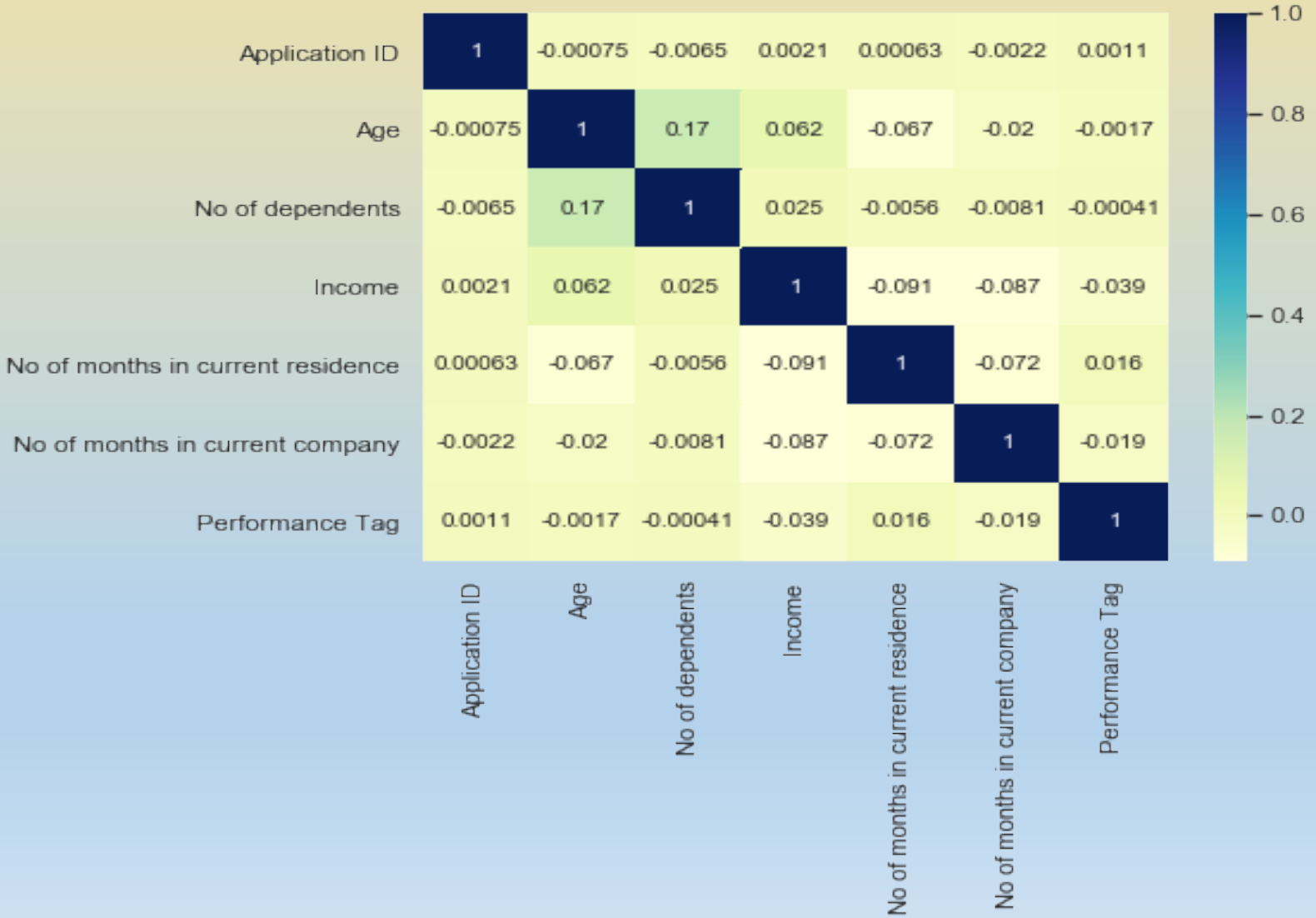






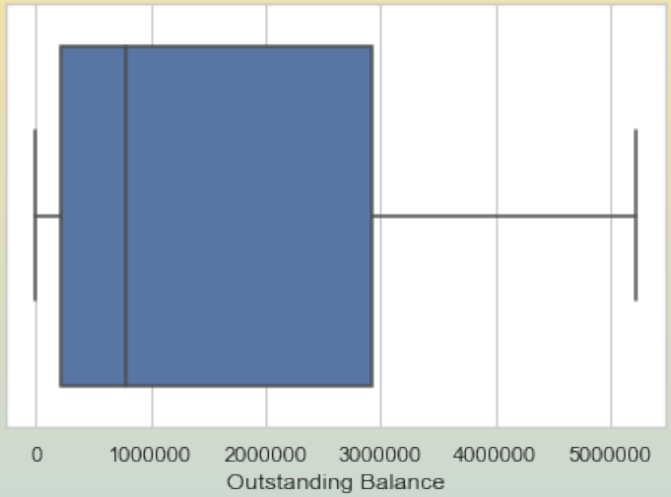
- We note few outliers in this column.
- Maximum credit card was taken in first few months of joining the company.
- Defaulters seem to have slightly lower number of months in their present company when compared to the non defaulters.
- There is this general trend of decrease in default rate when there is a increase in the number of month a person works in his/her present company.
- There is a slight increase in default rate in 5-7 years category, this may be due to the considerably lower proportion of population in these categories.

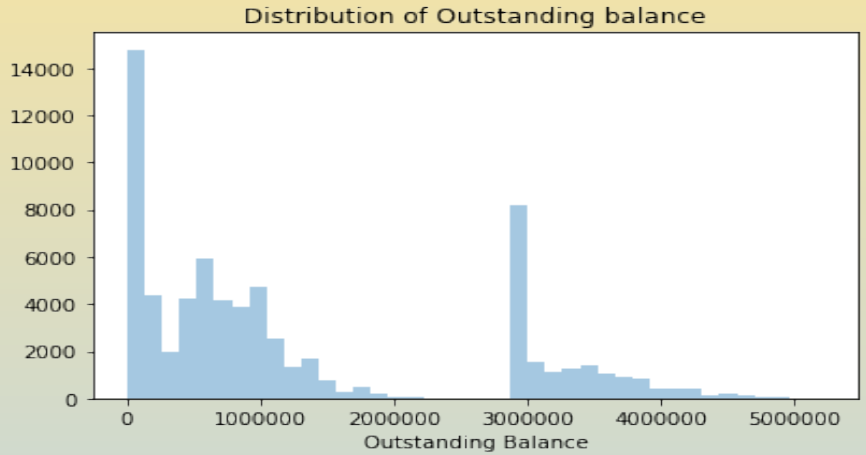
Correlation matrix of demographic Data Columns

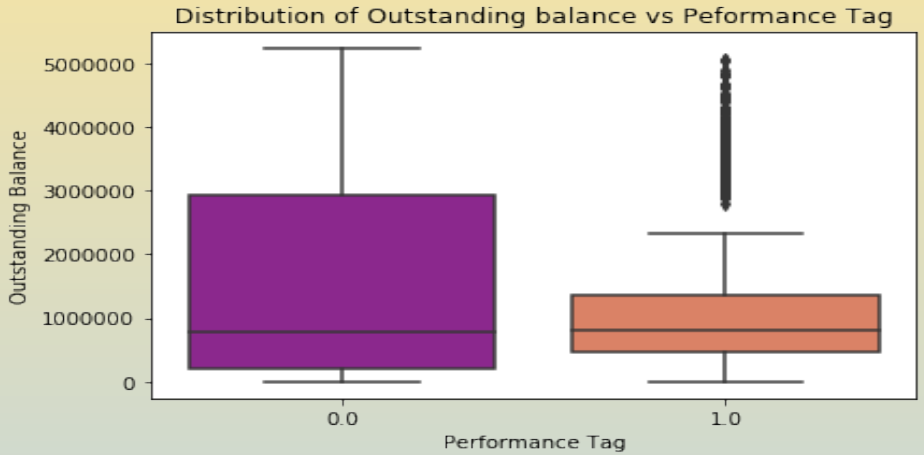


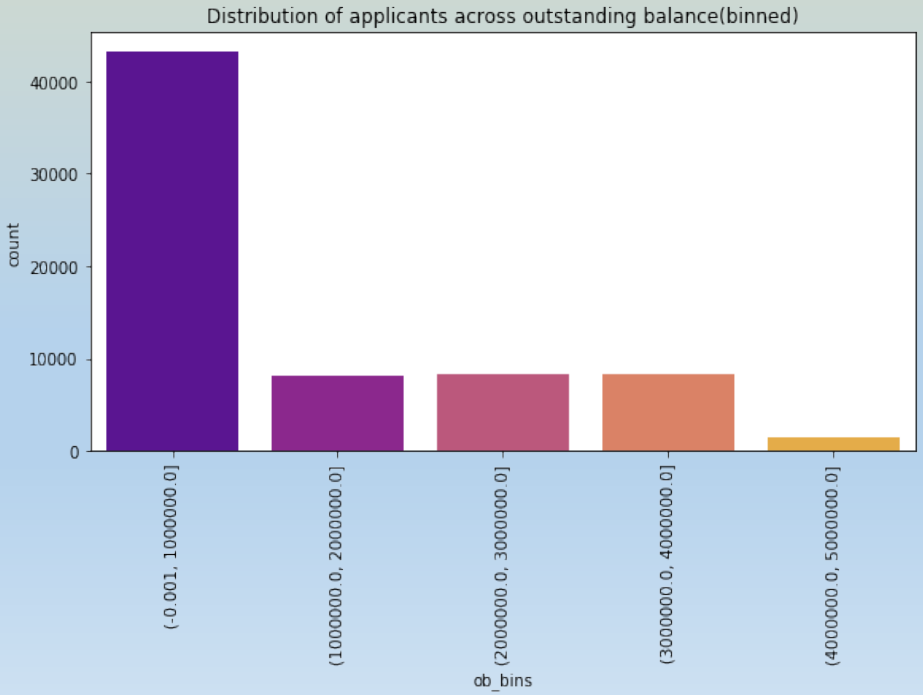
- We don't see any variable having strong correlation with 'Performance tag'.
- Age and 'no. of dependents' have slight correlation with each other.
- Income, no. of months in current residence and no. of months in current company are all negatively correlated to each other.

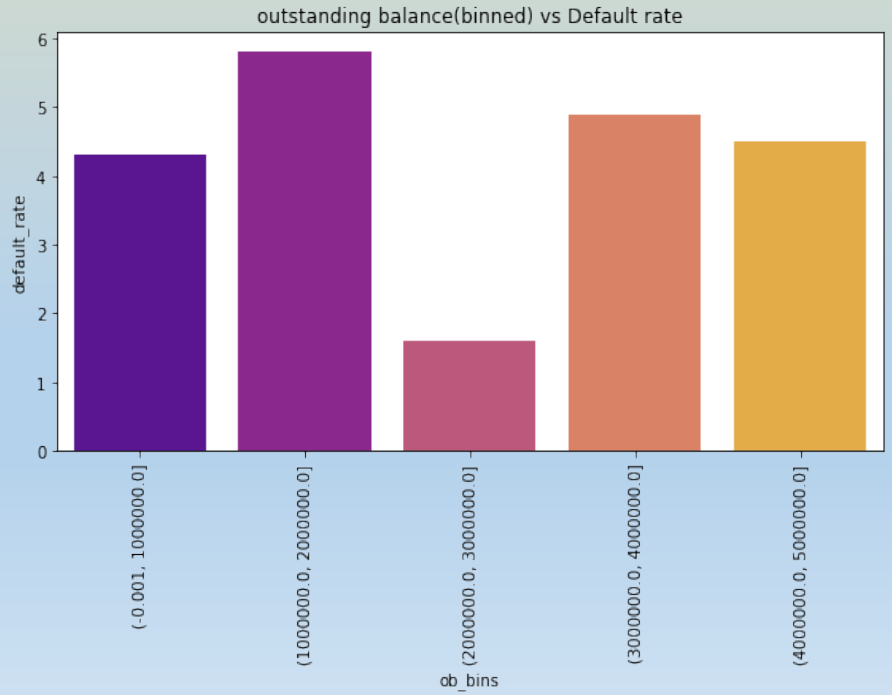
Credit Bureau Data Column: Outstanding Balance





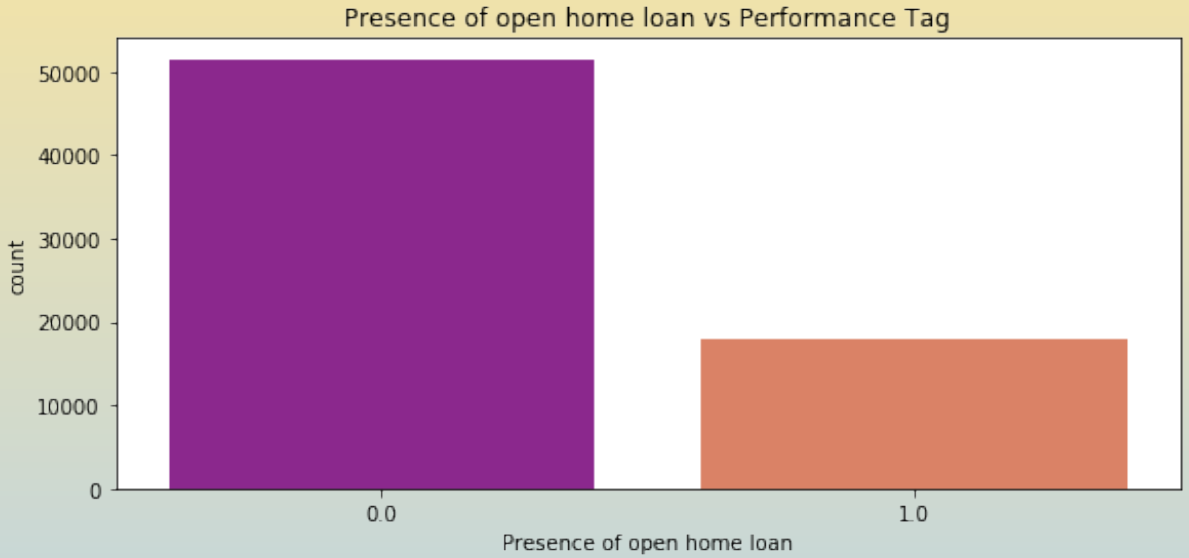


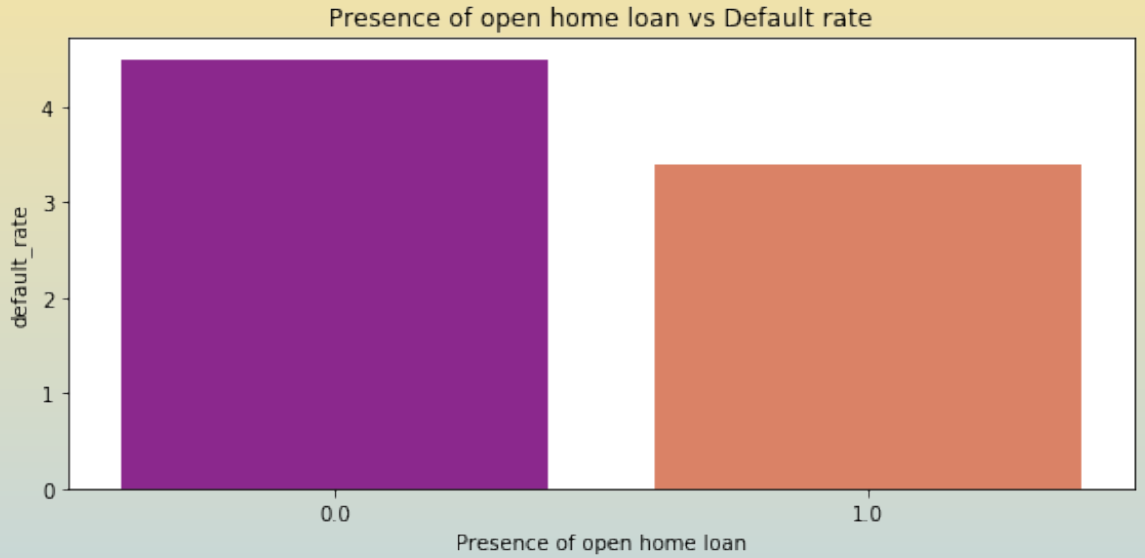


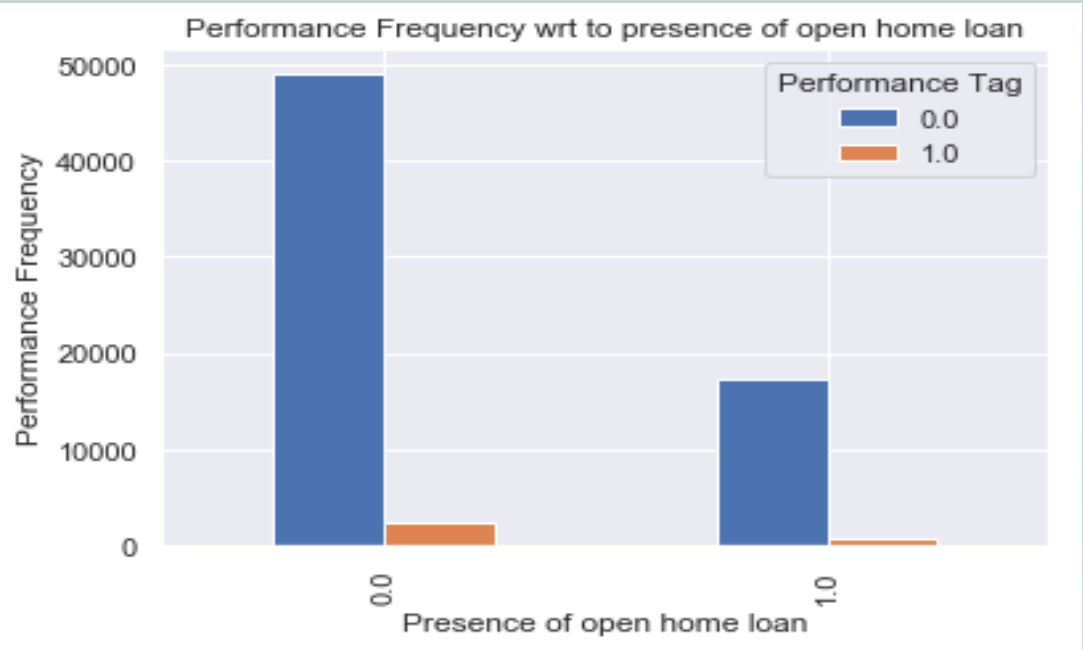


- We note that median is around 7,50,000.
- We see higher default rate among Outstanding Balance range (10,00,000 to 20,00,000). Maximum number of credit card holders fall under Outstanding Balance range (0 to 10,00,000).

Credit Bureau Data Column: Presence of open home loan

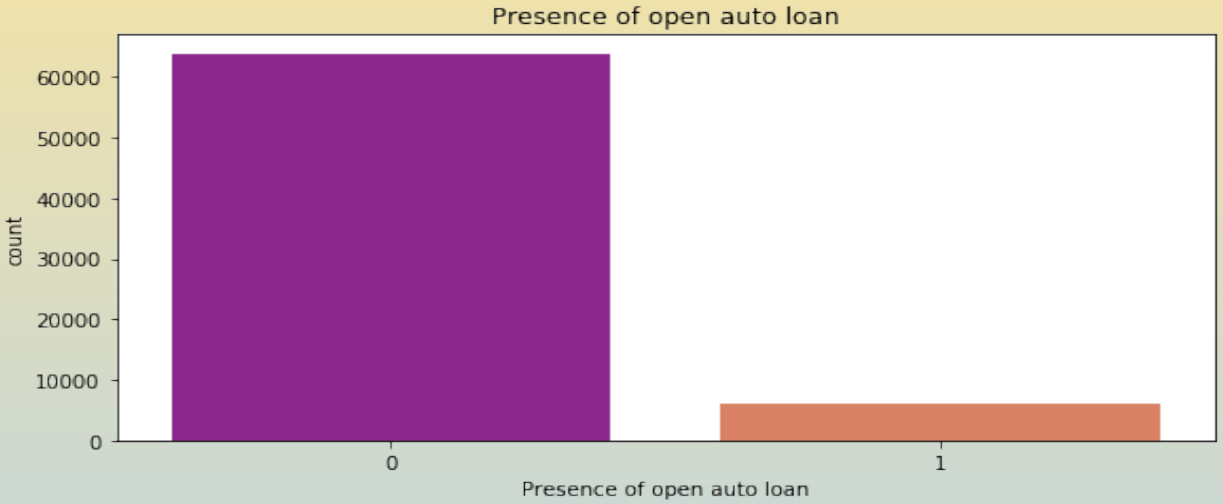


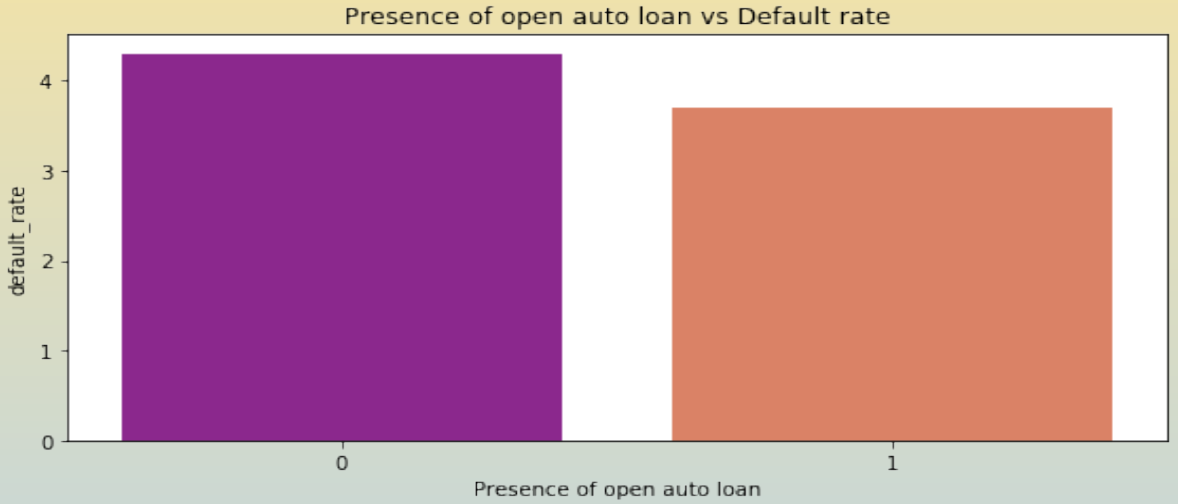


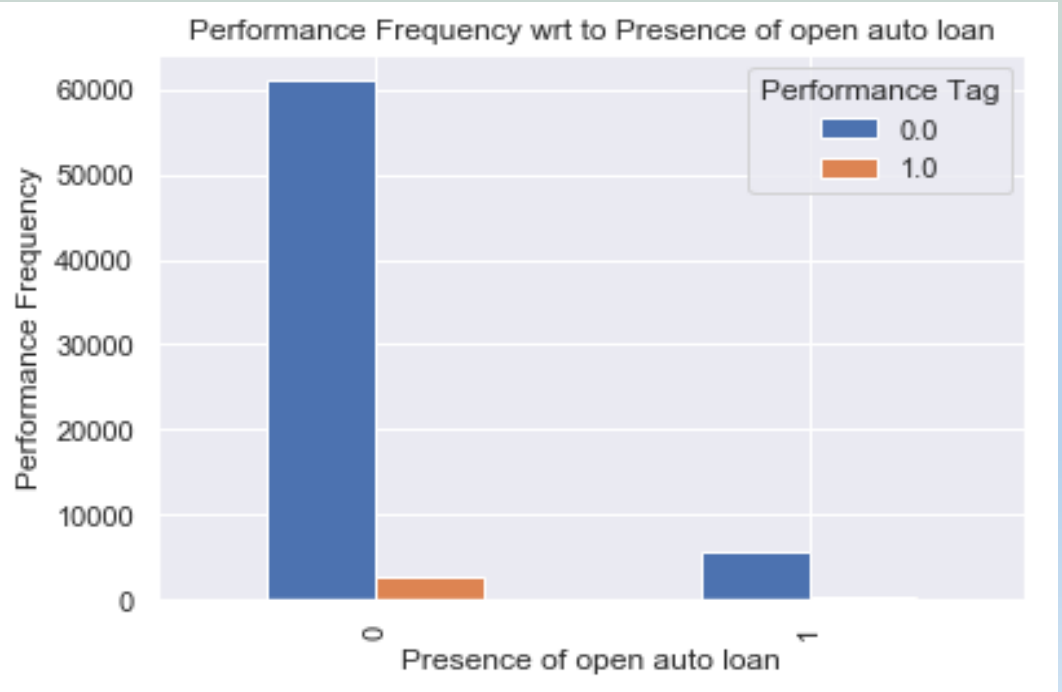


- Maximum credit card holders do not have open home loan.
- We see higher default rate among people who do not have open home loan compared to those who have open home loan.
- But let us also note that we have literally double the number of records for people having **no** open home loan.

Credit Bureau Data Column: Presence of open auto loan

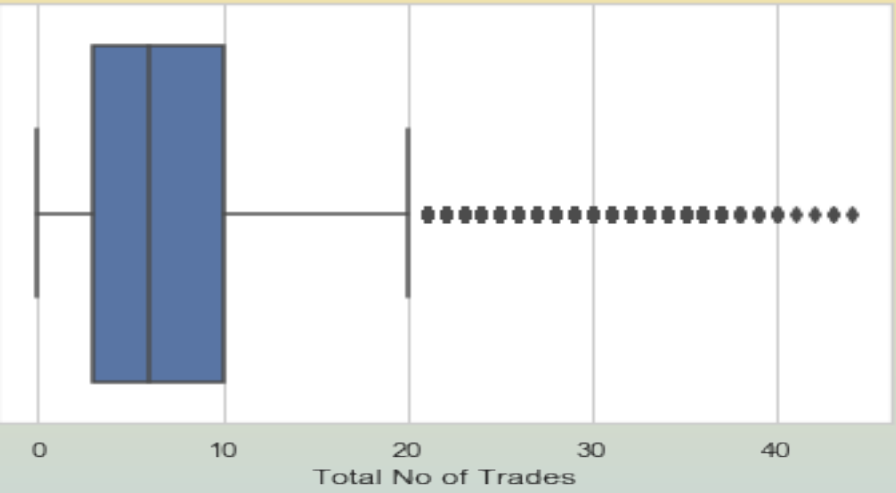


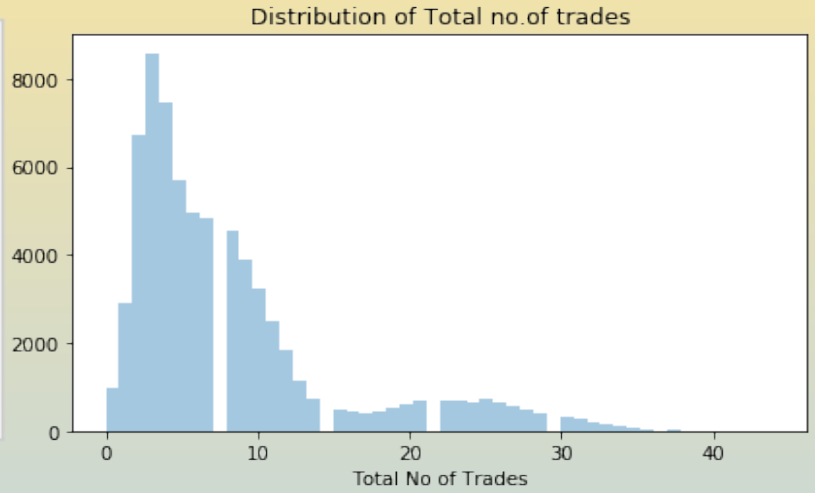


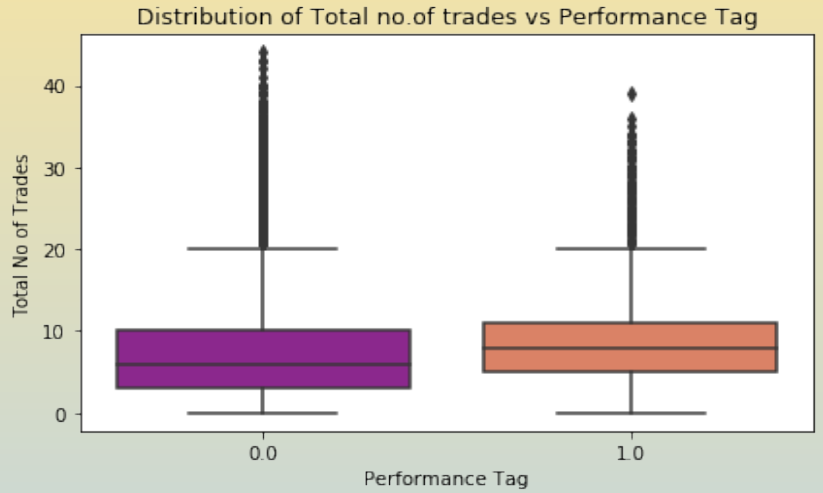


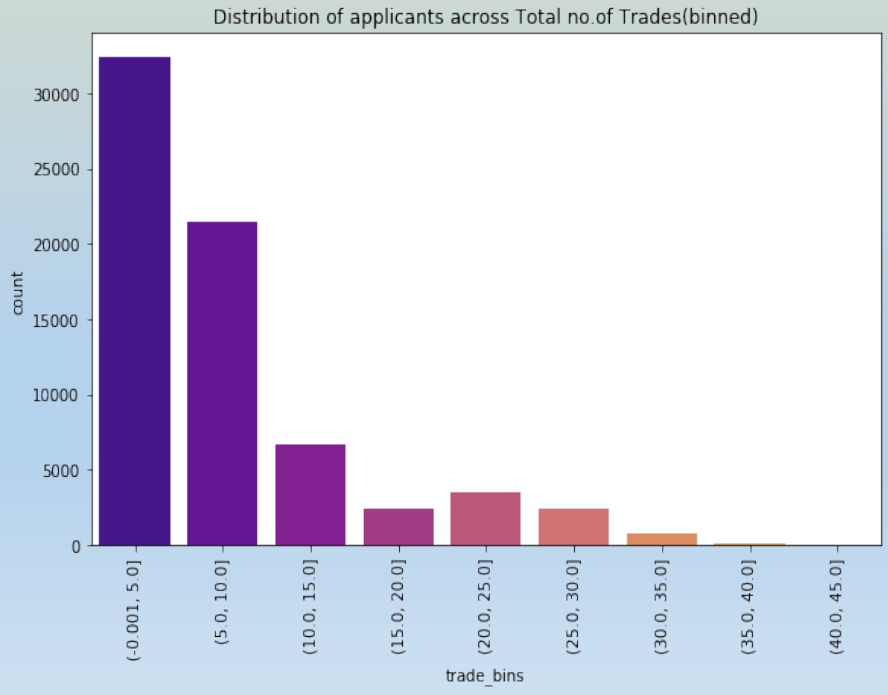
- Maximum credit card holders do not have open auto loan.
- We see higher default rate among people who do not have open auto loan compared to those who have open auto loan.
- But let us also note that we have literally triple the number of records for people having **no** open auto loan.

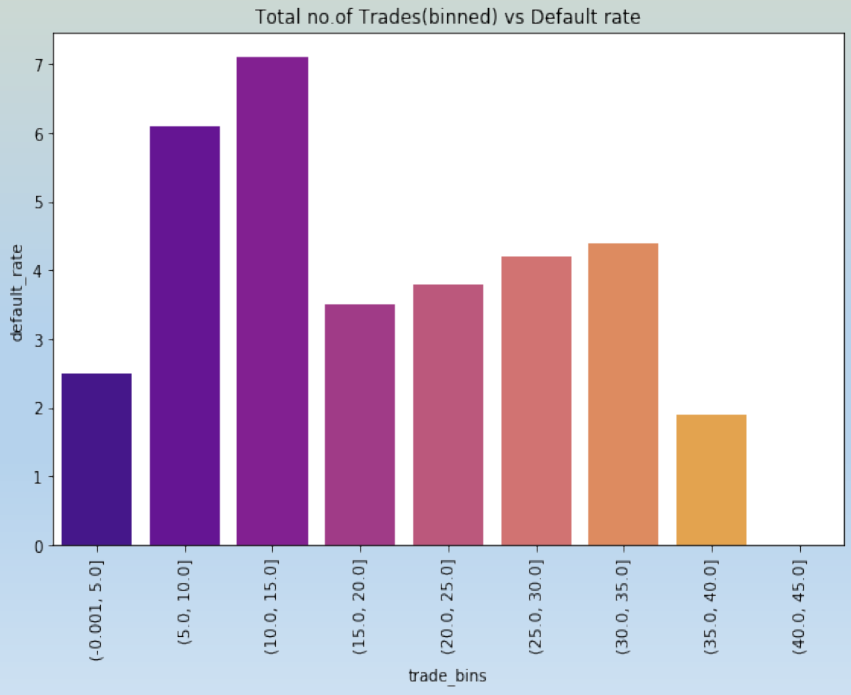
Credit Bureau Data Column: Total no. of trades





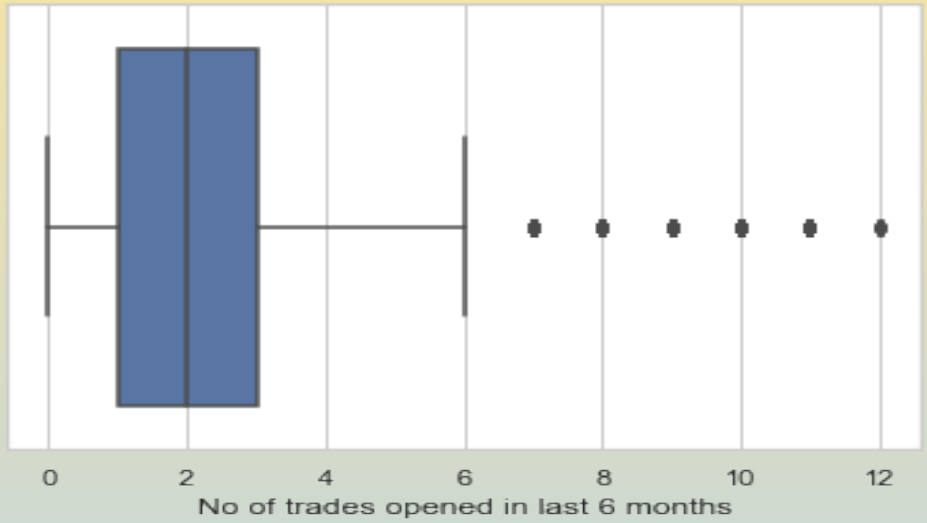


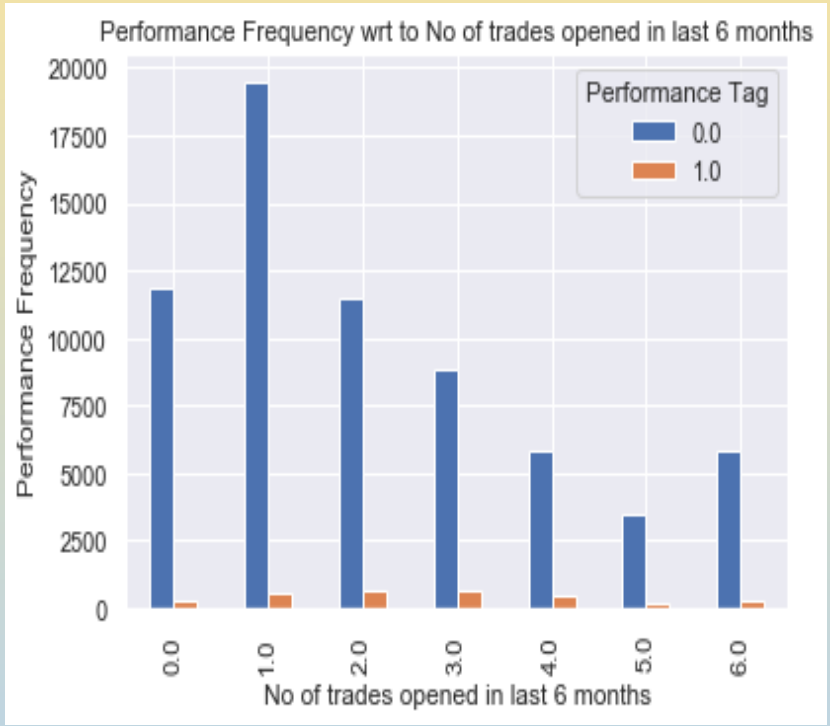


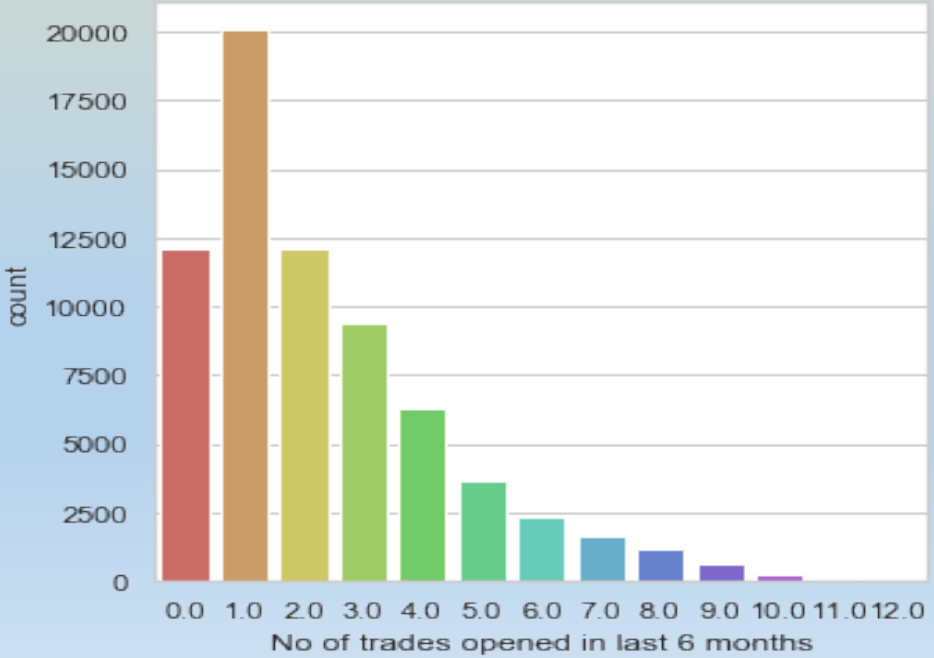


- We note few outliers in this column.
- Maximum credit card holders have done total of 0 to 5 trades.
- We note higher default rate for the total no of trades in the range (10 to 15).
- But as a side note - it is important to note (10 to 15) range have fewer data entries as well.

Credit Bureau Data Column: No. of trades opened in last 6 months

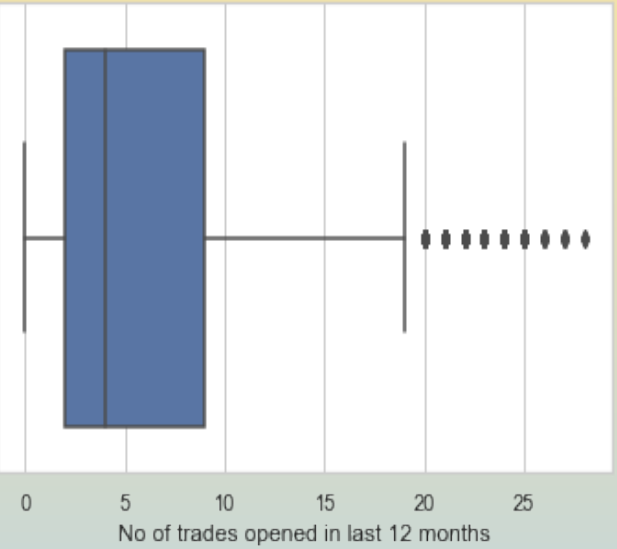


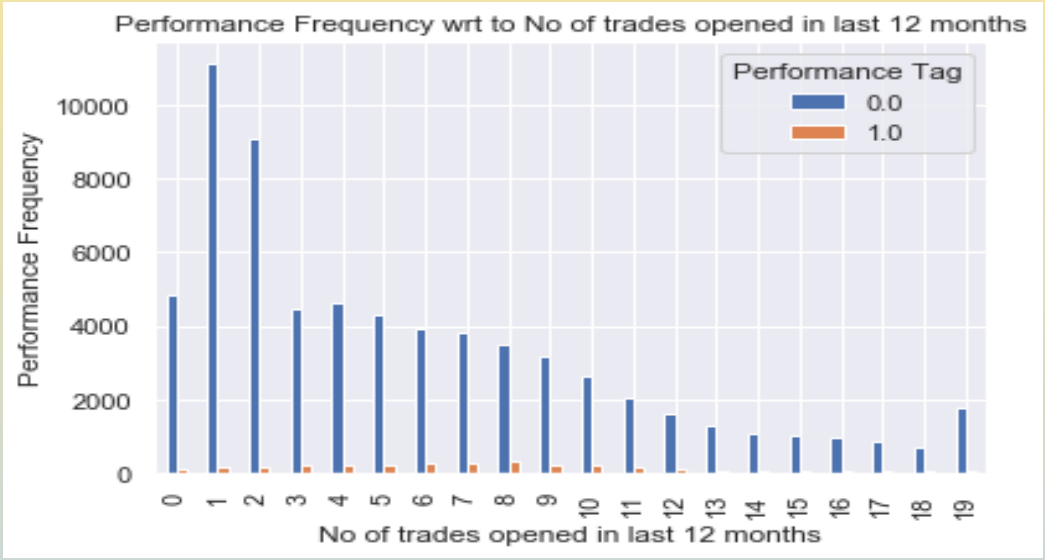




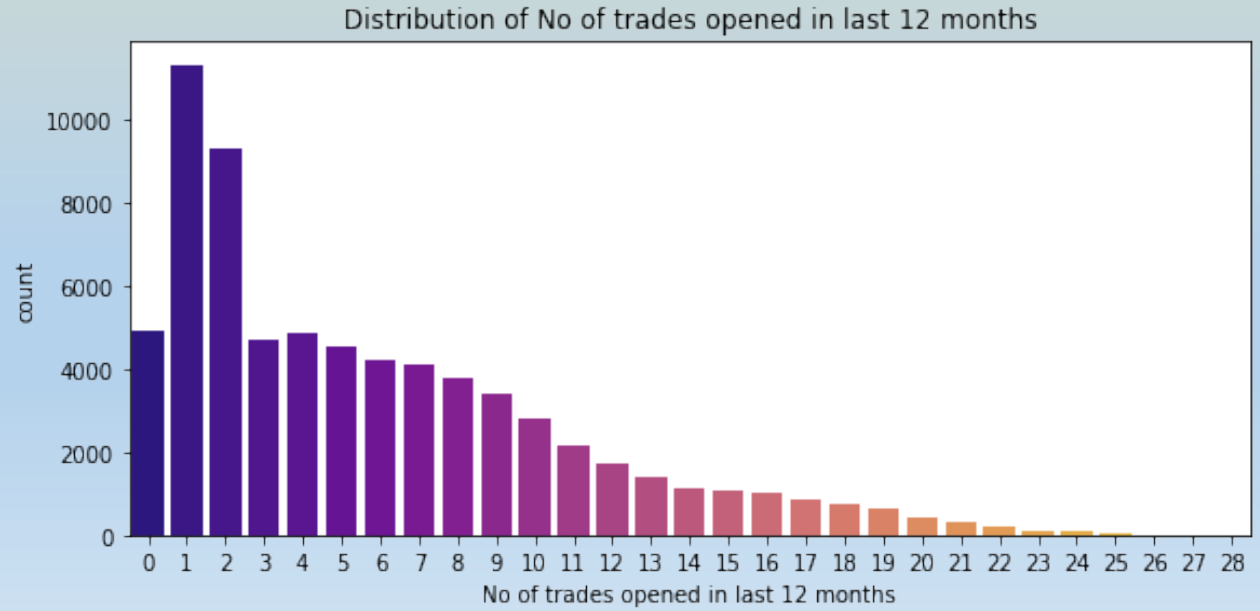
- We note few outliers in this column.
- Maximum credit card holders have opened 0 -4 trade in the last 6 months.
- We note default rate increase as number of trades opened increases in the last 6 months. This trend can be seen till 4 opened trades in the last 6 months.

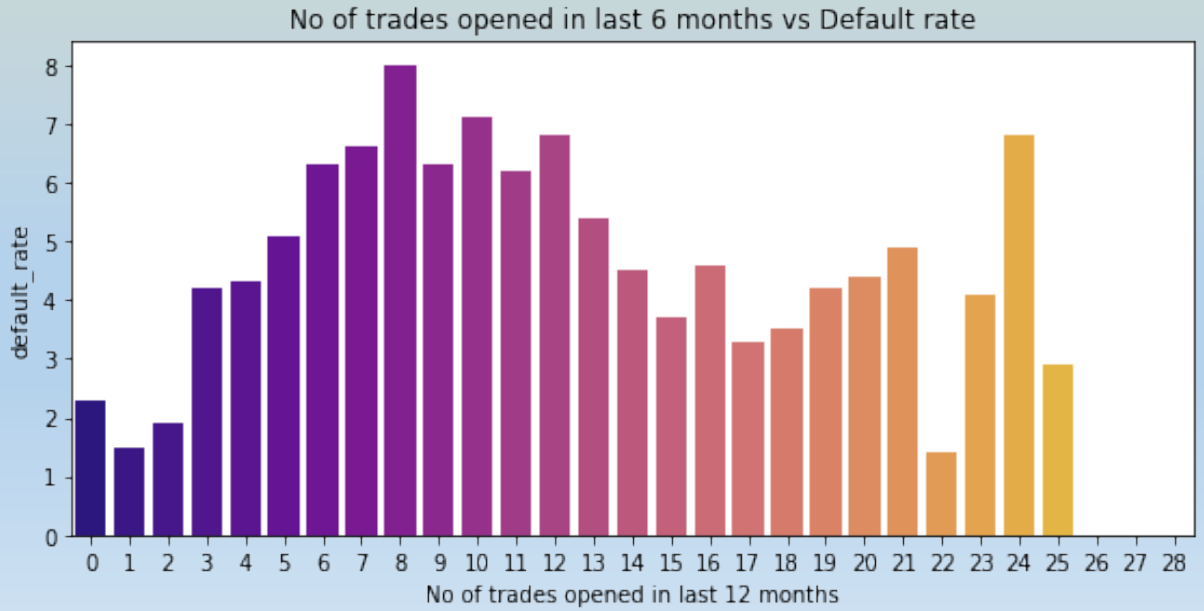
Credit Bureau Data Column: No. of trades opened in last 12 months



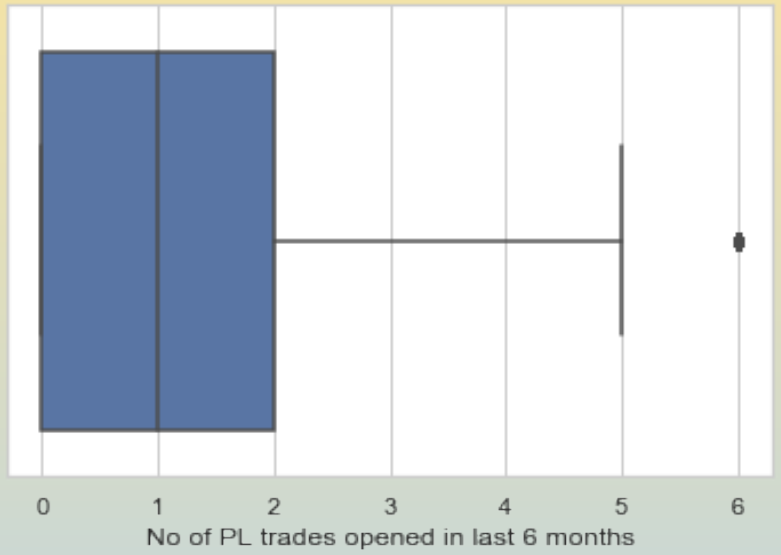


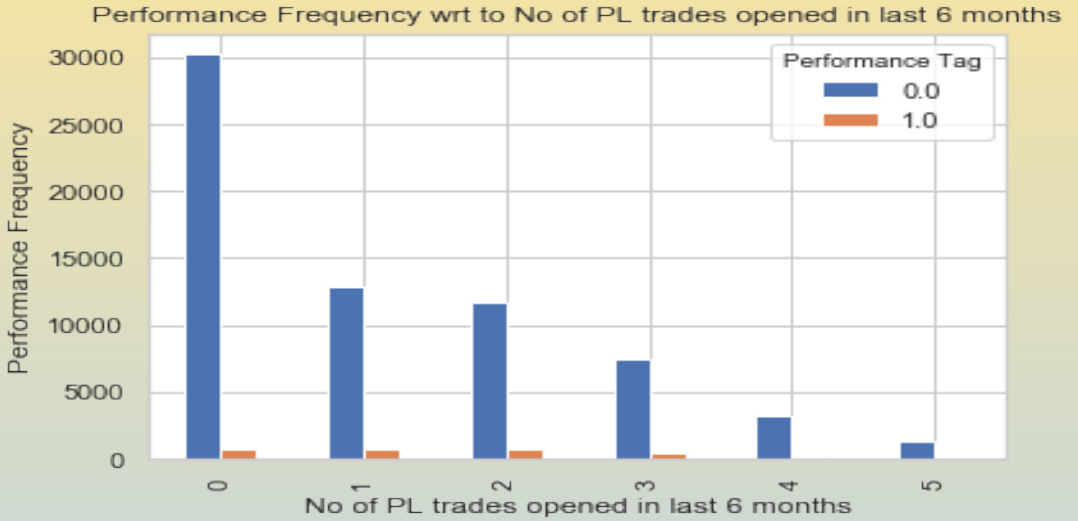
- We note few outliers in this column.
- Maximum credit card holders have opened 1-2 trade in the last 12 months.
- We note default rate increase as number of trades opened increases in the last 12 months. This trend can be seen till 8 opened trades in the last 12 months.

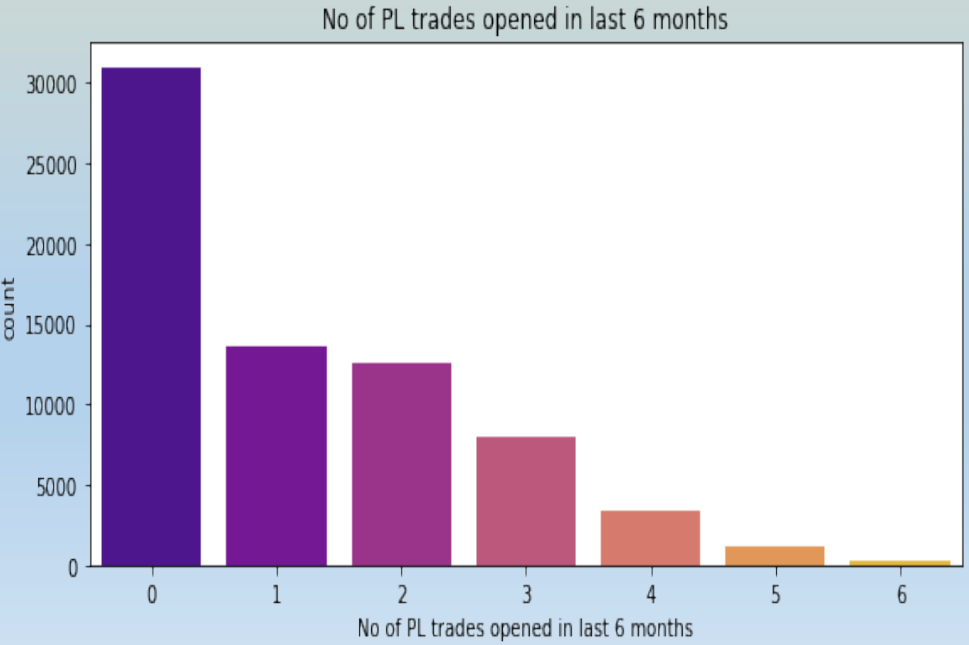


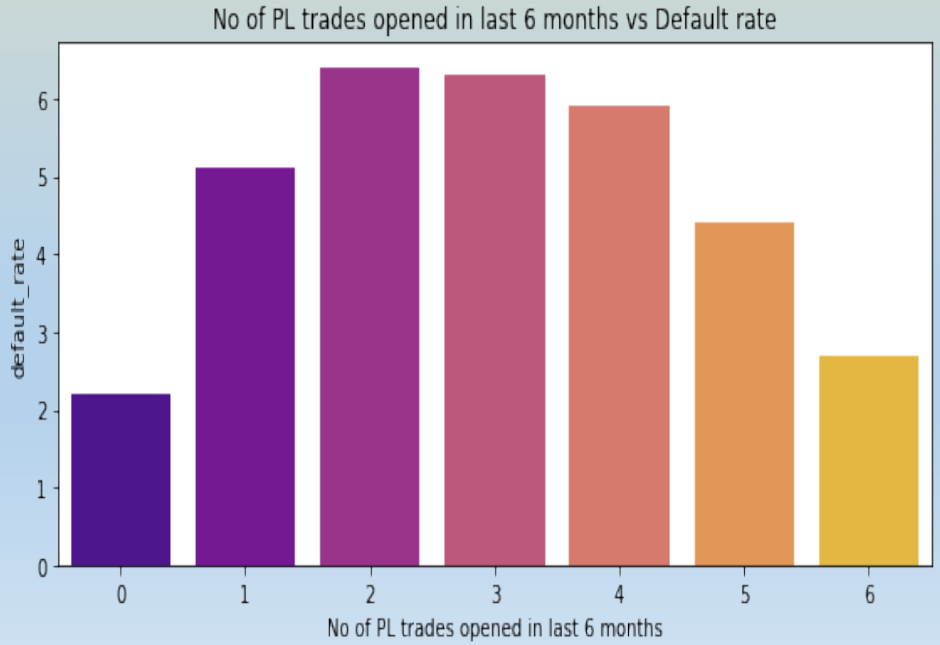


Credit Bureau Data Column: No. of PL trades opened in last 6 months



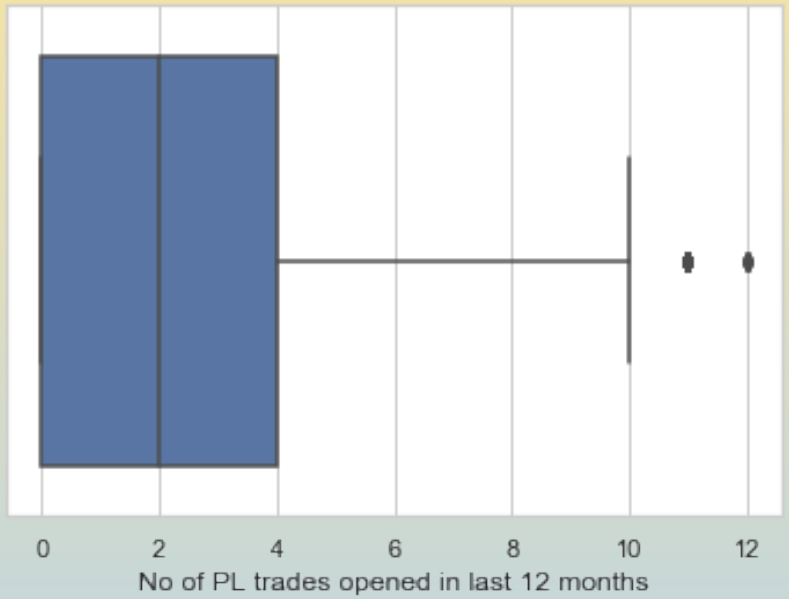


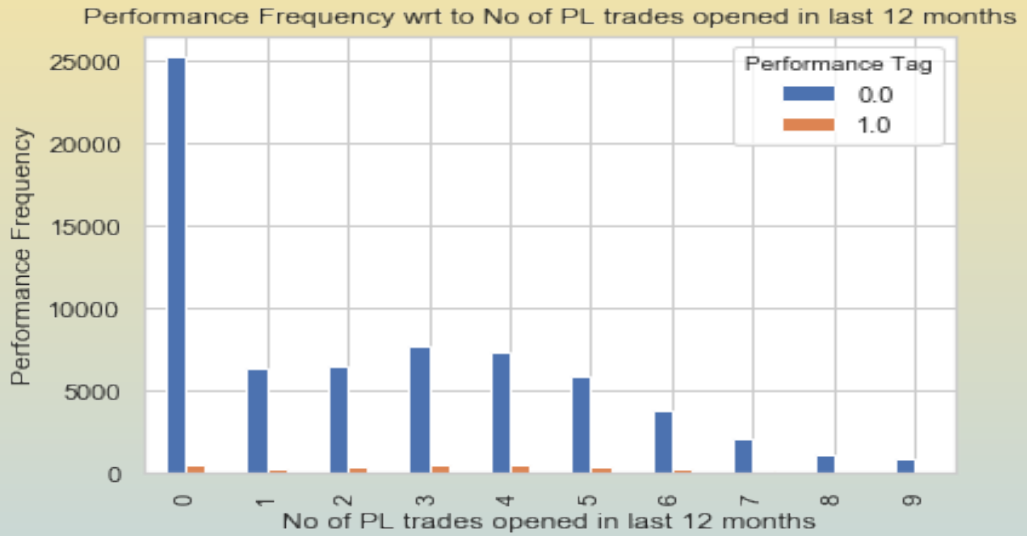


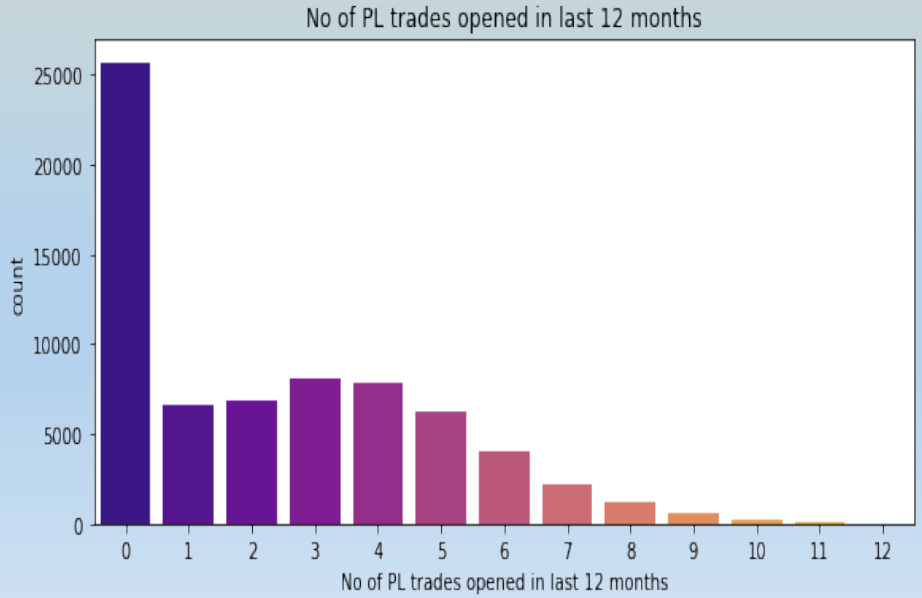


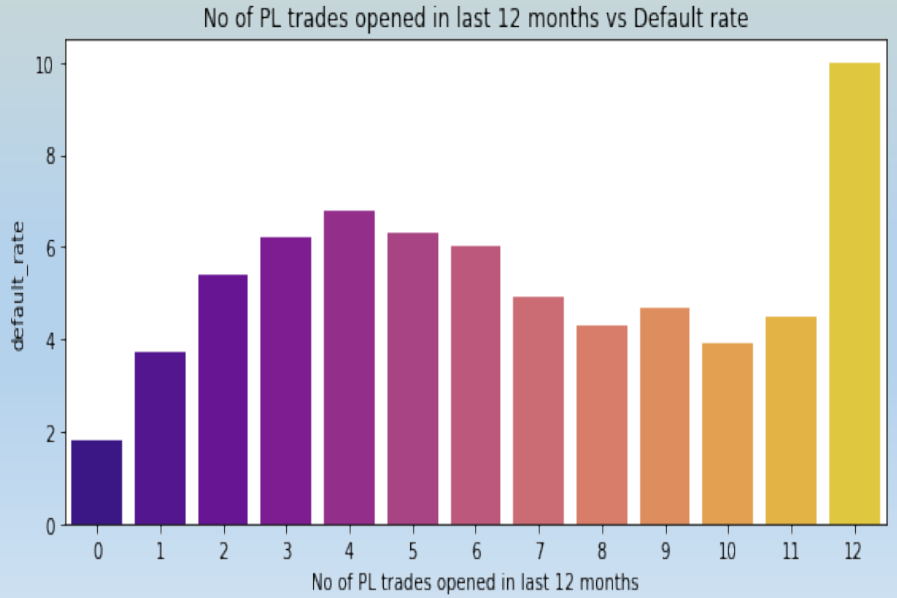
- We note few outliers in this column.
- Maximum credit card holders have not opened PL trade in the last 6 months.
- We note default rate increase as number of PL trades opened increases in the last 6 months. This trend can be seen till 2 opened PL trades in the last 6 months and then gradually decreases.

Credit Bureau Data Column: No. of PL trades opened in last 12 months



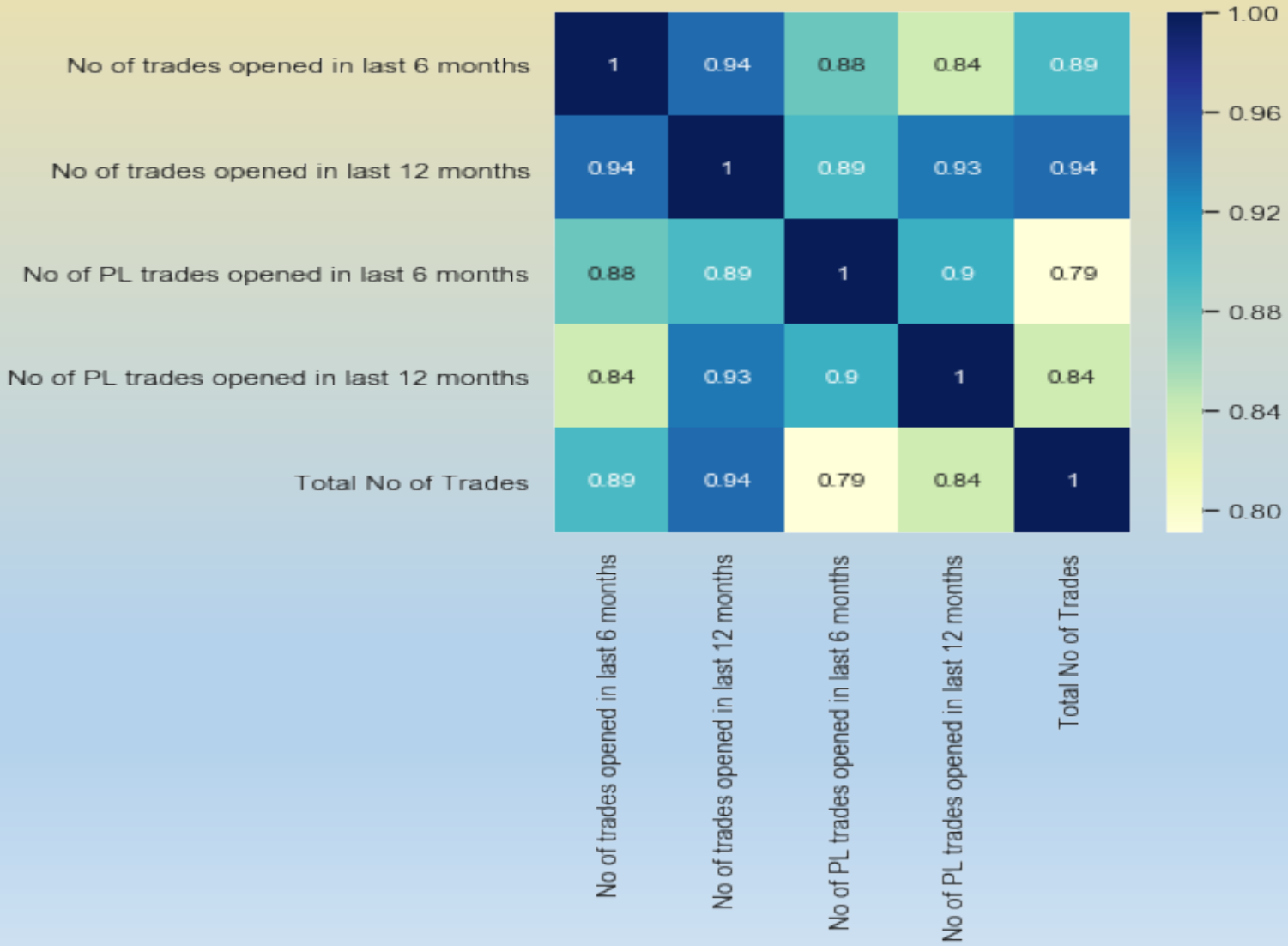






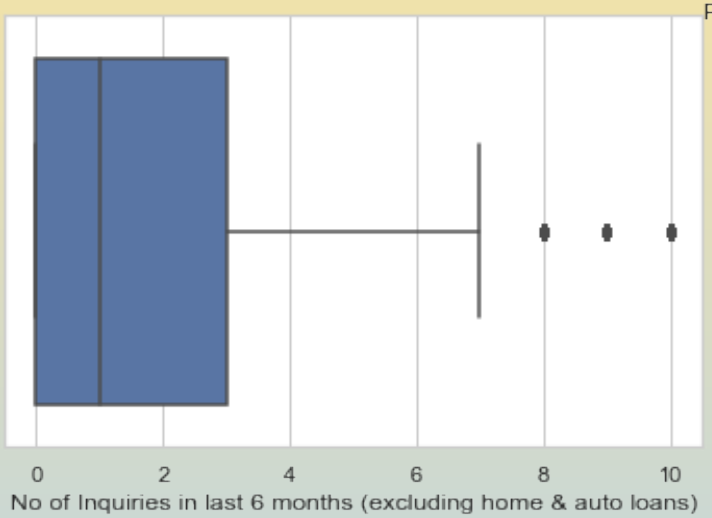
- We note few outliers in this column.
- Maximum credit card holders have not opened PL trade in the last 12 months.
- We note default rate increase as number of PL trades opened increases in the last 12 months. This trend can be seen till 4 opened PL trades in the last 12 months.

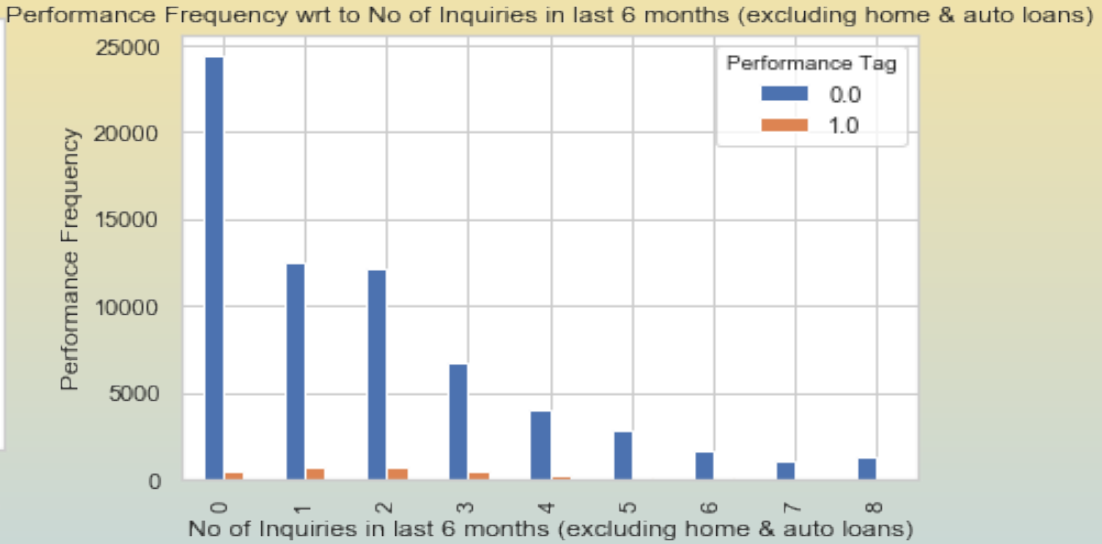
Correlation matrix of credit bureau trade columns

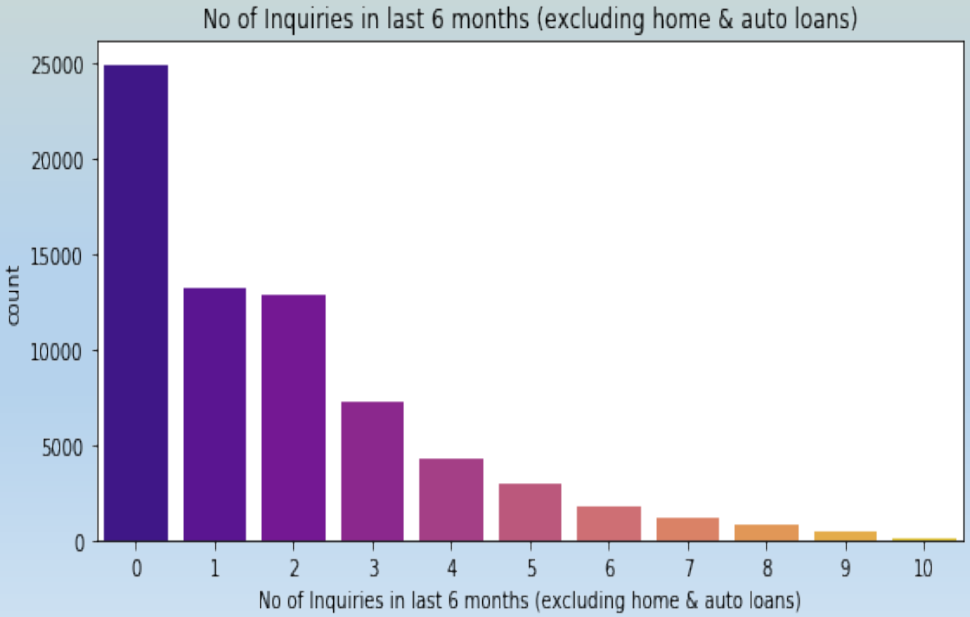


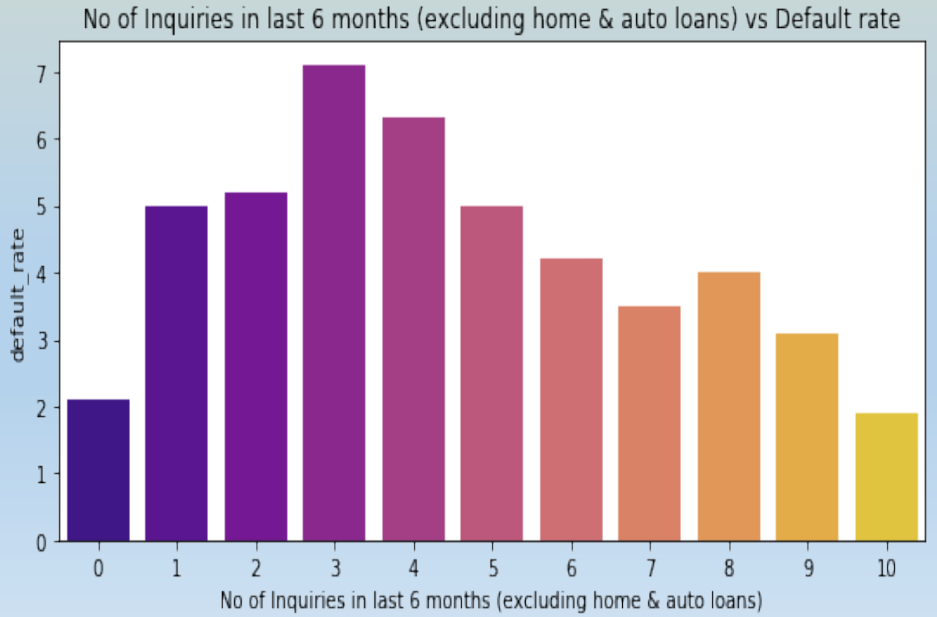
- As we note all the trade columns of credit bureau data are highly correlated to each other.
- And as a obvious behavior, the maximum correlation is among 'No of trades opened in last 6 months' with 'No of trades opened in last 12 months' and 'No of trades opened in last 12 months' with 'Total No of trades'. While model building either of column should suffice.

Credit Bureau Data Column: No. of inquiries in last 6 months (excluding home & auto loans)



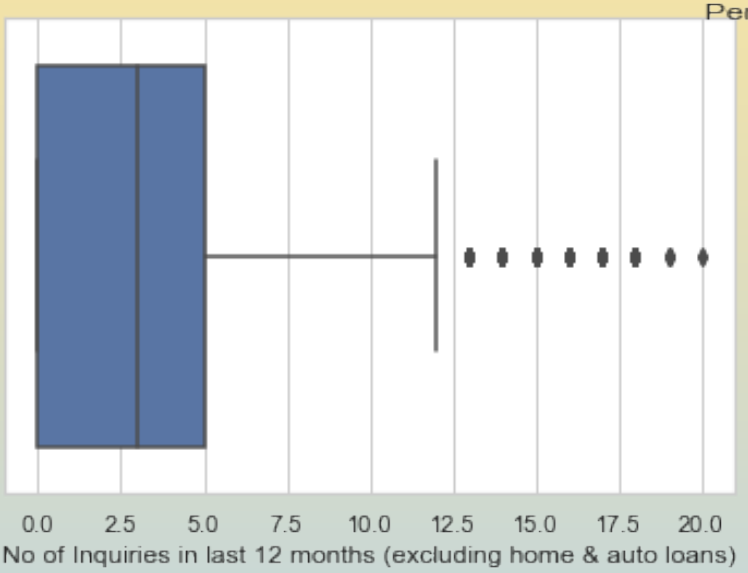


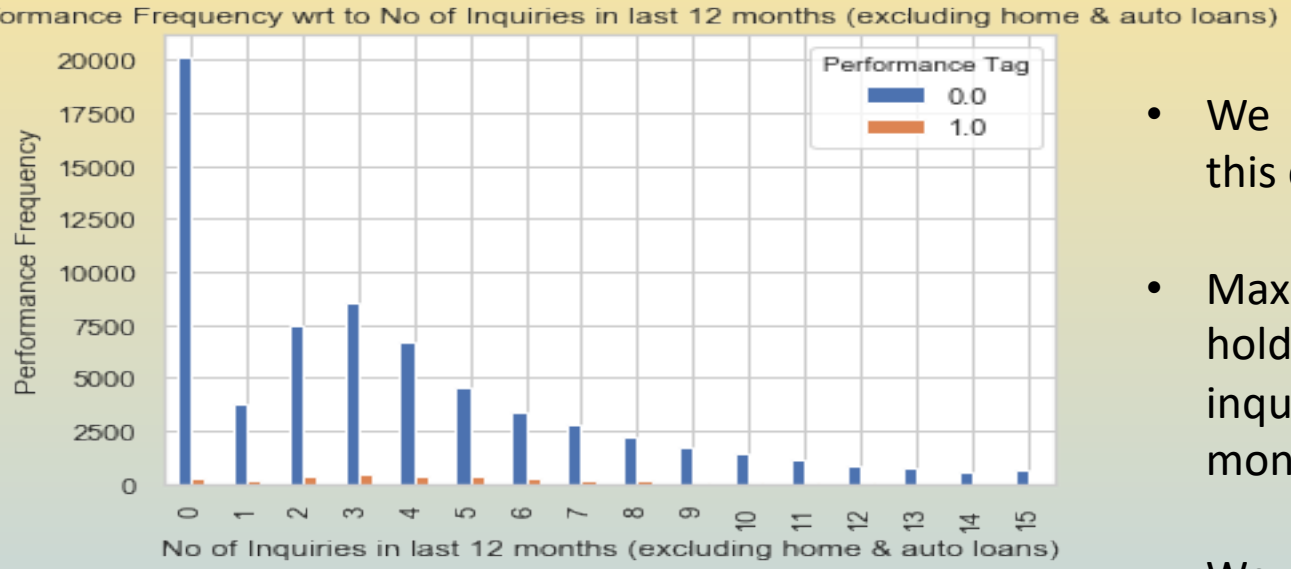


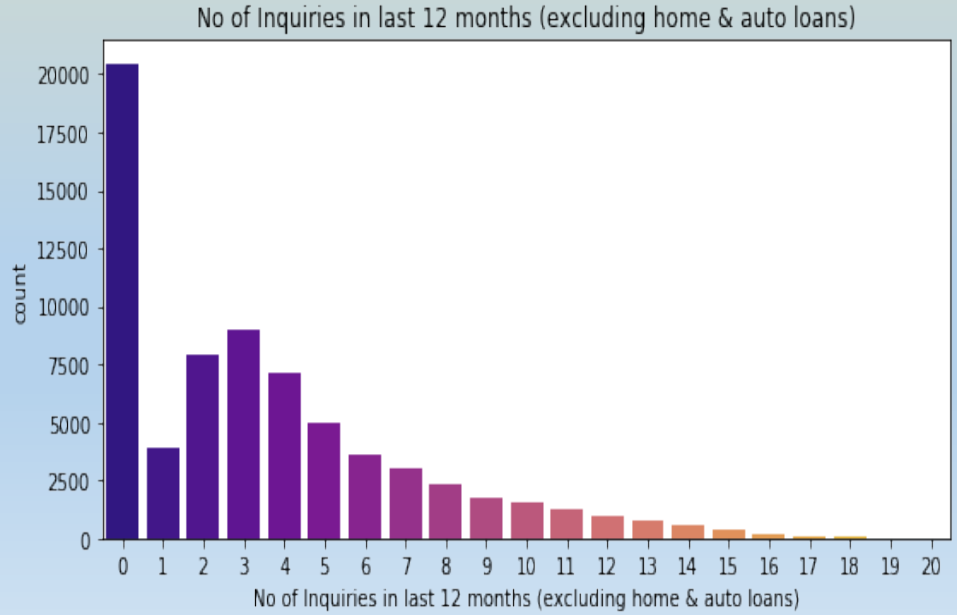


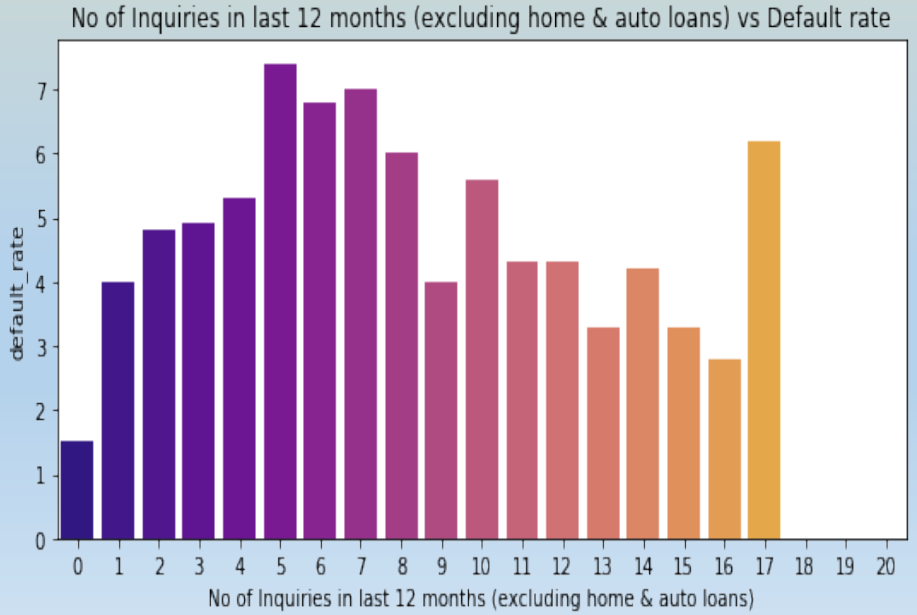
- We note few outliers in this column.
- Maximum credit card holders have not done inquiries in last 6 months.
- We note default rate increase as number of inquiries increases till 3 inquiries and gradually decreases.
- Also important to note there are few data entries beyond 1 to 2 inquiries done in last 6 months.

Credit Bureau Data Column: No. of inquiries in last 12 months (excluding home & auto loans)



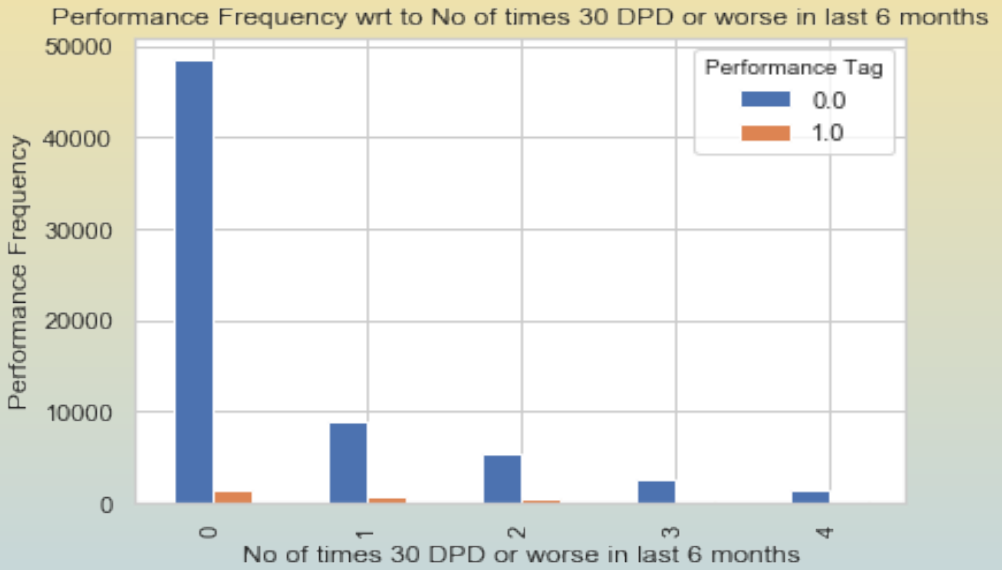


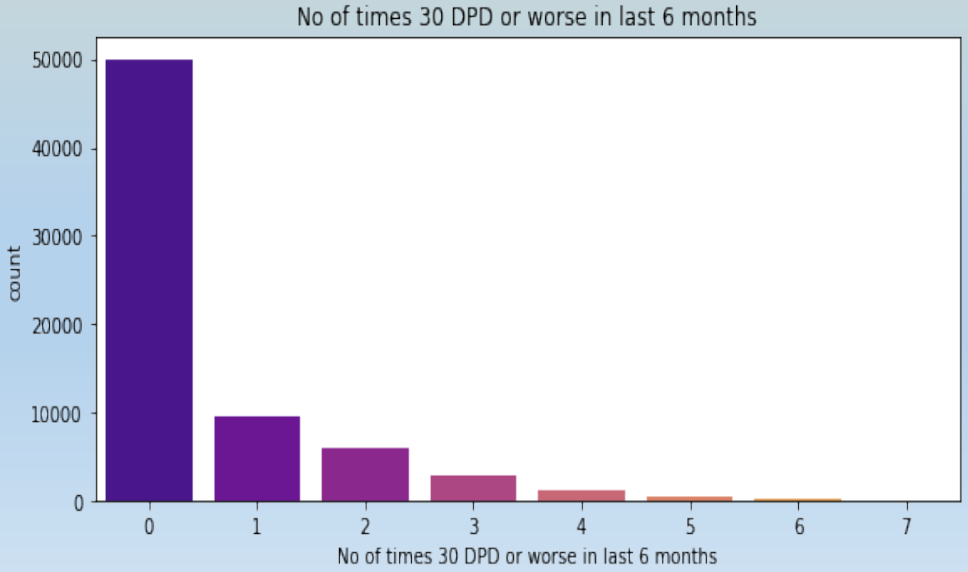


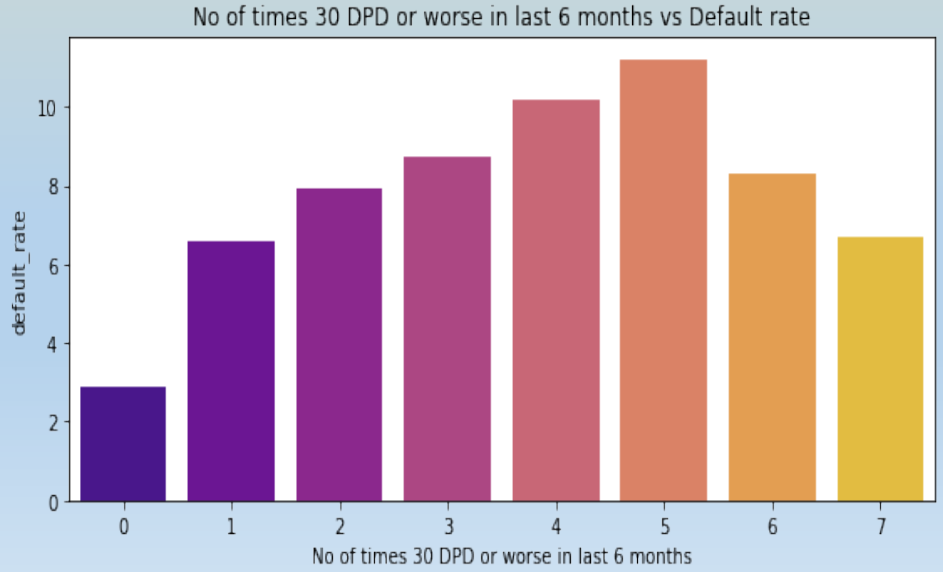


- We note few outliers in this column.
- Maximum credit card holders have not done inquiries in last 12 months.
- We note default rate increase as number of inquiries increases till 5 inquiries and gradually decreases.
- Also important to note there are few data entries beyond 4 to 5 inquiries done in last 12 months.

Credit Bureau Data Column: No. of times 30 DPD or worse in last 6 months

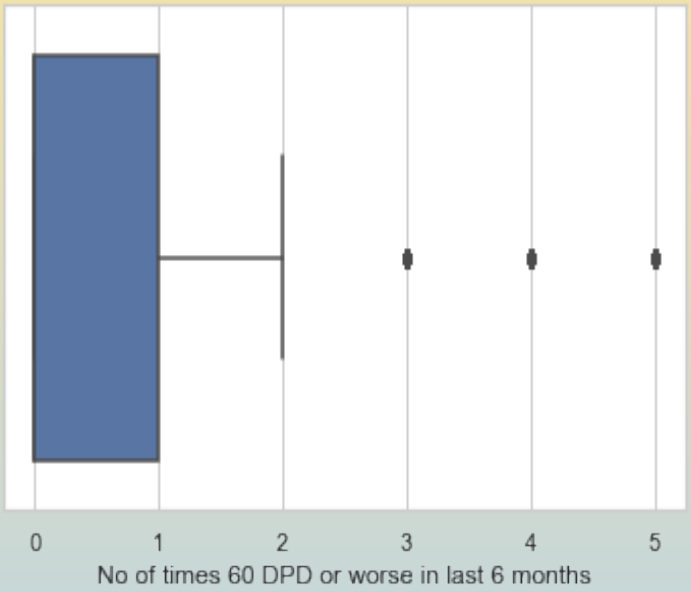


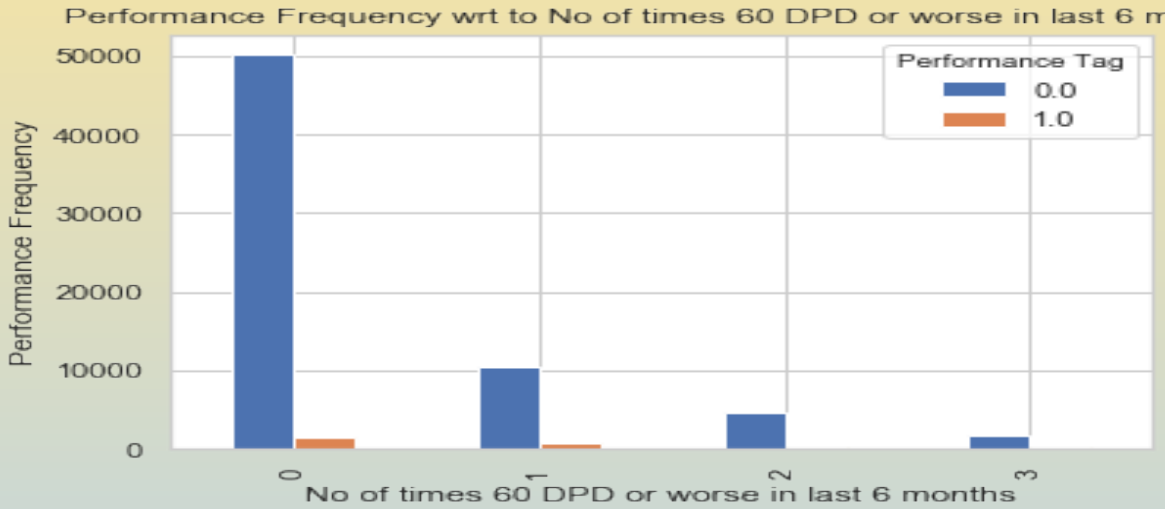




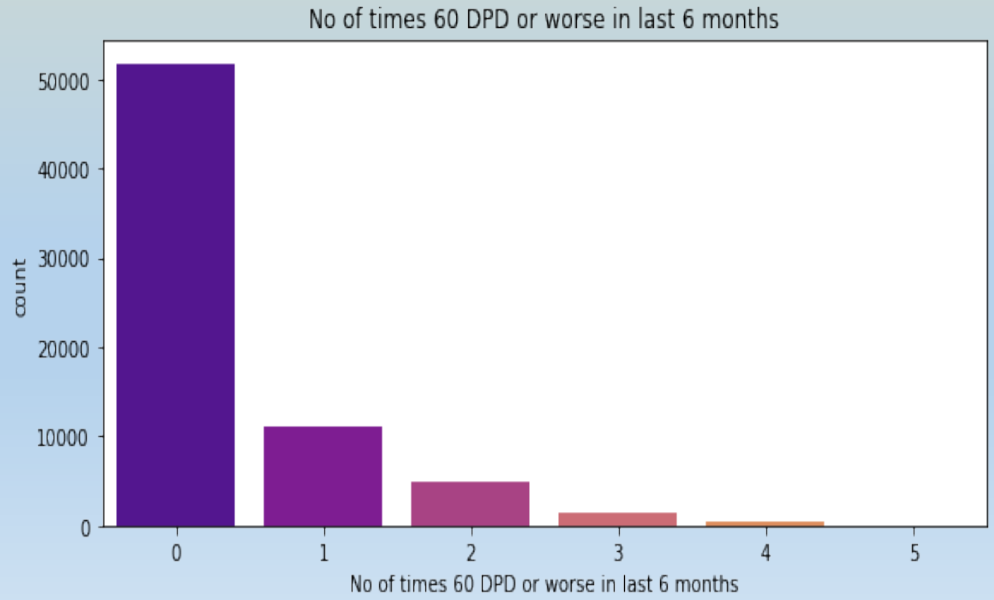
- We note few outliers in this column.
- Maximum credit card holders have paid their due well with 30 DPD in the last 6 months.
- The default rate increase as number of times customer has not paid dues since 30days in last 6 months increases.

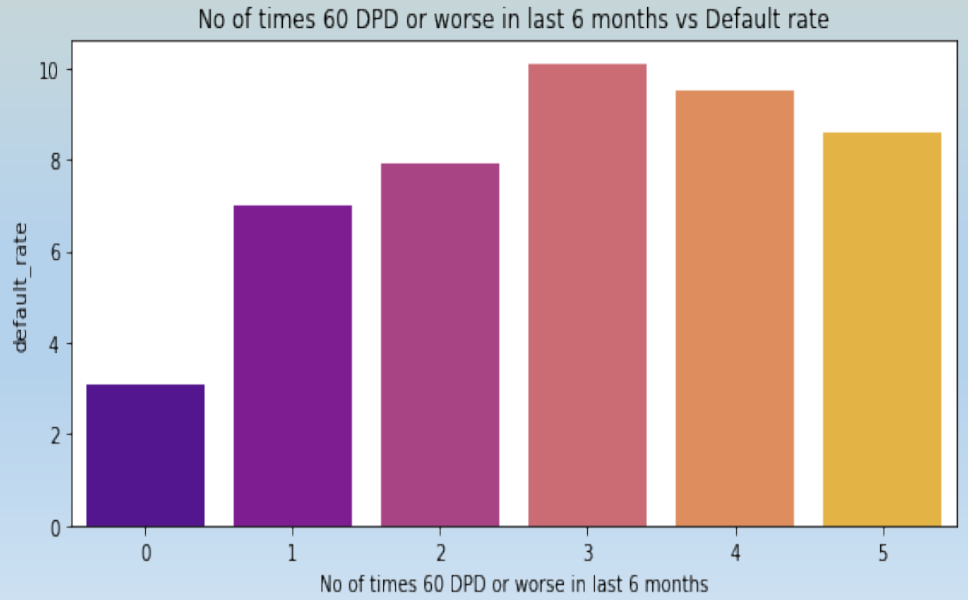
Credit Bureau Data Column: No. of times 60 DPD or worse in last 6 months





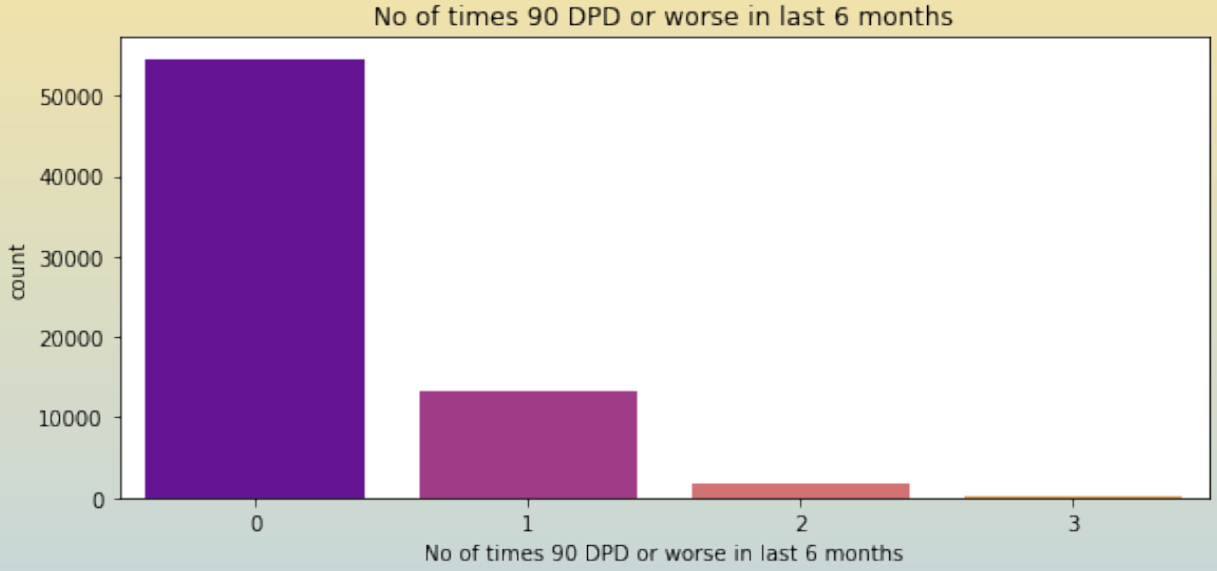
- We note few outliers in this column.
- Maximum credit card holders have paid their due well with 60 DPD in the last 6 months.

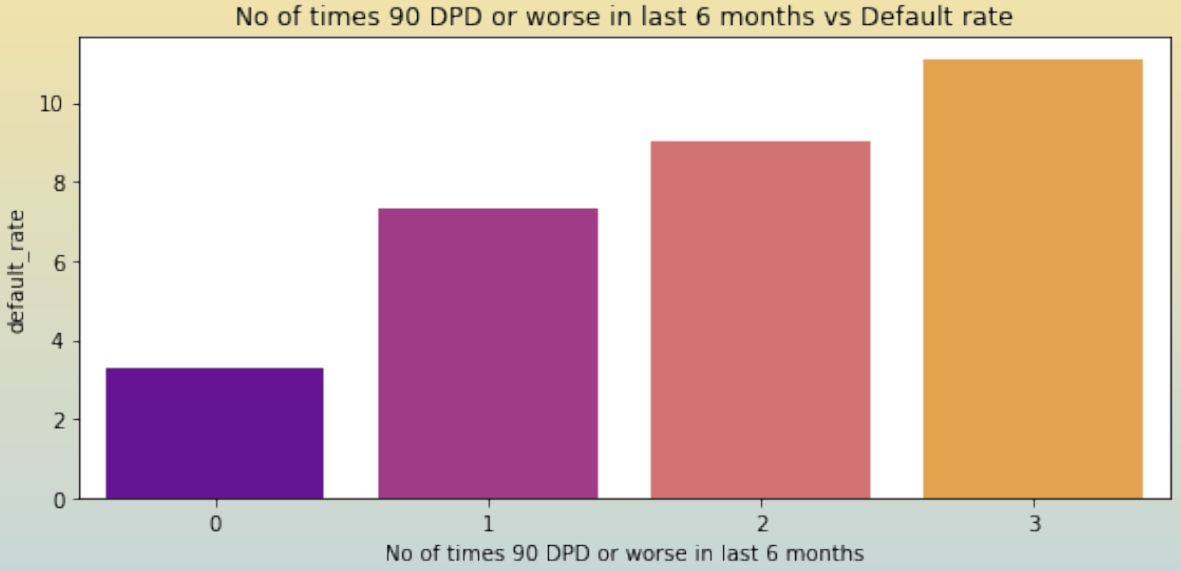


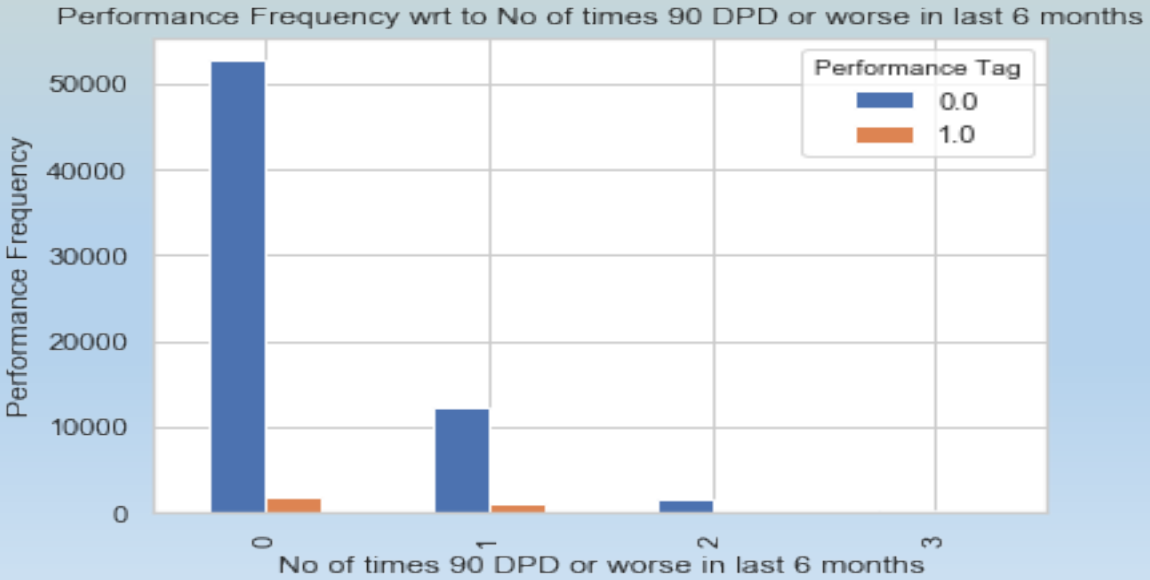


- The default rate increase as number of times customer has not paid dues since 60days in last 6 months increases.

Credit Bureau Data Column: No. of times 90 DPD or worse in last 6 months

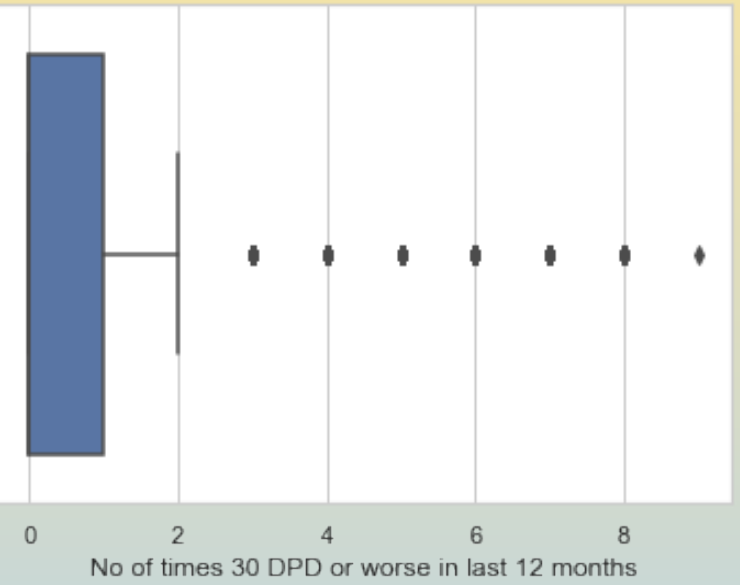


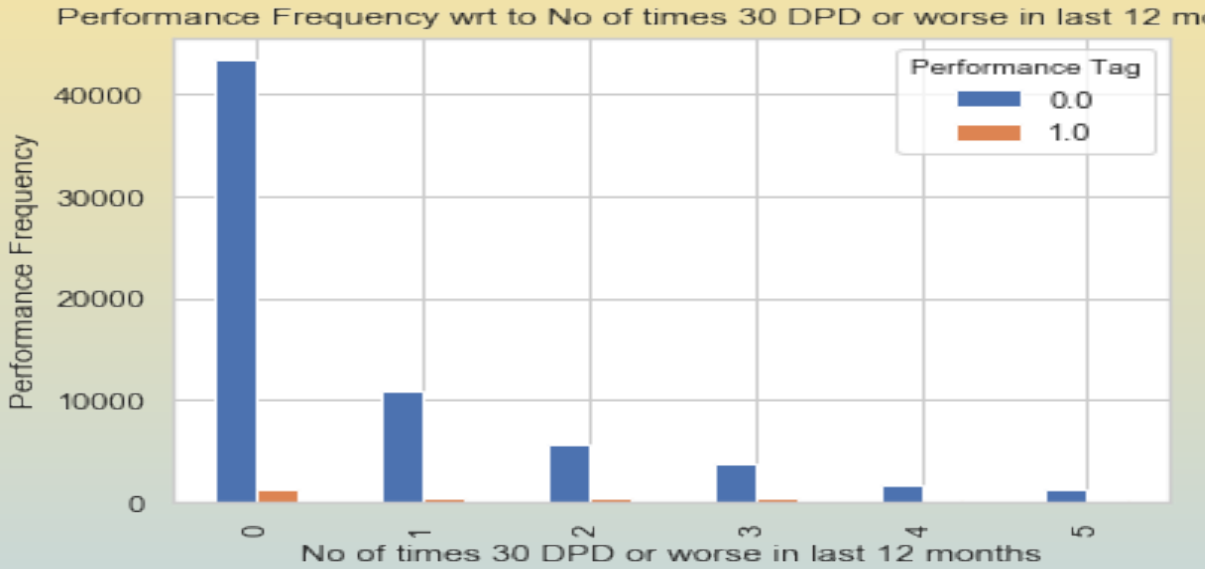




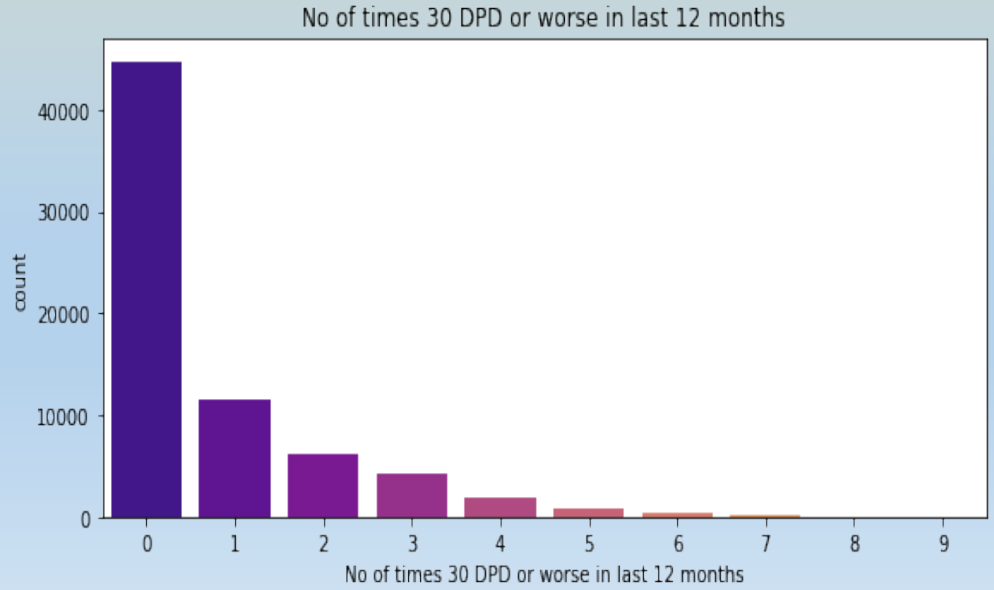
- Maximum credit card holders have paid their due well with 90 DPD in the last 6 months.
- The default rate increase as number of times customer has not paid dues since 90days in last 6 months increases.

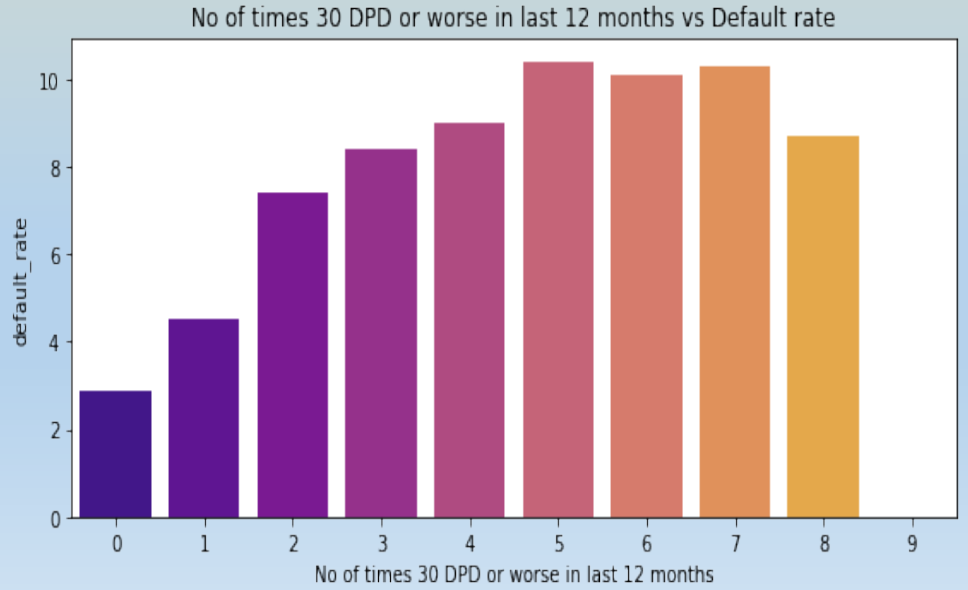
Credit Bureau Data Column: No. of times 30 DPD or worse in last 12 months



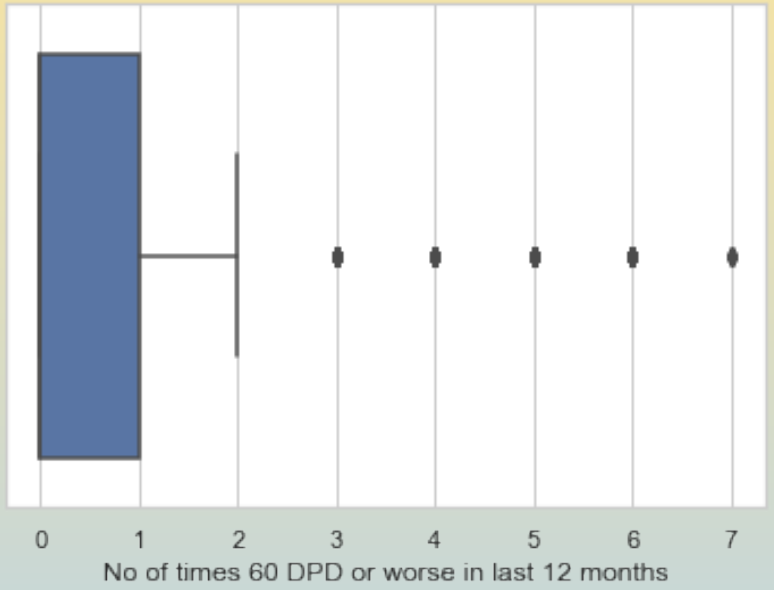


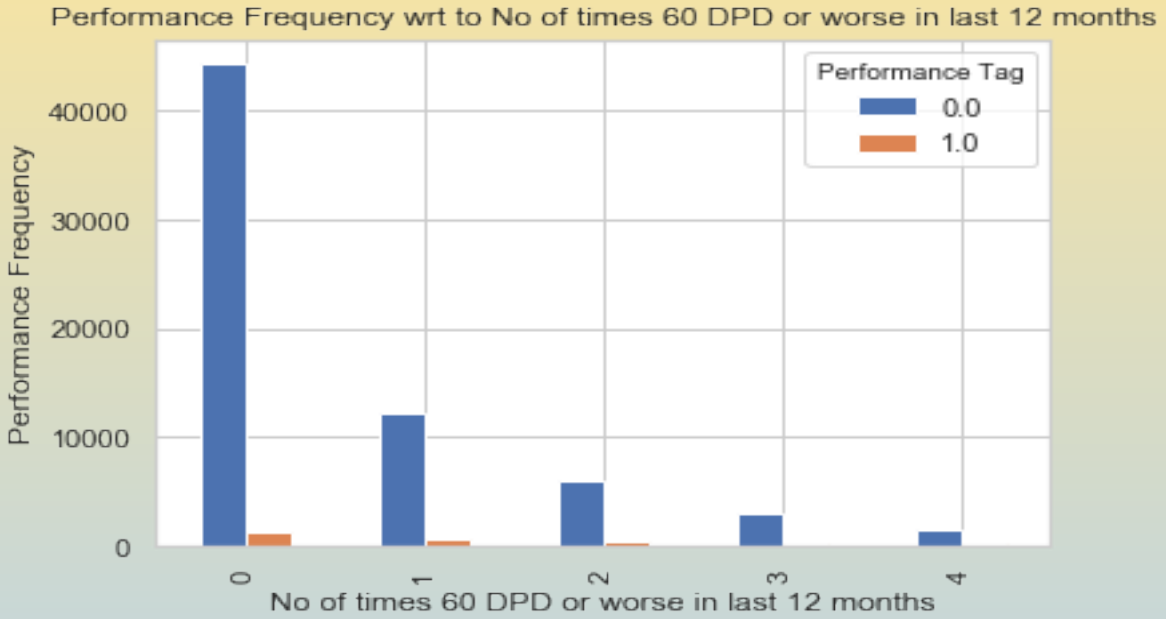
- We note few outliers in this column.
- Maximum credit card holders have paid their due well with 30 DPD in the last 12 months.
- The default rate increase as number of times customer has not paid dues since 30days in last 12 months increases.

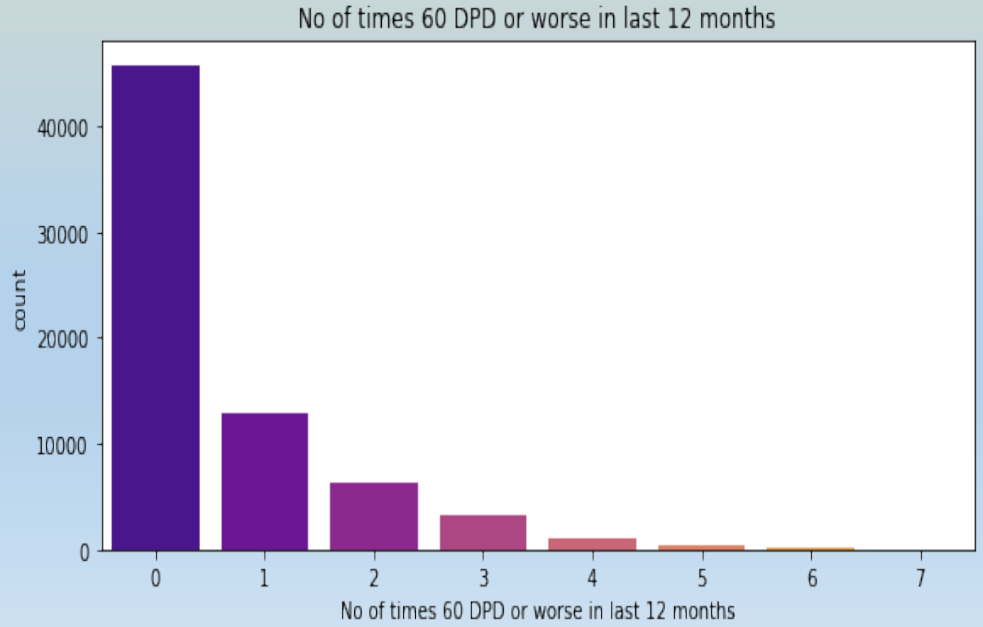


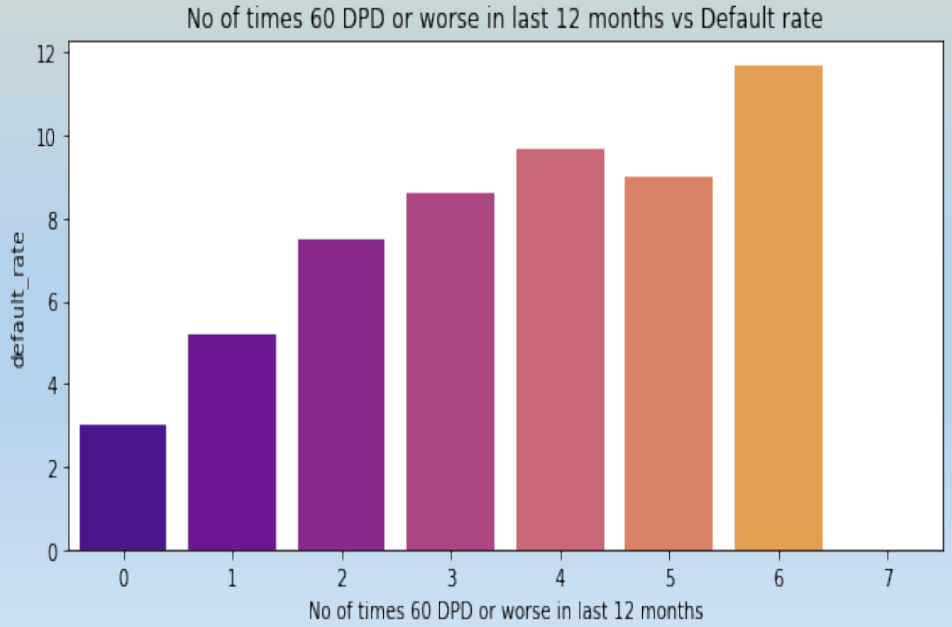


Credit Bureau Data Column: No. of times 60 DPD or worse in last 12 months



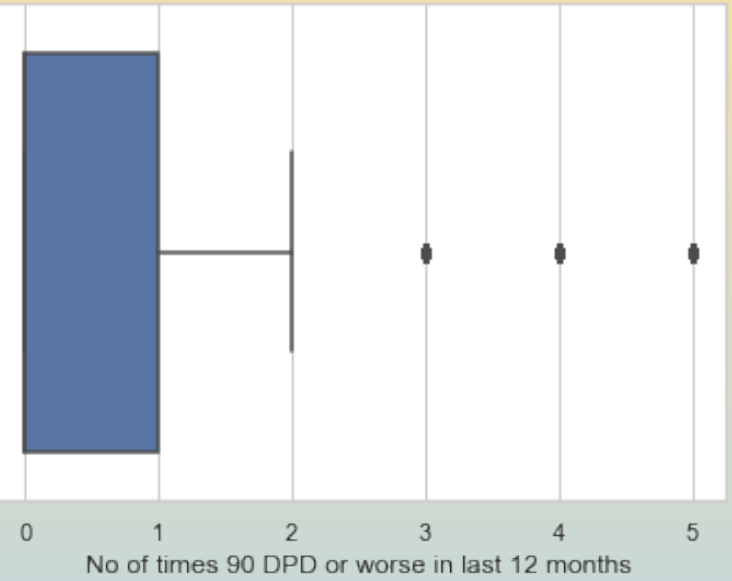


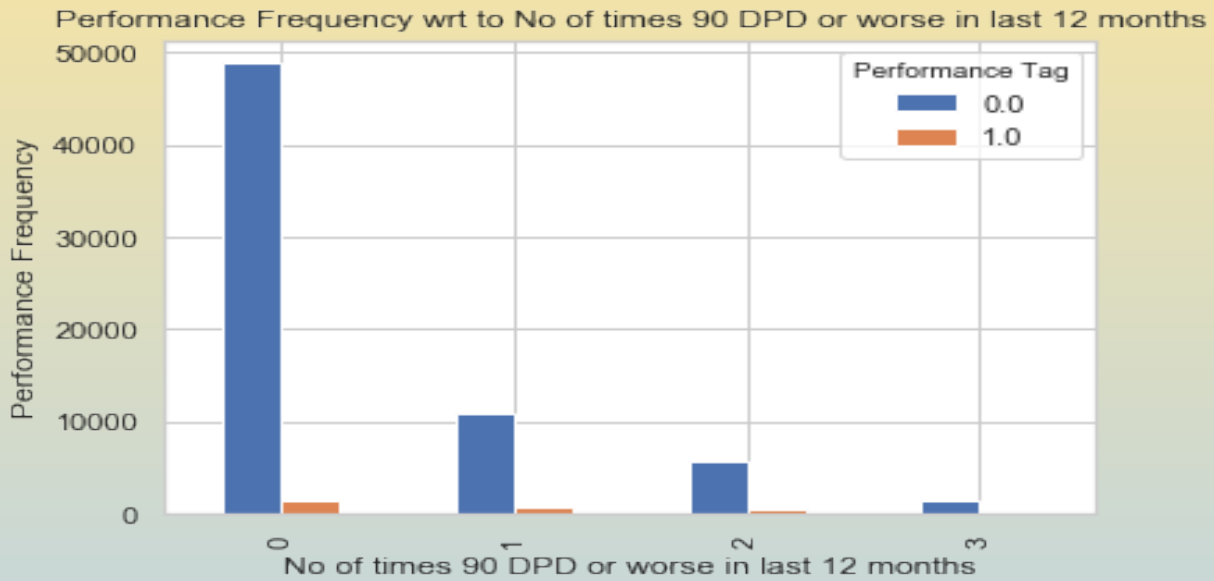


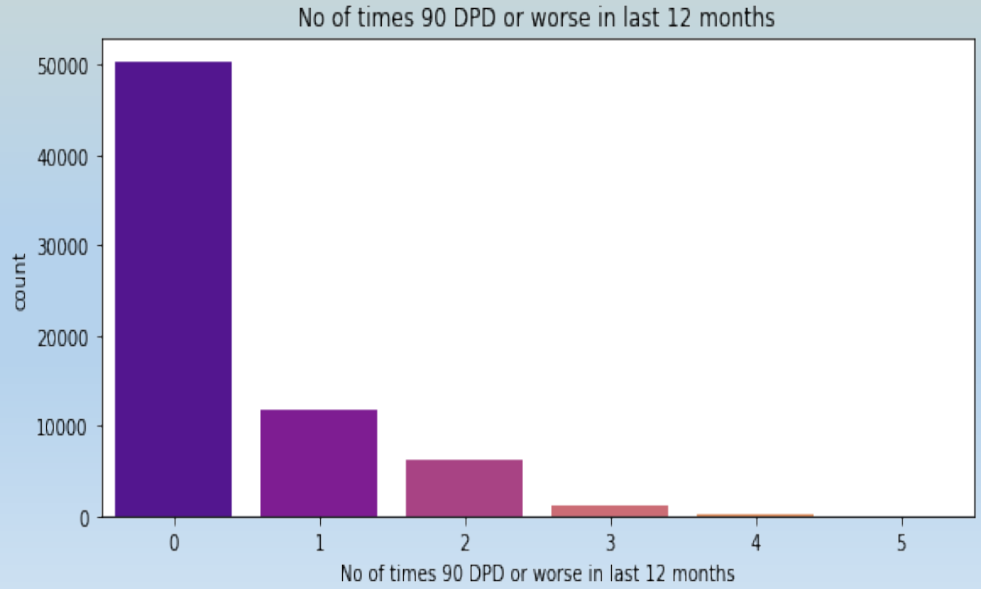


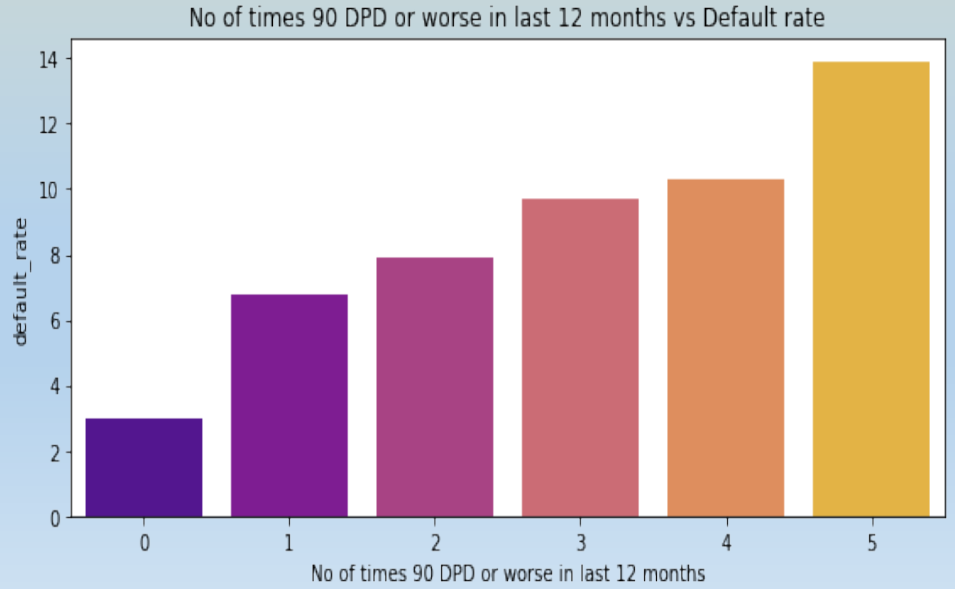
- We note few outliers in this column.
- Maximum credit card holders have paid their due well with 60 DPD in the last 12 months.
- The default rate increase as number of times customer has not paid dues since 60days in last 12 months increases.

Credit Bureau Data Column: No. of times 90 DPD or worse in last 12 months



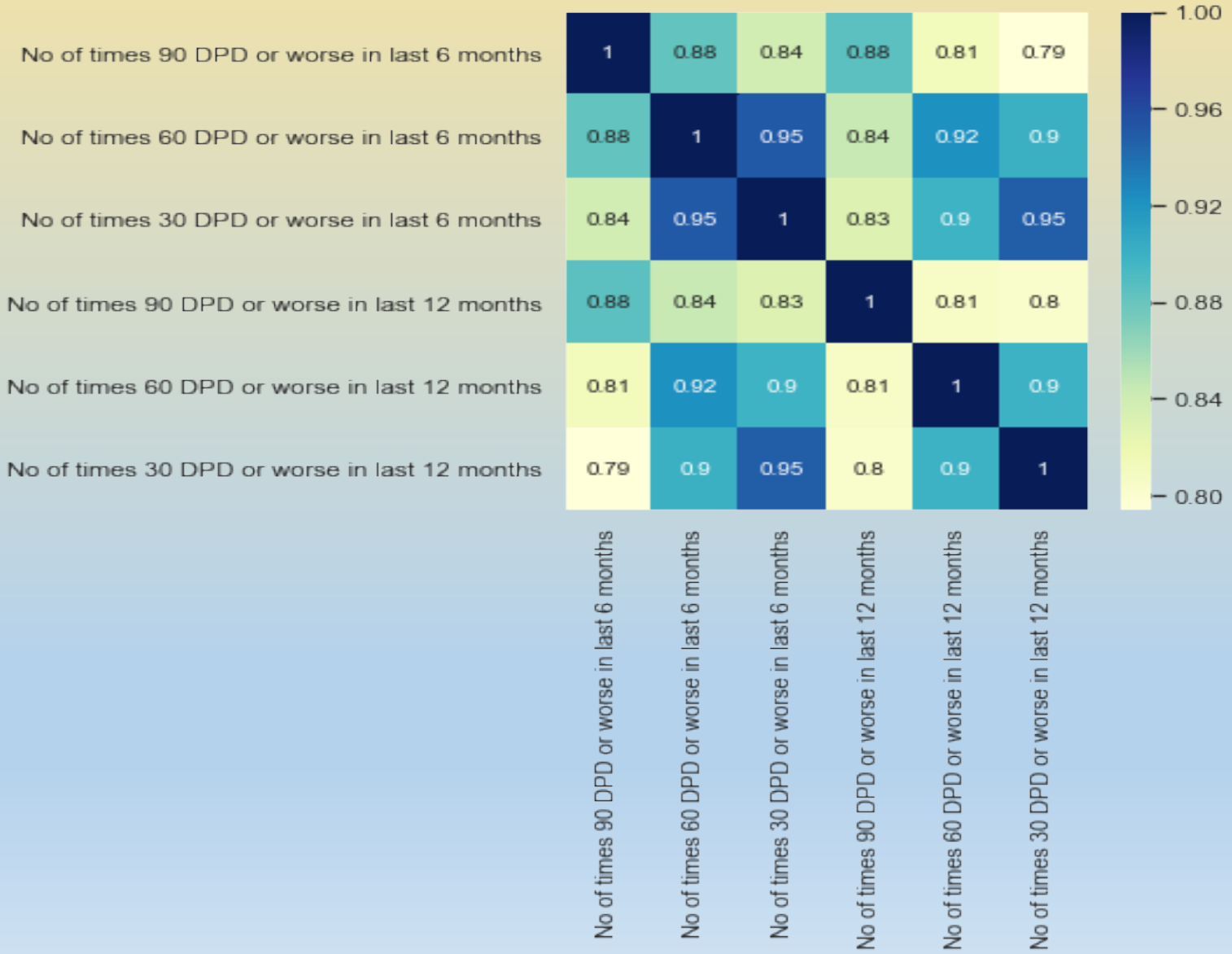






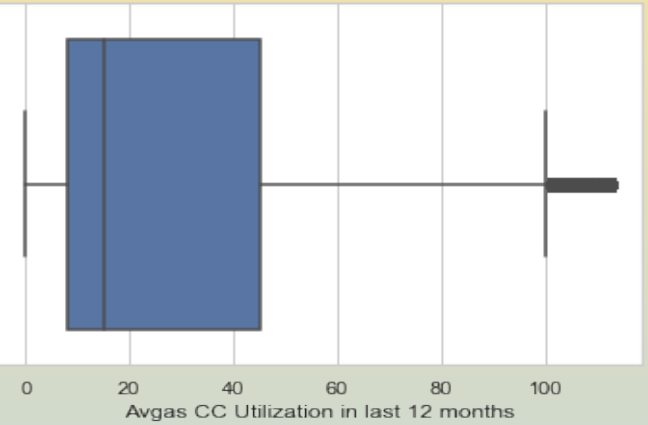
- We note few outliers in this column.
- Maximum credit card holders have paid their due well with 90 DPD in the last 12 months.
- The default rate increase as number of times customer has not paid dues since 90days in last 12 months increases.

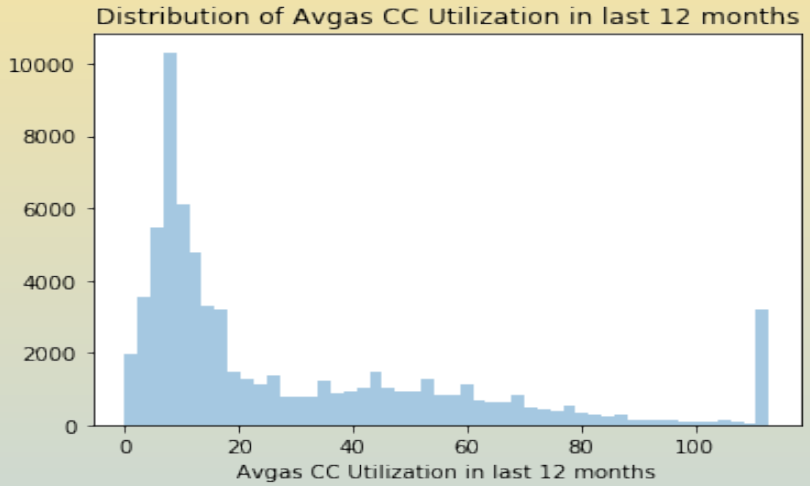
Correlation matrix of credit bureau DPD columns

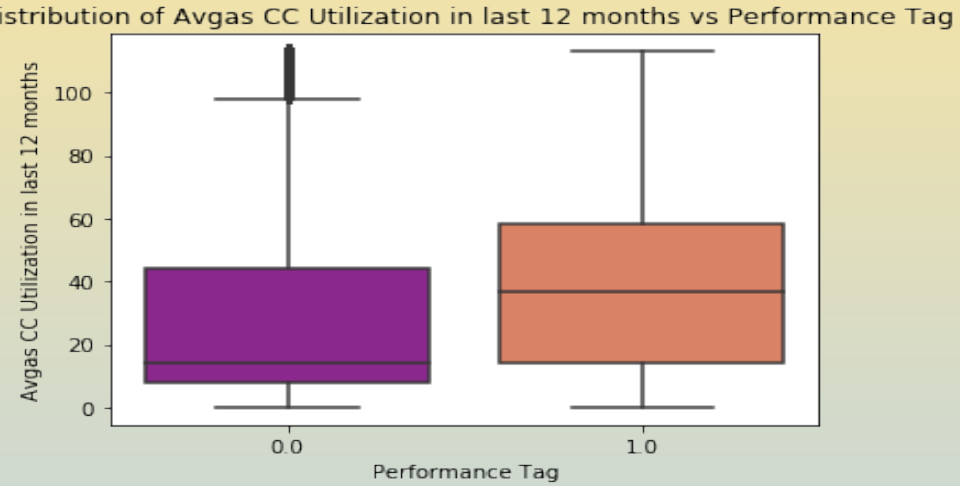


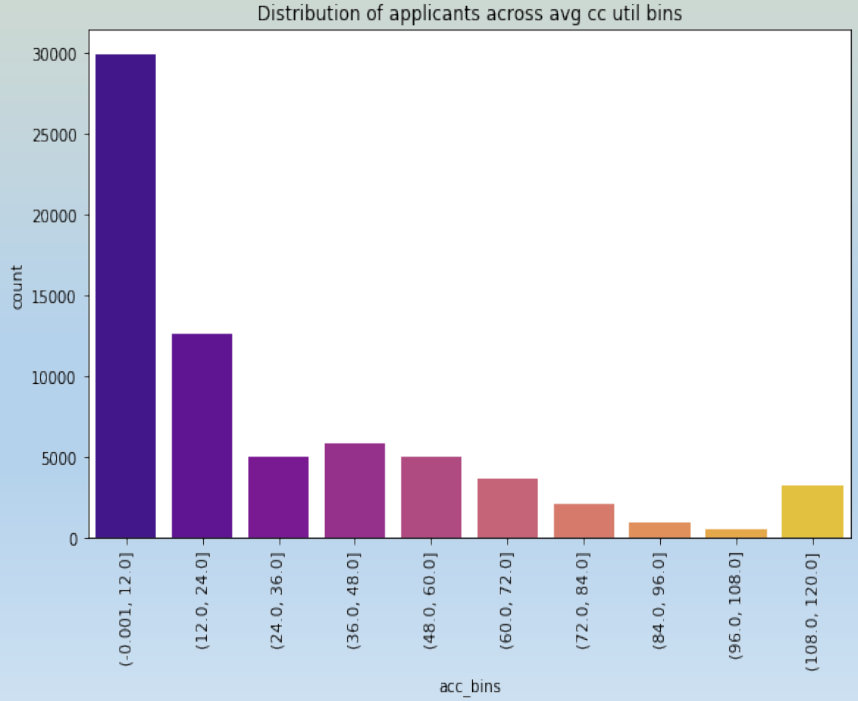
- As we note all the DPD columns of credit bureau data are highly correlated to each other.
- And as a obvious behavior- we see maximum correlation among:
  - No of times 30 DPD or worse in last 6 months with No of times 60 DPD or worse in last 6 months
  - No of times 30 DPD or worse in last 6 months with No of times 30 DPD or worse in last 12 months

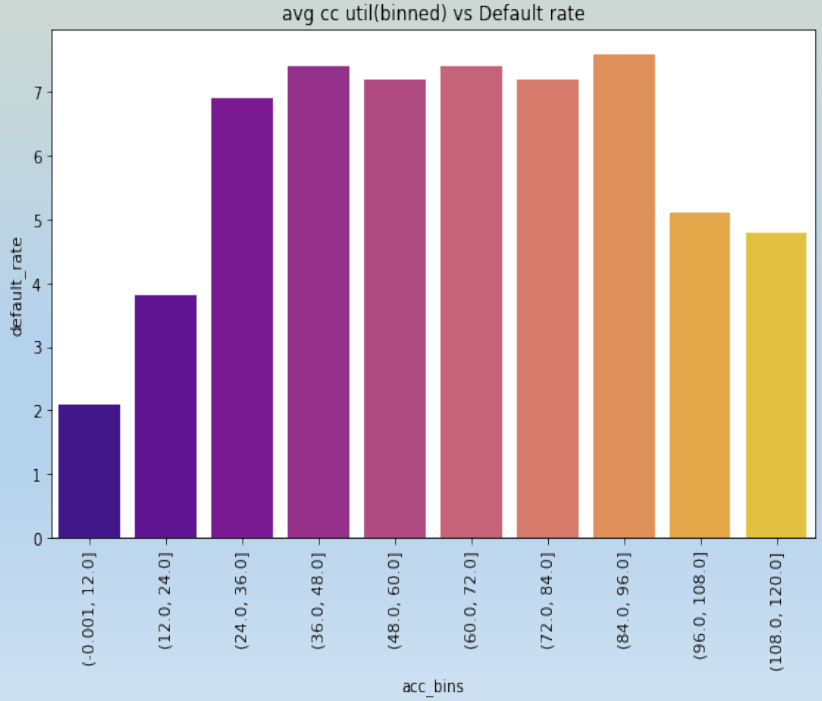
Credit Bureau Data Column: Avgas CC Utilization in last 12 months





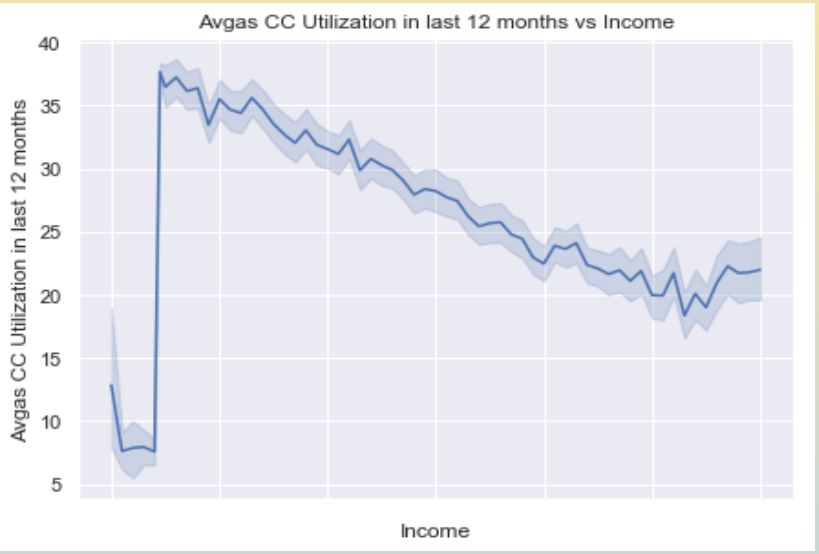


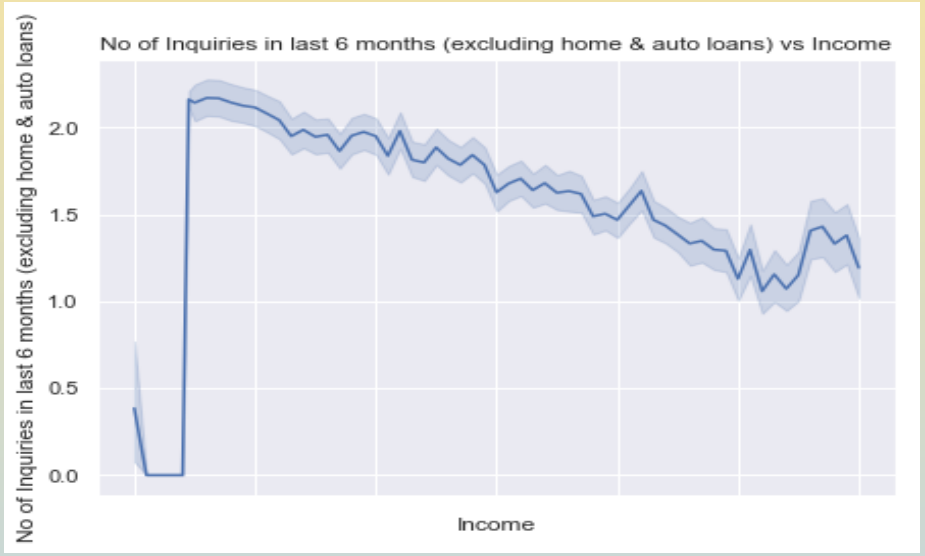


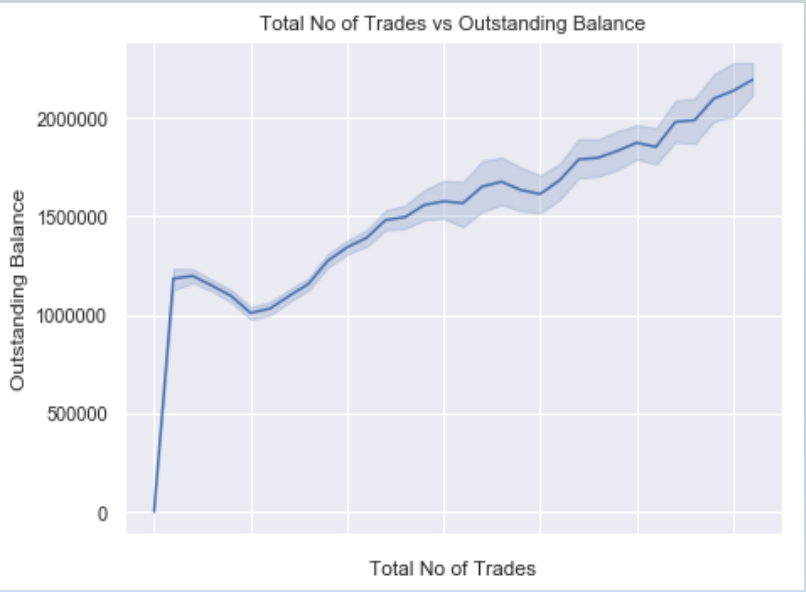


- We note few outliers in this column.
- Most of the credit card users have their avgas cc between 0 and 15.
- The median of avg. utilization for defaults is significantly higher than the non-defaulters.
- There is a positive correlation between default rate and average credit card utilization.

Few other observations







- As income increases, average utilization of credit card by customer decreases.
- As income increases, number of times the customer inquired decreases.
- Higher the outstanding balance, higher is the total number of trades.
- As income increases, number of times customer has not paid dues since 30 days in last 6 months decreases.

Important variables identified as per EDA

- **Demographic data columns:**

  - ✓ Income- Lesser the income, higher is the defaulted rate.
  - ✓ No. of months in current company – we can see a general trend of decrease in default rate when there is a increase in the number of month a person works in his/her present company.
  - ✓ No. of months in current residence - we can see a general trend of decrease in default rate with increase in number of months in current residence.

- **Credit bureau data columns:**

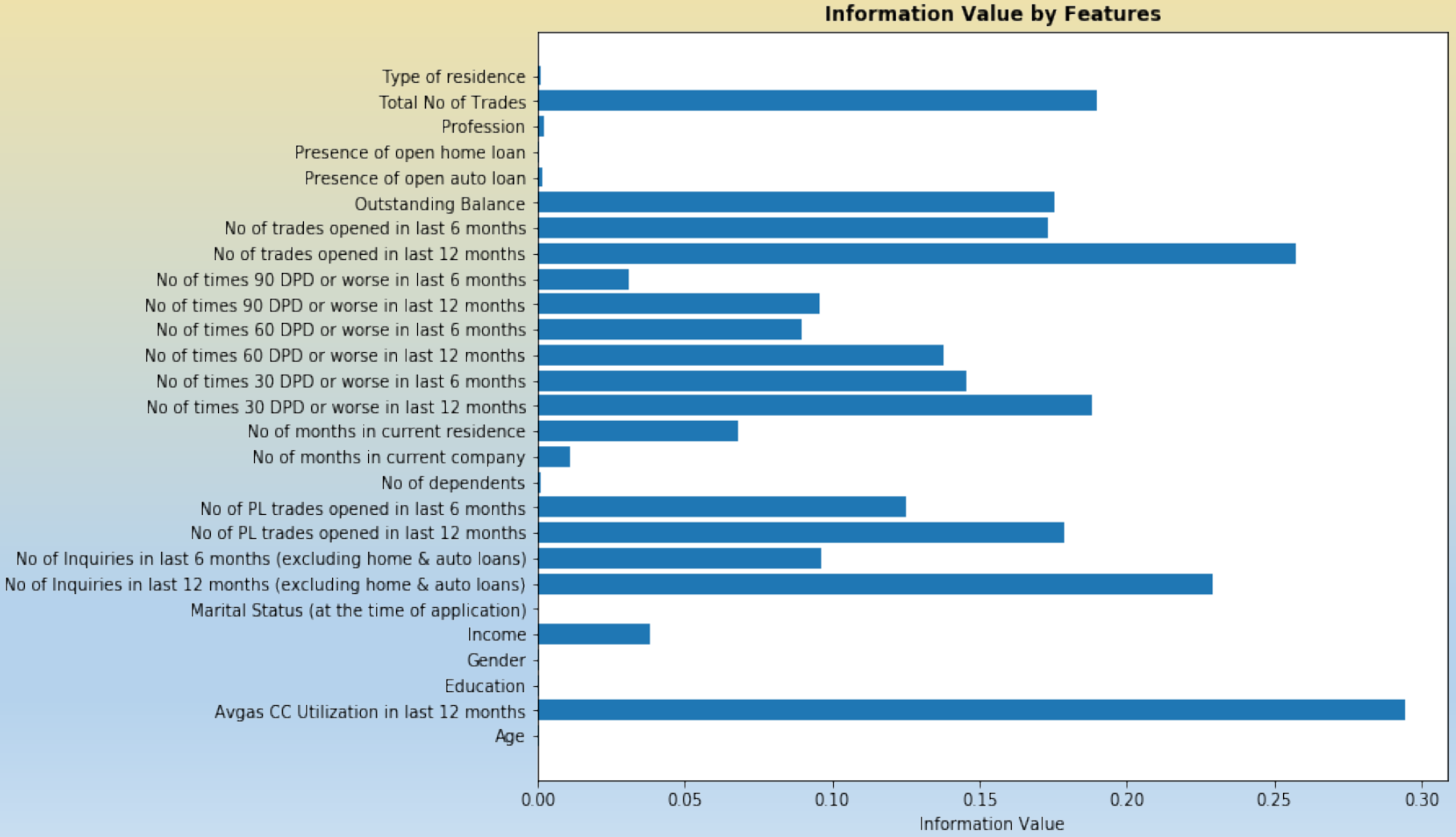
  - ✓ Avgas CC Utilization in last 12 months -There is a positive correlation between default rate and average credit card utilization, except when Avg CC is very high.
  - ✓ All DPD columns – As number of DPD increases, default rate increases. And all the DPD columns are highly correlated to each other so one of the columns in the model building should suffice.
  - ✓ All trades columns – As number of trading done increases, default rate increases. And all the trade columns are highly correlated to each other so one of the columns in the model building should suffice.

WOE and IV Analysis

	VAR_NAME	IV	Variable_Predictiveness
1	Avgas CC Utilization in last 12 months	0.294117	Medium predictive Power
19	No of trades opened in last 12 months	0.257558	Medium predictive Power
6	No of Inquiries in last 12 months (excluding h...	0.229287	Medium predictive Power
25	Total No of Trades	0.190025	Medium predictive Power
13	No of times 30 DPD or worse in last 12 months	0.188084	Medium predictive Power
8	No of PL trades opened in last 12 months	0.178891	Medium predictive Power
21	Outstanding Balance	0.175553	Medium predictive Power
20	No of trades opened in last 6 months	0.173375	Medium predictive Power
14	No of times 30 DPD or worse in last 6 months	0.145751	Medium predictive Power
15	No of times 60 DPD or worse in last 12 months	0.137659	Medium predictive Power
9	No of PL trades opened in last 6 months	0.125287	Medium predictive Power
7	No of Inquiries in last 6 months (excluding ho...	0.096079	Weak predictive Power
17	No of times 90 DPD or worse in last 12 months	0.095775	Weak predictive Power
16	No of times 60 DPD or worse in last 6 months	0.089620	Weak predictive Power
12	No of months in current residence	0.067782	Weak predictive Power
4	Income	0.038135	Weak predictive Power
18	No of times 90 DPD or worse in last 6 months	0.030744	Weak predictive Power
11	No of months in current company	0.010948	Not useful for prediction
24	Profession	0.002170	Not useful for prediction
22	Presence of open auto loan	0.001619	Not useful for prediction
26	Type of residence	0.000955	Not useful for prediction
10	No of dependents	0.000933	Not useful for prediction
2	Education	0.000774	Not useful for prediction
0	Age	0.000682	Not useful for prediction
23	Presence of open home loan	0.000465	Not useful for prediction
3	Gender	0.000348	Not useful for prediction
5	Marital Status (at the time of application)	0.000093	Not useful for prediction

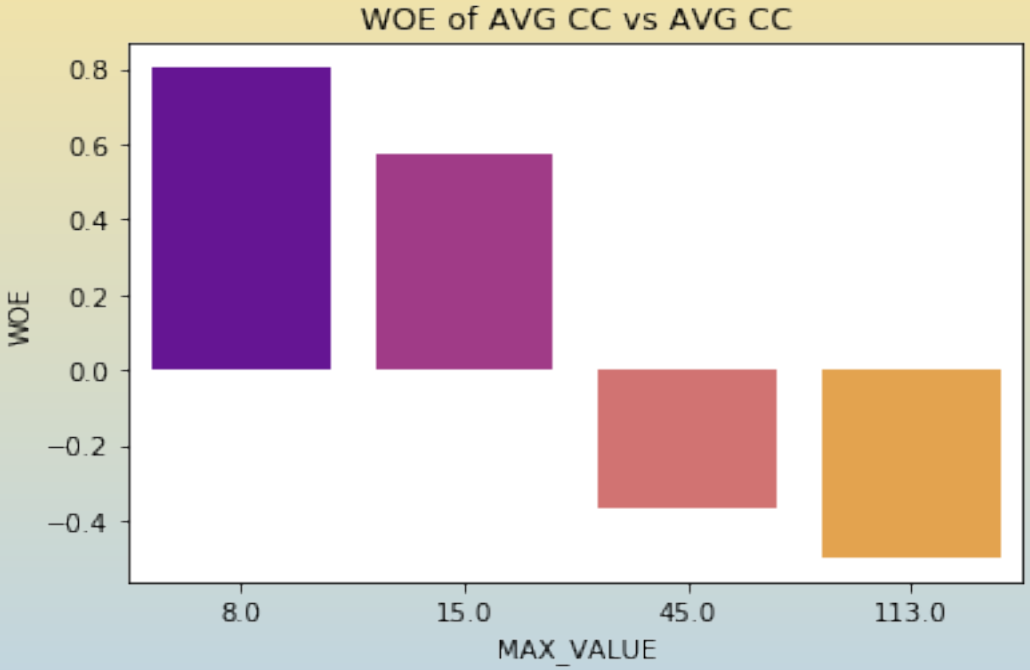
- In our EDA, we saw that there are outliers in few variables. Also few variables have null values. As WOE transformation has advantage of handling missing values and outliers – we replaced the null values by their corresponding WOE values. Outliers will be handled by the respective woe value.
- WOE and IV values are calculated for each of the attributes using information and woe binning.
- From the IV values we can conclude that parameters in the demographic data don't play much significant role in prediction and most of the significant variables are from Credit Bureau data.
- Top 11 Variables with IV value of 0.1 to 0.3 has medium predictive power and are considered significant. There is no variable with strong predictive power.

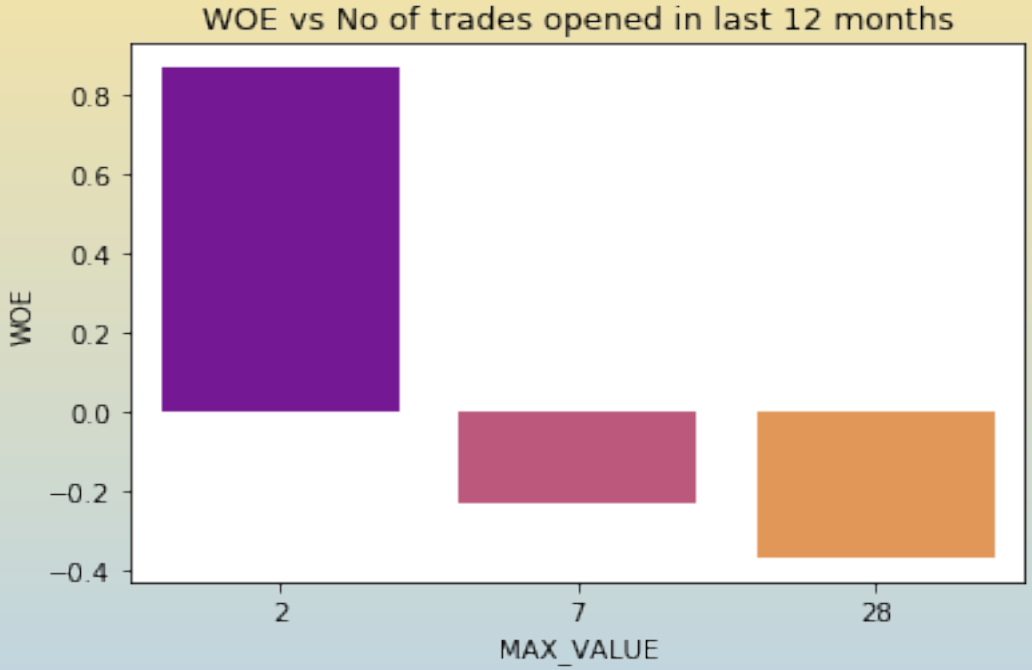
Information Value Plot

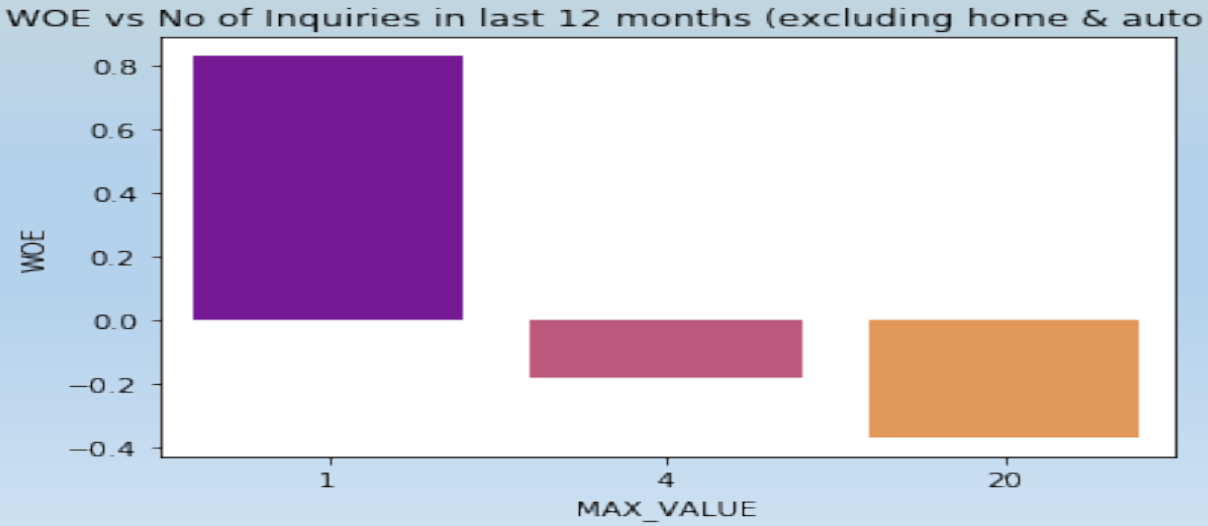


Average credit card utilization column, Trades columns and No. of inquires columns are the top variables that have high iv values.

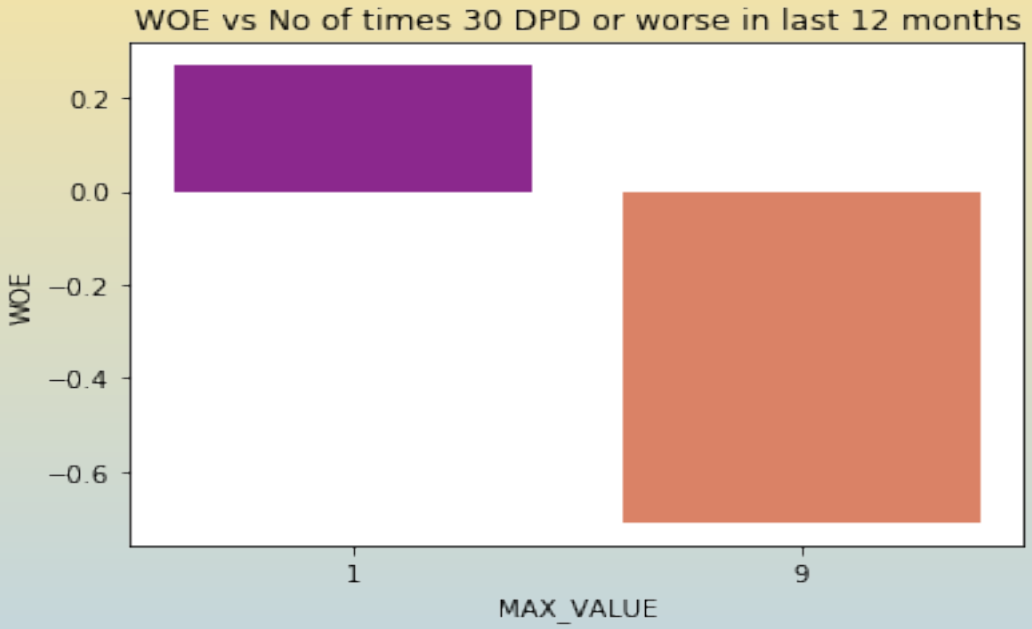
WOE plot

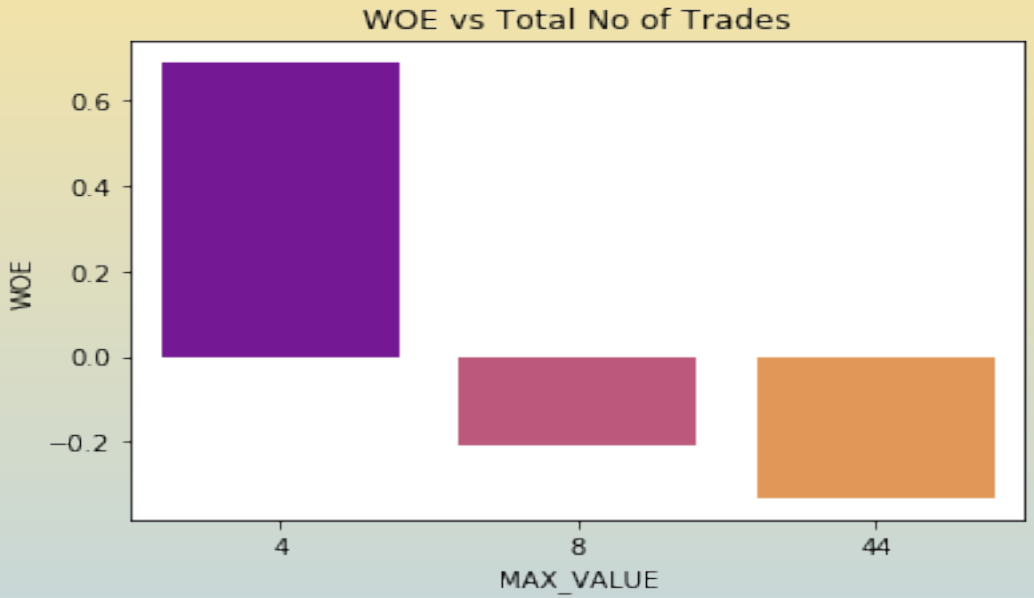

WOE of AVG CC vs AVG CC
WOE
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
8.0
15.0
45.0
113.0
MAX_VALUE

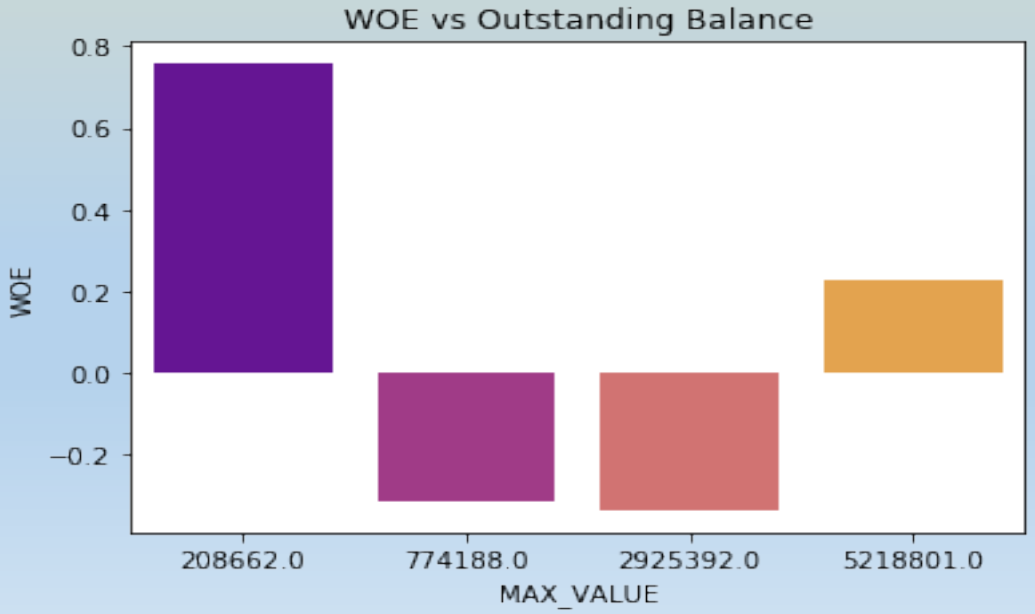

WOE vs No of trades opened in last 12 months
WOE
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
2
7
28
MAX_VALUE


WOE vs No of Inquiries in last 12 months (excluding home & auto loans)
WOE
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
1
4
20
MAX_VALUE

WOE plot







We see that almost all the variables have monotonic woe, except for outstanding balance

Model Building and Evaluation

General steps:

Outlier Treatment: Outlier detection is done using boxplot on continuous variables and quantiles function. The presence of outliers is treated by imputing the entire dataframe by WOE values.

Data Scaling: Though the variables are in different scale range initially, but since we are imputing the dataframe with woe values and woe values are log values of odds, all the features have woe values in the same scale (majorly between -1 and 1), the scaling is not required.

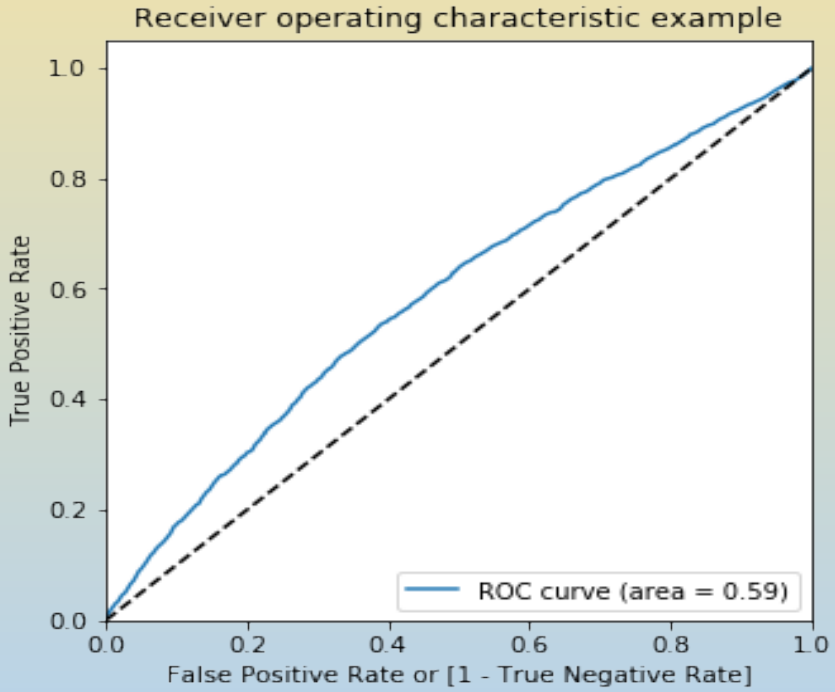
Data Split: The final dataset is split into Train and Test in 70:30 ratio for model building.

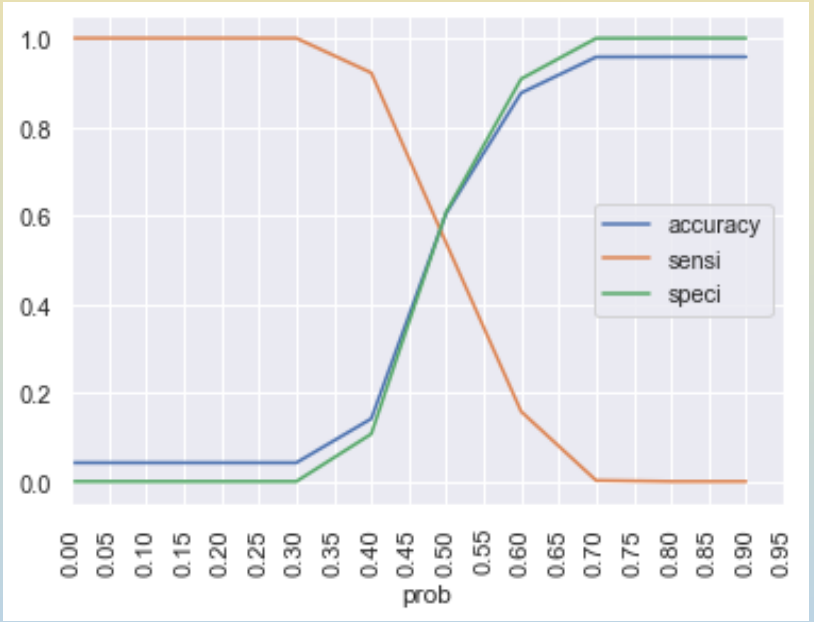
- All models are trained on training datasets and regularization was done by tuning of hyper parameters with cross validation on validation datasets.
- All the models are tested on test datasets that were kept separate from training and validation datasets.

Imbalance data: The given data is highly imbalanced. We have used class_weight = 'balanced' for balancing the training data sets in all our model.

Model Building and Evaluation

- **Demographic data model: Logistic Regression**

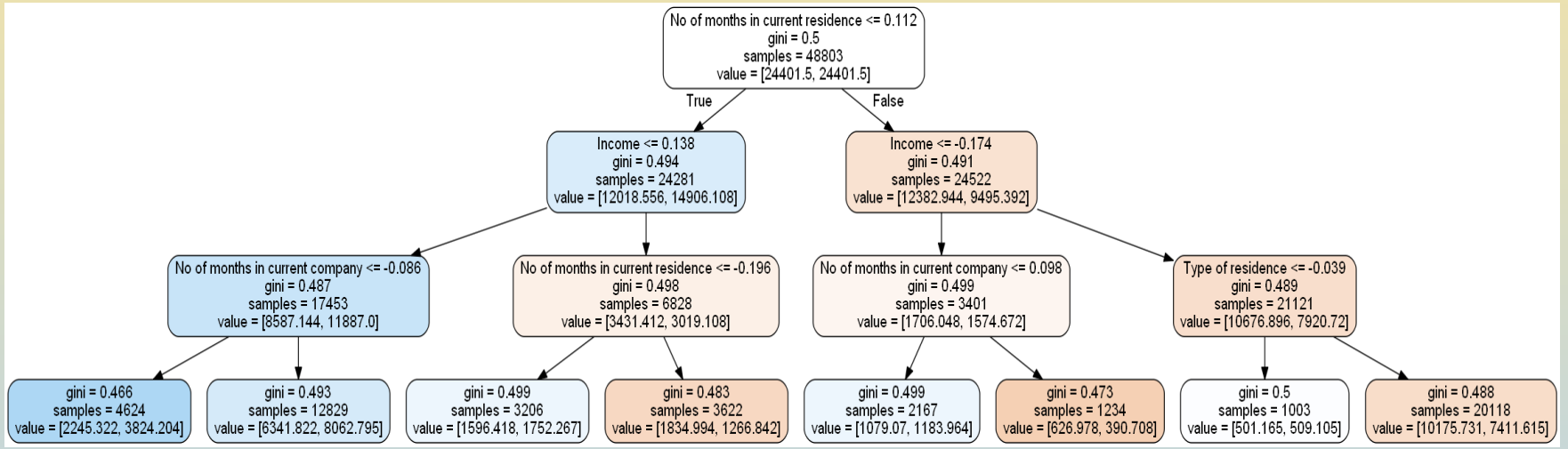




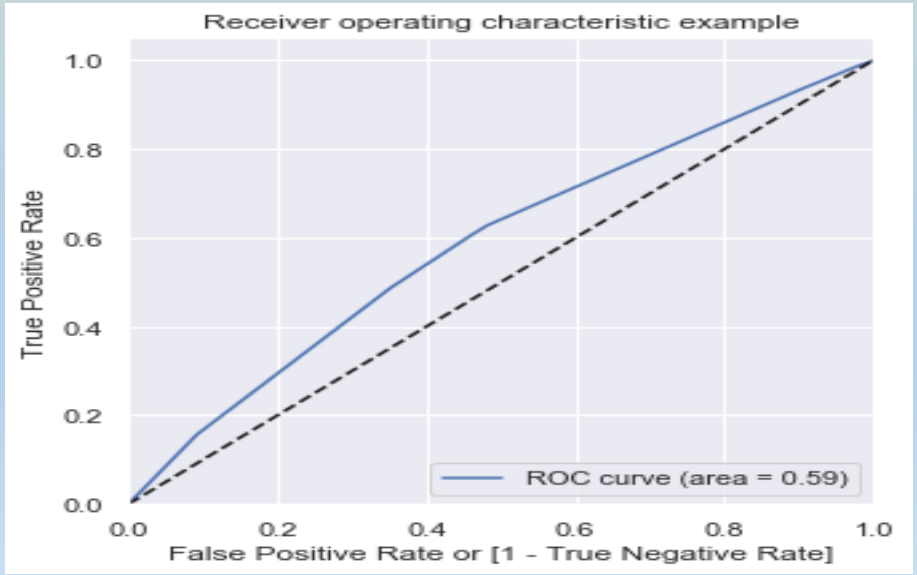
Logistic Regression	
Metrics	**Values**
AUC	0.59
Accuracy	55.63 %
Sensitivity	58.21 %
Specificity	55.52 %

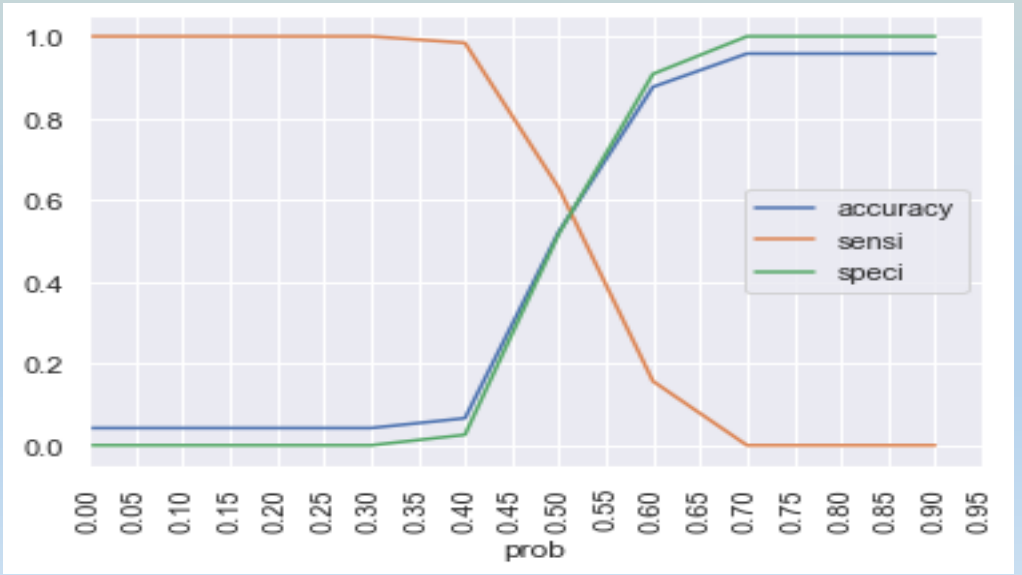
Model Building and Evaluation

- **Demographic data model: Decision Tree**



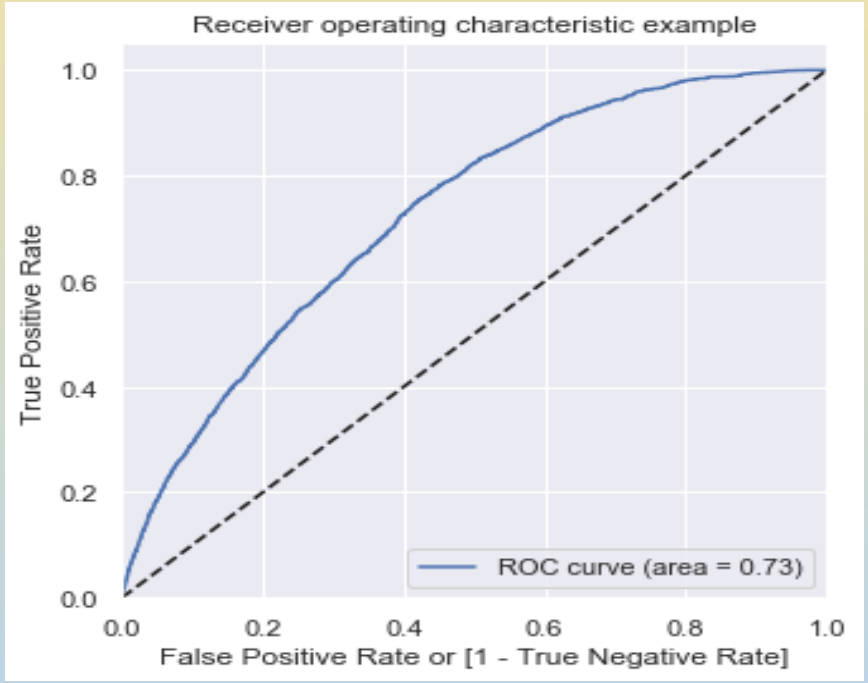
Decision Tree	
Metrics	**Values**
AUC	0.580
Accuracy	54.20 %
Sensitivity	59.90 %
Specificity	53.95 %

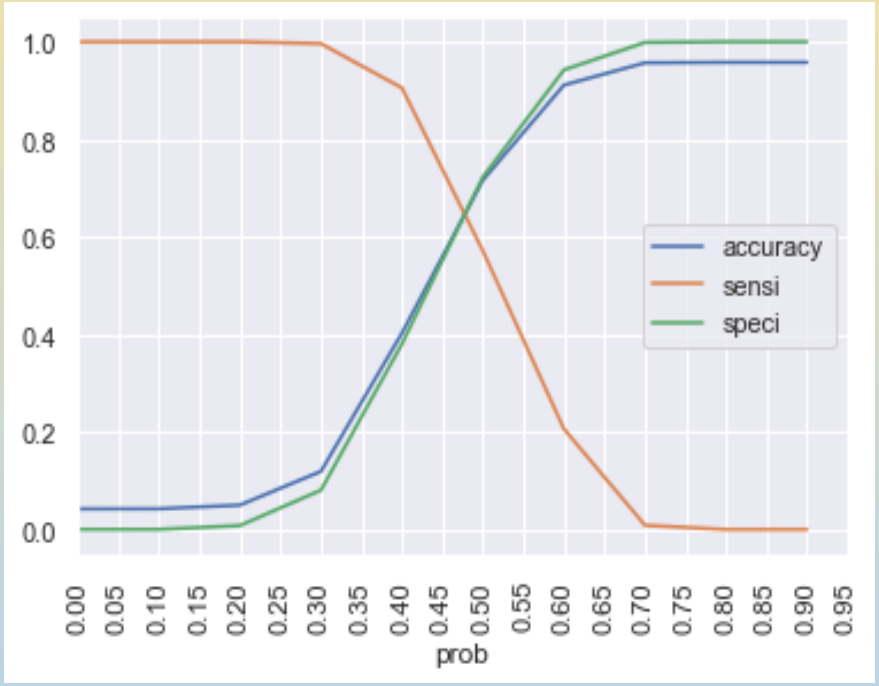




Model Building and Evaluation

- **Demographic data model: Random Forest**





Random Forest	
Metrics	**Values**
AUC	0.562
Accuracy	63.57 %
Sensitivity	45.18 %
Specificity	64.38 %

Model Building and Evaluation

Summary of demographic data model:

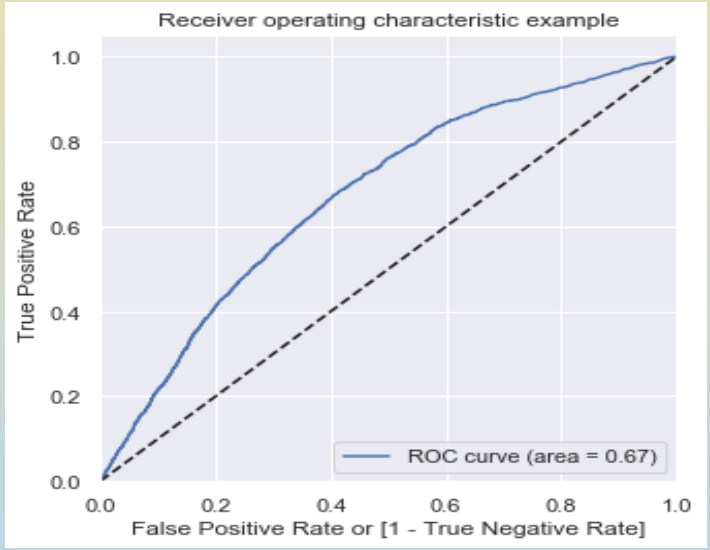
- Applied Logistic Regression, Decision Tree and Random Forest to demographic data.
- In logistic regression- we iteratively built the model by removing insignificant variables and variables exhibiting high levels of multicollinearity.
- We have tuned hyper parameters with cross validation.
- We have evaluated the model based on AUC, accuracy, sensitivity and specificity values.
- Random forest's test sensitivity score is very low when compared to logistic regression.
- Logistic Regression and decision tree model values are almost the same.
- However, overall performance of all the model is low. This shows demographic data alone is not sufficient for model building.

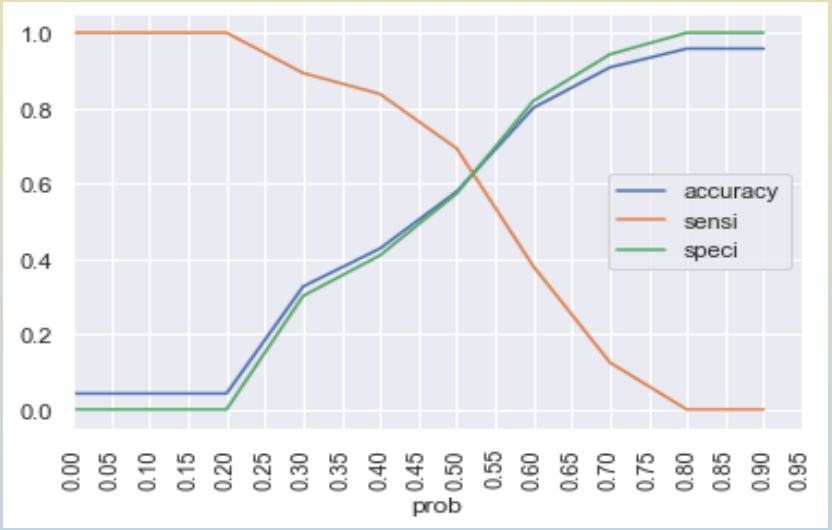
Metrics	Logistic Regression Values		Decision Tree Values		Random Forest Values	
	Train	Test	Train	Test	Train	Test
AUC	0.59	0.597	0.59	0.580	0.73	0.562
Accuracy	55.91%	55.63 %	54.13%	54.20 %	65.41%	63.57 %
Sensitivity	57.93%	58.21 %	60.74%	59.90 %	65.30%	45.18 %
Specificity	55.82%	55.52 %	53.84%	53.95 %	65.42%	64.38 %

Model Building and Evaluation

Complete data model: Logistic Regression

- For model building using logistic regression we have used IV for feature elimination
- We have selected all the variables with IV value greater than 0.02, i.e we have selected all the variables with IV predictiveness other than not useful
- The number of features selected is 17

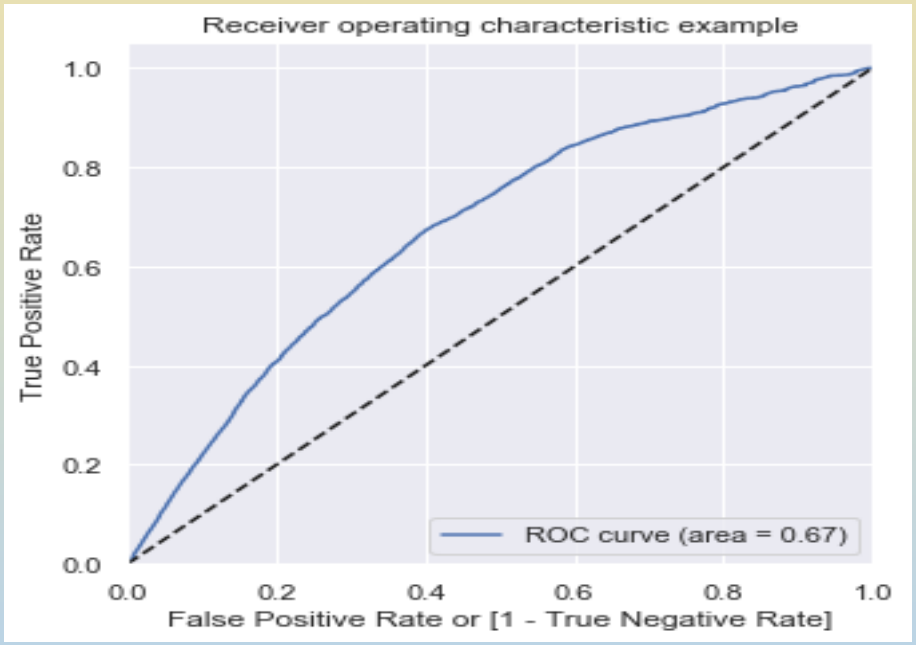


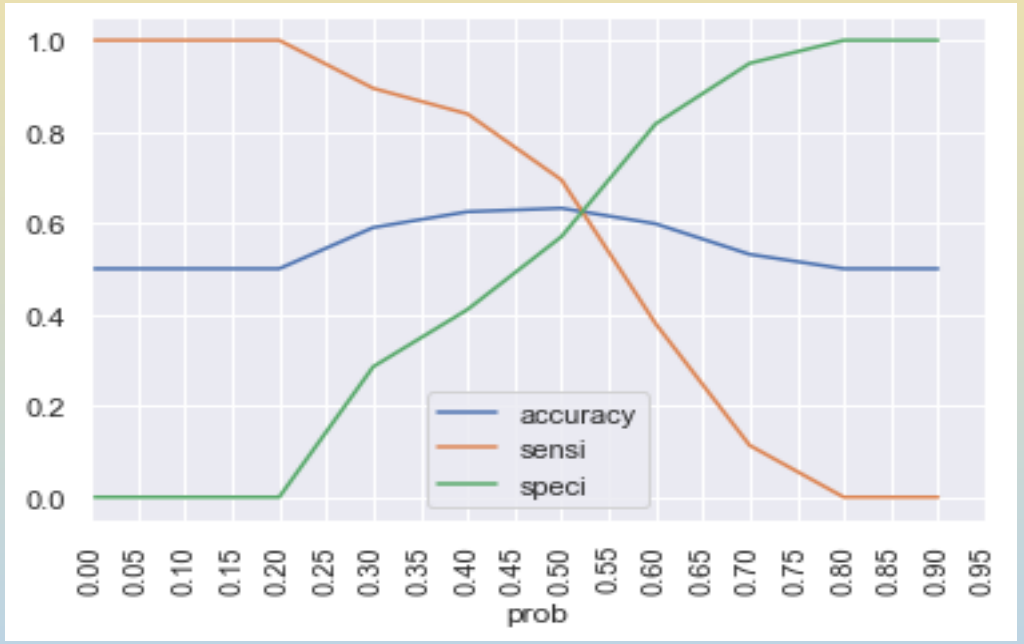


Logistic Regression	
Metrics	**Values**
AUC	**0.668**
Accuracy	**60.04 %**
Sensitivity	**66.47 %**
Specificity	**59.75 %**

Model Building and Evaluation

Complete data model: Logistic Regression with SMOTE

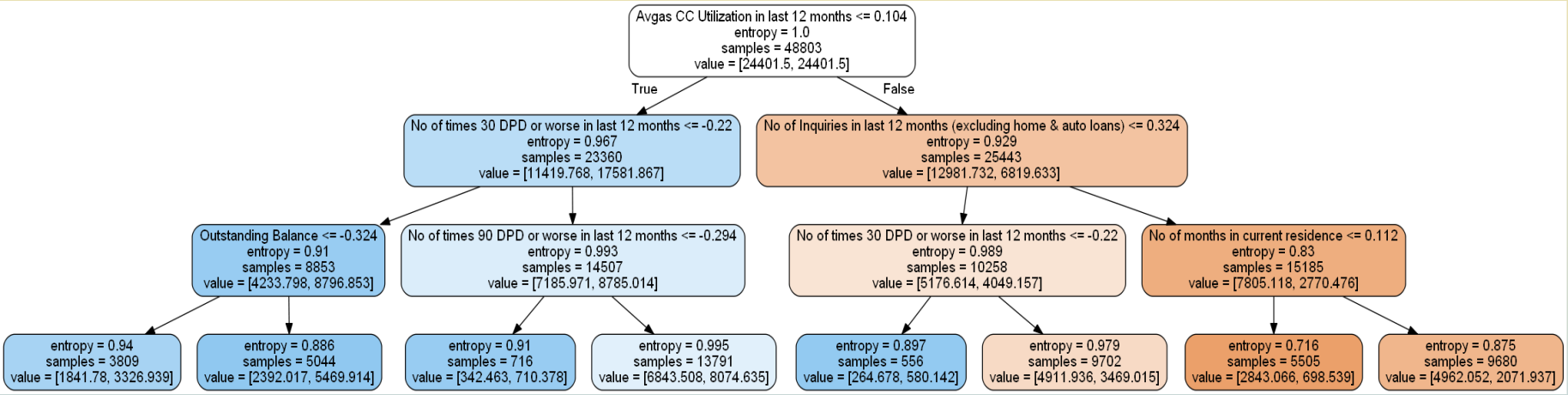




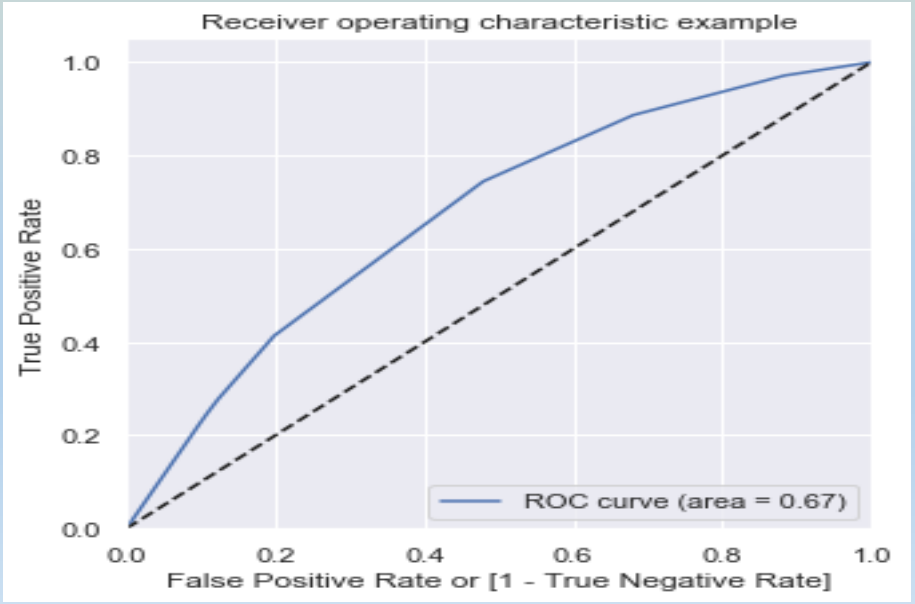
Logistic Regression with SMOTE	
Metrics	Values
AUC	0.667
Accuracy	59.77 %
Sensitivity	65.79 %
Specificity	59.50 %

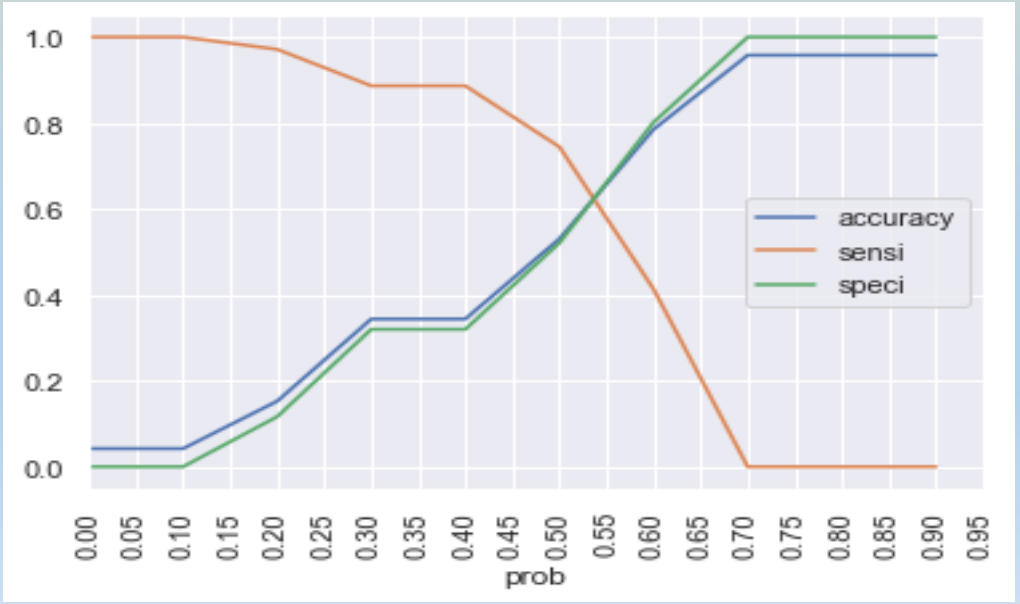
Model Building and Evaluation

Complete data model: Decision Tree



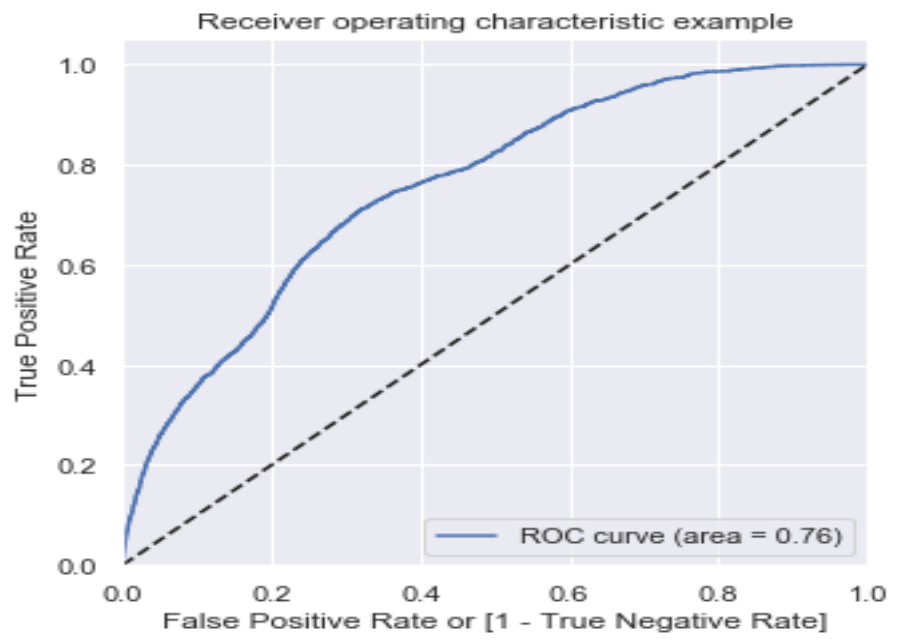
Decision Tree	
Metrics	**Values**
AUC	0.662
Accuracy	53.43 %
Sensitivity	74.40 %
Specificity	52.51 %

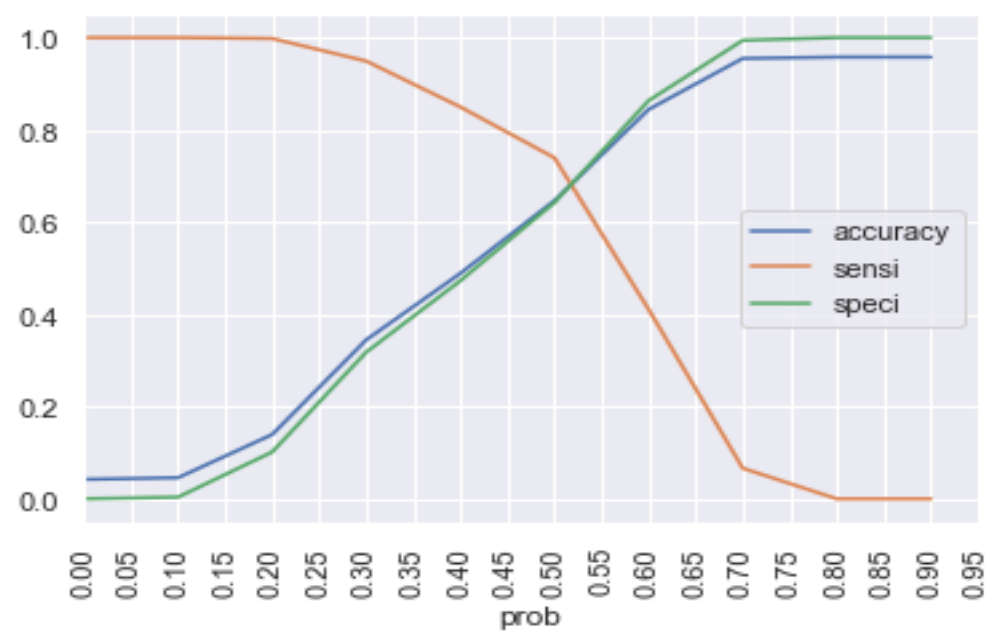




Model Building and Evaluation

Complete data model: Random Forest





Random Forest	
Metrics	**Values**
AUC	0.670
Accuracy	68.72 %
Sensitivity	54.13 %
Specificity	69.36 %

Model Building and Evaluation

Summary of complete data model(without rejected applicants):

- Applied Logistic Regression, Logistic Regression with SMOTE, Decision Tree, Random Forest and Support vector classification to complete data(without rejected applicants).
- For other models without SMOTE , we have applied class_weight = 'balanced' to handle imbalance dataset.
- We have tuned hyper parameters with cross validation.
- We have evaluated the model based on AUC, accuracy, sensitivity and specificity values.
- We have chosen logistic regression as our final model because of the following reasons:
  - ✓ Logistic regression with and without SMOTE has no much of a difference in values.
  - ✓ Logistic regression metrics like accuracy, sensitivity and specificity are consistent and are better compared to Decision tree model.
  - ✓ Though random forest had better training metrics, the test metrics has significant dip which shows that there is over fitting in the random forest model.

Metrics	Logistic Regression Values		Logistic Regression Values With SMOTE		Decision Tree Values		Random Forest Values	
	Train	Test	Train	Test	Train	Test	Train	Test
AUC	0.67	0.668	0.67	0.667	0.67	0.662	0.76	0.670
Accuracy	59.91%	60.04%	63.55%	59.77%	53.05%	53.43%	70.16%	68.72%
Sensitivity	67.29%	66.47%	67.75%	65.79%	74.42%	74.40%	68.17%	54.13%
Specificity	59.58%	59.75%	59.35%	59.50%	52.11%	52.51%	70.24%	69.36%

Performance of Final model

- The final model chosen is logistic regression because of the previously explained reasons
- Using the Logistic regression and the scorecard build we were able to identify 1974 defaulters (out of 2944)
- The percentage of defaulters identified = 1974*100/2944 = 67%

- Model rejected some non-defaulters because of their score being less than the cut-off
- Number of non-defaulters rejected = 26953
- Percentages of non-defaulters rejected = 26953*100/ 66775.0 = 40%

- The initial default percentage = 2944*100/69858 = 4.2%
- Default percentage after model = 970*100/40792 = 2.3%
- Therefore model has helped in **reducing** the defaulters percentage from **4.2% to 2.3%**

Performance on Rejected Applicants

- Out of 1425 rejected applicants the model identified 1411 applicants as defaulters
- The percent of identified rejected applicants = 1414*100/1425 = 99%
- The model has identified **99%** of the rejected applicants as defaulters

Note: While evaluating performance on rejected applicants we have imputed values with woe values that we have obtained for the non-rejected dataset

Factors affecting credit risk

Important variables which affect the credit risk as identified by our model are:

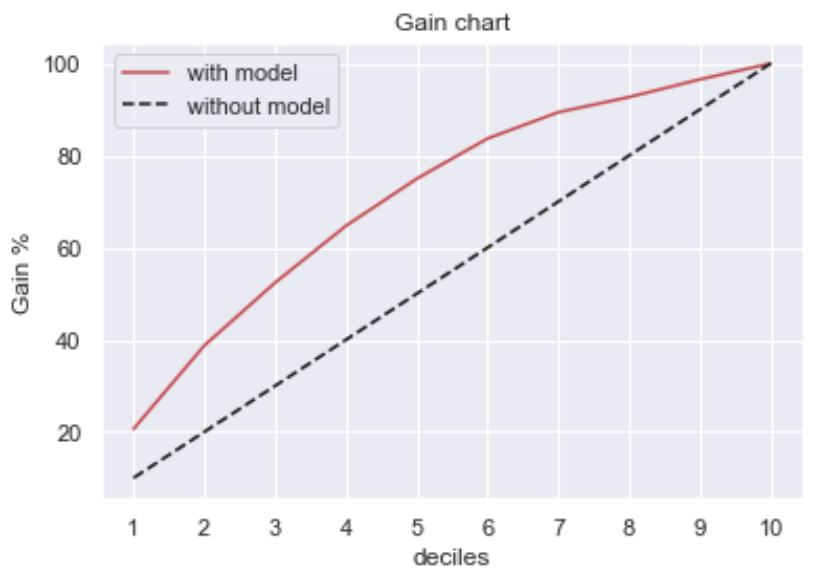
- Out of the 17 variables that we have chosen for building logistic regression model
  - No of Inquiries in last 12 months (excluding home & auto loans)
  - Avgas CC Utilization in last 12 months
  - No of times 30 DPD or worse in last 12 months
- The above are the top 3 variables that are affecting the default probability negatively
  - No of times 60 DPD or worse in last 6 months
  - No of Inquiries in last 6 months (excluding home & auto loans)
  - Total No of Trades
- The above are the top 3 variables the affect the default probability positively

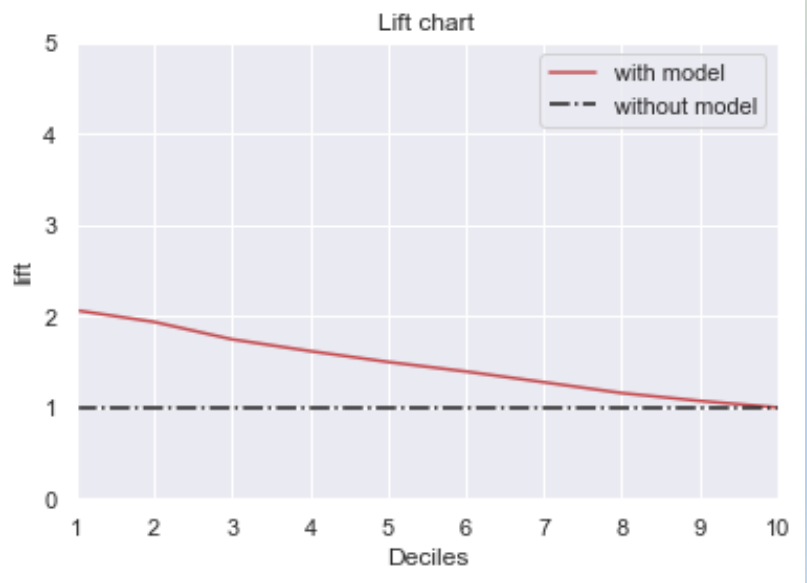
The equation for default probability:

default_prob = -0.42 x **Avgas CC Utilization in last 12 months**-0.2 x **Income**-0.43 x **No of Inquiries in last 12 months (excluding home & auto loans)**+0.16 x **No of Inquiries in last 6 months (excluding home & auto loans)**-0.11 x **No of PL trades opened in last 12 months**-0.11 x **No of PL trades opened in last 6 months**+0.01 x **No of months in current residence**-0.36 x **No of times 30 DPD or worse in last 12 months**-0.14 x **No of times 30 DPD or worse in last 6 months**-0.16 x **No of times 60 DPD or worse in last 12 months**+0.17 x **No of times 60 DPD or worse in last 6 months**-0.11 x **No of times 90 DPD or worse in last 12 months**-0.08 x **No of times 90 DPD or worse in last 6 months**-0.24 x **No of trades opened in last 12 months**+0.06 x **No of trades opened in last 6 months**-0.1 x **Outstanding Balance**+0.15 x **Total No of Trades**

Lift and Gain charts

- The deciles have decreasing order of probability of default, i.e decile 1 has higher probability of default than decile 2 and decile 2 has more than decile 10
- From the gain chart we can see that we are performing better at identifying defaulters than a random guess (dashed line)
- We can see that we are able to identify 83% of defaulter from top 6 deciles
- From the Lift chart we can see that we are able to perform twice better with the model when compared to random guess





Application Scorecard

- We have build application scorecard using our best **Logistic regression** model. Application score card was made with odds of 10 to 1 at a score of 400 doubling every 20 points.

- The scorecard was made using the following steps:
  - Probability of default for all applicants were calculated.
  - Odds for good was calculated. Since the probability computed is for rejection (bad customers)

    Odd(good) = (1-P(bad))/P(bad)
  - Log of odds i.e. ln(odd(good)) was calculated.
  - Used the following formula for computing score:

    400 + slope * (ln(odd(good)) -ln(10))

    Where, slope=20/(ln(20)-ln(10))
  - **Score card cutoff:**

    The cut-off for probability for our logistic regression is 0.521, hence the corresponding score in the scorecard will be the cut-off:

    odd(good) = 1-0.521/0.521 = 0.9193857964

    ln(odd(good)) = -0.084049444

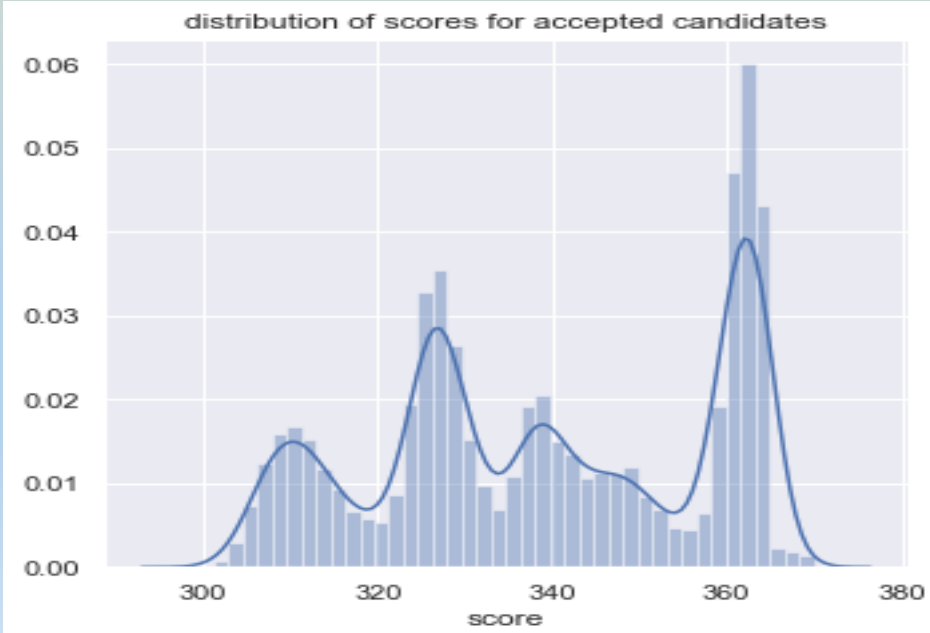
    Score = 400 + 20 *(-0.084049444-ln(10))/(ln(20)- ln(10)) = **331.13**

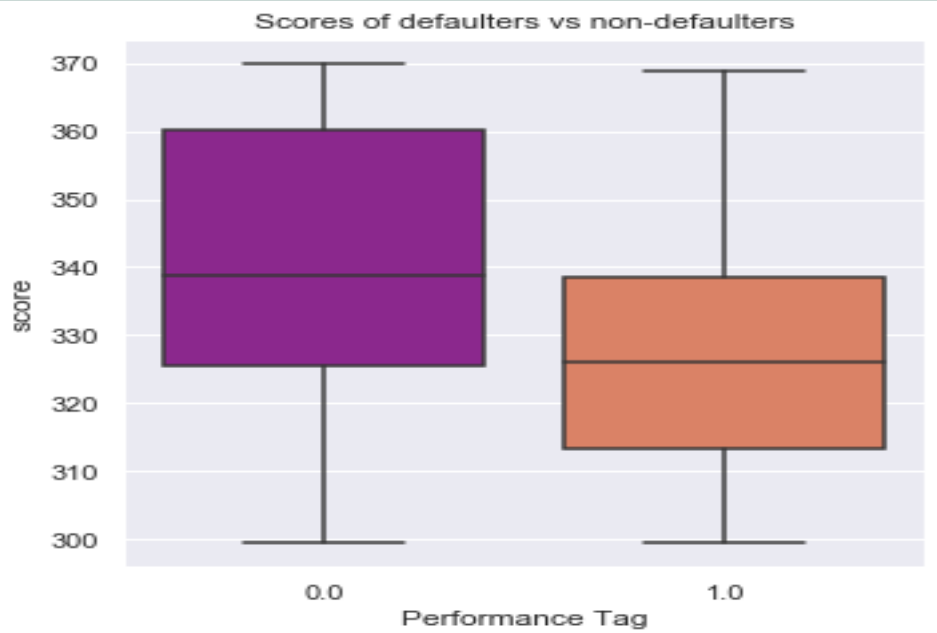
Application Scorecard

Summary of application score card values:

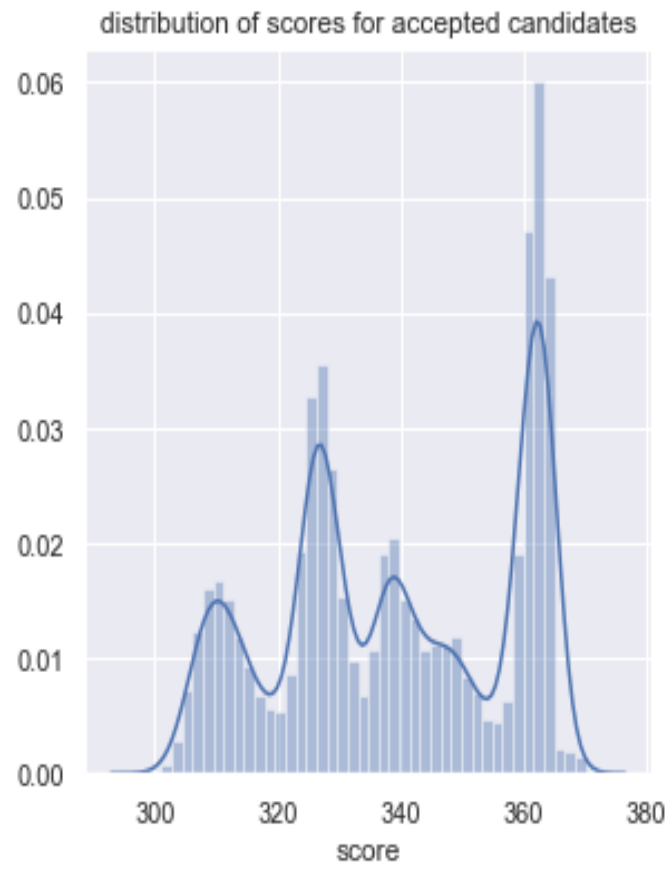
Performance Tag	count	mean	std	min	25%	50%	75%	max
0.0	66775.0	339.519084	18.859495	299.510973	325.620915	338.848553	360.219962	369.987940
1.0	2944.0	328.076234	16.765545	299.404681	313.411627	326.102752	338.378027	368.734719

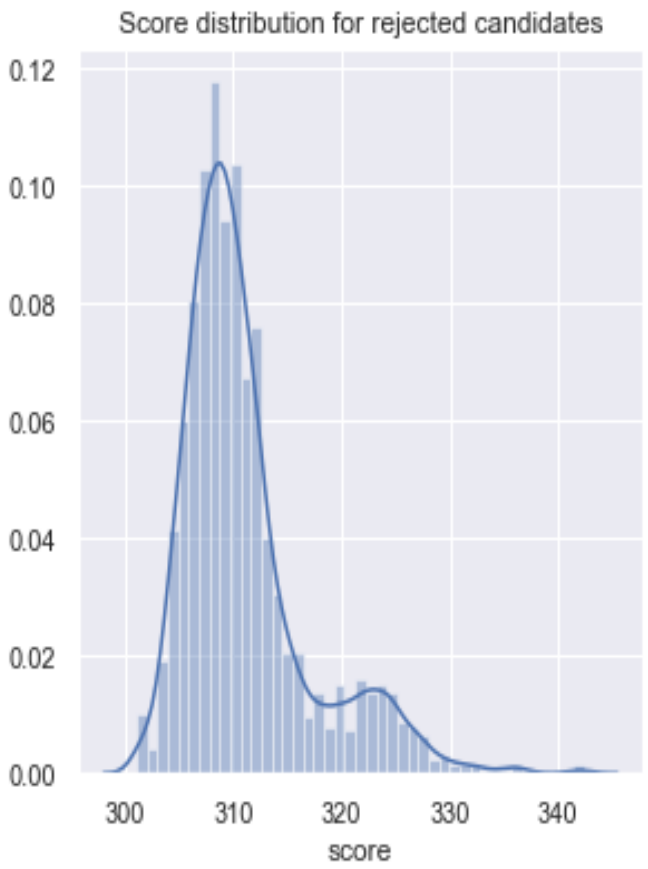
- ✓ Scores range from 299.4 to 369.98 for accepted applicants.
- ✓ The mean score of defaulter is 328, where as mean score of non defaulters is 339.5
- ✓ The cut-off score below which we would not grant credit cards to applicants is 331.13

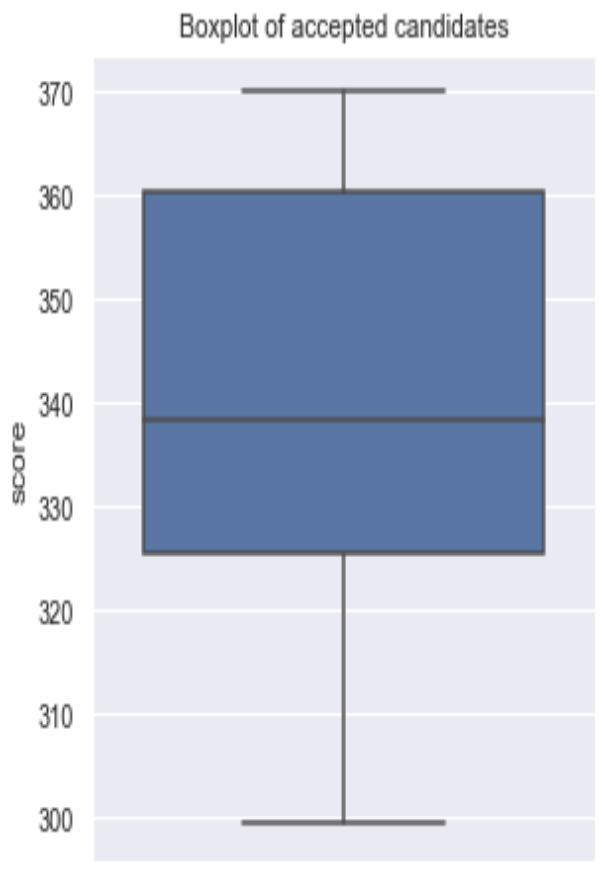


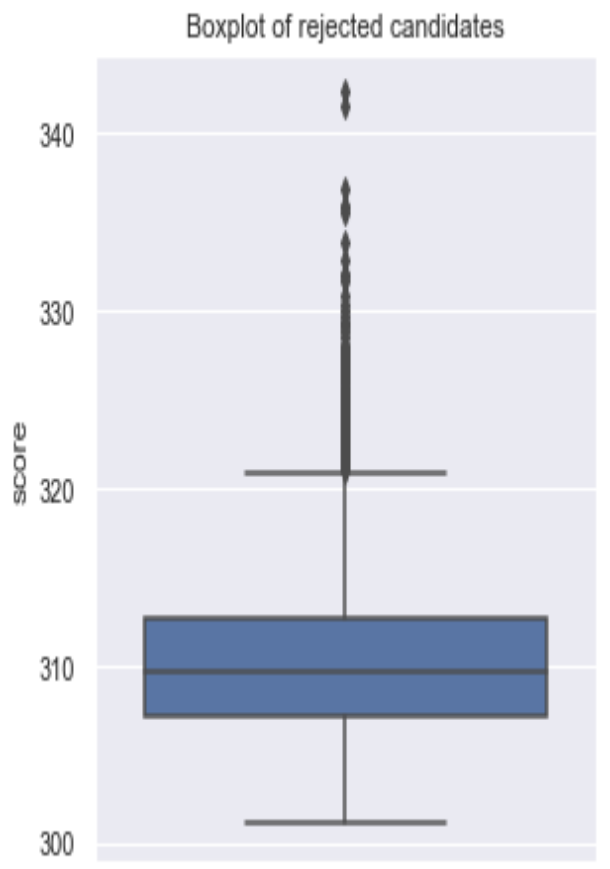


Performance of our logistic regression model and scores of rejected candidates









- We can clearly see that the scores for accepted candidates is much higher than the rejected candidates.
- The mean and median scores of accepted candidates are 339. 03 and 338.36 respectively.
- The mean and median scores of rejected candidates are 311.11 and 309.60 respectively.
- 99.22 % of the rejected applicants are identified correctly as defaulters by our model.

Financial Benefit Analysis

Financial benefit from the model can be divided into two buckets:

- The loss that can be prevented is by rejecting the bad customers using the cut-off from the model
- The loss incurred is due to rejecting the good customers with score less than the cut-off.

Assumptions:

- The variable 'Avg CC utilization' is taken as proxy for the amount earned because throughout EDA , IV and in our model we have seen that average credit card utilization is the important variable for identifying the defaulters.
- The unit for Avg CC is taken as thousands which is a fair assumption based on the dataset
- We are assuming that every non-defaulter will spend the avg. CC utilization every month and CredX earns a 10% on the amount spend.
- For every defaulter, we assume that CredX looses 5 times the avg CC utilization.

Based on the assumption,

- The avg CC for each defaulter is 39.9k (which is the mean avg CC with performance tag as '1')
- The avg CC for each non-defaulter is 28.3k (which is the mean avg CC with performance tag as '0')

- The amount that each defaulter will cost to the company = 39.9*5 = **199.5k (5times the avg CC util)**
- The amount that each non-defaulter will pay to the company = 28.3*0.1 = **2.8k (10% of avg CC util)**

Financial Benefit Analysis

Credit loss prevented:

- Credit loss is due to default customers not paying the amount taken
- Total credit loss prevented = no.of defaulters identified **x** cost due to each defaulter
- In approved candidates we had 2944 defaulters. Using our model, the number of defaulters identified i.e. the number of defaulters with score less than cut-off (331.13) is 1974.
- Therefore credit loss prevented = 1974*199.5k = **394 million**

Revenue Loss incurred:

- The loss incurred will by rejecting the good customers with score less than the cut-off.
- Total revenue loss incurred = No.of non-defaulters identified **x** loss due each non-defaulter
- Using our model, the number of non- defaulters having score less than cut-off (331.13) is 26953.
- Therefore revenue loss incurred = 26953*2.8 = **76 million**

Net saving achieved by the model is:

Net saving because of model = Credit loss prevented – Revenue loss incurred

= 394 – 76 **= 318 million**

Conclusion

- ✓ We have observed that Model with combined data **performed better** than the model with only demographic data; Logistic regression is chosen as the final model
- ✓ Application scorecard was build using the chosen logistic regression model
  - ✓ The maximum credit score was **370** (approx.); **331.13** is the obtained score cut-off
  - ✓ Significant difference in mean scores for rejected (**311**) and non-rejected applicants (**339**)
  - ✓ Model identified **99%** of the rejected applicants as defaulters
- ✓ Important features identified from the logistic regression are
  - ✓ No of Inquiries in last 12 months, Avgas CC Utilization ,No of times 30 DPD or worse in last 12 months; These are top 3 features effect the default probability negatively
  - ✓ No of times 60 DPD or worse in last 6 months, No of Inquiries in last 6 months, Total No of Trades; There are top 3 features that effect the default probability positively
- ✓ **Financial Benefit Analysis**
  - ✓ **Credit loss prevented** = No of defaulters identified * cost per each defaulter
    = 1974*199.5k = **394 million**
  - ✓ **Revenue loss incurred** = No of non-defaulters identified * loss per each non-defaulter
    = 26953*2.8k = **76 million**
  - ✓ Total loss prevented = Credit loss prevented – Revenue loss incurred
  - ✓ Therefore **Total loss prevented** = 394-76 = **318 million**

Thank You