BFS – Capstone Project

BY

Ramesh.R

Date – 10 June 2024

Business Objective

CredX is a leading credit card provider that gets thousands of credit card applications every year.

But in the past few years, it has experienced an increase in credit loss. The CEO believes that the best strategy to mitigate credit risk is to ‘acquire the right customers’.

In this project, our task is to help CredX identify the right customers using predictive models. Using past data of the bank’s applicants.

We need to determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of your project.

Data Understanding

There are two data sets provided for analysis:

- **Demographic** data
- **Credit Bureau** data.

Demographic/application data: This is obtained from 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 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.

Both files contain a **performance tag**, which 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.

In some cases, we have records for which all the variables in the credit bureau data are zero and credit card utilisation is missing. These represent cases in which there is a no-hit in the credit bureau.

There are also cases with missing credit card utilisation. These are the cases in which the applicant does not have any other credit card.

There are missing values in target variable **performance tag,** which indicates rejected applicants

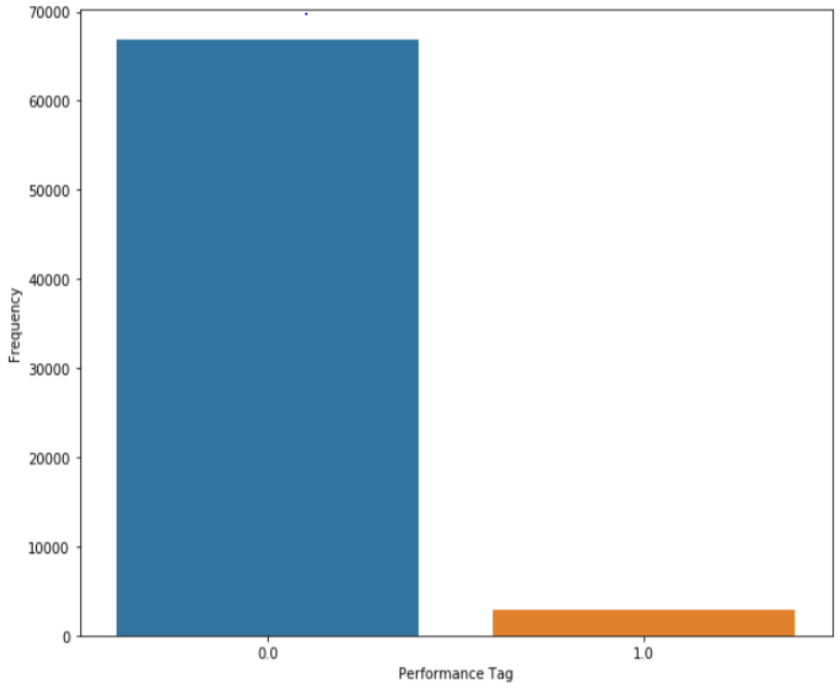
Data Understanding

Count of different values for Performance Tag which is dependant variable:

Customer	Performance Tag	Count
Good Customers	0	66920
Defaulters	1	2947
Rejected Applicants	NA	1425

Percentage of Defaulters: 4.23

There are missing values in target variable **performance tag,** which indicates rejected applicants



As it is a case of highly imbalanced class , this needs to be handled using balanced class during each model preparation.

Problem solving steps

- Perform data cleaning on both the data sets.
- Perform Exploratory Data Analysis on all the features of both the datasets to identify the important predictor variables.
- Perform Weight of Evidence (WOE) and Information Value (IV) analysis on Demographic dataset and create the WOE transformed datasets, Also find out the important variables based on the information values of the variables.
- Perform Logistic Regression model on the following datasets:
  - ✓ WOE Transformed Demographic Dataset
  - ✓ WOE Transformed Merged Dataset
- Perform some complex models like Decision tree or Random forest to see if the predictive power of the model is better than that of the logistic regression model.
- On the basis of the chosen model and significant variables in the model, score card would be prepared for the following:
  - ✓ Score card for the combined woe transformed dataset
  - ✓ Score card for the rejected applicants (records for which value of performance tag was missing)
- Access the financial benefits of the project by checking the underlying matrices that get optimized.
- Present all the results obtained in all the above steps to the management

Problem solving steps continued...

Important points while building the models:

- ✓ There is a class imbalance in the dataset. This needs to be handled using balanced class during each model preparation.
- ✓ Select the important variables based on higher Information Value(IV) and include those features while modelling
- ✓ Evaluate the accuracy of both the train and test sets and check for any overfitting
- ✓ Additional validation of data should be done on the dataset on the rejected applications (performance tag null) ignored for model building.
- ✓ Hyper parameters for tree models need to be optimized using GridSearch Cross validation and model with optimized parameter should be chosen.

Criteria for model evaluation

All the models will be evaluated on the following parameters :

- ✓ Confusion matrix for each model.
- ✓ Sensitivity, specificity, accuracy curve for each model with different cut-offs.
- ✓ AUC-ROC curve for the model using cut-off values for each model.
- ✓ Precision and Recall curve for cut-off should be generated.
- ✓ Gini-Index needs to be evaluated for Tree based models like decision tree and random forest.
- ✓ Within each model type evaluation using Grid Search based on recall values should be done to get models with optimized hyper parameters.
- ✓ For evaluation among models, the dataset for rejected applications (with performance tag missing), which were assumed as potentially defaulters should be considered for evaluations. Ideally, the output for all these applications should be defaulters.

Problem Solving Methodology

Data Cleaning – Demographic Dataset

Feature Name	Feature Description	Data Issues & Treatment
Application Id	Application id of the applicant	There were 3 duplicate application ids which were deleted from the dataset.
Age	Age of the applicant	Records where applicants age was less than 18, age was imputed with 18, assuming an applicant with age less than 18 years can not hold/apply for a credit card.
Gender	Gender of the applicant	Missing values for Gender were replaced with the more frequent value of the two i.e. Male.
Marital Status	Marital status of the applicant	Missing values for Marital Status were replaced with the more frequent value of the two i.e. Married.
No. Of Dependents	No. of dependents of the applicant	Missing values were replaced with the median value.
Income	Income of the applicant	Missing/Negative values were replaced with the median value.

Data Cleaning – Demographic Dataset Continued..

Feature Name	Feature Description	Data Issues & Treatment
Education	Education of the applicant	Missing values were replaced with “Unknown”.
Profession	Profession of the applicant	Missing values were replaced with the most frequent value “SAL”.
Type Of Residence	Type of residence applicant live in	Missing values were replaced with most frequent value 'Rented’.
No. Of Months in Current Residence	No. of Months in Current Residence	No Data Issues Found
No. Of Months in Current Company	No. Of Months in Current Company	No Data Issues Found
Performance Tag	Status of customer performance (" 1 represents "Default")	Records with missing performance tags were dropped as these records belonged to the customers who were never issued the credit card.

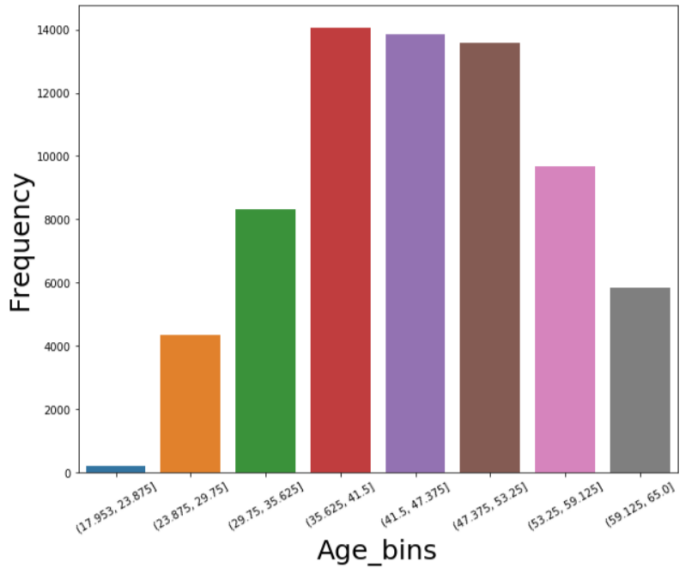
Data Cleaning – Credit Bureau Dataset

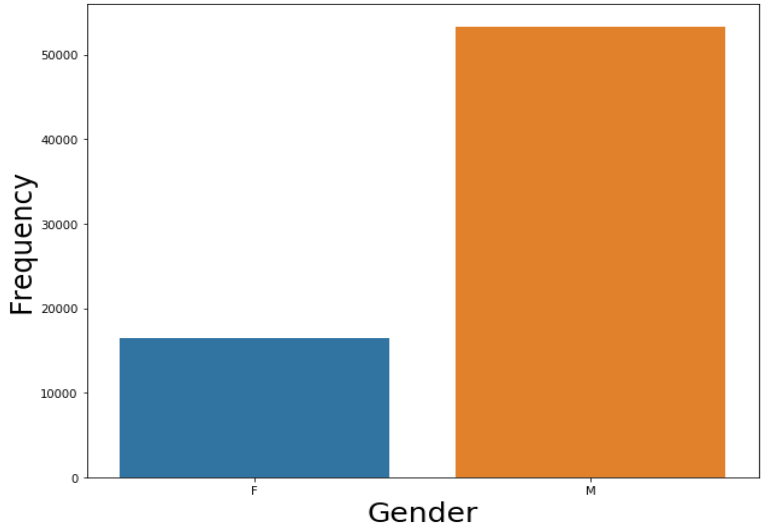
Feature(s) Name	Feature Description	Data Issues & Treatment
Application Id	Application id of the applicant	There were 3 duplicate application ids which were deleted from the dataset.
No of times 90/60/30 DPD or worse in last 6 months	No of times 90/60/30 DPD or worse in last 6 months	No Data Issues Found.
No of times 90/60/30 DPD or worse in last 12 months	No of times 90/60/30 DPD or worse in last 6 months	No Data Issues Found.
Avgas CC Utilization in last 12 months	Average utilization of credit card by customer	Missing values were replaced with 0 as null values for this field means credit card was never utilized in the last 12 months.
No of trades opened in last 6 months	Number of times the customer has done the trades in last 6 months	Missing values were replaced with the most frequent value i.e. 1.0.
No of trades opened in last 12 months	Number of times the customer has done the trades in last 12 months	No Data Issues found.
No of PL trades opened in last 6/12 months	No of PL trades in last 6/12 month of customer	No Data Issues found.

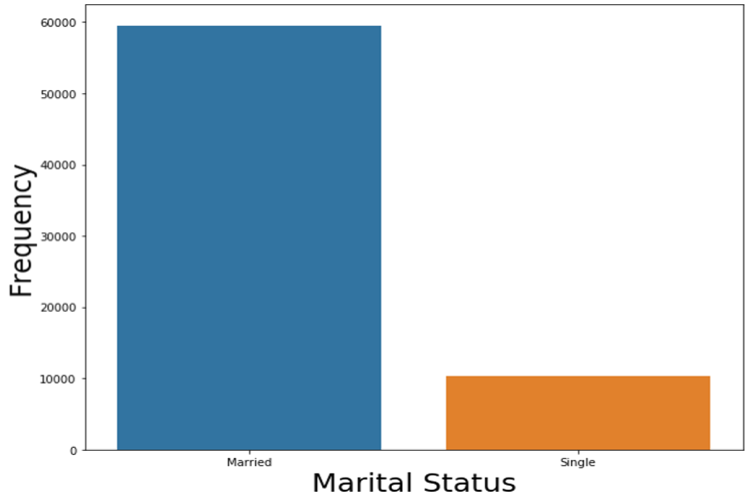
Data Cleaning – Credit Bureau Dataset

Feature Name	Feature Description	Data Issues & Treatment
No of Inquiries in last 6/12 months (excluding home & auto loans)	Number of times the customers has inquired in last 6/12 months	No Data Issues found.
Presence of open home loan	If the customer has home loan (1 represents "Yes")	Missing values were replaced with the most frequent one i.e. 1.
Outstanding Balance	Outstanding Balance of the applicant	Missing values were replaced with 0 assuming these are the records for which the credit card was never utilized as these records also had avg card utilization as null.
Total No of Trades	Number of times the customer has done total trades	No Data Issues found.
Presence of open auto loan	If the customer has auto loan (1 represents "Yes")	Missing values were replaced with the most frequent value i.e. 0.
Performance Tag	Status of customer performance (" 1 represents "Default")	Records with missing performance tags were dropped as these records belonged to the customers who were never issued the credit card.

EDA - Demographic Dataset-Socio(Univariate)

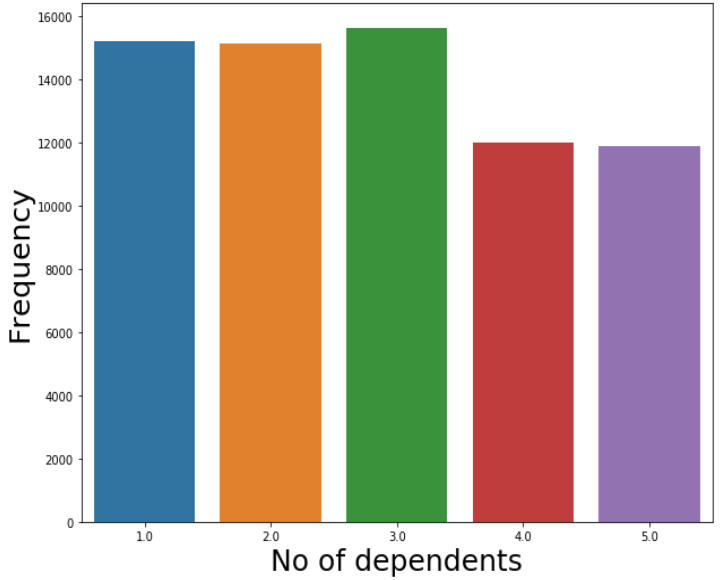


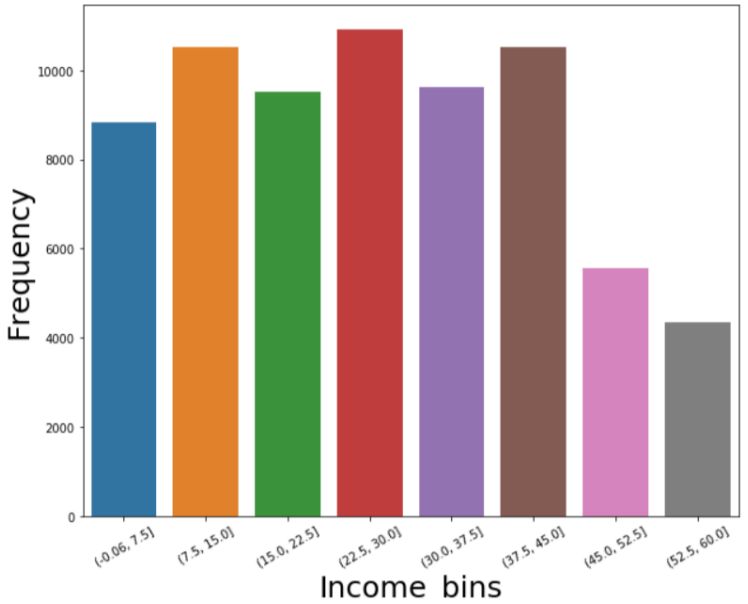


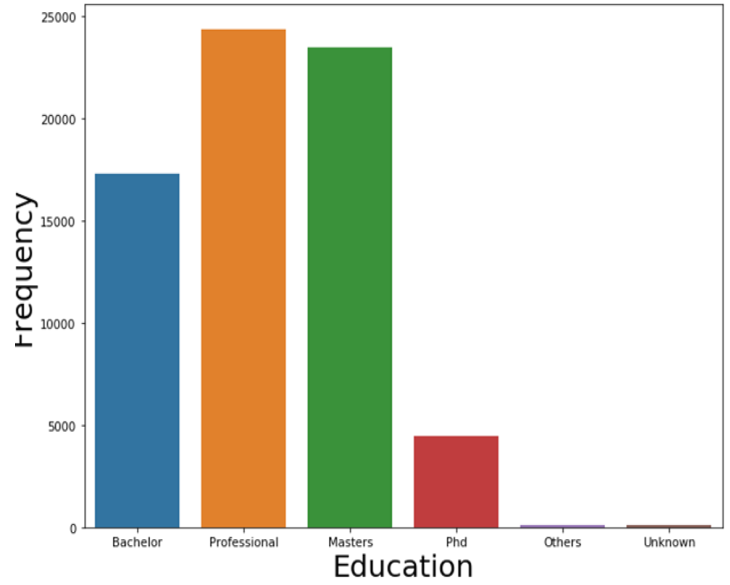


- Age groups: As expected middle aged people represent a significant portion of the customer base
- Gender: More number of Males compared to Females
- Marital Status: More Married customers in data than single

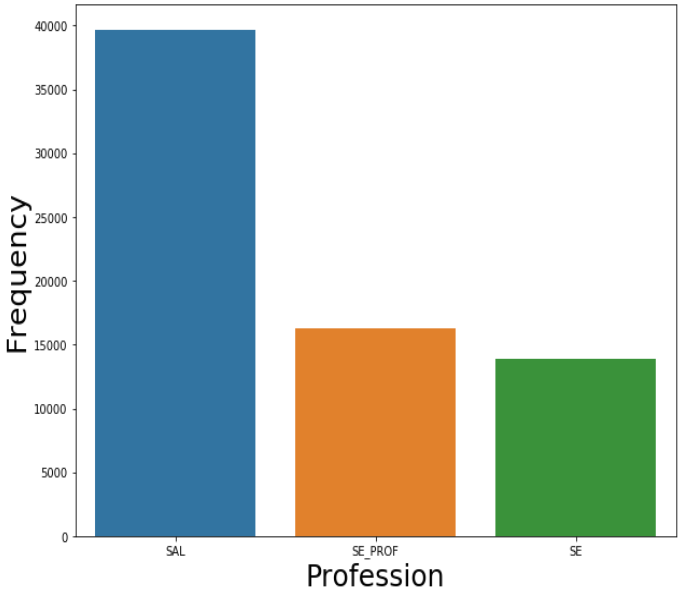
EDA - Demographic Dataset-Economic (Univariate)



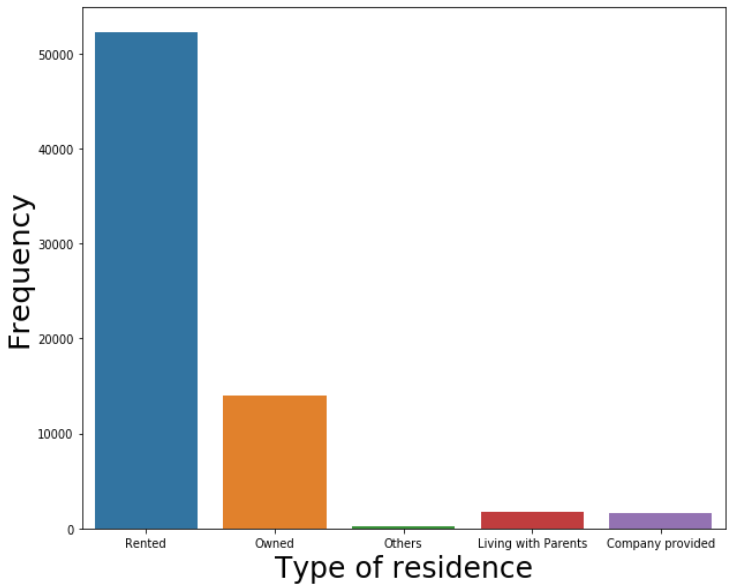


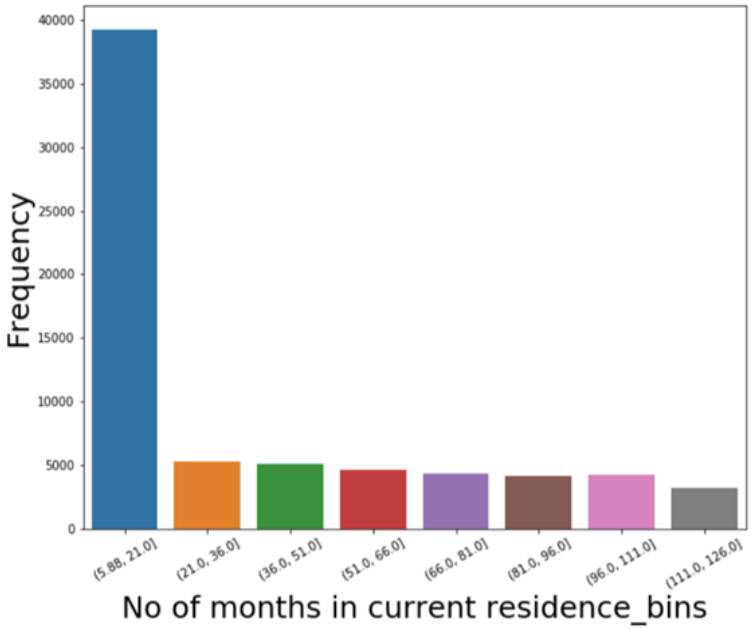


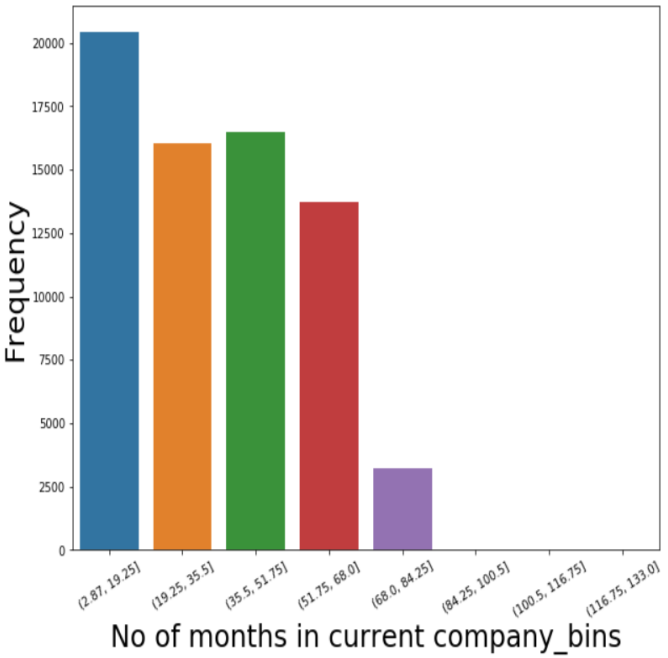
- No. of Dependants: Almost evenly distributed
- Education: Significant amount of Professionals/Masters
- Income Bins: High Income earners are low compared to low Income earners
- Profession: Significant number of Employees compared to self employed



EDA - Demographic Dataset (Univariate Analysis)

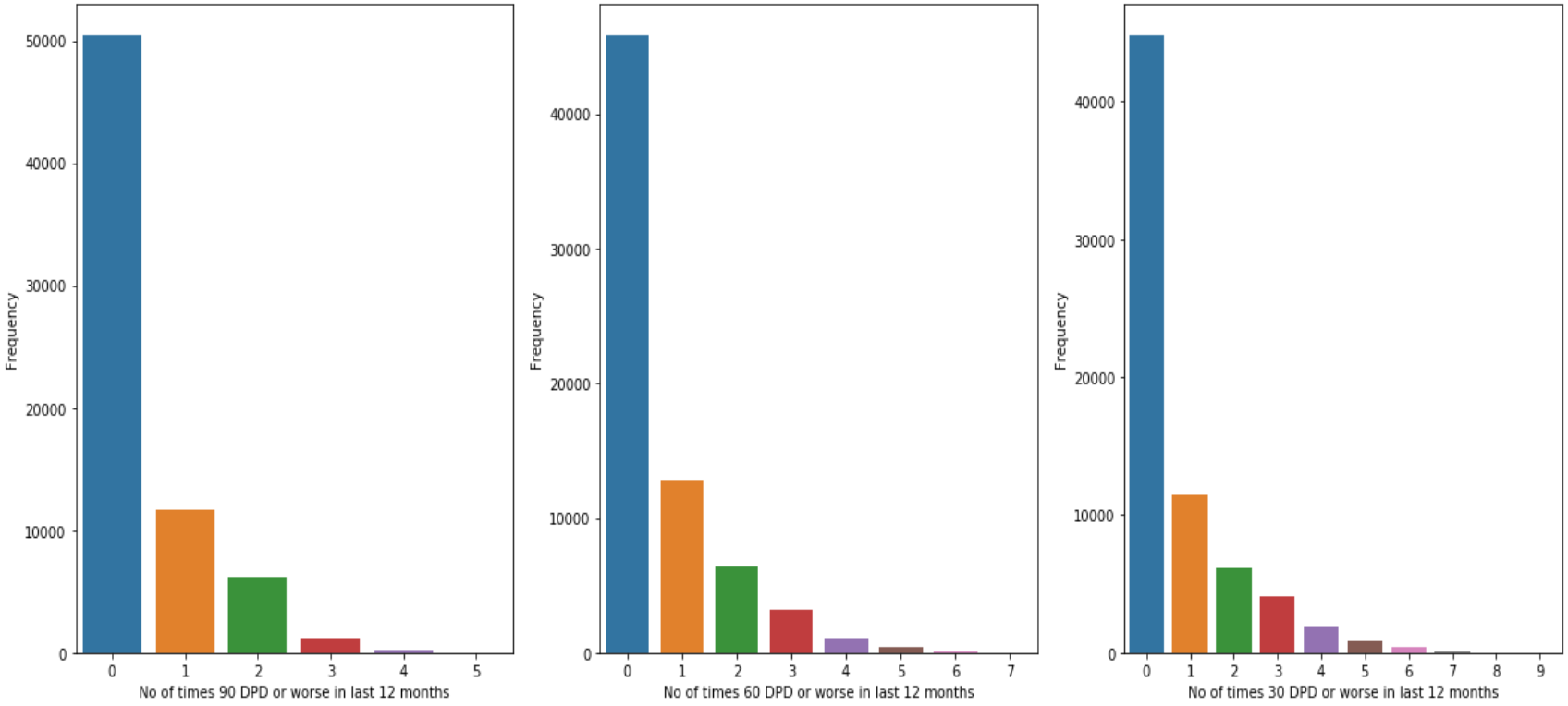






- Type of Residence: Significant amounts of Customers living in rented homes and then followed by owned houses
- No. of months in current residence: Short term residents are way high than long term residents
- No. of months in current company : Distributed well, a little more short term employees compared to others

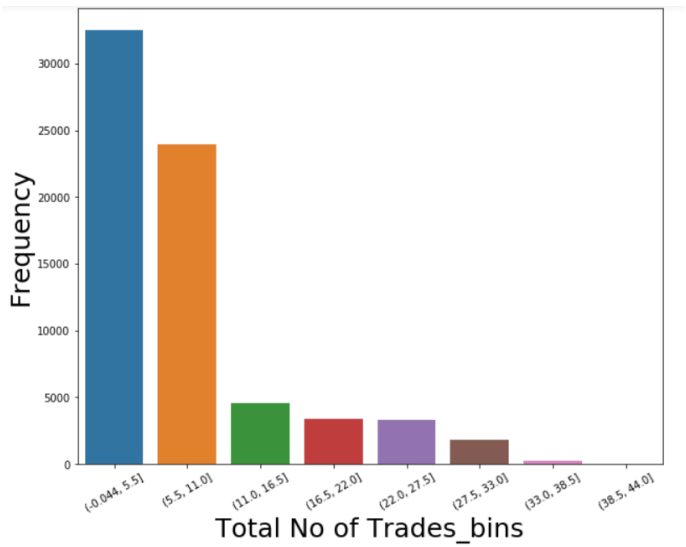
EDA – No. of times DPD(Univariate Analysis)

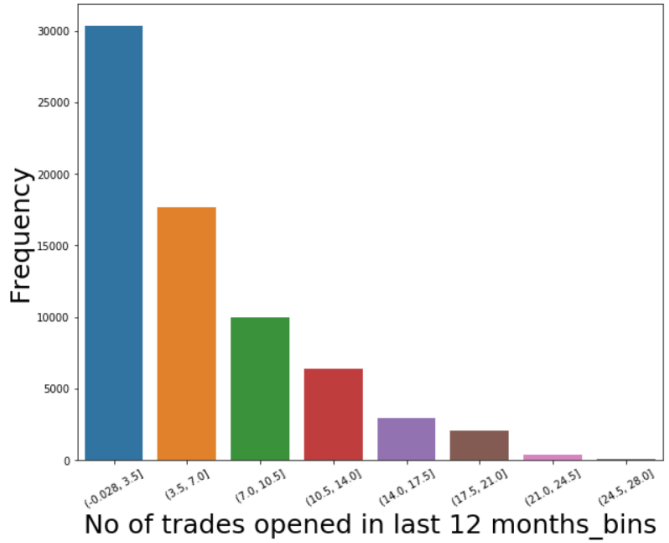


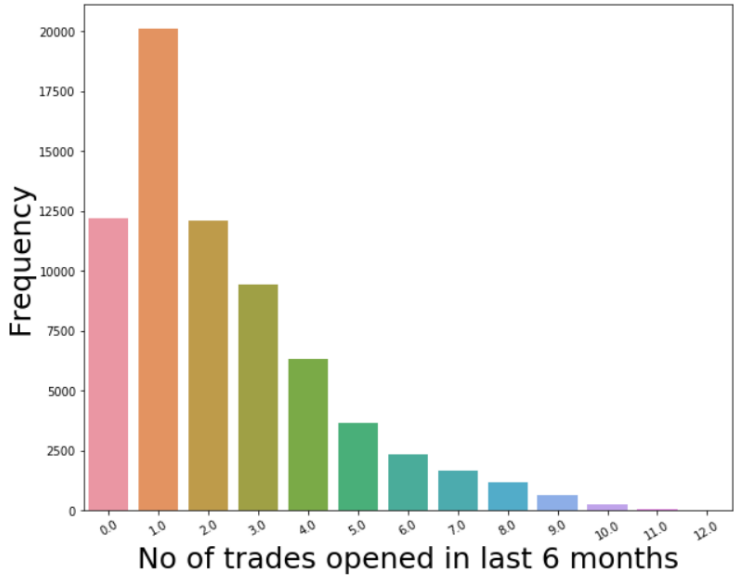
Observations:

No of times 90/60/30 DPD in last 12months - Majority of the customers belongs to 0 value, which means most of the customers have paid their dues on time.

EDA – Trades opened(Univariate Analysis)

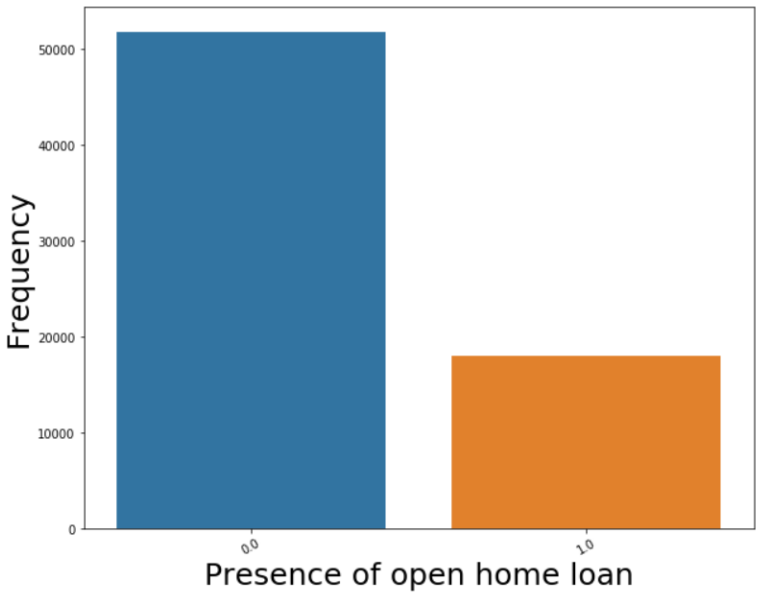


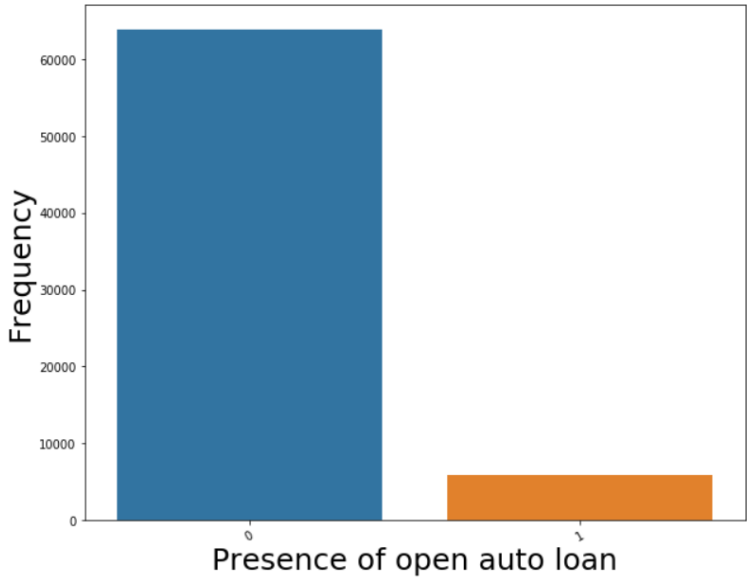




Total trades and trades opened in recent months follows a similar pattern

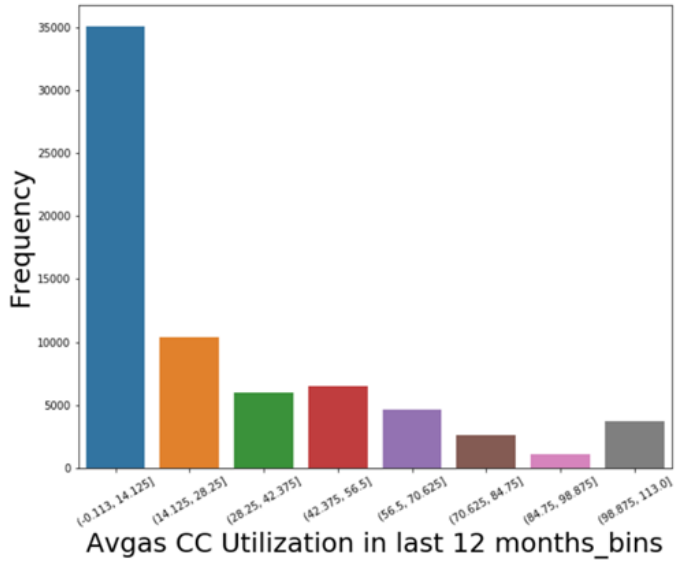
EDA – Presence of loan & CC utilization& Balance

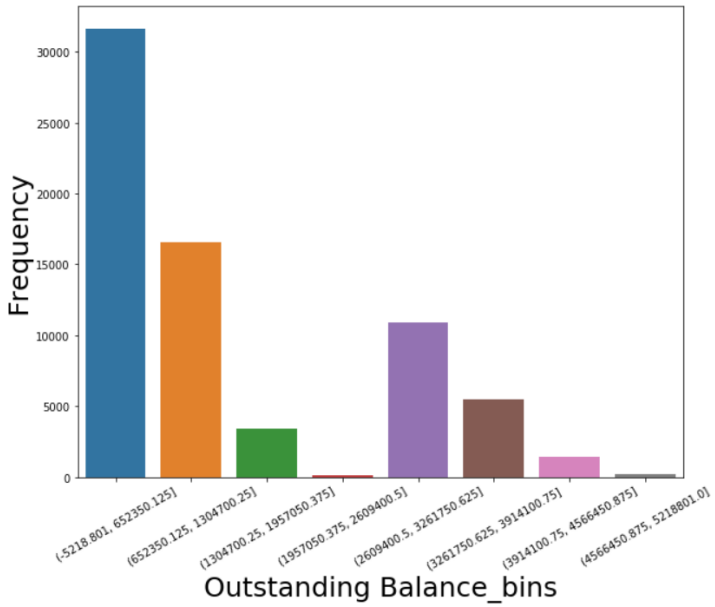




- Presence of auto loan, home loans:
  0 – no loan
  1 – loan

- Customers have more home loans
- than auto loans

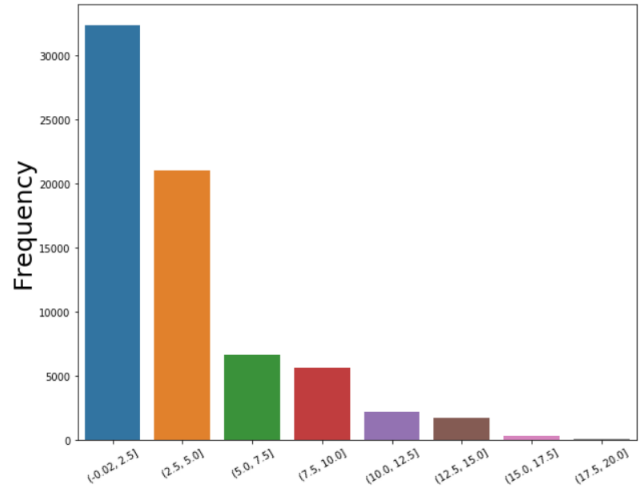




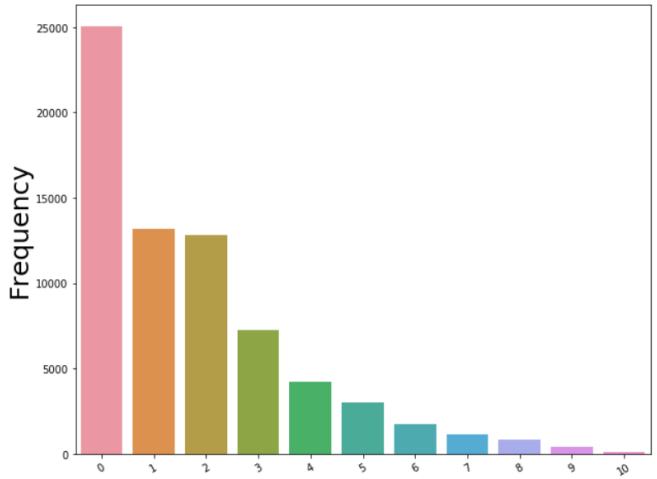
- Customers with low usage of Credit card are high and followed by others

- Customers with low outstanding balance are more

EDA – PL trades & Inquiries(Univariate Analysis)

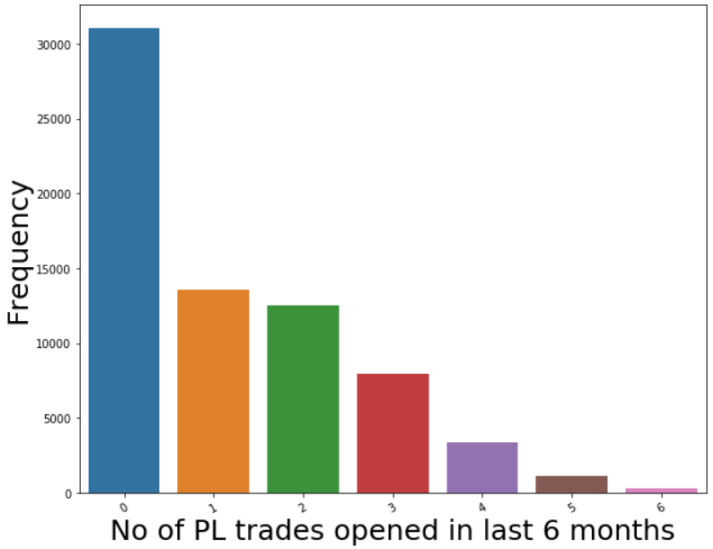


No of Inquiries in last 12 months (excluding home & auto loans)

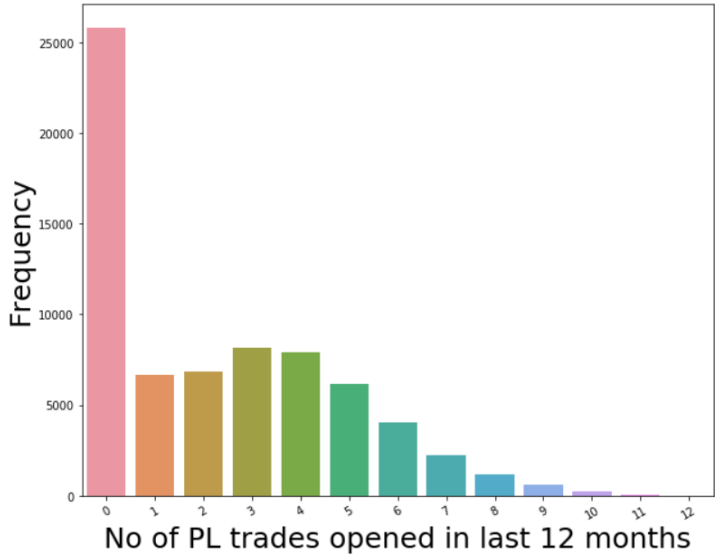


No of Inquiries in last 6 months (excluding home & auto loans)

- Customers with no enquiries are more and following
- Not much Inquiries overall



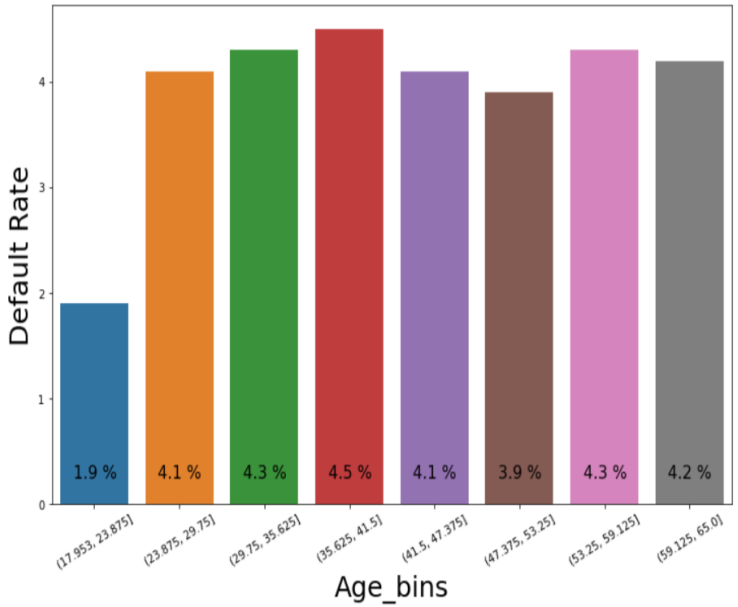
No of PL trades opened in last 6 months

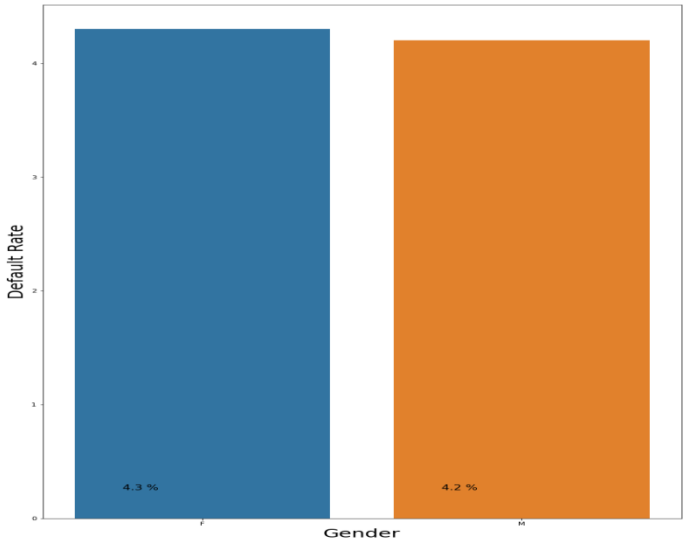


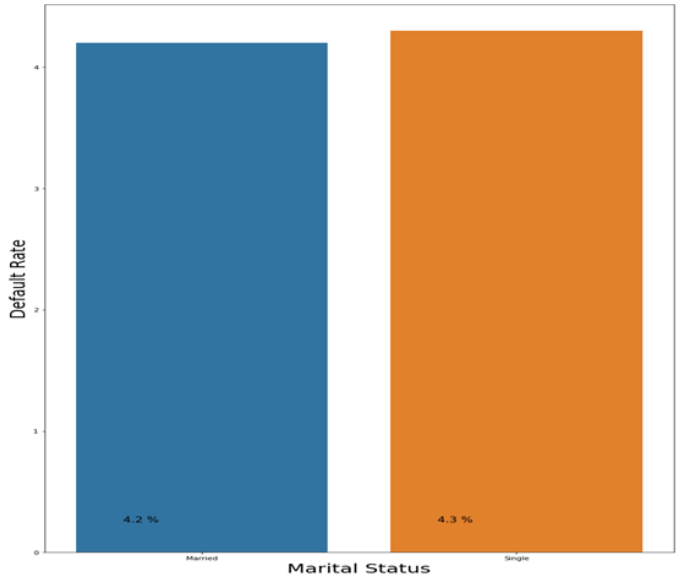
No of PL trades opened in last 12 months

- PL trades opened almost follows a similar pattern

EDA – Demographic-Socio Dataset (Bivariate Analysis)

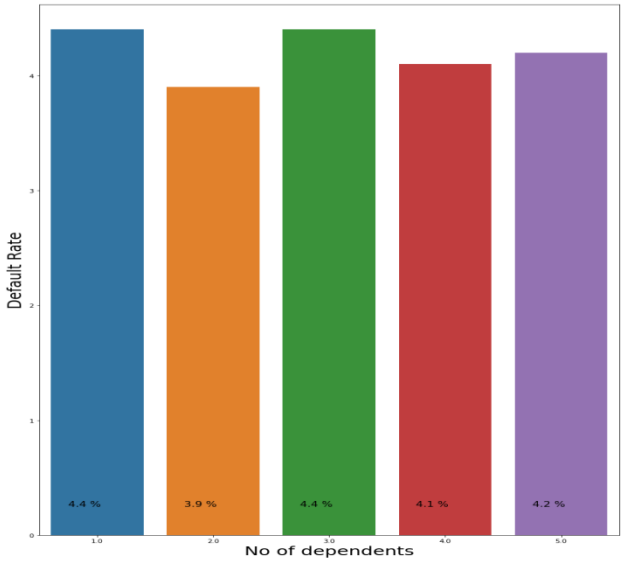


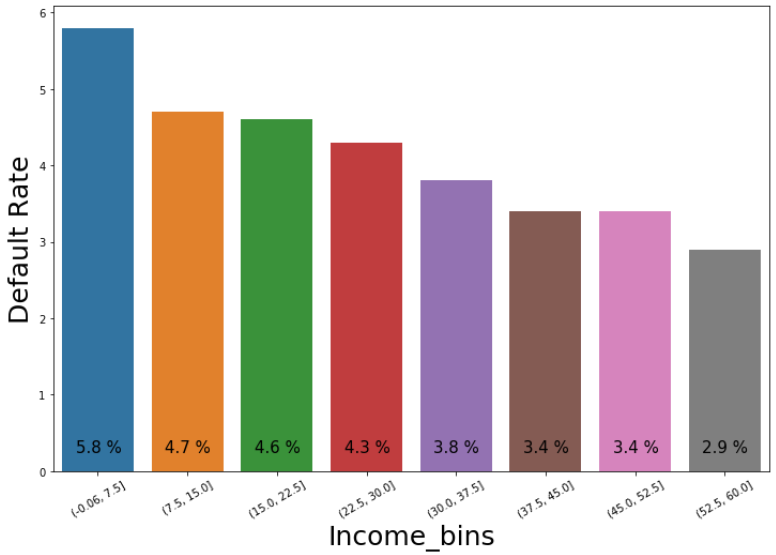


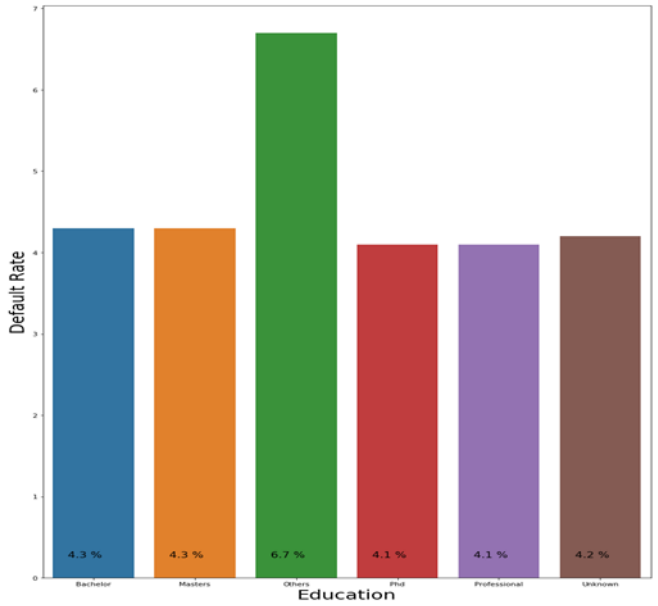


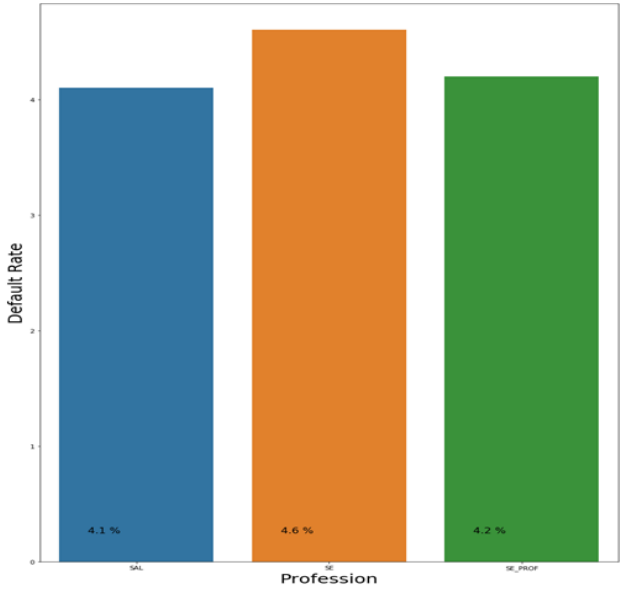
- Socio data of customers which include Age , Gender , Marital status does not show any significance with default rate

EDA – Demographic Dataset-Economic (Bivariate Analysis)



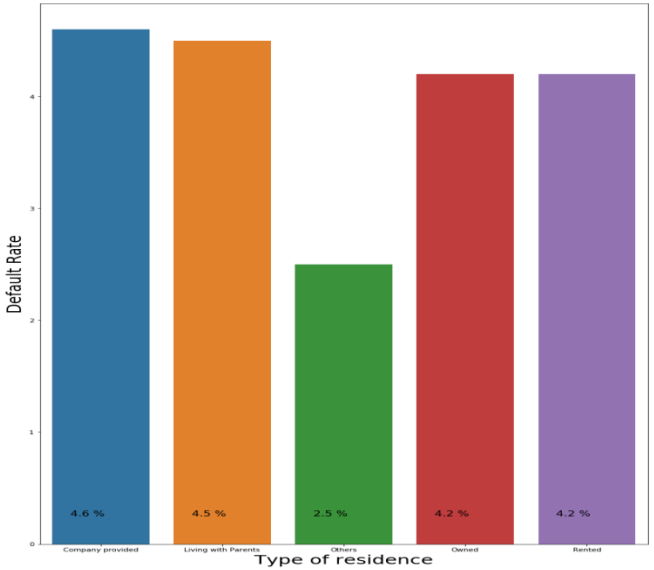


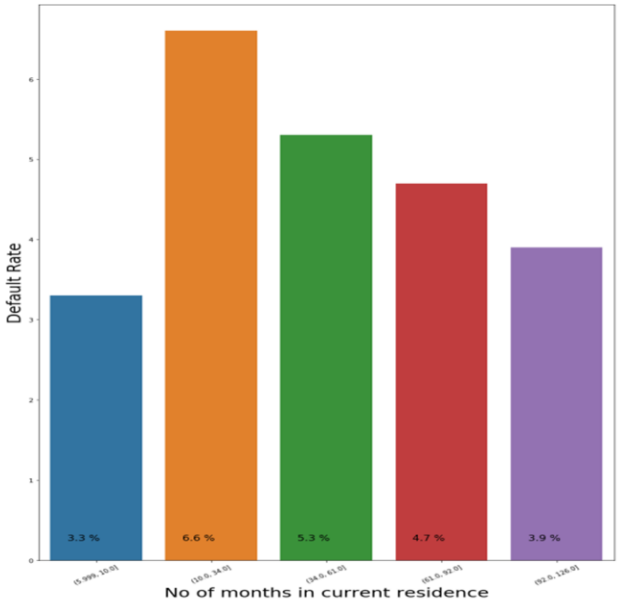


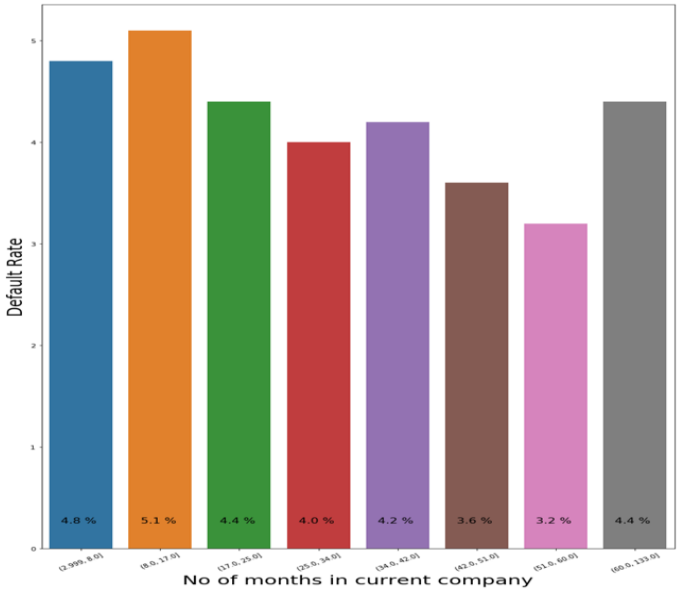


- No. of dependants and Profession does not have much significance
- Less Income has a positive effect on default rate
- Education : Others has effect on default rate

EDA – Demographic Dataset (Bivariate Analysis)

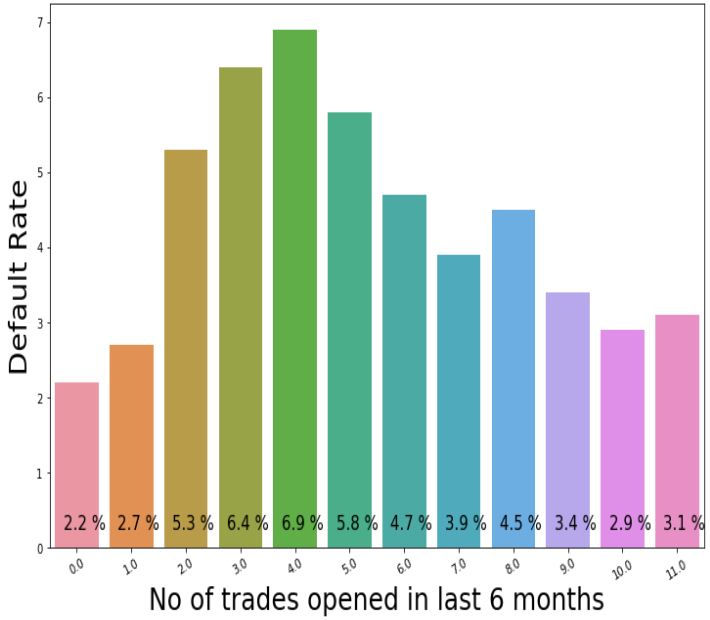


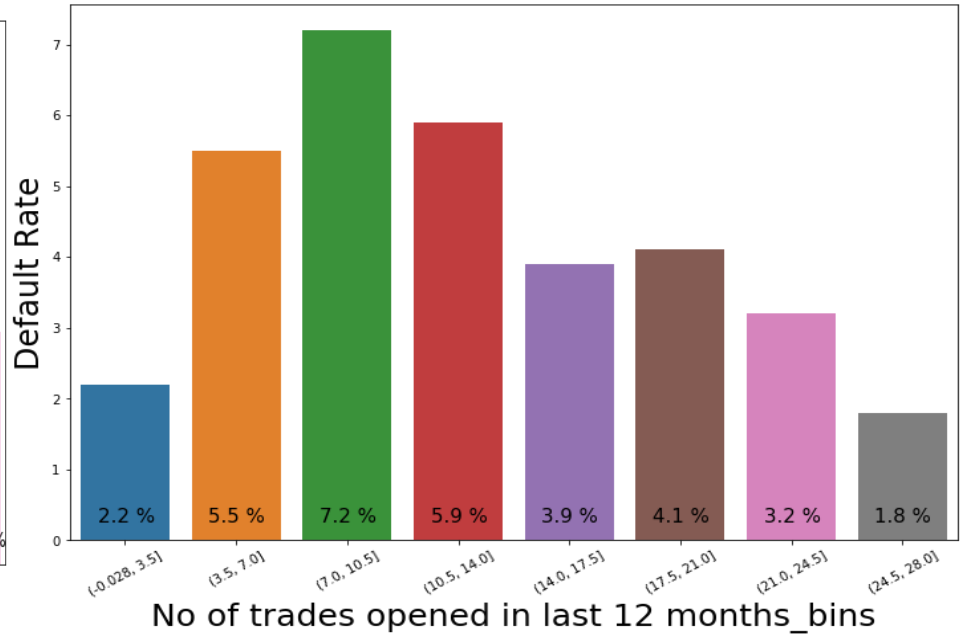


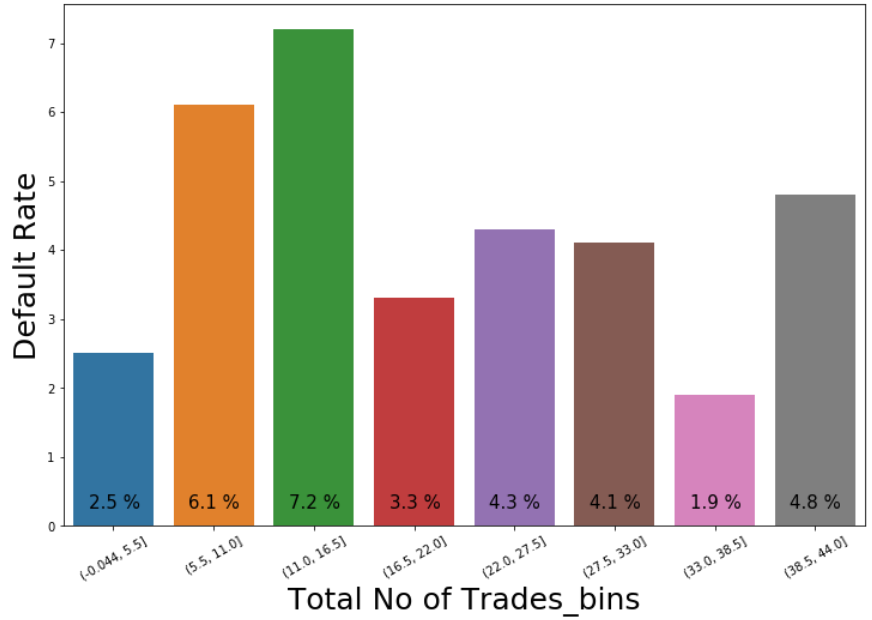


Type of Residence does not show any pattern it's random
No. of months in current residence/company are also having no pattern

EDA – No. of trades opened(Bivariate Analysis)

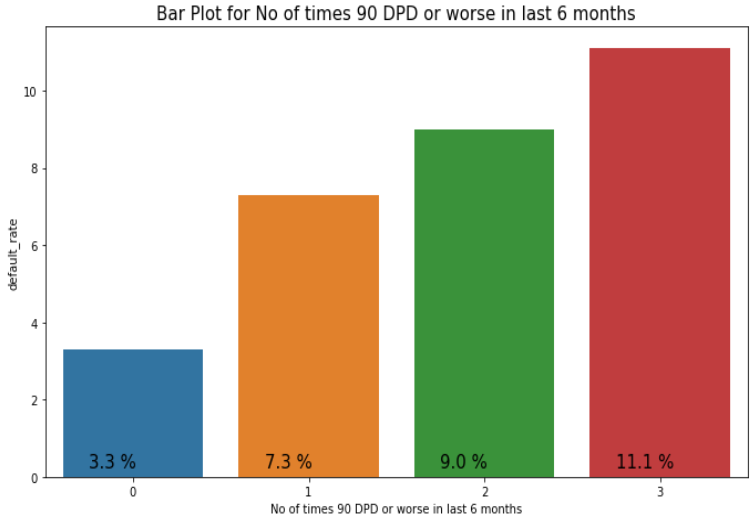


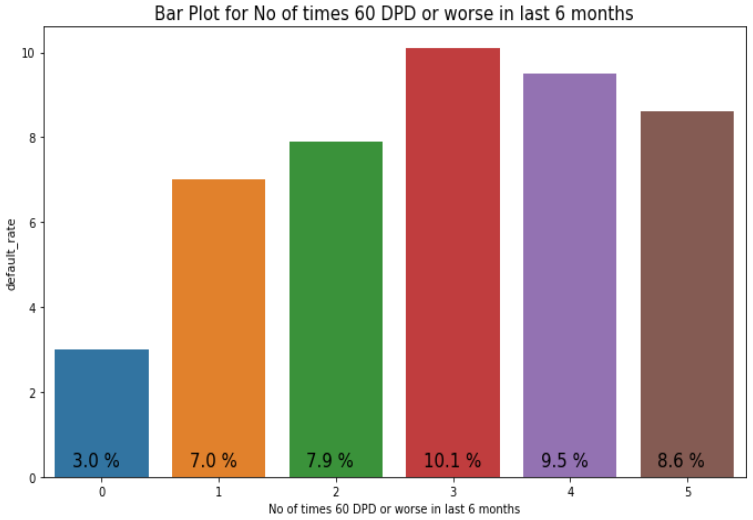


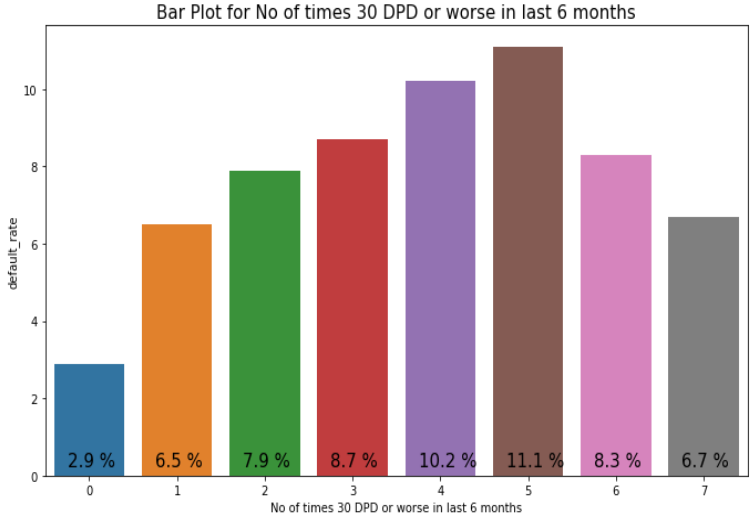


- No of trades opened in last 6 months – There is no clear trend for the default rate.
- No of trades opened in last 12 months - Default rate increases as the no. of trades increases Initially then it starts decreasing with one exception.
- No of trades also follows same as above

EDA – No of times DPD (Bivariate Analysis)

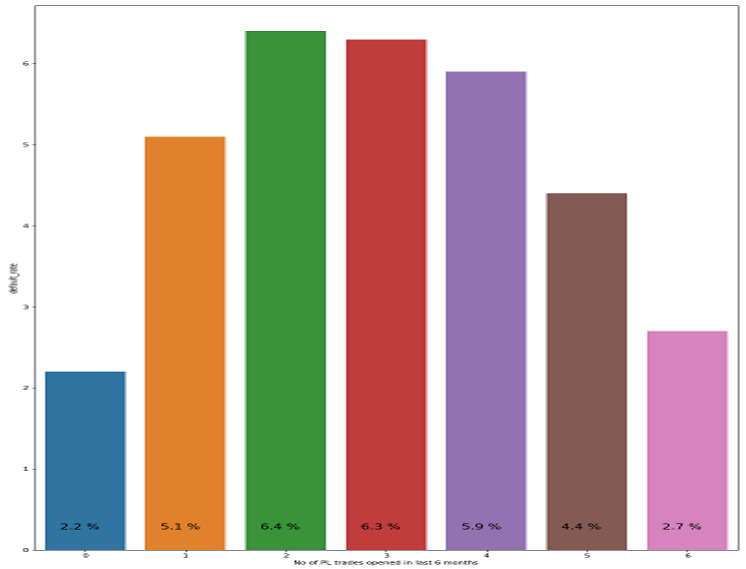


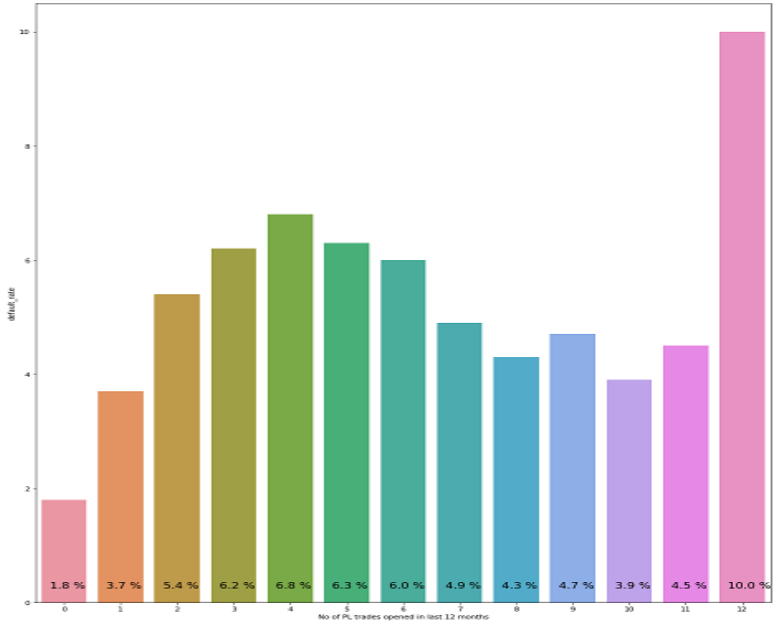




No of times 90/60/30 DPD in last 6 months – General trend is default rate increases as the no. of times 90/60/30 DPD increases.

EDA – No. of PL trades opened(Bivariate Analysis)

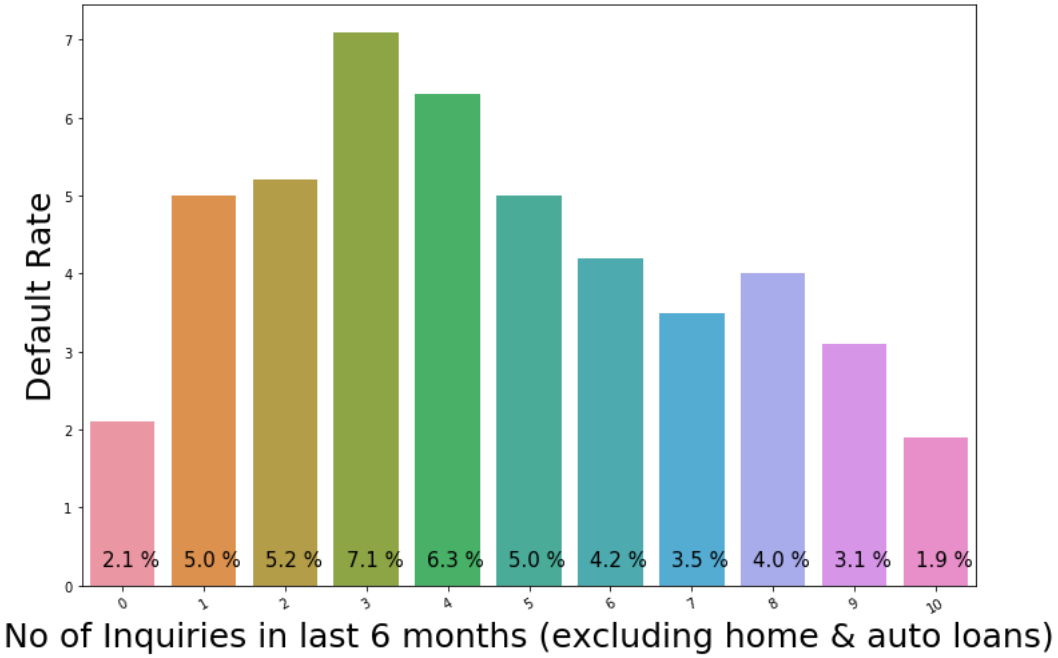


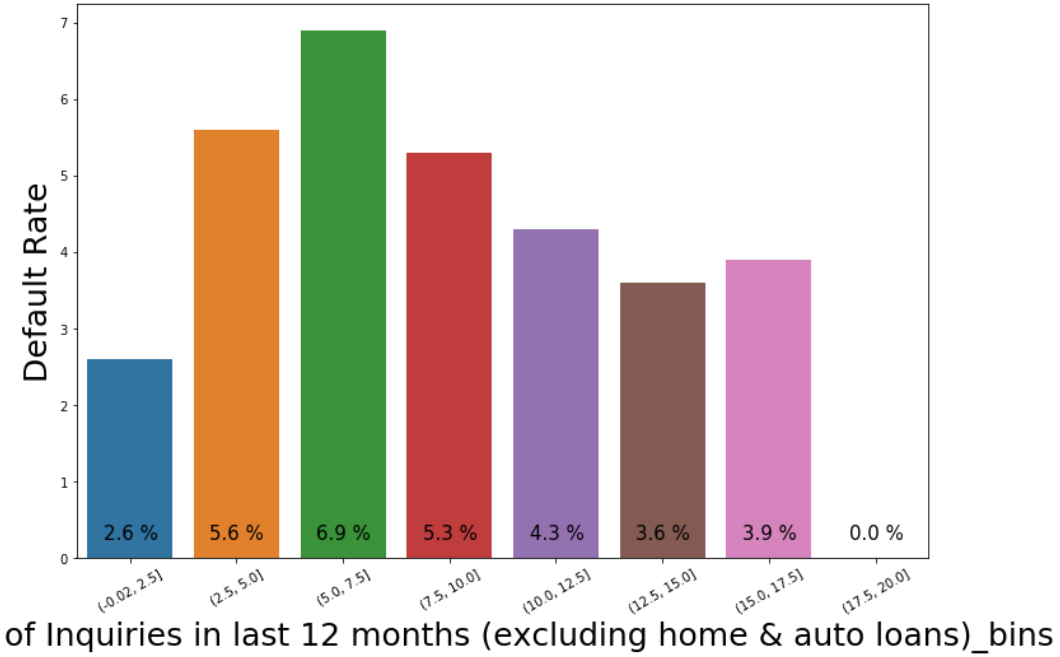


- No of PL trades opened in last 6 months – Default rate increases initially and starts decreasing later.

- No of PL trades opened in last 12 months - Default rate increases initially, it then decreases in the middle and ends up with a spike.

EDA – no of Inquiries(Bivariate Analysis)





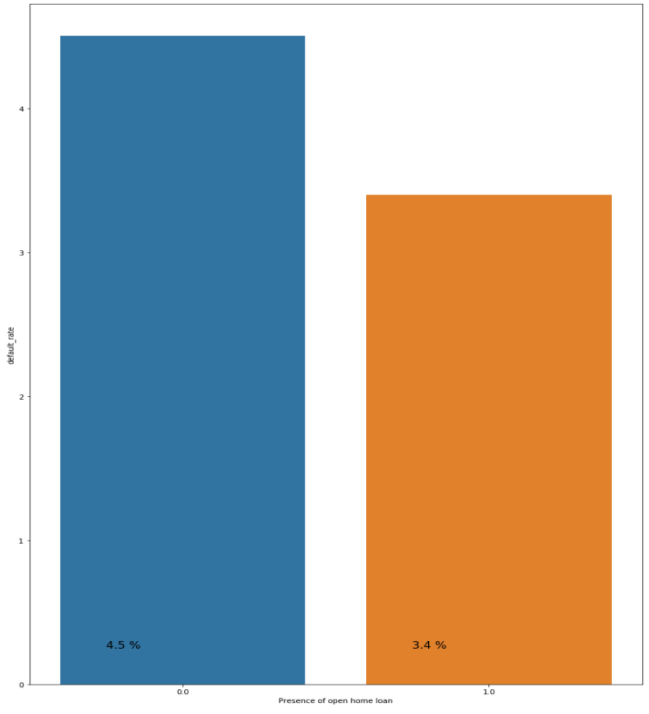
- No of inquiries in last 6 months (excluding home & auto loans) –

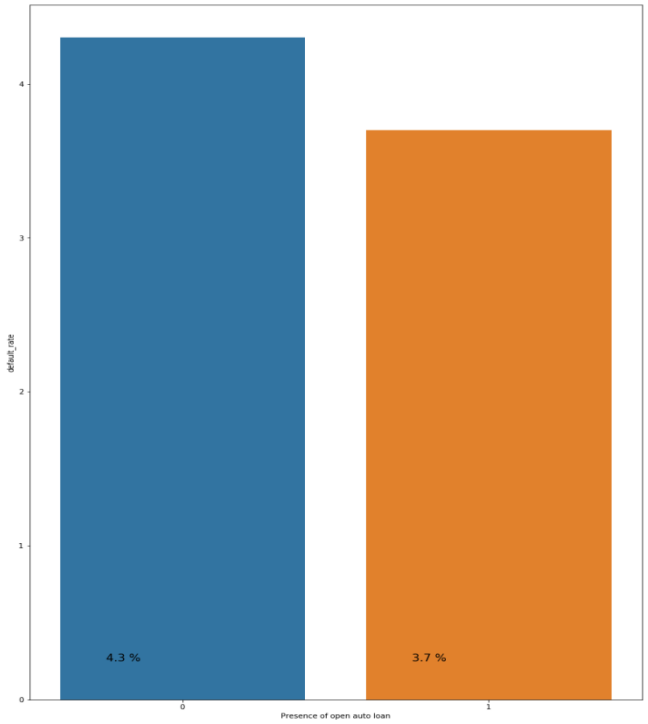
Default rates increases for the values 0 to 2, it then decreases for the higher values.

- No of inquiries in last 12 months(excluding home & auto loans) –

Default rate increases as the value of this variable increase except for the last bin

EDA – Presence of Loans(Bivariate Analysis)



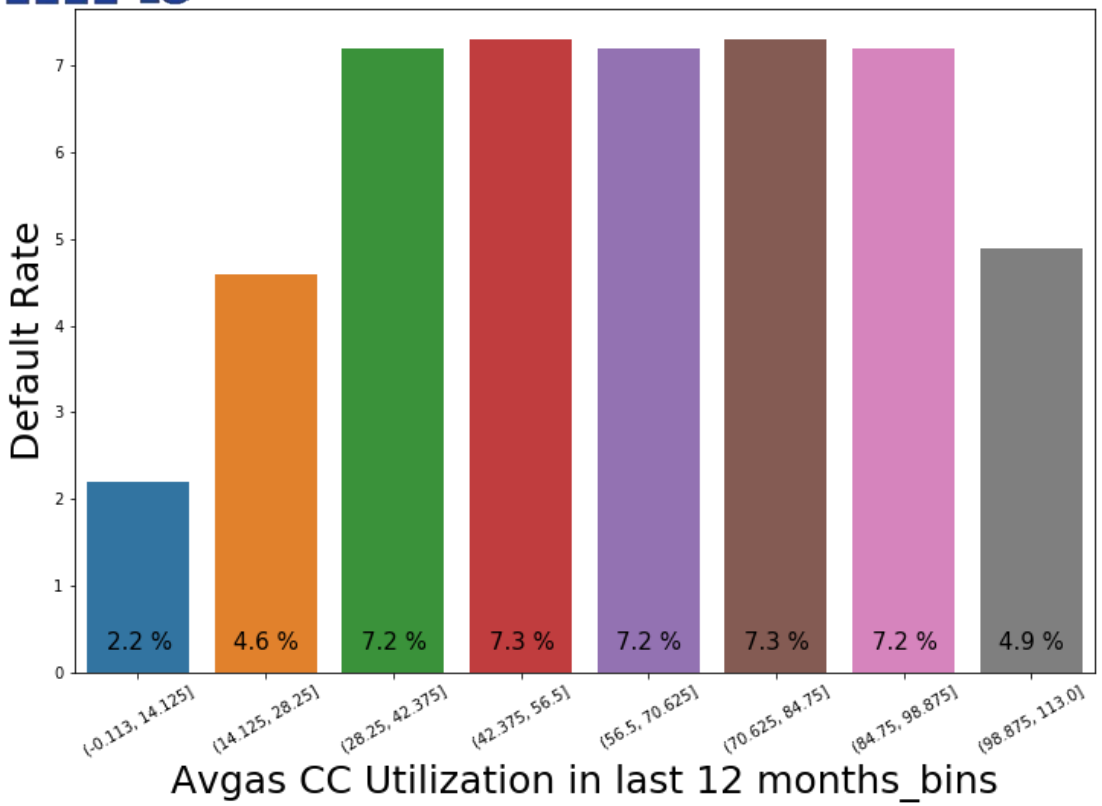


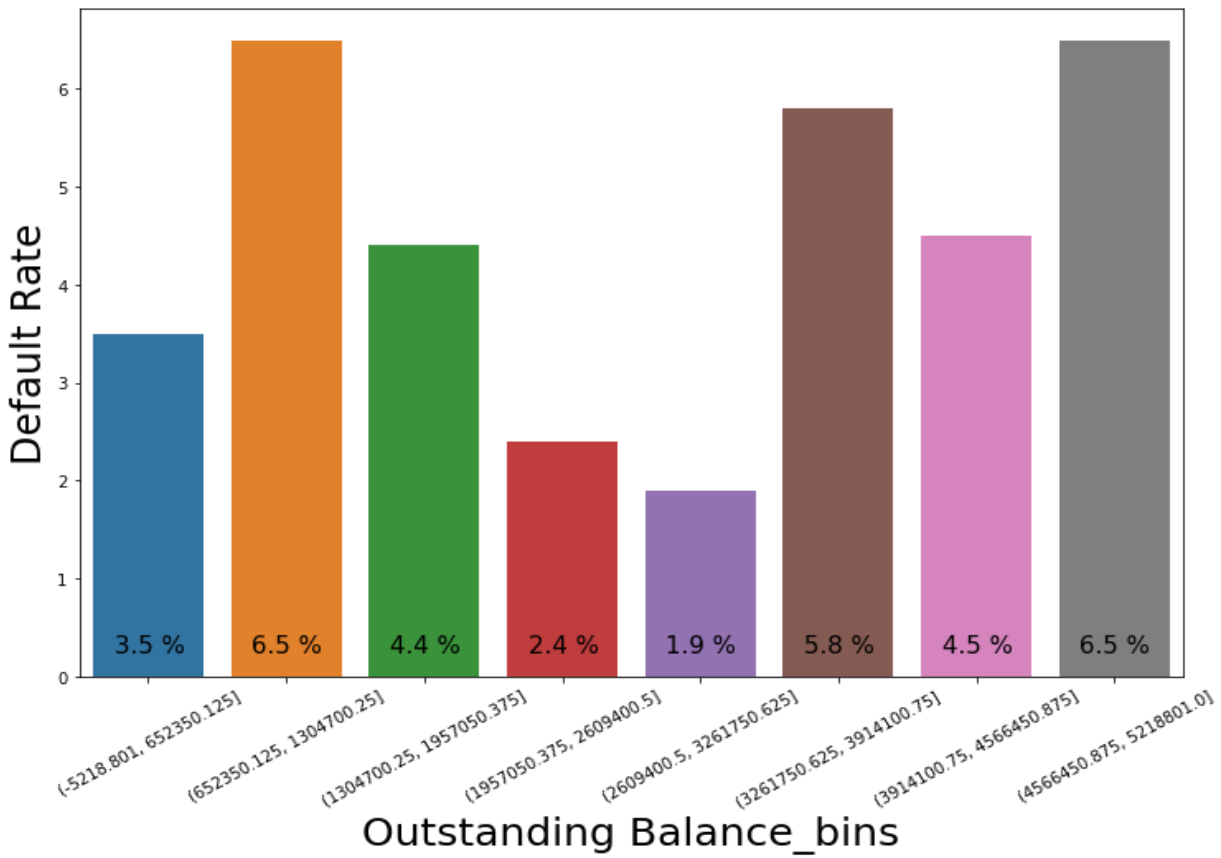
No loan

Loan

Presence of no home/auto loans have more default rate

EDA – Credit card utilization/balance (Bivariate Analysis)





- Credit card utilization follows a pattern, default rate increases with their usage of credit card

- Outstanding Balance : 1) Default rate is the lowest for couple of bins in the middle of the graph
  2) Default rate is maximum for the 2nd & 6th bins
  3) For rest of the bins, default rate varies from 3.5 to 4.5

EDA Results – List of important predictors

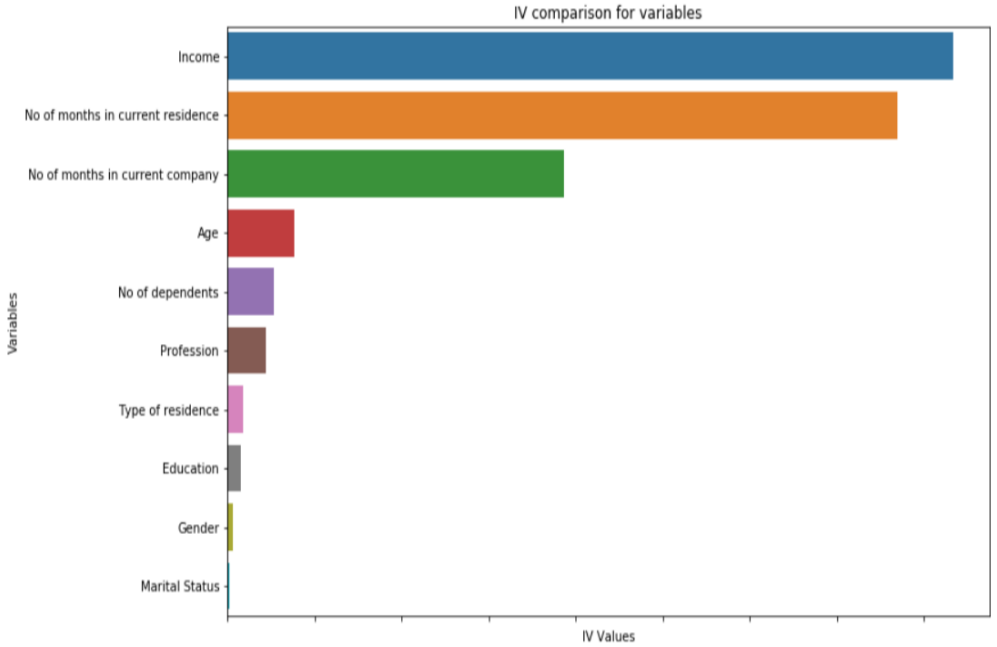
Demographic Dataset

Feature Name	Default Rate
Income	Default rate decreases as the income increases across the bins.
No. of months in current residence	Default rate decreases as the no. of months in the current residence increases.
No. of months in current company	Default rate decreases as the no. of months in the current company increases.

EDA Results – List of important predictors Credit Bureau Dataset

Feature	Default Rate
No of times 90/60/30 DPD or worse in last 6 months	Default rate increases as the no. of times 90/60/30 DPD or worse increases
No of times 90/60/30 DPD or worse in last 12 months	Default rate increases as the no. of times 90/60/30 DPD or worse increases
Avgas CC Utilization in last 12 months	Default rate increases as bin values for Avgas CC Utilization in last 12 months Increases
No of trades opened in last 6/12 months	Default rate increases as the value for No. of trades opened in the last 6/12 month increases
No of PL trades opened in last 6/12 months	Default rate increases as the value for No. of PL trades opened in the last 6/12 month increases
No of Inquiries in last 6/12 months (excluding home & auto loans)	Default rate increases as the value for No of Inquiries in last 6/12 months (excluding home & auto loans) Increases
Presence of open home/auto loan	Default rate is more for those who have not taken home/auto loan
Outstanding Balance	Default rate increases as Outstanding Balance increases
Total No of Trades	Default rate increases as Total No of Trades increases

Results of WOE & IV on Demographic dataset



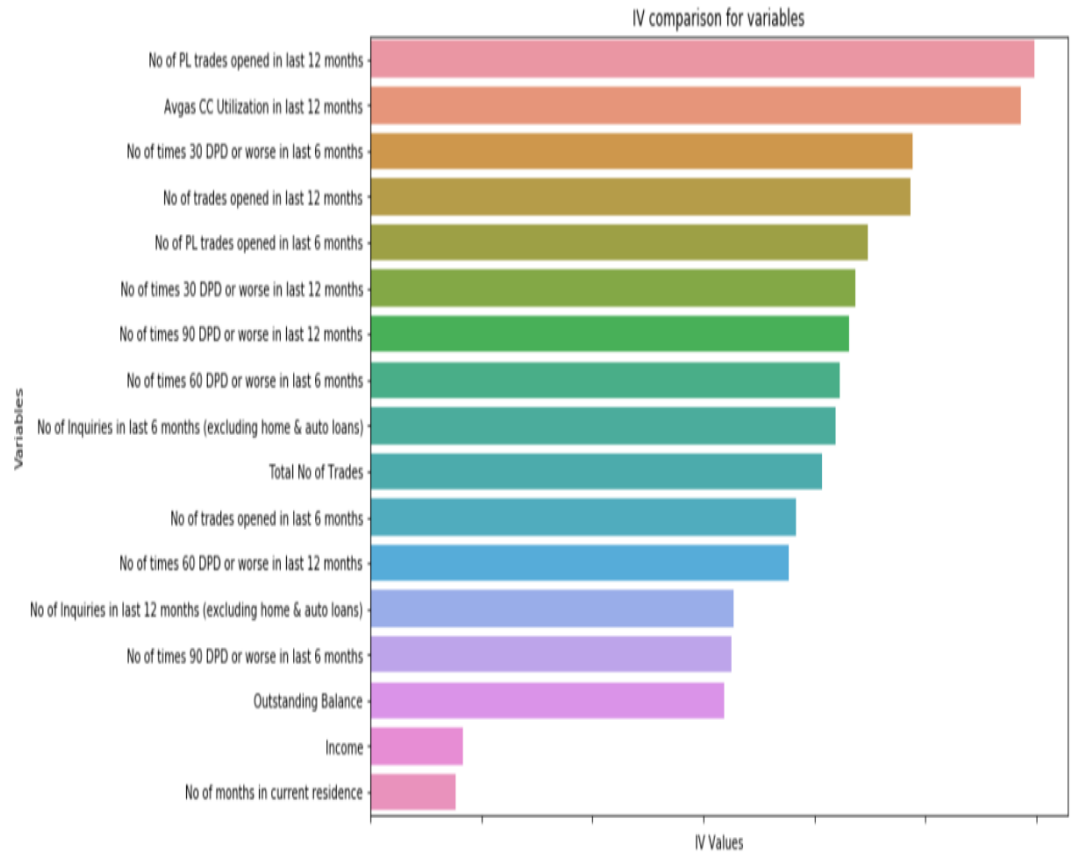
Variable	IV
Income	0.038354
No of months in current residence	0.017329
No of months in current company	0.010054
Age	0.002649
No of dependents	0.002649
Profession	0.002231
Type of residence	0.000921
Education	0.000783
Gender	0.000326
Marital Status	0.000095

- Based on the below predictiveness table, Demographic data set doesn't have variables where IV > 0.1, So not a significant dataset

- We will try to merge the demographic data set with credit bureau data and try to identify the presence of significant variables.

Information Value	Variable Predictiveness
Less than 0.02	*Not useful for prediction*
0.02 to 0.1	*Weak predictive Power*
0.1 to 0.3	*Medium predictive Power*
0.3 to 0.5	*Strong predictive Power*
>0.5	*Suspicious Predictive Power*

Important variables on Master Dataset



Variable name	IV value
No of PL trades opened in last 12 months	0.298981
Avgas CC Utilization in last 12 months	0.292840
No of times 30 DPD or worse in last 6 months	0.244237
No of trades opened in last 12 months	0.243495
No of PL trades opened in last 6 months	0.224242
No of times 30 DPD or worse in last 12 months	0.218599
No of times 90 DPD or worse in last 12 months	0.215644
No of times 60 DPD or worse in last 6 months	0.211263
No of Inquiries in last 6 months (excluding ho...	0.209320
Total No of Trades	0.203444
No of trades opened in last 6 months	0.191498
No of times 60 DPD or worse in last 12 months	0.188225
No of Inquiries in last 12 months (excluding h...	0.163425
No of times 90 DPD or worse in last 6 months	0.162650
Outstanding Balance	0.159489
Income	0.041686
No of months in current residence	0.038477

- These are the overall important variable with IV > 0.1 which we are considering for modelling
- Of all 17 variables, top 15 are from credit bureau, and only 2 are from demographic

Logistic Regression on Demographic data

- As expected from EDA, the model with demographic dataset show very low prediction power

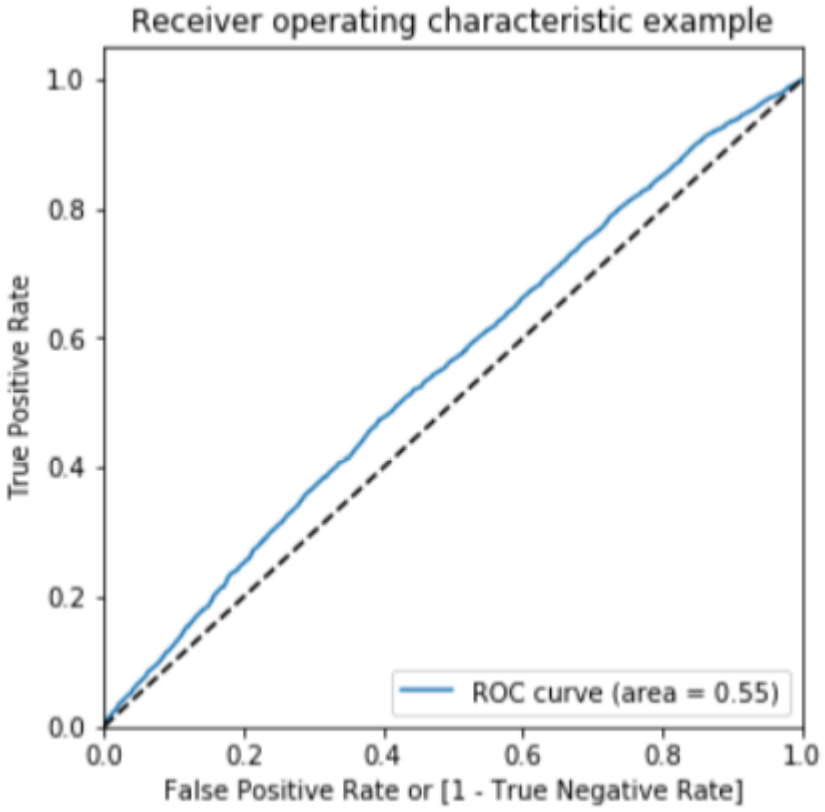
- Confusion Metrics:

  [[39391 7474]
  [1624   417]]

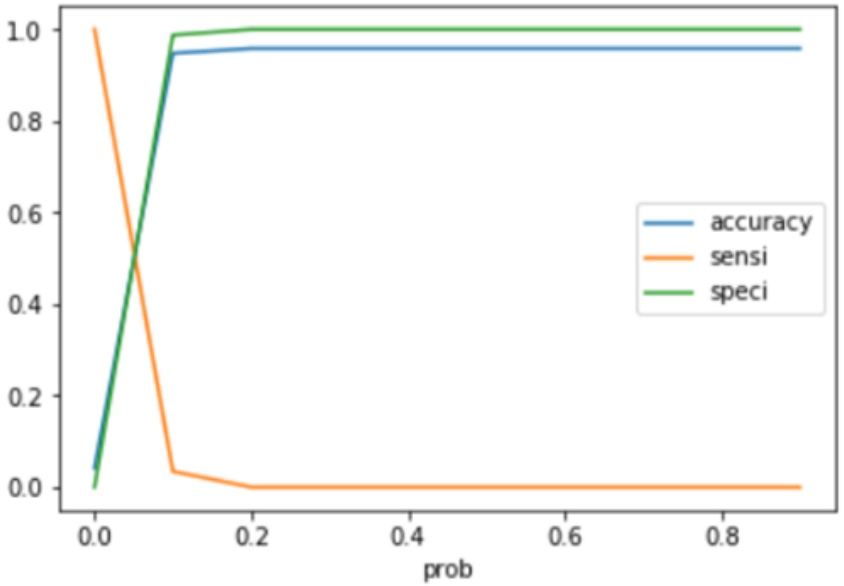
- Accuracy:0.814
- Sensitivity/Recall:0.204
- Precision : 0.0528
- Specificity:0.841
- AUC-ROC : 0.55

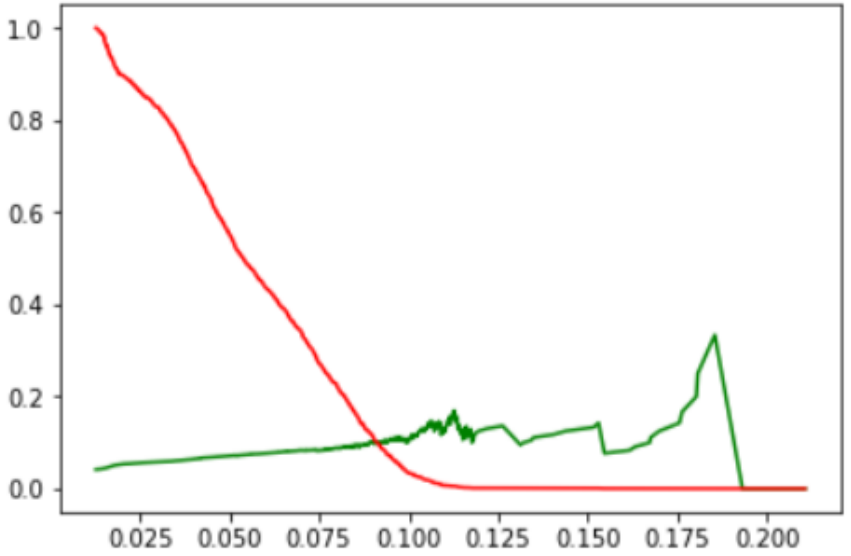
- Sensitivity is very low which is of utmost importance in this project.

Demographic Dataset Results



Optimal cut off values





For Sensitivity/Specificity optimal cut off is : 0.05

For Precision/Recall optimal cut off is : 0.09

Logistic Regression results-master dataset

- Model with master dataset has a greater predictive power

Auc-Roc : 0.67

Train set

- Confusion Metrics:

- [[32341 14524]
- [916 1125]]

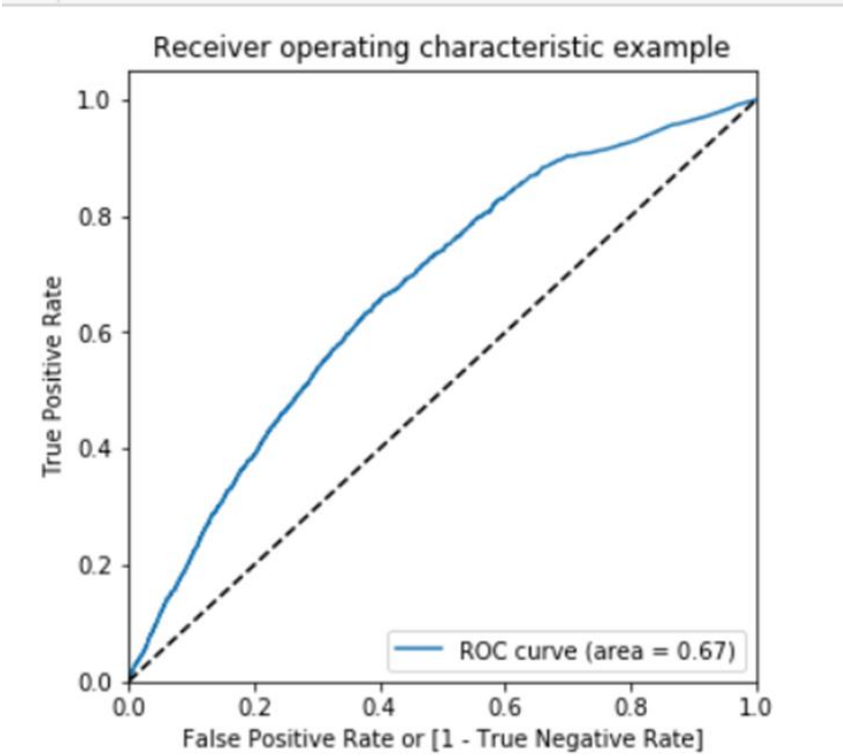
- Accuracy:0.684
- Sensitivity/Recall:0.551
- Precision :0.072
- Specificity:0.69

Test set

- Confusion Metrics:

[[13814 6241]
[397 509]]

- Accuracy:0.683
- Sensitivity/Recall:0.562
- Precision : 0.0754
- Specificity:0.689



Results are improved and same on both train and test which shows it's a **good model**

Regularized Logistic Regression –master dataset

- Best Parameter's for regularization

  ✓ C = 0.001
  ✓ Penality = l1

Train set

- Confusion Metrics:
  [[24337 22528]
  [554 1487]]

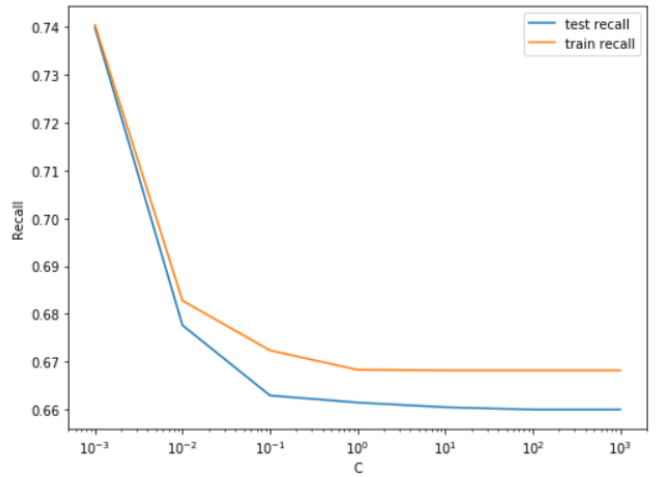
- Accuracy:0.528
- Sensitivity/Recall:0.729
- Precision : 0.0620
- Specificity:0.519

Test Set

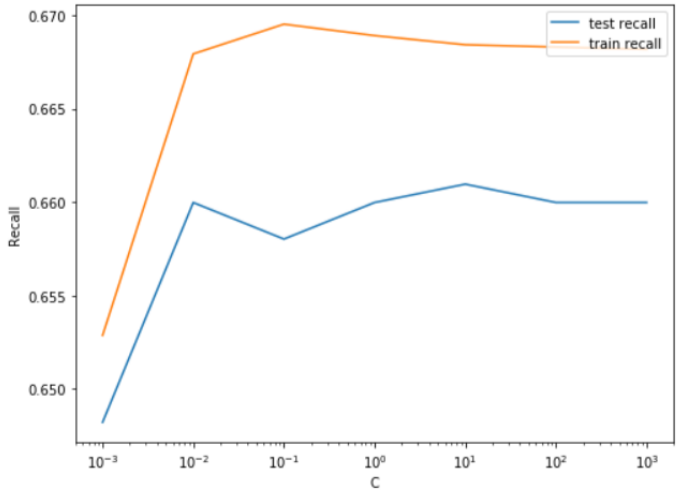
Confusion Metrics:
[[10337 9718]
[235 671]]

Accuracy:0.525
Sensitivity/Recall:0.741
Precision : 0.0645
Specificity:0.515

L1 Penalty



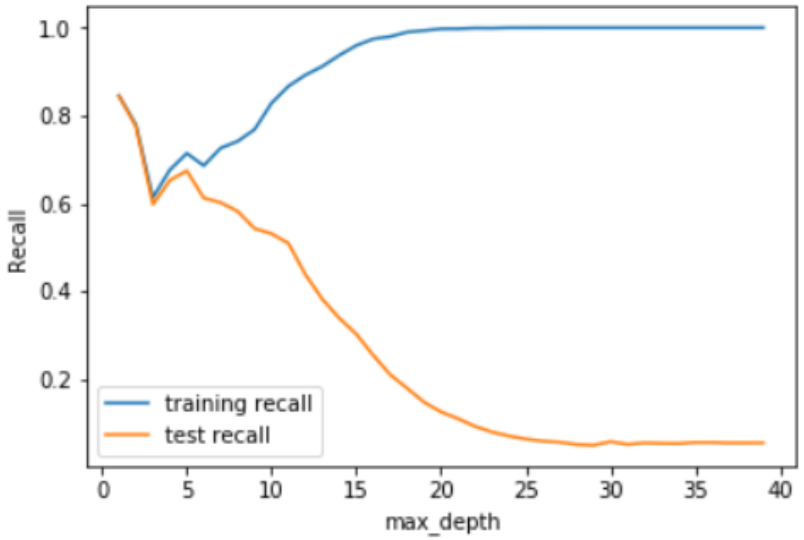
L2 Penalty

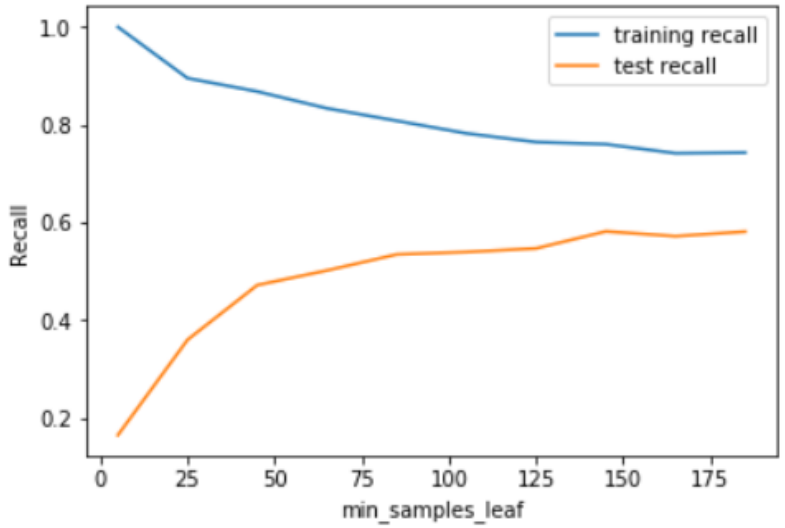


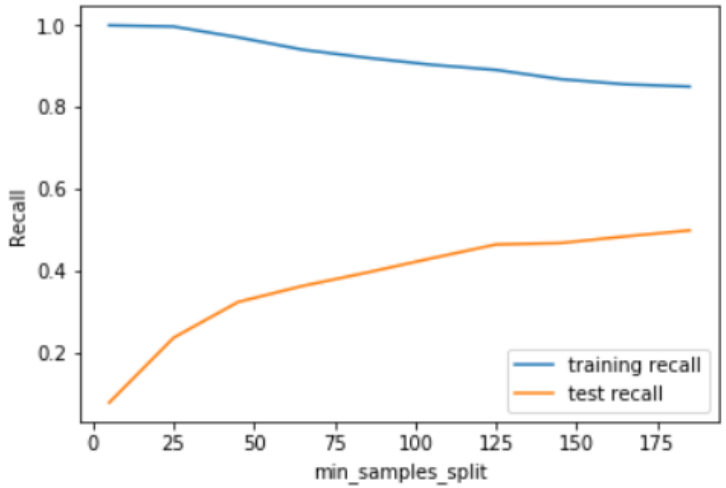
Decision Tree – master dataset

Best Hyper Parameters

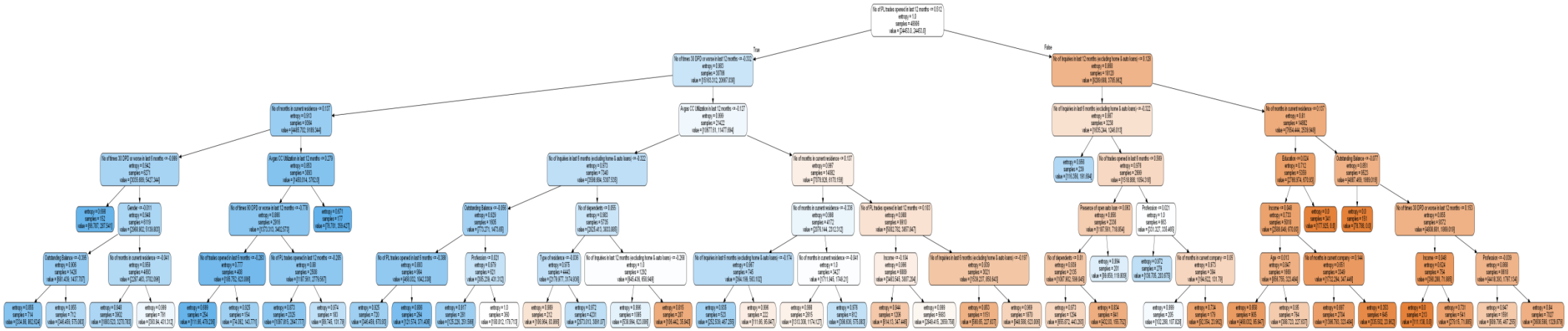
- Class weight=balanced
- Criterion='entropy',
- Max depth=6,
- Min samples leaf=150,
- Min samples split=100,







Decision Tree – master dataset



Train Set:

Confusion Metrics:

[[26992 19873]

[938 1103]]

Accuracy:0.574

Recall/Sensitivity: 0.5404

Precision:0.0525

Specificity:0.576

Test Set:

Confusion Metrics:

[[11315 8740]

[449 457]]

Accuracy:0.562

Recall/Sensitivity: 0.5044

Precision: 0.0496

Specificity:0.564

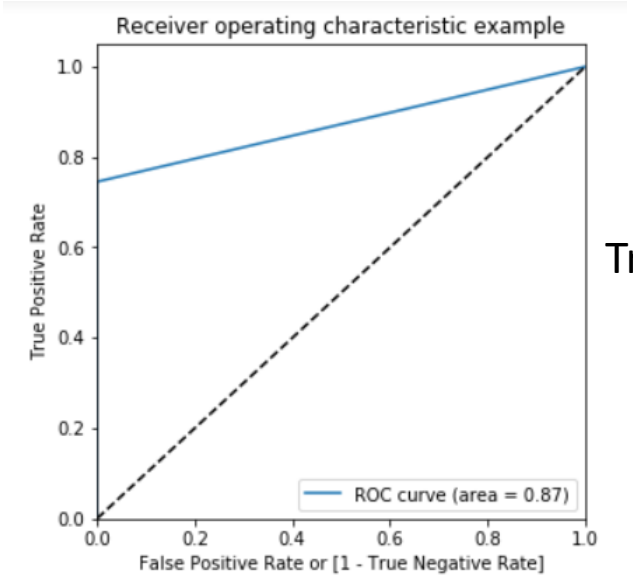
Random Forest(master dataset)-Default parameters

Train set:
Confusion Metrics:
[[46857 8]
[520 1521]]

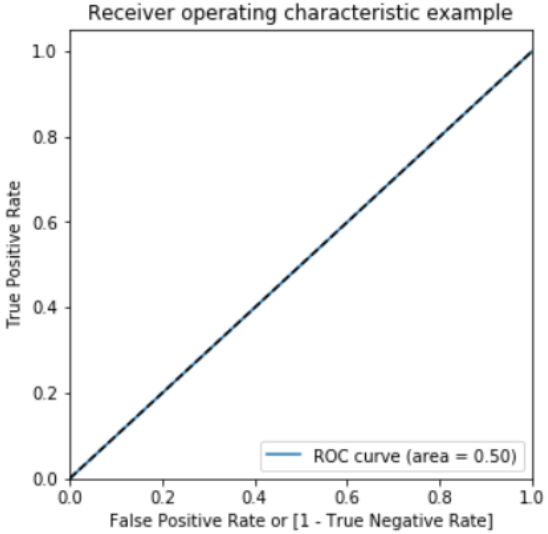
Accuracy:0.989
Sensitivity/Recall:0.745
Precision : 0.994
Specificity:1.0
Auc: 0.87

Test set:
Confusion Metrics:
[[20050 5]
[906 0]]

Accuracy:0.957
Sensitivity/Recall:0.0
Precision : 0.0
Specificity:1.0
Auc: 0.50



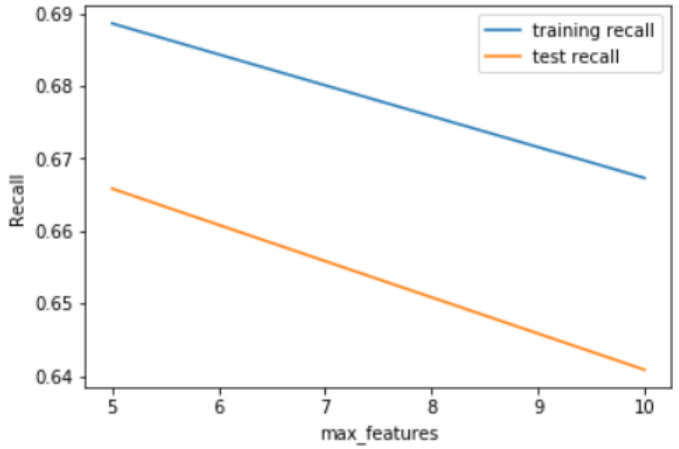
Train AUC

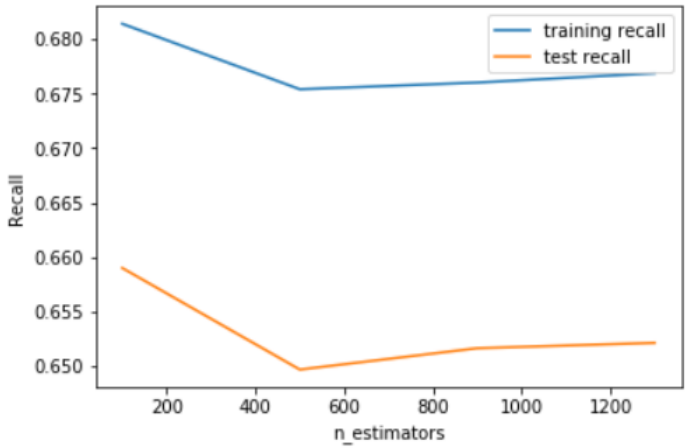


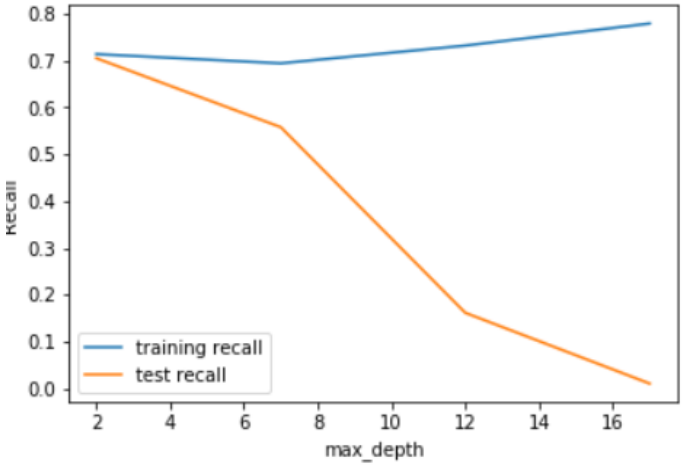
Test AUC

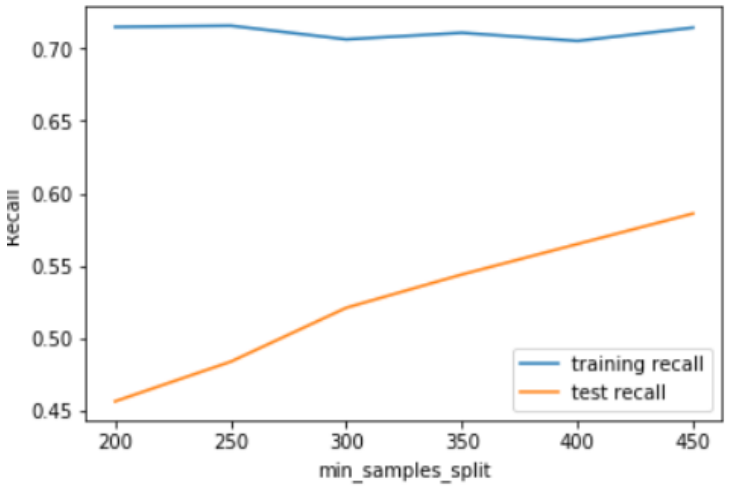
Here we can clearly see test set results are not good as train set, which clearly shows an **overfitting** problem with default parameters

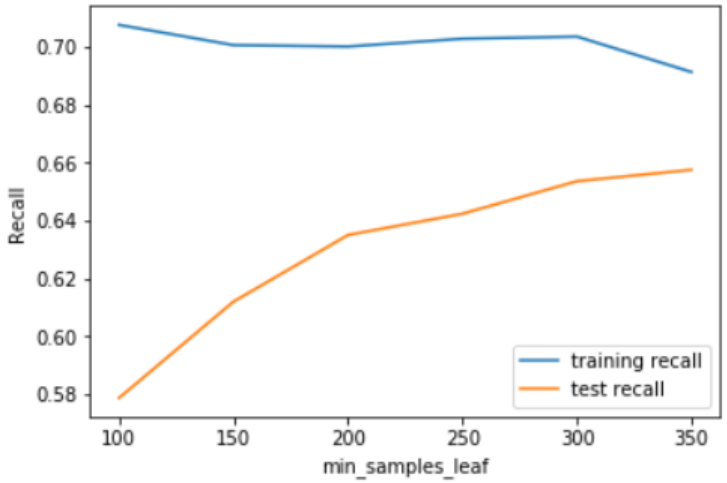
Optimal parameter tuning











Max features=10
N-estimators=200
Max depth=4
Min samples split=400
Min samples leaf=300

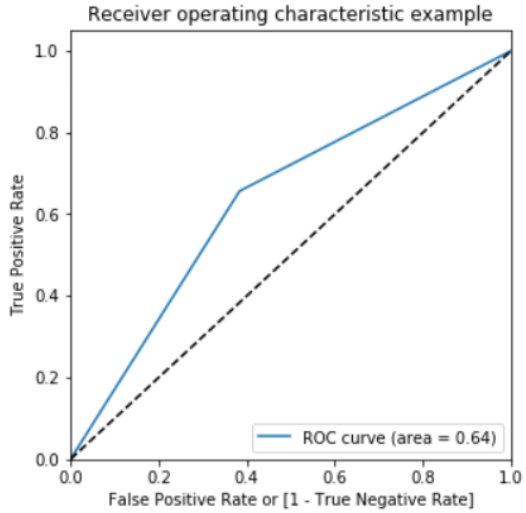
Random Forest Results –master dataset

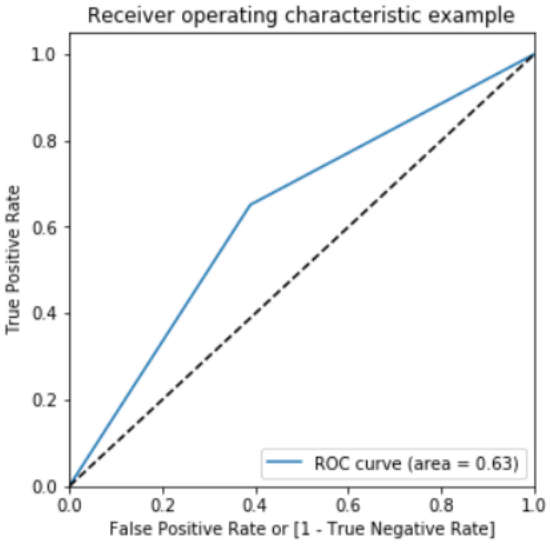
Train set

- ✓ Auc-Roc : 0.64
- ✓ Confusion Metrics:
  [[28648 18217]
  [681 1360]]
- ✓ Accuracy:0.614
- ✓ Sensitivity/Recall:0.666
- ✓ Precision : 0.0694
- ✓ Specificity:0.611

Test set

- ✓ Auc-Roc: 0.63
- ✓ Confusion Metrics:
  [[12137 7918]
  [307 599]]
- ✓ Accuracy:0.608
- ✓ Sensitivity/Recall:0.66
- ✓ Precision : 0.0703
- ✓ Specificity:0.605





Though the results are improved with hyper parameter tuning, logistic regression achieved better results than Random forest and Decision Tree

Summary of all the models built

Model Type	Training Data			Test Data			Rejected Applicants		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Logistic Regression	81.40%	5.28%	20.40%	80.70%	4.70%	18.30%	37.60%	100%	37.60%
Logistic Regression With Regularization	38.30%	3.67%	54.70%	38.70%	3.67%	52.20%	56.10%	100%	56.10%
Decision Tree	57.40%	5%	54.00%	56.00%	5%	49.00%	74.00%	100%	74.00%
Random Forest	62.00%	6.00%	54.00%	61.00%	5.20%	47.00%	72.00%	100%	72.00%
Logistic Regression	68%	7.10%	55.10%	68.30%	7.50%	56.00%	1.10%	100%	1.10%
Logistic Regression With Regularization	53.00%	6.10%	**73.00%**	52.50%	6.50%	**74.1**	99.60%	100%	**99.60%**
Decision Tree	64.00%	7%	65.00%	63%	7%	62%	99%	100%	99.00%
Random Forest	61%	7.00%	66.00%	61.00%	66%	7%	99%	1%	99%

Final Conclusion

- **Best Model (Regularized Logistic Regression on Master Dataset)**

Since Regularized Logistic Regression has produced the best results, we would go ahead with this model to create the scorecard and to calculate the financial benefits of the model.

Building application Score card

- Build an application scorecard with the good to bad odds of 10 to 1 at a score of 400 doubling every 20 points.

Inputs given:

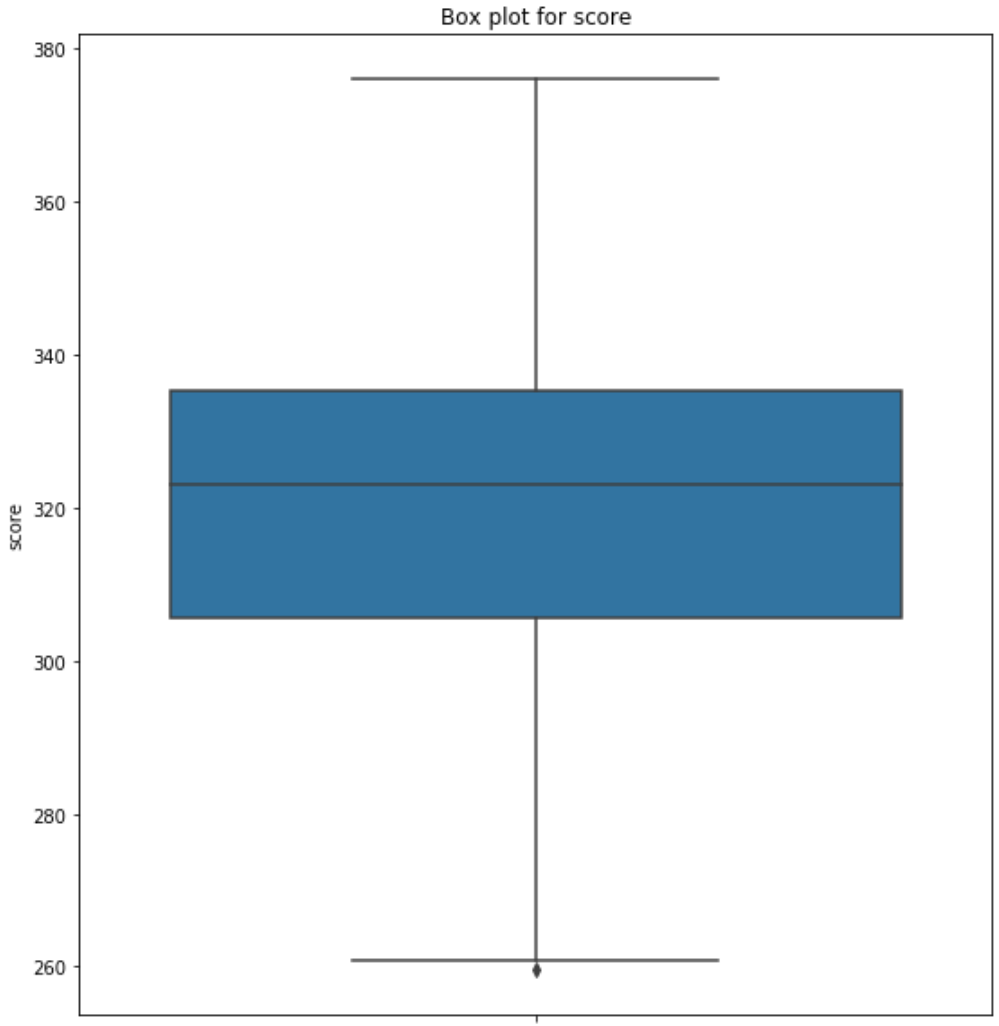
- 1.target_score = 400
- 2.target_odds = 10
- 3.pts_double_odds = 20

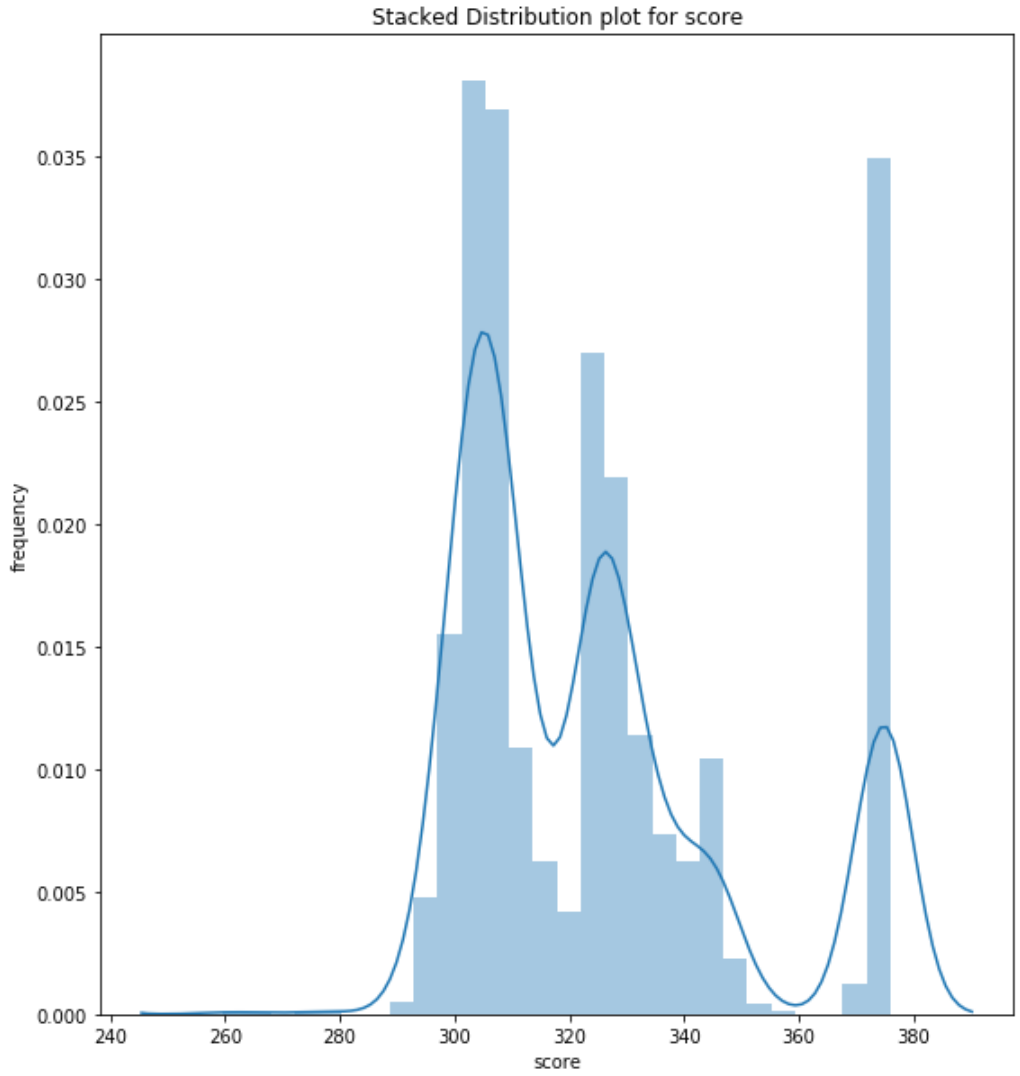
Calculation as follows:

- factor = pts_double_odds / log10(2)
- offset = target_score - factor X log10(target_odds)

After deriving the score card, the dataset of rejected applications (with performance tag missing), which were assumed as potential defaulters are compared with those of the approved ones.Ideally, the output for all these applications should be defaulters.

Application Score Card – Box plot & Stacked plot





Majority of the customers falls in the range of 290 to 350

Application Score Card – Recommended cutoff

- ✓ 1.Customers with a score less than 310 would not be granted credit card.
- ✓ 2.Cutoff of 310 correctly identifies almost 89% of the bad customers.
- ✓ 3.If we consider the scorecard built for the master dataset, then almost 21% of the good customers are not going to get the credit card.
- ✓ 4. If we reduce the cutoff from 310 to a lower number then it will defeat the purpose of doing this exercise of identifying the bad customers.
- ✓ 5. Though, if Bank is ready to take the risk they may reduce the cutoff by 5 points, keeping it to 305. A cutoff of 305 would correctly identify 76% of the bad customers, and will impact around 2.5% good customers.

Financial Benefits of the model

As mentioned in the problem statement, in the past few years CredX experienced an increase in credit loss. So, the main objective of doing this whole exercise was to mitigate this credit risk by acquiring the right customers.

Another important point is, Bank does not only loose money by giving credit card to the bad customers, it may also loose business(eventually money) by not giving credit cart to the good customers. So, the Machine Learning model should have strong predictive power to discriminate between good and the bad customers. Model should be be to correctly identify majority of the bad customers, at the same time, it should also ensure that good customers are not denied the credit card.

A good model will have following benefits:

- ✓ 1.Saves manual efforts of assessing each and every application as the model can process hundreds of applications in no time.
- ✓ 2.Prevents manual error as the whole process is automated.
- ✓ 3.No chance of underwriters taking bribes to approve an application, also model would not be biased towards any cast, religion etc.

Total number of Customers = 71292 (remember we removed three duplicate reports)

Approved Customers = 69867 (there were 1425 records with null values for performance tag, 71292-1425 = 69867)

Default Customers = 2947 (Customers with Performance Tag 1)

Financial Benefits of the model –contd..

Lets make some assumptions in order to calculate the actual profit and loss :

1) Customer Acquisition Cost (including paper work, phone calls cost, service tax etc.) - 50 USD

2) Credit Card Limit = 49,950 USD (taking odd number so that the money at risk is a round figure)

3) Money at Risk per customer = 49,950 + 50 = 50,000 USD

Total Money at Risk (Defaulted Customers) = 50,000 × 2947 = 14,73,50,000 USD

Money Machine Model can save: The best Model we built has a recall of 74%, hence it can save 74% of 14,73,50,000 USD:

Money Saved = (14,73,50,000 * 74)/100 = 10,90,39,000 USD

Money Lost = 14,73,50,000 USD - 10,90,39,000 USD = 3,83,11,000 USD

Model built is saving almost 35% of the Loss