BFS CAPSTONE PROJECT

FINAL- SUBMISSION

By,
Rohit Tater
Sowmya Sangeetham

Abstract

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.

Steps to Problem Solving

- ➢ Understand the underlying problem and the domain for the problem.
- ➢ Understand the dataset provided (both Demographic and Credit Bureau) and inspect each attribute of both.

- ➢ Model 1 – Building Model on Demographic Data :-
- ✓ Data importing & Data Understanding
- ✓ Data Cleaning
- ✓ Exploratory Data Analysis
- ✓ Data Preparation
- ✓ Test-Train Split
- ✓ Standardization
- ✓ Model Building
- ✓ Perform WOE on Original Demographic Dataset
- ✓ Model Building on WOE Dataset

Steps to Problem Solving (continued…)

- Model 2 – Building Model on Demographic Data + Credit Data :-
- ✓ Data importing & Data Understanding
- ✓ Data Cleaning
- ✓ Exploratory Data Analysis
- ✓ Data Preparation
- ✓ Test-Train Split
- ✓ Standardization
- ✓ Model Building
- ✓ Perform WOE on Original Master Dataset (Demographic + Credit)
- ✓ Model Building on WOE Dataset

- The apt two stable and optimized models (with stable characteristics) – one for demographics and second for combined data needs to be chosen.
- On the basis of the chosen model and significant variables in the model, two application scorecard should be prepared for the two models.
- Access the financial benefits of the project by checking the underlying metrics that get optimized.
- Present all the results obtained in all the above steps to the management.

Data Understanding

There are two data sets in this project : **Demographic** and **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:** This is taken from the credit bureau and contains variables such as 'number of times 30DPD 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, it is observed that all the variables in the credit bureau data are zero and credit card utilization is missing. These represent cases in which there is a no-hit in the credit bureau. The cases with missing credit card utilization are also observed. These are the cases in which the applicant does not have any other credit card.

Building Model on Demographic Data

Model Building on Demographic Data includes the below steps:

- Data Importing & Data Understanding
- Data Cleaning
- Exploratory Data Analysis
- Data Preparation
- Test-Train Split
- Standardization
- Modelling

Performing Data Cleaning on Demographic Data

- **Remove Duplicate Rows:**

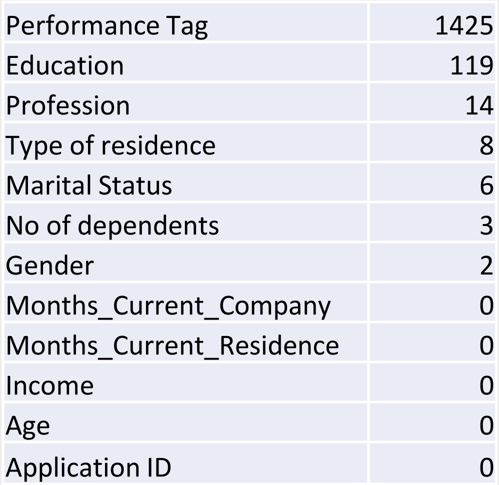
Total Number of records before dropping duplicate records :- 71295

Total Number of records after dropping duplicate records :- 71292

- **Column-wise Missing Values:**

➢ We can drop the Performance Tag variable as this is the dependent variable.

➢ Number of records having missing value in columns like Education, Profession, Type of residence, Marital Status, No. of dependents and Gender is very small compared to total number of records , hence we can drop these records as well.



Performance Tag	1425
Education	119
Profession	14
Type of residence	8
Marital Status	6
No of dependents	3
Gender	2
Months_Current_Company	0
Months_Current_Residence	0
Income	0
Age	0
Application ID	0

Percentage of Records left after Data Cleaning = 69718/71295 =97.78%

Due to Data Cleaning around 2.2% data is lost which is not much significant

Exploratory Data Analysis on Demographic Data

We will do Exploratory Data Analysis on individual features and look or patterns.

1. Univariate Analysis
   - Age : Age of Customer
   - Gender: Gender of customer
   - Marital Status: Marital status of customer (at the time of application)
   - No. of dependents: No. of children's of customers
   - Income: Income of customers
   - Education: Education of customers
   - Profession: Profession of customers
   - Type of residence: Type of residence of customers
   - No. of months in current residence: No. of months in current residence of customers
   - No. of months in current company: No. of months in current company of customers

2. Multivariate Analysis
   - Heatmap

Univariate Analysis on Demographic Data

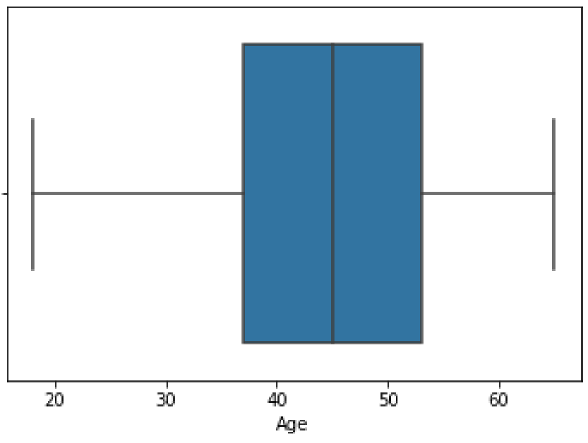
a. **Age:**

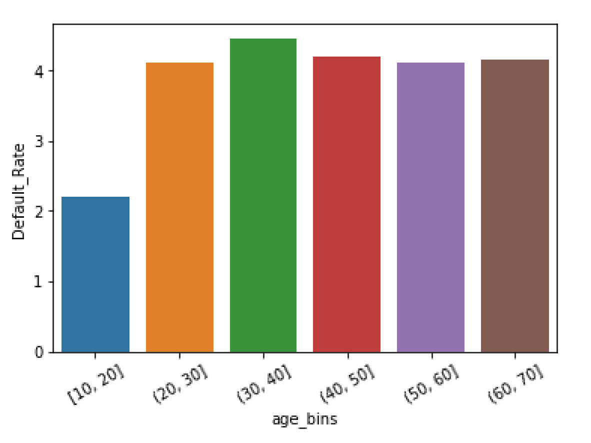
Since there are a fair few data entries with age < 18, we will club all of those ages at 18 .

Analysis of Impact of Age:

- People in age group 10-20 have lowest default rate but total number of records in that age group is very low.
- Default rate of People in age group 30-40 is highest however the difference is not significantly high.

There does not seem to be significant relation between default rate and Age of Applicant





Univariate Analysis on Demographic Data (continued…)

b. Gender:

Analysis of Impact of Gender:

- Default rate of Female is higher than Male, but the difference is not significantly high.

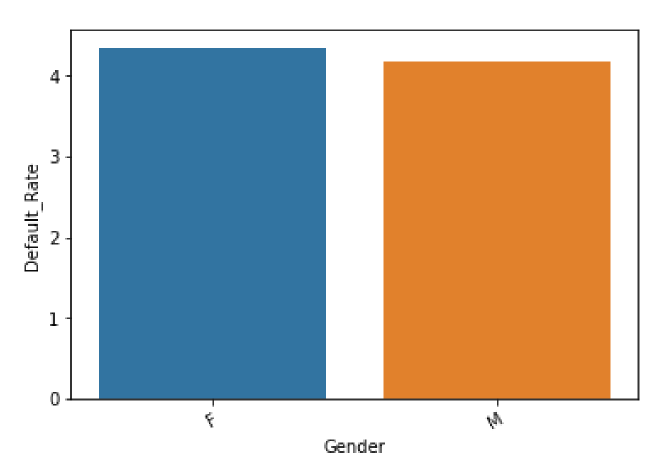
Thus, there does not seem to be significant relation between default rate and Gender of Applicant

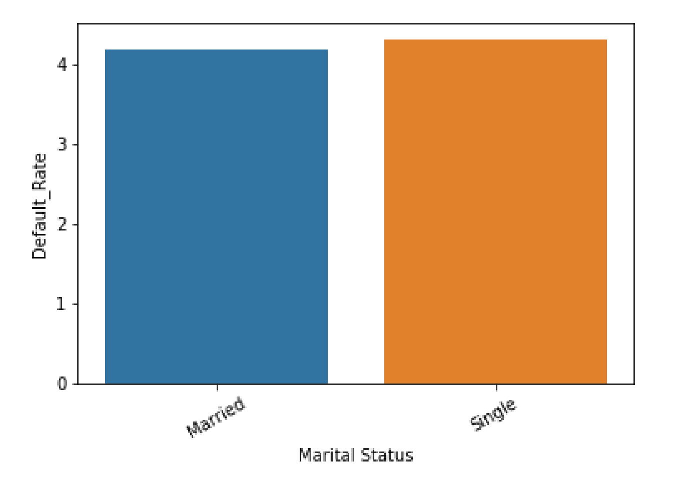
c. Marital Status:

Analysis of Impact of Marital Status:

- Default Rate of Single person is slightly higher than married person, but the difference is not significantly high.

Thus, there does not seem to be significant relation between default rate and Marital Status of Applicant





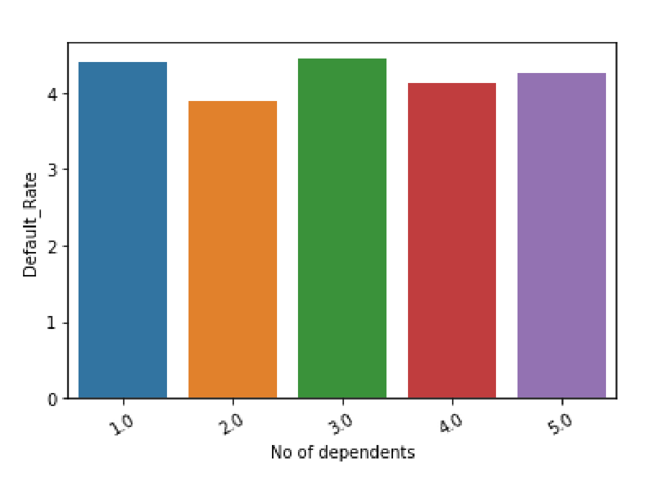
Univariate Analysis on Demographic Data (continued…)

d. No. of dependents:

Analysis of Impact of No. of dependents:

- Applicants with 2 dependent are less likely to default.
- Applicants with 3 dependent are more likely to default followed by applicants with 4 default.

Thus, there is no significant relation between Default Rate and No. of Dependents of Applicant.

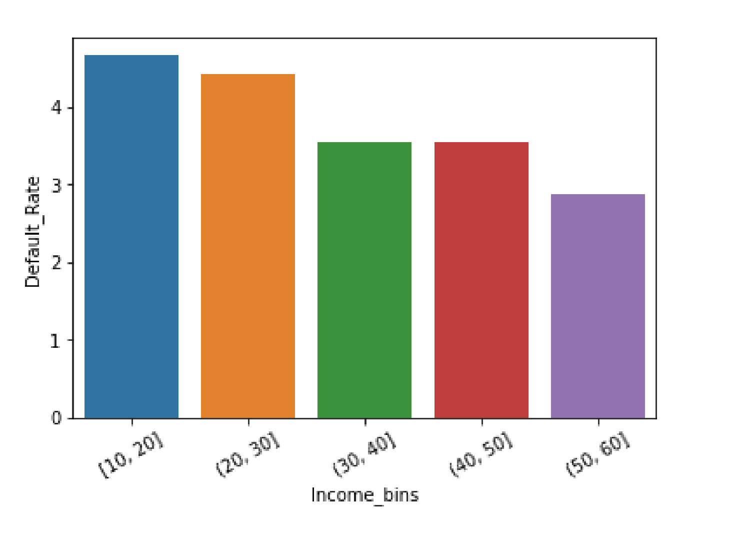


e. Income:

Analysis of Impact of Income:

- Default rate of applicants with low income is higher than the applicants with high income.
- This is obvious as people with low income may tend to default due to unavailability of money as compared to rich/higher income class.

There seems to be significant relation between Default Rate and Income of Applicant.



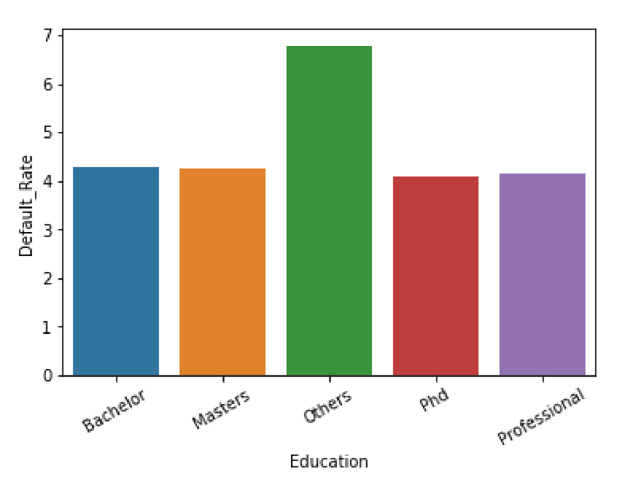
Univariate Analysis on Demographic Data (continued…)

f. **Education:**

Analysis of Impact of Education:

- The default rate of people with education type as Others have highest default rate but number of applicants in for that group is quite low. Hence it is not good idea to make any inference from it.

There seems to be significant relation between Default Rate and Income of Applicant.

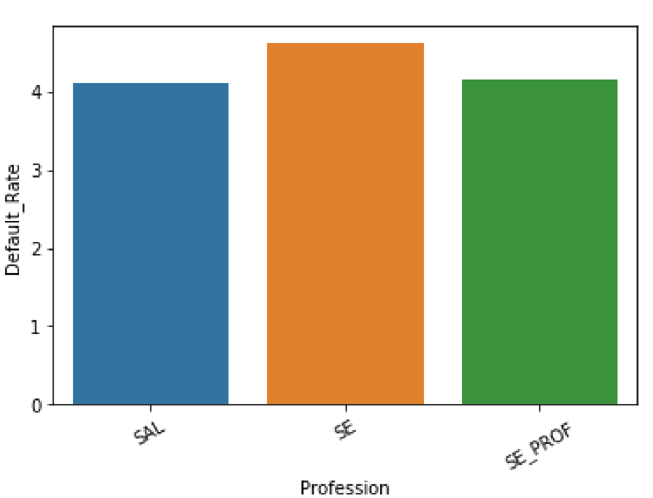


g. **Profession:**

Analysis of Impact of Education:

- The default rate of applicants who are self employed is higher compared to Salaried & Self-Employed Prof

There seems may be relation between Default rate and Profession of Applicant.



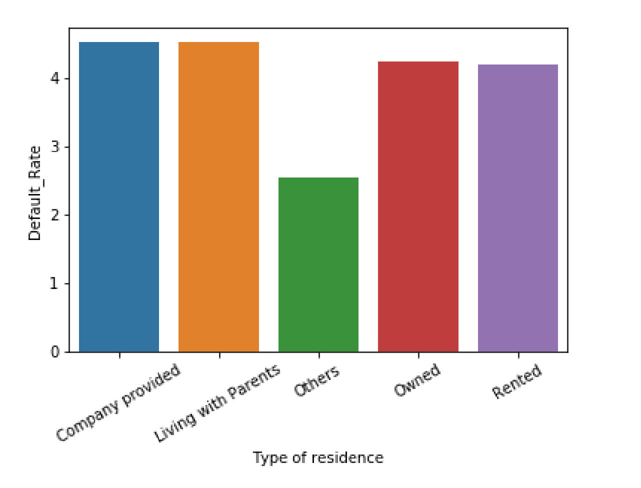
Univariate Analysis on Demographic Data (continued…)

h. **Type of Residence:**

Analysis of Impact of Residence:

- The default rate of applicants who are self emloyed is higher compared to Salaried & Self Employed Prof

There seems `significant` relation between `Default Rate` and `Profession` of Applicant

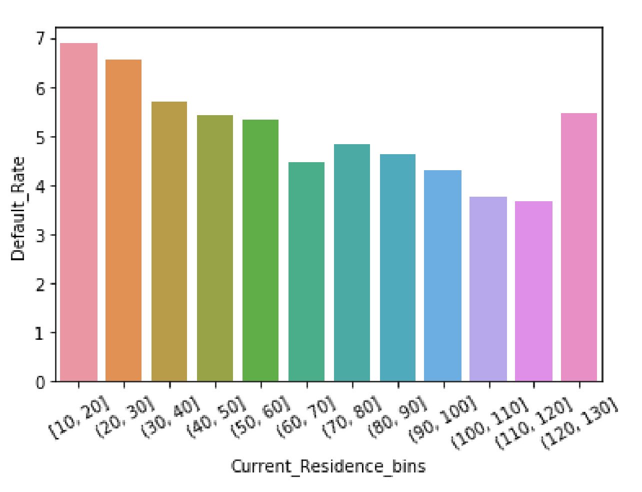


i. **Number of Months in Current Residence:**

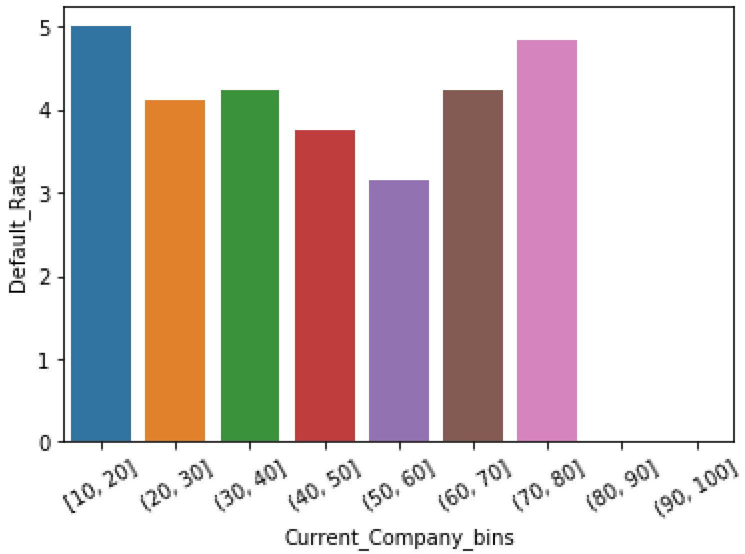
Analysis of Impact:

- The default rate of applicants who have a smaller number of months in current residence is higher than applicants having higher number of months

There seems to significant relation between Default rate and No. of Months in Current Residence of Applicant.



Univariate Analysis on Demographic Data (continued…)



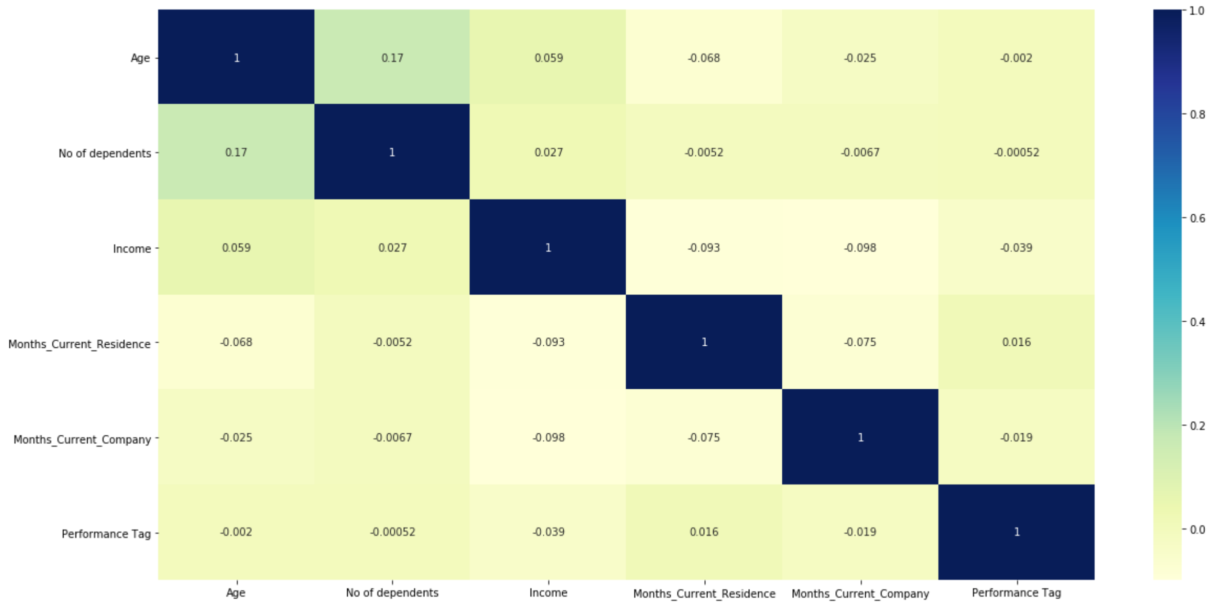
j. **No. of months in current company:**

Analysis of Impact:

- The default rate of applicants who have a smaller number of months in current company is higher than applicants having higher number of months in current company.
- The default rate of applicants with number of months in current company as 70-80 higher. But as the number of applicants in that group is lower, it is difficult to make any inference from it.

There seems to be significant relation between Default Rate and No. of Months in current company of Applicant.

Multivariate Analysis on Demographic Data



Analysis of Heatmap:

- The Performance Tag (Default chance) is highly co-related with Income, followed by Months_Current_Residence and Months_Current_Company.
- No. of Dependent doesn't have significant co-relation with the Performance Tag.

Important Predictor Variables for Demographic Data

The Important Predictor Variables obtained for Demographic Data(without woe Imputation) are :-

1.Income

2.Number of Months in Current Company

The Important Predictor Variables obtained for Demographic Data(with woe Imputation) are :-

1.Income

2.Number of Months in Current Company

3.Age

4. No of months in current residence

Building Model on Demographic Data + Credit Data

Model Building on Demographic Data includes the below steps:

- Data Importing & Data Understanding
- Data Cleaning
- Exploratory Data Analysis
- Data Preparation
- Test-Train Split
- Standardization
- Modelling

Data Importing & Data Understanding of Demographic + Credit Data

Data Importing & Data Understanding of Demographic + Credit Data includes the below steps:

- Importing CSV file
- Inspecting the various aspects of the dataframe
- Creating Master Dataframe

Performing Data Cleaning on Demographic + Credit Data

- **Remove Duplicate Rows:**

Total Number of records before dropping duplicate records :- 71301

Total Number of records after dropping duplicate records :- 71292

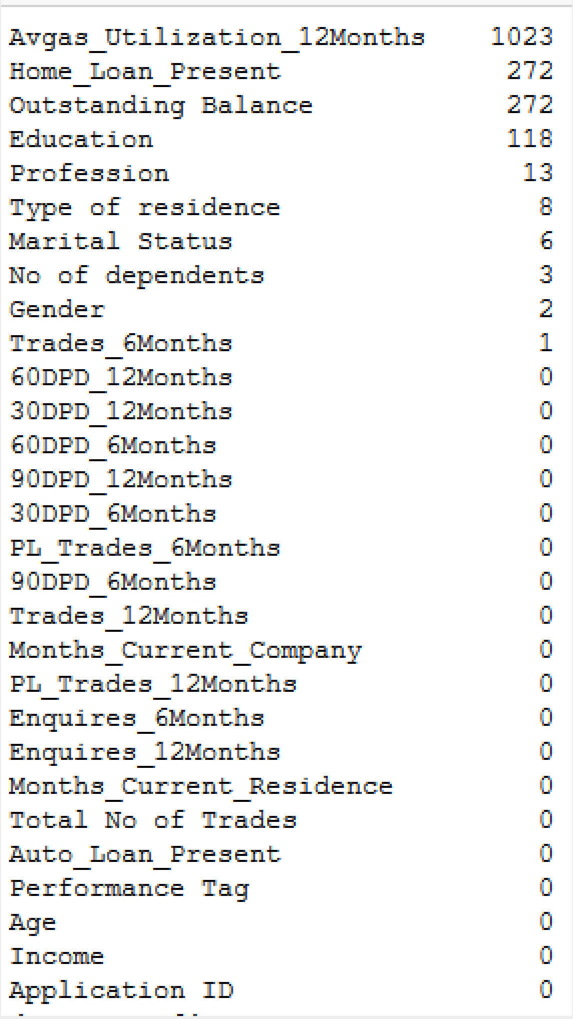
- **Column-wise Missing Values:**

➢ We can drop the Performance Tag variable as this is the dependent variable.

➢ Number of records having missing value in columns like Outstanding Balance, Profession Type of Residence, Marital Status, No of dependents & Gender is very small compared to total number of Records. Hence, drop these variables

➢ **Percentage of Records left after Data Cleaning = 68696/71292 =96.35%**

Due to Data Cleaning around 3.65% data is lost which is not much significant



```
Avgas_Utilization_12Months    1023
Home_Loan_Present              272
Outstanding Balance            272
Education                      118
Profession                      13
Type of residence                8
Marital Status                   6
No of dependents                 3
Gender                           2
Trades_6Months                   1
60DPD_12Months                   0
30DPD_12Months                   0
60DPD_6Months                    0
90DPD_12Months                   0
30DPD_6Months                    0
PL_Trades_6Months                0
90DPD_6Months                    0
Trades_12Months                  0
Months_Current_Company           0
PL_Trades_12Months               0
Enquires_6Months                 0
Enquires_12Months                0
Months_Current_Residence         0
Total No of Trades               0
Auto_Loan_Present                0
Performance Tag                  0
Age                              0
Income                           0
Application ID                   0
```

Exploratory Data Analysis on Demographic + Credit Data

We will do Exploratory Data Analysis on individual features and look or patterns.

1. Univariate Analysis
   - No of times 90 DPD or worse in last 6 months : Number of times customer has not payed dues since 90 days in last 6 months
   - No of times 60 DPD or worse in last 6 months : Number of times customer has not payed dues since 60 days in last 6 months
   - No of times 30 DPD or worse in last 6 months : Number of times customer has not payed dues since 30 days in last 6 months
   - No of times 90 DPD or worse in last 12 months : Number of times customer has not payed dues since 90 days last 12 months
   - No of times 60 DPD or worse in last 12 months : Number of times customer has not payed dues since 60 days last 12 months
   - No of times 30 DPD or worse in last 12 months : Number of times customer has not payed dues since 30 days last 12 months
   - Avgas CC Utilization in last 12 months : Average utilization of credit card by customer
   - No of trades opened in last 6 months : Number of times the customer has done the trades in last 6 months
   - No of trades opened in last 12 months : Number of times the customer has done the trades in last 12 months
   - No of PL trades opened in last 6 months : No of PL trades in last 6 month of customer
   - No of PL trades opened in last 12 months : No of PL trades in last 12 month of customer
   - No of Inquiries in last 6 months : Number of times the customers has inquired in last 6 months
   - No of Inquiries in last 12 months : Number of times the customers has inquired in last 12 months
   - Presence of open home loan : Is the customer has home loan (1 represents "Yes")
   - Outstanding Balance : Outstanding balance of customer
   - Total No of Trades : Number of times the customer has done total trades
   - Presence of open auto loan : Is the customer has auto loan (1 represents "Yes")

2. Multivariate Analysis
   - Heatmap

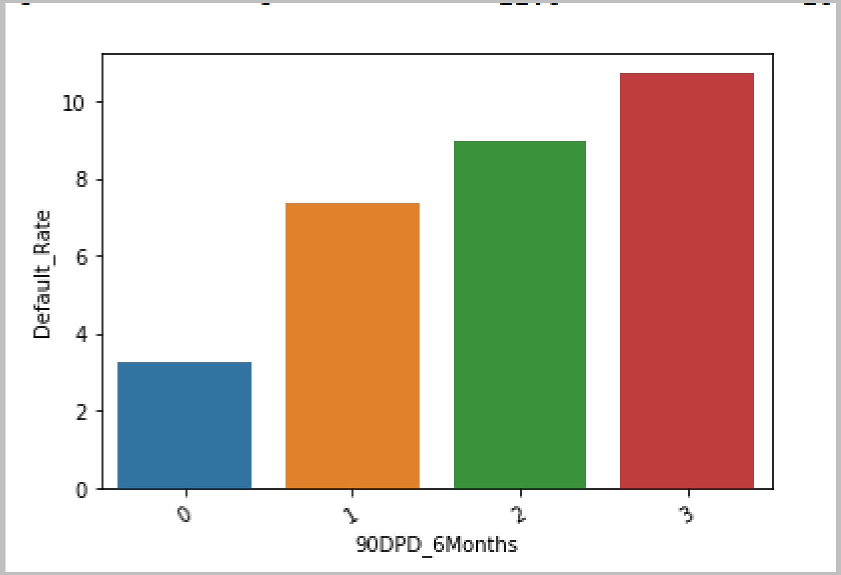
Univariate Analysis on Demographic + Credit Data

a. **No of times 90 DPD or worse in last 6 months:**

Analysis of Impact of No of times 90 DPD or worse in last 6 months:

- The default rate of people with `90DPD_6Months` as 3 is highest but number of applicant in for that group is quite low. hence it is difficult to make any inference from it.
- The Default rate increases with increase of Number of times 90 DPD or worse in last 6 months

There seems to be `significant` relation between `Default Rate` and `90DPD_6Months` of Applicant



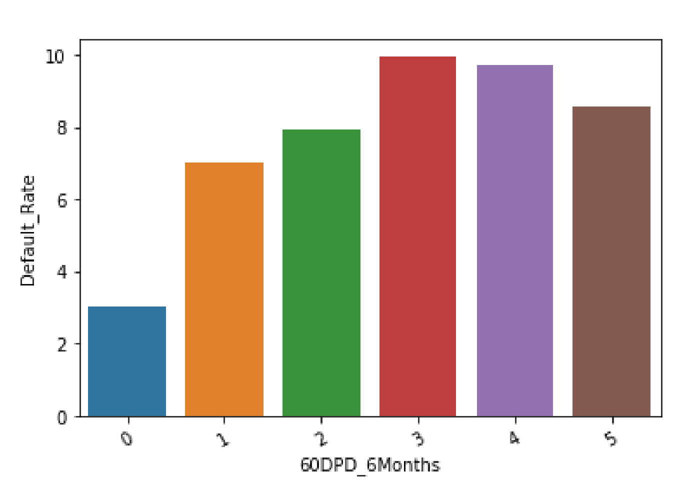
Univariate Analysis on Demographic + Credit Data (continued…)

b. **No of times 60 DPD or worse in last 6 months**

Analysis of Impact of No of times 60 DPD or worse in last 6 months:

- The default rate of people with `60DPD_6Months` as 4 & 5 is fairly high but number of applicant in for that group is quite low. hence it is difficult to make any inference from it.
- From Bar Chart, it seems that The Default rate increases with increase in Number of times 60 DPD or worse in last 6 months

There seems to be `significant` relation between `Default Rate` and `60DPD_6Months` of Applicant

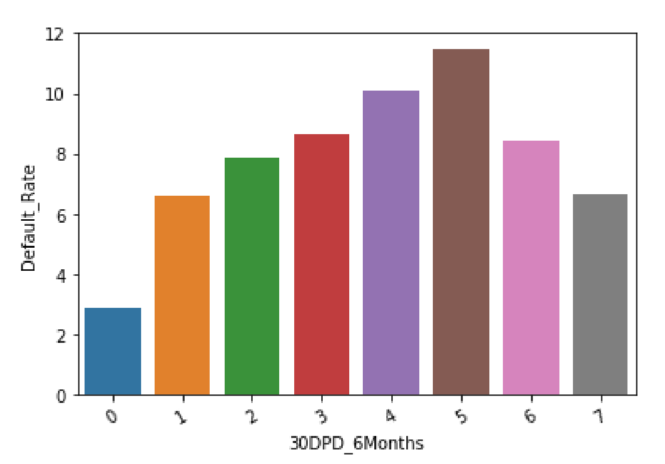


c. **No of times 30 DPD or worse in last 6 months:**

Analysis of Impact of No of times 30 DPD or worse in last 6 months:

- The default rate of people with `30DPD_6Months` as 5,6 & 7 is high but number of applicant in for that group is quite low. hence it is difficult to make any inference from it.
- From Bar Chart, it seems that The Default rate increases with increase in Number of times 30 DPD or worse in last 6 months

There seems to be `significant` relation between `Default Rate` and `30DPD_6Months` of Applicant



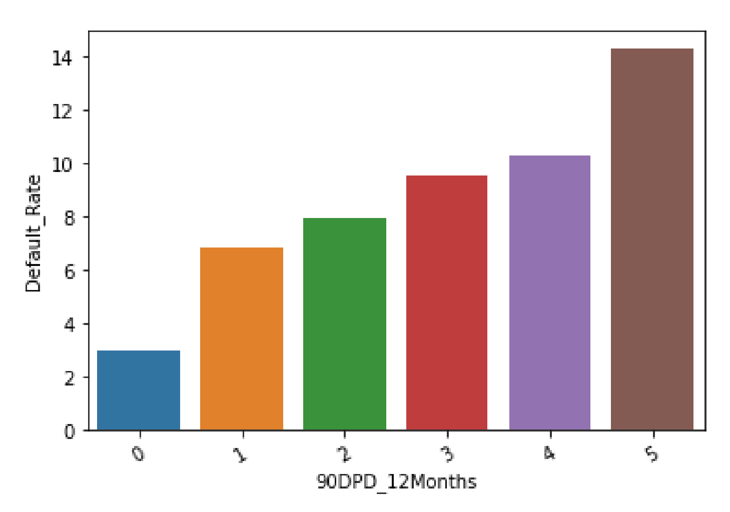
Univariate Analysis on Demographic + Credit Data (continued…)

d. **No of times 90 DPD or worse in last 12 months**

Analysis of Impact of No of times 60 DPD or worse in last 6 months:

- The default rate of people with `90DPD_12Months` as 5,6 & 7 is fairly high but number of applicant in for that group is quite low. hence it is difficult to make any inference from it.
- From Bar Chart, it seems that The Default rate increases with increase in Number of times 90 DPD or worse in last 12 months

There seems to be `significant` relation between `Default Rate` and `90DPD_12Months` of Applicant

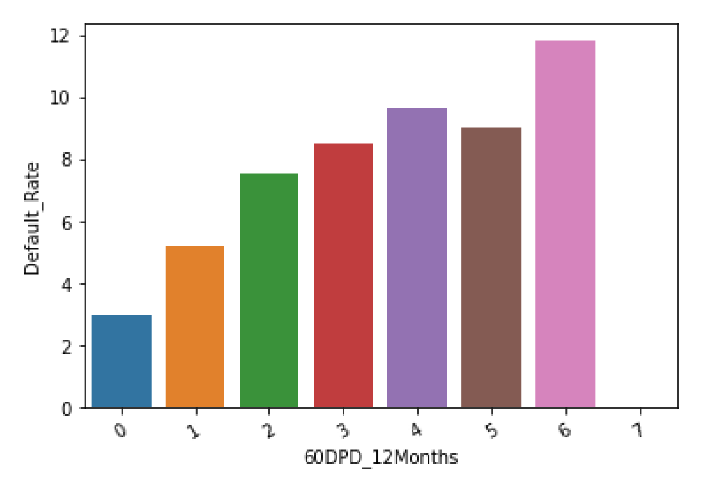


e. **No of times 60 DPD or worse in last 12 months**

Analysis of Impact of No of times 60 DPD or worse in last 12 months:

- From Bar Chart, it seems that The Default rate increases with increase in Number of times 60 DPD or worse in last 12 months

There seems to be `significant` relation between `Default Rate` and `60DPD_12Months` of Applicant



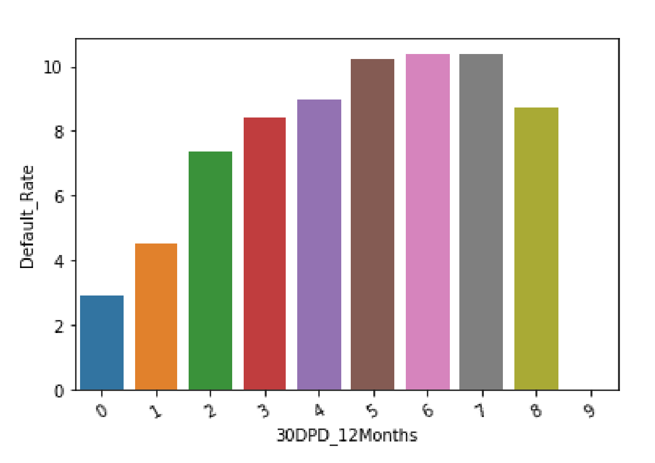
Univariate Analysis on Demographic + Credit Data (continued...)

f. No of times 30 DPD or worse in last 12 months

Analysis of Impact of No of times 30 DPD or worse in last 12 months:

- From Bar Chart, it seems that The Default rate increases with increase in Number of times 30 DPD or worse in last 12 months

There seems to be `significant` relation between `Default Rate` and `30DPD_12Months` of Applicant

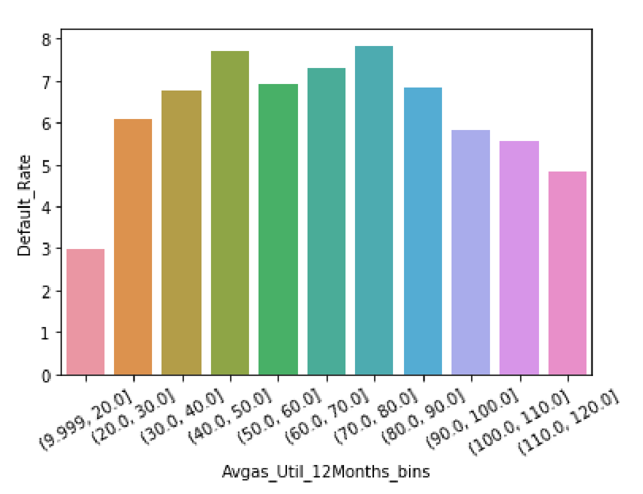


g. Avgas CC Utilization in last 12 months

Analysis of Impact of Avgas CC Utilization in last 12 months:

- The Default rate is highest for Applicants with Avgas Utilization value between 70-80 followed by Applicants with Avgas Utilization value between 60-70

There seems to be `no significant` relation between `Default Rate` and `Avgas_Utilization_12Months` of Applicant



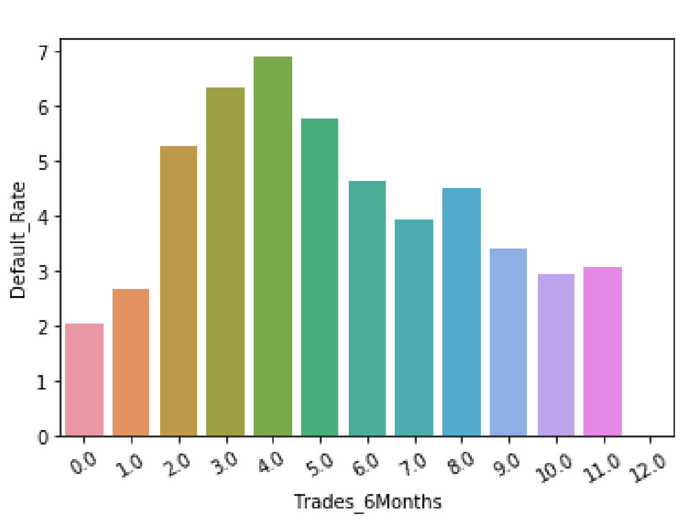
Univariate Analysis on Demographic + Credit Data (continued…)

h. No of trades opened in last 6 months

Analysis of Impact of No of trades opened in last 6 months:

- From Bar Chart, it seems that The Default rate is highest for Number of Trades as 4 followed by number of trades as 3 and number of trades as 5

There seems to be `significant` relation between `Default Rate` and `Trades_6Months` of Applicant

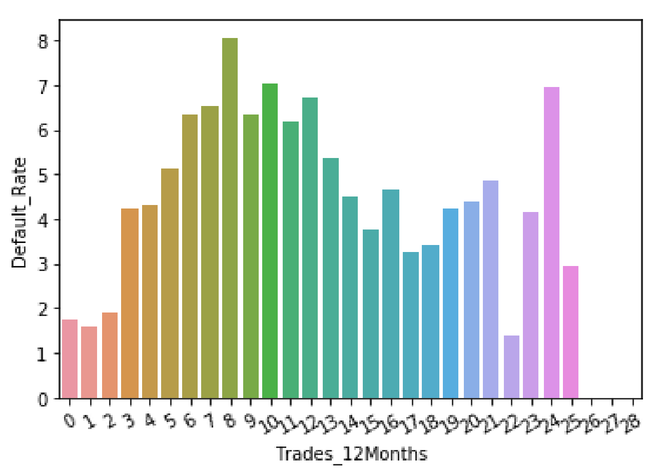


i. No of trades opened in last 12 months

Analysis of Impact of No of trades opened in last 12 months:

- From Bar Chart, it seems that The Default rate is highest for Number of Trades as 8 followed by number of trades as 10

There seems to be ` no significant` relation between `Default Rate` and `Trades_12Months` of Applicant



Univariate Analysis on Demographic + Credit Data (continued…)

l. **No of Inquiries in last 6 months**

Analysis of Impact of No of Inquiries in last 6 months:

- From Bar Chart, it seems that The Default rate is highest for Enquires_6Months as 3 followed by Enquires_6Months as 4

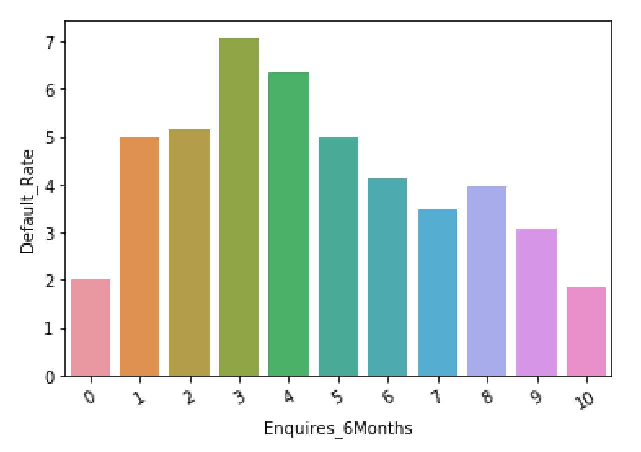
There seems to be `significant` relation between `Default Rate` and `Enquires_6Months` of Applicant

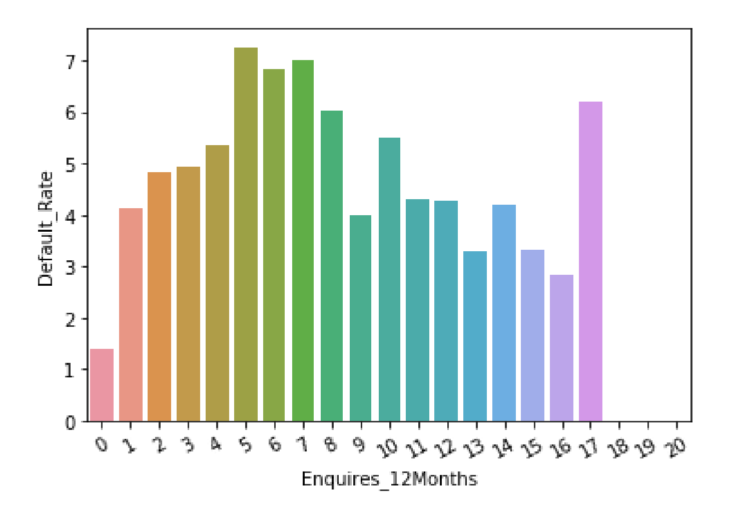
m. **No of Inquiries in last 12 months**

Analysis of Impact of No of Inquiries in last 12 months:

- From Bar Chart, it seems that The Default rate is highest for Enquires_12Months as 5 followed by Enquires_6Months as 7

There seems to be `significant` relation between `Default Rate` and `Enquires_12Months` of Applicant





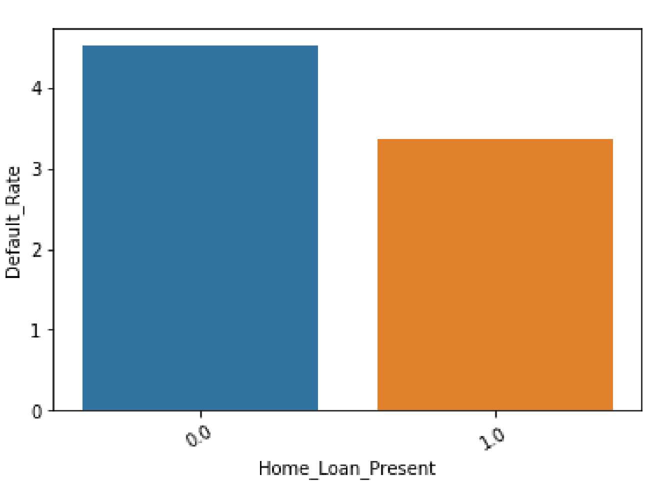
Univariate Analysis on Demographic + Credit Data (continued…)

n. Presence of open home loan

Analysis of Impact of Presence of open home loan:

- From Bar Chart, it seems that The Default rate of applicant with having home loan is lower compared to applicant not having home loan

There seems to be `significant` relation between `Default Rate` and `Home_Loan_Present` of Applicant

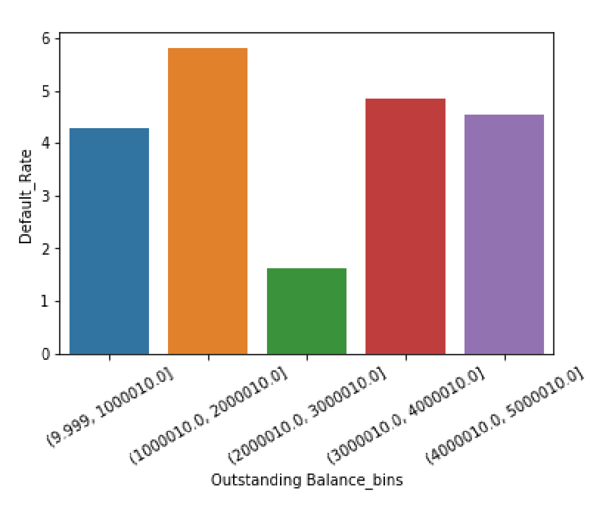


o. Outstanding Balance

Analysis of Impact of Outstanding Balance:

- From Bar Chart, it seems that The Default rate is highest for applicant having outsanding balance between 1million-2million followed by applicant having outstanding balance between 3million-4million

There seems to be `significant` relation between `Default Rate` and `Outstanding Balance` of Applicant



Univariate Analysis on Demographic + Credit Data (continued…)

p. Total No of Trades

Analysis of Impact of Total No of Trades:

- From Bar Chart, it seems that The Default rate is highest for If total no of trades is low

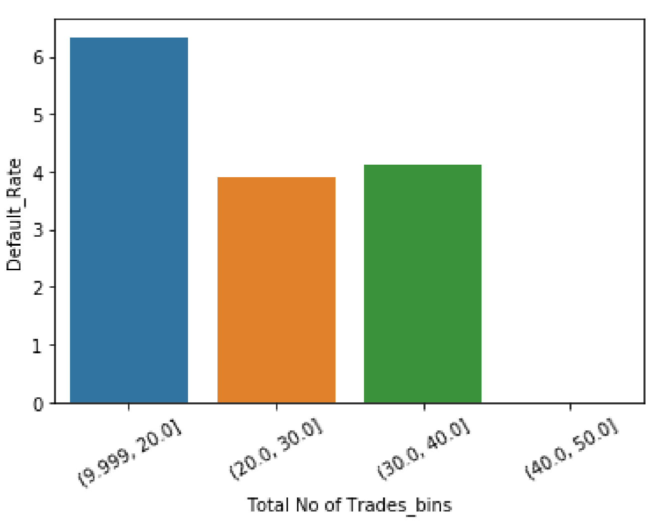
There seems to be `significant` relation between `Default Rate` and `Total No of Trades` of Applicant

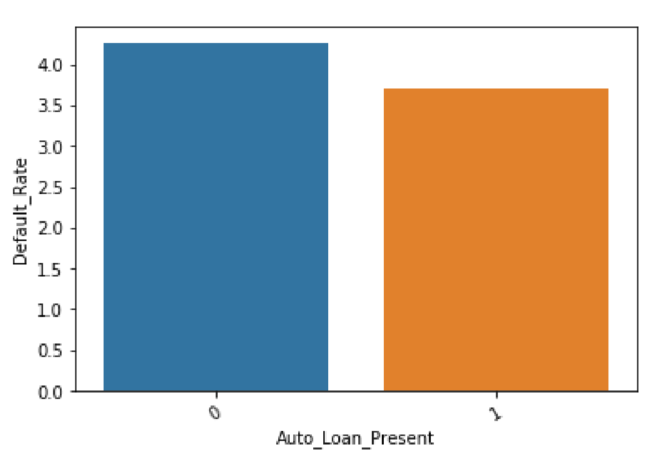
q. Presence of open auto loan

Analysis of Impact of Presence of open auto loan :

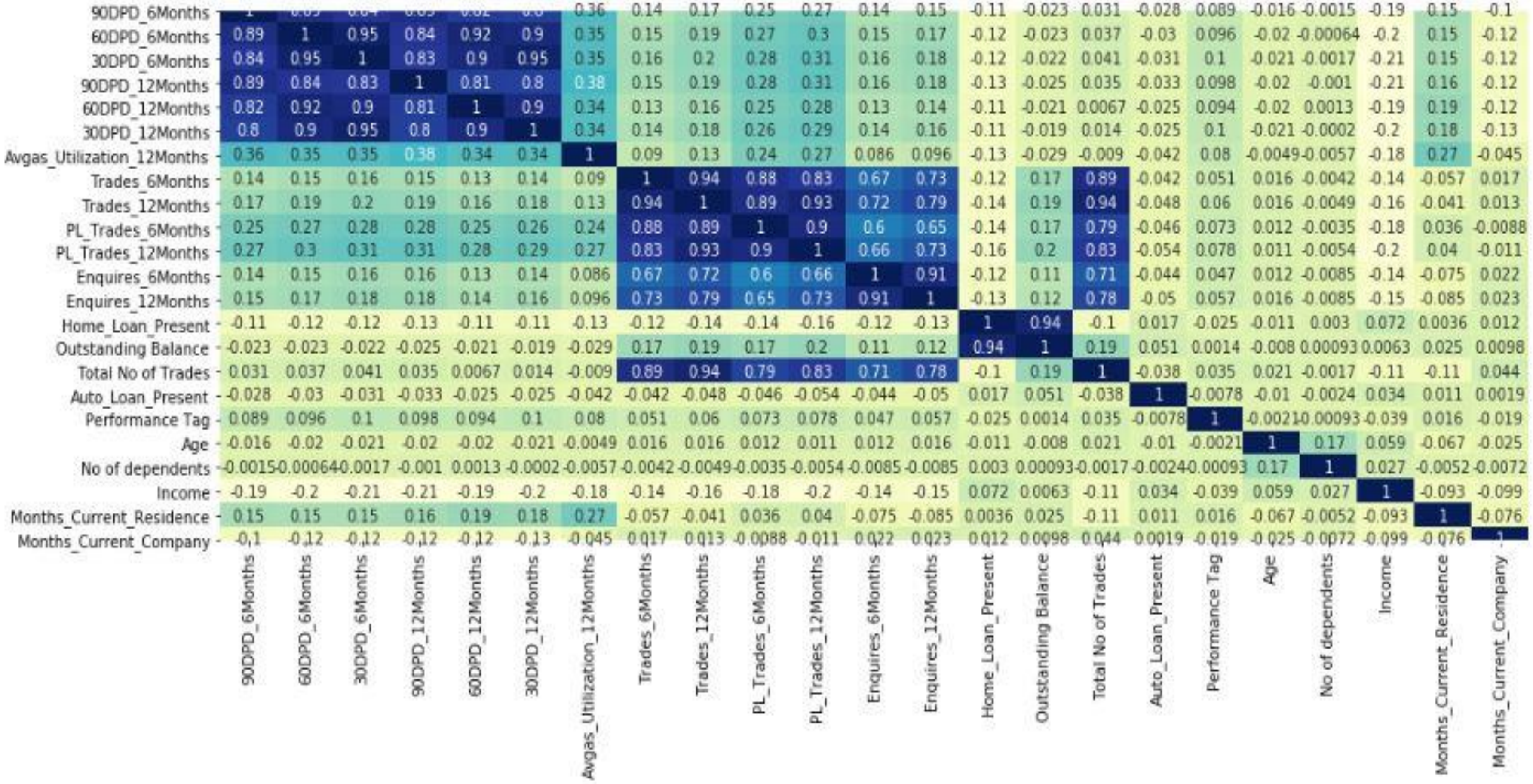
- From Bar Chart, it seems that The Default rate of applicant with having auto loan is lower compared to applicant not having auto loan

There seems to be `significant` relation between `Default Rate` and `Auto_Loan_Present` of Applicant





Multivariate Analysis on Demographic + Credit Data



	90DPD_6Months	60DPD_6Months	30DPD_6Months	90DPD_12Months	60DPD_12Months	30DPD_12Months	Avgas_Utilization_12Months	Trades_6Months	Trades_12Months	PL_Trades_6Months	PL_Trades_12Months	Enquires_6Months	Enquires_12Months	Home_Loan_Present	Outstanding Balance	Total No of Trades	Auto_Loan_Present	Performance Tag	Age	No of dependents	Income	Months_Current_Residence	Months_Current_Company
90DPD_6Months	1	0.89	0.84	0.89	0.82	0.8	0.36	0.14	0.17	0.25	0.27	0.14	0.15	-0.11	-0.023	0.031	-0.028	0.089	-0.016	-0.0015	-0.19	0.15	-0.1
60DPD_6Months	0.89	1	0.95	0.84	0.92	0.9	0.35	0.15	0.19	0.27	0.3	0.15	0.17	-0.12	-0.023	0.037	-0.03	0.096	-0.02	-0.00064	-0.2	0.15	-0.12
30DPD_6Months	0.84	0.95	1	0.83	0.9	0.95	0.35	0.16	0.2	0.28	0.31	0.16	0.18	-0.12	-0.022	0.041	-0.031	0.1	-0.021	-0.0017	-0.21	0.15	-0.12
90DPD_12Months	0.89	0.84	0.83	1	0.81	0.8	0.38	0.15	0.19	0.28	0.31	0.16	0.18	-0.13	-0.025	0.035	-0.033	0.098	-0.02	-0.001	-0.21	0.16	-0.12
60DPD_12Months	0.82	0.92	0.9	0.81	1	0.9	0.34	0.13	0.16	0.25	0.28	0.13	0.14	-0.11	-0.021	0.0067	-0.025	0.094	-0.02	0.0013	-0.19	0.19	-0.12
30DPD_12Months	0.8	0.9	0.95	0.8	0.9	1	0.34	0.14	0.18	0.26	0.29	0.14	0.16	-0.11	-0.019	0.014	-0.025	0.1	-0.021	-0.0002	-0.2	0.18	-0.13
Avgas_Utilization_12Months	0.36	0.35	0.35	0.38	0.34	0.34	1	0.09	0.13	0.24	0.27	0.086	0.096	-0.13	-0.029	-0.009	-0.042	0.08	-0.0049	-0.0057	-0.18	0.27	-0.045
Trades_6Months	0.14	0.15	0.16	0.15	0.13	0.14	0.09	1	0.94	0.88	0.83	0.67	0.73	-0.12	0.17	0.89	-0.042	0.051	0.016	-0.0042	-0.14	-0.057	0.017
Trades_12Months	0.17	0.19	0.2	0.19	0.16	0.18	0.13	0.94	1	0.89	0.93	0.72	0.79	-0.14	0.19	0.94	-0.048	0.06	0.016	-0.0049	-0.16	-0.041	0.013
PL_Trades_6Months	0.25	0.27	0.28	0.28	0.25	0.26	0.24	0.88	0.89	1	0.9	0.6	0.65	-0.14	0.17	0.79	-0.046	0.073	0.012	-0.0035	-0.18	0.036	-0.0088
PL_Trades_12Months	0.27	0.3	0.31	0.31	0.28	0.29	0.27	0.83	0.93	0.9	1	0.66	0.73	-0.16	0.2	0.83	-0.054	0.078	0.011	-0.0054	-0.2	0.04	-0.011
Enquires_6Months	0.14	0.15	0.16	0.16	0.13	0.14	0.086	0.67	0.72	0.6	0.66	1	0.91	-0.12	0.11	0.71	-0.044	0.047	0.012	-0.0085	-0.14	-0.075	0.022
Enquires_12Months	0.15	0.17	0.18	0.18	0.14	0.16	0.096	0.73	0.79	0.65	0.73	0.91	1	-0.13	0.12	0.78	-0.05	0.057	0.016	-0.0085	-0.15	-0.085	0.023
Home_Loan_Present	-0.11	-0.12	-0.12	-0.13	-0.11	-0.11	-0.13	-0.12	-0.14	-0.14	-0.16	-0.12	-0.13	1	0.94	-0.1	0.017	-0.025	-0.011	0.003	0.072	0.0036	0.012
Outstanding Balance	-0.023	-0.023	-0.022	-0.025	-0.021	-0.019	-0.029	0.17	0.19	0.17	0.2	0.11	0.12	0.94	1	0.19	0.051	0.0014	-0.008	0.00093	0.0063	0.025	0.0098
Total No of Trades	0.031	0.037	0.041	0.035	0.0067	0.014	-0.009	0.89	0.94	0.79	0.83	0.71	0.78	-0.1	0.19	1	-0.038	0.035	0.021	-0.0017	-0.11	-0.11	0.044
Auto_Loan_Present	-0.028	-0.03	-0.031	-0.033	-0.025	-0.025	-0.042	-0.042	-0.048	-0.046	-0.054	-0.044	-0.05	0.017	0.051	-0.038	1	-0.0078	-0.01	-0.0024	0.034	0.011	0.0019
Performance Tag	0.089	0.096	0.1	0.098	0.094	0.1	0.08	0.051	0.06	0.073	0.078	0.047	0.057	-0.025	0.0014	0.035	-0.0078	1	-0.0021	-0.00093	-0.039	0.016	-0.019
Age	-0.016	-0.02	-0.021	-0.02	-0.02	-0.021	-0.0049	0.016	0.016	0.012	0.011	0.012	0.016	-0.011	-0.008	0.021	-0.01	-0.0021	1	0.17	0.059	-0.067	-0.025
No of dependents	-0.0015	-0.00064	-0.0017	-0.001	0.0013	-0.0002	-0.0057	-0.0042	-0.0049	-0.0035	-0.0054	-0.0085	-0.0085	0.003	0.00093	-0.0017	-0.0024	-0.00093	0.17	1	0.027	-0.0052	-0.0072
Income	-0.19	-0.2	-0.21	-0.21	-0.19	-0.2	-0.18	-0.14	-0.16	-0.18	-0.2	-0.14	-0.15	0.072	0.0063	-0.11	0.034	-0.039	0.059	0.027	1	-0.093	-0.099
Months_Current_Residence	0.15	0.15	0.15	0.16	0.19	0.18	0.27	-0.057	-0.041	0.036	0.04	-0.075	-0.085	0.0036	0.025	-0.11	0.011	0.016	-0.067	-0.0052	-0.093	1	-0.076
Months_Current_Company	-0.1	-0.12	-0.12	-0.12	-0.12	-0.13	-0.045	0.017	0.013	-0.0088	-0.011	0.022	0.023	0.012	0.0098	0.044	0.0019	-0.019	-0.025	-0.0072	-0.099	-0.076	1

Analysis of Heatmap:

- - The Variable `**Age**` have highest co relation with `**Performance Tag**`(Defaulter Rate) Followed by variables `**30DDPD_12Months**` & `**30DDPD_6Months**`
- - Variables related to `**Trade**` are highly co related to each other
- - Similarly Variables related to `**DPD**` are also highly co related to each other

Important Predictor Variables for Demographic + Credit Data

The Important Predictor Variables obtained for Master Data(without woe Imputation) are :-

1.Income

2.Total Number of Trades

3.Avgas CC Utilization in last 12 months

4.No of months in current company

The Important Predictor Variables obtained for Master Data(with woe Imputation) are :-

1.No of times 30 DPD or worse in last 6 months

2.Avgas CC Utilization in last 12 months

3.Age

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

5.No of trades opened in last 12 months

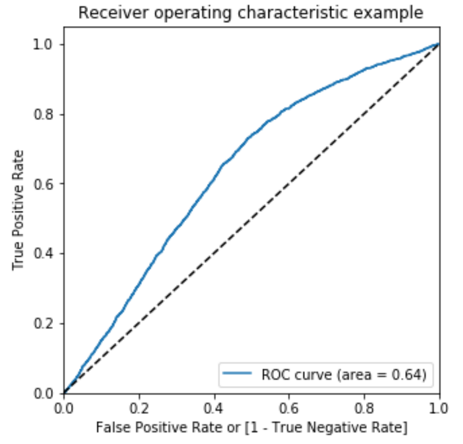
Building Model on Non-WOE Master Dataset

i. Logistic Regression on Non-Woe Master Dataset:

a) Analyzing top predictors:

1. Avgas_Utilization_12Months
2. Total No of Trades
3. Income
4. Months_Current_Company

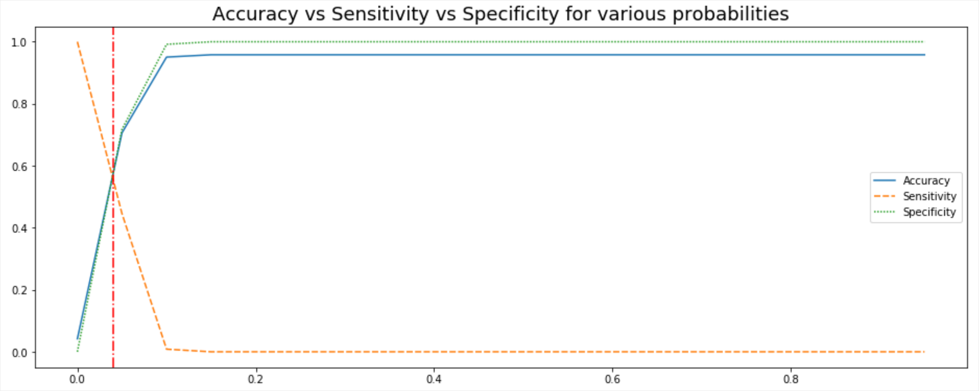
b) ROC Curve:



Building Model on Non-WOE Master Dataset (Contd..)

c) Obtaining Optimum Cutoff between Sensitivity, Specificity and Accuracy

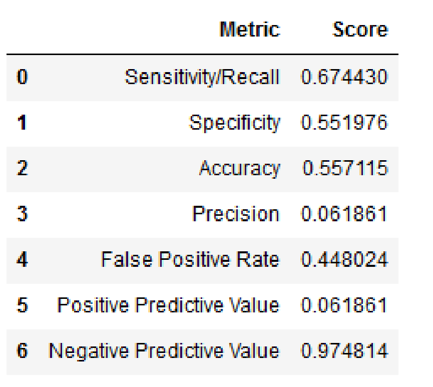
ROC Curve shows that 0.04 is the optimum point to take it as a cutoff Probability.


Accuracy vs Sensitivity vs Specificity for various probabilities

Metrics and their score for the Confusion Matrix:

Confusion Matrix:

[[25429 20640]
657 1361]]



	Metric	Score
0	Sensitivity/Recall	0.674430
1	Specificity	0.551976
2	Accuracy	0.557115
3	Precision	0.061861
4	False Positive Rate	0.448024
5	Positive Predictive Value	0.061861
6	Negative Predictive Value	0.974814

Analysis of Metrics of Logistic Regression on Non-Woe Master Dataset:

- Specificity & Accuracy of Logistic Regression is around 55% which is quite low.
- But Sensitivity/Recall of Logistic Regression is 67% which is also quite low.

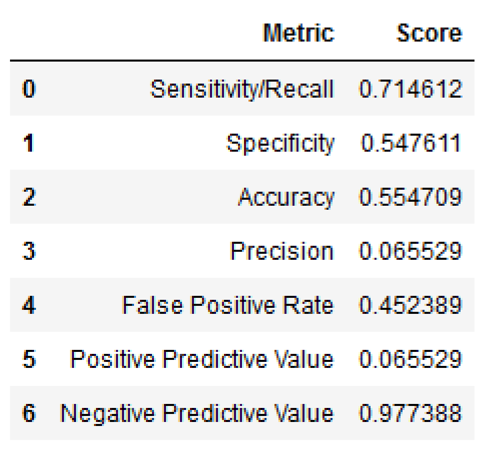
Building Model on Non-WOE Master Dataset (Contd..)

ii. Random Forest Model on Non-WOE Dataset

Confusion Matrix:

```
[[10806   8927]
    250     626]]
```

Metrics and their score for the above confusion Matrix:



	Metric	Score
0	Sensitivity/Recall	0.714612
1	Specificity	0.547611
2	Accuracy	0.554709
3	Precision	0.065529
4	False Positive Rate	0.452389
5	Positive Predictive Value	0.065529
6	Negative Predictive Value	0.977388

Analysis of Metrics of Random Forest for Non-WOE Master Dataset:

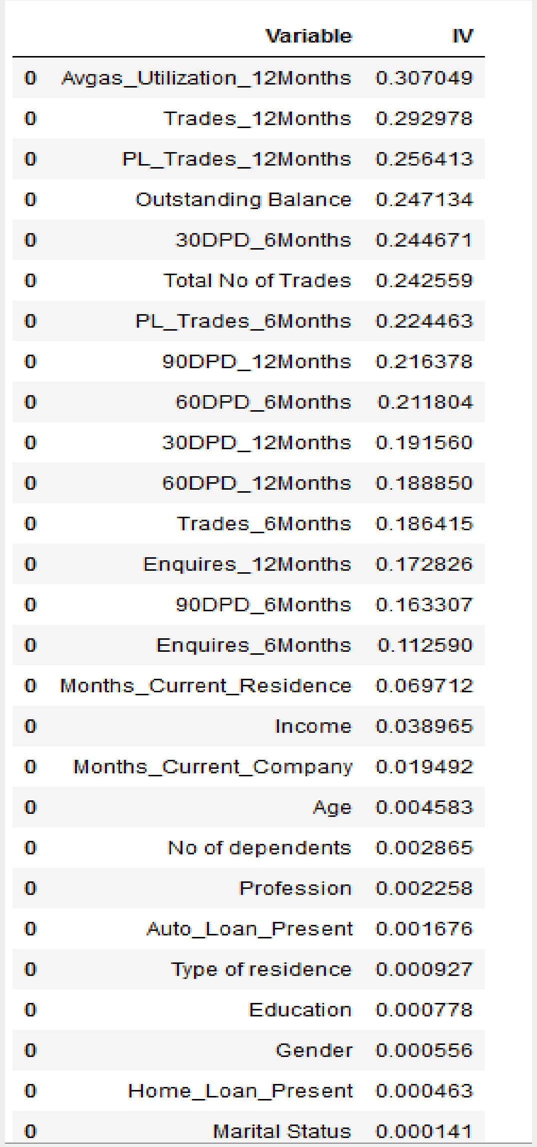
- Specificity & Accuracy of Random Forest is above 54% which is quite low.
- But Sensitivity/Recall of Random Forest is 70% which is quite good.

WOE /IV Analysis on Master Dataset

1. WOE & IV Value for Master Dataset

Information Value Analysis:

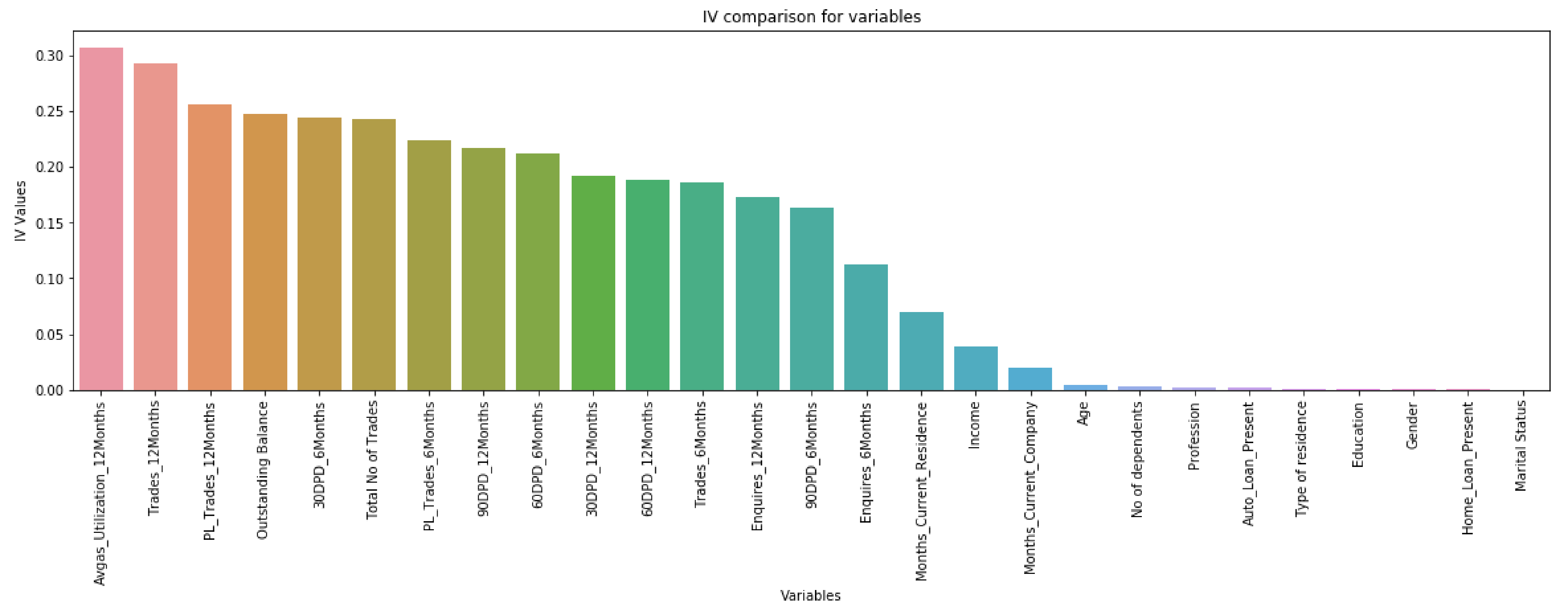
- The dataframe shows that Avgas_Utilization_12Months have highest information value followed by Trades_12Months & PL_Trades_12Months
- The information value of Demographic variables is lower than information value of Credit Variables.



	Variable	IV
0	Avgas_Utilization_12Months	0.307049
0	Trades_12Months	0.292978
0	PL_Trades_12Months	0.256413
0	Outstanding Balance	0.247134
0	30DPD_6Months	0.244671
0	Total No of Trades	0.242559
0	PL_Trades_6Months	0.224463
0	90DPD_12Months	0.216378
0	60DPD_6Months	0.211804
0	30DPD_12Months	0.191560
0	60DPD_12Months	0.188850
0	Trades_6Months	0.186415
0	Enquires_12Months	0.172826
0	90DPD_6Months	0.163307
0	Enquires_6Months	0.112590
0	Months_Current_Residence	0.069712
0	Income	0.038965
0	Months_Current_Company	0.019492
0	Age	0.004583
0	No of dependents	0.002865
0	Profession	0.002258
0	Auto_Loan_Present	0.001676
0	Type of residence	0.000927
0	Education	0.000778
0	Gender	0.000556
0	Home_Loan_Present	0.000463
0	Marital Status	0.000141

WOE /IV Analysis on Master Dataset (Contd…)

2. **Plotting all the consolidated IV Values of the Master variable:**



The Avgas_Utilization_12Months have more prediction power compared to other Master variable.

Building Model with Master Data (with WOE imputation)

1. **Logistic Regression with Master Data (with WOE imputation)**

a) **Analyzing top predictors:**

(i) 30DPD_6Months

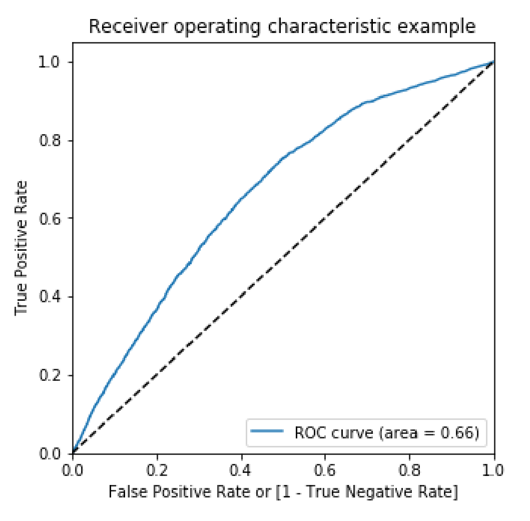
(ii) Avgas_Utilization_12Months

(iii) Trades_12Months

(iv) Enquires_6Months

(v) Age

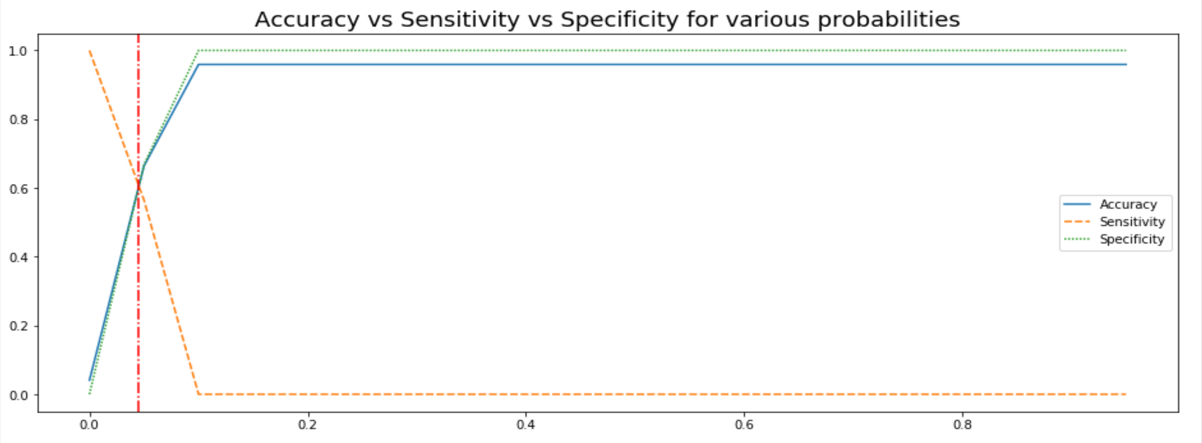
b) **ROC Curve:**



Building Model with Master Data (with WOE imputation) (Contd…)

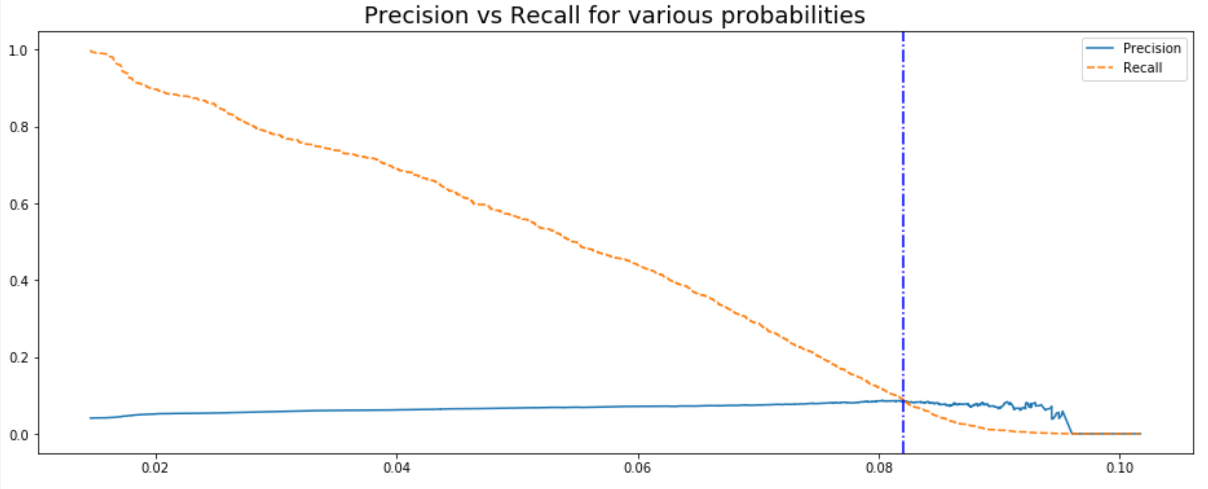
c) Obtaining Optimum Cutoff between Sensitivity, Specificity & Accuracy:

ROC Curve shows that 0.045 is the optimum point to take it as cutoff probability.



d) Precision and Recall Tradeoff:

The optimum value from Precision and Recall is 0.082



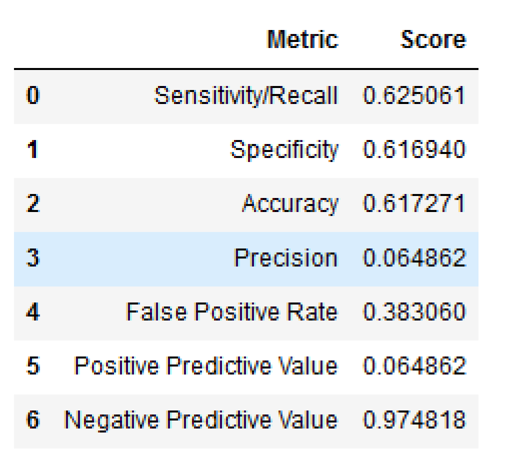
Building Model with Master Data (with WOE imputation) (Contd…)

Considering final cutoff probability as 0.045

Confusion Matrix:

```
[[29536   18339]
    763      1272]]
```

Metrics and their score for above confusion matrix:



	Metric	Score
0	Sensitivity/Recall	0.625061
1	Specificity	0.616940
2	Accuracy	0.617271
3	Precision	0.064862
4	False Positive Rate	0.383060
5	Positive Predictive Value	0.064862
6	Negative Predictive Value	0.974818

Analysis of Metrics of Logistic Regression on WOE Master Dataset:

- Sensitivity/Recall , Specificity & Accuracy of Logistic Regression is higher than 60% which is quite good.
- But Precision of Logistic Regression is 6.5% which is also quite low.

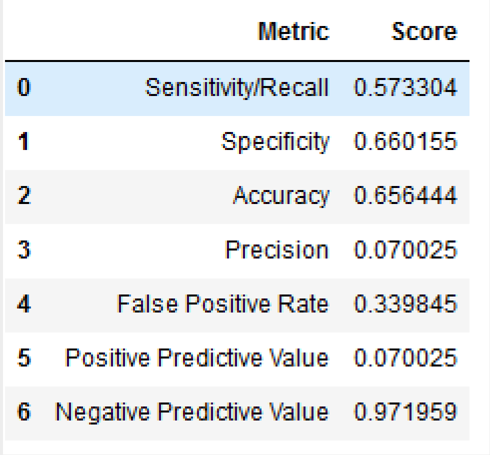
Building Model with Master Data (with WOE imputation) (Contd…)

2. Random Forest on WOE Master Dataset:

Confusion Matrix:

```
[[13518   6959]
    390    524]]
```

Metrics and their score for above confusion matrix:



	Metric	Score
0	Sensitivity/Recall	0.573304
1	Specificity	0.660155
2	Accuracy	0.656444
3	Precision	0.070025
4	False Positive Rate	0.339845
5	Positive Predictive Value	0.070025
6	Negative Predictive Value	0.971959

Analysis of Metrics of Random Forest on WOE Dataset:

- Specificity & Accuracy of Random Forest is higher than 60% which is quite good
- Sensitivity/Recall of Random Forest is lower than 60% which is not good (for final model)
- But Precision of Random Forest is 6.6% which is also quite low.

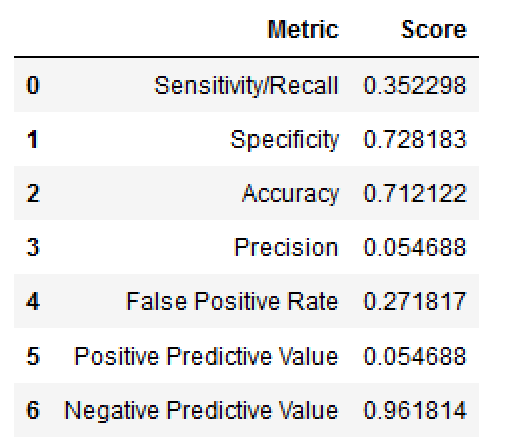
Building Model with Master Data (with WOE imputation) (Contd…)

3. Decision Tree on WOE Master Dataset:

Confusion Matrix:

```
[[14911  5566]
    592    322]]
```

Metrics and their score for the above confusion Matrix:



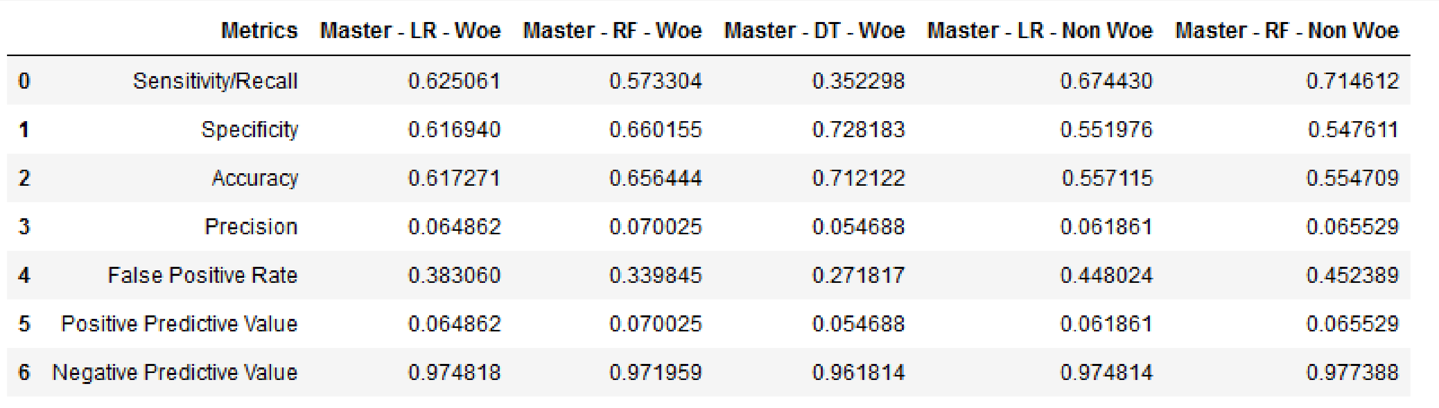
	Metric	Score
0	Sensitivity/Recall	0.352298
1	Specificity	0.728183
2	Accuracy	0.712122
3	Precision	0.054688
4	False Positive Rate	0.271817
5	Positive Predictive Value	0.054688
6	Negative Predictive Value	0.961814

Analysis of Metrics of Decision Trees on WOE Dataset:

- Specificity & Accuracy of Decision Trees is higher than 60% which is quite good.
- Sensitivity/Recall of Decision Tree is lower than 60% which is quite low.
- But Precision of Decision Trees is 5% which is quite low.

Model Selection & Model Evaluation

➢ Selection of Best Model:



	Metrics	Master - LR - Woe	Master - RF - Woe	Master - DT - Woe	Master - LR - Non Woe	Master - RF - Non Woe
0	Sensitivity/Recall	0.625061	0.573304	0.352298	0.674430	0.714612
1	Specificity	0.616940	0.660155	0.728183	0.551976	0.547611
2	Accuracy	0.617271	0.656444	0.712122	0.557115	0.554709
3	Precision	0.064862	0.070025	0.054688	0.061861	0.065529
4	False Positive Rate	0.383060	0.339845	0.271817	0.448024	0.452389
5	Positive Predictive Value	0.064862	0.070025	0.054688	0.061861	0.065529
6	Negative Predictive Value	0.974818	0.971959	0.961814	0.974814	0.977388

FINAL SELECTED MODEL – LOGISTIC REGRESSION ON WOE DATASET

Reasons for selection of this model:-

- Our Business Objective is acquire the right customers, hence we should select model with good Sensitivity & Specificity value.
- From above summary table we can see that, the model having good sensitivity & specificity is Logistic Regression model built on WOE Dataset.
- Hence we are selecting Logistic Regression model build on WOE Dataset as our final model.

Model Selection & Model Evaluation (Contd…)

➢ Model Evaluation:

- Minimum Probability of Default for Rejected Candidate :- 0.19262073358598422
- Maximum Probability of Default for Rejected Candidate :- 0.9915658041517841
- Average Probability of Default for Rejected Candidate :- 0.8220809193977572

- Number of Actual Defaulters :- 1425
- Number of Predicted Defaulters :- 1425
- Accuracy of Model :- 100.0

Thus all the rejected applicants are detected by our model as possible defaulters.

Model Selection & Model Evaluation (Contd…)

- **Application Scorecard:**

It is required to build an application scorecard with the good to bad odds. The ideal model that is suitable for such a case is Logistic Regression model with all variables(combined data of demographics and Credit Bureau). Similar representation of non-defaulters to defaulters we also have in our WoE transformed representation. So, we’ll need both our logistic regression coefficients that we got from fitting our model (WoE Logistic Regression model combined data model with L1 regularization prepared earlier) as well as our WOE dataset with the transformed WOE values.

The scorecard in this case will be evaluated as :

target_score = 400

target_odds = 10

pts_double_odds = 20

factor = pts_double_odds / log10(2)

offset = target_score - factor X log10(target_odds)

Finally,

scorecard['score'] = offset - factor X scorecard['logit’]

Note: The negative sign for the term of scorecard['logit'] is applied as model is evaluating probabilities ofodds of defaults (bad). However, we tend to provide a high score to the persons who are good. Therefore, the sign is negative

Model Selection & Model Evaluation (Contd...)

- Value of Factor = 66.43856189774725
- Value of Offset = 333.56143810225274

➢ Creation of Scorecard for Approved Applicants (Demo data + Credit data):

Checking the scores for all the values in performance tag:

Min value :- 275.239696093198

Max value :- 399.06066778417016

Mean value :- 344.34383739408673

Median value :- 344.7440257626507

Cut-Off Score value in Score card for Approved Applicants is 325

Model Selection & Model Evaluation (Contd…)

- **Creation of Scorecard for Rejected Applicants (Demo data + Credit data):**

Checking the scores for all the values in performance tag:

Min value :- 270.8868097284634

Max value :- 340.81895399616366

Mean value :- 288.94164470206226

Median value :- 285.6793023500219

Model Selection & Model Evaluation (Contd…)

- **Comparing Scorecard of Approved Applicants vs Rejected Applicants for Master data(Demo data +Credit data):**

For Rejected Candidates --> Min ScoreCard Value :- 270.8868097284634 & Max ScoreCard value :- 340.81895399616366

For Approved Candidates --> Min ScoreCard Value :- 275.239696093198 & Max ScoreCard value :- 399.06066778417016

Total Approved Applicants :- 71301

Total Approved Applicants having score lower than cut off score:- 27905

Percentage of Applicants having score lower than cut off score:- 0.3913689850072229

Total Rejected Applicants :- 1425

Total Rejected Applicants having score higher than cut off score:- 31

Percentage of Applicants having score higher than cut off score:- 0.9782456140350877

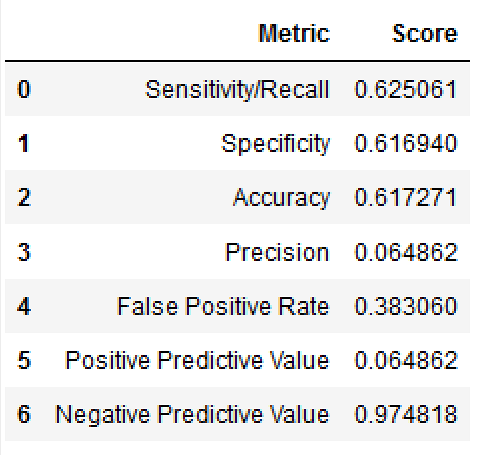
Scorecard Comparison - Approved Applicants vs Rejected Applicants:

- 97.82% of Rejected Applicants have score lower than cut off score (i.e 325).Hence If Cut Off score is kept as 325, we can detect 91.82% potential defaulters
- 39.13% of Approved Applicants have score lower than cut off score (i.e 325).Hence If Cut Off score is kept as 325, 39.13% applicants who are not defaulters will not get credit card

Model Selection & Model Evaluation (Contd…)

- **Assessing the financial benefit of selected Model**

1. **The model can detect 62% of defaulters.**

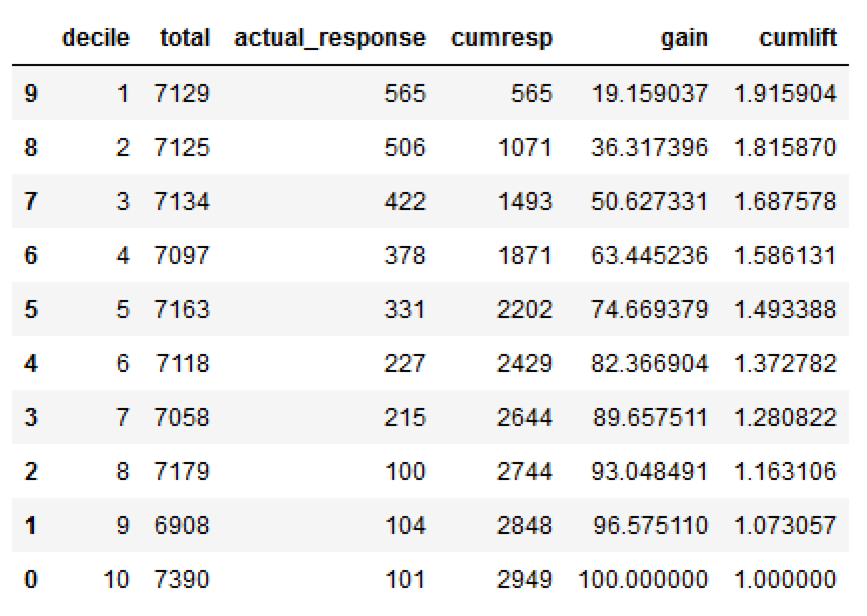


	Metric	Score
0	Sensitivity/Recall	0.625061
1	Specificity	0.616940
2	Accuracy	0.617271
3	Precision	0.064862
4	False Positive Rate	0.383060
5	Positive Predictive Value	0.064862
6	Negative Predictive Value	0.974818

- The Sensitivity/Recall of model is around 62%, which means it can detect 62% of potential defaulters from any given data.
- Hence the 62% of potential losses can be reduces by the bank.

Model Selection & Model Evaluation (Contd…)

2. **The model can detect 75% of defaulters by targeting 50% applicants.**



	decile	total	actual_response	cumresp	gain	cumlift
9	1	7129	565	565	19.159037	1.915904
8	2	7125	506	1071	36.317396	1.815870
7	3	7134	422	1493	50.627331	1.687578
6	4	7097	378	1871	63.445236	1.586131
5	5	7163	331	2202	74.669379	1.493388
4	6	7118	227	2429	82.366904	1.372782
3	7	7058	215	2644	89.657511	1.280822
2	8	7179	100	2744	93.048491	1.163106
1	9	6908	104	2848	96.575110	1.073057
0	10	7390	101	2949	100.000000	1.000000

From the above dataframe, we can see that 75% defaulters can be detected by targeting 50% applicants.

Thank You