BFS Capstone Project

CredX- Leading Credit Card Provider

Group Name:
1) Habeeb Rehman
2) Sanyam Goswami

CredX

Problem Statement:

CredX – a leading credit card provider is experiencing an increase in credit loss due to customer defaulting. Acquisition of right customers may be the best strategy to overcome this situation.

Objective:

- Identifying right customers using modelling
- Various variables affecting credit risk
- Creation of strategies to mitigate the acquisition risk
- Assess the financial benefit of the project

Steps for problem solving

As per the problem statement, It is binary supervised classification problem.

We build on Random forest and SVM to identify the customers having risk of defaulting

As per CRISP DM Framework, following steps were followed

1) Understanding the problem statement and the data set provided (demographic and Credit Bureau)
2) Data Cleaning. Using IV values for imputing values wherever necessary
3) Performing EDA on each data set individually to find significant variables
4) WOE(weight of evidence) and Information Value (IV value) on both the data set
   - With demographic transformed data set
   - Merging both the data set and prepare a list of significant variables
5) Using both the original data set and the transformed data set to prepare data models. We have tried to build the models like Random Forest ,SVM
6) Evaluate the models using sensitivity, specificity , accuracy features , precision and recall(Confusion matrix)
7) Creation of the application score card based on the models
8) Access the financial benefits
9) Present results to management

Approach Plan

Data Pre-processing

Merging the demographic data and Credit Bureau data into one master file (based on variable Application id)which gives us total 3 files for analysis

- Demographic- Total 71295 obervations with 12 variables
- Credit Bureau- Total 71295 Observations with 19 varaibles
- Master(Merged data set)

Data Cleaning Approach (Using the concept of WOE & Information Value)

The weight of evidence tells the predictive power of an independent variable in relation to the dependent variable

WOE = Ln(Distribution of Goods/Distribution of Bads)

Where **distribution of goods** - % of good customers in a group

distribution of bads - % of bad customers in a group

Data Quality issues

- Application id – Unique id in both data sets
- Removal of the duplicate application id after the merging of the two data set(3 rows)
- Where data is missing for continuous variable, WOE value can be imputed
- Performance tag have significant no of missing values which can be ignored – total 1425 rows
- For the missing values in variable Avgas CC Utilization in last 12 months , replace with 0
- For the missing values in Presence of open home loan, replace with 0

Data Quality issues

Variable Name	Data issue and approach
Application id	3 Duplicate application id has been excluded
Age	1 customer had negative age 19 customer with age as 0 45 customer had age less than 18. for all these records age was set as 18
Gender	2 Missing values were imputed
Marital status	6 Missing values were imputed
Income	81 negative values were imputed with median values
No of dependents	3 Missing value were imputed
Education	119 missing values were imputed with Unknown
Profession	14 Missing values were imputed
Type of residence	8 missing values were imputed
Performance tag	Total 1425 values were removed as there was no data for them

Steps for Calculating WOE(Our approach)

- Continuous variables in both the dataset set like **age**(range from 0 to 65) can be split into 5 parts(or buckets based on distribution)
- In age some values are wrong with negative values and the value below 18 as they are not eligible which can be ignored for WOE
- Calculation of the number of credit events and non-events in each group
- Calculation of the % of events and % of non-events in each group
- Calculate the WOE by taking natural log of the value in previous step
- WOE for age group between range 36-38 and 39-41 has high WOE which may be taken as one of the factor in predicting good customer

Similarly WOE can be utilised in predicting for other continous variables like

- Income
- Trade opened in last twelve months

WOE and IV Analysis

Variable	WOE Value
No of inquiries in last 12 months	0.27
Avg Credit card Utilization in last 12 months	0.26
No of times 30 DPD in last 6 months	0.24
No of times 90 DPD in last 12 months	0.21
No of times 60 DPD in last 12 months	0.19
No of trades opened in last 12 moths	0.19
No of times 60 DPD in last 12 months	0.18
Total No of trades	0.18
No of PL trades in last 12 months	0.17
No of times 90 DPD in last 6 months	0.16
No of PL trades opened in last 6 months	0,12

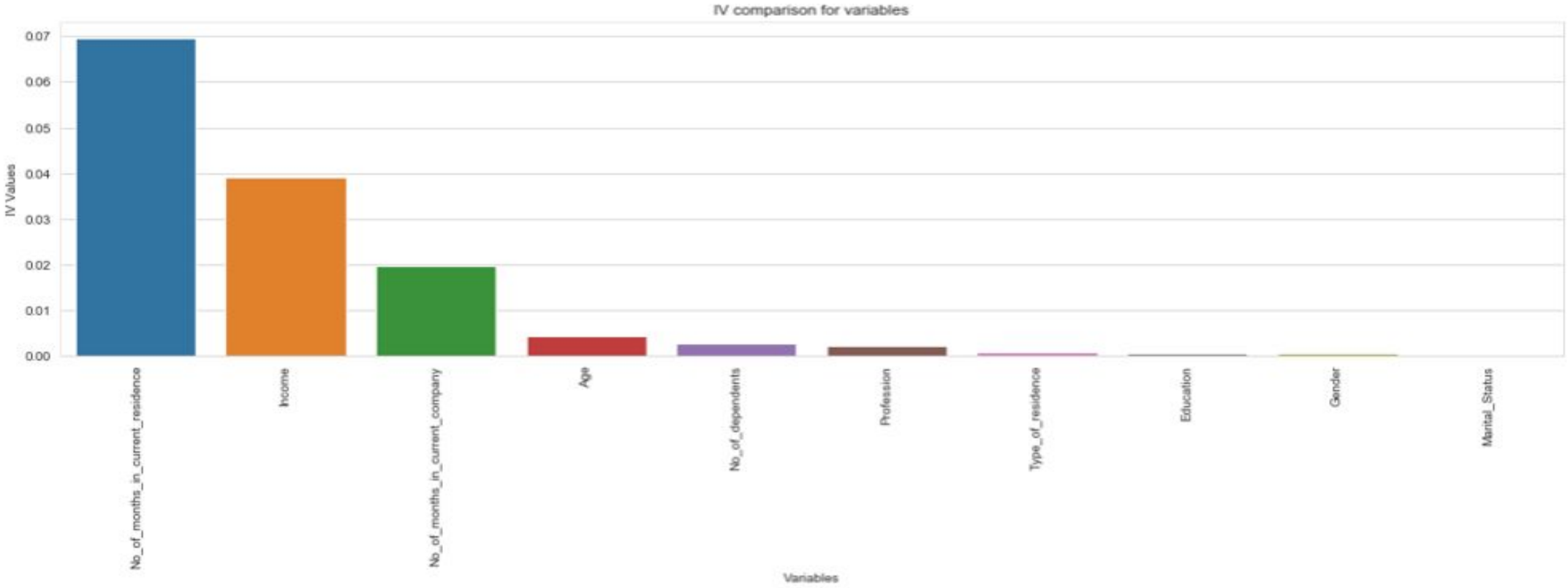
EDA(Exploratory and Data Analysis)

- Default rate w.r.t demographic data = 4.21
- As per data set , problem is classification supervised
- Both Univariate and Bivariate analysis is performed on whole data.

Some of the Observations:

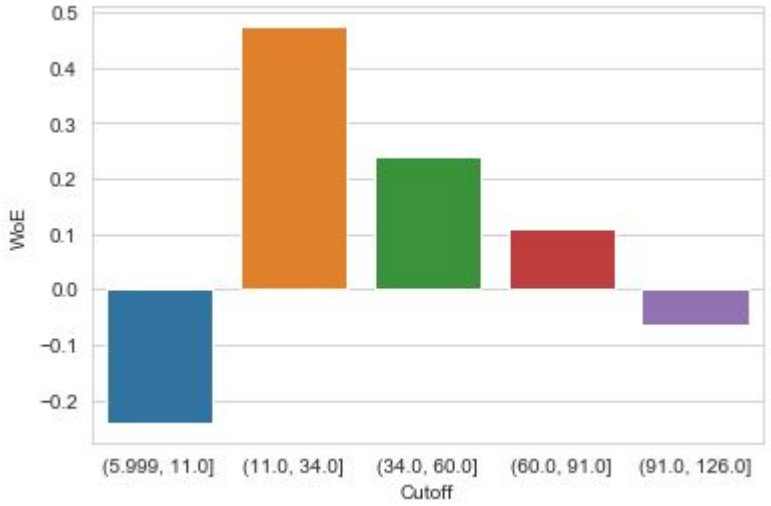
- More Male applicants than female with the same default rate
- More married applicant than single with the same default rate
- More salaried applicant but with same default rate
- No of defaulters are increasing with increase in no of 30/60/90 DPD or worst in last 6 months variable value
- Same is the no of defaulters is increasing with the no of 30/60/90 DPD or worst in last 12 months

Information Values comparison for variables

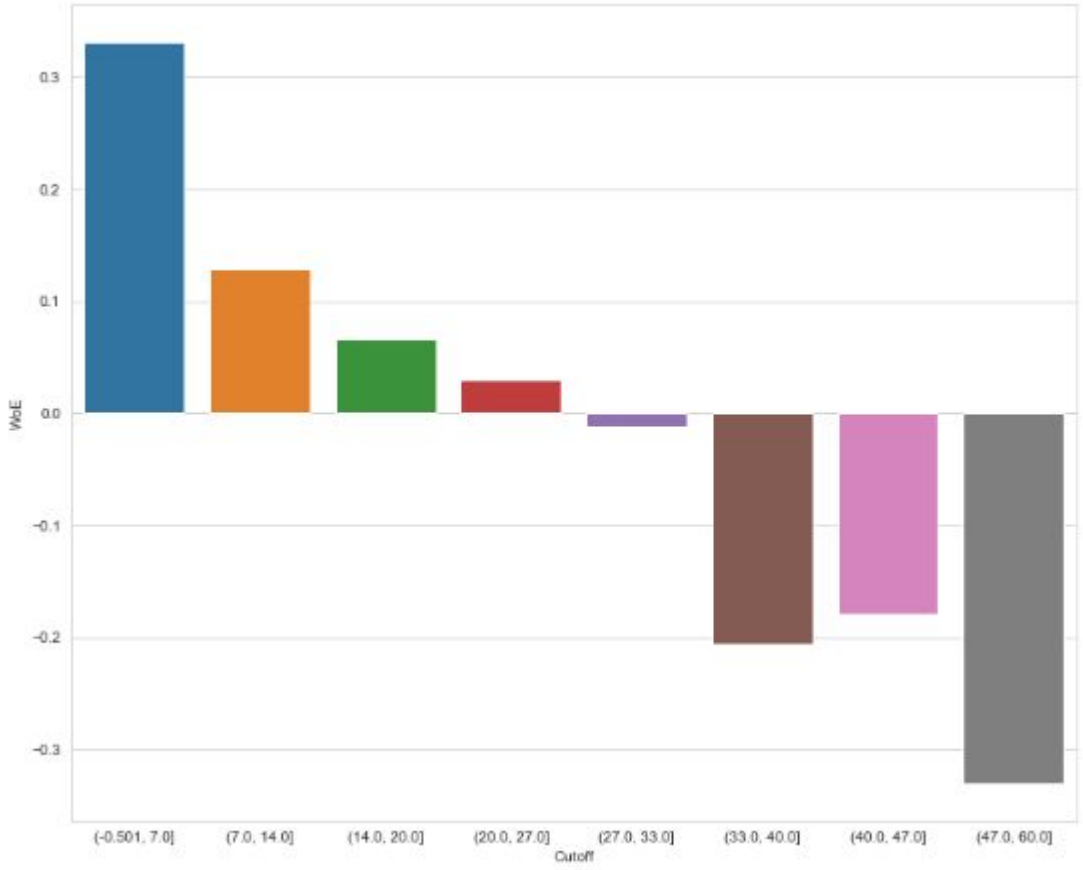


The variable No_of_months_in_current_residence is most significant in getting information about the default rate.

EDA(Cont..)

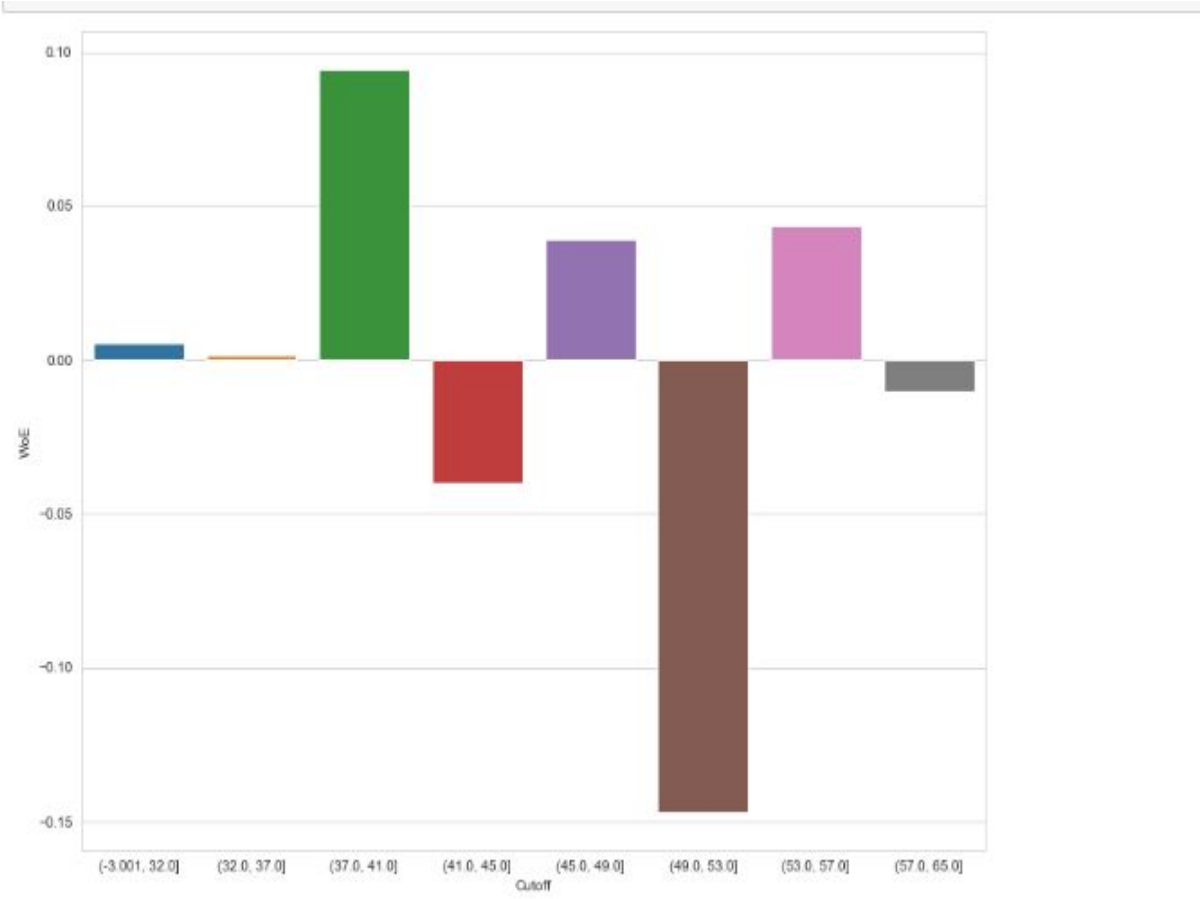


Plot of WOE w.r.t to No of months in current residence



Plot of WOE w.r.t income

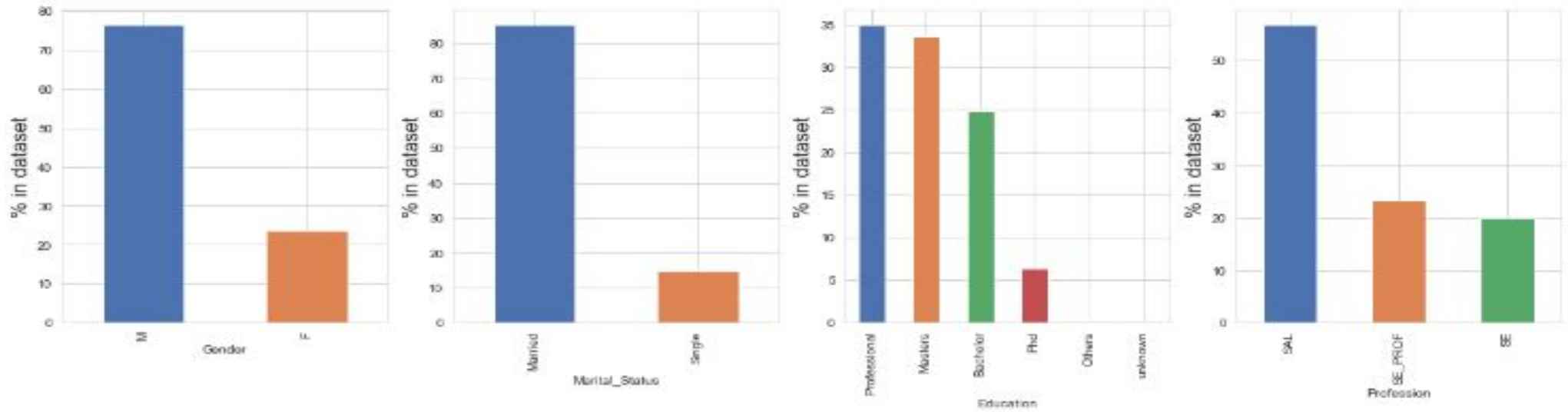
EDA(Cont..)



Plot of Age w.r.t WOE

EDA

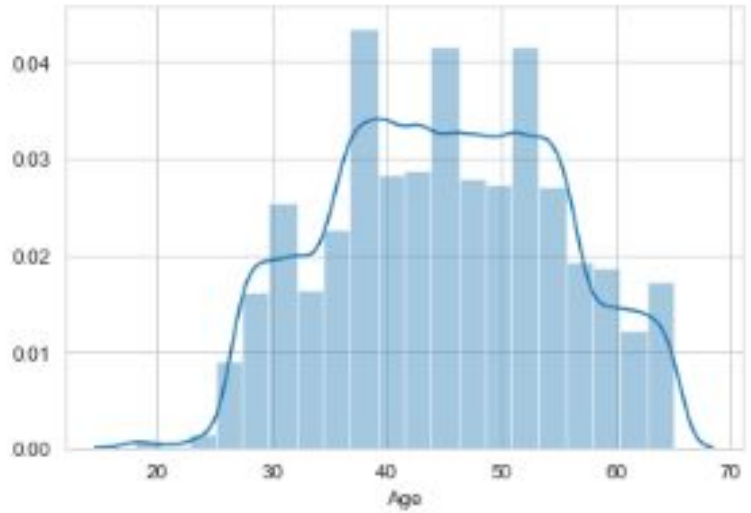
['Gender', 'Marital_Status', 'Education', 'Profession', 'Type_of_residence']



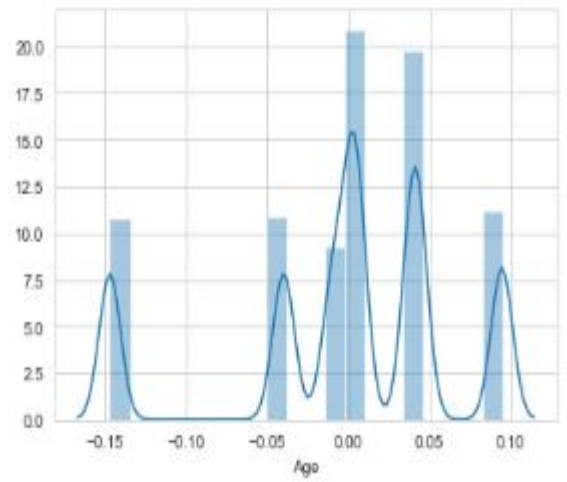
The above plot shows the Percentage of data sets among the variables Gender, Marital_Status, Education, Profession, Type of residence

EDA

Name: Age, dtype: float64

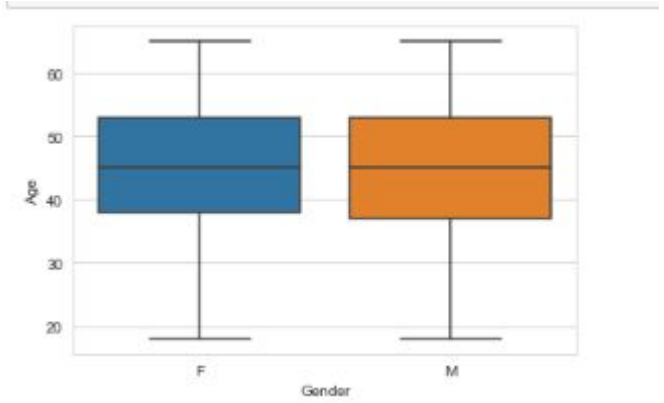


Name: Age, dtype: float64

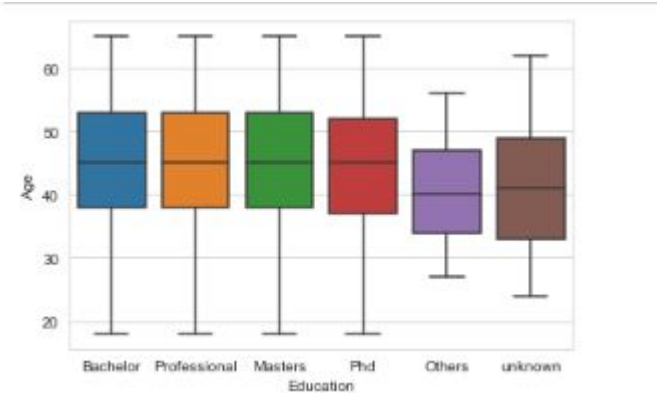


Univariate EDA of the variable age with the raw demographic data and WOE data respectivel

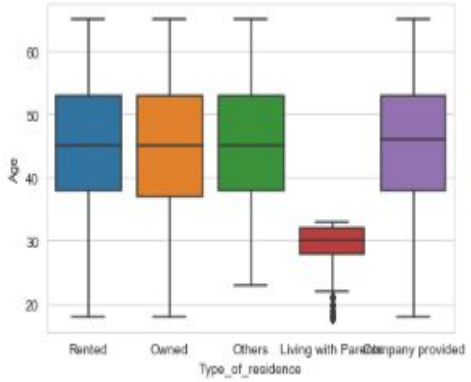
EDA



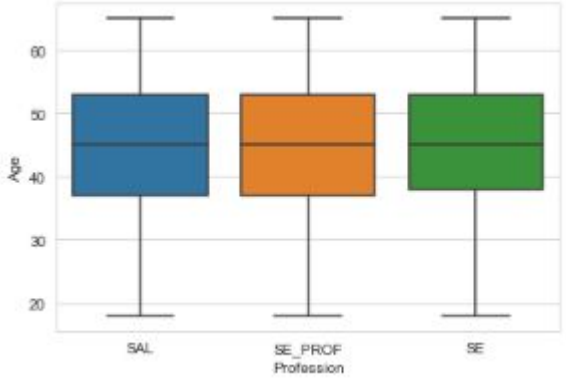
Bivariate analysis of age with Gender



Bivariate analysis of age with Education

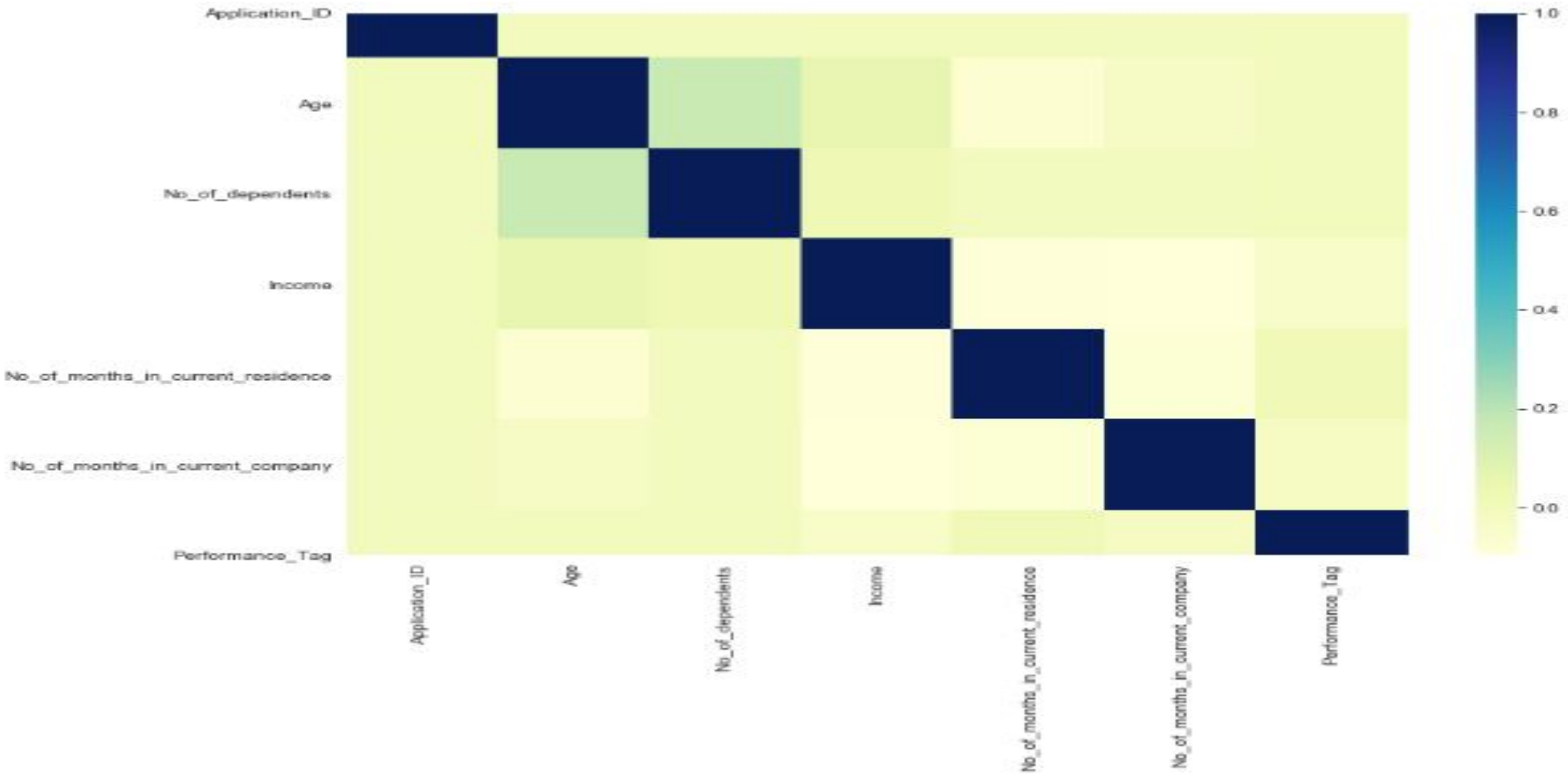


Bivariate analysis of age with type of residence

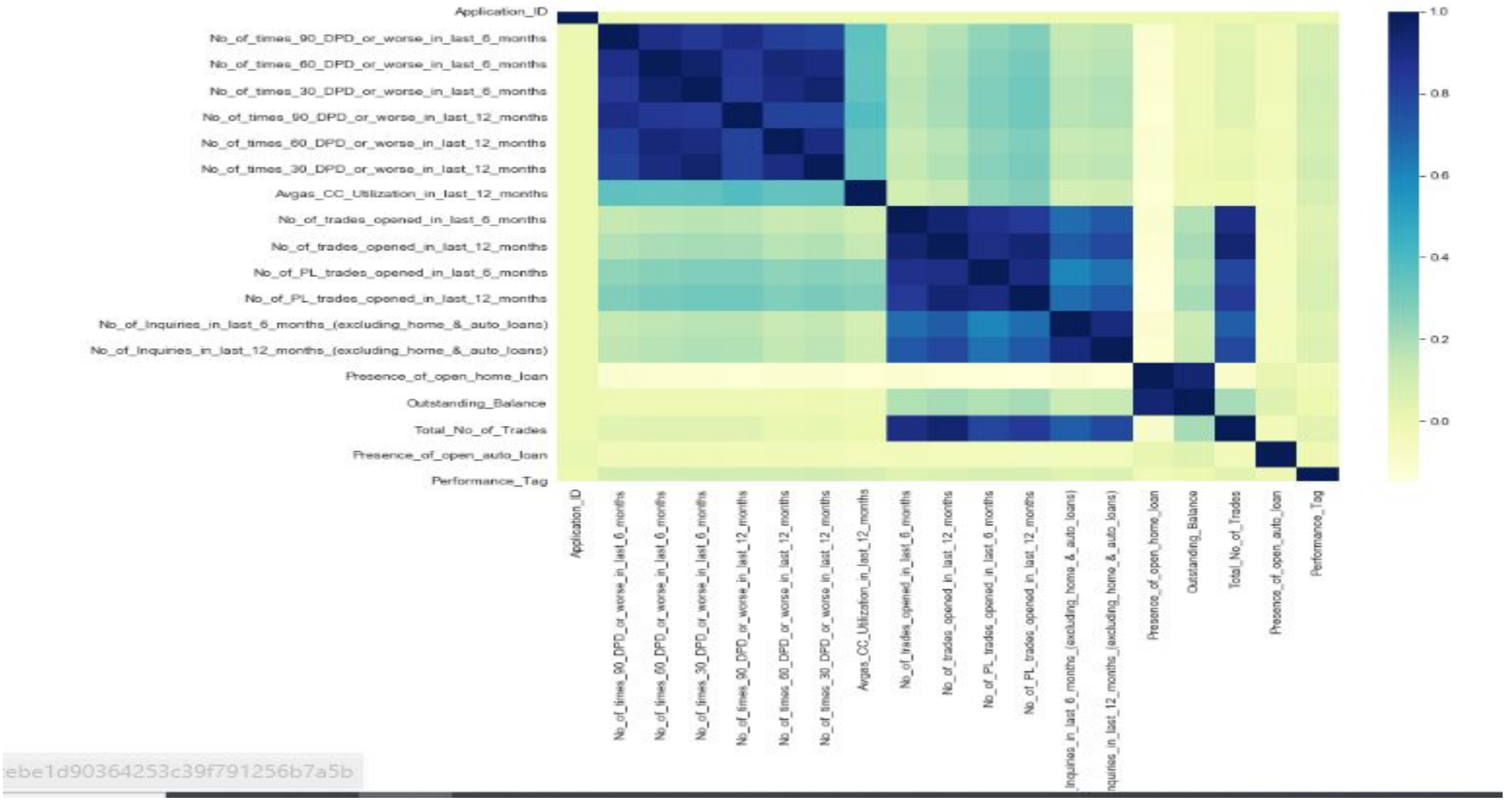


Bivariate analysis of age with profession

EDA(Demographic data)

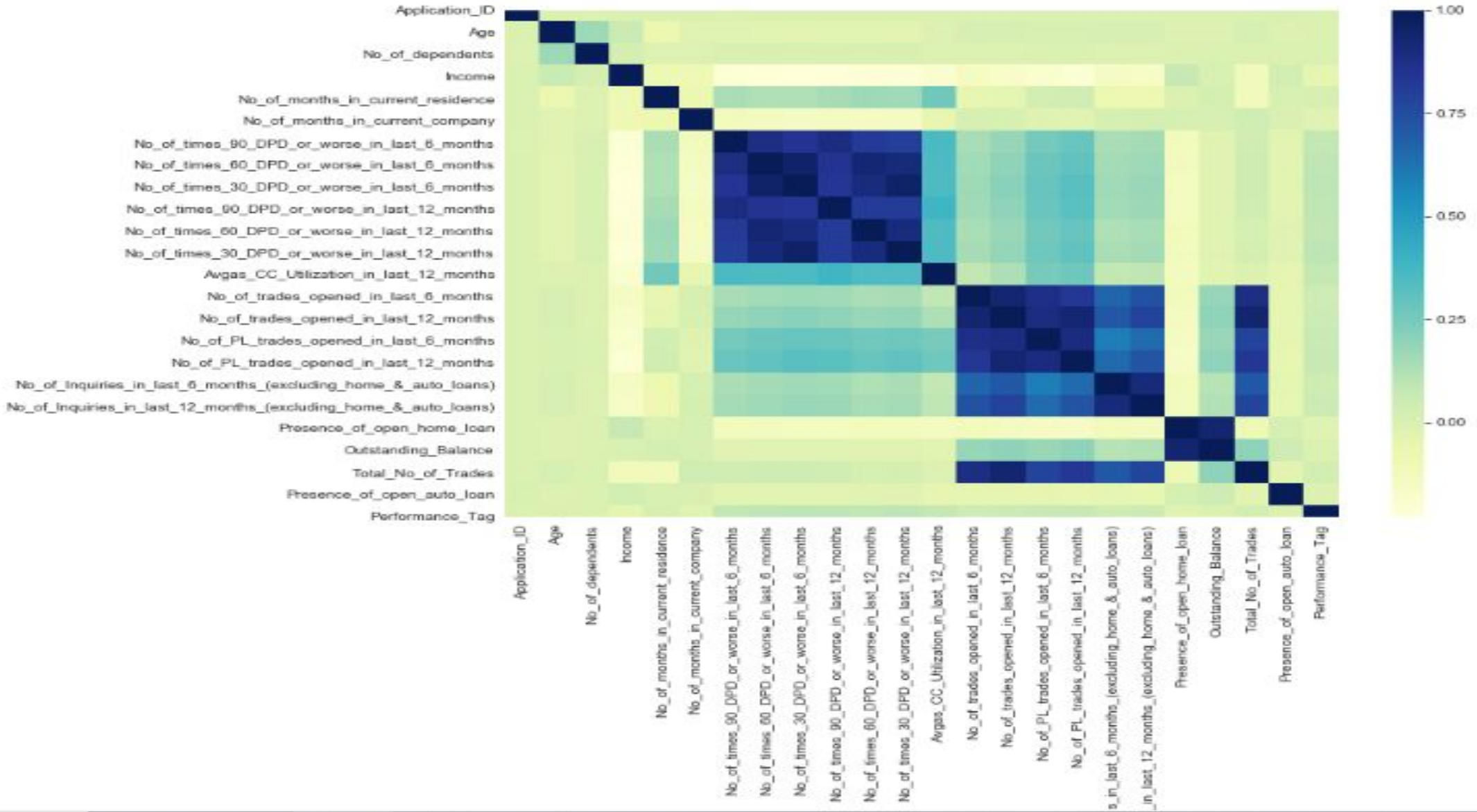


EDA(Credit Bureau Data)



ebe1d90364253c39f791256b7a5b

EDA(Combined Data set)



Model Building(Logistic Regression)

Generalized Linear Model Regression Results

Dep. Variable:	Performance_Tag	No. Observations:	48906
Model:	GLM	Df Residuals:	48895
Model Family:	Binomial	Df Model:	10
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-8494.5
Date:	Mon, 23 Dec 2019	Deviance:	16989.
Time:	12:10:55	Pearson chi2:	4.90e+04
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7657	0.092	-29.979	0.000	-2.946	-2.585
Age	0.0769	0.109	0.702	0.483	-0.138	0.291
Income	-0.7665	0.090	-8.480	0.000	-0.944	-0.589
No_of_months_in_current_residence	0.1963	0.072	2.728	0.006	0.055	0.337
No_of_months_in_current_company	-0.6454	0.146	-4.414	0.000	-0.932	-0.359
Marital_Status_Single	0.0687	0.064	1.080	0.280	-0.056	0.193
Education_Others	0.3696	0.463	0.799	0.424	-0.537	1.276
Education_Phd	-0.0275	0.093	-0.295	0.768	-0.210	0.155
Education_unknown	-0.6649	0.716	-0.928	0.353	-2.069	0.739
Profession_SE	0.1263	0.055	2.318	0.020	0.020	0.233
Type_of_residence_Others	-0.4081	0.508	-0.803	0.422	-1.404	0.588

Model Building(Logistic Regression)

Generalized Linear Model Regression Results

Dep. Variable:	Performance_Tag	No. Observations:	48906
Model:	GLM	Df Residuals:	48896
Model Family:	Binomial	Df Model:	9
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-8494.6
Date:	Mon, 23 Dec 2019	Deviance:	16989.
Time:	12:10:55	Pearson chi2:	4.90e+04
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7674	0.092	-30.061	0.000	-2.948	-2.587
Age	0.0769	0.109	0.702	0.483	-0.138	0.291
Income	-0.7665	0.090	-8.481	0.000	-0.944	-0.589
No_of_months_in_current_residence	0.1964	0.072	2.729	0.006	0.055	0.338
No_of_months_in_current_company	-0.6453	0.146	-4.413	0.000	-0.932	-0.359
Marital_Status_Single	0.0686	0.064	1.079	0.281	-0.056	0.193
Education_Others	0.3713	0.463	0.803	0.422	-0.535	1.278
Education_unknown	-0.6632	0.716	-0.926	0.355	-2.067	0.741
Profession_SE	0.1264	0.055	2.318	0.020	0.020	0.233
Type_of_residence_Others	-0.4081	0.508	-0.803	0.422	-1.404	0.588

Model Building & Evaluation(1)

Based on demographic data set

- **Logistic Regression model**

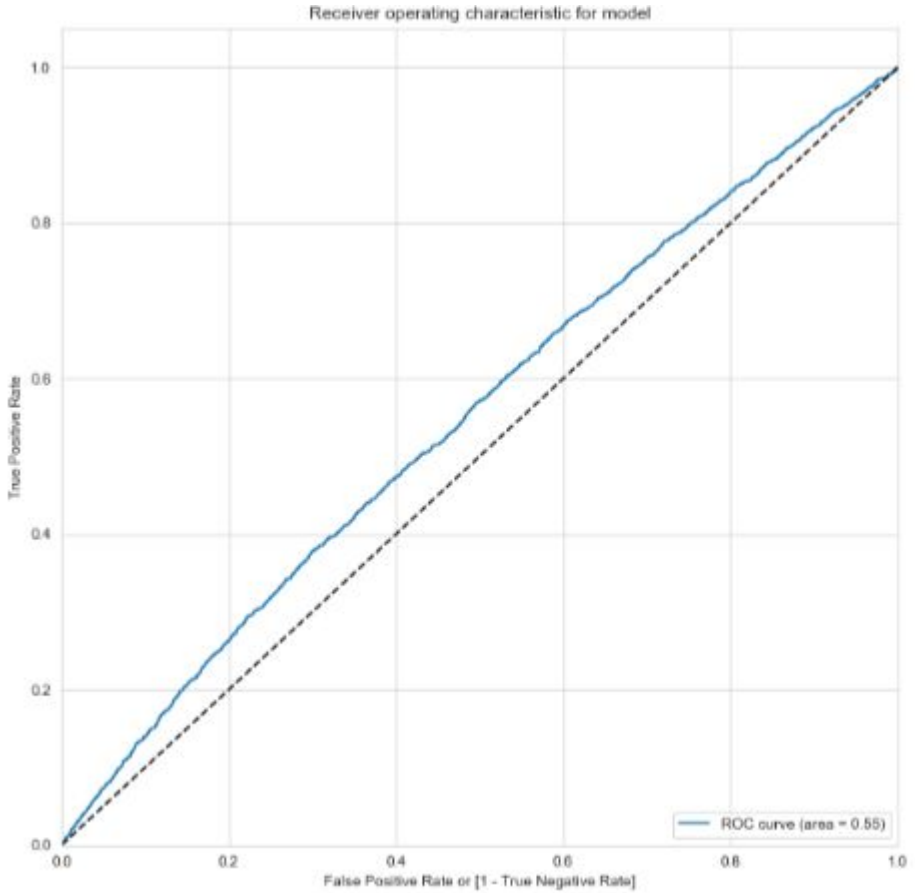
Confusion Matrix

```
[[46843     0]
 [ 2063     0]]
```

Overall Accuracy : 0.957

Precision : 0.0546

Recall : 0.26



ROC Curve based on the probability values generated by the model

Model Building & Evaluation(2)

Random Forest

Max_depth = 4

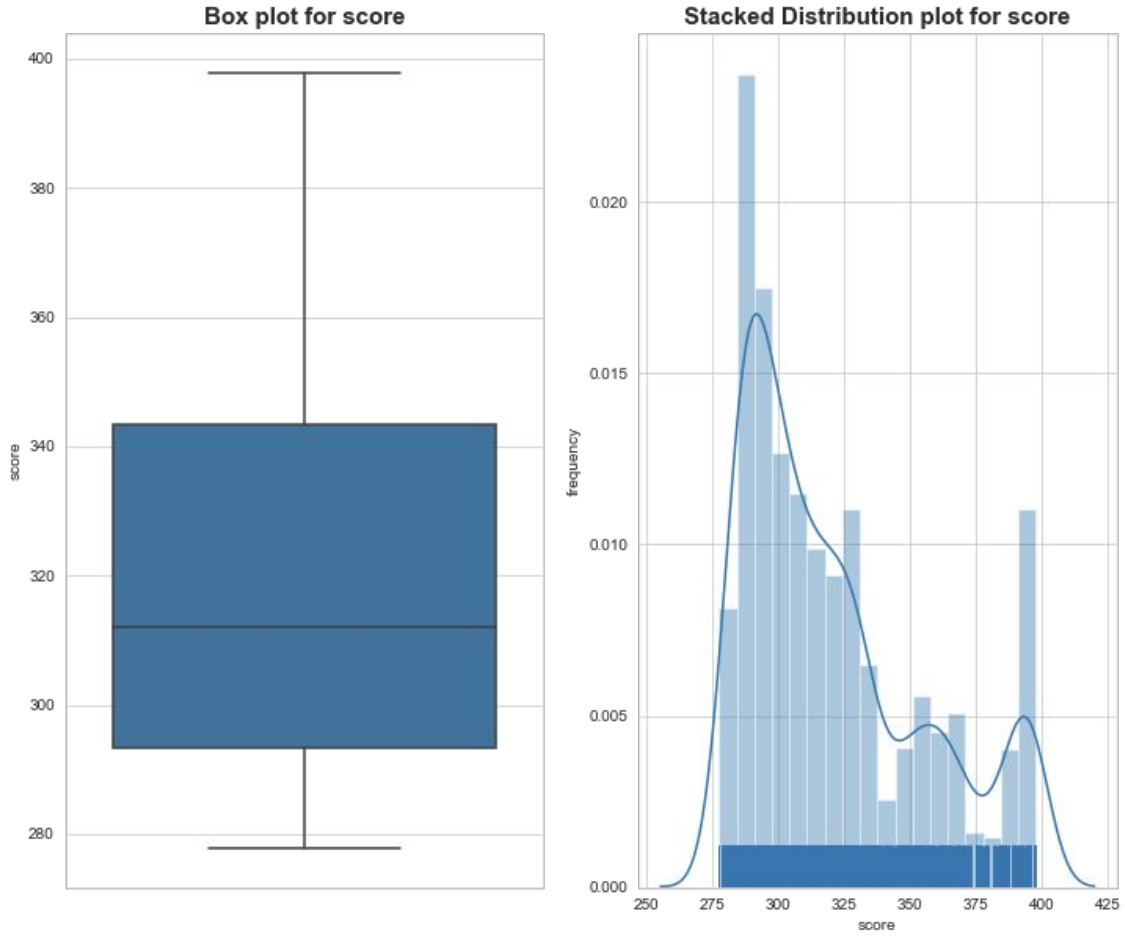
Min_samples_leaf = 350

n_estimators= 1000

Random Forest	
Accuracy	0.957
Precision	0.0
Recall	0.0

	precision	recall	f1-score	support
0	0.96	1.00	0.98	20077
1	0.00	0.00	0.00	884
accuracy			0.96	20961
macro avg	0.48	0.50	0.49	20961
weighted avg	0.92	0.96	0.94	20961

Score card box plot



Model Building & Evaluation(3)

Final Model Chosen : Random forest model based on the WOE transformed data

Reason to choose:

- Good recall values on the test data
- Model is expected not to overfit
- Model was already excluding all the rejected applicants
- Model seems to be stable

Financial Benefit

- Out of total 71295 applicants, 69870 were given credit cards
- 71295-69870= 2948 Applicants that made credit loss
- Assumption : On an average on a default of customer,lets assume we are in a loss of INR 80,000/-
- 80,000*2948 = 240 Million Approx(Total credit loss)
- Potential loss prevented using the model (given the recall value is 68%)

= 0.68 * 240 = 164 Million Approx

- Loss reduced by 240-164= 76 Million

Thank YOU