CREDEX

BFS CAPSTONE PROJECT – SEP 2018

RAM MUMMADAVARAPU

RISHI KANT MALVIYA

SARATHBABU SANKARAN

VIDHYA NAIR

BUSINESS UNDERSTANDING

- OBJECTIVE
  - CredX, a leading credit card provider is facing in increase credit loss
  - The objective is to mitigate credit risk by ‘acquiring the right customers’
- STRATEGY
  - Identify customers that present a credit risk using demographic and credit bureau data
  - Offer credit cards to only those customers that are deemed as less risky

DATA ASSESSMENT – UNDERSTANDING THE DATA AND DATA QUALITY

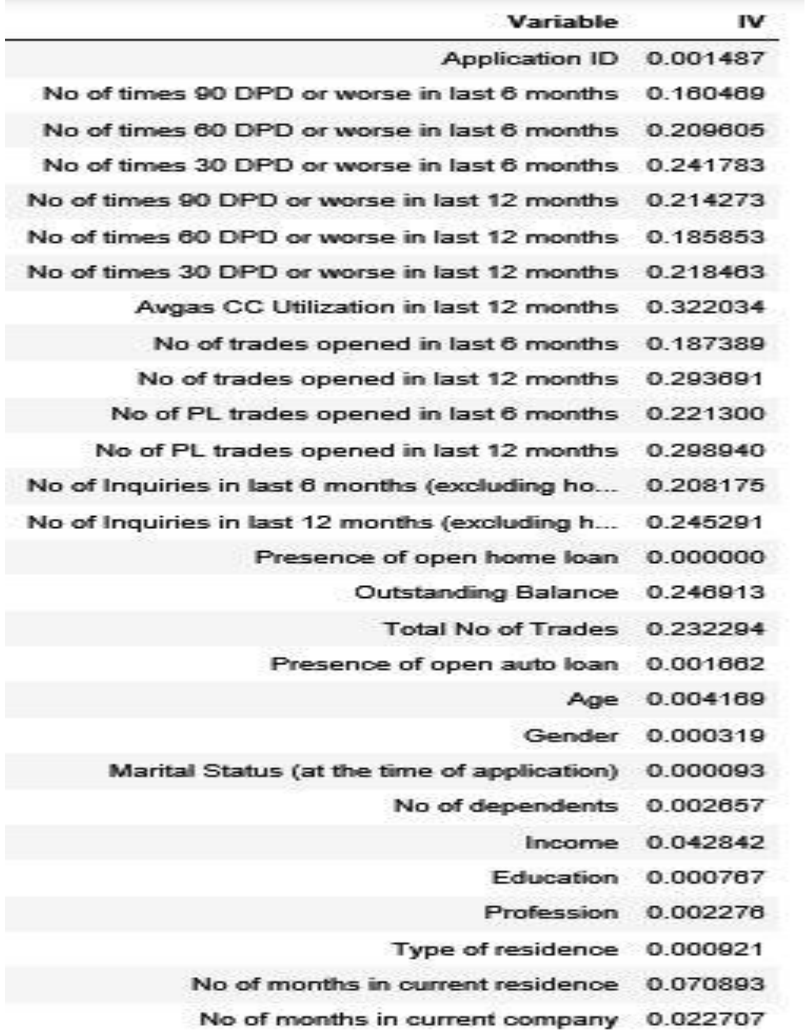
- UNDERSTANDING THE DATA
  - 2 datasets are available – Customer Demographic Data and Credit Bureau Data
  - 71289 unique customer data is present
  - The dependent variable is Performance Tag
  - The variable on which the datasets will be merged is Application ID
- DATA QUALITY
  - Duplicate Application ID entries are removed
  - Entries that have already been identified as defaulters or don't have any credit bureau data are also removed
  - 69867 unique customer data is ready for analysis
  - Outliers have been addressed

EXPLORATORY DATA ANALYSIS

- Univariate Analysis
  - Median income is 27 and median age is 45 years
  - ~24% of the customers are female, ~85% of the customers are married and 75% are living in rented residence
  - 57% are Salaried, 23% are Professionals and rest are Self employed
  - 28% customers defaulted 30 DPD or worse in last 6 months atleast once, 26% customers defaulted 60 DPD or worse in last 6 months atleast once, 22% customers defaulted 90 DPD or worse in last 6 months atleast once
  - 36% customers defaulted 30 DPD or worse in last 12 months atleast once, 34% customers defaulted 60 DPD or worse in last 12 months atleast once, 28% customers defaulted 90 DPD or worse in last 12 months atleast once
- Bivariate Analysis
  - Demographic data was not able to give very clear inferences as the number of defaulters were higher in large populations such as male customers, married customers, salaried customers etc.
  - Bureau dataset variables such as DPD data and trades data are better at predicting default behavior

DATA PREPARATION

- Information Value helps in reducing Dimension
- Weight of Evidence shows the effect of each independent variable on the dependent variable
- Data Balancing is carried out using SMOTE

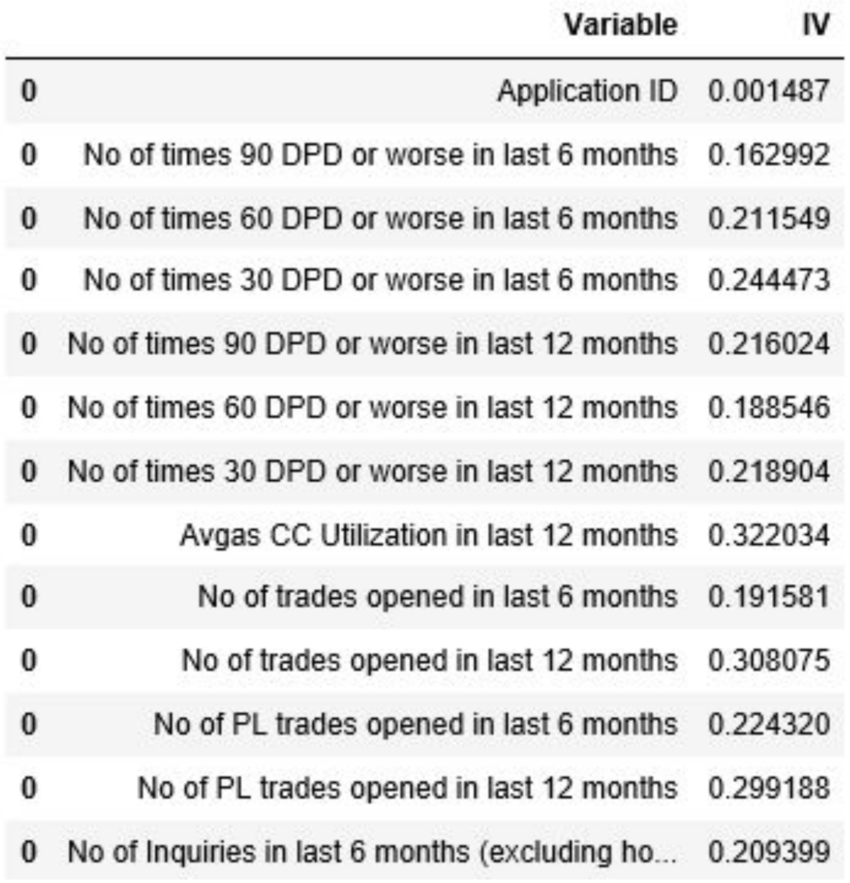


Variable	IV
Application ID	0.001487
No of times 90 DPD or worse in last 6 months	0.160469
No of times 60 DPD or worse in last 6 months	0.209605
No of times 30 DPD or worse in last 6 months	0.241783
No of times 90 DPD or worse in last 12 months	0.214273
No of times 60 DPD or worse in last 12 months	0.185853
No of times 30 DPD or worse in last 12 months	0.218463
Avgas CC Utilization in last 12 months	0.322034
No of trades opened in last 6 months	0.187389
No of trades opened in last 12 months	0.293691
No of PL trades opened in last 6 months	0.221300
No of PL trades opened in last 12 months	0.298940
No of Inquiries in last 6 months (excluding ho...	0.208175
No of Inquiries in last 12 months (excluding h...	0.245291
Presence of open home loan	0.000000
Outstanding Balance	0.246913
Total No of Trades	0.232294
Presence of open auto loan	0.001662
Age	0.004169
Gender	0.000319
Marital Status (at the time of application)	0.000093
No of dependents	0.002657
Income	0.042842
Education	0.000767
Profession	0.002276
Type of residence	0.000921
No of months in current residence	0.070893
No of months in current company	0.022707

WOE and IV Analysis

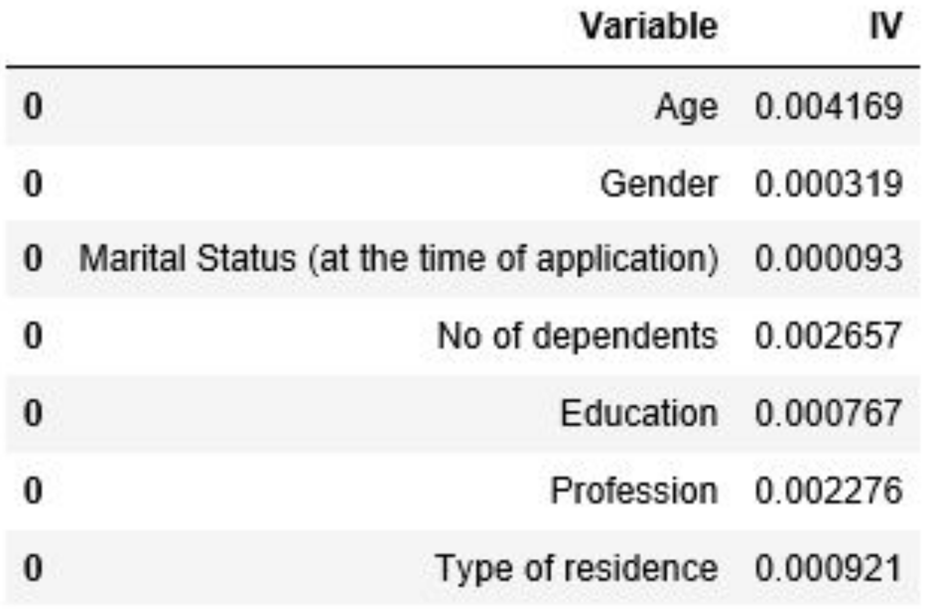
- ❑ WOE and IV values are calculated for each of the attributes using information and python functions.
- ❑ Attributes for which WOE values that were made by default using the python functions and were not monotonically changing across bins, coarser bins were made by decreasing the number of bins until monotonic behavior is observed across bins .
- ❑ For 9 variables with Missing values, the variable values were replaced by their corresponding WOE values.
- ❑ From the IV values we can conclude that parameters in the demographic data don't play much significant role in prediction and most of the significant variables are from Credit Bureau data.
- ❑ Top 12 Variables with IV value of 0.1 to 0.3 has medium predictive power and are considered significant. There is no variable with strong predictive power.

Top 12 variables with highest IV values



	Variable	IV
0	Application ID	0.001487
0	No of times 90 DPD or worse in last 6 months	0.162992
0	No of times 60 DPD or worse in last 6 months	0.211549
0	No of times 30 DPD or worse in last 6 months	0.244473
0	No of times 90 DPD or worse in last 12 months	0.216024
0	No of times 60 DPD or worse in last 12 months	0.188546
0	No of times 30 DPD or worse in last 12 months	0.218904
0	Avgas CC Utilization in last 12 months	0.322034
0	No of trades opened in last 6 months	0.191581
0	No of trades opened in last 12 months	0.308075
0	No of PL trades opened in last 6 months	0.224320
0	No of PL trades opened in last 12 months	0.299188
0	No of Inquiries in last 6 months (excluding ho...	0.209399

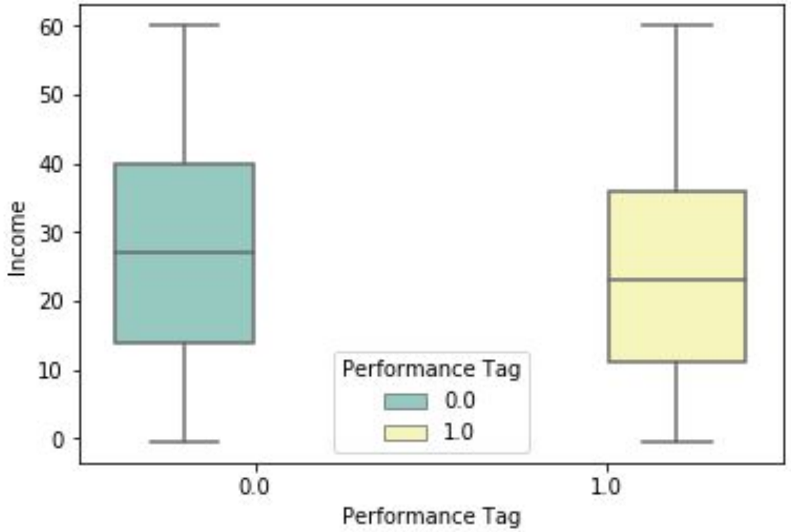
Variables name with less than .02 in Democratic data



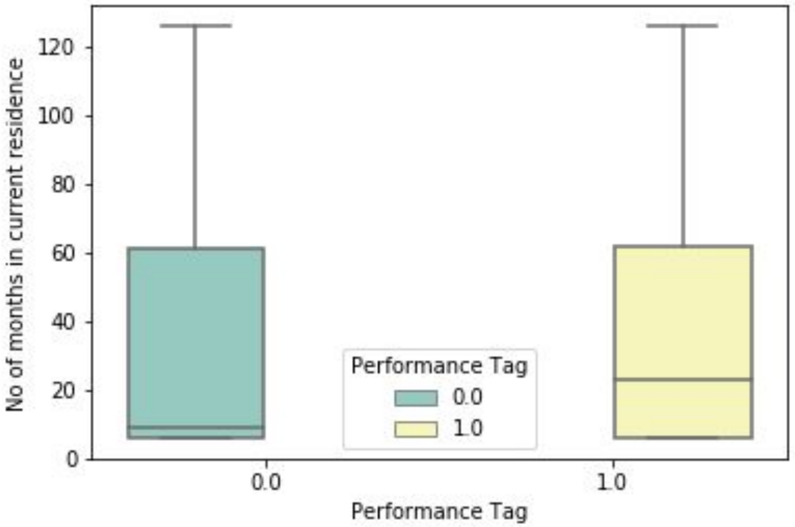
	Variable	IV
0	Age	0.004169
0	Gender	0.000319
0	Marital Status (at the time of application)	0.000093
0	No of dependents	0.002657
0	Education	0.000767
0	Profession	0.002276
0	Type of residence	0.000921

Exploratory Data Analysis

Insights derived by EDA : Both univariate and bivariate plots are made to get better insights of all variables of 2 datasets.



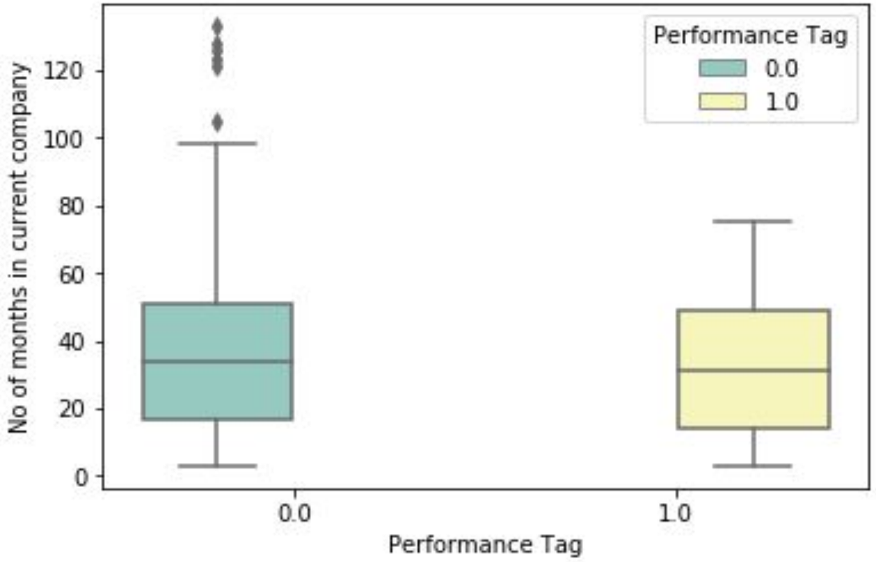
The median values for income of defaulters are lower than that of non-defaulters

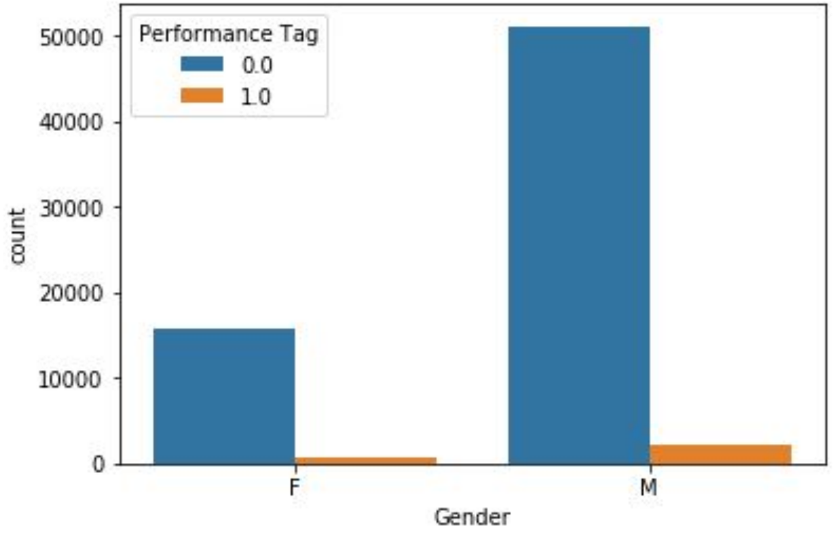


The median No.of.months.in.current.residence of non-defaulters are lower than that of defaulters.

Insights derived by EDA

Counts in the y axis of Plots in bivariate analysis of categorical variables were normalized so that they have equal heights.



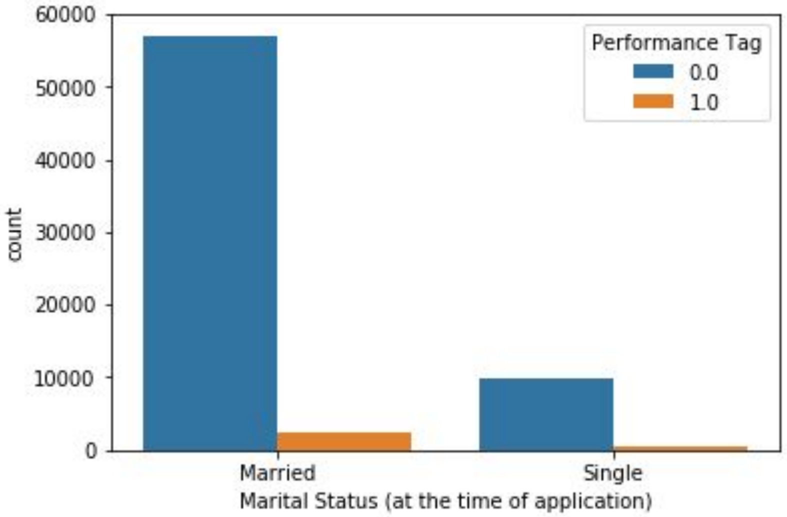


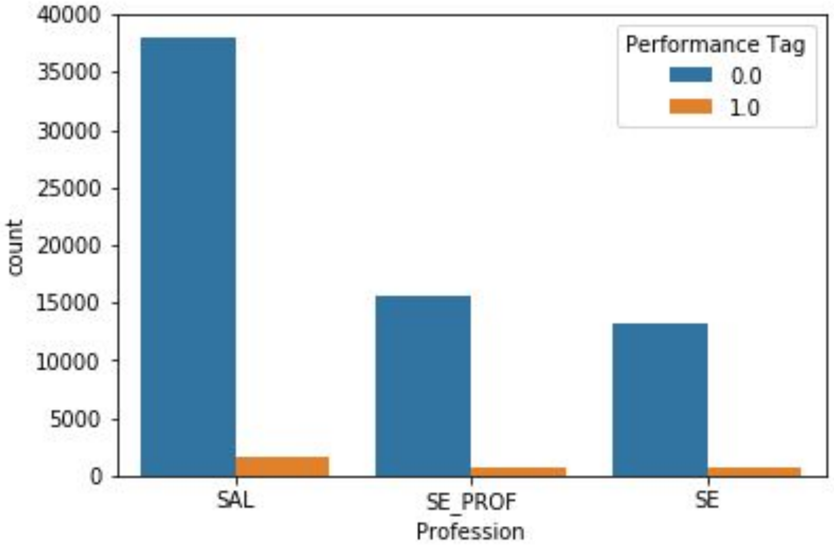
The median No of months in current Company of non defaulters is slightly lower than that of defaulters.

There are more male applicants than female applicants but there is no difference in default rates.

Insights derived by EDA

Insights derived by EDA : Both univariate and bivariate plots are made to get better insights of all variables of 2 datasets.

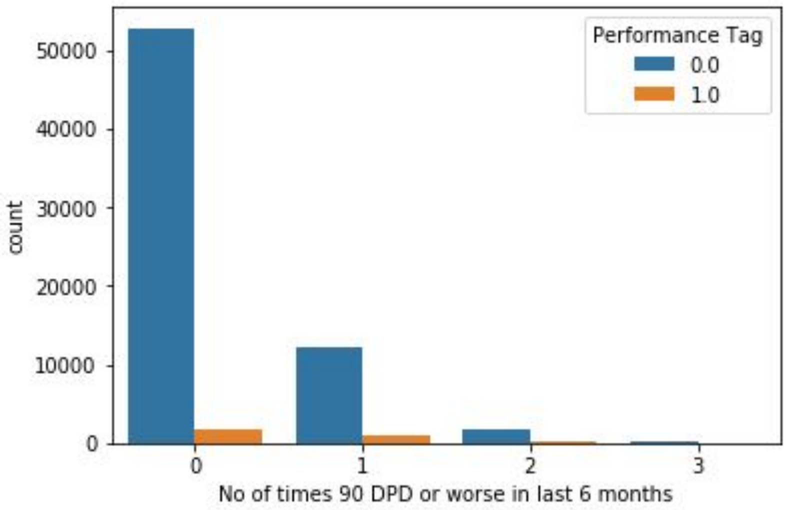


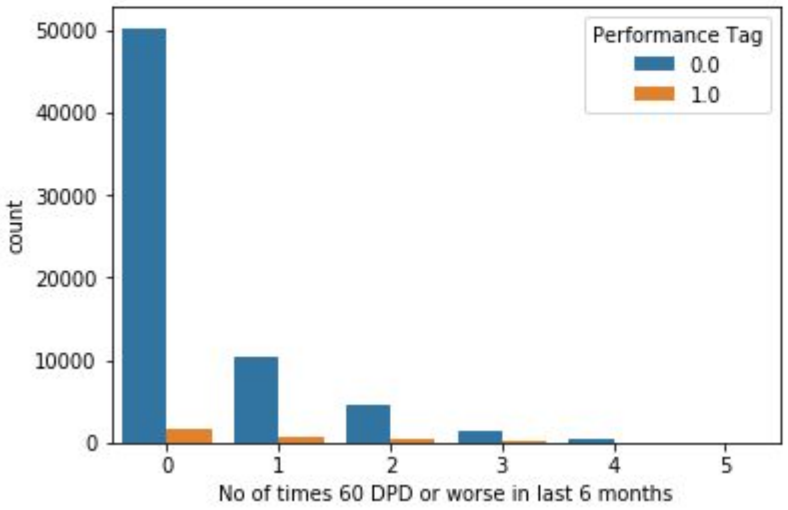


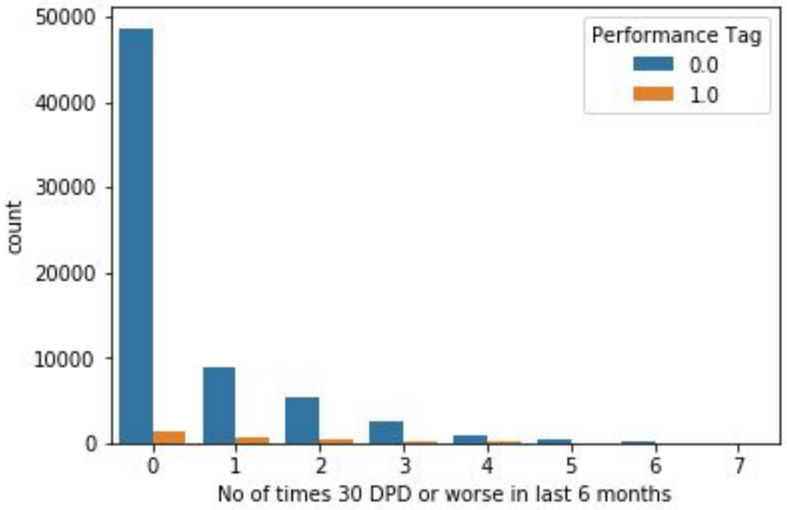
There are more **married** applicants than single applicants but there is no difference in default rates

There are more applicants whose **profession** is SAL but there is no difference in default rates.

Insights derived by EDA

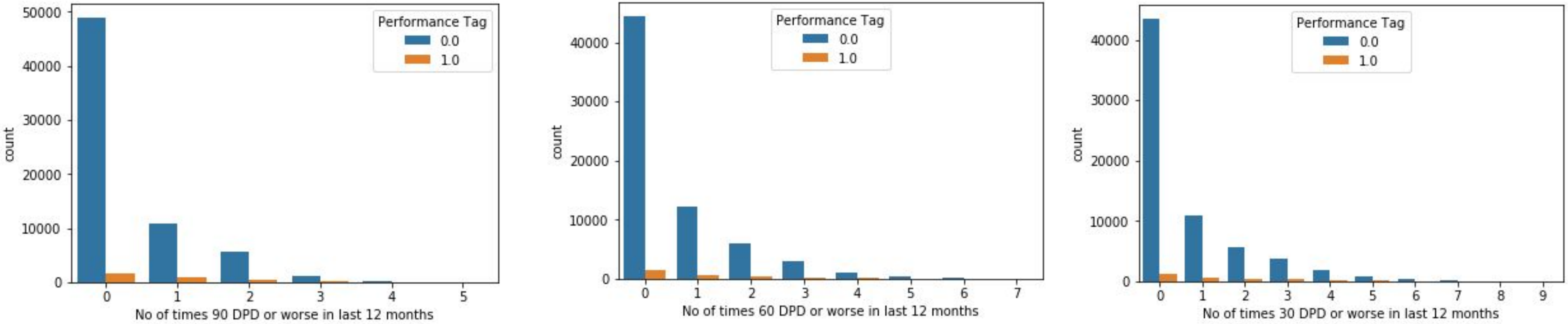






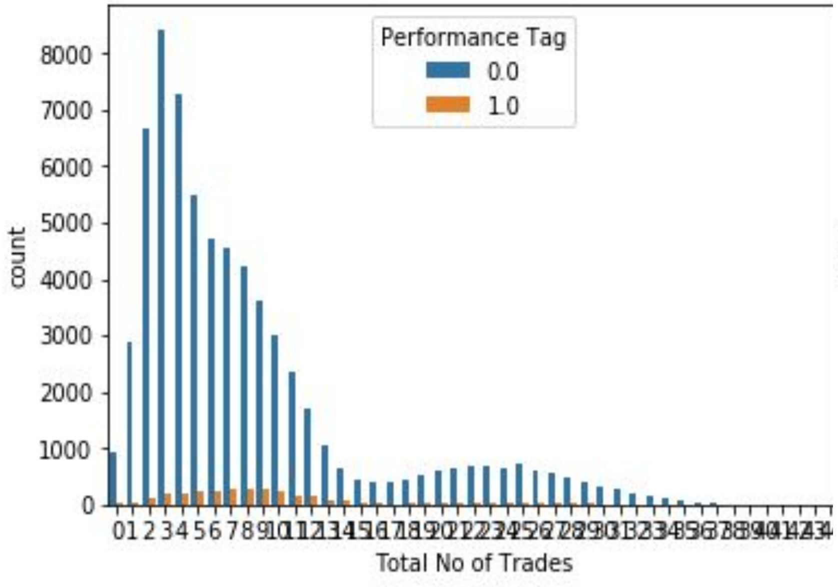
Percentage of defaulters is increasing with increase in Number of 30/60/90 DPD or worse in last 12 months variable values. Hence these variables can be important predictors.

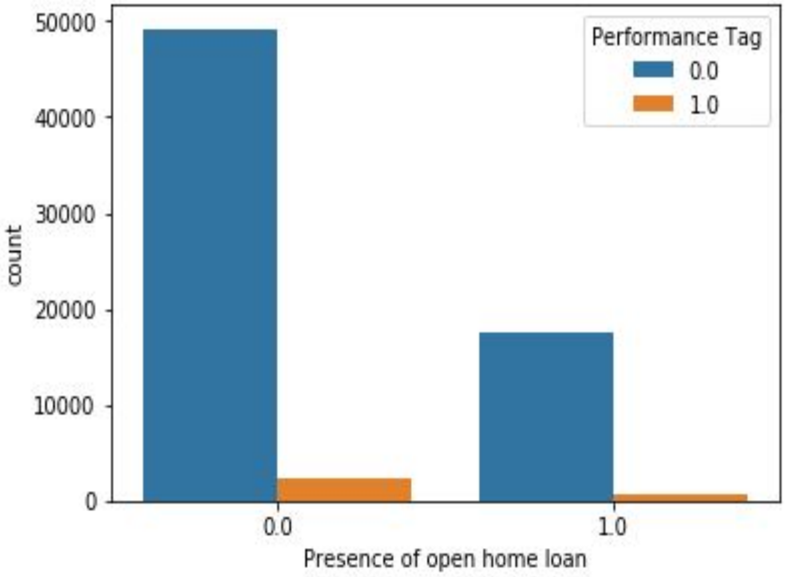
Insights derived by EDA

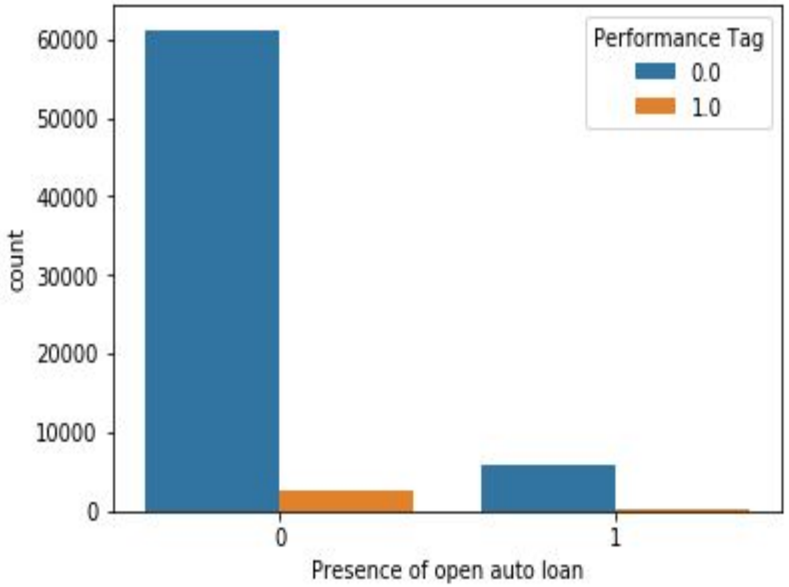


Percentage of defaulters is increasing with increase in Number of 30/60/90 DPD or worse in last 12 months variable values. Hence these variables can be important predictors.

Insights derived by EDA







No appropriate pattern found in no of defaulters in Total No of Trades, Presence of open home loan, Presence of open auto loan

Conclusions from EDA:

1. Demographic variables are not very good predictors of defaulting. Only below 3 variables seems significant.

Income
No of months in current residence
No of months in current company

.credit bureau dataset has many variables which seems like good predictors of defaulters .2

No.of.times.90.DPD.or.worse.in.last.6.months
No.of.times.60.DPD.or.worse.in.last.6.months
No.of.times.30.DPD.or.worse.in.last.6.months
No.of.times.90.DPD.or.worse.in.last.12.months
No.of.times.60.DPD.or.worse.in.last.12.months
No.of.times.30.DPD.or.worse.in.last.12.months
No.of.trades.opened.in.last.6.months
No.of.PL.trades.opened.in.last.6.months
No.of.PL.trades.opened.in.last.12.months

3. There is no correlation between numeric variables of demographic dataset.
4. Few numeric variables of Credit bureau dataset show strong positive correlation with other variables.

.The 6 variables – No of times 90/60/30 DPD.or.worse.in.last.6/12 months are highly correlated among themselves
.No of enquiries in last 6 months/12 months excluding home, auto loan variables are highly correlated
.No. of trade opened in 6/12 monts total number of trades, no of PL trades in 6/12 months are correlated

MODEL BUILDING APPROACH

OUTLIER TREATMENT:
We were using WOE for data preparation that has been take and care all outliers in the data set.

DATA SCALING:
Scaling is performed for all variables except Application ID and performance tag to standardize the data into common scale.

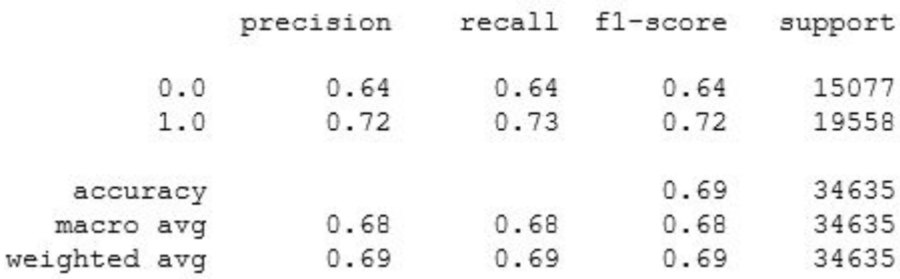
DATA SPLIT:
The final dataset is split into Train and Test in 70:30 ratio for model building. • All models are trained on training datasets and regularization was done by tuning of hyper parameters with cross validation on validation datasets. • All the models are tested on test datasets that were kept separate from training and validation datasets.

DATA SAMPLING:

- The given data is highly imbalanced. We have sampled data using SMOTEENNpackage for balancing the training data sets.
- The cutoff value for the probability of default was chosen such that model evaluation metrics like accuracy ,sensitivity and specificity were almost equal to each other.
- Logistic Regression was built by iteratively removing using these two algorithms

MODEL BUILDING APPROACH

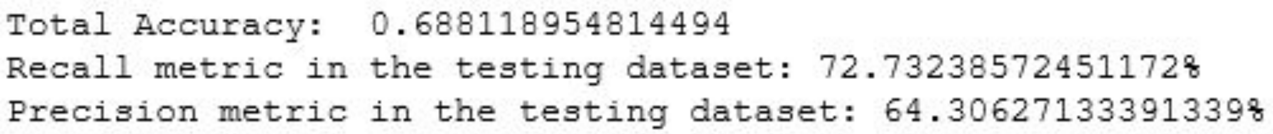
Logistic Regression:



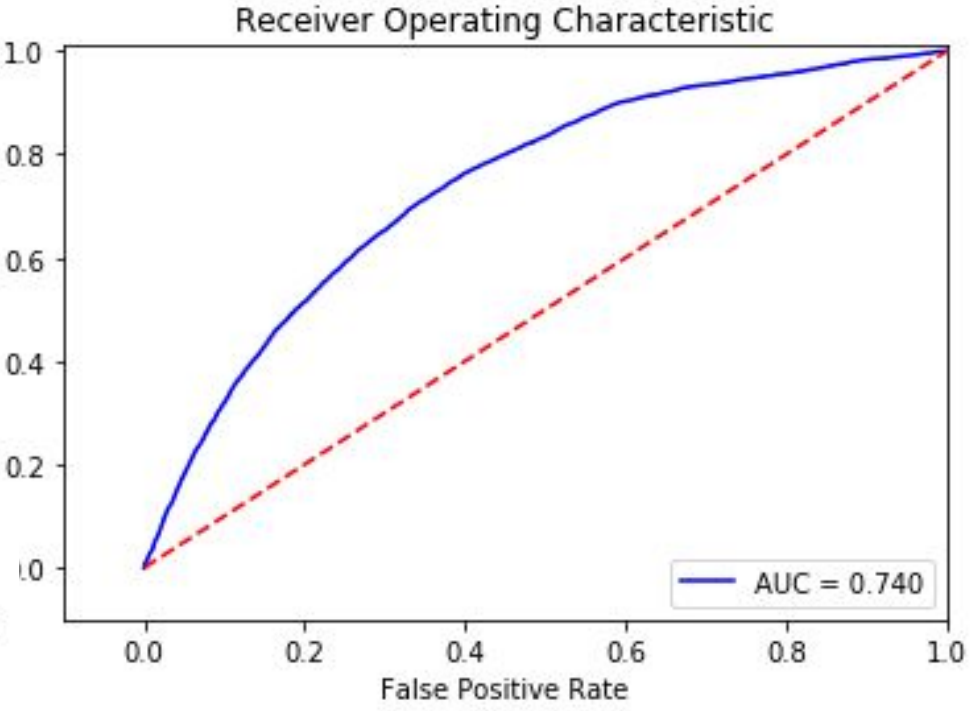
```
              precision    recall  f1-score   support

         0.0       0.64      0.64      0.64     15077
         1.0       0.72      0.73      0.72     19558

    accuracy                           0.69     34635
   macro avg       0.68      0.68      0.68     34635
weighted avg       0.69      0.69      0.69     34635
```



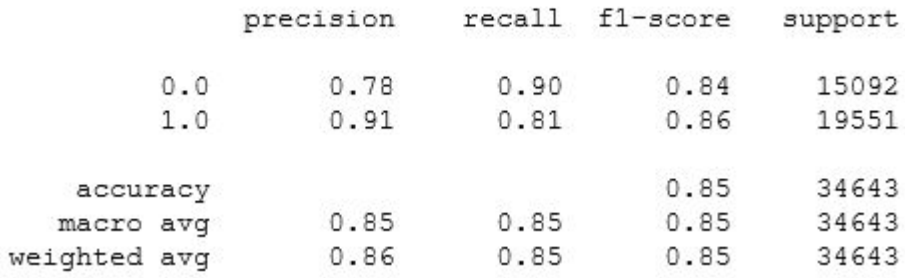
```
Total Accuracy:  0.688118954814494
Recall metric in the testing dataset: 72.73238572451172%
Precision metric in the testing dataset: 64.30627133391339%
```



the ROC Curve for Train Data

MODEL BUILDING APPROACH

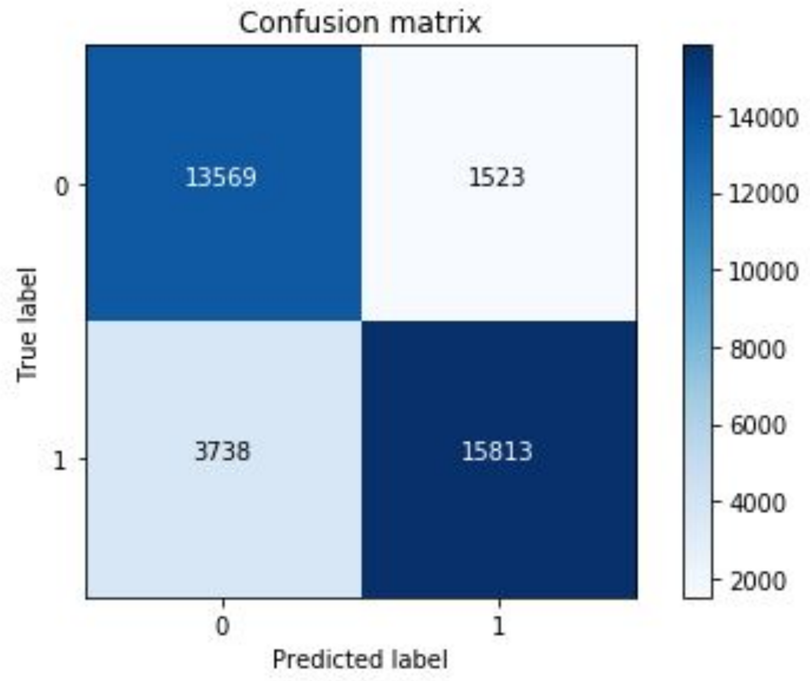
Decision tree analysis:

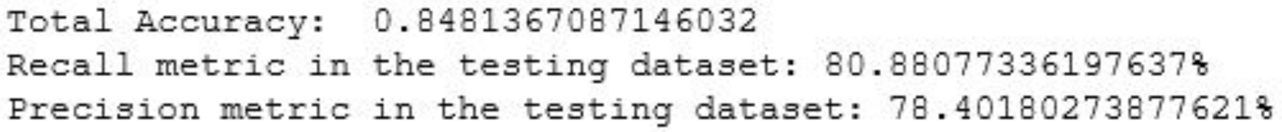


```
              precision    recall  f1-score   support

         0.0       0.78      0.90      0.84     15092
         1.0       0.91      0.81      0.86     19551

    accuracy                           0.85     34643
   macro avg       0.85      0.85      0.85     34643
weighted avg       0.86      0.85      0.85     34643
```

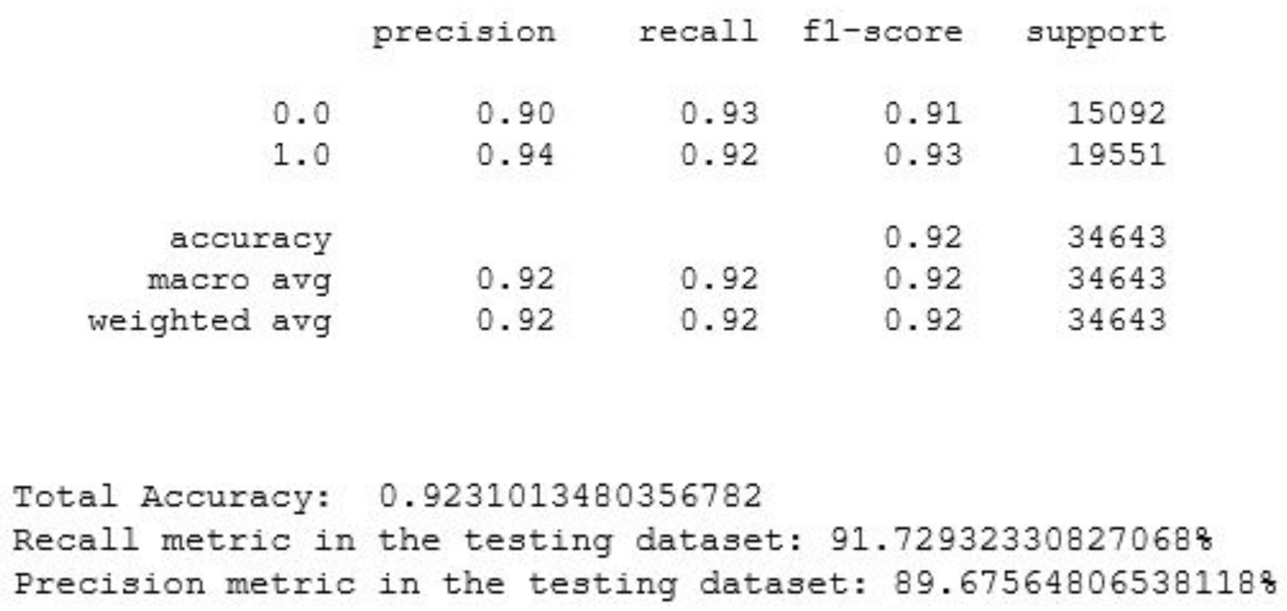




```
Total Accuracy:  0.8481367087146032
Recall metric in the testing dataset: 80.88077336197637%
Precision metric in the testing dataset: 78.40180273877621%
```

MODEL BUILDING APPROACH

Random Forest Classifier :

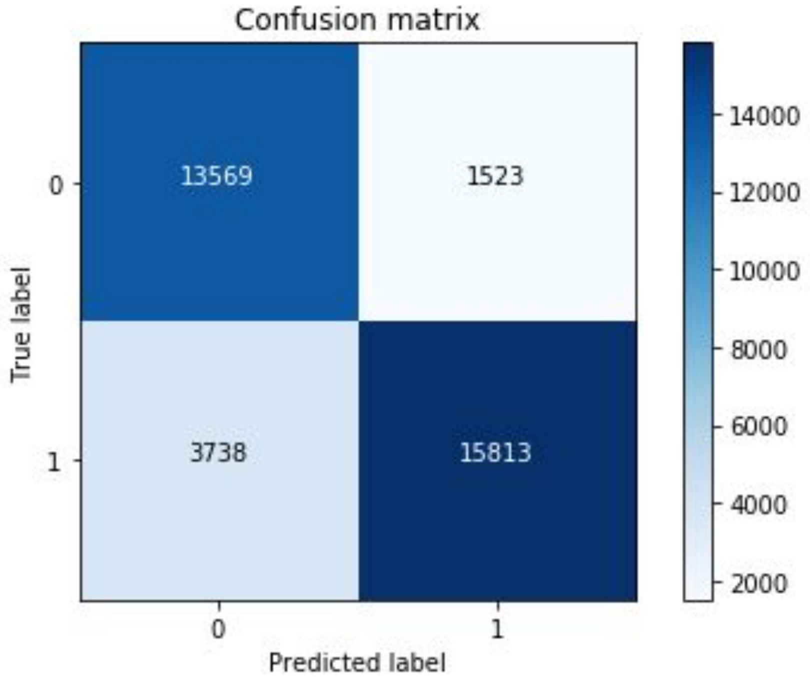


```
              precision    recall  f1-score   support

         0.0       0.90      0.93      0.91     15092
         1.0       0.94      0.92      0.93     19551

    accuracy                           0.92     34643
   macro avg       0.92      0.92      0.92     34643
weighted avg       0.92      0.92      0.92     34643


Total Accuracy:  0.9231013480356782
Recall metric in the testing dataset: 91.72932330827068%
Precision metric in the testing dataset: 89.67564806538118%
```



Application Scorecard

Final application scorecard was made using the Logistic regression model on the entire dataset in “Performance Tag” in 1425 records.

The logistic regression model was chosen since its evaluation metrics were comparable to other models as well it’s an easily interpretable simple model.

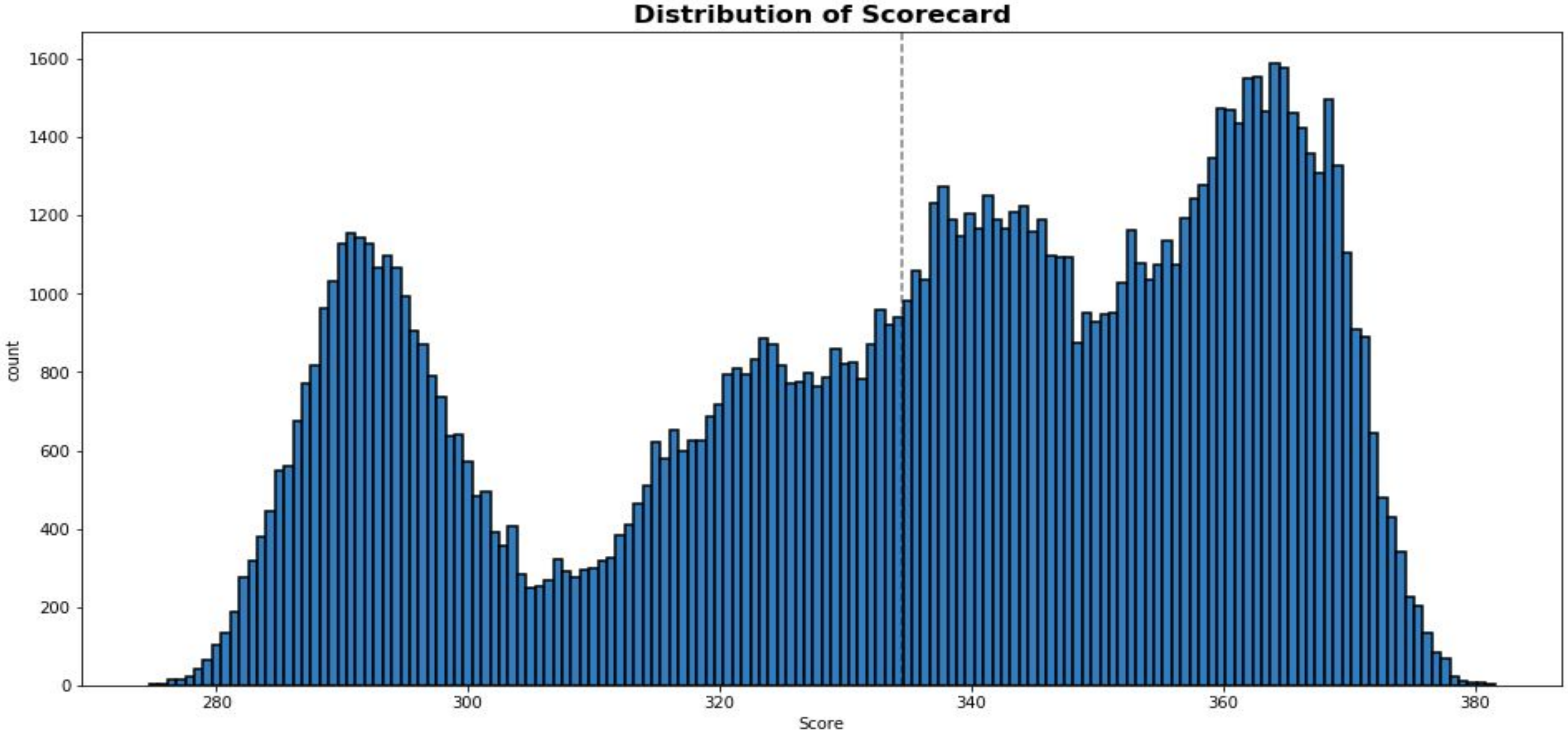
The scorecard was made using the following steps:

1. Application score card was made with odds of 10 to 1 being a score of 400. Score increases by 20 points for doubling odds.
2. Probability of default for all applicants were calculated
3. Odds for good was calculated. Since the probability computed is for rejection (bad customers), Odd(good) = (1-P(bad))/P(bad)
4. ln(odd(good)) was calculated
5. Used the following formula for computing application score card: 400 + slope * (ln(odd(good)) - ln(10)) where slope is 20/(ln(20)-ln(10)) Where, slope=20/(log(20)-log(10))

Summary of application score card values:

1. Scores range from 274.71 to 381.58 for applicants with median score being 344.43
2. Higher scores indicate less risk for defaulting

Application Scorecard



From the Above Graph we can make out that the CutOff Score is coming out to be about 335, below which Credit Card should not be granted to applicants.

Financial Benefits of the Model

The Confusion Matrix for calculating the Financial gain using our model was made on the dataset without missing Performance tag records, since we need to evaluate how much gain was achieved using our model for applicants who were provided with credit card compared to when no model was used.

Profit calculations – with model Vs without model

- We have considered an average profit of Rs.5000 from each non defaulters and
- an average loss of Rs.1,00,000 when each accepted applicant defaults
- Net Profit without model = Rs 3.9665 crores
- Profit using model will be total profit due to each true positive and each true negative minus loss from each false positive and each false negative prediction
- Profit with model = Rs16.41 crores
- Net financial gain with using our model = Rs. 12.44 crores
- Percentage financial gain = 313.72%

Financial Benefits of the Model

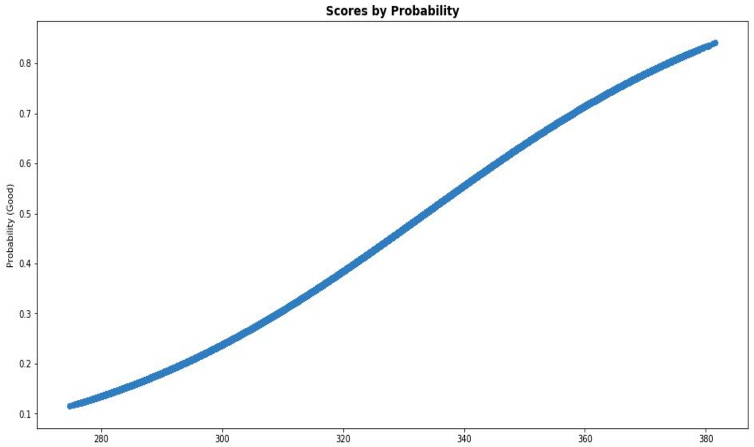
Revenue Loss : Occurs when good customers are identified as bad and credit card application is rejected.

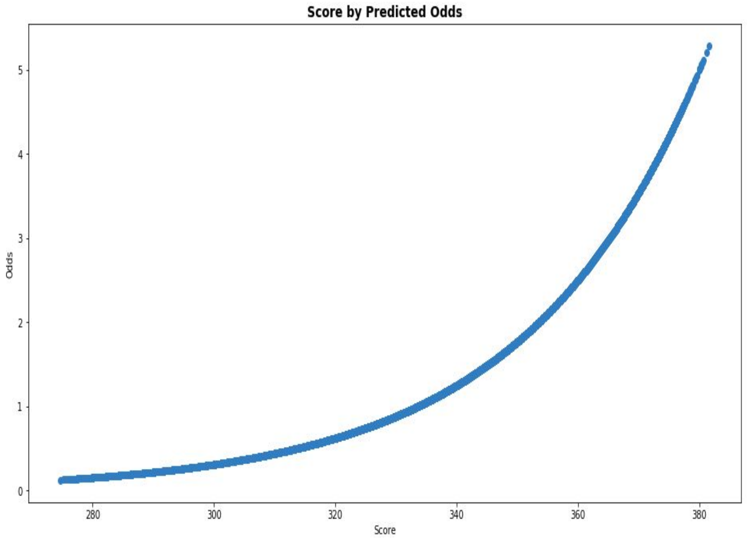
- No of candidates rejected by the model who didn’t default – 29549.
- Total No of candidates who didn’t default – 66919
- % of good candidates rejected by our model – 44.15%
- About 44.15% of the non defaulting customers are rejected which resulted in revenue loss.

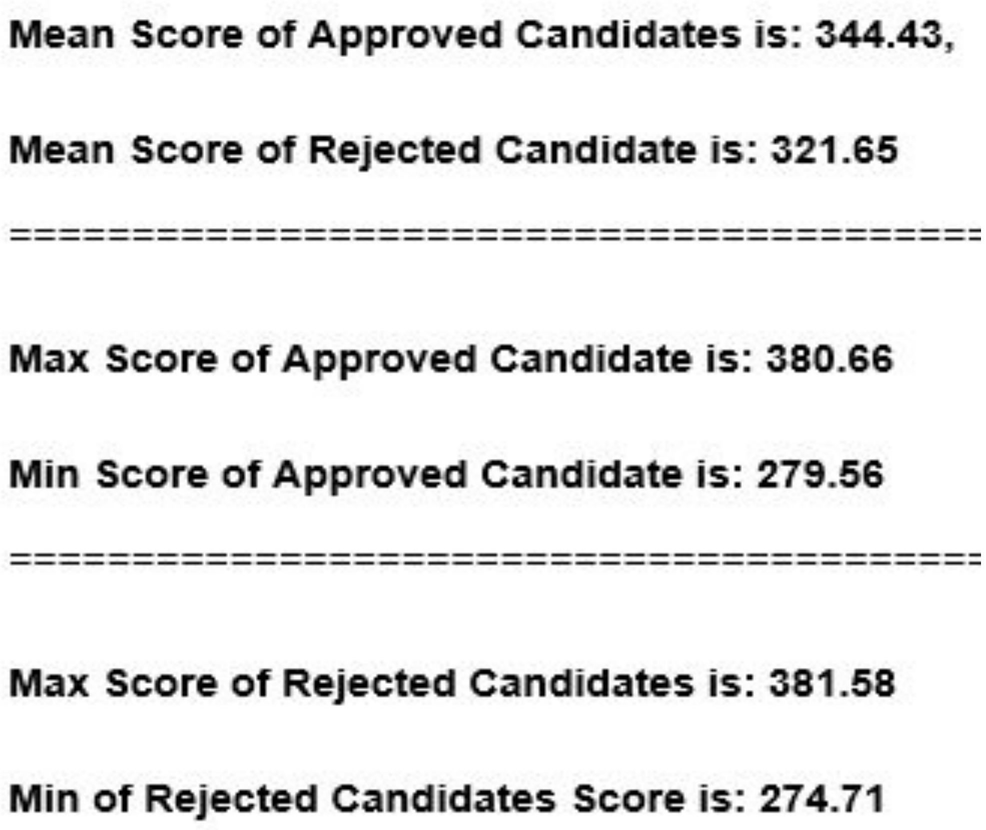
- Credit Loss Saved : The candidates who have been selected by the bank and have defaulted are responsible for the credit loss to the bank.
- % of candidates approved and then defaulted when model was not used = 4.2%
- % of candidates approved and then defaulted when model was used = 1311/69799 = 1.2%
- Credit loss saved => 4.2 – 1.2 = 3%

71% of percentage of Defaulters correctly identified

Application Scorecard







```
Mean Score of Approved Candidates is: 344.43,

Mean Score of Rejected Candidate is: 321.65

========================================

Max Score of Approved Candidate is: 380.66

Min Score of Approved Candidate is: 279.56

========================================

Max Score of Rejected Candidates is: 381.58

Min of Rejected Candidates Score is: 274.71
```

Conclusion

❑ Logistic regression model is chosen as the final Model with 68.9%. of Accuracy.

❑ Optimal score cut-off value of 327.21 is derived to approve and reject the applications.

❑ By this we found out that credit loss 3% was decreased when we used this model. Hence it is accurate in rejecting the candidate who may default in future.

❑ There is Net Financial gain of 313.72% after using the model.