Project Presentation

Rajasekhar Battula

BUSINESS UNDERSTANDING

Objective: To acquire the right customers thereby mitigate the credit risk.

Approach: Create binary classification models using demographic and credit bureau data of bank's past customers. Step by step procedure to achieve this is as follows.

- Data Understanding
- Data Cleaning and Preparation
- Univariate and Bivariate analysis
- Model Building and Evaluation
- Application scorecard creation
- Cost benefit assessment

Let's look at each of the above steps more in detail.

DATA UNDERSTANDING

Datasets provided are demographic and credit bureau data.

Demographic/application data: It is 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 information is gathered 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.

These datasets contain

	Demographic Data	**Credit Bureau Data**
No. of observations	71295	72195
No. of variables	12	19
No. of categorical variables	5	0
No. of numerical variables (excluding Application ID)	6	19
Target variable	Performance Tag	Performance Tag

DATA CLEANING

- There are 1425 rows with Performance Tag (Target variable) missing. Hence these records are copied as rejected data and used for model validation/scorecard validation.
- Number of rows with duplicate Application ID are 3, they are dropped from each dataset.
- There are 65 records whose age is less than 18 which is not the applicable age for issuing credit card hence those records are removed.
- There are 81 records whose income is negative. So, replaced negative income with income =0 for those records.
- Outliers are removed from columns 'No of months in current company', 'Avgas CC Utilization in last 12 months', 'No of trades opened in last 6 months', 'No of trades opened in last 12 months', 'No of PL trades opened in last 6 months', 'No of PL trades opened in last 12 months', 'No of Inquiries in last 6 months', 'No of Inquiries in last 12 months', 'Total No of Trades'.
- Final dataset is formed joining the datasets by Application ID and the final dataset contains 69,853 observations and 29 variables.

WOE Analysis:

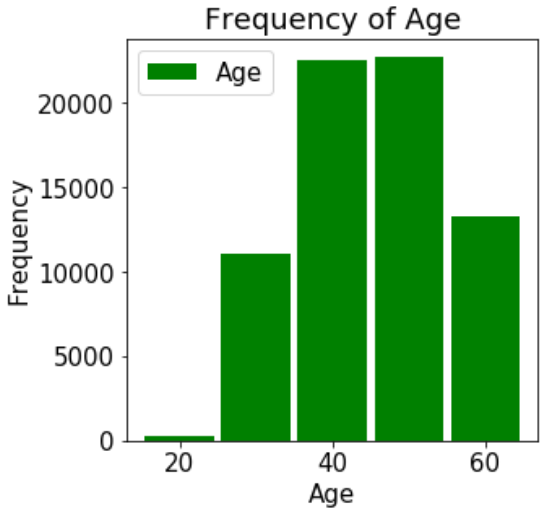
Weight of Evidence and Information Value values are calculated for each of the attributes in the final dataset.

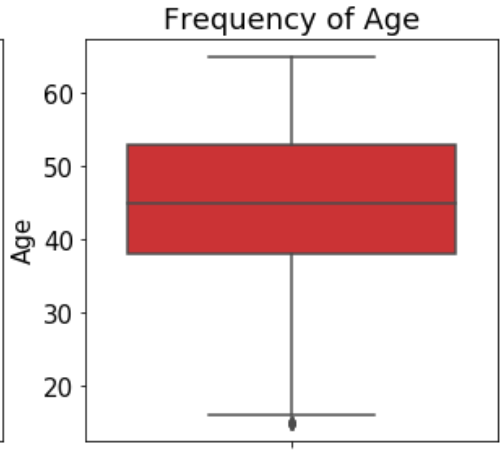
- From the IV values we can conclude that parameters in the demographic data don't play much significant role in prediction and most of the significant variables are from Credit Bureau data.
- Top 9 variables with IV value of 0.1 to 0.3 has medium predictive power and are considered significant. There is no variable with strong predictive power.

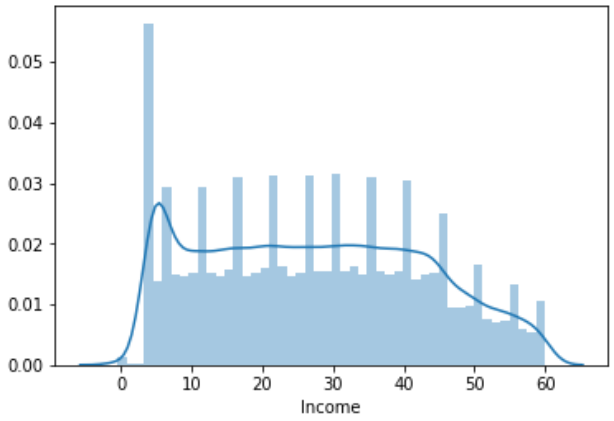
VAR_NAME	IV
Avgas CC Utilization in last 12 months	0.2710
No of trades opened in last 12 months	0.2574
No of Inquiries in last 12 months	0.2292
Total No of Trades	0.1899
No of times 30 DPD or worse in last 12 months	0.1880
No of PL trades opened in last 12 months	0.1766
No of times 30 DPD or worse in last 6 months	0.1457
No of times 60 DPD or worse in last 12 months	0.1377
No of PL trades opened in last 6 months	0.1247
No of times 90 DPD or worse in last 12 months	0.0957
No of trades opened in last 6 months	0.0953
No of Inquiries in last 6 months	0.0929
No of times 60 DPD or worse in last 6 months	0.0896
No of months in current residence	0.0521
Income	0.0376
No of times 90 DPD or worse in last 6 months	0.0307
Presence of open home loan	0.0170
No of months in current company	0.0127
Outstanding Balance	0.0084
Profession	0.0022
Presence of open auto loan	0.0017
Type of residence	0.0009
Education	0.0008
Age	0.0006
Gender	0.0003
Marital Status	0.0001
No of dependents	0.0001
Application ID	0.0000

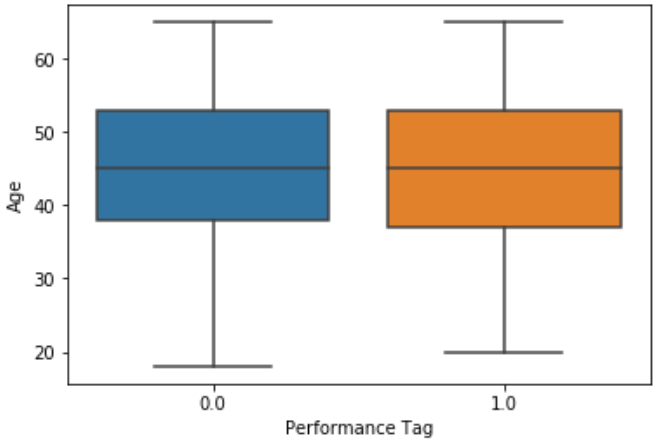
UNIVARIATE AND BIVARIATE ANALYSIS

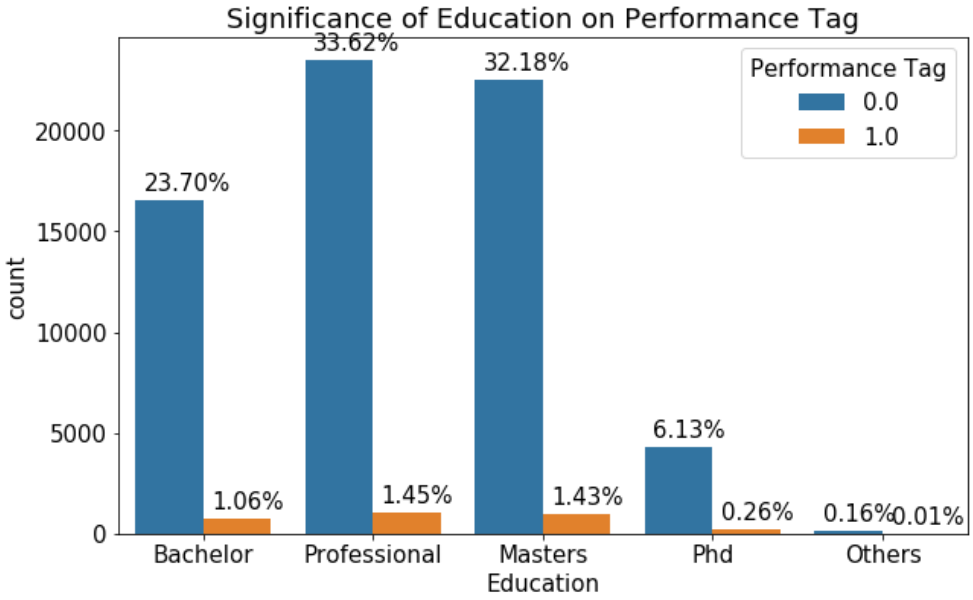
Some of the univariate and bivariate analysis performed on the dataset.

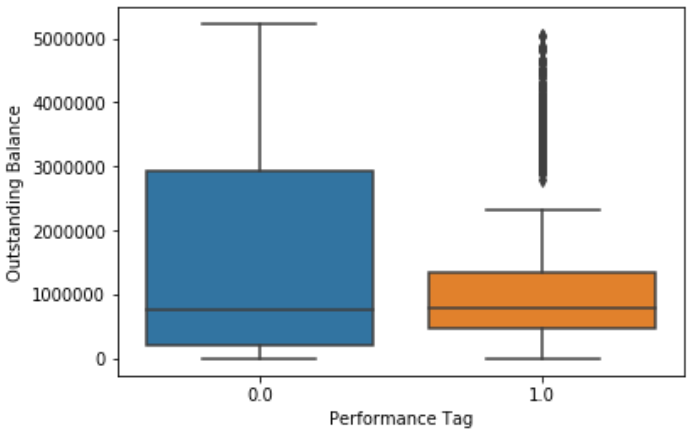


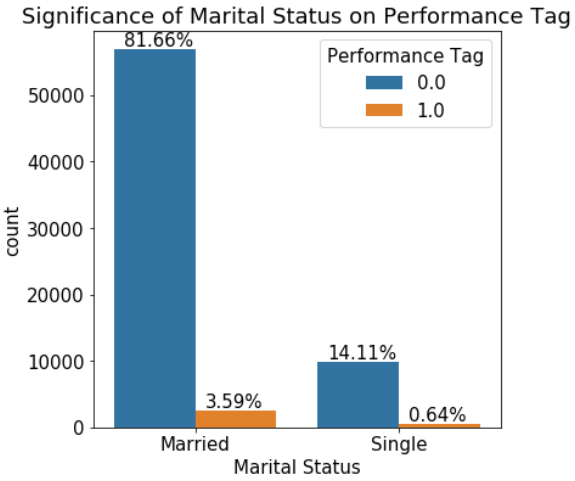


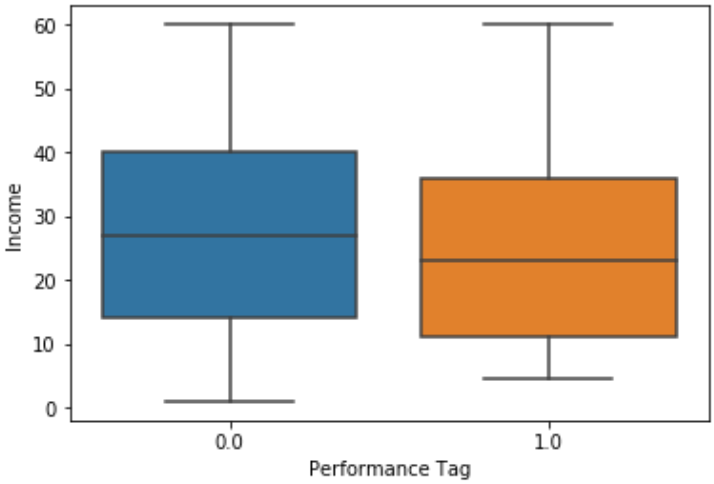


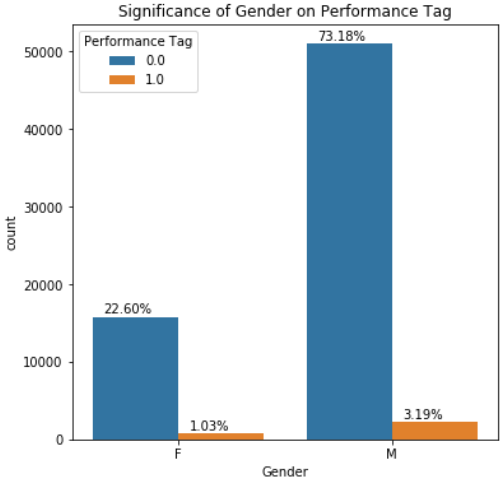


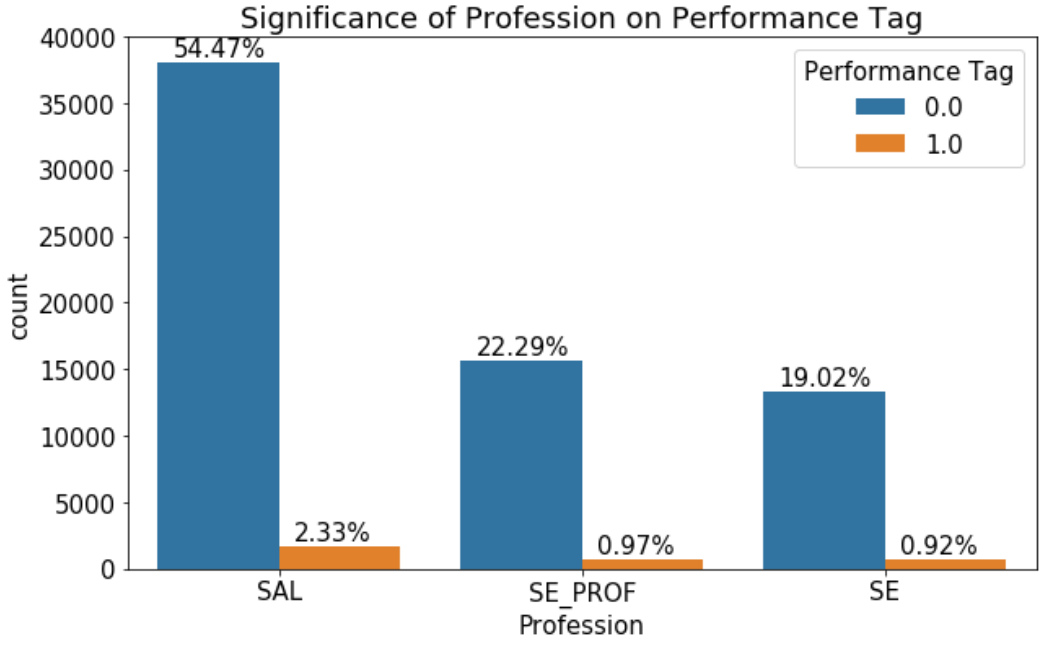


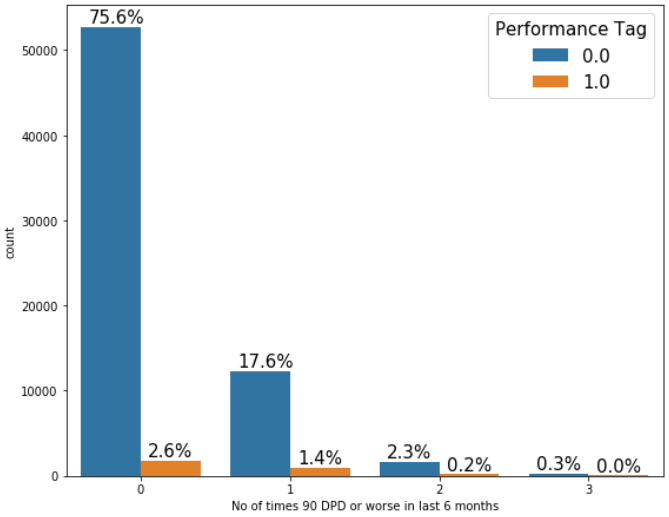

Outstanding Balance
5000000
4000000
3000000
2000000
1000000
0
0.0
1.0
Performance Tag

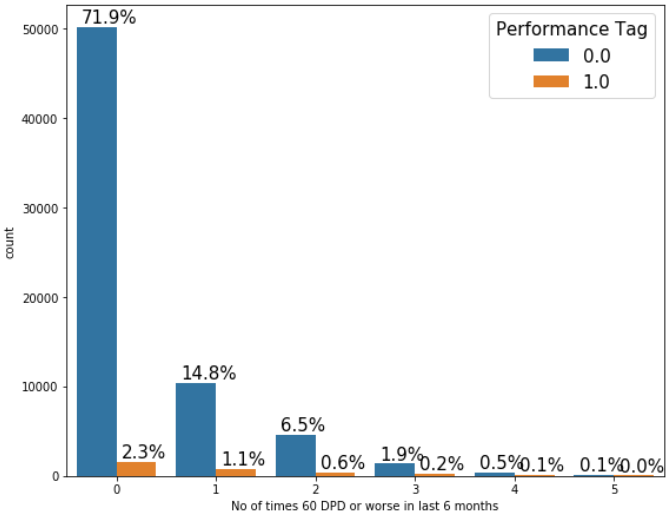

Significance of Marital Status on Performance Tag
Performance Tag
0.0
1.0
count
50000
40000
30000
20000
10000
0
81.66%
3.59%
14.11%
0.64%
Married
Single
Marital Status

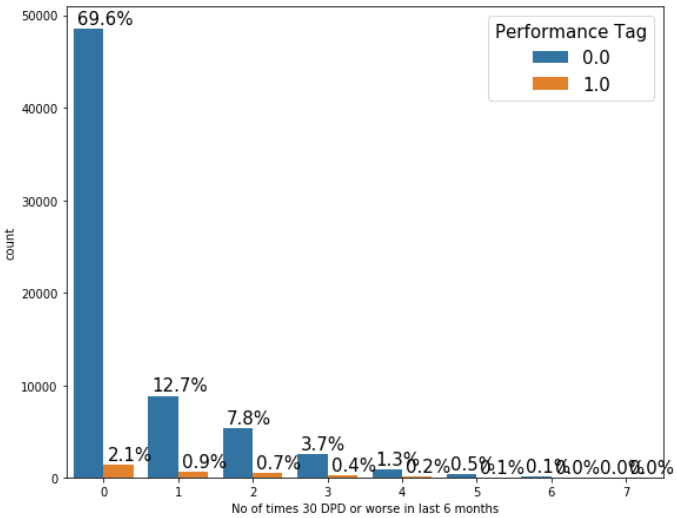

Income
60
50
40
30
20
10
0
0.0
1.0
Performance Tag

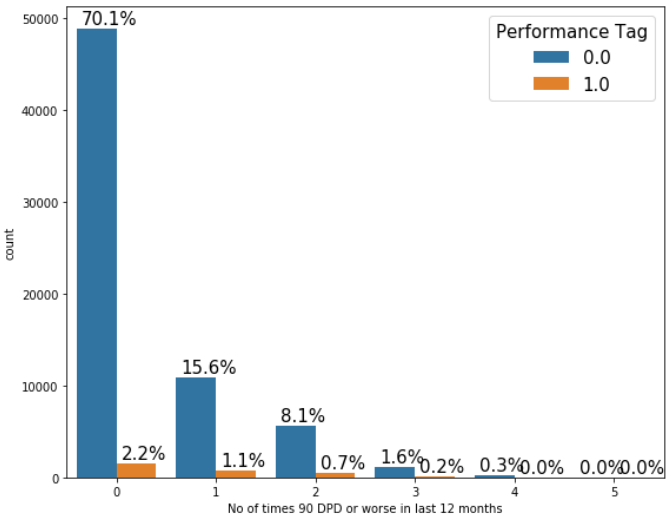

Significance of Gender on Performance Tag
Performance Tag
0.0
1.0
count
50000
40000
30000
20000
10000
0
22.60%
1.03%
73.18%
3.19%
F
M
Gender

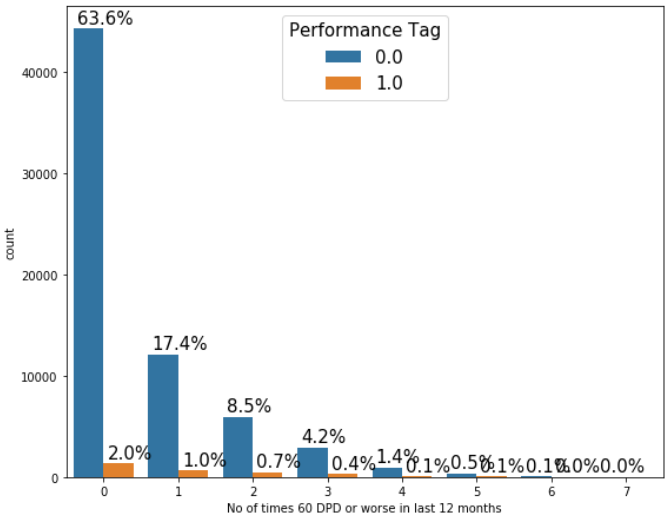

Significance of Profession on Performance Tag
Performance Tag
0.0
1.0
count
40000
35000
30000
25000
20000
15000
10000
5000
0
54.47%
2.33%
22.29%
0.97%
19.02%
0.92%
SAL
SE_PROF
SE
Profession

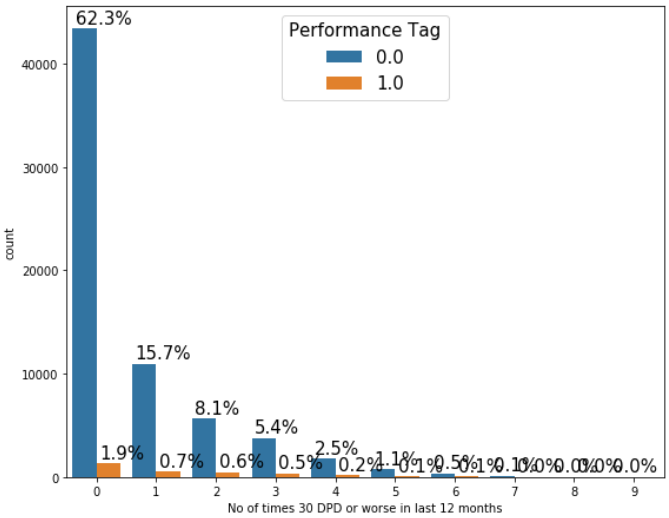

Performance Tag
0.0
1.0
count
75.6%
2.6%
17.6%
1.4%
2.3%
0.2%
0.3%
0.0%
No of times 90 DPD or worse in last 6 months

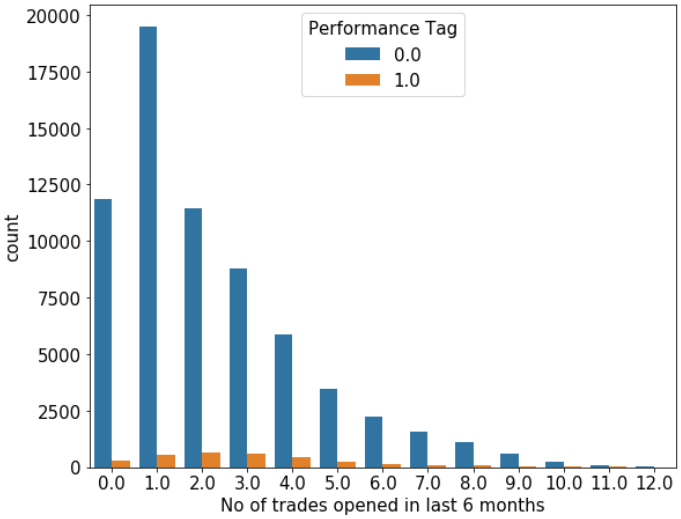

Performance Tag
0.0
1.0
count
71.9%
2.3%
14.8%
1.1%
6.5%
0.6%
1.9%
0.2%
0.5%
0.1%
0.1%
0.0%
No of times 60 DPD or worse in last 6 months

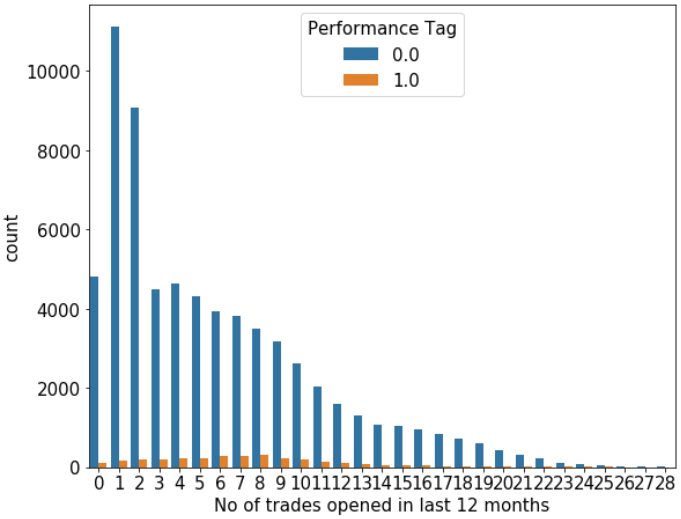

Performance Tag
0.0
1.0
count
69.6%
2.1%
12.7%
0.9%
7.8%
0.7%
3.7%
0.4%
1.3%
0.2%
0.5%
0.1%
0.1%
0.0%
0.0%
0.0%
No of times 30 DPD or worse in last 6 months

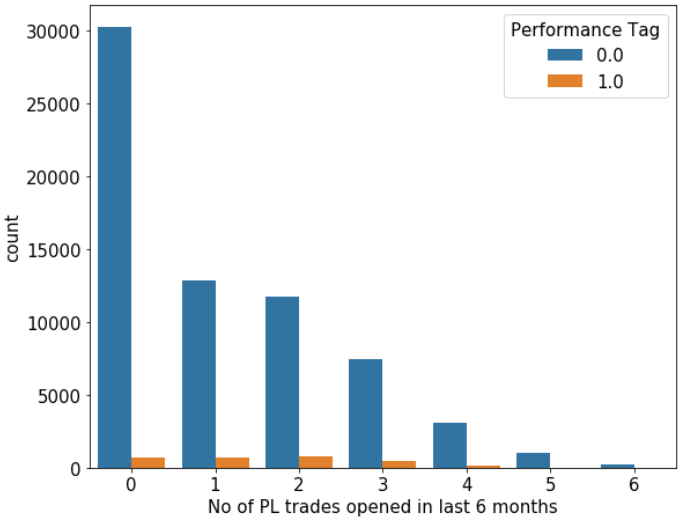

Performance Tag
0.0
1.0
count
70.1%
2.2%
15.6%
1.1%
8.1%
0.7%
1.6%
0.2%
0.3%
0.0%
0.0%
0.0%
No of times 90 DPD or worse in last 12 months

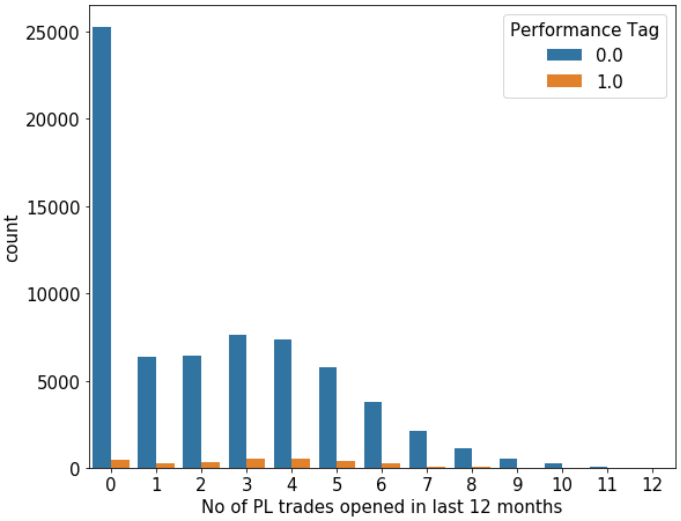

Performance Tag
0.0
1.0
count
63.6%
2.0%
17.4%
1.0%
8.5%
0.7%
4.2%
0.4%
1.4%
0.1%
0.5%
0.1%
0.1%
0.0%
0.0%
No of times 60 DPD or worse in last 12 months

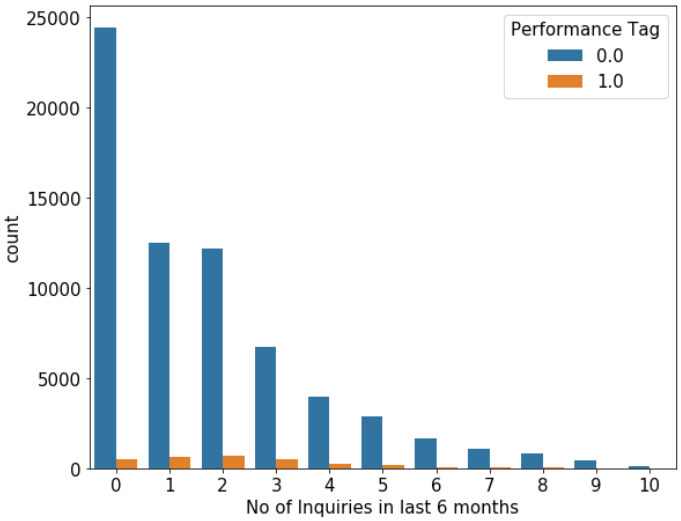

Performance Tag
0.0
1.0
count
62.3%
1.9%
15.7%
0.7%
8.1%
0.6%
5.4%
0.5%
2.5%
0.2%
1.1%
0.1%
0.5%
0.1%
0.1%
0.0%
0.0%
0.0%
No of times 30 DPD or worse in last 12 months

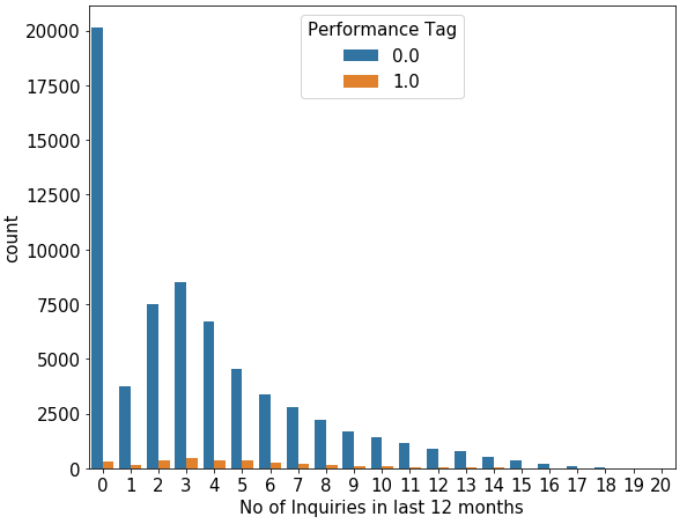

Performance Tag
0.0
1.0
count
No of trades opened in last 6 months


Performance Tag
0.0
1.0
count
No of trades opened in last 12 months


Performance Tag
0.0
1.0
count
No of PL trades opened in last 6 months


Performance Tag
0.0
1.0
count
No of PL trades opened in last 12 months


Performance Tag
0.0
1.0
count
No of Inquiries in last 6 months


Performance Tag
0.0
1.0
count
No of Inquiries in last 12 months

MODEL BUILDING

In this stage, we are building 2 models they are:

1. Model only using demographic data.
2. Model using full data (combination of demographic and credit bureau data).

Reiterating the objective of this project:

In the past few years, CredX has experienced an increase in credit loss. The CEO believes that the best strategy to mitigate credit risk is to 'acquire the right customers'

This can be achieved by identifying the customers with chance of default and rejecting their application. Assuming income is not major priority. In other terms, False Negatives has to be reduced i.e Recall/Sensitivity metric has to be high.

Considering this the Model valuation metrics are below,

After data cleaning and preparation, we have applied Logistic regression models on demographic data and observed the below results.

Demographic data model Metrics	Logistics Regression Values
Sensitivity	0.550
Specificity	0.566
Accuracy	0.565
AUC	0.558
F1 Score	0.100

After data cleaning and preparation, we have applied Logistic regression, Decision trees and Random forest models on full data and observed the below results.

Full data model	Logistics Regression Values	Decision Tree Values	Random Forest Values
Sensitivity	0.638	0.626	0.721
Specificity	0.603	0.567	0.518
Accuracy	0.604	0.570	0.527
AUC	0.620	0.597	0.620
F1 Score	0.124	0.113	0.118

Sensitivity is high for Random forest model. i.e 0.721 which will be used for this job.

Confusion Matrix

	Predicted Non-Default	Predicted Default
Actual Non-Default	10,398	9,639
Actual Default	256	663

APPLICATION SCORECARD CREATION

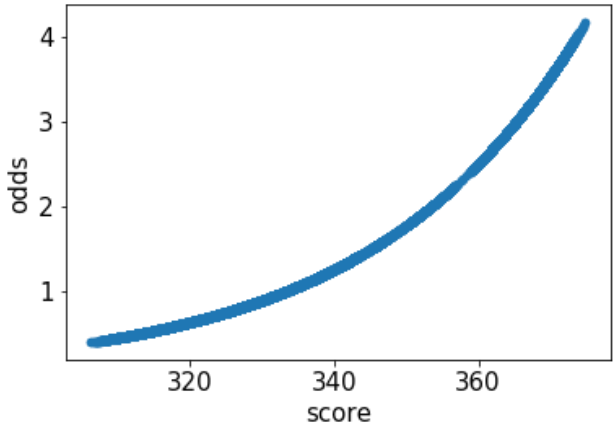
The application score for each applicant calculated using the Random forest model, ranges from 306.27 to 374.81. Score increases by 20 points for doubling odds for good customers. Application score for odds of 10 to 1 is 400. The higher the score, the better the customer is from a risk perspective.

Method used for computation of application scorecard:

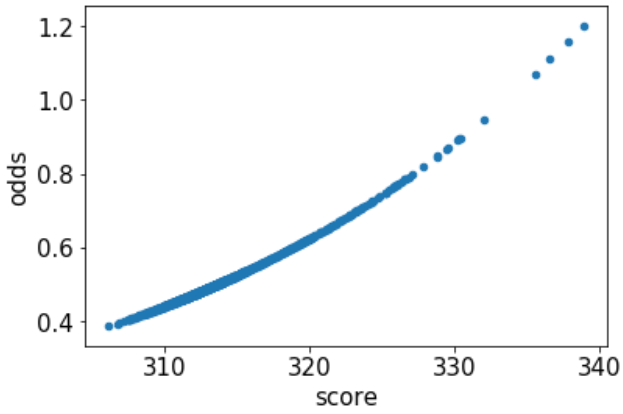
- Computed the probabilities of default for the entire population of applicants using the logistic regression model.
- Computed the odds for the good. Since the probability computed is for rejection (bad customers), Odd(good) = (P(good))/P(bad)
- Used the following procedure for computing application score

target_score = 400
target_odds = 10
double_odds = 20
slope = double_odds/ log(2)
offset = target_score - factor * log(target_odds)
Application score = offset + slope * log(odds)

For full data Application scorecard looks like below,



For rejected data (Source data without performance tag) Application scorecard looks like below,



When the Random forest model is applied on the rejected data (1425 applicants), 1421 applicants were predicted to be defaulted and 4 as non-default.

FINANCIAL BENEFIT OF THE MODEL

Considering the 20,956 applicants in test dataset for understanding financial benefit of the model. Applying Random forest model on this data has the following results.

1. The implications of using the model for auto-approval or rejection, i.e., how many applicants on an average would the model automatically approve or reject
   Ans: Among 20956 applicants (test data), 10654 will be approved by the model and 10302 will be rejected by the model.
2. The potential credit loss avoided with the help of the model

Ans: Credit loss avoided by the model is by rejecting applications of the 663 out of 919 i.e 72.14% of defaulters. Assume the average credit of each individual is 2 lakh rupees then total credit loss avoided = 13.26 crore

3. Assumptions based on which the model was built
   Ans: Assumption1 --> Generating income is less priority than mitigating credit loss.
   Assumption2 --> So model should be built on data where credit card was approved i.e Performance Tag is 0 or 1. Performance Tag = NA are considered as rejected data.
   Assumption3 --> NA value in 'Avgas.CC.Utilization.in.last.12.months' indicates no usage of CC by user. So assigned value 0.