CREDX RISK ANALYTICS CASE STUDY SUBMISSION

Group Name:

1. Omkar Daddikar
2. Amit Bhosale
3. Pradeep Doddarangaiah

CREDX RISK ANALYTICS CASE STUDY

Objective:

To acquire the right customers using the predictive models. We need to determine the factors affecting the credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of the project

Problem Statement:

CredX is a leading credit card provider that gets thousands of credit card applicants every year. But in the past few years, it has experienced an increase in credit loss due to increase in defaults.

Solution Approach:

This is a binary supervised classification problem. We aim at building models such as Logistic regression, Decision Tree Random forest to identify the customers who are at a risk of defaulting if offered a credit card using Application Scorecard.

CREDX RISK ANALYTICS CASE STUDY

Data Collection

- Read Data from Demographic data.csv File
- Read Data from Credit Bureau data.csv File

Data Cleaning

- Finding and Removing Duplicate Records
- Checking Outliers in Data and Treating them
- Explore data by Univariate and Multivariate Analysis
- Merging Demographic data & Credit Bureau Data

Information Value and WOE

- Compute Information Value
- Populate Feature with their WOE values
- Consider the Predictor Variable with IV > 0.1.
- Replacing NA values with values of nearest WOE bin.
- Replacing the Actual Data with WOE Values for the Predictor Variable

Model Building

- Split data into train and test
- Using SMOTE for balancing Data
- Building Logistic Model for Demographic Data
- Building Models for important predictor variables using Logistic Regression, Decision Tree and Random Forest

Model Evaluation

- We test the model against the test dataset using Confusion Matrix (determining accuracy, specificity and sensitivity)
- Identify the cutoff Values where accuracy sensitivity specificity curves meet

Building Application Scorecards

- Evaluate Model Metrics
- Build the Application Score Card for the Entire Dataset using the cutoff value, log odds and other parameters mentioned.
- Build Financial Strategies using the Optimum Model

CREDX RISK ANALYTICS CASE STUDY

DATA UNDERSTANDING

Two datasets are provided, demographic data and credit bureau data.

1. **Demographic/Application Data:** This dataset contains 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.

2. **Credit Bureau Data:** This information is taken from the credit bureau and contains variables such as 'number of times 30 DPD or worse in last 3/6/12 months', 'outstanding balance', 'number of trades', etc.

Demographic Data	Credit Bureau Data
Application ID	Application ID
Age	No of times 90 DPD or worse in last 6 months
Gender	No of times 60 DPD or worse in last 6 months
Marital Status (at the time of application)	No of times 30 DPD or worse in last 6 months
No of dependents	No of times 90 DPD or worse in last 12 months
Income	No of times 60 DPD or worse in last 12 months
Education	No of times 30 DPD or worse in last 12 months
Profession	Avgas CC Utilization in last 12 months
Type of residence	No of trades opened in last 6 months
No of months in current residence	No of trades opened in last 12 months
No of months in current company	No of PL trades opened in last 6 months
Performance Tag	No of PL trades opened in last 12 months
	No of Inquiries in last 6 months (excluding home & auto loans)
	No of Inquiries in last 12 months (excluding home & auto loans)
	Presence of open home loan
	Outstanding balance
	Total No of trades
	Presence of open auto loan
	Performance Tag

CREDX RISK ANALYTICS CASE STUDY

Nature of Data:

- The demographic data consists of 71295 observations with 12 variables.
- The credit bureau data consists of 71295 observations with 19 variables.
- Application ID is the common key between the two datasets for merging.
- Performance Tag is the target variable which says if customer is default or not. The values are 0(nondefault) and 1(default).

DATA CLEANSING AND PREPARATION

Data Quality Issues:

- Both occurrences of 3 duplicate Application ID records (765011468, 653287861, 671989187) has been excluded from the dataset.

Demographic Data Set

Variables	No of Missing Values/NA
Application ID	-
Age	-
Gender	-
Marital Status (at the time of application)	-
No of dependents	3
Income	-
Education	-
Profession	-
Type of residence	-
No of months in current residence	-
No of months in current company	-
Performance Tag	1425

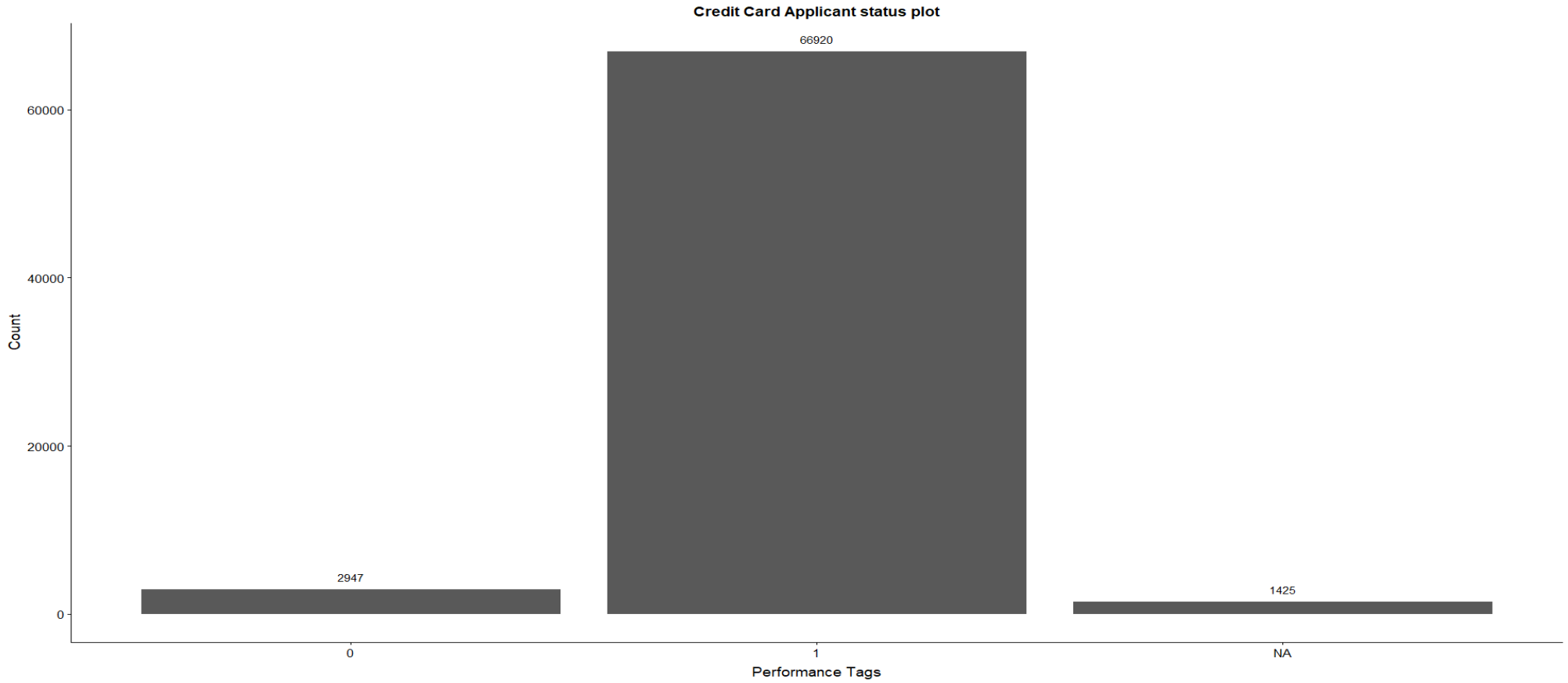
Credit Bureau Data Set

Variables	No of Missing Values/NA
Application ID	-
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	-
Avgas CC Utilization in last 12 months	1058
No of trades opened in last 6 months	1
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	-
Presence of open home loan	272
Outstanding balance	272
Total No of trades	-
Presence of open auto loan	-
Performance Tag	1425

CREDX RISK ANALYTICS CASE STUDY

EXPLORATORY DATA ANALYSIS:

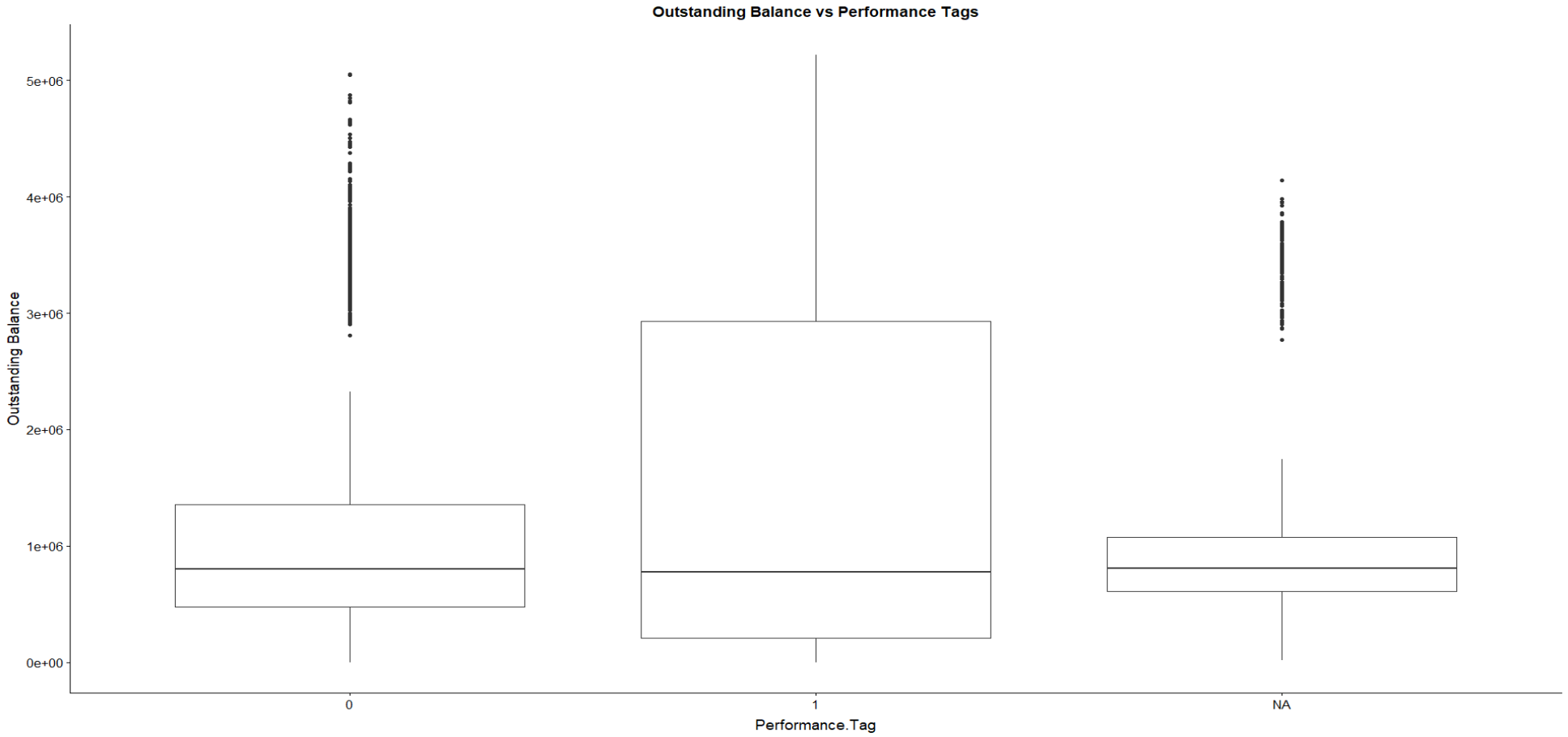
- Both Univariate and Bivariate analysis is performed on all the variables of the dataset.
- Around 4% of the dataset contain applicants which are defaulters.



CREDX RISK ANALYTICS CASE STUDY

EXPLORATORY DATA ANALYSIS:

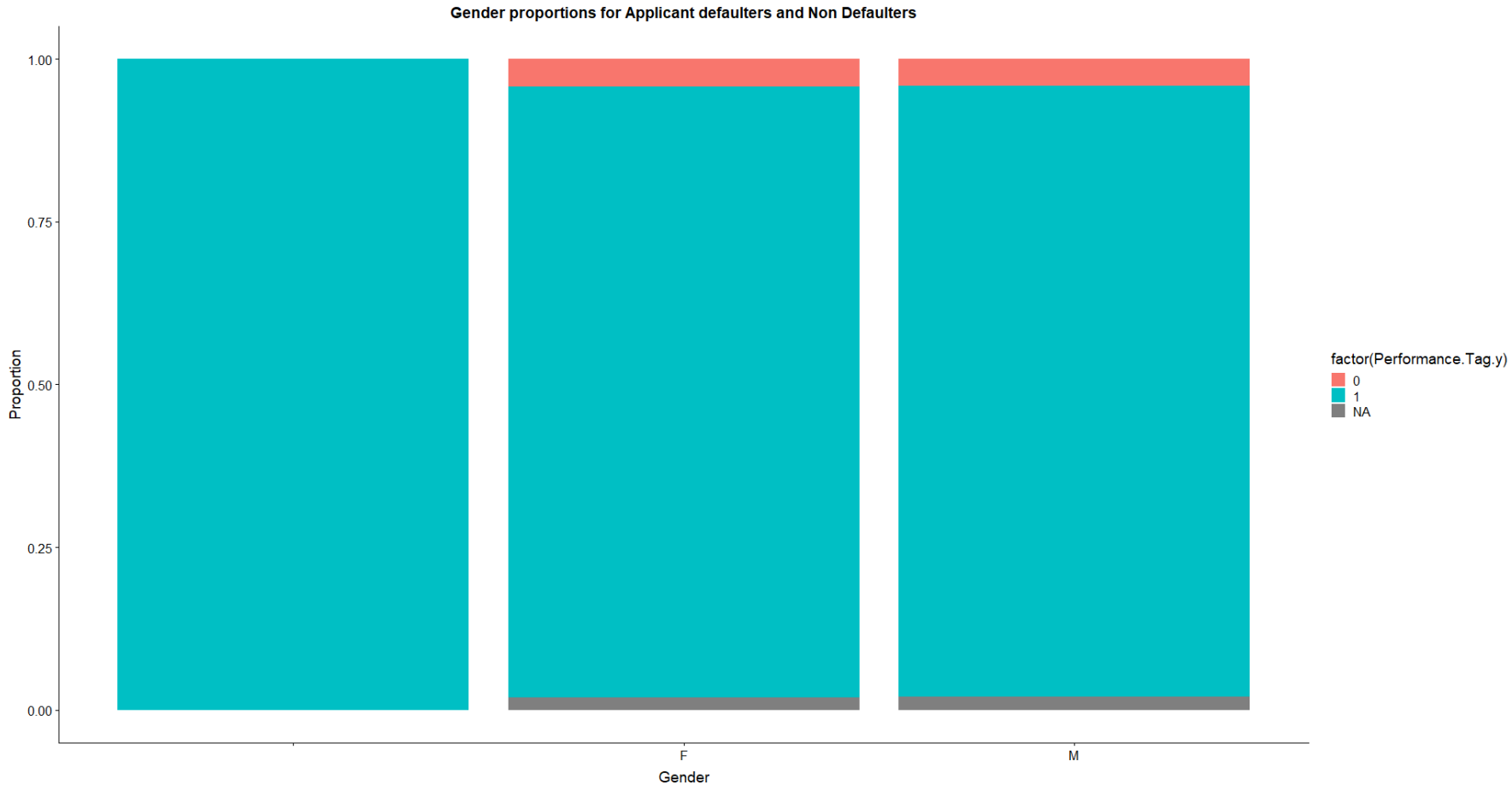
It signifies, higher the Outstanding balance, higher the chance of getting a defaulter



CREDX RISK ANALYTICS CASE STUDY

EXPLORATORY DATA ANALYSIS:

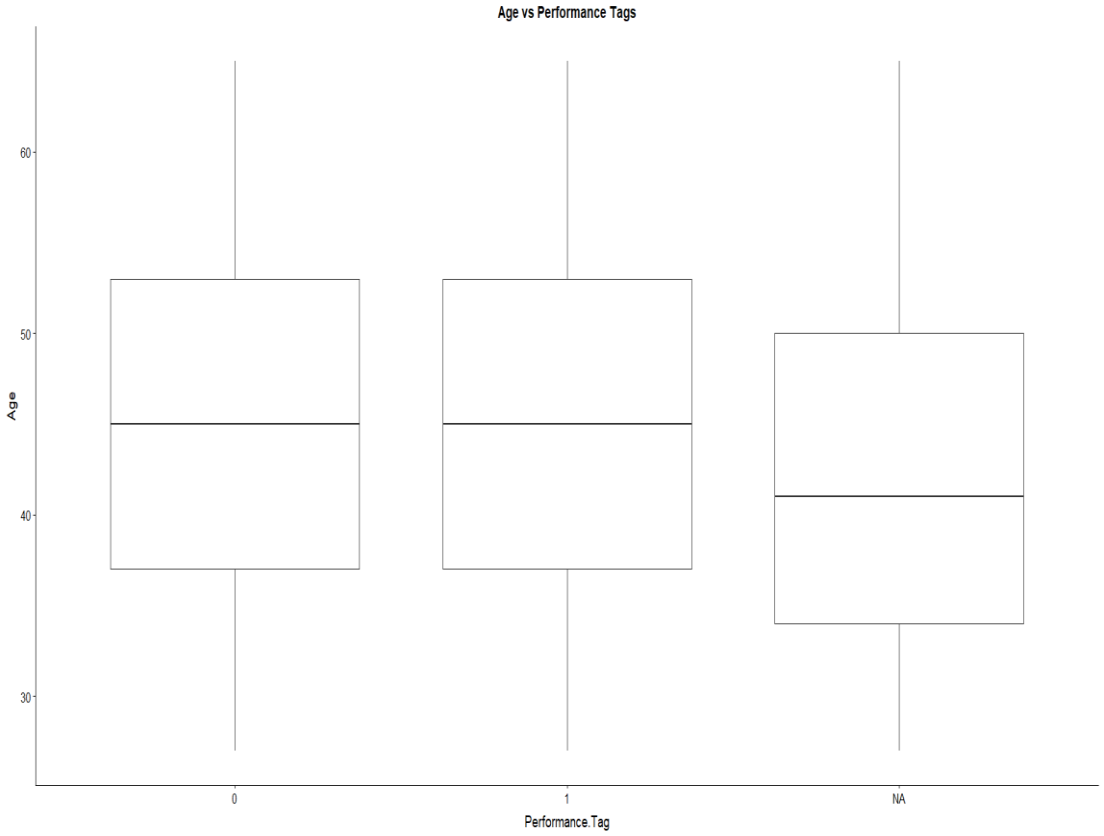
No significance of Gender attribute on the defaulters or non defaulters.

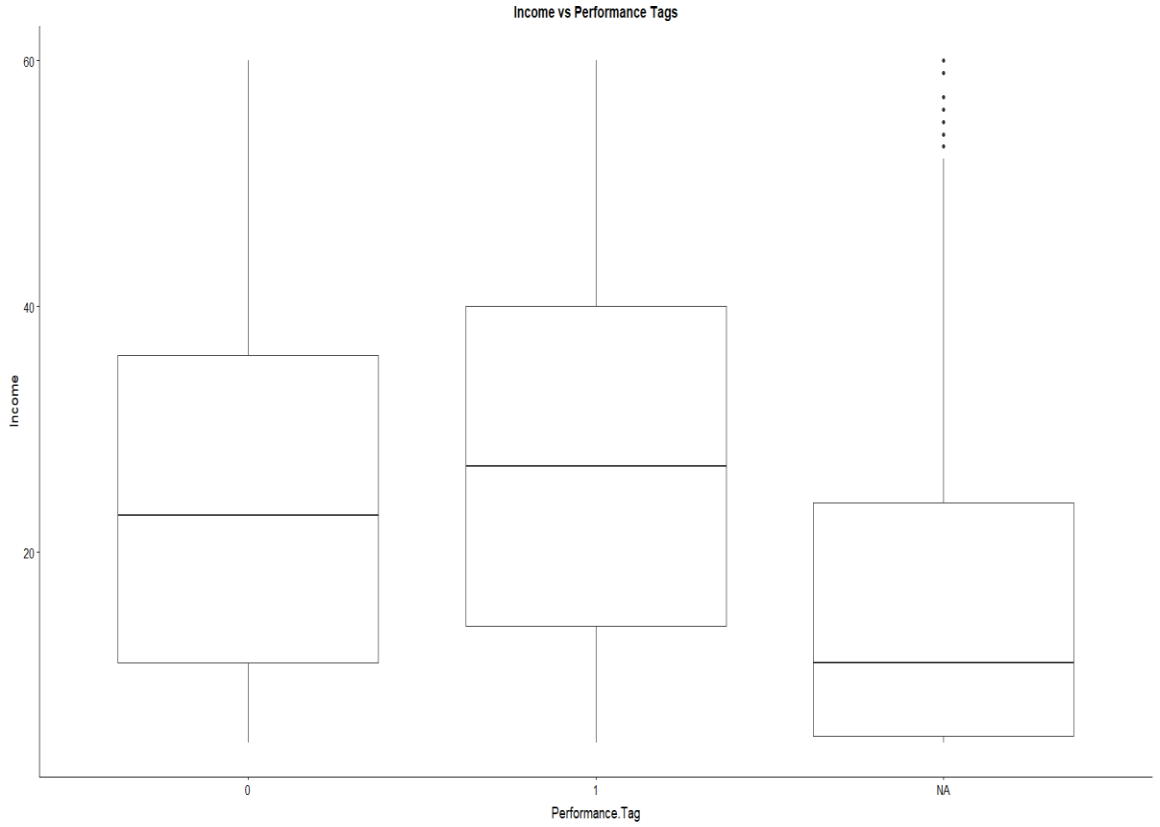


CREDX RISK ANALYTICS CASE STUDY

EXPLORATORY DATA ANALYSIS:

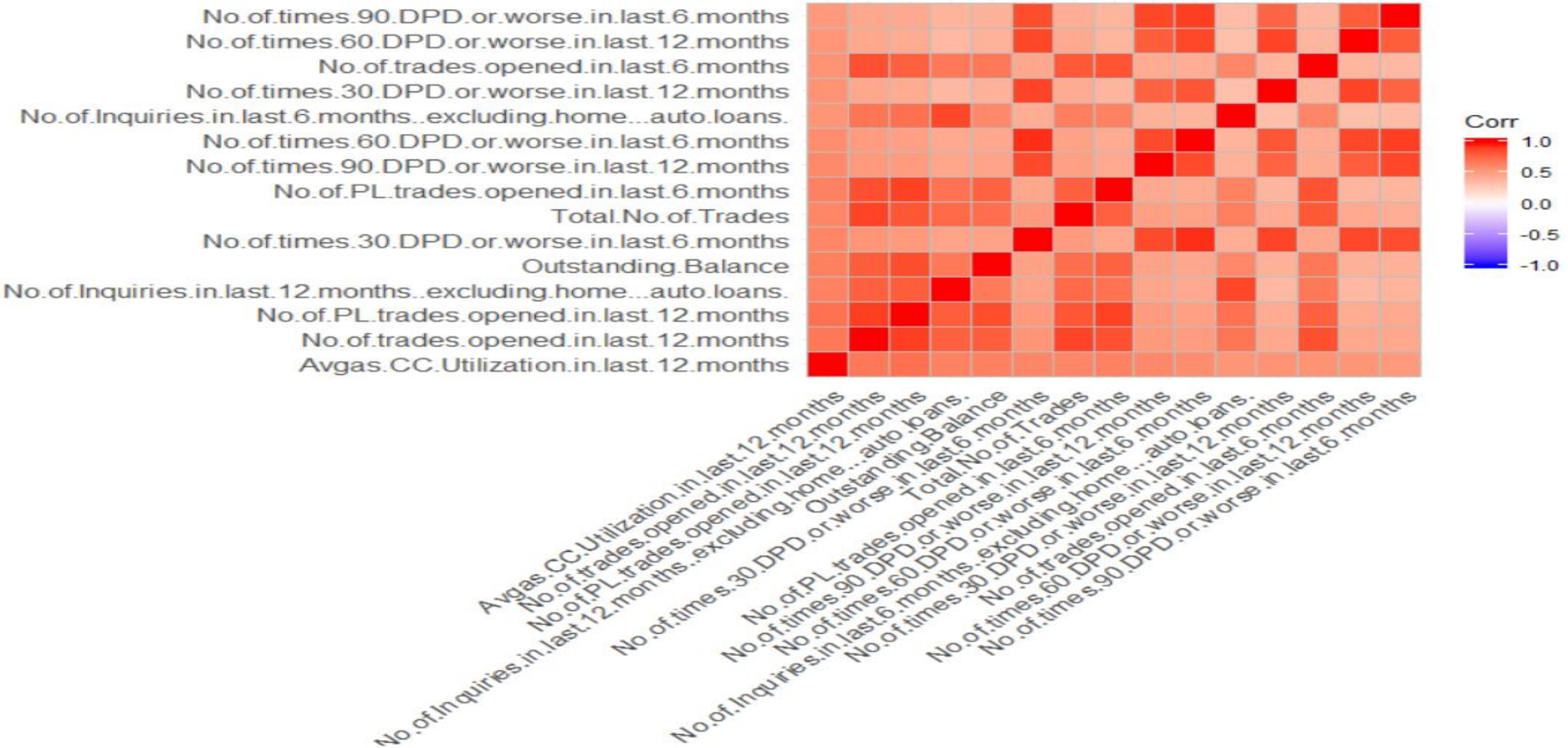
We can observe from the below boxplots that there is no significant impact on an applicant becoming a default based on their Age and Income values.





CREDX RISK ANALYTICS CASE STUDY

CORRELATION ANALYSIS: This was carried out for the important predictor variables identified in the dataset based on the IV values.



CREDX RISK ANALYTICS CASE STUDY

WOE(Weight of Evidence) AND IV ANALYSIS

- WOE and IV values are calculated for each of the attributes using information package. Continuous quantitative variables for which WOE values were not monotonically changing across bins, were made by the Information package in R by default.
- Since Information package treats 1 as 'good', we created a new variable - Reverse. Performance.Tag with inversed relationship for IV analysis.
- WOE(Weight of Evidence) tells the predictive power of an independent variable in relation to the dependent variable which can be given as follows

$$WOE = \ln\left(\frac{\text{Distribution of Goods}}{\text{Distribution of Bads}}\right)$$

- Below is the table showing the variable importance category based on the IV value it has.

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

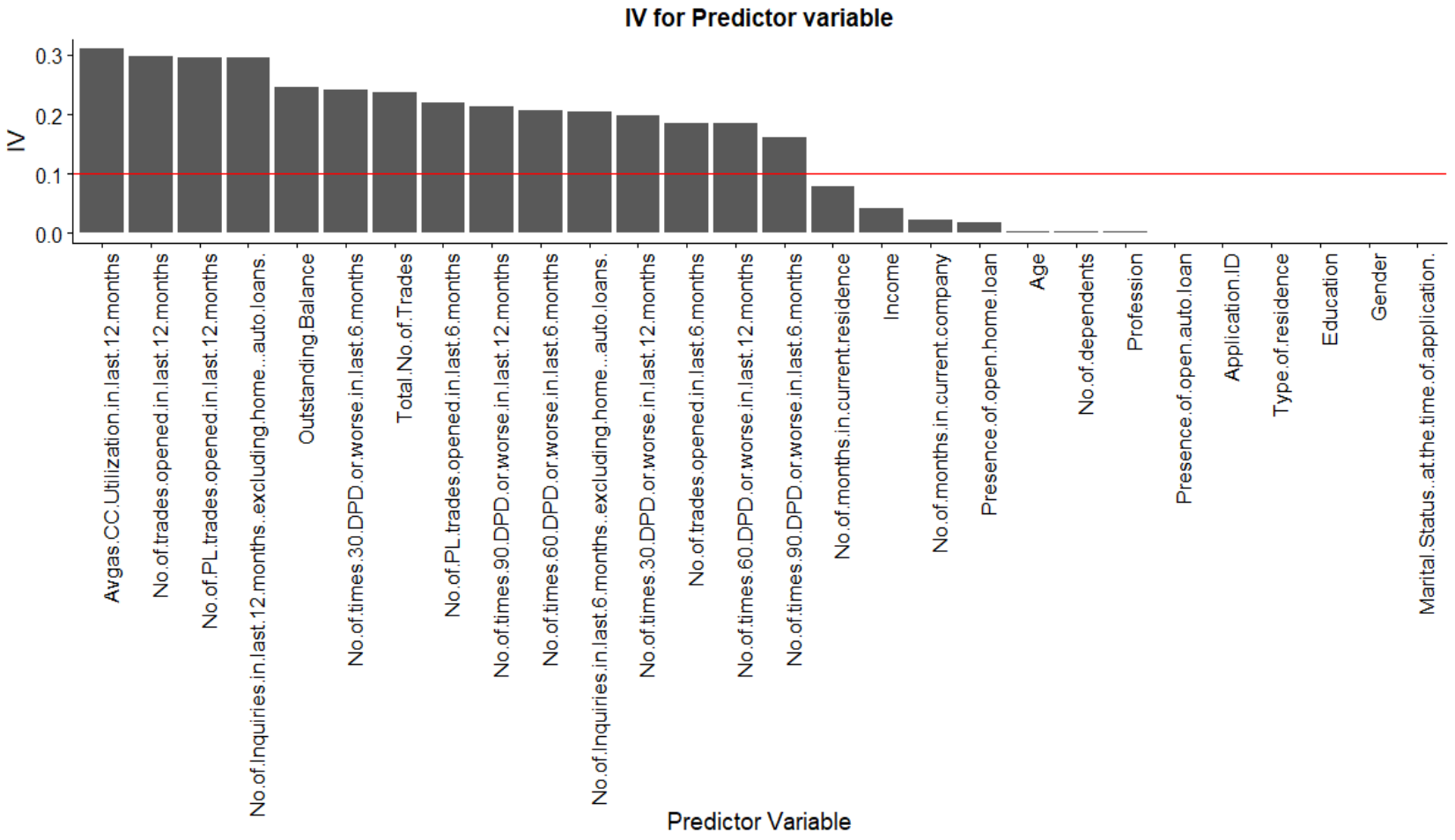
- From the IV values we can conclude that parameters in the demographic data don't play much significant role in prediction and most of the significant variables are from Credit Bureau data.
- Top 15 Variables with IV value of 0.1 to 0.5 has medium and strong predictive power and are considered significant.

CREDX RISK ANALYTICS CASE STUDY

WOE AND IV ANALYSIS AND IDENTIFYING IMPORTANT PREDICTOR VARIABLES (Highlighted in Green)

Variables of credit bureau dataset showed better insights than demographic variables.

Variables	IV
Avgas.CC.Utilization.in.last.12.months	3.10E-01
No.of.trades.opened.in.last.12.months	2.98E-01
No.of.PL.trades.opened.in.last.12.months	2.96E-01
No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.	2.95E-01
Outstanding.Balance	2.46E-01
No.of.times.30.DPD.or.worse.in.last.6.months	2.42E-01
Total.No.of.Trades	2.37E-01
No.of.PL.trades.opened.in.last.6.months	2.20E-01
No.of.times.90.DPD.or.worse.in.last.12.months	2.14E-01
No.of.times.60.DPD.or.worse.in.last.6.months	2.06E-01
No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.	2.05E-01
No.of.times.30.DPD.or.worse.in.last.12.months	1.98E-01
No.of.trades.opened.in.last.6.months	1.86E-01
No.of.times.60.DPD.or.worse.in.last.12.months	1.85E-01
No.of.times.90.DPD.or.worse.in.last.6.months	1.60E-01
No.of.months.in.current.residence	7.90E-02
Income	4.24E-02
No.of.months.in.current.company	2.18E-02
Presence.of.open.home.loan	1.76E-02
Age	3.35E-03
No.of.dependents	2.65E-03
Profession	2.22E-03
Presence.of.open.auto.loan	1.66E-03
Application.ID	1.50E-03
Type.of.residence	9.20E-04
Education	7.83E-04
Gender	3.26E-04
Marital.Status..at.the.time.of.application.	9.47E-05



CREDX RISK ANALYTICS CASE STUDY

DATA TRANSFORMATION

OUTLIER TREATMENT:

- Outlier detection is done using boxplot on continuous variables and quantiles function and the variables with outliers has been corrected by capping the outliers to the nearest non-outlier values.

DATA TRANSFORMATION:

- WOE Analysis is performed on all important predictor variables and the actual values are replaced with the WOE values pertaining to the bin they belong to.

DATA SPLIT:

- The Final dataset contains 69,867 records and the dataset is split into Train and Test in 70:30 ratio for model building.

DATA SAMPLING:

- The data is highly imbalanced. Only 4.2% of total data is about the defaulters. We have used SMOTE for balancing our data sets. It helps to generate artificial data based on sampling methods thereby improving the specificity of the models.

CREDX RISK ANALYTICS CASE STUDY

Predictive Modelling: Comparison of Model parameters for various models.

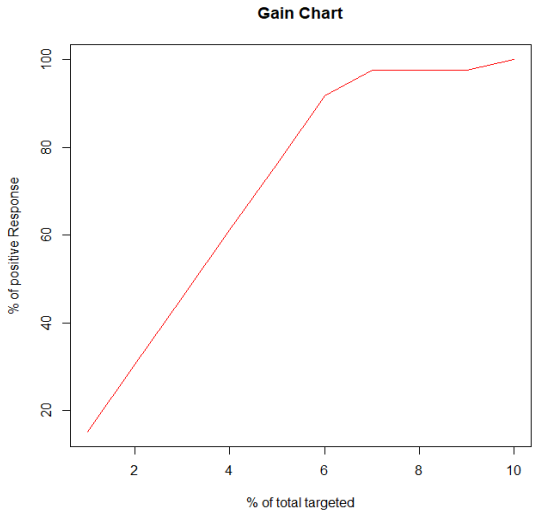
We had evaluated different models and found out the Accuracy, Sensitivity & Specificity values for these.

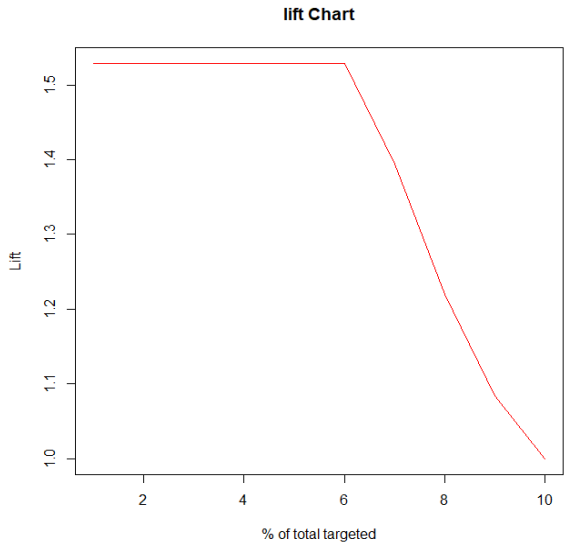
The best model was selected to be **Random Forest** as the model evaluation parameters were relatively higher and the models were stable.

Models	Regular data				Smoted data		
	Logistic Regression (Entire Data)	Logistic Regression (Demographic Data)	Decision Tree	Random Forest	Logistic Regression	Decision Tree	Random Forest
% Accuracy	67	58	95	61	62	61	66
% Sensitivity	68	58	99	61	62	61	62
% Specificity	55	54	0.3	62	65	65	66

CREDX RISK ANALYTICS CASE STUDY

Predictive Modelling: Comparison of Model parameters for various models.





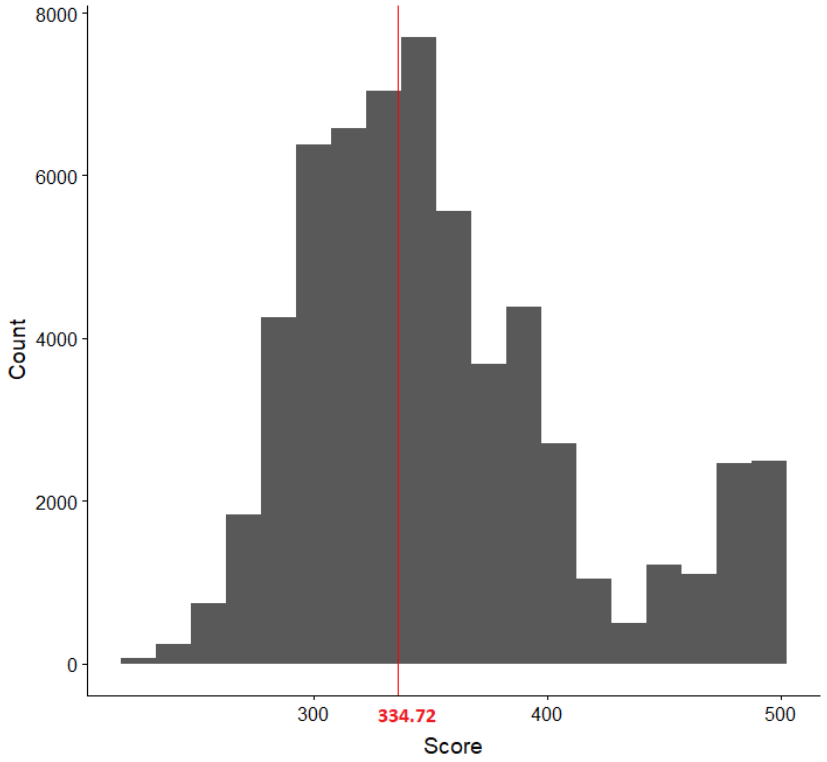
We can observe that in the 5th decile, we are able to target around 78% of the customers who are likely to be approved for the credit card service compared to 50% in case of a random selection.

Bucket	Total	Total Good	Cumulative Good	CumGain	CumLift
1	6987	6987	6987	15.57929	1.557929
2	6987	6987	13974	31.15858	1.557929
3	6987	6987	20961	46.73787	1.557929
4	6986	6986	27947	62.31493	1.557873
5	6987	6987	34934	77.89422	1.557884
6	6987	6987	41921	93.47351	1.557892
7	6986	1836	43757	97.56734	1.393819
8	6987	0	43757	97.56734	1.219592
9	6987	0	43757	97.56734	1.084082
10	6986	1091	44848	100	1

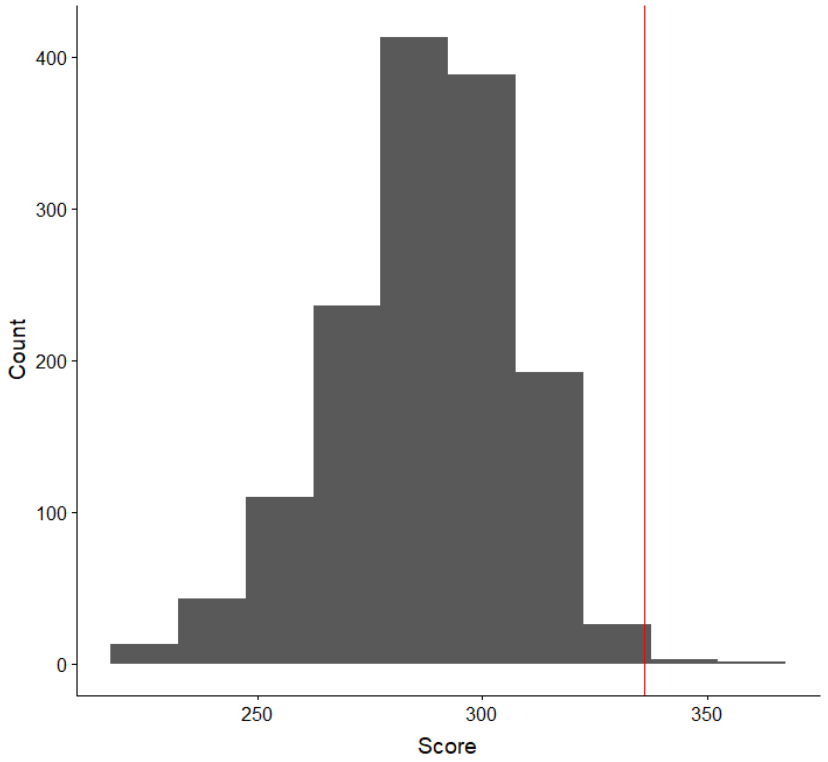
CREDX RISK ANALYTICS CASE STUDY

Application Scorecard: Score varies between 203 to 495; Cut-off score - 334

- Cut-off: 334 is the baseline for providing credit card to the customers where performance data is mentioned.



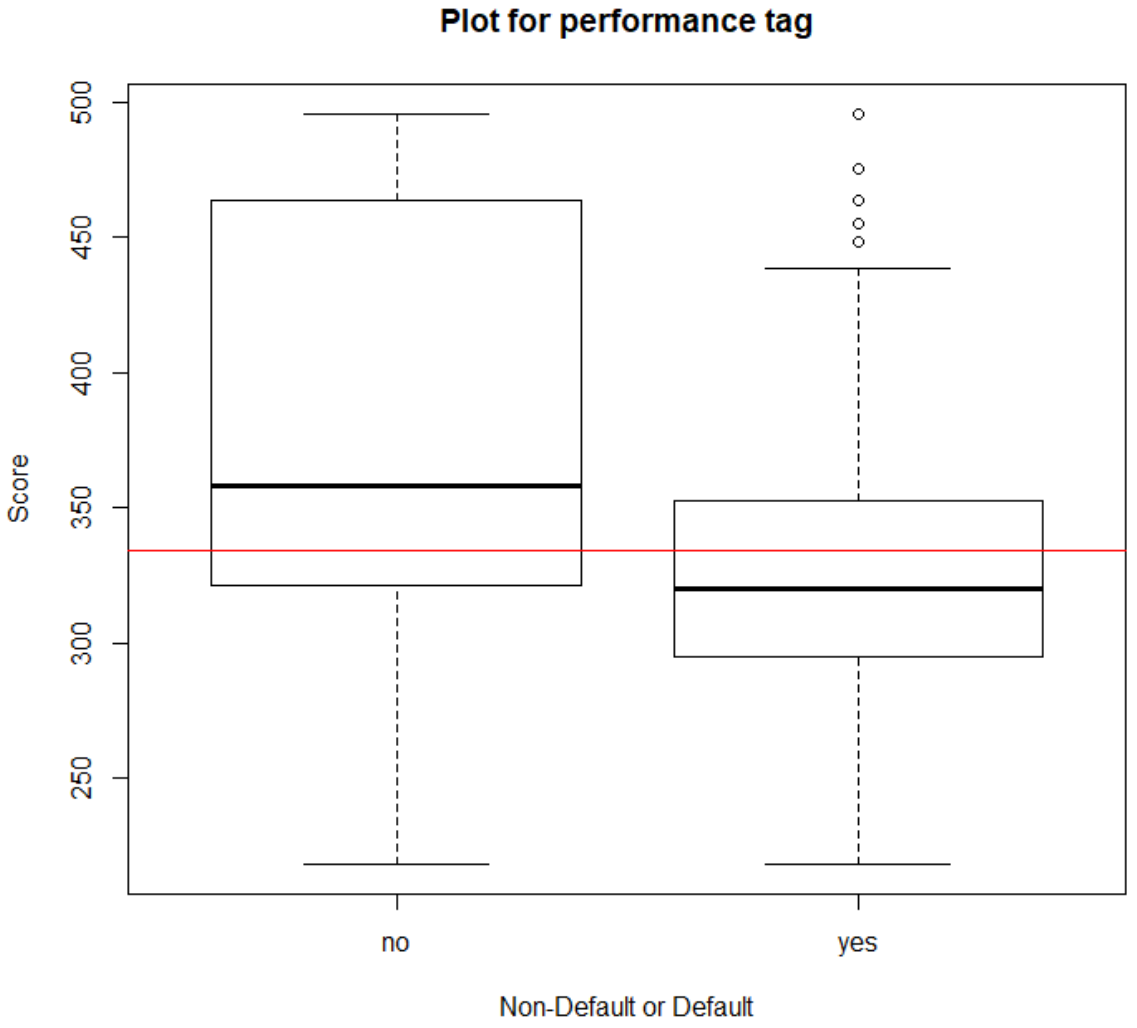
- Around 99.7% customers would had rejected for the credit card application for which we do not know if they would default or not (Performance.Tag is NULL)



CREDX RISK ANALYTICS CASE STUDY

Incorrect classifications before the model was applied to the dataset Rejected(yes) & Accepted(no)

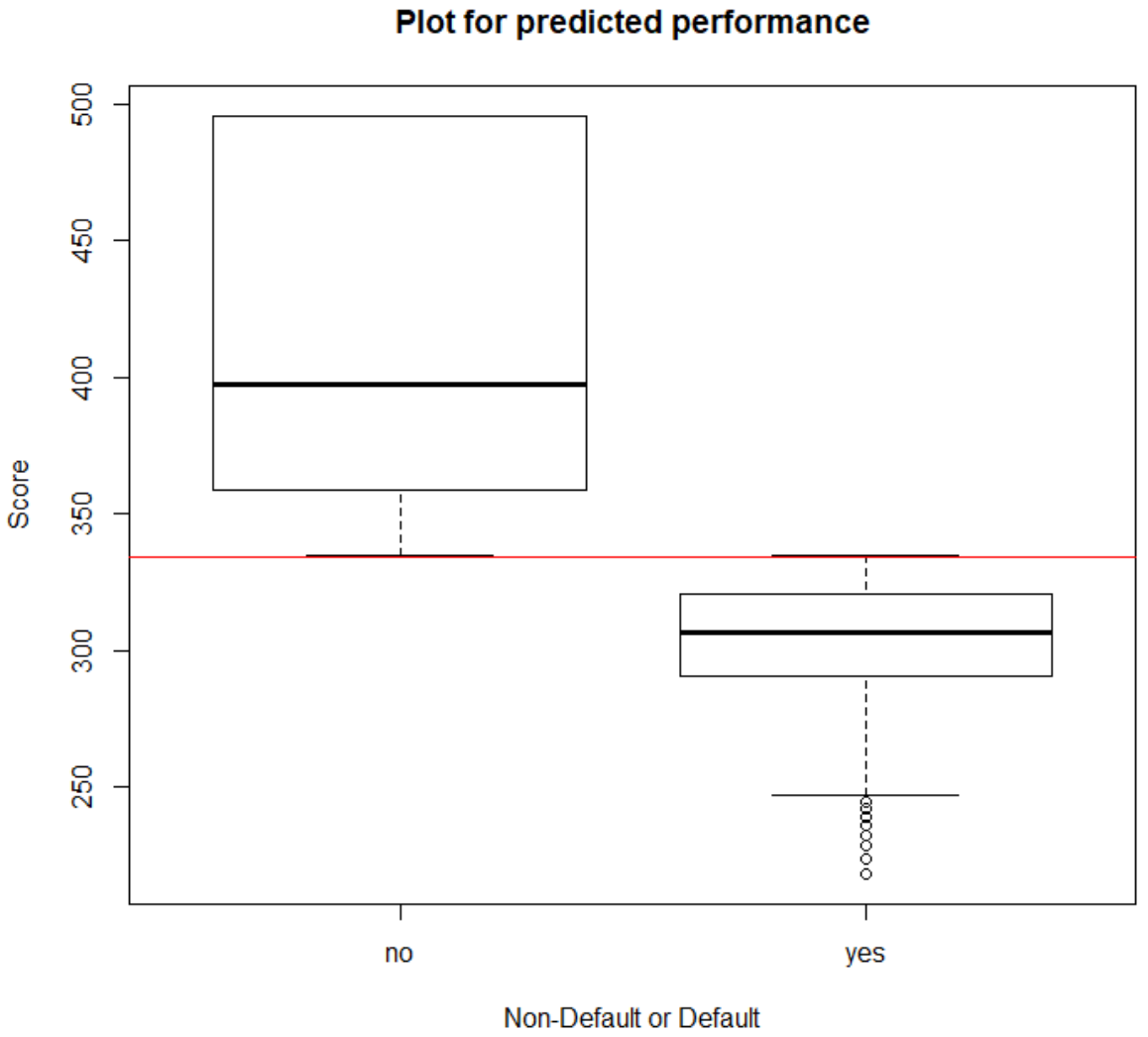
- Application scorecard cutoff was build from the final evaluated model.
- As per image, we can observe that there are lot of misclassifications on either sides for granting / rejecting a credit card applications leading to credit losses
- We shall reduce the misclassifications on using the model.



CREDX RISK ANALYTICS CASE STUDY

Rejected(yes) vs Accepted(no) Application Scorecard values.

- Application scorecard cutoff was build from the final evaluated model (Random Forest).
- As per image, we can observe that there are no or minimal misclassifications which help us with Auto Approval / Auto Rejection of the credit card applications.
- This helps us in correctly classifying the applicants which could default and which would turn out to be good customers.

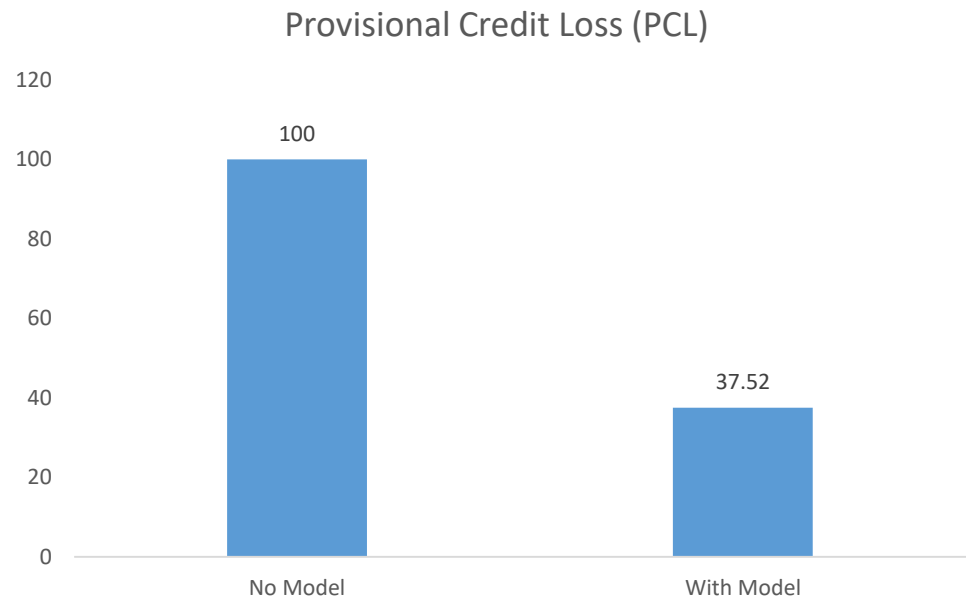


CREDX RISK ANALYTICS CASE STUDY

Impact on Provision for Credit Losses and Net Savings in Credit Losses

The provision for credit losses (PCL) is an estimation of potential losses that a company might experience due to credit risk.

- PCL lost without model = 100%
- PCL lost with model = 37.6%
- PCL Saved = 62.4%



- Assuming around 10000$ loss per customer, below are the findings.
  - Loss without Model: 29470000$
  - Loss with Model in place: 11080000$
  - Expected Loss saved: 18390000$(More than **18 million dollars**)
- Potential loss of revenue due to rejection of good customers, assuming around 500$ loss per customer.
  - Total number of Good Customers Rejected (22704), Expected net loss in revenue due to rejecting good customers (11352000) (Around **11.35 million dollars**)

Confusion Matrix		Actual Defaults	
		Good Customer	Bad Customer
Predicted	Good Customer	44216	1108
	Bad Customer	22704	1839

CREDX RISK ANALYTICS CASE STUDY

Conclusion

- We have to apply the model to reduce the potential credit loss to the company. Random forest model can predict with 66% accuracy on the customers who are likely to default or not.
- The credit loss incurred without using the model was 4.21% whereas on applying the predictive model, the credit loss incurred was reduced down to 1.58%.
- This would imply a reduction in funds kept aside for potential revenue losses by 2.6%
- For future scope, we can work on newer data as and when available to fine tune the model to increase the accuracy.