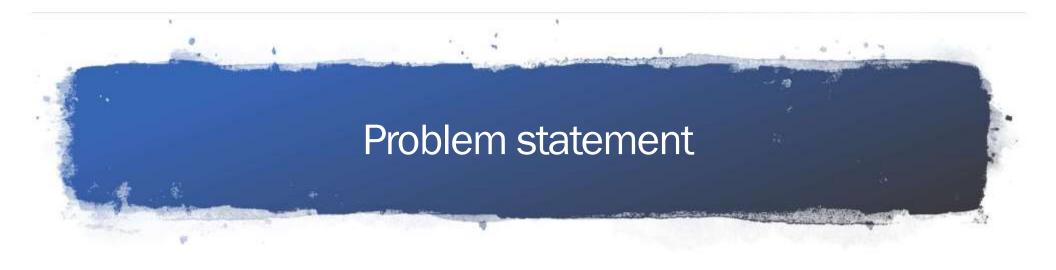


- Group Members:
  - Gaurav Tanwar
  - Harshita Mokkarala
  - Issac Abraham Alummoottil





To mitigate credit risk to 'acquire the right customers'.



To help CredX identify the right customers using predictive models.
Using past data of the bank's applicants



To determine the factors affecting credit risk

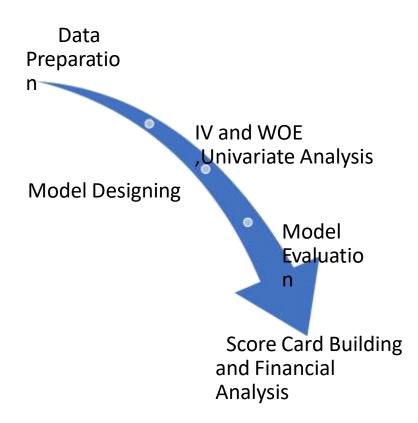


Create strategies to mitigate the acquisition risk and assess the financial benefit to CredX

# Data understanding

- Two data sets are present demographic and credit bureau data.
- Demographic/application data
  - 71295 records of 12 features is available
  - 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:
  - 71295 records of 19 features is available
  - This 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.
- Performance Tag is the target variable consisting of O(non-default) and 1(default)
- 4% of the data contains information about the defaulters i.e 2894/71295

## Methodology Followed





#### Data

# Preparation • Merged the Demographics and

- Merged the Demographics and Credit Bureau Data about Application Id
- Checking all the columns for NAs
- Checking all the columns for missing values
- Outliers detection using the quartiles
- Converting the categorical variables to factors
- Dummy variables creation for the necessary features



- Age contains values less than and Equal to Zero
- Income also contains negative values
- Other features seems to contain the outliers
  - NOTE: to avoid any loss of information contained in these variables, no capping will be done

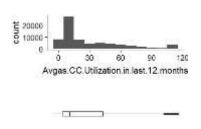


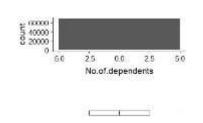
- Three duplicate Application IDs were removed
- 4.5% of the data contains NA values
- Target Variable is seen to contains most of the NA values, 2% of all the records i.e 1425/71295
- Few other features that contain NA values are Gender, Marital Status, Education, Profession, Type of Residence, Average CC utilization, No of trades opened in last 6 months, Presence of Open Home loan, Outstanding Balance
- There are no blank values present in the dataset
- All the records containing NA values are removed
- Data with 68696 records of 29 variables is left for analysis
- Minimum age is capped to 15 years
- Minimum income is capped to 0

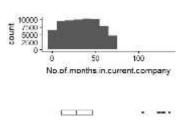
# ASSUMPTIONS TAKEN

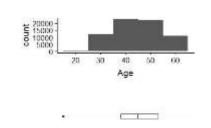
- The Data for only approved applications will be used for model creation i.e NAs will be removed from the target variable
- Removing all the NAs from the data is not causing any information loss
- Capping of the data in continuous features should not be performed as this might cause a loss of information present in them
- NA values in the target variable represent the rejected applicants

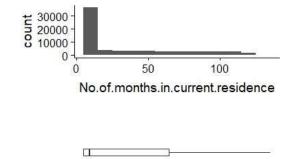


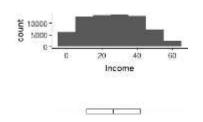




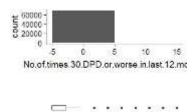


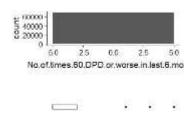


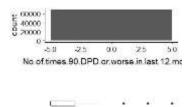




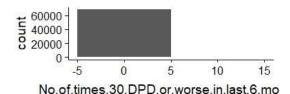
Univariate Analysis

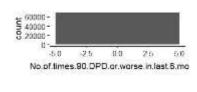




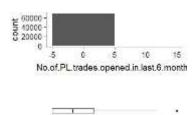




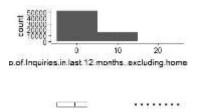


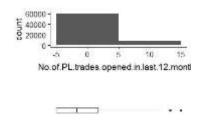


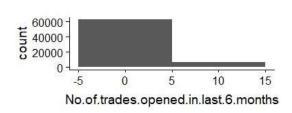
# UNIVARIA TE ANALYSIS CONT.

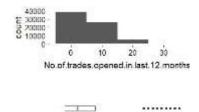








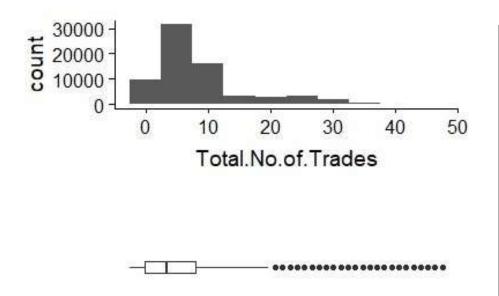


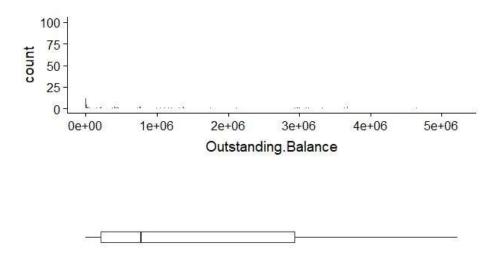


\_\_\_\_\_\_

# UNIVARIA TE ANALYSIS CONT.

#### UNIVARIATE ANALYSIS CONT.



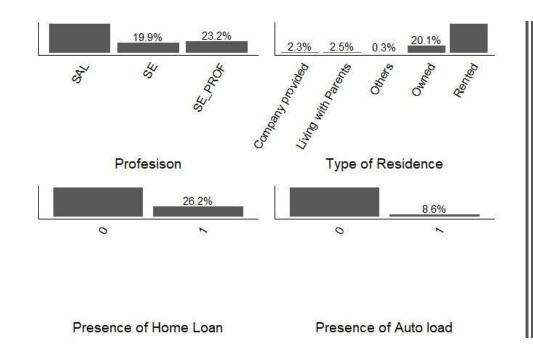


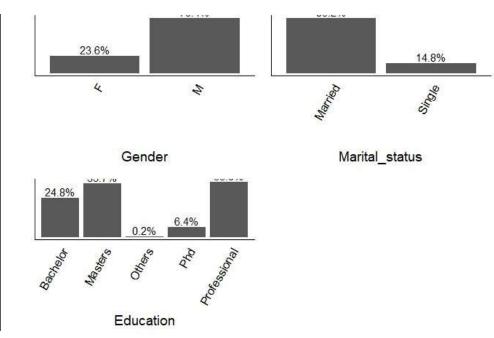
#### **Inference**

- Avg CC uilization (last 12 months): The graph shows that majority of people use 10-15 times Avg CC in last 12 months and this number is quite high compare to other utilization pattern.
- No. Of dependents: With no outliers and and straight horizontal lined graph shows a there is simmilar count for each no of dependant.
- No. of months in current company: With few outliers, majority of people are in 10-50 months range in current comapny.
- Age: At the range of 35-55 the count for the age is above 200000, on the other hand and at the two extremes, the count is less thaan 15000 for the range 25-35 and 55-65.
- No of months in current residence: A tower in the graph shows that there a very high number of records for 1-10 months in current residence adn rest are very few in numbers
- Income -From 10-45k Income shows alomst similar in pattern, except the two extremes which explains
  the low income for the frehser and old age.
- No. of times 30 DPD or worse in last 12 month, No. of times 30 DPD or worse in last 6 month, No. of times 30 DPD or worse in last 6 month, No. of times 60 DPD or worse in last 12 month, No. of times 60 DPD or worse in last 6 month, No. of times 90 DPD or worse in last 6 month : they have range consistent pattern from -5 to 5 and all have few outliers.
- No. of PL trade open in last 6 or 12 months: Both the graph shows similar trend from -5 to 5 except PL trade opened for the last 12 months shows few values for 5-15
- No. of enquiries in 6 month and 12 month: Both the graph shows similar trend i.e Very high in number in the range of -5 to 5 and very few value from 5 to 15
- No of trades in last 6 month, 12 months and total trades: Exceptionally all the three graphs shows
  different pattern, for last 6 month the patter is similar to pl trades and inquiries i.e high value from -5
  to 5 aand low value from 5 to 15. For 12 months there is step wise decrease can be seen but the total
  trades shows a tower for the range in between 0-10 i.e around 5 and have a lot of outliers.
- Outstanding balance: It shows a very distributed graph but a tower showing comparatively high number of values for the range near 0.



#### UNIVARIATE ANALYSIS





# Inference

- •Profession: The salaried profession outnumebered other by decent margin
- •Type of residence : Here, Rented outnumebered other by decent margin
- •Gender and Marital status: Male and married are very large in number compared to that of Females and unmarried respectively.
- •Presence of Auto or Home loan: High number of 0s than 1s shows the high number of absence of home and auto loadn respectively.
- •Education: Professionals, Masters and Bachelors degree holders are high in number on the other hand phd and others are less in number.

#### WOE Analysis

- No.of.Inquiries.in.last.12.months..excluding.home...auto.lo ans.
- Avgas.CC.Utilization.in.last.12.months
- No.of.PL.trades.opened.in.last.12.months
- No.of.trades.opened.in.last.12.months
- · Outstanding.Balance
- Total.No.of.Trades
- No.of.times.30.DPD.or.worse.in.last.6.months
- No.of.PL.trades.opened.in.last.6.months
- No.of.times.90.DPD.or.worse.in.last.12.months
- No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.
- No.of.times.60.DPD.or.worse.in.last.6.months
- No.of.times.30.DPD.or.worse.in.last.12.months
- No.of.trades.opened.in.last.6.months
- No.of.times.60.DPD.or.worse.in.last.12.months
- No.of.times.90.DPD.or.worse.in.last.6.months

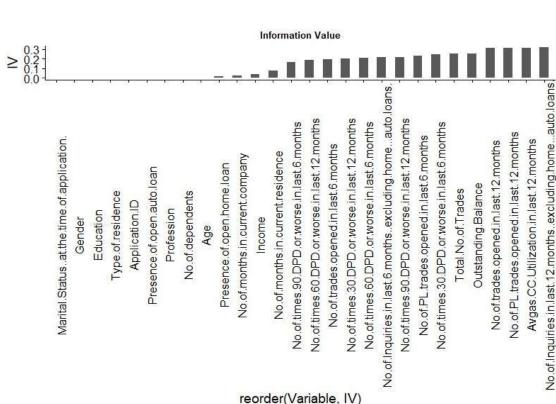
#### Interpretation of the IV value:

< 0.02 useless for prediction</li>

0.02 to 0.1 Weak predictor

0.1 to 0.3 Medium predictor

0.3 to 0.5 Strong predictor



#### Logistic Model - Result and Stats

#### Confusion Matrix and Statistics

#### Reference

Prediction No Yes No 13317 359 Yes 6445 488

Accuracy: 0.6699

95% CI: (0.6634, 0.6763)

No Information Rate : 0.9589

P-Value [Acc > NIR] : 1

Kappa: 0.0563

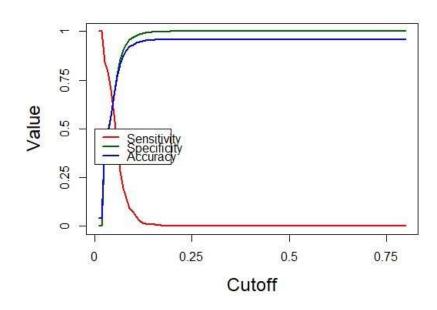
Mcnemar's Test P-Value : <2e-16

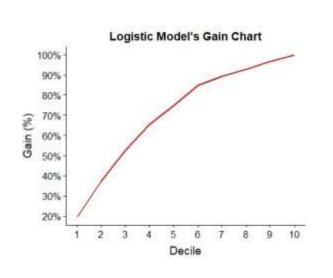
Sensitivity: 0.57615 Specificity: 0.67387 Pos Pred Value: 0.07039 Neg Pred Value: 0.97375 Prevalence: 0.04110

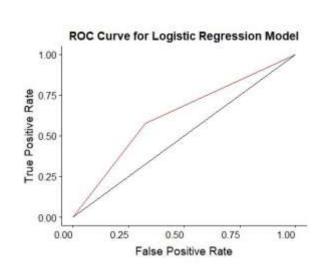
Detection Rate: 0.02368 Detection Prevalence: 0.33641

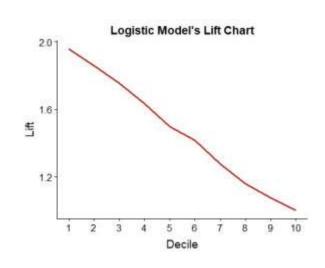
Balanced Accuracy: 0.62501

'Positive' Class: Yes



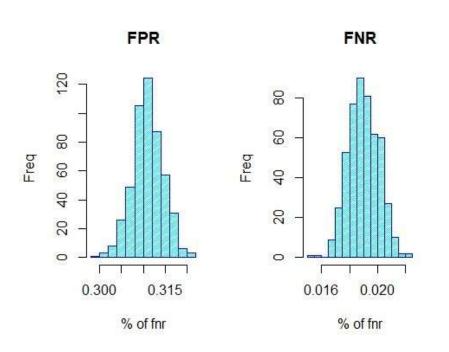


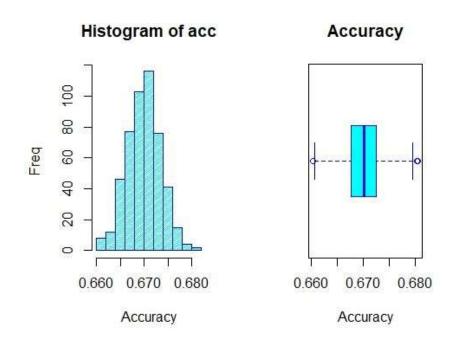




## **Logistic Model Evaluation**

#### Cross Validation Results On Logistic Regression





## Logistic Regression Using SMOTE

#### Confusion Matrix and Statistics

#### Reference

Prediction No Yes No 12347 329 Yes 7415 518

Accuracy: 0.6242

95% CI: (0.6176, 0.6309)

No Information Rate : 0.9589

P-Value [Acc > NIR] : 1

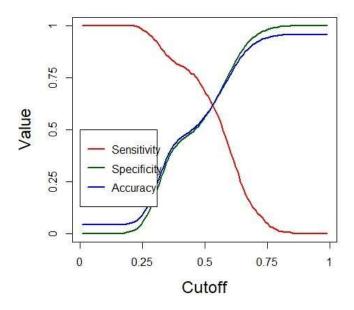
Kappa: 0.0472

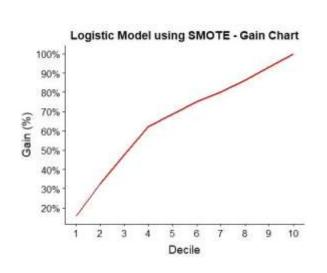
Mcnemar's Test P-Value : <2e-16

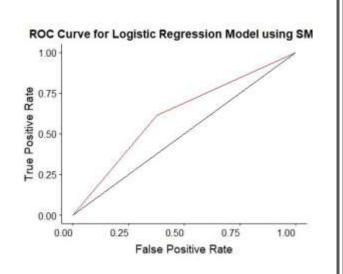
Sensitivity: 0.61157 Specificity: 0.62478 Pos Pred Value: 0.06530 Neg Pred Value: 0.97405 Prevalence: 0.04110 Detection Rate: 0.02513 Detection Prevalence: 0.38493

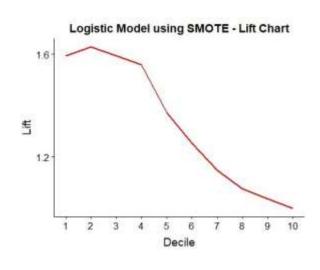
Balanced Accuracy: 0.61818

'Positive' Class: Yes



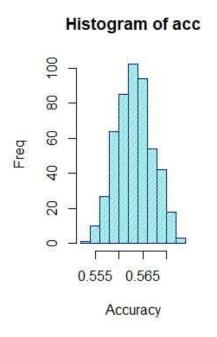


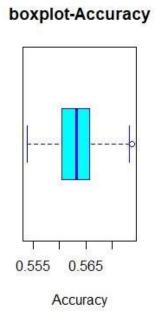


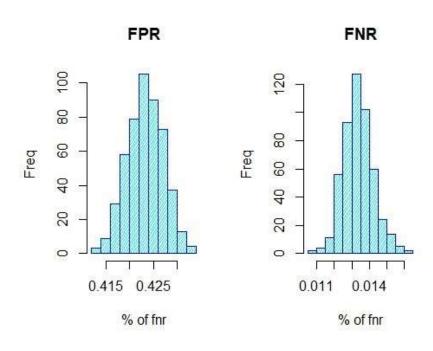


## Logistic Model Using Smote - Evaluation

#### Cross Validation Results On Logistic Regression Using SMOTE



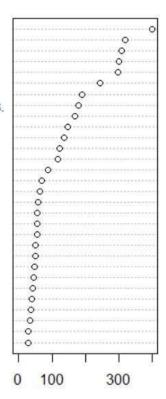




# Random Forest – Important Variables Predicted

#### model\_rf

```
Outstanding Balance
No.of months in current company
Avgas. CC. Utilization.in.last.12.months
Income
Age
No.of months in current residence
Total No.of. I rades
No.of. Inquiries in. last.12.months... excluding home... auto.loans.
No.of trades opened in.last.12.months
No.of trades opened in.last.6.months
No.of. Inquiries in.last.6.months... excluding home... auto.loans.
No.of. Itrades opened in.last.6.months
No.of. PL. trades opened in.last.6.months
No.of. Imes.30.DPD.or.worse.in.last.12.months
No.of. times.30.DPD.or.worse.in.last.12.months
No.of. times.30.DPD.or.worse.in.last.12.months
No.of. times.30.DPD.or.worse.in.last.12.months
No.of. times.30.DPD.or.worse.in.last.12.months
Profession.xSE.DPD.or.worse.in.last.6.months
Profession.xSE.DPD.or.worse.in.last.6.months
Profession.xSE.Type.of.residence.xRented
Marital Status...at.the time.of.application.
Type.of.residence.xOwned
No.of.times.60.DPD.or.worse.in.last.6.months
Presence.of.open.home.loan
Presence.of.open.home.loan
No.of.times.90.DPD.or.worse.in.last.6.months
```



MeanDecreaseGini

#### Random Forest - Result and Stats

#### Confusion Matrix and Statistics

#### Reference

Prediction no yes no 9762 223 yes 10000 624

Accuracy: 0.504

95% CI: (0.4971, 0.5108)

No Information Rate : 0.9589

P-Value [Acc > NIR] : 1

Kappa: 0.0354

Mcnemar's Test P-Value : <2e-16

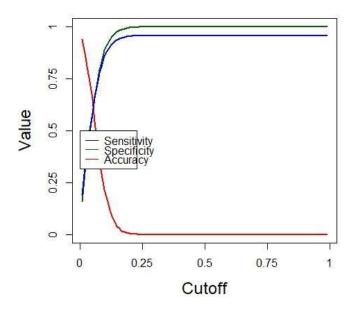
Sensitivity: 0.73672 Specificity: 0.49398 Pos Pred Value: 0.05873 Neg Pred Value: 0.97767

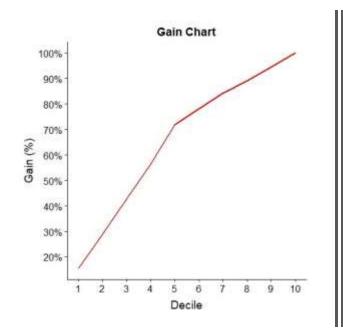
Prevalence: 0.04110 Detection Rate: 0.03028

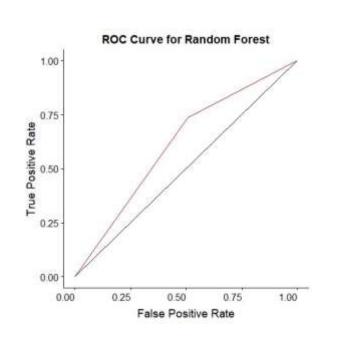
Detection Prevalence : 0.51550

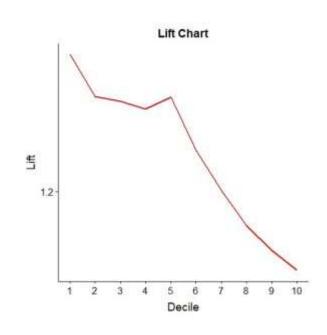
Balanced Accuracy : 0.61535

'Positive' Class : yes









#### **Random Forest Model Evaluation**

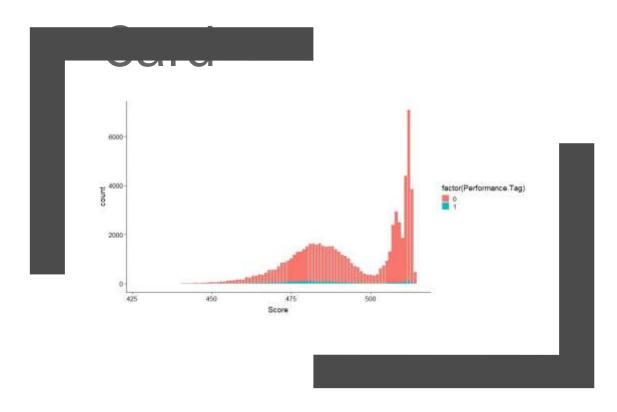
## Assumptions while Designing

Only Logistic Regression and Random Forest were used as the model building techniques because these are the simplest and fastest when the data size is huge

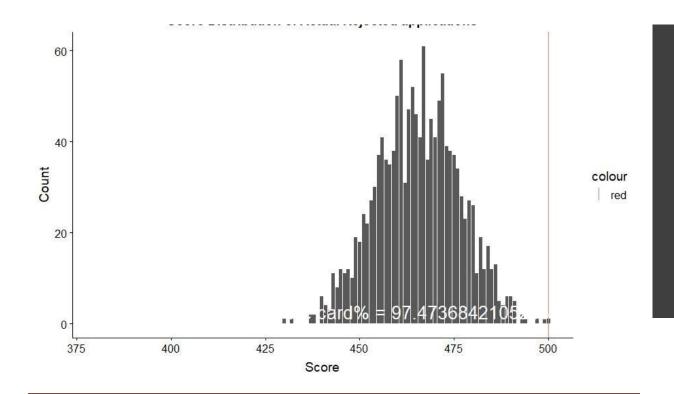
- Decision Trees are not used, an ensemble(RandomForest) is already applied
- It is never recommended to use SVMs in BFSI as the size of data can be really large and it would require a lot higher computing power
- Finally the most stable i.e Logistic Regression Model is used to design the score card



#### Score



- Good to bad odds vary from 10 to 1
- Scores lie in the range of 428 to 514
- There is a dip in the scores, at around 500
- 500 can be taken as the cutoff score for decision
- 82% of the defaulters are captured at 500



Score < 500 captures 97%
 of rejected population,
 who would have defaulted</li>

# Scores for the Rejected Application



#### Financial Benefits of Model

- Current approval rate 98%
- Credit loss 3689945678
- Using Model (using the optimal credit score of 480)
  - Approval rate 76%
  - Credit loss 2075459342
- The Credit loss after applying the model has slashed to half of the Original Credit loss without using any model
- There will be a tradeoff between the increase in approval rate and credit loss increase of one will lead to increase of other
- The current Optimal Score is 480 but can be tuned as per requirement