BFS Capstone Project
Mid-Submission

Group Members

1. Anuj Arya
2. Asim Pattnaik
3. Mohammed Suhail Y
4. Rakesh Bosu

Business understanding

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. The CEO believes that the best strategy to mitigate credit risk is to ‘acquire the right customers’.
In this project, you will help CredX identify the right customers using predictive models. Using past data of the bank’s applicants, you need to determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of your project.

There are two data sets in this project — demographic and credit bureau data.
Demographic/application data: 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: 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.

Both datasets have Performance Tag column which signifies whether the person has defaulted after getting a credit card.

Data Understanding

- Demographic Data
    - Rows - 71295
    - Columns- 12
- Credit Bureau Data

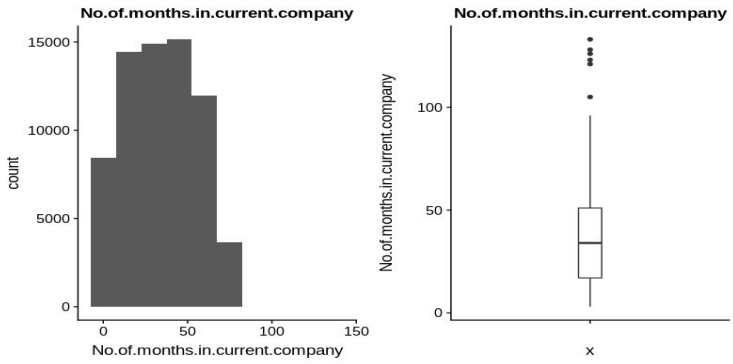
    - Rows - 71295
    - Columns- 19
- 71292 Unique values are found in both datasets. There are 3 Duplicate records in both.

Data cleaning and preparation

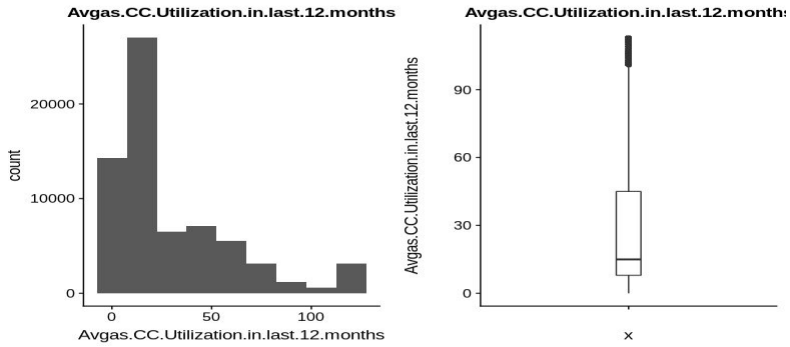
- Remove the duplicate rows in both the datasets.
- Merge both the datasets into one.
- Remove all the records with Age less than 18 as only people above age 18 can apply for credit cards.
- Remove all the records where Income less than or equal to 0 as it is not possible to have negative income. This constitutes only a negligible size of the entire dataset, hence it can be removed.
- Check for NA Values
    - Gender- 1
    - Marital Status- 5
    - No of Dependents- 2
    - Education- 119
    - Profession- 13
    - Type of Residence- 8
    - Avgas CC Utilization in last 12 months- 1053
    - No.of.trades.opened.in.last.6.months -1
    - Presence.of.open.home.loan- 272
    - Outstanding.Balance- 272
    - Performance.Tag- 1425
- Performance tag has many NA values which indicate they were rejected at the onset. Separate these records as a individual dataset and use it later to test the model.
- Remove the remaining NA values from the dataset as they are insignificant in size compared to the entire dataset.

Find Outliers

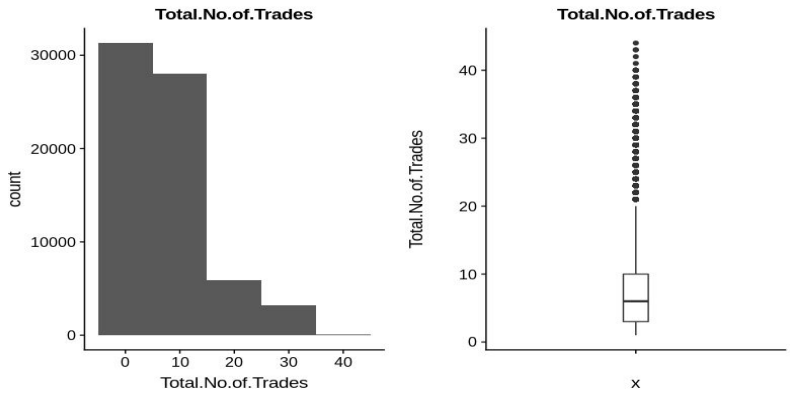
- Outliers were found in the following fields and were treated using the quantile method.
    - No.of.months.in.current.company



        - Sudden jump from 99% to 100%. Cap values to 74
    - Avgas.CC.Utilization.in.last.12.months



        - Sudden jump from 94% to 95%. Cap values to 91
    - Total.No.of.Trades



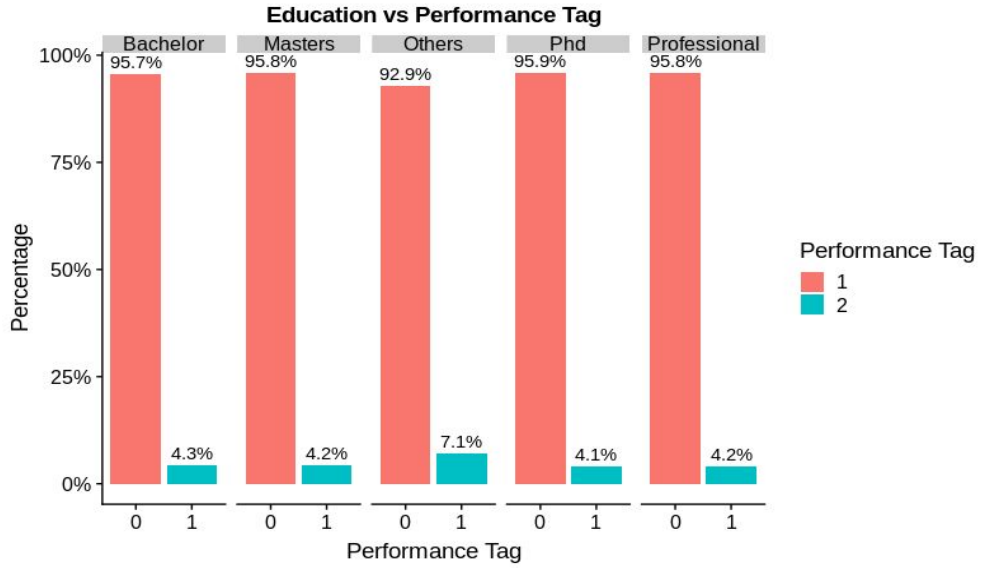
- Sudden jump from 99% to 100%. Cap values to 31

Exploratory Data Analysis

Salary Group vs Performance Tag

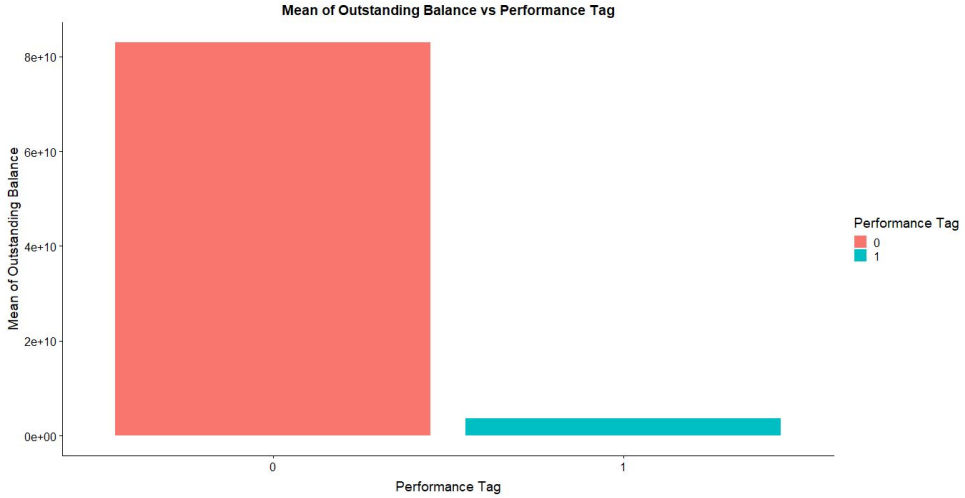
People belonging to Low Income group tend to default more than other groups.

- Education vs Performance Tag



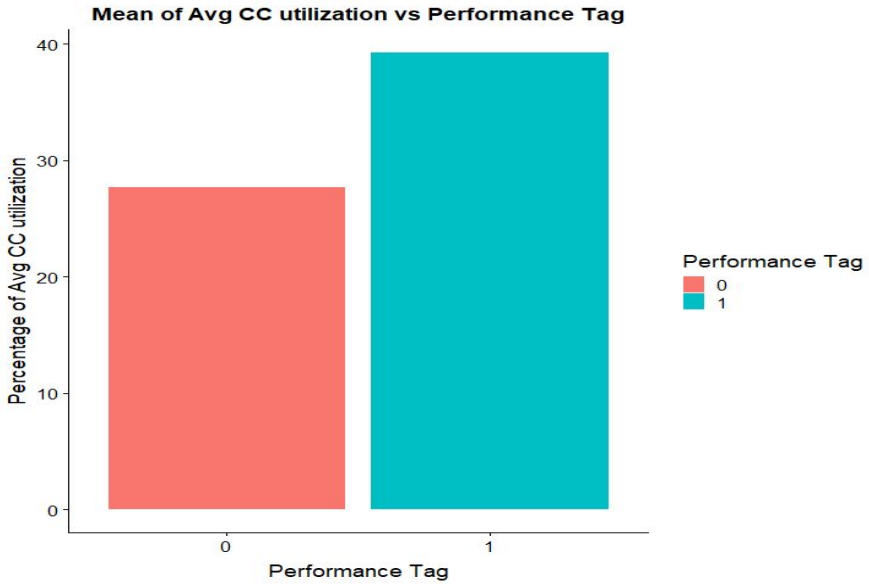
People with Education Group as Others tend to default more compared to remaining groups

- Outstanding Balance vs Performance Tag



  Defaulted customers tend to have much lesser balance than non-default customers

- Avg CC Utilization vs Performance Tag



  Default customers tend to use credit card more than non-default customers

- From EDA, we have identified some variables as strong predictors
  - Gender
  - Marital Status
  - Salary
  - Age

- No. of Dependents
- Education
- No of times 90 DPD or worse in last 6 months
- No of times 60 DPD or worse in last 6 months
- No of Trades
- Avg CC Utilization
- Outstanding Balance

WOE Transformation and Information Values

- WOE Transformation is carried out in all the variables and the Corresponding Information values are calculated to find the important variables.
- The following are the identified important variables based on the Information Value. Considering IV values greater than 0.20 as significant.

Variable	Information Value
Avgas.CC.Utilization.in.last.12.months	0.3196123
No.of.trades.opened.in.last.12.months	0.3075795
No.of.PL.trades.opened.in.last.12.months	0.2686921
No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.	0.2663039
Outstanding.Balance	0.2550174
Total.No.of.Trades	0.2533396
No.of.times.30.DPD.or.worse.in.last.6.months	0.2499031
No.of.PL.trades.opened.in.last.6.months	0.2334098
No.of.times.30.DPD.or.worse.in.last.12.months	0.2228656
No.of.times.90.DPD.or.worse.in.last.12.months	0.2204233
No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.	0.2188786
No.of.times.60.DPD.or.worse.in.last.6.months	0.2156838

Model Building Procedure

- For building the model we use both the Demographic Dataset as well as the Merged dataset of demographic and credit bureau data.
- First build the model using the cleaned dataset and check the accuracy.
- After that, apply the SMOTE method of synthetic data generation in order to handle the problem of class imbalance.
- Once the data is synthetically treated, use three different algorithms (Logistic Regression, Decision Tree and Random Forest) to build the models and determine the cutoff values to find the best values of accuracy, sensitivity and specificity for each model.
- Determine the best model based on R2, AIC, ROC, Accuracy, Sensitivity and Specificity values.

Application Scorecard and Financial analysis

- The Application Scorecard is used to determine the desired tradeoff between risk level and approval rate.
- Build an application scorecard with the good to bad odds of 10 to 1 at a score of 400 doubling every 20 points and find the balanced cutoff value for the score.
- This cutoff score would help us for financial analysis to identify the approval rate and net credit loss.
- Evaluate applicants and determine if they are eligible to get a credit card or if they should be rejected based on whether their score is greater than the balanced cutoff score.