BFS Capstone Project

submitted on June 17, 2019

Submitted by:

Ishan Savio Kerketta

Devanshi Kulshreshtha

Ayushi Gaur

Ajay Sharma

Business Objective

Problem Statement:

CredX, a credit card provider is facing a credit loss and wants to mitigate risk by acquiring the right customers. The objective is to identify the right customers using predictive models and techniques related to Acquisition Risk Analytics. We need to determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of the project.

Business Understanding:

The Credit card company wants to reduce the risk involved with its applicants for credit card.
Credit loss is of 2 types:

- Risky applicants given credit cards resulting in default in payments
- Non-risky applicants not given credit cards resulting in loss of revenue

The company wants to acquire the right customers based on this. This is a **Classification** problem.

Data Available

We have 2 structured datasets:

- **Demographic Data:** It has 71295 observations and 12 variables. The variables are related to an applicant's demographic details like Gender, Marital Status, Income, etc. The target variable is ‘Performance Tag’. If its value is 1 then an applicant defaults on credit card, else he does not.
- **Credit Bureau Data:** It has 71295 observations and 19 variables. It consists of variables informing if the customers have defaulted in previous history of credit cards, about their trades, etc. The target variable is Performance Tag.

Issues resolved in Data Cleaning Stage:

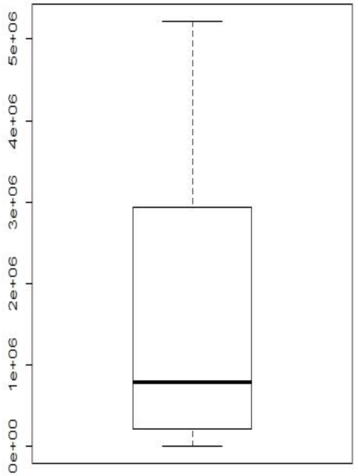
- Presence of 3 duplicates in identifier variable
- Presence of NA values – removed all
- Presence of invalid values e.g. negative age
- Outliers

Outlier Treatment

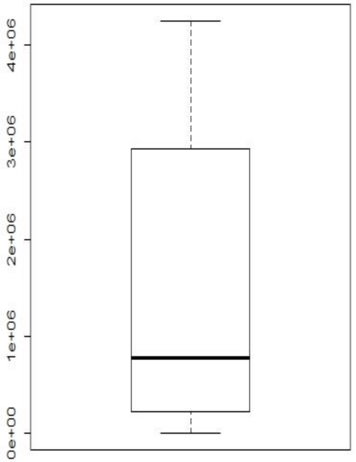
Outliers were found in before 1st percentile of income, after 99th percentile of number of months in current company, after 99th percentile of number of trades, after 99th percentile of Outstanding balance and after 99th percentile of number of trades

Outlier depiction with box plots and percentiles:

Outstanding Balance



Before treatment

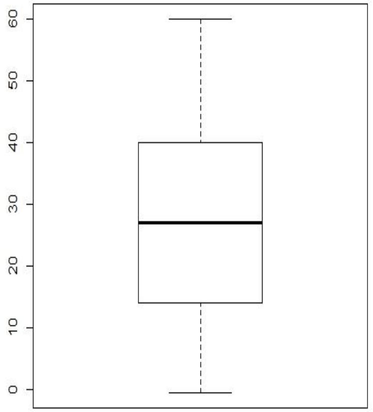


After treatment

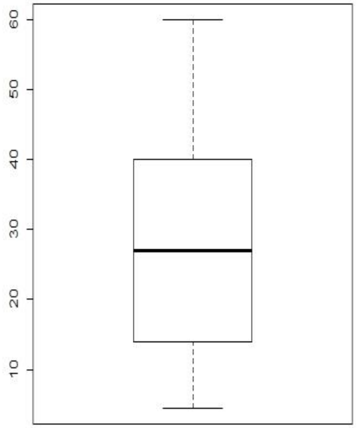
Percentiles

98%	99%	100%
4035188.90	4251676.10	5218801.00

Income



Before treatment



After treatment

Percentiles

0%	1%	2%
-0.5	4.5	4.5

Problem Solving Approach

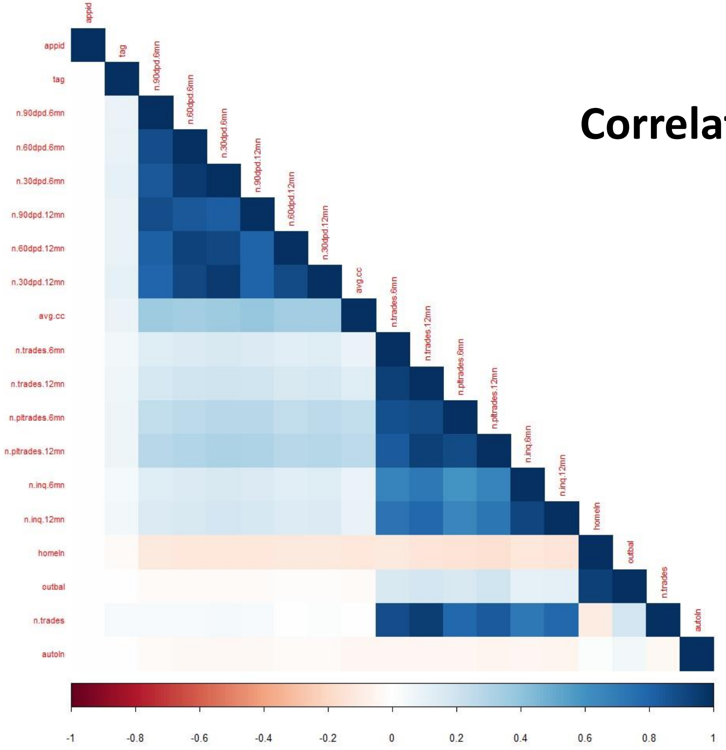
Steps

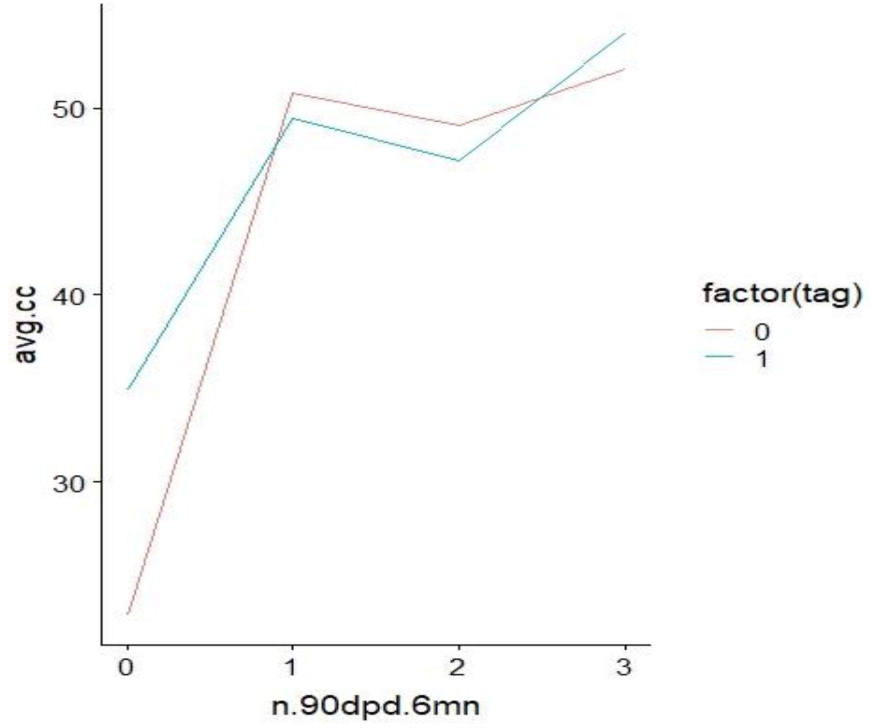
After data cleaning and merging the two datasets, we performed the following steps:

- Binning categorical values in Demographic data
- Extensive Exploratory Data Analysis to detect important predictors and insights
- Association rule mining for important predictors
- Weight of Evidence and Information Value Analysis
- Building an evaluation of Logistic Regression Model: It is a **classification problem** and this model gives us linear relationship
- Building a Random Forest Model: We will feed the important predictors obtain through logistic regression to a Random Forest model to check for **better performance**
- Creating an Application Scorecard
- Assessing Financial benefit of the project using Gain-Lift chart

Exploratory Data Analysis
Relationship between variables

Correlation Plot



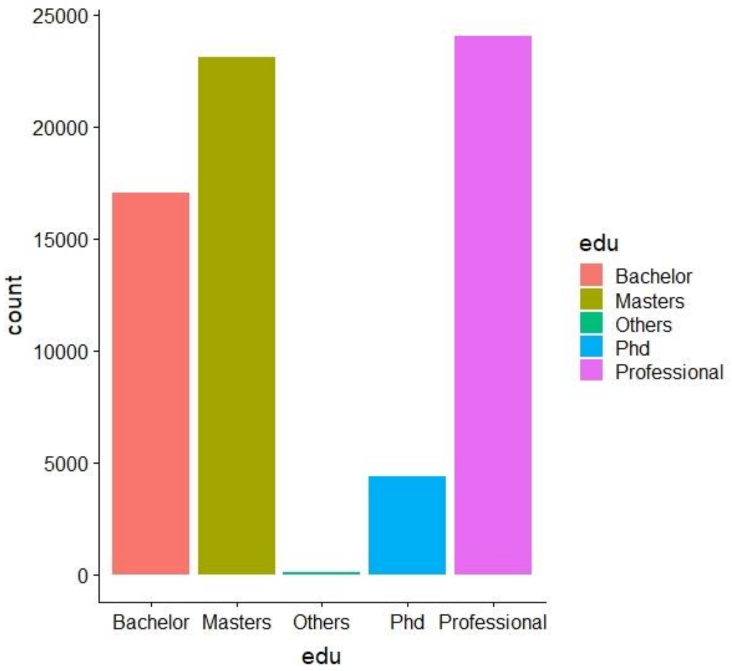


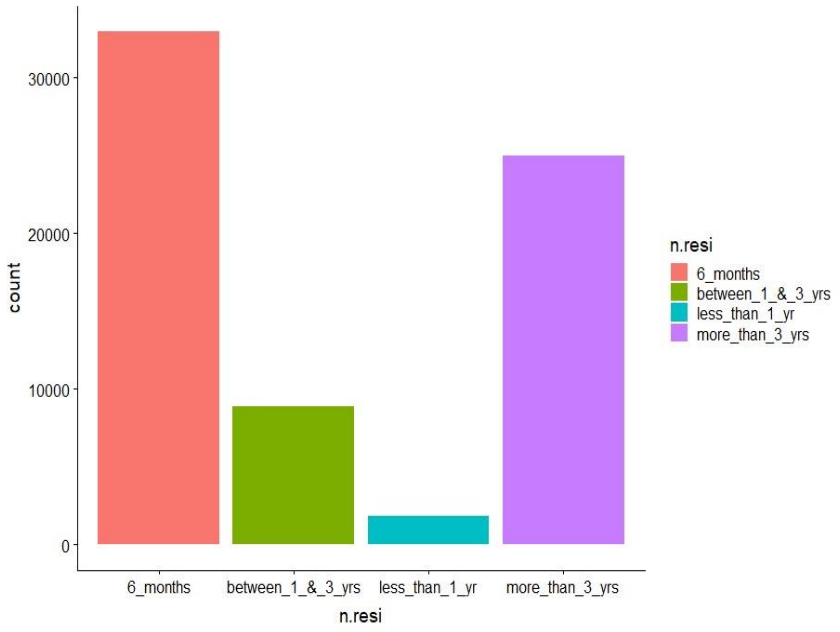
- The instances of 90/60/30 days past due in the last 6 to 12 months are correlated with each other.
- Outbalance is correlated with The Presence of home loan.
- Number of trades inquired are correlated with the number of inquiries made.

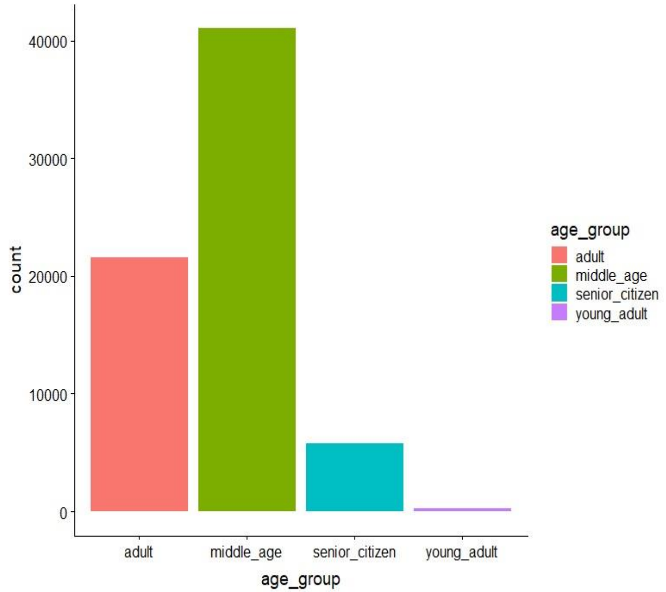
- Average Credit Utilization rises with an increase in the instances if 90 days past due instances in the last six months.

EDA - Univariate Analysis

Some of the important observations during EDA:







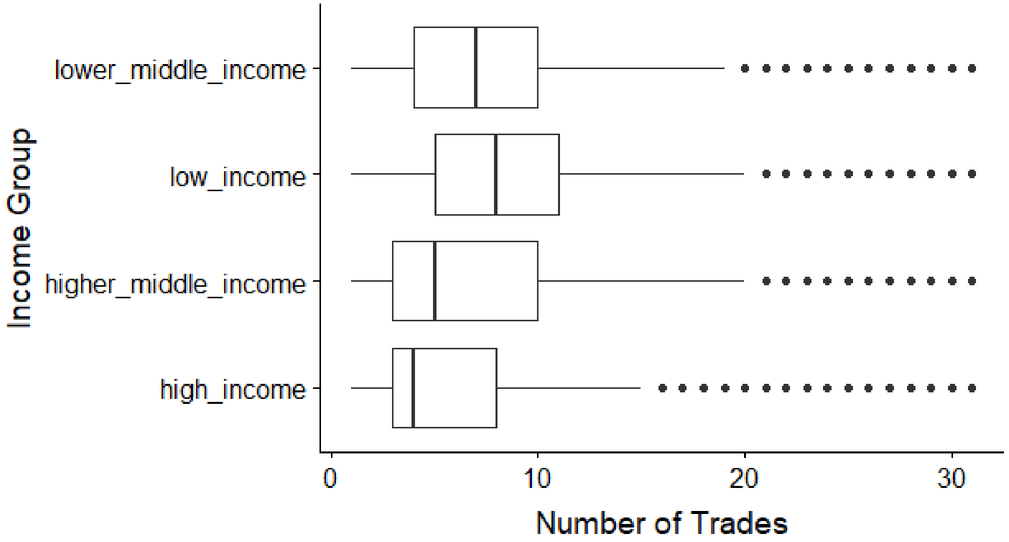
Most credit card applicants have a Professional education, followed by Masters and Bachelors

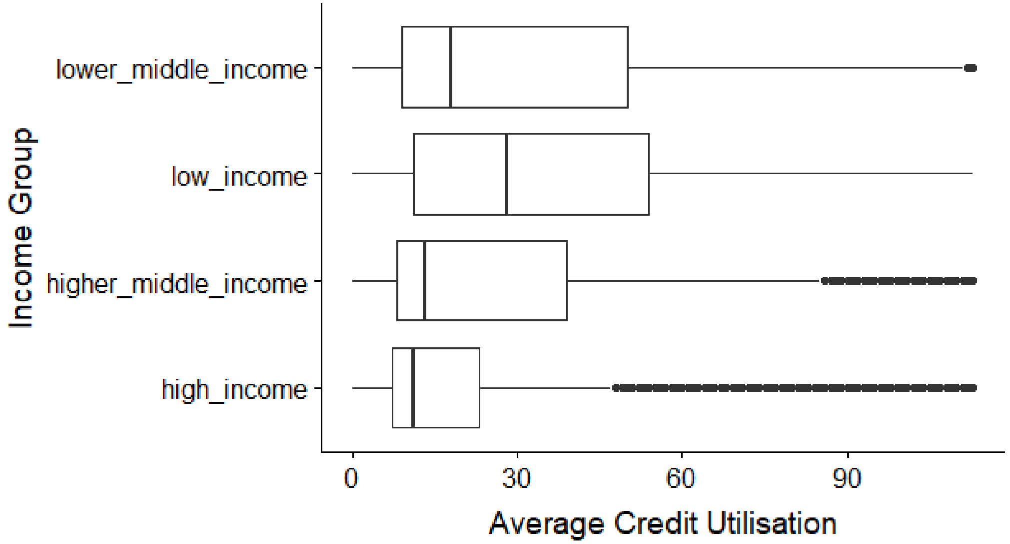
Most applicants are in their current residence since the last 6 months

Most applicants belong to middle age group

EDA - Bivariate Analysis

Some of the important observations during EDA:



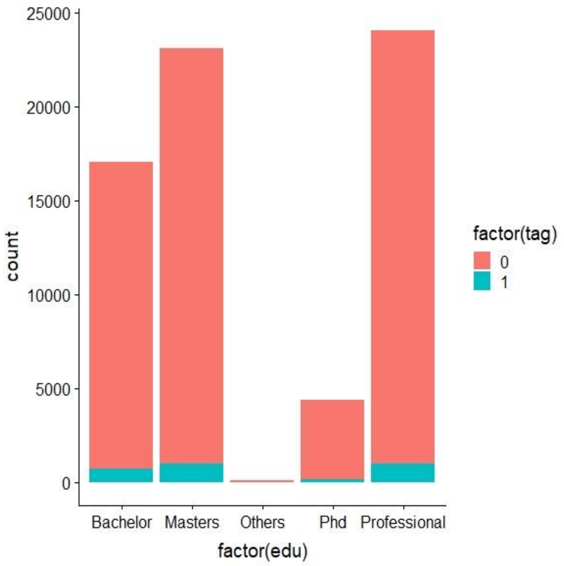


- Income Group has a bearing on the number of trades- low income groups make the most number of trades.

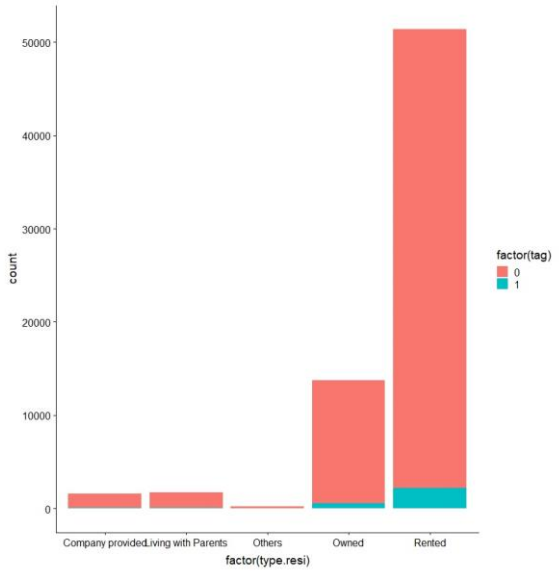
- Low income groups also have the highest credit utilization.

Predictors of Performance Tag

Some of the important observations during EDA:

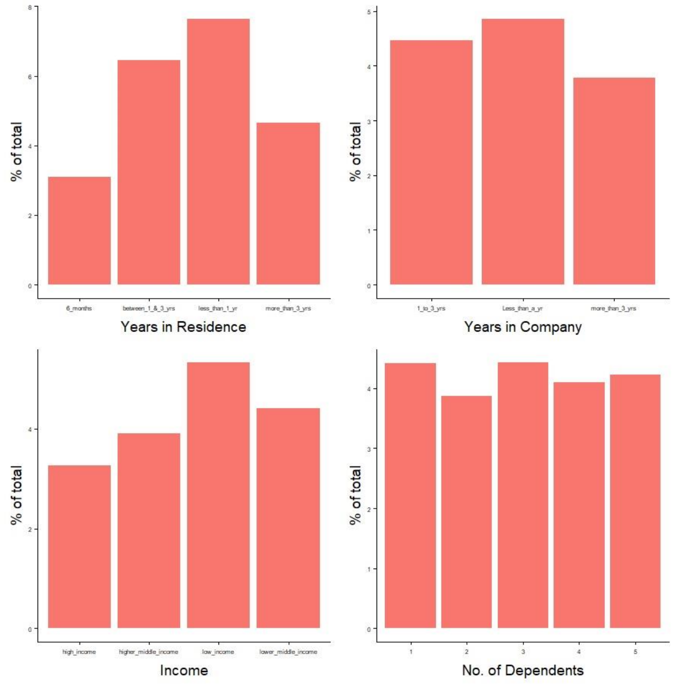


Education



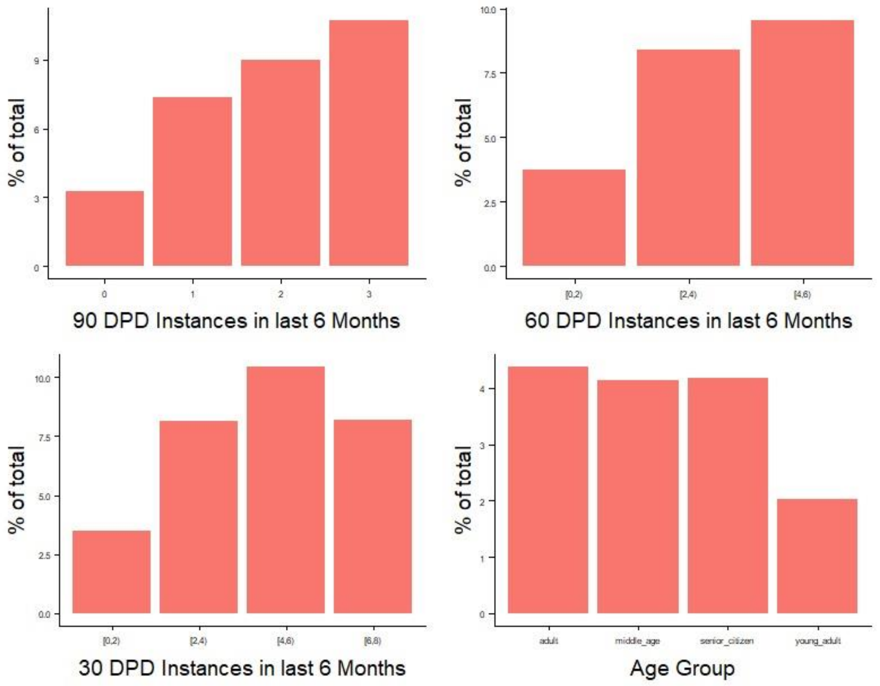
Type of Residence

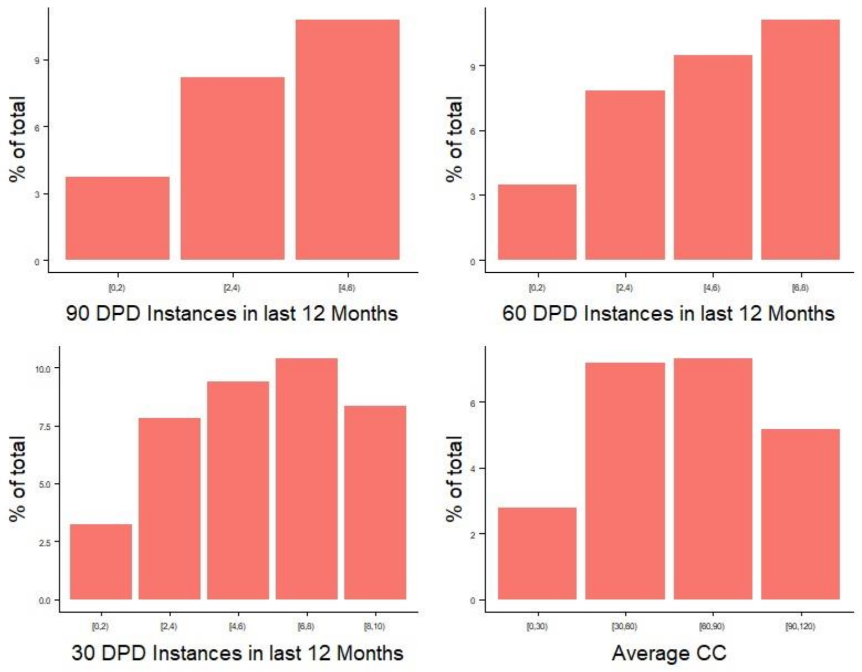
- Most applicants have a Professional Education, followed by Masters Education and Bachelors Education. Defaulters vary as per proportion.
- Most people live in rented houses and hence majority of the customer behavior can be traced here



- No clear linear trend is seen in the number of dependents and defaulting on a loan.
- Those at the same place of residence for 6months- 1year are most likely to default.
- Those at their company for less than an year are also more likely to default.
- Low income is also a good predictor of default.

Predictors of Performance Tag

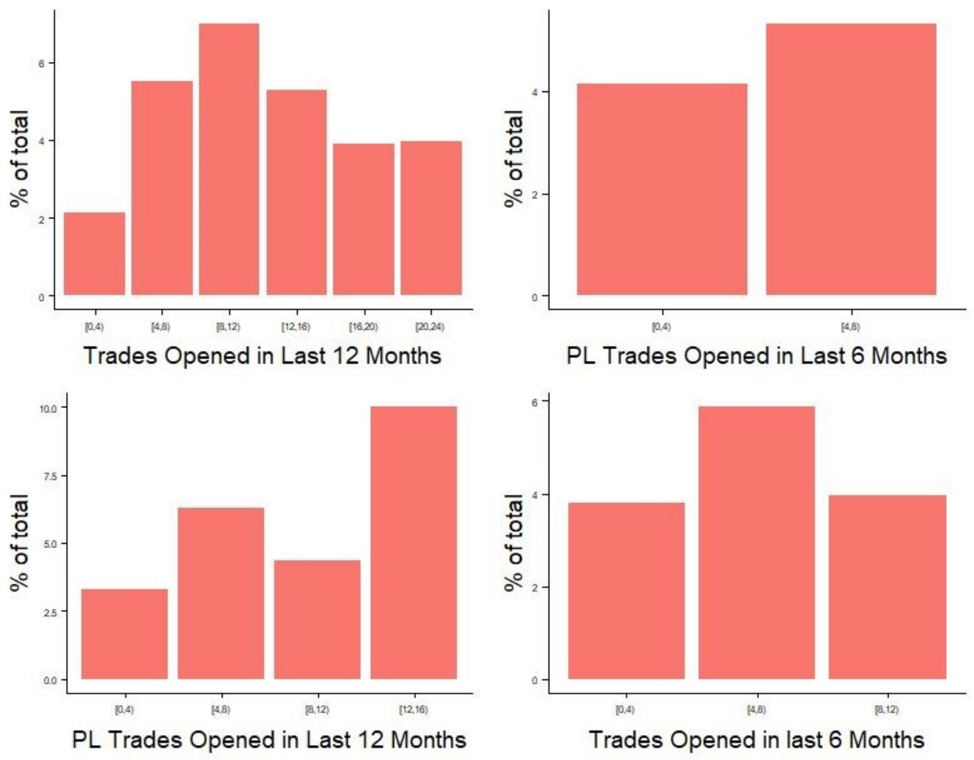


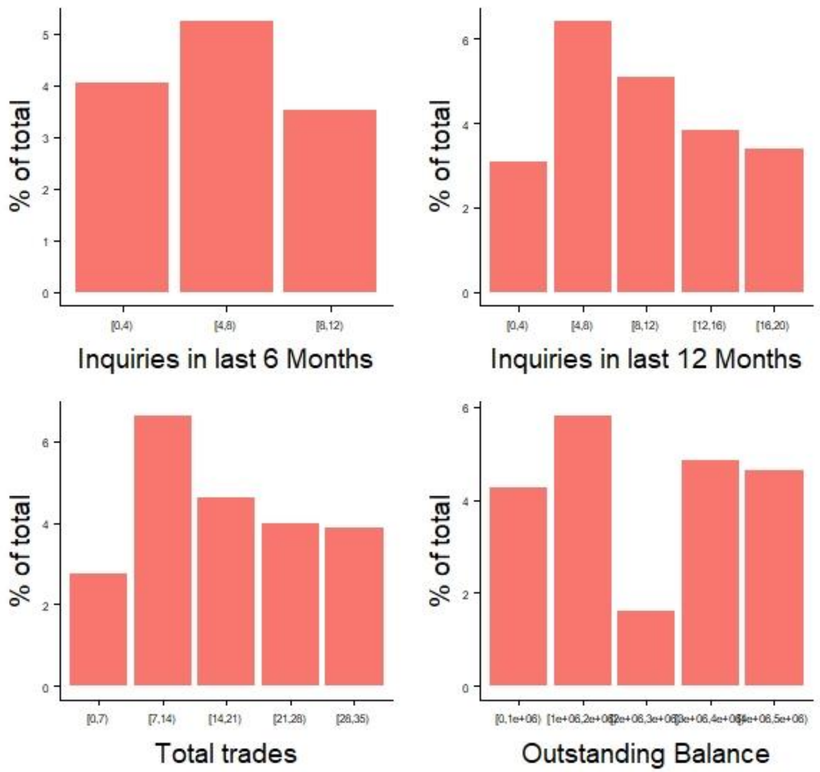


- Young adults are the safest best to extend credit cards to .
- As the number of DPD incidents increase, the chances of defaulting also increase.

- The trends with days past due continues when observed over a 12 month window.
- Those with average credit utilization between 30-90 are far more likely to default with others.

Predictors of Performance Tag



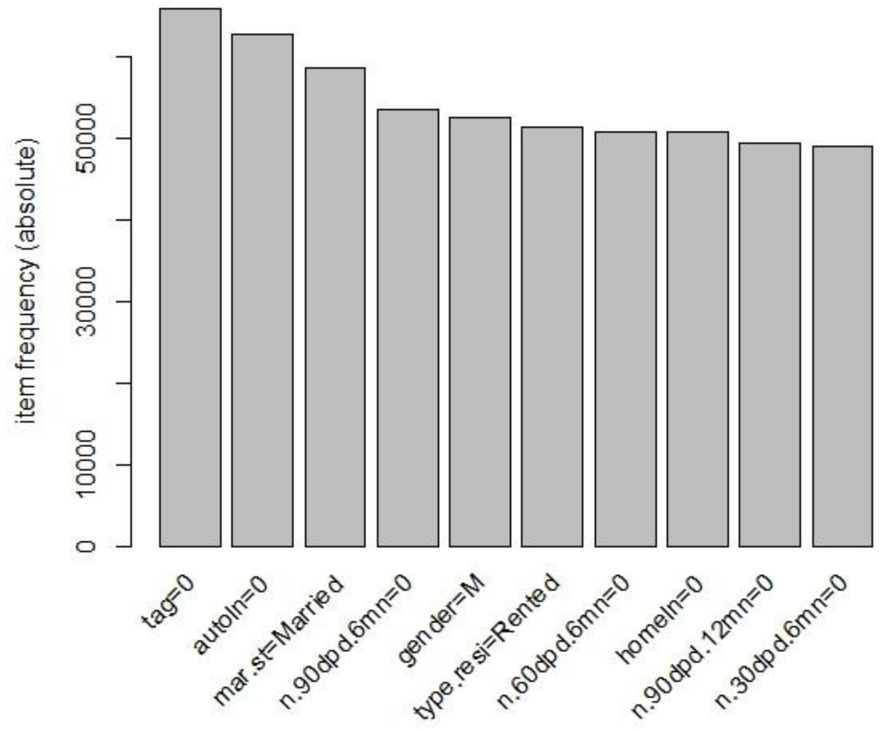


- Increasing number of PL trades increases the chances of defaulting.

- No clear linear trend of performance tag is seen against Inquiries made in the last six or twelve months, total number of trades or the outstanding balance.

Association Rule Mining

- More than 3 years in current residence, Zero PL trades opened in last 12 months and Zero Inquiries in last 6 months excluding home & auto loans paired with Performance Tag 0 shows highest lift
- More than 3 years in current residence, Zero PL trades opened in last 6 months, Zero PL trades opened in last 12 months and Zero Inquiries in last 6 months excluding home & auto loans paired with Performance Tag 0 shows second highest lift
- More than 3 years in current residence, zero times 90 DPD or worse in last 12 months and Zero PL trades opened in last 12 months paired with Performance Tag 0 shows third highest lift



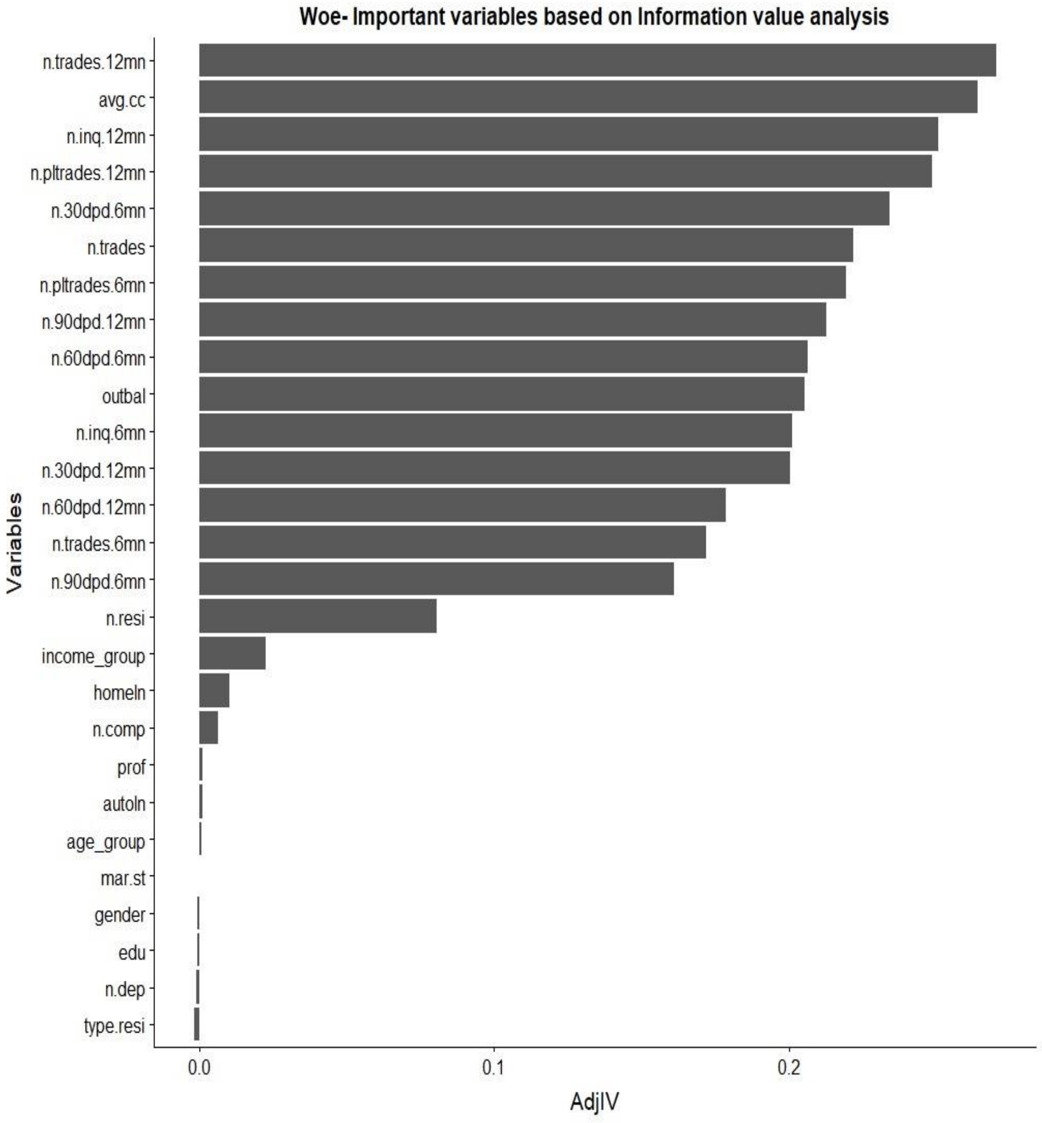
Item Frequency Plot

Note: Performance Tag 0 stands for non-defaulters

Weight of Evidence and Information Value

After performing an analysis on WOE and IV, we found that No. of trades opened, average credit card utilization and number of inquiries of the last 12 months are the top 3 predictors.

Variable	Information Value	Penalty	Adjusted IV
No.of.trades.opened.in.last.12.months	0.3121446	0.0421720	0.2699726
Avgas.CC.Utilization.in.last.12.months	0.3232310	0.0593382	0.2638928
No.of.Inquiries.in.last.12.months	0.2990086	0.0485914	0.2504172
No.of.PL.trades.opened.in.last.12.months	0.3016099	0.0531713	0.2484386
No. of rimes 30 DPD in last 6 months	0.2586435	0.0245893	0.2340541
Total No. of trades	0.2490031	0.0273583	0.2216448

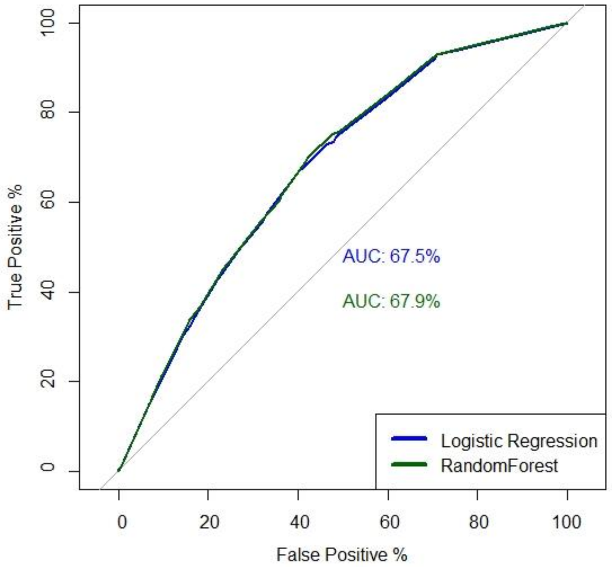


Model Building

After a sample ML model with demographic data, we built ML models with 2 versions of data:

1. Data with values replaced by corresponding Weight of Evidence value (WOE data)
2. Data with regular values (Regular data) – For Comparison

For WOE data

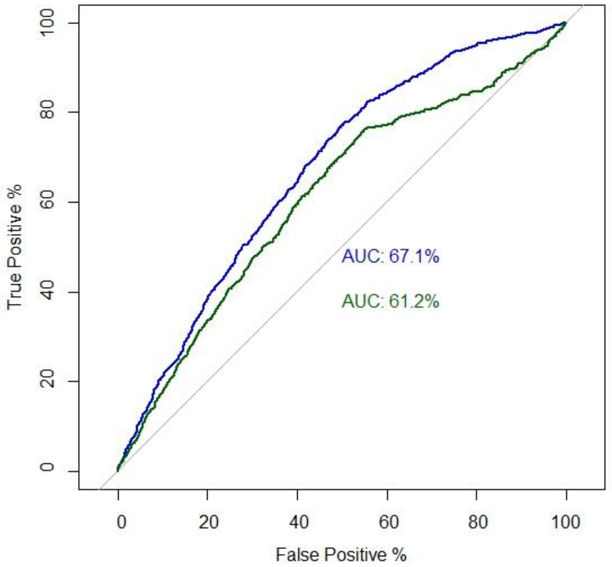


Accuracy

Log model:
55%
RF model:
63.4%

We developed a Logistic Regression model with an AUC score of 0.675
We developed a Random Forest model with an **AUC score** of **0.679**

For regular data



Accuracy

Log model:
47%
RF model:
62%

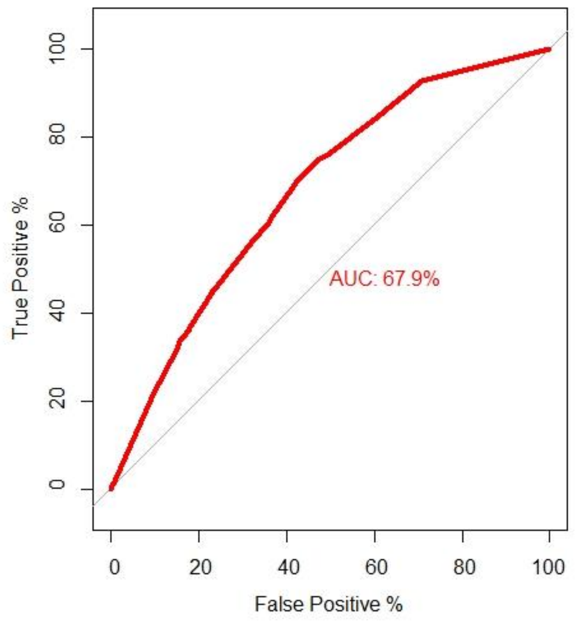
We developed a Logistic Regression model with an AUC score of 0.671
We developed a Random Forest model, with an AUC score of 0.612

Final Model

- We selected the Random Forest Model with WOE values as the Final Model.

- It has an **AUC score** of **0.679** and **Accuracy** of **63.4%**

- The **predictors** for Credit card default present in our **final model** were:
  - No. of times 30 DPD in the last 12 months
  - Average Credit Card utilization
  - No. of P/L Trades in the last 12 months
  - No. of Inquiries in the last 12 months



ROC curve of the final model

Application Scores – calculation & cut-off

Application Scores Calculation:

We calculated Application Scores based on the formula:

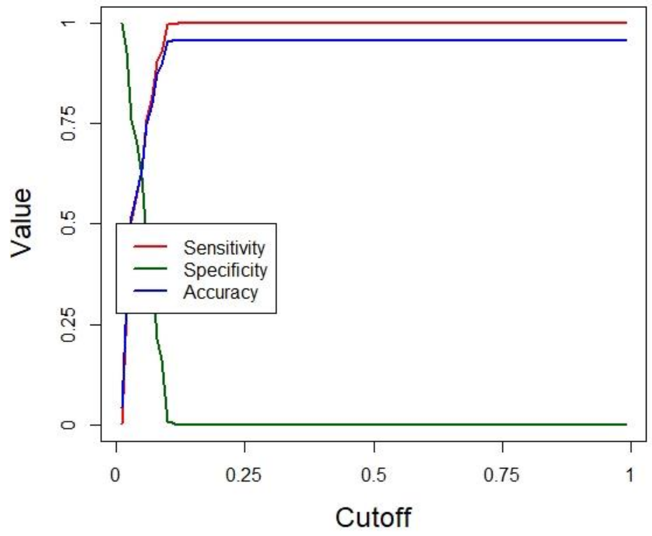
Score = (Offset - Factor)*(predicted probabilities)

where,

Offset = (Target Score Value - Factor)*log(Inverted Target Odds)

Factor = Points to double odds / log(2)

Application Scores Cut-off:



We considered the intersection between possible True Positives (Sensitivity), True Negatives (Specificity) and Accuracy of our final model as the cut-off %. Using this percentage (0.05%), we calculated the cut-off score as:
0.05% of Total of unique scores, i.e. 70 (approx.)

Looking at the scores across Good (not defaulting) and Bad (defaulting) Applicants, we concluded that:

➢Credit card applicants with an application **score below 70** can be **denied credit card**

Project Outcome

Financial Benefit of the Project:

To assess the financial benefit of this project we prepared a Gain-Lift table, which shows that:

➢After implementation of this project, the firm can detect 75% of Bad Applicants by targeting 50% of the Applicants.

➢Roughly, in money terms, the firm will be **saving** an average potential credit loss of **8.23 million USD**

Gain – Lift table

	bucket	total	totalresp	Cumresp	Gain	Cumlift
1	1	2060	183	183	21.08295	2.108295
2	2	2059	152	335	38.59447	1.929724
3	3	2059	122	457	52.64977	1.754992
4	4	2059	108	565	65.09217	1.627304
5	5	2059	93	658	75.80645	1.516129
6	6	2059	63	721	83.06452	1.384409
7	7	2059	74	795	91.58986	1.308427
8	8	2059	30	825	95.04608	1.188076
9	9	2059	20	845	97.35023	1.081669
10	10	2059	23	868	100.00000	1.000000