BFS Capstone Project: Mid Submission

Submitted on May 20, 2019 (Monday)

Group Members:

Ishan Savio Kerketta
Devanshi Kulshreshtha
Ayushi Gaur
Ajay Sharma

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.

Data Available

- Demographic Data
- Credit Bureau Data
- Data Dictionary (Metadata)

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.

Data Understanding

There are two 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.

APPROACH TO PROBLEM SOLVING

After loading the required packages in R and extracting the data, we will consider the following steps to solve the problem:

1. Demographic data >

Data cleaning
Data Preparation
Exploratory Data Analysis
Lift Analysis (Association Rule Mining)

2. Credit Bureau data >

Data Cleaning
Exploratory Data Analysis
Lift Analysis (Association Rule Mining)

3. Model Building: Demographic data >

Dummy variable creation
Splitting data into test and train
Running a Logistic Regression Model

4. Demographic + Credit Bureau data >

Merging demographic & credit bureau data (Inner Join)
Exploratory Data Analysis
Lift Analysis (Association Rule Mining)
WOE & Information Value Analysis
Model Building: Logistic Regression
Application Scorecard: Logistic Regression Model
Model Building: Random Forest
Application Scorecard: Random Forest Model
Financial Benefit Assessment

Note: We will pass the clause `na.strings = c("","NA")` while extracting the data so that all missing values are converted to NAs

1. Demographic Data

- Data Cleaning:
  - We found 3 duplicates in Application ID and removed them.
  - We removed all 1425 rows with NAs in the target variable (Performance.Tag)
  - We replaced the rows with NAs in Education column to 'Others'
  - We removed all other NAs in the dataset as their frequency is very small
  - We found invalid values like -3 and 0 in Age column. We also found unnatural values such as 15, 16, 17 since credit cards are not given to minors. We decided to remove them as the frequency is very small
  - We floored the Income variable at 4.5
  - We capped number of months in current company to 74
- Data Preparation:
  - We created bins 'Low Income', 'Lower Middle', 'Higher Middle' and 'High' for Income variable using the quartiles as reference
  - Created bins '6 months', 'less than 1 year', 'between 1 and 3 years', 'more than 3 years' for No. of months in current residence

    - Created bins 'less than 1 year', 'between 1 and 3 years', 'more than 3 years' for No. of months in current company
    - Created bins ‘young adult’, ‘adult’, ‘middle age’, ‘senior citizen’ for Age
- Exploratory Data Analysis:
    - We looked at a frequency distribution of all categorical variables using bar plots
    - We looked at frequency distribution and percentage distribution of all categorical variables in relation to the target variable Performance.Tag using cross tables and bar plots
- Lift Analysis (Association Rule Mining) using arules package in R:
    - High Income, 6 Months in Current residence and Middle age group paired with Performance tag showed the highest lift followed by High Income, 6 Months in Current residence and Rented residence paired with Performance tag. The third in the list was customers with salaried profession, living in rented residence, have been staying in current residence for 6 months and are middle aged paired with Performance Tag

2. Credit Bureau data

- Data Cleaning:
    - We found 3 duplicates in Application ID and removed them
    - We removed all rows with NAs in Performance.Tag
    - We also remove all the remaining rows which have NAs
    - We looked for outliers in continuous variables. We capped No. of trades opened at 99th percentile (i.e. 21)
    - We capped outstanding balance at 99th percentile
    - We capped Total No. of Trades at 99th percentile
    - We corrected the levels of 'Presence of open Home loan'
- Exploratory Data Analysis:
    - We created histograms for all continuous variables and bar plots for categorical variables
- Lift Analysis (Association Rule Mining)
    - Zero PL trades opened in last 6 months, Zero Inquiries in last 12 months excluding home and auto loans and presence of open home loan paired with Performance tag had the highest Lift
    - Zero PL trades opened in last 6 months, Zero Inquiries in last 6 months excluding home, Zero Inquiries in last 12 months excluding home and auto loans and presence of open home loan paired with Performance tag had the second highest Lift
    - Zero times 90 DPD or worse in last 12 months, Zero PL trades opened in last 6 months, Zero Inquiries in last 12 months excluding home & auto loans and Presence of open home loan paired with Performance tag had the third highest lift

3. Model Building with Demographic data only

- We will create dummy variables for the categorical independent variables in the dataset
- We will perform a 70-30 split of the data into train and test
- We will run a logistic regression model since the problem at hand is binary classification. We will omit Gender variable as fair lending excludes Gender
    - Firstly, we will use Stepwise selection using stepAIC function
    - Next, we will use backward selection removing variables based on multicollinearity (by checking VIF) and p-values
    - We will use the model to make predictions, use an optimum cutoff for binary classification and check Accuracy, Specificity and Sensitivity

4. Demographic data and Credit Bureau data

- We merged demographic data and credit bureau data by performing an inner join on Application ID
- Exploratory Data Analysis
    - We prepared box plots to see the relation between demographic variables and variables of credit bureau

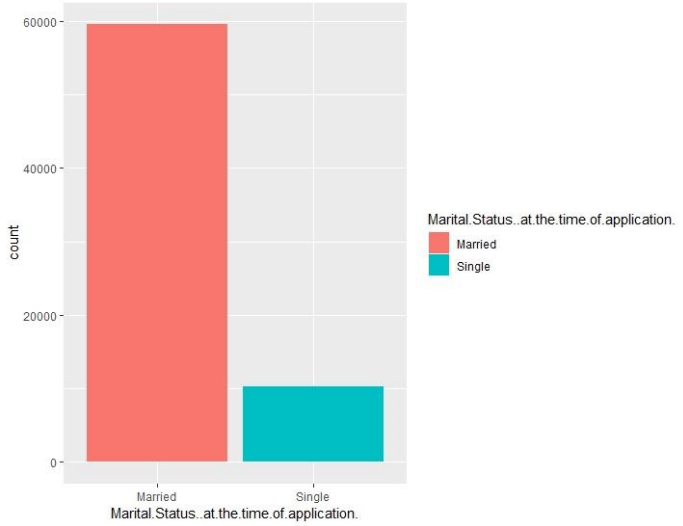
- Lift Analysis (Association Rule Mining)
    - More than 3 years in current residence, Zero PL trades opened in last 12 months and Zero Inquiries in last 6 months excluding home & auto loans paired with Performance Tag shows highest lift
    - More than 3 years in current residence, Zero PL trades opened 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 shows second highest lift
    - More than 3 years in current residence, zero times 60 DPD or worse 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 shows third highest lift
- Weight of Evidence and Information Value Analysis
    - We split the data into train and test. We used Information package in R to create WOE and IV tables. We used knitr package to look at top Information values. We also plotted WOE patterns for top 4 variables
    - Top variable as per IV are:
        - No. of Inquiries in last 12 months excluding home & auto loans
        - No. of PL trades opened in last 12 months
        - Avgas CC Utilization in last 12 months
        - No. of trades opened in last 12 months
        - Total No. of Trades
        - No. of times 30 DPD or worse in last 6 months
- Model Building: Logistic Regression
    - We will create dummy variables for the categorical independent variables in the dataset
    - We will perform a 70-30 split of the data into train and test
    - We will run a logistic regression model since the problem at hand is binary classification.
        - We will consider only those variables from demographic data that remained in the final model of logistic regression using demographic data only.
        - There are some variables that are clearly correlated with each other such as No. of times 30 DPD or worse in last 6 months, No. of times 30 DPD or worse in last 12 months. In such cases we will select the latter (i.e. 12 months) as it includes the data from former (i.e. 6 months)
- Application Scorecard: Logistic Model
    - We will build an application scorecard with the good to bad odds of 10 to 1 at a score of 400 doubling every 20 points
    - We will calculate log odds using predict function in R
    - We will calculate factor as points to double odds divided by the natural log of 2
    - We will calculate offset as (Target score value – factor) * log odds
    - We will calculate the score as (offset – factor) * log odds
    - We will use the rounded value of the score obtained as the final score
    - We identify the cut-off score beyond which we would not grant credit cards to applicants
- Model Building: Random Forest
    - Using the variables that remain in the final model of logistic regression using demographic and credit bureau data, we will create a Random Forest model and make predictions
    - We will also create optimal cutoff for binary classification and look at Accuracy, Specificity and Sensitivity
    - We will choose the Random Forest as the final model if the Accuracy, Specificity and Sensitivity values are higher than the Logistic Model
- Application Scorecard: Random Forest Model

- If the Random Forest Model turns out to be a better model, we will create an Application Scorecard in the same manner as described for the logistic model

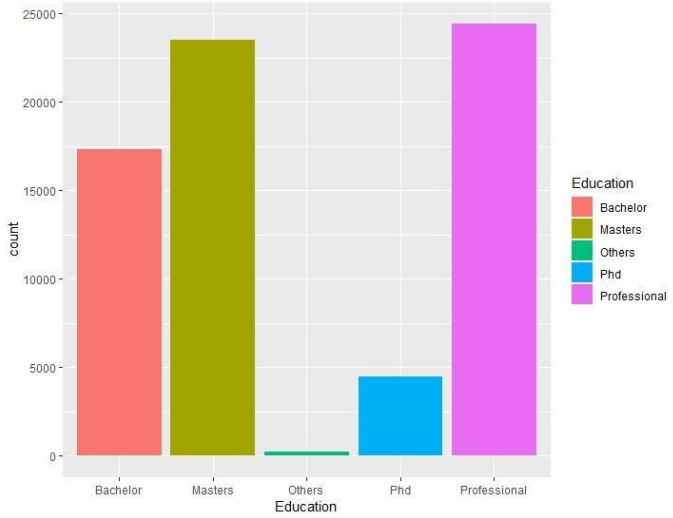
- Financial benefit Assessment
  - We will elaborate the effectiveness of our model
  - We will prepare a Gain chart and exhibit credit loss and how it can be reduced by targeting optimal percentage of applicants
  - We will explain the insights that we gather through the project

EXPLORATORY DATA ANALYSIS:

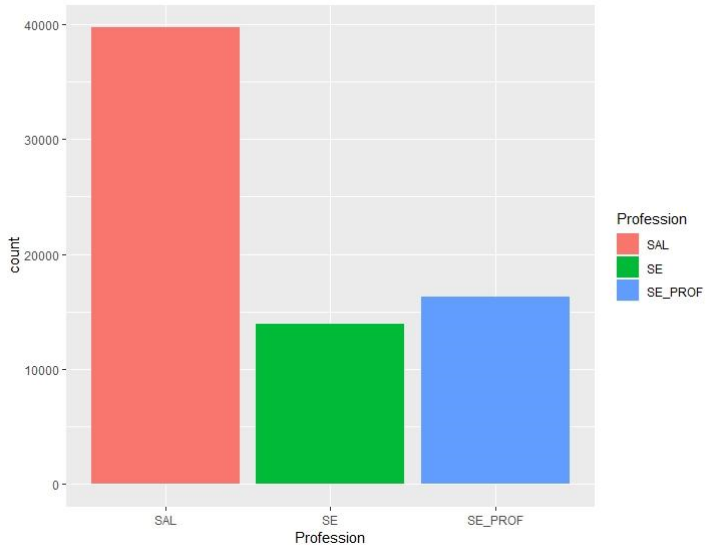
Univariate Analysis on Demographic data:



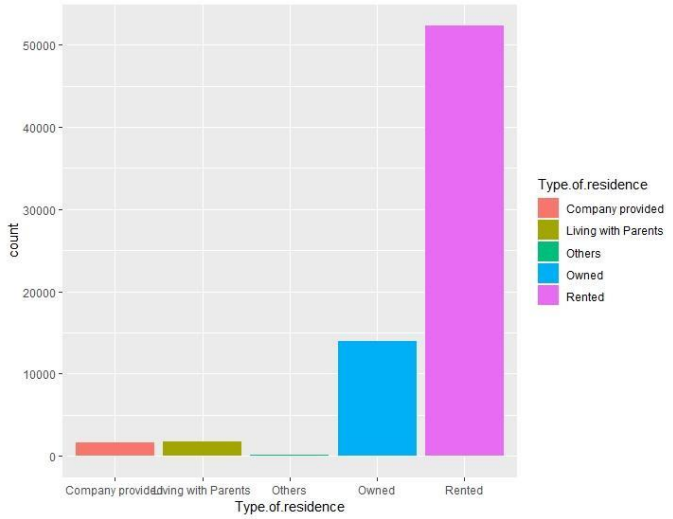
Most of the applicants are Married



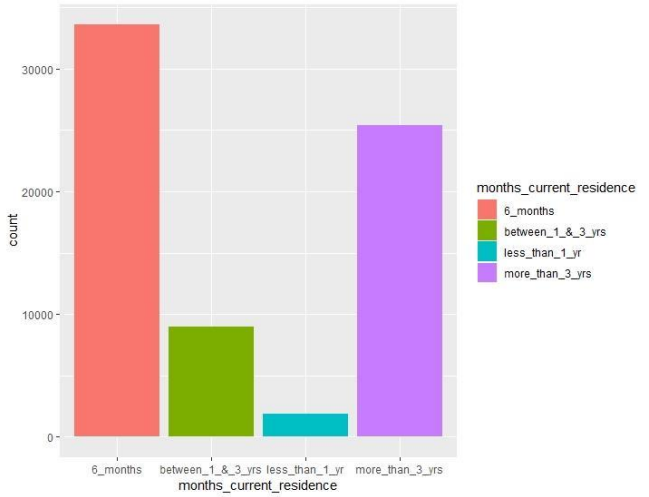
Most of the applicants have Professional Education, followed by those with a Master's Degree



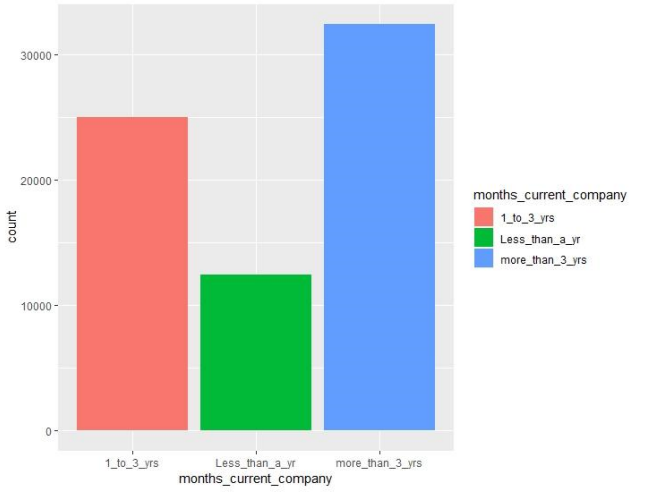
Most applicants have a salaried profession



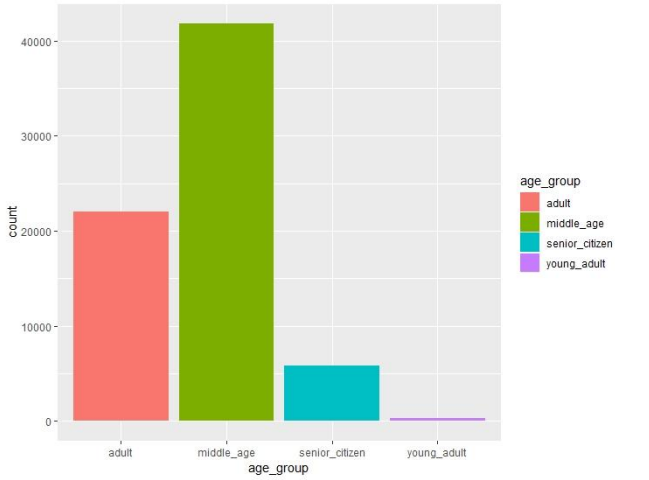
Most applicants lie in rented houses



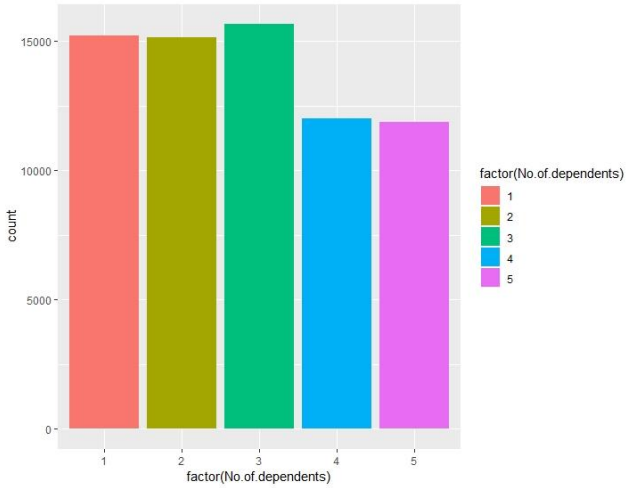
Most applicants have been living in the current residence since the last 6 months



Most applicants have been in the current company for more than 3 years



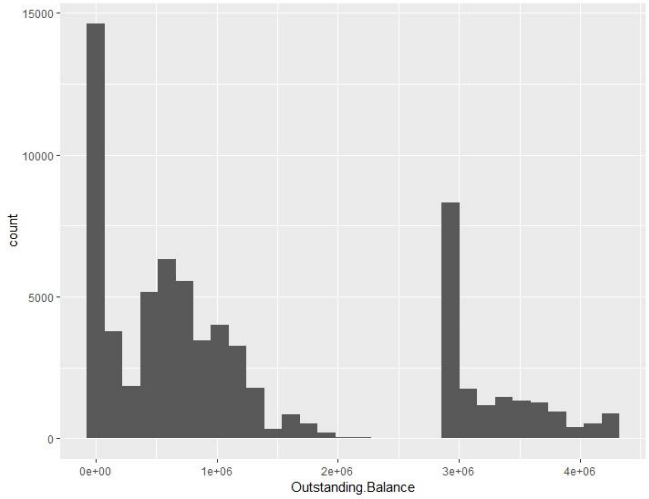
Most applicants are middle aged (i.e. between 40 & 60)



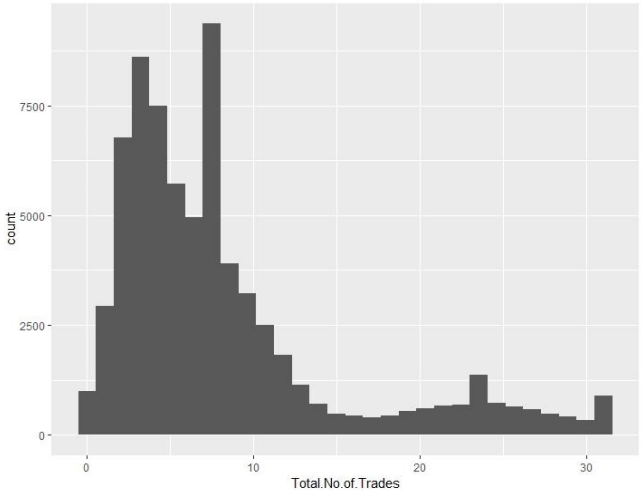
Most applicants have between 1 and 3 dependents

Note: Cross-tables prepared between categorical variables in demographic data do not show any relevant insights

Important histograms on variable in Credit Bureau data:



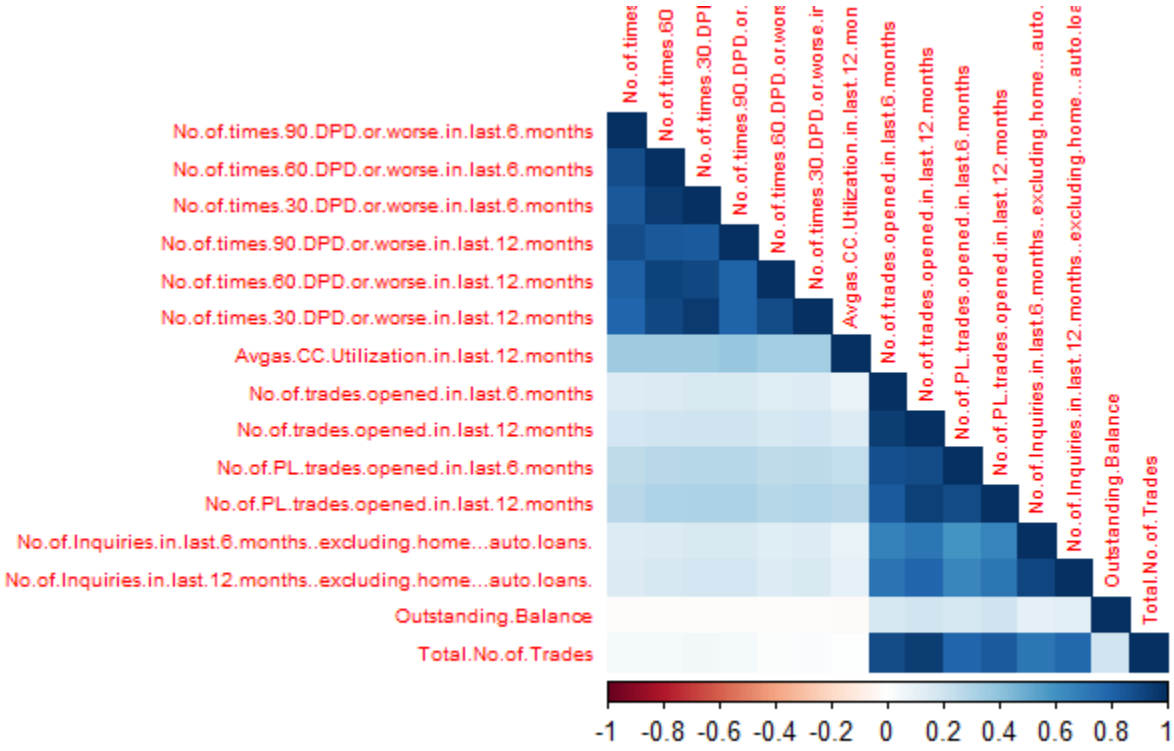
Frequency of Outstanding Balance is highest at low amounts and keeps decreasing. Again, there is a peak at around 3,000,000 and decreases after that.



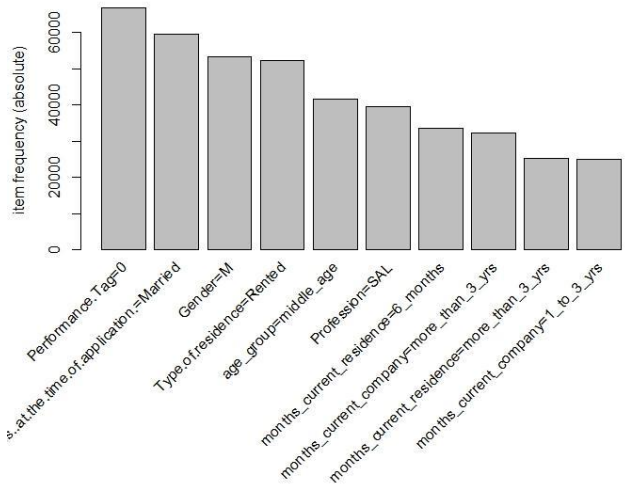
Frequency of total number of trades is highest 0 and 10

Note: Histograms of all the remaining variables show that most of the variables have a value 0 and the frequency keeps decreasing as the value increases

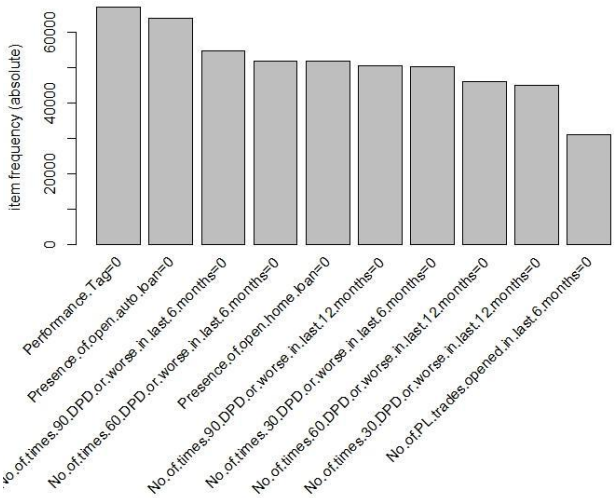
Correlation Between Variables in the Credit Bureau Data



Item Frequency Plots

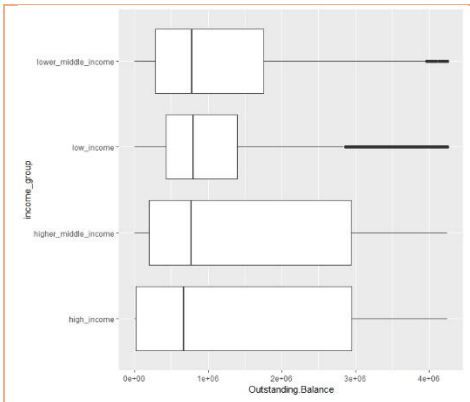


Item Frequency Plot for Demographic data – Top 10

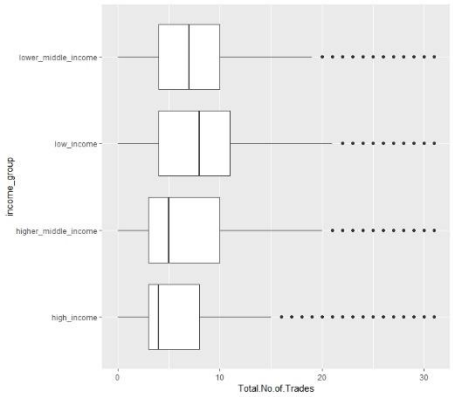


Item Frequency Plot for Credit Bureau data – Top 10

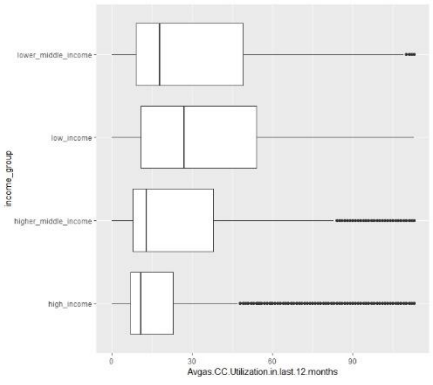
Box Plots of categorical variables in Demographic data VS continuous variables in Credit Bureau data:



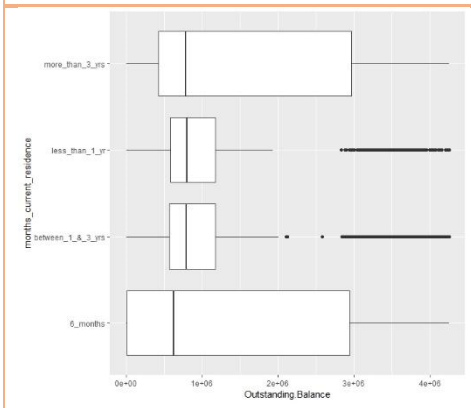
Applicants of high-income group **have lesser median outstanding balance, but the balance is widespread.**



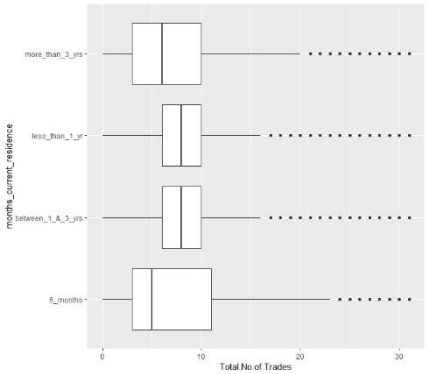
Applicants of low-income group **have the highest median number of trades.**



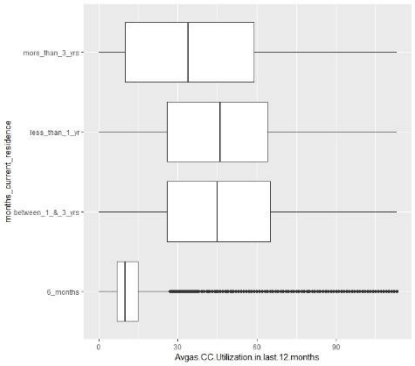
Applicants of low-income group **use credit cards the most and those of high-income use credit cards the least.**



Although applicants living in current residence for 6 months **have a widespread distribution of outstanding balance, the median is lesser.**

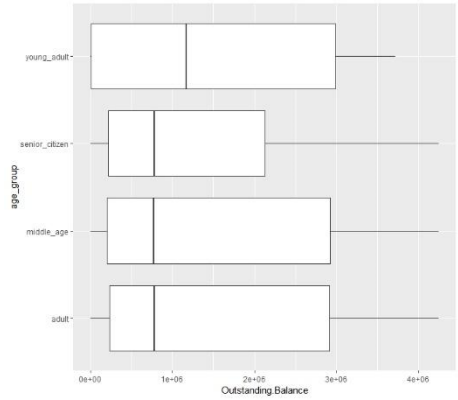


Applicants **living in current residence for 6 months** have lesser number of trades compared to others

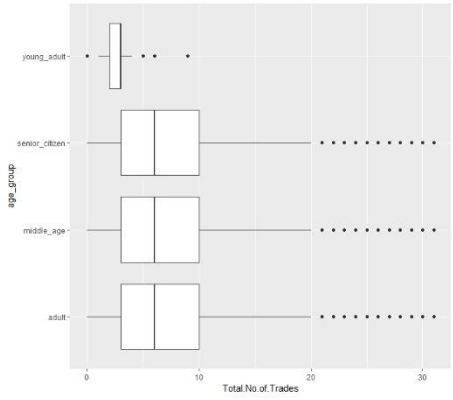


Most applicants **living in current residence for 6 months** use credit cards the least. However, there are some cases where they use more.

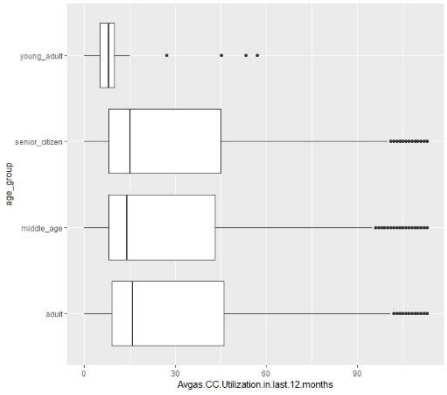
Age group



Young adults (less than age 25) **tend to have more outstanding balance on their credit cards**



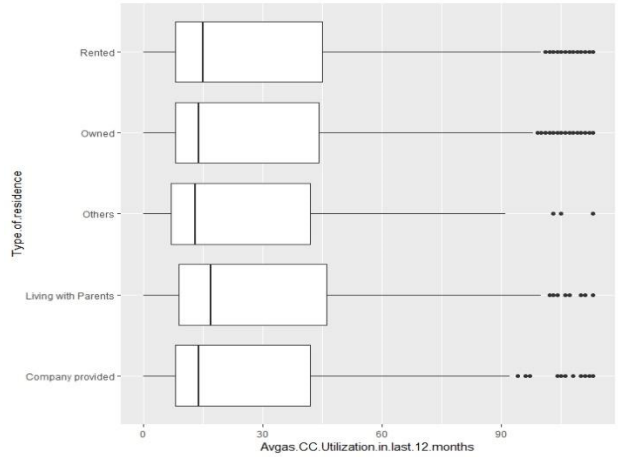
Young adults **don't induge in trade much**



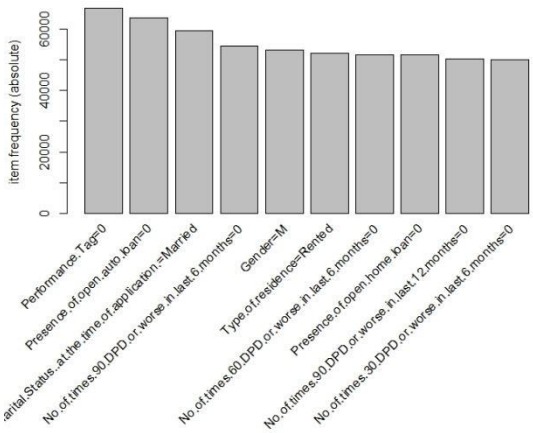
Young adults **don't use their credit cards much**

Type of residence

Overall Item Frequency



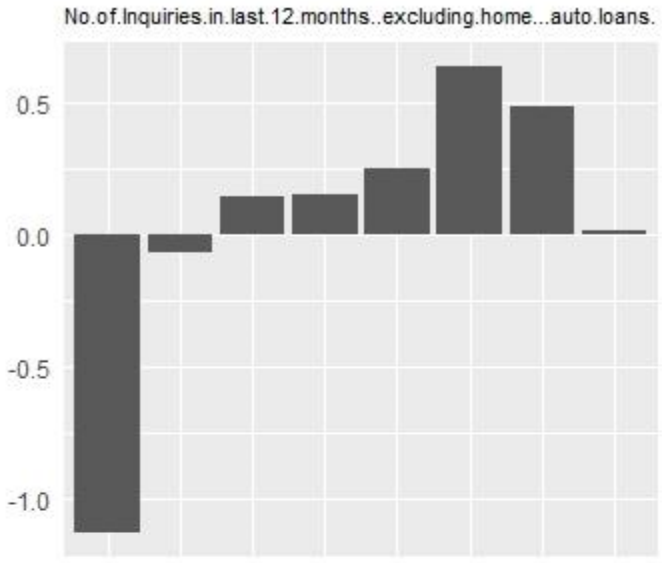
Applicants living with parents use credit cards more

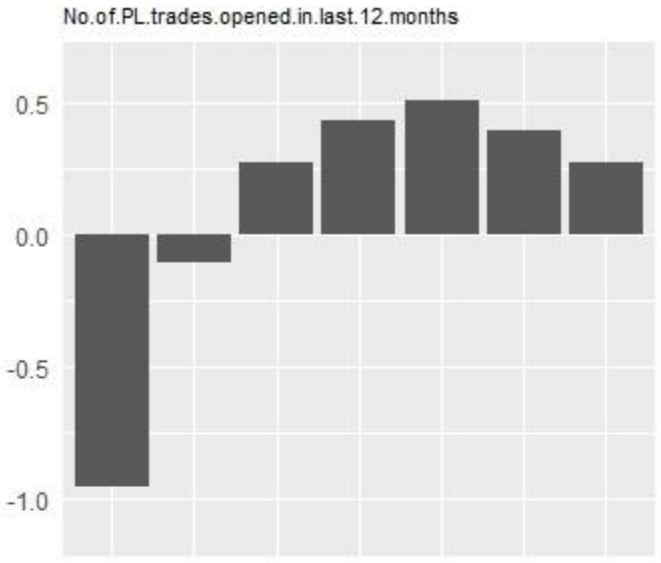


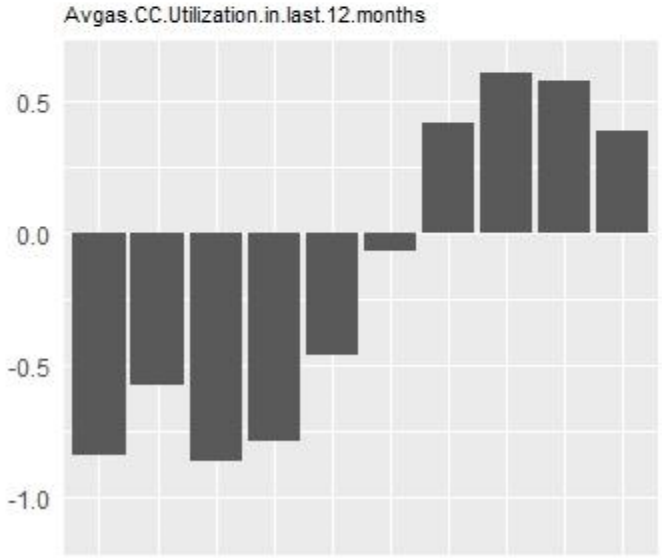
Item Frequency plot of overall variables – Top 10

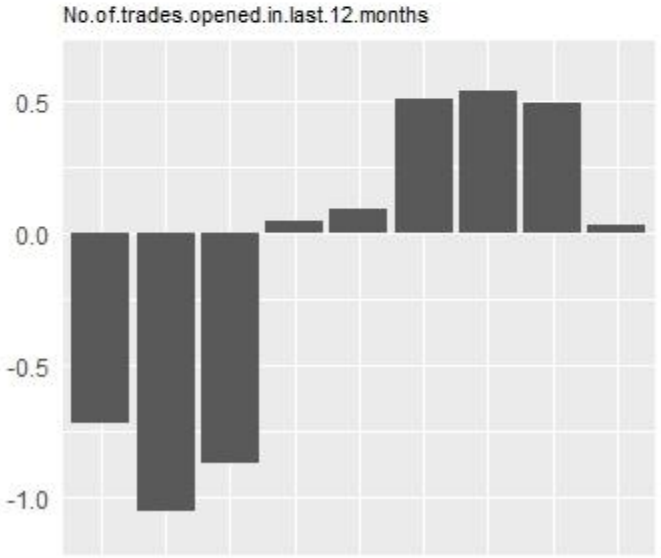
Weight of Evidence & Information Value Analysis

The following plots show weights of evidence of top four predictors of Credit card default as per Information Value:









Information values of Top variables:

Variable	Information Value	Penalty	Adjusted IV
No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.	0.3221598	0.0547036	0.2674561
Avgas.CC.Utilization.in.last.12.months	0.3143931	0.0477531	0.2666400
No.of.PL.trades.opened.in.last.12.months	0.3232161	0.0634933	0.2597228
No.of.trades.opened.in.last.12.months	0.3197047	0.0621917	0.2575129
No.of.times.30.DPD.or.worse.in.last.6.months	0.2438763	0.0167832	0.2270931