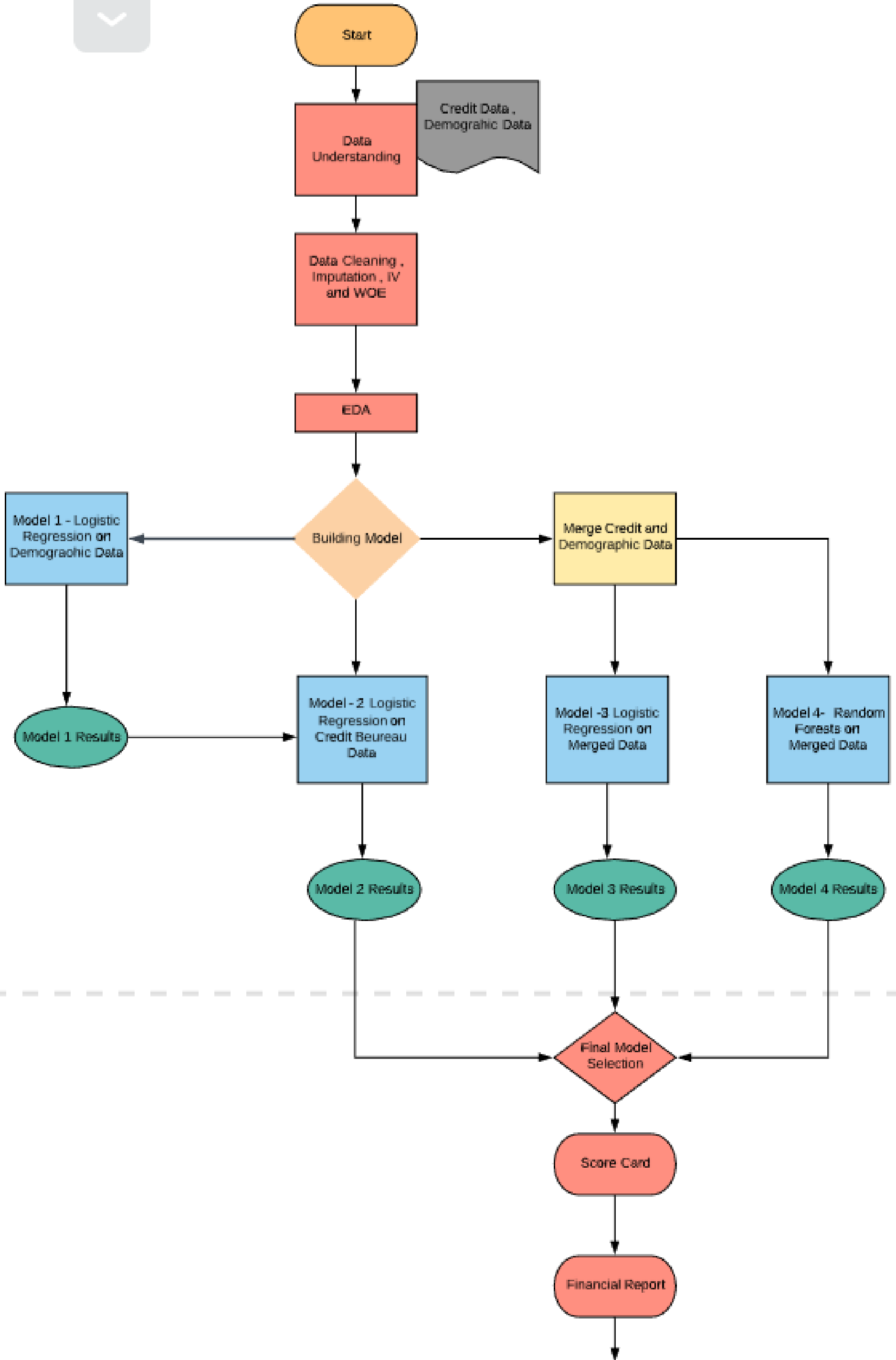
CredX Acquisition Analytics and Financial Assessment Report

- **Group Members**
  - Maya Kavuri
  - Puja Chakraborty
  - Pushpendra Sharma
  - Pavankumar Harathi

CredX Capstone Project

- **Business Understanding**
- CredX is a leading credit card provider that gets thousands of credit card applications 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
- **Business Objective**
- In this project, our task is to help CredX identify the right customers using predictive models. Using past data of the bank’s applicants, we need to determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of our project
- To identify variables which are strong indicators of default and potentially use the insights in approval /rejection decision making.


Start
Credit Data , Demograhic Data
Data Understanding
Data Cleaning , Imputation , IV and WOE
EDA
Model 1 - Logistic Regression on Demograohic Data
Building Model
Merge Credit and Demographic Data
Model 1 Results
Model - 2 Logistic Regression on Credit Beureau Data
Model -3 Logistic Regression on Merged Data
Model 4- Random Forests on Merged Data
Model 2 Results
Model 3 Results
Model 4 Results
Final Model Selection
Score Card
Financial Report

Approach and Process Flow

- Approach
- EDA : Univariate, Bivariate , WOE(weight of Evidence AND IV(Information Value)) Analysis.
- SMOTE: Imbalanced Classification Problem
- MODEL : Specificity , Sensitivity ad Accuracy Metrics
- Building Score Card
- Financial Report

DATA UNDERSTANDING

-

- There are two data sets with 71292 rows(records) in this project and named as **demographic** and **credit bureau** data.

- Both files contains “Performance Tag”(Target variable)which represents the applicants have gone
- 90 days Past due or Worse in past 12 months(i.e defaulted)after getting the credit card.
  - Total 1425 records (around 2%)are having NA “Performance Tag” column(They are considered as rejected applications)
  - Total 66922 records(around 94% are having “0” (Non-Defaulted) in “Performance Tag” column.
  - Total 2948 records(around 94% are having “1” (Defaulted) in “Performance Tag” column.

- **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

Data preparation And Assumptions

Duplicate ID treatment before merging the 2 dataframes.(3 Duplicates we removed

Outliers are present in the following variables :Age, No. of Months in current company ,No .of Times 30/60/90 PDP or worse in last 6/12 months ,No .of trades open in last 6/12 months ,Total No. of trades , Total inquiry in last 6/12 months.

Assumptions:Age below 18 years are minors hence removed them,and also applications with age negative were removed.

Income<=0 removal if less than 1% application present(106 applications)

NA treatment:~1.9%in the dependent variable were removed

NA treatment~1% in columns” Avg Credit Card Utilization” were removed

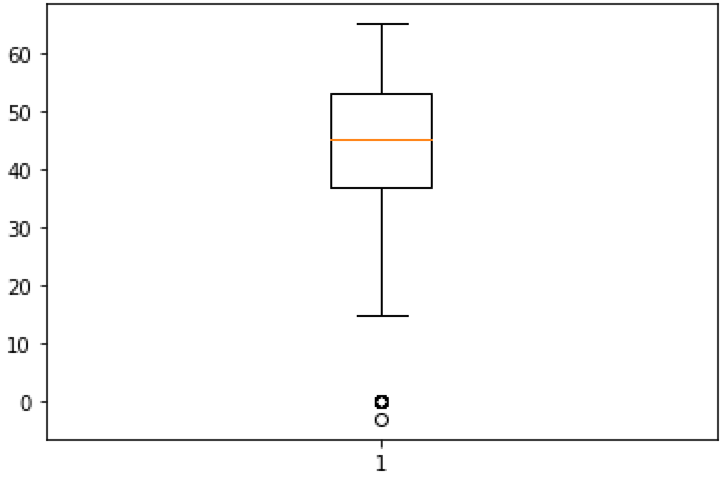
Dropped the rows where Gender,Marital status,Profession,Education and Type of residence

Dropped all the NA’s fromaverage cc utilization columns,No.of.trades.opened.in.last six months.,Presence of open home loans,Outstanding Balanceand Performance Tag column

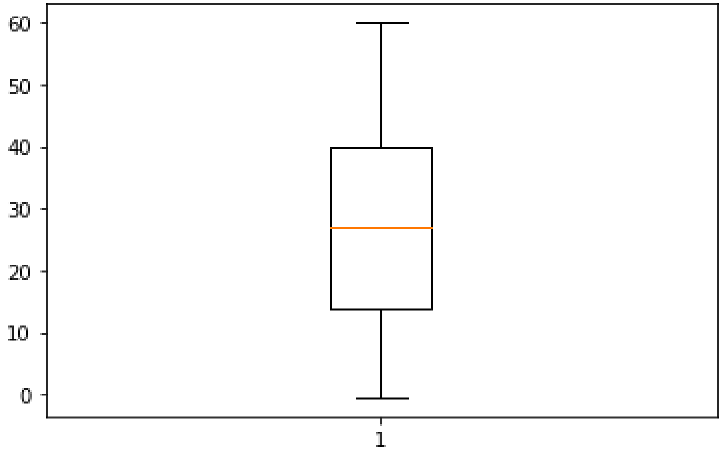
Used Weight of Evidence and Information valuation for treating missing values and imputing the values while building the Logistic Regression Model

Outlier treatment

Age



Income

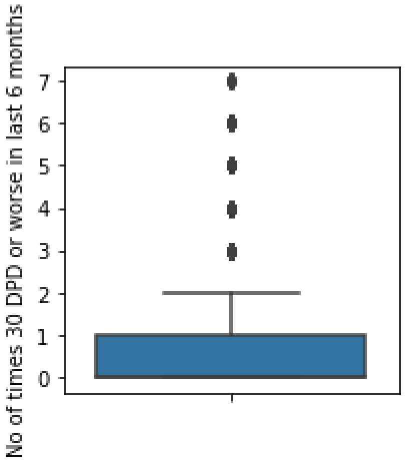


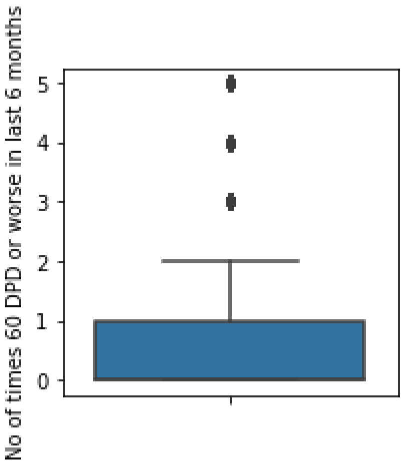
There are few outliers in Age and Income, having zero and negative values

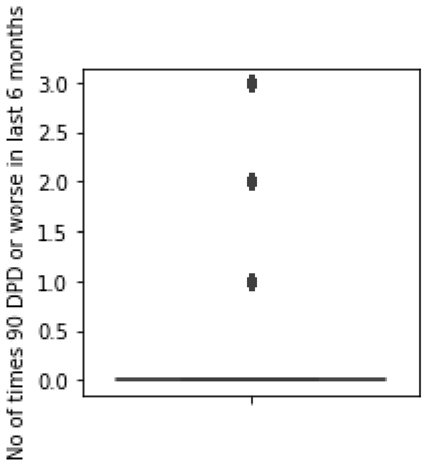
Other columns do not contain outliers,only certain valid value lies over the 3rd quartile

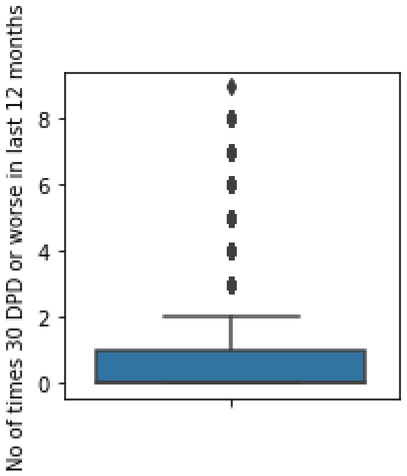
Outlier Detection using the Box Plots

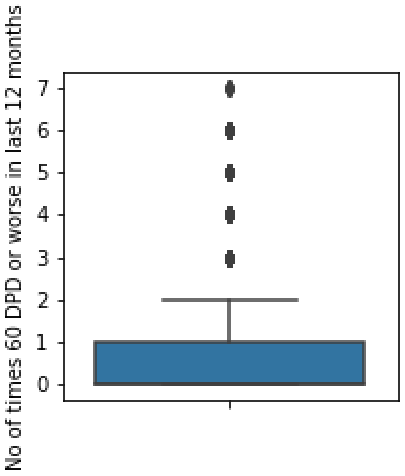
Below variables does not contain the outliers,only certain valid values lie outside the 3rd quartile

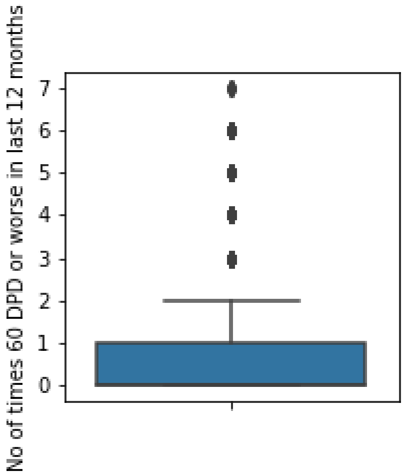








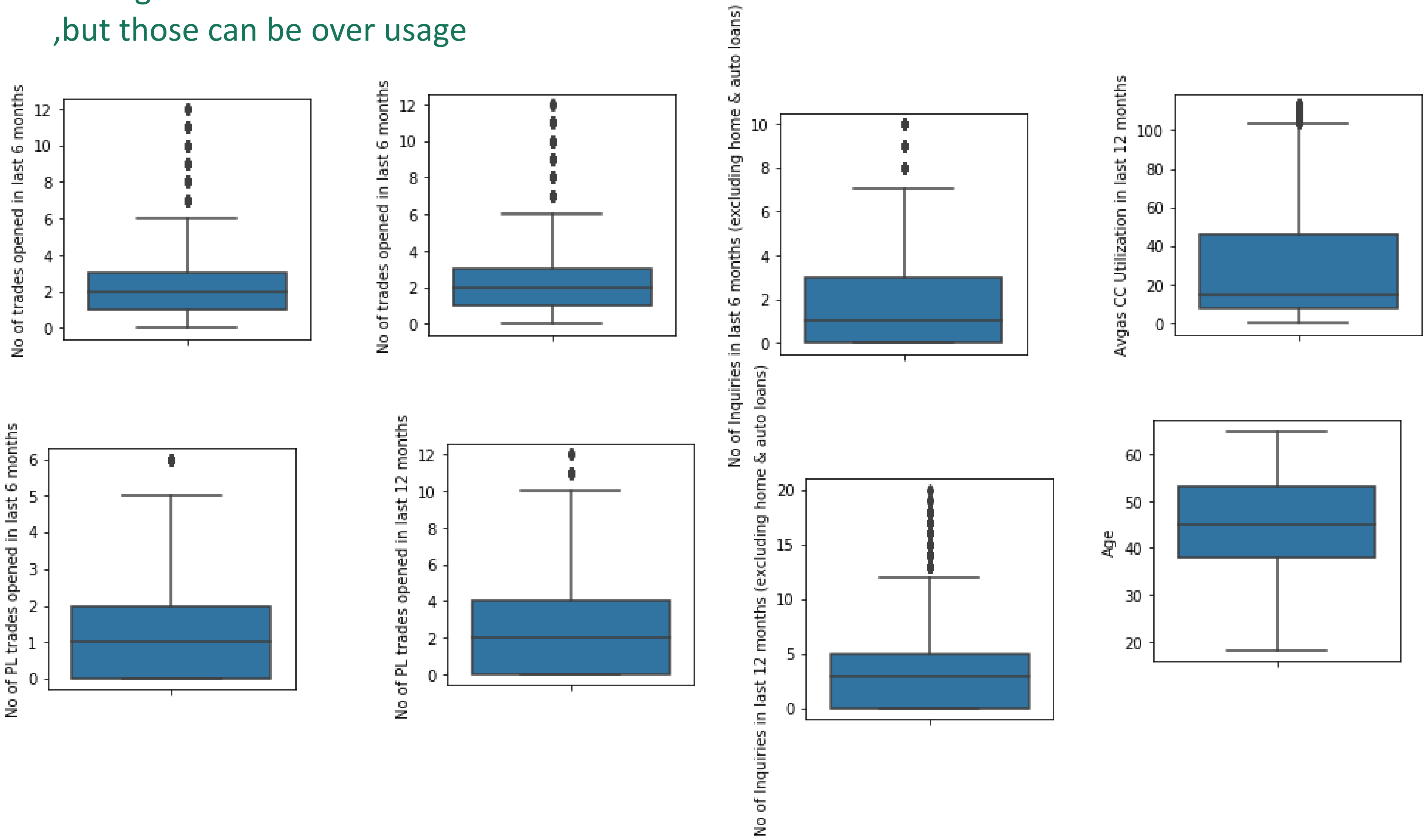




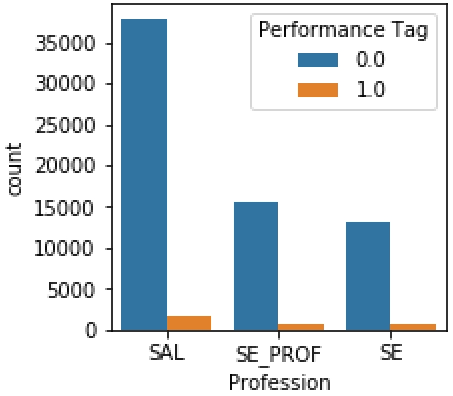
Outlier Detection using the Box Plots

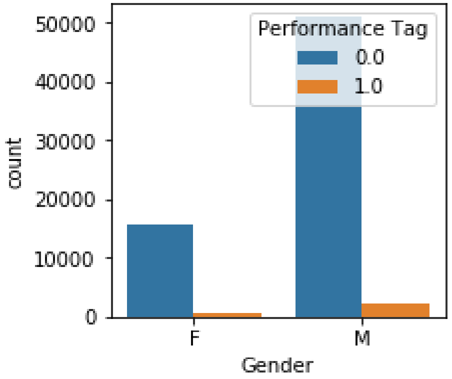
Few outlier in Age zero and negative values

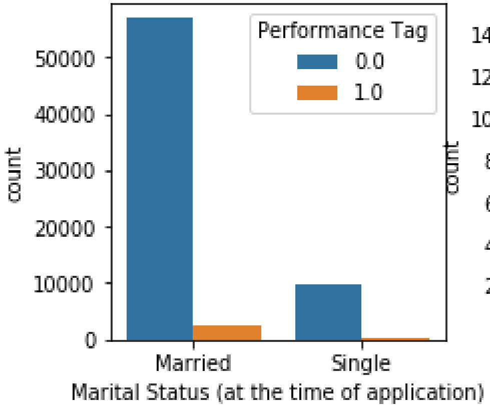
Average Credit card utilization contains values more than 100 ,but those can be over usage

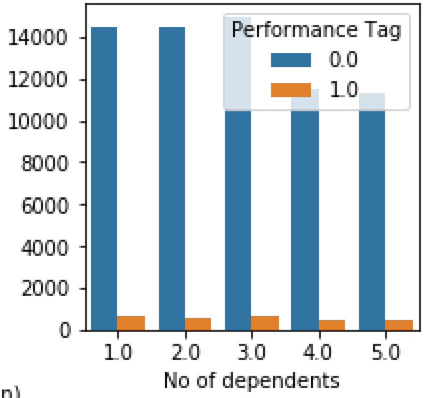


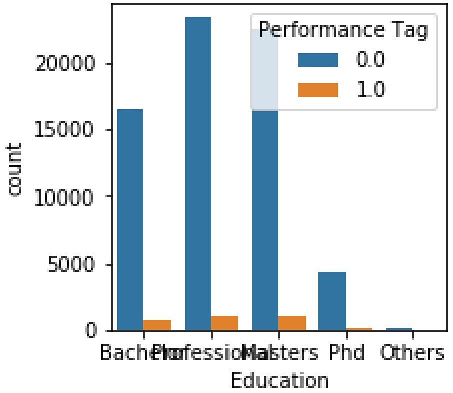
Univariate Analysis

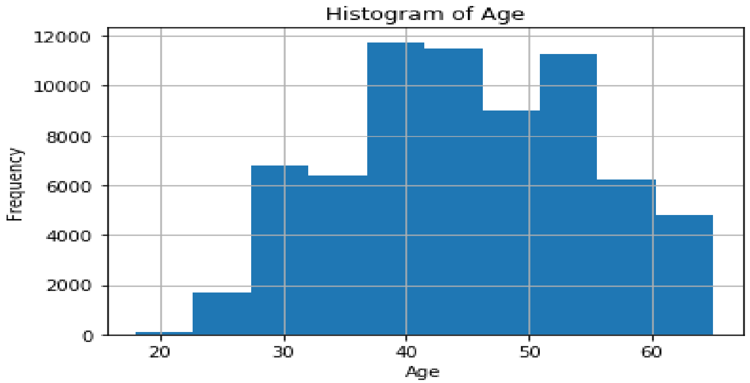












1.No of dependents have equal distribution of defaulter,so may not be a significant contributor

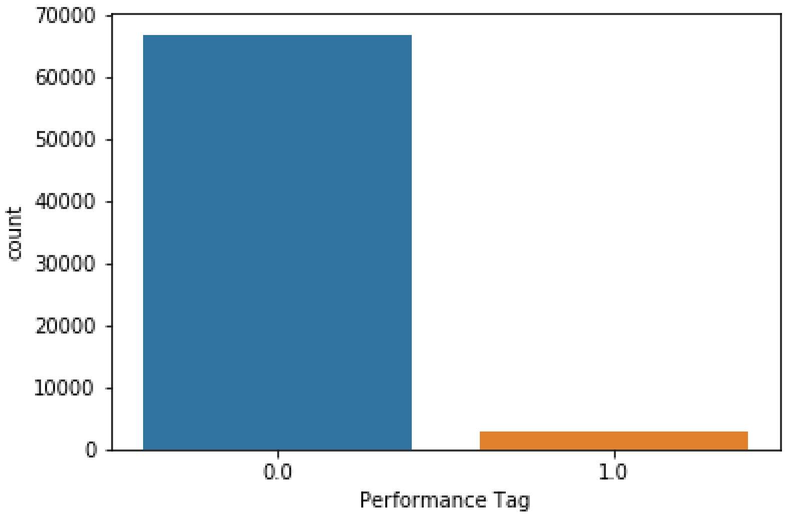
2.We can see defaulter in age between 30-60 soo its good contributor

3.Marital status and Gender also effects the default’s behaviour

4.Income is good contributor.Education ,type of residence seems to be good contributor.Absence of auto loan seems to have negative effect on the default behavio

Exploratory Data Analysis

Performance Tag



Percentage of default is 96%

Percentage of non default 4%

Our classes are imbalanced, and the ratio of no default to default instances is 96%.

We are using **SMOTE**: Synthetic Minority Over-sampling Technique to balance the dataset.

Summary of Exploratory Data Analysis

Following seems to be the contributing factors(after basic logic model)
Marital status
Age
Income
No.of times30 DPDor worse.in.last 6 months
No.of times 90 DPDor worse .in last 12 months
No.of PL trades opened in last 12 months
No.of inquiries in last 6 months excluding the home auto loans
No.of inquiries in last 12 months excluding the home auto loans
Presence of open home loan
Total No.of Trades
No.of dependents

Model and Final Model Selection

Logistic Regression(Demographic Dataset)	
Metrics	Values
Overall Accuracy	56%
Senstivity	55%
Specificity	57%
ROC	56%

Logistic Regression(Combined Dataset)	
Metrics	Values
Overall Accuracy	63%
Senstivity	69%
Specificity	56%
ROC	63%

Random Forest(Original Dataset)	
Metrics	Values
Overall Accuracy	62%
Senstivity	64%
Specificity	59%
ROC	62%

Above are the results/Accuracy /Sensitivity/Specificity and ROC obtained
Based on that we choose final model As Logistic Regression as its metrics are good
.WE used SMOTE technique to balance the data as its highly imbalanced data
We used Weight of evidence (as it treats,missing values,outliers,dummy variable creation etc)

Application Scorecard

- APPLICATION SCORE CARD • The application score for each applicant calculated using the logistic regression model,
- Score ranges from 297.6 to 365.3. Score increases by 20 points for doubling odds for good customers. Application score for odds of 10 to 1 is 400. We are decided
- final Cut-off score 339. Higher the scores indicate lesser risk for defaulting.
- **Scorecard for Rejected Applicants:**
- Ran the model on the Rejected application data frame(Woe imputed) none of them got score more than cut of
- Score ranged from 304 to 344 for the set of 1425(rejected applications)
- For the scores 355:100 percent applicants who defaulted were captured
- For the score 336:99% applicants who defaulted were captured

Financial Benefit Analysis

Final inferences made from the analysis are as below –

Objectives :

1. From P&L Perspective, the objective is to minimize “Net Credit Loss”.

2. Scorecard is used for determining desired trade-off between risk level and approval rate.

- With suggested optimal cut off score of 339, avg. 71% of applicants would be approved. Hence 29% applicants would be rejected.
- Assumption regarding credit loss- Outstanding balance of a defaulter is considered as credit loss for the specific user.
- Potential credit loss avoided by applying the model/scorecard implementation - 52%