CredX Acquisition Analytics

BUSINESS PROBLEM

Credx has experience an increase in the credit loss over the past few years. So, it wants to mitigate this risk by acquiring the right set of customers. Credx would need to do the following:

- Understand the factors affecting the credit the Credit Risk
- Create strategies to mitigate the acquisition risk
- Assess the financial benefit of the project

APPROACH TO THE PROBLEM

We will follow the CRISP-DM Framework:

- Business Understanding
- Data Understanding
- Data Preparation
- Model Building
- Model Evaluation
- Model Deployment

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'.

Also, the customers missing the payment due dates but ultimately closing the loans with the interest rate and penalty may be considered as the right customers. In fact, such customers are more profitable. But this also increases the risk of getting default. So, these must as well be considered while building the models and acquiring the customers.

DATA UNDERSTANDING

Data sets Available:

- ***Demographic/Application Data:*** Contains the data provided by the customers while applying for the credit card.
- ***Credit Bureau:*** Contains the data 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 Quality Check:

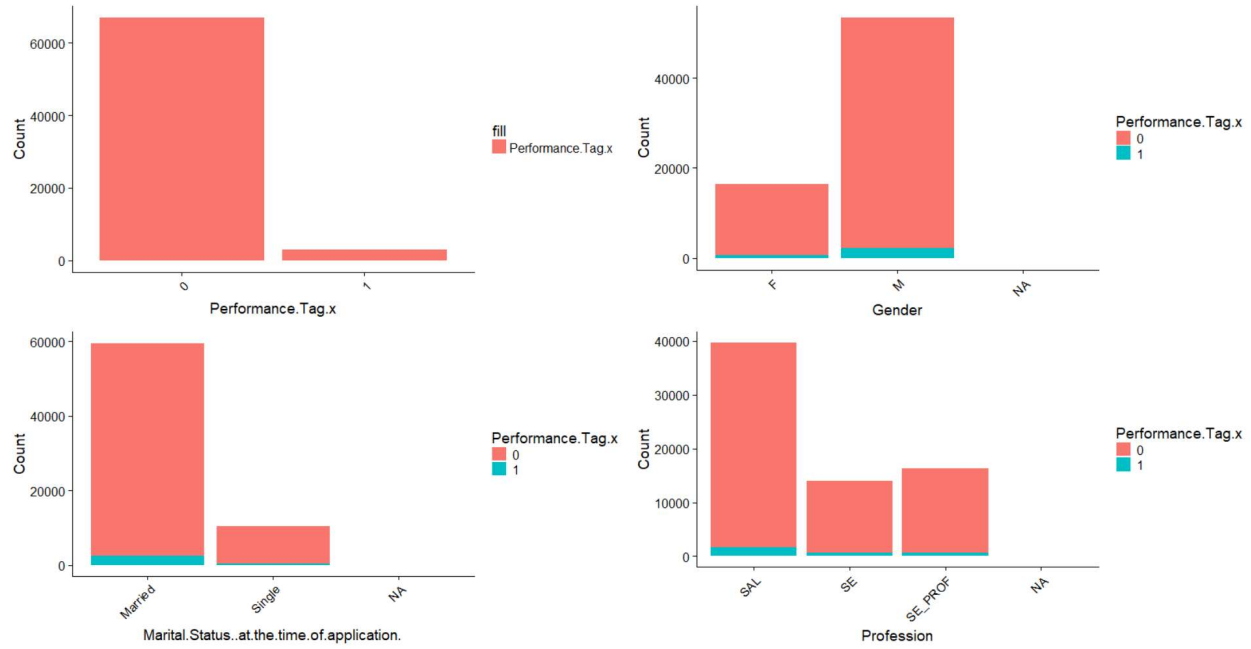
- ***Demographic Data:***
    - 71295 records having 12 attributes
    - Attributes with blank/NA values:
        - Gender - 2
        - Marital Status – 6
        - Type of Residence - 8
        - No of dependents - 3
        - Education - 119
        - Profession - 14
        - Performance Tags - 1425
    - Other data inconsistencies:
        - Application – 3 Duplicate records
        - Age – Invalid data (Age <=0)
        - Income – Invalid Income (Income <= 0)
        - No. of month in current company outlier detection – replaced all values greater than 74 to 74 as it covers 99% of the population

- ***Credit Bureau Data:***
    - 71295 records having 19 attributes
    - Attributes with blank/NA values:
        - Avg CC utilization in last 12 months – 1058
        - No. of trades opened in last 6 months – 1
        - Presence of open home loan – 272
        - Outstanding Balance - 272
        - Performance Tags - 1425
    - Other data inconsistencies:
        - Application – 3 Duplicate records
        - Total No. of trades in the current company - replaced all values greater than 31 to 31 as it covers 99% of the population

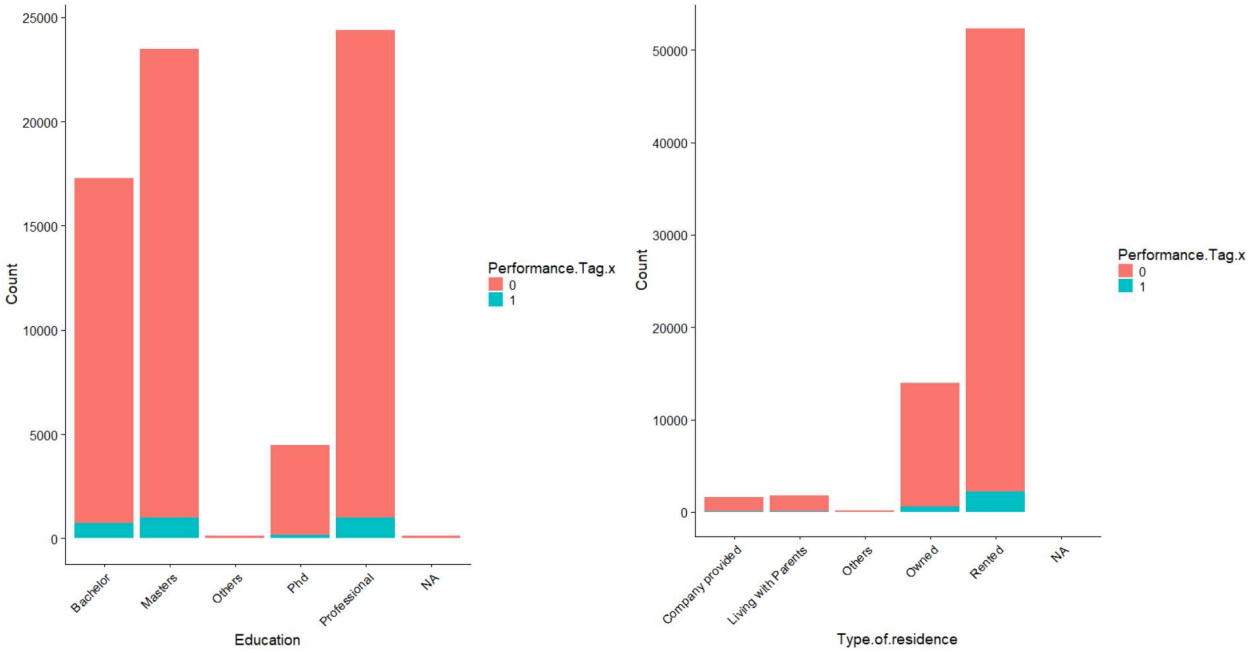
Assumption based on the Data quality check and data understanding:

- The records with Performance tag blank are considered as the Rejected Applicant and we would ignore it from our EDA and model building process. We would subset only the approved applicant for EDA and model building.
- NA values in Avg CC utilization in last 12 months are present the data imply that the customers does not having no other credit card. So, we will replace the Avg. Credit card utilization value from NA to 0. Also, we will create a flag as ***has_credit_card*** with value 1 for such customers.

EDA AND DATA PREPARATION



All the variables with missing values were imputed with their mode values.



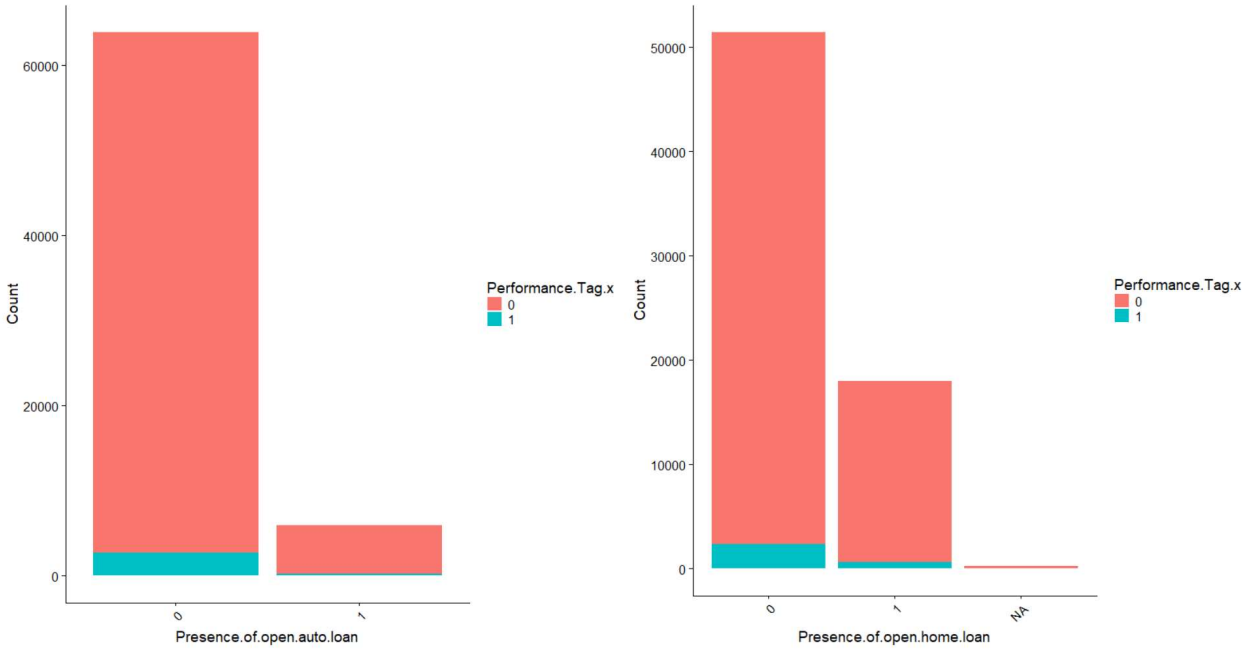
Education	N	Percent	WOE	IV
<NA>	118	0.001689	0.004186	2.97E-08
Bachelor	17304	0.247639	0.016619	6.90E-05

Masters	23481	0.336038	0.007375	8.73E-05
Others	119	0.001703	0.492048	6.06E-04
Phd	4465	0.063899	-0.01922	6.29E-04
Professional	24389	0.349033	-0.01872	7.51E-04

NA's in Education is imputed by Masters as its WOE values is close to Masters

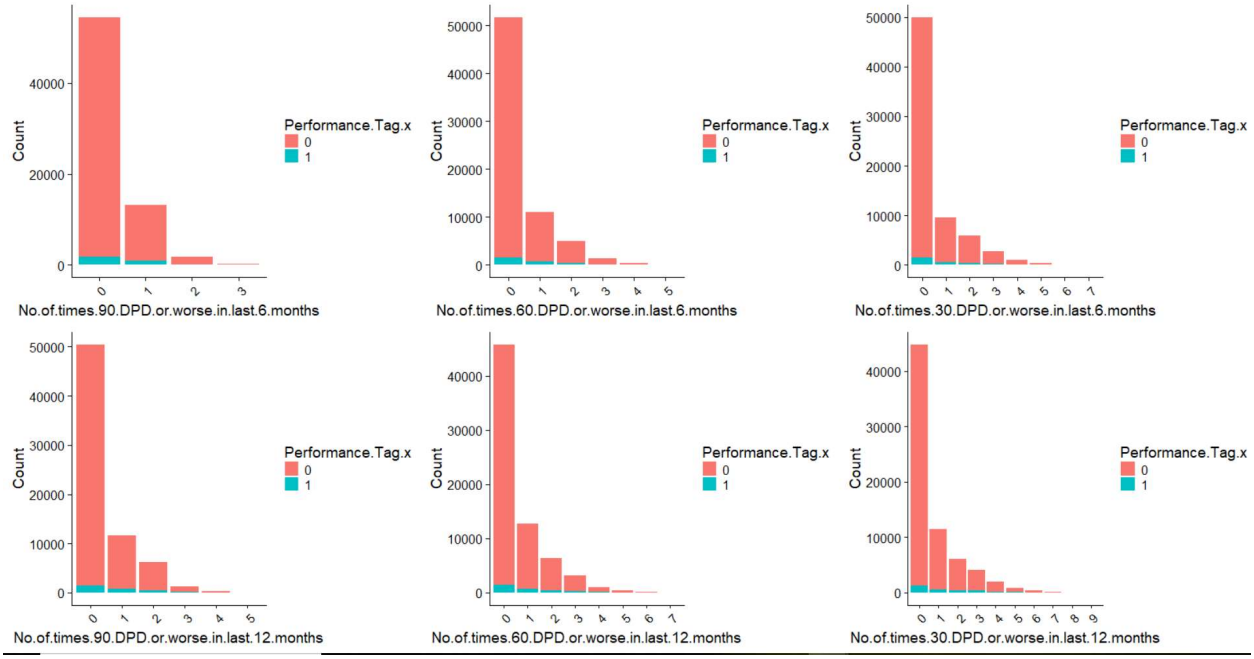
Type.of.residence	N	Percent	WOE	IV
<NA>	8	0.000114	0.00E+00	0
Company provided	1604	0.022955	7.89E-02	0.000148
Living with Parents	1779	0.025459	6.64E-02	0.000264
Others	198	0.002834	-5.31E-01	0.000895
Owned	14003	0.200398	3.57E-03	0.000898
Rented	52284	0.74824	-4.06E-03	0.00091

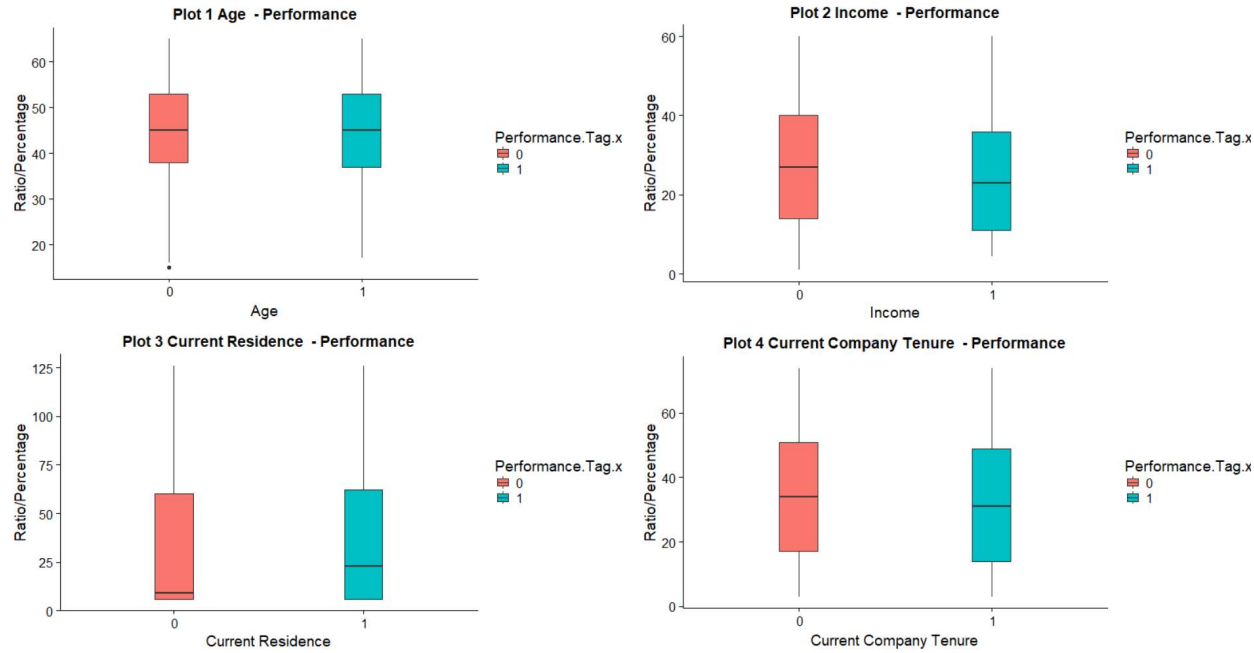
NA's is imputed by Owned based on the WOE values



Presence.of.open.home.loan	N	Percent	WOE	IV
<NA>	272	0.0039	-0.37556	0.000465
0	51465	0.73786	0.073141	0.004547
1	18012	0.25824	-0.23504	0.017374

NA's in the home loan are replaced by 1

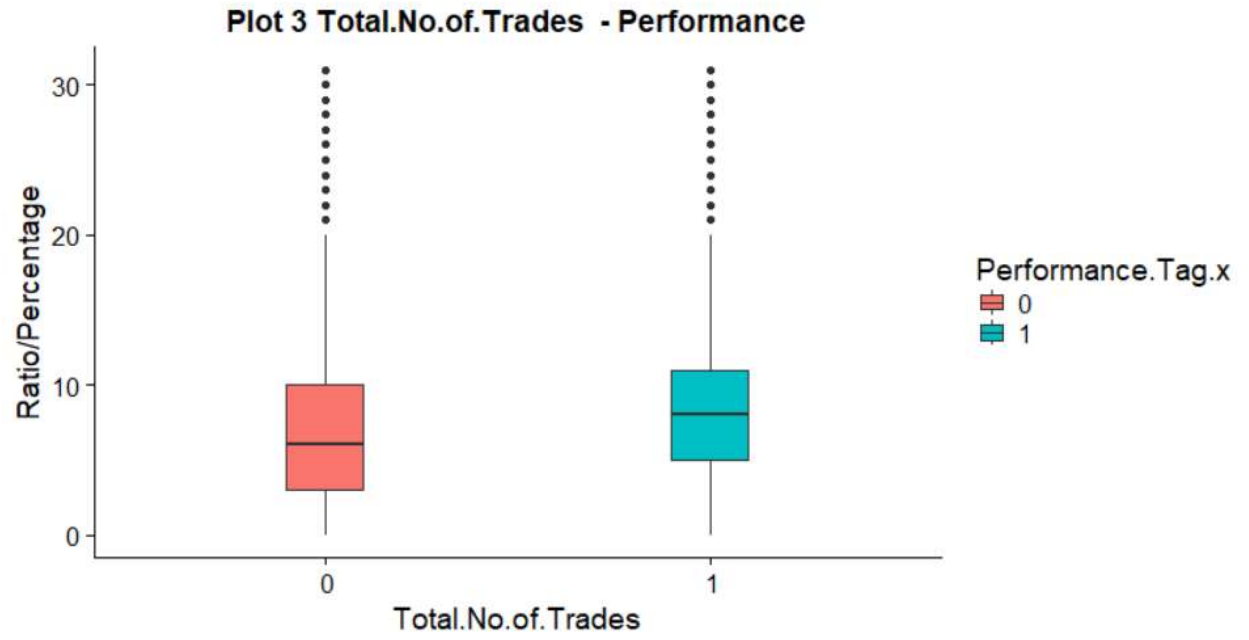




Maximum credit cards are approved for the customer between the age 35 to 55

As the number of months in the current residence increases the probability of getting default increases

As the number of months in the current company increases the probability of getting default decreases



As the number of trades increases the default value increases.

Summary of IV values (values greater than .001)

Variable	IV
No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.	0.3000220328
No.of.PL.trades.opened.in.last.12.months	0.2990278512
Avgcc.bin	0.2983927653
Total.No.of.Trades	0.2680068387
No.of.times.30.DPD.or.worse.in.last.6.months	0.2443179708
No.of.PL.trades.opened.in.last.6.months	0.2240952793
No.of.times.30.DPD.or.worse.in.last.12.months	0.2182590121
No.of.times.90.DPD.or.worse.in.last.12.months	0.2156178433
No.of.times.60.DPD.or.worse.in.last.6.months	0.2112879579
No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.	0.2091875050
No.of.trades.opened.in.last.6.months	0.1915765595
No.of.times.60.DPD.or.worse.in.last.12.months	0.1882095654
No.of.trades.opened.in.last.12.months.bin	0.1856445680
No.of.times.90.DPD.or.worse.in.last.6.months	0.1626467849
Income.bin	0.0348009471
NbrMntRes.bin	0.0278193041
NbrMntCmp.bin	0.0196844179
Presence.of.open.home.loan	0.0173203299
No.of.dependents	0.0026507311
Profession	0.0021605382
Presence.of.open.auto.loan	0.0016080634

EDA Summary

- 4% of the approved applicants gets default.
- Approval rate for male is higher and so is their default rate
- Approval rate for married applicants is higher and so is their default rate
- Salaried applicants have higher default rate
- Applicant with open auto loan have higher default rate
- Applicants with open home loan have higher default rate
- As the number of trades increases the default value increases.
- Maximum credit cards are approved for the customer between the age 35 to 55
- As the number of months in the current residence increases the default rate increases
- As the number of months in the current company increases the default rate decreases
- Based on the WOE analysis, the following are the top 10 variables that will have significant influence on the model
    - No.of.Inquiries.in.last.12.months..excluding.home...auto.loans
    - No.of.PL.trades.opened.in.last.12.months
    - Avgcc.bin (binned variable of Avgas.CC.Utilization.in.last.12.months)
    - 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.30.DPD.or.worse.in.last.12.months
    - No.of.times.90.DPD.or.worse.in.last.12.months
    - No.of.times.60.DPD.or.worse.in.last.6.months
    - No.of.Inquiries.in.last.6.months..excluding.home...auto.loans

MODEL BUILDING

Handling Imbalanced datasets:

The data set seems extremely imbalanced with only about 4% of the applicants getting default. With such imbalanced data set the model might miss out on the information required to do accurate prediction. We will consider the below techniques to handle the imbalanced dataset.

- ***Under sampling:*** Under-sampling balances the dataset by reducing the size of the abundant class.
- ***Oversampling:*** It tries to balance dataset by increasing the size of rare samples.

Initially we would try to build the model by using the original dataset, and then try to see if the under sampling and oversampling increases accuracy and stabilize the models.

Model Building Approach:

We will consider below techniques in building the acquisition model.

- ***Logistic Regression:*** This model will be used as the benchmark model, as it used heavily in the BFSI industry due to its interpretability. Ideally it will help us understand the predictive power of the variables.
- ***Decision Tree:*** Decision trees are easy to interpret results and explain the results to the executives. Also, the non-linear relationships between parameters would not affect the performance.
- ***Random Forest:*** Random forest helps overcoming the overfitting problem by decision trees. A random forest is not affected by outliers too much because of the aggregation strategy.

The best model will be chosen based on the predict power, discriminatory power and stability of the model.

MODEL EVALUATION

Model Performance Evaluation:

The metrics we would consider while evaluating the models are as below:

- ***Predictive Power:***
    - Accuracy
    - Sensitivity
    - Specificity
- ***Discriminatory power:***
    - Gain & Lift Chart

    - KS statistic
    - ROC Curve
- ***Model Validation:***
    - In-time Validation
    - K-fold Validation

APPLICATION SCORECARD

Steps to create the application score card:

- Calculate Odds(Good) for each candidate. Odds(Good) = P(Good)/P(Bad)
- Sort the applicants in the descending order of their odds.
- Develop the score card with good to bad of 10 to 1 at a score of 400 doubling every 20 points.
- Assess the application scores of the rejected applicants and compare it with the approve applicants to see the insights.
- Calculate the appropriate cut-off score below which applicants would not be granted credit cards.