Approach paper

CREDX- Credit Application
MID SUBMISSON

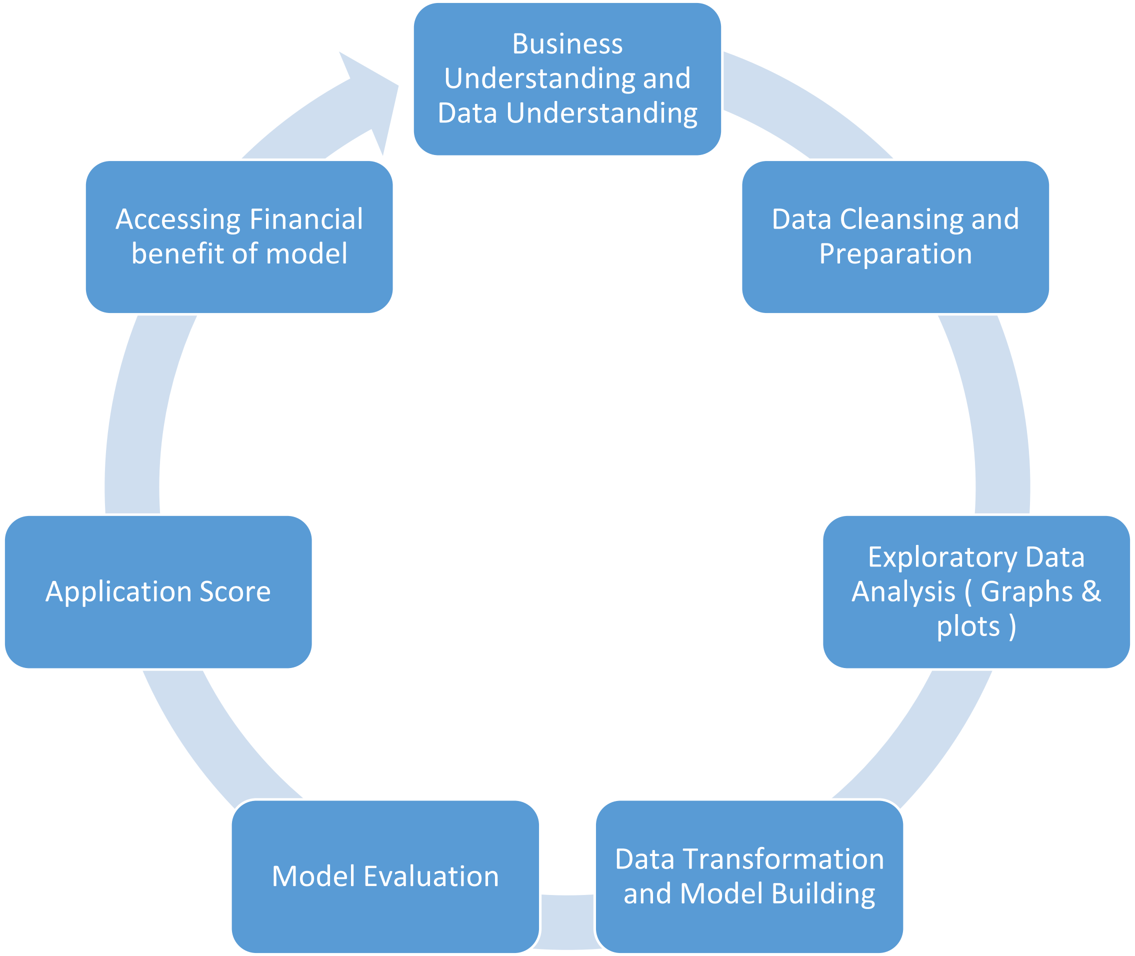
Business Understanding

Objective:
To identify the right customers using predictive models by determining the factors affecting credit risk and *c*reating strategies to mitigate them.

Problem Statement:
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 due to increase in defaults.

Solution Approach:
This is a binary supervised classification problem. We aim at building models such as Logistic regression, Random forest etc. to identify the customers who are at a risk of defaulting if offered a credit card. We have followed CRISP–DM framework. It involves the following series of steps:

Cross Industry Standard Process for Data Mining



Business Understanding and Data Understanding

Data Cleansing and Preparation

Exploratory Data Analysis (Graphs & plots)

Data Transformation and Model Building

Model Evaluation

Application Score

Accessing Financial benefit of model

Data Understanding

Two datasets are provided, demographic data and credit bureau data.

•1.Demographic/application data: This dataset contains 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.

•2. Credit bureau data: This information 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.

Nature of data:

- The demographic data consists of 71295 observations with 12 variables.
- The credit bureau data consists of 71295 observations with 19 variables.
- Application ID is the common key between the two datasets for merging.
- Performance Tag is the target variable which says if customer is default or not. The values are 0(nondefault) and 1(default).

Data Cleaning and Preparation

- The 1425 rows with no performance , we will use the data with null values as a test data
- Both occurrences of 3 duplicate Application ID records (765011468, 653287861, 671989187) has been excluded from the dataset.
- Since 18 is the minimum age to grant credit card, the 65 records with age <18 , we will impute binned WOE value.
- We will impute missing values with WOE.

Variable	Missing Values
Gender	2
Marital Status (at the time of application)	6
No of dependents	3
Education	119
Profession	14
Type of residence	8
Performance Tag_x	1425
Avgas 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 Tag_y	1425

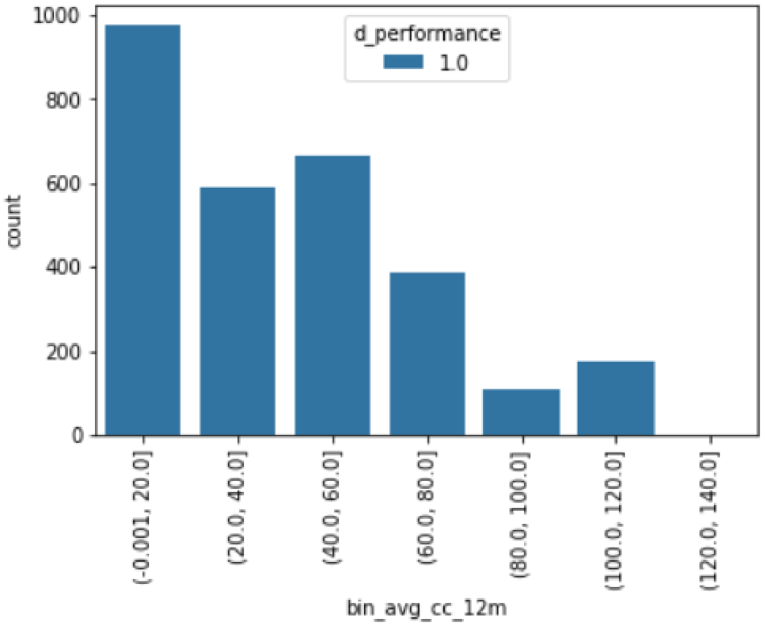
WOE and IV

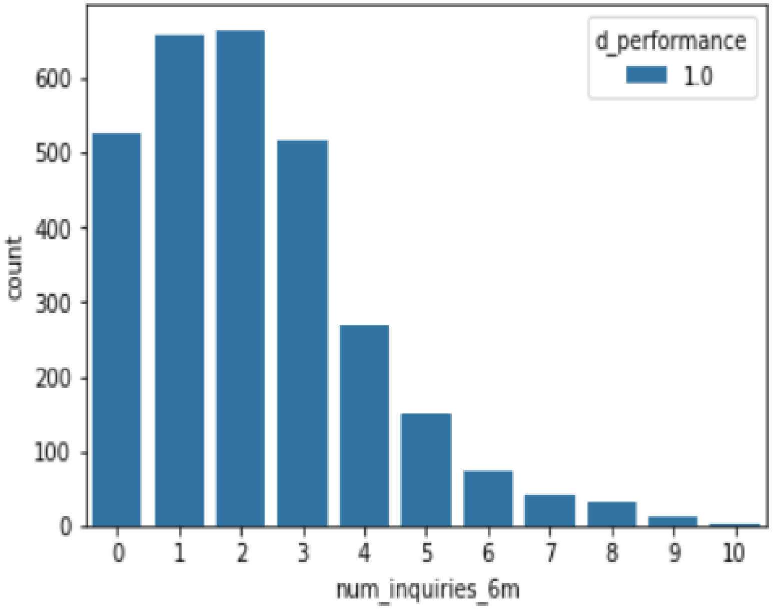
- The weight of evidence tells the predictive power of an independent variable in relation to the dependent variable. Since it evolved from credit scoring world, it is generally described as a measure of the separation of good and bad customers.
- **Benefits of WOE**
  It can treat outliers. Suppose you have a continuous variable such as annual salary and extreme values are more than 500 million dollars. These values would be grouped to a class of (let's say 250-500 million dollars). Later, instead of using the raw values, we would be using WOE scores of each classes.
- It can handle missing values as missing values can be binned separately.
- Since WOE Transformation handles categorical variable so there is no need for dummy variables.
- WoE transformation helps you to build strict linear relationship with log odds. Otherwise it is not easy to accomplish linear relationship using other transformation methods such as log, square-root etc. In short, if you would not use WOE transformation, you may have to try out several transformation methods to achieve this.
- **Information Value (IV)**
- Information value is one of the most useful technique to select important variables in a predictive model. It helps to rank variables on the basis of their importance.

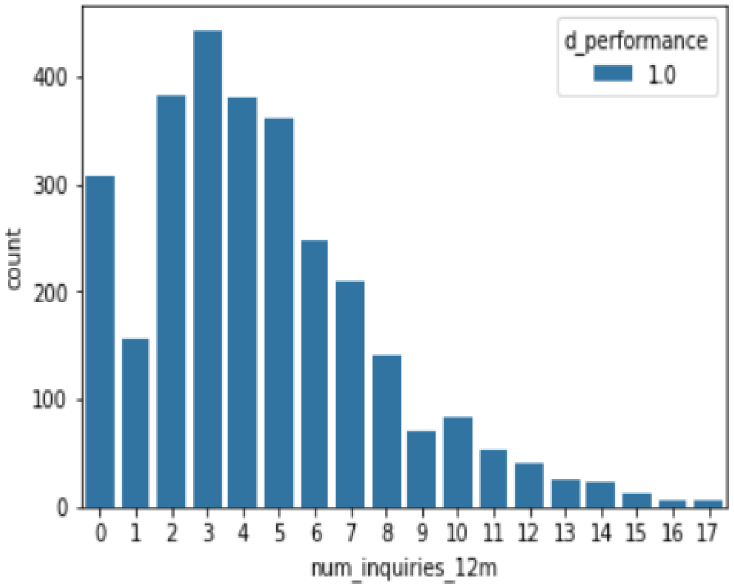
Top 9 Variables with IV value of 0.1 to 0.3 has medium predictive power and can be considered significant. There is no variable with strong predictive power with IV more than 0.5

	VAR_NAME	IV
2	Avgas CC Utilization in last 12 months	0.29352
20	No of trades opened in last 12 months	0.25743
7	No of Inquiries in last 12 months (excluding h...	0.22922
14	No of times 30 DPD or worse in last 12 months	0.18805
27	Total No of Trades	0.1873
9	No of PL trades opened in last 12 months	0.17664
15	No of times 30 DPD or worse in last 6 months	0.14571
16	No of times 60 DPD or worse in last 12 months	0.13768
10	No of PL trades opened in last 6 months	0.12474
18	No of times 90 DPD or worse in last 12 months	0.09571
21	No of trades opened in last 6 months	0.09533
8	No of Inquiries in last 6 months (excluding ho...	0.09294
17	No of times 60 DPD or worse in last 6 months	0.08957
13	No of months in current residence	0.04876
5	Income	0.03597
19	No of times 90 DPD or worse in last 6 months	0.03071
12	No of months in current company	0.01096
22	Outstanding Balance	0.00825
26	Profession	0.00222
24	Presence of open auto loan	0.00166
28	Type of residence	0.00092
3	Education	0.00078
0	Age	0.00063
25	Presence of open home loan	0.00046
4	Gender	0.00033
6	Marital Status (at the time of application)	9.5E-05
11	No of dependents	5.6E-05
1	Application ID	0.00003

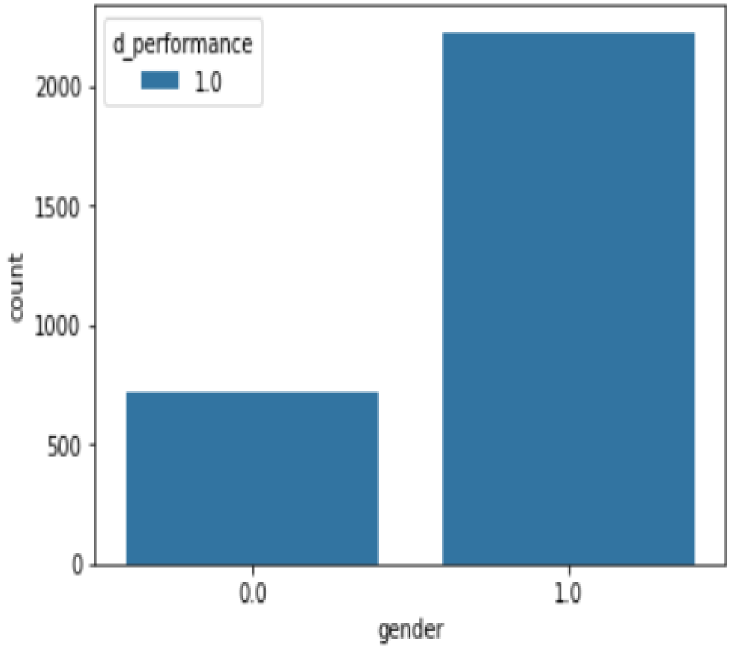
Defaulters behavior

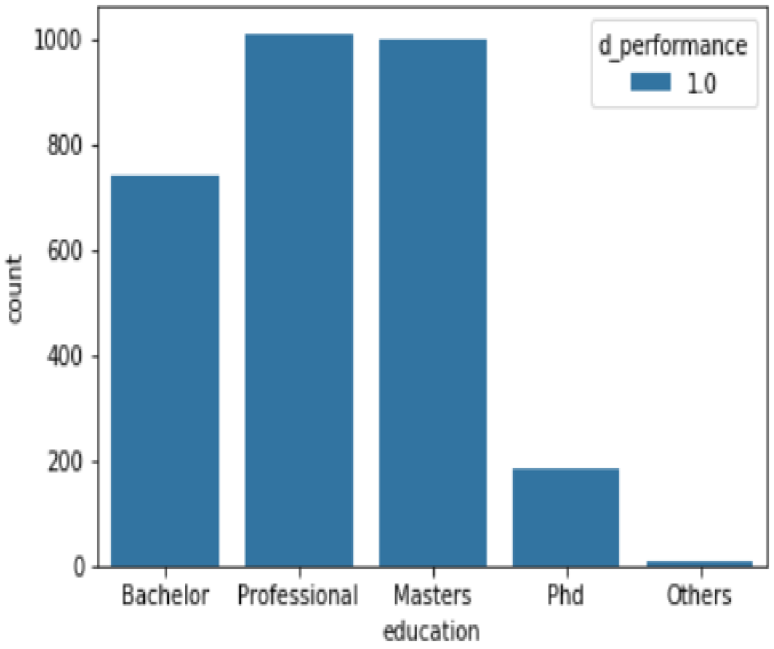


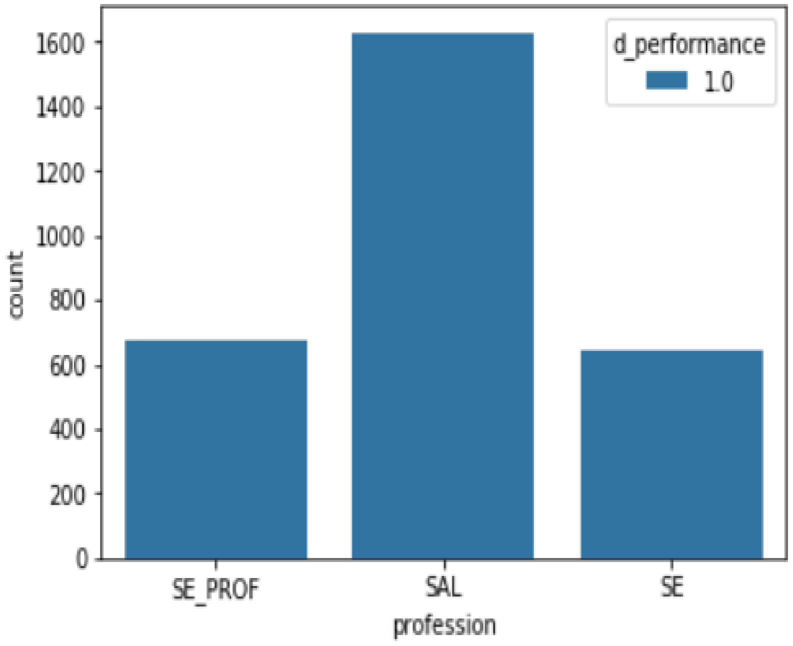




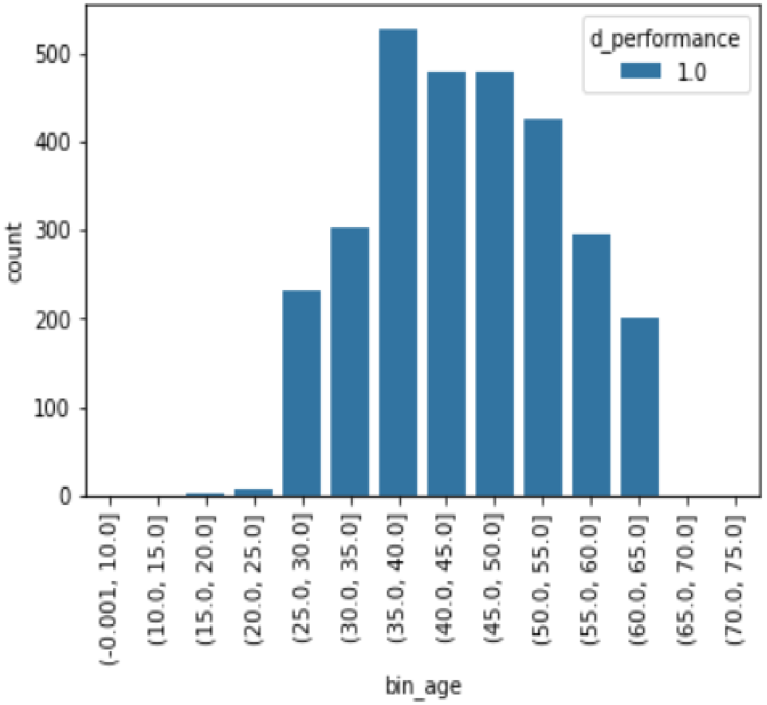
1. Shows behavior of people who uses credit limit upto 60k are likely to default
2. People who does upto 2-3 times inquiries for credit card within 6 months are likely to default
3. People who does upto 5-6 times inquiries for credit card within 12 months are likely to default

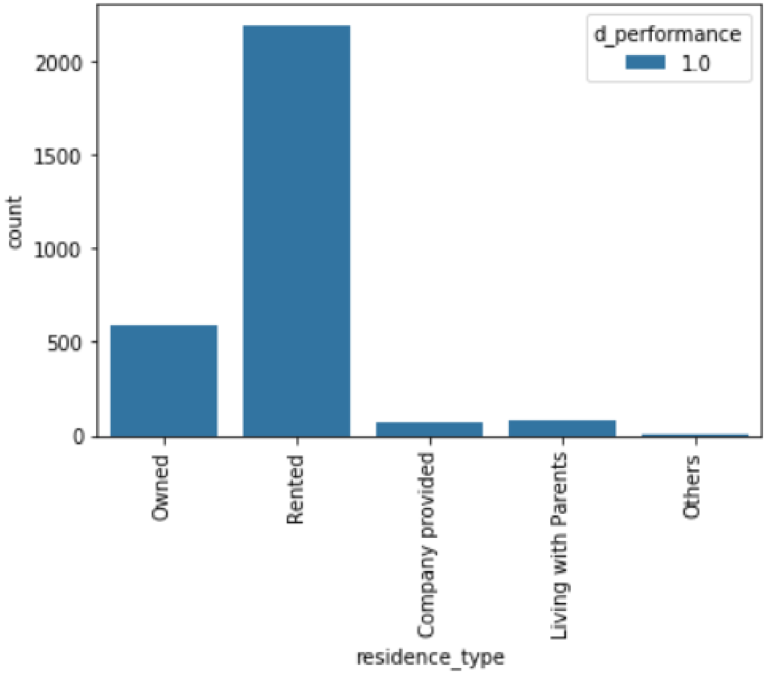


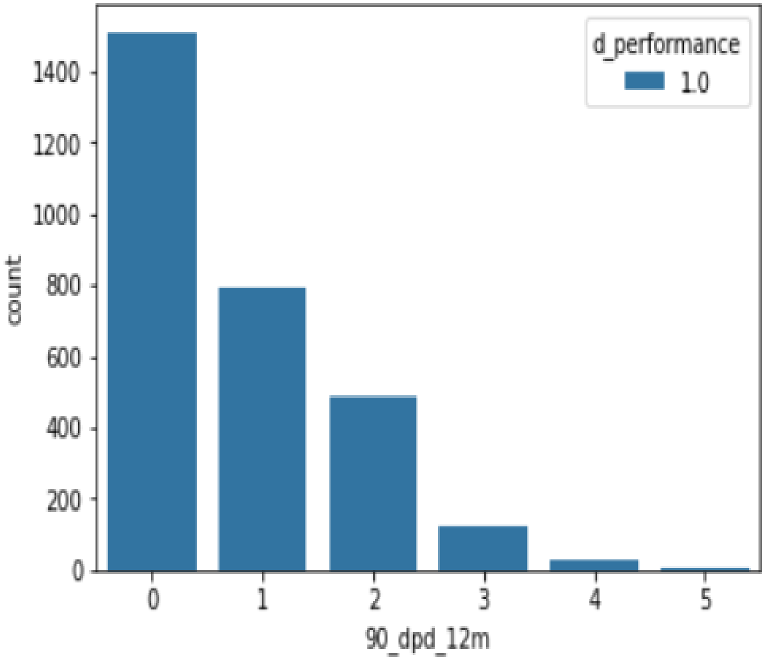




1. Males are more likely to default
2. Professionals and Master degree holder are more likely to default
3. Salaried people are more likely to default







1. People between age of 35 to 55 are more likely to default
2. People living in rented home are more likely to default
3. People who had late payment for 90 days are less likely to default.

Model Building

- Considering the classification problem of dividing the applicants in two categories based on the performance tag – Defaulters and Non Defaulters, we can use two different models.

1) Logistic Regression

2) Random Forest

3) SVM

4) Naive Bayes

- Segregating the data into test and train sets we will use the drill down approach to remove the non significant variables on the basis of VIF and p-values.
- In random forest we need to vary the number of trees, min number of buckets and min number of leaves in a node.

Model Evaluation Techniques

- Plotting the sensitivity, specificity and accuracy at various cut-off values
- Choosing the best cut-off value where all the three parameters are very high
- Plotting the confusion matrix for the best cut-off value
- Using the KS-Statistics and Lift-Gain chart to check for better performing model out of the two using the cross validation in case of logistics regression to fine tune the model

APPLICATION SCORE CARD

- The application score for each applicant calculated using the logistic regression model
- Method used for computation of application scorecard:

a) Need to compute the probabilities of default for the entire population of applicants using the model.

b) Need to compute the odds for the good. Since the probability computed is for rejection (bad customers),

Odd(good) = (1-P(bad))/P(bad)

Road Map

- We will make the model using various algorithms and will find the best model which perform better.
- We will propose to evaluate the model using different techniques like Confusion matrix, K-fold cross validation techniques, KS-Statistics, and based on that we would decide on the best model for our case.
- We will re-build the application scorecard on the final model to find the cut-off score and predict the potential
- We will find financial benefits for the company.
- We will predict the likelihood of default for the rejected candidates using the model.