BFS Capstone Project
CredX Credit Card Customers – Acquisition Analytics
Approach Document

BFS-12 Group Members

Sailendra Kona - DDA1720038
Maheshwara Reddy Golla - DDA1720125
Sumit Arora - DDA1720037
Krishnamani Ananthanarayanan - DDA1720224

Contents

Problem Statement

A leading credit card provider CredX receives thousands of applicants every year and experiencing increased credit loss over the last few years. The CEO wants to mitigate the credit risk by using the best strategy of acquiring the right customers.

Analytical Problem Solving Approach

To solve the business problem of acquisition analytics we are using CRISP-DM framework and involves the following steps

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

Business Understanding

Business Objective

CredX intends to mitigate their credit risk during acquisition by '*Finding The Right Customers*'. A *Right customer* would be the one who is not too risky to pay back their loan. In addition, right customer is the one who might miss one or two payment past due date but eventually closes the loan with interests and penalty.

Goals of Data Analysis

The objective solution to this problem and goals of data analysis are,

- Using past data of the bank's applicants identify the most important factors affecting credit risk
- Create strategies to mitigate the acquisition risk for new applications, by identifying right customers using predictive modelling to differentiate *Good Vs Bad* customer
- Build and application Score card and assess the financial benefit of project.

Data Understanding

We have two sets of data collected

- **Demographic data**: It contains customer-level information like *Age, Gender, Marital Status* and *Salary* etc.
- **Credit bureau:** This is taken from the credit bureau, contains past *Avg Credit Card Utilization, Outstanding balance and 30/60/90 DPDs in last 6/12 months etc*

Assumptions

- **Rejected Records**
  The given data is for approved loans and hence the *NA* values of *Performance Tag* considered as rejected applications.
- **Feature Engineering**
  The statistical methodology would suffice to be a standard for feature engineering
- **Data Cleaning**
  The records with very low percentage of incorrect and/or missing values have very low significance impact and hence they are excluded from analysis and modeling.

Describing Demographic Data and Quality Issues

Column Name	Type	Data Values / Range	Missing Values (NA, BLANK)	Invalid Data Values	Validation Checks / Rules
Application ID	Numeric		No		Duplicate Application ID values exist
Age	Numeric	Age Range 15-65	No	-3, 0	Age >= 18 is the eligible criteria
Gender	String	Values – F / M	Yes 2 Values		
Marital Status (at the time of application)	String	Married / Single	Yes 6 Values		
No of dependents	Numeric	Values Range 1-5	Yes 3 Values		
Income	Numeric	Value Range 0-60	No	-0.5	
Education	String	Bachelor / Masters / Other / PhD /Professional	Yes 119 Values		
Profession	String	SAL / SE / SE_PROF	Yes 14 Values		
Type of residence	String	Owned / Rented /Others / Company Provided /Living with Parents	No		
No of months in current	Numeric	6-126	No		

residence					
No of months in current company	Numeric	3-133	No		
Performance Tag	Numeric	0 / 1 1 represents "Default = Yes"	Yes 1.99% or 1425 BLANK Values		Default Yes – 4.1% Default No – 93.8%

Describing Credit Bureau Data and Quality Issues

Column Name	Type	Data Values / Range	Missing Values (NA, BLANK)	Invalid Data Values	Validation Checks
Application ID	Numeric	-	No	-	Duplicate Application IDs exist
No of times 90 DPD or worse in last 6 months	Numeric	0-3	No	-	
No of times 60 DPD or worse in last 6 months	Numeric	0-5	No	-	
No of times 30 DPD or worse in last 6 months	Numeric	0-7	No	-	
No of times 90 DPD or worse in last 12 months	Numeric	0-5	No		
No of times 60 DPD or worse in last 12 months	Numeric	0-7	No		
No of times 30 DPD or worse in last 12 months	Numeric	0-9	No		
Avgas CC Utilization in last 12 months	Numeric	0-113	Yes 1023 Values		
No of trades opened in last 6 months	Numeric	0-12	Yes 1 Value		
No of trades opened in last 12 months	Numeric	0-28	No		
No of PL trades opened in last 6 months	Numeric	0-6	No		

No of PL trades opened in last 12 months	Numeric	0-12	No		
No of Inquiries in last 6 months (excluding home & auto loans)	Numeric	1-10	No		
No of Inquiries in last 12 months (excluding home & auto loans)	Numeric	0-20	No		
Presence of open home loan	Numeric	0 /1	Yes 272 Values		
Outstanding Balance	Numeric	0-3884434	Yes 272 Values		
Total No of Trades	Numeric	1-44	No		
Presence of open auto loan	Numeric	0/1	No		
Performance Tag	Numeric	0 /1	Yes 1.99% - BLANK 1425 Values		93.8% - No 4.1% - Yes

Data Preparation and Feature Engineering

For data understating we would be doing the following checks

- Identification of categorical and continuous variables.
- Check for null values, missing values sanity check, duplicate records & outliers
- Univariate and bivariate analysis for categorical variables
- Univariate and Bivariate analysis for continuous variables

Data Cleaning

- **Removed Duplicates** records in both data sets i.e. rows with Application ID values 765011468, 653287861 and 671989187
- **Performance Tag** values as NA - Separated these rejected applications data records with and create another data frame as **rejected_applications**. This is used later to evaluate the final model.
- **Age** – Removed records with incorrect values -3 and 0. Also, remove with age <18 years, as credit card is not approved.

Data Integration

- Merged both Demographic data with Credit Bureau data using Application ID and create master data frame

Outlier Treatment

- **Income –** Outliers will be removed
- **Outstanding Balance –** Outliers will be removed

Missing Value Treatment

- **Demographic data**
  - Gender – Imputed with *Mode* value
  - Marital Status – Imputed with *Mode* value
  - No of dependents – Updated '0' for missing values
  - Education – Imputed with *Weight of Evidence*
  - Profession – Impute with *Mode* value
- **Credit Bureau data**
  - No of trades opened in last 6 months – Updated missing value as '0'
  - Presence of open home loan – Impute with *Weight of Evidence*
  - Outstanding Balance – Impute with *Weight of Evidence*
  - Avgas CC Utilization in last 12 months – Impute with *Weight of Evidence*

Feature Engineering

The approach involves treating few features differently (i.e. *ordinal vs nominal vs continuous*) and adding new features from existing features. Adding more features during data preparation makes different modeling techniques more flexible. For e.g. one-hot encoding is good for Linear models and label encoding works well with Tree models

One-Hot Encoding or Dummy Variables

- Gender – 2 values 0/1
- Marital Status – 2 values 0/1
- Profession – Dummy variables
- Typo of residence – Dummy variables
- Presence of open home loan – 2 values 0/1

Label Encoding

- Education – Ordinal categorical variable

Binning / Buckets

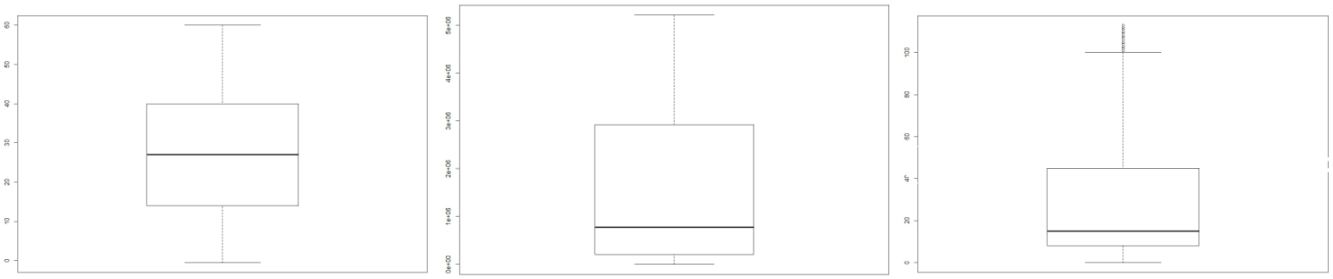
- Age
- Income
- No of months in current residence
- No of months in current company
- Avgas CC Utilization in last 12 months
- Outstanding Balance

Exploratory Data Analysis

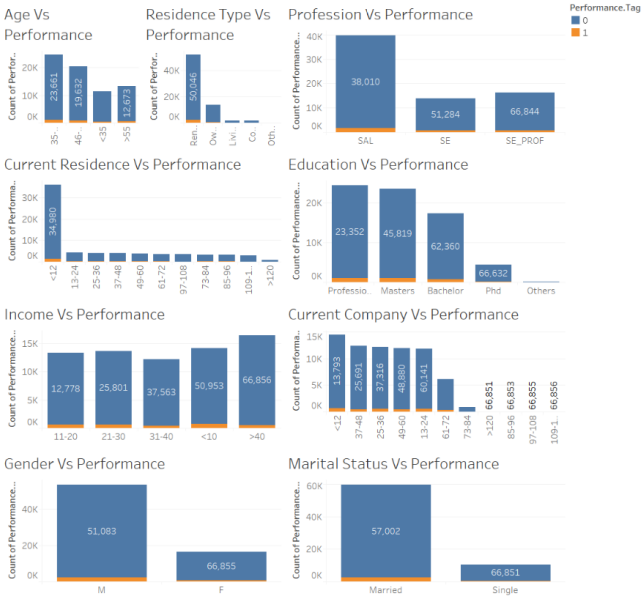
Following are some of the important plots for understanding data through exploratory analysis.

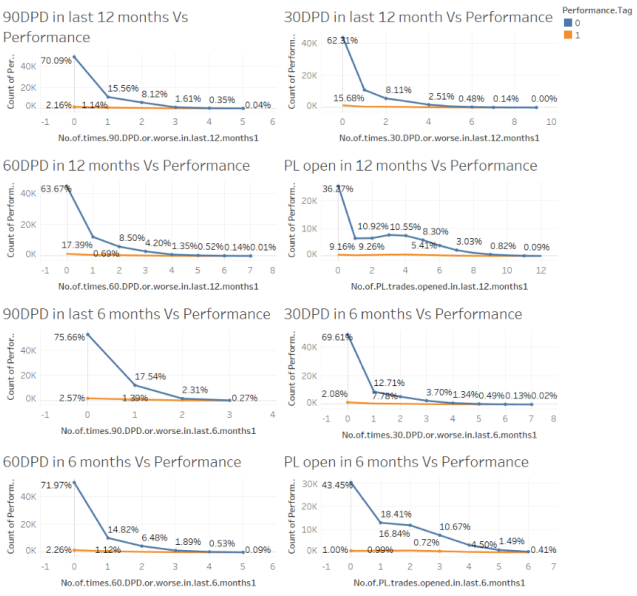
Univariate Analysis

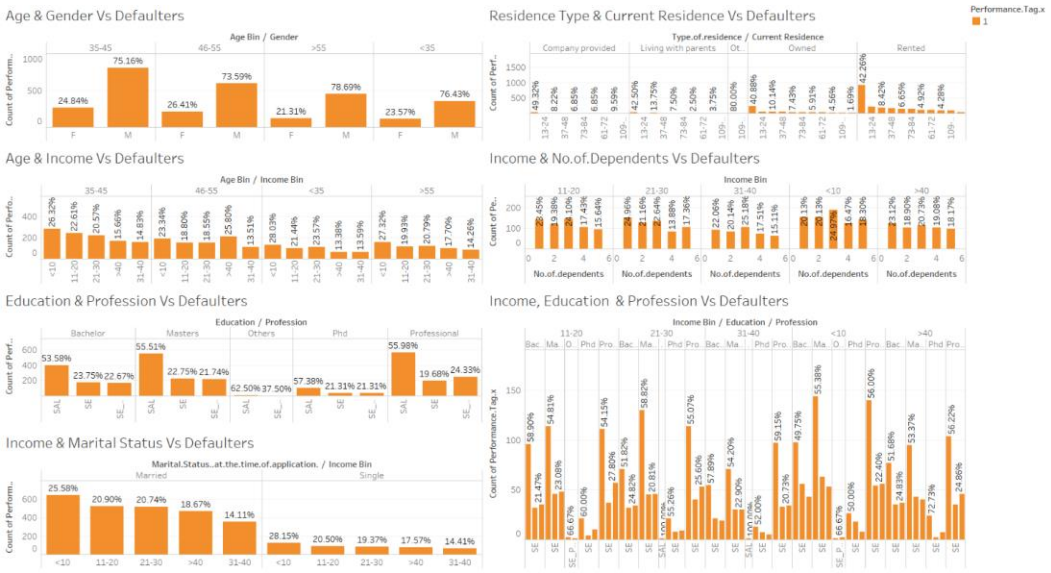
Income, Outstanding Balance and Average Credit Card Utilization



Bi-variate and Multivariate Analysis







Roll Rate Matrix

	current	0-29	30-59	60-89	90-119	120-149	150-179	180+
current	**0.96**	**0.04**	0.00	0.00	0.00	0.00	0.00	0.00
0-29	**0.33**	**0.55**	**0.12**	0.00	0.00	0.00	0.00	0.00
30-59	**0.08**	**0.46**	**0.40**	**0.07**	0.00	0.00	0.00	0.00
60-89	0.01	**0.26**	**0.47**	**0.24**	**0.03**	0.00	0.00	0.00
90-119	0.00	**0.06**	**0.34**	**0.43**	**0.15**	0.02	0.00	0.00
120-149	0.00	0.02	**0.13**	**0.36**	**0.35**	**0.13**	0.02	0.00
150-179	0.00	0.00	**0.04**	**0.19**	**0.39**	**0.29**	**0.09**	0.00
180+	0.00	0.00	0.00	**0.12**	**0.23**	**0.38**	**0.24**	0.04
Charged Off	0.00	0.00	0.00	0.00	**0.04**	**0.26**	**0.61**	**0.09**
Paid	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00

12 month roll rate matrix

	current	0-29	30-59	60-89	90-119	120+
current	**1.00**	0.00	0.00	0.00	0.00	0.00
0-29	**0.21**	**0.79**	0.00	0.00	0.00	0.00
30-59	0.02	**0.68**	**0.30**	0.00	0.00	0.00
60-89	0.00	**0.16**	**0.70**	**0.14**	0.00	0.00
90-119	0.00	0.00	**0.31**	**0.61**	**0.08**	0.00
120+	0.00	0.00	0.00	**0.43**	**0.52**	**0.04**
Charged O	0.00	0.00	0.00	0.00	**0.60**	**0.40**
Paid	0.00	0.00	0.00	0.00	0.00	**1.00**

6 month roll rate matrix

Weight of Evidence (Information Value) Analysis

The data contains status of customer performance through variable *Performance Tag* with value 1 representing *Default* and 0 for *Non-Default*. We leverage R Information package for computing the *Information Values (IV)*. But this package interprets 1 value for *Good* which is contradictory to business case here as Performance Tag value. So, we need another variable *Performance Tag* for IV with values 1 and 0 replaced with 0 and 1 respectively.

WoE & IV Formulas

Following is approach for computing WoE for variables as an example.

WoE = Log (No.of Goods/Total No.of Goods) – Log (No.of Bads/Total No.of Badds)
*IV = WoE * (No.of Goods/Total No.of Goods - No.of Bads/Total No.of Badds)*

Following is the standard reference table for interpreting variable predictive power based on IV values.

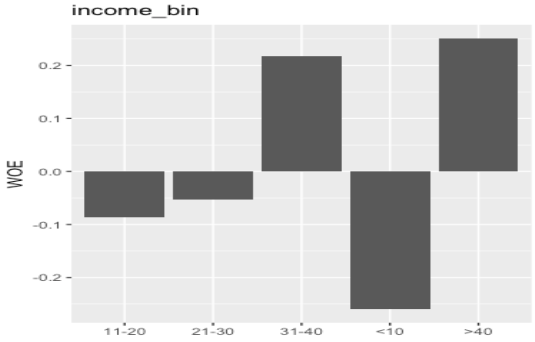
Information Value	Predictive Power
< 0.02	useless for prediction
0.02 to 0.1	Weak predictor
0.1 to 0.3	Medium predictor
0.3 to 0.5	Strong predictor
>0.5	Suspicious or too good to be true

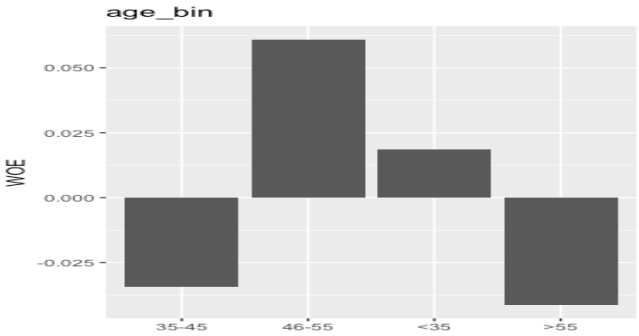
WOE & IV matrix for AGE_BIN and INCOME_BIN

	age_bin	N	Percent	WOE	IV
1	35-45	24740	0.3544311	-0.0342993	0.0004235771
2	46-55	20446	0.2929142	0.06086307	0.0014788933
3	<35	11361	0.1627604	0.01864939	0.0015350204
4	>55	13255	0.1898943	-0.04133416	0.0018656668

	income_bin	N	Percent	WOE	IV
1	11-20	13392	0.1918570	-0.0866074	0.0014975290
2	21-30	13628	0.1952380	-0.05284889	0.0020562160
3	31-40	12179	0.1744792	0.21745032	0.0095327010
4	<10	14155	0.2027879	-0.25970513	0.0249573500
5	>40	16448	0.2356379	0.25138452	0.038248765

WOE Plots for AGE_BIN and INCOME_BIN





WOE is computed and analyzed similarly across all other variables. Variables shortlisted using WOE and IV are used as a precursory set of variables that are further passed on to the next stages of the modelling.

Summary of IV Values (with only values > 0.02)

Rank	Variable	IV
1	Avgas.CC.Utilization.in.last.12.months	3.101586E-01
2	No.of.trades.opened.in.last.12.months	0.2982771000
3	No.of.PL.trades.opened.in.last.12.months	0.2960405000
4	No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.	0.2956044000
5	Outstanding.Balance	0.2460422000
6	No.of.times.30.DPD.or.worse.in.last.6.months	2.416795E-01
7	Total.No.of.Trades	2.371398E-01
8	No.of.PL.trades.opened.in.last.6.months	2.19731E-01
9	No.of.times.90.DPD.or.worse.in.last.12.months	2.139377E-01
10	No.of.times.60.DPD.or.worse.in.last.6.months	2.059261E-01
11	No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.	2.052399E-01
12	No.of.times.30.DPD.or.worse.in.last.12.months	1.984825E-01
13	No.of.trades.opened.in.last.6.months	1.859535E-01
14	No.of.times.60.DPD.or.worse.in.last.12.months	1.85638E-01
15	No.of.times.90.DPD.or.worse.in.last.6.months	1.601637E-01
16	No.of.months.in.current.residence	7.91399E-02
17	current_residence	6.080075E-02
18	current_company	6.080075E-02
19	Income	4.255551E-02
20	income_bin	3.824877E-02
21	No.of.months.in.current.company	2.175181E-02

Feature Selection using Inferences from EDA and Information Values

Following is the table with variable importance from general expectation, EDA results, Information Values and final selection with business intuition applied.

Column Name	Expectation	EDA Results	Information Value	EDA Inferences and Conclusion
Age	High	Medium	Low	**High** - *Age* group of 35-55 is significant
Gender	Medium	Low	Low	EDA also confirms not significant feature
Marital Status	High	Low	Low	**Medium-** EDA also confirms ***Married*** Significant
No of dependents	Medium	Low	Low	EDA also confirms not significant feature
Income	High	Low	Low	EDA also confirms not significant feature
Education	Medium	Low	Low	EDA also confirms not significant feature
Profession	Medium		Low	**High** - *Salaried* is significant with high frequency
Type of residence	High	High	Low	**High** - *Rented* is the most significant with high frequency

No of months in current residence	High	High	Low	**High -** < 12 months is high frequency
No of months in current company	High	Medium	Low	**Medium** - EDA also confirms **<24 Months** has significant default
No of times 90 DPD or worse in last 6 months	High	Low	Medium	**Medium –** Higher the number has default effect
No of times 60 DPD or worse in last 6 months	High	Low	Medium	**Medium –** Higher the number has default effect
No of times 30 DPD or worse in last 6 months	Low	Low	Medium	**Medium –** Higher the number has default effect
No of times 90 DPD or worse in last 12 months	High	Low	Medium	**Medium –** Higher the number has default effect
No of times 60 DPD or worse in last 12 months	Medium	Low	Medium	**Medium –** Higher the number has default effect
No of times 30 DPD or worse in last 12 months	Low	Low	Low	EDA also confirms not significant feature
Avgas CC Utilization in last 12 months	High	High	High	**High - Most of the utilization are <20**
No of trades opened in last 6 months	Medium	Low	Medium	EDA also confirms not significant feature
No of trades opened in last 12 months	Medium	Low	Medium	EDA also confirms not significant feature
No of PL trades opened in last 6 months	Medium	Low	Medium	EDA also confirms not significant feature
No of PL trades opened in last 12 months	Medium	Low	Medium	EDA also confirms not significant feature
No of Inquiries in last 6 months (excluding home & auto loans)	Medium	Low	Medium	EDA also confirms not significant feature
No of Inquiries in last 12 months (excluding home & auto loans)	Medium	Low	Medium	EDA also confirms not significant feature
Presence of open home loan	High	Low	Low	EDA also confirms not significant feature
Outstanding Balance	High	Low	Medium	EDA also confirms not significant feature
Total No of Trades	Medium	Low	Medium	EDA also confirms not significant feature
Presence of open auto loan	High	Low	Low	EDA also confirms not significant feature

Model Building

This problem belongs to supervised and binary classification problem with *Performance Tag* as the target variable.

Data Sampling

The data is high imbalanced with approximately 96% applicants are non-defaults and only 4% with defaulters. Therefore, we will apply these two sampling techniques to as balanced data set.

Under sampling
The method involves selecting equal (i.e. 4% or slightly more) number of samples of rare class, from abundant class and thus create a new balanced data set.

Over sampling
The method involved in this technique is SMOTE (Synthetic Minority Over-Sampling Technique).

Hybrid Sampling
In general, a combination of both under and over samplings does give better results and will choose best strategy for the business problem here.

Cross Validation

These two techniques are considered for cross validation

- K-fold
- Stratified K-fold

Model Selection

We selected following three different modelling techniques for this binary classification problem.

Logistic Regression
The model built with Logistic Regression will be our case base reference model. As this is the model highly used in BFS domain due to its interpretability and ease with which it can separate classes linearly.

Important Note: We will use ***stepAIC*** to deal with Multi-collinearity during the model development.

Decision Trees
The model built with decision trees is more helpful for both interpretability and deals well with imbalanced data sets and variables with label encoding.

Random Forest
This modeling technique helps over-fitting problem by decision trees and produces high efficiency than any single model by nature.

Model Evaluation

The evaluation of the models involves using relevant metrics and reporting the results. Following are the relevant metrics based on which the best model is selected.

Standard Metrics
- Sensitivity
- Specificity
- KS Statistic
- ROC Curve

Metrics for Imbalanced
- F1 Score
- MCC
- AUC

Credit Score
- Vintage Curve

Additionally a part of model validation, predict the likelihood of default for the rejected candidates and assess whether the results correspond to the expectations.

Model Deployment and Application Scorecard Preparation

Approach

In order to arrive at risk scores, we will follow the below steps:

1. Compute odds of a particular customer being 'good'
   $$Odds(good) = P(good) / P(bad)$$
2. Sort the customers from high to low odds (i.e., good to bad)
3. Calibrate the Odds (10 to 1) to a score of 400 doubling at every 20 points (points to double odds or PDO = 20). Below table will be updated

Table for reference

Application Scorecard					
Application ID	P(good)	P(bad)	Odds(good)	ln(Odds)	Score

4. Scores will then be computed on rejected population and the results will be compared and assessed

5. Analysis will be performed to assess the scores and the distribution of customers across scores to arrive at an appropriate cutoff
6. Discriminatory power (KS, Gini, Sensitivity, Specificity), accuracy and stability of the risk scores will be assessed

Assessment of Financial Benefit using the Model

The model will help the bank in making right acquisition decisions and thus reducing the credit loss. The implications of using the model for auto approval or rejection, i.e. how many applicants on an average would the model automatically approve or reject

Following performance measures will be computed to understand the impact of the model:

- Correctly Classified Instances (Rate) ≡ (TP + TN)/N
- Incorrectly Classified Instances (Rate) ≡ (FP + FN)/N
- True Positive (TP) Rate ≡ TP/(TP + FN)
- False Positive (FP) Rate ≡ FP/(FP + TN)
- Precision ≡ TN/(TN + FN)
- Recall ≡ TN/(TN + FP)
- F-Measure ≡ (2 × Recall × Precision)/(Recall + Precision)

The potential credit loss avoided / potential benefit of the model. Following metrics will be computed to denote the financial befit of the model.

Profit without Forecast = (TP + FN) ·Br ·Pm − (FP + TN) ·Bd
Profit with Forecast = TP ·Br ·Pm − FP ·Bd − TN ·Br .
Savings = TN(Bd − Br) − FN ·Br ·Pm

Where:

Br - running balance of all customers
Bd – balance of defaulting customers
Pm - Profit Margin Rate
TN(Bd − Br) = Savings due to correct decision
FN ·Br ·Pm = Opportunity cost due to incorrect decision .
Value Added ≡ (TN − FN ·Pm ·(Br / Bd−Br)) / TN + FP