Le Duc Bao

Data Science in   
Financial Service

Individual Assignment

Logo, company name

Description automatically generated

# Introduction

This assignment is to use Credit Scorecard to predict customer’s financial distress in the near future (2 years). The data provided from Singapore Management University, Data Science in Financial Service Course.

# Data description

The training dataset has 150,000 rows with 1 binary dependent variable ‘SeriousDlqin2yrs’ and 10 numerical independent variables (see **Appendix: Data dictionary**). The dependent variables is very skewed: 139,974 good customers and 10,026 bad customers (representing a 93%/7% distribution). Among the 10 independent variables, ‘MonthlyIncome’ has 29,731 and ‘NumberOfDependents’ has 3,924 missing values. Many of them also have skewed distribution and extreme outliers (**Appendix: Data distribution example**). We also found no strong correlation among the independent variables.

# Data pre-processing

Due to the findings that the independents are heavily skewed and contain extreme outliers, as well as missing data, we propose 3 procedures to clean the data. The 3 procedures may or may not be used in conjunction, depending on the data analysis method described in a later section. **Procedure 1**: Clipping extreme values & Natural logging; **Procedure 2**: Impute the missing data with the median value of the variable; **Procedure 3**: Variable binning. The detailed steps for each procedure can be found at **Appendix: Pre-processing procedures**

# Data analysis

In order to calculate the probability of default, we propose 2 methods:

## Method 1: Blackbox machine learning model Random Forrest

Text

Description automatically generated

We clip and impute most variables, storing the action we did in 2 separate columns (‘is\_clipped’ and ‘is\_imputed’) to preserve the original information. This is because in the financial setting, missing information or extreme values can carry some information that is relevant to the probability of loan default. Since the original distribution of default are skewed, we choose to perform oversampling on the bad samples to create a more balanced dataset, which can improve training. The probability of default is produced by the ensemble voting algorithm.

## Method 2: Whitebox scorecard model using LogisticRegression

Diagram, text

Description automatically generated

Since in data pre-processing procedure 3, we divide data into bins and create a separate bin for missing values, outliers and missing data does not affect our pipeline, thus procedure 1 and 2 are not necessary. After binning, we calculate the WoE and IV of each variable, and selecting only those with IV between 0.02 and 0.6, as is recommended by Naeem (2017)[[1]](#endnote-1). We replace each value from the original dataset with its WoE, run a logistic regression and use the coefficient and intercept to calculate the scoring for each attribute. Our setting is dpo = 20, factor of 600 points for a 50:1 good to bad ratio.

# Solution and result

For evaluation purpose, we split our training data into 60% training set and 40% dev set

Our RandomForestClassifier gives an AUC score of 0.86 on both training and dev set. Our LogisticRegression gives an AUC score of 0.78 on both training and dev set (**Appendix: ROC Charts)**. We also produced a scorecard table (**Appendix: Credit Scorecard**).

Both methods show no sign of overfitting but some sign of underfitting, especially for method 2. Due to the requirement to create a transparent model for regulatory purposes, we favour method 2 to method 1.

The distribution of scores vs ‘SeriousDlqin2yrs’ for the full training data below.

Chart, histogram

Description automatically generated

The score ranges from 500 to 640, which is considerably lower than the average FICO of 716[[2]](#endnote-2). This probably is because our training dataset is not representative of the population. Our training data set has 7% default rate, corresponding to a good:bad ratio of approximately 50:4, i.e., a 560 score.

However, we can still observe that our scoring can do a reasonably good job at separating good and bad customers, with good customers has much better score distribution than bad customers.

# Conclusion

In conclusion, we created a moderately good model to predict customer default, together with a detailed scorecard to explain our scoring.

Further work can be done to improve the performance of our model. First, domain knowledge should be applied to create a better binning strategy, instead of a n=6 bins in our current model. Second, we are using the recommended default pdo = 20 and a benchmark 600 points means 50:1 good:bad odds. This could be updated using expert opinion to fit the change in the macro environment. Third, expert opinion could be used to form better feature to for our models.

# Appendix

## Data dictionary



## Data distribution example

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

## Pre-processing procedures

* **Procedure 1a**: Clip the upper range of all variables with outliers by its 99th percentile (except for ‘DebtRatio’ at 70th percentile). I.e., all original values higher than the threshold of 99th or 70th percentile will be replaced by the 99th or 70th percentile. This will remove all extreme outliers.
* **Procedure 1b**: We will create a new binary column ‘is\_clipped’ for each clipped variable. E.g., if the ‘age’ variable is clipped, then the value of ‘age\_is\_clipped’ will be set to True, otherwise False. This is to retain the information value that is deleted when Procedure 1a is performed.
* **Procedure 2a**: Impute the missing data with the median value of the variable.
* **Procedure 2b**: Create a new binary column ‘is\_missing’ for each imputed value. E.g., if the ‘age’ variable is missing and imputed under procedure 2a, then the value of ‘age\_is\_missing’ will be set to True, otherwise False. This is to retain the information value that is deleted when Procedure aa is performed.
* **Procedure 3**: Bin each variable to 6 approximately equal sized bins. If the data is unable to be binned to n equal sized bins, we will reduce n until such binning is possible. Missing data is assigned a separate bin.

## Credit Scorecard



Figure 1: Excerpt of the Credit Scorecard

## ROC Charts

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Figure 2: ROC graph of RandomForestClassifier (left) and Logistic Regression (right)

1. Naeem, S. (2017). Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards (2nd ed.). Wiley. [↑](#endnote-ref-1)
2. https://www.fico.com/blogs/average-us-ficor-score-716-indicating-improvement-consumer-credit-behaviors-despite-pandemic [↑](#endnote-ref-2)