CS 6220 Summer '18 Project Report : Home Credit Default Risk

Chia Yi Liaw
liaw.c@husky.neu.edu

Mounica Subramani
subramani.m@husky.neu.edu

Sharyu Deshmukh
deshmukh.s@husky.neu.edu

Somya Bhargava
bhargava.so@husky.neu.edu

1 Abstract

Loan repayment ability is a supervised classification problem. The objective of this project is to use historical loan application data from the Kaggle Home Credit Default Risk dataset to explore effective methods and create classifier to predict either loaner would be able to repay based on statistic method and supervised machine learning.

2 Introduction

2.1 Project Description

Home Credit is a finance provider that focuses on serving the unbanked population. Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders. So, Home Credit tries to broaden financial inclusions for unbanked population. The Home Credit Default Risk challenge is a standard supervised machine learning task where the goal is to use historical loan application data to predict their clients' repayment abilities based on datasets provided.

- Supervised: The labels are included in the training data and the goal is to train a model to learn to predict the labels from the features
- Classification: The label is a binary variable, 0 (will repay loan on time), 1 (will have difficulty repaying loan)

2.2 Exploratory Data Analysis

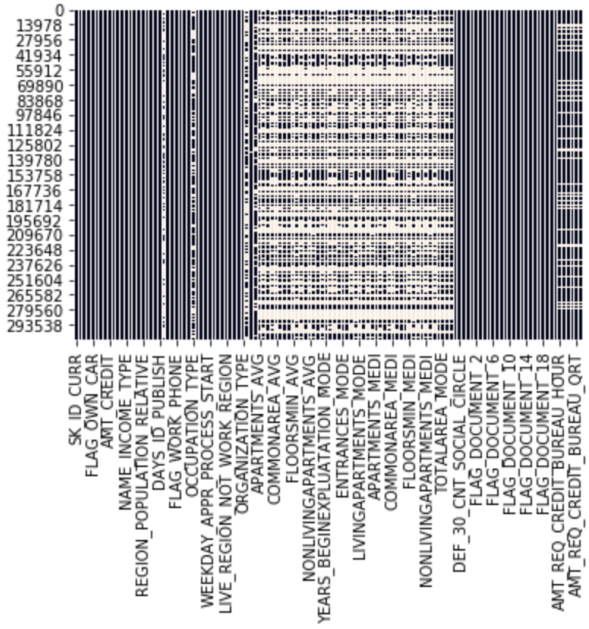

Figure 1: Missing Value Map

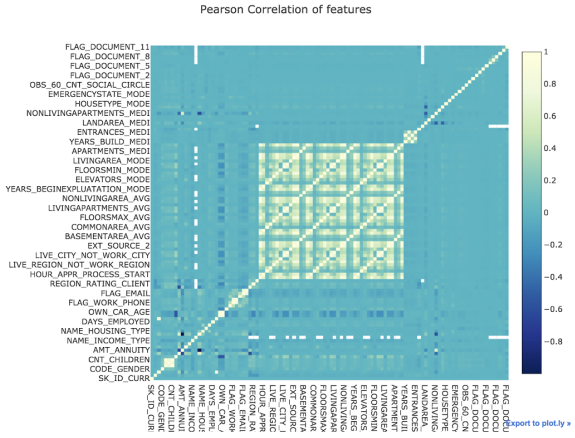

Figure 2: Feature Correlation Map

- Missing values: The Dataset contains 122 features with 307511 entries. As the respect of missing value, The distribution map gives better evaluation and understanding of missing data. From the visualization, we found out that missing value is concentrated in several features. In addition, there are 17 feature contains more than 60 percent of missing value. Imputation and deletion will be performed in the data pre-processing stage.

- Feature correlation : By calculating the r-square to determine the linear correlation between features. Pearson feature correlation quantifies the degree to which a relationship between two variables. By analyzing correlation feature, we implement the random forest method and LightGBM to improve calculation efficiency. Pearson correlation heat map shows the correlation between predictors, the lighter green indicates the higher correlation.

2.3 Data Preprocessing

- Feature selection : With the high dimensional and to reduce the variance of the model or chances of over-fitting. There are 17 features contain more than 60 percent missing value, hence we decide to remove those features as it's not suitable for training models. In addition, apply the LightGBM, removes the features that are collinear, low importance and zero importance, beside LightGBM, also run the random forest to rank the feature and then remove the features with the threshold equal to 0.0003. We have also used an ensemble technique where we used LightGBM gradient boosting model on the result of random forest. The outputs so generated are stored in csv format so that we can train them all and compare the results.

- Encoding the Categorical Variable : Encoding transforms categorical features to a format that works better with classification and regression algorithm. By using the one hot encoding for multiple categories and label encoding for binary categorical features, datasets would be better processed.

- Imputation of Missing Value: After feature selection , there are still some columns with missing data. Hence, will input the missing value with mode which is the most frequent value in the column, moreover, it would apply well on the categorical variable.

- Dealing with Imbalance Data: The below histogram shows the imbalanced distribution of loan repayment in the dataset. The target variable defines whether the loan was repaid by the borrower or not. From the graph, we found out that the data is highly imbalanced. Since, it might lead to incorrect result during modeling process, we needed re-sampling of the dataset. As a result, we have used both over and under-sampling. We will be using various methods to train the model. By looking at the evaluation, we will choose the better sampling method.

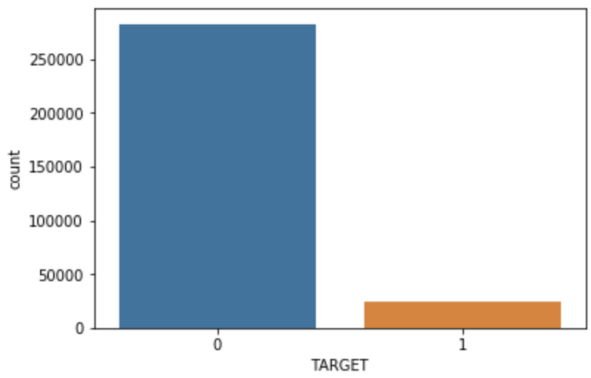

Figure 3: Distribution of target variable before sampling

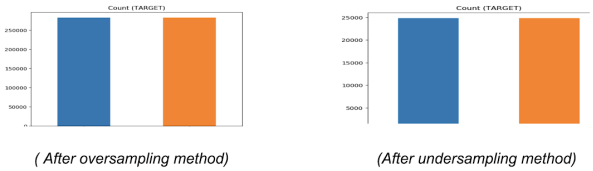

Figure 4: After undersampling and oversampling

3 Methodology

- Logistic Regression
- Support Vector Machine
- XGBoost
- LightGBM
- Ensemble models
- Catboost
- K-nearest neighbors
- Random Forest

4 Code

5 Results

- Visualizations

6 Discussion

7 Future Work

8 Conclusion

9 References