MAFS 6010Z Artificial Intelligence in AI : Project 1 - Home Credit Default Risk

Wong Hoi Ming (20641276) / WONG Sik Tsun (20038819)

1. Introduction

In this project, we attempt to build a classifier to predict whether a client will have difficulty for repayment via utilization of multiple datasets, including the 1) the static data of the current loan applications, 2) the details of previous applications and 3) the clients' history of repayment and installment. Our evaluation criterion will mainly be AUC score.

2. Dataset #1 – applications {train| test}

We treated numerical and categorical features separately. For categorical variables with many unique values (eg. organization or occupation), we use our domain knowledge to regroup the values so that the dimension of resultant dataset after one-hot encoding could be lower. For numerical features, we engineered 8 new features (eg. annuity/credit ratio, employed days/age ratio, geometric mean of EXT_SOURCE) which we believe can complement the existing variables. We also considered some illogical numerical inputs in some features to be missing values, such as "DAYS_EMPLOYED".

3. Dataset #2 – previous applications

We took a similar approach for the data set of previous applications. We regarded illogic inputs in the "DAYS_xxxxx" columns as missing values, and reclassify types of goods, sellers and suite associated with the clients or previous loans. We note that some "SK_ID_CURR" have multiple loan records. In order to merge with other datasets, we first aggregate multiple records of different variables in this dataset so that each "SK_ID_CURR" has only observation. We use max, min and mean to aggregate the numerical features, while using only mean for categorical ones.

4. Dataset #3 – installments

For the installment data set, we only use numerical variables pertaining to the repayment history of the clients. We engineered 3 new features, including 1) the number of days of delayed payment, 2) the overdue amount and 3) the ratio of overdue amount to the required installment. The ratio is set zero if there is no overdue payment. We then aggregate multiple records of the same ID into one by taking their max, min, mean and sum.

5. Model Specification

As we note that the fully aggregated dataset has numerous categorical features, we expect the relationship between the default probability or event and the features to be non-linear, and thus we use a more flexible model.

We fit our data to a gradient boosting decision tree and our model is specified below. At the m-step, let J_m be the number of its leaves for the model function F_m. The tree partitions the input space into J_m disjoint regions R_1m, R2m, … and predicts a value in each region.

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} \mathbf{1}_{R_{jm}}(x),\ \gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma)$$

The negative gradient is computed, and the model is updated to minimize the loss function. To check for overfitting, we will apply 10-fold cross-validation and compute the individual and average AUC score.

6. Results & Analysis

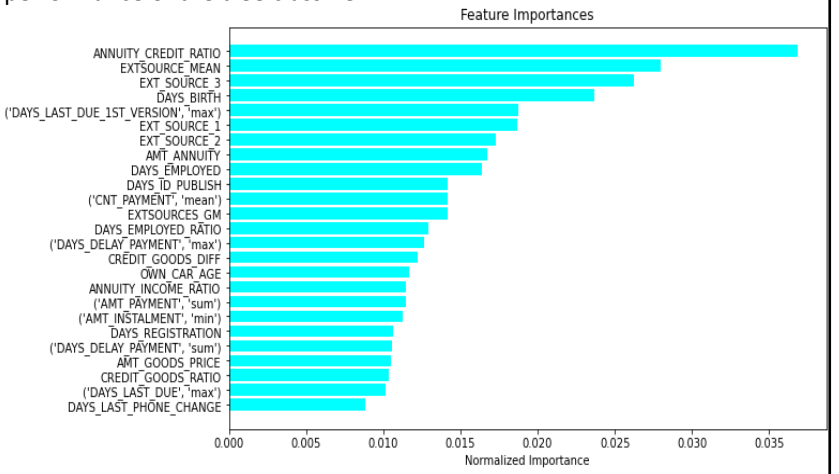
Our gradient boosting tree recorded an overall AUC of 0.845 and 0.783 respectively in the training set and validation set (shown in following table). It achieved similar AUC score at 0.782 in predictions on Kaggle.

2 submissions for math6010_wong_wong — Sort by

Submission and Description	Private Score	Public Score
submission1.csv a minute ago by Hoi Ming Wong	0.77651	0.78156

6. Results & Analysis (Continued)

Below illustrates the feature importance of the fitted model. The result is consistent with our understanding that 1) the ratio of loan annuity to the loan's credit amount, 2) scores from multiple external sources, 3) clients' ages & employment days and 4) the number of due days of the previous application have significant importance in predicting the clients' repayment abilities. The Kaggle score also indicated decent prediction performance of the tree classifier.



To improve the model, the following could be considered.

- Increase the number of iterations, but it will demand higher computational power and consume longer time
- Include more datasets for training the model, such as the data from credit bureau and credit card balances
- Include additional features or re-engineer the features, which may need deeper domain knowledge

7. Contribution

As we are a small group of two, both of us are involved in all the following steps – data cleansing & aggregation, model fitting and result analysis.