MATH 4995 Mini-Project 2: Home Credit Default Risk

Ngai Nok Yiu (20510180), Cheung Hang Yee (20514796)

1. Introduction

Home Credit default risk problem aims to predict whether or not an applicant will be able to repay a loan based on previous loan and application information. This is a supervised 2-class-classification problem. Our focus in this project is to explore different feature selection method, to save computation power, we use application and bureau data only.

2. Feature Engineering

First, we aggregate the numeric loan information made by each customer in in bureau and bureau_balance with count,mean, max, min and sum. For the categorical columns, we aggregate the value with sum and mean.

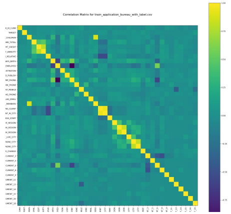
Second, Bureau and bureau_balance is merged together on 'SK_ID_BUREAU'. Then this joint table is merged to application so that one row in the final joint table represent each loan made by each customer. This final joint table contains 308 columns.

Fitting these columns to a light gradient boosting model and get a score of 0.75686.
The next step is to do feature selection so that we can save computation power and improve the generalization of the model.

Feature Selection - univariate selection

We calculate linear correlation each variable in the training table and then remove the highly correlated variables to remove collinearity.
A total of 71 columns is removed. Some examples of the removed columns are LIVINGAREA_AVG and LIVINGAPARTMENTS_AVG.

We fit this training data to a light gradient boosting model and get a score of 0.742

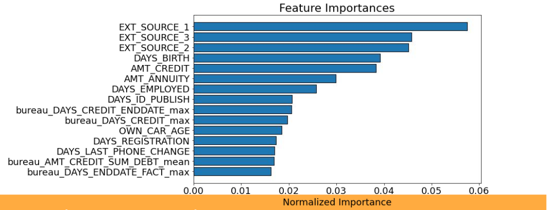


Correlation Matrix plot of training data

Feature Selection - linear model

The second method we test for feature selection is using the feature importance generated in the light gradient boosting model. We fit the training data to the model for 10 iteration and then calculate the average feature importance. There are 77 features with 0.0 importance. Some examples are FLAG_MOBIL which has the same value in the whole column.

This method get a score of 0.75717. The following is the feature importance graph.



Feature Selection - Random Forest

The third method we test for feature selection is using the feature importance generated in the Random Forest model, by using Mean decrease impurity. We compare the difference between removing the useless feature and the using full feature . The score the using the full feature is 0.610 and the score removing the useless feature is 0.723.

6. Conclusion

The disadvantage of using univariate selection is that it can only consider linear relationship between variables while in real world, the relationship are often more complicated.
Light gradient boosting model can handle null value thus reduce the effect on imputing a value to fillna.

7. Contribution

Feature Engineering, Feature selection: Ngai Nok Yiu

Feature selection: Cheung Hang Yee