MAFS 6010Z Project 1: Home Credit Default Risk

XIA Yiqiao yxiaat@ust.hk, Department of Mathematics, HKUST

1. Introduction

Home Credit Default Risk is a Kaggle project that aims to predict clients' repayment abilities with telco and transactional information. I managed to extract some features from various dataset and have gone through quite a few feature selection methods as well as machine learning models. The best model scored **0.75685** in Kaggle with team name **mafs6010z_Xia**.

2. Data Aggregation

Bureau balance status

- Turn into dummy variables and **exponentially weighted** by time, so that both time and status are reflected in the data.

Bureau

- Filter data within last 1000 days to ensure relevance.
- Aggregate data by using **mean**, so that categorical data frequency is reflected.

Credit card balance

- For each previous id, features are **exponentially weighted** by time (\sigma weight = 1) ;
- It is observed that most credit cards (87.2%) are **active** by the application time. And active credit card records are considered more important. Considering the nature of features, records of active credit card are **summed up** to generate features;
- Credit card status percentages are calculated.

Pos cash balance

- It is observed that active and completed constructed the majority of status. Features are **generated by these two status** respectively.

Previous application

- Credit card **status** percentages are calculated.
- **Credit** granted is summed up.

Instalment

- **LESS_PAY** is generated as the percentage that instalment is not paid.
- Instalment payment **time difference** is calculated.

3. Feature engineering

Feature generation

- Features that has top 4 correlation with default are used to generate polynomial features.
- Domain Knowledge Features inspired by Koehrsen.

Feature selection

- Generated features are filtered through Pearson correlation to ensure that the correlation between any 2 features is less than 0.5

Missing value handling

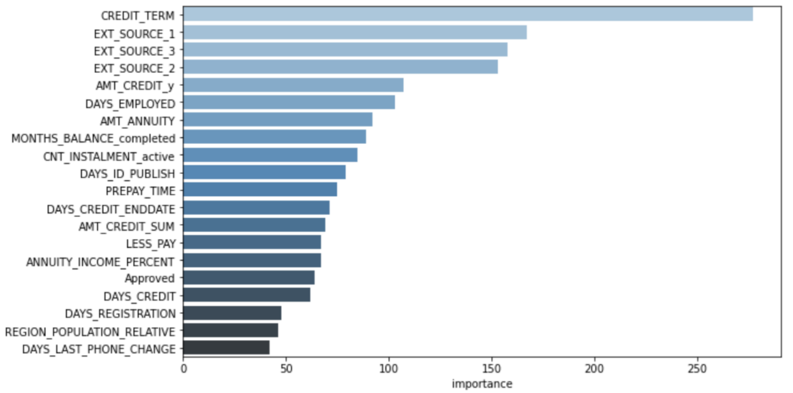
- Missing values are filled with median of train data.

3. Model Construction

Tested models

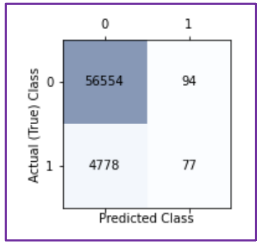
- KNN
- Logistics Regression
- Random Forest
- Light GBM
- Naive Bayes
- SVM
- Decision Tree

4. Feature Importance – Light GBM Model



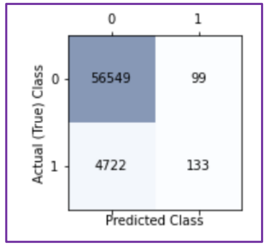
Top 20 feature importance

5. Model Performance – confusion matrix and ROC



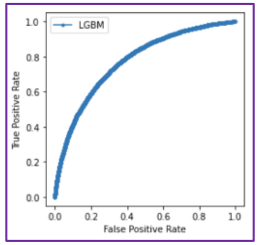
	0	1
0	56554	94
1	4778	77

Confusion matrix - logistics regression



	0	1
0	56549	99
1	4722	133

Confusion matrix – light GBM



ROC curve – light GBM

6. Analysis and Conclusion

Firstly, judgement from institutions is very helpful. CREDIT_TERM is the ratio of credit and loan annuity, which shows the distance from the credit. EXT_SOURCE also weighted high, which are scores from external data source. Credit amount, historical approved records,

Secondly, current personal states are important. For instance, DAYS_EMPLOYED shows the stableness of income and active instalment shows current instalment pressure. It's surprising that region population also matters and it's economic meaning may be further investigated.

Furthermore, historical transactions also help to understand clients. Clients default, or payed less than required are more likely to default in the future. And clients with longer credit days have more reliable credit records.

Generally, tree models perform better than linear models for this dataset. It can be further improved with more features with economic meanings.

7. References

Koehrsen, W. (2018). Start Here: A Gentle Introduction. Kaggle. https://www.kaggle.com/willkoehrsen/start-here-a-gentle-introduction