MATH 4995 Mini-Project 1: Home Credit Default Risk

Leung, Ko Tsun (20516287), Cheng, Tsz Yui (20595441), Yang, Po-Yen (20561878)

1. Introduction

Home Credit is trying to provide loans to people lack of credit history by predicting his/her ability to pay back with respect to his/her other alternative or historical data. In this project, we utilized logistic regression and light Gradient Boosting Machine (light GBM) as our prediction model.

2. Home Credit Risk Dataset (Kaggle Competition)

This dataset consists of five CSV tables that represents different financial data for Home Credit to predict his/her repayment ability. For logistic regression, we utilized only "application_train.csv", but for light GBM, we utilized "application_train.csv", "bureau.csv", "bureau_balance.csv", and "previous_application.csv".

Data Preprocessing

1. Aggregate bureau and balance data in training datasets
2. Convert categorical variables to dummy variables
3. Merge test dataset
4. Fill in missing values via SimpleImputer and MinMaxScalar

3. Feature Selection

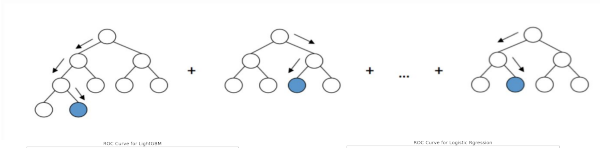
1. Filter features by the rate of missing value. Columns with more than 70% missing data are dropped as they contain insufficient information
2. Select features with sufficient corr1elation versus target column, as higher correlation indicates higher feature importance
3. Remove highly inter-correlated features to avoid multicollinearity which may weaken the statistical power of the model

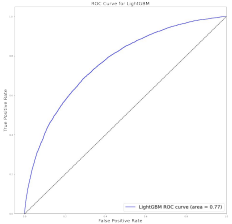
3.1. Methodology - Logistic Regression

Logistic regression is a basic statistical model that is often used to model a binary dependent variable. Since the prediction output is in the form of binary result instead of numerical value, logistic regression is chosen as its explanation ability is high.

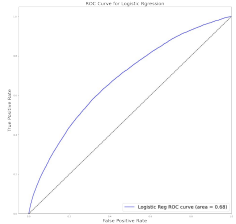
3.2. Methodology - Light GBM

Light GBM is an efficient gradient boosting framework that uses tree based learning algorithms. Light GBM grows its tree leaf-wise instead of depth-wise, meaning it chooses to grow the leaf with maximum delta loss. In addition, the important features score in the light GBM model is also very useful, and we have utilized it to find out the features that have the most significant impacts.





ROC Curve for Logistic Regression (application only)



ROC Curve for light GBM (merged)

5. Kaggle Result

Dataset & Features	Kaggle Score
Logistic Regression (application only)	0.66968
Logistic Regression (merged dataset)	0.67685
Light GBM (application only)	0.71793
Light GBM (merged dataset)	0.75763

6. Analysis

It is obvious that a simple model like logistic regression performed worse than a complicated model like light GBM. However, logistic regression is the most widely used model especially when the target variable is binary. Hence, we chose logistic regression model to be our base model and used it to compare with the performance of the light GBM model.

With light GBM model, we can easily identify which features are more important than the others, and this is useful when we are dealing with a huge and sparse dataset in high-dimension feature space. As we can see from the accuracy result, light GBM indeed performed better than logistic regression.

7. Conclusion

It is within our expectations that the light GBM model performs better the logistic regression model in terms of the prediction accuracy. On top of its model complexity, its ability to select more important features is also a key driver to this result.

We also observe that both models have better performance when the merged dataset is used as the input. This suggests that a set of more comprehensive and related data can lead to better prediction. Nonetheless, from the Kaggle result, we observed that the improvement from using the merged dataset is larger for the light GBM model.

8. References

Light GBM
https://lightgbm.readthedocs.io/en/latest/

9. Contribution

Handle Dataset
Leung, Ko Tsun

Logistic Regression
Cheng, Tsz Yui, Yang, Po-Yen

Light GBM
Yang, Po-Yen