MAFS 6010Z Project 1: Warm-up of Statistical Machine Learning: Home Credit Default Risk

Team Members: Aoran LI, Tianying ZHOU, Langting WENG, Yijia MA

1. Introduction

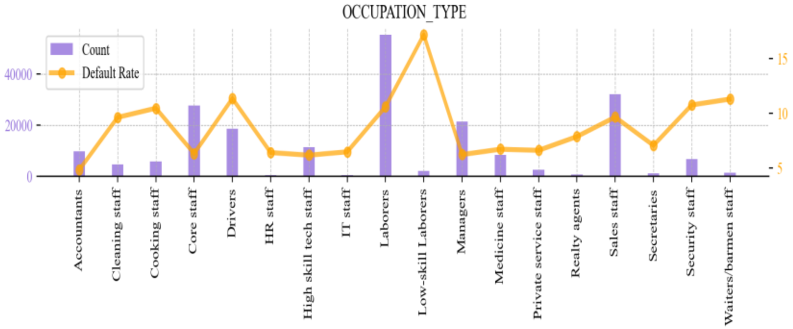
We have gone through datasets related to credit default and extracted useful information via exploratory data analysis. We run different models with processed data to analyze significant features for credit default and to predict the default probability in test data. To prove the model we selected is good enough, we also did some sensitivity analysis.

2. Exploratory Data Analysis

We used different test methods for different variables to analyze the effect of features on credit default probability.

How?

Chi-square Test, Point-Biserial Correlation, etc.



3. Feature Engineering

Previous default:

Match different "SK_ID_CURR" values based on 2 criteria：

Difference between two DAYS_DECISIONs of 1 = DAYS_DECISION of 2

DAYS_BIRTH of 1 - DAYS_DECISION 1 of 1 = DAYS_BIRTH of 2

Created a new variable called "previous default" to ascertain whether the user associated with this "SK_ID_CURR" had a previous default.

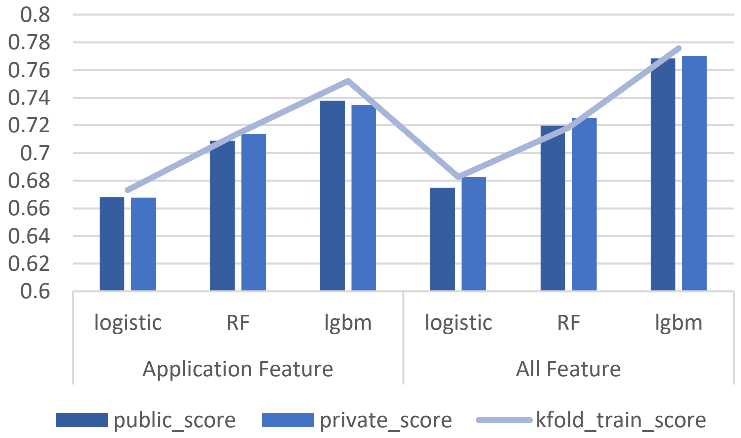
Bureau analysis: The significance of ‘bureau’ and ‘bureau_balance’ datasets is weak, but we extracted some significant features via some statistical analysis, like calculating the minimum, maximum, summation, and so on

Installments_payments analysis: Includes original features and new features constructed by simple operations， such as calculate the total default amount for each SK_ID_CIRR to conduct a new feature.

Previous_application analysis: Define functions that do one-hot encoding on a dataframe, then conduct some new features by performing a calculation on the grouped data ,and match the aggregated statistics to the appropriate client.

4. Model Construction

To show the performance of different models on different feature sets, we selected **logistic regression**, **random forest**, and **lightgbm** as comparison models and applied them to predict on different feature sets (one generated features using all data, and the other generated features using only part of the data). We obtained the scores corresponding to each model on Kaggle and conducted comparative analysis.



5. Analysis

Model Comparison:

It is obviously that when we extract features from other data tables and add them to the model, the performance of different models has significantly increased, indicating that the features we added have a significant effect on model recognition.

Feature Importance:

Both lgbm and random forest model give high score to EXT_SOURCE、YEAR_BIRTH、YEAR_PUBLISH, etc. On the contrary, logistic regression in the regression model has a significant divergence from the tree model in terms of variable importance.

6. Conclusion & Future Work

Conclusion:

- Feature engineering is of great help in improving the prediction performance of models, as new features often bring new information.
- Not all feature engineering can improve the model performance. Only when new features can provide new information to the model can the performance of the model be improved.
- Similar models have similarities in variable selection. Tree model and linear model capture different patterns of features.

Future Work:

- model hyperparameter tunning
- adding more domain relative feature

7. References

Will koehrsen. (n.d.). *Introduction to Manual Feature Engineering*. Kaggle.

8. Contribution

Exploratory Data Analysis: Tianying ZHOU
Feature engineering: Tianying ZHOU, Langting WENG, Yijia MA
Model Construction: Aoran LI