MAFS 6010Z Project 1: Home Credit Default Risk

WANG Lei, WANG Zhongchen, YE Xiaoyu and Zhang Quandi
{lwangcu, zwangfz, xyeak, qzhangcd}@connect.ust.hk
Department of Mathematics, HKUST

1. Introduction

We picked 4 more datasets besides the main table, namely “bureau”, “POS_CASH_balance”, “previous_application” and “credit_card_balance” for feature selection and feature engineering processes. After data encoding and consolidation, we regarded **logistic regression** as the baseline model for the whole project as the starting point and compared it with other 2 different models: **random forest** and **LightGBM**. To assess the effect of **PCA**, we compared the results for both data with/without selection for each model.

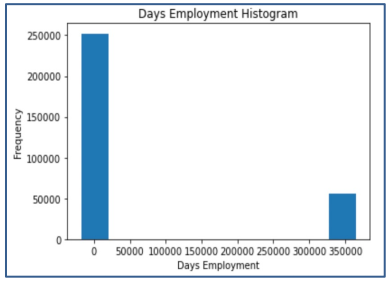
2. Data Consolidation

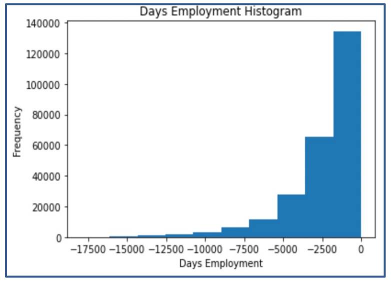
- **Encoding**

Firstly, use **label encoding** and **one-hot encoding** to encode all categorical variables by separating columns with less or equal to 2 and more than 2 unique values.

- **Anomaly Detection**

After screening, the anomaly was found in “DAYS_EMPLOYED” with histogram (left). Replacing anomaly with NaN, we obtained the updated histogram (right).





- **Imputation & Standardization**

We replaced the missing data with the median for all columns and standardize the dataset’s features onto a unit scale with mean=0 and variance = 1.

3. Feature Selection

- **Features from other datasets**

We selected “number of previous loans”, “days past due”, “normalized Interest rate”, “credit card limit during the month of the previous credit” and “balance during the month of previous credit” from other datasets as a complement of features in the mean table. For the first 3 features, we believe that the larger the values, the higher risk of default while the last 2 features are the opposite.

- **Domain Knowledge Features**

We constructed 2 domain knowledge features: ‘CREDIT_INCOME_PERCENT’, ‘ANNUITY_INCOME_PERCENT’, which stand for credit amount on the previous application/periodical payment of the previous application relative to client’s income. The higher the proportions under the condition of no default, the higher capacity of on-time repayment.

- **Principal Component Analysis**

We simply calculated the correlation between each pair of features and did not find significant linear relationship. Then we applied PCA for further feature selection, which retained 173 features (95% of variance).

4. Model Construction

For each of the 3 models, we used both data with/without PCA selection to test the effect of feature selection. The Kaggle scores for each training are as below:

Logistic Regression		Random Forest		LightGBM	
w/o PCA	w/ PCA	w/o PCA	w/ PCA	w/o PCA	w/ PCA
0.74354	0.73276	**0.73162**	0.69170	**0.75137**	0.72514

5. Conclusion

- **LightGBM performs the best while Random Forest performs the worst.**
  - As a simple model, logistic regression’s outperformance indicates a better fit of its model constriction to our data; As LightGBM is more advanced, the result meets our expectation.
- **Training without model selection by PCA has a better effect than those with PCA.**
  - PCA is a good method to select linear related features, the poor performance of PCA here indicates a non-linear relationship in between.

6. Future Work

- Apply more models, especially non-linear models;
- Further fine-turn the parameters of existing models;
- Add more reliable features based on deeper understanding of related datasets

7. References

- PCA: https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60
- Model: https://www.kaggle.com/willkoehrsen/start-here-a-gentle-introduction

8. Contribution

Code implementation

- WANG Zhongchen, WANG Lei

Model refinement

- WANG Zhongchen, YE Xiaoyu, ZHANG Quandi

Reporting

- WANG Lei, YE Xiaoyu, ZHANG Quandi