MATH 4995 Project 1: Supervised Classification with Full Spectrum or Premium Subset?

Zhihao SHAO[1,2] (zshaoac@connect.ust.hk)

[1]: Department of Life Science, HKUST; [2]: Department of Mathematics, HKUST

1. Introduction

Many people struggle to get loans due to insufficient or non-existent credit histories. In order to make sure this underserved population has a positive loan experience, Home Credit is challenging kagglers to makes use of a variety of alternative data to predict their clients' repayment abilities.

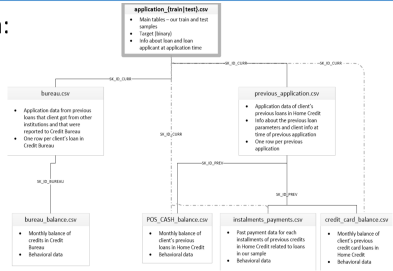
I always have a question in my mind upon being assigned a similar task: should I feed my model with a full set of features without selection, or a small subset after close examination? Which one would support a better prediction? If any difference, how large is it and is it worth the time I spend on selection?

In this project, I explored this problem on several common statistical and machine learning methods used in binary classification.

2. Dataset Description : Home Credit Default Risk

There are 7 different source of data:

- Application_train/test
- Bureau
- Bureau_balance
- Previous_application
- Pos_cash_balance
- Credit_card_balance
- Installments_payment

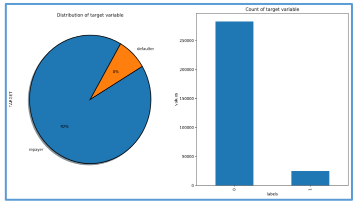


And you can find detailed description in the chart above.

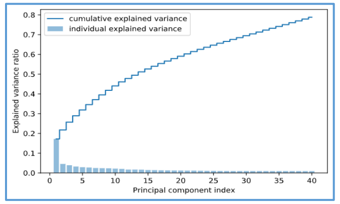
Due to the time limit, in this project I only used the main application training and testing data, which means that the models presented later were trained on incomplete data.
The result here is just a baseline prototype that we can then improve upon.

3. Exploratory Data Analysis

- This is a standard supervised classification task. The binary label 'TARGET' are included in the training data, with 0 (will repay loan on time) or 1 (will have difficulty repaying loan).
- I analyzed the predictors' data type, normality, multicollinearity, missing value status and their correlation with response.



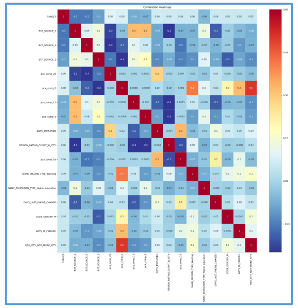
The distribution of 'TARGET' response in training dataset



Principal component explaining the data variance

4. Feature Engineering and Selection

The EDA suggests that most original features show weak correlation (absolute value < 0.1) with the response variable. In order to reduce the dimensionality, I conducted PCA analysis and extracted top 40 components which can explain about 80% of the data variance. Furthermore, 16 predictors with highest correlations were selected.

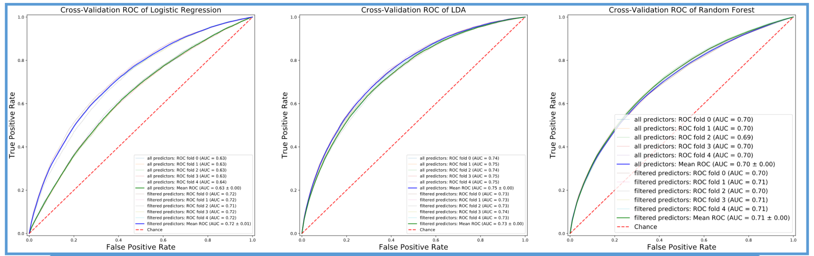


Correlation map of 16 filtered predictors

5. Modeling

- Three models were explored:
- Logistic regression;
- Linear discriminant analysis;
- Random forest.
- The models were built upon one set of 285 predictors and the other set of 16 predictors.
- 5-fold cross validation was performed for model selection.

6. Result



The best AUC score on Kaggle is (math4995_SHAO: 0.73731)

6. Analysis

- In logistic regression, the model built with filtered predictors outperforms the other one, indicating that the exclusion of multicollinearity phenomenon improves LR.
- such improvement is not significant in LDA or random forest, maybe because LDA is inherently based on stronger assumptions.
- As for random forest, one of its intrinsic advantages compared to bagging is to reduce inter-tree correlation, so it can accept many features.
- Relatively low AUC scores of my models may be attributed to
  I. Insufficient data retrieval
  II. Suboptimal number of predictors after filtering

7. Conclusion

Premium subset can improve the performance of logistic regression in this case, but no significant effect on methods that designed to mine information from numerous features.

8. References

Brownlee, Jason. 2016. "Bagging And Random Forest Ensemble Algorithms For Machine Learning". *Machine Learning Mastery*.