Modelling probability of default in credit risk

Julia Li

May 18, 2021

Strategy

Start of with googling what is the most common method for predicting default. After a quick research [1][2], logistic regression is the classical benchmark method and random forests are very common too. These are well known methods for binary classification problems and simple to implement with Python `sklearn` package, so I will try them out both. The data is exported into a csv file in order to import it in Python. Code written in Python.

Data preprocessing

See code in Appendix for basic data cleaning process. For data preprocessing, we 1) scale the bill statements and amount of previous payments with respect to the credit amount. Naturally, higher credit should allow for higher bill statements and repayments. Moreover, 2) new features are added - aggregated repayment status, aggregated bill statements, aggregated previous payments, and number of zero bill statements.

Method

Sklearn's `SelectFromModel` is performed recursively for feature selection. 10% of the data will be used as a test set. The rest of the data, which we call training data, will be applied in a 10-fold cross validation for calibrating the model parameters. For `LogisticRegression`, we tune `solver` and `max_iter`. For `RandomForestClassifier`, we tune `max_depth`, `min_impurity_decrease` and `n_estimators` and `criterion`.

Results

General result from `SelectFromModel` shows the repayment status has most significance and that the categorical features Gender, Education, Marital status and Age has the least impact on the response variable. See code in Appendix for more detailed results.

Model	Cross-val score	Test score	ROC AUC score
Logistic regression	0.81921	0.81554	0.65243
Random forests	0.81996	0.82027	0.66976

Table 1: Results

Final model choice:

```
RandomForestClassifier(random_state = 0,criterion = 'entropy', max_depth = 8,
min_impurity_decrease= 0,n_estimators = 300)
```

Respective feature selection:

```
feature_selection = ['x6', 'repay_status_sum']
```

Appendix

Link to code:
https://colab.research.google.com/drive/1CgOUXncTxLJFxmnYrYs2j0yHRc8SEu4u?usp=sharing

References

[1] Sarah Kornfeld. "Predicting Default Probability in Credit Risk using Machine Learning Algorithms". In: *KTH, School of Engineering Sciences (SCI)* (2020). DOI: https://www.diva-portal.org/smash/get/diva2:1437874/FULLTEXT01.pdf.

[2] Jinglun Yao, Maxime Levy-Chapira, and Mamikon Margaryan. *Checking account activity and credit default risk of enterprises: An application of statistical learning methods*. 2017. arXiv: 1707.00757 [q-fin.ST].