Data Analysis on Credit Card Approval Prediction -- A Brief Summary

Course: *Machine Learning and Data Science for Social Good (20S856137)*
Authors: *Boqin Cai (boqin.cai@stud.sbg.ac.at)*

A story happened to my father. He wanted to apply a credit account but he was refused many times. Finally the counter told my father it was a mistake of the risk control system and provide him a credit card. So the Credit Card Approval Prediction problem was chosen as my topic.

Basically, I used Python + Jupyter notebook to finish all the data analysis. Final report will be presented as a notebook. And the presentation will also be generated by Jupyter notebook (reveal.js.slides).

After data exploration, I found some problems in the dataset.

- Repeated ID
- Missing value
- Some variables are in text categories

So, I use pandas to clean the data. And I re-encode the whole dataset. All the fields after data cleaning are categorical variable, which is not linear. So I chose non-linear machine learning methods to solve the problem, decision tree and random forest.

The simple decision tree can't fit data well because the original data set is extremely imbalanced. So I used Synthetic Minority Over-sampling TEchnique (SMOTE) to oversample the dataset. Also, I used random forest based on bagging strategy to avoid overfitting or underfitting.

Usually, empirical parameters might not be the best choice for a model. So I used grid search cross validation to find the best combination of parameters. Here I chose 3 parameters for optimization, which are n_estimators, min_samples_leaf and max_depth. They will be tested in a range one by one and return a model with the highest accuracy.

Finally, the result of accuracy of random forest on training dataset reaches 0.88, and the test dataset also reaches 0.85, which is better than the decision tree.

Here are the key techniques I used for the experiment.

- Decision tree
- Random forest
- One-hot encoding
- SMOTE
- Grid Search Cross Validation