Ping An Case Study

January 10, 2019

1 Scenario:

- For Quantitative Investment, prediction is always important. If you can predict the condition of tomorrow based on today's data, it would be very helpful for investment adjustment. Our clients want to explore the probability of predicting the future performance based on the stock information given today. Thus, they have asked Ping An to work alongside them to identify the factors influencing the behavior, and to devise and implement a strategy to find out stocks which have better performance.

2 Your task:

- The stock exchange dataset of a certain day is given, you will seek factors (which have an effect on the performance of the next day) based on the dataset and given materials. Finding factors that are not mentioned in our materials is also highly encouraged! Then, you are supposed to apply these factors to predict what performance the stock of next day would be.
- We have scheduled a meeting in a week with our clients in which you will present our findings of the stock prediction issues (including the result you have got and the factors you recommended to use).
- You are in charge of building the model and of suggesting which stocks should have a better/worse performance as a result of the model's outcome.
- The first stage is to establish the viability of such a model. For training your model you are provided with a dataset which includes features of stock performance on a certain day. Of particular interest for the clients is how you frame the problem for training.
- Given that this is the first time the client is resorting to predictive modelling, it is beneficial to leverage descriptive statistics and visualization for extracting interesting insights from the provided data before diving into the model. Also, while it is not mandatory, you are encouraged to test multiple algorithms. If you do so it will helpful to describe the tested algorithms in a simple manner.
- Finally, using the trained model you shall 'score' and label stocks in the verification data set, and put them in descending order of the propensity of to have a good performance. You will submit this file with your presentation and your python files (in 'py' or 'ipynb' format), and your predictions will be scored with area under the ROC curve and Brier score which you shall be discussed during your presentation session.

3 Data fields and their description

- 'test_high_frequency_data' and 'train_high_frequency_data' files are given which show the details of stock prices changes in one day. For every stock, taking 15 minutes as a period, the open, close, high, low price, volume, and amount are provided, and you can calculate factors from these values. You shall name your own factors, and explain why you choose these factors as well as how you calculate them in your code or presentations. Also, you shall fill 'train_output' and 'test_output' files with these factors. After factors are chosen and calculated, you should apply these factors as features to build machine learning models for predicting the labels.
- The file named 'train_output' gives the label of every training stocks. The 'label' indicates how the performance of this stock is in the next day.
- 'test_output_template' gives an example of the finally result, you should submit a file like it.
- ATTENTION: 'stock_symbol' is the unique ID of the stock, it is formatted by 'str' type, so taking it as a number, or ignoring the front '0's is not acceptable. 'label' is an 'int' number, submitting it as a float number is also unacceptable.
- Table 1 describes all the data fields which are found in the data.

Table 1: Data fields and their description

Field name	Type	Description
stock_symbol	str	unique ID of each stock
date	object	date
time	object	end time of the stock in a period
open	float64	open price of the stock in a period
close	float64	close price of the stock in a period
high	float64	the highest price of the stock in a period
low	float64	the lowest price of the stock in a period
volume	float64	the total volume in a period
amount	float64	the total amount in a period
label	int64	shows the stock performance of the next day