Suvedei Soyol-Erdene
MTH4330
Professor Giulio Trigila
18 DEC 2024

Final Project:
Replicating the machine learning related paper " The Random Forest Model for Analyzing and Forecasting the US Stock Market in the Context of Smart Finance"

Overall, the paper illustrates the performance of predicting whether stock prices will go up or down in 3 time intervals 30, 60, 90 days. The paper introduced some technical indicators for features and implemented a random forest model for classification. In addition, there were some methods and steps that were not clearly stated or not stated at all.

Firstly, we were assigned to complete the paper by filling in the missing details for the paper. In 3.1 Data collection and preprocessing section of Methodology, Exponential smoothing method is used to remove the noise and reveal actual patterns in the historical data using following function:

$$S_0 = Y_0,$$
$$for\ t > 0,\ S_t = a * Y_t + (1 - a) * S_{t-1}$$

First issue was that alpha was not given but intuitively it is clear that alpha should not be less than 0.5, otherwise the new price would be totally different from smoothing price.

Another improvement could be the more clarification of why features given in table 1 are chosen.

RSI:

Relative Strength Index. It is measured based on average gain and average gain in a certain period of time. The number of days to calculate average loss and gain is not also mentioned. Industrial standard is 14 days.

$$RSI = 100 - (100 \div (1 + average\ gain / average\ loss))$$

Stochastic Oscillator:

Is a momentum indicator used in technical analysis based on closing time, lowest price and highest price in a certain time frame which is also not given in the paper. Assuming industry standard 2 weeks.

$$\%K = 100 \times (C - L_{14})/(H_{14} - L_{14})$$

where C is the recent closing price, L is lowest price, H is highest price in timeframe

William %R:

It measures the level of closing price relative to the highest high for a given look-back period, typically 14 days. It uses the same variable as Stochastic Oscillator.

$$\%R = (H_n - C) / (H_n - L_n) \times (-100)$$

where C is the recent closing price, L is lowest price, H is highest price in timeframe

MACD:

The Moving Average Convergence Divergence is also a trend-following indicator used in technical analysis. It is calculated based on the difference between the exponential moving average of certain two time intervals.

$$MACD = EMA_{12} - EMA_{26}$$

Price Rate of Change:

It measures the percentage change in price between current price and a certain number of periods ago.

$$ROC = ((Current\ Price - Price\ n\ periods\ ago) / Price\ n\ Periods\ ago) \times 100$$

On Balance Volume:

It is an indicator that measures buying and selling pressure as a cumulative indicator.

If the closing price is higher than the previous close then:

$$OBV = Previous\ OBV + Current\ Volume$$

If the closing price is lower than the previous close, then:

$$OBV = Previous\ OBV - Current\ Volume$$

In the prediction section, the Random forest model was implemented based on the features but there was no clear guide for how to separate the data into testing and training data sets.

Based on table 2. OOB error specific result, the sample size shows about 90 percent of the data of 7000. Therefore it is concluded that 90 percent of data was used for training.

What is the OOB error?

OOB stands for Out-of-Bag. The Out of Bag error is a method that estimates prediction error for machine learning models, particularly for bootstrap aggregating techniques such as Random Forest.

It first generates a new sample from the original data by making a subset with replacement and calculates the error based on the data that is not trained on the subset.

But for Time series, There are key considerations. Most important one is Temporal Dependencies : Time series data has an inherent order, the values are often dependent on previous values. That importantly applies to stock price prediction.

ROC:

ROC stands for Receiving Operating Characteristic curve that illustrates the performance of binary classification models considering True Positive Rate and False Positive rate. Ideal result of a good performing model would be higher True Positive Rate while keeping False Positive Rate.

Second part: Replicating the graphs and tables

For smoothing alpha:

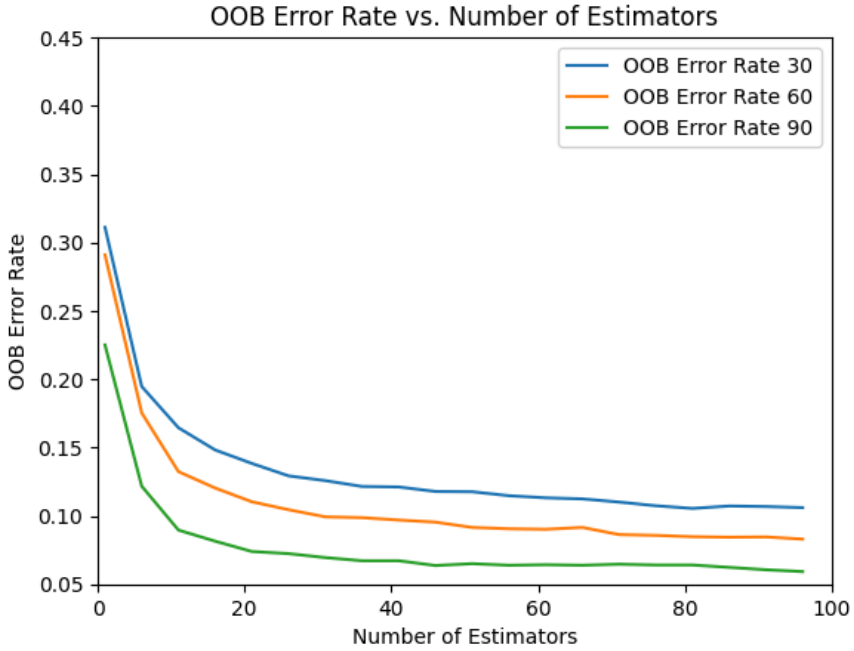
$$S_0 = Y_0,$$

$$for\ t > 0, S_t = a * Y_t + (1 - a) * S_{t-1}$$

I set alpha = 0.8 and 0.9, but did not show that much difference in performance

Methods I used to testify the result that were show in the paper:

Fig 2: Linear Separable test result (named same as the paper)



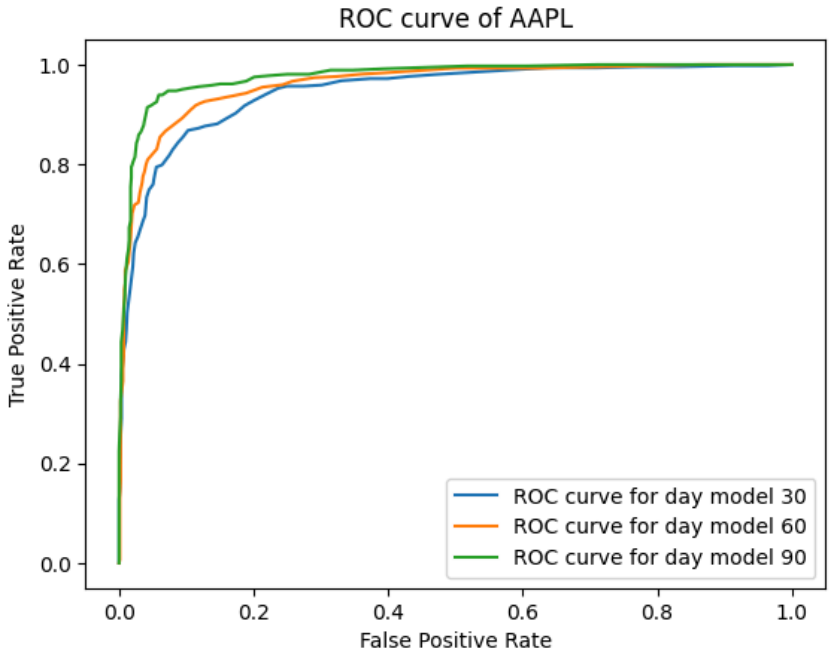
Comment: I used bagging method without considering the data is Time Series, and get similar result as Fig 2 in the paper

Table 2. OOB error Specific result

Trading_Period	Num_Trees	Sample_Size	OOB_Error
30	5	6219	0.217559
30	25	6219	0.133623
30	45	6219	0.120759
30	65	6219	0.113362
60	5	6219	0.180576
60	25	6219	0.103875
60	45	6219	0.095353
60	65	6219	0.093102
90	5	6219	0.124779
90	25	6219	0.069947
90	45	6219	0.062550
90	65	6219	0.059334

Illustrated some of the results from figure 2 based on the number of Estimators.
It turned out to have identically improved performance results as the paper shown in Table 2.
Even, replicated Table shows better performing result

Fig 4. ROC curve if Apple without considering it as Time series.



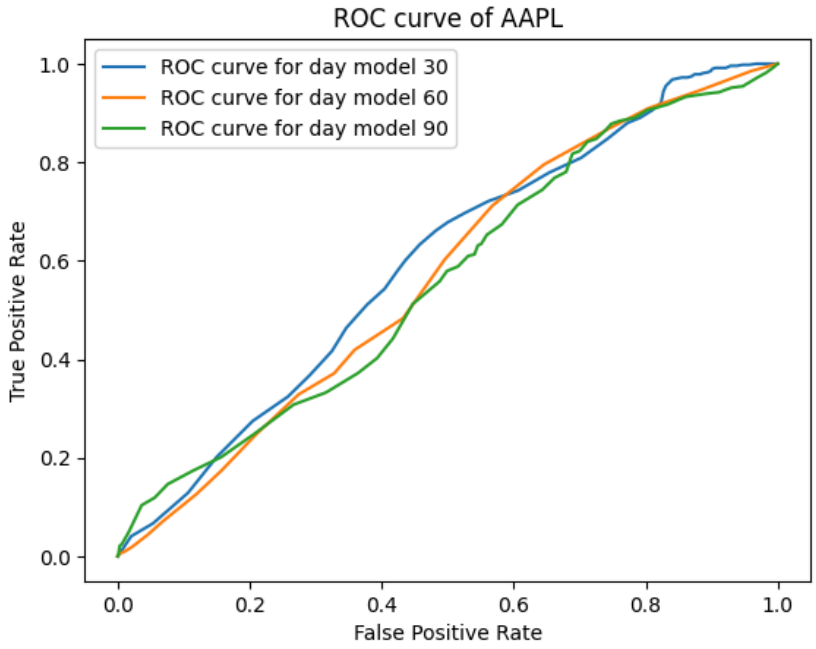
Comment: For ROC curve for AAPL, fig 4 is correctly replicated but using a randomly selecting method without considering the fact of time series.

Part 3: Analyzing the Method

From carefully studying the features, it was common that most features are engineered based on price difference and volume prices. Even though they illustrate different technical methods the variables hidden behind the function are the same price differences and volume of stock movement.

Here is the plotting same graph for Fig 4 but considering the better validation method regarding the time series characteristic of stock price prediction.

Fig 4. ROC curve of Apple and GM



In this figure, I considered the effect of the time series dataframe and trained data for the initial 80 percent of the data and tested the performance on the rest of the data (20%).

Comment: It turned out the Random Forest model performs poorly on the test.

It means **Fig 4** showing this ROC result from the paper is the test result of the model that does not consider Temporal Dependencies. Which made me suspicious that the model that tested on the paper does not consider the time series character of the data.

In conclusion, although in the paper, Random Forest method showed strong performance compared to other classification methods but without considering its major goal of predicting future price, the method fails because of not considering the time series qualities.

The new ROC curve figure based on a better validation method denies the performance of the method in the Paper.

Further Improvement for the paper would be introducing features that cover different aspects of data not just price movement and volume change. Secondly, implementing better validation method to assess performance of the model.