Predicting Daily Adjusted Close Stock Prices (SPY) using Time-series and XGBoost methods
Bao Nguyen

Domain background
Predicting stock prices has been a long-term interest in finance. For example, searching the phrase 'predict stock prices' on Google Scholar returns 1,020,000 results. There have been many machine learning and artificial intelligence methods used to predict stock prices. Wang, YF (2002) develops a fuzzy grey prediction to predict a large volume of stock prices in real-time. Gidófalvi, G (2001) uses a naïve Bayesian text classifier and leverages news articles to predict stock price movements. Roondiwala, M, Patel, H, and Varma, S (2015) predict stock prices using Long short-term memory (LSTM). Although predicting stock prices to match the market movements exactly might be impossible, attempts to predict stock prices are still valuable to understand and acknowledge the challenges.

Studying Economics motivates me to do this project. I have been using time-series and econometrics methods during my study and research in economics-related topics; however, applying those methods into predicting stock prices carry a lot of real-world applications.

Problem statement
I will build a stock price predictor that gives an estimation of 7-day stock prices for the S&P 500 Index (SPY). The dependent variable is the adjusted close prices.

Datasets and inputs
I used the yahoofinancials package to gather SPY stock prices from Yahoo Finance. The inputs are adjusted close prices of SPY with the period from 2019-01-01 to 2019-12-31. Stock prices are expected to be very noisy. Here are the first five lines of the data frame.

	SPY
2019-01-02	242.056915
2019-01-03	236.280746
2019-01-04	244.195160
2019-01-07	246.120560
2019-01-08	248.432953

This is the summary statistics of my dataset:

```
count    251.000000
mean     283.615941
std       15.948296
min      236.280746
25%      273.491623
50%      283.987000
75%      293.155945
max      318.371063
```

Solution Statement

I will use XGBoost (stands for eXtreme Gradient Boosting) to predict stock prices. According to Basak et al. (2018), given the nature of boosting (combining the results of various weak predictors to make a strong prediction), XGBoost approximates regressors to the training samples and finds the best split of the aggregate of the regressor functions.

Benchmark Model

I use a simple moving average as a benchmark model. A moving average is a technique to smooth out time-series data to reduce the "noise", taking the average of a certain number of previous periods to come up with a "moving average" for a given period. I used 7 lags for my benchmark model. The model uses the previous 7-day values to predict the value of the following day. This is the graph representing the actual values and the predicted values.

Evaluation Metrics

The evaluation metric is RMSE (Root mean squared error), which measures the errors between the stock's actual prices and the predicted adjusted close prices. Taking the square of the errors avoids the errors from underestimation and overestimation cancel out each other. The smaller the RMSE is, the better the prediction would be.

The RMSE value of the 7-day SMA is 2.76. While this is considered not a good value, this provides some benchmarking value so that we can compare the RMSE results of other models with this benchmarking RMSE.

Project Design

- Add features: Generate timestamp columns for the month, week, and year to incorporate trends and seasonalities. Find the appropriate number of lag values.
- Determine the train and validation dataset: exclude n number of ending values as validation data set. The remaining data points belong to the test dataset.
- Normalization: use mean and standard deviation values to normalize the variables.
- Fit the model.
- Predict the model and find RMSE value.

References

- Gidofalvi, Gyozo. Using News Articles to Predict Stock Price Movements. Department of Computer Science and Engineering, University of California, San Diego. 2001.
- Predicting stock price using fuzzy grey prediction system Y.-F. Wang Department of Information Management, Chang Gung Institute of Nursing, Kwei-Shan, Tao-Yuan, Taiwan
- Roondiwala, M., Patel, H. and Varma, S., 2020. Predicting Stock Prices Using LSTM. International Journal of Science and Research (IJSR), 6(4), pp.1754-1756.