Capstone 1 Proposal for Springboard Data Science Career track

Version: 1.0

Author: Jiaqi Xu

Coach: Kevin Glynn

Date: 05/12/2020

Forecasting S&P 500 index with Composite’s Daily Price Using Machine Learning Models

Background

S&P 500 is one of the most commonly followed equity indices which measures the stock performance of largest 500 companies listed on stock exchanges in US. The average annual total return of the index, including [dividends](https://en.wikipedia.org/wiki/Dividend), since inception in 1926 has been 9.8% (Wikipedia). From Berkshire Hathaway’s annual meeting 2020, one of the most famous investors Warren Buffet indicated “In my view, for most people, the best thing is to do is owning the S&P 500 index fund.” The index is also one of the factors to predict the direction of the economy. Because of above reasons, forecasting the movement of S&P 500 index is a long-time attractive topic.

Literature Review

Liu et at. (2016) implemented logistic regression, gaussian discriminant analysis (GDA), naïve Bayes and support vector machine (SVM) in forecasting S&P500 index and concludes that a SVM model with a Roseinadial basis function (RBF) kernel can achieve an accuracy rate of 62.51% for the future market trend of the S&P500 index. Niaki and Houseinzade (2013) used artificial neural network (ANN) to detect the statistically significant factors among 27 potential financial and economic variables which affect the movement of S&P500. However, the above researchers used S&P 500 index itself by data transformation or relative indices such as oil price, currency rates or other markets indices to achieve the prediction. Krauss (2017) developed a statistical arbitrage strategy based on deep neural networks (DNN), gradient-boosted trees (GBT), random forests (RAF), and their simple ensample and deployed it on the S&P 500 constituents. Lee and Kang (2020) trained the target Neural Networks (NNs) on the data of S&P 500 companies and Russel 3000 Index constituent companies to predict the future prices of the S&P 500 index. Krauss (2017) and Lee and Kang (2020)’s paper are close to my idea.

Goal

This project will compare different machine learning methods in forecasting the S&P 500 index quantitatively and qualitatively. Besides prediction accuracy, if time is available, a simple trading strategy (buy and hold) will be used to test the return of the investment by our prediction method.

Data

The dataset I will use for this project:

* Download historical S&P500 components on Wharton Research Data Services ([WRDS](https://wrds-www.wharton.upenn.edu/)) and [Wikipedia](https://en.wikipedia.org/wiki/List_of_S%26P_500_companies), clean them and get a mapping table indicates which stock will be included in S&P500 at the end of the specific time.
* Based on the components from above resources, I will use Python yahoo finance module [yahoo-finance-1.0.4](https://pypi.org/project/yahoo-finance/) to download daily data of all components and S&P 500 index from 2006/01/01 to 2020/01/01 (14 years data). Since yahoo finance will not keep the delisted stock’s information, several delisted stock’s data in historical S&P 500 will not be available. This will not affect the prediction too much because most of the stock’s information will be arrived.
* Other financial data such as currency rate, commodity prices, stock indices on other markets will also be considered as relative features in prediction process.

Approach

The cleaned dataset will be split into training, validation and testing parts. I will use daily price of components of S&P 500 as features to forecast S&P 500 index by quantitative and qualitative process:

* Quantitative
* Directly use the return of S&P 500, try 1,2,3…, n lags of features and make the corresponding 1,2,3…, n steps prediction of S&P 500 by following (supervised) machine learning methods:
* Linear Regression (by l1 (LASSO) and l2 (ridge-regression) adjusted
* Neural Network (ANN, CNN, LSTM or others)
* Reinforcement Learning (Q-learning or Actor-Critic method)
* Qualitative
* Classify the return of S&P 500 to different categories, such as:

|  |  |
| --- | --- |
| Return Range | Category |
| (-100.0%, -3.5%] | 0 |
| (-3.5%, -2.5%] | 1 |
| (-2.5%, -0.5%] | 2 |
| (-0.5%, 0%] | 3 |
| (0%, 0.5%] | 4 |
| (0.5%, 2.5%] | 5 |
| (2.5%, 3.5%] | 6 |
| (3.5%, ) | 7 |

* Try 1,2,3…, n lags of features and make the corresponding 1,2,3…, n steps prediction of S&P 500 by following (supervised) machine learning methods:
* Naïve Bayes
* Support Vector Machine
* Gradient-Boosted Trees
* Random Forests
* Decision Tree

Unsupervised methods such as k-nearest neighbor (KNN) and k-means will also be considered by not be took too much points on. I may select several of above methods by the limited time or will try all above methods if the research process is efficient.

Deliverables

* Jupyter notebook with relative codes on Github
* Final Report includes data acquisition, data transformation, model implementation and analysis, results comparation and conclusions
* Presentation Slides

Reference

Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, *259*(2), 689-702.

Lee, J., & Kang, J. (2020). Effectively training neural networks for stock index prediction: Predicting the S&P 500 index without using its index data. *Plos one*, *15*(4), e0230635.

Liu, C., Wang, J., Xiao, D., & Liang, Q. (2016). Forecasting s&p 500 stock index using statistical learning models. *Open journal of statistics*, *6*(6), 1067-1075.

Niaki, S. T. A., & Hoseinzade, S. (2013). Forecasting S&P 500 index using artificial neural networks and design of experiments. *Journal of Industrial Engineering International*, *9*(1), 1.