Gordon Lu

STAT 1223

Professor Nelson

14 March 2021

STAT 1223: Assignment 2

Contrary to centuries past, the world in the 21st century is growing steadily whether it be in terms of technological innovations, or developing new ways to fight viruses like COVID-19, it is beyond a shadow of a doubt that the world is competitive. With so much going on in the world, buzzwords like, “cryptocurrency”, “money”, “wealth” meet the eye well. Typically, with 9-5 jobs occupying most days, finding some leisure time that brings in a decent amount of money is hard to come by. Thus, the idea of “making easy money” by investing in the stock market becomes an attraction option. . In high-frequency trading firms, many complex algorithms are designed to optimally buy and sell stocks. For an average Joe, time is a sparse resource and spending hours upon hours on end trying to make profit from investing in the stock market while working is stressful. However, blindly investing into big name stocks is also dangerous.

Recent studies have shown that predicting stock prices is no stroll in the park to get done. With the number of confounding variables could exist, studies such as one written by Salm Lahmiri, looks to use a linear time series paired with Support Vector Regression and Singular Spectrum to predict the price of a stock. On the other hand, other approaches such as articles written by Seungwoo Jeon, Bonghee Hong and Victor Change as well as an article written by James Vanstone, Adrian Gepp and Geoff Harris involve stock prediction hinging upon using a neural network paired along with a linear times series. What is evident about each of the respective studies is the overarching goal to predict how well a model can predict stock prices. Holistically, at the base level, the response variable that I wish to learn more about is “buy or sell stock”, and how this is correlated with a stock’s previous price and other factors. What this paper seeks to do is not to describe trends in the stock market, but rather, given a certain stock, what determines how well a stock performs, and when should someone sell a stock versus buying the stock. One might wonder, “Why embark on trying to do something that seems to be a challenge to those who have tried?” I would refute and note that, I am doing this for the betterment of those, like me, who do not have the luxury of time, and have a genuine interest in investing and finance and learn how to predict time series data and potentially make a couple bucks. One way in which this paper may be appealing is the idea of having a computer decide whether to buy a stock or not, rather than spending hours on end, researching then deciding to buy or sell a stock. What I seek to learn is what variables influence the price of a stock, and given such variables, put together in a regression model, how well do said variables influence buying or selling decisions. If the result of overall regression turns out to be successful in predicting buying or selling, people would be able to boot up their computers, run a program, and watch their pockets grow (at some rate) and enjoy the show. It’s a win-win, both the general population would get to enjoy their wealth increase, and I would learn a great deal more about how stocks and time series work more closely.

I seek to use the Yahoo Finance API in order to collect data regarding stocks. In the Yahoo Finance API, each stock records the current listing for its price, as well as its historic data and other logistics such as the PE ratio, beta value and volume. The amount of data collected for a given stock from the Yahoo Finance API will vary depending on how frequent the price is updated as well as the date it was listed on the stock market. For example, a stock like Ford would have more observations than a stock like Apple, simply due to the fact that Ford has been around longer. For a given stock, I intend to use a small slice of the data, as with so many potential observations from when a stock is first listed, selecting data from a single year to the current date will be sufficient. As noted earlier, the observations will differ from stock to stock. I anticipate around 300-400 observations, which will be more than enough to conduct statistical tests.

In utilizing the data from the Yahoo Finance API, it is incredibly difficult to influence the price of a stock, and with a variety of other sources such as the TD Ameritrade API reporting similar numbers, to say that a stock price on Yahoo Finance does not reflect what it actually is would be similar to saying that billions of users are being lied to. With this, I am confident in saying that if Yahoo Finance API did use data that is influenced by external sources such as the government inflating the prices, then Yahoo Finance would not be as popular as it is now. With billions of users investing in the stock markets, and utilizing Yahoo Finance, along with other tools, it is incredibly difficult to say that the data that Yahoo Finance provides is incorrect, and biased. On the issue on whether the dataset is clean, it is beyond a shadow of a doubt that the data that the Yahoo Finance API presents regarding a stock is clean. It consistently updates the price of a stock with up-to-date information, and the only case in which a stock would be missing values would be if the stock were on the market, but then taken down, or simply isn’t on the stock market.

Overall, consolidating whether or not to buy a stock into a single response is the goal. The response, rather than being what the price of a stock could be, I seek to instead use a binary variable that will determine whether buying a stock at a certain day is worth the investment. I will call the response, “buy_stock”.

In an attempt to account for the potential influences on “buy_stock”, I will consider the following predictor variables:

- roe: Quantitative variable describing the return on equity for a given stock.
- stock_price: Quantitative variable determined by running the ARIMA model on the data, and will yield the optimal times to include for the price that will minimize volatility. The times will be based on days, so stock_price_1 would be yesterday’s price.
- pe: Quantitative variable determining the price-to-earnings ratio for a given stock.
- count_twitter_news: Quantitative variable determining the frequency of tweets regarding the stock.

References

Jeon, S., Hong, B., & Chang, V. (2018). Pattern graph tracking-based stock price prediction using big data. *Future Generation Computer Systems*, 80, 171–187. https://doi.org/10.1016/j.future.2017.02.010

Lahmiri, S. (2018). Minute-ahead stock price forecasting based on singular spectrum analysis and support vector regression. *Applied Mathematics and Computation*, 320, 444–451. https://doi.org/10.1016/j.amc.2017.09.049

Vanstone, B., Gepp, A., & Harris, G. (2019). Do news and sentiment play a role in stock price prediction*? Applied Intelligence (Dordrecht, Netherlands)*, 49(11), 3815–3820. https://doi.org/10.1007/s10489-019-01458-9