ECO421H1 F: Special Topics in Economics
Macroeconomic Finance (with machine learning applications)

2021 Fall

Professor Marlene Koffi

Technology Stock Price Predictions with Machine Learning Techniques

Wednesday, December 8, 2021

Jacob Yoke Hong Si, siyoke, jacobyh.si@mail.utoronto.ca

Yuhang Li, liyuhan9, yuhang.li@mail.utoronto.ca

Sicheng Wei, weisiche, sicheng.wei@mail.utoronto.ca

I. Introduction

Motivation and Research Questions

Stock price prediction has always been one of the key topics in finance. However, given fluctuations in the macroeconomic market and firm performance, a precise prediction is generally hard to be achieved. This research paper is motivated to examine how machine learning can be applied and contribute to stock price prediction given sufficient past information and strong upward movements. During the recession of the pandemic, stocks of leading technology firms have significantly grown in value despite the overall slowed growth and slack in the economy; with daily stock price data regarding publicly traded technology firms of high market value, we are motivated to utilize the past patterns of these tech companies stock prices and perform predictions for next periods.

As a general background, there are three schools of thought regarding stock movement prediction: The first school of thought believes that stock prices are unpredictable in nature given a highly competitive and transparent stock market, alluding to the Efficient Market Hypothesis (EMH). The second believes that firm level financial information and macroeconomic conditions affect the intrinsic value of a firm and thus can be used to forecast changes in stock returns in the long run as stock price converge to real value of a firm. The third believes that technical indicators incorporating momentum of past stock movement can be used to predict future prices in the short run. Our research is mainly based on the third school of thought, in that we use trends and volatility of past stock price movement in the context of Covid-19 to generate daily predictions for next periods. To borrow the idea of the general economy affecting stock returns from the second school of thought, as well as to capture some of the risks specific to covid-19, we incorporate a covid-19 stringency index and a VIX stock uncertainty index into our set of predictors. Naturally, our research involves a supervised learning experience that potentially captures both linear and non-linear behavior of stock prices.

Furthermore, we investigate how machine learning techniques assist the prediction of stock prices both qualitatively (i.e., direction estimation on whether price goes up or down) and quantitatively (i.e., the price level estimation). In machine learning jargon, qualitative and quantitative predictions are referred to as classification and regression tasks respectively.

Contribution

Within modern literature, relatively fewer analyses are conducted for data on daily and individual stocks. Given that Covid-19 has been a relatively new event, the analysis of all available daily data allows machine learning techniques to incorporate some of the most important information and contributes to the relevancy of stock price prediction in a more recent context. Such prediction is also more relevant to individual investors since they do not generally invest in a highly diversified portfolio of stocks. In addition, as our research applies a diversified set of parametric and nonparametric methods which stands on the two extremes of model flexibility (e.g., Neural Networks and Linear Regression), we can make comparisons and ultimately make suggestions on model choice for the two tasks. We also provide choices for two types of investors in terms of the best models they can utilize: The first type determines whether there is a positive or negative growth in the future stock price whereas the second type aims to predict the future trend of the stocks.

Literature Review

In their paper *The use of data mining and neural networks for forecasting stock market returns*, Enke and Thawornwong (2005) used various neural network models with economic indicators to capture non-linear movement in stock prices, and we use Multi Layer Perceptron, a member of the neural network family, along with uncertainty indexes to accomplish our task; the authors also incorporated trends and magnitude of variations to further account for the volatile nature of stock prices, which is similar to our methodology of using trend and volatility indicators. Given key predictions from the level estimates and the classification of signs of excess stock returns, the authors simulated a long-term trading strategy that alternates between stock and safe assets (treasury bills). They discovered that with the prediction from classification tasks, the investor gains the highest period-end return. Given the daily nature and short span of our testing data, a long-term investment in either stock market or treasury bills is not possible; however, we might still be able to give advice to potentially different types of investors with different tastes of information.

In the paper *Neural Networks for Technical Analysis: a Study on KLCI*, Yao, Tan, and Poh (1998), researchers identified several important technical indicators for trend and magnitude of stock price movements, namely the Moving Average (MA), Relative Strength Index (RSI), Momentum (M) and Stochastics (%K). For the ease of understanding and the popularity of use, we select MA and RSI as indicators for trends and magnitude of past stock price changes in our work.

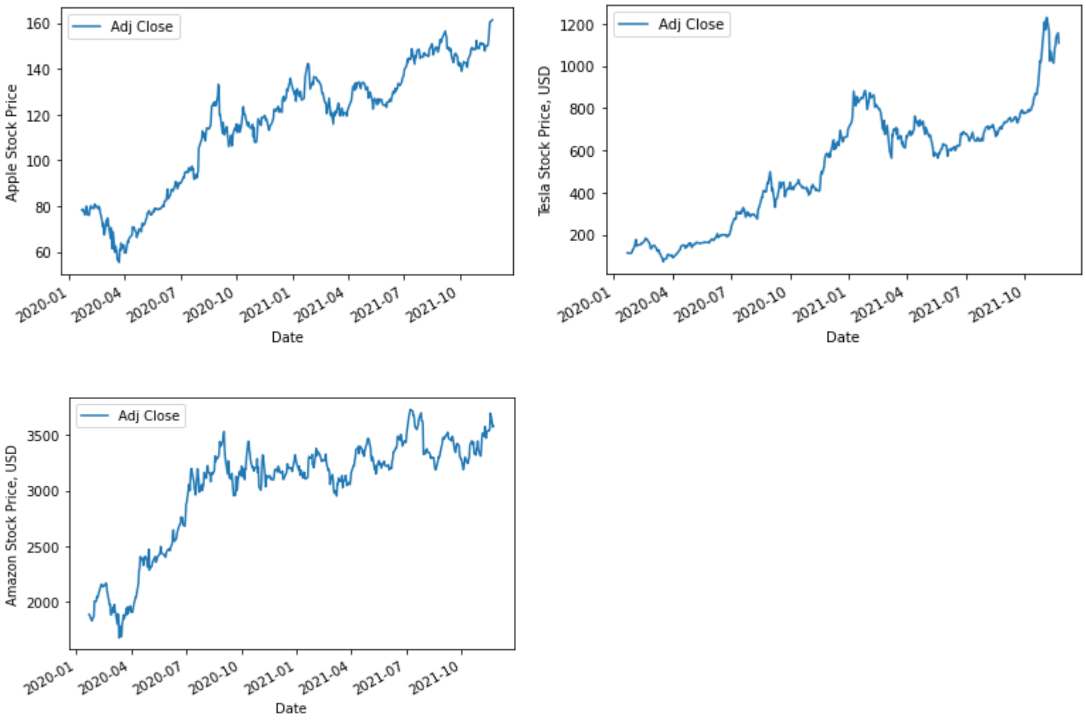
II. Data and Methodology

Data & Predictor choice

We obtained information on three stocks, Apple (APPL), Tesla (TSLA) and Amazon (AMZN) from Yahoo Finance for the period from 2020/01/21 to 2021/11/23.

All three stocks have strong upward movement, with amazon price exhibiting a rather non-linear trend and more occurrences of fluctuation. We take either percentage change in stock price (stock return on next trading day) as one of the predictors, for which the information regarding price change is effectively the same as incorporating both one period and two period lag of stock price; or directly use the past stock price.

Figure 1: Stock Movement, starting 2021/1/21:



To capture the overall trend as well as the zig-zagging movement along the trend, we apply two technical indicators, namely Relative Strength Index (RSI) and Exponential Moving Average (EMA）:

$$\mathrm{RSI} = 100 - \frac{100}{1 + (\text{Average Gain in past})/(\text{Average Loss in past})}$$

$$\mathrm{EMA_t} = \text{Adj closing price}_\mathrm{t} \times \alpha + \mathrm{EMA_{t-1}} \times (1 - \alpha)$$

where:

$$\alpha = \frac{2}{window\ length + 1}$$

RSI measures the relative strength of stock's upward movement and downward movement, which effectively translates into magnitude of price changes in the recent past. If we observe on average a high rate of positive change, RSI will be closer to 100. EMA is a moving average series that assigns higher weight to the more recent prices, which can be translated into a trend indicator that incorporates more recent price information. We decided to use a window length of 5 to be in line with the number of trading days per week. We obtained two measures of uncertainty, namely the VIX index obtained from CBOE and Covid-19 US stringency index from Oxford University for the same time period. VIX('VIX') is a widely known measure of 30-day ahead expected volatility of US stock market, and we use it to capture real time general stock market uncertainty that is present in our research; Covid-19 US stringency index('STRIN_US') is calculated based on nine indicators including workplace closures, travel bans, stay-at-home requirements, and other restrictions on public events and domestic movements. While some may suggest an uncertainty index derived from sentiment analysis of various media sources could better utilize the investor sentiment on certain tech stocks during Covid-19, different data sources (i.e., different media platforms we obtain user or critic comments from) could impact the results differently, and it is rather hard to disentangle the shocks of investor sentiment from that of others. Considering the completeness of stringency index data (maintained and updated daily) and given that the underlying government policy shock can affect our target firms' operation exogenously (e.g., transition of workplace and temporary shut-down of on-site business), we decide to use the stringency index in measuring uncertainty from the pandemic.

Models & Methodology

For level estimation (stock price prediction), we use the following general model form:

$$\mathrm{Y_{t+1}} = \mathrm{f}\,(\mathrm{Return_t}, \mathrm{VIX_t}, \mathrm{RSI_t}, \mathrm{EMA_t}, \mathrm{STRIN\_US_t}) + \varepsilon$$

where t is the time index and Y_t is the stock price in next period.

In linear regression, the equation can be explicitly specified as

$$\mathrm{Y_{t+1}} = \alpha_0 + \alpha_1 \mathrm{Return}_{\,t} + \alpha_2\, \mathrm{VIX_t} + \alpha_3\, \mathrm{RSI_t} + \alpha_4\, \mathrm{EMA_t} + \alpha_5\, \mathrm{STRIN\_US_t} + \varepsilon$$

We use in total four models for this task: MLP Regression, Random Forest Regression, Decision Tree Regression and Linear Regression (least flexibility).

For direction estimation task, we use the following general model form:

$$\mathrm{POS/NEG_{t+1}} = \mathrm{f}\,(\mathrm{Y_t}, \mathrm{VIX_t}, \mathrm{RSI_t}, \mathrm{EMA}_{\,t}, \mathrm{STRIN\_US_t}) + \varepsilon$$

Where POS/NEG is a binary variable indicating either upward (1) or downward movement (0) of stock price in the next period compared to the current period.

In logistic regression, it can be explicitly specified as

$$\mathrm{Pr}(\,\mathrm{POS/NEG_{t+1}} = 1) = \mathrm{g}\,(\mathrm{Y_t}, \mathrm{VIX_t}, \mathrm{RSI_t}, \mathrm{EMA_t}, \mathrm{STRIN\_US_t}\,)$$

Where g is the sigmoid function and error term is not included as we directly estimate the mean probability. We use past price of last period in this task as it produces better accuracy. In total we use four models: MLP classification, Random Forest Classification, Decision Tree Classification and Logistic Regression (least flexibility).

In terms of model flexibility, MLP Regressor and MLP Classifier have great tolerance for complex relationships: in each of its ‘hidden layer’, inputs are first transformed into weighted linear sum, then followed by non-linear activation function such as Rectified Linear Unit (ReLU) and Hyperbolic Tangent (TANH) to produce the output. The weights can change across different layers, allowing for more complex modelling. In our analysis, we utilize the default features (e.g., number of layers) of the MLP as well as ReLU as activation function.

To follow traditional practices in literature, we attempted to standardize our data before going into training and testing the model. However, as the standardization potentially changes the underlying distribution (data shape) of the predictors, the model performance deteriorates significantly for our level estimation while slightly improving for our direction estimation, so we decided to keep with the original data inputs for level estimation tasks in specific. In addition to the fundamental training and testing procedures, we utilize time series out of sample validation to further validate our result. Given our total 466 days of observation for each stock, we split our time series into five shares, and use a constant test window length of 93 with evolving training window length. In our first validation, the first 93 observations are used for training, and the next 93 observations are used for testing; in our last and fourth validation, all observations except for the last 93 are used for training, with the last 93 used for testing. We average the MSE and accuracy rate obtained from four validations to arrive at the final result for comparison of model performance.

III. Results

Table 1: Average Mean Squared Error for the regression prediction of AAPL, TSLA and AMZN stock price.

	Stock Ticker Symbols		
Average Mean Squared Error (USD^2)	AAPL	TSLA	AMZN
Multilayer Perceptron Regressor	91.0	3298.4	5741.9
Linear Regression	29.7	1370.3	10296.3
Decision Tree Regressor	299.8	32807.0	131148.1
Random Forest Regressor	256.8	31539.9	124692.1

Table 2: Average Accuracy Score for the classification prediction of AAPL, TSLA and AMZN.

	Stock Ticker Symbols		
Average Accuracy Score (in fractions)	AAPL	TSLA	AMZN
Multilayer Perceptron Classifier	0.58	0.51	0.53
Logistic Regression	0.52	0.54	0.55

Decision Tree Classifier	0.51	0.52	0.53
Random Forest Classifier	0.52	0.53	0.51

From Table 1, Linear Regression yielded the lowest average mean squared error (AMSE) at 29.7 and 1370.3 for both AAPL and TSLA respectively when compared to other models. The second-best model for AAPL and TSLA is MLP (Regressor) yielding an AMSE of 92.0 and 3298.4 respectively. Contrarily, for AMZN, the best model was MLP with an AMSE of 5741.9 and the second-best model happens to be the Linear Regression model yielding an AMSE of 10296.3. Therefore, our analysis seems to support the fact that the best two performing models for regression analysis are MLP and linear regression. This is expected since by design neural network models capture non-linearity of the fluctuations in our data, which coincide with literature. With regards to the rather linear trend for Apple and Tesla, the relative high performance of linear regression does not come as a surprise, and for Amazon that has a rather non-linear trend and many fluctuations, MLP performs relatively better.

In terms of our classification task, we determined that the logit model yielded the highest average accuracy score across four validations at 53.5% and 54.8% for both TSLA and AMZN respectively when compared to other models. Contrarily, for AAPL, the best model was MLP with an average accuracy score of 57.5%. Therefore, the consensus of our analysis agrees with the fact that the best two performing models for classification analysis are the Logit and MLP. Again, for Amazon stock that exhibit more non-linearity, neural network models seem to exploit past information better than other models, while logit model with more linear feature in its sigmoid function predicts better for stock prices with more linear trending behavior.

Figure 2: Predicted and Actual values of Adj Close of AAPL (left) and TSLA (right) using Linear Regression, first validation

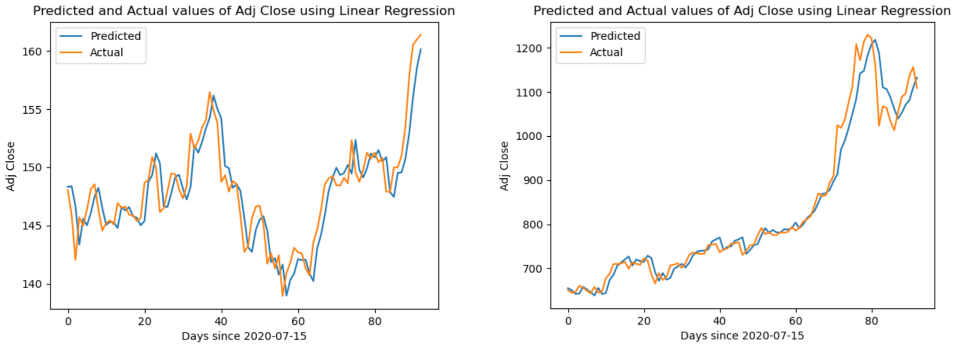
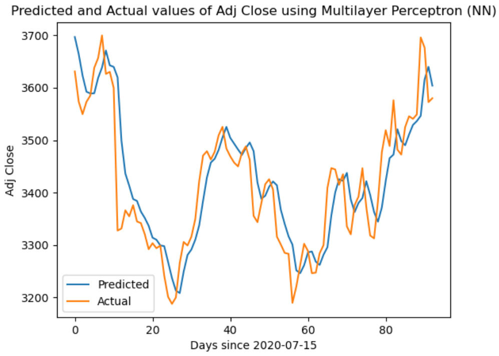


Figure 3: Predicted and Actual values of Adj Close of AMZN using MLP, first validation



In figures 2 and 3, we use the fourth validation set and plotted the predicted and actual values of AAPL, TSLA and AMZN using the models that provided the lowest AMSE (Linear Regression for AAPL and TSLA, MLP for AMZN), for the last 93 days in our data. In the plots, we observe that the predicted values tend to follow the trends of the actual predictions relatively well. While the predictions may not be able to map out every intricate detail of the actual trend such as sudden downturns or expansions as well as shocks, it is able to capture the general trend of the stock price. Finally, we look at the coefficients of Logistic Regression for fourth validation test sample for some information on relative importance of predictors. Since we did not scale our predictors for the level estimations task, the Linear Regression coefficients for Apple and Amazon could be misleading. For cases where MLP outcompetes other models, we do not have information on predictor importance due to the way neural networks changes its predictor importance (weights) in many of its hidden layers.

Table 3: Predictor Coefficients for Logit Model (Inputs Standardized)

	RSI_t	VIX_t	EMA_t	Return_t	$\mathrm{STRIN\_US}_t$
Logit_TSLA	-0.04	-0.31	0.09	-0.45	0.25
Logit_AMZN	-0.4	-0.22	-0.1	-0.35	0.29

From Table 3, one can find it hard to make definitive judgement. Nevertheless, it is safe to make the claim that past return might offer one with at least some predictive power, and one may also consider including other indicators into their analyses.

IV. Conclusions

Using the features we have selected, namely RSI, EMA, VIX index and Covid-19 stringency index for US, we constructed predictions on the Adjusted Closing Price and direction of stock price movement. From our findings, we managed to produce a relatively accurate set of predictions when compared to the actual values. Furthermore, with regards to classification, we obtained decent results with accuracy scores ranging between 50% and 60%. If Covid-19 were to persist, our strategy could be useful in the current context. We would advise type one investor with less information cost to stick with neural network models for stocks that have seen larger fluctuation as well as non-linear trends, and linear models to exploit information better for stocks with less fluctuations and a rather linear trend; For type two investors who favors only general stock movement information, We again advise them to use neural network models for predicting stocks with more fluctuation and non-linear trending behavior, and possibly logit model for stocks with rather linear past trends. Finally, judging from the Figure 2 and Figure 3, our predictions are still unable to account for relatively big shocks in stock prices, preventing us from achieving higher accuracy for both level estimation and direction prediction. To further improve our model fit and accuracy, we could potentially include more explanatory variables, as well as improving the design of our selected variables, for instance, constructing a more representative and firm-specific covid-related uncertainty index.

References

Enke, D., Thawornwong, S. (2005). *The use of data mining and neural networks for forecasting stock market returns*. Retrieved December 6, 2021.

Yao, J., Tan, C., Poh, H. (1998, October 7). *NEURAL NETWORKS FOR TECHNICAL ANALYSIS: A STUDY ON KLCI*. Retrieved December 6, 2021.