Predicting ETF prices using linear regression

Keqing Li*

St. Johnsbury Academy Jeju Jeju Si Jeju-Do Korea

*Corresponding author: s17012923@sjajeju.kr

Abstract. Machine learning has allowed computers to analyze data and make future predictions based on those dates. One of the most common and easiest to implement machine learning algorithms used to do this is simple linear regression. Simple linear regression finds trends in a data set by graphing a line that shows the relationship between two variables. This paper will show how Simple linear regression can predict future ETF prices by finding linear trends in two particular exchange-traded funds: Invesco QQQ and vanguard VGT, predict their value six months later using their five-year closing price in yahoo finance and compare their respective predicted growth rate.

Keywords: Machine learning; simple linear regression; exchange-traded funds.

1. Introduction

The Stock market is a place where investors can buy and sell bonds, options, funds, and shares of companies at varying prices, the prices changing over time, so the investors must be careful about when to buy or sell what to earn money. The problem is that there are items on the market with very similar trend lines. Invesco QQQ Trust (QQQ) and Vanguard Information Technology Index Fund (VGT) are two exchange-traded funds: combinations of stocks in the technology industry. Because of that, their portfolio has many overlaps, and they look remarkably similar. Investors encountering them may not know which to buy and choosing the wrong thing could lead to a higher opportunity cost. ETFs like these make ways to discover market trends and predict future ETF value more necessary. One method this problem can be solved is through machine learning. Machine learning allows artificial intelligence to find trends and patterns by analyzing data from the past and automatically learn and improve themselves. There are many machine learning algorithms, but simple linear regression is one of the most common and easy to implement. Simple linear regression shows a linear relationship between two variables. This model is not good at predicting short-term changes in prices because of the volatile and chaotic movement of ETF prices in short periods, but it is good at finding long-term trends.

Much research has been conducted about linear regression. M Umer Ghania, et al, (2019) successfully predicted the future stock price of amazon and apple using linear regression and improved the accuracy of their model by using a three-month moving average of their data and implementing exponential smoothing [1]. Vaishnavi Gururaj, et al, (2019) compared linear regression with support vector machines and provided pros and cons for both models [2]. Ishita Parmar, et al, (2018) predict the price of a stock using both regression and the LSTM model [3]. Lokasree B S (2021) produced a step-by-step guide to creating a simple linear regression [4]. E. Sreehari, et al, (2019) predicted future rainfall using simple linear regression and multiple linear regression [5]. Dastan Hussen Maulud, et al, (2020) compiled works on linear and polynomial regression by various researchers and compared their performance based on accuracy [6]. Harikrishnan R, et al, (2020) compiled various works on how machine learning can predict stock prices and did a comparative analysis [7]. Reza Gharoie Ahangar, et al, (2010) estimated stock prices in Tehran stock exchange using linear regression and artificial neural network methods and compared their results [8]. Alaa F. Sheta, et al, (2015) Used linear regression, support vector machine, and artificial neural networks to predict stock prices [9]. Wong Pik Har, et al, (2015) Found the relationship between ROA, ROE, and ROCE and stock return in plantation companies in Malaysia using simple linear regression [10].

Many papers are flawed because the predictions being made are often short-term. When using LSTM or multiple linear regression models, this offers greater accuracy. However, when it comes to the simple linear regression models, predicting the short term will be very inaccurate. In addition,

many papers do not make comparisons between the stocks studied, which could cause confusion as to which stock will generate more revenue. This paper differs from the previous research because this paper focuses more on ETFs rather than stocks. Although the two are somewhat similar, ETFs are less volatile than stocks because they allow investors to invest in many companies in a sector at once rather than particular companies, thus creating a diversified portfolio. Furthermore, this paper also compares the returns of the two ETFs discussed, QQQ and VGT, showing which one will have a greater rate of return and value in the future. This paper will show how to predict and compare future ETF prices using simple linear regression, assuming the price trends for those ETFs will not change, by finding out which among QQQ and VGT has a higher rate of return in six months and using the coefficient of determination to find the accuracy of our model.

The remainder of the paper is organized as follows. Section 2 explains the step taken to create the simple linear regression model for the two ETFs studied, QQQ and VGT. Furthermore, explain what dataset is used for the model; Section 3 shows the results of our model and the predictions made, including the estimated rate of return and ETF price in 6 months. Section 5 concludes the paper as well as providing limitations and possible improvements to this study.

2. Methodology

We obtained the historical data for QQQ and VGT, starting on February 21st, 2017, and ending on February 17th, 2022, almost five years in total, on Yahoo finance. We use the time, defined as the days after February 21st, 2017, and the closing price as variables. Using those variables, we can find how closing prices changed over time. The code and the graphs are made with python on Jupter notebook.

This paper aims to find and compare the price of QQQ and VGT and their perspective rate of return in the next six months if the trend they displayed during the last five years is to continue. We created a simple linear regression model; the first step is to split the data into test sets and training sets. We split them in a 4:1 ratio and randomized which data would be used as training sets, and which one would be used as test sets. The graphs below are scatter plots of the training sets and testing sets used in this research. Each point represents a dataset.

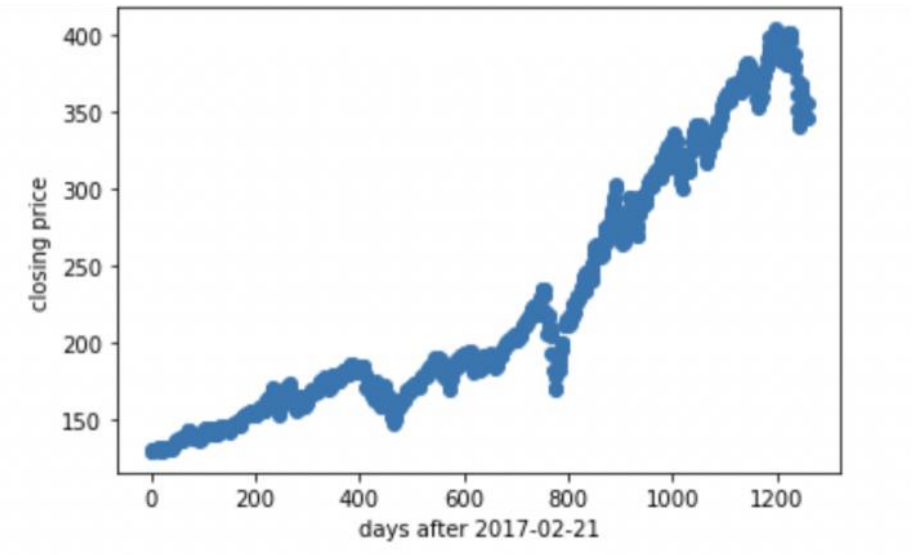

Figure 1. Illustration of training set for VGT.

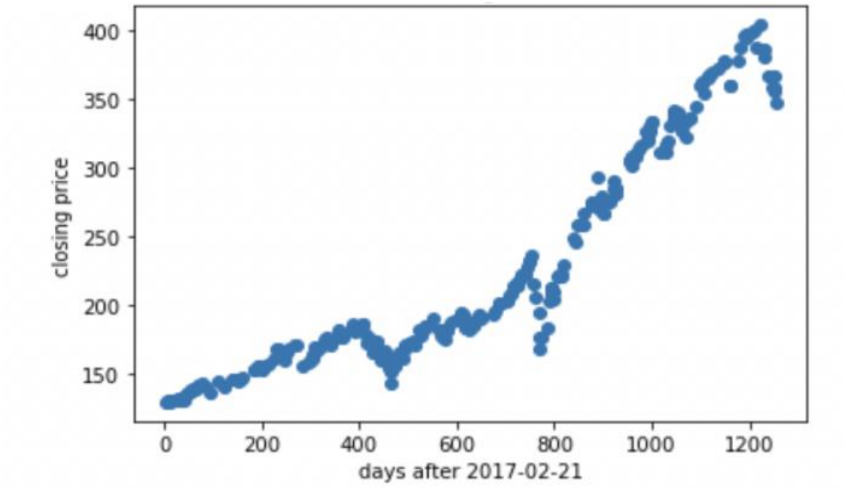

Figure 2. Illustration of test set for QQQ.

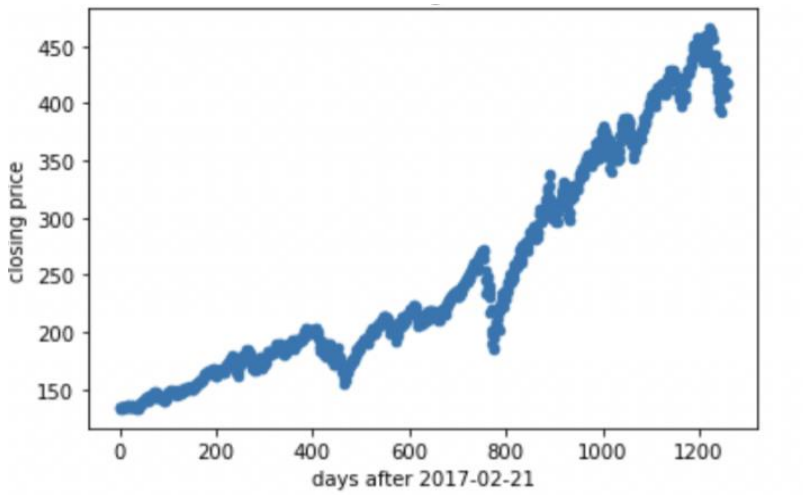

Figure 3. Illustration of training set for VGT.

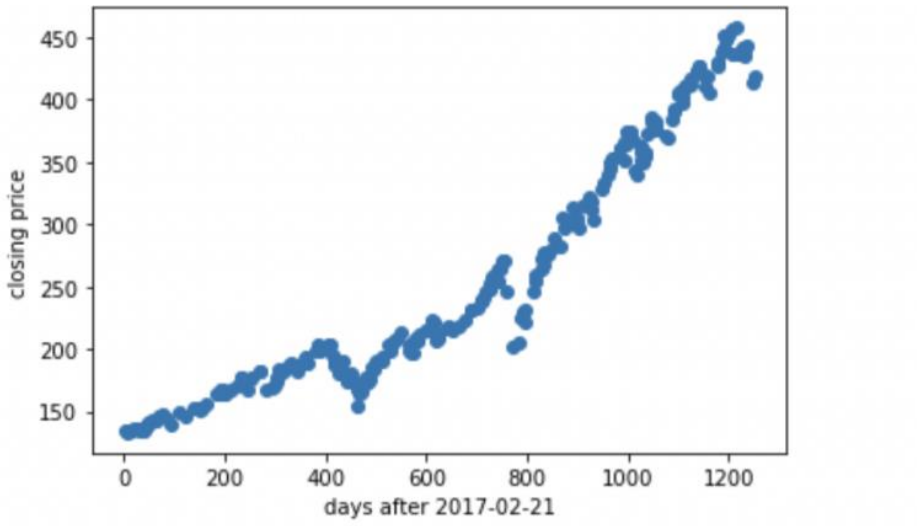

Figure 4. Illustration of test set for VGT.

The closing price, or the price at the end of each day, is used as the y axis. Days after the first day in the data, represented in the data imported as the index, are used as the x-axis. And the blue points each represent a data entry. We used the training set to fit the model of linear regression, which is modeled by:

$$y = mx + c \tag{1}$$

The y represents the closing price of the ETF, x represents the number of days after February 21st, 2017, the first day in the data set. c represents the y-intercept, or the closing price of the ETF at the beginning of the dataset, note that this variable is not the actual price of the ETF on February 21st, 2017, but a predicted price at that day made the algorithm based on the trend in the changes of all closing prices in the dataset. m represents the average change in price per day or the slope predicted by the model. We then extracted vital information from the model for calculation, namely the slope, the y-intercept, and the time and closing price at the last data point. Using them, we can find the price of the ETFs in 6 months, calculated by:

$$p = m(x + 180) + c \tag{2}$$

The meaning of m, x, and c is the same as above. p is the price of the stock 6 months from the present date, which in this study is February 17th, 2022. 180 is days within six months. based on the data provided to the model. After finding the predicted price in 6 months, we could find the rate of returns in 6 months, which is calculated by:

$$r = \frac{p-f}{f} \cdot 100\% \tag{3}$$

P has the same meaning as explained above. r is the rate of return after six months or the net gain of the investment over the six-month period expressed as a percentage. f is the closing price of the final datapoint in the data set, we can see this as the current price. After we find the rate of return, we can fit the testing sets to the coefficient of determination (R^2) to measure how well a linear regression model fits the data. If a R^2value is close to 0, there is little correlation, but if it is close to 1, there is a strong correlation. $R^2$is modeled by:

$$R^2 = 1 - \frac{RSS}{TSS} \tag{4}$$

The ESS or the explained sum of squares is the sum of squares of the difference between the predicted value and the mean value of a response variable. The TSS or the total sum of squares is the level of variation in the error of the regression calculated by the sum of squares plus the residual sum of squares.

3. Result

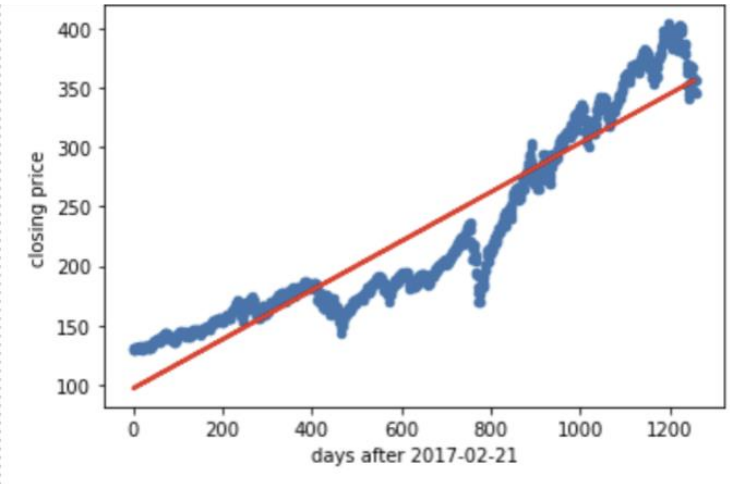

Figure 5. The linear regression model and predictions for QQQ.

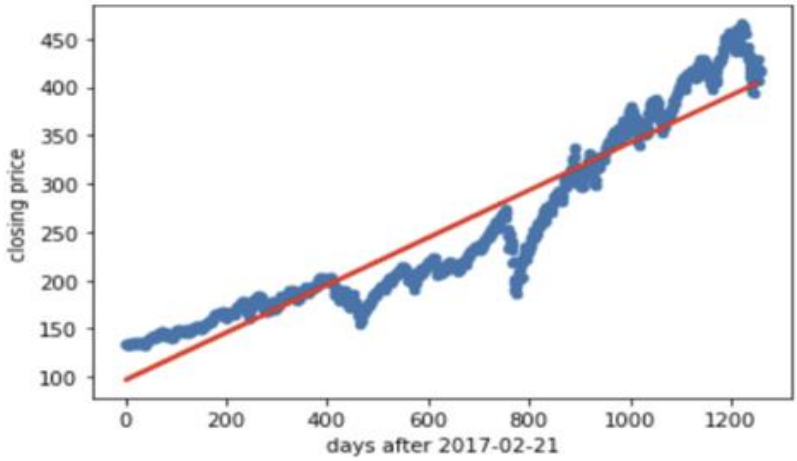

Figure 6. The linear regression model and predictions for VGT.

After implementing the steps outlined in the methodology section, we can get the following result. The blue scatter plot represents all the data for QQQ and VGT, with the y-axis being the closing price and the x-axis being the time. The linear function plotted on top of the scatterplot, shown as the red lines on in the graph, is the regression lines modeling the five-year trend of QQQ and VGT. We will use these lines to make predictions.

For QQQ, the y-intercept of the regression line is 96.93. The slope is 0.21. Plugging these numbers into the linear regression formula can give us the algebraic representation of the regression line. We add an additional six months or 180 days to the final date and use that for the x value. The result is 394.15 dollars, the predicted closing price six months from February 17th, 2022. This, in turn, is used along with the closing price on February 17th, 2022, to get the predicted rate of return for QQQ, which is 14.1 percent. The R^2score obtained for this regression line is 0.87. This is very close to one, meaning that the model fits the dataset for QQQ exceptionally well. From the information we obtained, we can see that QQQ is a sound investment in the long term with a healthy rate of return.

For VGT, the same processes and calculations used for QQQ are implemented again. The y-intercept of the regression line is 96.51. The slope is 0.25. We use the same x value as the one we used for QQQ since the length of time used is the same. The predicted price for six months from February 17th, 2022 is 416.84, and the rate of return is 7.93. The R^2score obtained for this regression line is 0.9. Once again, it is very close to one, which indicates that our prediction has few deviations from the actual dataset. From the predicted price and rate of return, we can see that VGT, much like QQQ, becomes more valuable over time, which makes it a good investment.

The linear regression model for both VGT and QQQ are upward sloping, they have similar slope and y-intercepts, but minor differences can be inferred based on the model. The slope of VGT is greater than that of QQQ, which means VGT has a higher growth rate. Its price will grow by a more significant margin over a set time period compared with QQQ. The rate of return for QQQ is higher than that of VGT, making it the better investment in a six-month period. VGT's R^2score is greater than QQQ's, meaning QQQ's model has a larger margin of error. Something surprising about our finding is that the model gives VGT a higher slope but also a lower rate of return in six months. However, if the slope is steeper, the rate of return should be as well. This is because the final closing price in the data sets for VGT is a bit higher than predicted, and the one for QQQ is a bit lower, meaning that, according to our model, the QQQ ETF is currently underperforming while VGT is over performing. This could explain why VGT has a lower rate of return while having a steeper slope. It also means that VGT could get a higher return if given more time.

4. Conclusion

In conclusion, we predicted and compared the future price and rate of return of QQQ and VGT with linear regression with data obtained in yahoo finance. The result shows that QQQ will have a

greater rate of return in six months, but VGT has a higher growth rate. This means that although QQQ will have to earn investors more profit in six months, investing in VGT is likely to be more profitable over a longer period of time. By using the data to train an algorithm to produce a linear regression model, this paper also demonstrated how machine learning could be used to see trends in the stock market and give investors a greater chance of profit. Although the two ETFs studied, QQQ and VGT are not very volatile, the simple linear regression model developed during this study is nevertheless very accurate.

Implementing machine learning in investing will significantly improve the decision-making ability of investors. It allows them to be better at predicting the future return on their investments and decide if the investment should be made. Simple linear regression, in particular, can help investors see price trends and notice small differences in figures that otherwise look very similar.

Linear regression still has many weaknesses. For example, it assumes that past trends reflect the future and predicts the future based on that assumption. While past performance is a good predictor of future trends, it is not always accurate. There could always be new events or developments that make the trends change direction. Another weakness is that if this model is used on highly volatile stocks or ETFs that do not have a clear trend, its accuracy will decrease, and it will no longer be helpful when trying to figure out the future price of an ETF. This model is also not good for short-term prediction. Differing from the LSTM model or multiple linear regression, the simple linear regression model is a straight line that does not change its slope over time. The stock market is highly volatile in the short term, the price of funds and stocks changes constantly, and if an impactful event has occurred in the market, prices often take a long time to stabilize. All of these make it almost impossible for prices to follow their past trend precisely in the short term and predicting them accurately using simple linear regression is impossible. Because of all the weaknesses and benefits of the simple linear regression model, more studies about how machine learning, both using simple linear regression and other models, can be used to predict items on the stock market should be conducted.

References

[1] Umer, M., Awais, M., & Muzammul, M. (2019). Stock Market Prediction Using Machine Learning (ML)Algorithms. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 8(4), 97–116.

[2] Vaishnavi Gururaj, Shriya V R, and Dr. Ashwini K "Stock Market Prediction Using Linear Regression and Support Vector Machines," International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 8 (2019) Pp. 1931-1934, 2019.

[3] Ishita Parmar, Navanshu Agarwal, Sheirsh Saxena, Ridam Arora, Shikhin Gupta, Himanshu Dhiman, Lokesh Chouhan "Stock Market Prediction Using Machine Learning", Department of Computer Science and Engineering National Institute Of Technology, Hamirpur – 177005, INDIA.

[4] S, Lokasree B. "Data Analysis and Data Classification in Machine Learning Using Linear Regression and Principal Component Analysis", Turkish Journal of Computer and Mathematics Education Vol.12 No.2 (2021), 835- 844.

[5] Sreehari, E., and G. S. Pradeep Ghantasal. "Climate Changes Prediction Using Simple Linear Regression", Journal of Computational and Theoretical Nanoscience Vol. 16, 1–4, 2019.

[6] Maulud, Dastan Hussen, and Adnan Mohsin Abdulazeez. "A Review on Linear Regression Comprehensive in Machine Learning", Journal of Applied Science and Technology Trends Vol. 01, No. 04, Pp. 140 –147, (2020)

[7] Harikrishnan, R., Gupta, A., Tadanki, N., Berry, N., & Bardae, R. (2021). Machine Learning Based Model to Predict Stock Prices: A Survey. IOP Conference Series: Materials Science and Engineering, 1084.

[8] Ahangar, R. G., Yahyazadehfar, M., & Pournaghshband, H. (2010). The comparison of methods artificial neural network with linear regression using specific variables for prediction stock price in Tehran stock exchange. arXiv preprint arXiv:1003.1457.

[9] Sheta, A. F., Ahmed, S. E. M., & Faris, H. (2015). A comparison between regression, artificial neural networks and support vector machines for predicting stock market index. Soft Computing, 7(8), 2.

[10] Har, W.P., & Ghafar, M. (2015). The Impact of Accounting Earnings on Stock Returns: The Case of Malaysia's Plantation Industry. International Journal of Biometrics, 10, 155.