UNIVERSITY OF ECONOMICS AND LAW
FACULTY OF FINANCE AND BANKING

The end of term report

Predict Ethereum price movements based on the news and blockchain information by Machine Learning

Instructors: ***Associate professor PhD. Nguyen Anh Phong, Master. Phan Huy Tam***

Student: ***Nguyen Duc Minh Tan***

Student ID: ***K194141745***

Ho Chi Minh City, June 20, 2022

Table of Contents

Abstract

Cryptocurrency market has potential for growth since Bitcoin or BTC emerged as one of the popular investments with extremely high yield. Besides Bitcoin, there is another altcoin which rapidly developing and dominating the cryptocurrency market, Ethereum. Therefore, this study aims to research and build a model to help investors forecast the price movement of Ethereum most accurately. In this research, the author forecasted Ethereum price movement based on Ethereum-related headline data, Ethereum-specific blockchain information, return of Bitcoin, and Litecoin during the November 2017 to May 2022 period. This study uses two machine learning regression models including random forest regression and support vector regression. The result found that the input variables help to increase the predictability of Ethereum's price movement, in which the Random forest is the best forecast model.

Keywords: alt-coin, bitcoin, blockchain information, cryptocurrency, ethereum, news.

1. Introduction

1.1. Introduction the cryptocurrency and Ethereum

A cryptocurrency or crypto is a form of digital cash that enables individuals to transmit value in a digital setting and it is not owned by any one party. The biggest difference between crypto and internet banking is that there is not a central bank or subset of users that can change the rules without reaching a consensus. The network participants or also called nodes run software that connects them to other participants so that they can share information between themselves.

Ethereum or ETH is a decentralized computing platform. It simultaneously runs on thousands of machines around the world, meaning that it has no owner. Ethereum allows people to transfer digital money. However, it is capable of a lot more, we can deploy your own code, and interact with applications created by other users. Because it is so flexible, all sorts of sophisticated programs can be launched on Ethereum.

1.2. Reason for choosing the topic

Firstly, Ethereum is a good project and this coin also brings many improvements compared to the most well-known coin of the Cryptocurrency market that is Bitcoin. Ethereum is considered a perfect alternative because it has overcome the disadvantages that Bitcoin exists such as being costly, consuming too much energy leading to environmental impact, slow transaction time, etc. The most prominent feature of this network is creating the smart contract. In addition, Ethereum also switched to the proof of stack consensus method, which has more advantages than the previous proof of work consensus method.

Second, Ethereum is the second-largest market capitalization coin, which has a huge impact on the market trend. In addition, there are many coins developed on the system of Ethereum, so accurately forecasting the price movement of Ethereum is also a part of forecasting coins in the market in general and coins developed based on the Ethereum network in particular.

Finally, cryptocurrency can contribute to effectively implementing the Blockchain system in areas where Blockchain is needed. The author hopes that the findings of this study will contribute to the expansion of knowledge in the field of cryptocurrency research. Moreover, cryptocurrencies are the trend of the times, but the number of papers on the topic is still relatively scarce.

2. Literature review.

In the stock market, any business having shares on the exchange is afraid of bad media information since the stock prices are directly influenced by media information about the business If a business is caught in bad media news such as tax evasion, bad products, etc will cause customers' confidence in the business to fluctuate, leading to investors selling short to avoid the risk that the company's stock will tend to decrease sharply.

The cryptocurrency market as well, this market depends much on market news. Investor psychology is a factor that greatly influences buying and selling decisions. Each time there is news of new operating of a cryptocurrency, there is a strong possibility of hitting investor psychology leading to the rise and fall of that coin. Some typical examples can be mentioned such as the increase in the price of Bitcoin after news that the Tesla company allowed customers to buy its electric cars by BTC in February 2021, or the ETH coin increased sharply in 2021 after the market received information that there will be Ethereum 2.0 forming to solve the shortcomings of the old system. Therefore, investors need to filter out the correct information and make investment decisions after making careful calculations.

Previous studies have shown that the impact of news does indeed affect the price movement of cryptocurrencies, including Ethereum. Research of Abraham, Higdon, Nelson, and Ibarra (2018) and Wołk (2020) found the method for predicting changes in Bitcoin and Ethereum prices utilizing Twitter data and Google Trends data. In addition, Smuts (2019) demonstrated that sentiment extracted from cryptocurrency investment groups on Telegram is found to be positively correlated to Bitcoin and Ethereum price movements. In addition, the author Naeem, Mbarki, Suleman, Vo, and Shahzad (2021) concludes a significant non-linear relationship between Twitter happiness sentiment and cryptocurrencies. Therefore, the use of machine learning models in forecasting the price movement of Ethereum is completely reasonable and expected to bring better results than traditional models. However, the experimental results show that using only sentiment analysis in predicting the return trend of Ethereum will bring the forecast performance not too high. Typically in the study Huang et al. (2021), the author outperformed the state of the art auto regressive based model by only 18.5% in the precision and 15.4% in the recall.

In order to get a good forecast result, we need to include in the model other factors that affect the return of Ethereum. According to the empirical study of Sabalionis, Wang, and Park (2021) on factors affecting price movements of Ethereum, the results show that variables like the price of Bitcoin, the price of Ethereum, news of Bitcoin, news of Ethereum, number of active addresses on the Bitcoin blockchain and the number of active addresses on the Ethereum blockchain. In that, Google search interest, number of tweets, and active addresses on the blockchain impact prices of Bitcoin and Ethereum over time. In addition, the author also gives evidence that the amount of active addresses is the most

significant variable among others influencing price movements in Ethereum. Based on spillover effects and GIRFs, the market news about the ETH, to a certain extent, have impacts on the Ethereum prices, but the impacts are weaker than that of active addresses in terms of magnitude and significance.

In addition, the supply and demand of one or more of our cryptocurrencies will affect the return of Ethereum. When a cryptocurrency is bought by many people, the price of that crypto will tend to increase and vice versa. Research by Kim, Bock, and Lee (2021) also shows that the inclusion of Ethereum-specific Blockchain information variables gives better predictive power because they represent the demand for Ethereum. In the blockchain network of this cryptocurrency, Ethereum's gas system consumes Ethereum to input information into the Ethereum Blockchain. Within the study of this author, variables such as gas limits, gas price, gas used, and uncle block variables were selected appropriately as Ethereum-specific Blockchain information to predict Ethereum prices. Research results show that blockchain information is significantly associated with Ethereum prices. In particular, Ethereum-specific Blockchain information contributes to the best performance when it was considered within the model along with macroeconomic factors, the generic Blockchain information of Ethereum, and Bitcoin's Blockchain information. Moreover, among the Blockchain information of other coins (i.e., Bitcoin, Litecoin, and Dashcoin).

Based on previous research, the author decided to perform sentiment analysis in combination with other factors affecting the return of Ethereum including Ethereum-specific blockchain information, return of Bitcoin, and return of Litecoin.

3. Implementation process

3.1. Collect Ethereum related news on Google News

Google News is a website collecting the auto information and news from many sources which are provided by Google. The advantage of getting headline news from this website is that you can get news from many different sources from different websites. At the same time, crawling from this data is also supported by a library built on python, making it easier for researchers to access the data. However, there are a few limitations to crawling data from this website that we can retrieve data from sites that are not reputable. Besides, the computer resources for crawling data are large. In this research paper, the author collects titles of articles by date from November 10, 2017 to May 29, 2022 and analyzes their sentiment to classify them as negative news and positive news about Ethereum. Headlines are considered the main idea of an article, therefore, they fully reflect the nature and content of an article. Besides, the title is the first thing readers see and read, and also the most read part of the article. Readers do not necessarily need to read the entire content of the article, but when looking at the headlines, readers also understand part of the content of an article with negative and positive information. During the crawling process, there were articles that were missing or had incorrect publication timelines. The author has processed these missing values and selected the attributes that are suitable for this study. However, the price movement of cryptocurrencies is not only affected by the news but also by many other factors, such as supply and demand information, information about the operating blockchain, etc. Therefore, the author has performed the Ethereum blockchain data collection in the following step.

3.2. Ethereum blockchain data collection

Blockchains are considered by experts to be the future of the world because of their transparency. Therefore, the information of the blockchain network is always open to the public, and anyone can collect data on public data from these networks. In this study, the author takes Ethereum information data including average difficulty, active address, average gas fee, transaction count, block count, and block size on the website "https://messari.io/asset/ethereum/historical". This website updates the information on Ethereum continuously from the official page "https://ethereum.org/en/" published by this cryptocurrency.

3.3. Collect prices of Ethereum, Bitcoin and Litecoin

Based on previous research papers, the author decided to include the return of Bitcoin and Litecoin model to serve to forecast the price movements of Ethereum. Bitcoin is the largest market capitalization crypto, which has a great influence on the entire cryptocurrency market. If Bitcoin is like a gold coin, Litecoin is considered a silver coin in this market. Besides, Litecoin is a cryptocurrency that is considered to solve problems that Ethereum has not solved in the past and present. That is why the author considers the impact of these two coins on Ethereum. At the step of collecting historical price data of Ethereum, Bitcoin, and Litecoin, the author uses the API method of the website "finance.yahoo" through the library "yfinance". The author collects data daily from November 10, 2017 to May 30, 2022.

3.4. Data processing

For the price data of three cryptocurrencies Ethereum, Bitcoin and Litecoin, the author only uses the adjusted close price "adj_close" value for model building. The author creates a data frame that includes the price of three cryptocurrencies, the blockchain information of Ethereum. To forecast the movement of this cryptocurrency, the author needs to calculate the return of these three coins with the formula:

$$Return_t = \frac{Price_t - Price_{t-1}}{Price_{t-1}}$$

The author processes the headline data that the author has crawled from Google News. The author performs sentiment analysis of the title using the library "TextBlob". However, the author wants to analyze the frequency of occurrence in the articles to get a rough estimate of this data at first. The author divides the title into 2 types of words, one is common words, the other is stop words in English. We can see that the top 5 of stop words

that appear most frequently in the author's dataset are “to”, “the”, “and”, “in” and “of” as shown in figure 1.

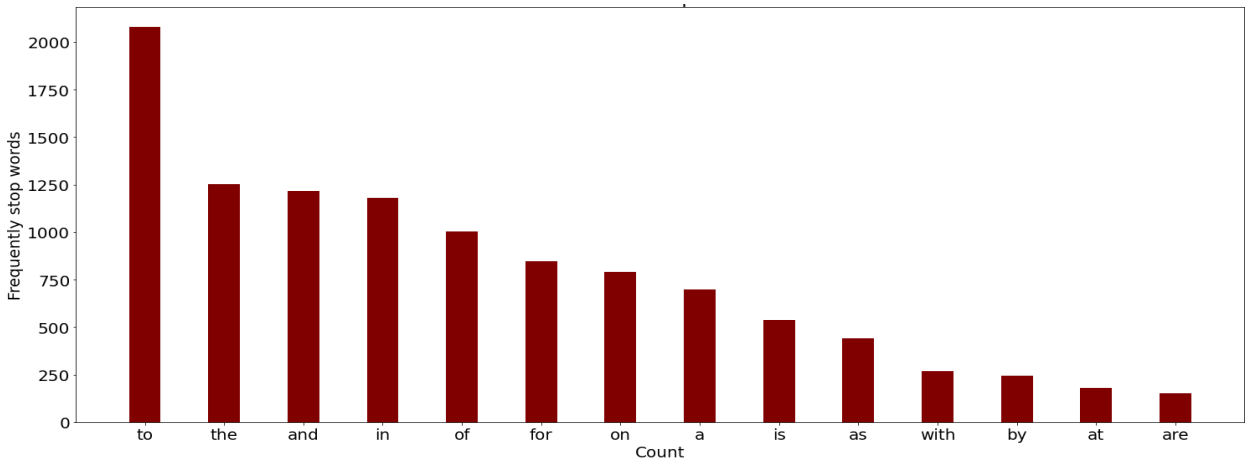

Figure 1. Frequency of occurrence of stop words

Next, the author visualizes the words that frequently appear in the titles represented in figure 2. The words that frequently appear in the titles that I collect are almost all related to Ethereum and the market which can be mentioned as Bitcoin, crypto, defi, blockchain, and some other altcoins, etc.

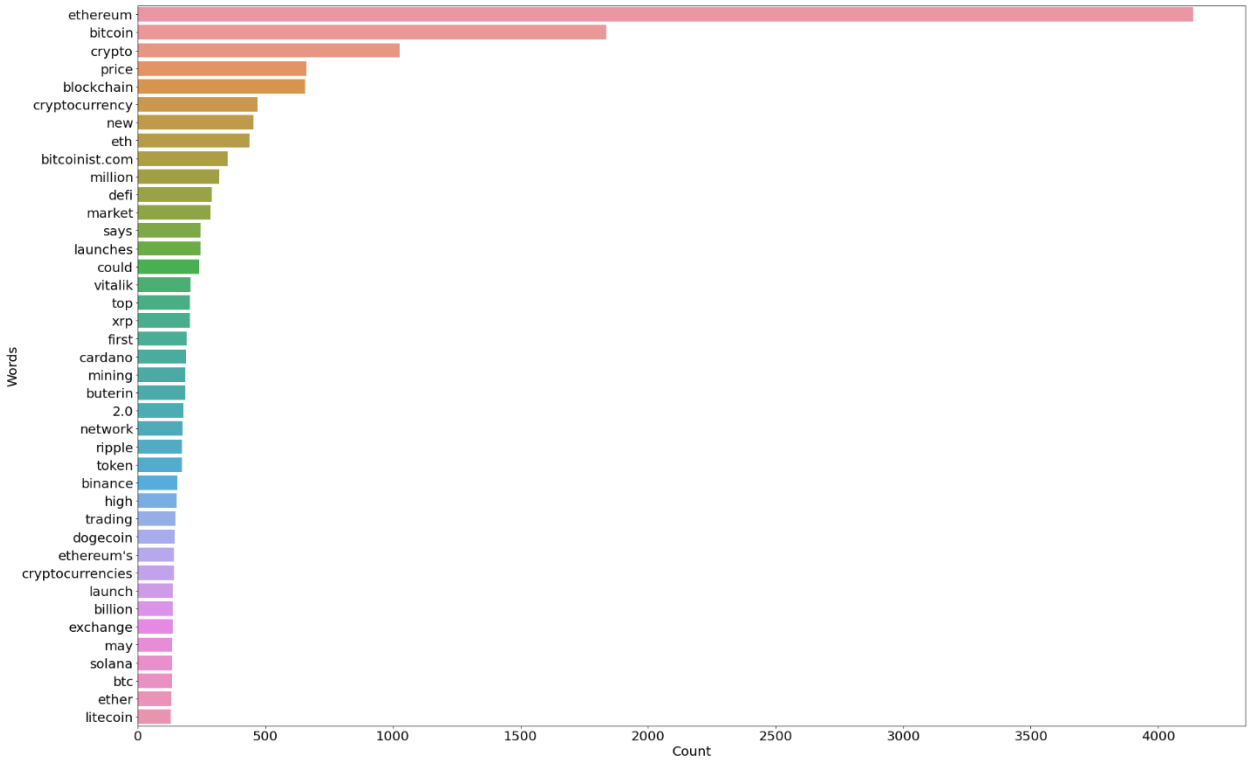

Figure 2. Frequency of occurrence of common words

The author analyzes the headlines of the articles and calculates two indexes of polarity, and subjectivity to assess the sentiment of the original newspaper headline. Polarity refers to the strength of an opinion. This indicator ranges from -1 to 1, representing positive, negative, and natural news. If something has a strong positive feeling or emotion associated with it will indeed have a certain orientation towards all other aspects of the existence of that object. If the title of an article has a negative calculated polarity, then the

paper carries negative information, if the polarity is greater than zero, the article content tends to be positive, and the title with this index of 0 is articles with natural information. Then, the author transforms this feature into 2 dummy variables for articles with negative and positive information. The author only separates 2 dummy variables instead of 3 variables to avoid perfect correlation between dummy variables. However, there are days when there will not be any information related to the market or to Ethereum, the author has filled the missing value with the forward fill method "ffill" because during the period of no new information, the old information will still have an impact on the return of the Ethereum.

Subjectivity refers to the degree to which a person is personally involved in an object. What matters the most here are personal connections and individual experiences with that object, which may or may not differ from someone else's point of view. Subjectivity has a value from 0 to 1, showing an increasing degree of subjectivity as this index gets closer to value 1. Sentiment analysis tools must be able to distinguish between both polarity and subjectivity in order to analyze opinions of users correctly. An opinion can have a high degree of subjectivity if it is expressed as a personal experience, whereas a low degree may indicate someone else's viewpoint on something else. In this study, the news about Ethereum that the author has collected during this period is mainly natural and objective as shown in figure 3. However, the number of articles with positive content is more than negative articles about ETH.

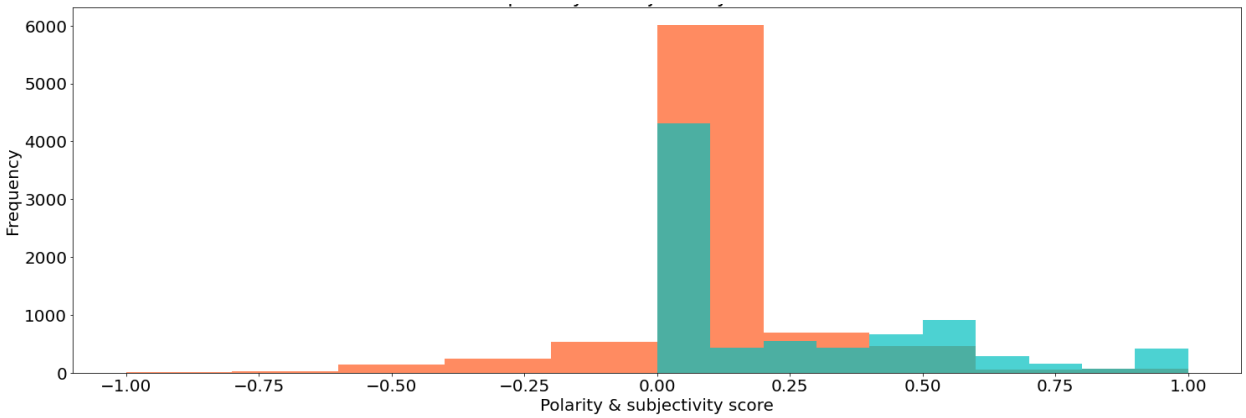

Figure 3. The polarity & subjectivity distribution

Although I collected data from November 10, 2017 to May 29, 2022, I divided this dataset into 2 parts. Part of the data from November 10, 2017 to November 10, 2022 is used to build predictive models. The remaining dataset, from November 11, 2022 to May 29, 2022 serves as an assessment of the model predictability on a new input dataset.

3.5. Descriptive Statistics

The author's dataset used for forecasting includes 1654 observations and 13 features, of which there are two dummy variables "Negative" and "Positive" described in table 1, and there are 11 continuous variables described in table 2.

Table 1. Statistical Description of dummy variables

	Negative	Positive
Count	1654	1654
Unique	2	2
Top	0	1
Freq	1335	1180

From the period November 2017 to the end of May 2022, there were a total of 1180 positive news about Ethereum, 319 negative news, and 155 natural content. This makes perfect sense as the price of Ethereum has also increased significantly during this period as shown in figure 4.

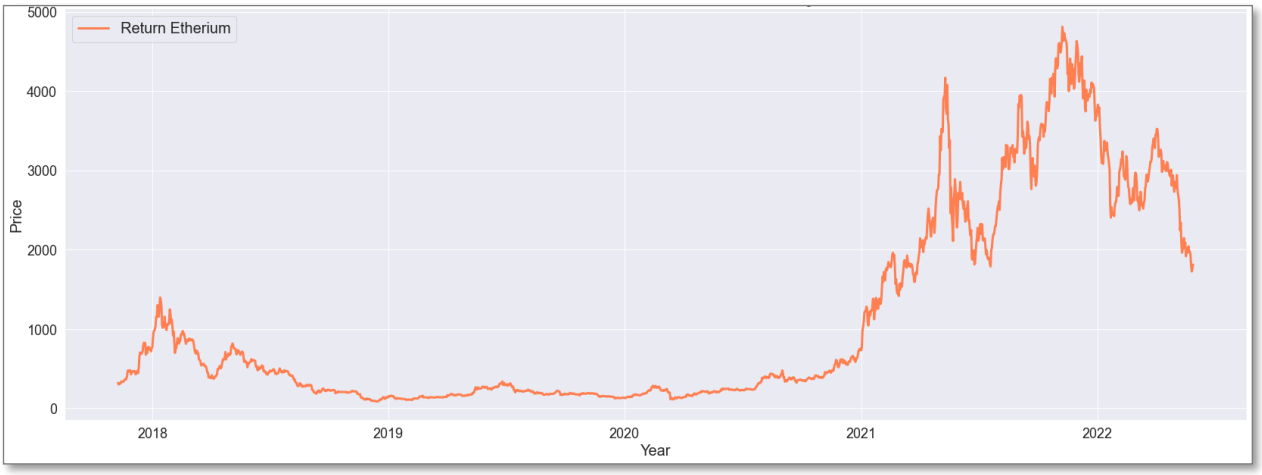

Figure 4. Price of Etherium from November 2017 to May 2022

Table 2. Statistical Description of continuous variables

	polarity	subjectivity	return_etherium	return_bitcoin	return_litecoin	avg_diff	act_address	avg_gas	trans_count	block_count	block_size
count	1654.000000	1654.000000	1654.000000	1654.000000	1654.000000	1.654000e+03	1654.000000	1654.000000	1.654000e+03	1654.000000	1.654000e+03
mean	0.060225	0.229316	0.002254	0.001534	0.001490	4.660133e+15	435454.021161	6.952772	9.070663e+05	6224.828900	2.431480e+08
std	0.099130	0.143044	0.051057	0.040767	0.057099	3.551804e+15	147010.239064	12.890340	2.865304e+05	399.193601	1.535289e+08
min	-0.600000	0.000000	-0.423472	-0.371695	-0.361773	1.390000e+15	171484.000000	0.050078	3.811510e+05	4139.000000	7.937030e+07
25%	0.000000	0.125000	-0.021807	-0.016276	-0.026834	2.300000e+15	305877.250000	0.145192	6.460402e+05	6038.250000	1.309455e+08
50%	0.053002	0.216667	0.001399	0.001477	-0.000090	3.130000e+15	404810.500000	0.586776	8.551020e+05	6404.000000	1.648381e+08
75%	0.101667	0.314559	0.027832	0.018983	0.027714	6.200000e+15	562800.500000	8.263823	1.173746e+06	6475.000000	3.037059e+08
max	0.750000	1.000000	0.264581	0.252472	0.475978	1.470000e+16	905234.000000	200.271677	1.716600e+06	6637.000000	7.664025e+08

Through Table 2, we can see that the volatility return of all three cryptocurrencies is relatively large. During the peak period, the return of ETH and BTC is 26,46% and 25,25% respectively. Meanwhile, LTC has a superior return when reaching 47,6%. However, during the weakening period, the return of these coins also declined at an alarming rate as the return of ETH fell 42,35%, BTC fell 37,17% and LTC fell 36,18%. Ethereum is one of the coins that users care about and use much. This coin has a large number of active addresses, averaging about 435454 accounts and having an average of 907066 transactions per day. The highest active address of ETH during this period was 905234 and the lowest active address was 171484. The gas fee of Ethereum is relatively high compared to other coins with the average gas fee of almost 7 ETH. At times when transactions are blocked due to network overload, the gas fee of this coin can amount to more than 200 ETH for a transaction.

3.6. Building model

First, we need to divide 80% of data into the train set, and 20% data into the test set. The train set is responsible for training the model to find out the predictive principles of the dependent variable while the test set is the set used to evaluate the predictive performance of this model. This work helps the author to evaluate the predictive performance of the model more objectively because the test set is a set of data not encountered by the model before. In this study, the author built two models including Random forest regression and Support vector regression to serve for forecasting the price movement of Ethereum. In the Support vector regression, we can generate multiple decision functions from different kernel functions to help improve the prediction results and Wang and Xu (2017) also showed that different kernel functions are chosen to form different support vector regression for each different data feature and give different prediction results. The SVR model handles non-linear data using a number of kernel functions such as RBF, linear, polynomial, sigmoid, and precomputed. A kernel is a function that takes the original non-linear problem and transforms it into a linear one, which is then handled by the algorithm in a higher-dimensional space. In this study, the author implements the Support vector regression model with the two most popular kernels, RBF and linear.

3.7. Evaluate the predictive performance of the model

To evaluate the predictive efficiency of a model, the author visualizes the actual results and the forecast results, combining comparisons on the Root Mean Square Error (RMSE), mean absolute error (MAE) and Mean absolute percentage error indexes (MAPE). The closer the above indicators move to zero, the better the model results.

Table 3. Error index

	Random forest	SVR RBF kernel	SVR linear kernel
RMSE	0,031	0,0362	0,0294
MAE	0,0196	0,0207	0,0191
MAPE	2,28%	1,88%	1,4%

The results from the table 3 show that the prediction abilities from the Random forest model and Support vector regression with the RBF kernel have more errors than the Support vector regression model with the linear kernel because this model all has RMSE, MAE and MAPE are lower. In addition, the author has visualized the forecast results of the three models above in figure 5. The better model is the one whose prediction line is closer to the actual line and has smaller error estimates. The results also show that SVR with linear kernel is the model that has the closest prediction to the real one. However, in order to have a more objective assessment of the predictability of the above models, the

author makes the price percentage change of Ethereum forecast on a new dataset from November 11, 2021 until May 29, 2022.

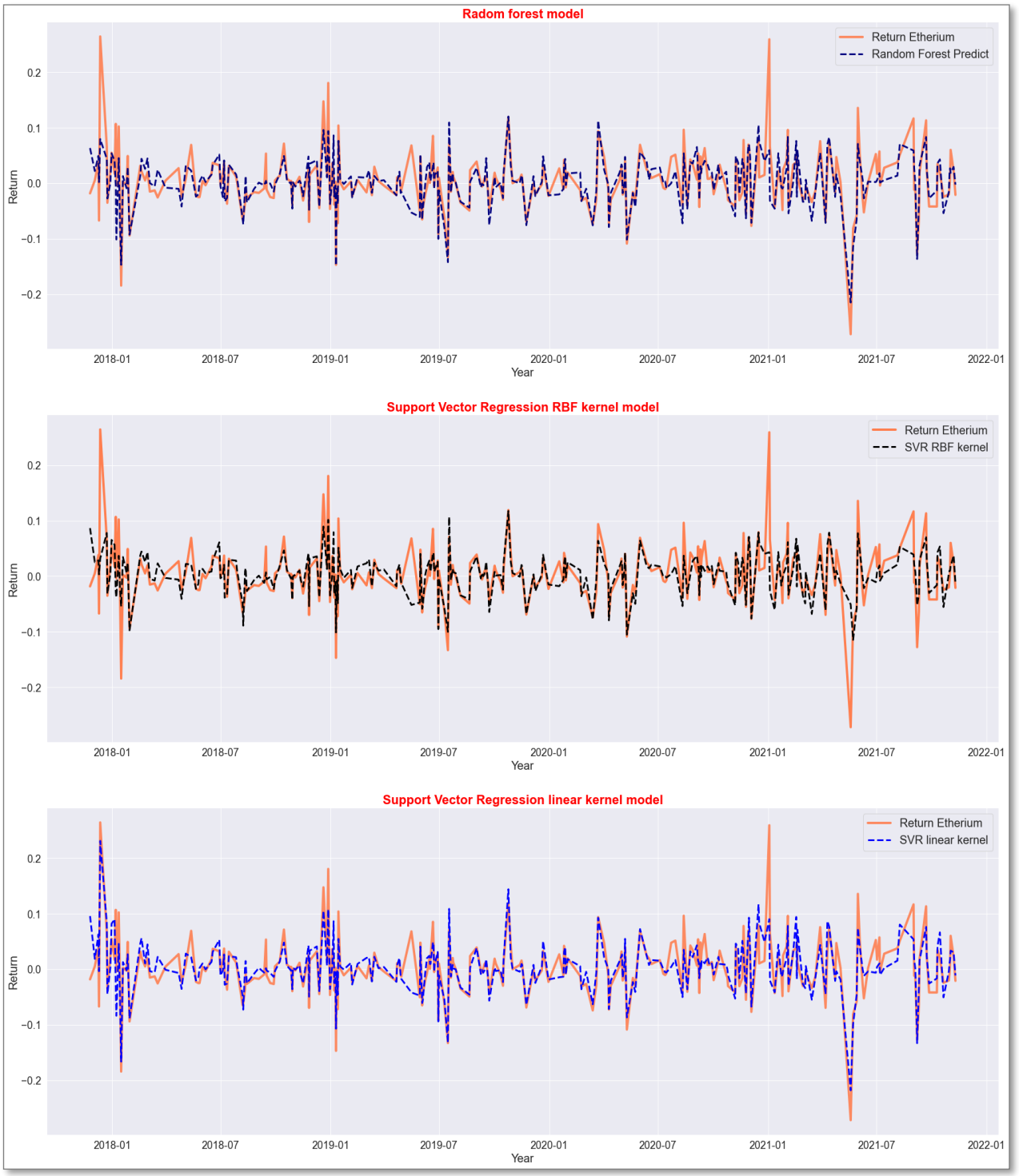

Figure 5. Random forest model & Support vector regression prediction

3.8. Model forecasting and evaluation on new datasets

After the forecast results are available, the author compares and evaluates them to find the best model used in this study. In the step evaluate the predictive performance of the model on the initial input data, SVR with the linear kernel is the model that does the best predicting task, however, the results from the table 4 show that the Random forest is

the model that gives the best predictive results when working on a new input dataset and has been visualized in figure 6. All three models are capable of accurately forecasting Ethereum's price movement but each has small deviations in the exact percentage change. More specifically, the Random forest model has RMSE = 0,02063 and MAE = 0,0154 which is the lowest error index of the three models. Besides, the forecasting results in the old dataset and the new dataset do not have too much difference in the predictive power of all three models. Thus, these models do not occur overfitting or underfitting.

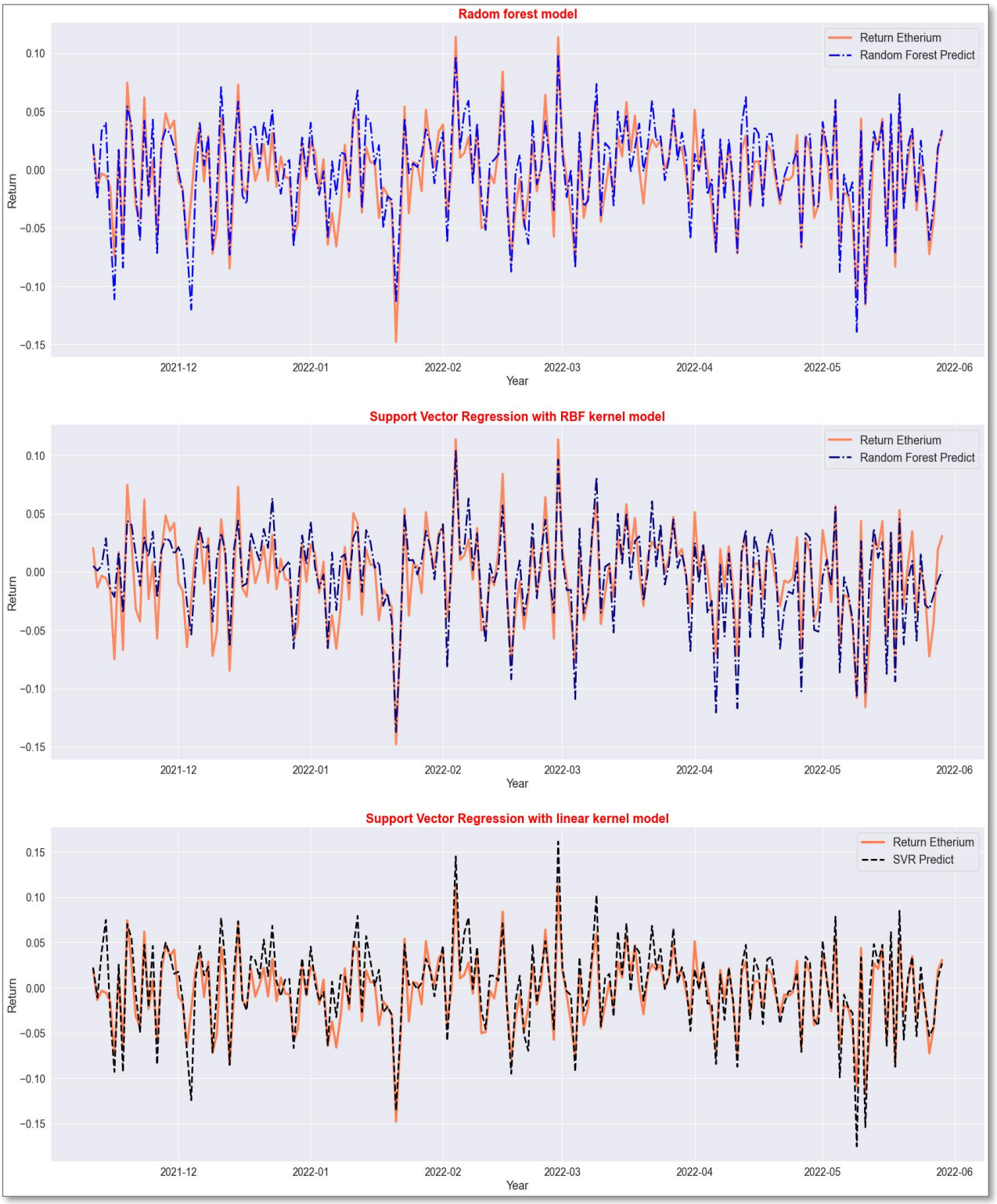

Figure 6. Predict Ethereum from 11/11/2021 to 29/05/2022

Table 4. Error index on new dataset

	Random forest	SVR RBF kernel	SVR linear kernel
RMSE	0,02063	0,0219	0,0224
MAE	0,0154	0,0175	0,0163
MAPE	1,94%	1,95%	1,83%

4. Conclusion

From the above forecast results, we can conclude that the input data such as sentiment of headline, the specific blockchain information, return of Bitcoin and Litecoin are appropriate and all have an impact on the price percentage change of Ethereum. Among the three models that the author has implemented in this study, the Support vector regression with the linear kernel model gave better prediction results than the others but the Random forest is the best model in the prediction on the new dataset. Although the cryptocurrency market is considered a high degree of risk market, the volatility of assets in this market is very large and is highly exposed to tail-risk Borri (2019), three of the above models give highly accurate forecast results and have similar results in the trend return of Ethereum and the prediction differences between these three models are tiny. Final conclusion, there is a different result in finding the most effective model for forecasting the price movement of Ethereum on two different datasets in this study and the author evaluates the Random forest model as the more effective model because it does better on a new dataset.

Reference:

Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. *SMU Data Science Review, 1*(3), 1.

Borri, N. (2019). Conditional tail-risk in cryptocurrency markets. *Journal of Empirical Finance, 50*, 1-19.

Huang, X., Zhang, W., Tang, X., Zhang, M., Surbiryala, J., Iosifidis, V., . . . Zhang, J. (2021). *Lstm based sentiment analysis for cryptocurrency prediction.* Paper presented at the International Conference on Database Systems for Advanced Applications.

Kim, H.-M., Bock, G.-W., & Lee, G. (2021). Predicting Ethereum prices with machine learning based on Blockchain information. *Expert Systems with Applications, 184*, 115480.

Naeem, M. A., Mbarki, I., Suleman, M. T., Vo, X. V., & Shahzad, S. J. H. (2021). Does Twitter happiness sentiment predict cryptocurrency? *International Review of Finance, 21*(4), 1529-1538.

Sabalionis, A., Wang, W., & Park, H. (2021). What affects the price movements in Bitcoin and Ethereum? *The Manchester School, 89*(1), 102-127.

Smuts, N. (2019). What drives cryptocurrency prices? an investigation of google trends and telegram sentiment. *ACM SIGMETRICS Performance Evaluation Review, 46*(3), 131-134.

Wang, H., & Xu, D. (2017). Parameter selection method for support vector regression based on adaptive fusion of the mixed kernel function. *Journal of Control Science and Engineering, 2017*.

Wołk, K. (2020). Advanced social media sentiment analysis for short-term cryptocurrency price prediction. *Expert Systems, 37*(2), e12493.