Analysis of Crypto based Sub-Reddit ’Ethereum’ and it’s influence on the Marker Price

Pushkar Deshpande, Sharanbasav Sumbad, Sowmya Reddy Kogilathota Jagirdar

Abstract

The Ethereum market is a mix of conflicting information from different social media platform such as twitter and reddit. Through the ideas of data mining form, the popular social media platform Reddit and machine learning, this paper tries to give a concise and straightforward representation of the data from the sentiment of the comments of a popular Ethereum subreddit such that even the most oblivious eye can form inferences. Ethereum trading is rapidly expanding and is becoming increasingly important in data research, Social Media Mining and economics. Machine Learning, on the other hand, is a branch of data science that focuses on creating algorithms that interpret the data supplied by the user and forecast the value appropriately. This article focuses on forecasting future market patterns for Ethereum using a machine learning approach called Vector Auto Regression (VAR). This is extremely beneficial since the investor may not only compare several equities at the same time, but he or she can also get highly precise data, with a reasonable evaluation metric rate. We want to take this a step further by offering insight into future market patterns and allowing investors to evaluate all currencies on a single platform before making their investment.

Introduction

Cryptocurrencies are digital assets that may be traded between people. Because they don’t need a central authority to govern them, they’re decentralized assets. As with other exchanges, issues emerge regarding what factors may influence their movement. The beginning of 2021 has provided us with instances of how social media influences the stock market and the crypto currency industry, such as the GameStop case and Elon Musk’s tweets advocating cryptocurrency investing. This paper looks at sentiment analysis on comments made on ’Ethereum’ subreddit to examine how it relates to the ethereum market price. Reddit is a social news aggregation, online content rating, and discussion website with a wealth of information for us to go through. It is a social media network like Twitter that collects information from numerous sources and organizes it in one place; the main distinction is that you don’t follow individuals on Reddit, but rather a topic of interest. We can tell whether a piece of writing is good, negative, or neutral by analyzing the emotion of words from certain postings. Sentiment analysis is a branch of natural language processing (NLP) that can be particularly effective in specific industries depending on the circumstances. For example, a corporation may be attempting to comprehend its consumers’ feelings based on product reviews. The company may use the sentiment to have a better understanding of how customers react to things and plan future business strategies. Cryptocurrency market analysis provides you with the latest information about trading tendencies, new players and analytics on recent cryptocurrency news. The market analysis includes the review of Ethereum,Bitcoin and other prominent altcoins, being excellent support for investment decision-making.
Technical analysis may be used to analyze the crypto market, focusing on statistical patterns and monitoring price movements, historical volume, and activity in order to produce price projections and predictions for the long and short term.Ethereum is based on blockchain technology, which means that it is operated by a decentralized network of nodes that collectively keep track of transactions. When someone submits a transaction to the network, the nodes verify it, and if there is consensus, the transaction is completed and recorded in a ledger. A transaction on the Ethereum network can also be the execution of a smart contract or dap. Proof of stake is a consensus technique used by Ethereum to validate transactions. This implies that the nodes that verify transactions are chosen based on their Ethereum system stake percentage. To host a node that bets Ethereum, you must currently have at least 32 ETH. Despite the high costs, Ethereum is still the definitive most active and most used blockchain r/Ethereum -The main Ethereum subreddits, r/ EtherMining - Ethereum Mining, and r/EthTrader - Trading/Marketplace discussion is three most used Ethereum subreddits platform. Ethereum’s price dynamics from January to December 201 9may be divided into three phases: optimism, pessimism, and finally, a consistent oscillation, demonstrating the importance of mood. As a result, studying price-sensitive social media discussions is extremely crucial for any crypto.

Related Work

A market is said to be efficient if the price has all type of information included in it, as future prices, past movements, etc., in this case, the market investors will not be able to predict it, as all investment strategies are already included in the price, being only independent and unexpected scenarios those that move the price (1).

In the paper (2) ,we see that VAR is the best among the forecasting algorithms such as ARIMA in forecasting the prices

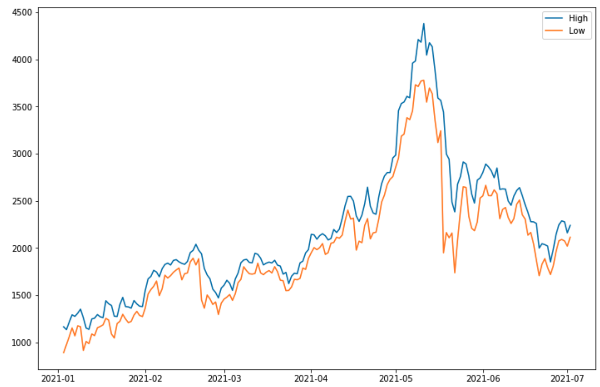


Fig. 1. High and low prices of Ethereum

based on the market data. The paper focuses on bitcoin price in comparison with other crypto currencies using VAR and VAR model performed better compared to ARIMA. ARIMA being univariate it can process only one time series at a time and takes more time to process multiple time series, whereas VAR being multivariate computes multiple time series at a time is faster with better performance compared to ARIMA. This also concludes that ARIMA model requires more computational power than VAR.

Existing works that tie online activity or sentiment to cryptocurrency prices have focused on online forums dedicated to a specific currency. (3), tweets containing a specific currency name or hashtag such as “bitcoin” or ” bitcoin” (4), (5), (6), (7), or only included the primary subreddit for a given currency (8).

Verma and Sharma (9) tried to predict the Bitcoin price using the sentiment from the Twitter social media using a ELMo embedding model and a SVM (Support Vector Machines) classifier, from which a main observation is highlighted for the present project which is that Bitcoin prices are not affected by investors sentiment, at least on Twitter, conclusion that is contradicted by Jahjah and Rajab (10), in which a strong relation between the Twitter sentiment and the Bitcoin returns is stated, and by Kraaijeveld and De Smedt (11), who used causality analysis to determined that Twitter sentiment can be used to predict Bitcoin, Bitcoin Cash and Litecoin prices.

Chevallier et al. (2021) stated that the state of the art investment strategies as the Buy and Hold, tend to show better performances than those methodologies based on Machine Learning. The prediction of the cryptocurrency prices is still a recent subject, by which we can find in the literature, research that contradict each other. For Reddit, which is on the scope of this project, it has too been used as a feature for cryptocurrency price prediction ((12)),and future price movements (like pump-and-dump schemes, (12)).

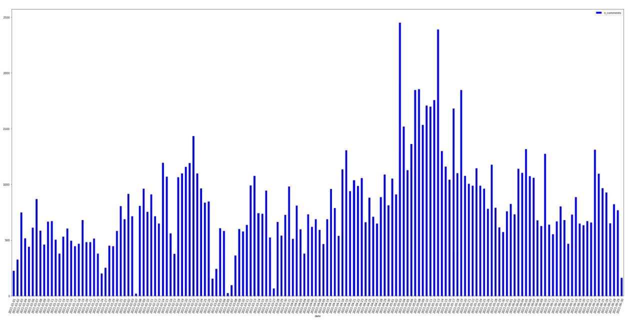


Fig. 2. Active users over time

Data

Data collection

Cryptocurrency price is extracted from the CryptoCompare API which is a free API. This API returns the current price of any cryptocurrency and all the trading info (price, vol, open, high, low etc) for the requested pairs. To extract the top subreddit for crypto we are using Reddit from IU server. The posts are scrapped considering the name of the cryptocurrency and the symbol of it; the corresponding keywords are bitcoin, BTC, ethereum, ETH, litecoin, and LTC. The reddit data used will be six months data from year 2021. We used the top subreddit community that focuses on discussing either Ethereum or cryptocurrencies in general. The subreddit being used is ‘r/Ethereum’. Figure 1 and Figure 2 shows the daily comments and submissions from the sampled Reddit community along with closing prices from CryptoCompare.

Changes in price and changes in levels of activity in these subreddits often accompany one another which suggests that the sampled Reddit data used here may exhibit a similar connection between closing price and activity to what has been found in previous works.

Data Preparation

The data preparation process starts by cleaning the corresponding posts from external links, deleted/removed comments, and English specific posts.Many research papers have approached the problem that social media bots impose on sentiment analysis methodologies on the cryptocurrency subject. We used the list of the bots from csv generated by [Peter] to filter out bots because bot comments interfere with the sentiment analysis.These bots comments are filtered out by removing the comments with the author names from the bot list.

The data that has observations with ’NA’ values and these need to be filtered as further processing of requires there to be no ’NA’ data. Hence, these are filtered by applying conditions on the processed data.

The comment data may contain things like urls,references to other subreddits,other users, and raw comment text that is used to format comments on the website, to get rid of this we filtered the data using regular expressions.

Then we performed language filtering by eliminating the

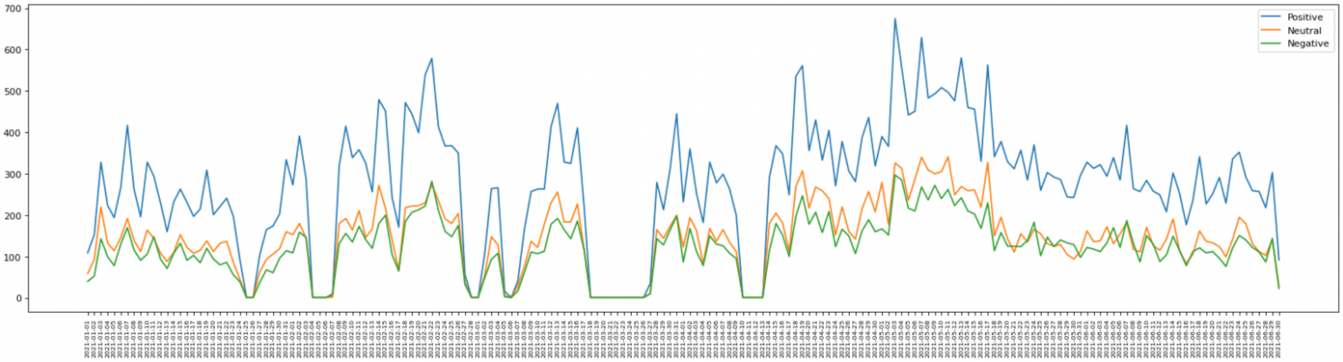


Fig. 3. Sentiment analysis

comments that did not contain English using regular expressions.

Methodology

Sentiment analysis

Sentiment analysis is a text analysis method that detects polarity within the text, whether a whole document, paragraph or sentence.We have VADER for detecting the sentiment of the users through comments. VADER (Valence Aware Dictionary for Sentiment Reasoning) is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion. It is available in the NLTK package and can be applied directly to unlabeled text data. VADER sentimental analysis relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores. The sentiment score of a text can be obtained by summing up the intensity of each word in the text.

VADER has a comprehensive lexicon, we can add sentiment words which are specific to our domain.The vaderlexicon.txt from the VADER package consists of all the sentiment scores with the their respective weights. As in this context,we are interested in financial sentiment.The csv file created by (13) has compiled the financial sentiment with columns for positive and negative words which we used for the analysis.The financial sentiment lexicon can be used from Loughran-McDonald financial sentiment dictionary. The combined lexicon from the csv file and the Loughran-McDonald file is then used.

We then apply the SentimentIntensityAnalyzer to the comment data's body column which provides sentiment polarity scores to the comments with negative, neutral and positive scores. These scores are then categorized into sentiment categories by assigning any value>= 0.05 to positive comment and value between 0.05 to -0.05 to neutral comment and rest to negative comment.

These categories are then used to assign then used to convert into binary variables which classifies the sentiment. The data is then merged with the processed reddit dataset.

Price prediction

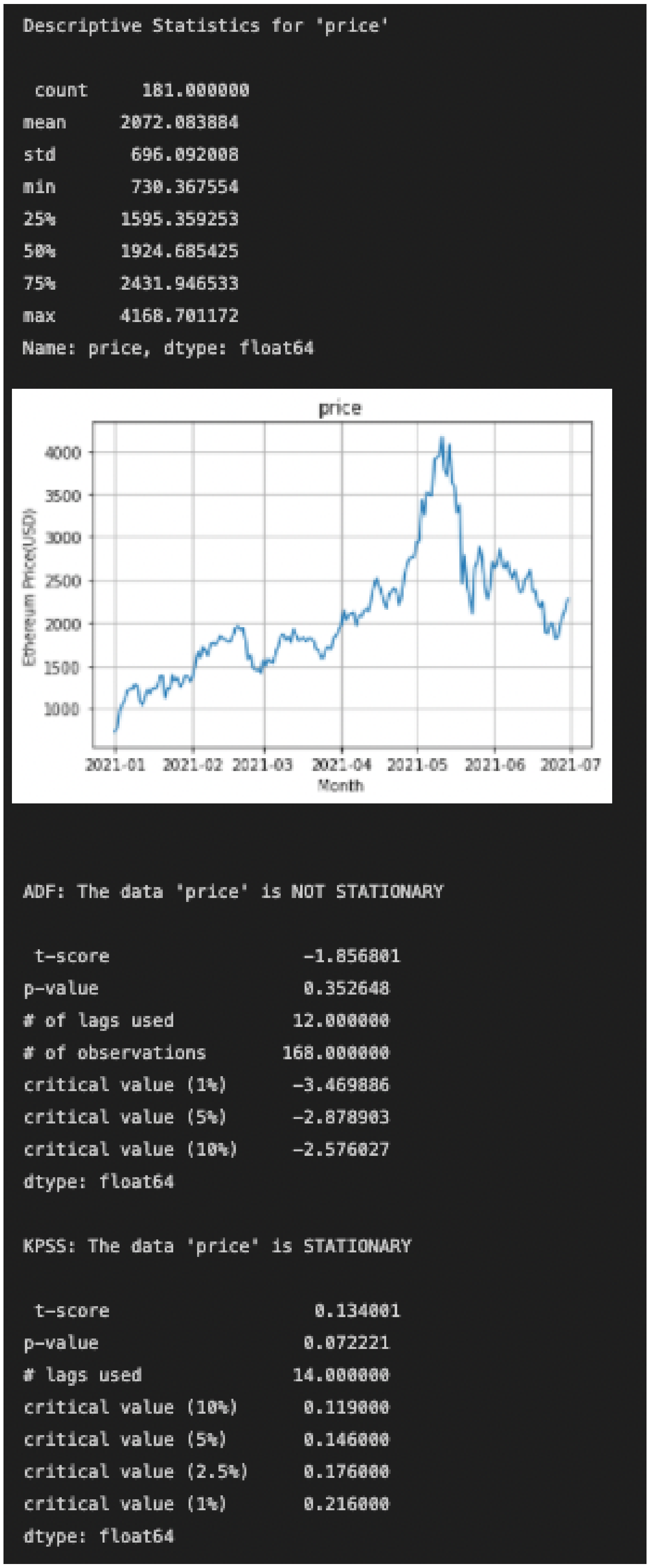
Formatting for Analysis:

For Analysis we need the data in the right format. As we obtained the data from the reddit it had comments we implemented the Sentiment analysis using the Vader sentiment analyzer. In our reddit data we convert the UTC time stamp to the Date format and proceed further , we then group by the date column to the total sentiment by month. We check if the date format is the same from the Reddit dataset and the price data that we extracted from the Cyprocompare API.Both the datasets are then merged with the 'Date' as the index.This provides us with a dataframe that contains the price, negative,positive and neutral comments counts for each month.We redefine the data based on the last date that we have the sentiment data for. We plotted a graph with the prices against the date and the negative,positive and neutral comments over time as seen in the figure 3 and 4.From the figures we can see that the neutral and negative comments are at the same level throughout the timeline whereas the positive comments seems to be higher for the cryptocurrency throughout.

Stationary transformation:

Our data must remain stationary for our time series forecasting to be effective. In order to remain immobile, we must maintain the series mean constant to 0. The variation should be consistent throughout time. Over time, the co-variance should be the same. Because a model describing the data without stationary would vary in accuracy at different time points, stationary is critical. As a result, for sample statistics like means, variances, and correlations to adequately characterize the data at all time periods of interest, stationarity is essential. Stationarity is important because a model explaining the data without it would vary in accuracy at different time points. As a result, stationarity is required for sample statistics like as means, variances, and correlations to appropriately characterize the data throughout all time periods of interest.

To check the stationarity in our data we make use of ADF, Augmented Dickey Fuller test, A test statistic is produced, and p-values are presented because of hypothesis testing using a null and alternate hypothesis. We may deduce if a particular time series is stationary or not based on the



```
Descriptive Statistics for 'price'

 count     181.000000
mean      2072.083884
std        696.092008
min        730.367554
25%       1595.359253
50%       1924.685425
75%       2431.946533
max       4168.701172
Name: price, dtype: float64
```

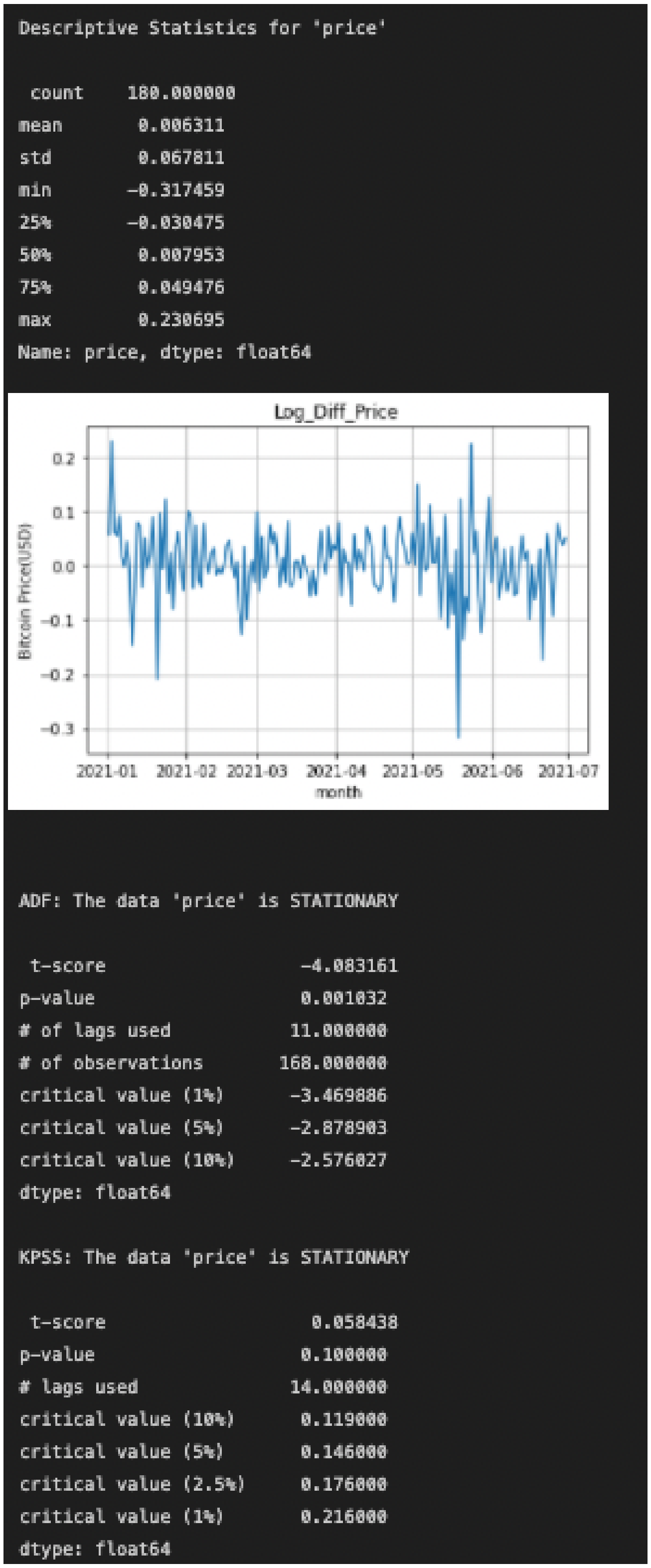
```
ADF: The data 'price' is NOT STATIONARY

 t-score                  -1.856801
p-value                   0.352648
# of lags used           12.000000
# of observations       168.000000
critical value (1%)      -3.469886
critical value (5%)      -2.878903
critical value (10%)     -2.576027
dtype: float64

KPSS: The data 'price' is STATIONARY

 t-score                   0.134001
p-value                   0.072221
# lags used              14.000000
critical value (10%)      0.119000
critical value (5%)       0.146000
critical value (2.5%)     0.176000
critical value (1%)       0.216000
dtype: float64
```

Fig. 4. Descriptive Statistics for 'price'



```
Descriptive Statistics for 'price'

 count    180.000000
mean       0.006311
std        0.067811
min       -0.317459
25%       -0.030475
50%        0.007953
75%        0.049476
max        0.230695
Name: price, dtype: float64
```

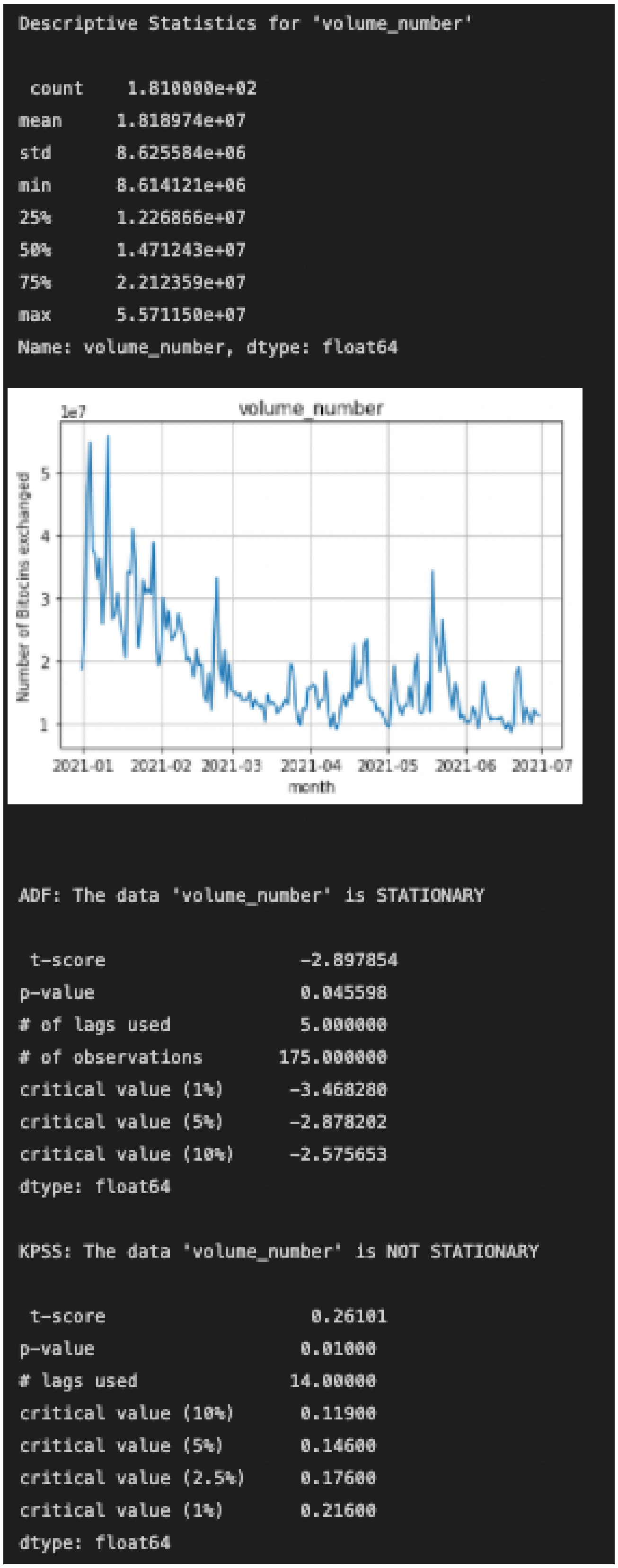
```
ADF: The data 'price' is STATIONARY

 t-score                  -4.083161
p-value                   0.001032
# of lags used           11.000000
# of observations       168.000000
critical value (1%)      -3.469886
critical value (5%)      -2.878903
critical value (10%)     -2.576027
dtype: float64

KPSS: The data 'price' is STATIONARY

 t-score                   0.058438
p-value                   0.100000
# lags used              14.000000
critical value (10%)      0.119000
critical value (5%)       0.146000
critical value (2.5%)     0.176000
critical value (1%)       0.216000
dtype: float64
```

Fig. 5. Descriptive Statistics for 'price' after log differencing



```
Descriptive Statistics for 'volume_number'

 count    1.810000e+02
mean     1.818974e+07
std      8.625584e+06
min      8.614121e+06
25%      1.226866e+07
50%      1.471243e+07
75%      2.212359e+07
max      5.571150e+07
Name: volume_number, dtype: float64
```

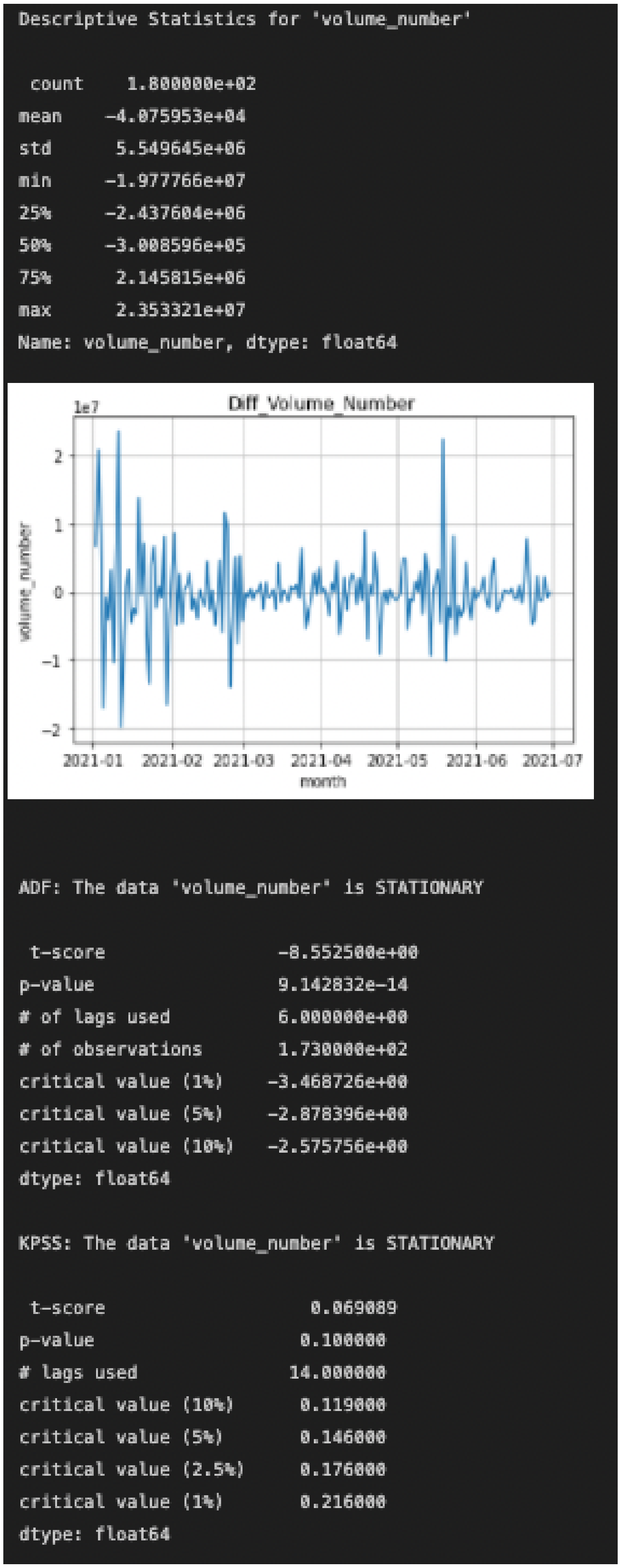
```
ADF: The data 'volume_number' is STATIONARY

 t-score                  -2.897854
p-value                   0.045598
# of lags used            5.000000
# of observations       175.000000
critical value (1%)      -3.468280
critical value (5%)      -2.878202
critical value (10%)     -2.575653
dtype: float64

KPSS: The data 'volume_number' is NOT STATIONARY

 t-score                   0.26101
p-value                   0.01000
# lags used              14.00000
critical value (10%)      0.11900
critical value (5%)       0.14600
critical value (2.5%)     0.17600
critical value (1%)       0.21600
dtype: float64
```

Fig. 6. Descriptive Statistics for 'volumenumber'



```
Descriptive Statistics for 'volume_number'

 count    1.800000e+02
mean    -4.075953e+04
std      5.549645e+06
min     -1.977766e+07
25%     -2.437604e+06
50%     -3.008596e+05
75%      2.145815e+06
max      2.353321e+07
Name: volume_number, dtype: float64
```

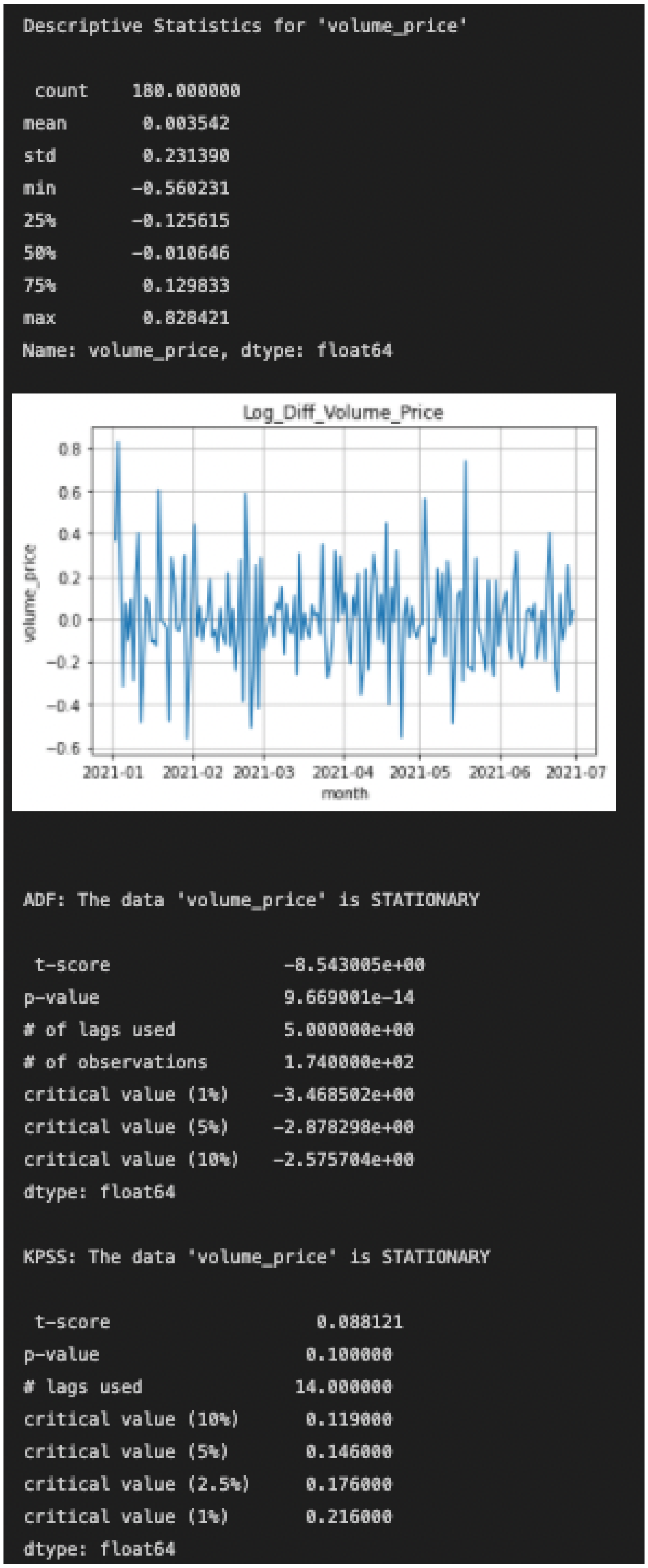
```
ADF: The data 'volume_number' is STATIONARY

 t-score                -8.552500e+00
p-value                 9.142832e-14
# of lags used          6.000000e+00
# of observations       1.730000e+02
critical value (1%)    -3.468726e+00
critical value (5%)    -2.878396e+00
critical value (10%)   -2.575756e+00
dtype: float64

KPSS: The data 'volume_number' is STATIONARY

 t-score                    0.069089
p-value                    0.100000
# lags used               14.000000
critical value (10%)       0.119000
critical value (5%)        0.146000
critical value (2.5%)      0.176000
critical value (1%)        0.216000
dtype: float64
```

Fig. 7. Descriptive Statistics for 'volumenumber' after log differencing



```
Descriptive Statistics for 'volume_price'

 count    180.000000
mean       0.003542
std        0.231390
min       -0.560231
25%       -0.125615
50%       -0.010646
75%        0.129833
max        0.828421
Name: volume_price, dtype: float64
```

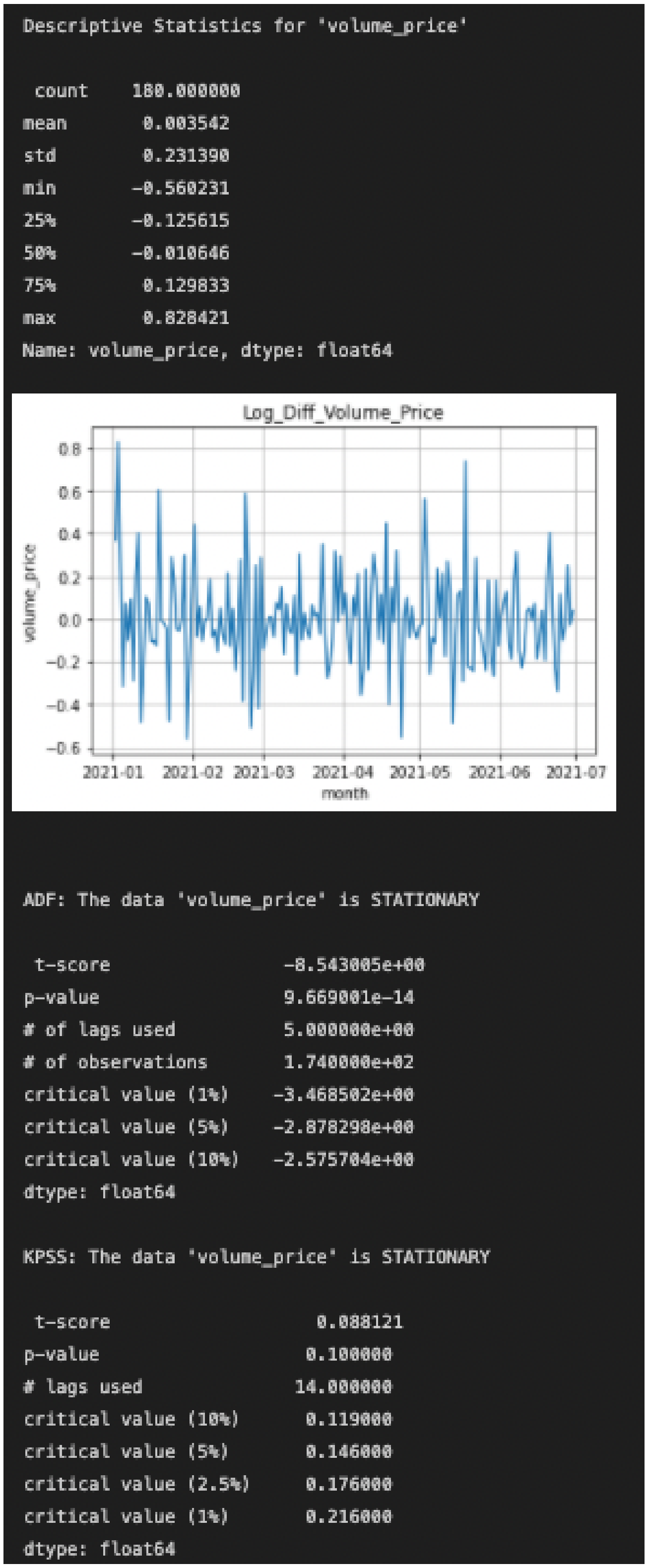
```
ADF: The data 'volume_price' is STATIONARY

 t-score               -8.543005e+00
p-value                9.669001e-14
# of lags used         5.000000e+00
# of observations      1.740000e+02
critical value (1%)   -3.468502e+00
critical value (5%)   -2.878298e+00
critical value (10%)  -2.575704e+00
dtype: float64

KPSS: The data 'volume_price' is STATIONARY

 t-score                  0.088121
p-value                  0.100000
# lags used             14.000000
critical value (10%)     0.119000
critical value (5%)      0.146000
critical value (2.5%)    0.176000
critical value (1%)      0.216000
dtype: float64
```

Fig. 8. Descriptive Statistics for 'volumeprice'



```
Descriptive Statistics for 'volume_price'

 count    180.000000
mean       0.003542
std        0.231390
min       -0.560231
25%       -0.125615
50%       -0.010646
75%        0.129833
max        0.828421
Name: volume_price, dtype: float64
```

```
ADF: The data 'volume_price' is STATIONARY

 t-score               -8.543005e+00
p-value                9.669001e-14
# of lags used         5.000000e+00
# of observations      1.740000e+02
critical value (1%)   -3.468502e+00
critical value (5%)   -2.878298e+00
critical value (10%)  -2.575704e+00
dtype: float64

KPSS: The data 'volume_price' is STATIONARY

 t-score                  0.088121
p-value                  0.100000
# lags used             14.000000
critical value (10%)     0.119000
critical value (5%)      0.146000
critical value (2.5%)    0.176000
critical value (1%)      0.216000
dtype: float64
```

Fig. 9. Descriptive Statistics for 'volumeprice' after log differencing

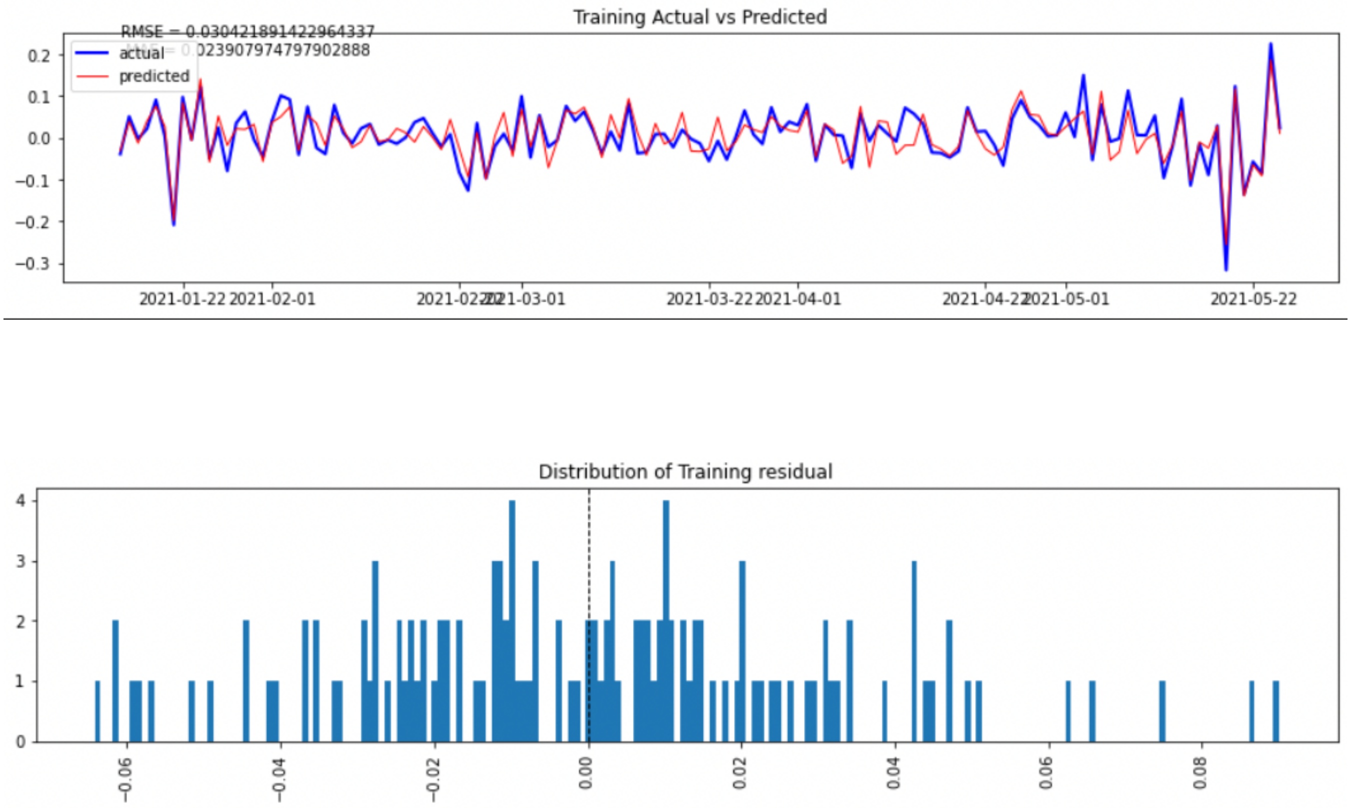


Fig. 10. Training results

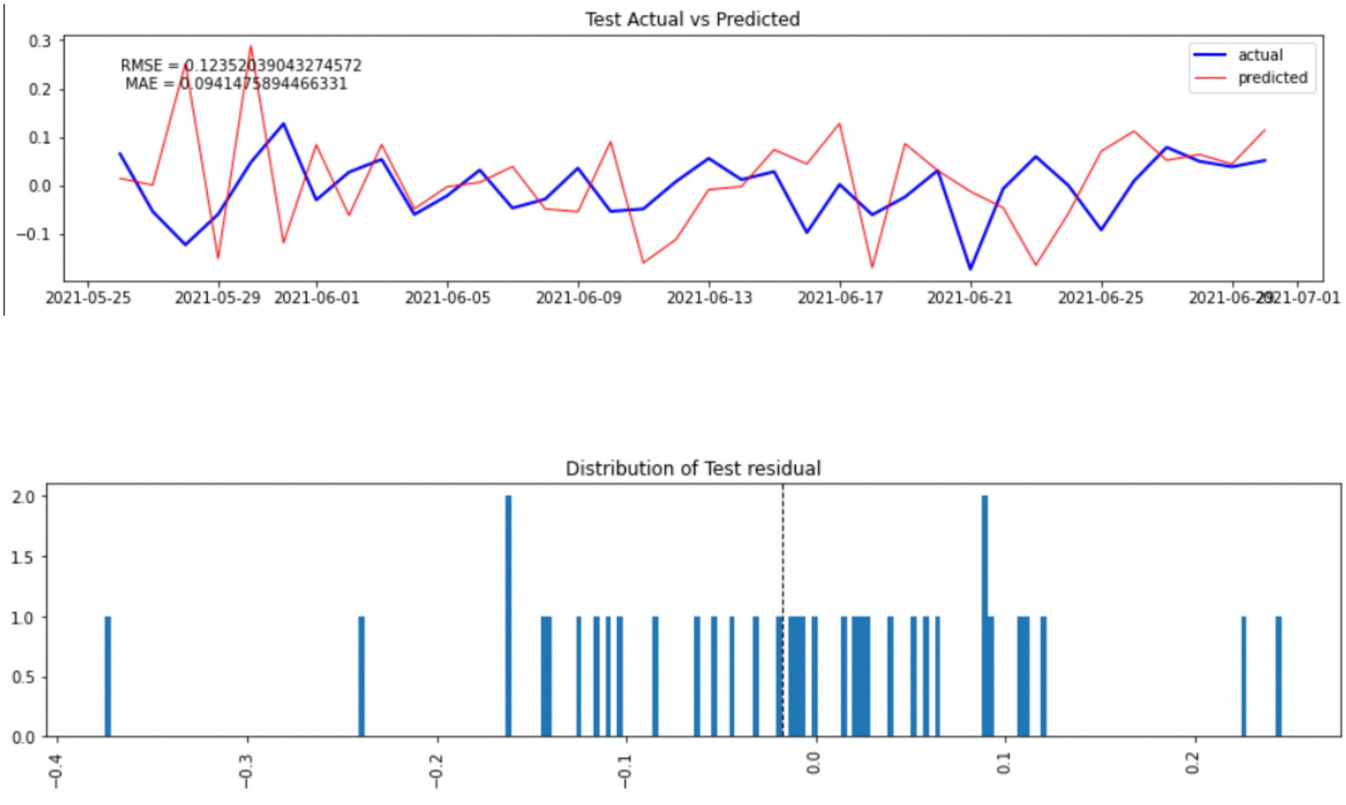


Fig. 11. Test results

statistic test and the p-values. The KPSS test, short for Kwiatkowski-Phillips-Schmidt-Shin (KPSS), is a sort of Unit root test that determines if a given series is stationary around a deterministic trend. To put it another way, the exam is comparable in spirit to the ADF test. The null hypothesis of the KPSS test is that the series is stationary, which is a fundamental distinction from the ADF test. So, in practice, p-value interpretation is the opposite of each other. That is, if the p-value is greater than the significance threshold (for example, 0.05), the series is non-stationary. In the ADF test, however, it means the tested series is stationary.
Many times, ADF and KPSS can give conflicting results. if so: [adf = stationary], [kpss = stationary] = series is stationary [adf = stationary], [kpss = NOT stationary] = series is difference stationary. use differencing to make it stationary [adf = NOT stationary], [kpss = stationary] = series is trend stationary. remove trend to make strict stationary [adf = NOT STATIONARY], [kpss = NOT STATIONARY] = series is not stationary For our Analysis we run create functions such as descriptive statistics ad ADF test and KPSS test to check for stationarity. these pythons and run the test for the price column and we obtain results as NOT STATIONARY for ADF and STATIONARY for KPSS, so we need to remove trend to make it strictly stationery
Fig1 and descriptive stats To remove the trend, we use Log and Differencing method, Log returns, or the difference in log over a time period, is the same as taking a log of data and then differencing it. For good reason, log returns are often used in finance. There are several significant distinctions between log returns and scaled percent returns. Scaled percent returns are not additive, but log returns are. To put it another way, the total of five one-day log returns equals the five-day log return. Furthermore, log returns are balanced in terms of profits and losses, whereas scaled percent returns are skewed toward gains. For example, if a stock drop 50Fig 2 and descriptive stats Once we obtain the Stationarity for Price, we check for the all the three types of comment positive, negative and neutral and find out that they are Stationary in our data the same can be said for the volatile column in our data, but Volume price and Volume were failing the Stationarity Hence we treated them and made them stationary. Same methods were applied shown in the figures.

Modelling

Vector autoregression, which is utilized for multivariate time series, was the model we used to forecast my data. We compare our fitted model's findings to test data as well as to a persistent model.

The vector auto-regression (VAR) time series model has a wide range of applications in econometric forecasting; VAR can capture the evolution and interdependencies of several time series. All the variables in a VAR are addressed symmetrically by incorporating an equation for each variable that explains its evolution using its own lags as well as the lags of all other variables in the model. This might be referred to as a scientific way for trading strategy. Advantage of using the scientific method for trading strategy design is that if the strategy fails after a prior period of profitability, it is possible to revisit the initial hypothesis and re-evaluate it, potentially leading to a new hypothesis that leads to regained profitability for a strategy.

We now must split the sample into training and validation sets. In time series we must be careful with this because we cannot simply randomly select a training and testing set, because of the time dependence. In practice you may need to estimate through a moving window, or you train your model using some of the history and then you keep moving forward in order to test your model. Our dataset is given within daily intervals and so we can predict up to a particular number of days.

VAR models (vector autoregressive models) are used for multivariate time series. The structure is that each variable is a linear function of past lags of itself and past lags of the other variables.
As an example, suppose that we measure three different time series variables, denoted by xt,1 and xt,2.
The vector autoregressive model of order 1, denoted as VAR (1), is as follows:
xt1=α1+11xt1,1 + 12xt1,2 + 13xt1,3+wt,1
xt2=α2+21xt1,1 + 22xt1,2 + 23xt1,3+wt,2

Each variable is a linear function of the lag 1 values for all variables in the set. In a VAR(2) model, the lag 2 values for all variables are added to the right sides of the equations, In the case of three x-variables (or time series) there would be six predictors on the right side of each equation, three lag 1 terms and three lag 2 terms. In general, for a VAR(p) model, the first p lags of each variable in the system would be used as regression predictors for each variable.
VAR models are a specific case of more general VARMA models. VARMA models for multivariate time series include the VAR structure above along with moving average terms for each variable. More generally yet, these are special cases of ARMAX models that allow for the addition of other predictors that are outside the multivariate set of principal interest.

We obtain the results from our modeling as shown in figure 10 and figure 11 respectively.

Evaluation

Once we have a model for our data, it's important to analyze how we can evaluate its quality. The first option is to use the residuals. Residuals basically the squared difference between the predicted values and actual values. This could be a simple option to judge the accuracy of our model.
We choose the RMSE and MAE for our modal evaluation

Mean absolute error (MAE)
The MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction. It measures accuracy for continuous variables. The equation is given in the library references. Expressed in words, the MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average.
Root mean squared error (RMSE)
The RMSE is a quadratic scoring rule which measures the average magnitude of the error. The equation for the RMSE is given in both references. Expressing the formula in words, the difference between forecast and corresponding observed values are each squared and then averaged over the sample. Finally, the square root of the average is taken. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE is most useful when large errors are particularly undesirable. The MAE and the RMSE can be used together to diagnose the variation in the errors in a set of forecasts. The RMSE will always be larger or equal to the MAE; the greater difference between them, the greater the variance in the individual errors in the sample. If the RMSE=MAE, then all the errors are of the same magnitude
Both the MAE and RMSE can range from 0 to . They are negatively oriented scores: Lower values are better.

Train: RMSE = 0.03042 MAE = 0.02390
Test: RMSE = 0.1235 MAE = 0.0294

Conclusion

We have made use of the first Half of 2021 data for both the comments and Price of Ethereum. And based on that our model has given acceptable results with low RMSE , we can further expand the study by considering the previous year data which you enable the algorithms to capture more variance and produce better results. This document makes it easier for the stockbroker to see patterns and boosts job productivity. Because stock trading is both time-consuming and diverse, our interface allows users to save time and energy while still making profitable deals. This work may be improved by improving the method used in the project and adding extensions such as machine learning prediction code and code such as neural networks, which can help the application perform better in the future when the user makes trades. To do forecasting, we may also experiment with alternatives such as the ARIMA model or Deep learning (LSTM Models) and assess their performance using diagnostics such as R-squared or RMSE. The existing VAR model could be improved by tuning the parameters in it based on the AIC, The VAR structure above, as well as moving average terms for each variable, are included in VARMA models for multivariate time series. More broadly, these are ARMAX models that enable the insertion of additional predictors that are not part of the multivariate set of major interest. And on the other end the sentiment analyses of the comments can be further improved based on the latest improvement in Machine Learning and Deep Learning fonts.

References

1. E. F Fama. Efficient capital markets a review of theory and empirical work. *The Fama Portfolio, pages 76–121.*, 2017.
2. S Anuradha G Gautam P N V Syamala Rao M, N Suresh Kumar. Bitcoin analysis prediction using var. *Vol. 28, No. 19, (2019), pp. 1141 - 1151*, 2018.
3. N. Park J. Choo J.-H. Kim Y. B. Kim, J. Lee and C. H. Kim. When bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation. *PloS one, vol. 12, no. 5, p. e0177630, 2017.*, 2017.
4. I. Lunesu M. Matta and M. Marchesi. Bitcoin spread prediction using social and web search media. *UMAP Workshops, 2015, pp. 1–10*, 2015.
5. J. Kaminsk. Nowcasting the bitcoin market with twitter signals. *arXiv preprint arXiv:1406.7577, 2014*, 2014.
6. L. Kristoufe. Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era. *Scientific reports, vol. 3, p. 3415, 2013.*, 2013.
7. J. Nelson J. Abraham, D. Higdon and J. Ibarra. Cryptocurrency price prediction using tweet volumes and sentiment analysis,. *SMU Data Science Review, vol. 1, no. 3, p. 1, 2018.*, 2018.
8. R. C. Phillips and D. Gorse. Cryptocurrency price drivers: Wavelet coherence analysis revisited. *PloS one, vol. 13, no. 4, p. e0195200, 2018*, 2018.
9. M. Verma and P. Sharma. Money often costs too much: A study to investigate the effect of twitter sentiment on bitcoin price fluctuation. 2020.
10. F. H. Jahjah and M. Rajab. Impact of twitter sentiment related to bitcoin on stock price returns. *Journal of Engineering, 26(6):60–71*, 2020.
11. O. Kraaijeveld and J. De Smedt. The predictive power of public twitter sentiment for forecasting cryptocurrency prices. *Journal of International Financial Markets, Institutions and Money, 65:101188.*, 2020.
12. Ramon Hinojosa Alejandro. Twitter and reddit posts analysis on the subject of cryptocurrencies. *Tecnologico de Monterrey, Campus Mty.*, 2021.
13. Patrick Canning. Predict bitcoin using reddit. 2018.