Cryptocurrency Trend Prediction Feature Engineering - Project Increment-1

Group 17 - Team Members
Ali Shah
Arsalan Aman
Ping Chun Lee

GitHub link: https://github.com/allenpclee/Group-17---Project-Increment-1

Goals & Objective

In this project, we would like to do the cryptocurrency price prediction. The most popular one is undoubtedly bitcoin, and it went crazy in 2021 as we know. The data is a time series data and we need to apply time series analysis. We would like to apply feature engineering techniques to help us get more useful information and knowledge before we go on our prediction model. There are 2 ways to do the prediction, one is the Recurrent Neural Network, another one is the linear regression model. We'll both part and summarize the result.

Motivation

The motivation behind this project was quite clear. We had seen various examples where neural networks and other models were used to predict stock prices. Stocks were a point of interest since they were widely used and had a steady fluctuation hence they were easier to predict and more accurate. On the other hand, crypto currency is still a new concept and not much work is being done on its prediction.

One major reason is the fact that it is extremely volatile. The haphazard fluctuation within mere minutes is what makes crypto so hard to predict, let alone be accurate. Hence, we are trying to experiment on what types of feature refinement and engineering we can apply to make our model more accurate.

Significance

The need for using feature engineering is important because it can extract necessary knowledge before doing the prediction. Besides, crypto currency prediction is a kind of time series analysis which contains necessary information. We think that applying time series analysis and extracting timing features is a key point to get high accuracy and low test error.

Objectives

In this project, we will do the stock price prediction step by step. First off, we will provide the data visualization such as using charts to depict trends in the data. Then we'll generate the time series data to do the data preprocessing task. After that, the most important step, which is feature engineering, will be used on numerical data and reduce the noise by using the method in the paper[2]. We will build the RNN model and use the long short term memory(LSTM) method. Finally we'll generate their time series prediction to compare with the original data.

Features

We will apply the sequence reconstruction method mentioned in paper[2]. This method can help us find similar motifs in different periods of time. The motif means the similar time series data waves exist in different time periods, which is the feature we would like to get. We will use the min-max scaler to scale our data. Use quantization to visualize which years had the greatest fluctuation,

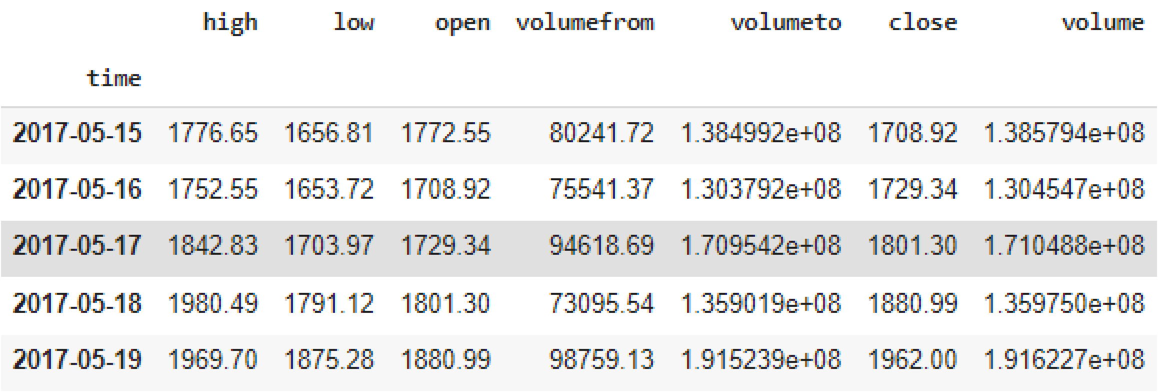
Related works

The time series analysis is the key point to our research, so we looked for the resources related to the knowledge of stock market price prediction and its feature engineering process. In the paper[4], it talked about three major techniques related to the time series, which are stationarity, seasonality, and autocorrelation. Because the trend of cryptocurrency prices changed rapidly, it's definitely a non-stationarity time series analysis. Then we need to focus on its seasonality. This feature indicates that the value changes toward the same direction periodically. We'll reference [3] to do the seasonality check. The autocorrelation is to identify the correlation of data points in different periods of time. We also reference the website in [2]. This author describes how to create new features and find the correlation between the features. Finally, it used the walk-forward method, and we're going to use this technique to predict the Bitcoin prices. We'll also use the technique in [4], which teaches us how to apply LSTM to the time series analysis and model prediction. The LSTM can remember and train the model based on the previous experience by using the memory inside the RNN.

Dataset

The dataset we used is from CryptoCompare website [1]. It provides the daily trading information of Bitcoin from 2017 to 2022. The column information it provides us shows below. We have daily high price, low price, market open price, close price, volume trade-in, volume trade-out, and conversion type and symbols. There are two things we need to do before we go on.

First, we don’t need the information about conversion type and symbol, so we drop these 2 columns. Then we sum up the “volumefrom” and “volumeto” together and add a new “volume” column to store it, because the volume mean the total trading volume, including buying and selling, in a day, which is an important information for us to analyze the data.



	high	low	open	volumefrom	volumeto	close	volume
time							
2017-05-15	1776.65	1656.81	1772.55	80241.72	1.384992e+08	1708.92	1.385794e+08
2017-05-16	1752.55	1653.72	1708.92	75541.37	1.303792e+08	1729.34	1.304547e+08
2017-05-17	1842.83	1703.97	1729.34	94618.69	1.709542e+08	1801.30	1.710488e+08
2017-05-18	1980.49	1791.12	1801.30	73095.54	1.359019e+08	1880.99	1.359750e+08
2017-05-19	1969.70	1875.28	1880.99	98759.13	1.915239e+08	1962.00	1.916227e+08

Detail of design features

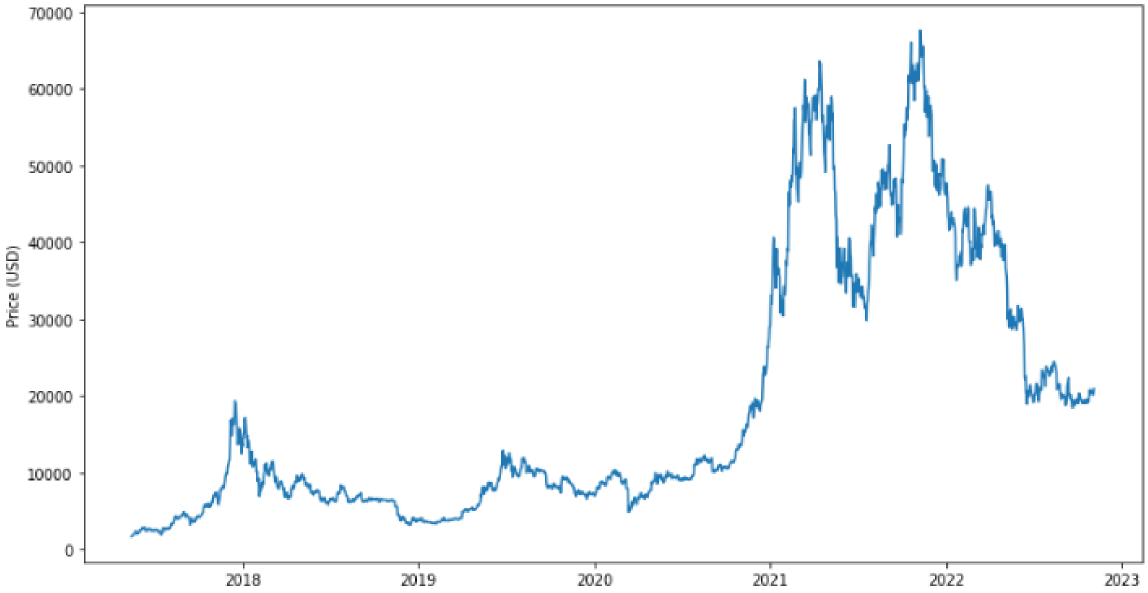
Firstly, we have done the data preparation we need in the previous section. We add the “volume” feature for future reference. Because there’s no invalid data in this dataset, we can just skip this process. Then we need to do the feature engineering to our dataset. The key point to feature engineering is discovery of new knowledge and useful features. In this case, we create 13 features from the volume, open, close, high, and low features in the original dataset. These features are statistical features that are common and useful when doing the time analysis.

Feature 1	Daily close price rate. The ratio of close and open.
Feature 2	Logarithm of volume. To avoid the large value and gap in volume.
Feature 3	Difference in close price everyday.
Feature 4	Difference in closing price of every week.

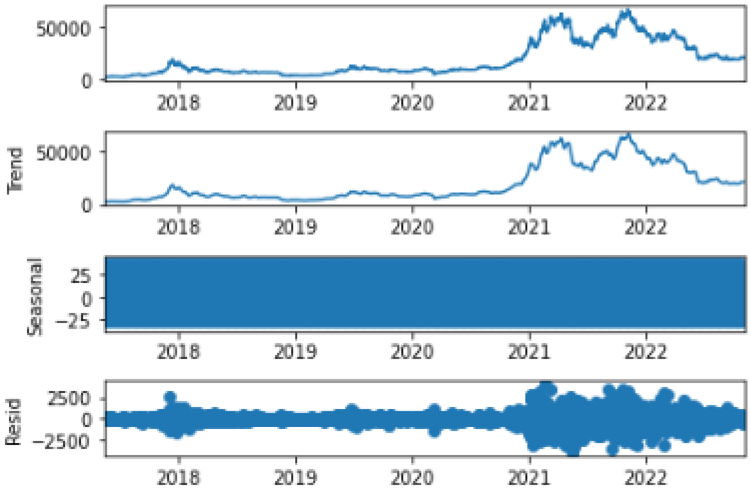
Feature 5	Difference in the closing price of every month. (30 days as approximation)
Feature 6	Change rate for everyday.
Feature 7	Change rate for every week.
Feature 8	Change rate for every month. (30 days as approximation)
Feature 9	Moving average of volume in 7 days.
Feature 10	Moving average of volume in 30 days.
Feature 11	daily volume vs. 200 day moving average.
Feature 12	daily closing price vs. 50 day exponential moving average.
Feature 13	Z-score of close price for every 200 days.

Analysis

This plot below shows the basic daily trend of the cryptocurrency. We can see the trend surge a little bit in 2018, but it went crazy in 2021.

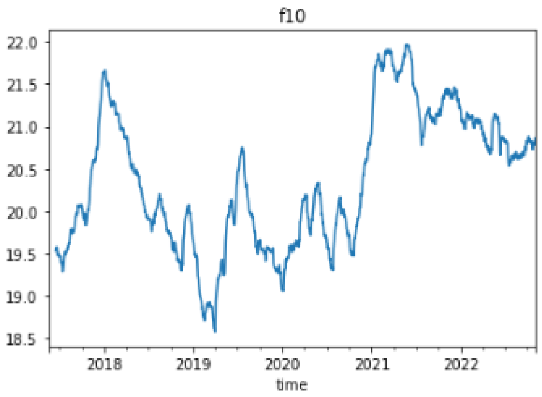


Now we need to analyze if there's a seasonality feature in this price trend. At the first peak at this figure, it's absolutely a non-stationarity trend, and we're not sure about if it's a seasonal trend. To check this feature, we apply the python package from stats.models.tsa.seasonal on the dataset. The result shows below.

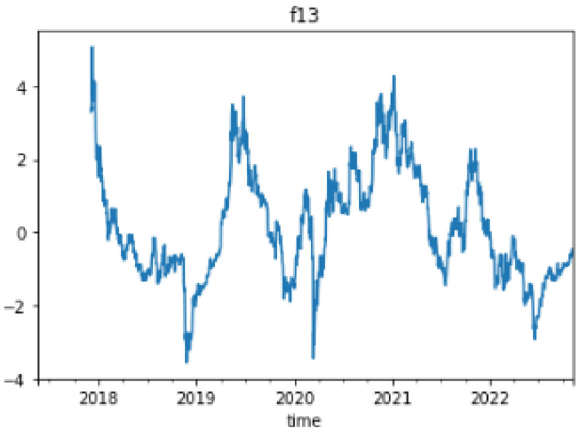


As we can see, there's no sign of seasonality in the price trend of Bitcoin. This makes prediction of the Bit-coin price trend harder. One important thing we learned from researching time analysis and prediction is that we cannot peak the future trend to do the stock price prediction. In order to conquer this difficulty, we need to analyze the correlation and trend on the new statistical features we created in the previous section.

For example, the features 10 shows Moving average of volume in 30 days, and the plot shows below. The trend tells us some peak points in 2018 and 2021, which matched the original price trend.

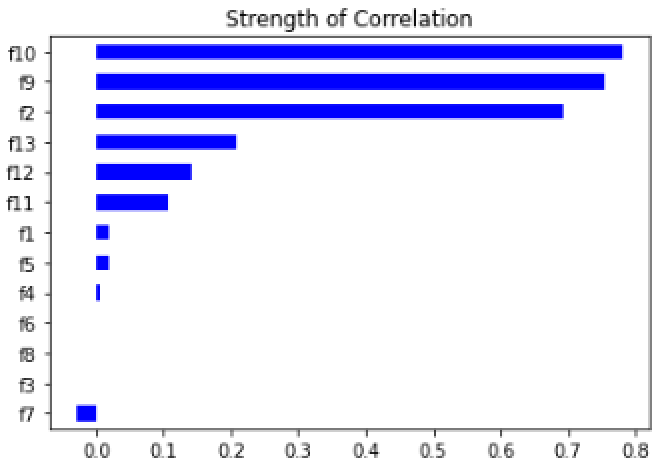


Another example for feature 13 shows the z-score. The plot trend below shows that the Bitcoin price is not stable and it always differs from the std a lot year to year.

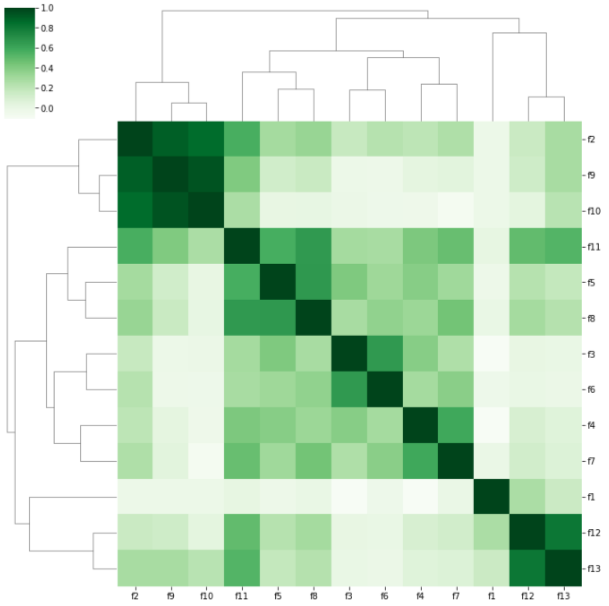


Implementation

According to the analysis above, we can start to find the correlation of these statistical features to the prices. We can also check if there's any feature that has a strong correlation to the prices which can help us do the prediction more accurately. The correlation shows below, we can see there are 6 features shown to have strong correlation to the prices, which are feature 2, 9, 10, 11, 12, 13. Other features have either negative relation or no relation to the prices.

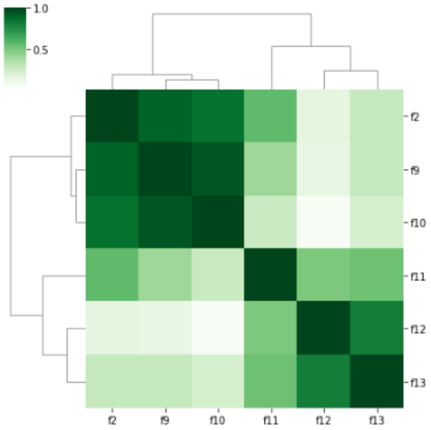


Based on the analysis above, there's one more thing we need to do is to find the relation of each feature. Although we have addressed some features that are good for us to train the model, we still need to remove the features with high correlation to each other. The reason is that we don't need to include similar features in the future prediction, and it can greatly avoid the dimensionality curse. The heatmap tells us the correlation of each feature. The figure below tells us the dark color with high relationship and light color with low relationship.

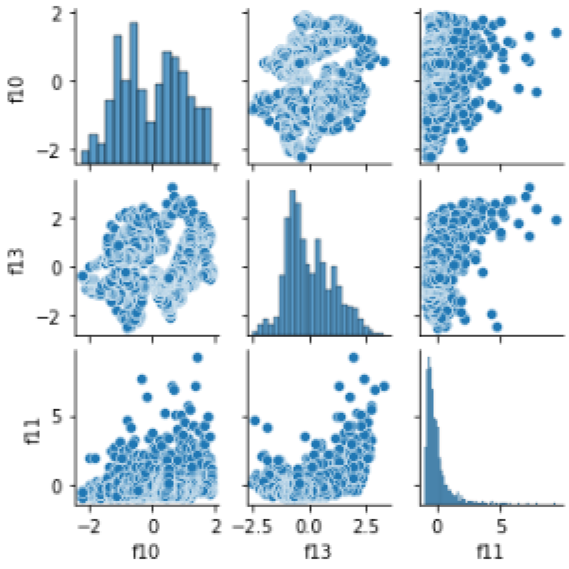


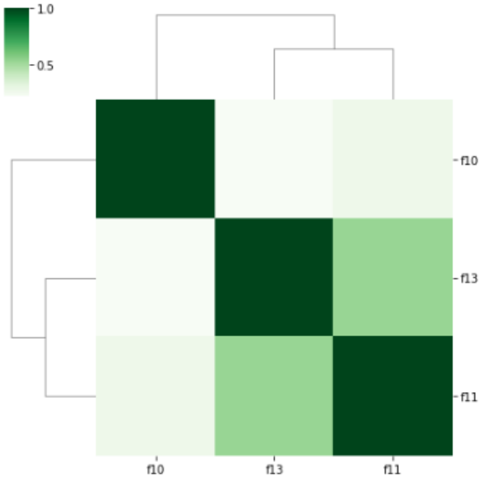
Preliminary result

Now we focus on what we found in the above section, the features 2, 9, 10, 11, 12, 13 are what we need and plot the heatmap again. We can see the figure below, there are features that have a high relationship to each other, such as feature 2 and feature 9. We need to remove them.



We finally choose our key features, which are features 10, 11 and 13. We can draw the scatter plot and heatmap to check their relationship. The 2 figures below tell us the features we select so far are safe and independent for us to train the model.





The key features, 10: log of 30-day moving average of volume, 11: daily volume to 200-day moving average, 13: z-score for every 200-day, are our target to train on the machine learning model. We'll focus on these key features in the next step.

Project management

Work completed

Description:

- We have completed the feature engineering and feature selection part for Bitcoin dataset. We also did the data visualization for the data analysis.
- For the LSTM, we have split the data and research for the model creation

Responsibility :

- Feature engineering: Ping-Chun Lee , Arsalan Aman, Ali Shah
- LSTM model:Ping-Chun Lee , Arsalan Aman, Ali Shah
- Documentation: Ping-Chun Lee , Arsalan Aman, Ali Shah

Contributions:

- Data preparation: Ping-Chun Lee (40%) Arsalan Aman (30%) Ali Shah (30%)
- Data analysis: Ping-Chun Lee (40%) Arsalan Aman (30%) Ali Shah (30%)
- Model Implementation: Ping-Chun Lee (30%) Arsalan Aman (35%) Ali Shah (35%)
- Feature engineering: Ping-Chun Lee (40%) Arsalan Aman (30%) Ali Shah (30%)
- Documentation: Ping-Chun Lee (35%) Arsalan Aman (35%) Ali Shah (30%)
- Video : Ping-Chun Lee (33%) Arsalan Aman (33%) Ali Shah (33%)

Work to be completed

Description:

- We need to finish the walk-forward modeling method and do the prediction and evaluation.
- For the LSTM, we need to finish the model creation and prediction.

Responsibility :

- Walk-forward modeling and evaluation: Ping-Chun Lee (30%) Arsalan Aman (35%) Ali Shah (35%)
- LSTM model: Ping-Chun Lee (30%) Arsalan Aman (35%) Ali Shah (35%)

References

[1] *CryptoCompare*. (n.d.). Cryptocurrency API, Historical & Real-Time Market Data | CryptoCompare. Retrieved November 4, 2022, from https://min-api.cryptocompare.com

[2] Gray, C. (2018, July 12). *Stock Prediction with ML: Feature Selection — The Alpha Scientist*. The Alpha Scientist. Retrieved November 4, 2022, from https://alphascientist.com/feature_selection.html

[3] Hayes, S. (2021, June 7). *Finding Seasonal Trends in Time-Series Data with Python*. Towards Data Science. Retrieved November 6, 2022, from https://towardsdatascience.com/finding-seasonal-trends-in-time-series-data-with-python-ce10c37aa861

[4] Valkov, V. (2019, April 25). *Cryptocurrency price prediction using LSTMs | TensorFlow for Hackers (Part III)*. Curiously. Retrieved November 4, 2022, from https://towardsdatascience.com/cryptocurrency-price-prediction-using-lstms-tensorflow-for-hackers-part-iii-264fcdbccd3f