Time-Series Forecasting Ethereum Prices

Business Understanding
- Ethereum is a decentralized, open-source blockchain with smart contract functionality. Its adoption in the financial world has grown exponentially over the past few years, and as a result, its price has skyrocketed
- Due to the uncertainty of decentralized finance, or DeFi, the value of Ethereum is highly volatile, which makes the use of traditional forecasting models such as ARIMA difficult
- The investment potential of Ethereum is unprecedented, and many different trading strategies are possible to take advantage of this. For this project, day-trading will be the trading strategy of focus.

Goal and Audience

- The goal of this project is to create a model that predicts prices for the Ethereum asset for the next day utilizing a machine learning model.
- The target audience for this project is an investment firm or an institutional invester that shows interest in day trading the Ethereum asset.

Data Source

This data was scraped from CoinMarketCap.com using the webscraper Octoparse. The webpages used ajax syntax for the "load page" button, and therfore ajax timeout time needed to be applied in order to properly extract the data. This data is only concerned with Ethereum, and no other coin or blockchain.

Features

The data includes the following features:

1. Open
2. High
3. Low
4. Close
5. Volume
6. Market Cap

This dataset provides a timeline of Ethereum prices and related data from August 7th, 2015 to June 8th, 2021.

Visualizations

`![image info](Visuals/Histograms.png)`

- High occurrence of low prices

- Very low occurrence of high prices

- Distribution of prices shows that the price remained relatively stationary, then spiked briefly several times

`![image info](Visuals/acf_plots.png)`

- Using the ACF and PACF plots shown above, we can conclude that an optimal value for p for an ARIMA model would be 1, and an optimal value for q for an ARIMA model would be 1 as well. You can identify this by the exaggreated correlation at the corresponding lag values

`![image info](Visuals/monthly_price_trends.png)`
- The visual representations of the monthly price movement shows no particular seasonal trend. There is no season that seems to have more activity than the others across the years.

EDA

 - The rolling averages calculated from three different windows (30, 90, 365) provide some more insight to the data. As the window increases in size, the rolling averages' values have very different values during the highly volatile periods of the price of Ethereum. This volatility resulted in each of these periods having wildly different minimum and maximum values, which results in rolling averages that also different by quite a lot. Unsurprisingly, the 30-day and 90-day rolling averages were the most closely related, especially during the first period of steep upwards trend. The prices did not reach magnitude differences during these windows that warranted such a drastic rolling average difference. However, at the end of our time period, the rolling averages end up differing in value by almost $500, which goes to show the extreme volatility that Ethereum experienced during this time period (the most recent months when Ethereum had a meteoric rise). In short summary, the 365-day moving average had the lowest average value because it generalized the most volatility, however its final value was very below the true price. The 30-day moving average had the highest value because it strongly accounted for the high volatility, and its final value was a little higher than the true price (the extreme upper values pulled the average upwards). The 90-day moving average was the closest to the true price, showing that it both accounted for and generalized the volatility the best of the three windows!

- From the year 2015 to the first quarter of 2017, the price of Ethereum remained quite stationary, with a very strong rise starting between March and April, which led to a strong upwards trend that lasted throughout the rest of the year of 2017, bring the price to a maximum value of 826.82 by the end of the year. This constituted a 10,106 percent price increase from the minimum price of 8.17 in the year of 2017, which is by all standards a very strong upwards trend. The volume of trades also followed this trend quite closely, matching the sentiment idea that as an asset shoots up in price, more people attempt to join in on the ride, and hence more trades are made. After the year 2017, the price of Ethereum immediately started a strong downwards trend beginning in January of 2018, and by the end of 2018 the price had settled to a minimum value of 84.30, roughly a 94 percent drop from its all time high at the very beginning of 2018. Volume for the rest of 2018 remained on average higher than the two years afterwards and the year before because at first people were participating in frequent trades due to the meteoric rise in price, and then people continued to sell their coins over the year as the price tanked. From 2019 to mid-2020, the price once again mostly resumed the stationary trend that it had exemplified from 2015 to about a quarter of the way through 2017, indicating that perhaps people lost interest in the Ethereum block-chain, doubted its potential, or simply moved on to different investments. There was a sharp rise in prise to a little over $250 during 2019, but it just as quickly fell back to close to the minimum value of that year, failing to breakout of its strong downwards trend. The volume from 2019 to mid-2020 would never drop to the levels seen before the coin's meteoric rise, most likely because such a note-worthy event put Ethereum on the map permanently. During 2019, there was a sharp rise and fall in volume that mirrored the trend of the quick rise and fall of price during that year. 2019-2021 would be the period of time when Ethereum would consistently reflect a yearly upwards trend. Volume was higher than its ever been, and the price rose to an unprecedented level of roughly $4000. During this upwards trend, there were several downwards trends that occured during certain months of the years. They seemed to be relatively random, with no predictability in their occurences, highlighting the unstationarity of the price of Ethereum, and also the idea that the price follows a "cyclical trend". There are very clear bull and bear markets, however the trickly part is timing these.

Stationarity

- The stationarity of the data was tested using an Augmented Dickey-Fuller Test. Before differencing, the ADF-Test gave a result that indicated that the data was strongly non-stationary. After first-order differencing, the ADF-Test gave a result that indicated the data was now stationary.

Modeling

Random-Walk
- RMSE = 323.097
- The best model constructed using Random-Walk had a relatively decent RMSE value, with only one model performing better; the One-Step-Ahead model.
`![image info](Visuals/random_walk.png)`

One-Step-Ahead
- RMSE = 72.12
- The One-Step-Ahead Forecast was very accurate! It was the most accurate of all of the models created.
`![image info](Visuals/zoomosa.png)`

ARIMA/AUTO-ARIMA
- RMSE = 915.550
- Using the calculated ACF and PACF graphs, along with the ADF-Test, the p, d, and q values of 1, 1, 1 were maually selected for the ARIMA model beore using AutoArima for automatic optimazation. The ARIMA model performed poorly for the data provided. This can almost certainly be attributed to the exaggerated volatility of Ethereum prices. Even when using the AutoArima package to optimize the values of p,d, and q, the model did not improve.
`![image info](Visuals/arima.png)`

SARIMAX
- RMSE = 658.58
- The SARIMAX model was run through a gridsearch in order to try and optimize its hypyer parameters
- With the optimization of the hyperparamters, an RMSE was achieved that was a significant improvement to the ARIMA model
`![image info](Visuals/sarimax.png)`

LSTM
- RMSE = 302.61
- For the LSTM models, I used a total of 25 epochs, allowing the MSE score to converge towards a value towards the final epochs. At first, a single LSTM layer was used, however the results were very poor, and so I opted to add in two extra LSTM layers to try and improve the results. The loss metric use was Mean-Squared-Error, and the optimizer was the 'adam' optimizer. The addition of the two LSTM layers aided in the model's predictions, and created a model that more closely followed the trend of the actual prices. The best RMSE score achieved was the third best of all the models tested.
- The RMSE of the LSTM model is 302.61, which is significantly better than that of the ARIMA and SARIMAX. However, it did not beat out the one-step-ahead model.
`![image info](Visuals/lstm.png)`

Conclusion
- Ethereum is an extremely volatile asset with a lack of seasonality. The best performing model was the one-step-ahead model, followed by the LSTM model, and then the random-walk model. Overall, it seems as if a one-step-ahead model may be the best option for predicting the price of Ethereum on a daily basis. However, when it comes to longer term predictions, the One-Step-Ahead model is not built for optimal predictions, and therefore the LSTM model or another model may be the best option.
- The business goal of this project was day trading, but a more long-term focused goal for the project might be beneficial to a broader audience and consequently be more useful.

Further Ideas
- Calculate profit as a more broadly understandable metric for the business audience of this project
- Improve model's through use of exogenous variables
- Test other model's for their effectiveness in longer term predictions, expanding the breadth of the project