Capstone Proposal Template

Topic: Predicting the Price of Ethereum Using On-Chain Metrics and Data Science Techniques

Business Understanding

Over the past several years, cryptocurrencies have been gaining a lot of attention as an emerging asset class. One of the biggest concerns facing investors at the moment is that the prices are known for being highly volatile. It is not uncommon to see massive price fluctuations in a single day, or in some instances, even just a few hours. Unlike stocks, the market for cryptocurrencies is open 24/7 and since they are built using blockchain technology, investors have complete visibility into what happens within the network. The aim of this project is to use on-chain metrics in an attempt to predict the future price of Ethereum through the use of statistical modeling and machine learning.

Data Understanding

There are several metrics I will be utilizing for my analysis and modeling which will be collected from glassnode.com. Glassnode allows you to extract various on-chain metrics such as the number of active wallet addresses, circulating supply, and the balance being held on exchanges vs. in cold storage as individual .json files. The target variable here is the price of Ethereum.

Data Preparation

Each of the variables in this data set are numerical. For pre-processing, I will be merging the .json files together into a single dataframe using the 'date' column. Since the cryptocurrency industry is relatively new and constantly evolving, certain metrics have more historical data than others. In order to ensure that everything is being captured in the final data set, null values will be dropped which leaves me with ~1,800 records to use for modeling purposes.

Modeling

The target variable for this project will be the price of Ethereum 30 days out. Since I will be working with time series data, I will build out several models including Linear Regression (baseline), XGBoost, LSTM (RNN), ARIMA and SARIMAX. The final product will be based on the best performing model.

Evaluation

Considering that error metrics in time series forecasting can be a bit misleading, there are several metrics that I would like to use in combination to assess the results of my models. The metrics I would like to focus on include the Mean Absolute Error (MAE), Median Absolute Error (MedAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Median Absolute Percentage Error (MDAPE). As an MVP, I would like to include a baseline model, as well as at least one more advanced model. To improve my project between the MVP and presentation, I would like to incorporate several other advanced models to be able to assess which strategy produces the optimal outcome.