Cryptocurrency Analysis and Forecasting

Group 7: Arnav Ashank, Atul Mohan Kamble, Pooja Satish Wankhede and Rachit Rahul Meshram
School of Engineering and Applied Sciences, University at Buffalo
EAS 503: Programming and Database Fundamentals for Data Scientists

Abstract

The current trend in the era of investment is trading in the cryptocurrency market. In the last couple of years, the cryptocurrency market went from being speculation to amassing enormous volumes of traffic of currency being traded every day. Even after being such a popular method of investment, there is still a lack of prediction methods for the extremely sensitive and volatile cryptocurrency market.

Introduction

Using this project, we plan to incorporate the parameter of prediction for the price of popular cryptocurrency "Ethereum" usually abbreviated as ETH in the markets. For this, an extremely vast dataset is being utilized which contains information about multiple cryptocurrencies. Time Series Forecasting methods will be used for the predictions.

Problem Statement

- Analyze the data corresponding to Ethereum and find out trends or interesting patterns
- Predict the closing price of ETH using Time Series Forecasting

daily_listing
timestamp
Asset_ID
Count
Open
High
Low
Close
Volume
VWAP
Target

daily_listing_agg
timestamp
Close

Fig1: Database schema structure

- The database schema consists of two tables: Dailylisting and Asset Details.

Data Information

The dataset downloaded from Kaggle contains information on historic trades for several crypto assets, such as Bitcoin and Ethereum. The dataset training file consists of 10 columns. The description of the columns is as below:

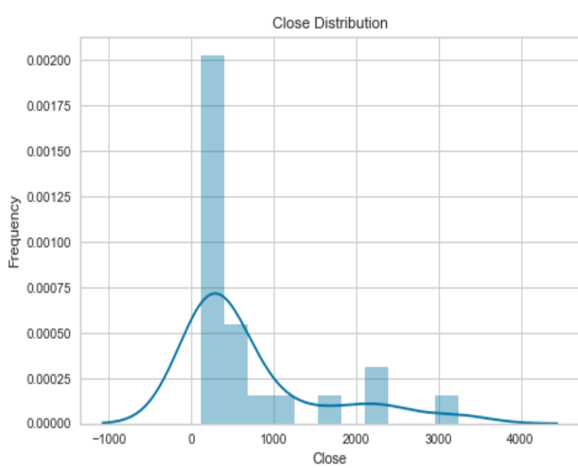
- timestamp - A timestamp for the minute covered by the row.
- Asset_ID - An ID code for the crypto asset. For eg. asset_id == 6 for ETH
- Count - The number of trades that took place this minute.
- Open - The USD price at the beginning of the minute.
- High - The highest USD price during the minute.
- Low - The lowest USD price during the minute.
- Close - The USD price at the end of the minute.
- Volume - The number of crypto asset units traded during the minute.
- VWAP - The volume-weighted average price for the minute.
- Target - 15 minute residualized returns

Database Operations

we have created a database called 'crypto' and dumped our dataset into table 'Dailylisting', then fetched two columns from table 'timestamp' as one of the features for prediction and 'Close' a target variable. Then performed predictions using time series algorithms on the fetched data.

Dataset Analysis

Distribution of Ethereum Data:



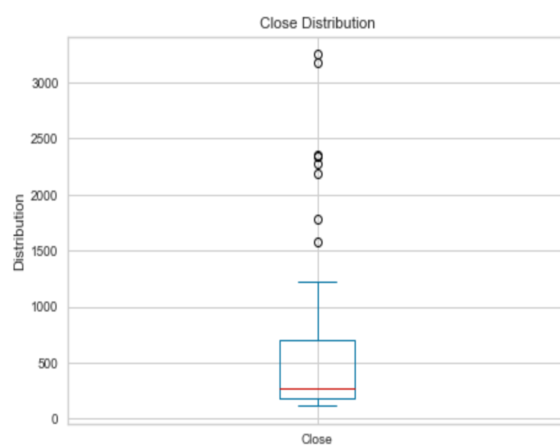


Fig2: Distribution of Ethereum Data

Cryptocurrency being as volatile as it is, we can see there is a lot of variances in the closing price of ETH. The data is right-skewed. This is further shown in the boxplot which has multiple outliers.

Visualization of Closing Price of Ethereum Time Series from 2018 to 2020:

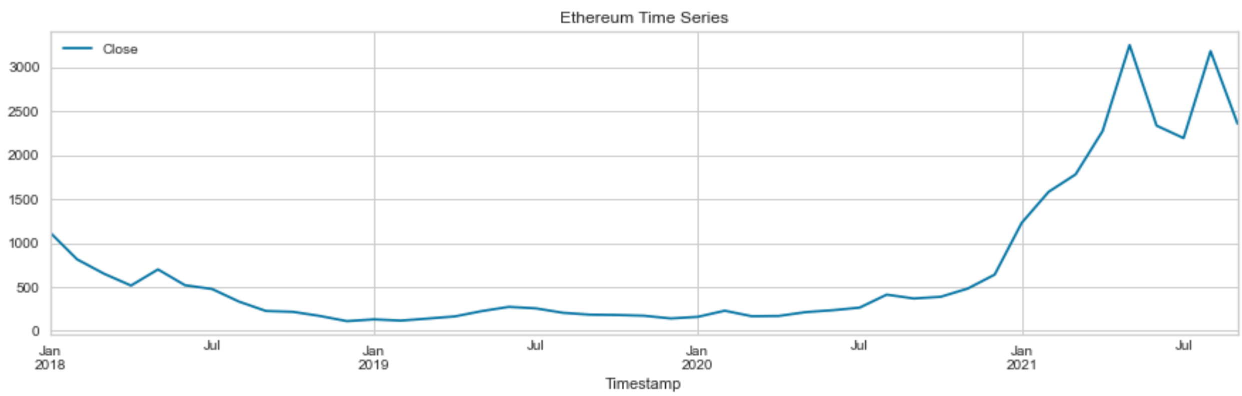


Fig3: Ethereum visualization from 2018 to 2020

The above graph shows the trend of ETH closing prices. We can see stable and low-lying values of the closing prices of ETH up until December of 2020 but see an exponential increase right after the end of the year 2020. This can be the cause of multiple external factors such as the COVID-19 Pandemic or sudden public uproar about cryptocurrency in general.

Results

- We first performed ETS (Error-Trend-Seasonality) decomposition on the time series data to check for trend, seasonality in the data. As it is evident from the graph, seasonality and trend are present in our data. Seasonality is found to be 12 months.
- We used pycaret package to build a time series model on the dataset, made predictions three time periods ahead (3 months ahead price of Ethereum), and compared their performance on the time series data. The two models used for prediction were exponential smoothing and auto_arima.
    - Exponential smoothing: Exponential smoothing is a technique for smoothing time series data using the exponential function. In contrast to simple moving average methods where the past observations are weighted equally. The farthest observations are assigned the least weights and observations close to present are assigned more weights.
    - Auto Arima: ARIMA, short for 'Auto-Regressive Integrated Moving Average' is a class of models that 'explains' a given time series based on its past values, that is, its lags and the lagged forecast errors, so that equation can be used to forecast future values.

- Any ‘non-seasonal’ time series that exhibits patterns and is not a random white noise can be modeled with ARIMA models.
- An ARIMA model is characterized by 3 terms: p, d, q. p is the order of the AR term, q is the order of the MA term, d is the number of differencing required to make the time series stationary
- The term ‘Auto Regressive’ in ARIMA means it is a linear regression model that uses its lags as predictors.

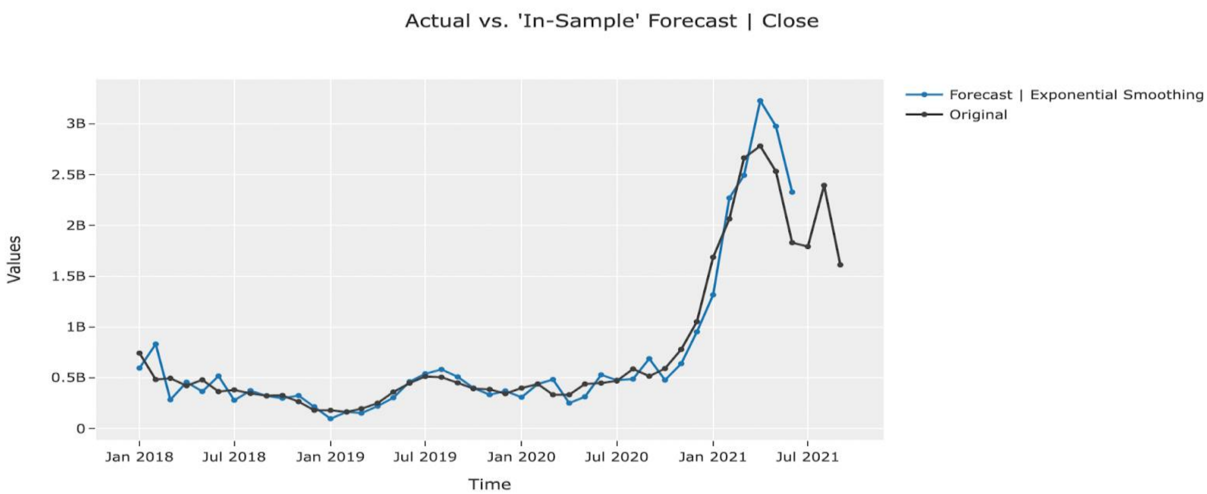


Fig4: Actual vs. ‘In-Sample’ Time Series Forecasting

Conclusion

We compared the metrics of the two models and found that **exponential smoothing** has much lower RMSE, MAE as compared to the auto_arima model. The exponential smoothing model predicted the prices which were quite following actual prices.
We have also used Streamlite which is an open-source framework to build web apps to visualize time-series decomposing which helped us to verify the various trends and volatility of Ethereum. We have currently performed a univariate analysis. In the future, we can explore multivariate analysis and can explore more algorithms that can give us better results in terms of future price predictions.

References / Useful Links:

- Dataset Link: https://www.kaggle.com/c/g-research-crypto-forecasting/data
- What is Cryptocurrency? https://en.wikipedia.org/wiki/Cryptocurrency
- What is Time Series Forecasting? https://en.wikipedia.org/wiki/Time_series
- Exponential Smoothing: https://en.wikipedia.org/wiki/Exponential_smoothing
- UBBox link to code: https://buffalo.app.box.com/folder/150455294788