BITCOIN PRICE AND TRADE STRATEGY PREDICTION

BAN 5600 - 01 ADVANCED BIG DATA COMPUTING AND PROGRAMMING

DR. JIN FANG

July 19, 2025

Sri Navya Guduru
Karan Chaudhary
Sayara Malla Thakuri
Swetha Pothu

Contents

Introduction

This research analyzes the challenging task of predicting the price variations of Bitcoin, a crucial component in the ever-changing world of cryptocurrencies. Using historical data that includes minute-by-minute updates on key market indicators from January 2012 to March 2021, the goal is to build prediction models that can foresee the erratic fluctuations ikan the price of Bitcoin.

Since its introduction in 2009 under the pseudonym Satoshi Nakamoto, Bitcoin has challenged established monetary systems and become a disruptive force in the financial industry. Its decentralized structure, supported by blockchain technology, transforms how we view and interact with money. Because Bitcoin price swings are unexpected, strong predictive models are required, which is why this research is important. Accurate insights are essential for academics, traders, and investors to manage and profit from the constantly shifting Bitcoin ecosystem.

The literature assessment includes foundational texts that illustrate the remarkable financial rise of Bitcoin. Nakamoto's seminal article from 2008, which described Bitcoin as a peer-to-peer electronic cash system, set the foundation for its revolutionary possibilities. Further research was prompted by studies conducted by Baur and Lucey (2015) that examined Bitcoin's status as a haven asset. The research conducted in 2016 by Ciaian et al. explored the complex economics influencing the formation of Bitcoin's price. Simultaneously, Kristoufek (2015) discovered key factors influencing the dynamics of Bitcoin's price, which made a substantial contribution to our understanding of cryptocurrency marketplaces.

There is a significant lack of integration between in-depth historical data analysis and advanced prediction modeling that is tailored to the price fluctuations of Bitcoin in this vast body of literature. By reliably forecasting price fluctuations using state-of-the-art analytical approaches, our research initiative seeks to close this gap. By doing this, it hopes to make a significant contribution to the rapidly developing fields of data-driven decision-making and cryptocurrency research.

In keeping with the fundamentals of the Advanced Big Data Computing and Programming course, this study combines sophisticated data analysis techniques with actual financial datasets. It captures the spirit of using data-driven insights to predict and understand the intricacies present in cryptocurrency markets, contributing to the class's learning objectives by providing real-world examples and insights into the finance industry.

Data Description

Source and Selection of Data: Our project leverages a unique dataset procured from Kaggle as in Appendix C.1; a platform acclaimed for its rich repository of data in various domains. The dataset we have chosen is specifically tailored to capture the nuances of Bitcoin's trading patterns, encapsulated in a well-structured CSV file format. This dataset is not just a collection of numbers but a narrative of Bitcoin's journey in the financial market over nearly a decade.

Temporal Span and Historical Context: Spanning from January 2012 to March 2021, the dataset offers a historical lens through which we can view the evolution of Bitcoin trading. This time frame is particularly significant as it covers a range of market conditions - from the nascent stages of cryptocurrency trading to periods of peak interest and market turbulence. Analyzing data across these varied market conditions provides a holistic view of cryptocurrency's behavior.

Granular Data Composition and Attributes:

- **Time-Stamped Records**: Each entry in the dataset is time-stamped with Unix time, marking the beginning of a 60-second trading interval. This high-frequency data allows for minute-by-minute analysis of market movements.
- **OHLC (Open, High, Low, Close) Metrics**:
    - The *Open Price* sets the stage for the trading window, offering insights into market opening sentiment.
    - The *High Price* within each interval highlights the peak of market optimism or bullish trends.
    - Conversely, the *Low-Price* points to the troughs of market activity, signaling bearish trends or sell-off pressures.
    - The *Close Price* provides a closure to the window's trading narrative, reflecting the market's final stance for that interval.
- **Volume Insights**:
    - *Bitcoin Volume* (Volume in BTC) mirrors the intensity of trading activity, serving as a barometer for market engagement.
    - *Fiat Currency Volume* (Volume in Currency) gives a perspective on the monetary scale of transactions, enhancing our understanding of market capitalization.

- **Weighted Bitcoin Price**: The VWAP (Volume Weighted Average Price), pivotal in this dataset, merges price with volume, offering a more balanced view of market valuation throughout the trading window.

Data Quality and Integrity Measures:

- Our dataset is meticulously curated to address data quality, with NaN values strategically placed to signify non-trading intervals. This aspect is crucial for analyzing market inactivity or pauses in trading.
- A thorough deduplication process has been employed, ensuring the uniqueness and authenticity of each data point, thereby fortifying the foundation of our analysis.

Dataset Dimensions and Technical Specifications:

- The dataset's vastness is evident, encompassing over 4.8 million rows and 8 distinct columns, making it a repository of extensive market data.
- The data types are thoughtfully chosen, with seven columns formatted as float64 for precision in price and volume measurements, and the timestamp column as int64 for accurate chronological analysis.
- Occupying approximately 296.5 megabytes of memory, the dataset's size is testament to its comprehensive nature and the depth of information it holds.

Plan for Detailed Statistical Examination:

- Our approach includes an elaborate statistical analysis to derive key metrics like mean, median, and mode, providing insights into the typical market conditions.
- We will delve into measures of variability and distribution such as standard deviation, variance, skewness, and kurtosis, to fully understand the range and nature of market movements.
- The analysis will also encompass advanced time-series examination to decode trends, cyclical behaviors, and potential seasonal patterns inherent in the Bitcoin market.

Dataset Schema Overview

Our analysis is predicated on a dataset with a well-defined schema that encapsulates key trading metrics of Bitcoin. This schema is composed of the following fields, each tailored to represent a specific attribute of the trading data:

- **Timestamp**: Stored as an integer, this field serves as a chronological marker for each data point, allowing for precise temporal analysis. The 'nullable = true' attribute indicates that there may be instances with missing timestamps, which will require careful handling during preprocessing.
- **Open, High, Low, Close Prices**: These fields are stored as double-precision floating-point numbers, reflecting the various price points of Bitcoin trading within defined time windows. The 'nullable = true' attribute for these fields suggests the potential for missing values in the dataset.
- **Volume (BTC) and Volume (Currency)**: These double-type fields capture the trading volume in Bitcoin and its corresponding fiat currency value within each time window. Their 'nullable' status indicates that there might be periods without trading activity.
- **Weighted Price**: Also, a double-type field, it provides the volume-weighted average price, offering a more nuanced view of the market price that accounts for the volume of trade.

Statistical Summary of the Dataset

The statistical summary as in Table A.1 provides an overview of the dataset's numerical characteristics over 3,613,769 data points, which represents a substantial subset of our entire dataset:

- **Mean Values**: The average (mean) values for the Open, High, Low, Close, Volume (BTC), Volume (Currency), and Weighted Price provide us with a central tendency of each metric. The mean prices range around 6,000 to 6,600 units, with the mean volume of Bitcoin trades at approximately 9.32 and the corresponding currency volume at around 41,762 units. The average weighted price closely mirrors the mean price values, suggesting a relative consistency in trading prices across the sampled time windows.
- **Standard Deviation**: The standard deviation for each price metric is significant, indicating an elevated level of volatility and variation in Bitcoin's price over time. A

similar observation can be made for the trading volume, which underscores the variable nature of Bitcoin trading activity.

- **Minimum and Maximum Values**: The minimum values for the prices and volume indicate that there have been low activity periods in the market, with prices and volumes close to zero. Conversely, the maximum values highlight the peaks of market activity and price spikes that Bitcoin has experienced.

Data Analysis

The correlation plot shown in Fig B.1 and Table A.2 between 'Open,' 'High,' 'Low,' and 'Close' is remarkably close to 1, indicating a strong positive correlation. This is expected, as these values are typically derived from the same financial instrument's price movements. The correlation coefficients between 'Volume_(BTC)' and other price-related metrics ('Open,' 'High,' 'Low,' 'Close') are negative and relatively small. This suggests a weak negative correlation between trading volume and price. The correlation coefficients between 'Volume_(BTC)' and 'Volume_(Currency)' as well as 'Volume_(BTC)' and 'Weighted_Price' are positive. This suggests a positive correlation between trading volume in BTC and trading volume in currency, as well as trading volume and the weighted average price.

Fig B.2 shows Bitcoin Candlestick Chart with Volume, in this each candlestick represents a week of trading. The rectangular "candle" has a vertical line (wick) extending above and below it. The top and bottom of the rectangle represent the open and close prices for the week. The top of the upper wick represents the highest price (High) during the week. The bottom of the lower wick represents the lowest price (Low) during the week. Green candles indicate that the closing price was higher than the opening price, suggesting a bullish week. Red candles indicate that the closing price was lower than the opening price, suggesting a bearish week. Each bar represents the trading volume for a week. The height of the bar corresponds to the total volume of Bitcoin traded during that week.

Monthly Trading Volume is shown in Fig B.3 The 'Timestamp' column is converted to a timestamp type, allowing for easier manipulation of time-based data. A new column 'month' is created by extracting the month from the 'Timestamp' using the date_format function. The data is then aggregated based on the 'month' column to calculate the total monthly trading volume of

Bitcoin. Each bar represents the total trading volume for a specific month. Highest Months is Nov 2013 with 965K and followed by Dec 2013 with 952K and Nov 2015 with 887K.

In Fig B.4 Monthly Volatility Rate is shown for which the 'Timestamp' column is used to extract the month, creating a new column 'Month' formatted as "yyyy-MM." Data is grouped by month. The standard deviation of the 'Close' prices is calculated for each month, representing the monthly volatility.

To analyze the distribution of columns Open, High, Low, Close, Volume_(BTC), Volume_(Currency), and Weighted_Price as shown in Fig B.5-11. The histograms have 30 bins, providing a detailed view of the distribution. There is a considerable spread in price-related features ("Open," "High," "Low," and "Close") with mean values around 6000 and high standard deviations, indicating substantial volatility. The trading volumes ("Volume_(BTC)" and "Volume_(Currency)") vary widely, with mean values of approximately 9.32 BTC and 41,762.84 currency units, respectively. The "Volume_(Currency)" column exhibits a notably higher standard deviation, suggesting significant variability in traded currency. The "Weighted_Price" column, representing the average weighted price, shows variability around a mean of 6008.93.

The Distribution of Capital as in Fig B.12 is a histogram displaying a relatively even distribution, suggesting that the values of capital are spread across different ranges without a pronounced skewness. The concentration around 100,000 could signify a common capital threshold or a particular investment strategy prevalent in the dataset. The uniform shape of the histogram suggests that capital values are not heavily concentrated in specific ranges, but rather distributed broadly across various values.

The time series plot of Close Prices in Fig B.13 displays the trend and fluctuations in Bitcoin closing prices over time. The line connects the data points, providing a visual representation of the price movements.

The Daily Closing Price Average Over Time shown in Fig B.14 shows that the plot provides a clear representation of how the average daily closing prices of Bitcoin have changed over time. The line depicts the trend, showing whether the average prices have been increasing, decreasing, or fluctuating.

The scatter plot between close vs volume in Fig B.15 in which each point on the plot represents a specific observation in the dataset, with the x-coordinate indicating the closing price and the y-coordinate representing the corresponding trading volume in Bitcoin. The points on the scatter plot are scattered across the graph, forming a pattern that illustrates the joint variation of closing prices and trading volumes. The downward slope pattern might suggest a negative correlation, indicating that as closing prices increase, trading volumes tend to decrease. This could be indicative of investors reacting to price changes, with higher volumes during periods of price decline. This pattern may reflect specific dynamics in the Bitcoin market, such as profit-taking during price increases or increased trading activity during periods of uncertainty or decline.

The scatter plot of High vs Low as shown in Fig B.16, a diagonal line from the bottom left to the top right suggests a positive linear relationship between high and low prices. This positive correlation indicates that when Bitcoin experiences higher prices, the corresponding low prices during the same period also tend to be higher. Similarly, lower soaring prices correspond to lower low prices.

The 5-Minute SMA of bitcoin price in Fig B.17 is a smoothed representation of the price, calculated by averaging the closing prices over a 5-minute interval. The blue line represents the actual closing prices of Bitcoin at different timestamps. The red line represents the 5-minute Simple Moving Average (SMA) of the close prices. Close Price crossover below the SMA could indicate a bearish signal. The blue line is consistently below the red line, it suggests that the actual close prices are generally below the 5-minute SMA. This configuration might indicate a potential downtrend or a period of lower volatility in the short term. The close prices are, on average, below the smoothed moving average. Peaks in both the Close Price and the 5-Minute SMA suggest a moment of heightened activity. The subsequent exponential rise in both the Close Price and the 5-Minute SMA indicates a period of strong bullish momentum.

In Fig B.18 to plot for showing the changes in capital over time and providing insights into the performance and fluctuations of the account balance. There is a single line on the plot, represented in green, which represents the capital values over different timestamps. By observing the line, we can identify the trends, patterns, or specific points where the capital experiences notable changes. Steep inclines may represent periods of profit, while declines may indicate losses.

To visualize the distribution of close prices categorized by different signals as in Fig B.19 where the boxplot displays the distribution of close prices for distinct categories or levels of a variable, specifically the "signal" variable. It can be interpreted from the plot that on average, the central tendency of close prices remains consistent regardless of the signal. On average, the central tendency of close prices remains consistent regardless of the signal. The presence of many outliers suggests that there are individual close prices significantly higher or lower than the majority within each signal category. These extreme values contribute to the variability in the data. It suggests a high degree of variability in close prices within each signal category.

to visualize the overall profitability over time as shown in Fig B.20, The overall profitability is calculated by taking the capital values over time, dividing by the initial capital, and subtracting 1. The line plot visually depicts how the overall profitability changes over the observed time period. Upward trends indicate periods of positive profitability, while downward trends suggest periods of negative profitability. Peaks and troughs in the line provide insights into the performance of the investment strategy. Positive trends indicate periods of growth, while negative trends suggest periods of decline.

Data Modeling

Random Forest Regression

The first model explored was the Random Forest Regression algorithm, an ensemble learning method that combines multiple decision trees to enhance predictive accuracy. The Root Mean Squared Error (RMSE) for this model was calculated at 1529.92, providing insights into the average prediction error. This model considered six features and comprised 20 trees in its ensemble. The evaluation metric, RMSE, is crucial for understanding the accuracy of predictions, with a lower RMSE indicating better model performance.

Gradient-Boosted Trees Regression

Subsequently, the Gradient-Boosted Trees Regression algorithm was employed, which builds decision trees sequentially to correct errors made by preceding trees. The RMSE for this model was calculated at 1177.55, highlighting a lower average prediction error compared to the Random Forest model. Like its counterpart, this model considered six features for prediction.

The comparison between the two tree-based models revealed that Gradient-Boosted Trees demonstrated superior predictive accuracy.

Linear Regression

Linear Regression, a classical statistical technique, was also applied to model the linear relationship between features and the target variable. Surprisingly, this model achieved an exceptionally low RMSE of 7.09, signifying highly accurate predictions. Like the tree-based models, Linear Regression utilized six features, providing interpretability by revealing linear relationships between predictors and the closing prices of Bitcoin.

Decision Tree Classifier

Additionally, a Decision Tree Classifier was employed, focusing on classifying instances into distinct categories. The model's performance was evaluated using the Area Under the Receiver Operating Characteristic (ROC) Curve, resulting in a value of 0.4599. However, it's important to note that ROC AUC is typically more relevant for classification tasks rather than regression problems.

Best Performing Model

The study identified the Random Forest Regression Model with the unique identifier RandomForestRegressor_36eb5daa5027 as the best-performing model among the ensemble methods. This model, utilizing 20 trees and considering six features, demonstrated favorable results in predicting Bitcoin closing prices.

In conclusion, the data modeling phase revealed varying levels of predictive accuracy among different algorithms. While the tree-based models, particularly Gradient-Boosted Trees, showed promise, the unexpectedly superior performance of Linear Regression suggests the importance of exploring diverse modeling approaches for robust predictions in the volatile cryptocurrency market. Further optimization and refinement of these models may enhance their accuracy and applicability in forecasting Bitcoin prices.

Predicted Vs Actuals

A line plot to visualize the predictions versus actual close prices over time as can be seen in Fig. B.22. By comparing the lines for actual and predicted close prices, we can visually assess the

model's performance in predicting the close prices over time. Ideally, the lines for actual and predicted close prices should closely align, indicating accurate predictions. Differences or deviations between the lines may highlight areas where the model performed well or areas that require improvement. The coincidence of the lines suggests that the model has made nearly perfect predictions. The predicted close prices align closely with the actual close prices. A close match between the lines suggests a high level of reliability in the predictive model. It is effectively capturing the underlying patterns and trends in the time series data.

Conclusion

Key Findings

The analysis brought to light notable variations in the price of Bitcoin, particularly in 2018 and late 2020, which were indicative of big shifts in the market. The analysis of Bitcoin prices revealed a steady increasing trend punctuated by erratic anomalies, providing insight into the intricate mechanics of the market. Except for trade volumes in currencies and Bitcoin, most factors showed strong correlations, suggesting a complex relationship and possible forecasting difficulties. Moving averages showed that there was little to no substantial divergence between the short- and long-term patterns in the price fluctuations of Bitcoin. Using Random Forest and linear regression models showed potential; Random Forest fit the time series data better than the other model.

Business Insights and Recommendations

When compared to more conventional approaches, advanced modeling techniques like Random Forest showed more encouraging findings, indicating its applicability in precise Bitcoin price forecasting. The fact that trade volumes in Bitcoin and other currencies have weaker connections highlights the necessity for complex models to effectively forecast the price of Bitcoin. The successful result of overlapping predictors and test data supported Random Forest's effectiveness in fitting the time series data.

The study's findings highlight the crucial role cutting-edge analytical techniques play in comprehending and predicting Bitcoin values in the context of the market's extreme volatility. As this study shows, using complex models like Random Forest yields better insights and more precise forecasts than using more conventional methods. Additionally, it underscores the

necessity for adaptable trading strategies and real-time decision-making tools to effectively navigate the dynamic fluctuations in Bitcoin prices. All things considered, this study adds a great deal to our knowledge about the mechanics of the Bitcoin market and highlights the significance of using sophisticated models to make accurate price predictions.

Appendix

Appendix A. Tables

summary	Open	High	Low	Close	Volume_(BTC)	Volume_(Currency)	Weighted_Price
count	3613769	3613769	3613769	3613769	3613769	3613769	3613769
mean	6009.02368	6013.357082	6004.488004	6009.013545	9.323249223	41762.8424	6008.9348
stddev	8996.247351	9003.521006	8988.778319	8996.359688	30.54989124	151824.7839	8995.991643
min	3.8	3.8	1.5	1.5	0	0	3.8
max	61763.56	61781.83	61673.55	61781.8	5853.852166	1.39E+07	61716.20534

A.1 Summary Statistics

	Open	High	Low	Close	Volume_(BTC)	Volume_(Currency)	Weighted_Price
Open	1	0.999999	0.999999	0.999999	-0.05186	0.344074	0.999999
High	0.999999	1	0.999998	0.999999	-0.051717	0.344498	0.999999
Low	0.999999	0.999998	1	0.999999	-0.05204	0.343541	1
Close	0.999999	0.999999	0.999999	1	-0.051877	0.344036	1
olume_(BT	-0.05186	-0.051717	-0.05204	-0.051877	1	0.352038	-0.051887
me_(Curre	0.344074	0.344498	0.343541	0.344036	0.352038	1	0.34401
eighted_Pr	0.999999	0.999999	1	1	-0.051887	0.34401	1

A.2 Correlation Matrix

Appendix B. Figures

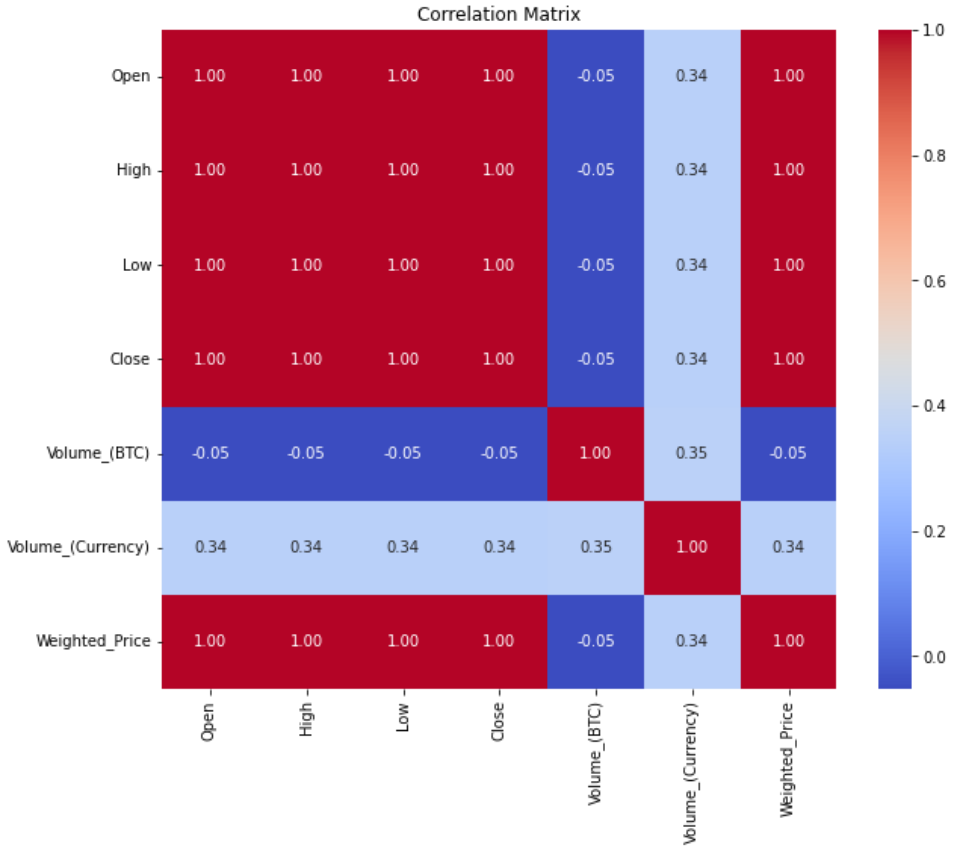

Figure B.1 Heatmap

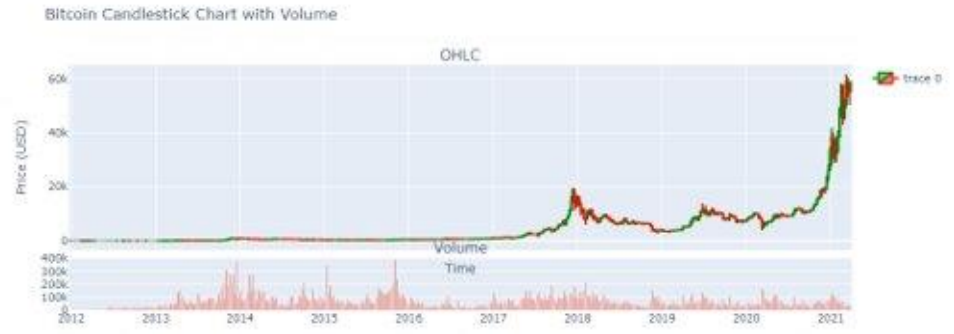

Fig B.2 Candlestick Chart with Volume

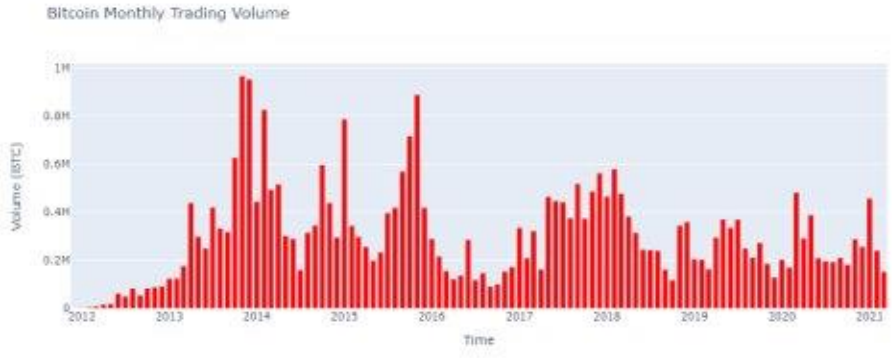

Fig B.3 Monthly Trading Volume

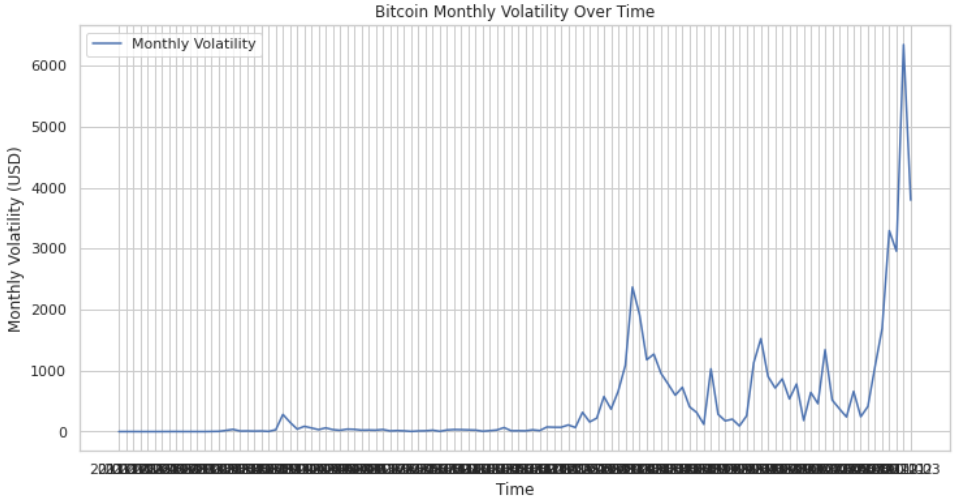

Fig B.4 Monthly Volatility Over Time

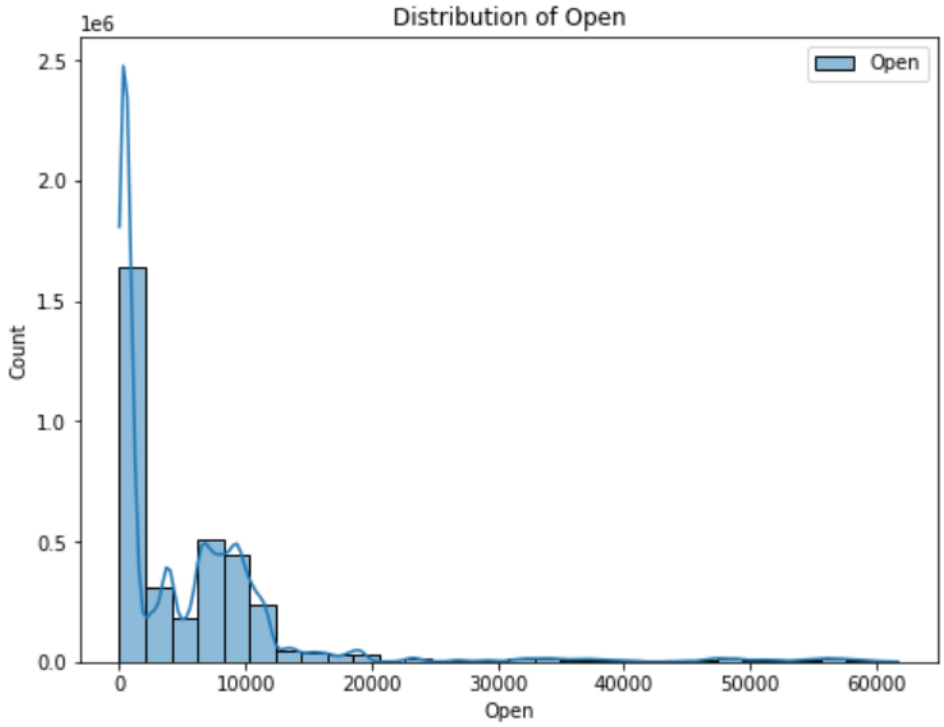

Fig B.5 Distribution of Open

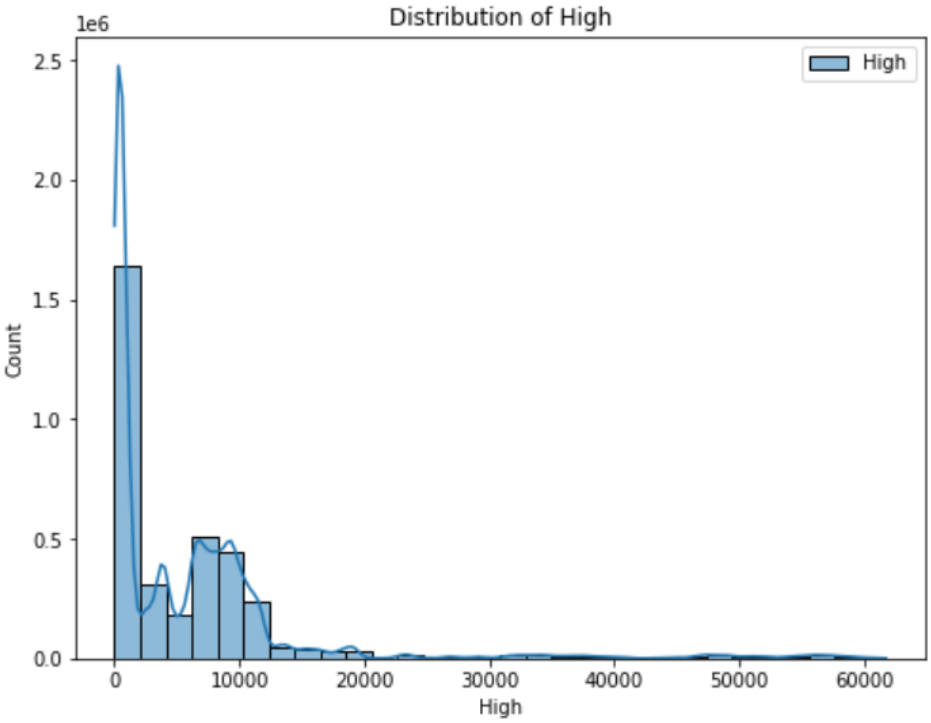

Fig B.6 Distribution of High

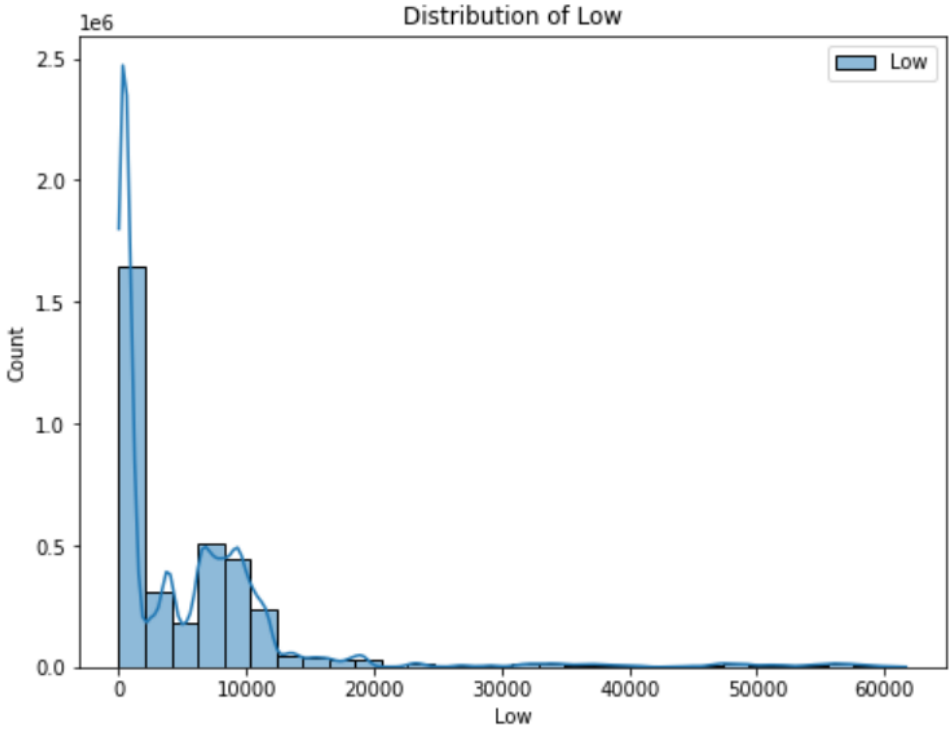

Fig B.7 Distribution of Low

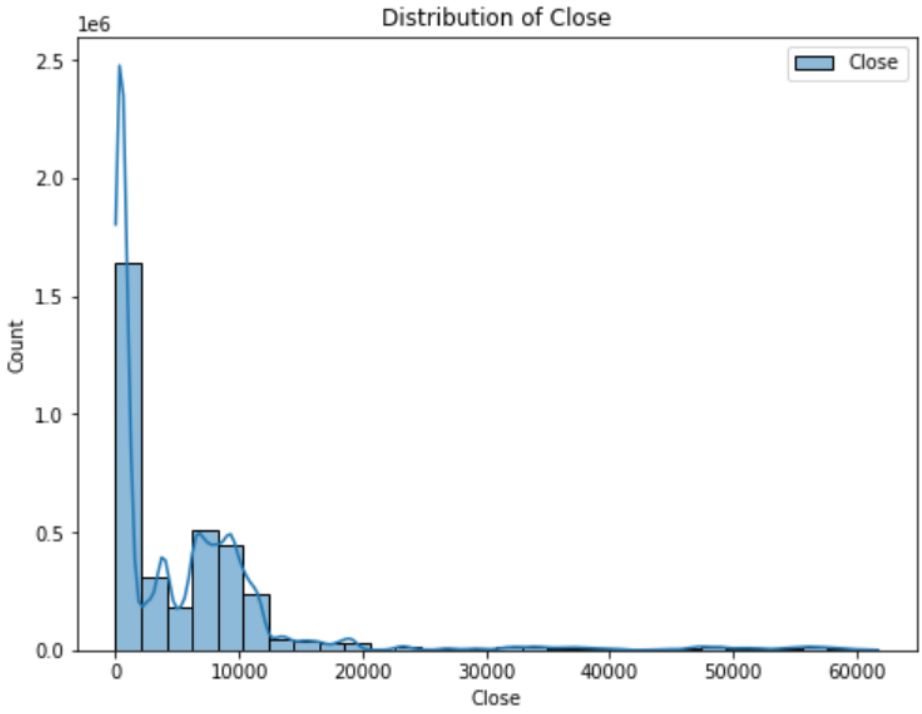

Fig B.8 Distribution of Close

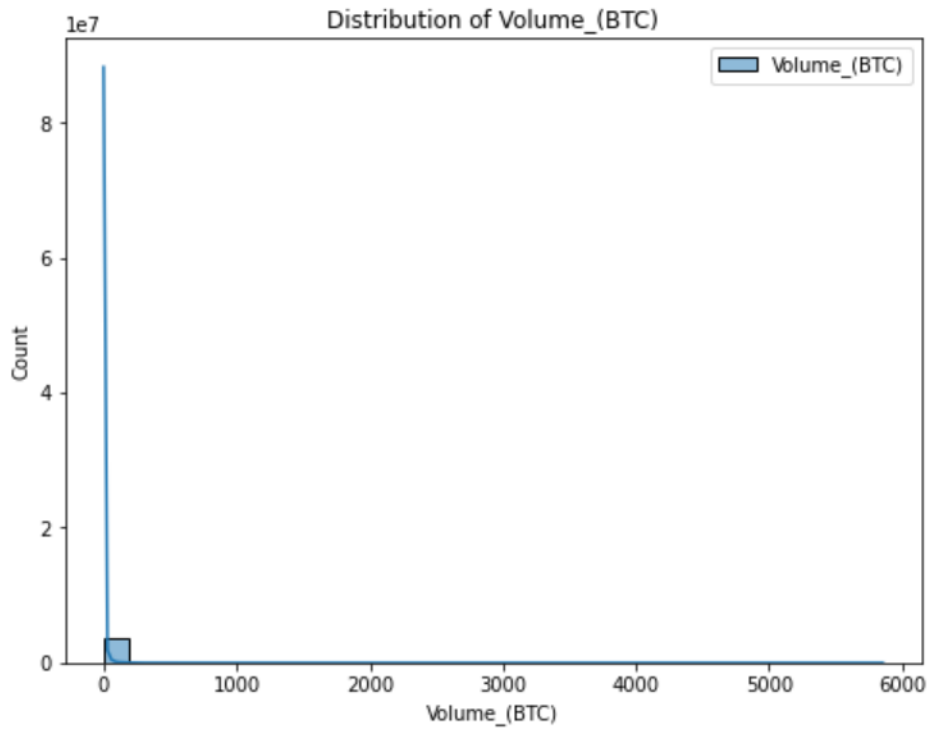

Fig B.9 Distribution of Volume_(BTC)

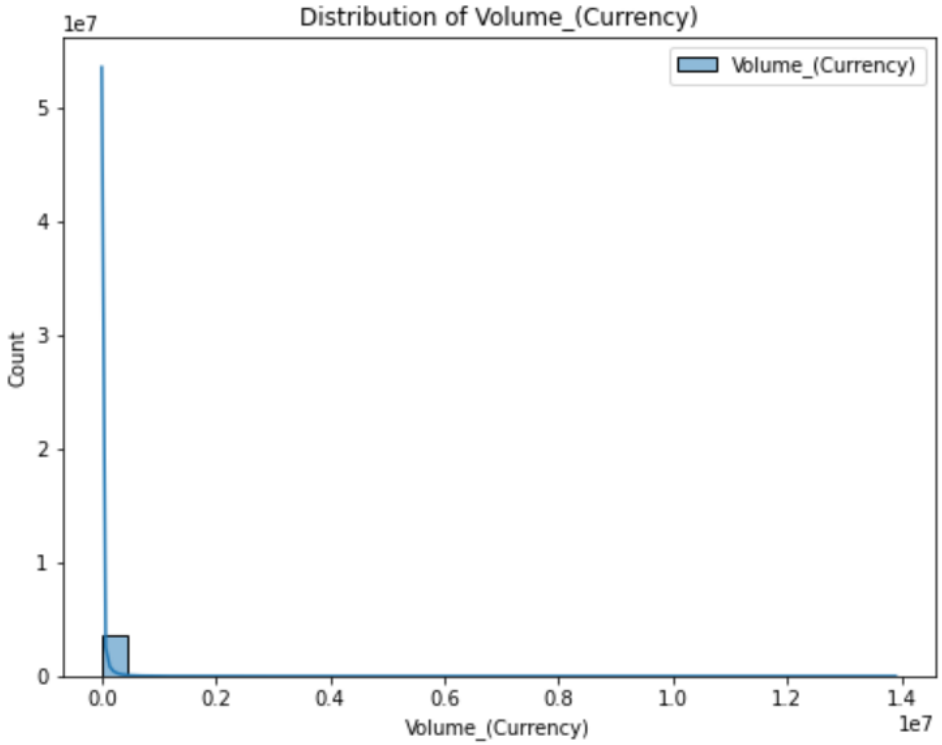

Fig B.10 Distribution of Volume_(Currency)

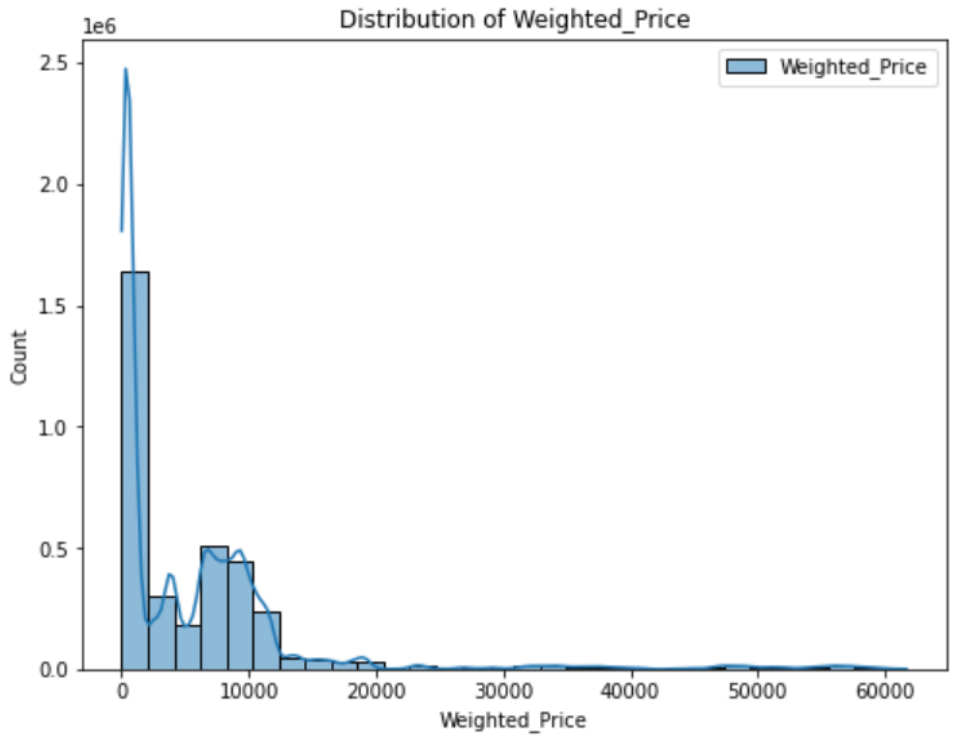

Fig B.11 Distribution of Weighted_Price

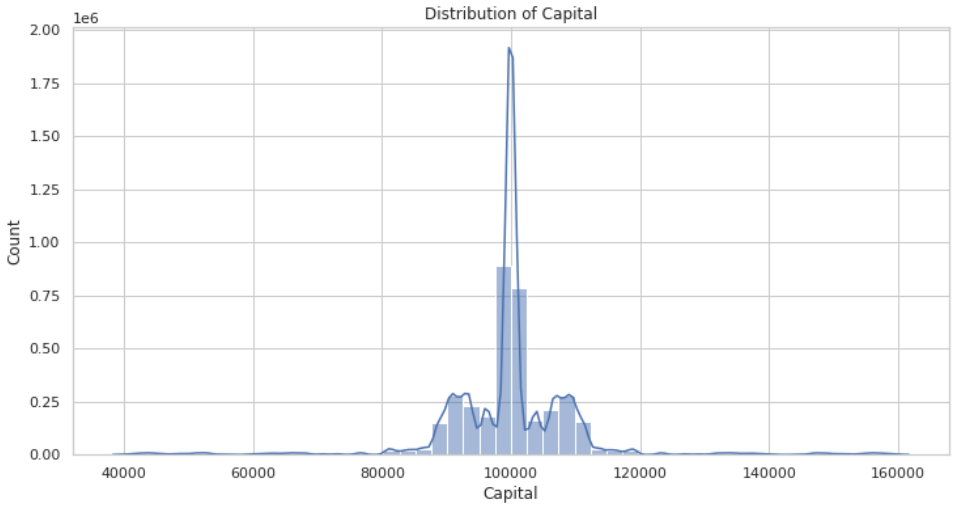

Fig B.12 Distribution of Capital

Fig B.13 Time Series Plot of Close Prices

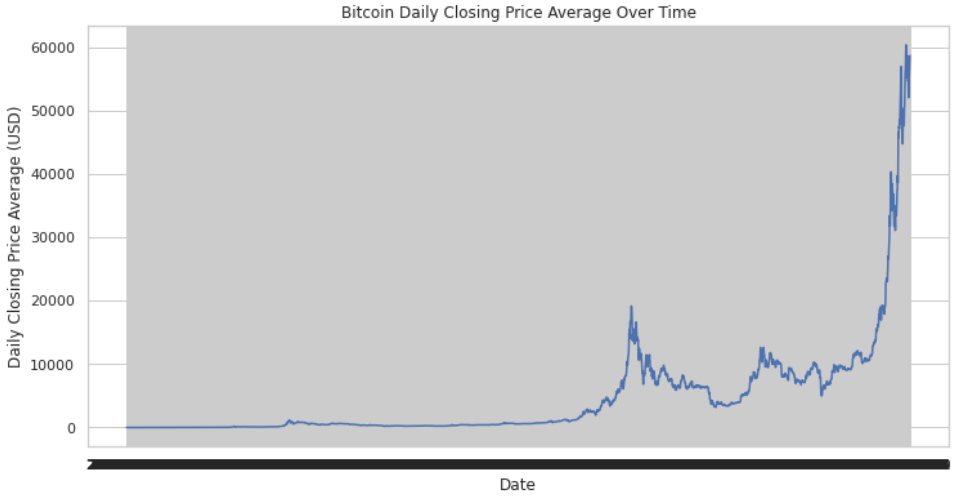

Fig B.14 Daily Closing Price Average Over Time

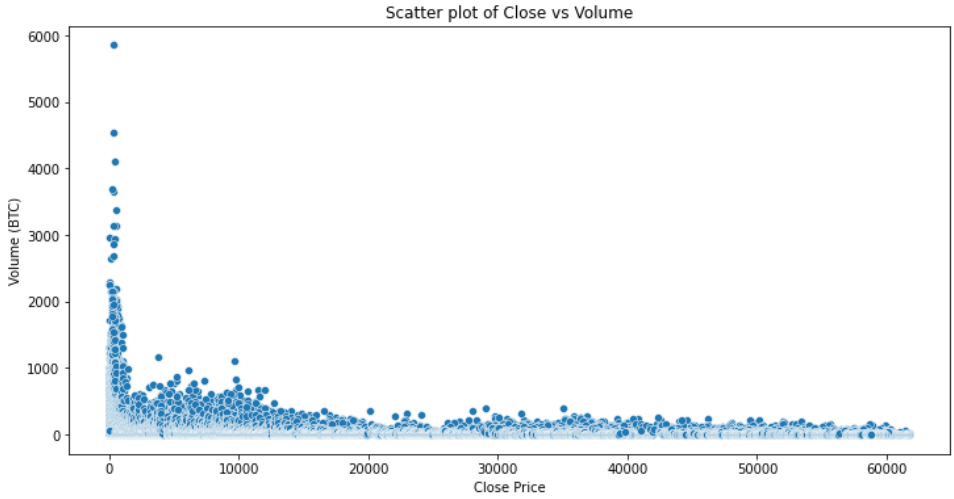

Fig B.15 Scatter plot of Close vs Volume

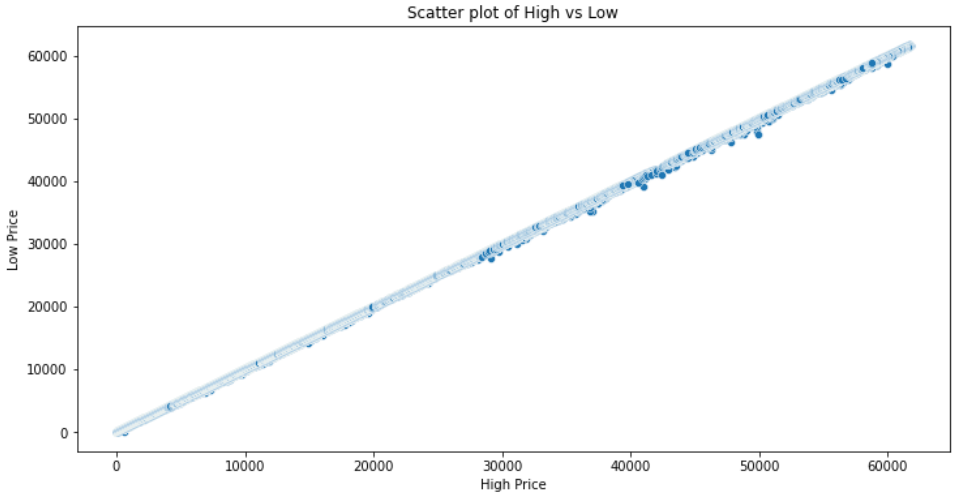

Fig B.16 Scatter plot of High vs Low

Fig B.17 Bitcoin Price and 5-Minute SMA

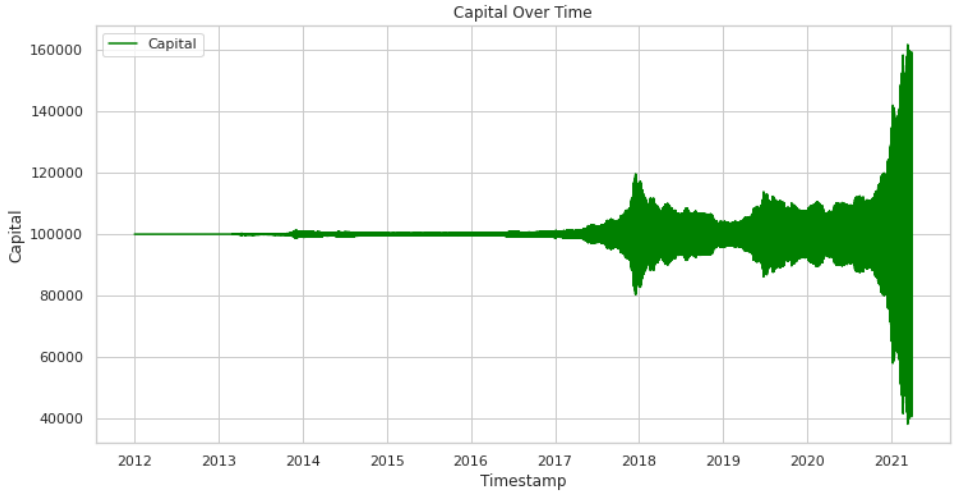

Fig B.18 Capital Over Time

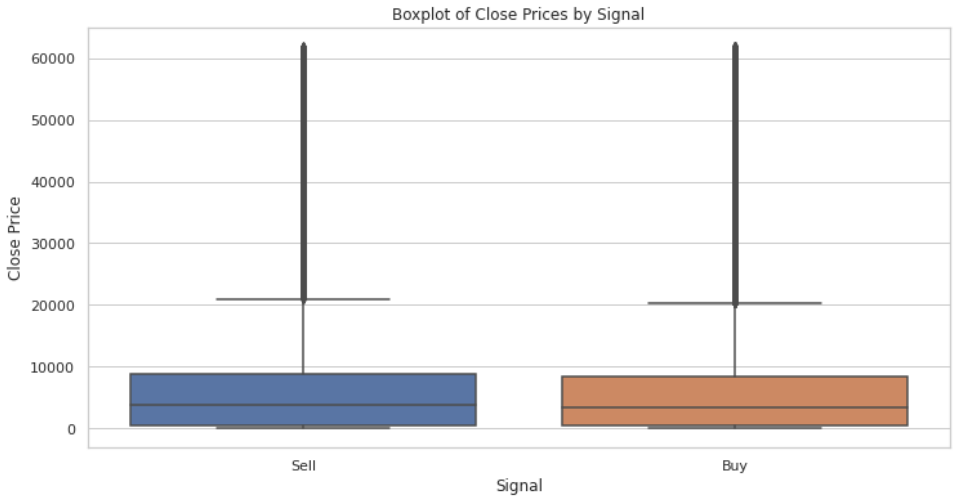

Fig B.19 Boxplot of Close Prices by Signal

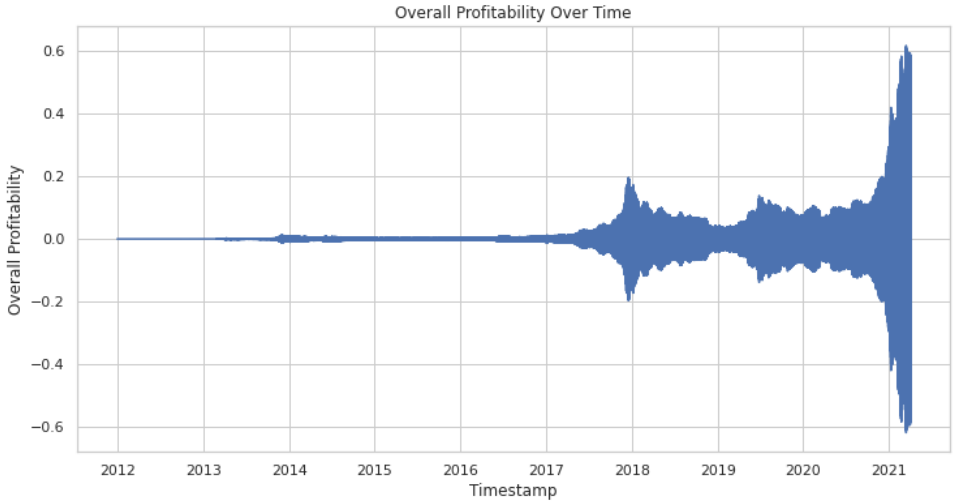

Fig B.20 Overall Profitability

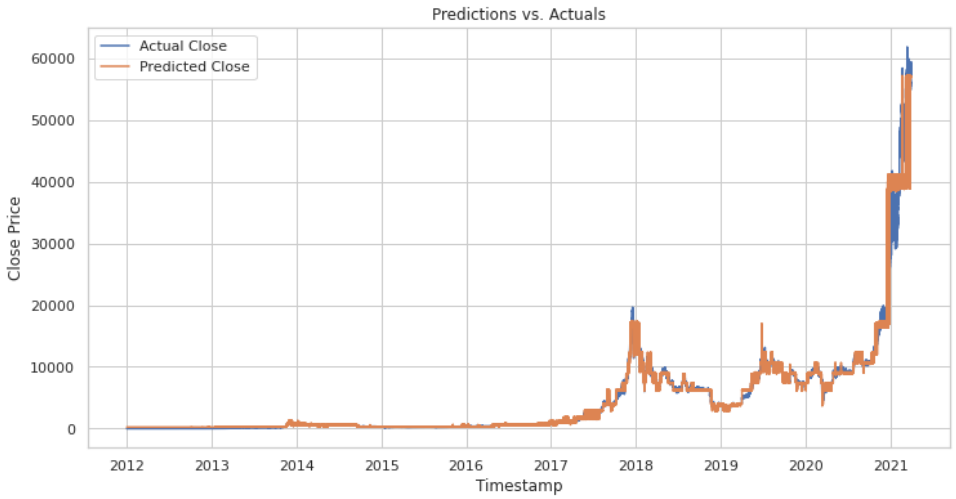

Fig B.22 Prediction Vs Actuals

Appendix C. File Links

C.1 Bitcoin Historical Dataset from Kaggle -

https://www.kaggle.com/datasets/mczielinski/bitcoin-historical-data/data?select=bitstampUSD_1-min_data_2012-01-01_to_2021-03-31.csv