BITPREDICT

A Bitcoin Price Prediction Model

CSE 158R FALL 2024 ASSIGNMENT 2

1 Introduction and EDA

1.1 Introduction

In this report, we analyse a comprehensive dataset of historical Bitcoin prices to gain insights into its behaviour over time. The dataset contains over 6.7 million records spanning from January 2012 to December 2024, providing minute-by-minute information on Bitcoin prices and trading volume. The key variables in the dataset include Open, High, Low, Close, Volume, and a newly created YearMonth column.

Bitcoin, a highly volatile digital asset, has been subject to frequent price fluctuations influenced by numerous economic and speculative factors. Our goal is to better understand these price behaviours and extract meaningful patterns that will ultimately inform our predictive modelling approach.

1.2 Exploratory Data Analysis

To start, we conducted a summary statistical analysis of the key features in the dataset, including Open, High, Low, Close, and Volume. The average closing price for Bitcoin has increased dramatically over the years, with substantial peaks and valleys indicating its volatile nature.

1.2.1 Time Range Analysis

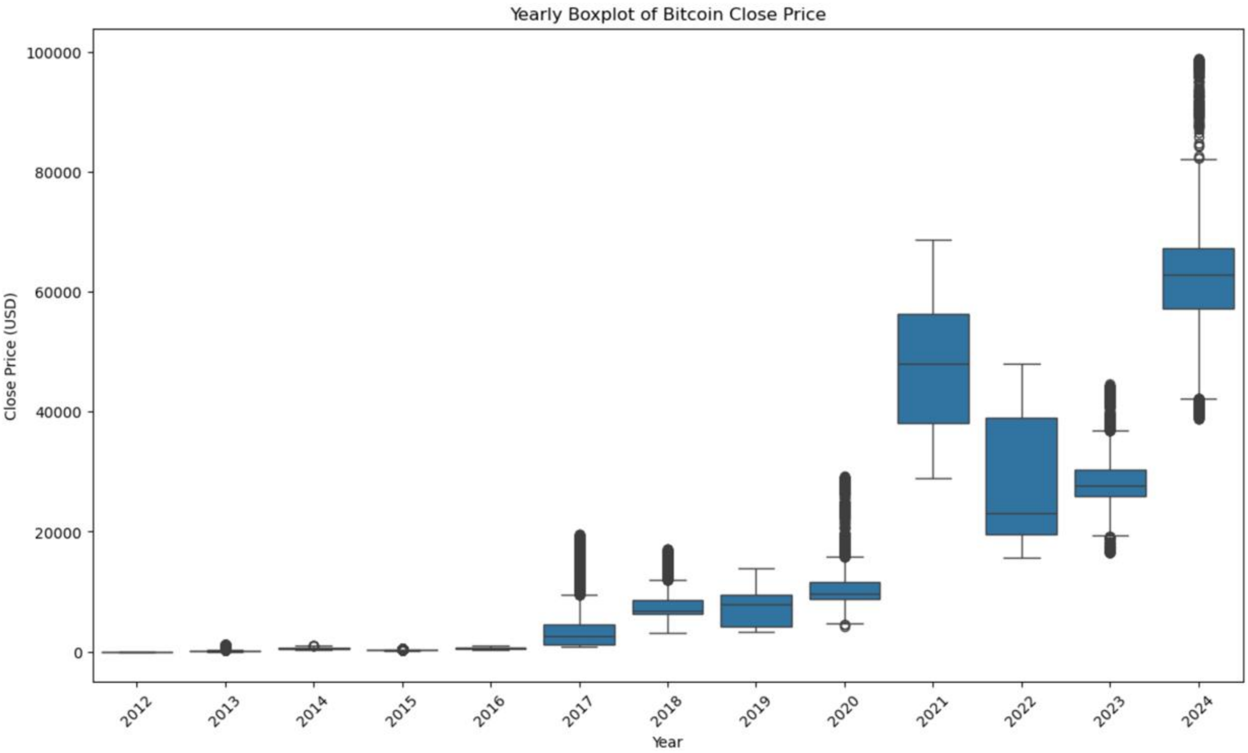


Figure 1: BitCoin Close Price Over the Years

The dataset contains minute-level data from 2012 to 2024, providing a detailed look at Bitcoin's market behavior over time. We visualized the Close price over this period, observing several significant periods of price growth, including the major rally in 2017 and the surge from late 2020 to 2021. We noted a general upward trend but also identified major periods of correction and consolidation.

1.2.2 Trend and Pattern Analysis

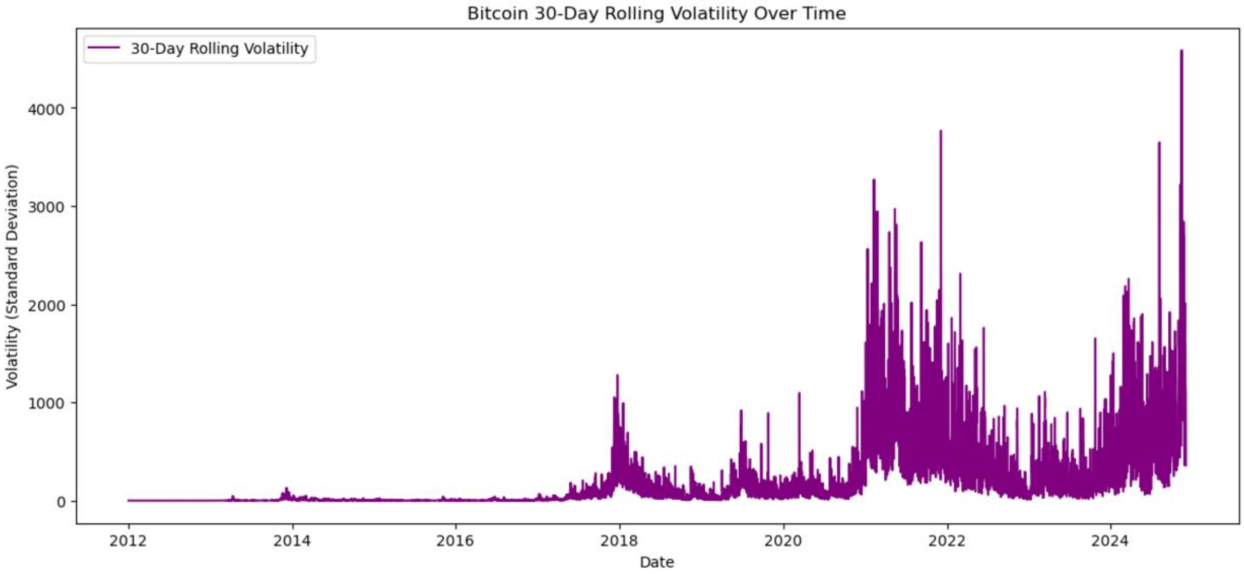


Figure 2: BitCoin 30 Day Rolling Volatility

To better understand Bitcoin's price trends, we calculated 7-day and 30-day rolling averages for the Close price. These moving averages helped in smoothing short-term fluctuations and highlighting long-term trends. The 7-day moving average captured the recent price dynamics, whereas the 30-day moving average gave a broader view of market movements. We also investigated the relationship between Volume and price movements. During periods of sharp price increases or decreases, there was often a corresponding spike in trading volume. This indicates that trading activity increases in times of high volatility, likely driven by both speculation and reaction to market news.

1.2.3 Distribution Analysis

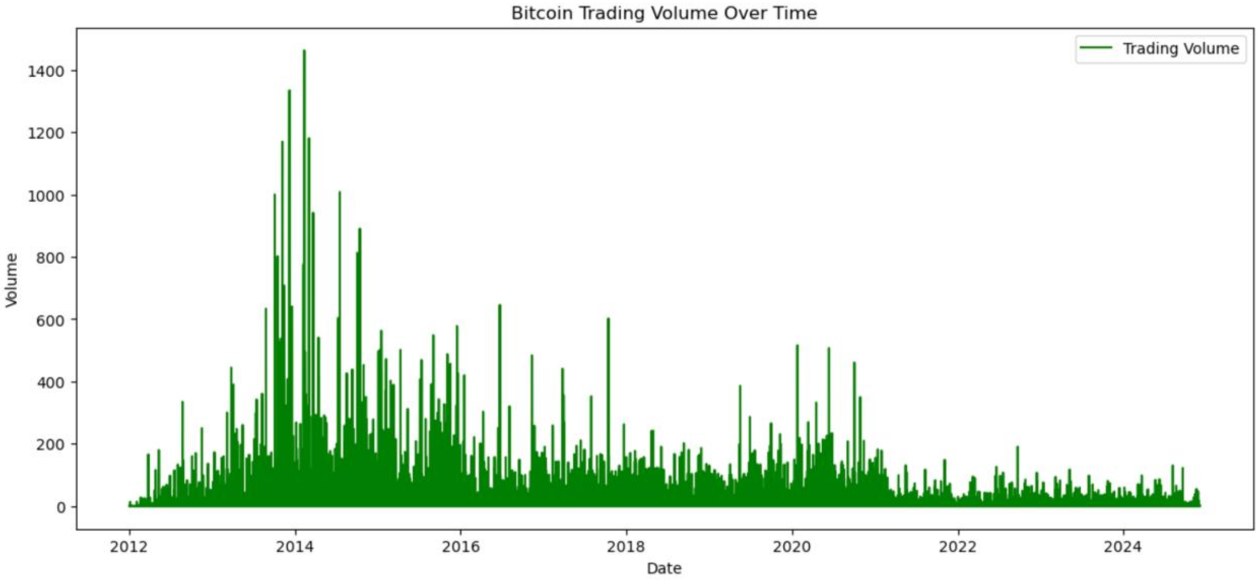


Figure 3: BitCoin Trading Volume

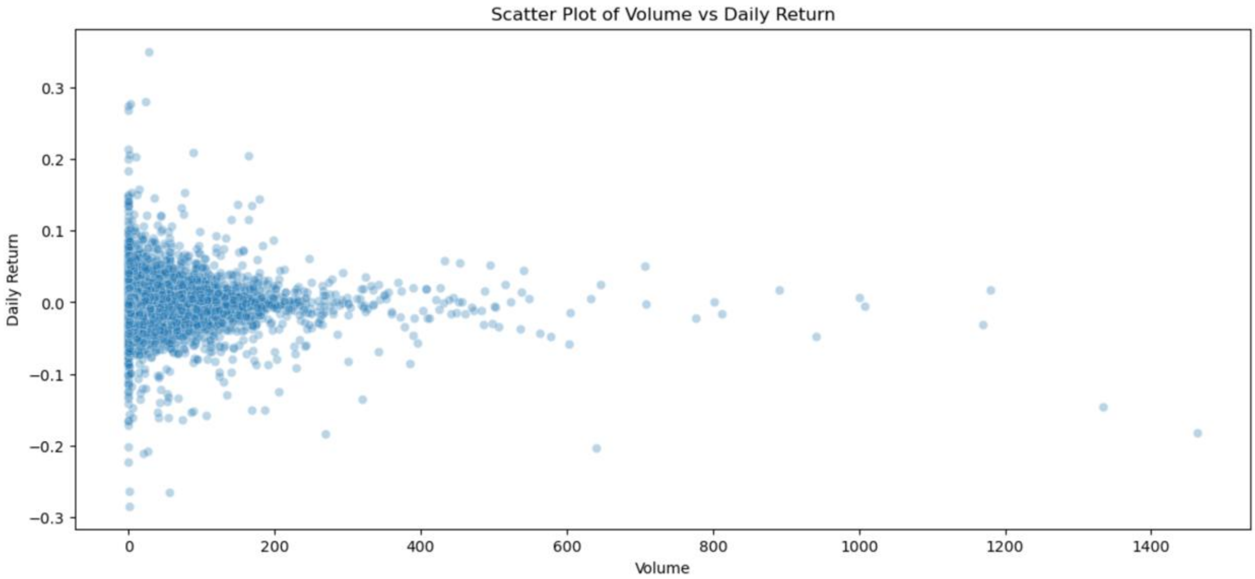


Figure 4: BitCoin Volume vs Daily Return

Most price variables exhibited a positively skewed distribution, which is consistent with Bitcoin's history of rapid growth followed by corrections. The Volume variable also showed a skewed distribution, with most trading occurring at lower volume levels but occasional very high spikes.

1.2.4 Volatility Analysis

We analyzed the volatility of the Close price over time, finding that certain periods, such as late 2017 and 2021, were marked by extreme price volatility. We also explored seasonality patterns by aggregating data on a monthly and yearly basis. Bitcoin's price showed no consistent monthly pattern, although some years exhibited seasonal peaks driven by broader market trends and news events.

1.2.5 Correlation Analysis

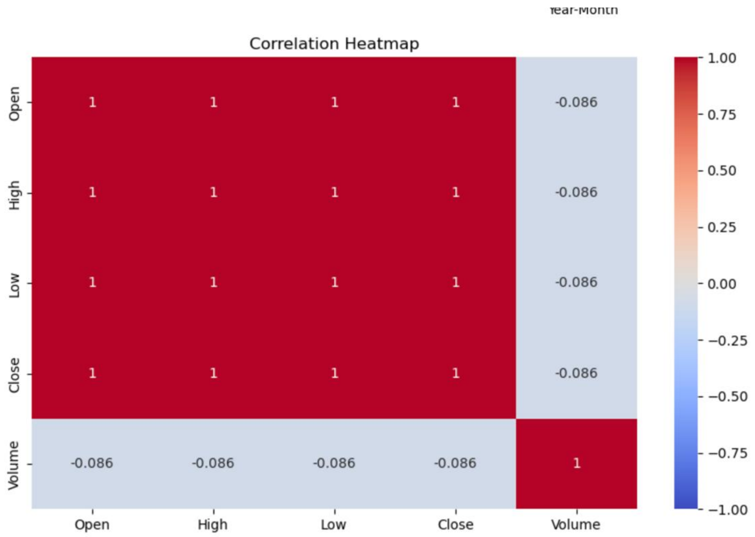


Figure 5: Correlation Matrix

We calculated the correlations between different features, such as Open, Close, High, Low, and Volume. The Open, Close, High, and Low prices were all highly correlated, as expected, due to the nature of financial time series data. The Volume was moderately correlated with price changes, suggesting that larger trading volumes could sometimes coincide with significant price movements.

2 The Predictive Task

2.1 BitCoin Price Prediction

The predictive task that we have chosen for this project is forecasting the Bitcoin closing price based on historical price data. This task is particularly relevant given the volatile nature of Bitcoin, which makes it challenging but rewarding to predict accurately. Our model aims to leverage historical trends and patterns in Bitcoin prices to provide useful price forecasts for both short-term and potentially long-term investment decisions.

2.2 Model Evaluation and Validity

To evaluate the predictive performance of our model, we will use Mean Absolute Error (MAE) as the primary metric. MAE helps us understand the average deviation of predicted prices from actual prices, which is critical for understanding how close our predictions are on a day-to-day basis. We chose MAE over other metrics like Mean Squared Error (MSE) because MSE values can become disproportionately large due to squaring errors, making the metric less representative in the context of volatile financial data like Bitcoin. Additionally, we are not using accuracy as a metric since it would require arbitrarily choosing an acceptable range of error, which could introduce unnecessary assumptions and bias into the evaluation.

By focusing on MAE, we aim to provide a more straightforward and interpretable assessment of our model's performance, particularly for a highly volatile asset like Bitcoin.

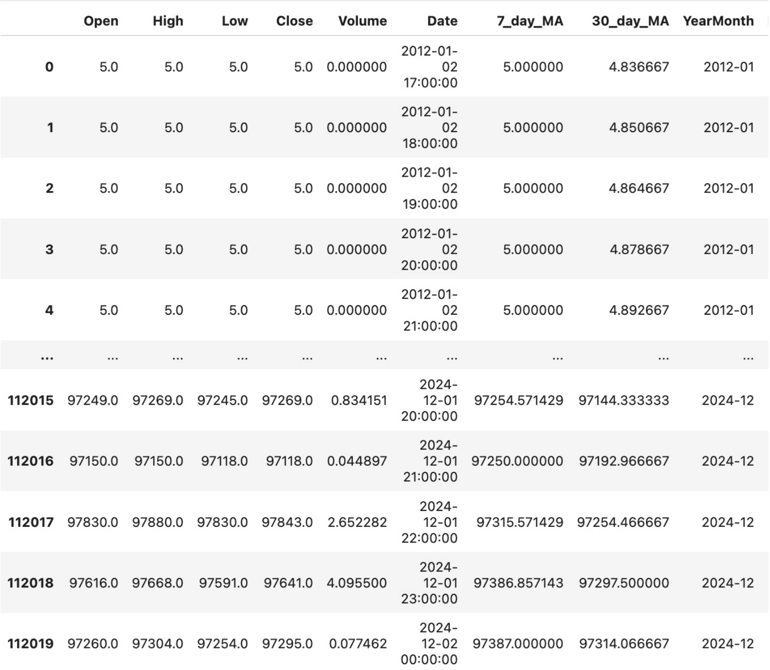
2.3 Data Preprocessing

Our initial dataset consisted of approximately 6.7 million rows of data, recorded every minute, offering a granular view of Bitcoin's historical price movements and trading volume. However, to make the data more manageable and computationally efficient for analysis, we aggregated the records into hourly intervals. This preprocessing step reduced the dataset size significantly, bringing it down to around 112,000 rows, while still preserving essential patterns and trends over time. By reducing the dataset, we ensured that the analysis would be faster and less resource-intensive without compromising the integrity of the data.

In addressing data quality, we removed rows containing missing or NaN values, which could have introduced inaccuracies or biases in the analysis. Additionally, the original Unix timestamp was converted into a human-readable Date column, allowing for more intuitive interpretation of time-based patterns. After the conversion, the original Unix timestamp column was removed to streamline the dataset and eliminate redundancy.

To further enhance the dataset's usability for trend analysis, we introduced new features, including a 7-day rolling mean and a 30-day rolling mean. These rolling averages provide insights into both short-term and long-term trends, smoothing out daily fluctuations and making it easier to identify patterns in Bitcoin's price movements over weekly and monthly periods. Such features are particularly valuable for detecting cycles and predicting future trends based on historical performance.

In addition, a YearMonth column was created by grouping the data by year and month. This new feature facilitates more effective exploratory data analysis (EDA), allowing for the examination of seasonal patterns and monthly trends. It also simplifies the aggregation of data for visualization and summary statistics, helping to highlight the evolution of Bitcoin prices and trading volumes over longer periods. Together, these preprocessing steps transformed the raw dataset into a structured and feature-rich format, enabling more effective analysis and model training.



	Open	High	Low	Close	Volume	Date	7_day_MA	30_day_MA	YearMonth
0	5.0	5.0	5.0	5.0	0.000000	2012-01-02 17:00:00	5.000000	4.836667	2012-01
1	5.0	5.0	5.0	5.0	0.000000	2012-01-02 18:00:00	5.000000	4.850667	2012-01
2	5.0	5.0	5.0	5.0	0.000000	2012-01-02 19:00:00	5.000000	4.864667	2012-01
3	5.0	5.0	5.0	5.0	0.000000	2012-01-02 20:00:00	5.000000	4.878667	2012-01
4	5.0	5.0	5.0	5.0	0.000000	2012-01-02 21:00:00	5.000000	4.892667	2012-01
...	...	...	...	...	...	...	...	...	...
112015	97249.0	97269.0	97245.0	97269.0	0.834151	2024-12-01 20:00:00	97254.571429	97144.333333	2024-12
112016	97150.0	97150.0	97118.0	97118.0	0.044897	2024-12-01 21:00:00	97250.000000	97192.966667	2024-12
112017	97830.0	97880.0	97830.0	97843.0	2.652282	2024-12-01 22:00:00	97315.571429	97254.466667	2024-12
112018	97616.0	97668.0	97591.0	97641.0	4.095500	2024-12-01 23:00:00	97386.857143	97297.500000	2024-12
112019	97260.0	97304.0	97254.0	97295.0	0.077462	2024-12-02 00:00:00	97387.000000	97314.066667	2024-12

Figure 6: Preprocessed Dataframe

3 Models Explored

3.1 The Unsuccessful Attempt - Support Vector Regressor (SVR)

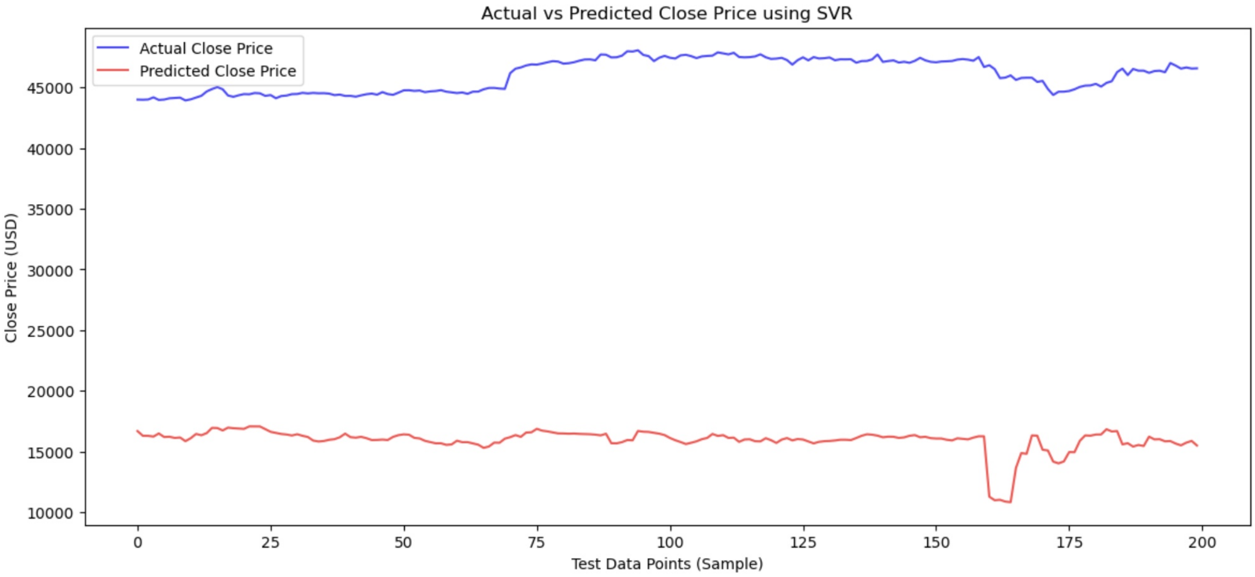


Figure 7: SVR Prediction for Small Sample

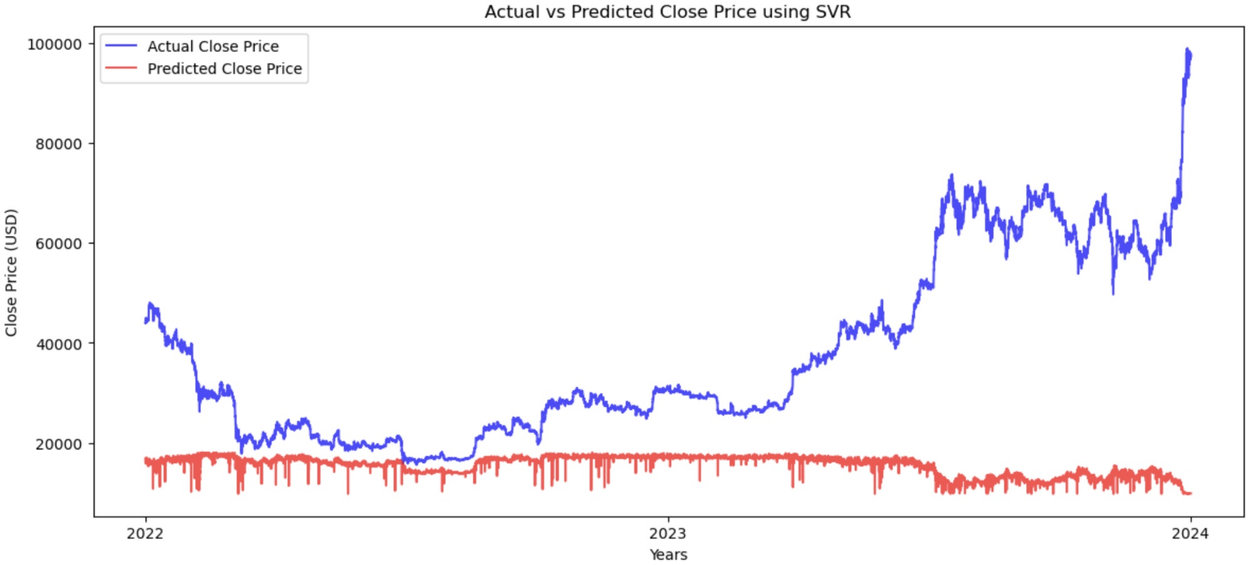


Figure 8: SVR Prediction for the Last 2 Years

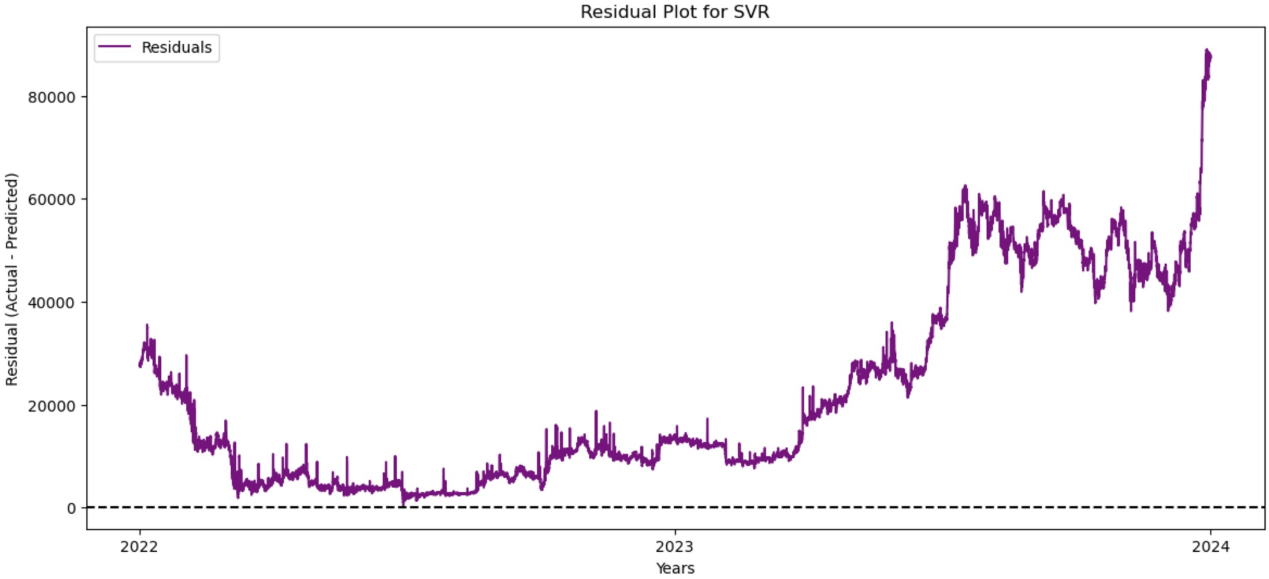


Figure 9: Residual Plot for SVR

The SVR (Support Vector Regressor) was our first attempt, but it turned out to be the least effective model in our evaluations. SVR struggled with both scalability and performance. Bitcoin price data is highly volatile, with frequent sharp changes, which makes it challenging for SVR to establish a consistent decision boundary.

One of the major limitations of SVR is its reliance on a kernel function to map input data into a higher-dimensional space to find an optimal separating hyperplane. This approach is better suited for datasets with well-defined patterns. However, for financial data like Bitcoin, which is non-stationary and has high volatility, SVR had difficulty adapting to the rapid fluctuations, resulting in poor predictive accuracy and a high MAE.

Another key issue was scalability. The computational complexity of SVR grows significantly with the size of the dataset, making it impractical for large time-series datasets like ours. The training times were considerably longer compared to other models, and this additional computational cost did not lead to improved predictive accuracy. The inability to handle the scale of the data effectively combined with its limited performance on volatile financial data made SVR an unsuitable choice for this predictive task.

3.2 A Better Attempt - Gradient Boost Regressor

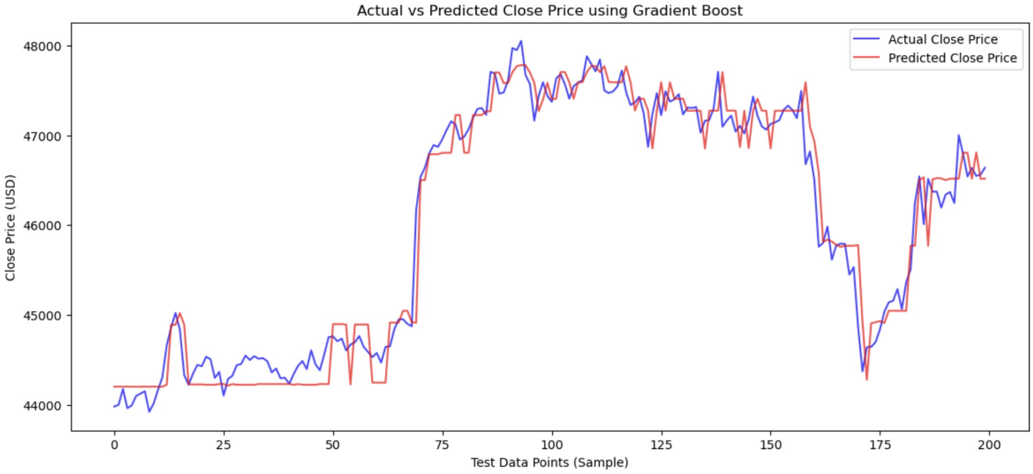


Figure 10: GBR Prediction for Small Sample

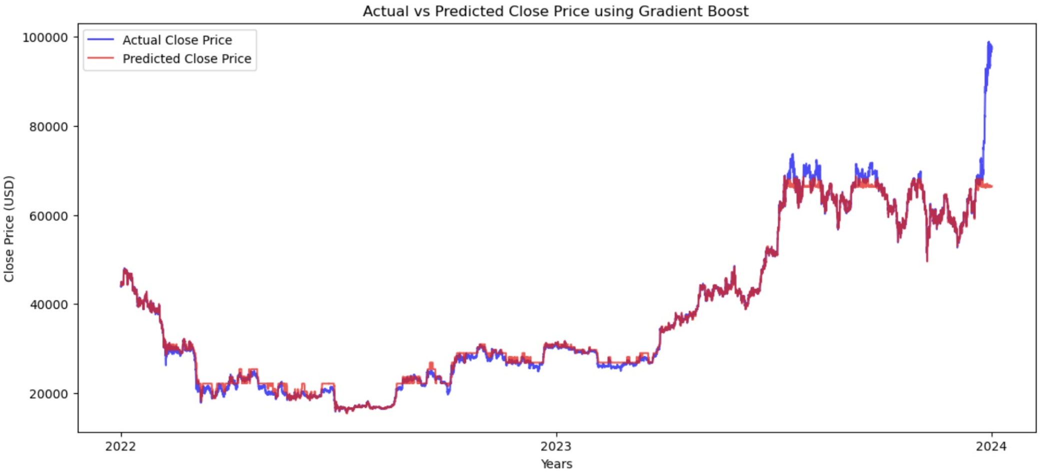


Figure 11: GBR Prediction for the Last 2 Years

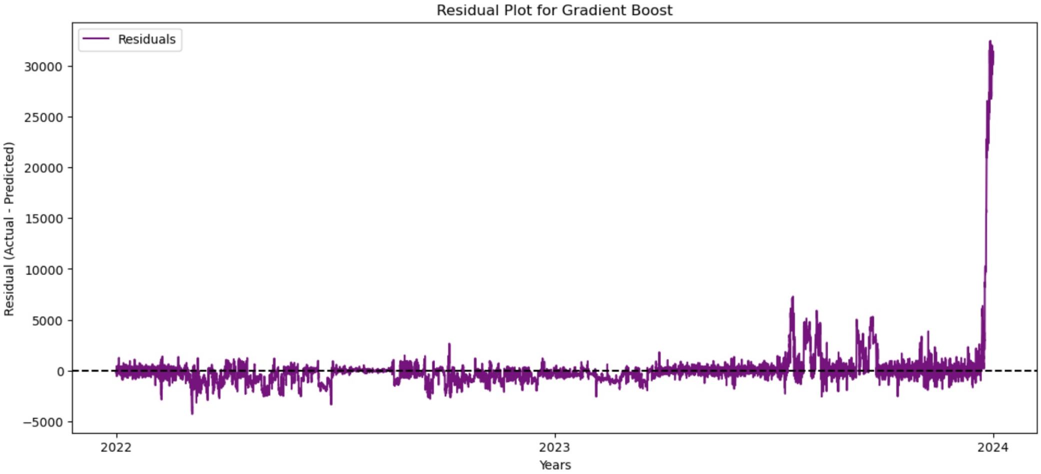


Figure 12: Residual Plot for GBR

The Gradient Boosting Regressor was another model we explored, and it performed better than SVR, achieving an MAE of 799.16. Gradient Boosting is well known for capturing non-linear relationships in the data by building an ensemble of decision trees iteratively, each correcting the errors of the previous one. This helped the model adapt to some of the fluctuations inherent in Bitcoin price data, resulting in a more accurate prediction compared to SVR.

One of the reasons Gradient Boosting performed better was its ability to learn complex, non-linear patterns, which are prevalent in financial datasets like Bitcoin. However, despite its improvements over SVR, it still lacked the temporal context that sequential models like LSTM provide. The lag features we used were helpful in providing some history to the model, but they were not sufficient for capturing the true sequential dependencies present in the data.Another limitation was overfitting. Gradient Boosting, despite its iterative nature, can easily overfit on noisy and highly volatile data like Bitcoin prices if not tuned properly. To mitigate this, we used techniques like limiting the number of estimators and tuning the learning rate, but the model's performance still did not reach the level of accuracy we desired.

3.3 The Best Model - Long Short-Term Memory(LSTM)

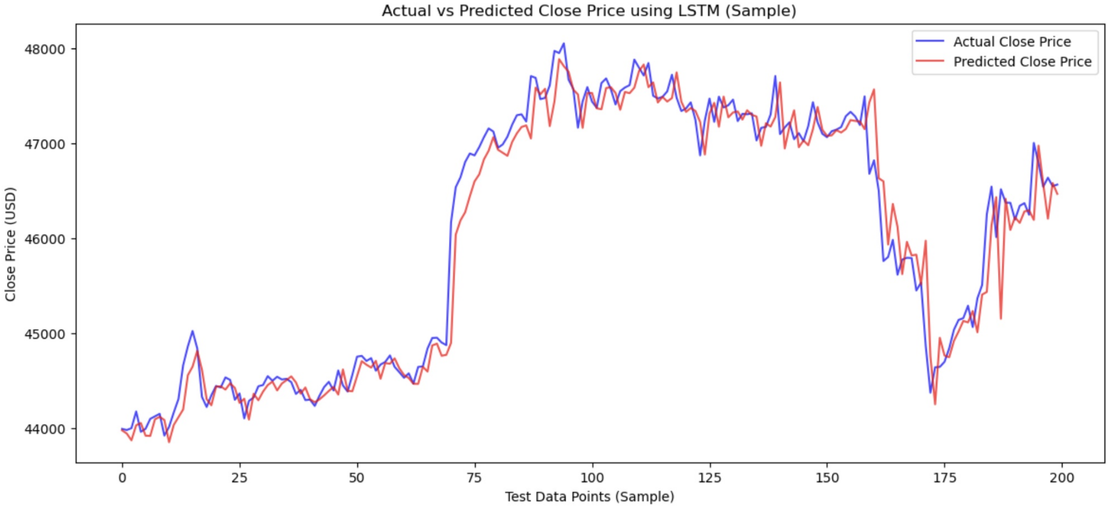


Figure 13: LSTM Prediction for Small Sample

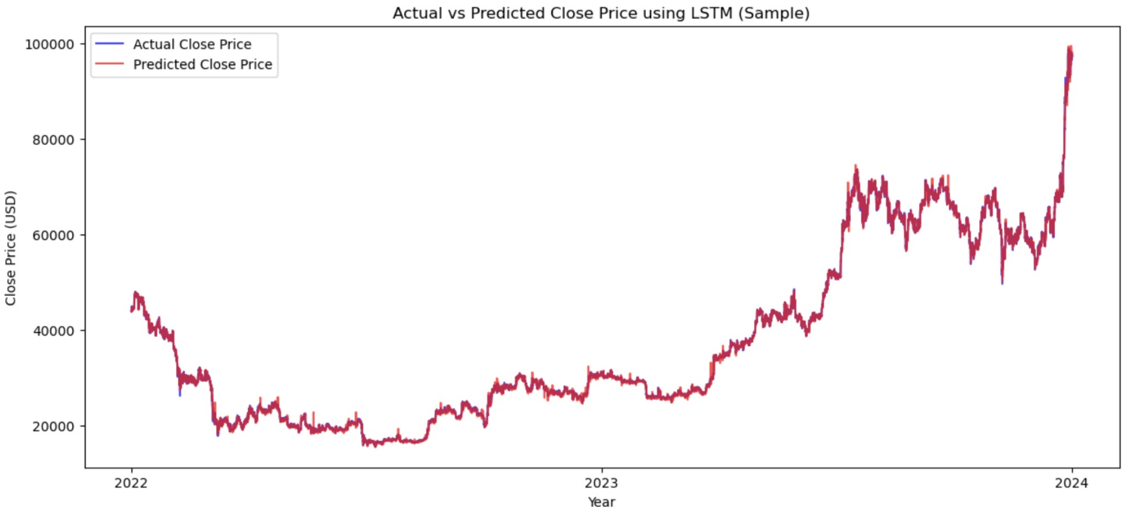


Figure 14: LSTM Prediction for the Last 2 Years

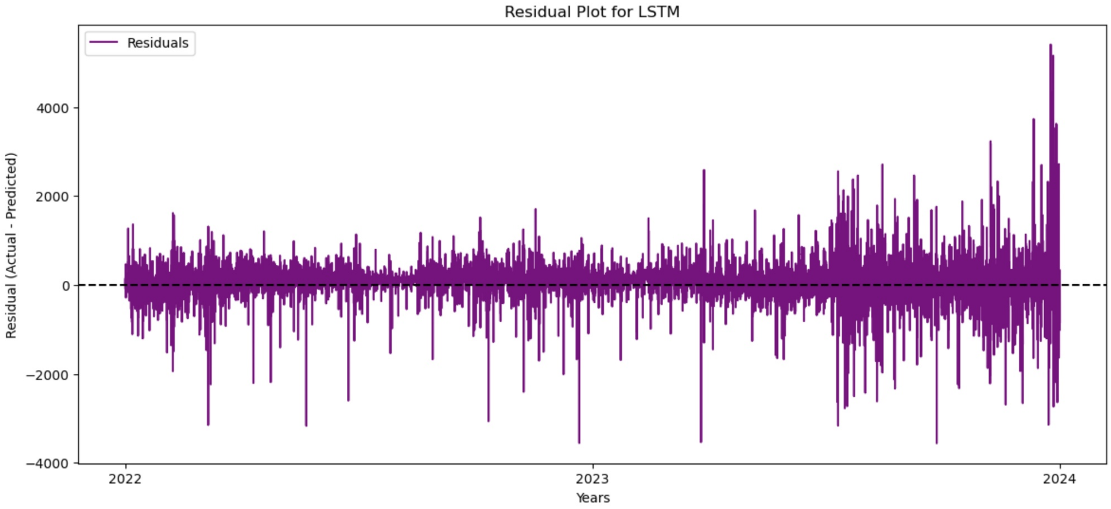


Figure 15: Residual Plot for LSTM

The LSTM (Long Short-Term Memory) model was ultimately chosen because it significantly outperformed the other models, producing the lowest MAE (221) among all the models we tested. LSTMs are a type of recurrent neural network (RNN) specifically designed to handle sequential data and capture long-term dependencies, making them ideal for time-series prediction tasks like Bitcoin price forecasting.

Bitcoin prices exhibit significant temporal dependencies, where past patterns and trends influence future prices. Unlike Gradient Boosting or SVR, which only considered lagged features or non-linear relationships, LSTMs are capable of retaining important information over extended periods through their memory cell structure. This allows them to effectively model the complex, volatile nature of Bitcoin prices, which is crucial for generating accurate predictions.

One of the key strengths of the LSTM model was its ability to adapt to the sequential characteristics of the data. By learning temporal dependencies, the LSTM model could make more informed predictions, reducing the overall prediction error. The LSTM also benefited from hyperparameter tuning, where we optimized the number of LSTM units, learning rate, and batch size. We found that using 50 LSTM units and the Adam optimizer provided the best results. Additionally, we incorporated early stopping to avoid overfitting, ensuring the model generalized well to unseen data.

However, LSTMs also come with certain challenges. They are computationally intensive, requiring more resources and longer training times compared to simpler models like SVR or Gradient Boosting. Overfitting was another concern, as the model could easily memorize the training data, especially given the noise and volatility of Bitcoin prices. To address this, we added dropout layers and used early stopping to prevent the model from overfitting.

In conclusion, while the LSTM model required careful tuning and additional computational resources, its ability to effectively capture the temporal dependencies in the Bitcoin price data made it the best choice for this predictive task. It provided the most accurate and reliable predictions, significantly reducing the MAE compared to the other models we explored.

4 Background

4.1 About BitCoin

Bitcoin is a decentralized digital currency created in 2009 by an anonymous entity known as Satoshi Nakamoto. Unlike traditional currencies, Bitcoin operates on a peer-to-peer network without the need for a central authority, such as a bank or government. Transactions are verified by network nodes through cryptography and recorded in a public ledger called the blockchain. Over the years, Bitcoin has grown to become the most popular cryptocurrency, and its price is known for being highly volatile. Factors such as market demand, regulatory news, technological developments, and macroeconomic trends heavily influence its price. Predicting Bitcoin's price is a challenging task due to its volatility, which makes it an ideal case study for exploring the capabilities of different predictive models.

4.2 The Dataset

The dataset we used for this project was sourced from Kaggle, specifically from the dataset titled "Bitcoin Historical Data". This dataset contains historical Bitcoin price data, including open, high, low, close, and volume information. It is a commonly used dataset for studying cryptocurrency price behavior due to its comprehensive historical coverage.

4.3 Methods Used Over the Years

Over the years, several methods have been used to predict Bitcoin prices, many of which have had limited success due to the unique volatility and non-stationary nature of cryptocurrency markets.

ARIMA (AutoRegressive Integrated Moving Average) has been one of the widely used models for time-series prediction, including financial data. While ARIMA can capture linear trends in historical data, it has significant limitations in dealing with the highly non-linear and volatile nature of Bitcoin prices. The performance of ARIMA is often hindered by its assumption of stationarity, which is not valid for Bitcoin data that shows sudden and sharp fluctuations. Consequently, ARIMA models typically result in poor predictive accuracy when applied to cryptocurrency price prediction.

Another category of models that have been used includes basic machine learning models like Linear Regression and Decision Trees. Linear Regression, which is designed to model linear relationships, falls short when dealing with complex financial time series like Bitcoin. The inherent non-linearity of Bitcoin prices makes Linear Regression ineffective, often resulting in large prediction errors. Decision Trees are slightly better than Linear Regression in capturing non-linear relationships, but as standalone models, they tend to overfit on small fluctuations and fail to generalize well to unseen data. This makes them unsuitable for predicting prices in a market as unpredictable as Bitcoin.

Many early approaches also used technical indicators such as Moving Averages and Momentum Indicators to predict cryptocurrency prices. While these indicators can provide some insight into the market's short-term behavior, they are not reliable for long-term predictions. Moving Averages often lag behind actual price movements, making them slow to react to rapid changes, which are very common in Bitcoin trading.

Simple models like the Random Walk model assume that future price changes are random and independent of past values. While this approach can sometimes provide a reasonable benchmark due to the unpredictable nature of Bitcoin, it is not suitable for precise predictions. The Naïve Forecast model, which assumes that the next period's price will be the same as the current period's, also fails to capture the complexities of the market and provides very limited predictive power.

4.4 Our Approach

The methods discussed above, including ARIMA, Linear Regression, and Moving Averages, have proven to be less effective in capturing the volatility and non-linear dependencies of Bitcoin prices. Our approach, which involved the use of Gradient Boosting Regressor and LSTM, aimed to overcome these limitations by employing models capable of learning complex relationships and temporal dependencies. While Gradient Boosting provided a more nuanced understanding of the data's non-linear patterns, LSTM took it a step further by leveraging sequential learning, ultimately providing more reliable predictions.

5 Conclusion

Our final model, the LSTM, achieved the lowest Mean Absolute Error (MAE) of 221, significantly outperforming both the SVR and Gradient Boosting Regressor models. The SVR model had a high MAE due to its inability to capture the complexity of Bitcoin's volatility and its scalability limitations. Gradient Boosting, while more effective than SVR, struggled to fully capture the temporal dependencies in the data, which was crucial for accurate predictions in a highly volatile market.The LSTM model succeeded primarily due to its architecture, which is designed to learn long-term dependencies in sequential data. This allowed it to effectively model the trends and fluctuations in Bitcoin prices, something that the other models could not achieve. The LSTM's memory cells retained relevant information from past price movements, enabling it to make more informed predictions.

In terms of feature representations, using lag features worked well for Gradient Boosting, as it helped provide some historical context to the model. However, the LSTM's ability to inherently capture sequential information without explicit lag features gave it a significant edge. The rolling averages we used during preprocessing helped smooth out short-term fluctuations, which was beneficial for all models, particularly in understanding longer-term trends.

The significance of our results lies in demonstrating the effectiveness of sequential models like LSTM in predicting highly volatile financial data. By comparing different models, we found that traditional statistical models and simpler machine learning approaches struggled with the inherent volatility and complexity of Bitcoin prices. The LSTM model, despite its computational intensity and risk of overfitting, provided the most reliable results, highlighting the importance of using models capable of capturing temporal dependencies in financial time-series data.

While the LSTM model outperformed the alternatives, it is important to note that the high volatility of Bitcoin remains a challenge, and there is always room for improvement. Future work could involve exploring hybrid models or incorporating external factors, such as macroeconomic indicators or sentiment analysis, to further enhance the predictive accuracy.

One significant limitation in all models remain : the inability to augment sentiment analysis and macroeconomic indicators into the dataframe. Bitcoin's price is heavily influenced by factors such as public sentiment, regulatory announcements, and broader economic trends. Without including these external variables, the model is only relying on historical price data, which can lead to unreliable predictions, especially for long-term forecasting. The exclusion of these factors limits the model's ability to account for abrupt market changes driven by news events or economic conditions, which are crucial for accurate long-term price predictions.