University of Strathclyde Glasgow

MS984: Data Analytics in Practice

Case Study 4: Predictive Analysis of Bitcoin and Recognising Limitations

Group 3

Aravinthkumar Pattaiyan | Shrishti Sridhar Manja | Hrutuja Mangesh Patkar |

Blaise Marvin Rusoke | Liyi Tan

Executive Summary

Our core aim for this case study is to explore and implement statistical and computational techniques such as machine learning algorithms to predict cryptocurrency prices, specifically Bitcoin. Despite being a recent technological development, cryptocurrencies offer vast amounts of data on various aspects of interests which will be useful for a plethora of financial experiments. Cryptocurrencies present a distinctive and intriguing opportunity for data analytics also it will be helpful to explore the full potential.

Multiple time series models (LSTM and XGBoost) were considered to forecast the future price of Bitcoin with only time as a predictor with appropriate preprocessing techniques. In addition, the XGBoost Regression model has been used to investigate the relationship between different features such as bitcoin variables, other crypto variables, and commodity and security variables with Bitcoin price.

1.0 Introduction

The combination of technological advancements, changing attitudes toward traditional finance practices, and economic uncertainties have led to the increased popularity of Bitcoin over the years leading to the introduction of cryptocurrencies. As the cryptocurrency ecosystem continues to evolve, there is a growing need to explore innovative statistical and computational techniques to forecast price movements accurately.

Being able to forecast bitcoin prices is important as traders often use asset price movements as part of their trading strategies. For example, trend-following traders may look for assets with high velocity as potential trading opportunities, aiming to capitalize on ongoing price trends. Understanding patterns in their movement is also crucial for risk management. Assets that are more volatile carry higher risks, thus, traders and investors need to account for this when making trading decisions and managing their portfolios.

By leveraging historical price data and relevant features such as gold pricing, transaction volume and other cryptocurrency pricing alongside the transaction date, we aim to develop robust predictive models. Our predictive models will be based on findings achieved from thorough exploratory data analysis feature selection and modelling and testing against a combination of univariate and multivariate models to find the optimal model.

2.0 Data Exploration

2.1 Introduction to Dataset

This report analyses a dataset comprising financial and cryptocurrency-related metrics. The dataset consists of 3772 observations and 17 variables, providing insights into various aspects of Bitcoin (BTC) and other cryptocurrencies, as well as related financial indicators. *Table 1* presents a dataset containing various financial and network metrics related to Bitcoin (BTC), as well as prices of other cryptocurrencies (Ethereum, Litecoin, Bitcoin Cash, and Cardano), gold prices, and stock market indices (Nasdaq composite index, Dow Jones Industrial Average). Sources for the data include Yahoo Finance, Quandl, Coin Metrics Community Data, World Gold Forum, and Google Trends as can be seen in *Table 1* shared below. Additional

data to support time-series model building was sourced from Yahoo Finance which provided 1756 additional instances of Opening Bitcoin Price.

Table 1: Dataset variables and sources

Variable	**Source**
BTC Price, Nasdaq composite index, DJI	Yahoo Finance
BTC network hashrate, Average BTC block size, NUAU - BTC, Number TX - BTC, Difficulty - BTC, TX fees - BTC, Estimated TX Volume USD - BTC	Quandl
Gold in USD	World Gold Forum
Ethereum Price, Litecoin Price, Bitcoin Cash Price, Cardano Price	Coin Metrics Community Data
Google trends	Google Trends

2.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) revealed insights into the relationships between Bitcoin (BTC) price and various factors. Initial data preprocessing involved handling missing values through linear interpolation to maintain data integrity. Our key visualisations included histogram analysis which provided an overview of the distribution of key variables, including BTC price, network hash rate, block size, transaction volume, fees, and prices of cryptocurrencies, gold, and stock market indices. We also deployed scatter plots which were used to assess the linearity between BTC price and selected features, offering insights into potential correlations and predictive relationships. Furthermore, the Spearman rank correlation coefficients method as seen in *Figure 1* in the appendix, was used to calculate to quantify the strength and direction of monotonic relationships between BTC price and other variables.

This helped us nullify the hypothesis of a correlation being observed between the BTC Price and Gold returns feature. Additionally, we observed multicollinearity between the Nasdaq composite index and Dow Jones Industrial Average feature due to a strong positive correlation between the two of them. Findings from our correlation matrix were applied in building a sound regressor model.

3.0 Methodology

Our target for this case study was to predict the future price of Bitcoin and to also understand how each of the different features in the dataset affected the Bitcoin Price. We employed three models, the LSTM (Long Short-Term Memory) and XGBoost models for univariate time series prediction of the Bitcoin Price and an XGBoost Regression model to study the relationship between the different features and the Bitcoin Price.

3.1. Regression Analysis

As illustrated by the Spearman Rank correlation coefficients in Figure 1 of the appendix, our exploratory data analysis revealed significant relationships between the Bitcoin price and some of the features in the dataset. Correlation however does not translate to causation and

to gain deep insight into the different features that are responsible for the rise and drop of the Bitcoin price, we employed a machine learning regression model. We particularly used an XGBoost Regression model, a boosting ensemble of decision trees, which at its core uses a greedy algorithm that splits tree nodes to branches using feature importance and is therefore good for illustrating feature importances. Table 2 in the appendix illustrates the model hyperparameters, performance metrics and datasets used for regression analysis.

As illustrated by Figure 2 in the appendix, the trained XGBoost model identified the Difficulty-BTC, DJI (Dow Jones Index), Nasdaq Composite Index and the Estimated Bitcoin Transaction Volume as the four most important features for predicting the Bitcoin price.

Difficulty-BTC, the most important feature, is concerned with the difficulty involved in attaining the target hash of the bitcoin network and this complexity increases when more miners join the network. As shown in Figure 3, the Difficulty-BTC and Bitcoin prices have followed a similar trend over the years, implying that an increase in Bitcoin popularity and public attention which in turn draws more miners is central to the rise of the Bitcoin price.

The second and third most important features, the Nasdaq composite index and the Dow Jones index have followed a similar trend with the Bitcoin Price over the years as illustrated in Figure 4. This implies that an increase in market confidence, economic growth, and investor risk appetite is central to the rise of the Bitcoin Price. As shown in Figure 5, the bitcoin transaction volume, another important feature, follows a similar trend with the bitcoin price, implying that an increase in the number of bitcoin transactions is central to the rise of bitcoin price.

3.2. Bitcoin Price - Time Series Prediction

Our approach for predicting the future price of Bitcoin involved the use of both an LSTM model because of the great accuracy provided by neural networks, and an XGBoost model because of its high interpretability that is absent in neural network approaches.

The LSTM's model hyperparameters and model performance results are shown in Table *3*, and the XGBoost hyperparameters, performance metrics and training features are shown in Table *4*. Figure 8 illustrates the data used for time series model training which comprised of provided dataset and extra data collected from Yahoo Finance that incorporated Bitcoin prices up to February 2024.

The Bitcoin Price time-series demonstrated autocorrelation which was observed through the correlation between the Bitcoin price and the four lagged Bitcoin prices (Bitcoin Price on the same date 1,2,3 and 4 years ago) shown in Table *5*. The lagged features together with the month, year and day of the year were used to train the XGBoost Time Series Model, and the model's feature importance rankings for predicting future bitcoin prices are shown in Figure 6. The future model's predicted future bitcoin prices from 22 February 2024 to 21 February 2025 are also illustrated in Figure 7. The feature importance values in Figure 6 demonstrate that the year and lagged (historical) bitcoin prices are central to predicting the future price of bitcoin.

Looking at the future Bitcoin price predictions from Feb 2024 to Feb 2025, the XGBoost model's future prediction values in *Figure 7* demonstrate that the bitcoin price will fall to

slightly below $40,000 in March 2024, recover and rise to about $48,000 in May, before further plummeting to a value below $30,000 between late June to early August. The price will however rise again to a value above $45,000 at the start of 2025.

3.3. Analysis and Discussion

Public Interest in bitcoin and the state of the economy are two factors that were identified to be central to the bitcoin price (as in Figure 2). Figure 9 illustrates the rise in Bitcoin price with an increase in public Google searches about Bitcoin. The XGBoost regression analysis model also revealed that the more people get interested in bitcoin and subsequently join the network either as miners, to solely transact in bitcoin or both, then the more the bitcoin price will rise. The state of the economy depicted by the Dow Jones and Nasdaq composite indices is also closely linked with the bitcoin price and is useful for making price predictions. Finally, our XGBoost time series prediction model demonstrated the importance of past bitcoin prices in predicting future bitcoin prices (as in Figure 6). Although the predicted bitcoin price (as in Figure 7) will have its peaks and troughs, the model predicts that the currency will still perform better than it did during the downturn period between 2022 and 2024 (as seen in Figure 8).

4.0 Limitations

Limitations of predicting bitcoin prices using multivariate regression and time series models are significant due to the cryptocurrency market's inherent volatility and susceptibility to external influences. Despite employing advanced models like regression models or LSTM, accurately forecasting bitcoin prices remains challenging due to the dependency on lots of external factors.

Despite LSTM's having a higher ceiling for generating more accurate future bitcoin price predictions, our LSTM model did not generalize well on our future prediction time scale (Feb 2024 to Feb 2025) as seen in Figure 10. This performance can be attributed to the lack of extensive hyperparameter tuning and model performance optimization which we would have done if we had more time.

Both time dependency models and multivariate dependency models encounter a trade-off between capturing temporal relationships or incorporating multiple variables. Future research will focus on enhancing model robustness by exploring innovative approaches to incorporate real-time data and improve adaptability to dynamic market conditions, thereby advancing predictive capabilities in bitcoin price forecasting and building capable multivariate time series models without having to trade-off.

5.0 Future Work

The current LSTM model was trained with standard hyperparameters, neglecting the use of hyperparameter tuning to optimize parameters such as the number of LSTM units, learning rate, batch size, and dropout rate. Hence, fine-tuning these parameters will be done to

enhance the model's ability to capture complex patterns and dependencies within the Bitcoin price time series data.

Applying smoothing techniques to the training dataset should also be tried as the inherent raggedness and volatility present in Bitcoin price data can introduce noise and hinder the model's ability to identify meaningful patterns. Techniques like exponential moving average (EMA) may help smooth the time series data, which could mitigate noise and highlight underlying trends and patterns.

Lastly, while the current implementation of the univariate LSTM time series model has provided valuable insights into Bitcoin price prediction, expanding the scope of the model to include additional features beyond time could enhance its predictive power. Integrating relevant multivariate features such as trading volume, market sentiment indicators, and technical analysis metrics alongside historical price data can provide valuable context and leverage the interdependencies between multiple variables to improve the model's ability to capture complex market dynamics influencing Bitcoin price movements.

5.0 Conclusions

In conclusion, the pursuit of accurately predicting bitcoin prices through advanced statistical and computational methods is crucial given the cryptocurrency market's evolving nature and growing significance. Despite the challenges posed by market volatility and external influences, such as significant world events that affect the stock market, the application of multivariate regression and time series models offers valuable insights into price movements. While these models provide a framework for analysis, their limitations underscore the need for ongoing research and innovation. Future endeavours should focus on refining models to better incorporate real-time data, enhance adaptability to dynamic market conditions, and mitigate the impact of noise on prediction accuracy. By addressing these challenges and leveraging emerging technologies, such as artificial intelligence and machine learning, we can advance our understanding of cryptocurrency dynamics and contribute to more informed decision-making in this rapidly evolving financial landscape.

Appendix:

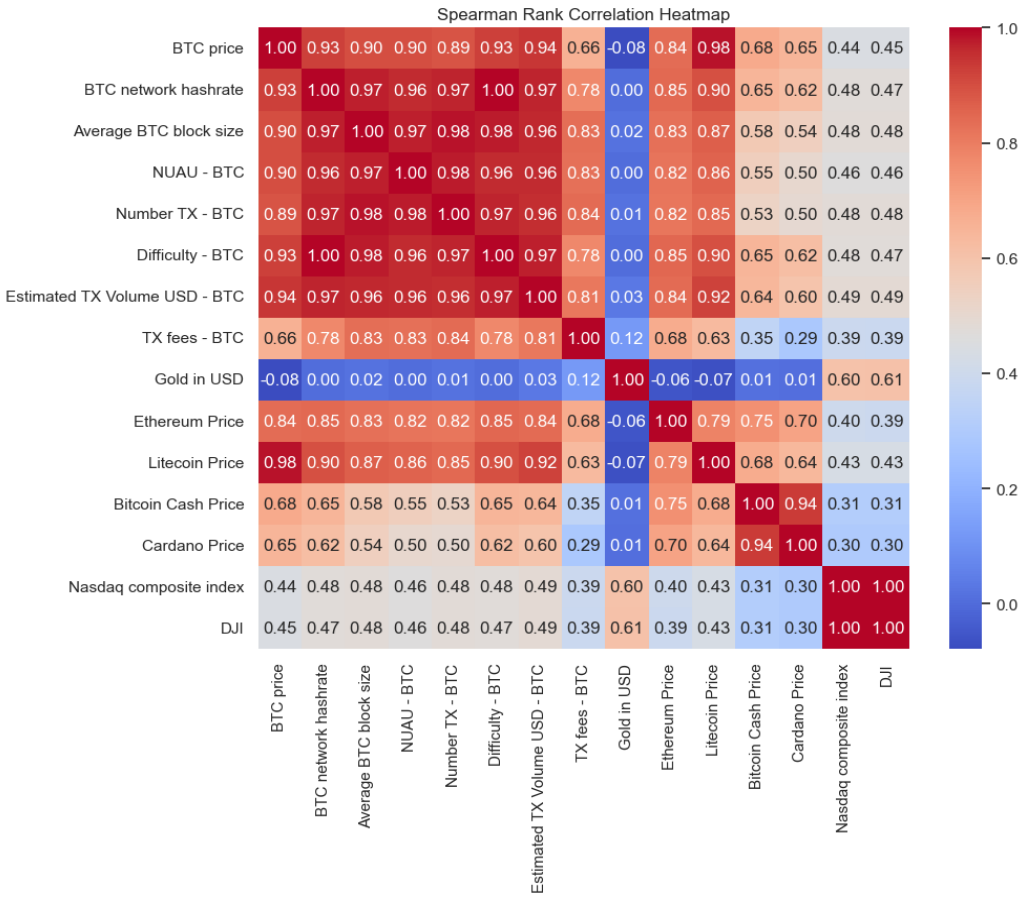

Figure 1: Spearman Rank Correlation between Bitcoin Price and other variables.

Table 2: XGBoost Regressor model hyperparameters and performance metrics

Model Hyperparameters				
Learning Rate	Max Depth	Number of Estimators	Lambda (Regularization Parameter)	Early stopping rounds
0.1	4	1200	5.0	50
Model Performance				
Training RMSE	46.35070			
Validation RMSE	263.07872			
Dataset				
Features used for regression analysis	BTC network hashrate, Average BTC block size, NUAU – BTC, Number TX - BTC, Difficulty - BTC, TX fees - BTC, Estimated TX Volume USD - BTC, Gold in USD, Nasdaq composite index, DJI (Dow Jones Index)			
Training Set Size	3017 instances			
Test Set Size	755 instances			

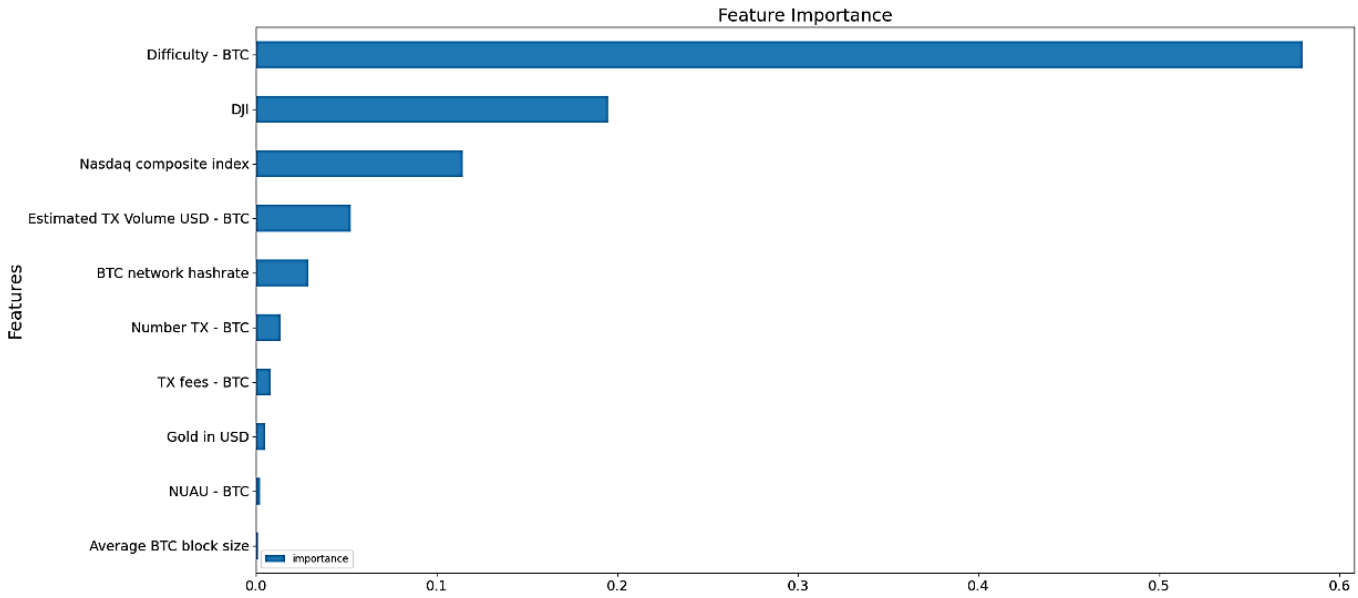

Figure 2: XGBoost Regression Analysis - Feature Importances

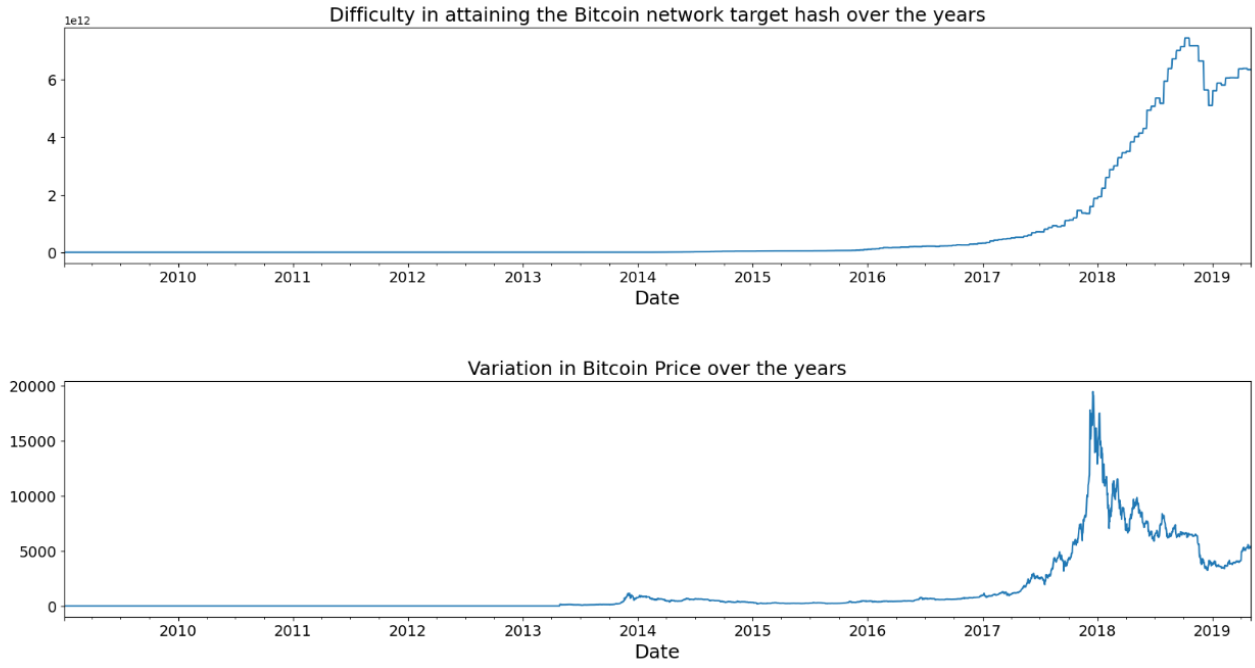

Figure 3: Variation of the Difficulty-BTC and Bitcoin Price over the years

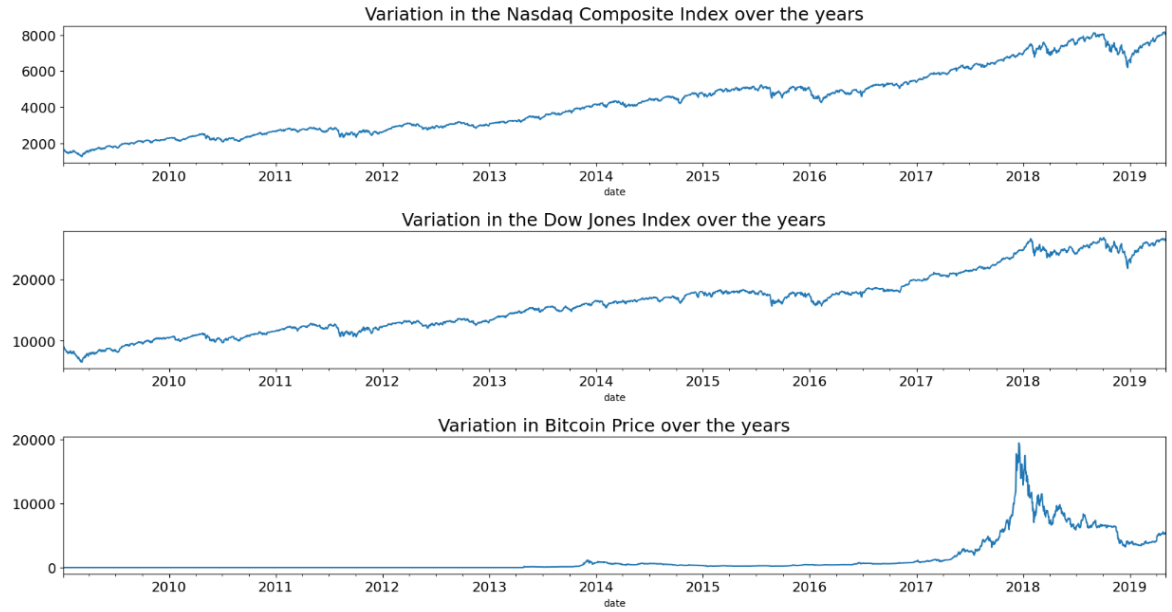

Figure 4: Variation of the Nasdaq Composite Index, Dow Jones Index and Bitcoin Price over the years

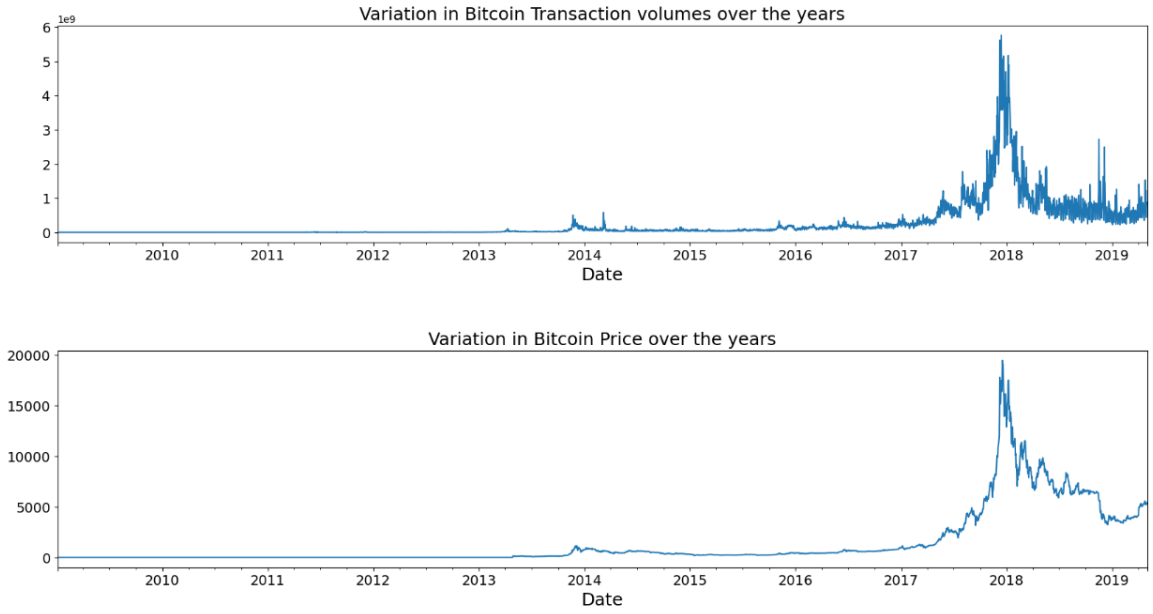

Figure 5: Variation in the Bitcoin Transaction Volumes and Bitcoin Price over the years.

Table 3: LSTM model hyperparameters and performance metrics

LSTM Model Hyperparameters				
Learning Rate	Optimizer	Loss	Model Layers	Units in Hidden Layers
0.01	Adam	Mean Squared Error	4	50
Model Performance				
Training RMSE (USD)		767.24		
Validation RMSE (USD)		784.41		

Table 4: XGBoost Time Series Model hyperparameters, performance metrics and features used to train the model

XGBoost Model Hyperparameters				
Learning Rate	Max Depth	Regularization Lambda	Estimators	Early Stopping Rounds
0.1	3	1.0	500	50
Model Performance				
Mean RMSE across the validation folds (USD)		8251		
Dataset				
Training Features used (Extracted from the date attribute)		Month of the year, Year, Day of the Year, Lag 1 (Bitcoin Price a year ago), Lag 2 (Bitcoin price 2 years ago), Lag 3(Bitcoin Price 3 years ago), Lag 4 (Bitcoin Price 4 years ago)		

Table 5: Pearson correlation coefficient between the Bitcoin Price and its lagged self, with a lag period of 1 year

	Lag 1 - Bitcoin Price on the same date 1 year ago	**Lag 2 - Bitcoin price on the same date 2 years ago**	**Lag 3 - Bitcoin Price on the same date 3 years ago**	**Lag 4 - Bitcoin Price on the same date 4 years ago**
Bitcoin Price	0.6114	0.5439	0.6975	0.7289

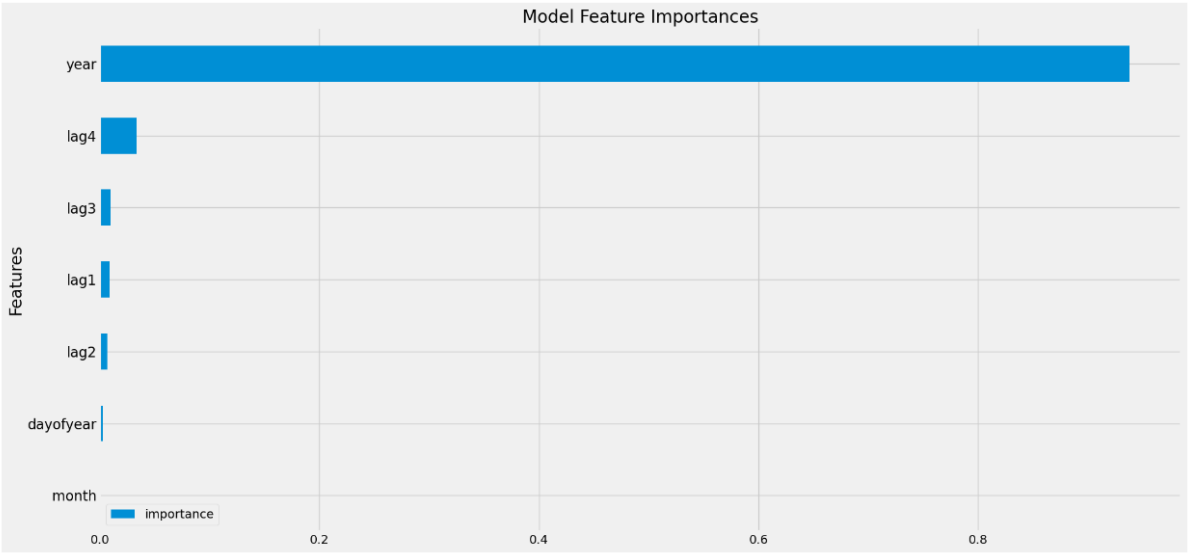

Figure 6: Feature Importances for predicting future Bitcoin Price

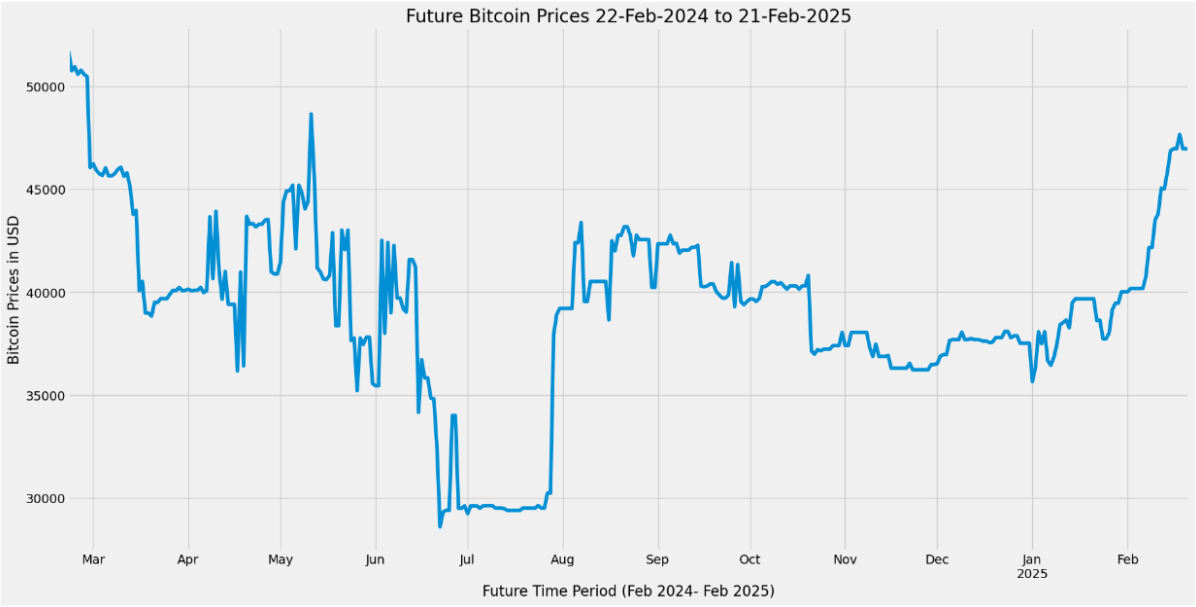

Figure 7: XGBoost Predicted Future Bitcoin Prices

Figure 8: Historical Bitcoin Prices

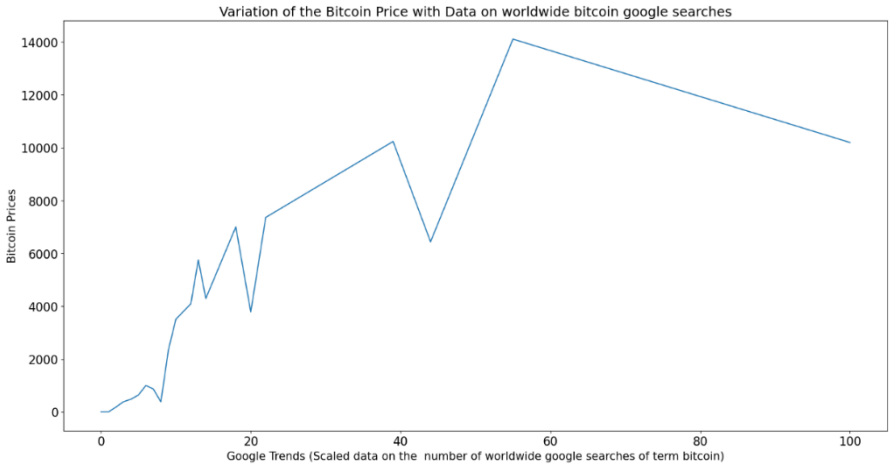

Figure 9: Variation in Bitcoin Price with google searches

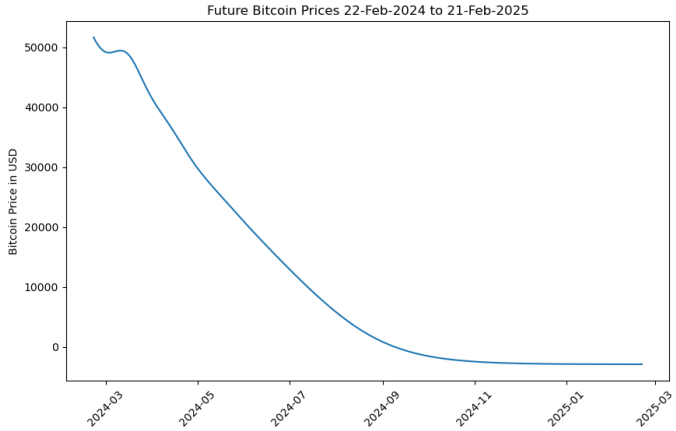

Figure 10: Predicted Future Bitcoin Prices by the LSTM model