University of Hertfordshire UH

School of Physics, Engineering and Computer Science

MSc Data Science Project
7PAM2002-0901-2024
Department of Physics, Astronomy and Mathematics

Data Science FINAL PROJECT REPORT

Project Title:

Bitcoin Price Prediction using ML

Student Name and SRN:

Hassan Faraz Khan

21089475

Supervisor: Peter Scicluna

Date Submitted: 6 Jan 2025

Word Count: 6965

GitHub Link: Click Here

Google Colab Link: Click Here

University of Hertfordshire UH

DECLARATION STATEMENT

This report is submitted in partial fulfilment of the requirement for the degree of Master of Science in Data Science at the University of Hertfordshire.

I have read the guidance to students on academic integrity, misconduct and plagiarism information at Assessment Offences and Academic Misconduct and understand the University process of dealing with suspected cases of academic misconduct and the possible penalties, which could include failing the project module or course.

I certify that the work submitted is my own and that any material derived or quoted from published or unpublished work of other persons has been duly acknowledged. (Ref. UPR AS/C/6.1, section 7 and UPR AS/C/5, section 3.6).

I have not used chatGPT, or any other generative AI tool, to write the report or code (other than where declared or referenced).

I did not use human participants or undertake a survey in my MSc Project.

I hereby give permission for the report to be made available on module websites provided the source is acknowledged.

Student SRN number: 21089475

Student Name printed: Hassan Faraz Khan

Student signature: HASSAN FARAZ KHAN

UNIVERSITY OF HERTFORDSHIRE

SCHOOL OF PHYSICS, ENGINEERING AND COMPUTER SCIENCE

SCHOOL OF PHYSICS, ENGINEERING AND COMPUTER SCIENCE

University of Hertfordshire UH

Abstract

This study investigates the application of machine learning models for predicting Bitcoin prices, aiming to address the challenges posed by the high volatility and complexity of cryptocurrency markets. The research evaluates the performance of three distinct models—Linear Regression, Random Forest, and Long Short-Term Memory (LSTM) networks—to provide a comprehensive understanding of their strengths and limitations in financial forecasting. By leveraging historical Bitcoin price data and feature engineering techniques, the study highlights the predictive capabilities of these models while offering valuable insights into their practical implications.

Linear Regression served as a baseline model, demonstrating strong predictive performance with a fine-tuned Mean Absolute Error (MAE) of 1181.13, Root Mean Squared Error (RMSE) of 1754.55, and R-squared (R^2) value of 0.9931. Random Forest, an ensemble learning method, excelled in identifying feature importance, with the 5-day moving average emerging as the most significant predictor. Despite its strengths in handling non-linear relationships, the model's overall accuracy was comparatively lower, with an MAE of 3888.63, RMSE of 8717.58, and R^2 value of 0.8286. LSTM networks, designed specifically for time-series forecasting, demonstrated superior performance by effectively capturing temporal dynamics and volatility. With an MAE of 1456.71, RMSE of 2087.55, and R^2 value of 0.9903, LSTM emerged as the most suitable model for predicting Bitcoin prices

The study also included 30-day future price predictions, providing a practical demonstration of the models' applicability. Linear Regression produced smooth, trend-aligned forecasts, while Random Forest exhibited greater variability in its predictions. LSTM balanced trend consistency with the ability to capture fluctuations, showcasing its adaptability to volatile market conditions.

The findings of this research underscore the importance of selecting models that align with the specific objectives and characteristics of the dataset. While Linear Regression provides a computationally efficient baseline, LSTM offers the nuanced understanding required for dynamic financial markets. Random Forest, despite its lower accuracy, contributes to feature interpretability and enriches the overall analysis.

Contents

1. Introduction

1.1 Background

Historically introduced in the year 2009, Bitcoin was the foremost advanced virtual currency available to users as conceived by its original creator, Satoshi Nakamoto (Edwards, 2024). As technology advanced, it became apparent that Bitcoin was going to revolutionize the concepts of money and finance altogether. This is because all the transacting parties could offer their transactions to be recorded by blocks, that did not require supervision by financial institutions such as banks, since every payment relied on the use of a distributed ledger technology blockchain. These qualities of Bitcoin- its decentralization, transparency, and independence of monetary authorities have contributed to the growing popularity of the digital currency amongst the advocates of untraditional financial mechanisms, and gradually bitcoin has become both a digital treasure and a currency (Weinstein, 2020). Still, the downside of this digital asset is that its value is subject to different and even strange influences, from market speculation, changes in regulation, and other external crypto economic processes, as well as the behavior of the investors, which can lead to high volatility of the asset and difficulties in forecasting its price.

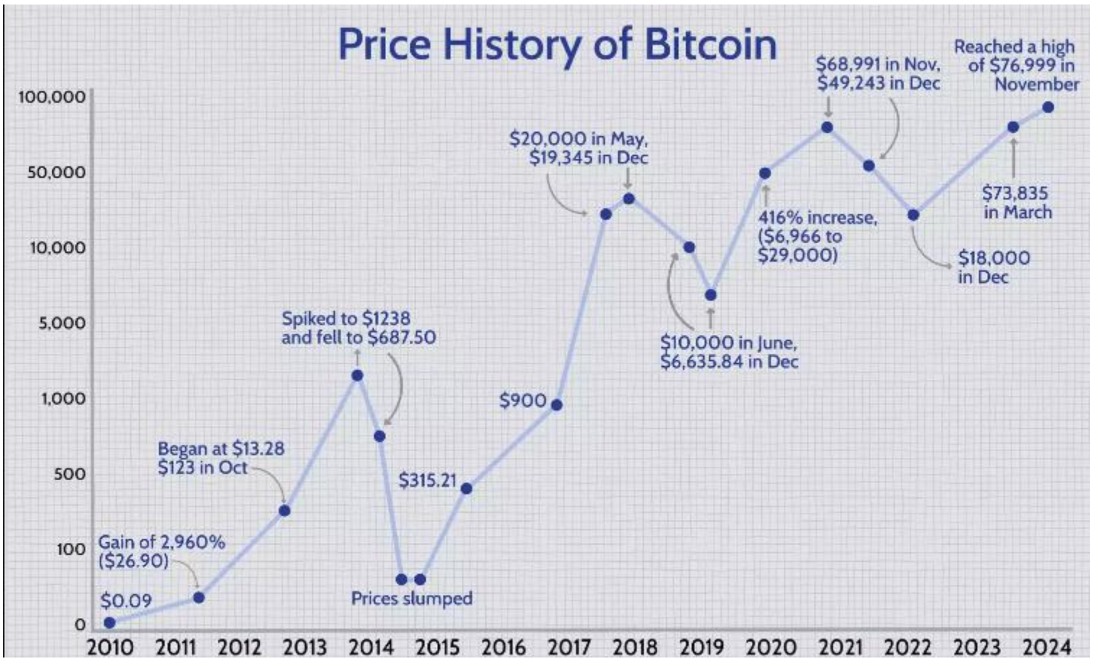

Figure 1: Price History of Bitcoin (2010-2024) (John Edwards, 2024)

Figure 1 illustrates the significant fluctuations in Bitcoin's price over time, highlighting key milestones and price peaks. The dramatic price changes underscore the volatility and unpredictability that make Bitcoin a challenging asset for predictive modelling.

1.2 The Volatility Challenge in Bitcoin Pricing

The price changes of Bitcoin are too fast and significant, making it one of its most prominent features. In contrast to legal tender which is generally stable and backed by the order of a central bank, the price of Bitcoin is subject to several underlying factors that cannot be predicted. For example, the negative or positive position of people with feelings of influence like Elon Musk, and some regulatory constraints can cause sudden value changes. One factor that causes Bitcoin volatility is market sentiment, which is mostly characterized by fear and greed. Another factor is that while stock markets open and close at fixed hours, bitcoin is a seller at any time of the day, all over the world, which brings about price hikes because as news and important matters occur, it directly affects investor behavior at that moment (Katsiampa, 2019).

The finite amount of Bitcoin available for mining is also one of the underlying causes of its price fluctuation. The total issuance is limited to no more than 21 million coins, with new coins being produced at an ever-diminishing pace through the method of mining (Manager, 2024). This supply is fixed, unlike currencies such as the US dollar and the Euro which can be expanded by authorities through inflationary monetary actions. As the currency reaches its maximum supply, however, the price swings due to this factor would likely become excessive, as the level of demand may be higher than the supply. These distinct features in turn generate an exceptionally difficult challenge for predicting the price of Bitcoin and hence its use in this study is very relevant for the advanced predictive modelling employed in this research (Manager, 2024).

1.3 The Role of Machine Learning in Financial Forecasting

The use of methods such as econometric and time-series models for financial forecasting is mostly not effective, particularly for unstable assets such as Bitcoin. Thus standard models for time series such as the ARIMA model, which is an autoregressive integrated moving average, are often constructed making linear assumptions based on the trends of the past data and hence it fails to represent the non-linear dynamics of behaviour such as the movement of Bitcoin price (Roy, Nanjiba and Chakrabarty, 2018). As a result of these obstacles, academicians and practitioners have turned to complicated methods of analysis, especially machine learning (ML) which is suitable for large and complex data sets with non-linear relations (CFI, 2023).

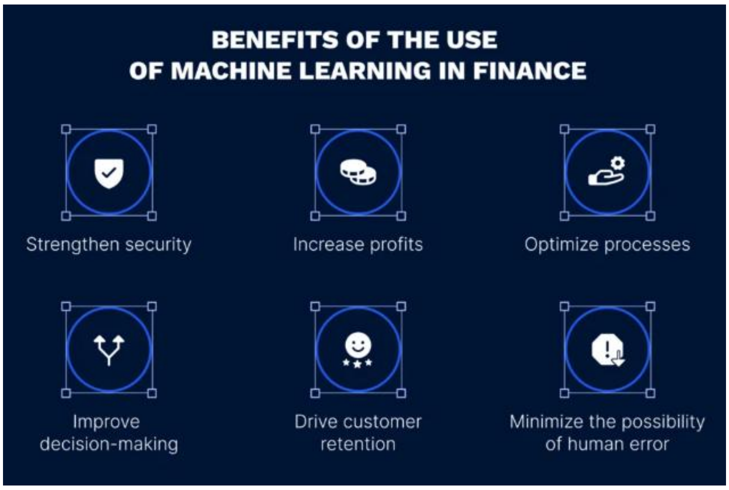

Figure 2: Machine learning usage advantages in finance (Danikovich, 2023)

Machine learning models offer a beneficial solution for advanced financial data analysis considering it is not limited to historical data such as prices but incorporates other regions such as trading volumes, social media, and economic variables. Since ML models are enabled to observe patterns in data and make some adjustments over time, they could be reasonably effective in predicting the price of Bitcoin and especially its sales (Samir Poudel et al., 2023). In this research, three ML models: Linear Regression, Random Forest Regressor, and Long Short-Term Memory (LSTM) neural network, have been chosen to investigate the Bitcoin price prediction and related performance comparison. All of these models have their advantages such as interpretability, resistance to noise, or ability to follow increments of time.

1.4 Research Problem

The behaviour of Bitcoin's price remains challenging, as it is influenced by several global economic drivers, sentiment shifts, and speculation. Traditional methods of financial forecasting struggle to account for the erratic and highly sensitive nature of Bitcoin’s price. Therefore, the research problem explored in this paper is:

How effective are different machine learning models, for predicting the price of Bitcoin and what is the cause of their differences in effectiveness?

The goal of this study is to evaluate three different machine learning techniques for their performance, explainability, and ability to model time series behaviors and fluctuations of bitcoin price. Understanding these differences will help guide the selection of suitable models for cryptocurrency forecasting, benefiting both investors and researchers in the field of financial technology.

1.5 Aims and Objectives

1.5.1 Aims

The objective of this research is to determine how successful certain machine learning algorithms, specifically Linear Regression, Random Forest Regressor, and LSTM neural networks, can be in forecasting the price of Bitcoin. The analysis will compare the models to explain the advantages and disadvantages of each of them in the course of performing research on Bitcoin price data.

1.5.2 Objectives

To achieve this aim, the research objectives include:

- *Data Collection:* A historical dataset of bitcoin prices should be acquired, along with other aspects, e.g., moving average, volatility, and daily returns, to construct a full dataset for training and testing the model.
- *Model Training:* On the given set of data, Linear regression, Random Forest Regressor, and LSTM models will be trained to create bitcoin price forecasting models.
- *Model Evaluation:* Each of those models will be assessed with the help of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score, which are several model performance metrics. This will enable assessments to be made on how accurate and how well each of these models can forecast the output.
- *Comparative Analysis:* Explore the advantages and disadvantages of each of the models about their performance in forecasting the dependent variable, Bitcoin which is known to be volatile, as well as trends and seasonality in the data.
- *Insights for Financial Forecasting:* Based on the above comparative analysis, inform which machine learning models are best suited for predicting cryptocurrencies and other areas of financial technology.

1.6 Significance of the study

Accurately forecasting the prices of Bitcoin has significant importance. This analysis can lead to enhanced forecasting models that can help investors make better decisions and lessen the chances of incurring huge losses, as well as aid in the stabilization of the market. Beyond the scope of this study, the results of this research may also apply to other currencies that are digital in nature as well as to other financial instruments characterized by extreme price fluctuations, where predictive analysis has some relevance. As it investigates different ML models in terms of their applicability to Bitcoin forecasting, this study advances the pursuit of durable financial forecasting strategies for new digital assets.

2. Literature Review

Machine Learning in Cryptocurrency Price Prediction

Historical Evolution of Machine Learning in Financial Markets

Machine learning (ML) in financial markets has significantly evolved over the last few decades. Presently, ML techniques are used for a much wider variety of applications: at the initial stage, they were used simply for a few specific applications such as fraud detection, credit scoring, and algorithmic trading-all these methods mainly employed structural data and normal patterns well processed by traditional algorithms. However, as the complexity of problems grew, with stock market prediction and portfolio optimization as examples, available data became vast, and computer capabilities increased.

As a result of the introduction of neural, ensemble, and advanced time-series models, applications of ML in finance are much more advanced in dealing with non-linear functions and interactions between financial variables, beyond simple prediction. This has been made possible by the very nature of ML, which can take huge datasets, manage heterogeneous features, and find patterns in very noisy and unstructured data. This will also ensure that they are well prepared for application in the highly dynamic domain typical of moving within cryptocurrency price forecasting.

Unique Challenges Cryptocurrencies Pose for Machine Learning Techniques

Cryptocurrencies, as financial instruments, present a distinct set of challenges for machine learning models. Unlike traditional assets, cryptocurrencies such as Bitcoin exhibit extreme price volatility, often reacting unpredictably to unconventional factors. These include social media sentiment, speculative market behavior, regulatory announcements, and technological developments within blockchain ecosystems.

Another challenge lies in the decentralized nature of cryptocurrencies, which operate independently of central regulatory authorities. This decentralization, combined with 24/7 trading across global markets, leads to price fluctuations that are more frequent and pronounced compared to traditional markets. Additionally, the limited historical data for newer cryptocurrencies and the inherent noise in available datasets further complicate prediction efforts. Such characteristics demand ML models capable of adapting to dynamic, complex, and often incomplete data environments.

Differentiation Between Traditional Financial Assets and Cryptocurrencies

Conventional financial assets like stocks and bonds have been determined by the macroeconomic factors influencing institutional regulations usually as well as historical performance trends. The great predictability surrounding assets stems from their governance by central authorities and market mechanisms. That made conventional econometric models as well as standard techniques from ML often enough for modelling the behaviour of those assets.

Unlike that, however, cryptocurrencies operate in an increasingly decentralized frame of operation and unregulated space with their value being most greatly influenced by the kinds of sentiment-driven factors, market adoption rates, and even external shocks like hacking incidents or crackdowns by regulators. Such unpredictability renders traditional models ineffective and paves the way for advanced ML techniques such as neural networks, ensemble methods, and so on along with deep learning. These methods will be very good at resolving the dynamic, non-linear multivariate characteristic of movements of prices in cryptocurrencies.

One principle manifested by the characteristics of cryptocurrencies is the establishment of strong ML models that must address volatility, non-linear dependencies, and unstructured data. Confronted by such challenges, ML would yield several insights regarding cryptocurrency markets, which could benefit investors and researchers.

Comparative Analysis of Machine Learning Models

Simple Models: Linear Regression

Linear Regression is one of the simplest and most interpretable models among all concerning financial forecasting. It identifies the relationship of a dependent variable with one or more independent variables by fitting a linear equation. The efficiency and simplicity of computation make it a widely used one; however, it fails to measure and handle the relationships which are complex, nonlinear, and often complicated as in the case of the extremely volatile and unpredictable behaviour of the prices of Bitcoin. In studies such as the one conducted by Samir et al. (2023), it has been shown that while Linear Regression sets a basis against which the prediction models are evaluated, its inflexibility to the rapid changes within the marketplaces at a less idealized position compared to having predictions made with a more advanced model.

Ensemble Methods: Random Forest

Random Forest is a popular ensemble learning technique that combines multiple decision trees to enhance predictive accuracy and control overfitting. This model does particularly well in high-dimensional and noisy data environments, such as those inherent to the cryptocurrency markers. Random Forest pools predictions from many trees, thereby reducing variance and increasing prediction stability. However, its feature-based importance rankings can sometimes overlook a temporal dependency, a critical aspect in financial time-series data. Random Forest has proved to be highly effective in many studies but may not match the performance of deep learning techniques for sequential or highly time-dependent tasks.

Deep Learning Approaches: LSTM

Long Short-Term Memory Networks, which are a specific type of Recurrent Neural Networks, are best applicable in the modeling of data dependent on time. The important memory cells ensure the retention of information for long periods, and hence LSTMs are excellent at capturing the sequential dependencies and patterns over a time scale. Katsiampa (2019) demonstrates that LSTM networks provide a forecast of Bitcoin prices that is superior to traditional and ensemble models and is particularly good at capturing its temporal dynamics. Of course, the requirement of Computation in LSTMs is much higher than those in other algorithms, formation sensitive to hyper-parameter tuning, and hence makes it complicated for implementation and optimization.

Dataset Ethical Considerations

1. **Data Source**:
   The dataset was retrieved from **Yahoo Finance**, a publicly accessible and reputable platform providing financial data for academic and personal use.
2. **Anonymization and GDPR Compliance**:
   The dataset contains no personal or sensitive information, such as names or images, and does not fall under GDPR regulations.
3. **Permission and Usage**:
   Yahoo Finance allows data use for academic purposes under its **Terms of Service**, provided it is not for commercial use.
4. **Ethical Collection**:
   The data is sourced directly from Yahoo Finance and has been ethically collected without involving personal data or social media content.
5. **UH Ethical Approval**:
   This project does not require UH ethical approval, as it does not involve surveys, personal data, or social media platforms.

3. Methodology

3.1 Data Description:

The dataset incorporated in the current study reflects various features of the Bitcoin market throughout a relatively long time. It has a time-indexed format, in which records information includes daily opening and closing prices, daily highest and lowest prices, and daily turnover. The availability of derived features such as the moving average for different time spans, daily returns, and volatility increases the data’s analytical value. These attributes give a three-in-one perspective of the price of Bitcoin as it shows the trends, short term volatility and the general stability of the market. The preprocessing steps undertaken in this work were useful in removing missing values which could have skewed the results obtained by the models used in the study as well as normalizing the features to enhance our ability to interpret the results obtained from the models.

The Bitcoin's price fluctuations over the time, the comprehensive dataset also captures the underlying nature of Bitcoin's markets. Hence, these characteristics make the dataset useful to provide input for sophisticated machine learning methodologies developed to predict proper price fluctuations.

3.2 Data Preprocessing

3.2.1 Data Collection

The information used for this research was excellent and collected with lot of care from the yahoo finance web site which is reliable and contains updated information. This plat form was used because of its effectiveness and ease in sourcing historical price data of Bitcoin. Indeed, the dataset is spread across several years allowing the analysis captures fluctuations and steady state of the market. Daily Open, Close, High, and Low Prices are important data attributes; there is also Trading Volume data attribute, which reflects the supply and demand in the markets. These raw data was complemented with engineered features to increase possibilities of the machine learning models when doing the analysis. For example, the Moving Averages for 5-day, 10-day, and 50-day periods were computed with a view to maintaining the average for the short and the long-term price fluctuation. Recall also that Volatility, a key variable in forecasting financial rates, was obtained with the help of the moving average of such main financial coefficients as the standard deviations of the daily returns for the subsequent 14 days. Further, affording timescale to vitality and other variables, it assists in quantifying volatility and final price inherent in Bitcoin trading. Further, Daily Returns which is the percentage change on closing prices from the previous day were also calculated in order to analyse daily fluctuations in the market.

The data was cleansed and transformed to make it ready for issue to the modeling stage. Whereby features are missing, measures were put in place to deal with this as the results could have been biases and or errors in the predictions. Noise observations that might be due to measurement errors or some other specific source of market disturbance were also addressed as to whether such observation should be considered as valuable data or as potential sources of distortions of result.

3.2.2 Data Preprocessing

Data preprocessing is a vital step to undertake when the goal is to feed and train the machines with a high-quality reliable dataset. In this study, various methods were used to clean the Bitcoin pricing data in order to making an analysis. First, the experiments detected missing values in the dataset and then corrected them. Inaccurate data, which may come from incomplete data records or errors in systems, will create biases or inaccuracies in even the most predictive models. To avoid this, wherever possible, methods of imputation were applied whereas where whole rows are heavily missing, such rows were either dropped to have clean data or imputed wherever it was unavoidable.

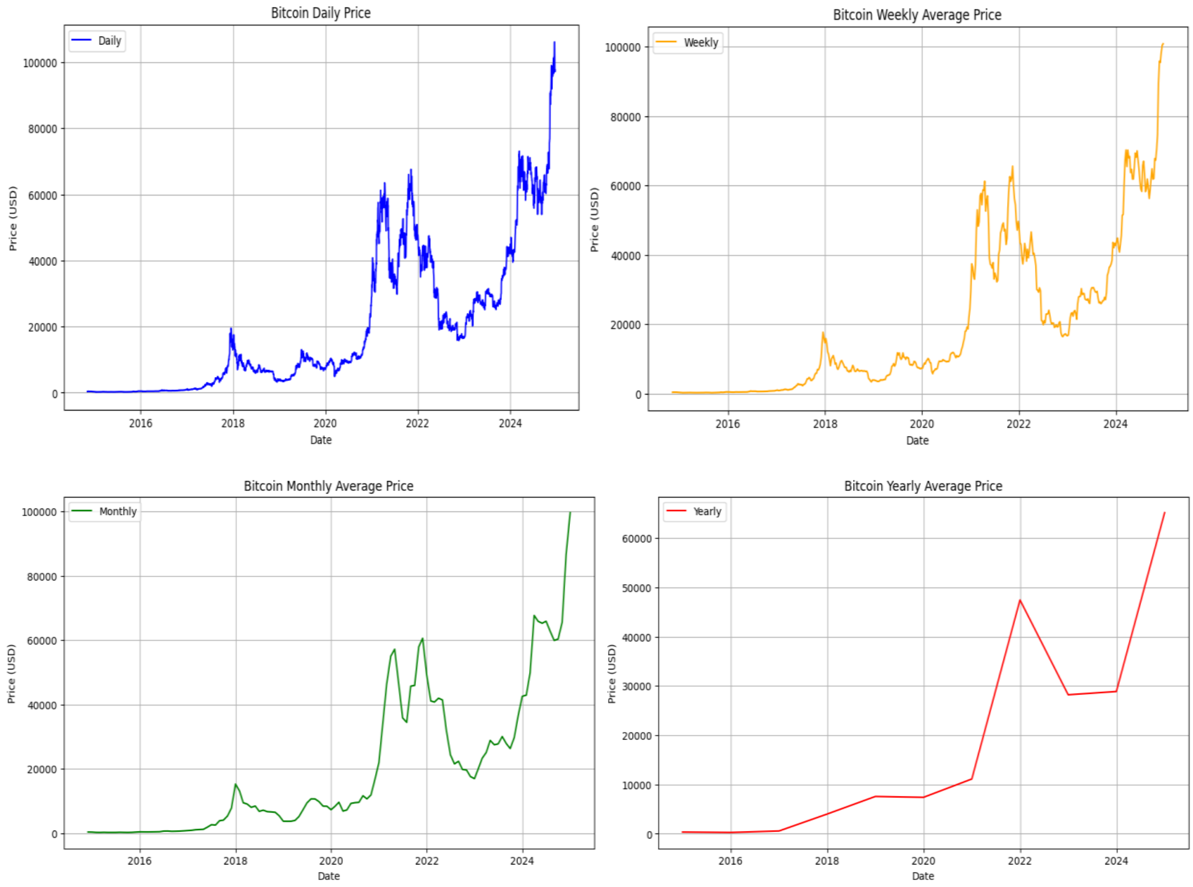

Figure 3*: Showing the average prices of Bitcoin on daily, weekly, monthly and yearly bases.*

Another topic of interest during the preprocessing stage was outliers which is a point or observation that is significantly different from the other points in a given data set. The volatility control of Bitcoin is relative to market events, and thus price outliers were tested to understand their contribution in the analysis. Instead of deleting these values out rightly, domain knowledge was used to distinguish between signals of market practice and noise. As such, we were able to retain the outliers that are relevant for the model in order to accommodate the inherent volatility of trading in Bitcoin. Normalization was done in an attempt to make the features of the dataset to be in the same distribution. To ensure equal importance of different numerical ranges, they normalized the values so that the larger numerical ranges would not overwhelm the overall prediction. In order to normalize the features, the standardization techniques like z-score were applied on the features to scale all the features into a similar range without altering their distributions.

Finally, as it will be discussed later, the set was divided into a training and a testing set which is paramount when it comes to model verification. The training set which contained data 80 % of the total dataset were used in developing the models while 20% of the entire data was used in the test set for accuracy determination. This division made sure that the measure used to gauge performance of the models was not over fitted and tried out on unseen data thus giving a true picture on the generality of the models.

3.3 Evaluation Metrics:

Evaluation metrics are an important means of assessing the effectiveness of models for machine learning. In this study, three prominent metrics were utilized to assess the models' accuracy and reliability in predicting Bitcoin prices. The assessment measures used herein are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination R^2. All these measures gave different about the deployed performance of the models in terms of prediction.

Mean Absolute Error (MAE):

MAE is one of the easiest to understand and interpret broader metrics since it is the average of the error's absolute value irrespectively of error direction. Mathematically, it is expressed as: stands for actual value stands for predicted value and stands for total number of predictions. Whereas the lower MAE value means that the predicted results are closer to the real results of the model.

Unfortunately, because it doesn’t differentiate between the magnitudes of the errors, MAE does not give large errors as much weight, and as such, is not very sensitive to large errors. In the case of Bitcoin, MAE played a crucial role in approximating the average variations between the forecasted and the real price. It revealed the overall predictive efficiency of each model, and thus, which of them acted uniformly throughout the entire dataset.

Root Mean Squared Error (RMSE):

RMSE is a popular measure in regression tasks used especially for models working with continuous data. It is calculated as: where , , and are the actual values, the predicted values and the number of predictions, respectively. RMSE sums up the squared differences between the predicted and actual values average i.e., it increases the weight of large errors. He then re-applys the square root to bring the value in the scale of the original data again. This characteristic of RMSE is convenient for the fact that it punishes large errors more often, but will not allow the model to be strongly affected by small errors. When this value is low they make fewer large predictive errors, and therefore the RMSE value is set at a lower value. In this study, RMSE was used as a measure that gave a deeper insight into the stability of the models in relation to the notorious volatility associated with Bitcoin prices by identifying its origin in outliers.

As it reduces the influence of small inconsistencies towards outliers, while dwelling on larger errors, RMSE was instrumental in determining the accuracy of predictions in situations where relatively small deviations equalled massive mismatches in economic terms.

Coefficient of Determination (R^2):

The R^2 metric, or coefficient of determination may be expressed as the number that describes the proportion of Total Variance in the dependent variable that is explainable by one or more independent variables in the model. It is calculated as: where represents the real condition, are the forecast condition, and is the average of the real condition. The coefficient R^2 varies from 0 to 1 and, by moving up to the value of 1, means that a model accounts for a higher proportion of variance in data. Negative values of the R^2 indicate that the model enjoys a worse predictive precision than the mere mean forecast, albeit this is not extremely frequent.

In this research, R^2 was crucial to determining how much of the underlying pattern in Bitcoin prices each model was able to explain. Higher ‘R^2’ value pointed toward the view that the applied model captured the paper’s dataset complexities better, and vice versa. Related to the error-based measures (MAE & RMSE), R^2 also provided a more wholesome approach to evaluate performance of the models especially in terms of explanatory power.

3.4 Model Implementation:

It was done to optimize the machine learning models which includes Linear Regression, Random Forest and LSTM for handling the complexity of the Bitcoin prices. Linear regression was used first as a norm with its straightforward and easily understandable logic form the basis for subsequent models. It was set up to make prognosis of Bitcoin prices depending on certain characteristics, which would give a benchmark for estimating more complicated models.

Since distinct features of the financial datasets are related in non-linear manners, Random Forest, an ensemble algorithm was applied. A multicenter of decision trees was used in this model where each tree was trained on a different subset of the data to increase the model’s precision without compromising its ability to generalize. The number of trees and the maximum tree depth parameters were fine-tuned by using a grid search.

Linear Regression

Linear Regression is one of the simplest Machine Learning Algorithms and therefore acts as a reference for this work. The main goal of using Linear Regression was to define the presence and nature of links between Bitcoin prices and the considered characteristics. This model goes behind the notion that assumes a linear relationship between independent variables and the dependent variable like certain features as moving averages, the daily return and the volatilities and the price of Bitcoin.

The implementation process started with the feature selection because it important to consider only those features that can provide meaningful insights into the model. These features were scaled to the interval of [0, 1] for better comparison because in regression, it is essential to work with features of the same scale. They were then divided into training and test data set with 80% and 20% respectively. The training process inc default the exogenous movable training process involved fitting of the linear model with the use of the least squares technique which gives minimized squared residual sum. By using this approach, the SL model shall be capable of finding accurate linear equation that explains the relationship of the input variable with the target variable.

However, there are certain benefit that can be derived from Linear Regression A simple model indeed has its benefits. It is very understandable as to how each of the features goes into the actual prediction process. This makes it relatively easy to implement, which may make it suitable for setting a base line against which actual sophisticated models can be compared. However, its major drawbacks are that it is designed for substantive variables rather than fluctuating variables, such as the Bitcoin price shown in Figure 1. Linear Regression is not good at capturing multivariate and thus it does not work well in problems that have high variance and noise. Therefore, the model offers the basic idea but does not have enough depth to construct pessimistic forecasts in the constantly changing conditions.

Random Forest Regressor

Random Forest Regressor is an accurate ensemble learning algorithm best suited for applications requiring high, non-linear interaction between predictors. This model was chosen due to its applicability in handling noisy and high-dimensional data, which are characteristics of Bitcoin since it is able to capture high-frequency data. Random Forest works in a similar way to Individual decision tree’s and AdaBoost with the difference that it builds several such trees during training and then arrives at the final class by consolidating the voting from all trees. This makes the method practicable through an ensemble approach that reduces prediction error and overfitting.

Feature extraction and preprocessing were performed as part of the preparatory process to the training of the Random Forest Regressor. The process of hyperparameter tuning was important in determining model optimization. The number of trees, maximum depth and minimum samples required split were tuned using grid search while cross validation was used to embrace all the models. Variable importance measures were also used in the study in order to determine the influence of each attribute to the prediction done by the model.

Nevertheless, overfitting is not a problem for Random Forest, especially for working with high-dimensional data since it is an ensemble of decision trees. Through voting means, the model lowers variance and generalizes since the odds coming from several trees are averaged. Furthermore, feature ranking ability of care helps in understanding the significance of features in the analysis of the data set.

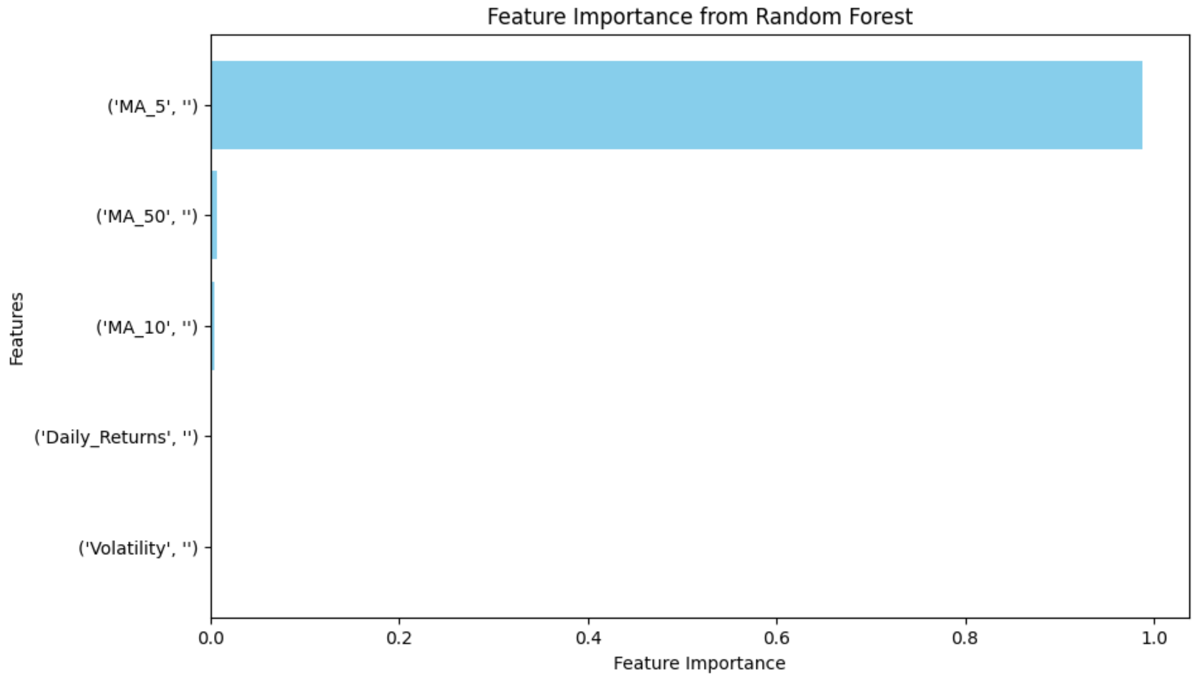

Figure 4*: Showing the feature importance for the Random Forest Model.*

Long Short-Term Memory (LSTM) Networks

Detailed investigations of temporal dependencies in Bitcoin price data required the ISO exclusion of general Recurrent Neural Network (RNN) and the use of Long Short-Term Memory (LSTM) networks that are optimized for sequential modeling of RNNs. LSTMs are particularly well-suited for the class of information that needs to be stored for longer periods which is exactly the case in the time-series forecasting. This capability is especially helpful for capturing detailed features and patterns of variation of Bitcoin prices.

We needed to create a specific architecture of the LSTM model to proceed with the complexities of data sets. The input data was converted into sequences through the technique of sliding window; historical data used in the current analysis served as basis for analysis on sequences to facilitate prediction on future outcomes. The model under consideration, LSTM network, had more layers; it imposed an input layer, one or several hidden layers containing LSTM units, and it also contained the output layer, which was fully connected. The remaining hyperparameters which include the number of hidden units the learning rate, the batch size and the sequence length were adjusted in the tuning phase for the best performance.

To reduce the risk of overfitting and improve model generalization the use of regularization techniques such as dropout layers was included. Adam was employed as the optimization algorithm for training the model and is one of the most popular techniques to be used in deep learning. For the specific purpose of loss function selection, the commonly used Mean Squared Error (MSE) was selected because of its relevance to reducing prediction errors that are being taken into consideration in the study at hand.

There are however a few disadvantages in using LSTMs given that they provide exaggerated benefits in modeling sequential data. Even the simple training of an LSTM model presents a number of high computational demands compared with other machine learning tools, which makes the deployment of LSTM models in existing architectures or with sufficient computational power a much more complex proposition.

4 Results and Discussions

The findings of the Bitcoin price prediction models—Linear Regression, Random Forest, Random Forest, and Long Short-Term Memory (LSTM) — show the strength and weakness of these methods while dealing with the chaotic nature of forecasting, specifically in the field of finance. These results are grouped by signal, fine-tuning outcome, and 30-day forecasts, so that we can compare the dynamics of each model.

Linear Regression had significantly higher accuracy with MAE = 1220.20, RMSE = 1830.65 and R² of 0.9924. These metrics point out a high degree of accuracy since the model is able to reflect the trends in the historical values as depicted below. Moreover, Mean Absolute Percentage Error; MAPE of 2.45 percent supports its capability to generate accurate estimates with minor variation from actual prices. When retrained further, Linear Regression got better with a lesser MAE of 1181.13, RMSE of 1754.55; the model has maximum adjusted R² of 0.9931 and minimum MAPE of 2.37%. Some of these enhancements were done after fine tuning of factors including the polynomial degree and the ridge regularization. These results confirm the applicability of the model in situations where the linearityprevails, however, the linear approach of the Linear Regression model is shown to be insufficient for capturing non-linear and fluctuating nature of Bitcoin prices.

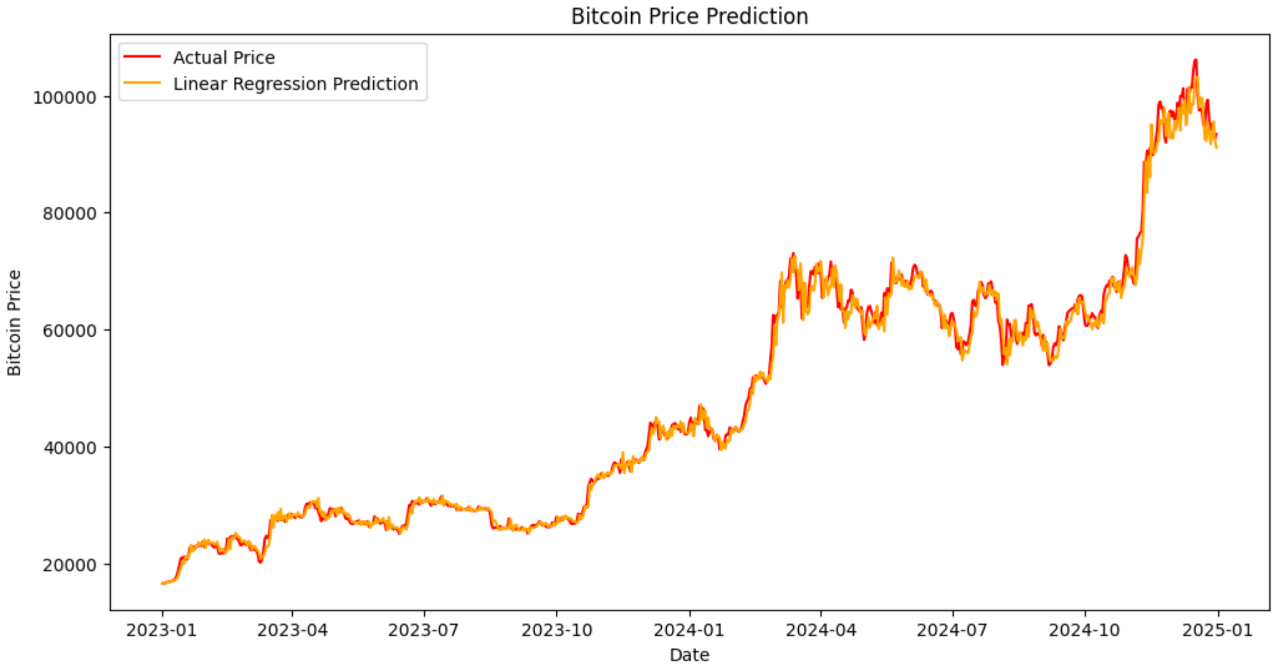

Figure 5*: showing line plot with difference of actual and predictions of Linear Regression.*

Analyzing the Random Forest which best fits for dataset with non-linearity and high level of noise, the result was also not very optimistic yet not very pessimistic, in other words it was average. Hence the MAE of 3888.63, realizing an RMSE of 8717.58 and an R² value of 0.8286, the generalized exposed model performed less than even Linear Regression. The MAPE of 6.02% supports the nature of the model that aiming for high levels of accuracy in predicting the Bitcoin price is challenging. Nevertheless, it provided good information in feature ranking where 5DMA represented a stronger value with a contribution of 98.80%, followed by 10DMA and 50DMA, daily returns, daily volatility at a considerably lower contribution. More tuning continued to enhance the parameter of counts of estimators, maximum depth and minimum samples for split nodes which in fact generated only fine-tuning enhancements. The higher volatility of Bitcoin price might have had an impact on the results because the model being built on a tree approach may not catch the temporal dependencies present in the time-series data well.

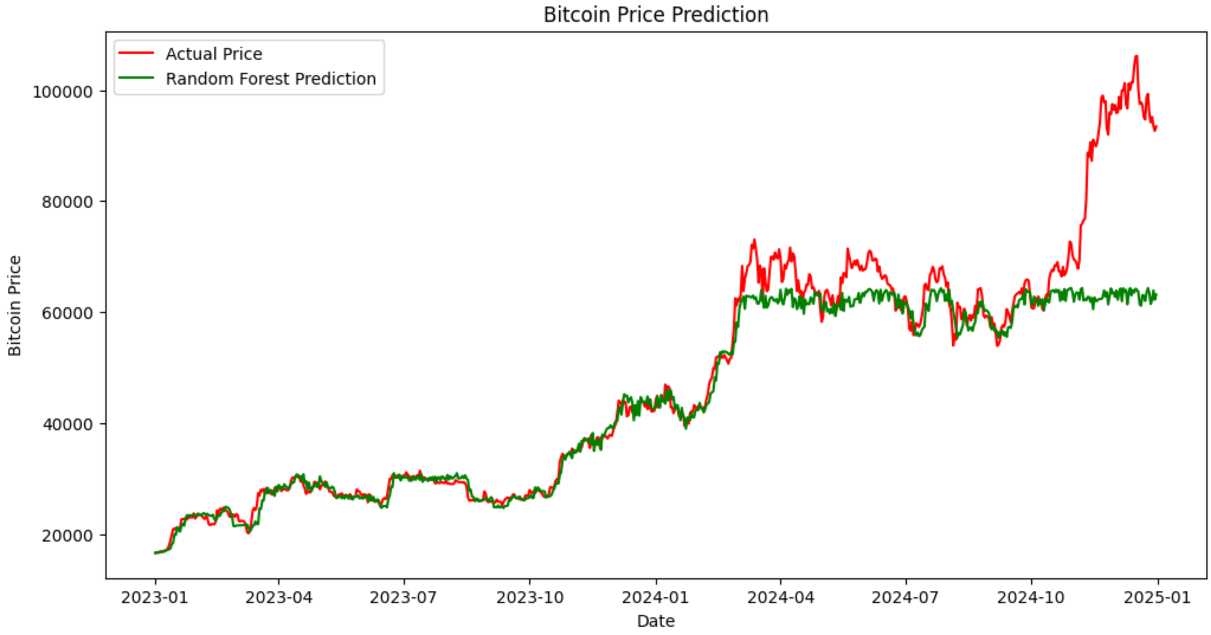

Figure 6*: Showing line plot with difference of actual and predictions Random Forest Models.*

Time series performance of the proposed LSTM model was ascertained by the fact that it was less erroneous than the Random Forest model. LSTM for its part, achieved an MAE of 1456.71, RMSE of 2087.55 and an R² value of 0.9903, thus proving that it is capable of modeling sequential data. The MAPE of 3.12% thus goes a long way in affirming the model's reliability even if it was a whisker lower than the fine-tuned Linear Regression model. The flow structure of LSTM, with cells of memory and transfer of information from one time step to the next allowed it to learn the temporal characteristics of Bitcoin price fluctuations, and the volatility that was more accurately captured compared to Random Forest. While training the model, however, other factors, for instance, the number of hidden units, learning rate and the number of factions in a sequence had to be adjusted to ensure that the model is capable to learn without overspecializing.

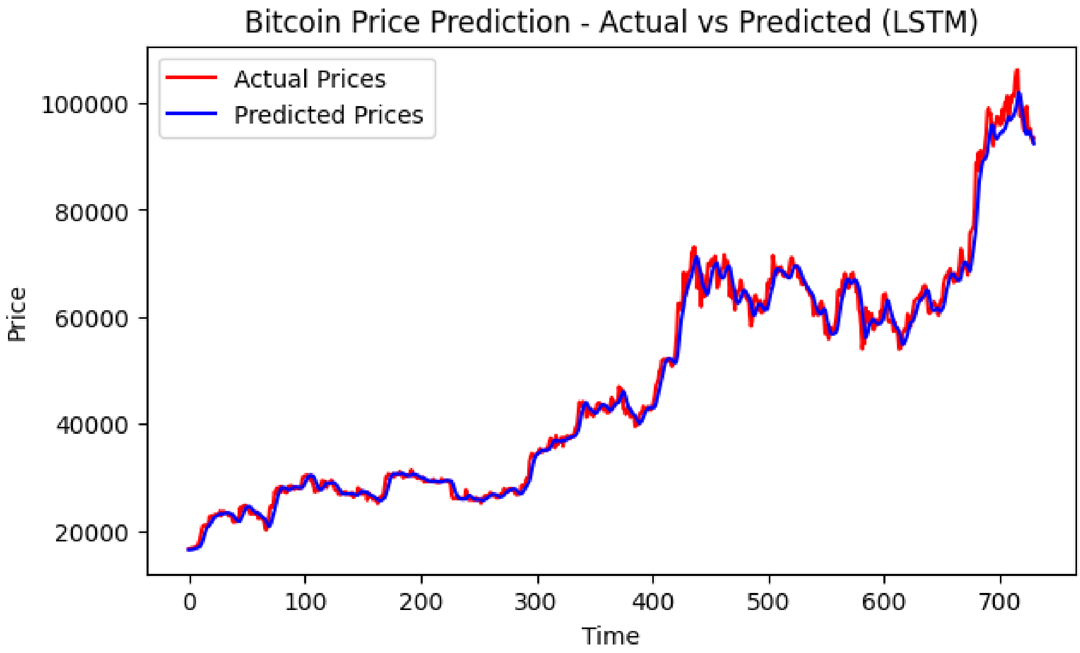

Figure 7*: Showing a line plot with difference of actual and predictions of LSTM model.*

Future Predictions:

Some foreseen uses of the models can also be ascertained from the future predictions study. As for Linear Regression, the prices to be expected for the next 30 days remained smooth curves that provide the model the capability of generalizing with the baseline patterns. For instance, prediction on the first day was $93,941.35, while that of the last day was $118,490.73. This steady progression is consistent with the linear assumptions of the model but it may well omit some occasional high or low prices. Random Forest, on the other hand, generated highly volatile projections: $100,402.21 of the first day, $85,554.33 of the third day, and $112,565.43 of the tenth day. They suggest that the model pays attention to the dataset's non-linearity and other intricacies but also its weakness to noise presence. Random Forest and LSTM had similar future projection but differed slightly on the daily projection starting from $100,402.21 on the first day and $101,348.33 on the 30th day and the between day transitions. The equilibrium between capturing volatility while sustaining consistent trends underlines LSTM's efficiency on dependence existing through sequences.

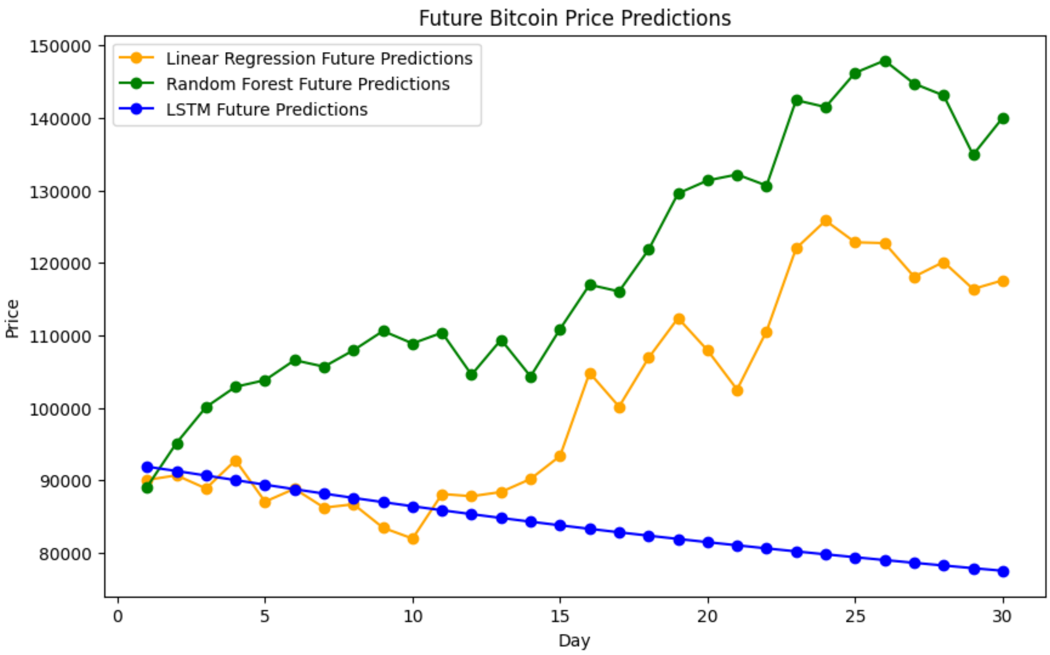

Figure 8*: Showing the future predictions of all three Models*

Consequently, a comparative analysis of these models will show the authors' intention to highlight certain advantages and deficiencies of the models being compared. SinceLinear Regression has high accuracy and also since its interpretation is easy, it is always useful to consider it as a good starting point, especially if the data being analyzed has mainly linear relationships. However, it does not capture nonlinearity and temporal dependencies and thus is suboptimal in highly volatile environment such as cryptocurrency markets. Random Forest was good at feature selection and handling interactions but did not perform well regarding sequential dependencies of Bitcoin prices. It was found out that LSTM type of model was the most suitable for this particular task due to its capacity to deliver both high accuracy and temporal features. However, its computational consumption raises practical implications especially when used in real-time forecasting.

Conclusion:

This study presented a comprehensive exploration of Bitcoin price prediction using three machine learning models: Linear Regression, Random Forest and Long Short-Term Memory or LSTM. Through these comparisons using performance evaluation criteria and forecast figures of the models, useful information was derived from the analysis regarding the abilities and drawbacks of the models when dealing with the dynamic and uncertain characteristic of cryptocurrency price information. It is vital for analysts to analyze and decide the missing links between the precision, computational time, and interpretability of the model while engaged in predicting in financial markets.

For Linear Regression, which is simple and easy to interpret the model delivered a fairly high accuracy with MAE of 1181.13 and RMSE of 1754.55 in the Fine-tuned model. The calculated R-squared (R^2) of 0.9931 confirmed good predictive power of linear tendencies within the dataset. The authors further stated that the failure of the model to account for nonlinear relationships and temporal dependence made its application challenging in the very volatile Bitcoin context. Nonetheless, it is high level of efficiency and low level of complexity that provide simple model as a good benchmark for comparison with other models.

Specifically, in terms of feature importance, Random Forest outperforms other nonlinear ensemble models without distortion to imply that the 5-day moving average is the most crucial predictor of Bitcoin prices, even though it had an MAE of 3888.63 and the R^2 of 0.8286, the predictive performance of the model was weak, being unable to handle the sequential characteristics of the time-series data appropriately. However, these limitations were seen with Random Forest, and this technique was useful in identifying feature interaction and non-linearity which is important in the analysis of data sets.

As for the capability, LSTM networks became claimed as the most preferable and efficient for dealing with the Bitcoin price since they are able to model sequential dependencies and capture inherent nature of volatility of the dataset. Using the MAE of 1456.71 and the RMSE of 2087.55, the LSTM model had the R^2 equals to 0.9903, this means that the proposed model is useful for time-series forecasting of energy consumption. Nevertheless, the computational requirement for training LSTM is a noteworthy issue pointing to the promising trade-off in-between the model sophistication and real-time computation capability.

Future Work:

As a result, the present study opens up several opportunities for subsequent research to improve ML models for more accurate Bitcoin price prediction. Another approach is the creation of a linear regression model, RF – Random Forest and LSTM – Long short-term memory.

Another area of work that can be taken up in future is to add more data into the model to make the prediction more comprehensive. Incorporating other external sources such as sentiments extracted from the social media platforms, macroeconomic indicators and blockchain transaction data might improve understanding of the market to build on the current features derived from historical prices namely moving averages and price volatility.

Another area that enhances recognition as a prerequisite to integration within the framework of deep learning architectures is another exciting research direction. It is also noteworthy that with current approaches and techniques, transformer models, which demonstrated high effectiveness when applied to NLP tasks, can be potentially adapted for time series forecasting in financial markets. Transformers' application in the overcoming long-distance relationships and the parallelism of the computations employed to work may solve several weaknesses seen in LSTM models, which include sensitiveness to hyperparameter optimization as well as higher computational demands.

Reference list

CFI (2023). *Machine Learning (in Finance)*. [online] Corporate Finance Institute. Available at: https://corporatefinanceinstitute.com/resources/data-science/machine-learning-in-finance/.

Danikovich, D. (2023). *Machine Learning in Finance: an Overview*. [online] EffectiveSoft. Available at: https://www.effectivesoft.com/blog/machine-learning-in-finance.html.

Edwards, J. (2024). *Bitcoin's Price History*. [online] Investopedia. Available at: https://www.investopedia.com/articles/forex/121815/bitcoins-price-history.asp.

Katsiampa, P. (2019). *Volatility Estimation for Bitcoin: a Comparison of GARCH Models Volatility Estimation for Bitcoin: a Comparison of GARCH Models*. [online] Available at: http://shura.shu.ac.uk/16526/1/Katsiampa-VolatilityEstimationforBitcoin(AM).pdf [Accessed 25 Apr. 2023].

Manager, B.D.Z.A.P. (2024). *Why Is Bitcoin Volatile? an Overview of Bitcoin Price Fluctuations | VanEck*. [online] Why is Bitcoin Volatile? An Overview of Bitcoin Price Fluctuations | VanEck. Available at: https://www.vaneck.com/us/en/blogs/digital-assets/bitcoin-volatility/.

Parsons, S. (2005). Introduction to Machine Learning by Ethem Alpaydin, MIT Press, 0-262-01211-1, 400 pp., $50.00/£32.95. *The Knowledge Engineering Review*, 20(4), pp.432–433. doi https://doi.org/10.1017/s0269888906220745.

Roy, S., Nanjiba, S. and Chakrabarty, A. (2018). Bitcoin Price Forecasting Using Time Series Analysis. *2018 21st International Conference of Computer and Information Technology (ICCIT)*. Doi:https://doi.org/10.1109/iccitechn.2018.8631923.

Samir Poudel, Rajendra Paudyal, Burak Cankaya, Sterlingsdottir, N., Murphy, M., Pandey, S., Vargas, J. and Khem Poudel (2023). Cryptocurrency Price and Volatility Predictions with Machine Learning. *Journal of marketing analytics*, 11(4), pp.642–660. doi:https://doi.org/10.1057/s41270-023-00239-1.

Weinstein, L. (2020). *Research Guides: Fintech: Financial Technology Research Guide: Cryptocurrency & Blockchain Technology*. [online] guides.loc.gov. Available at: https://guides.loc.gov/fintech/21st-century/cryptocurrency-blockchain.