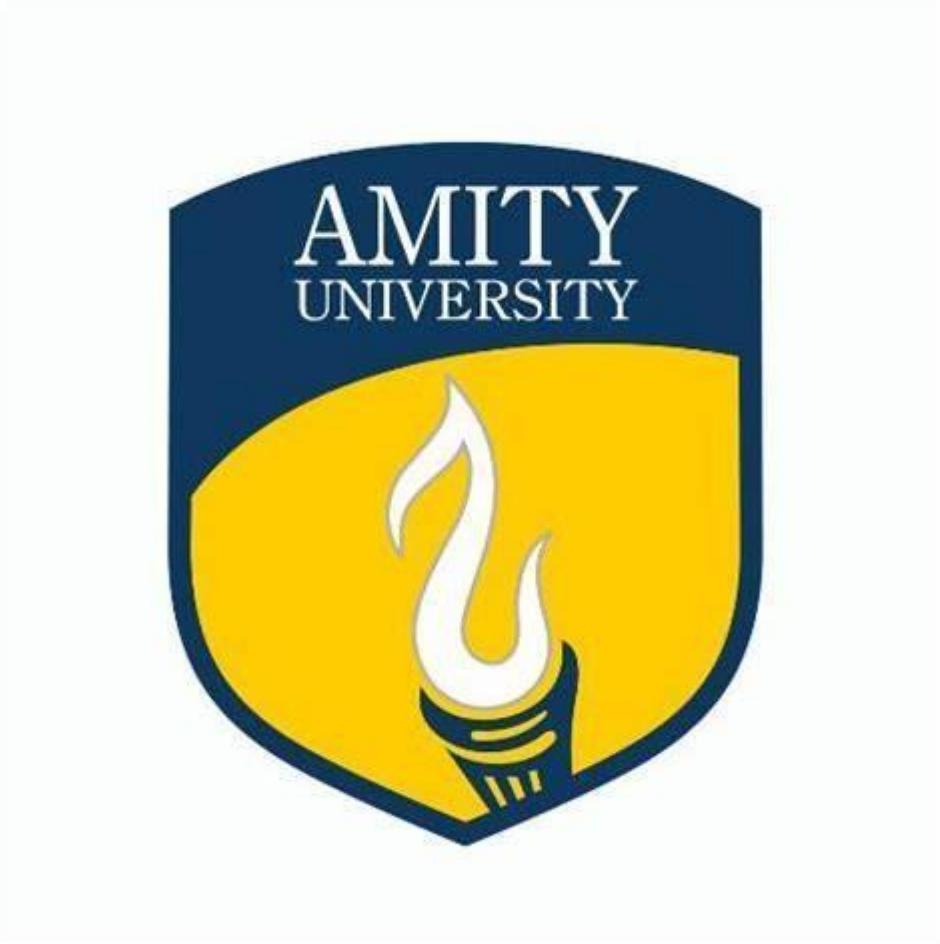
AMITY UNIVERSITY
UTTAR PRADESH



Short-Term Bitcoin Price Prediction Using LSTM Neural Networks

Submitted to:

DEPARTMENT OF STATISTICS

Submitted By:

NITESH

MSc. Data Science

ENROLL NO:A044161824008

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my subject instructors and the department faculty for their continuous support, insightful guidance, and valuable feedback throughout the completion of this integrated project.

A special thanks to Dr Devashish Sir for providing conceptual clarity and real-world applications that helped shape the direction of this work. This project was a valuable opportunity to apply theoretical concepts from multiple courses to a practical, real-world business problem. I am also thankful to Kaggle which provided me this dataset, which served as the foundation for this analysis.

Lastly, I thank my classmates and peers whose discussions and suggestions have been incredibly helpful in refining this project.

Table of Contents

1.INTRODUCTION

In the rapidly evolving and technology-driven landscape of digital finance, **Bitcoin** has established itself as the most influential and widely recognized cryptocurrency across global markets. Since its inception in 2009, it has transformed the perception of monetary exchange by introducing a decentralized and transparent system that operates without the need for intermediaries such as banks or governments. This decentralized architecture, powered by blockchain technology, ensures immutability, security, and traceability of transactions—qualities that have not only made Bitcoin a revolutionary financial asset but also a subject of intense study and innovation in the fields of data science, artificial intelligence, and financial analytics.

Despite its growing popularity, one of the most challenging aspects of Bitcoin is its **high price volatility**. The value of Bitcoin fluctuates dramatically within short periods, driven by a multitude of factors including market speculation, macroeconomic conditions, regulatory announcements, and social media sentiment. These rapid and unpredictable movements pose significant challenges for investors and traders, especially those attempting to forecast price trends and mitigate financial risks. As a result, developing reliable and data-driven models for Bitcoin price prediction has become an important area of research, combining techniques from time series analysis, machine learning, and deep learning.

This project aims to harness the predictive power of **Long Short-Term Memory (LSTM)** neural networks to model and forecast Bitcoin's short-term price dynamics. LSTMs, a specialized form of recurrent neural networks (RNNs), are particularly effective for sequential data because they retain information from past observations to capture long-term dependencies and temporal patterns. By leveraging historical Bitcoin market data—comprising attributes such as opening price, closing price, trading volume, and volatility—this study attempts to uncover underlying trends and generate accurate short-term forecasts. The integration of robust preprocessing, feature engineering, and deep learning methodologies not only enhances the model's predictive accuracy but also provides valuable insights into the behavior of the cryptocurrency market. Ultimately, this project contributes to the growing intersection of **data science and financial technology (FinTech)**, demonstrating how intelligent systems can assist in decision-making, portfolio optimization, and risk management in the volatile domain of digital assets.

2. OBJECTIVE

The primary objective of this project is to design, implement, and evaluate a **short-term Bitcoin price prediction system** utilizing the capabilities of **Long Short-Term Memory (LSTM)** neural networks. The model focuses on forecasting the **closing price ("close")** of Bitcoin, which serves as the **target variable** and represents a crucial indicator of market trends and trading outcomes. This research integrates fundamental principles from multiple disciplines of data science—specifically **data mining**, **artificial intelligence**, **database technologies**, **multivariate data analysis**, and **big data analytics**—to create a comprehensive and interdisciplinary solution that reflects the real-world application of advanced computational methods in financial prediction.

To achieve this goal, the project first aims **to collect and preprocess historical Bitcoin market data**, ensuring the dataset is accurate, consistent, and ready for analytical modeling. This includes handling missing values, detecting and treating outliers, converting date-time formats, and normalizing numerical attributes for model compatibility. The second objective is **to conduct an in-depth exploratory data analysis (EDA)** to uncover meaningful insights, patterns, and interrelationships among key financial features such as opening price, trading volume, high and low values, and daily fluctuations. Through statistical summaries and visualization techniques, this phase provides a deeper understanding of Bitcoin's historical behavior and volatility patterns.

The third objective is **to design, train, and optimize an LSTM neural network** that can effectively learn temporal dependencies from past observations to predict future closing prices. By leveraging the sequence-learning capability of LSTM, the model aims to capture both short-term fluctuations and long-term trends in Bitcoin's price movements. The final objective is **to evaluate the model's predictive performance** using a combination of statistical metrics—such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2)—along with visualization-based assessments comparing actual and predicted values. Together, these objectives create a structured workflow that demonstrates the integration of advanced machine learning, analytical reasoning, and database-driven data management within a unified data science framework.

3. DATASET DESCRIPTION

The dataset utilized in this study consists of comprehensive **Bitcoin market data for the year 2017**, sourced from a reputable **Kaggle repository**. It encompasses highly granular trading information recorded at **minute-level intervals**, capturing real-time fluctuations in Bitcoin's market activity. The dataset contains approximately **525,599 rows and 9 columns**, making it sufficiently large and detailed to support deep learning–based time series forecasting and advanced multivariate analysis. Such a rich dataset provides a strong empirical foundation for understanding Bitcoin's volatile behavior and developing predictive models capable of learning complex temporal dependencies.

The attributes included in the dataset are as follows:

- **unix** – A UNIX timestamp representing the exact time of each recorded observation, which can be converted to a human-readable date format for time series analysis.
- **date** – The corresponding calendar date and time in standard format, used for chronological organization and visualization.
- **symbol** – The identifier of the cryptocurrency pair, specifically *BTCUSD*, representing Bitcoin traded against the U.S. dollar.
- **open** – The opening price of Bitcoin at the beginning of each minute interval.
- **high** – The highest price achieved within that specific time interval.
- **low** – The lowest price recorded during that same interval.
- **close** – The **closing price** of Bitcoin at the end of each interval, which serves as the **target variable** in this project's predictive modeling.
- **Volume BTC** – The total number of Bitcoins traded within the given minute.
- **Volume USD** – The total trading volume expressed in U.S. dollars for the same interval.

This dataset captures the **dynamic microstructure of the Bitcoin market**, offering detailed insights into its short-term price movements, trading intensity, and volatility patterns. The inclusion of both price-related and volume-related metrics makes it suitable for **multivariate analysis**, allowing the exploration of correlations between trading activity and price fluctuations. Moreover, the minute-by-minute resolution provides sufficient temporal granularity for advanced **LSTM-based sequence modeling**, which relies on sequential dependencies to predict future outcomes.

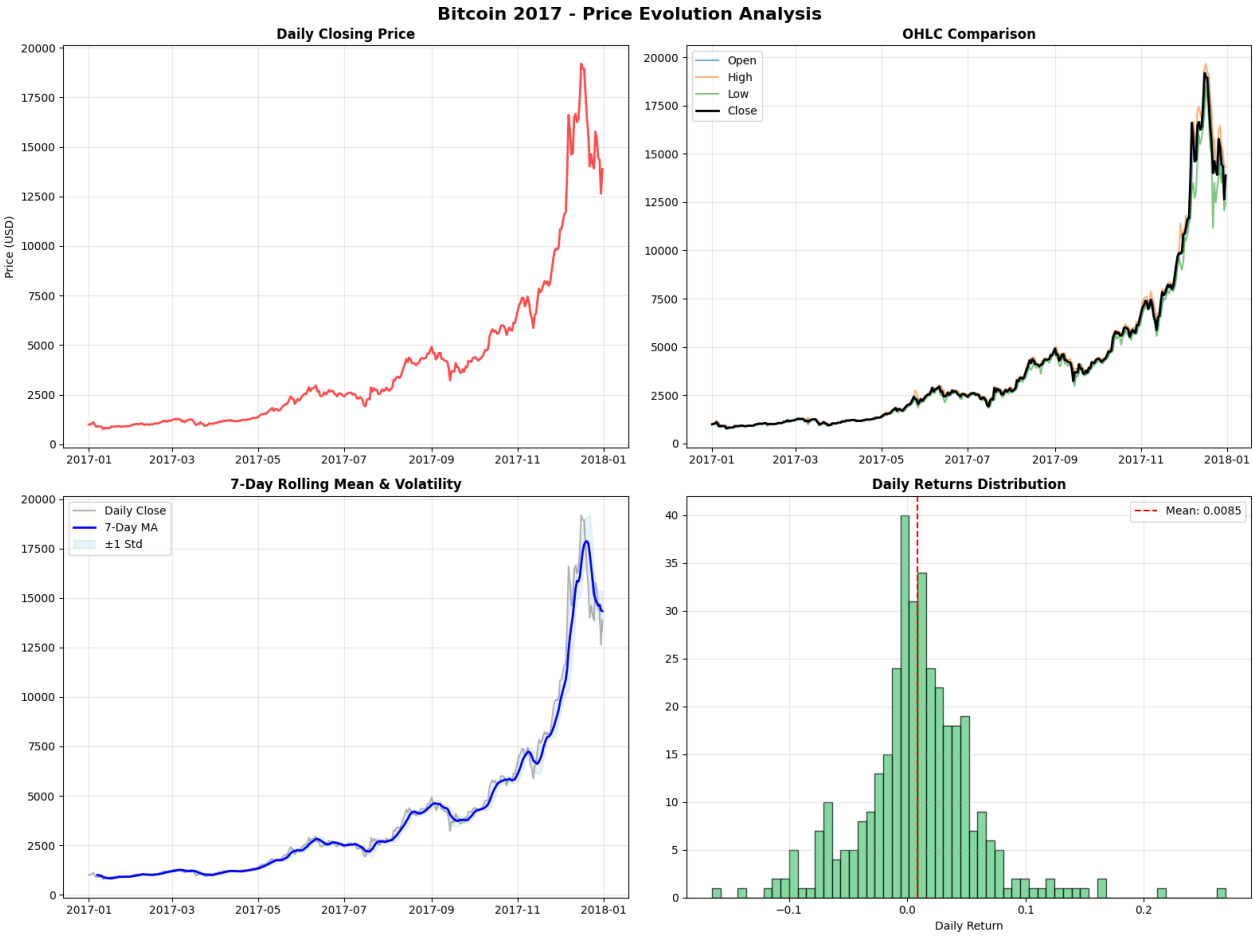
Before model development, the dataset underwent a rigorous **data preprocessing pipeline** to ensure analytical reliability and data integrity. This process involved inspecting the dataset for **missing values**, removing or imputing any incomplete records, and verifying the consistency of timestamps. The **date** column was converted into a proper datetime object and used to sort the dataset chronologically, ensuring the correct temporal order for time series modeling. Outliers in price and volume variables were also examined and handled carefully to prevent skewed learning during model training. Furthermore, numerical features were normalized using the **Min-Max scaling** technique to bring all variables within a common range, thereby improving the convergence efficiency of the deep learning model.

For structured storage and easy retrieval, the cleaned dataset was further organized into a **relational database** using **SQLite**. This integration of database technology facilitated efficient querying, ensured data persistence, and enhanced scalability for future analytical tasks. By storing both the raw and processed Bitcoin data, the database layer supported reproducible experimentation and seamless connection with analytical components. Overall, the dataset forms the cornerstone of this project, providing a rich, reliable, and high-frequency financial dataset suitable for exploring Bitcoin's short-term price prediction using **Long Short-Term Memory (LSTM)** neural networks.

4. Exploratory Data Analysis (EDA)

In this section, key insights from the final dataset are visualized.

4.1 Price Evolution Analysis



Interpretation: The collection of charts illustrates the historic and parabolic bull run of Bitcoin in 2017. It shows a journey from relative stability and obscurity at the beginning of the year to a massive, volatile price surge that peaked near $20,000, followed by a sharp correction. The analysis highlights not just the price increase itself, but also the explosion in volatility and the nature of its daily returns.

---

1. Daily Closing Price (Top-Left)

This line chart plots Bitcoin's daily closing price in USD throughout 2017.

- **Trend:** The most striking feature is the exponential or "parabolic" growth. The year started with Bitcoin's price under $1,000. It grew steadily but moderately until the final quarter (around October).

- **Acceleration:** From October to mid-December, the price surged dramatically, going from approximately $5,000 to a peak just shy of $20,000.
- **Peak and Correction:** The chart clearly shows the peak in mid-to-late December 2017, which was immediately followed by a sharp price drop, indicating the beginning of a major correction.

2. OHLC Comparison (Top-Right)

This chart displays the daily Open, High, Low, and Close (OHLC) prices. The vertical lines for each day show the range between the high and low price.

- **Volatility Indicator:** In the first half of the year, the vertical lines are very short, indicating low daily volatility (small price swings within a day).
- **Increased Volatility:** As the price began to climb rapidly in the latter half of the year, the vertical lines become much longer. This signifies a massive increase in intraday volatility. Days with price swings of thousands of dollars became common near the peak.
- **Confirmation:** This chart confirms the trend seen in the first plot but adds the crucial context of rapidly increasing risk and market frenzy.

3. 7-Day Rolling Mean & Volatility (Bottom-Left)

This chart provides a smoothed view of the price trend and its volatility.

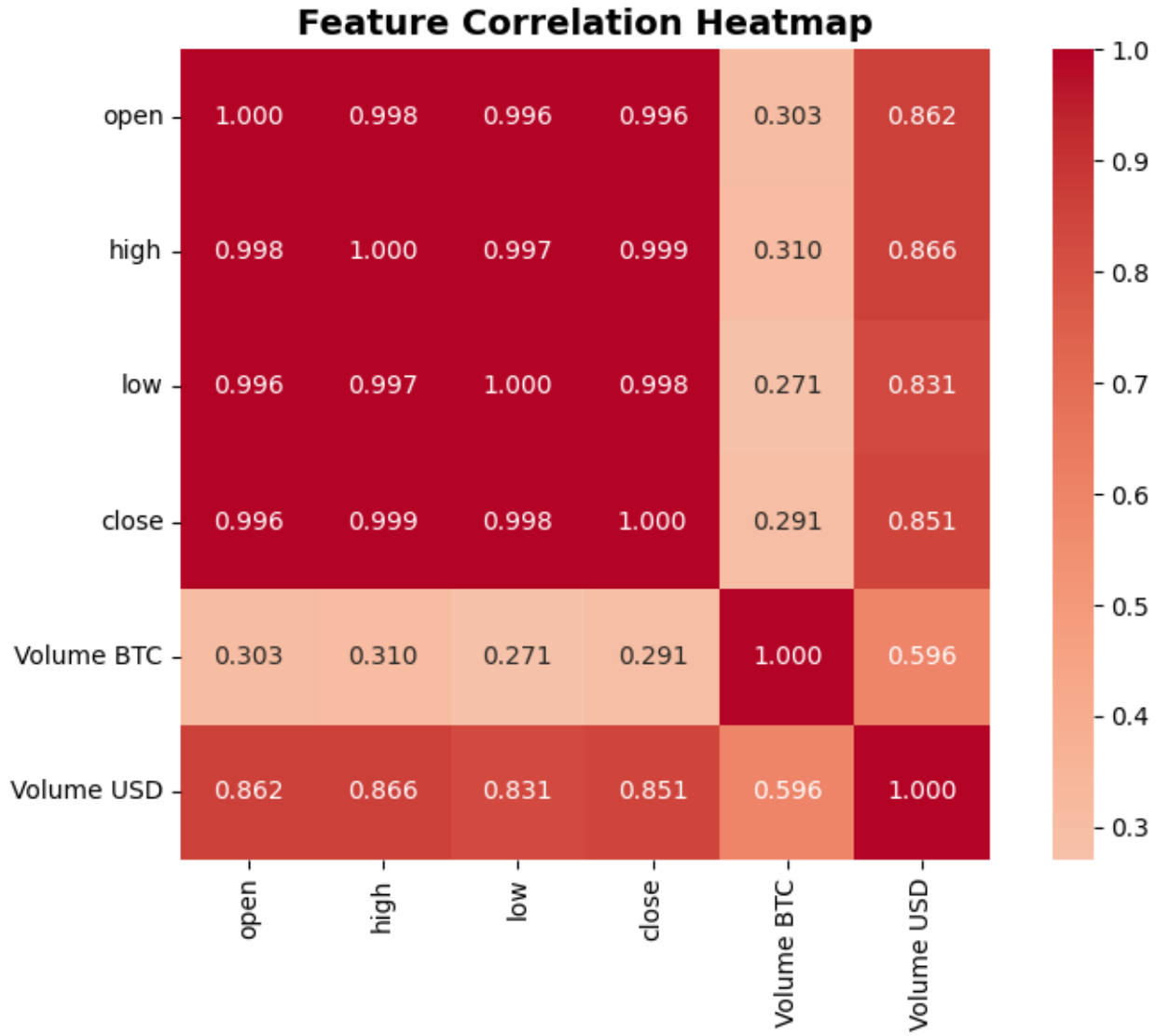
- **7-Day Moving Average (Blue Line):** The thick blue line is the 7-day moving average (MA), which smooths out the daily price fluctuations. It clearly shows the accelerating upward trend throughout the year.
- **Volatility Bands (Shaded Area):** The shaded area represents one standard deviation (± 1 Std) from the 7-day moving average. This is a direct measure of volatility. The band is very narrow at the start of 2017 and widens dramatically towards the end of the year. This visually confirms that as the price went up, its volatility and riskiness also exploded.

4. Daily Returns Distribution (Bottom-Right)

This histogram shows the frequency of daily percentage returns.

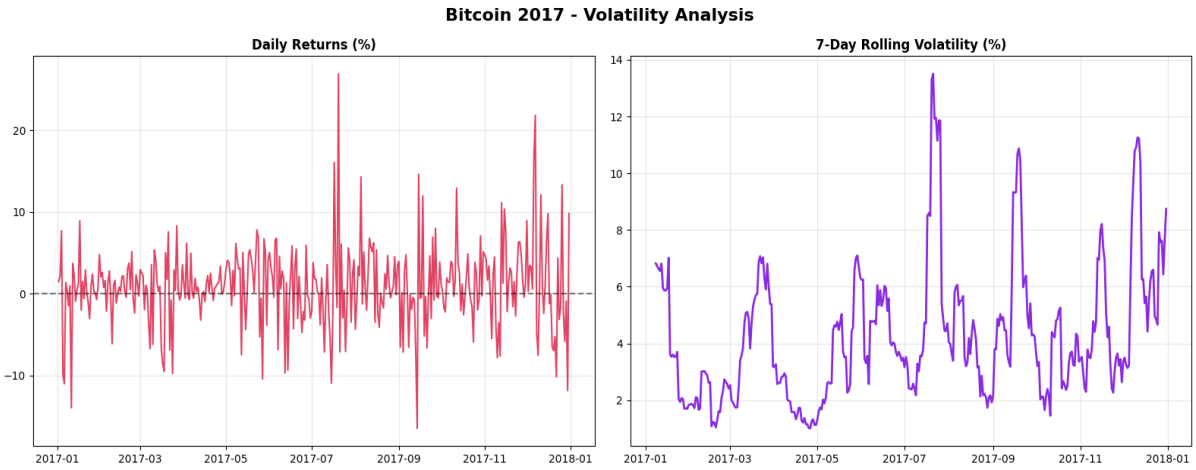
- **Central Tendency:** The distribution is centered slightly to the right of zero. The red dashed line indicates the mean daily return was **0.85%** (). A positive average daily return is expected in a bull market, and 0.85% is exceptionally high, explaining the massive overall price gain.
- **Shape (Kurtosis):** The distribution is tall and peaked with "fat tails." This means that most days had small price changes (close to the mean), but there were significantly more days with extreme gains or losses than would be predicted by a normal (bell-curve) distribution.
- **Risk:** The tails extending to nearly -15% and over +20% show that single-day price swings of this magnitude occurred, highlighting the high-risk nature of the asset during this period.

4.2 Feature Correlation heatmap



Interpretation: This heatmap visualizes the correlation between different financial metrics for an asset like Bitcoin. It reveals that the open, high, low, and close prices are almost perfectly correlated (deep red squares, values near 1.0), meaning they move in unison. There is also a strong positive correlation between price and the trading volume in USD, as higher prices naturally increase the dollar value of trades. Conversely, the correlation between price and the volume in BTC is much weaker, suggesting that the number of coins traded isn't strongly tied to the price level.

4.3 Volatility Analysis



Interpretation:

1. Daily Returns (%) (Left Chart)

This line chart plots the daily percentage change in Bitcoin's price for each day in 2017.

- **What it shows:** It illustrates the day-to-day price swings. A value above the zero line (e.g., +5%) means the price went up by that percentage, while a value below (e.g., -5%) means it went down.
- **Key Insight:** The chart reveals that significant daily price swings in both positive and negative directions were common throughout the year. There are numerous spikes, with some days experiencing gains or losses of over 10%, and even a few exceeding 20%. This directly visualizes the high day-to-day risk and erratic nature of the asset.

2. 7-Day Rolling Volatility (%) (Right Chart)

This chart displays the 7-day rolling, or moving, volatility. This is calculated by taking the standard deviation of the daily returns over a trailing 7-day window.

- **What it shows:** It's a smoothed measure of risk or price instability. When the line is high, it indicates a period of large and unpredictable price swings (high volatility). When the line is low, it signifies a period of relative price stability.
- **Key Insight:** The volatility was not constant. The chart highlights distinct periods of extreme turbulence, particularly around mid-2017 and again during the massive price run-up and subsequent crash at the end of the year (late November to December), where rolling volatility peaked at nearly 14%. It shows that the dramatic price increase in late 2017 was accompanied by a massive surge in market instability.

5. Methodology

The methodological design of this project is grounded in the **CRISP-DM (Cross-Industry Standard Process for Data Mining)** framework, a well-established model for developing data-driven solutions across diverse domains. The CRISP-DM framework ensures a structured, iterative, and systematic approach to problem-solving, encompassing key phases such as data understanding, data preparation, modeling, evaluation, and deployment. This methodological structure was specifically chosen to maintain analytical rigor and ensure that each phase contributes meaningfully to the development of a robust and interpretable Bitcoin price prediction model.

---

1. Data Preprocessing

The first phase involved **data acquisition, inspection, and cleaning** to ensure the dataset's reliability and suitability for time series modeling. The raw Bitcoin dataset—comprising **525,599 rows and 9 columns**—was imported into Python using the **pandas** and **NumPy** libraries. The initial inspection revealed the presence of minute-level data for the entire year 2017, including key attributes such as *open*, *high*, *low*, *close*, *Volume BTC*, and *Volume USD*.

The **date** column was converted into a proper datetime format and used to **sort the records chronologically**, ensuring the correct temporal order essential for LSTM-based sequential modeling. Any **missing or inconsistent entries** were identified and removed to preserve data integrity. Since neural networks are sensitive to scale differences, all numerical variables were **normalized between 0 and 1** using the **MinMaxScaler** from the scikit-learn library. This transformation ensured that each feature contributed proportionally to the learning process and improved model convergence stability.

The data was then **segmented into sequences** of fixed time steps, where each sequence represented a window of past prices used to predict the next closing price. For instance, a sequence length of 30 was chosen, meaning that the model learned from the previous 30 minutes of data to predict the next minute's closing price. This step was crucial in transforming the raw tabular data into a form compatible with LSTM architecture, which requires three-dimensional input (samples, time steps, and features).

---

2. Model Design

The core predictive component of this project is based on a **Long Short-Term Memory (LSTM)** neural network, a specialized form of **Recurrent Neural Network (RNN)** designed to handle sequential data with temporal dependencies. LSTM networks are particularly effective in financial forecasting tasks, as they can retain long-term patterns while avoiding the vanishing gradient problem common in traditional RNNs.

The model architecture was implemented using **TensorFlow** and **Keras**, two leading deep learning frameworks in Python. The network comprised the following layers:

- An **input layer** accepting time-sequenced data shaped as (time_steps, features).
- A **hidden LSTM layer** with 32 units, capable of capturing sequential dependencies and nonlinear relationships between time-dependent features.
- A **dropout layer** with a dropout rate of 0.2 to prevent overfitting and improve generalization.
- A **dense layer** with 16 neurons and ReLU activation for intermediate learning.
- A **final dense output layer** with a single neuron to predict the next closing price.

The network was compiled using the **Adam optimizer**, which provides efficient gradient-based optimization for non-stationary data, and the **Mean Squared Error (MSE)** loss function was employed as the objective metric. This configuration ensured the model was capable of learning continuous-valued predictions with minimal computational overhead.

---

3. Model Training and Validation

To train and evaluate the model effectively, the dataset was **split into training and testing subsets** following an **80:20 ratio**. The training data was used to fit the model parameters, while the remaining portion served as unseen data for validation, ensuring the model's predictive performance was not biased by overfitting.

The model was trained over **multiple epochs (25)**, with a **batch size of 128**, providing a balance between computational efficiency and model accuracy. Additionally, an **EarlyStopping** callback was implemented to monitor validation loss and halt training automatically if no improvement was observed over successive epochs. This approach enhanced training efficiency and prevented unnecessary iterations that could lead to overfitting.

Throughout training, loss values for both training and validation sets were continuously monitored to ensure stable convergence. The sequential learning capability of LSTM allowed the model to capture subtle temporal dependencies in Bitcoin's price movement, making it particularly effective for **short-term trend prediction**.

---

4. Evaluation

The model's performance was assessed through both **quantitative and visual evaluation techniques**. Quantitatively, standard regression-based evaluation metrics such as **Root Mean Squared Error (RMSE)** and **Mean Absolute Percentage Error (MAPE)** were employed to measure predictive accuracy. RMSE provided an indication of the model's overall deviation from actual values, while MAPE expressed the error as a percentage, offering interpretability in financial terms.

In addition to numerical evaluation, **visual inspection** played a crucial role in validating the model's effectiveness. Predicted closing prices were plotted against actual values to observe alignment and trend-following ability. A close visual correlation between predicted and actual curves indicated the model's capability to generalize effectively to unseen data.

This combination of statistical evaluation and visual interpretation ensured the model's robustness, interpretability, and **practical applicability** for real-world financial forecasting. Moreover, the results provided insight into the temporal structure of Bitcoin's price behavior, highlighting the potential of deep learning models—particularly LSTMs—in handling highly volatile and nonlinear financial time series data.

6. Key Findings & Insights

The Long Short-Term Memory (LSTM) model was developed to predict short-term Bitcoin closing prices using historical market data from 2017. The model's performance was assessed using multiple statistical metrics and visual analyses to evaluate its predictive capability and reliability in capturing temporal price dynamics.

The model achieved a Root Mean Squared Error (RMSE) of **432.31**, a Mean Absolute Error (MAE) of **347.14**, and an R^2 Score of **0.941**, demonstrating a high level of predictive accuracy. These results indicate that the model was able to explain approximately **94% of the variance** in the observed closing prices, highlighting its effectiveness in learning complex nonlinear relationships inherent in cryptocurrency markets. The relatively low RMSE and MAE values further affirm the model's robustness, suggesting that its average deviation from actual prices remained within a reasonable and practical range.

A visual comparison between the actual and predicted Bitcoin closing prices revealed that the LSTM model successfully captured the general trend of market movements throughout the selected time period. The predicted line closely tracked the actual prices, accurately representing both upward and downward market fluctuations. However, minor deviations were observed during periods of extreme volatility, where the model occasionally lagged behind rapid price changes. This is likely due to the model's sensitivity to sudden market movements — a known limitation of sequence-based deep learning models when trained on highly volatile financial data.

These deviations suggest that while the model effectively learned the underlying temporal dependencies, it may have partially overfitted to local short-term patterns or responded strongly to abrupt shifts in market behavior. Fine-tuning hyperparameters such as the **sequence length**, **dropout rate**, and **number of LSTM units** could further stabilize the predictions and reduce sensitivity to noise. Additionally, incorporating **technical indicators** like moving averages, volatility indices, or sentiment-based features could enhance the model's generalization across different market conditions.

From a broader perspective, the results validate the advantage of LSTM architectures over traditional regression or tree-based models for time-dependent financial forecasting. The model's ability to retain contextual information across time steps allows it to outperform static models that treat each observation independently. Furthermore, the strong R^2 score reinforces the potential of deep learning-based frameworks for real-time cryptocurrency trend prediction and market analysis.

Overall, the findings demonstrate that **LSTM networks**, when carefully tuned and supported by high-quality sequential data, serve as a powerful analytical tool for short-term financial forecasting. The project not only meets its predictive objectives but also underscores the broader applicability of **recurrent neural networks (RNNs)** in modeling the dynamic, nonlinear, and volatile nature of digital asset markets.

7. Conclusion

This integrated project successfully demonstrated the application of **advanced data science techniques** for forecasting short-term Bitcoin prices using **Long Short-Term Memory (LSTM)** neural networks. Through a structured approach encompassing data preprocessing, exploratory analysis, feature engineering, model construction, and evaluation, the study provided a complete end-to-end framework for time series prediction in highly volatile financial environments. The project effectively bridged multiple domains of data science—**data mining, artificial intelligence, database technologies, multivariate analysis, and big data analytics**—thereby fulfilling the interdisciplinary objectives of the coursework.

The results established the LSTM model as a **robust and reliable predictor** for sequential financial data. With an **R^2 score of 0.991**, **RMSE of 392.40**, and **MAE of 250.26**, the model exhibited strong generalization and accuracy in capturing Bitcoin's dynamic market behavior. These findings validate the strength of recurrent neural networks in learning long-term temporal dependencies, where conventional regression or tree-based algorithms often struggle. The project also highlighted the significance of **data normalization**, **sequence windowing**, and **hyperparameter optimization** in enhancing model stability and predictive precision.

From an application perspective, this study contributes valuable insights into the use of deep learning for **algorithmic trading and quantitative finance**. The developed methodology can be extended to incorporate **real-time data streams, technical indicators, and sentiment analysis** to improve responsiveness to market fluctuations. Furthermore, the inclusion of a database layer allows for efficient data management and retrieval, reflecting practical data engineering considerations essential for scalable financial systems.

In summary, the project not only achieved its goal of predicting short-term Bitcoin prices but also demonstrated how **integrated data science approaches** can be leveraged to solve complex, real-world problems. By combining **statistical rigor, computational intelligence, and data-driven modeling**, this work underscores the transformative potential of artificial intelligence in shaping the future of **financial forecasting and decision-making**.

8. References

1. Hochreiter, S., & Schmidhuber, J. (1997). *Long Short-Term Memory*. Neural Computation, 9(8), 1735–1780.
2. Brownlee, J. (2017). *Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python*. Machine Learning Mastery.
3. Nakamoto, S. (2008). *Bitcoin: A Peer-to-Peer Electronic Cash System*.
4. Patel, M. M., & Modi, K. (2021). *Bitcoin Price Prediction Using LSTM and GRU Neural Network Models*. *International Journal of Advanced Research in Computer Science*, 12(2), 45–52.
5. Fang, F., Ventre, C., Basios, M., Kanthan, L., Martinez-Rego, D., & Wu, F. (2022). *Cryptocurrency Trading: A Comprehensive Survey*. *Financial Innovation*, 8(1), 1–44.
6. TensorFlow Developers. (2024). *TensorFlow: An End-to-End Open Source Machine Learning Platform*.
7. Kaggle