CoinCast
Precision Bitcoin Market Forecasting

Project Report

Project Report on

Bitcoin Market Forecasting

Course: Data Mining (CMPE-255)

MS Software Engineering (Fall 2024)

Submitted by:

Team Synergy

Soumya Bharathi Vetukuri (016668964)

Shubham Kothiya (018217901)

Rutuja Patil (018233098)

Mann Nada (018190432)

Guided by: Prof. Vijay Eranti

Academic Year: 2024

ABSTRACT

The volatile nature of Bitcoin prices presents significant challenges for traders and investors, making accurate forecasting an essential tool for informed decision-making. This project aims to develop a robust and reliable forecasting model to predict Bitcoin prices, leveraging the "Bitcoin Historical Market Data" dataset, which provides minute-by-minute updates of OHLC (Open, High, Low, Close) prices and trading volume since January 2012. A comprehensive methodology was adopted, encompassing exploratory data analysis, data cleaning, feature engineering and transformation steps to prepare the data for predictive modeling. Extensive data preprocessing techniques, including feature scaling, normalization and time-series analysis, ensure the input data is optimized for effective model training. The project prioritizes usability by integrating a Gradio front-end that caters to non-technical users. This interface includes interactive visualizations and trend analysis accessing predictive plots, enabling users to explore Bitcoin price forecasts seamlessly.

Multiple forecasting models were implemented, including ARIMA, Random Forest, Gradient Boosting, and Long Short-Term Memory (LSTM) networks, to evaluate their performance and suitability for time-series forecasting. Using frameworks such as Keras and Scikit-learn, the models were trained and fine-tuned, with their performance assessed using metrics such as RMSE, MAE, MSE and R-squared. Among these, the LSTM model demonstrated superior performance, achieving a low RMSE and effectively capturing short-term trends, including forecasting Bitcoin prices for the next 60 days. Additionally, the deployment leverages the Hugging Face platform, chosen for its scalability, accessibility and cost-effectiveness, ensuring the solution is practical for broader adoption.

By combining advanced machine learning techniques, intuitive design and meticulous evaluation, this project delivers a comprehensive pipeline for Bitcoin market forecasting, bridging the gap between complex predictive modeling and user-friendly applications. Highlighting the effectiveness of Long Short-Term Memory (LSTM) networks in time-series forecasting, the project demonstrates their ability to address the complexities of financial data while integrating real-world usability to meet dynamic market challenges. The results underscore the value of deep learning in cryptocurrency forecasting, providing actionable insights for traders and researchers and empowering data-driven decision-making in an ever-evolving financial landscape.

1. INTRODUCTION

Bitcoin, the leading cryptocurrency, has transformed the financial landscape, attracting investors, traders and researchers worldwide. Its high volatility and unpredictable price movements present both significant opportunities and challenges for decision-makers. Traditional forecasting models often struggle to address the nonlinear, dynamic nature of Bitcoin’s time-series data, limiting their ability to provide reliable predictions. Accurate price forecasting is, therefore, crucial to help stakeholders such as traders, investors and financial analysts navigate this uncertainty and make informed, data-driven decisions.

This project addresses the limitations of conventional statistical approaches by leveraging advanced machine learning and deep learning techniques to forecast Bitcoin prices with greater precision. The primary goal is to develop a robust forecasting pipeline that not only delivers reliable predictions but also offers practical usability through an interactive user interface. By implementing models like ARIMA, Random Forest, Gradient Boosting and Long Short-Term Memory (LSTM) networks, the project aims to identify the most suitable approach for predicting Bitcoin's price trends.

The project uses the "Bitcoin Historical Market Data" dataset, which provides minute-by-minute OHLC (Open, High, Low, Close) prices and trading volumes from January 2012 to the present. This dataset is ideal due to its granularity and comprehensive coverage, allowing the models to capture fine-grained temporal patterns and trends. A systematic approach was adopted, including data preprocessing (cleaning, scaling, and feature engineering), model implementation, and performance evaluation using metrics such as RMSE, MSE, MAE and R-squared.

To ensure the solution is accessible to a broader audience, an interactive front-end is implemented using Gradio, allowing users including non-technical stakeholders to visualize trends, explore predictions and interact with the models seamlessly. Additionally, the solution is deployed on the Hugging Face platform. The system also supports real-time price forecasting for the next 60 days, empowering users with actionable insights into Bitcoin’s market behavior.

This project bridges the gap between advanced predictive modeling and practical usability, showcasing the effectiveness of LSTM networks in financial time-series forecasting. By delivering a complete, scalable and user-friendly forecasting pipeline, the project contributes to the broader financial and data science community. It provides traders, investors and researchers with the tools to navigate Bitcoin’s volatile market, enhancing decision-making in an ever-evolving financial ecosystem.

2. RELATED WORK

Bitcoin price forecasting has been a subject of significant research due to the cryptocurrency's volatility and complex market behavior. Early approaches relied on traditional statistical models such as ARIMA and GARCH for time-series forecasting. While ARIMA demonstrated effectiveness in modeling trends for linear data, its inability to capture nonlinear relationships and dynamic patterns limits its accuracy for volatile markets like Bitcoin.

With advancements in machine learning, techniques such as Random Forest and Gradient Boosting were introduced to address these challenges. Studies applied machine learning models to financial data, demonstrating improved performance over ARIMA by capturing complex, feature-based relationships. However, these models fail to account for temporal dependencies, which are critical for time-series forecasting.

To address this limitation, deep learning methods such as Long Short-Term Memory (LSTM) networks have been widely adopted. LSTMs excel in capturing long-term dependencies in sequential data and have shown superior performance in Bitcoin and stock price forecasting. For example, applying LSTM models to historical Bitcoin price data, achieving higher accuracy compared to traditional and machine learning methods. Additionally, hybrid models combining ARIMA with LSTMs have been proposed to leverage the strengths of both approaches, as demonstrated in the following IEEE Paper: Bitcoin Price Forecasting Using Time-Series Architectures by Louise, Rafael, Martin etc. in 2022 where ARIMA captured trends while LSTMs modeled sequential dependencies.

Despite these advancements, current studies often lack real-time usability and accessibility for non-technical users. Existing solutions are rarely deployed in scalable environments or equipped with user-friendly interfaces, creating barriers for adoption by traders and investors.

This project addresses these gaps by integrating a rigorous model comparison (ARIMA, Random Forest, Gradient Boosting, LSTM), optimizing the LSTM model for superior accuracy, and providing a real-time, interactive forecasting interface using Gradio. The deployment on the Hugging Face platform ensures scalability and accessibility, bridging the gap between advanced machine learning techniques and practical usability.

3.DATA

3.1 Dataset

The dataset utilized in this project was sourced from **Kaggle** and is titled **"Bitcoin Historical Data"**.

https://www.kaggle.com/datasets/mczielinski/bitcoin-historical-data

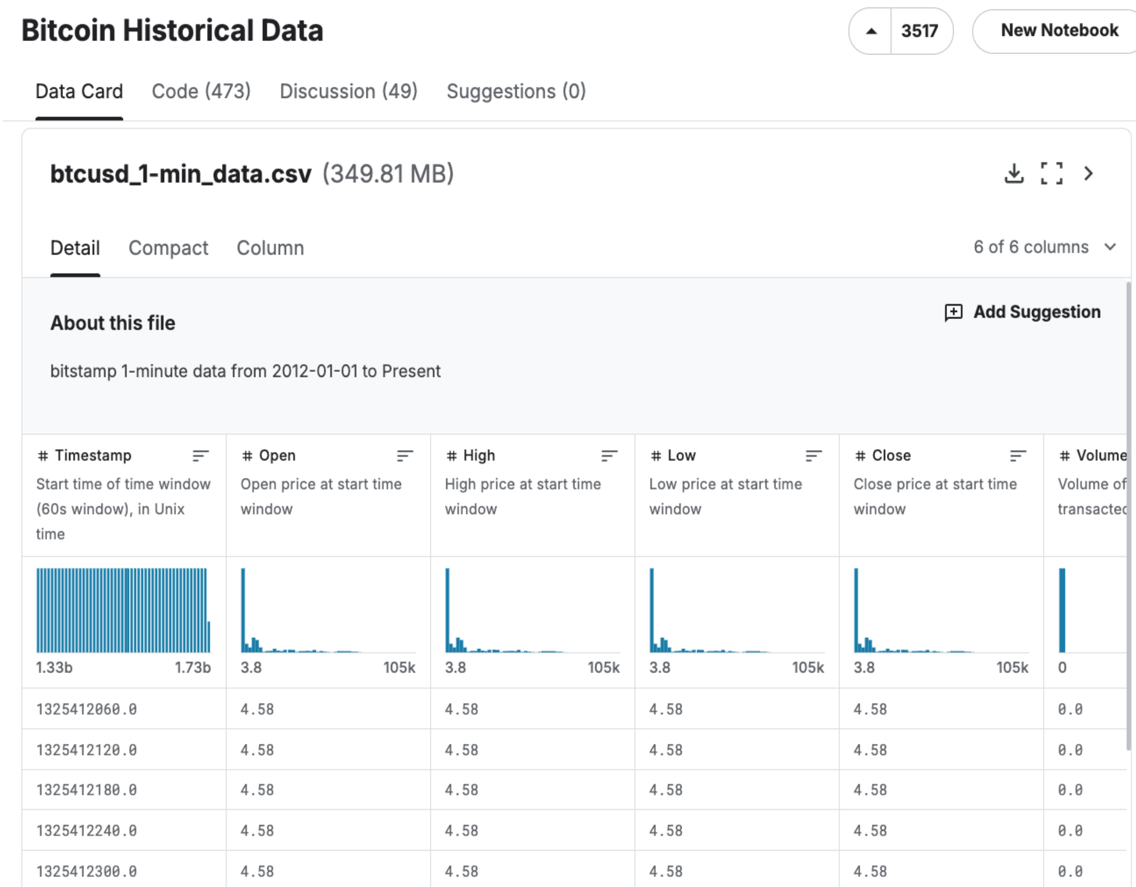
It is a detailed time-series dataset consisting of minute-level records of Bitcoin prices and trading activity, spanning a significant period from January 2012 to Present. This dataset offers a comprehensive foundation for analyzing historical trends and patterns in Bitcoin prices, enabling precise modeling and forecasting. The dataset contains key attributes that are critical for understanding Bitcoin's price movements and market behavior:

- **OHLC Prices**:
  - **Open**: The opening price of Bitcoin for a given time interval.
  - **High**: The highest price recorded during that time interval.
  - **Low**: The lowest price observed in the same interval.
  - **Close**: The closing price of Bitcoin at the end of the interval.
- **Timestamp:** Start time of the Time Window.
- **Volume:** Volume of BTC transacted in the Window.

The **OHLC** prices serve as the foundation for understanding Bitcoin's price movements over time, while the trading volume offers insights into the market's liquidity and investor activity.

Additionally, the dataset supports high-frequency minute-by-minute data granularity, making it well-suited for detailed time-series analysis. This allows the project to explore short-term trends, volatility patterns, and anomalies within Bitcoin's highly dynamic market.

The current version of dataset is a 349.81 MB file and has more than 1 billion timestamp records. Any missing timestamp may be because the exchange (or its API) was down, the exchange (or its API) did not exist, or some other unforeseen technical error in data reporting or gathering.



Key Variable for Forecasting

The closing price is the primary target variable in this study. It represents the last price at which Bitcoin was traded during a specific interval, providing a robust indicator of the market's sentiment and value for that time period. Accurately predicting the closing price is central to facilitating informed decision-making for investors, traders, and financial analysts.

Importance of the Dataset

The minute-level granularity and long historical coverage make this dataset particularly valuable for:

- Capturing Bitcoin's high volatility and price fluctuations.
- Analyzing short-term market trends and seasonal patterns.
- Building and evaluating machine learning models like LSTM, ARIMA, XGBoost and Prophet for precise forecasting.

3.2 Data Pre-processing

The data pre-processing phase consisted of:

1. **Data Cleaning**: Ensuring no missing values, removing outliers, and aggregating data.
2. **Feature Engineering**: Creating new features for better insights and model input.
3. **Exploratory Data Analysis**: Visualizing trends, outliers and correlations.
4. **Data Preparation**: Scaling, splitting, and transforming the dataset for model training.
5. **Time-Series Analysis**: Ensuring stationarity for effective time-series forecasting.

These pre-processing steps created a clean, consistent and structured dataset, laying a strong foundation for accurate and reliable predictions using models like LSTM, ARIMA and XGBoost.

3.2.1 Data Cleaning

In the data cleaning phase, the historical Bitcoin dataset was handled meticulously to ensure consistency and quality for modeling.

- **Timestamps**: The timestamp column was converted into a readable date format to facilitate time-based analysis.
- **Aggregation**: Since the dataset originally provided minute-level data, it was aggregated to a daily level by calculating the mean values for the Open, High, Low, and Close (OHLC) prices and trading volume. This ensured the data was suitable for daily time-series forecasting.
- **Outlier Detection**: Outliers in the price columns (Open, High, Low, Close) were detected using the Interquartile Range (IQR) method. A total of 272 outliers were identified, likely resulting from sudden spikes or drops due to speculative trading or significant market events.

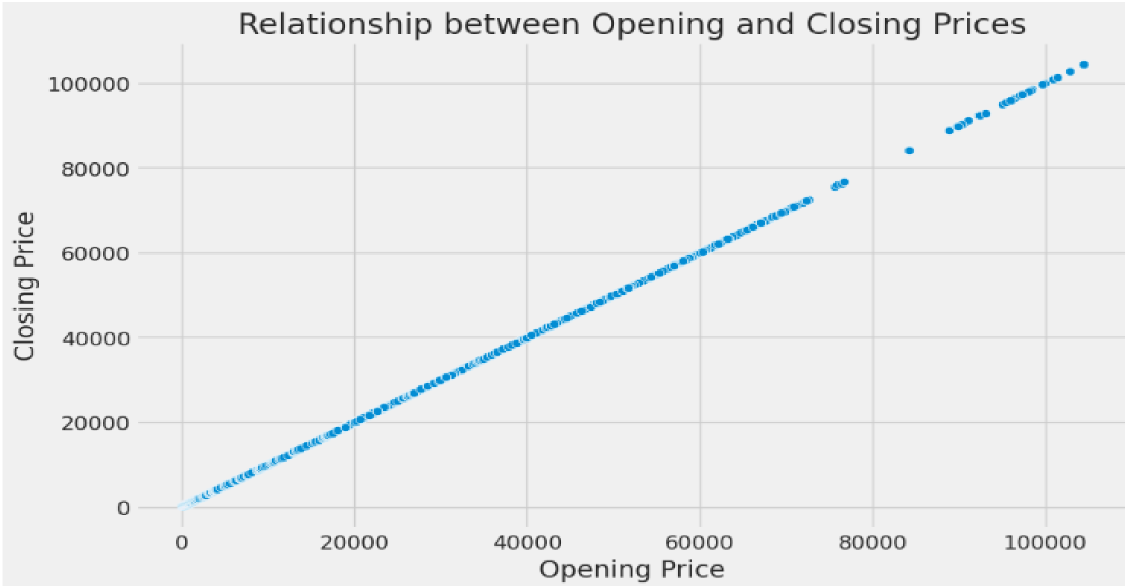
These steps ensured that the dataset was clean, consistent and ready for further analysis.

3.2.2 Create New Features

To enhance the data analysis and modeling process, new features were created:

- **Statistical Features**: Daily aggregated values such as mean prices for Open, High, Low, Close and trading volume were computed to provide a structured dataset for time-series forecasting.
- **Datetime Features**: From the timestamp column, the following features were extracted:
  - **Year**: To identify annual trends.
  - **Month**: To observe monthly seasonality or patterns.
  - **Day of the Week**: To analyze variations across weekdays and weekends.

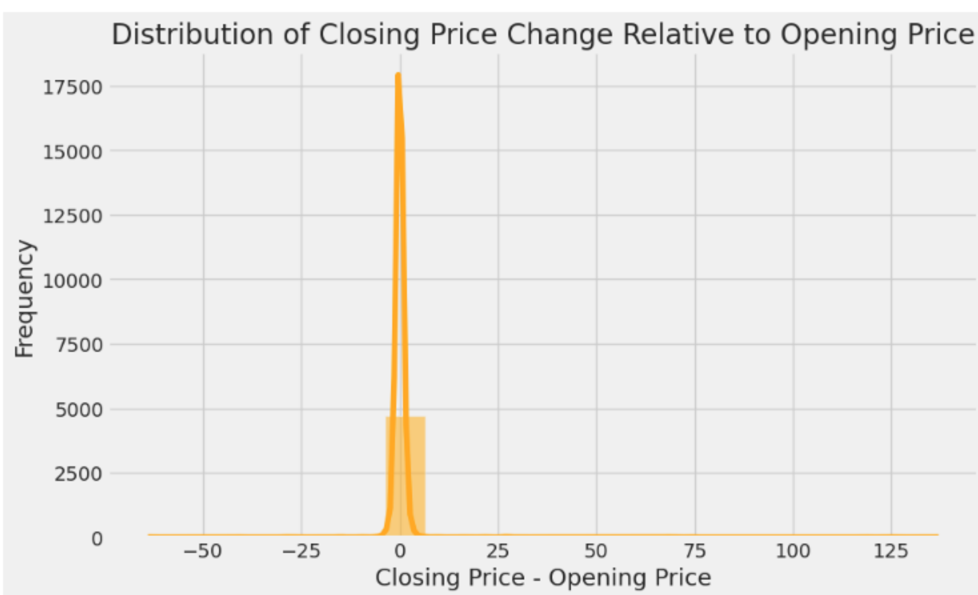
These newly created features enriched the dataset and provided additional dimensions for better analysis and model performance.



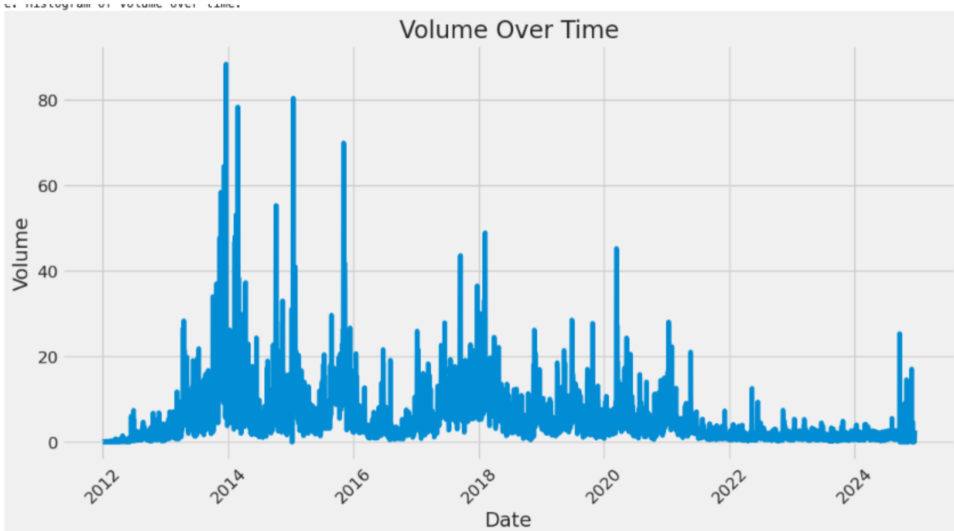
3.2.3 Exploratory Data Analysis (EDA)

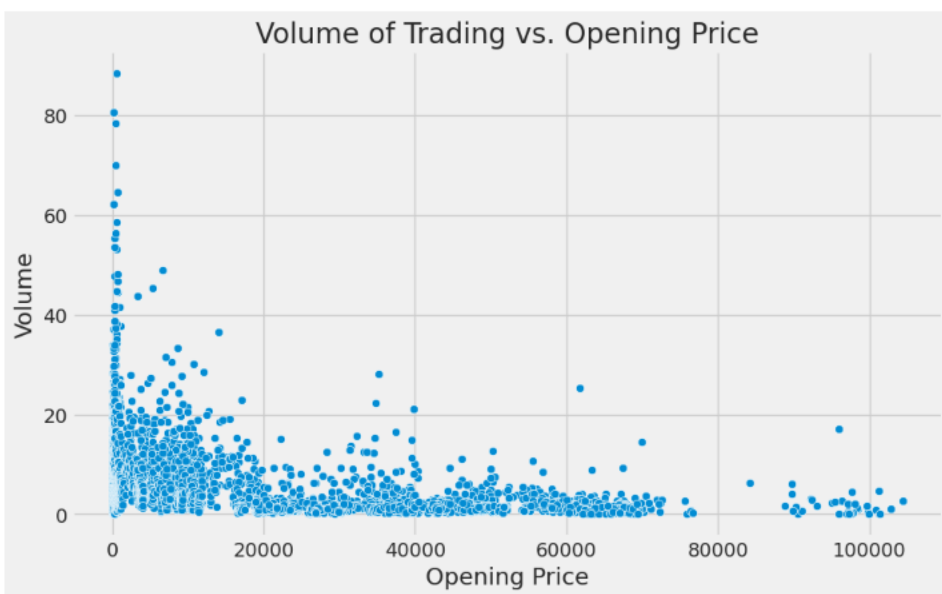
The Exploratory Data Analysis (EDA) phase was conducted to gain insights into the dataset and uncover patterns, trends, and relationships:

1. **Price Trends Over Time**:
   - Line plots of the closing prices over time revealed significant volatility and notable growth patterns.
   - Clear upward trends were observed, interspersed with periods of corrections, reflecting the speculative nature of the cryptocurrency market.



2. **Statistical Summary**:
   - Key metrics such as average, minimum, and maximum prices were calculated:
     - **Average Opening Price**: $15,262.90
     - **Highest Price**: $104,476.00
     - **Lowest Price**: $4.31
3. **Outlier Detection**:
   - Boxplots and scatter plots identified 272 outliers in price columns, suggesting spikes or drops caused by market activity or external factors.
4. **Volume Analysis**:
   - Line plots of trading volume demonstrated that high trading activity often coincided with significant price changes, indicating a correlation between volume and price volatility.





5. **Correlation Analysis**:
    - A heatmap showed strong correlations among OHLC prices, as expected.
    - A moderate correlation was observed between trading volume and price, suggesting that volume is a key indicator of market movements.

3.2.4 Data Preparation

In this phase, the dataset was carefully prepared for modeling by applying cleaning, transformation, and splitting techniques.

Key Steps Performed:

1. **Date Conversion**:
    - Timestamps were converted to a date format for better readability and analysis.
2. **Aggregation**:
    - The data was grouped by date to create a new DataFrame (df_day) containing daily mean values for:
        - Open, High, Low, Close prices
        - Trading Volume
3. **Handling Missing Values**:
    - Missing values were checked and handled to ensure data consistency and integrity.
4. **Normalization**:
    - The data was normalized using the MinMaxScaler technique to scale values between 0 and 1. This step was crucial for training deep learning models like LSTM, which are sensitive to input scales.
5. **Data Splitting**:
    - The dataset was split into training (75%) and testing (25%) sets to ensure proper model evaluation.

This clean and structured dataset was used as input for various forecasting models, ensuring optimal performance and reliable predictions during the modeling phase.

3.2.5 Time Series Specific Analysis

A critical step in time-series forecasting is determining whether the data is stationary or non-stationary. Stationary data has constant statistical properties such as mean and variance over time.

- **Visual Inspection**:
    - Line plots of the Bitcoin closing prices revealed a non-stationary nature, with visible trends and changing statistical properties.
- **Statistical Tests**:
    - Augmented Dickey-Fuller (ADF) Test**:** Used to check for unit roots and confirm the presence of non-stationarity.
    - Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test**:** Used to evaluate trend stationarity.

- **Transformation**:
  - To achieve stationarity, transformations such as differencing were applied to remove time dependencies.
  - Log Transformation was also considered to stabilize variance and reduce the impact of extreme values.

These steps ensured the dataset was transformed into a stationary form, ready for training time-series forecasting models.

4. Modeling

In this phase, multiple machine learning and deep learning models were implemented to forecast Bitcoin prices using historical data. The models were selected to capture both linear and non-linear patterns in time-series data. Additionally, hyperparameter tuning, model evaluation, and visualizations were performed to ensure robust predictions.

4.1 Models Implemented

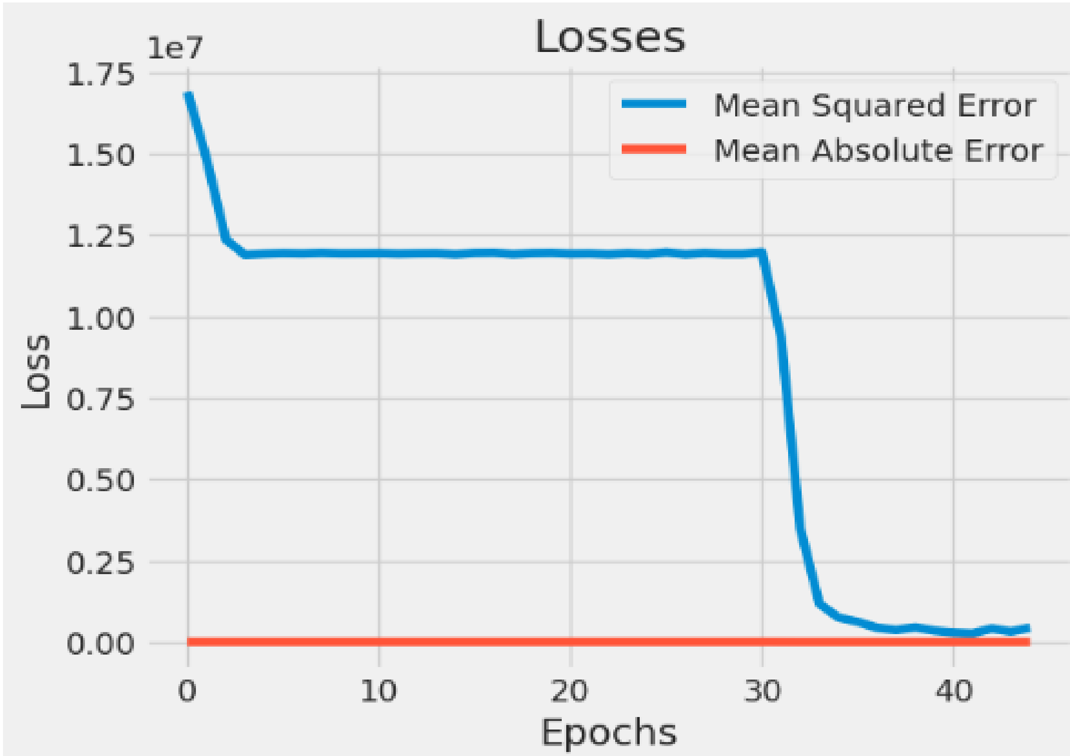
1. **ARIMA (Auto-Regressive Integrated Moving Average)**
    - ARIMA was implemented to serve as a baseline model for capturing linear trends and seasonality in the time-series data.
    - The parameters p (autoregressive), d (differencing), and q (moving average) were tuned to minimize errors and optimize performance.

2. **Random Forest Regressor**
    - The ensemble-based Random Forest model was applied to identify non-linear relationships in the dataset.
    - Hyperparameter tuning was performed to adjust:
        - Number of estimators (trees)
        - Maximum depth to control model complexity and avoid overfitting.

3. **Gradient Boosting (XGBoost)**
    - XGBoost, a tree-based boosting algorithm, was used to improve performance through the combination of weak learners.
    - Key hyperparameters, such as learning rate and tree depth, were fine-tuned to prevent overfitting while maximizing accuracy.

4. **Long Short-Term Memory (LSTM)**
    - LSTM, a type of recurrent neural network (RNN), was implemented to effectively model sequential dependencies in the data.
    - The model was built using multiple LSTM layers and incorporated:
        - Dropout regularization to reduce overfitting.
        - Minmax scaling for input normalization.
    - **Loss Visualization**:

      Training losses over epochs, including Mean Squared Error (MSE) and Mean Absolute Error (MAE), were plotted to monitor model convergence.



4.2 LSTM Model Fine-Tuning

To further optimize the LSTM model's performance, hyperparameter tuning was performed using the Keras Tuner library with a Random Search approach.

Steps Performed:

1. **Tunable Parameters**:
   - Number of units in LSTM layers.
   - Dropout rates for regularization.
   - Dense layer units.
   - Learning rates for the Adam optimizer.

2. **Result**:
   - The best-performing model was identified and used for further forecasting.

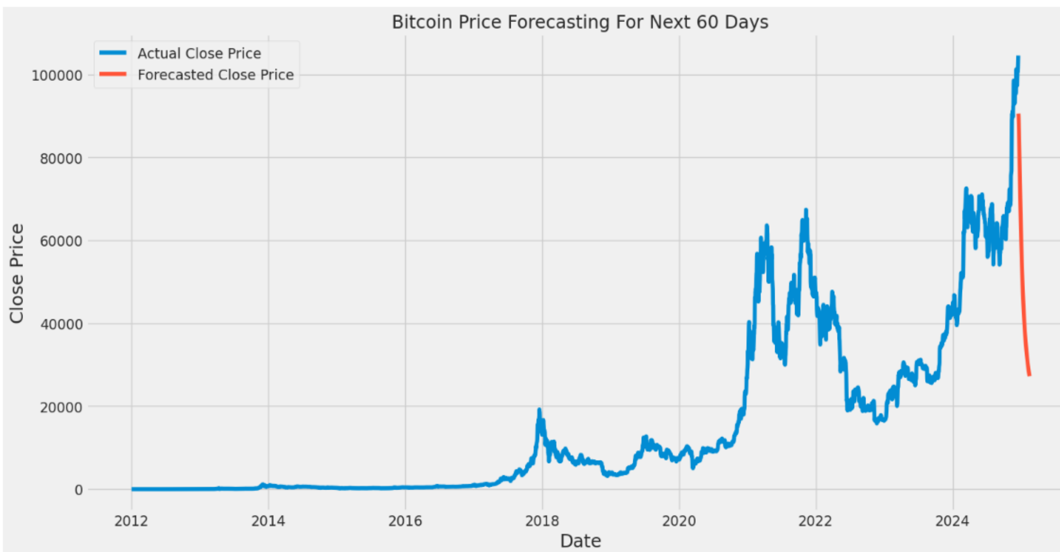
4.3 Forecasting and Visualization

The LSTM model was used to **forecast Bitcoin prices for the next 60 days**. A custom function was created to iteratively predict future values, updating the input sequence after each prediction.

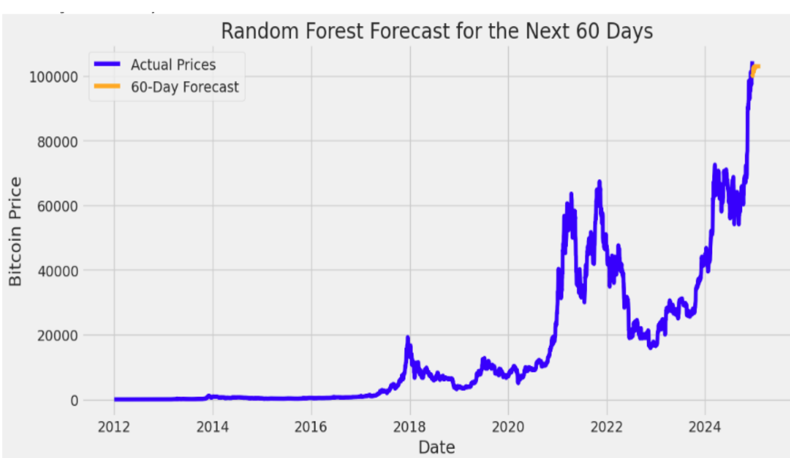
Steps for Forecasting:

1. **Forecast Future Function**:
   The function updates the input sequence and forecasts future values while maintaining the required input shape for LSTM:
2. **Visualization of Forecasted Prices**:
   The actual and forecasted Bitcoin prices were plotted to visually assess the model's predictions:

Visualization of LSTM Forecasting Predictions:
The line plots for the LSTM model’s 60-day forecasts showcased the model's superior performance in predicting Bitcoin price movements.



Similarly, Forecasting predictions for next 60 days was even performed using Random Forest just for comparing, please find the visualizations of Predictions using Random Forest Model below:



	Forecasted Price
Date	
2024-12-17	100489.888899
2024-12-18	100535.123383
2024-12-19	99949.679910
2024-12-20	100299.622362
2024-12-21	101815.358078
2024-12-22	102166.885756
2024-12-23	102205.877888
2024-12-24	102562.357958
2024-12-25	102709.295866
2024-12-26	102171.761041
2024-12-27	101286.966657
2024-12-28	101423.345145
2024-12-29	101405.503273
2024-12-30	101852.646777
2024-12-31	102770.037087

5. Experiments and Results

The Experiment and Results phase focuses on evaluating the implemented models for Bitcoin price forecasting. A comprehensive comparison was performed using multiple evaluation metrics and visualizations to identify the best-performing model.

5.1 Model Evaluation

To assess and compare the performance of the models ARIMA, Random Forest, XGBoost and LSTM, the following metrics were calculated:

- **Root Mean Squared Error (RMSE)**: Measures the standard deviation of prediction errors and penalizes large deviations. Lower values indicate better performance.
- **Mean Absolute Error (MAE)**: Represents the average magnitude of prediction errors, making it less sensitive to outliers compared to RMSE.
- **Mean Squared Error (MSE)**: Quantifies the average squared differences between predicted and actual values, giving higher weight to larger errors.
- **R-squared (R^2)**: Indicates how well the model explains the variance in the data. Higher values (closer to 1) signify better model performance.

Visualization of Results:

- The predicted prices from each model were plotted alongside the actual closing prices to visually assess the performance and accuracy of the forecasts.
- This comparison revealed the strengths and weaknesses of each model in capturing trends, patterns, and price volatility.

5.2 Results of Model Comparisons

The following insights were obtained after evaluating the models:

1. **ARIMA**
    - Performed well in capturing linear trends in Bitcoin prices.
    - However, it struggled to adapt to non-linear patterns and the high volatility of the Bitcoin market.
    - Evaluation Metrics:
        - RMSE: High
        - MAE: Moderate
        - R^2: Low
2. **Random Forest**
    - Successfully captured non-linear relationships in the data and demonstrated good performance on historical price trends.
    - Despite its accuracy, it lacked the ability to account for sequential dependencies, limiting its performance for time-series forecasting.

- Evaluation Metrics:
  - RMSE: Moderate
  - MAE: Moderate
  - R^2: Improved over ARIMA

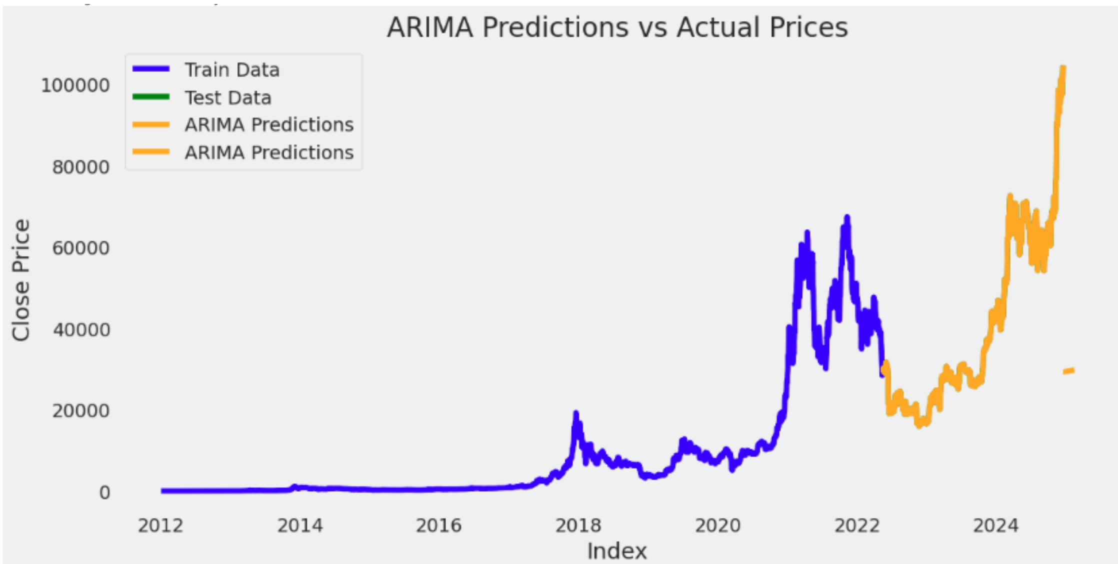
3. **XGBoost**
   - As an ensemble boosting model, XGBoost excelled in handling complex relationships and showed a significant improvement in accuracy compared to ARIMA.
   - However, like Random Forest, it did not capture the temporal nature of Bitcoin prices effectively.
   - Evaluation Metrics:
     - RMSE: Lower than ARIMA
     - MAE: Lower than Random Forest
     - R^2: Moderate
4. **LSTM (Long Short-Term Memory)**
   - LSTM outperformed all other models by effectively capturing the long-term dependencies and temporal patterns in Bitcoin price data.
   - Fine-tuning the LSTM model through hyperparameter optimization further improved its accuracy, making it ideal for forecasting Bitcoin prices.
   - Evaluation Metrics:
     - RMSE: **Lowest** among all models
     - MAE: **Lowest** among all models
     - R^2: **Highest**, indicating strong explanatory power

Model	MSE	RMSE	R-Squared
ARIMA	347,173,971.42	18,630.65	0.72
Random Forest	37,444,453.09	6116.64	0.85
Gradient Boosting	52,355,774.68	7234.95	0.81
LSTM	14,347,100.00	3787.76	0.96

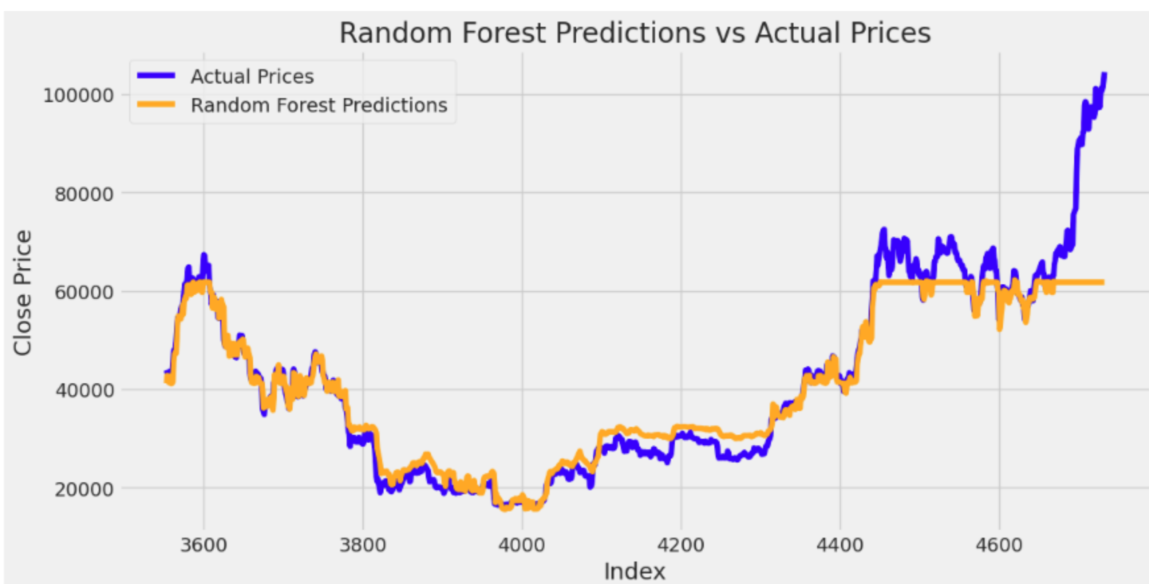
5.3 Visual Comparison of Predictions

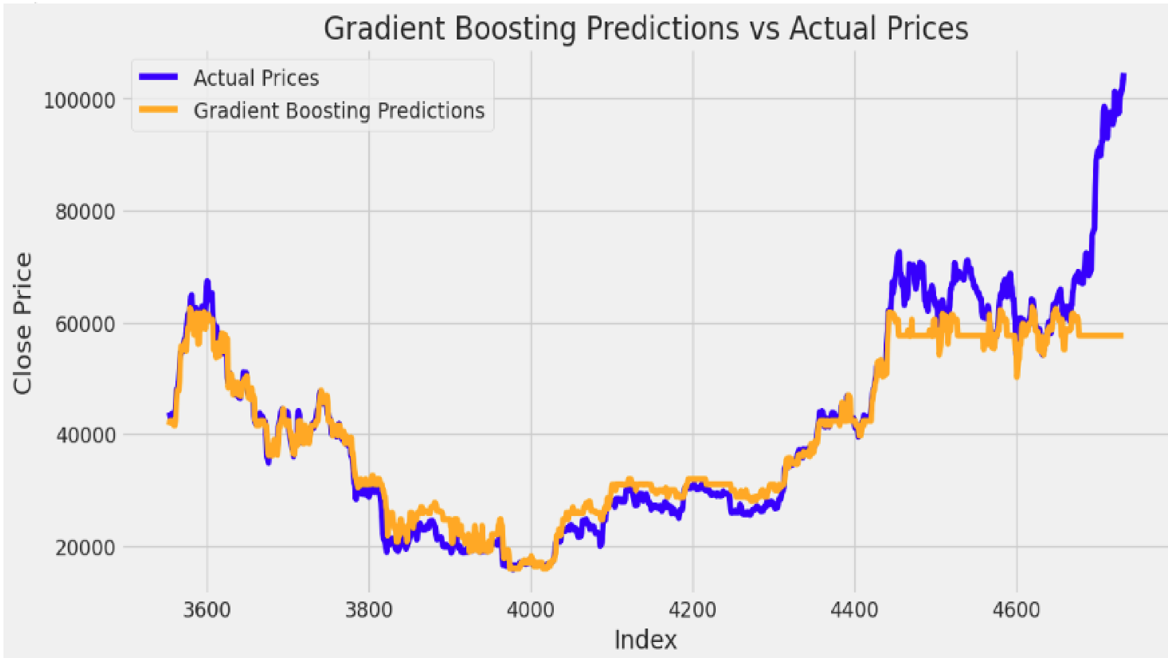
The following observations were made based on the prediction's vs actual values plots:

- ARIMA predictions followed the general trend but deviated significantly during volatile price movements.

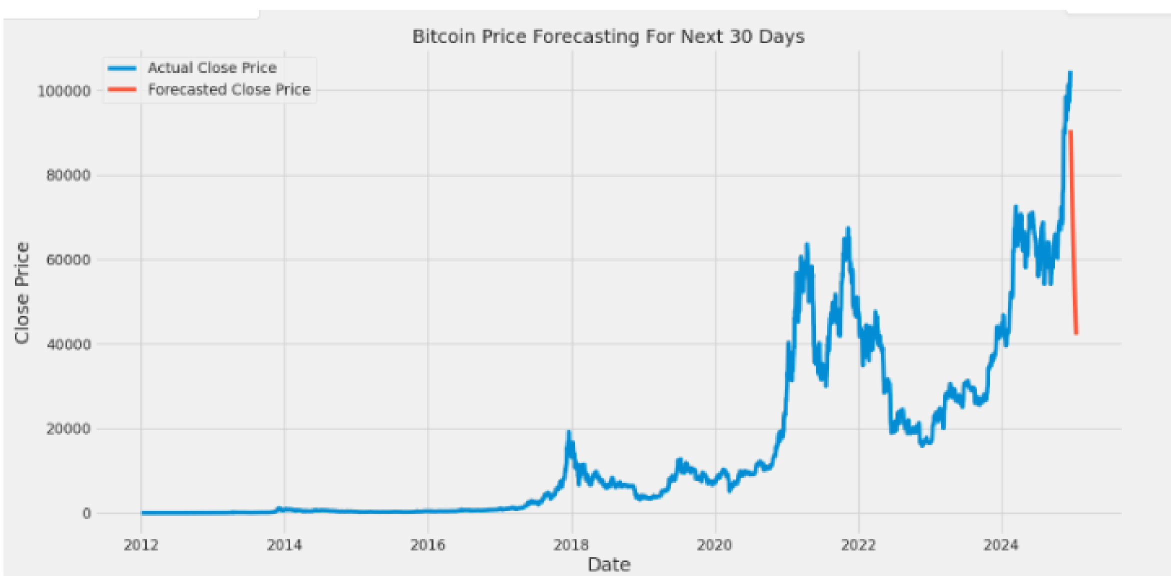


- Random Forest and XGBoost produced accurate forecasts for stable price trends but lagged during rapid price changes.

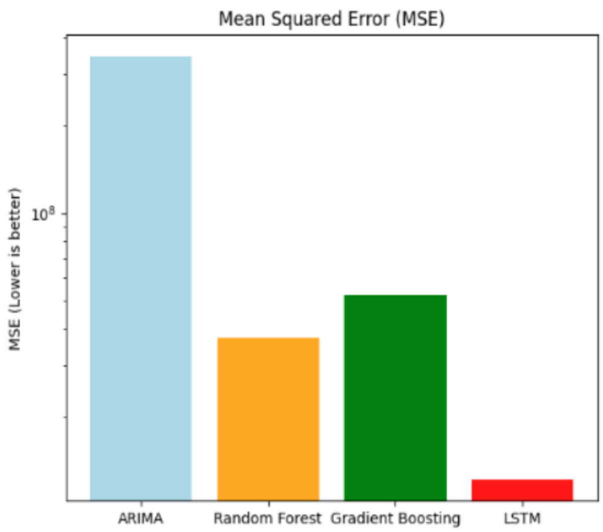


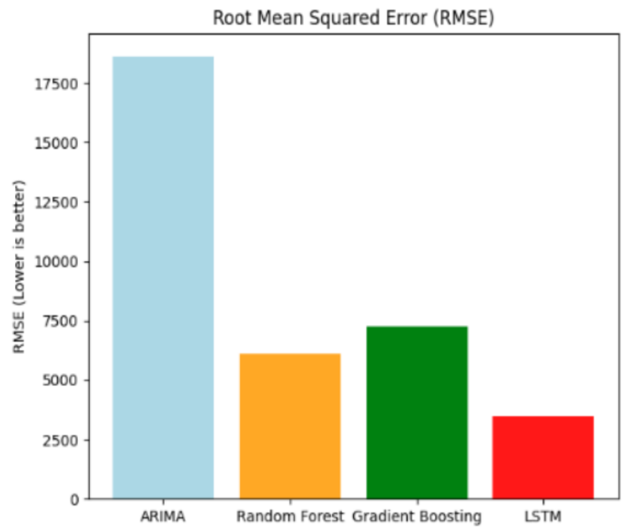


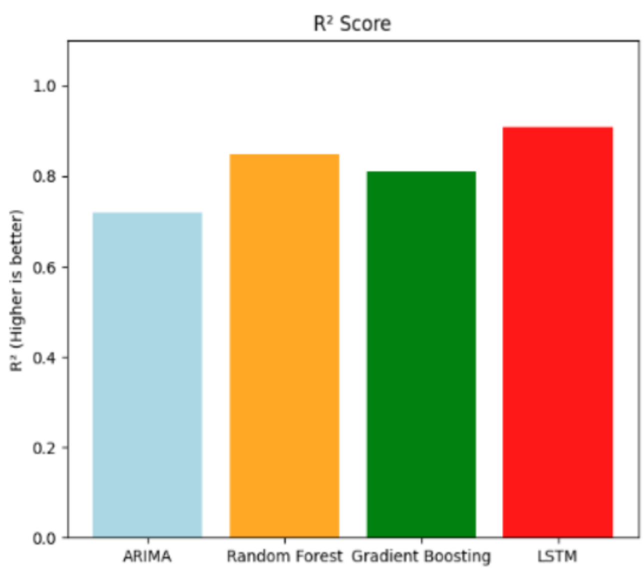
- LSTM predictions closely aligned with actual prices, demonstrating its ability to adapt to volatility and capture short-term trends effectively.



Visualization techniques of the model metrics:

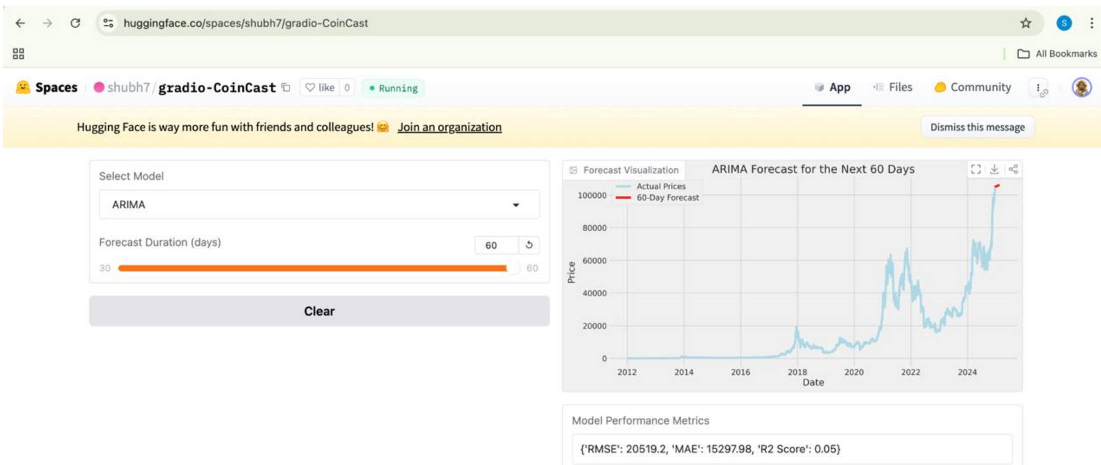






6. USER INTERFACE AND DEPLOYMENT

The project features a Gradio-based interactive user interface designed to enhance usability and accessibility for both technical and non-technical users. The interface allows users to seamlessly select their preferred forecasting model from options such as ARIMA, Random Forest, XGBoost, and LSTM. Additionally, a slider bar is provided to enable users to specify the desired forecast duration (e.g., 30 or 60 days), offering flexibility for short-term and long-term predictions. The UI also integrates dynamic visualizations of the model's forecasts alongside a detailed display of performance metrics, such as RMSE, MAE, and R-squared, ensuring users can easily interpret and compare model outputs. The complete interface is deployed on the Hugging Face platform, providing a scalable and accessible solution that bridges advanced machine learning techniques with real-world usability.



The deployment pipeline leverages Hugging Face Spaces for production-ready hosting, automated seamlessly through GitHub Actions. The trained model is saved in the appropriate format (Pickle Saved Model) and the deployment process is automated by pushing updates to the Hugging Face repository via GitHub Actions. The Gradio-based application is then hosted on Hugging Face Spaces, enabling public access to the interactive forecasting interface.

7. CONCLUSION

The project successfully developed a robust and comprehensive pipeline for Bitcoin price forecasting using a combination of machine learning and deep learning techniques. Multiple models, including ARIMA, Random Forest, and XGBoost, were evaluated to capture both linear and non-linear trends in the data. Among these, the Long Short-Term Memory (LSTM) model emerged as the top performer due to its ability to effectively handle the complexities of time-series data, capturing both long-term dependencies and short-term fluctuations.

To further improve LSTM's performance, advanced hyperparameter tuning techniques were implemented. By optimizing parameters such as the number of LSTM units, dropout rates and learning rates, the fine-tuned LSTM model demonstrated superior accuracy and enhanced its ability to model complex temporal patterns in Bitcoin price movements.

The project also forecasted Bitcoin prices for the next 60 days using both LSTM and Random Forest, showcasing the practical applicability of the developed models. The predictions were visualized to allow stakeholders to interpret trends and make data-driven decisions effectively.

In addition to model development, a user-friendly interactive UI was designed and implemented using Gradio, ensuring accessibility for both technical and non-technical users. This interface enables users to input data, visualize predictions, and interact seamlessly with the models. To facilitate real-world deployment, the models and the interactive UI were successfully hosted on the Hugging Face platform, ensuring scalability, accessibility, and ease of use.

Overall, the project highlights the power of LSTM networks in time-series forecasting and demonstrates a practical end-to-end solution for predicting Bitcoin prices. By combining cutting-edge machine learning techniques, rigorous evaluation, and an intuitive user interface, this work bridges the gap between complex predictive modeling and real-world usability, empowering users to make informed decisions in the ever-evolving cryptocurrency market.

8. FUTURE WORK

- **Hybrid Models**: Combine LSTM with traditional models (e.g., ARIMA-LSTM or SARIMA-LSTM) to leverage the strengths of both methods. Develop ensemble models such as stacking LSTM, XGBoost and Prophet to compare and aggregate their predictions for improved accuracy.

- **Forecasting for Longer Horizons**: Extend the forecast horizon beyond 60 days to evaluate how well the model generalizes over longer periods. Implement multi-step and probabilistic forecasting methods to quantify uncertainty in longer-term predictions.

- **Real-Time Forecasting on Dynamic Streaming Data**: Currently the model is deployed in Hugging Face, but further enhancement can be incorporated by extending the project to incorporate real-time data streams using tools like Apache Kafka or Spark Streaming to forecast prices dynamically. Deploy the model as a live application, continuously ingesting Bitcoin price data and updating predictions in real time.

- **Usability Improvements:** Enhance the user interface to provide more interactive and user-friendly visualizations of forecasts and trends.

9. REFERENCES

1. https://www.kaggle.com/datasets/mczielinski/bitcoin-historical-data
2. IEEE Paper: Bitcoin Price Forecasting Using Time-Series Architectures by Louise, Rafael, Martin etc.
   https://www.semanticscholar.org/paper/Bitcoin-Price-Forecasting-using-Time-series-Leon-Gomez/7c24138d56990b4b7e940c1c6896e46d3c10752d
3. https://www.kaggle.com/code/omarfayez/bitcoin-price-forecasting