Coin Cast

BitCoin Time-Series Forecasting

Presented By,
Team Synergy

Shubham Kothiya
Soumya Bharathi Vetukuri
Rutuja Patil
Mann Nada

Abstract

- This project aims to develop a robust and reliable forecasting model to predict Bitcoin prices.
- Originated from the idea of creating a solution aligned with a trending and impactful topic.
- Aimed to demonstrate various data preprocessing and preparation techniques.
- Develop multiple models, compare their performances and evaluate metrics to showcase the best-performing model.
- Integrate interactive UI and visualizations to present data insights.

Introduction

- Adopted CRISP-DM Methodology.
- Models explored include ARIMA, Random Forest, Gradient Boosting and LSTM.
- Hyperparameter tuning performed to optimize model performance.
- Price Forecasting for the next 60 days.
- Evaluation metrics: MSE, RMSE, MAE and R^2
- Gradio based interactive UI to present data insights.
- Hugging Face Deployment of Model.

Data and DataSet

- Source: [Bitcoin Historical Data]
- Bitcoin exchanges for the time period of Jan 2012 to Present with minute to minute updates of OHLC (Open, High, Low, Close) and Volume in BTC.

  Open:Open Price at Start Time Window

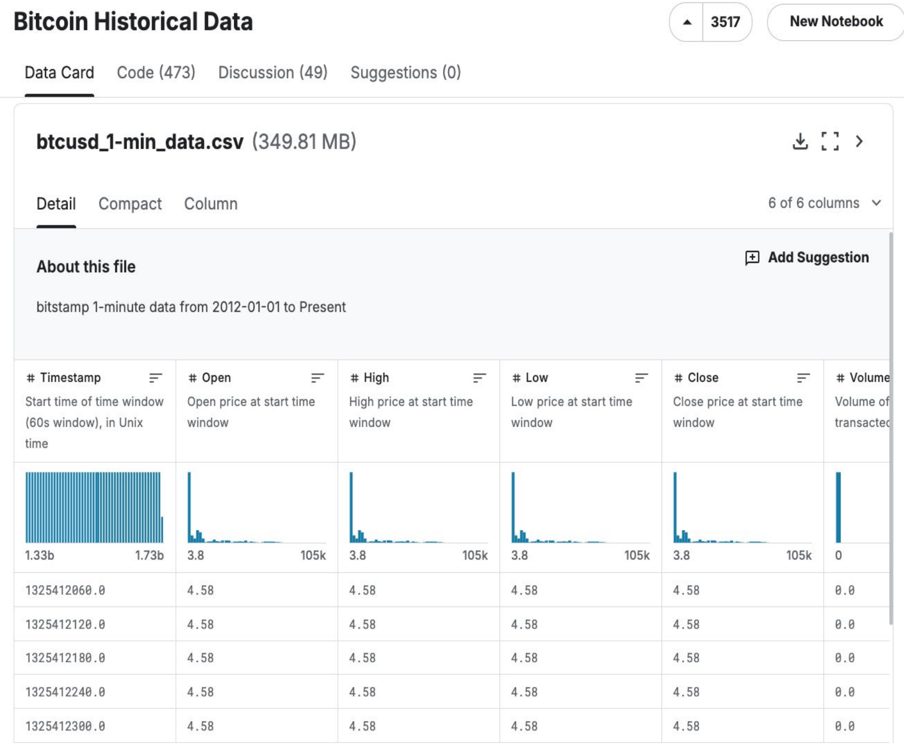
  High: High Price at Start Time Window

  Low: Low Price at Start Time Window

  Close: Closing Price at the end of Interval

  Timestamp: Time Window

  Volume: Volume of BTC transactions.


Bitcoin Historical Data

3517 New Notebook

Data Card Code (473) Discussion (49) Suggestions (0)

btcusd_1-min_data.csv (349.81 MB)

Detail Compact Column 6 of 6 columns

About this file

Add Suggestion

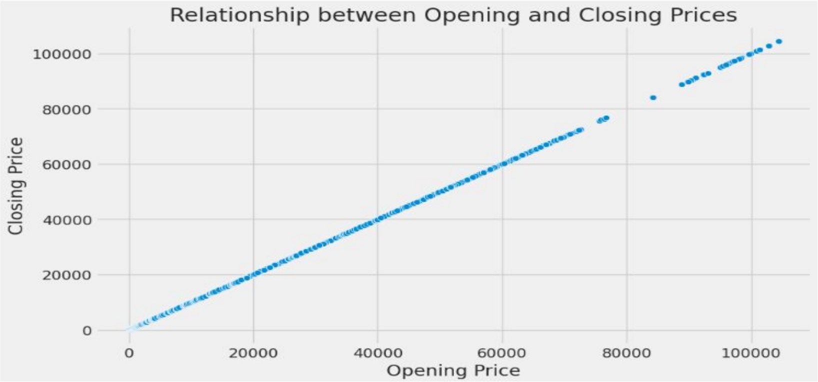
bitstamp 1-minute data from 2012-01-01 to Present

# Timestamp	# Open	# High	# Low	# Close	# Volume
Start time of time window (60s window), in Unix time	Open price at start time window	High price at start time window	Low price at start time window	Close price at start time window	Volume of transacted
1.33b 1.73b	3.8 105k	3.8 105k	3.8 105k	3.8 105k	0
1325412060.0	4.58	4.58	4.58	4.58	0.0
1325412120.0	4.58	4.58	4.58	4.58	0.0
1325412180.0	4.58	4.58	4.58	4.58	0.0
1325412240.0	4.58	4.58	4.58	4.58	0.0
1325412300.0	4.58	4.58	4.58	4.58	0.0

Data Analysis

- Explored the historical Bitcoin dataset to gain insights.
- Converted timestamps into a readable date format.
- Aggregated data to a daily level by calculating the mean values for OHLC prices and Volume.
- Conducted Univariate Analysis to identify trends:
  - Calculated average, minimum and maximum prices.
  - Checked for outliers using Interquartile Range (IQR).
- Visualized the data using line plots to observe price trends over time.

	Timestamp	Open	High	Low	Close	Volume
6729276	1.734307e+09	104510.0	104510.0	104257.0	104268.0	27.923539
6729277	1.734307e+09	104258.0	104298.0	104220.0	104298.0	1.929254
6729278	1.734307e+09	104350.0	104427.0	104350.0	104427.0	1.002842
6729279	1.734307e+09	104476.0	104664.0	104476.0	104536.0	2.709727
6729280	NaN	57854.0	57864.0	57835.0	57835.0	1.353466

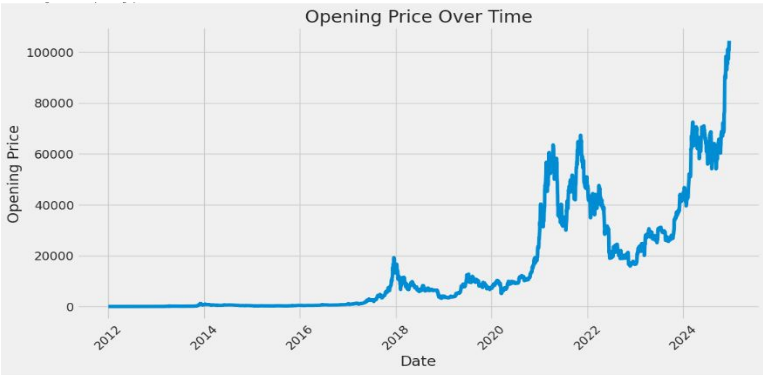


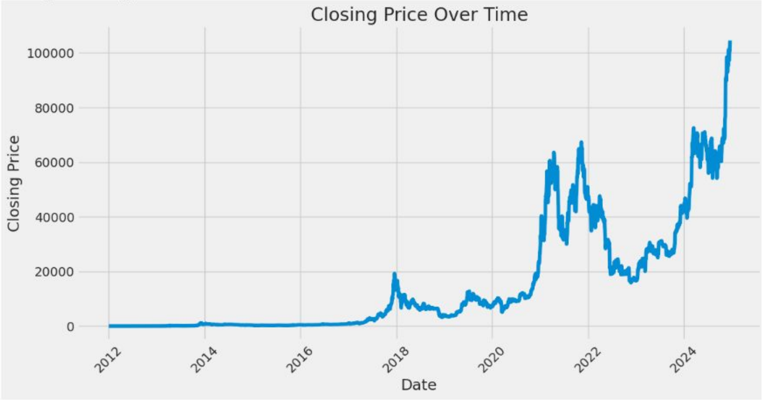
Data Preparation

- Prepared the dataset for modeling by cleaning and transforming the data.
- Converted timestamps to date format and grouped data by date.
- Handled missing values and ensure data consistency.
- Normalized the data using MinMaxScaler for deep learning models like LSTM.
- Split the dataset into training (75%) and testing (25%) sets to ensure proper model evaluation.
- Created a new DataFrame (df_day) containing daily mean values for Open, Close, High, Low prices, and Volume.

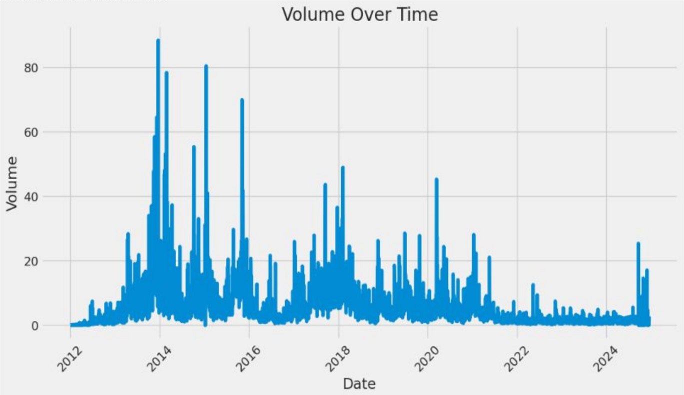
EDA and More..

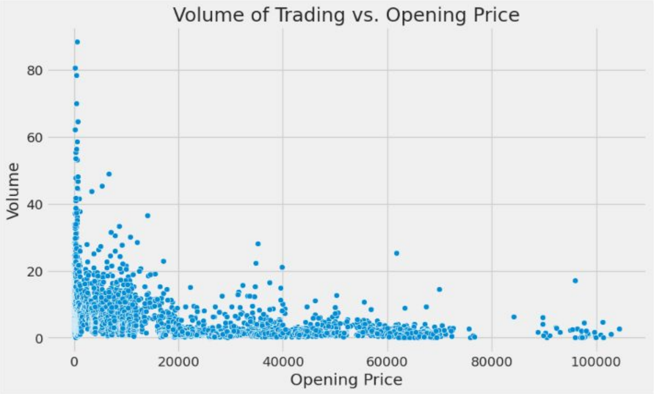
- Critical insights into the historical Bitcoin dataset
- Univariate Analysis and Visualizations.
- Bivariate Analysis and Visualizations.
- Exploratory Data Analysis (EDA) provided helped understand trends, patterns and anomalies in the data.
- Performed, visualized: Price Trends over Time, Statistical Insights, Outlier Detection, Volume Analysis, Correlation Analysis.
- Data Normalization and Preparation for modeling.





EDA and Visualizations





Modeling

ARIMA

- Auto-Regressive Integrated Moving Average
- Best suited for Linear Trends
- Parameters tuned to optimize performance.

Random Forest

- Captures Non-Linear relationships
- Handles Large Datasets Efficiently
- Less Ideal for Large Scale Time Series Forecasting

XGBoost

- Gradient Boosting
- Powerful Ensemble Model that boosts weak learners
- Fine-tuned learning rate and tree parameters to prevent overfitting.

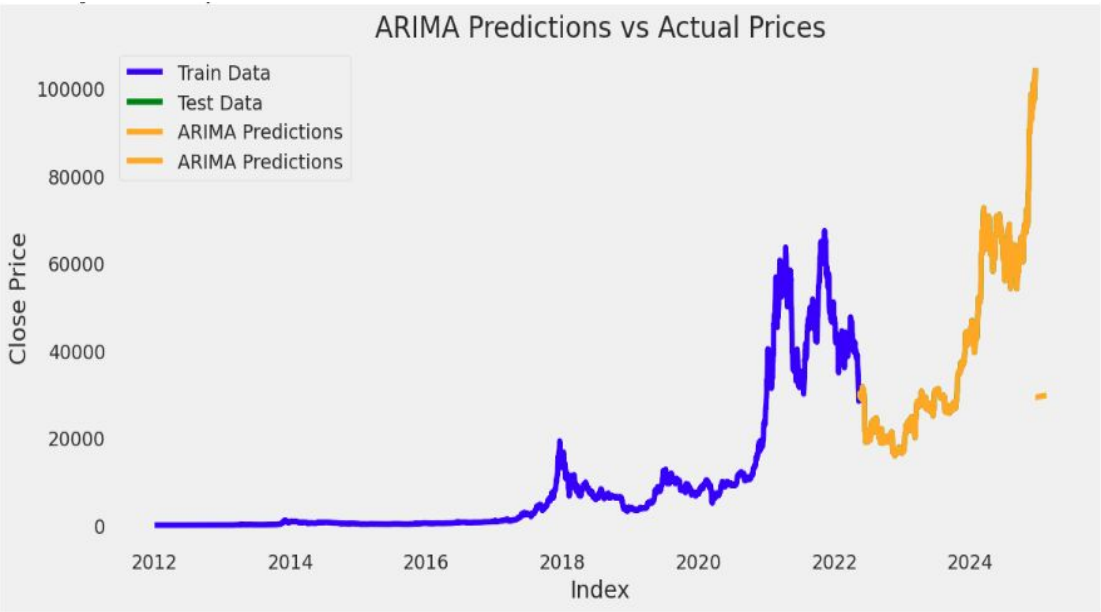
LSTM

- Long Short Term Memory
- Deep Learning Model for Sequential Data.
- Multiple LSTM layers - Improve Performance.

ARIMA - Metrics

ARIMA Model Evaluation Metrics:

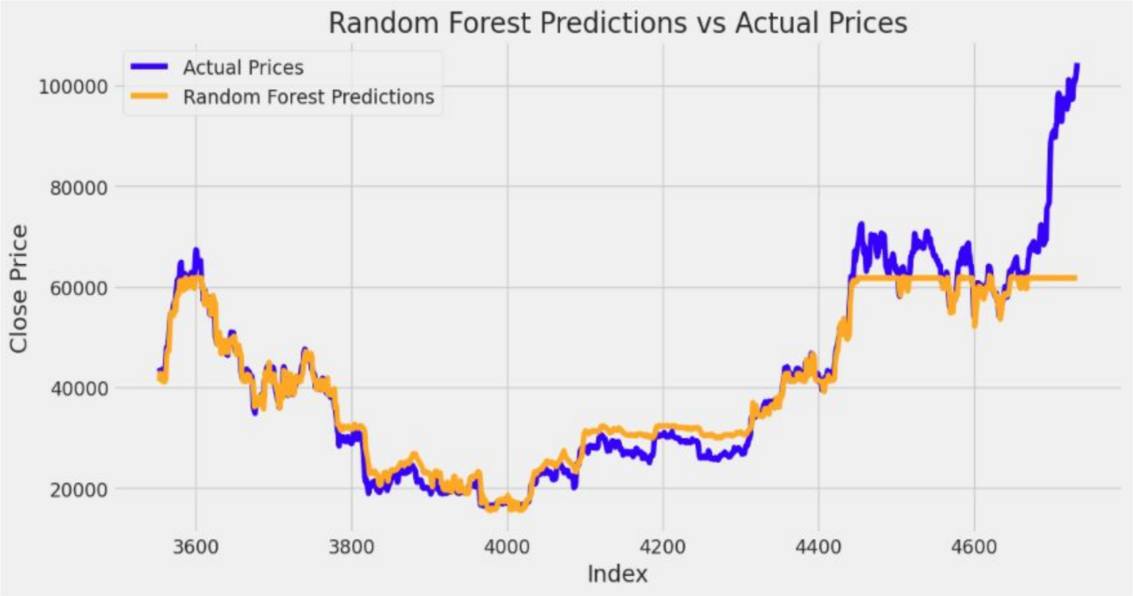
- Root Mean Squared Error (RMSE): 18,630.65
- Mean Absolute Error (MAE): 15297.98
- Mean Squared Error (MSE): 347,173,971.42
- R-squared (R2): 0.72



RANDOM FOREST - Metrics

Random Forest Evaluation Metrics:

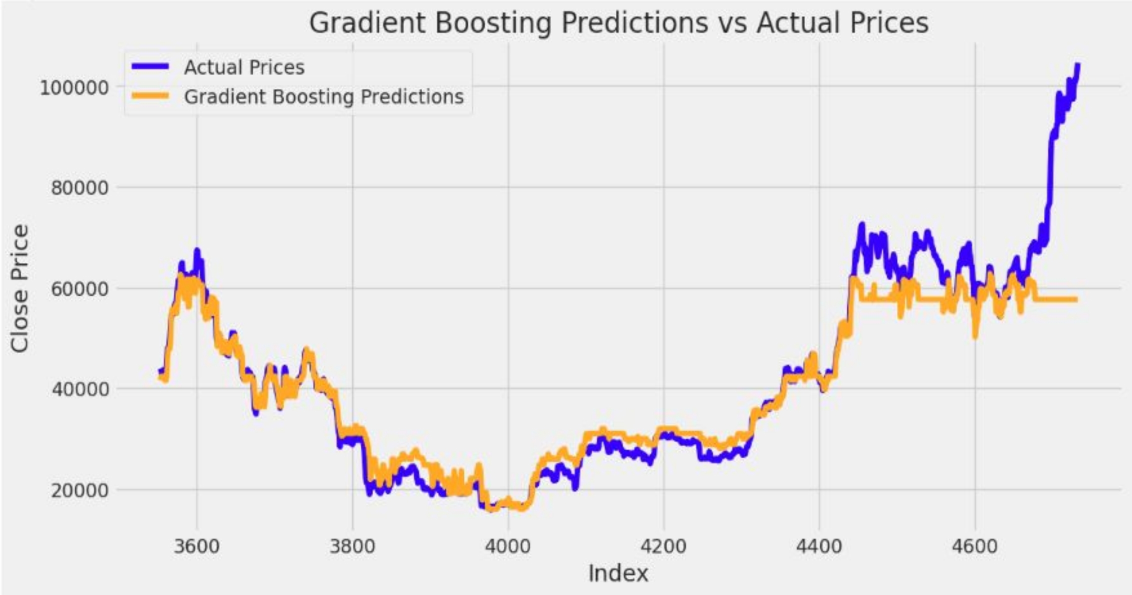
- Root Mean Squared Error (RMSE): 6116.64
- Mean Absolute Error (MAE): 3295.77
- Mean Squared Error (MSE): 37,444,453.09
- R-squared (R2): 0.85



GRADIENT BOOSTING - Metrics

Gradient Boosting Evaluation Metrics:

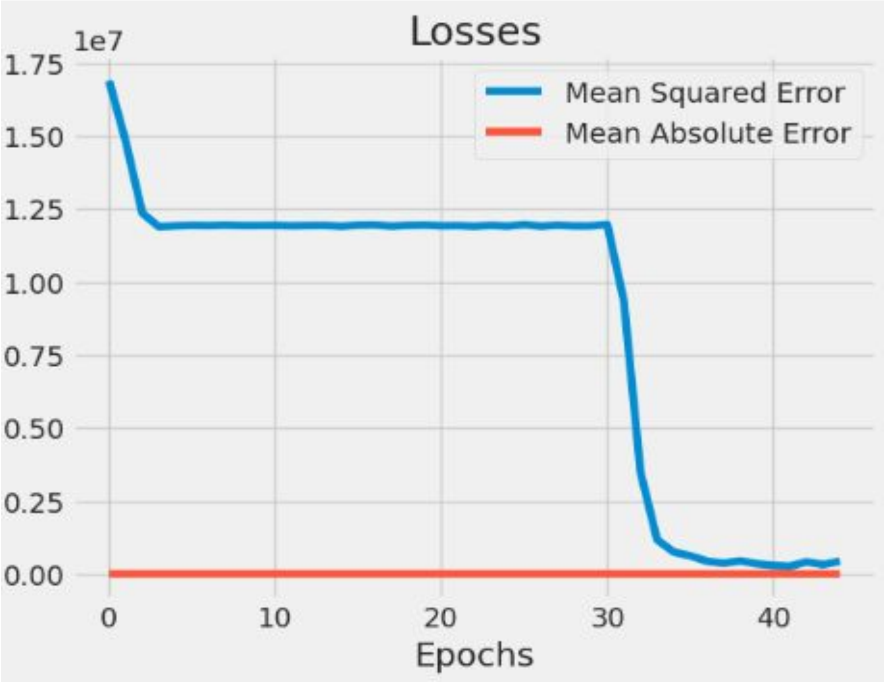
- Root Mean Squared Error (RMSE): 7234.95
- Mean Absolute Error (MAE): 3896.71
- Mean Squared Error (MSE): 52,355,774.68
- R-squared (R2): 0.81



LSTM - Metrics

LSTM Evaluation Metrics:

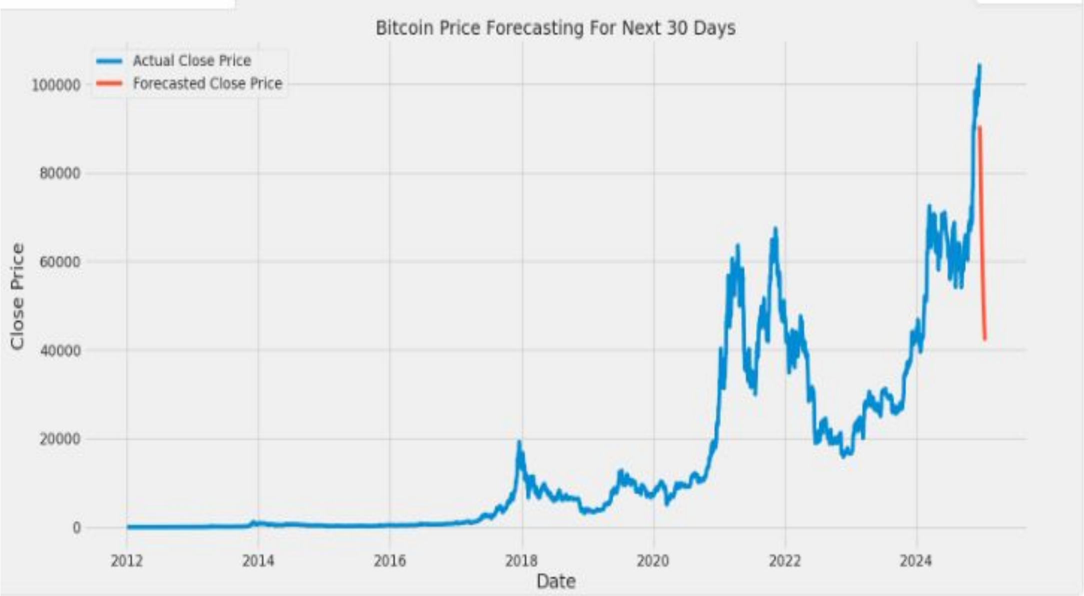
- Root Mean Squared Error (RMSE): 1101674407.65
- Mean Absolute Error (MAE): 1101674407.44
- Mean Squared Error (MSE): 1.2136865004652488e+18
- R-squared (R2): -2665919967.17



LSTM Model Tuning

LSTM Evaluation Metrics:

- Root Mean Squared Error (RMSE): 3787.76
- Mean Absolute Error (MAE): 2617.98
- Mean Squared Error (MSE): 14347100.75
- R-squared (R2): 0.96



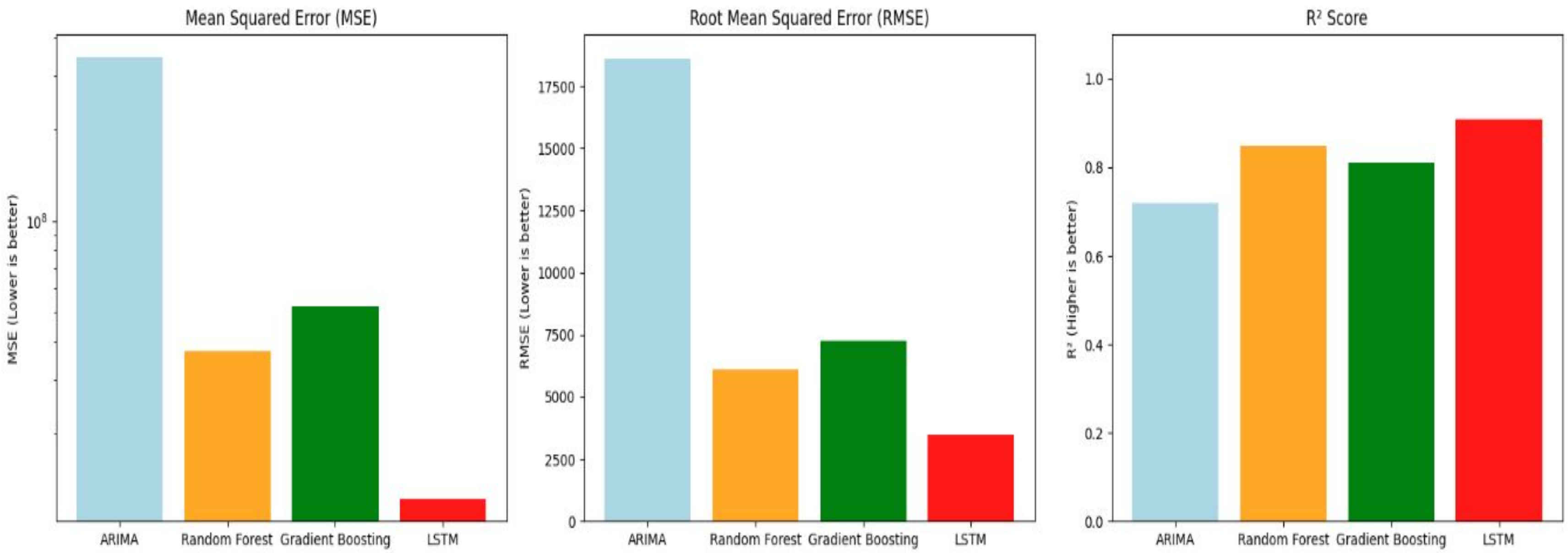
Model Performance Comparison

Evaluation Metrics:

- **Mean Squared Error (MSE)**: Measures the average squared difference between actual and predicted prices.
- **Root Mean Squared Error (RMSE)**: Provides the error in the same units as the target variable.
- **Mean Absolute Error (MAE)**: Measures the average magnitude of prediction errors.
- **R-squared (R^2)**: Represents the proportion of variance explained by the model.

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

Model Performance Visualization

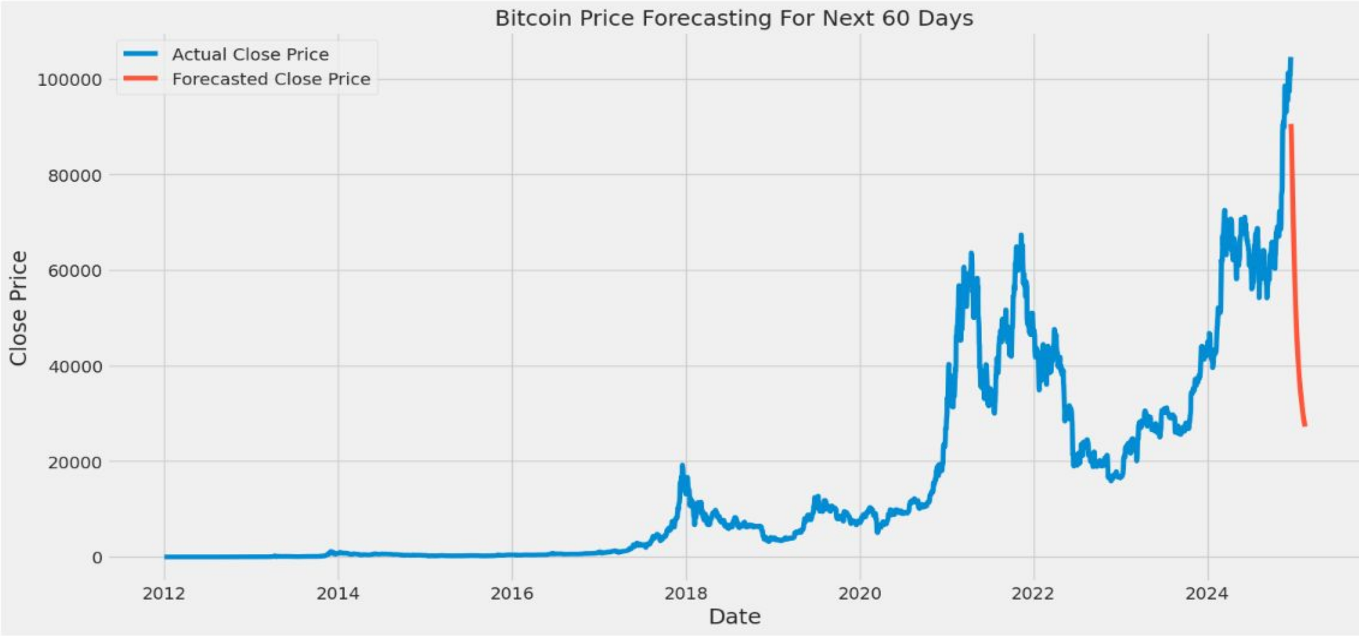


Random Forest Forecasting

Sample Forecast Prices

Date	Forecasted Price
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

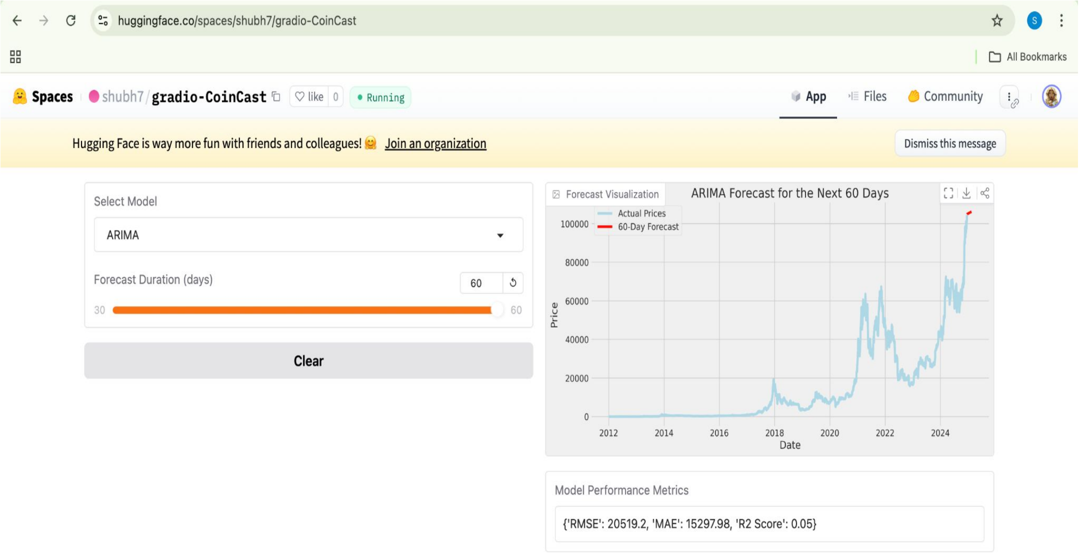
LSTM Forecasting



User Interface

UI Components:

- Gradio Based Interactive User Interface.
- Option to select the model (Eg: ARIMA, Random Forest, XGBoost, LSTM)
- Bar to select the duration of desired forecast window (Eg:30)
- Model's Forecast Visualization and Metrics display.
- Deployed in Hugging Face.



Deployment

Deployment Pipeline

The deployment pipeline is implemented using Hugging Face space for production-ready hosting, automated via GitHub Actions.

1. The trained model is saved in the required format (Pickle SavedModel).
2. GitHub Actions Pipeline automates the deployment process by pushing updates to the Hugging Face repository.
3. The Gradio-based application is hosted on Hugging Face Spaces for public access.

SCAN ME

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.

Conclusion

- Evaluated multiple machine learning models, including ARIMA, Random Forest, XGBoost and others, with LSTM identified as the top performer.
- Despite the success of the initial LSTM model, opted to implement model tuning techniques to further optimize its performance. By adjusting hyperparameters, the model demonstrated improved accuracy and enhanced the model’s ability to capture complex temporal patterns in Bitcoin price movements.
- Forecasted the Bitcoin prices for the Upcoming 60 days using Random Forest and LSTM.
- Designed and implemented an Interactive UI.
- Deployed the models and the gradio based user interface in Hugging Face.

THANK YOU!

Team
Synergy