Bitcoin Predictive Modeling

Your Data Science Team & Presenters

Noah Welgoss

Daniel Abrego

Jason Crespo

Why Bitcoin?

Historical Growth and Volatility

Since 2009, Bitcoin has seen an extraordinary price increase.

For instance, the price rose from under $1 in 2010 to an all-time high of over $65,000 in 2021, equating to a growth of over 6,500,000% in about a decade.

Interconnectedness with Other Markets

Bitcoin has increasingly shown a relationship with other financial assets, such as the stock market and gold.

Understanding these correlations can provide insights into Bitcoin’s behavior in response to broader economic trends.

Involvement in Emerging Financial Trends

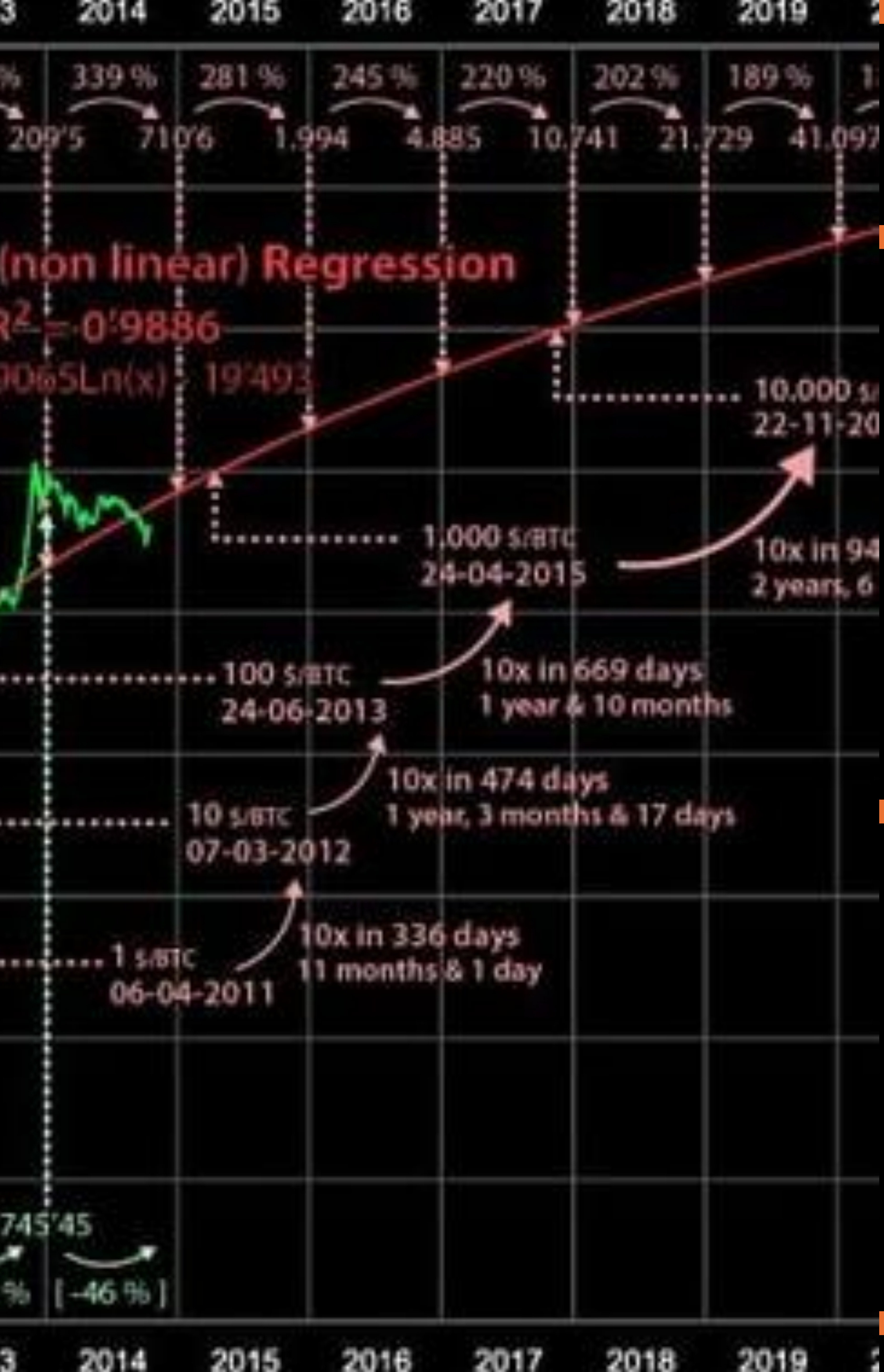
Studying Bitcoin’s evolving role helps predict future trends in both the cryptocurrency space and traditional financial markets.

Using predictive models, researchers can explore the impact of adoption rates, technological innovations, and policy changes on Bitcoin’s price

Availability of Data and Technology

Bitcoin operates on a blockchain, which is a publicly accessible ledger.

This transparency provides researchers with an extensive dataset of historical prices, transaction volume, hash rate, and more.



Objective:

To analyze and test the strength of the relationship between Bitcoin prices and selected variables, using a data science model built with scikit-learn.

Key Variables:

- **Nasdaq Index**: Represents the broader stock market and its correlation with Bitcoin as a digital asset.

- **M2 Money Supply**: Tracks the amount of liquid assets in the economy and its potential influence on Bitcoin as an alternative asset.

- **Ethereum (ETH)**: Provides insight into the interaction between Bitcoin and other major cryptocurrencies.

- Bitcoin Historical Pricing as the key variable

Statistical 7 Day Moving Averages - As a comparative value to the next day close price.

Methodology:

- **Data Segmentation *by Halving Dates***: We broke down Bitcoin's pricing data into distinct periods based on Bitcoin's halving events, aiming to capture shifts in behavior over time.

- **Modeling Approach**:
Using correlation and logistic regression, LSTM and Random Forrest we tested and analyzed how each variable correlated with Bitcoin and explored the evolution of these relationships over time and the possibility of confirming relationships for predicative pricing purposes.

Data Cleanup Efforts

Data Acquisition:

- The team identified and sourced four datasets from reputable platforms, including Kaggle, Bloomberg, and Coinbase. These platforms provided high-quality, relevant data necessary for our analysis of market trends related to Bitcoin, Ethereum, Nasdaq, and M2 Money Supply.

Analysis Tools Utilized:

- To effectively analyze the collected data for our predictive model, the team employed a range of tools and techniques. We leveraged Machine Learning (ML) methodologies, utilizing Scikit-learn for model building and evaluation. Time series analysis was conducted using Long Short-Term Memory (LSTM) networks to capture temporal dependencies in the data. Additionally, we used Python libraries such as Pandas for data manipulation and Matplotlib for data visualization. Our data was stored and queried using an SQL database, and we applied Random Forest algorithms for regression testing, enabling us to gain deeper insights into the relationships within the data.

Data Sources:

- Kaggle
- Bloomberg
- Coinbase

Files Included:

- Ethereum
- Bitcoin
- Nasdaq
- M2 Money Supply

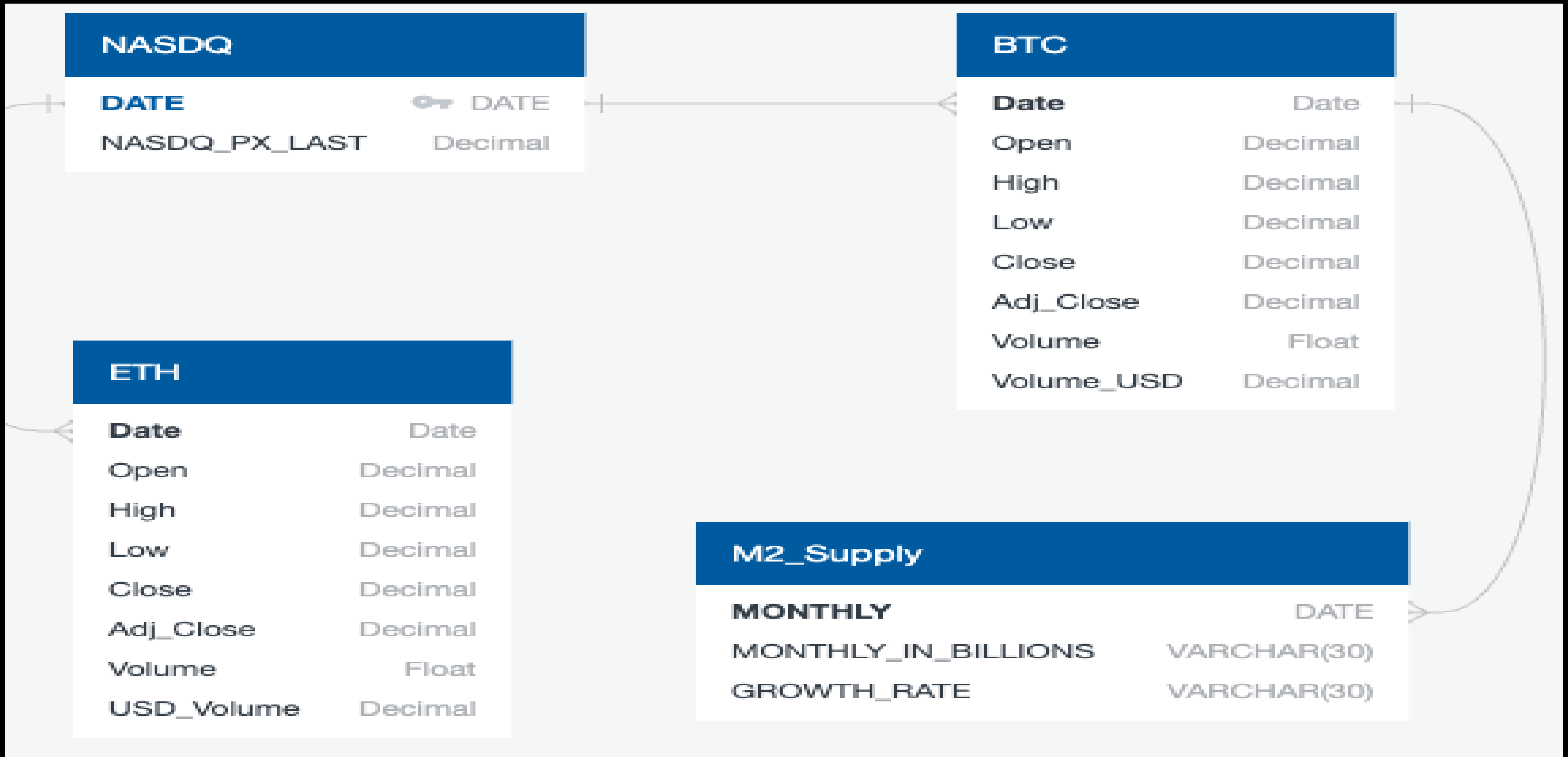
Dataset Overview:

- Contains over 3000 rows of data

Analysis Tools Utilized:

- Machine Learning (ML)
- Scikit-learn
- Time Series Analysis: LSTM
- Python Libraries:
  - Pandas
  - Matplotlib
- SQL Database

Postgres Schema



NASDQ	
DATE	🔑 DATE
NASDQ_PX_LAST	Decimal

BTC	
Date	Date
Open	Decimal
High	Decimal
Low	Decimal
Close	Decimal
Adj_Close	Decimal
Volume	Float
Volume_USD	Decimal

ETH	
Date	Date
Open	Decimal
High	Decimal
Low	Decimal
Close	Decimal
Adj_Close	Decimal
Volume	Float
USD_Volume	Decimal

M2_Supply	
MONTHLY	DATE
MONTHLY_IN_BILLIONS	VARCHAR(30)
GROWTH_RATE	VARCHAR(30)

- Schema Overview:
  - Comprised of 4 tables:
    - BTC (Bitcoin)
    - ETH (Ethereum)
    - Nasdaq
    - M2 Supply
- Data Utilization:
  - Nasdaq data serves as the key dataset for comparison
  - Facilitates the analysis of Bitcoin and Ethereum trends in the market
  - M2 Money Supply was also leveraged to identify correlations on BTC and ETH growth

Correlation and Logistical Regression

Reasons for Comparison

- Bitcoin and Ethereum are the two largest cryptocurrencies, accounting for approximately 61% of the sector's market cap.
- The correlation between Bitcoin and Ethereum can serve as an overall index for the cryptocurrency market, similar to the S&P 500 for the stock market.

Major Impacting Factors

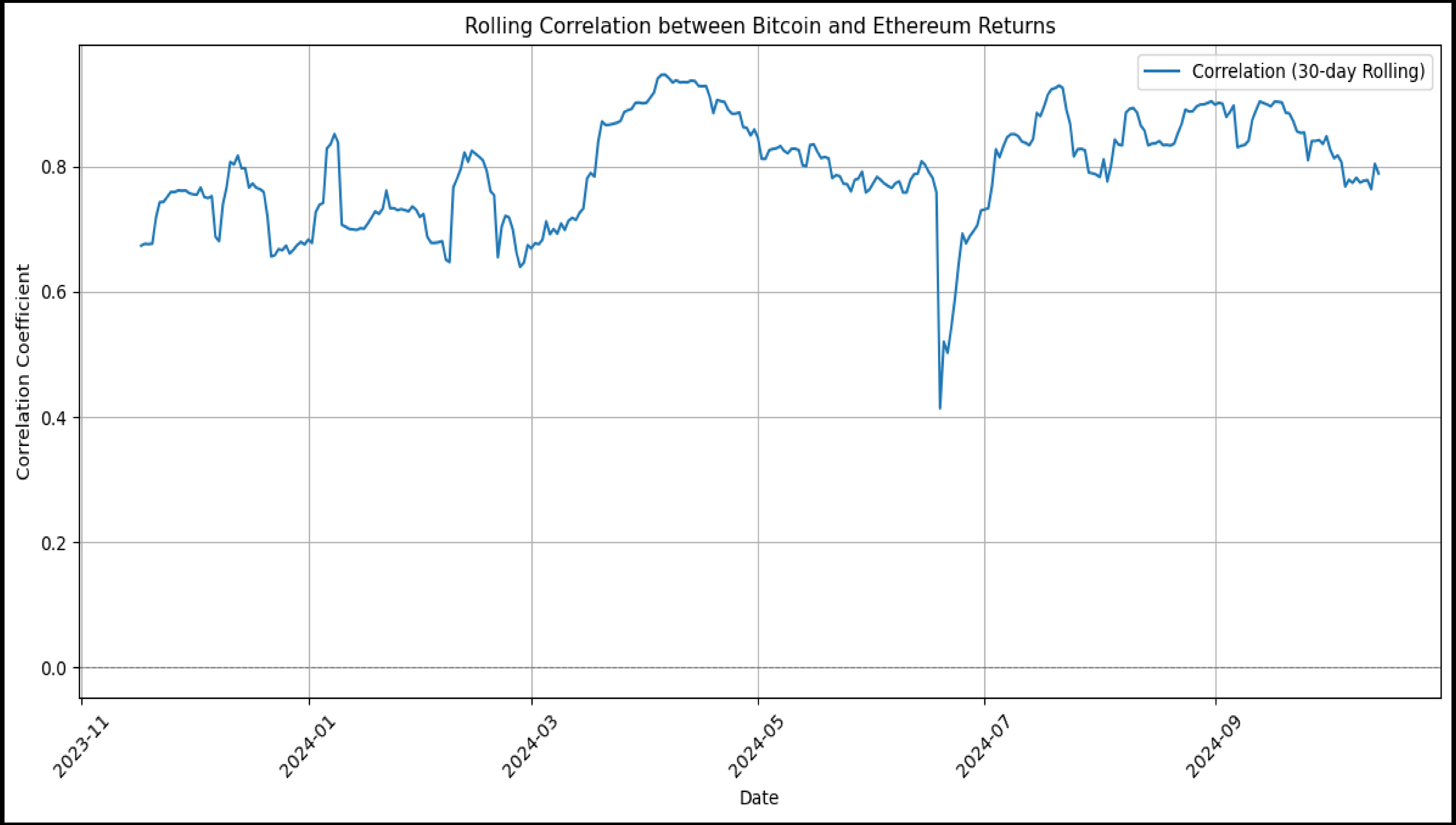
- The performance of technology stocks (NASDAQ) has a greater impact on Ethereum prices compared to Bitcoin prices.
- The US Dollar (M2 Money Supply) positively impacts Bitcoin and has a slightly negative impact on Ethereum.
- The ETH/BTC correlation may react more strongly to changes in Bitcoin, as it has a larger overall market cap than Ethereum.

Logistic Regression

- Compared the daily growth rates of prices from the NASDAQ and Ethereum to identify any correlation.
  - **Accuracy Score:** 60%
- Compared the monthly growth rates of prices from the M2 Supply and Bitcoin.
  - **Accuracy Score:** 66%

Market cap Correlation Analysis

Standardization of Data = Percent Change in Market Cap



Random Forest Prediction of Bitcoin and Ethereum

Random Forest

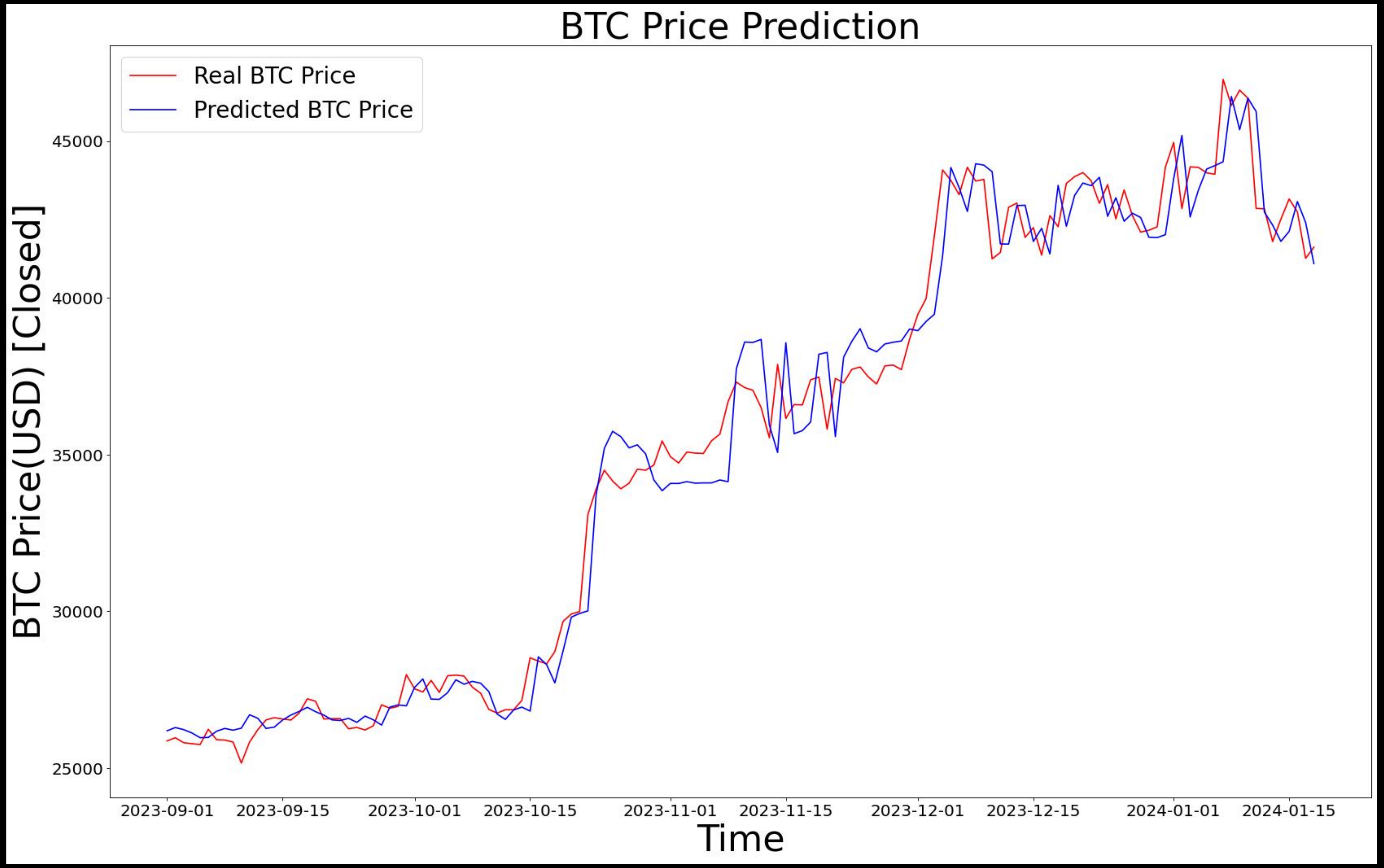
- Utilized both PySpark and Scikit-learn to build our model, using historical pricing data for Bitcoin and Ethereum from Kaggle.
- The model is a collection of decision trees used for classification and regression tasks. Each decision tree classifies an unlabeled point by casting a vote, and the random forest reports the label or value with the most votes.
- We trained our model to predict the next day's closing price, using 95% of our data for training and 5% for predictions.
- **Variables**: Open price, high price, low price, volume, volume in USD, 7-day moving average.
- **Accuracy Measurements**: RMSE (Root Mean Square Error) and r (Pearson's correlation coefficient).
- **Strengths**: Random Forest regression is effective for short-term predictions among investors.
- **Weaknesses**: Sudden volatility can lead to inaccuracies in measurements.

Results

- **BTC RMSE**: 1080
- **BTC R^2**: 0.98
- **ETH RMSE**: 61
- **ETH R^2**: 0.96

Random Forest Prediction of Bitcoin and Ethereum

Bitcoin

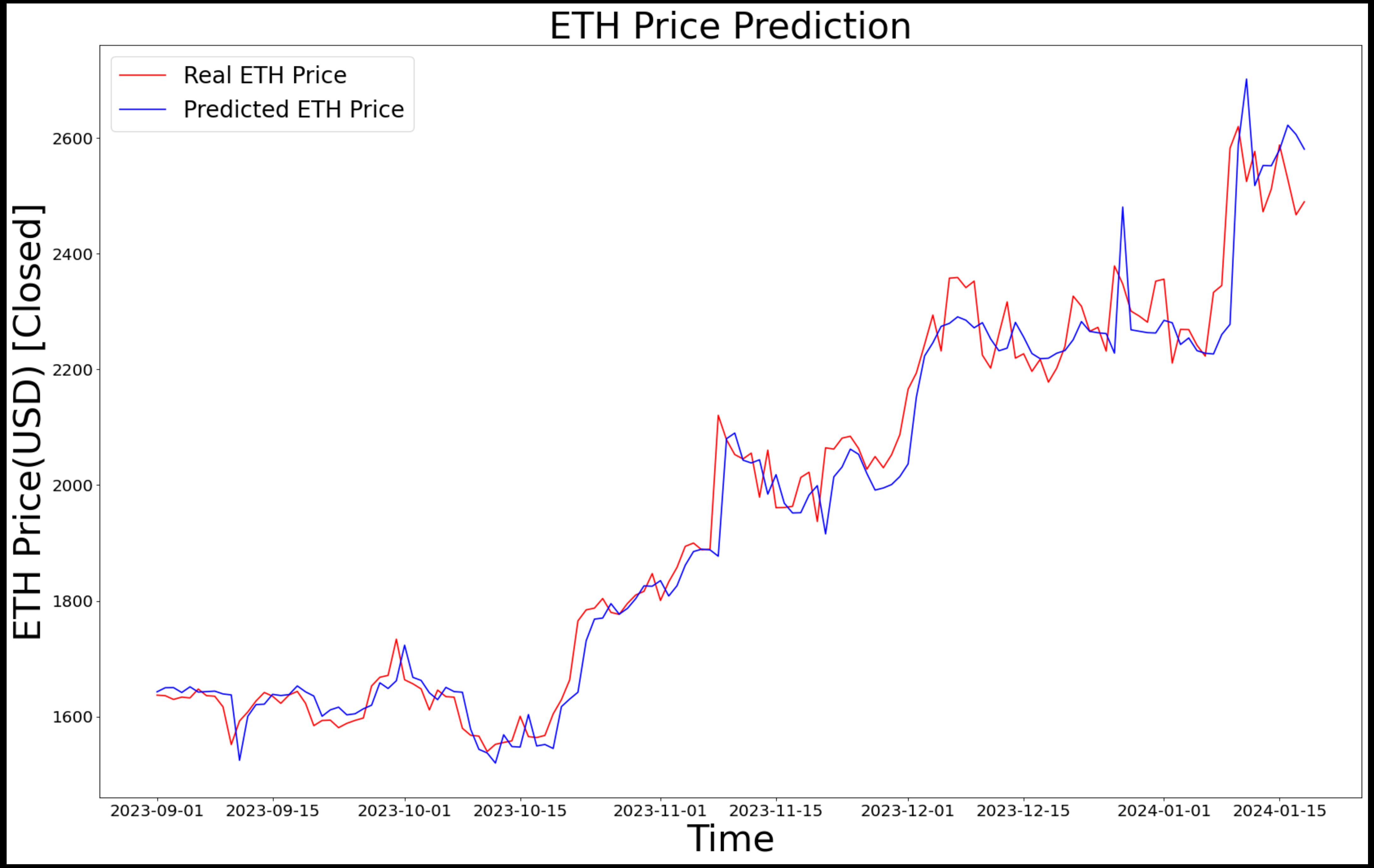


Pricing Data 2014 - 2024

Predictions are for 1st quarter 2024

VS

Ethereum



Pricing Data 2017-2024

Predictions are for 1st quarter 2024.

LSTM Prediction of Bitcoin and Ethereum

LSTM (Long Short-Term Memory)

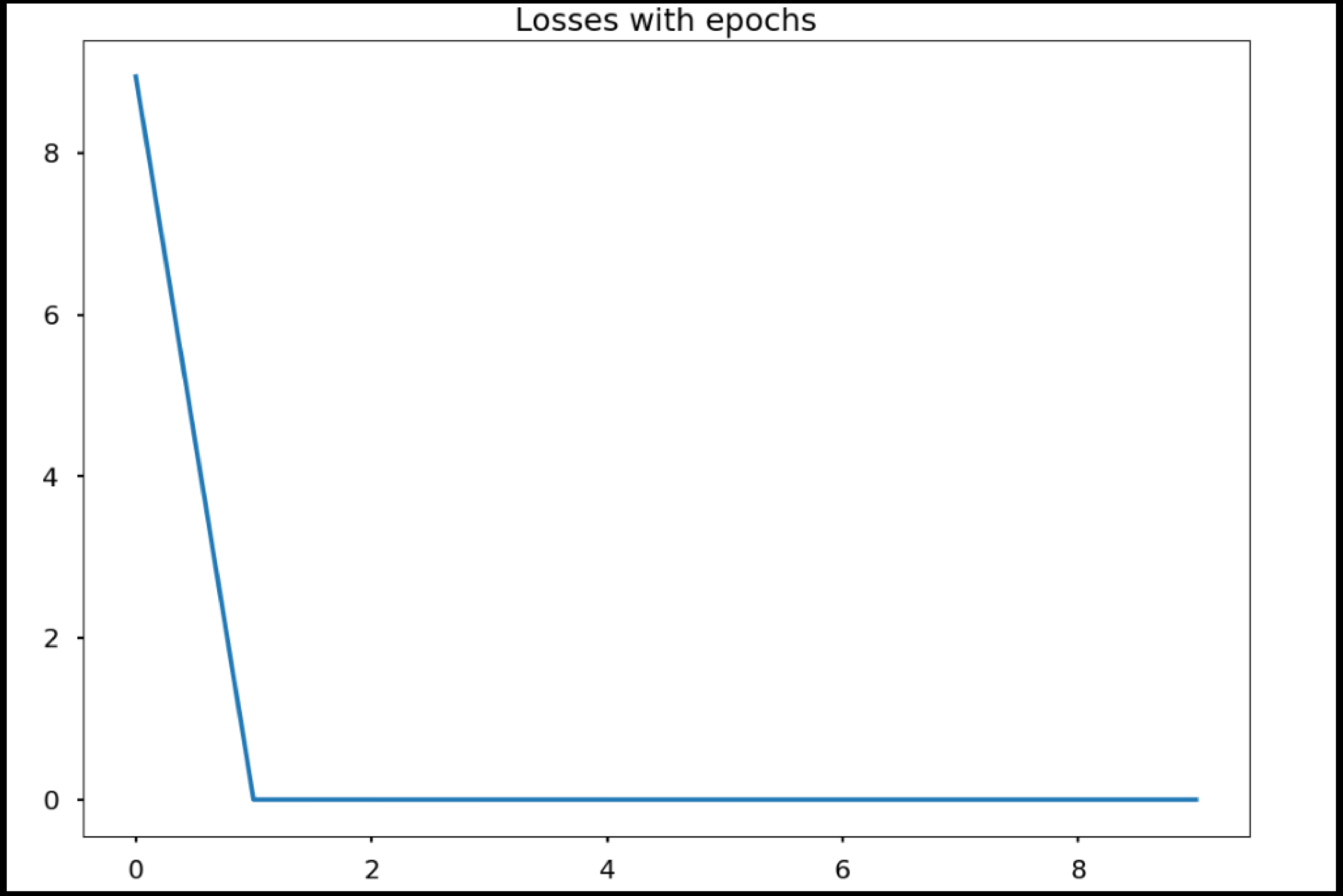
- Used Scipy, Scikit-learn, Statsmodels, and tensorflow.keras to build our model, using historical pricing data for Bitcoin and Ethereum from Coinbase.
- The model consists of memory cells, forget gate, input gate and an output gate. First the model analyses the time series, then the model continually processes the input along with the previous hidden state and cell state. During this processing the gates operate to update the cell state and hidden states based on current and previous states.
- We trained our model to predict the next day’s closing price, using 80% of our data for training and 20% for predictions.
- **Variables**: Close Price
- **Accuracy Measurements**: Optimizer adam and MSE (Mean Square Error) .
- **Strengths**: LSTM are strong performers for long-term time series forecasting data. Model become stronger with the larger the data set was.
- **Weaknesses**: Models take a long time to process and optimizing your model may be difficult

Difference between the currencies

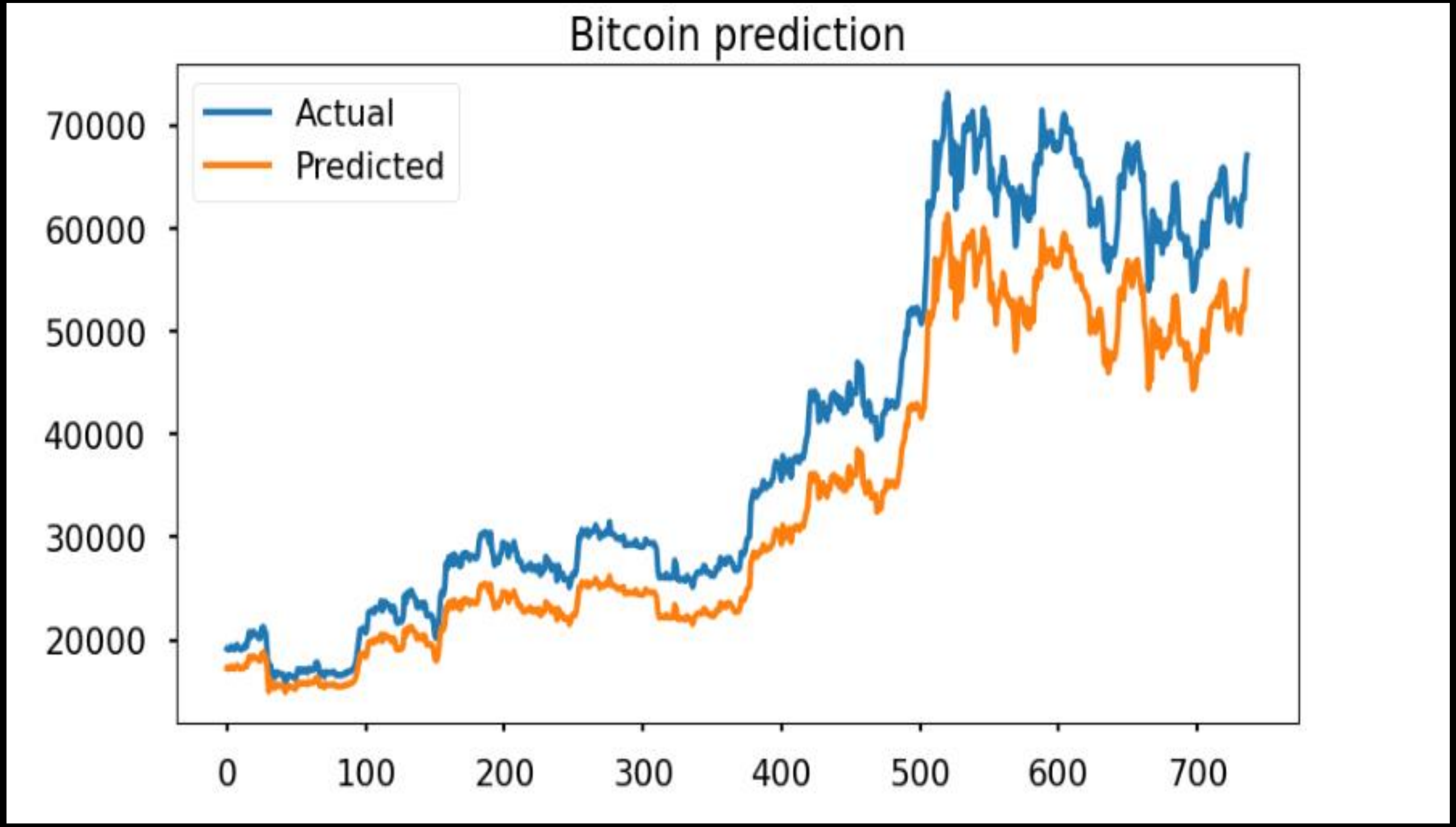
- **The BTC model had an activation layer and one hidden layer. We tested 10 epochs with a batch size of 50**
- **BTC Model had an optimal MSE of 2.1**
- **The ETH model had two activation layers and on hidden layer. We tested 12 epochs with a batch size of 50**
- **The ETH model had an optimal MSE of 3.9**

LSTM Prediction of Bitcoin

Charting the mean standard error for each epoch

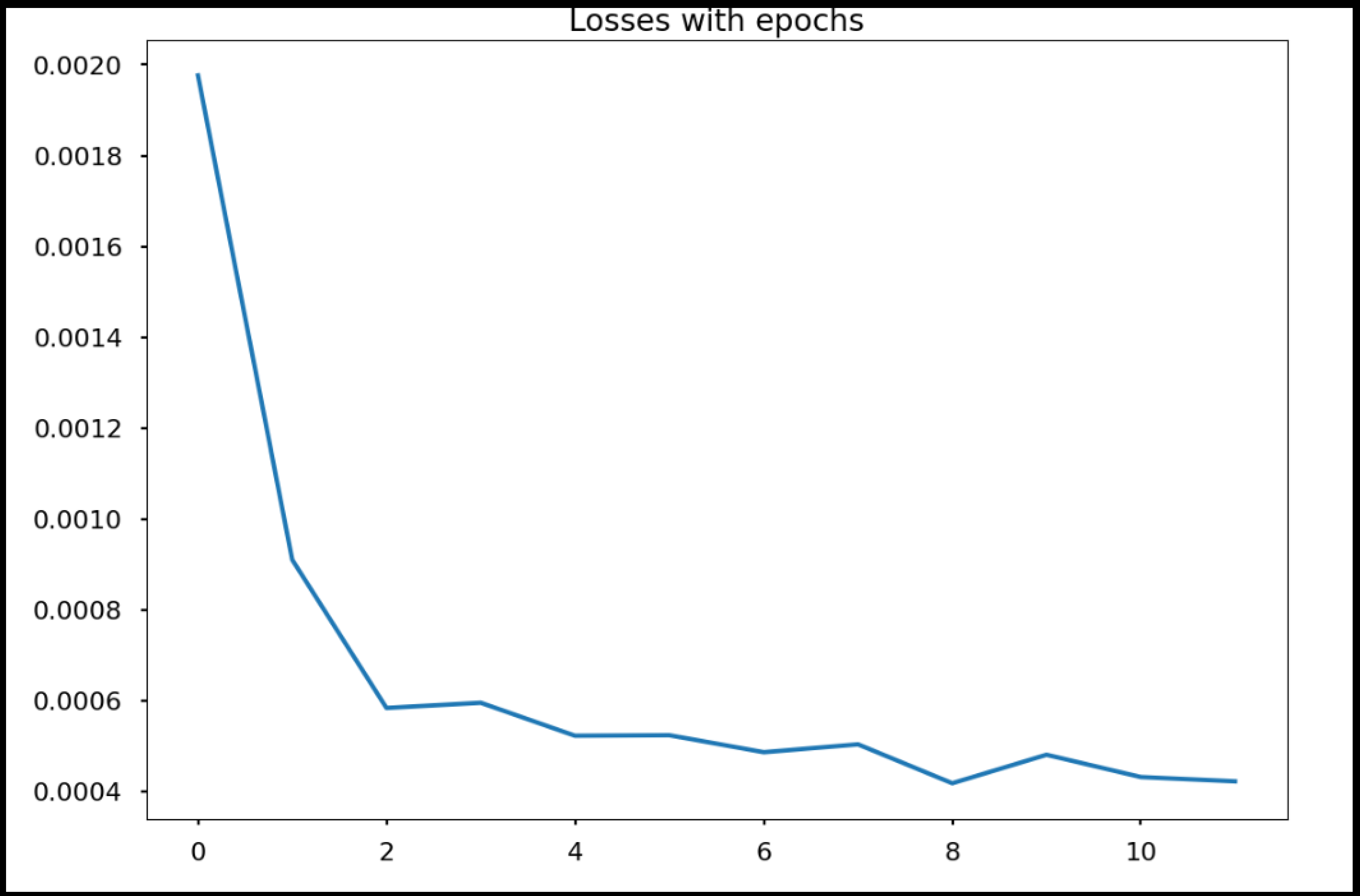


Comparing Predictions vs Actual from September 28, 2022 to October 16, 2024

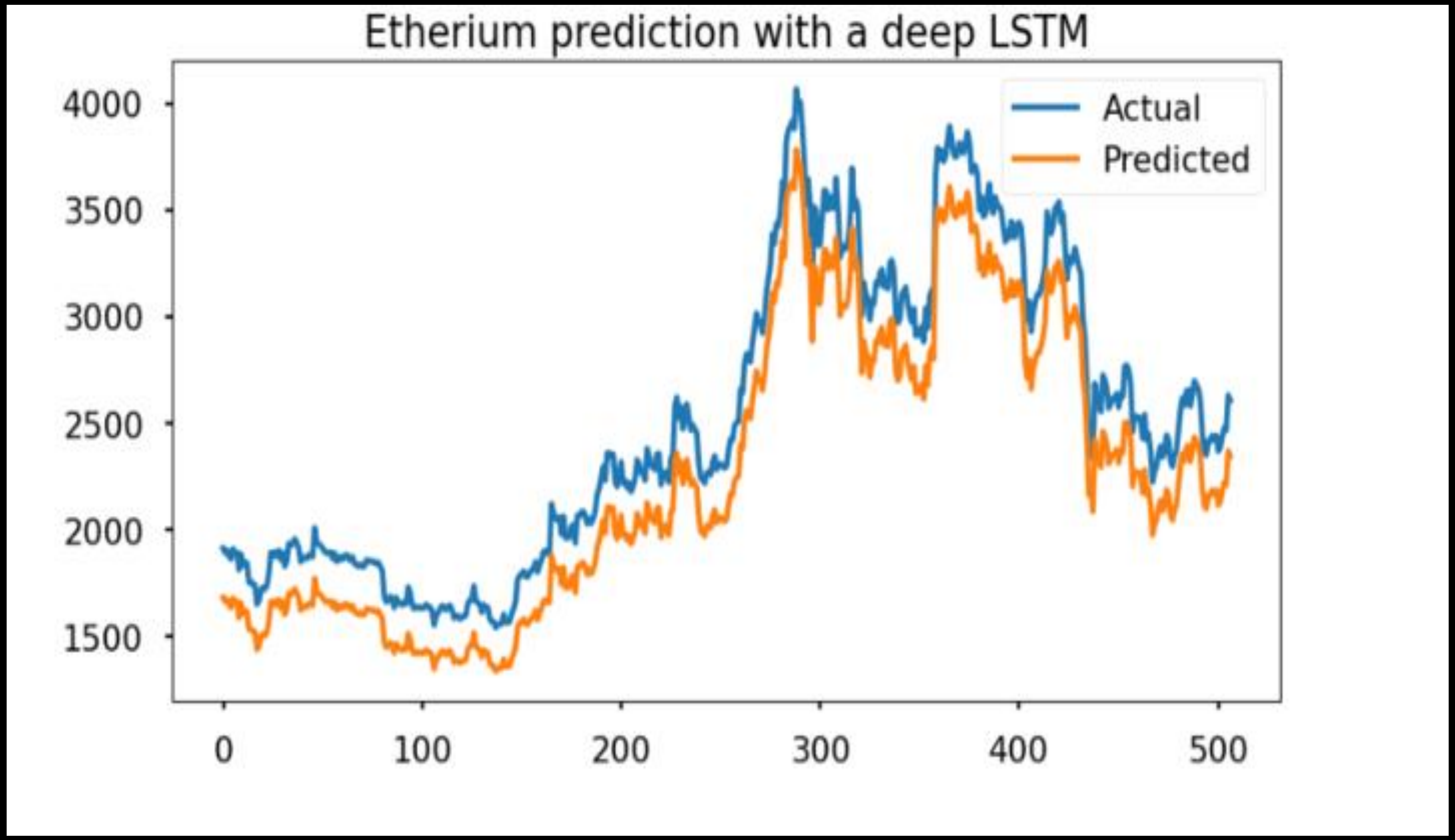


LSTM Prediction of Ethereum

Charting the mean standard error for each epoch



Comparing Predictions vs Actual from
April 16, 2023 to October 16, 2024



Conclusion

Conclusion

- Random Forest is a preferred model if the data set is limited and for a short-term forecast. Also, is a simpler model that is easier to train. Does not work well with high volatility data.

- LSTM is preferred if you have a large data set that wants to make long term predictions. It is more complex and is harder to train but can give accurate results.