**Cryptocurrency Price Forecasting with Sentiment Analysis**

**BTC, ETH, BNB — Dataset Preparation & Preprocessing Report**

**Objective**

This project aims to build forecasting models (ARIMA, Prophet, LSTM) for three major cryptocurrencies — **Bitcoin (BTC)**, **Ethereum (ETH)**, and **Binance Coin (BNB)** — and evaluate how **news sentiment** affects price predictions. The final goal is to integrate these models into a web application that predicts the next 7 days of prices based on recent market data and optional user-provided news.

**Datasets Used**

**1. Coin Price Data**

* **Bitcoin (BTC)**:  
  <https://coinmarketcap.com/currencies/bitcoin/historical-data/>
* **Ethereum (ETH)**:  
  <https://coinmarketcap.com/currencies/ethereum/historical-data/>
* **Binance Coin (BNB)**:  
  <https://coinmarketcap.com/currencies/bnb/historical-data/>

Each dataset consists of daily open, high, low, close, volume, market cap, and time-related columns from 2010 to 2025.

**2. News Dataset**

* **Crypto News Dataset**:  
  <https://github.com/soheilrahsaz/cryptoNewsDataset>  
  This includes headlines, news timestamps, sources, and associated cryptocurrency tags, as well as labeled sentiment reaction metrics (liked, disliked, etc.).

**Tools & Environment**

* **IDE**: JupyterLab (local environment)
* **Python Version**: 3.13.5
* **Working with:** .csv files (coin prices, news, sentiment scores)

**Libraries Used**

| **Library** | **Purpose** |
| --- | --- |
| pandas | Data loading, manipulation, and cleaning |
| numpy | Numerical operations and array handling |
| matplotlib / seaborn | Visual exploration (optional) |
| vaderSentiment | To compute sentiment polarity from news text |
| datetime | Date parsing and transformations |
| statsmodels | For ARIMA/ARIMAX time series modeling |
| prophet | Facebook's model for time series forecasting with trend/seasonality |
| tensorflow/keras | For building and training LSTM deep learning models |
| joblib / pickle | For saving and loading trained models |

**Preprocessing Steps (All Coins)**

1. **Loaded raw CSV files** for BTC, ETH, BNB from CoinMarketCap.
2. **Dropped irrelevant columns**:
   * timeOpen, timeClose, timeHigh, timeLow, name
3. **Parsed the timestamp column** to datetime.
4. **Removed timezone information** and normalized the timestamp to retain only the date.
5. **Set timestamp as the DataFrame index**, and sorted by date.
6. **Final Cleaned Columns**:
   * open, high, low, close, volume, marketCap, and timestamp (as index)
7. **Saved cleaned DataFrames**:
   * BTC\_cleaned.csv, ETH\_cleaned.csv, BNB\_cleaned.csv

**News Dataset Cleaning & Sentiment Extraction**

1. **Loaded news dataset** from GitHub.
2. **Filtered rows** by 'currencies' column to isolate news relevant to each coin.
3. **Handled missing description fields** by replacing them with empty strings.
4. **Combined title and description** as the input for sentiment analysis.
5. **Applied VADER sentiment analyzer** to calculate a compound sentiment score.
6. **Created per-coin sentiment CSVs**:
   * BTC\_sentiment.csv, ETH\_sentiment.csv, BNB\_sentiment.csv
7. **Merged sentiment data with cleaned price data** on the timestamp column.
8. **Filled missing sentiment values with 0** for days with no matching news.

**Final Cleaned & Merged Datasets (Ready for Modeling)**

* BTC\_final.csv
* ETH\_final.csv
* BNB\_final.csv

Each file contains:

* Daily: open, high, low, close, volume, marketCap
* Merged: sentiment score (ranging from -1 to 1)

**Model Development and Evaluation**

In this phase, I experimented with three key time series forecasting models — **ARIMA**, **Prophet**, and **LSTM** — on three major cryptocurrencies: **Bitcoin (BTC)**, **Ethereum (ETH)**, and **Binance Coin (BNB)**. Each model was trained using a train-test split approach where the **last 30 days of data were used as the test set**, and the rest of the data was used for training. The goal was to forecast the next 30 days of closing prices and evaluate the models using key performance metrics: **RMSE**, **MAE**, **MAPE**, and **R² Score**.

I also investigated the effect of incorporating **sentiment scores** as an external regressor and explored the impact of including additional features such as **volume** and **market capitalization** in LSTM models.

**Evaluation Metric Interpretation for Cryptocurrency Forecasting**

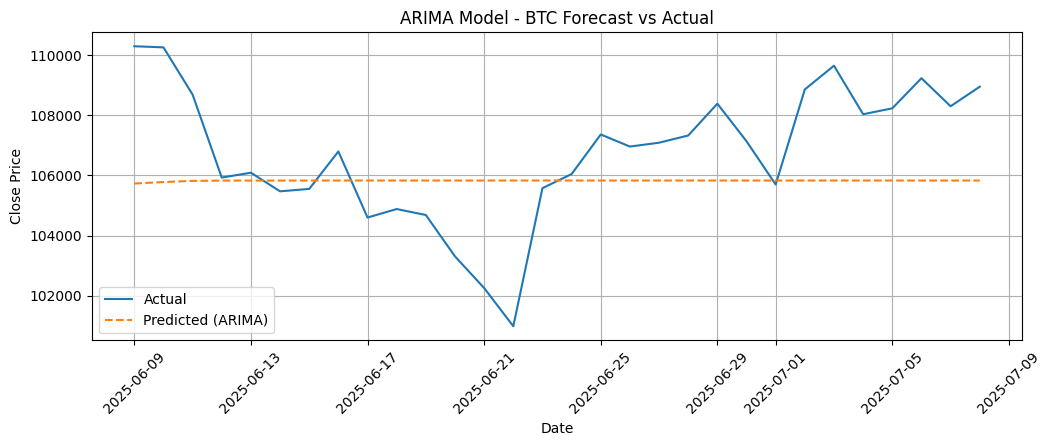
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | Good Performance | Acceptable Performance | Poor Performance | Interpretation |
| RMSE (Root Mean Squared Error) | - BTC: < 600 - ETH: < 40 - BNB: < 5 | - BTC: 600–1500 - ETH: 40–100 - BNB: 5–12 | - BTC: > 1500 - ETH: > 100 - BNB: > 12 | Indicates how far predictions deviate from actual prices. Lower is better; RMSE is sensitive to large errors. |
| MAE (Mean Absolute Error) | - BTC: < 450 - ETH: < 30 - BNB: < 3.5 | - BTC: 450–1200 - ETH: 30–75 - BNB: 3.5–9 | - BTC: > 1200 - ETH: > 75 - BNB: > 9 | Measures average magnitude of errors. Less sensitive to outliers than RMSE. |
| MAPE (Mean Absolute Percentage Error) | < 2% | 2% – 5% | > 5% | Shows error as a percentage of actual prices. Ideal for comparing across coins. Very sensitive when prices are low. |
| R² Score (Coefficient of Determination) | > 0.6 | 0.3 – 0.6 | < 0.3 or negative | Measures how well predictions explain variance in the data. Closer to 1 indicates stronger explanatory power. |

**ARIMA & ARIMAX Results**

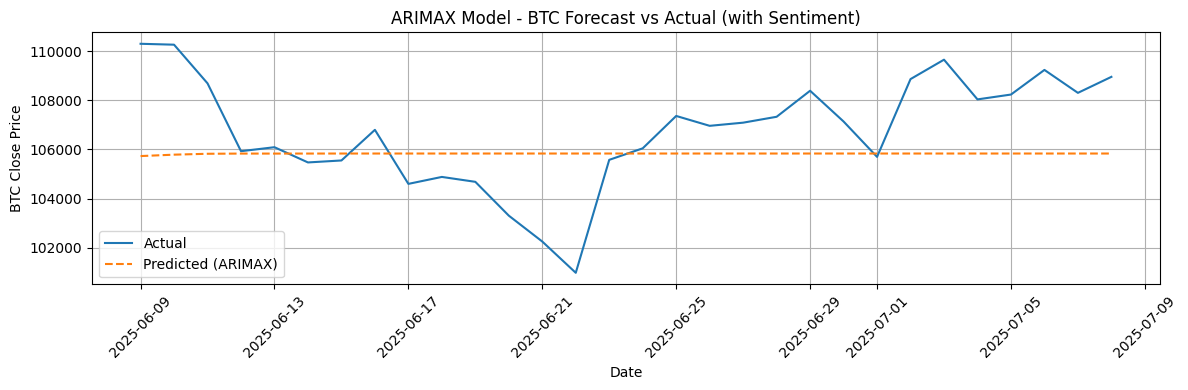
The **ARIMA model**, a classical time series approach, was trained on closing price data only, while **ARIMAX** extended this by including news sentiment as an exogenous variable.

**▪ BTC**

* **ARIMA**  
  R² Score: -0.1836  
  RMSE: 2409.43  
  MAE: 1948.18  
  MAPE: 1.82%

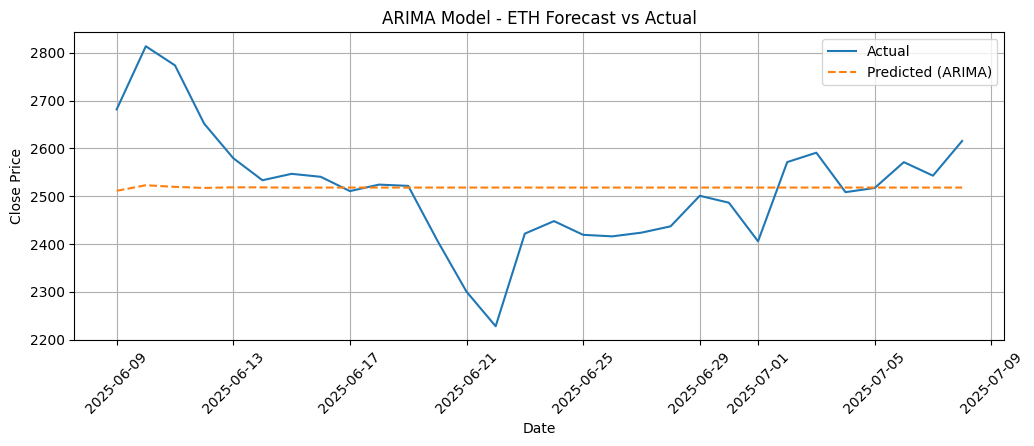


* **ARIMAX (with sentiment)**  
  R² Score: -0.1821  
  RMSE: 2407.95  
  MAE: 1946.92  
  MAPE: 1.82%

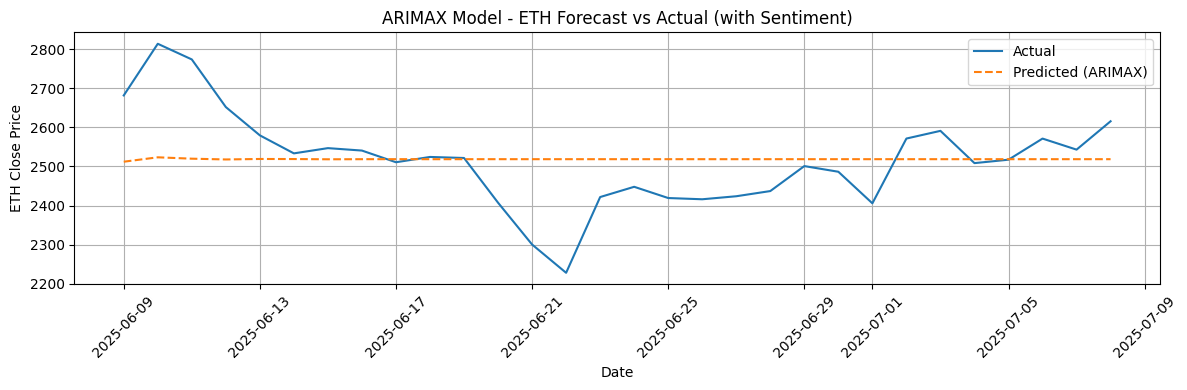


**▪ ETH**

* **ARIMA**  
  R² Score: 0.0025  
  RMSE: 119.71  
  MAE: 87.58  
  MAPE: 3.49%

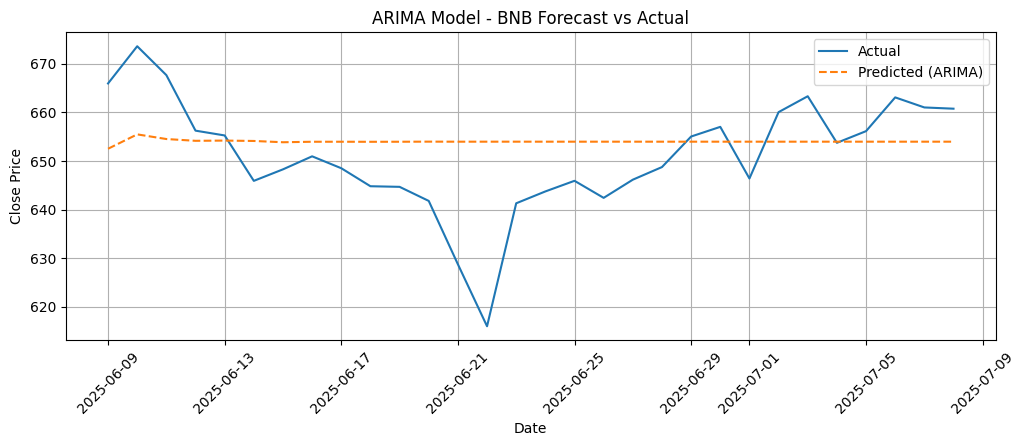


* **ARIMAX (with sentiment)**  
  R² Score: 0.0024  
  RMSE: 119.71  
  MAE: 87.58  
  MAPE: 3.49%

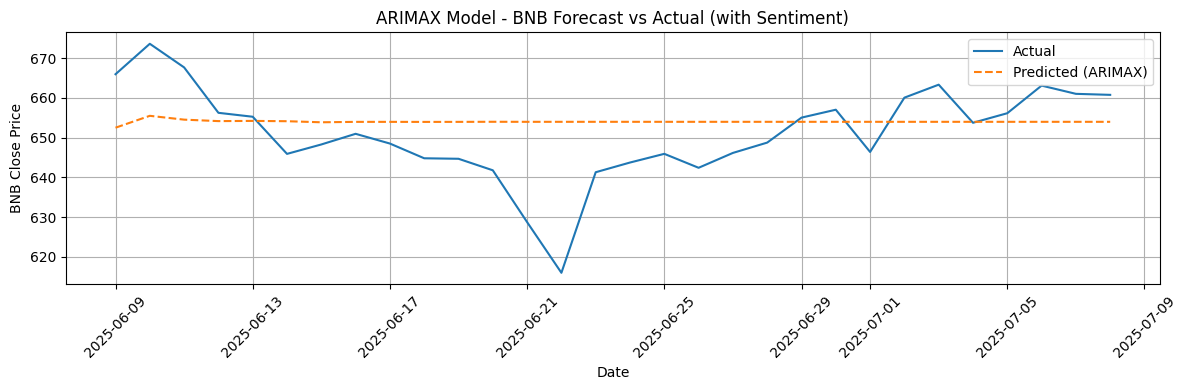


**▪ BNB**

* **ARIMA**  
  R² Score: -0.0535  
  RMSE: 11.76  
  MAE: 9.07  
  MAPE: 1.40%



* **ARIMAX (with sentiment)**  
  R² Score: -0.0540  
  RMSE: 11.76  
  MAE: 9.07  
  MAPE: 1.40%

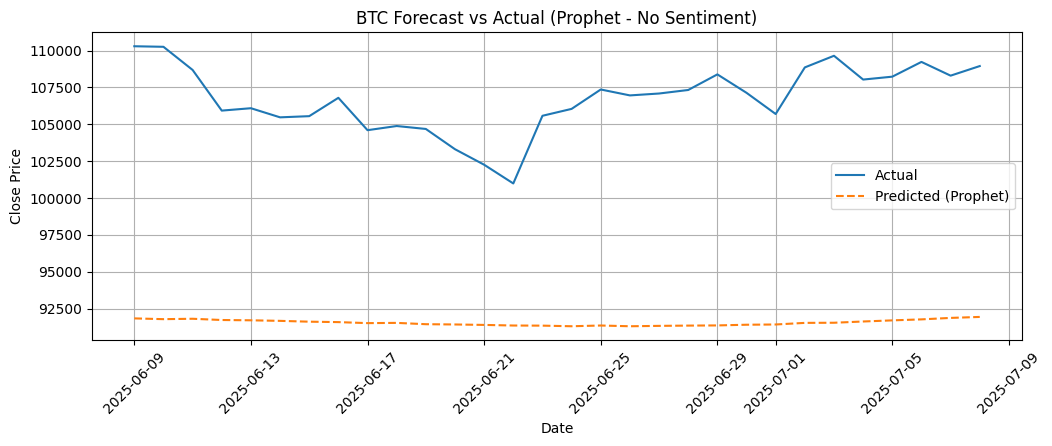


**Prophet & Prophet with Sentiment**

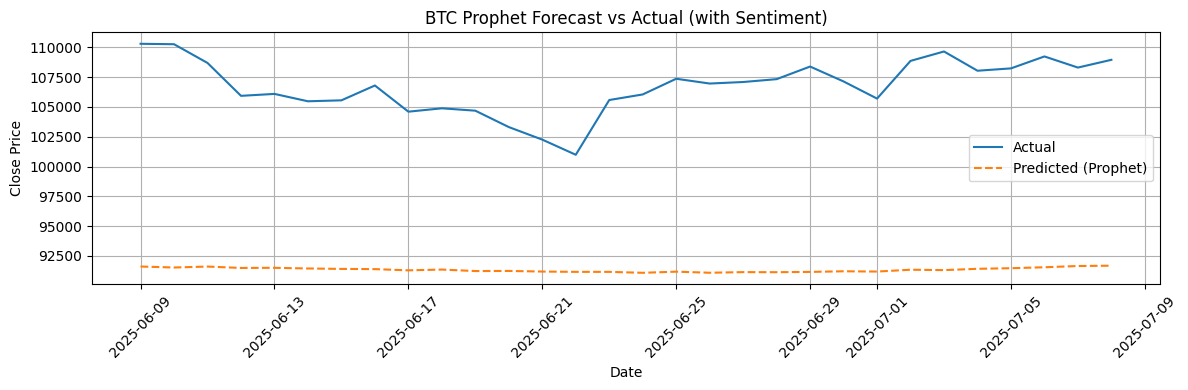
**Prophet**, developed by Facebook, was used both with and without sentiment as an additional regressor. While flexible in handling seasonality and trends, Prophet generally underperformed on this dataset.

**▪ BTC**

* **Prophet**  
  R² Score: -47.0557  
  RMSE: 15352.93  
  MAE: 15205.29  
  MAPE: 14.21%

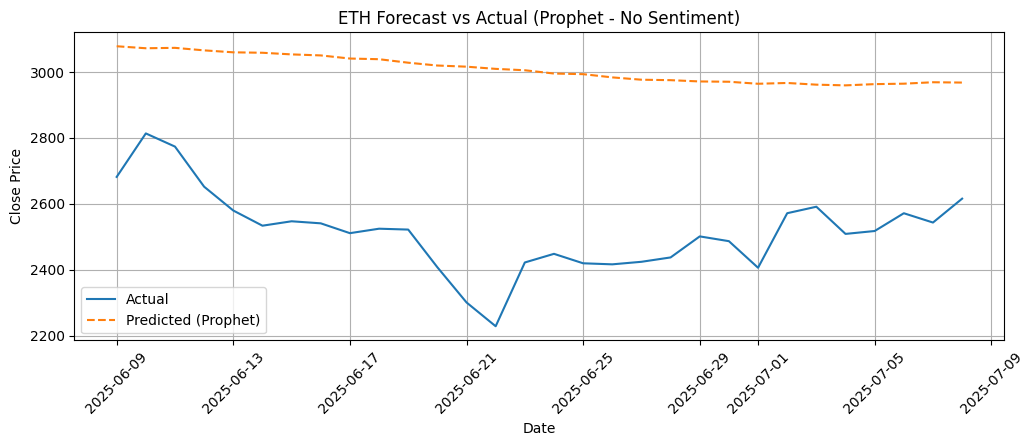


* **Prophet (with sentiment)**  
  R² Score: -48.328  
  RMSE: 15554.97  
  MAE: 15408.15  
  MAPE: 14.40%

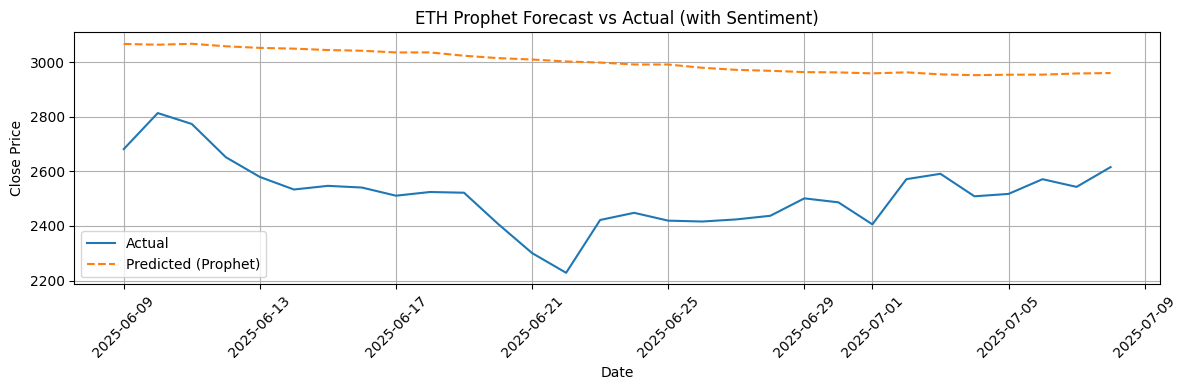


**▪ ETH**

* **Prophet**  
  R² Score: -16.6722  
  RMSE: 503.85  
  MAE: 491.98  
  MAPE: 19.79%

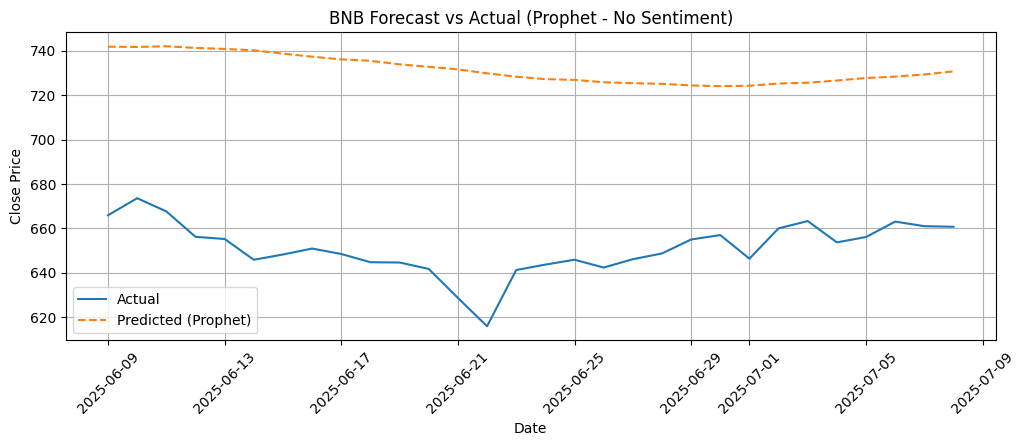


* **Prophet (with sentiment)**  
  R² Score: -16.2335  
  RMSE: 497.56  
  MAE: 485.35  
  MAPE: 19.53%

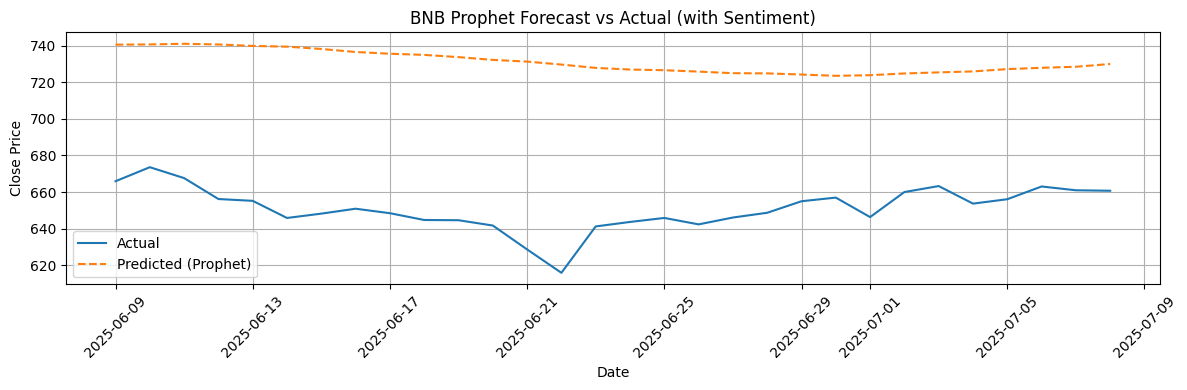


**▪ BNB**

* **Prophet**  
  R² Score: -49.3622  
  RMSE: 81.32  
  MAE: 80.48  
  MAPE: 12.39%



* **Prophet (with sentiment)**  
  R² Score: -48.6910  
  RMSE: 80.78  
  MAE: 79.92  
  MAPE: 12.31%



**LSTM (Long Short-Term Memory) Results**

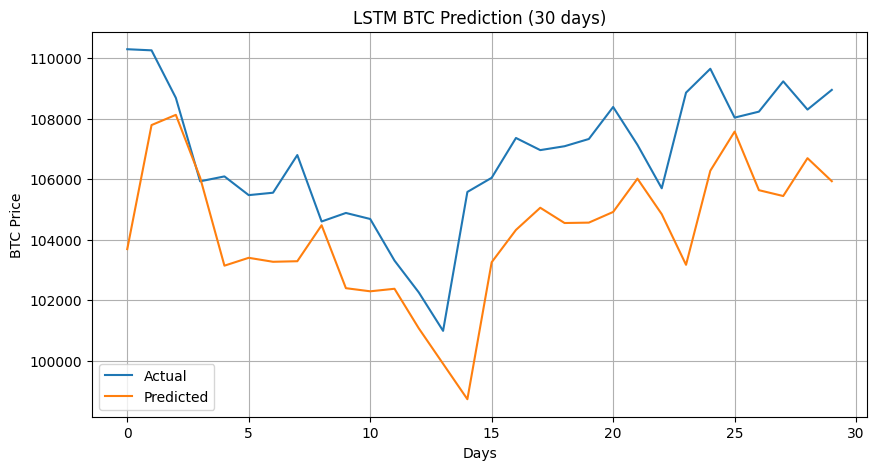
LSTM models were trained in three configurations:

* Using only the **closing price**
* Including **sentiment**
* Adding **sentiment**, **volume**, and **market cap**

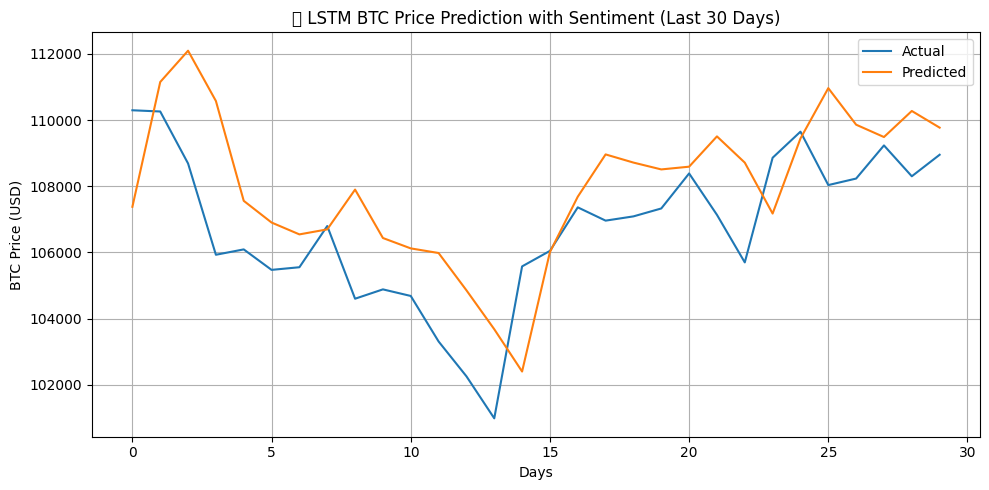
LSTM models significantly outperformed ARIMA and Prophet, especially when provided with more input features.

**▪ BTC**

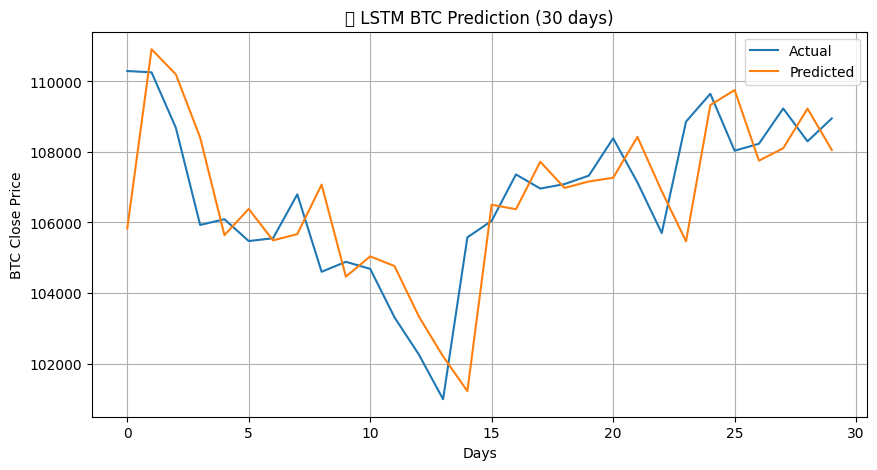
* **LSTM (Close only)**  
  R² Score: -0.8234  
  RMSE: 2990.62  
  MAE: 2487.72  
  MAPE: 2.32%



* **LSTM (with sentiment)**  
  R² Score: 0.0767  
  RMSE: 2128.14  
  MAE: 1783.99  
  MAPE: 1.68%

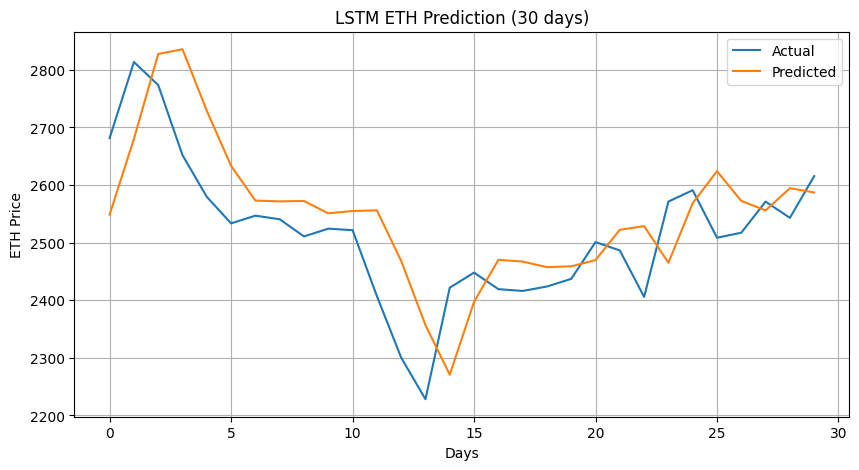


* **LSTM (with sentiment, volume, market cap)**  
  R² Score: 0.4214  
  RMSE: 1684.65  
  MAE: 1265.14  
  MAPE: 1.18%

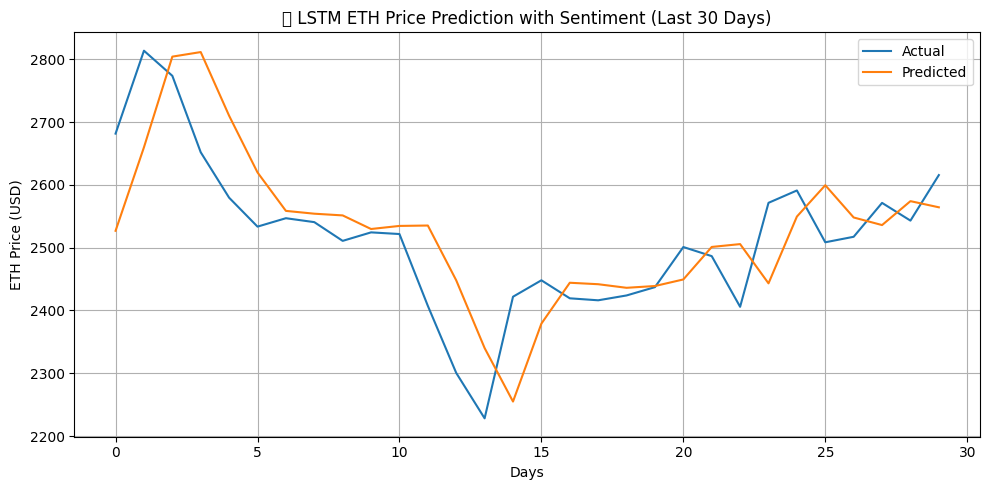


**▪ ETH**

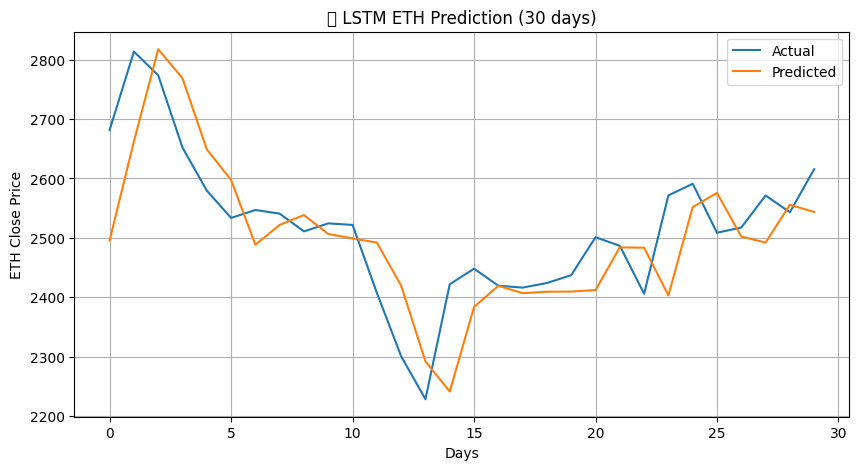
* **LSTM (Close only)**  
  R² Score: 0.3967  
  RMSE: 93.09  
  MAE: 77.36  
  MAPE: 3.09%



* **LSTM (with sentiment)**  
  R² Score: 0.4656  
  RMSE: 87.62  
  MAE: 68.73  
  MAPE: 2.73%

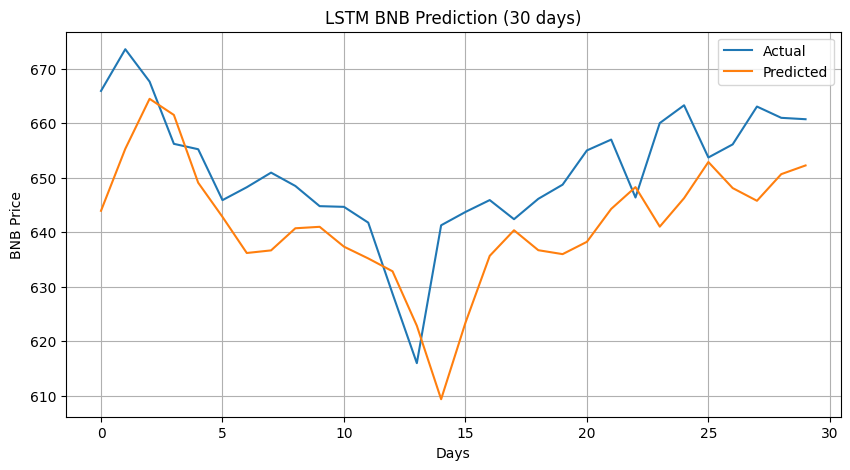


* **LSTM (with sentiment, volume, market cap)**  
  R² Score: 0.5111  
  RMSE: 83.81  
  MAE: 65.47  
  MAPE: 2.59%

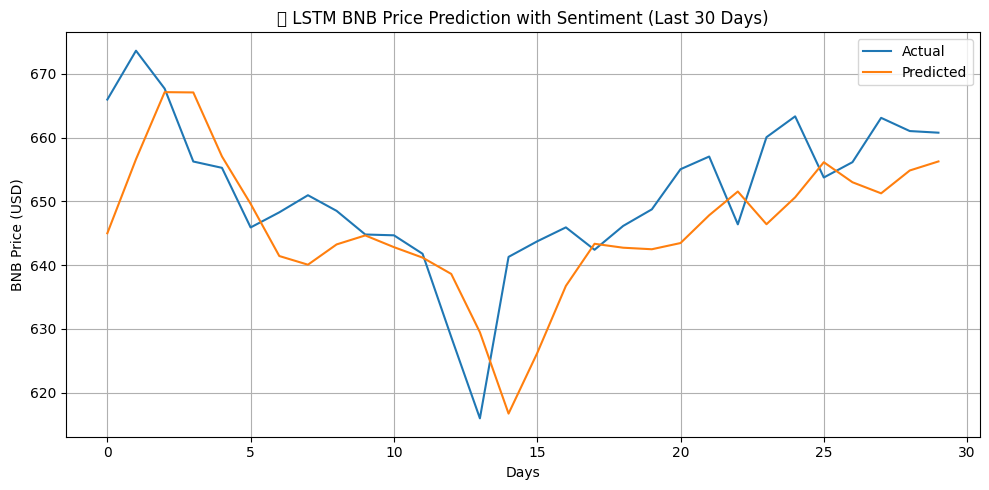


**▪ BNB**

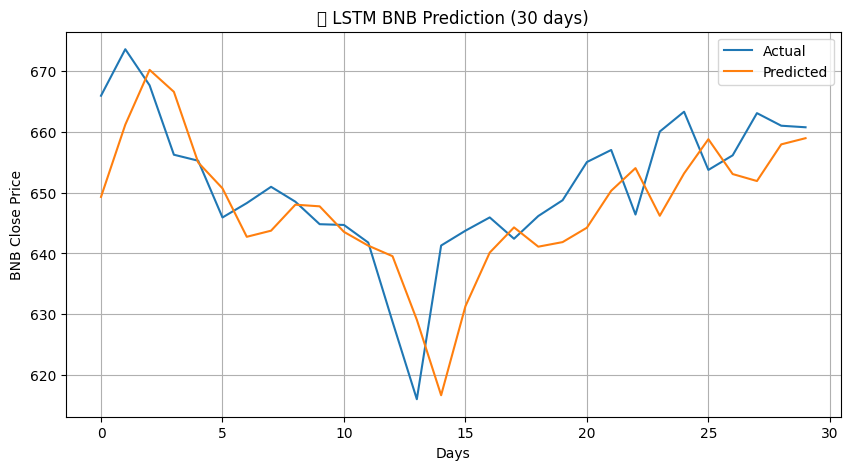
* **LSTM (Close only)**  
  R² Score: -0.2497  
  RMSE: 12.81  
  MAE: 10.66  
  MAPE: 1.63%



* **LSTM (with sentiment)**  
  R² Score: 0.1900  
  RMSE: 10.31  
  MAE: 8.20  
  MAPE: 1.26%



* **LSTM (with sentiment, volume, market cap)**  
  R² Score: 0.3629  
  RMSE: 9.15  
  MAE: 7.29  
  MAPE: 1.12%



**Summary of Results**

The goal of this study was to forecast cryptocurrency prices using historical data, incorporating optional sentiment and market indicators, and compare the performance of three different modeling approaches: **ARIMA/ARIMAX**, **Prophet**, and **LSTM**. Each model was evaluated on **Bitcoin (BTC)**, **Ethereum (ETH)**, and **Binance Coin (BNB)** using a consistent train-test split strategy, where the **last 30 days** of data were reserved for testing.

Below is a breakdown of key insights drawn from the experimental results:

**1. ARIMA vs. ARIMAX (with Sentiment)**

ARIMA served as a baseline model using only historical close prices, while ARIMAX incorporated **sentiment scores** as an external regressor.

* Across **BTC**, **ETH**, and **BNB**, **ARIMAX showed marginal improvements** over ARIMA in terms of RMSE, MAE, and R². However, the improvement was often negligible.
* The **R² scores** for both ARIMA and ARIMAX were **low or negative**, particularly on BTC and BNB, suggesting poor predictive power. This is likely due to the model’s linear assumptions and inability to adapt to the **non-stationary and highly volatile nature** of crypto markets.
* Despite including sentiment, **ARIMAX failed to capture complex price patterns**, making it insufficient for short-term high-volatility prediction, though still useful for simple trend forecasting.

**2. Prophet vs. Prophet with Sentiment**

Prophet, designed to capture seasonality and trend in time series data, was tested with and without sentiment.

* **Performance was consistently poor** for all three cryptocurrencies, with R² scores ranging from -16 to -49, indicating that Prophet **could not explain the variance** in the price series.
* While adding sentiment **slightly improved RMSE and MAPE**, the gains were not significant, and the models still **failed to generalize well** on unseen data.
* This confirms Prophet’s limitation when applied to financial time series that are **non-seasonal, abrupt, and noise-heavy** like cryptocurrency prices.

**3. LSTM (Univariate vs. Multivariate)**

LSTM networks, being deep learning models capable of modeling temporal dependencies, outperformed both ARIMA and Prophet across all evaluation metrics.

* **Univariate LSTM (using only close price)** already showed decent performance with moderate RMSE and MAPE, especially on **ETH**.
* Adding **sentiment as a feature improved performance further**, showing that sentiment has predictive value for price movements when processed non-linearly.
* The best performance was achieved by the **multivariate LSTM**, which combined **close price, sentiment, volume, and market cap** as inputs:
  + It achieved **the lowest RMSE and MAPE values across all three cryptocurrencies**, and the highest **R² scores**, indicating strong predictive accuracy.
  + **BTC’s R² improved from -0.82 (univariate) to 0.42 (multivariate)** — a substantial gain.