Name: Hashem Bader
Course: CS4440 – Data Science

Final Report: Cryptocurrency Forecasting

1. Context of the Problem

There are many cryptocurrency traders and businesses that deal in and use cryptocurrency, such as banks. For instance, some banks employ cryptocurrency to facilitate speedy transactions, "but because of its volatile pricing and the innovative blockchain technology it uses, they have trouble forecasting and assessing the cryptocurrency [1]". "It is much more difficult to predict the cryptocurrency because it is easily impacted by news about purchases, social media trends, and other cryptocurrencies [3]". Building trustworthy prediction models for cryptocurrency prices is crucial because they offer traders, investors, and banks useful information for successfully navigating the world of digital assets.

2. Dataset Description

"The analysis was performed using Cryptocurrency Historical Data from Kaggle by sudalairajkumar [2]". The dataset provided contains historical information about the performance of various cryptocurrencies, including 37,082 entries with 10 columns. Each record includes details such as a serial number (SNo) for each entry of each coin, the name of the cryptocurrency (e.g., Stellar), its symbol (e.g., XLM), and the date and time of the observation in the format YYYY-MM-DD HH:MM:SS. Price-related fields include the highest price (High), lowest price (Low), opening price (Open), and closing price (Close) for the given day. Additionally, it records the trading volume (Volume) and the total market capitalization (Marketcap) for the cryptocurrency on that day. This dataset can help with cryptocurrency price prediction applications.

3. Analysis and Modeling Approach

The project goal was to forecast Bitcoin prices using historical cryptocurrency data. Rather than using deep learning, the analysis was based on classical time series forecasting techniques implemented through the sktime Python library, which is well-suited for interpretable and statistically robust forecasting.

Key Steps:

- **Data Preprocessing**:
    - Focused mainly on the Bitcoin data from the dataset that has multiple cryptocurrencies.

- Cleaned and indexed the Date column to serve as the time series index.
    - The Close price column was selected as the target variable for prediction.
    - Checked for any missing or corrupted values.
- **Train-Test Split**:
    - The dataset was split into a training set and a test set, with the last 360 days of data used for testing.
    - A Forecasting Horizon was created using actual calendar dates for prediction evaluation.
- **Models Used**: A variety of common forecasting models were implemented and compared:
    - **Naive Forecaster** – Uses the last observed value for all forecasts.
    - **Seasonal Naive Forecaster** – "it is useful for highly seasonal data. it sets each forecast to be equal to the last observed value from the same season (weekly/monthly)[4]."
    - **Exponential Smoothing** – "Exponential smoothing is a broadly accurate forecasting method for short-term forecasts. The technique assigns larger weights to more recent observations while assigning exponentially decreasing weights as the observations get increasingly distant. This method produces slightly unreliable long-term forecasts[5]."
    - **Polynomial Trend Forecaster** – Fits polynomial curves to capture long-term movements.
    - **Ensemble Forecaster** – "Combines predictions from multiple models (e.g., naïve, exponential smoothing) to enhance forecast robustness.[6]"
- **Evaluation Strategy**:
    - Models were evaluated using Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE).
    - Applied sliding window cross-validation, where each window trained on 30 days and forecasted the next 7 days.
    - Results were averaged across all splits to determine the most consistent and best-performing model.
- **Forecasting Implementation**:
    - After model comparison, the Polynomial Trend Forecaster and Exponential Smoothing models showed strong performance.
    - The best model was retrained on the full Bitcoin dataset to forecast:
        - The next 7 days (short-term forecast)
        - The next 360 days (long-term forecast)
    - Forecast results were plotted to visualize trends.

4. Results and Interpretation

The evaluation demonstrated that traditional statistical models can provide solid predictive performance when applied thoughtfully to time series data like Bitcoin prices.

Performance Summary:

- Among all the models tested, the Polynomial Trend Forecaster delivered the best performance for long-term forecasting, achieving the lowest MAPE and RMSE scores. For short-term forecasting, Exponential Smoothing was chosen even though the Naive model had a slightly lower MAPE—because Exponential Smoothing provided more consistent and reliable results overall.
- These models were capable of capturing seasonal and long-term trends in the data.

Visual Insights:

- Forecast plots showed strong alignment between predicted and actual prices in most test periods in the short term but in the long term it was the opposite.
- The best model accurately tracked the general direction and trend of the market in stable periods.
- During more volatile or extreme price fluctuations, accuracy declined heavily, which is expected with statistical forecasting methods.

Practical Implication:

The models provide a good baseline for understanding price movement and can serve as helpful tools for short-term forecasting, especially in structured or seasonally influenced market conditions.

5. Limitations and Future Work

Limitations:

- The project focused only on Bitcoin, excluding other available cryptocurrencies in the dataset. This limited the model's generalizability to broader crypto markets.
- No external features (e.g., social media sentiment, news events) were included in the forecasting process.
- Forecast accuracy naturally declines during sharp market turns or highly volatile events — a common limitation in purely historical, univariate models.

Future Work:

- Expand the project to a multivariate time series model including variables like trading volume, moving averages, or news-based sentiment indicators.
- Compare classical models with more advanced ML or DL approaches like Prophet, ARIMA, or even hybrid neural-statistical models.
- Apply hyperparameter optimization techniques for fine-tuning.
- Extend the analysis to other top-traded cryptocurrencies like Ethereum and Solana, and conduct comparative performance evaluations because they were not included in the dataset.

References

[1] R. K. Rathore, D. Mishra, P. S. Mehra, O. Pal, A. S. Hashim, A. Shapi'i, T. Ciano, and M. Shutaywi, "Real-world model for bitcoin price prediction," Information Processing & Management, vol. 59, no. 4, 2022, Art. no. 102968. DOI: 10.1016/j.ipm.2022.102968. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030645732200084X

[2] The Cryptocurrency Historical Dataset: https://www.kaggle.com/datasets/sudalairajkumar/cryptocurrencypricehistory

[3] J. Liu and A. Serletis, "Volatility in the Cryptocurrency Market," Open Economies Review, vol. 30, no. 4, pp. 779–811, 2019. DOI: 10.1007/s11079-019-09547-5. [Online]. Available: https://link.springer.com/article/10.1007/s11079-019-09547-5.

[4] R.J. Hyndman and G. Athanasopoulos, "Some simple forecasting methods," *Forecasting: Principles and Practice*, 2nd ed., OTexts, 2018. [Online]. Available: https://otexts.com/fpp2/simple-methods.html

[5] Simplilearn, "An Introduction to Exponential Smoothing for Time Series Forecasting in Python," [Online]. Available: https://www.simplilearn.com/exponential-smoothing-for-time-series-forecasting-in-python-article

[6] T. Houtekamer and L. Mitchell, "Ensemble Kalman Filtering," *Journal of Computational Physics*, vol. 227, no. 1, pp. 244–266, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999107000812