Directional Forecasting Cryptocurrencies Project Overview Report

jiangqin.ma@gmail.com

1 INTRODUCTION

Cryptocurrency markets operate around the clock, presenting unique challenges and opportunities for traders and analysts. In this project, we undertake an exploratory data analysis (EDA) of a high-frequency dataset containing minute-by-minute OHLCV (Open, High, Low, Close, Volume) data for a specific cryptocurrency. The data captures essential details of price movements and trading volumes, providing a comprehensive view of market activity at a granular level (GenAI.Labs, 2024).

Beyond EDA, we aim to build predictive models capable of forecasting short-term price movements. This challenge offers a unique opportunity to delve into high-frequency trading insights within the volatile and fast-paced cryptocurrency market. It encourages innovative feature engineering, the application of cutting-edge machine learning techniques, and provides a realistic simulation of market impact.

The primary objective of this analysis is to understand the dataset's structure, uncover hidden trends, and identify anomalies that could impact predictive modeling. This EDA will serve as a foundation for developing robust models to predict short-term price movements, thereby contributing valuable insights into the dynamic and volatile cryptocurrency market.

2 EXPLORATORY DATA ANALYSIS (EDA)

2.1 Dataset Overview

The dataset provided for this project contains historical minute-by-minute OHLCV (Open, High, Low, Close, Volume) data for a specific cryptocurrency. It covers a time span from May 4, 2018, at 22:01 to May 17, 2022, at 19:58, comprising a total of 2,122,438 records across 11 columns. Each record represents one minute of trading activity, capturing essential metrics such as price movement, trading volume, and the number of trades, providing a detailed view of the cryptocurrency's market behavior.

The dataset includes columns like `timestamp`, which records the time of each data point, and price-related metrics (`open`, `high`, `low`, and `close`), which reflect the cryptocurrency's price changes during the minute. Additionally, it features `volume`, indicating the number of units traded, and `quote_asset_volume`, representing the total value of traded assets in USDT. Other columns, such as `number_of_trades`, `taker_buy_base_volume`, and `taker_buy_quote_volume`, provide insights into trading activity and market participation. The `target` column identifies the direction of the price movement in the next minute, where a value of 1 signifies an upward movement and 0 indicates no change or a downward movement.

2.2 Target Distribution

The target variable represents the direction of the cryptocurrency price movement in the next minute: 1 indicates an upward price movement, while 0 signifies no change or a downward movement. The dataset is reasonably balanced between the two classes, ensuring fair model performance across both outcomes.

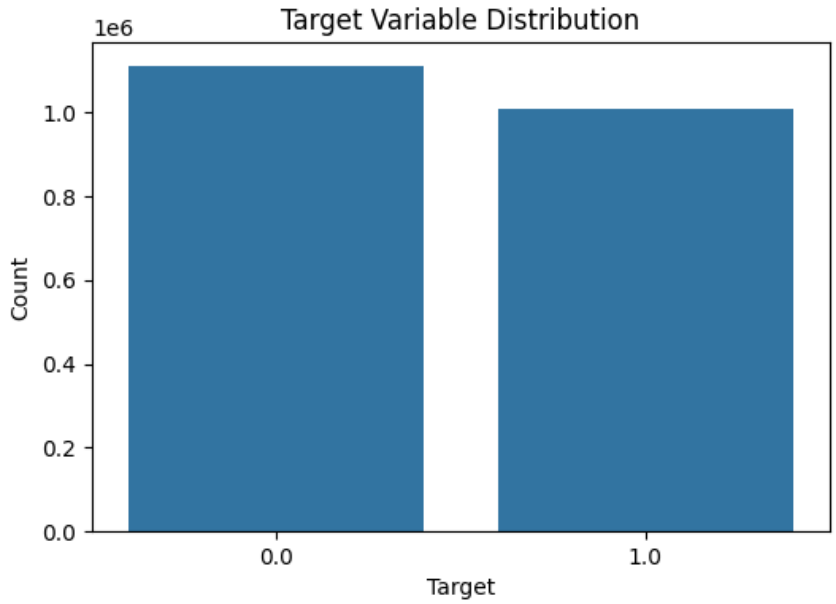

Figure 1—Target Variable Distribution

2.3 Correlation Analysis

A correlation analysis was conducted to identify relationships between features and the target variable. The results reveal minimal correlation between the `close`

price and the target variable. However, features such as `trading_volume` and the `number_of_trades` exhibit slight positive correlations with the target, suggesting they may hold some predictive value.

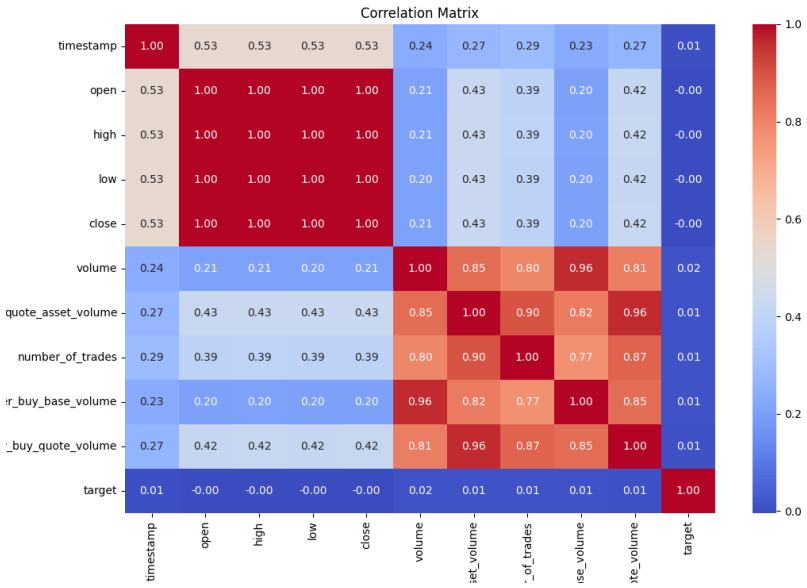

Figure 2—Correlation Matrix

2.4 Close Price Over Time

The time series of the `close` price reveals the overall price trends and volatility in the cryptocurrency market. Significant spikes and drops in the price indicate periods of high volatility, which are characteristic of cryptocurrency markets.

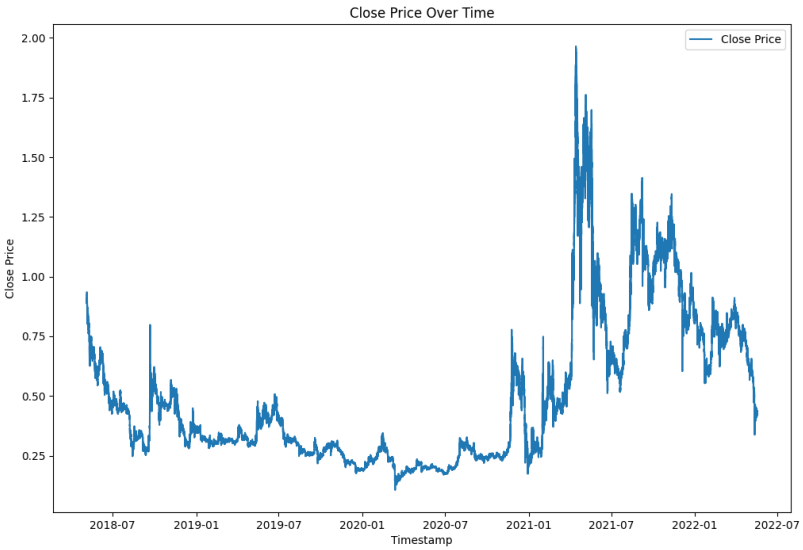

Figure 3—Close Price Over Time

2.5 Trading Volume Over Time

The `trading_volume` provides insights into market activity, with noticeable spikes pointing to periods of heightened trading activity that could signal increased market interest or external events.

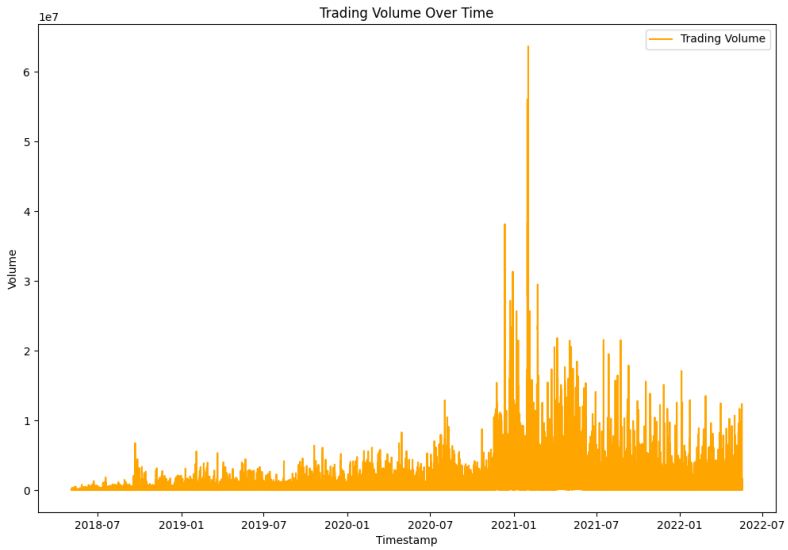

Figure 4—Trading Volume Over Time

2.6 Close Price by Target

The boxplot below shows how the `close` price varies with the target values. Both classes exhibit similar distributions, but higher prices appear slightly more common in upward movements (`target = 1`).

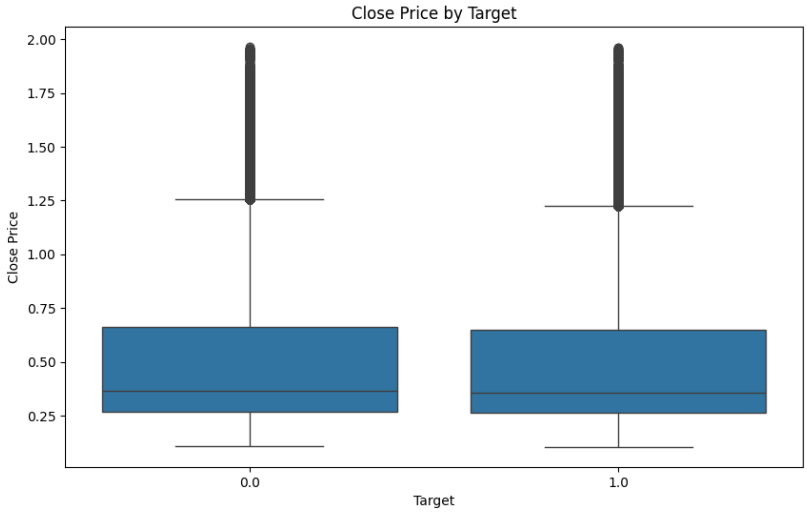

Figure 5—Close Price by Target

2.7 Volume Outliers

A boxplot of the `trading_volume` highlights the presence of extreme values. While most trading volumes fall within a reasonable range, the outliers could significantly impact model performance if not appropriately handled during preprocessing.

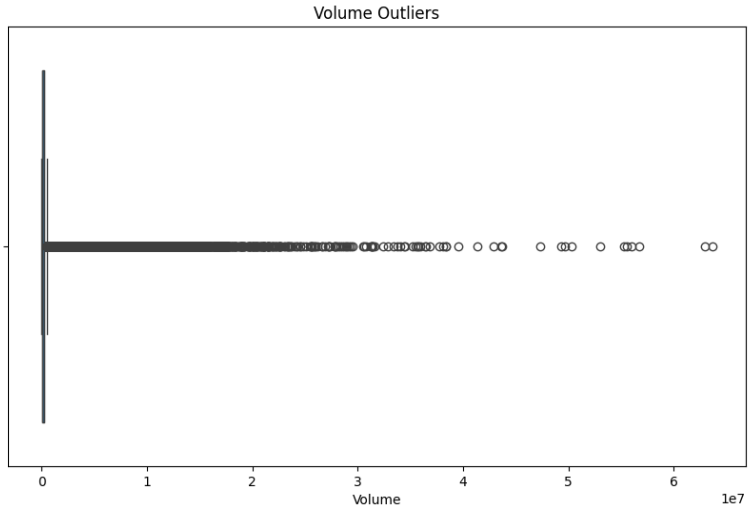

Figure 6—Volume Outliers

SUMMARY OF EDA

The EDA highlights several important aspects of the dataset. The balanced target variable distribution is promising for fair model evaluation. While individual feature correlations with the target are minimal, trading activity features like `volume` and `number_of_trades` may hold predictive power. The time series plots underscore the volatility and activity spikes characteristic of cryptocurrency markets. Additionally, the strong relationships between volume-related features provide opportunities for further feature engineering. These insights lay the groundwork for building robust predictive models.

3 METHODS

3.1 Overview

In this project, I employ three distinct modeling approaches to predict the next-minute price movement of a cryptocurrency based on historical data. The methods are as follows:

1. **SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors):** SARIMAX is a time-series forecasting model that incorporates exogenous variables to improve prediction accuracy. We use historical price features such as returns, volume, and volatility to predict future price movements.
2. **Hybrid SARIMAX and XGBoost Model:** This hybrid approach combines the strengths of SARIMAX for time-series prediction and XGBoost for capturing complex nonlinear relationships. SARIMAX outputs are used as features in XGBoost, creating a pipeline that leverages both statistical and machine learning methods.
3. **Spark Triple Ensemble Model:** This ensemble model integrates:
   - Gradient Boosted Trees (GBT)
   - Random Forest (RF)
   - Long Short-Term Memory (LSTM) Neural Networks

   The ensemble model is implemented in Apache Spark for scalability, making it suitable for high-frequency, large-scale data processing.

3.2 SARIMAX

The SARIMAX model is trained using historical time-series data, with exogenous variables derived from engineered features. The following steps are performed:

1. **Technical Indicator Calculation:** We calculate various technical indicators, including:
   - **Returns:** Percentage change in closing price.
   - **RSI (Relative Strength Index):** Measures the magnitude of recent price changes to evaluate overbought or oversold conditions (Investopedia, n.d.(c)).
   - **MACD (Moving Average Convergence Divergence):** Captures momentum by calculating the difference between fast and slow exponential moving averages (Investopedia, n.d.(d)).
   - **Rolling Averages:** Includes the 5-period rolling mean of closing prices and volume.

- **Price Range:** Difference between high and low prices normalized by the opening price.
- **Time Features:** Hour of the day and day of the week to account for temporal effects.

2. **Missing Value Handling:** Missing values in the dataset are addressed using forward and backward filling methods. Any remaining missing values are replaced with zeros to ensure model compatibility.
3. **Feature Preparation:** Exogenous features are selected and standardized using a 'StandardScaler'. The selected features include technical indicators and time features.
4. **Model Training:** The SARIMAX model is trained with the following configuration:
   - **Order:** $(1, 0, 0)$ — a simple AutoRegressive model of order 1 is employed.
   - **Exogenous Variables:** The standardized features prepared in the earlier step are used as exogenous inputs.
   - **Stationarity:** The model allows non-stationary data by setting `enforce_stationarity=False`.

This approach ensures that the SARIMAX model captures both temporal dependencies and the impact of exogenous variables on the target variable. Predictions are made for each future timestamp using the fitted model, and directional movement (up or not up) is determined by comparing the predicted price with the current price.

3.3 Hybrid SARIMAX and XGBoost

The Hybrid SARIMAX and XGBoost model combines the strengths of statistical modeling and machine learning to predict the next-minute price direction. The process involves:

1. **Feature Engineering:**
   - **Returns:** Percentage change in closing price.
   - **Momentum Indicators:** RSI, MACD, and ATR are calculated using shifted historical data to avoid label leakage. ATR (Average True Range) measures market volatility by averaging the true range over a specified number of periods. The true range is calculated as the greatest of:
     (a) The current high minus the current low,
     (b) The absolute value of the current high minus the previous close,
     (c) The absolute value of the current low minus the previous close.

ATR is particularly useful for assessing volatility and risk management in trading strategies (Investopedia, n.d.(a)).

- **Volume-Based Indicators:** Volume ratio and moving average are used to capture short-term trading activity.
- **Time Features:** Hour of the day and day of the week are included to account for temporal effects.

2. **SARIMAX Model:**
   - SARIMAX is trained on the most recent 50,000 rows of the training dataset.
   - Exogenous features (e.g., momentum and volume indicators) are standardized and included in the model.
   - Hyperparameters such as (p, d, q) and seasonal orders are tuned using a grid search over predefined parameter ranges.
3. **XGBoost Model:**
   - XGBoost is trained on the entire training dataset with standardized features and the target variable.
   - The model's hyperparameters, including learning rate, maximum depth, number of estimators, and subsample ratio, are fine-tuned through grid search.
   - The tree-based model captures complex nonlinear relationships in the data.
4. **Fine-Tuning Process:**
   - **Grid Search:** Both SARIMAX and XGBoost models are fine-tuned using a grid search approach. For SARIMAX, combinations of (p, d, q) orders and seasonal orders are explored. For XGBoost, parameters such as learning rate, maximum depth, number of trees, and subsample ratios are evaluated.
   - **Validation Split:** A temporal validation split is used to ensure that the model generalizes to unseen future data. The training set uses the most recent data for SARIMAX and the entire dataset for XGBoost.
   - **Evaluation Metric:** The Macro-Averaged F1 Score is used to evaluate model performance during fine-tuning, ensuring balanced performance across classes.
   - **Ensemble Weight Optimization:** A weight parameter is introduced to balance the contributions of SARIMAX and XGBoost predictions. The optimal weight is selected based on validation performance.
5. **Hybrid Prediction:**
   - Predictions from SARIMAX are based on directional price changes derived from its forecast.
   - Predictions from XGBoost are derived from the model's output probabilities.

- The final hybrid prediction is calculated as:

  $$\text{Final Prediction} = w \cdot \text{SARIMAX Predictions} + (1 - w) \cdot \text{XGBoost Predictions}$$

  where w is the optimized ensemble weight.

The hybrid model effectively combines the temporal insights from SARIMAX with the nonlinear predictive power of XGBoost, resulting in a robust approach to predicting cryptocurrency price directions.

3.4 Spark Triple Ensemble Model

The Spark Triple Ensemble Model combines the strengths of Gradient Boosted Trees (GBT) and Random Forest (RF) to predict the next-minute price direction of cryptocurrency. While the model includes an LSTM network for sequential pattern recognition, its contribution is excluded (LSTM weight = 0) in the final ensemble due to suboptimal performance during validation. Apache Spark ensures scalability for high-frequency data by enabling distributed computation.

3.4.1 *Feature Engineering*

Key features are calculated using Apache Spark and include:

- **Technical Indicators:** RSI, MACD, and Bollinger Bands. Bollinger Bands consist of a moving average (middle band) and two standard deviation bands (upper and lower) around it. Bollinger Bands help measure market volatility and potential overbought or oversold conditions (Investopedia, n.d.(b)).
- **Volume Metrics:** Volume ratio, calculated as the ratio of the current volume to its moving average over 10 periods.
- **Returns:** Percentage change in closing prices.
- **Target Variable:** A binary target indicating whether the price moves up (1) or stays the same/moves down (0) in the next minute. This is derived by comparing the current price to the next minute's closing price.

3.4.2 *Training Procedure*

1. **Data Preprocessing:**
   - Data is partitioned for distributed processing using Apache Spark.
   - Features such as RSI, MACD, Bollinger Bands, and volume metrics are calculated in Spark using sliding windows to ensure no label leakage.

- Missing values are filled using a forward and backward fill approach, and remaining NaNs are replaced with zeros.

2. **GBT and RF Training:**
   - A pipeline is created with a `VectorAssembler` to assemble feature vectors, followed by `StandardScaler` for feature normalization.
   - GBTClassifier and RandomForestClassifier are trained using features such as returns, RSI, MACD, Bollinger Bands, and volume ratio.
   - Hyperparameters such as maximum depth, number of trees, and learning rate are tuned for optimal performance.
3. **LSTM Exclusion:**
   - Although the LSTM network was trained with sequential data using a sliding window approach (sequence length of 10), its predictions were found to contribute negligible improvement to the overall ensemble performance.
   - Consequently, the LSTM predictions were excluded (LSTM weight = 0) from the final ensemble computation.

3.4.3 *Prediction*

- Predictions from GBT and RF are generated as probabilities for the positive class.
- The final ensemble prediction is computed as:

$$\text{Final Prediction} = 0.5 \cdot (\text{GBT Predictions}) + 0.5 \cdot (\text{RF Predictions})$$

This approach excludes LSTM predictions to optimize overall performance.

By focusing on tree-based models, the Spark Triple Ensemble Model effectively balances accuracy and scalability, leveraging GBT and RF for robust feature learning in high-frequency cryptocurrency price prediction.

4 EVALUATION

4.1 Evaluation Metric

The model's performance is evaluated using the **Macro-Averaged F1 Score**, which assesses the balance between precision and recall for each class (up and not up).

4.1.1 *Formula*

The F1 score for each class c is calculated as:

$$F1_c = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The Macro-Averaged F1 Score is:

$$F1_{macro} = \frac{1}{C} \sum_{c=1}^{C} F1_c$$

Where $C = 2$, representing the two classes (GenAI.Labs, 2024).

5 RESULTS

The performance of the three models was evaluated using the Macro-Averaged F1 Score, ensuring balanced performance across both classes (price moves up or does not move up). The results are summarized in Table 1.

Table 1—Performance of Models (Macro-Averaged F1 Score)

Model	**Macro-Averaged F1 Score**
SARIMAX	0.49092
Hybrid SARIMAX and XGBoost	0.49855
Spark Triple Ensemble Model	0.49654

5.1 Discussion

The Hybrid SARIMAX and XGBoost model achieved the highest Macro-Averaged F1 Score of 0.49855, outperforming the other approaches. The combination of SARIMAX for capturing temporal patterns and XGBoost for modeling nonlinear relationships proved to be the most effective.

The Spark Triple Ensemble Model, which integrates GBT, RF, and LSTM, achieved a Macro-Averaged F1 Score of 0.49654. While close to the hybrid model's performance, further tuning or more data might enhance its predictive capability.

SARIMAX, despite its simplicity, achieved a competitive Macro-Averaged F1 Score of 0.49092, but it struggled to model the complex nonlinear relationships inherent in the data.

5.2 Conclusion

The results demonstrate that the Hybrid SARIMAX and XGBoost model is the most effective for predicting cryptocurrency price movement in the next minute. Future research could focus on refining ensemble methods and incorporating additional features to further enhance performance.

6 REFERENCES

1. GenAI.Labs (2024). *Directional Forecasting Cryptocurrencies*. https://kaggle.com/competitions/directional-forecasting-cryptocurrencies. Kaggle.
2. Investopedia (n.d.[a]). *Average True Range (ATR) Indicator Explained With Formula*. https://www.investopedia.com/terms/a/atr.asp. Accessed: 2024-01-05.
3. Investopedia (n.d.[b]). *Bollinger Bands Definition and Uses*. https://www.investopedia.com/terms/b/bollingerbands.asp. Accessed: 2024-01-05.
4. Investopedia (n.d.[c]). "Relative Strength Index (RSI) Indicator Explained With Formula". In: *Investopedia Technical Analysis Resources*. Online: Investopedia. URL: https://www.investopedia.com/terms/r/rsi.asp.
5. Investopedia (n.d.[d]). "What Is MACD?" In: *Investopedia Technical Analysis Resources*. Online: Investopedia. URL: https://www.investopedia.com/terms/m/macd.asp.