**NAIROBI SECURITIES EXCHANGE:AI DRIVEN TIME SERIES EXPLORATION**

**CRISP-DM**

**AUTHORS:**

Brenda Mutai - Group Lead

Justin Mbugua

Sharon Momanyi

Stephen Munyiala

# **Table of Contents**

[**Table of Contents** 2](#_Toc197303200)

[**CHAPTER 1. INTRODUCTION.** 3](#_Toc197303201)

[**CHAPTER 2. BUSINESS UNDERSTANDING.** 4](#_Toc197303202)

[**CHAPTER 3. DATA UNDERSTANDING.** 6](#_Toc197303203)

[**CHAPTER 4. DATA PREPARATION**. 7](#_Toc197303204)

[**CHAPTER 5. EXPLORATORY DATA ANALYSIS.** 10](#_Toc197303205)

[**CHAPTER 6. FEATURE ENGINEERING** 22](#_Toc197303206)

[**CHAPTER 7. MODELING** 24](#_Toc197303207)

[**CHAPTER 8. MODEL EVALUATION.** 30](#_Toc197303208)

[**CHAPTER 9. DEPLOYMENT** 32](#_Toc197303209)

[**CHAPTER 10. CONCLUSIONS AND RECOMMENDATIONS.** 33](#_Toc197303210)

# **CHAPTER 1. INTRODUCTION.**

This project focuses on analyzing stock price movements for all publicly listed companies on the Nairobi Securities Exchange (NSE) over the years 2023 and 2024. By leveraging daily trading data—including stock prices, trading volume, sector classifications, and other relevant financial indicators—the study aims to uncover patterns and trends that influence market behavior.

Through time series analysis, statistical modeling, and machine learning techniques, the project will evaluate historical price movements to detect seasonal trends, price volatility, and sector-wide correlations. Additionally, outlier detection will be applied to assess extreme market events, ensuring a comprehensive understanding of price fluctuations without disregarding natural market dynamics.

A key focus of the project is to aid investors, financial analysts, policy makers and researchers in identifying potential opportunities and risks within the Kenyan stock market. By generating data-driven insights, the study will help stakeholders make informed decisions regarding portfolio diversification, market entry strategies, and investment planning.

Moreover, forecasting models—including ARIMA, XGBoost, and LSTM—will be deployed to predict future stock price movements, offering a quantitative basis for analyzing possible market directions. The study will also investigate trading volumes to understand liquidity trends, identifying stocks that exhibit high investor activity and those that remain relatively illiquid.

Ultimately, this project serves as a comprehensive exploration of Kenya’s stock market dynamics, aiming to enhance financial knowledge, support investment strategies, and contribute valuable insights to market participants.

# **CHAPTER 2. BUSINESS UNDERSTANDING.**

**2.1 Problem Statement.**

Investors and financial institutions operating on the Nairobi Securities Exchange (NSE) rely on precise information to make strategic decisions regarding stock trades. By forecasting future stock prices and identifying market trends, they can optimize investment strategies, improve portfolio management, and effectively mitigate risks. Such insights empower traders to seize profitable opportunities, avoid potential losses, and enhance overall financial stability.

Additionally, predictive models enable investors to understand market behavior better, adapt to changing conditions, and maintain a competitive edge. With tools for analyzing historical trading data, stakeholders can uncover patterns, assess the impact of external factors like economic policies or global market shifts, and make data-driven choices that align with their financial goals. These advancements are essential for thriving in a dynamic stock market environment like the NSE.

**2.2 Objectives**

The overall aim of this study is to develop time series models to forecast future stock prices.

Specifically, the study seeks to:

1. provide insights into which stocks might perform well based on historical trends and predictive models, which will allow for more informed decision-making.
2. Offer short-term predictions of stock prices or trends to support timely buy/sell decisions, potentially improving their profitability.
3. Democratize Access to NSE Analytics thus making it easier for ordinary people to track and engage with stocks and securities.

**2.3 Stakeholders**

**The study focuses and addresses the following stakeholders:**

1. Retail Investors: Use forecasts for making buy/sell decisions.
2. Financial Analysts: Integrate model outputs into broader financial analysis workflows.
3. Portfolio Managers: Optimize asset allocation strategies.

**2.4 Data Understanding.**

The data is a compilation of historical daily stock market price data relates to the Kenyan Nairobi Securities Exchange (NSE) for the years 2023 and 2024. It was sourced from *https://data.mendeley.com/(Kenya Nairobi Securities Exchange (NSE) All Stocks Prices 2023-2024)*

The data was scrapped from a publicly accessible website -<http://live.mystocks.co.ke/> - licensed by NSE by exporting raw web data to spreadsheets, then cleaned up to a final CSV.

Each stock data row has 13 data columns (1) Date (2) Stock Code (3) Stock Name (4) 12-month Low price (5)12-month High price (6) Day's Low price (7) Day's High price (8) Day's Final Price (9) Previous traded price (10) Change in price value (11) Change in price % (12) Volume traded (13) Adjusted price. One additional CSV file is also provided to show the stocks market sector, with three (3) columns as: (1) Market sector (2) Stock Code (3) Stock Name.

**2.5 Metric of success.**

1. Model Performance: Accuracy of stock price forecasts measured by metrics like Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) or Mean Squared Error(MSE).
2. Insight Utility: The relevance and actionability of the identified market trends, sector performance analyses, and trading patterns for stakeholder decision-making (Qualitative).
3. Completeness: Successful analysis covering the specified timeframe (2023-2024) and scope (all listed companies, key data points).

# **CHAPTER 3. DATA UNDERSTANDING.**

****

The dataset is structured in a tabular format with 14 columns and 35,393 rows.

Missing Data: "Previous" and "Adjusted Price" fields are often empty.

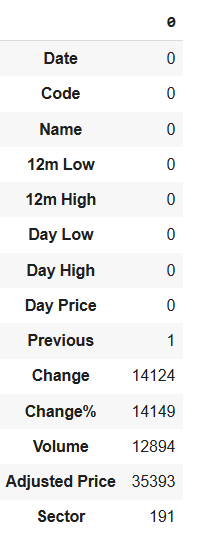
It contains variable such as:

* + **Date**: Trading date (e.g., "03-Jan-2023").
  + **Code**: Stock ticker symbol (e.g., "EGAD", "KUKZ").
  + **Name**: Company or index name (e.g., "Eaagads Ltd", "NSE 25-Share Index").
  + **12m Low/High**: 12-month lowest and highest prices.
  + **Day Low/High**: Daily trading range.
  + **Day Price**: Closing price for the day.
  + **Previous**: Previous day's closing price (missing for some entries).
  + **Change/Change%**: Absolute and percentage change from the previous day.
  + **Volume**: Trading volume (some entries missing or zero).
  + **Adjusted Price**: Not populated in the sample.
  + **Sector**: Classification (e.g., "Agricultural").

# **CHAPTER 4. DATA PREPARATION**.

The data appears to have several missing values. Next we check for the dataframe information to find out how many missing values are in each column. This is important when deciding how we will deal with the missing values.

On initial inspection, we observed the data to be clean but after removing the “-” it generated the following features.



The dataset contains missing values across several columns. The 'Previous' column has 1 missing entry. 'Change' and 'Change%' have a significant number of missing values, with 14,124 and 14,149 respectively. 'Volume' also has a substantial number of missing entries (12,894). The 'Adjusted Price' column exhibits the highest number of missing values at 35,393. Finally, the 'Sector' column has 191 missing values. All other columns ('Date', 'Code', 'Name', '12m Low', '12m High', 'Day Low', 'Day High', 'Day Price') have no missing values.

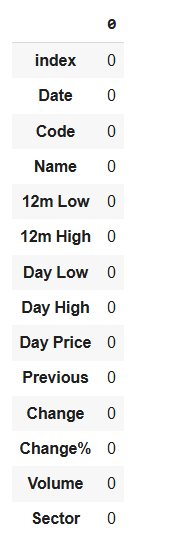
To prepare the numeric columns for analysis, we first need to clean them by removing any commas and spaces before converting them to numeric.

We also have to change the date from string format to datetime. Since the date in the 2023 dataset has the day first while the date in the 2024 dataset has the month first, we used a function that accounts for this.

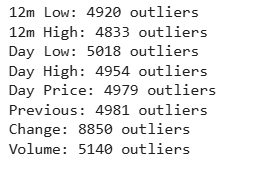
We then proceeded to do the following to the data.

* Drop the **Adjusted Price** column because it doesn't contain any values.
* Replace the missing values in the **Sector** column with 'Unknown'.
* Drop missing values from the **Previous** column which has a very small number of missing values.
* Fill the missing values in the **Change** and **Change%** columns by calculating the values.
* Replace the missing values in the **Volume** column with the median for each stock.

After cleaning this was the final generated clean data.



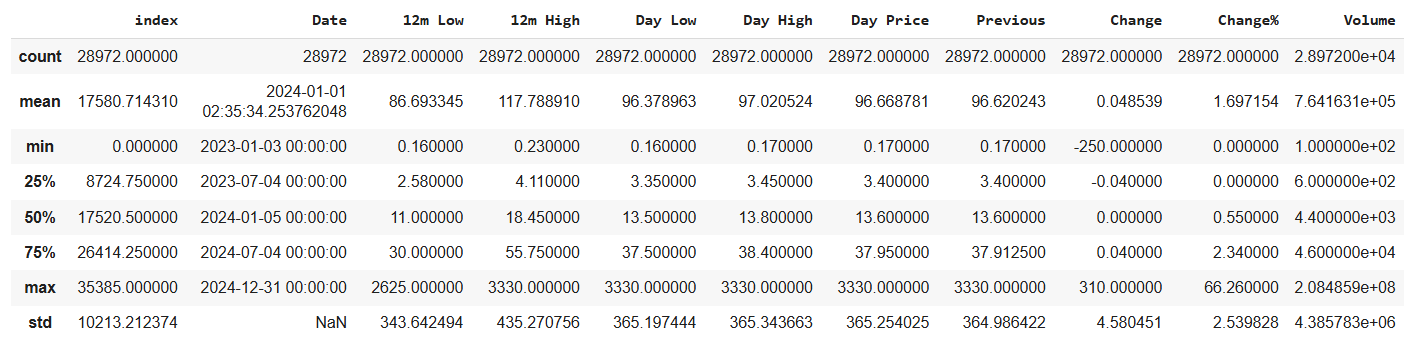
We analyzed the dataset for outliers to gain deeper insights, but given the inherent volatility of financial data, we have chosen to retain them. Most outliers reflect natural market fluctuations rather than errors, making them valuable for understanding real-world trends.



Several numerical columns contain a significant number of outliers. 'Day Low' has the highest number of outliers (5018), followed closely by 'Day Price' (4979) and 'Previous' (4981). '12m Low' (4920), '12m High' (4833), 'Day High' (4954), and 'Volume' (5140) also exhibit a substantial number of outliers. The 'Change' column shows the most outliers with 8850. This indicates potential data anomalies or high variability in these price and volume metrics.

# **CHAPTER 5. EXPLORATORY DATA ANALYSIS.**

We checked the data description to identify the numerical columns.



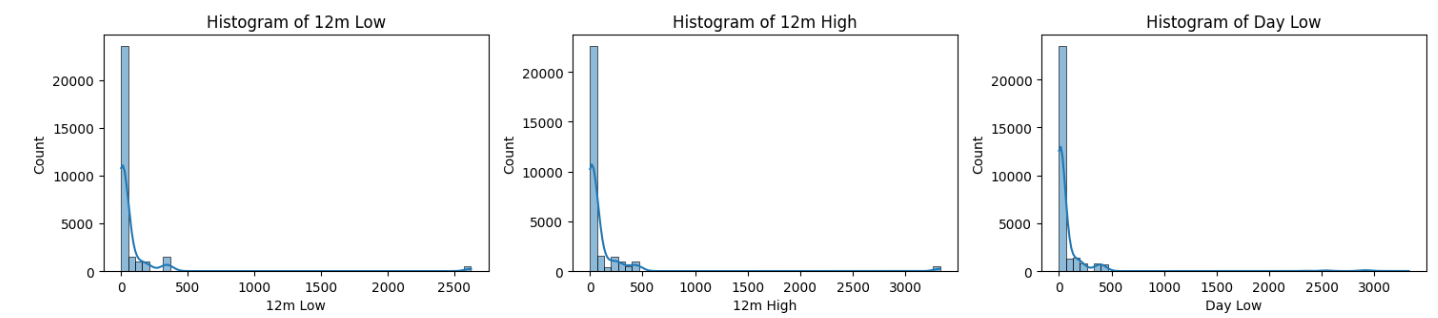
**Price Metrics ('12m Low', '12m High', 'Day Low', 'Day High', 'Day Price', 'Previous'):** Exhibit similar central tendencies (means around 86-97) and ranges, with maximum values reaching around 3330. The standard deviations are also comparable, indicating similar levels of price volatility.

**Change:** The average price change is slightly positive (0.0485), with a wide range from -250 to 310.

**Change%:** Shows a positive average percentage change (1.697), with a range from 0% to a maximum of 66.26%.

**Volume:** The average trading volume is approximately 7.64 million, with a significant range from 100 to over 20 billion. The high standard deviation indicates substantial variability in trading volume.

**5.1 Univariate Analysis**



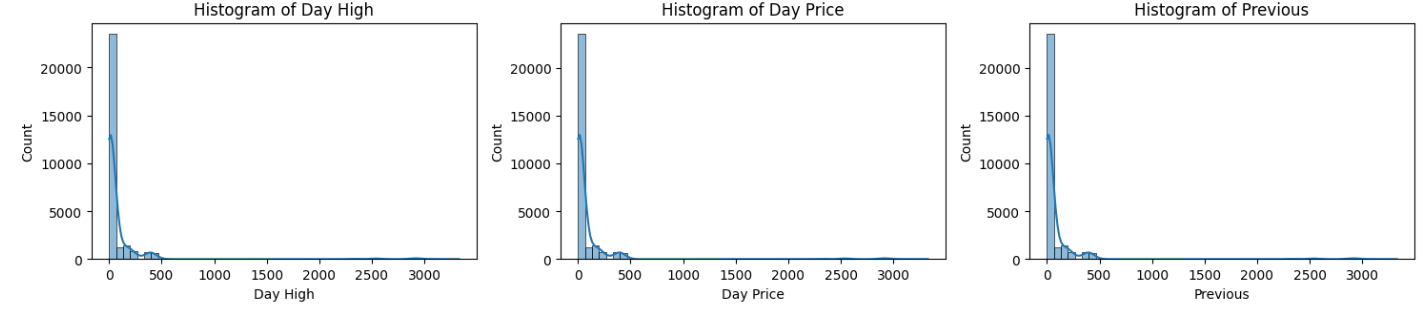
The histograms for '12m Low', '12m High', and 'Day Low' exhibit a strong positive skew.

**Majority of values are concentrated at the lower end of the price range.** A large number of observations have relatively low 12-month lows, 12-month highs, and daily low prices.

**A long tail extends towards higher values.** There are fewer observations with significantly higher prices, creating the right-skewed distribution.

**Potential for outliers on the higher end.** The extended tail suggests the presence of some extreme high values in these price features.

In essence, these price variables are not normally distributed, with a bias towards lower values and occasional spikes into higher price ranges.

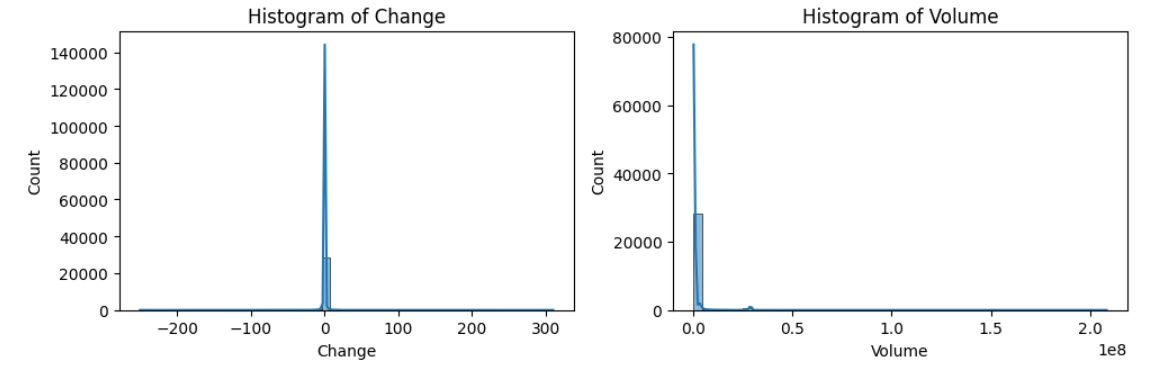


**Lower price ranges are more frequent:** The majority of the data points for the day's high, the closing price, and the previous day's closing price are clustered at the lower end of the price spectrum.

**Higher prices are less common:** The distributions have long tails extending towards higher values, signifying a smaller number of trading instances with significantly higher prices.

**Potential for right-skewed modeling:** Statistical models for these features might need to account for this non-normal distribution.

The consistent skew across these related price metrics suggests a common characteristic of the trading data where lower-priced assets or lower price movements are more prevalent.



**Histogram of Change:** This distribution is heavily concentrated around zero, indicating that the most frequent price change is no change. There are smaller frequencies of both positive and negative changes, with the range extending from approximately -250 to +300. The distribution appears somewhat symmetrical around zero, although the central spike dominates.

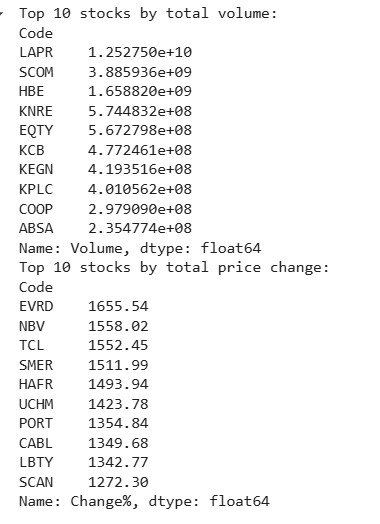
**Histogram of Volume:** This distribution exhibits a strong positive skew. The vast majority of trading volumes are low, clustered near zero. There is a long tail extending to the right, indicating that while infrequent, there are instances of extremely high trading volumes, reaching up to 2 x 10^8. This suggests that most trading days involve relatively low volume, with occasional periods of very high activity.

**5.2 Bivariate Analysis**

1. Top 10 Stocks by Volume and Price Change

**Top 10 Stocks by Total Volume:** LAPR leads significantly with a total volume of approximately 12.5 billion. SCOM follows as a distant second with roughly 3.9 billion. HBE, KNRE, EQTY, KCB, KEGN, KPLC, COOP, and ABSA also show substantial trading volumes, ranging from approximately 235 million to 574 million.

**Top 10 Stocks by Total Price Change (%):** EVRD exhibits the highest positive price change at 1655.54%. NBV and TCL also show very significant percentage gains at 1558.02% and 1552.45%, respectively. SMER, HAFR, UCHM, PORT, CABL, LBTY, and SCAN complete the top ten, all demonstrating substantial positive percentage price changes above 1270%.



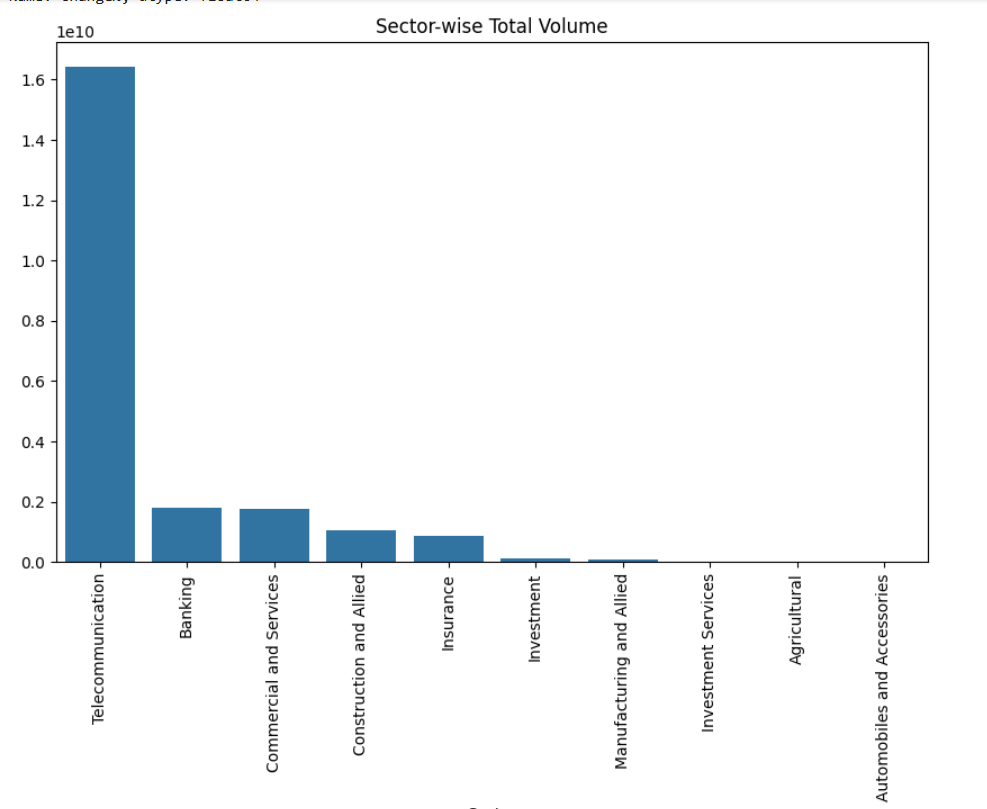
2. Total volume by sector

The **Telecommunication** sector exhibits the highest total trading volume, reaching approximately 16.4 billion. Following this, the **Banking** sector shows a significant volume of around 1.81 billion, closely trailed by **Commercial and Services** at approximately 1.76 billion.

Other sectors with notable total trading volumes include:

* **Construction and Allied:** ~1.05 billion
* **Insurance:** ~872 million
* **Investment:** ~134 million
* **Manufacturing and Allied:** ~73.6 million
* **Investment Services:** ~22.1 million
* **Agricultural:** ~7.35 million
* **Automobiles and Accessories:** ~463,900

The Telecommunication sector dominates in terms of overall trading activity, with a volume significantly higher than any other sector. The Banking and Commercial & Services sectors also show substantial trading interest. The remaining sectors have considerably lower total trading volumes in comparison.



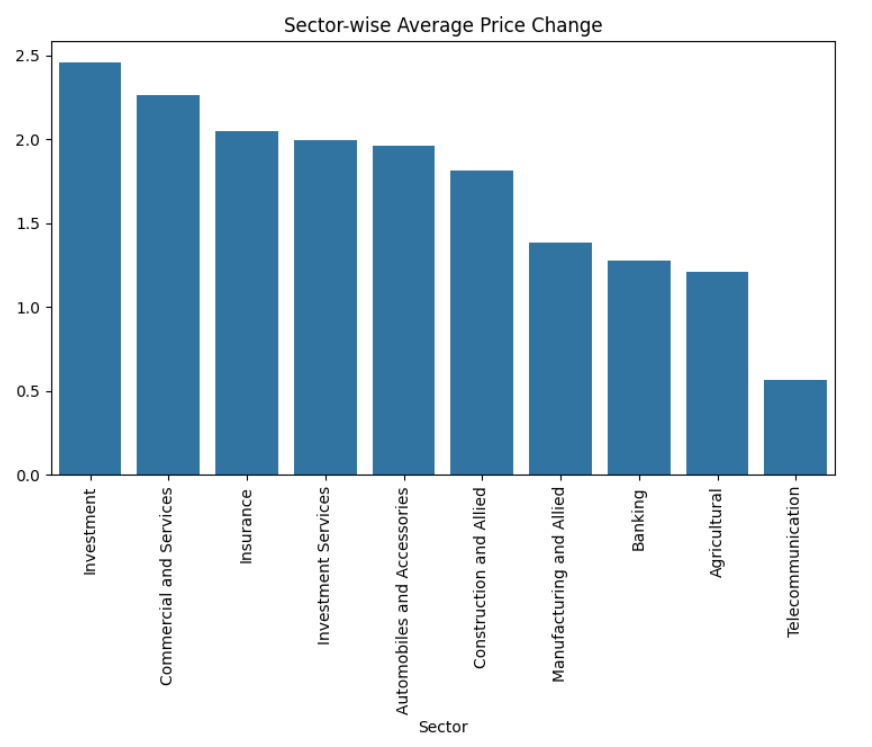
3. Sector wise average price change

The **Investment** sector exhibits the highest average percentage price change at approximately 2.46%. Following closely are **Commercial and Services** (around 2.27%) and **Insurance** (about 2.05%).

Other sectors with positive average percentage price changes include:

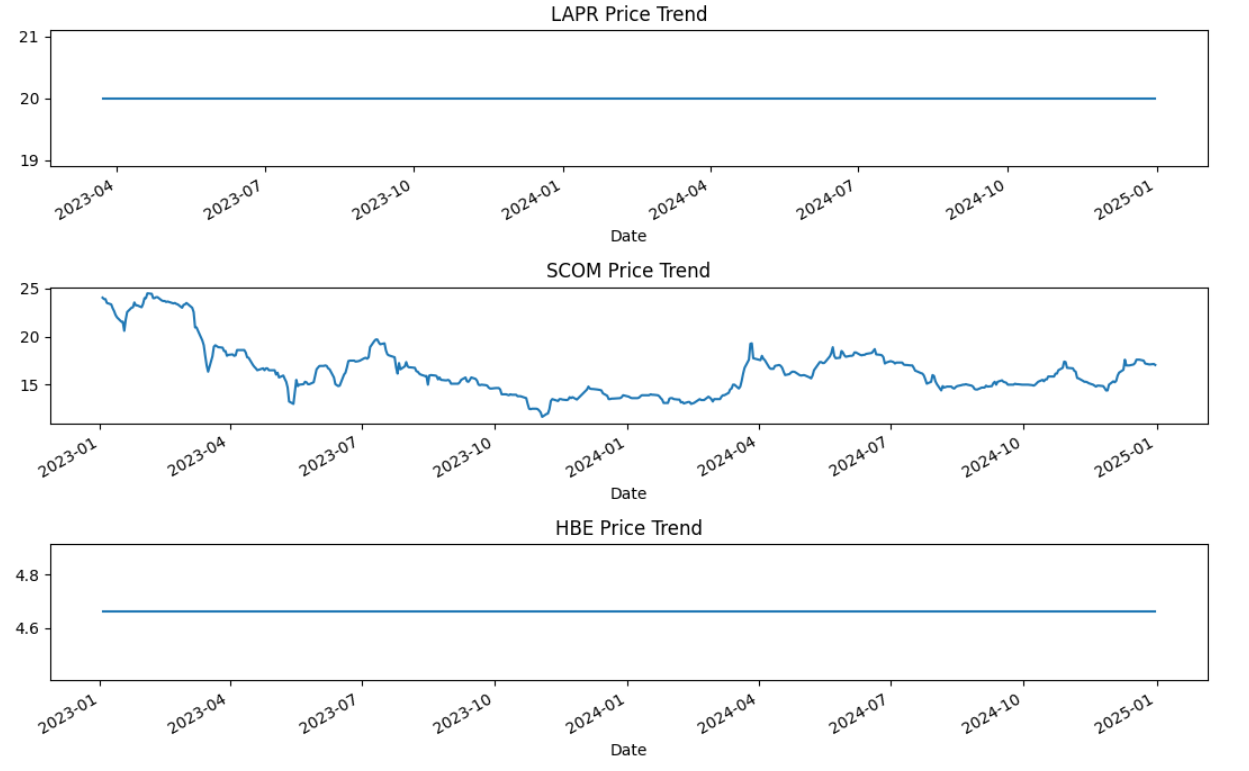
* **Investment Services:** ~2.00%
* **Automobiles and Accessories:** ~1.96%
* **Construction and Allied:** ~1.81%
* **Manufacturing and Allied:** ~1.38%
* **Banking:** ~1.28%
* **Agricultural:** ~1.21%
* **Telecommunication:** ~0.56%

The Investment sector shows the most significant average upward price movement. Interestingly, the Telecommunication sector, which had the highest total trading volume, displays the lowest average percentage price change among all listed sectors. This suggests that high trading volume doesn't necessarily translate to large average price increases.



1. Price and Volume Trends Over Time

Price trend

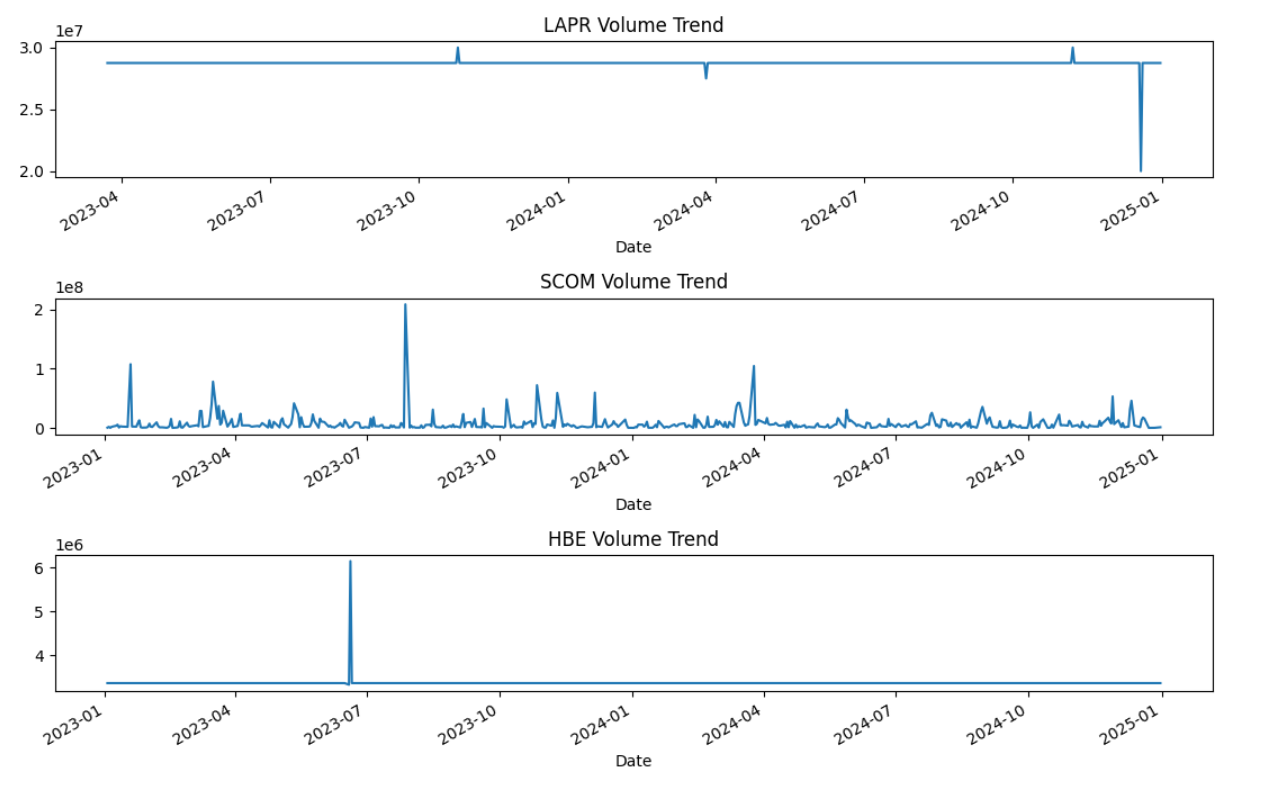


**LAPR Price Trend:** The price of LAPR appears to have remained constant at approximately 20 over the observed period from early 2023 to early 2025. There is no visible fluctuation in its price.

**SCOM Price Trend:** SCOM's price shows more volatility. Starting around 24 in early 2023, it experienced a significant decline throughout 2023, reaching a low of around 13. The price fluctuated between 13 and 20 during 2024 and shows a slight upward trend towards the beginning of 2025, closing around 17.

**HBE Price Trend:** Similar to LAPR, the price of HBE has remained relatively stable at approximately 4.65 throughout the observed period. There are no significant price movements.

1. Volume trend

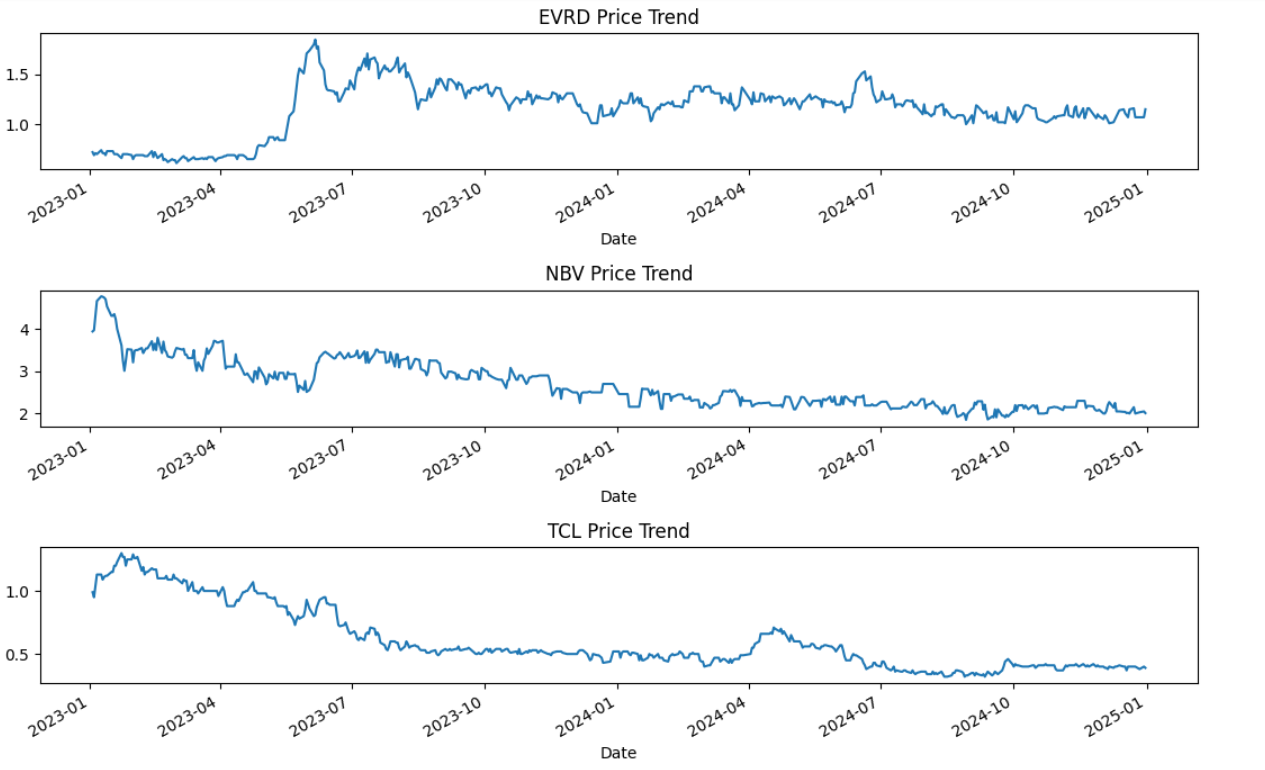


**LAPR Volume Trend:** The trading volume for LAPR has been consistently high, hovering around 2.8 x 10^7 throughout the period. There are a few noticeable spikes, particularly around late 2024 and early 2025, indicating periods of increased trading activity.

**SCOM Volume Trend:** SCOM exhibits more variable trading volume. While generally lower than LAPR, with most days showing volume below 2 x 10^7, there are significant spikes in volume at various points. Notably, there's a large spike in mid-2023, and several other smaller peaks throughout the observed period, suggesting intermittent periods of high trading interest.

**HBE Volume Trend:** The trading volume for HBE has been relatively low and stable, around 3.5 x 10^5 for most of the period. There is one very prominent spike in volume around mid-2023, reaching significantly higher than its typical trading volume, before returning to its baseline.

1. Price change trend

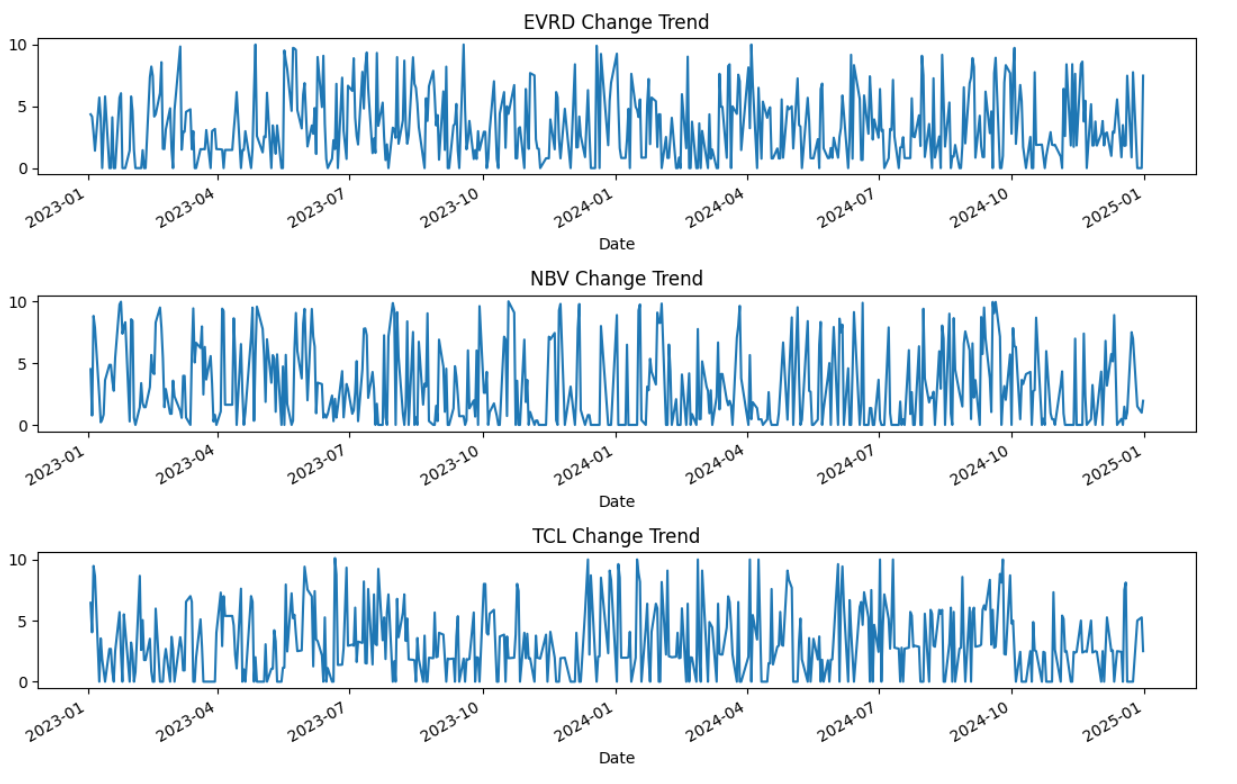


**EVRD Price Trend:** EVRD's price started around 0.8 in early 2023, remained relatively stable until mid-2023, then experienced a sharp increase, peaking above 1.6. Following this peak, the price gradually declined and fluctuated, ending around 1.1 in early 2025.

**NBV Price Trend:** NBV's price began around 4.5 in early 2023 and showed a declining trend throughout the observed period. There were some fluctuations, but the overall trajectory was downward, reaching approximately 2.2 by early 2025.

**TCL Price Trend:** TCL's price started around 1.1 in early 2023 and generally decreased over time. There were some minor rallies, but the dominant trend was downward, reaching a low of around 0.35 in late 2024 before showing a slight upward movement to around 0.4 by early 2025.

1. Volume change trend



EVRD:

* Displays consistent volume fluctuations throughout the period.
* No clear seasonality, but frequent spikes indicate periods of heightened trading interest.
* Suggests relatively stable market attention with periodic volatility.

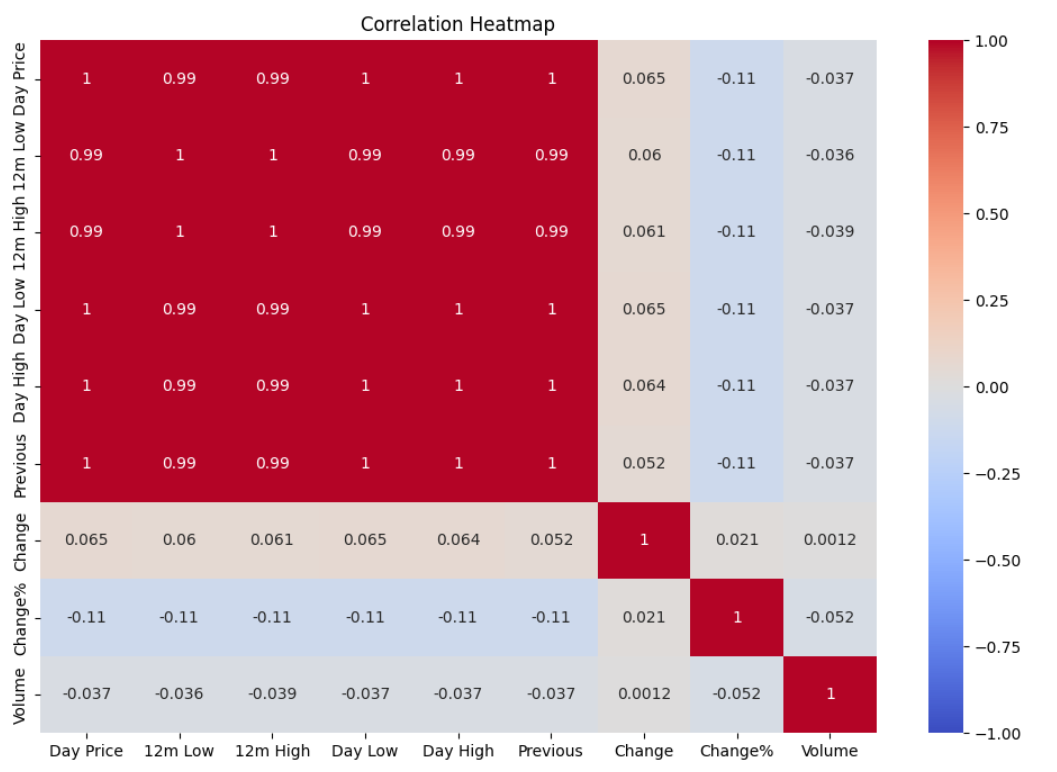
NBV:

* Shows higher volatility in early 2023, with activity tapering off mid-2023.
* Renewed volume bursts appear in late 2023 and Q4 2024, possibly due to company events or market news.
* Volume trend is irregular and event-driven.

TCL:

* Exhibits decreasing volume change over time.
* Highest activity was observed in early 2023, followed by a steady decline.
* Indicates waning investor interest or reduced market activity in this stock.

**5.3 Multivariate analysis**



**Strong Positive Correlations (Near 1):**

'Day Price', '12m Low', '12m High', 'Day Low', 'Day High', and 'Previous' are all very strongly positively correlated with each other (correlation coefficients around 0.99 to 1.00). This indicates that these price-related features move very closely together. When one of these prices is high, the others are also likely to be high, and vice versa.

**Weak Positive Correlation:**

'Change' shows a weak positive correlation with 'Day Price', '12m Low', '12m High', 'Day Low', 'Day High', and 'Previous' (correlation coefficients around 0.05 to 0.07). This suggests a slight tendency for larger price changes to occur when the overall price levels are higher, but the relationship is not strong.

**Weak Negative Correlation:**

'Change%' exhibits a weak negative correlation with 'Day Price', '12m Low', '12m High', 'Day Low', 'Day High', and 'Previous' (correlation coefficients around -0.11). This implies a slight tendency for larger percentage price changes to occur when the overall price levels are lower, but again, the relationship is weak.

**Very Weak Correlation (Near 0):**

'Volume' shows a very weak correlation (close to zero) with all the price-related features and 'Change'. There is a slightly stronger negative correlation (-0.052) between 'Volume' and 'Change%'. This suggests that trading volume is largely independent of the price levels and the absolute price change, but there's a very minor tendency for higher percentage changes to be associated with slightly lower volume.

**Moderate Positive Correlation:**

'Change' and 'Change%' have a moderate positive correlation (around 0.21), suggesting that larger absolute price changes tend to be associated with larger percentage price changes, as expected.

# **CHAPTER 6. FEATURE ENGINEERING**

* 1. Moving Averages (SMA and EMA)

We calculate 10-day and 50-day Simple Moving Averages(SMA), as well as the 10-day and 50-day Exponential Moving Averages(EMA), to highlight trends.

**SMA\_10** and **SMA\_50**: Simple moving averages over 10 and 50 trading days.  
**EMA\_10** and **EMA\_50**: Exponential moving averages over 10 and 50 days (more weight to recent prices).

These indicators help smooth out price fluctuations and identify overall market trends, making them essential for detecting potential reversals or continuations in stock movement. While SMA assigns equal weight to all past prices, EMA prioritizes recent price data, making it more reactive to short-term shifts in market conditions.

Moving averages are used for **trend identification, entry/exit points**, and **confirmation of breakout patterns**, helping traders refine their decisions.

* 1. Relative Strength Index (RSI).

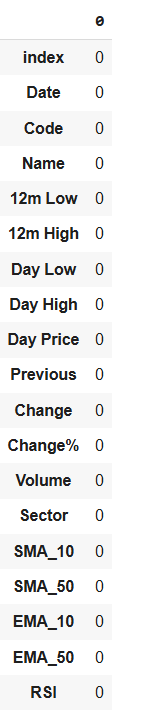
The RSI measures the speed and change of price movements. Values above 70 indicate that a stock is overbought while values below 30 indicate that it's oversold.

The Relative Strength Index (RSI) is a momentum indicator that helps assess the strength of price trends, offering insights into potential reversals or continuations in stock movement.

RSI is calculated based on average gains and losses over a specified period (typically 14 days). Readings above 70 suggest overbought conditions, meaning the stock might be due for a correction, while readings below 30 indicate oversold conditions, suggesting potential buying opportunities.

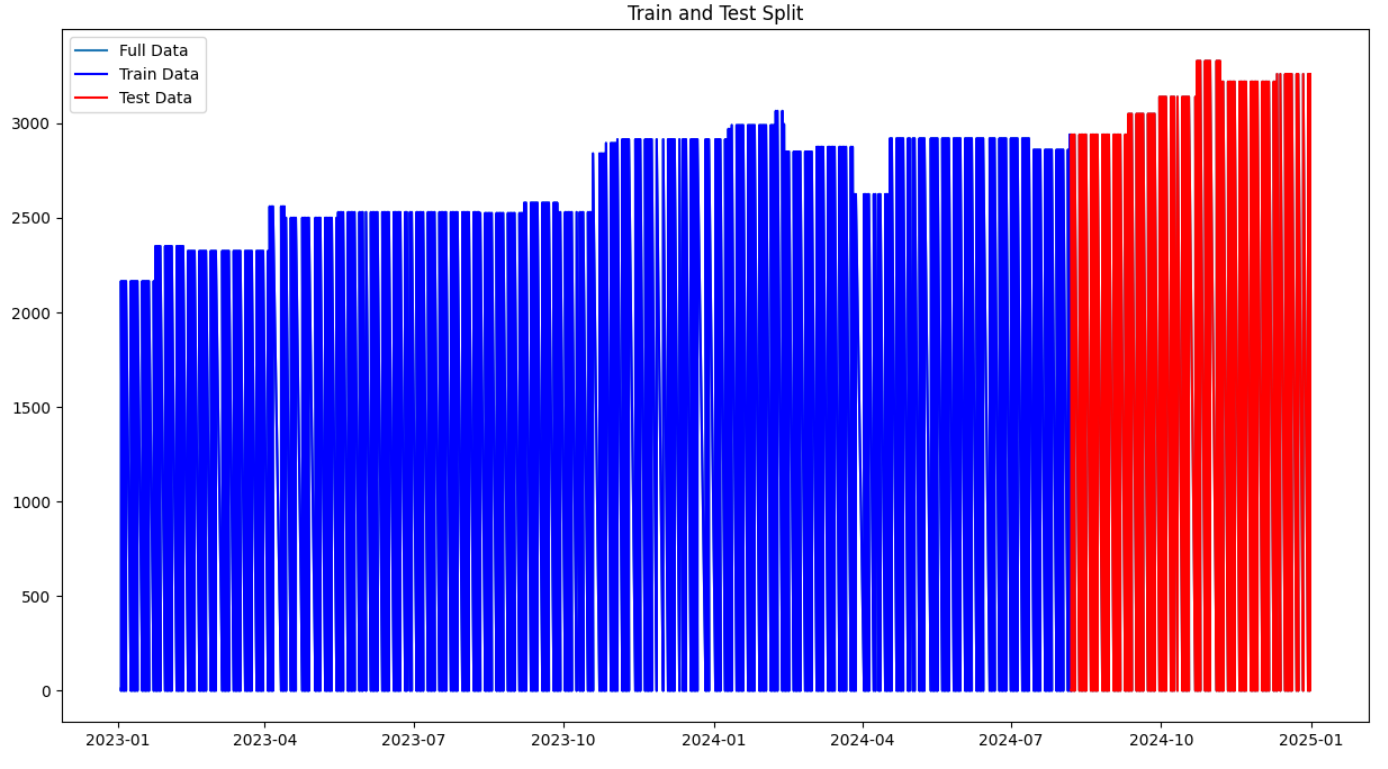
When price movements contradict RSI trends, it may signal trend reversals—for instance, if prices hit new highs while RSI forms lower highs, it could indicate weakening momentum.

This is how the data set looks with the new features.



# **CHAPTER 7. MODELING**

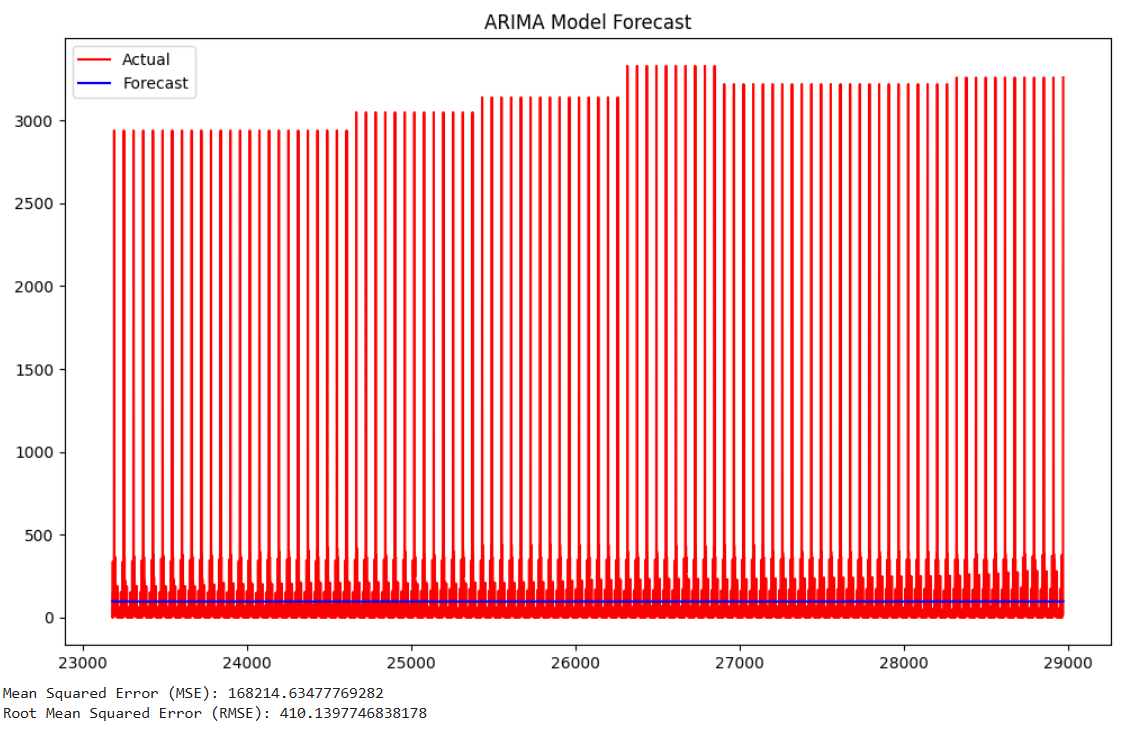
* 1. Train test split.



The bar plot illustrates the division of the dataset into training and testing sets over time.

* **Full Data:** Represented by a thin teal line, showing the complete time series data.
* **Train Data:** Displayed in blue bars, comprising the majority of the dataset and spanning from the beginning of the observed period (early 2023) up to approximately July 2024.
* **Test Data:** Shown in red bars, representing the portion of the data reserved for evaluating the model. This segment starts around July 2024 and extends to the end of the observed period (early 2025).

**7.1 ARIMA (AutoRegressive Integrated Moving Average )**

****

Actual Values (Red): Show a fluctuating pattern with a generally increasing trend over the depicted timeframe. There are consistent, high-frequency oscillations within the overall trend.

Forecasted Values (Blue): Present a relatively flat line at a low value, showing minimal variation and failing to capture the increasing trend and the oscillations present in the actual data.

**Mean Squared Error (MSE):** 168,214,634,776,9282. This is a very high value, indicating a significant discrepancy between the actual and forecasted values.

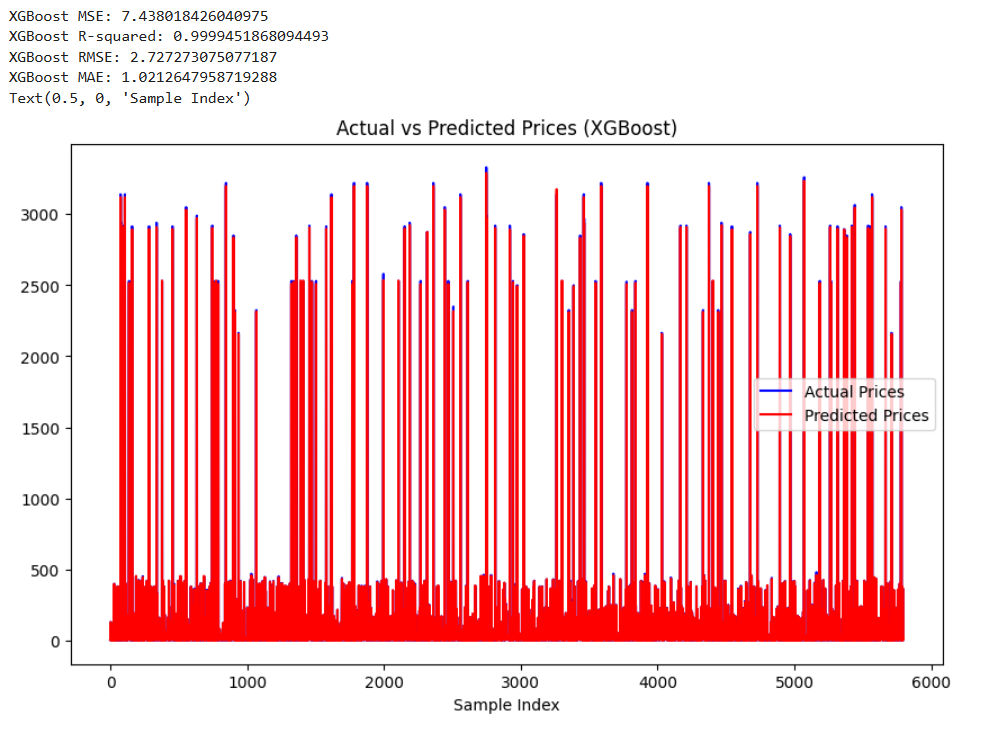
**Root Mean Squared Error (RMSE):** 410,139,774.6838178. The RMSE, being the square root of the MSE, is also very high, further highlighting the large average error in the model's predictions.

The ARIMA model, in this instance, appears to be a poor fit for the data. It fails to capture both the underlying increasing trend and the significant cyclical patterns present in the actual values.

The extremely high MSE and RMSE values quantitatively confirm the poor performance of the model in forecasting this particular time series.

Further investigation into different model parameters, other time series models, or the inclusion of exogenous variables might be necessary to achieve a more accurate forecast.

**7.2 XGBoost**



**XGBoost MSE:** 7.438018426040975. This is a very low Mean Squared Error, suggesting that the model's predictions have a small average squared difference from the actual values.

**XGBoost R-squared:** 0.9999451868094493. The R-squared value is extremely close to 1, indicating that the model explains approximately 99.99% of the variance in the actual prices. This signifies an exceptionally good fit.

The XGBoost model demonstrates outstanding predictive performance on this dataset.

The high R-squared value and low MSE indicate that the model accurately captures the relationship between the features and the target variable.

The tight clustering of points along the diagonal in the scatter plot visually confirms the model's high accuracy in predicting the prices.

* 1. Cross validation

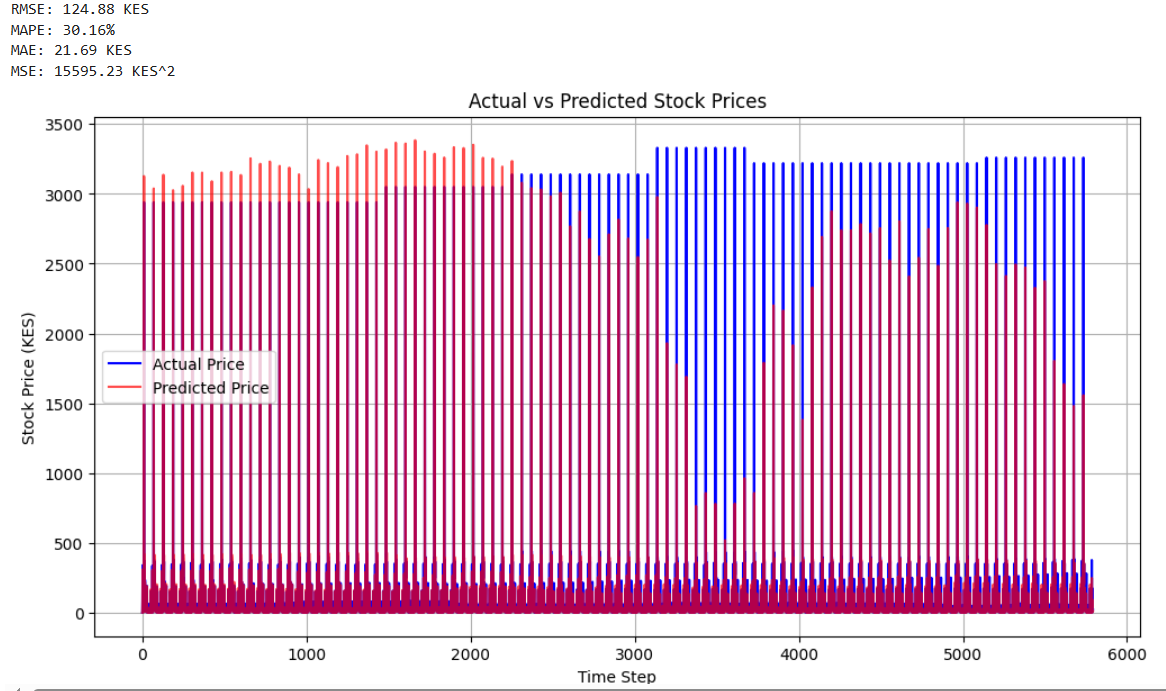


The cross-validation scores are consistently very high across all folds, ranging from approximately 0.9937 to 0.9999. This indicates that the model's performance is robust and generalizes well across different subsets of the training data.

The mean cross-validation score of approximately 0.9979 suggests that, on average, the model explains about 99.79% of the variance in the target variable during the cross-validation process.

This high mean score further reinforces the conclusion that the model (likely the XGBoost model from the previous image, given the excellent performance) exhibits strong predictive capabilities and good generalization ability.

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



**Actual Price (Blue):** Exhibits a fluctuating pattern with an overall upward trend and distinct cyclical variations.

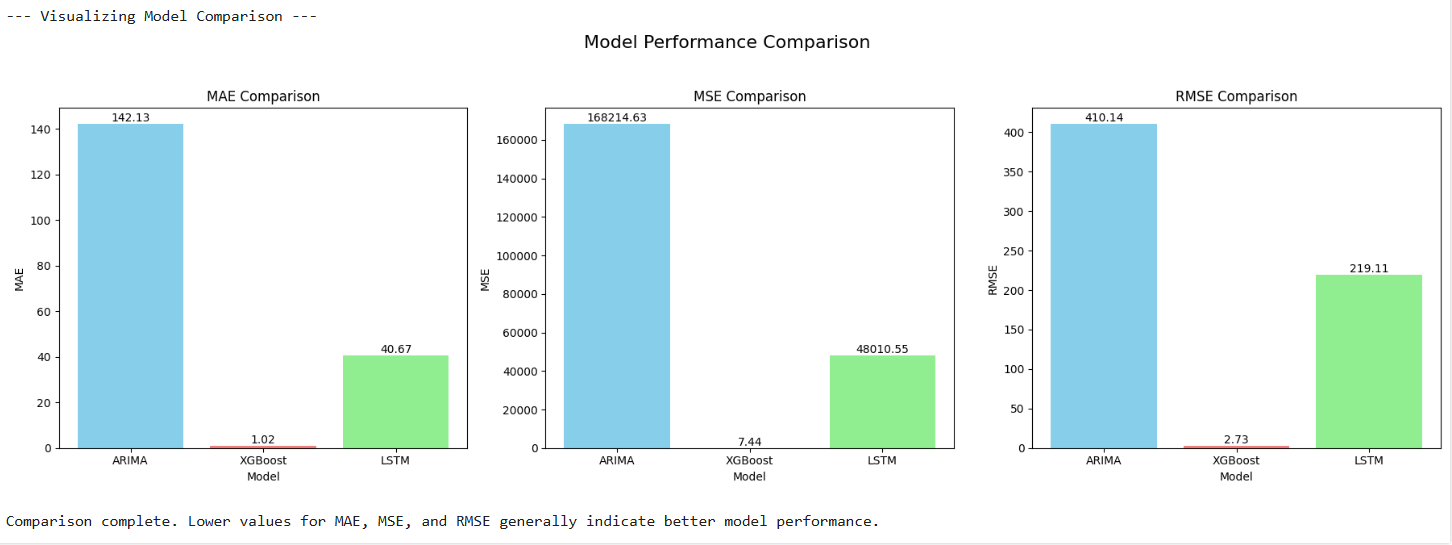
**Predicted Price (Red):** Generally follows the cyclical pattern of the actual prices but with significantly lower magnitude. The predictions capture the periodic highs and lows but consistently underestimate the actual price values.

**RMSE: 124.88 KES:** The Root Mean Squared Error is reported at 124.88 Kenyan Shillings. This suggests that, on average, the model's predictions deviate from the actual stock prices by approximately 125 KES. Given the scale of the stock prices depicted (reaching up to around 3,500 KES), this level of error indicates a relatively small average deviation.

**MAPE: 30.16%:** The Mean Absolute Percentage Error is calculated as 30.16%. This signifies that, on average, the absolute percentage difference between the predicted and the actual stock prices is about 30.16%. Observing the chart, the predicted price generally follows the cyclical behavior of the actual price, with peaks and troughs often aligning. However, the model appears to underestimate the magnitude of the price fluctuations, particularly during the earlier time steps. This underestimation contributes to the notable percentage error (MAPE of 30.16%). While the model captures the overall trend, the accuracy of the predicted price levels needs improvement. The RMSE of 124.88 KES quantifies the average price difference.

# **CHAPTER 8. MODEL EVALUATION.**

This section details the performance evaluation of the candidate models – ARIMA, XGBoost, and LSTM – based on key regression metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The objective is to identify the model exhibiting the superior predictive capability for the task at hand.

****

**1. Mean Absolute Error (MAE) Analysis:**

ARIMA: The ARIMA model shows a high MAE of approximately 142.13. This indicates that, on average, the model's predictions deviate from the actual values by a large amount, suggesting poor accuracy.

XGBoost: The XGBoost model exhibits an exceptionally low MAE of approximately 1.02. This signifies a very high degree of average accuracy, with minimal deviation between predictions and actual values.

LSTM: The LSTM model presents an MAE of approximately 21.69. While significantly better than ARIMA, its average absolute error is considerably higher than XGBoost's, indicating moderate accuracy.

**2. Mean Squared Analysis (MAE):**

ARIMA: The ARIMA model displays a very high MSE of approximately 168,214.63. This large value highlights the presence of substantial errors, likely including some very large deviations, contributing to high prediction variance.

XGBoost: The XGBoost model records a remarkably low MSE of approximately 7.44. This indicates that squared errors are minimal on average, suggesting highly accurate and consistent predictions with very low variance.

LSTM: The LSTM model shows an MSE of approximately 15,595.23. Although much lower than ARIMA's MSE, it is substantially higher than XGBoost's, indicating larger squared errors and more variability compared to XGBoost.

**3. Root Mean Squared Error (RMSE) Analysis**

ARIMA: The ARIMA model has a very high RMSE of approximately 410.14. This large value reflects the significant typical magnitude of the prediction errors, consistent with its high MAE and MSE.

XGBoost: The XGBoost model shows an exceptionally low RMSE of approximately 2.73. This reinforces the finding that XGBoost's predictions are highly accurate and consistently close to the actual values.

LSTM: The LSTM model has an RMSE of approximately 124.88. While considerably lower than ARIMA's RMSE, it is significantly higher than XGBoost's, confirming that its typical prediction errors are larger than those of XGBoost.

Among the three models used for forecasting, XGBoost stands out as the most effective, delivering superior accuracy, consistency, and reliability, as evidenced by its significantly lower error metrics across all evaluated categories. LSTM performs moderately well, surpassing ARIMA but still falling considerably short of XGBoost’s precision. Meanwhile, ARIMA exhibits the weakest performance by a wide margin, demonstrating the highest error rates and the least dependable predictive capability.

# **CHAPTER 9. DEPLOYMENT**

A key consideration in this project's deployment strategy is the decision to train and deploy an individual LSTM model for each specific NSE stock. This approach was explicitly chosen and facilitated by the nature of LSTM networks, which are well-suited for sequence-to-sequence forecasting on individual time series data. Unlike some traditional time series models or approaches that might struggle with or require complex aggregation or feature engineering across disparate series, LSTMs can effectively learn the unique temporal patterns and dependencies inherent in a single stock's price history. This allows for tailored models that capture the specific nuances of each stock's behavior, potentially leading to more accurate individual forecasts.

The choice of LSTM models for individual stocks directly impacts the deployment strategy by necessitating an infrastructure capable of managing a potentially large number of model artifacts and serving instances. The benefit, however, lies in the ability to provide highly specific forecasts for each stock, leveraging the LSTM's strength in capturing the unique time series characteristics of its dedicated data. This granular approach, enabled by the selection of LSTM, is a key highlight of this project's methodology and deployment design.

While XGBoost, a powerful gradient boosting framework, demonstrated strong predictive capabilities, particularly in handling tabular data with well-engineered features, its primary limitation in this project stemmed from the sequential nature of stock price data. Unlike LSTMs, which are designed to capture complex temporal dependencies and patterns across time steps, XGBoost requires extensive feature engineering to encode time-based relationships effectively. This extra preprocessing adds complexity and potentially limits the adaptability of the model across different stocks.

# **CHAPTER 10. CONCLUSIONS AND RECOMMENDATIONS.**

**10.1 CONCLUSIONS**

Superior Performance of XGBoost: The XGBoost model demonstrated a significantly higher level of accuracy compared to both ARIMA and LSTM, as evidenced by its substantially lower MAE (1.02), MSE (7.44), and RMSE (2.73). This suggests that XGBoost was best able to capture the underlying patterns and dynamics of the NSE time series data in this specific comparison.

Effectiveness in Handling Volatility and Non-linearity: Stock market data, including the NSE, is inherently volatile and exhibits complex non-linear relationships. The strong performance of XGBoost indicates its effectiveness in modeling these characteristics. As a tree-based ensemble method, XGBoost is well-suited to capture complex interactions between features and handle non-linear patterns that simpler linear models like ARIMA may struggle with.

LSTM's Potential, but Outperformed by XGBoost: The LSTM model, designed to handle sequential data and capture temporal dependencies, showed better performance than ARIMA but was clearly surpassed by XGBoost. While LSTMs are powerful for time series, their performance can be sensitive to architecture tuning, hyperparameter selection, and the amount of data. In this comparison, XGBoost proved to be the more effective model.

Limitations of ARIMA for Complex Patterns: The ARIMA model, a traditional linear time series approach, performed considerably worse than both XGBoost and LSTM. Its high MAE, MSE, and RMSE values suggest that the linear assumptions underlying ARIMA are likely insufficient to accurately model the intricate and potentially non-stationary nature of the NSE time series data. This highlights the advantage of more flexible machine learning models for this type of data.

**10.2 RECOMMENDATIONS**

1. Establish a Retraining Strategy

NSE market dynamics change over time. Define a strategy for regularly retraining the XGBoost model on the most recent data to ensure its predictions remain relevant and accurate. The frequency of retraining (e.g., daily, weekly, monthly) should be determined based on performance monitoring and computational resources.

1. Integrate with Risk Management

Recognize that forecasting models provide probabilistic estimates, not guarantees. Integrate the model's output into a broader trading or investment strategy that includes risk management protocols, position sizing based on confidence levels, and stop-loss mechanisms.

1. Monitor Model Performance Continuously

Once the model is deployed, establish a system for continuously monitoring its performance on live data using the chosen evaluation metrics. Set up alerts for significant degradation in performance, which would signal the need for retraining, model re-evaluation, or investigation into market regime changes.

1. Focus on high volume stocks

They have better liquidity and represent strong investor interest. This makes them suitable for short-term trading strategies due to the frequent price movement.

1. Incorporating Technical Indicators for Market Timing

Utilizing tools such as the Relative Strength Index (RSI) is crucial for determining optimal market entry and exit points, as RSI helps identify overbought or oversold conditions, aiding investors in making informed trading decisions.

1. Leveraging Real-Time Events for Improved Forecasting

Integrating live market news, economic reports, and geopolitical developments into predictive models allows for more dynamic and responsive stock price forecasts, ensuring that sudden shifts in market sentiment or external shocks are accurately accounted for in trading strategies.

**10.3 NEXT STEPS**

1. Enhancing Feature Engineering and Integration

Expanding feature sets beyond basic price data can improve prediction quality. This includes incorporating additional technical indicators like Bollinger Bands, integrating volume-based metrics for market liquidity insights, and adding fundamental and macroeconomic data such as earnings reports and sector trends to provide a broader market context.

1. Expanding Data Utilization and Continuous Retraining

Ensuring the model benefits from the most relevant data requires defining an optimal historical data horizon and systematically integrating new data. A retraining pipeline will automate data ingestion, processing, and model updates, while exploring higher-frequency trading data (e.g., hourly or intraday) can enhance short-term prediction accuracy.

1. Improving Model Evaluation and Performance Monitoring

Beyond basic error metrics like MSE and RMSE, more advanced financial metrics such as Sharpe Ratio, Maximum Drawdown, and Hit Rate will be used to assess trading strategy effectiveness. Implementing drift detection will help identify shifts in market conditions that may require model recalibration.