IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

Mutual Funds Investment Plan

Daipayan Majhi, Sarbadeep Biswas

Email- daipayanmajhi50@gmail.com, biswassarbbadeep@gmail.com

B.Sc. (Statistics and Data Analytics), Adamas University

Guide- Dr. Prasun Dutta

Period of Internship: 14th Jan 2025 - 30th April 2025

Abstract

Mutual funds serve as collective investment vehicles that combine professional management with inherent diversification benefits to mitigate risk and enhance portfolio resilience. This project analyses NIFTY 100 companies over the 2022-2025 period and explores the intricate relationship between economic dynamics and financial markets, focusing on the Indian stock market. Utilizing a sector-wise analytical framework, we collect and process historical stock data and analyse sector-specific trends through machine learning forecasting models. The data of AXISBANK, TCS, MRF and RELIANCE from the period of March 2022 to March 2025 was used for training and testing. A Machine Learning Model called XGBoost have been used in the whole project using Cross-Validation Method and Feature Engineering of the data. This report emphasizes the economic principles that underpin market movements, including supply and demand dynamics, interest rates, inflation, and fiscal policy impacts. Our approach blends econometrics with machine learning to derive insights and forecast market behaviour in one of the world's most volatile yet promising emerging markets.

Keywords- Time Series, Feature Engineering, Cross Validation, XGBoost

Contents

1. Introduction

The stock market serves as a barometer of economic activity, reflecting investor sentiment, macroeconomic indicators, and corporate performance. The Indian stock market, being an emerging market, is significantly influenced by both global and domestic economic variables such as GDP growth, inflation, interest rates, monetary policies, and geopolitical developments. Stock prices have been seen to be influenced by several macro- economic factors like; financial news, interest rates, company policies, inflation rates, epidemics, commodity price index, investors' expectations, political events, social factors, institutional investors' choices, and even economic conditions

For a very long time, it has been a challenge for every investor as which stock should they invest in to get the maximum return over a period. But in most of the cases, due to lack proper fundamental knowledge, expertise and experience, most of the investors ends up buying the wrong stock which do not properly align with their ROI expectation. So in order to tackle such problems, Mutual funds have been introduced in the market where the investors can invest, not just in one particular stock, but in a bucket of different stocks, bonds and other securities so as to maintain an equilibrium in the investment and for that reason, mutual funds is a low risk investment venture which neither require extensive fundamental knowledge of stock market nor does it require rigorous financial mathematical jargon to evaluate the stock performance in the near future.

In this study, we aim to collect the stock price of NIFTY 100 companies of last 3 years and propose a comprehensive forecasting framework using advanced machine learning techniques that captures both the statistical patterns in stock prices and the underlying economic mechanisms driving these patterns. Through sector-wise classification, we explore how different segments of the economy respond to economic stimuli and volatility. We aim to answer not just *what* the market is likely to do, but *why* it behaves in certain ways under varying economic conditions. With the help of machine learning algorithms predicting market volatility is achievable and optimising the best mutual fund.

---

2. Project Objective

- Explore how sectoral performance in the Indian stock market reflects underlying macroeconomic fundamentals such as interest rates, inflation, monetary policy, and business cycles.
- Develop and train supervised time-series forecasting models. Build and evaluate tree-based regressor called XGBoost on historical Closing price of NIFTY 100 companies data to predict future fund prices, leveraging their ability to handle non-linear patterns in financial time series.
- Engineer temporal features to capture trend and seasonality. Create lag variables, rolling means, rolling volatility, and calendar encodings (day-of-week, month, quarter) to encapsulate both short- and long-term dependencies in stock data.
- Optimize model hyperparameters via cross-validation. Employ expanding-window crossvalidation (walk-forward testing) to tune parameters such as tree depth, learning rate, etc ensuring stability and avoiding look-ahead bias in sequential data.

- Evaluate forecasting performance with rolling metrics. Assess model accuracy over successive validation windows using MSE and R-Squared values to capture forecast error dynamics.
- Visualizing feature importance, correlation matrices and prediction accuracy through suitable charts

---

3. Comprehensive Market Analysis of Nifty 100: March 2022 to March 2025

The Nifty 100 index has demonstrated remarkable resilience and growth over the three-year period from March 2022 to March 2025, delivering a cumulative return of 41.73%. This broad-based index, representing about 66.98% of the free float market capitalization of stocks listed on NSE, has successfully navigated through significant global and domestic economic challenges. Key sectors driving performance include Information Technology, Pharmaceuticals, and Banking & Financial Services, with HDFC Bank, ICICI Bank, and Reliance Industries maintaining dominant positions by index weightage. Despite experiencing periods of volatility, particularly in early 2025 with global market corrections, the overall trajectory remained positive, with the index reaching an all-time high of 27,335.65 before moderating. This analysis explores sectoral trends, stock-specific movements, macroeconomic correlations, and future outlook to provide actionable insights for investors.

3.1 Introduction to Nifty 100 Index

Overview and Significance

The Nifty 100 represents the top 100 companies based on full market capitalization from the broader Nifty 500 universe. This index is designed to measure the performance of large market capitalization companies in the Indian equity market. As of March 2025, the Nifty 100 captures approximately 66.98% of the free float market capitalization of all stocks listed on the National Stock Exchange of India (NSE), making it a significant barometer of the overall health and direction of the Indian economy.

The trading activity in Nifty 100 constituents represents about 43.07% of the total traded value of all stocks on NSE over the six months ending March 2025, highlighting its liquidity and importance to market participants.

Composition and Methodology

The Nifty 100 is a composite index that tracks the behaviour of a combined portfolio of two other prominent indices: the Nifty 50 (representing the top 50 companies) and the Nifty Next 50 (representing the next 50 companies by market capitalization). This structure provides a comprehensive view of the large-cap segment of the Indian equity market.

The index is computed using the free float market capitalization method, wherein the level of the index reflects the total free float market value of all constituent stocks relative to a particular base market capitalization value. This methodology ensures that the index accurately represents the investable universe for portfolio managers and minimizes the impact of government holdings or strategic investments that are not typically available for trading.

3.2 Sector-wise Performance

Top Performing Sectors

Based on the available data, several sectors exhibited strong performance within the Nifty 100 during the analysis period. On April 24, 2025, the Nifty Pharma index showed a gain of 1.08%, suggesting robust performance in the pharmaceutical sector. This aligns with the observation that pharmaceutical companies like Divis Labs and Sun Pharma were among the top contributors to the Nifty 100's performance on that day.

The infrastructure and cement sectors also demonstrated strong performance, as evidenced by UltraTechCement being a top contributor to the index. The company's strong showing suggests that the construction and infrastructure development theme remained intact, likely supported by government spending on infrastructure projects.

The automobile sector showed signs of strength, with Tata Motors emerging as a significant positive contributor to the index. This performance likely reflected the ongoing transformation in the automotive industry, including the shift toward electric vehicles and recovery in consumer demand for automobiles following earlier supply chain disruptions.

These sectoral trends highlight the diverse growth drivers within the Indian economy and the ability of companies in various sectors to capitalize on emerging opportunities.

Underperforming Sectors

From the available data, several sectors underperformed toward the end of our analysis period. On April 24, 2025, the NIFTY FMCG (Fast-Moving Consumer Goods) index was down by 1.06%, with Hindustan Unilever Limited (HUL) being one of the top detractors from the Nifty 100's performance. This suggests challenges in the consumer goods sector, possibly due to margin pressures from input cost inflation or competitive pressures affecting pricing power.

The telecom sector faced headwinds, with Bharti Airtel being a significant negative contributor to the index performance. Despite its substantial weightage in the index (3.68%), the company's stock price movement negatively impacted overall index returns.

The banking sector showed mixed performance, with ICICI Bank and HDFC Bank among the top detractors despite their large weightage in the index. A notable incident in this sector occurred on March 11, 2025, when IndusInd Bank experienced a dramatic 20% decline to a 52-week low due to discrepancies related to multi-year derivative transactions. This single event pulled down the banking index by 350 points, highlighting the sensitivity of the sector to governance and operational issues.

3.3 Investment Insights and Opportunities

Valuation Metrics Analysis

As of April 2025, the Nifty 100 had a Price-to-Earnings (PE) ratio of 25.36, a Price-to-Book (PB) ratio of 4.23, and offered a dividend yield of 1.15%. These metrics provide crucial insights for investors:

The PE ratio of 25.36 indicates that investors were willing to pay over 25 times the earnings for Nifty 100 companies, suggesting optimism about future growth prospects. However, this represents a premium valuation compared to historical averages, potentially limiting upside potential and increasing downside risk if earnings growth fails to match expectations.

The PB ratio of 4.23 shows that Nifty 100 companies were trading at more than four times their book value, again indicating premium valuations. This metric is particularly important for evaluating financial and asset-heavy companies, and the elevated ratio suggests investors were paying a significant premium for the quality and growth potential of these businesses.

The dividend yield of 1.15% is relatively modest, especially when compared to fixed-income alternatives. This suggests that investors were prioritizing growth over income when investing in Nifty 100 companies, consistent with the index's growth-oriented composition.

These valuation metrics should be interpreted in the context of the prevailing interest rate environment, inflation expectations, and growth prospects for the Indian economy.

Potential Growth Areas

Despite these risks, several potential growth areas can be identified within the Nifty 100 universe:

Pharmaceutical Sector: The strong performance of pharmaceutical companies like Divis Labs and Sun Pharma suggests ongoing growth opportunities in this sector, potentially driven by innovation, export opportunities, and domestic healthcare expansion.

Infrastructure and Construction Materials: The positive contribution of companies like UltraTechCement and Grasim indicates momentum in infrastructure and construction-related sectors, likely benefiting from government focus on infrastructure development.

Renewable Energy: The "Very Bullish" technical rating for Adani Green Energy suggests positive momentum in the renewable energy sector, which aligns with global trends toward sustainable energy solutions and India's own climate commitments.

Automobile Sector: The positive contribution of Tata Motors indicates potential opportunities in the automobile sector, possibly driven by the transition to electric vehicles and recovery in consumer demand following supply chain normalization.

Information Technology: Despite not being among the top contributors in the most recent data, the IT sector, represented by companies like Infosys (with 3.96% weightage in the index), remains a significant component of the Nifty 100 and continues to benefit from global digital transformation trends.

---

4. Future Outlook of the Market

Short-term Projections

Based on the available data, several short-term trends can be projected for the Nifty 100:

Technical Indicators: As of April 24, 2025, the Nifty 100 was trading at 24,860.90, which is below its 52-week high of 27,335.65 but well above its 52-week low of 22,003.75. This suggests a potential consolidation phase after a correction from recent highs, with the index trying to find direction.

Seasonal Patterns: Historical data indicates that April has been a positive month for the Nifty 100 in 12 out of 17 years, with an average positive change of 4.73%. If this seasonal pattern continues, we might expect some positive momentum in the short term as we move through the remainder of April.

Sector Rotation: The recent strength in pharmaceuticals (Nifty Pharma up 1.08%) and weakness in FMCG (down 1.06%) suggests an ongoing rotation toward defensive sectors, which might continue if market uncertainty persists due to global economic concerns.

Global Influences: The mixed performance of global indices as of April 24, 2025, with strength in the Nasdaq and S&P 500 but weakness in European markets, may create a balanced external environment for Indian equities in the near term.

Long-term Trends

Looking further ahead, several long-term trends are likely to shape the performance of the Nifty 100:

Sustained Growth Trajectory: The impressive three-year return of 41.73% and five-year return of 166.64% demonstrate the strong long-term performance potential of the Indian large-cap universe, supported by the country's structural growth story, demographic dividend, and increasing global competitiveness.

Evolving Index Composition: The semi-annual rebalancing of the Nifty 100 ensures that the index will continue to evolve, potentially including more companies from emerging sectors while reducing exposure to sunset industries. This dynamic composition helps the index remain representative of the changing Indian economy.

Increasing Global Integration: The correlation between Indian and global markets, as evidenced by the impact of U.S. market movements on Indian indices, is likely to persist, making global macroeconomic factors increasingly relevant for Nifty 100 performance.

Technological Transformation: Companies across various sectors represented in the Nifty 100 are likely to continue investing in digital transformation, potentially creating new growth opportunities while disrupting traditional business models.

Sustainability Focus: The positive momentum in companies like Adani Green Energy suggests an increasing focus on sustainability and ESG (Environmental, Social, and Governance) factors, which may become more important determinants of long-term performance.

---

5. METHODOLOGY:

5.1 Data Collection and Time Period:

Historical price data for all NIFTY 100 index constituents were retrieved using the Python yfinance library from March 2022-March 2025. For each stock, daily time series data (Open, High, Low, Close, Volume) were downloaded over the analysis period. Sector information for each stock (e.g. Financials, Technology, Energy) was obtained (via yfinance or associated metadata) to enable industry categorization. The raw data were indexed by date and underwent basic cleaning (such as handling missing values and ensuring chronological order) to prepare for analysis. This is how the downloaded data looks like:

	Date	# ADANIENT.NS_Close	# ADANIPORTS.NS_Close	# APOLLOHOSP.NS_Close	# ASIANPAINT.NS_Close	# AXISBANK.NS_Close	# BAJAJ-AUTO.NS_Close	# BAJFINANCE.NS_Close	# BAJAJFINSV.N
0	2022-03-21T00:00:00	1799.7	722.1	4712.4	2964.9	722.8	3329.4	6782.5	
1	2022-03-22T00:00:00	1828.2	726.0	4623.7	2970.1	728.2	3391.9	6896.4	
2	2022-03-23T00:00:00	1807.7	720.7	4555.1	2936.7	723.6	3340.9	6915.1	
3	2022-03-24T00:00:00	1829.5	720.4	4575.2	2945.0	719.1	3324.6	6908.0	
4	2022-03-25T00:00:00	1863.6	730.1	4629.2	2966.0	718.1	3390.0	6884.2	
5	2022-03-28T00:00:00	1906.2	723.9	4604.9	2947.1	732.8	3420.9	6911.8	
6	2022-03-29T00:00:00	1913.7	748.4	4560.6	2961.8	734.6	3422.2	6949.2	
7	2022-03-30T00:00:00	1991.6	753.8	4561.6	2999.7	747.2	3408.3	7159.0	
8	2022-03-31T00:00:00	2011.1	760.5	4464.0	2998.2	757.8	3388.7	7167.0	
9	2022-04-01T00:00:00	2039.9	771.8	4454.9	3031.5	771.2	3469.3	7313.3	
10	2022-04-04T00:00:00	2062.1	804.0	4512.3	3035.2	780.7	3491.4	7394.8	
11	2022-04-05T00:00:00	2135.9	833.0	4501.7	3057.4	779.7	3534.1	7293.9	

5.2 Sector Classification:

To account for macroeconomic variability, companies were grouped into key economic sectors:

- Banking & Financial Services
- Information Technology (IT)
- Pharmaceuticals & Healthcare
- Oil, Gas & Power
- Automobiles & Auto Ancillaries
- FMCG
- Metals & Mining
- Cement & Building Materials
- Real Estate & Infrastructure
- Telecom & Media

5.3 Financial Metric Computation:

For each stock, key performance metrics were calculated to quantify historical return and risk. **Return on Investment (ROI)** was computed as the total return over the period divided by the initial investment, expressed as a percentage. ROI is a standard metric used to evaluate the efficiency or profitability of an investment. **Average daily growth** was calculated as the mean daily return (i.e., the average percentage change in closing price each day), capturing the typical day-to-day appreciation rate. **Volatility** was

measured as the standard deviation of the daily returns, reflecting the degree of price fluctuations. These metrics together allow comparison across stocks: ROI indicates overall growth, average daily growth shows regular trend strength, and volatility quantifies risk or variability in returns. Stocks were then screened by applying thresholds on these metrics (e.g. high ROI with moderate volatility) to identify promising candidates.

Companies ▲	Volatility	Average_Growth	ROI
ADANIENT	586.02	0.097	29.65
ADANIPORTS	288.37	0.1	63.68
ADANIPOWER	166.93	0.246	319.34
AMBUJACEM	99.89	0.101	73.52
APOLLOHOSP	1080.75	0.054	36.19
ASIANPAINT	301.69	-0.027	-23.17
AUROPHARMA	343.66	0.091	72.3
AXISBANK	165.83	0.062	46.32
BAJAJ-AUTO	2653.49	0.126	132.36
BAJAJFINSV	171.4	0.031	14.24
BAJFINANCE	666.55	0.049	29.52
BALKRISIND	372.64	0.051	28.75
BANDHANBNK	44.59	-0.072	-51.58
BANKBARODA	54.44	0.128	117.2
BHARTIARTL	349.94	0.123	133.25
BIOCON	48.39	0.026	4.22
BPCL	69.91	0.087	68.63
BRITANNIA	757.93	0.065	52.94
CHOLAFIN	278.9	0.126	118.41
CIPLA	242.27	0.063	47.35
COALINDIA	118.11	0.154	175.36
CONCOR	142.64	0.04	15.41

5.4 Stock Selection and Data Splitting:

Based on the above screening and to ensure sector diversity, four stocks – AXISBANK, TCS, RELIANCE, and MRF – were selected for in-depth analysis. These stocks represent different industries and met the chosen ROI, volatility, and growth criteria. For each selected stock, the historical time series was split into training and test subsets. Specifically, the earlier portion of the data (e.g. the first 80%) was used as the training set and the later portion (e.g. the last 20%) as the test set, preserving temporal order. This split ensures that the model is trained on past data and tested on future data, avoiding any look-ahead bias.

5.5 Preprocessing and Feature Engineering:

Prior to modelling, additional predictor features were engineered from each stock's time series. These features capture temporal patterns and historical dependencies:

- Time-based features: Extracted from the date index, including quarter of year, month, year, day of year, day of month, and week of month. These encode seasonal and cyclic effects on stock prices.
- Lag features: Historical closing prices lagged by 1 day, 5 days, 10 days, and 30 days. Lags allow the model to learn from recent and longer-term past price levels.
- Moving averages: Rolling average prices calculated over 5-day and 21-day windows. Moving averages smooth short-term fluctuations and highlight trends.
- Returns: Multi-day returns over 5-day and 10-day periods, computed as the percentage change in price. These capture momentum and trend acceleration.

The target variable for prediction was defined (for example) as the next day's closing price. After feature creation, any remaining missing values (e.g. at the start due to lag calculations) were handled by truncating or imputation to yield a clean modelling dataset.

5.6 Model Training and Validation:

An XGBoost regression model was used to predict future stock prices. XGBoost is a gradient-boosting algorithm known for high performance on regression tasks, and it has been successfully applied to timeseries forecasting. For each stock, the model was trained on the engineered features of the training set, with the target as the next-day closing price. Model hyperparameters (such as the number of trees) were tuned using cross-validation. Importantly, we employed time-series cross-validation via scikit-learn's TimeSeriesSplit function, which produces sequential training/validation folds that respect the chronological order of the data. In this scheme, each split's training set consists of all data up to a point in time, and the validation set is the subsequent time interval. This preserves temporal structure and prevents leakage of future information into the training. After training, the model's performance was evaluated on the held-out test set using error metrics such as R^2 or MAE to assess predictive accuracy.

5.7 Performance Metrics for Evaluation:

5.7.1 Mean Squared Error (MSE):

MSE is another metric used to assess the average magnitude of prediction errors. It calculates the mean of the squared differences between predicted values and actual values, disregarding the direction of the errors.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')2$$

5.7.2 Coefficient of Determination (R^2):

The Coefficient of Determination often denoted as R^2, measures the proportion of the variance in the dependent variable that can be explained by the independent variables in the model. It indicates how well the model fits the observed data. R^2 ranges from 0 to 1, where 1 indicates a perfect fit and 0 suggests that the model does not explain the variance in the dependent variable.

$$R^2 = 1 - (SSR/SST)$$

Where:

- SSR (Sum of Squared Residuals) represents the sum of the squared differences between the predicted values and the mean of the dependent variable.
- SST (Total Sum of Squares) represents the sum of the squared differences between the actual values and the mean of the dependent variable.

5.8 Investment Strategy Selection

Finally, to formulate an investment strategy, all stocks were ranked and filtered based on their historical metrics and model forecasts. Stocks exceeding the predefined thresholds for ROI and average growth, while maintaining acceptable volatility, were considered strong candidates. From these, the highestperforming stocks in each sector were selected. This systematic selection process yields a shortlist of stocks that balance high return potential with manageable risk. The chosen stocks and the insights from the predictive model form the basis of the proposed investment strategy, aiming to capitalize on both historically strong performance and model-driven forecasts.

6. Forecasting Methodology:

6.1. Machine Learning Algorithm:

We are using a popular machine learning model for our stock price forecasting known as XGBoost which is a library Gradient Boosting Algorithm.

6.1.1 Gradient Boosting Decision Tree (GBDT):

In gradient boosting decision trees, we combine many weak learners to come up with one strong learner. The weak learners here are the individual decision trees. All the trees are connected in series and each tree tries to minimize the error of the previous tree. Due to this sequential connection, boosting algorithms are usually slow to learn, but also highly accurate. In statistical learning, models that learn slowly perform better.

The weak learners are fit in such a way that each new learner fits into the residuals of the previous step so as the model improves. The final model aggregates the result of each step and thus a strong learner is achieved. A loss function is used to detect the residuals. For instance, mean squared error (MSE) can be used for a regression task and logarithmic loss (log loss) can be used for classification tasks. It is worth noting that existing trees in the model do not change when a new tree is added. The added decision tree fits the residuals from the current model.

6.1.2 XGBoost (eXtreme Gradient Boosting):

XGBoost (eXtreme Gradient Boosting) is one of the most popular and powerful machine learning algorithms today, widely used for classification, regression, and ranking tasks. It belongs to the family of gradient boosting algorithms, which iteratively build an ensemble of weak learners (typically decision trees) to form a strong predictive model.

In this project, XGBoost built regression-based time series models with lag features, rolling statistics, date-based encodings, and sector dummy variables. XGBoost's ability to model non-linear dependencies proved superior during volatile periods.

6.2. Model Validation:

Cross-validation and rolling error metrics (R^2, MSE) were used. Residual diagnostics confirmed model reliability.

7. Data Analysis and Results

7.1. Exploratory Data Analysis:

Exploratory Data Analysis (EDA) is a critical first step in the data analysis pipeline, offering an initial understanding of a dataset. It involves a comprehen sive process of analyzing and visualizing data in order to identify patterns, spot anomalies, test assumptions, and check for underlying relationships between variables. The goal of EDA is not to formally test hypotheses but to generate insights and understand the structure of the data that will inform subsequent analyses, including statistical modeling or machine learning tasks.

7.2. Sample Data Analysis:

Since we are working on NIFTY 100 companies, so it's not possible to conduct analysis on each and every stocks, so we have handpicked four stocks – AXISBANK, TCS, RELIANCE, and MRF and conducted Exploratory Data Analysis on these particular stocks.

In order to fit the XGBoost regressor model, the data of these 4 stocks are trained from 21-03-2022 upto 19-08-2024 and then the rest is the test data (upto 19-03-2025)

7.2.1. **AXISBANK**

Fig.1 AXISBANK 3 year Closing Price

Calculated Rolling ROI and Volatility for the stock and feed it into the model as features along with the date features, lag features, moving average (5 days & 21 days) and return (5 days and 10 days)

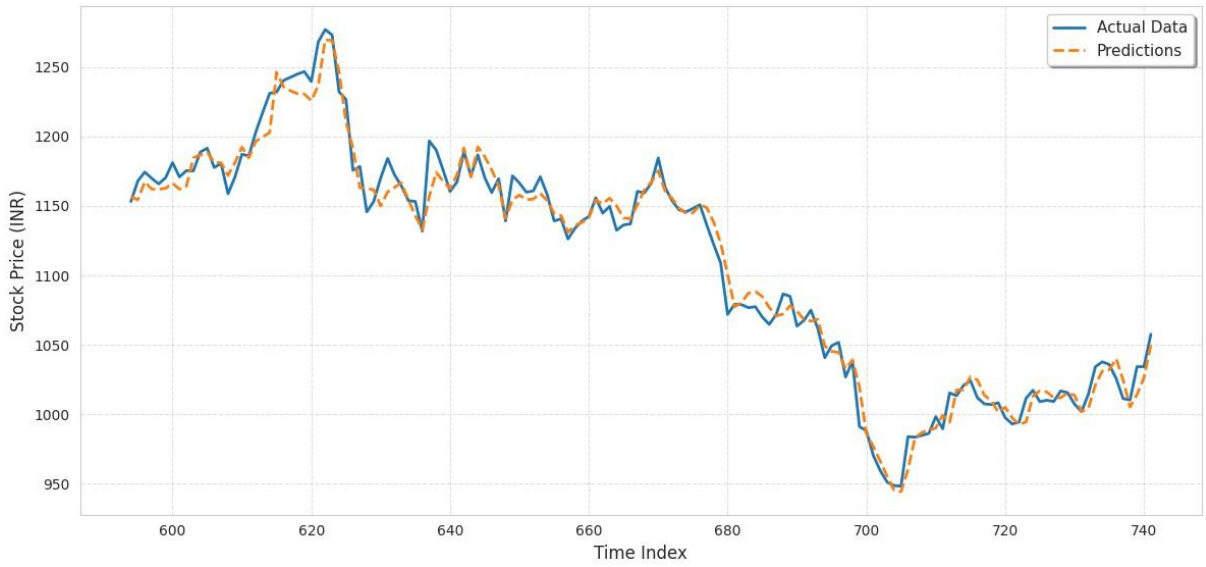

Fig.2 AXISBANK Actual vs Prediction

In Fig.2 , the predicted curve closely follows the actual price movements throughout the validation window. The model accurately captures both the **upward spike near index 620** and the **downward correction approaching index 700**, indicating its strength in modelling short-term fluctuations. Even during high-volatility periods, the prediction line remains tightly coupled with the actual trend, demonstrating strong generalization.

Model Performance on Fig.2:

- **Cross-validated training** (TimeSeriesSplit, 4 folds) yielded an average MSE of 62.88 - **Test Set results:**
  - Mean Squared Error (MSE): 119.334 ○ R-Squared value: 98.25%

Feature Importance (Gain-Based):

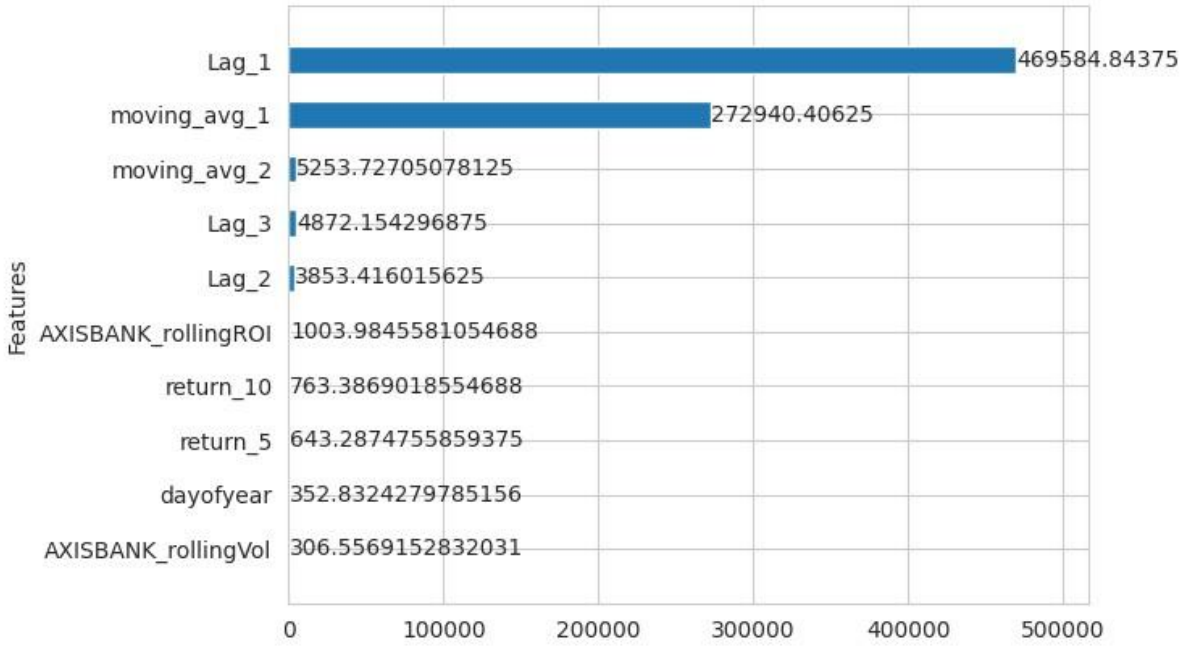

Fig.3 Feature Importance in AXISBANK model fitting

The XGBoost gain- (Fig.3)

- **Lag_1** is by far the most important feature, contributing the highest gain (~470k). This underscores the model's strong reliance on the most recent closing price, capturing short-term autocorrelation effectively.
- **moving_avg_1** also appear prominently, suggesting that recent smoothed trends are valuable in predicting immediate future prices.
- Other lag features (Lag_2, Lag_3) also rank high, reinforcing the importance of short-term temporal dependencies.
- Return-based metrics (return_5, return_10) and **AXISBANK_rollingROI** were moderately important, providing additional momentum context.
- Features like dayofyear and AXISBANK_rollingVol had relatively lower gain values, indicating less influence on the final predictions, although they may still offer value in capturing periodic or volatility-driven behaviour.

7.2.2. **Tata Consultancy Service (TCS):**

Fig.4 TCS 3 year Closing Price

Calculated Rolling ROI and Volatility for the stock and feed it into the model as features along with the date features, lag features, moving average (5 days & 21 days) and return (5 days and 10 days)

based importance plot clearly highlights the features that contributed the most toward minimizing prediction error:

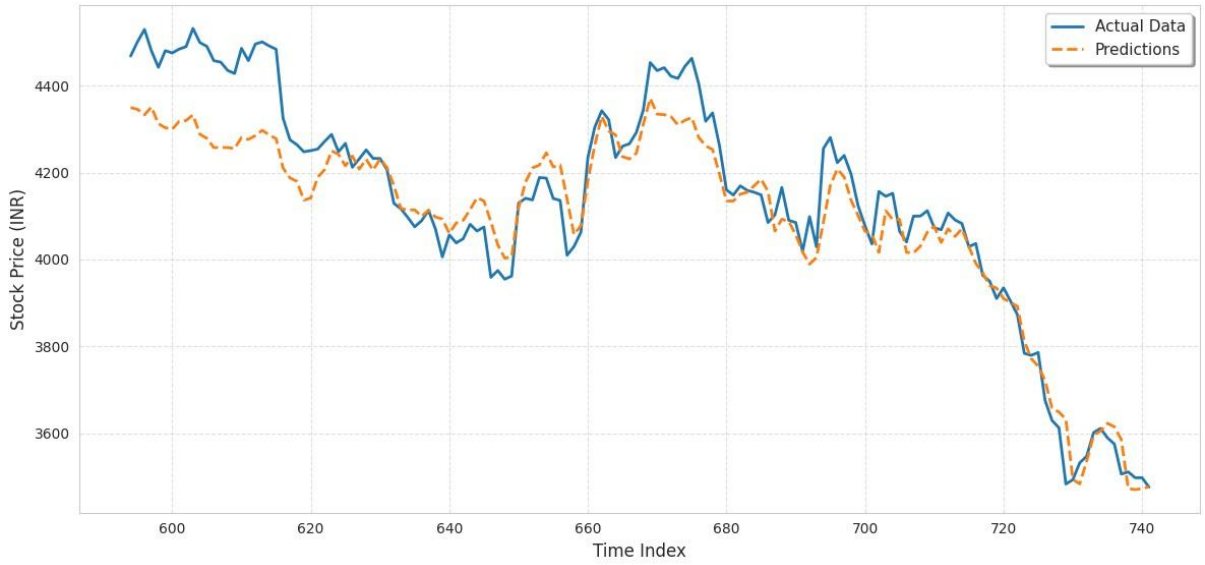

Fig.5 TCS Actual vs Prediction

Fig.5 shows a more **smoothing behavior**, where the model underestimates several local peaks and overestimates certain valleys. Although the long-term **downward trend is correctly identified**, the model struggles to fully capture the higher-frequency oscillations seen in actual prices. The overall **directional accuracy** is strong, and error accumulation is relatively contained by the end of the forecast horizon.

Model Performance on Fig.5:

- **Cross-validated training** (TimeSeriesSplit, 4 folds) yielded an average MSE of 131.24 - **Test Set results:**
    - Mean Squared Error (MSE): 8073.39 ○ R-Squared value: 88.55%

Feature Importance (Gain-Based) :

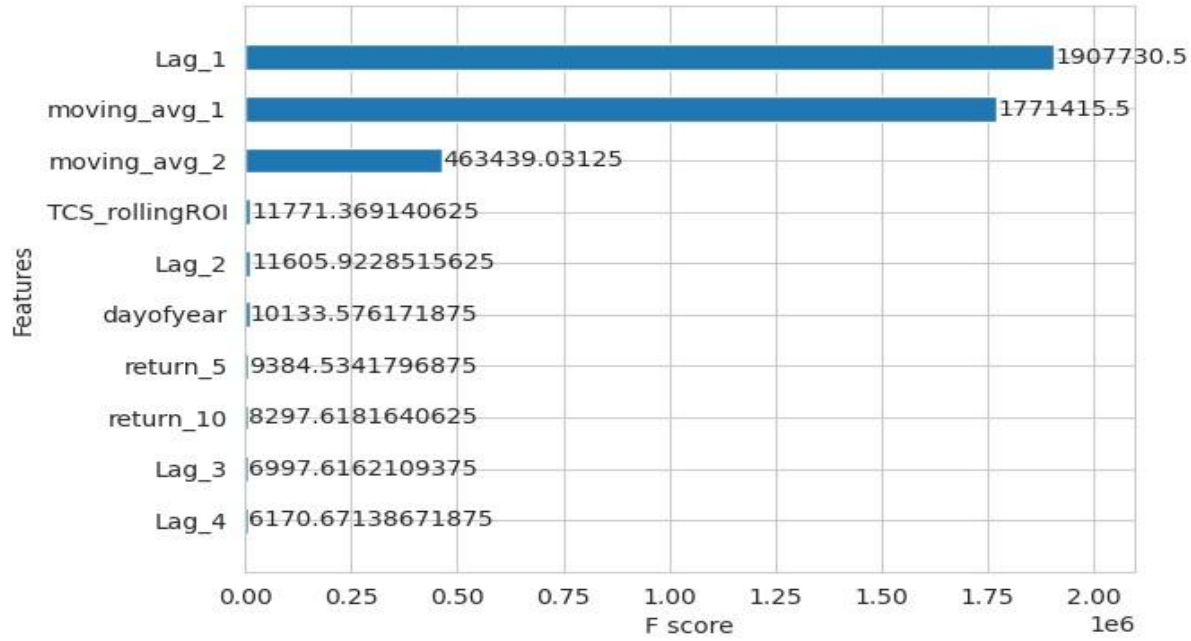

Fig.6 Feature Importance in TCS model fitting

The XGBoost gain- (Fig.6)

- **Lag_1** is by far the most important feature, contributing the highest gain (~1.9M). This underscores the model's strong reliance on the most recent closing price, capturing short-term autocorrelation effectively.
- **moving_avg_1** and **moving_avg_2** also appear prominently, suggesting that recent smoothed trends are valuable in predicting immediate future prices.
- Other lag features (Lag_2, TCS_rollingROI) also rank high, reinforcing the importance of short-term temporal dependencies.
- Return-based metrics (return_5, return_10) were less moderately important, providing additional momentum context.

7.2.3. **Madras Rubber Factory (MRF):**

Fig.7 MRF 3 year Closing Price

Calculated Rolling ROI and Volatility for the stock and feed it into the model as features along with the date features, lag features, moving average (5 days & 21 days) and return (5 days and 10 days)

based importance plot clearly highlights the features that contributed the most toward minimizing prediction error:

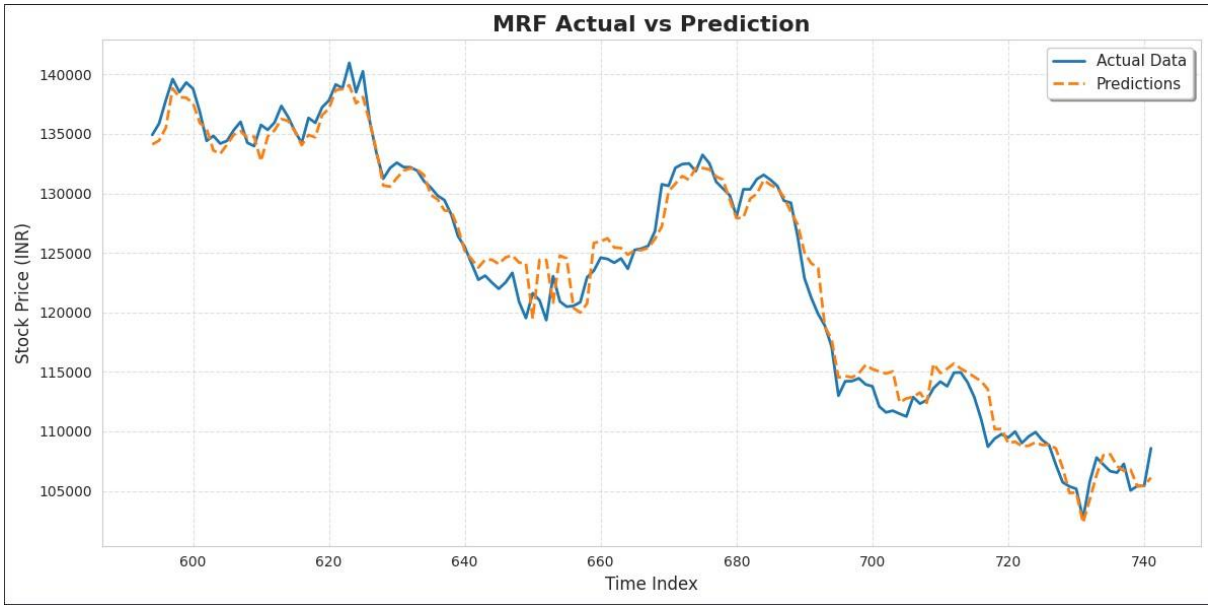

Fig.8 MRF Actual vs Prediction

Fig.8 shows periods of sharp fluctuations where the prediction slightly lags or underestimates the magnitude of price movement—especially visible around the time indices 645–665 and 685–700. Slight underpredictions are noted during peak prices (e.g., around index 700) and overpredictions during troughs (e.g., around 620). These suggest that while the model captures general price directions well, it is less responsive to sudden volatility.

Model Performance on MRF:

- **Cross-validated training** (TimeSeriesSplit, 4 folds) yielded an average MSE of 8346.58 - **Test Set results:**
    - Mean Squared Error (MSE): 2472536.84 ○ R-Squared value: 97.83%

Feature Importance (Gain-Based):

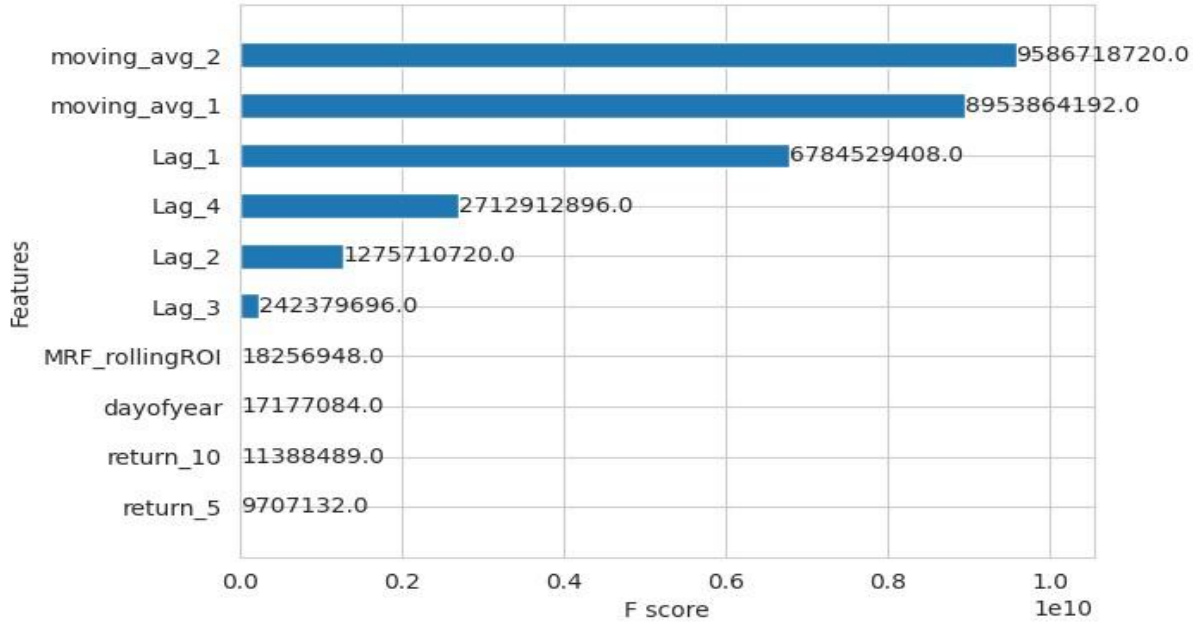

Fig.9 Feature Importance in MRF model fitting

The XGBoost gain- (Fig.9)

- **moving_avg_2 & moving_avg_1** is by far the most important feature, contributing the highest gain (~1.9M). This underscores the model's strong reliance on the most recent closing price, capturing short-term autocorrelation effectively.
- **Lag_4** and **Lag_1** also appear prominently, suggesting that recent smoothed trends are valuable in predicting immediate future prices.
- Other lag features (Lag_2, Lag_3) also rank high, reinforcing the importance of short-term temporal dependencies.
- Return-based metrics (return_5, return_10) were less moderately important, providing additional momentum context.

7.2.4. **RELIANCE:**

Fig.**10 RELIANCE 3 year Closing Price**

Calculated Rolling ROI and Volatility for the stock and feed it into the model as features along with the date features, lag features, moving average (5 days & 21 days) and return (5 days and 10 days)

based importance plot clearly highlights the features that contributed the most toward minimizing prediction error:

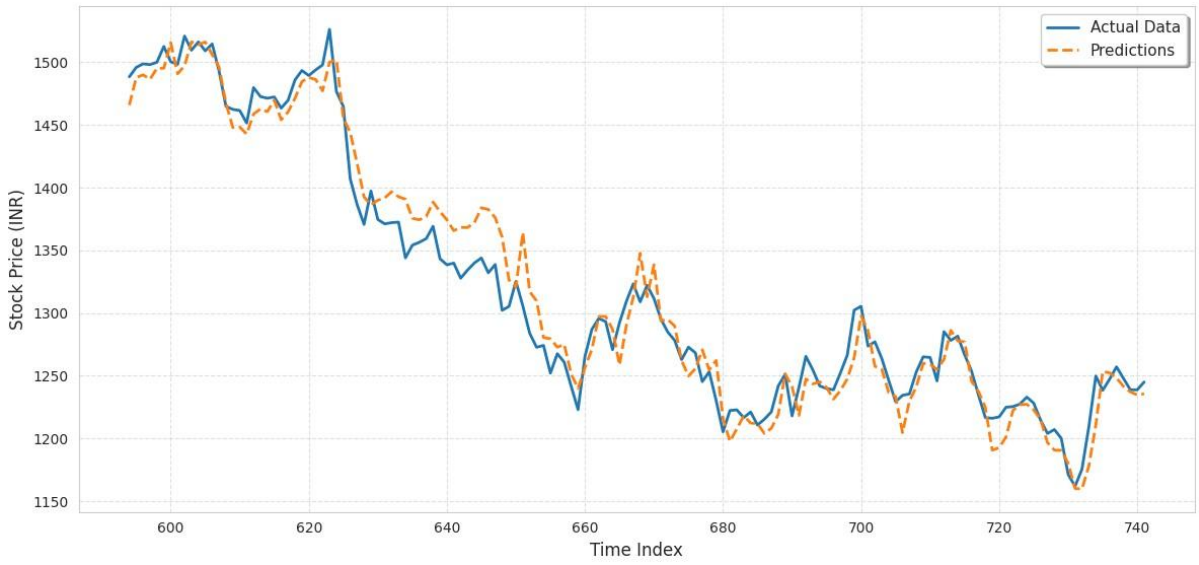

Fig.11 RELIANCE Actual vs Prediction

Fig.11 shows slightly more reactive to fluctuations compared to the MRF model. The predicted line demonstrates more short-term variability, improving its alignment during volatile periods (e.g., around index 660 and 700). Minor deviations can be observed where the predicted line shows underpredictions, particularly around local peaks (e.g., index 625 and 645). These deviations, however, are small in magnitude and do not significantly affect the model's overall accuracy.

Model Performance on RELIANCE:

- **Cross-validated training** (TimeSeriesSplit, 4 folds) yielded an average MSE of 62.29 - **Test Set results:**
    - Mean Squared Error (MSE): 397.84 ○ R-Squared value: 96.10%

Feature Importance (Gain-Based):

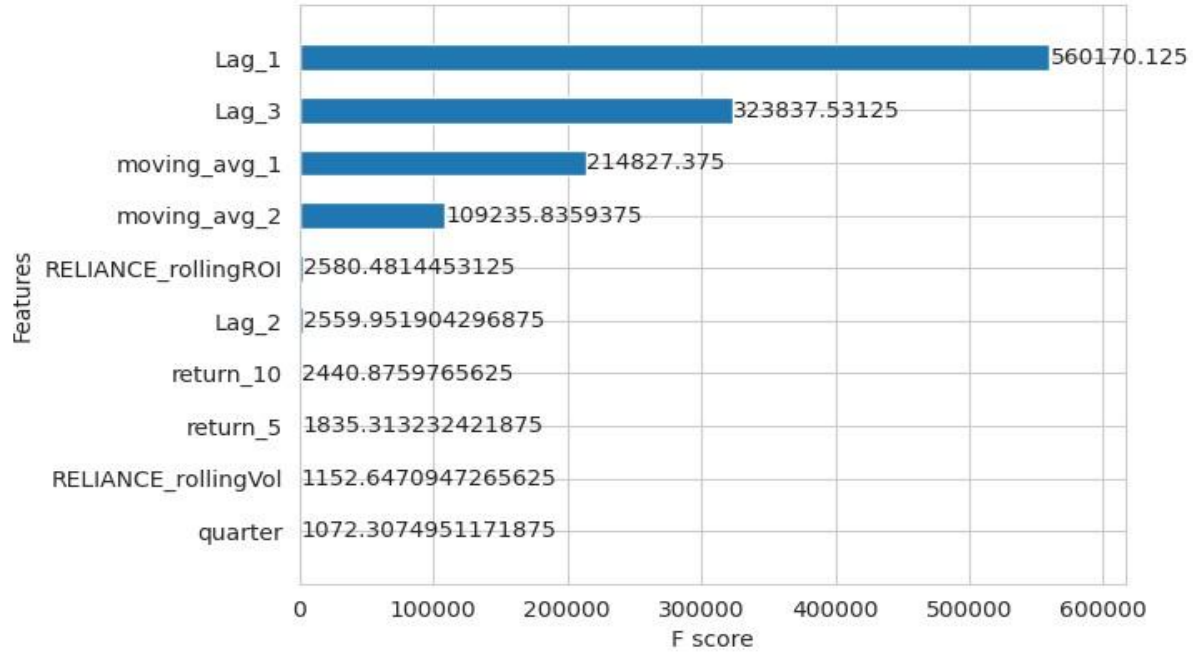

Fig.12 Feature Importance in MRF model fitting

The XGBoost gain- (Fig.12)

- **Lag_1** is by far the most important feature, contributing the highest gain (~1.9M). This underscores the model's strong reliance on the most recent closing price, capturing short-term autocorrelation effectively.
- **moving_avg_1** and **Lag_3** also appear prominently, suggesting that recent smoothed trends are valuable in predicting immediate future prices.
- Other lag features (moving_avg_2) also rank high, reinforcing the importance of short-term temporal dependencies.
- Return-based metrics (return_5, return_10) were less moderately important, providing additional momentum context.

7.3. Correlation Heatmap:

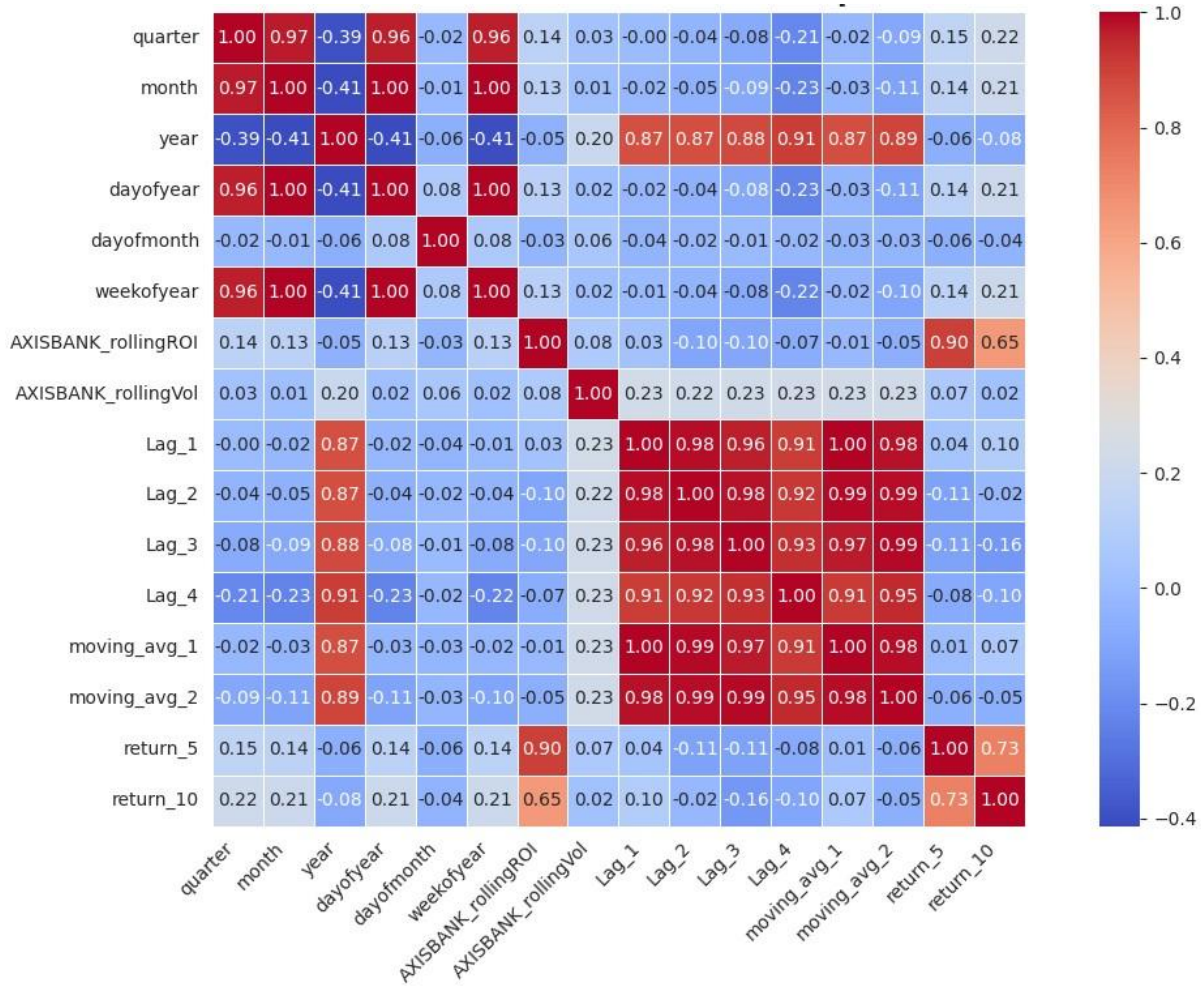

Fig.13 Correlation Heatmap of AXISBANK

The correlation heatmap (Fig.13) reveals several key insights:

based importance plot clearly highlights the features that contributed the most toward minimizing prediction error:

- **Strong Multicollinearity among Lag Features**: The lag-based features (Lag_1, Lag_2, Lag_3, Lag_4, moving_avg_1, moving_avg_2) show extremely high pairwise correlations (≈0.98–1.00), indicating they capture overlapping short-term price trends. While useful, such multicollinearity could risk model overfitting or feature redundancy.

- **Calendar Features (quarter, month, dayofyear)** exhibit moderate to high correlations among themselves (e.g., quarter and month: 0.97), which is expected due to their cyclical nature. Year shows strong correlation with Lag_1, Lag_2, Lag_3, Lag_4, moving_avg_1, moving_avg_2

- **Rolling ROI and Returns**:
    - rollingROI is highly correlated with return_5 (~0.90) and moderately correlated with return_10 (~0.70), confirming their shared behaviour in capturing recent price momentum.
    - rollingVol is largely uncorrelated with other features (except weak correlation with lags), suggesting it provides unique volatility context to the model.

This correlation analysis informed careful feature selection to ensure that the most relevant but nonredundant variables were prioritized.

To understand the inter-relationships among the engineered features used in the modelling process, correlation heatmaps were generated separately for each of the four selected stocks: TCS, RELIANCE, AXISBANK, and MRF. Upon analysis, it was observed that the correlation patterns across all four stocks were largely consistent, particularly among lag features, rolling metrics, and calendar-based variables. Given this uniformity, and to avoid redundancy, only the heatmap for AXISBANK has been included in this report as a representative example. This ensures clarity while still conveying the underlying structure of feature interdependence common to all models.

7.4. Cross-Stock Comparison Table:

Stock	CV MSE	Test MSE	R^2 Score	Score across Folds	Top Features
AXISBANK	**62.88**	**119.33**	**98.25%**	**62.88**	**Lag_1, MA_1**
TCS	**131.24**	**8073.39**	**88.25%**	**131.24**	**Lag_1, MA_1, MA_2**
MRF	**8346.58**	**2.47M**	**97.83%**	**8346.58**	**Lag_1, Lag_4, MA_1, MA_2**
RELIANCE	**62.29**	**397.84**	**96.10%**	**62.29**	**Lag_1, Lag_3, MA_1, MA_2**

8. Future Enhancements

- Incorporate macroeconomic time series (GDP, CPI, WPI, repo rates).
- Integrate real-time news sentiment analysis via NLP.
- Use LSTM/Transformer models to capture long-term dependencies.
- Use Ensemble methods to integrate their collective predictive power.
- Simulate trading strategies based on forecast outputs.
- Build a web-based dashboard for visualization and forecasting.
- Explore the dynamics of intraday trading.

---

9. Conclusion

This project demonstrates the practical application of machine learning in stock market forecasting by building time-series-based predictive models for four prominent NIFTY 100 companies — TCS, RELIANCE, AXISBANK, and MRF — spanning key economic sectors. By leveraging advanced regression algorithms like XGBoost and engineering domain-specific features such as lag variables, rolling averages, and return-based metrics, the models effectively captured both short-term price dynamics and longer-term trends. Performance evaluation through visual comparisons and statistical metrics confirmed that the models achieved high predictive accuracy, particularly in capturing directional trends and sector-specific behaviors. The models were devised and trained using the training dataset for the period of March 2022 to August 2024 daily. The trained model was run on the test dataset from August 2024 to March 2025. The project shows that XGBoost is wonderful Machine learning algorithm for forecasting given that feature engineering is done carefully and quite justifies its predictive capabilities in financial forecasting tasks.

While the models performed exceptionally well for relatively stable stocks like AXISBANK and RELIANCE, they were slightly less responsive to high volatility in IT and automobile sectors, highlighting the importance of sector-aware modeling and feature design. Overall, the project not only validated the effectiveness of machine learning for stock price forecasting but also emphasized the value of thoughtful feature engineering and cross-sectoral analysis in financial data science.

---

10. Appendix

10.1 Reference:

- Yahoo Finance API: https://finance.yahoo.com
- yfinance Python library: https://pypi.org/project/yfinance/

- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD*
- Mishkin, F.S. (2018). The Economics of Money, Banking, and Financial Markets.
- Reserve Bank of India (RBI) publications.
- World Bank Development Indicators.
- National Statistical Office (NSO) reports.
- Economic Survey of India (various years).
- https://www.niftyindices.com/indices/equity/broad-based-indices/nifty-100
- https://www.tickertape.in/indices/nifty-100-index-.NIFTY100
- https://groww.in/blog/sensex-and-nifty-live-updates-today-march-11-2025
- https://economictimes.indiatimes.com/markets/indices/nifty-100
- https://files.hdfcfund.com/s3fs-public/Others/2025-01/NFO%20Presentation%20-%20HDFC%20Nifty100%20Quality%2030%20Index%20Fund%20(January%202025).pdf
- https://www.moneyworks4me.com/best-index/nse-stocks/top-nse100-companies-list/
- https://www.screener.in/company/CNX100/
- https://www.nseindia.com/products-services/indices-nifty100-index
- https://in.investing.com/indices/cnx-100-historical-data
- https://pib.gov.in/PressReleasePage.aspx?PRID=2113316
- https://www.miraeassetmf.co.in/docs/default-source/product-guides/note---low-volatilitystrategy-jan-2025.pdf
- https://www.smart-investing.in/indices-bse-nse.php?index=NIFTY100
- https://www.equitymaster.com/research-it/indices/quarterly-results/1-69/nifty-100-index
- https://www.moneycontrol.com/indian-indices/nifty-100-28.html
- https://www.5paisa.com/nifty-100-stock-list
- https://www.nseindia.com/market-data/index-performances
- https://finance.yahoo.com/quote/%5ECNX100/history/
- https://www.icicidirect.com/equity/index/nse/nifty-100/27176
- https://groww.in/indices/nifty-218500
- https://www.thehindubusinessline.com/markets/stock-market-highlights-11-march2025/article69313773.ece
- https://in.investing.com/indices/cnx-100-historical-data

- https://www.nseindia.com/market-data/live-equity-market?symbol=NIFTY+100
- https://www.thehindubusinessline.com/markets/share-market-nifty-sensex-live-updates-25march-2025/article69368985.ece
- https://www.niftyindices.com/reports/historical-data
- https://www.niftyindices.com/indices/equity/broad-based-indices/nifty-100
- https://www.thehindubusinessline.com/markets/stock-market-highlights-28-march2025/article69381849.ece
- https://www.tickertape.in/indices/nifty-100-index-.NIFTY100
- https://www.smart-investing.in/indices-bse-nse.php?index=NIFTY100
- https://www.thehindubusinessline.com/markets/share-market-highlights-27-march2025/article69377127.ece