Agricultural Market Analytics

M.Sc. Agriculture Analytics

Castor Price Forecasting for Patan APMC (Gujarat)

Reported By

Group_no.: 6

Satyam Kumbhar (202319005)
Mitkumar Borda (202319008)
Kaushal Kathiriya (202319013)

Submitted To

Dr. Prity Kumari

Assistant Professor & Head
Department of Basic Science
Anand Agricultural University,Anand
Gujarat - 388 110

INDEX

TABLE AND FIGURES

ABSTRACT

This project focuses on applying advanced time series analysis techniques to forecast the price of castor crops in Patan district, Gujarat, using AGMARKNET data. The study evaluates the performance of multiple methodologies, including ARIMA, SARIMA, ARIMAX, SARIMAX, ARCH, GARCH, VAR, VARMAX, deep learning models like LSTM, and machine learning model like Random Forest. Historical data from 2010 to 2024 was analyzed to uncover trends, seasonality, and volatility in castor prices. A detailed comparison of these models is conducted using accuracy metrics such as AIC Score, BIC Score, MAE, RMSE, and % NRMSE. The research aims to identify the most reliable forecasting approach, aiding stakeholders such as farmers, policymakers, and traders in making informed decisions. The findings contribute valuable insights into market dynamics, enhancing price prediction accuracy and supporting strategic planning for the agricultural sector.

INTRODUCTION

Castor (***Ricinus communis L.***) is probably a native to north-eastern Africa. The castor crop belongs to the family of *Euphorbiaceae* and is developed in tropical and semi-tropical regions. Assisting farmers in their production and marketing decisions through price forecasts will enable them to realize better prices and the price forecast can be used as an extension strategy to achieve the goal of higher income by farmers from these crops. Castor is one of the oldest cultivated crops. However, it contributes to only 0.15 per cent of the vegetable oil produced in the world. The oil produced from this crop is considered to be of importance to the global specialty chemical industry because it is the only commercial source of a hydroxylated fatty acid *(Gohil. et. al (2023))*[1].

Agricultural price forecasting plays a pivotal role in reducing the inherent risks associated with farming by enabling producers to make informed decisions regarding crop production, storage, and marketing. However, price fluctuations due to factors like weather variability, market dynamics, export demand, and global economic conditions make it challenging to achieve reliable forecasts. Time series analysis has emerged as a robust approach to address this issue, offering methodologies to model and predict complex price patterns over time.

Castor's economic viability is heavily influenced by price trends in both domestic and international markets. Studies have shown that price volatility significantly affects farmers' incomes, especially in regions like Gujarat, where castor is a primary crop. For instance, research demonstrated how futures trading impacts price trends, highlighting the role of market forces, export demand, and government policies.(*Bansal et. al (2015)*)[2] Similarly, emphasized the practical benefits of price forecasts in enabling farmers to optimize their returns by delaying sales during low-price periods. (*Dhandhalya et al (2016)*)[3].

The district of Patan, a major castor-producing region, has exhibited unique price patterns characterized by high instability. Such volatility underscores the need for tailored forecasting models to accommodate regional nuances in price behavior. Time series analysis, with its capacity to incorporate seasonality, trends, and external factors, provides a robust framework for addressing these challenges *(Gohil et al. (2023))*[1].

Time series analysis involves the statistical study of sequentially ordered data to identify patterns and forecast future values. In agricultural economics, models such as ARIMA (Auto-Regressive Integrated Moving Average), SARIMA (Seasonal ARIMA), ARCH (Auto-Regressive Conditional Heteroskedasticity), and GARCH (Generalized ARCH) have been widely applied to predict price trends. ARIMA models are particularly notable for their ability to handle non-stationary data by differencing, as demonstrated by *Dhandhalya et al. (2016)*[3] in their study of castor prices in Gujarat.

Advanced methodologies like ARIMAX and SARIMAX extend the capabilities of ARIMA by incorporating external predictors, such as weather variables or export data, to enhance accuracy. Meanwhile, ARCH and GARCH models have proven effective in capturing price volatility, as evidenced persistent volatility in castor's spot and futures markets (*Bansal et. al (2015)*)[2].

The castor market in Gujarat is characterized by its reliance on both domestic demand and international

export trends. Patan district, known for its significant castor production, presents a complex price behavior influenced by seasonal indices, climatic conditions, and market dynamics. The importance of incorporating regional factors, such as local supply-demand imbalances and storage practices, into forecasting models *(Gohil et al. (2023))*[1].

This project applies a comprehensive suite of time series techniques to forecast castor prices in Patan using data from AGMARKNET (2010–2024). By employing ARIMA, ARIMAX, SARIMA, SARIMAX, ARCH, GARCH, VAR, VARMAX and deep learning model LSTM, the study seeks to identify the most accurate forecasting approach. Performance will be evaluated using metrics such as AIC Score, BIC Score, MAE, RMSE, and % RMSE, ensuring robust model validation. The integration of external predictors, such as weather variables or export data, further enhances the practical utility of these forecasts for stakeholders.

Effective price forecasting empowers farmers to optimize marketing strategies, reducing income instability and fostering better resource allocation. For policymakers, accurate predictions facilitate the development of targeted interventions to stabilize markets and ensure food security. By focusing on the castor crop in Patan district, this project contributes to the growing body of research that leverages time series methodologies to address the challenges of agricultural price forecasting. Furthermore, the comparative evaluation of traditional and machine learning models provides valuable insights into their relative strengths and applications, paving the way for more refined forecasting tools in the future.

In conclusion, this study underscores the critical role of time series analysis in mitigating the risks of price volatility in Gujarat's castor markets. By integrating statistical rigor with advanced computational techniques, it aims to deliver actionable insights that enhance the economic resilience of castor growers and stakeholders alike.

REVIEW OF LITERATURE

I. Time Series Analysis of Castor Crop for Price Forecasting in Gujarat

The study by *Gohil et al. (2023)*[1] evaluated castor price forecasting across six districts of Gujarat, including Patan. Using time series data (2007–2021), ARIMA models were applied to analyze seasonal indices and volatility. Results showed Patan had the highest instability (1.903) and significant seasonal price variations, with March witnessing the highest prices. The ARIMA (1,0,1) model was identified as best for Patan, demonstrating high precision based on metrics like RMSE and MAPE. This study highlighted the importance of tailored forecasting models for specific districts to address unique market dynamics

II. Growth and Volatility in Cash and Futures Market of Castor in India

Bansal and Zala (2015)[2] investigated the growth and volatility in castor's spot and futures markets from 1994 to 2013. They observed significant growth in wholesale price indices during the post-futures trading period, driven by demand, export patterns, and government policies. The ARCH and GARCH models indicated persistent volatility in spot and futures prices, suggesting that futures trading plays a crucial role in price discovery and risk management. The study emphasized the relevance of integrating futures trading insights into forecasting methodologies

III. Benefits of Price Forecast to Castor Growers in Gujarat

Dhandhalya et al. (2016)[3] examined the impact of price forecasts on castor farmers in Gujarat. By applying ARIMA (0,1,2) to Patan APMC data (1990–2013), they forecasted prices for March–May 2014 and advised farmers to delay sales for better returns. Farmers who followed the forecast achieved an incremental income of ₹35,826 per hectare. The study demonstrated the tangible benefits of price forecasts in improving farmer incomes and underscored the critical role of ARIMA models in agricultural decision-making

DATASET

Table 1 - Data Overview

	MARKET: A	MARKET: B
Commodity	Castor	Castor
State	Gujarat	Gujarat
District	Patan	Patan
Market	Patan	Siddhpur
Time period	1 Jan 2010 to 20 Nov 2024	1 Jan 2010 to 20 Nov 2024
Price/ Arrival	Both	Both
Data available	3902 Rows	3001 Rows
Minimum Price (Rs/Quintal)	2450	2685
Maximum Price (Rs/Quintal)	7615	7607

Preview of Dataset :

Table 2 - Preview and statistics of Dataset

Date	Arrival (tonnes)	Price (Rs/Quintal)
2010-01-01 00:00:00	84.8	2755
2010-01-02 00:00:00	75.8	2750
2010-01-03 00:00:00	80.3	2752.5
2010-01-04 00:00:00	87.4	2750
2010-01-05 00:00:00	122.1	2770
2010-01-06 00:00:00	80.3	2740
2010-01-07 00:00:00	74.3	2750
2010-01-08 00:00:00	110.1	2700
2010-01-09 00:00:00	122	2750
2010-01-10 00:00:00	99.36	2743.33
2010-01-11 00:00:00	163.7	2750
2010-01-12 00:00:00	108.5	2700
2010-01-13 00:00:00	133.2	2725
2010-01-14 00:00:00	127.5	2725
2010-01-15 00:00:00	116.9	2760

Castor Crop	Minimum	Maximum	Average	Standard Deviation
Price (Rs/Quintal)	2450	7615	4505.713616	1161.928912
Arrival (Tonnes)	0.23	2241.8	328.8302415	296.9375234

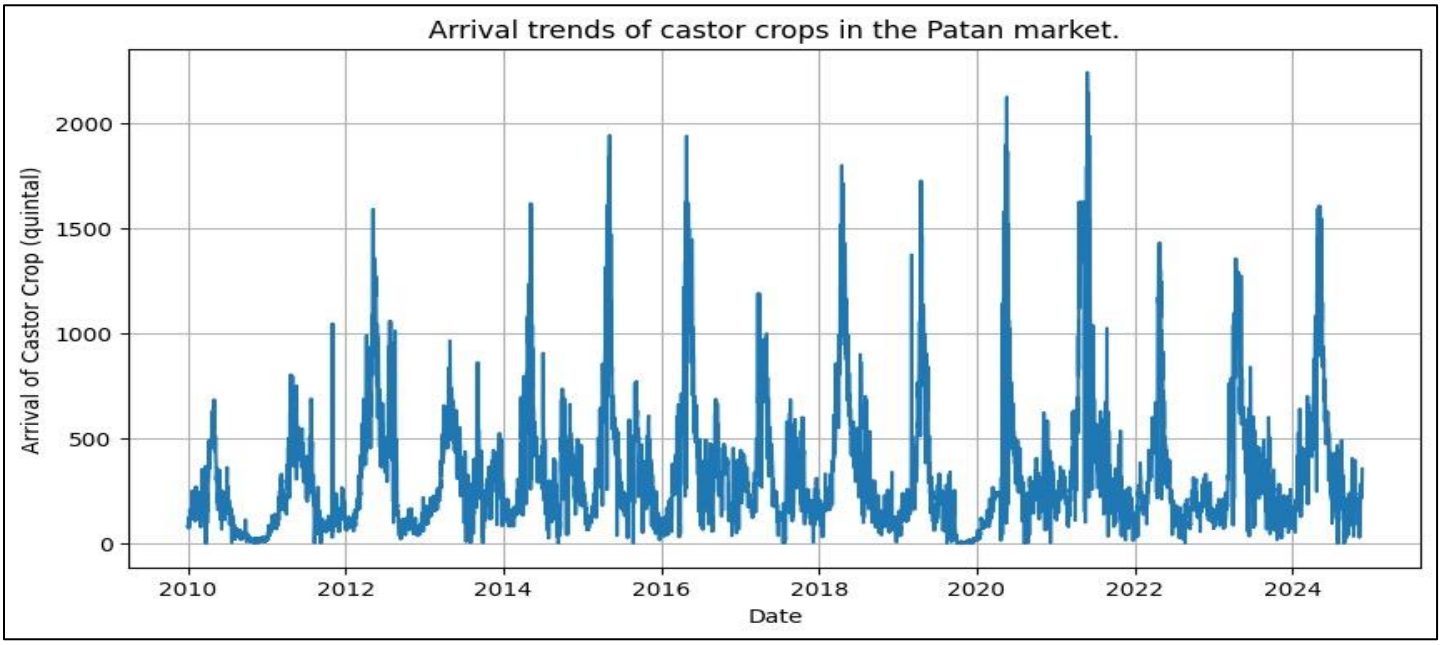


Figure 1 : Arrival trends of castor crop in Patan market

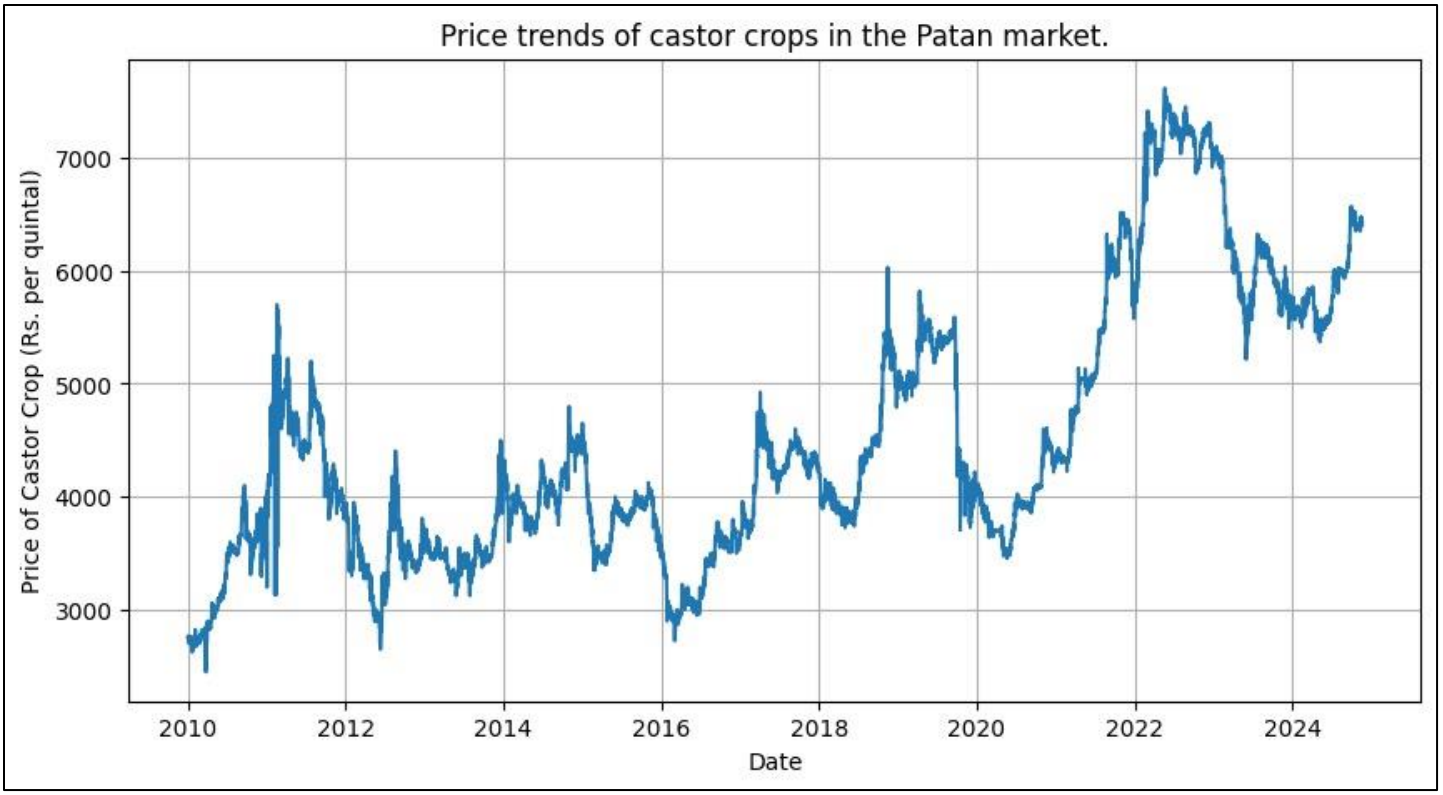
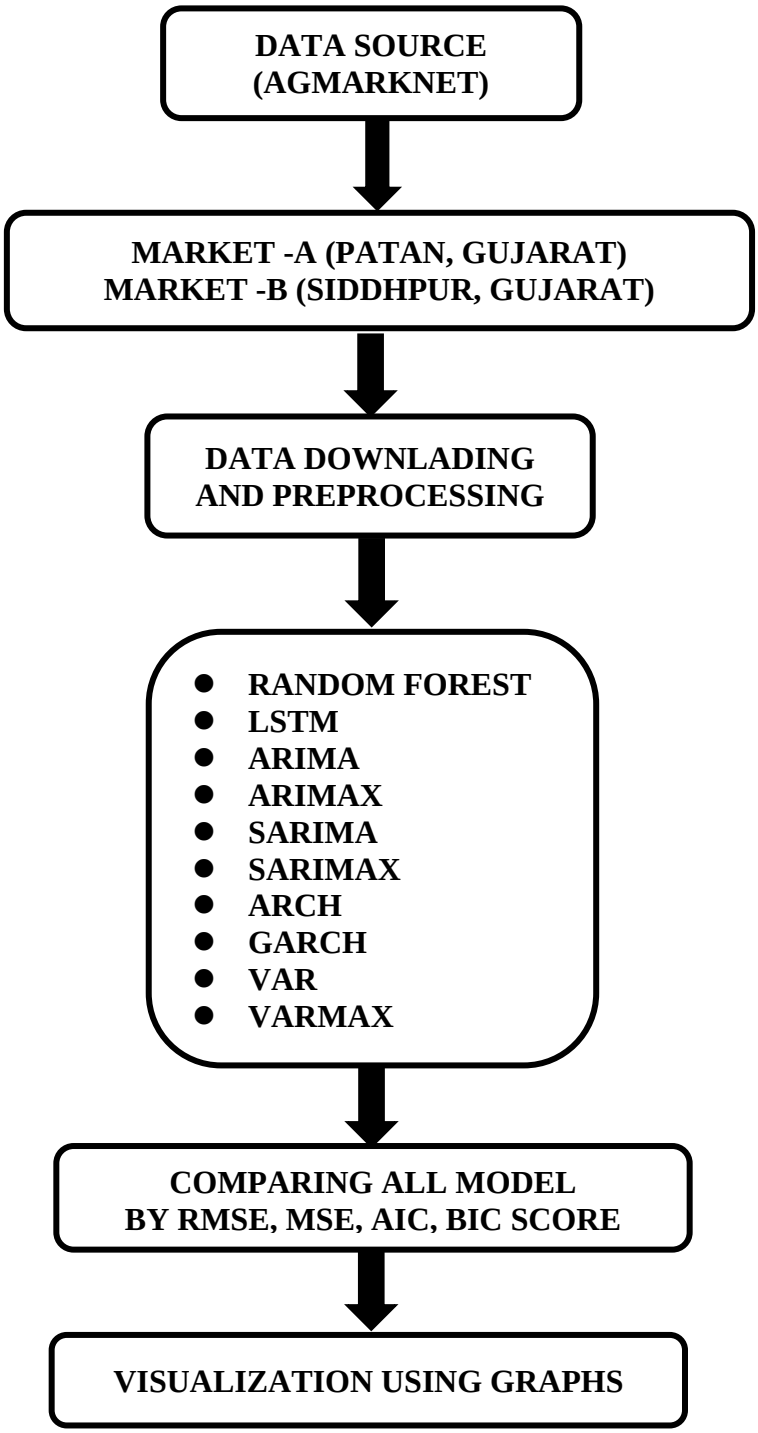


Figure 2 : Price trends of castor crop in Patan market

FLOWCHART


DATA SOURCE
(AGMARKNET)
MARKET -A (PATAN, GUJARAT)
MARKET -B (SIDDHPUR, GUJARAT)
DATA DOWNLADING
AND PREPROCESSING
RANDOM FOREST
LSTM
ARIMA
ARIMAX
SARIMA
SARIMAX
ARCH
GARCH
VAR
VARMAX
COMPARING ALL MODEL
BY RMSE, MSE, AIC, BIC SCORE
VISUALIZATION USING GRAPHS

METHODOLOGY

The methodology for castor price prediction integrates data preprocessing, exploratory data analysis, and a suite of statistical and machine learning models to identify trends, seasonality, volatility, and external influences. The stepwise approach ensures a comprehensive analysis and robust forecasting of potato prices in Patan, Gujarat. Below is the methodology with explanations of the models used.

1. Data Collection

- **Source :** Agmarknet, a repository for agricultural market data.
- **Data Type :** Historical castor price data for Patan, Gujarat.

2. Data Preprocessing

- Cleaning: Handle missing values, outliers, and duplicates.
- Stationarity Testing: Use methods like the Augmented Dickey-Fuller (ADF) test to ensure the time series data is stationary.
- Feature Engineering: Create lag variables, incorporate weather data, and add seasonal indices to enrich the dataset.

3. Exploratory Data Analysis (EDA)

Visualize trends, seasonality, and patterns in the data using line plots and decomposition methods.

4. Model Selection :

The following models were considered for forecasting :

4.1) ARIMA Model :

An ARIMA (**Autoregressive Integrated Moving Average Model**) is a form of regression analysis that gauges the strength of one dependent variable relative to other changing variables. It is invented by **George Box and Gwilym Jenkins.** The model's goal is to predict future securities or financial market moves by examining the differences between values in the series instead of through actual values.

An ARIMA model can be understood by outlining each of its components as follows:

- **Autoregression (AR) :** refers to a model that shows a changing variable that regresses on its own lagged, or prior, values.
- **Integrated (I) :** represents the differencing of raw observations to allow the time series to become stationary (i.e., data values are replaced by the difference between the data values and the previous values).
- **Moving average (MA) :** incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Each component in ARIMA functions as a parameter with a standard notation. For ARIMA models, a standard notation would be ARIMA with p, d, and q, where integer values substitute for the parameters to indicate the type of ARIMA model used. The parameters can be defined as:

- **p :** the number of lag observations in the model, also known as the lag order.
- **d :** the number of times the raw observations are differenced; also known as the degree of differencing.
- **q :** the size of the moving average window, also known as the order of the moving average [12][14].

Equation -

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t$$

Where,

- $\mathbf{y_t}$: The observed value of the time series at time t.
- $\boldsymbol{\mu}$: The mean (constant) term. It represents the long-term mean level of the series
- $\boldsymbol{\phi_i}$: The autoregressive coefficients
- $\mathbf{y_{t-i}}$: The past p values of the time series
- $\boldsymbol{\theta_j}$: The moving average coefficients
- $\boldsymbol{\epsilon_{t-j}}$: Past q error terms (or residuals)

Methodology:

1. Conduct ADF test to check stationarity.
2. Apply differencing if necessary to remove trends or seasonality.
3. Use PACF to determine the order of "p" and ACF for "q."
4. Fit the ARIMA model using Maximum Likelihood Estimation (MLE).
5. Compare models using AIC and BIC values to identify the optimal parameters.
6. Validate the model through residual diagnostics and forecasting accuracy.

4.2) ARIMAX Model :

An ARIMAX model, which stands for **AutoRegressive Integrated Moving Average with exogenous Variables**, is an advanced version of the ARIMA model. The ARIMAX model extends the ARIMA framework by integrating exogenous variables, which are external factors that can influence the time series being studied. This integration allows the model to leverage additional information that can significantly enhance forecasting accuracy.

The ARIMAX model retains ARIMA's core structure (p, d, q) but introduces a regression component for exogenous variables.

Exogenous Variables (X) are external predictors or factors not part of the time series but may have a significant impact on it. By incorporating these variables, the ARIMAX model can provide a more comprehensive analysis and better forecasting performance.

Key features include:

- Accounts for the influence of external factors on the dependent time series.
- Suitable for time series where predictors significantly impact future values.
- Requires careful selection of exogenous variables to avoid overfitting [11].

Equation -

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \sum_{k=1}^{m} \beta_k X_{t-k} + \epsilon_t$$

Where,

- $\mathbf{yt}$: The observed value of the time series at time t
- $\boldsymbol{\mu}$: The constant (intercept) term, representing the mean level of the series
- $\mathbf{p}$: The order of the AR part, indicating how many past values of yt are used
- $\boldsymbol{\phi_i}$: Coefficients of the autoregressive terms
- $\mathbf{y_{t-i}}$: Past values of the time series

- **q**: The order of the MA part, indicating how many past error terms are included
- **θ_j**: Coefficients of the moving average terms
- **ϵ_{t-j}**: Past error terms (residuals)
- **X_{t-k}**: External variables (predictors) that are not part of the time series but influence it
- **β_k**: Coefficients that measure the impact of the exogenous variables on yt
- **m**: The number of lags for the exogenous variables

Methodology:

1. Identify and preprocess exogenous variables.
2. Include these variables as predictors in the ARIMA model.
3. Fit the ARIMAX model and evaluate its performance using AIC, BIC, and residual diagnostics.

4.3) SARIMA Model :

SARIMA, which stands for **Seasonal Autoregressive Integrated Moving Average**, is a versatile and widely used time series forecasting model. It's an extension of the non-seasonal ARIMA model, designed to handle data with seasonal patterns. SARIMA captures both short-term and long-term dependencies within the data, making it a robust tool for forecasting. It combines the concepts of autoregressive (AR), integrated (I), and moving average (MA) models with seasonal components.

It is particularly effective for time series with recurring patterns at regular intervals, such as monthly crop prices or quarterly sales. SARIMA models are defined by six parameters: (p, d, q) for the non-seasonal component and (P, D, Q, m) for the seasonal component, where:

P, D, Q : Represent seasonal autoregressive, differencing, and moving average terms.

m: Indicates the seasonal period (e.g., 12 for monthly data).

SARIMA uses seasonal differencing to handle periodic fluctuations, making it a robust choice for forecasting time series with both trend and seasonality. Proper model selection involves analyzing seasonal and non-seasonal patterns through ACF and PACF plots [11].

Equation -

$$\Phi_p(B^s) \cdot (1-B)^d y_t = \mu + \Theta_q(B^s) \cdot \epsilon_t$$

Where,

- B = Backshift operator
- Φ_p = Coefficient for seasonal autoregressive (SAR) term
- $(1-B)^d$= Differencing operator of order dd for stationarity
- y_t= Time series value at time t
- μ = Mean of the time series
- Θ_q = Coefficient for seasonal moving average (SMA) term
- ϵ_t= White noise (error term)
- s = Seasonal period

Methodology:

1. Identify seasonality in the data using seasonal plots or decomposition.
2. Perform differencing to remove seasonality.
3. Extend ARIMA by including seasonal terms (P, D, Q, m).
4. Fit the SARIMA model and tune parameters using AIC and BIC.
5. Validate results using residual analysis and forecast performance.

4.4) SARIMAX Model :

The **Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors** (SARIMAX) model is a powerful time series forecasting technique that extends the traditional SARIMA model to account for seasonality and external factors. It's a versatile model that can accommodate both autoregressive (AR) and moving average (MA) components, integrate differencing to make the data stationary, and incorporate external variables or regressors. SARIMAX is particularly valuable when dealing with time-dependent data that exhibits recurring patterns over specific time intervals.

The model's structure is defined by parameters (p, d, q, P, D, Q, m) for seasonal and non-seasonal components, along with regression terms for exogenous variables. SARIMAX is particularly effective in agricultural or climate-sensitive contexts where seasonality and external factors like weather or market conditions drive trends [11].

Equation -

$$\Phi_p(B^s) \cdot (1-B)^d y_t = \mu + \Theta_q(B^s) \cdot \epsilon_t + \sum_{k=1}^{m} \beta_k X_{t-k}$$

Where,

- **B**: Backshift operator
- **s**: Seasonal period
- $\boldsymbol{\Phi_i}$: Coefficients of the seasonal autoregressive terms
- **d**: Order of differencing to remove non-seasonal trends and make the series stationary
- $\boldsymbol{\Theta_j}$: Coefficients of the seasonal moving average terms
- $\boldsymbol{X_{t-k}}$: External predictors (variables) that influence the dependent variable yt
- $\boldsymbol{\beta_k}$: Coefficients representing the impact of the exogenous variables
- **m**: Number of lags for the exogenous variables

Methodology:

1. Identify seasonality and external variables.
2. Combine ARIMA components with seasonal and exogenous terms.
3. Fit SARIMAX using appropriate (p, d, q) and (P, D, Q, m) values.
4. Validate the model with residual diagnostics and performance metrics.

4.5) ARCH Model :

The **autoregressive conditional heteroskedasticity** (ARCH) model was designed to improve econometric models by replacing assumptions of constant volatility with conditional volatility. Engle and others working on ARCH models recognized that past financial data influences future data—that is the definition of autoregressive.

Key features :

- Effective for modeling "clustering" of volatility (e.g., periods of high price volatility followed by calm periods).
- Useful in identifying risk and uncertainty in time series data.

ARCH models are widely applied in economic and financial data analysis, including forecasting crop price volatility [12][16].

Equation -

$$y_t = \mu + \epsilon_t$$

$$\epsilon_t = \sigma_t \cdot z_t$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2$$

Where,

- $\mathbf{y_t}$: The observed time series value at time t
- $\boldsymbol{\mu}$: The constant (mean) term, representing the average value of the time series
- $\boldsymbol{\epsilon_t}$: The error term or innovation at time ttt, representing deviations from the mean
- $\boldsymbol{\sigma_t}$: The **conditional standard deviation** of ϵt
- $\mathbf{z_t}$: A sequence of i.i.d. standard normal random variables (zt~N(0,1))
- $\boldsymbol{\sigma_t^2}$: The conditional variance of the error term at time t
- $\boldsymbol{\alpha_0}$: A constant term, representing the base level of variance
- $\boldsymbol{\alpha_i}$: Coefficients for the lagged squared error terms
- **q**: The order of the ARCH model, indicating how many past squared residuals are included in the model

Methodology:

1. Test for heteroskedasticity in residuals from ARIMA and SARIMA.
2. Fit the ARCH model to capture variance dynamics.
3. Validate the model with diagnostic tests and analyze volatility predictions.

4.6) GARCH Model:

GARCH (Generalized AutoRegressive Conditional Heteroskedasticity) extends ARCH by including past variances in addition to past error terms. Proposed by Bollerslev (1986), GARCH models capture long-term and short-term volatility dynamics, making them more robust for time series with persistent volatility.

Key features:

- Incorporates lagged variance terms to better model volatility.
- Commonly used in financial markets and commodity price analysis.

The model's effectiveness relies on parameters (p, q), representing the order of autoregressive and moving average terms in variance equations [12][16].

Equation -

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$$

Where,

- $\boldsymbol{\sigma_t^2}$: Conditional variance at time t
- $\boldsymbol{\omega}$: Constant term, providing a base level of variance
- $\boldsymbol{\alpha_i}$: Coefficients for the **lagged squared residuals**
- $\boldsymbol{\beta_j}$: Coefficients for the **lagged conditional variances**
- **q**: Order of the ARCH part, representing how many past squared errors are included
- **p**: Order of the GARCH part, representing how many past conditional variances are included

Methodology:

1. Fit GARCH to residuals from ARIMA/SARIMA models.

2. Optimize parameters for both ARCH and GARCH terms.
3. Validate the model using diagnostic tests and volatility forecasts.

4.7) VAR Model :

A **vector autoregression** (VAR) model is a multivariate time series model containing a system of n equations of n distinct, stationary response variables as linear functions of lagged responses and other terms. VAR models are also characterized by their degree p; each equation in a VAR(p) model contains p lags of all variables in the system. Unlike ARIMA, which focuses on a single series, VAR treats all variables as endogenous, making it suitable for analyzing complex systems like economic or market interactions.

Key features:

- Models dynamic relationships among variables without requiring external predictors.
- Requires stationary data, achieved through differencing or detrending.
- Applied in economic policy analysis, financial forecasting, and agriculture for examining interrelated market trends [9][15].

Equation -

$$Y_t = \mu + \sum_{i=1}^{p} A_i Y_{t-i} + \epsilon_t$$

Where,

- **μ**: A vector of constant terms
- $\sum_{i=1}^{p} A_i Y_{t-i}$**:** Lagged values of Yt
- ϵ_t**:** A vector of error terms

Methodology:

1. Identify the variables to include in the model and ensure that all are stationary using tests like ADF.
2. Determine the optimal lag order (p) using information criteria such as AIC or BIC.
3. Fit the VAR model with the selected lag order.
4. Perform Granger causality tests to analyze the causal relationships among variables.
5. Validate the model by examining residuals and forecasting accuracy.

4.8) VARX Model :

The VARX (**Vector Autoregressive Model with Exogenous Variables**) is an extension of the VAR model that incorporates exogenous variables, allowing for the analysis of systems where some variables are influenced by others. It is particularly useful in econometrics for modeling and forecasting multivariate time series data. The model is specified as VARX(p,s), where pp denotes the number of lags of the endogenous variables, and ss indicates the number of lags of the exogenous variables. This framework facilitates efficient estimation and testing for cointegration among variables, essential for understanding long-run relationships in economic data [9][17].

Equation -

$$Y_t = \mu + \sum_{i=1}^{p} A_i Y_{t-i} + \sum_{j=1}^{q} B_j \epsilon_{t-j} + \sum_{k=1}^{m} C_k X_{t-k} + \epsilon_t$$

Where,

- σ_t^2: Conditional variance at time t
- **ω:** Constant term
- α_i: Coefficients for the lagged squared residuals

- β_j: Coefficients for the lagged conditional variances
- **q**: Order of the ARCH (Autoregressive Conditional Heteroskedasticity) component
- **p:** Order of the GARCH (Generalized ARCH) component

Methodology:

1. Preprocess the data and ensure all variables (endogenous and exogenous) are stationary.
2. Determine the optimal lag order for endogenous variables and the inclusion of exogenous variables using AIC or BIC.
3. Fit the VARMAX model by including exogenous variables as additional regressors.
4. Evaluate the model using residual diagnostics and forecasting performance.
5. Use the model to make multivariate forecasts and assess how exogenous variables influence endogenous variables.

4.9) LSTM Model :

Long Short-Term Memory (LSTM) is a specialized type of recurrent neural network (RNN) designed to address the vanishing gradient problem, enabling it to learn long-term dependencies in sequential data. Developed by Hochreiter and Schmidhuber, LSTMs incorporate a memory cell and three gates: the input gate, forget gate, and output gate. These gates regulate the flow of information, allowing the network to selectively retain or discard data over time. LSTMs are widely used in applications such as speech recognition, language modeling, and time series prediction due to their effectiveness in handling sequences of varying lengths [11].

LSTM is a type of recurrent neural network (RNN) designed to capture long-term dependencies in sequential data. Its core mechanism relies on three gates: **forget gate, input gate, and output gate**.

Equation –

Forget Gate

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$

Input Gate

$$it = \sigma W i \cdot \mathrm{h} t-1, xt+bi$$

Cell Candidate

$$\mathrm{C}t = \tanh W C \cdot \mathrm{h} t-1, xt+bC$$

Cell State Update

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C_t}$$

Output Gate

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

Hidden State Update

$$h_t = o_t \cdot \tan \mathrm{h}(C_t)$$

Where,

- **σ:** sigmoid activation function
- $\mathbf{W_f}$**:** weight matrix associated with the forget gate
- $\mathbf{h_{t-1}}$**:** previous hidden state (output from the previous time step)
- $\mathbf{x_t}$**:** input at time t

- **b_f:** bias term for the forget gate
- **W_i:** weight matrix associated with the input gate
- **b_i:** bias term for the input gate
- **i_t:** output of the input gate
- **tanh:** hyperbolic tangent activation function
- **WC:** weight matrix for the cell candidate
- **bC:** bias term for the cell candidate
- **C_{t-1}:** previous cell state
- **W_o:** weight matrix for the output gate
- **b_o:** bias term for the output gate

Methodology:

1. Normalize data to accelerate model convergence.
2. Prepare data into sequences (X, y) for supervised learning.
3. Build the LSTM model with input, hidden, and output layers.
4. Compile the model with appropriate loss function and optimizer.
5. Train the model on the training dataset and validate it using test data.
6. Evaluate performance using metrics like MSE and MAE and visualize predictions.

4.10) Random Forest :

Random Forest is an ensemble learning method used for classification and regression tasks, which constructs multiple decision trees during training. The algorithm combines the predictions of these trees to improve accuracy and stability. Each tree is built using a random subset of data and features, reducing overfitting and enhancing generalization to new data. The final prediction is made by averaging the outputs for regression or by majority voting for classification. Random Forest is favored for its robustness, versatility, and ability to handle large datasets with high-dimensional spaces effectively [18].

Equation -

$$\boldsymbol{Gini = 1 - \sum_{i=1}^{C} p_i^2}$$

$$\widehat{y_t} = \frac{1}{T}\sum_{i=1}^{T} f_i(x_t)$$

Where,

- **C:** number of classes
- **p_i:** probability (or proportion) of class iii in the dataset
- **y_t:** The predicted or estimated value of yt at time t
- **T**: The total number of predictions or models being aggregated
- **$f_i(x_t)$**: The prediction made by the i-th model (or function) at time t using the input xt

Methodology:

1. Preprocess data by handling missing values and scaling.
2. Split data into training and testing sets.
3. Train the Random Forest regression model.
4. Evaluate model performance using R^2, MSE, and MAE metrics.

5. Model Training and Validation

- Split data into training and testing sets.
- Train each model using historical data.
- Validate performance using metrics such as:
- Root Mean Squared Error (RMSE): Measures the average magnitude of prediction errors.
- Mean Squared Error (MSE): Quantifies the mean of squared errors.
- Percentage Error (% Error): Reflects prediction error as a percentage.

6. Model Comparison

Evaluate the performance of all models.
Select the best-performing model(s) based on accuracy and interpretability.

7. Price Prediction

Use the chosen model(s) to forecast future potato prices.
Generate forecasts under various scenarios, incorporating external factors where applicable.

8. Visualization and Reporting

Present results using visual aids such as graphs and charts.
Highlight insights, trends, and recommendations for stakeholders.

RESULTS

I) ARIMA Model :

The model is best fitted at ARIMA (0,1,3).

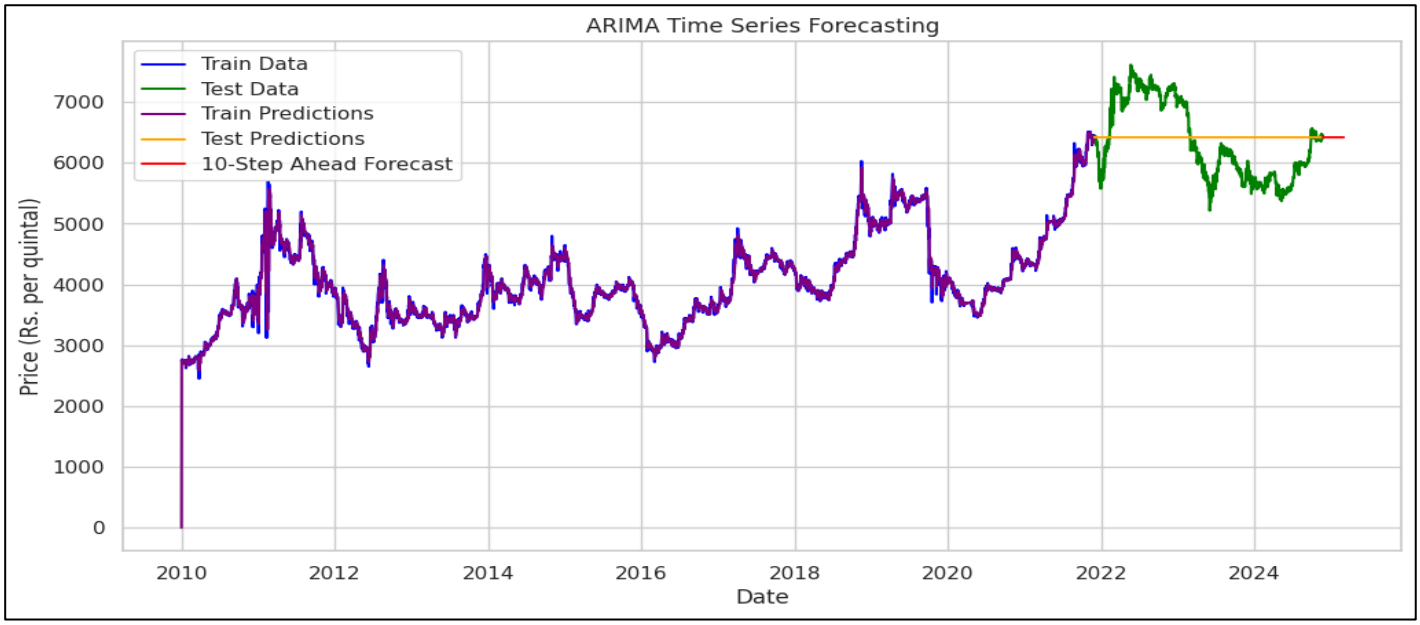


Figure 3 : ARIMA Time Series Forecasting

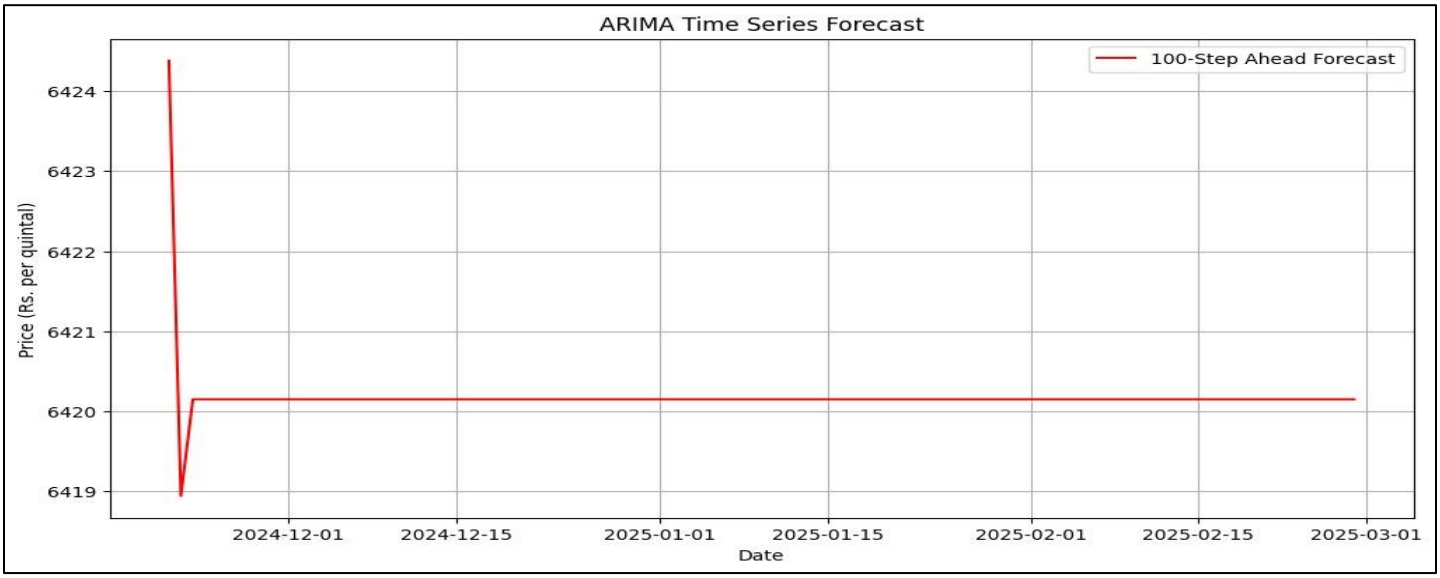


Figure 4 : ARIMA forecast

II) ARIMAX Model :

The model is best fitted at ARIMA (2,1,4).

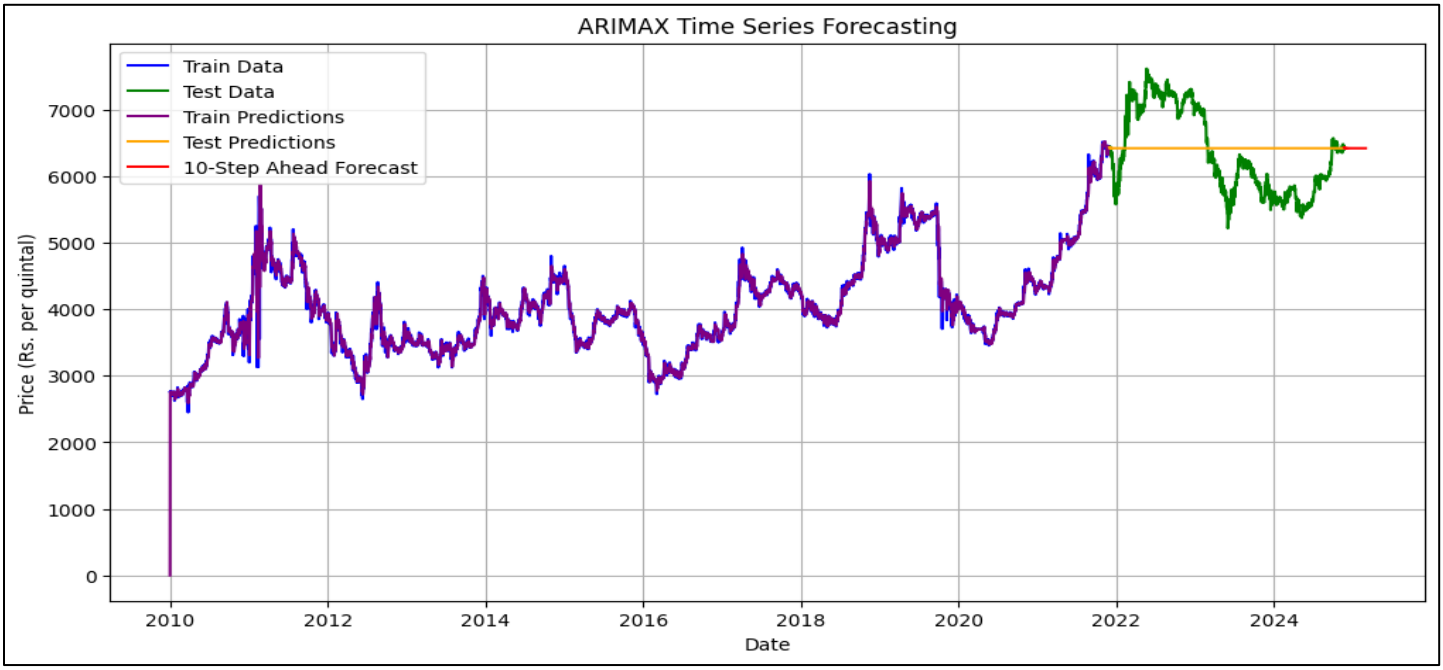


Figure 5 : ARIMAX Time Series Forecasting

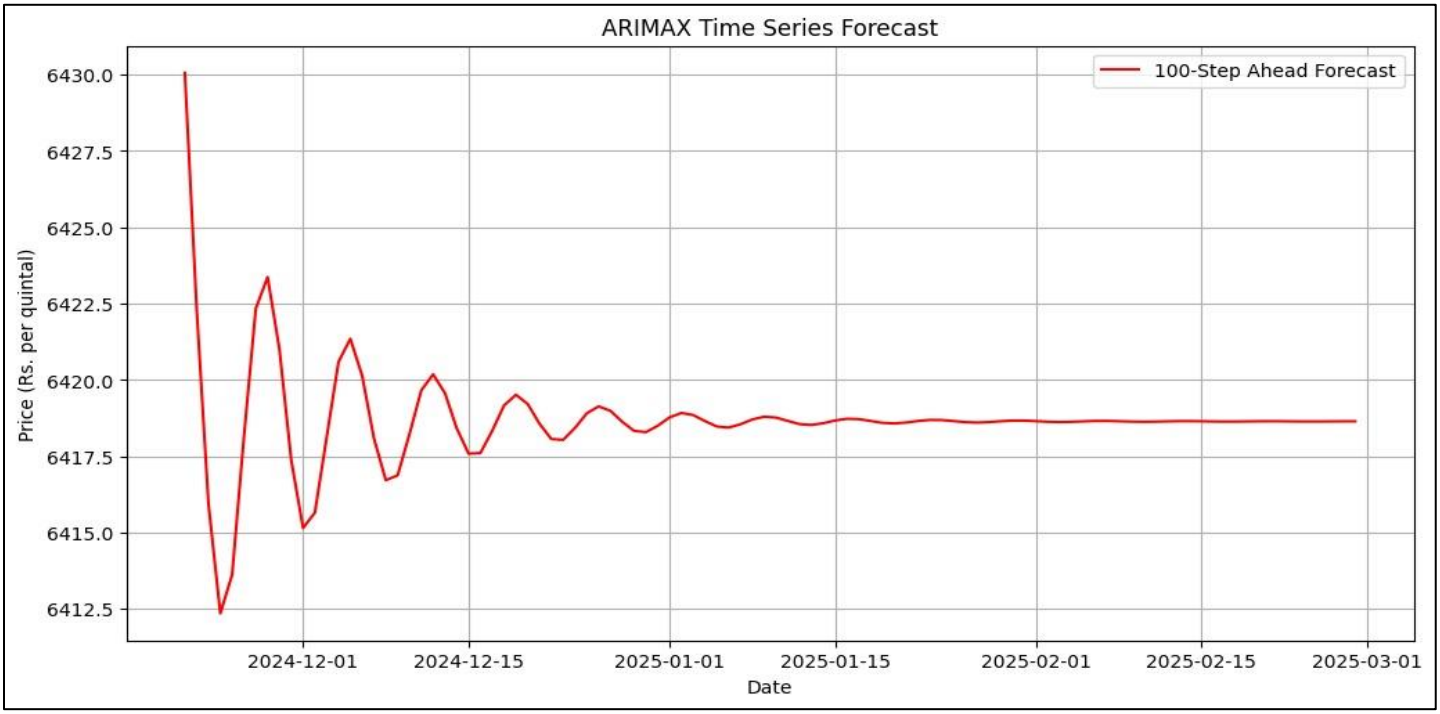


Figure 6 : ARIMAX forecast

III) SARIMA Model :

The model is best fitted at ARIMA (0,1,1)(2,0,2)[7].

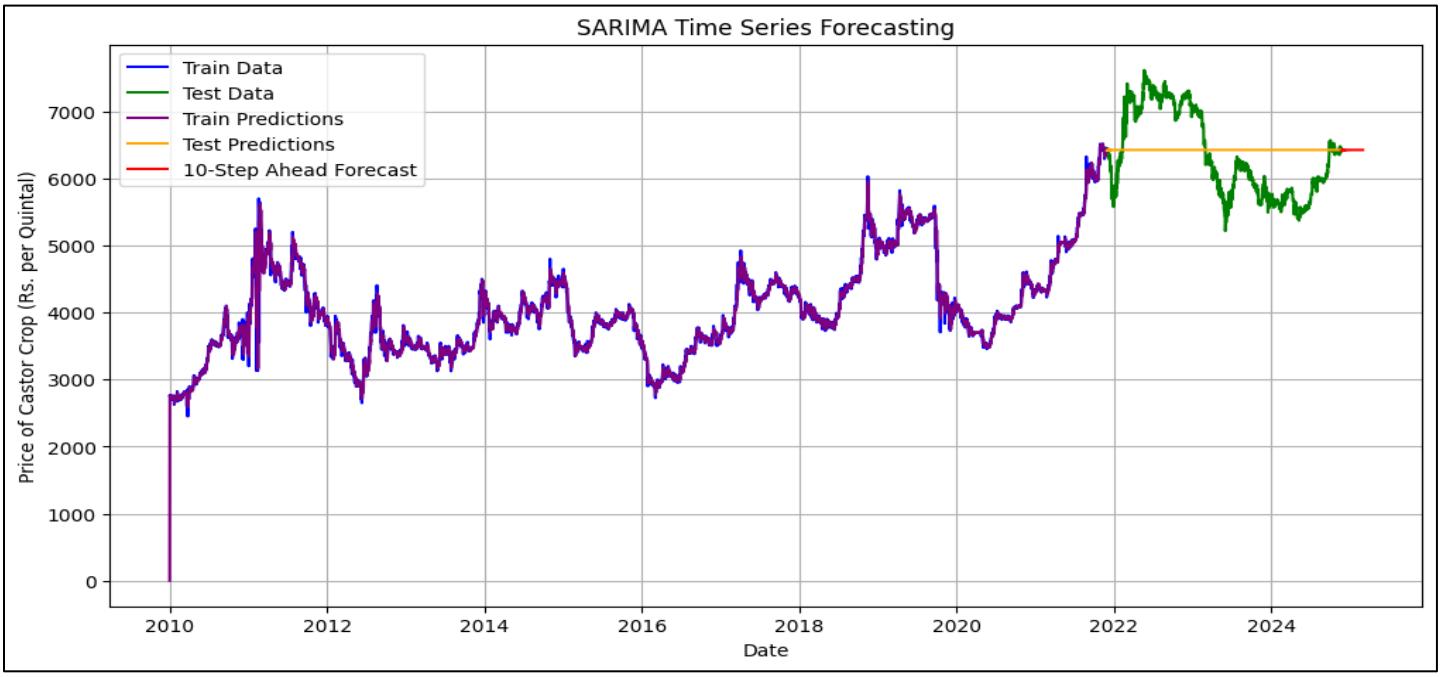


Figure 7 : SARIMA Time Series Forecasting

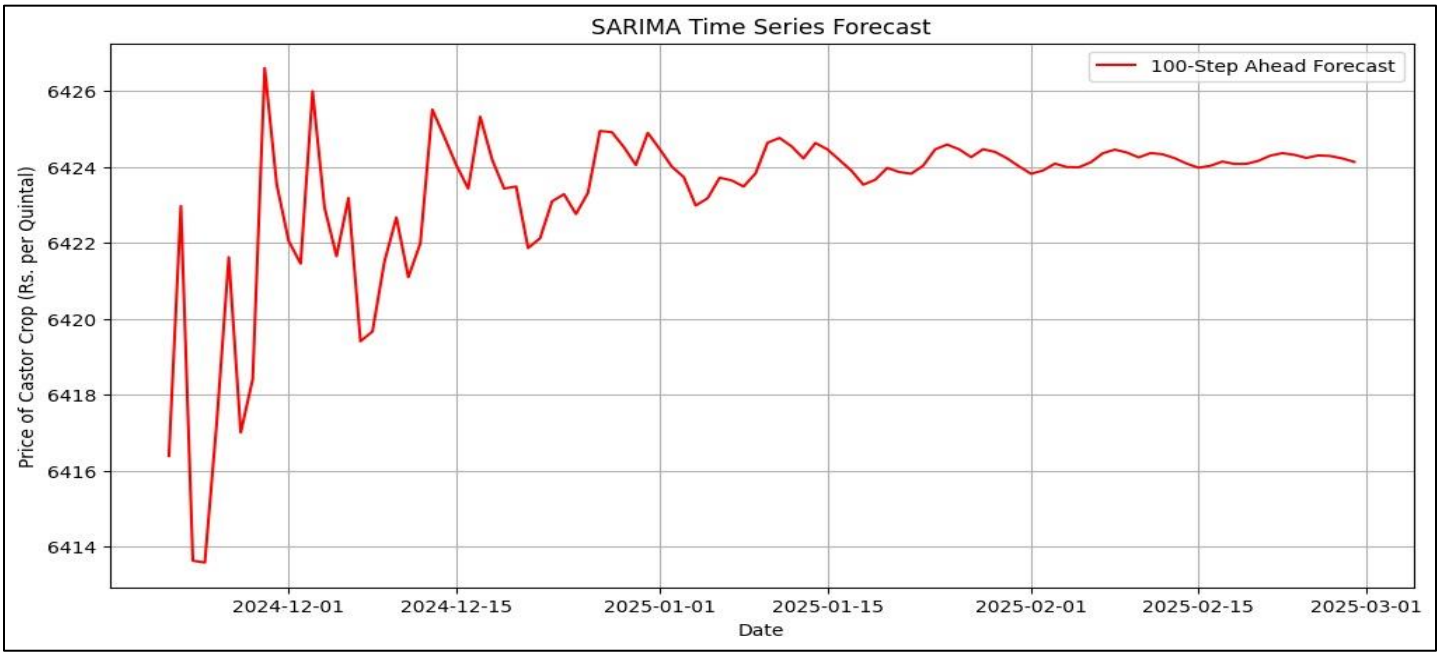


Figure 8 : SARIMA forecast

IV) SARIMAX Model :

The model is best fitted at ARIMA (3,1,0)(3,1,0)[7].

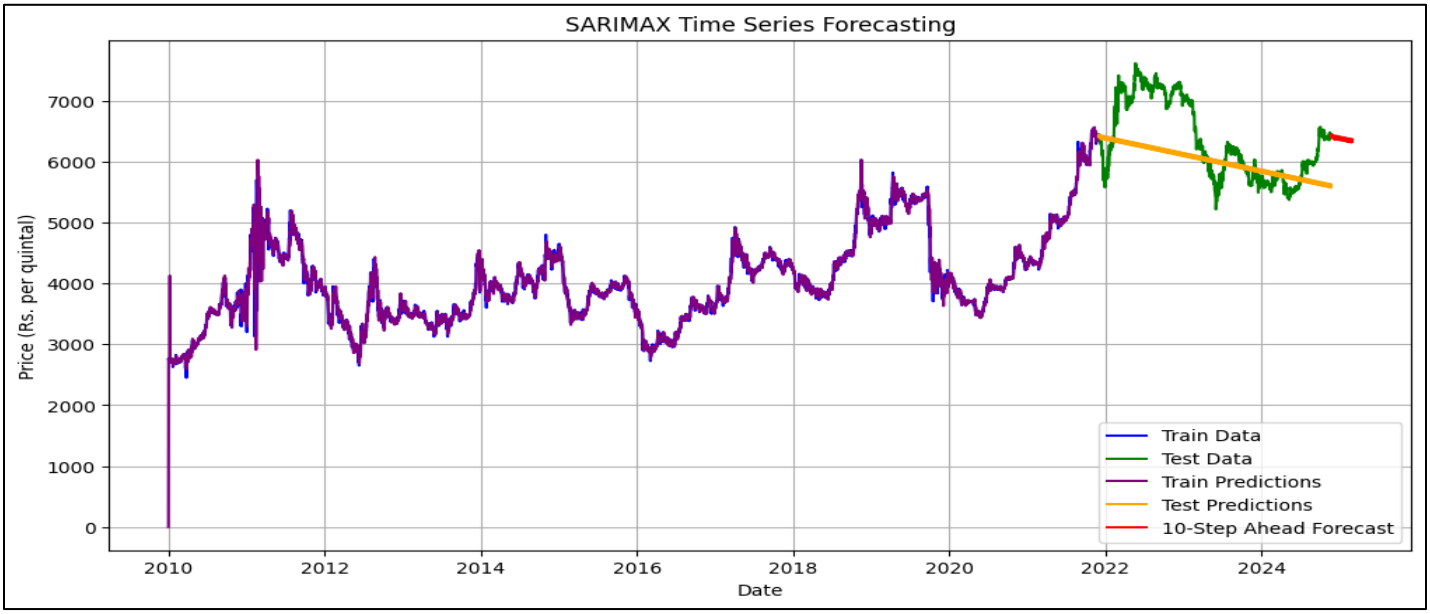


Figure 9 : SARIMAX Time Series Forecasting

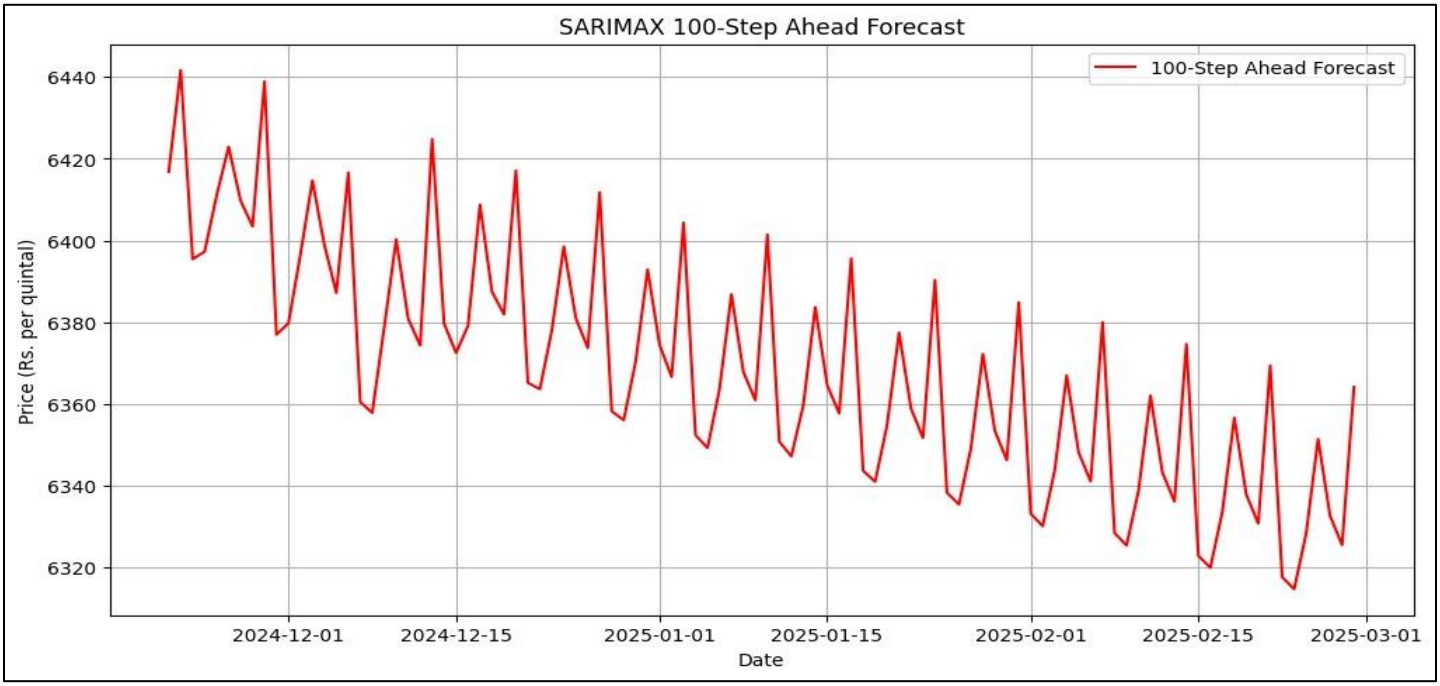


Figure 10 : SARIMAX forecast

V) ARCH Model :

The model is best fitted at ARIMA (3,1,0)(3,1,0)[7].

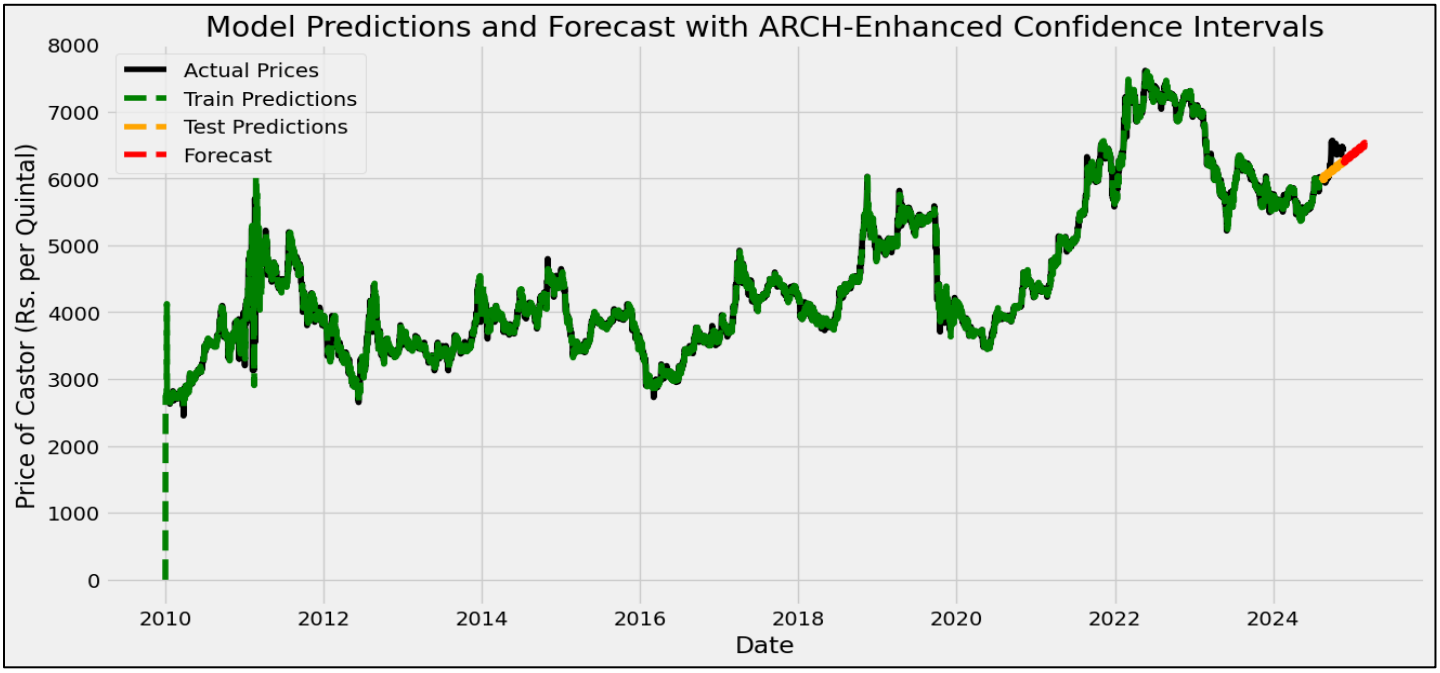


Figure 11 : ARCH Time Series Forecasting

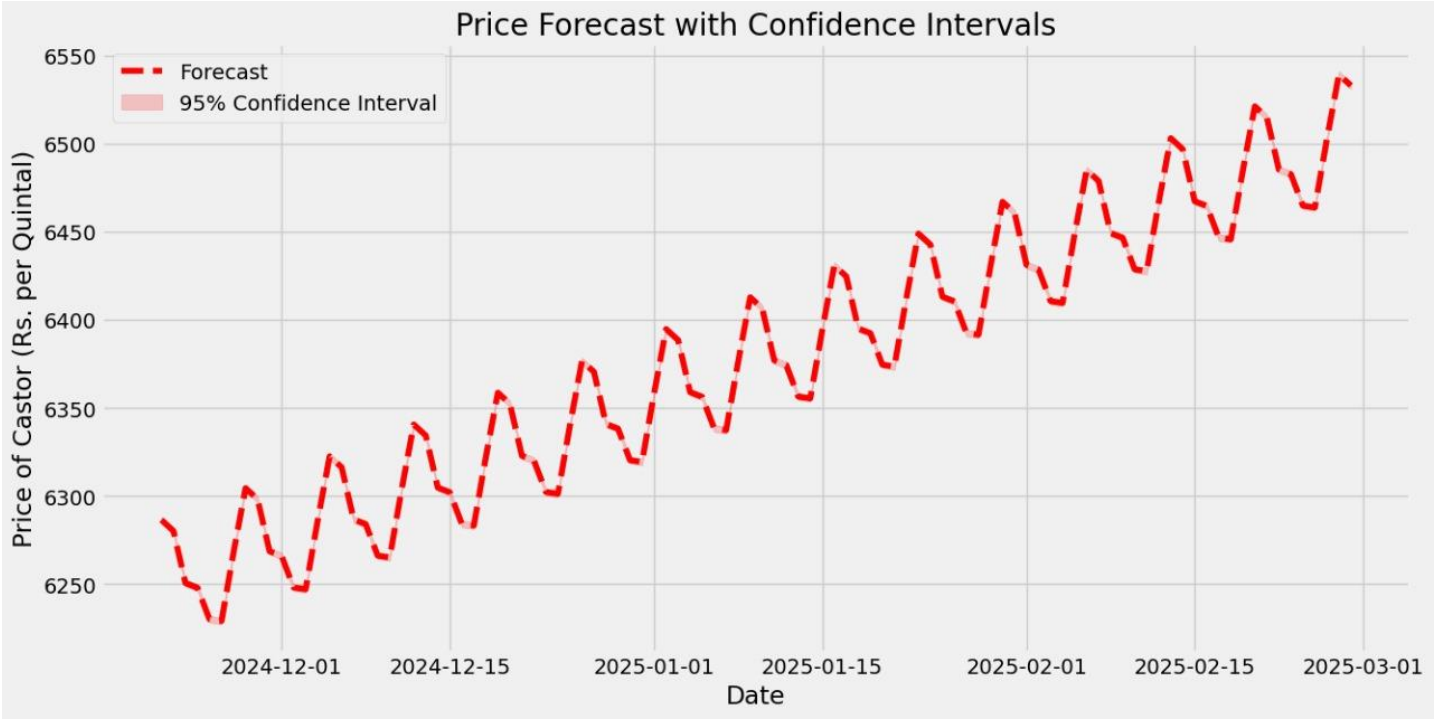


Figure 12 : ARCH Forecast with 95% confidence interval

VI) GARCH Model :

The model is best fitted at ARIMA (3,1,0)(3,1,0)[7].

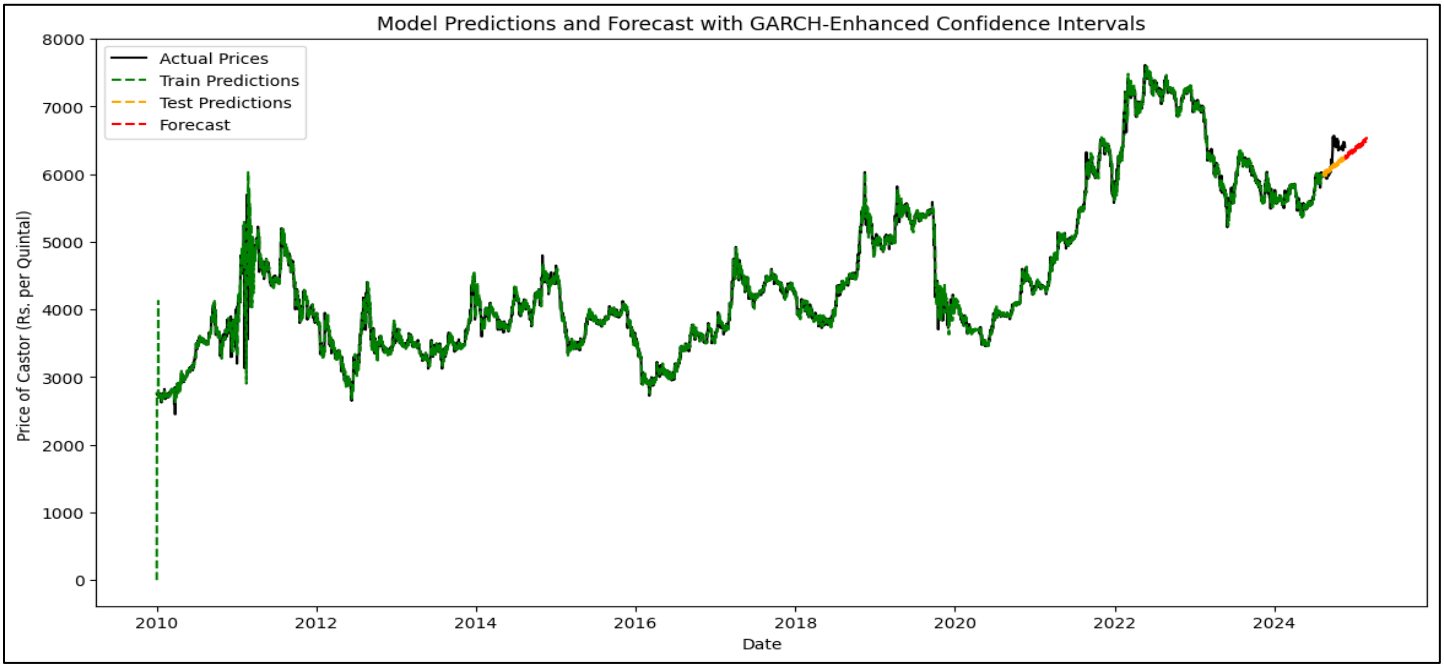


Figure 13 : GARCH Time Series Forecasting

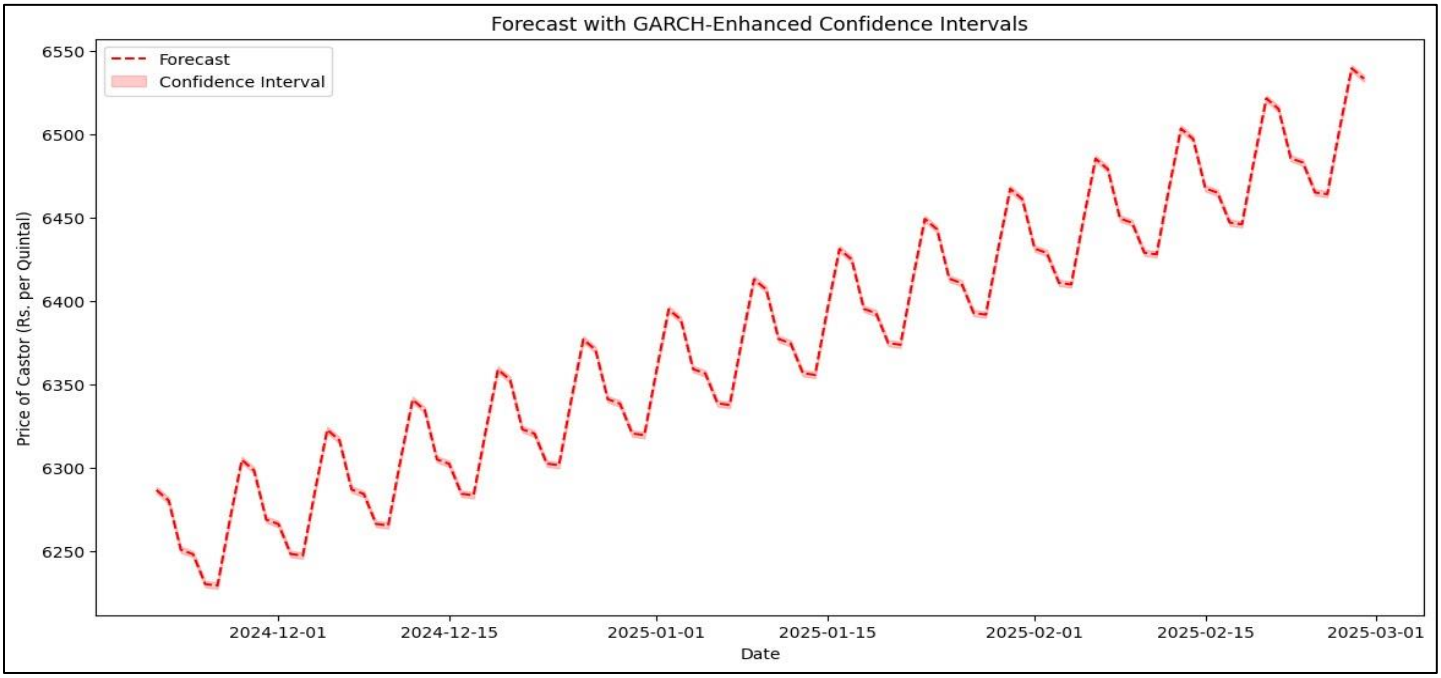


Figure 14 : GARCH Forecast with 95% confidence interval

VII) VAR Model :

Best fitted model : Order(13,20)

Figure 15 : VAR Model – Prediction of Price

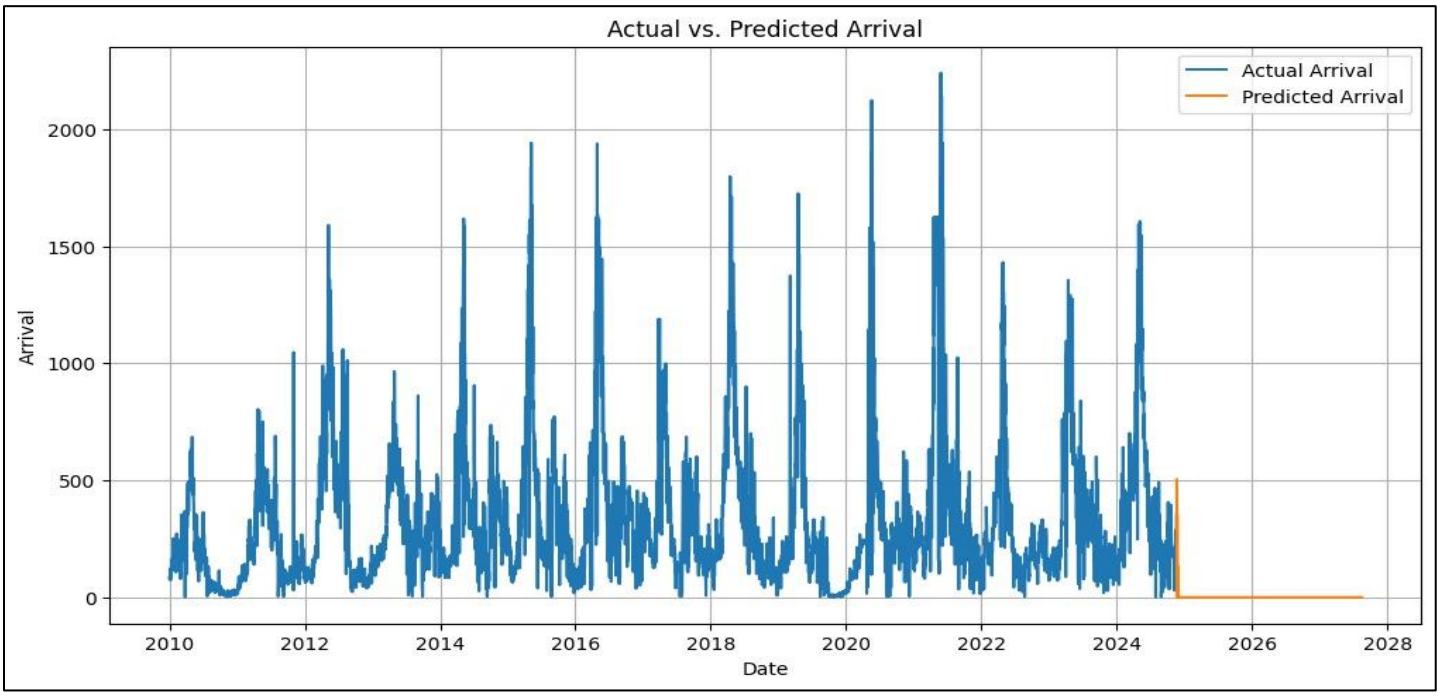


Figure 16 : VAR Model - Prediction of Arrival

VIII) VARMAX Model :
Best fitted Model = Order(13,20)

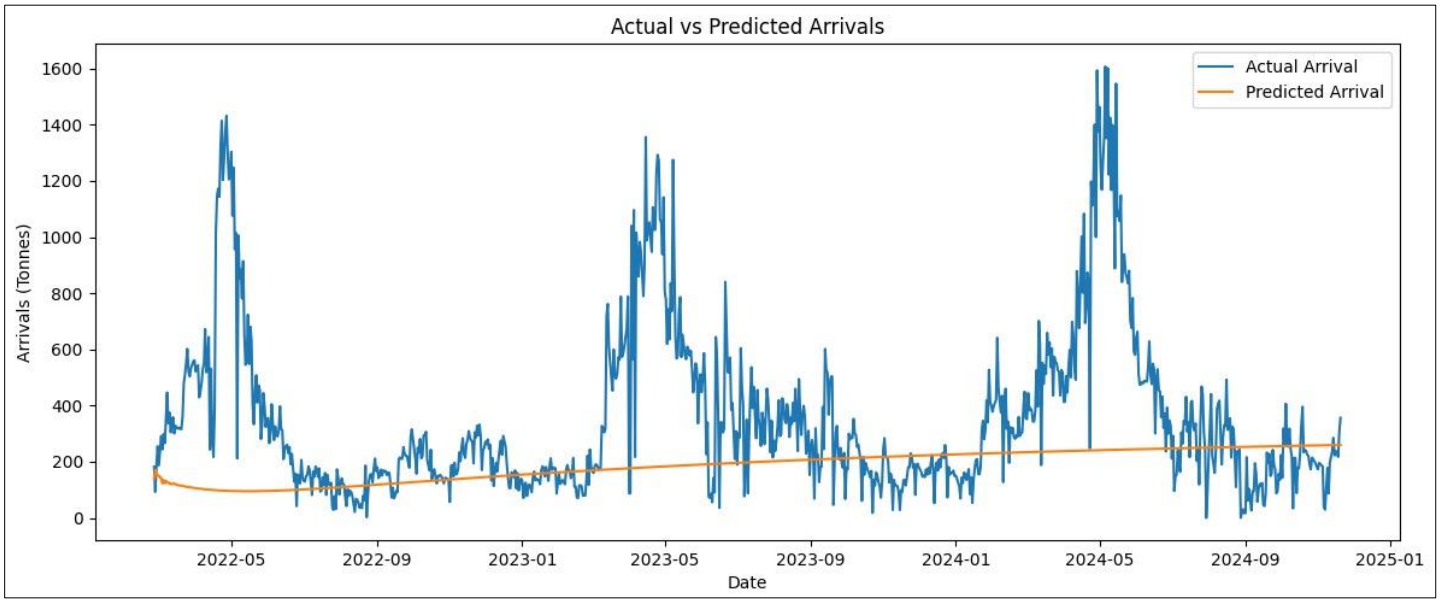


Figure 17 : VARMAX model - Prediction of Arrival

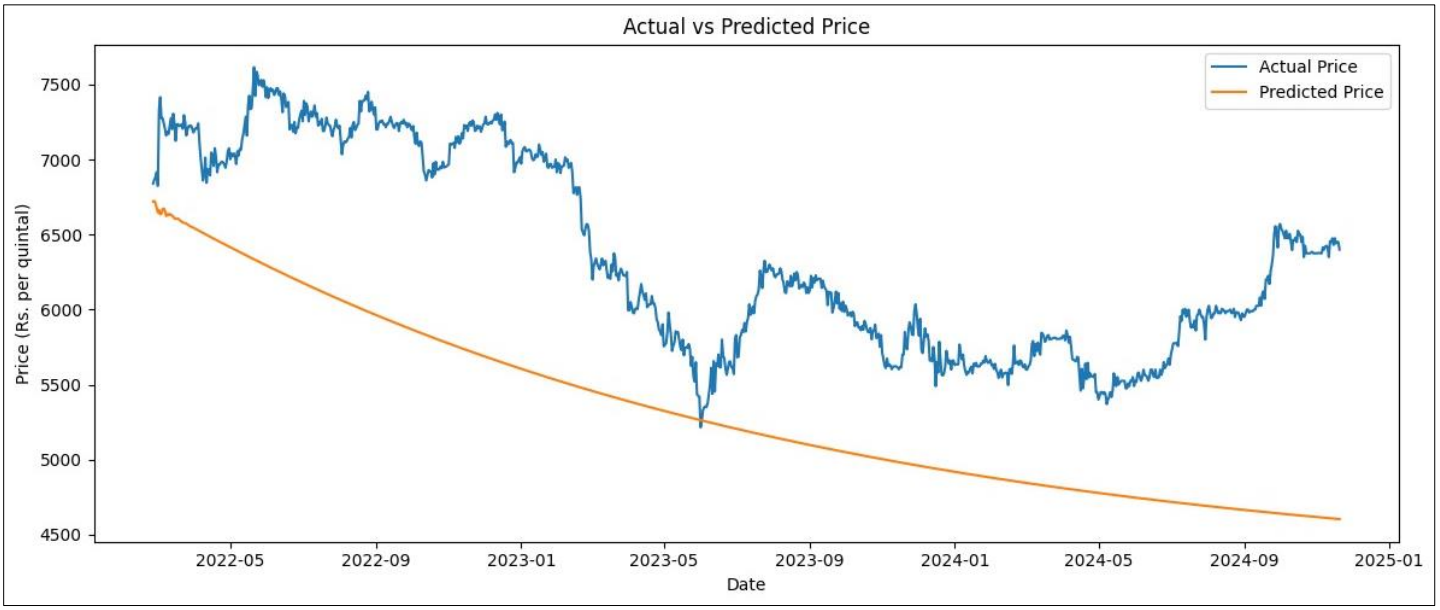


Figure 18 : VARMAX model - Prediction of Price

IX) LSTM Model :

The model is best fitted at {'optimizer'='Adam','learning_rate'='0.001','epochs'='50','batch_size'='64' 'loss'='mean_squared_error'}

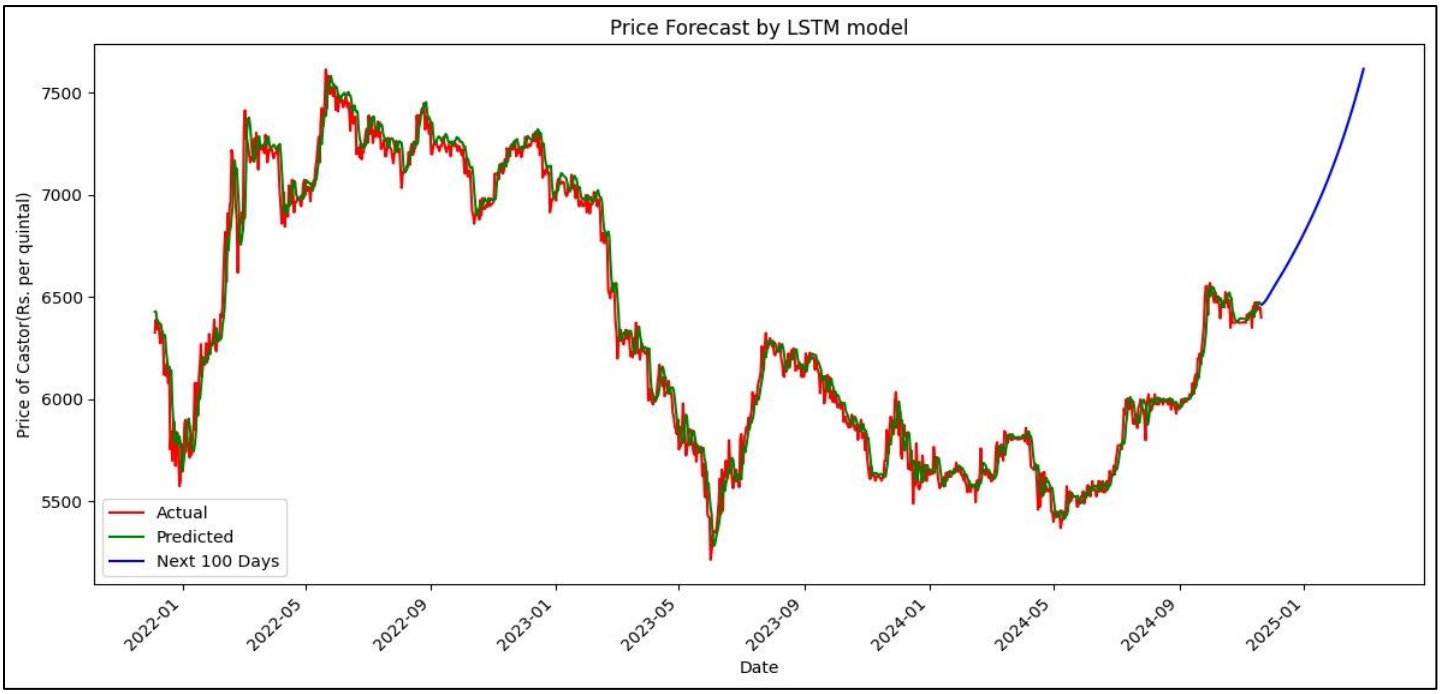


Figure 19 : Price Forecast by LSTM

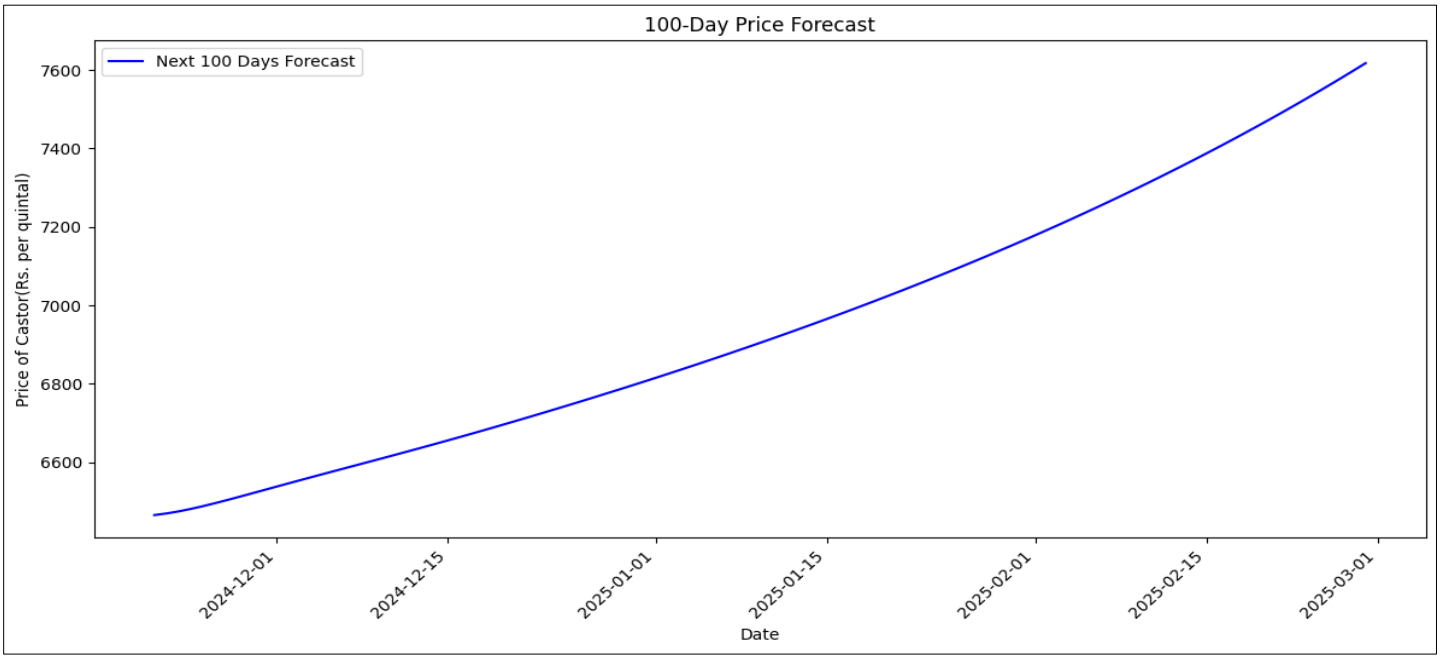


Figure 20 : LSTM forecast

X) Random Forest Model :

Best fitted model :{ 'lag':'5','bootstrap':True,'criterion':'squared_error','max_features':1.0,'min_samples_leaf':1, 'min_samples_split':2, 'n_estimators':100, 'random_state':42}

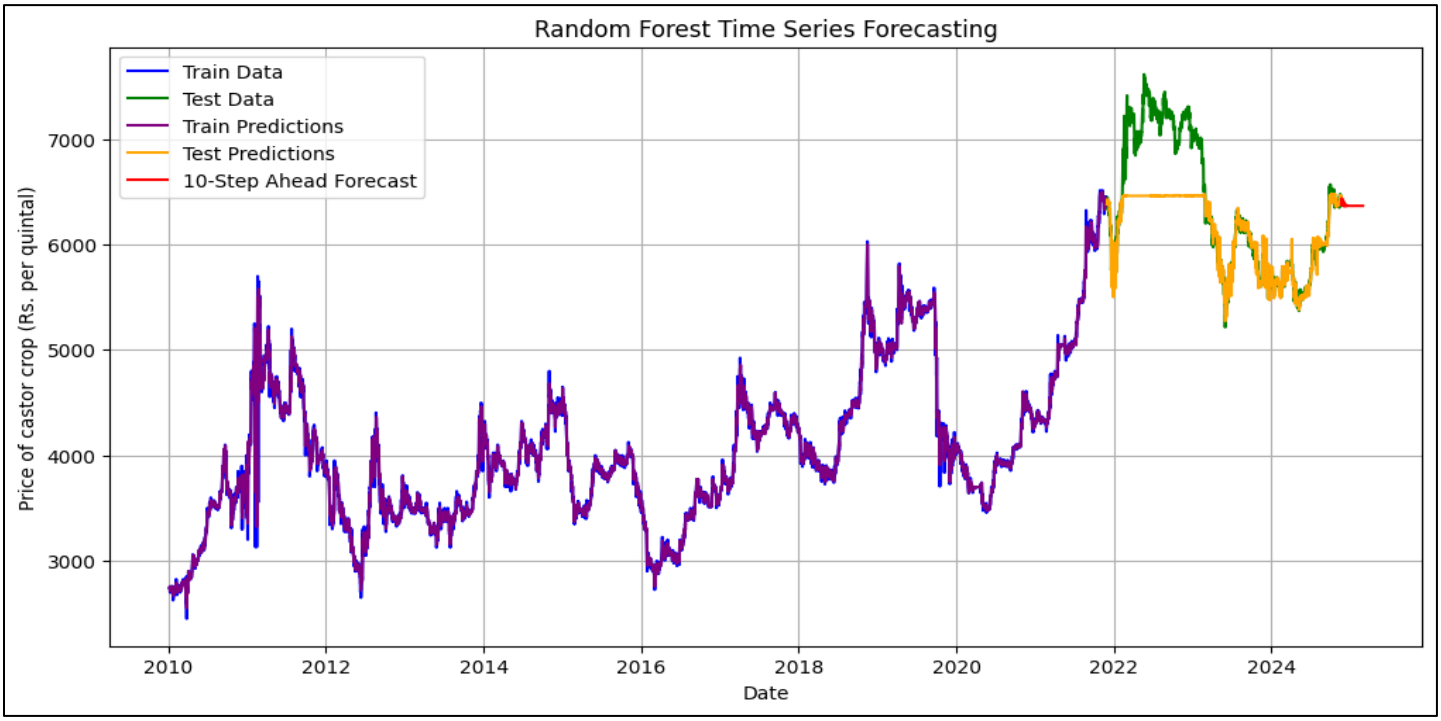


Figure 21 : Time Series Forecasting by Random Forest

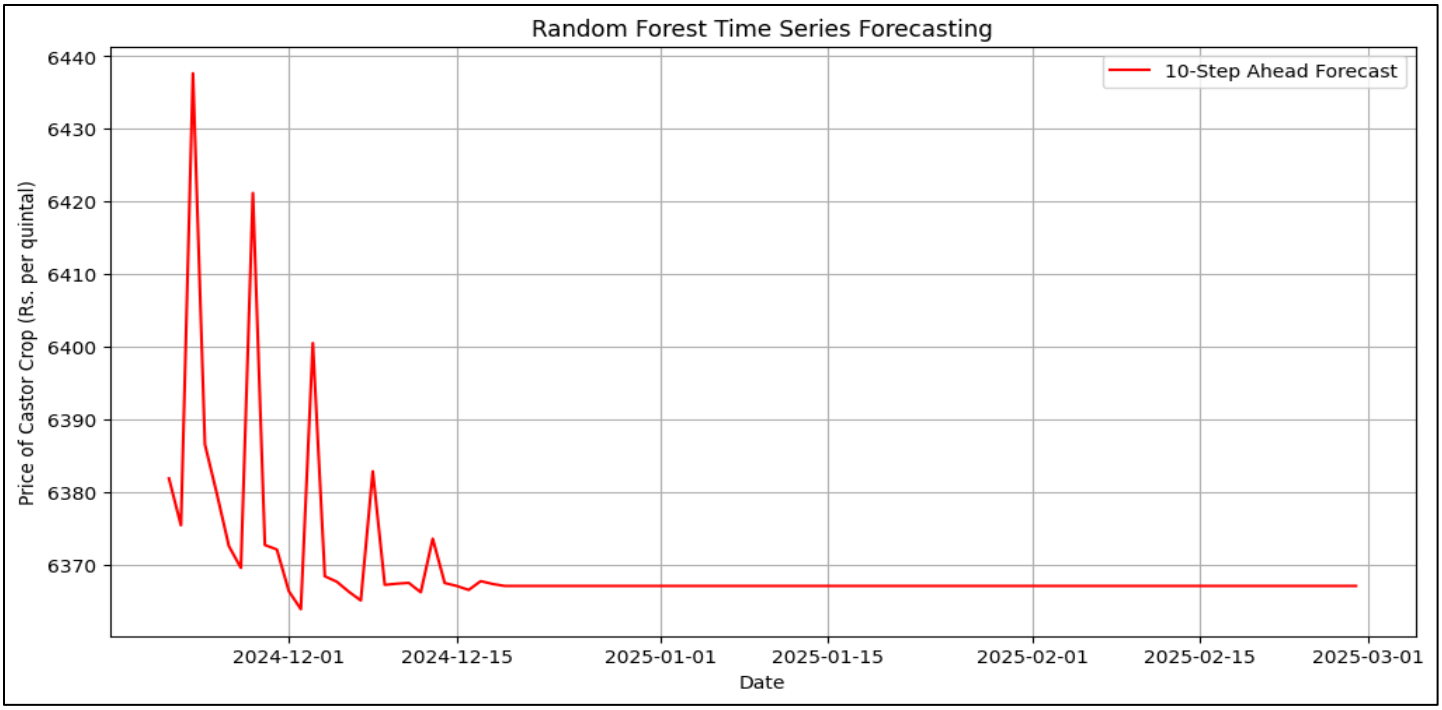


Figure 22 : Random Forest forecast

XI) Matrices :

Table 3 - Matrices

Model	AIC	BIC	MSE	RMSE	NRMSE %
ARIMA	51473.523	51499.034	420216.06	648.24	27.01
ARIMAX	51327.428	51372.071	419990.44	648.07	27.00
SARIMA	51456.233	51494.500	420826.47	648.71	27.03
SARIMAX	52162.208	52206.841	394303.05	627.94	26.16
ARCH	63599.795	63645.862	47898.47	218.86	34.20
GARCH	12352.4	12365.5	47898.47	218.86	34.20
VAR (Price)	19.0181	19.0268	40778463.08	6385.80	100.51
VAR (Arrival)			214078.969	462.686	129.98
VARMAX (Price)	108777.141	109141.824	6353.292	1080.926	17.01
VARMAX (Arrival)			355.97	342.449	96.20
LSTM	-	-	13025.904	114.13	1.80
RANDOM FOREST	-	-	178919.97	422.99	17.62

CONCLUSION

The primary objective of this project was to develop a reliable and robust price prediction system for castor crops in Patan, Gujarat, utilizing diverse statistical and machine learning models. After extensive experimentation and analysis, the following conclusions were drawn :

1) Performance of Models:

- **SARIMAX** emerged as the most accurate statistical model with the **lowest RMSE (627.94)** and **RMSE % (26.16%),** outperforming ARIMA, ARIMAX, and SARIMA. Its ability to incorporate exogenous factors made it particularly effective for forecasting castor prices.

- **LSTM** a deep learning model, demonstrated **excellent performance** with an **RMSE of 114.13** and an exceptionally low RMSE % of 1.80%. Its ability to capture nonlinear relationships and long-term dependencies makes it a promising choice for advanced forecasting tasks.

- **Random Forest** provided robust predictions, achieving an RMSE of 422.99 and RMSE % of 17.62%, indicating its effectiveness for scenarios.

2) Statistical Models :

- Traditional models like **ARIMA, ARIMAX**, and **SARIMA** provided moderate performance with RMSE values between 648 and 648.71 and RMSE % of approximately 27%. While effective for modeling short-term trends, they struggled to capture external influences and complex seasonality.

- **ARCH** and **GARCH** models excelled in volatility modeling, with both achieving low RMSE values of 218.86. However, their RMSE % (34.20%) indicates that they are more suited for risk management and market stability analysis than direct price prediction.

3) Multivariate Models :

- **VAR (Price)** and **VAR (Arrival)** demonstrated mixed performance. While VAR models captured interdependencies between variables, they had significantly higher RMSE values (6385.80 and 462.686, respectively), making them less effective for price forecasting.

- **VARMAX** models performed relatively better, especially for arrivals, achieving an RMSE of 342.449 with RMSE % of 96.20%. For prices, VARMAX produced an RMSE of 1080.926 with RMSE % of 17.01%.

4) Insights for Stakeholders :

- Farmers and traders can benefit from **LSTM** for proactive market decision-making, such as optimizing storage, timing of sales, and improving supply chain efficiency.

- Policymakers can leverage SARIMAX and Random Forest to monitor market trends, mitigate price volatility, and implement targeted interventions for price stabilization.

REFERENCES

[1] Gohil, V., Upadhyay, S.M., Patel, D.V., Delvadiya, J., & Patel, H. (2023).
Time series analysis of castor crop for price forecasting in Gujarat: A comprehensive study.
International Journal of Statistics and Applied Mathematics, 8(5), 194-203.
DOI: 10.22271/maths.2023.v8.i5c.1339

[2] Bansal, R., & Zala, Y.C. (2015).
A study on growth and volatility in cash and futures market of castor in India.
International Journal of Bio-resource and Stress Management, 6(5), 615-618.
DOI: 10.5958/0976-4038.2015.00094.9

[3] Dhandhalya, M.G., Chavda, H., Marviya, P.B., Tarpara, V.D., & Kumar, K. (2016).
Benefits of price forecast to castor growers in Gujarat.
International Journal of Agriculture Sciences, 8(52), 2414-2416.
Available at: http://www.bioinfopublication.org/jouarchive.php?opt=&jouid=BPJ0000217

[4] Box, G.E.P., & Jenkins, G.M. (1970).
Time Series Analysis: Forecasting and Control.
Holden-Day, San Francisco.

[5] Engle, R.F. (1982).
Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation.
Econometrica, 50(4), 987-1007.

[6] Bollerslev, T. (1986).
Generalized autoregressive conditional heteroscedasticity.
Journal of Econometrics, 31(3), 307-327.

[7] Kataria, P., & Chahal, S. (2007).
Commodity futures trading and price risk management: A study of Indian agricultural markets.
Agricultural Economics Research Review, 20, 53-71.

[8] Montgomery, D.C., Jennings, C.L., & Kulahci, M. (2003).
Introduction to Time Series Analysis and Forecasting.
John Wiley & Sons.

[9] Lutkepohl, H. (2005). New Introduction to Multiple Time Series Analysis. Springer.
This book provides an in-depth explanation of VAR models and their application.

[10] Shumway, R. H., & Stoffer, D. S. (2017). *Time Series Analysis and Its Applications. Springer*.
The book details multivariate time series analysis, including VAR and VARMAX frameworks.

[11] Geeks For Geeks - www.geeksforgeeks.org
[12] Investopedia - www.investopedia.com
[13] Simplilearn - www.simplilearn.com

[14] Research Gate - www.researchgate.net
[15] Mathworks - www.mathworks.com
[16] Medium = medium.com
[17] Academicoup = academic.oup.com
[18] Careerfoundry = careerfoundry.com