Price Forecasting for Tomato: A Statistical and Machine Learning Approach

M.Sc. Agriculture Analytics

MODULE-4

Agricultural Market Analytics

Submitted By:
GROUP-1
Harshil Ponkiya (202319002)
Jatin Satani (202319021)
Deep Rabadiya (202319029)

Submitted to:
DR. PRITY KUMARI
Assistant Professor & Head,
Anand Agricultural University

CONTENT

ABSTRACT

Price forecasting plays a critical role in commodity trading and price analysis, particularly within the agricultural sector. In this context, India stands as the second-largest producer of tomatoes globally, following China, contributing approximately 21 million tonnes to the market—accounting for 11.02% of the worldwide tomato supply. As the second most significant vegetable crop after potatoes, tomatoes highlight the importance of reliable pricing mechanisms due to their perishable nature and seasonal availability.

Recent trends indicate that tomato prices have undergone considerable fluctuations, which have escalated the financial risks encountered by growers. To address these challenges, the present study employs a comprehensive range of tomato price forecasting models. These include statistical methods such as ARIMA, ARIMAX, SARIMA, and SARIMAX, as well as volatility models including ARCH, and GARCH. Additionally, the study incorporates Vector Autoregression (VAR) and its extension (VARMAX), along with machine learning techniques like Long Short-Term Memory (LSTM) networks and Random Forest algorithms. This multifaceted approach aims to enhance the accuracy of price forecasts and provide valuable insights for stakeholders in the tomato supply chain.

INTRODUCTION

Tomato (Solanum lycopersicum L.) is a vital vegetable crop in India, ranking just behind potato, and it plays an important role in the Indian diet due to its versatility and nutritional benefits. Rich in antioxidants such as lycopene and carotene, tomatoes are a key ingredient in many dishes, from salads and sauces to processed products like ketchup and purees. India stands as the second-largest producer of tomatoes globally, after China, contributing over 11% to global tomato production with an annual yield of approximately 21 million tonnes (*Singh et al*., 2023).

Despite the year-round demand for tomatoes, their cultivation and pricing encounter significant challenges, primarily due to their perishable nature and seasonal production cycles. Tomatoes are predominantly grown during the rabi season in many regions, including our state, and they are also cultivated during the kharif season. Limited off-season production coupled with fluctuating supply often leads to significant price variations, with costs surging during periods of low yield.

This price volatility poses challenges for household budgets and affects the profitability of tomato growers, who often struggle to capitalize on advantageous market conditions . Accurate price forecasting is crucial in this context, enabling farmers to make informed decisions about their production and marketing strategies. By gaining insights into the factors driving price fluctuations, stakeholders can more effectively navigate market risks and enhance the economic stability of the tomato supply chain. This report highlights the necessity for precise tomato price predictions and provides insights into risk mitigation to facilitate better planning for both growers and consumers.

The present investigation has been undertaken to study the relationship between arrivals and prices of tomato, to assess the price volatility and co-integration of tomato among the selected markets in Maharashtra. It revealed that the maximum variability of tomato in arrivals was noticed in Pimpalgaon market among the months as compared to Mumbai, Nagpur, Nashik and Pune. The maximum variability of tomato in prices was noticed in Pimpalgaon market among the months as compared to Mumbai, Nagpur, Nashik and Pune. The overall significant negative correlation between arrivals and prices of tomato in Nagpur market was noticed. The market pair Nagpur-Pimpalgaon have bidirectional causality and Mumbai-Nagpur, Mumbai-Nashik, Mumbai-Pimpalgaon, Mumbai-Pune, Nagpur-Nashik, Pune-Nagpur, Pimpalgaon- Nashik and Pune-Pimpalgaon have a unidirectional causality. It is concluded that very high price volatility of tomato present in Mumbai, Nagpur, and Nashik and Pimpalgaon markets. It should be minimized and needs to protect price security for farming community (*Bhagat et al.,* 2023).

REVIEW OF LITERATURE

Tomato price forecasting has been the focus of numerous studies employing diverse methodologies, each with unique strengths and limitations.

ARIMA models are popular for time series forecasting due to their ability to capture trends and seasonality. *Sonvanee et al.* (2024) applied ARIMA to tomato price data and found it effective for short to medium-term forecasts. These models are particularly advantageous when market data exhibit linear trends and periodicity, although their predictive accuracy diminishes for highly non-linear or irregular data.

SARIMA extends ARIMA by incorporating seasonal components, making it well-suited for data with pronounced seasonal patterns. Studies like *Arunraj et al.* (2013) have shown SARIMA's effectiveness in capturing both seasonal fluctuations and underlying trends in tomato prices.

ARCH models are commonly used to model volatility in financial time series and have been applied to agricultural pricing to account for price fluctuations. For instance, *Bhagat et al.* (2023) demonstrated the utility of ARCH models in capturing tomato price volatility. However, these models are less suitable for long-term predictions as they primarily focus on short-term volatility clustering.

VAR models analyze interdependencies among multiple variables, making them valuable for understanding the dynamics between tomato prices and related factors like weather, transportation costs, and market demand. *Kumar et al.* (2024) highlighted VAR's capability to improve multi-variable forecasting accuracy, though its performance depends on the availability and quality of auxiliary data.

LSTM, a type of recurrent neural network, excels at capturing long-term dependencies and non-linear patterns in time series data. Studies such as *G. Avinash et al.* (2024) demonstrated LSTM's superior performance over traditional models when forecasting complex agricultural datasets. Its adaptability to include external variables like weather, crop yields, and market dynamics enhances its predictive power.

Random forest has also been applied to tomato price forecasting. These models are adept at handling non-linear relationships and high-dimensional data. *Tyralis et al.* (2017) observed that random forest models performed exceptionally well when enriched with diverse features such as climatic conditions and historical price trends. However, machine learning models require careful feature engineering and extensive datasets to achieve optimal performance.

Forecasting model effectiveness is often evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Despite advancements, challenges remain in capturing the full complexity of agricultural pricing, which is influenced by external factors like weather variability, policy changes, and global market trends. Cross-validation and out-of-sample testing are critical to ensure model reliability.

OBJECTIVES

The specific objectives of the present studies are,

1. To review the forecasting techniques used in time series analysis.
2. To forecast the future prices of TOMATO crop for the area under study.

DATASET

India's leading tomato-producing states include Andhra Pradesh, Madhya Pradesh, Karnataka, Gujarat, Bihar, Odisha, West Bengal, Telangana, Chhattisgarh, Haryana, Tamil Nadu, Uttar Pradesh, and Maharashtra. A dataset covering daily tomato prices has been compiled from the AGMARKNET website for the Pune market in Maharashtra.

The Pune tomato market was selected for its consistent supply, robust production [3], and strategic role in catering to central India's tomato demand. Additionally, its reliable data availability makes it ideal for analysis.

Reference Article: https://timesofindia.indiatimes.com/city/pune/tomato-prices-surge-due-to-supply-crunch/articleshow/111010046.cms

Data Source: https://agmarknet.gov.in/

Study Area:

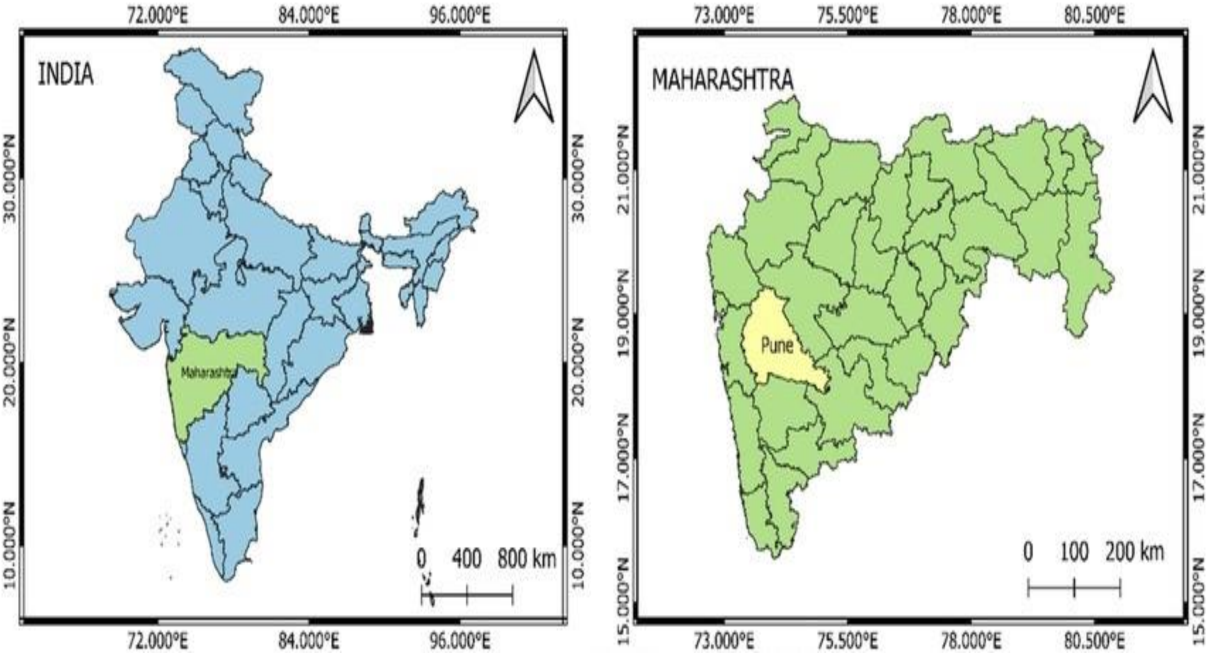


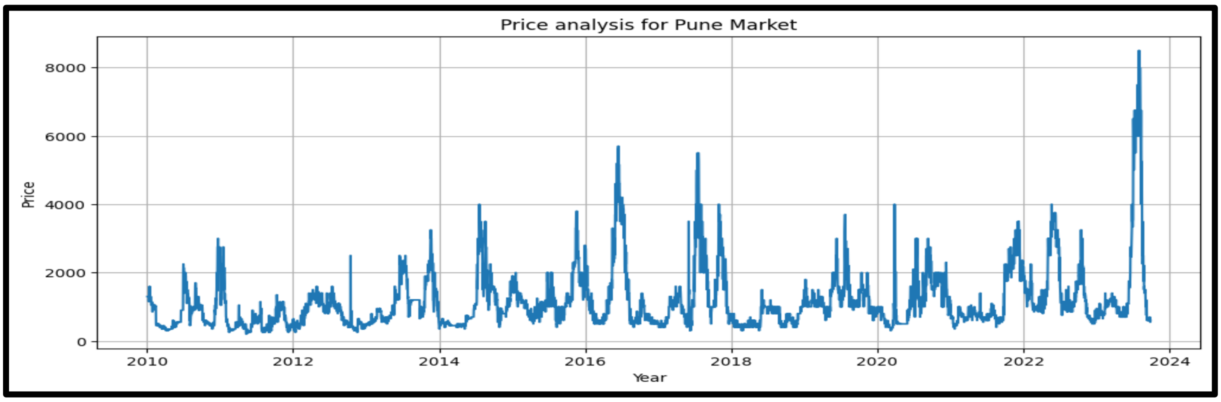
FIG: Pune District Highlighted within the State of Maharashtra, India

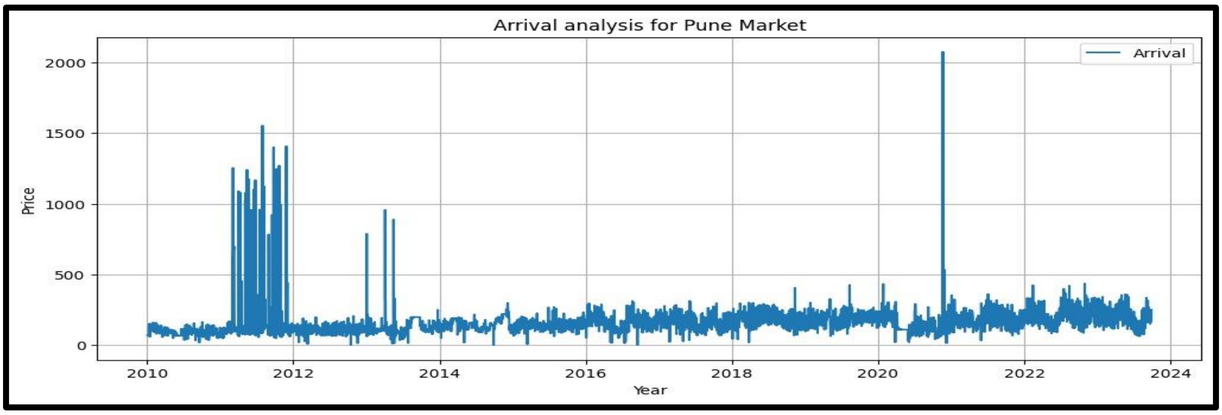
Table 1: About Data

Commodity	Tomato
State	Maharashtra
District	Pune
Market	Pune Market
Time Period	1st Jan 2010 to 12th Oct 2023
Price/Arrival	Both

Table 2: Descriptive Statistics of Arrival and Price Data

	Arrival	**Price**
count	5017	5017
mean	161.12	1161.18
std	97.56	896.34
min	0.1	200
25%	111	600
50%	147.7	950
75%	194	1350
max	2075	8500





METHODOLOGY

Different models like ARIMA, ARIMAX, SARIMA, SARIMAX, ARCH, GARCH, VAR, VARMAX, LSTM and Random Forest are used to analyze and predict the tomato price data.

1. ARIMA MODEL:

The ARIMA (Autoregressive Integrated Moving Average) model is a powerful statistical tool for time series forecasting that combines three key components: autoregression (AR), integration (I), and moving average (MA).

The ARIMA model, represented as ARIMA (p, d, q), uses three parameters: p, which indicates the number of lagged observations in the autoregressive (AR) component; d, which signifies the number of differences needed to achieve stationarity in the data; and q, which represents the number of lagged forecast errors in the moving average (MA) component. ARIMA model is also known as Box Jenkins models and it is expressed as follow [1]:

$$\mathbf{Y_t = \mu + \sum_{i=1}^{p} \varphi_i Y_{i-1} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_i}$$

Where:
$\mathbf{Y_t}$ **=** price
μ = mean of series
$\boldsymbol{\varphi_i}$ **=** parameters of AR model
$\boldsymbol{\theta_j}$ **=** parameters of MA model
$\boldsymbol{\varepsilon_t}$ **=** noise error term

Developing an ARIMA model starts with model identification, where data is differenced to achieve stationarity, and ACF and PACF plots are used to select p, d, and q values. Parameters are then estimated using methods like Maximum Likelihood Estimation, followed by validation through diagnostic tests, such as the Ljung-Box test, to ensure residuals are uncorrelated and normally distributed. If the model fails, parameters are adjusted and re-evaluated. Once validated, the model is used for forecasting, capturing patterns in the time series for accurate predictions, such as forecasting tomato prices.

2. ARIMAX MODEL:

The ARIMAX (Autoregressive Integrated Moving Average with eXogenous variables) model extends ARIMA by incorporating external variables, allowing for improved forecasting by considering additional influencing factors. ARIMAX model is expressed as follow [1]:

$$\mathbf{Y_t = \mu + \sum_{i=1}^{p} \varphi_i Y_{i-1} x_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_i}$$

Where:
$\mathbf{Y_t}$ = price
μ = mean of series
$\boldsymbol{\varphi_i}$ = parameters of AR model
$\mathbf{x_t}$ = exogenous variables
$\boldsymbol{\theta_j}$ = parameters of MA model
$\boldsymbol{\varepsilon_t}$ = noise error term

Model building starts with identifying a suitable ARIMA structure for the main variable, testing for stationarity, and applying differencing if necessary. Residual diagnostics, such as ACF and PACF plots, ensure the residuals resemble white noise, indicating a good fit. Parameters are estimated using nonlinear least squares methods.

The best model is selected based on performance metrics like maximum R², minimum RMSE, and minimum MSE. Once finalized, the ARIMAX model generates forecasts while accounting for external variables effectively, making it highly suitable for stationary time series data with external influences.

3. SARIMA MODEL:

Deseasonalization is performed to analyse and forecast time series data with seasonal patterns by dividing each observation by its seasonal index, removing fluctuations for clearer trend and cycle analysis.

The SARIMA (Seasonal ARIMA) model, following the Box-Jenkins approach, is ideal for time series with periodic seasonal patterns. Represented as ARIMA(p, d, q)×(P, D, Q)S, it integrates non-seasonal (p, d, q) and seasonal (P, D, Q, S) components, where S denotes the seasonal cycle period. The SARIMA model can be represented as [2]:

$$\mathbf{\Phi_p(B)\Phi_p(B^S)(1-B)^d(1-B^S)^D Z_t = \theta_q(B)}\Theta_Q\mathbf{(B^S)\varepsilon_t}$$

Where:
$\mathbf{\Phi_p(B)}$ = Seasonal autoregressive operator with p-order
$\boldsymbol{\theta_q}\mathbf{(B)}$ = Seasonal moving average operator with q-order
$\mathbf{(1-B)^d}$ = Seasonal autoregressive operator with p-order
$\mathbf{(1-B)^D}$ = Seasonal moving average operator with q-order
S = Seasonal length (e.g. in quarterly data s=4 and in monthly data s=12)
$\boldsymbol{\varepsilon_t}$ = Residual error in SARIMA model

SARIMA handles stationary and non-stationary data with seasonal elements, making it versatile for various forecasting scenarios. Addressing outliers, which represent anomalies or significant events, improves model accuracy and provides actionable insights. Incorporating external variables further enhances precision, allowing SARIMA to deliver robust and reliable forecasts for time series with seasonal characteristics, such as weekly tomato prices.

4. SARIMAX MODEL:

The SARIMAX (Seasonal ARIMA with eXogenous variables) model extends SARIMA by incorporating external variables for enhanced forecasting. The general SARIMAX model equation can be obtained by [2]:

$$\mathbf{Y_t} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\mathbf{X_{1,t}} + \cdots + \boldsymbol{\beta_k}\mathbf{X_{k,t}} + \left(\frac{\boldsymbol{\theta_q}(\mathbf{B})\Theta_{\mathbf{Q}}(\mathbf{B^S})\boldsymbol{\varepsilon_t}}{\boldsymbol{\Phi_p}(\mathbf{B})\boldsymbol{\Phi_p}(\mathbf{B^S})(\mathbf{1}-\mathbf{B})^{\mathbf{d}}(\mathbf{1}-\mathbf{B^S})^{\mathbf{D}}}\right)$$

Where:

$\mathbf{X_{k,t}}$ = observations of k number of external variables

$\boldsymbol{\beta_k}$ = regression coefficients of external variables

$\boldsymbol{\Phi_p}(\mathbf{B})$ = Seasonal autoregressive operator with p-order

$\boldsymbol{\theta_q}(\mathbf{B})$ = Seasonal moving average operator with q-order

$(\mathbf{1}-\mathbf{B})^{\mathbf{d}}$ = Seasonal autoregressive operator with p-order

$(\mathbf{1}-\mathbf{B})^{\mathbf{D}}$ = Seasonal moving average operator with q-order

S = Seasonal length (e.g. in quarterly data s=4 and in monthly data s=12)

$\boldsymbol{\varepsilon_t}$ = Residual error in SARIMA model

Represented as SARIMAX(p, d, q) × (P, D, Q)S, it unites seasonal and non-seasonal components along with external factors. Model development starts with achieving stationarity through seasonal and non-seasonal differencing, followed by analysing ACF and PACF plots to determine optimal lag orders.

Parameters are estimated using Maximum Likelihood Estimation, and residual diagnostics, such as the Ljung-Box Q test, ensure residuals are uncorrelated and normally distributed. External variables are integrated via linear regression, quantifying their impact through regression coefficients. The model is validated using out-of-sample data to confirm accuracy before being applied for forecasting.

5. ARCH MODEL:

The Autoregressive Conditional Heteroskedasticity (ARCH) model is widely used in financial econometrics to analyse and model time-varying volatility in time series data. It captures heteroskedasticity, where error variance changes over time, a common feature in financial markets characterized by volatility clustering. The ARCH model assumes that the variance of the error term depends on past squared residuals, modeled as an autoregressive process. Under conditions of w > 0; $\boldsymbol{\alpha_i} \geq \mathbf{0}$; $\sum_{\mathbf{i=1}}^{\mathbf{q}} \boldsymbol{\alpha_i} < \mathbf{1}$, the general ARCH (q) process is shown in equation format [3]:

$$\boldsymbol{\sigma}_{\mathbf{t}}^{\mathbf{2}} = \mathbf{w} + \sum_{\mathbf{i=1}}^{\mathbf{q}} \boldsymbol{\alpha}_{\mathbf{i}} \mathbf{u}_{\mathbf{t-i}}^{\mathbf{2}}$$

which means that the variance ($\boldsymbol{\sigma}_{\mathbf{t}}^{\mathbf{2}}$) of the time series today is equal to a constant ($\mathbf{w}$), plus some amount ($\boldsymbol{\alpha}$) of the previous variance ($\mathbf{u}_{\mathbf{t-i}}^{\mathbf{2}}$).

The model order, denoted as q, indicates the number of lagged squared residuals included. Parameters are estimated using Maximum Likelihood Estimation (MLE), optimizing the fit of the model to the data. By effectively capturing volatility dynamics, the ARCH model serves as a powerful tool for forecasting volatility and supporting financial decision-making and risk management.

6. GARCH MODEL:

GARCH (Generalized Autoregressive Conditional Heteroskedasticity) models are advanced tools for modeling conditional volatility in time series, especially in cases where volatility exhibits clustering, meaning periods of high or low volatility tend to follow one another. A GARCH model is typically of the following form [1]:

$$\boldsymbol{\sigma}_{\mathbf{t}}^{\mathbf{2}} = \mathbf{w} + \sum_{\mathbf{i=1}}^{\mathbf{p}} \boldsymbol{\alpha}_{\mathbf{i}} \in_{\mathbf{t-i}}^{\mathbf{2}} + \sum_{\mathbf{i=1}}^{\mathbf{q}} \boldsymbol{\beta}_{\mathbf{i}} \boldsymbol{\sigma}_{\mathbf{t-i}}^{\mathbf{2}}$$

which means that the variance ($\boldsymbol{\sigma}_{\mathbf{t}}^{\mathbf{2}}$) of the time series today is equal to a constant ($\mathbf{w}$), plus some amount ($\boldsymbol{\alpha}$) of the previous residual ($\in_{\mathbf{t-1}}$), plus some amount ($\boldsymbol{\beta}$) of the previous variance ($\boldsymbol{\sigma}_{\mathbf{t-i}}^{\mathbf{2}}$). Both ARCH (Autoregressive Conditional Heteroskedasticity) and GARCH models are widely used in analysing time-varying, non-linear volatility patterns in financial and economic data.

GARCH models extend ARCH by incorporating lagged conditional variances alongside lagged squared residuals, significantly reducing the number of parameters needed while maintaining the ability to effectively model complex volatility structures. These models are particularly valuable when the variance of a time series is not constant, as they explicitly address changes in variance over time, enabling accurate forecasting and analysis.

7. VAR MODEL:

Vector Autoregression (VAR) is a statistical tool for modeling and analysing the interdependence between multiple time series variables. Unlike univariate models, VAR examines variables collectively, capturing their dynamic interactions. The general form of the VAR model with the order p (VAR (p)) is as follows [4]:

$$\boldsymbol{Y_t} = \boldsymbol{\alpha}_{\mathbf{0}} + \boldsymbol{\beta_1 Y_{t-1}} + \boldsymbol{\beta}_{\mathbf{2}} \mathbf{Y}_{\mathbf{t-2}} + \cdots + \boldsymbol{\beta}_{\mathbf{m}} \mathbf{Y}_{\mathbf{t-p}} + \boldsymbol{\varepsilon}_{\mathbf{t}}$$

Where:

$\mathbf{Y_{t-i}}$ = a vector of size n × n containing n variables that are included in the VAR model at times t and t-i, i = 1,2, ..., p.

$\boldsymbol{\alpha}$ = intercept vector of size n × 1

$\boldsymbol{\beta_i}$ = a coefficient matrix of size n × n for every i = 1,2, ..., p

$\boldsymbol{\varepsilon_t}$ = error value vector of size n × 1

$\boldsymbol{p}$ = lag VAR

Each variable is modeled as a linear function of its past values and those of other variables, making it effective for understanding interconnected relationships. Represented in vector form, VAR describes the joint behaviour of variables through equations and includes error terms to account for unobservable influences.

8. VARMAX MODEL:

A VAR process can be influenced by observable variables determined outside the system, known as exogenous variables, which can be stochastic or non-stochastic. These variables, along with their lags, can impact the process. The general form of the VARX model (p) where p is the order (lag) of the endogenous variable and q is the lag of the exogenous variable can be written as follows [5]:

$$\mathbf{Y_t} = \boldsymbol{\alpha_0} + \boldsymbol{\beta_1}\mathbf{Y_{t-1}} + \boldsymbol{\beta_2}\mathbf{Y_{t-2}} + \cdots + \boldsymbol{\beta_m}\mathbf{Y_{t-p}} + \Theta\mathbf{X_t} + \boldsymbol{\varepsilon_t}$$

Where:

$\mathbf{Y_{t-i}}$ = a vector of size n × n containing n variables that are included in the VAR model at times t and t-i, i = 1,2, ..., p.

$\boldsymbol{\alpha}$ = intercept vector of size n × 1

$\boldsymbol{\beta_i}$ = a coefficient matrix of size n × n for every i = 1,2, ..., p

$\mathbf{X_t}$ = vector of the exogenous variable at time t-i, i = 1,2, ..., p

$\Theta_{\mathbf{j}}$ = matrix of exogenous variable sized parameters n × q for every i = 1,2, ..., q.

$\boldsymbol{\varepsilon_t}$ = error value vector of size n × 1

$\boldsymbol{p}$ = lag VAR

A VARMAX(p) model is designed to capture these dynamics, incorporating both endogenous variables and the effects of exogenous variables and their lags. Such models are valuable for explaining the dynamic relationships between endogenous and exogenous variables. Additionally, VARMAX models are widely used for predicting and forecasting time series data, providing insights into complex systems influenced by external factors.

9. LONG SHORT-TERM MEMORY (LSTM) MODEL:

Long Short-Term Memory (LSTM) is a specialized type of RNN designed to capture long-term dependencies in sequential data, overcoming issues like the vanishing gradient problem faced by traditional RNNs. Widely used in time series analysis, LSTMs maintain a memory unit called the cell state to store relevant information over extended sequences. The mathematical formulation of LSTM is represented by equations 1-6 [6]:

$$\mathbf{f_t} = \boldsymbol{\sigma}(\mathbf{W_f x_t} + \mathbf{U_f h_{t-1}} + \mathbf{b_f}) \qquad \textbf{(1)}$$

$$\mathbf{i_t} = \boldsymbol{\sigma}(\mathbf{W_i x_t} + \mathbf{U_i h_{t-1}} + \mathbf{b_i}) \qquad \textbf{(2)}$$

$$\acute{c}_\mathbf{t} = \Upsilon(\mathbf{W_c x_t} + \mathbf{U_c h_{t-1}} + \mathbf{b_c}) \qquad \textbf{(3)}$$

$$\mathbf{c_t} = \mathbf{f_t} * \mathbf{c_{t-1}} + \mathbf{i_t} * \acute{c}_\mathbf{t} \qquad \textbf{(4)}$$

$$\mathbf{o_t} = \boldsymbol{\sigma}(\mathbf{W_0 x_t} + \mathbf{U_0 h_{t-1}} + \mathbf{b_0}) \qquad \textbf{(5)}$$

$$\mathbf{h_t} = \mathbf{o_t} * \boldsymbol{\gamma}(\mathbf{c_t}) \qquad \textbf{(6)}$$

Where:
W = weight matrices for the corresponding connected input vectors
U = weight matrices of the previous time step
b = biases

The LSTM mechanism is centred around a cell state, denoted as $\boldsymbol{c_t}$, which serves as a storage unit for information. The LSTM architecture uses three gates to manage information flow: the forget gate ($\boldsymbol{f_t}$) decides which information to retain or discard from the cell state, the input gate ($\boldsymbol{i_t}$) determines what new information to add, and the output gate ($\boldsymbol{o_t}$) controls what information is passed to the next hidden state.

These gates leverage activation functions like sigmoid (σ) for decision-making and tanh for updating the cell state. By managing the flow of information, LSTMs effectively preserve context, enabling superior performance on tasks requiring long-term dependency tracking.

10. RANDOM FOREST REGRESSION MODEL:

Random Forest Regression is a machine learning approach that integrates multiple decision trees to predict continuous numerical values. Unlike classification tasks, where class labels are predicted, this method averages the predictions from multiple trees to generate the final output.

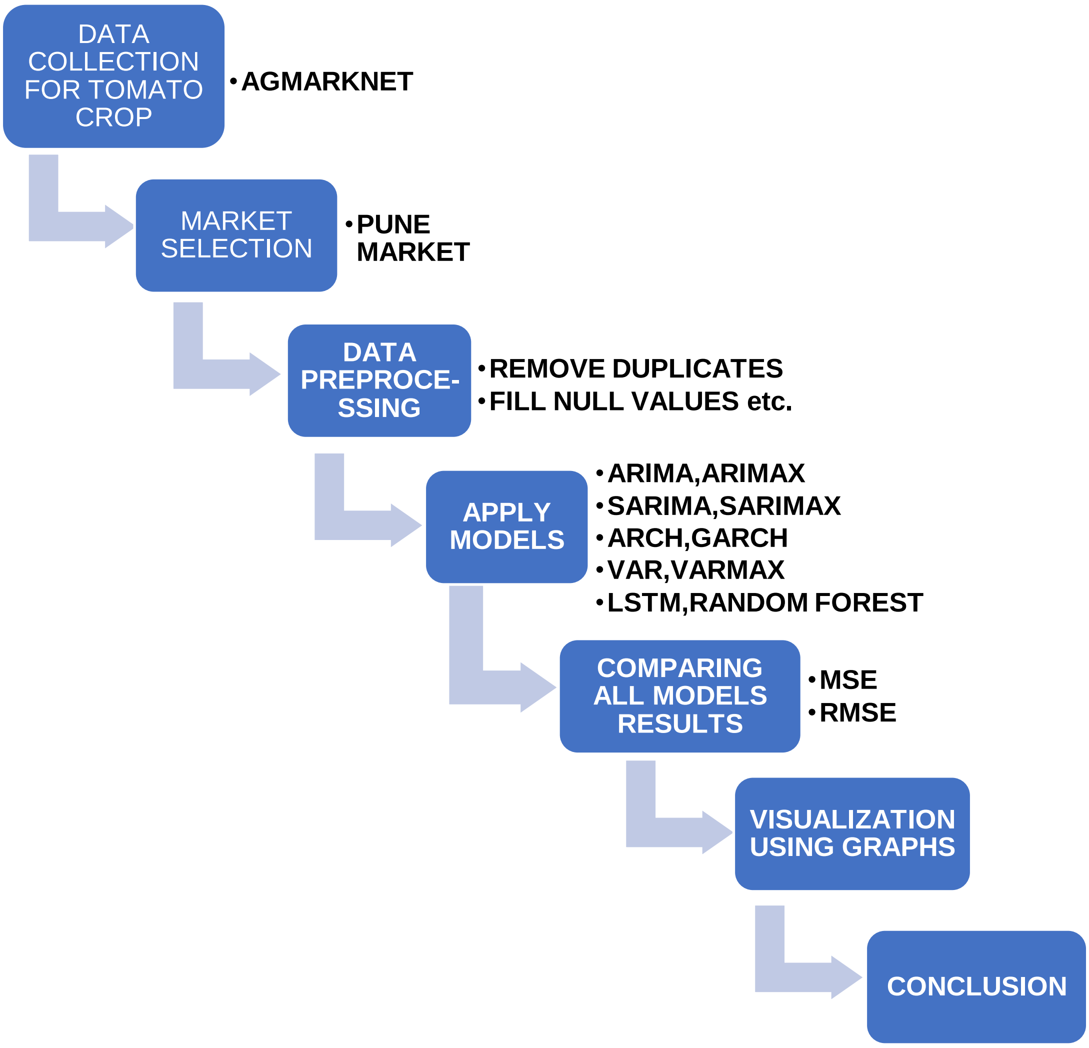
Using RF for one-step time series forecasting is straightforward and similar to the way that RF can be used for regression. Let g be the function obtained from the training of RF, which will be used for forecasting $\mathbf{x_{n+1}}$, given $\mathbf{x_1}$, …, $\mathbf{x_n}$. If we use k lagged variables then the forecasted $\mathbf{x_{n+1}}$ is given by the following equation for t = n + 1 [7]:

$$\mathbf{x_t} = \boldsymbol{g}(\mathbf{x_{t-1}}, \dots, \mathbf{x_{t-k}}), \boldsymbol{t} = \boldsymbol{k} + \mathbf{1}, \dots, \boldsymbol{n} + \mathbf{1}$$

The function g is not in closed form, but can be obtained by training the RF algorithm using a training set of size n − k. In each sample of the fitting set the dependent variable is $\mathbf{x_t}$, for t = k + 1, …, n + 1, while the predictor variables are $\mathbf{x_{t-1}}$, …, $\mathbf{x_{t-k}}$. When the number of predictor variables k increases, the size of the training set n − k decreases.

The algorithm excels in handling non-linear relationships and diverse data types, including numerical and categorical features. Its robustness comes from the ensemble approach, which reduces overfitting and improves generalization. The number of trees in the forest is a hyperparameter that is set during the training process. For predictions, each tree generates an output, and the final result is typically the average of these outputs.

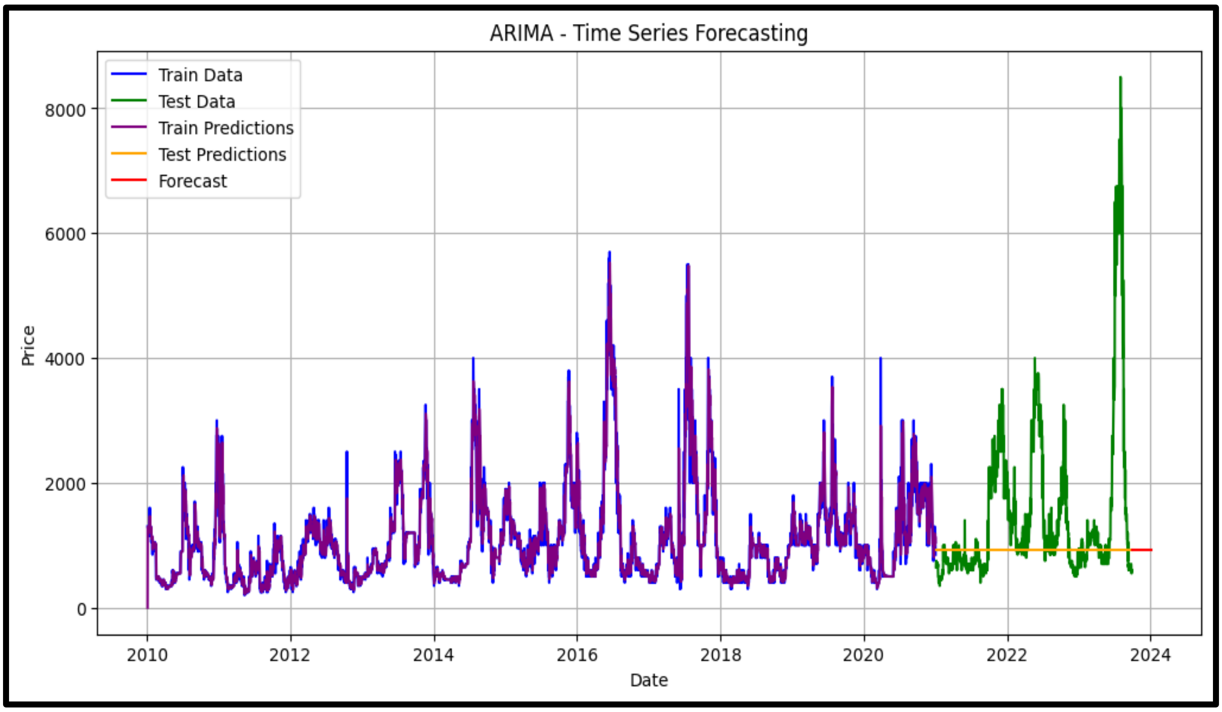
FLOWCHART


DATA COLLECTION FOR TOMATO CROP
• AGMARKNET
MARKET SELECTION
• PUNE MARKET
DATA PREPROCE-SSING
• REMOVE DUPLICATES
• FILL NULL VALUES etc.
APPLY MODELS
• ARIMA,ARIMAX
• SARIMA,SARIMAX
• ARCH,GARCH
• VAR,VARMAX
• LSTM,RANDOM FOREST
COMPARING ALL MODELS RESULTS
• MSE
• RMSE
VISUALIZATION USING GRAPHS
CONCLUSION

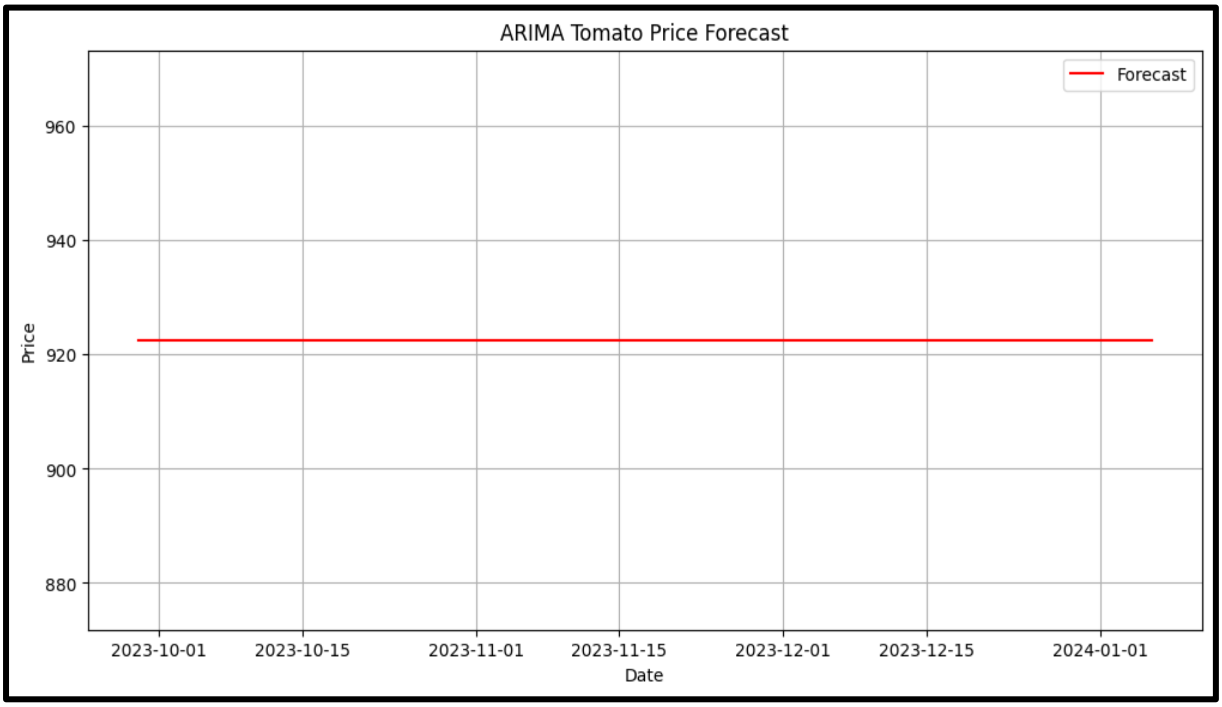
RESULTS

1. ARIMA MODEL

The best-fitting model is ARIMA (p,d,q)(P,D,Q,S) (0,1,1)(0,0,0,0).

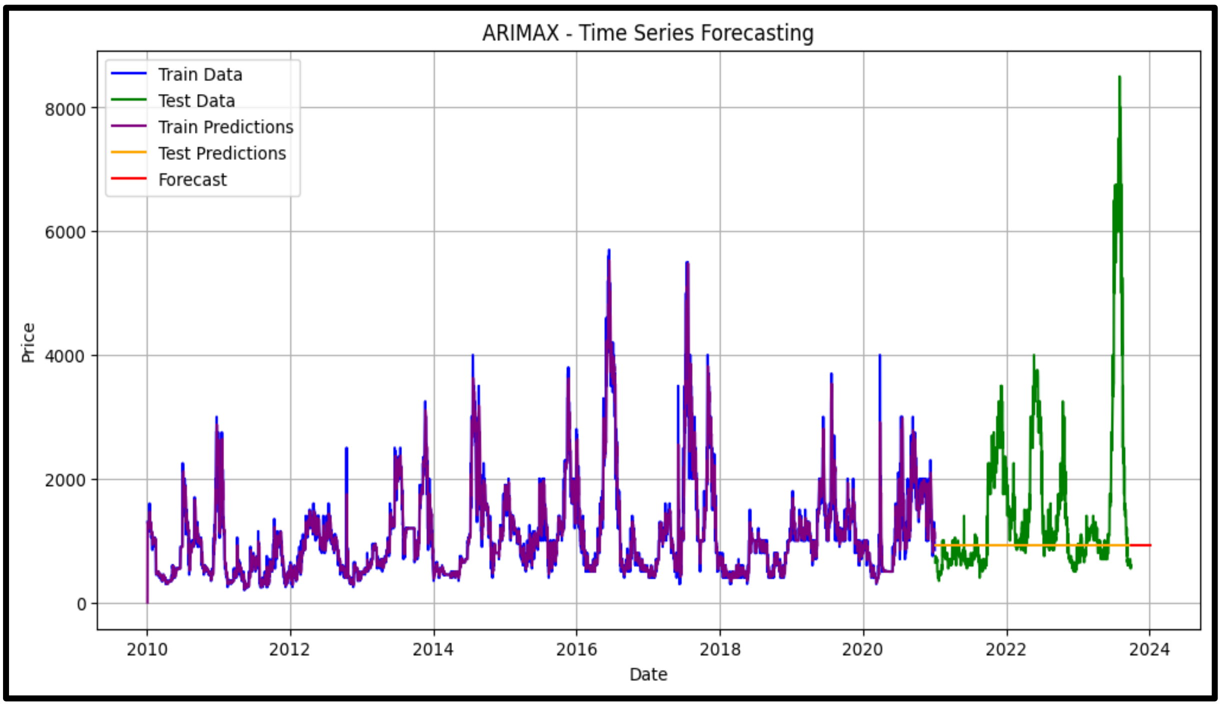


Price Forecast

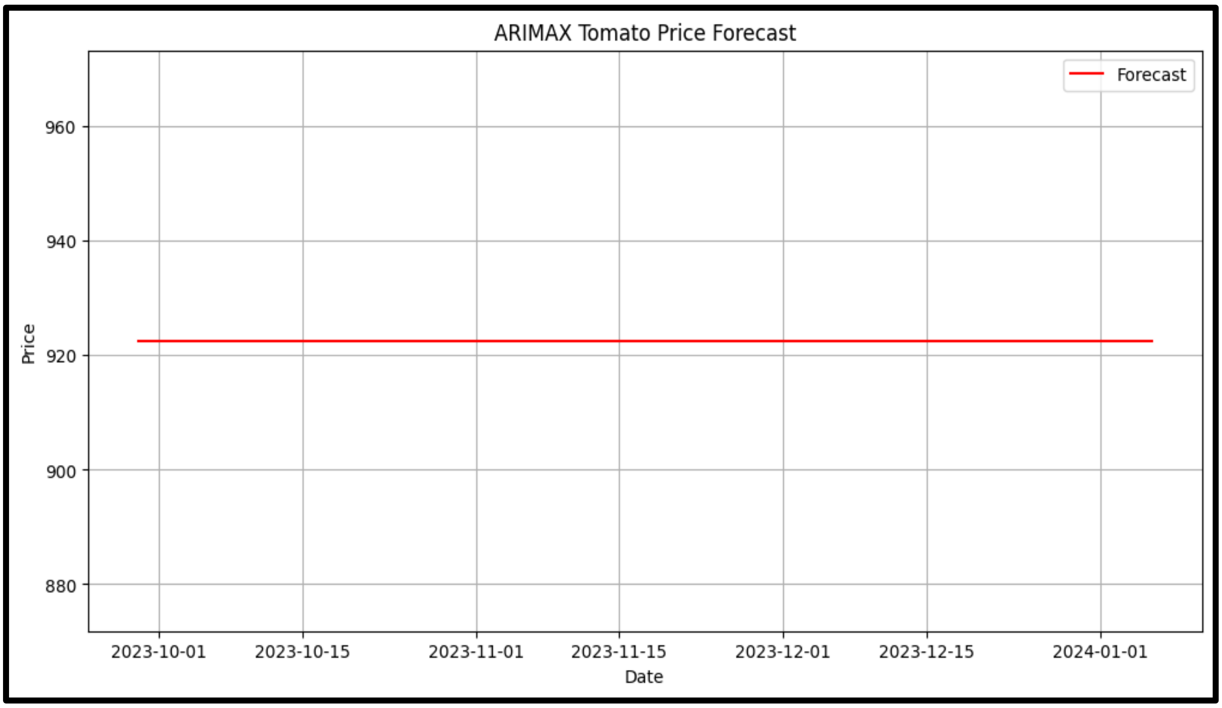


2. ARIMAX MODEL

The best-fitting model is ARIMAX (p,d,q)(P,D,Q,S) (0,1,1)(0,0,0,0).

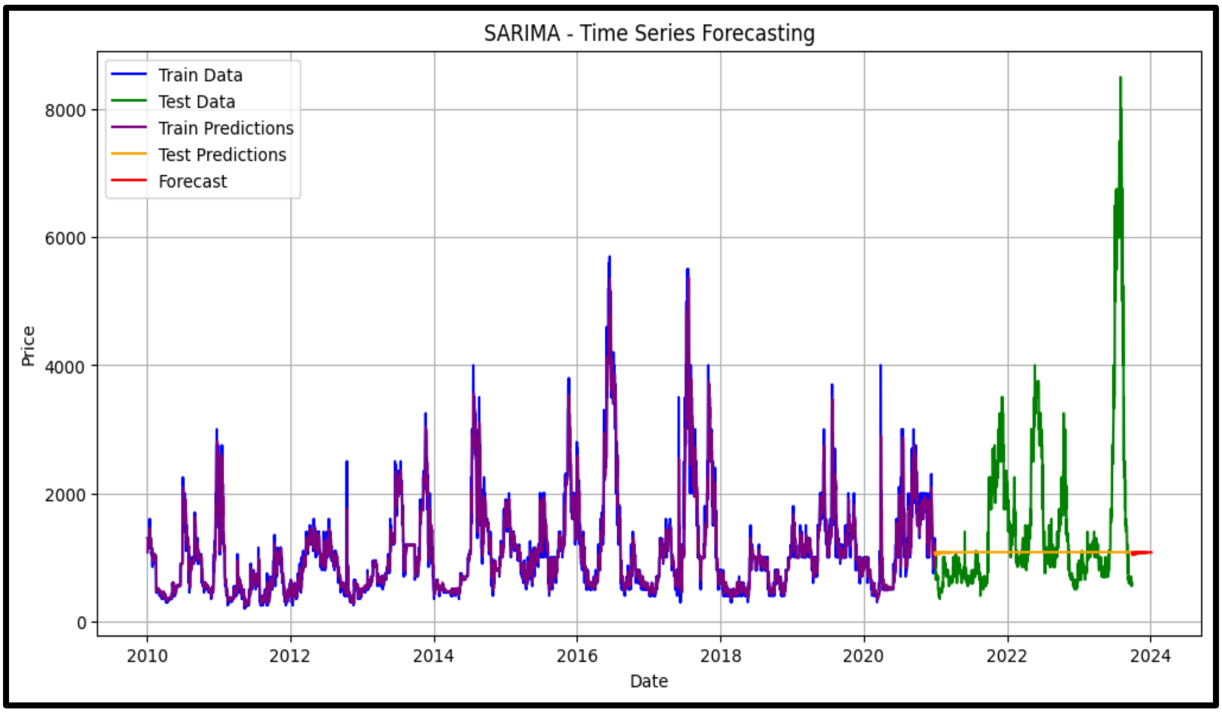


Price Forecast

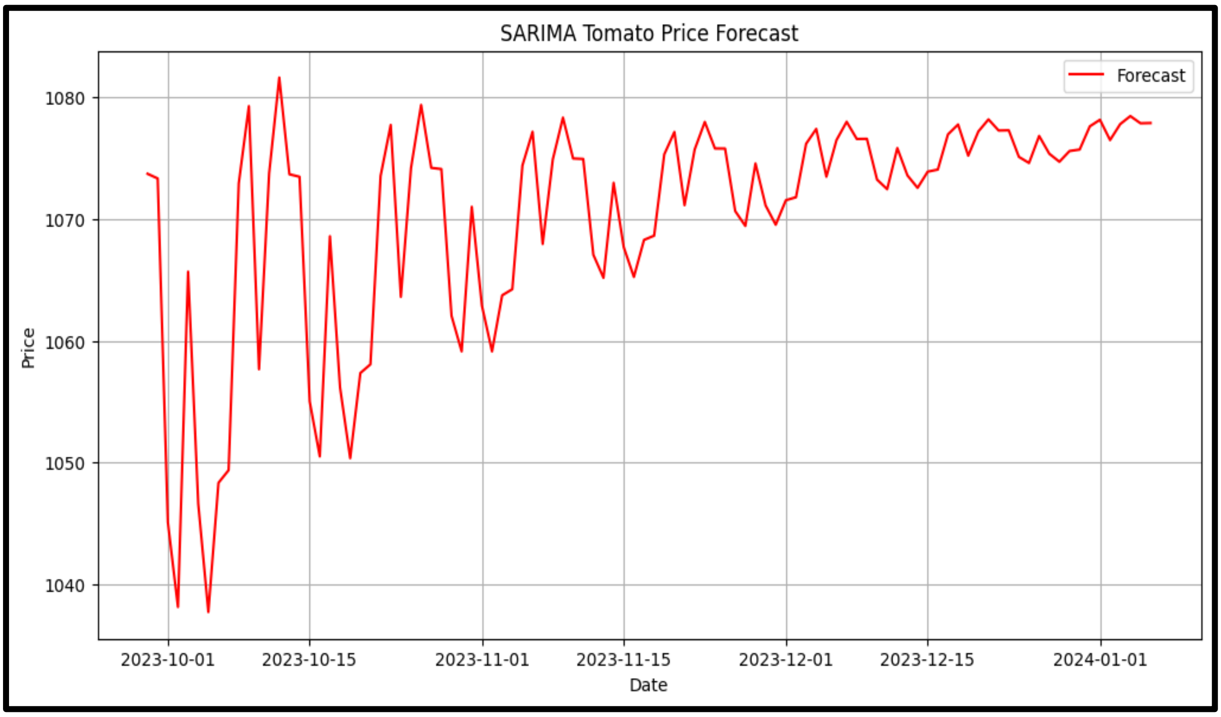


3. SRIMA MODEL

The best-fitting model is SRIMA (p,d,q)(P,D,Q,S) (1,0,1)(1,0,1,7).

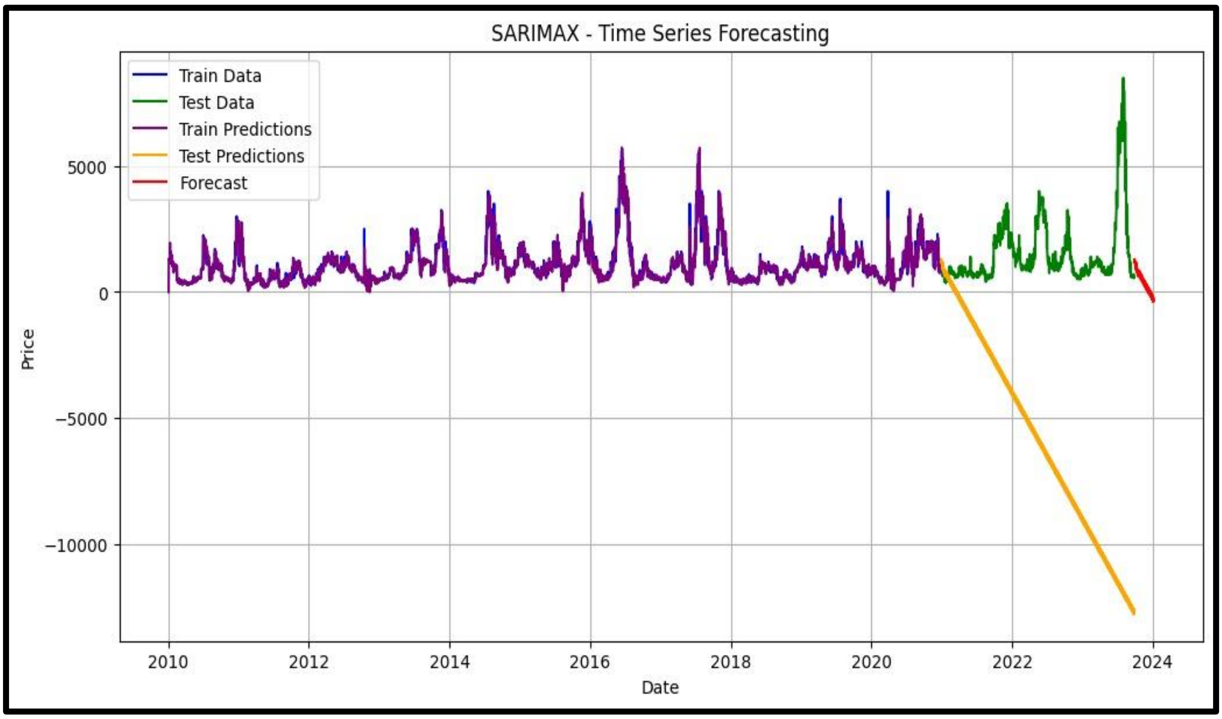


Price Forecast

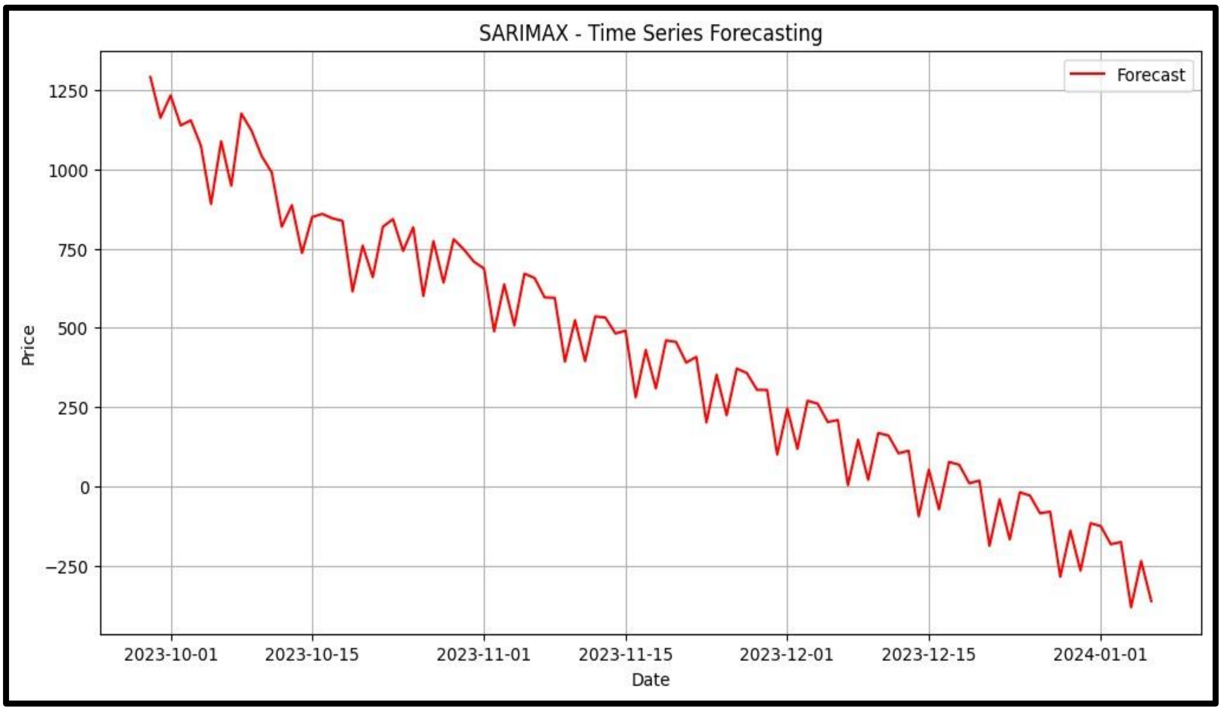


4. SRIMAX MODEL

The best-fitting model is SRIMAX (p,d,q)(P,D,Q,S) (1,0,1)(1,0,1,7).

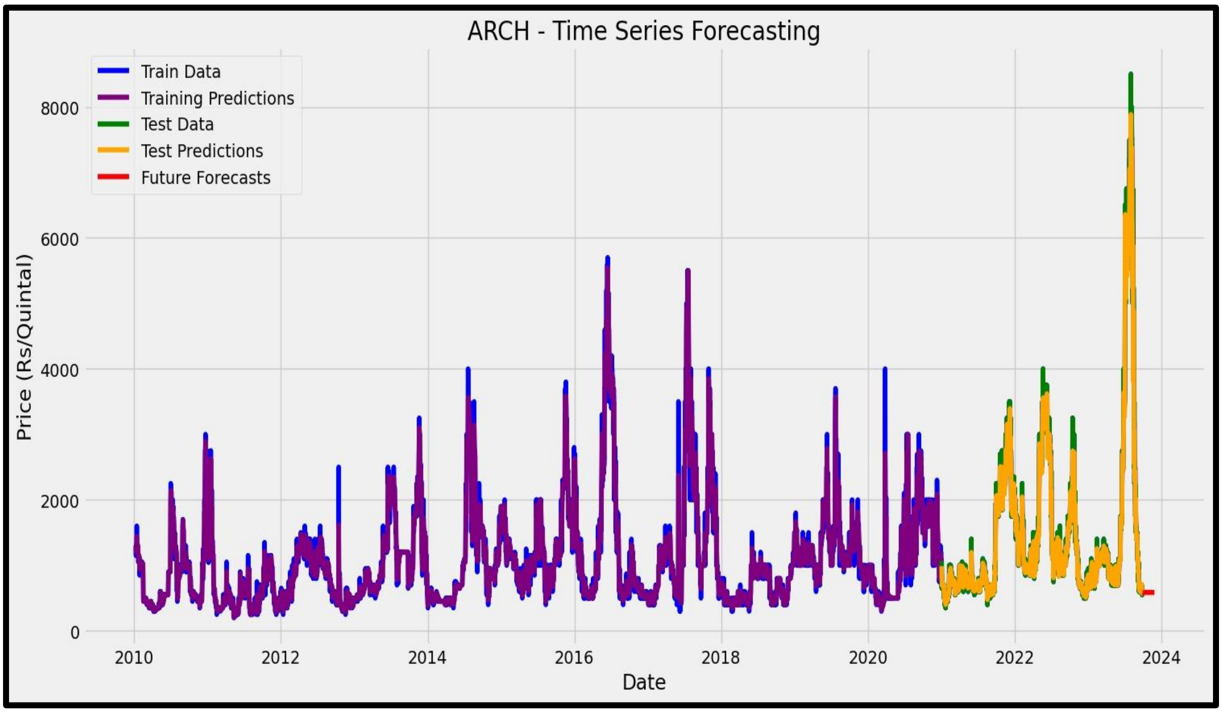


Price Forecast

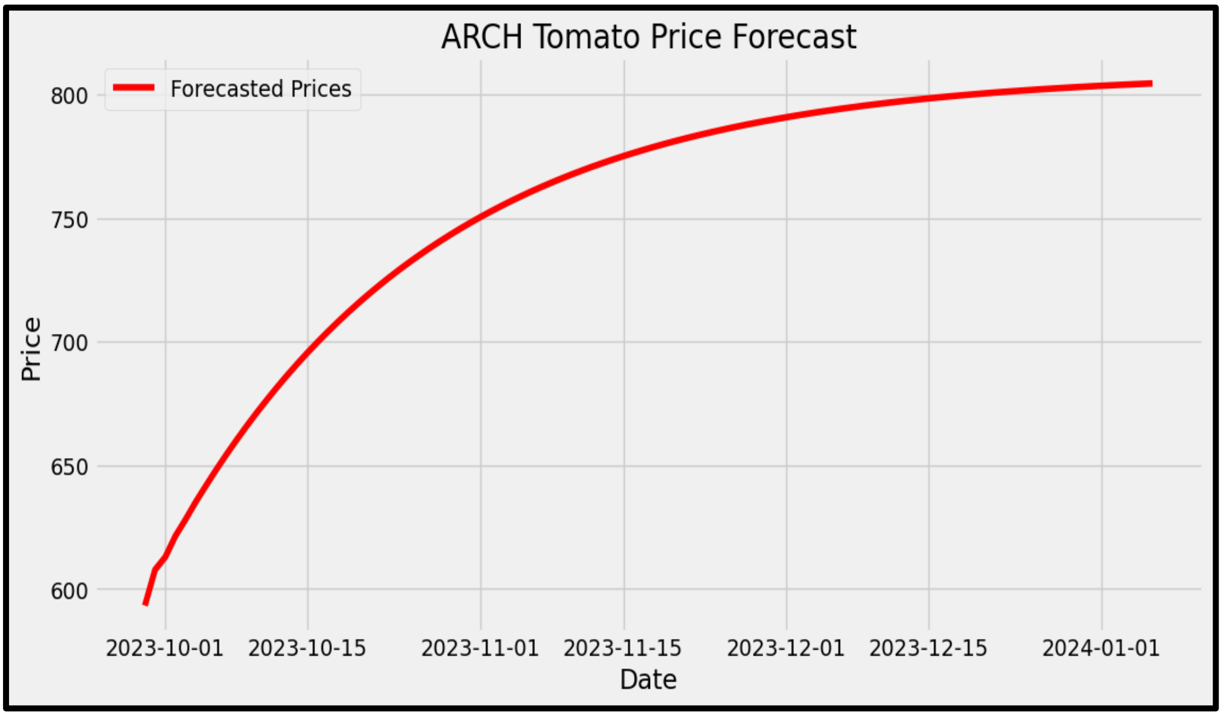


5. ARCH MODEL

Parameters	lags	vol	p,o,q	mean
Value	3	ARCH	2,1,2	AR

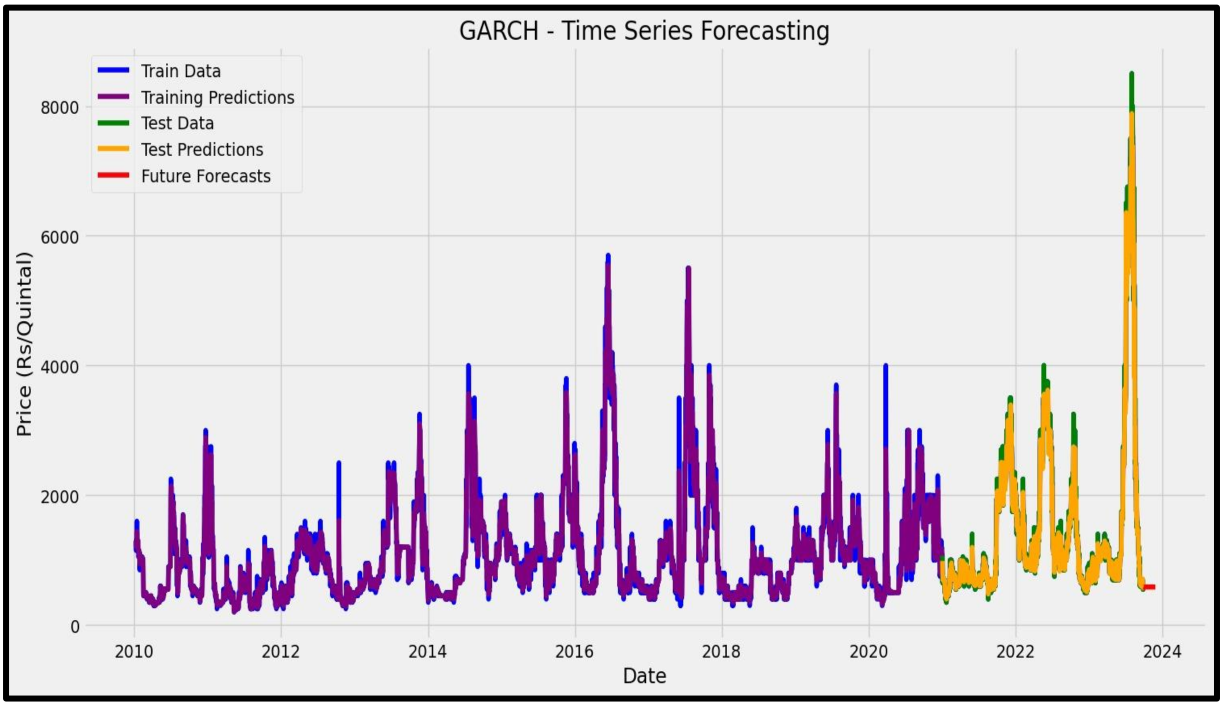


Price Forecast

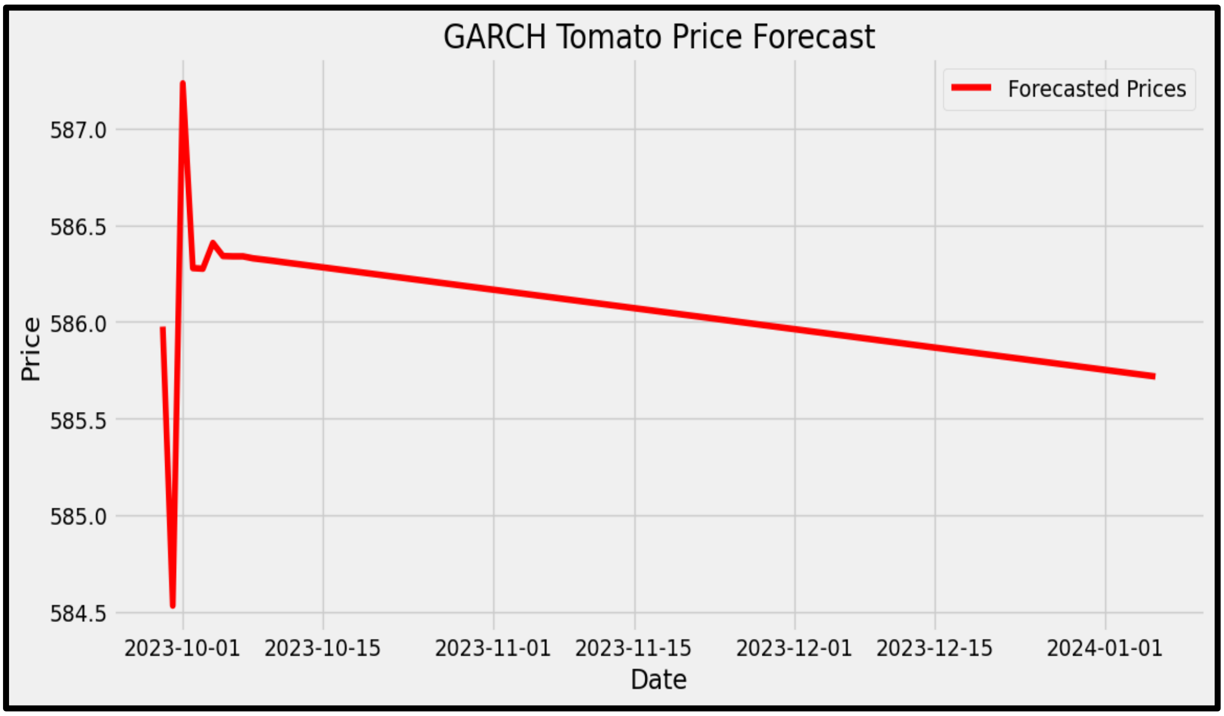


6. GARCH MODEL

Parameters	lags	vol	p,o,q	mean
Value	3	GARCH	2,1,2	AR



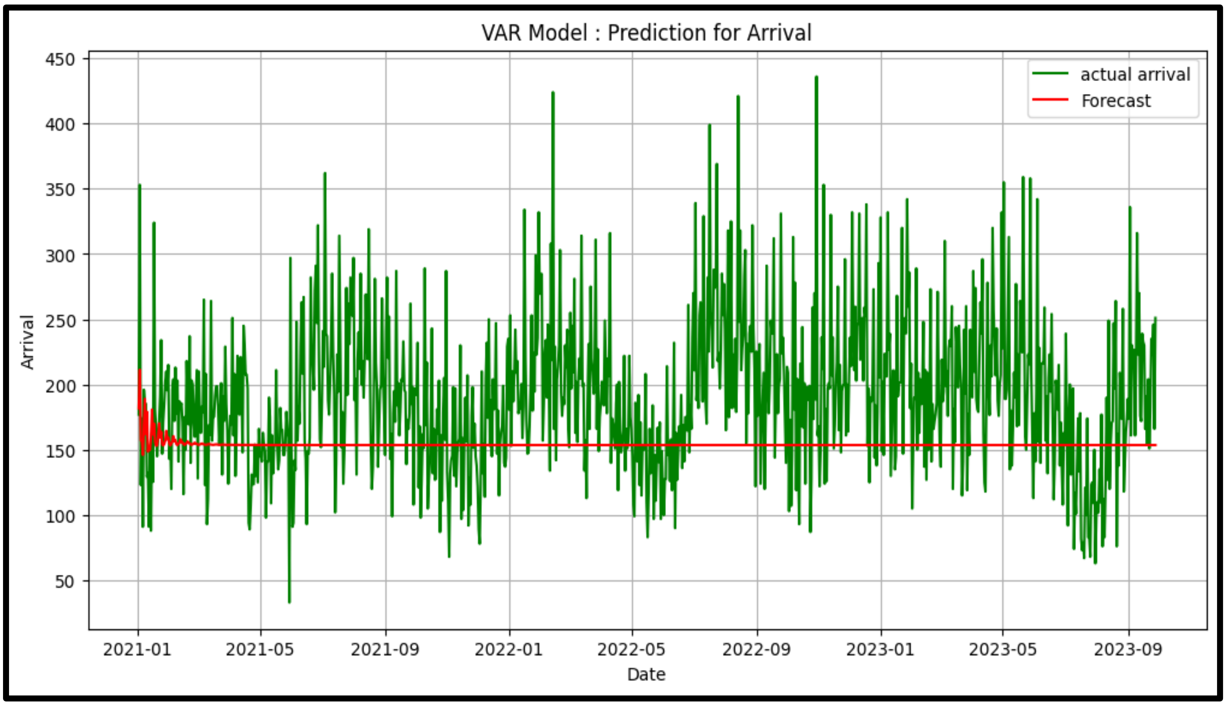
Price Forecast



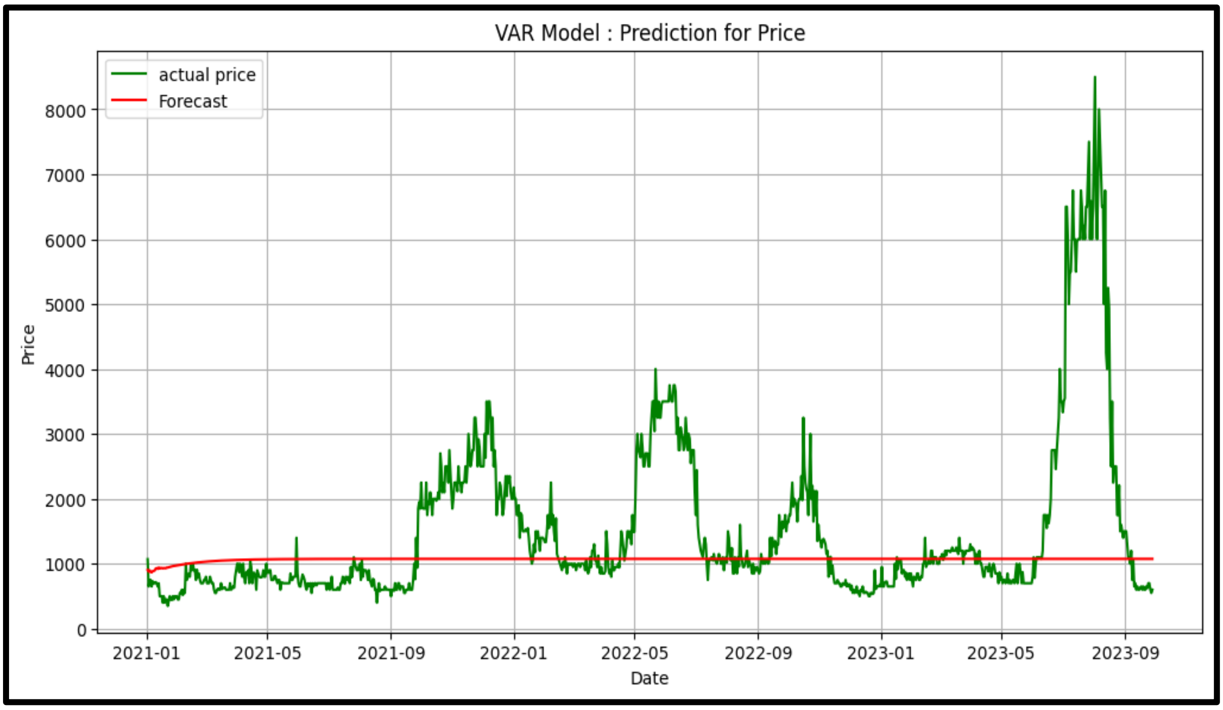
7. VAR MODEL

The selected VAR order is (20, 0).

Arrival Prediction



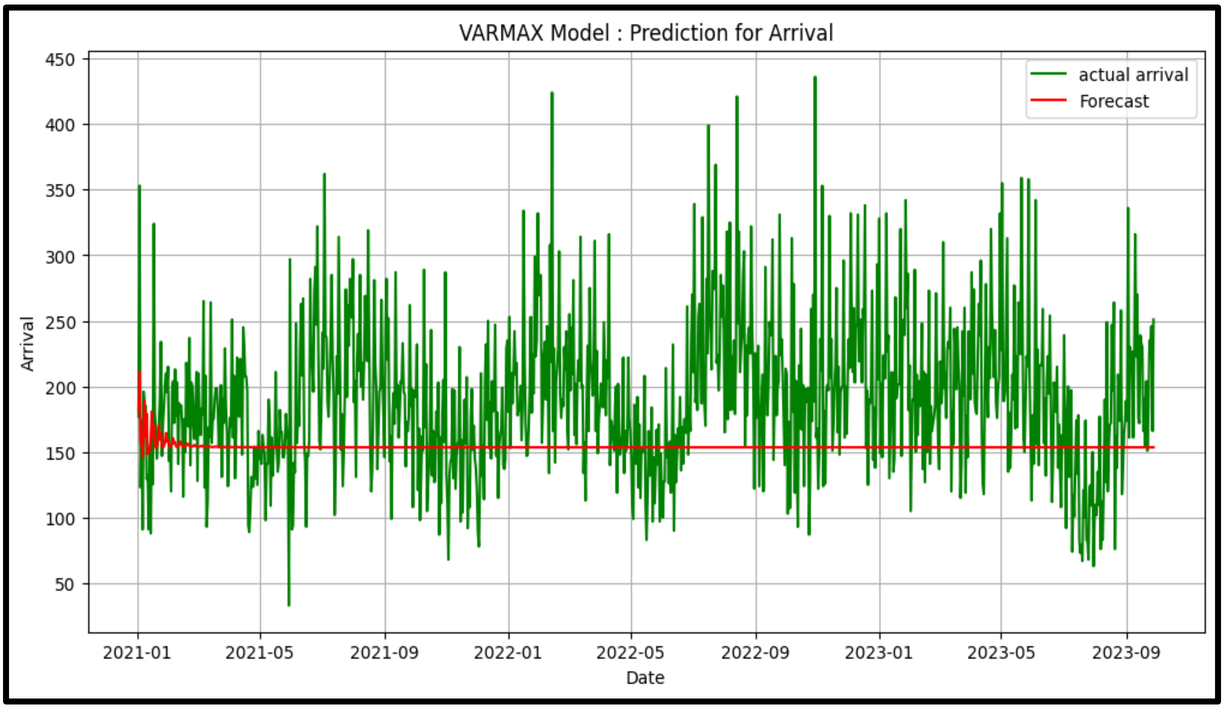
Price Prediction



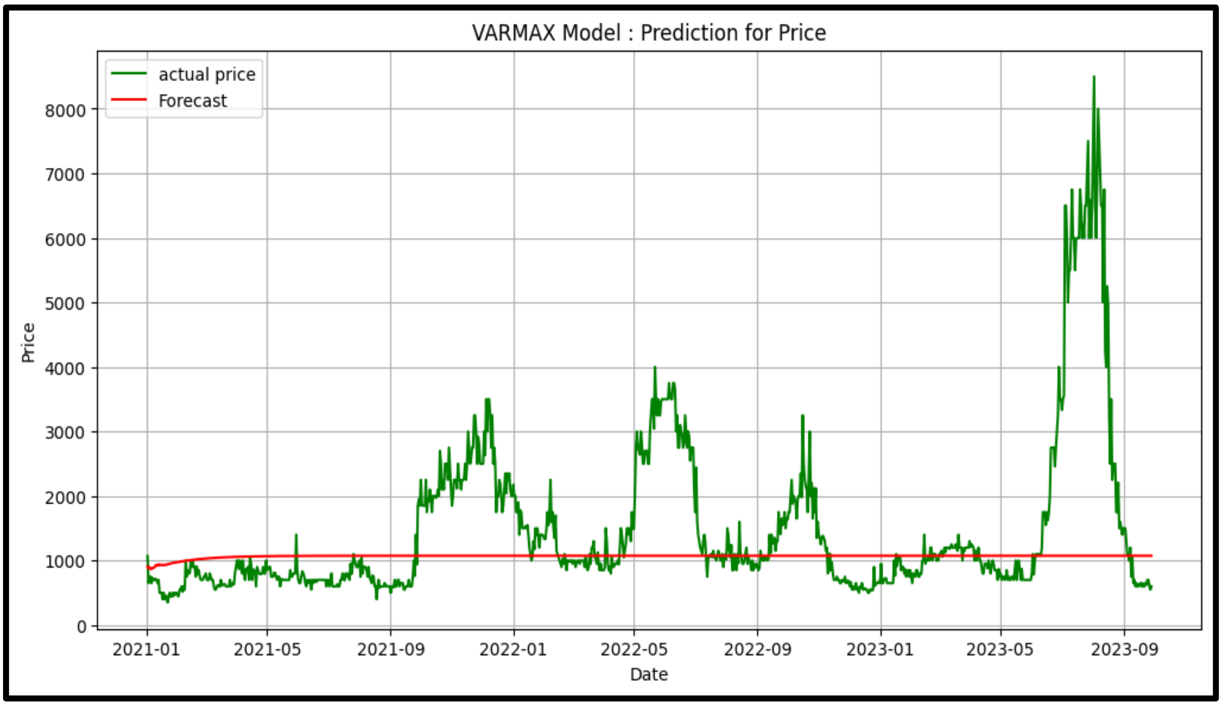
8. VARMAX MODEL

The selected VARMAX order is (20, 0).

Arrival Prediction

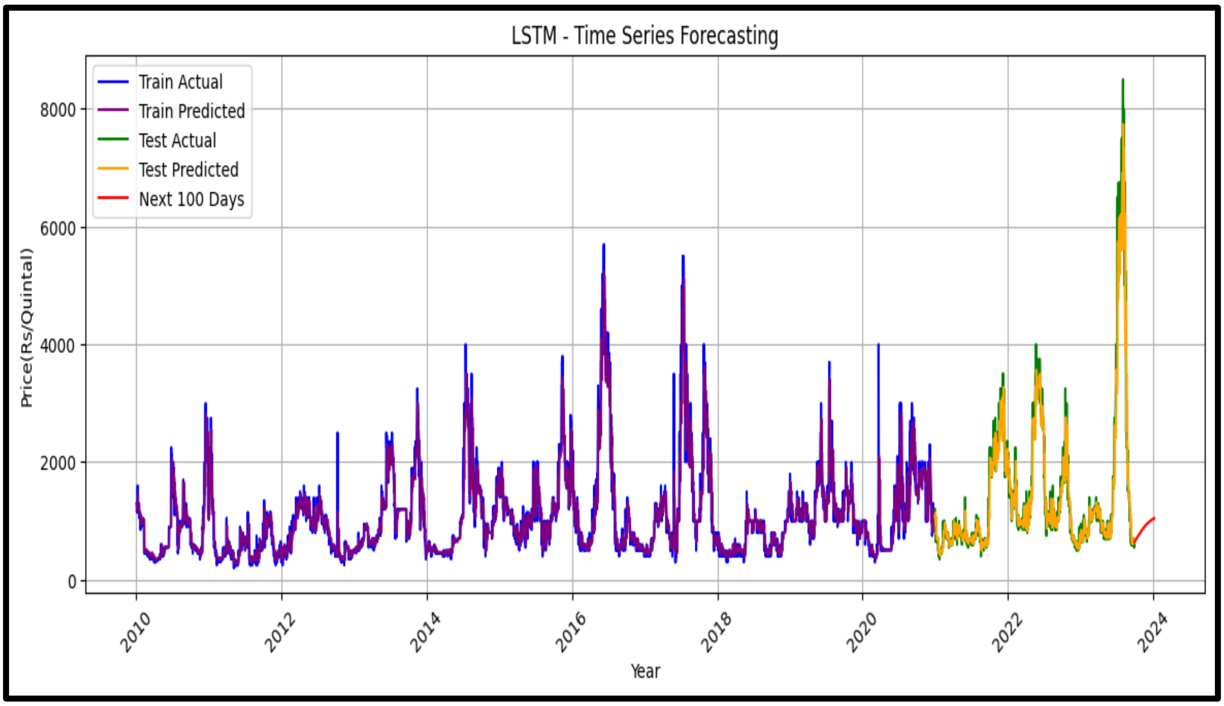


Price Prediction

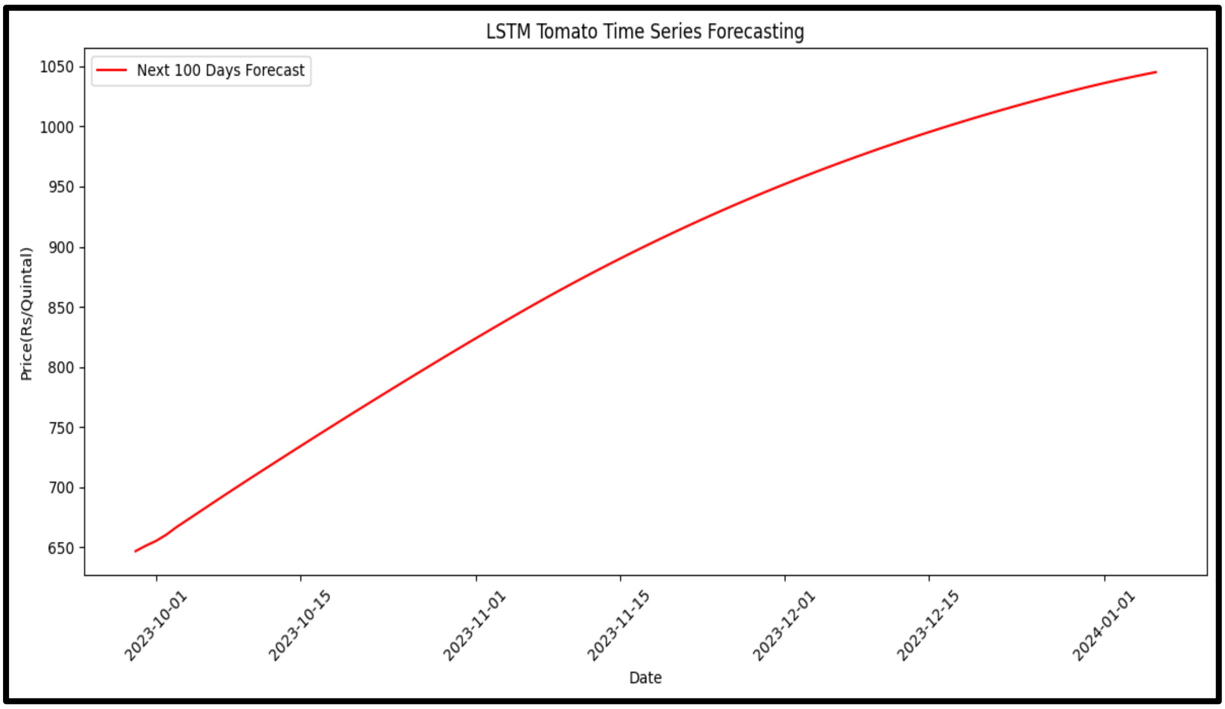


9. LSTM MODEL

Hyperparameter	Epochs	Activation Function	Optimizer	Loss Function
Value	50	ReLU	Adam	MSE, RMSE

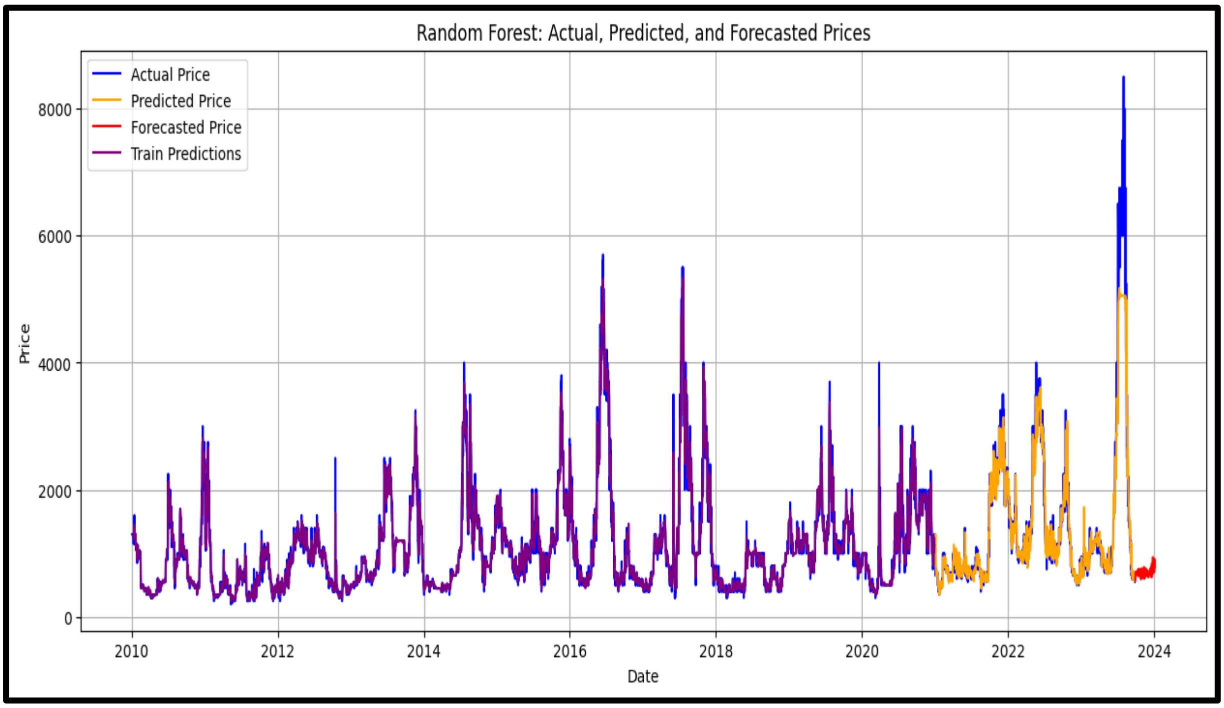


Price Forecast

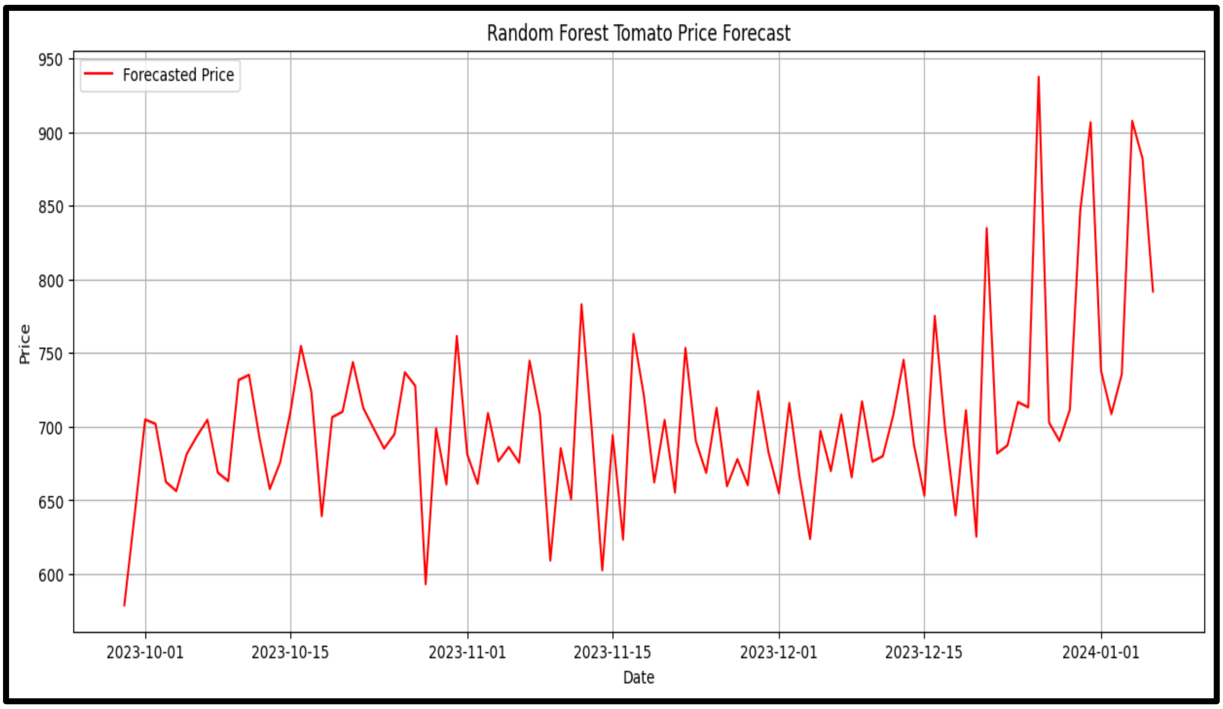


10. RANDOM FOREST MODEL

The model used 100 estimators (trees) to enhance predictive accuracy.



Price Forecast



➢ COMPARISON BETWEEN DIFFERENT MODELS

MODEL	MSE	RMSE	NRMSE	AIC	BIC
ARIMA	1991440.42	1411.18	0.94	54920.55	54933.14
ARIMAX	1991440.42	1411.18	0.94	54920.55	54933.14
SARIMA	1822751.36	1350.09	0.90	54839.50	54877.28
SARIMAX	74247134.82	8616.68	5.74	55638.49	55682.23
ARCH	14107.13	118.77	0.08	66856.30	66902.59
GARCH	14699.09	121.24	0.08	62889.40	62961.20
VAR PRICE	1828877.56	1352.36	0.90	102535.33	103070.68
VAR ARRIVAL	4619.92	67.97	0.36		
VARMAX PRICE	1830252.60	1352.86	0.90	102534.82	103070.17
VARMAX ARRIVAL	4604.97	67.86	0.36		
LSTM	120572.59	347.79	0.23	N/A	N/A
RANDOM FOREST	160475.85	400.59	0.27	N/A	N/A

CONCLUSION

Here are four key conclusions based on our study:

1. Effectiveness of Statistical Models:

   Among the traditional statistical models, SARIMA demonstrated the best performance with the lowest Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and RMSE (%) among ARIMA-based methods. This highlights SARIMA's suitability for handling seasonality in forecasting tomato prices.

2. Volatility Models for Risk Analysis:

   Volatility models (ARCH and GARCH) showed significantly lower MSE and RMSE values compared to other methods, indicating their effectiveness in capturing short-term price fluctuations and assessing related risks. However, their applicability is more focused on volatility rather than precise price forecasting.

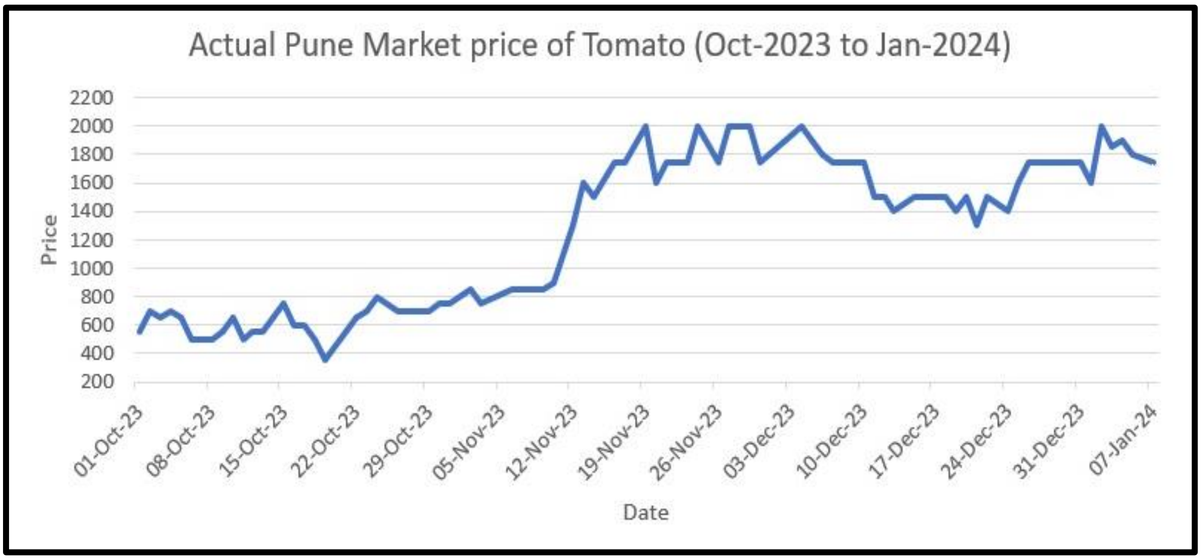
3. Machine Learning Superiority:

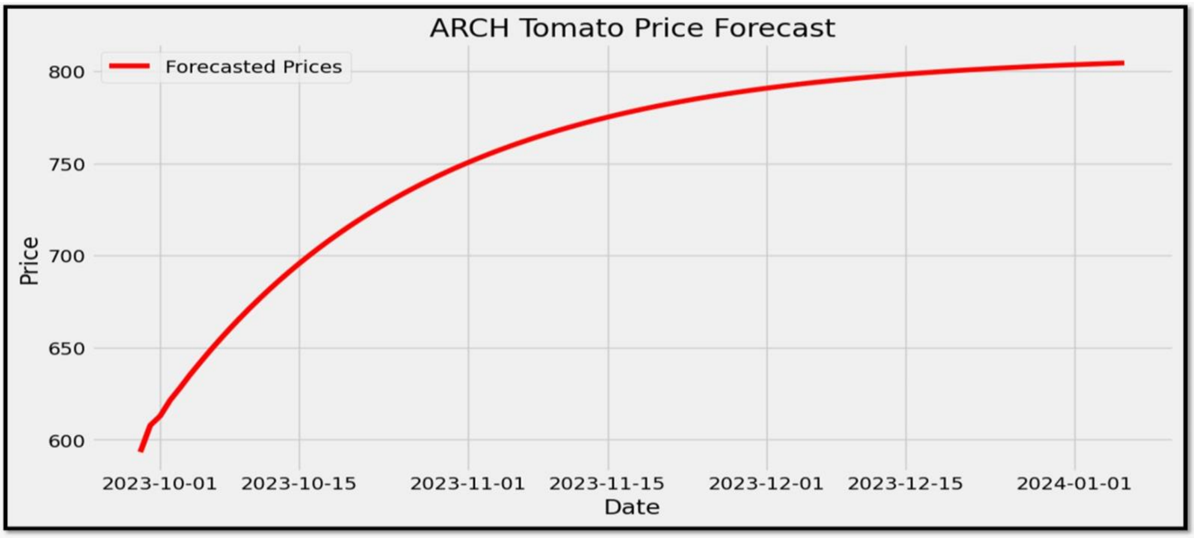
   The LSTM model achieved the lowest MSE and RMSE among all approaches, making it the most accurate in forecasting tomato prices. This underscores the strength of machine learning techniques, particularly LSTMs, in capturing complex patterns and long-term dependencies in time series data.

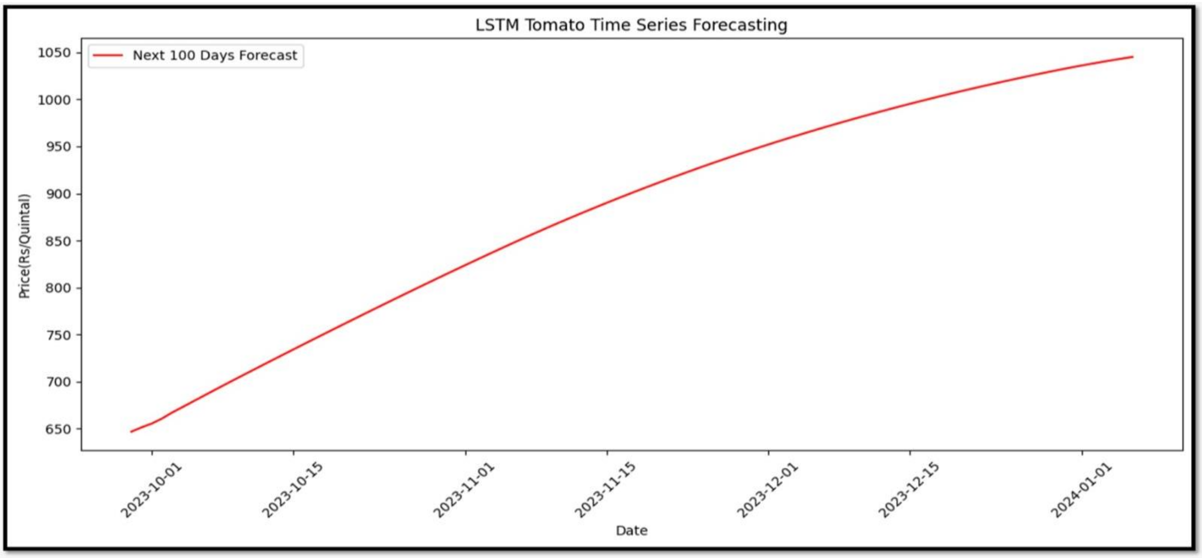
4. Price vs. Arrival Forecasting:

   Models predicting arrivals (VAR and VARMAX Arrival) performed better in terms of MSE and RMSE compared to price-focused models. This suggests that forecasting arrivals may be less complex and more predictable than price trends, likely due to fewer external influences or more stable patterns.

5. Comparison between actual market price and best models forecasted price:







The ARCH model's and LSTM model's forecasted prices for tomatoes show a steady increase over the period, predicting a gradual rise in prices without much fluctuation. In contrast, the actual market prices from Pune exhibit significant volatility, with sharp increases and decreases over time. The model effectively captures the general upward trend but underestimates the variability seen in real market conditions.

These findings highlight the complementary strengths of statistical, machine learning, and volatility models in addressing different aspects of price forecasting and risk analysis in agricultural markets.

REFERENCES

1. *Badal Prakash Singh.* 2023,"Article Tomato Price Forecasting - A Comparison between ARIMA, GARCH and ANN". http://doi.org/10.30954/2394-8159.01.2023.13

2. *Nari Sivanandam Arunraj, Diane Ahrens.* 2013,"Application of SARIMAX Model to Forecast Daily Sales in Food Retail Industry". http://doi.org/10.4018/IJORIS.2016040101

3. *Anju Bhagat, Rd Bansod.* 2023," Market integration and price transmission in major tomato markets of Maharashtra". https://www.researchgate.net/publication/374589599

4. *Pushpa Ghiyal, Joginder Kumar.* 2024,"An Application of multivariate time series models for forecasting the prices of tomato in Haryana". https://doi.org/10.22271/maths

5. *Erni Muschilati, Nursyiva Irsalinda.* 2020,"Forecasting Tourist Visit Using the Vector Autoregressive Exogenous Method (VARX)". http://doi.org/10.26555/konvergensi.v7i2.19608

6. *G. Avinash, Ramasubramanian V.* 2024,"Price Forecasting of TOP (Tomato, Onion and Potato) Commodities using Hidden Markov-based Deep Learning Approach". https://www.researchgate.net/publication/385631085

7. *Hristos Tyralis, Georgia Papacharalampous.* 2017," Variable Selection in Time Series Forecasting Using Random Forests". https://doi.org/10.3390/a10040114

8. *Sonvanee, O. P., and Pankaj Bhargav.* 2024, “Forecasting Tomato Price and Arrival Patterns in Krishi Upaj Mandis, Rajnandgaon, Chhattisgarh Using ARIMA Models”. https://doi.org/10.9734/ajaees/2024/v42i82528