Comparative Analysis of Price Prediction Models for Cotton:

A Statistical and Machine Learning Approach

## M.Sc. Agriculture Analytics

## MODULE-3

AGRICULTURAL MARKET ANALYTICS

Submitted by:

GROUP-4

*PRADIP PATELIA [202219018]*

*DHRUVI SHAH [202219019]*

*DIP PATEL [202219024]*

*NISHTHA AHUJA [202219025]*

*RACHIT PATEL [202219028]*

Submitted to:

*DR. PRITY KUMARI*

**ABSTRACT:**

The agricultural sector has long been the backbone of India's economy, employing a significant portion of its population and contributing substantially to its GDP. Cotton, a key player in Indian agriculture, takes center stage in this complex agricultural panorama. As the world's second-largest cotton producer, India has a historical association with this versatile crop spanning thousands of years. Approximately 5.8 million farmers derive their livelihoods from cotton cultivation, and millions more are involved in various aspects of the cotton industry. Cotton price prediction is a crucial aspect of agricultural economics, offering insights into the future movements of this vital global commodity. Various methods are employed, including traditional time series analysis models, machine learning algorithms (neural networks), fundamental analysis, and statistical models. The project aims to predict cotton’s futures prices by analyzing and modeling time series data on Cotton Prices. The study demonstrates how historical price data can be effectively utilized to predict future prices. We developed several models, including ARIMA, ARIMAX, ARCH, ARCHX, VAR, and LSTM models, employing a time series approach. The appropriate model selection was based on four performance criteria: the Akaike criterion, Schwarz Bayesian criterion, mean squared error, and root mean squared error. The results obtained underscore the model's efficacy in accurately modeling future cotton prices. This study offers valuable insights for stakeholders in the cotton industry, providing a robust methodology for making informed decisions grounded in historical pricing trends. It serves as a practical tool for various stakeholders along the cotton supply chain, including farmers, traders, the textile industry, and policymakers for risk management and strategic planning in the dynamic cotton commodity market.

**INTRODUCTION:**

Indian agriculture is a vital sector that has been the backbone of the country's economy for centuries. It plays a crucial role in providing livelihoods to a significant portion of the population, contributing to food security, and influencing the overall economic landscape of the nation. A substantial percentage of India's workforce is employed in agriculture, highlighting its significance in the country's employment landscape. As of now, approximately ~55% of the Indian population is engaged in agriculture, making it one of the largest employers in the country. This underlines the sector's pivotal role in sustaining rural livelihoods and supporting millions of families across the nation.

The recent Economic Survey for the year 2020-2021 indicates a noteworthy development in India's economic landscape: the share of agriculture in the country's GDP has reached almost 20%, a milestone not observed in the past 17 years. This underscores the substantial role played by the agriculture sector in shaping India's economic trajectory. Interestingly, this contribution surpasses the global average for agriculture in GDP at 6.4%. In contrast, India's industry and services sectors fall below their respective global averages, accounting for 30% and 63%, respectively. This economic dynamic highlights the continued significance of agriculture in India's overall economic composition, showcasing its resilience and impact on the nation's prosperity.

In the midst of India's vibrant agricultural panorama, cotton emerges as a key player, drawing attention due to its significant economic impact and strategic importance. As we delve into the diverse facets of Indian agriculture, where the sector has been the historical backbone of the nation's economy, cotton cultivation takes center stage. This fibrous crop contributes substantially to the country's agricultural output and plays a pivotal role in shaping economic dynamics. Cotton, an integral component of India's agricultural heritage, is paramount in the country's agricultural landscape. As the world's second-largest producer of cotton, India's historical association with this versatile crop dates back thousands of years, emphasizing its enduring significance. The cotton sector sustains the livelihoods of approximately 5.8 million farmers, with millions more engaged in various facets of the cotton industry.

Examining recent data, India's cotton production reveals a substantial contribution to the global cotton market. In the agricultural year 2022-23, an estimated 130.61 lakh hectares were devoted to cotton cultivation, yielding approximately 343.47 lakh bales with an average yield of 447 kgs per hectare, according to the Committee on Cotton Production and Consumption (COCPC) as of June 1, 2023. Despite India's substantial role in global cotton production, there remains room for improvement in yield efficiency, as the current yield of 447 kgs per hectare falls below the world average of 756 kgs per hectare. Beyond the fields, the impact of cotton extends into India's domestic textile industry, which stands as one of the largest in the nation. The industry has experienced remarkable growth in the last two decades, witnessing advancements in technology, such as installing open-end rotors and establishing export-oriented units. As a testament to its robustness, India has emerged as a significant consumer of cotton, accounting for about 22% of the world's cotton consumption.

This intricate interplay of historical significance, agricultural practices, and industrial utilisation underscores the multifaceted role of cotton in India's socio-economic fabric. As we navigate the intricate threads of cotton's journey in India, it becomes evident that this crop is not just a commodity; it's a cornerstone of livelihoods, a driver of industrial growth, and a reflection of the nation's dynamic agricultural prowess.  
  
Cotton price prediction is a critical component within the dynamic realm of agricultural economics, offering a predictive lens into the future movements of this vital global commodity. Various methods have been developed to anticipate cotton price trends, ranging from traditional time series analysis, such as ARIMA models, to more advanced machine learning algorithms like neural networks. Fundamental analysis, statistical models, and a holistic assessment of global economic conditions further contribute to the diverse toolkit employed for accurate predictions. The significance of cotton price forecasting cannot be overstated, especially considering its vast implications for stakeholders across the cotton supply chain. For farmers, these forecasts guide decisions on when to plant and harvest, aiding in risk management amidst price fluctuations. Traders and merchants benefit by optimizing trading strategies and managing inventory effectively. The textile industry relies on these forecasts for budgeting, production planning, and strategic sourcing decisions, as insights into future input costs are invaluable. Policymakers also turn to cotton price forecasts to inform agricultural policies and comprehend the potential socio-economic impacts of price shifts. In essence, it can be concluded that Cotton price forecasting is a vital tool for informed decision-making and sustainable practices within the broader cotton industry.

**REVIEW OF LITERATURE:**

*Mohanapriya M1 et.al.,* their study, compares the forecasting performance of two models, the ARIMA model and the Vector Auto Regression (VAR) model, for predicting the futures trading volume of cotton. The researchers used monthly data on various variables related to cotton futures trading. The study found that the ARIMA model, specifically ARIMA (2, 0, 0), was better at forecasting the trading volume compared to the VAR model.

*Racine Ly et al.* explored using Long-Short Term Memory (LSTM) models, a type of recurrent neural network, to forecast cotton and oil prices. The study compares the performance of machine learning methods with traditional methods like Autoregressive Integrated Moving Average (ARIMA) models. The results indicate that while machine learning methods fit the data well, they do not consistently outperform ARIMA models. However, combining the forecasts from both types of models leads to better results. The study also suggests using a forecast averaging method and expanding the analysis to other commodity prices.

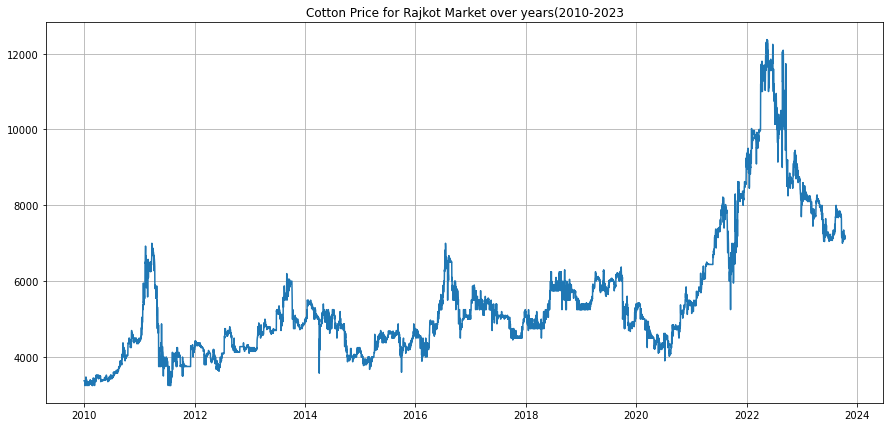
*Prema Borkar & P. M. Tayade* documented a study on modeling and forecasting cotton production in India using ARIMA models. The study concludes that the ARIMA (2, 1, 1) model is suitable for forecasting. The forecasts from 2015-16 to 2020-2021 can assist policymakers in predicting future grain storage, import, and export needs and enable them to take necessary actions.

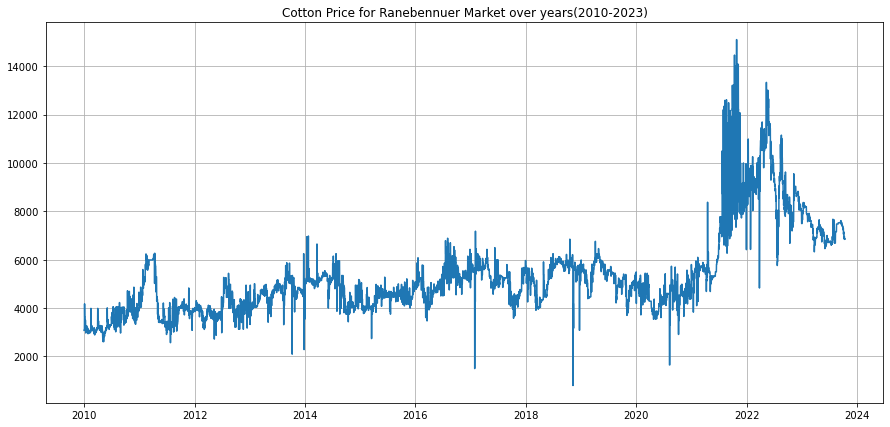
*Achal Lama et al.’s* research paper compares different models for forecasting price volatility in agricultural commodities. The study concludes that the exponential GARCH (EGARCH) model outperforms the autoregressive integrated moving-average (ARIMA) and generalized autoregressive conditional heteroscedastic (GARCH) models in predicting the international cotton price.

**DATASETS:**

Gujarat, Maharashtra, Karnataka, Andhra Pradesh, and Haryana are the largest cotton-producing states in India. As per availability, the time series data of daily prices of cotton from the AGMARKNET website for 13 years (from January 2010 to October 2023) has been collected for the Rajkot market of Gujarat and Ranebennur of Karnataka to predict the prices.

|  |  |  |
| --- | --- | --- |
|  | **MARKET: A** | **MARKET: B** |
| Commodity | Cotton | Cumin |
| State | Gujarat | Karnataka |
| District | Rajkot | Haveri |
| Market | Rajkot | Ranebennur |
| Time period | 1 Jan 2010 to 12 Oct 2023 | 1 Jan 2010 to 12 Oct 2023 |
| Price/ Arrival | Both | Both |
| Data available(days) | 3326 | 2737 |
| Minimum Price | 3250 | 800 |
| Maximum Price | 12370 | 15100 |





**METHODOLOGY:**

Different models like ARIMA, ARIMAX, ARCH, ARCHX, and LSTM are used to analyze and predict the cotton price data.

**ARIMA Model:**

ARIMA, which stands for Auto-Regressive Integrated Moving Average, is a widely used time series analysis and forecasting method. This model is particularly effective for predicting future values in a time series when there is a discernible pattern or trend. ARIMA combines autoregression (AR), differencing (I), and moving averages (MA) to capture different aspects of the time series data.

Components of ARIMA Model:

AutoRegressive (AR) Component:

The AR component refers to the regression of the series against its own past values. In simpler terms, it considers the linear relationship between the current value and its previous values. The "p" in ARIMA(p, d, q) denotes the order of the autoregressive component.

Integrated (I) Component:

The integrated component represents the differencing of the time series data to make it stationary. Stationarity is crucial for time series analysis, and differencing helps in removing trends or seasonality. The "d" in ARIMA(p, d, q) signifies the order of differencing.

Moving Average (MA) Component:

The MA component involves modeling the error term as a linear combination of past error terms. This part helps capture the data's random shocks or white noise. The "q" in ARIMA(p, d, q) represents the order of the moving average component.

The general form of an ARIMA model is denoted as ARIMA(p, d, q). The appropriate values for p, d, and q are determined through analysis of the autocorrelation and partial autocorrelation functions of the time series data.

PROCEDURE

To determine the order of the model to be fitted to the data, we need three variables: p, d, and q, which are nonnegative integers that refer to the order of the autoregressive, integrated, and moving average parts of the model, respectively. PACF and ACF are used to find out the values of p and q, respectively. Maximum Likelihood Estimation (MLE) to estimate the ARIMA model. Finding Akaike’s Information Criterion (AIC) and Schwarz Bayesian Information Criterion (BIC) for a set of models and investigating the models with the lowest AIC and BIC values. Along with AIC and BIC, we also need to watch those coefficient values closely and decide whether to include that component according to their significance level.

**ARIMAX Model:**

ARIMAX, or AutoRegressive Integrated Moving Average with eXogenous variables, is an extension of the ARIMA model. ARIMAX allows for the inclusion of additional external or exogenous variables that may influence the time series being analysed. This makes ARIMAX a more flexible and powerful tool, particularly when known factors outside the time series impact its behaviour.

Components of ARIMAX Model:

ARIMA Structure:

ARIMAX incorporates the autoregressive (AR), integrated (I), and moving average (MA) components of ARIMA. The ARIMA structure captures the temporal dependencies and trends within the time series.  
  
Exogenous Variables:

ARIMAX introduces exogenous variables, which are external factors not part of the time series but are believed to influence it. The model includes these variables to account for their impact on the dependent variable.

PROCEDURE

The procedure involves selecting appropriate orders (p, d, q) for the ARIMA component, including relevant exogenous variable (arrival), estimating the model, and evaluating its performance.

**VAR Model:**

Vector Autoregression (VAR) is a statistical model used in time series analysis to capture the interdependencies among multiple time series variables. Unlike univariate models that focus on a single variable, VAR models consider multiple variables simultaneously. VAR is a flexible and widely employed tool for understanding and forecasting the joint behavior of interconnected time series.  
  
Concepts of VAR Model:

Vector Formulation:

The VAR model is represented in vector form, where each variable is treated as a vector, and the model describes the evolution of the entire vector of variables over time.

Autoregressive Structure:

The VAR model assumes that each variable in the system is a linear function of its past values and the past values of all other variables in the system. This autoregressive structure is expressed in lagged terms.

Order of the Model (p):

The order of the VAR model, denoted as "p," represents the number of lagged observations considered for each variable. A VAR(p) model includes p lags for each variable.  
  
Error Term:

VAR models include an error term for each equation, capturing the unobserved factors affecting each variable that are not explained by the lagged values of the variables in the system.

PROCEDURE

First, the data are checked for stationarity using the Augmented Dickey-Fuller test at a 5% level of significance. The series is found to be non-stationary. We take the first difference of the data to make them stationary. Then fit a VAR model. In this case, models were estimated equation by equation using the principle of least squares. Coefficients were estimated using ACF and PACF, and the prediction was obtained.

**ARCH Model:**

Autoregressive Conditional Heteroskedasticity (ARCH) is a statistical model used to analyze and model a time series's volatility or variance, particularly in financial econometrics. These models are designed to capture the changing or conditional variance in a time series, recognizing that volatility can vary over time.

Concepts of ARCH Model:

Heteroskedasticity:

ARCH models address the issue of heteroskedasticity, which refers to the phenomenon where the variance of the errors in a time series is not constant over time. In financial markets, for example, volatility tends to exhibit clustering, with periods of high volatility followed by periods of low volatility.

Conditional Variance:

ARCH models assume that the variance of the error term at each time point is a function of past observations. The conditional variance is modeled as an autoregressive process, where past squared residuals contribute to the current conditional variance.

ARCH Model:

The order of the ARCH model, denoted as q, indicates the number of lagged squared residuals included in the model.

Parameter Estimation:

Parameters in ARCH models are typically estimated using maximum likelihood estimation (MLE). The process involves finding the parameter values that maximize the likelihood of observing the given data under the assumed model.

PROCEDURE  
To apply the ARCH model, you start by specifying the order of the model, and then estimate the model parameters. Once the model is estimated, you can analyze the results, including examining the conditional variance to assess the presence of ARCH effects in the data.

**ARCHX Model:**

Autoregressive Conditional Heteroskedasticity with eXogenous variable is an extension of the ARIMA model. ARIMAX allows for the inclusion of additional external or exogenous variables that may influence the time series being analysed.

Concepts of ARCHX Model:

ARCH Model:

The order of the ARCH model, denoted as q, indicates the number of lagged squared residuals included in the model.

Exogenous Variables:

ARCHX introduces exogenous variables, which are external factors not part of the time series but are believed to influence it. The model includes these variables to account for their impact on the dependent variable

Parameter Estimation:

Parameters in ARCH models are typically estimated using maximum likelihood estimation (MLE). The process involves finding the parameter values that maximize the likelihood of observing the given data under the assumed model.

PROCEDURE

The procedure involves selecting appropriate orders (p, d, q) for the ARCHX component, including relevant exogenous variable (arrival), estimating the model, and evaluating its performance.

**LSTM Model:**

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture designed to overcome the limitations of traditional RNNs in capturing and learning from long-term dependencies in sequential data. LSTMs are widely used in various applications, including natural language processing, time series analysis, and financial market predictions.

Components of LSTM:

Cell State:

LSTMs maintain a cell state, which serves as a memory unit that can capture information over long sequences. This enables LSTMs to remember important information and selectively update or forget it as needed.

Three Gates:

LSTMs have three gates—input gate, forget gate, and output gate—that control the flow of information through the cell state.

Input Gate: Regulates the flow of new information into the cell state.

Forget Gate: Manages the removal or "forgetting" of information from the cell state.

Output Gate: Determines the information to be output based on the cell state.

Hidden State:

The hidden state in LSTMs is responsible for carrying information throughout the sequence. It acts as a filtered version of the cell state and is used for making predictions.

Training and Backpropagation Through Time (BPTT):

LSTMs are trained using backpropagation through time, similar to traditional RNNs. However, LSTMs mitigate the vanishing gradient problem associated with RNNs, allowing them to learn long-range dependencies more effectively.

PROCEDURE

The model's training procedure involved using a sequence of the past 30 days of cotton prices as input to predict the price for the following day. For predicting cotton prices over the next 365 days, a rolling forecasting method was employed. The procedure commenced with the most recent 30 days of price data, from which the model generated a prediction for the subsequent day's cotton price. Now the predicted value along with the last 29 days of actual would be served as the basis for predicting the price for the day after the initial forecast. This iterative process was repeated until forecasts were generated for all 365 days within the forecast horizon.

**FLOWCHART:**

**DATA SOURCE**

**(AGMARKNET)**

**MARKET -A (RAJKOT, GUJARAT)**

**MARKET -B (RANEBENNUR, KARNATAKA)**

**DATA DOWNLADING AND PREPROCESSING**

**ARIMAX**

**ARCH**

**LSTM**

**VAR**

**ARCHX**

**ARIMA**

**COMPARING ALL MODEL**

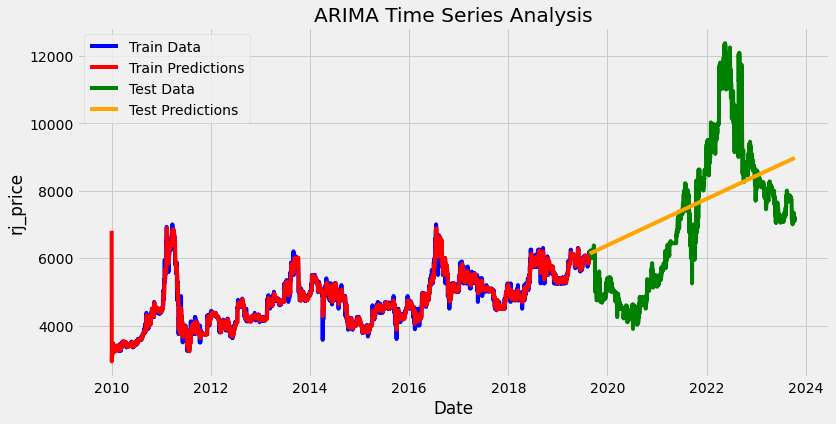
**BY RMSE, MSE, AIC, BIC SCORE**

**VISUALIZATION USING GRAPHS**

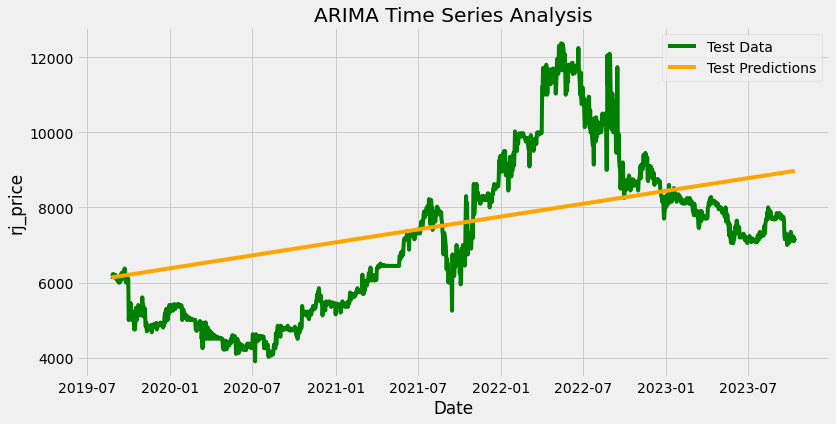
**RESULT:**

**ARIMA MODEL**

The model is best fitted at ARIMA (2,3,6).

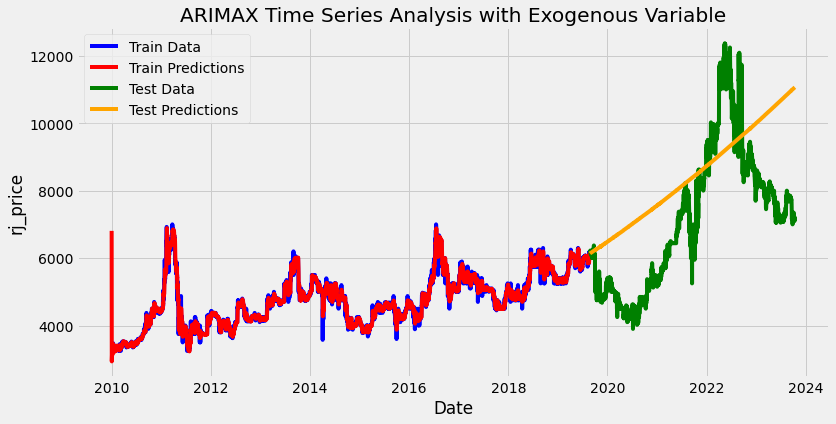


**Test prediction**

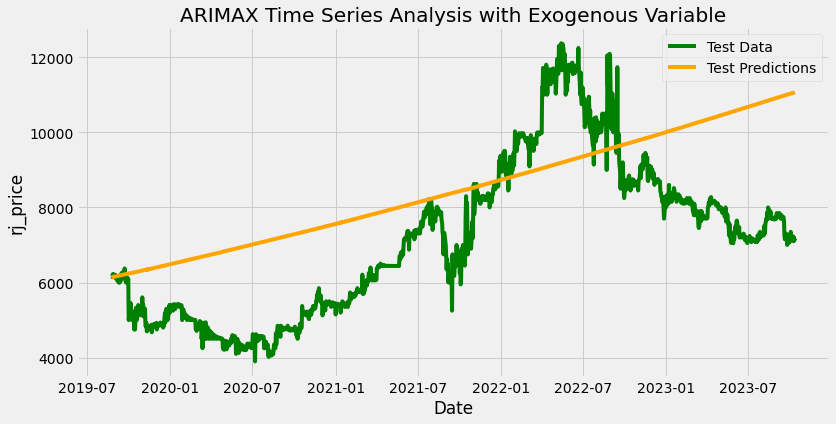


**ARIMAX MODEL**

The model is best fitted at ARIMAX (2,3,6)



**Test prediction**

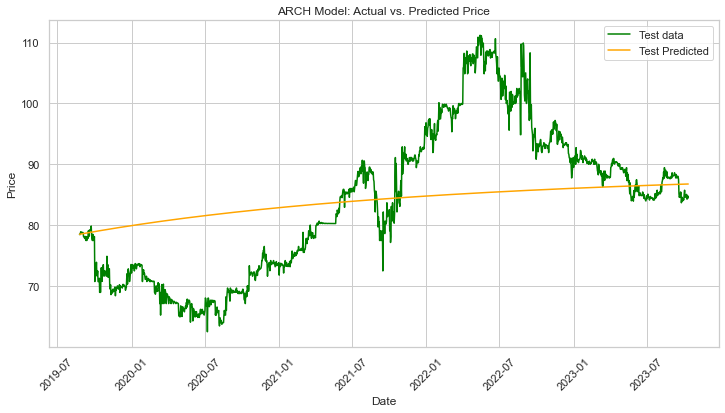


**ARCH MODEL**

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

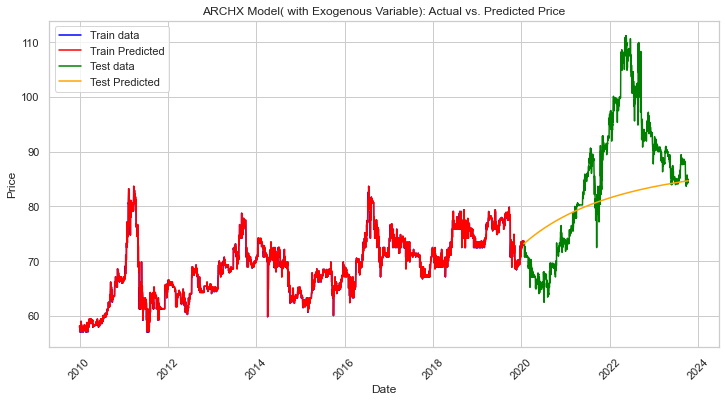


**Test prediction**

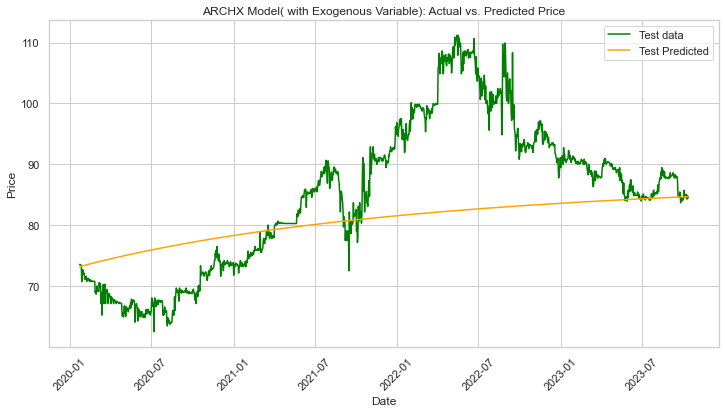


**ARCHX MODEL**

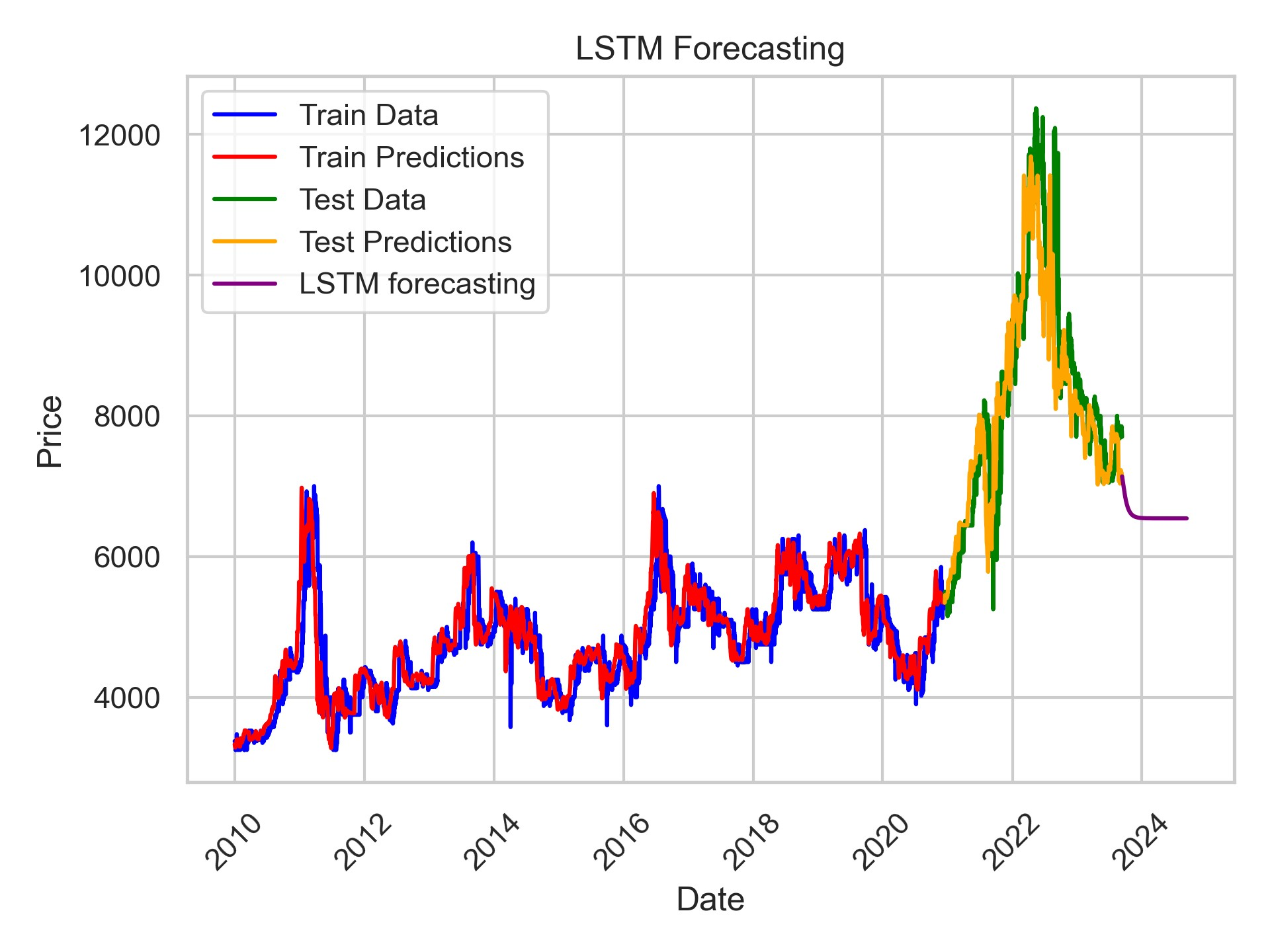
The model is best fitted at ARCHX (1,0,1).



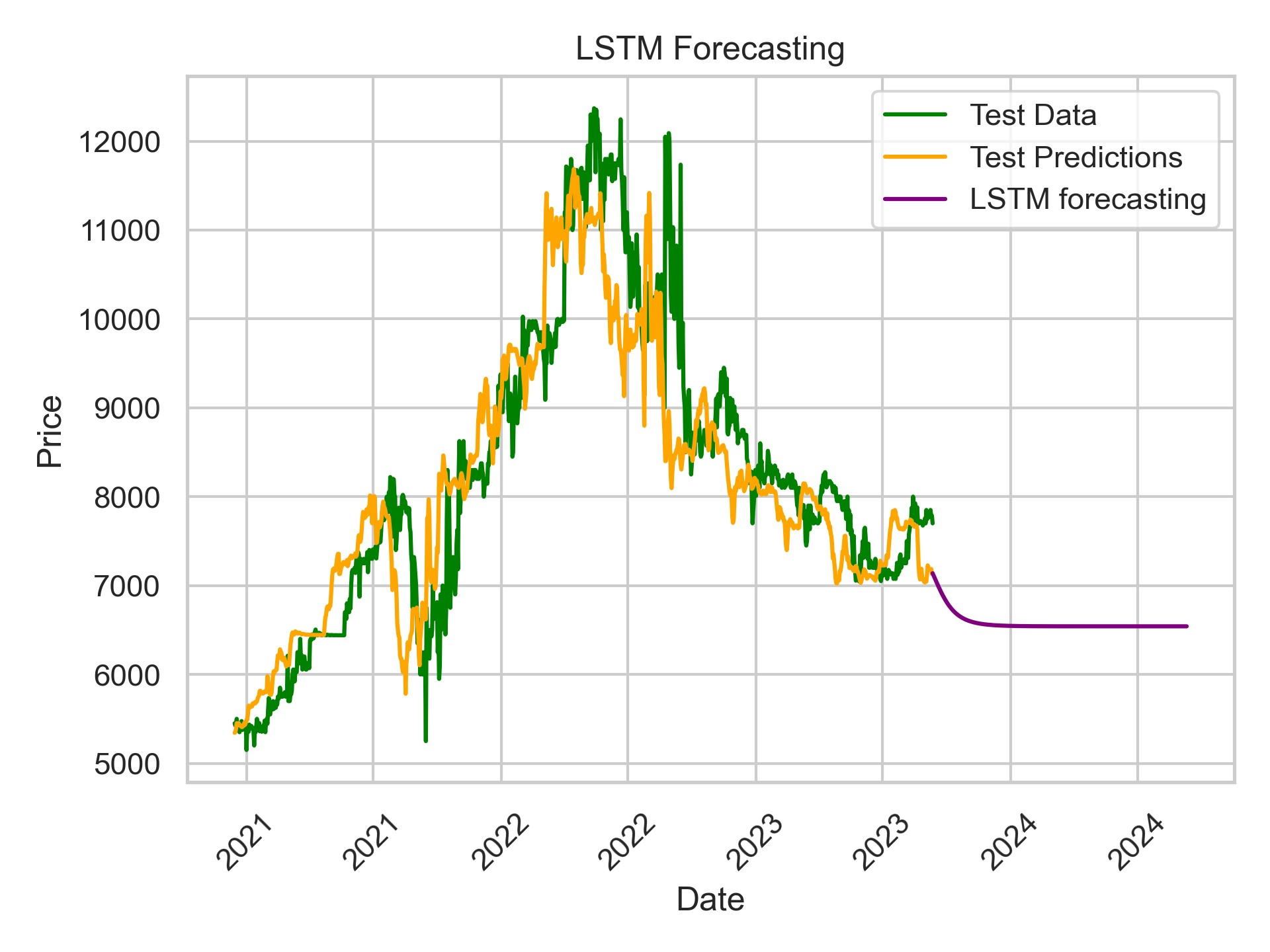
**Test prediction**



**LSTM MODEL**

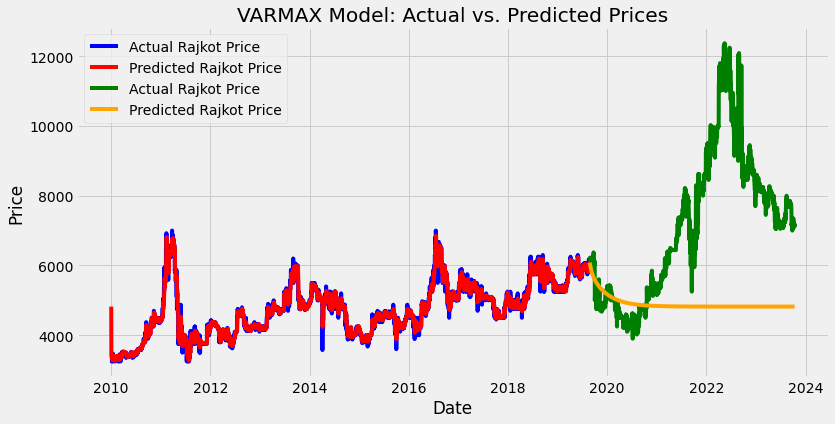


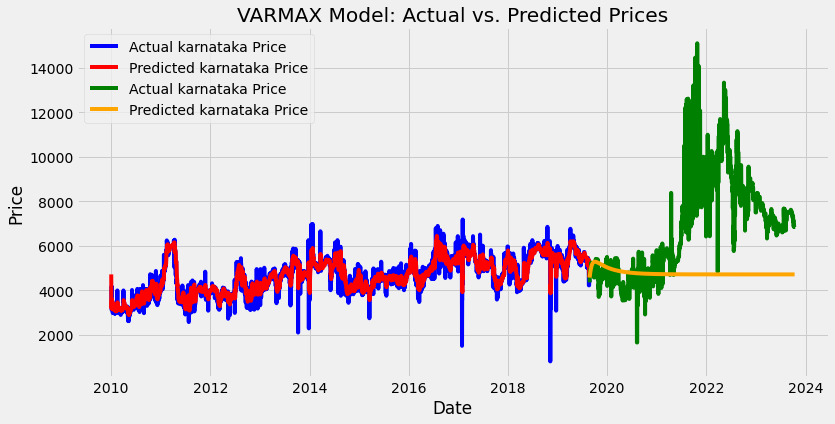
**Test prediction**

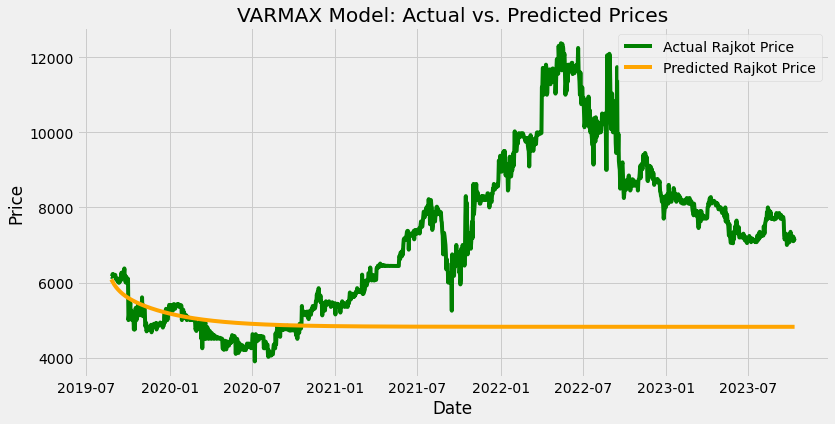


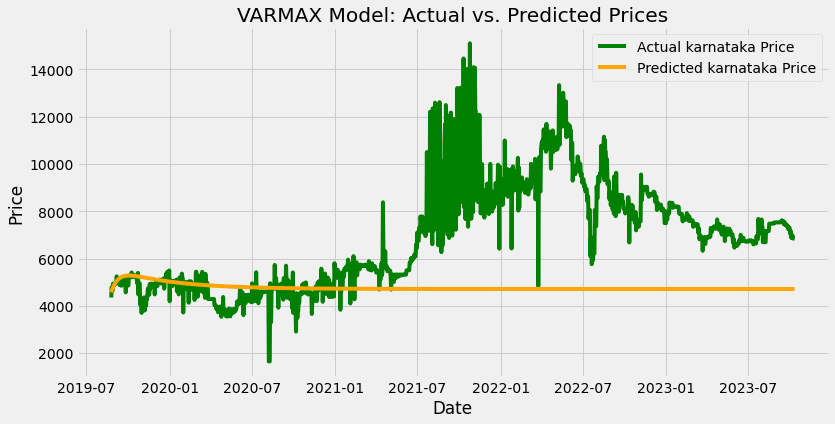
**VAR MODEL**

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









**COMPARISON BETWEEN DIFFERENT MODELS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MODEL** | **AIC** | **BIC** | **MSE** | **RMSE** |
| VAR | 70597.45 | 70673.84 | 7322895.4 | 2706.08 |
| ARIMA | 43984.29 | 44039.79 | 2772881.6 | 1665.19 |
| ARIMAX | 44024.07 | 44085.73 | 3986033.3 | 1996.50 |
| ARCH | 69569.6 | 69581.9 | 3326609.2 | 1823.89 |
| ARCHX | 72589.8 | 72602.3 | 3787475.3 | 1946.14 |
| LSTM | - | - | 582503.20 | 763.21 |

**CONCLUSION**

In the context of agricultural economics, cotton pricing is a vital factor affecting various stakeholders, from farmers and traders to policymakers and the textile industry. This comprehensive study delves into the intricacies of predicting cotton prices using a variety of time series and predictive models, encompassing ARIMA, ARIMAX, ARCH, ARCHX, VAR, and LSTM while focusing on two specific markets.

The results of this analysis reveal the unique strengths and characteristics of each model, providing valuable insights for decision-makers within the cotton industry. The LSTM model, a deep learning neural network type, stands out as a top performer, achieving exceptional predictive accuracy with an impressively low Root Mean Squared Error (RMSE). Its ability to capture complex patterns and nonlinear dependencies within the data makes it a compelling choice for those prioritizing predictive precision.

In contrast, the ARIMA model, a classical time series method, strikes a balance between model simplicity and performance. It boasts the lowest Akaike's Information Criterion (AIC) and Schwarz Bayesian Information Criterion (BIC) values, indicating a good trade-off between model fit and complexity. ARIMA is a solid choice when stakeholders require an accurate and interpretable model.

The ARCH (Autoregressive Conditional Heteroskedasticity) model shines in capturing the volatility patterns within the data. It displays a lower RMSE compared to traditional models, making it particularly useful for those interested in understanding and managing volatility in cotton prices. ARCH models are known for their effectiveness in capturing time-varying volatility, which is a crucial factor in financial and commodity markets.

The ARCHX model, a generalized version of ARCH, also excels in capturing volatility patterns. Its RMSE is competitive and serves as an extension of the ARCH model by considering both lagged squared residuals and lagged conditional variances, allowing it to capture more complex patterns in volatility.

The VAR (Vector Autoregression) model offers a different approach, considering the interdependencies among multiple time series variables. While the RMSE is higher than other models, VAR models are valuable when stakeholders need to understand and predict the joint behavior of interconnected variables.

In conclusion, the choice of which model to employ for cotton price forecasting depends on the specific needs of the analysis. Stakeholders looking for optimal predictive accuracy should consider LSTM, while those seeking a balance between accuracy and interpretability may opt for ARIMA. For understanding and managing volatility in cotton prices, both ARCH and ARCHX models offer robust options.

**REFERENCES:**

* [**http://masujournal.org/108/OfZ0RnBg6D4XqKm1B60rdV9980Sdvr.pdf**](http://masujournal.org/108/OfZ0RnBg6D4XqKm1B60rdV9980Sdvr.pdf)
* [**https://www.researchgate.net/publication/348379817\_Forecasting\_Commodity\_Prices\_Using\_Long\_ShortTerm\_Memory\_Neural\_Networks**](https://www.researchgate.net/publication/348379817_Forecasting_Commodity_Prices_Using_Long_ShortTerm_Memory_Neural_Networks)
* [**https://euroasiapub.org/wp-content/uploads/2016/10/1ESSMay-3546-1.pdf**](https://euroasiapub.org/wp-content/uploads/2016/10/1ESSMay-3546-1.pdf)
* [**https://www.researchgate.net/publication/273776515\_Modelling\_and\_Forecasting\_of\_Price\_Volatility\_An\_Application\_of\_GARCH\_and\_EGARCH\_Models**](https://www.researchgate.net/publication/273776515_Modelling_and_Forecasting_of_Price_Volatility_An_Application_of_GARCH_and_EGARCH_Models)
* [**https://cotcorp.org.in/**](https://cotcorp.org.in/)