International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2023; SP-8(5): 237-243
© 2023 Stats & Maths
https://www.mathsjournal.com
Received: 06-05-2023
Accepted: 11-06-2023

Riya Thakur
Department of Social Sciences, Dr. Y S Parmar University of Horticulture & Forestry, Nauni, Solan, Himachal Pradesh, India

Subhash Sharma
Department of Social Sciences, Dr. Y S Parmar University of Horticulture & Forestry, Nauni, Solan, Himachal Pradesh, India

Anmol Negi
Silviculture and Forest Management Division
Forest Research Institute, Dehradun, Uttarakhand, India

Corresponding Author:
Riya Thakur
Department of Social Sciences, Dr. Y S Parmar University of Horticulture & Forestry, Nauni, Solan, Himachal Pradesh, India

Apple price forecasting using different time series models in Himachal Pradesh

Riya Thakur, Subhash Sharma and Anmol Negi

Abstract

Apple is an important fruit crop of Himachal Pradesh, accounting for almost 49% of the total area under fruit crops and 85% of total fruit production. Fluctuations in the prices of agricultural crops affect supply and demand and have a significant impact on consumers. Accurate prediction of agricultural commodity prices would facilitate the reduction of risk caused by price fluctuations and is of great significance to the farmer's economies. Various time series models *viz*. ARIMA, ARCH-GARCH, and Recurrent neural network long short-term memory (RNN-LSTM) were used for efficient price prediction. The Solan market was selected purposively based on the highest arrival of apple produce in the state. To evaluate, these models daily price data was collected from AGMARKNET for the year 2012 to 2023. In all models, the best-fitted model was selected based on minimum information criteria. The results showed that using ARIMA (6, 1, and 1) and GARCH (1, 1) models were the best-fitted models. However, to confirm the validity of the models, the Root Mean Square Error value (RMSE) and Mean Absolute Percentage Error (MAPE) were compared which shows that the RNN (LSTM) model performed comparatively well over other models for forecasting apple prices. The prediction results based on the RNN model were better than those of the separate ARIMA and GARCH models. Furthermore, it best fits the actual price profile and has better generalizability.

Keywords: ANN, ARIMA, forecasting, machine learning

Introduction

India is an agriculture-dominated country where the horticulture sector has an important contribution to the overall growth of the economy. The various fruits grown in India are exported to different countries in the world. However, apple production is the most prominent one in India. In India, apples are grown as a commercial crop in the hilly areas of Jammu and Kashmir, Himachal Pradesh, Uttrakhand, and Arunachal Pradesh. Out of these Himachal is the second largest producer of apples and is known as the apple bowl of India. In Himachal Pradesh, apple production is one of the important sources of the state economy and it has a comparative advantage over the other crops grown in the region (Weinberger, Katinka, & Thoms A, 2007) [11]. The state has favorable agro-climatic zones and geographical conditions for apple cultivation (Singh, Kalia, & Lal, 2007) [10], (Panwar, 2011) [7]. To increase the farmer's income, the price prevailing in the market is important for the farmers to make decisions of their interest.

Therefore, price forecasting is an interesting area of research making researchers in the domain field always desire to improve existing predictive models. Price prediction is regarded as one of the most difficult tasks to accomplish in financial forecasting due to the complex nature of the market. This remains a motivating factor for researchers to evolve and develop new predictive models. Therefore, various time series model is there that helps in the prediction of future prices. The most common model used in the field of price forecasting was ARIMA, ARMA, ARCH-GARCH models, etc. ARIMA models are known to be more robust and efficient in financial time series forecasting especially short-term prediction than even the most popular ANN techniques. It has been extensively used in the field of economics and finance. Other statistics models are the regression method, exponential smoothing, and generalized autoregressive conditional heteroskedasticity (GARCH).

Methodology

The methodology used in this study to develop the ARIMA, ARCH-GARCH, and RNN (LSTM) model for Apple price forecasting is explained in detail in the subsections below. For data analysis purposes, the R-statistical package and Python software were used. The price data of Solan Market used in this research work is historical daily apple prices obtained from AGMARK net. The data consists generally of three elements, namely: max price, min price, and modal price respectively. In this research, the modal price is chosen to represent the price of the index to be predicted.

The first stage in time series analysis is to look at the stationary of price series data. A series is considered to be stationary if its statistical properties, such as mean and autocorrelation structures, remain constant over time. To determine the presence of a non-seasonal unit root in the price series, the Augmented Dickey-Fuller (Dickey and Fuller, 1979) [4] and Phillips-Perron (Phillips and Perron 1988) [12] tests were used. ARIMA model is one of the most popular approaches used in forecasting that follows the Box-Jenkins methodology (Box and Jenkins, 1976). In an Auto-Regressive Integrated Moving Average (ARIMA) model, the time series variable is assumed to be a linear function of previous actual values and random errors. The agricultural commodities price data are inherently noisy in nature and are volatile too therefore ARIMA model will not be enough to deal with such a series, as it is limited by assumptions of linearity and homoscedastic error variance.

ARCH and GARCH

The ARCH model aims to elaborate the variance clustering in the residuals as well as to indicate the squared errors in the nonlinear dependence of the first-moment model. Engle (1982) extracted the ARCH model from the ARIMA method to restrict the model for the conditional variance assumption for more accuracy in the prediction of the volatility. Researchers showed that the assumption of normality may not always be acceptable (Knief and Forstmeier, 2021). Therefore, non-normal distributions may be considered such as standard Cauchy distribution, Student-t distribution, and Generalized Error Distribution (Bollerslev, 1987; Braun *et al.*, 1995) [2, 3]. The Generalized Autoregressive Conditional Heteroskedasticity (GARCH) was extended by (Bollerslev, 1987) [2] to allow the conditional variance to follow the process of ARMA.

RNN (LSTM)

RNN is a type of artificial neural network that saves the output of a specific layer and feeds it back to the input of another to predict the result of the layer. It can memorize the previous inputs due to its internal memory. Long Short-Term Memory (LSTM) is a special kind of RNN that can learn long-time data dependencies (Hochreiter and Schmidhuber, 1997) [5]. The standard LSTM consists of three gates, and these gates are responsible for regulating the information and passing that information to the next unit. The forgotten value either forgets everything or does not forget the information based on the values of the forget gate (i.e., the forget gate forgets everything if the value is zero, and nothing if the value is one). The input gate controls the new information to add the next cell state, and it works in two parts. The first part of the input gate is the sigmoid layer, which controls the output value stored in the cell state. The input gate's second part is the Tanh layer, and it creates a vector of new feature values stored in the cell state. The output gates output the updated cell state information. Through the gates' structure, the statistics execute selectively and are handed through to update and hold the historical statistics and update the cell state. LSTM considers the previous historical values, analyses the present unknown patterns by adjusting itself according to the complete patterns, and makes future forecasts ahead.

An LSTM cell, h_{t-1}, is the previous memory output, and c_t is the current memory output. LSTM cell is explained as

It calculates the current memory (cg_t), the weight matrix (wt_{Cg}), and the bias is the (bs_{cg}).

$$cg_t = Tahn(wt_{Cg} \times [hd_{cg-1}, x_{Cg}] + bs_{cg}) \text{ ---------(1)}$$

The input gate manages the update of the current memory input data to the value of the memory cell, the weight matrix (wt_{ig}), and the bias (bs_{ig}) and the sigmoid function. The input gate is calculated as:

$$ig_t = \sigma(wt_{ig} \times [hd_{ig-1}, x_{ig}] + bs_{ig}) \text{ --------- (2)}$$

The forget gate controls the update of the previous memory data to the value of the memory cell, the weight matrix (w_{tf}), and the bias (bs_{fg}) and is the sigmoid function. The forget gate is calculated as:

$$fg_t = w_{tf}wt_{fg} \times [hd_{fgpedicle}] + bs_{fg}) \text{ --------- (3)}$$

lc_{t-1} is the last LSTM cell value, and the current memory cell can be calculated as:

$$cu_t = fit \times lc_{t-1} + cg_t \text{ --------- (4)}$$

Model Selection

When comparing among different specifications of ARMA-GARCH models, we select an appropriate model based on the Akaike Information Criteria (AIC) (Akaike, 1974) [1], the corrected Akaike Information Criteria (AICC), Schwarz's Bayesian Information Criterion (SBC) (Schwarz, 1978) and the Hannan-Quinn Information Criterion (HQC). The AIC, AICC, SBC, and HQC can be computed as:

$$AIC = -2\ln(L) + 2k \quad AICC = AIC + 2\, k(k+1)/\, N-k-1$$

$$SBC = -2\ln(L) + \ln(N)k \quad HQC = -2\ln(L) + 2\ln\$\ln(N)\%k$$

Where L is the value of the likelihood function evaluated at the parameter estimates, N is the
Number of observations, and k is the number of estimated parameters. The minimum value of AIC, AICC, SBC, and HQC was selected as the better model when comparing among models.

Model Evaluations

The performance of forecasting models is evaluated using three measures: Mean Square Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE), respectively. When comparing ARMA-GARCH models, the smallest value of MSE, RMSE, and MAPE are chosen as the most accurate forecast model.

Results and Discussion

For any time-series model establishment, stationary was a primary step. The price data collected from the Solan market on apple commodity used in this study covers the period from

5 Jan 2012 to 25 May 2023 having a total number of 2390 observations. Figure 1 depicts the original pattern of the series to give a general overview of whether the time series is stationary or not. From the graph below the time series exhibit stationary.

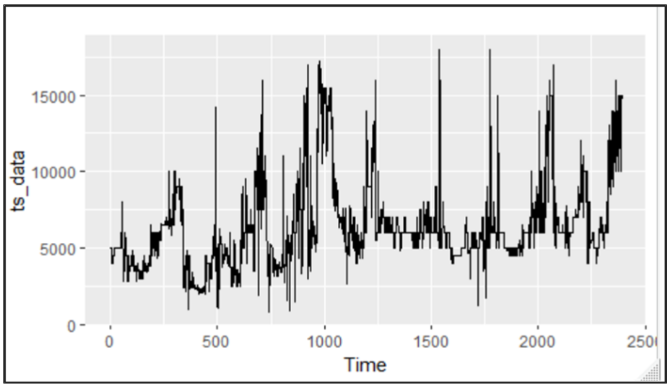

Fig 1: Time series plot of Solan market prices

Besides the time series plot, the ADF test and KPSS test were also used to confirm the stationary of the data. According to Table 1, the Augmented Dickey-Fuller unit root test and KPSS test, Apple's price for the Solan market shows that the time series data was stationary at p-value 0.01.

Table 1: Augmented dickey fuller unit root test and KPSS for apple in Solan

Market	Augmented dickey fuller	KPSS (unit root cointegration)
Solan	-5.01(0.01)	2.88

*Figure in the parenthesis is p-value

After stationary, the next step is to find the parameters of the AR and MA process by plotting the ACF and PACF graphs. Figure 2 shows the ACF dies down extremely slowly while the PACF has only a few significant spikes which shows that this time series data has an AR process of order p significant spike.

Figure 2 is the correlogram of the Nokia time series. From the Graph, the ACF dies down extremely slowly which simply means that the time series is non-stationary. If the series is not stationary, it is converted to a stationary series by differencing. After the first difference, the series "DCLOSE" of the Nokia stock index becomes stationary as shown in Figure 3 and Figure 4 of the line graph and correlogram respectively

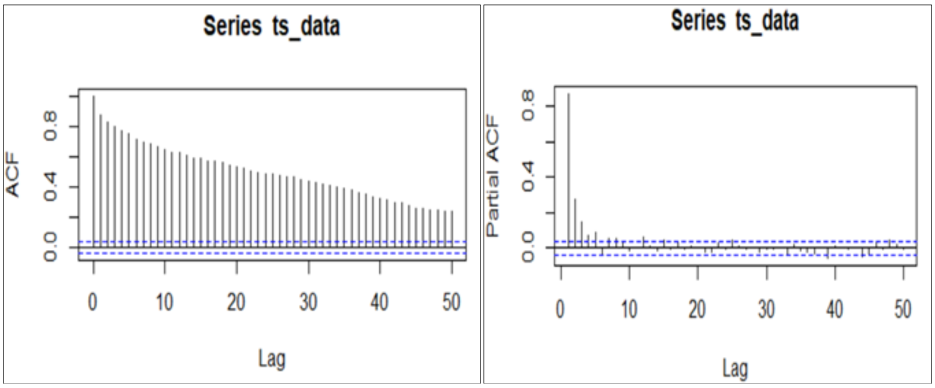

Fig 2: ACF and PACF plot of Apple price for Solan market

Minimum Akaike Information Criteria (AIC) was used to find the best-fitted model. Various parameters were presented in Table 2 of different ARIMA fit models. The results showed that the ARIMA model (6, 1, 1) was the best fitted with minimum AIC value with the lowest RMSE. Table 4 also shows the results of the Ljung-Box test which indicates the absence of autocorrelation for residuals of ARIMA models. The Ljung-Box statistics show that the ARIMA model (6, 1, and 1) has the lowest Q statistics value which is significant at a 5 percent level of significance indicating the absence of autocorrelation for residuals. The absence of auto-correlation in residuals is requisite for the best-fitted model. Figure 3 shows the autocorrelation of the residual of the ARIMA (6, 1, 1) model and parameter estimates along with the corresponding standard error for the selected (6, 1, 1) model presented in Table 3.

Table 2: Identification of the ARIMA (p, I, q) model for the Apple price series of the Solan market

Model	AIC	RMSE	MAPE	MASE	L Jung Box Q statistics	P value
ARIMA (8, 0, 0)	41082.82	1301.43	12.29	1.05	11.03	0.011
ARIMA (7, 1, 0)	41089.38	1309.15	11.94	1.03	14.72	0.002
ARIMA (6, 0, 0)	41095	1306.11	12.30	1.05	16.86	0.002
ARIMA (6, 1, 1)	41069.51	1303.69	12.039	1.04	-5.321	0.14**
ARIMA (5, 1, 0)	41112.58	1316.65	11.83	1.02	35.35	1.279e
ARIMA (2, 1, 3)	41075	1306.50	12.09	1.04	15.035	0.011

Table 3: Parameter estimation of ARIMA (6, 1, 1) by maximum likelihood estimation method for apple price series of Solan market

Parameter	AR1	AR2	AR3	AR4	AR5	AR6	MA1
Estimates	0.465	0.111	0.068	0.005	0.074	-0.080	-0.893
Standard error	0.040	0.027	0.024	0.023	0.023	0.022	0.036

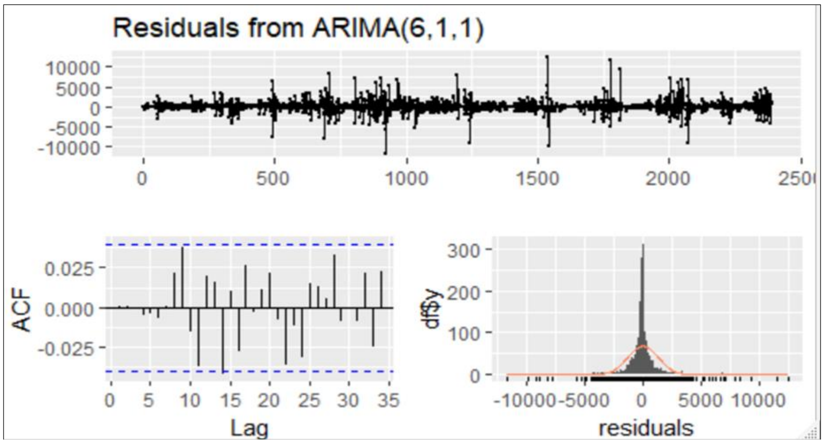

Fig 3: is the residual of the time series data

If the model is good, the residuals (difference between actual and predicted values) of the model are a series of random errors. Since there are no significant spikes of ACFs and PACFs, it means that the residual of the selected ARIMA model is white noise, with no other significant patterns left in the time series. Therefore, there is no need to consider any AR (p) and MA (q) further and this model can be used for prediction as also suggested by the Ljung test.

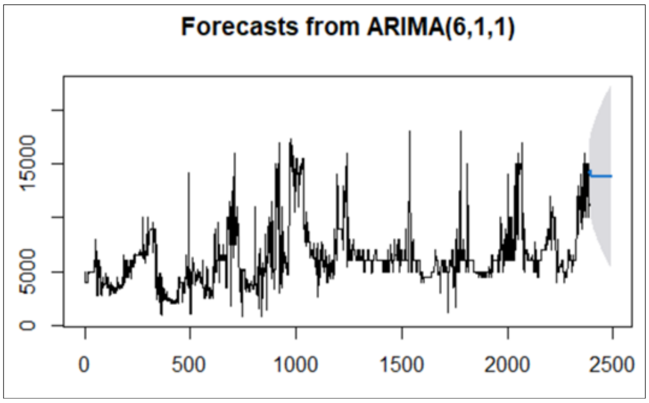

Fig 4: Time series plot of forecasted value of apple price for Solan market

For the ARCH and GARCH model

Although ACF & PACF of residuals have no significant lags the time series plot of residuals shows some cluster of volatility. It is important to remember that ARIMA is a method to linearly model the data and the forecast width remains constant because the model does not reflect recent changes or incorporate new information. In other words, it provides the best linear forecast for the series and thus plays little role in forecasting models nonlinearly. In order to model volatility, the ARCH/GARCH is used.

To apply the ARCH-GARCH model first step is to check if the time series data, its square value, and absolute value display any cluster of volatility. If volatility is present, the next step is to apply the ARCH effect to check that data is model for the ARCH and GARCH models. ARCH/GARCH should be used to model the volatility of the series to reflect more recent changes and fluctuations in the series.

Figure 5 shows that there is volatility clustering in the prices of the Solan market i.e., high changes are followed by high changes, and low changes are followed by low changes. This indicates the presence of heteroscedasticity in the data. Figure 6 shows that there is high autocorrelation present in the prices, the square value of prices, and the absolute value of prices. We can double-check the presence of auto-correlation in prices by applying the Ljung-Box test presented in Table 4. It shows that $p<0.05$ so the data is not independent. It means autocorrelation is present in the price data. When the ARCH test is applied it shows that the p-value is less than 0.05 which indicates that the null hypothesis (noarch effect) can be rejected. Therefore, the price data exhibit the ARCH effect.

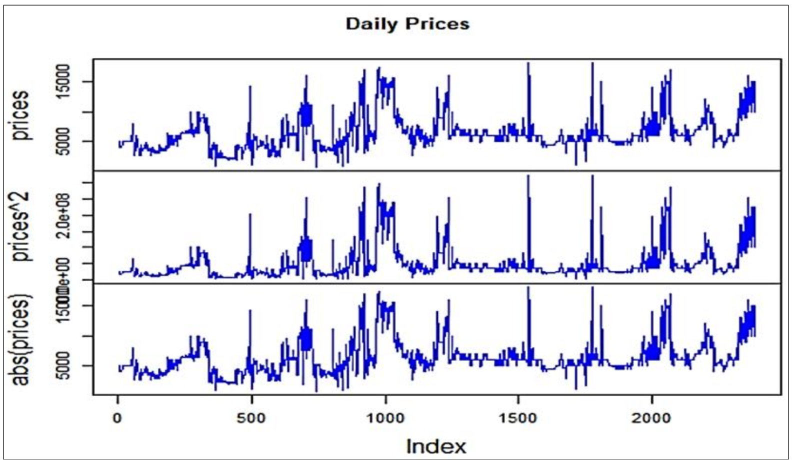

Fig 5: Volatility clustering of prices, square prices, and absolutes prices in the Solan market

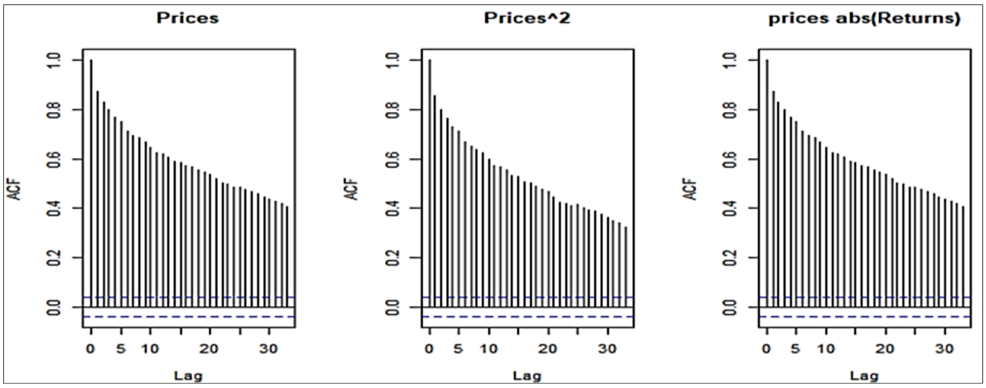

Fig 6: ACF plot of prices, square prices, and absolutes prices in Solan market

Table 4: Ljung-Box test and ARCH test to check the autocorrelation and ARCH effect

Test statistics	X squared value/Chi-square value	P value
Ljung-Box	13624	<2.2e-16
ARCH	1823.3	<2.2e-16

The next, step is to find the best-fitted ARCH and GARCH model based on the minimum value of Information criteria based on Akaike, Bayes, Shibata, and Hannan-Quinn. The results presented in Table 5 show that the GARCH (1, 1) distribution is the best-fitted model as it has the lowest Information criteria among all the other models. Therefore, the GARCH (1, 1) model is used for forecasting the next value. The parameters of this model are presented in Table 6 with values of estimates, standard error, and p values.

Table 5: Identification of the ARCH- GARCH model for the Apple price series of the Solan market

Information criteria	Akaike	Bayes	Shibata	Hannan-Quinn
ARCH (1,1)	17.960	17.969	17.960	17.963
ARCH (2,1)	17.962	17.973	17.961	17.965
ARCH (2,2)	17.960	17.969	17.960	17.963
GARCH (1,1)	16.762	16.777	16.762	16.767
GARCH (2,1)	16.767	16.784	16.767	16.773
GARCH (2,2)	16.765	16.784	16.765	16.772

Table 6: Estimation results of GARCH models

Parameters	Estimates	Standard Error	t-value	Pr(>|t|)
Mu	4807.363	4.9707e+02	9.6715	0.00*
Ar1	0.978	6.6320e-03	147.617	0.000**
Ma1	-0.484	2.5687e-02	-18.8554	0.000**
Omega	15669.316	1.7406e+03	9.002	0.000**
Alpha1	0.10008	6.460e-03	15.493	0.000**
Beta1	0.898	4.7200e-03	190.453	0.000**

*Significant at 5 percent, ** value is significant at 1 percent

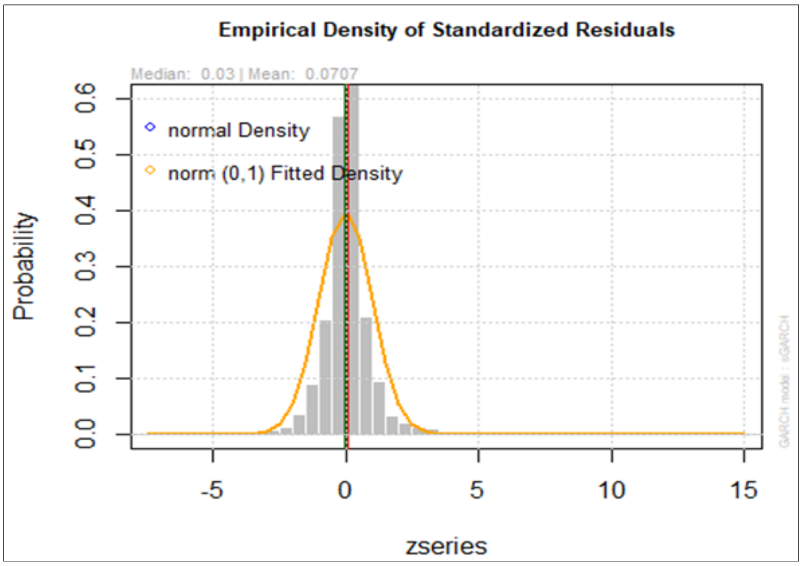

Fig 7: Residual plot of the best-fitted GARCH (1, 1) model

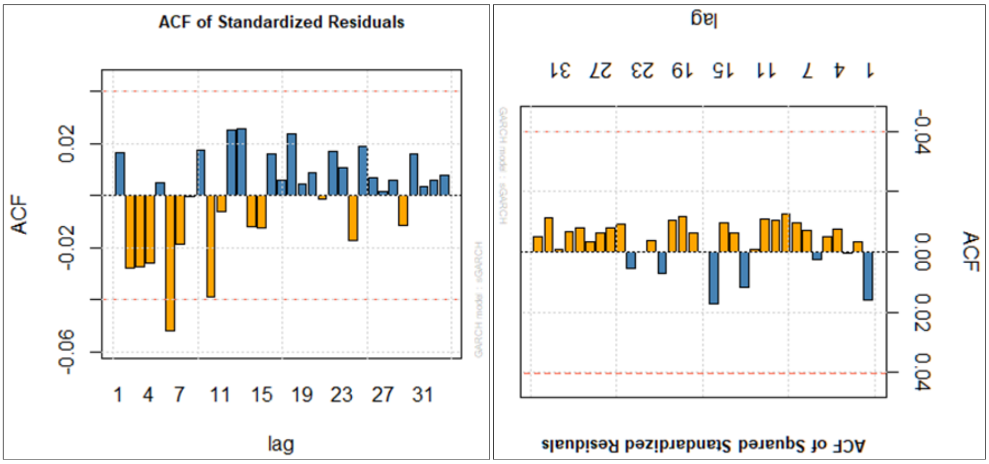

Fig 8: ACF plot of residuals and square residuals

The ACF & PACF plot of squared residuals confirms whether the residuals (noise term) are not independent and can be predicted. As mentioned earlier, strict white noise cannot be predicted either linearly or nonlinearly while general white noise might not be predicted linearly yet done so nonlinearly. If the residuals are strict white noise, they are independent with zero mean, normally distributed, and ACF & PACF of squared residuals display no significant lags. Therefore, the selected model can be used for forecasting.

RNN (LSTM)

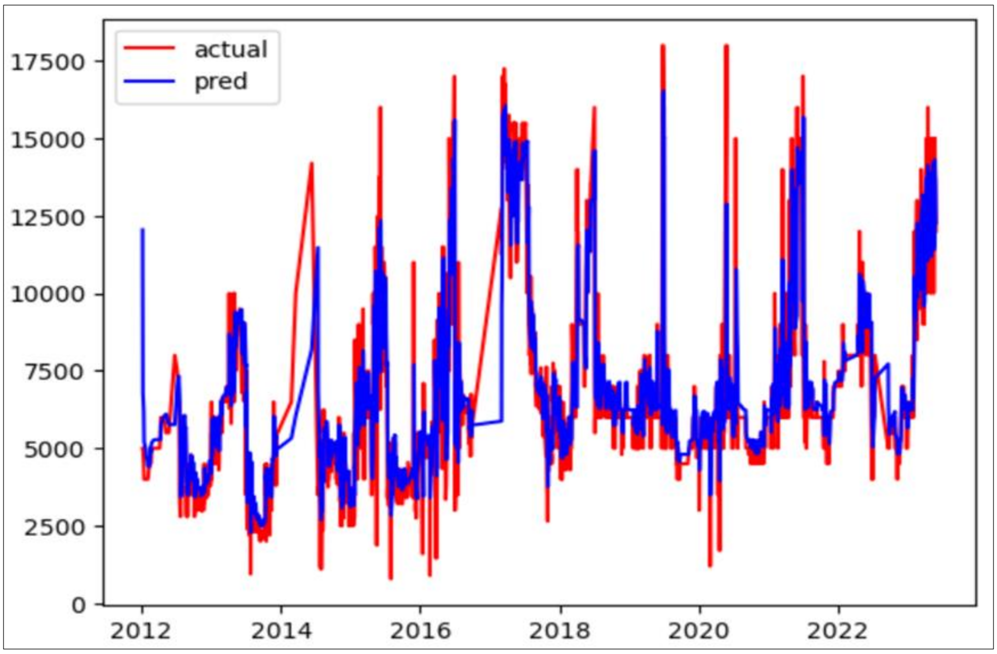

Fig 9: Time series plot of predicted and actual price data of Solan market

Table 5: Forecasting of daily Apple price of Solan market for 2023

Date	Actual prices	LSTM prices	ARCH-GARCH prices	ARIMA prices
26-05-2023	14000	14132	14144	14127
27-05-2023	14000	13640	14394	14312
29-05-2023	13000	13507	13756	14395
30-05-2023	12000	12886	13568	14119
31-05-2023	12000	12140	13384	14085
01-06-2023	12000	12886	13203	13939
02-06-2023	13000	13507	13027	13939
03-06-2023	11500	13640	12854	13902
05-06-2023	15000	14132	12685	13849
06-06-2023	15000	13735	12520	13839
07-06-2023	12000	13431	12358	13818

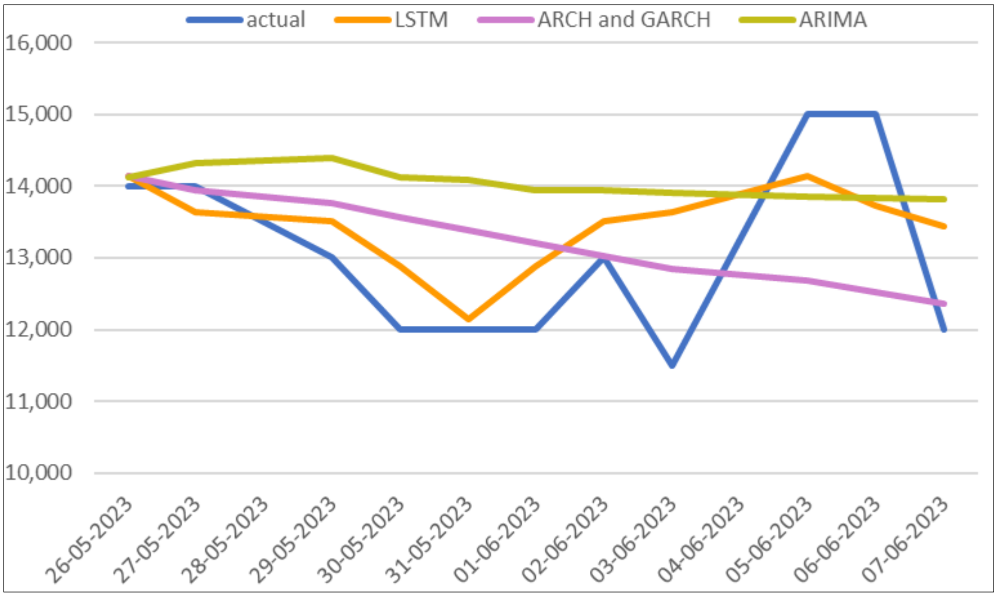

Fig 10: Time series plot of actual and predicted value using different time series model

Conclusion

The aim of this study was to forecast the price of apples for Solan markets using different time series models. The results showed that the ARMA (6, 1, 1), GARCH (1, 1), and RNN (LSTM) model was the best model for the time series data of a given period. These models were used to estimate and forecast the daily apple price for 12 days ahead in the future market effectively. The findings of this study suggest that the RNN (LSTM) model was the better model in comparison to ARIMA and GARCH models with the help of the smallest value of RMSE and the forecasted values from these different models. This finding also helps the state government to make policies with regard to relative price and also to establish relations with other neighbouring states of the country by making proper export plans based on price variation in the future.

References

1. Akaike H. A New Look at the Statistical Model Identification IEEE Trans. Autom. Control. 1974;19(6):716-723.
2. Bollerslev T. A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return, Review of Economics and Statistics, 1987;69:542-547.
3. Braun PA, Nelson DB, Sunier AM. Good News, Bad News, Volatility, and Betas, Journal of Finance. 1995;50:1575-1603.
4. Dickey DA, Fuller WA. Distribution of the Estimators for Autoregressive Time Series with a Unit Root. Journal of the American Stat. Association. 1979;74:427-431.
5. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computer. 1997;9:1735-1780.
6. Knief U, Forstmeier W. Violating the normality assumption may be the lesser of two evils. Behavior Research Methods. 2021;52(6):2576-2590.
7. Panwar T. Apple Production in Himachal Pradesh: An Impending Crises. Economic and Political Weekly, 2011;48(25):11-13.
8. Hossain ABMS, Al-Saif AM, Taha RM. Fruit growth, TSS and pH content development of water apple as affected by N-2-chloro-4-pyridyl-N- phenylurea (CPPU). Int. J Biol. Sci. 2021;3(2):06-11. DOI: 10.33545/26649926.2021.v3.i2a.29
9. Schwarz G. Estimating the Dimension of a Model. The Annals of Statistics. 1978;6(2):461-464.
10. Singh RR, Kalia V, Lal H. Impact of Climate change on Shift of Apple belt in Himachal Pradesh. ISPRS Archives XXXVIII-8/W3 Working Proceedings: Impact of Climate Change on Agriculture; c2007. p. 131-137.
11. Weinberger, Katinka, Thmas AL. Diversification into Horticulture and Poverty Reduction: A Research Agenda. World Development. 2007;35(8):1464-1480.
12. Phillips PCB, Perron P. Testing for a unit root in a time series regression. Biometrika. 1988;75(2):335-346.