International Journal of Statistics and Applied Mathematics 2024; SP-9(4): 123-125

ISSN: 2456-1452
Maths 2024; SP-9(4): 123-125

www.mathsjournal.com
Received: 06-07-2024
Accepted: 09-08-2024

Anita Sarkar
[1] The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi, Delhi, India
[2] ICAR-Indian Agricultural Statistics Research Institute, New Delhi, Delhi, India

Lalit Birla
[1] The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi, Delhi, India
[2] ICAR-Indian Agricultural Statistics Research Institute, New Delhi, Delhi, India

Ankit Kumar Singh
[1] The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi, Delhi, India
[2] ICAR-Indian Agricultural Statistics Research Institute, New Delhi, Delhi, India

Praveenkumar A
[1] The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi, Delhi, India
[2] ICAR-Indian Agricultural Statistics Research Institute, New Delhi, Delhi, India

Pushpendra Yadav
The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi, Delhi, India

Manoj Varma
[1] The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi, Delhi, India
[2] ICAR-Indian Agricultural Statistics Research Institute, New Delhi, Delhi, India

Corresponding Author:
Ankit Kumar Singh
[1] The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi, Delhi, India
[2] ICAR-Indian Agricultural Statistics Research Institute, New Delhi, Delhi, India

Comparison of machine learning models for oilseed price prediction

Anita Sarkar, Lalit Birla, Ankit Kumar Singh, Praveenkumar A, Pushpendra Yadav and Manoj Varma

DOI: https://dx.doi.org/10.22271/maths.2024.v9.i4Sb.1793

Abstract
Sunflowers are vital for agricultural economic growth, food security, and improving pollination for other crops. However, accurately forecasting sunflower prices is challenging due to factors such as fluctuating supply and weather conditions. This study evaluates the performance of various machine learning models *viz*., Artificial Neural Networks (ANN), Support Vector Regression (SVR), K-Nearest Neighbors (KNN), and Random Forest (RF) for predicting sunflower prices. The analysis uses monthly wholesale price data from January 2010 to June 2024 for the Bellary and Gadag markets in Karnataka, India, obtained from AGMARKNET. The findings reveal that the RF model outperforms the other models, demonstrating its superior effectiveness in predicting sunflower prices compared to the other machine learning approaches.

Keywords: Artificial neural network (ANN), k-nearest neighbours (KNN), Machine learning, Price prediction, Random forest (RF), Support vector machine (SVR)

Introduction
The sunflower (*Helianthus annuus*) is a highly cultivated crop with significant economic and agricultural value. Originally from North America, sunflowers are now grown across the globe, including in Europe, Asia, and South America. The plant is known for its large, vibrant yellow flowers, which are not only visually appealing but also serve as the main source of sunflower seeds-an important agricultural commodity. Sunflowers are versatile, primarily grown for their seeds, which are used to produce sunflower oil, a widely consumed edible oil. The seeds are also eaten as snacks or used in bird feed. Additionally, sunflowers play a role in renewable energy, as they can be used in the production of biodiesel. By-products of the plant, such as sunflower meal, are used in animal feed, while the stalks and leaves can be repurposed as livestock fodder or for mulching in agriculture. Sunflower seeds are highly nutritious, rich in essential fatty acids, especially linoleic acid, and are a great source of vitamin E, a powerful antioxidant. They also contain substantial amounts of protein, fibre, and essential minerals like magnesium, selenium, and zinc. The consumption of sunflower seeds and sunflower oil is linked to various health benefits, including improved heart health, enhanced skin quality, and reduced inflammation. Sunflower is a crucial oilseed crop worldwide, ranking fourth in global production after soybean, rapeseed, and palm oil. In the fiscal year 2022, India produced 190 thousand metric tons of sunflower crops (source: https://www.statista.com/statistics/report-content/statistic/1058517). The global market for sunflower oil is significant, fuelled by demand for cooking oil, especially in countries like India, China, and various European nations. Sunflower oil is highly valued for its mild flavour, high smoke point, and health benefits, such as being low in saturated fats and rich in vitamin E. These qualities make it a preferred option for both household cooking and industrial food processing. Sunflowers play a crucial role in advancing several Sustainable Development Goals (SDGs) by fostering economic growth, food security, and environmental sustainability. They provide essential nutrients and generate income for millions of smallholder farmers, helping to alleviate poverty (SDG 1) and reduce hunger (SDG 2). Sunflower cultivation also contributes to improved health (SDG 3) through better dietary quality and job creation, which in turn stimulates economic growth (SDG 8).

Sustainable farming practices in sunflower production support responsible consumption and production (SDG 12), while their role in biofuel production contributes to climate action (SDG 13). This study aims to forecast sunflower prices, enabling stakeholders to make informed decisions, such as government policy development and farmer planning, to stabilize markets and improve production efficiency. Accurate forecasts are essential for effective risk management and optimal resource allocation. Accurately predicting events and phenomena is essential in our daily lives, as it facilitates better decision-making under uncertain conditions. Modeling temporal price series helps in extracting valuable features from the data and allows for projecting future trends based on this information. Various stochastic processes are employed to model and forecast sunflower prices, enabling stakeholders to make informed decisions and manage risks effectively. Many researchers (Brandt and Bessler, 1983; Shahwan and Odening, 2007; Anggraeni *et al*., 2018; Kumar Mahto *et al*., 2019; Mohanty *et al*., 2023) [3, 8, 1, 5, 6] worked on prices prediction of different agricultural commodities. Paul *et al*. (2022) [7] evaluated the effectiveness of different ML algorithms, such as ANN, SVR, and RF, in forecasting wholesale brinjal prices, showing that ML techniques surpassed traditional models in performance. Jena *et al*. (2023) [4] concentrated on creating a low-complexity, adaptive ANN-based model for crop yield prediction. Few researchers (Wilson, 1985; Bal and Yayar, 2006; Turğut *et al*., 2023) [10, 2, 9] are worked with Sunflower price prediction. Although there is an increasing amount of research on using machine learning for agricultural price forecasting, several gaps still exist. Specifically, comprehensive studies dedicated to forecasting sunflower prices are scarce. Additionally, current research often overlooks the trade-offs between accuracy, interpretability, and computational efficiency across different machine learning models. This study aims to address these gaps by offering an in-depth comparative analysis of machine learning models for predicting sunflower prices. Accurately predicting sunflower prices is crucial for improving market efficiency and mitigating the economic risks posed by price volatility. While traditional forecasting methods have their uses, they often fall short in addressing the complexities of contemporary agricultural markets. Machine learning models present a promising alternative, as they can handle large datasets and detect intricate patterns within the data. However, given the numerous machine learning techniques available, it is unclear which model is most effective for predicting sunflower prices. This research aims to fill this gap by conducting a thorough comparative analysis of various machine learning models. This study has the objective to evaluate the performance of various machine learning models in predicting sunflower prices and to determine which model provides the best balance of accuracy, interpretability, and computational efficiency.

Methodology

k-Nearest Neighbour (k-NN)

The k-nearest neighbours (k-NN) algorithm, introduced by Fix in 1985, is a non-parametric method that leverages all available data points to predict numerical targets based on their similarity, typically measured using distance metrics. In kNN regression, a straightforward method involves predicting the target by averaging the numerical targets of the k closest neighbours. Alternatively, a weighted average can be applied, where weights are assigned based on the inverse distance of the k nearest neighbours. The kNN regression algorithm keeps a set of training instances, each characterized by a vector of n features $\left(F_1^i, F_2^i, \ldots, F_n^i\right)$ and a target vector with m attributes $\left(a_1^i, a_2^i, \ldots, a_m^i\right)$. When a new instance with a known feature vector but an unknown target is introduced, the algorithm identifies the k most similar training instances based on their feature vectors using a selected similarity or distance metric. Commonly used distance functions in k-NN regression include Euclidean, Manhattan, and Minkowski distances. k-Nearest Neighbour (k-NN).

Support Vector Regression (SVR)

In time series analysis, Support Vector Regression (SVR), which is derived from Support Vector Machines (SVM), is used to forecast future values by identifying patterns and relationships in historical data. SVR focuses on minimizing prediction errors while taking into account the temporal order of the data. As defined by Vapnik in 1999, the SVR model operates as follows:

$$f(y) = w \cdot \Phi(y) + \beta \quad (1)$$

Here, $\Phi(y)$ represents a mapping function that applies a non-linear transformation to convert non-linear input data into a linear format within a higher-dimensional feature space. In this context, $\boldsymbol{y}$ is the input vector, w is the weight vector associated with the model, and β is the bias term, both of which are determined by minimizing the regularized risk function, expressed as follows:

$$\mathcal{R}(x) = \frac{1}{2}\|w\|^2 + x\frac{1}{n}\sum_{i=1}^{N} l_\varepsilon(f(y_i), z_i) \quad (2)$$

Here, $l_\varepsilon(f(y_i), z_i)$ represents the empirical error, which is calculated using Vapnik's ε -insensitivity loss function (l_ε), and is defined as:

$$l_\varepsilon(f(y), z) = \begin{cases} |f(y) - z| - \varepsilon \; if \; |f(y) - z| \geq \varepsilon \\ 0 \; otherwise \end{cases} \quad (3)$$

where, z is the output vector.

Random Forest (RF)

Random Forest (RF) is an ensemble learning method built around decision trees, known for delivering strong performance across a wide range of applications. Decision trees are effective for both classification and regression tasks, and they are particularly well-suited for regression problems where the target variable is continuous. RF employs a technique called bootstrap aggregation, or bagging, where each decision tree is trained on a randomly selected subset of the entire training dataset. Let $R(\cdot)$ represent the function derived from training the RF model, which is used to predict $(\hat{y}_t)$ based on historical time series values and n lagged variables. The forecasted value $\hat{y}_t$ is then determined as follows:

$$\hat{y}_t = R(y_{t-1}, y_{t-2}, \ldots, y_{t-n}), t = n+1, \ldots,. \quad (4)$$

Artificial Neural Network (ANN)

The Artificial Neural Network (ANN) is a machine learning model composed of neurons arranged into three layers: the input layer, the hidden layer, and the output layer. The input layer receives the input data, the output layer generates the predicted outcomes, and the hidden layer captures complex relationships between the inputs and outputs. In the context of

time series, a nonlinear function *f* is applied to a range of time series values from y_{t-1} to y_{t-n}, represented as follows:

$$y_t = w_0 + \sum_{j=1}^{h} a_j f(a_{0j} + \sum_{i=1}^{n} a_{ij} y_{t-i}) + e_t \quad (5)$$

Here, a_{ij} and a_j are weights representing the relationships between nodes, h denotes the number of hidden nodes, n signifies the number of input nodes, and e_t represents the error term. ANN is highly regarded and widely used for forecasting the prices of agricultural commodities.

Result and Discussion

Monthly wholesale price data for sunflower, spanning from January 1, 2010, to June 31, 2024, has been gathered from the Bellary and Gadag markets in Karnataka, India, through the AGMARKNET portal (https://agmarknet.gov.in/). The machine learning models are evaluated empirically using three accuracy metrics: root mean square error (RMSE), mean absolute percentage error (MAPE), and mean absolute error (MAE), which are defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (5)$$

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \quad (6)$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (7)$$

where, y_i is the actual value, $\hat{y}_i$ is the predicted value, N denotes the total number of observations.

Table 1: Accuracy matrices of different machine learning models for predicting monthly wholesale prices of sunflower of different market.

Market	Bellary				Gadag			
Matrices	RF	SVR	KNN	ANN	RF	SVR	KNN	ANN
RMSE	665.44	723.42	1511.81	2031.90	341.07	616.79	1097.81	1967.68
MAPE	0.14	0.15	0.31	0.42	0.08	0.14	0.22	0.45
MAE	632.03	689.58	1408.66	1873.10	299.09	524.70	859.35	1740.43

Table 1 clearly shows that the RF model outperforms the other machine learning models, as it achieves the lowest values for all accuracy metrics in both markets. Following RF, SVR performs well, with k-NN and ANN ranking next. RF excels by effectively capturing complex patterns through its ensemble of decision trees, being less prone to overfitting, and requiring minimal pre-processing, which simplifies training and interpretation. Conversely, SVR is adept at modelling nonlinear relationships using kernel functions, offers strong generalization, and performs well on small to medium-sized datasets, without the high computational costs of ANN. While ANN can be very effective with large and complex datasets, RF and SVR provide a more balanced approach, particularly when interpretability, robustness, and efficiency are important. Given the relatively small size of the sunflower dataset, ANN does not deliver the expected results.

Conclusion

Accurate and timely forecasting of sunflower prices is crucial for farmers to select the best nearby markets for selling their produce at favourable prices. Furthermore, precise price forecasting aids stakeholders in making informed decisions, managing risks, and optimizing resource allocation, which contributes to market stability and enhanced production efficiency. This study provides a comparison of machine learning techniques for forecasting sunflower prices, especially in the dynamic and complex sunflower market. Among the models evaluated, the RF model shows superior performance across both markets, demonstrating its effectiveness in capturing price trends as indicated by RMSE, MAPE, and MAE metrics. The research highlights that RF methods offer greater accuracy compared to SVR, k-NN, and ANN when using sunflower price data from Karnataka. Future research could investigate advanced deep learning models to further improve sunflower price forecasting.

References

1. Anggraeni W, Mahananto F, Rofiq MA, Andri KB, Zaini Z, Subriadi AP. Agricultural strategic commodity price forecasting using artificial neural network. In: 2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI); c2018, p. 347-352. IEEE.
2. Bal HSG, Yayar R. Forecasting on sunflowers oil price in Turkey. J Appl Sci Res. 2006;2(9):572-578.
3. Brandt JA, Bessler DA. Price forecasting and evaluation: an application in agriculture. J Forecast. 1983;2(3):237-248.
4. Jena PR, Majhi B, Kalli R, Majhi R. Prediction of crop yield using climate variables in the south-western province of India: A functional artificial neural network modeling (FLANN) approach. Environ Dev Sustain. 2023;25(10):11033-1056.
5. KumarMahto A, Biswas R, Alam MA. Short term forecasting of agriculture commodity price by using ARIMA: Based on Indian market. In: Advances in Computing and Data Sciences: Third International Conference, ICACDS 2019; c2019, p. 452-61. Springer Singapore.
6. Mohanty MK, Thakurta PKG, Kar S. Agricultural commodity price prediction model: A machine learning framework. Neural Comput Appl. 2023;35(20):15109-15128.
7. Paul RK, Yeasin M, Kumar P, Kumar P, Balasubramanian M, Roy HS, *et al*. Machine learning techniques for forecasting agricultural prices: a case of brinjal in Odisha, India. PLOS One. 2022;17(7).
8. Shahwan T, Odening M. Forecasting agricultural commodity prices using hybrid neural networks. Comput Intell Econ Finance. 2007;2:63-74.
9. Turğut U, Güler D, Engindeniz S. The analysis of the relation between production and price of sunflower by Koyck model. Tarım Ekonomisi Dergisi. 2023;29(1):57-64.
10. Wilson WW. Price discovery and risk management in the sunflower market. In: NCR-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management, Chicago, Illinois; c1985 May, p. 2-3.