International Journal of Advances in
Soft Computing and Intelligent Systems (IJASCIS)
2023, Vol 02, Issue 02, 106-118

Original Paper

AGRICULTURAL COMMODITY PRICE FORECASTING WITH COMPETITIVE ENSEMBLE REGRESSION TECHNIQUE

R. Ragunath[a], R. Rathipriya[b,*]

[a] Ph.D. Research Scholar, Department of Computer Science, Periyar University, Salem-636011, rrcspu@gmail.com

[b] Assistant Professor, Department of Computer Science, Periyar University, Salem-636011, rathi_priyar@periyaruniversity.ac.in

ABSTRACT

This research study focuses on introducing a novel ensemble learning-based strategy using regression models to enhance the accuracy of forecasting Agricultural Commodity Price (ACP) trends. The main objective is to give farmers and traders better accurate pricing forecasts. The study uses data from India's rainfall data and the Wholesale Pricing Index (WPI) for essential commodities to test a variety of regression models, including ensemble regression models. The empirical results highlight the competitive ensemble approach's greater accuracy in capturing directional shifts in agricultural commodity pricing when compared to conventional regression models. As a result, this strategy has a lot of potential for assisting decision-making in the food and financial industries.

KEYWORDS

ACP, Price Forecasting, Ensemble learning-based approach, Ensemble Regression models, Competitive Ensemble Approach

1. INTRODUCTION

Farmers today often face significant losses due to their lack of understanding of market information, resulting in them selling their products at a loss. The prices of agricultural products are affected by various factors, such as weather, pests, production risks, global demand and supply, governmental policies, and economic considerations, leading to high price volatility [1-3]. As a result, accurate and dependable price forecasting methods are essential for managing pricing risks and making informed choices about selling products. While traditional ACP prediction methods that use various linear and non-linear forecasting models have been in use, machine learning has emerged as a superior approach [4].

Machine learning (ML) approaches have dominated the data science paradigm in recent years, as numerous empirical studies have shown their superiority over time series models for financial asset prediction [5]. Among the commonly used ML approaches are artificial neural networks (ANN), generalised neural networks (GRNN), support vector regression (SVR), random forests (RF), gradient boosting machines (GBM), and others, which are all data-driven, nonparametric methods for identifying stochastic relationships in data [6]. Traditional statistical methods like linear regression and Box-Jenkins processes are considered to perform worse than ANNs [7]. According to [8, 9], ML and deep learning algorithms are the best approaches for prediction problems, as neural networks have been found to be superior to statistical methods for forecasting agricultural prices [10]. Regression models like MLR [11], RF, Lasso, K-nearest neighbour (KNN), GPR, gradient boosting decision tree (GBDT), and SVR have also

successfully improved the prediction accuracy of Agri-products. However, each model has its own assumptions and limitations [12-13].

The shortcoming of the aforementioned prediction techniques is their reliance on a single model to forecast agricultural commodity data. To overcome this limitation, Ensemble learning algorithms have been used more and more in a variety of disciplines recently, particularly in agricultural research [14-17], as a result of advancements in computational technology and machine learning concepts. Ensemble learning primarily consists of combining multiple learners to create a stronger, more complete, and more comprehensive model. The core tenet of ensemble learning is that various base learners can still correct an error even if one base learner delivers a less accurate prediction. Some of the often-used ensemble learning techniques are the bosting, bagging, and stacking algorithms. Therefore, employing competitive ensemble learning, a novel methodology for ACP prediction is proposed in the current research.

This research work is formatted as follows" Section 2 expounds the machine learning methods and evaluation measures. Section 3 describes the proposed methodology for ACP prediction. Section 4 represents the result and discussions of the proposed work. Section 5 concludes the finding of the proposed research work.

2. METHODS

2.1 Predicting Techniques

Support Vector Regression (SVR), Decision Tree Regression (DTR), Gaussian Process Regression (GPR), Random Forest Regression, and Extreme Gradient Boost Regression (XGBoost Regression) are some of the ML techniques employed in this study. A summary of the various regression models is shown in Table 1. It's vital to keep in mind that there are some additional machine learning techniques available for this task. These machine learning approaches were chosen since they are frequently used to predict permeability in the literature.

Table 1: Regression Models

Algorithm	Description and Application
Support Vector Regression (SVR)	Although it operates on the same fundamental principles as SVMs, it optimises the cost function to fit the data points with the straightest line (or plane). By implicitly transforming their inputs into high-dimensional feature spaces, it is possible to do a non-linear regression with the kernel trick in an effective manner.
Gaussian Process Regression (GPR)	GPR uses a Bayesian approach that infers a probability distribution over the possible functions that fit the data. The Gaussian process is a prior that is specified as a multivariate Gaussian distribution.
Decision Tree Regression (DTR)	Decision Tress models learn on the data by making decision rules on the variables to separate the classes in a flowchart like a tree data structure. They can be used for both regression and classification.
Random Forest Regression (RFR)	Random Forest classification models learn using an ensemble of decision trees. The results of the random forest are determined by the decision trees' majority votes.
Extreme Gradient Boosting Regression (XGBoost Regression)	Extreme Gradient Boosting Regression, or XGBoost Regression, is an advanced machine learning algorithm used for predicting numerical values. It combines multiple weak models, like decision trees, to create a more accurate and powerful model. XGBoost Regression incorporates optimization techniques to improve performance and handles large

	datasets effectively.

2.2 Performance Measures

The absolute difference between the dataset's actual and forecasted values is averaged out to generate the Mean Absolute Error. It calculates the dataset's residuals' average in equation 1.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|actualY - predictedY| \quad \textbf{(1)}$$

The average of the squared difference between the data set's original and forecasted values is known as mean squared error. It calculates the residuals' variance in equation 2.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(actualY - predictedY)^2 \quad \textbf{(2)}$$

The square root of Mean Squared Error is called Root Mean Squared Error. It calculates the residuals' standard deviation in equation 3.

$$RMSE = \sqrt{MSE} = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(actualY - predictedY)^2} \quad \textbf{(3)}$$

R-squared, also known as the coefficient of determination, measures how much of the variance in the dependent variable is explained by the linear regression model is shown in equation 4. Since the score is scale-free, it will always be less than one regardless of how big or tiny the numbers are.

$$R^2 = 1 - \frac{\sum(actualY - predictedY)^2}{\sum(actualY - predictedY)^2} \quad (4)$$

MAPE is the total of all individual absolute deviations divided by the demand (each period separately). It is the average of the percentage errors is shown in equation 5.

$$\textbf{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{abs(actualY - PredictedY)}{actualY} \quad \textbf{(5)}$$

The ratio between the total absolute difference between the dataset's actual values and the predicted values, multiplied by 100, is known as the percentage of error (PE) is given in equation 6.

$$\textbf{Percentage of Error (PE)} = \frac{\sum_{i=1}^{n}|actualY - predictedY|}{\sum_{i=1}^{n} actualY} * 100 \quad \textbf{(6)}$$

3. Proposed Work

The proposed work is divided into many sections, including data collection, pre-processing, wholesale price prediction of crops using regression algorithms, and competitive ensemble learning utilising roulette wheel selection operator.

The main contribution of this paper is summarized as follows.

The proposed work competitive ensemble learning-based forecasting strategy is used to predict the price forecasting trend for crops in India, which consists of four basic forecasting models, and algorithms to achieve more accurate and robust performance under various change types.

3.1 Stage 1: Data Collection and Data Pre-Processing

Data were first gathered from datasets made available by the authorised website data.gov.in and pre-processed into forms that were appropriate for fitting in regression modelling. Many real-world datasets suffer from the missing value problem. Machine learning models' outcomes may be impacted by missing values, which may also cause the model's accuracy to decline. Managing missing values properly is essential. Many machine learning methods are ineffective when the dataset contains missing values. Therefore, in this study, the missing values are substituted with the proper imputation techniques.

3.2 Stage 2: Proposed Prediction Algorithms

The training and testing phases are the two key steps of a prediction algorithm as shown in Algorithm 1. Divide the given dataset into a training dataset and a testing dataset, each comprising 70% and 30% of the total. Using a training dataset and the chosen regressor modelling methodology, the training phase creates the prediction model first. The top 5 regressors, such as the eXtreme Gradient Boosting Regressor (XGB), Random Forest (RF), Support Vector regressor (SVM), Decision Tree Regressor (DT), and Gaussian Process Regressor (GPR) are employed. The performance validation of the training stage, which also guarantees the general performance of the regressor model, is used to avoid the overfitting issue. Using the testing dataset as input to the trained regressor, the trained regression model is then verified in the testing stage. Due to the fact that it is the other partitioned data from the initial dataset, this testing dataset shares the same characteristics as the training dataset. Performance metrics are used to evaluate both levels. By assessing how well training and testing performed, any overfitting issues can be found Iterative cross-validation is carried out k times. 10-time cross-validation was applied in this investigation. Figure 1 depicts the workflow of the proposed methodology.

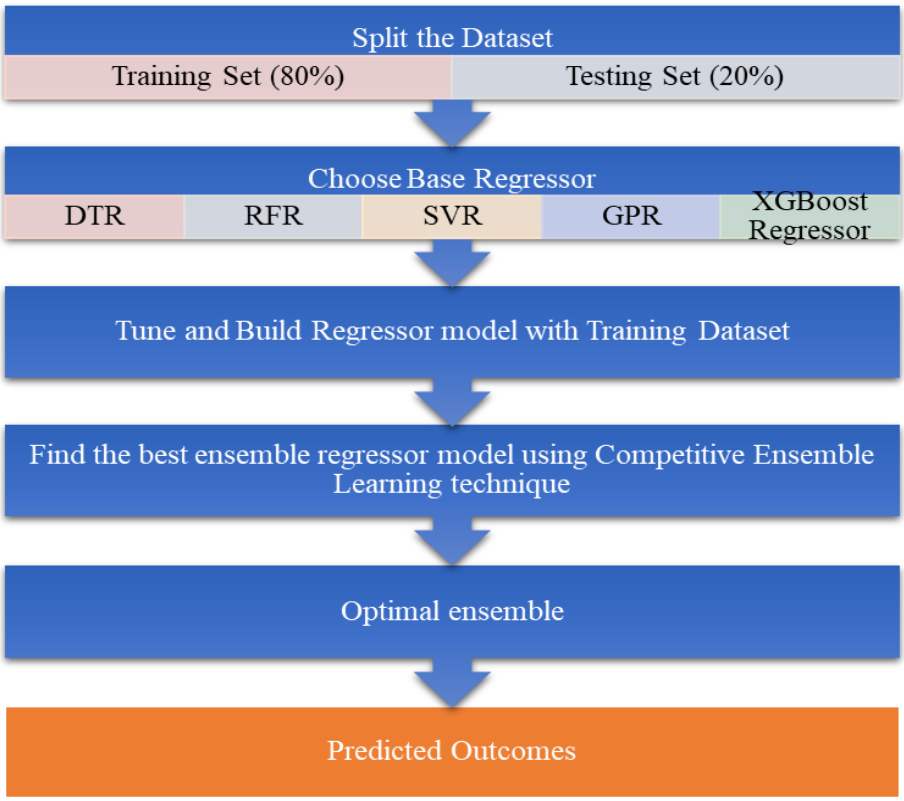

Figure 1: Overall Workflow of the Proposed Methodology

3.3 Stage 3: Competitive Ensemble Approach

The goal of a competitive ensemble approach is to choose the best regressor model out of a group of 'n' regressor models. To do this, the Roulette Wheel Selection (RWS) operator has to be adapted in this work to develop a competitive ensemble model. Therefore, the objective is to select a regressor model based on its correctness for forecasting.

Algorithm1: Agriculture Price Prediction using Competitive Ensemble Model

```
Input: Dataset (Rainfall, Crop_WPI)
Output: Forecasted Crop_WPI
Begin

// Initialize variables to track the best model and its fitness
        BestModel = null
        BestModelFitness = -infinity
        For each regressor model 'R' in modelList do
            Train the model 'R' using the training dataset    // Model Training
           // Training Data Evaluation
           WPI_Pred_tr = Predict Crop_WPI for the training data using model 'R'
           RMSE_tr = Calculate RMSE(TrainWPI, WPI_Pred_tr)
           R2_tr = Calculate R-squared(TrainWPI, WPI_Pred_tr)
           PercentError_tr = Calculate PercentError(TrainWPI, WPI_Pred_tr)
           Fitness_tr = CalculateFitness(RMSE_tr)
           // Testing Data Evaluation
           WPI_Pred_ts = Predict Crop_WPI for the testing data using model 'R'
           RMSE_ts = Calculate RMSE(TestWPI, WPI_Pred_ts)
           R2_ts = Calculate R-squared(TestWPI, WPI_Pred_ts)
           PercentError_ts = Calculate PercentError(TestWPI, WPI_Pred_ts)
           Fitness_ts = CalculateFitness(RMSE_ts)
           StoreModelFitness(modelList, Fitness_tr, Fitness_ts)
           if Fitness_ts > BestModelFitness then
              BestModel = 'R'
              BestModelFitness = Fitness_ts
           end if
        End For
        Return BestModel
      End
      Function CalculateFitness(RMSE)
        Fitness = (f(i) - min(f)) / (max(f) - min(f))
       End Function
```

Consider a set of 'n' regressor models with indexes ranging from 1 to n, each of which has a fitness value indicated as f(i), where 'i' stands for the model index. The selection probability for this competitive ensemble strategy can be rewritten as follows:

$$p(i) = (f(i) - \min(f)) / (\max(f) - \min(f))$$

$$\text{Where } f(i) = \text{sqrt}((1/n) * \Sigma(\text{actualy} - \text{predictedy})^2) \quad (7)$$

In this equation, f(i) represents the fitness value of regressor model 'i.' using, min(f) signifies the minimum fitness value among all regressor models, and max(f) indicates the maximum fitness value among all regressor models. The fitness values of the regressor models are scaled

to the range [0, 1] using this mathematical equation. It guarantees that models with higher fitness have a greater chance of being chosen while allowing all models a chance to be chosen.

4. OBSERVATIONAL STUDIES

The performance results of base regressors, such as SVR, DTR, RFR, GPR, and XGBoost Regressor, are shown in Tables 2,3,4,5 and 6 for the crops of paddy, maize, ragi, barley, and wheat. RMSE value, R2 Value, and % error for the training and test datasets are among the performance outcomes.

Table 2. Performance of Repressor Models for Paddy Dataset

Model	RMSE	MAE	R^2	MAPE	PE
XGB Regressor	0.931	1.66	0.985	0.01	1.205
Random Forest Regressor	1.15	1.38	0.985	0.009	0.0055
Decision Tree Regressor	0.75	1.23	0.99	0.008	0.0
Gaussian Process Regressor	68.015	130.76	-42.905	0.88	4.705
SVR	15.73	13.14	-0.01	0.085	0.085

Table 3. Performance of Base Regressor models for maize dataset

Model	RMSE	MAE	R^2	MAPE	PE
XGB Regressor	5.6985	6.24	0.855	0.04	1.86
Random Forest Regressor	7.09	6.39	0.835	0.04	0.0255
Decision Tree Regressor	7.475	8.44	0.765	0.06	0.03
Gaussian Process Regressor	63.69	120.98	-15.18	0.885	4.76
SVR	22.33	17.16	-0.024	0.12	0.12

Table 4. Performance of Base Regressor models for barely dataset

Model	RMSE	MAE	R^2	MAPE	PE
XGB Regressor	3.0	5.002	0.975	0.03	4.555
Random Forest Regressor	5.19	4.73	0.965	0.03	0.02
Decision Tree Regressor	4.41	5.69	0.955	0.03	0.015
Gaussian Process Regressor	69.48	131.28	-8.475	0.89	4.735
SVR	30.815	22.33	-0.012	0.145	0.145

Table 5. Performance of Base Regressor models for wheat dataset

Model	RMSE	MAE	R^2	MAPE	PE
XGB Regressor	3.19	4.36	0.92	0.025	1.41
Random Forest Regressor	3.68	3.65	0.935	0.02	0.014
Decision Tree Regressor	3.89	4.63	0.895	0.015	0.015
Gaussian Process Regressor	64.81	124.39	-27.81	0.89	4.725
SVR	18.1	14.88	-0.008	0.10	0.10

Table 6. Performance of Base Regressor models for ragi dataset

Model	RMSE	MAE	R^2	MAPE	PE
XGB Regressor	2.555	4.43	0.98	0.02	3.27
Random Forest Regressor	3.815	3.72	0.98	0.02	0.01
Decision Tree Regressor	2.48	4.26	0.98	0.02	0.01
Gaussian Process Regressor	88.075	164.20	-16.2	0.855	4.645
SVR	37.96	30.76	-0.345	0.155	0.155

The performance metrics of several regressor models for the Paddy dataset are shown in Table 2. Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R-squared (R2) coefficient, Mean Absolute Percentage Error (MAPE), and Prediction Error (PE) are used to evaluate the models. With the lowest RMSE, MAE, and outstanding R2 coefficient close to 0.99, the Decision Tree Regressor stands out among these models and demonstrates its greater predictive power for the Paddy dataset.

For the Maize dataset, the performance of several base regressor models is shown in Table 3. For the Maize dataset, the XGB Regressor shows the lowest RMSE and MAE, indicating greater prediction accuracy. The effectiveness of base regressor models for the Barley dataset is shown in Table 4. With a low RMSE, MAE, and a high R2 value of 0.975, the XGB Regressor distinguishes itself as the top performer and demonstrates its potency in forecasting the Barley dataset.

The performance metrics for different base regressor models used on the Wheat dataset can be found in Table 5. In the Wheat dataset, the Random Forest Regressor achieves the lowest RMSE and MAE, indicating its potential for precise predictions. The performance assessment of base regressor models using the Ragi dataset is shown in Table 6. Low RMSE values are shown by the XGB Regressor and Decision Tree Regressor, indicating high prediction skills for the Ragi dataset.

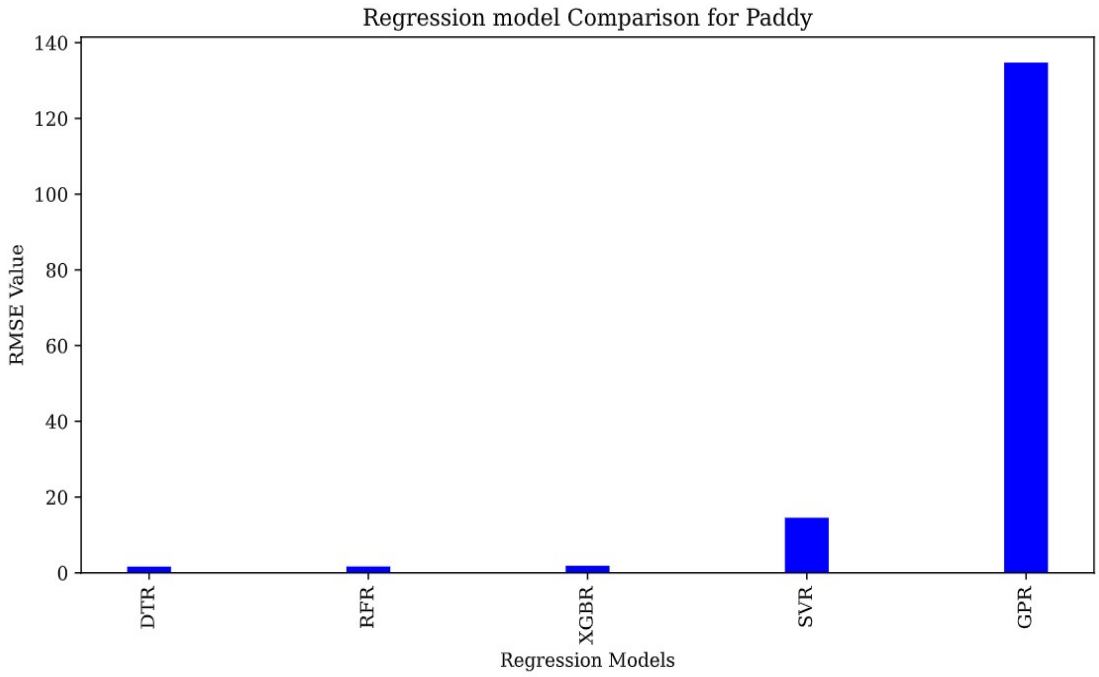

Figure (a). Paddy wholesale price data

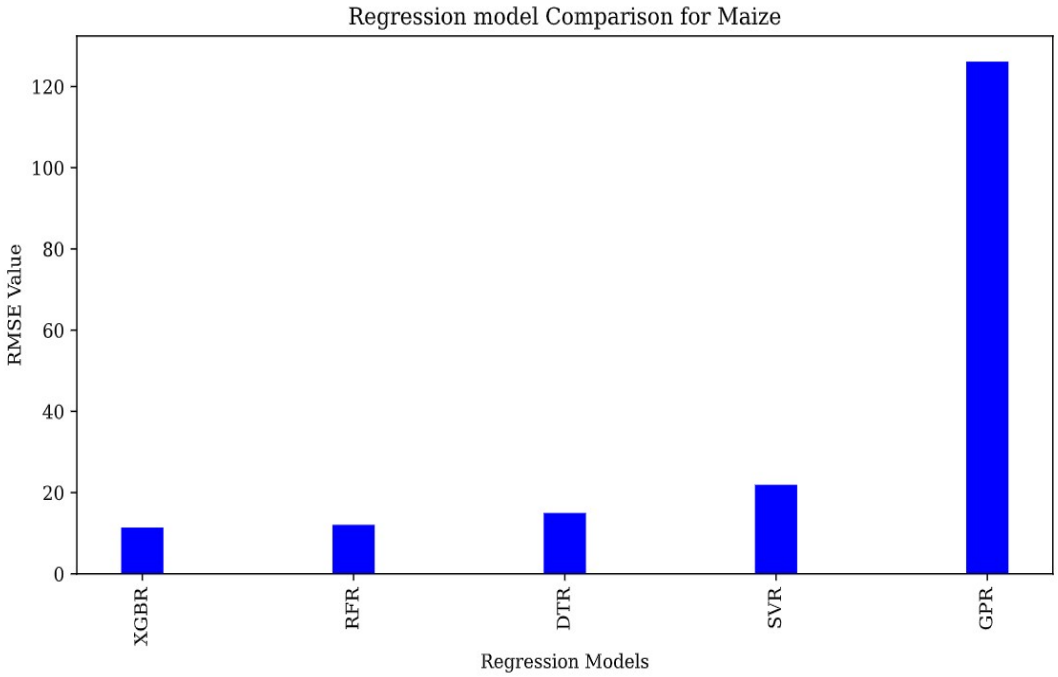

Figure (b). Maize wholesale price data

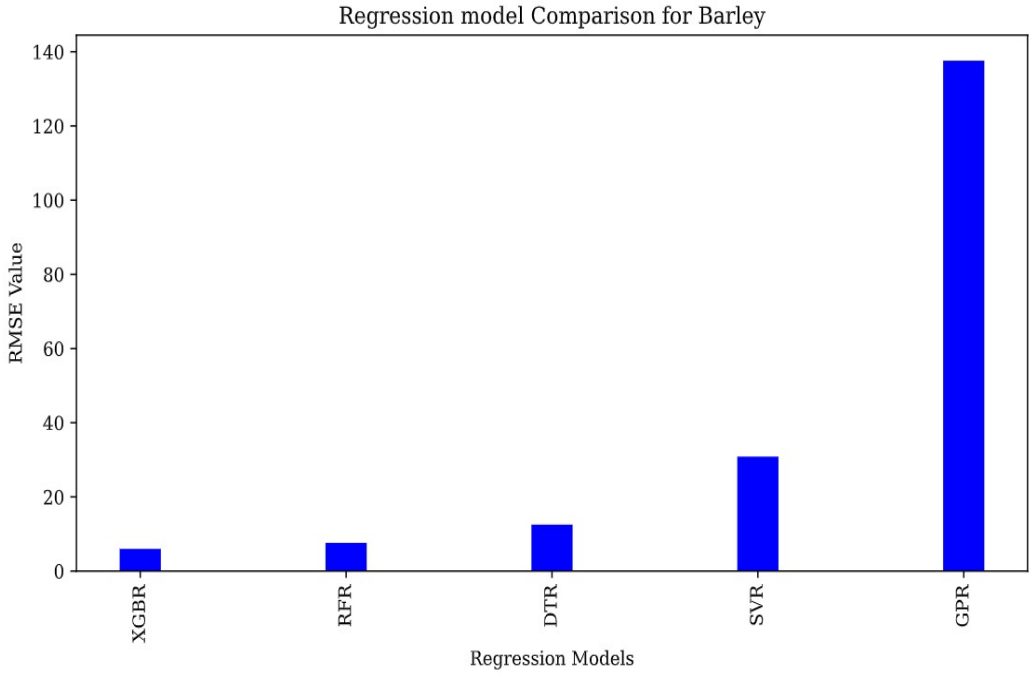

Figure (c). Barley wholesale price data

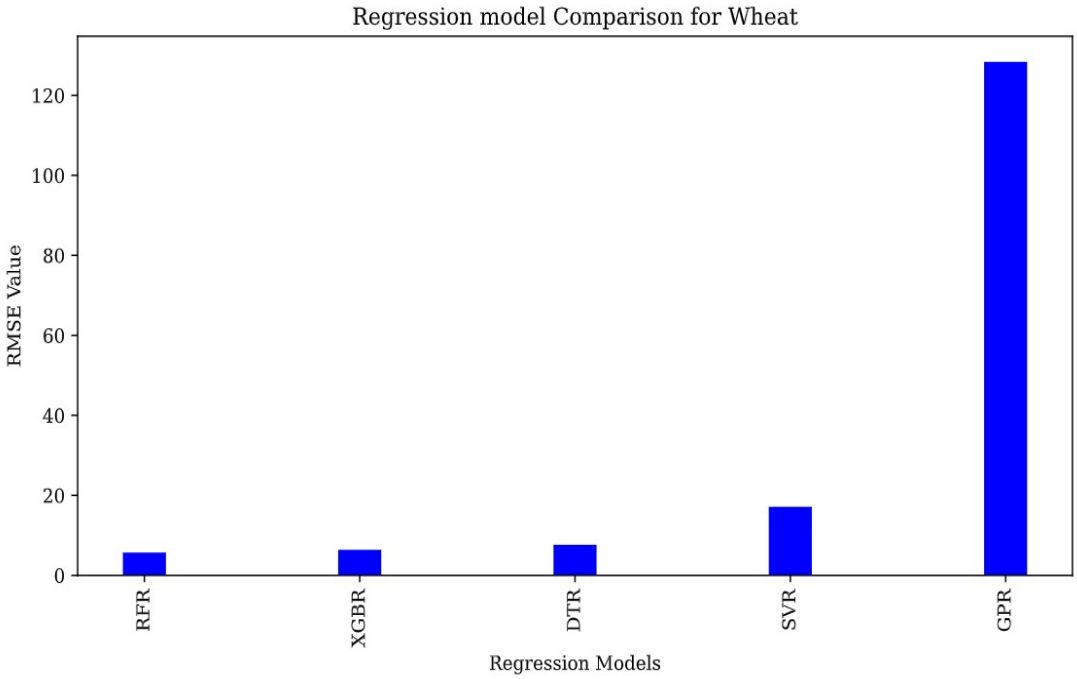

Figure (d). Wheat wholesale price data

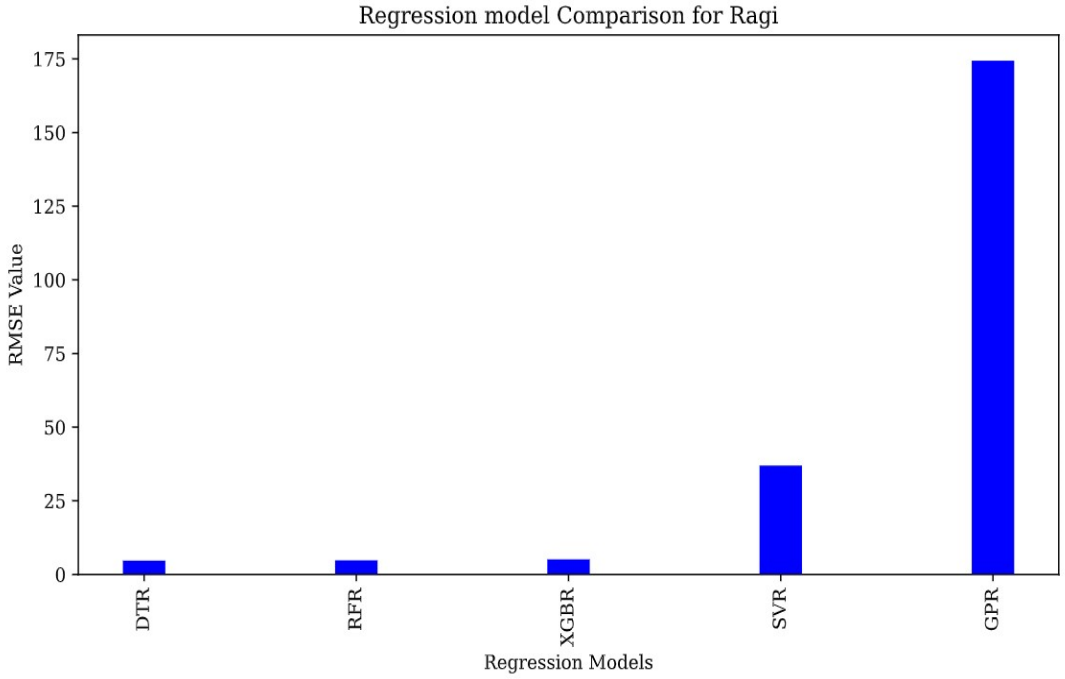

Figure (e). Ragi wholesale price data

Figures 2 (a), (b), (c), (d) and (e) depict the RMSE based comparison of regression models for all datasets. X- axis represent the base regression models. Y-axis represents RMSE value of the models. These bar charts clearly show the XGBoost Regression model has lower RMSE value for all crop datasets.

Table 7: Comparative Performance of Ensemble Methods

Crop	RMSE		Improvement Percentage
	Classical	Competitive	
Barley	22.579	5.69	74.79%
Maize	21.2567	0.75	96.47%
Paddy	17.3152	0.75	95.67%
Wheat	18.7354	3.0	84.03%
Ragi	26.977	2.48	90.81%

Table 7 show the comparative performance of classical ensemble method and competitive ensemble method for all crop datasets. The equation 8 calculates the percentage by which the RMSE has improved when transitioning from the classical ensemble model to the competitive ensemble model. It includes RMSE value for training datasets, test datasets and average RMSE value. It is clearly noted that competitive ensemble method has lower RMSE for all datasets.

RMSE Improvement Percentage = RMSE (classical ensemble model) – RMSE (competitive ensemble model) / RMSE (classical ensemble model) ×100% (8)

5. CONCLUSION

This study examined a very accurate agricultural price trend forecasting model using competitive ensemble learning using a RWS operator in the output of four regressor models for rainfall and crop datasets. Introducing competitive learning to each output of the forecasting model with multiple regressors instead of using a forecasting model with a single regressor was found to significantly improve model performance and increase the model's robustness. The outcomes of the proposed methodology demonstrated how competitive ensemble learning outperformed other single prediction models in terms of performance. The study's limitation due to the small number of characteristics in these datasets is a shortcoming. To counter this and improve accuracy going forward, more pertinent features can be added to the dataset in the future research.

ACKNOWLEDGEMENTS

The authors would like to thank everyone, just everyone!

REFERENCES

[1] D. Yang and Z. Liu, "Does farmer economic organization and agricultural specialization improve rural income? Evidence from China," Economic Modelling, vol. 29, no. 3, pp. 990–993, 2012.
View at: Publisher Site | Google Scholar

[2] S. Yao, Y. Guo, and X. Huo, "An empirical analysis of the effects of china's land conversion program on farmers' income growth and labor transfer," Journal of Environmental Management, vol. 45, no. 3, pp. 502–512, 2010.
View at: Publisher Site | Google Scholar

[3] L.-J. Chen, C. Ye, S.-W. Hu, V. Wang, and J. Wen, "The effect of a target zone on the stabilization of agricultural prices and farmers' nominal income," Journal of Agricultural and Resource Economics, vol. 38, no. 1, pp. 34–47, 2013.
View at: Google Scholar

[4] Flach P. Machine Learning: The Art and Science of Algorithms that Make Sense of Data. 2012: Cambridge University Press, Cambridge.

[5] Kumar D, Rath SK. Predicting the Trends of Price for Ethereum Using Deep Learning Technique. In: Dash S., Lakshmi C., Das S., Panigrahi B. (eds.) Artificial Intelligence and Evolutionary Computations in Engineering Systems. Advances in Intelligent Systems and Computing, 2020; 1056: 103–114. Springer, Singapore. https://doi.org/10.1007/978-981-15-0199-9_9

[6] Milunovich G. Forecasting Australia's real house price index: A comparison of time series and machine learning methods. Journal of Forecasting, 2020; 39(7): 1098–1118.

[7] Werbos PJ. Generalization of backpropagation with application to a recurrent gas market model, Neural Network. 1998; 1(4): 339–356.

[8] Zhang GP, Qi M. Neural network forecasting for seasonal and trend time series. Eur. J. Oper. Res. 2005; 160: 501–514.

[9] Zhang GP, Kline DM. Quarterly time-series forecasting with neural networks. IEEE Trans. Neural Netw. 2007; 18: 1800–1814.

[10] Weng Y, Wang X, Hua J, Wang H, Kang M, Wang FY. Forecasting Horticultural Products Price Using ARIMA Model and Neural Network Based on a Large-Scale Data Set Collected by Web Crawler. IEEE Trans. Comput. Soc. Syst. 2019; 6: 547–553.
[11] Sheehy JE, Mitchell PL, Ferrer AB. Decline in rice grain yields with temperature: Models and correlations can give different estimates. Field Crop Res. 2006;98(2–3):151–6.

[12] Zhang, Y. and Na, S. (2018) "A novel agricultural commodity price forecasting model based on fuzzy information granulation and Mea-SVM model," Mathematical Problems in Engineering, 2018, pp. 1–10. Available at: https://doi.org/10.1155/2018/2540681.

[13] Paul RK, Yeasin M, Kumar P, Kumar P, Balasubramanian M, Roy HS, et al. (2022) Machine learning techniques for forecasting agricultural prices: A case of brinjal in Odisha, India. PLoS ONE 17(7): e0270553. https://doi.org/10.1371/journal.pone.0270553.

[14] Chatzimparmpas, A.; Martins, R.M.; Kucher, K.; Kerren, A. StackGenVis: Alignment of Data, Algorithms, and Models for Stacking Ensemble Learning Using Performance Metrics. IEEE Trans. Vis. Comput. Graph. 2021, 27, 1547–1557.

[15] Li, Z.; Chen, Z.; Cheng, Q.; Duan, F.; Sui, R.; Huang, X.; Xu, H. UAV-Based Hyperspectral and Ensemble Machine Learning for Predicting Yield in Winter Wheat. Agronomy 2022, 12, 202.

[16] Razavi-Termeh, S.V.; Sadeghi-Niaraki, A.; Choi, S.M. Spatial Modeling of Asthma-prone Areas Using Remote Sensing and Ensemble Machine Learning Algorithms. Remote Sens. 2021, 13, 3222.

[17] Ustuner, M.; Sanli, F.B. Polarimetric Target Decompositions and Light Gradient Boosting Machine for Crop Classification: A Comparative Evaluation. ISPRS Int. J. Geo-Inf. 2019, 8.

Authors

Mr. R. Ragunath currently pursuing Ph.D. in Department of Computer Science, Periyar University, Salem India since 2022. I am published research papers in international conference including springer and it's also available online. Main research work focuses on A Responsible Artificial Intelligence System (RAIS) for Agriculture.

Dr. Ratipriya is an Assistant Professor in the Department of Computer Science, at Periyar University, Salem India. She is known for her exceptional contributions in the fields of data mining techniques and bio-inspired optimization. As an accomplished researcher, he has a strong publication record, with more than 60 research papers to her credit, indexed in prestigious databases. Apart from her many research articles, she has also authored more than 10 book chapters. Currently, she directs her research efforts to develop artificial intelligence-based predictive models.