International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Crop Price Estimation Using Stacking Ensemble Technique

[1]Dr. T Adilakshmi, [2]T. Jalaja, [3]M. Dheeraj Reddy, [4]Konduru Sai Kamal

Submitted: 28/01/2024 **Revised**: 06/03/2024 **Accepted**: 14/03/2024

Abstract: The agricultural sector in India, contributing approximately 17% to the GDP and engaging over 60% of the workforce, is at a pivotal juncture between traditional practices and modern technological advancements. This research explores the frontier of agricultural economics by employing an innovative approach to crop price prediction through the utilization of the Stacking Regressor algorithm. This advanced ensemble learning technique amalgamates the strengths of diverse regression models, harnessing their collective predictive power to yield superior accuracy. The study synthesizes historical crop data to construct a robust predictive framework. By integrating Random Forest and XGBoost regressors, the Stacking Regressor not only captures intricate patterns within the dataset but also adapts dynamically to the ever-changing agricultural landscape. This research aims to redefine the precision of crop price forecasting, offering a holistic and adaptable solution to stakeholders in the agri-business sector. As the global demand for sustainable and resilient agricultural practices intensifies, this research endeavors to empower farmers with a state-of-the-art solution, fostering informed decision-making and contributing to the advancement of a more resilient and efficient agricultural ecosystem.

Keywords: *Price Prediction, Stack Regressor, XGB Regressor, Random Forest Regressor.*

1. Introduction

Farmers, considered the backbone of our nation, have played a crucial role in India's history, dating back to the Indus Valley Civilization. Agriculture, deeply rooted in the economy, has contributed significantly to India's development by supplying food, delivering raw materials, and creating employment for a substantial part of the population. Despite its substantial role, the agricultural sector faces challenges, with more than 85% of the labour employed in this sector. While India's economic growth diversifies, approximately 50% of the population still relies on agriculture for their livelihood. The contribution of agriculture to India's GDP in terms of economy, is gradually diminishing, emphasizing the need for innovative solutions. Farmers encounter issues such as inadequate profits for their crops, highlighting the necessity for predictive, regression machine learning techniques for crop price forecasting. The well-being of crops directly impacts their value, with optimal conditions involving nutrient-rich soil and favourable climate conditions. India's agriculture encompasses diverse crops, including rice, wheat, pulses, cotton, coffee, maize etc. Despite the sector's vastness, crop yields in India fall significantly below international standards due to factors such as uneven rainfall, insufficient water management, and inefficiencies in harvesting and transportation. Tragically, farmer suicides have been on the rise in India for the past two decades, with nearly 300,000 farmers resorting to ingesting pesticides or hanging themselves due to crop price-related losses. In the 2011 census, there was a 45% rise in the suicide rate among Indian farmers, underscoring the urgency of addressing these challenges. The primary aim is to provide farmers with advanced crop price forecasts, enabling them to secure fair prices for their crops and, ultimately, mitigate the distressing issue of farmer suicides.

1.1. Abbreviations and Acronyms

RMSE- Root Mean Squared Error

MSE- Mean Squared Error

MAE- Mean Absolute Error

R^2- R-Square Error

RSS- Sum of Squares of Residuals

TSS- Total Sum of Squares

1.2. Related Works

[1] "Crop Price Prediction System using Machine Learning Algorithms" emphasizes the critical nature of price prediction in agriculture, particularly in forecasting crop prices for the future rotation based on real data. The paper underscores the importance of thorough data analysis, cleansing, and exploratory data analysis (EDA) to comprehend the various parameters within the dataset.

[1] *Professor & HOD - Computer Science and Engineering Department, Vasavi College of Engineering,
Hyderabad, India*
[2] *Assistant Professor- Computer Science and Engineering Department, Vasavi College of Engineering,
Hyderabad, India.*
[3,4] *Computer Science and Engineering Department, Vasavi College of Engineering, Hyderabad, India*
* *Corresponding Author Email t_adilakshmi@staff.vce.ac.in
jalaja.t@staff.vce.ac.in saikamalkonduru@gmail.com
dheerajreddymeka@gmail.com*

Employing diverse data mining techniques, the study aimed to construct an accurate model for precise price prediction. Several Machine learning algorithms, including Decision Trees, Linear regression, and XGBoost were employed for this purpose. Notably, XGBoost demonstrated superior performance among these algorithms in predicting crop prices, showcasing its effectiveness in the agricultural price prediction system.

[2] “Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector” suggested a way to predict the crop yield from the historical data that includes factors like ph., rainfall, crop, humidity, and temperature. They have used to algorithms random forest and decision tree, of which Random Forest has provided more accurate result.

The paper entitled [3] “Stacking Strong Ensembles of Classifiers” has clearly explained how the stacking technique can be leveraged to combine the predictions of multiple prediction models and generate a meta model that can be used trained and can be used to predict the results. It has also discussed the architecture of a stacking ensemble model which consisted of different components in order to build an effective model that one needs to take care of.

2. DATASET

The Crop Price Estimation dataset contains historical data of past 8 years which contain crop type, district, Profit price per quintal for each crop in each month for every year.

2.1. Dataset Description

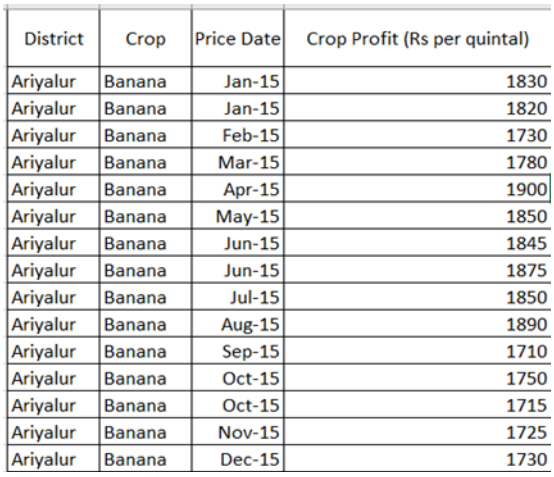
(I)District Attribute –Districts in India where the crop is predominantly grown.

(II) Crop Attribute – This data provides details about different crops that are considered for their profit price estimation

(III) Month Attribute– This data is considered to estimate the profit on a monthly basis for more clarity.

(IV) Crop Profit Price- This column provides insight about the profit obtained per quintal of that respective crop

Number of records in the dataset are 30000 rows and 4 columns. The Dataset is gathered primarily from two sources Home | Open Government Data (OGD) Platform India and Socio-Economic Statistics India, Statistical Data Figures Year-Wise (indiastat.com).



District	Crop	Price Date	Crop Profit (Rs per quintal)
Ariyalur	Banana	Jan-15	1830
Ariyalur	Banana	Jan-15	1820
Ariyalur	Banana	Feb-15	1730
Ariyalur	Banana	Mar-15	1780
Ariyalur	Banana	Apr-15	1900
Ariyalur	Banana	May-15	1850
Ariyalur	Banana	Jun-15	1845
Ariyalur	Banana	Jun-15	1875
Ariyalur	Banana	Jul-15	1850
Ariyalur	Banana	Aug-15	1890
Ariyalur	Banana	Sep-15	1710
Ariyalur	Banana	Oct-15	1750
Ariyalur	Banana	Oct-15	1715
Ariyalur	Banana	Nov-15	1725
Ariyalur	Banana	Dec-15	1730

Fig.1. Snapshot of Dataset.

The dataset contains details about 10 crops - Banana, Bengal gram, Black gram, Coconut, Coffee, Cotton, Green gram, Maize, Red Gram and Rice. The crop profit price information is collected every month and for every year for the period of 2015-2022 so as to predict with good accuracy using regression algorithm.

3. Proposed Methodology

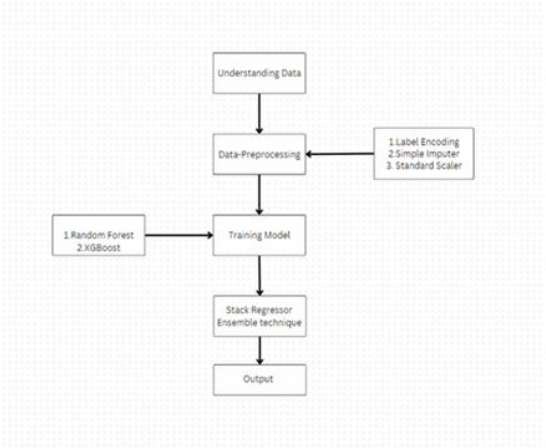


Fig. 2. Architecture of Machine Learning model built using Stacking Ensemble Technique

Data Preprocessing

Data preprocessing, also known as data preparation or cleaning, is the process of preparing data for analysis. The goal is to improve the quality of the data and make it more suitable for a specific task. Steps involved are:

1. Data Cleaning: This technique includes checking for any missing values, invalid formatting or duplicate records and removing them. Categorical variables, namely 'District' and 'Crop', undergo label encoding using Label Encoder, converting them into numerical format for compatibility with Machine learning algorithms. The 'Price Date' column is converted to a date-time format, and additional features, 'Month' and 'Year', are derived from it. The redundant 'Price

Date' column is subsequently dropped, streamlining the dataset.

2. Data Transformation: Data Transformations are mathematical operations that change the shape or scale of data. It allows us to understand the data more precisely and process it numerically by various techniques. Data Transformation includes Normalization, Standardization, Attribute Selection. Feature scaling is applied using Standard Scaler. This normalization process standardizes the numerical features ('Month' and 'Year') to have zero mean and unit variance, preventing certain features from disproportionately influencing the model training.

Stacking Ensemble Technique

Stacking is an important ensemble technique in which, a learner i.e the meta-learner/second level learner is trained to combine the first-level learners known as individual learners efficiently. First, Individual learners are trained and the obtained predictions are used as an input for the meta model, basically the obtained predictions are used as training data for the meta model. Now the meta model is trained and it gives the final prediction that leverages the power of all the individual learners to give the best possible output. The first level learners are often made up of different and diverse learning for better prediction.

The individual learners leveraged in the proposed solution are

1. **Random Forest Regressor**: Random Forest Regressor is an ensemble learning algorithm that excels in predictions for regression tasks. A "forest" of trees is constructed by combining the outputs of multiple decision tree regressors. Each tree predicts the target variable independently, and the final prediction is derived by averaging or taking a weighted sum of these individual predictions during its operation.. The strength of the Random Forest Regressor lies in its ability to mitigate overfitting, enhance accuracy, and handle complex relationships within the data. In the training process, every decision tree is crafted using a random subset of the training data. At each step of building the tree, we look at a random subset of features to decide how to split the nodes. This randomness injects diversity into the forest, preventing it from being overly influenced by noise in the data. The result is a robust and versatile predictive model that tends to generalize well to unseen data. Random Forest Regressor is widely employed in various domains, including finance, healthcare, and agriculture, where its ensemble approach contributes to more accurate and reliable predictions.

2. **XGBoost Regressor**: XGBoost, short for Extreme Gradient Boosting, is a versatile and powerful machine learning algorithm renowned for its effectiveness in regression tasks. It belongs to the family of gradient boosting algorithms and is characterized by its exceptional predictive performance and efficiency. XGBoost sequentially builds an ensemble of weak learners, typically decision trees, and combines their predictions to create a robust and accurate model. The algorithm employs a gradient boosting framework, where subsequent trees focus on minimizing the errors of the preceding ones. Key features of XGBoost include its regularization techniques, which help prevent overfitting, and its ability to handle missing data. XGBoost optimizes a specific objective function that combines the model's predictive accuracy and complexity, ensuring a balance between precision and simplicity. Additionally, it supports parallel processing, making it computationally efficient and scalable. XGBoost has found applications across diverse domains, including finance, healthcare, and competitive data science, where its superior performance has made it a popular choice for regression tasks.

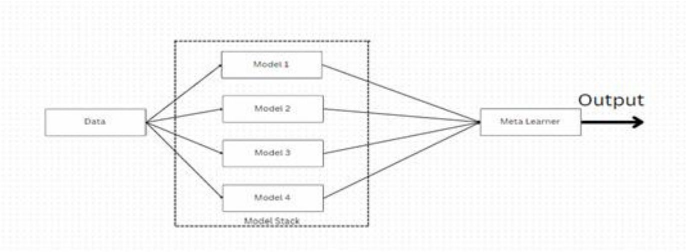


Fig. 3. General Architecture of a Stacking algorithm

4. Proposed Algorithm

1. Load the dataset and convert it into data frames.

2.Data preprocessing which includes encoding categorical variables by using LabelEncoder, Missing values in the dataset are handled using Simple Imputer, and Standard Scaler is used for data normalization.

3. Data is split into training and testing sets using train_test_split().

4. The Crop name whose profit price that needs to be estimated should be entered.

5. In the proposed solution we are using two regressors, Random Forest Regressor and XGBosst Regressor

6. RandomForestRegressor(n_estimators=200, max_depth=12, random_state=20) model with the respective hyperparameters is trained.

7. For Xgboost regressor , the parameters that are choosen are

```
params = {
        'n_estimators': [100, 150, 250],
        'max_depth': [3, 5, 7],
        'learning_rate': [0.01, 0.05, 0.1],
        'reg_lambda': [0.1, 0.5, 1.0]
```

}

In order to select the best combination of hyperparameters , GridSearch CV is used. The best parameters that are obtained from the GridSearch CV is { n_estimators: 150, reg_lambda: 1.0, max_depth: 7, learning_rate : 0.1}. XGBRegressor(best_params) is trained .

8. Using the Stacking Ensemble technique , the power of these two models is combined into a single ensemble model using the StackingRegressor i.e StackingRegressor(estimators=[('rf', rf), ('xgb', xgb)], final_estimator=meta_model) , where rf refers to the random forest regressor model and xgb refers to xgboost regressor respectively.

9. Now the StackingRegressor is trained on the dataset and is used to predict the Crop profit price estimation and is displayed so that the farmer has a clear idea of how much profit per quintal he can obtain by selling the crop thus enhancing his decision making to choose right crop before hand or to sell at a correct time in the market.

5. Results

The Machine Learning model that is built using Stacking algorithm with the combination of Random Forest Regressor and XGBoost Regressor, predicts the profit that can be obtained on per quintal every month for the next 3-4 months so that it provides a holistic view for a farmer to take an informed-decision keeping in view profit that can be gained.

Root Mean Squared Error:

$$RMSE = \sqrt[2]{\sum_{i=1}^{n} \frac{(Yi - Y{}^\wedge i)^\wedge 2}{n}}$$

Y1^,Y2^,……,Y^n are predicted values

Y1,Y2,…….,Y^n are observed values

n is the number of observations.

Mean Squared Error

$$MSE = 1/n \sum_{i=1}^{n} (Y_i - Y{}^\wedge{}_i)^\wedge 2$$

MSE = mean squared error

n = number of data points

Yi = observed values

Y^i = predicted values

Mean Absolute Error:

$$MAE = \frac{\sum_{i=1}^{n} abs(y_i - x_i)}{n}$$

MAE = mean absolute error

Yi = prediction

Xi = true value

n = total number of data points

R-Square value:

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 = coefficient of determination

RSS = sum of squares of residuals

TSS = total sum of squares

RSS = $\sum(Y_i - Y{}^\wedge{}_i)^\wedge 2$

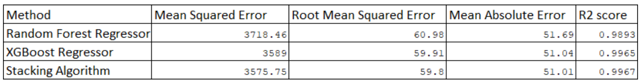
Where: Y_i is the actual value and $Y^\wedge{}_i$ is the predicted value.

TSS = $\sum(Y_i - Y^-{}_i)^\wedge 2$

Where: Y_i is the actual value and $Y^-{}_i$ is is the mean value of the variable/feature

Here , y_i is the actual crop profit price and $y^\wedge{}_i$ is the predicted price by our model.

Results:



Method	Mean Squared Error	Root Mean Squared Error	Mean Absolute Error	R2 score
Random Forest Regressor	3718.46	60.98	51.69	0.9893
XGBoost Regressor	3589	59.91	51.04	0.9965
Stacking Algorithm	3575.75	59.8	51.01	0.9967

Fig. 4. Results of Various algorithms based on the error score

It is clearly evident that Stacking Algorithm provided better results than Random Forest Regressor and XGBoost Regressor.

6. Conclusion and Future Scope

In conclusion, the integration of a sophisticated stacking ensemble technique, combining the strengths of Random Forest Regressor and XGBoost Regressor, into the realm of predicting crop profit per quintal presents a transformative approach for empowering farmers with informed decision-making capabilities. While traditional methods of profit prediction may fall short in capturing the intricate dynamics of agricultural markets, the proposed model leverages the collective power of two robust algorithms to overcome these challenges.

At the core of this predictive model lies a meticulous process of data preprocessing, model stacking, and comprehensive validation. The synergy between Random Forest and XGBoost harnesses the complementary strengths of both algorithms, enabling the identification of nuanced patterns within historical crop data that may elude individual models.

This ensemble approach ensures a more accurate and resilient prediction of profit per quintal, offering farmers a valuable tool for optimizing their cultivation strategies.

The envisioned benefits of this model are substantial. By providing farmers with a reliable estimate of expected profits beforehand, it empowers them to make strategic decisions in crop selection, resource allocation, and market engagement. The stacking ensemble's adeptness at capturing non-linear relationships in historical data enhances the precision of profit predictions, allowing farmers to navigate the uncertainties of the agricultural landscape more effectively.

The Proposed algorithm can be used as a backbone of a Web Application where a farmer can enter his soil profile and get a list of suitable crops that can be grown on his soil , with their detailed analysis of the profit prices of each crop , so that farmer can take a right informed decision before choosing the crop and estimate the profit that he can get. Moreover , by the inclusion of data of wide variety of crops also increases the options to grow a crop and the prediction can be based on season also. To make the web application more intuitive , a detailed guide on how to grow the crop , list of premium seeds that are available in the market that can be procured , list of best suitable pesticides and fertilizers that enhance the yield can be displayed so that it makes farmers job easier. A list of procurement centers that are near to him can be displayed so that farmer can sell his yield comfortably. Probably , if a new regression algorithm is developed it can also be used in combination with the current ensemble technique for a better output.

Furthermore, the adoption of this innovative ensemble technique aligns with the evolving needs of the agriculture sector. As farmers grapple with volatile market conditions and the need for sustainable practices, this model not only enhances predictive accuracy but also promotes a proactive and data-driven approach to farming. By embracing the combined power of Random Forest and XGBoost, the agriculture industry can anticipate a future where farmers are better equipped to make informed decisions, ultimately contributing to the sector's resilience and growth.

Author contributions

Meka Dheeraj Reddy: Conceptualization, Methodology, Software, Field study **Konduru Saikamal:** Data curation, Writing-Original draft preparation, Software, Validation., Field study **Dr. T. Adilakshmi:** Visualization, Investigation, Writing-Reviewing and Editing. **T. Jalaja:** Visualization, Investigation, Writing-Reviewing and Editing

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Pandit Samuel, B.Sahithi, T.Saheli, D.Ramanika, N.Anil Kumar “Crop Price Prediction System using Machine learning Algorithms." Quest Journals Journal of Software Engineering And Simulation, Vol. 06, No. 01, 2020,Pp. 14-20.

[2] Kumar, Y. Jeevan Nagendra, V. Spandana, V. S. Vaishnavi, K. Neha, and V. G. R. R. Devi. “Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector”.

[3] “Stacking Strong Ensembles of Classifiers” by Stamatios-Aggelos N. Alexandropoulos, Christos K. Aridas, Sotiris B. Kotsiantis, and Michael N. Vrahatis.

[4] Rohith R, Vishnu R, Kishore A, Deeban Chakkarawarthi, "Crop Price Prediction and Forecasting System using Supervised Machine Learning Algorithms", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 9, Issue 3, March 2020.

[5] R. Nagini, “Agriculture yield prediction using predictive analytic techniques”, 2nd International Conference on Contemporary Computing and Informatics, 2016.

[6] S.Veenadhari, Dr. Bharat Misra, Dr. CD Singh.2019. Machine learning approach for forecasting crop yield based on climatic parameters.

[7] Yung-HsinPeng, Chin-Shun Hsu, and Po-Chuang Huang Developing Crop Price Forecasting Service Using Open Data from Taiwan Markets, 2017 IEEE

[8] Bharati Panigrahi, Krishna Chaitanya Rao Kathala, M. Sujatha. "A Machine Learning Based Comparative Approach to Predict the Crop Yield Using Supervised Learning With Regression Models" , Procedia Computer Science, 2023