Crop Yield Prediction using Gradient Boosting Regression

Rahil Khan, Pratyush Mishra, B. Baranidharan

Abstract: ***Achieving greater crop yields remains a pressing challenge for both farmers and governments. This research examines the use and implementation of Gradient Boosting Regression in predicting crop yields for numerous districts in France. XGBoost, an efficient, optimized and flexible distributed gradient boosting library was used. Agricultural data was sourced from the CLAND Institute's 'Crop Data Challenge 2018', which contains approximately 38 years of maize data compiled by various departments from the months of January to September from regions in France. Average monthly climatic parameters such as evapotranspiration, maximum and minimum temperature, cumulative precipitation, yield anomaly, solar radiation levels and an irrigation coefficient were used as input variables. The best result obtained was an RMSE value of 0.755 and MAE error of 0.54 was obtained using a tuned XGBoost Regressor trained on original variables. This paper aims to compare various regression techniques in order to improve yield prediction thus giving farmers a chance to improve their cultivation with better insights as well as enabling them to harness the power of predictive analytics.***

Keywords: Gradient boosting, Yield prediction, Maize, Regression, XGBoost.

I. INTRODUCTION

The annual crop yield is the total volume of crop harvested per unit area in a given year. In the case of Maize, often expressed in tons per hectare, is the quantity of grain harvested per season. It mostly depends on the characteristics of the region where the Maize is grown and the climatic conditions of the year in this region (temperatures, radiation, rainfall etc.). The value of the yield is likely to vary greatly between regions and between years. For example, the yield may be abnormally low in a year with a significant water deficit at a key stage of crop development or, conversely, very high in a year with optimal climatic conditions throughout the season. Gopal, Bhargavi[1], surveyed the predictive ability of various AI/ML models, using feature subsets selected by variety of different algorithms. It was concluded that factors such as temperature, tank size, production area, open well and canal length gave better and more accurate models.

Thus, it is important to accurately predict pre-harvest performance. This is usually done in the fall, during the harvesting season. Reliable pre-harvest predictions provide the opportunity for regional economic operators to plan their Harvest, manage their stocks and optimize their contracts along with grain purchases and sales. Performance predictions are also used by organizations operating in international markets. Narasimhamurthy, Kumar [2] developed improved techniques to predict crop production under a variety of different weather conditions using Random Forests. Their study was done to aid farmers and associated stakeholders to make better agronomical and crop decisions. Abundant crop predictions or, on the contrary, predictions of significant losses, can strongly impact world agricultural market prices.

The main objective of this research is to develop tools to predict maize yields in France as accurately as possible, and then subsequently apply the tried and tested model to predict yields for other varieties of crops.

Using the parameters given in the dataset along with tuned regression models, analysis was done and subsequent predictions were made, with an emphasis on the Gradient Boosting Regression technique using the XGBoost library.

The remaining portion of this paper is organized into several sections. Section II describes the literature review, Section III details the methodology used, results obtained are discussed in Section IV and Section V summarizes and concludes the research findings in the paper.

II. LITERATURE REVIEW

Su, Xu, Yan [3] developed a support vector machine-based open crop model by assimilating various developmental stage models along with yield prediction models. An open input framework and scale-independent factors facilitated large-scale integration of data. The main objectives were to determine the hyperparameters, penalty coefficient and optimal kernel functions to enable investigations on three types of rice plantings.

Saad, Rusli [4] explored 4 supervised learning algorithms that predict rice yield based on weeds, diseases and pests in Kedah, Malaysia. It was found that Conjugate Gradient Descent algorithm exhibits the best performance as compared to Levenberg-Marquardt, Quick propagation and Back-propagation algorithms.

Ji, Sun, Yang [5] proposed and evaluated ANN models to predict rice yield for the mountainous regions in Fujian, China. A comparison between artificial neural networks and linear regression models was also conducted. For training, models were fed with historical yield data and field-specific rainfall data, including weather variables. It was found that neural network models outperformed linear regression models by producing more accurate yield predictions, but required huge volume of data.

Ramesh, Vardhan, [6] described crop Density based clustering and Multiple Linear Regression based crop yield prediction techniques.

Revised Manuscript Received on January 06, 2020.

***** Correspondence Author

Pratyush Mishra,* Department of Computer Science and Engineering, Kattankulathur Campus, SRM Institute of Science and Technology, Chennai, India. Email: pratyush.mishra64@gmail.com

Rahil Khan, Department of Computer Science and Engineering, Kattankulathur Campus, SRM Institute of Science and Technology, Chennai, India. Email: khan98rahil@gmail.com

Dr. B. Baranidharan, Department of Computer Science and Engineering, Kattankulathur Campus, SRM Institute of Science and Technology, Chennai, India. Email: baranidb@srmist.edu.in

The area of study was the East Godavari district of Andhra Pradesh in India.

Gandhi, Petkar, Armstrong [7] developed and used Support Vector Machines (SVM) to predict rice crop yield. Basically, it was built as a classifier model. It was found that the Multilayer perceptron, Bayesian and Naïve Bayes based networks performed considerably better by producing consistently accurate, sensitive and specific predictions of yield as compared to other classifiers.

Sellam, Poovammal [8] studied on the ways in which Area Under Cultivation, Annual precipitation and food price index affects the yield of crop using linear regression method to create the model and find the interdependence of these parameters.

Chemchem, Alin, Michael [9] created a classifier model using random forest in order to forecast the wheat yields. They used SMOTE as a pre-treatment in order to boost the accuracy of the model.

Charoen, Pradit [10] compared non-machine learning predictive models with machine learning predictive models and found a significant difference where machine learning models were able to predict yield with 20% more accuracy than the other traditional non-machine learning techniques.

III. PROPOSED METHODOLOGY

A. Gradient Boosting

XGBoost is an open-source library that implements gradient boosted decision trees that are efficient and highly optimized. In gradient tree boosting, models are not trained in isolation of each other, but rather in succession, where each model iteratively reduces errors made by previous models. The regular updating of the weights of leaves within a tree ensemble model allows for deriving an optimal model that can minimize the evaluation formula. As an outline, predicted values of the tree ensemble model are calculated as follows.

$$\hat{y_i} = \emptyset(x_i) = \sum_{k=1}^{n} K = f_k(x_i)\,, f_k \in \mathcal{F}\#(1)$$

It is noteworthy to mention that

$$\mathcal{F} = f(x) = w_{q(x)}(q : \mathbb{R}^m \to T, w \in \mathbb{R}^t)\ \#(2)$$

is regression tree space. x_i is input, $\hat{y_i}$ is output, q is the tree structure, and the number of leaves in the tree is identified as T. Each f_k matches the independent tree structure q and the weight w . w_i represents the score of the i^{th} leaf. This predicted value can be evaluated by

$$L(\emptyset) = \sum_i l(\hat{y_i}\,, y_i) + \sum_k \Omega\,(f_k) \qquad (3)$$

$$where\ \Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2\ \#(4)$$

where L is the loss function to find the difference between the predicted value $\hat{y_i}$ and the target value y_i. Ω represents the complexity of the model and is a regularization term that has the function of smoothening the weight to avoid overfitting.

B. Dataset and Variables

The first phase in the development of a prediction model is the compilation and preparation of sample data. An extensively compiled data is needed to suffice as input data For training the network. For this project we chose publicly available data from the CLAND Crop Data Challenge 2018. It contains data for maize crop and comes along with both training and testing files for analysis.

Table – 1: Describing variables in dataset

Variable Name	Description
Yield_anomaly	Predictor variable representing corn yield anomaly, expressed in tons per hectare
Year_harvest	Anonymous year of harvest (1-57)
NUMD	France Department number (anonymous)
ETP	Monthly average evapotranspiration by year
PR	Cumulative precipitation per year and dept.
RV	Average monthly radiation per year and dept.
SeqPR	Number of rainy days per year and dept.
Tn	Monthly Average minimum temperature
Tx	Monthly Average maximum temperature
IRR	Variable between 1 and 5 linked to the fraction of irrigated agricultural area in each department.

The variables contained within the dataset are described in Table 1.

The objective variable is the yield of maize that is to be predicted on the basis of prior data. It is given as 'yield anomaly', which is a measure of relative yield. A negative value indicates a decrease in yield with respect to prior years and a positive value indicates an increase in yield.

The dataset contains pertinent weather and rainfall statistics that greatly influence values of yield each year.

Data gathered is from the months of January till September and not the entire year as maize is not grown nor harvested during the other months.

C. Research Methodology

We begin by dropping relatively unimportant variables such as Department Number (NUMD) and anonymous year of harvest as they serve little to no value, except for visualization.

A 'DMatrix', an optimized and memory efficient matrix is created for use by XGBoost. The data is then split into two portions, one for testing and the other for training (70% for training, 30% for testing).

Using a tuned 'XGBRegressor' model we tested for optimal parameters that will allow for the most accurate predictions. After fitting and evaluating performance model, K-fold (k=10) cross validation is performed for unbiased performance results. The various parameters used for tuning are described in the next section.

Other regression techniques, namely Ada Boosting Regression, Random Forest Regression, Kernel Ridge Regression and K-Nearest-Neighbor (KNN) Regression are tested for the performance comparison.

Feature importance is also obtained for the top 2 best performing regression models and plotted for convenience. Python's 'Matplotlib' library is used to plot comparison graphs for the error values calculated for each regression model.

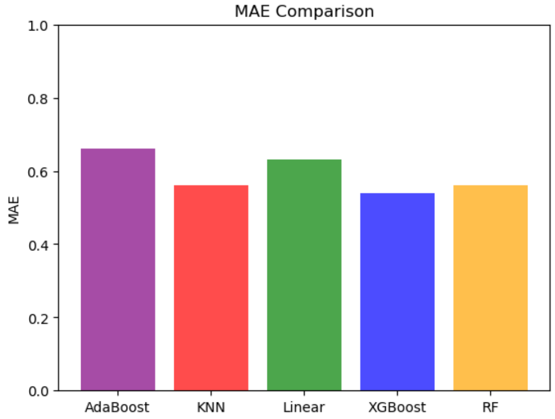

Fig. 1. Mean Absolute Error Comparison

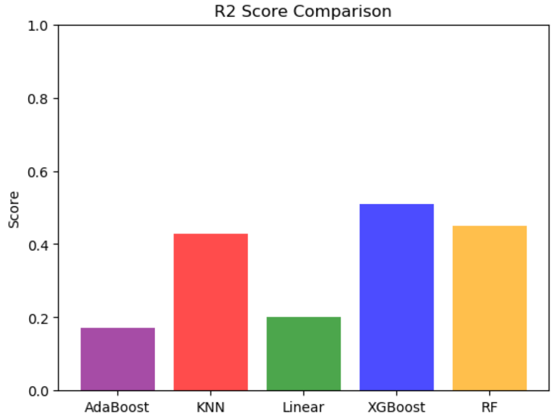

Fig. 2. R2 Score for each model

IV. RESULTS AND DISCUSSION

A. Experiment Description

Each model is trained utilizing original and unaltered values of the respective weather, rainfall and yield variables to train the predictive model. Each of the 6 main parameters have 9 columns with values for each month, totaling to 54 columns of data. Adding the irrigation coefficient, the columns increase to 55. Using the sklearn 'train_test_split' the data is distributed randomly for training and testing purposes.

B. Performance Evaluation

Frequently used regression metrics such as Root Mean Squared Error (RMSE), along with Mean Absolute Error (MAE) are used. An R^2 score is also calculated to ascertain how well the regression models approximate real data points. Each value is obtained after applying model to test data separated prior to experimentation.

For each equation, the value 'n' denotes number of values, while 'y_j' represents actual values and '$\hat{y}_j$' depicts predicted values. Mean Absolute Error does not consider their direction, but measures the average magnitude of errors in a set of predictions. Its formula is given as

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right| \#(6)$$

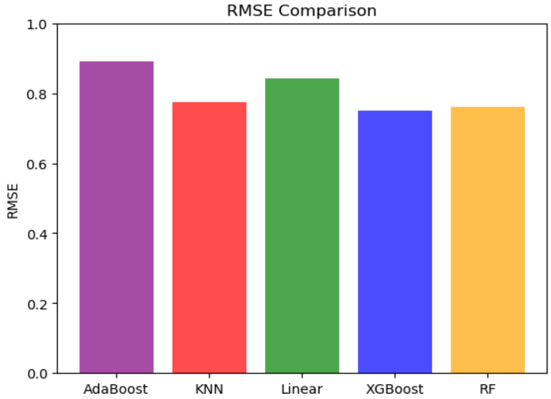

Fig. 3. Root Mean Squared Error Comparison

Root Mean Squared Error is a scoring rule that is quadratic in nature. It also measures the average magnitude of error. The square root of the average of squared differences between prediction and actual observation is the final value. Its formula can be described as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2 \#(7)}$$

Lesser values indicate better performance, as both metrics are negatively-oriented. RMSE gives a high penalty to large errors and will always be greater in magnitude than MAE.The R^2 score, also known as the coefficient of determination is the proportion of variance in the dependent variable with respect to independent variables. It is a measure of strength of how well the model relates to its predictions. Its formula is

$$R^2 \equiv 1 - \frac{SS_{res}}{SS_{tot}} \#(8)$$

$$where\ SS_{res} = \sum_{i}(y_i - f_i)^2 \#(9)$$

$$and\ SS_{tot} = \sum_{i}(y_i - \bar{y})^2 \#(10)$$

SS_{res} represents residual sum of squares and SS_{tot} represents total sum of squares. Table 2 depicts the performance comparison among various models.

Table – 2: Performance of regression models

ML Approach	RMSE	MAE	R2 Score
Ada Boosting	0.89	0.666	0.17
K-Nearest-Neighbors	0.774	0.561	0.427
Linear Regression	0.843	0.632	0.2
Gradient Boosting	**0.755**	**0.54**	**0.51**
Random Forests	0.757	0.55	0.49

Table – 3: XGBoost **parameters used for prediction**

Parameters	Values
learning_rate	0.2
max_depth	4
colsample_bytree	0.65
alpha	7.88
min_child_weight	2

C. Result Analysis

Table 2 outlines the results obtained by each regression model used. The parameters considered for the comparison of the different models where RMSE, MAE and R^2 score.

The tuned parameters for the best performing Gradient Boosting Regression model is listed in Table 2.

Figures 1, 2 and 3 depict in graphical format the error values of each model. There is only a slight difference in error values between the XGBoost and Random Forest models.

The tuned AdaBoost Regression model performed the worst in all error value tests. This could be attributed to its inflexibility in classifying weak and strong learners.

RMSE values compare the mean values hence it is the best criteria used to describe the model and hence Gradient Boosting method had the lowest value i.e. 0.755, Random Forest also had quite comparable values but taking into consideration the data size Gradient Boosting was considered as the more superior model. The other models had significantly high values of RMSE.

Hence, XGBoost can be used to create a predictive regression model based on the data to obtain the yield values of different crops.

D. Variable Importance

To determine the general importance of variables used in prediction we take the top 2 regression models, namely Random Forests and Gradient Boosting, which have nearly similar performance characteristics into consideration. Figures 1 and 2 describe the importance of each weather and rainfall characteristic in the form of a horizontal bar chart in descending order.

On observing the two graphs, we can understand that weather plays a pivotal role in prediction. Characteristics in relation to rainfall are shown to be vital for accurate predictions. For Gradient boosting, Evapotranspiration (ETP), which is the sum of plant transpiration and evaporation from the earth surface to the atmosphere is the most important variable.

In the Random Forest regression model, Precipitation (PR), which is basically water that reaches the ground from the atmosphere in the form of rain, is the most significant.

The irrigation coefficient (IRR) takes a backseat as it is classified as the least influential variable for both models. This may be attributed to the fact that the coefficient is not raw data, but instead an abstract tool to measure the area of irrigation. The relative importance of IRR is so low that it could hinder predictive accuracy, and could possibly be excluded from the training data.

Temperature holds the middle ground in terms of importance, while still being relevant to both models.

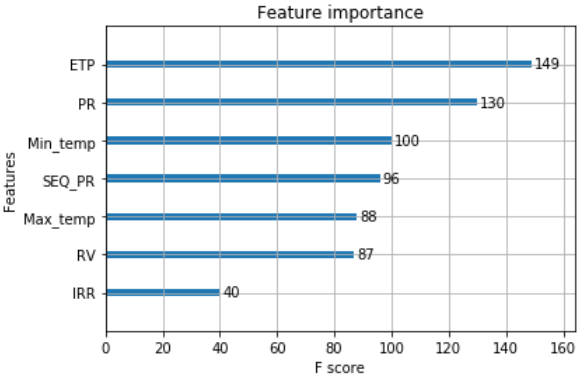

Fig. 4. Variable importance for Gradient Boosting

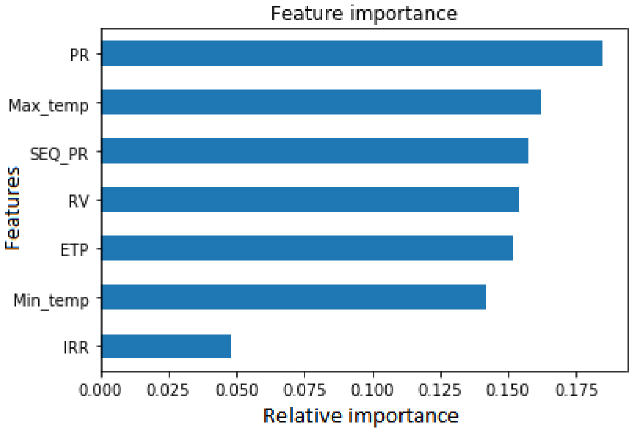

Fig. 5. Variable importance for Random Forests

On one hand, minimum temperature is highly influential for Random Forests and on the other hand minimum temperature edges out as a significant parameter for Gradient boosting.

Other variables such as average monthly radiation (RV) and number of rainy days (Seq_PR) hold lesser significance.

V. CONCLUSION

Crop yield prediction is essential for formulating a country's food policies. Timely and accurate predictions will be of great help to economic strategists. It would lead to proper export and import formulation based on the agricultural products. As we have identified in the literature survey, Machine learning models give consistently better results than non-machine learning models.

Based on the experiments done and the research methodology employed in this study XGBoost, also known as Extreme Gradient Boosting, was found to be the best model for prediction of yield. This study was able to achieve RMSE value of 0.755, MAE of 0.54 and R2 value of 0.51. In terms of variable importance, it was found that weather variables such as precipitation and evapotranspiration were vital for predictive accuracy.

Random Forest Regression also showed decent performance characteristics and could also be expanded on further to obtain better results.

For future research work, the aim would be to create improved models for different crops and try to enhance model accuracy by using larger amounts of descriptive data. The approach used in this study is intended for application on a variety of crops, such as wheat, rice, etc.

REFERENCES

1. Maya Gopal P. S. & Bhargavi R.: "Performance Evaluation of Best Feature Subsets for Crop Yield Prediction Using Machine Learning Algorithms", Applied Artificial Intelligence, 2019.
2. SML Venkata Narasimhamurthy, AVS Pavan Kumar, "Rice Crop Yield Forecasting Using Random Forest Algorithm", SSN: 2321-9653.
3. Ying-xue Su, Huan Xu, Li-jiao Yan, "Support vector machine-based open crop model (SBOCM): Case of rice production in China", Saudi Journal of Biological Sciences, Volume 24, Issue 3, 2017, Pages 537-547, ISSN 1319-562X.
4. Saad, Puteh & Jamaludin, Nor & Rusli, Nursalasawati & Bakri, Aryati & Kamaruddin, Siti. "Rice Yield Prediction - A Comparisonbetween Enhanced Back Propagation Learning Algorithms", 2004.

5. Ji, B., Sun, Y., Yang, S., & Wan, J. (2007). "Artificial neural networks for rice yield prediction in mountainous regions", Journal of agricultural science, 145, 249-261.
6. D. Ramesh, B. Vishnu Vardhan, "Analysis of Crop Yield Prediction Using Data Mining Techniques", IJRET Volume: 04 Issue: 01, Jan-2015.
7. Niketa Gandhi, Owaiz Petkar, Leisa J. Armstrong, Amiya Kumar Tripathy, "Rice Crop Yield Prediction in India using Support Vector Machines", IEEE 2016.
8. Sellam, Poovammal, "Prediction of Crop Yield using Regression Analysis", Indian Jounal of Science and Technology, Vol 9(38), 10.17485/ijst/2016/v9i38/91714, October 2016.
9. Chemchem, Alin, Krajecki, "Combining SMOTE sampling and Machine Learning for Forecasting Wheat Yields in France", 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE) 978-1-7281-1488-0/19
10. Phusanisa Charoen-Ung, Pradit Mittrapiyanuruk, "Sugarcane Yield Grade Prediction using Random Forest and Gradient Boosting Tree Techniques", 2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE) ©2018 IEEE

AUTHORS PROFILE

Rahil Khan is currently pursuing his Bachelor's degree in Computer Science and Engineering from SRM IST. His interests include Machine Learning, Data Science and Artificial Intelligence. He is dedicated to research further into the field of analytics, while simultaneously honing his technical and mathematical skillset.

Pratyush Mishra is a final year undergraduate student of SRM Institute of Science and Technology pursuing B.Tech in Computer Science and Engineering. He has a keen interest in the field of Artificial Intelligence and Machine Learning and has a vision to apply these concepts for the betterment of society.

Dr.B.Baranidharan has completed his Master of Technology in Computer Science and Engineering from SRM IST, Chennai and PhD in Wireless Sensor Networks (specialization) from SASTRA Deemed University, Thanjavur. Currently, he is working as Associate Professor in the department of CSE, SRM IST. He is having more than 10 years of academic experience and have published 22 papers in various International Journals and Conferences. Earlier his research involved about designing new clustering architecture for Wireless Sensor Networks and Internet of Things using various computational techniques. His current research includes Artificial Intelligence, Machine learning, Deep learning and Internet of Things.