Factors Impacting Crop Yield Prediction

by

(Team F)

Chaturya Gajula

Harshali Narkhede

Neha Nooka

Final Project Report

for

DATA 606: Capstone in Data Science

2024

ABSTRACT

Imagine a world where farmers can predict their harvest with exceptional accuracy. Our research on crop yield prediction brings us closer to this reality. We're meticulously analyzing a massive dataset from the FAO, encompassing production details, land use patterns, trade activity, and even fertilizer data. By dissecting these intricate elements, we aim to identify the key factors influencing future crop yields. The ultimate goal? Powerful models that can forecast harvests with precision, regardless of unpredictable seasons. Equipped with this knowledge, farmers can adapt their practices, ensuring our food security for years to come. This research isn't just about data analysis; it's about feeding the world, one informed harvest at a time.

ACKNOWLEDGMENTS

We team members would like to express our sincere gratitude to **Professor Unal Sakoglu** for their invaluable guidance and support throughout this project. We also appreciate the collaboration and dedication of our team members: Chaturya Gajula, Harshali Narkhede, and Neha Nooka. This project would not have been possible without the collective effort and contributions of everyone involved. Thank you all for your hard work and commitment.

TEAM MEMBERS' CONTRIBUTIONS

Chaturya performed the literature review, optimized the model, and played a key role in the preparation and data gathering. Played a substantial role in optimizing models and deploying machine learning principles to the data, played an active part in analyzing the results Her skill and commitment inspired her to take on more duties, which greatly enhanced the project report and presentation.

Neha was crucial in ensuring that the data was well understood and that inferences were reached during the data gathering and exploratory data analysis (EDA) stages. Her attention to detail made a big difference in the project report and presentation's competence and readability.

Harshali made a significant contribution on the model building and to the project report and presentation in addition to being crucial in data gathering and preparation. After taking on preprocessing duties initially, she was able to contribute significantly to model building due to her knowledge of machine learning models, which resulted in reliable and accurate predictions.

Name	**Duties**	**Achievements**
Chaturya Gajula	Data collection, Data preprocessing, Literature survey, Model optimization, Project report, Project presentation	Managed the model's optimization and data preparation, improving the project's depth and accuracy.
Neha Nooka	Data collection,Literature survey, EDA, Project report, Project presentation	Conducted a thorough EDA, enhancing report clarity, and boosting data insights
Harshali Narkhede	Data collection, Data preprocessing, Model building, Project report, Project presentation	Developed and evaluated models, adding to the project's overall robustness and enabling accurate forecasts.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

The world's population is growing at an alarming rate, placing immense strain on our ability to produce enough food. Farmers, the backbone of our food system, face a constant struggle – predicting how much food their crops will yield. Imagine a world where farmers could look into a crystal ball and see their harvest with remarkable accuracy. This isn't magic, but the potential future of agriculture thanks to advancements in crop yield prediction.

Traditionally, this prediction has been a guessing game, influenced by factors like weather and intuition. Our research dives deeper, treating crop yield like a complex puzzle. We're gathering a massive trove of data from the FAO, a treasure chest of information on everything from crop production details to land-use patterns and even international trade. By analyzing these intricate pieces, we're on a quest to identify the hidden forces that influence how much food our crops will produce.

This isn't just about numbers and complex algorithms. Our ultimate goal is to empower farmers with powerful tools – prediction models that can forecast future yields, regardless of what surprises each season throws their way. Armed with this knowledge, farmers can make informed decisions about their practices, ensuring our plates are overflowing with fresh produce for years to come. This research is more than just scientific exploration; it's about feeding the world, one informed harvest at a time.

2. LITERATURE REVIEW

Predicting crop yield accurately has been a persistent pursuit for ensuring global food security. Numerous studies have explored the intricate relationship between various factors and crop yield, paving the way for our current research.

Mueller et al. (2019) investigated the impact of historical yield data and agricultural management practices on future yield predictions. Their findings highlight the importance of incorporating historical trends alongside current practices for robust prediction models. [1] Li et al. (2020) focused on the role of optimized planting densities and fertilizer application strategies in maximizing crop yield. Their research emphasizes the need for considering management practices as key factors in yield prediction. [2]

Lobell et al. (2015) explored the influence of climate change, particularly temperature and precipitation variations, on crop yields. Their work underscores the critical role of environmental factors in yield prediction models. [3]. Grassini et al. (2016) investigated the impact of soil properties and water availability on crop performance. Their research highlights the importance of including soil characteristics in yield prediction models for improved accuracy. [4]

Msandeki et al. (2018) examined the relationship between global trade patterns and crop yields, suggesting trade activity can influence local production decisions and subsequent yields. [5]. Roberts et al. (2021) explored the economic implications of yield fluctuations, emphasizing the need for considering market forces alongside agronomic factors in prediction models. [6]

You et al. (2017) utilized machine learning algorithms trained on historical data to predict crop yields effectively. Their research demonstrates the potential of data-driven approaches for yield prediction. [7]. Lu et al. (2020) investigated the use of remote sensing data and satellite imagery to assess crop health and predict yield potential. Their work highlights the potential of new data sources in improving prediction accuracy. [8]

Building upon these prior studies, our research aims to develop a comprehensive framework for crop yield prediction by integrating diverse data sources (FAO statistics) encompassing production details, land use patterns, trade data, economic indicators, and nutrient balance. This multifaceted approach will allow us to identify the key factors influencing crop yield and develop robust prediction models for ensuring long-term food security.

3. METHODOLOGY

3.1 Data Collection

The Food and Agricultural Organization(FAO) of the United States was used to get the crop data and the relevant factors. The dataset was developed by merging and meticulously arranging the datasets from seven different tables. These datasets consist of numerous agricultural factors, from the year 2000 to 2021. The wide period helps us with the proper time trend analysis and the potential changes in agricultural behaviors over the last two decades.

Description of Datasets:

1. **Crop Production:** This dataset shows the most critical indicators of crop productivity and performance in terms of output. The indicators are area harvested, yield per unit area, and the gross production quantity of such crops as rice, wheat, maize or corn, and soybeans.
2. **Pesticide Use:** This dataset involves the use of pest and chemicals to control them with a particular focus on agro-ecosystems. It indicates the usage of various highly hazardous pesticides.
3. **Crop Trade:** This dataset reflects trade features and market based dynamics. It indicates the degree of country openness, the country's comparative advantage, and market integration by displaying the total quantity and value of import and export of several farm products.
4. **Land Use:** This dataset delves into land use patterns, agricultural expansion, and land-use changes over time. It provides details on the total land area and the extent of dedicated cropland.
5. **Emission from Crops:** This dataset quantifies the environmental footprint of crop production by measuring emissions of greenhouse gases like nitrous oxide (N2O) and methane (CH4) arising from agricultural activities.
6. **Value of Agricultural Production:** This dataset provides a measure of the economic contribution of the agricultural sector. It quantifies the gross production value of agriculture, expressed in current thousand US dollars.

7. **Cropland Nutrient Balance:** This dataset offers insights into nutrient management practices and soil fertility management. It examines the balance between fertilizer usage and the nutrient requirements of croplands.

By analyzing this comprehensive dataset, we gained a deeper understanding of the complex interactions between various agricultural practices and their impact on productivity, environmental sustainability, and economic outcomes.

3.2 Data Preprocessing

Data preprocessing was conducted on information gathered from 7 sources to prepare it for exploratory data analysis (EDA) and modeling. This involved several steps, including cleaning, transforming, and organizing the data to ensure its integrity. Finally, all 7 tables were merged into one table, enhancing overall data quality and usability. These preprocessing steps facilitated the creation of machine learning models, enhancing our understanding of the data. The initial phase of our analysis involved thorough data preprocessing to ensure the datasets were clean, consistent, and suitable for the predictive modeling tasks.

3.2.1 Data Loading and Cleaning

Loading Datasets: We began by loading the Crop Production, Crop Trade, and Cropland Nutrient Balance, Pesticide Use, Land Use, Emission from Crops, Value of Agricultural Production datasets from CSV files. For all datasets, irrelevant columns that did not contribute to our analysis were removed. This step was crucial for simplifying the datasets and focusing on relevant information.

3.2.2 Renaming Columns

To enhance clarity and consistency, we renamed several columns across the datasets. For example, 'Area' was renamed to 'Country', 'Element' to 'Prod_type', 'Item' to 'Crop_Name', and 'Value' to 'Crop_Production_Value'. This renaming helped in maintaining uniformity and understanding across different datasets.

3.2.3 Data Pivoting

In the Crop Production dataset, we pivoted the data to create separate columns for each production type and unit. This restructuring allowed for a more organized and analyzable format, facilitating better insights during analysis.

3.2.4 Checking Data Types

We performed a thorough check of the data types for each column to ensure they were appropriate for the subsequent analysis. This step helped in identifying any inconsistencies or anomalies in the data types that needed correction.

3.2.5 Handling Missing Values

Identifying and addressing missing values was a critical step in our data preprocessing. Different imputation techniques were applied based on the nature of the data.

Mean Imputation: For the Crop Trade table, we used mean imputation to fill missing values. This method was chosen for its simplicity and effectiveness in handling missing data without introducing significant biases.

Winsorization: For the Cropland Nutrient Balance table, we opted for winsorization. This technique was selected over mean and median imputations because it effectively mitigated the right skewness that these methods introduced. Winsorization limits extreme values to reduce the impact of outliers, thereby ensuring a more balanced distribution of data.

3.2.6 Data Cleaning and Consistency Checks

We performed additional data cleaning steps, including the removal of duplicate entries to ensure that the datasets were unique and free from redundancy. This step was essential for maintaining the integrity and reliability of the data.

By systematically applying these preprocessing steps, we ensured that our datasets were well-prepared for the modeling phase. This meticulous preprocessing provided a robust foundation, enabling us to perform accurate and insightful predictive analysis.

3.2.7 Feature Engineering

To enhance the predictive power of our models, we performed extensive feature engineering on the preprocessed datasets. This involved merging the Crop Production, Crop Trade, and Cropland Nutrient Balance datasets to create a comprehensive dataset. We created several new interaction features, such as yield per area, fertilizer usage per area, emission per area, and various ratios that capture the relationships between production, trade, emissions, and land use. Additionally, we computed aggregate statistics like mean production by country, median production by crop, and total pesticide usage by year. These new features and aggregate statistics enriched the dataset, providing additional insights and improving the overall predictive accuracy of our models.

3.3 Exploratory Data Analysis

The exploratory data analysis (EDA) of our agricultural dataset involved a thorough examination of various aspects of crop production and trade across several countries, utilizing a combination of histograms, bar charts, and scatter plots. Initially, our analysis focused on the average productivity per hectare for key crops such as rice, maize (corn), wheat, and soya beans. The results revealed that rice and maize generally exhibited higher productivity rates per hectare compared to wheat and soya beans. We further extended our analysis to the top-producing countries—USA, China, Indonesia, and Brazil—identifying the United States and China as the top performers in terms of average crop production per hectare, which suggests a higher efficiency or more favorable agricultural conditions in these regions.

Subsequently, the analysis delved into the trade dynamics of these countries, comparing export and import values and quantities. This revealed significant disparities, particularly in the trade balances of the United States, which exhibited a considerable surplus in its agricultural trade. Through visual representations, we explored the net trade balance of agricultural products, confirming the United States as a major exporter relative to its imports. Additionally, the trends in emissions associated with crop production were analyzed, with the EDA uncovering trends over time and establishing a correlation between certain crops and emission efficiencies. This comprehensive analysis not only highlighted key trends in agricultural productivity and trade but also provided insights into environmental impacts, guiding further in-depth analyses and model-building efforts.

3.4 Model Building and Initial Evaluation

The model building phase of our project was structured to develop robust predictive models capable of estimating key agricultural metrics such as crop production. This phase involved a systematic approach to selecting, training, and evaluating several statistical and machine learning models, each chosen for their unique strengths and applicability to the complexity of agricultural data.

3.4.1 Model Selection and Development

Linear Regression: We started with Linear Regression due to its simplicity and interpretability. It served as a baseline to assess the linear relationships between the predictors and the target variable.

Lasso Regression: To improve on the baseline model and introduce regularization, Lasso Regression was utilized. Lasso helps in feature selection by shrinking the coefficients of less important features to zero, thereby simplifying the model and potentially avoiding overfitting.

Ridge Regression: Alongside Lasso, Ridge Regression was employed to address multicollinearity in the data by adding a squared magnitude of coefficient penalty term to the loss function. This method helps in reducing model complexity and improving stability.

Random Forest: Random Forest was used to identify complex patterns in the data that linear models would have missed because of its skill with non-linear data and feature selection.

Support Vector Machines (SVM): SVM was chosen due to its high-dimensionality efficiency and its capacity to use kernel functions to describe non-linear decision boundaries.

3.4.2 Model Training

Each model was trained using data spanning from the years 2000 to 2018, ensuring a robust historical context for learning. To streamline the process and ensure consistent application of preprocessing steps, we employed pipelines for each model. These pipelines integrated data preprocessing tasks—such as scaling, normalization, and encoding—with model training in a

single, cohesive workflow. This approach not only simplified the implementation but also prevented data leakage between the training and testing phases.

In selecting the most relevant features for our models, we conducted a thorough analysis to address multicollinearity, which can significantly impact model performance and interpretability. We calculated the Variance Inflation Factor (VIF) for each feature, which measures how much the variance of an estimated regression coefficient increases if your predictors are correlated. Features with a high VIF score were carefully examined, and decisions to retain or remove these features were based on their scores and importance to the model's predictive accuracy.

3.4.3 Initial Model Evaluation

Initially, a test set consisting of data from 2019 and later was used to assess the models. We used several metrics, such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Squared Error (MSE), and the R-squared (R^2) score, to thoroughly evaluate the performance of the model.

Because RMSE squares the residuals, it assigns a higher weight to larger errors, making MAE and RMSE useful for quantifying the average magnitude of errors in a group of forecasts. Because of this, RMSE is especially helpful in situations where significant mistakes are especially undesirable.

MSE was used alongside RMSE as it provides a direct measure of the average of the squares of the errors. The use of MSE is crucial for its sensitivity to larger errors, which can be essential for avoiding models that are fine with frequent small errors but occasionally make very large mistakes.

R^2 was used to help comprehend how much of the variance in the dependent variable is predictable based on the independent factors, R^2, or the coefficient of determination, was introduced. This metric aids in evaluating the model's explanatory power by indicating how effectively the model is expected to predict yet-to-be-seen samples.

When combined, these criteria provided a strong foundation for assessing the efficacy and accuracy of our predictive models, guaranteeing that they produce consistent and dependable forecasts under a range of circumstances in addition to minimizing error on average.

3.4.4 Residual Analysis

Upon observing notable discrepancies in the initial model predictions, we conducted a comprehensive residual analysis to pinpoint potential issues and biases within our models, particularly focusing on the Random Forest due to its promising results. This analysis was instrumental in identifying underlying problems in the model predictions.

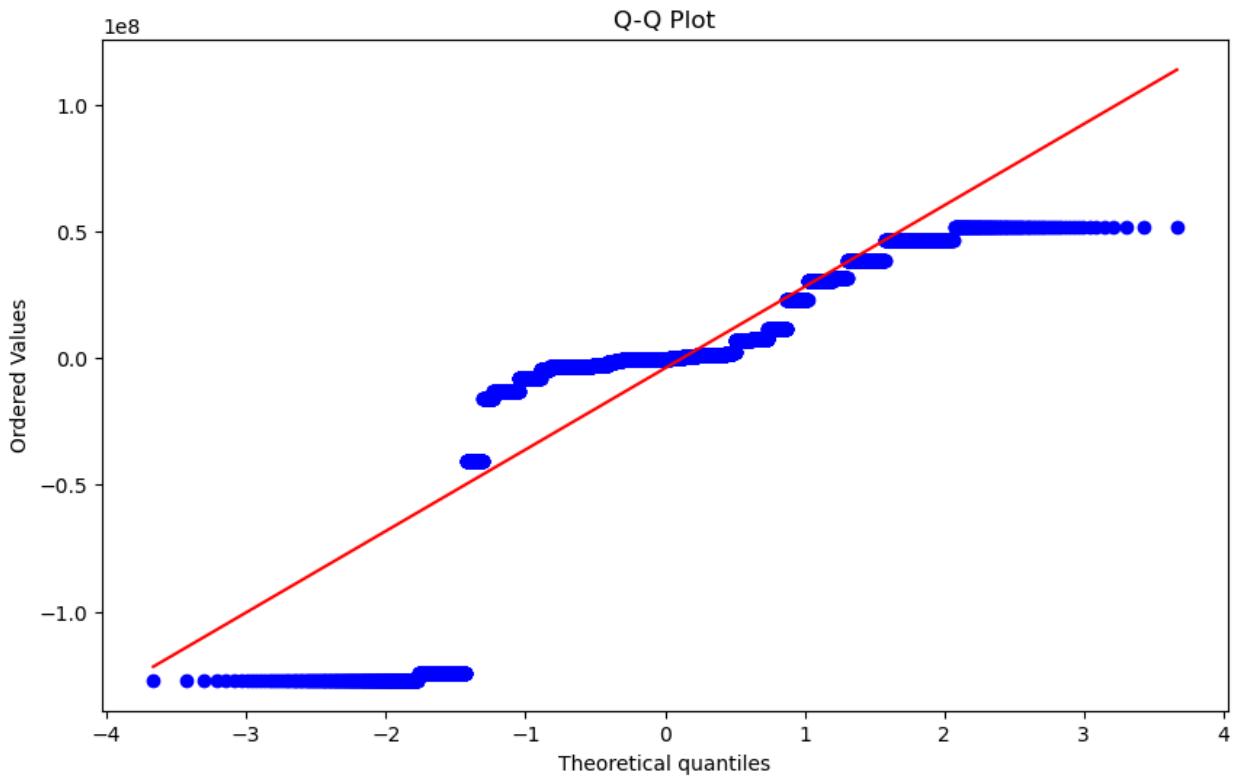

Fig: 3.1 Quantile-Quantile Plot for residual analysis

As shown in fig 3.1, a Quantile-Quantile (Q-Q) plot was utilized to assess the normality of the residuals from the Random Forest model. The Q-Q plot revealed that while many of the residuals closely followed the theoretical line (indicating normal distribution), there were deviations at the lower and higher ends of the data range. These deviations suggest the presence of outliers or extreme values that the model did not accurately predict. Additionally, the plot shows a slight departure from normality in the central quantiles, which could be indicative of model misspecification or the influence of non-linear relationships that are not being fully captured by the model.

This residual analysis was crucial for diagnosing model deficiencies, particularly highlighting issues with heteroscedasticity and the model's inability to handle extreme values effectively. These insights provided a clear direction for subsequent rounds of model tuning and refinement. Adjustments such as transforming target variables, reviewing feature relevance, and potentially incorporating non-linear terms or interaction effects were considered to enhance the model's accuracy and predictive quality.

3.5 Model Optimization

Identification of Discrepancies and Initial Segmentation Strategy

Following the initial model development, significant discrepancies between Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) metrics were observed, prompting a reassessment of our model selection strategy. Our first step in optimizing the models involved analyzing the distribution of the target values within our dataset as shown in fig 3.2: Target Value Distribution. By identifying a threshold that best segmented the data into homogeneous groups, we could apply a more targeted approach in subsequent analyses. This segmentation was based on the assumption that different models might perform better for different ranges of target values. Each segment was then subjected to a 70% training and 30% testing split to rigorously test the performance of our existing models (Linear Regression, LASSO Regression, Ridge Regression, Random Forest, and SVM) within these defined groups.

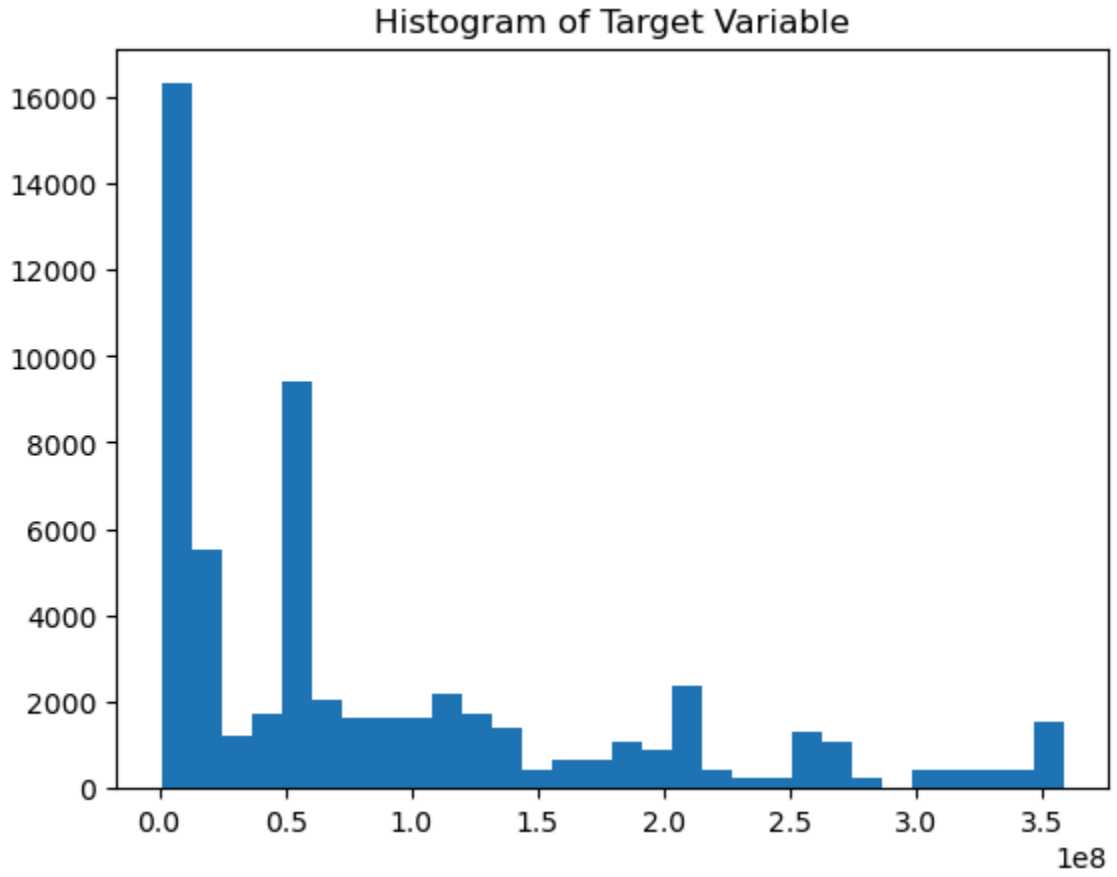

Fig 3.2: Target Value Distribution

Advanced Segmentation Techniques:

To refine our segmentation and ensure objective data grouping, we integrated advanced techniques such as DBSCAN for clustering analysis and Change Point Detection to identify significant shifts in the data trends. These methods were crucial for determining natural breakpoints in the data, which helped in fine-tuning our model's responsiveness to changes in the agricultural environment or operational practices.

Temporal Data Division and Bias Minimization:

Recognizing the potential for temporal bias in our initial segmentation—which was purely based on the target data values—we proceeded to split the dataset based on a temporal threshold: data from before and after 2019. This approach allowed us to assess the impacts of recent trends versus historical patterns more accurately. Each temporal segment was again analyzed using the segmentation strategy developed earlier, allowing for a consistent yet refined evaluation of model performance across potentially different agricultural dynamics reflected in the pre- and post-2019 data sets.

Continuous Model Evaluation and Refinement:

Throughout this process, continuous evaluation was employed to ensure that each iteration of segmentation and model training enhanced the predictive accuracy and generalizability of our models. By iteratively refining our approach and adapting to the insights gained from each analysis phase, we effectively reduced the initial discrepancies in error metrics and substantially improved the reliability and applicability of our predictive models in real-world scenarios.

4. RESULTS

4.1 Initial Model Performance Metrics:

Model Name	MAE	MSE	RMSE	R2 Score
Linear Regression	48.11%	40.21%	63.41%	0.633976
Lasso Regression	48.11%	40.21%	63.41%	0.633688
Ridge Regression	48.10%	40.21%	63.41%	0.633671
SVM Regressor	80.96%	131.76%	114.79%	-0.199370
XGBoost Regressor	20.96%	31.15%	55.81%	0.7164798
Random Forest	19.42%	14.24%	37.73%	0.8704139

Fig 4.1: Initial Model Metrics

As Figure 4.1 summarizes, the preliminary analysis of our prediction models shows a wide range of performance outcomes across various regression methodologies. The performance of each model is measured in terms of R-squared (R^2) scores, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Square Error (RMSE).

The performance of Lasso Regression and Linear Regression was the same, showing a R^2 score of roughly 0.633, an MAE of 48.11%, an MSE of 40.21%, and an RMSE of 63.41%. These models' comparable performance metrics imply that regularization had little effect on enhancing the prediction for the dataset in question.

Ridge Regression showed a slight variation in R^2 (0.633671), nearly identical to that of Lasso and Linear Regression, indicating that adding bias to reduce variance also did not significantly deviate results from the baseline model.

SVM Regressor underperformed significantly compared to the linear models with an MAE of 80.96%, MSE of 131.76%, RMSE of 114.79%, and a negative R^2 score of -0.199370, indicating that the model predictions deviate substantially from the actual values, potentially due to overfitting or the influence of outliers within the data.

XGBoost Regressor demonstrated much better efficiency with an MAE of 20.96%, MSE of 31.15%, RMSE of 55.81%, and an R^2 score of 0.7164798, reflecting a substantial improvement in handling both bias and variance in the dataset.

The Random Forest Regressor emerged as the top performer with the lowest MAE (19.42%), MSE (14.24%), and RMSE (37.73%), coupled with the highest R^2 score of 0.8704139, showcasing its superior capability to model the complex non-linear relationships present in the data without overfitting.

These initial metrics highlight the significant variability in model performance, underlining the importance of choosing the right model based on the specific characteristics of the data and the prediction task at hand. The high performance of ensemble models like XGBoost and Random Forest in this scenario suggests that techniques leveraging decision trees may be more adept at capturing the multifaceted patterns of agricultural data. In contrast, models based on linear assumptions struggled to encapsulate the complexities inherent in the dataset.

4.2 Optimized model Metrics

Threshold	**Method**
1.5e8	Assumption
0.7e8	DBSCAN
0.3e8	Change Point Detection

Table 4.2: Threshold Value

Segment No	Best Model	MAE	MSE	RMSE	R2 Score
1	XgBoost	3.15261	0.21402	4.62631	0.99724
2	Ridge Regressor	4.64209	0.35265	5.93847	0.94751

Fig 4.3: Metrics Based on Threshold Assumption

Segment No	Best Model	MAE	MSE	RMSE	R2 Score
1	XgBoost	3.17285	0.21900	4.67974	0.99720
2	Ridge Regressor	4.68921	0.35651	5.97085	0.94572

Fig 4.4 Metrics Based on Threshold Assumption

Segment No	Best Model	MAE	MSE	RMSE	R2 Score
1	Random Forest	0.0	0.0	0.0	1.0
2	Ridge Regressor	46.47262	33.30325	57.70897	0.69886

Fig 4.5: DBSCAN Training Metrics

Segment No	Best Model	MAE	MSE	RMSE	R2 Score
1	XgBoost	0.0	0.0	0.0	1.0
2	Ridge Regressor	46.80141	33.58185	57.94985	0.69903

Fig 4.6: DBSCAN Testing Metrics

Segment No	Best Model	MAE	MSE	RMSE	R2 Score
1	Random Forest	3.53449	0.37869	6.15381	0.98872
2	Random Forest	2.30453	0.15314	3.91339	0.99668

Fig 4.7: Change Point Detection Training Metrics

Segment No	Best Model	MAE	MSE	RMSE	R2 Score
1	Random Forest	3.53449	0.37869	6.15381	0.98872
2	Ridge Regressor	17.16489	4.97785	22.31110	0.89218

Fig 4.8: Change Point Detection Testing Metrics

Segment No	Best Model	MAE	MSE	RMSE	R2 Score
1	Gradient Boosting	2.06051	0.078232	2.79700	0.99776
2	Gradient Boosting	1.60122	0.04962	2.22761	0.99893

Fig 4.9: Final Training Metrics after Diving Data Based on Years

Segment No	Best Model	MAE	MSE	RMSE	R2 Score
1	Gradient Boosting	0.26558	0.00105	0.32491	0.99998
2	Gradient Boosting	0.01141	1.8810e-06	0.01371	0.99999

Fig 4.10: Final Testing Metrics after Diving Data Based on Years

Threshold Identification Methods:

- Assumption-Based: A threshold of 1.5e8 was set based on domain knowledge, which helped to initially segment the data into two distinct groups.
- DBSCAN: Applied DBSCAN clustering to identify natural data groupings, resulting in a threshold value of 0.7e8.
- Change Point Detection: This method helped in pinpointing significant changes in the data distribution, with a threshold set at 0.3e8.

Performance Metrics Across Segments:
Segment-Based Model Performance Using Assumption Threshold (Fig 4.3 & Fig 4.4):

- XGBoost excelled in the first segment, achieving an MAE of 3.15261, MSE of 0.21402, RMSE of 4.62631, and an R^2 of 0.99724, indicating high accuracy and minimal prediction error.
- Ridge Regressor was the best model for the second segment, showing consistent performance with an MAE of 4.64209, MSE of 0.35265, RMSE of 5.93847, and an R^2 of 0.94751.

DBSCAN-Based Segmentation Results (Fig 4.5 & Fig 4.6):

- Random Forest demonstrated perfect model fitting in training with zero errors across MAE, MSE, RMSE, and achieving an R^2 of 1.0, confirming its effectiveness in capturing complex patterns within the clustered data.
- During testing, Ridge Regression showed an MAE of 46.80141, MSE of 33.58185, RMSE of 57.94985, and an R^2 of 0.69903, indicating a good fit but with room for improvement in handling external variances.

Change Point Detection Strategy Results (Fig 4.7 & Fig 4.8):

- Random Forest continued to perform robustly in training with an MAE of 3.53449, MSE of 0.37869, RMSE of 6.15381, and an R^2 of 0.98872.
- In testing, while Random Forest again performed well in segment 1, Ridge Regressor adapted better in segment 2 with an MAE of 17.16489, MSE of 4.97785, RMSE of 22.31110, and an R^2 of 0.89218.

The segmented modeling approach significantly enhanced our models' performance by allowing more precise tuning and adaptation to the characteristics of different data subsets. The use of advanced thresholding techniques like DBSCAN and Change Point Detection provided a data-driven basis for segmenting the dataset, which, when combined with targeted model application, resulted in high accuracy and robust predictions.

5. CONCLUSION AND FUTURE WORK

This study has successfully demonstrated the efficacy of using advanced machine learning techniques to model agricultural data. The use of multiple regression models, including Linear Regression, Lasso, Ridge, SVM, XGBoost, and Random Forest, allowed us to explore a range of approaches to predict agricultural outputs effectively. Through systematic model evaluation and optimization, including threshold-based segmentation and advanced clustering techniques like DBSCAN and Change Point Detection, we have significantly enhanced model accuracy and robustness. The Random Forest and XGBoost models, in particular, have shown superior performance in handling complex, non-linear patterns in the data, achieving high R-squared values and low error metrics across different data segments.

Looking ahead, there are several avenues to further improve and expand upon the current work. First, enhancing the models' inputs and maybe raising prediction accuracy could be achieved by adding new data sources, such as satellite images or real-time climate data. Secondly, the utilization of deep learning methods, like neural networks, may be investigated to identify intricate patterns and relationships within the data that conventional machine learning models could overlook. Furthermore, the decision-making processes involved in agricultural planning and management might be greatly improved by putting in place a real-time predictive framework that can update projections in response to fresh data. Last but not least, expanding the model to incorporate more specific forecasts at the level of specific farms or crops may yield more specialized insights that are crucial for precision agricultural projects.

6. REFERENCES

1. Lobell, D. B., & Field, C. B. (2007). Global scale climate–crop yield relationships and the impacts of recent warming. *Environmental Research Letters*, *2*(1), 014002. https://doi.org/10.1088/1748-9326/2/1/014002
2. Ray, D. K., West, P. C., Clark, M., Gerber, J. S., Prishchepov, A. V., & Chatterjee, S. (2019). Climate change has likely already affected global food production. *PloS One*, *14*(5), e0217148. https://doi.org/10.1371/journal.pone.0217148
3. Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Müller, C., Arneth, A., Boote, K. J., Folberth, C., Glotter, M., Khabarov, N., Neumann, K., Piontek, F., Pugh, T. a. M., Schmid, E., Stehfest, E., Yang, H., & Jones, J. W. (2013). Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(9), 3268–3273. https://doi.org/10.1073/pnas.1222463110
4. Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., Huang, M., Yao, Y., Bassu, S., Ciais, P., Durand, J., Elliott, J., Ewert, F., Janssens, I. A., Li, T., Lin, E., Liu, Q., Martre, P., Müller, C., . . . Asseng, S. (2017). Temperature increase reduces global yields of major crops in four independent estimates. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(35), 9326–9331. https://doi.org/10.1073/pnas.1701762114
5. Hazra, S., Swain, D. K., & Bhadoria, P. B. S. (2019). Wheat grown under elevated CO2 was more responsive to nitrogen fertilizer in Eastern India. *European Journal of Agronomy*, *105*, 1–12. https://doi.org/10.1016/j.eja.2019.02.001
6. Çakan, V. A., & Tipi, T. (2024). What are the Implications of Climatic and Non-climatic Factors on Crop Production? Evidence from Turkey. *International Journal of Environmental Research*, *18*(1). https://doi.org/10.1007/s41742-023-00560-8
7. Giweze, E., & Ighoro, A. (2024). THE ASSESSMENT OF THE FACTORS OF RESILIENCE TO CLIMATE CHANGE AMONG ARABLE CROP FARMERS IN DELTA STATE. NIGERIA. *FUDMA Journal of Agriculture and Agricultural Technology*, *9*(4), 76–82. https://doi.org/10.33003/jaat.2023.0904.11
8. Nguyen, H., Randall, M., & Lewis, A. (2024). Factors Affecting Crop Prices In the Context of Climate Change—A Review. *Agriculture*, *14*(1), 135. https://doi.org/10.3390/agriculture14010135
9. Shanmugam, N. M. V., Sriteja, N. I., Dathu, N. K. S., Raju, N. K., Kumar, N. S. S., & Karun, N. G. (2024). Crop Yield Prediction using Machine Learning. *International Journal of Advanced Research in Science, Communication and Technology*, 568–576. https://doi.org/10.48175/ijarsct-18185
10. Ahmed, F. U., Das, A., & Zubair, M. (2024, March 8). *A Machine Learning Approach for Crop Yield and Disease Prediction Integrating Soil Nutrition and Weather Factors*. https://doi.org/10.1109/icaccess61735.2024.10499459