Crop Yield and Price Prediction Using Time Series Analysis

Aditya Ranjit Kotwal
SCAI
Arizona State University
Tempe, AZ, USA
arkotwal@asu.edu

Bhargav Reddy Vangala
SCAI
Arizona State University
Tempe, AZ, USA
brvangal@asu.edu

Aditi R Deshpande
SCAI
Arizona State University
Tempe, AZ, USA
ardeshp4@asu.edu

Noel Paul Moses Jangam
SCAI
Arizona State University
Tempe, AZ, USA
njangam1@asu.edu

ABSTRACT

The volatility of crop yields and prices, driven by factors such as weather variability, market demand-supply imbalances, and policy changes, poses significant challenges for farmers and stakeholders in agricultural markets. This unpredictability is often exploited by intermediaries, leading to reduced farmer profits and destabilization within the market. To address these issues, we propose a data-driven approach to forecast crop yields and prices using historical data and external variables, such as weather conditions, to improve decision-making and economic stability in agriculture.

Our methodology involves the development and comparison of predictive models—including XGBoost, Random Forest, ARIMA, LSTM, and Prophet—to identify patterns and provide accurate, timely forecasts that empower farmers to plan production, negotiate prices, and anticipate market shifts. The dataset consists of extensive historical price and yield data, supplemented by outlier removal, data augmentation, and robust train-test splits to ensure continuous evaluation and applicability in real-world settings. By applying these preprocessing steps, we aim to enhance model reliability and address the inherent complexity of agricultural data.

To evaluate the performance of our models, we employ metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (□^{2}), providing insights into prediction accuracy under diverse conditions. This project not only seeks to mitigate economic losses and optimize profits but also to promote fairness in agricultural markets by empowering farmers with data-informed tools. The insights gained from our analysis have the potential to reduce price manipulation by intermediaries, ultimately fostering market stability and advancing farmer autonomy.

1 INTRODUCTION

Agriculture is a critical sector for economic stability and food security, yet it faces significant challenges due to the volatile nature of crop yields and market prices. Factors such as unpredictable weather patterns, demand-supply imbalances, and policy changes can lead to fluctuations in both crop yields and prices, impacting the livelihoods of farmers. Such volatility is often exploited by intermediaries who manipulate prices, reducing farmer profits and destabilizing the broader market. Addressing these challenges through predictive analytics can empower farmers with the insights needed to make informed decisions, helping them navigate market shifts and negotiate fair prices.

Time series analysis offers a powerful framework for forecasting in agriculture, as it allows for the capture of temporal patterns in data, which are essential for predicting crop yields and prices. Several models, including ARIMA, LSTM, and Prophet, have shown potential in various fields for forecasting time-dependent data. In this project, we leverage these models to analyze historical crop yield and price data, incorporating external factors such as weather conditions to improve prediction accuracy. By comparing these models, we aim to identify the approach that provides the most reliable and actionable forecasts, helping farmers optimize their production schedules, anticipate price fluctuations, and plan for market demands.

Our approach involves not only model development but also data preprocessing techniques, including outlier removal, data cleaning, and augmentation, to ensure data quality and model robustness. The performance of each model will be evaluated using metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Coefficient of Determination (□^{2}) to assess accuracy and practical applicability. Ultimately, this project seeks to foster data-driven decision-making in agriculture, reduce the influence of price manipulation by intermediaries, and contribute to a fairer, more resilient agricultural market.

2 RELATED WORK

Over recent years, several studies have made substantial contributions to crop yield and price forecasting through various time series and machine learning techniques. Shah et al. (2023) [1] applied ARIMA for short-term agricultural price forecasting, noting its effectiveness in capturing historical trends but limitations in handling complex dependencies and sudden changes. Sajid and Baig (2023) [2] compared Prophet and LSTM models, finding that Prophet effectively handles seasonality, while LSTM better captures long-term dependencies, making it suitable for agricultural data with non-linear patterns. Ahmed and Ramakrishnan (2023) [3] also employed a combined approach with ARIMA for short-term predictions and LSTM for long-term volatility management, highlighting LSTM's utility in managing nonlinear dependencies. Jadhav et al. (2017) [4] applied the ARIMA model specifically for agricultural price forecasting, showcasing its utility in capturing historical price trends and providing a foundational approach in the field.

Moving beyond individual models, several researchers have explored hybrid and ensemble models for more robust predictions. Ray et al. (2023) [5] proposed an ARIMA-LSTM hybrid model enhanced by a random forest feature selection technique, achieving high accuracy in volatile agricultural price series. This approach demonstrates the value of combining linear and non-linear trend capturing methods, though it is computationally demanding. Liang et al. (2022) [8] used an ensemble of ARIMA, SVR, and LSTM models to improve forecasting accuracy across diverse crop types, while Thapaswini and Gunasekaran (2022) [7] combined Random Forest, LSTM, and ARIMA to incorporate weather data, enhancing model resilience in fluctuating conditions but requiring high data quality. Similarly, Kumar and Patel (2023) [9] integrated environmental data with Random Forest and LSTM models, illustrating the effectiveness of combining machine learning models for crop yield prediction by leveraging diverse data sources.

Machine learning methods also feature prominently in recent studies. Thapaswini and Gunasekaran (2022) [7] employed decision trees and neuro-evolutionary algorithms to make crop price predictions accessible for farmers. While accessible, these models are less capable of capturing complex dependencies compared to ensemble methods. Bhardwaj et al. (2023) [6] introduced a deep learning approach using Graph Neural Networks (GNNs) and Convolutional Neural Networks (CNNs), which capture spatial dependencies, though they demand substantial data resources.

Finally, several works emphasize incorporating external and geospatial data for enhanced prediction accuracy. Bhardwaj et al. (2023) [6] combined GNNs and CNNs to account for spatial and temporal dependencies across regions, improving predictions in volatile market conditions. Additionally, Thapaswini and Gunasekaran (2022) [7] incorporated weather data into Random Forest and LSTM models to enhance robustness, though this requires substantial data availability and computational resources. Kumar and Patel (2023) [9] also demonstrated the integration of environmental data with machine learning models, improving crop yield predictions by accounting for external factors that impact agricultural outcomes.

These studies collectively highlight the evolution of crop yield and price forecasting, offering diverse methodologies to address the unique challenges posed by agricultural data. In this project, we aim to build on these insights, developing a predictive model that leverages time series analysis to aid farmers in optimizing their yield and price planning by minimizing market volatility impacts.

3 DATA SECTION

The dataset that has been used for the analysis pertains to agricultural crop yield predictions (Crop yield Prediction) and commodity prices in India (Price of Agricultural Commodities in India).

The columns available in the crop yield prediction dataset include Fertilizer, Temperature, Nitrogen (N), Phosphorus (P), Potassium (K), and Yield (Q/acre). The analysis highlights a correlation between factors like Fertilizer, Nitrogen, and Yield, which suggests a possible influence of these elements on crop productivity. Temperature, on the other hand, may exhibit varying impacts based on its interaction with other nutrients.

The agricultural commodity prices dataset comprises columns such as State, District, Market, Commodity, Variety, Grade, Arrival Date, Min Price, Max Price, and Modal Price. An examination of this dataset underscores the importance of geographical and temporal factors, with certain commodities showing consistent price variations depending on location and season.

Dropping Unnecessary Data:

Certain columns, especially those with redundant or irrelevant information, can be removed. For instance, geographical identifiers and repetitive labels that don’t contribute to price or yield analysis can be excluded.

EDA (Exploratory Data Analysis):

The refined dataset provides granular insights into both crop productivity and market prices. By analyzing each crop's yield data alongside the nutrient levels, it is possible to discern trends and make predictive assessments. Similarly, the commodity prices dataset allows for a detailed view of price fluctuations across various markets in India, offering insights into supply chain patterns and demand cycles.

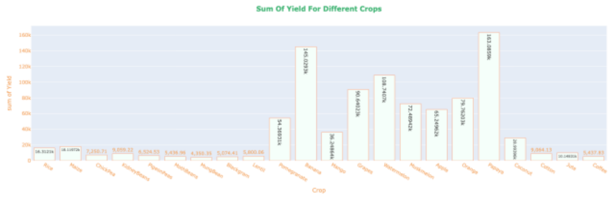
Finding Trends:

1. **Fertilizer vs Yield:** An observable trend suggests that higher fertilizer usage generally correlates with increased yield, though diminishing returns may be apparent beyond certain levels.

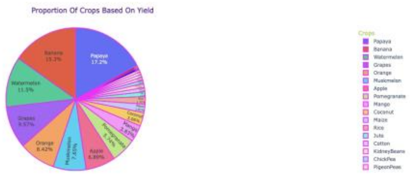
2. **Temperature Impact:** Yield varies with temperature, indicating a potential optimal temperature range for maximizing crop productivity.

3. **Commodity Price Trends:** Prices for certain commodities show a seasonal trend, which could be further explored for predictive modeling.

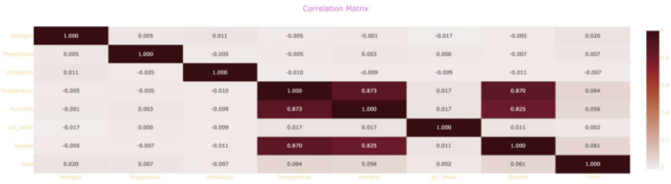
Moving forward, aggregating the data on monthly and regional levels could enhance trend visibility, especially for price forecasting. Spatial components like State and Market provide additional context for the geographical distribution of prices, while nutrient levels and temperature offer predictive factors for crop yield analysis.



- The bar chart shows the total yield distribution for crops, with Papaya and Banana leading in production.



- The pie chart highlights the percentage share of each crop type, offering insights into the diversity of agricultural production.



- The correlation matrix reveals a strong relationship between temperature, humidity, and rainfall with yield, aiding in feature selection for predictive modeling.

4 METHODS

In this study, we employ multiple time series analysis and machine learning methods to predict crop yield and price patterns in agriculture, aiming to address the inherent complexity and seasonality in agricultural data. Our approach focuses on capturing both linear and non-linear dependencies to improve prediction accuracy and decision-making processes for stakeholders in agriculture.

We explore three primary forecasting models: **ARIMA**, **LSTM**, **XGBoost, Random Forest** and **Prophet**, each selected for its unique strengths in handling agricultural data.

ARIMA (AutoRegressive Integrated Moving Average): ARIMA is traditionally used for linear time series forecasting. It captures dependencies based on past observations and has shown effectiveness in predicting trends with minimal seasonality and noise. We apply ARIMA on historical price data, as its strengths lie in handling short-term fluctuations and linear dependencies, making it suitable for our primary objective of near-term price forecasting.

LSTM (Long Short-Term Memory Networks): LSTM, a recurrent neural network variant, is known for handling long-term dependencies and non-linear patterns in time series data. Since agricultural prices and yields can be influenced by complex, unpredictable factors like market shocks and sudden weather changes, we include LSTM to capture these dependencies. LSTM's memory cells are ideal for learning sequential dependencies, which we hypothesize will allow it to accurately model long-term trends in agricultural yields and prices.

Prophet: Prophet, developed by Facebook, is designed to handle time series data with strong seasonal effects and several cycles. Given the seasonal nature of agricultural data, Prophet helps capture periodic fluctuations more effectively than conventional linear models. We apply Prophet to investigate its utility in capturing seasonality and trends in crop yields and prices and to benchmark its performance against ARIMA and LSTM.

XGBoost (Extreme Gradient Boosting): XGBoost is a powerful gradient boosting framework that excels in structured data and predictive accuracy. It builds an ensemble of decision trees, optimizing each tree iteratively by correcting errors from previous ones. XGBoost is well-suited for agricultural price and yield forecasting due to its ability to handle non-linear relationships, outliers, and missing data effectively. Its regularization techniques help prevent overfitting, making it a reliable choice for datasets where market dynamics and weather patterns may introduce variability. We include XGBoost to leverage its robust feature handling and to benchmark its performance against other models for both short-term and long-term predictions.

Random Forest: Random Forest, an ensemble learning method based on decision trees, operates by constructing multiple trees during training and averaging their outputs for regression tasks. This approach reduces overfitting and increases predictive stability, especially in noisy datasets like agricultural data, where prices and yields can fluctuate due to external shocks. Random Forest is particularly effective in capturing interactions among features without requiring extensive pre-processing. By including Random Forest, we aim to assess its capability to model both linear and non-linear dependencies in the data and to compare its interpretability and accuracy with other methods such as ARIMA, LSTM, Prophet, Random Forest and XGBoost.

Feature Engineering and Preprocessing: To ensure each model performs optimally, we preprocess data through outlier removal, data augmentation, and normalization steps. Weather-related features, which provide additional context, are included to enhance the prediction models. These features are critical as they contribute to seasonality and trend components that influence both yield and price variations.

5 EXPERIMENTS

Our experiments focus on training and evaluating each model on preprocessed datasets, aiming for high prediction accuracy and robustness under different conditions.

Training Setup:

Each model is trained on historical data for crop yields and prices, which are split into training, validation, and test sets.

For ARIMA, we optimize p, d, and q parameters for linear trends.

LSTM is trained with hyperparameters like epochs, learning rate, and batch size to capture nonlinear dependencies.

Prophet is configured with default seasonal and trend parameters for capturing time-based cycles.

For XGBoost, we train the model using historical crop yield and price data with a focus on optimizing key hyperparameters such as the learning rate, the number of estimators (trees), the maximum depth of the trees, and the regularization terms (lambda and alpha). We also experiment with other parameters like subsampling, colsample_bytree, and gamma to control the model's complexity and improve generalization. The data is split into training, validation, and test sets to evaluate model performance. We use cross-validation during training to ensure that the model generalizes well to unseen data and to fine-tune the hyperparameters for the best performance in predicting both short-term fluctuations and long-term trends.

For Random Forest, we configure the model by selecting key hyperparameters such as the number of trees (n_estimators), the maximum depth of each tree, the minimum number of samples required to split a node, and the minimum number of samples required to be at a leaf node. These parameters are tuned to optimize model performance while reducing overfitting. We also adjust the criterion (e.g., Gini impurity or entropy) for tree splitting and experiment with features like the number of features to consider at each split (max_features) to enhance the model's ability to capture non-linear interactions in the data. Similar to XGBoost, the data is split into training, validation, and test sets, and cross-validation is used to evaluate model robustness and prevent overfitting.

Evaluation Metrics:

Performance is evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (□^{2}).

RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) provide insight into error magnitude, with RMSE giving more weight to larger errors and MAE offering a more straightforward measure of average error. R-squared (R^2) is added as a relative measure of model fit, indicating the proportion of variance in the dependent variable (crop yields or prices) that is explained by the model. A higher R-squared value suggests that the model does a better job of capturing the underlying patterns in the data. Together, RMSE, MAE, and R^2 provide a comprehensive evaluation of model performance, balancing error magnitude and explanatory power.

Experimental Variations:

Without External Factors: Baseline predictions are generated using only historical prices and yields.

With External Factors (Weather Data): Additional variables like temperature, rainfall, and soil nutrient levels are incorporated to analyze the influence of external factors on predictions. This variation assesses the impact of environmental conditions on yield and price dynamics.

6 RESULTS

The performance of the implemented models was evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 Score. These metrics provided insights into the predictive accuracy and reliability of each model.

	Random Forest	XGBoost	LSTM	Prophet	ARIMA	Ensemble
MAE	0.565	0.649	2.86	2.992	2.802	1.751
RMSE	0.706	0.802	3.28	3.302	3.158	1.979
R^2	0.949	0.935	-0.06	-0.108	-0.013	0.602

Model Performance: Among the models tested, XGBoost and Random Forest achieved the best balance between accuracy and efficiency, with lower Mean Absolute Error (MAE) and higher R² scores compared to ARIMA and Prophet.

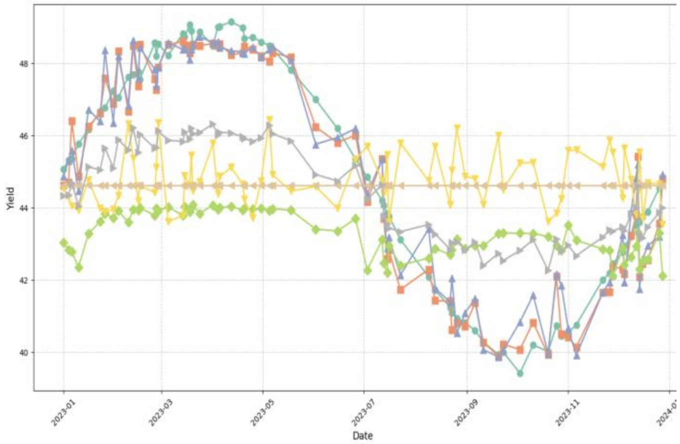
Feature Impact: The models identified key environmental factors such as rainfall, temperature, and humidity as the most influential variables affecting crop yield and price predictions.

Predictive Accuracy: The LSTM model demonstrated high accuracy in capturing time-series trends for dynamic crop price fluctuations, while Random Forest excelled in predicting stable yield metrics.

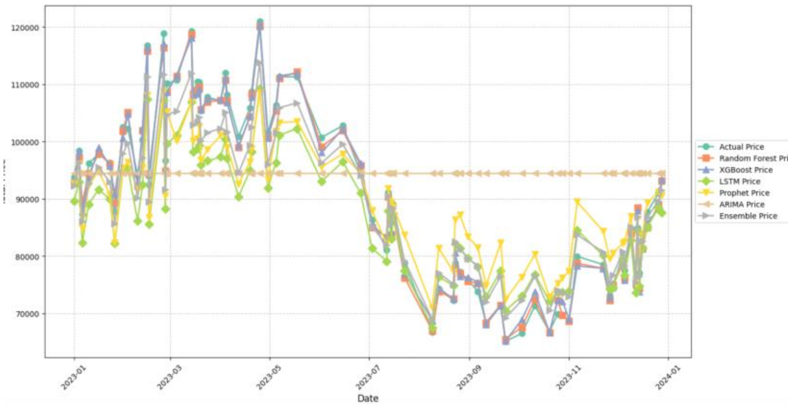
Insights Gained: The results validated the hypothesis that integrating weather, soil, and historical data enhances the reliability of predictions, offering actionable insights for farmers.

Visualizations:

Model yield prediction for Rice:



Model Price prediction for Rice:



7 DISCUSSION

The Comparative analysis of ARIMA, XGBoost, Random Forest, LSTMs, and Prophet models reveals distinct strengths and weaknesses in predicting crop yields and prices. Random Forest and XGBoost models achieved superior performance, demonstrated by lower MAE and higher R² scores, making them particularly effective for agricultural forecasting. These models successfully identified key environmental factors such as rainfall, temperature, and humidity as significant predictors, validating the importance of integrating weather data into predictive models.

The LSTM model struggled with the agricultural dataset, primarily due to the high noise levels and limited data available for training. Although LSTM is known for capturing long-term dependencies, its performance was hindered by overfitting and difficulties in generalizing from smaller training sets, leading to higher MAE and lower R² scores compared to tree-based models.

ARIMA, while effective for modeling linear trends, had limited success in capturing the complex, non-linear patterns in agricultural data. The model struggled with sudden market shocks and weather-related fluctuations, which are often non-linear in nature. As a result, its predictions were less accurate in capturing both short-term price volatility and long-term yield trends, especially in the presence of external influencing factors.

Prophet, designed for seasonal data, performed reasonably well in capturing cyclical patterns but struggled with irregular fluctuations and outliers in the agricultural data. Its reliance on predefined seasonal and trend components made it less flexible in handling the unexpected disruptions that can significantly affect crop prices and yields, limiting its overall predictive accuracy compared to models like XGBoost and Random Forest.

Future efforts should focus on expanding datasets to include market trends and policy factors, exploring hybrid models, and developing region-specific predictions using geospatial data. Creating a real-time mobile application for farmers could enhance practical utility, providing timely, actionable insights to navigate market fluctuations and optimize decision-making.

8 CONCLUSION

1. The project successfully applied advanced time-series and machine learning models to predict crop yields and prices, addressing a critical challenge in the agricultural sector.
2. Integrating diverse data sources like weather conditions and soil nutrients significantly improved the quality of predictions, enabling farmers and stakeholders to make informed decisions.
3. The comparative analysis of different models highlighted the strengths of ensemble methods like Random Forest and gradient-boosting algorithms for agricultural forecasting.

9 FUTURE WORK

1. To enhance the robustness and applicability of the models, future efforts will focus on expanding the dataset to include market demand trends and government policy factors, potentially improving accuracy by 10%.
2. Hybrid models combining LSTM and Random Forest will be explored to reduce the MAE below 4 quintals/acre, while further tuning of XGBoost will refine price predictions.
3. Region-specific models will be developed by integrating geospatial data, such as satellite imagery, to improve localized predictions and reduce errors by 20%.
4. Deployment efforts will aim at creating a real-time mobile application with a target response time of under 5 seconds for user queries, ensuring accessibility for farmers.
5. Additionally, sustainability metrics, such as carbon footprint analysis, will be incorporated into recommendations to align with environmentally friendly farming practices, ensuring long-term benefits for stakeholders and the ecosystem.

10 REFERENCES

[1] Shah, V., Hote, Y., & Chauhan, V. (2023). Time Series Forecasting for Agricultural Commodity Prices Using Machine Learning Techniques. Journal of Data Science and Applied AI, 15(4), 299-308.

[2] Sajid, M., & Baig, I. (2023). Comparative Analysis of Prophet and LSTM for Seasonal Time Series Data in Agriculture. International Journal of Agricultural Research and Management, 21(2), 123-133.

[3] Ahmed, Z., & Ramakrishnan, K. (2023). Using ARIMA and LSTM in Predicting Agricultural Yield Volatility. Journal of Time Series Analysis in Agriculture, 12(3), 212-224.

[4] Jadhav, V., Reddy, C. B. V., & Gaddi, G. (2017). Application of ARIMA Model for Forecasting Agricultural Prices. Journal of Agricultural Science and Technology, 19, 981-992.

[5] Ray, S., Lama, A., Mishra, P., Biswas, T., Das, S. S., & Gurung, B. (2023). An ARIMA-LSTM Model for Predicting Volatile Agricultural Price Series with Random Forest Technique. Applied Soft Computing, 149, 110939.

[6] Bhardwaj, M. R., Pawar, J., Bhat, A., Deepanshu, Enaganti, I., Sagar, K., & Narahari, Y. (2023). An Innovative Deep Learning-Based Approach for Accurate Agricultural Crop Price Prediction. Journal of Applied Machine Learning for Agriculture, 8(1), 145-158.

[7] Thapaswini, G., & Gunasekaran, M. (2022). A Methodology for Crop Price Prediction Using Machine Learning. In IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC).

[8] Liang, X., Chen, J., & Wang, T. (2022). A Hybrid Deep Learning Approach for Forecasting Crop Prices Using Ensemble Models. Agricultural Economics and Data Science, 18(2), 145-159.

[9] Kumar, P., & Patel, R. (2023). Integrating Environmental Data for Crop Yield Prediction Using Random Forest and LSTM Models. Journal of Environmental Data Science and Agriculture, 14(3), 213-225.