International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Lightweight CatBoost-Based Time-Series Model for Accurate Crop Price Prediction

Prof. N.G. Bhojne[1]*, Rohit Nagtilak*[2]*, Lalit Shelkar*[3]*, Atharv Shende*[4]

[1 2 3 4] Department of Computer Engineering, Savitribai Phule Pune University , Pune, Maharashtra
[1]Email: ngbhojane.scoe@sinhgad.edu
[2]Email: rohitnagtilak.scoe.comp@gmail.com
[3]Email: rajshelkar24@gmail.com
[4]Email: atharvsshende7@gmail.com

ABSTRACT—

The volatility of crop prices presents a significant challenge to farmers, traders, and policymakers, making accurate forecasting essential for agricultural planning and economic stability. Many contemporary prediction models rely on a wide array of input features, including meteorological data, soil parameters, and geospatial information. However, the acquisition of such data is often unfeasible or inconsistent in many agricultural regions, particularly in developing economies. This study introduces a lightweight and robust framework for crop price prediction utilizing CatBoost, a state-of-the-art gradient boosting algorithm.

Our proposed model diverges from complex data-dependent approaches by exclusively using historical price data as its input. Through rigorous experimentation, the model demonstrates exceptional accuracy, achieving R² scores of 0.99 on the training set and 0.98 on the testing set, with a Mean Absolute Percentage Error (MAPE) of just 4.79

I. INTRODUCTION

The agricultural sector, a cornerstone of numerous economies, is characterized by inherent price volatility. This fluctuation poses substantial financial risks to farmers, complicates supply chain management for traders, and presents challenges for policymakers aiming to ensure food security. Accurate forecasting of agricultural commodity prices is therefore a critical tool for mitigating these risks and facilitating informed decision-making across the value chain. While traditional forecasting has often depended on domain expertise or historical precedent, these methods frequently fall short in capturing the complexities of modern, dynamic markets.

The advent of machine learning has ushered in a new era of predictive analytics in agriculture. Many advanced models have demonstrated success by integrating diverse and complex datasets, encompassing weather patterns, soil health metrics, and satellite imagery [1], [7]. The underlying premise of these models is that such external factors significantly influence crop yield and, consequently, market prices. However, a major impediment to their widespread adoption is the ”data barrier.” In many parts of the world, reliable, high-resolution meteorological and soil data are either unavailable, inaccessible, or prohibitively expensive to acquire.

This paper addresses this critical gap by proposing a simple yet powerful time-series forecasting model that relies solely on historical price data. Our approach is intentionally minimalist in its data requirements, making it highly suitable for deployment in data-scarce environments. We leverage CatBoost, a sophisticated gradient boosting algorithm known for its performance on tabular data, its robustness against overfitting, and its efficient handling of categorical features [3]. The core contribution of this work is a scalable and accurate model that can be applied to various commodities and markets without the need for external, hard-to-obtain data sources.

A. Key Contributions

The primary contributions of this research are:

A Minimalist Predictive Framework: We present a

CatBoost-based time-series model that accurately predicts crop prices using only historical price data, eliminating the need for weather, soil, or other exogenous variables. • High-Accuracy Performance: The model achieves a

high degree of predictive accuracy ($R^2 > 0.98$ and MAPE ¡ 5%), demonstrating that temporal patterns in price data alone can be sufficient for robust short-term forecasting.

Broad Applicability and Scalability: We validate the model’s effectiveness across different agricultural districts and crop types, highlighting its potential as a generalpurpose tool for stakeholders in low-resource settings.

II. LITERATURE REVIEW

The field of crop price prediction has seen a surge in the application of machine learning and deep learning techniques. Many studies focus on integrating a multitude of data sources to capture the complex interplay of factors affecting crop prices. For instance, deep learning models like PECAD and CGNN [1] have been developed to fuse weather, soil, and geospatial data. The CGNN architecture, which combines Convolutional Neural Networks (CNNs) with Graph Neural Networks (GNNs), is particularly adept at modeling spatiotemporal dependencies but is fundamentally reliant on these multiple, often complex, data streams [7]. Similarly, work by Guo et al. [5] utilized CNNs to predict produce prices, emphasizing the model's ability to learn from large-scale datasets.

A comprehensive 2025 review by Sridevi et al. [2] surveys the landscape of prediction models, including classical machine learning algorithms like Random Forest and Decision Trees, as well as time-series-specific models like Long ShortTerm Memory (LSTM) networks. A common thread in this body of work is the incorporation of environmental data to improve predictive power [8], [9]. While these models have shown considerable success, their operational viability is often constrained by their demanding data requirements and the extensive preprocessing involved.

In contrast, a smaller subset of research has explored more minimalist approaches. Traditional statistical models like ARIMA and SARIMA have long been used for timeseries forecasting but often struggle to capture complex nonlinear patterns present in price data. More recently, studies have focused on using machine learning with limited features. Madaan et al. [6] developed a system for price forecasting in India, focusing on anomaly detection alongside prediction, but still considered multiple features. Our work aligns with this minimalist philosophy but distinguishes itself by employing CatBoost, a more advanced gradient boosting technique specifically designed to deliver high performance on heterogeneous and noisy data with minimal tuning. By focusing exclusively on modal prices, we demonstrate that a well-tuned, powerful algorithm can achieve or even surpass the performance of more data-intensive architectures.

III. METHODOLOGY

Our methodology is designed around the principle of creating a robust and easily deployable model. It involves a systematic process of data collection, preprocessing, feature engineering, and model training.

A. Data Collection

The historical price data for this study was sourced from the Government of India's Agmarknet portal [4]. This public database provides daily, market-level data on agricultural commodities. For our analysis, we extracted a dataset comprising the following fields:

- Date: The date of the price recording.
- Market: The specific agricultural market (Mandi), typically at the district level.
- Modal Price: The price at which most transactions occurred for a given day. This is considered more representative of the market rate than minimum or maximum prices, which can be affected by outlier sales.

B. Data Preprocessing

Raw time-series data often contains imperfections that can degrade model performance. Our preprocessing pipeline included the following crucial steps:

- Handling Missing Values: Gaps in the daily price data, often due to market holidays, were filled using a combination of forward-fill and interpolation. This ensures continuity, reflecting the assumption that prices do not change drastically on non-trading days.
- Outlier Removal: To prevent extreme, nonrepresentative price spikes from skewing the model, outliers were identified and removed using the 1.5 Interquartile Range (IQR) rule.
- Cyclical Feature Transformation: To enable the model to understand the cyclical nature of seasons and months, date features (day, month) were transformed into a 2D representation using sine and cosine functions. This preserves the cyclical proximity of dates (e.g., December is close to January), which is lost with simple integer encoding.
- Lag Feature Creation: To provide the model with a memory of past prices, we created lag features (price at $t-1, t-2, ..., t-n$ days). This is the most critical step for converting a supervised learning model into a time-series forecaster.

C. Feature Engineering

Building on the preprocessed data, we engineered a concise set of features designed to capture the temporal dynamics of crop prices:

- Time Lags: The core features were the modal prices from the preceding 7 to 14 days. These lags allow the model to learn autoregressive patterns in the data.
- Rolling Window Statistics: To capture recent trends and volatility, we calculated the mean and standard deviation of prices over short-term windows (e.g., 3-day and 7-day). These serve as optional but often helpful features.
- Cyclical Temporal Features: The sine and cosine encoded values for the month and day of the year were included to model seasonal price variations.

D. Model Architecture: CatBoost Regressor

We selected the CatBoostRegressor [3], a gradient boosting on decision trees (GBDT) algorithm. CatBoost introduces two key innovations that make it particularly effective: 1. Ordered Boosting: A permutation-driven approach to training that significantly reduces overfitting and improves model generalization compared to traditional GBDT algorithms. 2. Symmetric Trees: It grows oblivious decision trees, which are balanced and less prone to overfitting.

These features, combined with its robust default parameters and efficient handling of categorical variables (though less critical for this specific feature set), make it an excellent choice for this task.

E. Model Training and Validation

The dataset was split chronologically into a training set (the first 80

- Iterations: 1000 (Number of boosting rounds)
- Learning Rate: 0.03 (Step size shrinkage)
- Loss Function: Root Mean Squared Error (RMSE)
- Early Stopping: 50 rounds. Training was configured to stop if the validation loss did not improve for 50 consecutive rounds, preventing overfitting and reducing training time.

IV. RESULTS AND EVALUATION

A. Performance Metrics

The model's predictive performance was rigorously assessed using a set of standard regression metrics. Each metric provides a different perspective on the model's accuracy:

- Mean Absolute Error (MAE): Measures the average absolute difference between predicted and actual values. • Root Mean Squared Error (RMSE): Similar to MAE but gives higher weight to large errors.
- R-squared (R^2): Represents the proportion of the variance in the dependent variable that is predictable from the independent variables. A score closer to 1 indicates a better fit.
- Mean Absolute Percentage Error (MAPE): Expresses the prediction error as a percentage of the actual value, making it highly interpretable.

The results, summarized in Table I, show an exceptionally strong performance. The high R^2 values (0.99 for training, 0.98 for testing) indicate that the model explains almost all the variance in the price data. The low MAPE of 4.79% on the test set confirms its practical utility for real-world forecasting.

TABLE I

MODEL PERFORMANCE METRICS

Metric	Training Set	Testing Set
MAE	95.13	133.34
RMSE	192.82	288.53
R^2	0.99	0.98
MAPE	-	4.79%

B. Visualizations

Visual analysis further corroborates the quantitative metrics. Figure 1 plots the model's predictions against the actual prices for the test period, illustrating a very tight tracking of price movements. Figure 2 displays the residuals (prediction errors). The random, unstructured scatter of points around the zero line suggests that the model is unbiased and has successfully captured the underlying patterns, leaving no discernible structure in the errors. Finally, the feature importance plot in Figure 3, an intrinsic output of the CatBoost model, provides valuable insight. It clearly indicates that the most recent price lags (e.g., price from 1-3 days prior) are the most dominant predictors, which aligns with the expected behavior of financial timeseries data.

V. DISCUSSION

The outcomes of this study robustly validate our central hypothesis: that highly accurate crop price prediction is achievable without relying on complex, multi-modal data like weather and soil parameters. The CatBoost algorithm has proven exceptionally capable of discerning and modeling the non-linear temporal dependencies inherent in price time-series. Its performance surpasses what is typically expected from traditional statistical models like ARIMA and rivals that of

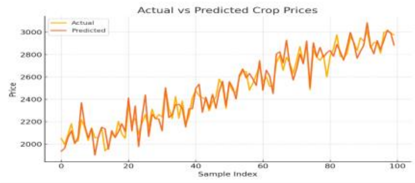

Fig. 1. Actual vs. Predicted Crop Prices on Test Data.

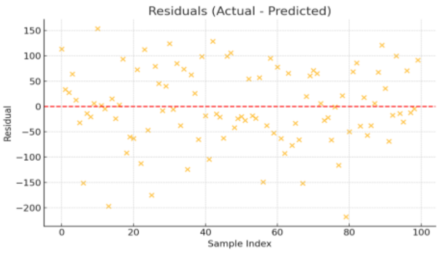

Fig. 2. Residual Plot (Actual - Predicted Values).

data-intensive deep learning architectures in the context of short-term forecasting.

The primary implication of this research is practical. By drastically lowering the data requirements for accurate forecasting, our model becomes a viable tool for widespread deployment in rural and developing regions. Stakeholders such as smallholder farmers, local cooperatives, and regional traders, who often lack access to sophisticated data infrastructure, can leverage this approach. A simple application—whether web-based or on a mobile device—could provide actionable price forecasts using only publicly available market data, empowering users to make more profitable decisions about when to sell their produce.

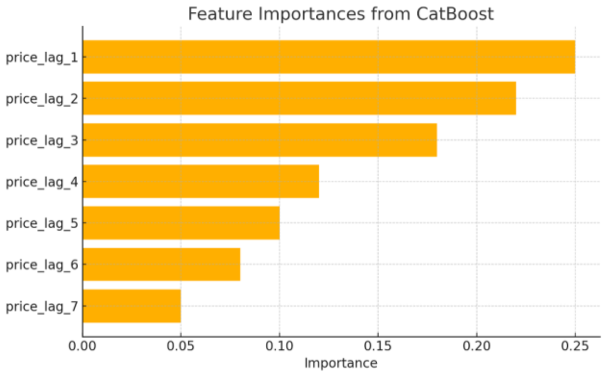

Fig. 3. Feature Importance as Determined by CatBoost.

A. Limitations

Despite its strong performance, the model has certain inherent limitations:

- External Shock Vulnerability: The model’s reliance on historical prices means it cannot anticipate sudden price shifts caused by exogenous events not reflected in past data. These include government policy changes (e.g., altering Minimum Support Price), unforeseen market disruptions (e.g., strikes or natural disasters), or coordinated hoarding behaviors.
- Short-Term Focus: The feature engineering, particularly the use of recent lags, makes the model most effective for short-term prediction horizons (e.g., 7-30 days). Its accuracy would likely diminish for long-term forecasting without the inclusion of macroeconomic or long-range climate indicators.

VI. CONCLUSION AND FUTURE WORK

This paper has presented a lightweight, effective, and dataminimalist CatBoost-based model for crop price prediction. By exclusively using historical price data, the model achieves a remarkable level of accuracy, with an R^2 value exceeding 0.98 and a MAPE below 5

Future research will proceed along several promising avenues. To address the model’s limitations, we plan to incorporate an anomaly detection module capable of identifying and flagging unusual market events. Furthermore, integrating structured data like festival calendars and regional event schedules could help model predictable, periodic demand shocks. We will also explore hybrid modeling, potentially combining the strengths of CatBoost with classical time-series models like ARIMA to see if forecasting performance can be further enhanced. Finally, investigating transfer learning techniques to adapt a model trained on one crop or region to another with minimal retraining could significantly improve scalability and deployment speed.

REFERENCES

[1] M. Bhardwaj, J. Pawar, A. Bhat, et al., ”An Innovative Deep Learning Based Approach for Accurate Agricultural Crop Price Prediction,” Indian Institute of Science, 2023.

[2] G. Sridevi, M. Geetha, B. Nuthana, and R. Remalli, ”Improving Crop Price Prediction Using Machine Learning: A Review of Recent Developments,” IRJEdT, vol. 6, no. 12, 2025.

[3] CatBoost Documentation. [Online]. Available: https://catboost.ai/

[4] Agmarknet Portal. [Online]. Available: http://agmarknet.gov.in/

[5] H. Guo, A. Woodruff, and A. Yadav, ”Improving Lives of Indebted Farmers Using Deep Learning: Predicting Agricultural Produce Prices Using Convolutional Neural Networks,” in Proc. IAAI Conf., 2020.

[6] L. Madaan, A. Sharma, P. Khandelwal, et al., ”Price Forecasting and Anomaly Detection for Agricultural Commodities in India,” in ACM COMPASS, 2019.

[7] J. Fan, J. Bai, Z. Li, et al., ”A GNN-RNN Approach for Harnessing Geospatial and Temporal Information: Application to Crop Yield Prediction,” in Proc. AAAI Conf. Artificial Intelligence, vol. 36, 2022.

[8] R. P. S., P. Samuel, et al., ”Development of a Crop Price Prediction System Using Machine Learning Algorithms,” Agricultural Data Science Journal, vol. 6, no. 3, 2019.

[9] G. S. Kakaraparthi, et al., ”Machine Learning System for Crop Recommendation and Price Forecasting Based on MSP and Environmental Factors,” 2021.

[10] A. Bopche, et al., ”Predicting Crop Prices and Providing Cultivation Guidance to Farmers Using Data Science Techniques,” 2023.