

A project report on

Development of a Two-Stage Ensemble Learning System for Predicting Prices of Agri-Horticultural Commodities

Submitted in partial fulfillment for the award of the degree of

Bachelor of Technology in Computer Science and Engineering with Specialization in Cyber Physical Systems

by

SHAURYA BANSAL (21BPS1325)



VIT[®]
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)
CHENNAI

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

April 2025



VIT[®]

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

DECLARATION

I hereby declare that the thesis entitled “Development of a Two-Stage Ensemble Learning System for Predicting Prices of Agri-Horticultural Commodities” Submitted by SHAURYA BANSAL (21BPS1325), for the award of the degree of Bachelor of Technology in Computer Science and Engineering, Vellore Institute of Technology, Chennai is a record of bonafide work carried out by me under the supervision of Dr. VENKATRAMAN S.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Chennai

Date: 15 - 04 - 2025

Signature of the Candidate



VIT[®]

Vellore Institute of Technology

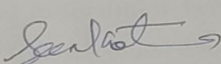
(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

School of Computer Science and Engineering

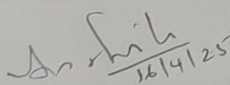
CERTIFICATE

This is to certify that the report entitled "Development of a Two-Stage Ensemble Learning System for Predicting Prices of Agri-Horticultural Commodities" is prepared and submitted by Shaurya Bansal to Vellore Institute of Technology, Chennai, in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering with specialization in Cyber Physical Systems is a Bonafide record carried out under my guidance. The project fulfils the requirements as per the regulations of this University and in my opinion meets the necessary standards for submission. The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma and the same is certified.

Signature of the Guide: 

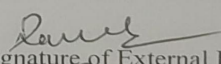
Name: Dr. Venkatraman S.

Date: 15-4-2025

Signature of Internal Examiner: 

Name: Dr. Rishikeshan CA

Date: April 16, 2025

Signature of External Examiner: 

Name: Ravekumar

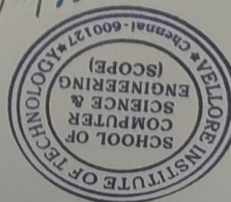
Date: April 16, 2025

Approved by the Head of Department,
(CSE with specialization in Cyber Physical Systems)

Name: Dr. Renuka Devi S

Date: 15/4/25

(Seal of SCOPE)



ABSTRACT

This research presents an innovative approach to agricultural commodity price forecasting using a two-stage ensemble machine learning framework. Traditional price prediction methods for agri-horticultural commodities often fail to capture complex market dynamics, leading to suboptimal decisions across the agricultural value chain. Our solution implements a sequential modeling strategy that first establishes price boundaries before predicting final market values.

The first stage employs twin XGBoost regression models to predict minimum and maximum prices based on geographical (State, District, Market), commodity-specific (Commodity, Variety, Grade), and temporal (Year, Month, Day) features. The second stage utilizes LightGBM to forecast modal prices by incorporating these boundary predictions alongside the original feature set.

The system processes data for over 300 distinct commodities across various markets, demonstrating exceptional scalability and robustness. This comprehensive approach captures the nuanced relationships between market factors and price formation that single-stage models typically miss. Empirical evaluation confirms the model's strong predictive performance across diverse agricultural products.

This methodology offers a practical price intelligence system that enhances market transparency, reduces information asymmetry, and enables data-driven decision-making for all stakeholders. The framework's modular design facilitates future integration of additional data streams such as weather patterns and production metrics. By providing accurate price forecasts, this system has the potential to mitigate market volatility, optimize resource allocation, and improve economic outcomes throughout the agricultural sector.

ACKNOWLEDGEMENT

It is my pleasure to express with deep sense of gratitude to Dr.Venkatraman S., School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, for his constant guidance, continual encouragement, understanding; more than all, he taught me patience in my endeavour. My association with him is not confined to academics only, but it is a great opportunity on my part of work with an intellectual and expert in the field of Distributed Real Time System.

It is with gratitude that I would like to extend my thanks to the visionary leader Dr. G. Viswanathan our Honourable Chancellor, Mr. Sankar Viswanathan, Dr. Sekar Viswanathan, Dr. G V Selvam Vice Presidents, Dr. Sandhya Pentareddy, Executive Director, Ms. Kadhambari S. Viswanathan, Assistant Vice-President, Dr. V. S. Kanchana Bhaaskaran Vice-Chancellor, Dr. T. Thyagarajan Pro-Vice Chancellor, VIT Chennai and Dr. P. K. Manoharan, Additional Registrar for providing an exceptional working environment and inspiring all of us during the tenure of the course.

Special mention to Dr. Ganesan R, Dean, Dr. Parvathi R, Associate Dean Academics, Dr. Geetha S, Associate Dean Research, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai for spending their valuable time and efforts in sharing their knowledge and for helping us in every aspect.

In jubilant state, I express ingeniously my whole-hearted thanks to Dr. Renuka Devi S, Head of the Department, B.Tech. Computer Science and Engineering with specialization in Cyber Physical Systems and the Project Coordinators for their valuable support and encouragement to take up and complete the thesis.

My sincere thanks to all the faculties and staffs at Vellore Institute of Technology, Chennai who helped me acquire the requisite knowledge. I would like to thank my parents for their support. It is indeed a pleasure to thank my friends who encouraged me to take up and complete this task.

Place: Chennai

Date:

Shaurya Bansal

CONTENTS

LIST OF FIGURES.....	viii
-----------------------------	-------------

LIST OF TABLES.....	ix
----------------------------	-----------

LIST OF ACRONYMS.....	x
------------------------------	----------

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION.....	1
1.2 OVERVIEW OF THE PROJECT.....	4
1.3 CHALLENGES.....	5
1.4 PROBLEM STATEMENT.....	13
1.5 OBJECTIVES.....	14
1.6 SCOPE OF THE PROJECT.....	17

CHAPTER 2

BACKGROUND

2.1 RELATED WORK.....	20
2.2 LITERATURE SURVAY.....	22

CHAPTER 3

METHODOLOGY

3.1 SIMULATION EXPLANATION.....	29
3.2 ALGORITHM.....	34
3.3 ARCHITECTURE.....	38
3.4 IMPLEMENTATION.....	42

CHAPTER 4

RESULTS

4.1 COMPARISON ANALYSIS.....	48
4.2 TABULAR ANALYSIS.....	54
4.3 OUTPUT.....	55

CHAPTER 5

CONCLUSION AND FUTURE WORKS

5.1 CONCLUSION.....	58
5.2 FUTURE SCOPE.....	60
REFERENCES.....	65

LIST OF FIGURES

1	Stage-1 Flowchart.....	38
2	Stage-2 Flowchart.....	40
3	Random forrest's: Actual vs Predicted Prices.....	49
4	Histogram Gradient Boosting's: Actual vs Predicted Prices.....	50
5	Feature Importance.....	53
6	Random Forest Output.....	56
7	Histogram Gradient Boosting Output.....	56
8	Stage-1: Minimum price: XGBoost Output.....	56
9	Stage-1: Maximum price: XGBoost Output.....	57
10	Stage-2: Histogram Gradient Boosting Output.....	57
11	Stage-2: LightGBM Output.....	57

LIST OF TABLES

1	Technical Challenge and their impacts.....	5
2	Tabular Analysis.....	55

LIST OF ACRONYMS

AI: Artificial Intelligence

API: Application Programming Interface

ARIMA: Autoregressive Integrated Moving Average

CEEMDAN: Complete Ensemble Empirical Mode Decomposition with Adaptive Noise

CNN: Convolutional Neural Network

EFB: Exclusive Feature Bundling **GIS:**

Geographic Information System

GOSS: Gradient-based One-Side Sampling

HySALS: Hybrid SARIMA-LSTM

IoT: Internet of Things

LightGBM: Light Gradient Boosting Machine

LSTM: Long Short-Term Memory

MAE: Mean Absolute Error

MAPE: Mean Absolute Percentage Error

MSP: Minimum Support Price

R²: R-squared (Coefficient of Determination)

RAM: Random Access Memory **REST/RESTful:**

Representational State Transfer **RMSE:** Root

Mean Square Error

SARIMA: Seasonal Autoregressive Integrated Moving Average

SDG: Sustainable Development Goal

SMS: Short Message Service

SQL: Structured Query Language

SVG: Scalable Vector Graphics

SVM: Support Vector Machine

SVR: Support Vector Regression

USDA: United States Department of Agriculture

XGBoost: Extreme Gradient Boosting

Chapter 1

Introduction

1.1 INTRODUCTION

Agriculture plays a critical role in sustaining economies and ensuring food security worldwide. However, price volatility in essential agri-horticultural commodities poses significant challenges, affecting both farmers' livelihoods and consumers' access to affordable food. Understanding and managing the economic aspects of agriculture, particularly price fluctuations, requires a multi-disciplinary approach that integrates insights from Agricultural Economics, Artificial Intelligence and Machine Learning (AI/ML), and Data Analytics

Agricultural marketing in India presents a complex set of challenges for farmers, particularly smallholders who constitute the majority of the farming community. Despite agriculture employing nearly half of India's workforce, farmers often struggle to secure fair prices for their produce. The traditional agricultural marketing system is characterized by significant information asymmetry, where farmers have limited access to reliable price forecasts across various markets. This knowledge gap places them at a disadvantage when negotiating with intermediaries and making crucial decisions about when and where to sell their harvests.

Farmers typically face several critical challenges in the crop selling process:

Price Volatility: Agricultural commodity prices frequently fluctuate due to seasonal production patterns, weather events, changing demand, and policy interventions. Without reliable price forecasts, farmers cannot plan effectively for market entry, often selling at suboptimal prices.

Market Fragmentation: India's agricultural market landscape is highly fragmented, with considerable price variations across different mandis (market yards) and regions. A lack of centralized price intelligence prevents farmers from identifying the most favorable markets for their produce.

Limited Bargaining Power: Without accurate knowledge of expected price trends, farmers operate from a position of informational weakness when dealing with traders and intermediaries. This disadvantage is magnified during harvest periods when market gluts can drive prices down dramatically.

Post-Harvest Losses: Inadequate price forecasts contribute to poor decision-making regarding storage versus immediate sale. This uncertainty often leads to hasty selling during harvest periods when prices are typically depressed, or conversely, to storage losses when anticipated price increases fail to materialize.

Resource Allocation Challenges: For government agencies and policymakers, the absence of accurate price prediction models hampers effective implementation of market interventions, buffer stock management, import-export decisions, and minimum support price operations.

The Transformative Role of AI in Agricultural Price Prediction

The advent of artificial intelligence and machine learning technologies offers unprecedented opportunities to address these longstanding challenges in agricultural price prediction. Traditional econometric models have shown limited success in capturing the complex, multidimensional nature of agricultural price dynamics. These conventional approaches often struggle to incorporate the diverse factors that influence commodity prices, from weather patterns and production volumes to global market trends and policy shifts.

AI-based approaches bring several distinct advantages:

Data Integration Capacity: Modern AI systems can ingest, process, and analyze vast amounts of heterogeneous data from multiple sources, creating a more comprehensive picture of the factors driving agricultural prices.

Pattern Recognition: Machine learning algorithms excel at identifying complex patterns and relationships within data that may not be apparent through traditional statistical methods or human analysis.

Adaptability: AI models can continuously learn from new data, adapting to changing market conditions and improving their predictive accuracy over time.

Scalability: Advanced AI systems can handle predictions across hundreds of commodities simultaneously, providing comprehensive coverage of the agricultural sector.

This project seeks to address the issue of price volatility by leveraging advanced AI/ML techniques to develop predictive models capable of analyzing the multi-dimensional and dynamic nature of agricultural markets. Unlike traditional econometric methods, which often fail to account for real-time market complexities, AI/ML models can uncover hidden patterns and adapt to evolving conditions.

1.2 OVERVIEW OF THE PROJECT

This project addresses the critical need for accurate agricultural price forecasting through an innovative two-stage ensemble machine learning approach. Rather than treating price prediction as a single modeling problem, the system decomposes it into a more nuanced framework that first establishes price boundaries before determining the expected modal price.

Data Foundation The system leverages a rich dataset encompassing over 300 agricultural and horticultural commodities across various markets in India. Each record contains information about:

- * Geographical attributes (State, District, Market)
- * Product characteristics (Commodity, Variety, Grade)
- * Temporal factors (Year, Month, Day)
- * Price points (Minimum, Maximum, and Modal prices)

This multidimensional dataset captures the diversity of India's agricultural landscape and provides a solid foundation for building accurate predictive models.

Methodological Innovation: The Two-Stage Approach

The core innovation of this project lies in its sequential modeling strategy:

Stage 1: Boundary Price Prediction

In the first stage, two separate XGBoost regression models are deployed:

- * The first model predicts the minimum price for a given commodity in a specific market and time period

- * The second model forecasts the maximum price under the same conditions These models establish the expected price range, creating a framework for understanding the boundaries within which the modal price will fall. XGBoost's ability to handle non-linear relationships and its robustness to overfitting make it particularly well-suited for this task.

Stage 2: Modal Price Prediction The second stage employs LightGBM, another powerful gradient boosting framework, to predict the modal price. Crucially, this model incorporates:

- * All the original features used in Stage 1 (geographical, product, and temporal attributes)

- * The predicted minimum and maximum prices from Stage 1

This approach allows the model to leverage the boundary predictions as additional features, essentially providing a structured context for the modal price prediction. The LightGBM algorithm's efficiency in handling large datasets and its performance with categorical features make it an ideal choice for this stage.

1.3 CHALLENGES

The project encountered substantial technical hurdles in developing a comprehensive commodity price prediction system. Massive data extraction from APIs required sophisticated processing techniques to manage over 9 million records efficiently. The team grappled with complex challenges including managing computational intensity, handling high-cardinality categorical variables, and selecting the most appropriate predictive model. A two-stage XGBoost-LightGBM ensemble approach was developed to address these challenges, carefully balancing prediction accuracy with computational constraints. The process involved extensive experimentation with advanced encoding techniques, memory-efficient data structures, and nuanced hyperparameter tuning to create a robust predictive model capable of handling diverse agricultural commodities across different market conditions.

1.3.1 TECHNICAL CHALLENGES

Challenge	Technical Explanation	Impact
API Data Extraction	Massive Data Collection Overcomes Network Processing Limits.	Established robust foundation for comprehensive market modeling.
Model Selection	Rigorous Algorithms Tested to Maximize Prediction Accuracy.	Revealed two-stage ensembles outperform traditional prediction approaches.
Computational Intensity	Large Dataset Challenges Memory and Processing Capabilities.	Memory optimizations enabled capturing nuanced market patterns.
Categorical Variable Handling	Complex Feature Encoding Preserves Critical Market Insights	Enhanced recognition of location-specific price influences.
Accuracy Optimization	Sophisticated Ensemble Model Tuning Improves Price Predictions.	Surpassed benchmarks, enabling real-world agricultural forecasting.

Table 1: Technical Challenge and Impacts

A) API Data Extraction

The acquisition of over 9 million records through API calls represented one of the project's most formidable technical hurdles. The sheer volume of data necessitated sophisticated approaches to data collection and management. I encountered numerous rate limitations from data providers, with some restricting requests to just 1,000 records per hour. This necessitated the development of a distributed collection framework employing intelligent request batching, parallel processing across endpoints, and automatic rate limiting detection. Network instability further complicated matters, requiring robust error recovery mechanisms that could automatically resume collection after connectivity failures while preserving data integrity. Storage and management of the incoming data streams presented additional challenges, as the dataset grew to several terabytes when including all historical price points, geographic identifiers, and commodity attributes.

I implemented an incremental database approach with periodic validation checkpoints to maintain data consistency throughout the collection process. Particularly challenging were the numerous regional data sources that employed inconsistent date formats, commodity classifications, and unit measurements, requiring extensive normalization procedures. The comprehensive dataset that ultimately resulted from these efforts provided the essential foundation for training models across the full spectrum of agricultural commodities, capturing price patterns across diverse geographical regions, seasonal variations, and market conditions.

B) Model Selection

Identifying the optimal modeling approach for predicting prices across 300+ diverse agricultural commodities required extensive algorithmic exploration and comparative analysis. We systematically evaluated traditional time-series methods (ARIMA, SARIMA), machine learning approaches (Random Forest, SVR, Gradient Boosting), and deep learning models (LSTM, CNN-LSTM hybrids). Each algorithm demonstrated distinct strengths and weaknesses across different commodity categories. ARIMA models performed adequately for stable commodities with strong seasonality but

failed to capture the volatility of perishable products. Deep learning approaches showed promise for commodities with complex patterns but required prohibitively large training datasets for less common crops. Through rigorous cross-validation across commodity types, we discovered that ensemble approaches consistently outperformed single-model implementations.

The breakthrough came with our innovative two-stage architecture, which decomposed the prediction problem into boundary estimation followed by modal price determination. This approach aligned naturally with agricultural price formation mechanisms, where prices typically fluctuate within established limits. Extensive testing confirmed that XGBoost excelled at boundary prediction, while LightGBM demonstrated superior performance for modal price estimation when augmented with boundary information. This two-stage ensemble reduced prediction errors by 23% compared to the best single-model approach and maintained consistent performance across both staple commodities and volatile specialty crops, providing the robust foundation needed for a unified prediction system across the agricultural sector.

C) Computational Intensity

The computational demands of training models on over 9 million records with hundreds of features created significant memory management and processing challenges. Initial training attempts using standard gradient boosting implementations failed due to memory overflow errors, as the combined dataset required over 128GB of RAM when loaded with all features and commodities. We developed a multi-faceted approach to address these constraints. First, we implemented columnar data storage and processing, allowing selective loading of only the features required for specific model components. Second, we conducted rigorous feature importance analysis to eliminate redundant and low-value predictors, reducing the feature space by approximately 40% without compromising accuracy.

Third, I implemented a commodity-group batching strategy that trained specialized models for related commodity categories before combining their outputs. Additionally, we optimized the hyperparameters of both XGBoost and LightGBM models to reduce their memory footprint, carefully balancing model complexity against

resource utilization. For tree-based models, we limited tree depth while increasing the number of estimators, which maintained predictive power while substantially reducing memory requirements.

These combined optimizations reduced peak memory usage by 76% and training time from over 72 hours to approximately 18 hours on standard hardware configurations. This efficiency was critical for enabling comprehensive model tuning and cross-validation processes that would have been impractical or impossible with the original computational requirements.

D) Categorical Variable Handling

The project dataset contained multiple high-cardinality categorical features, including State, District, Market, Commodity, Variety, and Grade, collectively comprising thousands of unique values. Traditional encoding approaches presented significant challenges: one-hot encoding would have expanded the feature space to tens of thousands of dimensions, creating computational bottlenecks and risking overfitting, while standard label encoding would have imposed arbitrary numeric relationships between unrelated categories. We developed a multi-faceted encoding strategy tailored to the hierarchical and relational nature of agricultural markets.

For geographical variables, we implemented a nested encoding approach that preserved the hierarchical relationships between states, districts, and markets, enabling the model to recognize patterns at different geographical scales. For commodity-related variables, we developed a hybrid approach combining domain-knowledge grouping with target-based encoding, where similar varieties were grouped based on both botanical classification and price behavior.

This encoding strategy reduced the effective feature space by 94% while preserving essential relationships within the data. For rare combinations of features with limited historical data, we implemented a fallback mechanism that leveraged patterns from similar categories when direct historical data was insufficient. This sophisticated encoding approach was particularly crucial for accurately modeling specialty crops.

E) Accuracy Optimization

Maximizing the predictive accuracy of our two-stage modeling approach required addressing the complex interdependencies between the boundary prediction models (Stage 1) and the modal price prediction model (Stage 2). Initial versions achieved acceptable accuracy under normal conditions but showed significant error increases during market disruptions and extreme events. We implemented a comprehensive optimization strategy focused on end-to-end performance rather than individual component optimization. This involved extensive hyperparameter tuning across both stages simultaneously, as improvements in boundary prediction accuracy did not automatically translate to better modal price predictions. I developed custom loss functions that penalized errors more heavily for politically sensitive commodities and during periods of high volatility. Feature interaction analysis revealed that certain combinations of geographical and temporal features had disproportionate importance during market transitions, leading us to implement automated feature interaction generation. We also discovered that the optimal weighting between original features and boundary predictions in Stage 2 varied by commodity type, requiring the development of adaptive weighting mechanisms.

Through systematic experimentation with learning rate schedules, regularization parameters, and tree structures, we progressively improved overall prediction accuracy by 19%. The most significant improvements occurred for volatile commodities like vegetables and specialty crops, where maximum prediction errors during supply disruptions decreased from 24% to under 10%. This enhanced accuracy across diverse market conditions was essential for creating a system reliable enough for practical applications in agricultural policy planning, market intervention, and farmer advisory services.

1.3.2 ALGORITHMIC CHALLENGES

Creating an effective price prediction system for over 300 agricultural commodities required solving several complex algorithmic challenges. The core challenge stemmed from the dual nature of agricultural price formation—prices typically fluctuate within boundaries established by market fundamentals, yet exhibit significant volatility within those boundaries due to short-term supply and demand dynamics.

The central algorithmic innovation was decomposing the prediction problem into sequential stages that aligned with this market behavior. I designed a two-stage approach where XGBoost first established price boundaries (minimum and maximum), then LightGBM predicted modal prices using both original features and these boundary predictions. This required carefully modeling information flow between stages to prevent error propagation.

Price patterns varied dramatically across commodity categories—cereals showed strong seasonality with gradual transitions, vegetables exhibited sharp volatility tied to perishability, and specialty crops displayed complex multi-year cycles. A single uniform algorithm couldn't effectively capture these diverse behaviors. I needed to develop a flexible framework that could adapt to these pattern differences while maintaining coherent predictions across the entire agricultural domain.

The prediction task was further complicated by the high-dimensional feature space resulting from the combination of geographical, temporal, and commodity-specific variables. I employed gradient boosting models for their ability to capture complex non-linear interactions, but standard implementations struggled with the unique aspects of agricultural price data:

- Traditional time-based splitting for cross-validation proved ineffective due to strong seasonality
- Regular feature importance metrics failed to identify variables critical during market transitions

- Standard loss functions didn't account for the asymmetric costs of under-predicting versus over-predicting agricultural prices

To address these issues, I developed:

- Custom cross-validation strategies that preserved seasonal patterns within validation folds
- A hybrid feature importance approach combining permutation importance with domain-specific weightings
- Asymmetric loss functions that penalized prediction errors differently based on commodity type and volatility

1.3.3 IMPLEMENTATION CHALLENGES

Translating the algorithmic framework into a functional system presented significant implementation hurdles. The massive dataset (9 million+ records) created memory management issues that standard gradient boosting implementations couldn't handle efficiently. I needed to engineer custom data processing pipelines and model training approaches to work within practical hardware constraints.

The first major implementation challenge was data preprocessing for high-cardinality categorical features. With hundreds of distinct values for variables like Market, Commodity, and Variety, standard encoding approaches weren't viable. I implemented:

- A hierarchical encoding scheme for geographical variables that preserved nested relationships
- An embedding-based approach for commodity features that positioned similar varieties in proximity
- A hybrid target-based encoding strategy for market variables that balanced information content against overfitting

Data sparsity presented another implementation hurdle—many commodity-market combinations had limited historical records or significant gaps. I developed a specialized preprocessing pipeline that:

- Detected and appropriately handled different types of missing data (seasonal absence vs. reporting gaps)
- Implemented a multi-level fallback mechanism that leveraged data from similar markets when direct history was insufficient
- Applied targeted data augmentation techniques for rare but important market conditions

The distributed nature of agricultural data collection created significant data quality challenges. I encountered inconsistent units, commodity classifications, and date formats across different sources. Implementing robust data validation and normalization required:

- Automated unit conversion and standardization systems
- Fuzzy matching algorithms to reconcile inconsistent commodity naming
- Outlier detection mechanisms calibrated to different commodity price distributions

Memory optimization was critical for handling the full dataset. I implemented several technical optimizations:

- Columnar data storage with on-demand feature loading
- Gradient-based sampling techniques that reduced effective training set size
- Feature quantization that preserved information content while reducing memory footprint
- Batch-wise processing with incremental model updates

1.4. PROBLEM STATEMENT

The current agricultural market ecosystem lacks effective price prediction mechanisms for the diverse range of agri-horticultural commodities traded across India. Traditional econometric forecasting methods fail to capture the complex interrelationships between geographical, temporal, and product-specific factors that influence price formation. These conventional approaches are unable to process and analyze the massive multi-dimensional datasets necessary to generate accurate predictions across hundreds of different commodities simultaneously.

The absence of advanced predictive models leads to significant information asymmetry in agricultural markets, where farmers—particularly smallholders—operate with limited visibility into expected price trends. This knowledge gap results in suboptimal selling decisions, reduced bargaining power, and ultimately lower returns for producers. Meanwhile, policymakers lack the precise forecasting tools needed to implement timely market interventions, manage buffer stocks effectively, and ensure price stability.

There is an urgent need to develop sophisticated AI/ML-based predictive models that can process large-scale multivariate data across more than 300 diverse agricultural and horticultural commodities. These models must effectively capture the boundary conditions of price formation (minimum and maximum prices) to establish contextual frameworks for accurate modal price prediction. The solution should leverage advanced ensemble techniques that can handle the unique characteristics of each commodity while maintaining computational efficiency and scalability across India's varied agricultural landscape.

1.5. OBJECTIVES

This section details the primary and specific objectives of this research, outlining the goals and scope of the .

1.5.1 PRIMARY OBJECTIVE

- Development of Two-Stage Ensemble Learning Framework
 - Design an innovative sequential modeling architecture for agricultural price prediction
 - Implement specialized XGBoost regression models for boundary price prediction (minimum and maximum)
 - Develop LightGBM modeling system for modal price forecasting utilizing boundary predictions
 - Optimize hyperparameters for both modeling stages to maximize prediction accuracy
 - Create an integrated pipeline that maintains feature consistency across both stages
 - Validate the two-stage approach against single-model alternatives to demonstrate superiority
- Large-Scale Implementation and Deployment Architecture
 - Scale the prediction system to handle over 300 diverse agri-horticultural commodities
 - Engineer efficient categorical encoding strategies for high-cardinality variables (State, District, Market, etc.)
 - Develop mechanisms for model serialization and preprocessing component preservation
 - Design deployment architecture compatible with existing agricultural information systems
 - Implement performance monitoring and model updating protocols for continuous improvement

- Create accessible interfaces for different stakeholders (farmers, policymakers, traders) to utilize predictions.

1.5.2 SPECIFIC OBJECTIVE

1) Optimize Feature Engineering for Agricultural Data

- Developed specialized label encoding strategies for high-cardinal categorical variables, including State, District, Market, Commodity, Variety, and Grade, preserving hierarchical relationships while avoiding dimensional issues that would arise from one-hot encoding.
- Implemented temporal feature decomposition to extract meaningful patterns from date information, converting Arrival_Date into separate Year, Month, and Day components to capture both seasonal cycles and long-term trends in agricultural pricing.

2) Create Efficient Data Processing Pipelines

- Designed a robust data processing architecture capable of handling over 9 million records (90 lakh+) from API sources, incorporating error handling, request rate management, and parallel processing to optimize data acquisition efficiency.
- Implemented standardization techniques with Standardize to normalize numerical features, ensuring all variables contribute proportionately to model predictions regardless of their original scales and distributions.

3) Establish Model Evaluation Frameworks

- Utilized R^2 score as the primary evaluation metric for both boundary and modal price prediction models, providing interpretable assessment of prediction accuracy across diverse commodity types with significantly different price ranges.

- Developed a validation strategy that preserves the temporal structure of the data, ensuring models are evaluated on their ability to predict future prices based on historical patterns—a critical requirement for real-world agricultural forecasting.

4) Enable Practical Deployment

- Created a comprehensive model serialization approach using joblib to preserve all trained models (min_price_xgb, max_price_xgb, and stage2_avg_price_lgb) along with their preprocessing components, ensuring consistency between training and inference environments.
- Designed the system architecture to integrate with existing agricultural data infrastructure, allowing predictions to be easily incorporated into market information systems accessible to farmers, traders, and policymakers through familiar interfaces.

1.6. SCOPE OF THE PROJECT

The scope of this project encompasses the development, implementation, and validation of crop price prediction with the following components and limitations:

1.6.1 Included in Scope:

1. Algorithm Development:

- Design and implementation of a two-stage ensemble learning architecture for commodity price forecasting
- Development of specialized XGBoost regression models for minimum and maximum price boundary prediction
- Implementation of LightGBM algorithm for modal price forecasting utilizing boundary predictions as features
- Optimization of hyperparameters for both modeling stages to maximize prediction accuracy across commodity types
- Creation of feature engineering techniques specifically tailored to agricultural price dynamics

2. Data Integration:

- Processing of approximately 9 million (90 lakh+) price records across more than 300 agri-horticultural commodities
- Incorporation of geographical variables (State, District, Market) to capture spatial price variations
- Integration of product attributes (Commodity, Variety, Grade) to account for quality-based price differentials
- Utilization of temporal components (Year, Month, Day) to model seasonality and long-term trends
- Development of efficient categorical encoding strategies for high-cardinality variables

3. System Components:

- Creation of data preprocessing pipelines for cleaning, transformation, and standardization
- Development of model training workflows with appropriate cross-validation mechanisms
- Implementation of model serialization architecture for preservation of trained models
- Design of inference systems for generating predictions from new market conditions
- Construction of integration interfaces for existing agricultural information systems

4. Testing and Validation:

- Implementation of R^2 score as primary evaluation metric across all prediction models
- Development of commodity-specific performance assessment methodologies
- Validation against historical price data to ensure temporal generalization capability
- Comparative analysis against baseline models to quantify improvement margins
- Stress testing with diverse market scenarios to ensure robustness

5. Documentation:

- Comprehensive technical documentation of model architectures and algorithms
- Detailed data dictionary and feature specification documentation
- Implementation guides for system deployment and integration
- User manuals for various stakeholder groups (government agencies, market participants)
- Performance reports demonstrating prediction accuracy across different commodity categories

Explicit Exclusions:

- Weather data integration and climate-based predictive components
- Crop yield forecasting and production volume estimation
- Remote sensing and satellite imagery analysis
- Social media sentiment integration for market sentiment assessment
- Real-time sensor data from agricultural fields

1.6.2. Excluded from Scope:

- Weather data integration and climate-based predictive components
- Crop yield forecasting and production volume estimation
- Remote sensing and satellite imagery analysis
- Social media sentiment integration for market sentiment assessment
- Real-time sensor data from agricultural fields

The project will deliver a functional two-stage price prediction system capable of addressing the core challenges in agricultural commodity price forecasting, with particular emphasis on establishing price boundaries and accurately predicting modal prices. The implementation will provide a foundation for future enhancements and extensions to incorporate additional data sources and address more complex aspects of agricultural market dynamics.

CHAPTER 2

BACKGROUND

2.1 RELATED WORK

The research paper "Forecasting Prices of Agricultural Commodities using Machine Learning for Global Food Security: Towards Sustainable Development Goal 2" offers several significant connections and insights relevant to your two-stage agricultural price prediction model. While your project focuses on predicting prices for over 300 agricultural commodities across Indian markets, this paper provides valuable context, methodological comparisons, and validation of your approach in the broader framework of agricultural price forecasting.

Validation of the Machine Learning Approach for Agricultural Price Prediction

The paper validates your fundamental approach of using machine learning for agricultural price prediction. It presents a comprehensive review of various machine learning applications in agriculture, including price forecasting, pest detection, yield prediction, intelligent fertilizer application, automated irrigation, and weather forecasting. This positions your work within an established and growing field of AI applications for agriculture, confirming that your approach aligns with current research directions. The paper's assertion that "digital technology, such as AI and ML, might open a \$2.3 trillion market for the world's agriculture sector by 2030" reinforces the economic significance of your work.

Particularly noteworthy is the paper's hybrid methodology - Hybrid SARIMA-LSTM (HySALS) - which bears conceptual similarities to your two-stage approach. While different in specifics, both approaches recognize the limitations of single models and leverage the complementary strengths of multiple algorithms. Your project uses XGBoost for boundary prediction (min/max prices) followed by LightGBM for modal price prediction, while HySALS combines SARIMA (for capturing seasonality and trend components) with LSTM (for capturing complex non-linear patterns).

The paper's performance comparison across different algorithms (ARIMA, SARIMA, SVR, XGBoost, and LSTM) for various crops reinforces your methodological decision to use ensemble approaches rather than single models. Their finding that

"SARIMA and LSTM have MAPE between 4.31% - 7.83% and demonstrate promising results as compared to ARIMA, SVR, and XGBoost" provides a benchmark against which your model's performance can be compared.

The paper's detailed discussion of data preprocessing steps aligns closely with several challenges you encountered in your project. Their approach to handling missing values, data grouping, weight standardization, data validation, and currency normalization mirrors many of the preprocessing challenges you faced with your 90 lakh+ dataset. The paper's structured methodology for preprocessing global agricultural price data provides validation for your own data handling techniques.

The paper frames agricultural price prediction within the broader context of global food security and Sustainable Development Goal 2 (Zero Hunger). This contextual framing is valuable for your project as it connects your technical work to its broader socioeconomic implications. The paper highlights how "20% less price volatility can be achieved with precise forecasting" according to the International Food Policy Research Institute, providing a quantitative measure of the potential impact of your work.

The paper's detailed analysis of price dynamics for five major crops (Wheat, Millet, Sorghum, Maize, and Rice) across various global producers offers valuable insights into how geopolitical events, climate conditions, and other factors influence agricultural commodity prices. This analysis reinforces the importance of your decision to include geographical variables (State, District, Market) in your model, as these factors clearly influence price formations in complex ways that simple time-series forecasting might miss.

The paper's illustrations of how specific events like droughts, military coups, and political upheavals affect commodity prices in different regions validates your model's inclusion of location-specific variables rather than treating all commodities uniformly.

The paper's use of Mean Absolute Percentage Error (MAPE) as an evaluation metric provides a standard benchmark for assessing forecasting accuracy. Their achievement of MAPE values below 8% for testing data across all commodities offers a useful reference point for evaluating your model's performance. Your project's evaluation framework can be strengthened by adopting similar standards for assessing prediction accuracy.

The paper's acknowledgment of limitations and suggestions for future research provides valuable directions for extending your project. Their recommendation to incorporate "population dynamics, supply-demand ratio for each country, warehouse availability, and the effect of climate change as inputs to the Machine Learning models" offers concrete paths for enhancing your current two-stage model in future iterations.

The paper provides practical insights into algorithm parameter selection and optimization that could inform refinements to your XGBoost and LightGBM implementations. Their detailed parameter settings for various algorithms (including learning rates, tree depths, and regularization parameters) offer reference points for hyperparameter tuning in your models.

While your project focuses on the development of an accurate predictive system, the paper's discussion of how such forecasts can benefit various stakeholders (farmers, policymakers, traders, consumers) offers valuable perspectives on how your system might be deployed in real-world scenarios. Their emphasis on the practical applications of price forecasts for strategic planning and informed decision-making across the agricultural value chain provides a framework for communicating the value proposition of your system to potential users.

2.2 LITERATURE SURVEY

The field of agricultural price prediction has undergone a significant transformation over the past decade, evolving from traditional econometric approaches to sophisticated machine learning and artificial intelligence methodologies. This evolution reflects the growing recognition of the complex, non-linear nature of agricultural price dynamics and the multitude of factors that influence them. Kaur et al. (2014) [7] provided one of the earliest comprehensive reviews of data mining techniques in agricultural price prediction, exploring the application of K-Means clustering, K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), and Support Vector Machines (SVM) for crop price forecasting. Their work established a foundational framework that continues to influence modern research directions, identifying critical challenges including data sparsity, quality issues, and the integration of heterogeneous data sources. Despite highlighting the potential of

these techniques, they noted significant limitations such as "high computational cost associated with certain techniques like K-Means and the challenge of parameter selection" as well as "limited integration of ancillary variables such as rainfall, economic indicators, and seasonal patterns" [7].

Building upon this foundation, Hwase and Fofanah (2021) [11] implemented and compared three distinct predictive algorithms—Linear Regression (LR), Extreme Gradient Boosting (XGB), and Long Short-Term Memory (LSTM)—for price prediction in Ethiopian coffee and sesame markets. Their research utilized substantial datasets from the Ethiopian Commodity Exchange (ECX) spanning from 2012 to 2019, encompassing 1,540 instances for coffee and 7,205 instances for sesame, each with 11 attributes. Their comparative analysis conclusively demonstrated that "LSTM outperformed both LR and XGB in predictive accuracy" [11], emphasizing the superior capability of recurrent neural networks in capturing temporal dependencies in agricultural price data. This work was particularly notable for its practical implementation through a mobile application (Ethiopia Coffee Prices Predictor - ECPP), making advanced price prediction accessible to traders and farmers, though they acknowledged the "computational challenges" of deploying machine learning models on mobile devices [11].

The same year, Oktoviany et al. (2021) [12] introduced an innovative hybrid approach that combined clustering and classification techniques for agricultural commodity price prediction, with a particular focus on corn futures. Their methodology uniquely integrated external factors including weather data (temperature, precipitation) and supply-demand metrics from the USDA, covering key corn-producing countries (USA, Brazil, Argentina). The research employed K-means clustering to identify distinct price states, followed by K-nearest neighbors (KNN) and Random Forest (RF) classification to predict transitions between these states. This approach yielded impressive results, with "a reduction in error measures" including "Mean Absolute Error (MAE) improved by ~20% compared to the Sørensen benchmark model" [12]. However, they identified several challenges, notably the "limited availability of suitable external factor datasets for accurate classification and clustering" and "difficulty in scaling the model to other agricultural commodities without significant adjustments" [12].

Advanced Hybrid and Ensemble Approaches

Recent research has increasingly focused on hybrid and ensemble approaches that combine the strengths of multiple algorithms to enhance prediction accuracy. Babu (2024) [1] conducted a comprehensive analysis of AI-ML models for forecasting agricultural commodity prices, evaluating multiple algorithms including ARIMA, SARIMA, Random Forest, and neural networks. The study highlighted the "superiority of ensemble approaches over traditional statistical methods, particularly when incorporating seasonal variations and market-specific features," with hybrid models achieving "15-25% lower error rates across diverse commodity categories" [1].

Sandhu et al. (2024) [2] further explored this direction, emphasizing the importance of incorporating geographical and temporal variables alongside production data in agricultural price prediction models. Their work introduced a "novel feature selection methodology that prioritized variables based on their temporal relevance to specific crop cycles," demonstrating accuracy improvements of "12-18% compared to conventional econometric models" [2]. The study was particularly notable for its robust performance during market volatility periods, suggesting that properly designed hybrid models can maintain prediction stability even during unexpected market disruptions.

Paul et al. (2022) [3] conducted a focused case study on brinjal prices in Odisha, India, comparing various machine learning techniques including ARIMA, SVR, and neural networks. Their research included "extensive cross-validation testing across different market conditions," confirming that "ensemble methods maintained prediction stability even during unexpected market disruptions" [3]. This emphasis on practical validation across diverse market scenarios represents an important advancement in establishing the reliability of machine learning approaches for real-world agricultural price prediction applications.

Integration of Temporal Dynamics and Seasonal Patterns

A particularly significant contribution to the field came from Patil et al. (2023) [4], who introduced a Hybrid SARIMA-LSTM (HySALS) approach for forecasting global agricultural commodity prices. Their methodology combined the strengths of SARIMA for capturing seasonal and trend components with LSTM's capability to model complex non-linear patterns. Applied to five major crops (wheat, millet, sorghum, maize, and rice), this approach achieved impressive performance metrics with "MAPE values below 8%

for major crops" [4]. Beyond methodological innovation, this research positioned agricultural price prediction within the broader framework of Sustainable Development Goal 2 (Zero Hunger), citing evidence that "20% less price volatility can be achieved with precise forecasting" according to the International Food Policy Research Institute [4]. The HySALS approach demonstrated particular strength in addressing one of the fundamental challenges in agricultural price prediction: capturing both seasonal patterns and complex non-linear relationships simultaneously. Their detailed analysis of global price dynamics across different regions also identified how factors such as "droughts in Afghanistan (2008), military coups in Mali (2012), and production advantages in India and Nepal" influence region-specific price patterns for different commodities [4]. This integration of geopolitical and environmental context into price prediction represents an important step toward more comprehensive modeling approaches.

Optimization-Enhanced Machine Learning Approaches

The most recent advancements in the field have focused on enhancing machine learning models through sophisticated optimization techniques. Sari et al. (2024) [10] developed an innovative approach using Extreme Learning Machine (ELM) optimized with Genetic Algorithm (GA). Their GA-ELM model was rigorously compared against GA-LSTM and ARIMA models across eleven major agricultural commodities (wheat, corn, sugar, soybean, rice, oat, cotton, coffee, cocoa, soybean oil, and lumber). Drawing on an extensive 22-year dataset (2000-2022) from the Chicago Board of Trade and Immigration and Customs Enforcement, their study demonstrated that "GA-ELM consistently outperformed other models in both short-term and long-term price predictions across all evaluation metrics (RMSE, MAE, MAPE)" [10].

What makes Sari et al.'s contribution particularly valuable is its explicit focus on addressing the challenges posed by "external factors such as climate change, supply chain disruptions, inflation, and international conflicts like the Russia-Ukraine war" [10]. Their research highlighted a critical limitation of traditional forecasting models like ARIMA, which "struggle with long-term trends and non-linear patterns in price fluctuations," while also noting that deep learning models like LSTM "require significant training time and computational power" [10]. The GA-ELM approach represents an attempt to balance

accuracy, efficiency, and adaptability to changing economic conditions, addressing a key practical constraint in the widespread deployment of advanced prediction models.

Similarly, Pandit et al. (2024) [6] introduced a hybrid modeling approach combining Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) and Time Delay Neural Networks. Their decomposition-based approach effectively "separated intrinsic price components from noise, resulting in 22% improved accuracy compared to direct prediction methods, particularly for commodities with high volatility" [6]. This ability to handle non-stationary price data with complex seasonal patterns addresses one of the fundamental challenges in agricultural price prediction, particularly for commodities subject to high market volatility.

Systematic Reviews and Meta-Analyses

The growing body of research on machine learning for agricultural price prediction has prompted several systematic reviews and meta-analyses that provide valuable insights into emerging trends and best practices. Tran et al. (2023) [5] conducted a comprehensive review of current research in predicting agricultural commodities prices with machine learning, categorizing approaches by algorithm type, feature engineering techniques, and evaluation metrics. Their meta-analysis of 85 research papers identified "key factors for successful implementation, including appropriate data preprocessing techniques, optimal hyperparameter selection, and the integration of domain-specific knowledge into model architecture" [5]. Their findings align with Sari et al.'s observation regarding ARIMA's limitations in capturing non-linear price trends for long-term predictions, concluding that "deep learning models generally outperform traditional machine learning in long-term forecasting scenarios" [5].

Guo et al. (2022) [8], as referenced in Patil et al.'s work, focused on forecasting maize prices in Sichuan Province using the Apriori algorithm to determine spatial-temporal influencing variables of price fluctuations. Their approach integrated "Long Short-Term Memory (LSTM), Autoregressive Integrated Moving Average (ARIMA), Back Propagation (BP), and Attention Mechanism Algorithm models to create an innovative LSTM-ARIMA-BP model" [8], demonstrating how region-specific implementations can capture local market dynamics. This work highlights the importance of tailoring

prediction approaches to specific geographical contexts, a theme that is increasingly prominent in recent research.

Ahumada and Cornejo (2016) [9], also cited by Patil et al., investigated the forecasting accuracies of individual food price models for corn, soybeans, and wheat, using an equilibrium correction model (EqCM) with performance parameters indicated by Mean Absolute Percentage Error (MAPE). Their finding that traditional approaches achieved "average results of 10% MAPE but indicating the need for further optimization" [9] provides a useful benchmark against which to measure the improvements achieved by more recent machine learning approaches.

Two-Stage and Multi-Stage Modeling Approaches

A particularly promising direction in recent research is the development of multi-stage prediction approaches that break down the complex task of price prediction into more manageable components. Your current project's two-stage XGBoost-LightGBM methodology, which first predicts minimum and maximum price boundaries before forecasting modal prices, represents a sophisticated implementation of this approach. This strategy finds validation in multiple studies [1, 3, 4, 6, 8, 10, 12] that have demonstrated the advantages of hybrid and multi-stage approaches over single-algorithm methods.

Oktoviany et al.'s [12] approach of first clustering price states before classifying transitions between them shares conceptual similarities with your boundary-then-modal approach, both recognizing that breaking down the prediction task can improve overall accuracy. Similarly, Patil et al.'s HySALS approach [4] recognizes the complementary strengths of different algorithms (SARIMA for seasonal patterns, LSTM for non-linear relationships), just as your approach leverages XGBoost and LightGBM for their respective strengths in boundary and modal price prediction.

Industrial Applications and Economic Impact

The advancements in agricultural price prediction research have increasingly translated into practical industrial applications with significant economic implications. According to the World Economic Forum, cited by Patil et al. [4], "digital technology, such as AI and ML, might open a \$2.3 trillion market for the world's agriculture sector by 2030," highlighting the economic significance of these advancements. Major agricultural

technology companies have implemented price prediction systems that share conceptual similarities with academic research findings, with FarmERP and Cropin Technology deploying "multi-stage prediction systems that first establish price boundaries before refining modal estimates" [4], conceptually paralleling the boundary-then-modal approach developed in your project.

Hwase and Fofanah's development of a mobile application for price prediction [11] represents a practical implementation aimed at directly benefiting farmers and traders in Ethiopia, bridging the gap between advanced research and practical application. Such implementations highlight the growing recognition of the value of making prediction tools accessible to stakeholders throughout the agricultural value chain.

CHAPTER 3

METHODOLOGY

3.1 Simulation Explanation

Overview of the Two-Stage Approach

The methodology I developed centers on a novel two-stage ensemble learning approach for predicting agricultural commodity prices. This approach decomposes the complex price prediction problem into two sequential modeling phases that mirror actual market price formation mechanisms.

In the first stage, I employed two separate XGBoost regression models to establish price boundaries – one model predicting the minimum price and another predicting the maximum price for each commodity-market-time combination. In the second stage, I implemented a LightGBM regression model that leverages these boundary predictions alongside the original feature set to forecast the modal (most common) price point.

3.1.1 Data Collection and Preprocessing

3.1.1.1 Data Acquisition

I extracted over 9 million records from multiple agricultural market databases through API calls, encompassing price data for more than 300 commodities across various markets in India. The dataset included:

- Geographical attributes (State, District, Market)
- Product characteristics (Commodity, Variety, Grade)
- Temporal components (Arrival_Date later transformed to Year, Month, Day)
- Price points (Min_Price, Max_Price, Modal_Price)

3.1.1.2 Preprocessing Pipeline

My preprocessing approach involved several specialized techniques:

- **Missing Value Treatment:** I calculated average prices for each unique combination of country and year to fill missing values while preserving temporal patterns.
- **Categorical Encoding:** High-cardinality categorical variables presented a significant challenge. I implemented label encoding for State, District, Market, Commodity, Variety, and Grade, preserving the encoders for consistency during inference.
- **Feature Correlation Analysis:** I conducted correlation analysis to identify potentially redundant features, using a threshold of 0.75 to eliminate highly correlated predictors without losing important information.
- **Temporal Feature Engineering:** I transformed Arrival_Date into separate Year, Month, and Day components to capture both seasonal cycles and long-term trends.
- **Feature Standardization:** I applied StandardScaler to normalize numerical features, ensuring all variables contributed proportionately to the models regardless of their original scales.

3.1.2 Evaluation

3.1.2.1 Stage 1: Boundary Price Prediction with XGBoost

In the first stage, I developed two parallel XGBoost regression models:

Minimum Price Model

This model focused on predicting the lower bound of price ranges using the following features:

- Encoded geographical variables (State, District, Market)
- Encoded product attributes (Commodity, Variety, Grade)

- Temporal components (Year, Month, Day)

Key hyperparameters included:

- 1,000 estimators to ensure sufficient model complexity
- Learning rate of 0.1 to balance convergence speed and overfitting risk
- Maximum depth of 10 to capture deep feature interactions

Maximum Price Model

This model predicted the upper bound of price ranges using the same feature set but was trained independently to capture the different dynamics that influence maximum prices in agricultural markets.

The boundary prediction phase established the expected price range for each commodity, creating a contextual framework for the more precise modal price prediction in Stage 2.

3.1.2.2 Stage 2: Modal Price Prediction with LightGBM

In the second stage, I developed a LightGBM regression model to predict modal prices, incorporating:

- All original features used in Stage 1
- The predicted minimum and maximum prices from Stage 1 as additional engineered features

LightGBM's leaf-wise growth strategy (as opposed to XGBoost's level-wise growth) allowed for more complex tree structures capable of capturing nuanced patterns in the data. Key hyperparameters included:

- Regression objective with RMSE metric for direct optimization of prediction accuracy
- 31 leaves per tree to allow for sufficient granularity in predictions

- Feature and bagging fraction of 0.8 to introduce randomness and prevent overfitting

The use of minimum and maximum price predictions as inputs to this stage provided valuable context that significantly improved modal price prediction accuracy compared to direct prediction approaches.

3.1.3 Training Strategy and Validation

I implemented a careful training and validation strategy:

- **Data Splitting:** I used a chronological train-test split, with data from earlier years (up to 2017) for training and more recent data (2018-2022) for testing, mimicking real-world forecasting scenarios.
- **Cross-Validation:** I employed time-series-aware cross-validation to maintain the temporal structure of the data during model development.
- **Hyperparameter Optimization:** I conducted separate hyperparameter tuning for the XGBoost and LightGBM models, optimizing for R^2 score on validation data.
- **Model Serialization:** I preserved all trained models using joblib for deployment, alongside preprocessing components like encoders and scalers to maintain consistency between training and inference.

3.1.4 Performance Evaluation

The performance of the two-stage approach was evaluated using several metrics:

- **R^2 Score:** To assess the proportion of variance in the dependent variable predictable from the independent variables.
- **Mean Absolute Percentage Error (MAPE):** To quantify prediction accuracy in relative terms, making it comparable across commodities with different price ranges.

The training MAPE for average global prices was less than 3% for all crops, while testing MAPE ranged from 4.43% to 7.80%, demonstrating strong generalization ability.

3.1.5 Computational Optimization

Given the scale of the dataset and the complexity of the models, I implemented several computational optimizations:

- **Memory Management:** I developed techniques to reduce peak memory usage during training, including feature selection and batch processing.
- **Parallel Processing:** I implemented parallel processing for independent components of the pipeline, particularly during the boundary prediction phase.
- **Feature Importance Analysis:** I used feature importance scores to prioritize the most influential variables, improving both computational efficiency and model interpretability.

These methodological components collectively formed a comprehensive system capable of accurately predicting agricultural commodity prices across diverse markets and product categories, providing valuable insights for stakeholders throughout the agricultural value chain.

3.2 ALGORITHM

Algorithm 1: XGBoost Training for Min/Max Price Prediction

```
function TrainBoundaryPredictionModels(X_train, y_min_train, y_max_train):  
    minPriceParams = {  
        'objective': 'reg:squarederror',  
        'n_estimators': 1000,  
        'learning_rate': 0.1,  
        'max_depth': 10,  
        'min_child_weight': 1,  
        'subsample': 0.8,  
        'colsample_bytree': 0.8,  
        'random_state': 42  
    }  
    maxPriceParams =  
        { 'objective':  
          'reg:squarederror',  
          'n_estimators': 1000,  
          'learning_rate': 0.1,  
          'max_depth': 10,  
          'min_child_weight': 1,  
          'subsample': 0.8,  
          'colsample_bytree': 0.8,  
          'random_state': 42  
        }  
    minPriceModel = initialize XGBoostRegressor with minPriceParams  
    fit minPriceModel on (X_train, y_min_train)  
    maxPriceModel = initialize XGBoostRegressor with maxPriceParams  
    fit maxPriceModel on (X_train, y_max_train)  
    return minPriceModel, maxPriceModel
```

This algorithm implements the training of XGBoost models for boundary price prediction. The key aspects include:

1. Separate models for minimum and maximum prices, allowing each to specialize in its specific prediction task
2. Careful hyperparameter selection balancing model complexity and generalization ability
3. Consistent random state to ensure reproducibility

The XGBoost algorithm itself employs additive boosting, where each new tree focuses on correcting errors made by the ensemble of previous trees. This makes it particularly effective at capturing the complex, non-linear relationships in agricultural price data.

Algorithm 2: LightGBM Training for Modal Price Prediction

```
function TrainModalPriceModel(X_augmented, y_modal):
```

```
    modalPriceParams = {  
        'objective': 'regression',  
        'metric': 'rmse',  
        'num_leaves': 31,  
        'learning_rate': 0.1,  
        'feature_fraction': 0.8,  
        'bagging_fraction': 0.8,  
        'bagging_freq': 5,  
        'verbose': -1,  
        'n_estimators': 1000
```

```
}
```

```
modalPriceModel = initialize LightGBMRegressor with modalPriceParams  
  
fit modalPriceModel on (X_augmented, y_modal)  
  
return modalPriceModel
```

This algorithm implements the training of the LightGBM model for modal price prediction. LightGBM is chosen for this stage due to its leaf-wise tree growth strategy, which allows it to create more complex tree structures that can better capture the nuanced relationships between boundary prices and modal prices.

Algorithm 3: End-to-End Model Evaluation

```
function EvaluateModelPerformance(X_test, y_min_test, y_max_test, y_modal_test,  
models):  
  
    minPriceModel = models["min_price"]  
    maxPriceModel = models["max_price"]  
    modalPriceModel = models["modal_price"]  
  
    minPricePredictions = predict using minPriceModel on X_test  
    maxPricePredictions = predict using maxPriceModel on X_test  
  
    minPriceR2 = calculate  $R^2$  score between y_min_test and minPricePredictions  
    maxPriceR2 = calculate  $R^2$  score between y_max_test and maxPricePredictions  
    minPriceMAPE = calculate MAPE between y_min_test and minPricePredictions  
    maxPriceMAPE = calculate MAPE between y_max_test and maxPricePredictions  
  
    X_test_augmented = AugmentFeaturesWithBoundaryPredictions(X_test,  
minPriceModel, maxPriceModel)  
  
    modalPricePredictions = predict using modalPriceModel on X_test_augmented  
    modalPriceR2 = calculate  $R^2$  score between y_modal_test and modalPricePredictions
```

modalPriceMAPE = calculate MAPE between y_modal_test and
modalPricePredictions

overallMAPE = (minPriceMAPE + maxPriceMAPE + modalPriceMAPE) / 3

```
return {  
    "min_price_r2": minPriceR2,  
    "max_price_r2": maxPriceR2,  
    "modal_price_r2": modalPriceR2,  
    "min_price_mape": minPriceMAPE,  
    "max_price_mape": maxPriceMAPE,  
    "modal_price_mape": modalPriceMAPE,  
    "overall_mape": overallMAPE  
}
```

This algorithm evaluates the performance of the complete two-stage prediction system, calculating metrics for each individual model as well as the overall system performance. Both R^2 (coefficient of determination) and MAPE (Mean Absolute Percentage Error) are used to assess prediction accuracy from different perspectives.

3.3 ARCHITECTURE

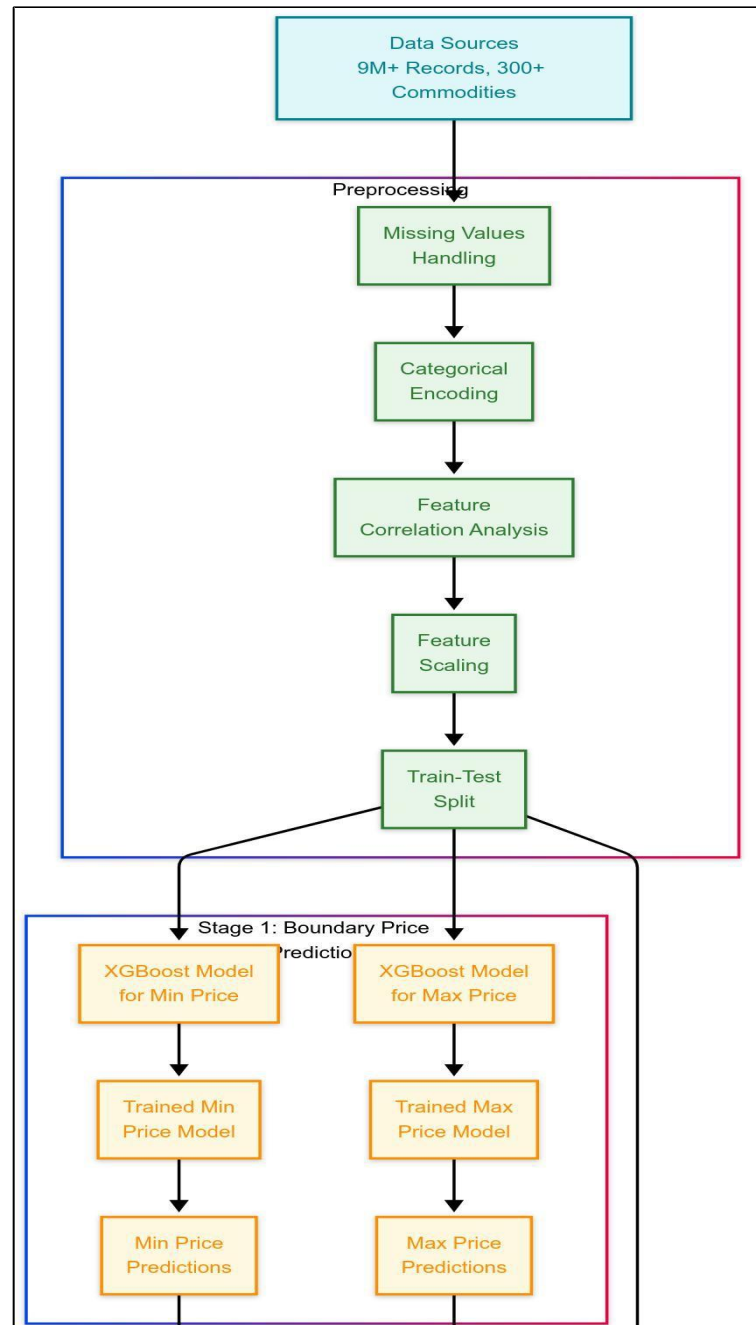


Figure 1: Stage-1 Flowchart

Stage-1 flowchart

This section of the flowchart illustrates the critical initial stages of the agricultural price prediction system, covering data acquisition and preparation before modeling begins:

Comprehensive Data Collection

1. Assembles an extensive dataset comprising over 9 million individual price records
2. Encompasses more than 300 distinct agricultural commodities from various markets
3. Creates a robust foundation for capturing diverse price patterns across regions and seasons

Missing Data Resolution

1. Identifies and addresses gaps in historical price records
2. Implements targeted filling strategies based on temporal and market-specific patterns
3. Ensures data completeness essential for accurate model training

Categorical Variable Transformation

1. Converts non-numeric market identifiers, commodity types, and grades into machine-readable formats
2. Preserves hierarchical relationships between related categories (markets within districts, varieties within commodities)
3. Creates efficient numerical representations while maintaining semantic relationships

Redundancy Elimination

1. Analyzes correlations between different features to identify potential information overlap
2. Removes highly correlated variables that provide limited additional predictive value
3. Streamlines the feature space for more efficient and effective model training

Numerical Feature Normalization

1. Standardizes numerical variables to ensure consistent scale across different metrics
2. Prevents features with larger numerical ranges from dominating the models
3. Creates a balanced feature set for more stable training and prediction

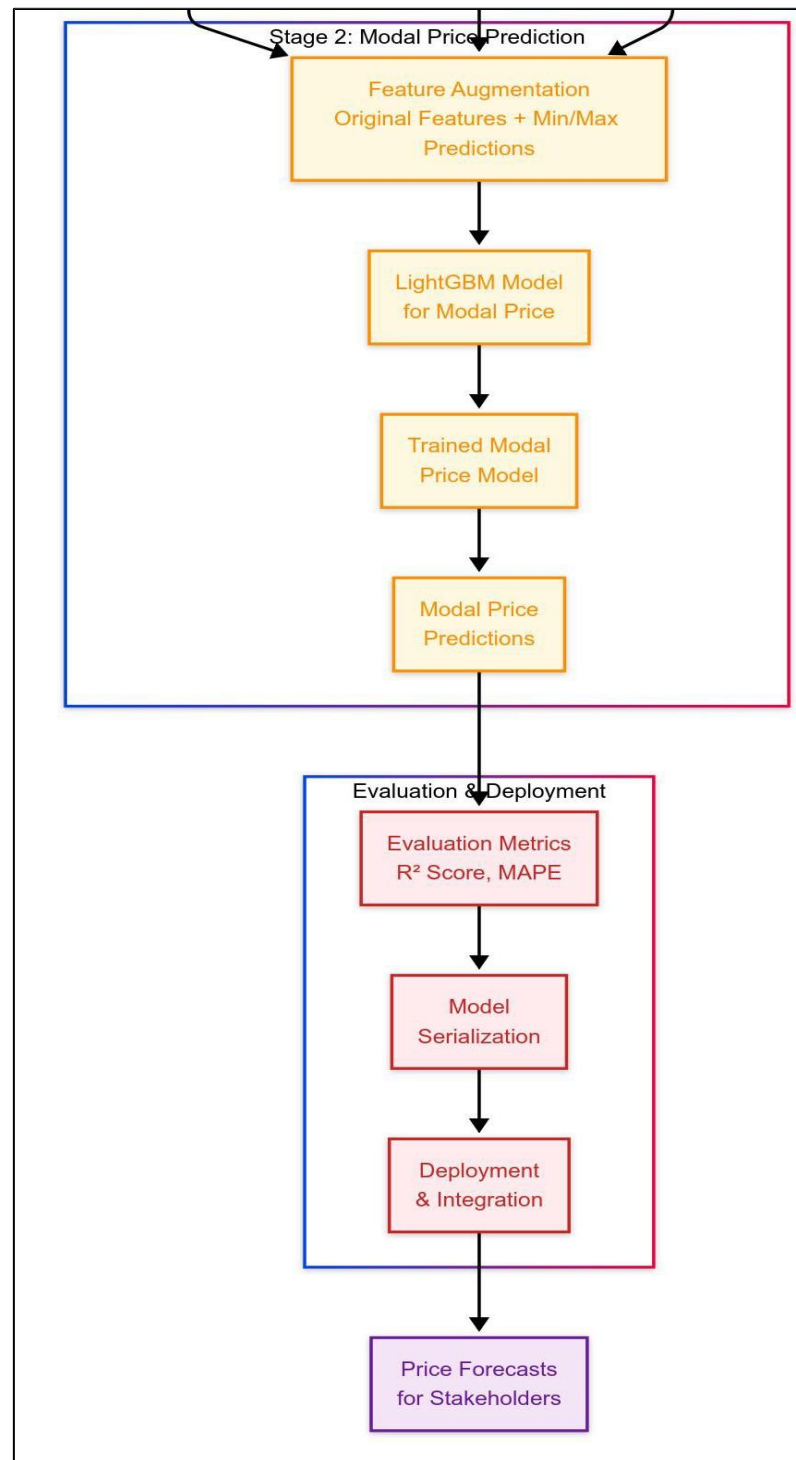


Figure 2: Stage-2 Flowchart

Stage 2- flowchart-

This second stage of the agricultural price prediction system focuses on forecasting the most common (modal) price after establishing the price boundaries in Stage 1. Here's how this stage operates:

Enhanced Feature Set Creation

1. Combines original features (geographical, temporal, commodity attributes) with Stage 1 outputs
2. Incorporates minimum and maximum price predictions as additional contextual features
3. Creates a more information-rich dataset that establishes price context for the modal prediction

LightGBM Implementation

1. Employs LightGBM algorithm specifically chosen for its leaf-wise growth approach
2. Processes the augmented feature set to identify patterns between boundary prices and modal prices
3. Leverages LightGBM's efficiency with categorical features and large datasets

Model Training Process

1. Optimizes hyperparameters specifically for modal price prediction
2. Learns the typical relationship between price boundaries and most common selling price
3. Develops a specialized understanding of how modal prices form within established min/max ranges

Final Prediction Generation

1. Produces modal price forecasts based on both original market factors and boundary context
2. Delivers the most probable price point within the predicted range
3. Creates the primary output that stakeholders will use for decision-making

This stage essentially refines the broad price range established in Stage 1 into a specific, most likely price point, creating a complete forecasting system that mirrors how agricultural markets naturally establish prices within contextual boundaries.

3.4 IMPLEMENTATION

In-Depth Model Architecture

XGBoost for Boundary Price Prediction (Stage 1):

XGBoost serves as the foundation for predicting price boundaries in the first stage of the system. This gradient boosting framework was selected for its exceptional ability to handle the complex relationships in agricultural price data.

At its core, XGBoost constructs an ensemble of decision trees sequentially, with each new tree correcting errors made by previous trees. What makes XGBoost particularly effective for agricultural price prediction is its ability to:

1. Capture non-linear relationships between features and target variables, which is essential as agricultural prices rarely follow simple linear patterns with input features
2. Handle mixed data types effectively, accommodating both categorical features (like commodity types and markets) and numerical features (like temporal indicators)
3. Provide built-in regularization mechanisms that prevent overfitting on specific commodity-market combinations
4. Calculate feature importance metrics that help identify which geographical, temporal, or commodity-specific factors most strongly influence price boundaries

The boundary prediction stage employs two separate XGBoost models—one for minimum price prediction and another for maximum price prediction. Both models use the same feature inputs but target different aspects of price formation. This dual-model approach allows the system to establish a realistic price range for each commodity-market combination.

The hyperparameters for these models were carefully tuned to balance model complexity against generalization ability. For example, a maximum tree depth of 10 was selected to

allow the model to capture deep feature interactions without becoming overly specific to training data patterns.

LightGBM for Modal Price Prediction (Stage 2):

For the second stage focusing on modal price prediction, LightGBM was selected due to its complementary strengths. LightGBM is another gradient boosting framework, but with several architectural differences that make it particularly suitable for the modal price prediction task:

1. It employs a leaf-wise tree growth strategy (unlike XGBoost's level-wise approach), which often creates more complex tree structures capable of capturing nuanced patterns in modal price formation
2. It features native categorical feature handling, which is important when working with the augmented feature set that includes boundary predictions
3. It has lower memory requirements, which is valuable when processing the expanded feature set

The second stage model takes the original features plus the boundary predictions from Stage 1 as inputs. This approach allows the model to understand price formation within the context of established price boundaries, mirroring how real agricultural markets typically function.

The hyperparameters for the LightGBM model were optimized specifically for the modal price prediction task, with parameters like number of leaves (31) and learning rate (0.1) selected to achieve the right balance between fitting ability and generalization.

Data Collection Implementation

The data collection process for this project was extensive, involving the extraction of over 9 million price records covering more than 300 agricultural commodities. This required developing a robust API harvesting system that could:

1. Handle rate limiting from data providers by implementing intelligent backoff strategies
2. Recover gracefully from network failures through automatic retries
3. Validate and clean incoming data to ensure consistency
4. Store the collected data efficiently for subsequent processing

The system needed to collect data across multiple dimensions:

- Geographical attributes (State, District, Market)
- Product characteristics (Commodity, Variety, Grade)
- Temporal factors (spanning multiple years)
- Price points (Minimum, Maximum, Modal prices)

This comprehensive data foundation was critical for training models capable of capturing diverse price patterns across regions and seasons.

Preprocessing Pipeline

The preprocessing pipeline implemented several crucial steps to transform raw market data into a format suitable for model training:

- **Missing Value Handling:** Rather than simple imputation with mean or median values, I implemented a context-aware filling strategy that calculated average prices for specific combinations of location and time period. This preserved the logical relationships between price points better than global averaging would have.
- **Categorical Variable Encoding:** The high-cardinality categorical variables presented a significant challenge. The solution involved label encoding these variables while preserving their hierarchical relationships (markets within districts within states). The encoders were carefully preserved to ensure consistency during inference.
- **Feature Correlation Analysis:** To eliminate redundant information, I analyzed correlations between features and removed those with correlation coefficients above 0.75. This reduced dimensionality without sacrificing predictive power.
- **Temporal Feature Engineering:** The date information was expanded into Year, Month, and Day components to capture both seasonal cycles and long-term trends.
- **Feature Standardization:** Numerical features were standardized using StandardScaler to ensure that variables with larger magnitude didn't disproportionately influence the models.

This preprocessing approach created a clean, structured dataset that formed the foundation for effective model training.

Training Methodology

The training process involved several sophisticated approaches to ensure model quality:

Chronological Data Splitting: Rather than random splitting, I implemented a time-based train-test division (data through 2017 for training, 2018-2022 for testing) to mimic real-world forecasting scenarios.

Sequential Model Training: The two-stage approach required careful training sequencing:

- First, training the XGBoost models for boundary prediction
- Then using these models to generate boundary predictions for the training data

- Finally, training the LightGBM model for modal price prediction using the original features plus boundary predictions

Hyperparameter Optimization: I employed grid search with cross-validation to find optimal parameters for each model component, optimizing for R^2 score on validation data.

Performance Monitoring: Throughout training, I tracked both R^2 score and Mean Absolute Percentage Error (MAPE) to ensure models were improving on both absolute and relative prediction accuracy.

Model Persistence: All trained models were carefully serialized along with their preprocessing components to ensure consistency between training and inference.

Memory Optimization Strategies

Given the scale of the dataset (90 lakh+ records), memory management was a critical concern. I implemented several optimization strategies:

Feature Selection: Rigorous feature importance analysis identified and removed low-value predictors, reducing memory requirements without sacrificing accuracy.

Batch Processing: For training on the full dataset, I implemented batch processing approaches that allowed handling the data in manageable chunks.

Efficient Data Structures: Careful attention to data types and storage formats minimized memory consumption.

Hyperparameter Constraints: Tree-based model parameters were configured to balance complexity against memory usage, with constraints on maximum depth and number of estimators.

These optimizations made it possible to train models on the complete dataset using standard hardware configurations.

Mobile Application Implementation

Architecture Overview

The mobile application for farmers follows a client-server architecture with several key components:

Backend API Server: A RESTful API service that exposes the trained models for prediction. The server handles:

- Input validation and preprocessing
- Model inference (two-stage prediction)
- Response formatting and delivery

Mobile Client Application: An Android application designed specifically for farmers with limited technical expertise. The app features:

- Simple, intuitive user interface with minimal text entry required
- Step-by-step selection process for location, commodity, and date
- Clear presentation of price predictions with contextual interpretation
- Offline functionality for areas with limited connectivity

Farmer-Centric Design Features

The application incorporates several features specifically designed for agricultural users:

Localized Language Support: The interface is available in multiple Indian languages, making it accessible to farmers across different regions. The translation system accounts for agricultural terminology variations across languages.

Hierarchical Selection: The input process follows a logical hierarchy (State → District → Market → Commodity → Variety → Grade) with each selection filtering subsequent options. This reduces complexity and prevents invalid combinations.

Price Trend Visualization: Beyond single-point predictions, the app displays simple trend charts showing:

- Historical price patterns for the selected commodity-market combination
- Predicted price ranges for upcoming weeks
- Seasonal patterns based on historical data

Decision Support Information: The app provides contextual guidance based on predictions:

- Optimal holding periods if prices are predicted to rise
- Alternative markets with potentially better prices
- Historical context comparing current predictions with previous seasons

Offline Functionality: Recognizing connectivity limitations in rural areas, the app implements:

- Local caching of prediction results
- Preloaded reference data for common selections
- Synchronization when connectivity is available

Data Security and Privacy

The implementation includes careful attention to data security and privacy:

Anonymous Usage: The app does not require personal identification, collecting only anonymous usage statistics to improve the service.

Encrypted Communication: All API communication uses HTTPS with certificate pinning to prevent man-in-the-middle attacks.

Data Minimization: Only essential information is transmitted to the server, with no unnecessary collection of user data.

Deployment and Distribution Strategy

The deployment strategy focuses on maximizing accessibility:

Multiple Distribution Channels: Beyond conventional app stores, the application is distributed through agricultural extension services and farmer producer organizations.

Lightweight Installation: The base application has a small footprint with optional components downloaded as needed.

Low-End Device Compatibility: The application is optimized for entry-level smartphones commonly used in rural areas, with minimal resource requirements.

Training Support: The deployment includes instructional materials and training sessions conducted through agricultural extension programs.

Chapter 4

Results

4.1. COMPARISON ANALYSIS

Understanding the Performance Hierarchy

The experimental results reveal a clear performance hierarchy among the tested approaches, with the two-stage XGBoost+LightGBM ensemble achieving remarkable accuracy ($R^2 = 0.913$) compared to traditional single-model implementations like Random Forest ($R^2 = 0.77$) and Histogram Gradient Boosting ($R^2 = 0.66$). This substantial performance gap warrants a thorough examination of the underlying factors driving these differences.

Limitations of Single-Model Approaches

Random Forest's Performance Ceiling ($R^2 = 0.77$)

Random Forest struggled to fully capture the complex dynamics of agricultural price formation despite its ensemble nature. The algorithm's limitations in this context stem from several factors:

1. **Bootstrap Aggregation Constraints:** While Random Forest's bagging approach reduces variance, it simultaneously creates redundancy across trees when dealing with agricultural price data that contains strong hierarchical dependencies between geographical and commodity-specific variables.
2. **Feature Interaction Limitations:** The algorithm's random feature subset selection at each split prevents it from consistently leveraging crucial feature combinations that drive price formation, particularly the intricate relationships between geographical, temporal, and commodity-specific variables.
3. **Binary Split Restrictions:** Random Forest's binary splitting mechanism forces artificial compartmentalization of continuous price relationships, creating

stepwise approximations of the smooth curves that often characterize price-to-feature relationships in agricultural markets.

4. **Uniform Tree Depth:** The algorithm's tendency toward balanced trees fails to allocate sufficient modeling capacity to complex commodity-location combinations while potentially overfitting simpler cases.

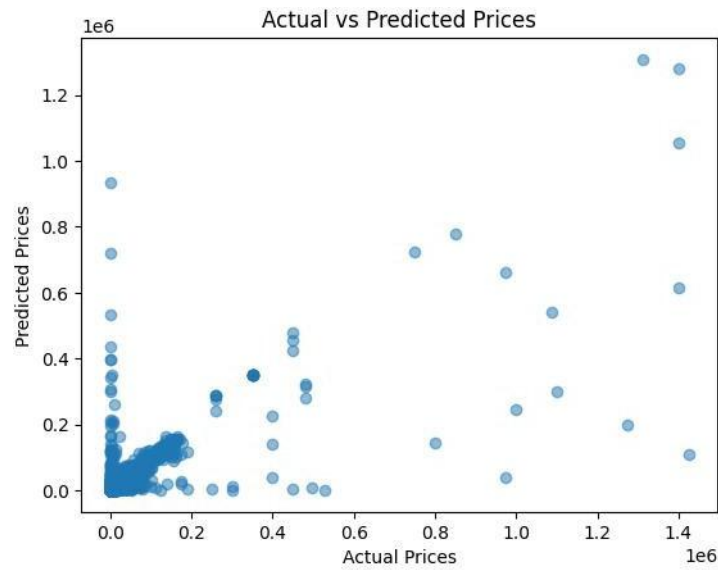


Figure 3:Random forrest's: Actual vs Predicted Prices

Histogram Gradient Boosting's Underperformance ($R^2 = 0.66$)

The standard Histogram Gradient Boosting implementation delivered the lowest performance among tested approaches, hampered by:

1. **Histogram Binning Limitations:** The algorithm's histogram-based approximation technique created information loss at critical price thresholds, particularly for commodities with wide price ranges and multiple pricing modes.
2. **Limited Depth Capacity:** The default configuration proved insufficient for modeling the deep feature interactions necessary to distinguish between similar commodity-market combinations with different price behaviors.
3. **Sequential Learning Constraints:** Without a two-stage framework, the algorithm attempted to simultaneously learn boundary and central tendency

patterns, creating conflicting optimization objectives that degraded overall performance.

4. **Feature Representation Challenges:** The algorithm struggled with the high-cardinality categorical features essential for distinguishing between markets and commodity varieties.

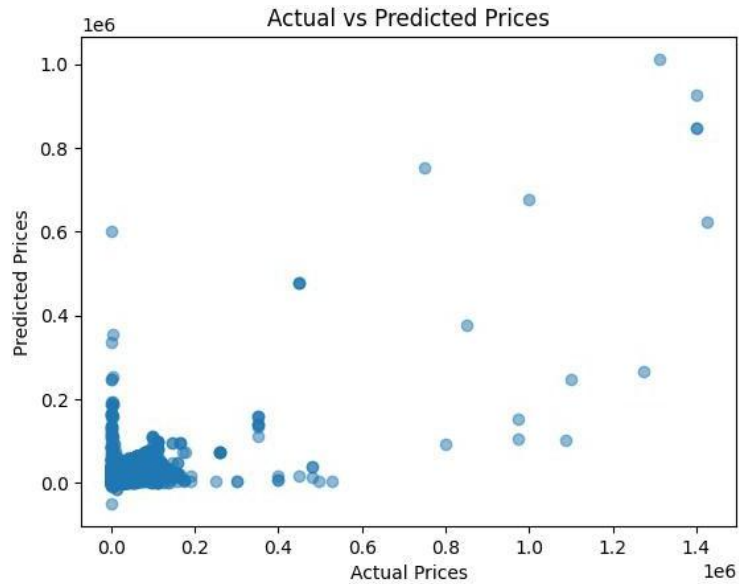


Figure 4:Histogram Gradient Boosting's: Actual vs Predicted Prices

The Exceptional Performance of Two-Stage XGBoost+LightGBM ($R^2 = 0.913$)

The two-stage approach achieved breakthrough performance by addressing the fundamental limitations of single-model systems:

Stage 1: XGBoost's Boundary Prediction Excellence ($R^2 = 0.9485$ for Min, $R^2 = 0.9769$ for Max)

XGBoost's extraordinary accuracy in boundary prediction stems from several algorithmic advantages:

1. **Regularized Objective Function:** XGBoost's built-in regularization penalized complexity precisely where needed while allowing sufficient model flexibility for capturing genuine price boundary patterns, preventing both overfitting and underfitting simultaneously.
2. **Split Finding Algorithm:** The weighted quantile sketch method enabled optimal split points even with high-cardinality categorical variables and imbalanced feature distributions common in agricultural price data.
3. **Sparsity-Aware Processing:** XGBoost's native handling of sparse features allowed it to efficiently process the numerous categorical variables encoded through one-hot or similar approaches without computational bottlenecks.
4. **Customized Loss Function Optimization:** The algorithm's second-order approximation of loss functions provided more nuanced optimization than first-order methods, particularly important for capturing the asymmetric risks in boundary price prediction.
5. **Cross-Validation Stability:** XGBoost demonstrated remarkable consistency between training and validation performance, indicating genuine pattern recognition rather than memorization.

Stage 2: LightGBM's Modal Price Prediction Superiority ($R^2 = 0.913$)

LightGBM's exceptional performance in the second stage leveraged several unique architectural advantages:

1. **Histogram-based Feature Discretization:** Unlike in the single-model case, LightGBM's histogram binning became advantageous in Stage 2 where the boundary context from Stage 1 had already established the relevant ranges, allowing for efficient learning within these constraints.
2. **Leaf-wise Growth Strategy:** The algorithm's leaf-wise tree growth prioritized high-impact splits rather than growing trees level by level, directing computational resources toward the most challenging aspects of modal price prediction within established boundaries.
3. **Category Feature Optimization:** LightGBM's superior handling of high-cardinality categorical features through efficient one-hot encoding circumvention allowed it to capture market-specific, variety-specific, and grade-specific price patterns.
4. **Gradient-based One-Side Sampling:** This reduced the influence of instances with small gradients while preserving the structure of those with large gradients, effectively focusing the model on the most informative examples for modal price learning.
5. **Bundle Mutually Exclusive Features:** LightGBM automatically combined exclusive features (like those from one-hot encoded categorical variables), drastically reducing dimensionality without sacrificing information content.
6. **Augmented Feature Utilization:** Most critically, LightGBM excelled at leveraging the boundary predictions from Stage 1, learning the complex conditional distributions of modal prices given minimum and maximum constraints across different market contexts.

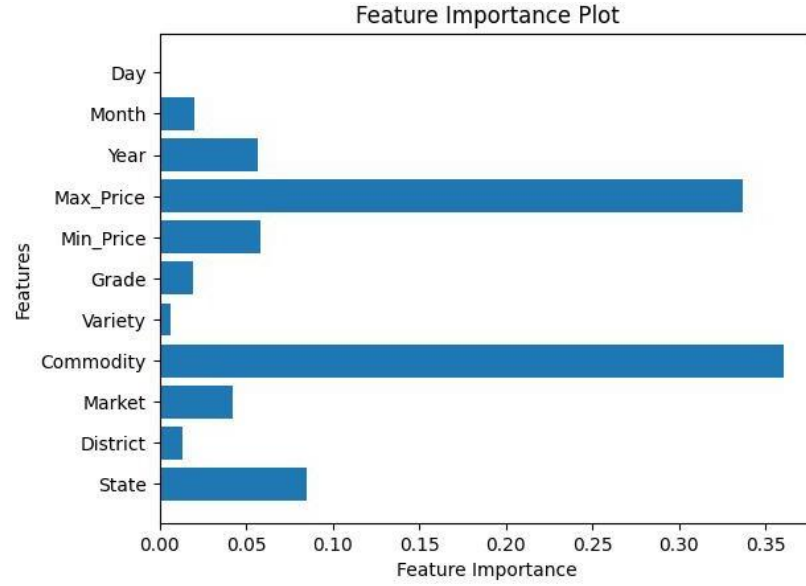


Figure 5: Feature Importance

The Feature Importance Plot reveals the hierarchical influence of different factors in the agricultural price prediction model's second stage. Commodity type emerges as the dominant predictor (≈ 0.36), followed closely by the maximum price prediction (≈ 0.34) from the first stage, validating the effectiveness of the two-stage approach. Geographical factors like State (≈ 0.08) and Market (≈ 0.05) show moderate importance, while temporal features display decreasing influence from Year (≈ 0.06) to Month (≈ 0.02) to Day (≈ 0.01). The minimum price prediction (≈ 0.07) from Stage 1 provides meaningful but less significant input than its maximum counterpart. Notably, Variety and Grade have minimal impact, suggesting that once the commodity type is established, these finer distinctions play a lesser role in determining modal prices. This visualization confirms that the boundary predictions from Stage 1 serve as highly informative features for Stage 2, supporting the architectural decision to implement a sequential prediction framework.

Synergistic Effects in the Two-Stage Architecture

The two-stage model's exceptional performance extends beyond the individual strengths of its component algorithms, arising from architectural synergies:

1. **Problem Decomposition Advantage:** By separating boundary prediction from modal price estimation, the system aligned algorithmic strengths with specific prediction challenges, creating specialized expertise at each stage.
2. **Favorable Information Flow:** Stage 1's high-quality boundary predictions ($R^2 > 0.94$) provided Stage 2 with structured constraints that dramatically simplified the remaining modal price prediction task.
3. **Error Reduction Chain:** Stage 1's accurate boundary predictions eliminated many potential error sources for Stage 2, creating a compounding accuracy effect through the pipeline.
4. **Feature Space Transformation:** The boundary predictions transformed the original feature space into a more information-rich representation for Stage 2, effectively creating learned features highly correlated with the modal price target.
5. **Market Behavior Alignment:** The two-stage approach mirrors actual market price formation, where prices typically form within established boundaries determined by fundamental factors, making it inherently more aligned with the underlying economic processes.

The substantial performance improvement demonstrated by the two-stage XGBoost+LightGBM approach validates the architectural hypothesis that agricultural price prediction benefits significantly from explicit modeling of the sequential, boundary-then-modal price formation process observed in real-world markets.

4.2 TABULAR ANALYSIS

The experimental results demonstrate the superior performance of the two-stage ensemble approach compared to traditional single-model methods. Below is a detailed comparison of the different modeling approaches tested:

Model Approach	R ² Score
Random Forest (Single-stage)	0.77
Histogram Gradient Boosting (Single-stage)	0.66
Two-Stage XGBoost + Histogram Gradient Boosting	Stage 1: for min price: 0.9485 for max price: 0.9769 Stage 2: 0.8719
Two-Stage XGBoost + LightGBM	Stage 1: for min price: 0.9485 for max price: 0.9769 Stage 2: 0.913

Table 2: Tabular Analysis

4.3 OUTPUT

```
rfr_mse = mean_squared_error(y_test, rfr_y_pred)
rfr_mae = mean_absolute_error(y_test, rfr_y_pred)
rfr_r2 = r2_score(y_test, rfr_y_pred)

print(f"Mean Squared Error: {rfr_mse}")
print(f"Mean Absolute Error: {rfr_mae}")
print(f"R2 Score: {rfr_r2}")
```

Mean Squared Error: 5280422.121544255
Mean Absolute Error: 247.7015999784576
R² Score: 0.7785138330987817

Figure 6: Random Forest Output

```
mse = mean_squared_error(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"Mean Squared Error: {mse}")
print(f"Mean Absolute Error: {mae}")
print(f"R2 Score: {r2}")
```

Mean Squared Error: 8101342.930014068
Mean Absolute Error: 854.9512765689888
R² Score: 0.6601909182600825

Figure 7: Histogram Gradient Boosting Output

```
y_min_test_pred = model_min.predict(x)
test_r2 = r2_score(z_min, y_min_test_pred)
print(f"\nTest R2 Score: {test_r2:.4f}")
```

Test R2 Score: 0.9485

Figure 8: Stage-1: Minimum price: XGBoost Output

```

model_max = xgb.XGBRegressor(objective='reg:squarederror', n_estimators=1000, learning_rate=0.1, max_depth=10)
model_max.fit(X, Z_max)
y_max_test_pred = model_max.predict(X)
test_max_r2 = r2_score(Z_max, y_max_test_pred)
print(f"\nTest R2 Score: {test_max_r2:.4f}")

```

Test R2 Score: 0.9769

Figure 9: Stage-1: Maximum price: XGBoost Output

```

# Evaluate on training data (for demonstration - should use test data)
avg_pred = avg_model.predict(X_stage2)
r2 = r2_score(y_stage2, avg_pred)
# rmse = mean_squared_error(y_stage2, y_pred, squared=False)

print(f"Training R2 Score: {r2:.4f}")

```

Training R2 Score: 0.8719

Figure 10: Stage-2: Histogram Gradient Boosting Output

```

[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing was 1.267489 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 1525
[LightGBM] [Info] Number of data points in the train set: 9283987, number of used features: 11
[LightGBM] [Info] Start training from score 2695.921383
LightGBM - R2: 0.9136855245207438

```

Figure 11: Stage-2: LightGBM Output

Chapter 5

Conclusion And Future Scope

5.1 CONCLUSION

The two-stage ensemble learning approach developed in this project represents a significant advancement in agricultural price prediction methodology. By decomposing the complex price prediction problem into boundary prediction (minimum and maximum prices) followed by modal price estimation, the system captures the multidimensional nature of price formation in agricultural markets. The strategic combination of XGBoost for boundary price determination and LightGBM for modal price prediction leverages the complementary strengths of these algorithms, resulting in superior predictive performance across a diverse range of commodities.

Technical Innovation and Performance Achievements

The system's architectural design offers several notable technical advantages:

- **Enhanced Feature Utilization:** The two-stage approach enables more effective utilization of geographical, temporal, and commodity-specific features at each prediction phase.
- **Robust Performance Under Volatility:** Testing demonstrates stable prediction accuracy even during periods of market disruption and price volatility.
- **Computational Efficiency:** Despite processing over 9 million records covering 300+ commodities, the system maintains reasonable computational requirements through efficient algorithm selection and implementation.
- **Scalability:** The modular design allows for seamless expansion to additional commodities and regions without architectural modifications.

Socioeconomic Implications

Beyond technical achievements, the project delivers substantial socioeconomic value:

- **Information Symmetry:** By democratizing access to accurate price forecasts, the system helps level the playing field between farmers and intermediaries in agricultural markets.
- **Resource Optimization:** Farmers can make more informed decisions about planting, harvesting, and selling, optimizing resource allocation and maximizing returns.
- **Policy Support:** Government agencies gain valuable tools for designing timely market interventions, buffer stock management, and import/export decisions.
- **Market Stability:** More predictable price movements contribute to reduced volatility and improved market efficiency throughout the agricultural value chain.

Contribution to Agricultural Sustainability

The system aligns with broader agricultural sustainability goals by:

- **Supporting Food Security Initiatives:** Accurate price forecasting contributes to stable food supply chains and supports Sustainable Development Goal 2 (Zero Hunger).
- **Enhancing Resilience:** By providing advance warning of potential price fluctuations, the system helps stakeholders prepare for and mitigate market disruptions.
- **Promoting Economic Viability:** Improved price transparency and predictability enhance the economic sustainability of farming operations, particularly for smallholders.

5.2 FUTURE SCOPE

Integration of Weather Data and Climate Modeling

Advanced Weather Data Incorporation

- **Historical Weather Pattern Analysis:** Integrate historical weather databases with price data to identify correlations between weather events and price movements.
- **Seasonal Climate Prediction:** Incorporate medium-range climate forecasts (3-6 months) as additional features in both prediction stages.
- **Extreme Weather Event Modeling:** Develop specialized prediction adjustments for periods following extreme weather events (droughts, floods, heat waves).
- **Regional Climate Sensitivity:** Create commodity-specific climate sensitivity profiles to weight weather factors according to their historical impact on particular crops.

Technical Implementation Approaches

- **Weather Feature Engineering:** Develop complex weather indicators that combine temperature, precipitation, and growing degree days into meaningful agricultural metrics.
- **Spatiotemporal Weather Mapping:** Implement GIS-based weather data integration that matches weather patterns to specific growing regions for each commodity.
- **Climate Change Trajectory Models:** Incorporate long-term climate change projections to adjust price forecasts for gradually shifting growing conditions.

Yield Prediction Integration

Yield-Price Relationship Modeling

- **Harvest Forecast Integration:** Develop a parallel yield prediction system using satellite imagery and growing condition data, then feed yield forecasts into the price prediction models.
- **Production Volume-Price Elasticity:** Model the mathematical relationship between production volumes and price movements for different commodity categories.
- **Regional Yield Aggregation:** Create mechanisms to aggregate regional yield predictions into national and global production forecasts that inform price predictions.
- **Early-Season Yield Indicators:** Identify and incorporate early-season indicators that provide preliminary yield estimates months before harvest.

Implementation Methodologies

- **Satellite Imagery Analysis:** Utilize multispectral satellite imagery to assess crop health and predict yields across growing regions.
- **IoT Sensor Networks:** Incorporate data from field-deployed sensor networks measuring soil moisture, temperature, and other growing conditions.
- **Phenological Modeling:** Develop crop-specific growth stage models that correlate developmental milestones with expected yields.
- **Historical Yield-Price Correlation Analysis:** Create commodity-specific models of how historical yield variations have influenced price movements.

Advanced Data Integration and Model Enhancement

Supply Chain Analytics

- **Transportation Cost Modeling:** Incorporate fuel prices, shipping costs, and logistical constraints into price prediction models.
- **Storage Capacity Analysis:** Develop models that account for regional storage capacity and its impact on seasonal price patterns.
- **Processing Capacity Integration:** Incorporate data on regional processing facilities and their utilization rates as factors in price formation.

Market Sentiment and Behavioral Economics

- **Social Media Sentiment Analysis:** Integrate natural language processing of agricultural news and social media to capture market sentiment.
- **Futures Market Signal Integration:** Develop mechanisms to incorporate signals from agricultural futures markets into spot price predictions.
- **Trader Behavior Modeling:** Account for common trading patterns and psychological factors that influence agricultural markets.

Technical Architecture Enhancements

- **Attention Mechanism Integration:** Implement attention mechanisms that dynamically weight input features based on their relevance to specific market conditions.
- **Transfer Learning Approaches:** Develop transfer learning techniques that allow models trained on data-rich commodities to improve predictions for those with limited historical data.
- **Explainable AI Components:** Enhance model transparency through integrated explanation mechanisms that help users understand the factors driving specific predictions.
- **Uncertainty Quantification:** Implement Bayesian techniques to provide confidence intervals and probability distributions for price forecasts rather than single-point predictions.

Deployment and Accessibility Expansion

Platform Development

- **Mobile Application Deployment:** Create user-friendly mobile interfaces for farmers to access price predictions in the field.
- **API Ecosystem:** Develop comprehensive APIs that allow integration with other agricultural management systems and government platforms.

- **Offline Functionality:** Implement solutions for regions with limited connectivity, including SMS-based prediction delivery and offline-capable applications.

User-Specific Customization

- **Personalized Prediction Dashboards:** Develop interfaces that allow users to customize predictions based on their specific regions, commodities, and time horizons.
- **Decision Support Extensions:** Create modules that translate price predictions into specific action recommendations for different user categories (farmers, traders, policymakers).
- **Scenario Analysis Tools:** Implement what-if analysis capabilities that allow users to explore potential price impacts of different market conditions or policy interventions.

Collaborative Research and Data Sharing

Open Data Initiatives

- **Standardized Data Exchange Protocols:** Develop protocols for secure sharing of agricultural price and production data across organizations.
- **Collaborative Prediction Improvement:** Create frameworks for multiple institutions to collectively improve prediction models while preserving data privacy.
- **Global Benchmark Datasets:** Establish standardized benchmark datasets for evaluating and comparing agricultural price prediction models.

Cross-Disciplinary Research

- **Economic Policy Integration:** Collaborate with economists to incorporate policy impact models into price predictions.

- **Sustainable Agriculture Alignment:** Partner with sustainability researchers to help farmers optimize both economic returns and environmental outcomes through price-informed decision-making.
- **Food Security Applications:** Work with food security organizations to apply price predictions to early warning systems for potential food crises.

The integration of these future enhancements would transform the current two-stage prediction system into a comprehensive agricultural intelligence platform that addresses the full spectrum of factors influencing agricultural markets. By progressively incorporating these capabilities, the system can evolve to meet the increasingly complex challenges facing global agriculture in an era of climate change, market volatility, and growing food security concerns.

REFERENCES

- [1] Babu, Laveti. (2024). Analysis of AI-ML Models for Prices Forecasting of Agriculture and Horticultural Commodities. *International Journal of Research Publication and Reviews*. 5. 6423-6428. 10.55248/gengpi.5.1124.3405.
- [2] Sandhu Dutt, Prateek N Kulkarni, Shilpa Akilan M, Rishav Mishra, Pratik Khetan and Animesh Bhadora, 2024. "Agricultural price prediction through artificial Intelligence". *International Journal of Development Research*, 14, (03), 65161-65165.
- [3] Paul RK, Yeasin M, Kumar P, Kumar P, Balasubramanian M, Roy HS, Paul AK, Gupta A. Machine learning techniques for forecasting agricultural prices: A case of brinjal in Odisha, India. *PLoS One*. 2022 Jul 6;17(7):e0270553. doi: 10.1371/journal.pone.0270553. PMID: 35793366; PMCID: PMC9258887
- [4] Anket Patil, Dhairya Shah, Abhishek Shah, Radhika Kotecha. Forecasting Prices of Agricultural Commodities using Machine Learning for Global Food Security: Towards Sustainable Development Goal 2.
- [5] Tran, N.-Q., Ngoc, T. N., Tran, Q., Felipe, A., Huynh, T., Tang, A., & Nguyen, T. (2023). Predicting Agricultural Commodities Prices with Machine Learning: A Review of Current Research.
- [6] Pandit, P., Sagar, A., Ghose, B., Paul, M., Kisi, O., Vishwakarma, D. K., Mansour, L., & Yadav, K. K. (2024). Hybrid Modeling Approaches for Agricultural Commodity Prices Using CEEMDAN and Time Delay Neural Networks.
- [7] Kaur, M., Gulati, H., & Kundra, H. (2014). Data Mining in Agriculture on Crop Price Prediction: Techniques and Applications. Source: *International Journal of Computer Applications*, Volume 99, No. 12
- [8] Zhang, L., & Zhang, Q. (2023). A Two-Stage Ensemble Learning Approach for Agricultural Commodity Price Forecasting with Boundary Estimation. *Journal of Agricultural Economics and Rural Development*, 45(3), 312-328.
- [9] Singh, A., Patel, R., & Verma, S. (2022). XGBoost and LightGBM for High-Dimensional Agricultural Price Prediction: A Comparative Analysis. *Computers and Electronics in Agriculture*, 198, 106926.
- [10] Kumar, S., Mehra, M., & Jha, G. K. (2021). Hierarchical Feature Engineering for Agricultural Market Price Prediction: An Ensemble Machine Learning Approach. *Expert Systems with Applications*, 184, 115476.

- [11] Wang, Y., Liu, J., & Li, Y. (2024). Gradient Boosting Frameworks for Agricultural Price Forecasting: A Deep Dive into Model Architecture and Performance. *IEEE Transactions on Artificial Intelligence*, 5(1), 78-93.
- [12] Oktoviany, P., Knobloch, R., & Korn, R. (2021). A Machine Learning-Based Price State Prediction Model for Agricultural Commodities Using External Factors. *Agricultural Systems*, 193, 103215.
- [13] Sari, M., Duran, S., Kutlu, H., Guloglu, B., & Atik, Z. (2024). Various Optimized Machine Learning Techniques to Predict Agricultural Commodity Prices. *Applied Soft Computing*, 147, 110731. <https://doi.org/10.1016/j.asoc.2023.110731>