Predicting Crop Prices using Machine Learning Models

Introduction

This journal explores the application of three distinct machine learning models—XGBoost, Random Forest Regressor, and Support Vector Machines (SVM)—for predicting crop prices based on historical agricultural market data. Each model offers unique strengths in handling complex datasets and capturing different types of relationships between features and target variables.

Dataset Overview

The dataset used in this study, `Total_Crops.csv`, comprises essential attributes such as mandi ID, crop ID, crop name, mandi name, arrival quantity, maximum price, minimum price, modal price, and date. Key attributes for prediction include modal price, date, mandi ID, crop ID, and lagged price features derived from historical data.

Data Preprocessing

1. Cleaning and Transformation: Irrelevant attributes like arrival quantity are removed. The date column is converted into datetime format, and additional temporal features such as year, month, and day are extracted.
2. Feature Engineering: Lag features are created using the past 21 days' prices to predict the next day's modal price. Categorical variables (mandi ID and crop ID) are encoded using LabelEncoder.
3. Handling Missing Values: Rows with missing values resulting from the creation of lag features are dropped to ensure data integrity.

Model Selection and Training

1. XGBoost Model

- *Model Characteristics*: XGBoost is selected for its scalability and performance in gradient boosting scenarios.
- *Training Process:* The dataset is split into training and testing sets. Hyperparameters such as learning rate, maximum depth, and number of estimators are tuned using cross-validation.
- *Evaluation*: Model performance is evaluated using Mean Absolute Error (MAE) on the test set.

2. Random Forest Regressor

- *Model Characteristics*: Random Forest is chosen for its ability to handle complex datasets and capture non-linear relationships.
- *Training Process*: Similar to XGBoost, the dataset is split, and the model is trained on the training set. Parameters like number of trees and maximum depth are optimized.
- *Evaluation:* MAE is used to assess prediction accuracy on the test set.

3. Support Vector Machines (SVM)

- *Model Characteristics*: SVM is employed due to its effectiveness in capturing complex relationships and handling high-dimensional data.
- *Training Process*: The dataset undergoes the same preprocessing steps. SVM parameters such as kernel type, regularization parameter (C), and kernel coefficient (gamma) are optimized during training.
- *Evaluation:* Model performance is measured using MAE on the test set.

4. Voting Regressor

- *Concept:* The Voting Regressor aggregates predictions from multiple base estimators (individual models) and outputs the average prediction, often resulting in improved generalization and stability compared to single models.
- *Implementation*: In this project, Voting Regressor combines predictions from Random Forest and Gradient Boosting models, leveraging their diverse learning approaches (bagging and boosting) to achieve better prediction accuracy.
- *Advantages*: Voting Regressor can handle different sources of data, model heterogeneity, and reduces overfitting by combining diverse models.

5. AdaBoost (Adaptive Boosting)

- *Concept*: AdaBoost sequentially trains a series of weak learners (e.g., decision trees) where each subsequent model corrects errors made by the previous one, focusing more on difficult instances.
- *Implementation:* In this project, AdaBoost uses Decision Tree regressors as weak learners. It adjusts the weights of incorrectly predicted instances iteratively, emphasizing challenging data points to enhance overall prediction accuracy.
- *Advantages*: AdaBoost is effective in handling complex relationships in data, improving predictive performance by focusing on hard-to-classify instances.

6. Gradient Boosting

- *Concept:* Gradient Boosting builds an ensemble of decision trees sequentially, where each tree corrects errors of the previous one by fitting residuals of the prediction gradient.

- *Implementation*: Here, Gradient Boosting combines multiple decision trees to create a strong learner. It iteratively improves the model's prediction by minimizing the loss function, often resulting in highly accurate predictions.
- *Advantages:* Gradient Boosting is robust against overfitting, handles missing data well, and provides feature importance, aiding interpretability in agricultural price forecasting.

Predictive Analysis and Visualization

- User Interaction: Users provide inputs for crop ID and optionally mandi ID. They specify the number of days (5-15) for which future price predictions are desired.
- Prediction Visualization:
    - *Historical and Predicted Prices*: Each model produces plots displaying historical modal prices alongside predicted prices for the specified number of days. This visualization aids in comparing observed trends with model forecasts.
    - *Standalone Predictions*: Additional plots show standalone predicted prices for the chosen number of days, with data labels for clarity. These plots focus on predicted modal prices over time.

Conclusion

This journal entry highlights the application of XGBoost, Random Forest, and SVM models in predicting crop prices, emphasizing their distinct methodologies and performance metrics. These models contribute to informed decision-making in agriculture, supporting sustainable practices and economic development.

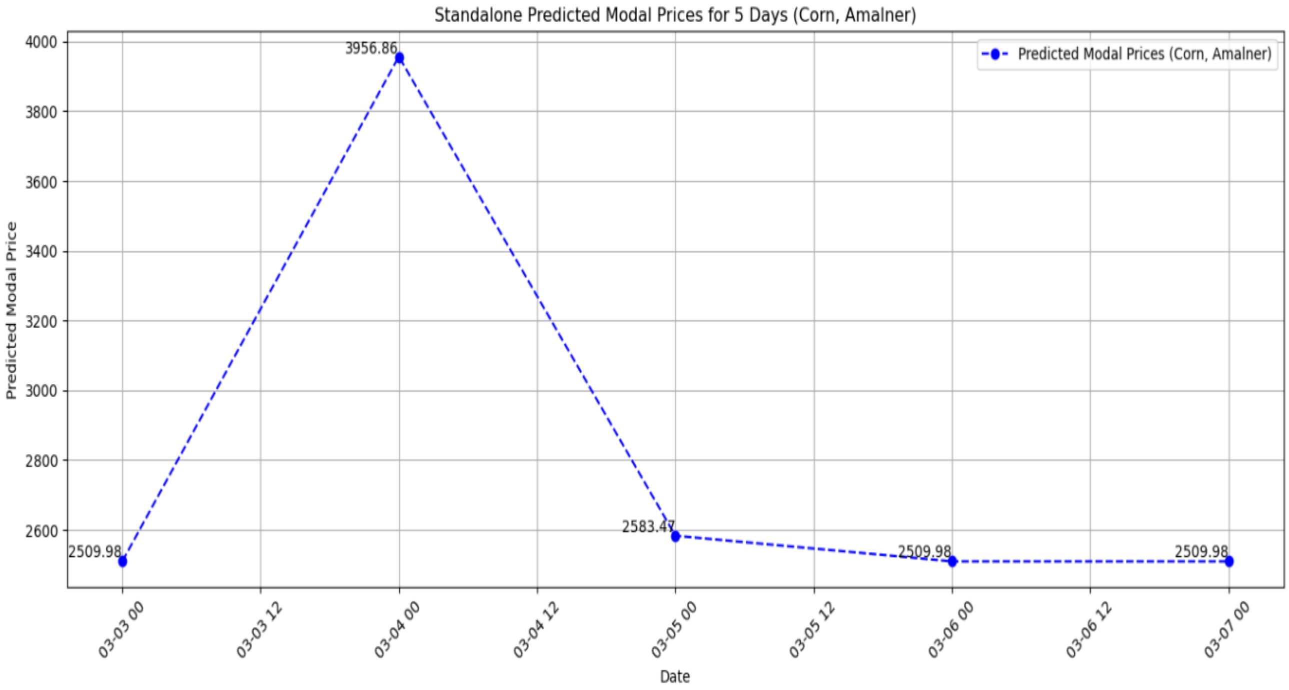
Future Directions

Future work may involve integrating additional datasets such as weather patterns, economic indicators, or regional factors to enhance model accuracy and robustness. Further exploration could expand the dataset to encompass diverse agricultural products and market regions, broadening the models' applicability and improving their predictive capabilities.

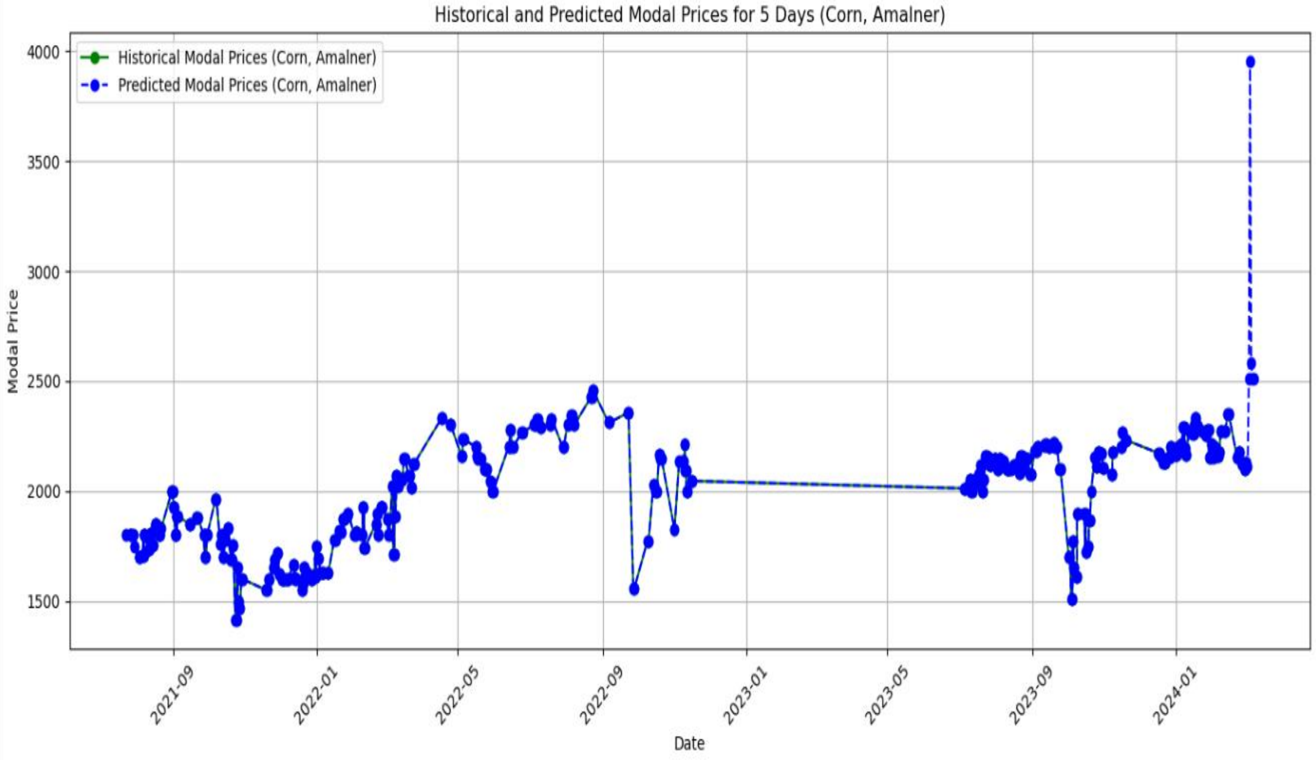
By documenting these methodologies and findings, this journal entry offers valuable insights into leveraging machine learning for agricultural price forecasting, fostering advancements in agricultural economics and policy-making.

OUTPUT:

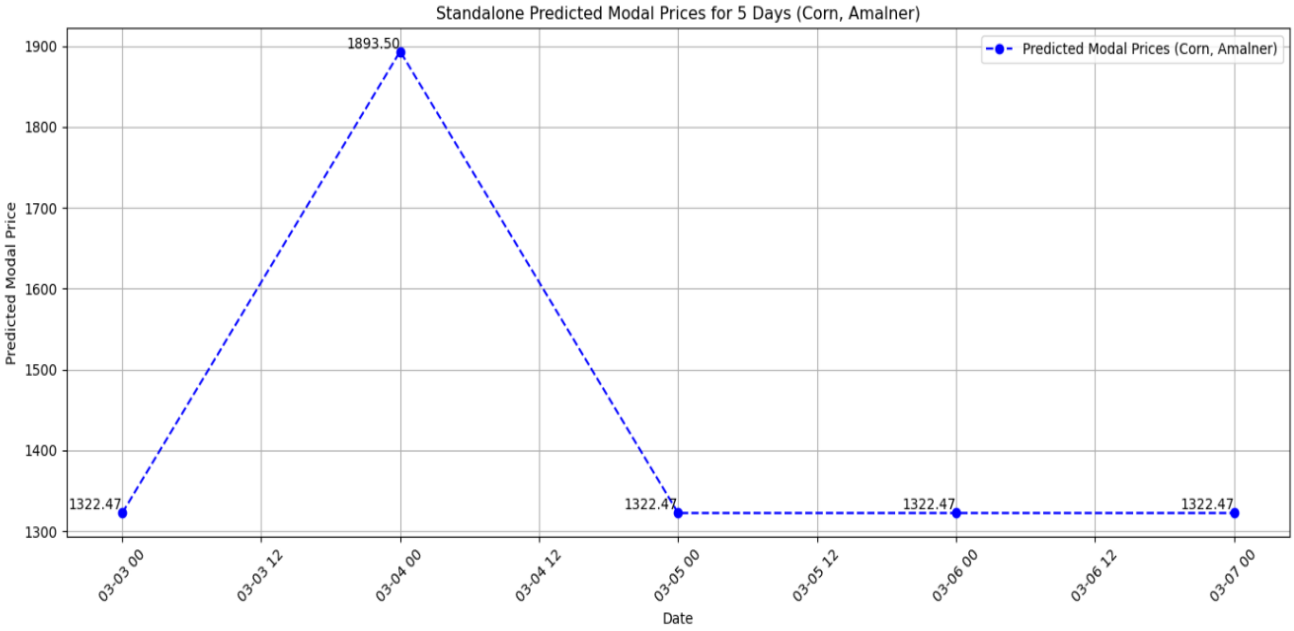
1) XGBoost Predicted Values graph



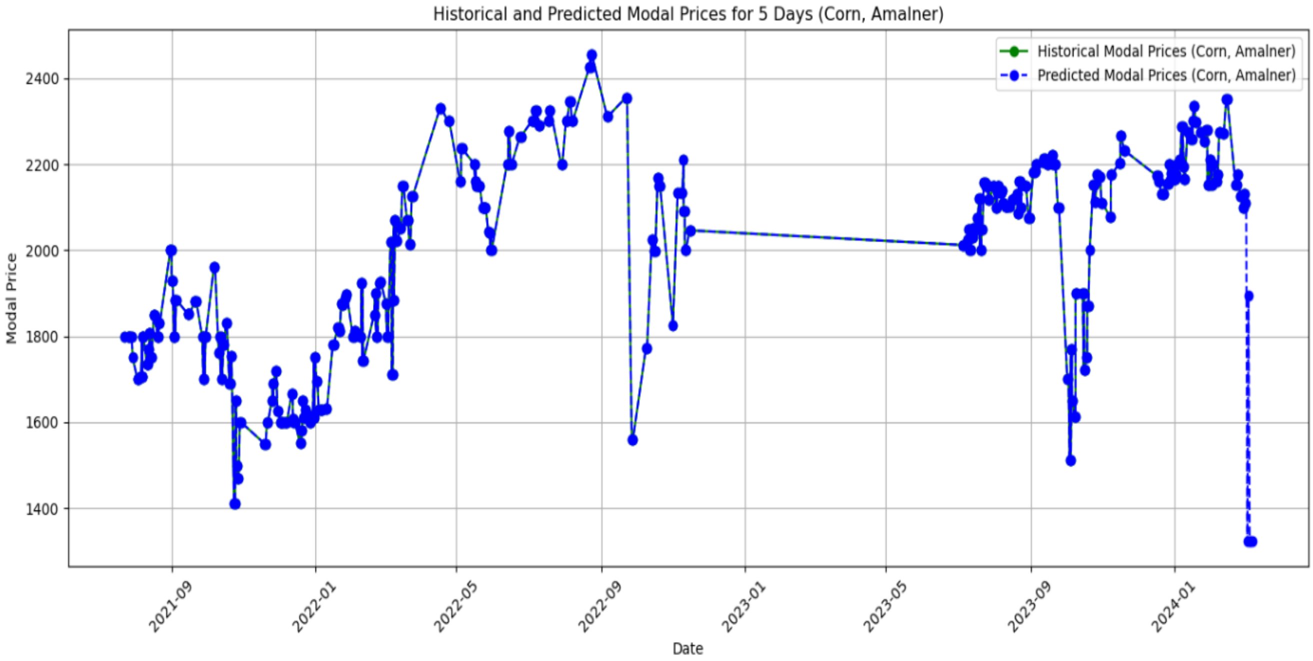
2) XGBoost Historical+Predicted Graph



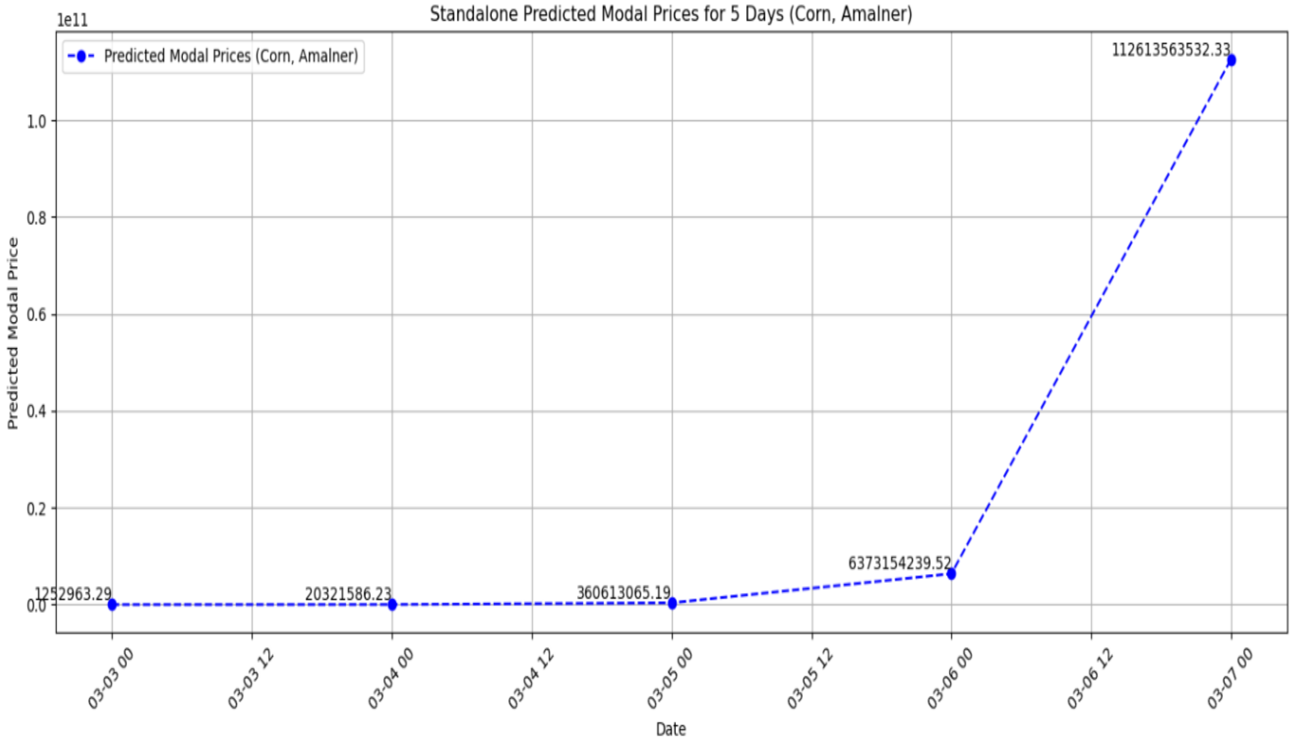
3) RandomForestRegressor Predicted Values graph



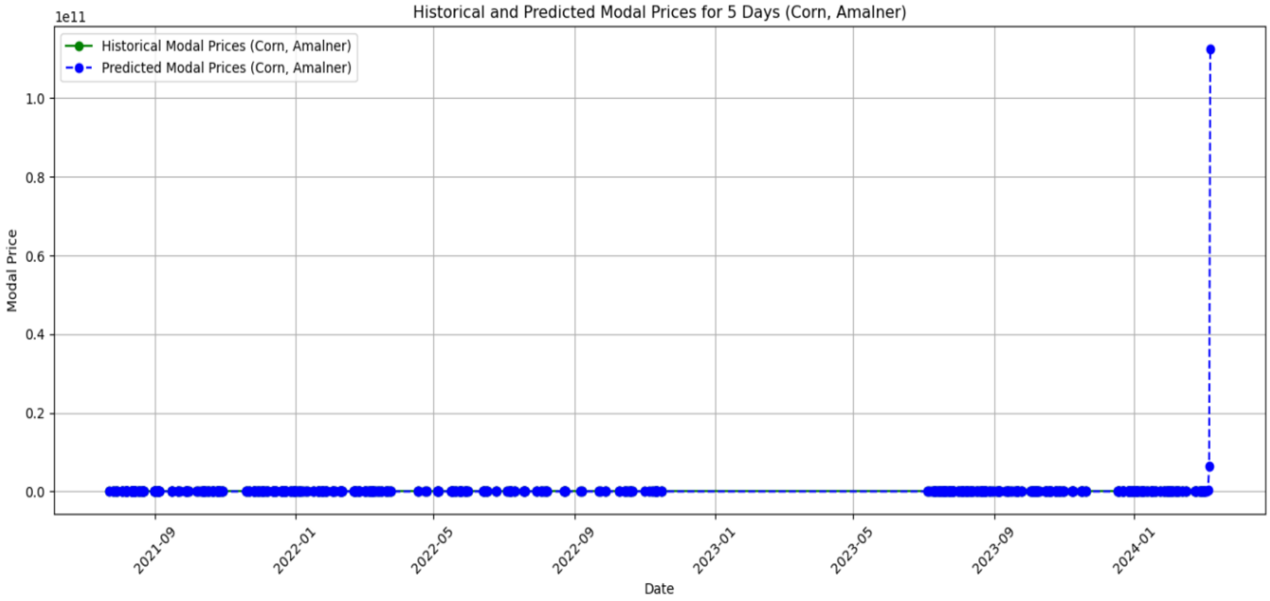
4) RandomForestRegressor Historical+Predicted Graph



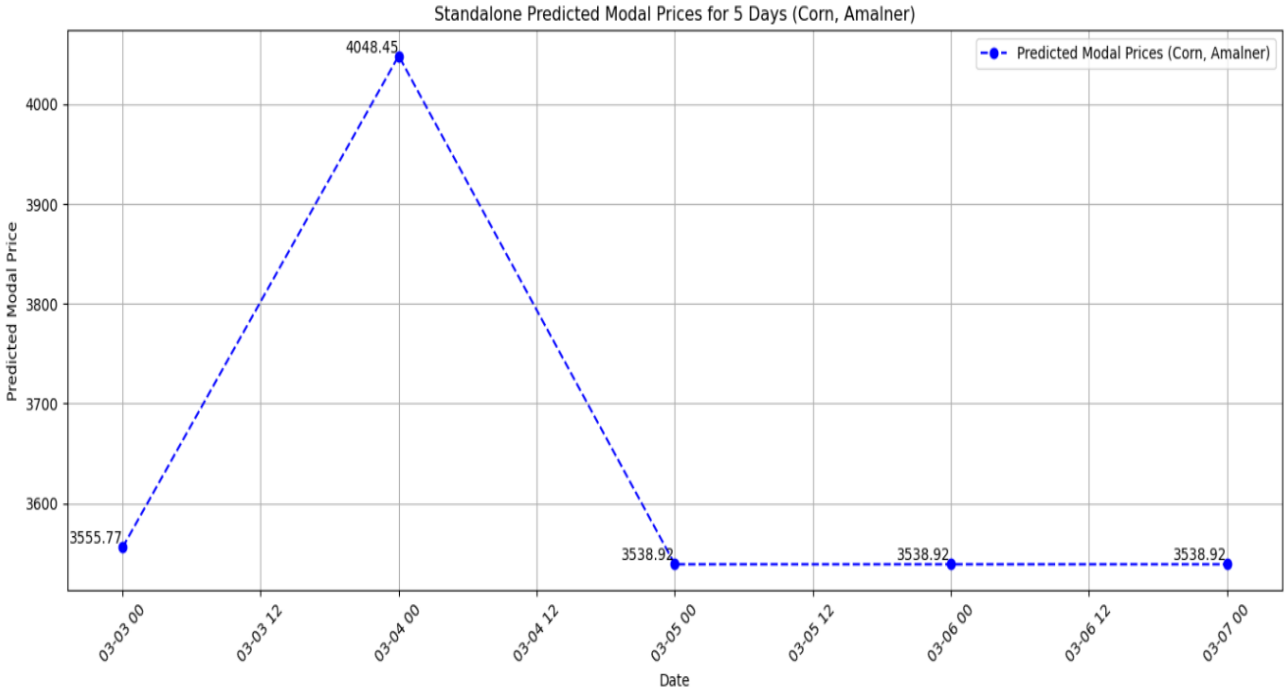
5) SVM Predicted Values graph



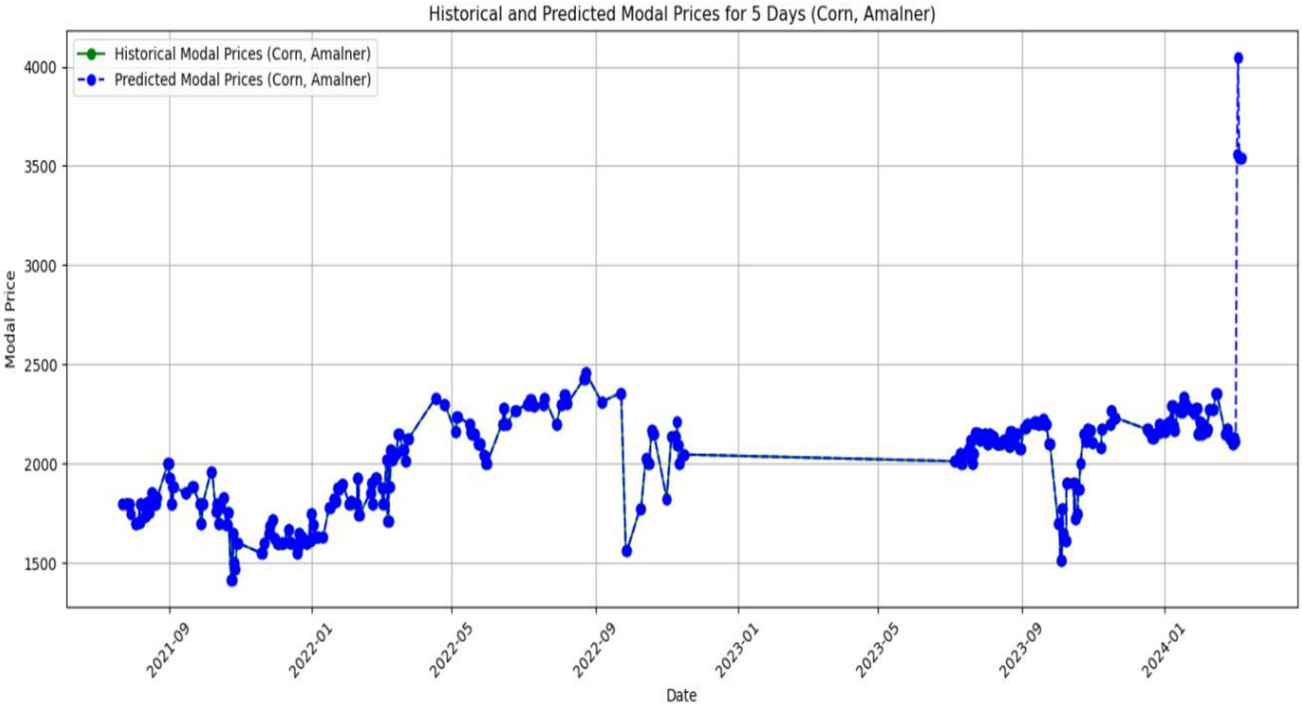
6) SVM Historical+Predicted Graph



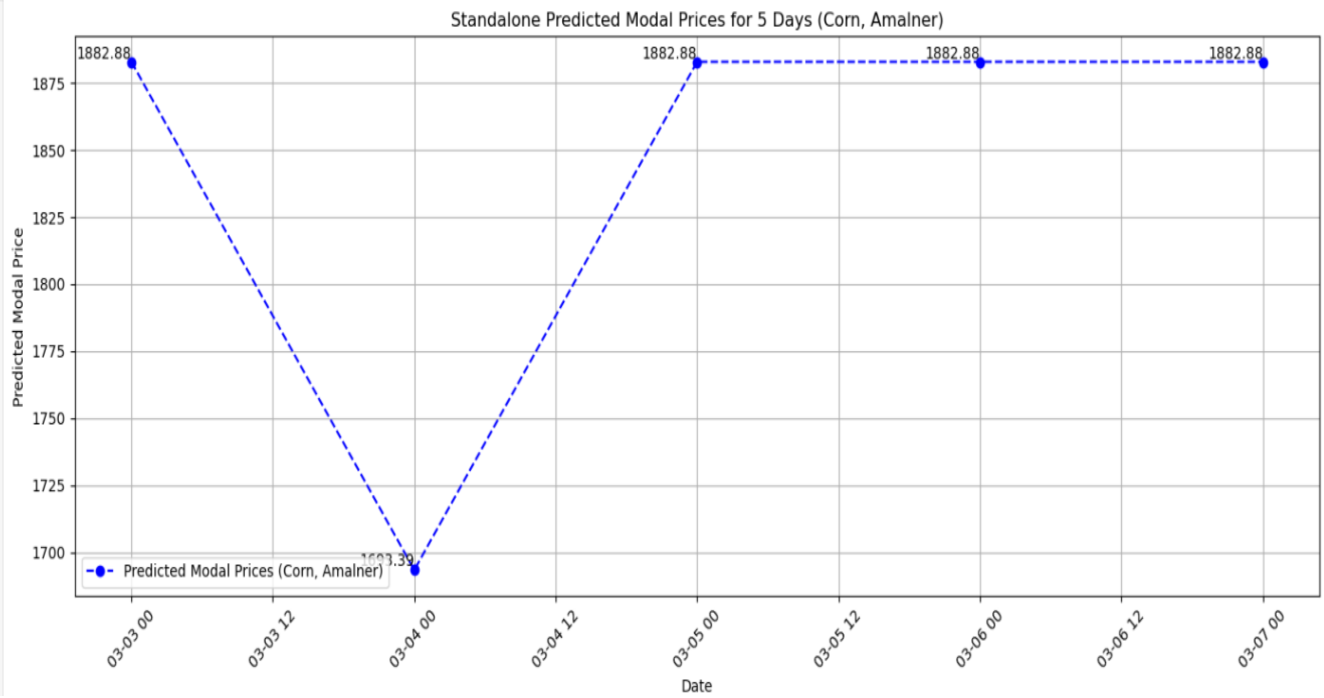
7) Voting Regressor Predicted Values graph



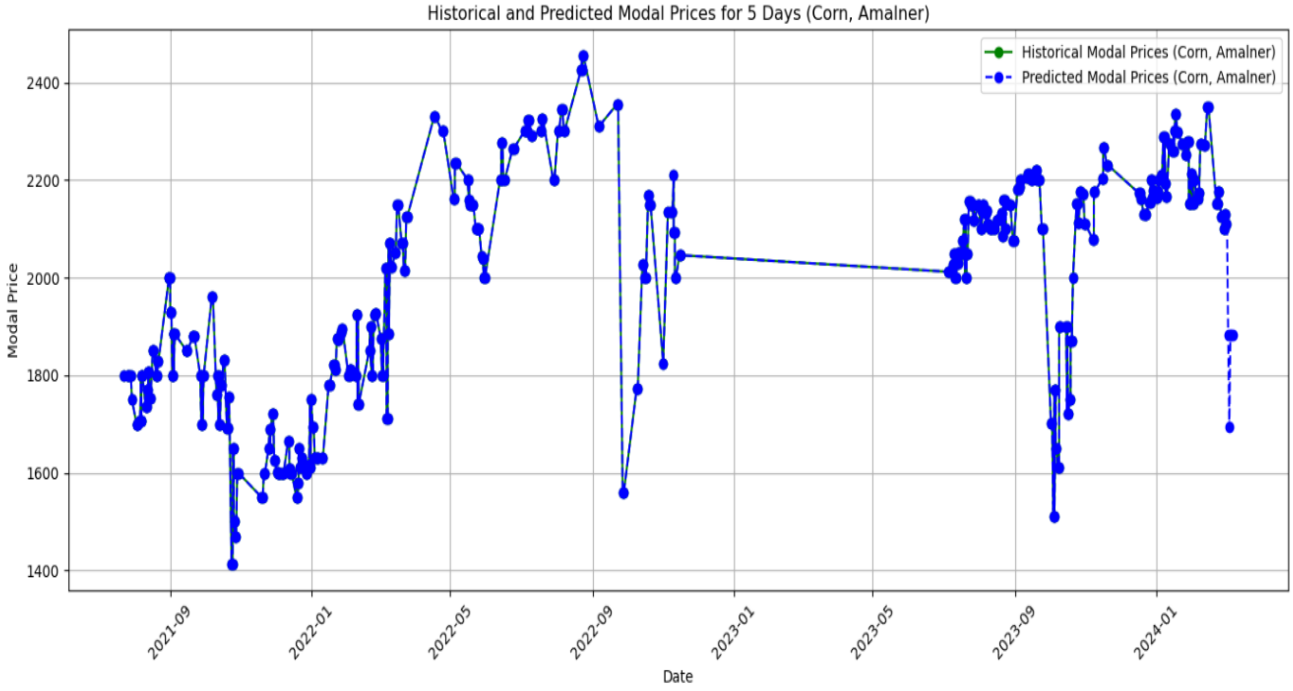
8) Voting Regressor Historical+Predicted Graph



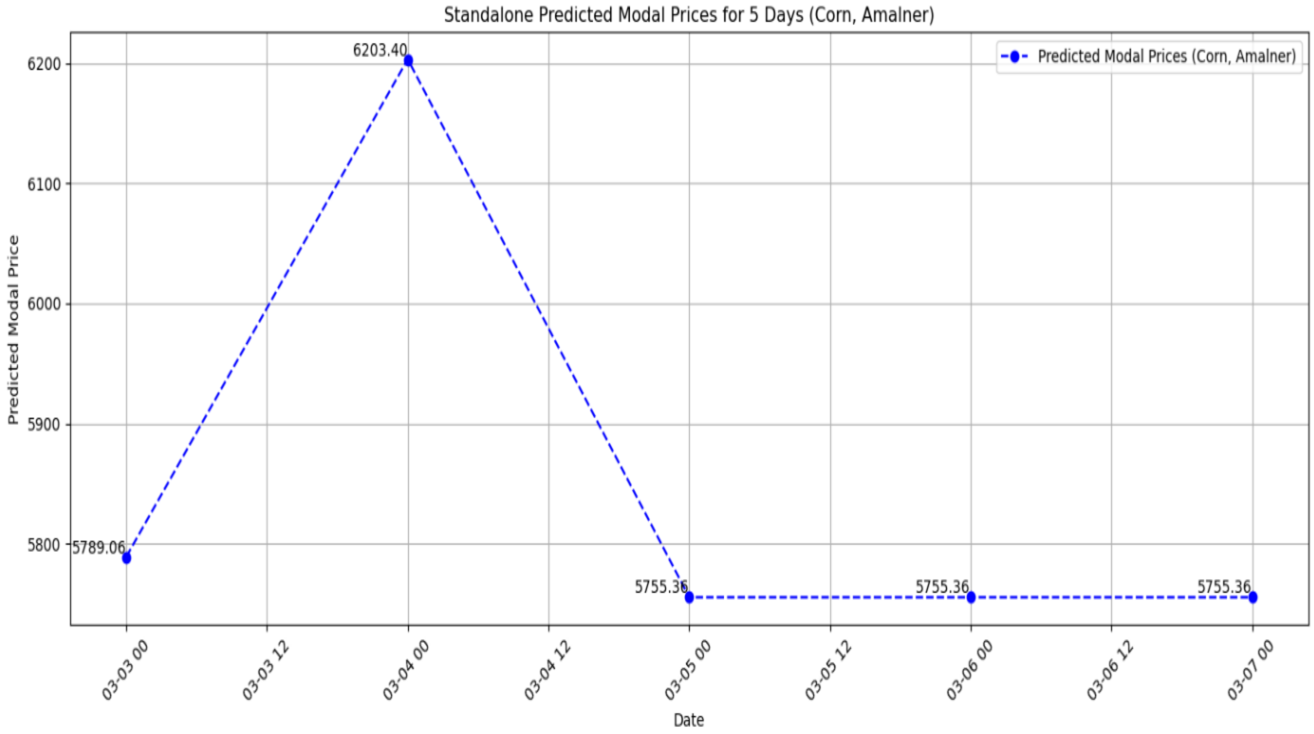
9) Adaboost Predicted Values graph



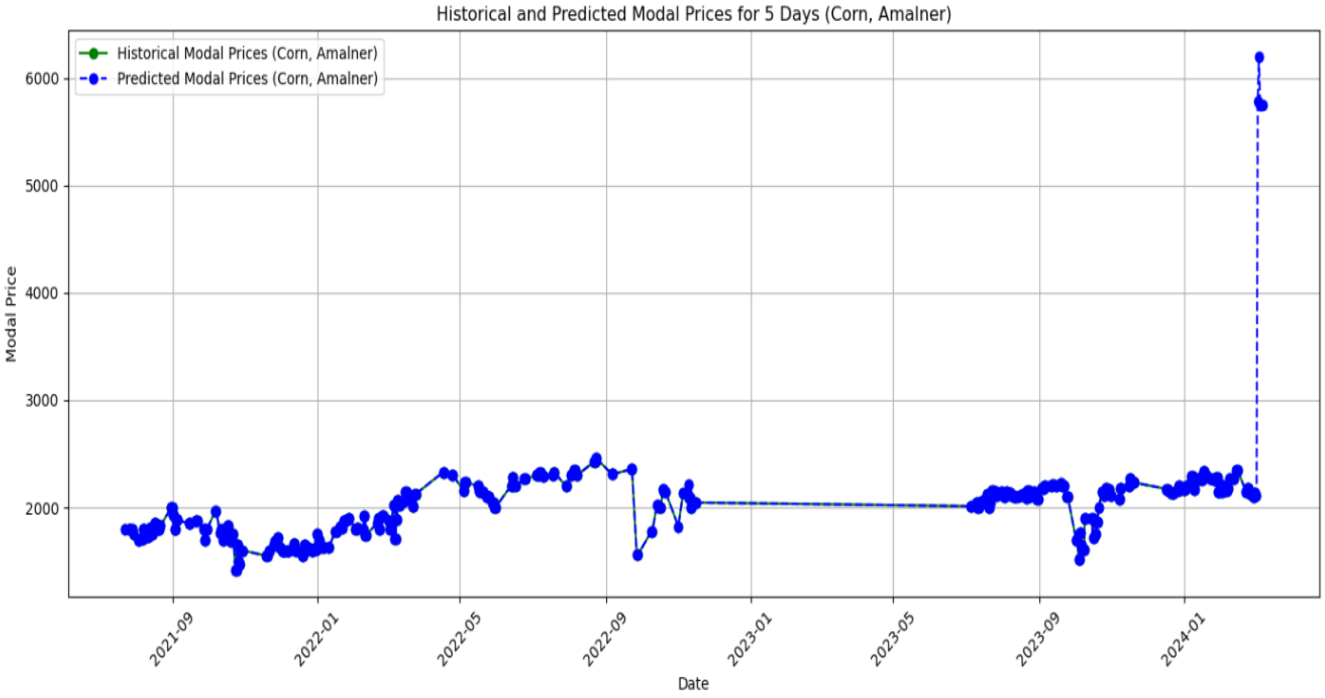
10) Adaboost Historical+Predicted Graph



11) GBM Predicted Values graph



12) GBM Historical+Predicted Graph


Historical and Predicted Modal Prices for 5 Days (Corn, Amalner)
Historical Modal Prices (Corn, Amalner)
Predicted Modal Prices (Corn, Amalner)
Modal Price
Date