Comparing Models of Predicting Food Prices in Uganda

Belinda NAKANYIKE, Jusper TWINAMASIKO, Charles OWORI

22 November 2025

Abstract

The lack of a reliable food price forecast in Uganda is addressed by this project as the primary requirement for food safety and strategic planning. After a thorough data preprocessing of the historical data obtained from the World Food Programme, the study proceeded to train, tune, and evaluate six different machine learning models. Four metrics(R^2, RMSE, MAE, MSE) assessed the performance of the models. The Random Forest model was the top performer, achieving an R^2 of 0.9471. The research shows that although the results are impressive, the model's accuracy is limited by the absence of external features like seasonal and weather data. The project then implemented an interface to allow users to input data using Gradio.

1 Introduction

The goal of this project was to build and compare machine learning models capable of predicting food prices in Uganda based on historical data from the World Food Programme Database. For consumers, farmers, vendors, and policymakers, predicting price changes is important because it facilitates planning, food security monitoring, and informed decision-making. To find the model that generates the best predictions, a range of algorithms was trained and assessed using different metrics.

2 Methodology

2.1 Data Preprocessing

The dataset was preprocessed to make it clean and ready for analysis. This involved the following steps:

1. Removing outliers
2. Removing non-foods from the category column
3. Feature engineering by extracting the month and year from the date feature
4. Encoding categorical features
5. Selecting features to be used in the prediction
6. Train-test split
7. Scaling

2.2 Data Analysis

The monthly bean price analysis reveals a strong and consistent seasonal pattern with lower prices in the first few months of the year, most likely from post-harvest market saturation, followed by a gradual rise to a peak in the dry months as supplies become lower, and then a decrease towards the end of the year.

2.3 Algorithms Used

We employed and evaluated six different machine learning techniques in order to develop predictive models for Ugandan food prices. The algorithms were carefully selected to represent a variety of modeling paradigms, such as random forests, neural networks, k-nearest neighbors, decision trees, gradient boosting machines, and linear regression models.

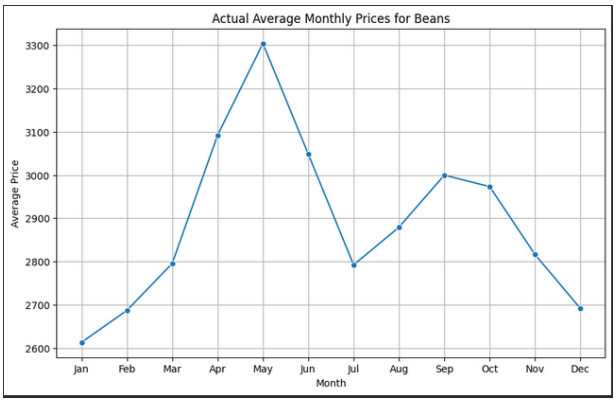

Figure 1: Average monthly prices of beans across months of the year

2.3.1 Random Forest

Before the Random Forest model was improved by hyperparameter tuning, a baseline with default parameters was established using RandomizedSearchCV. This process carefully examined the optimal conditions of significant parameters using the core Random Forest ideas of bagging and feature randomization. Additionally, the model provides feature importance ratings that indicate how much each feature contributed to the predictions. These scores offer useful information about the underlying reasons that enhance the model's interpretability in addition to its predictive effectiveness.

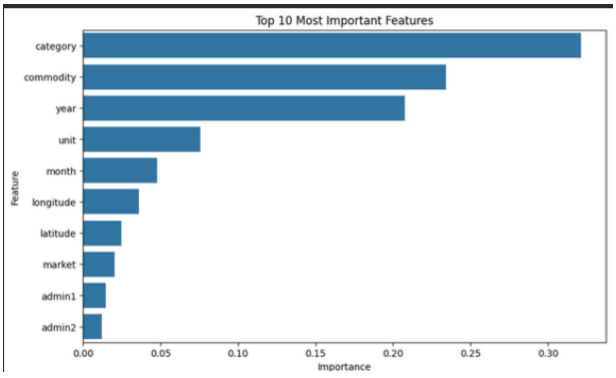

Figure 2: Top 10 most important features

2.3.2 Neural Network

The neural network approach involves building a sequential model with several dense layers, ReLU activation, and MSE loss. Standardizing input features and implementing early stopping to prevent overfitting by tracking validation loss and reestablishing ideal weights were the key steps.

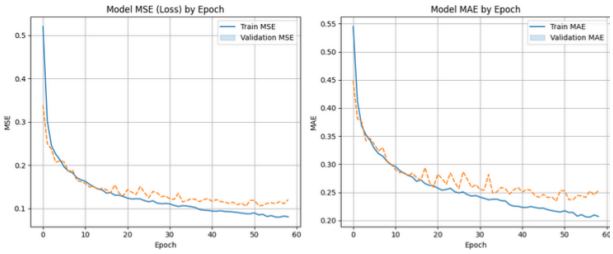

Figure 3: Model loss and Model MAE by epoch

2.3.3 Decision Tree

Using default parameters, the Decision Tree Regressor was set up. After that, an exhaustive hyperparameter tuning session took place by RandomizedSearchCV with 50 iterations and 5-fold cross-validation to explore the main parameters such as the depth of the tree, the rules for splitting the sample, and the criteria for selecting the features. While the overall performance was assessed using a variety of statistical techniques to verify the model's prediction accuracy, the optimizer used a negative mean squared error as the metric to assess the efficiency of the model configuration.

2.3.4 KNN

In order to ensure fair distance calculations, the K-Nearest Neighbours methodology used a two-step process beginning with crucial feature standardization using StandardScaler, followed by systematic hyperparameter optimization using RandomizedSearchCV to determine the best combination of neighbours, weighting scheme, and distance metric. The model's ability to correctly generalize patterns

from the scaled features is ensured by adjusting key parameters.

2.3.5 linear regression

One-hot encoding for categorical variables and standardization of both features and target were among the several preprocessing steps used in the implementation of linear regression. The model established linear relationships on the processed data and was evaluated using standard regression metrics to provide a baseline for price prediction.

2.3.6 Gradient boosting Regressor

Gradient boosting is an ensemble machine learning technique that develops models sequentially, with each model fixing the mistakes of its predecessors. A complete hyperparameter optimization was carried out by RandomizedSearchCV with 25 iterations and 5-fold cross-validation. The model was trained with the preprocessed data, and its performance was assessed with metrics to allow a comprehensive performance check and accuracy of prediction.

3 Results

3.1 Model Performance Metrics

The project utilized a set of performance metrics to precisely assess the effectiveness of the model. It included Mean Squared Error(MSE) that was used to penalize larger prediction errors, Root Mean Squared Error(RMSE) that was used to indicate the error size, Mean Absolute Error(MAE) that was used as a measure of error which is robust and insensitive to outliers, and R-squared (R^2) that used to measure the proportion of variance explained by the models. These metrics yielded different but complementary insights into the prediction accuracy, characteristics of error distribution, and overall explanatory power of the model, which constituted a complete understanding of the model's performance for making reliable decisions.

MODEL	R^2	RMSE	MAE	MSE
Random forest	**0.947**	**0.241**	**0.158**	**0.058**
Neural network	0.912	0.310	0.222	0.096
KNN	0.838	0.424	0.299	0.179
Decision tree	0.915	0.306	0.203	0.093
Linear regression	0.792	0.479	0.352	0.230
Gradient boosting	0.941	0.256	0.169	0.066

Table 1: Comparison of results for the models

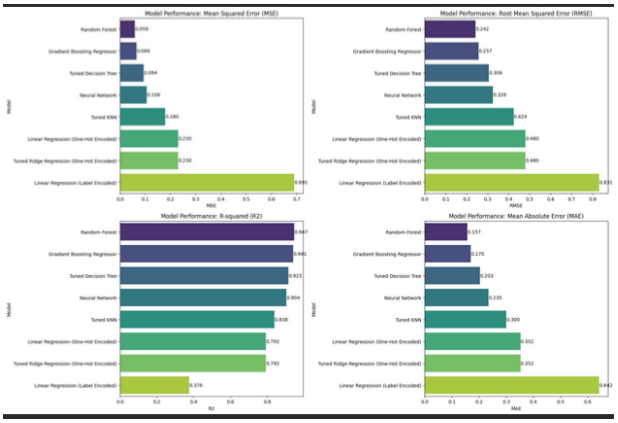

Figure 4: Comparison of the models using metrics

3.2 Discussion of results

Among the different models tried, Random Forest was the most successful in terms of performance because of its capability to recognize non-linear relationships. Nevertheless, all models exhibited limited accuracy to a greater extent due to the gaps in the dataset, particularly the lack of seasonal, weather, and demand-related features that have a strong influence on food prices. The outcomes emphasize that adding more potential features would not only increase the model's performance but also enable it to make more accurate predictions.

3.3 Prediction Interface

The prediction interface was implemented using Gradio. It enables users to input features, process them, and then utilize the best model to predict prices based on the input.

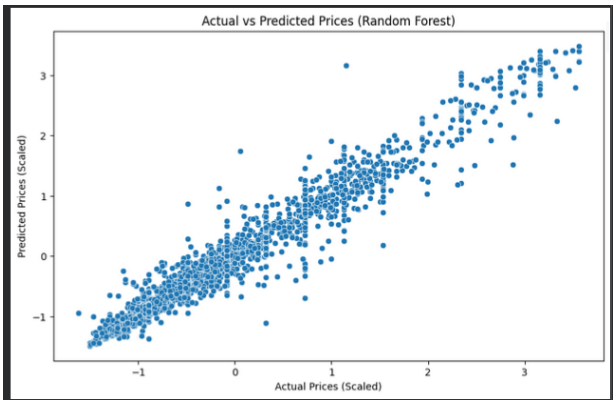

Figure 5: Actual vs predicted scaled prices using random forest

s

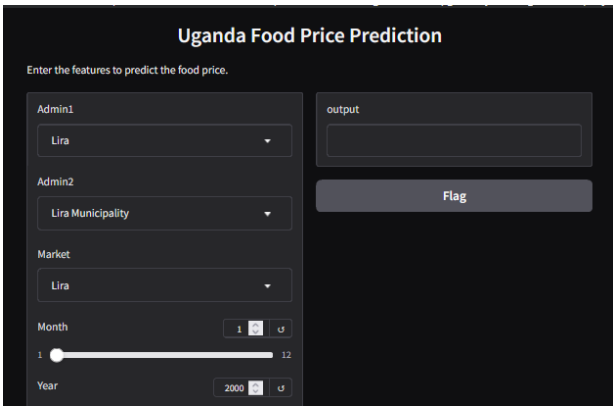

Figure 6: Prediction Interface

4 Challenges faced

The dataset was missing necessary external factors like rainfall, temperature, harvest periods, and seasonal production cycles, which have a major impact on food availability and prices in Uganda.

The dataset did not have consumer demand features such as holiday demand, urban and rural consumption differences.

5 Recommendations for future work

The model can better account for the economic drivers of price changes by including weather and seasonal data in the dataset. Accurate forecasting is very important as factors like temperature, rainfall, and extreme weather can still have a very large impact on crop yields and the availability of supply.

The model can more effectively explain the changes in prices of food if it uses demand-related features such as market demand, consumption patterns, and population growth, along with other features. This is done by fully integrating supply and demand into the dataset, thus leading to an increased overall accuracy.

The installation of a real-time forecasting system that can give real-time updates of the prices of food to farmers, merchants, and policymakers.

.

6 Conclusion

Through extensive experimentation, the project finds that machine learning, more specifically the Random Forest algorithm, is capable of predicting Ugandan food prices with high accuracy, thus paving the way for data-driven decision making. The model's success is attributed to its ability to capture complex nonlinear patterns in the data. However, the model's capability to predict is hindered by shortcomings of the dataset, especially the lack of the most important real-world factors such as climate, season, and demand. The final model was turned into an interactive Gradio interface. For future improvements, we strongly recommend integrating additional external datasets and exploring more sophisticated models to further enhance accuracy and robustness for food security applications.