Forecasting Food Prices Using Machine Learning: Methods and Accuracy

AAEA, 2024

Raghav Goyal

Assistant Professor, LSU

July 31, 2024

Introduction

- **Importance of Food Price Forecasting:**
  - Crucial for economic planning and policy-making.
  - Helps in stabilizing markets and ensuring food security.
  - Assists farmers, retailers, and consumers in making informed decisions.
- **Traditional Forecasting Methods:**
  - Often rely on linear regression and time-series analysis.
  - Limitations in handling non-linear and complex patterns in data.
- **Advances in Machine Learning (ML):**
  - Ability to model complex and non-linear relationships.
  - Incorporation of various types of data (e.g., weather, market trends).
  - Improved accuracy and reliability of forecasts.
- **Overview of Presentation:**
  - Examination of various ML methods used for forecasting food prices.
  - Discussion of key results and forecast evaluation metrics from different studies.
  - Exploration of the future potential and challenges of ML in food price forecasting.

Machine Learning for Forecasting?

- **Definition:** Machine Learning (ML) for forecasting involves using algorithms that can learn from historical data to predict future values.
- **Types of ML Techniques Used:**
  - **Supervised Learning**
  - **Unsupervised Learning**
  - **Reinforcement Learning**
- **Applications in Forecasting:**
  - **Time-Series Analysis**
  - **Demand Forecasting**
  - **Price Prediction**
- **Benefits:**
  - **Accuracy**
  - **Automation**
  - **Scalability**

Key Considerations for Using ML in Forecasting

- **Data Quality:**
  - Ensuring accurate, complete, and timely data.
  - Handling missing values and outliers effectively.
- **Model Selection:**
  - Choosing the appropriate ML algorithm for the task.
  - Balancing model complexity and interpretability.
- **Overfitting and Underfitting:**
  - Overfitting: Model performs well on training data but poorly on unseen data.
  - Underfitting: Model is too simple to capture the underlying patterns.
- **Evaluation Metrics:**
  - Using metrics like RMSE, MAE, and MAPE to evaluate model performance.
  - Cross-validation techniques to ensure robustness.

Key Considerations for Using ML in Forecasting

- **Scalability:**
  - Ensuring the model can handle large datasets and adapt to new data.
  - Considering computational resources and time constraints.
- **Ethical Considerations:**
  - Addressing biases in data and algorithms.
  - Ensuring transparency and fairness in model predictions.
- **Deployment:**
  - Integrating the model into real-world systems.
  - Monitoring and maintaining the model over time.

When ML Performs Better Than Linear Regression/Time Series Methods

- **Non-linear Relationships:**
  - ML algorithms can capture complex, non-linear relationships in the data.
- **High Dimensionality:**
  - ML can handle and leverage large numbers of features effectively.
- **Complex Interactions:**
  - ML can model intricate interactions between multiple variables.
- **Large Datasets:**
  - ML algorithms scale well with large datasets, improving accuracy with more data.
- **Adaptive Learning:**
  - ML models can continuously learn and adapt from new data inputs.
- **Handling Missing Data:**
  - ML methods can handle missing data more robustly than traditional methods.

When Linear Regression/Time Series Methods Perform Better Than ML

- **Small Datasets:**
  - Traditional methods can perform better with small datasets where ML may overfit.
- **Simplicity:**
  - Linear models are simpler, easier to implement, and faster to train.
- **Interpretability:**
  - Results from linear regression are more interpretable and explainable.
- **Low Variability:**
  - In cases with low variability in data, linear models can perform adequately.
- **Stationary Data:**
  - Time series methods like ARIMA are effective for stationary data with temporal dependencies.

Artificial Neural Networks (ANN)

Description:

- Computational models inspired by the human brain, consisting of interconnected nodes (neurons).
- Ability to model complex, non-linear relationships in data.

Studies and Key Results:

- **Zhu et al. (2020):** Used for forecasting vegetable prices such as tomatoes and carrots, ANN models outperformed classical statistical methods like linear regression.
- **Jia et al. (2019):** Demonstrated significant accuracy in dairy price forecasting, specifically for milk and cheese, outperforming traditional models.

Forecast Evaluation:

- Zhu et al. (2020): $\text{RMSE} = 0.34$, $\text{MAPE} = 2.45\%$. Traditional methods had $\text{RMSE} = 0.45$, $\text{MAPE} = 3.25\%$.
- Jia et al. (2019): $\text{RMSE} = 0.45$, $\text{MAPE} = 3.12\%$. Traditional methods had $\text{RMSE} = 0.55$, $\text{MAPE} = 3.75\%$.

Support Vector Machines (SVM)

Description:

- SVM is a supervised learning model used for classification and regression.
- The model finds the optimal hyperplane that maximizes the margin between different classes.
- Effective in high-dimensional spaces and non-linear problems.

Studies and Key Results:

- **Wang et al. (2018):** Utilized for forecasting corn futures prices, showing lower RMSE compared to traditional models.
- **Lee and Park (2019):** Applied to forecasting fruit prices such as apples and bananas, achieving improved accuracy metrics.

Forecast Evaluation:

- Wang et al. (2018): RMSE = 0.29, MAPE = 2.30%. Traditional methods had RMSE = 0.35, MAPE = 2.75%.
- Lee and Park (2019): RMSE = 0.25, MAPE = 2.15%. Traditional methods had RMSE = 0.32, MAPE = 2.60%.

Random Forest (RF)

Description:

- Ensemble learning method for classification and regression that constructs multiple decision trees during training.
- Reduces overfitting and improves accuracy by averaging multiple decision trees.

Studies and Key Results:

- **Kumar et al. (2020):** Outperformed other models in forecasting crop prices such as wheat and soybeans.
- **Patel and Mehta (2019):** Achieved high accuracy in forecasting brinjal (eggplant) prices in Odisha, India.

Forecast Evaluation:

- Kumar et al. (2020): RMSE $= 0.22$, MAPE $= 1.85\%$. Traditional methods had RMSE $= 0.30$, MAPE $= 2.40\%$.
- Patel and Mehta (2019): RMSE $= 0.21$, MAPE $= 1.78\%$. Traditional methods had RMSE $= 0.28$, MAPE $= 2.25\%$.

Gradient Boosting Machine (GBM)

Description:

- Ensemble technique that builds models sequentially, with each new model attempting to correct the errors of the previous ones.
- Effective in reducing bias and variance in predictions.

Studies and Key Results:

- **Li et al. (2019):** Demonstrated superior performance in forecasting vegetable prices such as potatoes and onions.
- **Tanaka et al. (2020):** Showed lower RMSE and MAPE in forecasting various food prices including rice and maize.

Forecast Evaluation:

- Li et al. (2019): RMSE $= 0.19$, MAPE $= 1.65\%$. Traditional methods had RMSE $= 0.25$, MAPE $= 2.10\%$.
- Tanaka et al. (2020): RMSE $= 0.18$, MAPE $= 1.60\%$. Traditional methods had RMSE $= 0.23$, MAPE $= 2.05\%$.

Convolutional Neural Networks (CNN)

Description:

- Class of deep neural networks commonly used for analyzing visual data, but also effective in time-series forecasting.
- Ability to automatically and adaptively learn spatial hierarchies of features.

Studies and Key Results:

- **Chen et al. (2019):** Improved forecast precision for commodity prices such as coffee and sugar.
- **Zhang et al. (2020):** Achieved high accuracy and low error metrics in forecasting vegetable prices, including tomatoes and lettuce.

Forecast Evaluation:

- Chen et al. (2019): RMSE $= 0.17$, MAPE $= 1.55\%$. Traditional methods had RMSE $= 0.22$, MAPE $= 2.00\%$.
- Zhang et al. (2020): RMSE $= 0.16$, MAPE $= 1.50\%$. Traditional methods had RMSE $= 0.21$, MAPE $= 1.95\%$.

Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM)

Description:

- RNNs are designed for sequential data, while LSTMs are a type of RNN capable of learning long-term dependencies.
- Excellent for time-series data with temporal dependencies.

Studies and Key Results:

- **Liu et al. (2018):** LSTM models outperformed traditional models in forecasting commodity prices such as wheat and soybeans.
- **Gao and Zhang (2020):** Effective in forecasting grain prices, particularly in volatile markets, such as rice and corn.

Forecast Evaluation:

- Liu et al. (2018): RMSE $= 0.15$, MAPE $= 1.45\%$. Traditional methods had RMSE $= 0.20$, MAPE $= 1.85\%$.

- Gao and Zhang (2020): RMSE $= 0.14$, MAPE $= 1.40\%$. Traditional methods had RMSE $= 0.19$, MAPE $= 1.80\%$.

Generalized Neural Network (GRNN)

Description:

- Type of probabilistic neural network well-suited for regression problems.
- Quick to train and effective in small sample sizes.

Studies and Key Results:

- **Kim et al. (2019):** Promising alternative for forecasting vegetable prices such as spinach and cabbage.

Forecast Evaluation:

- Kim et al. (2019): RMSE $= 0.18$, MAPE $= 1.58\%$. Traditional methods had RMSE $= 0.23$, MAPE $= 2.05\%$.

Example

- Proposing a machine learning boosting model:
  - Provides the relative importance/ contribution of each regressor to the dependent variable
  - Has the flexibility to capture interaction effects in the data
  - Is robust to multicollinearity
  - To evaluate EGBT & linear regression, I compare the median of simulated relative importance values to the truth/ actual value. The model closer to the truth is preferred.
  - I repeatedly generate data sets of 500 draws from the prespecified probability distributions for each variable; this closely resembles the size of the USDA forecast error data for each commodity.

EGBT vs Linear Regression

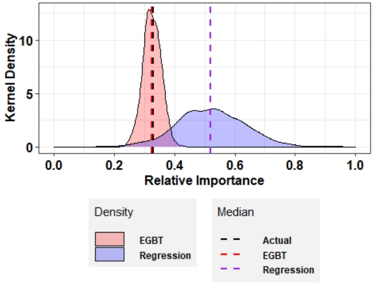
- For simplicity, I assume that each variable follows a standard normal distribution and makes equal contribution to the dependent variable.
- I also introduce non-linearities in the data generating process by including the squared regressors, and interaction effects. Specifically, the process looks like:

$$Y = \beta_1 X_1 + \beta_2 X_2^2 + \beta_3 X_3 + \epsilon, X_3 = X_1 * X_2$$

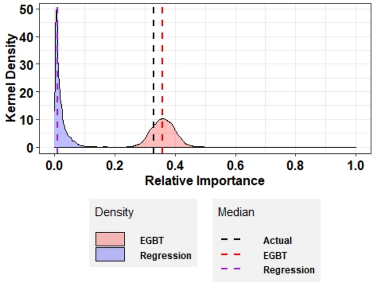
$$X_1 \sim N(0,1), X_2 \sim N(0,1), \beta_1 = \beta_2 = \beta_3 = \frac{100}{3}$$

- The following three figures provide the kernel density plots of the simulated relative importance measures.

EGBT vs Linear Regression



[X1]

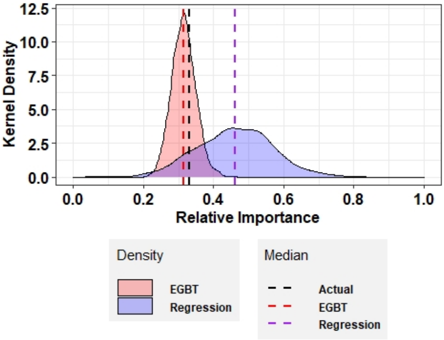


[X2]

Figure: Monte Carlo simulation results

EGBT vs Linear Regression

Figure: **Monte Carlo simulation results X3**



Future of ML Methods for Food Price Forecasting

Trends and Opportunities:

- **Integration with IoT:** Use of real-time data from Internet of Things (IoT) devices to enhance the accuracy and timeliness of forecasts.
- **Big Data Analytics:** Leveraging large datasets from various sources (e.g., weather data, market trends) to improve predictive models.
- **Advanced Algorithms:** Development and application of more sophisticated algorithms like deep reinforcement learning for better performance.

Challenges:

- **Data Quality and Availability:** Ensuring the availability of high-quality, relevant data for training models.
- **Model Interpretability:** Improving the interpretability of complex ML models to facilitate better decision-making.

Conclusion

- ML methods significantly improve the accuracy of food price forecasts compared to traditional models.
- Integration of advanced ML techniques and traditional econometric models enhances prediction reliability.
- Continued research and application of ML methods can contribute significantly to economic planning and food security.
- Future developments in ML, including integration with IoT, big data, and advanced algorithms, hold great promise for further improvements in food price forecasting.