AGRICULTURAL PRODUCT PRICING

James Bootsma, Aditya Manickam Arumugam, Krishang Karir, Abdul Aziz Essam, Mohd Daiyaan Nasir

***Abstract* – Farm product prices are influenced by an intricate interplay of climatic conditions and economic variables, making their prediction a challenging yet critical task for various stakeholders in the agricultural sector. This study aims to develop a predictive framework that forecasts future farm product prices using machine learning models, including Long Short-Term Memory (LSTM) networks, and traditional regression approaches, such as Linear Regression and Random Forest Regression. Leveraging historical climate data and economic indicators, such as GDP and oil prices, from Environment and Climate Change Canada and Statistics Canada, we aim to analyze their impact on farm product price prediction. Advanced feature engineering, along with ensemble modeling techniques, are applied to enhance prediction accuracy. Our goal is to generate actionable insights that can help farmers, policymakers, and consumers make informed decisions, thereby improving financial stability and sustainability in the agricultural sector.**

Keywords: Agricultural Price Forecasting, LSTM, Random Forest, Linear Regression, Climate and Economic Variables

1. INTRODUCTION

The agricultural sector plays a vital role in the global economy, contributing to food security, employment, and economic stability. However, farm product prices often fluctuate due to various climatic and economic factors[9]. Price volatility not only affects farmers' income stability but also impacts consumers' affordability and complicates decision-making for policymakers[11]. Understanding and forecasting these price fluctuations are therefore crucial.

Traditional forecasting methods such as Linear Regression have been used to predict agricultural prices. However, these methods struggle to model non-linear relationships and temporal dependencies in data[3]. With the rapid advancement in machine learning, specifically Long Short-Term Memory (LSTM) networks, there is now an opportunity to overcome these limitations. LSTM networks are recurrent neural networks designed to handle time-series data, making them particularly suitable for capturing the long-term dependencies and non-linear relationships that are inherent in agricultural price forecasting[4].

This study focuses on comparing three different models for farm product price prediction: LSTM, Linear Regression, and Random Forest Regression. By integrating climate and economic data into these models, we aim to uncover insights into which approach offers the best predictive performance. The findings of this research have the potential to enhance forecasting accuracy and provide valuable support to farmers, policymakers, and other stakeholders in the agricultural sector.

2. DATA COLLECTION & PREPROCESSING

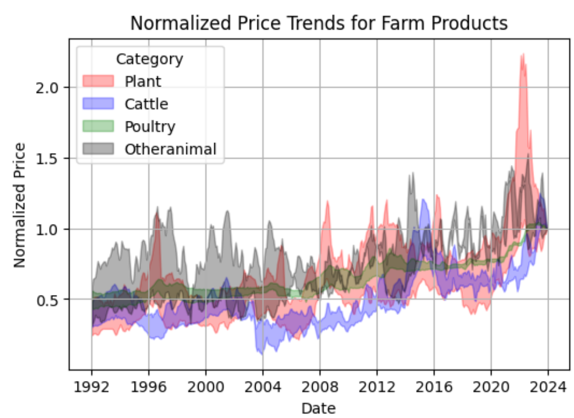

FIGURE 1: NORMALIZED PRICE TRENDS FROM 1992 TO 2024

The core data for this study was collected from reputable sources, including Environment and Climate Change Canada (ECCC)[2] and Statistics Canada[1]. The data covers a span of over 30 years, offering a comprehensive basis for model development. The primary variables included in the analysis are farm product prices, climate data (temperature and precipitation), economic indicators (GDP), and oil prices.

2.1 Data Sources

- Farm Product Prices: Monthly price data for 37 agricultural products were sourced from Statistics Canada[1]. Given the consistency and robustness of the data, 19 products were selected for the analysis, ranging from grains to livestock. See

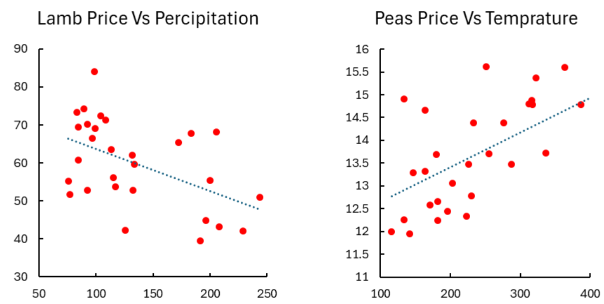

FIGURE 2:(LEFT) PRICE OF LAMB WITH RESPECT TO PRECIPITATION. (RIGHT) PRICE OF PEAS WITH RESPECT TO TEMPERATURE

Figure 1 for a spread of the prices based on the categories we defined.

- Economic Data: Gross Domestic Product (GDP) data was obtained from Statistics Canada[1], specifically focusing on agricultural production and filtering it to reflect the economic activities related to agriculture.
- Oil Prices: Diesel and gasoline prices were aggregated from provincial data to assess their influence on the transportation and production costs in agriculture[1].
- Climate Data: Temperature and precipitation data were provided by ECCC[2] , which were crucial to understanding the environmental factors affecting crop yields and farming conditions. Figure 2 show the impact temperature and precipitation have on the agricultural prices.
- Price data from Statistics Canada[1] reveals significant regional variations across provinces. Figure 3 shows a heatmap which visually demonstrates these differences, with color gradients indicating the extent of price variation in each province.

2.2 Data Cleaning & Transformation

The data underwent several preprocessing steps to ensure its quality and relevance for modeling:

- Handling Missing Data: Missing values in agricultural price data were imputed based on national averages for the month, adjusted by the percentage difference between national and provincial prices. Missing values in other variables were handled using linear interpolation, ensuring continuity in the dataset.
- Feature Engineering: We created lagged variables to account for temporal dependencies. These lagged features allowed the model to account for delayed effects, such as the impact of GDP or oil prices on prices for farm products in subsequent months.
- Log Transformation: To normalize the distributions and handle price fluctuations more effectively, log transformations of product prices were applied. Additionally, year-over-year (YoY) changes were calculated to facilitate comparisons across different periods.
- Response Variable: The YoY percentage change in the log-transformed price was used as the target variable to be predicted by the models.

FIGURE 3: HEATMAP OF STEER PRICES IN CANADA, 2022

These preprocessing steps resulted in a clean, comprehensive dataset that includes monthly data from 1992 to the present, ensuring robust analysis and reliable results.

3. METHODOLOGY

This study compares three different modeling approaches to predict agricultural prices: Linear Regression, Random Forest Regression, and Long Short-Term Memory (LSTM) networks. The goal is to assess the suitability of these models for forecasting farm product prices by considering factors such as computational complexity, predictive performance, and interpretability.

3.1 Linear Regression

Linear Regression was used as the baseline model for comparison. This method assumes a linear relationship between the predictor variables (temperature, precipitation, GDP, and oil prices) and the target variable (price change). Although simple, it provides valuable insights into the impact of individual features and serves as a performance benchmark for more complex models[3].

Model Implementation:

- Feature Selection: LASSO (Least Absolute Shrinkage and Selection Operator) regularization was applied to feature selection, ensuring that only the most important predictors were used in the model.
- Interaction Terms: Interaction terms between the categorical and numerical variables were introduced to capture different trends between categories, such as the combined effects of oil prices on cattle versus plant prices.
- Evaluation Metrics: The model's performance was evaluated using metrics such as Root Mean Squared Error (RMSE) to measure prediction accuracy.

3.2 Random Forest Regression

Random Forest Regression was chosen due to its ability to model non-linear relationships and handle large, high-dimensional datasets. It is a powerful ensemble technique that aggregates the predictions of many decision trees to improve accuracy and reduce overfitting[5].

Model Implementation:

- Parameter Tuning: The model's hyper-parameters, including the number of estimators, maximum depth, and minimum samples per leaf, were optimized using grid search.
- Handling Categorical Variables: Label encoding was used for categorical variables, such as the month of the year and geographical region, to incorporate them into the model effectively.
- Evaluation Metrics: Similar to Linear Regression, the Random Forest model was evaluated based on RMSE, comparing its predictive accuracy with the baseline model.

3.3 Long Short-Term Memory (LSTM) Networks

LSTM networks were implemented to capture the temporal dependencies and non-linear patterns inherent in the agricultural price data. LSTMs are particularly suited for time-series forecasting tasks due to their ability to maintain memory across time steps, which allows them to model complex trends and seasonality in the data[4].

Model Implementation:

- Sequence Generation: Data was organized into sequences, where each sequence corresponds to a series of historical monthly observations used to predict the next month's price.
- Model Architecture: The LSTM model was built with multiple layers, including input, LSTM, and dense layers, with dropout regularization to prevent overfitting.
- Optimization: The model was trained using the Adam optimizer, and early stopping was implemented to prevent overtraining and to ensure optimal model performance.

4. RESULTS & DISCUSSION

Each of the models was evaluated using a hold-out validation dataset, which consisted of data from 2020 onward. The results were analyzed based on their ability to generalize to unseen data and their performance in predicting the future price of agricultural products.

4.1 Linear Regression:

The LASSO regression results are as follows:

$$
\begin{aligned}
p_t = \beta_1 p_{t-1} + \beta_2 p_{t-3} + \beta_3 p_{t-4} + \beta_4 p_{t-12} \\
+ \beta_6 c_{cat} * p_{t-5} + \beta_7 c_{pla} * p_{t-1} \\
+ \beta_8 c_{cat} * d_{t-1} + \beta_9 c_{cat} * d_{t-2} \\
+ \beta_0 + \epsilon_t
\end{aligned}
$$

p represents the change in price
c denotes the category (cattle, plant, other, poultry)
d is the price of diesel

As anticipated, the previous month's price (p_{t-1}) emerged as a strong predictor, along with lag variables from three and four months prior (p_{t-3}) and(p_{t-4}). We also included the price from two months prior (p_{t-2}) to capture any intermediate effects.
The influence of previous prices varied by product category, with cattle prices notably affected by diesel prices due to the significant role of transportation and feed costs in cattle farming. Interestingly, climate and GDP variables did not significantly impact the regression outcomes.

The refined regression equation is:

$$
\begin{aligned}
p_t = 0.76 p_{t-1} - 0.001 p_{t-2} + 0.03 p_{t-3} + 0.06 p_{t-4} - 0.13 p_{t-12} \\
+ 0.04 c_{cat} * p_{t-1} + 0.10 c_{pla} * p_{t-1} \\
+ 0.05 c_{cat} * d_{t-1} + 0.01 c_{cat} * d_{t-2} \\
+ 0.004 + \epsilon_t
\end{aligned}
$$

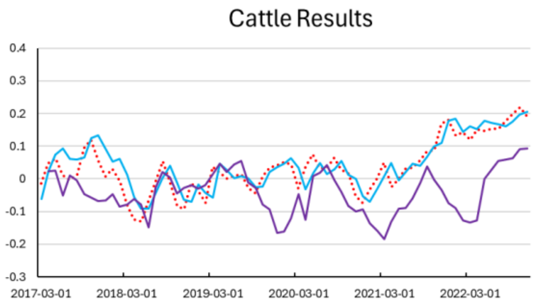

FIGURE 4: CATTLE PRICE TRENDS FROM 2017 TO 2022

Linear Regression performed adequately for products with relatively stable price patterns. However, it struggled with products experiencing higher volatility, often producing overly simplistic predictions that assumed future prices would closely follow the previous month's trend, with minor adjustments based on external factors. This is evident in Figure 4 and Figure 5, where the linear regression model's predictions closely track the previous month's prices but fail to capture more significant fluctuations.

4.2 Random Forest Regression

Random Forest Regression did not perform as well as simply predicting the previous month's value. This likely indicates that there was too much noise in the data for the model to capture meaningful patterns. The model's predictions were often less accurate than the baseline, suggesting that the complexity of the data overwhelmed the model's ability to generalize.

4.3 LSTM Networks

Similarly, the LSTM model struggled to outperform the simple baseline of predicting the previous month's value. The high level of noise in the data likely contributed to the model's inconsistent performance. The LSTM's predictions were often erratic, indicating that the model had difficulty learning from the noisy data. This is evident in Figure 4, where the LSTM model's predictions follow somewhat similar patterns to the actual prices but also exhibit random fluctuations that deviate significantly from the true values. This inconsistency highlights the challenges the LSTM model faced in capturing the underlying trends amidst the noise.

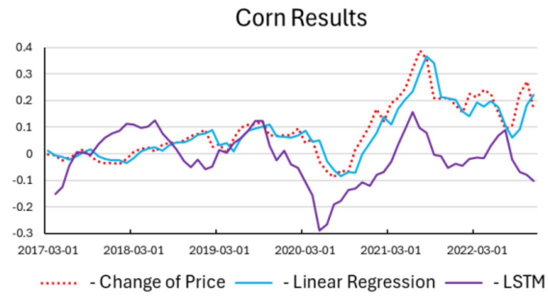

FIGURE 5: CORN PRICE TRENDS FROM 2017 TO 2022

4.3 Future Work

Future work will involve running these models on the residuals (errors) of a linear regression model built solely on the lag features of the previous price. By first modeling the primary trend using a simpler linear regression approach, we can isolate and analyze the residuals, which represent the unexplained variance. Applying more complex models like Random Forest and LSTM to these residuals may help capture any remaining patterns in the data that the linear model might miss. This two-step approach aims to improve the overall predictive performance by leveraging the strengths of both linear and non-linear modeling techniques. Additionally, further feature engineering and data preprocessing steps, such as noise reduction and the inclusion of more relevant external variables, will be explored to enhance model accuracy and robustness.

5. CONCLUSION

This study assessed the effectiveness of Linear Regression, Random Forest Regression, and Long Short-Term Memory (LSTM) networks in forecasting agricultural prices using climate and economic data. Linear Regression performed well for stable price patterns but struggled with volatility. Random Forest and LSTM models showed potential in capturing complex relationships but were hindered by high noise levels in the data, leading to inconsistent predictions. The findings highlight the need for more advanced models and feature engineering techniques to better capture the complexities of agricultural price forecasting. Future research should focus on refining these models, incorporating additional relevant variables, and exploring hybrid approaches that combine the strengths of different methodologies. Enhancing model accuracy and reliability will support better decision-making for farmers, policymakers, and other stakeholders in the agricultural sector.

References

1. **Statistics Canada**. (n.d.). *Agricultural and food statistics*. [Online]. Available: https://www.statcan.gc.ca/eng/start
2. **Environment and Climate Change Canada**. (n.d.). *Climate Data Online*. [Online]. Available: https://climate.weather.gc.ca/
3. Li, S., & Li, X. (2020). *Prediction of agricultural prices based on machine learning: A comprehensive review*. *Journal of Agricultural Economics*, 71(4), 912-926.
4. Amram, S., & Zhang, Y. (2019). *Application of Long Short-Term Memory (LSTM) in time series forecasting for agricultural price predictions*. *IEEE Access*, 7, 131584-131596.
5. Yang, Y., & Zhang, J. (2020). *Ensemble forecasting for agricultural price prediction using random forest regression*. *Journal of Forecasting*, 39(3), 459-468.
6. Pandey, D., & Sharma, D. (2021). *A machine learning approach for forecasting agricultural price volatility*. *Computers and Electronics in Agriculture*, 185, 106107.
7. Ghosh, S., & Ghosh, S. K. (2021). Exploring the effects of climatic and economic factors on agricultural prices using machine learning techniques. Agricultural Systems, 187, 103027
8. **OECD-FAO Agricultural Outlook 2020-2029**. (2020). *Agricultural price forecasting*. [Online].
9. **World Bank**. (2021). *Global Economic Prospects: The impact of oil price fluctuations on agriculture*. [Online]. Available: https://www.worldbank.org/en/publication/global-economic-prospects
10. **United Nations Food and Agriculture Organization (FAO)**. (2020). *Food price index*. [Online]. Available: https://www.fao.org/worldfoodsituation/foodpricesindex/en/
11. **U.S. Department of Agriculture (USDA)**. (2020). *Economic Research Service: Agricultural Prices*. [Online].
12. **National Renewable Energy Laboratory (NREL)**. (2019). *Energy consumption in agriculture and its relationship with oil and gas prices*. [Online]. Available: https://www.nrel.gov/research/
13. Babcock, B. A. (2012). The impact of US biofuel policies on agricultural price levels and volatility. China Agricultural Economic Review, 4(4), 407-426. DOI: 10.1108/17561371211284786.