AALBORG UNIVERSITY BUSINESS SCHOOL

Data-Driven Business Modelling and Strategy

Project Report

CITRUS FRUITS PRICE PREDICTION

A Focus on the EU Market

Supervisor:	Roman Jurowetzki
Group members:	Thi Minh Nguyen Karolina Motyka Siyao Zhang

Aalborg, December 2024

Table of Contents

1. Introduction

1.1 Background

Citrus is one of the world's major fruit crops. Citrus fruits, such as oranges, lemons, mandarins, and clementines, are wide-cultivated and well-accepted because of their pleasant flavor, aroma and important nutrients.

The EU is one of the largest global markets for citrus fruits, which is both a major producer and consumer of citrus fruits[1]. EU neighboring citrus producers, such as Egypt, South Africa mainly supplement EU production, ensuring a steady supply throughout the year, especially during off-seasons in the EU.

Predicting crop prices is a critical aspect of agricultural market stability and risk management. Citrus fruits dominate the EU's fresh fruit imports, they are pivotal in the retail market, especially in organic and sustainably grown produce. Predictive analytics provides a tool for distributors and retailers to plan logistics, avoid overstocking, and reduce food waste by matching imports or storage with market demand. To consumers, this helps stabilize citrus prices by avoiding sudden supply shortages or gluts.

Traditional crop price predictive approaches focus primarily on yield predictions based on environmental and biotic variables. However, crop prices are influenced by additional factors, such as market trends, policy changes, and unforeseen events like natural disasters.

In EU, citrus imports are highest during the winter months when local EU production cannot fully meet demand[2]. Southern Hemisphere countries, such as South Africa supply citrus during their harvest seasons, complementing EU production cycles. The dynamics of supply chain also effects citrus fruits prices in the EU. For instance, events like the COVID-19 pandemic and the Russia-Ukraine war have disrupted global supply chains, affecting the availability and cost of citrus imports.

The policy changes, including changes in both agricultural policies and trade regulations, could affect the citrus prices in the EU. The Common Agricultural Policy (CAP)

[1] https://fas.usda.gov/data/european-union-citrus-semi-annual-4

[2] https://citrusindustry.net/2024/07/12/european-citrus-production-update/

in the EU[3] subsidies and support for local producers may reduce the competitiveness of imports. Stricter EU regulations, such as cold treatment requirements to prevent pests, can increase the cost of citrus imports from certain countries or regions[4].

Natural disasters in citrus-exporting countries can significantly impact citrus fruits prices in the EU due to disrupted supply chains and reduced yields. In recent years, South Africa faced frost events that affected citrus crops[5], limiting exports to the EU. Severe flooding in Valencia and other Spanish regions in October 2024 caused over €192 million in damage to the citrus industry[6]. These natural disasters may also cause chain effects in the trading market, contributing to price volatility.

1.2 Research Objectives

Based on the above background information, the primary goal of this research is to develop a data-driven, predictive model for citrus fruit prices in the export-import context in the EU.

Oranges, lemons, mandarins, and clementines are among the most representative of the citrus family due to their global popularity, versatility, and nutritional benefits. Oranges are one of the most widely cultivated fruits, as well as the most traded citrus fruit in the EU. Lemons are a key export crop for Spain[7], and are integral to European cooking, used in marinades, dressings, and desserts. Mandarins are a major part of the EU citrus market, especially in Spain, and they are especially popular during winter holidays in Europe, symbolizing prosperity and well-being. Clementines, a hybrid variety of mandarin, are a prized citrus fruit in the EU due to their seedless nature and sweetness, and are strongly associated with Christmas in many EU countries. Therefore, this research chooses these fruits to represent the overall citrus family.

[3] https://agriculture.ec.europa.eu/common-agricultural-policy_en

[4] https://iifiir.org/en/news/importing-citrus-fruits-into-the-eu-cold-treatment-is-now-mandatory

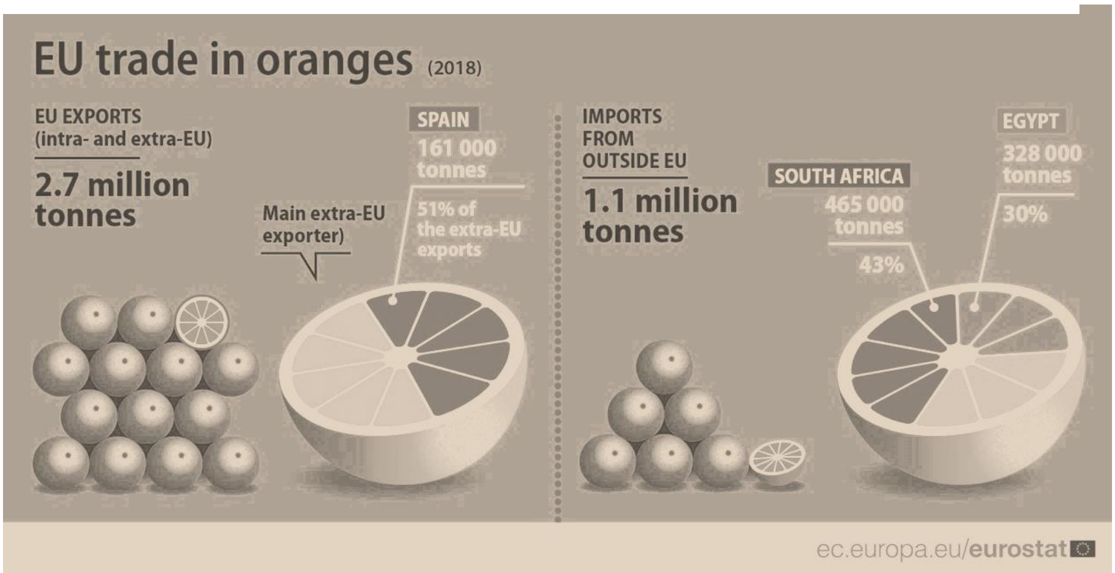
[5] https://www.freshplaza.com/north-america/article/9643760/exceptional-cold-causes-damage-of-hundreds-of-millions-of-rands-to-south-africa-s-winter-veg-and-citrus/

[6] https://www.freshplaza.com/north-america/article/9681107/severe-floods-and-crop-losses-in-spain-disrupt-global-citrus-and-orange-juice-supply/

[7] https://fas.usda.gov/data/european-union-citrus-annual-1

This research use Spain, South Africa and Egypt as the main suppliers of citrus fruits to the EU. Spain is the dominant citrus producer in Europe, particularly for oranges, mandarins, and lemons, due to its Mediterranean climate and advanced agricultural practices. South Africa, with an efficient citrus industry focused on high-quality exports, has citrus season that aligns with the EU's off-season (summer months), making it an essential supplier when European production is low. Egypt is known for its cost-competitive citrus exports, especially oranges, which have gained a significant share in the EU market. Egypt's location near the Mediterranean also ensures efficient logistics and faster delivery times to EU markets.

Figure 1 EU Trade in Oranges



(Source: Eurostat)

Specific objectives in this research include:

1. Identifying key variables affecting citrus fruits prices. Citrus fruit prices in the EU are influenced by a range of variables, including environmental, economic, and geopolitical factors. This objective aims to comprehensively identify and analyze these variables to form the foundation of the predictive model. Key factors include climatic conditions in producing countries (e.g., droughts in Spain, frost in South Africa), global market trends (e.g., demand for organic citrus), production costs, and trade policies like EU tariffs on each exporting country. By leveraging data from agricultural reports, weather databases,

and historical price statistics, this objective seeks to quantify the impact of each variable on price fluctuations.

2. Testing and tuning machine learning models for predictive accuracy. This objective focuses on building and refining machine learning (ML) models to predict citrus fruit prices in the EU accurately. The research will begin by testing various algorithms. Each model’s performance will be measured using metrics like RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and R-squared. Particular emphasis will be placed on hyperparameter tuning to optimize model performance. Techniques like grid search, random search, and Bayesian optimization will be employed to fine-tune parameters such as learning rate, tree depth, or regularization strength. To improve generalization, cross-validation strategies will ensure the model is robust to unseen data. Furthermore, this objective will evaluate the interpretability of the ML models, particularly in identifying the most impactful variables on price changes. By iteratively testing and refining models, the goal is to strike a balance between accuracy, computational efficiency, and real-world applicability.

3. Proposing applications for large-scale agricultural planning and sustainability. The final objective is to translate the model’s predictions into actionable insights for stakeholders in the citrus supply chain, including farmers, distributors, policymakers, and consumers. Accurate price forecasts can guide agricultural planning by helping farmers decide optimal harvest times and crop choices. For policymakers, the model can serve as a tool for mitigating market shocks by adjusting trade policies or providing subsidies during crises. Additionally, the research will explore the role of price predictions in advancing sustainability goals, such as reducing food waste by aligning production with demand and promoting resilient agricultural practices in the face of climate change. The model's output could also enhance supply chain logistics, minimizing transportation inefficiencies and carbon footprints. Ultimately, this objective seeks to position the predictive model as a critical decision-support system that fosters economic stability and environmental sustainability in the EU citrus market. Partnerships with agricultural agencies and industry stakeholders will be proposed to ensure practical implementation of these applications.

1.3 Significance

This research on predicting citrus fruit prices in the EU bridges a gap in crop price prediction by integrating localized and global trade variables. It holds significant importance across economic, agricultural, environmental, and policymaking domains.

Accurately predicting citrus prices can stabilize markets by reducing price volatility, benefiting both producers and consumers. Farmers and exporters often face economic losses due to unexpected price drops or spikes caused by natural disasters, trade restrictions, or demand fluctuations. By providing accurate forecasts, this research will enable stakeholders to make informed decisions about pricing, production volumes, and export strategies. Additionally, for import-dependent markets in the EU, particularly during off-seasons, better price predictions can help optimize procurement and reduce costs associated with overstocking or shortages.

Price forecasts directly impact farmers' decisions about crop selection, planting schedules, and resource allocation. By aligning production with market demand, farmers can improve profitability while minimizing waste. For example, if the model predicts a price increase for oranges due to supply shortages, farmers may prioritize planting or harvesting those crops. Furthermore, predictions can assist in mitigating risks associated with climate change, such as shifts in growing seasons or increased susceptibility to pests and diseases, by offering timely insights for adaptive agricultural practices.

The research aligns with the EU's sustainability goals, such as those outlined in the European Green Deal and Farm-to-Fork Strategy. Accurate predictions of price and demand can encourage sustainable farming practices, such as optimizing water usage, reducing pesticide reliance, and adopting precision agriculture. Additionally, by promoting demand-driven production, this research can help minimize the environmental footprint of overproduction or unnecessary imports.

For policymakers, price prediction models offer a tool to anticipate and mitigate market shocks. Governments can implement trade policies, subsidies, or strategic reserves to stabilize markets and protect consumers. The model can also inform long-term policy decisions related to trade agreements, import quotas, and climate adaptation measures.

2. Literature Review

Predictive analytics in agriculture can benefit producers, consumers as well as the trading market. Consequently, numerous studies have focused on crop yield and price prediction using machine learning methods, exploring diverse algorithms and input variables.

2.1 Existing Approaches

Research on crop yield prediction has primarily focused on understanding environmental and biotic factors that influence production. Since crop yield significantly affects crop prices, these studies provide a foundational understanding for price prediction.

Title	Algorithm	Input Data
Random forests for global and regional crop yield predictions[8]	Random forest, linear regression	Temperature, precipitation, nitrogen fertilizer application rate, soil density or content, latitude, irrigation etc.
Accurate prediction of sugarcane yield using a random forest algorithm[9]	Random forest	Temperature, precipitation, radiation etc.
An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning[10]	Random forest	Temperature, precipitation, crop type, soil density or content, radiometric etc.

[8] Jeong JH, Resop JP, Mueller ND, Fleisher DH, Yun K, et al. (2016). Random Forests for Global and Regional Crop Yield Predictions. *PLOS ONE*, 11(6): e0156571.

[9] Everingham, Y., Sexton, J., Skocaj, D., & Inman-Bamber, G. (2016). Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for Sustainable Development*, 36(2), 27.

[10] Filippi, P., Jones, E. J., Wimalathunge, N. S., Somarathna, P. D. S. N., Pozza, L. E., Ugbaje, S. U., . . . Bishop, T. F. A. (2019). An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. *Precision Agriculture*, 20(5), 1015-1029.

Building on crop yield prediction, research on crop price forecasting introduces additional complexities by incorporating market dynamics, such as trade policies, market access, and local pricing trends.

Title	Algorithm	Input Data
Crop price forecasting system using supervised machine learning algorithms[11]	KNN	Precipitation, maximum-trade, minimum support price, yield
Crop price prediction using machine learning[12]	Random forest, linear regression	Season, state, temperature, soil etc.

Because variables related to market dynamics are difficult to access directly from existing datasets, text-based data becomes essential. By leveraging Natural Language Processing (NLP) and Large Language Models (LLMs), unstructured text can be transformed into actionable features for crop price prediction.

Title	Algorithm	Input Data
Meta-Learning based adaptive crop price prediction for agriculture application[13]	LSTM	Market trend, season, irregularity, crop yield etc.
Large Language Models for Crop Yield Prediction[14]	LLM, GPT-4	Weather, soil, crop yield, remote sensing data, GIS
News Event-Driven Forecasting of Commodity Prices[15]	RNN	Events extracted from news articles

[11] https://www.irjet.net/archives/V6/i4/IRJET-V6I41037.pdf

[12] https://journal.esrgroups.org/jes/article/view/3961/3406

[13] D. K, R. M, S. V, P. N and I. A. Jayaraj, "Meta-Learning Based Adaptive Crop Price Prediction for Agriculture Application," *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2021, pp. 396-402.

[14] https://assets-eu.researchsquare.com/files/rs-4750823/v1_covered_a60f8e13-a03b-41f1-93e8-20f4efdb074a.pdf

[15] Chakraborty, S., Jagabathula, S., Subramanian, L., & Venkataraman, A. (2024). Frontiers in Operations: News Event-Driven Forecasting of Commodity Prices. *Manufacturing & Service Operations Management, 26*(4), 1286-1305.

While there has been significant progress in developing predictive models for staple crops such as wheat, rice, and maize, the focus on fruits, including citrus, remains relatively underexplored. Crops have been prioritized due to their global importance in food security and trade, while fruits are often seen as supplementary in diet and economy. This has resulted in a lack of sophisticated models tailored to the unique characteristics of fruit markets, such as perishability, seasonality, and greater sensitivity to consumer demand trends.

2.2 Gaps in Current Research

Existing predictive models for agricultural pricing often isolate factors like environmental conditions, historical pricing, or market dynamics instead of integrating them into a comprehensive framework. For instance, models focusing on climatic factors such as precipitation or drought often fail to include critical economic variables like trade volumes or consumer demand. Similarly, economic models using historical price trends rarely consider real-time influences like sudden weather anomalies or geopolitical events. This siloed approach results in predictions that may be accurate for specific variables but lack robustness and practical application in dynamic markets like citrus fruits. For citrus, where multiple factors interact (climatic disasters, tariffs) is essential. An integrated model combining these variables can better reflect real-world price dynamics, improving decision-making for stakeholders such as farmers, distributors, and policymakers. Such integration requires advanced machine learning techniques and access to diverse, high-quality datasets to capture interdependencies effectively.

Many existing studies on crop prices are regionally focused, emphasizing localized factors like domestic production costs or local weather conditions, while overlooking the global trade networks that increasingly influence agricultural markets. Citrus fruits in the EU, for instance, are heavily dependent on imports from countries like South Africa and Egypt during certain seasons. Price fluctuations in the EU are often driven by cross-regional dynamics, such as the cost of shipping, global tariffs, or competition from other trade blocs. Furthermore, trade policies are rarely accounted for in predictive models. By ignoring the interconnectedness of global trade, current models fail to capture important external influences on citrus prices. Addressing this gap requires integrating data on trade flows, tariffs, and geopolitical events into prediction systems. Such models can better account for the complexities of citrus trade, offering insights into how disruptions in one region might ripple across EU markets.

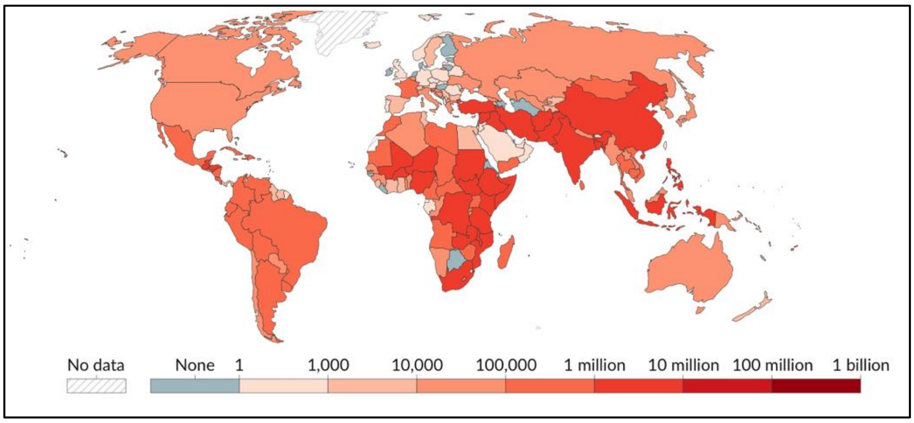
2.3 Research Contributions

As stated above, traditional models often focus narrowly on specific datasets, such as historical prices or weather patterns, offering only a fragmented understanding of market dynamics and neglecting broader variables that significantly impact pricing. This model bridges that gap by introducing a comprehensive forecasting system that integrates diverse data sources, incorporates underutilized external factors like disasters and policy changes, and employs advanced time-series forecasting techniques. By considering a wider spectrum of data, this study captures the multifaceted dynamics underlying agricultural markets, offering a comprehensive understanding of the factors influencing price fluctuations. This integrative approach enhances the model's ability to make accurate predictions under diverse and volatile conditions.

In recent years, our climate has undergone significant changes, with natural disasters such as floods, wildfires, droughts, and hurricanes becoming increasingly common. These events frequently have a devastating impact on agricultural systems, disrupting supply chains and reducing yields. While floods often devastate crops and hinder transportation, droughts represent an equally significant threat. Droughts can trigger cascading effects across multiple sectors, including agriculture, forestry, and water resources. They inhibit crop growth, reduce harvest yields, and contribute to price volatility in agricultural markets.[16] On the map below, we can observe that the scale of natural disasters is increasing, and more and more people are being affected by these events.

[16] https://climate.ec.europa.eu/climate-change/consequences-climate-change_en

Figure 2 Number of people affected by disasters in 2020



(Source: Our World in Data based on EM-DAT)

Similarly, policy changes and tax adjustments can lead to significant market shifts. By incorporating these external predictors, the model can more accurately reflect real-world disruptions, improving its adaptability in rapidly changing markets.

Traditional forecasting methods, such as ARIMA, often struggle with larger datasets, higher dimensionalities data or long-term forecasts. Using advance time-series forecasting techniques model provides precise, long-term price forecasts that align with market demands, helping businesses manage inventory, logistics, and procurement strategies effectively. Its ability to integrate all relevant variables into the forecasting process ensures it can adapt to market volatility, offering a competitive edge to users.

In conclusion, this study contributes to the field of agricultural price forecasting by introducing an innovative framework that integrates diverse data sources, incorporates critical external predictors, and leverages advanced time-series models.

3. Methodology

3.1 Data Sources

For this model, several key data sources are collected, including citrus fruit price data, weather data, natural disaster data, and tax-related information. By integrating these diverse datasets, the model provides a more comprehensive understanding of the factors driving market dynamics, significantly enhancing its predictive accuracy and practical relevance.

3.1.1 Historical Price Data:

Prices for citrus fruits such as lemons, mandarins, clementines, and oranges in European region are included. These historical prices, gathered from the European Commission databases, are an essential factor for identifying trends and making predictions about future market behavior.

Source: EU citrus price data

3.1.2 Weather Data:

Weather data plays a crucial role in understanding the environmental conditions that influence citrus cultivation. For this study, we collected data on variables such as precipitation, humidity, and temperature, which were sourced from meteorological services providing API access. The data focuses on key citrus-growing regions, including South Africa, Spain, and Egypt. These climate-related insights allow the model to assess how weather conditions impact crop growth, yield, and overall market dynamics.

Source: Open Meteo

3.1.3 Natural Disasters Data:

Natural disaster data is an essential component for understanding disruptions that affect citrus production and supply chains. We used a dataset covering global natural disasters since 2000, with a specific focus on citrus-growing regions such as South Africa, Spain, and Egypt. The data includes detailed information about the type of disaster, such as floods, droughts, or hurricanes, and its intensity, offering insights into the potential impact on agricultural outputs and market fluctuations.

Source: EM-DAT The International Disaster Database

3.1.4 Tariffs

Tax variations play a significant role in shaping the citrus market. We collected data on trade regulations, including import and export taxes, which directly impact production costs, trade dynamics, and market prices. By leveraging this information, the model demonstrates how fiscal policies contribute to price fluctuations, offering valuable insights for stakeholders.

Source: MAC MAP

3.2 Data Preprocessing and Features Engineering

3.2.1 Data Preparation

Preparing raw data for analysis is essential to ensure its consistency, accuracy, and compatibility with the modeling framework. This stage involves cleaning, transforming, and structuring data to optimize its usability. For our analysis, data processing focused on harmonizing multiple data sources, including historical price data, weather variables and disaster records. The data preparation process included key steps such as:

- *Historical Price Data*

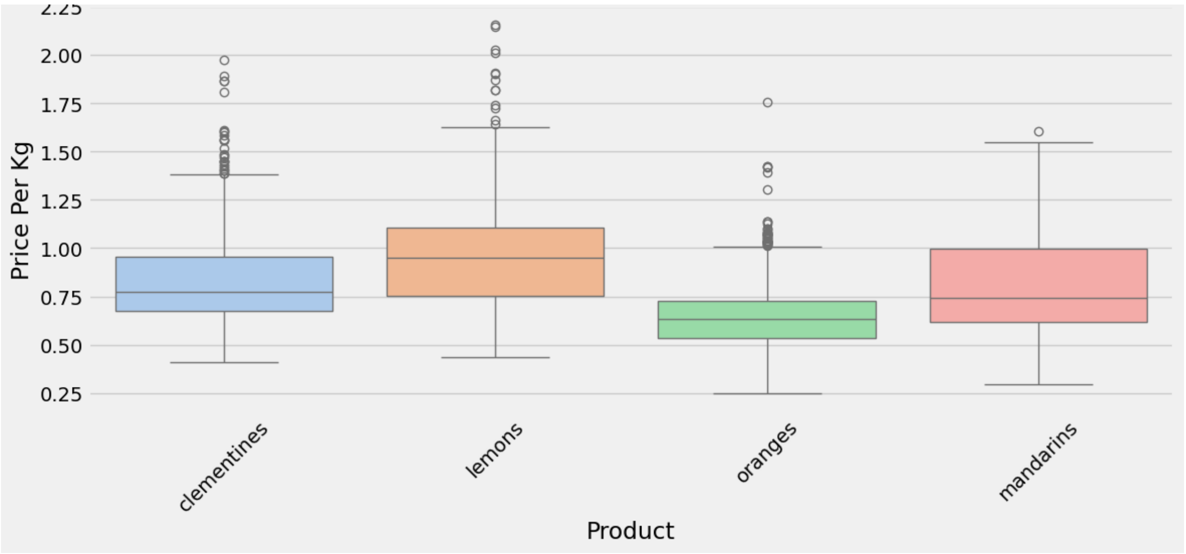
Following the extraction of data, the preparation process began with the historical price dataset. The data spans from 2004 and includes weekly prices for citrus fruits such as oranges, clementines, lemons, and mandarins in European region. To standardize the information, the original price values, provided per 100 kg, were converted to reflect the price per kilogram.

- *Handling Missing Weekly Data During Off-Season Periods*

In the citrus price dataset, missing weekly data during off-season periods naturally occurs when certain fruits are not available. To maintain data integrity, these weeks were excluded from the training set, as their absence reflects a genuine lack of market activity rather than a data collection issue.

When calculating lag and rolling features (such as past-week or past-month prices), these seasonal gaps posed a methodological challenge. Our approach was to fill missing off-season values with 0, which accurately represents the fruit's market unavailability. This method prevents the alternative approach of using prices from distant, potentially irrelevant time periods, thus ensuring consistent and meaningful lag and rolling feature calculations across all data points.

Figure 3 Correlation Matrix



- *Weather and Disasters*

Based on the fact that crop yield significantly impacts crop prices, making it essential to identify the top regions that supply citrus fruits to the EU. Incorporating data on weather conditions and disasters affecting these regions can provide valuable insights and enhance the accuracy of the model. These regions include:

Spain	Castellon, Valencia, Alicante, Murcia, Seville, Huelva, and Almería
Egypt	Beheira, Dakahlia, Fayoum, and Ismailia in Egypt
South Africa	Limpopo, Mpumalanga, and KwaZulu-Natal

Moreover, the weather condition and disasters happing during the growing and harvesting will also be taken into account. For each type of citrus fruit, this period is different and as follow:

Orange	Mandarin	Clementine	Lemon
9 months	9 months	8 months	6 months

- *Time and Lag Features*

Time-based features help capture temporal trends and seasonality in crop prices. Meanwhile, lag features help the model recognize price momentum and delays in market response to changes in supply or demand. There will be 2 type of lag features for 2 models: monthly and weekly.

3.2.2 Feature Set

Base features, used for both models

Table 1 Base features set, used for both models

Feature Name	Data Type	Description	Remark
quarter	int	Quarter of the predicted time.	
month	int	Month of the predicted time.	
product	string	Type of citrus fruit: orange, mandarin, clementine, lemon.	
_dsubgroup	string	Type of disaster with the most significant impact during the growing period. Takes the value "Nothing" if no disaster occurred.	These features are calculated for each country, meaning there are 3 times the number of features.
_ppl_affected	int	Total number of people affected by disasters during the period.	
_dcount	int	Number of regions growing citrus fruit affected by disasters during the period.	
_no_disaster	int	Total number of disasters during the period.	
_tem	float	Average daily temperature during the period.	
_rain	float	Average daily precipitation during the period.	
_sunshine	float	Average daily sunshine duration during the period.	
_tariff	float	Average tariff value (AVE) of the product in the year of prediction.	

- *Weekly features*

Table 2 Weekly features

Feature Name	Data Type	Description	Remark
lag_1w	float	Price of week n-1	There are weeks without price data for specific types of fruit, typically due to the off-season when those fruits are unavailable. In such cases, the missing price is filled with 0
lag_2w	float	Price of week n-2	
lag_3w	float	Price of week n-3	
lag_4w	float	Price of week n-4	
lag_5w	float	Price of week n-5	
lag_6w	float	Price of week n-6	

- *Monthly features*

Table 3 Monthly features

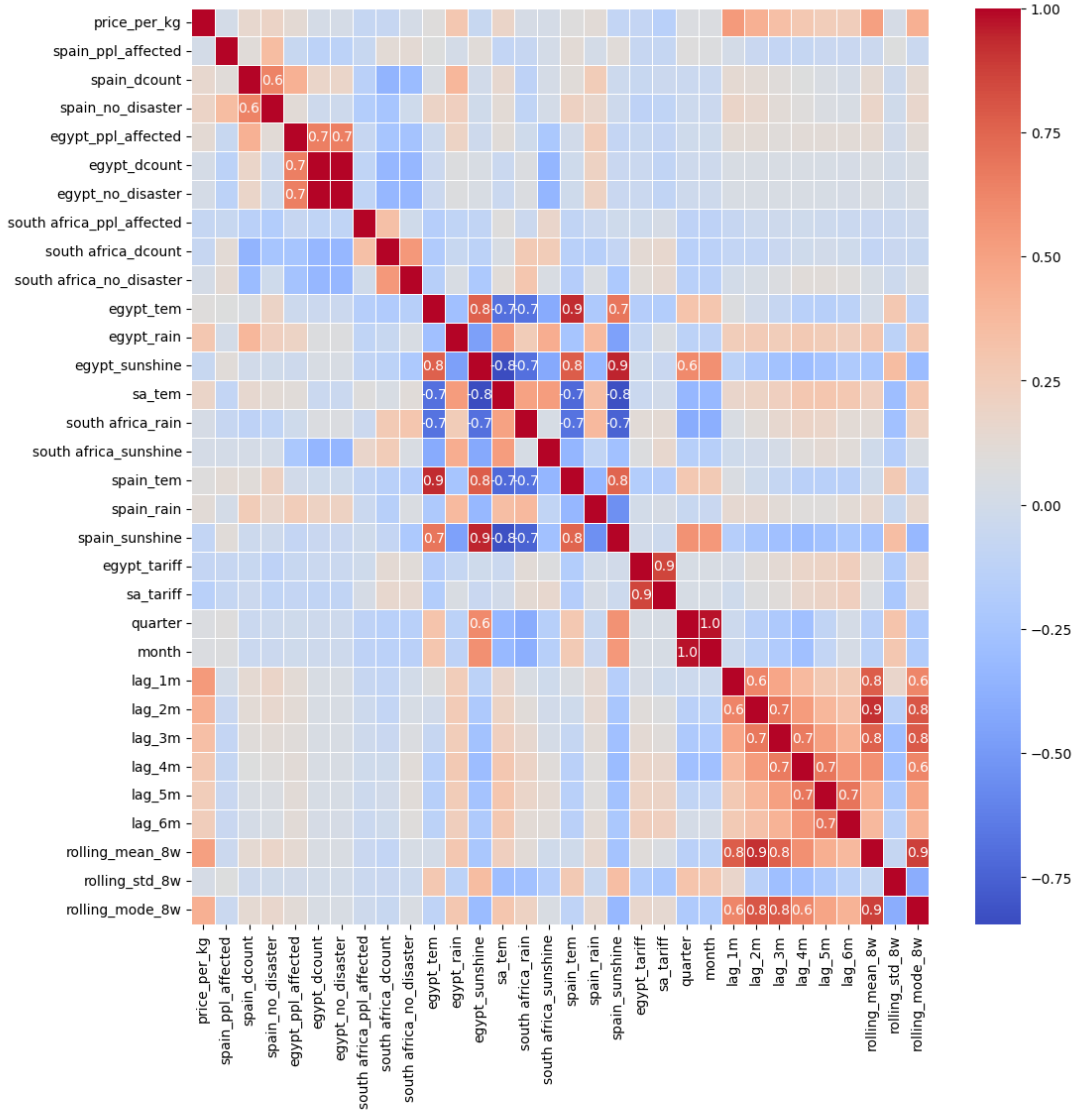
Feature Name	Data Type	Description	Remark
lag_1m	float	Price of week n-4	There are weeks without price data for specific types of fruit, typically due to the off-season when those fruits are unavailable. In such cases, the missing price is filled with 0
lag_2m	float	Price of week n-8	
lag_3m	float	Price of week n-12	
lag_4m	float	Price of week n-16	
lag_5m	float	Price of week n-20	
lag_6m	float	Price of week n-24	
rolling_mean_8w	float	Average price from week n-4 to n-12	
rolling_mod_8w	float	Median price from week n-4 to n-12	
rolling_std_8w	float	Std of prices from week n-4 to n-12	

The inclusion of rolling features in the monthly model, but not the weekly model, reflects the different time horizons and patterns being captured:

Weekly Model: Lag features alone are sufficient to capture short-term fluctuations and immediate price dynamics. The granularity of individual lags effectively represents rapid variations.

Monthly Model: Lag features capture longer-term trends, typically reflecting seasonal patterns that are more stable over time. These lags help the model understand broader, recurring market behaviors. However, by introducing ***rolling features (mean, median, standard deviation)***, the model gains insights into more recent fluctuations, offering a comprehensive view of the most current market dynamics. This combination allows the model to capture both stable seasonal trends and recent market volatility, improving its ability to predict prices based on both long-term patterns and short-term changes.

Figure 4 Correlation Matrix



The high correlation between lag features and weather conditions across different regions can be attributed to shared climatic patterns, geographic proximity, and synchronized seasonal trends. These factors often lead to similar weather impacts on crop yields and market dynamics in multiple regions. Despite the high correlation, these features should still be retained in the model as they provide valuable insights into temporal trends and regional dependencies, which are crucial for accurate predictions. Retaining them ensures the model captures the broader climate influences on crop prices and yield fluctuations, thus improving its predictive power.

3.3 Model Training

While the weekly model effectively detects immediate fluctuations and short-term patterns, the monthly model offers a more comprehensive perspective by capturing long-term trends, market cycles, and macroeconomic influences that impact prices over time. By using both models, we can ensure more accurate forecasting for both short-term and long-term decision-making, enhancing the robustness of price predictions across different time frames.

- **Algorithms:**

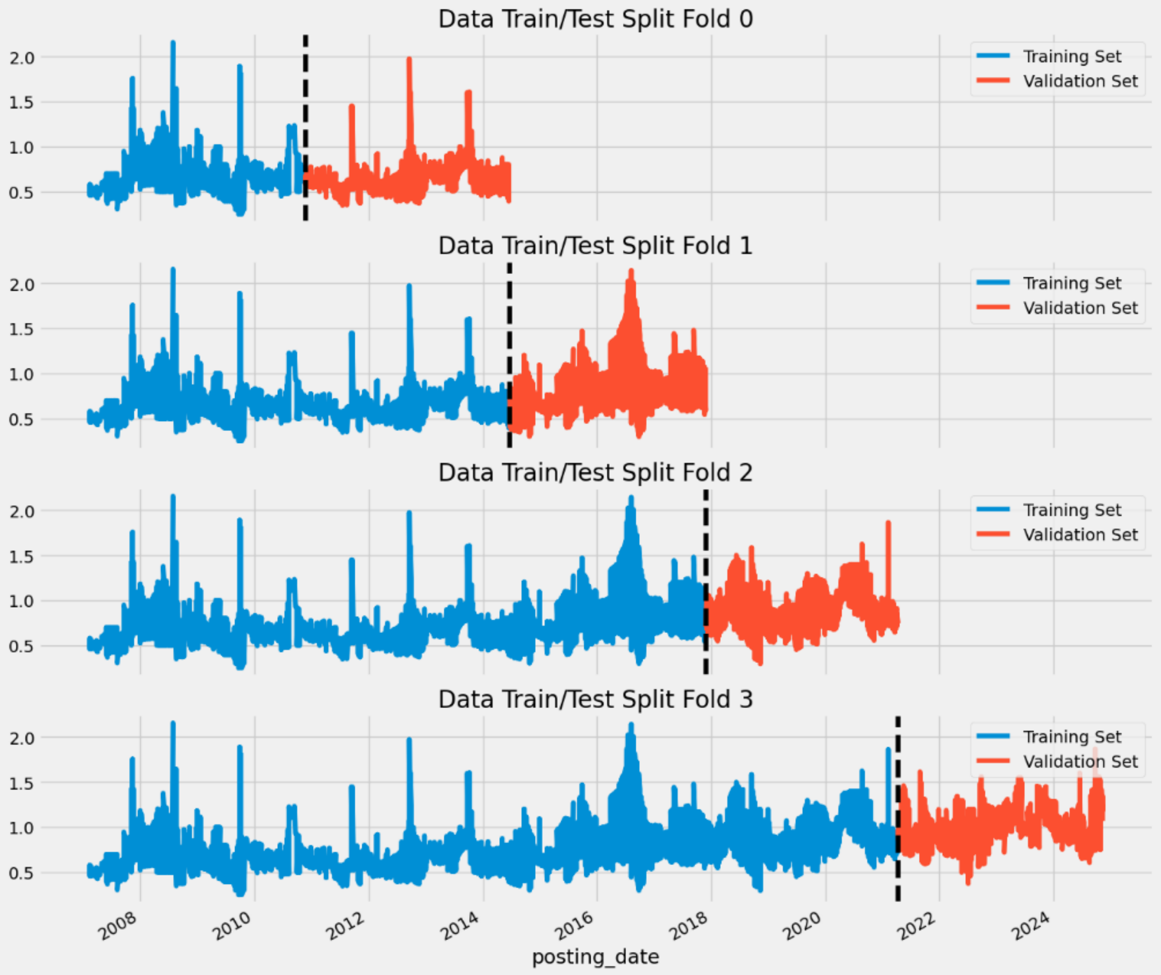
Baseline Model: Random Forest Regressor

Gradient Boosting: XGBoost Regressor, Catboost Regressor

- **Train-Test split for using 4-fold cross validation**

Initial test size = 20%. Due to limited data, 4 is the maximum number of folds to test.

Figure 5 Train/Test Split using 4-fold cross validation



3.3.1 Weekly Model

Table 4 Weekly Model Performance Evaluation

Model Performance Metrics:	**RandomForest**	**XGBoost**	**CatBoost**
MAE	0.099291	0.083208	0.082478
MSE	0.026509	0.018273	0.018100
RMSE	0.160113	0.134737	0.132355
R^2	0.503532	0.643363	0.667103

3.3.2 Monthly Model

Table 5 Monthly Model Performance Evaluation

Model Performance Metrics:	**RandomForest**	**XGBoost**	**CatBoost**
MAE	0.148678	0.145131	0.144214
MSE	0.040853	0.039960	0.038815
RMSE	0.202121	0.199901	0.197016
R^2	0.161241	0.1795680	0.203077

The results show that CatBoost consistently outperformed the other models, exhibiting the best accuracy and error metrics across the board, along with the highest R^2 value, which indicates a stronger capability to explain the variance in the data.

CatBoost performed better in predicting citrus fruit prices compared to Random Forest and XGBoost due to its efficient handling of categorical and numerical features. The dataset includes categorical features about fruit type and the category of disasters for each country. Using One-Hot Encoding will result in 19 more columns for Random Forest and XGBoost to handle. CatBoost, on the other hand, encodes them natively using target-based and permutation-driven encoding methods, preserving data compactness and avoiding sparsity.

Additionally, CatBoost captures complex interactions between categorical and numerical features, such as how disaster types influence citrus prices across countries and time. This is particularly beneficial for this dataset, which contains about 15 numerical features and categorical features calculated for multiple countries. By dynamically creating feature combinations during training, CatBoost effectively models these dependencies without manual intervention.

The model's ordered boosting mechanism also ensures robust generalization in our time-sensitive data, such as weekly and monthly price predictions, by reducing overfitting. These capabilities make CatBoost an ideal choice for this study, offering superior performance with minimal preprocessing compared to Random Forest and XGBoost.

3.4 Fine-tuning

CatBoostRegressor was selected as the primary predictive model due to its high performance. The model was retrained on the entire dataset, and Random Search was employed to identify the optimal parameter configuration.

3.4.1 Best Parameter Set

Table 6 Best Parameter Set

Parameter	Weekly	Monthly
random_strength	2	N/A
random_seed	42	42
od_wait	20	N/A
n_estimators	500	N/A
min_data_in_leaf	5	N/A
max_leaves	40	N/A
learning_rate	0.02	0.01
l2_leaf_reg	5	4
grow_policy	'Lossguide'	N/A
depth	4	4
verbose	100	100
loss_function	'RMSE'	'RMSE'
iterations	N/A	1500
random_state	N/A	42

3.4.2 Model Performance after Fine-tuning

Table 7 Model Performance after Fine-tuning

Metric	Weekly	Monthly
MAE	0.0769	0.1293
MSE	0.0152	0.0313
RMSE	0.1233	0.1769
R^2	0.6870	0.3599

Figure 6 Weekly Model - Plotting the prediction of CatBoostRegressor

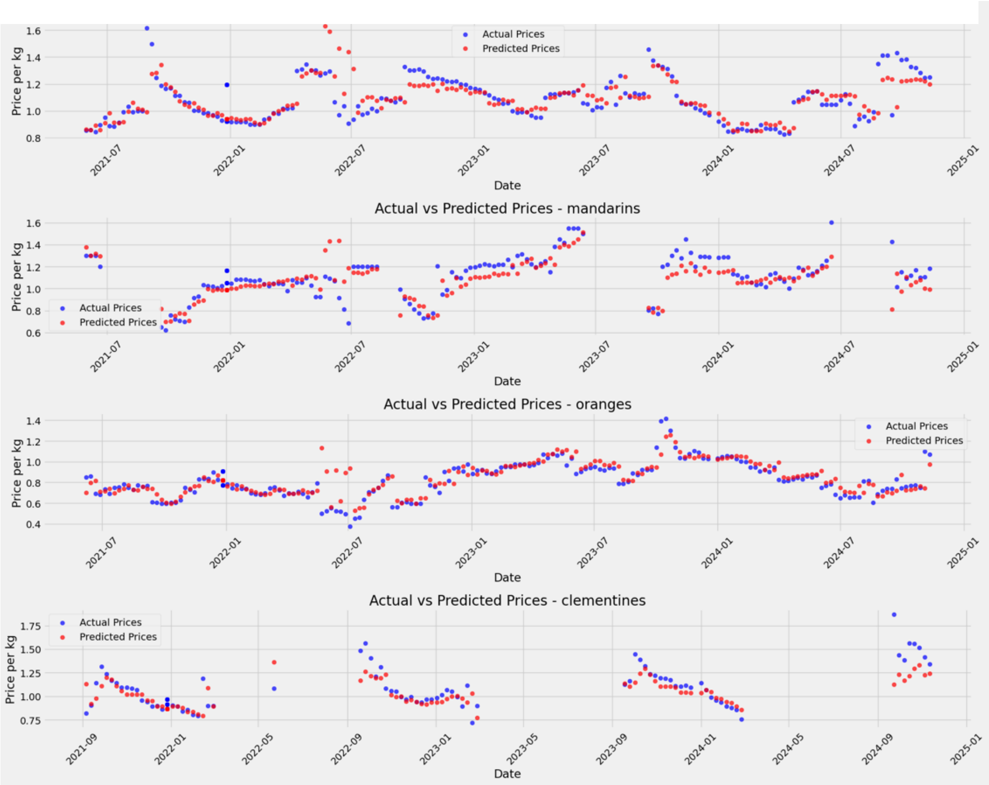
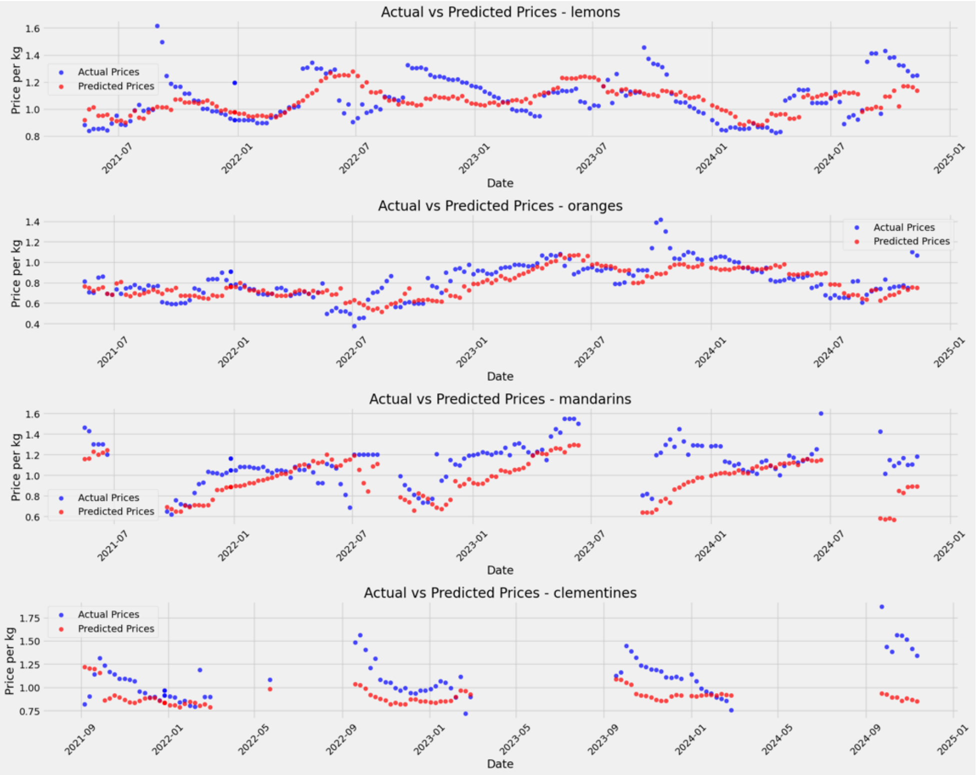


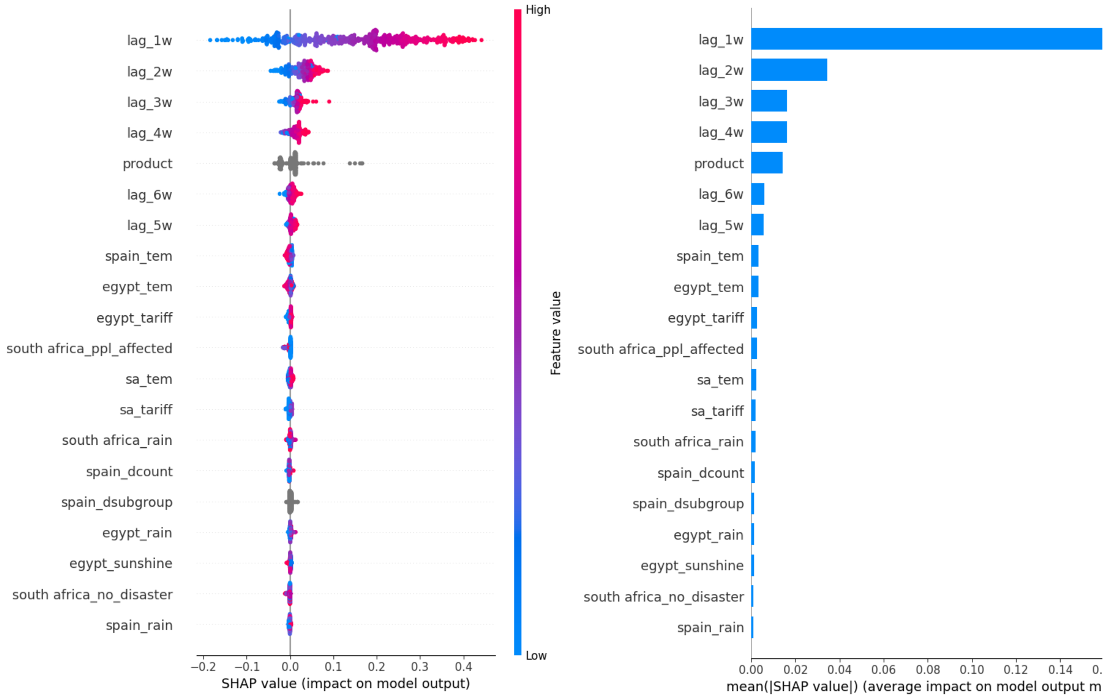
Figure 7 Monthly Model - Plotting the prediction of CatBoostRegressor



4. Model Explanation and Evaluation

4.1 Model Explainability

Figure 8 Feature Impact on Model Output: SHAP Value Analysis



The lag features (lag_1w, lag_2w, lag_3w) have the strongest positive influence on the model's predictions, highlighting the importance of recent historical prices as critical factors in forecasting future prices. The consistently high positive SHAP values for these lag features indicate their essential role in capturing short-term price dynamics.

In contrast, features related to weather (e.g., _tem, _rain) and tariffs exhibit a mixed impact. Some of these features positively influence predictions, while others show a negative relationship, reflecting their varied effects on pricing.

Features reflecting disaster events, such as _no_disaster and _dcount, generally show a negative SHAP relationship. This suggests that natural disasters tend to offset prices, albeit with a relatively small impact. This finding highlights the potential to expand and refine such features to enhance predictive performance.

4.2 Model Performance Compared to Other Researches

Predictive models for agricultural prices in existing literature mostly focus on staple crops (e.g., grains, soybeans) or industrial crops (e.g., cotton), and often achieve R^2 scores in the range of 60-80% for weekly or short-term forecasts, but rarely address highly perishable goods like citrus fruits, where market volatility is higher. This model, specifically targets the citrus fruit market which is underrepresented in predictive modeling literature, however, achieves a strong R^2 of 68.70% for weekly predictions, demonstrating a high level of accuracy in explaining short-term price variance. The performance is therefore competitive and contextually significant.

Table 8 Model Performance vs Existing Studies

Title	Crop type	R^2	RMSE	AVG target var
Random forests for global and regional crop yield predictions	Wheat	Not mentioned	0.32	2.68
	Maize	Not mentioned	1.13	6.77
	Potato	Not mentioned	2.77	19.93
Accurate prediction of sugarcane yield using a random forest algorithm	Sugarcane (January)	0.72	Not mentioned	
An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning	Wheat, barley and canola (July)	0.91	Not mentioned	
Crop price prediction using machine learning	Multiple crops	Not mentioned	485.07	2551.06

Many crops price predictive models in the literature focus only on one temporal scale, often failing to accommodate users with different forecasting needs. By separately addressing weekly and monthly forecasts, this model offers flexibility and better resolution for stakeholders who need price predictions at different timeframes. For example, short-term predictions (weekly) cater to traders and distributors, while long-term predictions (monthly) benefit policy-makers and large-scale agricultural planners. By combining climate, price history and trading factors, the model addresses the needs of diverse stakeholders, from farmers and traders to policymakers and distributors.

4.3 Model Performance without Disasters and Tariffs Data

To further investigate the impact of the disaster and trade tariff-related features, the models were retrained with these features excluded from the dataset. Both the weekly and monthly models were evaluated on this reduced feature set.

The results show a decrease in the R-squared (R2) value for both models:

- Weekly model R2 decreased from 0.6870 to 0.6760
- Monthly model R2 decreased from 0.3599 to 0.3418

This indicates that weekly price fluctuations are likely driven more by immediate and short-term factors, such as recent prices (lag features), rather than macroeconomic or external events like tariffs and disasters. Meanwhile, monthly price trends are more likely to be affected by broader market conditions, including seasonal patterns, tariffs, and the long-term impact of disasters. This is understandable, as such changes typically take time to fully manifest in the market.

Additionally, this presents an opportunity to enhance the study by incorporating more external factors to improve long-term predictions. These insights would be more valuable for both businesses and farmers, as they would help in better evaluating and anticipating market trends over extended periods.

5. Innovation

5.1 Business Model Innovation

In the context of Business Model Innovation, the project primarily targets small and medium-sized enterprises (SMEs) within the agriculture sector, which often face resource constraints when it comes to implementing advanced technologies such as predictive modeling. These businesses typically rely on traditional methods of forecasting, which are time-consuming, prone to errors, and insufficient in handling uncertainties such as weather changes and economic fluctuations. This project introduces a predictive model for citrus price forecasting, aiming to empower SMEs by integrating more accurate, data-driven decision-making processes that were previously out of their reach.

Service Blueprint: Price Forecasting Process for SMEs

- *Before the Predictive Model (Current Process)*

- Data Collection:

Manual Collection: SMEs rely on manual data collection, typically based on past sales data and market trends. This process is labor-intensive, often without any comprehensive integration of external factors.

Limited Variables: Data sources are usually limited to basic historical price data and seasonality patterns. Weather and disaster variables, which can significantly impact agricultural production and market prices, are typically not included in the planning process.

- Decision-Making:

Reactive Decisions: Decisions regarding pricing, production, and distribution are based on historical data and ad-hoc judgment. There is no formal predictive model in place, so decisions are made reactively rather than proactively.

Lack of Forecasting Tools: SMEs often do not have access to sophisticated forecasting tools, which results in an inability to anticipate market fluctuations and adjust strategies accordingly.

- Customer Value Proposition:

The value proposition focuses on offering competitive prices based on current market trends, but this approach lacks the foresight and stability needed to navigate price volatility or plan for future demand effectively.

- After Implementing the Predictive Model

- Data Integration:

Automated and Comprehensive Data Collection: The predictive model integrates multiple data sources, including historical prices, weather data, and disaster reports. This integration allows businesses to track factors that influence prices, such as temperature changes or rainfall in key citrus-growing regions, which were previously neglected in their planning.

Use of External Data: The model considers variables that were not typically included, such as the potential impact of extreme weather or natural disasters, improving forecasting accuracy and helping SMEs adapt to these external factors.

- Predictive Decision-Making:

Proactive Pricing and Production Strategy: By using the predictive model, SMEs can receive accurate price forecasts, allowing for proactive adjustments to pricing, production volumes, and inventory management. This reduces the risk of overproduction or underpricing and enables businesses to plan for the future more effectively.

Improved Risk Management: The model helps businesses anticipate price fluctuations, helping them mitigate risks associated with unforeseen events, such as weather disruptions or market downturns.

- Customer Value Proposition:

The value proposition evolves to emphasize data-driven strategies. SMEs can now offer more stable, reliable pricing to their customers based on accurate predictions of market trends, which builds trust and loyalty. By optimizing production and pricing, businesses can avoid waste and maximize efficiency, improving overall sustainability.

In this way, integrating a predictive model into an SME’s business model addresses their resource limitations by providing a cost-effective solution for better decision-making. It enables them to compete more effectively with larger companies by offering data-driven insights without requiring substantial investment in expensive technologies or external expertise. This approach fosters innovation by empowering SMEs to adopt more advanced and sustainable practices, ultimately improving their competitiveness in an uncertain and dynamic market environment.

5.2 Boosting Sustainability

5.2.1 Economic Resilience

The predictive model enhances economic sustainability by enabling SMEs to make more informed decisions in a volatile market. By accurately forecasting citrus prices and production trends, businesses can adjust pricing strategies and production volumes accordingly. This helps SMEs avoid losses due to market fluctuations and ensures that they can meet demand without overproducing, leading to more stable revenues and long-term viability. Such resilience is especially crucial for agricultural SMEs that face unpredictable external factors like price volatility and changing consumer demand.

5.2.2 Long-Term Viability through Data-Driven Innovation

The use of predictive modeling fosters a culture of innovation by providing SMEs with data-driven insights that help them anticipate market shifts and adapt to new trends. This technological adoption not only improves short-term decision-making but also positions businesses for long-term success. With more accurate predictions, SMEs can fine-tune their operations, innovate in response to new consumer preferences, and stay competitive in the market. This ability to innovate continuously ensures that businesses remain adaptable in the face of evolving challenges, contributing to their sustainability over time.

5.2.3 Social Sustainability: Supporting Local Communities

Sustainable business practices, informed by accurate forecasting, lead to more stable and fair employment within local communities. By optimizing production and reducing waste, SMEs can ensure that workers in agriculture have steady jobs and fair wages. Additionally, predictable production helps ensure a stable supply of goods to local markets, which benefits consumers by providing more affordable and consistent access to essential agricultural products. This enhances the overall social sustainability of the region, as businesses contribute to local economic growth and community stability.

5.2.4 Sustaining Agricultural Practices

Predictive modeling plays a crucial role in sustaining agriculture by helping SMEs anticipate the impact of environmental factors such as weather, disasters, and seasonal fluctuations. By incorporating these predictions, businesses can avoid overproduction or underproduction, reducing waste and minimizing the use of unnecessary resources. Moreover, better forecasting of climate conditions allows for the implementation of sustainable farming practices, such as adjusting planting and harvesting schedules to avoid extreme weather events. This proactive approach helps mitigate the environmental impact of agriculture, ensuring that practices remain sustainable while meeting market needs.

Bibliography

Articles

1. Chakraborty, S., Jagabathula, S., Subramanian, L., & Venkataraman, A. (2024). Frontiers in Operations: News Event-Driven Forecasting of Commodity Prices. *Manufacturing & Service Operations Management*, 26(4), 1286-1305.

2. D. K., R. M., S. V., P. N., & I. A. Jayaraj. (2021). Meta-Learning Based Adaptive Crop Price Prediction for Agriculture Application. *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 396-402.

3. Everingham, Y., Sexton, J., Skocaj, D., & Inman-Bamber, G. (2016). Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for Sustainable Development*, 36(2), 27.

4. Filippi, P., Jones, E. J., Wimalathunge, N. S., Somarathna, P. D. S. N., Pozza, L. E., Ugbaje, S. U., et al. (2019). An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. *Precision Agriculture*, 20(5), 1015-1029.

5. Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., et al. (2016). Random Forests for Global and Regional Crop Yield Predictions. *PLOS ONE*, 11(6): e0156571.

Webpages

1. https://agriculture.ec.europa.eu/common-agricultural-policy_en
2. https://climate.ec.europa.eu/climate-change/consequences-climate-change_en
3. https://citrusindustry.net/2024/07/12/european-citrus-production-update/
4. https://www.freshplaza.com/north-america/article/9643760/exceptional-cold-causes-damage-of-hundreds-of-millions-of-rands-to-south-africa-s-winter-veg-and-citrus/
5. https://www.freshplaza.com/north-america/article/9681107/severe-floods-and-crop-losses-in-spain-disrupt-global-citrus-and-orange-juice-supply/
6. https://iifiir.org/en/news/importing-citrus-fruits-into-the-eu-cold-treatment-is-now-mandatory
7. https://www.irjet.net/archives/V6/i4/IRJET-V6I41037.pdf
8. https://journal.esrgroups.org/jes/article/view/3961/3406
9. https://open-meteo.com
10. https://ourworldindata.org/explorers/natural-disasters?tab=map&time=2020&Disaster+Type=All+disasters&Impact=Total+affected&Timespan=Decadal+average&Per+capita=false&country=~OWID_WRL
11. https://assets-eu.researchsquare.com/files/rs-4750823/v1_covered_a60f8e13-a03b-41f1-93e8-20f4efdb074a.pdf
12. https://fas.usda.gov/data/european-union-citrus-annual-1
13. https://fas.usda.gov/data/european-union-citrus-semi-annual-4
14. https://agridata.ec.europa.eu/extensions/DashboardCitrus/CitrusTrade.html
15. https://www.emdat.be

List of Tables

List of Figures