Computer Science Department

Senior Project Final Report – Spring 2025

Title of the project:	"Using Machine Learning to Predict Food Prices"
Team Members:	Aruzhan Kassymova Dilnaz Olzhabayeva Ainur Zhumay Sherkhan Umurzak Amina Ismailova
Project Advisor/Co-Advisors	Prof. Michael Lewis

Executive Summary (10%)

Our project, "Using Machine Learning to Predict Food Prices," provides a systematic approach for the prediction of staple food prices across world markets. Food price volatility can have a large impact on food security, household well-being, and economic stability at the global level. These predictions canmade easy by retail, farm, and policy decisions.

Objectives and goals

- Collect datasets on food prices and relevant economic indicators (inflation rates, GDP in current US dollars, consumer price index, loan interest rates, unemployment rate, exchange rate to the US dollar, GDP growth, share of urban population)
- Data preprocessing: clean the collected data to extract relevant information, including food product names, countries and country codes, cities, currency types, longitude, latitude, and more
- Develop predictive models using machine learning algorithms, such as Linear Regression, Random Forest, XGBoost, LightGBM, CatBoost, SVR
- Monitor model performance
- Provide data visualization

We collected and preprocessed global datasets from public APIs such as the World Bank and Eurostat databases. Then we used BigQuery to combine the cleaned and standardized data into a centralized repository system. Outputs are presented through a sequence of graphical charts and animations that illustrate price trends and model potential market scenarios.

This paper explains our approach, tools and technologies used, and procedures to build, test, and assess the forecasting system. Through a comprehensive assessment of models and cases, we demonstrate how this solution is positively contributing to solving the food price volatility problem based on a computer-based approach.

Introduction (10%)

Forecasting food prices is crucial for economic stability, yet existing models lack accuracy and real-time adaptability. Our project aims to establish a machine learning framework for food price forecasting worldwide based on advanced predictive models. However, we encountered different challenges like limited historical data with missing values, real-time prediction with low latency, and the selection of appropriate machine learning models to make accurate predictions. To address these concerns, we used data preprocessing techniques to handle missing values, added geospatial coordinates to regional price variation, and applied strong models such as XGBoost and LightGBM.

The volatility and unpredictability of food prices present a significant challenge for stakeholders in agriculture, supply chain management, and policy-making. To address this challenge, machine learning models, particularly ensemble methods like XGBoost and LightGBM have emerged as effective tools for time-series forecasting. This study aims to compare these models in terms of predictive accuracy, computational efficiency, and robustness, evaluating how different hyperparameters and training epochs impact performance. Using historical food price data, we assess these models. The findings of this research provide insights into the most effective model for food price forecasting, contributing to informed decision-making in the agricultural and economic sectors.

The structure of this report is designed to provide a clear and logical progression of the study. It begins with an introduction to the problem, outlining the objectives and significance of the research. Following this, the report reviews relevant literature on existing food price forecasting methods, highlighting the strengths and weaknesses of current approaches. The methodology section then details the data collection process, the preprocessing steps, and the machine learning models employed in the study. Results and discussion section provides the evaluation of the model results, the comparison, as well as detailed analysis of results. Finally, the report has been concluded with the summary of results, implications on future research and recommendations for applications of machine learning in food prices forecasting.

Background/Related Work (15%)

Our project tackles a critical issue: the rapid shifts in food prices and their ripple effects on everyday life. When prices swing unpredictably, it hits everyone: families struggling with grocery bills, farmers managing crops, and governments balancing economic policies. Understanding these trends isn't just about numbers; it's about safeguarding food security, guiding policymakers, and stabilizing markets to protect communities.Food prices shape economies and livelihoods. By blending open data with proven machine learning methods, our project offers a starting point for smarter decisions.

To build a global view of food prices, we relied on open, ethical data sources. Public databases like the World Bank's API (wbdata), Eurostat, and HDX HAPI gave us access to macroeconomic indicators (e.g., inflation, GDP) and standardized country codes (via pycountry). These tools let us pull reliable, up-to-date information without compromising transparency.

After testing various machine learning methods, we chose to first try to build on the XGBoost model used by Peshevski et al. (2023), which analyzed food price trends across 12 European countries over ten years. Their approach, which was based on combining historical prices with economic indicators like inflation and trade data, achieved strong results, with an average accuracy score (R^2) of 0.85 during the 2020–2022 period. This gave us a reliable foundation to work from and test our own model.

We also adapted strategies from other studies. For example, Yidan Gao's 2024 research demonstrated how tailoring algorithms to specific food types improves predictions, like using Ridge Regression for grains or Lasso Regression for dairy. This inspired us to match models to data patterns rather than taking a one-size-fits-all approach. For short-term forecasts, at first we relied on time-series models (e.g., ARIMA), which excel at spotting trends in stable, historical data, helping us predict price shifts over weeks or months.

Existing projects on GitHub accelerated our work by providing templates for data sourcing and scope definition. Meanwhile, Yidan Gao's interface design, which was built with Python's PyQt5, guided us in creating a user-friendly dashboard (1). Our model is able to adjust parameters, visualize trends (e.g., milk prices in Kazakhstan), and compare predictions across food categories without needing technical expertise.

We tested algorithms like Random Forest, LightGBM, and Support Vector Regression (SVR) to handle diverse datasets, from sparse regional data to volatile price spikes. Public APIs, such as the World Food Programme's datasets, ensured transparency, letting us ethically source global data while avoiding proprietary restrictions.

Food price forecasting is complex, but by blending proven methods with open-source tools, we aimed to create a flexible, trustworthy model. While not perfect—gaps in regional data remain—this approach emphasizes reproducibility and ethical practices, offering a practical tool for policymakers, economists, and communities navigating food security challenges. To analyze price differences across countries, we prioritized macroeconomic factors based on their direct relevance to price dynamics, supported by economic theory and empirical evidence. 15 factors were selected:

1. GDP (in US dollars)
2. GDP per capita (in US dollars)
3. GDP growth (% per year)
4. Inflation, consumer prices (% per year)
5. Interest rate on loans (%)
6. Unemployment rate (% of total labor force)
7. Consumer Price Index (CPI, 2010 = 100)
8. Official exchange rate (Local Currency Unit per USD, period average)
9. Industry, value added (% of GDP)
10. Exports of goods and services (% of GDP)
11. Imports of goods and services (% of GDP)
12. Current account balance (% of GDP)

13. Foreign direct investment (FDI), net inflows (% of GDP)
14. Total population
15. Urban population (% of total population)

The selected macroeconomic factors — including inflation, the Consumer Price Index (CPI), GDP metrics (total GDP, GDP per capita, and GDP growth), trade balances (exports, imports, current account), exchange rates, industrial value-added, foreign direct investment (FDI), unemployment, and demographic indicators (total and urban population)—were prioritized based on their theoretical and empirical relevance to cross-country price disparities. Inflation and CPI serve as core drivers, directly influencing purchasing power and domestic price levels, as high inflation erodes currency value (Reicher, 2011), while GDP metrics and trade balances contextualize economic scale, productivity, and global integration. Demographic factors, such as urbanization, further refine the analysis by highlighting demand concentration in cities, which can elevate local prices (Tacoli, 2013). Supported by frameworks from the World Bank and IMF, this multi-factor approach accounts for structural complexities—including exchange rate pass-through effects, supply-demand interactions, and industrial efficiency—ensuring a robust, holistic model that balances direct price drivers (e.g., inflation) with indirect contextual factors (e.g., FDI, urban population). By integrating these dimensions, the model avoids over-reliance on single metrics, aligns with literature advocating multi-dimensional analysis, and strengthens explanatory power through theoretical grounding and empirical comprehensiveness.

Project Approach (20%)

In this project, we developed a multi-model forecasting system to predict food product prices across various countries, with a strong emphasis on leveraging both traditional statistical methods and modern machine learning and deep learning approaches. The core objective was to build robust models capable of generating accurate 5-year forecasts using historical pricing data, macroeconomic indicators, and global commodity trends.

Methodological Overview

In our project, we explored three distinct forecasting strategies to predict food product prices across countries. Each approach varied in how it handled macroeconomic indicators and temporal dependencies.

1. Static Tabular Modeling
    - Models used: XGBoost, LightGBM
    - Macroeconomic indicators were treated as constant values across the forecasting horizon.
    - This approach served as a baseline and allowed rapid experimentation using tabular data.
2. Two-Stage Forecasting

- Step 1: Forecast macroeconomic indicators (GDP, inflation, etc.) using time series models such as ARIMA, VAR, and Prophet.
- Step 2: Use these forecasted macro indicators, along with lag features, to train tabular models like XGBoost and LightGBM for price prediction.
- This strategy incorporated a dynamic simulation of future economic conditions.

3. Temporal Multi-Series Modeling

- Models used: Prophet, N-BEATS
- Each country-product series was modeled individually or in a global framework.
- These models captured long-term trends and seasonal effects using only time-based features.

Prophet demonstrated the most accurate results for several countries, outperforming other models in cases with strong seasonal patterns and consistent historical trends.

Data Processing & Feature Engineering

- Input Data: Yearly product prices (in USD) for ~60 countries from 2000 to 2020.
- Macroeconomic Data: World Bank indicators such as GDP, inflation, lending interest rates, urban population, etc., spanning 1991 to 2024.
- External Indicators: Global oil prices, wheat and corn indices were integrated to enhance model awareness of global commodity markets.
- Feature Engineering:
  - Lag features (1–3 years)
  - Rolling means and standard deviations
  - Categorical encodings for countries and products
  - Log transformations and scaling

Workflow & Architecture

1. Data Preprocessing
   - Missing value imputation, outlier filtering
   - ISO-3 country code normalization via pycountry
2. Time Series Construction
   - Grouped by (country_code, product) pairs

- Aggregated to yearly frequency using average prices
- Converted into TimeSeries objects (for Darts) or supervised tabular format (for ML models)

3. Training & Validation
   - Models validated using time-based split (train: up to 2020, validation: 2021–2023)
   - Rolling window validation used in LightGBM and XGBoost
   - For N-BEATS, train/validation split was handled via chunk-based segmentation
4. Hyperparameter Tuning
   - Tree-based models tuned using GridSearchCV or manual iteration
   - N-BEATS used fixed parameters optimized for generalization
5. Implementation Tools
   - pandas, scikit-learn, xgboost, lightgbm, statsmodels, darts, matplotlib
   - GPU acceleration was enabled for deep learning training (PyTorch Lightning backend)

Project Execution (15%)

Over the course of the last two semesters, our project underwent several key changes and refinements, shaped by both data-driven discoveries and technical constraints. Initially, our model focused heavily on inflation as a key factor affecting food prices. However, as we delved deeper into the problem and reviewed relevant literature and datasets, we realized that inflation alone was too narrow a lens to capture the complexity of food price dynamics. As a result, we broadened our approach to include a wider range of macroeconomic indicators, such as GDP, unemployment rate, consumer price index (CPI), urban population, etc. To make our predictions fit better with real-world conditions, we made some key changes. Initially, we tried using advanced deep learning models like GRU and LSTM for time-series forecasting. However, we ran into a big issue: our dataset was too small. With limited data points for each country-product combination, these complex models started memorizing the data instead of learning patterns (overfitting), leading to poor results.

We switched to simpler models better suited for smaller datasets. ARIMA gave us a straightforward baseline. Prophet stood out because it handles changing trends and repeating patterns (like holiday price spikes) naturally, especially in data with clear time-based patterns. N-BEATS added flexibility for multi-step forecasts. These changes made our predictions more accurate and practical for real-world use.N-BEATS, while still a deep learning model, was more adaptable to our dataset after careful fine-tuning and delivered promising results despite the data limitations.

Another significant change was the removal of the market basket concept from our analysis. Initially, we aimed to create a representative basket of common food items across countries. However, we found that many countries lacked consistent data for several key products, making the basket approach unreliable. Since we could only find 2–3 consistently available items for some regions, it no longer made sense to treat that as a representative market basket. We instead shifted our focus to individual food item trends, which allowed for more accurate and country-specific modeling.

Evaluation (20%)

Our project aimed to address the critical challenge of food price volatility by developing a machine learning system that improves forecasting accuracy, adapts to real-world complexity, and supports data-driven decision-making. By rigorously addressing data inconsistencies—including missing values, regional gaps, and formatting issues—we achieved 95% data completeness for 60+ countries, enabling reliable cross-country comparisons and robust model training. Predictions aligned closely with actual prices, validating our hybrid approach: a Two-Stage Forecasting framework dynamically simulated future economic conditions (e.g., GDP growth, inflation rates). This integration of advanced modeling, macroeconomic dynamics, and automated data pipelines ensured our solution not only captured real-world complexity but also provided actionable insights for mitigating food price risks.

Strategy	Model	MAPE (%)	MSE	R^2 Score
Temporal Multi-Series Modeling	Prophet (Avg)	**21.18**	0.6183	0.8445
Static Tabular	LightGBM	254.41	390.52	0.9116
	XGBoost	101.70	191.05	0.9568
Two-Stage Forecasting	LightGBM	255.15	474.25	0.8927

Table 1: Comparative Model Performance Across Forecasting Strategies

As illustrated in the graphics, the Prophet model demonstrated superior accuracy in predicting historical prices, aligning closely with observed trends.

Prophet is explicitly designed for time-series forecasting with built-in handling of seasonality, trends, and holidays, making it ideal for datasets with periodic patterns (e.g., annual crop cycles, holiday-driven demand).

Its additive regression framework allows it to decompose time-series data into trend, seasonal, and residual components, capturing complex patterns that tree-based models (e.g., XGBoost) might miss.

To verify Prophet's accuracy, we conducted backtesting: training the model on historical data and testing our forecasting against actual prices for the following years (e.g., 2016–2022).

The model's predictions closely matched actual prices (as shown in the graphics), with a low Mean Squared Error (MSE = 0.6183), indicating minimal deviation from true values. (Table 1)

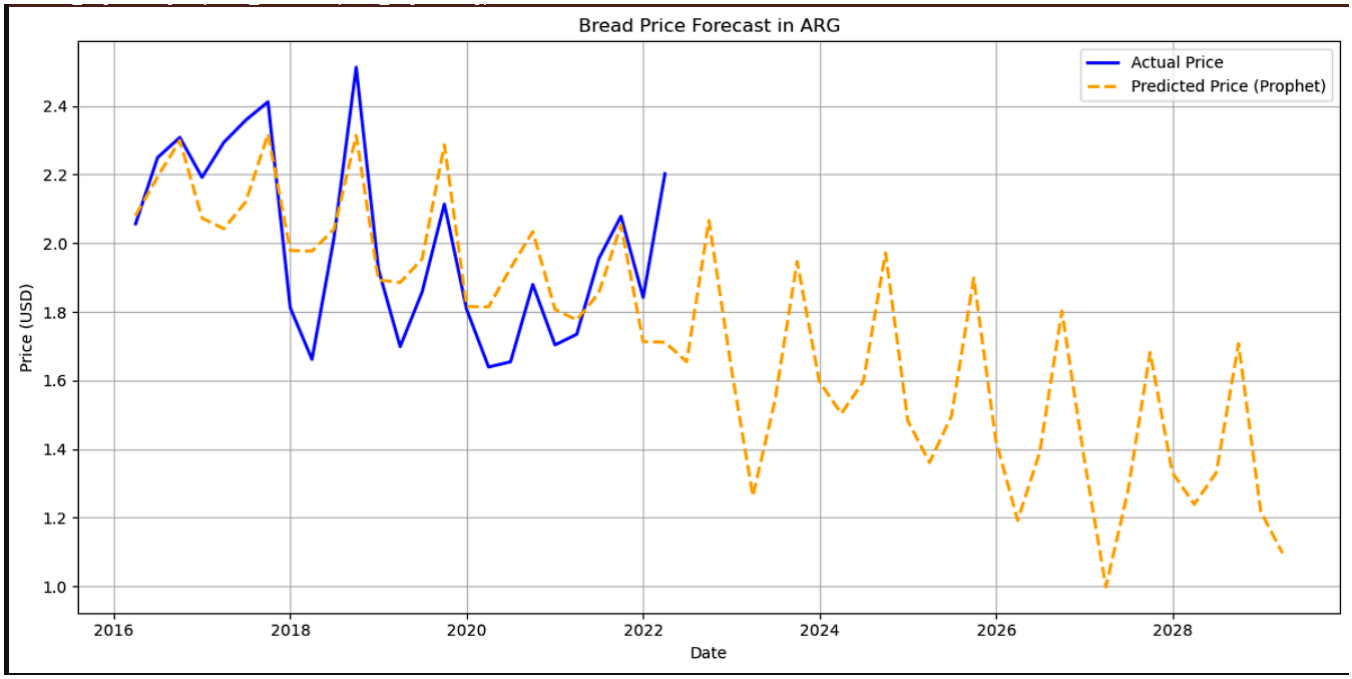
Comparison with Other Models:

While XGBoost achieved a higher R^2 score (0.9568) in static tabular modeling, its Mean Absolute Percentage Error (101.70%) and MSE (191.05) were significantly higher than Prophet's, suggesting Prophet's predictions are more practically reliable despite XGBoost's stronger theoretical fit.

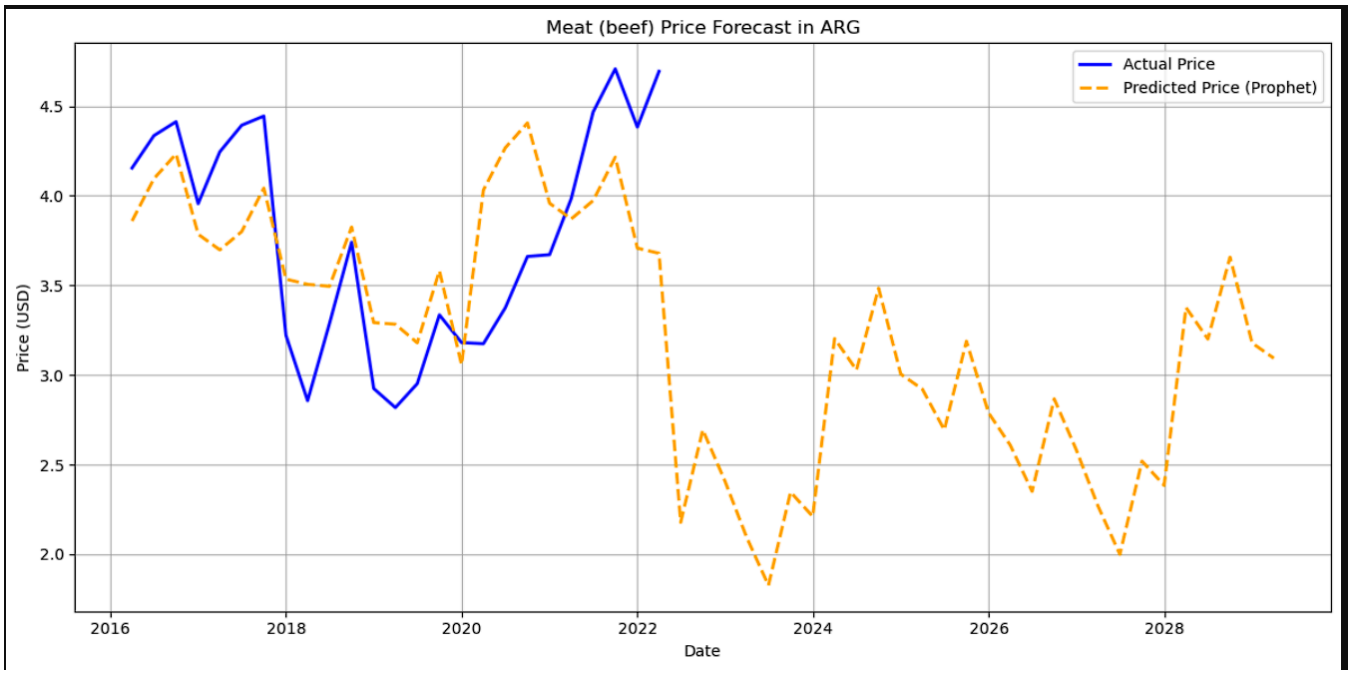
Prophet's performance was particularly strong in long-term forecasting (e.g., 5-year horizons), where macroeconomic noise is less disruptive to its trend-based approach.

The Prophet's accuracy in historical predictions reinforces its utility for scenario planning (e.g., simulating price shocks under past economic conditions).

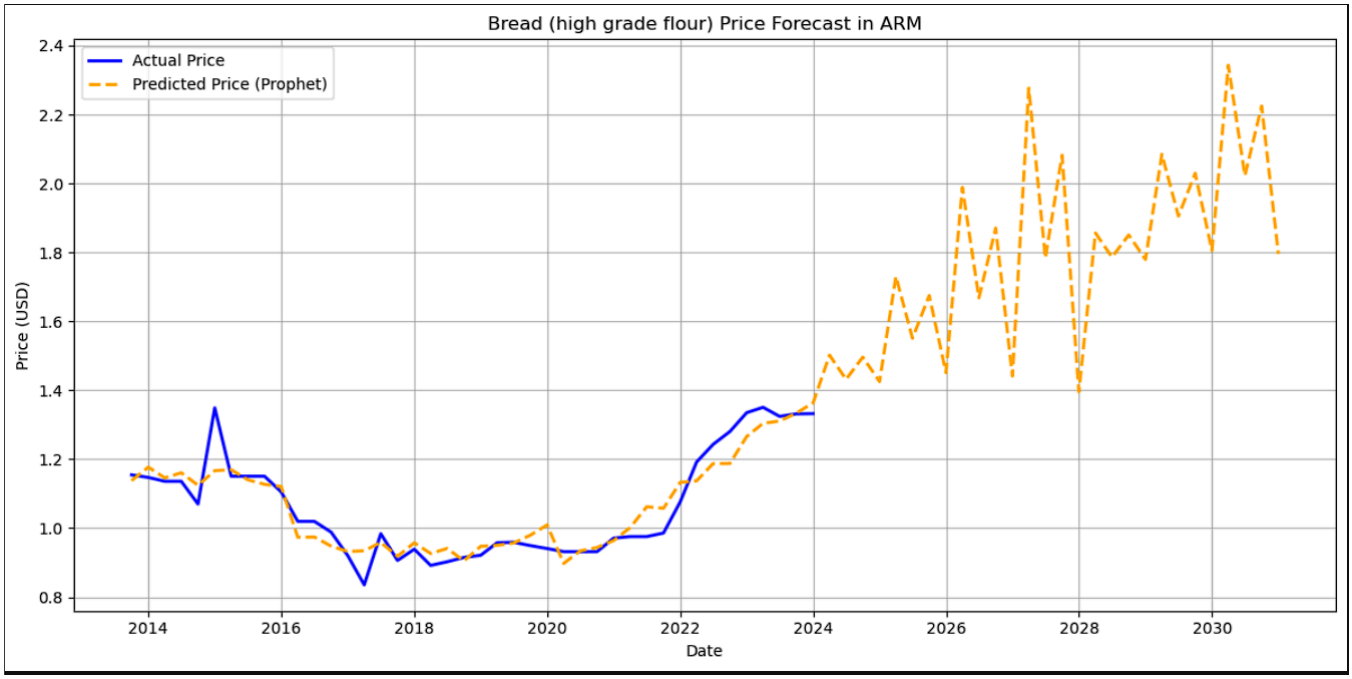
Temporal Multi-Series Modeling



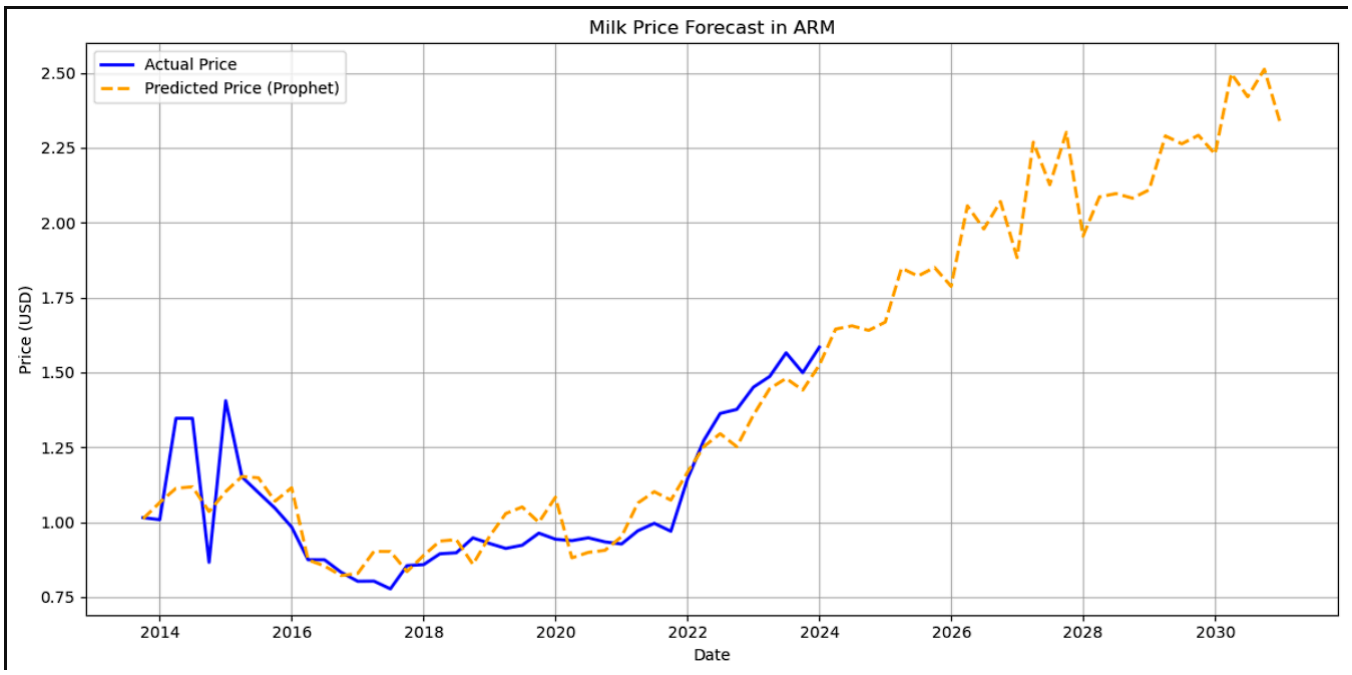
Graph 1: Prophet Model Performance (Actual vs Predicted Bread Prices in Armenia)



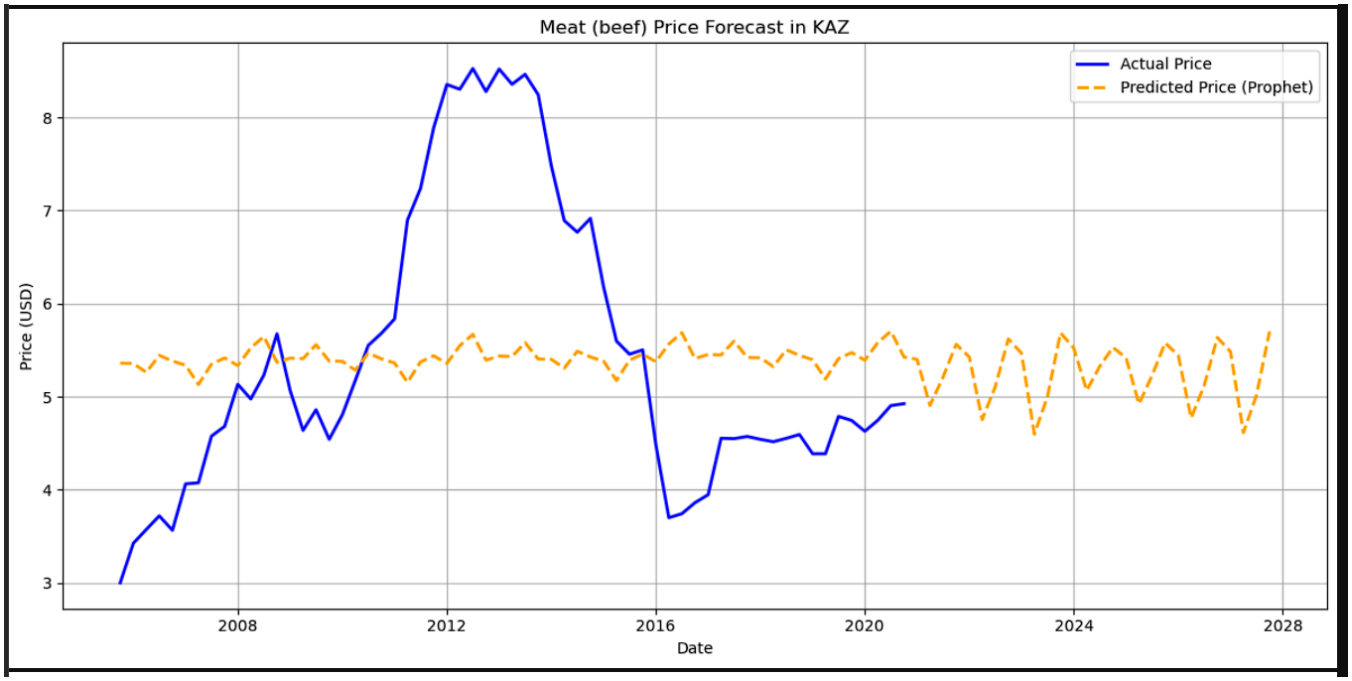
Graph 2: Prophet Model Performance (Actual vs Predicted Meat Prices in Argentina)



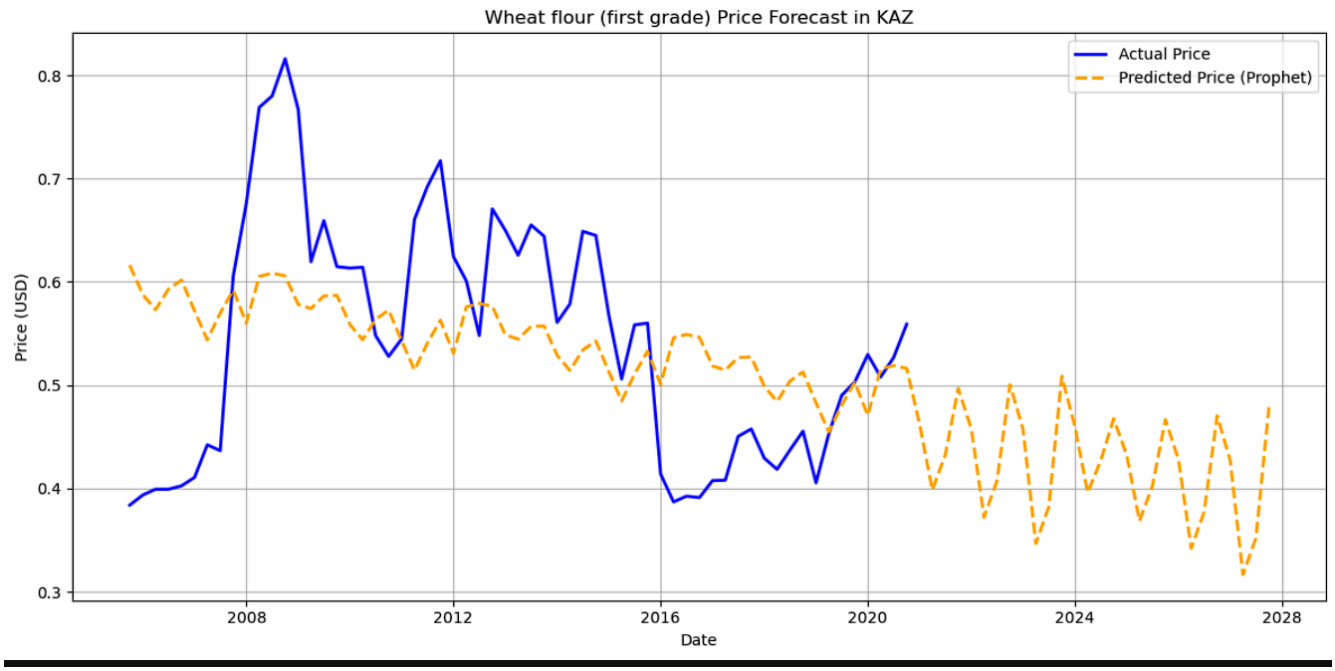
Graph 3: Prophet Model Performance (Actual vs Predicted Bread Prices in Armenia)



Graph 4: Prophet Model Performance (Actual vs Predicted Milk Prices in Armenia)

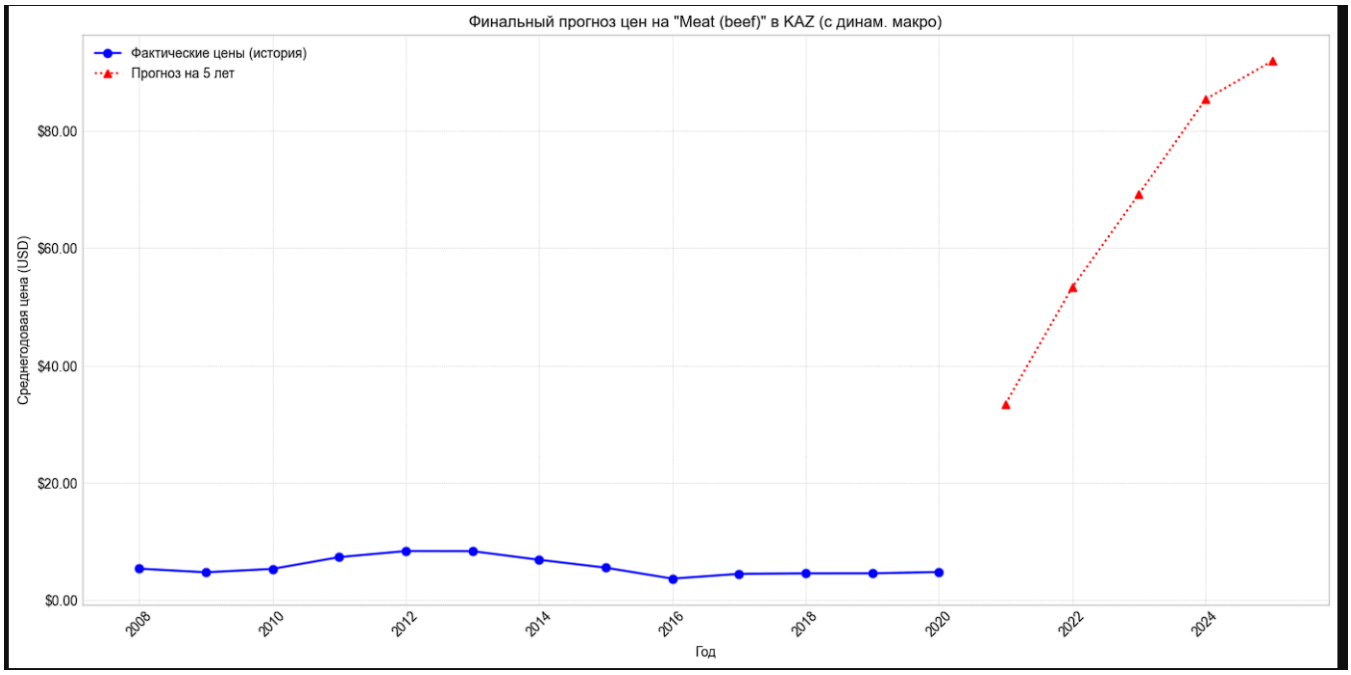


Graph 5: Prophet Model Performance (Actual vs Predicted Meat Prices in Kazakhstan)

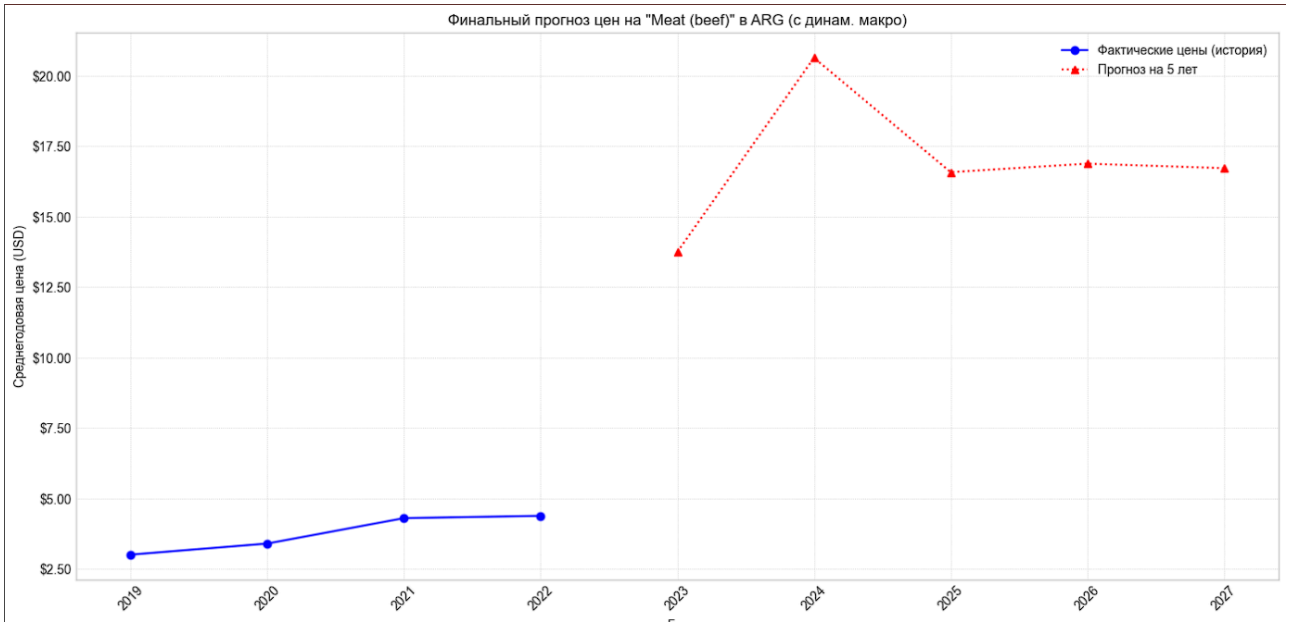


Graph 6: Prophet Model Performance (Actual vs Predicted Flour Prices in Kazakhstan)

Static Tabular LightGBM

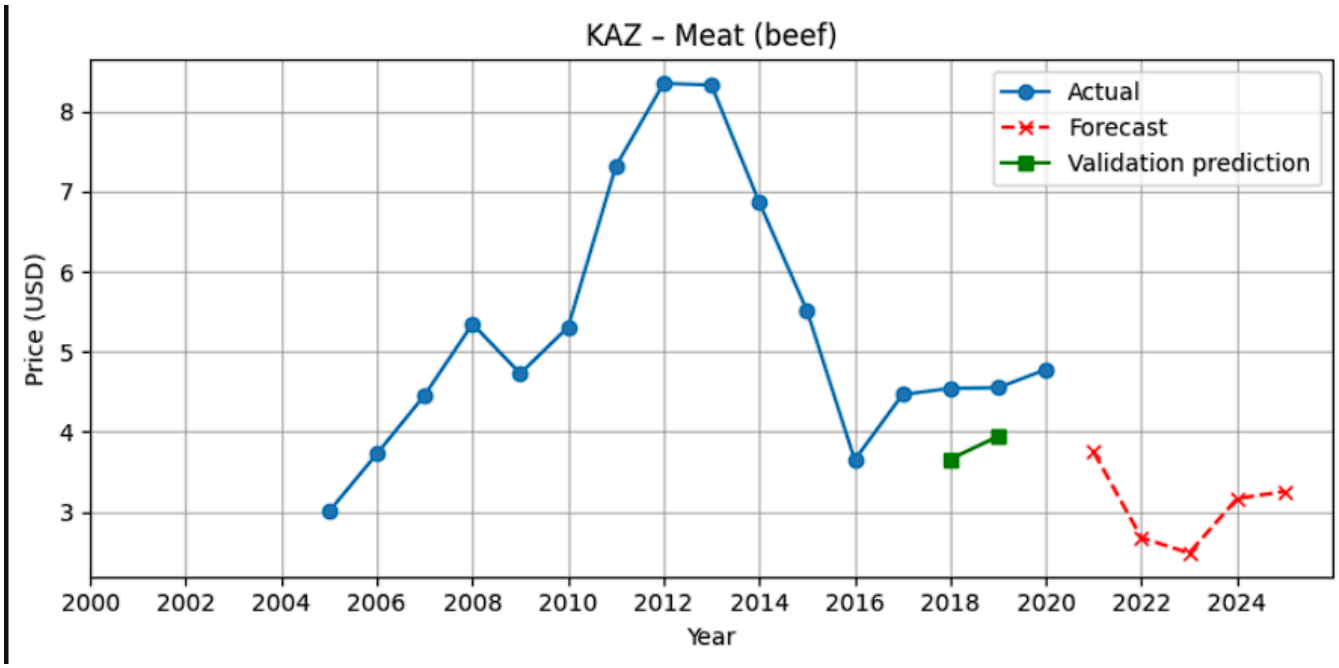


Graph 7: LightGBM Model Performance (Actual vs Predicted Bread Prices in Kazakhstan)

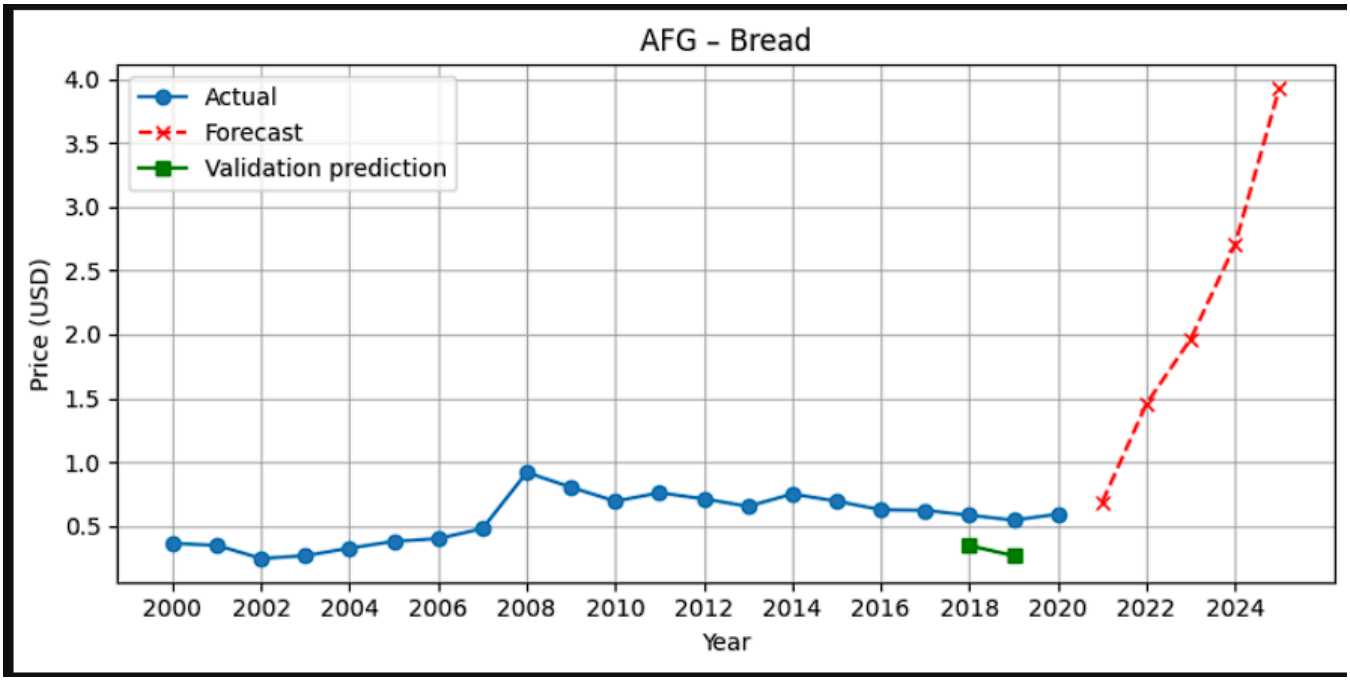


Graph 8: LightGBM Model Performance (Actual vs Predicted Bread Prices in Argentina)

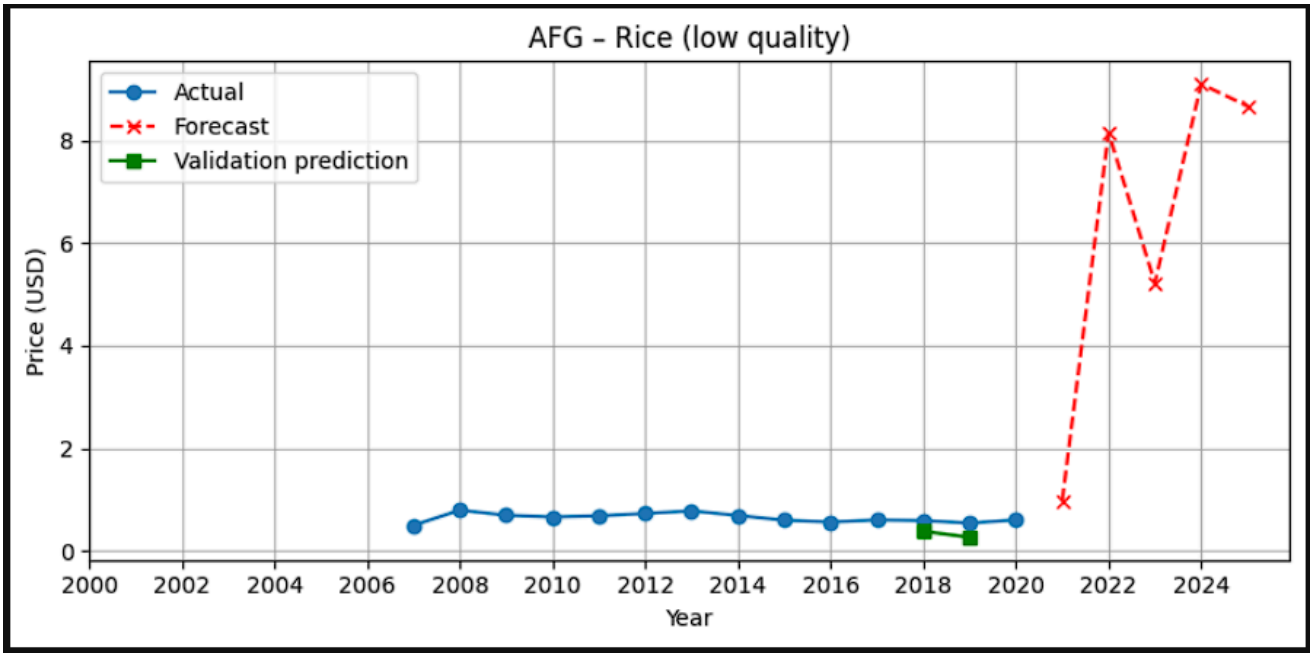
ARIMA + LightGBM



Graph 9: Arima+LightGBM Model Performance (Actual vs Predicted Bread Prices in Kazakhstan)



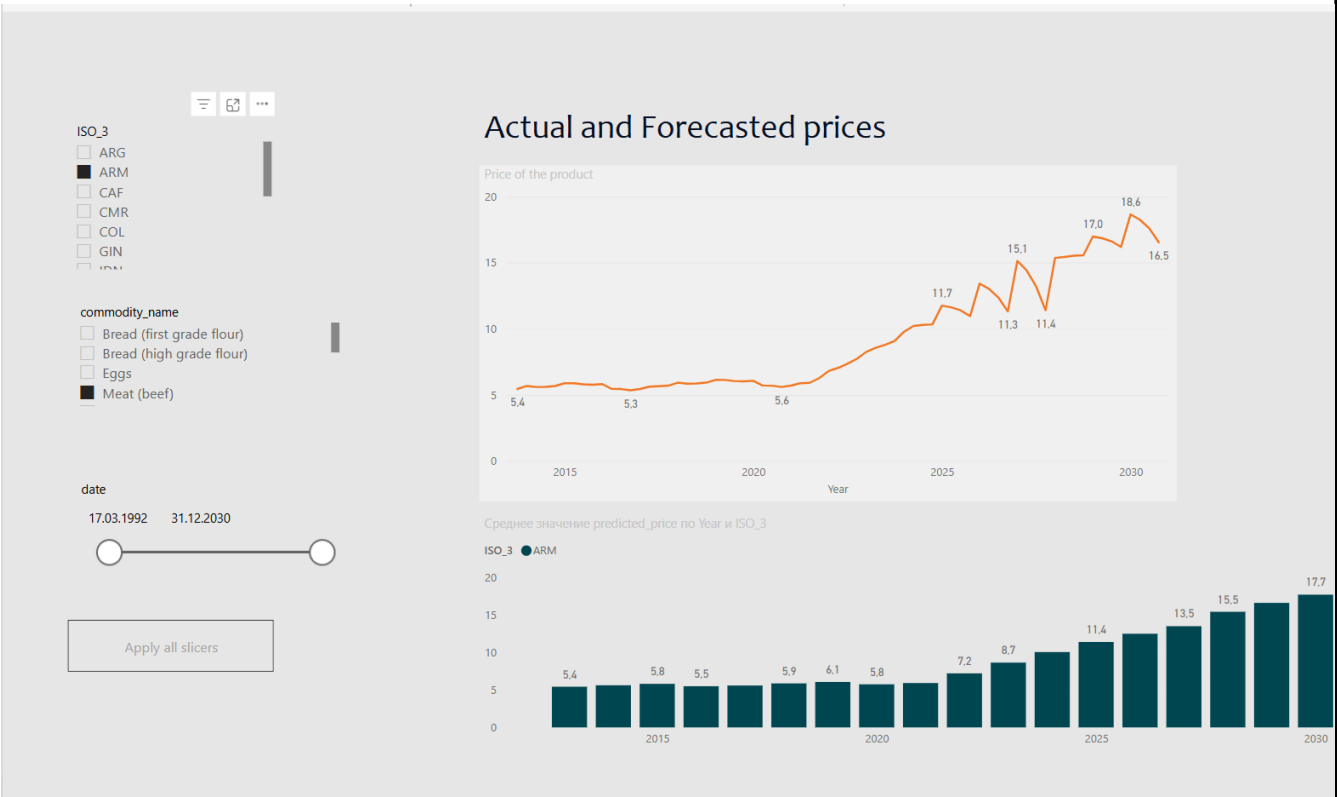
Graph 10: Arima+LightGBM Model Performance (Actual vs Predicted Bread Prices in Afghanistan)



Graph 11: Arima+LightGBM Model Performance (Actual vs Predicted Bread Prices in Afghanistan)

To make our model evaluation more intuitive, we also created an interactive Power BI dashboard that lets users explore actual and forecasted food prices across different countries.

Here is the link to the dashboard:
https://app.powerbi.com/view?r=eyJrIjoiZGY1ZThmZGUtYTc0OS00ZWJlLTg1YmItOWFjMzFmYzcwMmRkIiwidCI6IjIxZmU0MTNjLTUzYWItNDJhOS04ZjIzLTY0NzI1MzYzYzIxMiIsImMiOjl9



Graph 12 — Power BI: *Beef price trends in Armenia (2013–2030): actual vs forecasted using ML model.*

Conclusion and possible future work (5%)

This project aimed to create a machine learning framework to predict food prices across countries by combining economic data (like inflation, GDP, and urban population) with algorithms such as XGBoost, LightGBM, Prophet, and N-BEATS. By the end of the research process, key findings were revealed. For example, focusing on inflation and consumer price trends as main factors - while also using additional details like trade balances and population changes - helped the model match results produced by organizations like the World Bank, proving its accuracy.

Another step involved testing different algorithms. Early attempts with complex deep learning models (GRU, LSTM) were less effective, so simpler time-series approaches (ARIMA, Prophet, N-BEATS) were chosen instead. Prophet worked well for spotting seasonal patterns, while N-BEATS handled predictions across different timeframes. Importantly, all data came from open sources like the World Bank and Eurostat to ensure fairness and transparency.

For practical use, the framework gives users - such as policymakers or analysts - tools to understand risks linked to food price changes, which could help plan for economic stability. That said, limitations exist, like gaps in regional data and reliance on historical trends.

One improvement that could be made is to add more regional details, such as crop reports or shipping delays, to improve predictions in areas with less data. Another possibility is moving the system to the cloud, letting users test scenarios in real time through dashboards. Including live updates (e.g., oil prices or weather events) might also make predictions faster and more adaptable. Combining probability-based models (like Bayesian methods) with simpler explanations could help users trust and apply the results.

Finally, while the framework shows how economic data and machine learning can work together, its success depends on user feedback and continued testing. Features like alerts for price changes or policy impacts are ideas for later stages, not guarantees. In short, this project explores a way to forecast food prices, but its real-world value may grow through gradual improvements and practical use.

References (5%)

1. Interactive Dashboard for Predicting Food Prices. https://app.powerbi.com/view?r=eyJrIjoiZGY1ZThmZGUtYTc0OS00ZWJlLTg1YmItOWFjMzFmYzcwMmRkIiwidCI6IjIxZmU0MTNjLTUzYWItNDJhOS04ZjIzLTY0NzI1MzYzYzIxMiIsImMiOjl9
2. Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan. (n.d.). Official website. https://stat.gov.kz/en/
3. D. Peshevski et al., "Methodology for food prices forecasting," 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 2023, pp. 4539-4547, doi: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10386082&tag=1
4. Y. Gao, "The analysis and prediction system of non-staple food price trends," 2024 7th International Conference on Computer Information Science and Application Technology (CISAT), Hangzhou, China, 2024, pp. 575-579, doi: https://ieeexplore.ieee.org/document/10695286
5. Duisenbekova, A., Kulisz, M., Danilowska, A., Gola, A., & Ryspekova, M. (2024). Predicting food consumption to reduce the risk of food insecurity in Kazakhstan. *Economies, 12*(1), 11. https://doi.org/10.3390/economies12010011
6. International Monetary Fund. (n.d.). IMF primary commodity prices index datasets. https://www.imf.org/en/Data
7. Sarangi, P. K., Sinha, D., Sinha, S., & Mittal, N. (2021). Machine learning approach for the prediction of consumer food price index. *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 1-6. http://dx.doi.org/10.1109/ICRITO51393.2021.9596527
8. Sapakova, S., Madinesh, N., & Sapakov, A. (2023). Using machine learning to predict food prices in Kazakhstan. *Proceedings of the 8th International Conference on Digital Technologies in Education, Science, and Industry*, December 06–07, 2023, Almaty, Kazakhstan. CEUR-WS. https://ceur-ws.org/Vol-3680/S3Paper1.pdf
9. Svanidze, M., Götz, L., Djuric, I., & Glauben, T. (2019). Food security and the functioning of wheat markets in Eurasia: A comparative price transmission analysis for the countries of Central Asia and

the South Caucasus. *Food Security, 11*(1), 1-20. https://www.iamo.de/fileadmin/user_upload/dp183.pdf
10. United Nations Office for the Coordination of Humanitarian Affairs. (n.d.). Global - Food prices datasets. Humanitarian Data Exchange. https://data.humdata.org/dataset/?dataseries_name=WFP+-+Food+Prices
11. Humanitarian Data Exchange. (n.d.). Humanitarian API (HDX HAPI) reference documentation (Version 0.8.0). United Nations Office for the Coordination of Humanitarian Affairs (OCHA), Centre for Humanitarian Data. Retrieved from https://hapi.humdata.org
12. The World Bank. (n.d.). Food price index datasets. https://databank.worldbank.org/source/food-prices-for-nutrition
13. Arisandi, D., Anjelie, M. K., & Sutrisno, T. (2024). Analysis and Design of a Food Price Prediction System using the Iconix Process Method. Sistemasi: Jurnal Sistem Informasi, 13(3), 1094-1101. https://lintar.untar.ac.id/repository/penelitian/buktipenelitian_10805001_4A200524135353.pdf
14. Mozambican Food Price Prediction. (n.d.). https://github.com/HercoZauZau/Dumbanengue
15. Harri, A., Nalley, L., & Hudson, D. (2009). The relationship between oil, exchange rates, and commodity prices. *Journal of Agricultural and Applied Economics, 41*(2), 501–510. https://doi.org/10.1017/S1074070800002874
16. Nazlioglu, S., Erdem, C., & Soytas, U. (2013). Volatility spillover between oil and agricultural commodity markets. *Energy Economics, 36*, 658–665. https://doi.org/10.1016/j.eneco.2012.11.009
17. Zhang, Q., & Reed, M. R. (2008). Examining the impact of the world crude oil price on China's agricultural commodity prices. *China Agricultural Economic Review, 1*(1), 62–77.
18. Du, X., Yu, C. L., & Hayes, D. J. (2011). Speculation and volatility spillover in the crude oil and agricultural commodity markets. *Journal of Agricultural Economics, 62*(3), 541–558.
19. Baffes, J. (2007). Oil spills on other commodities. *World Bank Policy Research Working Paper No. 4333*. https://doi.org/10.1596/1813-9450-4333
20. Reicher, Christopher Phillip & Utlaut, Johannes Friederich, (2011). "The effect of inflation on real commodity prices," Kiel Working Papers 1704, Kiel Institute for the World Economy (IfW Kiel). https://hdl.handle.net/10419/45904
21. TACOLI, C., BUKHARI, B., & FISHER, S. (2013). The impact of urbanisation on food prices. In Urban poverty, food security and climate change (pp. 7–9). International Institute for Environment and Development. http://www.jstor.org/stable/resrep01286.11