Time Series Forecasting Report: Corporación Favorita Grocery Sales

1. Introduction This project focuses on forecasting grocery sales for Corporación Favorita, a retail company based in Ecuador. The dataset was obtained from Kaggle and includes transactional sales data across multiple stores and product families. To make the task more focused and computationally feasible, the data was filtered to include only the Guayas region and items from the top 3 most-sold product families, with a random sample of 2 million rows taken for analysis.

2. Data Preparation and Feature Engineering The data underwent a thorough cleaning and preparation process:

- **Handling missing values** and filling gaps in sales dates.
- **Removing outliers** and correcting negative sales entries.
- **Extracting time-based features** including year, month, day of week, and weekend indicators.
- **Creating lag features** (e.g., sales 1 and 7 days ago) and **rolling statistics** (7-day moving averages and standard deviations).

These steps enhanced the dataset's temporal structure, enabling models to learn from past trends and seasonality.

3. Baseline and Advanced Models Three models were trained to compare forecasting performance:

- **Naive Model**: Served as a baseline using the assumption that the next day's sales are the same as the previous day's.
- **SARIMAX**: A statistical model that captures autoregression, integration, moving averages, and seasonality.
- **XGBoost**: A tree-based machine learning model trained on the engineered features.

4. Tools and Deployment

- A **Streamlit application** was developed to allow interactive, user-friendly forecasting.
- An **MLflow platform** was set up for experiment tracking, model versioning, and reproducibility.
- The final XGBoost model was saved for reuse and deployment.

5. Evaluation Results The models were evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2) metrics:

Model	MSE	MAE	R^2 Score
Naive	7.14	0.22	-0.0067
SARIMAX	1,378,970	905.28	-0.1829
XGBoost	6.95	0.40	0.0203

6. Interpretation and Discussion

- The **Naive model**, despite its simplicity, performs surprisingly well on this sample. Its low MAE indicates strong short-term tracking, though its R^2 is close to zero.
- The **SARIMAX model** struggles due to its assumption of linear relationships and the noisy, complex structure of retail data. Its high error and negative R^2 indicate a poor fit.

- The **XGBoost model** marginally outperforms the Naive baseline in terms of MSE and R^2, though its MAE is higher. This suggests it captures broader trends but may not track fine-grained changes as closely.

7. Recommendations for Improvement

- Introduce additional external features (e.g., oil prices, holidays, promotions).
- Fine-tune model hyperparameters using cross-validation.
- Segment models per store or item to capture localized trends.
- Test deep learning models such as LSTMs for more complex pattern recognition.

8. Conclusion This project demonstrates a full-cycle time series forecasting pipeline: from data cleaning and feature engineering, through modeling, evaluation, and deployment. While SARIMAX underperformed, the XGBoost model showed promise and could be improved further. The inclusion of Streamlit and MLflow ensures that the solution is both interactive and reproducible, laying the groundwork for a robust production forecasting system.