# Data Science Project Report: Retail Demand Forecasting for Corporación Favorita

## 1. Business Problem Overview

Corporación Favorita, a leading Ecuadorian grocery retailer, faces challenges in accurately predicting product demand across its multiple store locations. The ability to forecast demand effectively is crucial for:

- **Optimizing inventory management** to prevent stockouts and reduce waste.
- **Improving supply chain efficiency** by ensuring the right products are available at the right time.
- **Enhancing promotional strategies** to maximize sales and customer satisfaction.
- **Reducing financial losses** due to overstocking or missed sales opportunities.

The company currently relies on **subjective forecasting methods** with minimal data utilization and lacks automation in its decision-making process. This project aims to leverage **machine learning techniques** to develop a robust, data-driven demand forecasting model.

### Data Source

The dataset used for this project is derived from the publicly available **"Corporación Favorita Grocery Sales Forecasting"** dataset on Kaggle (source).

## 2. Data Extraction, Filtering, and Exploratory Analysis

### Data Filtering Strategy

To ensure a focused and efficient analysis, we selected:

- **Region:** Stores in the state of **Guayas**.
- **Products:** Two representative items (**106716, 1158720**).
- **Date Range:** Data before **April 1, 2014**.

**Filtering Process:** The dataset was processed in chunks to optimize memory usage. Data was filtered based on store locations, item numbers, and date constraints, then aggregated to provide a time-series view of unit sales per product.

## Exploratory Data Analysis (EDA)

Several key steps were conducted to understand the dataset:

**1. Seasonality & Trends**

- The **time-series decomposition plot** indicates **strong weekly and monthly seasonality**.
- **Weekend sales are significantly higher** than weekday sales, especially on Saturdays, as seen in the sales trends visualization.

**2. Promotions and Demand Volatility**

- Items on **promotion** exhibit significantly **higher volatility** in sales, as demonstrated in the **promotion vs. non-promotion sales plot**.
- The impact of promotions varies across products and store locations, suggesting an interaction effect.

**3. Holiday Effects on Sales**

- Sales spikes around **national holidays**, as indicated in the **holiday impact analysis plot**, confirm that holidays strongly influence demand.
- Incorporating holiday indicators into forecasting models could improve accuracy.

**4. External Factors Correlation**

- **Oil prices** show a weak negative correlation with demand, suggesting minor economic influence.
- **Holidays and promotions** emerge as the strongest external predictors of demand fluctuations.

**5. Stationarity Test (Augmented Dickey-Fuller Test)**

- The **ADF test results** indicate that the raw time series is **non-stationary** (p-value > 0.05), meaning that it has a trend or seasonality component.
- **First-order differencing** was applied to make the series stationary, confirmed by a **post-differencing ADF test** showing a p-value < 0.05.
- This transformation is necessary for ARIMA modeling to ensure reliable forecasting.

# 3. Machine Learning Models & Performance Comparison

## Models Evaluated

The project evaluated two major approaches:

1. **ARIMA (Autoregressive Integrated Moving Average)** – A traditional time-series model.

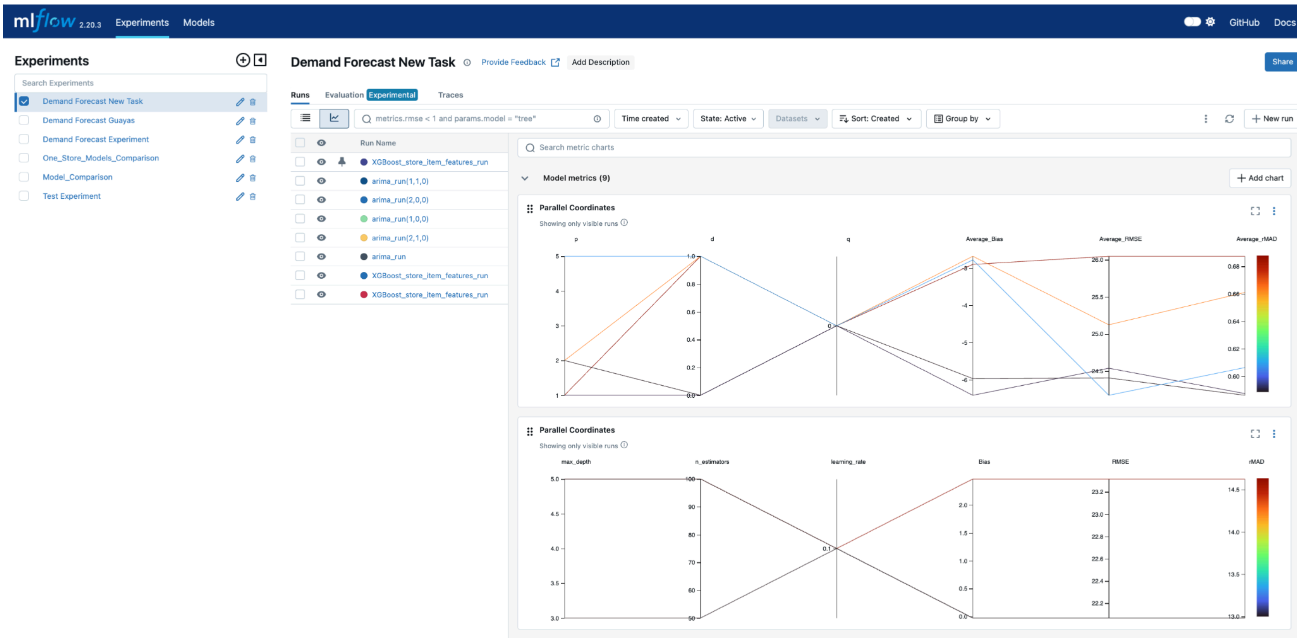
2. **XGBoost (Extreme Gradient Boosting)** – A tree-based machine learning model optimized for time-series forecasting.

## Performance Metrics Used

- **Root Mean Squared Error (RMSE)**: Measures the absolute forecast accuracy.
- **Relative Mean Absolute Deviation (rMAD)**: Evaluates forecast error relative to demand.
- **Bias**: Measures systematic over/under-prediction.

## Results Summary

| Model | RMSE (Lower = Better) | rMAD (Lower = Better) | Bias (Closer to 0 = Better) |
|---|---|---|---|
| **ARIMA(1,0,0)** | 24.54 | 0.587 | -6.41 |
| **ARIMA(2,0,0)** | 26.05 | 0.586 | -5.97 |
| **XGBoost (50 trees, depth=3)** | **22.07** ✅ | **12.98** ✅ | **-0.028** ✅ |




## Key Insights from Results

- **XGBoost outperforms ARIMA** in all key metrics, providing the lowest **RMSE (22.07)** and nearly zero **bias (-0.028)**.
- **ARIMA models systematically under-predict sales** (bias = -6.41 and -5.97), making them less reliable.

# 4. Business Recommendations

## Best Model Selection Based on Business Needs

| Business Task | Best Metric to Optimize | Best Model |
|---|---|---|
| **Daily/Weekly Demand Forecasting** | RMSE (accuracy) | ✅ **XGBoost** |
| **Inventory & Supply Chain Planning** | RMSE + Bias (avoid stock issues) | ✅ **XGBoost** |
| **Promotional Impact Forecasting** | rMAD (relative accuracy) | ✅ **XGBoost** |
| **Seasonal Demand Forecasting** | rMAD (proportional shifts) | ✅ **XGBoost** |
| **Financial & Budget Planning** | Bias (avoid systematic errors) | ✅ **XGBoost** |
| **Long-Term Strategic Forecasting** | Bias (prevent long-term miscalculations) | ✅ **XGBoost** |

## Implementation Strategy

1. **Deploy XGBoost for demand forecasting** across all stores.
2. **Integrate automated forecasting into inventory and procurement systems** to optimize stock management.
3. **Develop a monitoring dashboard** to track model performance and fine-tune forecasts.
4. **Consider hybrid models (ARIMA + XGBoost)** for special cases with limited data.

# 5. Conclusion

This project successfully demonstrated that **XGBoost is the best forecasting model** for Corporación Favorita's grocery demand prediction. By implementing this model, the company can:

- Reduce forecasting errors and minimize stock issues.
- Improve supply chain efficiency.
- Optimize marketing and promotional strategies.

## Next Steps

- Fine-tune the XGBoost model with additional features (seasonality, holidays, economic factors).
- Automate data pipelines for real-time forecasting.
- Deploy the model in production with business dashboard integration.

By leveraging **machine learning**, Corporación Favorita can shift from subjective forecasting to **data-driven decision-making**, ensuring better inventory management and increased profitability.

**Prepared by**

**Svitlana Kovalivska, PhD**

**12.03.2025**