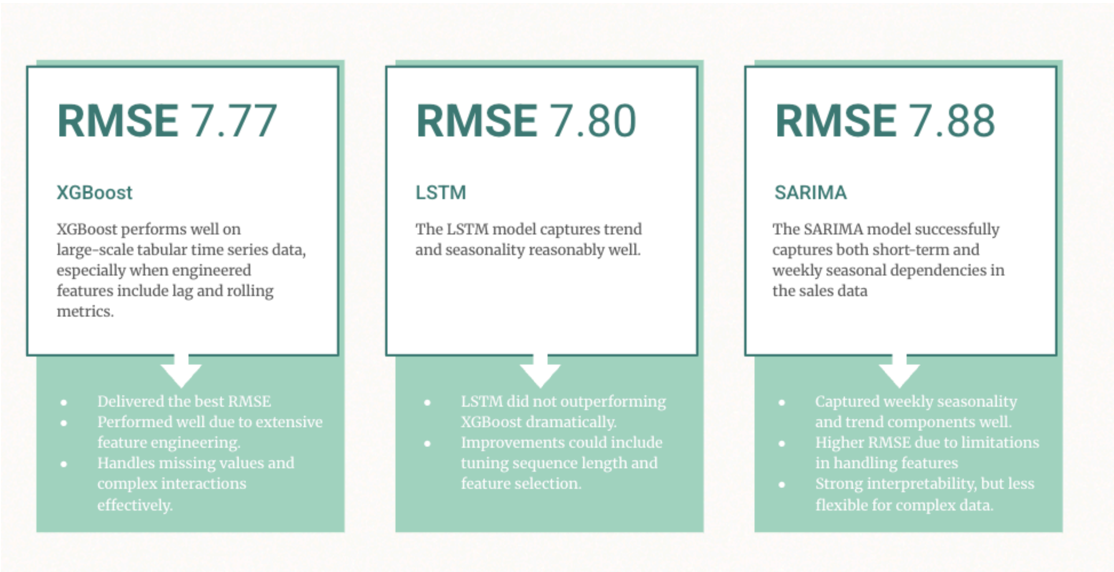
Retail Forecast Project

Executive Summary

From Checkout to Forecast: A Retail Data Deep Dive

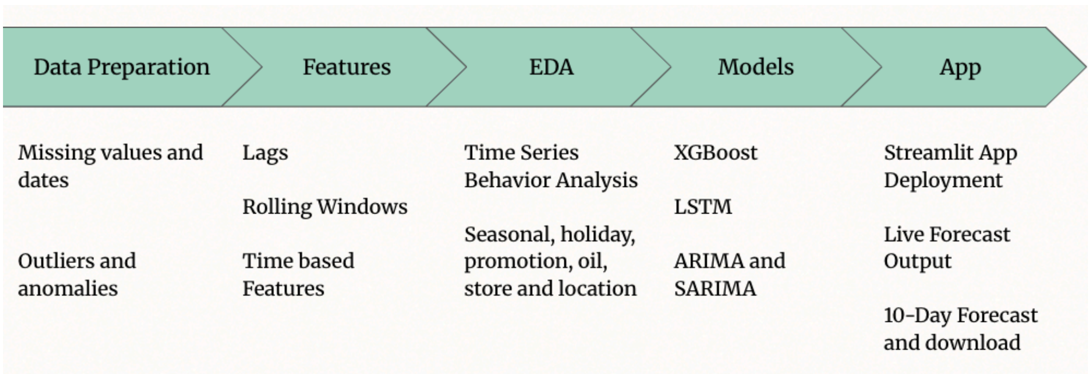
This project aimed to improve inventory and promotion planning by forecasting daily unit sales per store-item-date for a large Ecuadorian grocery retailer.

To address this challenge, we engineered time-based features, handled a large and sparse dataset with over 3.3 million rows, and compared three forecasting approaches: a tree-based machine learning model (XGBoost), a deep learning model (LSTM), and a classical statistical model (SARIMA).



XGBoost delivered the best overall performance, achieving the lowest RMSE and bias. A forecasting app was deployed using Streamlit to make predictions easily accessible. It features user input controls, a 10-day forecast output, and a download option.

This end-to-end pipeline demonstrates the power of combining advanced modeling techniques with practical tools for decision support in retail.



Retail Forecast Project

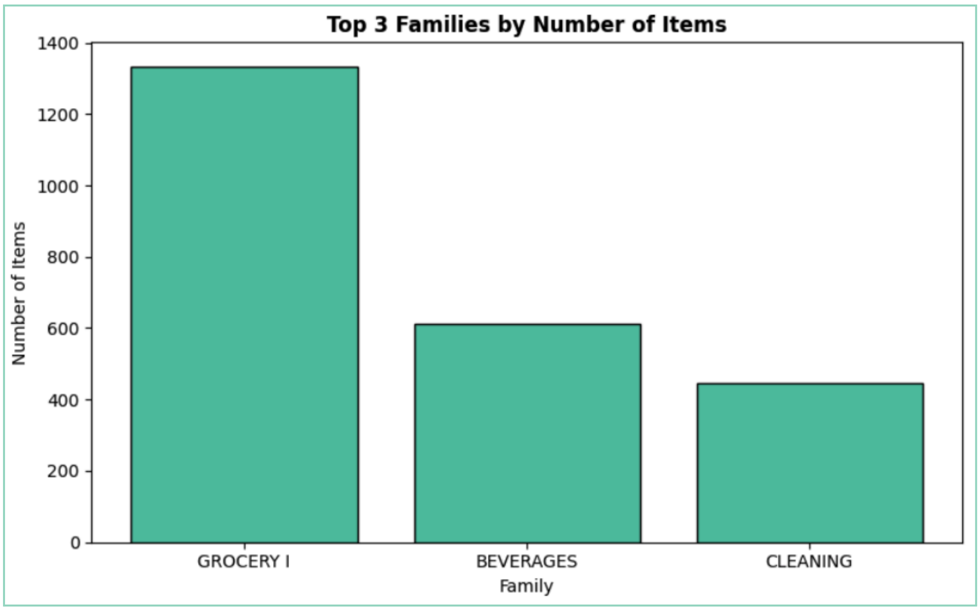
Detailed Report

Business Context

- **Client**: Ecuadorian supermarket chain
- **Goal**: Support demand planning and promotions by forecasting daily unit sales for store-item-date combinations.
- **Challenge**: Highly granular and sparse data (47% zero sales), seasonality, holiday effects, and store heterogeneity.

Data Overview

- **Dataset source**: Kaggle - Corporación Favorita
- **Data merged**: Items, stores, holidays, oil prices, transactions
- **Final dataset**: 3.39M rows × 41 columns
- **Time**: Jan 2013 – Mar 2014
- **Focus**: Guayas region and top 3 item families



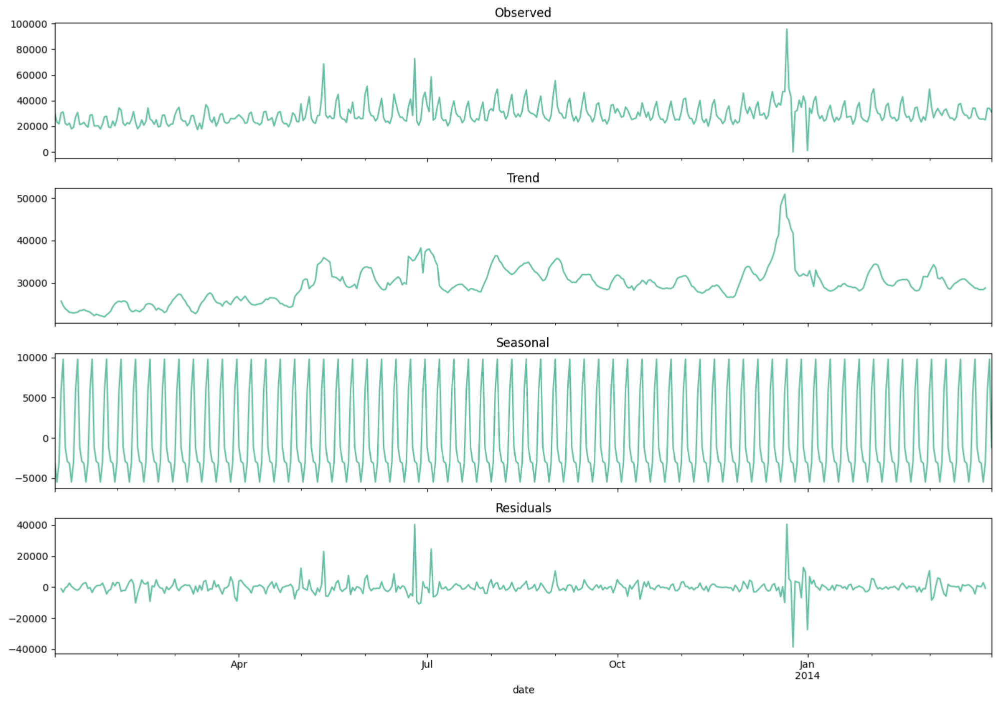
Retail Forecast Project

Preprocessing and Feature Engineering

- **Cleaning**: Removed anomalies, filled missing dates, handled nulls
- **Feature creation**:
    - **Time features**: day_of_week, is_weekend, day_of_year, month, etc.
    - **Lag features**: lag_1, lag_7, lag_14, lag_28
    - **Rolling windows**: 7d, 14d, 30d averages and std
    - **Holiday indicators**: national, local, transferred

EDA Highlights

- Time series decomposition revealed strong seasonality and trend components
- Sales peak around local holidays and promotions
- Zero-inflation and variance across stores and item types
- Top families: Grocery I, Beverages, Cleaning



Retail Forecast Project

Modeling

XGBoost (ML)

- Performed best (RMSE: 7.77, Bias: 0.173)
- Strengths: Handles sparse/tabular data, strong with lag features
- Top features: unit_sales_14d_avg, unit_sales_30d_avg, unit_sales_7d_avg, unit_sales_7d_std

LSTM (DL)

- Performed similarly (RMSE: 7.80), best MAD & SMAPE
- Captured trend/seasonality well, but more sensitive to tuning and data size

SARIMA (Statistical)

- Captured seasonal components effectively
- Performance slightly worse (RMSE: 7.89), but highly interpretable

Evaluation Metrics

Model	RMSE	MAD	SMAPE	Bias
XGBoost	7.77 ✅	3.52	120.69%	0.173 ✅
LSTM	7.80	3.49 ✅	121.55% ✅	0.195
SARIMA	7.89	3.53	122.40%	0.189

Retail Forecast Project

Forecast App

- **Built with**: Streamlit
- **Functionality**:
  - User selects store, item, and date
  - Displays forecasted sales
  - 10-day forecast with download option
  - Historical trend visualization
- **Backend**: Preprocessed data + trained XGBoost model

Key Takeaways

- Time-based engineered features significantly boosted performance
- ML models outperform classical methods on sparse tabular data
- XGBoost offers a solid tradeoff: accuracy + interpretability + speed
- LSTM potential grows with larger data + hyperparameter tuning

Recommendations

- Deploy the app for operational planning
- Automate data refresh and model retraining
- Explore advanced DL models (Transformers, N-BEATS)
- Expand to multi-step and regional forecasts