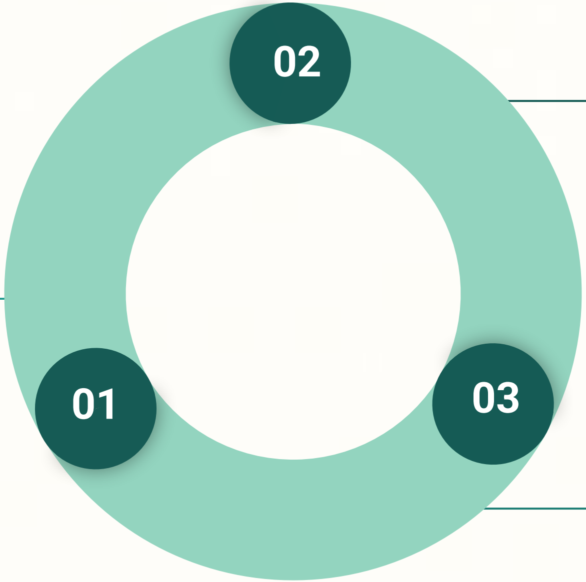
From Checkout to Forecast:

A Retail Data Deep Dive

Dido De Boodt
June 2025

Project Overview



01 Business Goal

Improve stock and promo planning by predicting sales per store-item-date

02 Dataset

Ecuadorian retailer with over 120 million rows.

03 Focus

Guayas region and top 3 item families

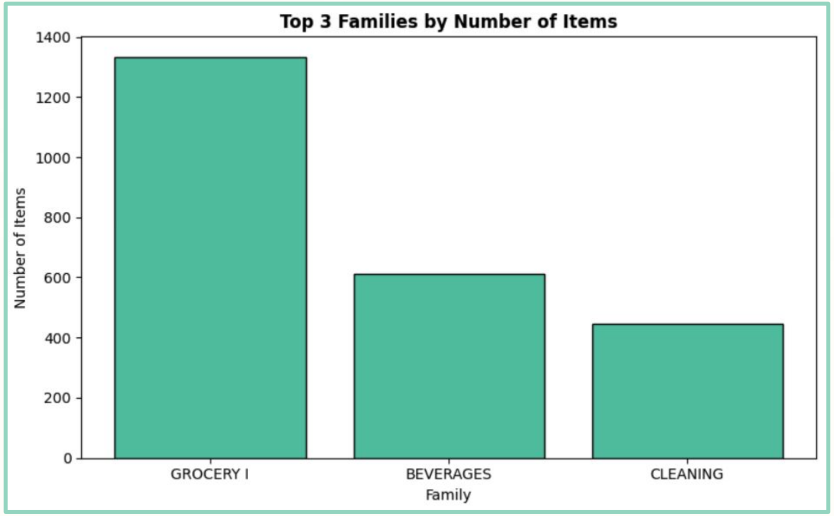
Project Steps

Data Preparation	Features	EDA	Models	App
Missing values and dates	Lags	Time Series Behavior Analysis	XGBoost	Streamlit App Deployment
	Rolling Windows		LSTM	
Outliers and anomalies	Time based Features	Seasonal, holiday, promotion, oil, store and location	ARIMA and SARIMA	Live Forecast Output
				10-Day Forecast and download

About the Data

Key Steps

- Merged: items, stores, and holidays
- Cleaned missing values & outliers
- Created lag, rolling avg, calendar features



Shape: 3.39M rows × 41 cols

Date range: Jan 2013 – Mar 2014

Stores: 10

Items: 1,127

Avg daily sales: 3.97

Zero sales: 47.25%

From Patterns to Predictions

RMSE 7.77

XGBoost

XGBoost performs well on large-scale tabular time series data, especially when engineered features include lag and rolling metrics.

- Delivered the best RMSE
- Performed well due to extensive feature engineering.
- Handles missing values and complex interactions effectively.

RMSE 7.80

LSTM

The LSTM model captures trend and seasonality reasonably well.

- LSTM did not outperforming XGBoost dramatically.
- Improvements could include tuning sequence length and feature selection.

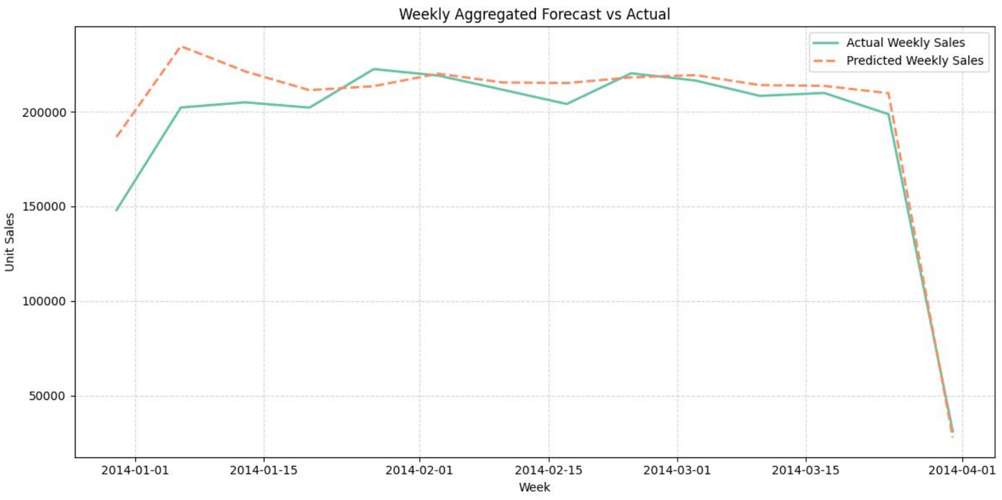
RMSE 7.88

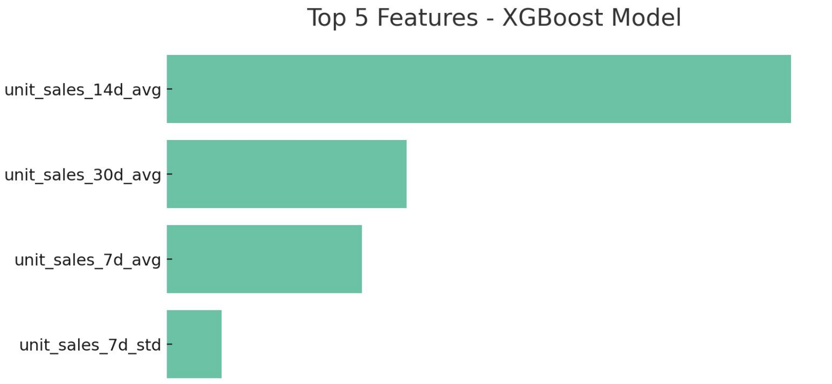
SARIMA

The SARIMA model successfully captures both short-term and weekly seasonal dependencies in the sales data

- Captured weekly seasonality and trend components well.
- Higher RMSE due to limitations in handling features
- Strong interpretability, but less flexible for complex data.

XGBoost Deep Dive





Why XGBoost Stood Out

- Lowest RMSE (7.77) and lowest bias among all models
- Effectively leveraged lag-based and rolling window features
- Adapted well to sparse & skewed data (47% zero sales)
- Weekly predictions aligned closely with true values
- Transparent model: top predictors are interpretable

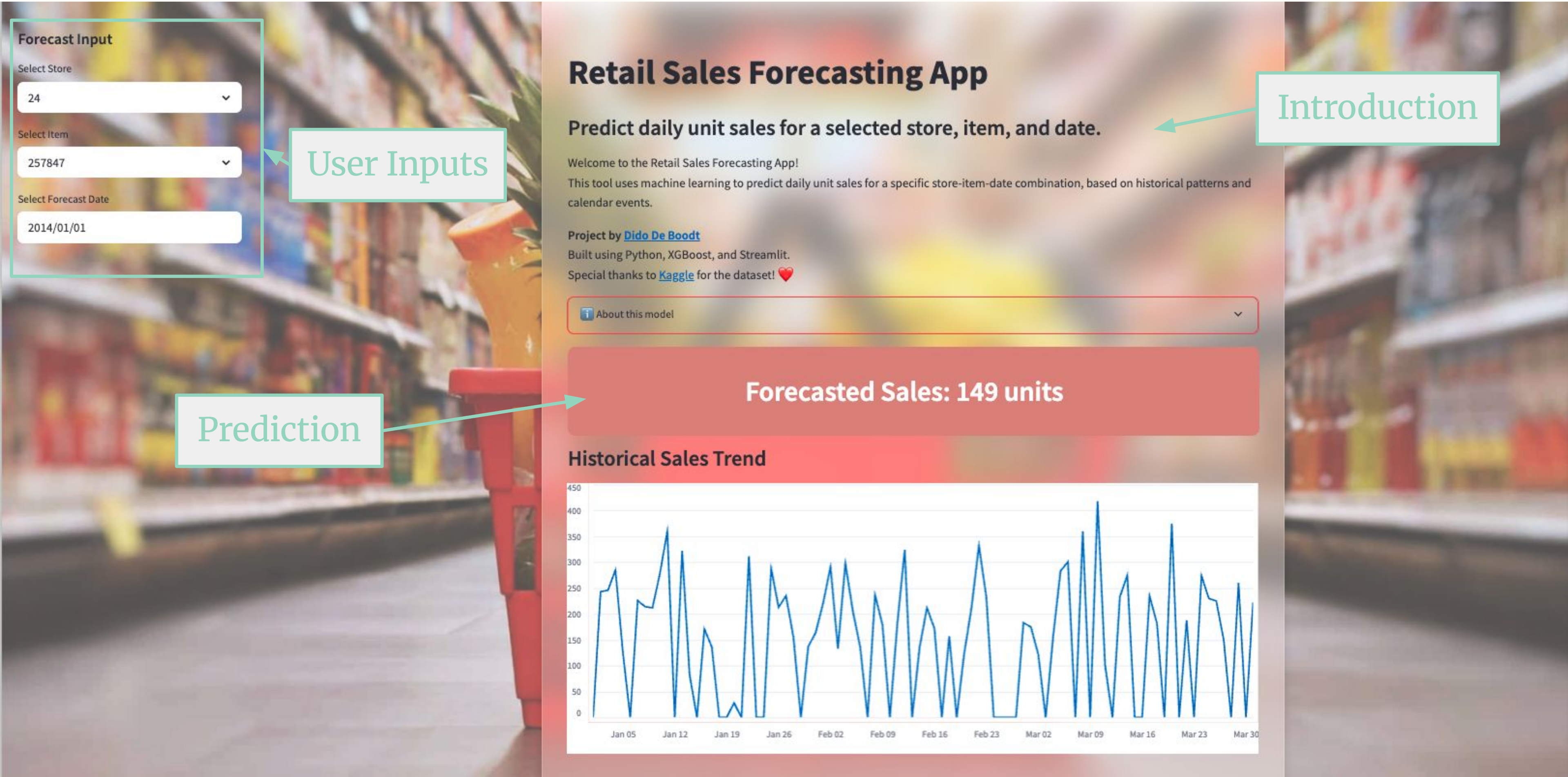
Evaluation Metrics

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

Key Takeaways:

- **XGBoost** had the best RMSE and Bias → most accurate and least skewed.
- **LSTM** achieved lowest MAD and SMAPE → most consistent error.
- **SARIMA** performed well capturing seasonality, but had slightly higher errors overall.

App Demo (Click here for the app)



Final Thought & Recommendations

Key Takeaways

- Time-based features significantly improved results
- ML models outperform classical ones
- LSTM requires tuning & memory usage
- XGBoost has best trade-off accuracy vs explainability

Recommendations

- Deploy app to support demand planning
- Automate updates and retraining
- Explore Prophet or attention-based models like Transformers

Thank you for your attention!

[LinkedIn]
[GitHub Repo]