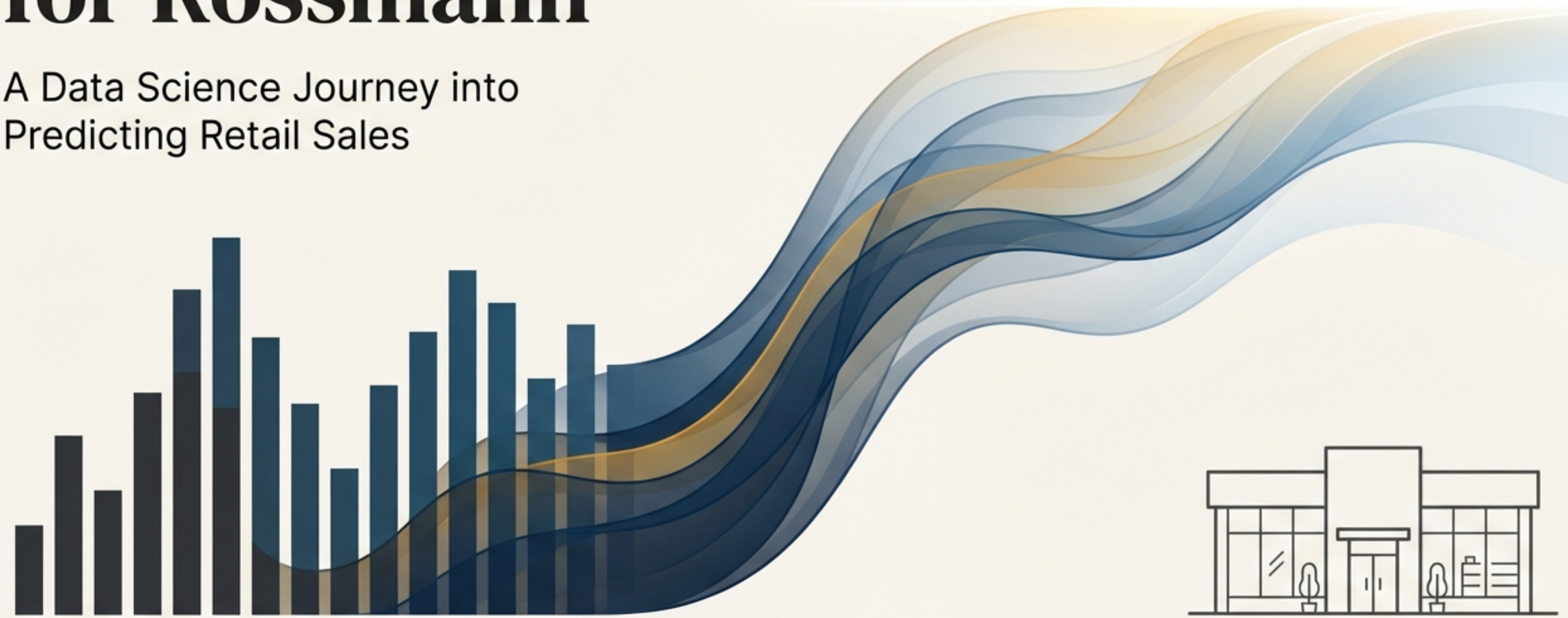
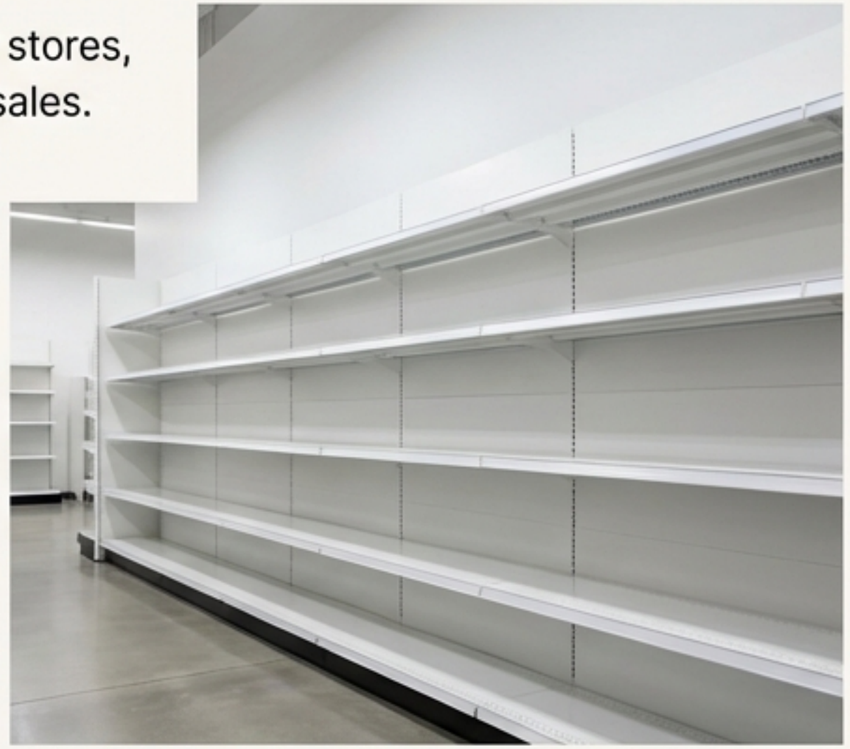
Forecasting the Future for Rossmann

A Data Science Journey into Predicting Retail Sales



The High-Stakes World of Retail Forecasting

Rossmann, a European chain with over 3,000 stores, faces a critical challenge: predicting daily sales.

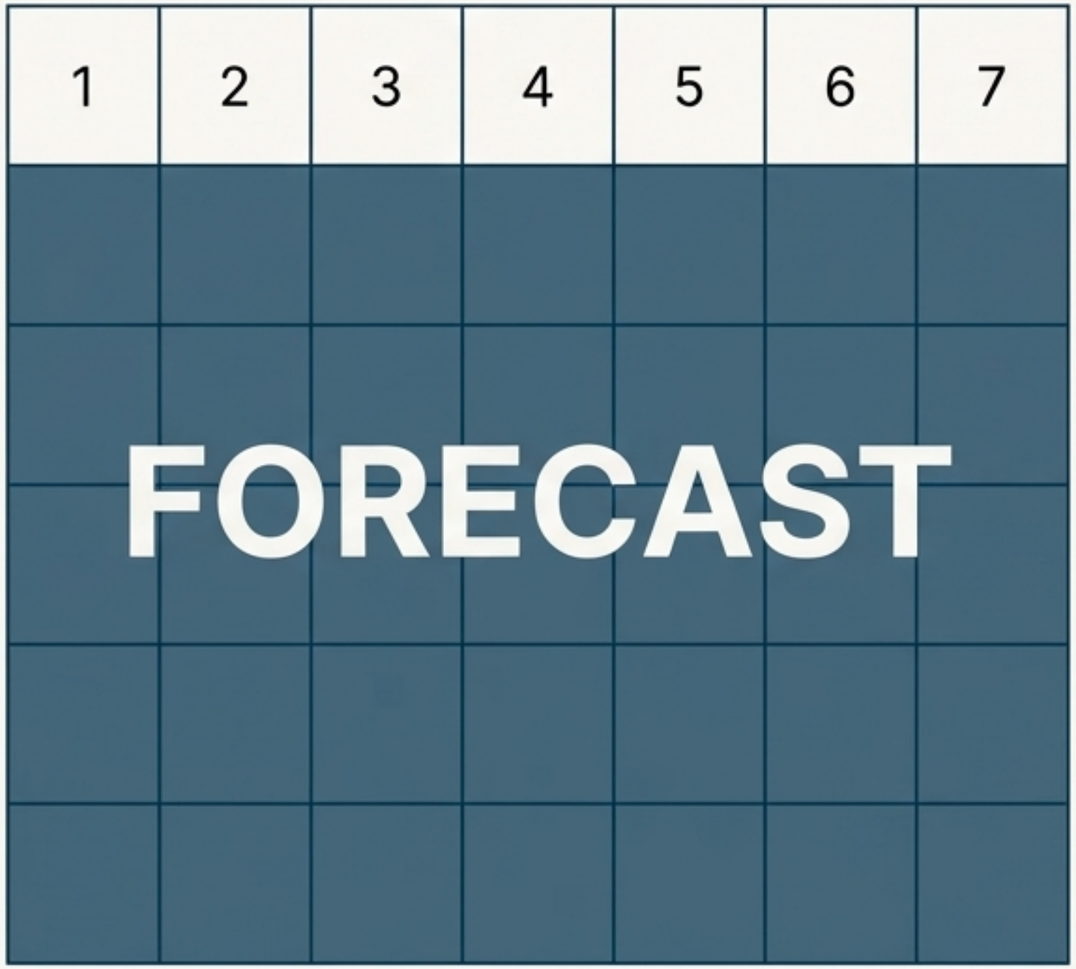


Many products have a short shelf life.

Over-stocking wastes resources; under-stocking loses sales.

Accurate forecasting is critical for operations.

Our Mission: Predict Sales for the Next 6 Weeks



Primary Goal

Forecast daily sales for key, high-value stores over a 42-day period.

Key Questions to Answer

- What is the impact of customer footfall?
- How do promotions affect sales?
- Which model tells the best story: classical time-series or modern machine learning?

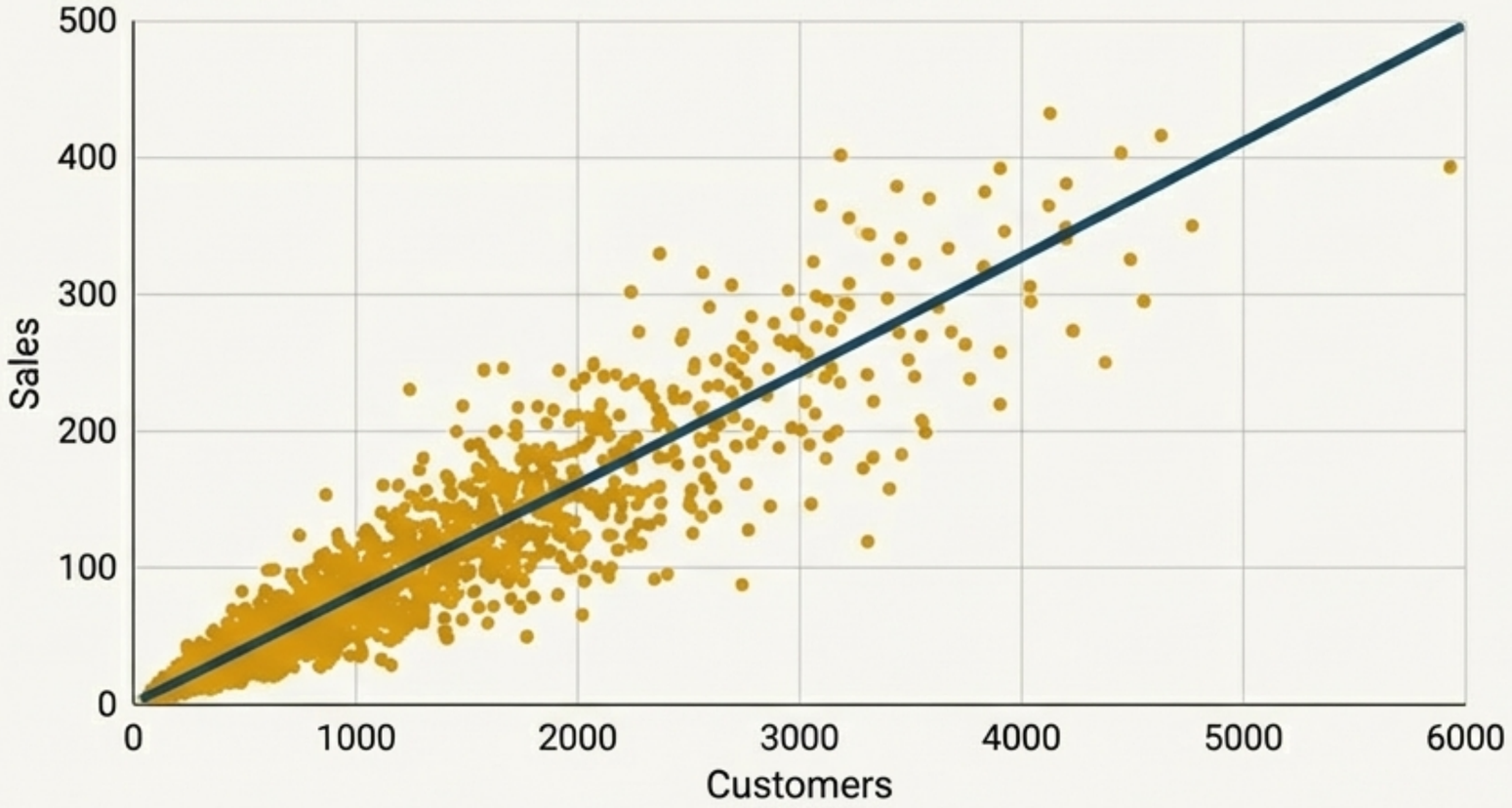
MAPE < 5%

The Measure of Success

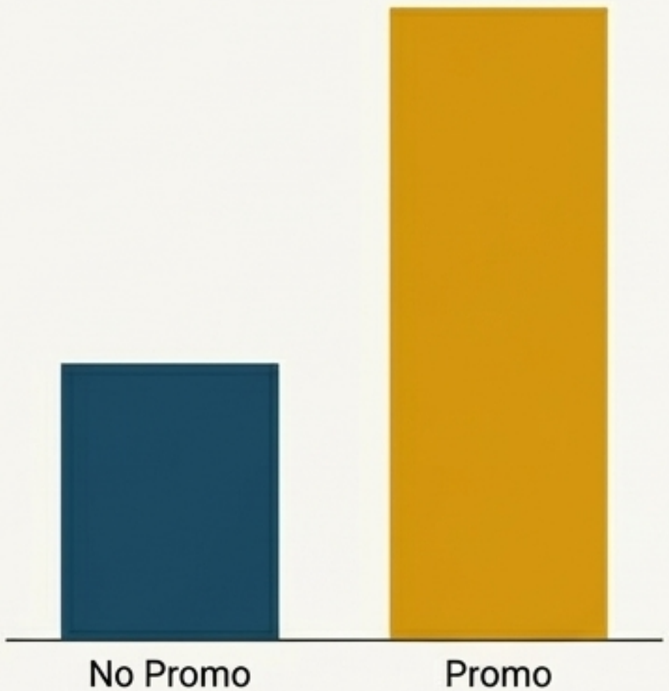
Uncovering the Patterns in the Data

Source of Truth: Two datasets were used:`train.csv` (daily sales data) and `store.csv` (store metadata).

Customers vs. Sales: A Powerful Connection.



Promo vs. Sales: A Consistent Uplift.

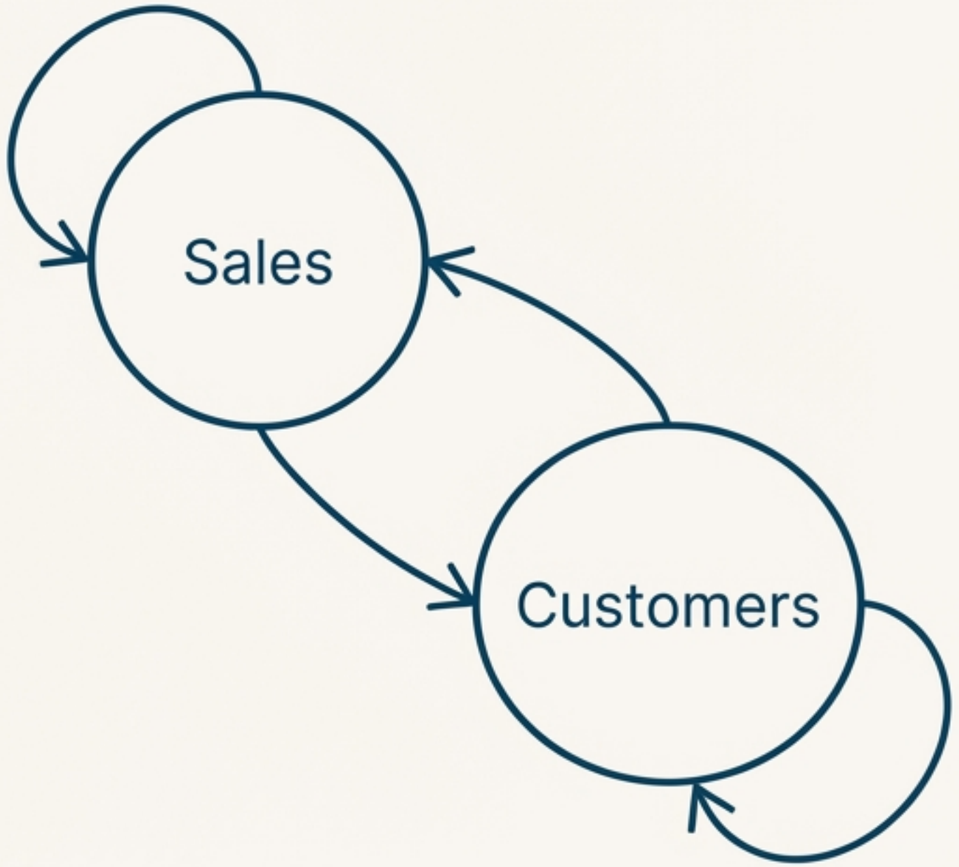


Our First Approach: A Classic Time-Series Model

The Contender:
VAR (Vector Autoregression).

The Logic:
VAR was chosen to model the relationship between multiple time-series variables simultaneously (like Sales and Customers).

The Prerequisite:
We performed stationarity checks (ADF test) and differencing to prepare the data, following best practices.



A Dead End: When Good Theory Meets Messy Reality

Numerical Instability

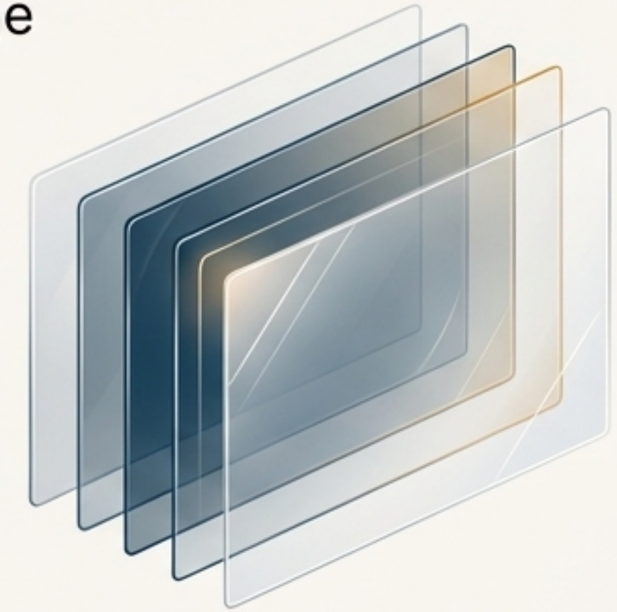
The VAR model was abandoned due to “numerical instability”.

The relationship between Sales and Customers was *too* strong and deterministic. **This is a known** limitation in some retail datasets where **footfall directly drives sales. The model couldn’t handle it.**

We needed a model pivot.

Two New Paths Forward

The Sage

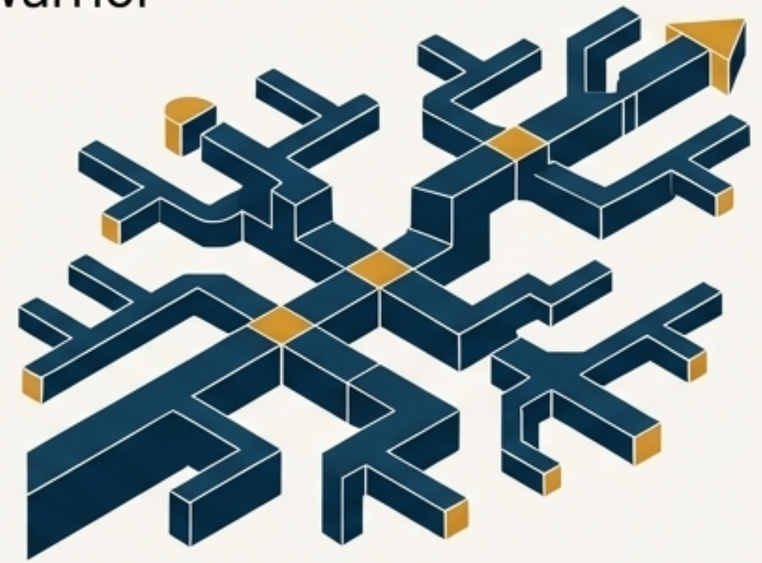


SARIMAX

The wise and transparent sage. Captures time, understands external drivers, and explains its reasoning clearly.

Interpretability | Stability | Temporal Structure

The Warrior



XGBoost

The powerful but mysterious warrior. Uses complex techniques to find hidden patterns and deliver maximum predictive power.

Accuracy | Non-linear | Feature Engineering

The Sage's Approach: Understanding the Why with SARIMAX

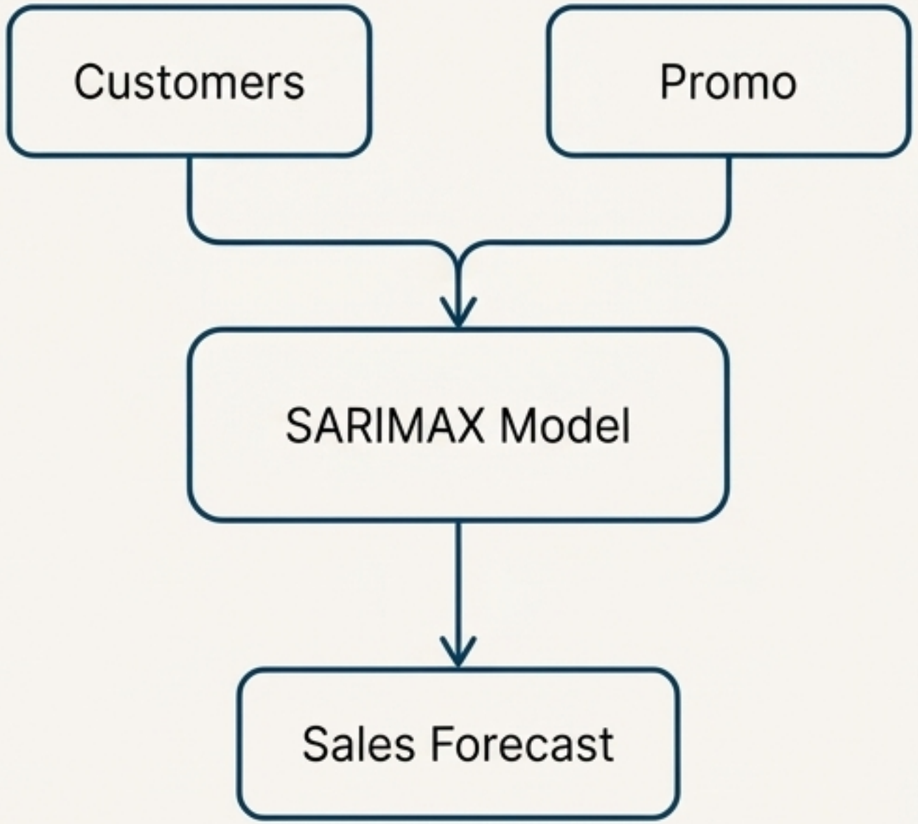
Model Configuration:
`SARIMAX(1,1,1)`

How it Works:

- It captures the inherent temporal structure (seasonality, trend).
- It directly incorporates external business drivers (`Customers`, `Promo`) as exogenous variables.

Key Strengths:

- **Highly Interpretable**: We can see exactly how much promotions or customer traffic impact sales.
- **Stable & Production-Ready**: Produces reliable and smooth forecasts.



The Warrior’s Power: Maximizing Accuracy with XGBoost

Model Configuration:
A Gradient-Boosted Regression Tree model.

How it Works:

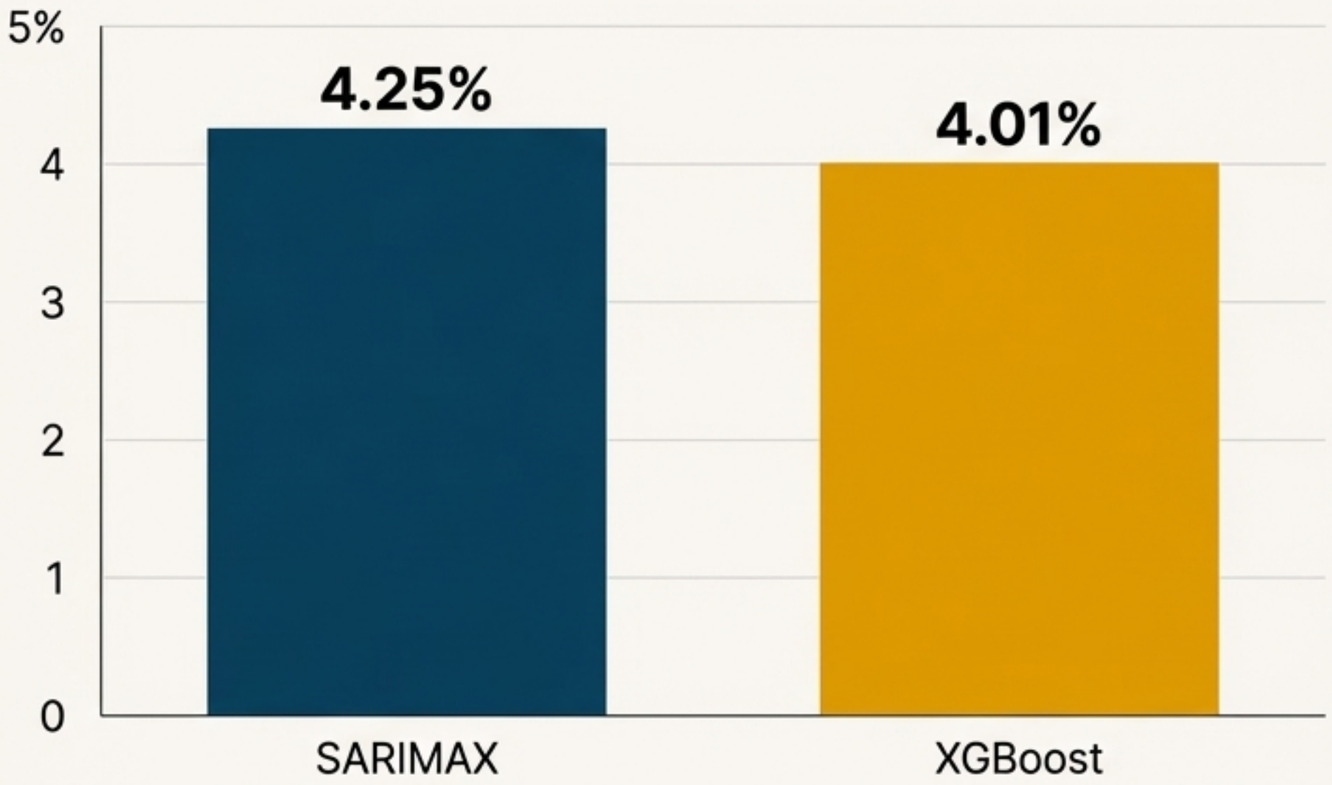
- It uses **Lag-based feature engineering** (e.g., sales from 1, 7, and 14 days ago).
- It understands **calendar features** like Day of Week and Month.

Key Strengths:

- **Captures Non-linear Patterns**: Finds complex relationships the other models might miss.
- **High Accuracy**: Often a top performer in forecasting competitions.

The Final Showdown: Accuracy on a 42-Day Holdout Set

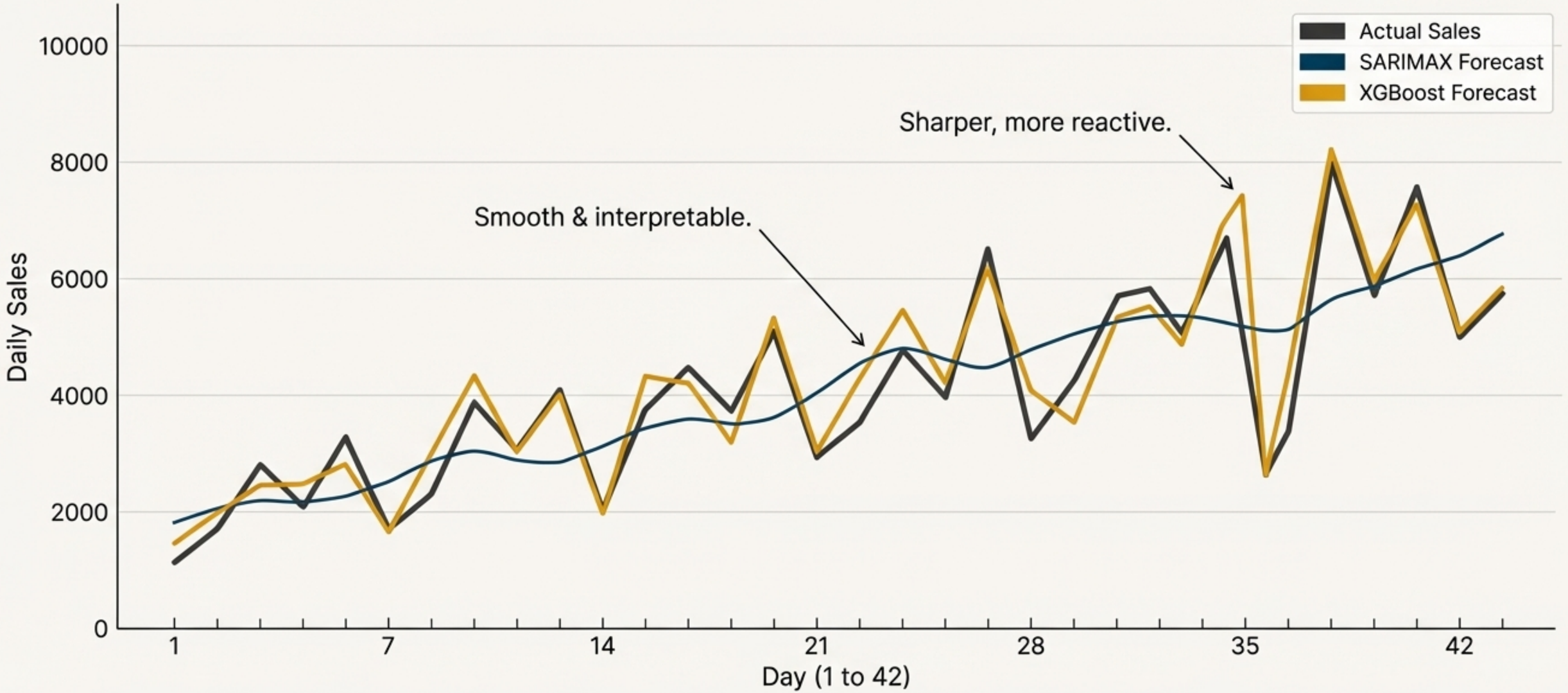
The Test: Both models were used to predict sales for the last 42 days of data, which they had never seen before.



Model	MAPE
SARIMAX	**4.25%**
XGBoost	**4.01%**

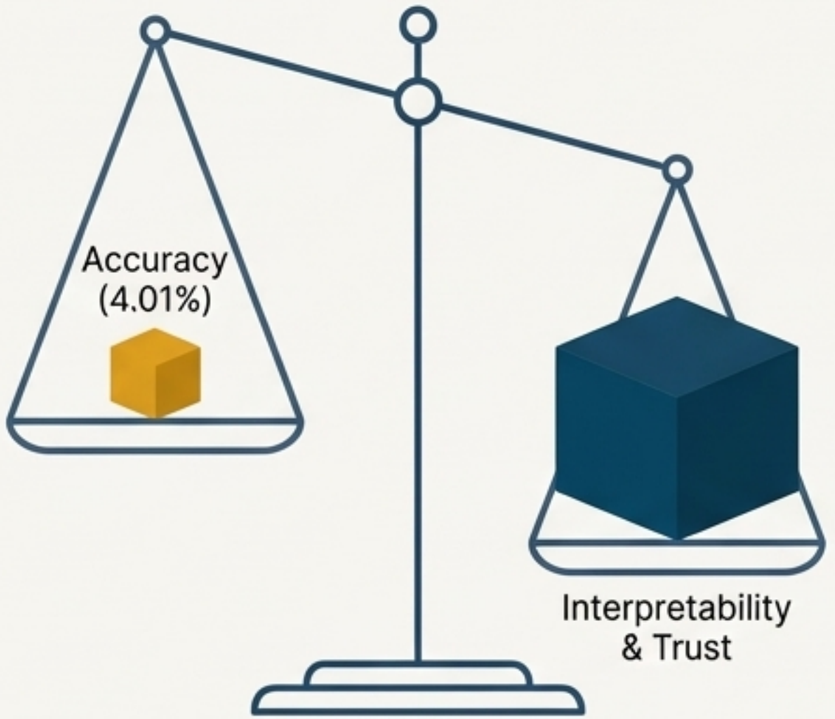
The Verdict: Both models achieved excellent accuracy (<5%). XGBoost is marginally more accurate.

Visualizing the Forecast: Smooth vs. Sharp



The Winning Decision: Why Interpretability Trumped Raw Accuracy

The Choice: **SARIMAX was selected as the final model.**



1. **Comparable Accuracy:** The small 0.24% accuracy gain from XGBoost was not significant enough to justify its complexity.
2. **Superior Interpretability:** SARIMAX provides a clear linkage to business drivers. We can tell stakeholders, “A promotion increases sales by X%.”
3. **Easier Explainability:** Simpler to explain, deploy, and maintain.

XGBoost’s Role: Retained as a high-performing benchmark for future comparisons.

Key Lessons from the Journey

Finding #1: Drivers

Customer footfall is the strongest predictor of sales. Promotions provide a consistent, measurable lift.

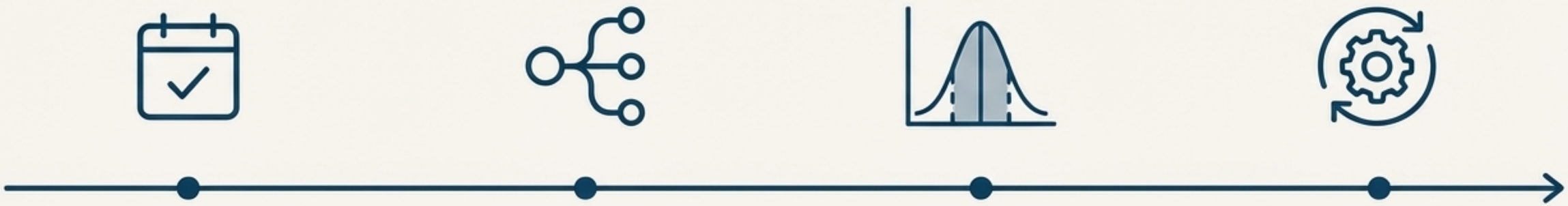
Finding #2: Methodology

Combining classical time-series models (like SARIMAX) and machine-learning models (like XGBoost) provides a robust forecasting system.

Finding #3: Philosophy

The ultimate model selection must always balance **Accuracy** with **Interpretability**. This is the core principle.

The Journey Continues: Future Enhancements



Incorporate detailed holiday calendars explicitly.

Extend the model to multi-store hierarchical forecasting.

Add probabilistic forecasts to provide confidence intervals (e.g., "we are 95% confident sales will be between X and Y").

Deploy the system as an automated forecasting pipeline.