Walmart Weekly Sales Forecasting

Capstone Project

Pooja Suryavanshi

vpoojasuryavanshi@gmail.com

Table of Contents

1. Problem Statement

Retail forecasting is a balancing act: stock too much, and capital is wasted; stock too little, and customers leave empty-handed. For Walmart, operating 45 stores across diverse U.S. regions, this challenge is compounded by local economic conditions, weather, fuel costs, holidays, and consumer behavior.

This project uses the **Walmart weekly sales dataset** (6,434 records) to answer two core questions:

1. **What drives sales?** How do unemployment, temperature, or holidays impact revenue?
2. **Can we accurately forecast the next 12 weeks of sales per store?**

The dataset includes 8 features: `Store`, `Date`, `Weekly_Sales`, `Holiday_Flag`, `Temperature`, `Fuel_Price`, `CPI`, and `Unemployment`. Our goal is to transform this data into actionable insights and reliable predictions for inventory planning.

2. Project Objective

This project has two primary goals:

1. Exploratory Analysis

- Quantify the impact of macroeconomic factors (unemployment, CPI, fuel price) on sales.
- Analyze how temperature and holidays influence store performance.
- Identify seasonal trends and their timing.
- Rank stores by historical performance and measure the gap between best and worst.

2. Predictive Modeling

- Develop and compare forecasting models:
  - **Machine Learning**: Random Forest, XGBoost
  - **Time Series**: Prophet, SARIMAX
- Build **ensemble forecasts** (e.g., XGBoost + Prophet) for robustness.
- Deliver **12-week sales forecasts** for all 45 stores.

The ultimate aim: provide Walmart with a data-driven forecasting system to optimise inventory and reduce waste.

3. Data Description

The dataset (`Walmart DataSet.csv`) contains 6,434 weekly records from 45 U.S. Walmart stores.

Column	Description
Store	Store ID (1–45)
Date	Week-ending date (DD-MM-YYYY)
Weekly_Sales	Total weekly sales ($)
Holiday_Flag	1 = holiday week, 0 = regular
Temperature	Avg weekly temperature (°F)
Fuel_Price	Regional fuel cost ($)
CPI	Consumer Price Index (inflation)
Unemployment	Regional jobless rate (%)

Key Characteristics:

- Temporal coverage: ~2.7 years of weekly data → ideal for short-term forecasting.

- Store-level granularity: Enables localized modeling.

- External variables: Supports analysis of economic impacts.

This structure supports both multivariate ML models and univariate time-series methods.

4. Data Pre-processing Steps and Inspiration

To ensure model reliability, we executed the following steps:

1. Data Cleaning

- Converted `Date` to datetime using `dayfirst=True`.
- Sorted by `Store` and `Date` to preserve time order.
- Filled missing values with column means (minimal missingness).

2. Feature Engineering

- Created lag features (`Lag1`–`Lag4`): Sales from previous weeks.
- Added 4-week rolling average to smooth volatility.
- These features help models learn short-term demand momentum.

3. Train-Test Split

- Used time-based split (first 80% train, last 20% test).
- Prevents data leakage, future doesn't inform past.

Inspiration:

In retail forecasting, time order is sacred. Lag features are industry-standard for capturing recent trends, while time-based splits mimic real-world deployment.

5. Choosing the Algorithm for the Project

We evaluated a blend of ML and time-series models:

1. **Random Forest**: Robust to noise, handles non-linear relationships.
2. **XGBoost**: Gradient boosting is fast, accurate, and feature-rich.
3. **Prophet**: Built for business time-series with holidays.
4. **SARIMAX**: Classical stats model for trends and seasonality.

This mix allows comparison of feature-driven ML vs. pattern-driven time-series.

6. Motivation and Reasons for Choosing the Algorithm

XGBoost & Random Forest

- Excel at tabular data with mixed features.
- Use lag features and external variables (temp, CPI) effectively.
- XGBoost won for speed and accuracy.

Prophet

- Automatically models U.S. holidays via `add_country_holidays('US')`.
- Requires minimal tuning, ideal for business use.

SARIMAX

- Pure time-series model, no external features needed.
- Captures cycles and seasonality statistically.

Ensemble Approach

- Combined XGBoost + Prophet and XGBoost + SARIMAX.
- Leverages strengths of both paradigms → more reliable forecasts.

This strategy balances accuracy, interpretability, and practicality.

7. Assumptions

Key assumptions made:

1. **Feature Stability**: Future values of temperature, fuel price, CPI, and unemployment are approximated using the last known week's values.

2. **No Structural Shifts**: No store closures, pandemics, or new competitors during the forecast horizon.

3. **Store Independence**: Each store is modeled separately no cross-store influence.

4. **Holiday Flag Accuracy**: The provided `Holiday_Flag` correctly marks major U.S. holidays.

These assumptions are standard in short-term retail forecasting.

8. Model Evaluation and Techniques

Evaluation Metrics

- MAE: Average dollar error.
- MAPE: % error intuitive for business.
- R^2: Variance explained.

Results (Time-Based Test Set)

Model	MAE	R^2	MAPE
Random Forest	$4,300	0.48	10.2%
XGBoost	$3,950	0.51	9.2%

XGBoost was selected as the best ML model.

Time-Series Models

- Prophet: Captured holiday spikes well (e.g., Dec 2010 surge).
- SARIMAX: Strong for stores with clear seasonal cycles; slower but stable.

Final Forecasts

- Built two ensembles:
  - XGBoost + Prophet
  - XGBoost + SARIMAX
- Visualized for Store 1, both tracked historical trends and projected future sales realistically.

9. Inferences from the Same

Business Insights

- Unemployment negatively affects sales (r = -0.15). Store 12 is the most sensitive.
- December shows massive sales spikes driven by holidays.
- Warmer weather slightly boosts sales (r = 0.18).
- CPI has a weak negative correlation (r = -0.08).
- Top stores: 20, 15, 10. Worst: Store 45. Gap: ~$36,000/week.

Model Insights

- XGBoost outperformed Random Forest by leveraging lag features.
- Prophet excelled at holiday modeling without manual feature engineering.
- SARIMAX provided a stable statistical baseline.
- Ensembles smoothed individual model weaknesses.

No single model dominated, hence the value of hybrid forecasting.

10. Future Possibilities of the Project

1. Real External Forecasts

   Use actual weather and economic forecasts for the next 12 weeks.

2. Store Clustering

   Group stores by behavior (e.g., “holiday-sensitive”) and build cluster-specific models.

3. Automated Dashboard

   Build a Streamlit or Power BI dashboard for real-time forecasts.

4. Weighted Ensembles

   Use MAE-based weights instead of simple averaging.

5. Longer Horizon

   Extend to 26 or 52 weeks for strategic planning.

6. Promotion Data

   Add marketing campaign data to explain sales bumps.

These steps would turn this into an operational forecasting system.

11. Conclusion

This project successfully combined exploratory analysis and advanced forecasting to tackle Walmart's sales prediction challenge.

We found that XGBoost, enhanced with lag features, delivers the most accurate ML-based forecasts. When combined with Prophet (for holidays) or SARIMAX (for time patterns), ensemble models provide robust, business-ready predictions.

The analysis confirmed that holidays dominate seasonal trends, unemployment hurts sales, and store performance varies widely, insights that can guide Walmart's inventory and marketing strategies.

This work demonstrates how practical data science can solve real retail problems with clarity and impact.

12. References

1. Walmart Dataset – Publicly available for educational use.

2. Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. KDD.

3. Taylor, S. J., & Letham, B. (2018). *Forecasting at Scale*. The American Statistician.

4. Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts.

5. Statsmodels Documentation – SARIMAX Implementation. https://www.statsmodels.org

6. Scikit-learn & XGBoost Official Documentation.

7. Prophet GitHub Repository. https://facebook.github.io/prophet/