Sales Synergy: Predicting Walmart Weekly Sales

Cheri Beda

DSC680-T301 Applied Data Science

Professor Amirfarrokh Iranitalab

Bellevue University

5/16/2025

Business Problem

The primary business problem addressed in this project is how to optimize inventory management at Walmart stores by predicting weekly sales. By improving forecast accuracy, Walmart can reduce stockouts and overstocks, enhancing customer satisfaction and operational efficiency.

Background/History

Walmart is one of the largest retail chains in the world, and efficient inventory management is crucial to its success. This project uses historical sales data combined with external factors like weather and economic indicators to explore patterns and predict future sales.

Data Explanation

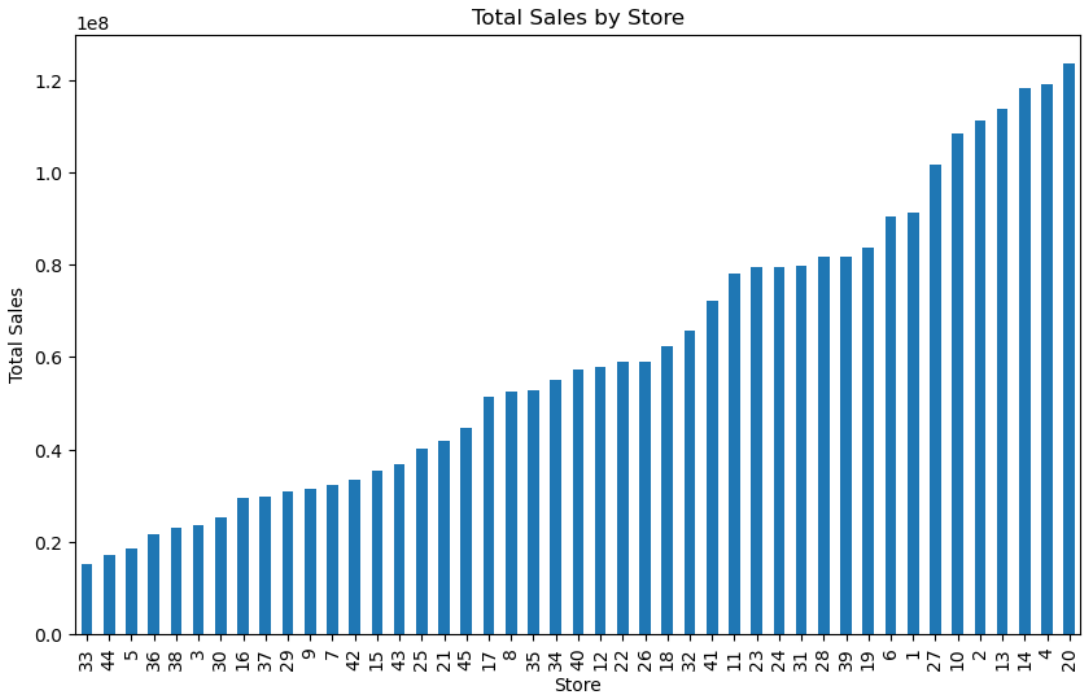
The dataset used is the publicly available Walmart Sales Forecast dataset from Kaggle. It includes weekly sales data for multiple Walmart stores with associated variables: store ID, weekly sales, holiday flag, temperature, fuel price, CPI, and unemployment rate. Data was preprocessed using Python's pandas library to standardize date formats, manage missing values, and create additional time-based features (year, month, week). Lagged features (sales from 1, 2, and 3 weeks prior) were added to enhance model performance.

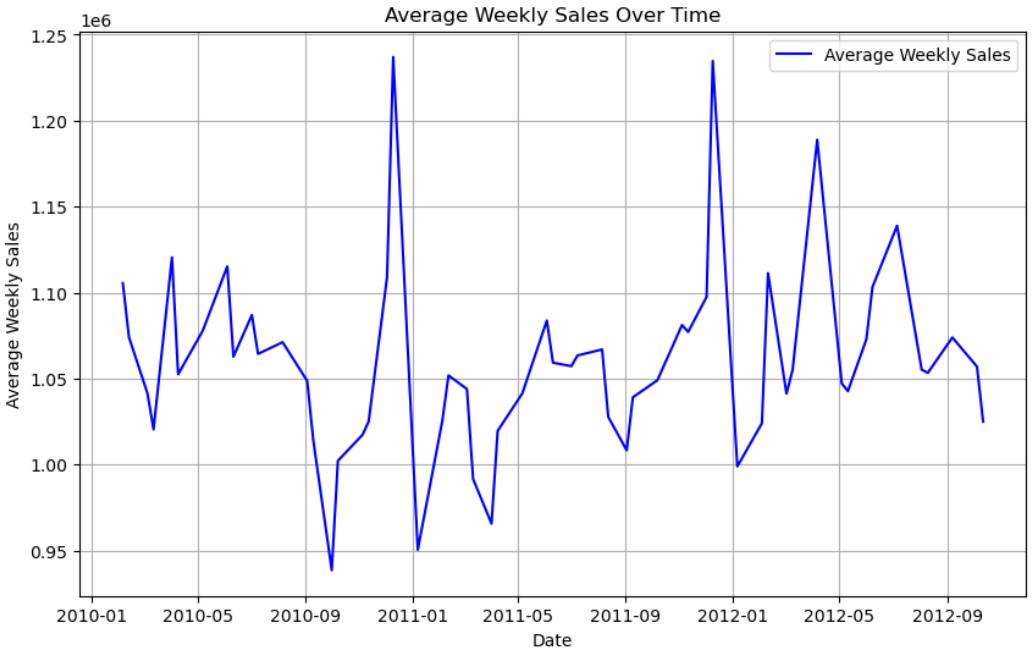
Methods

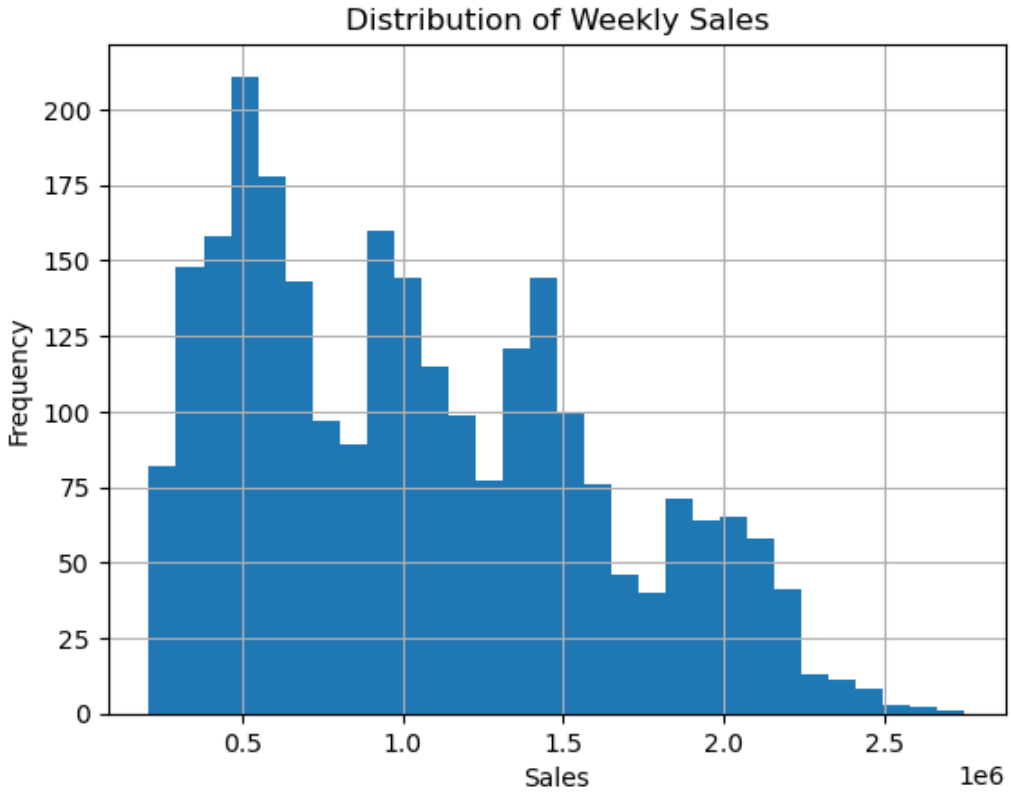
The primary predictive model used was XGBoost, selected for its strength in structured/tabular data. Additional methods included exploratory data analysis using Matplotlib and Seaborn for line plots, histograms, and heatmaps. Feature engineering and lag creation were central to improving the model's performance.

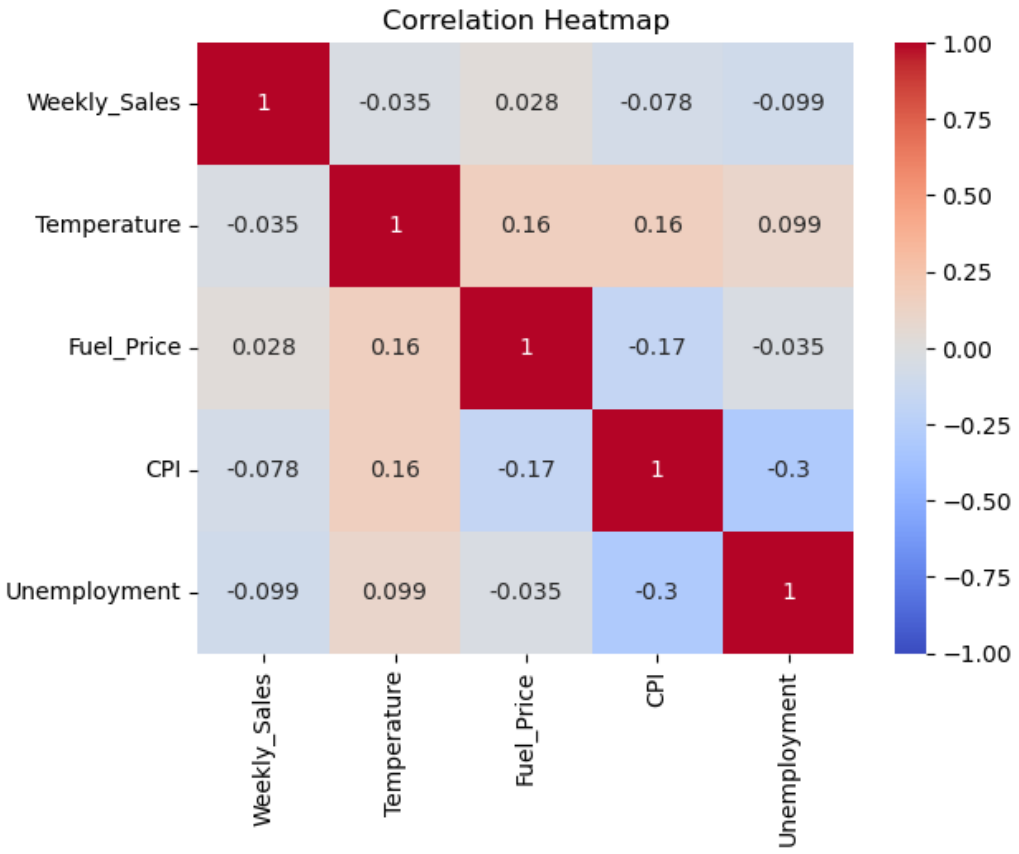
Analysis

The XGBoost model achieved an R^2 value of 0.9329 when lagged features were included, indicating strong explanatory power. Mean Squared Error (MSE) was significantly lower with lagged features. Visualizations revealed seasonal trends, holiday spikes, and regional store performance variability. A baseline comparison without lagged features showed a drastic drop in accuracy (R^2 = 0.1338), highlighting their importance.




Average Weekly Sales Over Time
1e6
1.25
1.20
1.15
1.10
1.05
1.00
0.95
Average Weekly Sales
Average Weekly Sales
2010-01
2010-05
2010-09
2011-01
2011-05
2011-09
2012-01
2012-05
2012-09
Date


Distribution of Weekly Sales
200
175
150
125
100
75
50
25
0
Frequency
0.5
1.0
1.5
2.0
2.5
Sales
1e6



Conclusion

The model successfully predicts weekly sales at Walmart and provides insights into sales trends and contributing variables. This enables better forecasting and decision-making around inventory, logistics, and promotions.

Assumptions

It is assumed that the provided data is accurate and that lagged weekly sales are a reliable predictor of future performance. It is also assumed that economic indicators and temperature affect consumer behavior.

Limitations

The dataset does not include promotional data, regional events, or online sales. These missing elements may impact prediction accuracy. Additionally, macroeconomic indicators might not fully capture local purchasing behavior.

Challenges

Key challenges included handling inconsistent date formats, generating lagged features without introducing data leakage, and selecting the appropriate model parameters. Future scalability and regional customization also present challenges.

Future Uses/Additional Applications

This model can be adapted for other retail chains, extended to daily or online sales, or used to forecast product-level demand. It may also support dynamic pricing and supply chain optimization strategies.

Recommendations

Walmart should incorporate lagged feature models into its inventory planning tools. It should also explore integrating promotional, weather, and competitor data into future models.

Implementation Plan

1. Deploy the model store-by-store as a forecasting tool.
2. Train managers on interpreting forecasts and adjusting inventory.
3. Monitor model accuracy and update weekly with new sales data.
4. Expand the model by integrating with real-time data streams.

Ethical Assessment

Though the dataset is public and anonymized, real-world application raises concerns around over-reliance on automation and potential bias in modeling. Ethical use should involve human oversight, regular validation, and transparency in decision-making.

Appendix: Supporting Documentation:

Jupyter Notebook with Python Code

Audience Questions

Why was XGBoost selected over other models?

XGBoost was chosen due to its high performance on structured/tabular data, its ability to handle missing values, and its efficiency in capturing non-linear relationships through gradient boosting, making it a good choice for retail sales forecasting.

What impact did lagged features have?

Lagged sales features significantly improved model performance, boosting the R^2 score from 0.1338 (without lag) to 0.9329. This highlights their critical role in capturing temporal patterns and trends in weekly sales behavior.

How often does the model need retraining?

Weekly retraining is recommended to incorporate the most recent data, especially for responding to seasonality, holidays, and shifting consumer behavior. Regular updates help maintain forecast accuracy.

What variables could improve the model further?

Including promotional calendars, product-level data, local events, weather forecasts, and online sales could provide deeper insights and enhance model accuracy. Customer sentiment data may also offer predictive value.

How would this scale to other retailers?

The approach is scalable across retailers that have historical sales and external variables. Customization would be needed for store-specific factors, regional behavior, and product mix, but the core model framework is transferable.

Were there ethical concerns in your feature selection?

Since the dataset is anonymized and public, ethical risks were minimal. However, in real-world applications, transparency and fairness must be ensured, especially if predictive outputs affect employment, pricing, or resource allocation.

Did you try a neural network or deep learning model?

Neural networks were considered, but XGBoost was prioritized due to its interpretability, lower data requirements, and better performance with the available tabular dataset. Deep learning may be explored in future iterations.

What are the risks of relying on automated forecasts?

Risks include over-reliance without human oversight, propagation of bias from historical data, and poor adaptability unpredictable events. Forecasts should supplement not replace expert judgment and contextual analysis.

How would Walmart integrate this in real-time?

Walmart could integrate the model into its existing inventory systems through a real-time data pipeline, leveraging APIs to update predictions weekly and automating stock-level adjustments based on forecast outputs.

Could this be used to forecast other KPIs (e.g., staffing, returns)?

Yes, similar models could be trained to forecast staffing needs, return volumes, or customer traffic, using appropriate historical data and temporal features, thereby expanding predictive capabilities beyond sales.

References

Abbott, D. (2014). *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst*. Wiley.

Ahmedov, A. (2023). Walmart Sales Forecast [Data set]. Kaggle. https://www.kaggle.com/datasets/aslanahmedov/walmart-sales-forecast/data

Brownlee, J. (2016, June 13). A gentle introduction to XGBoost for applied machine learning. *Machine Learning Mastery*. https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

NetSuite. (2021). Inventory analytics: Best practices for smart, data-driven decisions. https://www.netsuite.com/portal/resource/articles/inventory-management/inventory-analytics.shtml

Siegel, E. (2016). *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. Wiley.

Velvetech LLC. (2023). Retail predictive analytics: 10 game-changing use cases. https://www.velvetech.com/blog/retail-predictive-analytics/