# Walmart Sales Forecasting: A Time Series and Machine Learning Approach

Vasavi Kullanakoppal

Bellevue University

DSC 630: Predictive Analytics

Prof. Andrew Hua

May 23, 2025

**Walmart Sales Forecasting: A Time Series and Machine Learning Approach**

## Introduction of Topic/Problem

Retail businesses consistently seek to improve operational efficiency, inventory management, and staffing decisions through accurate sales forecasts. Walmart, as one of the largest retail chains globally, presents a particularly relevant case for forecasting due to its large-scale operations and highly seasonal sales patterns.

The core objective of this study is to develop a predictive model that forecasts weekly sales using historical data and external macroeconomic factors such as temperature, fuel prices, CPI, unemployment rates, and holiday events. Accurate forecasting models can help retail managers make informed decisions regarding demand planning and promotional strategies.

## Overview of Data Used

The data used for this project was obtained from Kaggle's publicly available Walmart dataset, which spans from February 5, 2010, to November 1, 2012. The dataset includes weekly sales figures for multiple Walmart stores along with the following attributes: Store, Date, Weekly Sales, Holiday Flag, Temperature, Fuel Price, CPI, Unemployment, and Holiday Events. The dataset enables the analysis of time series trends, holiday effects, and macroeconomic indicators on sales performance.

## Methods of Analysis

The data preparation involved transforming the dataset into a long format to facilitate time series modeling. Temporal features such as lagged sales values, encoded month and year, and macroeconomic indicators were engineered. Exploratory data analysis included line charts, boxplots, heatmaps, and scatter plots to identify patterns and correlations.

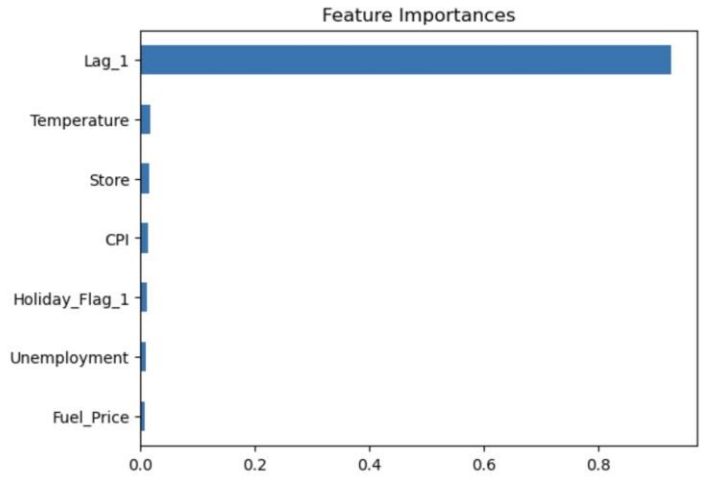
For modeling, several techniques were considered: ARIMA, SARIMA, Random Forest, and LSTM. Ultimately, a Random Forest Regressor was selected due to its strength in modeling nonlinear relationships and handling multicollinearity. A time-based split was used for training and validation to maintain the chronological integrity of the data. Evaluation metrics included Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to assess model accuracy.

**Evaluation Metrics**

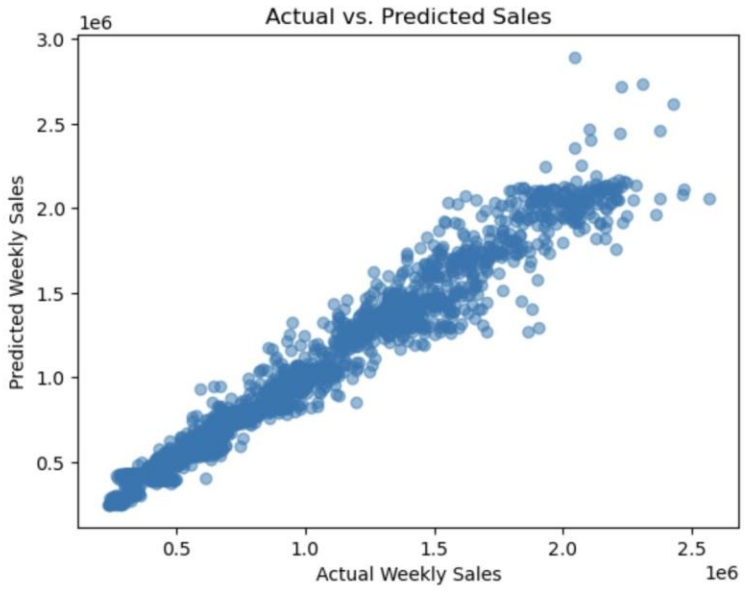
Mean Absolute Error (MAE): ~76,546 and Root Mean Squared Error (RMSE): ~110,366. From these results, it seems that the model can do quite well at predicting overall trends, but deviations may still occur, particularly during periods of large volume sales.

**Model Interpretation**

Feature Importance plot indicated that Lag_1 (Sales in the Previous Week) was by far the strongest predictor. All other features (e.g., CPI, Temperature, and Holiday Flag) had little or no impact on model performance in this setup.




The Actual vs. Predicted Plot shows a strong correlation between predicted and actual values, verifying the model's capacity to learn from history. Some variation in higher sales values suggests extra tuning will help improve predictions during peak weeks




**Results & Findings Explained**

The Random Forest model yielded a MAE of approximately $76,546 and RMSE of approximately $110,366, demonstrating a reasonable level of accuracy in capturing the overall sales trend. A feature-important analysis revealed that the most predictive feature was Lag_1, or the sales figure from the previous week. This finding reinforces the idea that recent sales data is a strong indicator of future performance.

Other variables such as CPI, Temperature, and Holiday Flags had relatively little impact, suggesting that macroeconomic indicators might not significantly influence short-term sales predictions at the weekly level, or that their effects are already baked into historical sales trends. A visual inspection of actual vs. predicted sales plots showed strong alignment for most weeks, with deviations occurring primarily during peak holiday weeks.

**Conclusion**

This study confirms that historical sales data is the most effective predictor for future sales, underscoring the importance for businesses to invest in robust data collection and forecasting infrastructure. While external variables such as economic indicators and temperature had limited predictive power in this context, incorporating specific holiday-related flags or promotional campaigns could improve peak season accuracy.

Future enhancements may include refining the feature set to flag specific holidays like Black Friday or integrating store-level demographic or marketing data. Additionally, tuning the model hyperparameters and testing deep learning models like LSTM could offer further performance gains. The model developed in this study can be effectively used by retail managers to improve demand forecasting, staff scheduling, and promotional planning.

**References**

Yasserh. (n.d.). *Walmart dataset*. Kaggle. https://www.kaggle.com/datasets/yasserh/walmart-dataset?resource=download

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). *Statistical and Machine Learning forecasting methods: Concerns and ways forward*. PLoS ONE, 13(3), e0194889. https://doi.org/10.1371/journal.pone.0194889

Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324