# Walmart Sales Forecasting – Report

# Abstract

This project focuses on forecasting Walmart’s weekly sales using historical data that includes store information, holiday events, and economic indicators. Two machine learning models—Linear Regression (baseline) and XGBoost Regressor (advanced)—were applied and compared. While the linear model struggled to capture complex patterns, XGBoost achieved excellent accuracy with an R² of ~0.95.

# 1. Introduction

Sales forecasting is critical for retail giants like Walmart to optimize supply chain management, staffing, and inventory control. This project aims to build machine learning models to predict weekly sales using historical sales data and store-level features.

The dataset, obtained from Kaggle, includes train.csv (weekly sales data), features.csv (economic and holiday features), and stores.csv (store-related metadata).

# 2. Data Preparation

## Data Cleaning:

- Dropped unused columns: MarkDown1–5

- Removed records with zero or negative sales

- Extracted year, month, day from the Date column

- Encoded categorical variables (IsHoliday, Type)

## Final Dataset:

- Features (X): Store, Dept, Type, Size, IsHoliday, Fuel Price, CPI, Unemployment, Temperature, Year, Month, Day

- Target (y): Weekly\_Sales

# 3. Baseline Model – Linear Regression

## Setup:

- Train/Test split: 80/20 (random)

- Model: Linear Regression

## Results:

- Mean Squared Error (MSE): 464,612,749

- Mean Absolute Error (MAE): 14,523

- R² Score: 0.09

Interpretation: The linear model performed poorly, failing to capture non-linear relationships and seasonality in sales.

# 4. Advanced Model – XGBoost Regressor

## Time-Aware Split:

- Training Data: Before 2012

- Testing Data: 2012 and beyond

## Hyperparameters:

- n\_estimators=500

- max\_depth=8

- learning\_rate=0.05

- subsample=0.8

- colsample\_bytree=0.8

## Results:

- Mean Squared Error (MSE): 22,122,995

- Mean Absolute Error (MAE): 2,703

- R² Score: 0.95

Interpretation: XGBoost significantly outperformed Linear Regression, capturing complex dependencies and seasonal trends effectively.

# 5. Results Comparison

## Linear Regression:

- MSE: 464,612,749 | MAE: 14,523 | R²: 0.09 (Poor fit)

## XGBoost Regressor:

- MSE: 22,122,995 | MAE: 2,703 | R²: 0.95 (Excellent fit)

# 6. Conclusion

- Linear Regression was unable to capture the variability of Walmart’s sales data, resulting in poor performance.

- XGBoost Regressor achieved an excellent fit with R² ≈ 0.95, highlighting its ability to model complex, non-linear patterns.

- Future Improvements: Introduce lag features and rolling averages, apply seasonal decomposition, and explore advanced time-series models (e.g., LSTM, Prophet).