Walmart Sales Forecasting

Brian Reppeto

Bellevue University

DSC 680

Prof. Neugebauer

December 15, 2024

Walmart Sales Forecasting

**Business Problem**

Predict future sales for Walmart to optimize inventory, allocate resources, and meet customer demands. Accurately forecasting sales helps to minimize overstocking, reduce costs, and ensures demand for products is met. This project will address the challenges of forecasting department-level weekly sales across Walmart stores utilizing historical data and external factors.

**Background/History**

Founded in 1962, Walmart has grown into the world’s largest retailer (Walmart, 2024). Walmart has a vast store network and diverse product range. This creates challenges in demand forecasting due to seasonal trends, holiday periods, economic fluctuations, and promotional campaigns. Traditional methods often fail to capture these nuances, prompting the need for advanced machine learning approaches to improve forecasting accuracy.

**Data Explanation**

**Source and Structure**: The dataset has three key files (Cukierski, 2014):

1. Stores: Details such as store size and type.
2. Features: Weekly economic data (CPI, unemployment), markdown information, and holiday indicators.
3. Sales: Weekly department-level sales for each store.

**Data Preparation**

* Missing values in CPI and unemployment were imputed with column means.
* Missing markdown values were replaced with zero.
* Date columns were converted to datetime format, and derived features such as “Year” and “Days to Next Christmas” were added.
* One-hot encoding was applied to the “Type” column.

**Data Dictionary**

* Store: Unique identifier for each store.
* Dept: Unique identifier for each department.
* Date: Date of observation.
* Weekly\_Sales: Weekly sales revenue (target variable).
* Size: Physical size of the store.
* Type: Store type (categorical: A, B, C).
* CPI: Consumer Price Index.
* Fuel\_Price: Regional fuel price.
* Unemployment: Regional unemployment rate.
* IsHoliday: Boolean indicating a holiday week.

**Methods**

A variety of machine learning models were employed:

1. Decision Tree Regressor: For interpretability and handling non-linearity.
2. Random Forest Regressor: For robust predictions via ensemble learning.
3. XGBoost: For its superior handling of feature importance and regularization.
4. KNN Regressor: As a baseline to capture localized trends.

The analysis includes data standardization for numeric columns, train-test splits (80/20 ratio), and performance evaluation using MAE, RMSE, and accuracy metrics.

**Analysis**

* EDA: Highlighted strong seasonality in sales and a correlation between CPI, unemployment, and markdowns with weekly sales.
* Model Performance
  + Random Forest achieved the highest accuracy (92%) with a low RMSE.
  + XGBoost closely followed but required higher computational resources.
  + An ensemble of top models further reduced errors.

**Conclusion**

The utilization of machine learning significantly improved Walmart’s sales forecasting accuracy. Random Forest and XGBoost emerged as the top performers, with ensemble methods offering additional robustness.

**Assumptions**

* Economic indicators (CPI and unemployment) are reliable predictors.
* Holiday impact is consistent year-over-year.
* Promotional markdowns influence customer behavior uniformly.

**Limitations**

* Limited coverage of external factors like competitor pricing.
* Potential overfitting in complex models like XGBoost.
* Sparse markdown data limits its predictive power.

**Challenges**

* High computational cost for large datasets.
* Balancing model complexity with interpretability.
* Handling seasonality and irregular sales spikes during holidays.

**Future Uses/Additional Applications**

* Dynamic Pricing: Adjusting prices in real-time based on forecasted demand.
* Inventory Optimization: Automating restocking processes to minimize waste.
* Promotional Planning: Tailoring markdown strategies to maximize ROI.

**Recommendations**

* Prioritize Random Forest and XGBoost models for deployment.
* Incorporate external data like weather and local events.
* Continuously monitor model performance and retrain periodically.

**Implementation Plan**

1. Phase 1: Data pipeline setup and model deployment.
2. Phase 2: Train staff on model usage and interpretation.
3. Phase 3: Monitor real-time performance and fine-tune parameters.

**Ethical Assessment**

* Transparency: Ensure stakeholders understand the forecasting process.
* Fairness: Avoid over-reliance on variables that may lead to bias (socioeconomic data).
* Privacy: Secure customer data used in forecasting models.
* Impact: Avoid decisions that could negatively affect employees or suppliers.

**Questions for Milestone 4**

1. What variables most influence the sales predictions?
2. How are holiday effects accounted for in the model?
3. Why were Random Forest and XGBoost chosen as primary models?
4. How does the ensemble method improve accuracy?
5. What steps were taken to prevent overfitting?
6. Can the models be generalized to other retailers?
7. How does data imputation affect model performance?
8. What ethical concerns arise from using sales forecasts?
9. How are extreme outliers handled in sales data?
10. What additional features could improve forecast accuracy?
11. Distribution of Weekly Sales (Histogram)

A graph of a sales distribution

Description automatically generated with medium confidence

1. Store Type Distribution (Pie Chart)

A pie chart with numbers and a few different colored circles

Description automatically generated

1. Correlation Heatmap

A graph with red and blue squares

Description automatically generated

1. Model Accuracy Comparison (Bar Chart)

A graph showing different colored bars

Description automatically generated

References

Cukierski, W. (2014). Walmart recruiting - store sales forecasting. Retrieved December 10, 2024, from <https://kaggle.com/competitions/walmart-recruiting-store-sales-forecasting>

Walmart. (2024). Walmart History. Retrieved December 15, 2024, from <https://corporate.walmart.com/about/history>