Walmart Sales Forecasting Project Report

Project by: Parth Chaudhari

Introduction

Walmart, being one of the largest retail chains globally, faces the challenge of managing inventory and sales effectively across its stores. Predicting weekly sales is crucial for optimizing operations, inventory allocation, and enhancing customer satisfaction. This report explores the data, methodologies, visualizations, and model evaluation used to predict Walmart's weekly sales.

This project utilizes three datasets provided by Walmart:

1. **Train.csv**: Weekly sales data for various stores and departments.
2. **Features.csv**: Additional information like markdowns, CPI, fuel price, and unemployment.
3. **Stores.csv**: Metadata about the stores, such as type and size.

Link to Datasets: [Kaggle - Walmart Sales Forecasting]

Methodology

Data Preprocessing

To ensure the datasets are clean and usable for modeling, the following steps were undertaken:

1. **Merging Datasets**: Combined Train.csv, Features.csv, and Stores.csv using common columns like Store and Date.
2. **Handling Missing Values**:
    a. Imputed missing values in CPI and Unemployment using the median.
    b. Replaced null or negative values in MarkDown columns with 0.
3. **Feature Engineering**:
    a. Added time-based features such as Year, Month, and Week.
    b. Combined markdown columns into a single feature: Total_MarkDown.
4. **Outlier Removal**: Removed extreme values using Z-score filtering.
5. **Normalization**: Scaled numeric features using MinMaxScaler.
6. **One-Hot Encoding**: Encoded categorical variables like Store, Dept, and Type for compatibility with machine learning models.

Model Building

Four machine learning models were evaluated:

1. **Linear Regression**: A baseline model to compare performance.

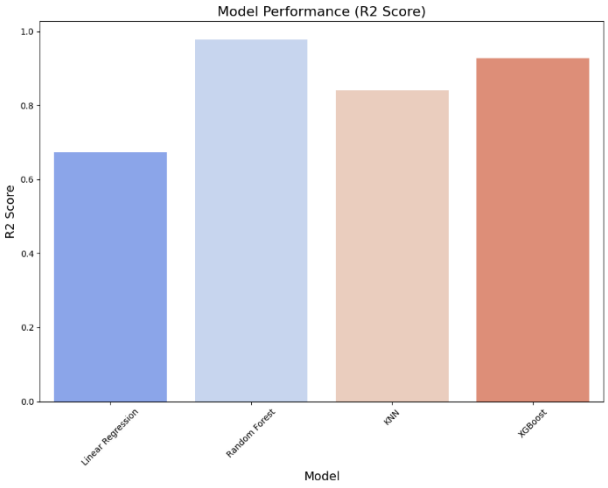
2. **Random Forest Regressor**: A robust ensemble method capturing non-linear relationships.
3. **K-Nearest Neighbors (KNN)**: Predictions based on proximity to historical data.
4. **XGBoost Regressor**: A high-performance gradient boosting model.

The models were evaluated using metrics like Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2).

Visualizations and Insights

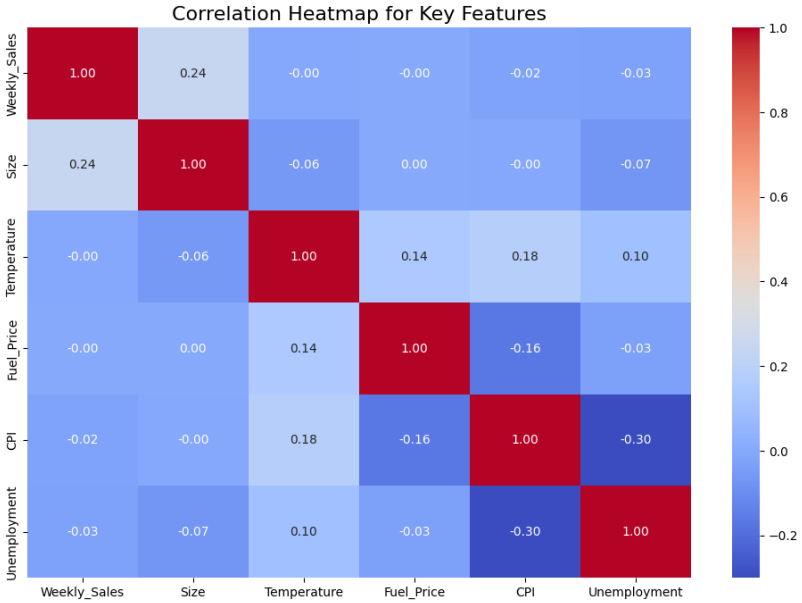
1. Model Performance (R^2 Score)

The bar plot below compares the R^2 scores of the models. **Random Forest** achieved the highest score (0.978), followed by **XGBoost** (0.927).



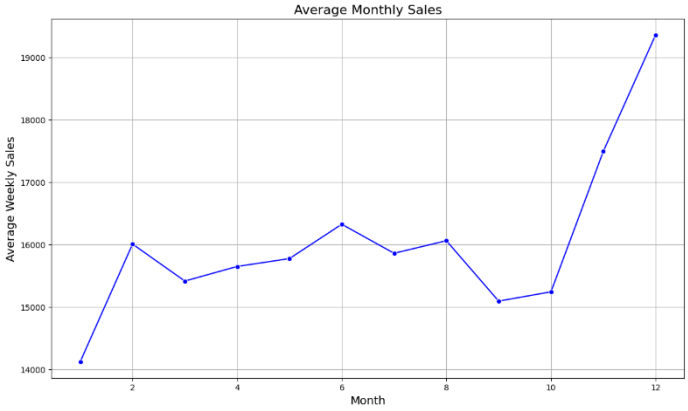
2. Correlation Heatmap

The heatmap highlights the relationships between key features. Features like Size showed a weak positive correlation with Weekly_Sales, while economic indicators like Fuel_Price and Unemployment had minimal impact.



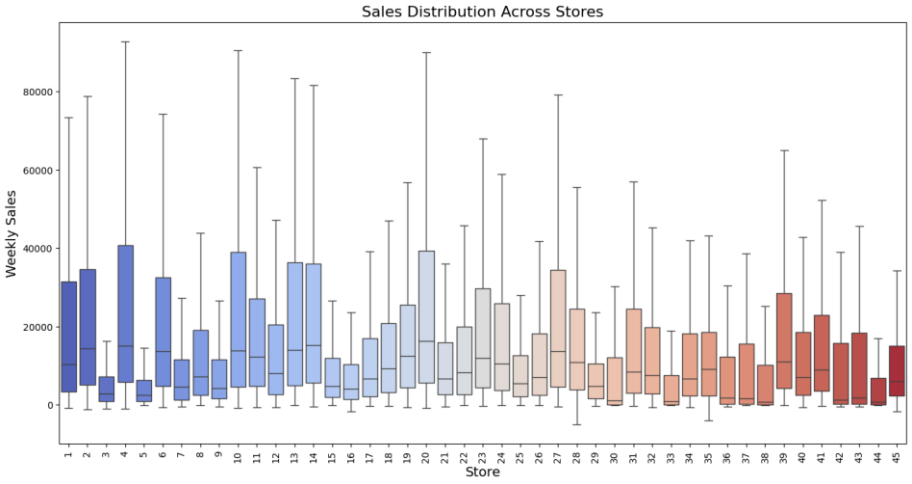
3. Average Monthly Sales

The line plot below reveals strong seasonality, with December having the highest average sales due to the holiday season.



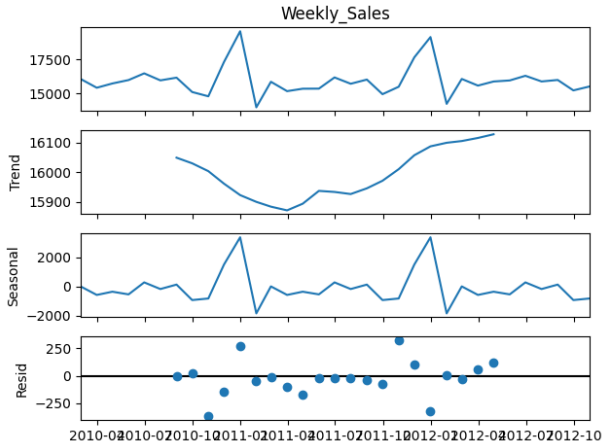
4. Sales Distribution Across Stores

The box plot shows significant variability in sales across stores. Some stores consistently outperform others, likely due to factors like location and size.



5. Time Series Decomposition

The time series decomposition highlights the trend, seasonal, and residual components of Weekly_Sales. Seasonal peaks align with major shopping periods.



Results

Model Evaluation

The table below summarizes the performance of the models:

Model	RMSE	MAE	R^2
Linear Regression	0.120606	0.084675	0.673215
Random Forest	0.031119	0.015706	0.978244
KNN	0.084425	0.031736	0.839874
XGBoost	0.057102	0.036970	0.926748

Key Findings

1. **Best Model**: Random Forest outperformed all other models, achieving the lowest RMSE and highest R^2.
2. **Seasonality**: Sales peak in December, indicating the importance of holiday-driven demand.
3. **Economic Indicators**: Features like Fuel_Price and Unemployment showed weak correlations with sales.
4. **Store-Level Insights**: Sales variability across stores suggests location-specific strategies could improve overall performance.

Conclusion

This project successfully predicted Walmart's weekly sales using machine learning models. The key takeaways include:

- Random Forest and XGBoost are the most reliable models for sales forecasting.
- Strong seasonality patterns necessitate proactive inventory management during peak months.
- Sales variability across stores highlights opportunities for location-specific improvements.

Recommendations

1. **Inventory Optimization**: Allocate more inventory during holidays and peak seasons.
2. **Markdown Analysis**: Reassess markdown strategies to ensure their effectiveness.
3. **Store-Specific Strategies**: Focus on high-performing stores to replicate successful practices.

Future Scope

1. **External Factors**: Incorporate weather conditions and local events to improve forecasts.
2. **Regional Models**: Develop models tailored to specific store types or regions.
3. **Advanced Techniques**: Explore deep learning models like LSTMs for better time series predictions.

References

- Dataset: Kaggle - Walmart Sales Forecasting
- Libraries: Pandas, Seaborn, Matplotlib, Scikit-learn, XGBoost, Statsmodels