Retail Forecasting Insights with CRISP-DM: A Case on Walmart Sales

Chayan Shah

25th October 2024

Abstract

Accurate retail sales forecasting is essential for optimizing inventory, refining supply chain operations, and improving decision-making frameworks. This paper utilizes the Cross-Industry Standard Process for Data Mining (CRISP-DM) on Walmart's weekly sales data from Kaggle, aiming to both predict sales and classify them by category. Employing Linear Regression and Random Forest models, the study assesses model effectiveness through R-squared scores, confusion matrices, and classification metrics. The findings suggest that while Random Forest yields better predictive accuracy, further enhancement can be achieved via feature engineering and integrating time-series analysis.

Keywords: CRISP-DM, Retail Forecasting, Machine Learning, Regression, Random Forest, Classification

1. Introduction

Retailers rely on dependable sales forecasts to maintain optimal inventory levels, enhance customer satisfaction, and ensure seamless operations. As one of the largest global retail chains, Walmart generates substantial data across its stores, regions, and varying seasonal trends. By analyzing sales trends, Walmart can make informed decisions related to promotions, staffing, and supply chain efficiency.

In this study, we apply the CRISP-DM framework to Walmart's sales data, using it as an end-to-end guide for data mining. This includes regression for sales prediction and classification for segmenting stores based on sales volume. The study employs Linear Regression as a foundational model and Random Forest for non-linear scenarios, culminating in an assessment of their effectiveness and suggested future improvements.

2. Literature Review

The evolution of retail forecasting has seen both traditional models, like ARIMA, and contemporary machine learning approaches gaining prominence. Machine learning models such as Random Forest, Gradient Boosting, and Neural Networks offer greater flexibility in handling complex patterns within large datasets, as observed in recent studies.

CRISP-DM, initially proposed by Shearer, has gained traction due to its structured, iterative process, guiding analysts through each phase from business understanding to deployment. Compared to alternative methodologies like SEMMA and KDD, CRISP-DM's emphasis on iteration and business context has made it more suitable for retail applications requiring adaptability and robust model refinement.

3. Data and Methodology

The Walmart dataset, publicly available on Kaggle, encompasses weekly sales data across multiple stores, including variables such as temperature, fuel prices, and economic indicators. Following the CRISP-DM framework:

- **Business Understanding**: Forecasting sales to guide Walmart's inventory and staffing.
- **Data Understanding**: Initial exploration of data trends and distribution across metrics like sales, temperature, and holidays.
- **Data Preparation**: Data preprocessing includes addressing missing values, normalizing metrics, and generating time-based features.
- **Modeling**: Linear Regression serves as a baseline model, complemented by Random Forest for more nuanced feature interactions.
- **Evaluation**: Model performance is gauged through R-squared, MSE, and classification accuracy for sales categorization.

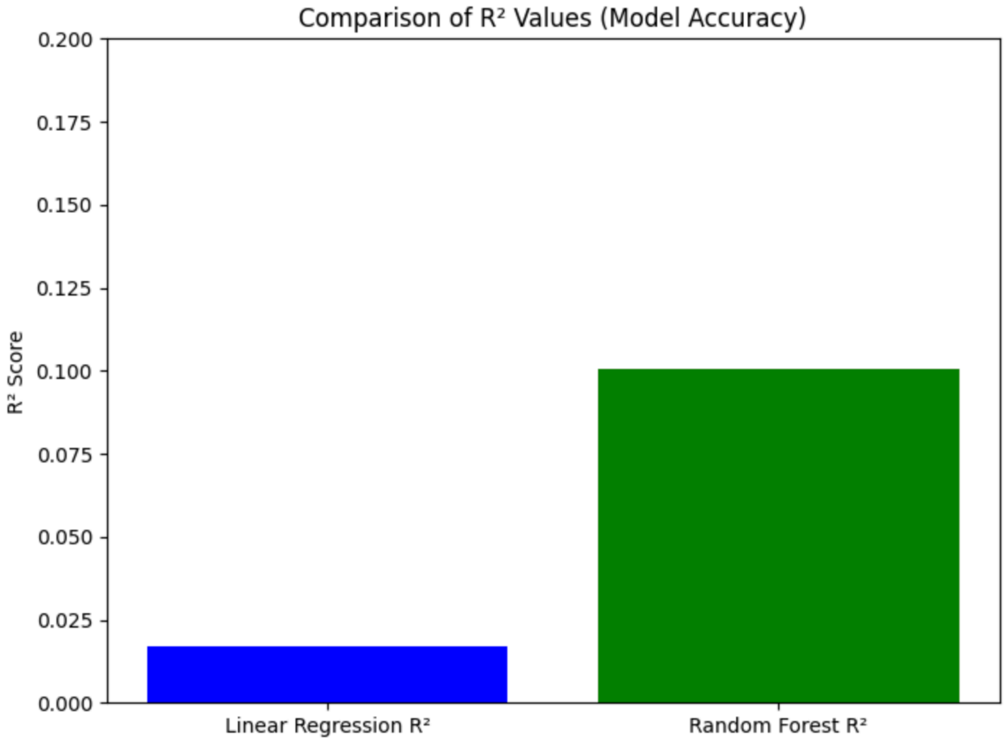
4. Modeling

4.1 Linear Regression

Linear Regression provides a straightforward method for analyzing relationships between sales and factors like fuel price and temperature. However, the model yields limited predictive power (R-squared = 0.017), suggesting that linear assumptions do not fully capture the sales dynamics.

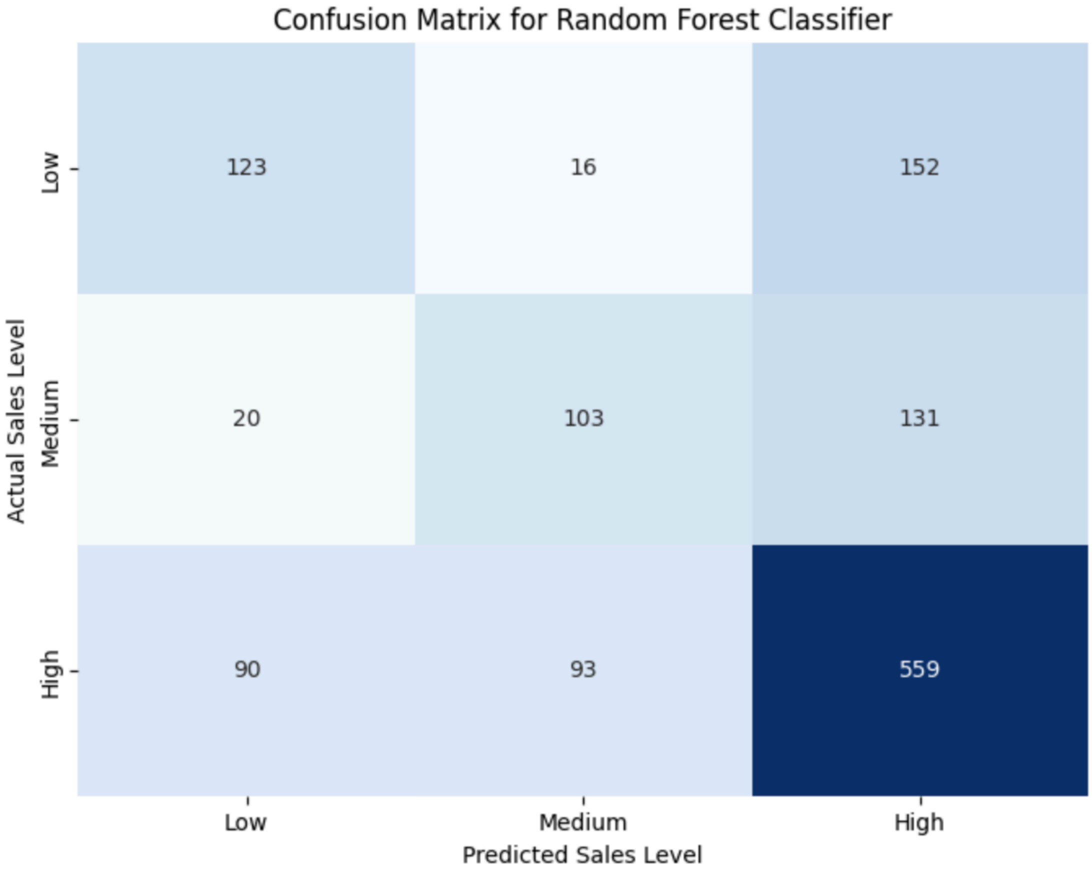
4.2 Random Forest Regressor

As a more advanced alternative, Random Forest can handle non-linear interactions. After hyperparameter tuning, the model achieved an R-squared of 0.101, a marked improvement over linear regression.



4.3 Classification Task

Sales were classified into low, medium, and high categories, with the Random Forest Classifier outperforming in accuracy, especially in high sales prediction.



5. Results and Discussions

The Random Forest model demonstrated superior predictive capabilities, particularly in high-sales weeks. Key variables, such as year, temperature, and unemployment, were identified as significant predictors. Despite Random Forest’s advantage over linear models, classification performance indicates potential for further refinement, especially for medium and low sales segments.

6. Conclusion and Future Work

In applying CRISP-DM to retail sales forecasting, Random Forest proved effective for Walmart’s sales data. However, the analysis could benefit from more granular data, such as store-specific attributes or promotional activities. Future research may explore time-series forecasting with ARIMA or LSTM models to capture temporal dependencies, enabling more accurate and actionable sales insights.