Comprehensive Retail Sales Analysis Using CRISP-DM on Walmart Sales Data

Rutuja Nemane

October 21, 2024

Abstract

Retail sales forecasting plays a critical role in supply chain management, inventory control, and decision-making. This paper applies the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to the Walmart Sales Dataset from Kaggle to predict weekly sales and classify sales into different categories. Linear Regression and Random Forest models are applied for regression and classification tasks, and the results are analyzed using R-squared, confusion matrix, and classification metrics. The models show that while Random Forest outperforms Linear Regression, further improvements can be made through feature engineering and time series analysis.

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

1 Introduction

Retail businesses are heavily reliant on accurate sales forecasting to optimize inventory, improve customer satisfaction, and ensure smooth operations. Walmart, being one of the largest global retailers, generates massive amounts of sales data across different stores, locations, and times of the year. Predicting sales trends can help Walmart in making informed decisions regarding promotions, staffing, and supply chain management.

This study utilizes the CRISP-DM methodology, a robust framework for data mining, to conduct an end-to-end analysis of Walmart's weekly sales. Two approaches are employed: regression for sales prediction and classification for categorizing stores based on sales levels. We use Linear Regression as a baseline model and Random Forest for non-linear relationships. Finally, we evaluate the models' effectiveness and suggest improvements.

2 Literature Review

Numerous studies have been conducted on retail sales forecasting using machine learning techniques. Traditional statistical models like ARIMA have been commonly used for time-series forecasting in retail. More recently, machine learning models such as Random Forest, Gradient Boosting, and Neural Networks have gained popularity due to their ability to capture non-linear relationships in large datasets [1].

CRISP-DM, introduced by Shearer [2], has been widely adopted as a framework for structuring data mining projects, particularly in retail. It provides a systematic approach to understanding the business problem, preparing data, building models, and evaluating their performance. Alternative methodologies like SEMMA and KDD focus more on model-building phases but lack CRISP-DM's iterative approach, making CRISP-DM more suitable for real-world business applications [3].

3 Data and Methodology

3.1 CRISP-DM Methodology

CRISP-DM consists of six phases, as shown below:

- **Business Understanding**: The objective of this study is to forecast sales based on historical data and to classify sales levels to inform inventory and staffing decisions.

- **Data Understanding**: The dataset contains weekly sales data for Walmart stores, including features such as temperature, fuel price, holiday indicators, Consumer Price Index (CPI), and unemployment rates.

- **Data Preparation**: Data preprocessing involves handling missing values, normalizing numerical features, and creating new time-based features.

- **Modeling**: Linear Regression and Random Forest models are built to predict sales and classify sales levels.

- **Evaluation**: Model performance is evaluated using R-squared, Mean Squared Error (MSE), and classification metrics such as precision and recall.

- **Deployment**: The model can be deployed in a retail decision-support system for real-time sales forecasting.

3.2 Dataset Description

The dataset used for this analysis is publicly available on Kaggle and contains 6,435 records from Walmart stores. Features include:

- **Store**: Store number.

- **Date**: The week-ending date.

- **Weekly Sales**: Sales amount in USD.

- **Holiday Flag**: Indicator for holiday week.
- **Temperature**: Regional temperature in Fahrenheit.
- **Fuel Price**: Regional fuel price in USD.
- **CPI**: Consumer Price Index.
- **Unemployment**: Regional unemployment rate.

4 Modeling

4.1 Linear Regression

We first applied a simple Linear Regression model to predict weekly sales using features such as temperature, fuel price, CPI, and unemployment. Although Linear Regression is easy to interpret, the R-squared value of 0.017 indicated that it was insufficient for capturing the complexity of retail sales data.

4.2 Random Forest Regressor

Next, we used a Random Forest Regressor, which provides flexibility in handling non-linear relationships and can account for interactions between features. After tuning hyperparameters, we achieved an R-squared value of 0.101, demonstrating a noticeable improvement over linear regression.

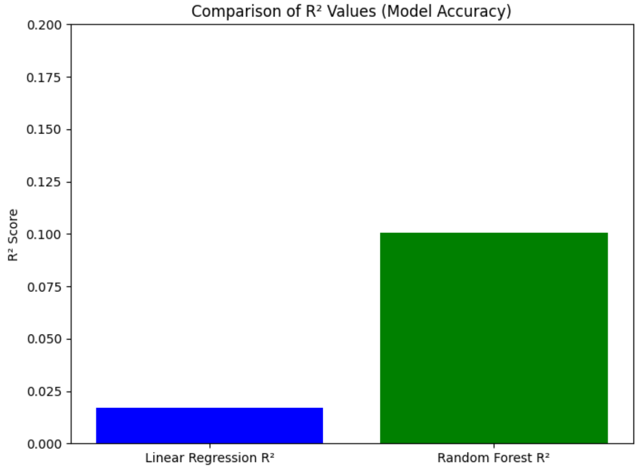

Figure 1: Comparison of R-squared values between Linear Regression and Random Forest

4.3 Classification Task

For classification, we categorized sales into three levels: low, medium, and high. The Random Forest Classifier was used to predict these categories, and the resulting confusion matrix is below.

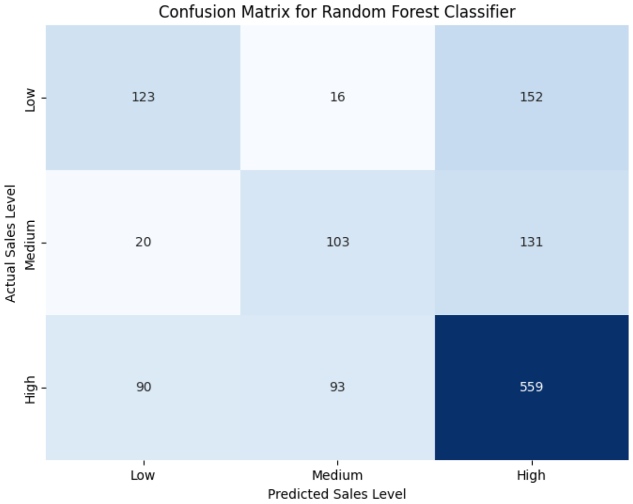

Figure 2: Confusion Matrix for Sales Classification

5 Results and Discussions

The Random Forest Regressor showed better performance in predicting weekly sales, as seen from the R-squared value. In the classification task, the model performed well in predicting high sales but struggled with medium and low sales. Feature importance analysis revealed that year, temperature, and unemployment were significant predictors of weekly sales.

5.1 Feature Importance

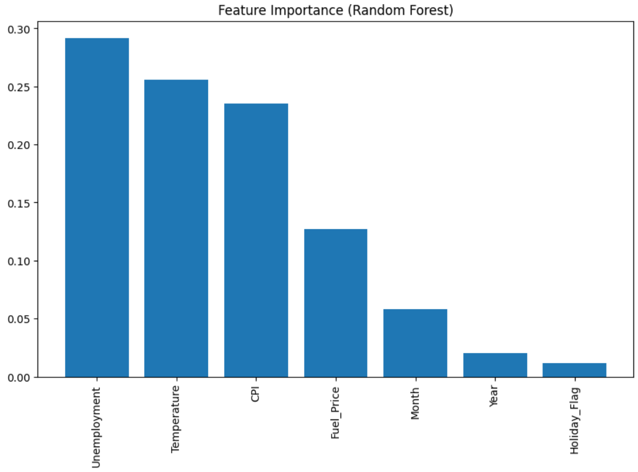

Figure 3: Feature Importance for Sales Prediction

6 Conclusion and Future Work

This study demonstrates the effectiveness of the CRISP-DM methodology in retail analysis. The Random Forest model outperformed linear regression in predicting sales. However, classification performance can be improved with more granular data, such as promotions or store-specific information. Future work will involve time-series forecasting techniques like ARIMA or LSTM to capture the temporal patterns in sales.

References

[1] Sezgin, M., Sankur, B. (2004). Survey over image thresholding techniques and quantitative performance evaluation. Journal of Electronic Imaging, 13(1), 146-165.

[2] Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Mining, 5(4), 10-20.

[3] Wirth, R., Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 29-39.