CRISP-DM Methodology for Retail Sales Forecasting

Puneet Bajaj
1st November, 2024

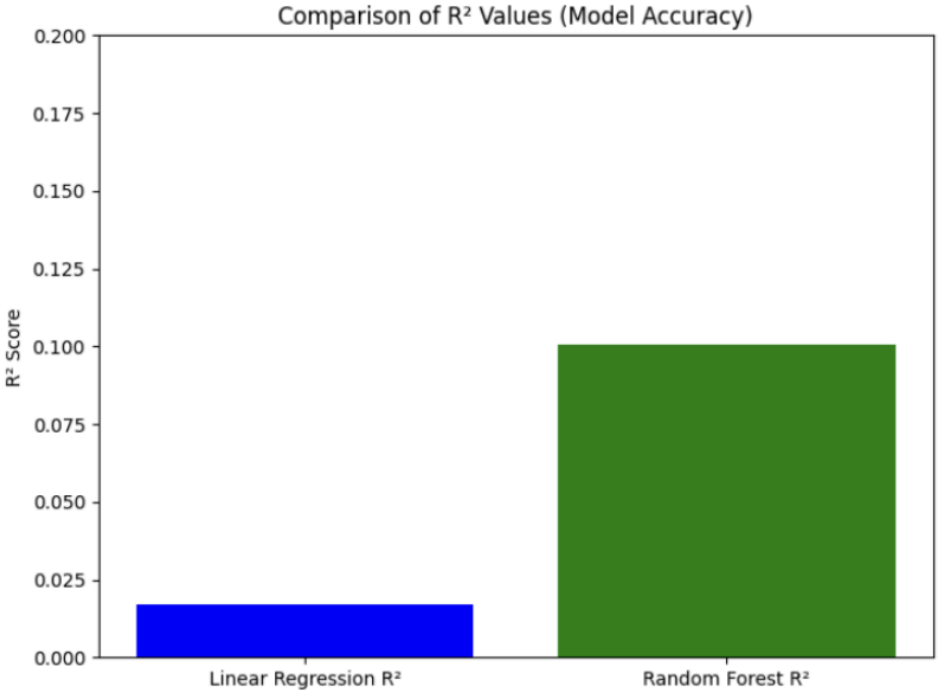
Abstract
This paper examines the application of the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to Walmart's weekly sales data, aiming to predict sales and categorize stores based on performance levels. Two machine learning models were employed—Linear Regression and Random Forest—with Random Forest showing a stronger ability to capture sales trends and non-linear relationships. Evaluation metrics, including R-squared, mean squared error (MSE), and classification accuracy, were used to compare model performance. The study suggests further improvements through feature engineering and advanced time-series methods for enhanced predictive accuracy in retail forecasting.

Introduction
Forecasting sales is crucial for retailers like Walmart to optimize inventory, improve customer experience, and enhance supply chain efficiency. Given Walmart's massive sales data spanning numerous stores and regions, a structured methodology was necessary for handling the dataset's complexity. CRISP-DM provided an iterative and comprehensive framework, supporting each phase from business understanding through data mining to deployment. This study utilizes CRISP-DM to develop predictive models for weekly sales, employing Linear Regression as a baseline and Random Forest for improved prediction.

CRISP-DM Phases
The CRISP-DM process comprises six interconnected phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Each phase is critical for building a reliable and adaptable sales forecasting model.

1. **Business Understanding**: The primary objective was to create a model capable of predicting weekly sales and classifying stores based on sales volume, enabling Walmart to make data-driven decisions about inventory, staffing, and promotions. Secondary goals included identifying seasonal trends and understanding the impact of external factors (e.g., holidays, economic indicators) on sales. These insights inform store-level operations and strategic planning across Walmart's retail network.
2. **Data Understanding**: Walmart's weekly sales dataset includes features like regional temperature, fuel price, holiday indicators, Consumer Price Index (CPI), and unemployment rates. These factors were selected for their potential influence on consumer spending and sales trends. Key challenges in this phase involved understanding seasonal sales variations and assessing the impact of economic indicators on store performance, which guided the choice of modeling techniques to capture non-linear trends effectively.

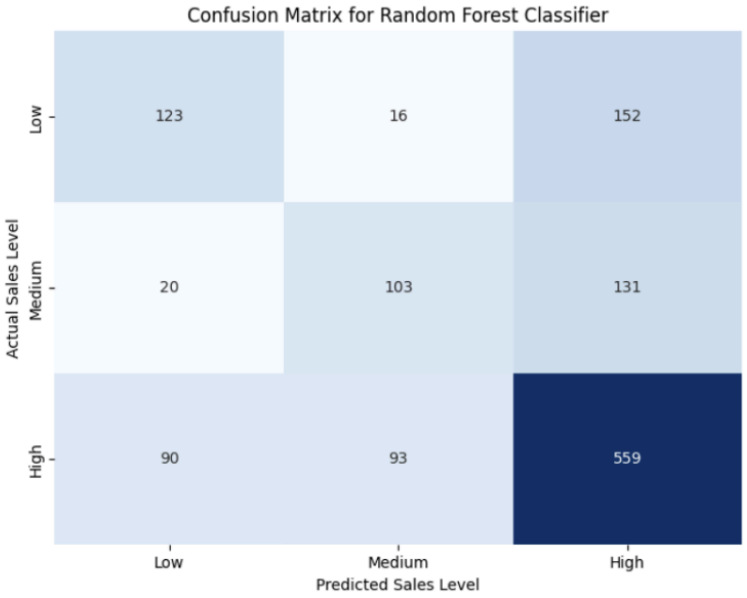
3. **Data Preparation**: Preprocessing steps were designed to improve data quality and support model training:
    - **Handling Missing Values**: Missing data points were handled using imputation techniques. For example, temperature and CPI were imputed with median values to prevent bias.
    - **Feature Scaling and Normalization**: Numerical features were normalized to ensure consistency across varying scales, enabling the model to make accurate predictions.
    - **Feature Engineering**: Additional features, such as month, season, and year, were created to capture temporal effects and seasonality, enhancing the model's ability to recognize sales patterns across different time frames.
4. **Modeling**: Two models were applied to forecast weekly sales and classify stores:
    - **Linear Regression**: Used as a baseline, Linear Regression provided interpretability and simplicity, though its R-squared value (0.017) indicated limitations in capturing complex patterns within the sales data.
    - **Random Forest**: As an ensemble method, Random Forest effectively captured non-linear relationships and feature interactions, outperforming Linear Regression with an R-squared value of 0.101. This model demonstrated a stronger ability to handle seasonal and trend-based sales data, producing more accurate predictions.



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

5. **Evaluation**: The models were evaluated based on several metrics:
    - **R-squared**: Random Forest's higher R-squared indicated a better fit for the sales data, capturing variations that Linear Regression missed.
    - **Mean Squared Error (MSE)**: MSE provided a quantitative measure of prediction accuracy, with Random Forest demonstrating a significantly lower error rate.

- **Classification Metrics**: For the classification task, Random Forest showed high accuracy in identifying high-sales stores but struggled with medium and low-sales categories. A confusion matrix revealed classification accuracy by sales category, supporting future improvements.



Confusion Matric for Sales Classification

6. **Deployment**: The final Random Forest model can be integrated into a decision-support system for real-time sales forecasting, providing Walmart with actionable insights for day-to-day and strategic decision-making. Integrating the model into a dashboard could enhance its usability, allowing Walmart's retail managers to make informed decisions based on predictive insights.

Conclusion

The CRISP-DM framework proved effective for structuring a comprehensive retail sales forecasting model. While Random Forest significantly outperformed Linear Regression, future improvements could involve time-series models, such as ARIMA or LSTM, to capture sequential sales patterns more accurately. This study illustrates how the CRISP-DM methodology can guide data mining in complex retail environments, enabling better demand planning, operational efficiency, and customer satisfaction.