1]Hasan Mhowwala [1]San Jose State University

A Data-Driven Exploration of E-Commerce Sales Patterns and Predictions

[

September 30, 2023

Abstract

With the surge in e-commerce growth, understanding sales patterns becomes paramount. This paper offers an in-depth exploration of an e-commerce dataset, culminating in a predictive model for product sales. The study emphasizes the importance of a systematic data analysis and modeling approach, highlighting significant predictors for e-commerce sales.

1 Introduction

In the modern commercial landscape, e-commerce platforms are integral. Gleaning insights from sales data paves the way for effective marketing and inventory management. This research aims to identify patterns in e-commerce sales data and create an accurate sales prediction model.

2 Data Understanding

Our dataset encapsulates details of various products: their prices, stock levels, reviews, advertising budget, and corresponding sales. Initial observations revealed a distinction among three product categories: Electronics, Home Appliances, and Clothing.

3 Data Visualization

Visual explorations discerned a differential relationship between product categories and sales. Notably, Electronics and Home Appliances manifest a more extensive sales range compared to Clothing, suggesting potential variations in demand or marketing efficacy.

4 Data Preparation

Ensuring the data is well-suited for machine learning models required preprocessing:

- **Encoding**: The 'Category' column underwent one-hot encoding, translating its categorical nature to a format amenable for algorithms.
- **Scaling**: Numerical attributes were standardized, targeting a mean of 0 and a variance of 1.

5 Clustering Analysis

The K-means algorithm was leveraged to cluster data points based on similarities. Using the silhouette score as a metric, the optimal cluster count was identified to be three. This segmentation provided a nuanced perspective of the dataset.

6 Regression Modeling & Evaluation

Our primary objective was predicting sales. Three regressors were tested: Linear Regression, Decision Tree, and Random Forest. The Random Forest model outperformed its counterparts, as evidenced by the lowest MAE and RMSE values. Its ensemble-based methodology, which amalgamates predictions from various decision trees, rendered it most effective for this dataset.

7 Results

Scatter plots comparing actual versus predicted sales were generated:

- The left plot (in blue) illustrates the default Random Forest model.
- The right plot (in green) elucidates the optimized Random Forest model.

These plots emphasized the proximity of predictions to actual sales, with the optimized Random Forest model displaying marginal improvements.

8 Hyperparameter Tuning

To further hone the model, hyperparameter tuning was conducted using GridSearchCV. This exercise yielded a marginally enhanced model.

9 Conclusion

This study accentuates the significance of structured data analysis in e-commerce. Key findings include:

- Stock levels, product prices, and advertising budgets are cardinal predictors for sales.

- The Random Forest regressor is exceptionally suited for e-commerce sales predictions.
- Periodic model retraining with updated data is essential for maintaining predictive accuracy.

10 Acknowledgments

Gratitude is extended to all contributors and facilitators of this research.