NONE

A PREPRINT

September 27, 2023

ABSTRACT

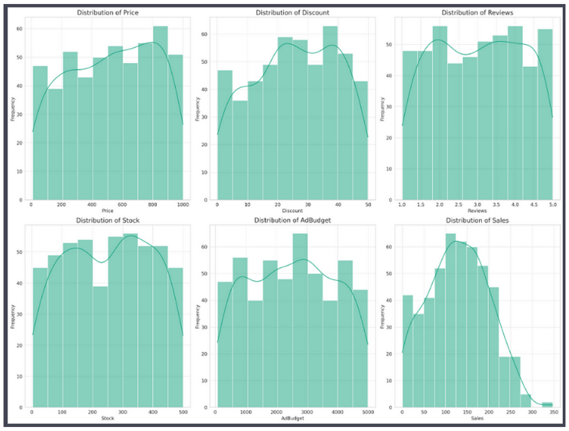
None

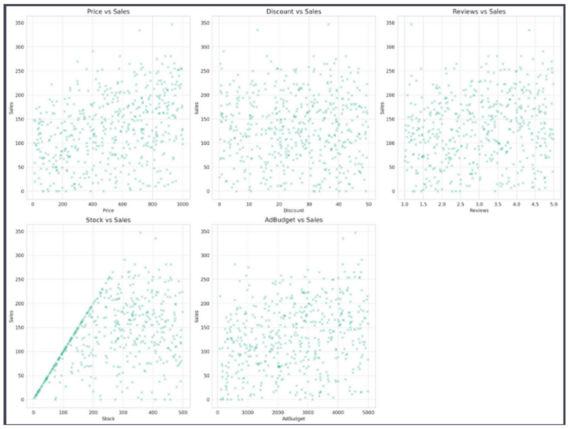
Keywords

0.0.1 Exploring E-Commerce Sales: A Data-Driven Approach

Hello, fellow data enthusiasts! Today, I'm diving deep into an intriguing e-commerce dataset to uncover insights and build predictive models. Let's get started!

Data Understanding: The dataset contains information about different products, including their prices, stock levels, and sales. We have features like ProductID, Category, Price, Discount, Reviews, Stock, AdBudget, and Sales. A quick overview shows that the dataset covers three product categories: Electronics, Home Appliances, and Clothing.





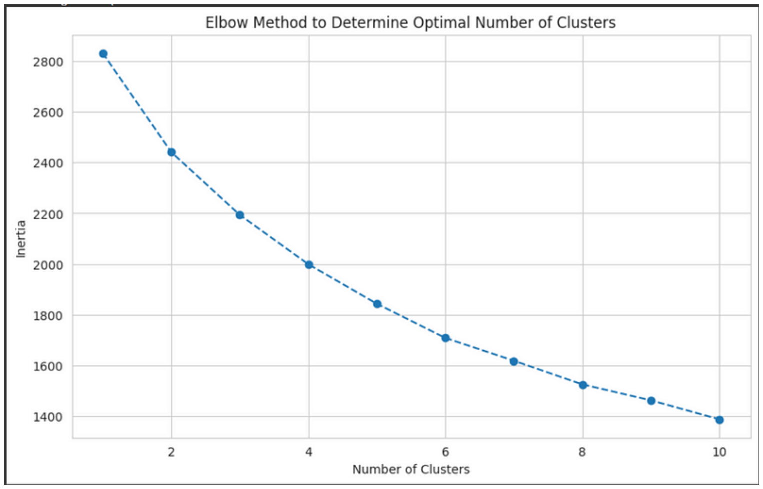
Data Visualization: A key observation from the data is the relationship between product categories and sales. Electronics and Home Appliances have a wider range of sales compared to Clothing, indicating possible higher demand or better marketing strategies.

Data Preparation: The data underwent several preprocessing steps:

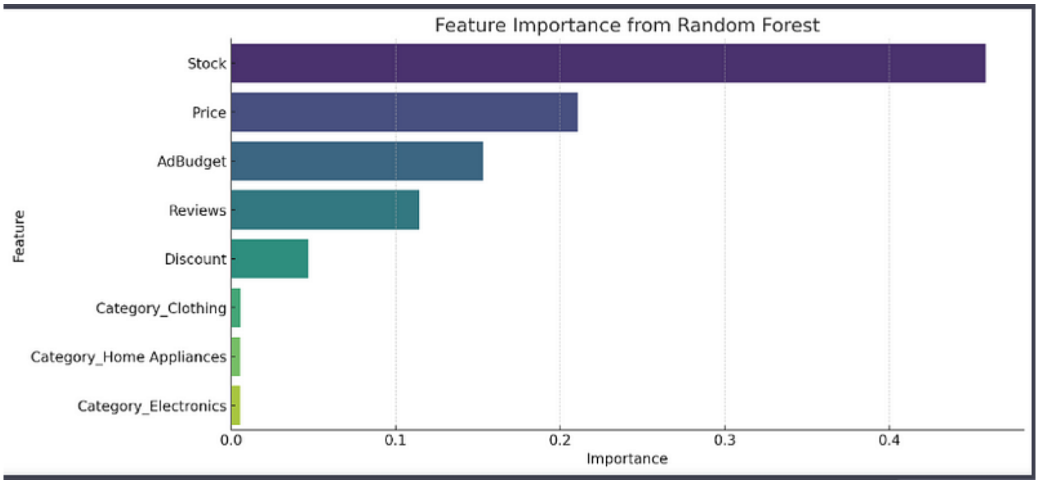
- Encoding: The 'Category' column was one-hot encoded to transform the categorical data into a format suitable for machine learning algorithms.
- Scaling: Numerical features were standardized to have a mean of 0 and variance of 1, which is crucial for algorithms sensitive to feature scales.

Data Preprocessing:
Standardization was applied to numeric columns, ensuring they're on a comparable scale. The categorical category column underwent one-hot encoding, converting it into a binary matrix.

Clustering Analysis:
K-means clustering was employed to group similar data points. A silhouette score curve identified the optimal number of clusters, which was three. This clustering provided a segmented view of the dataset.

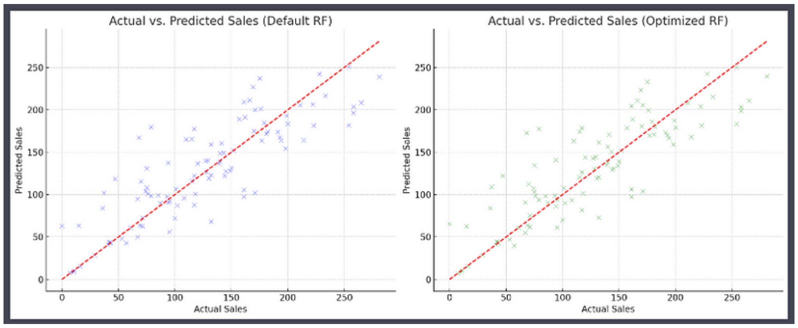


Regression Modeling & Evaluation:
The primary goal was to predict Sales. I experimented with Linear Regression, Decision Tree, and Random Forest regressors. Among these, the Random Forest model emerged victorious, boasting the lowest MAE and RMSE values.



As seen from the charts, the Random Forest Regressor stood out, achieving the best performance across all metrics. Its ensemble approach, aggregating predictions from multiple decision trees, proved effective for this dataset.

RESULTS



Here are the scatter plots comparing the actual vs. predicted sales:

- Left plot (Blue): Represents the default Random Forest model.
- Right plot (Green): Represents the optimized Random Forest model.

In both plots:

- The x-axis represents the actual sales values, while the y-axis represents the predicted sales values.
- The red dashed line represents the line of perfect prediction. Ideally, all points should lie on or close to this line.
- The closer the points are to the red line, the more accurate the predictions.

From the plots, we can observe:

- Both the default and optimized Random Forest models provide reasonably accurate predictions, as most points are close to the red line.
- The optimized model does not show a significant improvement over the default model in this visual representation, which is consistent with the earlier performance metrics.

5. Hyperparameter Tuning:
To squeeze out maximum performance, the Random Forest model underwent hyperparameter tuning using GridSearchCV. The tuning resulted in a slightly refined model, though improvements were modest.

Conclusion:
This data-driven adventure underscored the importance of a structured approach in data science. Key takeaways include:

- Stock, Price, and Adbudget are pivotal predictors for sales.
- The Random Forest model, with its inherent robustness, is apt for sales predictions in e-commerce contexts.
- Regular model retraining with fresh data ensures continued accuracy.