Predicting Customer Transaction Amounts: A Machine Learning Approach

1. Problem Statement

As part of this data science assignment, I aimed to predict whether a supermarket transaction would be above or below the average transaction value. Understanding these transaction patterns could enable more targeted marketing and promotional strategies, enhancing overall customer spending.

2. Approach: Understanding the Dataset and Inference after EDA

The dataset provided for this project includes 1000 transactions with various attributes like customer details, payment method, and product information. During the exploratory data analysis (EDA), I have made sure that there were no missing values, which facilitated smooth data processing. The analysis revealed distinct spending patterns across different product lines, which could be crucial for devising effective marketing strategies.

3. Techniques Used in the Process

I have employed several data preprocessing techniques:

- **Feature Encoding**: I have used LabelEncoder to transform categorical variables into a format suitable for model input.
- **Feature Scaling**: This was essential to normalize data, ensuring models that depend on the scale of features would function optimally.
- **Data Splitting**: Finally divided the data into an 80-20 split for training and testing, allowing us to validate our models effectively.

4. Model Performance of All the Algorithms and the Best Performance Among Them

I have performed and explored multiple models for this classification task:

- **Logistic Regression**: This model achieved an accuracy of 78%, with relatively high precision but moderate recall, indicating it was better at predicting true positives but less reliable for identifying all positive instances.
- **Decision Tree Classifier**: It had an accuracy of 72%, showing it was less effective compared to logistic regression in handling this dataset.

- **Linear Regression**: This model was not ideal for a classification task as it's typically used for regression, and thus, the performance metrics like R-squared were low.
- **Random Forest Classifier**: The best model with an accuracy of 77%, showing a balanced approach in handling both precision and recall.

5. Feature Importance and Inference

From the Random Forest model, we can learn that:

- **Quantity** and **Unit Price** are crucial predictors, implying transactions with higher quantities and more expensive items are likely to exceed the average value.
- **Product Line** influences spending, suggesting that strategic product placement and promotion could drive sales.

6. Strategies for increasing customer transaction amounts based on my findings:

1. Targeted Promotions for High-Spending Categories

Evidence:

- **High Average Transaction Values**: The 'Home and lifestyle', 'Sports and travel', and 'Health and beauty' categories have higher average transaction values than other categories. For instance, 'Home and lifestyle' averaged 336.64, 'Sports and travel' 332.07, and 'Health and beauty' 323.64, which are all above the overall average transaction value of 322.97.

2. Enhanced Marketing for Low-Spending Categories

Evidence:

- **Below Average Spending**: Categories like 'Electronic accessories', 'Food and beverages', and 'Fashion accessories' have lower average transaction values (below 322.97). This insight is crucial for targeting these categories with strategies aimed at increasing their transaction value.

3. Optimize In-Store and Online Experience

Evidence:

- **Product Popularity and Traffic**: Analysis shows that certain times of day or specific days generate higher traffic for high-spending categories. This can be correlated with sales data to optimize product placement both in-store and online.