Assignment: Data Analytics Process and Interpretation

E-commerce Customer Behavior Analysis
20020602 - L.H.K.S. Lakshan
IS4116 Business Intelligence Systems
Repository Link : BIS_Assignment

1 Introduction

The rapid growth of e-commerce has led to an increased focus on understanding customer behavior to enhance satisfaction and retention. This report analyzes a dataset containing information about customer demographics, purchasing patterns, and satisfaction levels in an e-commerce platform. The primary objective is to address the business question: **"What factors influence customer satisfaction levels?"**

1.1 Dataset Overview

The dataset used for this analysis is titled **"E-commerce Customer Behavior"** and contains **450 rows** (customers) and **11 columns** (features). It includes the following attributes:

- **Customer ID**: Unique identifier for each customer.
- **Gender**: Gender of the customer (Male/Female).
- **Age**: Age of the customer.
- **City**: City where the customer resides (e.g., New York, Los Angeles, Chicago, etc.).
- **Membership Type**: Type of membership (Gold, Silver, Bronze).
- **Total Spend**: Total amount spent by the customer.
- **Items Purchased**: Number of items purchased.
- **Average Rating**: Average rating given by the customer for their purchases.
- **Discount Applied**: Whether a discount was applied to their purchase (TRUE/FALSE).
- **Days Since Last Purchase**: Number of days since the customer's last purchase.
- **Satisfaction Level**: Customer satisfaction level (Satisfied, Neutral, Unsatisfied).

This dataset provides valuable insights about consumer behavior and helps businesses make data-driven decisions. For example:

- The **Total Spend** column can reveal spending patterns among different customer segments.
- The **Satisfaction Level** column serves as the target variable for predicting customer satisfaction.
- Categorical variables such as **City** and **Membership Type** allow for segmentation analysis to identify trends across demographics.

By applying statistical methods and machine learning techniques, this report aims to uncover key drivers of customer satisfaction and provide actionable recommendations to improve customer experiences. The insights gained from this analysis can guide businesses in tailoring their strategies to enhance customer loyalty and retention. (Dataset : E-commerce Customer Behavior)

2 Methodology

The data analysis process followed in this analysis consists of several key steps:

2.1 Data Preprocessing

The dataset was preprocessed to ensure its quality and compatibility with analytical tools. The missing values in the `Satisfaction Level` column were filled with the default value "Neutral." Categorical variables such as `Gender`, `Discount Applied`, `Membership Type`, and `City` were encoded into numerical formats using the label encoding and one-hot encoding. Additionally, irrelevant columns like `Customer ID` were removed to streamline the analysis.

2.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to understand the structure of the data set and identify patterns. Summary statistics (mean, median, standard deviation, etc.) were calculated for numerical variables, while visualizations such as count plots, box plots, and correlation heatmaps were created to explore relationships between variables. Key findings from the EDA include:

- Trends in customer spending across different cities and membership types.
- Distribution of satisfaction levels (Satisfied, Neutral, Unsatisfied) across demographic groups.
- Correlations between features such as Total Spend and Satisfaction Level .

2.3 Statistical Methods

To predict customer satisfaction levels, a Random Forest Classifier was trained using features such as `Total Spend`, `Age`, `Average Rating`, and others. The data set was divided into training sets (80%) and testing sets (20%) to evaluate the model performance. Feature importance analysis was also performed to identify the most influential factors that affect satisfaction levels.

2.4 Visualization Tools

Visualizations were created using Python libraries such as Matplotlib and Seaborn. These tools enabled the creation of clear and informative charts, including bar graphs, heatmaps, and box plots, which are included in this report to support the findings.

3 Results

3.1 Customer Satisfaction Distribution

Figure 1(a) shows the distribution of customer satisfaction levels in the three categories: **Satisfied**, **Neutral**, and **Unsatisfied**. The data is relatively balanced, with approximately 120 customers classified as Satisfied, 110 as Neutral, and 115 as unsatisfied. This balanced distribution ensures that the analysis is not skewed toward any particular category.

3.2 Correlation Heatmap

Figure 1(b) presents a correlation heatmap that reveals relationships between different variables. Notable correlations include

- A strong positive correlation ($\tilde{0}.97$) between **Total Spend** and **Items Purchased**, indicating that customers who spend more tend to purchase more items.
- A high positive correlation ($\tilde{0}.94$) between **Average Rating** and **Satisfaction Level**, confirming that higher scores lead to greater satisfaction.
- A moderate negative correlation ($\tilde{-}0.42$) between **Discount Applied** and **Days Since Last Purchase**, suggesting that discounts may delay repeat purchases.

3.3 Total Spend vs Satisfaction Level

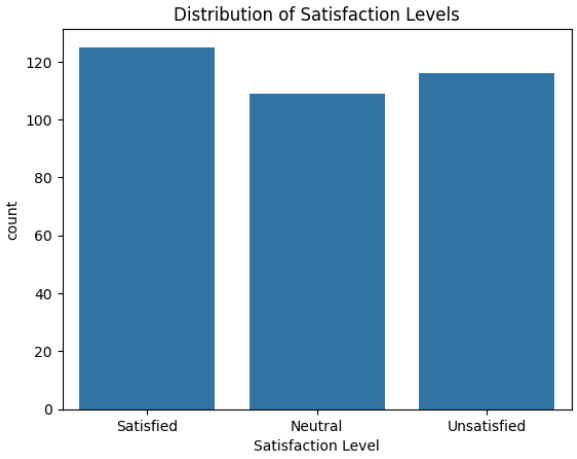
Figure 1(c) illustrates the relationship between total spending and customer satisfaction levels using a boxplot. Key observations include:

- **Satisfied customers** tend to spend significantly more, with median values around $1,200 and a range extending up to $1,500.
- **Neutral customers** exhibit moderate spending, with median values around $450.
- **Unsatisfied customers** spend the least, with median values around $600. This suggests that higher spending is strongly associated with higher satisfaction levels.

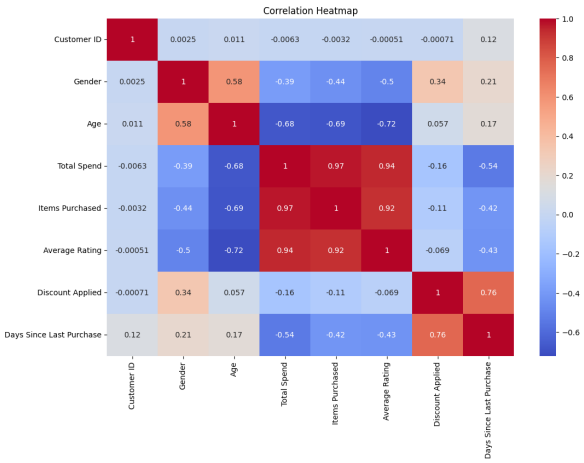
3.4 Feature Importance

Figure 1(d) highlights the most important features influencing customer satisfaction based on the Random Forest model. The top factors are:

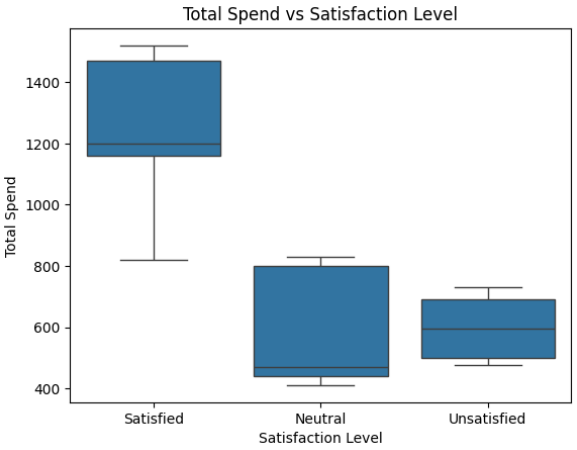
- **Days Since Last Purchase**: Recent activity plays a crucial role in determining satisfaction.
- **Total Spend**: Higher spending is a significant predictor of satisfaction.
- **Items Purchased**: The number of items purchased also contributes to satisfaction.



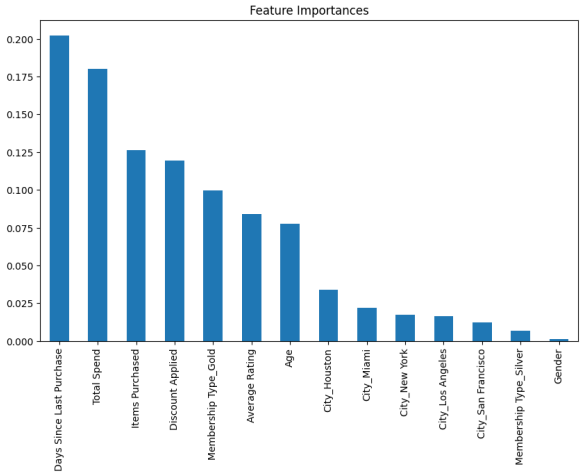
(a) Distribution of Satisfaction Levels



(b) Correlation Heatmap



(c) Boxplot: Total Spend vs Satisfaction Level



(d) Feature Importance Bar Chart

Figure 1: Key visualizations from the analysis. (a) Distribution of satisfaction levels, (b) Correlation heatmap, (c) Boxplot showing total spend vs satisfaction level, and (d) Feature importance bar chart.