Data Mining for Strategic Business Insights: A Comprehensive Analysis

Bereket Gebremariam

University of the Cumberlands

MSCS634

Professor Penmasta

8/10/2025

# Project Summary

This project involved a comprehensive data mining analysis of a customer behavior dataset to uncover strategic business insights. The process followed a structured approach involving data preparation, predictive modeling, and pattern discovery. The dataset was cleaned, engineered, and modeled to predict customer spending, identify high-value customers, and uncover meaningful behavioral patterns.

# Dataset Selection and Preprocessing

Dataset Used. The dataset consisted of customer behavior records with variables including Customer ID, Gender, Age, Membership Type, Total Spend, Items Purchased, and Satisfaction Level. This dataset was chosen for its rich combination of demographic, behavioral, and transactional data, which provided an ideal basis for applying diverse data mining techniques to solve a practical business problem.  
  
Preprocessing and Feature Engineering. Several steps were undertaken to prepare the dataset:  
- Column Cleaning: Column names containing spaces were standardized (e.g., Membership Type became Membership\_Type) to prevent coding errors.  
- Missing Data: Missing values in Membership\_Type were imputed with the mode, and records with missing Customer\_ID were removed.  
- Feature Engineering: A new feature, Shopping\_Frequency, was derived from Days\_Since\_Last\_Purchase to create a more intuitive measure of customer engagement.

# Predictive Modeling Results

Regression Analysis. The primary objective was to predict Total\_Spend. Two models—Multiple Linear Regression and Ridge Regression—were evaluated using R-squared (R²), Mean Squared Error (MSE), and cross-validation. Both models performed well, with Ridge Regression slightly outperforming Multiple Linear Regression (R² ≈ 0.78). The results indicated that Items\_Purchased, Discount\_Applied, and the engineered Shopping\_Frequency were the most influential factors in predicting customer spending.  
  
Classification Analysis. The goal was to classify customers as High\_Spender or Low\_Spender based on a spending threshold. Two models were developed: Decision Tree and k-Nearest Neighbors (k-NN). Hyperparameter tuning with GridSearchCV significantly improved the Decision Tree model’s performance. The tuned model demonstrated high accuracy and F1-score, as confirmed by its confusion matrix and ROC curve. The analysis revealed that high-value customers typically purchased more items and had higher Average\_Rating values.

# Pattern Discovery Results

Clustering Analysis. K-Means clustering segmented the customer base into three distinct clusters based on numerical features such as Total\_Spend and Items\_Purchased:  
1. High-Value Customers: High average spending, frequent purchases, and a high proportion of Premium members.  
2. Core Customers: Representing the majority, with moderate spending and purchase frequency.  
3. Low-Engagement Customers: Low average spending and infrequent purchases, often new or less engaged customers.  
A scatter plot of Total\_Spend versus Items\_Purchased visually confirmed clear separation between these clusters.  
  
Association Rule Mining. Using the Apriori algorithm, relationships between categorical features were identified. For example, the rule {City=New York} → {Membership\_Type=Premium} indicated that customers in New York were more likely to have premium memberships. This insight can guide targeted marketing and regional sales initiatives.

# Practical Recommendations

Based on the findings, the following recommendations were developed:  
1. Targeted Marketing: Utilize the classification model to proactively identify likely high spenders and offer them exclusive promotions or loyalty programs.  
2. Customer Retention: Tailor marketing strategies to each cluster, with personalized perks for high-value customers and introductory offers for low-engagement customers.  
3. Strategic Discounts: Implement tiered discount strategies, providing greater discounts to high-value customers and smaller introductory offers to new customers.  
4. Regional Insights: Use association rules to inform geo-targeted campaigns, such as promoting premium membership upgrades in cities with strong correlations.

# Ethical Considerations

Ethical considerations were prioritized throughout the analysis:  
- Data Privacy: No personally identifiable information beyond Customer ID was used. Analysis focused on aggregated behavioral patterns.  
- Fairness and Bias: The dataset was reviewed for bias, particularly in demographic attributes such as gender and age. Behavioral factors, rather than protected characteristics, guided segmentation and recommendations.  
- Transparency: The entire workflow—from preprocessing to model evaluation—was documented in a Jupyter Notebook. Interpretable models like Decision Trees were used to ensure the decision-making process was understandable and to support bias detection.

# Appendix

This appendix includes the output of the jupyter notebook code.

--- First 5 rows of the dataset ---

Customer\_ID Gender Age City Membership\_Type Total\_Spend \

0 101 Female 29 New York Gold 1120.20

1 102 Male 34 Los Angeles Silver 780.50

2 103 Female 43 Chicago Bronze 510.75

3 104 Male 30 San Francisco Gold 1480.30

4 105 Male 27 Miami Silver 720.40

Items\_Purchased Average\_Rating Discount\_Applied \

0 14 4.6 True

1 11 4.1 False

2 9 3.4 True

3 19 4.7 False

4 13 4.0 True

Days\_Since\_Last\_Purchase Satisfaction\_Level

0 25 Satisfied

1 18 Neutral

2 42 Unsatisfied

3 12 Satisfied

4 55 Unsatisfied

--- Dataset Info ---

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 350 entries, 0 to 349

Data columns (total 11 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Customer\_ID 350 non-null int64

1 Gender 350 non-null object

2 Age 350 non-null int64

3 City 350 non-null object

4 Membership\_Type 350 non-null object

5 Total\_Spend 350 non-null float64

6 Items\_Purchased 350 non-null int64

7 Average\_Rating 350 non-null float64

8 Discount\_Applied 350 non-null bool

9 Days\_Since\_Last\_Purchase 350 non-null int64

10 Satisfaction\_Level 348 non-null object

dtypes: bool(1), float64(2), int64(4), object(4)

memory usage: 27.8+ KB

--- Descriptive Statistics ---

Customer\_ID Age Total\_Spend Items\_Purchased Average\_Rating \

count 350.000000 350.000000 350.000000 350.000000 350.000000

mean 275.500000 33.597143 845.381714 12.600000 4.019143

std 101.180532 4.870882 362.058695 4.155984 0.580539

min 101.000000 26.000000 410.800000 7.000000 3.000000

25% 188.250000 30.000000 502.000000 9.000000 3.500000

50% 275.500000 32.500000 775.200000 12.000000 4.100000

75% 362.750000 37.000000 1160.600000 15.000000 4.500000

max 450.000000 43.000000 1520.100000 21.000000 4.900000

Days\_Since\_Last\_Purchase

count 350.000000

mean 26.588571

std 13.440813

min 9.000000

25% 15.000000

50% 23.000000

75% 38.000000

max 63.000000

--- Handling Missing Values ---

Missing values before cleaning:

Customer\_ID 0

Gender 0

Age 0

City 0

Membership\_Type 0

Total\_Spend 0

Items\_Purchased 0

Average\_Rating 0

Discount\_Applied 0

Days\_Since\_Last\_Purchase 0

Satisfaction\_Level 2

dtype: int64

Missing values in 'Membership\_Type' imputed with mode: 'Gold'

Rows with missing 'Customer\_ID' have been dropped.

Missing values after cleaning:

Customer\_ID 0

Gender 0

Age 0

City 0

Membership\_Type 0

Total\_Spend 0

Items\_Purchased 0

Average\_Rating 0

Discount\_Applied 0

Days\_Since\_Last\_Purchase 0

Satisfaction\_Level 2

dtype: int64

--- Exploratory Data Analysis (EDA) ---

A graph of blue and white bars

AI-generated content may be incorrect.

A graph of blue rectangular bars

AI-generated content may be incorrect.

A red and blue squares with white text

AI-generated content may be incorrect.

A blue rectangular object with black lines

AI-generated content may be incorrect.