Introduction to Data Science

Project Report

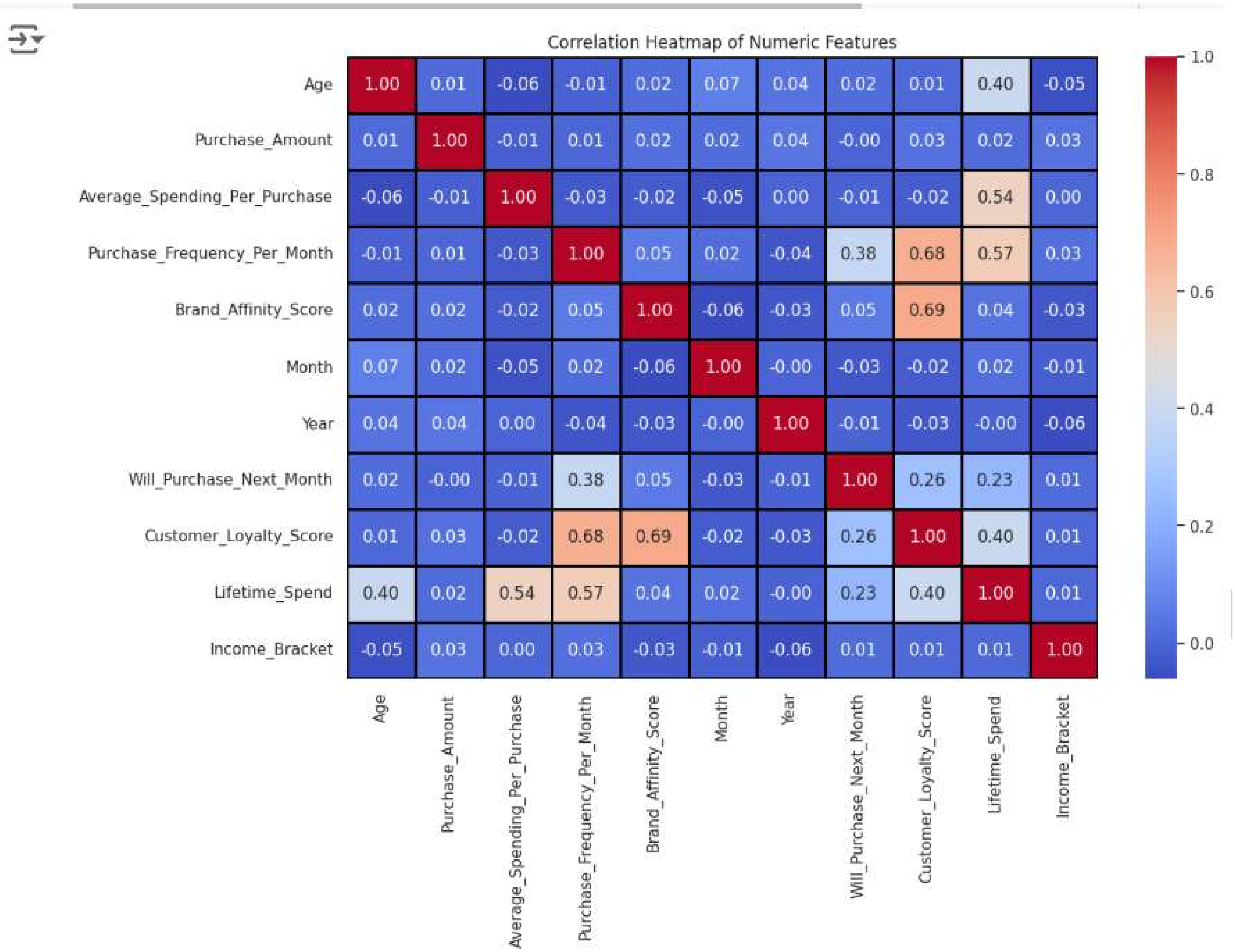
23i-2622 23i-2628

Imtiaz Mall Electronics Section Analysis

Introduction

Imtiaz Mall, a renowned department store chain, is facing declining sales and a significant number of non-recurring customers in its electronics section. To tackle this challenge, this project aims to conduct a comprehensive data analysis and develop strategies for customer retention and sales growth.

Investigating Correlations



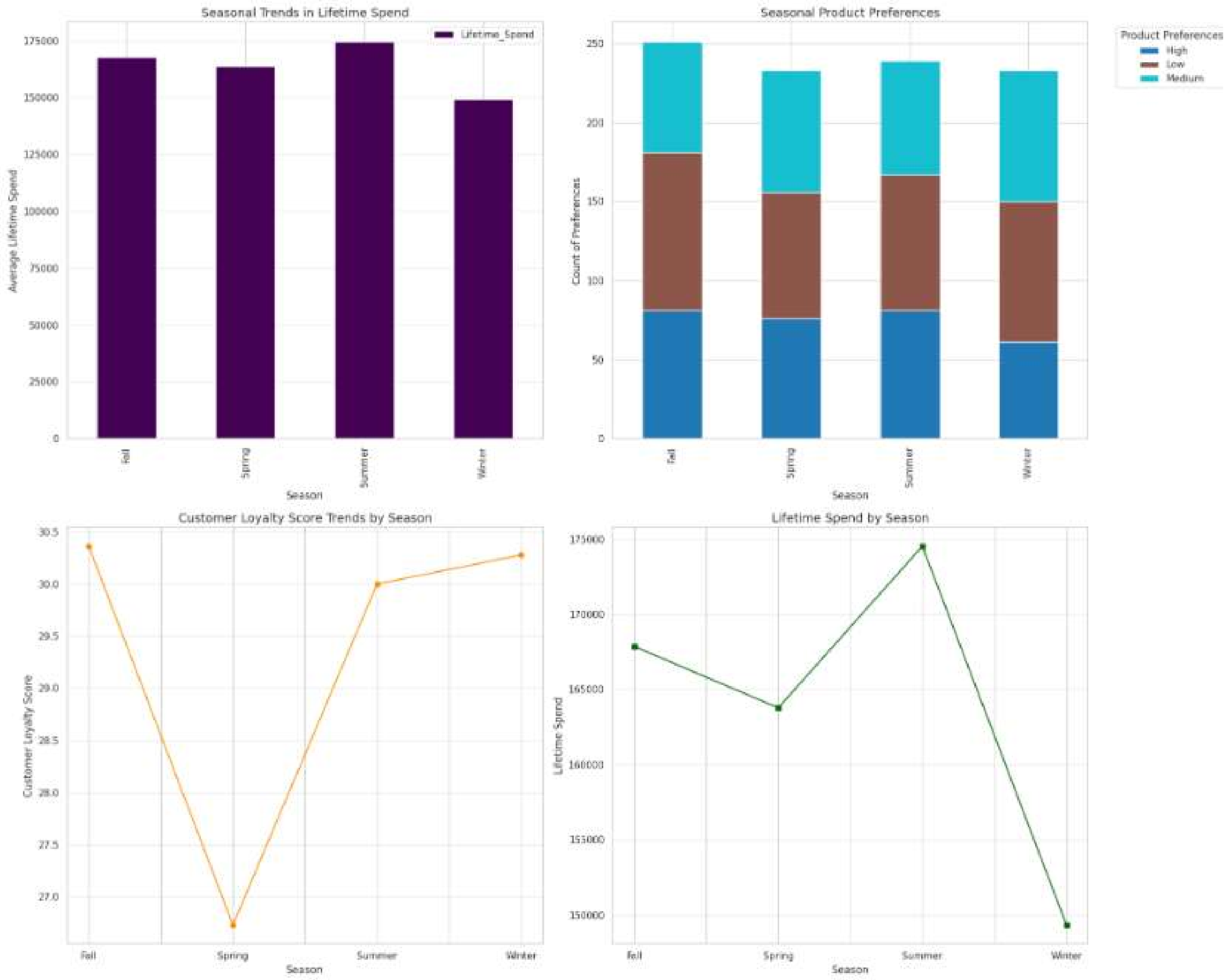
	Age	Purchase_Amount	Average_Spending_Per_Purchase	Purchase_Frequency_Per_Month	Brand_Affinity_Score	Month	Year	Will_Purchase_Next_Month	Customer_Loyalty_Score	Lifetime_Spend	Income_Bracket
Age	1.00	0.01	-0.06	-0.01	0.02	0.07	0.04	0.02	0.01	0.40	-0.05
Purchase_Amount	0.01	1.00	-0.01	0.01	0.02	0.02	0.04	-0.00	0.03	0.02	0.03
Average_Spending_Per_Purchase	-0.06	-0.01	1.00	-0.03	-0.02	-0.05	0.00	-0.01	-0.02	0.54	0.00
Purchase_Frequency_Per_Month	-0.01	0.01	-0.03	1.00	0.05	0.02	-0.04	0.38	0.68	0.57	0.03
Brand_Affinity_Score	0.02	0.02	-0.02	0.05	1.00	-0.06	-0.03	0.05	0.69	0.04	-0.03
Month	0.07	0.02	-0.05	0.02	-0.06	1.00	-0.00	-0.03	-0.02	0.02	-0.01
Year	0.04	0.04	0.00	-0.04	-0.03	-0.00	1.00	-0.01	-0.03	-0.00	-0.06
Will_Purchase_Next_Month	0.02	-0.00	-0.01	0.38	0.05	-0.03	-0.01	1.00	0.26	0.23	0.01
Customer_Loyalty_Score	0.01	0.03	-0.02	0.68	0.69	-0.02	-0.03	0.26	1.00	0.40	0.01
Lifetime_Spend	0.40	0.02	0.54	0.57	0.04	0.02	-0.00	0.23	0.40	1.00	0.01
Income_Bracket	-0.05	0.03	0.00	0.03	-0.03	-0.01	-0.06	0.01	0.01	0.01	1.00

The correlation heatmap provides insights into the relationships between various numerical features in the dataset. Key findings include:

1. **Customer Loyalty and Purchase Behavior:**
    a. *Customer Loyalty Score* exhibits a strong positive correlation with *Purchase Frequency Per Month* (0.68) and *Will Purchase Next Month* (0.26). This indicates that loyal customers are more likely to make frequent purchases and continue purchasing in the future.
2. **Lifetime Spend and Spending Patterns:**
    a. *Lifetime Spend* correlates positively with *Average Spending Per Purchase* (0.54) and *Purchase Frequency Per Month* (0.57). Customers who spend more per transaction and shop more often contribute significantly to overall spending.
3. **Age and Lifetime Spend:**
    a. A moderate positive correlation (0.40) exists between *Age* and *Lifetime Spend*, suggesting older customers tend to spend more over time.
4. **Weak Relationships:**
    a. Features like *Income Bracket* and *Purchase Amount* have minimal correlations with most other variables, indicating a weaker direct influence on overall customer behavior.

These correlations highlight actionable relationships, such as targeting high -loyalty customers and understanding spending habits to tailor retention strategies.

Seasonal Variations in Customer Behavior



1. Seasonal Trends in Lifetime Spend

- Customers tend to spend the most during *Summer* and the least during *Winter*. This suggests a seasonal peak in activity that can inform promotional strategies during off-peak seasons.

2. Seasonal Product Preferences

- Product preferences remain consistent across seasons, with a higher count of *Medium*-preferred products followed by *High* and *Low*. This consistency implies that promotions and marketing efforts should cater to these preferences year-round.

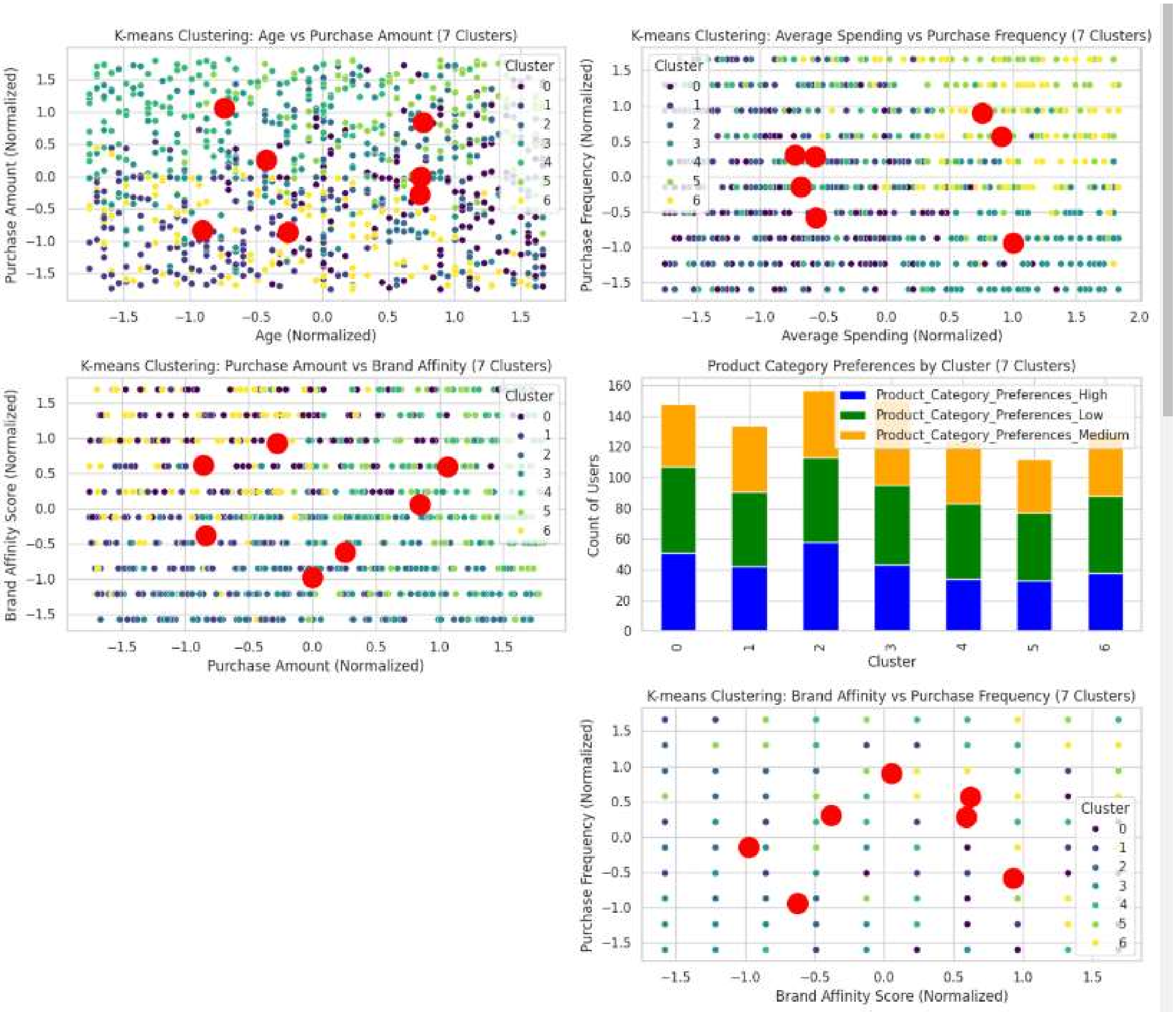
3. Customer Loyalty Score Trends by Season

- Loyalty scores drop significantly in *Spring* but recover during *Summer* and *Winter*. This indicates potential opportunities to improve engagement strategies during Spring to retain customer loyalty.

4. Lifetime Spend by Season

- Lifetime spend trends align closely with loyalty patterns, peaking in *Summer* and declining in *Winter*. These trends emphasize the importance of enhancing customer retention efforts during the low-spending seasons.

Clustering Analysis



1. Differences Between Clusters

- **Age vs. Purchase Amount**:
  - Some clusters (e.g., Cluster 0) represent users with **higher purchase amounts** but span across a wide range of ages, suggesting that age may not directly correlate with spending.
  - Clusters like Cluster 2 exhibit **lower purchase amounts**, indicating a different spending behavior or budget limitation.

- **Average Spending vs. Purchase Frequency**:
  - Clusters differ in spending behavior. For example:
    - Cluster 3 shows **higher average spending** but **lower purchase frequency**, indicating occasional large purchases.
    - Cluster 6 has **low average spending** but **higher purchase frequency**, suggesting frequent but smaller transactions.
- **Purchase Amount vs. Brand Affinity**:
  - Clusters like Cluster 4 display **high brand affinity** and **high purchase amounts**, possibly indicating loyalty to premium brands.
  - In contrast, Cluster 1 shows **low brand affinity** despite moderate purchase amounts, suggesting a preference for non-branded or generic products.
- **Product Category Preferences**:
  - Some clusters have a strong preference for specific categories:
    - Cluster 5 has a high proportion of users with "High" product category preferences.
    - Cluster 2 has more users with "Low" preferences, indicating they are less likely to explore diverse product categories.

2. Similarities Between Clusters

- **Purchase Frequency Patterns**:
  - Clusters such as Cluster 0 and Cluster 1 exhibit **medium purchase frequency**, showing stable shopping behaviors regardless of other differences like brand affinity or spending.
- **Age Group Distribution**:
  - Several clusters span similar ranges of age, indicating that user age is not always a determining factor for specific spending habits (e.g., clusters 3 and 4).
- **Brand Affinity Overlap**:
  - Some clusters (e.g., Cluster 1 and Cluster 2) show overlap in brand affinity scores, indicating shared preferences for brands despite differences in purchase frequency or spending.

Key Takeaways:

1. **High-Spending Clusters** (e.g., Cluster 3) differ significantly from **low-spending clusters** (e.g., Cluster 2), both in terms of purchase frequency and product preferences.
2. **Brand Affinity** plays a crucial role in defining customer behavior, with some clusters (e.g., Cluster 4) heavily influenced by it.
3. Certain clusters (e.g., Cluster 1) exhibit mixed behaviors, with no strong affinity or extreme values, representing general shoppers.
4. Overlapping age ranges and frequency patterns highlight **behavioral clustering** rather than demographic-based clustering.

MODULE 5

Linear Regression

- **Performance Metrics**:
  - Mean Absolute Error (MAE): **23.20**
  - Mean Squared Error (MSE): **742.73**
  - R-squared (R^2): **-0.01** (indicates poor fit; the model fails to explain variance in the data effectively).
- **Key Observations**:
  - The negative R^2 value suggests that the linear regression model is performing worse than a simple mean-based model.
  - Indicates that customer purchase behavior may not have a linear relationship with the independent features.
- **Strengths**:
  - Provides basic insights into the linear relationships between features.
  - Useful for understanding directional effects of predictors.
- **Limitations**:
  - Fails to capture non-linear and complex patterns in customer behavior.
  - Poor accuracy for predictive purposes.

Decision Tree Classifier

- **Performance Metrics**:
  - Accuracy: **0.92**
  - Precision: **0.95**
  - Recall: **0.96**
  - F1 Score: **0.95**
- **Key Observations**:
  - The decision tree performs exceptionally well, especially in predicting customers who are likely to purchase (class 1).
  - High precision and recall demonstrate that the model minimizes false positives and false negatives.
  - The classification report indicates some imbalance in predicting non-purchasers (class 0), but overall weighted performance is strong.
- **Strengths**:
  - Highly interpretable and actionable.
  - Captures non-linear patterns and interactions between features effectively.
  - Accurate in identifying purchasing behavior based on features like brand affinity, income level, and frequency.
- **Limitations**:
  - Risk of overfitting if not properly tuned (e.g., tree depth).
  - Sensitive to noise in the data.

K-Means Clustering

- **Performance Summary**:
  - Identified **7 clusters** with distinct customer behaviors.
  - Significant findings include:
    - **Cluster 4**: High-spending customers with strong brand affinity.
    - **Cluster 6**: Value-conscious customers with frequent but low purchases.
    - **Cluster 1**: Moderate spenders with low brand loyalty.
- **Strengths**:
  - Effective segmentation of customers into actionable groups.
  - Provides strategic insights for targeted marketing and product recommendations.
- **Limitations**:
  - Requires predefined cluster count (optimal K selection).
  - Interpretation of clusters is subjective and may require domain expertise.

1. Comparison of Predictive Models

1.1 Regression Analysis

- **Strengths**:
  - Ideal for identifying and quantifying linear relationships between customer features (e.g., age, spending habits) and target variables like purchase amounts.
  - Provides interpretable coefficients that indicate the direction and magnitude of influence of features.
- **Limitations**:
  - Limited in handling non-linear relationships, which are common in customer behavior.
  - Prone to overfitting if the data has multicollinearity or too many irrelevant features.
- **Applicability**:
  - Useful for predicting continuous outcomes, such as estimating average customer spending.

1.2 Decision Tree Classifier

- **Strengths**:
  - Handles non-linear relationships well, making it suitable for complex customer segmentation.
  - Easy to interpret with clear decision rules for actionable insights.
  - Can model interactions between features effectively.
- **Limitations**:
  - May overfit if not pruned or tuned properly, leading to poor generalization.
  - Sensitive to small changes in data, resulting in different splits.

- **Applicability**:
    - Well-suited for classification problems like predicting whether a customer will make a purchase (binary or multiclass).

1.3 K-Means Clustering

- **Strengths**:
    - Effective for unsupervised learning, identifying customer segments without predefined labels.
    - Captures similarities and differences in customer behavior across multiple dimensions.
    - Simple and efficient to implement and scale for large datasets.
- **Limitations**:
    - Assumes spherical clusters, which may not represent real-world data distributions.
    - Requires pre-specification of the number of clusters, which can be subjective.
- **Applicability**:
    - Useful for customer segmentation, providing insights into different user groups and their characteristics.

2. Recommendations for the Electronics Section

2.1 Marketing and Personalization

- **High-Spending, High-Brand-Affinity Customers (Cluster 4)**:
    - Focus on premium electronics and offer loyalty rewards or early access to high-end products.
    - Create targeted campaigns highlighting brand-specific deals and exclusive items.
- **Frequent Purchasers with Low Spending (Cluster 6)**:
    - Bundle offers or discounts on bulk purchases to increase transaction value.
    - Promote cost-effective products and emphasize value-driven messaging.
- **Low Brand Affinity, Moderate Spending Customers (Cluster 1)**:
    - Focus on educational campaigns emphasizing product quality to build brand loyalty.
    - Offer comparisons to competitors to influence brand-affinity-driven purchases.

2.2 Inventory Management

- Align product assortment with clusters showing preferences for specific product categories (e.g., clusters with "high product category preferences").
- Stock high-demand items based on predictive models of purchase frequency and average spending.

2.3 Customer Retention

- Use decision tree insights to identify key factors driving purchase decisions (e.g., product availability, price range).
- Develop personalized follow-ups, such as product recommendations, based on cluster profiles.