Technical Report

Personalized Marketing through Customer Segmentation

Data Storm 5.0 – Semi Final Round

- Team Name: 3_Amigos

- Team Members:

Pairavi Thanancheyan

Sangaran Thevarasa

Nithursika Kalanantharasan

- GitHub Repository Link: https://github.com/Pairavi/Data-Storm-5.0

1. Problem Description

The company has identified that conventional marketing strategies have been ineffective in engaging their current customer base over the past few years. To address this, they plan to adopt a personalized marketing strategy tailored to individual customer preferences. The primary objective is to enhance marketing strategies by identifying high-end customers, categorized into three groups: Premium, Loyal, and Frequent, while also considering average and low-performing customers.

To support this initiative, the company has provided historical customer data, including monthly spend, visit frequency, and basket size. The goal is to develop a rule-based approach to identify high-end customers by utilizing clustering techniques to analyze the provided data. This analysis will help establish a clear framework based on percentile rankings for classifying customers into the defined segments.

The company aims to define target customer segments and create effective intervention strategies for each segment. The predefined promotion goals include increasing customer retention, customer spend, and visit frequency. By analyzing each segment and applying a percentile approach to the given variables, the company seeks to develop and deliver targeted mass promotions tailored to each high-end customer segment.

2. Data Preprocessing

2.1 Handling Missing Values

- **Detection and Initial Analysis:**

  A thorough examination revealed significant missing values in key columns: average_monthly_spend**,** average_monthly_visit_frequency**,** and average_monthly_basket_size**.**

The counts of missing values were as follows:

➔ average_monthly_spend: 9386 missing values
➔ average_monthly_visit_frequency: 9328 missing values
➔ average_monthly_basket_size: 9286 missing values

- **Imputation Techniques:**

1. Categorical to Numeric Conversion:

Some values were represented as non-numeric text. These were converted to numeric values using predefined mappings. For instance:

➔ 'one point two' was mapped to 1.2
➔ 'nine hundred' was mapped to 900

This step ensured that all values were in a numeric format suitable for further analysis. Additionally, the columns were converted to the appropriate data types to facilitate numerical operations and analysis.

```
freequency_mapping = {
    'one point two': 1.2,
    'nine point five': 9.5,
    'two': 2,
    'twenty two': 22,
    'three point four': 3.4
}

df['average_monthly_visit_frequency'] =
df['average_monthly_visit_frequency'].replace(freequency_mapping)
```

```
spend_mapping = {
    'nine hundred': 900,
}

df['average_monthly_spend'] =
df['average_monthly_spend'].replace(spend_mapping)
```

2. Regression-Based Imputation:

- **Dataset Preparation:**

The dataset was divided into four subsets:

- ➔ Train dataset: Rows where none of the three columns (average_monthly_spend, average_monthly_visit_frequency, average_monthly_basket_size) had missing values.
- ➔ Test dataset 1: Rows where average_monthly_spend was missing but the other two columns had values.
- ➔ Test dataset 2: Rows where average_monthly_visit_frequency was missing but the other two columns had values.
- ➔ Test dataset 3: Rows where average_monthly_basket_size was missing but the other two columns had values.

It was confirmed that there were no rows with more than one missing value, simplifying the imputation process.

- **Linear Regression Models:**

Separate linear regression models were trained for each missing value scenario:

- ➔ For average_monthly_spend, a model was trained using average_monthly_visit_frequency and average_monthly_basket_size.
- ➔ For average_monthly_visit_frequency, a model was trained using average_monthly_spend and average_monthly_basket_size.
- ➔ For average_monthly_basket_size, a model was trained using average_monthly_spend and average_monthly_visit_frequency.

Before putting these datasets into the models, Min-Max Scaling was applied to ensure all features were on a similar scale.

- **Imputation Process:**

- ➔ Predictions were made for the missing average_monthly_spend in Test dataset 1 using the trained model, and the values were filled in.
- ➔ Predictions were made for the missing average_monthly_visit_frequency in Test dataset 2 using the trained model, and the values were filled in.
- ➔ Predictions were made for the missing average_monthly_basket_size in Test dataset 3 using the trained model, and the values were filled in.

After model prediction, some rows contained negative values. These negative values were replaced with their absolute values to maintain data integrity.

- **Final Dataset Compilation:**

The filled Test datasets were combined with the original Train dataset to reconstruct the complete dataset, restoring the initial row count and ensuring no missing values.

2.2 Handling Duplicates

- **Detection and Removal:**

Duplicate records were identified and removed based on the customer_id field. This step was crucial to ensure that the dataset was free from redundant entries that could skew the analysis and affect the accuracy of the clustering results.

2.3 Handling Outliers

- **IQR Method for Outlier Detection:**

Outliers were detected using the Interquartile Range (IQR) method. This method identifies outliers as values below the first quartile (Q1) - 1.5IQR or above the third quartile (Q3) + 1.5IQR. This approach robustly identifies extreme values without being affected by the overall distribution shape.

- **Capping Outliers:**

Instead of removing outliers, they were capped at the 1st and 99th percentiles. This method ensures that extreme values do not disproportionately influence the clustering process while retaining as much data as possible. Capping outliers rather than removing them preserves data volume and structure, ensuring a balanced dataset for analysis.

Outlier Counts:

➔ average_monthly_spend: 8 outliers removed
➔ average_monthly_visit_frequency: 3 outliers removed
➔ average_monthly_basket_size: 4 outliers removed

3. Feature Scaling and Standardization

- **Min-Max Scaling:**

Min-Max Scaling was applied to normalize the data, transforming features to a fixed range, typically [0, 1]. This scaling technique was chosen for its ability to maintain the relationships between variables while ensuring that all features contribute equally to the clustering process.

- **Impact on Model Performance:**

➔ Uniform Contribution to Distance Metrics:

For clustering algorithms like K-Means, which rely on distance calculations, feature scaling ensures that all features contribute equally to the distance metrics. Without scaling, features with larger ranges would dominate the distance calculation, leading to biased clustering results.

➔ Enhanced Convergence Speed:

Scaled features lead to faster convergence in algorithms like K-Means. When features are on a similar scale, the algorithm can more quickly find the optimal cluster centers, improving computational efficiency.

➔ Data Preparation Consistency:

Consistent application of scaling across the dataset ensures that model performance is not adversely affected by differing feature scales during prediction or further analysis. This consistency is crucial for reliable and reproducible results.

4. Challenges Faced During Model Training

- **Handling Missing Values**

In our dataset, we observed that only 3% of the rows had missing values. Traditionally, in machine learning, we might ignore such a small percentage of missing data, considering it negligible. However, given the business context, where each data point can significantly impact the clustering and subsequent segmentation of high-end customers, we decided to impute these missing values.

Before proceeding with imputation, we conducted a thorough check to identify any rows where two out of the three key fields (monthly spend, monthly visit frequency, and monthly basket size) had missing values. Fortunately, we did not encounter any such cases. This

allowed us to confidently impute the missing values for individual fields without worrying about the compounded effect of multiple missing fields in the same row.

- **Feature Scaling**

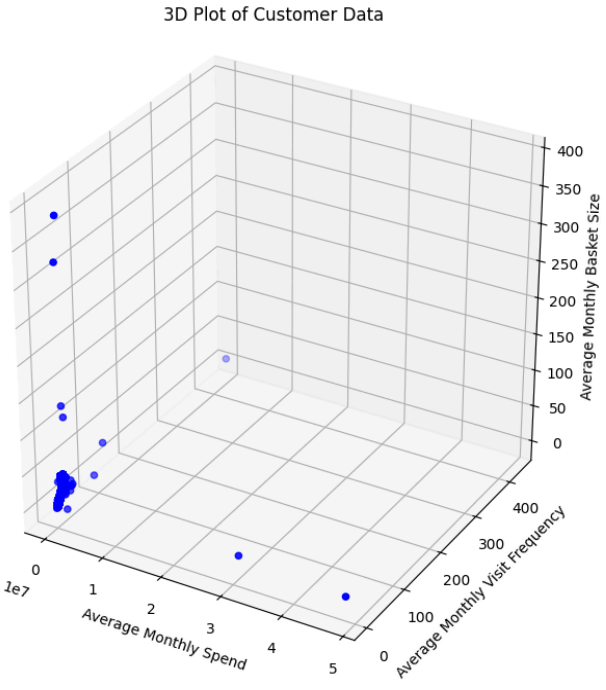
During the preprocessing stage, we faced a challenge in selecting the most suitable scaling method for our features. To make an informed decision, we examined the distribution of each column (monthly spend, monthly visit frequency, and monthly basket size) in the dataset.

Our analysis revealed that none of these features followed a normal distribution. Given this observation, we opted for the Min-Max scaling method. The rationale behind this choice was to transform the data into a uniform range between [0, 1], which is particularly effective for algorithms like K-means clustering and DBSCAN algorithm that are sensitive to the scale of the data.

- **Dealing with Outliers**

Outliers can significantly impact clustering results, especially when dealing with customer segmentation in a retail context. We observed a small number of customers (fewer than 10) with substantial deviations in their data, which could distort the clustering process. However, we also recognized that these outliers might represent valuable high-spending or frequent customers, making it crucial to handle them appropriately.

Visualizing the Data

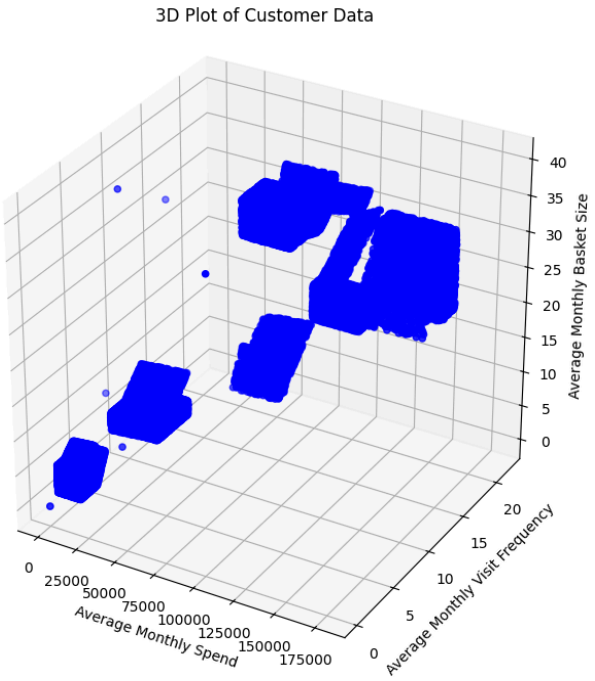


We visualized the data in a 3D plane using the three key features: monthly spend, monthly visit frequency, and monthly basket size. This step helped us understand the distribution and identify the presence of outliers visually

By visualizing the data, we assessed the impact of the outliers on the overall clustering process. This step ensured that we understood whether the outliers were skewing the clustering results or if they were forming distinct clusters on their own.

Removal of Outliers: After careful consideration, we opted to remove the extreme outliers to ensure the integrity of the clustering results. This decision was made because the outliers were significantly skewing the clustering process and not forming meaningful clusters.

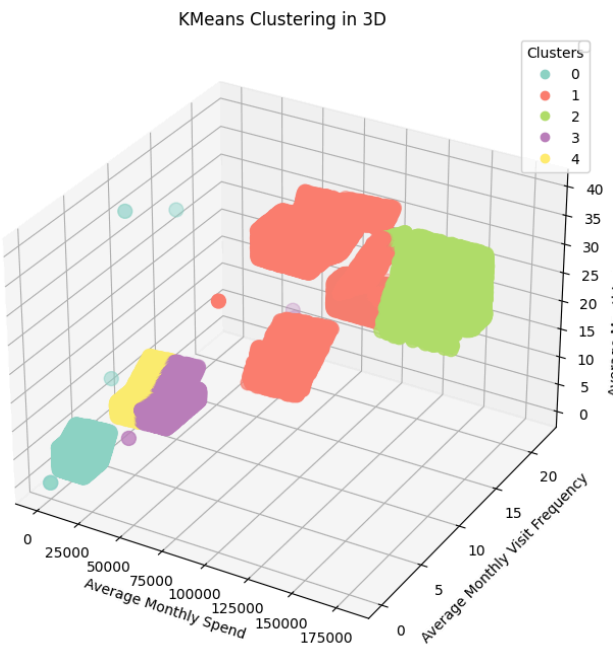
Resulting Visualization



- **Algorithm Selection**

Selecting the appropriate clustering algorithm was a significant challenge due to the different types of clustering results produced by each algorithm. Initially, we experimented with the K-Means clustering algorithm, but the results did not meet our requirements. Here's a detailed account of our approach and the rationale behind our final algorithm choice:

Initial Attempt with K-Means



We started with K-Means, a popular clustering algorithm that partitions data into K clusters based on minimizing the variance within each cluster.

After applying K-Means, we visualized the clustering results. However, the clusters formed by K-Means did not align well with our business requirements

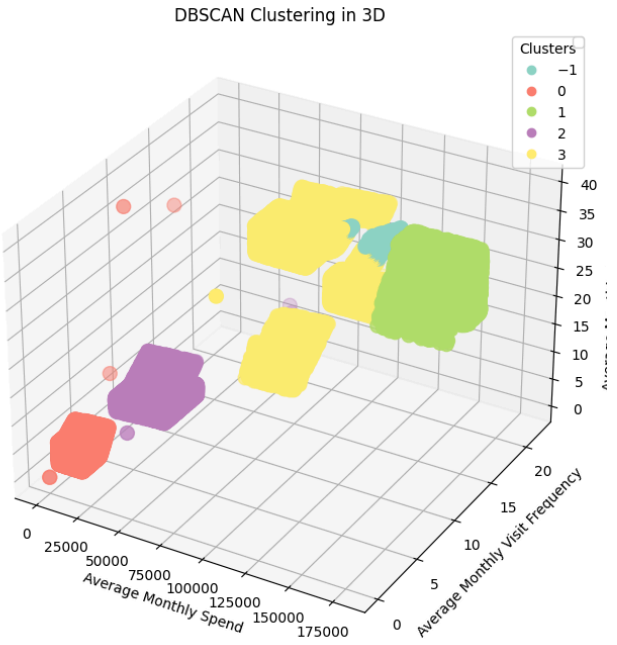
After further analysis and internal discussions, we recognized that our data exhibited density-based characteristics, with regions of varying point densities and some noise

Given these observations, we decided to explore a density-based clustering algorithm that could better handle these data properties.

Switch to DBSCAN

We moved to the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm, which is well-suited for datasets with varying densities and can identify clusters of arbitrary shapes.

After applying DBSCAN, we visualized the clustering results. The clusters formed were more meaningful than those produced by the K-Means clustering algorithm and aligned well with the natural data groupings and our business requirements.



4. Clustering Analysis and Customer Segment Interpretation

- **Cluster Naming and Description**

Based on the DBSCAN clustering analysis, the clusters were identified and named according to their spending habits, visit frequencies, and basket sizes as follows. To get an accurate range for all three columns of each cluster, some exceptional few points were removed:

- **Cluster 1**: High-End - Premium Customers
  - ➔ Monthly Spend: 140,000.0 and above
  - ➔ Monthly Visit Frequency: 11.17 - 15.99
  - ➔ Monthly Basket Size: 25.0 - 35.99

  Characteristics: Premium customers spend significantly high amounts, have high visit frequencies, and maintain high basket sizes, representing the top tier of high-end customers.

- **Cluster 2**: High-End - Loyal Customers
  - ➔ Monthly Spend: 120,367.21 - 133,296.05
  - ➔ Monthly Visit Frequency: 11.39 - 15.99
  - ➔ Monthly Basket Size: 29.56 - 39.98

  Characteristics: This cluster includes loyal customers who have slightly lower spend than Cluster 1 but still within the very high range, with high to very high visit frequencies and basket sizes.

- **Cluster 3**: High-End - Frequent Customers
  - ➔ Monthly Spend: 80,000.0 - 120,337.34
  - ➔ Monthly Visit Frequency: 7.7 - 22.0
  - ➔ Monthly Basket Size: 15.0 - 39.99

  Characteristics: Frequent customers exhibit high spend and basket sizes with very high visit frequencies, indicating they shop often with significant purchases.

- **Cluster 4**: Medium-Level Customers
  - ➔ Monthly Spend: 20,000.0 - 49,999.0
  - ➔ Monthly Visit Frequency: 4.0 - 10.0
  - ➔ Monthly Basket Size: 8.51 - 14.95

  Characteristics: These customers have average monthly spend, visit frequencies, and basket sizes, indicating moderate engagement and spending patterns.

- **Cluster 5:** Low-Level Customers
  - ➔ Monthly Spend: 0.0 - 14,999.0
  - ➔ Monthly Visit Frequency: 0.0 - 4.18
  - ➔ Monthly Basket Size: 0.0 - 6.87

  Characteristics: These customers exhibit low behavior across all three metrics, with lower spend, visit frequency, and basket size.

5. Rule-Based Conditions for Each Segment

The rule-based conditions for each segment were established based on observed data ranges and customer behavior:

- **Premium Customers (Cluster 1)**
  - ➔ Monthly Spend: 140,000 and above
  - ➔ Monthly Visit Frequency: 11.17 - 15.99
  - ➔ Monthly Basket Size: 25 - 35.99

- **Loyal Customers (Cluster 2)**
  - ➔ Monthly Spend: 120,367.21 - 133,296.05
  - ➔ Monthly Visit Frequency: 11.39 - 15.99
  - ➔ Monthly Basket Size: 29.56 - 39.98

- **Frequent Customers (Cluster 3)**
  - ➔ Monthly Spend: 80,000 - 120,337.34
  - ➔ Monthly Visit Frequency: 7.7 - 22
  - ➔ Monthly Basket Size: 15 - 39.99

- **Medium-Level Customers (Cluster 4)**
  - ➔ Monthly Spend: 20,000 - 49,999
  - ➔ Monthly Visit Frequency: 4 - 10
  - ➔ Monthly Basket Size: 8.51 - 14.95

- **Low-Level Customers (Cluster 5)**
  - ➔ Monthly Spend: 0 - 14,999
  - ➔ Monthly Visit Frequency: 0 - 4.18
  - ➔ Monthly Basket Size: 0 - 6.87

Methodology for Defining Conditions

To define these conditions:

- Data Analysis: The dataset was analyzed to understand the distribution of spending, visit frequency, and basket sizes across all customers.
- Clustering: DBSCAN clustering was applied to segment the customers based on their behaviors.
- Range Identification: The value ranges for each cluster were determined, and segments were matched to business-defined customer categories.
- Rule Formulation: Percentile-based rules were created to classify customers into these predefined segments accurately.

6. Marketing Interventions and Strategies

1. Premium Customers

- Promotion Goals: Increase customer spend.
- Strategies:
    - ➔ Offer exclusive high-end product promotions and early access to sales.
    - ➔ Provide personalized shopping experiences and dedicated customer service.
    - ➔ Implement tiered loyalty programs with significant rewards for continued high spending.

2. Loyal Customers

- Promotion Goals: Increase visit frequency.
- Strategies:
    - ➔ Launch targeted campaigns to encourage more frequent visits, such as double points days.
    - ➔ Provide incentives for referring friends or family members.
    - ➔ Offer personalized product recommendations based on past purchases.

3. Frequent Customers

- Promotion Goals: Maintain high visit frequency and increase basket size.
- Strategies:
    - ➔ Implement a subscription service for frequently purchased items.
    - ➔ Offer bundle deals and discounts on bulk purchases.
    - ➔ Use in-app notifications and emails to remind customers of deals and restocks.

4. Medium-Level Customers

- Promotion Goals: Increase engagement and spend.
- Strategies:
    - ➔ Offer introductory promotions and discounts to encourage higher spending.
    - ➔ Implement loyalty programs with achievable rewards to increase engagement.
    - ➔ Use personalized marketing to highlight products that match their spending patterns.

5. Low-Level Customers

- Promotion Goals: Increase engagement and spending.
- Strategies:
    - ➔ Offer significant discounts and promotions to attract higher spending.
    - ➔ Implement referral programs to encourage customers to bring in new clients.
    - ➔ Use targeted marketing to raise awareness of product ranges and benefits.

These strategies aim to leverage the unique characteristics of each customer segment, ensuring tailored and effective marketing interventions that align with the business's goals and customer needs.

Assumptions:

- Assumes removal of outlier data points to improve clustering accuracy without significantly affecting segment characteristics.
- Assumes customer data accuracy and consistency.
- Assumes willingness of customers to engage with promotional offers.

Conclusion

This report outlines a data-driven approach to customer segmentation and personalized marketing for KJ Marketing. By leveraging clustering techniques and rule-based conditions, the company can effectively target high-end segments and implement tailored interventions to enhance customer engagement, retention, and spend. The proposed strategies align with the company's goals of optimizing marketing efforts and delivering value to its diverse customer base.

*****