Data Storm v5.0

— Preliminary Round —

Organized By

Rotaract Club of University of Moratuwa

May 13, 2024

Powerd By

OCTAVE

1 Background To The Business

KJ Marketing is a leading retail supermarket chain in Sri Lanka, with a network of outlets across both urban and suburban areas. They offer an extensive selection of products, including dry goods, fresh items, and luxury products, catering directly to end consumers.

Figure 1: Grocery Items

2 Background To The Problem

The company has recently identified that during the past few years using conventional and standard marketing strategies has not been effective to their current customer base. Therefore, in an effort to enhance its marketing strategies, the company is interested in adopting a personalized marketing strategy tailored to individual customer preferences. To support this initiative, they have provided you with historical sales data, which includes average monthly sales per customer. You are tasked with developing a sophisticated analytical method to identify groups of similar customers based on this data.

The company operates 22 outlets located in both urban and suburban regions, offering a diverse range of products spanning dry goods, fresh produce, and luxury items.

- **Dry items** typically refer to non-perishable goods that do not require refrigeration or freezing to maintain their quality and safety. These items can include a wide range of products such as grains, cereals, pasta, canned goods, snacks, baking ingredients, spices, condiments, and beverages like coffee and tea. They are called "dry" because they do not contain high levels of moisture and can be stored at room temperature for extended periods without spoiling.

- **Fresh items** refer to perishable goods that have a limited shelf life and require refrigeration or special storage conditions to maintain their quality and safety. These items typically include fruits, vegetables, dairy products, meat, poultry, seafood, and bakery items like bread and pastries. Fresh items are valued for their nutritional content, flavor, and texture, and they are often purchased for immediate consumption or use within a few days. Maintaining the freshness of these items is crucial to ensure they meet consumer expect.

- **Luxury items** items are high-end products that are characterized by their premium quality, exclusivity, and often higher price point compared to standard or mass-market alternatives. These items are typically associated with superior craftsmanship, exceptional materials, unique designs, and prestigious brand names. Luxury items can span across various categories including fashion, accessories, jewelry, watches, electronics, pet foods , automobiles, home goods, and gourmet food and beverages.

(a) Dry Items

(b) Fresh Items

(c) Luxury Items

Figure 2: Item Categories

Consumers are drawn to luxury items for various reasons, including their perceived status symbol, exceptional performance or functionality, exquisite aesthetics, and the overall experience they offer. Owning luxury items can signify wealth, sophistication, and discerning taste, making them desirable to individuals seeking to indulge in a higher standard of living or express their personal style and identifications and health standards.

Through an initial analysis, **6 customer segments** of customers were identified. Your task is to classify the new customer to relevant customer segments.

3 Datasets and Variable Descriptions

You are provided with two types of datasets for your analysis as follows.

1. **train_kaggle.csv** – You are required to use this data set for model training and testing purposes.
    - **customer_id**: unique identifier of the customers.
    - **outlet_city**: identifiers of the cities where the outlets are located.
    - **luxury_sales**: average monthly sales per customer for luxury good.
    - **fresh_sales**: average monthly sales per customer for fresh goods.
    - **dry_sales**: average monthly sales per customer for dry goods.
    - **cluster_category**: categorizes the customers into different clusters.
2. **test_kaggle.csv** – You are required to use this data set to evaluate the model.
    - **customer_id**: unique identifier of the customers.
    - **outlet_city**: identifiers of the cities where the outlets are located.
    - **luxury_sales**: average monthly sales per customer for luxury good.
    - **fresh_sales**: average monthly sales per customer for fresh goods.
    - **dry_sales**: average monthly sales per customer for dry goods.

Teams should identify the clusters related to customers available in the following data set and submit them to the Kaggle in the format provided in section 4.

4 Deliverables and Evaluation Matrices

4.1 Analytical Solution - 40% of the total evaluation

For the customer_id in the test_kaggle data set, you are required to create an analytical solution to predict the relevant cluster_catgeory. Predictions should be in the following format as a CSV file and submitted to the Kaggle competition.

Customer_ID	cluster_category
4379	XXX
4368	XXX
4456	XXX
4234	XXX
4232	XXX
...	...

For the evaluation of the analytical solution you needs to use the **Accuracy Score**. i.e., You can use a function like sklearn.metrics.accuracy_score which computes subset accuracy when the set of labels predicted for a sample must **exactly** match the corresponding set of labels in y_true.

4.2 Technical Report - 60% of the total evaluation

Based on your analytical solution, you are required to create a report of your solution **(this needs to be in a proper report format)** with clearly defining the steps, features, feature engineering steps, modelling approaches, evaluation metrics, all the necessary plots/figures and interesting business findings that you can derive from this analysis while answering the following questions.

1. Elaborate on the methodologies implemented to address missing values, duplicates and outliers within the dataset? Please describe any specific techniques used for imputation or exclusion, and the rationale behind these choices.

2. Explain the features you chose for the above task. How did you determine their relevance to the problem?
3. Has feature scaling or normalization been applied to the data? If so, which methods were utilized and explain how these techniques improve the performance of the model?
4. Have you used any encoding strategies? Provide a comprehensive explanation of the chosen encoding methods and their impact on the model's input requirements and performance.
5. How do the features correlate with the target variable, and are there any notable inter-feature relationships?
6. Describe the target variable and interpret each category within it, detailing the characteristics that define the different customer segments.
7. What are the algorithms you considered for this problem, and why did you choose the final algorithm?
8. Were there any challenges faced during model training, such as over-fitting or computational constraints
9. Briefly define and explain all the classified clusters while providing appropriate names.
10. How can your solution enhance the effectiveness of the company's marketing strategies based on the classified clusters?