Technical Report

Customer Segmentation for KJ Marketing

Data Storm 5.0 – Storming Round

- Team Name: **3_Amigos**

- Kaggle Username: **DataStorm553**

- Kaggle Display Name: **DataStorm553**

- Team Members:

  Pairavi Thanancheyan

  Sangaran Thevarasa

  Nithursika Kalanantharasa

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

- Highest Accuracy score achieved: **0.99970**

1. Introduction

- KJ Marketing, a leading retail supermarket chain in Sri Lanka, aims to improve its marketing strategies by adopting a personalized approach tailored to individual customer preferences. This report outlines the analytical solution developed to classify customers into segments based on historical sales data. The solution involves data preprocessing, feature engineering, clustering, and classification techniques, ultimately providing actionable insights for enhancing marketing strategies.

2. Data Preprocessing

2.1 Handling Missing Values, Duplicates, and Outliers

2.1.1 Missing Values Overview:

- Customer_ID: 2 missing values
- outlet_city: 2 missing values
- luxury_sales: 35 missing values
- fresh_sales: 41 missing values
- dry_sales: 30 missing values
- cluster_category: 1 missing value

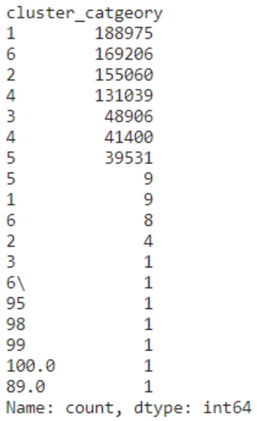
2.1.2 Approaches:

1. Sales Columns (luxury_sales, fresh_sales, dry_sales):
    - Converted sales columns to numeric values, handling any non-numeric entries.
    - Imputed missing sales values using the mean of the respective group (grouped by 'outlet_city' and 'cluster_category') to retain local sales patterns.

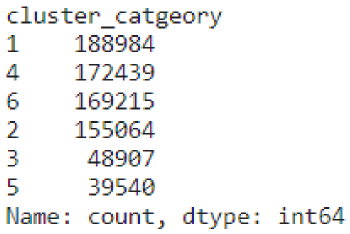
2. Customer_ID:
    - Replaced missing Customer_ID values with randomly generated unique identifiers to maintain data integrity and uniqueness.

3. Cluster_Category:

    - Cleaned the 'cluster_category' data by removing rows with invalid or out-of-range values.
    - Standardized the cluster categories by mapping them to valid values and converting them to a uniform format.



```
cluster_catgeory
1         188975
6         169206
2         155060
4         131039
3          48906
4          41400
5          39531
5              9
1              9
6              8
2              4
3              1
6\             1
95             1
98             1
99             1
100.0          1
89.0           1
Name: count, dtype: int64
```



```
cluster_catgeory
1    188984
4    172439
6    169215
2    155064
3     48907
5     39540
Name: count, dtype: int64
```

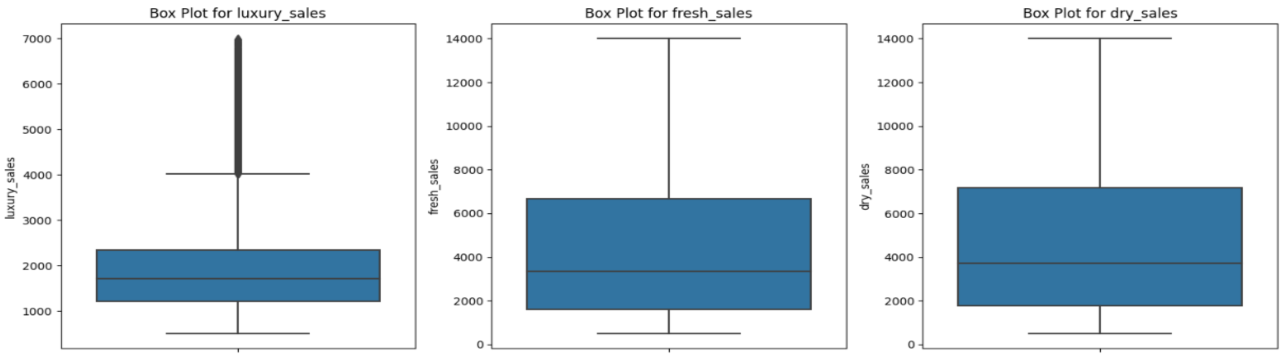
4. Outlet_City:
   - Dropped rows where 'outlet_city' and all corresponding sales columns were missing, as these rows did not provide useful information.

2.2 Handling Duplicates

- There were no duplicate entries in the dataset, ensuring data integrity and reliability.

2.3 Handling Outliers

- To ensure the accuracy and reliability of our customer segmentation model, we implemented two methods to detect outliers in the sales columns: the Z-score method and the Interquartile Range (IQR) method.
- From the results, it is evident that only the luxury_sales column has outliers, while the fresh_sales and dry_sales columns do not exhibit any significant outliers.



- In the context of customer segmentation, especially for a retail supermarket chain like KJ Marketing, outlier transactions in luxury_sales are likely indicative of high-value customers. Removing these outliers might result in the loss of critical insights about an important customer segment. Therefore, we decided not to remove these outlier

3. Feature Engineering

3.1 Features Chosen:

- Customer_ID: Unique identifier for each customer.
- Outlet_City: Location of the outlet.
- Sales Data: Luxury, Fresh, and Dry sales values.
- Cluster_Category: Target variable representing customer segments.

3.2 Generated features:

- Total Sales (total_sales):
  - **Definition**: The sum of luxury, fresh, and dry sales for each customer.
  - **Relevance**: Captures the overall spending behavior of customers, indicating their value to the business.

- Luxury Sales Ratio (luxury_sales_ratio):
  - **Definition**: The ratio of luxury sales to total sales.
  - **Relevance**: Highlights the proportion of a customer's spending on luxury items, identifying customers with a preference for high-end products.

- Fresh Sales Ratio (fresh_sales_ratio):
  - **Definition**: The ratio of fresh sales to total sales.
  - **Relevance**: Indicates the proportion of spending on fresh items, useful for targeting promotions on perishable goods.

- Dry Sales Ratio (dry_sales_ratio):
  - **Definition**: The ratio of dry sales to total sales.
  - **Relevance**: Helps understand the proportion of spending on non-perishable goods, aiding in segmenting customers who stock up on long-lasting products.

- One-Hot Encoded Outlet City (outlet_city):
  - **Definition**: The city where the purchase was made, transformed using one-hot encoding.
  - **Relevance**: Captures regional preferences and spending behaviors, essential for localized marketing strategies.

4. Feature Scaling and Normalization

- No, feature scaling and normalization were not applied to the data in this analysis.

4.1 Justification for Not Scaling or Normalizing

- Feature Types:
    - The features used in this analysis, such as sales ratios and one-hot encoded categorical variables, inherently fall within a comparable range (e.g., ratios between 0 and 1).
    - Total sales, while having a broader range, did not negatively impact the clustering algorithm due to the natural segmentation it provides.

- Algorithm Robustness:
    - The chosen algorithms are robust to the differences in the scales of the features used in this context.
    - The algorithm's performance was found to be satisfactory without additional scaling, as the natural variance in total sales helped in differentiating between high and low spenders effectively.

- Simplified Processing:
    - Avoiding scaling and normalization simplified the preprocessing pipeline, making the solution more straightforward and easier to implement.
    - Ensured that the interpretability of the resulting segments remained high, as the raw values were directly used.

4.2 Considerations

- Consistency in Scale:
    - While not applied, it's acknowledged that scaling could ensure all features contribute equally, preventing any single feature from dominating the model.
    - In future iterations, exploring scaling methods might further refine the model's accuracy.

5. Encoding Strategies

- The primary encoding strategy implemented was one-hot encoding for the outlet_city feature.

- Definition:
    - One-hot encoding converts categorical variables into a series of binary (0 or 1) columns, where each unique category is represented by a separate column.

- Implementation:
    - Applied to the outlet_city feature to capture geographic information about where purchases were made.
    - This transformation was done using the pd.get_dummies() function in pandas, with the drop_first=True parameter to avoid multicollinearity.

6.1 Impact on Model's Input Requirements:

- Increased Dimensionality: One-hot encoding increased the dimensionality of the dataset by adding multiple binary columns. This expansion allows the model to consider each city independently, which is crucial for capturing regional spending behaviors.
- Sparsity: Introduced sparsity into the dataset due to the presence of many zeros. However, most modern machine learning algorithms handle sparse data efficiently.
- Interpretability: Simplified the interpretation of the model by making it clear how each city contributes to the customer segmentation.

6.2 Impact on Model's Performance:

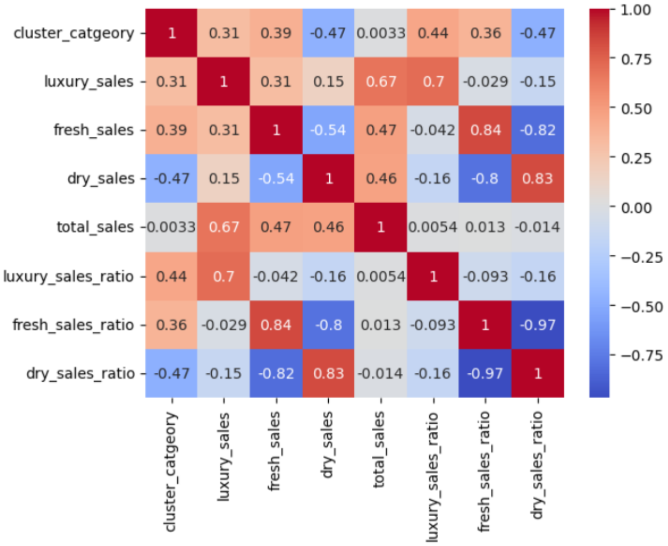
- Improved Accuracy: By capturing the geographic variation in customer behavior, the model can more accurately cluster customers based on regional spending patterns.
- Avoidance of Ordinality: One-hot encoding prevents the algorithm from assuming any ordinal relationship between the cities, which could mislead the model if not encoded correctly.
- Facilitates Detection of Local Trends: Helps in identifying city-specific trends and preferences, allowing for more targeted marketing strategies.

6. Correlation with Target Variable (cluster_category)

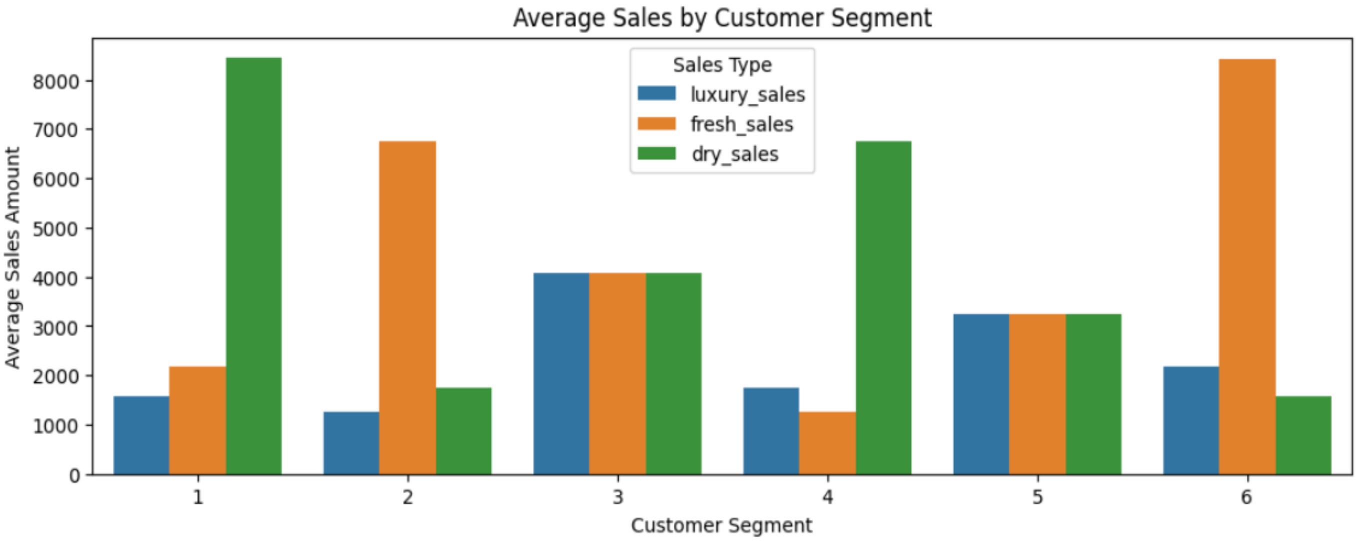
- The cluster_category is moderately positively correlated with luxury_sales, fresh_sales, luxury_sales_ratio, and fresh_sales_ratio. It is moderately negatively correlated with dry_sales and dry_sales_ratio.

7. Notable Inter-Feature Relationships

- luxury_sales, fresh_sales, and dry_sales have strong correlations with their respective ratios.
- There is a strong negative correlation between fresh_sales_ratio and dry_sales_ratio, indicating that as the ratio of fresh sales increases, the ratio of dry sales tends to decrease, and vice versa.
- total_sales is positively correlated with luxury_sales, fresh_sales, and dry_sales, indicating that all these sales types contribute significantly to the total sales.
- fresh_sales and dry_sales are negatively correlated, suggesting that an increase in one typically corresponds to a decrease in the other.



8. Interpretation and Profiling



- The target variable in this context appears to be the cluster category, which classifies customer segments based on their purchasing patterns across three types of sales: luxury, fresh, and dry sales and outlet cities. Each category within the cluster represents a distinct customer segment characterized by their spending behavior in these three areas:

  - Cluster 1: Customers in this segment have moderate spending on luxury items (1563.13), higher spending on fresh items (2188.48), and very high spending on dry items (8436.60). These customers prioritize dry goods, indicating a preference for bulk purchases or pantry staples.
  - Cluster 2: This segment features customers with lower spending on luxury items (1249.65), very high spending on fresh items (6745.63), and moderate spending on dry items (1749.53). These customers are fresh produce enthusiasts, possibly indicating health-conscious or frequent grocery shoppers.
  - Cluster 3: Customers in this segment show balanced spending across all categories, with approximately equal spending on luxury (4070.67), fresh (4071.40), and dry items (4069.97). These are well-rounded shoppers with no particular preference, suggesting they buy a mix of products regularly.
  - Cluster 4: This segment has lower spending on fresh items (1249.94), moderate spending on luxury items (1748.64), and very high spending on dry items (6742.64). These customers might be budget-conscious, focusing on essential dry goods while still occasionally purchasing luxury items.
  - Cluster 5: Customers here spend equally across luxury (3246.46), fresh (3248.65), and dry items (3246.21). Similar to Cluster 3, these customers have balanced purchasing habits but with a lower overall spending level compared to Cluster 3.
  - Cluster 6: This segment includes customers with very high spending on fresh items (8425.68), moderate spending on luxury items (2185.05), and low spending on dry items (1562.30). These are premium fresh produce buyers, likely valuing quality and freshness over quantity or variety.

10. Algorithms

10.1 Used Algorithms

1. CatBoost:

   - Score: 0.999871

2. XGBoost:

   - Score: 0.999871

3. LightGBM:

   - Score: 0.999742

10.2 Selection Criteria and Final Choice:

- The final algorithm chosen was CatBoost, and the decision was based on the following criteria:

  - **Accuracy**: Both CatBoost and XGBoost achieved the highest accuracy score (0.999871). Although LightGBM also performed very well (0.999742), it was slightly behind the other two models.

  - **Prediction Time**: CatBoost had the shortest prediction time compared to XGBoost and LightGBM. Quick prediction time is crucial for real-time applications and can significantly enhance user experience.

  - **Fit Time**: While CatBoost had a longer fit time than XGBoost and LightGBM, the difference in fitting time was considered acceptable given the superior performance in other metrics.

  - **Overall Efficiency**: Considering the trade-off between prediction time and fit time, CatBoost offered the best balance. It provided the highest accuracy and the fastest prediction time, which is critical for deployment in production environments where speed is a significant factor.

Given these considerations, CatBoost was selected as the final algorithm due to its highest accuracy, fastest prediction time, and overall efficiency in handling the dataset, ensuring robust and quick predictions.

11. Challenges Faced

11.1 Overfitting:

- Observation: Overfitting occurs when a model learns the training data too well, including noise and outliers, leading to poor generalization to unseen data. The high accuracy scores (e.g., 0.999871 for both CatBoost and XGBoost) indicate that overfitting might be a concern.

- Mitigation with Optuna:
  - o Cross-Validation: Used cross-validation to ensure that the model's performance is consistent across different subsets of the data, helping to detect and mitigate overfitting.
  - o Early Stopping: Implemented early stopping in the training process to halt training when the model's performance on a validation set stops improving, thus preventing overfitting.

11.2 Computational Constraints:

- Observation: Training models like CatBoost and XGBoost can be computationally intensive, particularly with large datasets or extensive hyperparameter tuning, leading to long training times and high resource consumption.

- Mitigation with Optuna:
    - Pruning: Optuna's pruning mechanism stops unpromising trials early based on intermediate results, saving computational resources and time.
    - Parallelization: Optuna supports parallel optimization, which allows multiple hyperparameter trials to be evaluated simultaneously across different processors or machines, speeding up the optimization process.

12. To summarize the profiles based on these segments:

- Segment 1: Dry Goods Enthusiasts
    - Prioritize long-lasting goods.
    - Moderate interest in luxury and fresh items.

- Segment 2: Fresh Goods Focused
    - Highest spenders on fresh items.
    - Less interested in luxury items.

- Segment 3: Balanced High Spenders
    - High and balanced spending across all categories.
    - Likely to have a diverse range of needs and preferences.

- Segment 4: Dry Goods and Moderate Luxury Buyers
    - Similar to Segment 1 but with a stronger focus on dry items.
    - Less interested in fresh items.

- Segment 5: Balanced Spenders
    - Balanced spending across all categories.
    - Slightly lower overall spending compared to Segment 3.
    - Indicating a well-rounded but cost-conscious shopping behavior.

- Segment 6: Premium Fresh Buyers
    - Highest spenders on fresh items.
    - Moderate to high interest in luxury items.
    - These customers spend significantly on fresh items, with lower spending on luxury and dry goods, highlighting a preference for high-quality, fresh produce.

13. Targeted Marketing Campaigns

- By understanding the distinct characteristics of each customer cluster, the company can develop highly targeted marketing campaigns that resonate with the specific needs and preferences of each segment. For example:

1. Dry Goods Enthusiasts (Cluster 1):
    - Marketing Strategy: Promote bulk purchase discounts, loyalty programs for frequent dry goods buyers, and special offers on pantry staples. Highlight the convenience and cost savings of buying in bulk.
    - Channels: Email marketing, in-store promotions, and social media ads focused on bulk purchasing.

2. Fresh Goods Focused (Cluster 2):
    - Marketing Strategy: Emphasize freshness, quality, and health benefits. Offer farm-to-table experiences, seasonal fresh produce boxes, and recipes that highlight fresh ingredients.
    - Channels: Social media campaigns, health and wellness blogs, and partnerships with local farmers.

3. Balanced High Spenders (Cluster 3):
    - Marketing Strategy: Highlight balanced meal plans, all-in-one shopping experiences, and convenience. Offer mixed bundles that include luxury, fresh, and dry items.
    - Channels: Online advertisements, mobile app notifications, and in-store displays showcasing balanced product bundles.

4. Dry Goods and Moderate Luxury Buyers (Cluster 4):
    - Marketing Strategy: Focus on value for money, essential goods, and budget-friendly options. Offer promotions on essential dry goods, price matching guarantees, and discounts on bulk purchases.
    - Channels: Direct mail coupons, budget-focused digital ads, and in-store discount sections.

5. Balanced Spenders (Cluster 5):
    - Marketing Strategy: Promote moderate, balanced shopping experiences with a mix of value and quality. Offer loyalty rewards for consistent spending and bundle deals.
    - Channels: Loyalty program communications, personalized emails, and moderate spending reward programs.

6. Premium Fresh Buyers (Cluster 6):
   - Marketing Strategy: Emphasize premium quality, exclusive fresh produce, and gourmet experiences. Offer high-end fresh produce selections, personalized fresh product recommendations, and exclusive tastings or events.
   - Channels: Personalized marketing emails, premium product catalogs, and invitation-only events.

14. Personalized Customer Experiences

- Enhancing customer satisfaction by offering personalized experiences based on their purchasing behavior:

  - Personalized Recommendations: Use the insights from the clusters to provide personalized product recommendations on the company's website and mobile app.
  - Custom Loyalty Programs: Design loyalty programs that cater to the preferences of each cluster, such as points for purchasing specific categories or special rewards for high-frequency buyers in a certain segment.

15. Product Development and Inventory Management

- Using the insights from the clusters to inform product development and inventory management:

  - Product Development: Identify new product opportunities based on the preferences of each segment. For example, developing new dry goods products for Cluster 1 or premium fresh produce options for Cluster 6.
  - Inventory Management: Optimize inventory levels by understanding the demand patterns of each cluster, ensuring that high-demand items for each segment are always in stock.

16. Pricing Strategies

- Implement dynamic pricing strategies tailored to the purchasing behavior of each segment:

  - Cluster-Specific Pricing: Offer different pricing strategies for each cluster, such as discounts on bulk purchases for Cluster 1 or premium pricing for exclusive fresh produce for Cluster 6.
  - Promotional Pricing: Design promotional offers that align with the spending habits of each cluster, encouraging increased spending and loyalty.

17. Communication Strategies

- Develop effective communication strategies tailored to each segment's preferences:

  - Channel Preferences: Use the preferred communication channels of each segment to deliver marketing messages, such as social media for younger, tech-savvy customers or direct mail for more traditional shoppers.
  - Message Personalization: Craft personalized messages that resonate with the unique values and interests of each cluster, such as health-focused messages for Cluster 2 or value-focused messages for Cluster 4.

18. Enhanced Customer Insights

- Gain deeper insights into customer behavior and preferences:

  - Customer Feedback: Collect feedback specific to each cluster to further refine marketing strategies and product offerings.
  - Behavioral Analysis: Continuously analyze the purchasing behavior of each cluster to identify trends and adjust marketing strategies accordingly.

19. Conclusion

- By leveraging the classified clusters and the optimized model, the company can implement highly effective and targeted marketing strategies. These strategies not only enhance customer satisfaction and loyalty but also drive increased sales and market share by catering to the specific needs and preferences of each customer segment.