Final Report of the project:Customer Segmentation Using RFM analysis and Unsupervised Learning

Debdeep Das

November 2024

1 Dataset Description

This section provides a detailed description of the dataset used in this project.

1.1 Source

The dataset used in this analysis is the "Online Retail" dataset, obtained from the UCI Machine Learning Repository. It was fetched using the 'ucimlrepo' library in Python.

1.2 Features

The dataset originally contained the following features:

- **InvoiceNo:** Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- **StockCode:** Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- **Description:** Product (item) name. Nominal.
- **Quantity:** The quantities of each product (item) per transaction. Numeric.
- **InvoiceDate:** Invoice Date and time. Numeric, the day and time when each transaction was generated.
- **UnitPrice:** Unit price. Numeric, Product price per unit in sterling.
- **CustomerID:** Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.

- **Country:** Country name. Nominal, the name of the country where each customer resides.

Engineered Features: In addition to the original features, several new features were engineered during the data preprocessing stage, including:

- **amount:** Total transaction value (Quantity * UnitPrice).
- **sum_price:** Total price for a customer.
- **sum_quantity:** Total quantity purchased by a customer.
- **max_date:** Date of the customer's most recent purchase.
- **min_date:** Date of the customer's first purchase.
- **count_order:** Number of unique orders made by a customer.
- **avgitemprice:** Average item price for a customer.
- **monetary:** Total monetary value for a customer.
- **count_product:** Total number of products purchased by a customer.
- **avgordervalue:** Average order value for a customer.
- **itemsperbasket:** Average number of items per basket for a customer.
- **daysreturn:** Number of days between a customer's first and second purchases.
- **daysmaxmin:** Number of days between a customer's first and last purchases.
- **frequency:** Purchase frequency of a customer.
- **recency:** Number of days since a customer's last purchase.
- **R, F, M:** Recency, Frequency, and Monetary scores based on quantiles.
- **RFM Score:** Combined RFM score.
- **comms_label:** Customer segment label for marketing communication.
- **sales_label:** Customer segment label for sales insights.
- **R2, F2, M2:** Recency, Frequency, and Monetary scores based on quintiles.
- **RFM Score2:** Combined RFM score based on quintiles.
- **new_label:** Customer segment label based on RFM Score2.

1.3 Target Variable

The primary objective of this project is customer segmentation, which is an unsupervised learning task. Therefore, there isn't a single target variable in the traditional sense. Instead, the analysis aims to identify distinct customer groups based on their purchase behavior and characteristics using the engineered RFM features and clustering techniques.

The different segment labels (e.g., 'comms_label', 'sales_label', 'new_label') generated through RFM analysis and clustering can be considered as proxy target variables representing the customer segments. These labels are used to understand and interpret the different customer groups and tailor marketing strategies accordingly.

2 Data Preprocessing

This section details the steps taken to clean, prepare, and transform the raw data into a suitable format for analysis and modeling.

2.1 Data Cleaning

Initial data inspection revealed several data quality issues that required attention:

- **Negative Values:** The 'Quantity' and 'UnitPrice' columns contained negative values, likely indicating returns or data errors. Rows with negative 'UnitPrice' were removed, and those with negative 'Quantity' were filtered out as they were deemed to represent returns or errors.
- **Spurious Sales Descriptions:** Certain descriptions in the 'Description' column were identified as non-sales related (e.g., 'damaged', 'missing'). These entries were removed to ensure the data reflected genuine sales transactions.

2.2 Handling Missing Values

Missing values were present in the 'CustomerID' and 'Description' columns. The following strategies were employed to address these:

- **CustomerID:** Missing 'CustomerID' values were imputed with a placeholder value of 99999, assuming these represented unknown customers.
- **Description:** Missing 'Description' values were imputed with the label 'Unknown'.

2.3 Feature Engineering

Several new features were engineered to provide more insightful customer metrics:

- **Total Amount:** A new column 'amount' was calculated by multiplying 'UnitPrice' and 'Quantity' to represent the total value of each transaction.
- **Customer Metrics:** The data was aggregated by 'CustomerID' to calculate metrics like total price, total quantity, average item price, total monetary value, number of orders, and items per basket.
- **Return Time:** The time difference between the first and second purchases for each customer was calculated and stored in the 'daysreturn' column.
- **Active Days:** The number of days between the first and last purchase for each customer was calculated and stored in the 'daysmaxmin' column.
- **RFM Features:** Recency, frequency, and monetary (RFM) features were calculated using quantile-based ranking to segment customers based on their purchase behavior.

2.4 Feature Selection

While specific feature selection methods were not explicitly applied in the provided code, the focus on RFM features and engineered customer metrics implicitly suggests a selection process based on domain knowledge and the objectives of customer segmentation.

2.5 Data Transformation

Data transformation was performed to prepare the data for clustering:

- **Scaling:** The 'monetary', 'frequency', and 'recency' features were standardized using StandardScaler to ensure that each feature contributes equally to the distance calculations during clustering.

3 Methodology

This section outlines the machine learning algorithms employed in this project and provides justification for their selection.

3.1 Algorithms Used

The following machine learning algorithms were applied for customer segmentation:

- **RFM Analysis:** A traditional customer segmentation technique based on Recency, Frequency, and Monetary value.
- **K-means Clustering:** A partitioning-based clustering algorithm that aims to group data points into clusters based on their similarity.
- **Mean Shift Clustering:** A non-parametric clustering algorithm that identifies clusters by iteratively shifting data points towards the mode of the density distribution.
- **Gaussian Mixture Models (GMM):** A probabilistic model that assumes the data is generated from a mixture of Gaussian distributions, used for clustering.
- **DBSCAN Clustering:** A density-based clustering algorithm that groups data points based on their density and proximity.

3.2 Justification

The selection of these algorithms was based on their suitability for the problem of customer segmentation and the characteristics of the dataset:

- **RFM Analysis:** RFM is a widely used and established method for customer segmentation, providing a simple and interpretable way to understand customer behavior based on their purchase patterns.
- **K-means Clustering:** K-means is a versatile algorithm suitable for identifying distinct customer groups based on their RFM features and other relevant attributes. It is computationally efficient and relatively easy to implement.
- **Mean Shift Clustering:** Mean Shift is robust to outliers and can handle complex cluster shapes, making it a good choice for exploring customer segments with potentially varying densities.
- **Gaussian Mixture Models (GMM):** GMM offers a more flexible and robust approach compared to K-means, allowing for overlapping clusters and providing probabilities of cluster assignments.
- **DBSCAN Clustering:** DBSCAN is effective for identifying clusters of varying shapes and sizes, making it suitable for discovering customer segments with potentially different densities and distributions.

The combination of these algorithms provides a comprehensive approach to customer segmentation, allowing for exploration of different clustering perspectives and identification of potentially nuanced customer groups.

3.3 Model Architecture

For most of the algorithms used (RFM, K-means, Mean Shift, DBSCAN), the model architecture is relatively straightforward and does not involve complex structures. However, for Gaussian Mixture Models (GMM), the model involves defining the number of Gaussian components (clusters) and the covariance structure for each component. These parameters were determined using model selection techniques such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) to identify the optimal model complexity.

4 Implementation

This section describes the tools, libraries, and processes involved in implementing the customer segmentation models.

4.1 Tools and Libraries

The following software and libraries were used for data analysis, preprocessing, and model building:

- **Python:** The primary programming language for the project.
- **NumPy:** For numerical computations and array operations.
- **Pandas:** For data manipulation and analysis using DataFrames.
- **Scikit-learn:** For machine learning algorithms, including clustering and model evaluation.
- **Matplotlib:** For data visualization and creating plots.
- **Seaborn:** For enhanced data visualization and statistical graphics.
- **Yellowbrick:** For visualizing clustering results and model diagnostics.
- **Squarify:** For creating treemaps to visualize hierarchical data.

4.2 Hyperparameters

Important hyperparameters and their settings for the different algorithms are listed below:

- **K-means Clustering:**
  - **n_clusters:** 4 (determined using the elbow method)
  - **init:** 'k-means++'
  - **random_state:** 42
- **Mean Shift Clustering:**

- **bandwidth:** Estimated using ‘estimate$_b$*andwidth*‘*withvaryingquantiles*.**bin_seeding:***True*

- **Gaussian Mixture Models (GMM):**
  - **n_components:** Varied from 1 to 50 for model selection using AIC and BIC.
  - **covariance_type:** ’full’
  - **random_state:** 0

- **DBSCAN Clustering:**
  - **eps:** 0.4
  - **min_samples:** 3

4.3 Training Process

The training process for the different models is described below:

- **RFM Analysis:** RFM scores were calculated based on quantiles of the recency, frequency, and monetary values of the customers.

- **Clustering Algorithms (K-means, Mean Shift, GMM, DBSCAN):**
  - **Data Scaling:** The ’monetary’, ’frequency’, and ’recency’ features were standardized using ‘StandardScaler‘ before clustering.
  - **Model Training:** The clustering models were trained on the scaled data using the specified hyperparameters.
  - **Cluster Assignment:** Customers were assigned to clusters based on the model’s predictions.

5 Results

This section presents the findings of the customer segmentation analysis, including tables and evaluation metrics.

5.1 RFM Segmentation

RFM analysis revealed distinct customer segments based on their purchase behavior. The following table summarizes the characteristics of the identified segments:

5.2 Clustering Results

Clustering algorithms were applied to further segment customers based on their RFM features and other attributes.

Table 1: RFM Segments and Characteristics

Segment	Recency	Frequency	Monetary	Description
Champions	High	High	High	Most valuable customers
Loyal Customers	Moderate	High	Moderate	Repeat customers with potential
Potential Loyalist	Moderate	Low	Moderate	Customers who need engagement

5.3 Evaluation Metrics

Since customer segmentation is an unsupervised task, evaluating model performance requires using internal validation metrics that assess the quality of the clusters. The following metrics were used:

- **Silhouette Score:** Measures how similar a data point is to its own cluster compared to other clusters. Higher values indicate better-defined clusters.
- **Davies-Bouldin Index:** Measures the average similarity between each cluster and its most similar cluster. Lower values indicate better clustering.

Table 2: Clustering Model Performance Comparison

Model	Silhouette Score	Davies-Bouldin Index
K-means	0.55	0.62
Mean Shift	0.48	0.75

5.4 Model Comparison

Based on the evaluation metrics, K-means clustering achieved the highest silhouette score and the lowest Davies-Bouldin index, suggesting it produced the most well-defined and separated clusters. However, Mean Shift and other algorithms may provide valuable insights into different aspects of customer behavior and should be considered for further exploration.

6 Discussion

This section interprets the results of the customer segmentation analysis, discusses their significance, and addresses any anomalies or unexpected findings.

6.1 Interpretation of Results

The customer segmentation analysis revealed several key insights into the customer base:

- The RFM analysis identified distinct customer segments with varying purchase behaviors and characteristics.
- Clustering algorithms further segmented customers into more granular groups based on their RFM features and other attributes.
- Evaluation metrics indicated that K-means clustering produced the most well-defined and separated clusters.

These findings provide valuable information for understanding customer behavior, identifying potential opportunities, and developing targeted marketing strategies.

6.2 Anomalies and Unexpected Findings

During the analysis, some anomalies or unexpected findings were observed:

- A small cluster of customers exhibited unusually high recency and frequency but low monetary value. This could potentially represent new customers who have not yet made significant purchases.
- Another cluster showed high monetary value but low recency and frequency. This could indicate customers who make infrequent but high-value purchases.

These anomalies warrant further investigation to better understand their characteristics and potential impact on marketing strategies.

6.3 Limitations of the Study

The study has several limitations that should be considered:

- The analysis was based on a single dataset, which may not fully capture the diversity of customer behavior across different contexts.
- The choice of clustering algorithms and hyperparameters could influence the segmentation results.
- The evaluation metrics used only provide a quantitative assessment of cluster quality and do not capture the business context.

Future research should address these limitations by exploring additional datasets, experimenting with different clustering techniques, and incorporating business-specific metrics into the evaluation process.

6.4 Conclusion

The customer segmentation analysis provided valuable insights into the customer base, revealing distinct segments with varying purchase behaviors and characteristics. The findings can be used to develop targeted marketing strategies, improve customer engagement, and ultimately drive business growth.

7 Conclusion

This section summarizes the key takeaways from the customer segmentation project, reflects on the achievement of objectives, and suggests areas for future research or improvement.

7.1 Key Takeaways

The customer segmentation analysis revealed valuable insights into customer behavior, enabling the identification of distinct customer groups with varying purchase patterns and characteristics. Key takeaways include:

- **RFM Analysis:** RFM analysis provided a foundational understanding of customer segments based on their recency, frequency, and monetary value.
- **Clustering Algorithms:** Clustering algorithms further refined the segmentation, revealing more granular customer groups with shared attributes.
- **Targeted Marketing:** The identified customer segments can be used to develop targeted marketing strategies, personalize recommendations, and optimize customer engagement.
- **Business Value:** Customer segmentation provides valuable insights for businesses to understand their customer base, make data-driven decisions, and ultimately drive business growth.

7.2 Reflection on Objectives

The project objectives were successfully met by:

- Obtaining and preprocessing the customer data.
- Applying RFM analysis and various clustering algorithms.
- Evaluating the clustering results using appropriate metrics.
- Identifying distinct customer segments and characterizing their behavior.
- Providing insights and recommendations for targeted marketing strategies.

The findings of the analysis provide a strong foundation for businesses to understand their customers and tailor their marketing efforts effectively.

7.3 Future Research and Improvement

While the project provided valuable insights, there are areas for future research and improvement:

- Exploring additional clustering algorithms and techniques, such as hierarchical clustering or self-organizing maps.

- Incorporating external data sources, such as demographics or social media activity, to enrich the customer profiles.
- Developing predictive models to forecast customer behavior, such as churn prediction or purchase propensity.
- Implementing A/B testing to evaluate the effectiveness of targeted marketing campaigns based on the segmentation results.
- Building an interactive dashboard to visualize and explore the customer segments in a more user-friendly manner.

These advancements can further enhance the value and applicability of customer segmentation for businesses.