Abraham Abate
DSC680 – Applied Data Science
Professor Amirfarrokh Iranitalab
First Project – Milestone 1

Project Proposal: Comparing Clustering Algorithms for Customer Segmentation

Topic: This project aims to compare different clustering algorithms for customer segmentation based on purchasing behavior using RFM analysis. By evaluating various clustering techniques, I believe we can determine which method provides the most meaningful customer groups for business decision-making.

Business Problem

Retail businesses rely on effective customer segmentation to tailor marketing strategies and improve customer retention. However, different clustering algorithms may yield different segmentations, making it essential to compare their performance. This project will address the following questions:

- How do different clustering algorithms compare in segmenting customers based on RFM analysis?
- Which clustering approach provides the most interpretable and actionable customer segments?

- How do segmentation results impact customer lifetime value (CLV) and churn prediction models? By comparing clustering methods, businesses can choose the most suitable approach to enhance customer engagement and revenue management.

Datasets

The dataset comes from UCI Machine Learning Repository containing transactional data with attributes such as invoice number, product description, quantity, unit price, customer ID, and country. This dataset contains 8 features and 541,909 instances. The dataset contains transactions from a UK-based, registered non-store online retailer, recorded between December 1, 2010, and December 9, 2011. The data will be used to analyze purchasing patterns and predict future behavior.

The attributes in the dataset include:

- InvoiceNo: Unique identifier for each transaction
- StockCode: Product code
- Description: Product description
- Quantity: Number of units purchased
- InvoiceDate: Date and time of purchase
- UnitPrice: Price per unit of the product
- CustomerID: Unique identifier for each customer
- Country: Customer's country of residence

Methods

For this analysis, I am planning to perform four separate analyses:

1. RFM Analysis:

Recency, Frequency and Monetary Value (RFM) are key metrics used for customer segmentation. Recency measures the number of days since a customer's last purchase, Frequency represents the total number of purchases made, and Monetary Value indicates the total amount spent. RFM scores will be calculated based on these metrics and used as input for clustering algorithms to group customers into meaningful segments for analysis.

2. Clustering Algorithms to Compare:
    - K-Means Clustering: A centroid-based approach that groups customers by minimizing intra-cluster variance.
    - Hierarchical Clustering: A tree-based approach that builds nested clusters based on similarity.
    - DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Identifies clusters based on density, making it useful for detecting outliers.
    - Gaussian Mixture Models (GMM): A probabilistic clustering method that assigns customers to multiple clusters with different probabilities.
    - Self-Organizing Maps (SOMs): A neural network-based clustering method for high-dimensional data.

3. Evaluation Metrics for Clustering:

Clustering performance will be evaluated using multiple metrics to ensure meaningful and well-defined customer segments. The Silhouette Score will measure how well-separated the clusters are by assessing the similarity of data points within their assigned clusters. Additionally, the Dunn Index or Davies-Bouldin Index will be used to evaluate cluster compactness and separation, ensuring that clusters are both distinct and internally cohesive.

4. Impact on CLV and Churn Prediction:

To evaluate the business relevance of different clustering approaches, CLV prediction and churn classification models will be applied to each segmentation method. This will help determine which clustering technique provides the most meaningful insights for customer behavior and retention strategies. For CLV prediction, regression-based models such as XGBoost will be used to estimate the total revenue expected from a customer over time.

For the Churn prediction, logistic regression and random forests will be utilized to identify customers at risk of discontinuing purchases. By analyzing transaction history and customer behavior, these models will enable targeted retention efforts to reduce churn and improve long-term customer value.

Ethical Considerations

Ethical concerns in customer data analysis include data privacy, ensuring transaction data is anonymized and securely stored, and bias in predictive models, which must be carefully managed to prevent discrimination against specific customer groups. Implementing strong security measures, selecting unbiased training data, and continuously monitoring models are essential for responsible data use.

Additionally, transparency is crucial to help stakeholders understand segmentation and predictive insights, preventing misinterpretation. Fairness in retention strategies should also be prioritized, ensuring that marketing efforts do not solely favor high-spending customers but also support loyal, lower-spending ones, fostering trust and long-term engagement.

Challenges/Issues

During this analysis, several challenges may arise that could impact the accuracy and effectiveness of the results. One major issue is data quality, as missing CustomerIDs, incorrect transaction records, and outliers can distort insights and lead to unreliable predictions. Addressing these issues requires thorough data cleaning, validation, and imputation techniques to ensure the dataset is accurate and complete.

Feature engineering is another critical challenge, as developing meaningful features for CLV and churn prediction requires in-depth domain expertise. Identifying the right variables and transformations can significantly influence model performance, making it essential to experiment with different feature selection and extraction methods.

Selecting the right clustering method is also a key consideration, as different techniques may produce varying customer segments. K-Means, hierarchical clustering, and other methods must be evaluated carefully to determine which best captures meaningful patterns in the data. Proper validation and interpretation of clustering results are necessary to ensure actionable customer segmentation.

Finally, model interpretability is crucial to ensure that business stakeholders can understand and act upon model predictions. Complex models like gradient boosting or deep learning may offer high accuracy but can be difficult to interpret. Providing clear explanations, using interpretable models where possible, and leveraging visualization techniques can help bridge the gap between data science and business decision-making.

References

- Fader, P. S., Hardie, B. G., & Lee, K. L. (2005). "RFM and CLV: Using Past Customer Behavior to Predict Future Behavior."
- Han, J., Kamber, M., & Pei, J. (2011). "Data Mining: Concepts and Techniques."
- Chen, Da., Sain, S.L., Guo, K. (2012)“Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining.”
- UCI Machine Learning Repository