DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

CLUSTERING REPORT

Group 43

Filipe Esteves, 20250470

Mariana Freitas, 20250369

Miguel Caramelo, 20250378

Miguel Silva, 20250379

Fall/Spring Semester 2024-2025

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

The following report will detail the findings of our clustering analysis performed on the dataset provided of 16,574 customers. The process focused on customer's flight behavior and purchasing habits to measure frequency, purchasing power, tenure and engagement, leading to distinct customer segments as insights for AIAI's marketing strategy.

The results show a three-segment within the customers, including: Loyal and High-Value Customers; Growth Potential Customers; and Low Loyalty Customers. Each segment has distinct patterns going over engagement, value contribution and retention potential, allowing us to build the next procedures. Key recommendations are retention and endorsement towards the Loyal and High-Value customers, personalized strategies for Growth Potential customers, and reactivation strategies for Low Loyalty Customers.

2. INTRODUCTION

In today's business environment, analyzing data is crucial to achieve success and improved outcomes, especially when dealing with a very competitive industry.

The initial explanatory analysis revealed less flights with companions, a high rate of points accumulated with low redemption showing possible engagement. This suggests latent engagement and unrealized value of the customer base, while also demonstrating how capable this company is of retaining potential sowed by the lack of cancellation rates, along with the discovery of customers that are a great financial target through outliers.

These findings prompt the use of Clustering to group customers based on their similarities to support targeted marketing, to move beyond the potential of the airline. This technique will expose the differences in engagement intensity, loyalty behavior as well as financial value, which will also enable us to build meaningful and actionable profiles to reach monetization opportunities.

Our explanatory data analysis prepared the data preprocessing and feature engineering, to get an effective clustering performance and interpretability.

Within the *flights* dataset, we created two ratio-based features: *PointsRedeemRatio* and *FlightsWithCompanionsRatio.* These ratios normalize the customer behavior, measuring the customer's engagement with the points system and how frequently they fly with companions relative to their total activity. To capture customer recency, we created a *Recency_Month* defined as the number of months between the most recent flight and a fixed reference date (most recent month available on the dataset), providing a time-based measure of engagement where lower values indicate recent activity.

Within the *customer* dataset, we created *Income Class* where we'll capture the distribution of the income among the education levels. *CustomerTenure* which encapsulates the length of the relationship between the customer and the airline, helping to distinguish long-term loyal customers from those with shorter relationships. The last feature engineered, *Active_Client*, is a binary indicator of retained customer from those who need to enter the reactivation procedure.

3. METHODOLOGY

Upon data preprocessing and feature engineering complete, we followed along with feature selection to ensure the best comparability across variables for clustering.

To proceed with the feature selection, a heatmap was used with Pearson Correlation to check the relationship between every variable to select the most crucial. Features with low variance were removed as they provided limited insights in distance-based clustering algorithms, for example *Gender*, *Country* (this dataset reflects exclusively in Canada), *Location Code*, *First Name*, *Last Name* and *Customer Name*, as well as some numeric features mirrored by other more significant features, such as *DistanceKM* and *DollarCostPointsRedeemed.*

Highly correlated features were removed as well to avoid over weighting similar behavior and to prevent distortion of the result of the clusters. When encountering variables in this situation, an evaluation is thought out against the business relevancy of the variables in question to further translate them into an actionable insight.

After feature selection, the dataset is merged and all retained variables were standardized, using OneHotEncoder for categorical features and StandardScaler for numeric features. The resulting set of features provides a comprehensive representation of customer engagement and behavior ready to build customer segments.

To build the customer segments with Clustering Algorithms we implemented the following 3 techniques.

3.1. Agglomeration Hierarchical Clustering

Agglomeration Hierarchical Clustering is a distance-based approach able to detect unusual data points present in the dataset, hierarchical clustering allows for a visual inspection of the customer's similarities through dendrograms. The dendrogram illustrates the successive merging of observations based on the increase in within-cluster variance.

3.2. K-Means Clustering

K-Means Clustering is a method appropriate for the features that are continuous, numeric, standardized and as we are dealing with a moderate-to-large customer base, offering computational efficiency and clear interpretability through cluster centroids. These characteristics are usually better to translate into actionable customer segments easily communicated to stakeholders.

3.3. DBSCAN Clustering

We also implemented DBSCAN Clustering that accounts for the fact that customer distributions may include dense behavioral cores as well as sparse, irregular outliers. By identifying clusters based on local density rather than predefined shapes or sizes, DBSCAN naturally isolates noise and reveals organically formed groupings, making it especially effective for detecting niche segments and boundary cases that traditional methods might overlook.

In terms of tuning, the number of clusters was evaluated across several values, using silhouette scores, the elbow method and dendrogram structure, which together these techniques provided complementary perspectives: hierarchical clustering made a clear structural exploration but appears to be less scalable, K-Means represents a more stable interpretable segmentation.

4. RESULTS AND VALIDATION

The multiple clustering approaches were applied to determine the suitable number of clusters, hence number of customer segments.

The approach Agglomerative Hierarchical Clustering suggests a three-cluster solution through the dendrogram analysis as it revealed a clear distinction into three major branches before the distances decreased substantially, therefore additional splits would force artificial and not reveal meaningful structures.

The approach K-Means Clustering was evaluated across multiple values of k using both the elbow curve and the silhouette scores. The elbow curve clearly shows a pronounced inflection at k=3, following that point reduction within the cluster variance diminish. The silhouette analysis supports this result as it showed higher average values for three clusters in comparison with a by other options.

The DBSCAN clustering approach was evaluated using the k-distance graph for the 12th nearest neighbor. The curve exhibits a clear elbow at approximately **eps = 1.2**, where the slope sharply increases, indicating a transition from dense regions to sparse noise. This inflection point suggests a suitable threshold for distinguishing core clusters from outliers. The visual analysis supports the selection of this eps value, as it balances cluster compactness with noise minimization, aligning with DBSCAN's density-based principles.

With our three-cluster solution selected, we are proceeding with defining our customer's segments.

The first segment is the **Loyal and High-Value Customers**, long tenured customers that have demonstrated consistently high engagement with AIAI throughout the observation period, and actively participate in the loyalty program by accumulating and redeeming their points regularly. Usually consist of high flight frequency, low recency values, an active status, high customer tenure, high spending and revenue and high usage of loyalty program benefits. This segment is the most valuable

for AIAI as it reflects brand attachments and trust which is essential for long-term profitability, therefore the airline must maintain the service quality as well as perceived value due to the financial weight this segment contains.

The second segment is **Growth Potential Customers**, where customers have less consistent engagement compared to the highly loyal group, with moderate tenure and somewhat interact with the loyalty program, however, through the points redeemed variable, it is noticeable the incomplete usage of the available benefits. Usually consists of moderate flight frequency, mixed recency patterns, medium customer tenure, moderate spending and revenue, and accumulation of loyalty points with low redeemed points. This segment translates into opportunity, as customer display sufficient engagement to become a high-valued status, for that reason it is crucial to create targeted incentives, personalized offers and improve communication of the former benefits to reach their full potential.

The third segment is **Low Loyalty Customer**, where limited engagement prevails, there's shorter customer relationships and higher likelihood of inactivity or cancellation along with a very minimal loyalty program participation. Usually consists of low flight frequency, high recency, low spending and limited financial contribution and minimal engagement with the loyalty program. This segment illustrates lower economic value where the airline should rebuild a relationship with the customer through reactivation procedure, while also being cost-effective.

5. STRATEGIC RECOMMENDATIONS

Based on the three segments and behavior profiling, in these sections we'll be suggesting targeted and actionable strategies tailored to each customer segment. These recommendations are set to improve retention, increase the customer lifetime value, and make their marketing resources more efficient.

For our Loyal and High-Value Customers, the airline should focus on maintenance of the perceived value by introducing or strengthen VIP loyalty programs with exclusive benefits such as priority services and early access offers. For example: offering points accumulated on every customer's anniversary increasing depending on how many birthdays this customer has spent with the airline; milestone-based bonuses by accumulating extra points when reaching 1000+ points accumulated, also depending on how long this customer has used the AIAI's services, those who reach 3000+ points accumulated get airport lounge clearance (within Canadian airports); perks on flights with companions by having points accumulated per person (over 18 years of age) traveling, available for a maximum of 2 flights per year.

For the Growth Potential Customers, the airline should focus on deepening engagement to convert them into high-value customer by advertising cross-selling campaigns based on past preferences as well as increase the communication towards the customer. For example: messaging the customer on how many points accumulated they have available to redeem via notifications and/or emails; tailored benefits, such as seat selection, discount on luggage and meal offers for the next flight using the points accumulated available; if recency increases, offer a time-limited discounts.

For the Low-Loyalty Customers, the airline should proceed with a reactivation procedure, whilst being cost-efficient. For example: start off by offering a quick return incentives, such as, "fly with us this summer, get 3000 points" to try getting the customer into the points system; small points bonus for each flight along with an reinforcement for holiday seasons; increase communication through notifications and emails to remind the customer of the points system perks along with offers to re-enter the loyal program. If this procedure is successful, move these customers to the second segment to increase the value of the loyalty program perceived by the user.

6. CONCLUSION

This segmentation analysis provides the airline AIAI with a clear, data-driven understanding of their customer's loyalty. The three-cluster solution integrates behavioral indicators, value-related features and loyalty indicators, in order to gather new actionable marketing strategies for business stakeholders.

Some limitations we encountered along the analysis include the sensitivity from the features chosen on the feature selection and aggregation made for the outcome of the cluster, and relying on the distance-based algorithm for the clusters capture the relationship of the features adequately. For the future, it could be interesting to explore the customer lifetime value forecasting to enhance the segmentation of the customers along with the strategic impact.

Overall, we were able to build a strong foundation for the marketing of the program, the improvement of the customer's experience, and loyalty optimization at AIAI.

7. AI STATEMENT

Artificial Intelligence was used to support and refine our notebooks, specially assisting in the making of the extra points. AI was also used for debugging code.

8. CONTRIBUTION

Filipe Esteves 20250470

- K-means Clustering
- DBSCAN Clustering
- Clustering-Based Recommender System
- Report Analysis/Results Interpretation

Mariana Freitas 20250369

- Report Analysis/Results Interpretation
- Video
- Hierarchical Clustering
- Deep Embedded Clustering

Miguel Caramelo 20250378

- EDA's
- Report Analysis/Results Interpretation
- Hierarchical Clustering
- Clustering with perspectives

Miguel Silva 20250379

- DBSCAN Clustering
- Deep Embedded Clustering
- Clustering-Based Recommender System
- Report Analysis/Results Interpretation

We, the group members listed above, certify that this report represents our original analytical work and interpretations. While AI tools were used as specified before, all insights, conclusions, and recommendations are the result of our independent analysis and critical thinking. We take full responsibility for the accuracy and quality of this submission.