NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 1: Hotel customer segmentation

David, Psuik, number: 20230818
Peter, Falterbaum, number: 20230956
Luis, Penteado, number: 20230441
Noah, Campana, number: 20230996

Group A

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

March, 2024

INDEX

1. Executive Summary

This research aims to refine Hotel H´s marketing strategies through effective customer segmentation by leveraging data-driven insights to replace the currently used marketing segments bases only on the sales origin. The objective is to gain deeper customer knowledge, which will enable the identification of specific customer patterns through the usage of a diverse range of variables including geographical, demographic, and behavioral characteristics.

The focus on existing customers is an important component of this strategy, which is driven by insights showing that keeping current clients is more cost-effective than acquiring new ones. Currently only 4% of the customers revisit Hotel H, which illustrates the potential of improving customer loyalty. The business criteria are centered on measurable increases in unique monthly customers.

The segmentation model is implemented with the k-means algorithm, utilizing the variables *Age*, *AverageLeadTime, LodgingRevenue, BookingsCheckedIn* and *PersonNights*. The success of this model is assessed by statistical measurements which evaluate the quality of the generated clusters and the ability of gaining valuable business insights.

Four separate clusters were generated with the provided dataset which ranged from young travelers to experienced travelers, most valued customers, and early birds. The recommendations made for each cluster include specific tactics like personalized offers and leveraging digital channels for communication. By customizing marketing strategies to the preferences of each customer segment Hotel H might improve customer loyalty.

Furthermore, a significant opportunity lies in engaging the high number of individuals that created an account but never stayed at the hotel. This research suggests a more detailed data collection during the account creation, which would facilitate deeper customer analysis for potential customers.

To conclude, through the implementation of a detailed analysis and a customized approach, Hotel H can target their customers more precise and efficient to subsequently increase their revenue.

2. Business Understanding

2.1. Business Objectives

This project has the main goal to rebuild Hotel H's marketing strategies based on an advanced understanding of its customers by developing a clustering segmentation model. Currently, their decision-making relies on a simplistic segmentation based on a single information about the customer: the sales origin. By using only one feature the hotel fails on capturing different needs and patterns of their clientele. To address this critical question, the hotel should categorize its customers based on a wider variety of factors including geographical, demographic, and behavioral characteristics. The expected outcome is to achieve a significant improvement in customer knowledge and be able to use these insights to increase marketing efficiency and increase revenue streams for the hotel. Furthermore, this problem can be issued based on a more tangible perspective. Currently, only 4% of H's customers book more than once in the hotel, leaving the most customers one-time visitors. This ratio shall be increased leveraging the segmentation in scope.

During this research we will focus on Hotel H´s current customers, guiding our strategies towards strengthening the loyalty of the existing clientele and refining the marketing approaches for each segment. The emphasis of focusing on existing customers is based on the insights from the Harvard Business Review, which highlights that acquiring a new customer can be 5 to 25 times more expensive than retaining existing ones.[1] This issue can be addressed by a thorough customer segmentation to understand customers better, leading to improved services and marketing.

2.2. Business Success Criteria

The key success criterion will be achieved if our marketing initiatives and customer engagement strategies engage customers to more frequent bookings or extended use of Hotel H's services, which lead to overall increased revenue. The business impact measure of the success criteria will be conducted by the marketing team in collaboration with the analytics department. This process ensures a precise assessment of the business's performance.

2.3. Determine Data Mining Goals

The data mining goals for this project are about how we can translate the business situation into actionable technical objectives. The main goal is to replace Hotel H's current single feature segmentation using a clustering technique, specifically the k-means algorithm. The technical approach purpose is using several variables but not limited to geographic, demographic, and behavioral characteristics, and find patterns within the customers unrecognized previously. To evaluate the cluster quality, the silhouette score and inertia will be applied. The process will be considered successful if the clusters have capacity to clearly generate actionable insights and how effective the actions can be applied to our business purposes and strategic objectives. Practically that means considering the distribution of the clusters among all customers to ensure separate groups.

[1] Gallo, 2014

3. Methodology

3.1. Data Understanding

The dataset for Hotel H contains 111,733 entries across 29 distinct columns. This data contains demographic details such as age and nationality from 199 different nationalities, from which 90% of the customers are European. The most frequent nationalities that stayed in the hotel are Germany, Spain, France, Portugal, and Great Britain. The total obtained lodging revenue is the highest from German customers and the lowest from Portuguese which can be seen in Appendix 1.

Furthermore, it includes a spectrum of customer preferences and specific information about their bookings and spendings. The data was gathered over a period of more than three and a half years.

Within the dataset, there are notable gaps in age and Document ID, with 3.73% and 0.89% missing values, respectively. The distribution of several columns is skewed, with a substantial presence of outliers that will need to be addressed to avoid distortion in analysis. It was also found some seasonality in the days since creation feature.

Over 30,000 entries (29,71%) represent loyalty program members who have never completed a stay at the hotel. Furthermore, customers below the age of 18 constitute 6.65% of the dataset entries. The occurrence of a single Document ID for roughly 3,000 entries suggests a default value placeholder for missing information, which will be excluded from the analysis to prevent the introduction of noise.

Cancellations and no-shows represent a small percentage of the total bookings, implying a potentially strong customer commitment once a booking is made. Moreover, the booking behavior shows a preference towards travel agents, with about 80% of bookings made through this channel. There seems to be a seasonal cycle in the observed data as every 365 days a similar a spike appears in the creation of new customers. Lastly, the data suggests a trend where customers who book early in advance, specifically more than 400 days ahead, are typically around 71 years old and have a lower likelihood of cancelling their bookings.

3.2. Data Preparation

For clustering, as for most machine learning algorithms, data preprocessing and careful transformation are necessary steps to ensure the coherency and reliability of the data, which, in turn, guarantees the validity of the clustering results. This chapter lies out the performed treatments to address issues such as correlations, data inconsistencies - including missing or duplicated data - and outliers. Subsequently the data transformation to align with format needed for applying the clustering is presented.

Data Inconsistency and Missing Values

The dataset contains a certain document id being used over 3000 times for different customers. Without additional information which justifies the repetition, the use of the id was considered as a test case. Consequently, customers associated to this document id have been removed from the dataset. Same treatment was applied to entries lacking any document id. This led to a removal of around 3.5% of the data. Furthermore, missing values on age have been imputed with the KNN imputation method. This approach fills missing values based on similar entries, maintaining the dataset's overall distribution.

Several customers occurred multiple times in the dataset. Customers were considered duplicates if the same document id, name, and nationality was assigned. These customers have been aggregated, with numerical features being averaged or summed, depending on the business context. Binary or categorical features have been combined using the mode. This ensures that each customer appears only once in the dataset which removes redundancies and improves the quality.

6 % of the data set are minors under the age of 18. In the analysis, these minors were regarded as accompanying guests who neither actively participate as customers nor generate revenue for the hotel. Marketing strategies would not be effective on them. Therefore, they were removed from the data and not included in the clustering.

Correlation

According to Pearson correlation method a correlation matrix was plotted. It was found a high correlation between Room nights and another three variables: *PersonsNights* (aprox. 0.86), *LodgingRevenue* (aprox. 0.71) and *BookingsCheckedIn* (aprox. 0.69). By excluding *RoomNights* correlation of several features is avoided, considering a Pearson's coefficient threshold of 0.7.

Outliers

Boxplots have been employed to manually identify skewedness and outliers. Several features were indicated with outliers which have been processed with adequate approaches to deal with their outliers. In this study different approaches have been tested for outlier treatment, with its success measured subsequently by the clustering and its quality.

Creating a baseline for the assessment, the first approach removed manually most extreme outliers by setting thresholds such that at most 1% of one feature's data is classified as outliers. In total, this led to a removal of 4% of extreme entries. Comparing to this approach other treatments were evaluated, whereas all other approaches were based on the first. The second approach capped the values manually so that values extremer than a threshold has been set to the threshold. Binning the features was the third approach used to engage with outliers. The bins have been defined by a 10-fold frequency separation, filled with the median of each bin. Therefore, the distribution was kept the same, balancing the effect of outliers by using the median of the bins.

Categorical Variables

To normalize the distribution of the 199 different nationalities we kept the most frequent ones and group the other into a separate class called "others". After this transformation a one hot encoder was employed.

Feature Engineering

AnyFloorPreference was created to aggregate any preferences for a specific floor as low and medium floor requests have been small. This summed up 3 features.

A new feature *Signups* was introduced to flag persons who did not check in at least once. These people are considered potential customers as they already signed up to the hotel.

The *Companion* feature labels persons who do not have any revenue, but at least one check in. These persons are considered not responsible for the booking.

The feature *Seasonality* categorizes each customer based on the time of the year they were created, dividing the year into four equal parts to represent the four seasons. This allows to analyze patterns across the different seasons.

Scaler

Given the presence of skewedness in the data, the robust scaler was used for normalization. The scaler applies the statistical distribution, especially the median of the values to scale. This results in possible negative values for certain scaled quantities, despite the original values being positive. The robust scaler is known to be effective against the leverage of outliers, while ensuring the integrity of the central data distribution. Therefore, it was chosen for this analysis.

4. Modeling

4.1. Segmentation Design

We will focus on creating actionable insights regarding customer retention, by separately storing records of individuals who enrolled in the loyalty program but never stayed at Hotel H. Therefore, the performed clustering will be based specifically on clients that had at least one check-in. In addition, the selection of useful features to develop the clusters are important to guarantee meaningful and targetable clusters. The choice of features is supported by technical analysis. The selected features to apply the clustering model are described in the following.

Age is a crucial feature that might allow identifying generational trends in travel preferences. *Average Lead Time* for example possibly helps to distinguish customers based on planning habits. *Lodging Revenue* measures the spending of the customers, which could help identifying high-value clusters. Furthermore, the number of check-ins might help to identify loyal customer groups. The feature *PersonsNights* might reveal patterns which are typical for group or business travelers and BookingsheckedIn could capture variations in customer loyalty while ensuring that the derived marketing actions are practical and implementable.

With the given dataset, we will implement the k-means clustering algorithm to identify customer segments. The applied method enables the classification of customers into different groups based on similarity in the selected features. The resulting segments should provide a deeper understanding of the customers to allow more personalized marketing strategies.

4.2. Technical Evaluation of the Model

We are aiming to group customers into clusters to differentiate distinct profiles from each other. To be more precise, segments should differ from each other when looking at single features to derive actionable insights. To achieve this, several preprocessed datasets have been used in the clustering process.

To determine the optimal number of clusters for the segmentation, we made use of statistical measurements which compute and compare several distances across different clusters. Metrics like

the Silhouette Score, inertia, distortion, and the inter-cluster distance were calculated for ranging number of clusters and evaluated each one of the different approaches. The following proposed number of clusters were analyzed and further evaluated to ensure business relevant segments which bring value to the marketing team. Furthermore, we intend on attaining balanced distribution among the clusters.

In addition, several visualization techniques have been leveraged to assess the quality of the resulting clusters and their separation from each other. This manual evaluation reaffirms that the clusters identified are not only statistically significant in their distribution and separation, but practically valuable for targeted marketing efforts.

5. Evaluation of Segmentation Results

This study targeted a customer segmentation of Hotel H to provide insights into different profiles to enrich the understanding of the dynamics and characteristics of the customers. As such a more detailed and personalized service and improvements in marketing shall be made possible.

The employed segmentation approach spotted 4 different clusters. The clusters show their highest difference in *Age*, *Average*LeadTime, *LodgingRevenue* and *PersonNights*. Using these features, the customers were presented in well distributed clusters, with approximately 15,000 in the smallest segment, and 20,000 in the biggest.

Figure 1 visualizes the average distributions across the four defined clusters. The y-axis values have been scaled to capture a range of measures within the feature space. It is important to note that the presence of negative values on the scale does not inherently indicate negative quantities, due to the robust scaling. All features contain only positive values in their initial state.

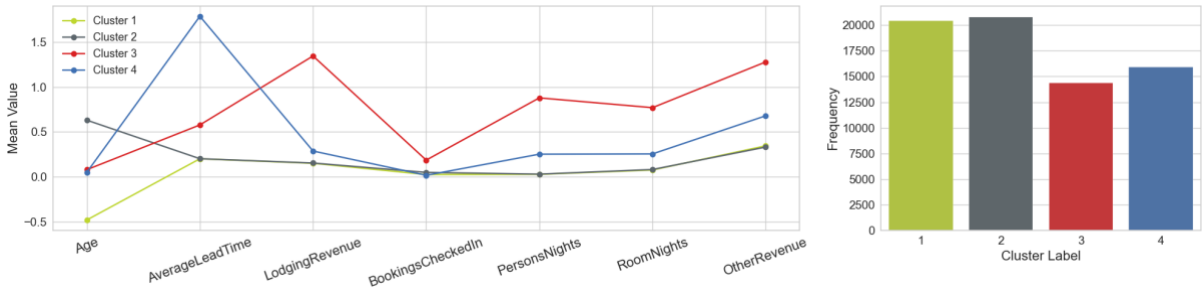

Figure 1 - Average feature values of clusters.

The subsequent section will lay out the unique attributes and distinguishing features of the four clusters, underscoring their differences in characteristics.

Young Travelers. Cluster 1 contains the youngest customers of the hotel which book their stay relatively shortly before arriving. Their average age is 36 years, and they usually bring slightly less revenue than customers from the third segment. Because the feature *RoomNights* is low compared to other segments, we assume they have shorter stays. This assumption holds throughout this report.

Senior Travelers. The second cluster shows similar characteristics to the first segment apart from being predominantly made up of an older demographic. All other averaged attributes – also distribution of nationalities and season of the signups – are very closely different from the Young Travelers. The

similarity of the numerical features is clearly visible in Figure 2, where the distribution among the features is shown.

Most Valued. Cluster 3 consists of customers which are mostly middle aged (around 49) and book their stay a common number of days in advance of their arrival (lead time). Compared to the other clusters they have the highest lodging revenue and therefore spend the most for their accommodation. Additionally, they are characterized by around 18% higher frequency for the number of check-ins such as a longer stay with more companions.

Early Birds. Cluster 4 is characterized by customers with an average age of 48 years. The customers that belong to the fourth segment book their stay in the hotel earlier than all other clusters in advance of their arrival. Furthermore, these customers show a slightly higher average of lodging revenue and total of *PersonNights* spent at the hotel compared to the clusters 1 and 2. Compared to the other segments where German customers make up around 10% of the customers, the early birds consist of around 25% Germans, who on average spend more for their stay.

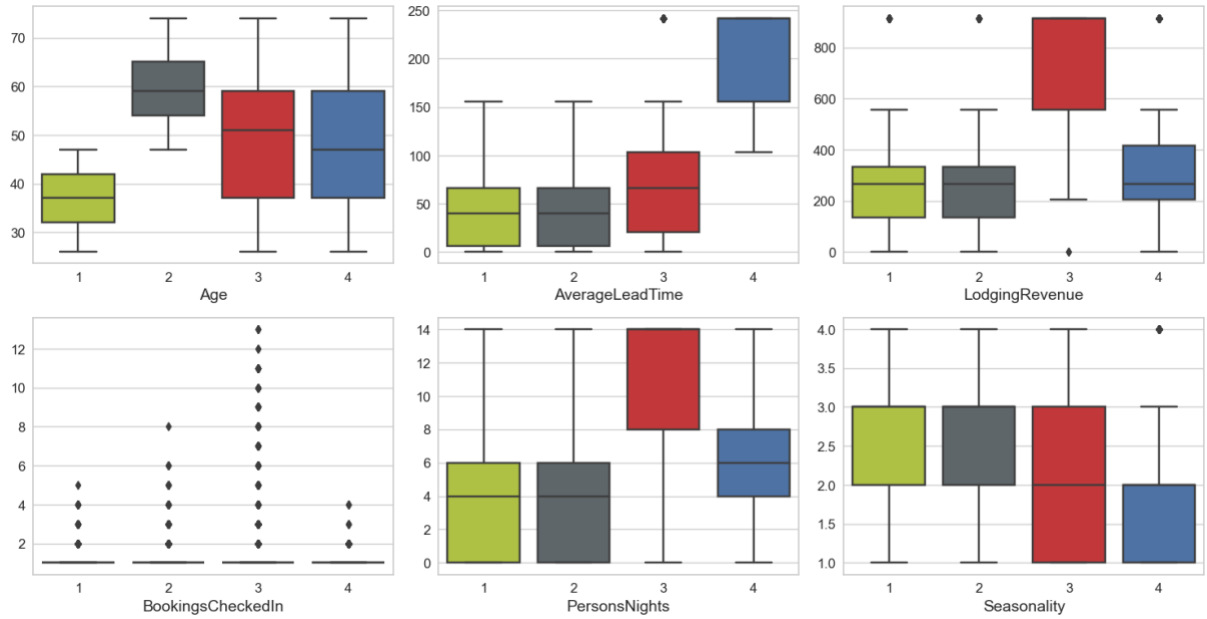

Figure 2 – Feature distributions across clusters.

6. Recommendations

This section presents an ensemble of business opportunities arising from the identification and examination of the clusters. The insights are focused on targeted marketing strategies and action items that aim at enhancing key metrics critical to customer retention and loyalty. These include the ratio of new bookings and overall lodging revenue. Leveraging strategic segmentation and a data-driven approach, the hotels performance can be boosted in key areas.

In alignment with the business objective, its economically more efficient to increase the revenue generated from existing customers than acquiring new ones. For the proposed customer segmentation, this means focusing engagement on two different clusters that have the highest potential for reservation and tend to spend the most money. Thus, the focus should concentrate on the high-value customers who generate the most revenue for the hotel. In addition, the focus should

center on the early birds, which are the most clearly differentiated from the other segments and offer potential for highly targeted strategies. The following recommendations relate to all the clusters mentioned, but it should be emphasized that clusters 3 and 4 should be given higher priority than clusters 1 and 2.

The Youths. For cluster 1, the youngest and spontaneous bookers, we propose launching dynamic last-minute deals to leverage their preference of short-term planning and increase bookings during off-peak seasons. Furthermore, by communicating through mobile apps, younger customers could be provided with exclusive in-app offers to enforce re-visiting the hotel. By partnering with travel influencer promoting the hotel during their stay, the hotel could approach young customers once again.

The Experienced Travelers. Cluster 2, which mirrors the spontaneity of Cluster 1 but consists of an older demographic, could be leveraged by offering comfort packages that emphasize on relaxation and cultural experiences. Additionally, the marketing team could introduce rewards for repeated stays such as special offers for relaxation experiences and room upgrades. If filling empty rooms with spontaneous bookings yields higher revenue than the marketing effort required, it should be focused on experienced travelers or youths, who are the most likely to book spontaneously.

Most Valued. For cluster 3, the most valuable middle-aged segment, which is known for higher spending and longer stays, we propose to introduce special offers for extended visits, such as complementary nights or exclusive services after a certain stay duration. Additionally, recognizing this segments tendency to travel with companions, offering packages which include discounts for multiple rooms or group bookings could engage these customers to revisit the hotel.

Early Birds. Cluster 4 is characterized by their planning and early booking habits. Therefore, we propose Early Bird specials that offer attractive discounts for reservations made well in advance. This approach rewards their proactive planning and assists in improved managing of occupancy. To further deepen the loyalty and engage early booking customers, the hotel could offer exclusive previews of upcoming features for booking before a certain point in time. The Early Birds show a high proportion of two seasons were around 80% of the customer segment signed up. Subsequently, it could be highly beneficial to offer discounts for booking a long time in advance as these customers tend to do that.

To generally elevate customer loyalty and satisfaction, the hotel´s strategy could focus on delivering personalized offers to each customer segment through digital channels like mobile apps and email marketing. Customizing these promotions according to the preferences identified across segments 1 to 4 ensures a personalized approach that resonates with each group´s needs.

Around 33 thousand of the provided entries are created accounts that never experienced a stay at the hotel. These show a high potential for being future clients, but more data collection at the point of the creation is essential to enable deeper analysis and better segmentation for future personalized marketing efforts. For now, using the individual´s email as a direct channel with effective campaigns and welcome offers could motivate them to make their first booking at the hotel.

7. Conclusions

In conclusion, this research demonstrated the significant benefits that clustering and targeted marketing strategies offer to Hotel H. Through a more sophisticated segmentation and the usage of behavioral, demographic, and geographic factors in addition to sales origin, the hotel can acquire deeper insights into its customer. With the implementation of the k-means clustering algorithm, four distinct customer segments with distinct characters and booking behaviors have been identified. With these insights, customized marketing strategies are proposed to the hotel. These strategies include the introduction of dynamic pricing, personalized offers for each segment´s preferences and leveraging digital channels for more effective communication. Furthermore, the recommendation to prioritize on existing customers is supported by evidence indicating that retention is markedly more cost-effective than acquisition, thus offering an opportunity to enhance Hotel H´s currently poorly existing customer loyalty. In conclusion, incorporating these suggestions into Hotel H´s operation is expected to improve the marketing efficiency, revenue growth and customer loyalty.

8. References

Gallo, A. (2014, October 29). The value of keeping the right customers. Harvard Business Review. https://hbr.org/2014/10/the-value-of-keeping-the-right-customers

McIlroy, A., & Barnett, S. (2000, December 1). Building customer relationships: Do discount cards work?. Managing Service Quality: An International Journal.

9. Appendix

Appendix 1: Lodging Revenue and Count of Individuals by Selected Nationalities.