Capstone Three: Final Project Report *Javier Nuno*

01/07/2021

Hotel Customer Segmentation

Problem statement

- *How can we group hotel customers to create targeted marketing strategies and tailored services? The case of an urban hotel in Lisbon, Portugal.*

Customer segmentation is the marketing process of separating your customers into groups based on certain standard features and is a precursor to targeting and positioning. Customer segmentation becomes critical to address successful marketing strategies and create a compelling service portfolio.

Effective customer segmentation should be:

- Differentiable: customer categories should be mutually exclusive,
- Measurable: the size of the target segment should be meaningful and delimited,
- Accessible: customer groups should be addressable,
- Substantial: they must be profitable and with growth potential,
- Actionable: the company should have a compelling message and product portfolio for them
- Compatible: customer groups should share the company's values and goals.

Historically, customer segmentation has been performed by sorting and classifying customers by standard features or behavior patterns. Clustering algorithms can unveil underlying trends and associations hidden from the classic analysis and help create a more effective customer segmentation.

This project presents the customer/guest segmentation for an urban hotel in Lisbon, Portugal.

Objective

The present project's objective is to find patterns in the customer booking behavior or preferences using unsupervised machine learning techniques so that it is possible to label customers into 4-6 segments.

Methodology

The Dataset

We have used a dataset containing records of bookings of a city hotel, with 31 variables describing 79,330 observations between the 1st of July of 2015 and the 31st of August 2017[1].

The steps followed in this project are:

1. Data wrangling

1.1. Data collection: I selected the original dataset's appropriate features that represent the behavior and patterns of the customer's reservations.

1.2. Feature extraction: I created some new attributes from the existing variables that contain relevant insights about booking references.

1.3. Feature analysis and profiling: I conducted a thorough investigation to understand each variable's distribution and missing values.

1.4. Quality assessment and data cleaning: The missing values were imputed with meaningful data when possible or removed if the observation could distort the result.

2. Exploratory data analysis (EDA)

2.1. Feature transformation: I will carry out a series of category regrouping in some features. The objective is to reduce the attributes' cardinality, so when they are transformed into numerical variables, we do not end up with a highly sparse dataset.

2.2. Exploring associations among categorical variables: Most of the variables are categorical, so we cannot use a Pearson's correlation matrix to explore their relationship. Instead, I used Cramer's V algorithm to assess the association between categorical attributes. Cramér's V (sometimes referred to as Cramér's phi) is a measure of association between two nominal variables, giving a value between 0 and +1 (inclusive). It is based on Pearson's chi-squared statistic and was published by Harald Cramér in 1946[2]

2.3. Pairwise exploration of correlation in selected variables: I compared selected categorical variables versus other numerical variables to explore the correspondence between them.

3. Preprocessing

3.1. Sample selection: In this step, I took a reduced sample of the whole dataset for computation feasibility purposes. There are ~80,000 observations, making it very time-consuming to run different tests with different clustering algorithms. Therefore, our approach will select 20% of the rows in the dataset, stratifying them by the variable *ArrivalDateMonth* to get the most representative sample possible.

Similarly, there are 37 variables in the original dataset, some of which are not relevant for the customer segmentation analysis. Those irrelevant variables are related to hotel processes rather than customer choices.

[1] *Hotel booking demand datasets*. Nuno, Antonio; De Almeida, Ana; Nunes, Luis. Data In Brief n22 (2019), Pag 41-49
[2] Cramér's V. (2020, April 14). In Wikipedia. https://en.wikipedia.org/wiki/Cram%C3%A9r%27s_V#cite_note-1

3.2. Categorical to cardinal transformation: In present section, we will encode all categories with numeric variables to prepare the dataset for the scaling step. We will assign a new variable for each category but one to avoid collinearity. We can see that we end up with 91 variables in a very sparse dataset.

3.3. Feature scaling: Finally, I unified and recentered the distribution of the variables by scaling the whole dataset to a mean of 0 and unit variance.

4. Modeling and results

4.1. Dimensionality reduction: This step entailed reducing features to allow faster computation without losing much information from those features eliminated. I used the Principal Component Analysis algorithm (PCA) to find synthetic attributes representing the most extensive variability in the data and contain the most valuable information.

4.2. Modeling with K-Means Clustering: K-Means Clustering was the first algorithm used to classify our observations. Initially, I defined the number of desired clusters I was later going to use as input to the model. I carried out three different methods to figure out a priori this number:

- The Elbow Method
- The Silhouette Coefficient Method
- Clustering observation with t-SNE

4.3. Modeling with alternative algorithms: In this step, I used other alternative algorithms to classify our samples:

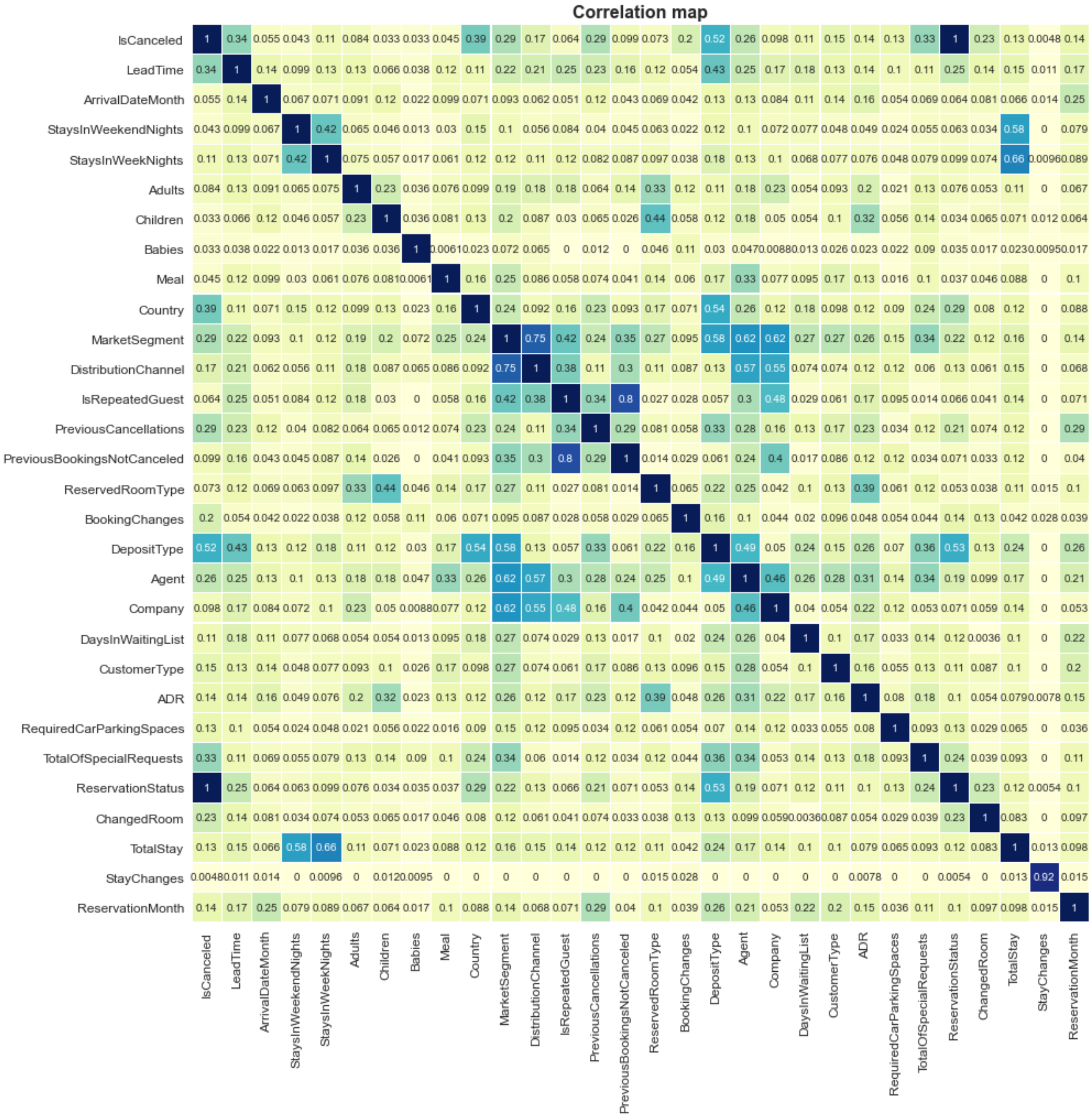
- Agglomerative Hierarchical Clustering
- DBSCAN Clustering
- Mean-Shift Clustering
- Expectation-Maximization Clustering

4.4. Visualize results: I plotted the dataset in 2D using PCA and labeled each sample with the results from the best performing model: K-Means clustering.

4.5. Customer segments profiling: Finally, I compared some statistics of the features grouped by cluster to understand the main attributes that differentiate the customer segments and define a profile for each of them.

Analysis

Association matrix

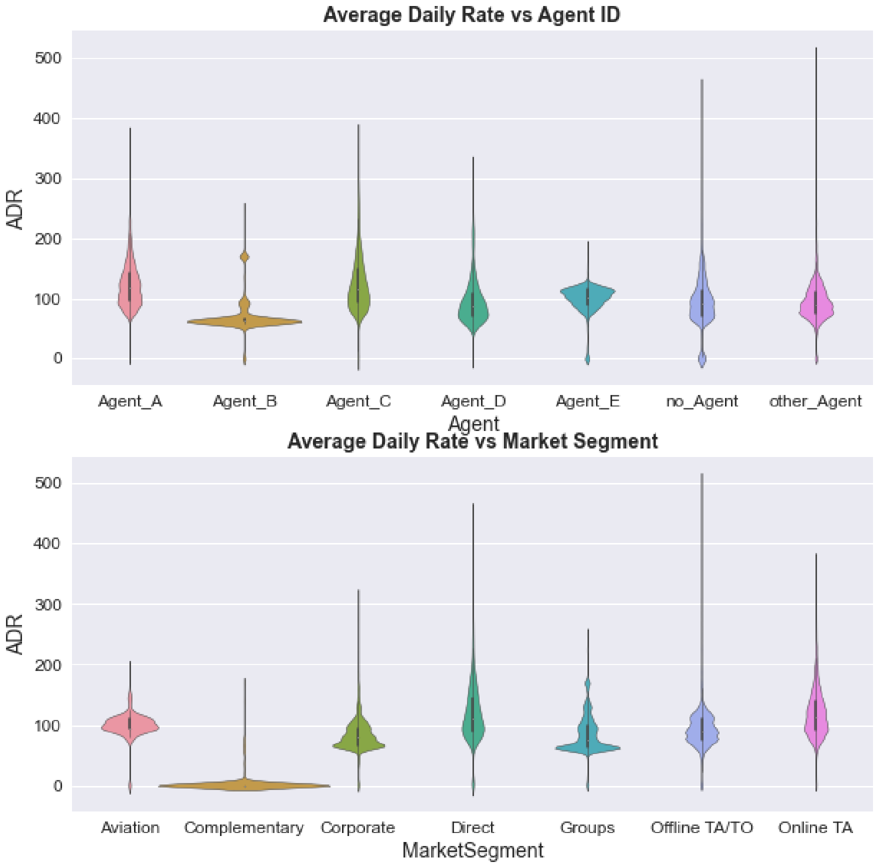


Graph 1 - Feature correlation map

In graph 1, we can observe that in general, there is no association among the features in the dataset, except for some exceptions:

- *IsRepeatedGuests* and *PreviousBookingsNotCancelled* appears correlated with a 0.8 Cramer's V ratio
- *DistributionChannel* and *MarketSegment* are also associated with a 0.75 ratio, making sense since both are related to sales channels. Also, these features are moderately correlated with the sales *Agent* and the *Company* making the reservation.

- *ReservationStatus* and *IsCancelled* are completely correlated, and the information they contain is wholly duplicated.
- Finally, another interesting association is *Country* and *IsCancelled*, meaning that guests from certain countries are more likely to cancel their reservation.

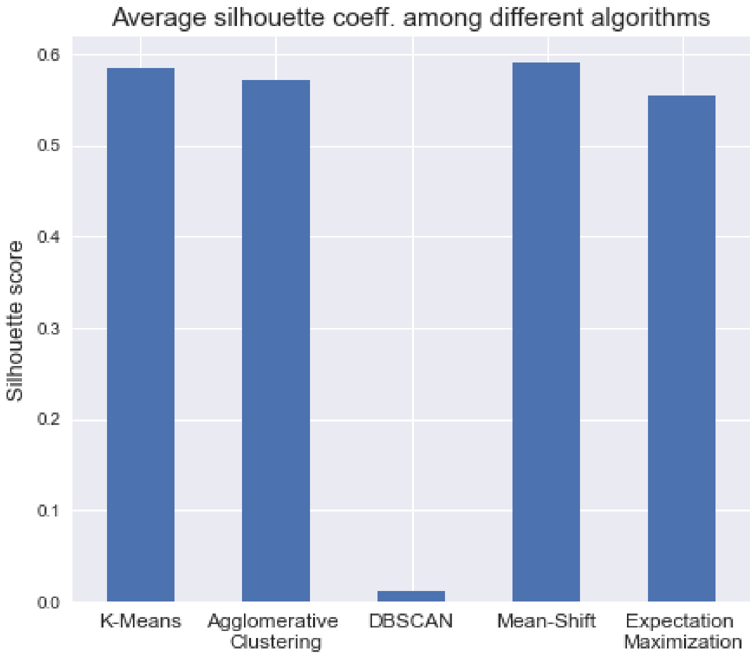


Graph 2 - Average daily rates vs. sales agent and market segment

In the graph 2, we can see that reservations made through different agents had in general different average daily rates. For example, Agent B accounts for the distribution with the lower rates with a median value of EUR 80, whereas Agent C had the bookings with the higher price range averaging ~EUR 100.

Regarding Market Segment, Corporate and Groups enjoyed the lower rates with roughly EUR 70 on average, contrasting with Online and Direct reservations which had EUR 100 rates on average.

Findings



Graph 3 - Silhouette score for various models

Graph 3 shows that Mean-Shift produced the best silhouette score with 0.59. Still, the algorithm selects only three clusters, which is inadequate for the customer segmentation - we need at least four for the segmentation to be useful and accurate. K-Means is the second-best algorithm with a 0.58 average silhouette score.

It is worth noting that DBSCAN performed poorly with this dataset, which is expected since our dataset is very sparse, and this algorithm works better for complex spatial structures but tighter datasets.

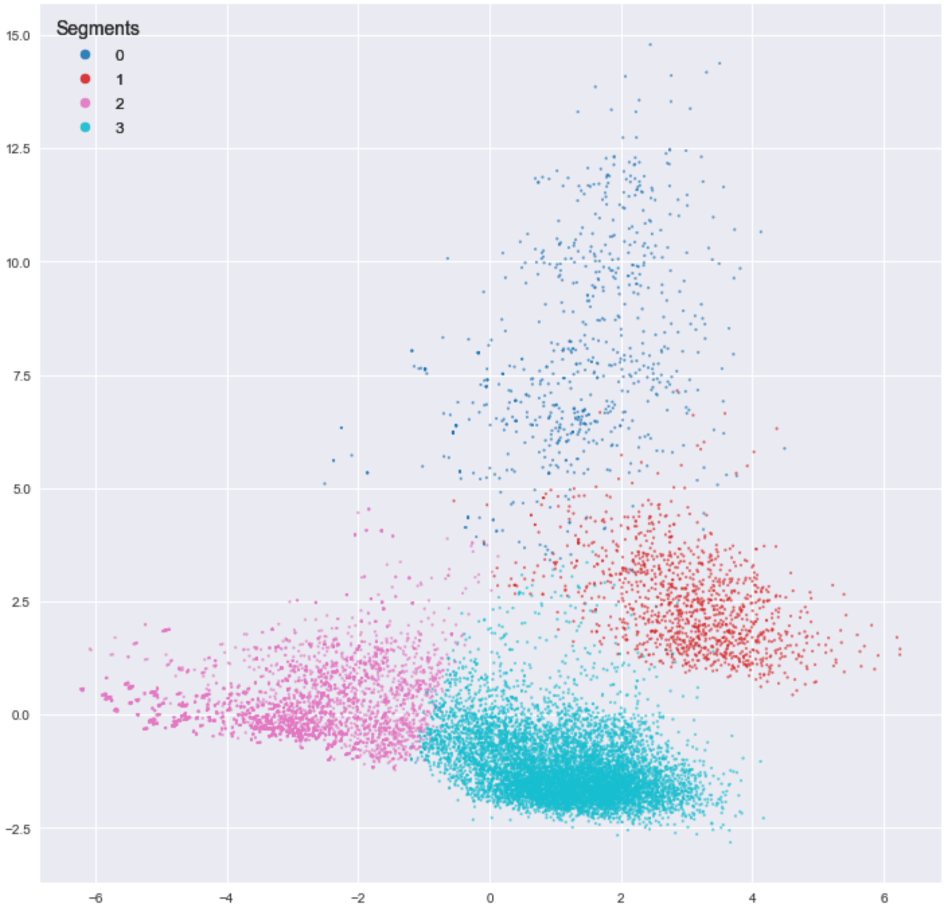
Table 1 below from an article of the University of Berkeley[3] shows that all algorithms except DBSCAN found some group structure in the data, showing that there are some customer segments with similar behaviors and preferences. However, the structure is not extraordinarily strong, and it is necessary to visualize it further to understand these figures.

Range	Interpretation
0.71 - 1.0	A strong structure has been found.
0.51 - 0.7	A reasonable structure has been found.
0.26 - 0.5	The structure is weak and could be artificial.
< 0.25	No substantial structure has been found.

Table 1 - Silhouette score interpretation

[3] https://www.stat.berkeley.edu/~spector/s133/Clus.html

Conclusion



Graph 4 - Scatterplot of samples

The present study has shown an underlying segment structure in the data, and it is possible to establish four well-defined customer groups with similar characteristics. However, not all these groups have the same consistency and cohesion. It can be seen in the scatter plot 4, samples from segments 2 and 3 are closer, meaning that their attributes are more homogeneous than in segments 0 and 1, which can even be considered to not forming a group.

A customer profile based on the segments is presented below:

- Customer Segment 0: National business guest who arrives in May, stays one night. This client usually makes the reservation a week in advance, has a contract with the hotel, and pays a discount rate.
- Customer Segment 1: Tourist national guest who arrives in August and stays one night. He/She usually reserve three weeks in advance directly with the hotel or uses the Agent C, meaning that he/she has done previous research or is a returning customer. This client is the least sensitive to prices and is willing to pay a high rate.
- Customer Segment 2: National group guest who arrives in September and stays two nights, usually makes the reservation through a travel agency with 5-6 months in advance. This guest is very sensitive to prices and is very likely to cancel the reservation.

- Customer Segment 3: Tourist guest from a European country who arrives in August and stays two nights, usually makes a reservation two months in advance through an online agency (Agent A). This customer is not sensitive to prices, so he/she is willing to pay more for the room or ancillary services.

Recommendations

The hotel management can use this study to confidently segment their clients into four groups as a starting point to understand its customers. They can also use the segment profile to target marketing campaigns, define the portfolio of services, and refine the pricing strategy.

However, the profile of these segments is incomplete and may require delving into their inherent characteristics. Namely, attribute values of segments 0 and 1 vary significantly among them, and it is critical to discriminate between them.

Further research

In the hotel industry, it is quite common for customers to change their booking attributes either at the time of their check-in or during their stay. Understandably, the distribution of some variables differs between non-canceled and canceled bookings. Consequently, the use of these datasets may require this difference in distribution to be considered for further studies.

This study can be complemented with a detailed breakdown of each customer segment's profile, and further split them in smaller groups, especially for the most numerous segments (segments two and three).