Team 3 Final Report

8th March 2021

Selma Sentissi El Idrissi	selmasen@bu.edu
Shuyi Zhu	shuyizhu@bu.edu
Jaya Nagesh	jnagesh@bu.edu
Shamika Kalwe	shamika@bu.edu

The problem:

The problem we are trying to solve is to understand the behavior of hotel customers and group them accordingly. This way, we can better advise the hotel owner on which groups of customers are worth pursuing in order for hotels to remain profitable. We will give hotel owners better insights on how to market and run promotions.

This is especially important in today's era with the rise of Airbnb and other housing platforms, hotels are no longer one of the only lodging options. This dataset about a hotel in Lisbon, Portugal is a good way to understand and get good insights on past customer's behavior in order to predict and target potential future ones. This dataset is ideal since it comprehends three years of customer behavior, geographic, and demographic information.

In addition, as this dataset is customer centered it will give us insights into Customer Analytics which is a prominent use case of Analytics in today's business world.

The Data Set:

Hotel's Customers Dataset
https://www.kaggle.com/nantonio/a-hotels-customers-dataset with its clarification on https://www.sciencedirect.com/science/article/pii/S2352340920314645?via%3Dihub The dataset contains 83,590 records of customers and 31 columns. Each variable (column) represents a characteristic or description of the customer. In addition to personal and behavioral information, the dataset also contains demographic and geographical information.

Methodology:

Association Rules

We applied the association rules to explore and understand the hotel preference of customers. We considered the association between travelers' hotel booking behavior and the demographic factors and hotel attributes. In this practice, we focused on two things in association rules: 1) understand which customers are more likely to return to the same hotel for repeated stays, therefore holds substantial value to the hotel. 2) understand the stay length patterns of customers, and provide possible suggestions to increase hotel profits.

First of all, since our original dataset is mostly numeric value-based, we selected several attributes (age, average daily price, nationality, repeated bookings, length of stay, and distribution channels), renamed the numeric value in label groups, and converted them into category type.

In the first part of association rules mining, we specified the consequence control in repeated bookings; for the antecedents control, we selected Age, Nationality, and Length of Stay. We first applied one antecedent to one consequent methodology in finding the association

between the consequence and each antecedent (e.g. {Age Group = Adults 40-59} => {Repeated Bookings = Book 2 Times}). Apriori generated all the possible rules for customer loyalty, yet we filtered the table by only looking at the rules with lift value >1. So we are sure that the two occurrences are dependent on one another, and the rules are potentially useful. After linking the associations for each antecedent, we combined all three antecedents (Age Group, Nationality, and Length of Stay) to predict the consequence of Repeated Bookings. Below is the association rules table (partial) generated (for consequence = {Book 3 Times and above} and {Book 2 Times} separately):

antecedents	consequents	support	confidence	lift
(Adult 40-59, Holiday)	(Book 3 Times and above)	0.034837	0.211394	1.613117
(Adult 40-59, Others Countries)	(Book 3 Times and above)	0.054751	0.176771	1.348911
(Adult 40-59)	(Book 3 Times and above)	0.086574	0.172611	1.317166

antecedents	consequents	support	support	confidence	lift
(Holiday, Senior 60+)	(Book 2 Times)	0.641301	0.054603	0.838392	1.307330
(Weekend Trip, Senior 60+)	(Book 2 Times)	0.641301	0.089687	0.795007	1.239679
(Weekend Trip, Others Countries, Senior 60+)	(Book 2 Times)	0.641301	0.056530	0.794444	1.238802

We then conclude that customers aged from 40-59 and 60 above are likely to rebook the hotel 2 and more than 2 times. What's more, customers in these two age groups are likely to stay longer as they are more likely to book for holiday or weekend trips.

Moving on to the second part of association rules mining, we specified the consequence control in the length of stay; as for antecedents, we selected Age, Nationality, Average Daily Price Group, and Distribution Channel. In this part, we mainly focus on one antecedent to one consequence associations. Below is one of the association rules table (partial) generated (for consequence in Length of Stay, antecedent in Average Daily Price):

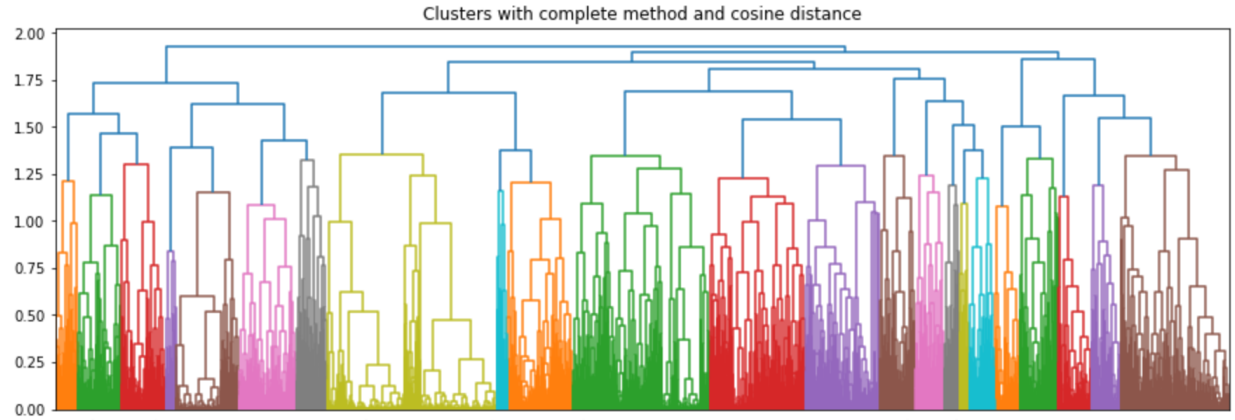
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
8	(Single Day)	(High)	0.178732	0.022088	0.006918	0.038706	1.752339	0.002970	1.017287
9	(High)	(Single Day)	0.022088	0.178732	0.006918	0.313199	1.752339	0.002970	1.195788
6	(Medium)	(Single Day)	0.157434	0.178732	0.040470	0.257062	1.438255	0.012332	1.105433
7	(Single Day)	(Medium)	0.178732	0.157434	0.040470	0.226431	1.438255	0.012332	1.089192
0	(Holiday)	(Low)	0.325443	0.820477	0.280872	0.863043	1.051879	0.013853	1.310794

The most interesting findings among the four association rules tables are: 1) seniors are most likely to book the hotel for weekend trips and holidays; adults aged 18-59 tend to book for single trips more. 2)As for nationality, German and French people are likely to book for holiday purposes; Portugal and Spanish people visit Single Day stay, and British is likely for weekend trips. 3) As for the average daily price, the single night booking has a higher average price, whereas the holiday has, on average, the lowest daily price. 4) In terms of the different distribution channels, most single-day bookings are ordered via corporate or direct channels. In contrast, people who plan to stay on holiday or weekend tend to book from Travel Agents.

Therefore, by joining the two perspectives' findings, we suggest the hotel consider providing coupons/rewards to customers aged from 40 above when they book for holiday or weekend trips because these are the potential loyal customers who are highly likely to rebook several times. With the coupon options that lower the average price per night, more previous customers will likely book the hotel again. On the other hand, the hotel could consider raising the average price per stay for customers who book for only one night's stay.

HCluster

We first try to choose which distance we would use by computing the Euclidean and the Cosine distances. Following their visualization, we could not choose which one would be the best to evaluate the models so we then tried to compute all methods with both. First, we used hierarchical clustering with the 4 main methods: single, complete, average, and ward which were all computed with Euclidean distance. Only the ward method gives us visible clusters. We can count 7 clusters in this case. Next, we tried to use cosine distance. Even though the results do seem better, these are still not the most accurate. We can count less than 3 clusters for the single method, 22 for the complete method, and 11 clusters for the average method. The ward method, which seemed the most promising with the Euclidean metric, does not compute at all with the cosine metric.



The best model on the H cluster was the complete method with cosine distance. However, we concluded that this model was not the best for our study, thus, we did not go further into the analysis.

KMeans

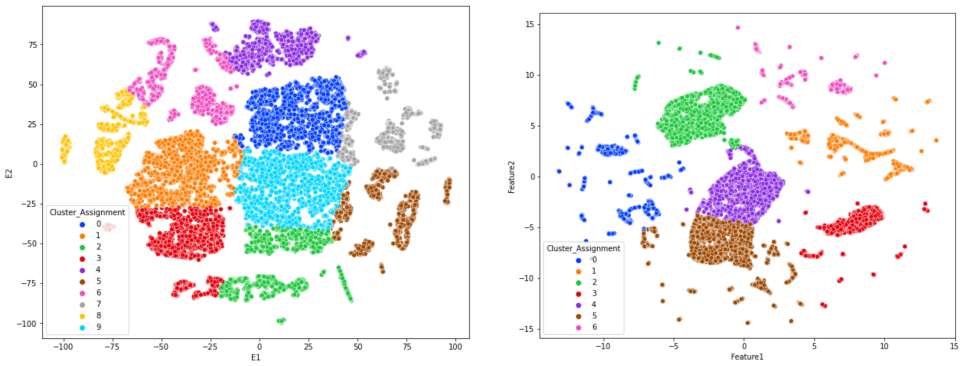
First a KMeans baseline was established. This baseline was just executing the KMeans algorithm without any dimensionality reduction techniques. The number of clusters was selected based on minimizing inertia and maximizing silhouette score as well as the business problem of hotel customer segmentation. 13 clusters were selected as the optimal, however the clusters turned out to be imbalanced. Six out of 13 clusters have less than 50 observations with three of those six having observations less than 10 observations. From the hotel perspective, it could potentially be difficult for a hotel to formulate their strategy based off of less than 10 customers being in a group as those customers could also be outliers/rare cases.

Next we started exploring dimensionality techniques. This overall pipeline is as follows: apply dimensionality reduction technique, obtain inertia and silhouette plots, select number of clusters, and then run the KMeans algorithm. The first dimensionality reduction technique explored was PCA. Our goal was to try and maintain 95% of the variance in the principal components. Based on the cumulative explained variance graph it seemed that 19 principal components were optimal. However the eigenvalue bar plot that was constructed depicted that around eight components would be optimal. Considering the trade-offs, we decided to take 14 principal components (around 83% of the variance explained) and perform clustering. We ended up with 10 clusters. Similar to the previous example, the clusters are still imbalanced however not as badly as just the KMeans without any dimensionality reduction.

Lastly, we explored TSNE and UMAP. For the TSNE and KMeans model we choose ten clusters and for the UMAP and KMeans model we chose seven clusters. From the plots of the features and the clusters, in

TSNE we can see that the clusters are overall less dense/compact (however still compact) overall than UMAP where a few clusters are very tightly packed/dense and other clusters are very spread apart. From the practical perspective, TSNE followed by KMeans will be a good model to pursue for our analysis since clusters are relatively compact and there seems to be a decent number of observations in each cluster, hence more balanced clusters than UMAP.

TSNE with KMeans clusters (left) vs UMAP with KMeans clusters shown below (right):



DBSCAN

Setting up DBSCAN models requires two key parameters eps and min_samples. Below we have the definition of these parameters and the methods we use to select them.

- eps: It is the maximum distance between two samples for one to be considered as in the neighbourhood of the other. We got a range for epsilon (0.5 to 1.5) by plotting NearestNeighbours distance curve and getting the elbow point. Then we iterated through this range and got the epsilon which maximizes the Silhouette score.
- min_samples: We use the eps from above and the number of clusters generated in KMeans (or desired clusters) method as reference to select appropriate min_samples.

However, one of the issues we faced with this method was, the silhouette score for the eps range was very low and on increasing eps beyond the range we were able to maximize silhouette score further while making the desired number of clusters. Another problem was that this method required dependency on the number of clusters generated in KMeans method. As we were happy with the clustering results of KMeans we did not see the value in continuing with the DBSCAN analysis.

Analytical findings, conclusions and recommendations as they relate to your business problem

We decided that the best model to use for analysis is the KMeans with tSNE and PCA. From that model, we gathered some observations:

Age: The fourth cluster has the highest age mean with people averaging almost 52 years old whereas the 6th cluster has the minimum age average with people's age mean being 34 years old. Also, clusters 1, 8, and 9 all have an average mean of 48 years old.

Days Since Creation: People from cluster 8 have created their hotel account more than 2 years ago whereas people from the 4th cluster have on average a 6 months old account.

Average Lead Time: People from the 4th cluster made their reservation on average less than 2 days before their arrival day whereas people from the third cluster have made their reservation 3 and a half months in advance on average.

Lodging (Room) Revenue & Other Revenue & Persons Night & Room Nights: The 4th cluster has the lowest lodging revenue with 10.23 Euros per stay on average as well as the lowest other revenue expenses

with 2 euros, whereas the 1st cluster has the highest lodging revenue with more than 650 Euros per stay by customer as well as the highest other revenue expenses with 159 euros. However, the 1st cluster also has almost 9 persons per reservation as well as the highest number of nights per reservation whereas the 4th one only has a fraction of 1 on average for both variables.

Booking Checking rate: This variable is the number of bookings the customer made that end up with a staying which can be defined to valid bookings. The highest valid booking rate is 1.2 and is detained by the 7th cluster whereas the 4th cluster has once again, the lowest valid booking rate, which is equal to almost zero. This could explain the 4th cluster's low scores on lodging and other revenue.

Customer Requests: Customers from the 6th cluster have requested the highest floor the most amongst customers from all clusters when they make their reservation whereas the 3rd cluster has some customers requesting the lower floor. Only the 0th cluster has ever requested an accessible room amongst all clusters. The 6th cluster has a tendency to request cribs. Nine out of ten clusters have some preferences for king size beds and 8 out of ten have also some preferences for twin size beds. Also, none of the clusters requested no alcohol in the minibars. Finally, the 8th cluster has 87% of its people requesting a quiet room.

Additional Note: Clusters 2 and 5 were not mentioned since they do not have any of the lowest nor the highest scores on any of the variables. They are average and blend perfectly amongst all the clusters.

Overall findings

- Among the 10 clusters, the average age of overall customers is around 48.
- More than 3 clusters have customers who created their account more than 2 years ago.
- Among the 10 clusters profile, no cluster shows the value 1 in Cancelling/ Not Showed for the reservation.
- With our analysis of lodging revenues and room nights, we realized that the people who spent the least were the people who booked their stay the closest to the actual date.
- People who booked their stay the closest to the actual date were also the most likely to not check in with their reservation afterward.
- The most frequent requests: asking for a king size or a twin size bed, as well as for a quiet room. All the other requests were not used a lot on our sample.

Recommendations

- Since the average customer is middle aged, the hotel should target these people especially (Adults 40-59 are most likely to rebook).
- The hotel greatly depends on loyal customers; it is reasonable to suggest the hotel to provide more rewards to attract and maintain these previous customers.
- Since people did not show a pattern of canceling or not showing their reservation, this means that the hotel does not have to adopt a credit card mandatory fill form at the moment of the reservation.
- Adjust the price/night based on association rules, provide a higher daily price for customers booking only for 1 night, but a lower price for customers booking for holiday trips (Could add some coupon for customers booking from Travel Agent/Operator).