Team 3 Proposal

21st Feb 2021

Team members:

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

(a) 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.

(b) 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.

Some relevant attributes (columns in the data set) we are planning to focus in our analysis are:

Column Name	Data Type	Description
Age	Numeric	Customer's age (in years) at the last day of the extraction period.

LodgingRevenue	Numeric	Total amount spent on lodging expenses by the customer (in Euros). This value includes room, crib, and other related lodging expenses.
BookingsCancelled	Numeric	Number of bookings the customer made but subsequently canceled (the customer informed the hotel he/she would not come to stay).
BookingsCheckedIn	Numeric	Number of bookings the customer made, and which end up with a stay.
BookingsNoShowed	Numeric	Number of bookings the customer made but subsequently made a "no-show" (did not cancel, but did not check-in to stay at the hotel).

(c) Analysis Methodology

(i) We have some null values (4%) in the 'Age' column which we will have to either impute. We have some outliers like negative values for 'Age' 'AverageLeadTime' columns etc which we will have to look into.

(ii) The column "NameHash" and the column "DocIDHash" both have values that are not interpretable so we will drop them from the dataset as part of our data cleaning process.

(iii) Exploratory graphs to get comfortable with variables

(iv) We will generate the correlation matrix to see if there are any correlations in the input variables. For e.x. We have columns like SRHighFloor, SRLowFloor which probably would be highly correlated. This would set the base for our next step of PCA

(v) We will then use PCA to reduce our feature space since there are 31 columns. We will target to maintain about 90% of the variance within the dataset through the PC components.

(vi) For clustering we will first try the Hierarchical model and get some insights into the natural layout of the customer base. We will try different distance metrics (Euclidean and cosine) as well as methods (complete, single, average, ward linkage) and see which best suits the data

(vii) Next we will try KMeans clustering. We will evaluate our model looking at inertia and silhouette scores. We aim to pick the optimal value of clusters that minimizes inertia and maximizes PC.

(viii) We also will take into account business considerations. For example, we cannot have a very high number of clusters as it will be impractical and costly to have many customer segmentations.

(ix) Conduct profiling (STAT summary of customers per cluster, e.g. mean number of night_stayed)

(x) Visualizations to present key findings