Group Project (Clustering)

BookMe

MACHINE LEARNING 2021 / 2022

May 4, 2022

1 Introduction

Welcome to the BookMe company. This organization is a well-established company operating in the hospitality sector. BookMe provides accommodation to tourists and travellers, delivering necessary lodging services to those who travel the world, whether for leisure or business. It provides an international website where citizens can book their accommodation. Presently they have around 30,000 registered customers and serve more than 100,000 consumers a year. The website offers a variety of services, but they are focused in providing rooms with the best conditions possible. In order to control the quality of the services, every time a client makes a reservation, at the end of the stay, a survey is sent to complete on how the guest perceived the provided services. A scale of 0 to 5 is used to rate multiple aspects of the services, in this way, customers can reveal how satisfied they are regarding location, price, amenities provided, and others.

Globally, the company had stable revenues and a healthy bottom line in the past three years, but the profit growth perspectives for the next three years are fickle. A few strategic initiatives are being considered to invert the situation. One of those is a Marketing efficiency program to improve marketing activities, focusing on boosting the marketing campaigns' efficiency tremendously.

1.1 At the Marketing Department

The marketing department is under pressure to spend more wisely its annual budget. The CMO knows the importance of having a more quantitative approach to marketing decisions. The department requested a small team of 5 data scientists (your group) with a clear objective in mind: try to cluster the different types of customers that the company has to create more efficient campaigns. Desirably, these activities' success will prove the approach's value and convince the more skeptical within the company.

2 Objective of the project

The team's objective in this project is to identify actionable segments within the company's Customer base. These segments must be determined by looking at data available and through the usage of quantitative techniques. A priori, two visions are considered essential – the customer satisfaction and customer characteristics. Nonetheless, other perspectives will be valued. This project's main output will be a report that identifies the main customer segments and a first draft of a marketing plan.

3 Datasets

You have access to one dataset: a csv file with historical data corresponding to 15589 customers. The dataset file is the same used in the predictive project for training (train.csv).
The data contains the following attributes:

Attribute	Description
Cust_ID	Customer's identification number
Name	Customer's name
Year_Birth	Customer's birth year
Longevity	Whether the customer registered more than 1 year ago or not (yes or no)
Churn	Whether the customer churned or not (churn or nochurn)
TypeTravel	Customer's reason for travelling (business or leisure)
RoomType	Type of room reserved
RewardPoints	Customer's rewarding point for loyalty
Comfort	Satisfaction level of customer regarding comfort of the room (0 to 5)
ReceptionSchedule	Satisfaction level of customer regarding reception schedule (0 to 5)
FoodDrink	Satisfaction level of customer regarding food and drink available (0 to 5)
Location	Satisfaction level of customer regarding accommodation location (0 to 5)
Wifi	Satisfaction level of customer regarding wi-fi service (0 to 5)
Amenities	Satisfaction level of customer regarding accommodation amenities(0 to 5)
Staff	Satisfaction level of customer regarding staff (0 to 5)
OnlineBooking	Satisfaction level of customer regarding online booking ease(0 to 5)
PriceQuality	Satisfaction level of customer regarding price quality relationship (0 to 5)
RoomSpace	Satisfaction level of customer regarding room space (0 to 5)
CheckOut	Satisfaction level of customer regarding check-out (0 to 5)
CheckIn	Satisfaction level of customer regarding check-in (0 to 5)
Cleanliness	Satisfaction level of customer regarding cleanliness (0 to 5)
BarService	Satisfaction level of customer regarding bar service (0 to 5)

4 Deliverables

1. A Jupiter notebook with all the needed code implemented to obtain the results presented in the report and to obtain the results explored in the report.
   The file naming format should be "202122_Cluster_GroupXX_Notebook.ipynb", where "GroupXX" should be your group number.

2. A report that describes the analytical processes and the conclusions obtained, with at most 8 pages:

- **Heading 1:** Arial, Size 12 pt, in bold
- **Heading 2 (if needed):** Arial, Size 11 pt, in bold and italic
- **Text:** Arial, Size 10 pt, line space of 1.5 points.
- **Margins:** The default ones in word (Top, Bottom, Left and Right as 1").

All the figures and tables should be included in the Annexes (at the end of the document) and referenced in the body text, and are not included on those 8 pages mentioned previously.
The reports that do not follow the specified conditions will suffer penalisation on the grade.
The file naming format should be "202122_Cluster_GroupXX_Report.pdf", where "GroupXX" should be your group number.

4.1 Notes

- We will evaluate all the topics mentioned based on the report - a well-structured and succinct report will have a big weight on the evaluation.

- The jupyter notebook will be analyzed only if some doubt arises during the report evaluation. If some steps were done in the Jupyter notebook but not described in the report, we will not evaluate those. As an example, imagine you check the outliers, and at the end of your project, you decide to keep them. In the report, you should mention how you check if you had outliers, what the steps were to remove them and why you decide to keep them at the end, among other insights that can be relevant. The jupyter notebook should be delivered with all the cells already run.
- The report and the code will pass through a process of plagiarism checking.

5 Evaluation Criteria

The following table quantifies the major evaluation criteria.

Criteria	Percentage	Maximum Grade (out of 20)
Report-quality and Story-telling	10%	2
Introduction and Methodology	5%	1
Exploration	10%	2
Pre-processing	15%	3
Modelling	15%	3
Description of Customer Segments	15%	3
Marketing Plan	10%	2
Conclusions	5%	1
PCA	2.5%	0.5
KNN Imputer or other imputation method	2.5%	0.5
Other Clustering technique	5%	1
Creativity & Other Self-Study	5%	1
TOTAL	100%	20

A project that focus only on the techniques and methodologies approached during the practical classes will have at most 17 values. The remaining 3 values are possible to achieve if contributions based on self-study and creativity are applied, and clearly explained on the report.

This bullet-list provides some details about each aspect:

- **Report-quality and Story telling:** Each report should follow the provided report structure and describe the steps and main insights along the process. Clarity, synthesis, objectiveness, and business-contextualization are very welcome. Your decisions and steps must be reasonably justified by the previous

findings (when this is possible and feasible), your hypothesis and findings must be related to the problem's business-context, etc..

- **Introduction and Methodology:** The introduction presents the general topic and main goal of the project. The methodology consists in the overall approach that underpins your work, and includes descriptions of the typical phases of the project.

- **Exploration:** Describe the studied population using statistical measures, business insights and visualizations representative of the major insights.

- **Pre-processing:** Includes all the needed steps to transform the raw data into the data prepared to cluster. Involves all the steps for cleaning, transform and reduce the dataset. It also involves the business-related transformations of the original input features and the explanation of those.

- **Modelling:** Implementation and reasoning behind any clustering model used in the project and addressed in classes. Two perspectives are obligatory: Customer perspective and product usage perspective. More perspectives are optional and considered as points in "Creativity and other self-study". You should explain the feature selection for each perspective for the sake of each perspective segmentation.

- **Description of Customer Segments:** Each segment for each perspective should be explored (statistically and visually) and described, focusing on the main aspects that differentiate each one.

- **Marketing Plan:** You should provide a succinct but well-oriented marketing plan that will answer the main insights obtained during clustering.

- **Conclusions:** Summarizes the key supporting ideas you discussed throughout the work.

- **PCA:** A theoretical explanation of the algorithm should be provided in the annex (not included in the 8 pages). This algorithm allow dimensionality reduction, and even if not used in the final solution, the application of PCA implies interpreting the results and the clear identification of the number of components to be used (and the process used to quantify the final number of components).

- **KNN Imputer or other Imputation methods:** A theoretical explanation of the algorithm should be provided in the annex (not included in the 8 pages)..

- **Other Clustering technique:** A theoretical explanation of the algorithm should be provided in the annex (not included in the 8 pages). Involves the depth and the quality of the comparative analysis between the clustering solutions provided by the different algorithms, the existence of a short comparative study between different customers' profiles obtained by different methods, etc.;

- **Creativity and Other Self-Study:** If other algorithms not given during classes are applied, a theoretical explanation of the algorithm should be provided in the annex (not included in the 8 pages). This topic includes not only the application of different techniques but also aspects of creativity, such as the creation of other perspectives during clustering besides the obligatory ones.

All topics are evaluated through a comparison of the work provided by the different groups.