Text

Description automatically generated

DATA SCIENCE AND MACHINE LEARNING

Supervised Machine Learning Group Project

**BookMe**

June 2022

**Data Science and Machine Learning**

**Supervised Machine Learning Group Project**

**BookMe**

***Group Members:***

Mariana Dias Gil, *nº 20211043*

Ricardo Daniel Pacheco De Almeida, *nº 20201335*

Soraia Andreia Anselmo Alves, *nº 20201840*

Table of Contents

[Table of Methodologies’ Description 5](#_Toc106290104)

[Table of Figures 6](#_Toc106290105)

[1. Introduction and Methodology 7](#_Toc106290106)

[2. Data Exploration 7](#_Toc106290107)

[2. Data Pre-processing 7](#_Toc106290111)

[3. Modelling 7](#_Toc106290112)

[4. Performance Assessment 7](#_Toc106290113)

[5. Conclusions 7](#_Toc106290114)

[6. Bibliography 8](#_Toc106290115)

[7. Annexes 9](#_Toc106290116)

[7.1. Figures 9](#_Toc106290117)

[2.1. Methodologies 23](#_Toc106290119)

# Table of Methodologies’ Description

[Methodology 1 – KNN 17](#_Toc100600227)

[Methodology 2 - PCA Analysis 17](#_Toc100600228)

[Methodology 3 – K-Means Clustering 18](#_Toc100600229)

[Methodology 4 – Hierarchical clustering 18](#_Toc100600230)

# Table of Figures

[Figure 1 - Summary of dataset's attributes and description 11](#_Toc100944649)

[Figure 2 - First three entries of BookMe dataset 11](#_Toc100944650)

[Figure 3 - Duplicated entries removed 12](#_Toc100944651)

[Figure 4 - Information on BookMe dataset 12](#_Toc100944652)

[Figure 5 - Descriptive statistics for numerical variables 13](#_Toc100944653)

[Figure 6 - Countplot of "RoomType" and "TypeTravel" 13](#_Toc100944654)

[Figure 7 - Countplot of "Longevity" and "Churn" 13](#_Toc100944655)

[Figure 8 - Skweness values 14](#_Toc100944656)

[Figure 9 - Kurtosis values 14](#_Toc100944657)

[Figure 10 - Boxplots for 'Amenities' and 'Staff' 14](#_Toc100944658)

[Figure 11 - Boxplots for 'PriceQuality' and 'CheckOut' 14](#_Toc100944659)

[Figure 12 - Boxplot for 'Cleanliness' 15](#_Toc100944660)

[Figure 13 - Missing values distribution (after transformation of underage people to missing) 15](#_Toc100944661)

[Figure 14 - 'Year\_Birth' descriptive statistics before imputation 15](#_Toc100944662)

[Figure 15 - 'Year\_Birth' descriptive statistics after imputation 15](#_Toc100944663)

[Figure 16 - 'Wifi' values before correction 15](#_Toc100944664)

[Figure 17 - 'Wifi' values after correction 16](#_Toc100944665)

[Figure 18 - Distribution of new variable 'Gender' 16](#_Toc100944666)

[Figure 19 - Distribution of new variable 'Age' 16](#_Toc100944667)

[Figure 20 - Distribution of new variable 'satisfaction\_avg' 16](#_Toc100944668)

[Figure 21 - Heatmap 17](#_Toc100944669)

[Figure 22 - PCA Analysis 17](#_Toc100944670)

[Figure 23 – Customers’ Clusters 18](#_Toc100944671)

[Figure 24 - Customers’ Clusters: Age Distribution 18](#_Toc100944672)

[Figure 25 - Customers’ Clusters: Descriptive Statistics 19](#_Toc100944673)

[Figure 26 - Customers' Satisfaction Clusters 19](#_Toc100944674)

[Figure 27 - Customers' Satisfaction Clusters: Accommodation 20](#_Toc100944675)

[Figure 28 - Customers' Satisfaction Clusters: Reception 20](#_Toc100944676)

[Figure 29 - Accommodation vs Reception 20](#_Toc100944677)

[Figure 30 - Reception vs Catering 20](#_Toc100944678)

[Figure 31 - Customers' Satisfaction Clusters: Descriptive Statistics 21](#_Toc100944679)

[Figure 32 - Clients' Retention Level 21](#_Toc100944680)

[Figure 33 - Average Satisfaction 21](#_Toc100944681)

[Figure 34 - Clients' Retention Level: Descriptive Statistics 22](#_Toc100944682)

# Introduction and Methodology

BookMe is a well-established company operating in the hospitality sector by delivering lodging services to those who travel the world, whether for leisure or business. It provides a global website to book accommodation, with more than 100.000 consumers accesses per year, and around 30.000 registered customers. The website offers multiple services though it is focused on providing rooms with the best conditions possible. BookMe implemented a survey to obtain their customer’s insights on the quality of the service at the end of each stay. They are asked to evaluate their satisfaction regarding a variety of factors related to the service on a scale of 0 to 5.

Even though the company has had stable revenues and a healthy bottom line in the past 3 years, **profit growth perspectives for the next three years are feeble**. Besides investing in a Marketing efficiency program to improve marketing activities, focused on boosting campaign efficiency, BookMe also hired a team of data scientists to analyze customer behavior and satisfaction, and predict which customers have high probability of churn depending on behavior and/or satisfaction.

In order to proceed, the data science team chose the **Cross Industry Standard Process for Data Mining** (CRISP-DM) as a methodological approach. It is comprised of six key steps: **understand the business need**, including its objectives; **understand the data**, how it is extracted loaded, described, explored and its overall quality; **prepare the data**, including selecting and cleaning, and formatting; **model the data**, by selecting modelling techniques, creating a test design, building the models and assessing them; **evaluate the results**, review the process and determine next steps; and, finally, **deploy the model**, monitor the results, produce a final report and review the project.

To help kick-start the project and reduce overheads, BookMe provided three different datasets: a training set (“train.csv”) to build the machine learning models and assess their performance if needed, a test set (“test.csv”) to see how well the models perform on unseen data, and a final test set (“test\_samplesubmission.csv”) for cloud-based deployment. The data science team chose Python along with a Jupyter Notebook interface to develop the analysis.

# Data Exploration

As a general first step to using Python for Data Science, code needs to be imported from multiple libraries. These are collections of related modules containing bundles of code that can be used repeatedly in multiple programs. This project uses pandas, numpy, seaborn, matplotlib, datetime, sklearn, and scipy (Attachment A). After importing data from “train.csv” and “test.csv”, exploration can begin. Basic exploration was performed by checking the shape (number of records and attributes) of both datasets, column names and types, duplicate values, head and tail records, descriptive statistics, skewness and kurtosis, unique values per column, and creating multiple visualizations for observing relations between variables. From this step, the data science team drew the following conclusions: there are no duplicate values, the training dataset has 15,589 rows and 21 columns, about three times as many as the test dataset, which has 5,195 rows and 20 columns, which is a proper proportion for developing a supervised machine learning model. “train.csv” has 16 integer variables, 4 object, and 1 float variable. “test.csv” is equal, minus the float variable. Although neither has duplicate values, there are 195 missing values in the training dataset’s column *Year\_Birth*.

For the training dataset, the average customer satisfaction is 3. There is a high *RewardPoints* standard deviation, indicating that its values are spread out. Multiple services were rated 0, at least once, expect for *Location*, *Staff*, *PriceQuality, CheckOut*, *CheckIn* and *Cleanliness*. The median rate for *Amenities, Staff, OnlineBooking, PriceQuality, RoomSpace, CheckOut and Cleanliness* is 4*.* Likewise, the median age is 41. The Maximum value for Wifi is 6, which was addressed later in the process.All services have been rated 5 at least once. The test dataset provided similar results.

Regarding skewness, the training set only has moderately skewed variables: *Amenities, Staff, PriceQuality, CheckOut, and Cleanliness.*Parallelly, there are no high kurtosis values. Kurtosis describes the “tailedness” of the data. Most of the attributes, except for *RewardPoints*, are negative, meaning that their distribution is flatter than a normal curve with the same mean and standard deviation.

From analysing categorical variables there are no unary variables nor missing values. Name can be used to create a binary variable called gender. From the visualizations, the data science team inferred that clients with churn are the ones with more longevity and reward points.

# Data Pre-processing

s

# Modelling

# Performance Assessment

# Conclusions

Briefly summarize the important results and conclusions presented in the paper. What are the most important points illustrated by your work? How will your results improve future research and applications in the area?

# Bibliography

* James, G., Witten, D., Hastie, T. J., & Tibshirani, R. J. (2017). *An introduction to statistical learning: With applications in R*. Springer;
* Patel, A. A. (2019). *Hands-on unsupervised learning using Python: How to build applied machine learning solutions from unlabeled data*. O'Reilly;
* Harrison, M. (2019). *Machine learning pocket reference: Working with structured data in Python*. O'Reilly.

# Annexes

## Figures

Figure 1 - Summary of dataset's attributes and description



Figure 2 - First three entries of BookMe dataset

Uma imagem com mesa

Descrição gerada automaticamente

Figure 3 - Duplicated entries removed

Uma imagem com mesa

Descrição gerada automaticamente

Figure 4 - Information on BookMe dataset

Uma imagem com mesa

Descrição gerada automaticamente

Figure 5 - Descriptive statistics for numerical variables

Uma imagem com mesa

Descrição gerada automaticamente

Figure 6 - Countplot of "RoomType" and "TypeTravel"

Chart, bar chart

Description automatically generated

Figure 7 - Countplot of "Longevity" and "Churn"

Chart, bar chart

Description automatically generated

Figure 8 - Skweness values

Uma imagem com texto

Descrição gerada automaticamente

Figure 9 - Kurtosis values

Uma imagem com texto, recibo

Descrição gerada automaticamente

Figure 10 - Boxplots for 'Amenities' and 'Staff'

Chart, box and whisker chart

Description automatically generated

Figure 11 - Boxplots for 'PriceQuality' and 'CheckOut'

Chart, box and whisker chart

Description automatically generated

Figure 12 - Boxplot for 'Cleanliness'

Chart

Description automatically generated

Figure 13 - Missing values distribution (after transformation of underage people to missing)

Text

Description automatically generated with medium confidence

Figure 14 - 'Year\_Birth' descriptive statistics before imputation

Text

Description automatically generated

Figure 15 - 'Year\_Birth' descriptive statistics after imputation

Text

Description automatically generated

Figure 16 - 'Wifi' values before correction

Table

Description automatically generated

Figure 17 - 'Wifi' values after correction

Table

Description automatically generated

Figure 18 - Distribution of new variable 'Gender'

Text

Description automatically generated with medium confidence

Figure 19 - Distribution of new variable 'Age'

Text, letter

Description automatically generated

Figure 20 - Distribution of new variable 'satisfaction\_avg'

Text, letter

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidenceFigure 21 - Heatmap

Figure 22 - PCA Analysis

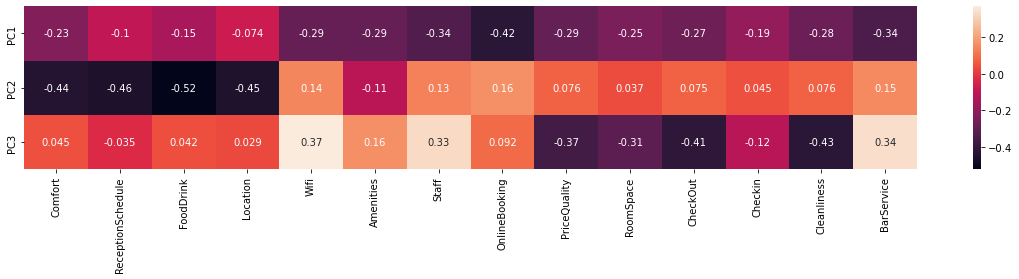


Figure 23 – Customers’ Clusters

Table

Description automatically generated

Figure 24 - Customers’ Clusters: Age Distribution

Chart, histogram

Description automatically generated

Figure 25 - Customers’ Clusters: Descriptive Statistics

Text

Description automatically generated Graphical user interface, text, application, email

Description automatically generatedText

Description automatically generated with medium confidence

Figure 26 - Customers' Satisfaction Clusters

Graphical user interface

Description automatically generated with medium confidence

Figure 27 - Customers' Satisfaction Clusters: Accommodation

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedFigure 28 - Customers' Satisfaction Clusters: Reception

Figure 29 - Accommodation vs Reception

Chart, scatter chart

Description automatically generated

Figure 30 - Reception vs Catering

Scatter chart

Description automatically generated

Figure 31 - Customers' Satisfaction Clusters: Descriptive Statistics

Table

Description automatically generatedGraphical user interface, text

Description automatically generatedGraphical user interface, text, application

Description automatically generated

Figure 32 - Clients' Retention Level

Table

Description automatically generated

Figure 33 - Average Satisfaction

Chart, histogram

Description automatically generated

Figure 34 - Clients' Retention Level: Descriptive Statistics

Graphical user interface, text

Description automatically generated Text, table

Description automatically generated Graphical user interface, text, application

Description automatically generated

## Methodologies

Methodology 1 – KNN

**K-Nearest** **Neighbours[[1]](#footnote-1)**

The KNN (k-nearest neighbours) is the most common neighbourhood-based method is k-nearest neighbours (KNN). To label each new point, KNN looks at a k number (where k is an integer value) of nearest labelled points and has these already labelled neighbours vote on how to label the new point. In Python, by default, KNN uses Euclidean distance to measure what is closest.

The choice of k is very important. If k is set to a very low value, KNN becomes very flexible, drawing highly nuanced boundaries and potentially overfitting the data. If k is set to a very high value, KNN becomes inflexible, drawing a too rigid boundary and potentially underfitting the data.

*Strengths:* Unlike linear methods, KNN is highly flexible and adept at learning more complex, nonlinear relationships. Yet, KNN remains simple and interpretable.

*Weaknesses:* KNN does poorly when the number of observations and features grow. KNN becomes computationally inefficient in this highly populated, high-dimensional space since it needs to calculate distances from the new point to many nearby labelled points in order to predict labels. It cannot rely on an efficient model with a reduced number of parameters to make the necessary prediction. Also, KNN is very sensitive to the choice of k. When k is set too low, KNN can overfit, and when k is set too high, KNN can underfit.

Methodology 2 - PCA Analysis

**Principal component analysis[[2]](#footnote-2)**

Principal component analysis (PCA) refers to the process by which principal component are computed, and the subsequent use of these components in understanding the data. PCA is an unsupervised approach, since it involves only a set of features X1, X2,...,Xp, and no associated response Y . Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization (visualization of the observations or visualization of the variables).

Suppose that we wish to visualize n observations with measurements on a set of p features, X1, X2,...,Xp, as part of an exploratory data analysis. We could do this by examining two-dimensional scatterplots of the data, each of which contains the n observations’ measurements on two of the features. However, there are such scatterplots. If *p* is large, then it will certainly not be possible to look at all of them; moreover, most likely none of them will be informative since they each contain just a small fraction of the total information present in the data set. Clearly, a better method is required to visualize the n observations when p is large. In particular, we would like to find a low-dimensional representation of the data that captures as much of the information as possible. For instance, if we can obtain a two-dimensional representation of the data that captures most of the information, then we can plot the observations in this low-dimensional space. PCA provides a tool to do just this. It finds a low-dimensional representation of a data set that contains as much as possible of the variation. The idea is that each of the n observations lives in p-dimensional space, but not all of these dimensions are equally interesting. PCA seeks a small number of dimensions that are as interesting as possible, where the concept of interesting is measured by the amount that the observations vary along each dimension. Each of the dimensions found by PCA is a linear combination of the *p* features.

Methodology 3 – K-Means Clustering

**K-Means1**

*K*-means clustering is a simple and elegant approach for partitioning a data set into *K* distinct, non-overlapping clusters. To perform *K*-means clustering, we must first specify the desired number of clusters *K*; then the *K*-means algorithm will assign each observation to exactly one of the K clusters.

It optimizes the grouping by minimizing the within-cluster variation (also known as inertia) such that the sum of the within-cluster variations across all k clusters is as small as possible. To speed up this clustering process, k-means randomly assigns each observation to one of the k clusters and then begins to reassign these observations to minimize the Euclidean (or others) distance between each observation and its cluster’s center point, or centroid. As a result, different runs of k-means—each with a randomized start—will result in slightly different clustering assignments of the observations. From these different runs, we can choose the one that has the best separation, defined as the lowest total sum of within-cluster variations across all k clusters.

Methodology 4 – Hierarchical clustering

**Hierarchical Clustering1**

An alternative clustering approach—one that does not require us to precommit to a particular number of clusters—is known as hierarchical clustering. One version of hierarchical clustering called agglomerative clustering uses a tree-based clustering method and builds what is called a dendrogram. A dendrogram can be depicted graphically as an upside-down tree, where the leaves are at the bottom and the tree trunk is at the top.

The leaves at the very bottom are individual instances in the dataset. Hierarchical clustering then joins the leaves together—as we move vertically up the upside-down tree—based on how similar they are to each other. The instances (or groups of instances) that are most similar to each other are joined sooner, while the instances that are not as similar are joined later. With this iterative process, all the instances are eventually linked together forming the single trunk of the tree.

This vertical depiction is very helpful. Once the hierarchical clustering algorithm has finished running, we can view the dendrogram and determine where we want to cut the tree—the lower we cut, the more individual branches we are left with (i.e., more clusters). If we want fewer clusters, we can cut higher on the dendrogram, closer to the single trunk at the very top of this upside-down tree. The placement of this horizontal cut is similar to choosing the number of k clusters in the k-means clustering algorithm.

1. Patel, A. A. (2019). *Hands-on unsupervised learning using Python: How to build applied machine learning solutions from unlabeled data*. O'Reilly [↑](#footnote-ref-1)
2. James, G., Witten, D., Hastie, T. J., & Tibshirani, R. J. (2017). An introduction to statistical learning: With applications in R. Springer [↑](#footnote-ref-2)