Group Project

BookMe

DATA SCIENCE AND MACHINE LEARNING 2022

April 26, 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.

Focused in detecting possible churn situations, BookMe hired a team of data scientists (your group) to analyze the behavior and satisfaction of their customers and to predict which customers have a high probability of churn depending on their behavior/satisfaction.

2 Objective of the project

Your goal is to build a predictive model that answers the question "Which customers are more likely to churn?" using the small quantity of data accessible from the customers data base that contains general information about the customers behaviour and their satisfaction.

3 Datasets

You have access to three different datasets:

1. The training set should be used to build your machine learning models and assess their performance if needed. In this set, you also have the ground truth associated with the customer behavior, i.e., if the user was considered churn or not.

2. The test set should be used to see how well your model performs on unseen data. In this set, you don't have access to the ground truth, and the goal of your team is to predict that value (0 or 1) by using the model you created based on the training set. The predicted values in the test set should be submitted on Kaggle. The score of your predictions will be evaluated using the F1 Score.

The available 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 - 1; or nochurn - 0)
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.
   The file naming format should be "202122_Pred_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 penalizations on the grade.
   The file naming format should be "202122_Pred_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)
Kaggle performace	15%	3
Report-quality and Story-telling	7.5%	1.5
Introduction and Methodology	5%	1
Exploration	10%	2
Pre-processing	15%	3
Modelling	17.5%	3.5
Performance Assessment	10%	2
Conclusions	5%	1
Other predictive models (not given during classes)	5%	1
Creativity & Other Self-Study	10%	2
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:

- **Kaggle performace:** The performance obtained on Kaggle, on the submission selected (F1 Score).
- **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, meaningful 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 model. Involves all the steps for cleaning, transform and reduce the dataset.

- **Modelling:** the implementation of different predictive algorithms and the process of fine-tuning those models. The application of additional models not given during classes are optional and considered as points in "Other predictive models".

- **Performance Assessment:** The comparison of different models and their performance.

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

- **Other predictive models:** 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 provided by the different algorithms,

the theoretical explanation of the algorithm itself and the justification of the chosen parameters;

- **Creativity and Other Self-Study:** If other techniques not given during practical classes are applied, a theoretical explanation of the algorithm / technique 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 the quality of visualizations, plots and others.

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