OCTOBER 2022

HOTEL CALIFORNIA

GROUP PROJECT
MACHINE LEARNING I 2022/2023

I. INTRODUCTION

Welcome to the new and improved Hotel California
Such a lovely place

First inaugurated in 1977, Hotel California was a fairly small, albeit charismatic hotel in downtown Lisbon. Its unique location and history made it a cultural and touristic landmark that attracted people from all over the world. In the old days, vacancies at Hotel California were rare and successful bookings had to be made months in advance.

However, times changed. Touristic operators, overbooking and free cancellation policies became standard practice in the industry. Cancellations started to be more and more frequent throughout the 2010's and neither management nor staff had the means or the know-how to properly operate in the new landscape. Unsurprisingly, the Hotel filed for insolvency during the Covid-19 lockdowns and its doors have not re-opened to the public.

A new group of investors purchased the old hotel and are now making arrangements for a grand reopening in 2025.

II. PROJECT GOALS

The development of an appropriate overbooking strategy to minimize cancellation vacancies is the top priority for the new management.

The first step is to develop a predictive model that is able to identify whether a given Booking will be cancelled or not. Hotel management is encouraging the most talented teams of machine learning engineers to participate. **That's where you come in**:

You will be provided with a representative sample of Hotel California's bookings in 2016. **The data is divided between a training set (close to 14000 bookings) and a testing set (close to 5400 bookings).**

Your goal is to train a predictive model with the training set and use the model to make predictions on the test set.

III. DATASET

You have access to two different datasets:

In the **training set**, you will find the features and the ground truth associated with each Booking instance, i.e. whether it was cancelled (1) or not (0). Use it to build your machine learning models. The goal will be to use the model you created and make predictions on unseen data (i.e. your test set).

In the test set, you will see the same features presented in the training set. However, you will not have access to the ground truth of the test set. Your goal will be to predict the ground truth value ("0" or "1") by using the model you created using the training set.

The available data contains the following attributes:

ATTRIBUTE	DESCRIPTION
BookingID	Unique identifier
ArrivalYear	Year of guest check-in date
ArrivalMonth	Month of guest check-in date in numeric format (January = 1, ..., December = 12)
ArrivalWeekNumber	Number of week in a calendar year of guest check-in date
ArrivalDayOfMonth	Day of month scheduled for check-in
ArrivalHour	Hour of the day guest intends to arrive on the check-in date
Adults	Number of adults (+14 years old) registered in the booking
Children	Number of children (4 to 14 years old) registered in booking
Babies	Number of children (between 0 and 4 years old) registered in the booking
FirstTimeGuest	Boolean feature worth 1 if this is the guest's first booking
AffiliatedCustomer	Boolean feature worth 1 if the guest is part of the hotel's affiliate program
PreviousReservations	Number of previous reservations made by guest
PreviousStays	Number of previous uncanceled reservations made by guest

03

ATTRIBUTE	DESCRIPTION
PreviousCancellations	Number of previous cancelled reservations made by guest
DaysUntilConfirmation	Days elapsed between the booking date and confirmation of the reservation date
OnlineReservation	Boolean feature worth 1 if the reservation was made online
BookingToArrivalDays	Days elapsed between the booking date and check-in date
ParkingSpacesBooked	Number of parking spaces requested in the reservation
SpecialRequests	Number of special requests made by the guest (e.g. baby crib, extra towels, etc...)
BookingChanges	Number of changes made by guest after the initial reservation
CompanyReservation	Boolean feature worth 1 if the reservation was made in the name of a company
FloorReserved	Floor of the room where the guest intends to stay
FloorAssigned	Floor of the room where the guest was assigned to stay
DailyRateEuros	Average price per day in Euros
DailyRateUSD	Average price per day in US Dollars
PartOfGroup	Boolean feature worth 1 if the current reservation is associated with at least another reservation
%PaidinAdvance	Percentage of total amount paid in advance by guest (non-refundable)
OrderedMealsPerDay	Number of daily meals (Breakfast, Lunch, Dinner) included in the reservation
CountryofOriginAvgIncomeEuros (Year-2)	GDP per capita of the guest's country of origin 2 years before the booking
CountryofOriginAvgIncomeEuros (Year-1)	GDP per capita of the guest's country of origin in the year before the booking
CountryofOriginHDI (Year-1)	Human Development Index value for the guest's country of origin in the year before the booking
Canceled	Outcome variable worth 1 if the booking was cancelled

IV. DELIVERABLES

Upon the project's deadline, you will be required to submit a zip file containing:

- A report that describes the analytical processes and the conclusions obtained with, at most, 15 pages (excluding cover, but including annexes). The file naming format should follow ***ML1_GroupXX_Report.pdf***, where ***GroupXX*** should be your group number. **Use the template provided in the file *Report_Template.docx*** with the following settings:
    - **Heading 1: Calibri, Size 14 pt, in bold**
    - **Heading 2 (if needed): Calibri, Size 13 pt, in bold**
    - **Text: Calibri, Size 11 pt, line spacing of 1.15 pt and paragraph spacing of 6 pt**

- A Jupiter notebook with your code implementation. The file naming format should be ***ML1_GroupXX_Notebook.ipynb***, where ”GroupXX” should be your group number

- A csv file containing your test set's BookingID and your predictions for column Canceled. The file should only contain 2 columns.

V. NOTES

1. We will disregard steps and results that are not mentioned in your report
2. When in doubt, we will run your Jupyter Notebooks. Therefore, make sure we can run the notebook from start to finish in one go. Notebooks that do not fulfil this condition will be penalized.
3. All the code that is not needed to obtain your final solution should be commented
4. The report and code will pass through a process of plagiarism checking
5. Your predictions will be evaluated using the metric F1-Score