MATH 11205: Machine Learning in Python 2022-2023
Project 2 Description

For this assignment we will be using data collected on hotel bookings (Antonio, Almeida and Nunes, 2019) https://sciencedirect.com/science/article/pii/S2352340918315191#f0010. These data come from the booking systems of two real hotels and reflect bookings made between July 1st, 2015 . through August 31st 2017. This is a large (119,390 observations) and real world (i.e. messy) data set, care will be needed to check for and to address underlying data quality issues

Assignment Goal

For the purposes of the project, consider yourself a Data Scientist who has been hired by a large hotel operator who has provided these data in the hope of better understanding why customers cancel their reservations. The two hotels present in the data are individual representative examples of their resort and city properties. Using these data they would like you to construct a predictive model that can accurately classify when a booking will be canceled. In this case they are primarily interested in the accuracy of the model but they are also interested in understanding what aspects of a booking affect the likelihood of it being canceled.

As such, you need to develop an **understandable, validated** model for the 'is_canceled' variable as the outcome of interest using features derived from the data. As before it is important that this model be **accurate** and **reliable** and any conclusions you draw well supported. We explicitly do not want a blackbox model - you should be able to explain and justify your modeling choices and your model's predictions.

Your model may use as few or as many of the provided features, and you may transform and manipulate these features in any way that you want to generate additional features. Also note as these data are directly derived from the hotels' booking system there is the possibility of incorrect / unreasonable data being present - performing basic sanity checks and validation on the features is strongly recommended as part of your exploratory data analysis. You are welcome to exclude as much or as little of the data as you would like, but these choices should be clearly justified in your text.

We have covered a number of models and modeling approaches in the lectures and workshops and you should explore a variety of different approaches for this particular task. However, your ultimate goal is to deliver a **single model**. These are competing interests and it is up to you to find a reasonable balance between the two - some of your marks will be based on how well you accomplish this.

Client Requirements

Your report (in the discussion section) must include the following:

- Some discussion of the features that are most important for predicting a cancellation - we do not need discussion of specific coefficient values but direction of the effect should be clear (e.g. the earlier a booking is made the more likely it is to be canceled).
- A validated assessment of your model's performance, but this must be specifically discussed in the context of bookings and running a hotel.
- It is not sufficient to report summary statistics like the accuracy or AUC - you must address the performance in terms of potential gains and losses for the hotel (e.g. think about what happens if your model predicts a cancellation that does not actually occur and a room ends up being double booked or vice versa).
- Explain why you think your particular model would or would not be economically viable.

Required Structure

A Jupyter notebook template called 'project.ipynb' has been provided. It includes the required sections along with brief instructions on what should be included in each section. Your completed assignment must follow this structure - **you should not add or remove any of these sections, if you feel it is necessary you may add extra subsections within each**. Please remove the instructions for each section in the final document.

All of your work must be contained in the 'project.ipynb' notebook, we will only mark what is included in this file (both the write-up and relevant coding). You may work on the notebook in whichever environment you prefer, but please ensure that the final pdf file includes all necessary parts of your writing. For the name of your .ipynb file, use the convention of 'name_surname.ipynb (ie. ozan_evkaya.ipynb).

Our expectation is that most projects will be roughly 25-30 pages in length at most including text & figures, but excluding the related code. Overall, there is an **upper limit of 35 pages** including the coding part. Your notebook must include all of your work, but make sure that you are only retaining required components, e.g. remove unused code and figures (if a figure is not explicitly discussed in the text it should not be in the final document). **So, there is a trade-off between the length of your text and coding snippets while constructing your report.** Overall, your project will be partially assessed on your organization / presentation of the document - it should be as polished and streamlined as possible. **Try to be as concise as possible while creating your write-up. We highly recommend that you check the appearance of your rendered PDF before submitting, as its appearance can differ significantly from the notebook**.

You are expected to submit your completed work. For this, please submit your final PDF of project report (generated from a Jupyter notebook) to the Project assignment on Gradescope. For a group submission - all contributors should be added to the assignment on Gradescope. At the beginning of project '.ipynb' notebook, the contributor name places are available to fill out !

About Data Set

The data set contains the following variables:

Table 1: Description of variables in the data set

VARIABLE	DESCRIPTION
is_canceled	Value indicating if the booking was canceled (1) or not (0)
hotel	Hotel (Resort Hotel or City Hotel)
lead_time	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
arrival_date_year	Year of arrival date
arrival_date_month	Month of arrival date
arrival_date_week_number	Week number of year for arrival date
arrival_date_day_of_month	Day of arrival date
stays_in_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
stays_in_week_nights	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
adults	Number of adults
children	Number of children
babies	Number of babies
meal	Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)
country	Country of origin. Categories are represented in the ISO 3155–3:2013 format
market_segment	Market segment designation. In categories, the term TA means 'Travel Agents' and TO means 'Tour Operators'
distribution_channel	Booking distribution channel. The term TA means "Travel Agents" and TO means "Tour Operators"
is_repeated_guest	Value indicating if the booking name was from a repeated guest (1) or not (0)
previous_cancellations	Number of previous bookings that were cancelled by the customer prior to the current booking
previous_bookings_not_canceled	Number of previous bookings not cancelled by the customer prior to the current booking
reserved_room_type	Code of room type reserved. Code is presented instead of designation for anonymity reasons
assigned_room_type	Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons
booking_changes	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
deposit_type	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made; Non Refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total cost of stay.
agent	ID of the travel agency that made the booking
company	ID of the company/entity that made the booking or responsible for paying the booking ID is presented instead of designation for anonymity reasons
days_in_waiting_list	Number of days the booking was in the waiting list before it was confirmed to the customer
customer_type	Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated to it Group – when the booking is associated to a group Transient – when the booking is not part of a group or contract, and is not associated to other transient booking Transient-party – when the booking is transient, but is associated to at least other transient booking—
adr	Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
required_car_parking_spaces	Number of car parking spaces required by the customer
total_of_special_requests	Number of special requests made by the customer (e.g. twin bed or high floor)