Introduction to Machine Learning (67577)
Hackathon 2023 - Challenge 1

Teachers: Dr. Gabriel Stanovsky, Gilad Green

TAs: Eitan Wagner, Gili Lior, Noa Moriel, Reshed Mintz

Tzars: Itai Alon, Nitay Alon

Second Semester, 2023

1 Hackathon Challenge 1: Hotel Cancellations

Cancellation behavior analysis holds immense value for Agoda. By leveraging machine learning, we can revolutionize our ability to make quick decisions and provide our partners, who may not have access to extensive booking records, with invaluable insights into cancellation patterns. Your task is to harness the power of machine learning to create predictive models that forecast whether a guest will stay at a hotel or cancel their booking. But it doesn't stop there! We also need to determine the most likely timing of cancellations.

Your solution will empower Agoda to optimize the net room nights at the booking level, resulting in improved decision-making, enhanced efficiency, and, ultimately, happier guests. So, gear up, get creative, and let your passion for machine learning transform how we understand and manage hotel cancellations.

1.1 Dataset

The *training* dataset Agoda_training.csv holds 58,659 records with 38 features determined upon booking and the unknown variable of interest, the "cancellation_datetime" which features the date of cancellation when it is such, see Table 1. To evaluate your predictions, we test you on 7,818 booking records, split into two equal size files, for which you are given only *part* of the booking information - the missing data is specified in each task.

The *test* data sets Agoda_Test_1.csv and Agoda_Test_2.csv correspond to tasks 1.2.1 and 1.2.2

Column Name	Description
h_booking_id	Booking id
booking_datetime	Date and time the booking occurred
checkin_date	Date of check-in
checkout_date	Date of checkout
hotel_id	The id of the booked hotel
hotel_area_code	A random code representing the area the hotel is at (e.g., Gush Dan)
hotel_city_code	A random code representing the city where the hotel is (e.g., Tokyo)
hotel_country_name	ISO2 of the hotel's country
hotel_brand_code	A random code representing the hotel's brand (e.g., Ibis Hotels)
hotel_chain_code	A random code representing the hotel's chain (e.g., Marriot)
hotel_live_date	When the hotel was first created in the system
hotel_star_rating	Hotel's star rating, between 0 and 5 by 0.5
accommadation_type_name	Hotel, Hostel, Apartment, etc.
charge_option	When does the customer need to pay
h_customer_id	Hashed customer_id
customer_nationality	Country name of the customer's passport
guest_is_not_the_customer	Boolean indicating whether the person who bought the bookings.
guest_nationality_country_name	Country name of the guest's passport
no_of_adults	
no_of_children	
no_of_extra_bed	
no_of_room	How many rooms were booked
origin_country_name	Name of the country inferred from the IP address
language	Language used at the website
original_selling_amount	Cost of booking for the customer
original_payment_method	
original_payment_type	
original_payment_currency	ISO3 of currency code the customer was charged in
is_user_logged_in	Was the user logged in when they made the booking?
cancellation_policy_code	See below
is_first_booking	Was this the first booking made by this customer?
request_nonesmoke	Did the user request a none smoking room?
request_latecheckin	Made the user request to check in late?
request_highfloor	Did the user request a high floor?
request_largebed	Did the user request a large bed?
request_twinbeds	Did the user request twin beds?
request_airport	Did the user request an airport transfer?
request_earlycheckin	Made the user request to check in early?
cancellation_datetime	Optional: date of cancellation. If Null, the booking was not canceled.

Table 1: *Agoda data column descriptions*

1.2 Tasks

Please note that the tasks are independent. Partial submission will grant you a partial grade. Please read the submission guides carefully - if your output doesn't exactly match these definitions it will regarded as no submission.

1.2.1 Cancellation prediction

Given the booking information, we would like to predict **whether** this order will or will not be canceled.
The input A dataframe of the booking information – all columns except "cancellation_datetime".
The output A csv file named "agoda_cancellation_prediction.csv" with two columns:- *id* (h_booking_id) and *cancellation*: where 1 indicating that a cancellation is predicted, and 0 otherwise.
Evaluation We will evaluate your predictions according to their F1 macro metric. An example of the output:

ID	cancellation
111	1
222	0
333	1

Table 2: *Cancellation prediction output*

1.2.2 Cost of cancellation

If a cancellation occurs, we would like to **estimate** the expected money loss. This is an important metric, as it helps us estimate the income.
The input A dataframe of the booking information – all colums, excluding "original_selling_amount"
The output A csv file named "agoda_cost_of_cancellation.csv" with two columns: *id*(h_booking_id) and *predicted_selling_amount* column: the predicted selling amount of that order. Provide a prediction for **all** records. That is, also for those orders who are *unlikely to be cancelled* - in this case the date entry should be -1. An example of the output:
Evaluation We will evaluate your predictions according to their RMSE score from the true value

ID	predicted_selling_amount
111	44.5
222	-1
333	290

Table 3: *Cancellation prediction output*

1.2.3 Churn prediction Model

For this part, you are asked to find the most relevant feature for cancellation prediction. That is, what features predict cancellation "the best". This is very useful when building a model that flags orders at risk. *Note* - there's no correct answer to this section: your goal is to find a set of features that is informative about cancellation and as small as possible.
The input The training data set

The output A pdf file named "agoda_churn_prediction_model.pdf". Provide at most 2 written pages describing the features. Add between 3-4 figures to reinforce your claims. Make sure that you justify your suggestions rigorously.

1.2.4 Suggest the optimal cancellation and pricing policy

For this part, you get to be the CEO of Agoda for a few hours. First, suggest a cost function of cancellations, e.g. a linearly increasing cost as a cancellation nears the check-in date. Next, try to think of the optimal cancellation and pricing policy based on the data, minimizing the cancellations but still allowing customers flexibility. Can you suggest a more generic framework that fits the pricing policy given various cancellation cost functions?
There is no correct answer to this question. Be as creative as you can and suggest data-driven solutions that are well-reasoned.
The input The training data set
The output A pdf file named "agoda_cancellation_policy.pdf". Provide at most 2 written pages describing the policies. Add between 3-4 figures to reinforce your claims.

1.3 Code Requirements

You must submit a separate file for each task, named `task_i.py`, where *i* is the task number. In addition you have to submit a `main.py` file that has an executable code
(`if __name__ == "__main__" ...`).
In this file, you implement a method called `main` that inputs a relative path to the task 1 test set and task 2 list of dates (in `pd.date` format - YYYY-MM-DD).
Each task code should run **independently** when the main file is executed (that is, if task one fails, task two should still run).
NOTICE Since we are not going to train your model, you are required to load your trained model in this method. Please verify this method works as expected (with all the required additional files - s.a., the weights file).