TEST: PREDICTING BAGGAGE LIKELIHOOD

OVERVIEW

At eDreams Odigeo we're always looking for ways to improve customer satisfaction. With this objective in mind, we would like to predict whether a new customer is interested in buying additional baggage in order to speed up the booking process.

GOAL

The goal of this task is to predict which new customers are going to purchase additional baggage for their trips using historical from previous customers. The code can be developed in any of the following languages: R, Python or Java.

DATA DESCRIPTION

Two files are attached with the training and test datasets. The training dataset contains 50000 bookings and the test dataset 30000 bookings. The data fields are the following ones:

TIMESTAMP: *(date)* Date when the booking was bought.

WEBSITE: *(string)* Website where the trip was purchased. It is composed of a prefix that stands for the website ("ED" = Edreams, "OP" = Opodo, "GO" = Go Voyage) and a suffix for the country (for example: ES = Spain)

GDS: *(integer)* Number of flights bought through the Global Distribution System

NO GDS: *(integer)* Number of flights bought though other channels.

DEPARTURE: *(date)* Departure date

ARRIVAL: *(date)* Arrival date

ADULTS: *(integer)* Number of adults

CHILDREN: *(integer)* Number of children

INFANTS: *(integer)* Number of infants

TRAIN: *(boolean)* Whether the booking contains train tickets or not

DISTANCE: *(float)* Distance travelled

DEVICE: *(string)* Device used for the purchase

HAUL TYPE (string): Whether the trip was "Domestic", "Continental" or "Intercontinental".

TRIP TYPE: *(string)* Trips can be either "One Way", "Round Trip" or "Multi-Destination"

PRODUCT: *(string)* Bookings can contain only travel ("Trip") or the travel and a hotel ("Dynpack").

SMS: *(boolean)* Indicates if the customer has selected a confirmation by SMS

EXTRA BAGGAGE: *(boolean)* Variable to predict, only in the train dataset. Indicates if the customer has purchased extra baggage for the trip or not.

EVALUATION

The evaluation will be based on the quality and explanation of the source code as well as the prediction score. For the prediction score, the metric will depend on the output: in case a binary output is submitted, the evaluation method will be the F1 Score, in case a probability is given, the evaluation method will be the AUC ROC.

SUBMISSION FORMAT

The submission must contain the source code and the predictions for the 30000 bookings in CSV format, for instance:

ID	EXTRA_BAGGAGE
0	True
1	False
[…]	

Or the probabilities:

ID	EXTRA_BAGGAGE
0	0.35
1	0.78
[…]	

Additionally, the submission should include comments or a document file explaining the solution or any insights the candidate might have found.