

Shanshu Interview Question

Problem Description: A room sharing company (e.g., Airbnb) wants to help room providers set a reasonable price for their rooms. One of the key steps is to build a model to predict the purchase probability of a room (described by certain features as well as the date) under certain prices. We have the following historic data:

ID	The data ID
Region	The region the room belongs to (an integer, taking value between 1 and 10)
Date	The date of stay (an integer between 1-365, here we consider only one-day request)
Weekday	Day of week (an integer between 1-7)
Apartment/Room	Whether the room is a whole apartment (1) or just a room (0)
#Beds	The number of beds in the room (an integer between 1-4)
Review	Average review of the seller (a continuous variable between 1 and 5)
Pic Quality	Quality of the picture of the room (a continuous variable between 0 and 1)
Price	The historic posted price of the room (a continuous variable)
Accept	Whether this post gets accepted (someone took it, 1) or not (0) in the end

The training data is posted at: http://www.menet.umn.edu/~zwang/files/case2_training.csv

The testing data is posted at: http://www.menet.umn.edu/~zwang/files/case2_testing.csv

(There are 50,000 training and 20,000 testing data.)

Goal: Build a model to predict the purchase probability of each test data. We will evaluate the model by the AUC of your result (thus please give a probability for each test data).

Format of the answer: Please send back the following items to job@shanshu.ai

- 1) The prediction result: A csv file containing the result. To make the file smaller, you only need to keep the ID and answer columns.
- 2) The code you used to produce the answer. Python code is preferred. However, R or other languages are also acceptable.
- 3) A short description of your approach (no more than half a page) and the main findings in your analysis.