Introduction

In Ruangguru we always strive to give the best learning experience to our student, so we give personalized recommendations to student what to study next and what video to watch each time student completing a test, in this task you're given historical data of student performances on test and you're asked to predict probability of correct answer for the next question

Task

1. Create an Exploratory data analysis (EDA) and create a model to predict probability of is_correct in test.csv file given the training data
2. Briefly Explain your thought process and approach

Topic - Subtopic - Learning Node Relations

1. One topic can contains multiple subtopics
2. One subtopic can contains multiple learning nodes
   One question can have multiple learning nodes and one learning node can be tagged into different questions

Evaluation:

Submissions are evaluated on area under the ROC curve between the predicted probability and the observed target.

Dataset

Files

**please find the files for train.csv , test.csv and sumbission_example files here*

train.csv

- user_id (int64)
  ID code for the user. Unique per user
- session_id (int64)
  ID code for the session of the test. Unique per session
- session_no (int64)
  Nth test session of the user_id.
- topic (int64)
  ID code for the question Topic. Unique per topic

- sub_topic (int64)
  ID code for the question Sub Topic. Unique per subtopic
- learning_node (int64)
  ID code for the question Learning Node (or you can think of it as sub-sub-topic). Unique per learning node.
- question_id (int64)
  ID code for the question (same question has same question_id). Unique per question
- question_type (String)
  Type of question (Single Choice = only have 1 correct answer, Multiple = you have to choose all correct answer to be scored)
- session_question_no (int64)
  Nth Question for the Test Session
- learning_node_question_no (int64)
  Nth Question for the learning node for the test session
- question_difficulty (String)
  The Question Difficulty
- question_number_of_choice (int64)
  Number of Choice for the Question
- question_number_of_correct_choice (int64)
  Number of Correct choice for the Question
- question_number_of_correct_selected (int64)
  Number of Correct Answer user select for the question
- question_number_of_wrong_selected (int64)
  Number of Wrong Answer user select for the question
- ms_first_response (int64)
  Time in milliseconds the user need to answer the question
- is_correct (int64)
  1 if question_number_of_correct_selected = question_number_of_correct_choice, 0 otherwise
- row_id (int64)
  ID code for the row

test.csv

- For all column, the detail is same as train.csv
- Except in test set set we don't have this information:
  - is_correct
  - ms_first-response
  - question_number_of_wrong_selected
    question_number_of_correct_selecte
    d
- And you're asked to predict probability of is_correct

Test Submission Format

Here are several things you need to know about the tasks:

1. Please write your answers in the with the following format and make sure to send in,
    a. **submission.csv** with the same format of **submission_example.csv**
        i. For each row_id in the test dataset, you need to provide is_correct probability prediction result, the submission csv file should contain a header and have the following format

row_id	is_correct
0	0.01442211494
1	0.5250734431
2	0.3175731267

    b. Jupyter Notebooks or Script with explanation of your thought process and way you decide to use that approach
2. The deadline for the submission is **2 days** since you received this test assignment.
3. Kindly send back the result by replying to this email.

Please do not hesitate to contact the Ruangguru team should you have any questions.