

PeerSIST: Case Study for Analytics Role Applicants

Context

You are just hired as a data analyst in the newly formed analytics department of [Riiid](#), which is a leading AI startup company specializing in providing learning resources and adaptive practices to English learners in South Korea. On the first day of your job, you are invited to attend a meeting with the business operation team in which you are briefed on the company's platform. Then you are handed over a dataset ('EdNet') collected from this platform over the last two years from over 700K users. This dataset logged detailed user activities while they were interacting with systems. Intrigued by the sheer amount of data collected, your manager is interested in how the analytics department can help to support the company's missions to reimagine the learner's experience using AI/ML/data analytics techniques. You are asked to look into the data and prepare a brief for your manager. Specifically, your manager is looking for answers to the following questions.

Tasks

1. **Who are the users?** To answer this question, you will need to compile a user profile table (Table 1) with information about users, including
 - a. Overall practice volume and performance (e.g. # of questions answered, % of questions answered correctly)
 - b. Learning activity (e.g. # lectures watched, # explanation read)
 - c. Add three additional metrics you would like to compute to describe users

Create a few plots to illustrate the information in the tables. Feel free to choose the type of plot you think is appropriate. *If it takes a long time to process all users, you may choose to process only a subset of the users. We are not looking for the correctness of the end result, but rather your process.*

2. **What are the questions/items?** To answer this question, you will need to compile a question profile table (Table 2) with at least the following information including
 - a. Question ID
 - b. Question Type
 - c. Number of times being practiced
 - d. Number of times answered correctly

Create a few plots to illustrate the information in the tables. Feel free to choose the type of plot you think is appropriate.

3. Design a modified metric of "accuracy" to fairly describe users' ability by considering the difficulty level derived from Table 2. Be creative. Describe the procedure to compute the metrics. Be sure to be specific so that interns can use your pseudo code to implement the metrics without much trouble. Implement the proposed metrics and plot a histogram

of the metrics across all users (or a subset of users of your choice) and interpret the results.

4. Pick a user with a reasonable amount of activity (you will need to define the “reasonableness” and specify the criteria) and create a dashboard that consists of a series of plots to tell a story of this user’s activity patterns. For inspiration, you may look at the user dashboard for fitness trackers such as Fitbit.

Deliverable:

- A google slide deck summarizing the above findings in the format of plots or tables (those are not table 1 or table 2, but small tables you decide to use to present information) or other contents as requested by your manager (such as those pertaining to question 3 or 4). Please label clearly on the slide which question you are answering. There are no lower/upper limits of the number of slides. Always keep the message concise and effective, and keep your audience in mind. In this case, it is your manager. Please change the slide's permission to editable; we will make comments on your slides.
- Please create a GitHub account and upload codes to the repo (e.g. python notebook) as well as the result table from tasks 1 and 2
- Please submit your two links (one link to your google slide deck and another link to your github repo) to this google form <https://forms.gle/bn9KvpyJDpKx77d17>
- The deadline for submission is **3 pm Friday, Jan 20th**; however, if you also work simultaneously on the design task, you may submit by **3 pm Sunday, Jan 22nd**. **Late submissions will not be accepted.**

Dataset and Background Readings:

Please download the dataset from the following link. For this task, you may only use KT4 (uncompressed size 6.4GB). But you may need to download other small lookup tables(e.g. those in the contents folder) for the purpose of this assignment. **Note: it may take a few hours to download/unzip the data files**
<https://github.com/riid/ednet>

Please refer to this paper for details of dataset (mainly Section 1 and 2, up to page 7)
<https://arxiv.org/abs/1912.03072>

Software and Tools

You are free to use any software tools you feel comfortable with, which include but are not limited to Python, R, WEKA or Tableau.

Q&A (continuously updated)

====