RETTL Project Summer 2022 Sprint 2 Report

Tianze (Steven) Shou

7/22/2022

Abstract

During Sprint 2, the analytics work mainly involved decomposing and generating EDA statistics for the tutor log data. Some EDA tools for raw Datashop format tutor log data were developed for potential future uses, and basic statistics of the different periods and days were generated with the tutor log data in-hand. Tutor events (*correct attempt*, *incorrect attempt*, and *hint request*) were distilled and merged with the two other modalities (position events and observation events) to make a master event file. This master event file was shared with the ENA/TMA team for further analysis.

Introduction

The tutor logging data exported from Lynette follows the Datashop by-transaction format, where each `transaction` between the intelligent tutor and the student constitutes a row. Each transaction has unique ID and is accompanied with information on `anonnymous student ID`, `student response type`, `tutor response type`, `time`, `student input`, etc..

Just like the observation data and position data, the tutor log data come from the 3-day experiment session where 5 class periods were included. The class periods are usually consisted of teacher's explanation of class material followed by individual student's practice with the intelligent tutor where teacher's help may be provided individually.

Tutor Data Overview

During the first week of Sprint 2, some exploratory data analysis was done to the tutor data. Major time commitment was devoted to creating a tool box for generating summary statistics from tutor data. The tool box includes functions that extracts mean correct response rate, mean KC-level correct rate, mean time taken to solve a question, mean problem level, etc. from a given group of students and a given time period. For more information on the tool box, please refer to the `tutor_log_summary_stat.ipynb` file.

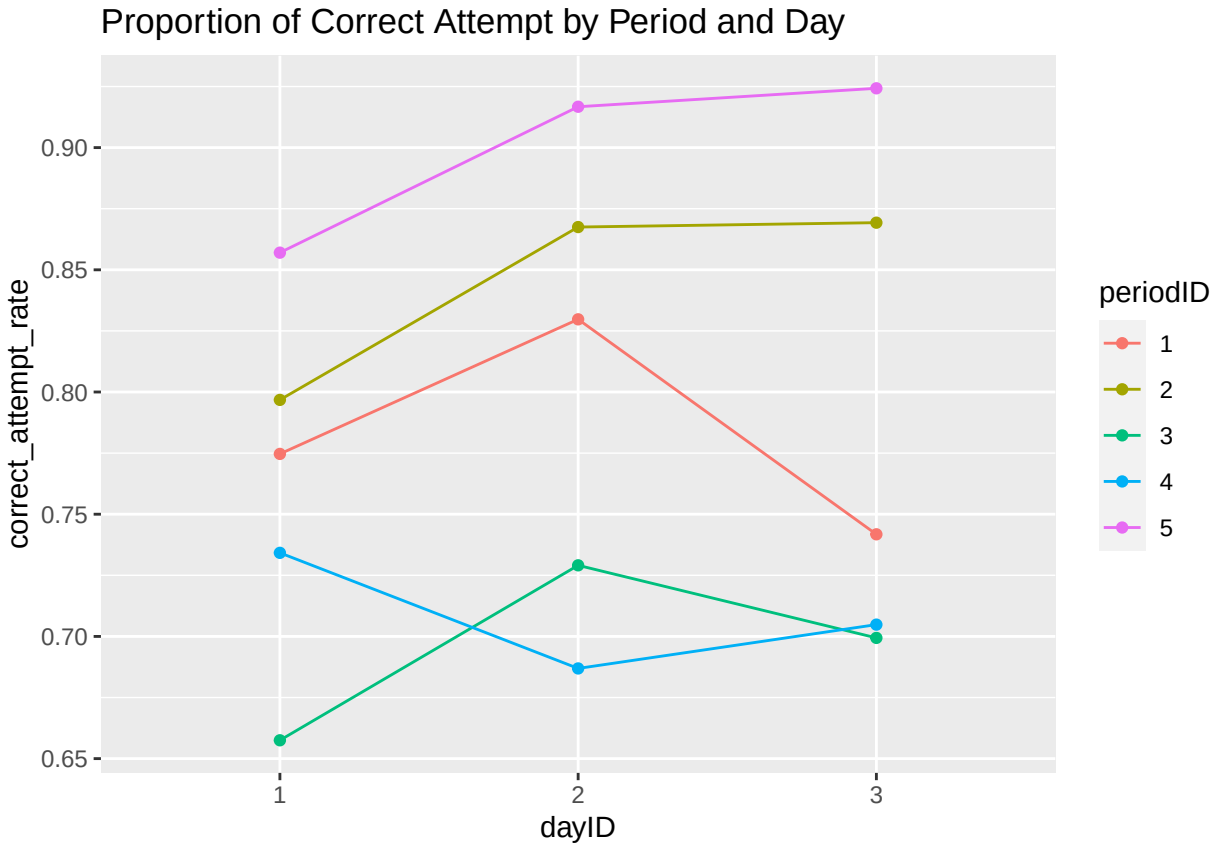
Below are some example summary statistics visuals that showcase these tools and provide a basic idea on the tutor data.

This is a portion of the tutor log summary data file where summary statistics is generated for each day/period.

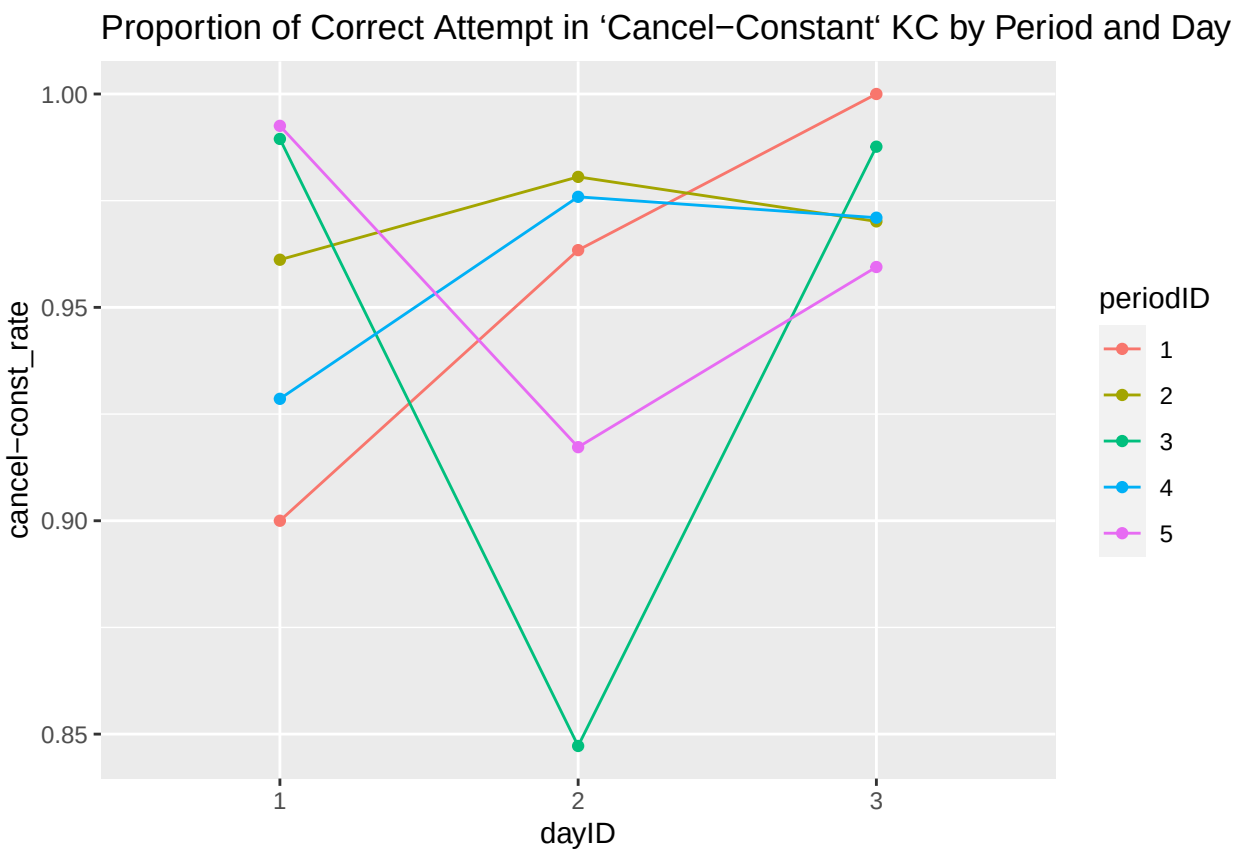
```
## # A tibble: 6 x 20
##   dayID periodID correct_attempt_rate total_n_problems total_n_hints ave_n_hint
##   <fct> <fct>                  <dbl>            <dbl>         <dbl>      <dbl>
## 1 1     1                      0.775               86           190      2.21
## 2 1     2                      0.797              173           342      1.98
## 3 1     3                      0.658              193            77      0.399
## 4 1     4                      0.734              156           142      0.910
## 5 1     5                      0.857              233            29      0.124
## 6 2     1                      0.830              146           327      2.24
```

```
## # ... with 14 more variables: time_per_problem_mean <dbl>,
## #   time_per_problem_sd <dbl>, problem_level_mean <dbl>,
## #   problem_level_std <dbl>, `cancel-const_rate` <dbl>,
## #   `division-simple_rate` <dbl>, divide_rate <dbl>,
## #   `subtraction-const_rate` <dbl>, `combine-like-const_rate` <dbl>,
## #   `subtraction-var_rate` <dbl>, `combine-like-var_rate` <dbl>,
## #   `cancel-var_rate` <dbl>, `distribute-division_rate` <dbl>, ...
```

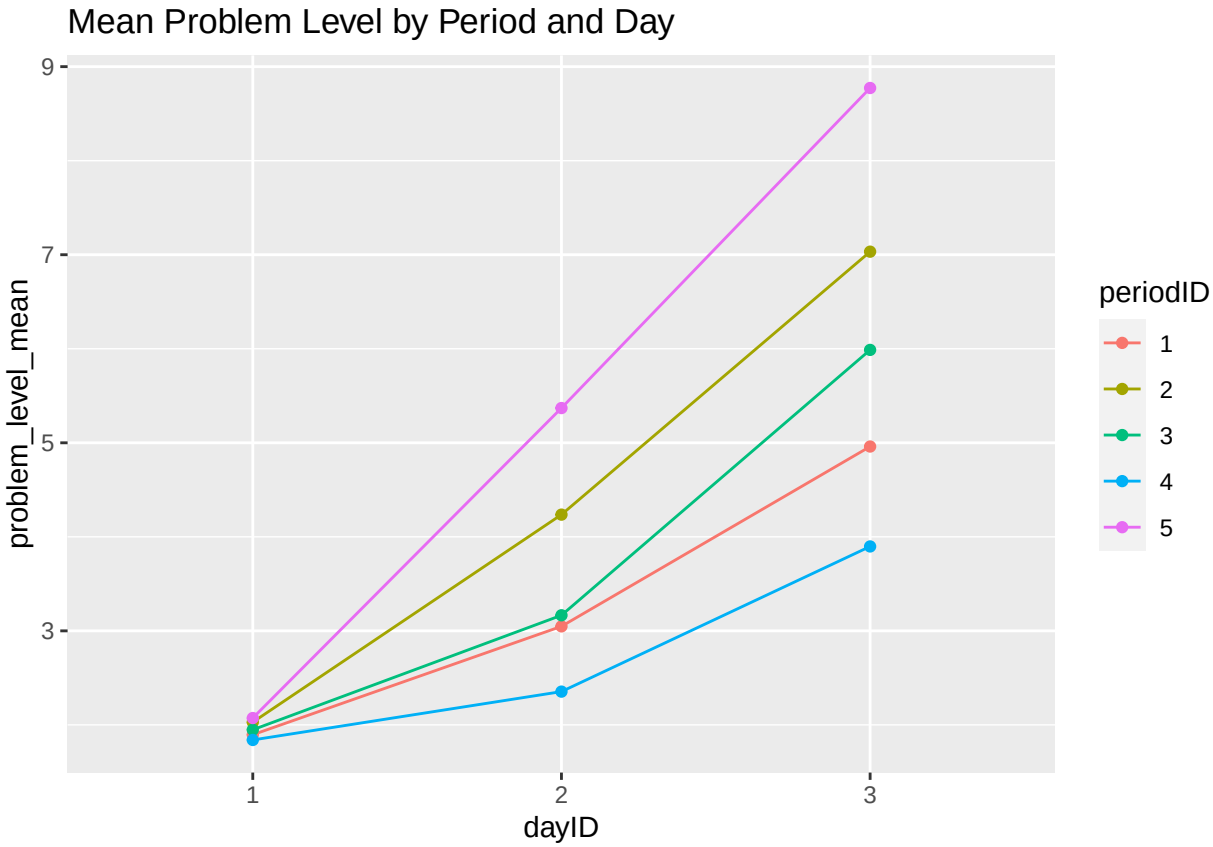
Below is the correct attempt rate for each of the 5 periods during the 3 days. Period 5 is high above other periods since it is considered the "honors" class.



Below is the mean correct attempt rate for each period for the "cancel-constant" KC.



Below is the mean problem level during the 3 days for the 5 periods. The higher the level, the more difficult problems are. Here one can tell that students are given more complex problems over the days.



Events from Tutor Data

Events from the tutor log data are of two types: *attempt* and *hint request*, where *attempt* was further decomposed into *correct attempt* and *incorrect attempt* later when merged with other modalities. These events all have invididual students (identified by `anon_user_id`) as `actor` and `tutor` as `subject`.

For *correct attempt* and *incorrect attempt* events, `content` column would specify corresponding outcome, i.e., correct and incorrect, student text input, and peoblem level in the Lynette system. For *hint request* event, the `content` column would include information on the hint message displayed for the student and the problem level.

Merging Three Modalities

Once the tutor events are distilled, they are ready to be merged with the events from position and observation data. The overall merging process is simple, which involves concatenating the three dataframes within Pandas. However, one aspect in the observation event data was modified to better accommodate the needs of ENA model.

Previously, the `talking to small group` event in the observation data takes multiple students as `subject`. Whereas the ENA team suggested splitting this event into multiple lines where all other information is the same but with one student per rows so that the ENA model can better process this event.