[redacted] - March 2022

1 Introduction

Online learning has become an increasingly critical medium for students to engage with a wide variety of subjects. In addition to supporting traditional compulsory education at an unprecedented capacity during pandemic-related lock downs, this medium underpins substantial industries offering educational opportunities at a variety of levels to a myriad of audiences. Exploring data generated from student interactions with learning material offers the promise of better understanding of mediated learning interactions leading to better results. A key source for useful data of this nature is found in interaction logs of Learning Management Systems (LMS). An LMS is a (typically online) interface facilitating a student's interaction with learning and assessment materials (Li et al., 2018).

One challenge is defining and capturing data meaningful to operational definitions that reflect results in learning experiences. For example, in cases where structured input provide a clear way to validate accuracy, a 'correct' input may be less demonstrable of understanding and/or successful learning (e.g. a true/false question). On the other hand, even meaningfully parsing unstructured input can be difficult even before any attempt is made to describe accuracy (e.g. a written response). Beyond this, there are measuring ambiguities which make answering meaningful questions with data about mediated learning experiences incredibly noisy problems. For example, a count of the seconds it takes a pupil to provide input to a question cannot reliably adjust for distractions that might cause a delay - distractions that are all the more prevalent considering that the content is being accessed in a relatively uncontrolled environment (e.g. on a device in the home).

Analysis of this type represents a big data challenge for a number of reasons. First, modern platforms regularly engage with tens or hundreds of thousands of learners simultaneously. These interactions generate a massive amount of unstructured data in addition to tabular logs like what is used here. The granularity of available data and thus potential feature space is truly massive. At the same time, efficient processes for producing actionable information from this glut of data hold potential for optimizing learning processes and experiences. Parsing and grouping these data in meaningful ways is influenced by a myriad of cultural and platform-specific characteristics.

In this analysis, we will first perform some exploratory analysis then profile students experiences using the Junyi Academy online learning platform by aggregation and grouping data that appear in the record of LMS content interactions. We will then deploy a interpretable machine learning model in order to elucidate the weight different facets of student experience appear to contribute to various result profiles. Technology used includes Hadoop Distributed File System for data storage and access. Pyspark is employed in exploratory data analysis, aggregation, and normalization (namely, pyspark sql functions and dataframes). For vectorization, pipeline creation, and modeling tasks, various modules of pyspark MLlib are employed. For modeling, a Random Forest classifier is used in order to allow for interpretable feature importance.

1.1 Data Overview and Summary

The data used in this analysis consist of two tables, one containing data on 16,217,311 attempts to answer questions in an online learning platform between August 2018 and July 2019, and the second containing profile and demographic information about the 72,630 students. This data includes unique

student identifiers, though it is sufficiently psudonymized so as to not be personally identifiable. The uncompressed tabular data of the problem attempts is 3.02 gb in size. It is released under the creative commons licence (version 4.0) and is available on Kaggle.

Selected summary statistics and notes are provided below (see included ipynb file for calculating these):

1.1.1 User Data Summary Tables

Metric	Gender	Points	Badges Count	User Grade
Count	32905	72758	72758	72758
Mean	NA	63047	9.54	5.62
Stddev	NA	124204	19.03	2.04
Min	NA	1	0	1
Max	NA	4047528	760	12

Table 1: Summary statistics for user data.

User Grade	Count
1	1388
2	1815
3	8324
4	10052
5	14720
6	11568
7	12760
8	6533
9	3436
10	1068
11	570
12	524

Table 2: Frequency table for user grade.

The high proportion of null values in gender indicate that this field may not be particularly useful for analysis.

1.1.2 Log Data Summary Tables

Metric	Exercise Session Repeat	Total Seconds	Hints Used	Level
Count	16217311	16217311	16217311	16217311
Mean	1.26	44.38	0.53	0.483
Stddev	1.17	100.54	1.16	0.94
Min	1	0	0	0
Max	119	1800	65	4

Table 3: Summary statistics for log data

Answer	Count
Correct	11412558
Incorrect	4804753

Table 4: Count of correct and incorrect answers

Times Repeated	Count
1	13809714
2	1640868
3	424514
4	159510
5	72507
6+	104546

Table 5: Count of repeated exercise sessions

1.2 Hypotheses

Initial exploration of the LMS and user records indicate that there is a wide range of level of interaction. Specifically, some students appear to use the platform once, or for a very short period. This is of interest, both for questions related to optimizing educational outcomes as well as business and user experience concerns. For this purpose, we will profile and label students based on if they interact with the platform only once, more than once but fewer than 5 times, and 5 or more times (to disambiguate interrupted study sessions, we consider a single interaction a distinct day where the student has logged in and attempted at least one question). We aim to test the following hypotheses based on this:

1. The standardized length of time a student takes to answer a particular question is inversely correlated to accuracy.
2. Motivational features (i.e. "points" and "badges") are a more prominent factor in frequent interaction with lower levels.
3. Using hints in answering questions is an important feature in modeling students type of interaction and correlates to more frequent interaction.

1.3 Planned Analysis

In order to elucidate these hypotheses, we will initially aggregate data from the LMS logs and join these to the user table into level-specific data sets. Aggregations will summarize student interaction patterns and usage relevant to our hypotheses, including number of distinct days they studied, the percent of their attempts that are on repeated material, their average usage of hint functionality over the course of their studies, and the average length of time taken to answer questions (standardized based on the average time taken on the particular question for all attempts in the logs).

Afterwards, we will classify student interaction frequency according to our operational definitions by applying a labeling function to the joined user data. To investigate feature importance, we will then use a Random Forest Classifier to model features to the interaction pattern label using a parameterized grid search to limit number of trees and forest depth. Finally, the best trained models are visualized for the data overall and by level for feature importance and distinctions relevant to our hypothesis are discussed.

2 Implementation

See the attached Jupyter notebook for implementation code.

3 Summary & Conclusions

3.1 Length and Accuracy

Time taken standardized by question	Incorrect	Correct
25% or less	20%	80%
26 - 50%	25%	75%
51 - 75%	30%	60%
76 - 100%	34%	66%
101 - 125%	37%	63%
126 - 150%	39%	61%
151 - 175%	41%	59%
176 - 200%	43%	57%
more than 200%	47%	53%

Table 6: Accuracy by time taken.

Measuring the rate of correct answers by standardized completion time reveals that our first hypothesis appears correct. The longer a pupil takes to answer a question (as compared to the average completion time for the question), the lower their accuracy tends to be. This suggests potential for live monitoring for intervention support may be useful. For example, as a pupil exceeds the average completion time for a question, unprompted hints or contextual prompting may be beneficial. It may also be the case that lengthier completion times correspond to distracted pupils.

3.2 Motivational Features

User Grade	No badges	One Badge	Multiple Badges
1	27%	19%	54%
2	22%	18%	60%
3	32%	15%	53%
4	32%	17%	52%
5	23%	14%	63%
6	27%	13%	60%
7	27%	13%	60%
8	25%	11%	63%
9	27%	12%	61%
10	34%	13%	53%
11	31%	12%	56%
12	33%	11%	56%

Table 7: Badges count by grade level.

There does not appear to be much pronounced difference in users' "badges" count by grade level. This does not support our hypothesis, suggesting rather that badges as a motivational feature are achieved more or less evenly across grade levels.

3.3 Distinct Interactions & Feature Importance

Hints usage appears to have more impact for distinct usage frequency for questions of level 2 and 3, dropping off for lower and higher level questions. Average timing on the other hand, appears to increase in significance with levels, though the importance of this feature is not overly pronounced. Repeated sessions appear to be significant in influencing pupils engaging with material in lower and higher levels.

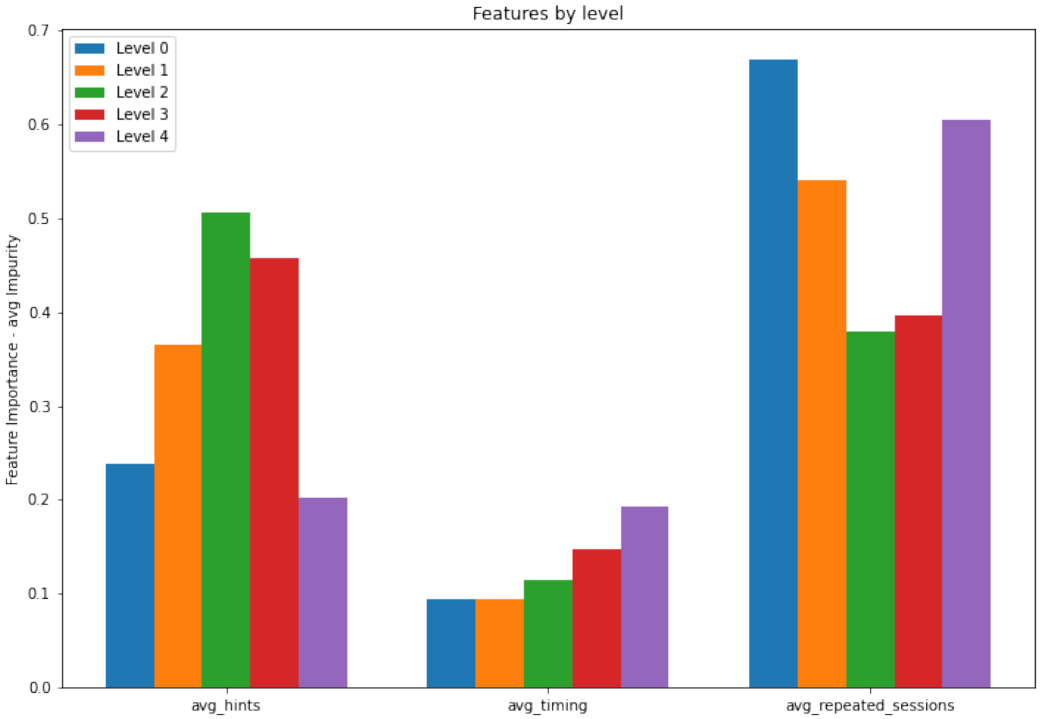

Figure 1: Feature Importance - distinct days of interaction by question level

3.4 Discussion

This analysis utilized pyspark and hadoop distributed file system to address a set of hypothesises made regarding a "big" data set consisting of log files and user data from a commercial LMS. These analysis suggest a number of further enquires related to pupil performance and interaction patterns with materials. First, monitoring the standardized length of time a pupil takes to complete a question may offer opportunities to intervene with appropriate support as their accuracy tends to decrease with time. This possibly corresponds to lack of understanding, distraction, and/or frustration. The "hint" system in use on this platform, however, did not appear to contribute to student success, as suggested by the sharp decline in accuracy rate as hint usage increased. The number of sessions where material was repeated appeared to be the most significant factor in classification models profiling students' tendency to use the platform regularly. The nature of this relationship is likely of critical importance for questions of usability, as well as learning concerns, and customer retention. Distinctions however in how students interact with materials of different levels appears quite pronounced. Modeling predictive of distinct student interactions varies across levels when considering the same features. This suggests that educational and business concerns related to consistent product use should consider different levels of material differently.

One limitation to consider in this investigation is the interpretation of feature importances. Average impurity decrease by feature may show bias towards features with higher cardinality. It may be useful to confirm trends shown here with feature permutation analysis on holdout test sets of smaller datasets. Another notable limitation is the limited time window of available data. Though users who joined in the latter quarter of the available time window were excluded from modeled data to limit the bias towards non-repeat users they might add, it is not infeasible that some bias remains in users who joined within the fist three quarters and returned to using the platform later.

3.5 Works Cited

Li, R., Singh, J. Bunk, J. (2018). Technology Tools in Distance Education: A Review of Faculty Adoption. In T. Bastiaens, J. Van Braak, M. Brown, L. Cantoni, M. Castro, R. Christensen, G. Davidson-Shivers, K. DePryck, M. Ebner, M. Fominykh, C. Fulford, S. Hatzipanagos, G. Knezek, K. Kreijns, G. Marks, E. Sointu, E. Korsgaard Sorensen, J. Viteli, J. Voogt, P. Weber, E. Weippl O. Zawacki-Richter (Eds.), Proceedings of EdMedia: World Conference on Educational Media and Technology (pp. 1982-1987). Amsterdam, Netherlands: Association for the Advancement of Computing in Education (AACE). Retrieved March 27, 2022 from https://www.learntechlib.org/primary/p/184436/.

P. J. Chen, M. E. Hsieh, T. Y. Tsai. Junyi Online Learning Dataset: A large-scale public online learning activity dataset from elementary to senior high school students., 2020. Available from https://www.kaggle.com/datasets/junyiacademy/learning-activity-public-dataset-by-junyi-academy