ASSISTments Longitudinal Data Mining Competition 2017: A Preface

Thanaporn Patikorn
Worcester Polytechnic Institute
100 Institute Road
Worcester, MA, 01609
tpatikorn@wpi.edu

Neil Heffernan
Worcester Polytechnic Institute
100 Institute Road
Worcester, MA, 01609
nth@wpi.edu

Ryan Baker
University of Pennsylvania
3700 Walnut Street
Philadelphia, PA 19104
ryanshaunbaker@gmail.com

ABSTRACT

This proceeding includes papers from some of the leading competitors in the ASSISTments Longitudinal Data Mining Competition 2017. In this competition, participants attempted to predict whether students would choose a career in a STEM field or not, making this prediction using a click-stream dataset from middle school students working on math assignments inside ASSISTments, an online tutoring platform. At the conclusion of the competition on December 3rd, 2017, there were 202 participants, 74 of whom submitted predictions at least once. The three winners were announced at the NorthEast Big Data Spoke Meeting at MIT on February 16th, 2018. In this workshop, some of the leading competitors presented their results and what they have learned about the link between behavior in online learning and future STEM career development.

1. INTRODUCTION

During the 10th International Conference on Educational Data Mining in Wuhan, China, the ASSISTments Longitudinal Data Mining Competition was announced by the Big Data for Education Spoke of the Big Data Northeast Innovation Hub, a research hub funded by the U.S. National Science Foundation. This competition used a longitudinal dataset collected on students using ASSISTments, a free online tutoring platform, in 2004 - 2006. The ASSISTments team tracked those students to see who graduated from high schools, who went on to college, what their majors were, and finally if they chose a career in STEM (Science, Technology Engineering and Math) for their first job, post-college. Several papers have shown that behavior in ASSISTments in middle school can predict high school and college outcomes [4] [7][8]. The task given to the participants in this competition was to use deidentified click-stream data to try to predict the whether the student pursued a career in STEM or not. This data was provided to participants to analyze before it was used by the research team themselves, an unusual step that enabled participants in the competition to gain first access to a cutting-edge research data set.

In recent years, there has been increasing interest by school districts and state education agencies in predicting student success and dropout [1][2]. These detectors are used to give early warnings to teachers, guidance counselors, and school leaders when students show signs that they are losing interest or experiencing difficulties. These detectors support teachers making targeted interventions to take necessary actions to help students before it's too late. However, there has thus far been relatively less work to drive K-12 early warning based on students' risk of dropping out the STEM pipeline. This is particularly problematic, given the current economic context. While there is increasing demand for STEM workers, substantial numbers of students lose interest in STEM subjects and fields or are insufficiently prepared to participate in these careers [9]. Developing automated detection of STEM career participation may help us to identify students who could benefit from an intervention to help to support their interest and readiness for STEM [6].

2. ASSISTments Longitudinal Data Mining Competition 2017

The competition ran from June 27, 2017 to December 3, 2017. Registration for the competition and the dataset were entirely free, in line with the goals of promoting 1) STEM education, 2) educational data mining, and 3) open science. The primary condition of accessing the dataset was to not take any action to deanonymize the dataset. Even though the competition has already been concluded, we still welcome interested researchers to sign up for the competition dataset[1].

2.1 Dataset

The dataset in this competition was the ASSISTments clickstream dataset collected during 2004 - 2006. This dataset contained actions middle-school students took while working on their mathematics assignments. In addition to raw recorded actions, participants were also provided with several distilled measures, for instance, measures of the student's affective state and disengaged behaviors (bored, concentrating, confused, frustrated, off-task, and gaming). These measures were obtained by collecting student affect observations in real classroom and then using machine learning techniques to train models that replicated those judgments within a clickstream dataset [5]. The detectors were validated to ensure that they applied effectively to unseen students from urban, rural, and suburban settings [3]. The dataset contains 78 clickstream data predictor variables and the target

[1] You can sign up for the dataset here:
https://goo.gl/forms/seAyF0aHUOxevhfF3
The description of the dataset can be found in the competition website:
https://sites.google.com/view/assistmentsdatamining/data-mining-competition-2017

variable "isSTEM": whether the student's career of choice was in the STEM fields or not, defined using the NSF guidelines for STEM careers. There are 942,816 action-level data rows collected from 1,709 students in total. For the competition, the dataset was split into 3 sets: the training set, the validation set, and the test set.

2.1.1 Training Set

The training set contained the majority of the students from the full dataset. For each student in this dataset, both the students' action-level ASSISTments usage data and their "isSTEM" variable were available. Participants, as well as any researchers who are interested in STEM education, could make full use of this dataset, using any state-of-the-art data mining technique they chose to find the relationships between the student actions and their career choice (as long as it does not violate the terms of use).

During the data collection, there were many students for whom we collected ASSISTments usage data, but we were unable to retrieve their career information. Specifically, we know the isSTEM for only 591 students out of 1,709 students. We decided to include the ASSISTments usage data of these students in the training set since there are many co-training machine learning approaches that could train a model by using unlabeled data along with labeled data. The training set contains 514 labeled students and 1,118 unlabeled students.

2.1.2 Validation Set

The validation set was mainly used for the public leaderboard. This leaderboard let participants know how well they were doing compared to other participants. All clickstream data from students in the validation set were made available to participants. Participants, however, were unable to directly access the "isSTEM" variable for the students in the validation set. When ready, participants could submit their prediction for the validation set’s isSTEM students. The system would then evaluate the predictions, inform participant of their scores, and then update the participant's best scores on the leaderboard. The evaluation scheme will be further discussed in the later section.

2.1.3 Test Set

The only purpose of the test set was to be used to determine the winner of the competition. Like the validation set, participants could only access the clickstream data of students in this set and not their isSTEM. The difference between the validation and the test set was that the test set was not used to calculate the leaderboard scores; the results were not visible until after the competition was complete. The reason we chose to separate the test set from the validation set was to make sure that the winners of the competition were not simply participants who overfit using the leaderboard, but who genuinely could predict entirely unseen data.

2.2 Evaluation

For the evaluation of models, participants were required to submit their predictions for students in both the validation set and the test set. Participants, however, were not informed as to which students were in which set. Once a day at noon EST, new submissions were evaluated on the validation set. While participants could submit as many predictions as they wanted, only the participant's latest submission was evaluated, to discourage them from overfitting to the leaderboard. The system then updated each participant's personal submission log with their latest submission's scores as well as the public leaderboard, where each participant's best scores were shown compared to other participants' best scores.

2.2.1 Evaluation Criteria

Both the leaderboard scores and the final scores were calculated by using a linear combination of the area under the ROC curve (AUC) and the root mean squared error (RMSE). Since isSTEM was observed and collected as binary values, AUC was initially chosen as the evaluation criterion. AUC captures the model’s ability to differentiate students in the two categories from each other, based on the relative confidence in the predictions. It is most suitable when the variable being predicted is binary and the predictions are numerical. However, after testing, we found that AUC, or any single metric, could be easily overfit to, especially given the small sample size.

Thus, we selected a second evaluation criterion: RMSE. While RMSE is designed for comparing two numbers, it provides an assessment that rewards models that are more certain when they are correct and punishes models that are uncertain with high confidence. It also maps to a context of use where the model provides different recommendations when it is uncertain than when it is highly confident.

For the sake of the competition, we decided to aggregate the two metrics, AUC and RMSE, into one score so that we could determine the winners. Since AUC ranges from 0 (reverse ranking) to 1 (perfect ranking) and RMSE, in this case, ranges from 0 (perfect predictions) to 1 (total opposite predictions), we define Aggregated Score as a linear combination of the two metrics, with one metric inverted:

$$\text{Aggregated Score} = \text{AUC} + (1 - \text{RMSE})$$

2.3 Different Population from Training to Validation and Test Sets

In October 2017, we discovered that the distribution of isSTEM within the training set was not the same as that of validation and test set. Specifically, the ratio of isSTEM = true and isSTEM = false of the validation set and test set were the same, but that ratio of the training set was more than double that of the validation set and test set. We investigated the issue and decided to keep the three sets as they were and announced this information to all participants. The reasons we decided to keep the data sets unchanged were 1) it is not uncommon for models to be applied to a context with different distribution and/or population from the training set. The difference between the sets, while they were not intended, did emulate this possible real application issue. 2) the isSTEM ratio of the validation set and the test set were the same, meaning participants could use the result from the validation set to adjust for the discrepancies between the training and the validation set, which would be reflected in the test set, since the isSTEM distribution of the validation and test sets were the same.

3. Conclusion of the Competition

The competition was concluded on December 3rd, 2017. At the conclusion of the competition, 202 participants had signed up for the competition, 74 of whom submitted predictions at least once.

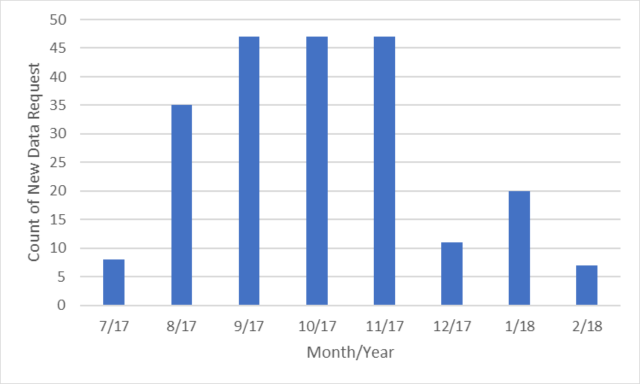

Figure 1: the number of new unique emails that signed up for the competition dataset in each month from July, 2017 to February 2018.

3.1 Data Request Over Time

Most of the requests for the dataset were from August 2017 to November 2017. Since one of our main goals is to promote research in this area, we were glad to see that requests for the dataset continued even after the competition ended in December.

3.2 Submissions Over Time

At the first glance, the number of submissions peaked during November 2017, which was the last full month before the competition concluded. However, since the competition concluded on December 3rd, 2017, December 2017 was the month with the most submissions per day of 19.33, more than double the rate in November 2017 (9.19 submissions per day).

Among all participants who submitted predictions at least once, about two-third of them submitted more than once, and only about one-sixth submitted more than ten times. Only 8 participants submitted more than 20 times.

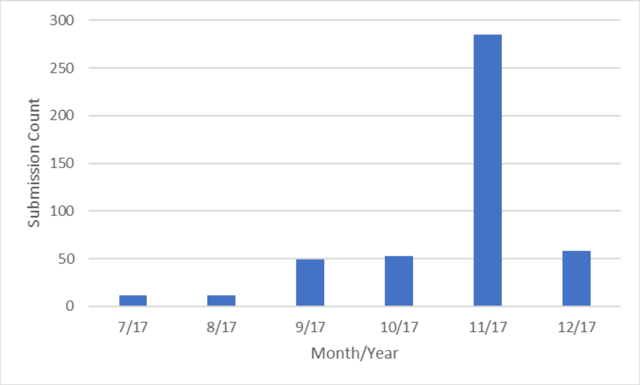

Figure 2: the number of submissions evaluated by the system in each month from July, 2017 to December 2017.

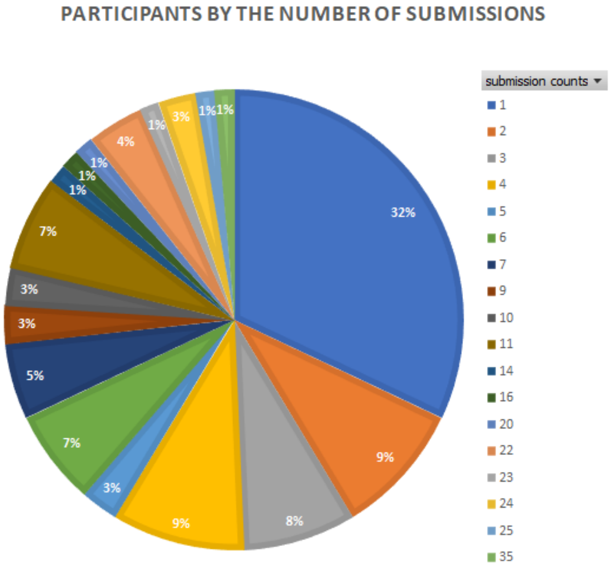

Figure 3: the percentage of participants by the number of submissions they made during the competition.

3.3 Submissions Scores Over Time

Overall, the quality of submitted predictions averaged across all participants appeared to increase slightly over the months as shown in Figure 4. While the average scores seemed to plateau after October, it is important to note that there were many participants who joined later in the competition. Their scores were averaged together with other participants who had already worked on the competition. We further investigated by looking at the aggregated score of the 1st, 2nd, 3rd, etc. submissions averaged across all participants, which is shown in Figure 5. A similar increasing trend to Figure 4 can also be observed in Figure 5. It is important to note that there were only 8 participants who submitted more than 20 times, which could be one of the reasons why the graph fluctuates a lot when x > 20.

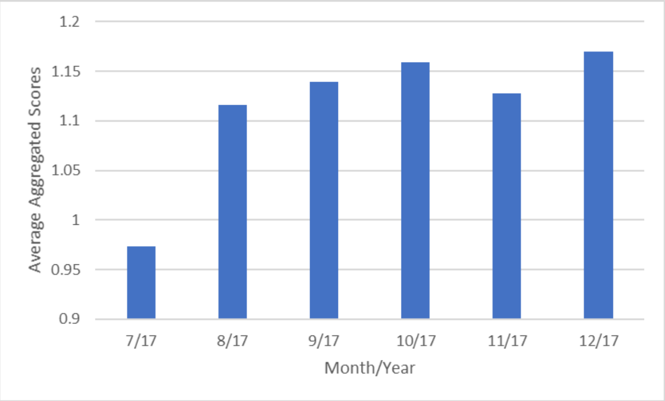

Figure 4: the aggregated scores averaged across all participant predictions submitted and evaluated in each month from July, 2017 to December 2017.

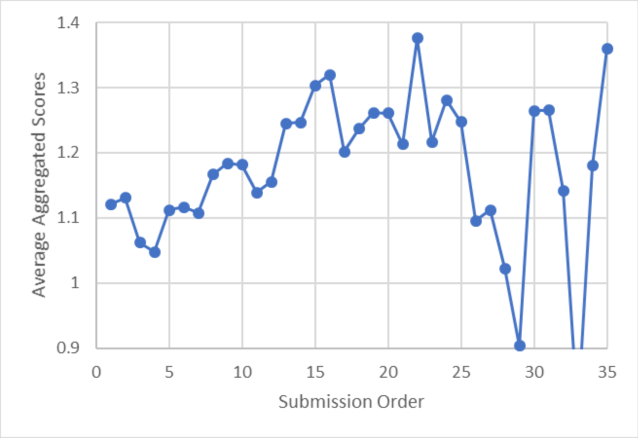

Figure 5: the aggregated scores by the submission order of each participant, averaged across participants from July, 2017 to December 2017. For example, the average aggregated scores of everyone's second submission is the data point at x = 2.

3.4 Winners

The three winners were announced during the NorthEast Big Data Spoke Meeting at MIT on February 16th 2018. The first place winning team of Chun Kit Yeung, Kai Yang, and Dit-yan Yeung is from the Hong Kong University of Science and Technology, who participated in the workshop. The second place winner was Makhlouf Jihed from Japan's Kyushu University, who also participated in the workshop. The third place honors went to the University of Michigan Data Science Team, a group that regularly competes in data competitions like this one.

4. ACKNOWLEDGMENTS

This competition was supported by NSF grant #DRL-1661987. Much of the data collection was supported by NSF grant #1031398.

The ASSISTments platform is made possible via multiple grants from NSF grants (ACI-1440753, DRL-1252297, DRL-1109483, DRL-1316736, DGE-1535428, OAC-1724889, OAC-1636782 & DRL-1031398), the US Dept. of Ed (IES R305A180401, R305A120125 & R305C100024 and GAANN), and the Office of Naval Research.

5. REFERENCES

[1] Bowers, A. J. (2010). Grades and graduation: A longitudinal risk perspective to identify student dropouts. *The Journal of Educational Research*, 103(3), 191-207.

[2] Knowles, J. E. (2015). Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin. *Journal of Educational Data Mining*, 7(3), 18-67.

[3] Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., Heffernan, C. (2014) Population validity for Educational Data Mining models: A case study in affect detection. *British Journal of Educational Technology*, 45 (3), 487-501.

[4] Ocumpaugh, J., San Pedro, M.O., Lai, H-y., Baker, R.S., Borgen, F. (in press) Middle School Engagement with Mathematics Software and Later Interest and Self-Efficacy for STEM Careers. To appear in *Journal of Science Education and Technology*.

[5] Pardos, Z.A., Baker, R.S.J.d., San Pedro, M.O.C.Z., Gowda, S.M., Gowda, S.M. (2014) Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. *Journal of Learning Analytics*, 1(1), 107–128.

[6] Reider, D., Knestis, K., & Malyn-Smith, J. (2016). Workforce education models for K-12 STEM education programs: Reflections on, and implications for, the NSF ITEST program. *Journal of Science Education and Technology*, 25(6), 847-858.

[7] San Pedro, M., Baker, R., Bowers, A. & Heffernan, N. (2013) Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School. In S. D'Mello, R. Calvo, & A. Olney (Eds.) *Proceedings of the 6th International Conference on Educational Data Mining*

[8] San Pedro, M.O., Baker, R., Heffernan, N., Ocumpaugh, J. (2015) Exploring College Major Choice and Middle School Student Behavior, Affect and Learning: What Happens to Students Who Game the System? *Proceedings of the 5th International Learning Analytics and Knowledge Conference*. pp 36-40.

[9] Sass, T. R. (2015). Understanding the STEM Pipeline. Working Paper 125. *National Center for Analysis of Longitudinal Data in Education Research (CALDER).*