Predication and Analysis of STEM and Non-STEM College Major Enrollment with ASSISTments

Guimin Dong
Department of Computer Science, 2301 Vanderbilt Place
Nashville, TN 37235-1826 USA

***Abstract*—ASSISTments, developed by researchers years ago, is a mathematics tutoring system designed for middle school students. Many middle school students have used this software, and a fraction of them are now attending colleges. Previously, researchers created a logistic regression model to predict students' choice of college majors (STEM or NON-STEM), and have done some detailed statistical analysis. In this project, we built models for predicting students' college majors (STEM or NON-STEM) by training Support Vector Machine and Multilayer Perceptron with different combination of features generated from log files of ASSISTments and achieved higher accuracy, and constructed Unweighted Pair Group Method with Arithmetic Mean (UPGMA)to cluster five different groups by using similar features from previous research. From the results, we found that our new set of features selected empirically by ensembled feature selection method can increase accuracy of our classifier and makes such feature set more explanatory to predict STEM or Non-STEM majors enrollment in college, and unsupervised UPGMA clustered students into five groups providing some insight of personal characteristics in choosing STEM or Non-STEM majors.**

Keywords: ASSISTments; logistic regression; Random Forest; support vector machine; Multi-Layer Perceptron; Unweighted Pair Group Method (UPGMA) with Arithmetic Mean; Educational Data Mining.

I. Introduction

Research shows that middle school is a crucial juncture for a student to start thinking about his or her academic achievement, college attendance, and future career. And college and university degree programs in science, technology, engineering and mathematics (STEM) are considered STEM degrees, and they are in high demand across many industries. Several years ago, educators and industry analysts detected a trend that indicated an academic deficiency in STEM areas for students entering college. It is important to let the students to be well-prepared with stem skills for stressful and rigorous college-level stem major courses [1]. Given the fact that increasing numbers of students are getting access to computers and even mobile phones these days, researchers designed various educational software and conducted research using data from the middle school students' interaction with the software. Using such data, researchers did a large amount of statistical analysis. 66% accuracy of predicting whether students will choose STEM or non-STEM major was achieved by using logistic regression model[2]. Although disengagement, which includes boredom, off-task, and frustration, has negative impact on attitude of higher education and causes low learning achievement[3, 4, 5] and these features were used to predict STEM and Non-STEM major selection[1], this features detected and extracted by the Intelligent Tutoring System, ASSISTments, have small contribution to predict major selection in our analysis.

Our motivation for this project, firstly, logistic regression model has rigid assumption about multicollinearity, linearity of independent variables and log odds, and independent observations, and previous studies performed logistic regression without verifying such assumptions. Secondly, better representation of samples will lead more accurate and more robust classification model, previous study did not examine broader selection of features to train machine learning models. Thirdly, hyperparameter optimization to success higher accuracy was not discussed in previous studies. What's more, unsupervised learning methods were not utilized. Last but not the least, in previous study, disengaged behaviors (carelessness, gaming the system, and off-task behavior), and educationally-relevant affecting states (boredom, engaged concentration, confusion, frustration) were considered to predict STEM or Non-STEM major, however these superficial negative learning behaviors may not reflect students' real state: some student who choose STEM major and are good at mathematics may feel bored about the questions showed in computer, and some other students prefer discovery and experiments will not cooperate very well with ASSISTments.

The contribution of our analysis in educational data mining: an ensembled method to empirically extract most 35 important features to train our machine models was performed, and we constructed SVM with 0.826 mean test score and 0.105 standard deviation in the 10 fold grid search cross validation to predict STEM or Non-STEM major selection with 76.35% accuracy in our test data set. Psychological point of view of students' STEM or Non-STEM majors selection was discovered by using UPGMA that careful learners and experimenters have higher probability to choose STEM major when they are in college.

II. Methodology

A. ASSISTments: Mathematics Educational Software

This study is based on students' log files generated from ASSISTments system which is an intelligent tutoring system students can interact with to learn knowledge[7], Specially, ASSISTments is a free web-based mathematics tutoring system for middle-school mathematics, developed by Dr. Neil Hefferman, was used in a series of research to understand

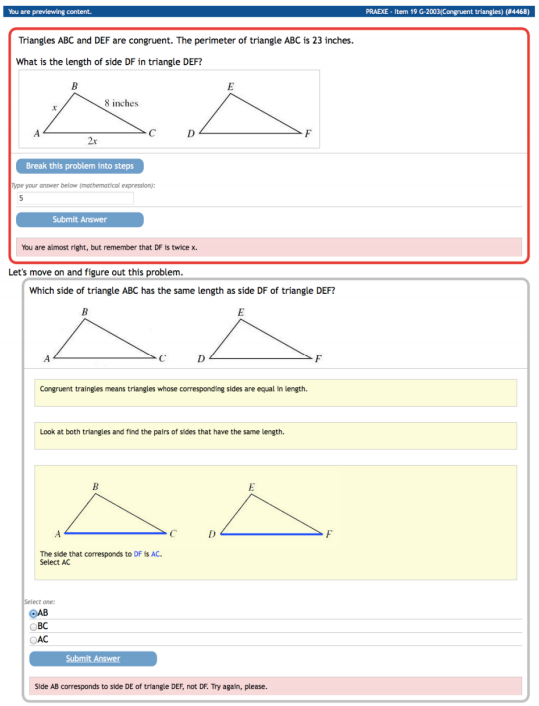

Fig. 1. ASSISTments

	RF	RFECV	lda	logReg	svc	Mean
AveCarelessness	1.0	0.0	0.48	0.0	0.0	0.3
AveCorrect	0.4	0.0	0.46	0.01	0.0	0.17
AveKnow	0.34	0.0	0.07	0.0	0.0	0.08
AveResBored	0.08	0.0	1.0	0.0	0.0	0.22
AveResConf	0.01	0.0	0.09	0.0	0.0	0.02
AveResEngcon	0.0	0.0	0.03	0.01	0.0	0.01
AveResFrust	0.34	0.0	0.16	0.0	0.0	0.1
AveResGaming	0.0	0.0	0.16	0.0	0.0	0.03
AveResOfftask	0.06	0.0	0.1	0.0	0.0	0.03
Ln	0.27	0.0	0.59	0.0	0.0	0.17
Ln-1	0.72	0.0	0.59	0.0	0.0	0.26
NumActions	0.01	0.6	0.0	0.56	0.41	0.32
RES_BORED	0.01	0.0	1.0	0.0	0.0	0.2
RES_CONCENTRATING	0.0	0.0	0.03	0.01	0.0	0.01
RES_CONFUSED	0.05	0.0	0.09	0.0	0.0	0.03
RES_FRUSTRATED	0.03	0.0	0.16	0.0	0.0	0.04
RES_GAMING	0.0	0.0	0.16	0.0	0.0	0.03
RES_OFFTASK	0.01	0.0	0.1	0.0	0.0	0.02
attemptCount	0.15	0.0	0.06	0.05	0.02	0.06
bottomHint	0.22	0.0	0.56	0.0	0.0	0.16
consecutiveErrorsInRow	0.04	0.0	0.13	0.0	0.02	0.04
correct	0.29	0.0	0.46	0.01	0.0	0.15
endsWithAutoScaffolding	0.0	0.87	0.77	0.0	0.0	0.33
endsWithScaffolding	0.0	0.0	0.21	0.01	0.01	0.05
frIsHelpRequest	0.04	0.0	0.32	0.01	0.0	0.07
frIsHelpRequestScaffolding	0.0	0.0	0.1	0.01	0.01	0.02
frPast5HelpRequest	0.38	0.0	0.24	0.03	0.0	0.13
frPast5WrongCount	0.16	0.0	0.05	0.01	0.01	0.05
frPast8HelpRequest	0.18	0.0	0.15	0.04	0.01	0.08
frPast8WrongCount	0.0	0.0	0.04	0.01	0.01	0.01
frTimeTakenOnScaffolding	0.05	0.27	0.0	0.24	0.55	0.22
frTotalSkillOpportunitiesScaffolding	0.04	0.0	0.0	0.05	0.0	0.02
frWorkingInSchool	0.0	0.0	0.07	0.02	0.01	0.02
helpAccessUnder2Sec	0.02	0.0	0.71	0.0	0.0	0.15
hint	0.28	0.0	0.49	0.01	0.0	0.16
hintCount	0.14	0.0	0.1	0.03	0.03	0.06
hintTotal	0.13	0.0	0.01	0.05	0.04	0.05
manywrong	0.27	0.07	0.36	0.02	0.0	0.14
original	0.32	0.0	0.12	0.01	0.0	0.09
past8BottomOut	0.07	0.0	0.02	0.01	0.01	0.02
responseIsChosen	0.0	1.0	0.0	0.0	0.0	0.2
responseIsFillIn	0.0	0.0	0.28	0.0	0.0	0.06
scaffold	0.0	0.2	0.63	0.0	0.01	0.17
stlHintUsed	0.0	0.8	0.45	0.0	0.0	0.25
sumTimePerSkill	0.05	0.53	0.0	0.47	0.46	0.3
timeGreater10SecAndNextActionRight	0.1	0.0	0.79	0.0	0.0	0.18
timeGreater5Secprev2wrong	0.0	0.93	0.18	0.0	0.0	0.22
timeOver80	0.04	0.0	0.02	0.0	0.0	0.01
timeSinceSkill	0.07	0.73	0.0	0.0	0.0	0.16
timeTaken	0.05	0.33	0.0	0.26	0.42	0.21
totalFrAttempted	0.02	0.4	0.0	0.72	1.0	0.43
totalFrPastWrongCount	0.06	0.0	0.02	0.02	0.03	0.03
totalFrPercentPastWrong	0.0	0.0	0.08	0.0	0.0	0.02
totalFrSkillOpportunities	0.03	0.0	0.0	0.1	0.04	0.03
totalFrSkillOpportunitiesByScaffolding	0.02	0.0	0.05	0.01	0.01	0.02
totalFrTimeOnSkill	0.04	0.47	0.0	1.0	0.62	0.43
totalTimeByPercentCorrectForskill	0.28	0.67	0.0	0.25	0.04	0.25
MCAS	0.39	0.13	0.0	0.23	0.52	0.25

Fig. 2. Feature Selection by Coefficients

how students behavior is related to their academic performance, college attendances, and college major choices[3]. ASSISTments evaluates a students knowledge level, detects and records a student's interaction, and assists students in understanding concepts, using knowledge to solve problems and transferring learning, providing teachers with insights of students' knowledge development and evaluation of learning performance on specific knowledge components. When learning with this intelligent tutoring system, students' cognitive skill will be assessed by several mathematical problems. If students answer correctly, they proceed to the next problem. If they answer incorrectly, the system scaffolds instruction by dividing the problem into component parts, leading the students to divide and conquer each smaller components of the original problem with hints. (Fig 1). Once the original problem is correctly answered, the students will step into the next.

B. Data

This project used the data from a longitudinal study, now over a decade long, led by Professor Ryan Baker and Professor Neil Heffernan. This study, funded by multiple grants from the National Science Foundation, tracks students from their use of the ASSISTments blended learning platform in middle school in 2004-2007, to their high school course-taking, college enrollment, and first job out of college. There are 87,8110 interactions stored in the log files with 76 features at ASSISTments from 517 students and the following-up STEM or Non-STEM majors chosen by these students when they were in colleges.

C. Data Prepossessing and Feature Engineering

As the log files records the interaction sequence between students and ASSISTments, for each student, differentiated with student ID, we extracted the mean values for each feature of all sequence interaction. Then we dropped the samples without specific STEM or Non-STEM major claim. Then we are left with 492 samples with STEM or Non-STEM major claim. In the feature MCAS, which is Massachusetts Comprehensive Assessment System, -999 indicates the missing value of the score. So, we imputed the missing value with ridge regression with 76.53% adjusted R-squared value. We selected the 54 features out of the whole feature set, for complete description of features please refer APPENDIX. Then we performed supervised learning algorithms including Random Forest(RF), Linear Discriminant Analysis(LDA), logistic regression(LogReg), svm, and Recursive feature elimination with cross-validation(RFECV) by setting the response variable as STEM which is 1, or Non-STEM majors which is 0, and the mean of each coefficient generated from these five algorithms was collected. The features with mean of coefficient less than 0.05 were dropped empirically, as shown in Fig 2. In Fig 2., clearly, concentration, confusion, frustration, off-task, and gaming, where are reflected by rescaled of the confidence of student affect prediction, have small contribution to classify STEM or Non-STEM major chosen. Although the mean coefficient of boredom is 0.2, where LDA assigned 1.0, it is hardly to determine the effect of boredom in the learning process for classification of major chosen. There are

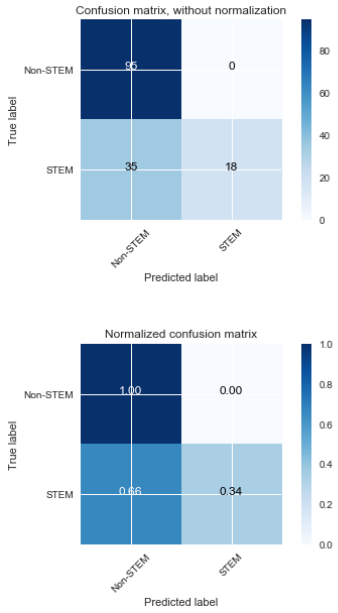

Fig. 3. Confusion Matrix of Trained SVM

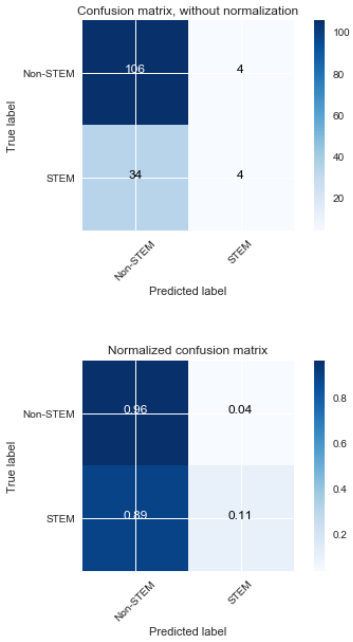

Fig. 4. Confusion Matrix of Trained MLP

some interesting phenomenon that test related features, such as correctness, scaffolding, hint, MCAS, and first response time spent on knowledge component across all problems, have greater impact to major classification, implying that, indeed, STEM major, comparing with Non-STEM major, has higher requirement in test and students who perform well on quantitative field test will have higher probability to choose STEM major and vice versa. Thus different from precious study which used disengaged behaviors and educationally-relevant affecting states to train logistic regression model[2], we included test related effects and also include boredom passing into our machine learning algorithm to train the predicting models. In the next section, we executed supervised and unsupervised learning algorithms to train the models.

III. Supervised and Unsupervised Learning Methods for STEM or Non-STEM Major Classification

A. Supervised Learning

1) Support Vector Machine(SVM): In SVM, binary classification problem was modeled in this case to predict STEM or Non-STEM majors. When we train SVM model, which involved hyperparameters, we cannot decide which combination of hyperparameters will achieve highest accuracy, thus we performed exhaustive grid search with 10 fold cross validation and training data size 0.7 (scikit-learn via python) in the hyperparameters search space, including, kernels(linear, Gaussian, polynomial, sigmoid), penalty for regularization C within the sequence [0.001, 0.01, 0.1, 1, 10, 100, 1000], gamma, which determine how close the points to the hyperplane will be considered, in the sequence of [0.01, 0.1, 1, 10, 100, 1000].

In the grid search with 10 fold cross validation, we achieved highest mean test score 0.826 wih standard deviation 0.105. Then we used the trained SVM to predict the 30% test data set, finally we got 76.35% accuracy, as shown the confusion matrix in Fig 3. In normalized confusion matrix, we can observe that 100% Non-STEM major was predicted correctly, and 34% STEM major was predicted correctly, implying the test related features help to determine the students who will not choose STEM major but cannot help to predict the students who will choose STEM major.

2) Multi-Layer Perceptron(MLP): In MLP scenario, via Keras and scikit learn in python, our hyperparameters search space constituted of: activation : ['identity', 'logistic', 'tanh', 'relu'], solver : ['sgd', 'adam'], learning-rate : ['constant', 'invscaling', 'adaptive'], max-iter : [800, 900, 1000], momentum : [0.5, 0.6, 0.9]. Similarly, in the grid search with 10 fold cross validation, highest mean test score 0.677 was succeed with rectified linear unit(relu) activation function, MLP dimension of $35 \times 70 \times 1$, adaptive learning rate, 800 maximum iteration, 0.5 momentum and stochastic gradient solver. In the 30% test data set, 74.32% accuracy with such MLP model was achieved and the confusion matrices are shown in Fig 4. Comparing to SVM model, we got similar results, and SVM perform better in classify both STEM and Non-STEM major students.

B. Unsupervised Learning

Different from supervised learning, in unsupervised learning part, we use the same set of features as in [2], including knowledge level, correctness, number of actions, boredom,engaged concentration,confusion, frustration, off-task behavior and gaming.

1) Unweighted Pair Group Method with Arithmetic Mean(UPGMA): We built the clustering model using the RapidMiner agglomerate clustering operator with measure

Fig. 5. UPGMA Clustering Dendrogram

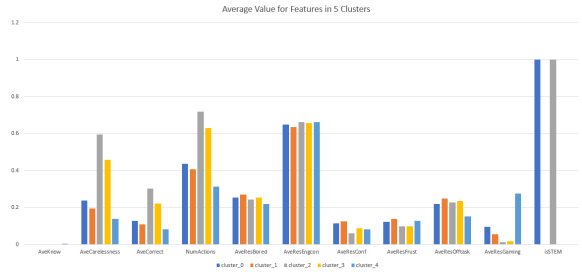

Fig. 6. Statistics of Features for 5 Clusters

type as numerical values on Euclidean distances. By running the UPGMA algorithm, we got the dendrogram shown in Fig 5. To get the actual clusters and their corresponding data points, we need to estimate and choose a proper level in the dendrogram to split the hierarchical data structure into smaller clusters. With careful consideration, we decided to divide the data into five clusters of sizes 99, 189, 18, 61 and 100 [8] (the level at which we split the dendrogram is also shown in Fig 5.

For five clusters we have divided, we could characterize them as careful learners, confused and frustrated learners, experimenters that have chosen STEM majors, experimenters that have chosen NON-STEM majors and gamers. The statistics of features for five clusters is shown in Fig 6.

The first group of 99 students can be characterized as careful learners. They share the relatively low carelessness, which means they tend not to make mistakes if they know the skill required to solve a specific problem. However, since they take every step carefully (thus may be slowly), these students did not complete as many problems in the tutoring system as other students did. As a result, the total number of correct steps they took was not as high as other students. The second group of 189 students can be characterized as confused and frustrated learners. This group of students, just like careful learners, take each step carefully, resulting in a relatively smaller number of actions. They are not careless and would spend more time at each step, but this may be because they are confused about the corresponding knowledge. This group of students is also more frustrated than others according to the statistics, which may lead to the off-task behaviors. The third group of 18 students can be characterized as experimenters that have chosen STEM majors. This groups of students are substantially more careless than other students. However, since they would complete much more actions when using the tutoring system, they could also achieve much more correct steps than others. Considering these students would spend most of their time working on problems in the tutoring system, they showed very few behaviors of gaming the system (playing with non-training part of the tutoring system). The fourth group of 61 students can be characterized as experimenters that have chosen NON-STEM majors. Comparing to the experimenters that have chosen STEM majors, students in this group are more moderate regarding the number of actions they completed. These students did complete substantially more actions than average, but less than the students in the previous group. Besides, students in the fourth group are more careful than the students in the group three. The only apparent difference between students is the fact that students in the group three have finally chosen STEM major in the college, while the students in the other group have chosen NON-STEM major. The last group of 100 students can be characterized as gamers. Students in this group tend to spend much more time gaming the system, rather than working on mathematics problems. They had the most knowledge on average comparing to other groups and showed the least carelessness, which might indicate they are quite talented regarding the mathematics skills. However, students in the last group had the least number of actions in training and spent the most time in gaming the system, which indicates that these students might not be interested in the mathematics, thus not choosing the STEM major. Among the five clusters we got by running the UPGMA algorithm, all the students in group one (careful learners) and group three (experimenters that have chosen STEM majors) finally chose STEM major in the college, and all the students in the left three groups chose NON-STEM major. We can infer from the result that solving mathematics problems carefully and spending most of the time in actual training process (instead of gaming the system) could indicate such students are interested in the study of mathematics or its related subjects, thus finally choosing the STEM major. Besides, students that are willing to take actions when learning mathematics and are not afraid of making mistakes also have relatively higher chances of enrolling in the STEM major in the college.

IV. Discussion and Conclusion

In this project, we examined three different methods for predicting students STEM major enrollment in the college and analyzing datasets inner structure. In supervised learning paradigm, both support vector machine model and multi layer perceptron model improved the prediction accuracy from the previous logistic model with accuracy 66% [2] to 76.35% (SVM) and 74.32% (MLP). Different feature set, which test skill mostly related, was extracted to train SVM and MLP with hyperparameter optimization. The students whose major are Non-STEM can be predicted in a very high precision 100% for SVM and 96% for MLP. However, STEM enrollment of

students cannot be correctly classified in both SVM and MLP, further more, SVM performed better than MLP in classifying STEM major. Thus, we can conclude that less competitive test performance and poor related test skill can help people to predict the students' enrollment in Non-STEM major when they are going into colleges. When a student has considerable amount of time to interact with ASSISTments during middle school, the students' instructors can understand the probability whether this student will choose STEM or Non-STEM major in the future by using the SVM model, provide some advice to the student about his/her future academic plan or career path. What's more, there are still some students' major can not be correctly predicted by our models, Non-STEM major will be predicted in most of cases.

The unsupervised learning method (Unweighted Pair Group Method with Arithmetic Mean) provided us some insight into the detailed structure of the dataset. By dividing the data into five groups according to the dendrogram generated from the algorithm, we know that for some similar groups (the third and forth group in UPGMA) it is hard to tell if a student would finally choose the STEM or NON-STEM major only based on their feature values.

The limitation of our project: First of all, we have imbalanced classes in our data set, where total number of isSTEM is 151 and total number of nonSTEM is 341. The number of our samples is still small. Secondly, as the original log records of students obtained the sequence interaction between students and ASSISTments, there should be much more insight can be discovered.

In the future study, sophisticated resampling methods can be executed to improve prediction accuracy, such as Modified synthetic minority oversampling technique(MSMOTE) and algorithmic Ensemble Techniques, such as Bootstrap Aggregating. We also could extract large amount of samples from web based version of ASSISTments with the help of big data techniques to improve our models. Further more, with the sequence interaction records which have specific structural architecture, we could combine time series and deep learning algorithm to discover the variation of knowledge level of students in the process of interaction with ASSISTments by using convolution neutral network or recurrent neural network and other more interesting implication which can be inferred from the records

V. Acknowledgments

The authors express their thanks to Ningyu Zhang for the help of choosing the class projects topic.

References

[1] Wang, X. 2013. Why Students Choose STEM Majors Motivation, High School Learning, and Postsecondary Context of Support. *American Educational Research Journal*, 50(5), 1081-1121.

[2] San Pedro, M., Ocumpaugh, J., Baker, R., and Heffernan, N. 2014. Predicting STEM and non-STEM college major enrollment from middle school interaction with mathematics educational software. *Educational Data Mining 2014*.

[3] San Pedro, M., Baker, R., Bowers, A., and Heffernan, N. (2013). Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In *n Proceedings of the 6th international conference on educational data mining*.

[4] Cocea, M., Hershkovitz, A., and Baker, R.S.J.d. 2009. The Impact of Off-task and Gaming Behaviors on Learning: Immediate or Aggregate? *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 507-514.

[5] Fredricks, J. A., Blumenfeld, P. C., and Paris, A. H. 2004. School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74, 59109.

[6] Lau, S. and Liem, A.D. 2007. The role of self-efficacy, task value, and achievement goals in predicting learning strategies, task disengagement, peer relationship, and achievement outcome. *Contemporary Educational Psychology*, 3, 1-26.

[7] Razzaq, L. M., Feng, M., Nuzzo-Jones, G., Heffernan, N. T., Koedinger, K. R., Junker, B., ... and Rasmussen, K. P. (2005, May). Blending Assessment and Instructional Assisting. In *AIED*, pp. 555-562.

[8] Segedy, J. R., Kinnebrew, J. S., and Biswas, G. (2015). Using coherence analysis to characterize self-regulated learning behaviours in open-ended learning environments. J. In *Journal of Learning Analytics*, 2, 1348.

[9] Segedy, J. R., Kinnebrew, J. S., and Biswas, G. (2015). Using coherence analysis to characterize self-regulated learning behaviours in open-ended learning environments. J. In *Journal of Learning Analytics*, 2, 1348.

[10] Trevor H., Robert T, Jerome F., *The Elements of Statistical Learning: Data Mining, Inference, and Predication*, Springer, New York, 2009.

[11] Christopher M. B., *Pattern Recognition and Machine learning*, Springer, New York, 2006

[12] Christopher D. M., Hinrich S., *Foundations of Statistical Natural Language Processing*, The MIT Press Cambridge, Massachusetts London, England, 1999

APPENDIX

	Column name	example	description
1	ITEST_id	8	a deidentified ID/tag used for identifying an individual student
2	SY ASSISTments Usage	2004-2005	the academic years the student used ASSISTments
3	AveKnow	0.35241648	average student knowledge level (according to Bayesian Knowledge Tracing algorithm -- cf. Corbett & Anderson, 1995)
4	AveCarelessness	0.18327565	average student carelessness (according to San Pedro, Baker, & Rodrigo, 2011 model)
5	AveCorrect	0.48390152	average student correctness
6	NumActions	1056	total number of student actions in system
7	AveResBored	0.20838904	average student affect: boredom (see Pardos, Baker, San Pedro, Gowda, & Gowda, 2014)
8	AveResEngcon	0.67912589	average student affect:engaged concentration (see Pardos, Baker, San Pedro, Gowda, & Gowda, 2014)
9	AveResConf	0.11590539	average student affect:confusion (see Pardos, Baker, San Pedro, Gowda, & Gowda, 2014)
10	AveResFrust	0.11240808	average student affect:frustration (see Pardos, Baker, San Pedro, Gowda, & Gowda, 2014)
11	AveResOfftask	0.15650305	average student affect: off task (see Pardos, Baker, San Pedro, Gowda, & Gowda, 2014 and also Baker, 2007)
12	AveResGaming	0.196561	average student affect:gaming the system (see Pardos, Baker, San Pedro, Gowda, & Gowda, 2014 and also Baker Corbett Koedinger & Wagner, 2004)
13	actionId	9950	the unique id of this specific action
14	skill	properties-of-geometric-figures	a tag used for identifying the cognitive skill related to the problem (see Razzaq, Heffernan, Feng, & Pardos, 2007)
15	problemId	104051118	a unique ID used for identifying a single problem
16	assignmentId	20405010	a unique ID used for identifying an assignment
17	assistmentId	104051118	a unique ID used for identifying an assistment (a instance of a multi-part problem)
18	startTime	1096470301	when did the student start the problem (UNIX time, seconds)
19	endTime	1096470350	when did the student end the problem (UNIX time, seconds)

20	timeTaken	49	Time spent on the current step
21	correct	0	Answer is correct
22	original	1	Problem is original not a scaffolding problem
23	hint	1	Action is a hint response
24	hintCount	1	Total number of hints requested so far
25	hintTotal	1	total number of hints requested for the problem
26	scaffold	0	Problem is a scaffolding problem
27	bottomHint	0	Bottom-out hint is used
28	attemptCount	1	Total problems attempted in the tutor so far.
29	problemType	textfieldquestion	the type of the problem
30	frIsHelpRequest	1	First response is a help request
31	frPast5HelpRequest	0	Number of last 5 First responses that included a help request
32	frPast8HelpRequest	0	Number of last 8 First responses that included a help request
33	stlHintUsed	0	Second to last hint is used – indicates a hint that gives considerable detail but is not quite bottom-out
34	past8BottomOut	0	Number of last 8 problems that used the bottom-out hint.
35	totalFrPercentPastWrong	0	Percent of all past problems that were wrong on this KC.
36	totalFrPastWrongCount	0	Total first responses wrong attempts in the tutor so far.
37	frPast5WrongCount	0	Number of last 5 First responses that were wrong
38	frPast8WrongCount	0	Number of last 8 First responses that were wrong
39	totalFrTimeOnSkill	0	Total first response time spent on this KC across all problems
40	timeSinceSkill	0	Time since the current KC was last seen.
41	frWorkingInSchool	1	First response Working during school hours (between 7:00 am and 3:00 pm)
42	totalFrAttempted	0	Total first responses attempted in the tutor so far.
43	totalFrSkillOpportunities	0	Total first response practice opportunities on this KC so far.
44	responseIsFillIn	0	Response is filled in (No list of answers available)
45	responseIsChosen	0	Response is chosen from a list of answers (Multiple choice, etc).

46	endsWithScaffolding	0	Problem ends with scaffolding
47	endsWithAutoScaffolding	0	Problem ends with automatic scaffolding
48	frTimeTakenOnScaffolding	0	First response time taken on scaffolding problems
49	frTotalSkillOpportunitiesScaffolding	0	Total first response practice opportunities on this skill so far
50	totalFrSkillOpportunitiesByScaffolding	0	Total first response scaffolding opportunities for this KC so far
51	frIsHelpRequestScaffolding	0	First response is a help request Scaffolding
52	timeGreater5Secprev2wrong	0	Long pauses after 2 Consecutive wrong answers
53	sumRight	0	
54	helpAccessUnder2Sec	0	Time spent on help was under 2 seconds
55	timeGreater10SecAndNextActionRight	0	Long pause after correct answer
56	consecutiveErrorsInRow	0	Total number of 2 wrong answers in a row across all the problems
57	sumTime3SDWhen3RowRight	0	
58	sumTimePerSkill	49	
59	totalTimeByPercentCorrectForskill	0	Total time spent on this KC across all problems divided by percent correct for the same KC
60	prev5count	0	
61	timeOver80	0	
62	manywrong	0	
63	confidence(BORED)	0.59786477	the confidence of the student affect prediction: bored
64	confidence(CONCENTRATING)	0.23429359	the confidence of the student affect prediction: concecntrating
65	confidence(CONFUSED)	0	the confidence of the student affect prediction: confused
66	confidence(FRUSTRATED)	0	the confidence of the student affect prediction: frustrated
67	confidence(OFF TASK)	0.83870968	the confidence of the student affect prediction: off task
68	confidence(GAMING)	0.00852182	the confidence of the student affect prediction: gaming
69	RES_BORED	0.37642746	rescaled of the confidence of the student affect prediction: boredom
70	RES_CONCENTRATING	0.32031737	rescaled of the confidence of the student affect prediction: concentration
71	RES_CONFUSED	0	rescaled of the confidence of the student affect prediction: confusion

72	RES_FRUSTRATED	0	rescaled of the confidence of the student affect prediction: frustration
73	RES_OFFTASK	0.78558547	rescaled of the confidence of the student affect prediction: off task
74	RES_GAMING	0.0002642	rescaled of the confidence of the student affect prediction: gaming
75	Ln-1	0.13	baysian knowledge tracing's knowledge estimate at the previous time step
76	Ln	0.06119041	baysian knowledge tracing's knowledge estimate at the time step
	schoolID	1	the id (anonymized) of the school the student was in during the year the data was collected
	MCAS		Massachusetts Comprehensive Assessment System test score. In short, this number is the student's state test score (outside ASSISTments) during that year. -999 represents the data is missing