Predicting wheel spinning in students using ASSISTment dataset.

Ayush Kumar
NC State University
akumar23@ncsu.edu

Dharmang Bhavsar
NC State University
dmbhavsa@ncsu.edu

Dr Collin Lynch
NC State University
cflynch@ncsu.edu

ABSTRACT

This paper depicts a novel method to try to predict wheel spinning in students and tries to find the right value of the number of attempts on a question to predict wheel spinning[1] (a term coined by J. Beck and Y. Gong) effectively. The predictor takes into account certain values such as time taken, hints asked for, incorrect answers among others to model student learning. Then using some assumptions on these modelling values the predictor learns and tries to predict the whether a student is wheel spinning or not. Models are developed for every attempt between 5 to 10 per student per skill. The paper also uses different values for mastery learning cutoff for student models. Thus a lot of parameter tweaking is involved. At last, the best model is depicted which shows that using the value of BKT threshold as (0.99), but after tweaking some internal parameters of the BKT the results are comparable to lower threshold values as well. The experiment suggests that it is better to try to predict wheel spinning early on, but it has its own drawbacks which are mentioned at the end of the paper.

Keywords

Student modeling; mastery learning; wheel-spinning; behavioral detector; prediction;

1. INTRODUCTION

Cognitive tutors are used to make students achieve mastery in a certain knowledge domain[6]. The underlying assumptions that a cognitive tutor takes for mastery learning is that, the more questions you give to a student about a KC, the more likely the student is to learn about it. This learning of students is modelled using student knowledge modelling techniques such as Bayesian Knowledge Tracing[3] among others.

The underlying assumption of such cognitive tutors is a matter of concern because all students are not able to pick up a skill just by solving problems about it. There might be some external factors affecting the student who is not able to master a skill. But, as the tutor keeps on giving problems to students, it might be counterproductive as a student becomes frustrated with not being able to solve such problems. So, it is important for cognitive tutors to predict or detect wheel spinning in students.

There have been previous attempts at detecting and predicting wheel spinning in students. This paper takes a different approach to model and predict wheel spinning.

The higher level research question is that if such systems are at all capable of predicting wheel spinning in students? The main research questions is if by using some assumptions can we model a system that tries predicts wheel spinning in students without any kind of interference by humans? And if wheel spinning can be detected, is logistic regression the correct algorithm to do so?

2. DATASET

The ASSISTments data from the year 2009-10 was chosen for this study. The initial dataset contained 401756 rows with 30 columns which included information about the student, question type, answer type, hints, opportunities amongst other points, details of which can be found on the dataset website[5]. Each row consists of a student attempt for a particular problem. Here, a problem is associated with a skill, which was the base of our study. It was seen that there were multiple rows where the skill ID was NaN. These rows were removed, leaving 338001 rows. Other NaN values in the opportunities original column were set to zero.

The next step was to remove the columns that would have no bearing on the results. These were the various ID columns including assignment_id, school_id, teacher_id etc. It should be noted that student and skill_id were kept as they were the basis used in our classification. Some features like answer type were carefully observed and their correlation values with other columns were calculated and finally we decided to keep answer_type in the data. There were two columns, opportunity and opportunity_original which were highly correlated with a value greater than 0.99. Keeping this in mind, the opportunity_original column was removed from the modelling.

There were also a few scaffolding problems comprising of less than 8% of the dataset. These rows were also not considered in our model building process.

The data was sorted in chronological order for each student-skill, hereafter referred to as the key. There were 123 skills present and 4163 students after the processing steps mentioned in the above paragraphs. These together formed 41981 key values.

It was observed that for some keys the number of attempts was quite less. As Bayesian Knowledge Tracing was to be used to generate our ground truth for values up to the 10th attempt. It was decided that only keys having more than or equal to 10 attempts will be used. This left 9125 unique key values, together having 182690 rows.

3. METHODOLOGY

Detecting whether a student is wheel spinning involves finding out when a student stops learning a skill even though he/she is still attempting problems. Thus, establishing the ground truth for this becomes all the more difficult. Earlier studies have considered Mastery learning[2] as the benchmark to decide whether a student is wheel spinning or not. This study uses Bayesian Knowledge Tracing to find out whether a student has learnt a skill or not.

3.1 Bayesian Knowledge Tracing

In Bayesian knowledge tracing we attempt to monitor the student's changing knowledge state during practice. Each time the student has an opportunity to answer a question in the model, the tutor updates its estimate of whether the student knows the skill, based on the student's action.[3]. The model requires initial parameters for guess, slip, transition as well as the initial knowledge level. It iterates through each opportunity a student has for a skill and updates his/her knowledge state based on their actions(whether they were correct or not). The following values were assumed for the purpose of this study.

L_0	0.5
Guess	0.3
Slip	0.1
Transition	0.2

Table 1. Parameters for BKT.

Once the Bayesian Knowledge level was was calculated, the ground truth was established by first deciding the threshold n value which would be used to classify whether a student has learnt the skill or not. Thus, it would be nth attempt for that particular key. This paper varies n from 5 to 10. Based on the knowledge level of the key, the separate column was created, setting 0 if the knowledge level was above a particular threshold and 1 otherwise. The following values of the threshold were used, 0.75, 0.8, 0.85, 0.9, 0.95 and 0.99. It was observed that 0.99 gave the best results. The corresponding graphs for it are located in the results section.

3.2 Logistic Regression

For each value of n, a Logistic Regression model was built for each cumulative attempt. That is, Model 1 considered data only of the first attempt of that particular key. Model 2 considered data for the first 2 attempts and so on. Because the nth attempt was set as the ground truth, Models upto n-1 were built. 12 columns were used to fit the Logistic Regression Model on the data. StratifiedKFold with 5 folds was used to calculate the performance of the models.

3.3 Support Vector Machine

On observation of the results of the Logistic Regression Model, Support Vector Machine with C=1 and a linear kernel was used with circumstances similar to the Logistic Regression Model.

4. RESULTS

Earlier the ground truth was set as 0 when the student had learnt the skill and 1 otherwise. The goal of this study is to detect wheel-spinning in students as early as possible. The below graphs are a measure of precision and recall based on the negative values.

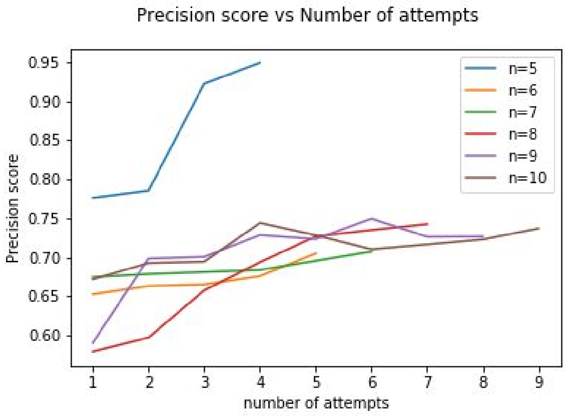


Figure 1: Precision when BKT threshold = 0.99 (Logistic Regression Model)

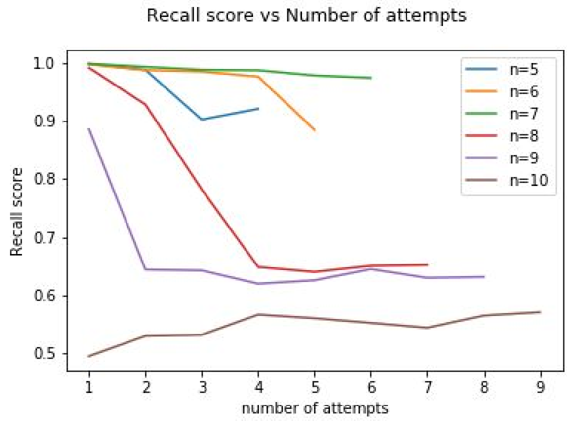


Figure 2: Recall for BKT threshold = 0.99 (Logistic Regression Model)

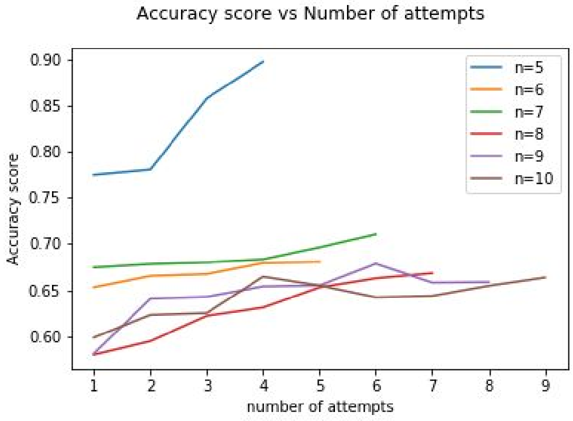


Figure 3: Accuracy for BKT threshold = 0.99 (Logistic Regression Model)

Each line indicates the score for different models keeping the same n value. Thus, it can be seen how predictions vary across different n values.

It is seen that precision has a high value for n=5, indicating that there are not many false positives for the model when considering the fifth attempt. For other values of n, Model 1 gave a precision of around 0.65 which gradually increased to around 0.70 as models with more attempts were tested.

Recall had a perfect score of 1 for n=5,6,7. This is expected, as reaching a BKT threshold of 0.99 is highly unlikely with the parameters which were used for BKT (see Table 1). The recall score decreases for Model 1 as we increase the value of n. This makes sense, as it would get difficult to predict wheel spinning at the 10th attempt looking just at the first attempt parameters.

There is a reduction in the recall score from Model 2 to 4 across different n values, before it tends to stabilize.

It was noted that keeping a lower BKT threshold drastically reduced the performance of the LR models, with recall values lower than 0.1.[Appendix 1]. This was due to a large number of false positives generated by the model.

It was seen that tweaking the L_0 value to 0.3 with a BKT threshold of 0.99 gave slighter better recall values based on different values of n.[Appendix 2]. It was observed that the values for n, from 5 to 8 were more consistent.

Similar to Logistic Regression graphs, different lines in the graph for SVM models correspond to different values of n.

It is observed that the curves of the SVM model are not as smooth as the Logistic Regression Model, although the values are approximately similar in both the cases. The reason for the fluctuations in the graphs of SVM maybe because SVM is not able to model the forgetting of students because the data takes forgetting into account and hence such fluctuations come into play.

Below are the graphs generated using the SVM models -

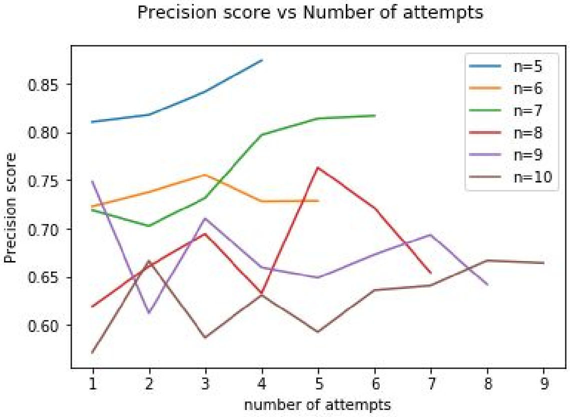


Figure 4: Precision when BKT threshold = 0.99

(Support Vector Machine Model)

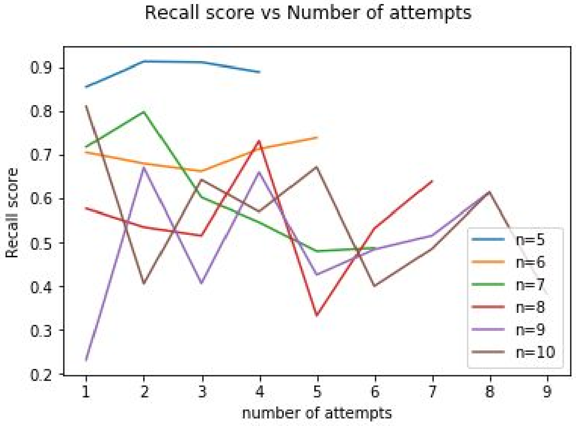


Figure 4: Recall when BKT threshold = 0.99

(Support Vector Machine Model)

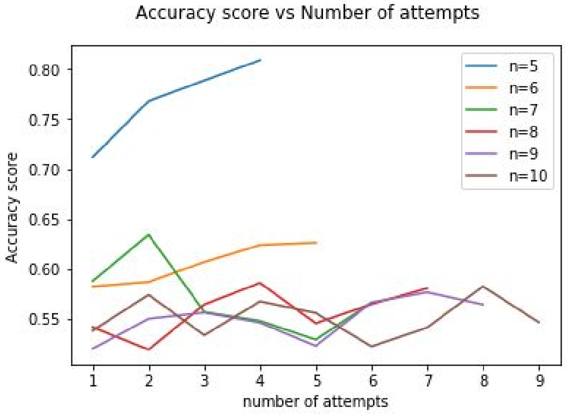


Figure 6: Accuracy when BKT threshold = 0.99

(Support Vector machine Model)

The general conclusion for the precision and recall values to be falling as n increases can be attributed to the timeline of prediction. For n=5 we are trying to predict the near future compared to n=10.

Currently, no conclusion was found for the fluctuations in the SVM model prediction metrics. It might be the case that SVM was not able to model the forgetting of students properly because the BKT values tend to fluctuate for such students.

Further tweaking the BKT Lo values to as low as 0.2 improved the measures or recall and accuracy. The reason behind this could be that more students would end up wheel spinning due to the low initial knowledge state. The increase would have led to a better model and reduce the scope of the false positives.

5. CONCLUSION

Previous work[4] indicated that Artificial Neural Networks gave a high recall and low precision. This study using Logistic Regression indicates otherwise. However, it must be noted that there were differences in the approach of how the ground truth was calculated.

Another important aspect to be considered for the BKT modelling is the difficulty level of the questions. The level will have an impact on the initial parameter values, which in turn affects the ground truth.

Overall, it was observed that predicting wheel spinning for a smaller n value was better. This indicates that it would be easier to detect whether a student would learn the skill or not using an earlier threshold. It must be noted that tweaking the initial parameters of BKT can lead to widely different results. Thus, it is important to set them carefully based on the various domain factors which would be relevant. Also, it was observed that it was better to predict wheel spinning early because from the BKT values it was seen that most students were wheel spinning. So the model gave high metric values for that, which does not mean that those students were necessarily wheel spinning. But what it does mean is that the model is able to fit to imbalanced dataset in a good way. Though improvements can be made, but this is what was the general conclusion from the result analysis.

6. ACKNOWLEDGMENTS

Our thanks to Dr Collin Lynch for his invaluable guidance in the research behind this paper. Also, the ASSISTments team, whose dataset was used in this study.

7. REFERENCES

[1] J. E. Beck and Y. Gong, "Wheel-Spinning:Students Who Fail to Master a Skill," in Proceedings of 16th International Conference, AIED 2013, Memphis, TN, USA, 2013.

[2] Y. Gong and J. Beck, "Towards Detecting Wheel-Spinning: Future Failure in Mastery Learning," in Learning at Scale 2015, 2015.

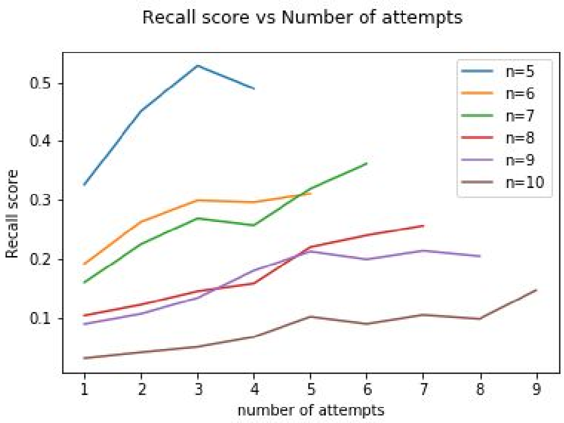
[3] A. T. Corbett and J. R. Anderson., "Knowledge tracing: Modeling the acquisition of procedural knowledge". User Modeling and User-Adapted Interaction, 1995.

[4] N. Matsuda, S. Chandrasekaran and J. Stamper, “How Quickly can wheel spinning be detected?”. Educational Data Mining 2016, 2016.

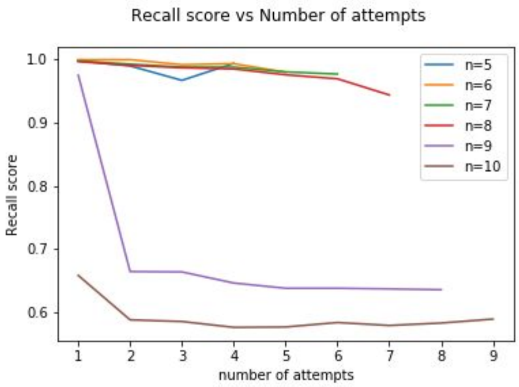
[5] ASSISTment 2009-10 Skill Builder Dataset. (https://drive.google.com/file/d/0B3f_gAH-MpBmUmNJQ3RycGpJM0k/view?usp=sharing)

[6] A. C. Sales and J. F. Pane, “The Role of Mastery Learning in Intelligent Tutoring Systems: Principal Stratification on a Latent Variable”. Annals of Applied Statistics, 2017.

8. APPENDIX



Appendix 1 Precision when BKT threshold = 0.85 (Logistic Regression Model, L_0 = 0.5)



Appendix 2 Precision when BKT threshold = 0.99 (Logistic Regression Model, L_0 = 0.3)