* Reviewer1:

[C1] It is unclear to me what “practices opportunities”, “skill”, “attempt” exactly means when stating the problem. The authors are suggested to give more concrete examples to clarify these concepts. Moreover, it is also important to make clear if there are any temporal

dependencies among the sequence of opportunities.

* The explanation of “practices opportunities”, “skill”, “attempt” will be put on page x.
* Are there any temporal dependencies among the sequence of opportunities?

[C2]The authors point out that their detector can have **negative examples** of the M/W features. That is, if a student already achieves M before the 10th opportunity (which is already observed), the detector can still predict W on him/her on the 10th opportunity. How could that happen? Isn’t there a naïve way to overcome this deficiency? How about training the detector based on only the remaining five features and then combining the output with the “M/W on each opportunity”, as the latter is actually a definite indicator of M.

* The gradient boosted decision tree model needs more explanations. Maybe a more formal problem formulation and some figures can help understand the adopted method.
* The contribution of this paper is somewhat limited as it is mainly on feature engineering. Both logistic regression and the gradient boosted decision tree model are standard machine learning methods. Do the authors improve these methods in novel ways that are specific to the considered problem?
* Reviewer2:
* The data includes records that might be dependent (for example when they are related to the same student in different times). This implies that iid modelling (Logistic regression, gradient boosted trees) proposed in this paper is not adequate for the data. Instead, longitudinal models have to be applied
* The cross validation procedure proposed to estimate the performance of the models cannot be used due to the mentioned dependency present in the data. Since it was used in the experiments, the performance estimation is optimistic: for example when records for a student can be simultaneously in the training folds and test fold.
* the problem considered is imbalanced. This implies the need for balancing techniques such as cost sensitive learning, ROC analysis, etc.
* the problem considered is imbalanced. This implies the need for balancing techniques such as cost sensitive learning, ROC analysis, etc.
* Reviewer3:
* In terms of related work, the focus is too narrow. The authors cite their previous work (Matsuda et al.,2016) and also the work of another research group - Beck and Gong (2013), Gong and Beck (2015) and Gong et al. (2016). The research should be situated in the broader area of predicting student performance.
* The paper is not self-content. It assumes that readers are familiar with previous work on wheel-spinning. It was not clear to me what “skill”, “attempt” and “practice opportunities” are. These concepts need to be clearly defined and illustrated with examples; they are very important for understanding the paper.
* I am also not sure what the definition of “cognitive tutor” is and what the relation between a cognitive tutor and the proposed method is. The title says “wheel-spinning detection in student performance on cognitive tutors”. Is the method applicable only to a cognitive tutors? I think that wheel-spinning is a problem for ITS in general.
* The task is classification – predict M or W. What is the distribution of the two classes? The majority class will determine the baseline for the classification performance; it is very important to compare with a baseline.
* The skills are not independent – the more complex ones depends on the simpler ones; if a student hasn’t mastered the simpler ones, he/she will fail the more complex ones. This skill (and time) dependency is not taken into account when constructing the training data. How would it affect the training data and the task? Is it possible that the later data is mainly from class W (as the students haven’t mastered the previous required skills) and hence the task in the later time period becomes very imbalanced?
* The authors need to properly compare with previous work, not simply mention previous results as 70-79% precision and 25-50% recall. This means that the results should be presented for each opportunity (from opp3 to opp9).
* Why is the logistic regression model called a “generic” model and the boosted trees called an “upgrated” model? Both are equally generic and a boosted tree classifier is not an upgrated versions of logistic regression.
* There are other many problems with the use of terminology and description of the methods, e.g. just to give 2 examples:--The summary of boosted trees on p. 9 is very poor - not clear, not coherent; it looks like sentences taken from various sources. Which are the “weak algorithms”? --On the same page – “We train gradient boosted decision trees with a ten-fold cross validation by each opportunity…”. Cross validation is not a training algorithm but an evaluation procedure.
* Not likely unless the evaluation is extended to other datasets and the issues are addressed.
* Define “skill”, “attempt”,“practice opportunities” and give an example to illustrate them. Define “cognitive tutors”.
* Table 2 requires more explanation – I did not find it useful in its current form.
* Abstract:

-low recall is not "lack of detecting power"

-generic and upgraded - incorrect terminology, see my comments above

-more simplified and fast - a simpler (in what sense - smaller number of features), fast – I actually didn't see results supporting this claim