Response to reviewers: *"Kaggle-in-class Data Challenges Can Boost Student Learning"*

We would like to thank the reviewers and the Associate Editor for their thoughtful review of the manuscript. The reviews have been helpful for improving the manuscript.

Please find below the *referees' comments (italics)* and our response (blue).

Reviewer: 1

I enjoyed this paper and think it makes a contribution to the literature. However, I would like to see some improvements made.

1. *Briefly discussing the experience from the professor side of things would be helpful. So in addition to the student survey it would be helpful to read how hard it was to set up and work with the Kaggle competitions and any logistical challenges.*

   Additional information has been added to a "Teachers corner", Section 5. In addition a video explaining step by step how to create the competition in the Kaggle platform, including data files preparation, was created. The link to the video is also add to the "Teachers corner" section.

2. *Also more discussion of study limitations is needed. For example, in the ETC class the "treatment" is confounded with level of student (graduate/ undergraduate). Other limitations should be briefly discussed as well. I also would like to see suggestions for future research.*

   The "Participants" (2.3) and "Performance"' (3.1, 4.1) sections now contain more detailed explanations on why this is reasonable. Future work has been expanded.

3. *Sentence misworded in line 29 page 9*

   Fixed

4. *Also in Figure 3, it would seem there is some non-linearity in a couple of the graphs. I would like to see that mentioned and discussed. It seems there is a point of diminishing return, in graphs E and F.*

   Added.

5. *I also think it is a study weakness that all the survey questions are worded in a positive manner and this could be leading. I realize given the study is done this can't be changed but it would be worth noting in the discussion ways the survey might be improved.*

   It might be more productive in future work to directly ask for level of engagement first, then one could also ask students who did not do the competition as well the question.

   "How would you rate your level of engagement in this course?"

   Not at all engaged--- up to extremely engaged with choices in between.

   Fair comment, we can't change the question wording at this stage, but this limitation is mentioned as part of the survey results analysis in "Interest" section (4.3).

Reviewer: 2

Major Comments

1. *The article does not provide a motivation for incorporating a predictive competition in a coursework. What is the educational aim of the competitive element as compared to more "standard" projects where students do not compete? Relatedly, what are the advantages and limitations of the proposed competitive format? A literature review, which currently is completely missing from the article, would be more than appropriate, summarizing thoroughly the existing work on project-based learning and putting the proposed format of Kaggle competitions vis-à-vis the project-based pedagogy.*

   Revised the introduction to address this. Expanded the literature review.

2. *The article does not provide details regarding the implementation/administration of the competition in a classroom setting, and lacks practical guidelines for teachers of statistics and data science. Thus, in the current form it has little to offer if the reader considers replicating Kaggle competitions in her/his classroom. Some of the important details that are left out of the article include but are not limited to the time and involvement required from the instructor from start to finish; the format of instructor's involvement with students/teams; time in the semester (start? middle? end?) when the projects get assigned; presence of a teaching assistant (TA), etc.*

   Same comment as reviewer 1. Material has been added accordingly.

3. *Many of the datasets on Kaggle are in need of pre-processing/cleaning prior to the application of predictive models. In authors' view, should these tasks be carried out by instructors/TAs or left to students instead? What are the pros and cons of each of these approaches?*

   The new section "Teacher's Corner" addresses this.

4. *For many Kaggle datasets there are existing (at least preliminary) solutions on the web. Reliance on those resources may drastically deteriorate student involvement, interest, and takeaways from the projects. In authors' view, what steps can be taken by instructors in that regard?*

   These data sets are provided for self-learning. We suggest using data that lecturer have collected so that the students do not have access to the full data. This point added to "Teachers corner", Section 5.

5. *In the first paragraph of section 2.3 authors mention that students were allowed to first submit individually and then form groups. It would be helpful if authors could explain the motivation behind that format.*

   Motivation has been added to the "Participants" section (2.3).

6. *In the second paragraph of section 3.1, as well as in section 4.1 authors note/imply that the postgrad cohort in ETC2420 took part in the Kaggle challenge while the undergraduate cohort was used as the control group. This could introduce confounding, as the average quantitative preparedness in the postgraduate cohort could be significantly higher than that of the undergraduate group prior to the Kaggle challenge. As such, the choice of the undergraduate cohort could be inappropriate as a control group.*

This has been explained in the revised experimental setup where participants are described.

7. *To understand the assessment of student performance more thoroughly, it would be helpful if authors could provide the complete final exam at least for one of the two courses (perhaps in supplementary materials), clearly marking the questions pertaining to the Kaggle challenges.*

   Both exams have been added in supplementary materials.

8. *In section 3.1 authors introduce the metric that they use in the assessment of student performance. As authors discuss in the third paragraph of section 3.1, the constructed metric measures the consistency of students' performance on a specific set of questions in relation to the overall exam performance. Subsequently authors use the defined metric for the assessment of performance, and attribute change in medians to improved performance (e.g. second paragraph of section 4.2). In reviewer's view, change in medians shows change in consistency of performance, and that does not necessarily translate to improvement in learning. It is possible to construct a hypothetical example where the defined metric increases without any change in the performance on the regression- (or equivalently classification-) related part. It is not clear why authors used the "ratio" metric, as opposed to, for example, simply the percentage of the possible points received on a specific subset of questions (e.g.regression- or classification-related). As such, the use of the proposed measure does not seem to be justified as being appropriate for the assessment of performance. If the currently used metric is retained, then the authors should explain the construction of it more clearly, exemplify it using numbers, and justify why an increase in the metric signals improvement in learning (rather than improvement in consistency).*

   This has been addressed as part of "Performance" section (3.1).

9. *The entries in the "Median difference" column in Table 1 do not correspond to the differences in the median bars implied from Figure 1 (e.g. the difference between the median bars in the "regression" box from the left sub-figure and the "regression" box from the right sub-figure is not 0.1 as entered in Table 1).*

   They do match. It doesn't look like it from the scales on the boxplot, but double checking the calculations, they are correct. These are automatically calculated, and the code is in the Rmd document that generates the paper. The numbers in the table are manually entered, but they match what the code for the boxplot produces.

Minor Comments

10. *It was not clear whether the training and testing sets noted in section 2.2 were the same across the two institutions.*

    This has been explained as part of "competition data" section (2.2).

11. *It is unclear if the survey mentioned in section 4.3 was anonymous. If it was, please stress that, and if it was not then address the bias that would be introduced in survey responses due to the lack of anonymity.*

    This has been explained as part of "Interest" section (3.3).

12. *It is mentioned in section 2.3 that there were 63 students randomized to regression and classification competitions, while in section 4.3 the total number goes down to 61. Please clarify.*

    One student eventually didn't sit the exams and one student didn't participated in the competition. This has been explained as part of "Participants" section (2.3).

13. *Further details behind Kaggle including but not limited to examples of popular datasets/competitions that it has hosted, data contributors, the motivation for contributing data, and the consequences of winning a competition or scoring high – would be helpful in the introduction.*

    This has been explained as part of "Introduction" section.

Associate Editor

1. *There is no literature review. Even a short review of papers that have used projects or competitions would put this paper in context. Also some review (if it is out there) of your performance metric might help the readers to understand why you measured learning gains in this way.*

   Further literature review had been added to the "Introduction". We provided more explanation about the choice of performance measure in section "Performances" (3.1).

2. *Discussion of the logistics of setting up such a competition would be very helpful to the reader. The Kaggle link should be more prominently placed in the article as well as your personal insight/experience to what kind of investment a professor needs to make in order to make this happen.*

   This is a comment of both reviewers, and a new section has been added with this information.

3. *There needs to be a better explanation of why undergraduates were used as a control group at one of the universities. This jumped out to several reviewers.*

   This has been explained.

4. *I would argue for a data display/analysis that shows for each individual student the difference between their performance score on their "treatment" questions versus their "non-treatment" questions, i.e. the classification students that did "better than expected" on classification questions, how many of those also did "better than expected" on the regression questions and so on.*

   In the revised version Figure 3: Students performance in classification and regression questions by competition type, has been add and discussed as part of "Performance" section (3.1).