Kaggle-in-class Data Challenges Can Boost Student Learning

Abstract

Kaggle is a data modeling competition service, where participants compete to build a model with lower predictive error than other participants. Several years ago they released a simplified service that is ideal for instructors to run competitions in a classroom setting. This paper describes the results of an experiment to determine if participating in a predictive modeling competition enhances learning. The evidence suggests it does. In addition, students were surveyed to examine if the competition improved engagement and interest in the class.

Keywords: instructional technology, statistical modeling, data science, statistics education, data mining

1 Introduction

Kaggle (The Kaggle Team 2018) is a platform for predictive modelling and analytics competitions where participants compete to produce the best predictive model for a given data set. It is well-known for its competitions (e.g. Rhodes (2011)), some of which come with rich monetary prizes (e.g. Howard (2013)). There are also learning competitions (Agarwal 2018), designed to help novices hone their data mining skills. Winners are typically expected to share their code, and occasionally new algorithms have emerged, for example, deep neural networks (Hinton & Dahl 2012) and XGBoost (Chen & Guestrin 2016).

In 2015, Kaggle InClass was introduced, as a self-service platform to conduct competitions. These competitions can be private, limited to members of a university course, and are easy to setup. This is an opportunity for educators to provide a vehicle for students to objectively test their learning of predictive modelling. As a competition, with an independent clear performance metric, along with a dynamic leader board, students can see how their model predictions compare with the models produced by other students. Being able to make multiple submissions over a several week time frame, enables them to try out approaches to improve their models. This paper examines the educational benefits of conducting predictive modeling competitions in class on performance, engagement and interest.

In the past few years, the educational community started to collect positive evidence on including competitions in the classroom. None of these were data analysis competitions. Van Nuland et al. (2015) ran a competition assessing anatomical knowledge, as part of an undergraduate anatomy course. Calnon et al. (2012) discusses robotics competitions as part of computer science education. Carpio Cañada et al. (2015) discusses the participation of students in externally run artificial intelligence competitions. All of these studies found significant improvement in student exam marks accredited to participation in competition.

Classroom competition is an example of active learning, which has been shown to be pedagogically beneficial. Prince (2004) surveyed the literature and found that all forms of active learning have positive effect on the learning experience and student achievement. The magnitude of the effect of different approaches, though, varies. What's more, Freeman et al. (2014) examined 158 studies published in about 50 STEM educational journals. The

authors found that student exam scores increased by almost half a standard deviation through active learning. Moreover, students in classes with traditional lecturing were 1.5 times more likely to fail than their peers in classes with active learning.

A competition, like any other assessment method has its advantages and disadvantages. It brings the 'game' feeling, increases the interest level among students and motivates for higher performance (Shindler (2009), p 105). But, it may have negative influence if constructed poorly. Among the negative influences are increased stress and anxiety, induced by fearing a low ranking, failure or technology barriers. In addition, students may invest a disproportionate amount of time and effort into competition. Despite some received criticism, a properly set competition can benefit the students greatly. The competition should be relatively short in duration to avoid consuming undue energy. Students should be clear about the rules and the goal. They should be properly rewarded and most important, feel that they have reasonable chance to win or achieve high mark (Shindler 2009).

Shelley et al. (2009*b*) raised the need for more quantitative and statistical analysis of evidence in science education. This paper contributes to this call by offering statistical analysis of theffects on learning of classroom data competitions.

2 Experimental setup

2.1 Data collection

The experiment was conducted during Semester 2 2017. Data was collected during two classes, one at the University of AB (MAST90083), and one at CD University (ETC2420/5242).

2.2 Competition data

Two data sets were compiled for the kaggle challenges: Melbourne property auction prices and spam classification. The Melbourne auction price data was collected by extracting information from real estate auction reports (pdf) collected between Feb 2, 2013 and Dec 17, 2016. The spam classification data was compiled by graduate students at Iowa State University as part of a data mining class, in 2009. Data was compiled by monitoring and

extracting information from their emails by class members, over a period of a week, and manually tagging them as spam or ham.

Both data sets were split into training and test sets, for the kaggle challenge. Students had access to the true response variable only for the training data. For the Melbourne housing data, students were expected to predict price based on the property characteristics. For the spam data, students were expected to build a classifier to predict whether the email as spam or not.

Both data sets are challenging for prediction, with relatively high error rates.

The training and the testing data sets of the Melbourne auction price data were similar but not identical across the two institutions. Some of the variables in the data set were simulated, e.g. property land size and house size. The simulated data was generated slightly differently for different institutions. This was done deliberately to prevent students passing answers from one institution to another.

2.3 Participants

Computational Statistics and Data Mining (MAST90083) is designed for postgraduate level students with math, statistics, information technology or actuarial backgrounds. It covers modelling both continuous (regression) and categorical (classification) response variables. The 63 students were randomized into one of two kaggle competitions, one focused on regression (R) and the other classification (C). This setup mimics randomized control trials, which is the 'gold standard, in experiment design (Shelley et al. (2009*a*), ch. 1). Students built prediction models and made submissions individually for 16 days, and then were allowed to form groups to compete for another 7 days.

The reason for this strategy was first to motivate each of the students to think about modelling and be actively engaged in the competitions through individual submission. The lecturer allowed participants to create groups towards the end of the competition to illustrate the advantages of group work and mixing models. Another motivation for this stratagem was the university policy, requiring a strategy to assign students individually in group assignments.

One of the 63 students elected not to take part in the competition, and another student

didn't sit the exam, producing a final sample size of 61.

Statistical Thinking (ETC2420/5242), covers regression, but not classification, and has a mix of undergraduate and postgraduate students. Only the 34 postgraduate (5242) students were required to participate in the kaggle competition, and competed in the regression (R) challenge. This was run independently from the MAST90083 competition. The 145 undergraduate (2420) students were used as controls for examining performance of the postgraduate students. Although, it may be surprising at first glance, the undergraduate students provide a reasonable control for the graduate students. The class is taught to both cohorts simultaneously. The entry requirements to the Bachelors of Commerce at Monash is high, and these students have strong mathematics backgrounds. In the years prior to this experiment, the undergraduate scores on the final exam are indistinguishable from those of the graduate students. The competition ran for one month. Students formed their own teams of 2-4 members to compete.

3 Methodology

3.1 Performance

Better performance is equated to better understanding of the material, as measured in the final exam. MAST90083 and ETC2420/5242 included questions, with several parts, on the final exam related to kaggle challenges. These questions were identified prior to data analysis.

For all questions in the exam, difficulty and discrimination scores were computed, using the mean and standard deviations. Of the questions pre-identified as being relevant to the data challenges, only the parts that corresponded to high level of difficulty and high discrimination were included in the comparison of performance.

Scores for the relevant questions were summed, and converted into percentage of the possible score. The total exam score was converted to a percentage. One can expect, that on average, student's success rate for each question will be about the same as the success rate in the total exam. Understanding one topic better than another will result in higher success rate for questions asking about the better understood topic compared to the scores

for other topics. For example, we would expect from a student with a 70% exam mark to get 70% marks on each of the questions in the exam, if she has similar knowledge level on all the exam topics. If in some topic, say regression, the student has better knowledge, she will perform better on the regression questions. Her success rate on regression question will be higher than 70%. Consequently, her performance on some other questions should be below 70% which is associated with lesser understanding of these topics.

Therefore, performance for each student was computed as the ratio of these two numbers, percentage success in the regression (classification) questions and percentage success in the total exam. A value of 1 would indicate that the student's performance on that set of questions was consistent with their overall exam performance, greater than 1 that they performed better than expected, and lower than 1 meant less than expected on that topic.

Using only the percentage of successes for each set of questions, instead of the proposed ratio, will not differentiate between a better performance and just a better student. Especially in the case of ETC2420/5242 that have a mixed population of master and undergraduate student.

The distribution of the performance scores by group is shown as a boxplot. Focus is on the difference in median between the groups. Permutation tests were conducted to examine difference in median scores for students participating or not in a competition.

3.2 Engagement

The students were allowed to submit at most one prediction per day, while the competitions were open. The frequency of submissions, and the accuracy (or error) of their predictions, made by individual students, is recorded as a part of the kaggle system. To examine whether engagement improved performance, scores on the questions related to the competition normalised by total exam score (as computed in the performance section) is examined in relation to frequency of submissions during the competition. In addition, performance in the competition as measured by accuracy or error, is also examined in relation to the number of submissions. Scatterplots, correlation and linear models are used to examine the associations.

3.3 Interest

Students in MAST90083 and ETC5242 were invited to give feedback about the course, in particular about the data competitions, before the final exam. This information was voluntary, and students who completed the questionnaire were rewarded with a coupon for a free coffee. The data from this survey was viewed by the researchers after all course grades had been reported. To reduce potential bias in students replies, we emphasize this point as part of the instruction at the beginning of the survey.

4 Results

4.1 Performance

Figure 1 shows the data collected in MAST90083. Performance is plotted against type of question, separately for the competition they completed. The difference in median scores indicates performance improvement. Students who completed the classification competition (left) performed relatively better on the classification questions than the regression questions in the final exam. Conversely, students who participated in the regression competition performed relatively better on the regression questions.

Question Set	Median difference	Permutation p-value
Classification	0.250	0.033
Regression	0.104	0.000

Table 1: Comparison of median difference in performance by competition group, for MAST90083 students, using permutation tests. Both sets of medians are significantly different, indicating improved scores for questions on the topic related to the Kaggle competition.

Table 1 shows the results of permutation testing of median difference between the groups. Generally the results support the competition improved performance. Students who participated in the kaggle challenge for classification scored higher than those that did the regression competition, on the classification problem. Using a permutation test, this

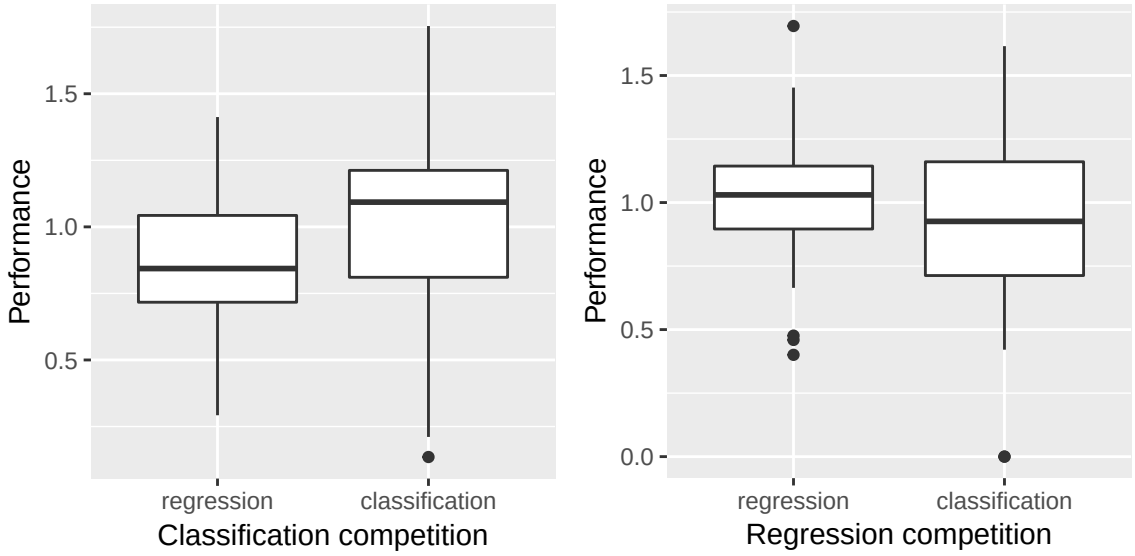

Figure 1: Boxplots of performance on regression and classification questions in the final exam, by type of data competition completed in MAST90083. Students generally performed better on the questions corresponding to the competition they participated in.

corresponds to a significant difference in medians. Similarly the results show that students who did the regression challenge, performed better on these exam questions.

Figure 2 shows the results for students ETC2420/5242. The boxplots suggest that the students who participated in the challenge performed relatively better than those that didn't on the regression question than expected given their total exam performance.

Only the post-graduate students participated in the regression competition, as their additional assessment requirement. Scores for the question on regression (Q7a,b,c) in the final exam were compared with the total exam score (RE). On these question parts, a, b, c, over all the students all three were in the top 10 of difficulty, with students scoring less than 70%, on average. Parts b, c were in the top 10 for discrimination, and part a was at rank 13.

Based on the median, the students who participated in the kaggle challenge scored 0.09 higher than those that didn't, a median of 1.01 in comparison to 0.92. Using a permutation test, this corresponds to a significant difference in medians, with p-value of 0.015.

Figure 3 presents students' scores for classification and regression questions. A score over 1 is considered outperformance (relative to the expectation). Quarters one and three

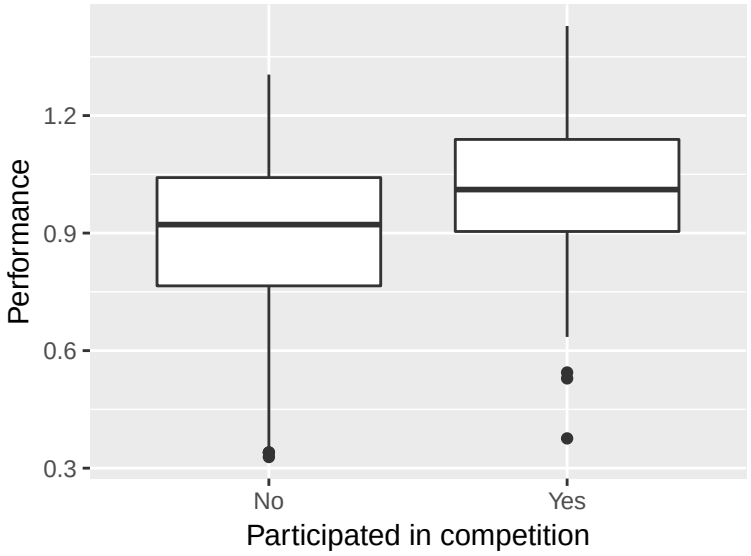

Figure 2: Performance for regression question relative to total exam score for students who did and didn't do the regression data competition in ETC2420/5242.

include students that underperform or outperform on both types of questions, respectively. In both cases the number of students that participated in the classification competition is very close to the number of students that participated in the regression competition (excluding a few regression students on the border of score 1). Students in quarters two and four outperform on one type of questions but not on the other type. We can see that more regression students outperform on regression questions than classification students (12 vs. 7). Similarly, classification students do better on classification questions (11 vs. 3). This is another evidence towards positive influence of the data competition on student's performances.

4.2 Engagement

The number of submissions that a student made, may be an indicator of performance on the exam questions related to the competition. A student who is more engaged in the competition may learn more about the material, and consequently perform better on the exam. Figure 4 (top row) shows performance on the classification and regression questions, respectively, against their frequency of prediction submissions for the three student groups

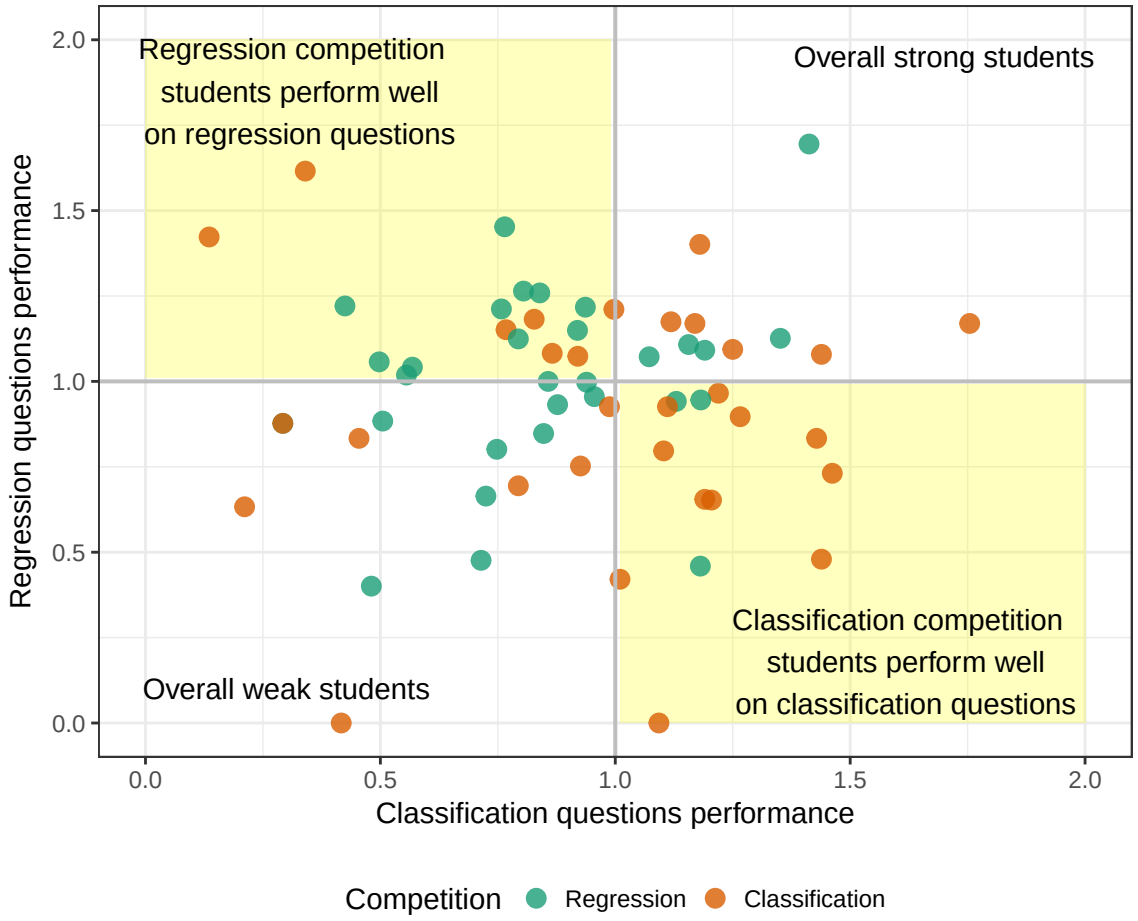

Figure 3: Students performance in classification and regression questions by competition type.

(MAST90083 classification and regression, ETC5242 regression) competitions. The relationship is weak in all groups, and this mirrors insignificant results from a linear model fit to both subsets. On the other hand, the predictive accuracy improved with the number of submissions for the regression competitions. There appears to be some nonlinearity present in these plots, suggesting reduced returns. That is reasonable to expect. Also, some students strategically make very poor initial predictions, to get a baseline on error equivalent to guessing.

The competition performance relative to number of submissions is shown in plots (d)-(f). Each point corresponds to one student, and accuracy or error of the best predictions submitted is used. The regression competition seemed to engage students more than the classification challenge. Students submitted more predictions, and their models improved with more submissions.

4.3 Interest

Figure 5 shows the survey responses related to the kaggle competition, for MAST90083 and ETC5242. The response rate for MAST90083 was 55%, with 34 of 61 students completing the survey. The response rate for ETC5242 was 50%, 17 students out of 34 completed the survey. Overwhelmingly, students reported that they found the competition interesting and helpful for their learning in the course.

After collecting the survey from the students we were told that the questions about students engagement were positively worded. This may have potentially led to some bias. An improved way could be to ask directly about student's engagement, e.g. "How would you rate your level of engagement in this course?" with set answer options of "not at all engaged" – up to "extremely engaged" with several choices in between. Another improvement could be asking ETC2420 students that didn't take part in the competition about their level of engagement and compare the answers with other students of ETC5242. We acknowledge that the differences in the engagement levels may not necessarily be a result of participation in the competition but it is still an interesting aspect.

The survey was not anonymous. However, the results became available to the lectures only after all the grades were realised to the students. This point was emphasized in the

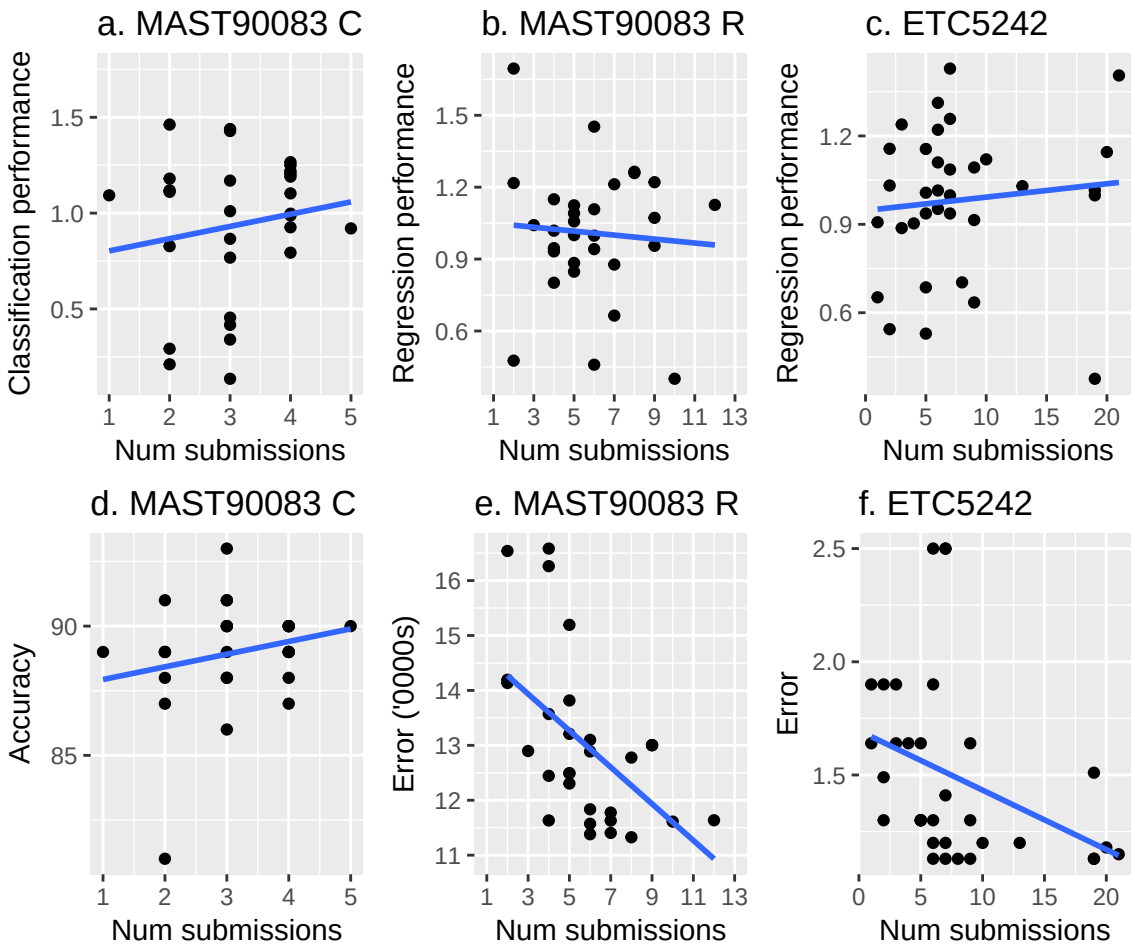

Figure 4: Scatterplots of the exam performance (a-c) and competition performance (d-f) by number of prediction submissions, for the three student groups. The relationships with exam performance are weak. For the MAST90083 and ETC5242 regression competitions, a clear pattern is that predictions improved substantially with more submissions. (House price in ETC5242 were divided by 100,000, explaining the difference in magnitude of error between two competitions.)

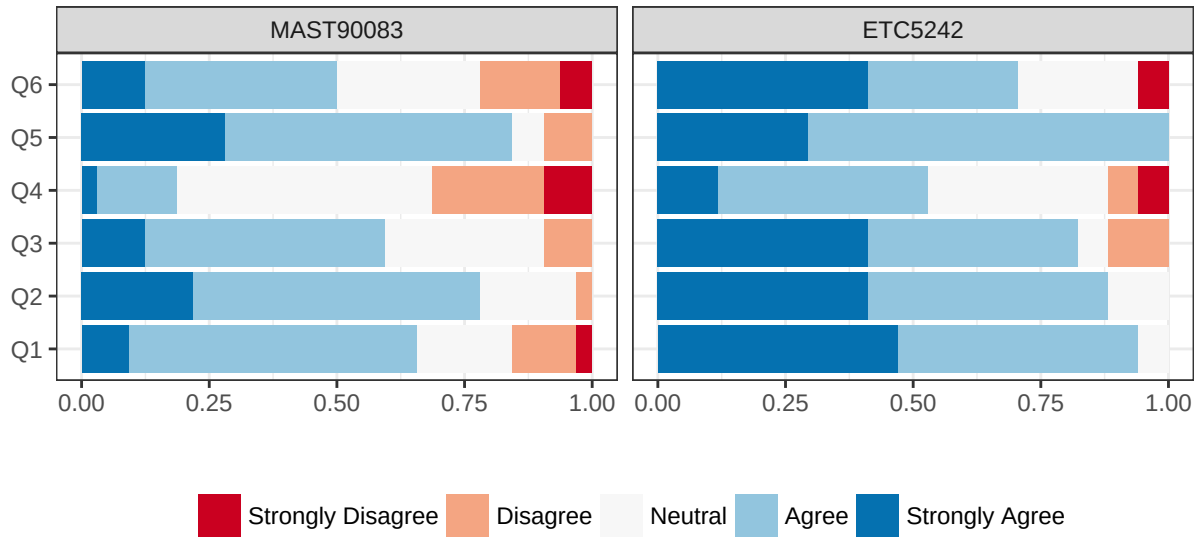

Figure 5: Summary of responses to survey of kaggle competition participants. Overwhelmingly the response to the competition was positive in both classes, especially the questions on enjoyment and engagement in the class, and obtaining practical experience. (Table 2 lists the questions.)

Label	Question
Q1	I found the data competition is great fun.
Q2	Taking part in the data competition contributed a lot to my engagement with the subject.
Q3	Taking part in the data competition improved my confidence in my understanding of the covered material.
Q4	Taking part in the data competition improved my confidence in my success in the final exam.
Q5	Taking part in the data competition improved my confidence in my ability to use the acquired knowledge in practical applications.
Q6	I feel that the required time investment in the data competition was worthy.

Table 2: Questions asked in the survey of competition participants.

instructions to the students at the beginning of the survey.

5 Teachers' corner

Creating a new competition is surprisingly easy. Information on setting up a Kaggle In-Class challenge is available on the service's web site (`https://www.kaggle.com/about/inclass/overview`). There is a setup wizard for step-by-step guidance on getting your competition underway. We have created a short video illustrating the steps to establish a new competition, available on the web (`https://www.youtube.com/XXXX`).

The kaggle service provides some data sets, primarily for student self-learning. These are not suitable for use in a class challenge, because all the data is available, and solutions are also provided. We recommend providing your own data for the class challenge. Finding a suitable data set for a competition can be a difficult task. The criteria for a good data set are:

1. the full set is not available to the students, to avoid cheating.
2. the data is not too easy, or too hard, to model so that there is some discriminatory power in the results.
3. data should be relatively clean, to the point where the instructor has tested that a model can be fitted.
4. contains some challenges, that make standard off-the-shelf modeling less successful, like different variable types that need processing or transforming, some outliers, large number of variables.
5. if it is a classification challenge, it will work better with relatively balanced classes, because the overall accuracy is the easiest metric to use.

Choosing the metric upon which to evaluate the model is another decision. Our advice is to keep it simple, so you, and the students, can understand the student scores. If you are running a regression challenge, then "Root Mean Squared Error (RMSE)". It is is a classification challenge, then "Categorization Accuracy", the percentage correct is reasonable.

The data needs to be split into training and testing sets. Kaggle will then split your test set into two, a public set that is used to provide ongoing scores to participants, and a private set, on which performance is revealed only after the competition closes. If you have categorical variables in the data set, you will want to make sure that all categories are present in both training and test sets. The training set will have both predictors and response, but the test set will have the response variable removed. Each observation needs to be assigned an id, because this will be needed to evaluate predictions. The solution file, containing the id and the true response, is provided to the system for evaluating submissions, and is kept private. A sample submission file needs to be provided. Participants will submit their solutions in the same format.

It is a good idea to build a basic model yourself on the training data, and predict the test data. The performance of this model can be provided to the participants as baseline to beat.

The competition needs to run without any intervention from the instructor. The instructor can monitor students progress: the number of submissions, student scores and even the uploaded data at any time. When the competition ends the "Leaderboard" page provides a list of students ordered by the final score. It also provides all the scores from all past submissions (under "Raw Data" on "Public Leaderboard").

It may be recommended to limit students to one submission per day. It encourages students to think about more efficient improvement of their model before the next submission. It also prevents the student spending too much time building and submitting models. Some students will become so engaged in the competition that they might neglect their other coursework. About halfway through the coompetition, students might be allowed to form teams, to learn how averaging models can boost performance.

In awarding course points to student effort, we typically align it to performance. Participant ranks based on their performance on the private part of the test data are recorded. Performance scores that are pretty close to each other should be given the same rank, reflecting that there may not be a significant difference between them. The best gets perhaps, 5 points, then a half a point drop until about 2.5 points, so that the worst performing students still get 50% for the task. Kaggle does not allow you to download participants

email addresses. All you see is their kaggle name. Record the student names in kaggle to match with your class records.

Along with the competition, students were expected to submit a report that explained their modeling strategy, and what they had learned about the data beyond the modeling. The overall score for this part of the course was a combination of the mark for their report and their performance in the challenge. In both courses this accounted for 10% of the final mark.

From an instructor perspective, its very rewarding watching the students participate in the competition. It provides a truly objective way to assess their ability to model in practice. Students are often motivated to consult with the instructor about why their model is underperforming, or what other approaches might produce better results.

6 Discussion

This paper has described the setup and results of an experiment to examine the effectiveness of data competitions on student learning, using Kaggle InClass as the vehicle for conducting the competition. The experiment was conducted in the classroom settings as part of the normal teaching of the courses, which imposes limitations on the design. However, with both MAST90083 and ETC2420/5242, either a randomized assignment of students to two topic groups, or a control group, was possible, which enabled comparing performance.

The primary finding, is that participating in a data challenge competition, produces a statistically significant improvement in the learning of the topic, although the effect size is small. Secondarily, anecdotally, the competitions enhanced interest and engagement in the course.

7 Future work

This work is one of a few quantitative analyses of data competition influences on students' performance. More evidence needs to be collected from other STEM courses to explore consistent positive influence. Moreover, future investigation is required to understand the influence of the different aspects of data competition implementation on the magnitude

of the performance improvement. For example, the competition duration, availability and accessibility of additional material, requirement of writing a final report or giving a short oral presentation are elements worth investigating. Prior and post testing of students might provide a better experimental design.

8 Acknowledgments

This project (title: Effect of Data Competition on Learning Experience) has been approved by the Faculty of Science Human Ethics Advisory Group University of AB (ID: 1749858.1 on September 4, 2017) and by CD University Human Research Ethics Committee (ID: 9985 on August 24, 2017).

This document was produced in R (R Core Team 2017) with the package knitr (Xie 2015). Data cleaning was conducted using tidyr (Wickham & Henry 2018), dplyr (Wickham et al. 2017) and plots were made with ggplot2 (Wickham 2016). The materials to reproduce the work are available at `https://github.com/XXX`.

References

Agarwal, N. (2018), 'Use kaggle to start (and guide) your ml/ data science journey - why and how', `https://towardsdatascience.com/use-kaggle-to-start-and-guide-your-ml-data-science-journey-f09154baba35`.

Calnon, M., Gifford, C. M. & Agah, A. (2012), 'Robotics competitions in the classroom: Enriching graduate-level education in computer science and engineering', *Global Journal of Engineering Education* **14**(1), 6–13.

Carpio Cañada, J., Mateo Sanguino, T., Merelo Guervós, J. & Rivas Santos, V. (2015), 'Open classroom: enhancing student achievement on artificial intelligence through an international online competition', *Journal of Computer Assisted Learning* **31**(1), 14–31. **URL:** *https://onlinelibrary.wiley.com/doi/abs/10.1111/jcal.12075*

Chen, T. & Guestrin, C. (2016), Xgboost: A scalable tree boosting system. ArXiv e-prints. **URL:** *https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf*

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H. & Wenderoth, M. P. (2014), 'Active learning increases student performance in science, engineering, and mathematics', *Proceedings of the National Academy of Sciences* **111**(23), 8410–8415.

Hinton, G. & Dahl, G. (2012), 'Deep Learning How I Did It: Merck 1st place interview', *No Free Hunch by Kaggle Team* .
URL: *http://blog.kaggle.com/2012/11/01/deep-learning-how-i-did-it-merck-1st-place-interview/*

Howard, J. (2013), 'POWERDOT awarded $500,000 and Announcing Heritage Health Prize 2.0', http://blog.kaggle.com/2013/06/03/powerdot-awarded-500000-and-announcing-heritage-health-prize-2-0/.

Prince, M. (2004), 'Does active learning work? a review of the research', *Journal of engineering education* **93**(3), 223–231.

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Rhodes, J. (2011), 'Competition shines light on dark matter', `https://obamawhitehouse.archives.gov/blog/2011/06/27/competition-shines-light-dark-matter`.

Shelley, M. C., Yore, L. D. & Hand, B. (2009*a*), Education research meets the gold standard: Evaluation, research methods, and statistics after no child left behind, *in* 'Quality Research in Literacy and Science Education', Springer, pp. 3–15.

Shelley, M. C., Yore, L. D. & Hand, B., eds (2009*b*), *Quality research in literacy and science education: international perspectives and gold standards*, Springer Science & Business Media.

Shindler, J. (2009), *Transformative classroom management: Positive strategies to engage all students and promote a psychology of success*, John Wiley & Sons.

The Kaggle Team (2018), 'The home of data science & machine learning', `https://www.kaggle.com`.

Van Nuland, S. E., Roach, V. A., Wilson, T. D. & Belliveau, D. J. (2015), 'Head to head: The role of academic competition in undergraduate anatomical education', *Anatomical sciences education* **8**(5), 404–412.

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
URL: *http://ggplot2.org*

Wickham, H., Francois, R., Henry, L. & Muller, K. (2017), *dplyr: A Grammar of Data Manipulation*. R package version 0.7.4.
URL: *https://CRAN.R-project.org/package=dplyr*

Wickham, H. & Henry, L. (2018), *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. R package version 0.8.0.
URL: *https://CRAN.R-project.org/package=tidyr*

Xie, Y. (2015), *Dynamic Documents with R and knitr (2nd edition)*, Chapman and Hall/CRC.