Kaggle-in-class Data Challenges Can Boost Student Learning

Julia Polak
Department of Statistics, University of Melbourne
and
Dianne Cook
Department of Econometrics and Business Statistics, Monash University

May 27, 2018

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 reduced service that enables instructors to run competitions in a classroom setting. This paper describes the results of an experiment to determine if the 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 well-known for the data competitions, some richly funded. It provides a platform for predictive modelling and analytics competitions where participants compete to produce the best predictive model for a given data set. 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 paper examines the educational benefits of conducting predictive modeling competitions in class on performance, engagement and interest.

2 Experimental setup

2.1 Data collection

The experiment was conducted during Semester 2 2017. Data was collected during three classes, one at the University of Melbourne (MAST90083), and two at Monash University (ETC2420/5242 and ETC3250).

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 compiled 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.

2.3 Participants

MAST90083 is titled Computational Statistics and Data Mining, is designed for postgraduate level, for 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). Students individually built prediction models and made submissions for 16 days, and then were allowed to form groups to compete for another 7 days.

ETC2420/5242, titled Statistical Thinking, covers regression, and has a mix of undergraduate and postgraduate students. Only the 34 postgraduate (5242) students were required to participate in the kaggle competition focused on regression (R). The 145 undergraduate (2420) students are considered control for examining performance. The competition ran for one month. Students formed their own teams of 2-4 members to compete. Several undergraduates also chose to compete individually. The material on regression methods, particularly the computational methods needed to successfully predict housing prices was new to both groups of students.

ETC3250, called Business Analytics, is an undergraduate course focusing on data mining. All students participated in a kaggle competition on the classification challenge. Because this group had no comparison group, it was difficult to assess performance.

2.4 Platform

MAST90083 used `https://inclass.kaggle.com/c/XXX`. ETC2420/5242 used `https://inclass.kaggle.com/c/vitticeps`.

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. Performance for each student was computed as the ratio of these two numbers. 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.

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 theire 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.

4 Results

4.1 Performance

Figure 1 shows the data collected in MAST90083. Normalized scores for the classification and regression questions are plotted as boxplots against type of competition participation. The normalized scores were computed based on overall test score (CE, RE), in plots A, B, as well as by overall question score (CQ, RQ), in plots C, D. The difference in median scores indicates performance improvement. In plot A, the normalized scores for the classification questions were better for students who participated in the classification competition. From plot B, a very small increase in median score on the regression questions can be seen for the students who participated in the regression competition. Plots C, D show the scores normalized by total question score. This increases the difference for the regression performance.

Question Set	Median difference	Permutation p-value
Classification	0.250	0.012
Regression	0.104	0.00

Table 1: Comparison of median difference in performance by competition group.

Table 4.1 shows the results of permutation testing of median difference between the groups. Generally the results support the competition improved performance. Students

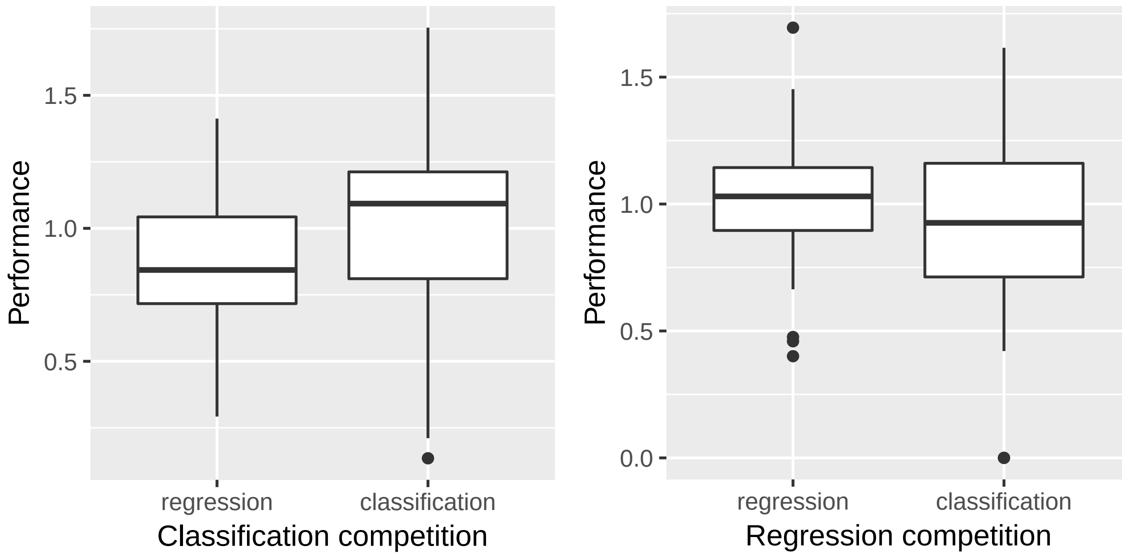

Figure 1: Performance on regression and classification questions relative to total exam score (A, B) and overall question (C, D) for students by type of data competition in MAST90083. Differences in medians indicate improved performance.

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 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 0.02 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 0.031.

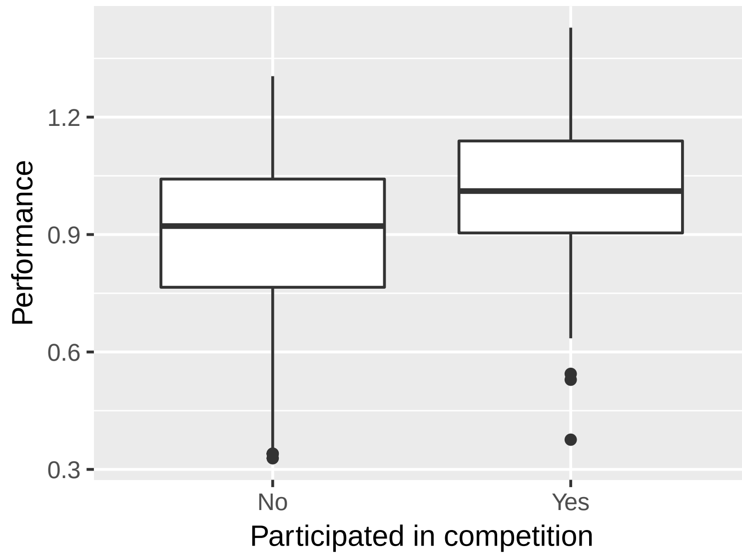

Figure 2: Performance for regression question relative to total exam score for students who did and didn't do the regression data competition in ETC2420/5242.

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 3 (top row) shows performance on the classification and regression questions, respectively, against their frequency of prediction submissions for the three student groups (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.

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 predictiona, and their models improved with more submissions.

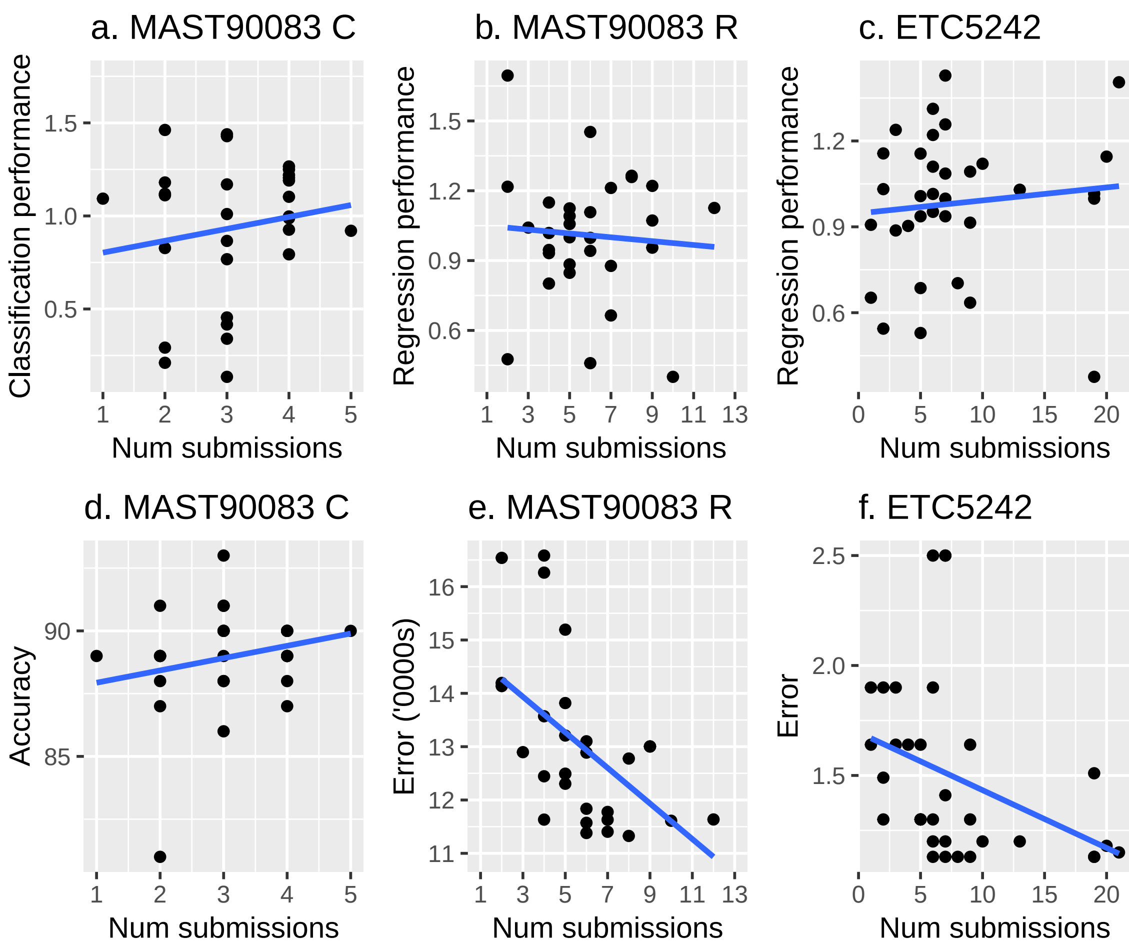

Figure 3: Scatterplots of the exam performance (a-c) and competition performance (d-f) by number of prediction submissions, for the three student groups. The relationships with exam performance are weak. For the MAST90083 and ETC5242 regression competitions, a clear pattern is that predictions improved substantially with more submissions.

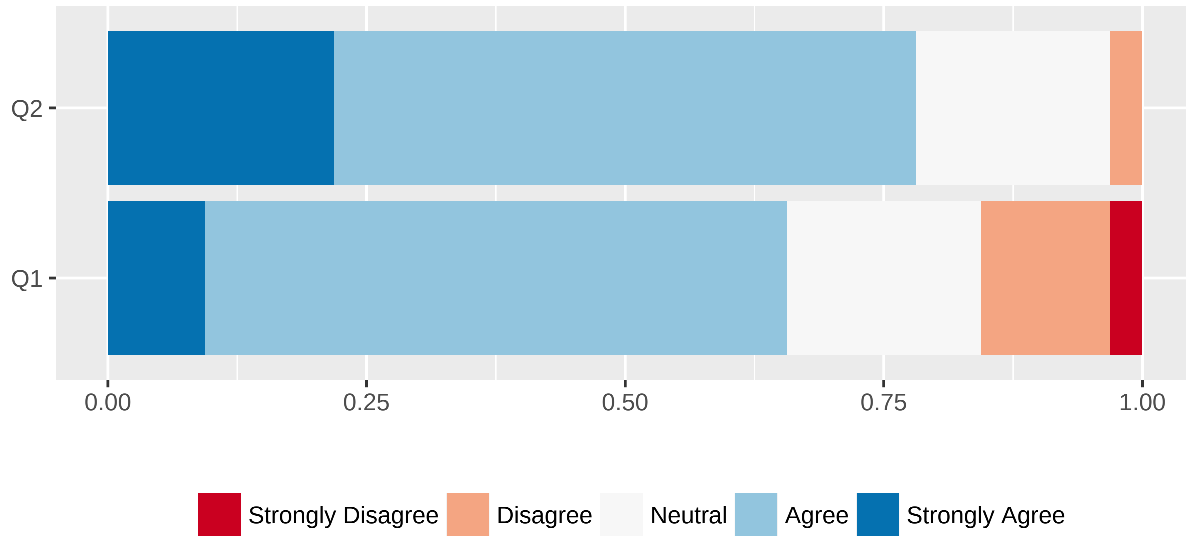

Figure 4: Stacked bar charts of responses to survey.

4.3 Interest

4.3.1 MAST90083

Response rate was 55%, with 34 of 61 students completed the survey.

4.3.2 ETC5242

The response rate was 50%, 17 students out of 34 completed the survey.

4.3.3 ETC3250

5 Discussion

This paper has discussed results from an experiment to examine the effectiveness of data competitions on student learning, using Kagggle InClass as the vehicle for conducting the competition. The evidence suggests that participating in competitions enhances learning.

6 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 Melbourne (ID: 1749858.1 on September 4, 2017) and by Monash 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).

7 Supplementary material

The following material is provided in addition to the main paper. - Code to reproduce the results is in the Rmd document - De-identified data - Additional details of analysis - Copies of relevant exam questions

References

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/*

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

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.