Do data competitions improve learning?

A study of student performance, engagement, and experience with Kaggle InClass data challenges

Julia Polak

Dianne Cook

CAUSE/JSDSE webinar series

Welcome from our host and moderator

Leigh Johnson (Capital University)

What's new in the journal?

Latest articles

Article

How to Get Away with Statistics: Gamification of Multivariate Statistics ›

Jacopo Di Iorio et al.

Published online: 26 Oct 2021

Article

Student Developed Shiny Applications for Teaching Statistics ›

Sabrina Luxin Wang et al.

Published online: 19 Oct 2021

Article

Diagnosing Data Analytic Problems in the Classroom ›

Roger D. Peng et al.

Published online: 11 Oct 2021

Article

Using Team-Based Learning to Teach Data Science ›

Eric A. Vance

Published online: 1 Oct 2021

No webinar in December 2021.
We will return in the new year!

Consortium for the Advancement of Undergraduate Statistics Education

https://www.causeweb.org/cause

Undergraduate Statistics Project Competition(Class Project and Research Project)

More info at https://www.causeweb.org/usproc

Undergraduate Statistics Research Conference (eUSR), Nov. 5, 2021
https://www.causeweb.org/usproc/eusrc/2021

Julia Polak

Julia Polak is a lecturer in Statistics at the University of Melbourne. She has a broad range of research interests including nonparametric methods, forecasting and data visualisation. In addition, Julia has many years of experience in teaching statistics and data science for different audience.

Dianne Cook

Di Cook is a Professor in Econometrics and Business Statistics at Monash University in Melbourne. Her research is in the area of data visualisation, especially the visualisation of high-dimensional data using tours with low-dimensional projections, and projection pursuit. A current focus is on bridging the gap between exploratory graphics and statistical inference.

Outline

- Why data competitions?
- What is Kaggle-in-Class?
- Our experiment
- Findings
- How to do this yourself?

Why data competitions?

- In the past few years, the educational community in the classroom started to collect positive evidence
- Large part of the evidence is of anecdotal nature
- Moreover, none of it was data analysis competitions
- In our work, we are offering a statistical analysis of the classroom data competitions' effect on learning

What is Kaggle-in-Class?

- Platform for running predictive analytics competitions
- The competition can be limited to pre-defined members
- Simple interface to set up, and for students to upload predictions
- Flexible to different response variables, and has choice of loss function, number of submissions
- Active and immediate public leaderboard, with hidden private board for final performance to prevent gaming predictions

Our experiment: design

UniMelb:

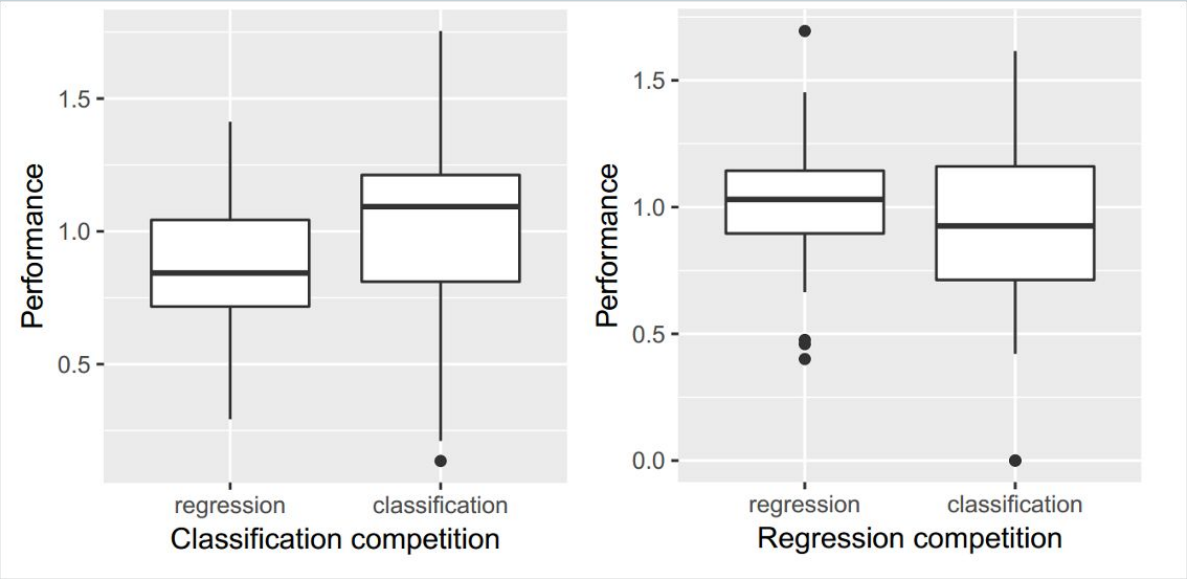
- Computational statistics & data mining subject
- Postgraduate level, for students with math, statistics, information technology or actuarial backgrounds
- Covers regression & classification modelling

- *n = 61*
- Randomly allocated to regression (R) or classification (C) competitions
- Mimic randomized control trials
- Duration 16 + 7 days
- Individual & teaming

Our experiment: competition data

- Melbourne property auction prices
  - Collected by extracting information from real estate auction reports (pdfs)
  - Students were expected to predict price based on the property characteristics
- Spam classification
  - Students' emails were monitoring over a period of a week, and manually tagging them as spam or ham (compiled by graduate students at Iowa State University)
  - Students were expected to build a classifier to predict whether the email is spam or not
- Both data sets are challenging for prediction, with relatively high error rates

Our experiment: performance



Taking part in the data competition improved the performance during the exam in the relevant topic.

In each competition, the median performance score is higher for the questions about the competition topic than the score for the other topic and higher then 1.

Difference in median scores were statistically significant (permutation tests).

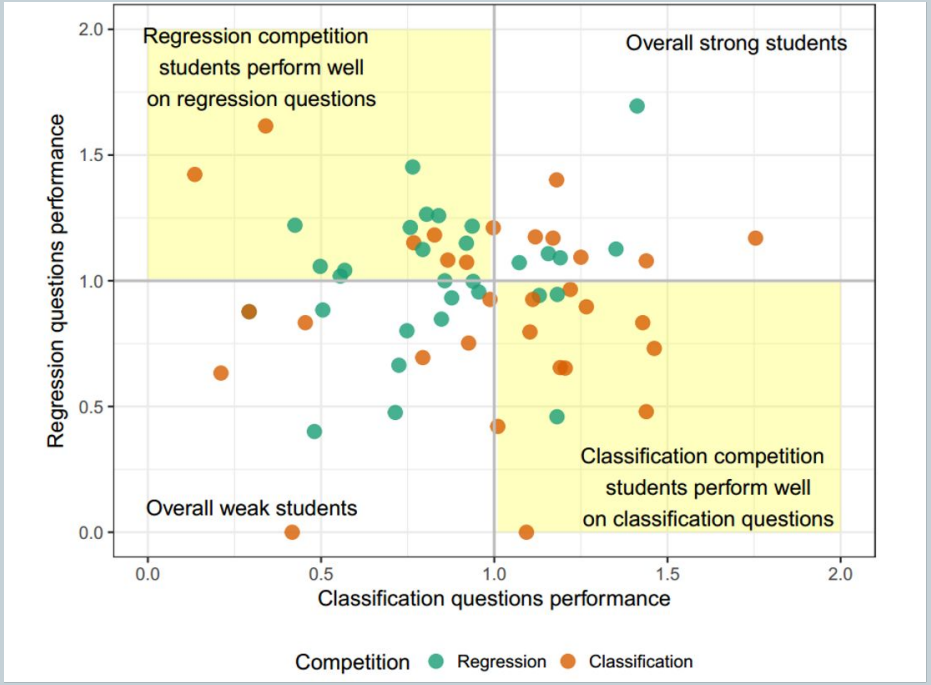
How is performance calculated?

$$\text{Performance} = \frac{\%\ \text{success in regression (or classification) questions}}{\%\ \text{success in the final exam}}$$

- 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

Value > 1	*Value = 1*	*Value < 1*
Meant that they performed better than expected.	*Indicate consistent performances.*	Meant less than expected on that topic.

Our experiment: performance



More regression students outperform on regression questions than classification students (12 vs. 7).

Similarly, classification students do better on classification questions (11 vs.3).

Additional evidence towards positive influence of the data competition on student's performances.

Our experiment: engagement

A student who is more engaged in the competition may learn more about the material, and consequently perform better on the exam.

- Engagement was measured by frequency of submissions
- We used scatterplots, correlation and linear models

Positive correlation between frequency of submissions and performance in the competition.

The relationship between frequency of submissions and performance in the exam is weak in all groups.

The regression competition seemed to engage students more than the classification challenge.

Our experiment: Monash University

- Covered regression only
- A mix of undergraduate (UG) & postgraduate (PG) students
- 34 PG competed
- UG (141) used as controls (UG entry requirements are really high). Performance on Qs related to regression, relative to Qs on other topics.

Results at the Monash class support findings from Unimelb

Taking part in the data competition improved the exam performance in the relevant topic.

Positive correlation between frequency of submissions and performance in the competition.

Weak relationship between frequency of submissions and performance in the exam.

Our experiment: interest

Is great fun

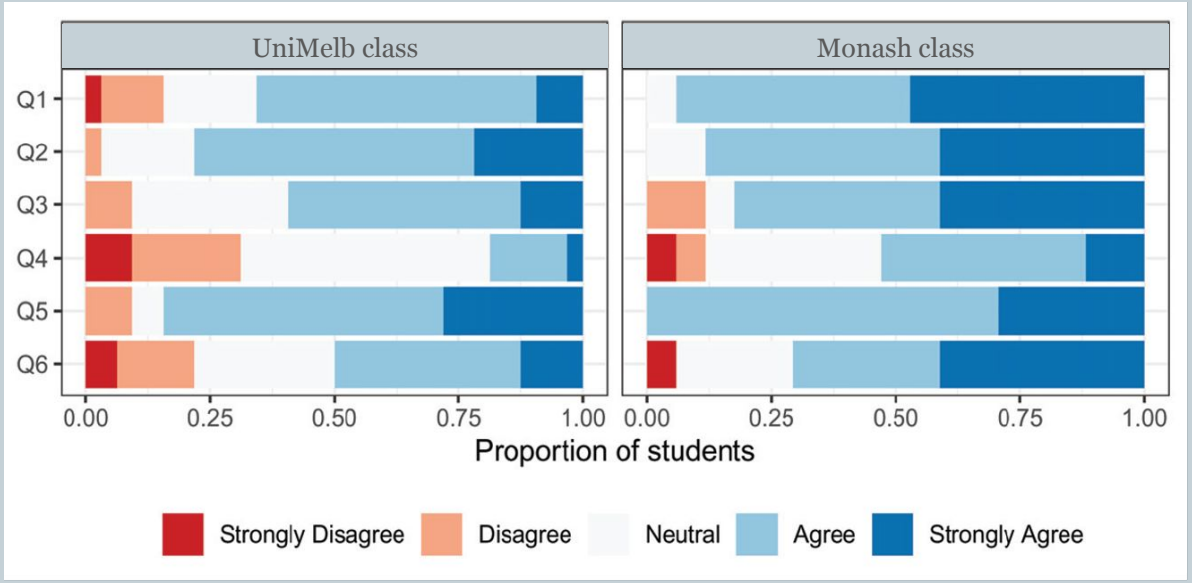
↑ my engagement

↑ understanding of the material

↑ confidence in exam success

↑ confidence in implementation

Time investment was worthy



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.

Our experiment: students testimonies

"I found the data competition quite useful for learning the material."

"I think the competition was a great success and lots of fun!"

"Thanks for a good and inspiring experience!"

"Data Competition was the best part of the course!"

"Would really enjoy this type of learning in other units."

"The Kaggle competition made the assignment particularly engaging."

"...I found the scoring on Kaggle to be a great motivator to continue working on the project (and try to beat the other teams)..."

How to do this yourself?

- Video step-by-step: https://www.youtube.com/watch?v=tqbps4vq2Mc&t=32s.
- Information on setting up: https://www.kaggle.com/about/inclass/overview.
- Our recommendations:
  - Data
  - Randomly training/test
  - Loss function & submissions
  - Teams vs. individual
  - Provide a baseline model
  - Extra resources
  - Additional report / video to explain participant's model and what they have learned additionally about the data
  - Final grades & timing

Examples of data from recent competitions

- 2021 Victorian bushfire causes
- 2020 Pictionary
- 2019 How Points End in Tennis
- 2017 Melbourne auction prices (Statistical thinking class)
- 2017 Spam or Ham (Machine learning class)
- 2016 Ames housing prices (Statistical thinking class)
- 2016 Who's speaking (Machine learning class)
- 2015 Happy paintings (Machine learning class)

See `https://resources.numbat.space/setting-up-a-kaggle-competition.html`

Thank you!

Questions ?

Outline