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

March 24, 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 Fall semester 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. Students were expected to predict price based on the property characteristics. The spam classificaation data was compiled by graduate students at Iowa State University as part of a statistical computing class by Dr Heike Hofmann, in XXX. Data was compiled by monitoring and extracting information from emails over a period of a week, and manually classifying them as spam or ham. Students were expected to classify the email as spam or not.

Both data sets provide substantial challenge for prediction.

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 students were required to participate in the kaggle competition focused on regression (R). The 145 undergraduate 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.

ETC3250, called Business Analytics, is an undergraduate course focusing on data mining. All students participated in a kaggle competition on a classification problem. Because this group had no comparison group, it was difficult to assess performance. This data was primarily used to examine engagement and interest based on a follow-up questionnaire.

2.4 Platform

MAST90083 used `https://inclass.kaggle.com/c`. ETC2420/5242 used `https://inclass.kaggle.com/c/vitticeps`. ETC3250 used .

3 Methodology

4 Results

4.1 Performance

4.1.1 MAST90083

We have examined two normalizations. Once, we normalized the score for the question (or group of questions) by the maximum possible score for the question (or the group of questions), denoted as PTQ. We also normalized by the total exam score, denoted as PTE. The PTQ and the PTE scores were calculated for each student for the following four clusters of questions

- Questions related to classification methods
- Questions related to regression methods
- Questions related to the classification and regression methods
- Questions related to other topics that covered during the subject but unrelated to the data competition

In addition, the PTQ and the PTE scores were calculated to each of the questions in the first and the second clasters.

Figure 1 shows the boxplots for the PTQ scores and the PTE scores for each group of students. The left boxplot is related to the students that took part in the data competition related to the regression methods, the Melbourne Price competition. The right boxplot is related to the students that took part in the data competition related to the calssification methods, the Spam classification competition.

The plots in the left column summarize the PTQ scores and the plots in the right column summarize the PTE scores. The four questions clusters are corresponding to the four rows in the figure.

Examining the figure, we can see clearly the positive correlation between student's scores and the type of the data competition. Namely, the median score (PTQ and PTE)

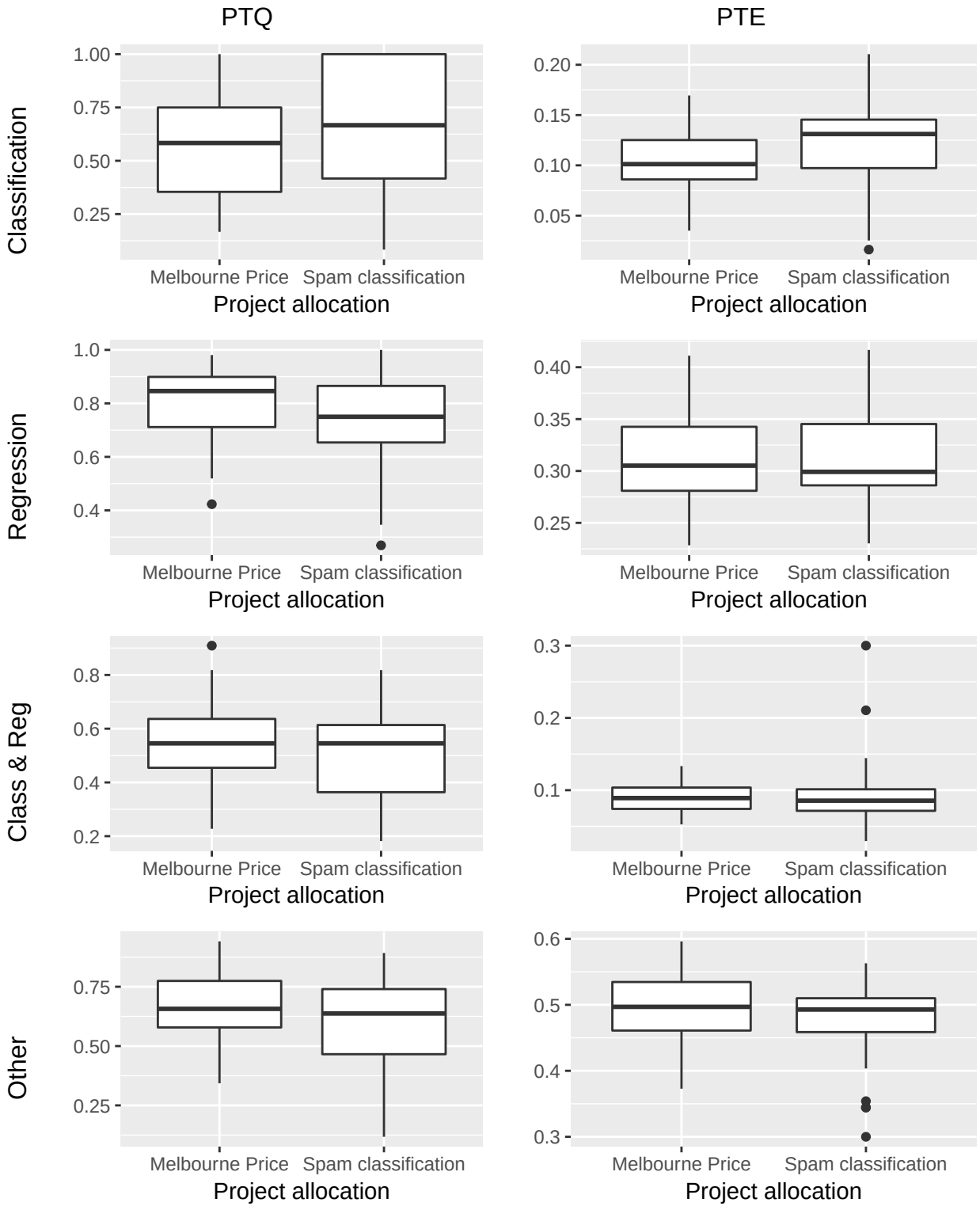

Figure 1: PTQ and PTE scores distributions over different question groups

of the students that took part in the regression related competition is higher for regression questions. Similarly, for the students that took part in the classification related competitions. There is no significant different in the medians of the two groups of students for the question related to the general question covered both the classification and the regression methods. As well as for the all other question related to other topics covered during the subject.

The statistical significance of this results was examined via the "permutation tests" . . . [TO COMPLETE].

```
## [1] 0.015
```

```
## [1] 0.262
```

Next we examined the statistical significance of the mean differences between the performances of the two group of students.

Examining individual questions

In the final exam two questions were related to the classification methods (Q1 and Q10) and five questions were related to regression methods (Q5, Q6, Q8, Q15 and Q16).

Figure 2 shows the scores distribution for the classification methods questions: Q1 and Q10.

Q1 was a multi-chose question, worth only 2 points (out of 100). Both plots, of the PTQ and the PTE scores show no difference between the two groups of students.

Q10, was a relatedly large question, worth 10 points. Examining the medians of the PTQ scores indicate only mild advantage for the students from the classification competition. However, looking on the PTE scores suggests that the classification questions was much easier for the students from the classification competition comparing to all the other questions in the exam.

Next we examined the statistical significance of the mean differences between the performances of the two group of students for Q10.

Figure 3 shows the scores distribution for the regression methods questions: Q5, Q6, Q8, Q15 and Q16. There was no difference in performances of the two groups in the Q5, a small 2 point, question. Interestingly, for question Q6, 6 point question that asked to

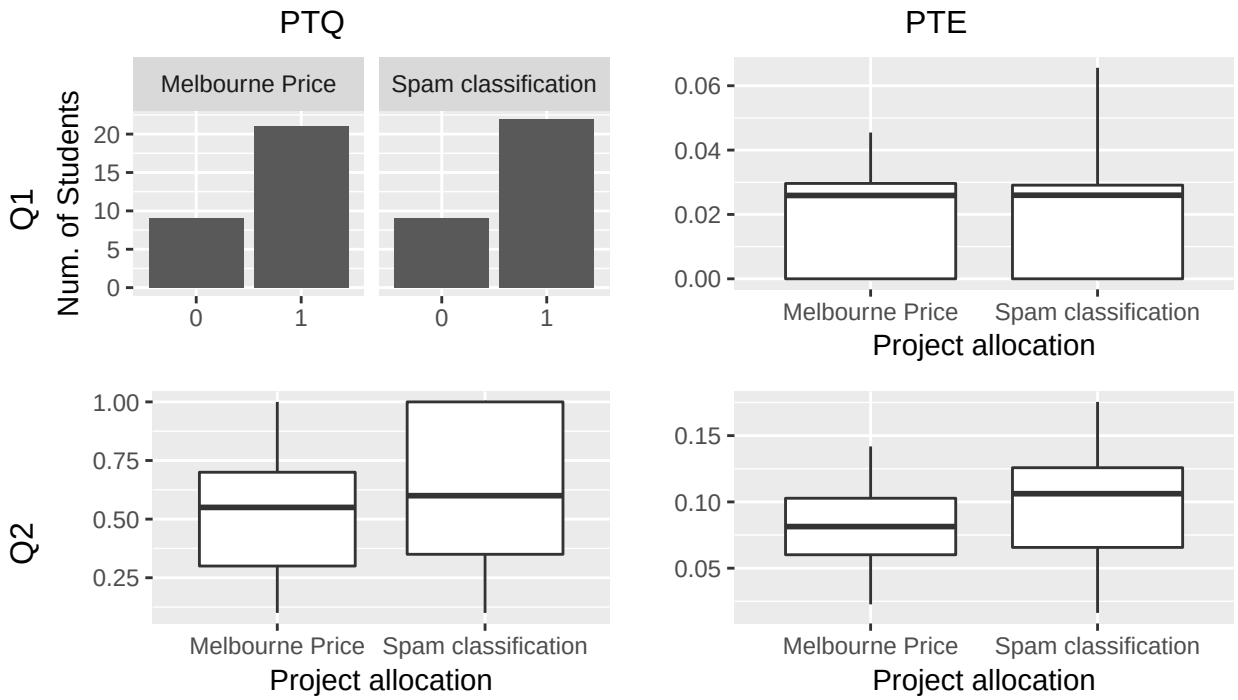

Figure 2: PTQ and PTE scores distributions for classification questions

explain different line of code, the median PTQ scores were similar in the two groups. However, the lower 50 percent of student's scores was spread over the much lower scores. In fact, the second quartile of the scores of students participated in the classification competition covered the range of the scores of the 50 percent of student's scores participated the regression competition. This suggest that participation in the competition help the student to remember the R command better. Similar behaviour can be observed in the performances in Q15, and easy 8 point question, dealing with the differences between Lasso and Ridge regressions.

In the performances in Q8, a 8 point fairly easy question, that dealing with the regression trees, a 'mirror' behaviour can be observed. The median PTQ scores were similar in the two groups, however the third quartile of the scores of students participated the regression competition spread in the range of the 3^{rd} and the 4^{rd} quartile of the students' scores participated the classification competition.

Finally, looking on the PTQ scores, in the Q16, 4 point question, that required a deep understanding of the GAM method, one of the regression technic, the students participated the regression competition had a clear advantage over their pears participated the

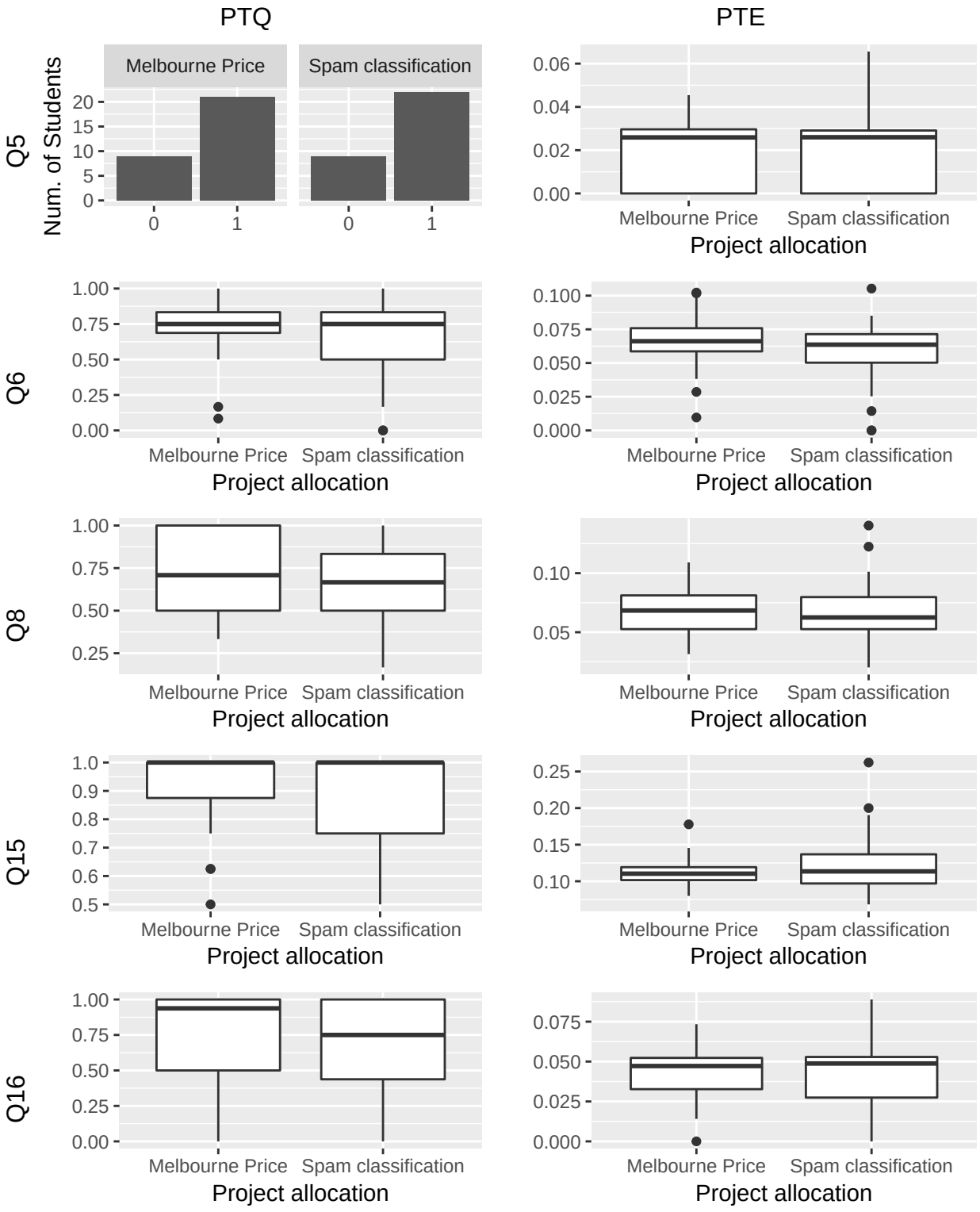

Figure 3: PTQ and PTE scores distributions for regression questions

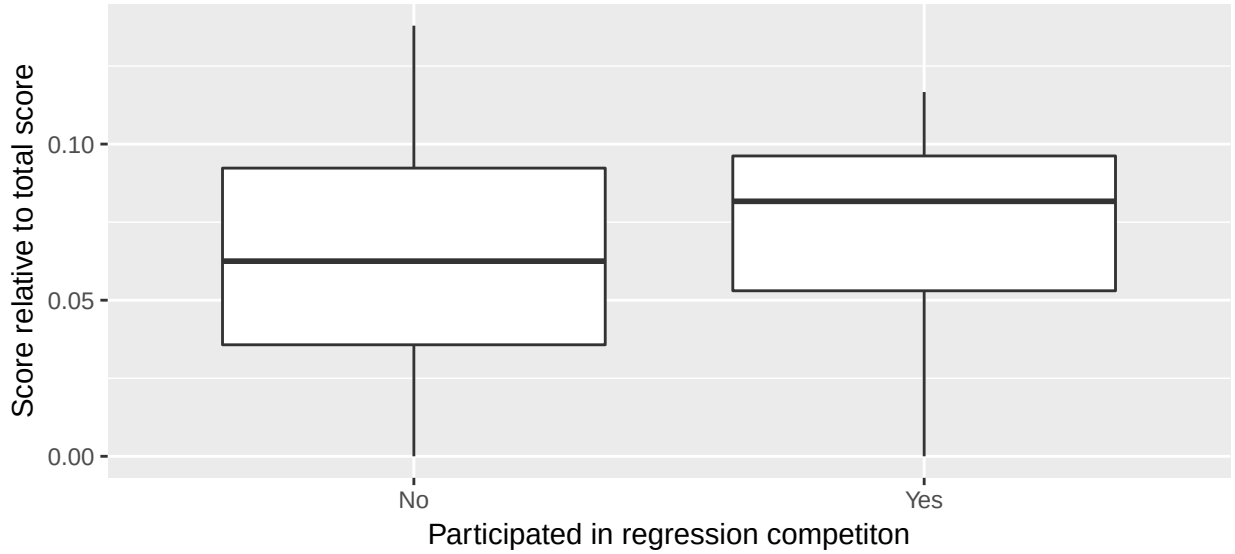

Figure 4: Performance for regression question relative to total exam score for students who did and didn't do the regression data competition in the Statistical Thinking course at Monash.

classification competition.

Next we examined the statistical significance of the mean differences between the performances of the two group of students for Q6, Q8, Q15 and Q16.

Examining all the remind exam questions weren't arise any different between the two groups of students.

4.1.2 Monash students

Figure 4 shows the results for students in the Statistical Thinking course at Monash University. Only the post-graduate students participated in the regression competition, as their additional assessment requirement. Scores for the question on regression in the final exam were compared with the total exam score (PTE defined earlier). The boxplots suggest that the students who participated in the challenge performed realtively better on the regression question than expected given their total exam performance: the median is higher and there is less variability.

On average the students who participated in the kaggle challenge scored 0.0191667 higher than those that didn't, a median of 0.0720037 in comparison to 0.0629408. Using

a permutation test, this corresponds to a significant difference in medians, with p-value of 0.031. A conventional two-sample t-test, of mean difference, yields similar results with a p-value of 0.0797066.

4.2 Engagement

To examine the correlation between student's engagement levels and the performances in the exam we plot the number of submission during the competition versus the performances in the exam, Figure 5. Once again we exam the performances based on two normalizations. Once normalizing by the total possible marks for the relevant cluster of questions (PTQ) and once normalizing by total exam marks (PTE). For the students participated in the Melbourne Price competition is the cluster of regression questions. For the students participated in the Spam Classification competition is the cluster of questions about classification methods.

In Figure 5 we can clearly see a weak positive correlation between the number of submissions during the data competition and the scores normalized to the total possible marks for the relevant cluster of questions. This suggest that as more engaged the student was with the competition, the question about the methods relevant to her competition were easier to her. There is no correlation between the number of submission and the marks for the relevant cluster of questions normalized by the total exam marks (PTE). *BECAUSE ... the questions that unrelated to the data competitions (51 points)? students put less effort to learn other material? Harder questions?*

We didn't found any evidence for correlation between the performances in the competition (final score) and the performances in the exam. This suggest that the single fact of participation improve the students marks in the exam. Not necessarily better students in the competition have grater chances to success in the exam.

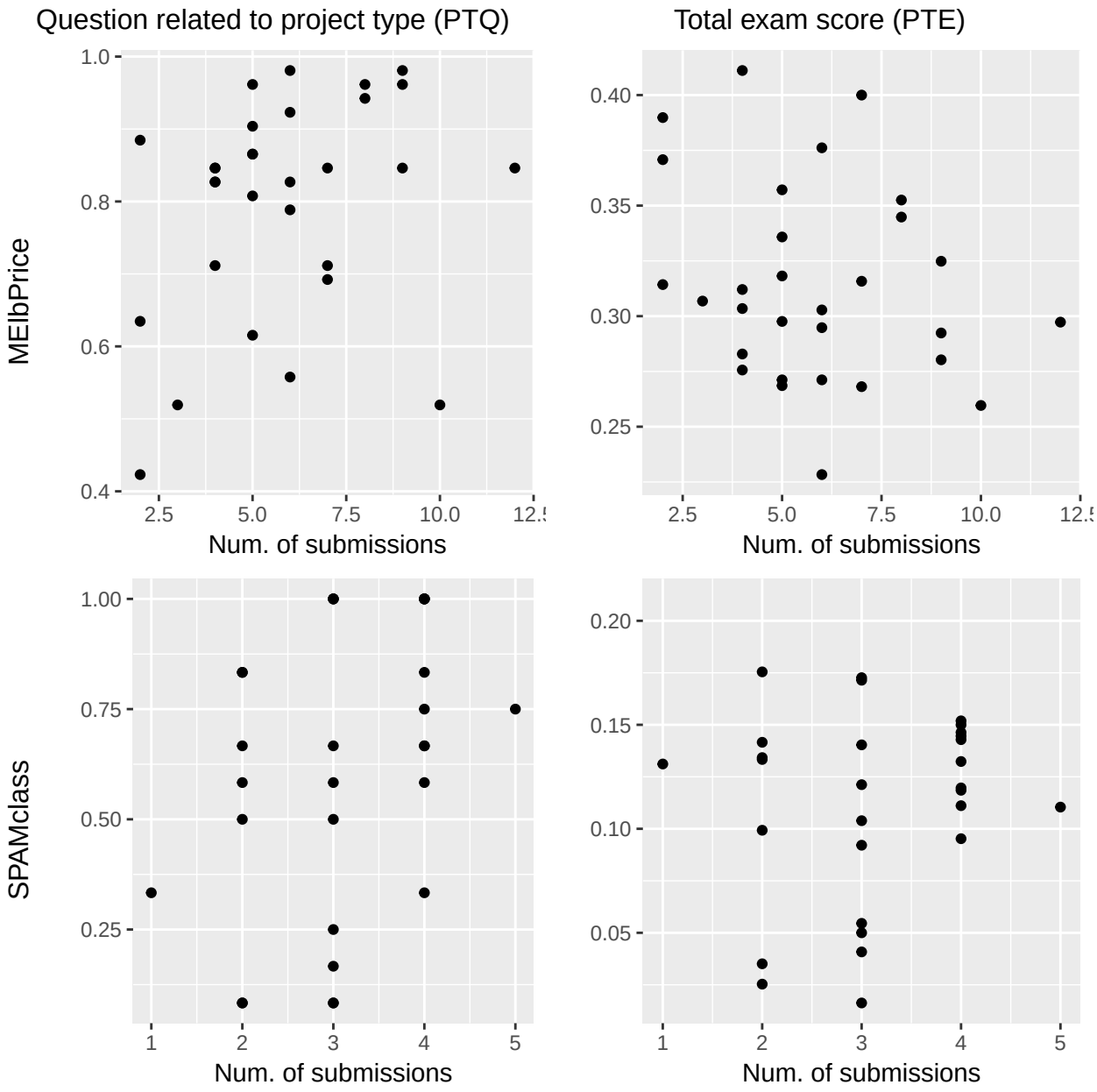

Figure 5: Number of submission vs exam grades

4.3 Interest

5 Discussion

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 the 4th, 2017) and by Monash University Human Research Ethics Committee (ID: 9985 on 24/08/2017).

7 References

8 Stuff to hang onto for now

9 Verifications

This section will be just long enough to illustrate what a full page of text looks like, for margins and spacing.

Campbell & Austin (2002) Schubert et al. (2013; Chi et al. 1981)

10 Appendix

There were two data sets used for competitions, one was a regression problem and the other a classification problem.

10.1 Melbourne price data set

This data contained records of auction prices for residentia properties in Melbourne. The data was compiled from auction reports collected between Feb 2, 2013 and Dec 17, 2016.

Auction reports were published weekly by Domain and compiled by Home Price Guide (R) *COMMENT: Found the symbol* in pdf format. To read them into R we used the XXX

package *COMMENT: reference to our new package for reading those reports.* The reports contained the following information about the properties sold during the week: full address, number of bedrooms, property type (house, unit/ duplex, townhouse, development site or other residential), sold price, type of sale (sold, sold prior, passed in, no bid, vendor bid, withdrawn prior to auction, sold after auction and N/A - price or highest bid not available) and agent name. We also removed all the records without the price. Over all we had 75,367 sold properties for the data competition.

To enrich this data, to make it more realistic we added several additional features, number of visitors, the average rating given to the property by the visitors, number of car spaces, number of baths, land size and house size. It is possible that some of this information might have been extracted from other web sites, given that we had the full address of the property, but the time required to research that was too daunting to tackle for the upcoming teaching period.

10.2 Spam data set

References

Campbell, J. I. & Austin, S. (2002), 'Effects of response time deadlines on adults' strategy choices for simple addition', *Memory & Cognition* **30**(6), 988–994.

Chi, M. T., Feltovich, P. J. & Glaser, R. (1981), 'Categorization and representation of physics problems by experts and novices', *Cognitive science* **5**(2), 121–152.

Schubert, C. C., Denmark, T. K., Crandall, B., Grome, A. & Pappas, J. (2013), 'Characterizing novice-expert differences in macrocognition: an exploratory study of cognitive work in the emergency department', *Annals of emergency medicine* **61**(1), 96–109.

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