Work in Progress: Finding the Right Questions: Using Data Science to Close the Loop with Classroom Response Systems

Asuman Cagla Acun Sener, University of Louisville

Asuman Cagla Acun Sener holds B.S. and M.S. degrees in Computer Science and Computer Engineering. She is currently pursuing a doctoral degree in Computer Science at Knowledge Discovery & Web Mining Lab, Department of Computer Science and Computer Engineering, University of Louisville. She is also working as a graduate assistant. Her research interests are educational data mining, visualization, predictive modeling and recommender systems.

Prof. Olfa Nasraoui , University of Louisville

Olfa Nasraoui is Professor of Computer Engineering and Computer Science, Endowed Chair of e-commerce, and the founding director of the Knowledge Discovery and Web Mining Lab at the University of Louisville. She received her Ph.D. in Computer Engineering and Computer Science from the University of Missouri-Columbia in 1999. From 2000 to 2004, she was an Assistant Professor at the University of Memphis. Her research activities include Data Mining/ Machine Learning, Web Mining, Information Retrieval and Personalization, in particular in problems involving large multiple domain, high dimensional data, such as text, transactions, and social network data. She is the recipient of the National Science Foundation CAREER Award, and the winner of two Best Paper Awards, a Best Paper Award in theoretical developments in computational intelligence at the Artificial Neural Networks In Engineering conference (ANNIE 2001) and a Best Paper Award at the Knowledge Discovery and Information Retrieval conference in Seville, Spain (KDIR 2018). She has more than 200 refereed publications, including over 47 journal papers and book chapters and 12 edited volumes. Her research has been funded notably by NSF and NASA. Between 2004 and 2008, she has co-organized the yearly WebKDD workshops on User Profiling and Web Usage Mining at the ACM KDD conference. She has served on the program committee member, track chair, or senior program committee of several Data mining, Big Data, and Artificial Intelligence conferences, including ACM KDD, WWW, RecSys, IEEE Big Data, ICDM, SDM, AAAI, etc. In summer 2015, she served as Technical Mentor/Project Lead at the Data Science for Social Good Fellowship, in the Center for Data Science and Public Policy at the University of Chicago. She is a member of ACM, ACM SigKDD, senior member of IEEE and IEEE-WIE. She is also on the leadership team of the Kentucky Girls STEM collaborative network.

Dr. Jeffrey Lloyd Hieb, University of Louisville

Jeffrey L. Hieb is an Associate Professor in the Department of Engineering Fundamentals at the University of Louisville. He graduated from Furman University in 1992 with degrees in Computer Science and Philosophy. After 10 years working in industry, he returned to school, completing his Ph.D. in Computer Science Engineering at the University of Louisville's Speed School of Engineering in 2008. Since completing his degree, he has been teaching engineering mathematics courses and continuing his dissertation research in cyber security for industrial control systems. In his teaching, Dr. Hieb focuses on innovative and effective use of tablets, digital ink, and other technology and is currently investigating the use of the flipped classroom model and collaborative learning. His research in cyber security for industrial control systems is focused on high assurance field devices using microkernel architectures.

WIP: Finding the Right Questions: Using Data Science to Close the Loop with Classroom Response Systems

Introduction

This work in progress paper explores the use of data science to analyze classroom response system (CRS) data. A CRS is an educational technology tools that when paired with an appropriate pedagogy, such as team-based learning, provide increased classroom engagement in support of improved teaching and learning [1]-[4]. They do this by leveraging technology to allow every student to respond to instructor posed questions. Many of these systems, such as Learning Catalytics and clickers, collect and store a wealth of individual student response data that is aggregated to provide instructors with real time (in-class) student response data [5]. Several efforts for the analysis of CRS data have been reported [6]. Some of them focused on comparing traditional and team-based approaches [7]. Other studies have performed data analysis on student surveys [8]-[10], or combined historical student grades with survey responses [11]-[13]. One open challenge in this setting is how to glean insights from all of the collected response data to identify activities, specific questions, and combinations of questions that associate with student performance. A data driven analysis of student response data collected by a CRS combined with student performance data will enable individual instructors to refine and adapt their use of a CRS, thus closing the loop from an instructional design perspective. This paper presents a data science methodology and preliminary results of analyzing CRS data accumulated from daily activities in two sections of an a calculus I course taken exclusively by engineering students. The data is collected from the CRS Learning Catalytics where students respond to questions in two rounds following a team-based learning model. In the first round, students answer questions individually; in the second round, they answer the same questions as a team, reviewing each other's answers from round 1 and receiving feedback about correctness of their answers. The CRS stores each student's responses from both rounds along with a timestamp.

Objectives

The objective of this study is to develop and employ a data science methodology to aggregate and explore the data collected by the use of a CRS, with the final goal to help answer two questions: 1) examine the effect of the difference between individual and team-based responses on student performance, and 2) identify which activity question scores, individual and team-based, are associated with better exam performance, thus possibly allowing the reduction of the number of questions.

Methods

The data consists of classroom activity scores and the exam score for a single unit. Class activity data was collected through the classroom response system Learning Catalytics. The score data thus consisted of the unit exam score and the class activity scores for 30 questions for each of 53 students. For each question, students receive two scores as described above, with each score being in the range (0,4). Table A1 in Appendix A shows sample data for four students, showing the round 1 and round 2 scores for two questions. Using the original question scores, we constructed three sets of data to use in building predictive models for the unit exam score.

Dataset 1 has all the round 1 and round 2 scores for each of 30 questions for each student. This data thus consists of 61 variables (the scores for all 30 questions and the first unit exam) for each student (a 53X61 data matrix). *Dataset 2* has only a difference factor (df) scores, calculated for each student as the difference between the round 2 score (team based) and round 1 score (individual). This data thus consists of 31 variables (the round 2 score minus round 1 score difference for each of the 30 questions and the first unit exam) for each student (a 53X31 data matrix). *Dataset 3* is the combination of dataset 1 and 2. This data thus consists of 91 variables (the scores in round 1, round 2, and the round 1 minus round 2 differences, for all 30 questions, in addition to the first unit exam) for each student (a 53X91 data matrix).

In this preliminary work, we report on a limited number of methods. We first use exploratory analysis, then build predictive models of exam performance to help explore which in-class activity questions play an important role in student performance. We use scikit-learn version (0.20.2) [14], a Python (version 3.6) machine learning library, to build predictive models. For all three predictive models, we use as target the score for the first unit's exam. We split the data into two subsets, with 80% (of the students' scores) used for training the predictive models and the remainder (20%) used for testing the models.

We created heat maps of the students' class activity and exam scores, as well as the constructed features obtained by aggregating these scores. We then built a random forest model, a powerful predictive model that has a built-in mechanism to filter the set of possibly correlated predictor variables to only the most predictive ones [18]-[19]. We then computed the feature importance scores to select the score-based features which contribute most to predict the unit's exam score, and thus the activities that are most predictive with the exam score.

Results

In the following, and due to space limits, we report only a selection of the results that seemed to be interesting on the entire combined data set (Dataset 3). We have 30 questions for each of 54 students in this dataset, thus a total of 1620 individual class activities in Unit 1, of which 1076 remain after absences are excluded.

Table 1: Summary of Individual and Team-Based scores on class activities

Total number of questions answered (absences excluded)	Number of times that students got the same score on each round	Number of times that students got a lower score on round 2	Number of times that students got a higher score on round 2
1076	497	58	521
100%	46.2%	5.4%	48.4%

From Table 1, we see that 48.4% of activities in-class, resulted in getting higher scores after the team-based discussions. When we compare this with the 5.4% of students whose scores decreased after team discussions in round 2, we may see that the benefit of working in a team-based model outweighs the negatives.

Appendix B - Figure 1 shows the heatmap of several aggregated constructed features for each student, such as averages of round 1 and round 2 scores, and average of improvements from round 1 to round 2, as well as the proportion of activities where the round 2 team-based scores

improve, deteriorate or remain the same compared to the individual score (round 1). A quick glance at the entire visualization shows that the majority of students improve their activity scores as a team (round 2). Furthermore, the first and last columns in this visualization show a tendency for a better exam score with higher improvements in team-based activities relative to the individual score. The conclusions from the visualization may be limited by small samples.

Appendix B - Figure 2 shows the correlation matrix between aggregated scores and exam score. The exam scores are positively correlated with the round 1 and round 2 (stronger correlation) scores and they are even stronger correlated with the number of improvements from round 1 (individual score) to round 2 (team score). This may indicate that team-based activities in the classroom are important. Further analysis would be needed to study the separate effect of round 1 performance and team-based improvement.

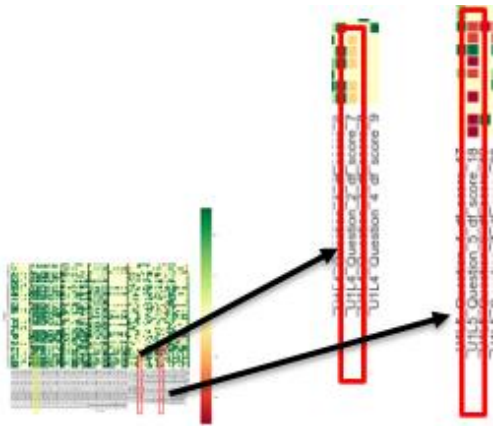


Figure 1: shows a heatmap of Dataset 3's activity scores, df scores (round 2 minus round 1) and Exam 1 Score, in addition to the constructed feature (Number of df score above zero/ Number of df). Data is sorted in ascending order of unit 1 exam score. Most activities result in improved scores after the team discussions, compared to individual work. However, Unit 1 Lesson 5 Question 5 results in an interesting anomalous pattern in Figure 1 (See Appendix C - Figure 3 for the entire heatmap) where the round 2 minus round 1 difference scores are negative for several of the top performing students (the bottom rows in the heatmap), indicating a lower score after team discussions. A similar pattern is observed for Unit 1 Lesson 4 Question 2. Although the numbers are low, such an anomaly is easy to spot in this figure and may result in further study as a follow up with the students.

We built a Random Forest predictor (RF1) which resulted in an accurate model on the left out test set with an RMSE of 0.089. Then we selected the top features based on the computed feature importance and built increasingly complex models, thus increasing the number of features used to build the model by gradually adding the features in the order of the feature importance until the RMSE on the test set stabilizes to a reasonably close value to the full RF1 model. This stabilization occurred at the reduced model (RF2), which resulted in a testing RMSE of 0.092, obtained with the top 25 features. The top 10 and bottom 10 features are listed in Appendix E - Table 1.

Discussion and Conclusions

The features sorted by their importance, show that certain class activity questions (e.g. the top 10 features in Appendix E - Table 1 and sample questions are listed in Appendix D) are highly predictive of the exam score at the end of the unit; while certain questions (e.g. the bottom 10 features in Appendix E - Table 1, a sample of which is listed in Appendix B) seem to have little impact on the Exam 1 score. Furthermore, for some of these questions (e.g. Unit 1 Lesson 4 Question 2), the difference between the team and individual round scores seems to be important determinants of the unit's exam score in addition to the round 1 score. This, in combination with the direction of impact on the exam score, may indicate that team-based activities are particularly effective for certain questions. However, Appendix B - Figure 3 (see yellow outline annotations in the figure) shows that overall most students' round 1 and round 2 scores were similar for this question, except for a small group of students who got the top scores in Exam 1. Surprisingly their scores in round 2 decreased compared to round 1 (see red outline around the orange colored cells on the left in Appendix B - Figure 3), a similar pattern observed for Unit 1 Lesson 5 Question 5 (see red outline around the red colored cells on the right in Appendix B - Figure 3) which motivates examining these cases closely and scrutinizing the question or the team formations.

Making conclusions based on the importance order of the features corresponding to the questions is not straightforward since it is possible that round 2 and round 1 scores for the same questions may be correlated, which would eliminate one of the rounds from the top features. However we can quickly glance at the heatmap visualization in Appendix B - Figure 3 to verify that the top feature, hence the most important question (Unit 1 Review 1 Question 1) round 2 is significantly higher than round 1, and furthermore that Round 2 scores for this question are positively associated with the exam score in Unit 1 (easy to see since the exam score is sorted in Appendix B - Figure 3 - last column) for most students. Yet the comparative order (ranking based on feature importance) of the questions remains relevant for gauging question importance. Future analysis should consider extending our predictive model to be able to analyze local association between questions and performance for groups of similar students instead of the entire class, paving the way towards a *personalized* approach to question design.

To conclude, our preliminary study based on visualization offers one way to explore disparate data using a shallow but fast process that relies on visual perception to spot patterns, trends, anomalies and dependencies. This approach, although simple, is limited by the need to become familiar with the specific visualizations and their interpretation within a particular context. Even after sufficient familiarity is established with the visualizations, there is a risk of subjectivity in interpreting visualizations. Despite these limitations, we emphasize that visualization should only be a preliminary step to spot patterns and help the domain expert formulate hypotheses that can later be analyzed using rigorous methods in order to answer research questions. Our preliminary study of predictive models and feature importance is a proof of concept that analyzing score based features may shed some light on which questions hold higher prediction power of the students' performance in the exam at the end of the unit. An analysis of the coefficients of a linear model or another interpretable predictive model such as a decision tree regression model, might be able to better understand how each of the questions relates to the unit exam score. Our analysis has additional limitations, the most important of which may be the effect of the unit's

exam questions on which classroom questions turn out to be important. This can be circumvented by a content-based analysis that pays attention to the actual content of questions, and by an analysis that hones in on the student's performance in each individual exam question, in addition to the total exam score. Despite its limitations, the type of analysis performed in this paper may one day allow the instructor to fine-tune the choice of questions in order to design optimal classroom activity questions, thus closing the loop in classroom response systems. Future work will address the limitations and will expand the analysis in both scale (units and courses) and methods.

References

[1] Rezaee, R., Moadeb, N., & Shokrpour, N. (2016). Team-based learning: A new approach toward improving education. *Acta Medica Iranica*, 54 (10), 679-683.

[2] Carmichael, Jeffrey. 2009. "Team-based Learning Enhances Performance in Introductory Biology." *Journal of College Science Teaching* 38(4):54–61.

[3] Filer, Debra. 2010. "Everyone's Answering: Using Technology to Increase Classroom Participation." *Nursing Education Perspectives* 31(4):247–50.

[4] Fies, Carmen, Marshall, Jill. 2006. "Classroom Response Systems: A Review of the Literature." *Journal of Science Education and Technology* 15(1):101–109.

[5] Middleditch & Moindrot, Cogent Economics & Finance (2015), 3: 1119368. Doi: 10.1080/23322039.2015.1119368

[6] H., Paul et al. "Analysis of the Team-Based Learning Literature: TBL Comes of Age" *Journal on excellence in college teaching* vol. 25,3-4 (2014): 303-333.

[7] J. Johnson, E. Bell, M. Bottenberg, D. Eastman, S. Grady, C. Koenigsfeld, E. Maki, K. Meyer, C. Phillips, L. Schirmer. A Multiyear Analysis of Team-Based Learning in a Pharmacotherapeutics Course. *American journal of pharmaceutical education*. 78. 142. Doi: 10.5688/ajpe787142.

[8] M. Awatramani and D. Rover, "Team-based learning course design and assessment in computer engineering," 2015 IEEE Frontiers in Education Conference (FIE), El Paso, TX, 2015, pp. 1-9. doi: 10.1109/FIE.2015.7344227

[9] Borges NJ, Kirkham K, Deardorff AS, Moore JA. Development of emotional intelligence in a team-based learning internal medicine clerkship. Medical Teacher. 2012;34:802–806.

[10] Jacobson TE. Team-based learning in an information literacy course. Communications in Information Literacy. 2011;5(2):82–101.

[11] Conway SE, Johnson JL, Ripley TL. Integration of team- based learning strategies into a cardiovascular module. *American Journal of Pharmaceutical Education*. 2010;74(2):1–7.

[12] Dana SW. Implementing team-based learning in an introduction to law course. *Journal of Legal Studies Education*. 2007;24(1):59–108.

[13] Letassy NA, Fugate SE, Medina MS, Stroup JS, Britton ML. Using team-based learning in an endocrine module taught across two campuses. *American Journal of Pharmaceutical Education*. 2008;72(5):1–6.

[14] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." *Journal of machine learning research* 12.Oct (2011): 2825-2830.

[18] L. Breiman, “Random Forests”, Machine Learning, 45(1), 5-32, 2001.

[19] L. Breiman, J. Friedman, R. Olshen, and C. Stone, “Classification and Regression Trees”, Wadsworth, Belmont, CA, 1984.

Appendix A:

Table A1: Sample data for four students round 1 and round 2 scores for two class activity questions in Unit 1 Lesson 1 (U1L1)

Index	_U1L1_Question_1_(round_1)_score	_U1L1_Question_1_(round_2)_score	_U1L1_Question_2_(round_1)_score	_U1L1_Question_2_(round_2)_score
0	4	4	0	4
1	4	4	0	4
2	4	4	4	4
3	4	2	4	1

Appendix B: Figure 1, Figure 2 and Annotations on Figure 3 (heatmap visualization accompanying the discussion of results)

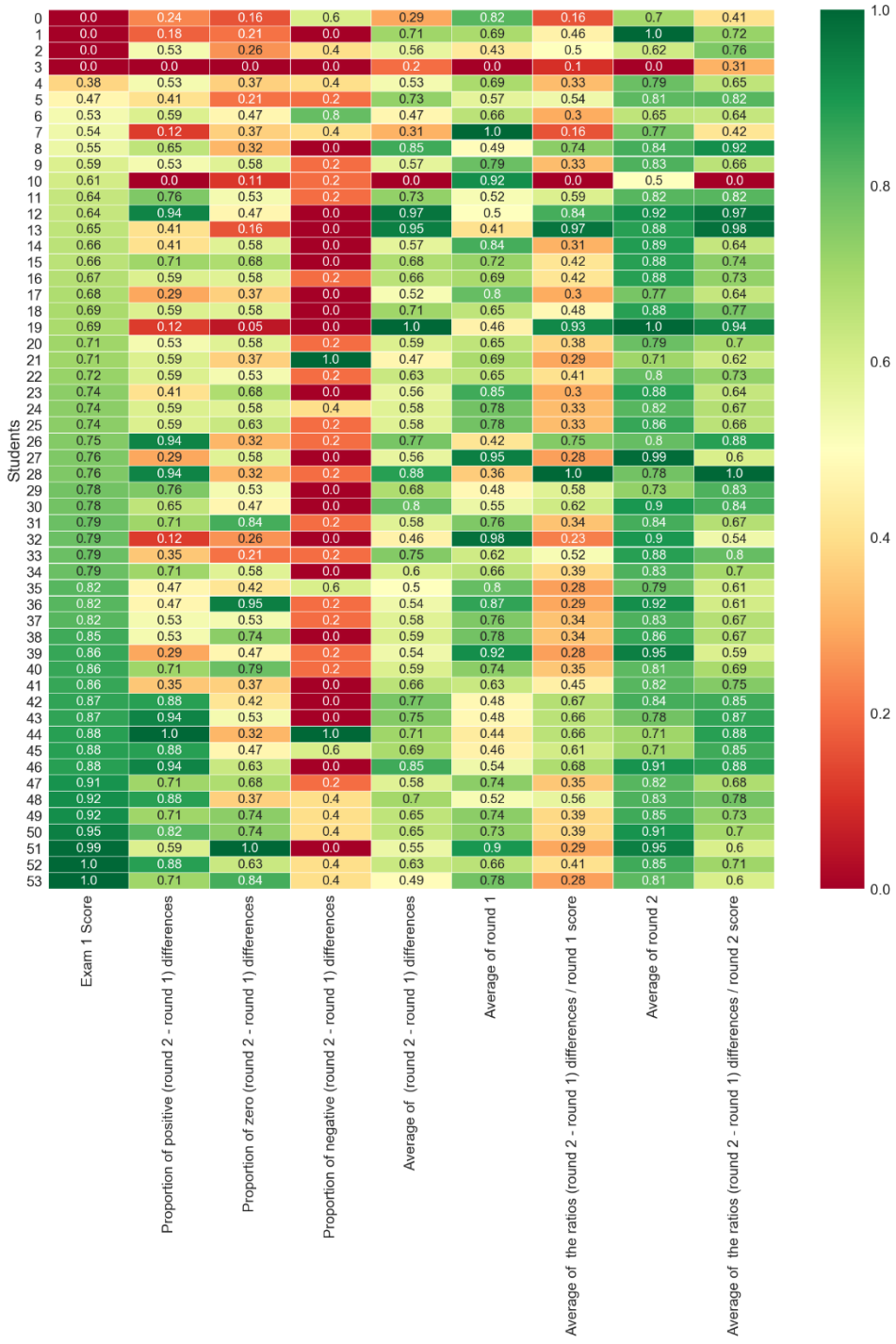


Figure 1: Heatmap of the calculated aggregates: columns: Exam 1 Score, Proportion of positive (round 2 - round 1) differences, Proportion of zero (round 2 - round 1) differences, Proportion of negative (round 2 - round 1) differences, Average of (round 2 - round 1) differences, Average of round 1, Average of the ratios (round 2 - round 1) differences / round 1 score, Average of round 2, Average of the ratios (round 2 - round 1) differences / round 2 score. Rows: students. All ranges are in [0,1].

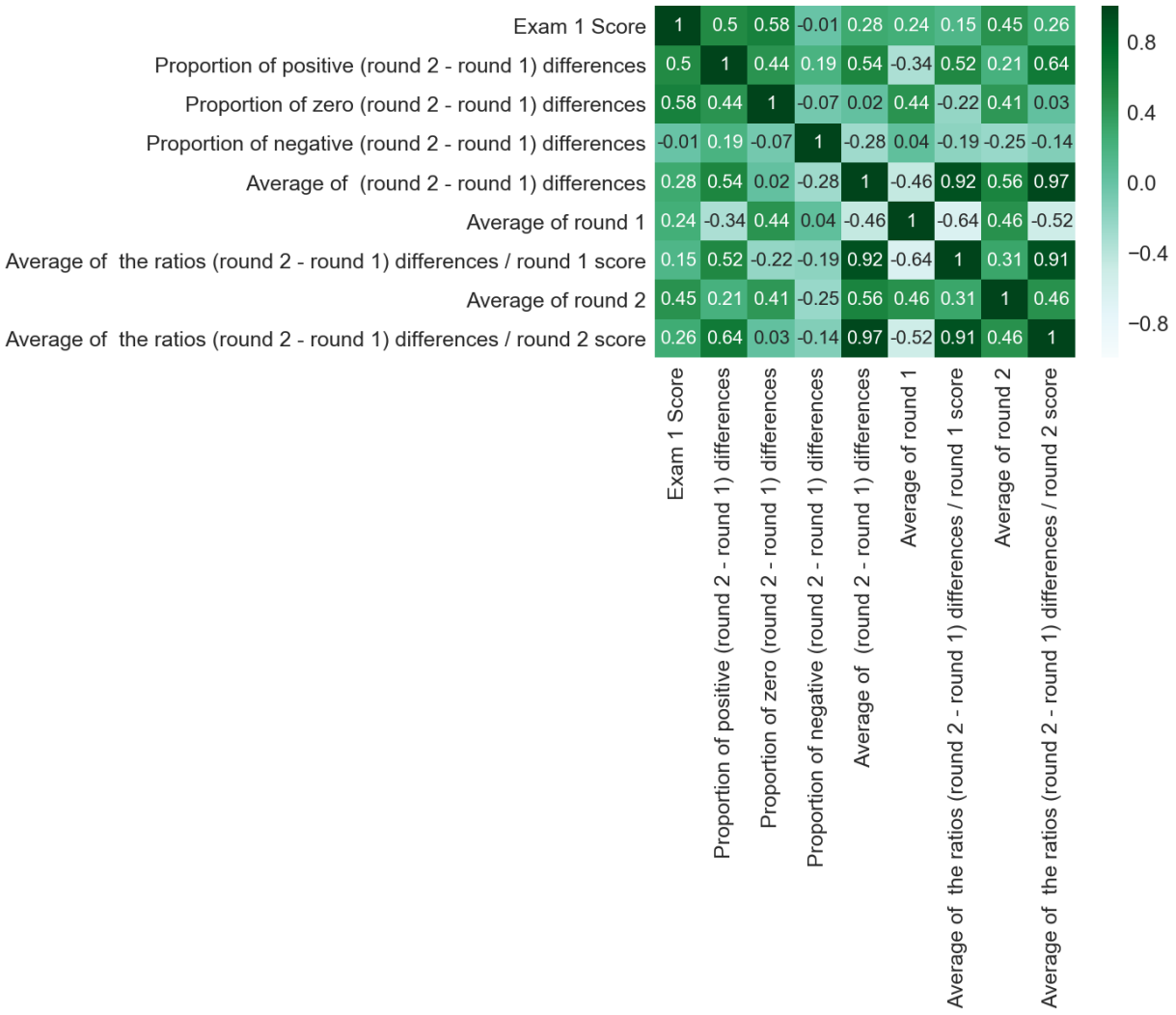


Figure 2: Correlation matrix between aggregated scores and exam score.

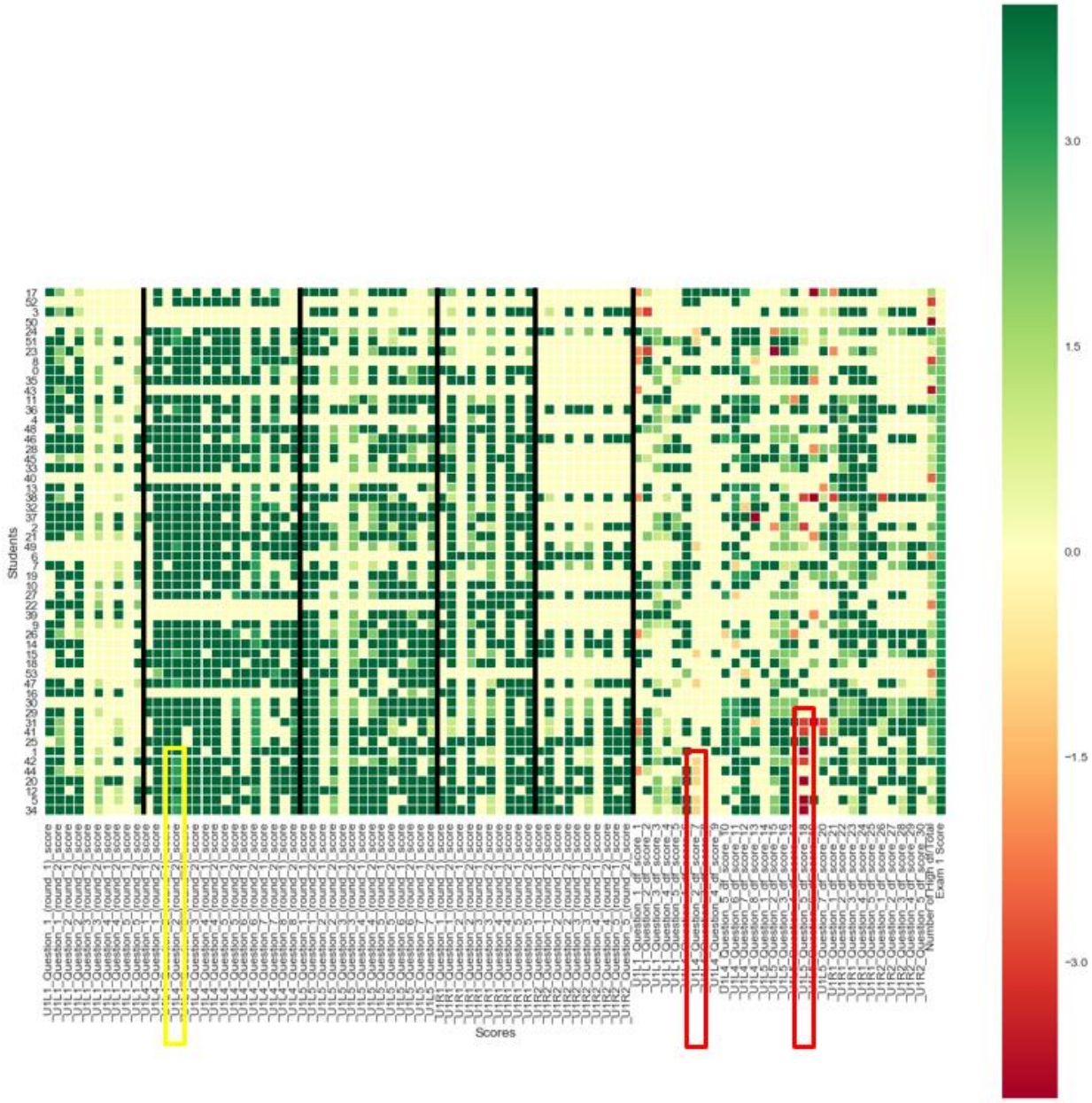


Figure 3: Data set 3: columns are activity scores, df scores (round 2 minus round 1) and Exam 1 Score, in addition to the constructed feature (Number of df score above zero/ Number of df). Rows are students. Exam 1 score range (0, 1). Df scores range (-4, +4). Data is sorted in ascending order of unit 1 exam score.

Appendix C: A sample of top 5 questions

1. Unit 1 Review 1 Question 1

Does the following limit exist, and if it does evaluate the limit.

$$\lim_{x \to 2} \frac{|4x - 8|}{x - 2}$$

Hint: Evaluate $f(0)$ and $f(4)$

A. -4

B. 0

C. 4

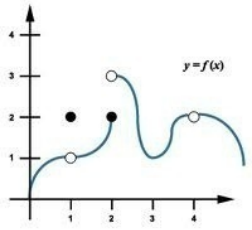
D. This limit does not exist

E. None of the above

2. Unit 1 Lesson 4 Question 2

Use the graph of f(x) to answer the following:

$$\lim_{x \to 2} f(x)$$



A. 2

B. 1

C. 3

D. Positive infinity

E. Does not exist

3. Unit 1 Lesson 1 Question 2

Solve the following equation (give only the value, or values for x)

$$\frac{4}{x+3} - \frac{2}{x-3} = \frac{5x}{x^2-9}$$

4. Unit 1 Lesson 4 Question 8

Evalaute the following limit

$$lim_{x \to -5^-}(x + 11)\frac{|x+5|}{x+5} = ?$$

Appendix D: A sample of bottom 5 questions

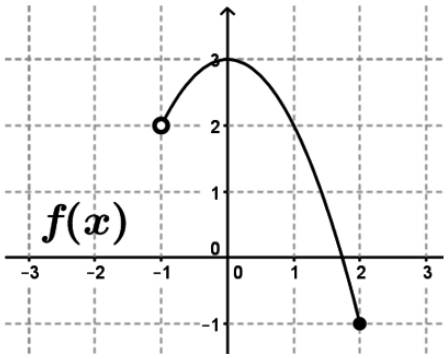
1. Unit 1 Lesson 1 Question 3

Solve the complex equation. (Find z, where z=a+bi).

Enter your answer in the form a+bi (or a-bi).

$(\frac{1+i}{1-i})^2 + \frac{1}{z} = 2 + 5i$

2. Unit 1 Lesson 5 Question 1



For what values of *x* is *f(x)* continuous?

A. [-1, 2]

B. (-1, 2)

C. [-1, 2)

D. (-1, 2]

E. None of the above

3. Unit 1 Lesson 4 Question 3

Evalute the limit:

$lim_{x \to -3}(2x - 5)$

4. Unit 1 Review Question 4

A force of $\vec{F} = 6\hat{i} + 7\hat{j}$ is used to pull a ramp (from the botton to the top). The ramp has an incline, or grade of 30%. IF the base of the ramp is 10 feet how much work is done pulling the box up the ramp. (just enter your answer, appropriate units of work are implied).

5. Unit 1 Review Question 5

Find the equation of the line parallel to the vector $2\hat{i} + 3\hat{j}$ passing through the point P(2-1).

Put your answer in point slope form (y=mx+b) and enter the right hand side as your answer (mx+b)

y=?

Appendix E:

Table 1: Top 10 and bottom 10 features based on feature importance scores in the trained random forest model that predicts the exam score in Unit 1.

Top features	**Importance score**	**Bottom features**	**Importance score**
U1R1 Question 1 (round 2)	1.00	U1L1 Question 3 (round 1)	0.0
U1L1 Question 2 (round 2)	0.85	U1L1 Question 5 (round 1)	0.0
U1L4 Question 8 (round 2)	0.54	U1L1 Question 4 (round 1)	0.0
U1L4 Question 2 df score 7	0.53	U1L4 Question 3 (round 2)	0.0
U1L4 Question 2 (round 1)	0.37	U1L4 Question 5 (round 2)	0.0
U1L5 Question 3 df score 16	0.36	U1L5 Question 1 (round 1)	0.0
U1L1 Question 2 df score 2	0.32	U1L5 Question 1 df score 14	0.0
U1L5 Question 2 df score 15	0.30	U1R1 Question 4 (round 1)	0.0
U1L4 Question 3 (round 1)	0.26	U1R2 Question 5 df score	0.0
U1R1 Question 5 (round 2)	0.19	U1R2 Question 3 (round 1)	0.0