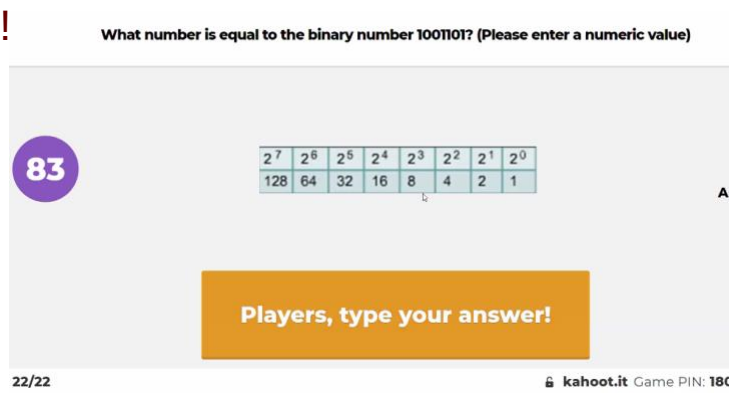
Department of Mathematics

February 22, 2021

Math Club Trivia Quiz a Big Success!



Last Thursday, seven students matched mathematical wits in the Math Club's first-ever online trivia quiz, administered over Zoom and Kahoot. President **Lily Dong** created a fun, 20 question contest with problems like fill-in-the-sudoku box, make-this-number, and some quick computations that required speed and mental agility. The top finishers, **Aidan McAuliffe, Will Grimwood,** and **Uri Tomer**, each won prizes for their efforts. Congratulations!

The next Math Club meeting will be **Thursday, February 25 at 5:30 pm.** To get on the mailing list, contact club President **Lily Dong** (dongl@union.edu)

Math Club meeting Zoom link: **https://union.zoom.us/j/7365779779**

Senior Writing and Pieces from Thesis

Steven Crouch *wrote his senior thesis this past fall term under the direction of Professor Roger Hoerl.*

I wrote my senior thesis titled *Comparison of Big Data Classification Methods: Random Forests and Neural Networks* with Professor Hoerl in the Fall 2020 term. Modern technologies have given rise to an increased reliance on machine learning algorithms to process loan applications, diagnose diseases, filter through job applications, and many more tasks that are otherwise more difficult for manual processing. Therefore, it is vitally important that we identify the strengths and the shortcomings of these algorithms to understand the instances in which their applications are most appropriate.

Random forests and neural networks are considered "black box" algorithms in that the inputs and outputs are well-understood, but the models themselves are not entirely interpretable. We included logistic regression in the study to provide a more transparent alternative as a baseline for comparison. Using the R programming language, we evaluated the effectiveness of each classification method in the context of three datasets with binary (True vs. False, Default vs. Repayment, etc.) response variables: bank marketing data, Lending Club data, and online shopping data. Since we were interested in the algorithms' abilities to classify new data, we performed train-test splits in which the models are trained with 80% of the data and then we measured the classification strength on the "new" 20% of the remaining data. In addition, we performed 10-fold cross validations which are essentially extended versions of train-test splitting intended to boost the rigor of model evaluation.

After using neural networks (and deep learning variations), random forests, and logistic regression to classify the three datasets, we drew conclusions regarding the pros and cons of these methods. One key finding was that random forest tends to overfit data, meaning that it fits the training data nearly perfectly, but when presented with new data, significant reductions in classification performance were noted. On the other hand, the performance lag converged with that of the other two methods, showing that random forest is an effective classifier despite its overfitting. In general, "black box" algorithms are considered stronger methods of classification, but logistic regression in this study proved to be a reliable, transparent alternative. Although some predictive performance is lost in applying logistic regression, it is a strong option for domains in which transparency is necessary, such as processing bank loan applications.

I approached my senior thesis with a nervous optimism as knew it would be a challenging experience. However, I can reflect with gratitude having seized this opportunity and I am thankful for Professor Hoerl's guidance throughout the term. I encourage all math majors to consider writing a thesis as it is a rewarding opportunity to perform in-depth research of your interests and challenge yourself in ways that are certain to promote personal growth.