**Knewton Machine Learning Challenge**

You are a visiting Professor in the first and only University of Mars, which admits exclusively Martians (off-planet tuition would be too expensive for Earthlings anyway). You have joined the Department of Astronomy and are in charge of teaching Astrometrics. Unfortunately the university is quite overcrowded and each semester there are close to 13,000 martian students that take Astrometrics.

Given the size of your class, and a disturbing lack of teaching assistants, you are forced to limit the number of questions given to each student on the mid-term exam to only five multiple choice questions. To further complicate matters the department head has mandated that all test questions be pulled from a bank of approximately 400 “approved” test questions. To increase the robustness and ease of creation of your midterm exam you decide to randomly and uniformly choose 5 questions from the question bank for each student. This means that each student’s exam consists of disjoint, intersecting, or identical sets of 5 questions.

To assist you with running such a large class, the university has provided you with last year’s student performance data (where the professor created exams using a process similar to your own). Additionally, the department head has admitted to you that she thinks that not all the questions in the “approved” question bank are actually useful for evaluating each student’s relative understanding of Astrometrics. She also warns you that while it is OK not to use all the questions from the bank, you must, at minimum, use 50% of them in order to achieve sufficient coverage of the curriculum.

With the mid-term exam only a few days away, which questions should you exclude from the exam generation algorithm such that the exam results will provide the most meaningful ranking of the students in the class? Why? The data set you are given by the department has one record per line, where each record consists of:

1. A unique identifier for a question
2. A unique identifier for a student
3. The correctness of a student's response. A correct answer is marked as a 1; an incorrect response is marked as a 0

Please explain both your reasoning and methodology: we are more interested in the process of your solution than in the final product. Also, we’ll want to read through and be able to run the code you used. (We prefer you use a production-oriented language such as Java, Python, or C/C++ rather than a research platform such as Matlab, R, SAS, or SPSS.) Finally, please include an estimate of the time you spent solving this problem.

Good luck!

- The Knewton Data Science Team