ORIE 4741 - PROJECT PROPOSAL

Jolene Mei (xm87), Charlie Ruan (cfr54), Joe Ye (ty357)

Available Dataset

The available dataset consists of three semesters' pre-Add/Drop and post-Add/Drop enrollment information for Cornell students. Each excel sheet contains all Cornell students' enrolled courses on a specific date of the pre- or post-Add/Drop period. Each row of the excel sheet is a student-course pair that contains the student id, course subject, course catalog, course enrollment capacity, and other relevant information. In our final project, we will protect students' privacy by refraining from disclosing individual student's course schedule.

Problem Overview

We are interested in exploring the following question with the dataset:

- Given the pre-Add/Drop student enrollment information of one semester, can we predict that semester's post-Add/Drop student enrollment?

A group of undergraduate students advised by Professor David Shmoys (named Scheduling Team below) has been helping the University Registrar to schedule final exams via integer programming models. Currently, the scheduling optimization model requires the student-level course enrollment information after the Add/Drop period as input. This data is typically not available until a month after the semester begins. However, this process could be completed much earlier in advance (even before the semester begins) if the team could take advantage of the pre-Add/Drop data instead of waiting until the end of the Add/Drop period.

Besides, knowing such changes in the number of students allows the University Registrar to adjust the number of open spots for each course. In other words, knowing that there will be a certain number of students dropping the course, we can increase the capacity at the beginning of the semester so that the final number of students can match the capacity, allowing more students to take the course they want.

Our Approach

In order to solve the problem with the available dataset, we need to predict the number of students that will drop the course post-Add/Drop given each student's pre-enrollment schedule for a semester. We plan to train a model for each course. Specifically, for a course i, the data sample will be all the students that have enrolled in course i before Add/Drop, across all historical semesters. The feature of each data point will encompass the student's schedule for the semester, consisting of specific classes they intend to take, as well as various class-related attributes such as the number of credits, presence of a final exam, and class format, etc. The data point's label will be 1 if course i is no longer in the student's schedule after Add/Drop, and 0 otherwise. Using logistic regression, the model will output the probability that a student will drop course i. The intuition is that the probability that a student will drop a course is directly related to the difficulty of their schedule.

After the pre-enrollment period for each course, we can use this prediction model to calculate the probability that each student will drop that course. The summation of the output will be the expected number of students dropping the course given their pre-Add/Drop schedules. By knowing such an estimated figure, the University Registrar and the Scheduling Team can solve the problem of early final scheduling and seat offering.