A Review of Course Outcome Scoring

Introduction

In order to provide students with a better understanding of the challenge that may be presented within courses, we aimed to build an informative metric for UW courses based on the performance of students within each course relative to their historic performance.

While one might assume a student's course GPA relative to some central measure of course GPA could be an indicator of difficulty, this can lead to a multitude of issues when instructors grade courses differently across colleges, terms, or even within individual course offerings and sections. GPA measures also neglect students who do not utilize the standard grading option, or withdraw from the course prior to grading. Another option could be to simply look at the proportion of students who have an adverse outcome out of the total number of students who have taken a course, however this neglects the nature of some courses (i.e. first year courses with high withdrawal rates due to freshman enrollments). One might then suggest course-evaluation based measures of course difficulty, although we confront the previous issue of students who withdraw from courses prior to evaluation; as well as various sampling biases like self-selection bias.

Our idea is to build a score for each course indicative of an expectation of adverse outcomes relative to the actual outcomes using data composed of all students enrolled past census day.

Thus, we approach this task by building a model which generalizes well when predicting a course outcome for a given student's enrollment in a quarter, only utilizing data available at and prior to the tenth day of the quarter. The outcome of interest is a binary indicator informing whether the enrollment resulted in a failing grade or withdrawal (FW) from the course. Then, grouping predictions by course offering (e.g. MATH 124, BIOL 180, etc., NOT by section or instructor), we can compute the expected FW rate of a course. The difference between the ground truth FW rate and our expected FW rate is used as a means of scoring the difficulty of a course. In a given course, if more students are receiving a FW than we would expect from our model's estimates, then we propose the course is more difficult. Similarly, if less students are receiving a FW than we estimate, then we propose the course is less difficult.

Data

We have defined a failing grade (GPA<0.7) and withdrawals past census day as adverse outcomes. Hardship withdrawals, incomplete, and in progress grades were ignored for training our model and for generating predictions. However, features related to a student's course load did utilize enrollments that resulted in such outcomes.

As mentioned, data used to create a course outcome prediction is composed of information that would be available at the beginning of a quarter. This means each observation consists of data related to a

student's course load *(i.e. total number of credits, number of courses, course types, etc.)*, historical data about recent and all-time student performance *(e.g. total number of VLPA courses, last quarter's GPA, cumulative GPA in STEM courses for the last 4 quarters, etc.)*, and data related to the course the student is enrolled in *(e.g. number of offerings in the current quarter, historic cumulative proportion of FW outcomes, last quarter's student GPA variance, etc.)*.

Model

The model combines mixed effects with tree-boosting using the GPBoost algorithm in order to handle the temporal and grouped nature of our observations, simultaneously removing the need to manually implement interactions and handling naturally missing values (such as missing a student's numeric grade for the previous quarter when exclusively non-standard grading options are used).

We optimize our model with cross-entropy. We use 5-fold cross validation with a randomized grid search then continue with a localized grid search to determine hyperparameters.

The resulting model is defined as $y = F(X) + Zb + \epsilon$. y is our vector of outcomes, X is our matrix of predictors, Zb is our design matrix multiplied by our vector of random effects, and ϵ is a vector of error terms. Our model defines F as a set weighted trees each mapping our feature space to our output, although when $F(X) = X^T\beta$ with β as a coefficient vector, then the model is just a linear mixed effects model.

Results & Interpretation

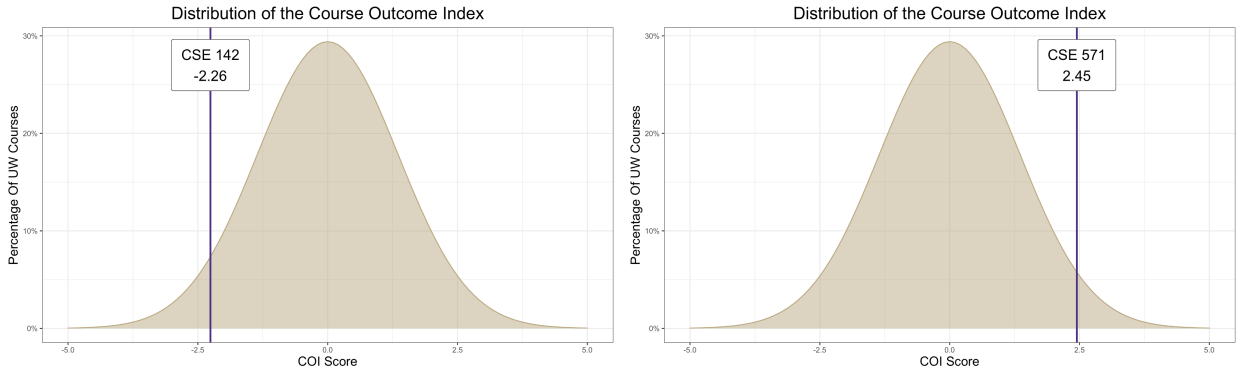
After building our predictions for each student and grouping enrollment data by course, we calculate actual FW rates subtracted from expected FW rates, meaning higher values indicate less difficult courses.

Rate differences are then normalized, and then scaled using the following formula: $x_i = (e^{-\frac{Q_1(n)}{2n_i}})(r_i)$ where r_i is the normalized rate difference for a given course i, n is the set of total course enrollments for all evaluated courses, $Q_1(n)$ is the first quartile of n, and n_i is the number of enrollments in course i.

Scores are adjusted by $e^{-\frac{Q_1(n)}{2n_i}}$ due to the observation that prediction error can cause courses with a very small number of enrollments to have inflated estimated rate differences, leading to a potentially misrepresentative score. This function scales the rate difference such that courses with very few enrollments per year are shrunk towards zero, but courses with at least several enrollments per year are minimally impacted. We finally apply ordered quantile normalization and rescale values to [-5,5].

As desired, using Spearman's rank correlation our course scores are almost entirely unrelated to course GPA, however our scores have a weak positive correlation with actual FW rate (0.20).

Investigating individual courses, we can look at two different CSE courses. Although the subject matter of CSE 142 may be less intense than that of CSE 571, we can see that our method ranks CSE 571 as an "easier" course because our method is intended to indicate a notion of difficulty for students who take these courses. Our score indicates that a CSE 142 student enrollment would be more likely to result in an adverse outcome than a CSE 576 enrollment.



We see many courses in the school of medicine in the lower end of our scores. We also see many other notoriously difficult courses at the more difficult end of our scale, including the PHYS 12X, MATH 12X, and CSE 14X series. However, we also see many non-STEM courses, including multiple courses in the school of social work, college of education, linguistics, and Japanese (just to name a few).

Conclusion

Since the score is intended to represent the likelihood of an unexpected adverse outcome for a typical student enrollment, we feel this score is a valuable indicator of the subjective concept that is a course's difficulty, but of course cannot completely represent the idea. This concept is still a work in progress, and there are a variety of data sources and feature engineering methods that could be utilized to better inform our model.

Updated on 2/8/2023