Understanding Duke Course Popularity

Ethan Holland

Abstract

Understanding college course registration behavior not only gives insight into the student decision making process but is also useful for universities planning courses and professors trying to encourage students to take their courses. This project models undergraduate course fullness at Duke University to identify factors associated with course popularity and to be able to predict whether out-of-sample classes will fill-up. Data is collected from DukeHub and partial-pooling is used to account for small categories and new categories.

Data

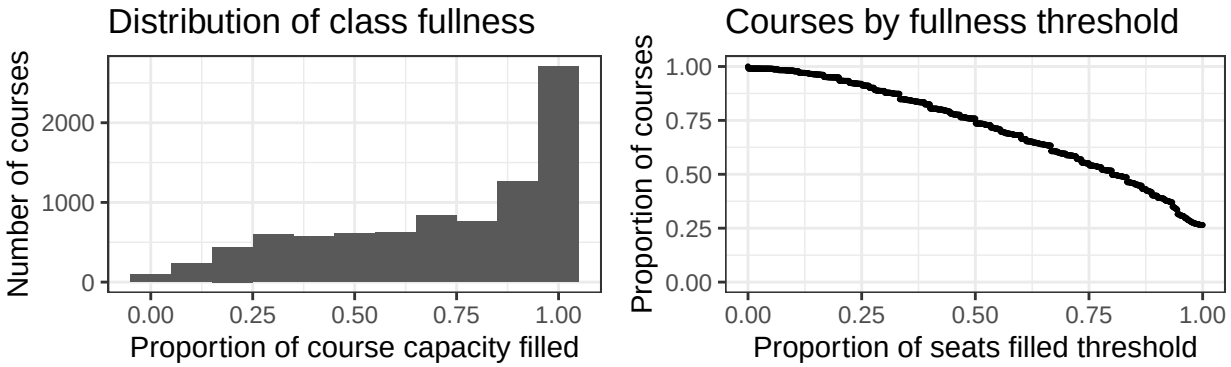
Data was collected from DukeHub, Duke's website for all things from course registration to transcript requests. Although the current form of DukeHub requires login to access information and renders data via iframes, the beta version of DukeHub 2.0 (beta.dukehub.duke.edu) does not require authentication to view courses and uses an API to communicate between the frontend (webpage) and backend (server), simplifying data collection. Thus, data was collected via requests to two endpoints of the DukeHub 2.0 API, the first of which gives a list of courses for a given term and the second of which gives information on the sections of a given course.

Since this analysis is interested in the choices that Duke students make, only those classes which any Duke student can chose to take are used For example, this excludes courses taken abroad, courses at the Duke Marine Lab, and classes only open to Focus program students. Although undergraduates can take graduate courses, only undergraduate courses are included to keep the focus on the target population. Additionally, courses created so that individual students can get credit for 1-on-1 work with a professor are excluded from the analysis. For example, *Introduction to Jazz* is considered but *Jazz Guitar*, the course representing individual private lessons in jazz guitar with a music professor, is not. Finally, to preserve independence, each course is counted exactly once using its primary department, but a record is kept of the number of crosslistings.

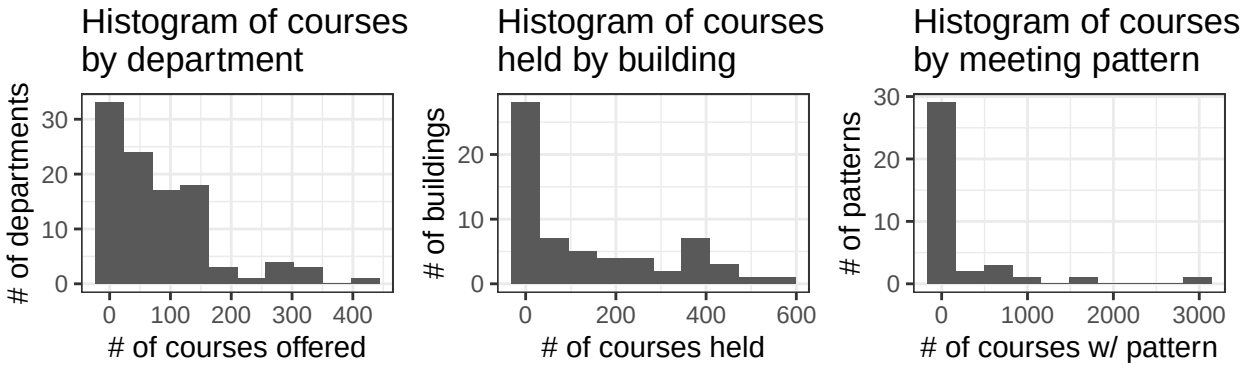
The resulting dataset includes 8999 courses over 11 semesters (2014 Fall to 2019 Fall). Data for the upcoming Spring 2020 term is available but is not used since registration numbers may be inflated before the end of the drop/add period. Information on meeting location/time is missing for some courses but, since this represents only 2.9% of records, these courses are simply removed.

EDA

We begin with an exploratory data analysis.



The plot on the left shows a histogram of class fullness and the plot on the right shows the proportion of courses which meet a given fullness threshold for each threshold. Although 75.8% of courses are at least half full, only 26.5% of courses are completely full. In general, the trend between fullness threshold and number of courses which meet the threshold appears to be non-linear.



As can be seen in the histograms above, the distributions for department, building, and meeting pattern are all heavily skewed right and include many small categories. For example, 23 out of 100 departments offered fewer than 15 courses (as the primary department) over the span of the 11 semesters under consideration. The prevalence of these small categories suggests that use of partial-pooling may be necessary. Additionally, some categories, such as in the Rubenstein Arts Center building and Urdu department, only exist in later years. Partial-pooling would also allow for prediction on such new categories.

Model

Although the data includes information on the proportion of seats which are full for each course, we model whether or not a course fills its capacity (or exceeds its capacity). Based on the EDA, partial-pooling is used for department, building, and meeting pattern with all other variables being fixed effects.

Since prediction is one the goals of this analysis, model selection was performed based on out-of-sample predictive power. Models were trained on the first 10 semesters of data, and tested on the 11th. Model performance was compared based on the area under the receiver operating characteristic curve (AUC), a metric which quantifies the distinguishing ability of a model as its prediction threshold is varied. Out-of-sample AUC results for a selection of models tried are shown in the table below.

	AUC
All variables	0.7324
Without buildings	0.7311
Without depts.	0.6751
Without meetings	0.7327
Without meetings, with dept. specific semester coefficient	0.7352
Without meetings & Pratt, with dept. specific semester coefficient & quadratic start time	0.7357

Based on the model selection process, the final model (shown in the last row of the table) is:

$$\text{logit}\left[\Pr(\text{full}_i = 1)\right] = \beta_0 + \vec{\beta}\cdot\vec{x}_i + \alpha^{dept}_{j[i]} + \beta^{term}_{j[i]}\text{term}_i + \alpha^{building}_{k[i]}$$
$$\alpha^{dept}_j \sim \mathcal{N}(\mu^{dept},\ \sigma^2_{dept})$$
$$\alpha^{building}_j \sim \mathcal{N}(\mu^{building},\ \sigma^2_{building})$$

where $j[i]$ is the index of the primary department for course i, $k[i]$ is the index of the building in which course i takes place, and $\vec{x}_i$ holds term, level, seminar status, number of credits, crosslistings, capacity, whether or

not the course has a discussion and/or a lab, start time, number of weekly meetings, total lecture time, and campus for course i.

Results & Discussion

The fixed effects coefficients for the final model, after retraining on the entire dataset, are as follows:

Variable	Estimate	Std. Error	z value	p-value
(Intercept)	0.6816	1.2280	-1.8662	0.0620
Term	0.9932	1.0123	-0.5586	0.5764
Level: < 100	1.7537	1.1526	3.9561	0.0001
Level: 200-299	0.6825	1.0865	-4.6043	0.0000
Level: 300-399	0.5177	1.0920	-7.4800	0.0000
Level: 400-499	0.3081	1.1182	-10.5404	0.0000
Seminar	1.3506	1.0726	4.2878	0.0000
Credits: 0	2.6244	1.9270	1.4710	0.1413
Credits: 0.5	0.0311	2.0526	-4.8259	0.0000
Credits: 2	1.9486	1.8621	1.0730	0.2833
Crosslistings	1.1411	1.0276	4.8560	0.0000
Capacity (normalized)	0.7413	1.0477	-6.4276	0.0000
Has discussion	0.5859	1.1422	-4.0194	0.0001
Has lab	0.7832	1.1375	-1.8970	0.0578
Start time (min) (quadratic)	0.2159	14.9403	-0.5670	0.5707
Start time (min) (linear)	0.0354	15.9023	-1.2080	0.2270
Number of meetings	0.7412	1.0650	-4.7561	0.0000
Total lecture time (normalized)	0.9070	1.0399	-2.4947	0.0126
Campus: central	0.7671	1.3298	-0.9300	0.3524
Campus: east	0.7440	1.1155	-2.7055	0.0068
Campus: other	0.6666	1.2007	-2.2172	0.0266

The baseline values are a 100-level, single credit course held on west campus.

The strongest predictors are level (a proxy for difficulty), capacity, and number of crosslistings. Expected probability of fullness decreases as course level goes up. All else held constant, each additional crosslisting is expected to multiply the odds ratio by a factor of 1.1411. Having a lab or a discussion decreases expected probability of fullness whereas seminars are expected to be full more often than non-seminars. Courses which meet fewer times per week and for less total time are expected to fill at higher rates, although it is interesting that the meeting days themselves does not add any additional information (as seen in the model selection). A course being held on West campus is expected to fill with higher probability than a comparable course held elsewhere. All else held equal, the Old Chemistry Building is expected to have the highest probability of fullness and Gross Hall is expected to have the lowest. The Physical Education course is expected to fill with highest probability while a German course is expected to fill with lowest probability.

Visualization

In order to make this analysis more accessible to the general public, a published visualization can be found at ethanjholland.shinyapps.io/dukecoursepopularity/. The visualization uses the above model to compare the expected probability of fullness for two classes, side-by-side. Course information can be entered by hand or loaded from past classes. As values are changed, boxes on the right appear which are green when the current settings of a variable favor the modified course (right course) and red if the variable favors the original course (left course). This tool can help departments and professors make educated descisions to maximize popularity when determining the details of a course.