Capstone 2 - College Scorecard

Kirsten Regier

1 Introduction

Main question: What are the primary factors that have contributed to the closing of 4-year undergraduate colleges and universities in the US in the last decade?

According to the https://www.statista.com/ data 'Number of higher education institutions in the U.S. 1980-2017' data, between 2012 and 2017, the number of higher ed institutions decreased by roughly 400. It could be useful to have a model that is able to predict colleges and universities that are in danger of closing. Colleges and universities could use information about common features of schools that have closed to examine their own operational systems, particularly as they face financial difficulties from the coronavirus, which compound the previously existing demographic trends that challenge college enrollment. Insight into which factors that have contributed most significantly in the closing of other colleges could allow current college administrators to make informed decisions about how to protect the future of their institution.

2 Approach

2.1 The Data

The data was gathered from the US Department of Education's College Scorecard, which is available from https://collegescorecard.ed.gov/data/. The database contains information about schools participate in the Title IV program and accept or distribute federal financial aid to students. Annual data is available from 1996 to 2013; the current analysis focusses on data from 2010 and beyond.

The database includes information about the institution, its finances, federal financial aid and student debt, completion and retention rates, and student demographics. A subset of 15 features were selected for analysis, and narrowed further based on missing data and preliminary modeling. The target variable was a boolean variable that indicates whether the school is currently operating.

2.2 EDA

Several of the basic features that distinguish different types of schools are their governance structure and the predominant degree offered by the institution. The distribution of the schools by these variables are shown in Table **??**. A majority (60%) of the schools in the dataset are **private for-profit** schools, and most of these are primarily certificate granting institutions. There are approximately equal numbers of public schools and private non-profit schools, though they differ in the types of degrees offered. The majority of the **private non-profit** schools offer bachelor's degrees or higher, while more than half of the **public** schools offer primarily certificates and associate's degrees.

The other variable of interest in Table **??** is whether the schools is currently operating or not in a particular year. The data in Table **??** includes all schools coded as **not** currently operating in the database from 2010 - 2013, while the currently operating schools are taken only from the 2013 data. Given the prominence of private for-profit, primarily certificate granting institutions in the data overall, it is not surprising that the majority of the closed schools (57%) are also from this category.

		Governance structure							
		Public		Private nonprofit		Private for-profit		Total	
	Operating	No	Yes	No	Yes	No	Yes	No	Yes
Degree	Not classified	27	75	35	89	191	309	253	473
	Certificate	117	594	147	199	1577	2300	1841	3093
	Associate	42	780	26	163	257	545	325	1488
	Bachelor	13	588	94	1241	134	271	241	2100
	Graduate	1	15	40	236	58	35	99	286
	Total	200	2052	342	1928	2217	3460	2759	7440

Table 1: Distribution of schools by predominant degree awarded, governance structure, and whether they are currently operating.

The boxplots in Figure **??** show the differences in undergraduate student enrollment (left) and cost (right) by governance structure and whether the schools are currently operating. Currently operating public schools have the highest median enrollment and the lowest cost. Private non-profit schools have the highest median cost and relatively small enrollment numbers, though there are private non-profit schools with relatively large enrollments. Half of the private for-profit schools have enrollment of less than 200 students, but the four schools with the enrollment over 100,000 students are also private for-profit schools. Likewise, while the mean cost of the private for-profit schools is lower than that of private nonprofit schools, the schools with the highest cost are also private for-profit schools.

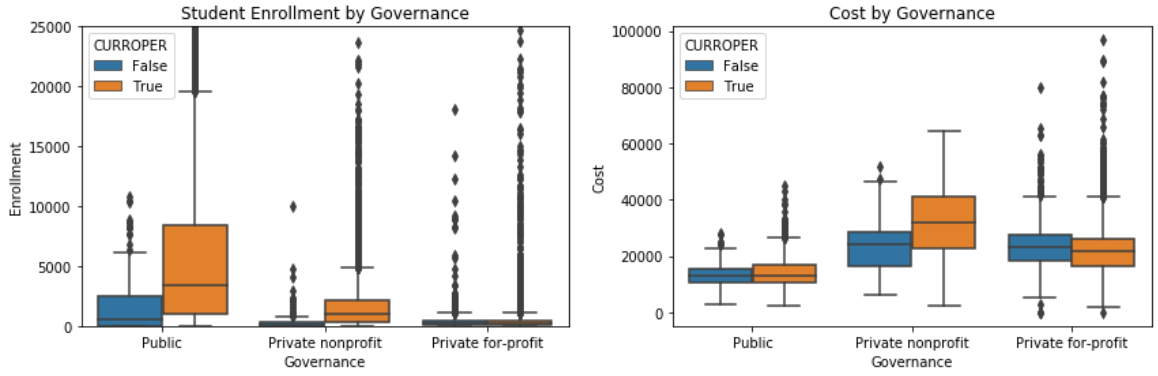

Figure 1: Enrollment (left) and cost (right) of schools by governance structure and whether they are currently operating (CURROPER). In general, public schools are larger and less expensive than private schools. Private non-profit schools are slightly larger and more expensive than private for-profit schools, though private for-profit schools show more variation in enrollment and cost than private nonprofit schools.

In the swarmplot in Figure **??**, the cost of currently operating schools is arranged by predominant degree offered and governance structure. This graph highlights that public and private for-profit schools offer mostly certificates and associates degrees, while most private nonprofit schools offer bachelor's degrees (or higher). One reason that private nonprofit schools have the highest median cost could be that they offer higher degrees than most of the public and private for-profit schools. It is also interesting to note that public schools (blue) have lower cost than

private for-profit schools (green) regardless of the degree type.

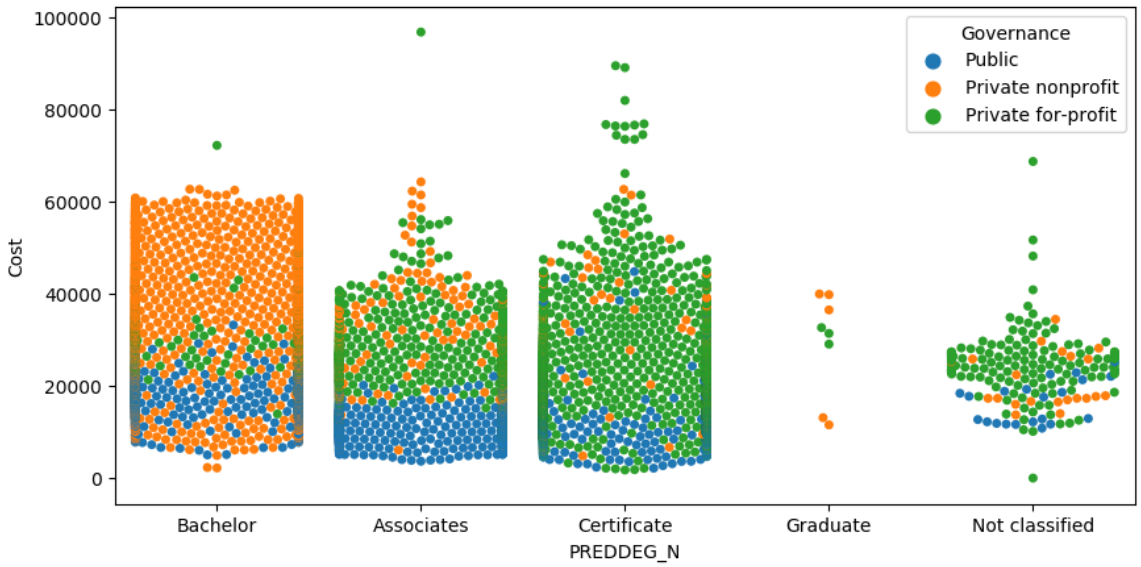

Figure 2: Cost of currently operating schools by predominant degree offered and governance structure. Private nonprofit schools have higher costs in general, but offer higher degrees than public and private for-profit schools. Public schools generally have lower costs in all degree categories.

While student cost and student debt are important financial metrics for students to consider when choosing a school, other financial measures may be informative when considering the financial health of an institution. Several of these features include a school's net tuition revenue per full-time equivalent (FTE) student and instructional expenditure per FTE student. The net tuition revenue per FTE is calculated by subtracting all discounts from tuition revenue, and dividing by the total number of FTE students. The instructional expenditure per FTE is calculated by dividing the instructional expenditure by the number of FTE students. Another potentially important factor could be faculty pay, which is reported in the database as the average monthly faculty salary.

Figure **??** shows heatmaps of the Pearson correlation coefficients between various financial metrics and enrollment for institutions with each type of governance structure. Overall, private for-profit schools show the weakest correlations between most of the financial variables, with the exception of a very strong correlation (.81) between tuition revenue and instructional expenditure. While public and private nonprofit schools do show fairly strong correlations between tuition revenue and instructional expenditure (0.49 and 0.6, respectively), they are less strong than that of private for-profit schools. The most highly correlated features for public schools are cost and debt, while for private nonprofit schools, the highest correlation is between cost and faculty salary, followed closely by cost and tuition revenue.

2.3 Other graphs to include

- Students older than 25
- Completed (withdrawn, still enrolled)

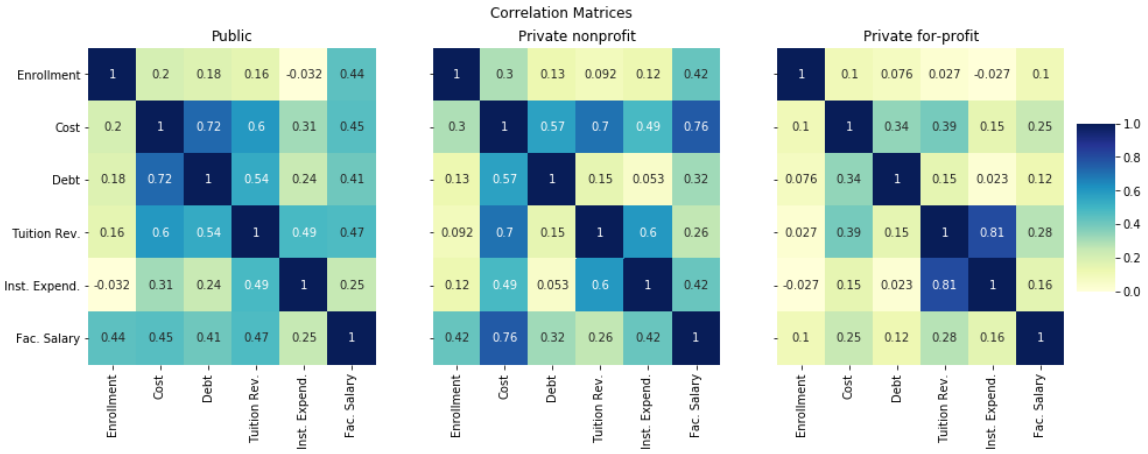

Figure 3: Pearson correlation matrices for various financial metrics by governance type. Darker colors indicate stronger correlations.

2.4 Data manipulation

In any given year, there are many fewer schools that are not currently operating than schools that are operating, which led to an unbalanced dataset from the perspective of the target variable. Several methods were explored for balancing the data.

Schools that were not currently operating were extracted from all years (2010 - 2013), while currently operating schools were selected only for the most recent year (2013). This yielded 2759 observations for closed schools and 7440 observation for currently operating schools.

The resampling method that yielded the best balance of overall accuracy and precision & recall of the closed schools doubled the number of closed schools via random upsampling (with replacement). This resulted in 5518 observations for closed schools, and a total of 12958 observations.

2.5 Modeling

The final model included two supervised ML models in series - an initial Decision Tree model that formed the basis for an AdaBoost model.

3 Findings

General findings of the models

- Important features (of different model options) -

Insert feature importance graph Decision tree important feature - UGDS (schools with larger enrollments are less likely to be closed), CONTROL (40 % of schools in database are private for profit schools, and X% of the closed schools are private for profit schools.

AdaBoost important features - none really stand out a having a huge effect; they all have some effect, but there is not one that seems to be overly important.

It would be very interesting to see which schools have actually closed since 2013, which should help refine the model and improve the accuracy of the predictions.

3.1 3 Concrete Recommendations for client

The Council of Christian Colleges and Universities (CCCU) provides data, support and insight for 180 member colleges and universities. There are several ways that they could use this model to serve their member institutions.

1. Recommendation one - Look at the schools in the test set that were "false negative" results - the ones that the model predicted to be closed, but are actually currently operating. Are there any uniting characteristics, or any special circumstances or institutional changes not accounted for in the model to explain why the schools are still operating, or why the model made the wrong prediction?

   For example, the model based on the 2013 data predicted that 7 private nonprofit primarily bachelors institutions would be closed. An internet search for those schools showed that they are all currently operating in 2020. It would be interesting to learn more about the individual schools, and see if they were close to closure at any point in the last 7 years, and what changes, if any, they have made recently that would influence their institutional organization or financial situation.

2. Recommendation two - Run each of the member schools through the model and see what it predicts. If the model predicts that the school is closed, it could be a warning to school administrators that they need to examine their current operational practices, and make strategic changes.

3. Recommendation three - In addition to predicting a binary/boolean value for the schools operating status, the model generates the probability that each school is open (or closed). Even for schools that are predicted to be current operating, it could be insightful to look at the probability associated with the prediction. A school with a 51% probability of being open could be in more danger of closing than one with a 99% probability of being open. Examining the probability of individual client schools could provide useful information to the school's administrators.

4 Future Research

- The most recent data available is from 2013, so there should be five or six years worth of more current data somewhere. It would be interesting to compare the predictions made by the 2013 model to the actual current data, and see how the model and predictions would change if retrained on the more current data. If the goal is of using the model is to help schools weather the financial and economic uncertainty surrounding the coronavirus pandemic, having accurate *current* information about the schools and their financial status would be important.

- While the current scorecard data reports on several institutional financial metrics, like tuition revenue, instructional expenditure, and average faculty salary, other financial metrics may provide useful information about the financial status of the institution, that would help make more accurate predictions. Suggested additional data include:

  - the size of the institutional endowment (if any),
  - the amount of institutional debt, or the proportion of institutional debt to the operating budget,

  - the unfunded discount rate - how much financial aid the school provides students apart from federally funded aid

- Another popular resource for evaluating colleges and universities is the annual US News and World Report rankings. Including information from these rankings, or accessing the data on which the rankings are based, could provide additional metrics and useful features for predicting the closure of institutions.

5 Potential graphs

Figure 4: Cost and Median Debt of schools by control and current operating status. In general, Private schools have higher cost and higher debt than the others. Public schools and for profit schools have similar debt ranges, but public schools have generally lower costs.

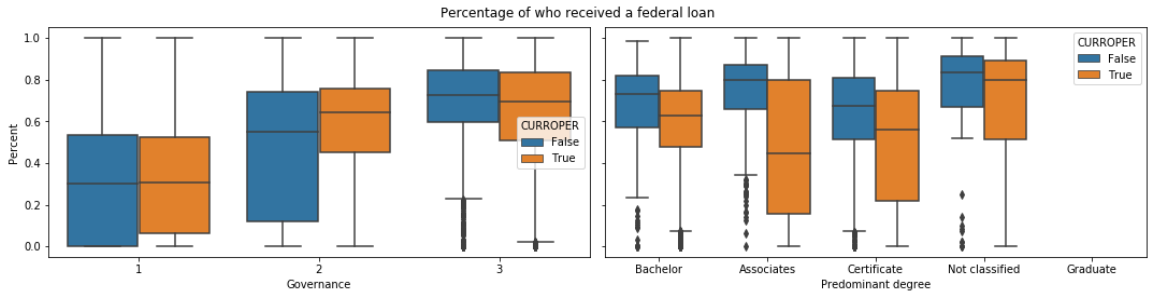

Figure 5: Cost of schools by control and current operating status. In general, Private schools have higher cost and higher debt than the others. Public schools and for profit schools have similar debt ranges, but public schools have generally lower costs.

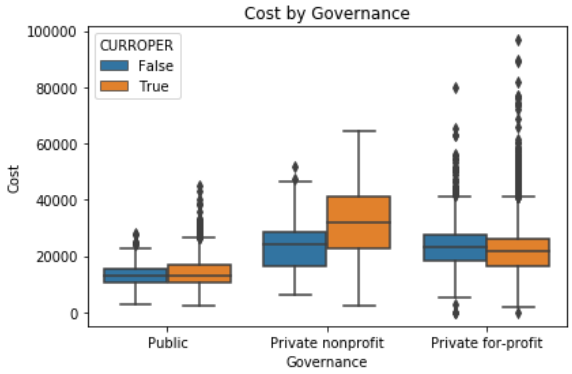

Figure 6: Cost of schools by control and current operating status. In general, Private schools have higher cost and higher debt than the others. Public schools and for profit schools have similar debt ranges, but public schools have generally lower costs.

	Median debt	Tuition revenue	Instructional expenditure	Faculty salary
Public	0.6754	0.4353	0.1642	0.4304
Nonprofit	0.5181	0.2204	0.1564	0.5098
For-profiit	0.2952	0.2257	0.0814	0.2391

Table 2: Pearson correlation coefficients for Cost and other financial variables, by governance type.

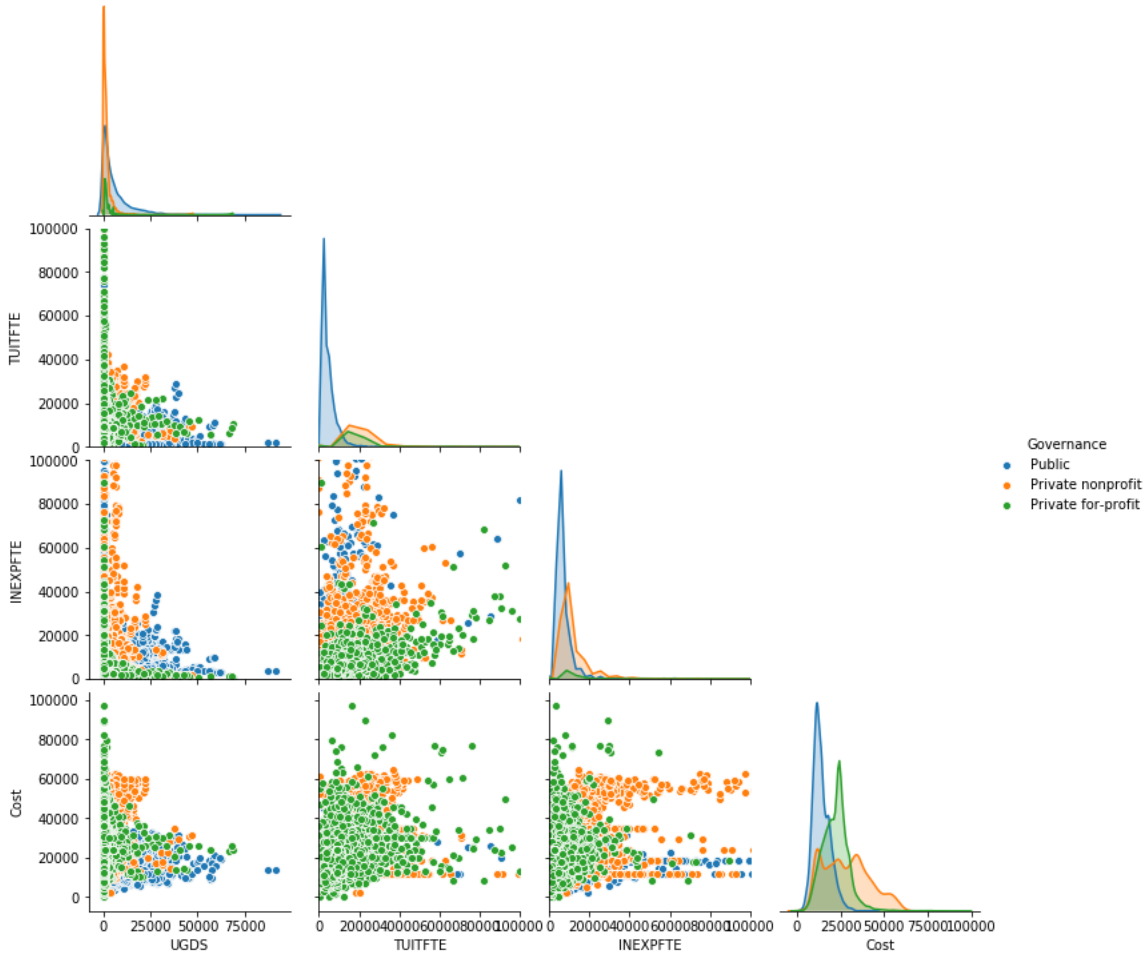

Figure 7: Pairplots of enrollment (UGDS), tuition revenue (TUITFTE), instructional expenditure (INEXPFTE), and cost by governance structure. The diagonal plots show the kernel density estimates for each feature by governance structure.

Enrollment and Cost by Financial factors