Capstone 2 - College Scorecard Data

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 12 features was selected for the final analysis. The target variable was a boolean variable that indicates whether the school is currently operating.

2.2 EDA

2.2.1 Governance structure and Predominant degree offered

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 1. 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 1 is whether the schools is currently operating or not in a particular year. The data in Table 1 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	272	241	2101
	Graduate	1	15	40	236	58	35	99	286
	Total	200	2052	342	1928	2217	3461	2759	7441

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

2.2.2 Enrollment and Cost

The boxplots in Figure 1 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 nonprofit 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 only school with the enrollment over 100,000 students is also a private for-profit school. Likewise, while the median 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.

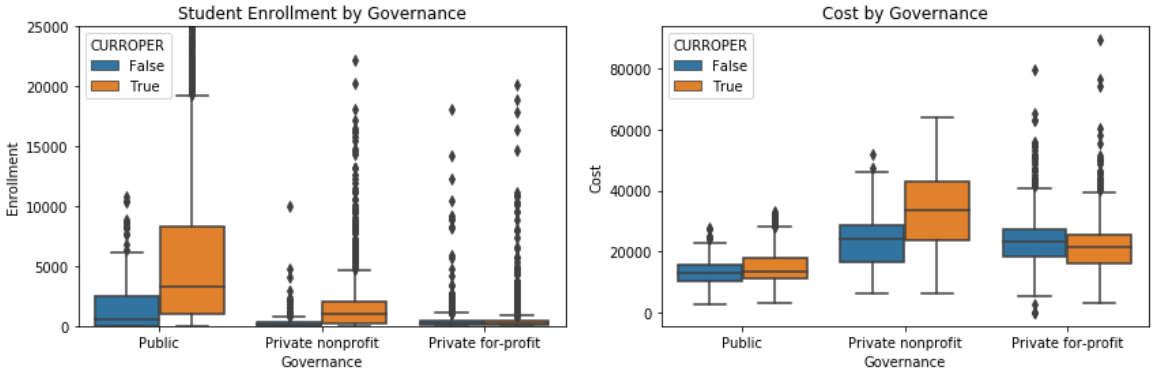

Figure 1: Enrollment (left) and cost (right) of schools by governance structure and whether they are currently operating (CURROPER). In general, public schools are larger and less expensive than private schools. Private non-profit schools are slightly larger and more expensive than private for-profit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

In the swarmplot in Figure 2, 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.

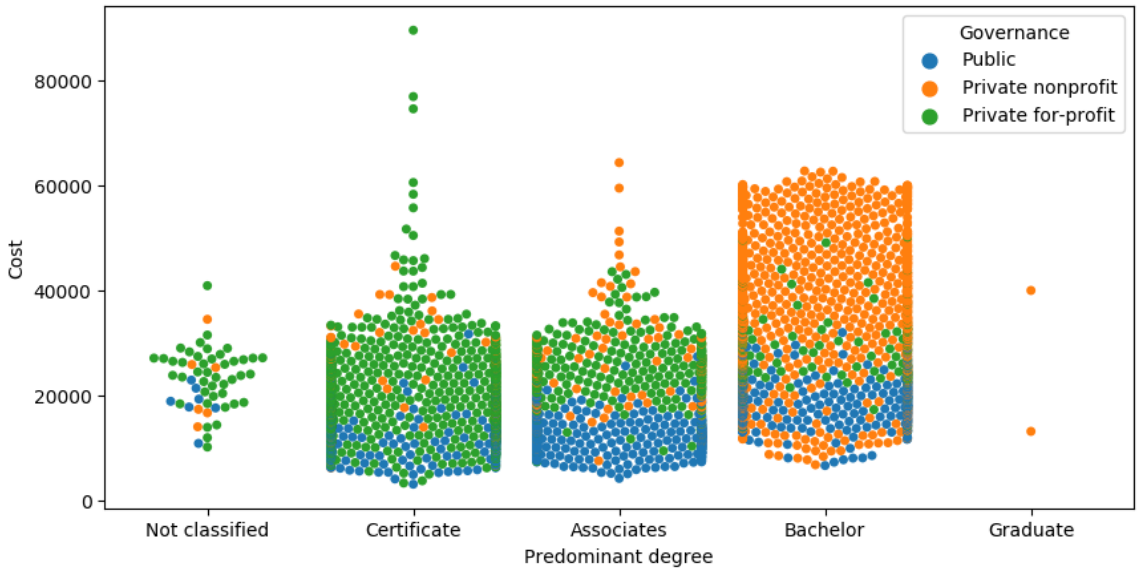

Figure 2: Cost of currently operating schools by predominant degree offered and governance structure. Private nonprofit schools have higher costs in general, but offer higher degrees than public and private for-profit schools. Public schools generally have lower costs in all degree categories.

2.2.3 Other financial metrics

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 3 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. 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.

2.2.4 Student characteristics

While the original plan was to include a variety of information about student demographics, like average SAT/ACT scores, much of this data was either missing or suppressed in the original data.

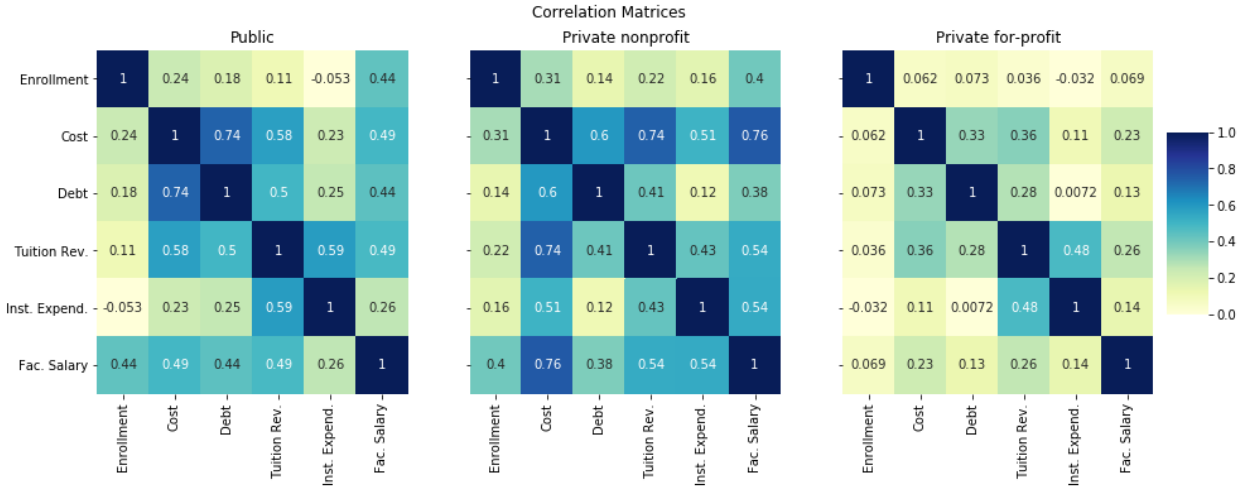

Figure 3: Pearson correlation matrices for various financial metrics by governance type. Darker colors indicate stronger (positive) correlations.

As a result, the only student characteristic that was maintained was the percentage of the student population over the age of 25. As shown in left panel of Figure 4, private for-profit schools have the highest percentage of students over age 25, while private nonprofit schools tend to have a slightly lower median percentage than public schools. When divided by predominant degree type (right panel), institutions that offer predominantly bachelor's degree have the youngest populations, and graduate schools have the oldest populations. Since most private nonprofit schools also offer predominantly bachelor's degrees, it is not surprising that these two categories have similarly low percentages of students over age 25. Interestingly, schools that are not currently operating tend to have higher percentages of students over age 25 than currently operating schools.

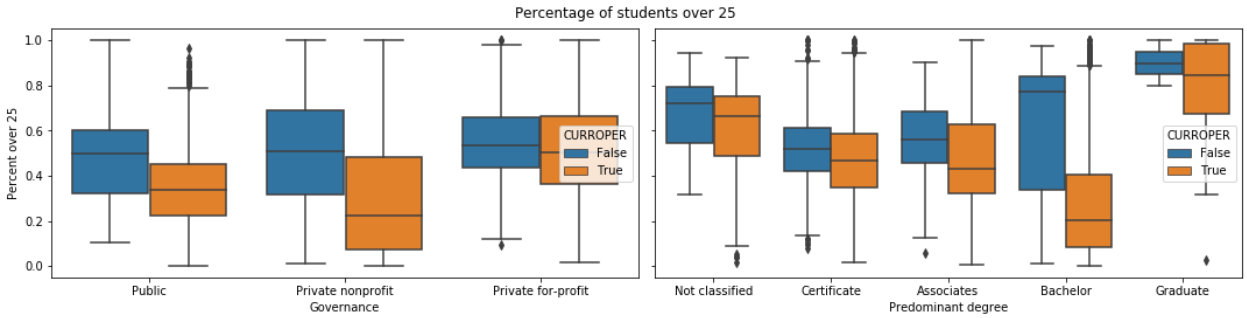

Figure 4: Boxplots of the percentage of students over age 25 by school governance structure (left) and predominant degree offered (right).

2.3 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

(CURROPER). 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 7441 observations 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 12959 observations.

2.4 Modeling

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

2.4.1 Decision Tree

A grid search was performed to find the optimal hyperparameters (tree depth and node purity method) for the initial decision tree model. The optimal depth was set to 22 and the node purity method was 'gini'. With the threshold at 50% for assigning boolean predictions to the target variable (Currently operating), the overall accuracy for the initial Decision Tree model was 87%, with precision of 80% and recall of 93% for the 'Not currently operating' class, as shown in Table 3.

Table 2: Confusion matrices and Evaluation metrics

Table 3: Decision Tree Model

Currently operating		Predicted		Recall
		No	Yes	
Actual	No	1533	123	0.93
	Yes	380	1852	0.83
Precision		0.80	0.94	
Accuracy				0.87

Table 4: AdaBoost Model

Currently operating		Predicted		Recall
		No	Yes	
Actual	No	1538	118	0.93
	Yes	195	2037	0.91
Precision		0.89	0.95	
Accuracy				0.92

2.4.2 AdaBoost Model

After the initial Decision Tree model was trained, it formed the base for an AdaBoost model. Again, the hyperparameter of the AdaBoost model (n-estimators = 122) was chosen empirically. The AdaBoost algorithm improves model performance by focussing on mis-classified samples and adjusting the model weights to improve their classification. The final AdaBoost model had an overall accuracy of 92% with a precision of 89% and recall of 93% for the 'Not currently operating' class, as shown in Table 4.

3 Findings

3.1 Feature weights

The weights of different features for the each of the models are given in Figure 5. The initial Decision Tree model results (left) suggest that the most important feature in predicting whether a school is currently operating is undergraduate enrollment (UGDS), suggesting that the more students a school has, the more likely it is to be operating. The next two features are the governance

structure (CONTROL) and the percentage of students over age 25 (UG25abv). The percentage of students over age 25 is a continuous, numerical variable, which means that schools with higher percentages of older students are more likely to be operating. The governance structure (CONTROL) is a categorical variable. In this case, the model suggests that private for-profit schools have a highest probability of being open, followed by private nonprofit schools, with public schools having the lowest probability of being open.

The least important features are the predominant degree (PREDDEG) and the percent of faculty who are employed full time (PFTFAC). While the other features have weights greater than zero, which suggests that they do have predictive power in the model, there are not clear divisions between features with lesser or greater importance.

The only feature that was important in the AdaBoost model was governance structure (CONTROL).

Figure 5: Feature importance levels based on the Decision Tree model (left) and the AdaBoost model (right).

3.2 Recommendations

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

1. 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 final model based on the 2013 test set predicted that 3 private nonprofit bachelor's institutions would be closed. An internet search for those schools showed that 2 of them are still operating in 2020, and one closed in 2016. It would be interesting to do more research about the closed school, to see if and how its impending closure were represented in the data, particularly if there were financial factors that were not included in the model. It would also be interesting to learn more about the operating 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. Run each of the member schools through the model and examine the predictions. 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.

   For example, because I know of several CCCU member institutions with the name 'Bethel', I ran all of the schools whose name starts with 'bethel' through the model. While all eleven of them were correctly predicted, for both currently operating and not currently operating schools, it is likely that some of the 180+ currently operating CCCU member institutions would be predicted to be closed.

3. In addition to predicting a boolean value for the schools operating status, the model generates the probability that each school is open (or closed). The final AdaBoost model used a threshold of 50% for assigning a school as operating or not. Adjusting this threshold would change the predictions for various schools, especially those close to the threshold boundary.

   Even without adjusting the threshold used by the model, the probabilities assigned by the model could be useful. 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.

   For example, for the eleven 'Bethel' schools referenced above, all the correctly predicted, currently operating schools had over a 99% predicted probability of being open, while the correctly predicted, closed schools had a predicted probability of less than 3% of being open.

   There were eleven private nonprofit schools with a predicted probability between 50-60%. Of these schools, four were predicted to be open, but were actually closed (false positive responses). Of the schools that were currently operating in 2013, five appear to be still operating in 2020, and two appear to have either been renamed or merged with other schools, and are thus not currently operating under the same name.

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 and instructional expenditure, 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.