Effect of Economic Recessions on Educational Activity

Blake Stanford <bstanfor@bu.edu>, Vincent Wahl <vinwah@bu.edu>, Chenyang Yu <cyu1221@bu.edu>

1. Project Objective

The goal of our project is to find out how economic recessions affect educational activity. To determine this, we will look at how schools respond to economic recessions as well as how students (or prospective students) respond to economic recessions. To give a more insightful analysis of this, we also have to look at how state and federal government educational spendings change in response to economic recessions. Specifically, in part due to the availability of data, we will be analyzing the 2008 recession, with the goal of answering the following questions: How do schools respond to economic recessions? Which schools respond similarly in economic recessions? How do students respond to recessions?

Our initial hypothesis was that economic recessions would lead to a decrease in educational activity because students would not be willing to or would be unable to cover the cost of going to university. However, after researching these questions, we have found that because of financial stimulus packages to public schools and increased financial aid, the hypothesized relationship is more complicated than what we would expect. Furthermore, the relationship varies among schools. These results are further described in section 6 of this report.

Our project repository can be found at: https://github.com/cyu1221/CS506

2. Data sets

The primary data sources used in this project are: The National Center for Education Statistics (NCES), Yahoo! Finance (Yahoo)[7], the Federal Reserve Bank of St. Louis (FRED), U.S. News & World Report – National University Rankings, and Google's GeoCoding API. Yahoo is used to retrieve data on four major stock indices and bond yields. The stock indices initially used in this project were: Dow Jones Industrial Average (DJIA), Standard & Poor's 500 (SNP), NASDAQ (NSD), and Russell 2000 (RUT). After further consideration, we decided to only use DJIA as it is comprised of stocks from major companies in the US, better reflecting the overall economy compared to the other indices which are comprised of smaller companies or a subset of stocks conforming to some selection rule. Although NSD represent all stocks traded at Nasdaq, the index is heavily weighted by performance of tech companies as the stock exchange has many major and minor tech companies traded there. Since NSD is so heavily affected by this, we chose to use DJIA over NSD. To reflect bond yields, the CBOE 10-year US treasury yields (TNX) will be used. Bond yields, representative of the real interest rate, help provide insights into financial aid and the borrowing environment. The St. Louis Federal Reserve has provided the unemployment statistics. These statistics were used in order to gain a basic understanding of the characteristics of a recession.

Yahoo's financial data makes available much more specific data than needed, both intra and inter day, but we have chosen to only extract the adjusted closing price of each week for the period 1992-2018, since we are only looking for extended periods of depressions in the market. For unemployment, the data available is sampled at a monthly frequency.

NCES is a part of the United States Department of Education, and it provides per institution, yearly, educational statistics. The data of interest from NCES are number of newly enrolled students, number of newly graduated students, amount of financial aid, amount of student loans, resources allocated to research, and the schools addresses. The street addresses are used to retrieve the geographical location (longitude and latitude) of each institution using google's GeoLocation API.

The data retrieved from NCES was highly variable in terms of quality and completeness – many schools had only reported on some of the variables we wanted to look at or did not report each year. To account for this, we have chosen to only look at schools that has consistently reported data on our selected attributes over the period 2001 and 2016. In total, we have data collected from 2173 post-secondary schools.

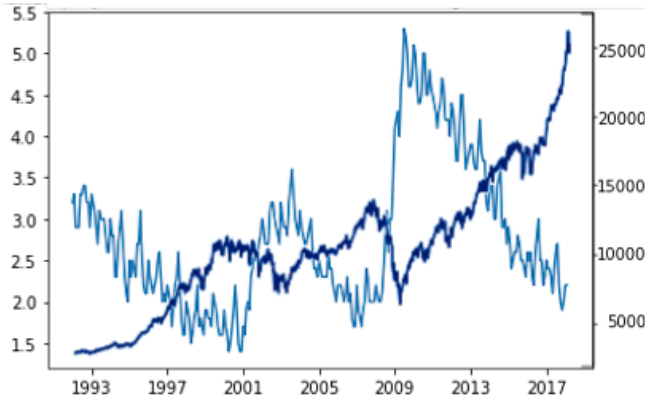
The national universities ranking data, together with the educational data, offers insights into whether rankings affects schools' and students response during economic recession. We use the 2017 National University Rankings dataset, which was extracted from U.S. News & World Report. The U.S. News & World Report publishes annual university rankings, comparing the academic quality of U.S.-based schools. The ranking of a school is based on factors like academic excellence, graduation rates, faculty resources and freshman retention rates. We chose National Universities category because it offers a full range of undergraduate majors, plus master's and doctoral programs.

3. Initial Data Analysis

Financial Data:

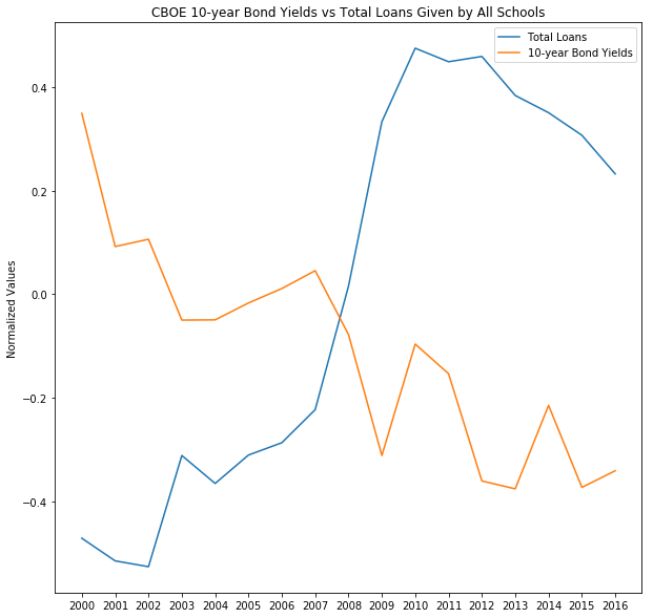
The financial data was used initially to determine periods of economic recessions and to explore the relationship between the amount of student loans given and interest rates.

Looking at the daily closing price of the DJIA and the unemployment data, both graphed below, we can see that the two are heavily negatively correlated. This is something we would expect given that unemployment increases in economic recessions and decreases in stable and strong economic periods. Furthermore, we can see dramatic changes to both DJIA and unemployment in the periods 2001-2002 and 2008-2009.



Closing Price of DJIA (dark blue, right side scale)
US Unemployment Rate (light blue, left side scale)

In order to explore the relationship between bond yields and the total amount of student loans given by year, we first ran some simple statistical tests on the datasets.



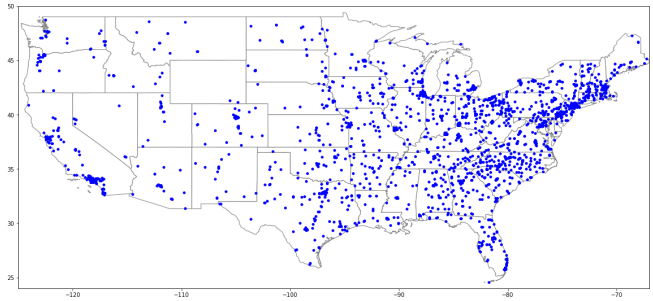
Normalized aggregate sum of all students loans (blue) and normalized CBOE 10-year bond yields (orange)

Firstly, we found that the normalized bond yields and normalized total amount loaned to students are strongly negatively correlated, having a correlation coefficient of -0.8264. This can be confirmed visually in the chart above and is as expected: A decrease in bond yields – representative of a decrease in the real interest rate – should logically lead to an increase in the amount loaned to students, as the cost of borrowing is reduced with lower interest rates.

With these preliminary results, we can make some conjectures about how financial aid should change given a recession. We would expect that financial aid would generally decrease given a stressful economic environment.

Educational Data:

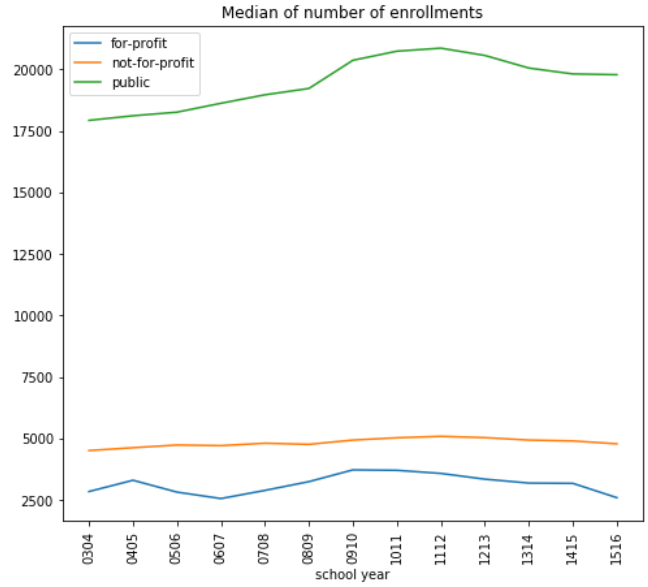
Initially we had data covering more than 4000 schools in the US. However, after trying to join each retrieved variable with their University ID, we saw that many of the schools had poor consistency in reporting statistics to the NCES. To resolve this, we decided to only look at the schools that were consistent. This resulted in a data set that covered 2173 schools. On the map below, displaying the schools, we can see that the most schools are located on the eastern parts of the US and with some dense areas in california (San Francisco and Los Angeles).



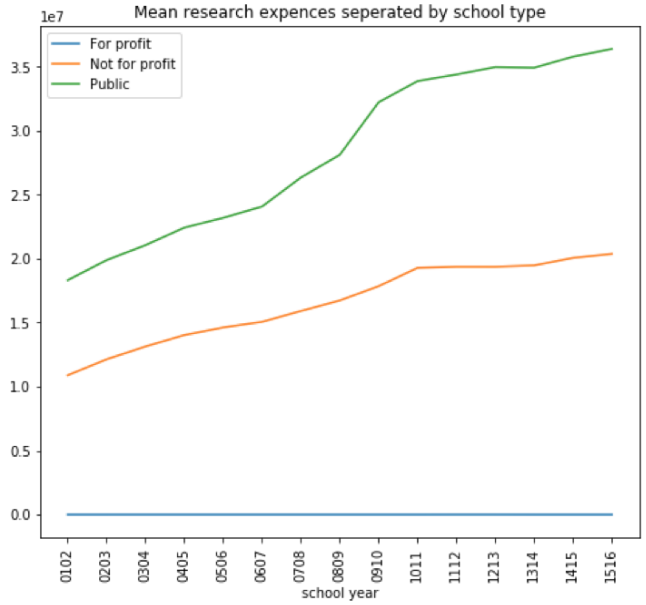
Map of The U.S. with blue plots indicating geographic location of each school in our data set

From the National Universities rankings dataset, we tried to retrieve the ranking of each of the schools for 2016. We decided to only use ranking from one year, as the rankings rarely changes by much from year to year, and we are primarily using the ranking as an indication of prestigious / unprestigious. Unfortunately, we were only able to retrieve rankings for the 180 best ranked schools. To resolve the rankings for the remaining schools, we decided to give each school that was not ranked a ranking of 181. This was chosen because we know that they were not ranked among the top 180, but couldn't conclude a specific ranking. Thus, each non-ranked school was assumed to be ranked the same.

To get an initial understanding of our data, we started by plotting the aggregate values over all schools for each year for each variable.



Median number of number enrollments across all schools of same type for each year.



Median research expenses across all schools of same type for each year.

Above, you can see the graphs of the median number of enrollments from 2003-2015 and mean research expenses from 2001 to 2015. By exploring the two graphs, we can see that both enrollments and schools allocation to research increases during the period of 2008-09. This was a result that initially seemed to conflict with our hypothesis. However, in both graphs we see that public schools are clearly dominating this increase. In terms of the increase in enrollment, this could be due to the fact that students choose to go to public schools with lower tuition fees, as we will further explore in section 5. After further research into the federal response to the 2008 financial crisis, we discovered that the stimulus packages put together by the US in response to the 2008 recession included an injection of funds to the education sector. This stimulus package is a likely cause for the increase in research allocations across public and not-for-profit schools.

4. Algorithm

Clustering of Schools and Students by similar behavior:

To make our variables capture how schools and students respond to financial recessions, we decided to project the data into percentage change from the previous year. We use the change from the previous year to capture how the behavior changes when entering an experiencing a recession. Percentage

change is used to allow for a reasonable comparison between schools of different sizes.

To cluster schools on similar behavior, we made a vector representation of each school on the variables in question. And then applied a Gaussian Mixture Model (GMM) to all the schools. GMM is a probabilistic unsupervised learning model that can divide a dataset into normally distributed subsets that are similar in terms of some metric. To achieve this, GMM defines k normal distributions, each with a mean μ and a variance Σ, in the euclidean-space of the data points. The model is then optimized to try to maximize the probability of a datapoint x given the k distributions. Given that we do not know how to assign points to clusters initially and both μ and Σ for each is unknown, an expectation maximization algorithm is used to estimate the models parameters. This method first assigns points in a simple way, or randomly, to clusters, then steps back and forth between optimizing the probability of a cluster assignment to a data point given the data point and maximizing the likelihood that a sample from the distributions is a point assigned to the cluster with that distribution.

To determine the number of clusters to use, the number k in the description of GMM above, we clustered schools for all k between 2 and 12, inclusive. Then we calculated the silhouette score for each of the points in the 12 clusterings, and picked the clustering with the highest average score. The score for each point is determined by:

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

Where s(x) is the score for a datapoint x. b(x) is defined as the minimum, for all d, in: average distance from x, assigned to cluster c, to all points in cluster d, where d is a different cluster than c. a(x) is defined as the average distance from x to all other points in the same cluster. s(x) is then, more intuitively, the average distance from x to all points in the “next best” cluster minus average distance from x to points in its own cluster, divided by the larger average distance of the two. The silhouette score ranges from -1 to 1 where a high value indicates that a datapoint is well matched to its own cluster and poorly matched to other clusters.

In our model we have decided to use euclidean-distance as our metric, and all the data was normalized before GMM was applied. Normalization was done to prevent some of the fields in each datapoint to be dominating just because the range of values taken on by that field is greater compared to others.

We decided to do a Gaussian Mixture Model (GMM) clustering over a subset of the variables that corresponded to the behavior we were looking for in the time periods we want to look at. To cluster student responses to economic recessions, we used the change in the number enrolled, number graduated, and amount of loans, for 2008-2009. To cluster schools’ response to economic recessions we used the change in allocations to research and grants given by the school for the same period.

The clustering of schools is used to better determine how a subset of students or schools behave in response to economic recessions by looking at how the students or schools within the same cluster behave.

Logistic Regression:

Given the results from the above algorithms we were interested to see if we could create a predictive model that would determine whether or not a school could expect an increase or decrease in enrollment. In our first attempt at this, we decided to use a logistic regression to classify each school as either increasing or decreasing in admission. We set up this regression by shuffling our data before setting aside 30% of the data as test data. The remaining 70% of the data was used to train the model.

In logistic regression, the dependent variable is a binary variable that contains data coded as 1, increase in admission, or 0, decrease in admission. The logistic regression model predicts P(Y=1) as a function of X. After implementing logistic model, if the p-value for a coefficient is smaller than 0.05, we consider it to be significant to the model.

We perform cross validation to avoid overfitting and to produce a prediction for each observation dataset. We are using 20-fold Cross-Validation on the entire dataset to train Logistic Regression model. If the average accuracy remains close to the Logistic Regression model accuracy, the model generalizes well. We use this model for future studies. Besides cross validation, confusion matrix is used to describe the performance of a classification model. It contains four test statistics: true positive, true negative, false positive and false negative test result. From the confusion matrix, we are

able to obtain precision, accuracy and F Score, which is a weighted average of the true positive rate and precision.

Artificial Neural Network (ANN) For Binary Classification:

To further improve upon our results in our logistic regression, we implemented an ANN in Keras, trained for binary classification. Our ANN was implemented as a fully connected Feed-Forward neural network, consisting of three hidden layers, first layer with 144 neurons and then for each subsequent layer the number of neurons is 66% of the layer before. The final output layer is one neuron. In the hidden layers, the neurons are implemented with Rectified Linear Unit activation function and in the output layer a sigmoid activation is used. For training, binary cross entropy was used as the loss functions, a dropout rate of 0.4 was used for the input layer, a dropout rate of 0.3 was used after the first and second hidden layer, for optimization the adam optimizer with a learning rate of 10^{-5} was used, batch size was set to 64, number of epochs was 500, and each hidden layer had max norm regularization of 4.

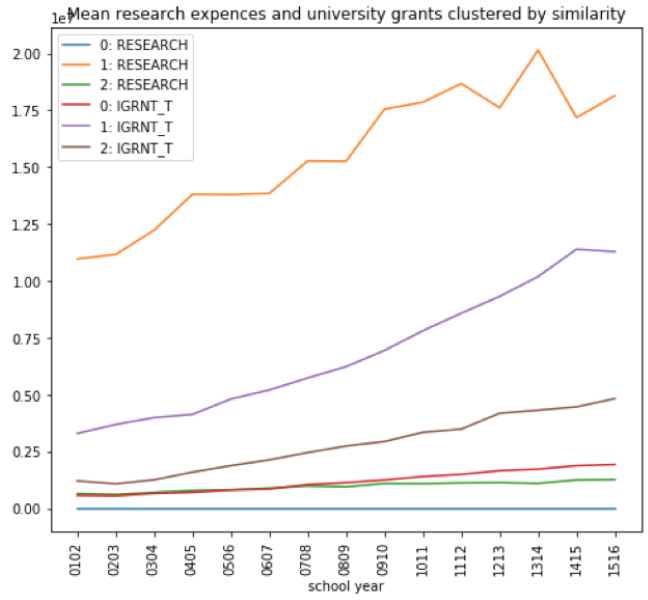
We decided to use max norm and dropout during training to prevent overfitting to the data. The implementation was successful in preventing overfitting, with training error very close to testing error.

To optimize hyperparameters, network topology, and regularization, we initially trained the network on 50% of the data and cross-validated on 20%. After determining how the network would be implemented based on the cross-validation, we trained our final network on the training data and cross-validation data, before testing it on the 30% remaining, untouched, data.

5. Results

Clustering

After clustering schools on similar behavior during the 2008 recession, we got some interesting results. The GMM clustering, together with silhouette score, resulted in the schools being divided into 3 clusters all of which were of similar size. After assigning each school to a cluster, we graphed the attributes for all the years, and found that one of the clusters behaved differently from the others. This group shows a decrease in median amount of funds allocated to research during the period of 2008, which is what we initially expected overall. However, for this group of schools, we see a great variability in mean amount allocated to research, so the significance of this is questionable. Common denominators for schools within this cluster is that they allocate a large amount to research, they give out the most institutional grants, and the cluster consists of very close to 50% public schools, 50% not-for-profit schools, and no for-profit schools. Furthermore, the schools in this cluster are almost exclusively top tier schools. The amount allocated to research for the clusters can be viewed below. The interesting cluster explained above is denoted with the number 1. In this chart, each cluster is represented by two lines. Cluster 1, for example is then represented by the orange and purple lines.

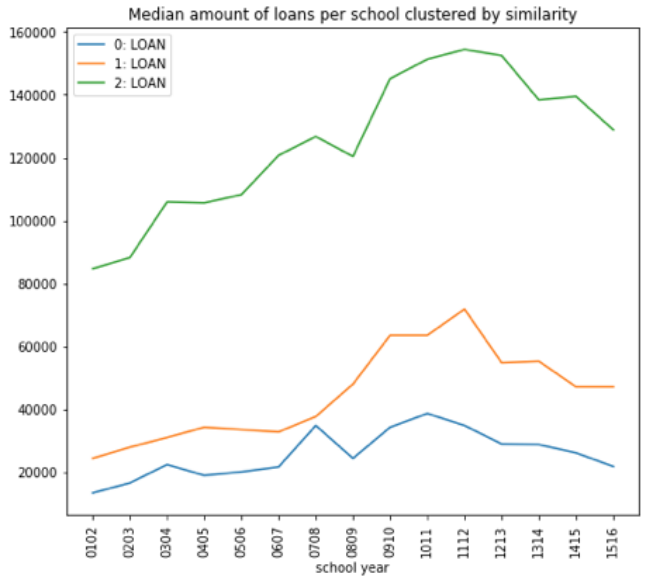


Mean research expenses across all schools with similar changes in student grants and research expenses during recessionary periods.

To address how students, or prospective students, react to economic recessions, we want to use the number of students enrolling, the number students of graduating, and the amount of student loans taken up.

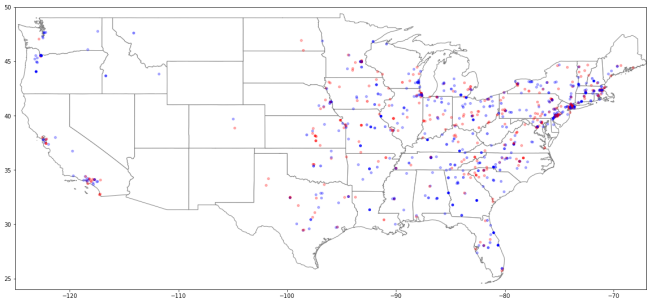
If we group schools by similarity of student behavior during the financial crises, the same way we did for schools' behavior, we get the following results for the different groups. Group 2 is about 4 times the size of the two other groups and Group 0 is the primary group for for-profit schools, while the two others have about equal distribution of public and not-for-profit. Here the ranking of the schools were about equally distributed among the clusters. The results from this clustering did not give much insight into what types of schools we are looking at, but, as one can see on the graph below,

group 1 is primarily responsible for the increase in median loan amount during the financial recession.

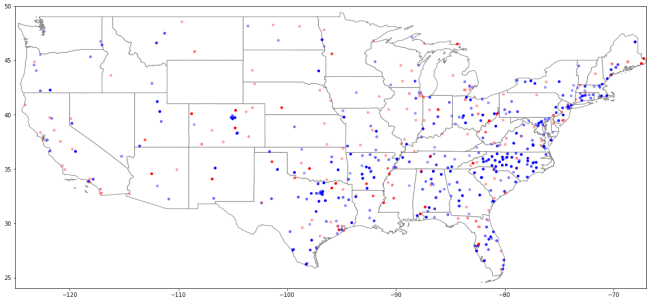


Median amount of student loans across schools with similar change in number enrolling, number graduating, and student loans during recessionary periods

Since it was difficult to see any clear assignment of students behavior to clusters, we tried to plot the geolocation of not-for-profit and public schools with color assigned according to the change in enrollments during the financial recession in 2008. On the plot, blue corresponds to an increase in enrollment and red corresponds to a decrease. The more red/blue a point is, the greater the change was during 2008-2009. This led to some very interesting results:



Change in enrollment during 2008 for not-for-profit schools. Blue indicates an increase and red indicates a decrease



Change in enrollment during 2008 for public schools
Blue indicates an increase and red indicates a decrease

On the plots above, we can see that map for not-for-profit contains a much higher percentage of red points than the map for public schools. Furthermore, we can see that many of the stats that are primarily red in the map for not-for-profit schools, are primarily blue in on the map over public schools. This suggest that students have chosen to attend public schools over not-for-profit schools during the 2008 recession.

To further quantify this, we frame this as a classification problem. Initially we tried a logistic regression model, with input being, for each school, the percentage change from the school year 07-08 to 08-09 for each of our variables except school type which was changed into an indicator vector. All the data was normalized before fitting our logistic regression model to 70% of the data, and then tested on 30%. The model ended up having an accuracy on test data of 67%. Below, you can see a description of the fitted model:

```
Optimization terminated successfully.
         Current function value: 0.571777
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:       dummy_enrollment   No. Observations:                 1106
Model:                         Logit   Df Residuals:                     1092
Method:                          MLE   Df Model:                           13
Date:                Fri, 27 Apr 2018   Pseudo R-squ.:                0.08084
Time:                        17:33:31   Log-Likelihood:               -632.39
converged:                       True   LL-Null:                      -688.00
                                        LLR p-value:                1.071e-17
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Rank          -0.0032      0.002     -1.359      0.174      -0.008       0.001
0809_GR        0.0014      0.002      0.578      0.563      -0.003       0.006
0809_ASSETS    0.0009      0.001      0.775      0.438      -0.001       0.003
0809_EXPENSES  0.0453      0.011      4.034      0.000       0.023       0.067
0809_REVENUE 9.742e-06      0.001      0.011      0.992      -0.002       0.002
0809_RESEARCH  0.0006      0.001      0.702      0.482      -0.001       0.002
0809_ANYAIDN   0.0178      0.005      3.929      0.000       0.009       0.027
0809_LOAN_T   -0.0001   9.64e-05     -1.160      0.246      -0.000    7.71e-05
0809_FGRNT_T   0.0023      0.002      1.458      0.145      -0.001       0.005
0809_SGRNT_T   0.0015      0.001      1.611      0.107      -0.000       0.003
0809_IGRNT_T  -0.0005      0.000     -1.645      0.100      -0.001    9.63e-05
for-profit     0.9271      0.635      1.460      0.144      -0.317       2.171
not-for-profit 0.6935      0.425      1.634      0.102      -0.139       1.525
public         1.4124      0.444      3.178      0.001       0.541       2.283
==============================================================================
```

The fitted model had 3 variables, marked in green, that were determined to be significant. The binary variable ‘public’, whether the school is a public school or not, was one of these with a positive coefficient. This corresponds to what we concluded from the geoplots above, that enrollment to public schools increased during economic recessions. Another important variable was ‘ANYAIDN’, also with a positive coefficient. This variable corresponds to the change in number of students receiving any kind of aid (federal, state or institutional grants, or student loans). This is also something we would expect, given that an increase in number of students receiving aid is something that can be a determining factor in whether a student chooses to go to a school or not. The third variable, also with positive coefficient was

'EXPENSES', which corresponds to the change in expenses by the school. Although not being a variable we expected to be important, it makes sense that students choose to go to schools that are determined to increase their spendings as this might be associated with increase in physical capital, human capital, research, etc.

The performance achieved with this model on the test data is given below:

	precision	recall	f1-score	support
0.0	0.62	0.11	0.19	163
1.0	0.67	0.96	0.79	312
avg / total	0.66	0.67	0.59	475

From the results above we can see that the model has a much higher precision than recall and an overall very low recall rate for predicting a decrease in enrollment, i.e predicting 0. This implies that the model is very selective in predicting a decrease in enrollment. For the the 1 prediction, an increase in enrollment, we see the opposite. This indicates that it frequently predicts 1, getting most of the cases where that is true, but also many cases where it is supposed to be 0. Given that the test data consists of 163 decreases in enrollment and 312 increases in enrollment, and the precision and recall is as indicated above, it is obvious that the prediction model has converged on a state where, if in doubt, it predicts increase in enrollment since in most scenarios this is the case.

Although our model did not have a fantastic classification performance, it still provided insight into important variables in determining change in enrollment.

To try to get a better performance we then applied our ANN as defined in section 4 to all the same variables as in our logistic regression model with the addition of two new variables being the squared of 'ANYADIN' and 'EXPENSES'. The accuracy of our trained model ended up being 71% on the test data. The precision, recall, and F1 score is given below:

	precision	recall	f1-score	support
0	0.53	0.22	0.31	144
1	0.73	0.92	0.81	330
avg / total	0.67	0.70	0.66	474

The results of our ANN model ended up being a little better than what we achieved with the logistic regression. This is something we would expect as the ANN allows for non-linear fitting. Although, we still see the same type of bias towards predicting an increase in enrollment, it is not as bad as before. The recall rate for predicting a decrease in enrollment has drastically increased.

6. Conclusions

The first thing that was immediately clear to us after our analysis is that the relationship between the economy and the education sector is much more complicated than we originally expected. Sources of complication include major variance in the behaviors of individual schools. Attributes such as allocation to research, financial aid given, and the size and demographics of a school's population all make it hard to track the exact changes caused solely by the economy. In particular with financial aid, the 2009 stimulus package in conjunction with the lack of education data from past recessions made it particularly difficult to tease out underlying relationships between the two entities. Despite this, we were still able to make steps towards answering our original questions.

Cluster analysis revealed that some of our initial hypothesized relationships about schools and the economy do indeed hold for at least the cluster with the highest amount allocated to research. These schools experienced the highest volatility in research spending following closely with the volatility of the financial markets with the exception of 2009 where the relationship strays due to the aforementioned stimulus package. Other clusters, however, seem to respond minimally or not at all to economic trends.

In looking at the behavior of students during recessions, we looked primarily at enrollment and loan statistics. It is our opinion that it is highly likely that the large increase in loans taken by students was due to the 2009 stimulus package, combined with very low interest rates, this created a desirable environment for borrowing. We then explored changes in enrollment across all of the schools in our dataset. We found an almost ubiquitous increase in enrollment across the country at public schools, while private schools did not show as strong of a trend. We think that it is possible that this means there was a movement of students originally attending expensive not-for-profit universities to more affordable public universities.

7. References

"^TNX Historical Prices | CBOE Interest Rate 10 Year T No Stock." *Yahoo! Finance*, Yahoo!, 14 Feb. 2018, finance.yahoo.com/quote/^TNX/history?period1=694242000&period2=1518584400&interval=1wk&filter=history&frequency=1wk. Data sampled from 1/1/92 to 2/14/18 at weekly intervals.

"^DJI Historical Prices | Dow Jones Industrial Average Stock." *Yahoo! Finance*, Yahoo!, 29 Apr. 2018, finance.yahoo.com/quote/^DJI/history?period1=694242000&period2=1518584400&interval=1wk&filter=history&frequency=1wk. Data sampled from 1/1/92 to 2/14/18 at weekly intervals

Li, Susan. "Building A Logistic Regression in Python, Step by Step." *Towards Data Science*, Towards Data Science, 29 Sept. 2017, towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8.

"Preprocessing Data¶." *Preprocessing Data - Scikit-Learn 0.19.1 Documentation*, scikit-learn.org/stable/modules/preprocessing.html.

"Sklearn.mixture.GaussianMixture¶." *Sklearn.mixture.GaussianMixture - Scikit-Learn 0.19.1 Documentation*, scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html.

"Sklearn.linear_model.LogisticRegression¶." *Sklearn.linear_model.LogisticRegression - Scikit-Learn 0.19.1 Documentation*, scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

"Sklearn.linear_model.LinearRegression¶." *Sklearn.linear_model.LinearRegression - Scikit-Learn 0.19.1 Documentation*, scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.

"Sklearn.model_selection.cross_val_score¶." *Sklearn.model_selection.cross_val_score - Scikit-Learn 0.19.1 Documentation*, scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html.

"The Sequential Model API." *Sequential - Keras Documentation*, keras.io/models/sequential/.

"Use the Data: Access IPEDS Data Submitted to NCES through Our Data Tools or Download the Data to Conduct Your Research." *The Integrated Postsecondary Education Data System*, National Center for Education Statistics, nces.ed.gov/ipeds/use-the-data.
IPEDS data submitted to NCES was accessed by selecting "survey data" and "complete data files" on the webpage. Under "All Years" and "Graduation Rate" tabs, we downloaded "Graduation rate data, 150% of normal time to complete - cohort year 4-year and 2-year institutions" datasets from 2001 to 2016. Under "All Years" and "12-Month Enrollment" tabs, we downloaded "12-month unduplicated headcount" datasets from 2001 to 2016. Under "All Years" and "Institutional Characteristics" tabs, we downloaded "Directory information" from 2016. Under "All Years" and "Student Financial Aid and Net Price" tabs, we downloaded "Student financial aid and net price" from 2001 to 2016. Under "All Years" and "Finance" tabs, we downloaded "Public institutions - GASB 34/35", "Private not-for-profit institutions or Public institutions using FASB" and "Private for-profit institutions" for Fiscal year 2001 to 2016.