Goodgrant Gives Grant Good Grants

Team #54737

February 1, 2016

1 Introduction

Investing in the education of students is paramount to the prosperity of a nation. It allows for the advancement of science and the production of productive members of society. For some, graduating from college is the ultimate dream. For most first generation children of immigrants, whose parents work tirelessly to give their children the means to live a better life, attending college would be awesome. Given the current upward trend of college tuition, it is becoming more and more difficult for people to attend college.[1]

With the gravity of this issue in mind, it obvious that the system needs a financial lift to help eager students afford college. The Goodgrant Foundation would like to donate $100 million USD to enhance student performance at colleges and universities. As such the Goodgrant Foundation would like to donate its money over a period of five years such to have greatest effect on student performance.

To assist in this decision, data collected by the U.S. National Center on Education Statistics and the College Scorecard data set was distributed. The data consisted of a list of 7804 institutions described by 122 variables. As an indication towards the genre of data there within, the data dictionary included provided a list of categories into which all of the data fell (Table 1).

completion	school	student	admissions	academics
cost	earnings	repayment	aid	root

Table 1: The categories of the data variables, known as "dev-categories".

2 Assumptions

The following assumptions were deemed necessary and appropriate to allow the creation of the models.

1. Each school will use the donation to the best of its ability. It is assumed that there would be no negligence or waste of the donation.

2. All schools considered are a Title IV school. Title IV refers to the institution's status for receiving federal funding. SOURCE: Because some variables pertain only to Title IV schools, it is necessary for this assumption to rely on the data.

3. Schools with less than 300 studdents will be not receive grants due to a large amount of data for smaller schools being interpolated.

3 Data Munging and Cleaning

The data as it was given was in a very unusable state: much of the data was missing, expressed as "NULL", or withheld due to privacy concerns, expressed as "PrivacySuppressed". Thus it was necessary for the data to be cleaned before it could be used in any sort of analysis. The initial matrix contained 7804 schools described by 122 variables.

The first step was to remove the obviously useless variables: metadata stored as strings or other descriptors. These included the OPE IDs, Institute Name, City, State, URL, Net Price Calculator URL, predominant degree awarded, control of the institution (public, private for-profit, or private nonprofit), and the locale. These variables only offered information at a very high level. If the control and locale variables were expanded to flags, it would have been feasible to include them, as the default was expressed as an single integer. It was decided that this information was not relevant to the models.

Now the strings had been removed, the data could be treated as a numeric structure, as the "NULL" and "PrivacySuppressed" values were replaced with NaN (Not a Number). To address the overabundance of missing data, any rows that had more than 50% of data missing were immediately thrown out. Since not enough information about these schools was provided, it is not reasonable that our models would be able to recommend them with any degree of confidence. Furthermore, utilizing one of the given variables, any school that was not currently operating at the time these data was collected will not be considered for any donation. This left 6766 schools for consideration.

Next, each variable was examined to determine its usability. If more than 50% of the schools did not have information for any given variable, then that entire variable was thrown out. Because so much of the data for these variables was missing, it was not feasible to impute the missing values. This left 74 viable variables.

For remainder of the missing data, the existing values were used to impute the missing data. Because the data in the "school" and "academics" dev-categories appeared to express something akin to numeric metadata, these data were used to determine the likeness between schools. These data were used to calculate

the nearest neighbors (for $k = 301$) for each school. For each school that was missing data, the median value its nearest neighbors was calculated and imputed for that variable. This completes the data cleaning and munging: only valid, numeric data remains for viable candidate schools.

4 Exploratory Analysis and Graphs

Principle Component Analysis was employed to discover the vectors of maximum variance that could be used to describe the cleaned dataset. The bar graph of cumulative explained variance, shown below, shows that the variance of the dataset can be described almost entirely using a very select few vectors. Specifically, the first three principle components describe 89.98% of the data's variance.

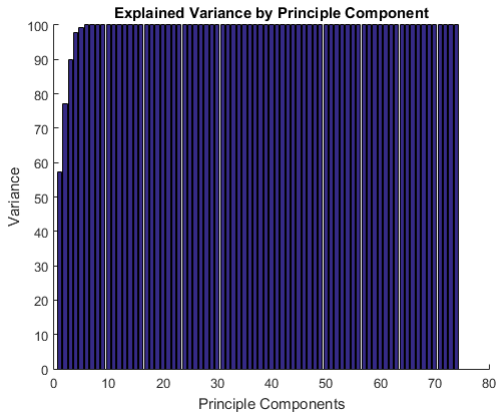

Figure 1: Explained Variance by Principle Component [2]

The dataset was clustered using the k-means algorithm in order to search for a meaningful classification of schools in order to aid the identification of attributes which could identify a school as being worthy of investment dollars. Silhouette evaluation showed that the optimal number of clusters for this dataset was k=2. In order to evaluate this clustering the data was projected onto the first and second principle components and color coded by cluster, as shown in the figure below. As is visible, there is no intuitive clustering scheme visible under these projections.

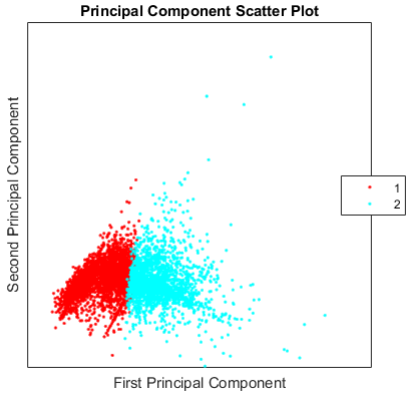

Figure 2: Clustering Projected Onto the First and Second Principle Components[2]

The data was then projected onto the coordinate system defined using the first and third principle components. The figure below shows the arrangement of data using this coordinate system. Again the clustering appears to be arbitrary without any well separated clusters.

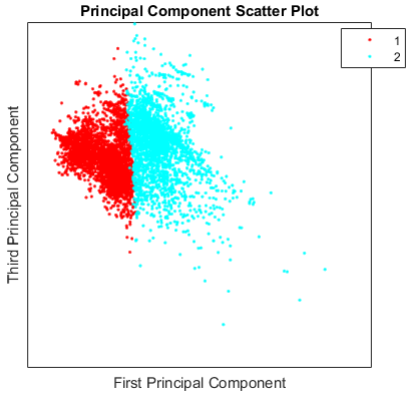

Figure 3: Clustering Projected Onto the First and Third Principle Components[2]

Finally, the data was projected onto the coordinate system defined by the second and third principle components. The figure below shows the data projected onto this vector space. In this space there appears to be some meaningful arrangement but again there aren't any well separated clusters.

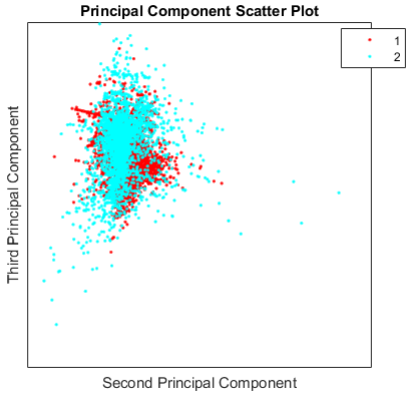

Figure 4: Clustering Projected Onto the Second and Third Principle Components[2]

5 Models

Given the loose definition of return on investment for the context of this problem, multiple avenues were pursued with the objective of maximizing the return.

5.1 Model A

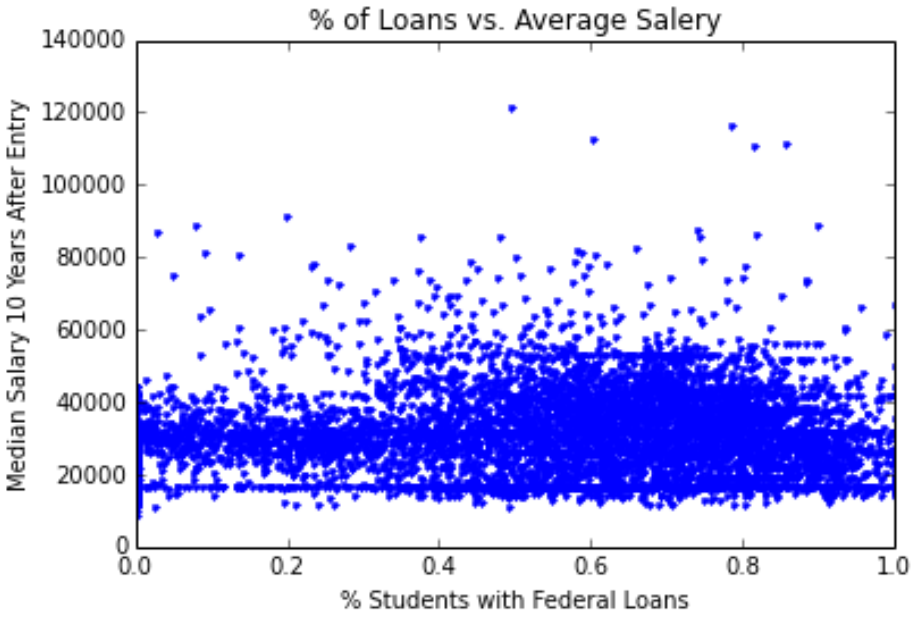

Figure 5: Comparison of % of federal students receiving federal loans and median salary of students ten years after entry.

The objective of Model A was to determine schools that have qualities to be successful if given more funding. The model first categorizes schools into three different categories based on the percent of all federal undergraduate students receiving a federal student loan. The categories are: schools with zero to forty percent, schools with forty to eighty percent, and schools with receiving eighty to one-hundred percent of all federal undergraduate students receiving a federal student loan. The group of schools where the percent of federal students receiving federal loans is zero to forty percent is the target group.

Once these groups were determined, Figure 5 was generated and it indicates that schools receiving forty to eighty percent of federal students receiving federal loans had on average higher salaries. The mean salaries between the target and forty to eighty percent groups were determined to be significantly different by running a t-test comparing the mean salaries of the groups.

From this observation it was determined that some schools in the target group could have students receive higher salaries if more funding was given to the school. Therefore, this model will only recommend schools to receive grants if they are in the target group.

Next, model schools were picked from the forty to eighty group based on two qualifications: one, the school must have a median salary fifty percent over the average of the group; and two, the students from that school must have a high ability to pay off debt. Ability to pay of debt was scored by:

$$\text{Ability to Pay Debt} = \left(\frac{\text{Median Salary After 10 Years}}{\text{Median Debt}}\right) \tag{1}$$

The top 5 schools based on the ability to pay and being above the median were pulled and assigned as model schools.

Next, the data from the schools, both the target group and the model schools, were normalized and the euclidean distance from each target school to each model school was calculated; the mean was taken for each target school and the set of distances to the model school. Lower scores indicate that the school is more likely to succeed when given money, and this was how the target schools were ranked. Lastly any school in the target group that had less than three-hundred students was ignored as being too small for the money to be used effectively.

The strength of this model is that it focuses on schools that do not receive much aid. The top schools selected will be ones that do well without aid and have attributes that should lead to high student success by receiving grants. The weakness of this model is that it ignores schools that already receive a certain amount of federal loans. If schools do well when given a certain amount of aid they are ignored due to their students already receiving federal assistance. These schools have the opportunity to have a high ROI given more grants; instead schools that don't have a high amount of students receiving federal assistance are preferred.

Rank	**UUID**	**University Name**	**Similarity Score**
1	426314	Embry-Riddle Aeronautical University	6.314794
2	437097	University of Management and Technology	6.552895
3	455512	Woodland Community College	6.881402
4	115001	Glendale Community College	6.939552
5	122384	San Diego Miramar College	6.986423
6	190512	CUNY Bernard M Baruch College	6.988203
7	187046	Thomas Edison State College	6.999098
8	140331	Chattahoochee Technical College	7.049751
9	117247	Laney College	7.060236
10	109907	Barstow Community College	7.088742

Table 2: The top 10 schools from Model A.

5.2 Model B

The premise of Model B is that the median salary of graduates from a school reflects that school's ability to foster a strong educational environment. Assuming that the graduate will accept the highest salary offered, it is therefore a representation of how the economy values the graduate's education. So it was concluded that it would reasonable to use the salary variable as the expected function output in a supervised learning application.

The input data to this analysis excluded the data in the "school" and "academics" dev-categories, leaving "completion", "admissions", "student", "cost", "earnings", "repayment", and "aid" from which to learn. This categories include the data that were previously indicated as strong factors for positive ROI. In an exercise in supervised learning and to validate this approach, the data was randomly shuffled and split into a 500-school training set and a 7304-school testing set. Each set was independently normalized and centered to have a mean of zero. Using the training data, linear regression was used to produce a model with a regularization parameter chosen with cross-validation.

$$\mathbf{w} = [(X \cdot X^{\intercal} + \lambda \mathrm{I})^{-1} \cdot X^{\intercal}] \cdot \mathbf{y} \tag{2}$$

In equation (1), $\mathbf{X}$ is the data matrix, $\mathbf{y}$ is the function output, and λ is the regularization parameter.

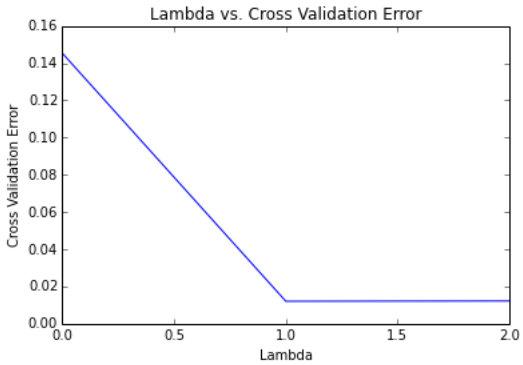

Figure 6: A plot of λ against the cross validation error. The lowest cross validation error was found to be 1.00.

Using the optimal hypothesis, $\bar{g}$, with $\lambda = 1.00$, the mean squared error within the training data was calculated to be 1.340%. The mean squared error of $\bar{g}$ for the testing set was 5.949%. This indicates that the linear model does a relatively good job modeling these data. However, since it is of no importance how well $\bar{g}$ models data outside of the given dataset, there exists no concept of over-fitting the given data. Thus, with the reassurance from the training and testing errors, the data was combined to train the final linear hypothesis, g.

Table 1 displays the ten schools with the highest output, indicating that they are ideal recipients of a donation because these schools are able to produce high performing graduates that are strongly valued by the economy.

Rank	**UUID**	**University Name**	$g(\mathbf{x}) = y$
1	215062	University of Pennsylvania	0.450461104
2	164739	Bentley University	0.448561854
3	115409	Harvey Mudd College	0.443981011
4	198419	Duke University	0.434522028
5	112260	Claremont McKenna College	0.414340125
6	166683	Massachusetts Institute of Technology	0.411835533
7	211440	Carnegie Mellon University	0.409733392
8	131496	Georgetown University	0.408210174
9	168148	Tufts University	0.382029155
10	243744	Stanford University	0.377695422

Table 3: The top 10 schools from the linear regression model.

5.3 Final Model

The final model was designed to select those schools that were targetted by Model A as being worthy of receiving funds that could maximize our Return on Investment as calculated using the equation below.

$$\text{ROI} = \left(\frac{\text{Investment} * \text{Median Salaray} * 20}{(\text{Net Tuition} * 4)} \right)^{y} \tag{3}$$

Equation (2) is designed with the idea that value is generated by allowing a student to successfully graduate from a school and enter into the workforce in a way which will enable them to find work with an increased salary. The y value, derived from our linear regression model allows the model to adjust these numbers based on the amount of value added by the completion at a student's school based on the schools attributes.

In order to generate a final list of schools in this way, the best schools identified by Model A were fed into Model B to give a final ordering. Table 2 displays this final list of schools along with their ROI as calculated using the given formula.

Rank	**UUID**	**University Name**	$g(\mathbf{x}) = y$
1	190512	CUNY Bernard M Baruch College	0.092531348
2	426314	Embry-Riddle Aeronautical University	0.067320162
3	437097	University of Management and Technology	0.046711849
4	187046	Thomas Edison State College	0.024132616
5	475565	Stella and Charles Guttman C.C.	0.016430553

Table 4: Recommendations from Model A, ranked by Model B, showing top five.

6 Analysis and Recommendations

Approach A created a model that identifies schools that have potential to improve if given more funding. The schools that have potential to improve are determined by how closely they compare to schools that do well and have more funding: the target schools. The amount of funding was determined by the percentage of federal students who receive some federal loans. Schools that do well are determined by the median salary of students ten years after entry and by the student's ability to pay off debt. Once target schools have been determined, the model determines which school that receive little funding are similar to the target schools. The idea is that given more funding, the lower funded schools can improve to become more like model school. The Model A top results consist of state and community colleges that the model indicates have a high potential for growth.

Model B, linear regression, created a model that fairly precisely calculated the worth of attending each school, using the median salary ten years after entry as the ultimate goal. Interestingly, the rankings produced by this model seem to resemble the average college rankings of various college ranking groups. Several of the top ten schools are in the Ivy League, whose graduates tend to be very successful. So choosing to invest in these schools would be investing in schools that are, quite obviously, very good investments in students because they are known to perform very highly. However, these schools also tend to be very wealthy as well, making a donation to those schools less impactful. As such, it is concluded that this would not make a good recommendation for the Goodgrant Foundation.

By using the insightful rankings from the first model and calculating the respective outputs from the linear regression, an indicator is obtained that represents the quality of the school in a secondary manner. This was how Table 4 was created. From these five optimal schools, the recommended amount of investment is calculated using the number of undergraduates at each school.

UID	**Undergrads**	**% of group**	**Total Award, USD**	**Yearly, USD**
190512	13698	30.95	3.095E7	6.19E6
426314	9689	21.89	2.189E7	4.378E6
437097	781	1.76	1.76E6	3.52E5
187046	19596	44.28	4.428E7	8.856E6
475565	493	1.11	1.11E6	2.22E5

Table 5: The breakdown of the $100 million donation between the five optimal schools.

Each top school was given an annual sum proportional to the number of undergraduates at the school. With this data, the annual return on investment was calculated for each school and the results placed in Table 6.

UID	Annual ROI
190512	5.409
426314	3.388
437097	2.049
187046	1.568
475565	1.270

Table 6: The calculate Return on Investment for the optimal schools.

In summation, the total Return on Investment for this strategy is 12.414 annually. Over the period of five years, this is a 62.070 return.

7 Conclusion

The final model suggests a methodology for finding those schools who have strong characteristics but a low degree of external funding. Using this scheme, schools who are both in need of funding, and capable of utilizing funds in an effective manner can be targeted. These schools are theorized to have the capability to achieve the same degree of success as those schools originally targeted by Model B. Injecting these schools with additional grant money should push them in the direction of those schools that are already successful and allow them to grow towards this end.

8 Further Remarks

These models were built to maximize the given ROI equation with respect to investment dollars. Expanding the feature space provided in the data set could have allowed for ROI to defined in different, and potentially more accurate, ways. Many features that were defined in the data dictionary but were missing from the data set were theorized to be useful in defining the degree of success with which additional money could be used at a given school. For instance, many of the features pertaining to graduation rates of different subgroups of students were desired. The theory was that comparing the graduation rates of students who had received PELL Grants or Federal Loans with those students who had not received funding would illustrate the relationship between funding and increased academic success. The data on giving graduation rates for various income brackets was desired with similar reasoning. For future research in the area, these features should be sought out in order to develop more robust and accurate modeling schemes.

References

[1] CollegeBoard, *Average net price over time for full-time students*, Jan 2016.

[2] MATLAB, *(R2015b)*, The MathWorks Inc., Natick, Massachusetts, 2015.