Georgia Institute of Technology®

Investigation & Classification of Median Income

Based on US Gov't Scorecard Data

Toyya Pujol-Mitchell
Chris Shartrand

Problem Description

In Fall of 2015, President Obama announced the release of the US Department of Education's College Scorecard. The goal of the College Scorecard was to allow American families to make better and more informed decisions when choosing a college. The raw data was posted on the www.data.gov for public use.

As the cost of college in the United States continues to rise, more American families are looking at college as a financial investment. Income measured after attendance is a practical assessment of one's return on the investment of college. Hence, the goal of this project is to assess the College Scorecard data with respect to median income. First, we hope to find what college characteristics are most important to a student's income six and ten years after enrolling. Second, determine if we can accurately classify these colleges into groups that produce high income earners and low income earners.

Data Description and Preparation

The 2014 financial and college data together consisted of over 1700 variables for the over 7800 post-secondary institutions in the US and its territories. The data attempts to provide a comprehensive view of the schools and their students. For example, the financial type data included financial information of the college and its students. Examples would be average instructional expenditure for a full time equivalent and average family income of dependent students. Non-financial data included data on the competitiveness of the school and strength of enrolled students (average SAT scores, acceptance rate, etc). The data also included other relevant information about the school such as regional data (State and zip code), Accreditation Agency, demographics of the schools, and highest degree offered.

The data preparation was extensive and involved

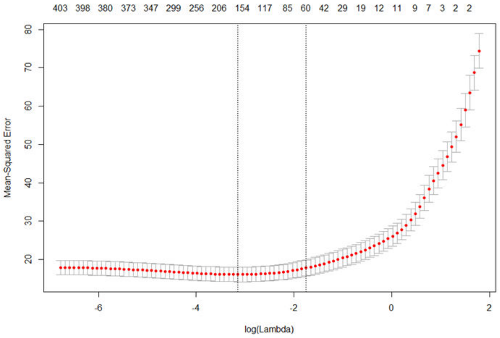
- Merging the financial and non-financial data
- Removing rows with more than 20% missing data
- Removing columns that contained the same value for every row
- Identification of factor variables and converting them into binary variables for lasso regression
- Removing columns with over 1000 factors

Methods

Model Development

We first performed multiple linear regression as a baseline for the model. The large number of variables created an Adjusted R-squared of **80.9%** for the median income 6 years after enrollment and **84.4%** for 10 years after. This model included the 481 variables that survived the data cleaning.

Due to the high number of variables, we performed Lasso for variable selection. We decided against stepwise, since the complexity of the data was too high for stepwise to run efficiently. Lasso was able to reduce the data about by 55%, which we believed was still too large for practical use. Hence, we chose eighteen variables based on the MSE vs Variables chart (see below), since it provided a low number of variables without exponential error growth. The model of eighteen variables had an Adjusted R-squared of **73.4%** for six-years and **76.2%** for ten-years. Additional fitting was performed (transformation of data for normality correction, addition of interaction terms via stepwise regression, and outlier and influential point removal) resulting in an Adjusted R-squared of **82.1%** and **83.9%** for six and ten-year respectively. This is quite close to the initial regression with all the variables, and hence a good fit given the large number of variable reduction. The results of the selected models (not including interaction terms) as well as the description of the variable names are below.



Looking at the chart below, we can evaluate the coefficients of the selected variables. The six and ten-year models have many overlapping variables, which demonstrates consistency. The majority of the variables seem to be financial and debt data of the student such as family income and cumulative debt. Family income for both independent and dependent students have a positive effect on median income. The impact of family income can also be seen indirectly in that the percent of students who received a Pell Grant, a federal grant for low income students, is a significant variable. The percent of majors awarded also comes has an effect. The percent of engineering, mechanic and repair technologies, transportation/materials moving, business, and social sciences degrees also have a positive effect on median income. On the other hand, the percent of visual/performing arts and culinary/personal services have a negative effect. The last theme is financial data of the college. Tuition revenue and expenditures per a full-time equivalent student (FTE) both have positive effects on income.

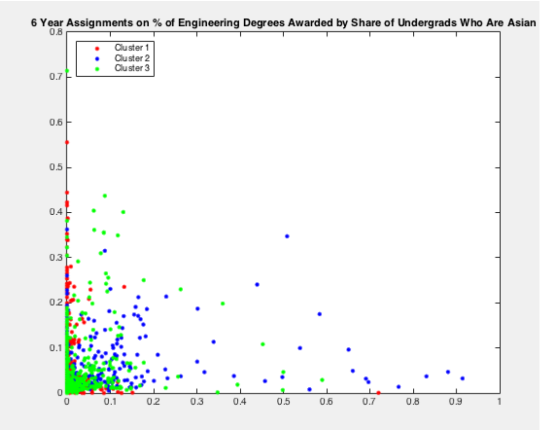
	6 Year Model	10 Year Model
(Intercept)	9.675	16.63
DEP_INC_AVG	7.23E-05	9.71E-05
RPY_3YR_RT_SUPP	7.806	4.909
WDRAW_DEBT_MDN	1.96E-04	1.86E-04
INEXPFTE	1.90E-04	1.86E-04
PCIP14	25.93	33.65
PCIP12	-5.6	-8.546
IND_INC_AVG	2.03E-04	1.45E-04
UGDS_ASIAN	23.11	31.54
CUML_DEBT_P25	2.89E-04	3.24E-04
PCTPELL	-2.268	-4.716
PCIP50	-9.481	-7.664
TUITFTE	6.12E-05	
PCIP47	5.77	8.479
GRAD_DEBT_N	-6.83E-05	3.44E-04
PCIP49	10.55	
NOTFIRSTGEN_DEBT_	-5.48E-05	
PCIP39	-6.064	
IND_INC_N	2.27E-04	
HIGHDEG		-0.1474
APPL_SCH_PCT_GE4		7.217
PCIP52		6.612
PCIP45		8.235
CIP11BACHL		0.3769

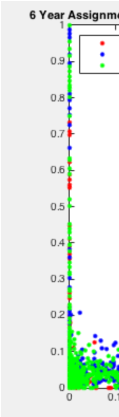
Variable Name	Variables Description
DEP_INC_AVG	Mean Family Income for Dependent Student
RPY_3YR_RT_SUPP	3 Year Loan Repayment Rate
WDRAW_DEBT_MDN	Median debt of students not completing school
INEXPFTE	Instructional expenditure per FTE
PCIP14	Percent of Engineering Degrees Awarded
PCIP12	Percent of Personal And Culinary Services Degrees Awarded
IND_INC_AVG	Mean Family Income for Independent Student
CUML_DEBT_P25	Cumulative loan debt at the 25th percentile
UGDS_ASIAN	% undergraduate degree-seeking students who are Asian
PCTPELL	Percent of Undergraduates Who Received Pell Grant
PCIP50	Percent of Visual And Performing Arts Degrees Awarded
TUITFTE	Net tuition revenue per full-time equivalent student
PCIP47	Percent of Mechanic And Repair Technologies/Technicians Degrees Awarded
GRAD_DEBT_N	Median Debt Completers Cohort
PCIP49	Percent of Transportation And Materials Moving Degrees Awarded
NOTFIRSTGEN_DEBT_N	No. of Students in the Median Debt Not 1st Generation Students Cohort
PCIP39	Percent of Theology And Religious Vocations Degrees Offered
IND_INC_N	No. of Students in the Family Income Independent Students Cohort
HIGHDEG	Highest Level of Degree Offered
APPL_SCH_PCT_GE4	No. of schools on FAFSA applications >= 4
PCIP52	Percent of Business, Management, Marketing, And Related Support Services Degrees Offered
PCIP45	Percent of Social Sciences Degrees Offered
CIP11BACHL	Bachelor's degree in Computer And Information Sciences And Support Services Offered

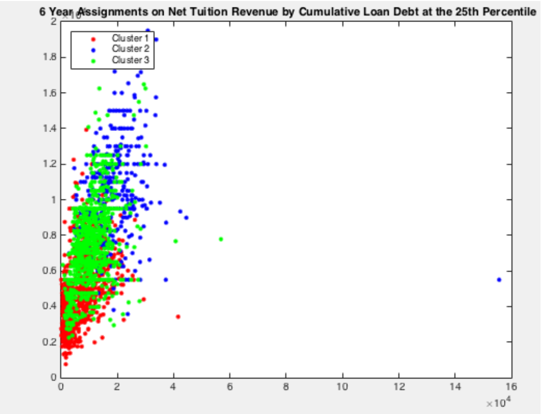
Using Lasso Regression and Linear Regression, we were able to determine the most important variables in predicting median income. Next, we want to see if clustering the data will help us assess the similarities of the colleges and any trends among the data.

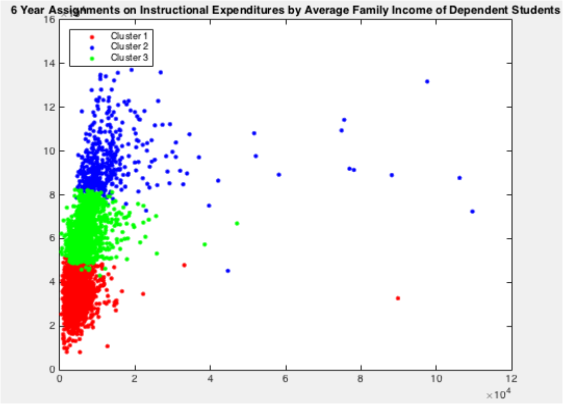
Clustering

After developing the linear models through Lasso Regression for income six-years out and ten-years out, it was imperative to understand which of the eighteen variables for each model affected the income. We also wanted to assess whether or not there were any grouping trends for the data. As a result, we decided to run a k-means clustering on the data for both six and ten years out. We chose a k = 3 groups in the hope that they would uniformly cluster into distinct groups of low, medium and high income. It immediately became clear that full visualization of the data would be impossible, as 18-dimensional space that would be necessary. Initially, we decided to deal with this issue by just producing plots of the clustering based on pairs of variables in order to view trends. Four of these plots can be seen below all corresponding to income six years out. Plots for income ten years out display similar trends.







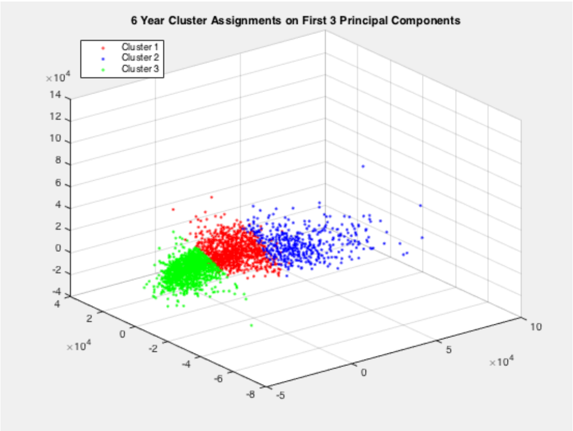


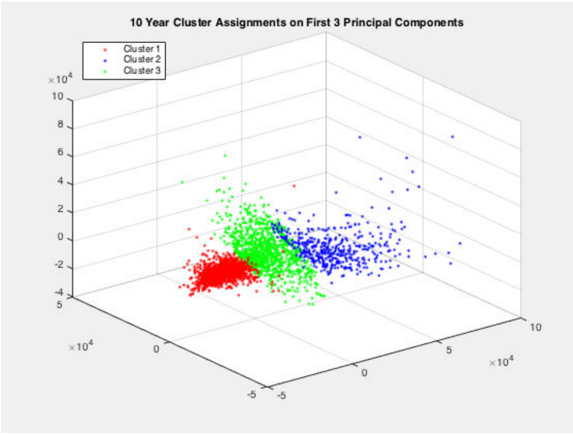
Evident from the plots was the lack of cluster predictability from the majority of the eighteen variables. The last of the four plots, instructional expenditures per full time student against average family income of dependent students, was one of the few perspectives that yielded clear cluster boundaries. We additionally found the three cluster centroids and computed the predicted income based on the linear model to discover if we had successfully grouped the data into clusters of low, medium and high income. For income six years out, we found the predicted income of the three centroids to be $\mathbf{\$22,000}, \mathbf{\$30,000}$, and $\mathbf{\$38,000}$ respectively. In the case of income ten years out, the income of the centroids was found to be $\mathbf{\$27,000}, \mathbf{\$37,000}$, and $\mathbf{\$48,000}$ respectively. These results were a key factor in confirming our intuition that college attendees can expect to have a higher income when they are ten years out of school versus six years out of school.

Despite these somewhat promising observations, we were unable to fully confirm our original goal. While the plot of instructional expenditure against average family income showed distinct cluster groups, it was only one perspective of an 18-dimensional space. Hence we still could not yet say confidently that we were able to uniformly cluster the data into three groups of low, medium and high income. Furthermore, it was also evident from the plots that visualization could not solely aid us in understanding which of the eighteen variables contributed most to the differences in income six and ten years out. Therefore, we deemed dimension reduction via principal component analysis to be necessary.

We began by reducing the size of the data for income six years out. Because we still wanted to be able to visualize the clustering, a goal of using the first three principal components was set. Analysis of the explained variance found that $\mathbf{91.25\%}$ of the variance in income six years out could be explained solely by the first three components, which was a level that was satisfactory for using only the first three principal components. Similarly for income ten years out, $\mathbf{97.45\%}$

of the variance could be explained by the first three principal components. Using the three principal components for both six-year and ten-year income, we reran the k-means clustering for three groups. The 3-dimensional plots for both six year and ten year can be seen below.





As reflected in the two plots, we were successfully able to cluster the data into three groups of low, medium and high income for both the six year and ten-year data. It is also interesting to observe the larger variation in the clusters for the ten-year data. This observation follows an intuitive sense that it is harder to model the amount of income you would make the longer out of college that you are. By analyzing the principal component scores, we were able to find which of the underlying variables most greatly affected income. For income six years out, average family income of dependent students, average family income of independent students and institutional tuition revenue per student were the most heavily weighted variables for the first, second, and third principal component respectively. Therefore, this tells us that the amount of money that your family makes and the amount of money that you spend to attend college are the most important determining factors for the amount of income you would expect to make six years out of college. In the case of income ten years out, average family income of dependent students, average family income of independent students and institutional expenditure per student were the most heavily weighted variables for the first, second, and third principal component respectively. Comparing the two results against each other found that even after an extra four years out of college, the amount of money your family makes is still the most important factor in determining the amount of money you make. However, there is a key shift in the change from institutional tuition revenue to institutional expenditure. This phenomenon indicates that in the short term, your earnings depend on the amount of money you paid to your college, as in you may accept a lower paying job after graduation to start to pay off student debt versus staying unemployed for a longer period of time in order to find a higher paying job. However, in the long term, your earnings depend on the amount of money that your college paid to educate you, indicating that an institutions willingness to fund the educational process greatly affects their students' future wage earnings.

Classification

Following the cluster and principal component analysis, we found that while we had gained monumental insight into the factors that affect income earnings we still lacked predictive capabilities on classifying whether a college is likely to produce a low or high income student based on their institutional data. The creation of a logistic regression was therefore warranted. To generate the binomial data of low or high income, we found the mean income from the institutional data and classified all 3,046 institutions into low income if they were below the mean and high income otherwise. This process was run for both the six-year income data and ten-year income data. Based on the logistic model that was produced for both sets of data, we computed the odds that an institution was to be included in the high income classification. If the odds were below 0.5, we placed that institution into the low income classification and all others were placed into the high income classification. Following this, we created a confusion table to quantify the predictive capabilities of our classification model. The table for both six and ten-year logistic regression models and the K-Nearest Neighbors (KNN) models, to be described following, can be seen below.

Classification Model	Correct	Incorrect	Percent Correct	Percent Incorrect
Logistic 6 Years Out	2714	332	89.10%	10.90%
Logistic 10 Years Out	2700	346	88.64%	11.36%
KNN 6 Years Out	631	131	82.81%	17.19%
KNN 10 Years Out	624	138	81.89%	18.11%

In the clustering section, we discussed the limitations of K-Means. Nevertheless, we wanted to attempt to use a nearest neighbor algorithm for classification. KNN was a good choice. We chose a K=4, which was determined by a leave-one-out cross validation. For the KNN implementation, **75%** of the data was used for training, which is why the nominal number correct is lower than the logistic. The results of the correct rates can be seen in the table above. The KNN performed quite well, but slightly worse than that of the logistic regression.

Results and Future Work

The results of our analysis can be summarized below

- We were able to create a linear model with an Adjusted R-squared of **82.1%** for 6 years from start of enrollment and **83.9%** for 10 years
- Utilizing the model selection process of the linear model, we were able to find the eighteen most important variables for predicting median income. These eighteen variables were the basis for the clustering and classification
- We were able to successfully cluster our data into groups of low, medium and high income for both six years out and ten years out. The predicted income for the centroids of the groups were **$22,000, $30,000, and $38,000** and **$27,000, $37,000, and $48,000** respectively.
- By using principal component analysis, we were able to discover that six years from enrollment, the amount of money that your family makes and the amount of money that your institution makes per student were the most important factors in determining the amount of money that you make. While for ten years out of college, the amount of money that your family makes and the amount of money that your institution spent on educating its students were the most important factors in determining the amount of money that you make.
- By fitting a logistic regression on our data to classify institutions into producing low and high income earning students, we were able to correctly classify **89.10%** of institutions six years out but that number dropped slightly to **88.64%** for ten years out.
- KNN classification yielded slightly worse results than that of the logistic regression with correct rates of **82.81%** and **81.89%** respectively.

Some suggestions for additional work in this project is described in this section. One suggestion is changing the number of K for the KNN algorithm. There are methods to find an optimal K, via cross validation, which may improve our classification accuracy. In addition, other classification methods such as Linear/Quadratic Discriminate Analysis, Principle Component Analysis or Support Vector Machines could be explored to determine if they yield better results.

For the logistic regression, we chose the default probability of 0.5 as the determining cut-off. A range, such as 0.4 to 0.6, could be explored to see if the correct rate could be increased via the logistic regression. The number of variables could be increased to 30, which as the smallest number before the MSE growth rate was no longer linear.