The University of Chicago Booth School of Business

BUSE 41201– Big Data – Section 01
Spring Quarter 2017 – Veronika Rockova

Higher Education in America:

Exploring the Costs and Benefits of Attending Four Year Colleges

8 June 2017

By Maayan Aharon, Sruti Balakrishnan, Nicholas Lilovich, James Mack

Honor Code

We pledge our honor that we have not violated the Booth Honor Code during this assignment.

1. Executive Summary

In this report, we analyzed United States Department of Education "College Scorecard" data from 1996-2014 in order to understand the characteristics, benefits, and costs of the various higher education institutions in America. We were thus able to reach the following insights:

In our first analysis, we use a series of regression techniques to understand the institutional or student characteristics that drive school admission rates. In the end, we found that we are indeed able to predict highly selective schools with a reasonable degree of accuracy, with the random forest technique having the most predictive power. This result is unsurprising, since colleges likely form into natural "groups" based upon selectivity, with characteristics unique to each group – a data structure that lends itself well to a non-parametric model like the random forest. Two of the most meaningful variables in predicting these types of schools were 1) the college completion rate of past students, and 2) the percentage of students receiving federal loans.

Second, we attempted to divide the types of institutions that offer four year degrees into various categories (e.g., demographics, admissions criteria) using non-parametric categorical analysis, and use those categories to predict future student earnings. We found that one can indeed create meaningful groupings of schools based upon academic programs, demographics, and admissions criteria, and those groups are highly predictive of future earnings. This analysis is extended to use these groupings as a filter in exploring colleges that best outperform their predicted earnings.

We next investigated whether historically black colleges and universities (HBCUs) helped improve black student outcomes. We found that when controlling for student demographics and financial information, the HBCU designation alone did not appear to be predictive of improved black student completion – the variables that are most predictive of student completion appear to be related to college selectivity and wealth. However, it is worth noting that HBCUs are known for characteristics like a high black undergraduate population, a variable that our analysis suggested was a contributor to black completion.

Finally, we examined what schools and degree programs provide the best return on investment. The results of this analysis were unsurprising; schools that offer a large number of engineering and math degrees, have high completion rates and SAT scores tend to offer the best ROI (e.g., Ivy League colleges and engineering schools).

2. Introduction to Dataset and Analysis

In September 2015, President Barack Obama and the U.S. Department of Education revealed the "College Scorecard," an extensive database containing federal data for over 7,000 U.S. higher education institutions. This dataset contains records from the U.S. Treasury, the IRS, and the Department of Education, and covers a wealth of information pertaining to college performance, including admissions rates, student demographics, college cost, student debt, and student outcomes from 1996-2014. Although the federal government avoided explicitly ranking these schools, this data was intended to provide consumers with greater transparency into the costs and benefits of higher education. Accordingly, we used this data to perform our own investigation into the various characteristics of colleges.

We limited our analysis to four-year colleges only, for better comparability (since two-year colleges tend to behave quite differently). Even so, the pruned dataset still contains over 2,000 U.S. colleges (including both nonprofit and for-profit institutions), and over 1,600 variables. (Please see Appendix I for a data dictionary containing explanations of the most common variables used in this analysis.) Our research questions follow:

A. **What institutional or student characteristics drive school admission rates?** Specifically, what characteristics are predictive of a "highly selective" institution?
B. **How can we categorize the types of institutions that offer four year degrees?** How do those categories vary in their ability to help students earn high salaries? What schools transcend their "inputs" and help students earn higher salaries than would be expected?
C. **Do schools that target specific types of students (e.g., historically black colleges and universities) result in better outcomes for those students, compared to their performance at more generalized higher education institutions?**
D. **What schools/degree programs provide the best return on investment?**

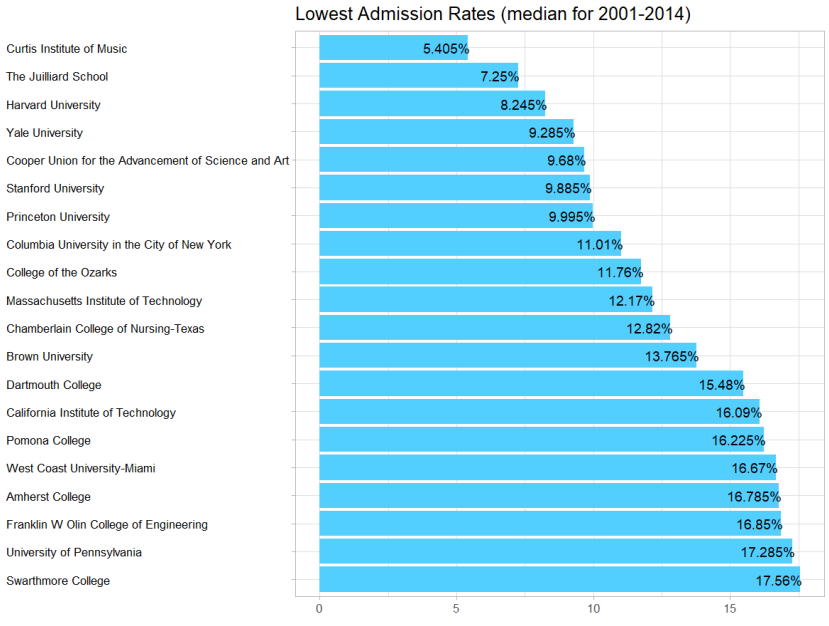
Below, we will conduct some exploratory analysis to better understand this dataset, and then address each of these research questions in order. We will end by sharing some preliminary conclusions from our research.

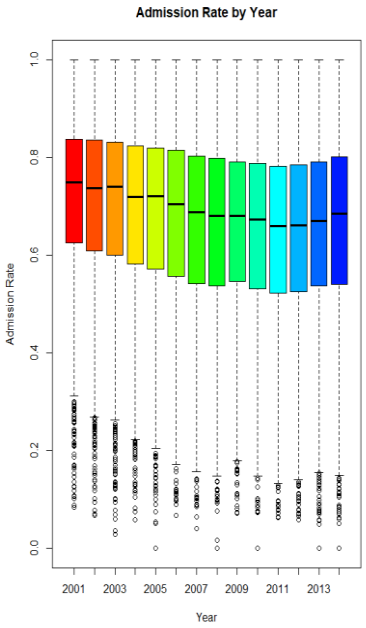
3. Exploratory Analysis

Looking at the full set of data, we can see some interesting trends with regards to admission rate, completion rate, student demographics, SAT score, and debt/aid.

A. Admission Rate

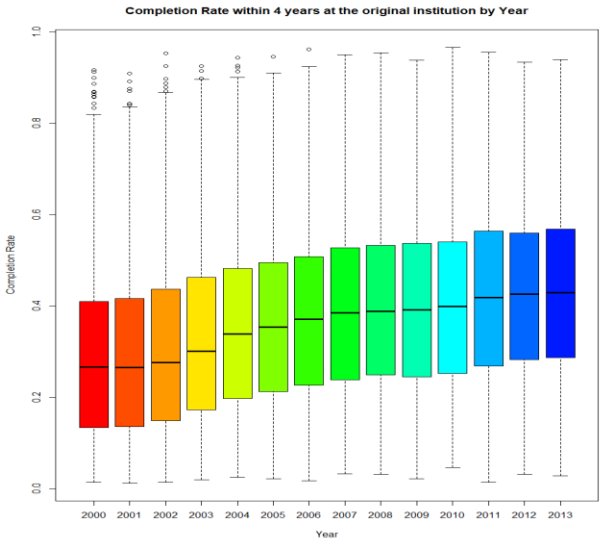
Overall, aggregating for all institutions, admission rates are high and usually fall between 60% and 80% percent. From 1996-2014, admission rates fell from an average of 71.2% to an average of 66.2%. On the right, you can see the most selective schools, on average, across 2001-2014.

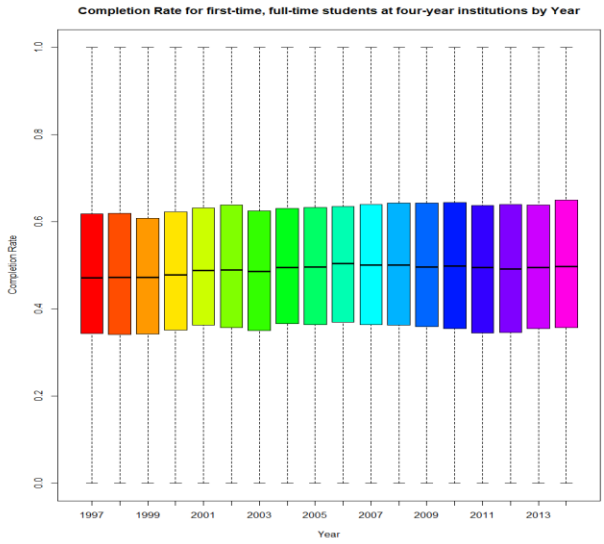




B. Completion Rate

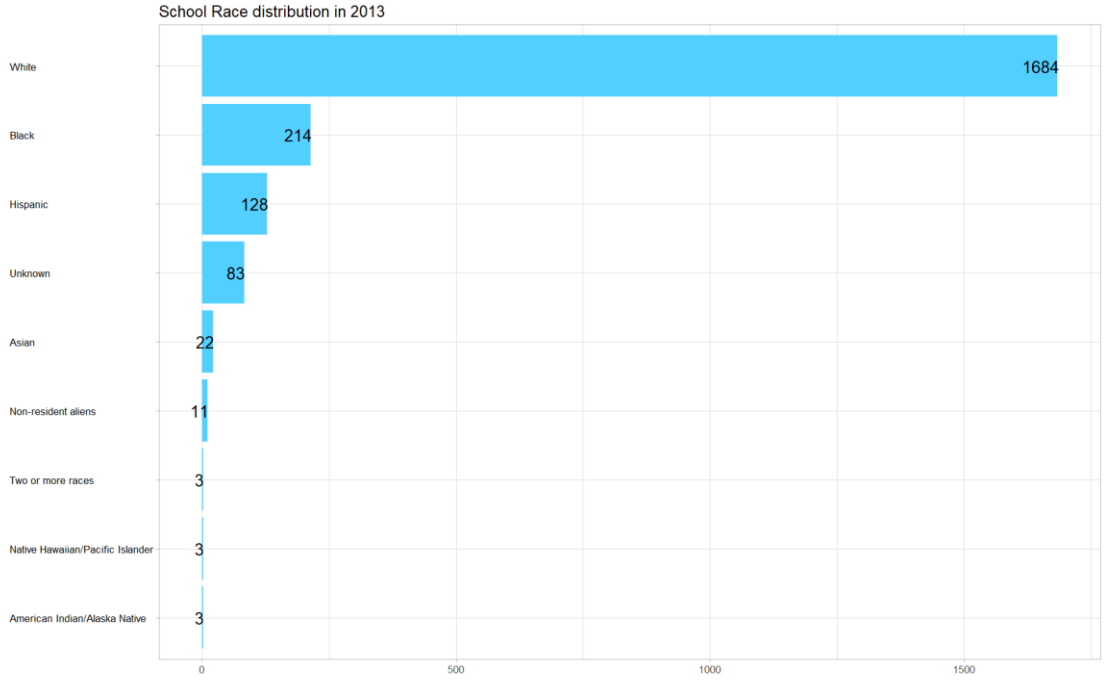
Aggregating for all institutions, the percent of students who completed within 4 years at the original institution rose between 2000 and 2013 from an average of 29.4% to an average of 43.8%. Looking at completion rate for first-time, full-time students at four-year institutions we can see a relatively steady compilation rate, around 50% between 1997 and 2014.

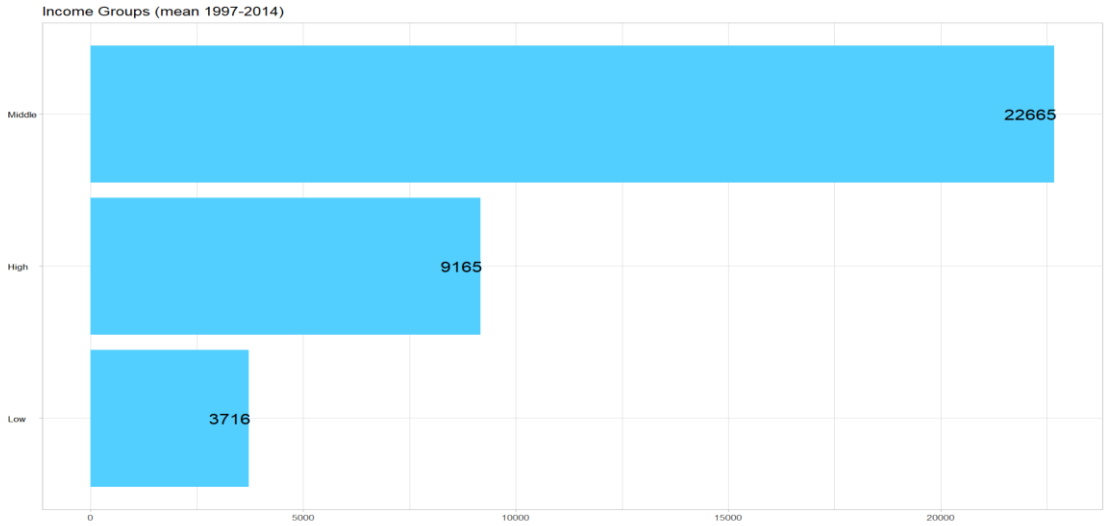




C. Students Demographics

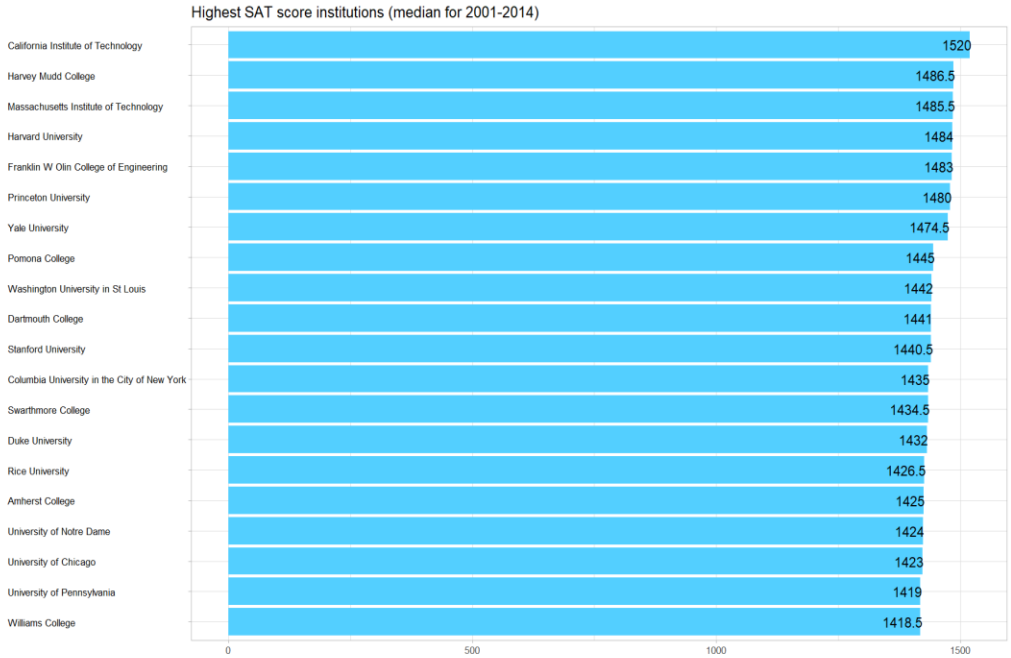
We categorized schools by race according to the predominant race of share of enrolled students. In 2013, the predominant school race was White followed by Black and Hispanic Schools. Looking at the average student family income groups between 1997-2014, we can see that most students come from medium (USD 30,000 to USD 75,000) or high (over USD 75,000) income groups.

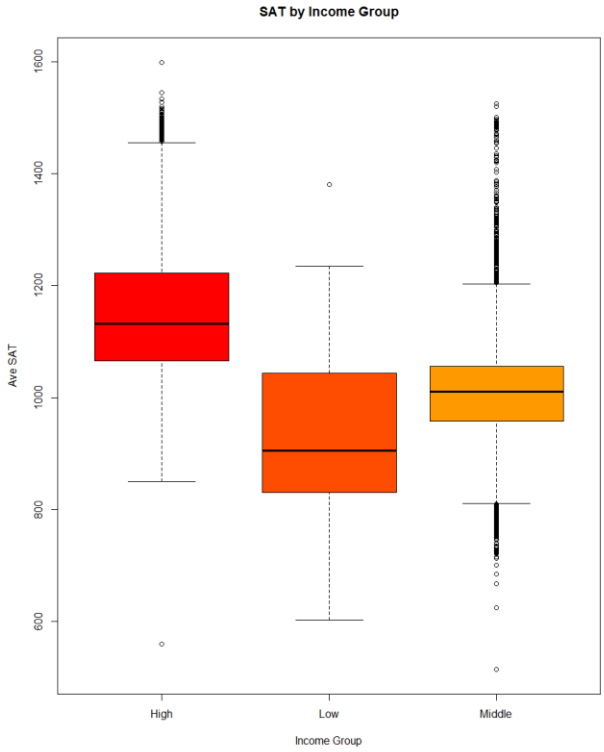




D. SAT Scores

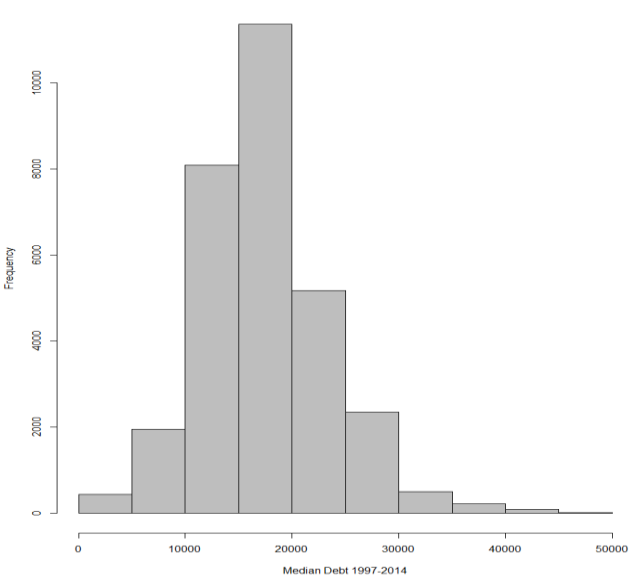
Highly selective schools correlate with higher average SAT scores. Twelve out of the twenty most selective schools also have the highest SAT scores. The other 8 are most likely selective on other characteristics (like music talent for the Juilliard School). Average SAT scores also correlate with income groups. As the typical student at the school comes from a higher income group, the school's average SAT score is higher. Looking at SAT scores by the type of degree awarded (across all years), we can see that SAT score increases as a higher percentage of the degrees at the school are Engineering, Math or Gender Studies-related. Surprisingly, the SAT score decreases as a higher percentage of the degrees awarded are Computer Science or Legal related. SAT score by degree type graphs are presented in Appendix 3. We investigate some of these curiosities in the analysis that follows.

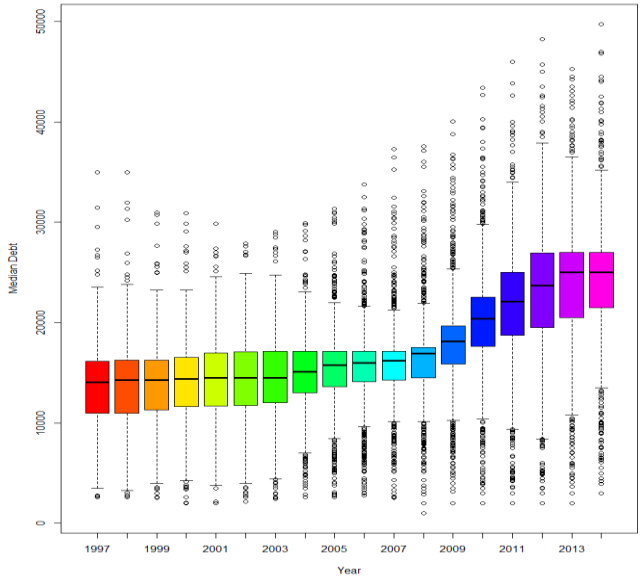


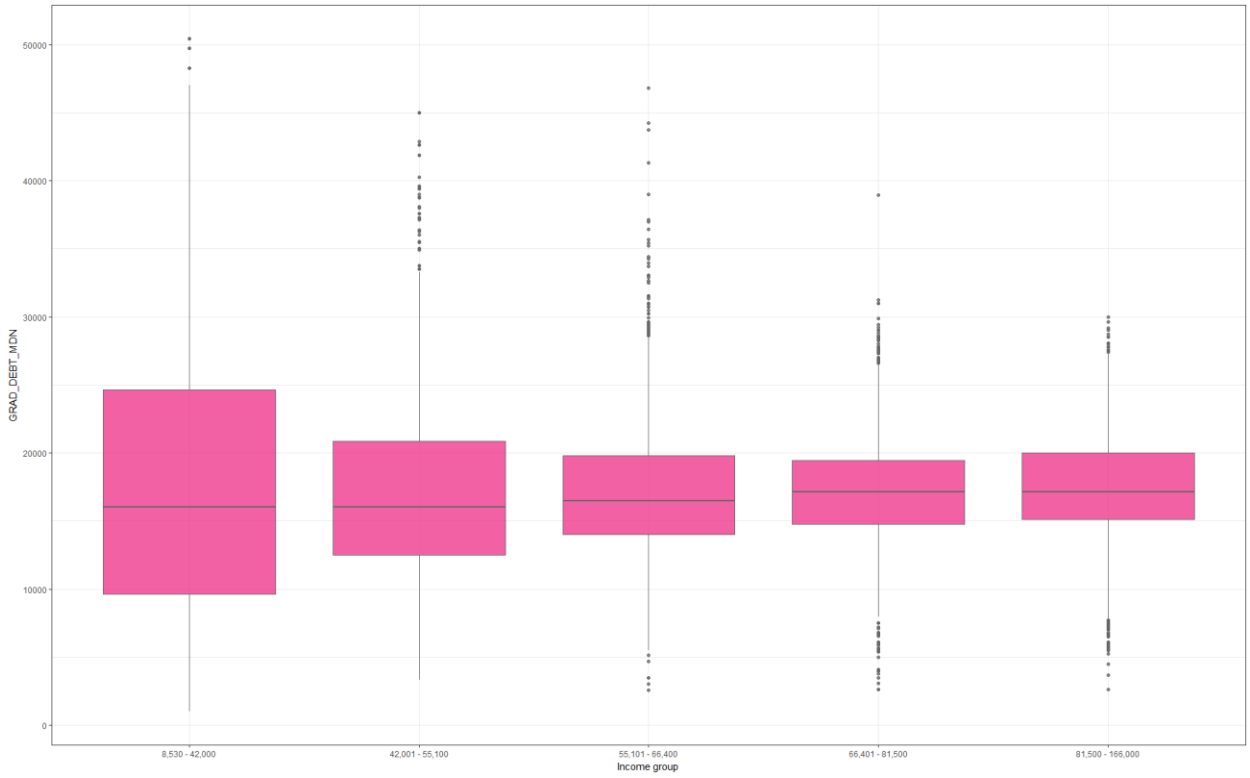


E. Debt

We wanted to look at the initial debt students have upon graduation to later analyze the return on investment by institution. The median debt between 1997 and 2014, for most students, was between USD 5,000 and USD 30,000. Nevertheless, the overall average debt has risen over the years from an average of USD 13,497 to an average of USD 24,281 (44.4%). When we explore the conditional distribution of the median debt at a school (over the several years in the data set) relative to the average student family income reveals that the average debt is relatively constant among the different groups, but lower income schools have higher variance in their debt. Comparing at median debt by institution between 1997 and 2014 shows that the investment on education varies significantly among schools. Appendix 3 show the top and bottom 20 institutions by their student's debt.


Frequency
0
2000
4000
6000
8000
10000
0
10000
20000
30000
40000
50000
Median Debt 1997-2014


Median Debt
0
10000
20000
30000
40000
50000
1997
1999
2001
2003
2005
2007
2009
2011
2013
Year


GRAD_DEBT_MDN
0
10000
20000
30000
40000
50000
8,530 - 42,000
42,001 - 55,100
55,101 - 66,400
66,401 - 81,500
81,500 - 166,000
Income group

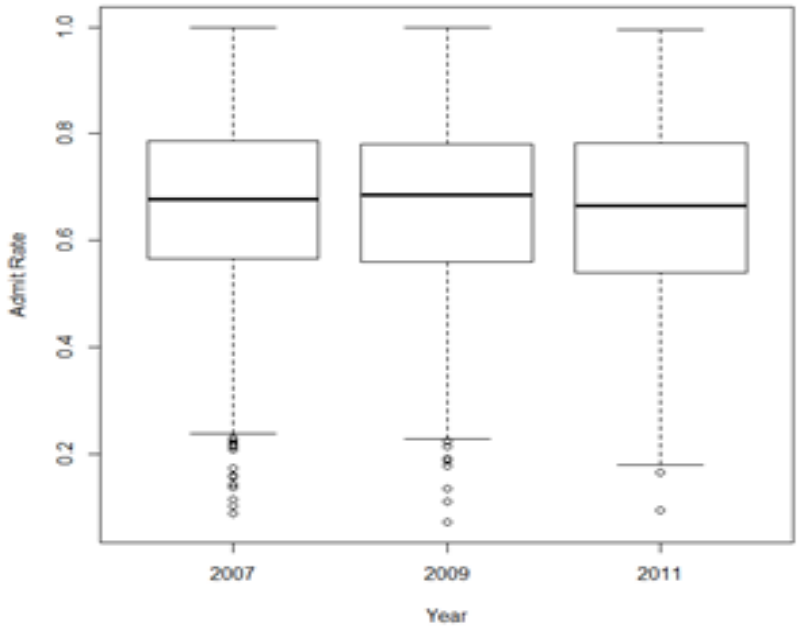
4. What institutional or student characteristics drive school admission rates?

We were struck by the number of four-year colleges that had admission rates of 100%, and were interested to investigate our ability to predict highly selective schools.

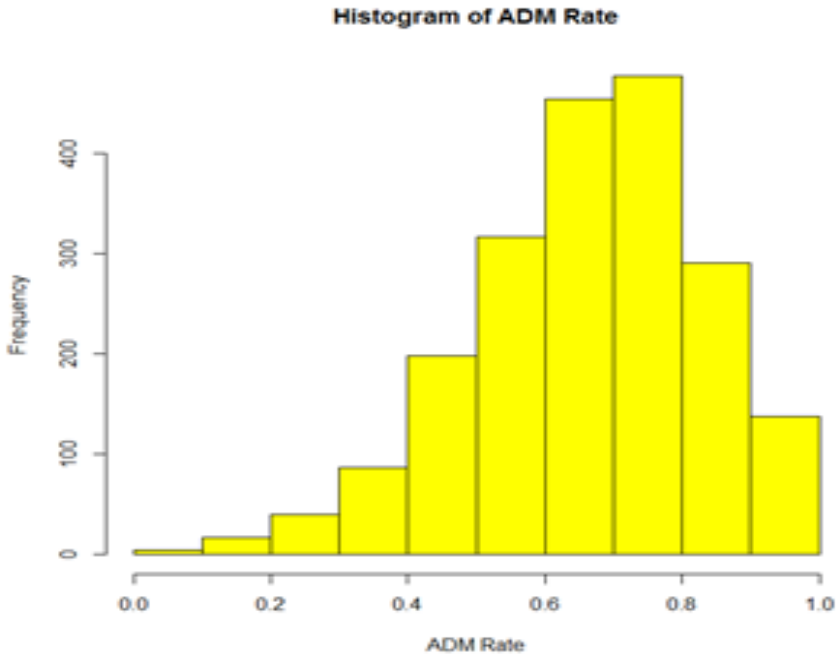
A. Data Preparation and Initial Investigation

We subsetted the data to 77 independent variables that we believed could be important in our analysis. These factors included: SAT scores for admitted students, student demographics, school characteristics, completion rates, proportion of students studying various topics, cost of tuition, student earnings 10 years after graduation, etc. As these variables were not available for all years (and not all schools had data for all variables) our subsetted data is composed of 2,025 school-year observations (783 in 2007, 752 in 2009, and 490 in 2011).

Admit Rate by Year:

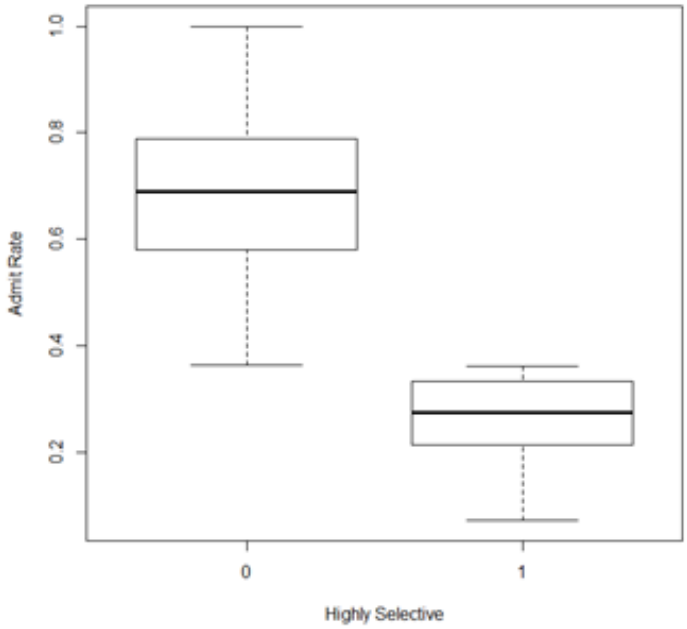


While we will include year as a factor variable, it appears that admit rate is pretty stable across the years we are examining. This is not surprising as this analysis is spanning only four years, and it is unlikely that there could be structural changes in such a short period of time.



The admit rate has the greatest density between 60%-80%. We define a "highly selective" school as having an admit rate of 5% the observed admit rate distribution. Therefore, in our subsetted dataset, a school is "highly selective" if it has an admit rate below 36.15% (the 5th percentile in the admit rate distribution). This "highly selective" category will be used as the y-variable in the bulk of our analysis.

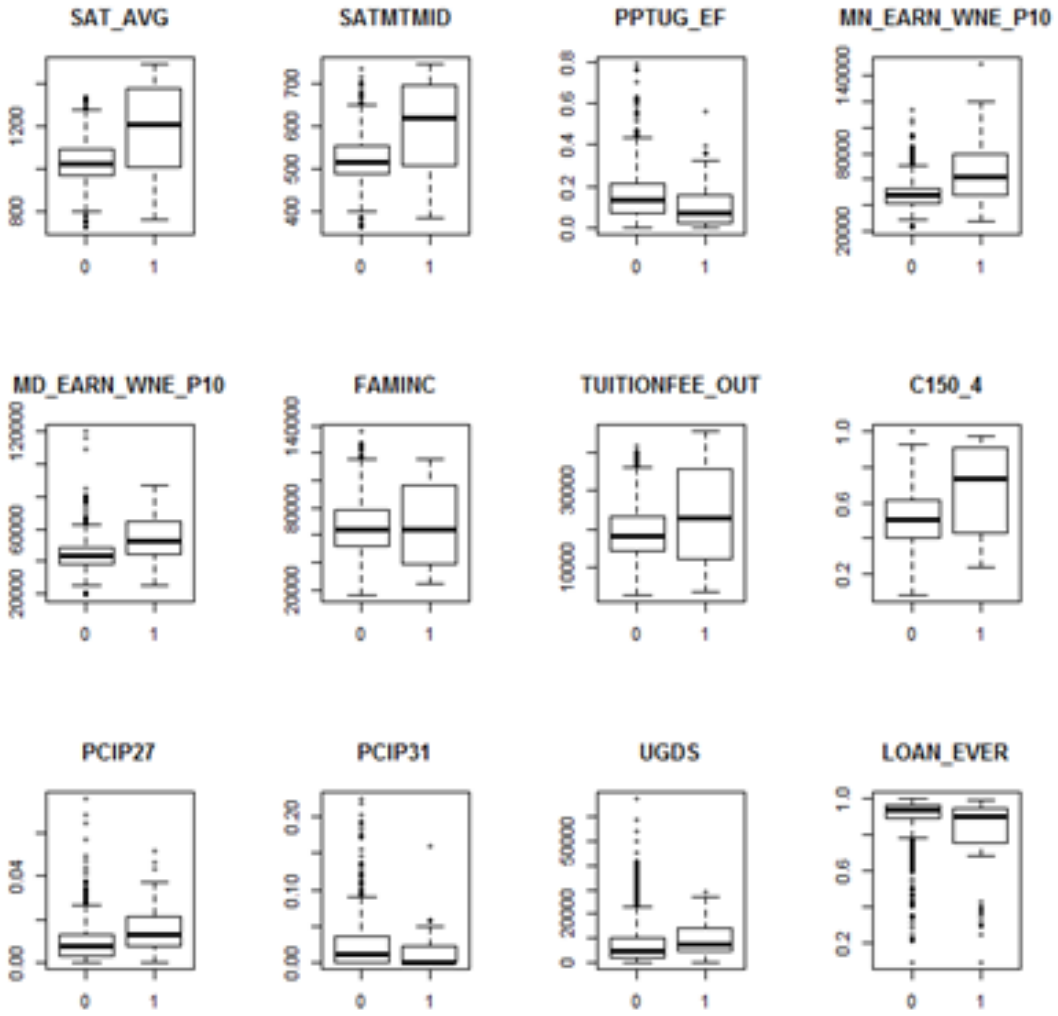
Box Plot of Highly Selective Schools and Admit Rate:



The non-highly selective schools have a median admit rate of approximately 70%, while the highly selective schools have a median admit rate of approximately 30%. As these non-highly selective and highly selective schools have

very different admit rates (by design), it will be interesting to see if these differences can be explained using the other variables in the dataset.

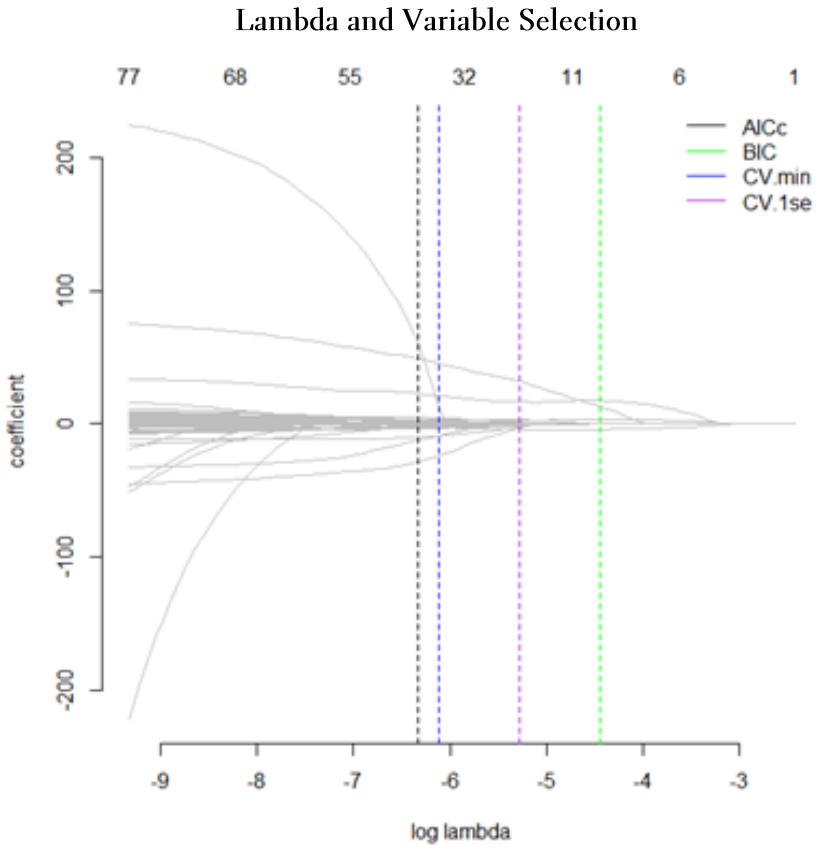
Conditional Distribution of Highly Selective and a Sample of Independent Variables



As seen above, highly selective, and non-highly selective schools appear to have different distribution of values for some of the various independent variables, which suggests that our efforts to predict "highly selective" may be successful.

B. Analysis

First, we fitted Lasso models on our binary "highly selective" variable using the 77 independent variables.

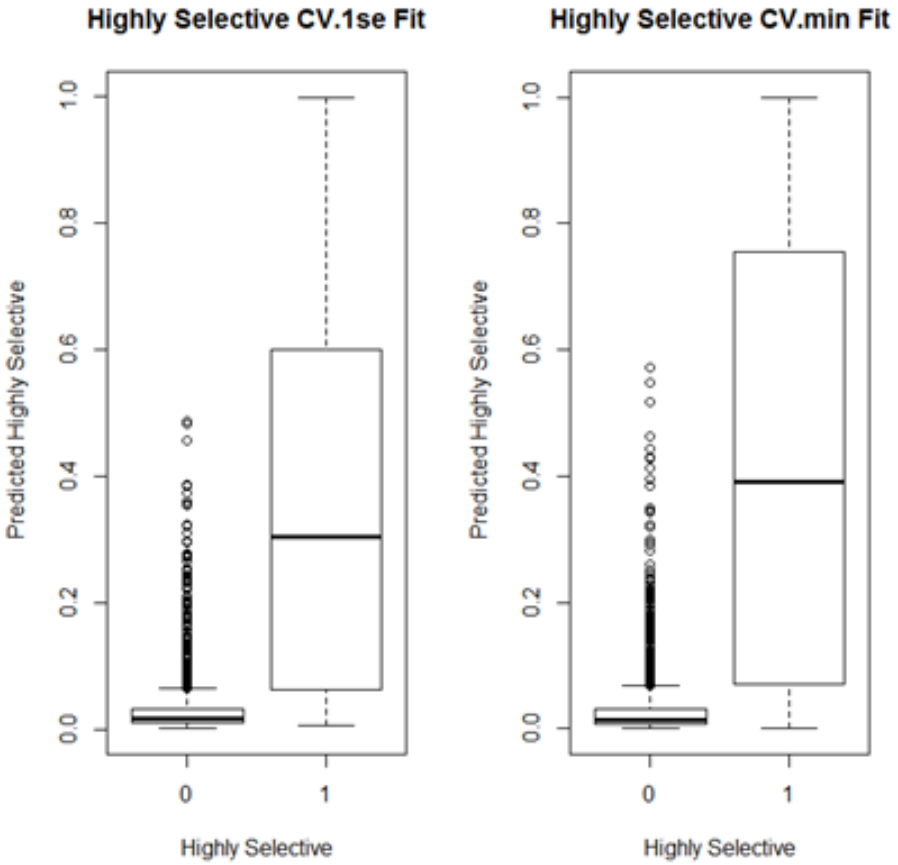


Variable	CV.min	CV.1se
PCIP05	20.6717	15.2698
PCIP45	0.8588	3.2976
PCIP27	44.5610	1.0576
SATVR25	0.0087	0.0002
SD_EARN_WNE_P10	0.0000	0.0000
PCT90_EARN_WNE_P10	0.0000	0.0000
LOAN_EVER	-2.8878	-3.4995
region_5	0.6233	
region_6	-1.3787	
region_7	-0.6133	
region_8	0.5779	
MAIN	0.5129	
PCIP01	0.5335	
PCIP03	-7.6933	
PCIP04	0.0450	
PCIP09	1.5385	
PCIP14	-3.0566	
PCIP15	-0.4222	
PCIP22	-22.3687	
PCIP23	3.8291	
PCIP26	-8.5189	
PCIP43	1.1201	
PCIP44	4.3529	
PCIP51	-0.3900	
UGDS	-0.0001	
DEP_INC_AVG	0.0000	
IND_INC_AVG	0.0000	
PELL_EVER	2.1628	
FEMALE	0.7192	
MARRIED	-2.2650	
FAMINC_IND	0.0000	
C150_4	1.6561	

As expected, the number of selected variables follows the usual pattern: AICc > CV.min > CV.1se > BIC. Below, the coefficients selected using the for CV.min and CV.1se rules are listed.

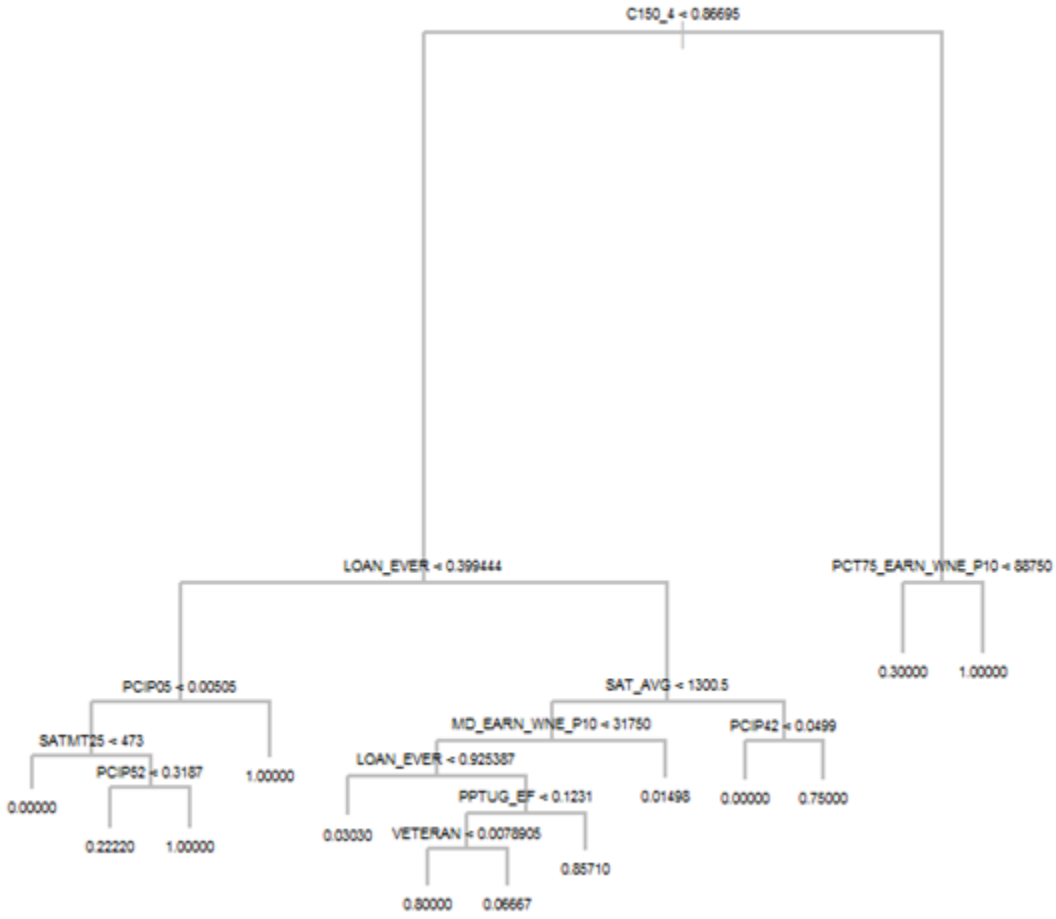
In this version, our CV.1se selects 7 variables, while CV.min selects 32 variables. We were surprised that the proportion of students pursuing certain courses of study compromised such large number of the selected variables (our PCIP variables). Particularly, the larger the proportion of students studying ethnic, cultural, and gender studies (PCIP05) or math and statistics (PCIP27), the more likely that it is a highly selective school. Conversely, the larger the proportion of students studying legal profession and studies (PCIP22), the less likely that it is a highly selective school. This is likely because these are four-year bachelor programs, this legal training is to be a paralegal rather than a lawyer, which intuitively is less competitive.

Next, we plotted y (highly selective) and $\hat{y}$ (predicted highly selective) for our CV Lasso models to compare their predictive ability.



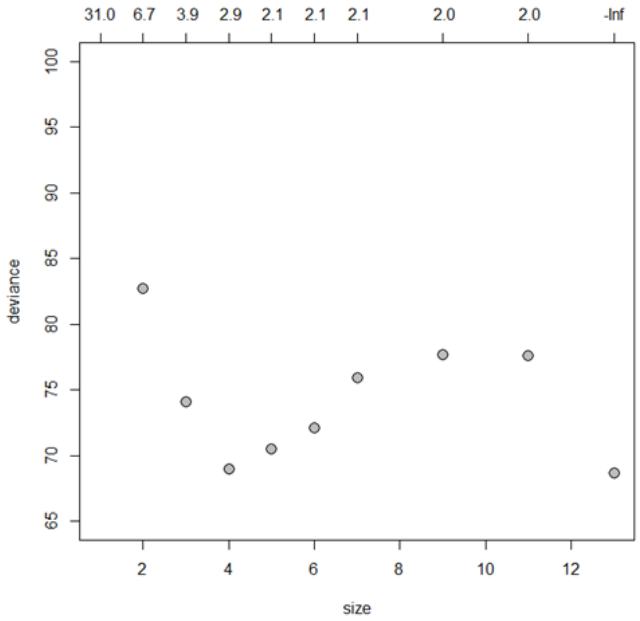
Unsurprisingly, because it is selecting more independent variables, CV.min performs a little better in prediction; CV.min has an R^2 of 0.466 and CV.1se has a R^2 of 0.414. Next, we used trees and random forests analyses as another technique to predict highly selective schools. Below is a non-parametric tree model when we fit our binary highly selective variable on our 77 independent variables.

Full Tree Model



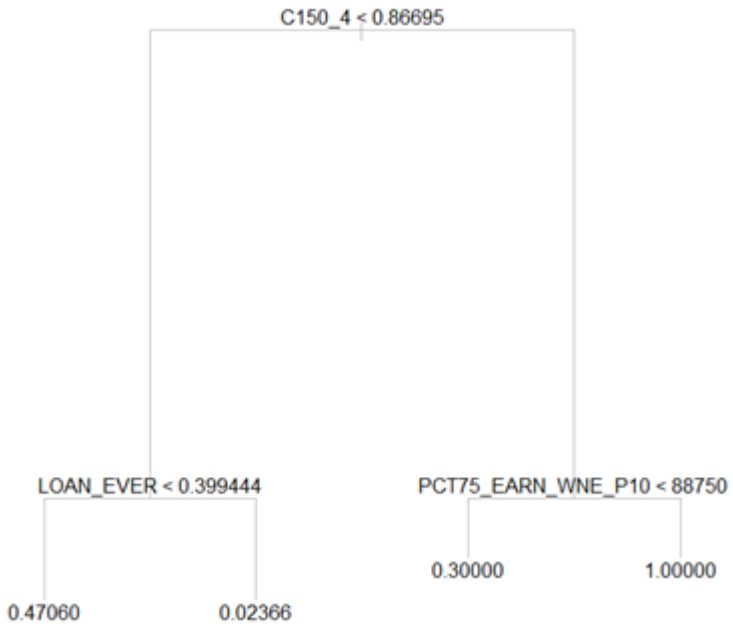
The model appears to perform pretty well in sample; the proportion of highly selective schools at each of the leaves is typically a very high or very low number. The in-sample R^2 for our full tree is 0.625. However, as the model selects 13 leaves, it is very likely that the model is overfitting and will not perform well out of sample. In order to combat this overfitting, we investigated pruning our tree.

Deviance and Tree Size



As you can see, while the 13-leaf tree has the lowest deviance, the deviance for a 4-leaf tree is only marginally higher. Thus, we believe it is worthwhile to investigate a pruned tree with 4 leaves.

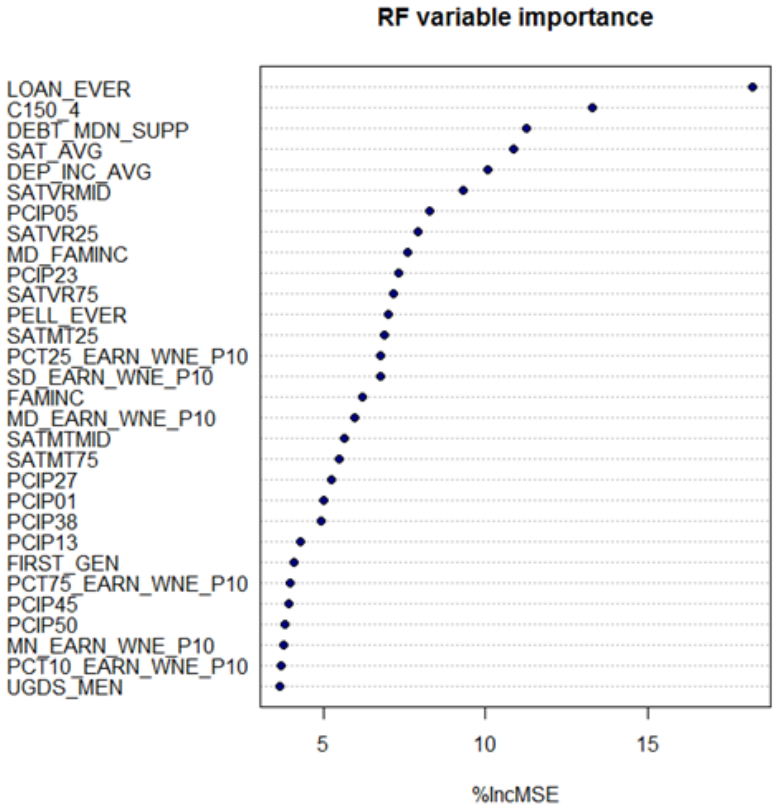
Pruned Tree



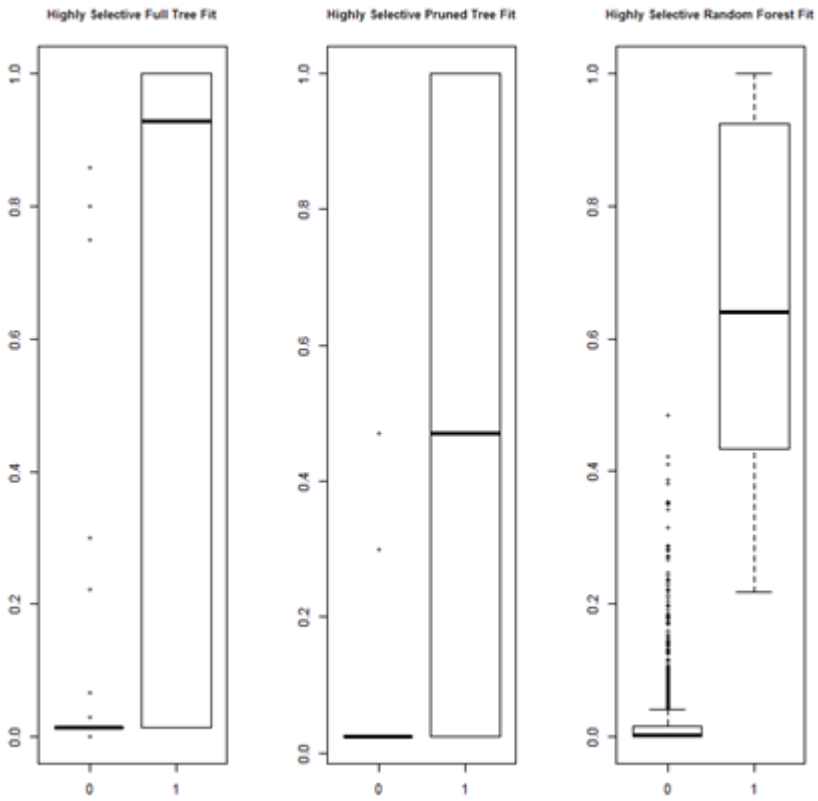
While our model's R^2 does drop to 0.427 with our pruned tree, it is still doing a pretty good job in predicting highly selective school and the reduced size of the model lends itself to interpretability. Furthermore, we suspect in an OOS analysis that the pruned tree may outperform the full tree.

Our pruned tree tells us that if a school's 4-year completion rate is above 86.69% and if the 75th median earning percentile 10 years after college is greater than $88,750, it is a highly selective school (100% highly selective schools). Conversely, if a school's 4-year completion rate is below 86.69% and if the percent of students receiving federal loans is greater than 39.94%, than it is almost definitely not a highly selective school (2.36% highly selective schools).

Finally, the last model that we fit on our data is a random forest. We were particularly interested in this model, since this type of model has been recently performing very well in prediction data analytics competitions.

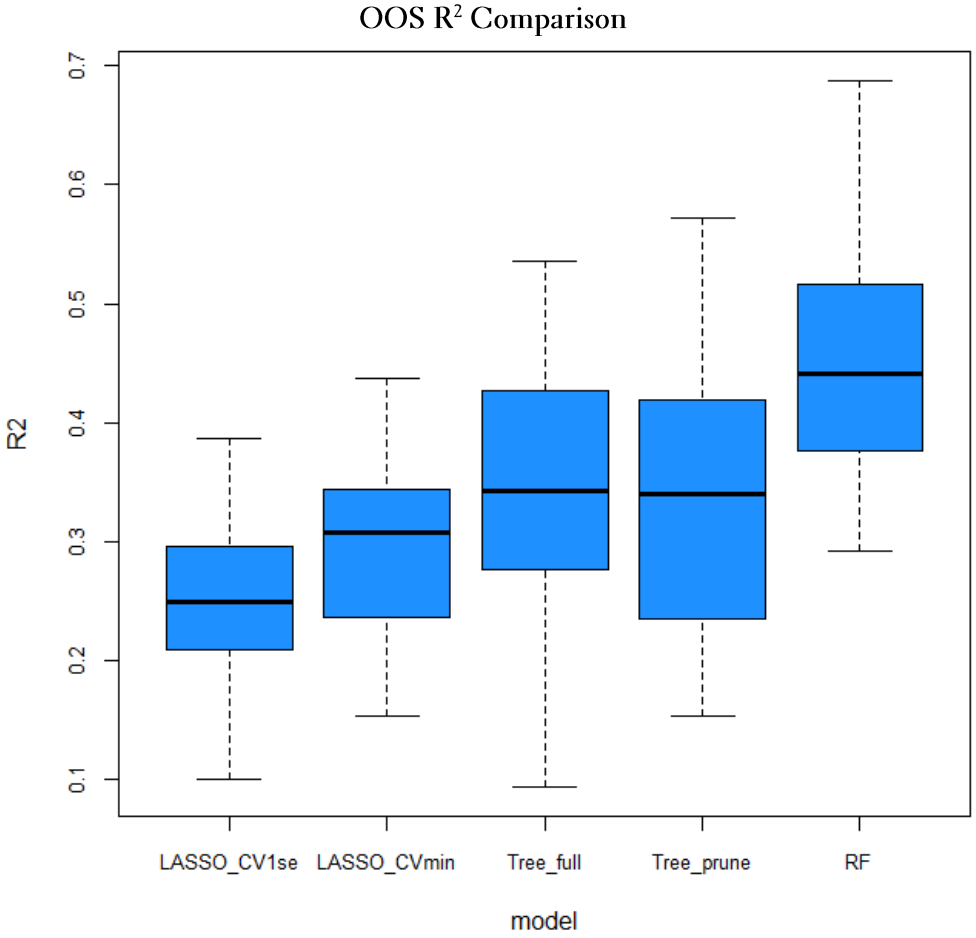


The random forest selects the % of students receiving federal loans and the completion rate as the two most important variables in predicting whether or not a school is highly selective. The random forest performs the best in R^2, with a value of 0.788. Now, after having modeled a full tree, pruned tree, and random forest model, we wanted to visually see their in-sample fit.



As our R^2 value suggested, the random forest performs the best and has quite a bit of separation between non-highly selective and highly selective schools. While we can visually see that the pruned tree does not perform as well as the full tree in-sample, we suspect that this result could change out of sample. We have made CV.lasso.1se, CV.lasso.min, full tree, pruned tree, and random forest models to predict whether a school is highly selective. Thus far, we have compared in-sample fit; however, the best test of their predictive ability is to see how the models perform out of sample, so we can more confidently assert which model performs the best.

To this end, in a loop that we run 20 times, we randomly split the data into a training dataset of 1,600 observations to estimate the models, and a 425-observation testing dataset to evaluate the models. We evaluate the models by calculating the out-of-sample R^2. Thus, at the end of the analysis we have 20 OOS R^2 values for each model. Below is a boxplot of our findings.



Our random forest model performs the best in predicting highly selective schools, with a median OOS R^2 of 0.42. As we suspected, our pruned tree outperforms our full tree OOS, indicative of the full tree's overfitting. We are impressed by the power of the random forest model and are looking forward to using it in our professional careers.

5. Categorical Analysis and Earnings Residuals of Four Year Colleges?

This analysis had several goals. First, colleges were categorized into "buckets" that would be relevant to a prospective student, using three categories of data: academic programs, demographics, and strictness of admissions. Next, the principal components for each category were calculated, and used to compare groups to determine whether they represent meaningful variation in the college data set. With these complete, median earnings were predicted for each college using the entire set input data (using and comparing three estimation methods). Finally, a few examples in college selection are presented.

A. Description of Data Used

Income

The income variable used (MD_EARN_WNE_P10) corresponds to the median income earned by Title IV students ten years after graduation. We used the 2012 value, which refers to students who graduated in 2002. Title IV students are all students who received Federal grants or Federal loans. While this corresponds to a majority of all undergraduates,[1] these students are poorer than those who did not receive grants or loans. Further, in colleges with many wealthy students, our model will not necessarily represent the typical student experience (though it should represent the loan/grant receiving student experience well).

Academics

The 38 program variables (PCIP##), which each correspond to the percentage of degrees awarded in a particular subject, were used to group colleges by academic programs. A 2002 subset of college data was used. When matched with the available 2012 earnings data, 1,917 observations of individual college campuses were available.

Admissions

The admit rate, Median ACT, and Mean SAT were used to group schools by selectivity. Two factor variables corresponding to the availability of ACT and SAT information were also generated, so that colleges without SAT or ACT information could still be included in further analyses. Admissions data was sparse in 2002, so the 2006 subset of colleges was selected, which resulted in a matched set of 1,607 observations of individual college campuses.

Demographics

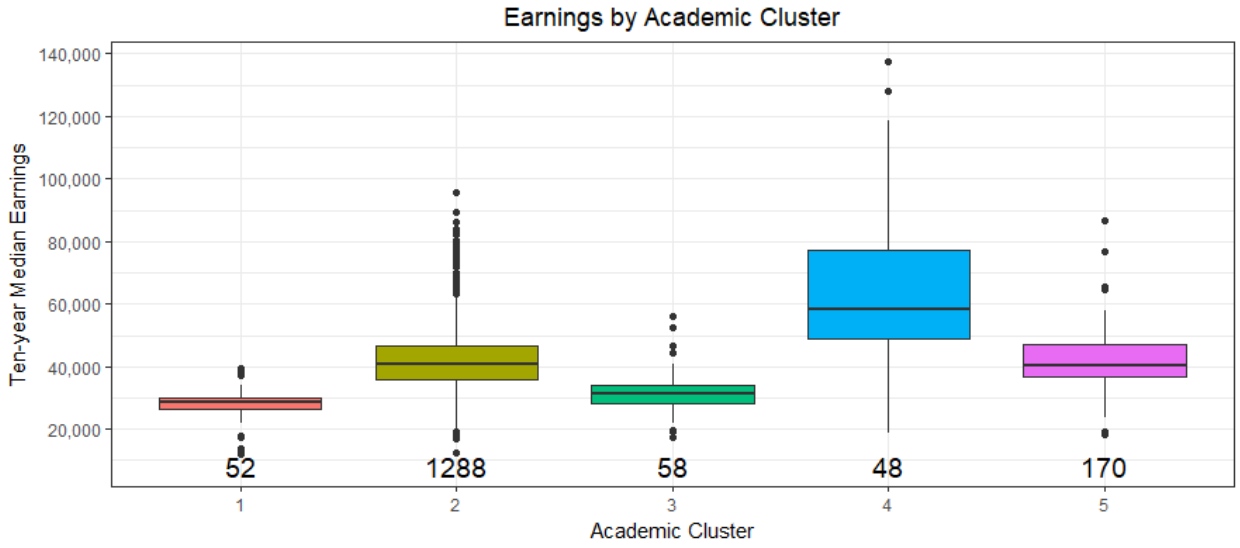
Demographic data was inconsistently collected up to about 2010, and in the several years before this there was a transition in the reporting terminology and categories. That shift was not completed by 2008, but that year still includes a substantial amount of demographic information mixed between two different reporting standards. Both standards were included in the analysis, since several comparisons suggest that this results in the best categorization and earnings predictive performance. Fourteen total variables were included, which resulted in a set of 2,055 observations of individual college campuses.

[1] The median college rate of Federal loan uptake is over 90% in most years.

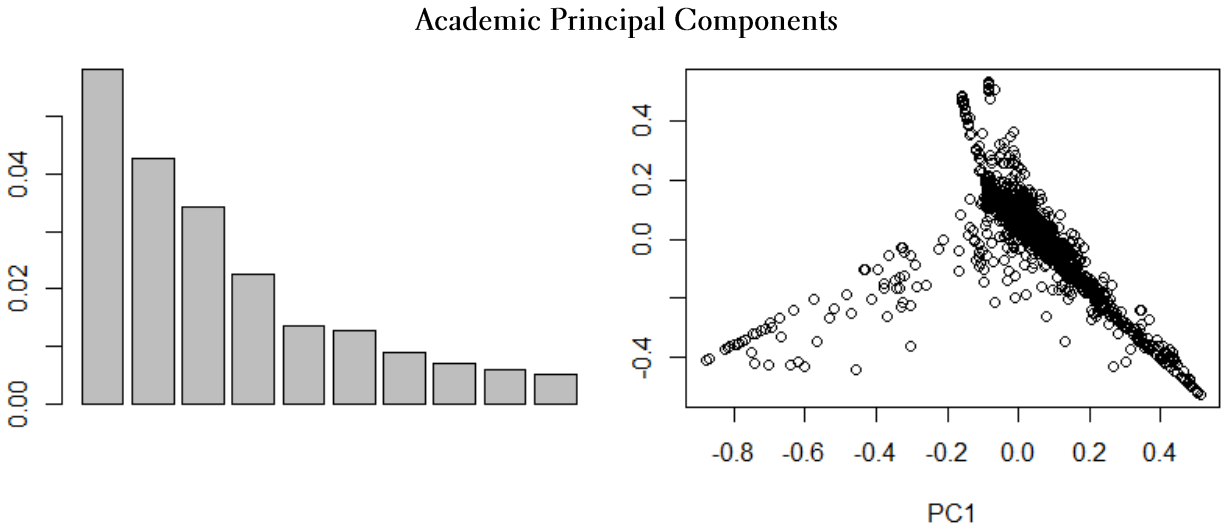
Cluster Analysis

Academics

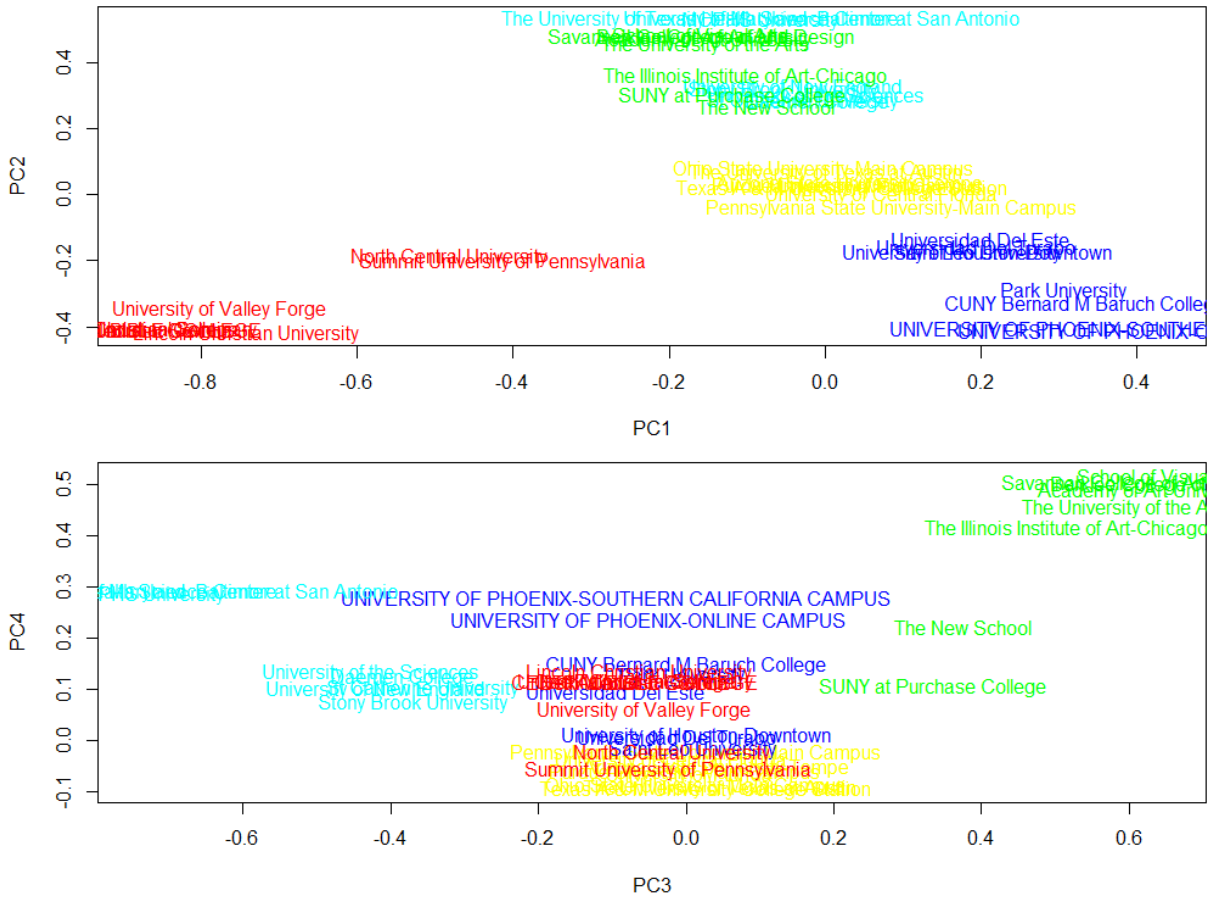
The K-Means technique was used to cluster schools into groups that shared program characteristics. Five clusters were used in order to maintain interpretability (information criteria suggested we use hundreds). Right away, we see that some groups earn more income than others in the chart below. Note that the total membership in each group is listed at the bottom of the chart.



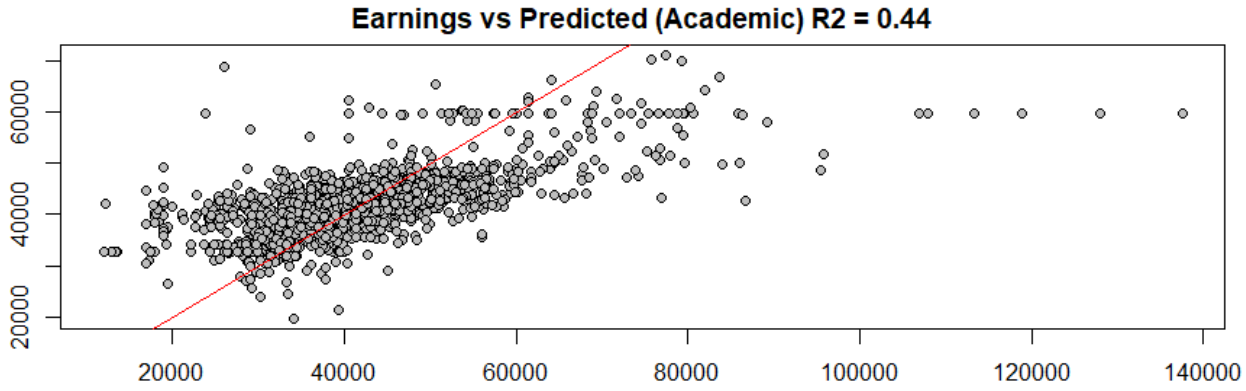
Next, to interpret these groups, a principal component analysis was performed on the academic variables. There is a definite decline in explained variance after the first few components.



Next, we plotted the largest colleges in each group against the principal components to get a sense of their differences, and to help identify the characteristics of these groups. This analysis revealed the following group definitions: Group 1: Religious (Red), Group 2: Generalist (Yellow), Group 3: Audio/Visual and Performing Arts (Green), Group 4: Medical and Technical (Light Blue), Group 5: Non-Traditional Student (Blue).



With this done, a preliminary cv.lasso regression was performed to validate the hypothesis that earnings can be predicted.

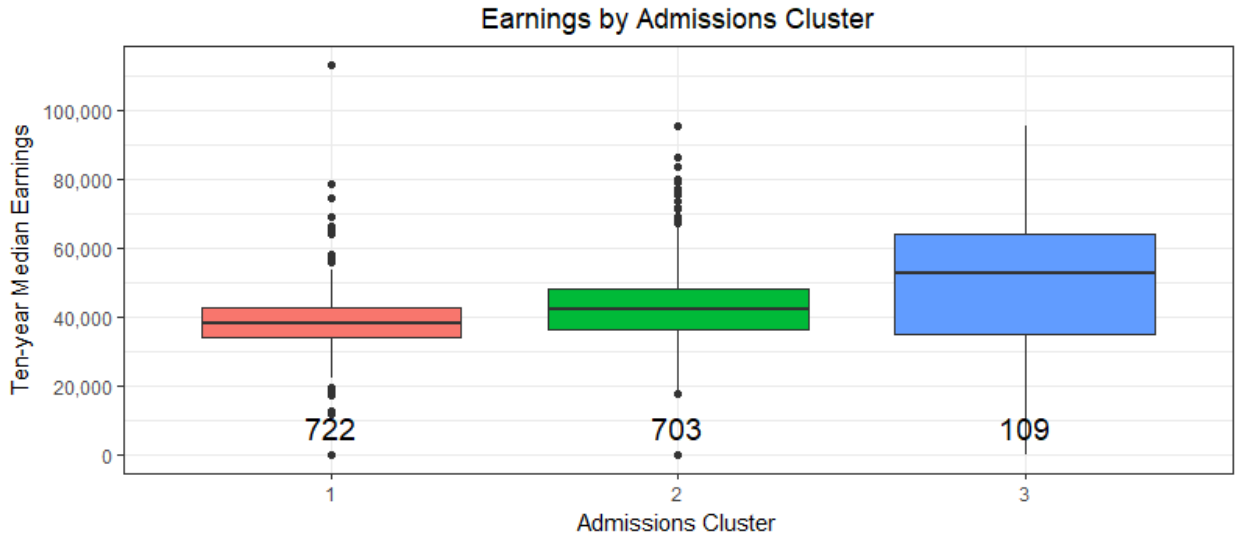


While there are some prediction artifacts caused by schools with very similar academic programs, it is clear that the program mix at a college is strongly predictive of eventual earnings. Next, the admissions selectivity was evaluated.

Admissions

After a number of attempts to use statistical programs to generate meaningful groups from admissions criteria, an intuitive approach was substituted. Three groups ("not selective," "selective," and "very selective") were generated based on the following rules: if a college admissions rate was below the 35th percentile, OR their mean SAT was above the 65th percentile, OR their median ACT was above the 65th percentile, the school was marked as "selective." If the same was true for the 5th percentile, and 95th percentiles respectively, the school was marked as "very selective."

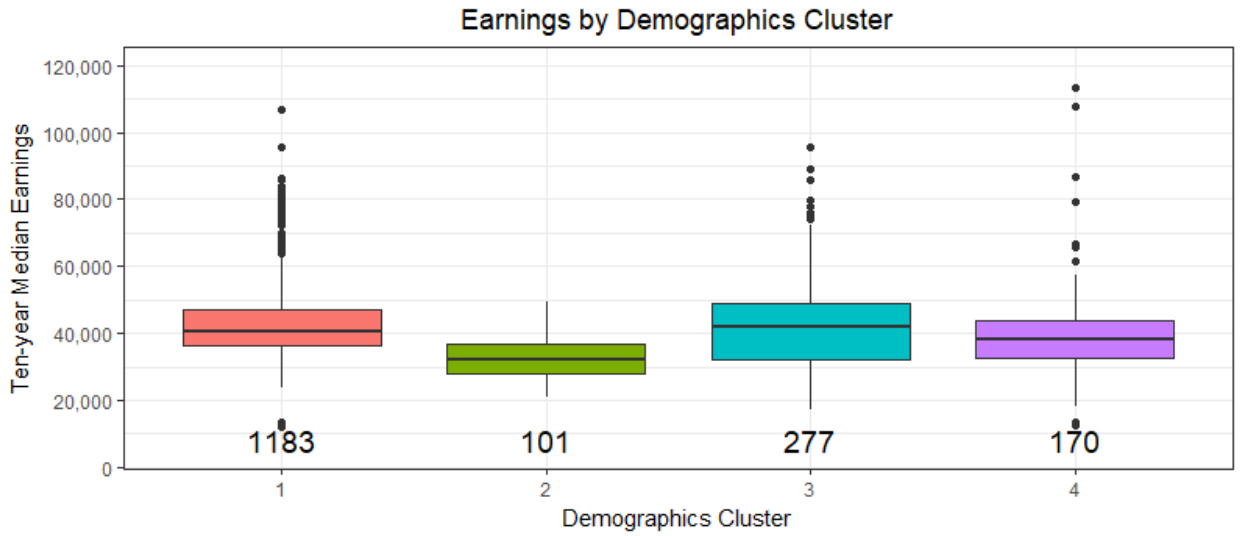
These groups should still be meaningful for students since they correspond closely to a student's academic aspirations and performance in high school. Further, there is a strong conditional relationship with earnings.



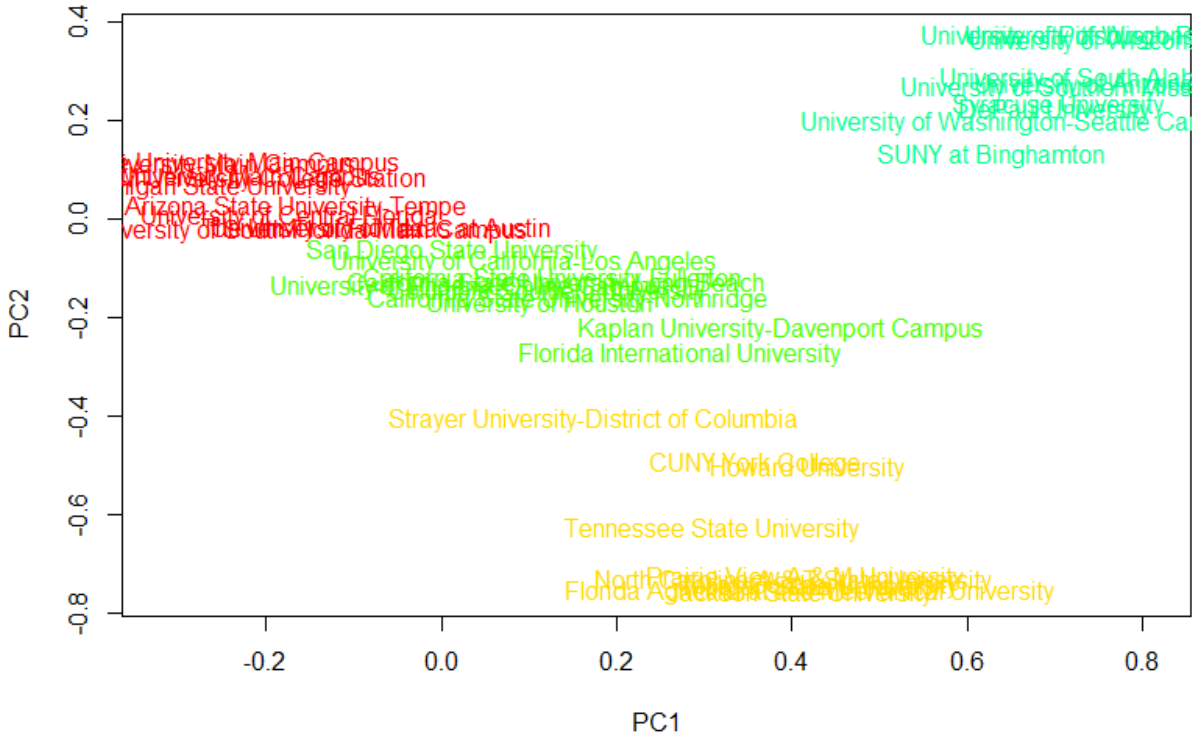
The principal components analysis was not meaningful for these data, since the admissions rate and test scores are so closely correlated. That said, these data are still useful in predicting earnings in an ensemble with the other data sets.

Demographics

The K-means analysis was meaningful for these data. The conditional distribution shows a strong relationship with earnings.



Next, we plotted the largest colleges in each group against the principal components to get a sense of their differences, and to help identify the characteristics of these groups.

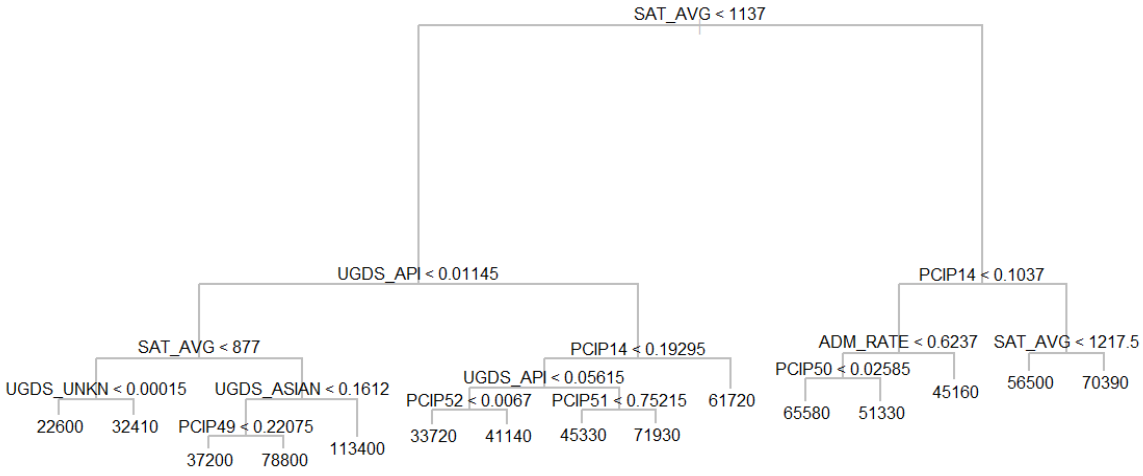


This analysis revealed the following group definitions: Group 1 (Red) and Group 4 (Light Blue): High White Group 2: High Black (Yellow), Group 3: High Hispanic or Asian (Green). PC2 might be defined as a "whiteness quotient," while PC1 is an artifact of the two variable types included in this data set.

B. Earnings Prediction and Analysis

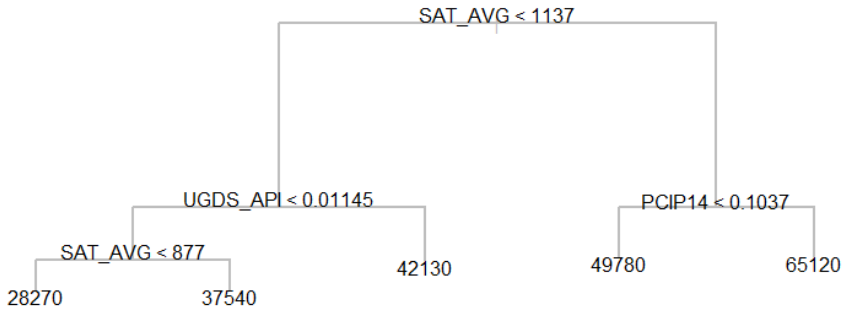
In order to predict earnings using all of the inputs from the previous three clusters analyses, colleges with incomplete data were removed. This resulted in a data set of 1,411 observations of individual college campuses. Once that was completed, a tree analysis was performed to get a sense for the most important drivers of earnings at colleges.

Unpruned Tree

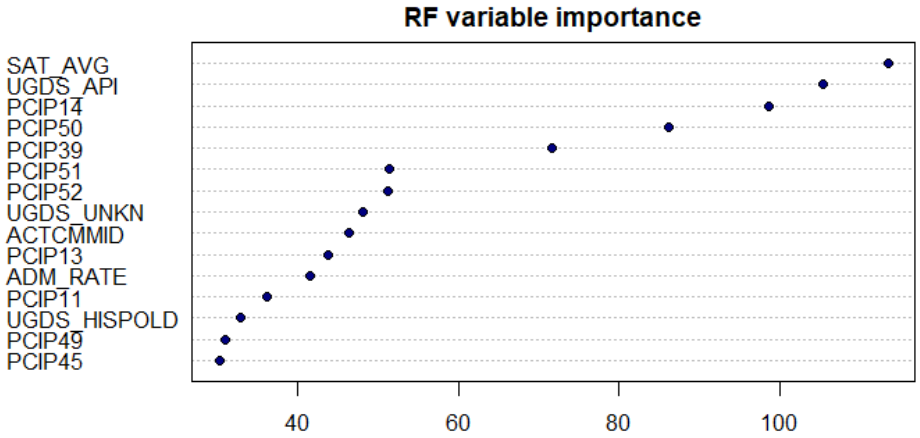


Here, we see that a number of variables are important, but none more so than the SAT_AVG at the school. The proportion of Asians at a school (UGDS_API) and the proportion of Engineering Degrees (PCIP14) are also quite relevant. A cross-validated tree function was used to prune this tree. While the lowest deviance occurred at the largest size, a 5-leaf tree was nearly as good.

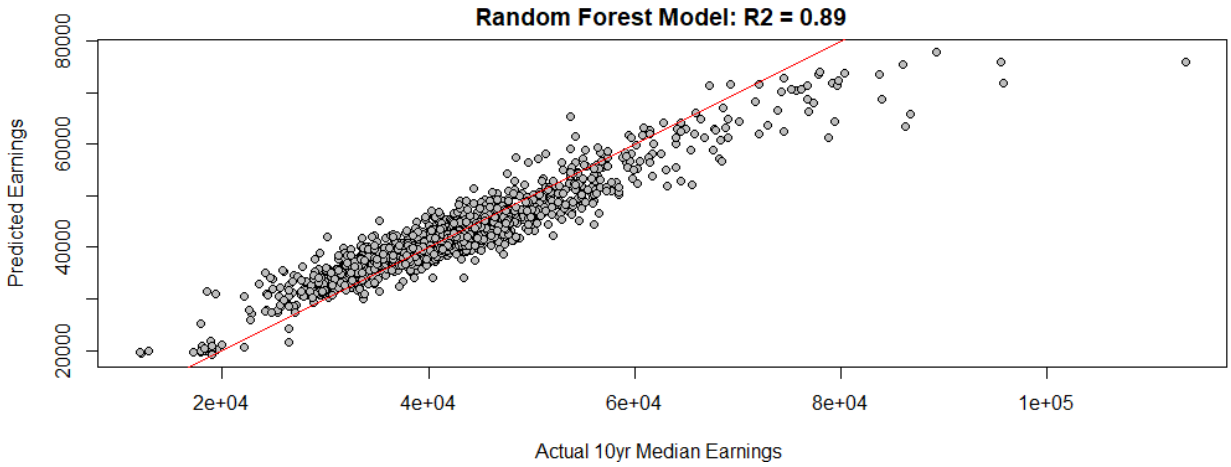
Unpruned Tree



Finally, a 25,000-tree random forest (node size 20) was used to estimate the earnings of each school. These variables were estimated to be important based on that algorithm.



As before, the Average SAT, Asian proportion, and Engineering proportion are strongly predictive. Further, the "Theology And Religious Vocations" proportion (PCIP39) and the "Visual And Performing Arts" proportion (PCIP50) proportion are strongly predictive of earnings (negatively, as it turns out). We can now compare the random tree prediction performance against the actual reported earnings.



With all of the variables included, the random forest very closely predicts the median earnings at a particular college after 10 years. That said, there is still a bit of skew in the prediction - the model overestimates the earnings of schools at the bottom of the distribution, and underestimates the income of schools at the top of the earnings distribution. It is possible that confounding variables or signaling effects exacerbate the innate ability of a college to assist graduates in earning a salary.

C. Group Selection and Evaluation

In this section, we'll consider three different students with contrasting goals to show how this data can be useful in helping a student chose a college.

Student 1

This is a first-generation black student with a relatively strong academic focus, but they're not quite sure what degree they want yet. This student wants to choose Academic Group 2 (Generalist), Admissions Group 2 (selective), and Demographic Group 2 (high black). The schools that overperform their prediction most in this group are below. A number of HBCUs are part of this list, which reinforces the idea that they may be the right fit for some.

Institution	Median Earnings	Residual
Lincoln University	$33,500	$2,426
Hampton University	$41,200	$2,350
Howard University	$46,000	$1,113
Morehouse College	$41,600	$847

Student 2

This is a white student that really wants some diversity, and has an extremely strong academic background. They're pretty sure they'd like to be an engineer, but they're open to medicine as well. This student wants to choose Academic Groups 2 and 4 (Generalist and Medicine/Technical), Admissions Groups 2 and 3 (selective and highly selective), and Demographic Group 3 (high Hispanic and Asian). The schools that overperform their prediction most in this group are below. Some of the usual suspects appear, but we also see the Univ. of the Pacific, which manages an impressive $10k more income than would be expected for an institution with its input profile.

Institution	Median Earnings	Residual
Harvard University	$95,500	$19,582
MIT	$89,200	$11,291
Univ. of the Pacific	$68,200	$10,938
Stanford University	$86,000	$10,545

Student 3

This is a Hispanic student with a passion for music, and a strong connection to their faith. They'd prefer a school which is not majority white, and believe that their academic credentials are not likely to be a strong factor in deciding where to go to school. This student wants to choose Academic Groups 1 & 3 (Religion and Arts), Admissions Groups 1 and 2 (not selective and selective), and Demographic Groups 2 & 3 (high Black and High Hispanic/Asian). In this case, the Arts schools win out, with the Art Center College of Design in Pasadena, California showing the best residual income.

Institution	Median Earnings	Residual
Art Center College of Design	$56,100	$11,768
Otis College of Art and Design	$46,800	$5,642
The Illinois Institute of Art-Schaumburg	$33,300	$1,497
Savannah College of Art and Design	$35,100	$1,278

D. Further work

The analysis performed above may be extended to evaluate an individual student's ideal schools, based on family income group, demographics, completion rates, and other variables.

6. Do schools that target specific types of students (e.g., historically black colleges and universities, women's, and men's-only colleges) result in better outcomes for those students?

Another issue that interested us was whether "specialty" colleges serve their students more effectively than institutions that admitted all types of students. Here, we define "specialty" colleges as those which are focused on maintaining a particular sort of student demographic. We were motivated to explore this question for a few reasons. First, women's colleges have faced increasing pressure over the past few decades to become coeducational. Additionally, historically black colleges and universities (HBCUs) have been featured quite a bit in the media as alternative pathway to success for black students. Unfortunately, the data on women's only colleges is quite limited – there are only 38 women-only colleges, and most of them have begun accepting men. Therefore, we chose to focus our analysis upon HBCUs exclusively.

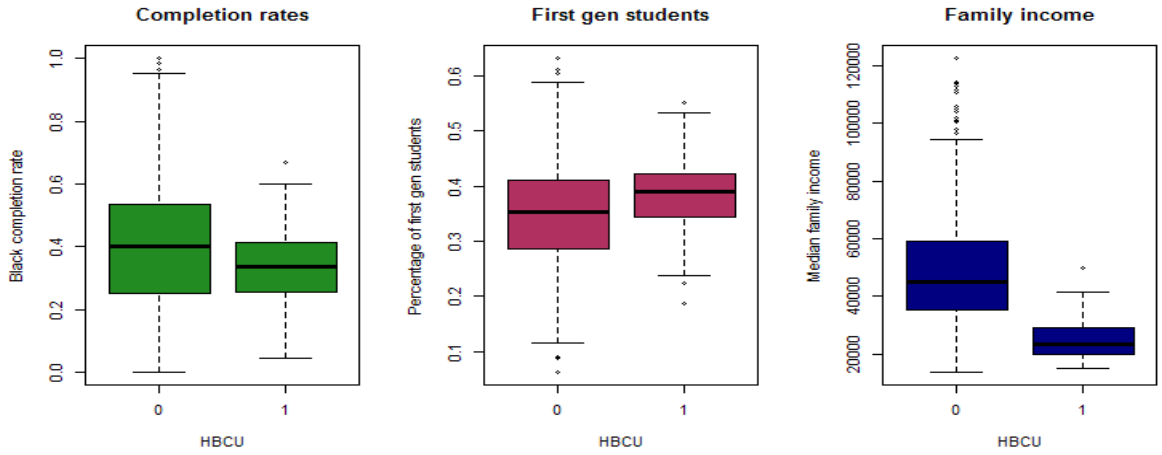
A. Data Preparation

We restricted our analysis to the year 2014, which provided the most recent and complete set of data. Since many of the original 1600 variables were rather duplicative, we manually narrowed down the set of potential explanatory variables to approximately 70 variables that measured demographic, financial, student outcomes, and admissions statistics. Here, we had to make a trade-off between data availability and the inclusion of variables in the dataset – many variables that would have been interesting to include had very low reporting rates among the schools in the sample.

We chose C150_4_BLACK to be our dependent variable – this refers to the percentage of black students who graduated within 6 years at that college or university. We then removed observations from the dataset that did not report on the variables we chose to include. Finally, we chose to remove non-HBCUs from the sample set for which black students accounted for less than 4% of the student body, because we felt that colleges below this threshold would not be able to provide very credible evidence on their ability to graduate black students. Eventually, we were able to narrow the dataset down to a sample that included 60 HBCUs and 957 non-HBCUs.

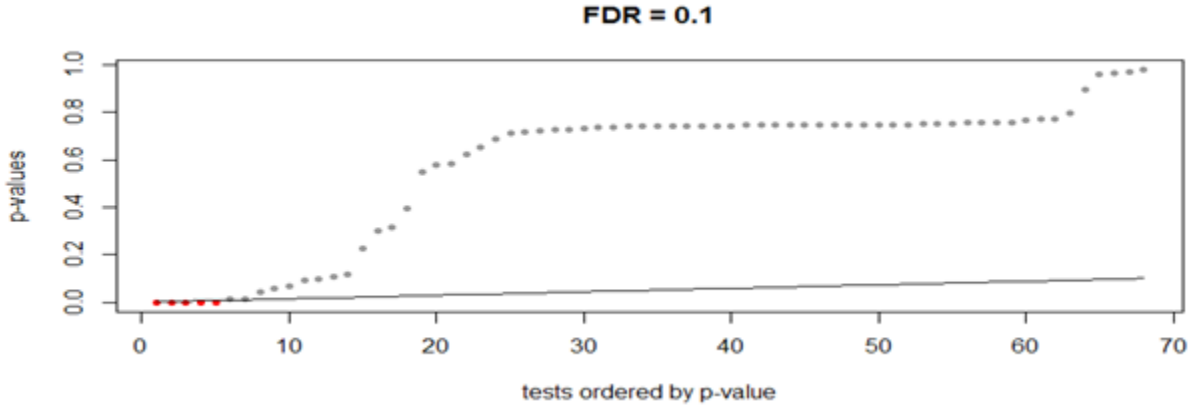
B. Analysis

At first glance, HBCUs appear to perform worse than the average non-HBCU in terms of graduating students on time. However, HBCUs also appear to admit many more low-income and first-generation black college students than the average non-HBCU:

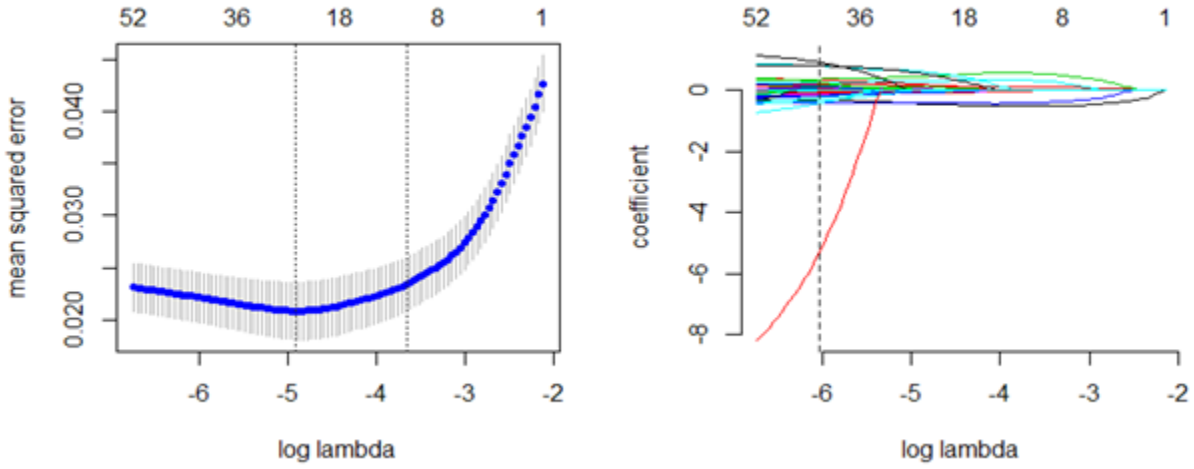


Based on this information, our initial hypothesis follows: when controlling for factors such as number of first generation college students, family income, and other demographic/financial factors, HBCUs perform better than non-HBCUs when graduating black students on time. We decided to structure our analysis by first using OLS, LASSO, and Random Forests techniques to assess what factors most impact black student completion rates at higher education institutions. We then conducted a causal LASSO, using the HBCU indicator as a treatment effect, to assess the causal effect of a college designated as an "HBCU."

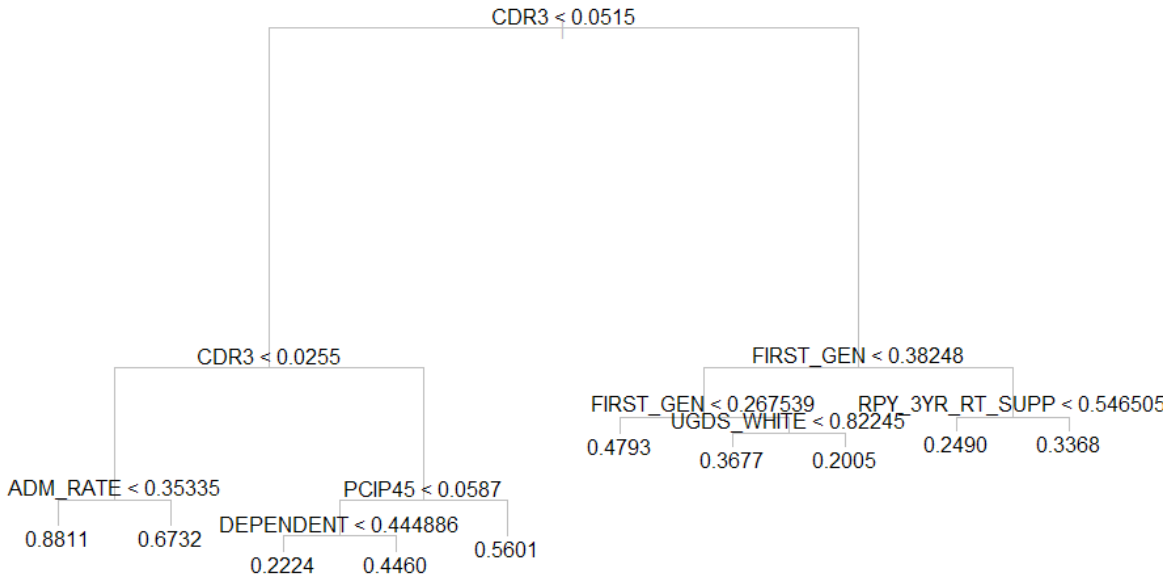
Our first OLS regression, run on all 70 variables, had an in-sample R^2 value of 0.58 – quite high, considering the many idiosyncratic issues that affect college outcomes. We then ran an FDR cut (setting q= 0.1) to reduce the risk of false discovery, and found that relatively few variables were significant enough to fall below this threshold: Only admit rate (ADM_RATE), the number of first generation students (FIRST_GEN), the number of undergraduates (UGDS), the share of part-time students (PPTUG_EF), and the three-year repayment rate (RPY_3YR_RT_SUPP) were found to be significant using this method. (See Appendix for a full list of p-values for this and other regressions conducted in this analysis.) This restricted regression had an in-sample R^2 value of 0.44.



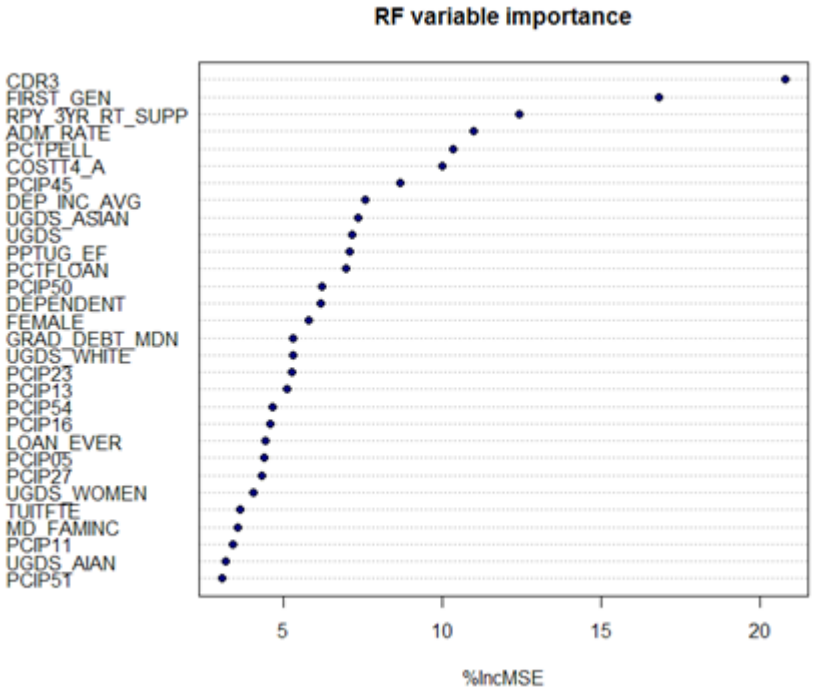
We then ran a LASSO regression, using cross-validation to select the coefficients (using the 1se lambda). This analysis found 12 coefficients to be significant, and had an R^2 value of approximately 0.5. The HBCU indicator, once again, was not found to be a significant coefficient.



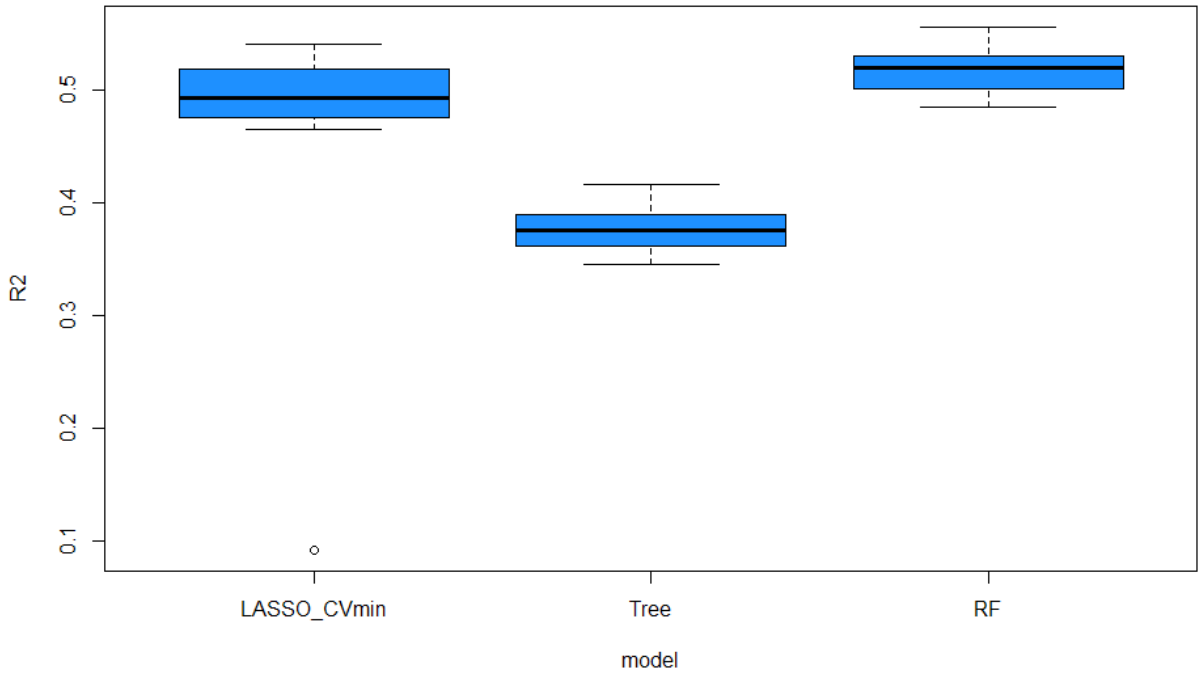
Next, we employed both a regression tree (pruned using the CART algorithm) and random forest techniques to predict black completion. The pruned tree (size ten), is shown below:



In short, the CART model places high importance on some of the same variables as the previous regressions, but also considers CDR3 (the 3-year default rate of graduates) to be extremely important in predicting black completion. This makes sense, because a low default rate is generally indicative of a school that prepares its graduates for success. The in-sample R^2 for this technique is approximately 0.52. The random forest model also considers CDR3 to be extremely important:

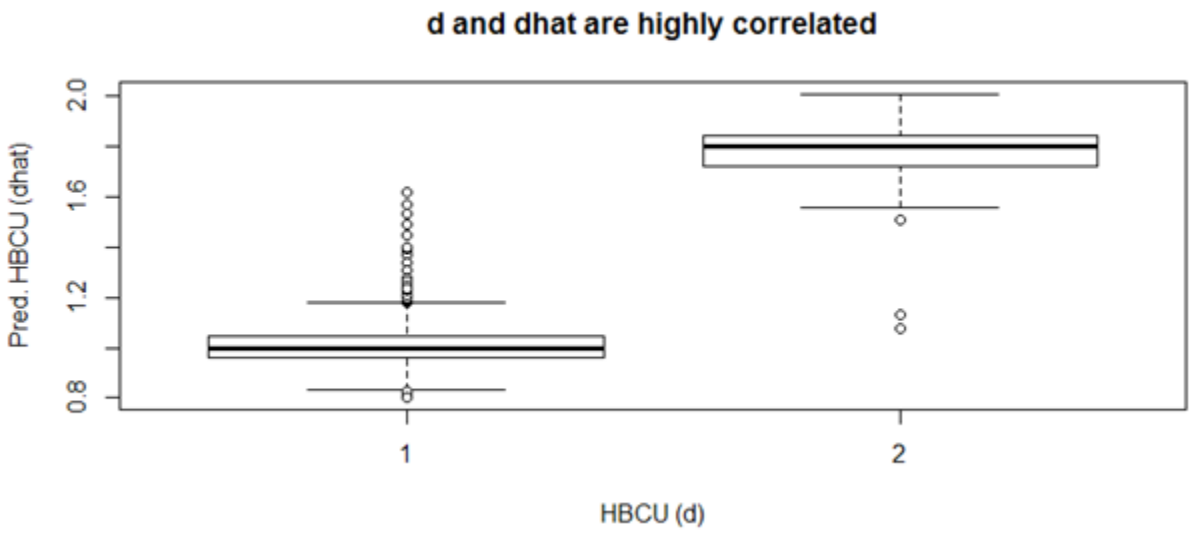


The in-sample R^2 for this model is extremely high – 0.79. Therefore, we decided to estimate out-of-sample R^2 for the LASSO, full tree, and random forest models. We did this by conducting a cross-validation exercise (as in section three) in which we ran each of the 3 models 20 times, randomly splitting the data into a "training" and "test" dataset each time (similar to our first research question in Section III). The R^2 values calculated for each of the three models is summarized below:



The random forest model appears to perform consistently better than the other two models, and is hence the most effective at predicting black student outcomes. None of the three models consider "HBCU" to be a significant variable when predicting these outcomes.

Finally, we conducted a causal LASSO to assess whether designating a college an HBCU had an effect upon black student outcomes. When we left "d" (the indicator of whether a college as an HBCU) unpenalized, the effect of an HBCU was measured as 0.04 percentage points – in short, an HBCU designation had a very mildly positive effect upon black student outcomes. However, when we penalized d like other variables in the analysis, the effect of d was zero. This is likely because the HBCU designation was highly correlated to the other explanatory variables in the analysis, such as the percentage of black undergraduates in the school (R^2(d, dhat) = 0.76).



Therefore, when controlling for student demographics and other factors, merely designating a college as an HBCU does not appear to improve black student outcomes relative to other schools.

C. Conclusion

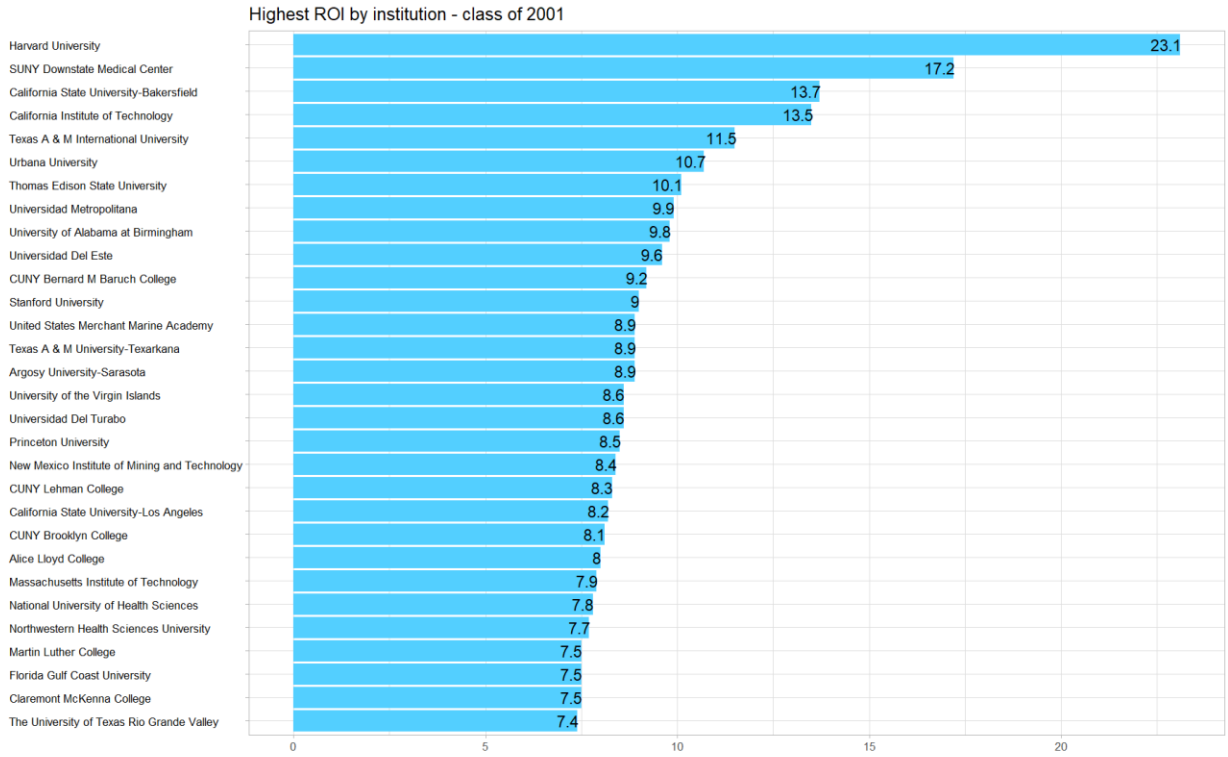
In short, it appears that HBCUs do not appear to have a significant impact upon black student outcomes in the United States, after controlling for other factors such as the percentage of black students in the undergraduate school's population, and the number of first generation students. Unsurprisingly, the schools that appear to be the most selective, and are good at placing their students into high-paying careers, appear to be best at guaranteeing good outcomes for black students.

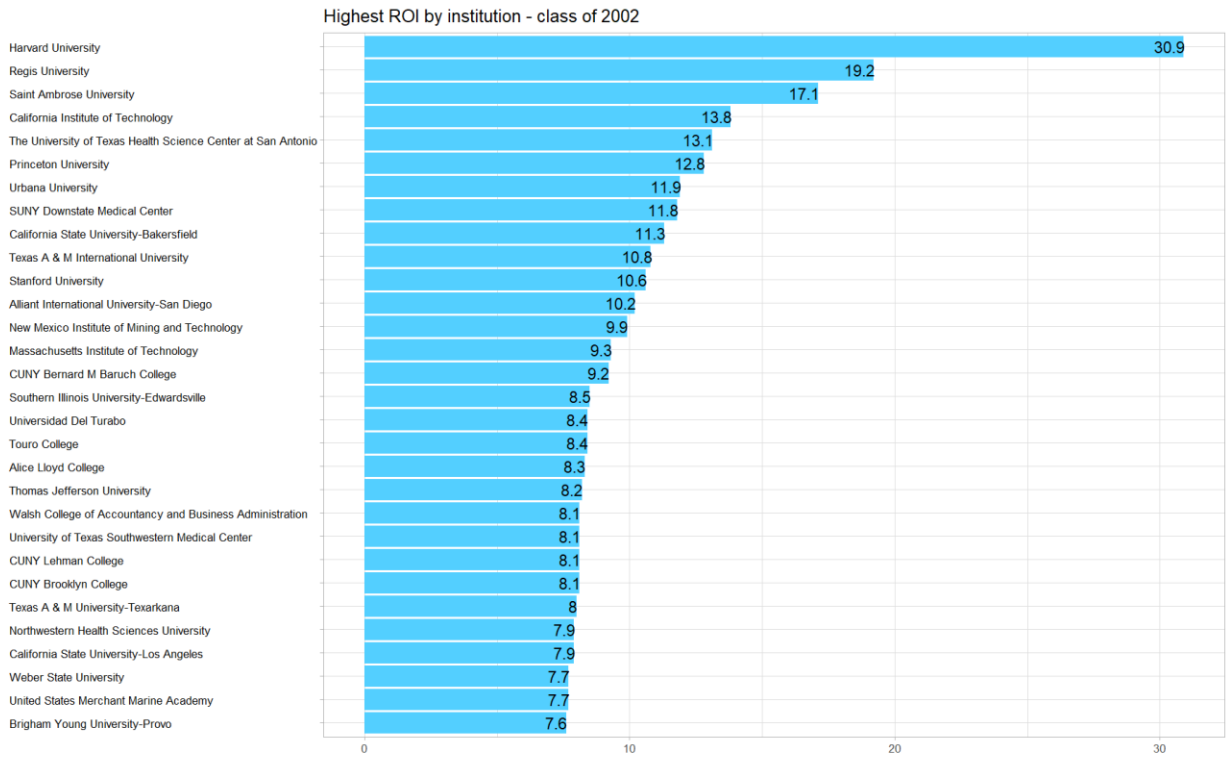
7. What schools/degree programs provide the best return on investment?

We were interested to learn what is the return on investment for the different undergraduate programs. In order to investigate this question, we choose to look at median debt upon graduation, as an indicator for investment, and median earnings after 10 years, as an indicator for return.

A. Data preparation

We restricted our analysis to the years 2001 and 2002 for debt, and 2011 and 2012 for earnings. This was due to the fact that 10-year earnings data was available only for the years 2011 and 2012. The ROI estimator we created is a deviation of median earnings by median debt for each class (i.e. med earnings 2011 / med debt 2001). The tables below show the top 30 schools' ROI for the class of 2001 and 2002. The results were not surprising and reflect higher ROI for IVY league schools as well as medical and engineering schools. we can also see a trend where the ROI increases over the years.

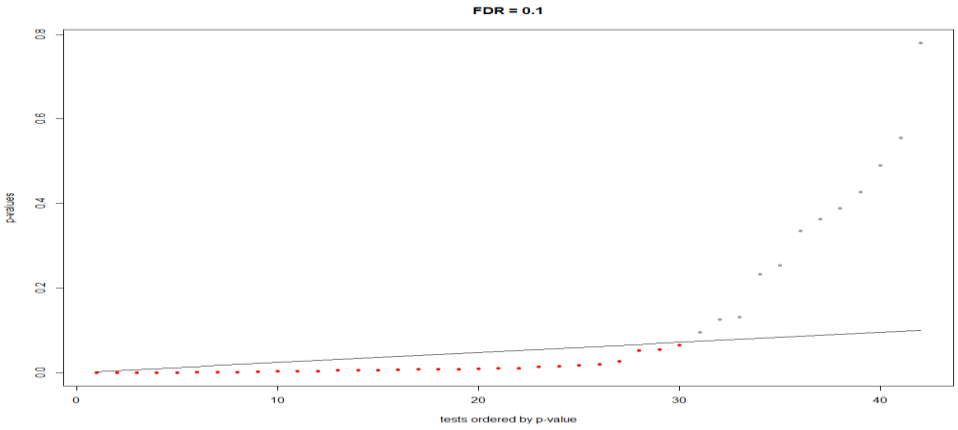




For our regression analysis, we choose to focus on the class 2002 ROI since it had more observations. We chose ROI to be our y-variable and removed observations with no debt or earning information. We created a subset with variables related to topic learned (engineering/medical/etc.), family income, first generation students, tuition, and completion rate.

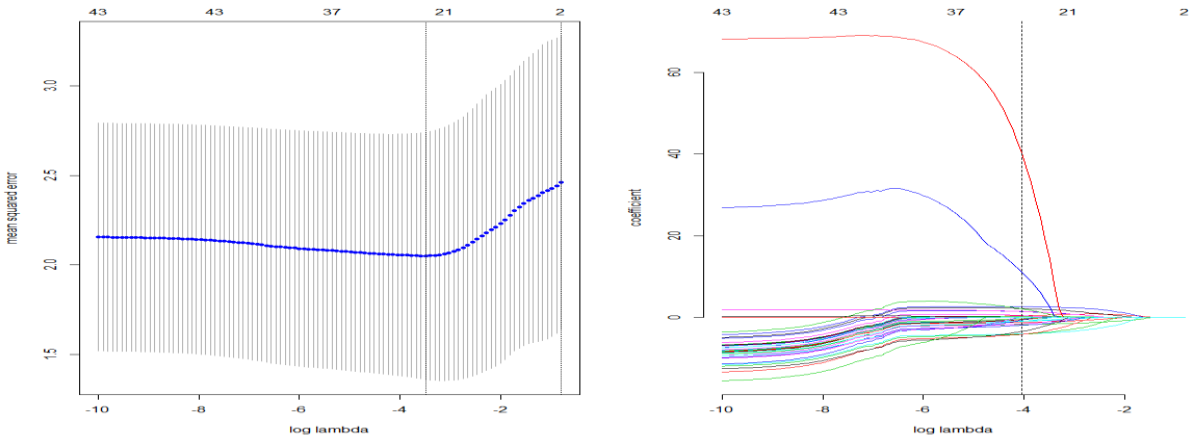
B. Analysis

Our first OLS regression, run on 42 variables, had an R^2 value of 0.26. We then ran an FDR cut (setting q= 0.1) to reduce the risk of false discovery, we found out that most variables were significant - 30 out of 42. These variables include the tuition cost, SAT score, completion rate, average family income and different degree types. This restricted regression had an R^2 value of 0.245 but has only 13 significant coefficients. The significant coefficients are presented in the table below.



	Estimate	Std. Error	t value	Pr(>|t|)	
PCIP01	-3.88E+00	2.31E+00	-1.675	0.094307	Percentage of degrees awarded in Agriculture, Agriculture Operations, And Related Sciences.
PCIP09	-2.35E+00	1.34E+00	-1.756	0.079377	Percentage of degrees awarded in Communication, Journalism, And Related Programs.
PCIP13	-2.64E+00	9.25E-01	-2.851	0.004448	Percentage of degrees awarded in Education.
PCIP23	-6.42E+00	2.10E+00	-3.057	0.002297	Percentage of degrees awarded in English Language And Literature/Letters.
PCIP26	-5.74E+00	1.61E+00	-3.571	0.000374	Percentage of degrees awarded in Biological And Biomedical Sciences.
PCIP31	-5.80E+00	1.94E+00	-2.991	0.002853	Percentage of degrees awarded in Parks, Recreation, Leisure, And Fitness Studies.
PCIP38	-3.49E+00	1.75E+00	-1.998	0.046011	Percentage of degrees awarded in Philosophy And Religious Studies.
PCIP39	-2.44E+00	9.23E-01	-2.643	0.008358	Percentage of degrees awarded in Theology And Religious Vocations.
PCIP50	-3.13E+00	8.98E-01	-3.49	0.000505	Percentage of degrees awarded in Visual And Performing Arts.
FAMINC	-2.77E-05	3.89E-06	-7.129	1.98E-12	Average family income
SAT_AVG	5.95E-03	7.24E-04	8.224	6.32E-16	Average SAT score
C150_4	1.80E+00	4.72E-01	3.819	0.000142	Completion rate for first-time, full-time students at four-year institutions
TUITIONFEE_IN	-2.61E-05	8.60E-06	-3.041	0.002423	In-state tuition and fees

We then ran a LASSO regression, using cross-validation to select the coefficients. This analysis found 27 coefficients to be significant, which include tuition cost, SAT score, completion rate, average family income and different degree types. The significant coefficients are presented in the table below.



	Estimate	
PCIP01	-9.14E-02	Percentage of degrees awarded in Agriculture, Agriculture Operations, And Related Sciences.
PCIP04	-1.96E+00	Percentage of degrees awarded in Architecture And Related Services.
PCIP03	-1.43E+00	Percentage of degrees awarded in Communication, Journalism, And Related Programs.
PCIP09	-6.53E-01	Percentage of degrees awarded in Communication, Journalism, And Related Programs.
PCIP11	2.51E+00	Percentage of degrees awarded in Computer And Information Sciences And Support Services.
PCIP13	-1.07E+00	Percentage of degrees awarded in Education.
PCIP14	5.58E-01	Percentage of degrees awarded in Engineering.
PCIP22	-1.84E+00	Percentage of degrees awarded in Legal Professions And Studies.
PCIP23	-3.19E+00	Percentage of degrees awarded in English Language And Literature/Letters.
PCIP24	2.31E-01	Percentage of degrees awarded in Liberal Arts And Sciences, General Studies And Humanities.
PCIP25	1.41E+01	Percentage of degrees awarded in Library Science.
PCIP26	-3.63E+00	Percentage of degrees awarded in Biological And Biomedical Sciences.
PCIP27	7.18E-01	Percentage of degrees awarded in Mathematics And Statistics.
PCIP30	1.13E+00	Percentage of degrees awarded in Multi/Interdisciplinary Studies.
PCIP31	-3.81E+00	Percentage of degrees awarded in Parks, Recreation, Leisure, And Fitness Studies.
PCIP38	-1.07E+00	Percentage of degrees awarded in Philosophy And Religious Studies.
PCIP39	-7.61E-01	Percentage of degrees awarded in Theology And Religious Vocations.
PCIP43	8.21E-02	Percentage of degrees awarded in Homeland Security, Law Enforcement, Firefighting And Related Protective Services.
PCIP44	-1.30E-03	Percentage of degrees awarded in Public Administration And Social Service Professions.
PCIP45	1.88E+00	Percentage of degrees awarded in Social Sciences.
PCIP46	2.49E+00	Percentage of degrees awarded in Construction Trades.
PCIP50	-1.65E+00	Percentage of degrees awarded in Visual And Performing Arts.
FAMINC	-2.23E-05	Average family income
SAT_AVG	5.24E-03	Average SAT score
C150_4	1.15E+00	Completion rate for first-time, full-time students at four-year institutions
TUITIONF	-1.60E-05	In-state tuition and fees

C. Conclusion

It appears that factors like family income, degree type, completion rate, SAT scores and tuition cost can predict the return on higher education. Our OLS model help to predict mainly the degree types which will lead to lower ROI (Education, English, Biology). Our CV model also show some degree types that will predict higher ROI (Engineering, Math, Homeland Security). We can see that SAT score and completion rate have significant positive effect in both models. The one surprising outcome is Average Family Income which has negative effect on the ROI. This may be occurring due to interactions with another variable, or it may simply show that students from wealthy families may be more comfortable choosing “unprofitable” majors, such as Gender Studies.

8. Appendices

A. Appendix 1: Data Dictionary

The most common variables used in the analysis are as follows:

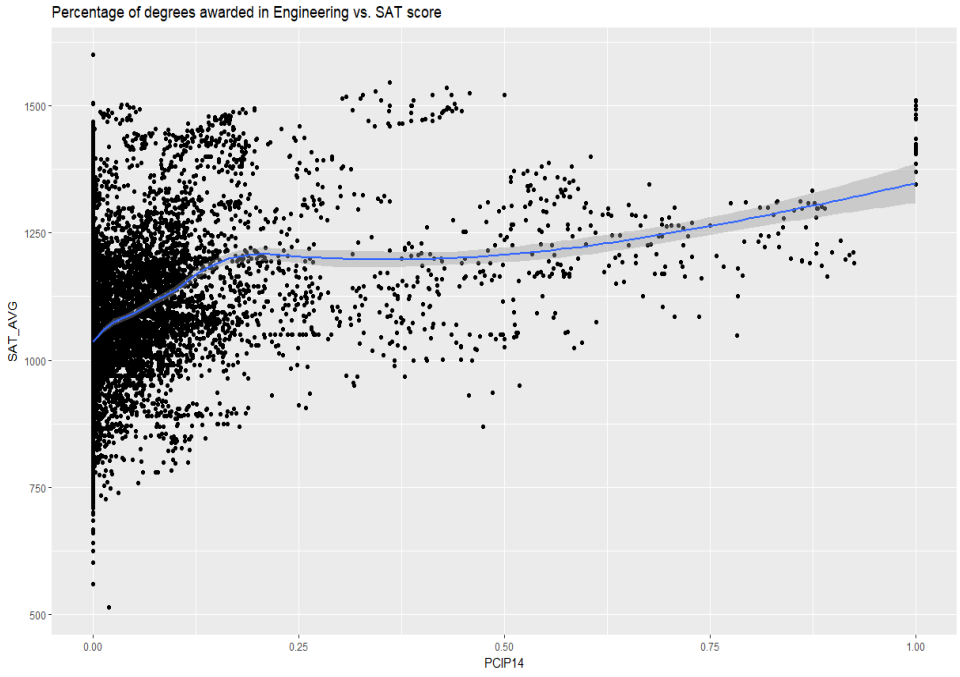
VARIABLE NAME	NAME OF DATA ELEMENT	dev-category	Data type
OPEID	8-digit OPE ID for institution	root	character
INSTNM	Institution name	school	character
CITY	City	school	character
STABBR	State postcode	school	character
MAIN	Flag for main campus	school	factor
CONTROL	Control of institution	school	factor
HBCU	Flag for Historically Black College and University	school	factor
ADM_RATE	Admission rate	admissions	numeric
SATVR25	25th percentile of SAT scores at the institution (critical reading)	admissions	numeric
SATVR75	75th percentile of SAT scores at the institution (critical reading)	admissions	numeric
SATMT25	25th percentile of SAT scores at the institution (math)	admissions	numeric
SATMT75	75th percentile of SAT scores at the institution (math)	admissions	numeric
SATWR25	25th percentile of SAT scores at the institution (writing)	admissions	numeric
SATWR75	75th percentile of SAT scores at the institution (writing)	admissions	numeric
SATVRMID	Midpoint of SAT scores at the institution (critical reading)	admissions	numeric
SATMTMID	Midpoint of SAT scores at the institution (math)	admissions	numeric
SATWRMID	Midpoint of SAT scores at the institution (writing)	admissions	numeric
PCIP01	Percentage of degrees awarded in Agriculture, Agriculture Operations, And Related Sciences.	academics	numeric
PCIP03	Percentage of degrees awarded in Natural Resources And Conservation.	academics	numeric
PCIP04	Percentage of degrees awarded in Architecture And Related Services.	academics	numeric
PCIP05	Percentage of degrees awarded in Area, Ethnic, Cultural, Gender, And Group Studies.	academics	numeric
PCIP09	Percentage of degrees awarded in Communication, Journalism, And Related Programs.	academics	numeric
PCIP10	Percentage of degrees awarded in Communications Technologies/Technicians And Support Services.	academics	numeric
PCIP11	Percentage of degrees awarded in Computer And Information Sciences And Support Services.	academics	numeric
PCIP12	Percentage of degrees awarded in Personal And Culinary Services.	academics	numeric
PCIP13	Percentage of degrees awarded in Education.	academics	numeric
PCIP14	Percentage of degrees awarded in Engineering.	academics	numeric
PCIP15	Percentage of degrees awarded in Engineering Technologies And Engineering-Related Fields.	academics	numeric
PCIP16	Percentage of degrees awarded in Foreign Languages, Literatures, And Linguistics.	academics	numeric
PCIP19	Percentage of degrees awarded in Family And Consumer Sciences/Human Sciences.	academics	numeric
PCIP22	Percentage of degrees awarded in Legal Professions And Studies.	academics	numeric
PCIP23	Percentage of degrees awarded in English Language And Literature/Letters.	academics	numeric
PCIP24	Percentage of degrees awarded in Liberal Arts And Sciences, General Studies And Humanities.	academics	numeric
PCIP25	Percentage of degrees awarded in Library Science.	academics	numeric
PCIP26	Percentage of degrees awarded in Biological And Biomedical Sciences.	academics	numeric
PCIP27	Percentage of degrees awarded in Mathematics And Statistics.	academics	numeric
PCIP29	Percentage of degrees awarded in Military Technologies And Applied Sciences.	academics	numeric
PCIP30	Percentage of degrees awarded in Multi/Interdisciplinary Studies.	academics	numeric
PCIP31	Percentage of degrees awarded in Parks, Recreation, Leisure, And Fitness Studies.	academics	numeric

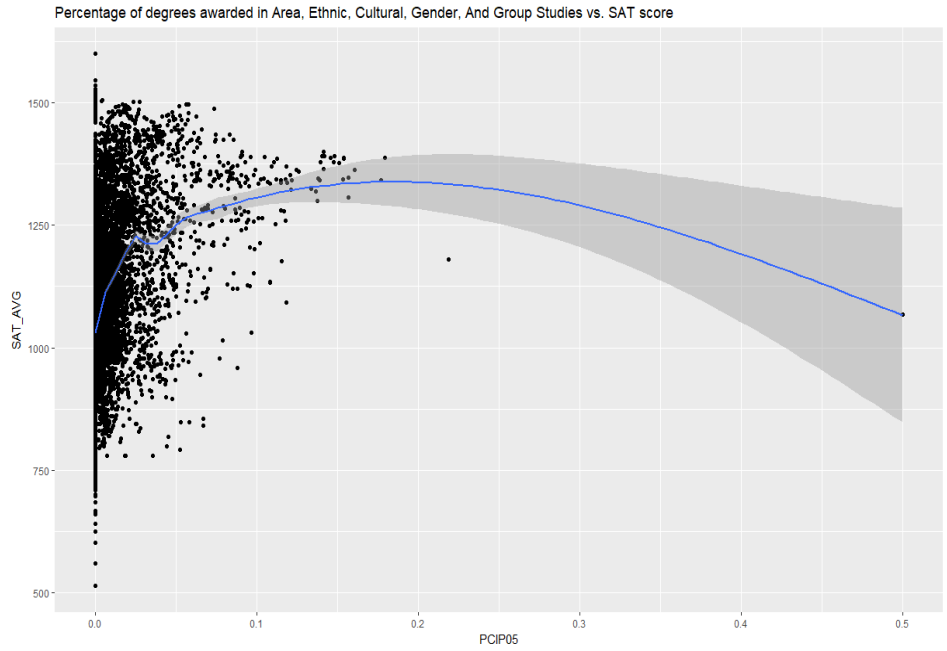
VARIABLE NAME	NAME OF DATA ELEMENT	dev-category	Data type
PCIP38	Percentage of degrees awarded in Philosophy And Religious Studies.	academics	numeric
PCIP39	Percentage of degrees awarded in Theology And Religious Vocations.	academics	numeric
PCIP40	Percentage of degrees awarded in Physical Sciences.	academics	numeric
PCIP41	Percentage of degrees awarded in Science Technologies/Technicians.	academics	numeric
PCIP42	Percentage of degrees awarded in Psychology.	academics	numeric
PCIP43	Percentage of degrees awarded in Homeland Security, Law Enforcement, Firefighting And Related Protective Services.	academics	numeric
PCIP44	Percentage of degrees awarded in Public Administration And Social Service Professions.	academics	numeric
PCIP45	Percentage of degrees awarded in Social Sciences.	academics	numeric
PCIP46	Percentage of degrees awarded in Construction Trades.	academics	numeric
PCIP47	Percentage of degrees awarded in Mechanic And Repair Technologies/Technicians.	academics	numeric
PCIP48	Percentage of degrees awarded in Precision Production.	academics	numeric
PCIP49	Percentage of degrees awarded in Transportation And Materials Moving.	academics	numeric
PCIP50	Percentage of degrees awarded in Visual And Performing Arts.	academics	numeric
PCIP51	Percentage of degrees awarded in Health Professions And Related Programs.	academics	numeric
PCIP52	Percentage of degrees awarded in Business, Management, Marketing, And Related Support Services.	academics	numeric
PCIP54	Percentage of degrees awarded in History.	academics	numeric
DISTANCEONLY	Flag for distance-education-only education	school	factor
UGDS	Enrollment of undergraduate certificate/degree-seeking students	student	numeric
UGDS_WHITE	Total share of enrollment of undergraduate degree-seeking students who are white	student	numeric
UGDS_BLACK	Total share of enrollment of undergraduate degree-seeking students who are black	student	numeric
UGDS_HISP	Total share of enrollment of undergraduate degree-seeking students who are Hispanic	student	numeric
UGDS_ASIAN	Total share of enrollment of undergraduate degree-seeking students who are Asian	student	numeric
UGDS_AIAN	Total share of enrollment of undergraduate degree-seeking students who are American Indian/Alaska Native	student	numeric
UGDS_NHPI	Total share of enrollment of undergraduate degree-seeking students who are Native Hawaiian/Pacific Islander	student	numeric
PPTUG_EF	Share of undergraduate, degree-/certificate-seeking students who are part-time	student	numeric
CURROPER	Flag for currently operating institution, 0=closed, 1=operating	school	factor
COSTT4_A	Average cost of attendance (academic year institutions)	cost	numeric
TUITFTE	Net tuition revenue per full-time equivalent student	school	numeric
PCTPELL	Percentage of undergraduates who receive a Pell Grant	aid	numeric
C150_4_BLACK	Completion rate for first-time, full-time students at four-year institutions (150% of expected time to completion/6 years) for black students	completion	numeric
PCTFLOAN	Percent of all federal undergraduate students receiving a federal student loan	aid	numeric
CDR3	Three-year cohort default rate	repayment	numeric
DEP_INC_AVG	Average family income of dependent students in real 2015 dollars.	student	numeric
GRAD_DEBT_MDN	The median debt for students who have completed	aid	numeric
WDRAW_DEBT_MDN	The median debt for students who have not completed	aid	numeric
LOAN_EVER	Share of students who received a federal loan while in school	aid	numeric
FEMALE	Share of female students, via SSA data	student	numeric
MARRIED	Share of married students	student	numeric
DEPENDENT	Share of dependent students	student	numeric
VETERAN	Share of veteran students	student	numeric
FIRST_GEN	Share of first-generation students	student	numeric
MD_FAMINC	Median family income in real 2015 dollars	student	numeric
RPY_3YR_RT_SUPP	3-year repayment rate, suppressed for n=30	repayment	numeric
UGDS_MEN	Total share of enrollment of undergraduate degree-seeking students who are men	student	numeric
UGDS_WOMEN	Total share of enrollment of undergraduate students who are women	student	numeric

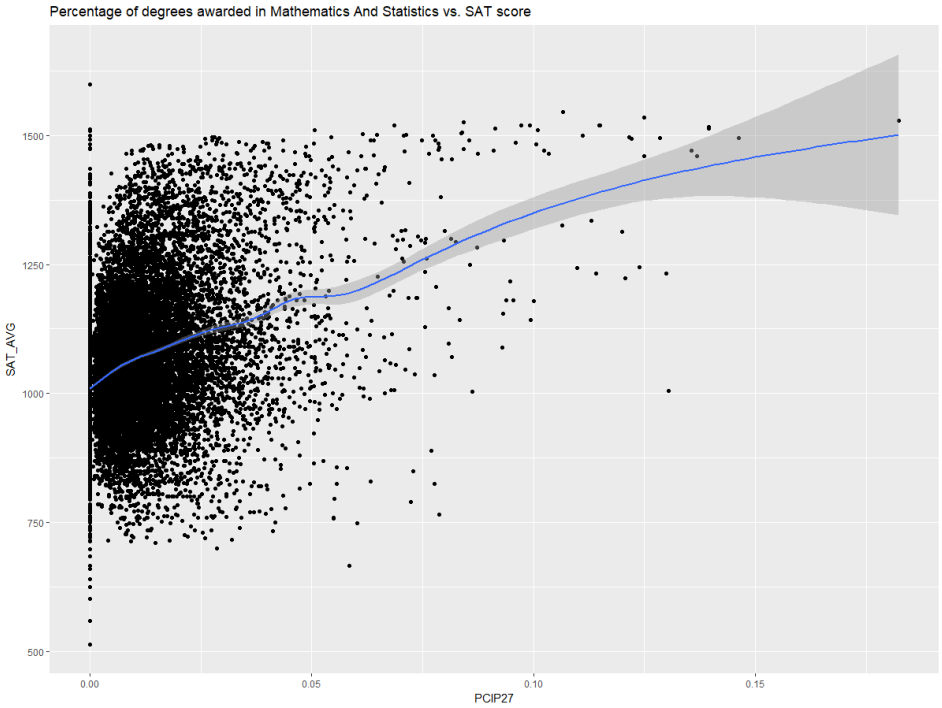
B. Appendix 2 - Exploratory Analysis Graphs

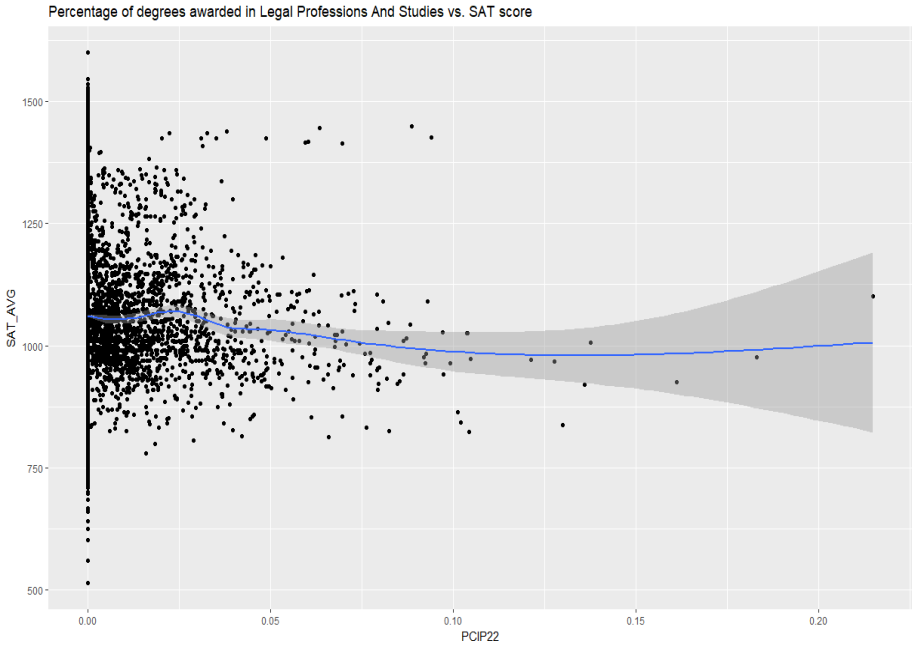
SAT scores by the type of degree awarded (across all years)

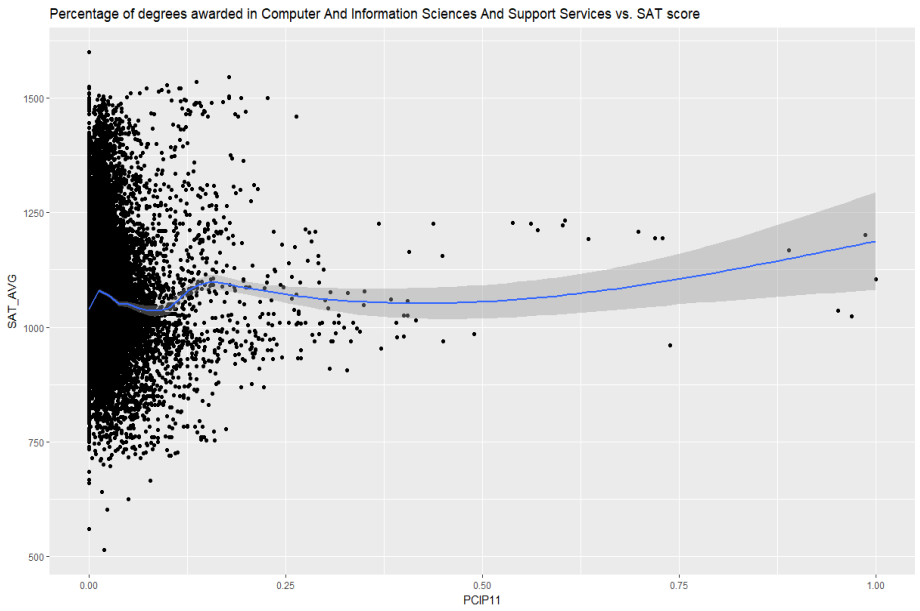
The trendline is a non-parametric smoother which shows the local trendline at each section of the graph.



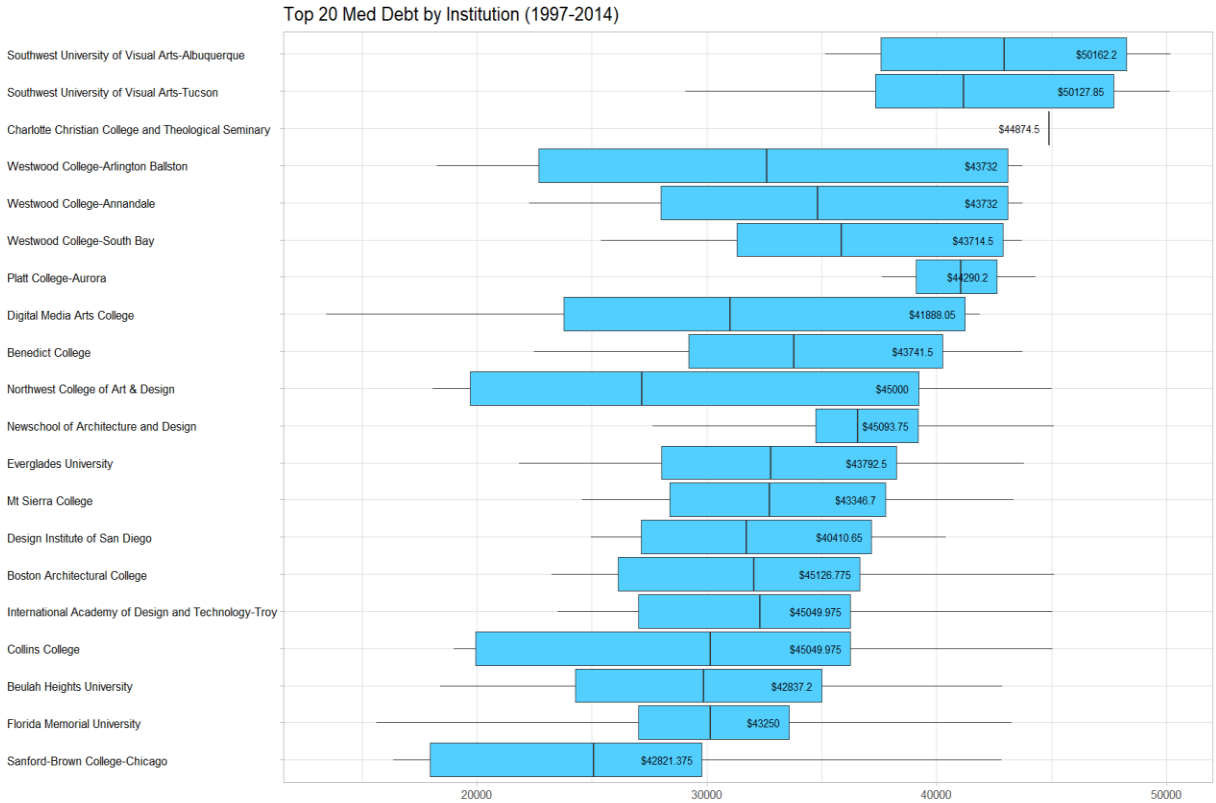


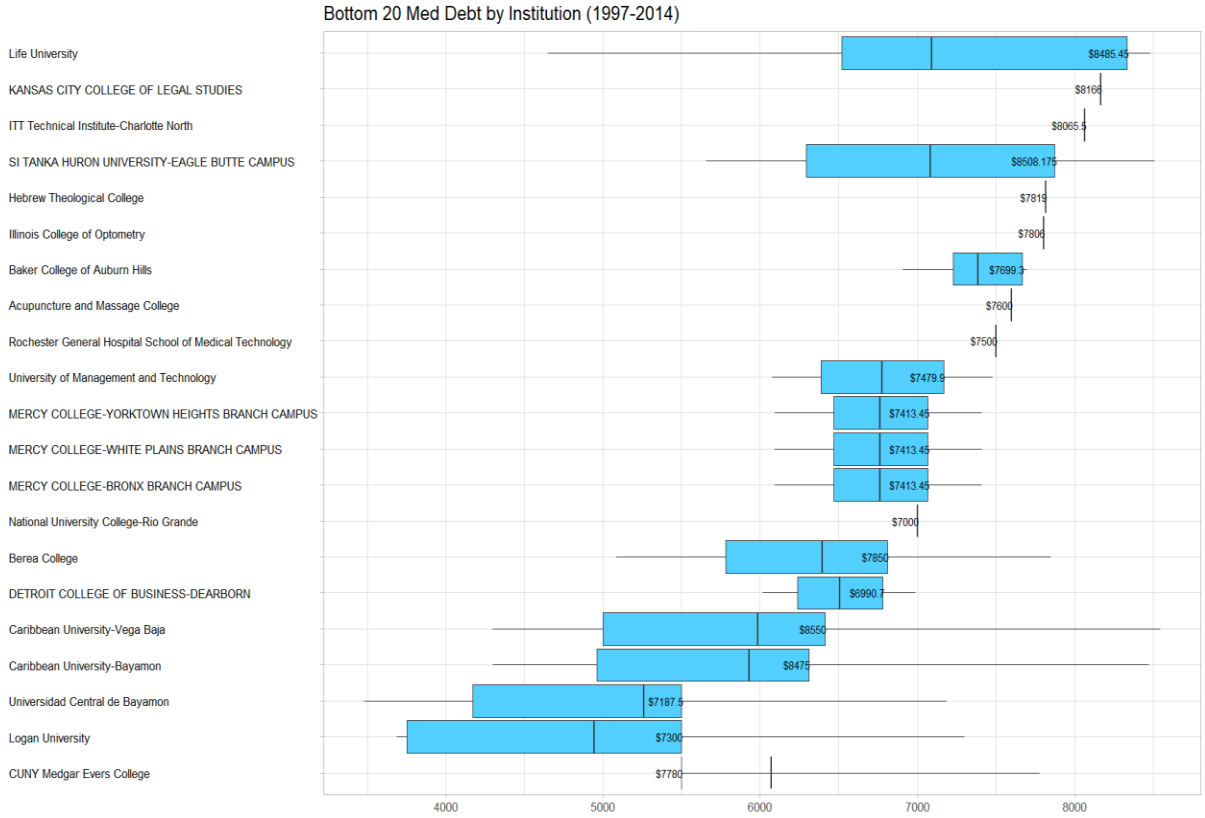

Percentage of degrees awarded in Mathematics And Statistics vs. SAT score
SAT_AVG
1500
1250
1000
750
500
0.00
0.05
0.10
0.15
PCIP27


Percentage of degrees awarded in Legal Professions And Studies vs. SAT score
SAT_AVG
1500
1250
1000
750
500
0.00
0.05
0.10
0.15
0.20
PCIP22



Top and bottom 20 institutions by their student's debt





C. Appendix 3: Regression results for HBCU Analysis

The results of the OLS and LASSO regressions for the HBCU analysis:

	input	est_OLS	p_OLS	est_OLS_cut	p_OLS_cut	est_LASSO	est_LASSO_1se
1	(Intercept)	-9.74E+00	7.53E-01	5.51E-01	3.58E-32	*NA*	*NA*
2	ADM_RATE	-1.06E-01	3.42E-04	-2.18E-01	9.47E-15	-8.77E-02	-6.32E-02
3	CDR3	-4.16E-01	4.13E-02	*NA*	*NA*	-4.15E-01	-4.39E-01
4	CONTROL2	-3.89E-02	9.11E-02	*NA*	*NA*	0.00E+00	0.00E+00
5	CONTROL3	-1.89E-02	6.52E-01	*NA*	*NA*	0.00E+00	0.00E+00
6	COSTT4_A	2.57E-06	1.22E-02	*NA*	*NA*	1.34E-06	1.04E-06
7	CURROPER1	-1.03E-01	9.73E-02	*NA*	*NA*	-1.64E-02	0.00E+00
8	DEPENDENT	-9.26E-02	1.17E-01	*NA*	*NA*	0.00E+00	0.00E+00
9	DEP_INC_AVG	5.12E-07	5.46E-01	*NA*	*NA*	4.76E-07	3.03E-07
10	DISTANCEONLY1	-5.48E-03	9.58E-01	*NA*	*NA*	0.00E+00	0.00E+00
11	FEMALE	7.52E-02	5.79E-01	*NA*	*NA*	1.12E-01	2.62E-02
12	FIRST_GEN	-3.73E-01	6.44E-04	-6.66E-01	1.97E-17	-4.61E-01	-5.13E-01
13	GRAD_DEBT_MDN	-2.23E-07	8.97E-01	*NA*	*NA*	0.00E+00	0.00E+00
14	HBCU1	3.31E-02	3.95E-01	*NA*	*NA*	2.64E-02	0.00E+00
15	LOAN_EVER	-8.01E-02	3.17E-01	*NA*	*NA*	0.00E+00	0.00E+00
16	MAIN1	2.84E-02	3.01E-01	*NA*	*NA*	0.00E+00	0.00E+00

	input	est_OLS	p_OLS	est_OLS_cut	p_OLS_cut	est_LASSO	est_LASSO_1se
17	MD_FAMINC	-4.61E-07	5.82E-01	*NA*	*NA*	0.00E+00	0.00E+00
18	PCIP01	9.55E+00	7.58E-01	*NA*	*NA*	0.00E+00	0.00E+00
19	PCIP03	1.06E+01	7.33E-01	*NA*	*NA*	0.00E+00	0.00E+00
20	PCIP04	9.70E+00	7.54E-01	*NA*	*NA*	0.00E+00	0.00E+00
21	PCIP05	1.10E+01	7.24E-01	*NA*	*NA*	4.95E-01	1.01E-01
22	PCIP09	1.03E+01	7.41E-01	*NA*	*NA*	8.34E-02	4.00E-03
23	PCIP10	1.04E+01	7.37E-01	*NA*	*NA*	0.00E+00	0.00E+00
24	PCIP11	1.03E+01	7.41E-01	*NA*	*NA*	0.00E+00	0.00E+00
25	PCIP12	9.95E+00	7.48E-01	*NA*	*NA*	0.00E+00	0.00E+00
26	PCIP13	9.88E+00	7.50E-01	*NA*	*NA*	-8.98E-02	-8.84E-03
27	PCIP14	1.01E+01	7.44E-01	*NA*	*NA*	0.00E+00	0.00E+00
28	PCIP15	1.02E+01	7.41E-01	*NA*	*NA*	0.00E+00	0.00E+00
29	PCIP16	9.06E+00	7.70E-01	*NA*	*NA*	0.00E+00	0.00E+00
30	PCIP19	1.00E+01	7.47E-01	*NA*	*NA*	0.00E+00	0.00E+00
31	PCIP22	9.63E+00	7.56E-01	*NA*	*NA*	0.00E+00	0.00E+00
32	PCIP23	1.02E+01	7.43E-01	*NA*	*NA*	0.00E+00	0.00E+00
33	PCIP24	1.01E+01	7.45E-01	*NA*	*NA*	0.00E+00	0.00E+00
34	PCIP25	-1.37E+00	9.66E-01	*NA*	*NA*	0.00E+00	0.00E+00
35	PCIP26	1.01E+01	7.45E-01	*NA*	*NA*	0.00E+00	0.00E+00
36	PCIP27	1.11E+01	7.21E-01	*NA*	*NA*	5.77E-01	2.59E-01
37	PCIP29	9.89E+00	7.50E-01	*NA*	*NA*	0.00E+00	0.00E+00
38	PCIP30	1.02E+01	7.43E-01	*NA*	*NA*	0.00E+00	0.00E+00
39	PCIP31	1.01E+01	7.45E-01	*NA*	*NA*	0.00E+00	0.00E+00
40	PCIP38	1.01E+01	7.44E-01	*NA*	*NA*	0.00E+00	0.00E+00
41	PCIP39	1.01E+01	7.44E-01	*NA*	*NA*	0.00E+00	0.00E+00
42	PCIP40	9.72E+00	7.54E-01	*NA*	*NA*	0.00E+00	0.00E+00
43	PCIP41	9.10E+00	7.70E-01	*NA*	*NA*	0.00E+00	0.00E+00
44	PCIP42	9.81E+00	7.52E-01	*NA*	*NA*	0.00E+00	0.00E+00
45	PCIP43	1.02E+01	7.43E-01	*NA*	*NA*	0.00E+00	0.00E+00
46	PCIP44	1.00E+01	7.46E-01	*NA*	*NA*	0.00E+00	0.00E+00
47	PCIP45	1.04E+01	7.37E-01	*NA*	*NA*	4.45E-01	5.42E-01
48	PCIP46	1.17E+01	7.18E-01	*NA*	*NA*	0.00E+00	0.00E+00
49	PCIP47	8.07E+00	7.95E-01	*NA*	*NA*	0.00E+00	0.00E+00
50	PCIP48	1.26E+01	6.85E-01	*NA*	*NA*	0.00E+00	0.00E+00
51	PCIP49	1.03E+01	7.40E-01	*NA*	*NA*	0.00E+00	0.00E+00
52	PCIP50	1.01E+01	7.46E-01	*NA*	*NA*	0.00E+00	0.00E+00
53	PCIP51	9.98E+00	7.47E-01	*NA*	*NA*	0.00E+00	0.00E+00
54	PCIP52	1.01E+01	7.44E-01	*NA*	*NA*	0.00E+00	0.00E+00
55	PCIP54	1.01E+01	7.45E-01	*NA*	*NA*	0.00E+00	0.00E+00
56	PCTFLOAN	3.17E-02	6.24E-01	*NA*	*NA*	0.00E+00	0.00E+00
57	PCTPELL	-1.55E-01	1.05E-01	*NA*	*NA*	-1.15E-02	0.00E+00
58	PPTUG_EF	-2.32E-01	6.21E-05	-9.27E-02	2.65E-02	-6.48E-02	-2.61E-02
59	RPY_3YR_RT_SUPP	4.27E-01	5.84E-07	3.62E-01	1.13E-17	1.66E-01	8.18E-02

	input	est_OLS	p_OLS	est_OLS_cut	p_OLS_cut	est_LASSO	est_LASSO_1se
60	TUITFTE	5.68E-07	7.13E-01	*NA*	*NA*	8.05E-07	1.13E-06
61	UGDS	4.89E-06	5.55E-08	5.36E-06	3.33E-14	3.64E-06	2.46E-06
62	UGDS_AIAN	-1.01E-02	9.79E-01	*NA*	*NA*	0.00E+00	0.00E+00
63	UGDS_ASIAN	1.90E-01	2.28E-01	*NA*	*NA*	2.78E-01	3.38E-01
64	UGDS_BLACK	2.01E-01	1.38E-02	*NA*	*NA*	0.00E+00	0.00E+00
65	UGDS_HISP	3.20E-03	9.70E-01	*NA*	*NA*	0.00E+00	0.00E+00
66	UGDS_NHPI	1.43E+00	6.03E-02	*NA*	*NA*	0.00E+00	0.00E+00
67	UGDS_WHITE	-2.55E-02	7.26E-01	*NA*	*NA*	-4.39E-02	0.00E+00
68	UGDS_WOMEN	2.57E-01	6.64E-02	*NA*	*NA*	0.00E+00	0.00E+00
69	WDRAW_DEBT_MDN	6.51E-07	7.66E-01	*NA*	*NA*	0.00E+00	0.00E+00