Data Science Final Project

Goutham Yegappan

12/16/2021

Predict 10 Year Median Income of Universities

Introduction

In a democracy, more so than other forms over governance, it is consequential that each individual has access to the tools of literacy, arithmetic reasoning, problem solving, and critical thinking. As each individual has the opportunity to not only shape legislation, but to also lead the country themselves, the effectiveness of a democracy lies in its ability to educate its population. By articulating the importance of education through this lens, it becomes more vital to have a clear and concise definition of what a high-functioning education system looks like, and the virtues it should strive towards reaching.

For the purposes of this paper, I draw these ideals to be those upon which an education system should measure its own success.

```
A system of education should be evaluated based on its students' ability to:

1.  Acquire a good job in a constantly evolving job market
    a.  High quality benefits, and a life sustaining salary
    b.  High levels of security and professional growth
2.  Critically think and question established forms of government and authority
3.  Be creative and push social and intellectual boundaries
4.  Engage in civic discourse and participate in politics
5.  Sustain a physically healthy body and mentally healthy mind
6.  Develop personal philosophies (ethical, metaphysical, and existential) that are consistent with their actions
thus understanding real world rewards and consequences
7.  Form meaningful relationships with the people and the world around them
```

These ideals push for a more holistic approach towards education. One that strays away from subjugating students to vigorous examinations that seem to serve arbitrary goals, discouraging students from finding what interests them in this world. As each one of these ideals are extremely difficult to measure, I wanted to start this examination of the education system from the highest levels, at the 4-year American university.

Overview

From the ideals for a school listed above, I wanted to focus on a universities' ability in assisting its graduates in landing a good job. In this project I aim to predict the 10-year median income of graduates from a given university based on its admission, academics, location, completion, and financial statistics. Through this exploration I will then analyze the factors which contribute to universities with high paying graduates.

Problem Statement

Previous research done in predicting average income of universities (Center of Education and the Workforce) have often used linear regression models for prediction. The downsides of these models are that they were overly simplified, and returned extremely large variations between their predicted value and the actual value. The data these papers used were also preliminary, as source such as the College Scorecard now provides fine grained data on every school in the nation. This project looks to improve upon the predictions, while providing useful analysis through measuring variable importance inside our machine learning models.

High Level Roadmap

```
1.  The data pertaining to universities are collected from the College Scorecard. To account for variation across
graduation classes, I use data from the 2007, 2011, and 2014 years.
2.  Shrink the universities selected for the model to consist of only 4-year not-for-profit universities.
3.  Add state and county data from the U.S census.
4.  Remove outliers and pre-process data.
5.  Select variables used in earlier models and add features that are correlated with income.
6.  Convert relationship status, and other categorical data into dummy variables
7.  Run machine learning pipeline to return model that best predicts the median income of graduates 10 years afte
r they begin their program.
```

Data

Data Decisions

Which Colleges to Include?

As the College Scorecard contains over 7,000 schools, I chose to use a ranking system provided by 'UniRank' to select all colleges that were non-profit 4-year universities. This reduced the number of rows to around 1,700 universities.

Which Years to Select?

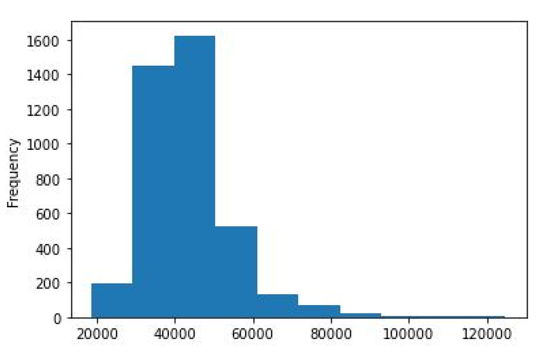
In making the decision between creating a model with only one year and using multiple years, I had to take into consideration the variance between universities stats between the years. Given that an entire new batch of students are present at a university every 4 years I tested to see if the same school varied significantly every four years. After concluding that it did, I used the years that had data for our outcome variable, which was only present in only 8 of the last twenty years. I then picked the three years that closest resembled a new batch of students, ending with 2007, 2011 and 2014. That being said, one drawback to this, is as I don't use time-series data I don't take into consideration the relationship over time for the same school.Below is a chart of all the years with the percent of values that were Nan for our outcome variable.

```
2000 1.0
2001 1.0
2002 1.0
2003 1.0
2004 1.0
2005 1.0
2006 1.0
2007 0.10744512982355472
2008 1.0
2009 0.11126506858805597
2010 1.0
2011 0.1265169119545572
2012 0.10900534215212414
2013 0.10725632227729064
2014 0.10494463044038115
2015 1.0
2016 1.0
2017 1.0
2018 1.0
2019 1.0
```

Percent of Nan Values By Year

Which Outcome to Use?

The 10 year median income indicates the true value of a university more than a 1 year median income, as it compounds on communication skills, networking provided by the school, and extended involvement of the university. In addition, it is important that people find their work interesting, thus stay in that field for a longer period of time, as opposed to someone who gets an amazing job straight from undergrad, but never found the spark and leaves right after. One downside to this that can be argued is that there will be a lot more external variables that impact the 10 year median that the 1 year median would not necessarily have. Here is the distribution of the outcome variable.



Outcome Distribution

Which Variables to Select?

In the earlier literature from the CEW aforementioned, they used the following variables.

Predominant degree, race composition, share of students majoring engineering or high earnings, graduation rates, faculty expenditures and median family income.

A critique of these selections is that they are not broad enough as they share a lot of overlap. For example, the predominant degree and share of students majoring in engineering or high earnings are highly correlated. This is also true of median family income and race composition. Keeping these qualms in mind, I looked to improve my model by introducing a variety of factors across the following fields. These were the factors I added along to the previous ones.

```
Admission: Admission Rate
Academics: Average SAT Score and ACT Score
Location: Average and Median Income By State and County
Completion: Retention Rate, Completion Rate After 5 Years, Number of Graduates
Financial: Tuition Rates, Housing Costs, Net Tuition Revenue
```

The values before show the correlation of some of these variables to the income outcome.

```
ADM_RATE -0.292472456768344
ADM_RATE_ALL -0.2876946058603706
SAT_AVG 0.63368002714841
SAT_AVG_ALL 0.634232621823138
PCIP13 -0.2920882357543315
PCIP14 0.44538009023234476
PCIP45 0.3043089251935598
TUITIONFEE_IN 0.43871155104862297
TUITIONFEE_OUT 0.5402712906082172
TUITFTE 0.4969835194525041
INEXPFTE 0.5009625127601597
AVGFACSAL 0.6684412008358194
C150_4 0.6338943653569683
C150_4_NRA 0.294679875126957
RET_FT4 0.6003089597621118
C100_4 0.543526477643689
D150_4_NRA 0.320721355619533
GRADS 0.35681269057483866
G12MN 0.3218348608492102
ROOMBOARD_ON 0.47059879692201334
ENDOWBEGIN 0.3470397015889839
ENDOWEND 0.34824215192979313
Average Income 0.3509004190494937
Median Income 0.3291259037158245
median_individual_income_age_25plus_2019 0.33434250686915545
CIP14BACHL_0.0 -0.30247356967885475
CIP14BACHL_1.0 0.29966347184692127
```

Correlation Chart

Analysis of Data Science Toolkit

Cosine-Similarity

In matching the universities from the College Scorecard data to the colleges in the 'UniRank' website, one difficulty I ran into was that the names were all slightly different. This same issue occurred when matching the county names to their respective county ID. Using cosine-similarity however, I was able to pair these values by measuring how accurate values were to each other as opposed to checking for perfect matches. I was able to get a near 98% match rate which was more than enough for this project.

Web Scrapers

To collect both the 'UniRank' and Census data I used BeautifulSoup to design a web scraper that retrieved the pertinent data. As these packages allow for excellent traversing of websites, I was easily able to enter sub links to gather information on specific universities as well, such as the year of establishment.

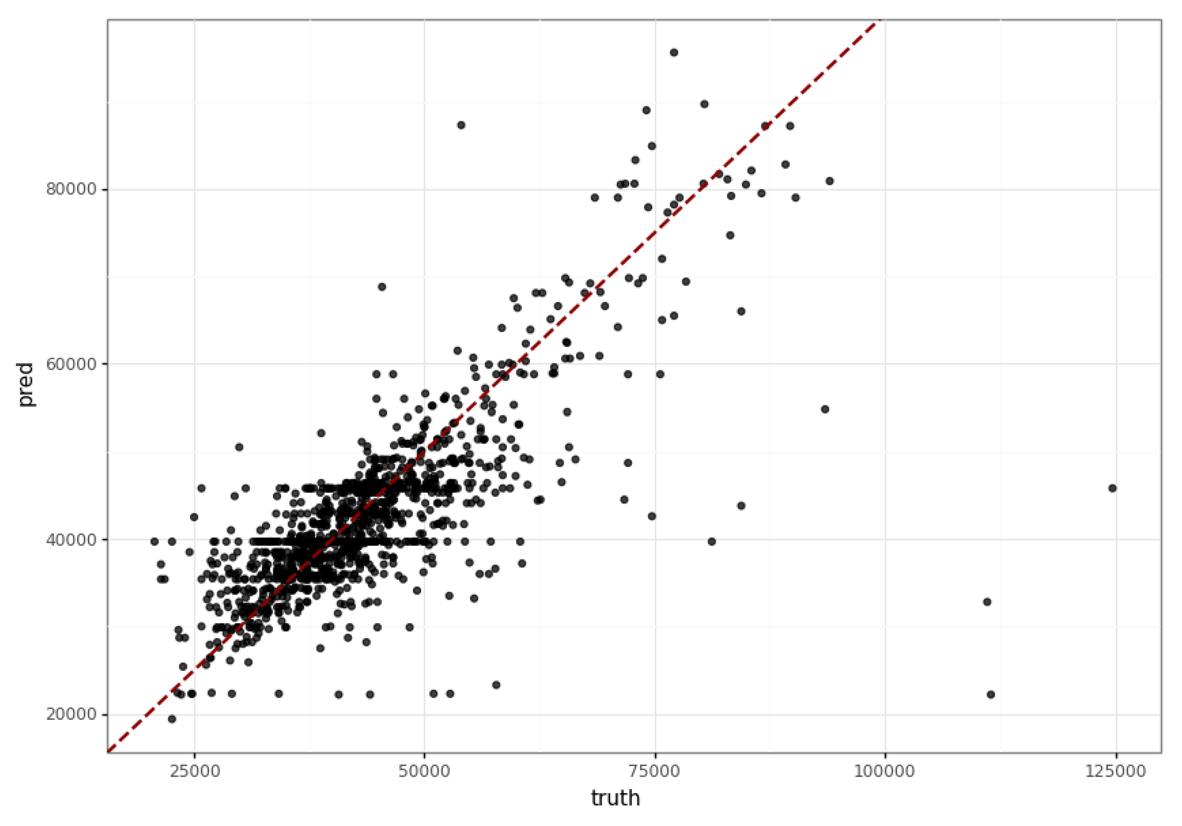
Machine Learning Models

In my machine learning pipeline I used a random forest, linear regression, and a nearest neighbor model to test which model would be the best predictor. These are all examples of supervised machine learning models, meaning that they rely on the outcomes being provided to make their predictions. A linear regression model finds the best fit line between the dependent and independent variables. The nearest neighbor model assigns data points outcomes based on the values of the 'n' nearest data points. The random forest is a model that consists of various decision trees which then average the outcomes of each tree to produce a singular final prediction.

Results

Model

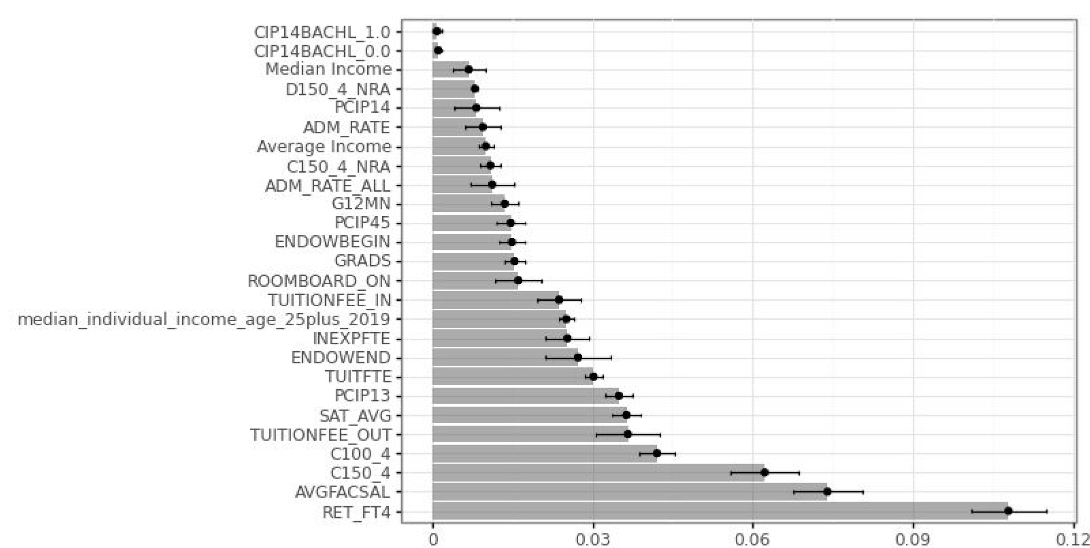
After running these three models, and using the negative mean squared error as the ranking metric, it became evident that the random forest performed significantly better than the other models. This is most likely the case due to the fact that random forests rely on many uncorrelated models to make a decision while the other models don't. This model returned a mean absolute error of $4,555 on the test data. In the graph before we can see the points graphed with the prediction on the y axis and the actual on the x axis. We can see that as we deviate from the mean we get a poorer model.



Truth vs Pred Chart

Variable Importance and Explanation

1. RET_FT4 (Retention Rates of 4 Year Students): This variable is different from completion rates, in that is shows how many students returned after their first year in school. While it is quite surprising that this variable held the most predictive power, it makes sense that schools that provide the right environment for success, keep students around. This is important because for one, if a student at any point feels that their investment may not pay off as they once expected, they can always pull out. This is also true for those students whose tuition increases at a rate they were not anticipating. An interested analysis of this being the most predictive factor could also speak to the fact that a student may actually get the most benefits from a university in their first year. This could be due to networking reasons, as well as connections to the universities hiring portals. This would speak quite highly to the fact that there could be sheep skin effect of going to school and future income.
2. AVGFACSAL (Average salary of faculty): This could impact average income of graduates for these following reasons. Better teachers will be attracted to institutions that offer them more salary. These teachers are better because they either have more teaching experience or are experts in their field, allowing them to have an insight that many other professors would not. Having these better instructors would then inspire students as well as teach them the underlying skills needed to perform well in the workspace. These instructors may also be better connected with the industry, creating a smooth transition for students from the university to the work force. The salary of faculty also speaks to the priorities of the university, as universities who take good care of their faculty, most likely also take good care of their students.
3. (Also 4) C150_4 & C100_4 (Completion rate within 150% of normal time and 100% of normal time): The next two most influential variables are the completion rates for the school. Separate from retention rates these variables speak to the importance of students fully completing their education. It is most interesting that these variables seem to be less predictive than retention rates, but I predict a possible reason for this is that there are multiple variables representing completion. Completion rates speak to the importance of finishing all four years of school. This could be due to either the reputation of the school or the skills learned throughout the four years.
4. TuitionFee_Out (Out of State Tuition): The relationship between out of state tuition and high income could be because of these reasons. The students that are coming from out of state, are paying more, thus have more to lose and are putting in more effort in school. Another reason could be that these students are getting a more well rounded education and are also learning that the world is bigger than just their community. By traveling out of state, and maybe even out of the country, they are experiencing a new society and culture. Another reason is that students who can afford out-of-state tuitions are also more likely to come from families with wealth, thus enabling them to get better paying jobs. Another reason, from the school perspective is that schools with higher out-of-state tuitions are able to use this to fund other aspects of their university, such as faculty spending. This however, seems to be more unlikely because then schools with higher in state tuition would also be a highly predictive factor.
5. SAT AVG: The next highly predictive variable is the SAT average. It is hard to tell whether there could be an introduction of a collider bias, as SAT scores can be linked to university acceptance, and that might have a large connection with getting a high income, but there could in fact be no connection between SAT score and ability to acquire a high paying job. This variable would show that students who are already performing better at the high school level are more likely to work in higher paying jobs.
6. PCIP13: Percent of Program That in Education This is the first highly predictive feature that is negatively correlated with a high income. This is quite ironic, in that programs in education, are poorly rewarded by the education system. This, I think speaks more loudly to the fact that schools with more students in the education program, are most likely liberal arts schools which host a large number of majors that are not well paid.



Importance By Variable

Other Notable Results

I find it most interesting that the location that the university is located in, admission rates and the types of degrees a school offers are not highly predictive factors.

Results

Looking at the variable importance list the largest finding is that there seems to be much more importance placed on the school itself than on the students these schools accept. There seems to be no strong predictive power between admission rates and the income of these graduates which is quite surprising. What this speaks to is the fact that what happens at the university helps students find these high paying jobs. Whether it be through their reputation or the material taught can not be determined from this paper. However, this should reflect upon schools to pay more attention to the way in which their administration and students are taken care of more than they focus on refining their admission process.

Moving on to the actual predictive power of these results, we can see that the average cost from the actual value is approximately 4,500 off. While, future implementations (improvements can be seen in the following section) will make these predictions more accurate, this is a solid start which puts a college's 10 year median income in a relative $9,000 confidence interval.

From these findings, students can better prepare to realistically understand what the likely outcomes and benefits of going to a certain college will be. Students can make better judgement regarding their choice of university by inquiring from professors how they feel that they are being treated by the university, and speak to students to see if they feel that the university provides for them financially, mentally, and physically.

Places For Future Improvement

From the processes taken to arrive at these results, there are a few areas in which major improvement could have been done. To begin a deeper knowledge in the area of expertise would have been helpful in engineering variables that predicted income. In addition to this, it it important to mention that the large amount of missing variables definitely made this project significantly less reliable as the number of data points we had significantly diminished after removing rows with many nan variables.

In replacing the remaining nans with either median, mean, 0, or some other formula, there is a lot of intuition that was missed here. A better approach to replacing nans, or even a better approach towards scraping the actual values would have been extremely helpful here.

There were also computational issues as my computer could not process these many data points quickly for a more complex model. Therefore I was not able to test a larger range of parameters which would have definitely helped my model. I have a strong hunch that this model quite over fit the training data, and did not account to the variance in the test data.

I would also have liked to double check the college scorecard data to make sure that their values were accurate. One hunch that I had was that even the 2014 college data had a 10 year md income, which at first did not make sense to me. The only way this could be possible is if the measurement was started when one graduated school and the 4 years of invested time was already accounted for. This is a very odd way to calculate this data, and more investigation into this would need to be done.

Future Steps

While this model can be continuously improved, connecting back to the first section, there are a few more places of improvement that can be facilitated, which improves the overall quality of the education sector. To accomplish this there needs to be a metric measuring each one of those ideals. When these future models are built, we can have a more holistic approach of measuring the success of universities. This project is the first step of many to accomplishing this. The largest takeaway from this project is that a student's financial success relies more on the facilities that are provided to them at the university than the skills they already possess when entering these.

Bibliography

1. The college payoff: Education, occupations, Lifetime earnings. CEW Georgetown. (2021, August 13). Retrieved December 17, 2021, from https://cew.georgetown.edu/cew-reports/the-college-payoff/