Predicting Student Success

Taimur Ghani and Michael Liu

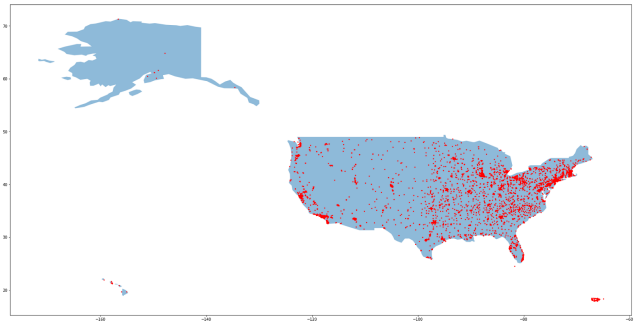
Overview

The project had two main portions to it and two main goals. The first portion set out to evaluate what factors in a college are important when students are trying to determine which colleges are better. The definition used for a better college here is a college with a higher post-graduate salary than another. It is often assumed that things such as higher SAT scores, lower admission rates, location, etc. are the most important factors for a school. College data was used to determine if these factors were the most important factors for maximizing a colleges post-graduate salary and if there were also other important factors that are overlooked.

The second goal of this project was to predict the post graduation salary of a student. Since college tuition is increasing at an astronomical rate, it is important to know whether or not the investment in a higher education is worth it. There are many factors that a student must think about when they attend college or university. In contrast to the first goal, this was looking at specific student data to estimate an actual specific student data rather than trying to find which factors increase the median salary of a school in general. Hence, we set out to create a model that would be able to estimate a specific students post-graduate salary based on some specific characteristics about them and their university.

Part 1: What Makes a College Better?

For this we used a dataset from the US department of education. This dataset included data about many colleges in the United States. This data included many features about various colleges. Some of the key ones we used were institution name, median-post graduation salary, average SAT score, number of undergraduates, median student debt, average faculty salary per month, cost of attendance per year, admission rate, latitude, longitude, and state abbreviation.

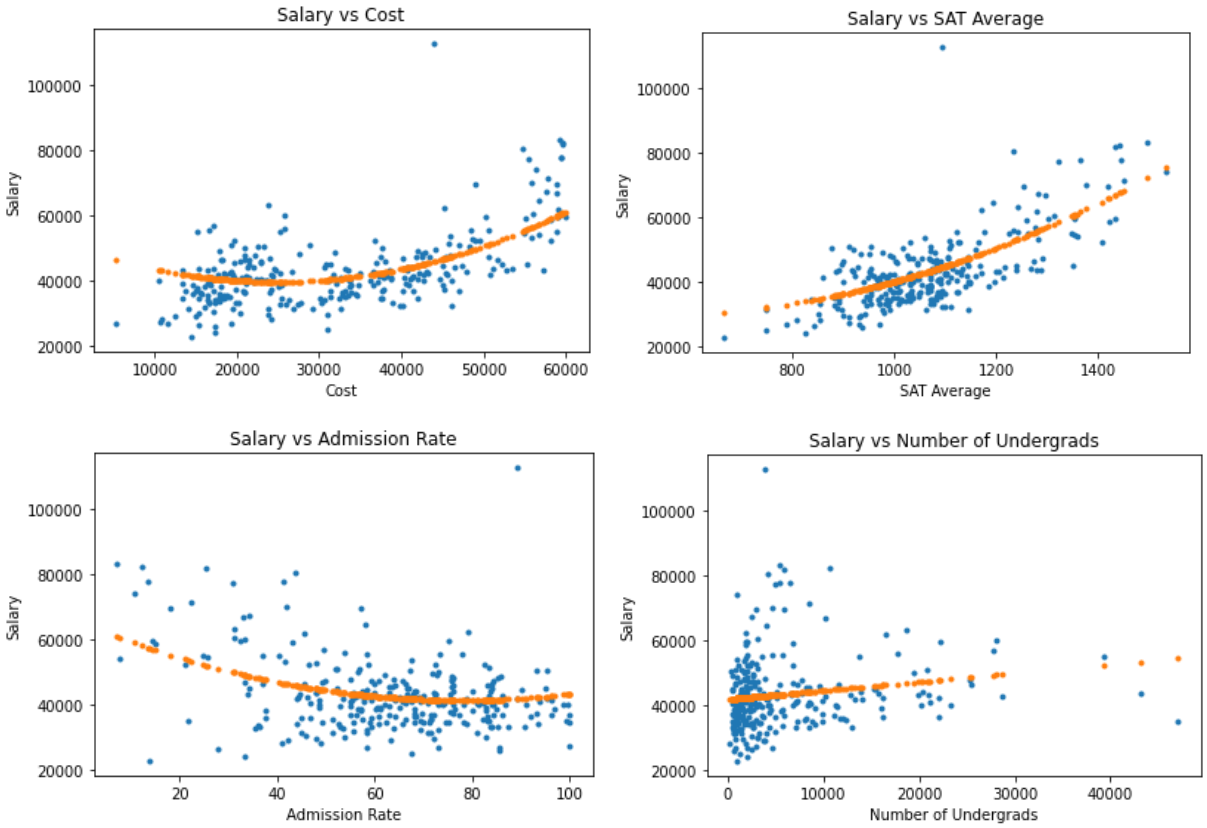


On the right is a map of the United States with a red dot representing one college. This plot was made with the latitude and longitude from the dataset with the python library geopandas. As the plot shows, there are colleges from all across the nation represented in the dataset, including Alaska and Hawaii.

Many of the entries were NULL or “Privacy Suppressed” so these rows and colleges were removed to have clean data which ended up having data for approximately 1500 colleges. Since the faculty salary was given on a per month basis, we multiplied this by 12 to turn this into a per year salary. Also, the admission rate was given as a decimal, so this was multiplied by 100 to become a percentage.

There are some biases in the data that must be kept in mind and might potentially have an impact on the results. The data is limited to schools that receive financial aid. Most schools do receive financial aid, hence this is not a major factor, but it is still a bias. It is also important to note that the college data was for 2013-2014. The salaries were adjusted for inflation (not done by us but rather the dataset came like this). For our purposes, this should not be a major factor since we are just trying to see which factors make a college “better”. Also, it is not that old, so many colleges would not be too different. Newer data would be used, but the data from newer years do not have the median salaries for colleges aggregated yet.

To start, single variable regression with a 80-20 train-test split was used on various predictors with a constant predictor variable: Median Post Graduate Salary. This was done to see which features seem to be good estimators for what is a better school (again, defining better as a school that has a higher post-graduate salary). Below, some of these results are shown:



These are only a few of the predictors that were plotted to determine which features seemed to be good predictors. For some features, polynomial regression was used such as admission rate and for others simple linear regression was used such as debt (not shown). As seen above, this process eliminated some features, such as the number of undergraduates which did not seem to be a good estimate for which school is better. For each of these the mean absolute error was calculated on the test set to determine how good of a predictor this was with just one variable. The best single variable test-set loss was “Salary vs. SAT average” which results in a mean absolute error (MAE) of about $6000. For just one single variable (using polynomial regression), this was quite good. Our baseline was using just the median salary and that resulted in a mean absolute error or approximately $8000, hence this was a $2000 MAE difference. Using all the data we had gathered from these single variable regressions - which features are good predictors and which seem to be bad predictors - we decided to try to perform multivariable regression to try to find a better model.

For the multivariable regression, the variables used were cost (which was divided by 1000 to be cost by thousand), admission rate, SAT average, and average faculty salary of a school. Some extra columns were added to the predictor matrix, such as some squared terms for terms that were quadratic in the single linear regression. This would make the final model:

$$Salary = \beta_0 + \beta_1 a + \beta_2 a^2 + \beta_3 b + \beta_4 b^2 + \beta_5 b + \beta_6 c + \beta_7 c^2 + \beta_8 d$$

Where “a” is cost, “b” is admission rate, “c” is SAT average, and “d” is average yearly faculty salary.

Using sklearn, we were able to solve for the coefficients. This model was tested against the test-set to determine the loss of this multivariable regression model. The MAE was approximately $4800. This was $3200 better than the baseline of using the median salary, and it was still an improvement of $1200 from using single variable regression on just SAT average.

This seemed to give us a good understanding on if the assumptions most students make such as higher SAT averages and lower admission rates were accurate and what outliers were out there. We decided to test one more important factor which was adding the colleges state. Using one-hot encoding we were able to do the same multivariable regression as above, except in this case we added the state data. Using this we were able to get a MAE of approximately $4400. This was a better loss but was not super significant when it was including showing that state might not have as much as many people often assume

```
Salary vs Multivariable with States:
Baseline Mean Absolute Error: 7755.234657039711
Mean Absolute Error: 4444.555004695511
```

In conclusion, there were many things that were found from this part of the project and many things that if given more resources and time we could have changed. First, we did see that in general schools with higher SAT averages and higher faculty salaries did generally tend to have higher post-graduate salaries. Number of undergraduates did not matter much when deciding if a school was better than another. One key was to look at outliers and see what outliers exist. We saw that often schools of medicine were outliers with higher salaries than other schools with the same features. This makes sense due to the fact the job market for these schools offer higher salaries. Also, some outliers with high salaries were ivy leagues, as well as schools that were mainly engineering schools. Some schools that were outliers having low salaries based on their predictors were schools that were mainly all liberal arts - this also makes sense due to the job market present for some of these majors. Lastly, some many of the outliers we found with high salaries were also found on many of the “Best Value” college lists online, showing it was consistent with many of those as well.

If more time was given, possibly aggregating more data together and looking at a few more variables that do not seem as likely would be interesting. Some things such as looking at the type of location of the school such as rural or urban would be interesting to look at. This, along with looking at the number of internships on average on students, could help determine if being in a more urban environment increases the chances of getting an internship, hence increasing post-graduate salary. The methods would most likely stay the same, but potentially there could be an addition of accounting for the fact that some variables are probably related within the model and accounting for that when coming up with the equation. In general, the main improvement would be to delve a little deeper to get some more “explanations” for why some features act in the way they do when compared to salary.

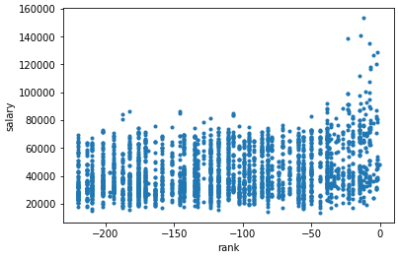
Part 2: Estimating Post-Graduate Salary

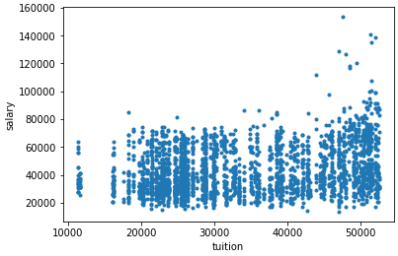
We took the dataset from the US department of education on student outcome by field of study, which included the level of education of an individual student, post graduate salary, field of study. Then we took another data set that had rankings of the top 200 universities in the U.S. and their locations(state), and matched the rankings and location to the university of each student, merging the data into a single csv file to predict post graduate salary. Many of the entries in the U.S. department of education data set had fields that were left blank, so when we

processed the CSV, we checked to ensure all the fields that are required for the ML model are present and the college the individual goes to is in the ranking data set too so they could be matched up. The major was matched up by using a dictionary that had a list of keywords in the most popular majors in the United states. After processing these two CSVs, we were able to get close to 4000 entries of usable data.

There are some biases in the data that must be kept in mind and might potentially have an impact on the results. The data is limited to students who received Federal Financial Aid. Hence, this does not include all students.

In the machine learning model, one hot coding was applied to categorical fields such as major and location. We first plotted salary against rank and tuition.





We were able to see some trends and positive correlation between those variables. Then we tested the baseline, checking the loss against the average/median salary of all students, and we were able to get an average L1 loss between $13000-14000. We then applied multiple linear regression to the data set to predict salary. The average L1 loss for the training set is $6650, and the average L1 loss for the test set is $7140, a significant improvement from before.
We then calculated the correlation coefficient for each variable to try to determine what is the most significant factor for salary.

```
Correlation coefficients: (Average of absolute value for each major and location)
Rank:  0.27516991288998127
Tuition:  0.27379003470916397
Major Average:  0.09794138646649324
Location Average:  0.027263949812578698
```

We saw that rank and tuition had the highest correlation coefficient with salary, rank makes sense as you would expect higher salary after attending a better school, tuition probably has a strong correlation with the rank of the school(many top universities are private and highly expensive). Location on average has basically no correlation with the post graduation salary, this could possibly be explained by the fact that many people don't work in the same state they went to school in post graduation. And while some majors have a strong (both positive and negative correlation) with salary, the majority of them are not strong correlation variables.

```
The 20 strongest positive correlation variables:
('Major__Computer Science', 0.27750605575247667)
('Rank', 0.27516991288998127)
('Tuition', 0.27379003470916397)
('Major__Electrical', 0.21041802256897066)
('Major__Mechanical', 0.20158330446918055)
('Major__Information', 0.1993688291503737)
('Major__Nursing', 0.18949459123925194)
('Major__Computer Engineer', 0.15359268928644704)
('Location__PA', 0.1290309895297291)
('Major__Civil', 0.12817229862823157)
('Major__Industrial', 0.1019892293395976)
('Major__Chem', 0.09332481849073107)
('Major__Finance', 0.09161343334551993)
('Location__MA', 0.09024913459482714)
('Major__Accounting', 0.08570107750355382)
('Major__Aerospace', 0.08031775457003525)
('Major__Management', 0.08017123459009073)
('Major__Biomedical', 0.07743168466727679)
('Major__Economics', 0.07538763449112489)
('Major__Math', 0.06781922280322908)
```

```
The 20 strongest negative correlation variables:
('Major__Art', -0.24868254624353117)
('Major__Language', -0.16604624441967622)
('Major__Bio', -0.1570856324783877)
('Major__Psychology', -0.1508678179197493)
('Major__English', -0.13132900574731995)
('Major__Anthropology', -0.1070513265734893)
('Major__Health', -0.1020680120084168)
('Major__Sociology', -0.10175567517802471)
('Major__Music', -0.0969468147193498)
('Major__History', -0.09347074669013604)
('Major__Social', -0.08424434662822672)
('Major__Journalism', -0.07858061086859504)
('Location__OR', -0.07574970734294881)
('Location__FL', -0.06901194638999004)
('Major__Politic', -0.06556126273897572)
('Major__Government', -0.06556126273897572)
('Major__Criminal', -0.06214420878216978)
('Location__VA', -0.04959628931964158)
('Location__GA', -0.04915066901017824)
('Location__MS', -0.048814250193346266)
```

Above is the list of the strongest positive and and strongest negative correlation variables. We saw the majors like computer science, electrical engineering and mechanical engineering, along with rank and tuition, had a very strong positive correlation variable. And majors such as art, language are the strongest negative correlation variables.

We then tried to make the parameters polynomial to see if that would make the fit better, the loss on the training set decreased significantly, while the loss on the test set increased significantly. This is most likely due to overfitting of the polynomial data. As a result of this finding, we went back to our original model.

If we had more time for this project, we would have tried to look for more data that could be used, such as the individual's SAT score and the location they are from. We could also use data such as the student's family income, as that could have some impact on the student's resources for education. The methods used would most likely still be the same, and it's possible with additional data we could lower the loss. The main improvement would come from more variables that have high correlation coefficients.

In conclusion, rank and tuition have a strong correlation with post graduation salary, and some majors have strong correlations to salary as well, while many of them have very little correlation and can not act as good prediction variables. Location is a very weak predictor variable.