Where it Pays to Attend College

Demi Tu
Student at UW
Seattle, WA USA
demigod@uw.edu

Matthew A. Wong
Student at UW
Seattle, WA USA
mattw98@uw.edu

Xinyu Chen (Rhea)
Student at UW
Seattle, WA USA
xinyuc2@uw.edu

Zhiqi Lin (ZK)
Student at UW
Seattle, WA USA
zlin2016@uw.edu

ABSTRACT

When prospective college students deciding their college and majors, one of the factors that affects their decisions is the future salary after graduation, since student loan debt is quickly becoming a national problem [3]. Our project is demonstrated to find out the correlations between college, major, and salary by machine learning. By implementing linear regression, we are able to explore the relationship between college and future salary in terms of school rankings, types, and regions. Through our report, we have found out that the choice of college does have a significant impact on mid-career median salary. By exploring the major and salary dataset using K means cluster, we have found out the salary throughout a person's career is highly impacted by the choice in undergraduate major.

INTRODUCTION

Our research address problems that are not only close to home but are also issues on a national level. First of all, does it matter where students go to college? The answer is yes when it comes to students' future paychecks, and this is supported by years of research [5]. Of course, factors that differ among individual students will make the outcome to vary, but research has supported the statement that the more elite a school, the better its alums' paychecks [1]. The rankings of the universities matter a great deal, and the effect is more evident over time.

Settling on a college is not the only difficult decision that needs to be made by students. According to the book named "The Undecided College Student: An Academic And Career Advising Challenge" written by Virginia N. Gordon, about 20 to 50 percent of students enter college as "undecided", and an estimated 75 percent of students change their major at least once before graduation. It is also important to note that "decided" students are not necessarily basing their decision of major on factual research. According to a College Student Journal survey, of more than 800 students who were asked to elaborate on their career decision-making process, factors that played a role included a general interest the student had in the subject he or she chose; family and peer influence; and assumptions about introductory courses, potential job characteristics, and characteristics of the major [2].

While students might fall into a never-ending cycle when choosing colleges and majors, another problem that we are addressing, on the other hand, concerns the topic of student loan debt which directly relates to our target variable, salary. "Even among those who borrowed only for their undergraduate education...only half of students had paid off all their federal student loans 20 years after beginning college in 1995-96." Instead, average borrowers in "this group still owed approximately $10,000 in principle and interest, about half of what was originally borrowed, 20 years after beginning college" (Velez, 2018). The fact that it is very difficult for recent graduates to find well-paying jobs at the outset of their careers, significantly limited their ability to repay their student loans.

The overarching purpose of our research project is to provide an additional, but important, piece of information to students when they contemplate on which college to attend and what they major in college. By revealing potential relationships between colleges and/or majors and how much in salary people earn with their degrees, it will help kickstart the process of deciding on a college and major so students can have more time to learn rather than struggling to find a major of interest.

It is an important undertaking because it is a topic that is closely related to us, college students who are actively seeking jobs and/or internships. We also know we have a broad audience because there are so many incoming/current college students, just like us. If such information were accurately presented, it might affect students' choices on colleges and majors because salary is, perhaps sadly, a very important factor when selecting a career.

RELATED WORK

First we wanted to get a better idea about the relationship between college, undergraduate major and income. In order to do this we referenced by Scott L. Thomas and Liang Zhang which examined the rate of to which salary growth for early career college graduates can be attributed to college and major choice. In their research the two computed the statistical difference between different college graduate incomes based on the inclusion of various factors. Their findings indicated that there was a statistical advantage to going to a prestigious college when it came to post grad careers, but that advantage didn't really grow until several years into the career. Furthermore, they found that choice in major was the biggest contributing factor to an individual's ultimate salary. Another point in this study is that factors

such as ethnicity, gender and socio-economic background influence a person's income. However, our data does not include these variables, but there inclusion would make a good branch for additional research [6].
Several key related studies were analyzed in order to inspire our own work. Since predicting salary was a key component in our study we referenced a submission in an open competitive study hosted by Kaggle with the goal of predicting salary from job postings (Kaggle). This paper analyzed the salary predicting power of four regression techniques: linear models, random forest regressors, gradient boosted tree regression and neural networks. Their final results found that the best stand-alone model was the random forest regressors, but that their best performance came from an aggregate model that linearly combines the four best random forest models. While this was not the approach, we ended up adopting for our model, the analysis of different regression techniques was particularly useful and defining our own.

METHODS

In United States, college education consists of 4 years of undergraduate schooling. A college type can either be State, Party, Ivy League, Engineering and Liberal Arts. When researching for source data, we weren't able to find a comprehensive dataset that contains salary data related to both *majors* and *colleges*. Thus, we chose to look at these two individual factors separately.

College vs. Salary

Due to the limit of data source, we combine that data sets of salaries by college type, salaries by college region, and salaries by college ranking together to create a comprehensive dataset. Because the data sets concerning colleges do not come from the same source, not all colleges in our comprehensive college data set has rankings. Rather than removing all the rows of colleges that do not have rankings, we decide to take this opportunity to separate our data set into two part: college with ranking, and college without ranking so that we can test if ranking is a deciding factor that impacts salary. We noticed that there were still missing values in our data sets after the split. However, those are not independent variables or features that we will use in modeling, so we chose to leave them as that.

Since we are dealing with a lot of categorical data like school type and school region., we employ One-Hot-Encoding to transform those categorical data into meaningful numeric numbers for further data analysis. Because we are only concerned with the strength of correlations, not whether they are positive or negative correlations, correlation squared values work the best for us to find out the variable correlations.

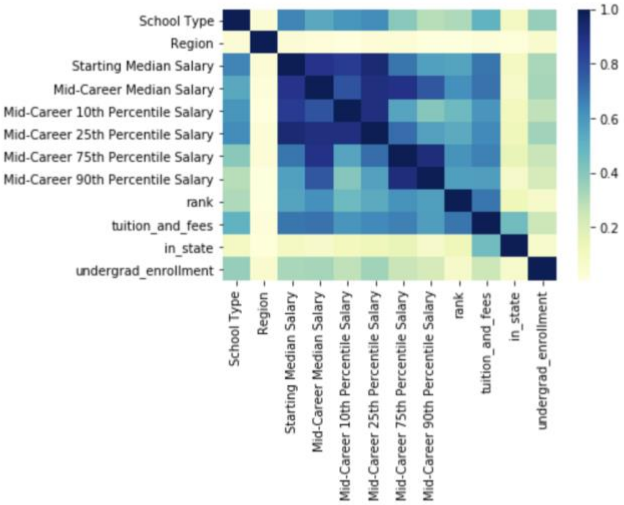
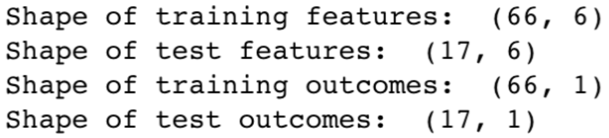


Figure 1. Shows that there are some notable correlations between School Type and salary in general. This does not seem like the case for Region though. We will need to use more than one variable to make our predictions because more than one variable has strong correlation with salary.

Instead of creating new features based on our domain knowledge, we incorporated an external data set that provides important insights of what we are trying to generate. We chose to perform feature selection using *VarianceThreshold* which removes all features whose variance is low and does not have much effect on the target variable. We didn't think we such step is needed before performing this feature selection because all features we have in hand seem to have relationships with Salary. We only input the relevant features that can be interpreted by the learning-based approach, and they are *School Type*, *Region*, *Starting Median Salary*, *rank*, *tuition_and_fees*, and *undergrad_enrollment*.



```
Shape of training features:  (66, 6)
Shape of test features:  (17, 6)
Shape of training outcomes:  (66, 1)
Shape of test outcomes:  (17, 1)
```

Figure 2. We split data into two train/test sets: *training features*, *test features*; *training outcomes*, and *test outcomes*

We performed data visualization as our first step because different features are scaled differently. For example, the number of *categories* is different for *School Type* and *Region*. Therefore, we used a *MinMaxScaler* that subtracts the the minimum, and divides by the difference between the minimum and the maximum of each column. In order to not leak information from out test data into our training data, we normalized our data after splitting.

We then fit a multiple linear regression model with the scaled training data. We chose linear regression because the dependent variable, Mid-Career Median Salary, is measured on a continuous measurement scale and the independent variables consist of categorical and continuous measurements. *LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)*. After fitting the model, we use it to make predictions on the test data, and then we hit a Cross validation score of *0.7616423311095187*. We try to optimize the (hyper) parameters of the model using a grid-search. We got a performance score of *0.9064408556353608,* with indications: Best parameters {'copy_X': True, 'fit_intercept': True, 'normalize': True}. Although the cross validation score and the performance of grid-search can vary due to the randomness of splitting training and testing data, it seems like we were able to improve the accuracy of our model by using grid-search.

Major vs. Salary

After acquiring the data from kaggle.com, we remove all strings representing salaires with integers which enables us to perform analysis on the data. Then, we created an interactive bar chart here so users can visualize the relationship between different undergraduate majors and their corresponding earnings at different stages into their careers.

When we finish cleaning the data, we introduce two new variables for feature engineering. First, we calculate the rank of each undergraduate major for each *percentile column* and *mid-career salary* and then use that to compute a new variable *Rank_Sum*. By taking the sum of all percentile ranks, we hope control for undergraduate majors that have extremely high salaries in a single percentile. We also introduce another new variable *Greatest_Range* that is the difference between the 90th percentile and the 10th percentile in hopes to account for the variability within the salaries.

In modeling section, we selected *Greatest_Range*, *Mid-Career Median Salary*, *Starting Median Salary*, and *Rank_Sum* as the columns to be used to compute our clusters. The range is used in order to capture the amount variability within the salaries for each major. In order to get a sense of what the salary is like for an average person, we also include mid career and starting median salaries. Finally we include the sum of all the ranks. We believe this variable will help account for any major that has its median salary affected by especially high or low salaries within a percentile.

After scaling the data, we then create an plot using the *elbow method*. This method determines the optimal number of clusters to use by showing where the increase in clusters does not account for any more explanation in the variance. Based on the plot we chose to use five clusters for our *K means model*. Due to the small size of the data set, a larger number of clusters may subdivide the Undergraduate majors into indiscernible clusters.

We use KMeans model and fit it to our data by taking in cluster Ids and attach them back onto the original data set and then we split the original data set by cluster id (see Figure 3 – 7).

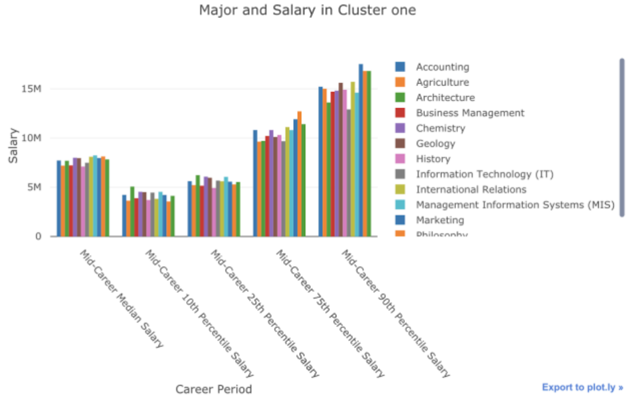


Figure 3. Cluster One has good, but not great median salary and good room for growth. Containing majors such as Business Management, Accounting, IT, and Marketing, this cluster seems to be consistently above average.

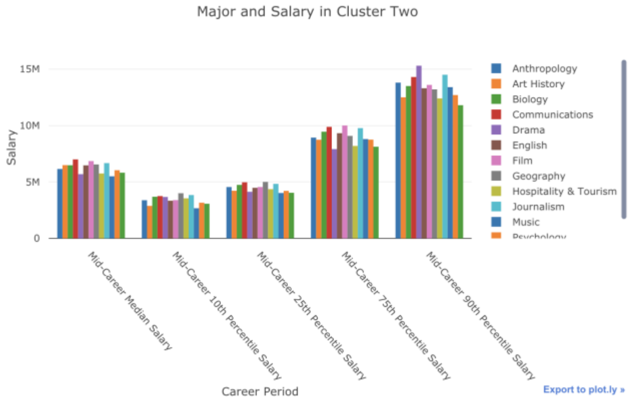


Figure 4. Cluster Two contains the largest range between the percentiles of all the clusters. Containing undergraduate majors like drama, music, film, and journalism, which can pay exceptionally well if you're successful, but overall pay less that other options.

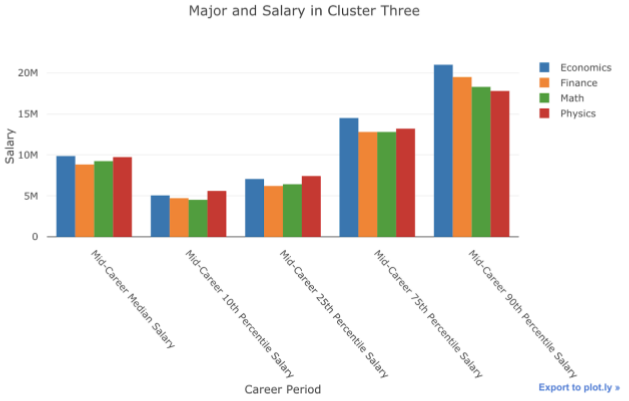


Figure 5. Cluster Three is the smallest cluster in terms of number of majors. It contains degrees all related heavily to mathematics and the second highest average mid career salary

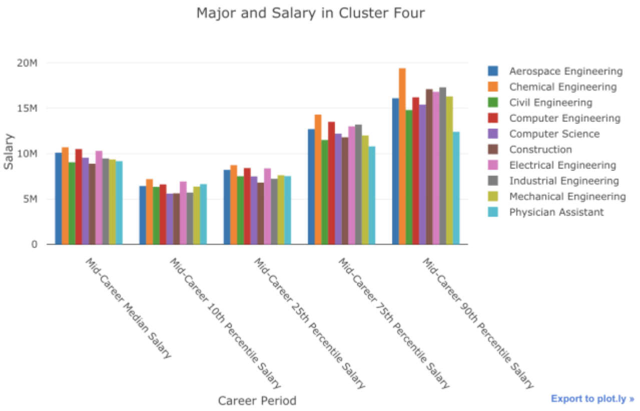


Figure 6. Cluster Four is the easiest to categorize, this cluster contains all of the engineering majors with the addition of physician assistant and construction. This cluster has a relatively high median mid-career salary with good room for growth

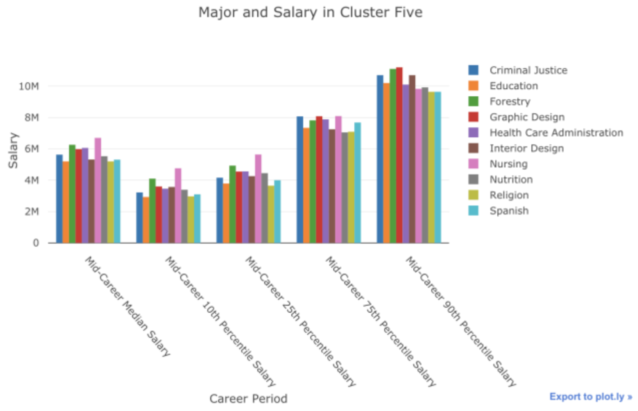


Figure 7. Out final cluster, Cluster Five, has a relatively low mid-career median salary, but unlike the first these majors produce careers with a much smaller range between the percentiles

RESULTS

For College vs. Salary analysis, we access the accuracy of the model by calculating the Root Mean Squared Error (RMSE), which is the square root of the mean of the squared errors. This is the correct approach for our linear regression model because it is interpretable in the "y" units -- 6332.022512031576. We believe this high RMSE is due to our large salary units. It might be easier to understand our results by visualizing the predictions.

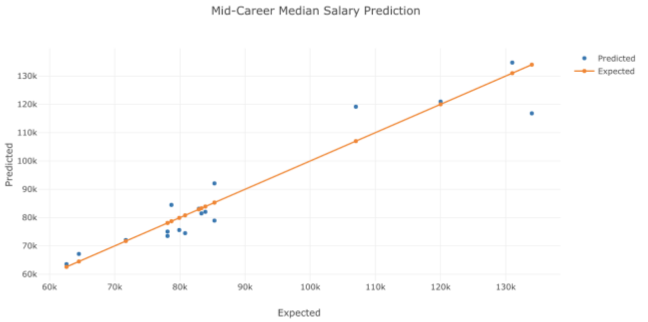


Figure 8. We can see that our model in fact performed well because our predictions are fairly close to the line of expectation

After exploring relationships between features that represent the choices of college and salary graduates make after college, as well as making predictions specifically about the mid-career median salary using our multiple linear regression model, we can conclude that the choice of college does have a significant impact on mid-career median salary.

For Major vs. Salary, we set out to discover how the choice in Undergraduate Major impacted salary throughout a person's career. Through our approach we created five clusters. From these clusters we can see several interesting details. Firstly, if you want a career with a high paying salary, a major that has its basis in math seems to the way to go. This can be seen in cluster two, containing predominantly engineering majors, and cluster five with contains majors like Math, Physics, and Economics. Both of these cluster sport high mid-career median salaries as well as some of the highest areas for growth. Other undergraduate majors can be riskier, such as Drama, Journalism, and Music. These majors and others like them in cluster four, have relatively low mid-career salaries, but have a drastic amount of potential growth as there 90th percentile has some of the highest salaries. Based on these clusters, we do believe that Undergraduate Major has an impact on salary.

DISCUSSION

Because our results support the statements that the choice of college has a significant impact on mid-career median salary and the choice of undergraduate major has a significant impact on overall salary, our research implicates that if students really want to earn a high salary in their future career, they should attend Ivy League schools. These highly ranked colleges are located in the Northeastern region, which are factors that we found to have significant impact on the graduates' wages. In terms of majors, one that has its basis in math seems like the way to go.

Our models have a lot of room for improvement, and there are still other factors as well as noise that our study did not capture, so we do not advise students to rely solely on our study to choose a college to attend or majors to pursue. This should serve as a guidance of what to expect in terms of the relationships between colleges and salary and between majors and salary. As a matter of fact, it might be more beneficial to increase the value in terms of salary of the colleges and majors that are not paid as much as the others. Rather than promoting for certain colleges and majors, it is more right to have the society to recognize and acknowledge the value of those that are not mainstream. We believe it is still more important for students to pursue a career about which they are passionate.

FUTURE WORK

To better understand the relationship between the college and salary in the future, we would like to drop the starting salary feature from our linear model regression. Currently, we are including the starting salary in the model to do better prediction. However, we hope to learn whether our prediction will be changed if we are only considering the features related to the colleges (such as types, region, etc.). Therefore, one of our future steps will be building another model that's not including the starting salary as a feature.

When we explored our dataset before modeling, we have found out that there's a strong correlation between school rankings and salary. Therefore, we have filtered out the colleges that are without ranking information in the dataset to train the better model. But the filtering decreased our data sizes. In the future, we are planning to find another dataset that includes more college ranking data, so that we will be able to train our model with a larger dataset that contains more colleges.

Currently, our major and college datasets are completely separated. We are hoping to find out another useful dataset to combine the original datasets together. In this case, we are able to find out whether a students' major or college have a larger impact on his/her future salary.

REFERENCES

1. Brewer, D., Eide, E., & Ehrenberg, R. (1999). Does It Pay to Attend an Elite Private College? Cross-Cohort Evidence on the Effects of College Type on Earnings. The Journal of Human Resources, 34(1), 104-123. doi:10.2307/146304
2. Freedman, L. (2013, June 28). The Developmental Disconnect in Choosing a Major: Why Institutions Should Prohibit Choice until Second Year. Retrieved from https://dus.psu.edu/mentor/2013/06/disconnect-choosing-major/
3. Lindsay, T. (2018, May 24). New Report: The U.S. Student-Loan Debt Crisis Is Even Worse Than We Thought. Retrieved from https://www.forbes.com/sites/tomlindsay/2018/05/24/new-report-the-u-s-student-loan-debt-crisis-is-even-worse-than-we-thought/#2b55f989e438
4. Velez, E. D. (2018, May 09). Newly Released Student Loan Data Bust Several Myths about Student Loan Repayment. Retrieved from https://evolllution.com/attracting-students/todays_learner/newly-released-student-loan-data-bust-several-myths-about-student-loan-repayment/
5. Weissmann, J. (2012, May 17). Does It Matter Where You Go to College? Retrieved from https://www.theatlantic.com/business/archive/2012/05/does-it-matter-where-you-go-to-college/257227/
6. Thomas, Scott L. and Zhang, Liang (2005, June). POST-BACCALAUREATE WAGE GROWTH WITHIN 4 YEARS OF GRADUATION: The Effects of College Quality and College Major

   https://www-jstor-org.offcampus.lib.washington.edu/stable/pdf/40197375.pdf?refreqid=excelsior%3Aa56d019a1a01a27b54305828484d07bd