Analysis of the Financial benefits of attending College in the USA

Mark Makris (19230705)

School of Computer Science

National University of Ireland, Galway

Supervisors

Dr. Desmond Chambers

In partial fulfillment of the requirements for the degree of

MSc in Computer Science (Data analytics)

July 4, 2020

DECLARATION I, MARK MAKRIS, do hereby declare that this thesis entitled analysis of the financial benefits of attending college in the USA is a bonafide record of the research work done by me for the award of MSc in Computer Science (Data analytics) from the National University of Ireland, Galway. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

Signature: ____

Abstract

In the United States obtaining a higher education is culturally considered a necessity to a financially profitable life. Given the data provided by the United States Department of Education, an analysis of the profitability of college is proposed. The first objective is to discover if the earnings after college outweigh the cost to attend. The analysis showed a clear increase in cost over time and earnings six years after enrolling in college did not show any increase in value. The second objective proposed is to understand what attributes impact earnings the most, first with regression of unique elements, and then with feed forward networks to evaluate groups of variables and their impacts. The conclusion of the thesis demonstrates that the regression and neural network analyses found that there is a general sense of increased earnings and that there are groups of attributes that are better at predicting future earnings. It would appear to be possible to predict earnings after college based on given inputs about the students and institutions they attend.

Keywords: Regression, Machine Learning, Feed forward, College, Debt, Earning

Contents

List of Figures

1. Introduction

The rising cost in tuition of four-year colleges in the United States is a concern to many. This raises the question of whether or not it is worth it to pay these huge amounts to attend college and incur great debt, as it is commonly believed that you need a four-year degree in order to succeed in life in America. I am looking to identify how much tuition is increasing and see whether or not it is in line with how much students are financially benefiting from college. Given that being able to earn money after college is one of the primary goals if not the primary goal, it is important to determine if college will pay for itself. It is also important to determine if profitability can be predicted. More specifically if a certain set of attributes could determine if the cost of college is worth the investment and outweighs the significant amount of debt generated during college. The likelihood of one attribute having a direct impact on earnings is unlikely but finding certain groups of attributes that are good predictors is more likely. Using a combination of statistics and machine learning models my goal is to find these attributes that can help make more informed decisions on whether or not college is worth the debt one is likely to incur, before making such a large investment.

2. Related Research

Below I have collected information from multiple different sources on the predicted earnings for students after college, default rates on student loans, and models to predict spending on loans.

Earnings

In an earlier study of the difference of salary earnings, after having taken student loans, there is a clear difference in the amount of money that will be earned depending on the academic majors or professional fields chosen, even though the cost of academic majors does not differ. For instance, in a least-squares means analysis for industry salary earnings, it was found that a real estate agent averaged out to be nearly one hundred and fifty thousand dollars a year, where amusement and recreation services workers was merely 22,000 dollars a year. This shows a vast range in how much one can earn based on industry chosen after college. Similarly, in the same study, there was evidence that based on the academic majors chosen, that business administration majors averaged about 59,000 dollars where anthropology averaged 23,000 dollars. (Donhardt, G. L., 2003)

In a study from New Zealand, they found that the amount of money earned at a median level position in 2001 was higher as the level of the position increased. There is also at a consistent rate of average future earnings from an intermediate vocational level degree to a higher degree. There is also evidence that the possible wages are higher as you progress in the field which will lead to more income over time. The rate at which income increases also falls off the older you get so one will most likely not be moving to a substantially higher income from where one begins. The study also showed it is evident that the amount of money earned between job fields can have large inequalities, and therefore some fields of study are much more profitable than others. The job fields with higher income may be ones with higher degrees, but this does not always show that there are ways to increase possible income with less possible debt as well. (Maré, D. C., & Liang, Y., 2006)

Default Rates

The number of students who default on their school loans may be increasing more than people realize. It could be possible that at the current rate of increase the default rate for the current year 2020 could be at forty percent. This is not an even forty percent across the board of students, as students in private institutions have twice the default rate of students at public institutions. For instance, after twelve years, private borrowers' default on their loans at 52% versus 26% for public borrowers. There is evidence that overall student debt numbers accumulated by students are somewhat irrelevant to determining if an individual will have a certain amount of debt as the influencing factors are more along the lines of student and institutional factors. The study also found that black students are defaulting at five times the rate of white students, and trends seen in for-profit colleges are the most concerning. (Scott-Clayton, J., 2018)

Another study showed that college debt is accrued by two thirds of graduating American university seniors in 2017, at an average of about thirty thousand dollars. This is one of the main reasons for research around American cost of college, debt, and expected profit of their degrees. The United States has been seeing a steady increase in the amount of debt being accrued in recent years and the study wants to see where it has been impacting the most. The study shows a default rate on student loans of five percent for students with a bachelor’s degree. The rate is even higher for associate degrees at 12%, 29% percent for certificates, and 23% for students who do not complete their course. Families with lower income are affected more by the debt accrued than others with higher incomes, along with different average debt amounts on a state by state basis.   (Student Debt and the Class of 2017, 2019)

Effects of Debt

A study completed in 2019, showed that the highest average loan amounts for students are found in the United Kingdom with an average above 50,000 pounds or 62,457 dollars followed by the United States. This has caused the study to investigate the mental effect on borrowers of having these substantial debts. It was found that debt can influence the future net worth of individuals, while also putting more strain on students while in school, which can lead to more dropouts. Ultimately it can widen the gap of economic inequalities. They also found that students with high student debt, tend to have much worse mental health scores, which can impact every aspect of their life, with an increase in loneliness and difficulty building social networks. The study concluded that debt could have an impact on future economic profitability and showed that there are more aspects to debt than just academic performance.   (Nissen, S., Hayward, B., & McManus, R., 2019)

Another study showed the rise in amounts of loans needed to pay for college is steadily increasing, and how the debt impacts groups differently. Some demographics that are impacted the most by student debt, are detailed in the report. From this study one can see that black and low-income students are some of the most impacted by loans, as they finish with debt more often than others. More than half of black students getting an associate degree finish with debt. The highest debt burdens are found at for-profit schools. All these debts have been found to decrease work satisfaction when entering the workforce, because of the debt weighing over them as well. When students incur more than 10,000 dollars in debt, it starts to have negative impacts on their life. On the other hand, having debt below 10,000 dollars can help push students in a positive manner while in college. The study also finds that students without debt are more likely to own homes and have lower mortgage rates as well. (Huelsman, M., 2015)

Debt Models

Another study showed it is always expected that some debt is expected going to college, but it has now become the primary way for Americans to go to college, but how much debt is too much, and how it’s impacting the way that people are living compared to their prior debt free life? The life cycle model is built on the future amount of money that the student will be making. This can have an underestimation effect on the true effect of the debt. This model does not take into account that life is unpredictable, and assumes that future income will never change, or that life expenses can change rapidly. Another model the study shares, is the eight percent rule. This model is built on the idea that students should not allocate more than eight percent of their future gross income to paying off their student loans, so that they can still afford other monthly expenses. Knowing that mortgage rates are expected to be 29% of gross income and 41% of gross income for all combined debt expenses. The part where this model does not completely account for all variables is that now these other debt rules are changing and becoming more diverse with time and may not apply to everyone or as many as it used to. (Baum, Sandy|Schwartz, & Saul., 2005)

Benefits of college

Some studies are that the benefits of college do not only come from the monetary or economic outcomes. A study from the university of Maine investigated some of these benefits. My research will not go in depth on topics outside of economics, but it is still relevant to bring in multiple perspectives. One of the arguments is that it can help a community as a whole and not just the individual who received the degree. Some of the benefits to the community are indirect in the form of increased tax sources and the ability to contribute more to the community through having a job, and not being dependent or draining on social systems. Some of the other community impacts are greater charitable donations and volunteer work, along with lower crime rates. (Trostel, 2015)

The study had many points of interest with their increase or decrease in performance based on college degrees earned. In terms of earnings they found that the average annual earnings are 134% higher and the average lifetime earnings is 114% higher. Some of the benefits include a 3.5 times lower chance of being in poverty, being 47% more likely to have health insurance from employment, and a 72% greater chance of having a retirement plan from employers. (Trostel, 2015)

A lot of these benefits come mainly from having a job and the study shows that job safety is greater with higher employment rates and greater job prestige. On top of all that they claim an increased probability of being happy. Although these benefits are all very good being able to earn money is a key factor in almost all of these which is another reason why I am looking at the financial impact of college. (Trostel, 2015)

Financial Aid

A study looking at the impact of financial aid on student debt found that the elasticity of school cost is only .229 which means that for every percent increase in cost than debt increases by .229 percent as well. They also investigated the impact of state grants on debt which resulted in a negative relation of .05, which means that grants decrease debt but by nearly no amount. In terms of simple debt numbers, the average debt found for students in 2011 was 22,799 dollars. When they included the non-debt students it was lowered to just 13,772 dollars. This study was also able to find that for every 100-point increase on the SAT score there was a 15.7 percent decrease in average student debt. (Monks, 2014)

In a study on student debt and hardship they found there is a need for aid, to prevent debt, to increase the quality of life and reduce health care risks of students. It has been found that students with over 25,000 dollars of debt have greater financial hardships. The students who are struggling to pay their debt also have different social and economic profiles. These students are more likely to be black, supporting dependents, and be unemployed. In regard to predicting who will have debt after college is an important defining factor. Completing college will not guarantee any future financial difficulties but the possibility is less. (Despard, 2016)

3. Source Data

College Scorecard

The data used for my research is coming from the United States Department of Education (DoED), which produces what they call the College Scorecard. They collect information, such as academics, students, and costs on colleges across America in order to develop new government policies, as well as helping high school graduates select which colleges to apply to. The DoED provides a website where you can compare different schools based on a multitude of factors, or simply by selecting a specific school to get a deeper understanding of what the schools offer. Most importantly, and for the purpose of this thesis the DoED also provides raw data files to the public going all the way back to 1996, on all the schools, with thousands of information fields for each school. Having these raw data files, will allow me to collect all the foundation data needed to complete my analysis.

Data breakdown

In order to use the data, I first need to have a complete understanding of the available data to break it down in order to identify the needed data points. There are somewhere around two thousand fields available, but not all of these are useful for our goal. To help with the analysis, these fields can be broken down into groups like academics, which are directly related to degrees awarded, programs offered, or earnings which indicate how much money is being made after leaving school. Only ten of these fields will be needed for my research.

Some other notable aspects of this information from the DoED is that the data is collected on a yearly basis, and therefore one will find separate sheets for each academic year, which can be aggregated to gather into a complete data set of 22 years of data. This data collection goes all the way back to the fall of 1996 and continues until the spring of 2018.

Another consideration reading these data points is that there are three categories for the data provided, either it is present, null, or privacy suppressed. Unfortunately, there is no explanation of why some data is suppressed or missing, which could lead to some bias in what data is available versus hidden for some unknown reason. An example of possible bias could be schools with unreasonably high costs could be hiding costs or something of that sort. That being said there is still plenty of data available to allow me to fully proceed with my research.

When it comes to data on the amount of debt coming out of school, the DoED calculate median for nearly all of their fields, which makes them somewhat useless when it comes to predicting desired future earnings as the median of a data set is not representative of the data as a whole. In addition, the only means for debt information are the rates of defaulting on loan payments within two years before 2011 and the default rate within three years after 2011. On the other hand, however, there is good data pertaining to how much students are making shortly after college, which will be used for most of the analysis. When it comes to values for defining schools and students, there is plenty of useful data that can be used to test ideas and show trends or predictions. Finally, when calculating the mean of each individual field it is important to take into account the number of students from each university in order to not skew the data towards smaller schools.

4. Preliminary Analysis

Averages over time

The simplest way to look at whether college is financially beneficial, is to calculate the average amount of money required to attend college, against the amount of money one will likely earn after finishing college. In order to do this, I looked at the years which had data on the average cost of school and took an average for each unique school year, by unique IDs. Then, I was able to plot that on a bar chart along with a linear plot overlaid. I follow the same process for the average earnings of students six years after enrolling which assumes two years in the workforce. While there is a slight difference in which years have cost data and future earnings data, I was still able to see the data trends as the overall trend is more important than small changes from one year to another.

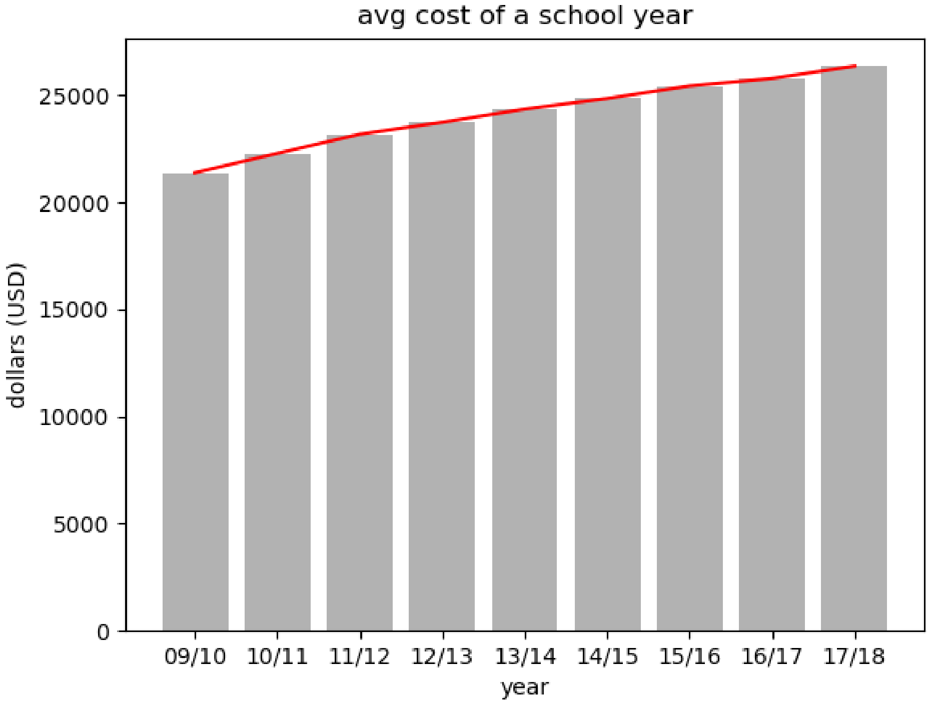


Figure 4.1

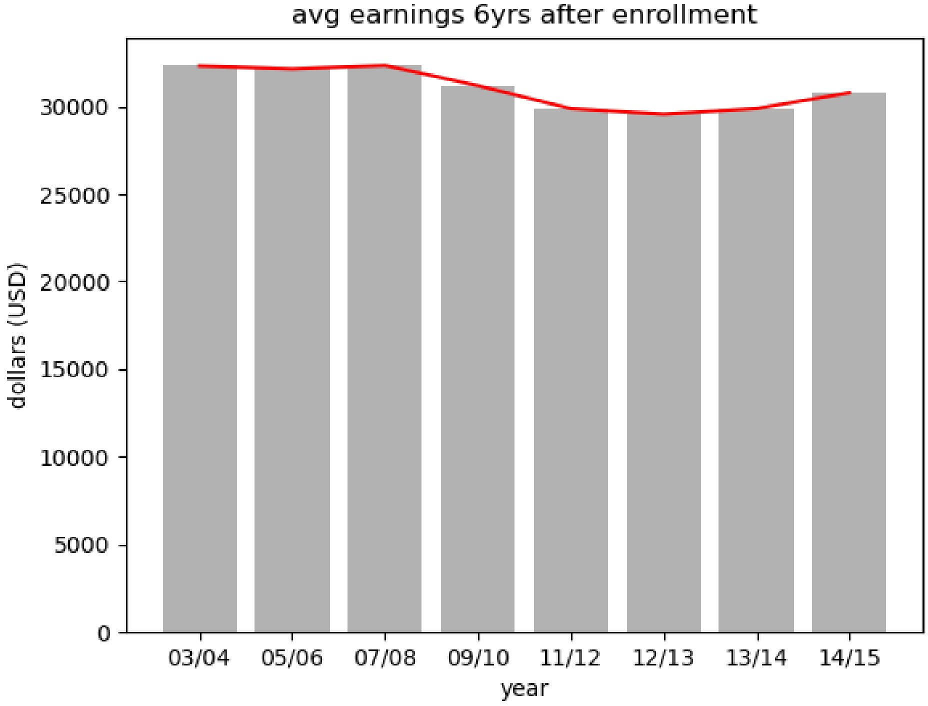


Figure 4.2

Figures 4.1 and 4.2 show the result of my first analysis, that over the past nine years one can see a steady increase in the cost to attend school for one academic year, while the amount one will likely earn shortly after leaving college is not, in fact it actually decreases. In other words, as the investment one puts into college increases, the benefits seem to be decreasing over time. This however is an analysis at a very high level, and does not get into any of the nuances of what parts of college are important to invest, in in order to receive the most financial benefit from college which I will look into further into this paper.

Now when looking at the change in default rate one does not see much of a change over the period of the analysis. Figure 4.3 looks like it is decreasing but one only sees an overall change of maybe one percent.

Key: 0.14 = 14%

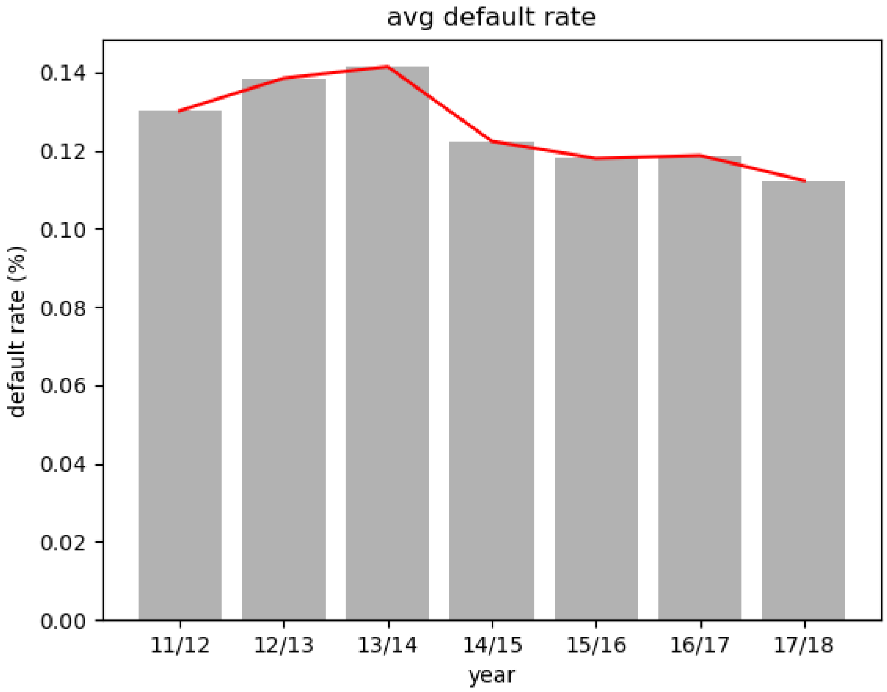


Figure 4.3

Paying off the cost of school

The average cost of one academic year in 2018 was $26,364 which includes tuition, room, board, and books. If we multiply this by four, which is the standard length of years it takes to complete a program, the total average bill is $105,456. In addition, fifty percent of students graduate from college with loans of between $4,107 and $23,198 to be paid off. This can be found using the average cumulative debt at the 75th percentile and the 25th percentile.

Next, we can calculate the time it would take to pay off that bill using the eight percent model (Baum, Sandy|Schwartz, & Saul, 2005) using the average amount of money earned six years after college enrollment, which would normally be two years after graduation, and be $30,771. If eight percent of that income is to be used for paying off debt, as proposed by the previous model, one can look at how long it will take to pay off the debt based on how large the remaining loan was. Obviously, this estimation does not take into account increase in pay or interest rate, I am just assuming those are similar enough to cancel each other out for this.

Seventy-fifth percentile:

23,198 / (30,771/100*8) = 9.4 years

Twenty-fifth percentile:

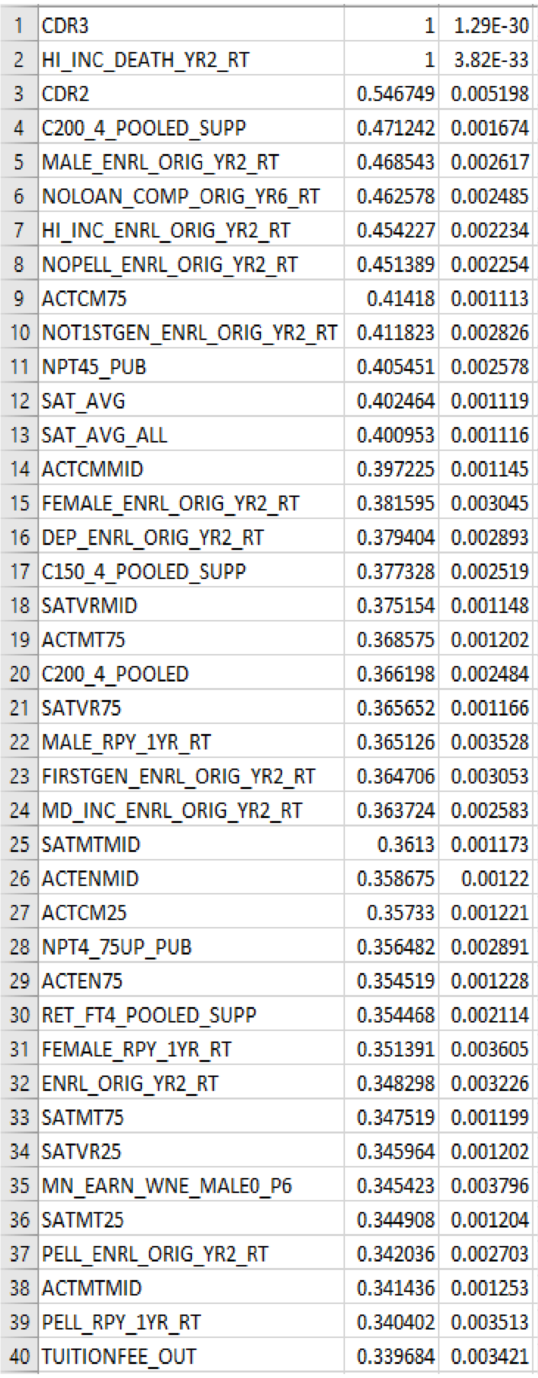
4,107 / (30,771/100*8) = 1.7 years

What is seen here is that fifty percent of students should be taking between 1.7 and 9.4 years to pay off their student loans, which may be financially worth it for some but not others depending on the outcome of the study.

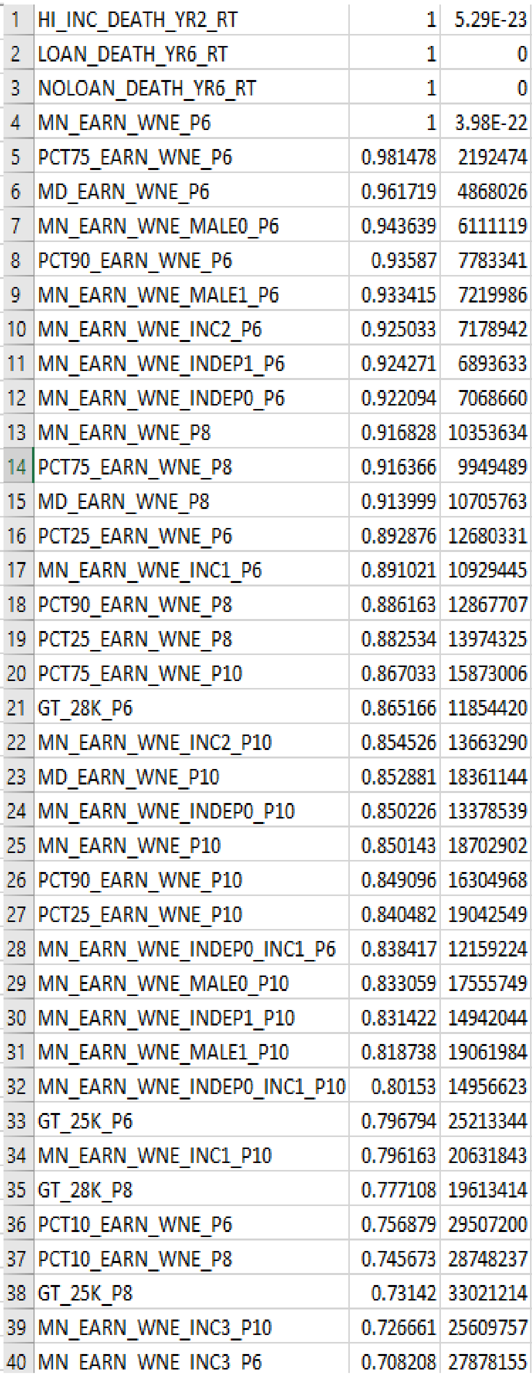
Regression

When looking at which fields like academic majors, demographics, or any other provided by the DoED will result in the highest return on investment, linear regression is one of the simplest ways to review the data to find out. If one or even multiple given fields can be shown to have the ability to increase the amount of money one is likely to earn after college, then one can learn to prioritize those fields when searching to find the right college to attend, or whether to attend at all. Therefore, I am looking to find a regression line resulting in which one variable can explain where a student will land in terms of future income with low variation.

In order to determine which fields are the most important to predicting different factors I had to run a regression algorithm on every other available field to see which ones predicted the desired output the best. This was implemented by using the Scikitlearn linear model and metrics packages. First, I had to aggregate all the years of data together into one file. Then I pulled every instance where the independent variable and the dependent variable for the current model both exist so that it would be usable in a linear model. Then with that set of available data, I split into a standard 60 / 40 split of training and testing, where the training had the first 60% and the testing had the last 40% of the data. From there I was able to use Scikitlearn to build a linear model and fit it to the training data. The fitted model is then used to predict the text data and compute some performance metrics. I chose to use mean square error and r squared. The mean square error explains the average amount of distance any point is from the given linear model and the r squared shows how much of the data is explained by the model. Then every attribute is added to a list which consists of the independent variable of the given model, its r squared value, and the mean square error. The list is then written to a csv file which can then be reordered to see which variables have the highest r squared values. The goal is to see which variables have over at least a 50% r squared value, otherwise if it is less than that it is not good at explaining the majority of the dependent variable and the linear model is somewhat useless. I expect the mean squared error should be smaller as the r squared increases in general. The results of the analysis are below in figure 4.3 and 4.4.



1	CDR3	1	1.29E-30
2	HI_INC_DEATH_YR2_RT	1	3.82E-33
3	CDR2	0.546749	0.005198
4	C200_4_POOLED_SUPP	0.471242	0.001674
5	MALE_ENRL_ORIG_YR2_RT	0.468543	0.002617
6	NOLOAN_COMP_ORIG_YR6_RT	0.462578	0.002485
7	HI_INC_ENRL_ORIG_YR2_RT	0.454227	0.002234
8	NOPELL_ENRL_ORIG_YR2_RT	0.451389	0.002254
9	ACTCM75	0.41418	0.001113
10	NOT1STGEN_ENRL_ORIG_YR2_RT	0.411823	0.002826
11	NPT45_PUB	0.405451	0.002578
12	SAT_AVG	0.402464	0.001119
13	SAT_AVG_ALL	0.400953	0.001116
14	ACTCMMID	0.397225	0.001145
15	FEMALE_ENRL_ORIG_YR2_RT	0.381595	0.003045
16	DEP_ENRL_ORIG_YR2_RT	0.379404	0.002893
17	C150_4_POOLED_SUPP	0.377328	0.002519
18	SATVRMID	0.375154	0.001148
19	ACTMT75	0.368575	0.001202
20	C200_4_POOLED	0.366198	0.002484
21	SATVR75	0.365652	0.001166
22	MALE_RPY_1YR_RT	0.365126	0.003528
23	FIRSTGEN_ENRL_ORIG_YR2_RT	0.364706	0.003053
24	MD_INC_ENRL_ORIG_YR2_RT	0.363724	0.002583
25	SATMTMID	0.3613	0.001173
26	ACTENMID	0.358675	0.00122
27	ACTCM25	0.35733	0.001221
28	NPT4_75UP_PUB	0.356482	0.002891
29	ACTEN75	0.354519	0.001228
30	RET_FT4_POOLED_SUPP	0.354468	0.002114
31	FEMALE_RPY_1YR_RT	0.351391	0.003605
32	ENRL_ORIG_YR2_RT	0.348298	0.003226
33	SATMT75	0.347519	0.001199
34	SATVR25	0.345964	0.001202
35	MN_EARN_WNE_MALE0_P6	0.345423	0.003796
36	SATMT25	0.344908	0.001204
37	PELL_ENRL_ORIG_YR2_RT	0.342036	0.002703
38	ACTMTMID	0.341436	0.001253
39	PELL_RPY_1YR_RT	0.340402	0.003513
40	TUITIONFEE_OUT	0.339684	0.003421



1	HI_INC_DEATH_YR2_RT	1	5.29E-23
2	LOAN_DEATH_YR6_RT	1	0
3	NOLOAN_DEATH_YR6_RT	1	0
4	MN_EARN_WNE_P6	1	3.98E-22
5	PCT75_EARN_WNE_P6	0.981478	2192474
6	MD_EARN_WNE_P6	0.961719	4868026
7	MN_EARN_WNE_MALE0_P6	0.943639	6111119
8	PCT90_EARN_WNE_P6	0.93587	7783341
9	MN_EARN_WNE_MALE1_P6	0.933415	7219986
10	MN_EARN_WNE_INC2_P6	0.925033	7178942
11	MN_EARN_WNE_INDEP1_P6	0.924271	6893633
12	MN_EARN_WNE_INDEP0_P6	0.922094	7068660
13	MN_EARN_WNE_P8	0.916828	10353634
14	PCT75_EARN_WNE_P8	0.916366	9949489
15	MD_EARN_WNE_P8	0.913999	10705763
16	PCT25_EARN_WNE_P6	0.892876	12680331
17	MN_EARN_WNE_INC1_P6	0.891021	10929445
18	PCT90_EARN_WNE_P8	0.886163	12867707
19	PCT25_EARN_WNE_P8	0.882534	13974325
20	PCT75_EARN_WNE_P10	0.867033	15873006
21	GT_28K_P6	0.865166	11854420
22	MN_EARN_WNE_INC2_P10	0.854526	13663290
23	MD_EARN_WNE_P10	0.852881	18361144
24	MN_EARN_WNE_INDEP0_P10	0.850226	13378539
25	MN_EARN_WNE_P10	0.850143	18702902
26	PCT90_EARN_WNE_P10	0.849096	16304968
27	PCT25_EARN_WNE_P10	0.840482	19042549
28	MN_EARN_WNE_INDEP0_INC1_P6	0.838417	12159224
29	MN_EARN_WNE_MALE0_P10	0.833059	17555749
30	MN_EARN_WNE_INDEP1_P10	0.831422	14942044
31	MN_EARN_WNE_MALE1_P10	0.818738	19061984
32	MN_EARN_WNE_INDEP0_INC1_P10	0.80153	14956623
33	GT_25K_P6	0.796794	25213344
34	MN_EARN_WNE_INC1_P10	0.796163	20631843
35	GT_28K_P8	0.777108	19613414
36	PCT10_EARN_WNE_P6	0.756879	29507200
37	PCT10_EARN_WNE_P8	0.745673	28748237
38	GT_25K_P8	0.73142	33021214
39	MN_EARN_WNE_INC3_P10	0.726661	25609757
40	MN_EARN_WNE_INC3_P6	0.708208	27878155

Figure 4.4

Figure 4.5

Each row of the figures 4.3 and 4.4 show the id, variable name, r squared value, and error. Given the output of the linear model ordered from highest to lowest r squared values in figures 4.3 and 4.4 one can learn about the effectiveness of the linear model in these instances. The image only shows the top 40 variables as there are not any variables further down that explain their data well. One can see that on the left where the regression is predicting the loan default rate within three years there are no variables that can truly explain what is seen in loan defaults. Unless it is predicting itself or using a variable with almost no present or available data, everything is only explaining under fifty percent of the data, and it cannot truly be used to accurately predict what I am looking for. The same goes for modeling the earnings six years after enrollment. The only difference is that there are a lot of variables getting very good coverage with high r squared values. The problem with those variables is that they are also variables pertaining to how much a student will earn so they would be considered dependent variables not independent which is why they are not useful.

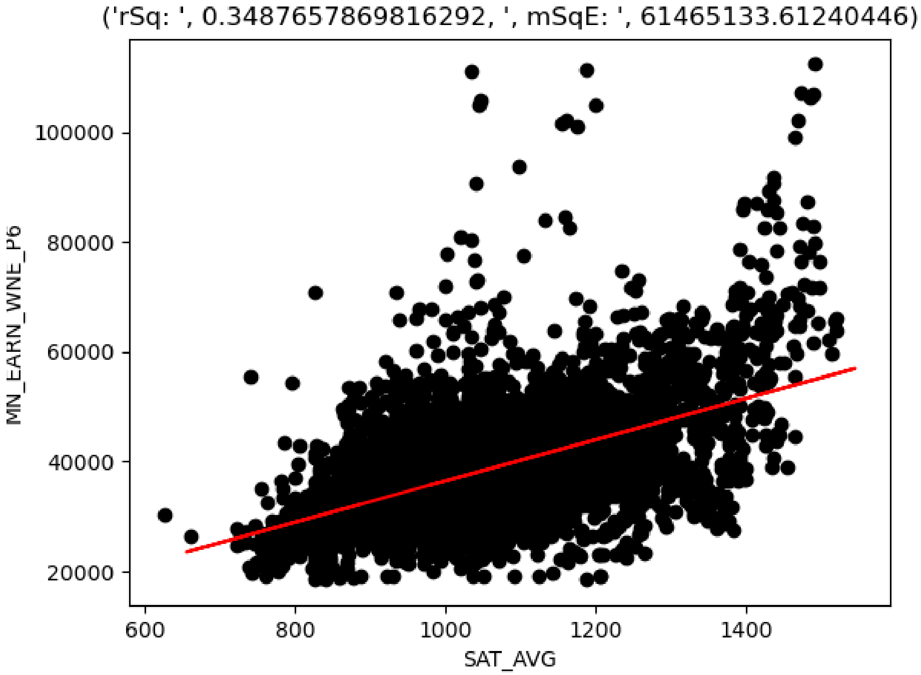


Figure 4.6

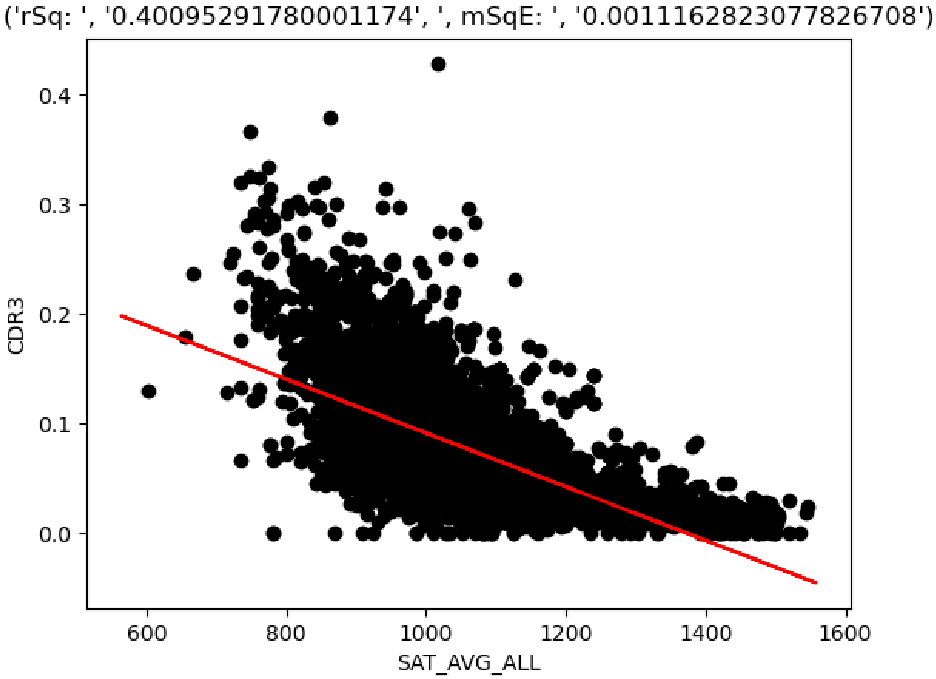


Figure 4.7

Even though the linear model is too simple to explain loan default rates or future expected earnings there is still useful information that can be gleaned from visualizing the models. A good example of this is the average SAT scores for students in predicting the future earnings in figure 4.5 and the default rates in figure 4.6. Starting with figure 4.5 one sees that the linear model goes straight through the selected data, but there is only a 40% r squared value so it could never fully explain the future earnings. That does not mean that it does not explain anything. Seeing that the line understands the direction or flow of the data means that it can be used as one piece of understanding the data along with many other variables that understand at least one piece of what I am trying to predict. We see from figures 4.5 and 4.6 that using multiple linear regression would not be helpful either. Normally when there is some curve or movement that is more complex than a simple linear model, more complex multiple linear models should be used. However, one can still see that there is a very large amount of variation in the data that would never fully be able to be explained.

5. Detailed Analysis and Results

Neural networks

Given that there are many variables that explain some part of the given DoED data as shown with the regression models from the previous section, I should be able to combine groups of those fields, together to get a better understanding of the DoED data. Trying to combine all the variables together into one function could include many variables that are not adding anything a neural network can learn from. It is more reasonable to create a smaller network which can focus in on smaller groups of variables that can pinpoint where the information is coming from within the DoED data. That is why I decided to create functions based on the subcategories of fields given by the Department of Education. There are nine listed categories given:

- Academics
- Admissions
- Aid
- Completion
- Cost
- Earnings
- Repayment
- School
- Student

Not all of these groups are useful for predicting future earnings. For example, the earnings category would be predicting itself which gives us no information to learn from. Also, the repayment category will not be used as a predictor as it would be better used as something the earnings would be predicting. Another unneeded category would be the school section as it does not contain information that would be useful as it pertains to school websites, school names, and geographical data.

Given that a neural network is meant to classify into classes rather than output unique values I will use the DoED's income brackets as my classes. The income brackets are broken down into the three classes listed below:

- Low income: < $30,000
- Middle income: $30,001-$75,000
- High income: $75,001 +

The DoED's data needs to be given new columns to store the newly defined brackets above. This new column will be used to classify the data instead of outputting a numerical value like regression would. By doing this it will allow me to understand the income bracket one would most likely be put into after college. The incomes are defined as integers to be classified in three new variables called lowClass, middleClass, and upperClass. Each row is then noted with a 1 for the matching income bracket and a 0 for the remaining two.

A fully connected neural network is an extremely demanding task on processing time, all the years of data needs to be reduced as much as possible. To begin, all years which have information on future earnings missing were removed. This still left 41,000 rows of information which is more than enough to build a good network. Also, because there are a lot of missing data points within the reduced data set, these missing data points were replaced with the average of its respective field as is standard practice, and should have little to no effect on the outcome of the network. This operation allows for a complete dataset to be processed, which can then be subdivided into smaller sets using the DoED's categories previously listed.

When trying to understand how well a neural network performs one must understand the random chance of choosing the correct class. Given that there are three classes some may assume there is a thirty three percent chance of choosing the correct class. Although not all three classes are filled equally. The probability distribution of the three classes are listed below:

- Low class: 53.9%
- Middle class: 45.3%
- Upper class: 0.8%

The upper class is nearly empty, and the network can easily learn to never guess the upper class is the correct class. So, when analyzing how well it is performing, a good indicator is an accuracy of above 54% is better than random.

These models will be using a combination of rectified linear units and SoftMax activation functions as these had the best overall performance. The hidden layers use the rectifier and the output layer uses SoftMax to choose the income bracket. The loss function used will be categorical cross entropy and the Adam optimizer. The performance metric is accuracy, and the batch size will be sixty-four for best performance. The structure of the model's hidden layers will be mostly what changes in the different models created along with how many epochs are used. Each subset of data will be analyzed individually with accuracy plots and loss plots.

Academics

The academics category contains information about the percentage of each academic program held within each school along with the average lengths of time it would take to complete the courses described. The descriptions for the five largest programs are recorded with their program IDs and respective titles. As this title is uniquely tied to the program ID, and because I need floats to process a neural network the string titles can be removed.

The academic category data set had 240 input variables and 3 output variables for the income brackets. The best accuracy for this data was found with a feedforward architecture that had 2 hidden layers. The first hidden layer had to have 100 nodes and the second had to have 30 nodes. This produced a model with 84% accuracy over 100 epochs.

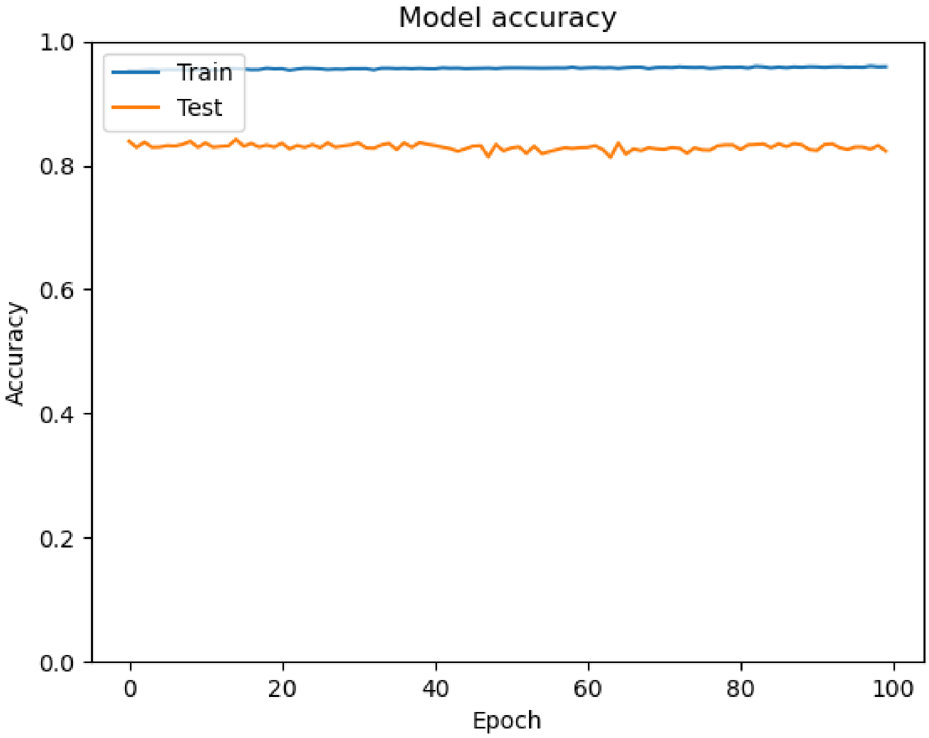


Figure 5.1

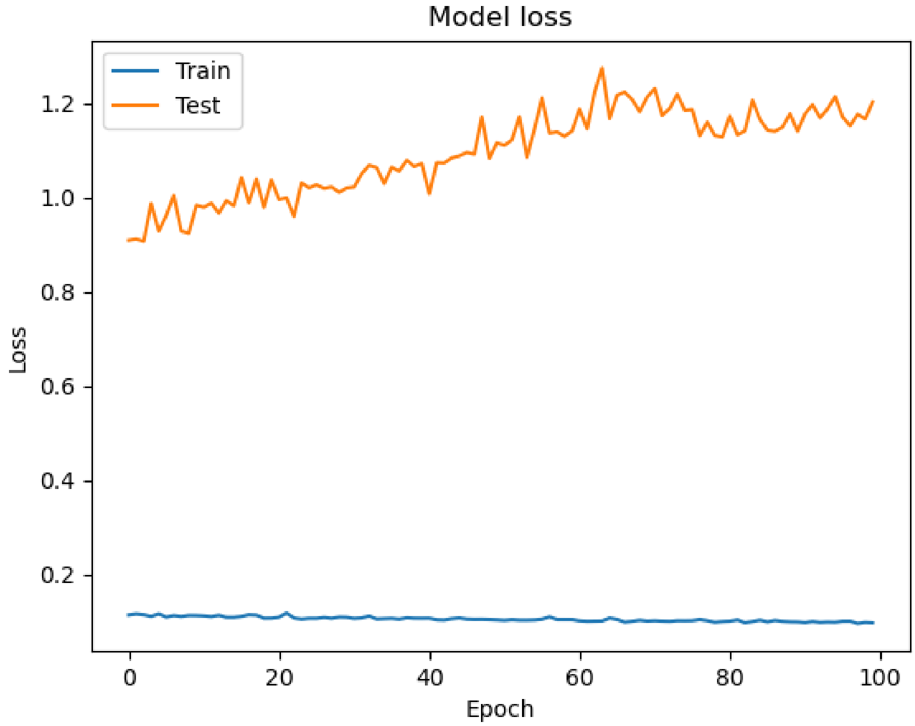


Figure 5.2

Admissions

The admissions category covers information on standardized tests used to judge whether a student is qualified to enter an institution. This includes SAT scores, ACT scores, and admission rates of institutions. This is the smallest set used.

The admissions category data set had 24 input variables and 3 output variables for the income brackets. The best accuracy for this data was found with a feedforward architecture that had 1 hidden layer. The hidden layer had to have 12 nodes. This produced a model with 74% accuracy over 100 epochs.

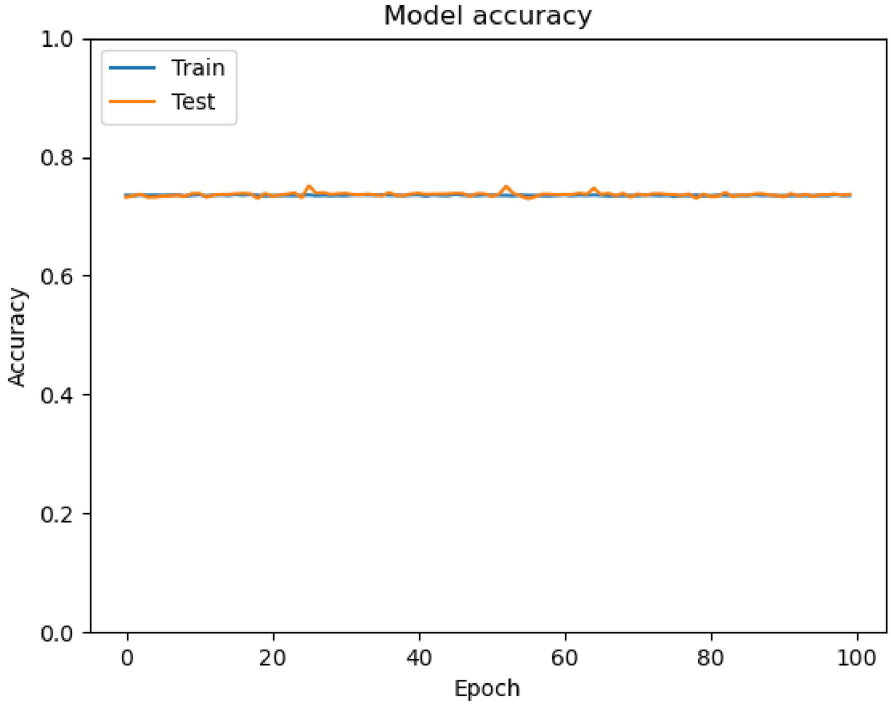


Figure 5.3

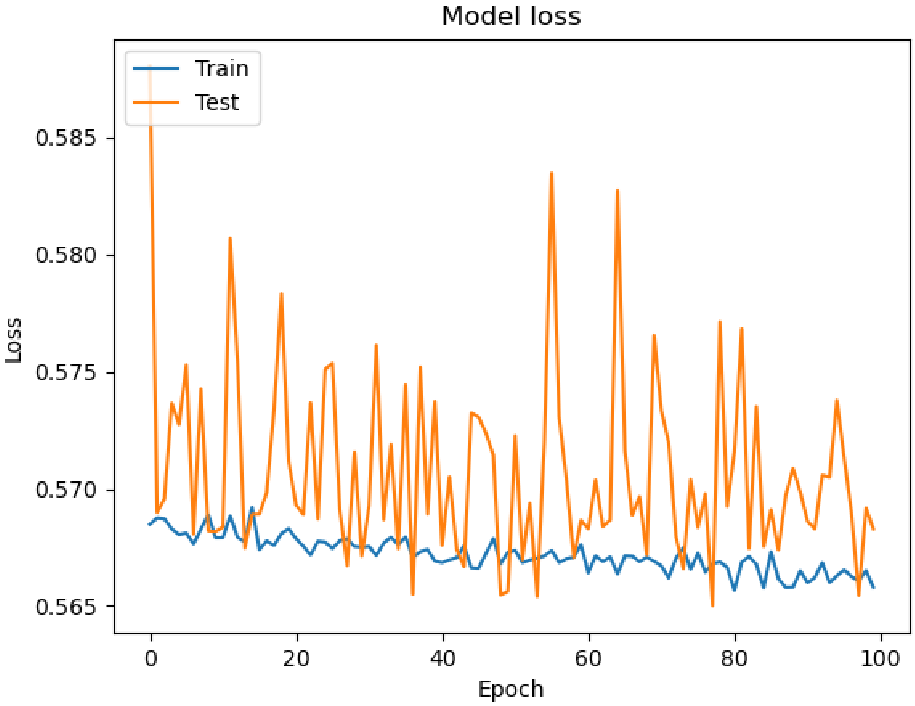


Figure 5.4

Aid

The aid category has information on students that use loans to pay for school and the amount of debt accrued by the end of college. This includes grant rates and median debts of different categories of students. It also looks at aid based on demographics or income levels.

The academic category data set had 39 input variables and 3 output variables for the income brackets. The best accuracy for this data was found with a feedforward architecture that had 1 hidden layer. The hidden layer had to have 20 nodes. This produced a model with 76% accuracy over 20 epochs.

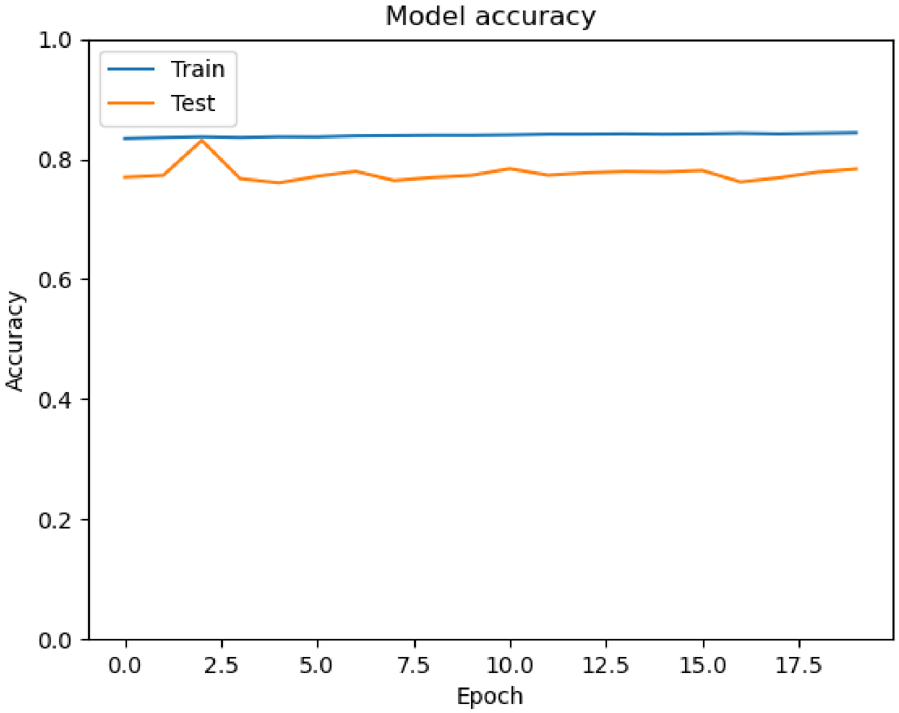


Figure 5.5

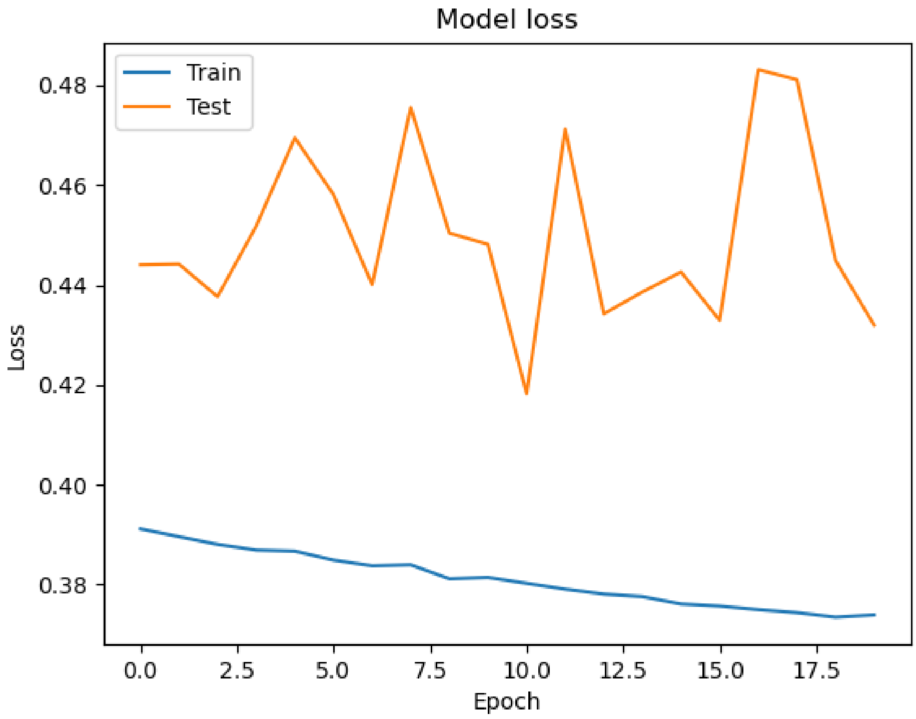


Figure 5.6

Completion

The completion category consists of information on how many years it takes students to complete. This is by far the largest category, larger than all the rest combined. The only value that was removed was the year of completion which was a date as the length of time was already recorded and this would have been duplicate information. There is data on demographics, school types, and income types in regard to completion of school.

The completion category data set had 1212 input variables and 3 output variables for the income brackets. The best accuracy for this data was found with a feedforward architecture that had 1 hidden layer. The hidden layer had to have 20 nodes and the second had to have 20 nodes. This produced a model with 85% accuracy over 50 epochs.

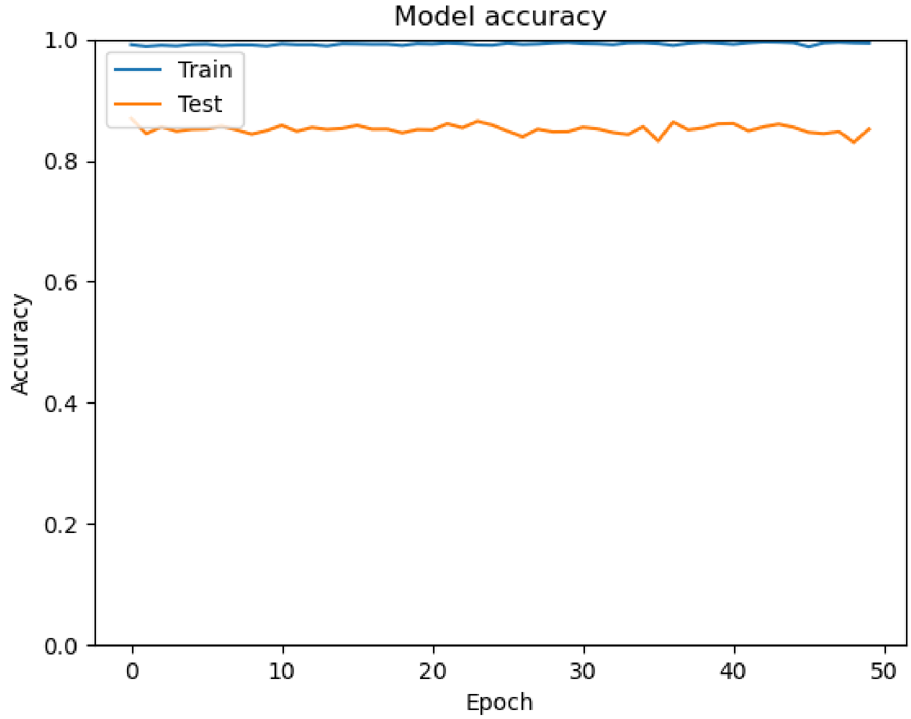


Figure 5.7

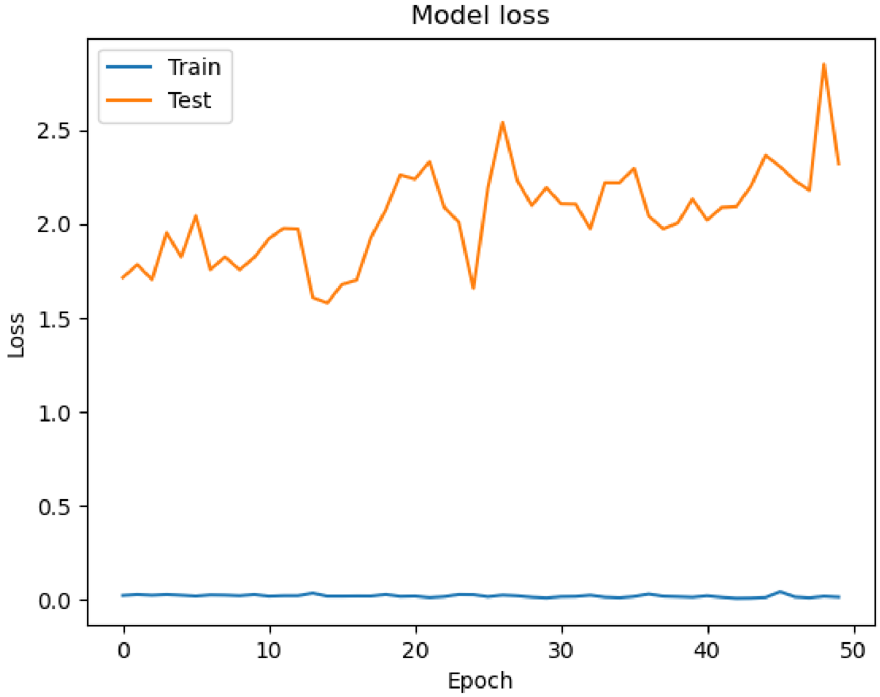


Figure 5.8

Cost

The cost category includes the price of schools in different categories, the price of books, and other supplies for an academic school year. This is one of the smaller sets as it is only concerned with pricing of schools and not student demographics or income levels.

The cost category data set had 76 input variables and 3 output variables for the income brackets. The best accuracy for this data was found with a feedforward architecture that had 2 hidden layers. The first hidden layer had to have 50 nodes and the second had to have 15 nodes. This produced a model with 75% accuracy over 100 epochs.

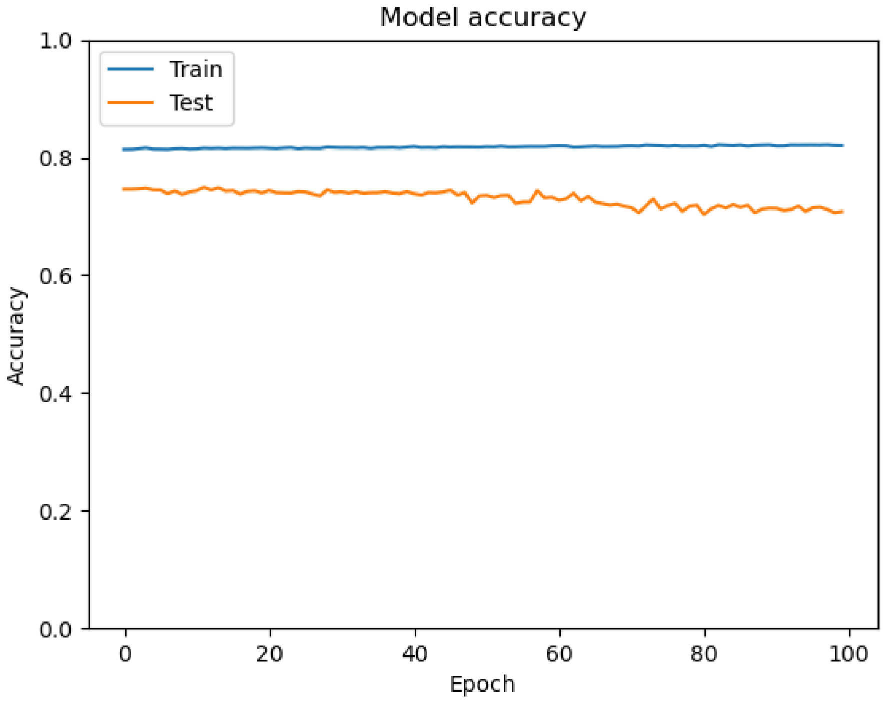


Figure 5.9

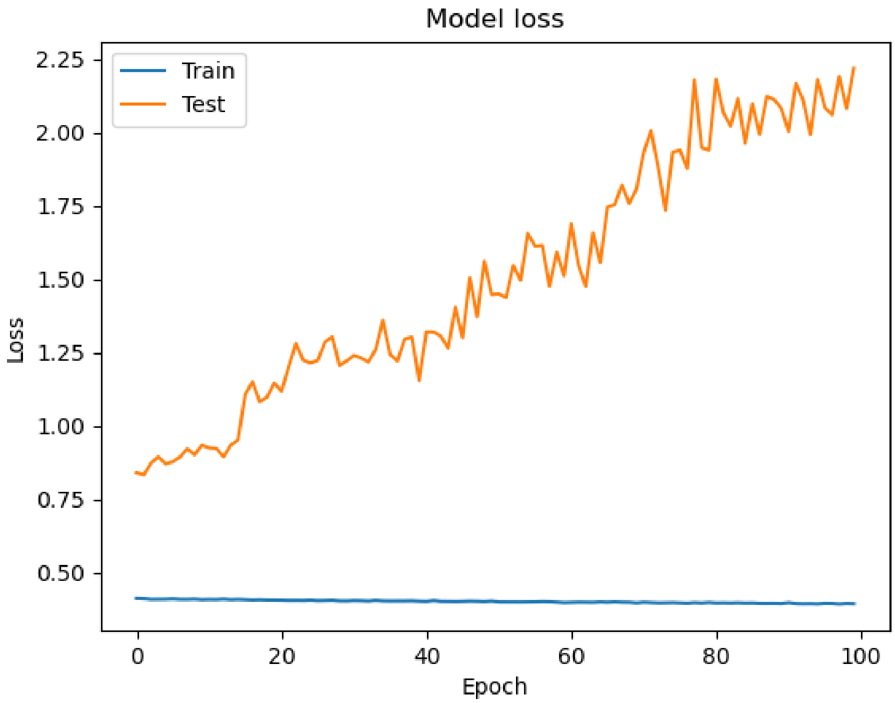


Figure 5.10

Student

The student category is the last category which includes information on the students themselves who are going into these institutions. It covers information such as: first-generation student, ethnicity, and backgrounds. It also covers income of the student's family when entering.

The student category data set had 116 input variables and 3 output variables for the income brackets. The best accuracy for this data was found with a feedforward architecture that had 2 hidden layers. The first hidden layer had to have 60 nodes and the second had to have 30 nodes. This produced a model with 84% accuracy over 50 epochs.

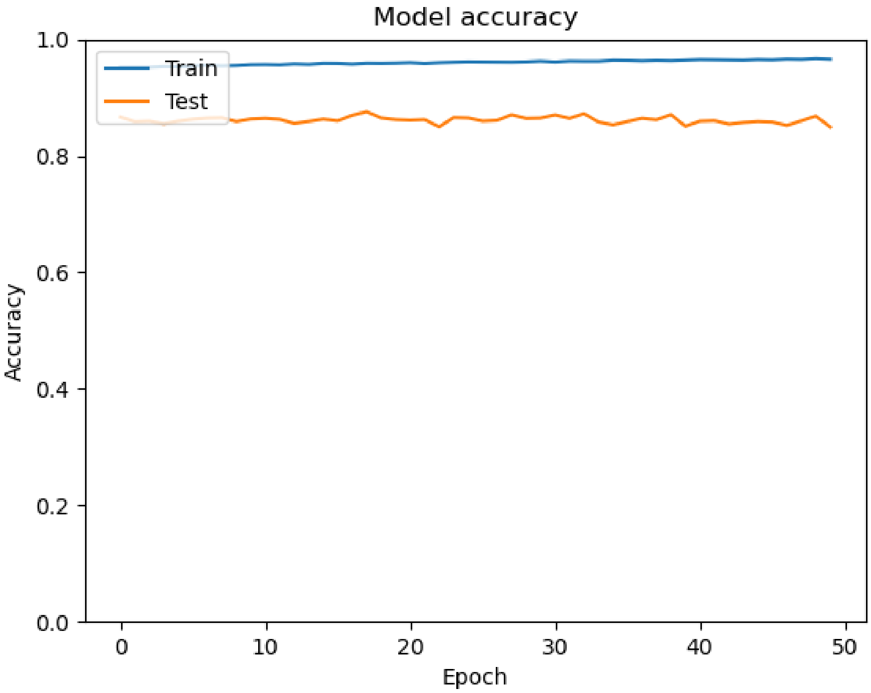


Figure 5.11

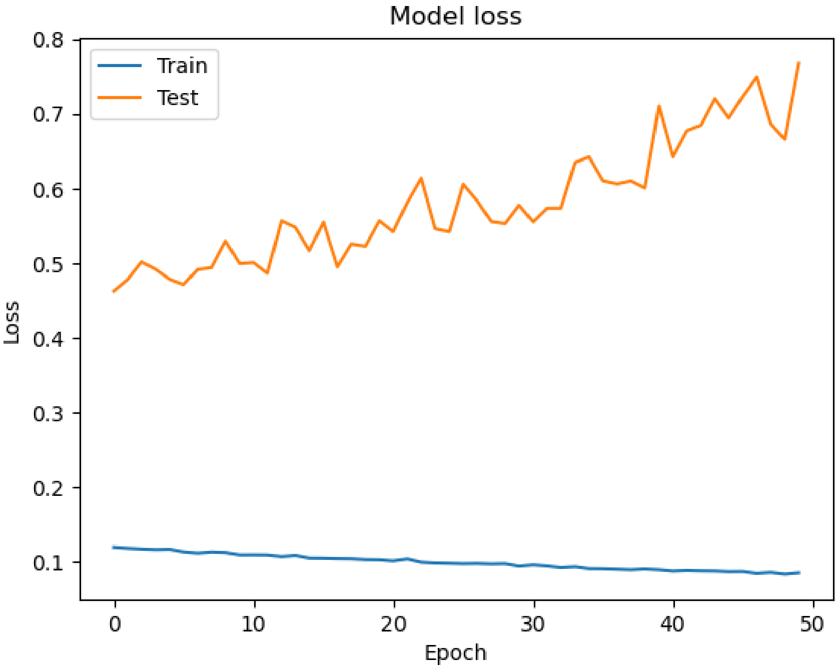


Figure 5.12

Neural Networks Results

Not all the data sets were able to produce the same accuracy in class prediction. The ranking of the classes from best to worst prediction accuracy are listed below:

1. Student: 86.0%
2. Completion: 85.4%
3. Academics: 84.5%
4. Aid 75.6%
5. Cost: 74.8%
6. Admissions: 73.9%

All of these datasets varied in size from having just a few dozen inputs to over a thousand. Although, there was not a clear correlation to the size of the dataset to the performance of the models. This allows me to know that the accuracy is not based on the number of variables but on the content of the variables. It can also be seen that the accuracies of the datasets are able to be broken down into two main groups. The top three sets student, completion, and academics have about a 1% difference and the bottom three aid, cost, and admissions have about a 1.5% difference between them. Although, from the top three to the bottom three there is a large 10% difference which is why it makes sense to group the top three sets together and the bottom three sets together.

The lower performing group consisted of the datasets of aid, cost, and admissions which show good performance compared to a random guess with an outcome of about 75% accuracy. This is still 25% off of the unrealistic set goal of 100% accuracy. This shows that there is a connection between these groups and the future earnings after college and should not be ignored when applying for college, but the other group did even better. With student, completion, and academics having a significantly higher performance at 85% accuracy I see much more influence from these groups on the future earnings compared to the lower performing group. Although all data sets show influence on earnings, this shows us that there is reason to focus on these higher performing groups slightly more when determining the financial worth of college.

6. Evaluation of Results

Beginning with the preliminary analysis here was an increase in school costs and consistent earnings after college with no clear increase or decrease over the past few years. This lines up with much of the research done in terms of where we see higher prices with less return on investment in future earnings. It would make sense that debt may be getting a bit out of control like some studies have said, because students are not making enough earnings to pay off school debt. The point where the initial analysis is different is the downward trend in default rates within two years. The research pointed towards increases in default rates which could have discrepancies with our data, because of the time frame or scope of the studies which could be different.

The research on how much debt Americans are coming out of school with is nearly identical to what the data found, at an average debt of about 30,000 dollars. This will then take students years to pay off with many taking over a decade to repay using one of the many models like the 8% model for debt payoff. Having these matching figures gives more credibility to our DoED's data outputs as it is the same as multiple studies.

There were some differences in one of the studies which showed some lower debt rates than what I found. The average debt of 22,799 dollars (Monks, 2014) is about 76% of the 30,000 dollars found could be from a few reasons. One of which is the difference in the study using an older year of data and from a different source than the DoED. The fact that it is a few years old would only explain maybe a few percent difference at most. There are many of factors that influence the debt accrued by students and it is difficult to determine why the number is much lower than what I discovered by the true value is likely somewhere between the number they found and what I have found.

The regression analysis revealed that there were a lot of factors that have partial significance in predicting factors like earnings and default rates. Most of the best explanatory variables had a lot to do with test performance, which was not brought up in the related research, but it could have some correlation that cannot be seen in the regression model with attributes like race or income levels.

The network models showed that there are groups of fields that will have some impact on what future earnings will be. This does support some of the related research done regarding inequalities in earnings. I found that attributes about the student in terms of their demographic or income, completion of programs, and what academic programs were taken had the highest income bracket prediction accuracy. If black and low-income students are really at such a disadvantage, then it would make sense that those fields would be good at predicting the future earnings. Then the performance of the academics set shows that there are some differences in earnings from different academic programs like engineering compared to a liberal art. This backs up the idea that choosing the right major will have a great impact on future earnings. There was no related research to compare to the completion of the programs, but it would make sense that completing the degree one is pursuing would make one more money from being able to get a job in that field.

When comparing these results to studies done in other countries, we can see that the United States is not the only country that may need some analysis on whether college is financially worth the investment. We saw prices for schools in the United Kingdom as an example of a location that has rising prices. If they were able to find information similar to the

DoED for other countries, we could likely do similar tests as done in this paper. We could then look at what groups are unique to each country that take on much more debt like we find in the United States where one's background impacts the profitability of college for them.

The results in accuracy of my networks also have some interesting comparisons to the elasticity found in different fields and the amount of debt it influences. For example, the cost of a school has elasticity of on .229, the impact of grants only had a negative elasticity of .05, and the performance on an SAT score has an elasticity of negative 15.7 for every 100 points scored. (Monks, 2014) What we see in this is that categories like cost and aid had a low elasticity and had lower accuracies in predicting earnings where the category of student academics we see a larger elasticity and higher performance in future earning accuracies. This shows some correlation in the elasticity of certain categories and the ability to use those categories to predict with higher accuracies.

Regarding future research on this data there is more that can be done. If the right variables are chosen, then a better network should be able to be made to predict the future earnings of graduates. This goes beyond the scope of this project, but if enough time is spent trying different combinations of variables from different groups then accuracy should be able to increase potentially substantially.

Another option for future research would be looking into other outputs of college participation like the research done on the benefits of college that are not simply future earnings. The only problem with this would be the limit in the available variables in the DoED's dataset. It lacks the large variety of output variables that would allow for an analysis of benefits outside of future earnings. A lot of the benefits listed as outside of just economic variables like loans and community benefits are related to the amount of money an individual earns. This is another reason why focusing the scope of this paper on economic outputs is still a good measure.

7. Conclusion

What is seen from the initial analysis is that the return on investment of college is losing value over time. This is the first sign that college may not be worth the money to attend an institution and the debt that comes along with it. It would appear that there are some variables that are clearly important in defining the worth of college for students. One can see the most influential groups from out neural network sets were information about the student themselves, their completion of programs, and the programs chosen by the students. From this one can advise students to make sure that they are choosing a good program and making sure they complete it if they wish to make earnings from their degree. The negative effects of having accumulated lots of debt from school can have lasting effects and more time should be spent deciding whether college is financially worthwhile and should be attended. Going to a four-year college is not always the right plan for everyone financially.

8. References

Baum, Sandy|Schwartz, & Saul. (2005, November 30). How Much Debt Is Too Much? Defining Benchmarks for Manageable Student Debt. Retrieved from https://eric.ed.gov/?id=ED562688

"College Scorecard Data." *College Scorecard*, collegescorecard.ed.gov/data/.

Despard, Mathieu R, et al. "Student Debt and Hardship: Evidence from a Large Sample of Low- and Moderate-Income Households." *Children and Youth Services Review*, vol. 70, 2016, pp. 8–18., doi:10.1016/j.childyouth.2016.09.001.

Donhardt, G. L. (2003, November 30). Post-Graduation Economic Status of Master's Degree Recipients: A Study of Earnings and Student Debt. Retrieved from https://eric.ed.gov/?id=EJ965769

Huelsman, M. (2015). The Debt Divide: The Racial and Class Bias Behind the "New Normal" of Student Borrowing. Retrieved from https://www.demos.org/research/debt-divide-racial-and-class-bias-behind-new-normal-student-borrowing

Maré, D. C., & Liang, Y. (2006). Labour Market Outcomes for Young Graduates Part A: Main Report. Retrieved from http://motu-www.motu.org.nz/wpapers/06_06.pdf

Monks, James, et al. "The Role of Institutional and State Aid Policies in Average Student Debt." *The ANNALS of the American Academy of Political and Social Science*, vol. 655, no. 1, 2014, pp. 123–142., doi:10.1177/0002716214539093.

Nissen, S., Hayward, B., & McManus, R. (2019). Student debt and wellbeing: a research agenda. Retrieved from https://www.tandfonline.com/doi/pdf/10.1080/1177083X.2019.1614635

Scott-Clayton, J. (2018, May 15). The looming student loan default crisis is worse than we thought. Retrieved from https://www.brookings.edu/research/the-looming-student-loan-default-crisis-is-worse-than-we-thought/

Student Debt and the Class of 2017. (2019, July 15). Retrieved from https://ticas.org/affordability-2/student-debt-and-class-2017

Trostel, P. A. (2015). LUMINA ISSUE PAPERS. Retrieved July 17, 2020, from https://www.luminafoundation.org/files/resources/its-not-just-the-money.pdf