Effect of Household Income on a Student's Educational Attainment

Aiden Chiang

ABSTRACT

There is no doubt that a student's background and household characteristics have a major impact on their education. Access to technology, a quiet place to study and access to extracurriculars outside of school all contribute to the successful development of a student. Often, a student's household income is cited as the largest contributing factor to their educational achievements when compared to all other characteristics of a student, yet rarely is the gap quantified. Using statistical analysis and feature importance analysis on a regression with deep neural network model, this study finds that a student's household effect has a significant effect on a student's educational attainment and plays a significantly larger role than nearly all other characteristics of a student.

INTRODUCTION

The mantra that all children in the US are entitled to an equal education is one that has been spread wide and far by numerous public figures. Indeed, it is a very agreeable statement, as a society we believe that all children should have the opportunities to develop and learn, to become who they choose to be. Yet despite all this talk, it remains just that, talk. Inequality in education is still widespread in the US and though attempts are made to address this, they are few and far between. Whether it be the inability to agree on a solution or the inability to find the cause of the problem, inequality in education seems to have been forgotten, having fallen out of the public's attention and the politicians' agendas. This study serves two different purposes. The first purpose of this paper is to quantify the gap between economically disadvantaged (ECD) students and non-economically disadvantaged (NEC) students, in addition to quantitatively comparing household income as a factor affecting a student's educational attainment to other characteristics defining a student. The second purpose of this study is to raise awareness for the rampant inequality that exists within the US education system in addition to the significant damages it has on US students. Previous studies have shown that a student's family's income has a statistically significant effect on a student's performance (CITE SOURCES HERE). These studies focus solely on the relationship between the student's family's income and their educational achievement; however, they fail to consider effects of other characteristics of students such as race, gender, the percentage of similar peers, etc. This study intends to address the other characteristics with a feature importance analysis on a regression with deep neural network model in order to quantitatively measure the size of the effect household income has on a student's educational performance when compared to other characteristics.

Data

Two different datasets were used in this study. The first dataset used was obtained from Harvard. Each row contains the specific family income for a student, along with various measures of reading and mathematical attainment. The second dataset was obtained from the Stanford Education Data Archive (SEDA). The SEDA dataset contains the average test scores for students in grades three through eight in reading and math for the 2008-09 through the 2017-18 school years. Each row in the dataset corresponds to a specific school, subject, grade, and year followed by the average test scores, standard error, and total number surveyed for the entire school, and each subgroup (race, gender, economically disadvantaged). The average test scores in the SEDA dataset were scaled such that the average test score value corresponds to the grade level of the same value. For example, a test score of three, would be the average score for all students in third grade.

In the first portion of our study, we directly analyzed the relationship between a student's household income and their educational attainment. To prepare the Harvard dataset for our analysis we decided to use the SAT reading sub score, SAT math sub score, and total SAT score due to the widespread use and popularity of the SAT test. The other measures for reading and mathematical attainment were then dropped along with the remaining rows that contained NaN values. As for the SEDA dataset, in preparation for our statistical analysis we dropped all columns except for subject, grade level, and mean test scores for three groups (all, economically disadvantaged, and non-economically disadvantaged).

For the second portion of our study, we utilized a deep neural network. To prepare the dataset, the average test score for the entire school, all school identifiers and standard error columns were removed from the dataset. The percentage for each subgroup was then calculated by taking the column containing the number surveyed for each subgroup and dividing it by the total surveyed for the school. Each column corresponding to a subgroup percentage was then concatenated onto the end of the DataFrame, and the total surveyed for each subgroup and the entire school was then dropped.

Using one-hot encoding, a new DataFrame was created from the SEDA DataFrame with each row corresponding to a specific subject, grade, year, and subgroup with its respective average test score. The remaining rows with NaN values were dropped.

Methodology

In this section, we will take an in-depth look into how we analyzed and measured the impact of a student's household income on their educational attainment. There are multiple approaches to measuring this effect, in this study two approaches were used. In the first approach, we directly analyzed the datasets for a relationship between household income and educational attainment. We used Pandas to analyze the data and MatPlotLib to visualize the data. In the second approach we used TensorFlow to construct a regression with deep neural network model. We then trained the model using our SEDA dataset which we one-hot encoded, and then preformed a feature importance analysis on our trained model to identify the factors that played the largest

role in determining a student's educational attainment.

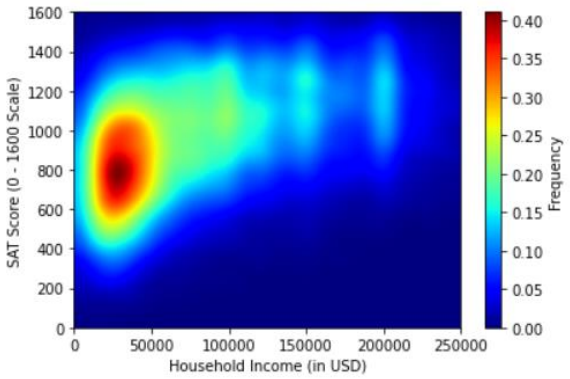

Figure 1: Household Income vs SAT Score

In the first part of this study, we wanted to directly compare the relationship, if any, between household income and educational attainment. Using the Harvard dataset, we were able to create a heatmap as shown in Figure 1 by plotting household income against SAT score. From the heatmap, we can see a central concentration of datapoints at around the halfway mark between $0 and $50,000 household income and the concentration of data points decreasing as income increases. There is also a visible upward trend with SAT score as household income increases. This is evident when we compare the average SAT scores of students with under 50,000 USD Household Income, 1140, with the average SAT scores of students with $200,000 to $250,000 USD Household income, 1316. This is an increase of 176 points out of the SAT's scale of 1600, a 15.43859% increase in SAT score.

With the SEDA dataset, we were able to create three different graphs. The first two graphs compared the math and reading scores between economically disadvantaged students, non-economically disadvantaged students, and all students. The third graph shows the mean test scores for all students by grade level, as well as the standard deviation for the mean test scores in each grade level.

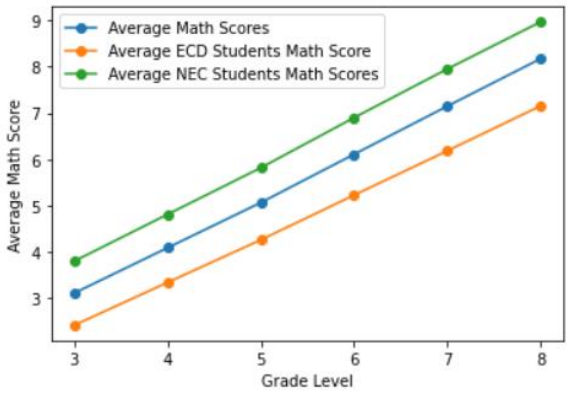

Figure 2: Average Math Scores for NEC, ECD, and all Students

As evident in Figure 2 and Figure 3, there is a larger gap in mean test scores between the economically disadvantaged students represented by the orange line, and non-economically disadvantaged students represented by the green line in both mathematics and reading.

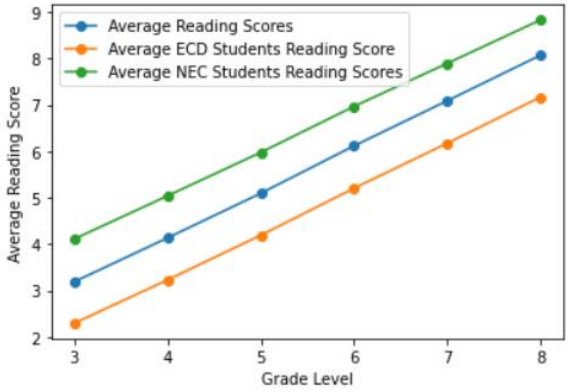

Figure 3: Average Reading Scores for NEC, ECD, and all Students

We then calculated the average gap in mean test scores across grade levels for both reading and mathematics. In mathematics, the average gap in mean test scores for grades three through eight was 1.61162. This represents a gap of over one and a half school years. In reading, the average gap in mean test scores for

grades three through eight was 1.75950. This represents a gap of over one and three quarters of a school year.

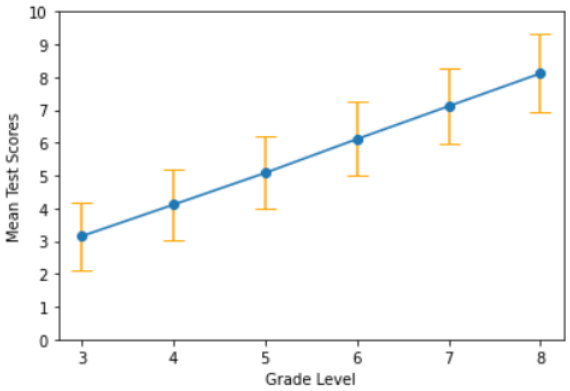

Figure 4: Mean Test Scores for all Students by Grade Level with Standard Deviation Error Bars

The third graph created from the SEDA dataset depicts the mean test scores of students in both reading and mathematics combined, along with the standard deviation as error bars as shown above in Figure 4. There is a standard deviation of a little over 1 grade year with an upward trend in the standard deviation from the lowest standard deviation being 1.04786 in grade three, up to a standard deviation of 1.15078 in grade eight. This upward trend in standard deviation as grade level increases is a common and known trend that occurs due to the divergence of learning rate as students age. Other than that, we can also see that there is a strong one-to-one relationship between the grade level and mean test score which is a function of the dataset as explained in the "Data" section of this paper.

In the second approach, we built a regression model using a deep neural network (DNN) in order calculate the feature importance of each feature in the dataset which we obtained through the one-hot encoding during the preprocessing of the data. Before we look at the model, we must first define some vocabulary relating to our model. We must first understand what exactly a DNN model is? Every neural network consists of at least two layers, an input layer and output layer. Each layer has a different number of nodes depending on the model. In this case, our model has a normalization layer as described in the "Data" section of this paper as the input layer with 13 nodes, one for each feature. Since we want to predict a single value, the score of a student, the output layer consists of a singular node. The difference between a normal neural network (NN) and a DNN, is that the DNN has hidden layers, layers between the input and output layer. In a NN, every node is interconnected by a value called the weight. The weight is a number that the value in the node is multiplied by to generate the number in the next node. The weight starts out random, and as the model is trained, an algorithm called backpropagation, which calculates the gradient of the loss function, will adjust the weights of the neural network based on the learning rate that is given. What exactly is a loss function you may ask? A loss function is an algorithm that calculates the error between the value predicted by the model and the actual value. In this case, we chose to use the mean absolute error (MAE) loss function since it does not punish the model as heavily for outliers when compared to the mean squared error loss function. We want to avoid punishing our model too heavily for outliers since we want to predict as close as possible to the average student for each feature. Additionally, the MAE loss function only takes the absolute value of the loss, whereas the MSE will square the loss. By not preforming a square on the loss, we are more easily able to compare the loss to the standard deviation and other statistical

values that we investigated in the first portion of this study. When we compile our model, we also need to provide an optimizer function. An optimizer function essentially adjusts the learning rate, so the model is more efficient and reduces the loss of the model. For our model, we utilize the Adam optimizer which is considered one of the best optimizers as it combines the advantages of other optimizers such as maintaining a history of updates to the weights as well as an adaptive learning rate that increases when dealing with sparse data and decreases when dealing with dense data. The final piece of our DNN is the activation function that is a parameter of the hidden layer. The activation function controls the range of outputs our layer can produce. In this model, our hidden layer uses the rectified linear unit (ReLU) activation model. What this activation function does is return the value zero for any negative value inputted, and for any positive value inputted, it returns the same exact positive value.

Having a better understanding of our dataset, we can now begin to build our model. We start by using a 0.8 split to separate our data into a training set and testing set. We then pop the value we want to predict on, in this case the score, and store it in a separate variable for both the training and testing set. Finally, we normalize our dataset in order to standardize the scales and ranges of our dataset. Our dataset is now ready to be used for training our DNN model. For this project, we will be using TensorFlow and Keras in order to build our model. We start by building a Keras Sequential Model, with one input layer, one Dense hidden layer, and one Dense output layer. Our Dense hidden layer has 7 neurons, which was chosen due to it being the average of the 13 nodes in the input layer, and the one node in the output layer. Our model uses the Rectified Linear Unit (ReLU) activation function since we want our output, the score of a student, to be a non-negative number. We then compile our model with the Mean Absolute Error (MAE) loss function, which is the best loss function in this case since we want to limit the effect of outliers on our regression model and associate the loss values with the statistical values generated in the first part of the study. We use the Adam optimizer which is widely considered the best optimizer in machine learning, with the standard learning rate of 0.001. Finally, we trained our model using 3 epochs. This was determined by trial and error. After training on 10 epochs a significant decrease in loss reduction was experienced. By training on only 3 epochs, the model can capture nearly all of the datasets information while also avoiding overfitting the model as shown in Figure 5 below.

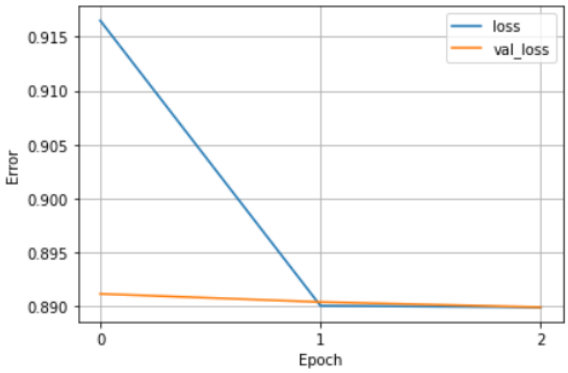

Figure 5: Error (Loss) of our DNN Model after X Epochs

After training our model, we use feature importance analysis to quantitatively calculate the effect each feature has on model's prediction for a student's score. In order to preform feature importance, we choose a single column, which represents a feature, and we permutate that column. We then use the model and the permutated

column along with the other non-permutated columns in the dataset to predict a student's score. The loss compared to the true values is then computed and recorded alongside the name of the permutated column. This is then repeated for each feature column in the dataset and all the results are recorded as shown in Figure 6 below.

Permutated Column	Loss
asian	1.005049705505371
male	1.0069162845611572
subject	1.0070534944534302
year	1.0094729661941528
native american	1.010225772857666
african american	1.0131306648254395
hispanic	1.0136717557907104
female	1.0364086627960205
cohort_pct	1.0512055158615112
ECD	1.0609605312347412
nec	1.22321057319641111
white	1.2750582695007324
grade	2.189927339553833

Figure 6: Table Generated from our Feature Importance Analysis Showing the loss when each Associated Column was Permutated and fed into our DNN Model

The table is ranked by loss. The more loss associated with a permutated feature column means that the feature is more important in determining the student's score. This is because by permutating a column we are essentially putting random values in that column, if it is an important feature for the model when predicting a student's score, then by assigning random values the model will be unable to accurately predict the student's score. From the table, we are able to see that the economic features, ECD and NEC have the fourth and third largest loss values respectively compared to all other features. The interpretation for this is that being non-economically disadvantaged is the third most important factor when determining a student's score, behind only being white, and the grade level. Similarly, being economically disadvantaged is the fourth most important factor in determining a student's score, behind being white, NEC, and grade level.

Limitations

The largest limitation we faced during this study was the datasets that we used, especially for the DNN model. Datasets with various household characteristics and measures of educational attainment are not only exceedingly rare to find, but also extremely expensive to create. In order to even use the SEDA dataset for our DNN model, we had to use one-hot encoding since the rows in the dataset did not actually represent an individual with various features, but rather a whole school and average test scores based on subgroups. This meant that our dataset contained no information about student's with more than one feature, but rather each row was relating to a specific feature. If possible, it would be best to obtain and use a dataset in which each row represents an individual and columns represent a feature of the associated individual. However, due to the time constraints and financial constraints it was impossible to find or obtain such dataset for this project.

Conclusion

Recent studies on the factors that allow students to succeed academically, have led to a greater understanding of the relationship between wealth and the ability of a student to learn and academically preform. Our study suggests that economic status is one of the leading determinants of whether a student will succeed or not, above other factors such as race and gender. The education of students in the present, is the success of our nation in the future.