**Education and Income**

Kayla Thompson

DSC 680 - Fall 2021

<https://github.com/kaymaethomp1>

**Abstract**

In this project I will create a model that will use gender and level of education to predict the level of income a person receives. Pursuing higher levels of education is expensive and time consuming. This project is important to prove the perception in society that more education means more income. I will use a linear regression model and a random forest model as the frameworks for this model and will test each to determine the best model for this problem. This analysis is important because it can serve as a means for potential students to weigh the positives and negatives for pursing higher levels of education. I plan to use Python to create my model and test my model. Prior to building the model data clean up and exploratory data analysis will also be done in Python.

**Problem Statement**

Education is touted as being the path out of poverty and toward financial freedom. For many, this promise of a better future is the reason to take on massive amounts of debt. The cost of education is a hot topic in the United States and many political figures have taken stances on this topic. In schools across the United States, college is advertised as the next logical step for students to take after completing their high school degree. Government agencies such as the Social Security Administration have made claims stating: “Men with bachelor's degrees earn approximately $900,000 more in median lifetime earnings than high school graduates. Women with bachelor's degrees earn $630,000 more. Men with graduate degrees earn $1.5 million more in median lifetime earnings than high school graduates. Women with graduate degrees earn $1.1 million more” [1]. The US Bureau of Labor Statistics has also published documents with similar messaging, an example of which is below. [2]

Chart

Description automatically generated

All the agencies we have come to trust, seem to be pushing the message that if a young adult wants to ensure a lifetime of wealth, higher education levels should be pursued. This claim goes highly uncontested, with this project we will aim to further prove the claim that regardless of age and sex, higher levels of education lead to higher levels of income.

**Methods and Results**

I started this project by completing exploratory data analysis. In this analysis I created a histogram of each variable to get a sense of the spread within each variable. Figure 1 is the resulting histogram of hourly wages. This histogram shows that wages are pretty evenly spread however, there are major spikes just below ten dollars per hour and just below twenty dollars per hour. The average hourly wage in our data set is fifteen dollars and fifty cents. Figure 2 is the histogram for age. Age appears to be normally distributed with the average age in our data set being 37. Figure 3 is the distribution of years of education. This distribution is normal but takes on a slight skew to the right. The average number of years of education for this data is 13. Figure 4 is the histogram for sex, which is a fairly even split between male and female.

Next, using ThinkPlot I created a scatter plot. I put years of education on the x axis and hourly wages on the y axis. What I noticed from this scatterplot is that there doesn’t appear to be much of a relationship between the two variables. However, what is apparent is that there is a spread of income attributed to each grouping of years of education. To finish out the exploratory data analysis, I created a correlation heat map. Figure 6 in the appendix is the heatmap for this dataset. The strongest correlation present is between age and hourly wage, followed by years of education and hourly wage and lastly sex and hourly wage.

Next, I used sklearn to divide my data up into a test set and a train set. I chose to have 25% of my data set aside into the test set and the remaining 75% as the training set. With this step I also named Composite Hourly Wages as the Y that the model should predict. The first model built and tested was a linear regression. I did not eliminate any variables and I used the Linear Regression package from sklearn, Linear Model. The model produced a root mean squared error of 6.47 on the training set and a root mean squared error of 6.49 on the test set. The variance for this linear regression model was 0.0249. Next, I used the Random Forest Regressor package from sklearn to create a random forest model for this data. I also used root mean squared error to evaluate the performance of this model. The root mean squared error of the random forest was 7.07.

**Conclusions**

Based on the exploratory data analysis, the results from the linear regression and random forest models, I conclude that years of education is not a good indicator to determine income level. I conclude this because of the variation seen in income at all levels of education achieved and because there is only a small correlation between the amount of hourly income a person makes and their level of education. Additionally, both models has fairly large root mean square error, indicating a high level of variance in the residuals of the model.

**Appendix**

Figure 1

Chart, histogram

Description automatically generated

Figure 2

Chart, bar chart, histogram

Description automatically generated

Figure 3

Chart, histogram

Description automatically generated

Figure 4

Chart, bar chart

Description automatically generated

Figure 5

Chart, scatter chart

Description automatically generated

Figure 6

A picture containing square

Description automatically generated

**Sources**:

*Average salary by education level: Value of a college degree*. Bachelor's Degree Completion. (2021, September 10). Retrieved October 1, 2021, from https://www.northeastern.edu/bachelors-completion/news/average-salary-by-education-level/.

Coe - annual earnings by educational attainment. (n.d.). Retrieved October 1, 2021, from https://nces.ed.gov/programs/coe/indicator/cba.

*Education and income: How learning leads to better salaries*. The Learning Agency. (2021, April 13). Retrieved October 1, 2021, from https://www.the-learning-agency.com/insights/education-and-income-how-learning-leads-to-better-salaries/.

Porter, E. (2014, September 10). *A simple equation: More education = more income*. The New York Times. Retrieved October 1, 2021, from https://www.nytimes.com/2014/09/11/business/economy/a-simple-equation-more-education-more-income.html.

[1] *Social Security Administration*. Research Summary: Education and Lifetime Earnings. (n.d.). Retrieved October 1, 2021, from https://www.ssa.gov/policy/docs/research-summaries/education-earnings.html.

Strauss, S. (2017, December 7). *The connection between education, income inequality, and unemployment*. HuffPost. Retrieved October 1, 2021, from https://www.huffpost.com/entry/the-connection-between-ed\_b\_1066401.

*Strong link between education and earnings*. World Bank Blogs. (n.d.). Retrieved October 1, 2021, from https://blogs.worldbank.org/education/strong-link-between-education-and-earnings.

[2] U.S. Bureau of Labor Statistics. (2021, April 21). *Education pays*. U.S. Bureau of Labor Statistics. Retrieved October 1, 2021, from https://www.bls.gov/emp/chart-unemployment-earnings-education.htm.

U.S. Bureau of Labor Statistics. (n.d.). *Measuring the value of education : Career outlook*. U.S. Bureau of Labor Statistics. Retrieved October 1, 2021, from https://www.bls.gov/careeroutlook/2018/data-on-display/education-pays.htm.

Wolla, S. A., & Sullivan, J. (n.d.). *Education, income, and wealth*. Economic Research - Federal Reserve Bank of St. Louis. Retrieved October 1, 2021, from https://research.stlouisfed.org/publications/page1-econ/2017/01/03/education-income-and-wealth/.