Exploring Predictive Models for Income

Nicholas Liagridonis, Sophie Liu, Lawson Vitosh
DATA 11900 Midterm Project

I. Overview & Motivation

Our group was curious about how we can predict income, and to what extent a variety of factors affect income. For our project, we were interested in investigating how these factors individually and cooperatively could predict income. Some factors we planned to study included: Educational Attainment, Citizenship Status, Age, Sex, among others. We planned to use the US Census Bureau Current Population Survey's March 2022 Annual Social and Economic (ASEC) Supplement[1] to create a collection of linear regression models, histograms, and other visualizations to demonstrate our findings. With the aid of these visualizations, we intended on finding out more about the factors that inform income, and see how accurately we could predict the income of each person using specific factors.

II. Related Work

One piece of media that one of our group members had found and talked about was a Forbes article that had come out the previous summer about the average salaries of 2023 College Graduates.[2] It predicted average salaries by major for 8 different categories. The article included the average salaries in different fields based on other variables such as race and gender, but we were curious what income levels we would be able to predict if we included college majors, but also tried to include a variety of other factors, to try to accurately predict future income, rather than just calculate the current average incomes.

[1] https://www.census.gov/data/datasets/time-series/demo/cps/cps-asec.html
[2] https://www.forbes.com/advisor/student-loans/average-salary-college-graduates/

III. Initial Questions

At first, we wanted to try to predict the future income of college students. As college students, we were quite curious about what factors affect future income, including college major, background, sex, etc. However, we found difficulty making a predictive model due to the nature of the data we had at our disposal. Our data was very static in nature; we were looking at Census data that captured students at a particular point in time, but couldn't see how they progressed throughout time. Trying to predict future income for college students today based solely on adults' attributes when they were in college was especially difficult to quantify properly, given that quite a bit has changed since many adults were in college. There would be differences in how certain variables affected future income even between decades, so to create a predictive model in that fashion did not prove feasible.

Following our same line of interest, we opted to attempt to create a predictive model for current income. Our plan was to continue to adjust a regression model with factors that we believed would be relevant and see how well they predicted the income of the individuals represented in the dataset. We would continue to fine-tune this model until we could get relatively close to predicting the income of a person based on other factors, or at the very least, break income into a series of brackets and be able to accurately predict an individual's income bracket.

In addition to regression analysis, we also explored kNN classification to try to predict what income brackets people would fall into based on the characteristics they possessed.

IV. Data Cleaning & Exploratory Analysis

To start, we first attempted some data cleaning, only to find that this was in fact a very clean dataset. We tried to drop all rows with Null values, to find that there were in fact no Null values in the dataset. We also checked to see if there were any repeated ID numbers to see if any people were repeated in the dataset, but found none.

Next, we created a new dataframe, a subset of the original dataframe, including only the columns that we thought would be most useful to study. We were concerned that with 832 columns and 152,732 rows, we would have issues with long run times on some of our code. We looked through the data dictionary provided by the Census Bureau and chose 31 columns to start. We then explored a variety of characteristics within those 31 columns to see what data we had to work with. We had a mix of categorical and numerical data to work with.

Plotting the distributions of ages, hours worked per week, and total income brackets, we realized that there were a lot of people in the dataset that fell into the edge categories (there were over 21,000 people earning under $2,500 annually, most of whom were under the age of 20, as well as a lot of people who earned over $100K). To reduce error in our model, we first decided to filter out current students by their enrollment status flag. However, after doing this we were still left with a lot of low earners. We decided that a better way to filter our data would be to only include people who are full-time workers in the labor force. This reduced the number of people in the lowest income bracket significantly, and we were left with a dataframe of only full time workers that had around 80,000 rows.

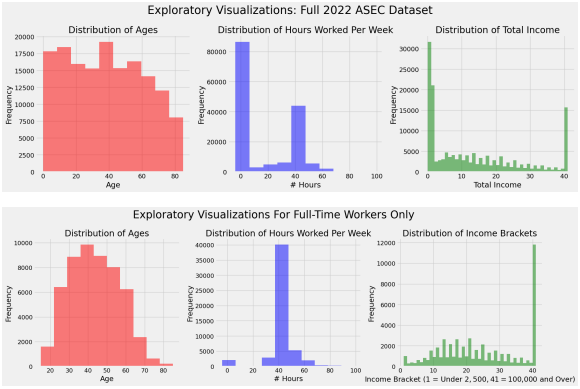

Figure 1. Above: Exploratory visualizations on full ASEC dataset. Below: exploratory visualizations on subset of full-time workers.

After we were satisfied with our data, we were then able to dummy variables for all of our categorical variables, so we could regress on that data. After this, we were able to run our first regression model.

V. Regression Analysis

After selecting our features and filtering the dataset to just full-time workers, we started to run linear regression to predict incomes.

Running the linear regression on the full time workers, we were able to predict incomes, but they had a high error. There were two sources of this error: extremely low earners and high earners. The low earners skewed the average percent error because the survey had many people listed as earning $0 or $1, even though they were supposed to be working full time. These low numbers also created huge percent errors. The model was also not able to predict high incomes well because there are fewer high earners to train the model on, and the difference between people in the high income groups is much larger than in the middle ranges. We reran the regression without training on people with incomes above $100,000. This was motivated by two reasons. First, people with incomes above $100,000 seem to be distinct from those earning under that number. Sorting by incomes, incomes

above $100,000 grow much more rapidly than those under $100,000, showing that there might be two distinct groups: over and under $100,000. Second, the survey we based the project off used $100,000+ as the overflow group for incomes, meaning that they thought there was a difference between those with incomes over and under $100,000. After removing people with incomes over $100,000, we reran the regressions and then filtered out the predictions that were under $1,000. The model had high errors for predicted incomes that low because it is supposed to be predicting full time incomes, and full time incomes cannot legally be that low. Our final model can be used to predict incomes for the middle-range, full-time workers in America.

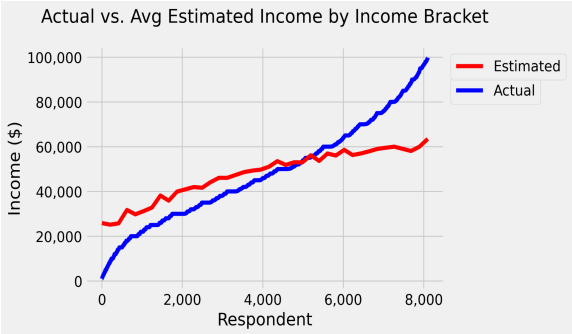

Figure 2. Actual vs. average estimated income by income bracket from multiple linear regression.

Figure 2 shows the actual and the estimated average incomes in an income bracket. The model does well predicting in the mid-range, but struggles on the high and low ends. The average error is about $15,000 and the average percent error is 54%.

However, the model is good at predicting average income for a group in a category. For example, Figure 3 shows that the model is accurate when predicting average income for each age.

Figure 4 is an example of the predictions made for a categorical variable. Again, it shows that the model is capable of predicting average income for a group of people.

Figure 3. Actual vs. predicted income by age, from linear regression model

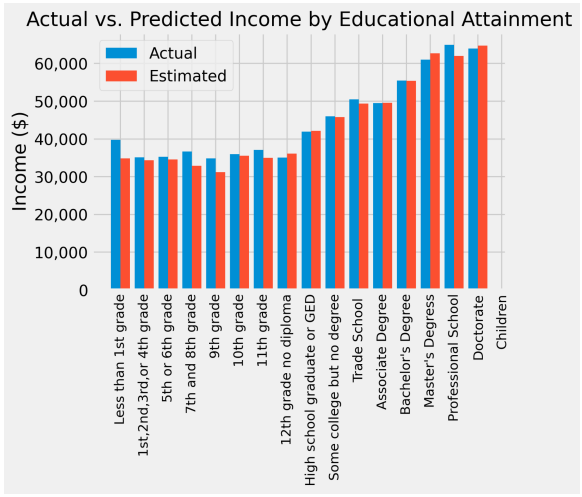

Figure 4. Actual vs. predicted income by educational attainment category, from linear regression model

As Figure 3 shows, the model is overfit. We attempted to correct this by using Lasso and Ridge regressions to identify the most useful variables for the prediction. However, after using cross-validation to find the tuning parameter, the models with the lowest error were those that kept all the variables.

VI. kNN Classification

In addition to multiple linear regression, we also attempted to create a predictive classification model using the k-nearest-neighbors algorithm, to see if we could accurately predict what income bracket an individual falls into. Similar to the final regression analysis, we performed

the kNN classification on the subset of only full time workers.

Since there were 41 income categories with minute differences in income, we decided to regroup the income brackets into 10 categories instead. On an initial run with k=3, the model had an accuracy of around 27.3%. Looking more closely, we saw that it had an average absolute error of 2.085 categories, with a mix of over- and under-estimates (the mean error was -0.223). A similar level of error persisted with k=5, but for k=7 absolute error improved to 1.889 categories off, and absolute error for k=9 improved to 1.876 categories off.

We then performed kNN on the subset of workers with total incomes above $1,000 and under $100,000, as we did with the linear regression model. The error was better for all iterations of k that we tested: for k=3, the predictions were off by 1.968 categories, 1.876 categories for k=5, 1.788 categories for k=7, and 1.759 categories for k=9.

VII. Next Steps

In order to improve the accuracy of our models, we would need to reselect variables and incorporate new information. First, within the survey data, we can find different variables that help with the prediction. However, the most important step in our new feature selection would be to use less of them. Our regression model is overfit, meaning we might be able to get better out-of-sample predictions by only using the most useful variables.

Second, we need more information about both high and low income workers, since that is where our model has the most error, especially for people with incomes above $100,000. Our model is for predicting average earners, so finding new information about low and high earners would increase our accuracy for these groups. In particular, it would be helpful if we had more information about college majors, roles within a company, and other income outside of salaries. The data contained overall categories for these, but just about what level of degree they had or what industry they worked in. Looking inside these broad categories would help predict incomes.

Additionally, information on how people with high incomes make money would be useful. We had the most difficulty fitting our predictive model to the highest earners, so optimizing that area of weakness would significantly improve our model. On top of having jobs with higher and faster growing salaries, people with high income also earn money outside of their jobs, such as with investments. Learning about how high earners make money would be helpful.

VIII. Group Member Contributions

Nic headed data collection and cleaning, in addition to variable selection and research outside of the main survey that was used. Sophie performed the kNN predictions, and Lawson worked on the regression analysis. All members collaborated on the report.

IX. Bibliography

1. US Census Bureau (2022, September 8). *Annual Social and economic supplements*. Census.gov. Retrieved February 24, 2023, from https://www.census.gov/data/datasets/time-series/demo/cps/cps-asec.html

2. McGurran, B. (2022, July 28). *Average salaries of college graduates 2023*. Forbes. Retrieved February 24, 2023, from https://www.forbes.com/advisor/student-loans/average-salary-college-graduates/