Ridge and Random Forest Models To Predict an Individual's Wage Income and Gender

Introduction

The gender pay gap has been a longstanding issue in the United States. In 2022, women earned roughly 82% of what men earned. This is similar to 2002, when women made roughly 80% of what men earned, showing that the gender pay gap has persisted at a similar rate over the past 20 years **(Pew Research Center 2023)**. This machine learning model is motivated by this fact, and looks to see whether the gender of an individual plays a significant role in predicting an individual's wage income.

This project utilizes the 2021 American Community Survey, an annual demographic survey conducted by the U.S. Census Bureau. This survey provided numerous characteristics of individuals in the U.S., including but not limited to the gender characteristics that were incorporated into the machine learning models in the project.

The machine learning models I chose to use for this project were the SciKit Learn's Random Forest Classifier and Random Forest Regressor, and Scikit Learn's Ridge Regression. The results of these models showed that while gender ranks highly on the feature importance measure in the Random Forest Regressor model and is above 50% accuracy for the Random Forest Classifier model, it does not pair well with the Random Forest model to create a successful model for predicting wage income.

Data

Data on the 2021 American Community Survey was obtained from the Integrated Public Use Microdata Series (IPUMS), the world's largest individual-level population database. I selected a 0.1% sample size of the 2021 American Community Survey dataset and included the individual characteristics of age, gender, race, cognitive difficulty status, ambulatory difficulty status, citizenship status, English proficiency, area of residence, number of children, recency of childbirth, and wage income.

While the focus of this project was the effect of gender on wage income, I selected a variety of variables that either had potential to affect an individual's wage income, or had been thoroughly researched in the past and proven to have correlation to wage income, to measure against the effects of gender.

First, thorough processing of the dataset was necessary to make it usable for analysis. In the IPUMS database format, the values for each variable either represented a numerical value, or a numerical value that represented a categorical value that was noted in the dataset's description upon download. For categorical variables, I transformed all numerical values into categorical descriptors in order to accurately represent their meaning. Certain numerical values also represented instances of illegibility, inapplicability to the individual, or missing answers, and had to be removed from the dataset. Particularly for the wage income variable, although the variable was numerical, the 999999 and 999998 values did not actually indicate the wage income earned by an individual, but instead represented missing values and inapplicability for an individual, and had to be removed. The wage income variable was also chosen over the total income variable because the total income variable included negative values, and when running a logistic regression model, values must be greater than zero for a logarithm to be taken. Additionally, a very small non-zero value (0.000001) had to be added to all wage income values because this variable included values of 0, and logarithms of 0 cannot be taken as well.

Only individuals aged 16-49 were included in our analysis based on findings presented below, revealing that the age of peak income in the 2021 American Community Survey occurred at 49 years old. Beyond this age, income demonstrated a consistent decline. Individuals under the age of 16 were not included because minors under the age of 16 can only work limited hours, and would naturally have wage income limits because of this.

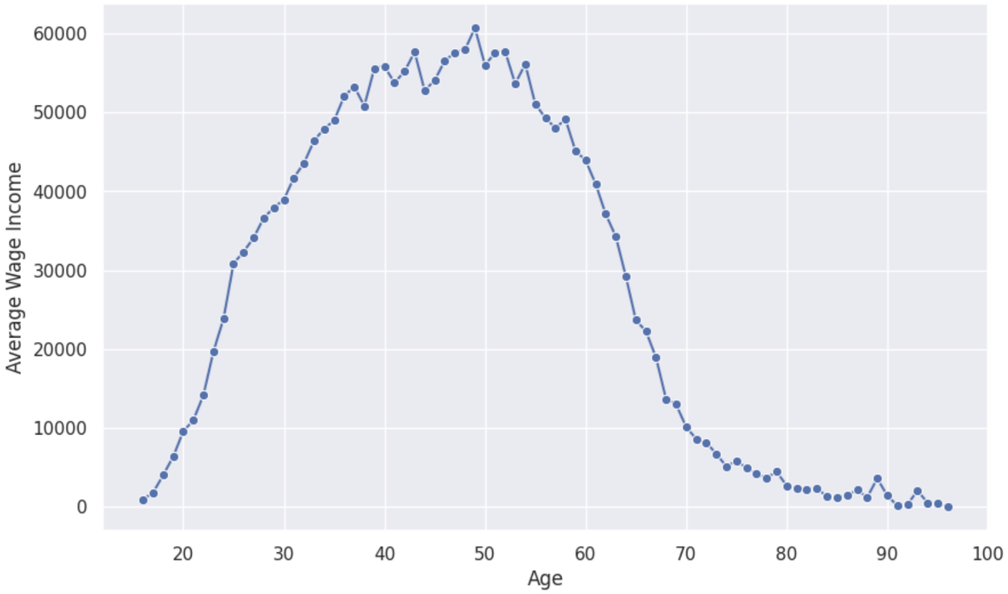

Figure 1: Average Wage Income by Age

I first examine the wage income distribution by gender, as this is the primary variable being studied in the project. Figure 2 shows that there is a clear difference in average wage income for men and women. I also acknowledge that gender is not binary, however, for this project, gender is defined only as man and woman because of the constraints of the data. In the American Community Survey, gender is not differentiated from sex, and therefore only two options are available for an individual to choose when answering the survey - male or female. Therefore, for the purposes of this project, gender is defined as man or woman.

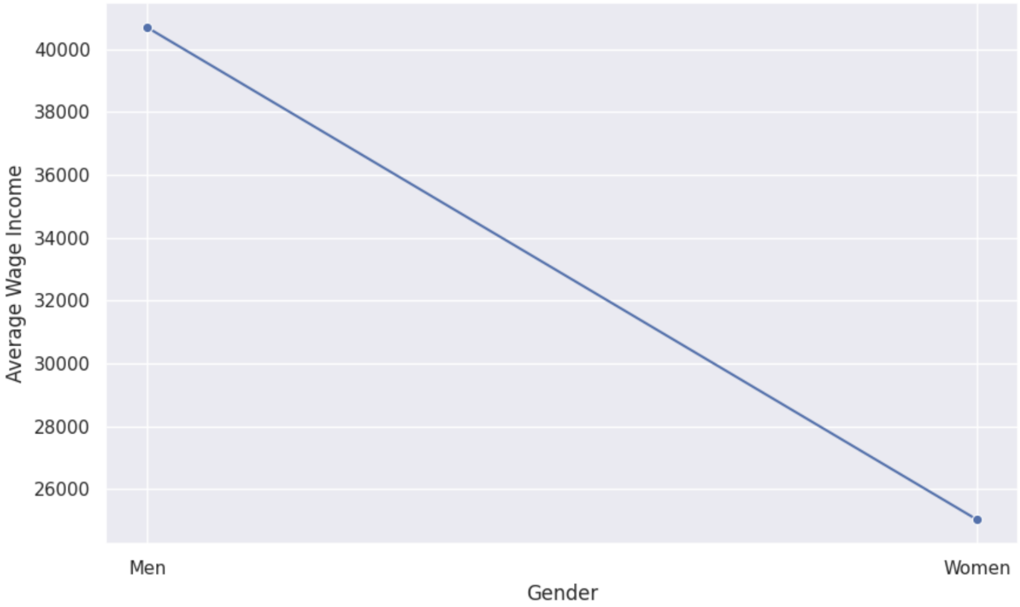

Figure 2: Average Wage Income by Gender

I also examined the distribution of educational attainment encompassed high school completion, undergraduate degree completion, and advanced degree completion. The figure below illustrates a marked increase in wage income as years of schooling, and therefore completion of more advanced degrees, increases. This is important to note because while the main focus of this project was to examine the effects of gender on wage income, education has been widely studied as a mechanism of increased wages. The U.S. Bureau of Labor Statistics reported in 2020 that median usual weekly earnings increase and rates of unemployment decrease as level of degree attainment increases **(U.S. Bureau of Labor Statistics)**.

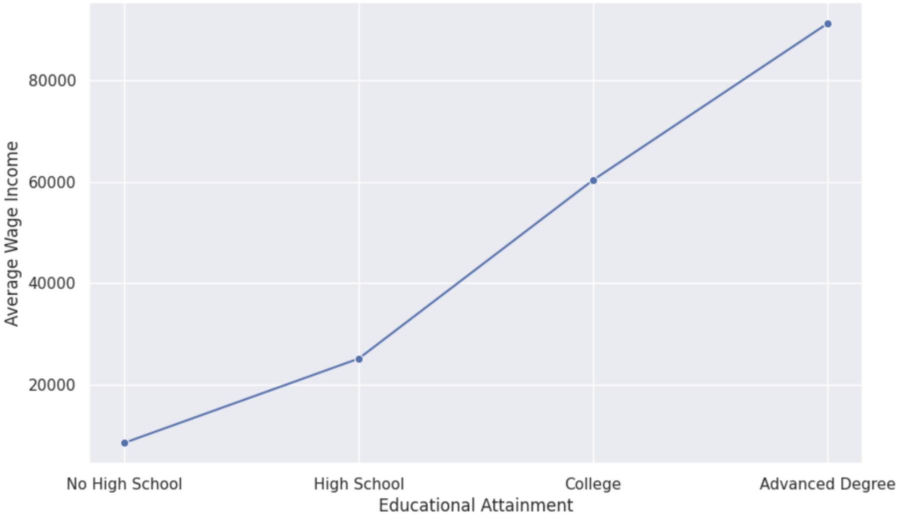

Figure 3: Average Wage Income by Educational Attainment

Although wage income by state naturally varies based on type of work prevalent in the state, and individual state economies, it is interesting to note that Washington, D.C., denoted by State FIP code 11 in Figure 3, has a much higher average wage income compared to all other states, each of which are denoted by a specific state FIP code. It is also clear that average wage income can vary greatly by state, with variations of over $20,000 excluding Washington, D.C., and variations of over $40,000, including Washington, D.C. Given the wide variation in average wage income by state, I chose to include state as one of the predictors of the project's models.

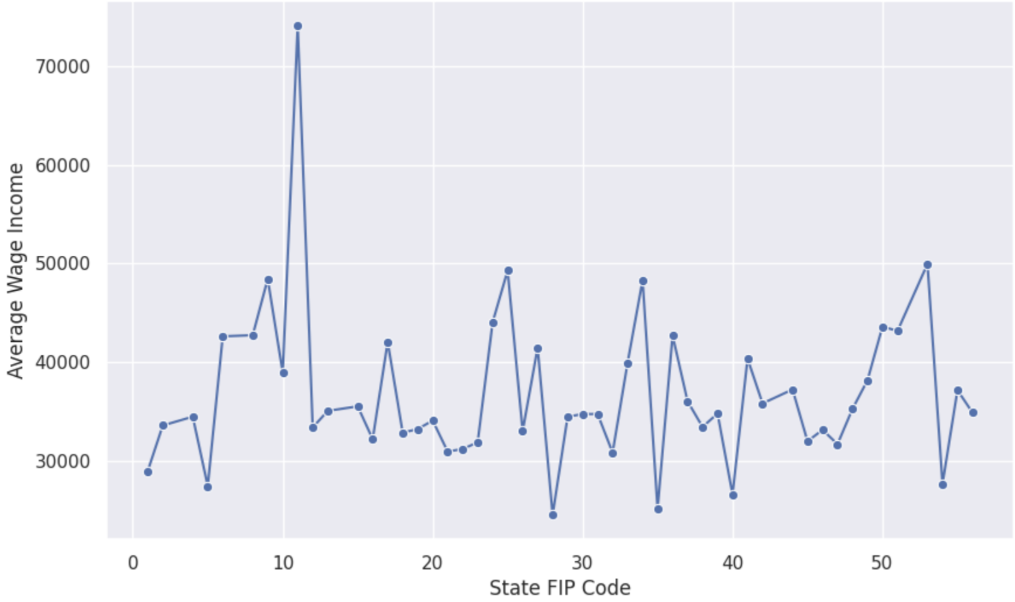

Figure 4: Average Wage Income by State

Modeling

After cleaning the dataset, I chose to run three models.

The first model I ran was a Ridge Regression, using wage income as the target variable, and Here are some more details about the machine learning approach, and why this was deemed appropriate for the dataset.

The model might involve optimizing some quantity. You can include snippets of code if it is helpful to explain things.

This is how the method was developed.

Results

Figure X shows… [description of Figure X].

Discussion

From Figure X, one can see that… [interpretation of Figure X].

Conclusion

Here is a brief summary. From this work, the following conclusions can be made:

first conclusion
second conclusion
Here is how this work could be developed further in a future project.

References