ISM 370
Jalyn Buthman

Business Analytics Project 3

The census income dataset provides data on individuals, including if they make above $50,000 per year, or equal to/below $50,000 per year. Within the dataset the majority of people make $50,000 or less annually(Figure 1). In the following analysis I will explore which demographic information can be used to predict if an individual makes more than $50,000 per year.

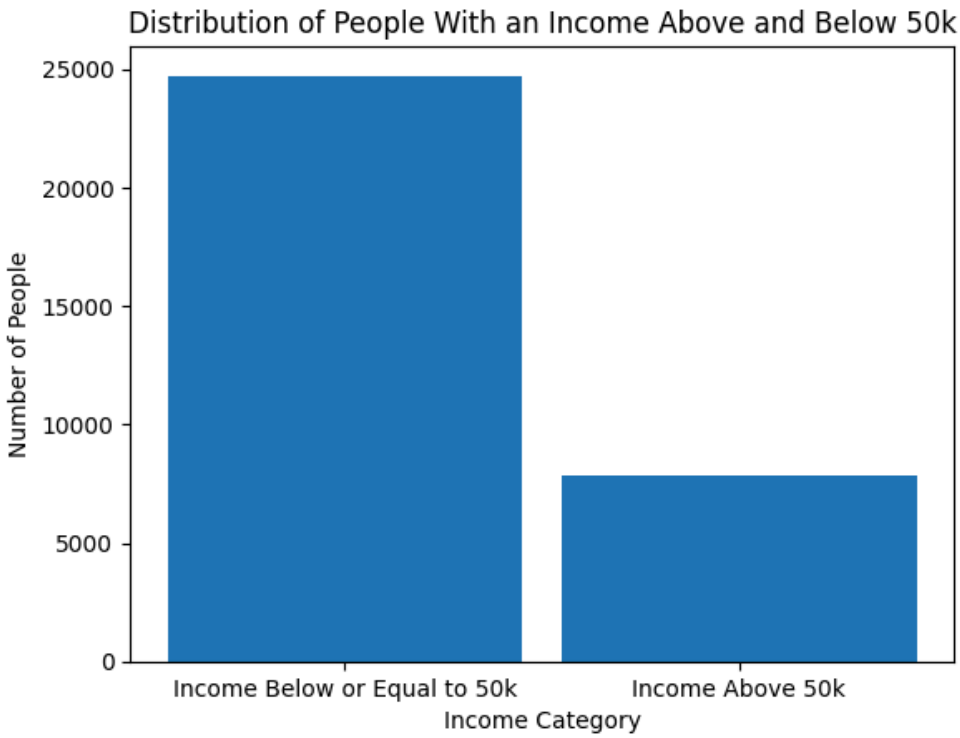

Figure 1

Throughout my analysis I explored the impact of age, education level,hours worked per week, and sex to determine how they could be used to predict individual income. For age, the distribution of individuals making more than $50,000 annually is roughly normally distributed, with a slight right skew. People with an income above $50,000 peak around the age of 40, with a sharp decline in the middle of the peak (Figure 2).

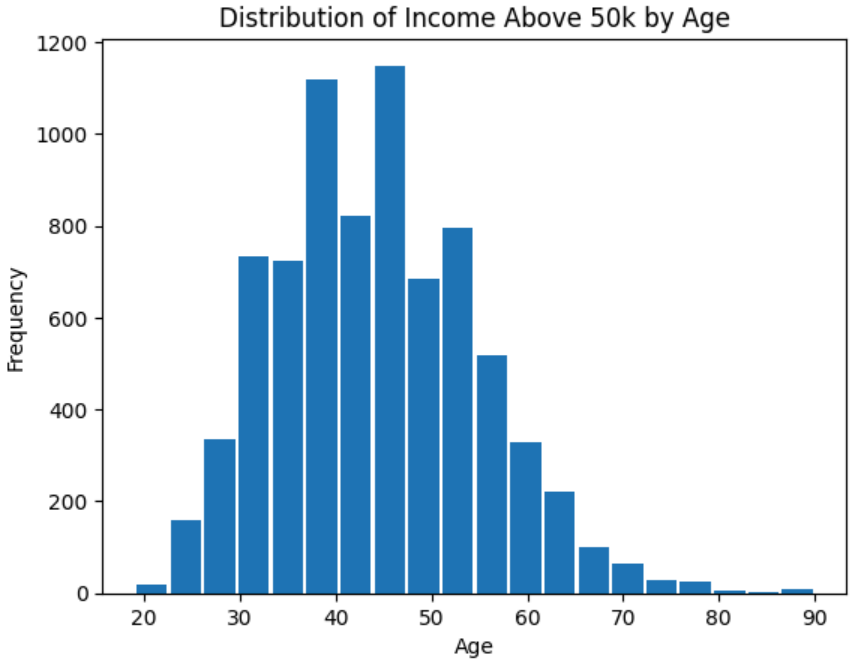

Figure 2

The distribution of individuals making more than $50,000 per year is much more scattered when sorted by education level, with a left skew of the data. There are however two peaks around education levels 9 and 13 (Figure 3).

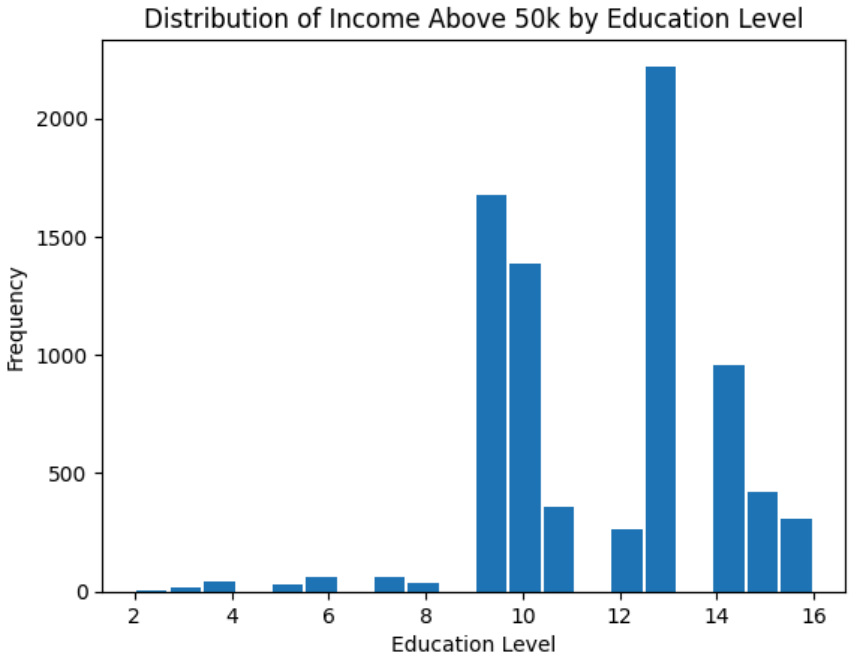

Figure 3

For the metric hours worked per week, the distribution of people who make more than $50,000 annually has a sharp peak right before 40 hours per week. There are then two lower peaks around 50 and 60 hours per week respectively.

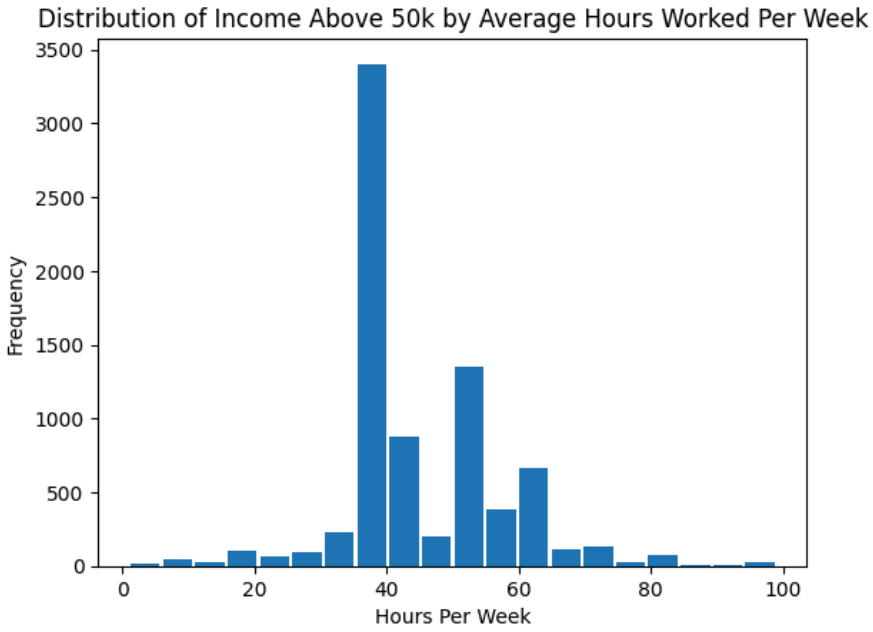

Figure 4

When comparing people with an income above $50,000 per year by sex, we can see that there are far more men than women in this category (Figure 5). Since there is such a large gap, sex may be a good predictor of an individual making more than $50,000 annually.

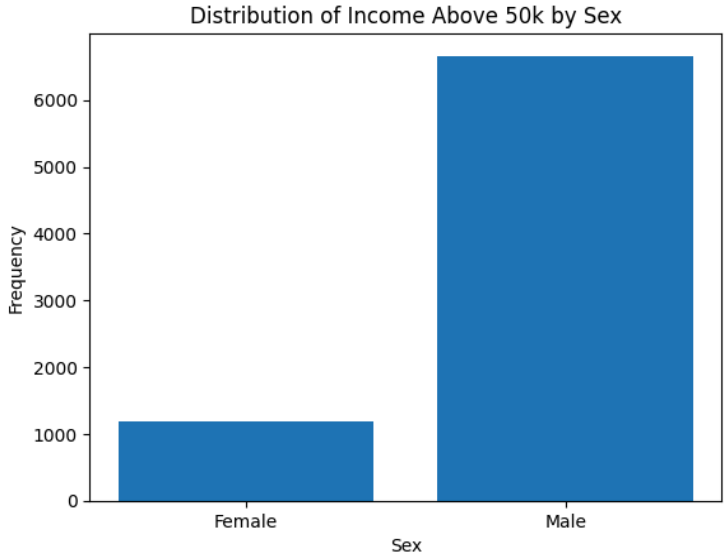

Figure 5

I used the K Nearest Neighbor, Decision Tree, Random Forest, and Neural Network supervised learning models to further analyze how age, education level, hours worked per week, and sex could be used to predict if an individual makes more than $50,000 per year. Of these models, the Random Forest and Neural Network consistently receive a tie for the lowest error score, making them the best models for predicting an individual's income category (above or equal to/below $50,000 annually) from this dataset. These models' error prediction score is 0.13 $\pm$ 0.01 for the test variables used for this report.

The first prediction was made using the test variables listed below:
Age: 35
Education Level: 10
Hours Worked Per Week: 30
Sex: Female

The second prediction was made using the test variables listed below:
Age: 47
Education Level: 12
Hours Worked Per Week: 40
Sex: Male

The first set of test variables resulted in a prediction that the individual would make $50,000 or less annually from both the Random Forest and Neural Network models. For the second set of test variables the Random Forest model predicted that the individual would make $50,000 or less, while the Neural Network model predicted that the individual would make more than $50,000 annually. This means that one of the models' predictions is wrong for the given set of test variables.

The low error prediction scores from the Random Forest and Neural Network models suggests that they can be useful in predicting whether or not an individual makes more than $50,000 per year. Although their error rates are low, we must keep in mind that these models are not perfect at their predictions. We can even see this with the test variables for the second prediction above, where the two models had opposite predictions. While the Random Forest and Neural Network models can be used to make predictions of an individual's income category, the resulting predictions should be used mindfully.