Alexander BeVier
Dr. Korede Ajogbeje
QMSS 301 003
3 November 2023
Project 2: Predictive Analysis - Report

Predictive Analysis: Using Census Data to Predict the Likelihood of "High Income"

METHODS

The script used is split into four sections: set up, StatModel, SKLearn, and visualizations. All code was run in Google Colab using Python. Additional libraries used include Pandas, Numpy, Matplotlib, SKLearn, Seaborn, and StatModel. The current data set used is census data comprised of demographic and occupational information designed to predict whether annual income exceeds $50,000 per year, or for the purposes of this analysis, "high-income". The dataset was sourced from the University of California-Irvine Machine Learning Repository.

Set Up. Data was first loaded in and processed using Pandas. The variable 'hispanic' (boolean) was created by using the 'native_country' variable and filtered all unique values (countries) that are of Hispanic origin (according to the 2023 World Population Review) into a list of 'Hispanic countries'. For any given individual, if the value of 'native_country' is in the list of Hispanic countries, the value of 'hispanic' returned True. Anything else, the value returned False. 'Income_class' was converted from string to binary.

Education (education_num), Race, Sex, and Hispanic were the variables selected to interpret income for this study. Education was chosen because it often has a strong correlation with income. Other variables were chosen to investigate whether there are gender, racial, and/or ethnic differences in income rates. Variables in question were filtered out from the loaded data frame and stored in a separate data frame.

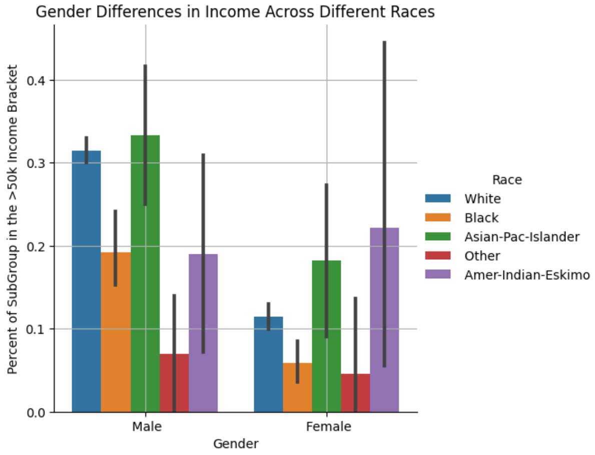
Using StatModel. Dependent variable data used the newly created data frame. The Independent variable was assigned as the income class variable. The 'train_test_split' model from the SKLearn library was used to prepare the data. The ratio of the training data size to the testing data size was 80:20, meaning that 80% of the data was used for training and 20% of the data was used for testing. The random state was set to 16.

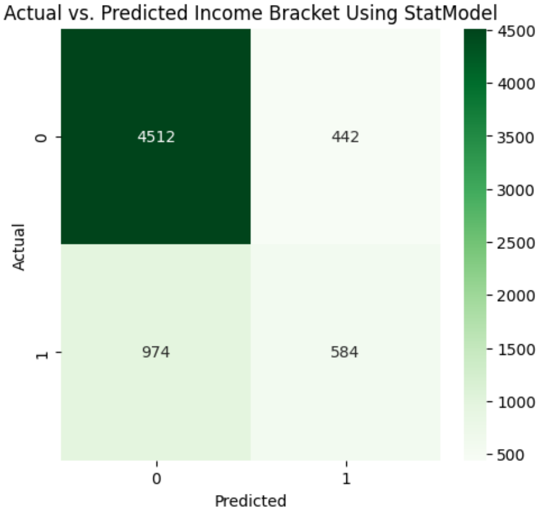
Logistic regression was performed using the ‘logit’ function from SKLearn. The logistic regression produced statistical information on all of the selected variables, omitting the value in each variable with the lowest coefficient and treating it as a baseline value. Following the logistic regression, predicted probabilities were created of the likelihood of each observation being in the high-income bracket. Predictions were then rounded and converted to binary values and checked for accuracy.

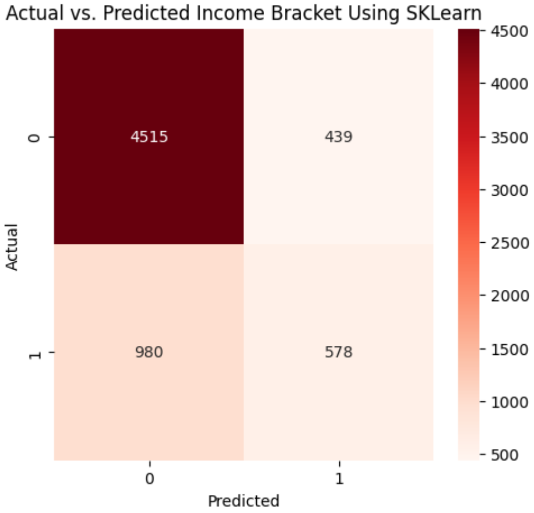
Using SKLearn. Seeing as logistic regression through SKLearn won’t automatically convert categorical/string variables into numeric/float/binary variables as StatModel does, additional cleaning was necessary. ‘Sex’ was converted into a binary variable, and ‘Race’ was converted to a numeric/integer variable. Values in ‘Race’ were ordered by ascending order of the coefficients of the logistic regression using StatModel. In other words, unique values (races) that had lower coefficients were given lower index positions in the reordering of the variable, and vice versa.

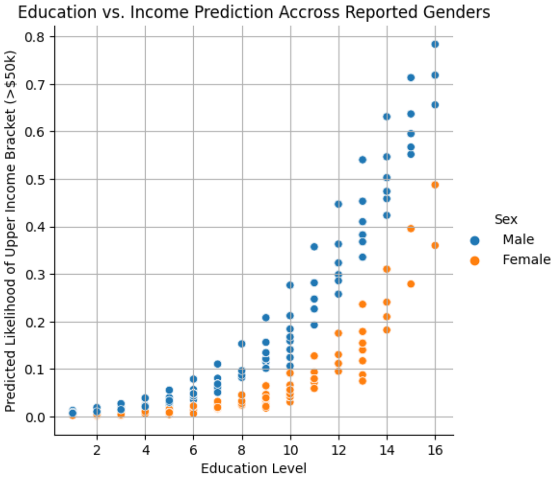
Regarding the logistic regression, all parameters of regression are the same as that of StatModel. Coefficients and intercepts are created using the LogisticRegression() function from SKLearn. It should be noted that due to the cleaning of ‘Sex’ and ‘Race’ variables, only four coefficients were created for this regression, compared to the seven coefficients created during the previous regression. Training accuracy is reported again as before, yet this time accuracy score, precision score, recall score, and f1 score are reported via classification report.

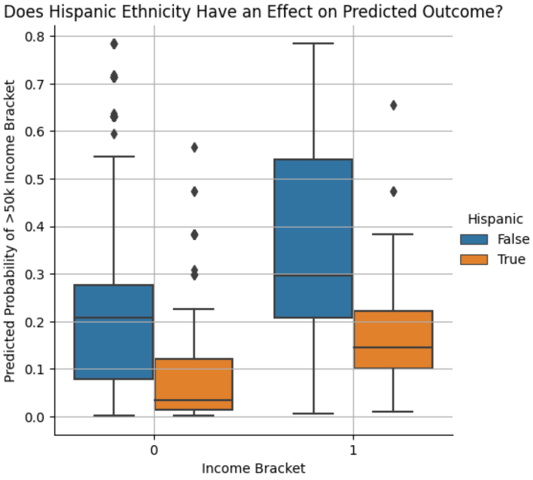
Visualizations. Four different plots were created, all using Seaborn. First, a bar chart was created to picture the data before the regression. The bar chart shows the percentage of a group who are in the high-income bracket, grouped by both race and gender (under a 95% confidence interval). After the regression, a side-by-side heatmap is produced using the heatmaps of the two confusion matrices created by the accuracies of the two regressions. Next, a relationship plot (or relplot) is constructed, comparing how education level correlates with predicted income bracket across gender. A relationship plot is very similar to a scatter plot but provides specific emphasis on visualizing how variables relate to each other. Finally, a box plot is created to compare how people who are and are not Hispanic are predicted to be in the high-income bracket, compared with the current income bracket they were in when the data was collected.


Gender Differences in Income Across Different Races


Actual vs. Predicted Income Bracket Using StatModel


Actual vs. Predicted Income Bracket Using SKLearn


Education vs. Income Prediction Accross Reported Genders


Does Hispanic Ethnicity Have an Effect on Predicted Outcome?

RESULTS

StatModel. Results of the logistic regression show an intercept of -6.8985. This means that holding all other variables at null, or zero, the likelihood of getting "High Income" is -690%. The coefficient of Hispanic[True] is -0.6413, of Sex[Male] is 1.3365, and of Education is 0.3752. If you are Hispanic, the probability of getting "High Income" decreases by 0.64. If you identify as male, the probability increases by 137%, and as education level increases by 1 unit (or 1 grade), the probability increases by 0.37. Coefficients for Race[Asian-Pac-Islander, Black, Other, White] are 0.4958, 0.3193, 0.1166, and 0.8456, respectively. This means that the probability of "High Income" is 0.5 for Asian or Pacific Islanders, 0.32 for Blacks, 0.12 for others, and 0.85 for Whites. Prediction accuracy is 0.782.

SKLearn. The coefficients of the logistic regression using SKLearn are 0.37415324, -0.62584343, -1.33273, and 0.19213024. This means that the likelihood of getting "High Income" is 0.37 with each increasing unit of education, -0.63 if Hispanic, -1.33 if female, and 0.19 with each increasing unit of race most likely to be in the "High Income" bracket (based on coefficients of the previous regression). Training accuracy is 0.782.

DISCUSSION

Using education level, sex, Hispanic identification, and race, the model described above is 78.2% correct in predicting the chances that an individual will be placed in the >$50k, or "high income" annual income bracket. However, the results of the side-by-side heatmap seem to suggest a higher accuracy rate. The model identified that being white is the factor most influential in predicting placement in the high-income bracket. While the plot above shows that Asians or Pacific Islanders have the highest percentage of individuals in the high-income bracket, the sample size of those who identify as white (n = 27,815) is much higher than that of Asians and Pacific Islanders (n = 1,039), thus suggesting evidence supporting the regression data. The box plot shows that those who identify as Hispanic are less likely to be in the high-income bracket than those who don't identify as Hispanic. The relationship plot shows that females are less likely to be in the high-income bracket than males. Data from both box and relationship plots are in alignment with the regression data. Trends identified from the relationship plot additionally show a positive correlation between education level and likelihood of being in the high-income bracket, which is reflected in the education coefficients from the regression data.

CONCLUSION

In the present analysis, data on education level, sex, gender, and ethnic background were pulled from a census dataset to predict income bracket. Logistic regression was used as a means to perform predictive analysis, along with many other analytical tools and libraries in Python to provide insightful visualizations. Results of the analysis showed that white males with high education and no Hispanic background had the highest probability of being in a high-income bracket. Future analysis might use alternative information such as age, occupation, or marital status to predict income bracket.

EXTERNAL RESOURCES USED (REFERENCES)

https://archive.ics.uci.edu/dataset/20/census+income
https://worldpopulationreview.com/country-rankings/hispanic-countries