MODEL PERFORMANCE ANALYSIS

Kirsten Childs

01/15/2022

MODEL PERFORMANCE ANALYSIS

1. OVERVIEW

RTI is trying to predict whether individuals make over $50,000 a year using US Census data. After exploring the data, all continuous variables were grouped into categories, a logistic regression model was created, and odds ratios were calculated for all variables. The odds ratios revealed that individuals with at least a bachelor's degree are between 3 and 8 times more likely to have an income over $50,000 a year than individuals with only a high school diploma.

2. METHODOLOGY AND ANALYSIS

2.1 DATA USED

The data contain 19 attributes for 48,842 individuals from the US Census. For a first look at the variables, statistical tests were utilized to assess which variables were significant indicators for whether someone has an income over $50,000 a year. Next, bar graphs were used to visualize all categorical variables and box plots to visualize all continuous variables.

2.2 LINEARITY OF CONTINUOUS VARIABLES

Each continuous variable was tested to determine if it met the linearity assumption for logistic regression. All continuous variables failed the test and were converted into categorical variables using the previously created visualizations. For capital gains and capital losses, if the individual did not have any, they were grouped into a 'No' category and a 'Yes' category if they had gains or losses. The number of hours worked per week was grouped into less than 40 hours a week, 40 hours a week, and greater than 40 hours a week. And finally, the ages of individuals were grouped according to previous census data standards into less than 24 years of age, 25 - 44 years of age, 45 - 64 years of age, and 65 + years of age.

2.3 DATA CLEANING AND MODEL SELECTION

The last modifications to the dataset were to address limited sample sizes. Individuals with only a pre-school education were grouped with individuals who only had a 1^{st}-4^{th} grade education. Individuals who had never worked were grouped into the missing work status category. And individuals from Honduras, Laos, and Holland-Netherlands were grouped into the missing country status category.

Model selection began by using a logistic regression model with all the variables present and then a backward selection method was used to eliminate insignificant variables. This process produced two models, which both predict the probability of whether an individual has an annual income over $50,000 a year. The first model consisted of eight predictors: age, working-class, education level, marital status, race, hours/week working, capital gains, and capital losses. The second model had identical predictors but replaced working-class with occupation. Both variables offered similar information to the model.

3. RESULTS

The models were tested on the unseen *validation* data to see how well they predicted whether individuals make over $50,000 a year. The first model produced a concordance value of 89%, and the second model, 90%. Therefore, from the second model, the now final model consisted of eight predictors: age, occupation, education level, marital status, race, hours/week working, capital gains, and capital losses. This model can predict if an individual's income is over $50,000 with 89% accuracy on the *test* dataset.

3.1 ODDS RATIOS

The odds ratios were calculated for each variable to determine the strength of association between the variable and making over $50,000 a year. Figure 1 displays an individual's education level by descending education level with associated odds ratios. The odds are all compared to an individual who is a high school graduate.

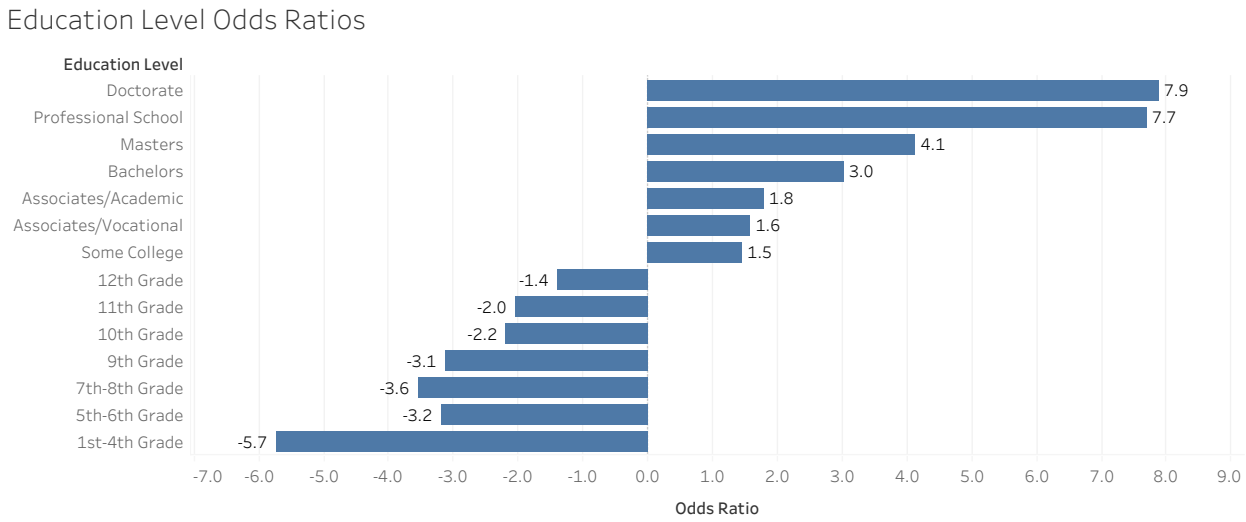

***Figure 1**: Odds Ratios*

As seen in Figure 1, an individual with a doctorate is 7.9 times more likely to have an income over $50,000 a year than an individual who is a high school graduate.