Income Classification Based on the 1994 US Census

Chong Li, Amanda Tsai, Lily Wang

5/9/2021

Introduction

With growing inequality in the American society, we are interested in seeing what factors contribute to divergence in individuals' income level. Understanding these factors is an important first step in working toward a more equitable distribution of income. We will be investigating these factors using data from the 1994 US census bureau database. In this specific case, we are looking to see what factors contribute to one's income to be greater than $50k (approximately $80k in 2021) in 1994.

The dataset contains 32561 observations and 15 variables. The outcome variable is a binary variable, `income`, which represents whether or not a person makes greater than or less than/equal to $50k a year. The predictor variables are: `age`, `fnlwgt`, `education.num`, `capital.gain`, `capital.loss`, `hours.per.week`, `workclass`, `education`, `marital.status`, `occupation`, `relationship`, `race`, `sex`, and `native.country`. The first 6 of these are continuous and the latter 8 are categorical.

Data source: https://www.kaggle.com/uciml/adult-census-income/

Data Cleaning

Many categorical predictors contained numerous levels (e.g. `native.country` contained 42). This meant that some levels ultimately have very few observations, especially in cases where the distributions were skewed. This poses an issue in data partitioning later on because oftentimes, the observations of a level could be allocated entirely to the training or testing data set. To remedy this issue, the levels of the predictors that contained more than 5 levels were grouped based on logical sense.

For `workclass`, all government jobs were combined into one category, and all self-employed jobs into another.

Similarly, `occupation` was grouped into 6 categories, according to the 2018 US Census Occupation code list. The categories were MBSA (management, business, science and arts), Service, Sales, NCM (Natural Resources, Construction and Maintenance), PTM (Production, Transportation and Material Moving), and Military.

All levels in `marital.status` that started with "Married" were grouped into one "Married" category and all countries in `native.country` that were not "United-States" into one "Others" category.

insert missing data visualization here

As seen from the table above, only 0.9% of the data was missing from just three predictors, so the missing observations were dropped. After dropping the missing observations, the "Without-pay" category of `workclass` and "Armed-Forces" category of `occupation` had very few observations and did not fit in with any of the other categories, thus all observations of those two categories were dropped as well. `fnlwgt`, an estimate derived such that people with similar demographic characteristics have similar weights, was dropped as it was decidedly unrelated to income.

After the cleaning process outlined above, our final dataset contained _____ observations and _____ variables.

Exploratory Analysis and Visualization

Models

Model Selection

All categorical predictors were turned into dummy variables and the data set was split into 70:30 training to test data. The training data with all predictors was then trained on a variety of models with ranging flexibility and assumptions: logistic regression, elastic net, LDA, QDA, MARS, and Random Forest (including bagging and boosting). The cross-validation and test AUC results are shown below:

Limitations

Conclusions