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 data set used 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 Armed-Forces.

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 logically fit in 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.

We also experimented with changing the continuous predictors `capital.gain` and `capital.loss` into binary (gain/loss or not) due to an excess of 0's, as well as categorizing `education` and `hours.per.week`. However,

none of these transformations improved cross-validation performance, possibly due to the loss of information from categorizing numeric data. Thus we performed a log(1+x) transformation on `capital.gain` and `capital.loss` to attempt to correct for skew, kept `hours.per.week` as is, and did not include `education` (due to its redundancy with `education.num`) in the final data set.

After the cleaning process outlined above, the final data set that we will perform EDA and modeling on contains 30139 observations and 13 variables.

Exploratory Analysis and Visualization

Among the continuous predictors, we can see that the age for those with incomes less than 50k per year is skewed younger. The age for those with incomes greater than 50k per year is relatively more normally distributed and the median age is higher than that of the other group. Within capital gain and capital loss, we can see that the values for these two predictors are both heavily skewed towards 0 for both income groups. Looking at the number of years of education, we can see that for those with incomes greater than 50k, the number of years is skewed higher, with the majority of observations having at least 8 years of education and close to half having more than 12. The education years for people with incomes less than 50k is more normally distributed, with the median being around 8 to 9 years of education, and less observations with over 12 years of education. We can also observe that those with incomes greater than 50k tend to work more hours per week.

Among the categorical predictors, we can see greater discrepancies in martial status, occupation, race, relationship and sex between the two income groups. Those with incomes greater than 50k per year tend to be married, with occupations in the MBSA category, white, male and husband of the household. Though the majority of these people are also native to the US and working in private companies, the ratio between the two income groups is similar for US and other country natives, as well as government, private and self-employed jobs, so a clear trend can't be identified for the `native.country` and `workclass` predictors.

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

Some predictors were non-normally distributed and heavily skewed, or had highly imbalanced classes. This may have affected the model performances for LDA and QDA, as both models assume the observations to be normally distributed.

Conclusions