Income classification based on the 1994 US census

Lily Wang

3/6/2021

Introduction

The dataset used in this project was extracted from the 1994 US Census Bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). 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. `education` was dropped as it was redundant with `education.num`.

Because many of the categorical predictors had multiple levels (e.g. `native.country` contained 42 unique observations), this meant that some levels ultimately have very few observations especially in cases where the distributions were skewed. This posed an issue in data partitioning later on because oftentimes, all the observations of a level could be allocated to the training or testing dataset.

To remedy this issue, the levels of the predictors that contained more than 5 levels were grouped based on logical sense. All the different "married" categories in `marital.status` were grouped together, as well as all the different "self-employed" categories in `workclass`. `occupation` was grouped according to the 2018 census occupation classification list. The abbreviations are as follows: MBSA = Managerial, Business, Science, and Arts, NCM = Natural Resources, Construction, and Maintenance, and PTM = Production, Transportation, and Material Moving.

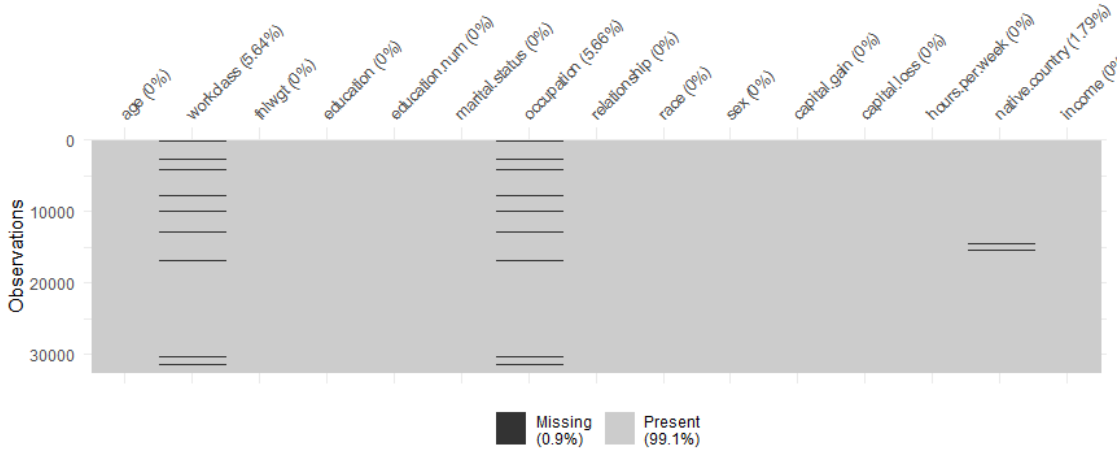

Figure 1. *Visualization of missing observations in the census dataset*

As seen from Fig 1 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.

After the cleaning process outlined above, our final dataset contained 30139 observations and 14 variables. Ultimately, we are interested in answering the following questions:

1. How accurately can we classify income based on the information we have?
2. Which variables are the most important in classifying income?

Exploratory Analysis and Visualization

Continuous Predictors

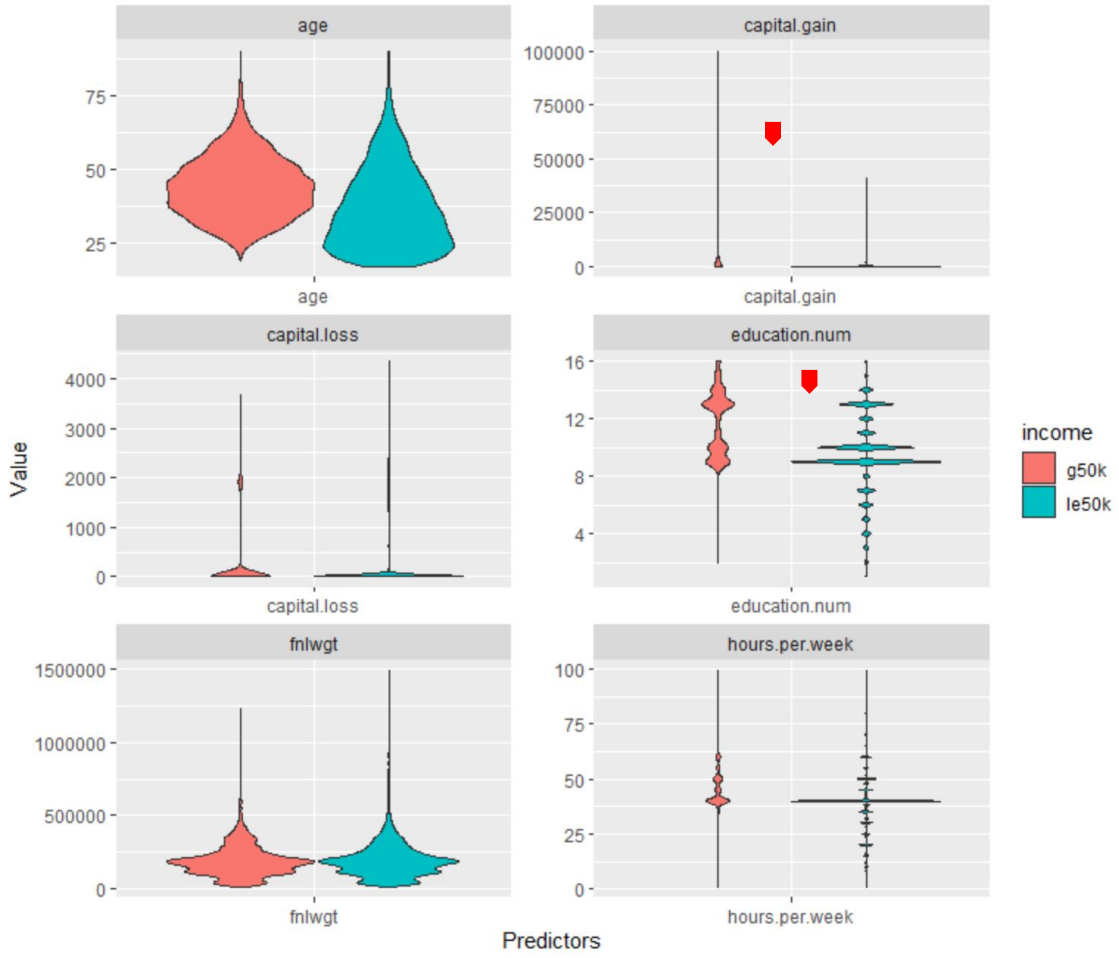

Figure 2. *Violin plots depicting the distributions of all continuous predictors by income*

Among the continuous predictors, overall, it seems that the age of people whose income is less than 50k is skewed younger, while the age of people whose income is more than 50k is more normally distributed and older. People whose income is more than 50k also tend to

be more highly educated and many of them work more than 40 hours per week. People whose income is less than or equal to 50k mostly work around 40 hours per week and have 0 capital gains. Furthermore, aside from age, all other predictors are not close to being normally distributed.

Categorical Predictors

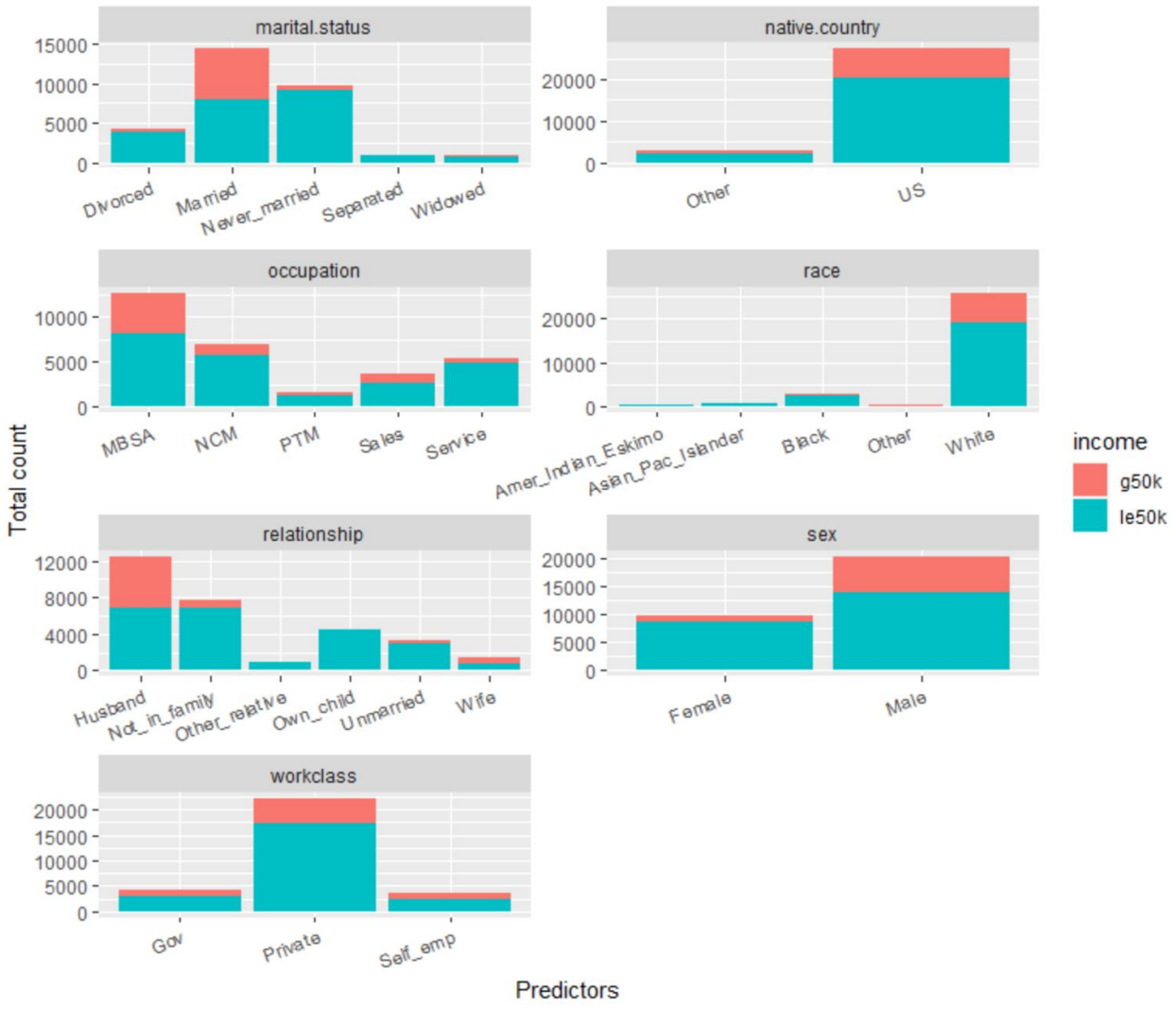

Figure 3. *Stacked bar charts of all categorical predictors by income*

Among the categorical predictors, it seems that, in comparison to all other categories and predictors, a greater proportion of people who make over 50k are married, white, in MBSA, male, or native to the US. Interestingly, although the proportion of over 50k to under 50k is smaller in females than males, that proportion looks to be around 50% in both the husband and wife categories in the `relationship` predictor.

Models

Model Selection

All categorical predictors were turned into dummy variables and the dataset 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, penalized logistic regression, LDA, QDA, KNN, and decision tree. Naive-Bayes and GAM were tried as well but they were met with errors/warnings. The cross-validation and testing AUC results are shown below:

Table 1. Cross-validation AUC of all models trained

Model	Mean AUC
GLM	0.901
GLMNET	0.901
LDA	0.889
QDA	0.865
KNN	0.883
RPART	0.888

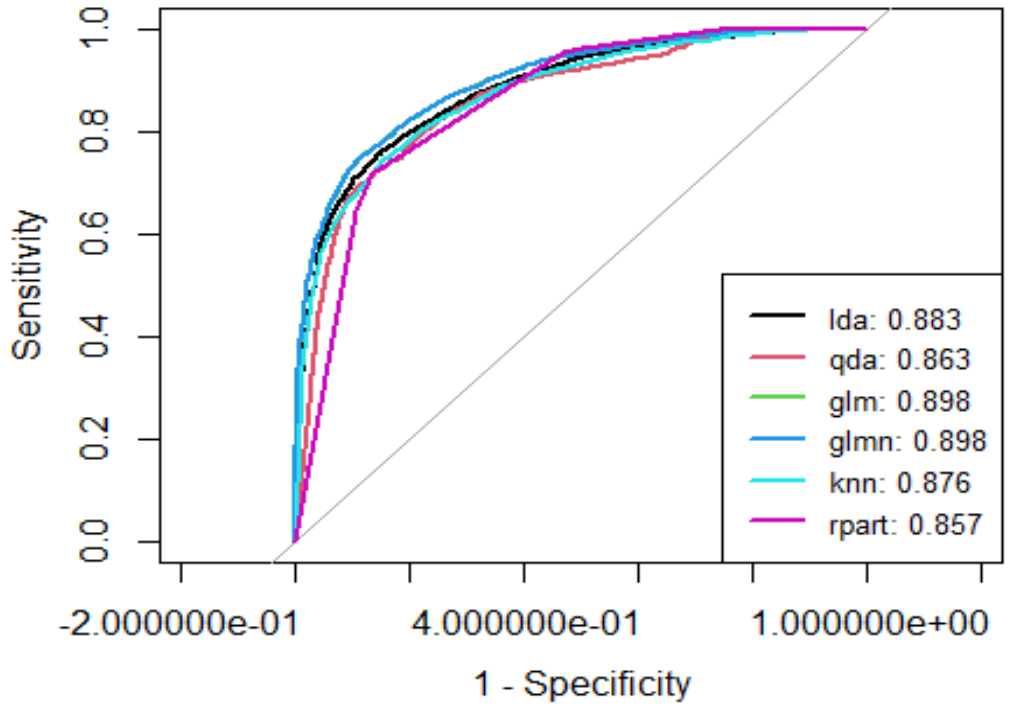

Figure 4. *ROC curves of all trained models on the test dataset*

While all models performed very well in classifying income on the training and testing datasets with AUC > 80, logistic regression and penalized logistic regression had the best performance on both the training and testing datasets. These two models are less flexible than some of the other models that were trained, like KNN and decision tree.

The logistic regression did not require parameter tuning, but fitting the penalized regression model required tuning both alpha and lambda, so an optimal grid was picked by graphing the cross-validated AUC and tuning the ranges of the grid so that the point at which AUC was maximized was contained within the graph.

Variable Importance

The important variables in the logistic regression and penalized regression model shared similarities. As seen from Fig 5 below, both models' top 10 most important variables contained the predictors `sexMale`, `occupationNCM`, `occupationService`, `occupationPTM`, and `relationshipWife`. However, the logistic regression model deemed all of the continuous predictors except `fnalwgt` as more important while the penalized regression model did not.

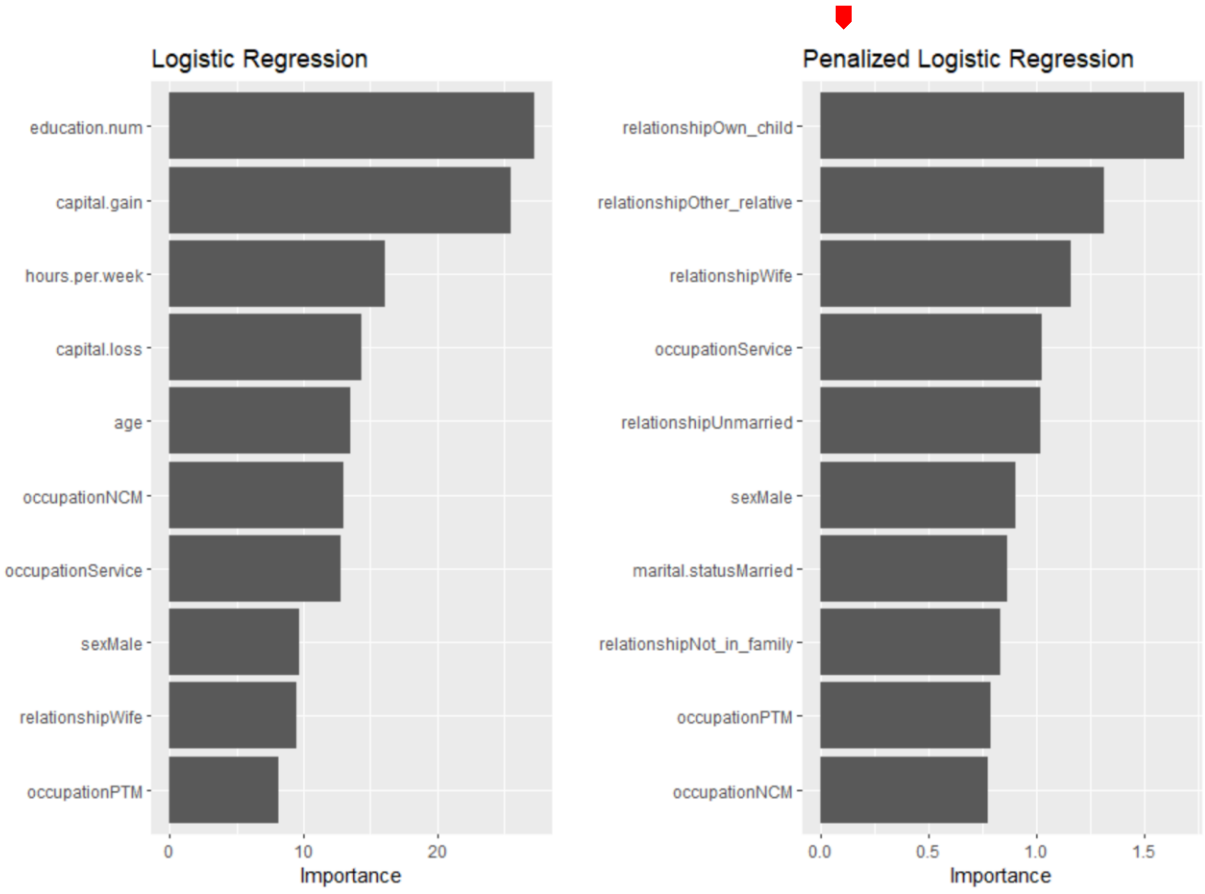

Figure 5. *Variable importance plots of the two best performing models*

Limitations

Running the logistic regression resulted in multiple warning messages: `glm.fit: fitted probabilities numerically 0 or 1 occurred`. These messages indicate that some of the classes could be well-separated, which in turn, would make the parameter estimates for the model unstable. However, while LDA handles well-separated classes better, it did not end up performing any better.

Additionally, due to the highly skewed, non-normal distribution of the predictors, some of the models with normal distribution assumptions like LDA and Naive-Bayes may not have performed as well despite their being quite robust to assumption violations. In fact, Naive-Bayes could not even finish running.

Finally, as we did decide to drop the observations that were missing as well as levels of predictors that did not have enough observations, doing so may have introduced some bias into our final dataset.

Conclusions

We found that from the 1994 US census dataset, logistic regression and penalized logistic regression performed the best at classifying income to less than/equal to 50k or above 50k. It was somewhat expected that logistic regression would perform better than LDA and Naive-Bayes given the extreme violation of normality assumptions in the predictors.

Some of the predictors discussed during EDA did end up becoming important predictors of income in the models, such as being male and being a wife (and years of education, hours worked per week, and age in the logistic regression model), which mostly make sense logically and have been found to have an effect on income through various studies.