Midterm Project

Chong Li

Contents

Introduction

Data and Motivation

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

The set includes 15 variables: 'age','workclass','fnlwgt','education','education.num','marital.status','occupation', 'relationship','race','sex','capital gain', 'capital loss', 'hours per week', 'native country' and 'income.

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

Data Description and Cleaning

The dataset 'adult.csv' had 32,561 observations with close to 2413 NA entries. Rows with missing values were first removed. Three columns 'fnlwgt', 'education' and 'relationship' were also removed before further cleaning because 'fnlwght', the population weight of each entry, was not relevant for our analysis and 'education'/'relationship' was shown to heavily correlate with 'education.num' and 'marital_status' respectively, which may lead to multicollinearity issues.

Aside from 'age', 'work hours' and 'education.num', all other variables were coded into numerical dummy variables with at most 5 classes. For variable 'workclass', all values containing 'gov' were merged to be the same level, while all values containing 'self-empl' were merged as well. A similar process was carried out for ‘marital_status’ and ‘occupation’. Furthermore, the ‘capital_gain/loss’ variables were transformed into zero/non-zero categorical variable as more than 80% of observations had zero as entry. Lastly, entries with occupation ‘armed-forces’ or workclass ‘no-income’ were excluded as both had very few (<30,0.1%) entries.

Exploratory analysis/visualization

For the exploratory analysis, I first examined the summary of the dataframe. It is heavily unbalanced with a majority being white (race=0), married (maritla_status=1), male(sex=1), non-immigrants (native_country=0). The median years of education was 10 and median work hours per week was 40. For the binary outcome variable ‘income’, about 75% of entries was <=50k while 25% was above >50k.

I also plotted bar graph for categorical variables and scatter plot for the three continuous variables. The bar graph (figure 1) shows having occupation in management, business and science significantly (occupation=0) increase the chance of making over 50k. For people with both capital gain/loss (capital_gain/loss=1), there is a higher chance in higher earnings. In addition, males also have a higher chance for elevated income than females and people who are married seem to be much more likely to have income over 50k. For continuous variables (figure 2), number of education and hours per week both have an influence on improving income from glancing at the graph.

Models

Methods

For the final predicted models, 11 predictors were included (8 categorical, 3 continuous)in total. All categorical predictors were number coded and continuous predictors were integers. Because this is a large data set, it was partitioned into training/test by a ratio of 80/20. The standard training control was set as 5 times repeated, 10-fold cross-validation.

The techniques used for the classification task including penalized logistic regression, regression trees, linear discriminant analysis and quadratic discriminant analysis. he discriminant analysis requires the assumption that the predictors follow a normal distribution, however, for classification tasks, discriminant analysis can still be quite robust even when the assumption is violated. While logistic regression requires no assumption for predictors’ feature, it’s still best for the independent variables to not highly correlate with each other. Similarly, tree-based models does not require any predictor normality assumptions.

For the penalized logistic regression, the tuning parameter is selected at where the AUC value is the largest. The best tune selected is when alpha = 1, indicating lasso penalty, with the other tuning parameter lambda at 0.0006. The lambda value close to 0 suggests that few beta parameters should be reduced to 0 from the regression. For the regression tree, the ideal ‘cp’ value is chosen when the increase size of splits no longer decrease RSS. From figure 3, the ‘cp’ value is approximately 0.004.

Results

For all four models that were used, as shown in figure 4, they yielded comparable AUC at close to 0.85, indicating high level of prediction accuracy. Logistic regression was the best performing model with prediction accuracy at 0.8369, 95% CI (0.8273,0.8461). From the variable importance measure (figure 5) of the logistic regression, we can see that the variables that the model relies the most on to make predictions are whether a person is married (marital_status=1), having capital gain/loss (capital_gain/loss = 1) and not having management occupation (occupation!=0).

The regression tree tells a similar story, being married and capital gains were the first two splits in the trees. In addition to the variables shown in the logistic regression, education.num > 13 also proves to be an important factor in predicting income higher than 50k.

My analysis have a number of limitations due to the dataset being unbalanced. Due to the fact that an overwhelmingly number of observations were white and other minority groups make up a small total percentage, it is likely that the analysis does not sufficiently capture the relationship between race and likelihood for high income. In addition, in the process of data cleaning, I dropped roughly 5% of total observations, which may lead my analysis to ignore some potential associations.

Conclusions

From the above results, some predictors for income >=50k align well with traditional wisdom. For instance, having gone to college (education.num>13) is considered to be a gateway for higher earning; the gender wage gap is very apparent with males much likely to earn more; the type of occupation also generally corresponds with earnings. In addition, results such as having non-zero capital gain OR loss contribute to higher earnings is surprising but reasonable: people who make more money tend to have an investment portfolio which would lead to passive income (or losses). Lastly, individual’s age does positively correlate with income potential.

In addition, the regression results also give me a number of results that we were not expecting. For starters, staying in marriage being the most important predictor of income greater than 50K. One potential cause of this relationship that I can think of is that married couples are generally entitled to more favorable tax laws. People who are never married are also less likely to make more than 50K than divorced/separated/widowed individuals. Part of this might be explained that unmarried individuals are on the younger end (increase age does positively correspond to earnings), but it could also be for other reasons that warrant further investigation.

Lastly, there are a number of relationships that I expected but was not demonstrated in the models. There does not seem to be a significant race income gap between “Black”, “White” and “Asian (although notable for native Americans). Being an”immigrant" only mildly reduces ones’ chance to have higher earning, which is also contrary to popular stereotypes that immigrants work in low-end jobs.

In conclusion, our analysis through using the 1994 income data provided us with many interesting insights. Being in a stable marriage turns out to be the most important determining factor in predicting one’s income. In addition, at the time this data was gathered, there existed a sizable gender gap in terms of income. Furthermore, people who have the ability and awareness to participate in capital management, even when return is negative, tend to be more likely to earn more money. Due to the nature of the data set, the relationship between race and income needs to be futher evaluated.

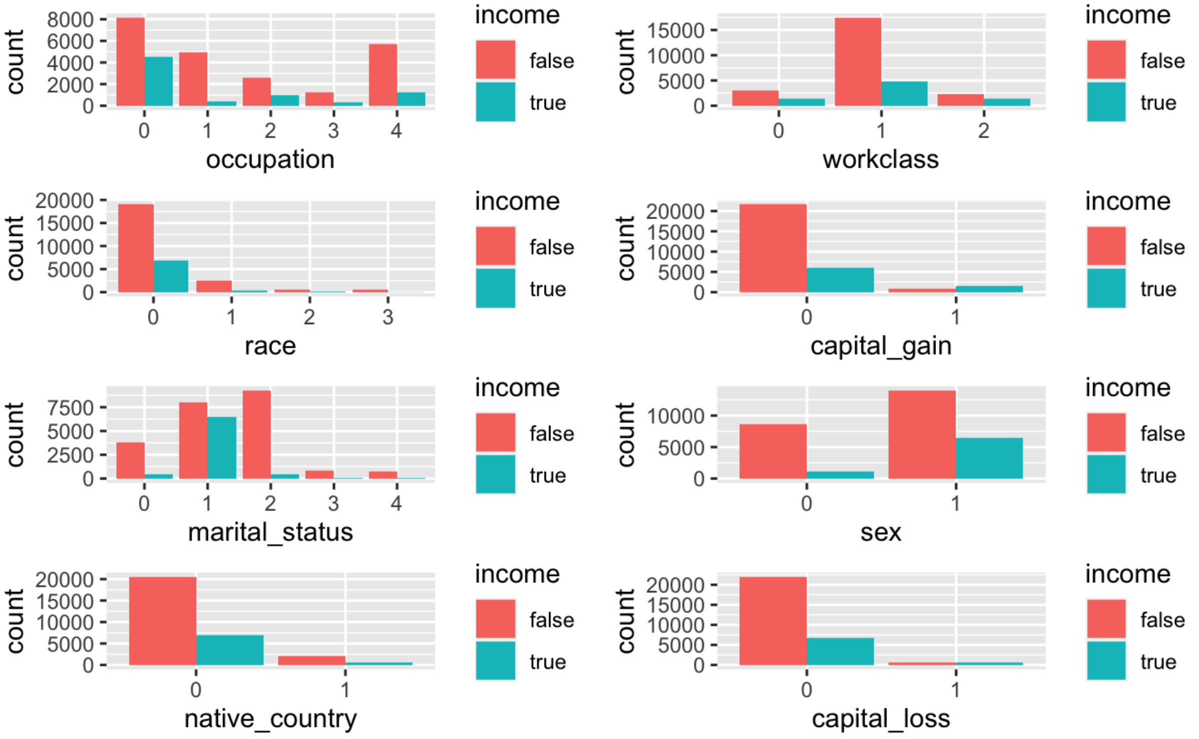


Figure 1: Categorical Variable vs Income

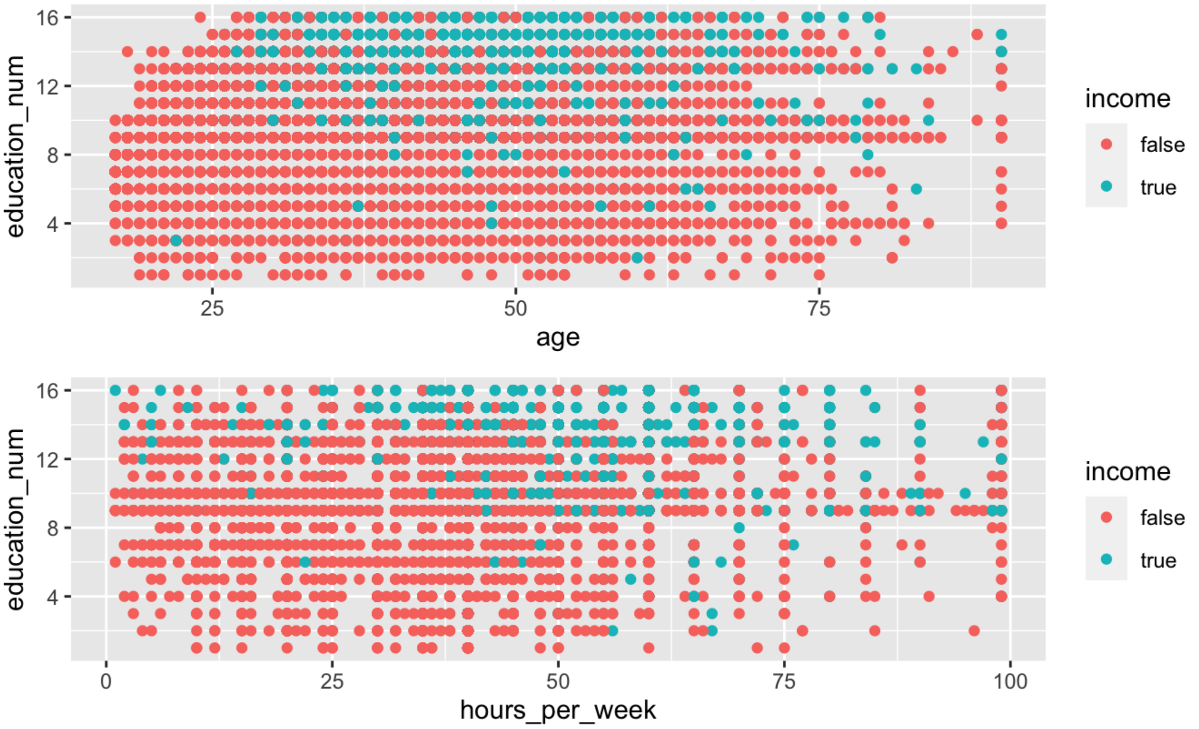


Figure 2: Continuous Variable vs Income

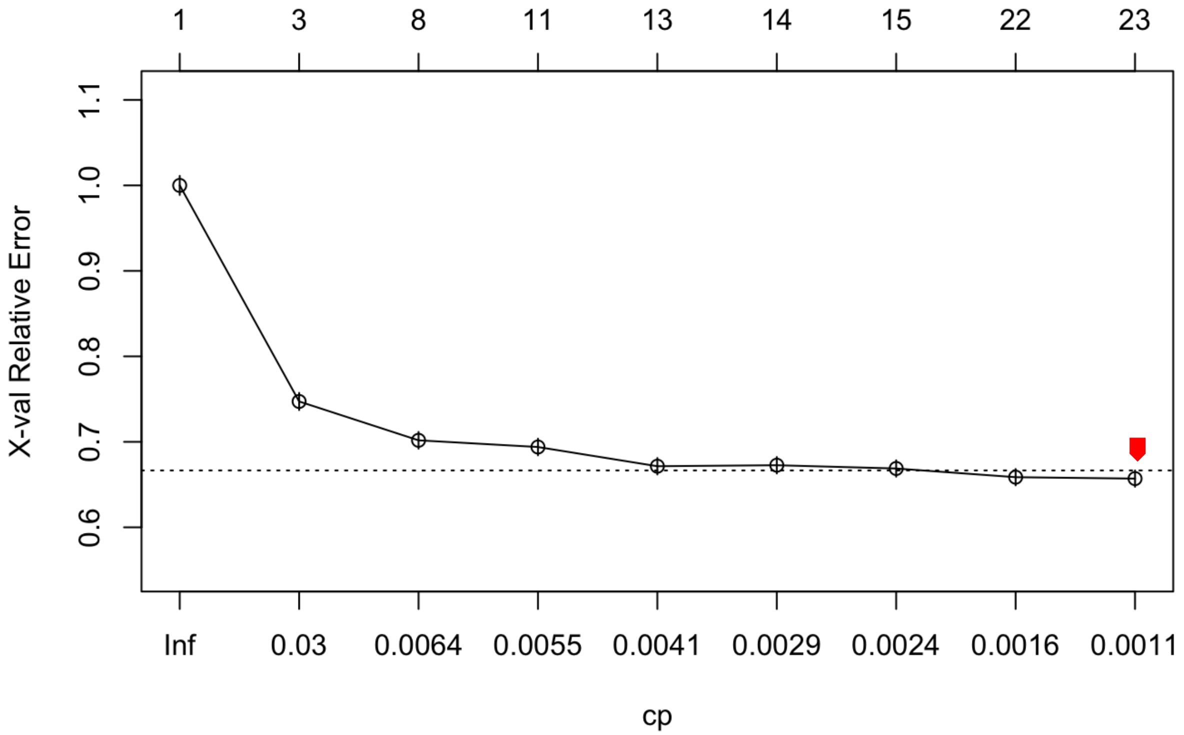


Figure 3: Regression Tree Tuning Selection

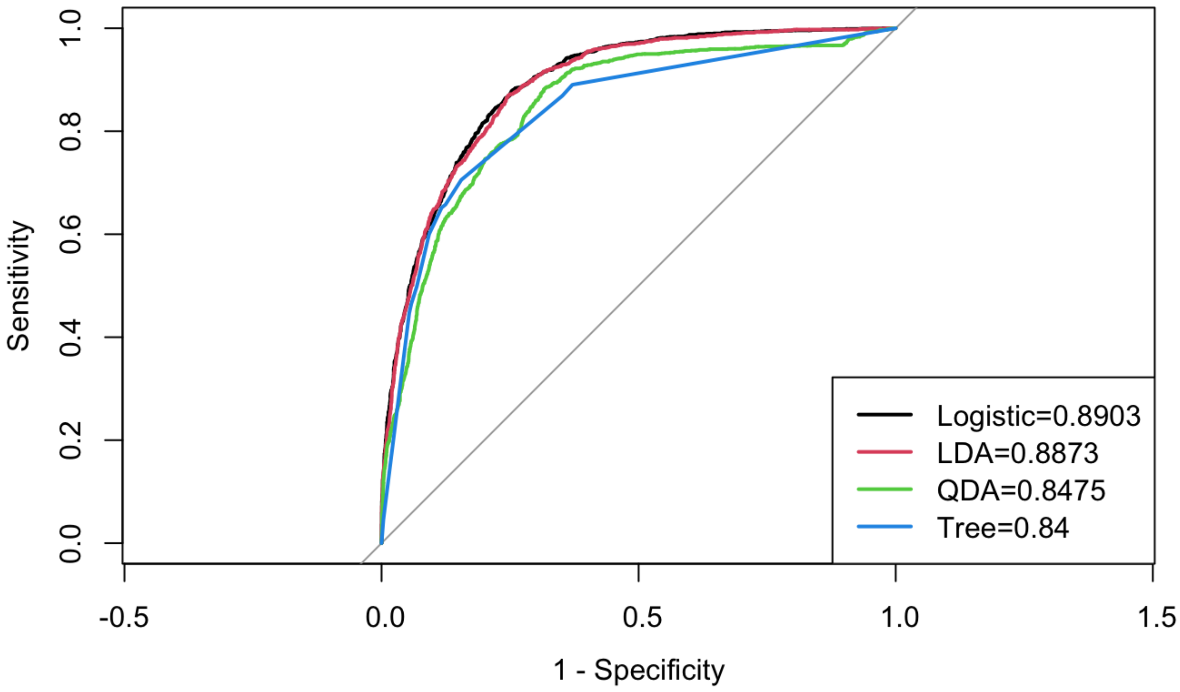


Figure 4: ROC Curves Comparison

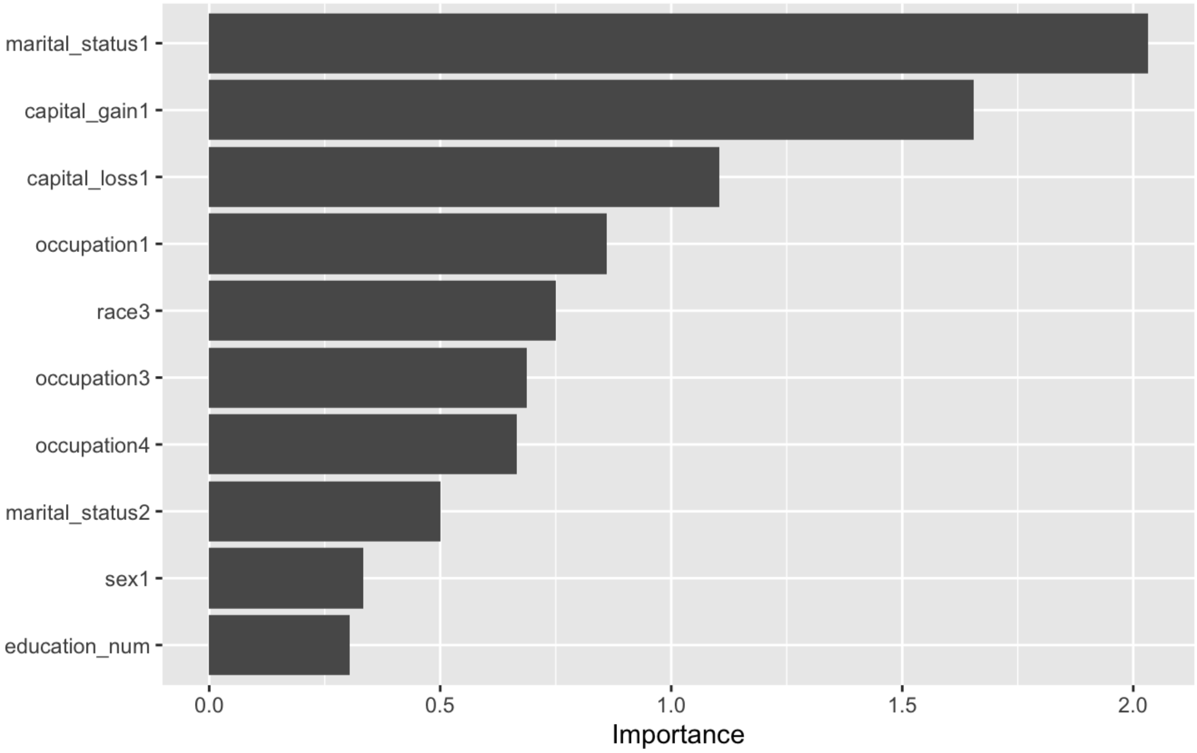


Figure 5: Variable Importance Measure

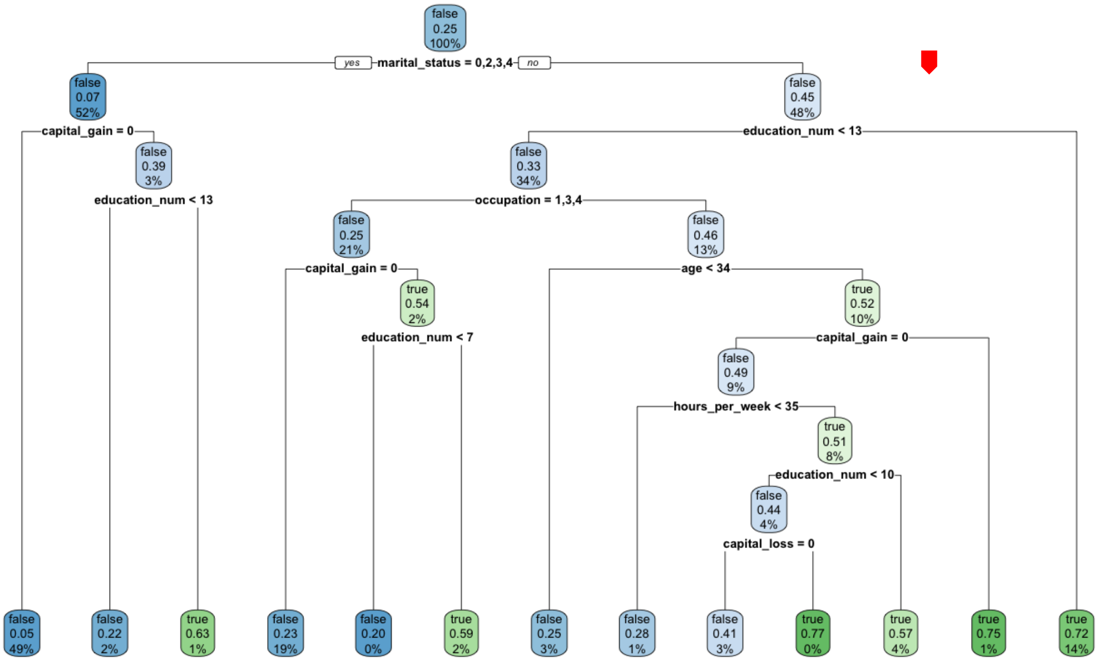
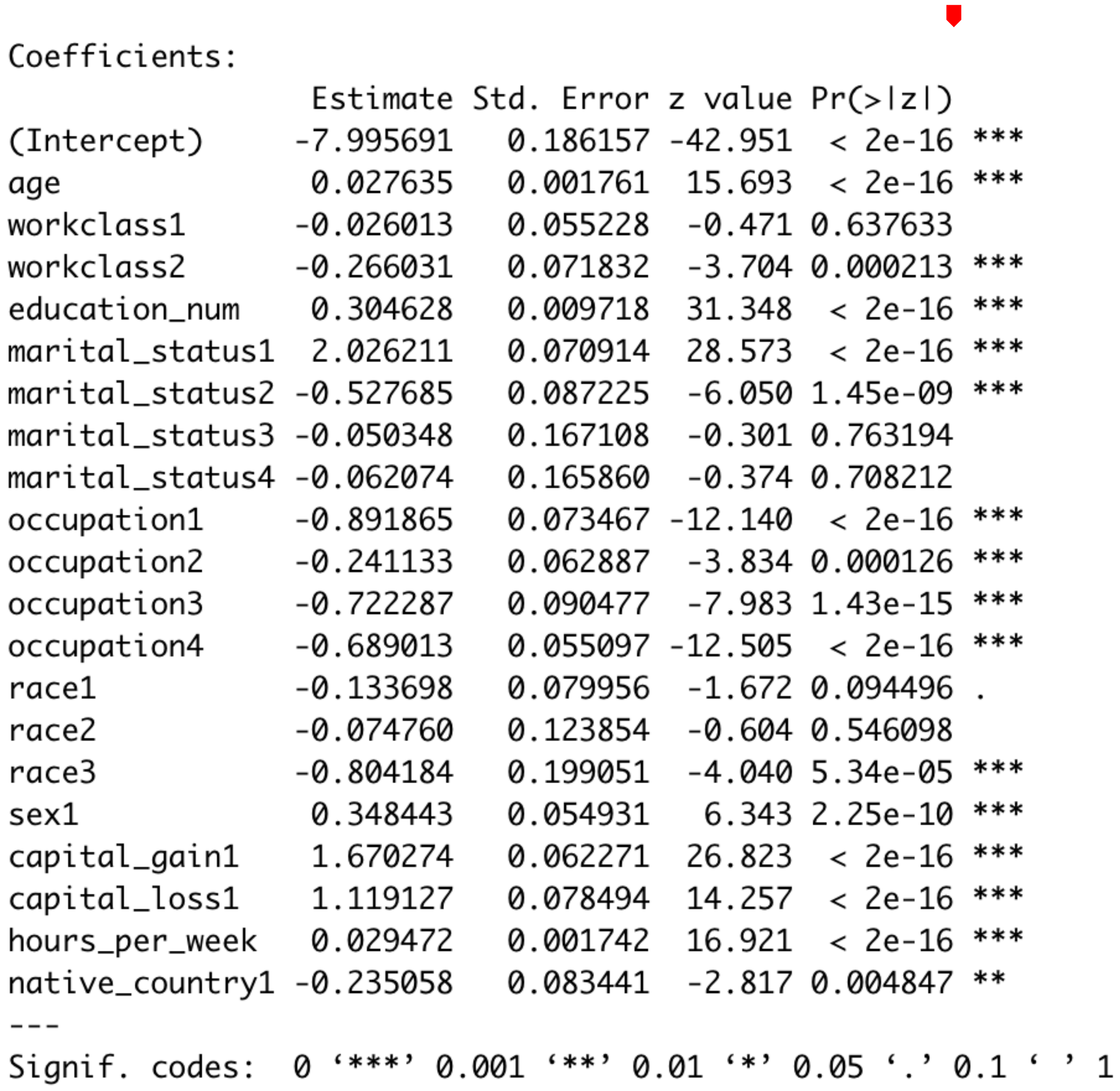


Figure 6: Regression Tree Graph



```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)     -7.995691   0.186157 -42.951  < 2e-16 ***
age              0.027635   0.001761  15.693  < 2e-16 ***
workclass1      -0.026013   0.055228  -0.471 0.637633    
workclass2      -0.266031   0.071832  -3.704 0.000213 ***
education_num    0.304628   0.009718  31.348  < 2e-16 ***
marital_status1  2.026211   0.070914  28.573  < 2e-16 ***
marital_status2 -0.527685   0.087225  -6.050 1.45e-09 ***
marital_status3 -0.050348   0.167108  -0.301 0.763194    
marital_status4 -0.062074   0.165860  -0.374 0.708212    
occupation1     -0.891865   0.073467 -12.140  < 2e-16 ***
occupation2     -0.241133   0.062887  -3.834 0.000126 ***
occupation3     -0.722287   0.090477  -7.983 1.43e-15 ***
occupation4     -0.689013   0.055097 -12.505  < 2e-16 ***
race1           -0.133698   0.079956  -1.672 0.094496 .  
race2           -0.074760   0.123854  -0.604 0.546098    
race3           -0.804184   0.199051  -4.040 5.34e-05 ***
sex1             0.348443   0.054931   6.343 2.25e-10 ***
capital_gain1    1.670274   0.062271  26.823  < 2e-16 ***
capital_loss1    1.119127   0.078494  14.257  < 2e-16 ***
hours_per_week   0.029472   0.001742  16.921  < 2e-16 ***
native_country1 -0.235058   0.083441  -2.817 0.004847 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 7: Logistic Regression Output