Income Prediction

Can we predict someone’s earnings based on demographic information?

Adil Iqbal, Angelita Krepel, Dosbol Aliev, Udochukwu Amaefule, Malik Taylor

11/22/2022

### Introduction

This data was extracted from the U.S. Census in 1994. Our goal will be to predict whether an individual earns more than $50,000 USD per year (class) based on data collected by the census alone. Examples of data collected by the census include education level (education), marital status (marital.status), and country of origin (native.country) among others. All features are either categorical or continuous integer values.

#### Features

age: continuous. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked. fnlwgt: continuous. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool. education.num: continuous. marital.status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse. occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handl sdasdasders-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black. sex: Female, Male. capital.gain: continuous. capital.loss: continuous. hours.per.week: continuous. native.country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

#### Response Variable

class: >50K, <=50K.

#### Question

Can we predict if an individual will make 50K or more based on a list of predictors?

#Dosbol  
set.seed(1)  
data = read.csv("adult.csv", stringsAsFactors = T)  
adult = read.csv("adult.csv")  
summary(data)

## age workclass fnlwgt   
## Min. :17.00 Private :33906 Min. : 12285   
## 1st Qu.:28.00 Self-emp-not-inc: 3862 1st Qu.: 117550   
## Median :37.00 Local-gov : 3136 Median : 178144   
## Mean :38.64 ? : 2799 Mean : 189664   
## 3rd Qu.:48.00 State-gov : 1981 3rd Qu.: 237642   
## Max. :90.00 Self-emp-inc : 1695 Max. :1490400   
## (Other) : 1463   
## education education.num marital.status   
## HS-grad :15784 Min. : 1.00 Divorced : 6633   
## Some-college:10878 1st Qu.: 9.00 Married-AF-spouse : 37   
## Bachelors : 8025 Median :10.00 Married-civ-spouse :22379   
## Masters : 2657 Mean :10.08 Married-spouse-absent: 628   
## Assoc-voc : 2061 3rd Qu.:12.00 Never-married :16117   
## 11th : 1812 Max. :16.00 Separated : 1530   
## (Other) : 7625 Widowed : 1518   
## occupation relationship race   
## Prof-specialty : 6172 Husband :19716 Amer-Indian-Eskimo: 470   
## Craft-repair : 6112 Not-in-family :12583 Asian-Pac-Islander: 1519   
## Exec-managerial: 6086 Other-relative: 1506 Black : 4685   
## Adm-clerical : 5611 Own-child : 7581 Other : 406   
## Sales : 5504 Unmarried : 5125 White :41762   
## Other-service : 4923 Wife : 2331   
## (Other) :14434   
## sex capital.gain capital.loss hours.per.week   
## Female:16192 Min. : 0 Min. : 0.0 Min. : 1.00   
## Male :32650 1st Qu.: 0 1st Qu.: 0.0 1st Qu.:40.00   
## Median : 0 Median : 0.0 Median :40.00   
## Mean : 1079 Mean : 87.5 Mean :40.42   
## 3rd Qu.: 0 3rd Qu.: 0.0 3rd Qu.:45.00   
## Max. :99999 Max. :4356.0 Max. :99.00   
##   
## native.country class   
## United-States:43832 <=50K :24720   
## Mexico : 951 <=50K.:12435   
## ? : 857 >50K : 7841   
## Philippines : 295 >50K. : 3846   
## Germany : 206   
## Puerto-Rico : 184   
## (Other) : 2517

#### Removing missing values.

In this data set, missing values are marked with a “?” character. Since all the missing data occurs in our categorical variables, it does not make sense to replace them with a median or a mean. So we are opting to remove them entirely for our analysis. After removing the observations with missing variables, we are left with 45,222 observations.

#Adil  
cols\_missing\_vals <- c("occupation", "workclass", "native.country")  
for (column in cols\_missing\_vals) {  
 data <- data[!grepl("\\?", data[, column]),]  
}  
summary(data)

## age workclass fnlwgt   
## Min. :17.00 Private :33307 Min. : 13492   
## 1st Qu.:28.00 Self-emp-not-inc: 3796 1st Qu.: 117388   
## Median :37.00 Local-gov : 3100 Median : 178316   
## Mean :38.55 State-gov : 1946 Mean : 189735   
## 3rd Qu.:47.00 Self-emp-inc : 1646 3rd Qu.: 237926   
## Max. :90.00 Federal-gov : 1406 Max. :1490400   
## (Other) : 21   
## education education.num marital.status   
## HS-grad :14783 Min. : 1.00 Divorced : 6297   
## Some-college: 9899 1st Qu.: 9.00 Married-AF-spouse : 32   
## Bachelors : 7570 Median :10.00 Married-civ-spouse :21055   
## Masters : 2514 Mean :10.12 Married-spouse-absent: 552   
## Assoc-voc : 1959 3rd Qu.:13.00 Never-married :14598   
## 11th : 1619 Max. :16.00 Separated : 1411   
## (Other) : 6878 Widowed : 1277   
## occupation relationship race   
## Craft-repair : 6020 Husband :18666 Amer-Indian-Eskimo: 435   
## Prof-specialty : 6008 Not-in-family :11702 Asian-Pac-Islander: 1303   
## Exec-managerial: 5984 Other-relative: 1349 Black : 4228   
## Adm-clerical : 5540 Own-child : 6626 Other : 353   
## Sales : 5408 Unmarried : 4788 White :38903   
## Other-service : 4808 Wife : 2091   
## (Other) :11454   
## sex capital.gain capital.loss hours.per.week   
## Female:14695 Min. : 0 Min. : 0.00 Min. : 1.00   
## Male :30527 1st Qu.: 0 1st Qu.: 0.00 1st Qu.:40.00   
## Median : 0 Median : 0.00 Median :40.00   
## Mean : 1101 Mean : 88.59 Mean :40.94   
## 3rd Qu.: 0 3rd Qu.: 0.00 3rd Qu.:45.00   
## Max. :99999 Max. :4356.00 Max. :99.00   
##   
## native.country class   
## United-States:41292 <=50K :22654   
## Mexico : 903 <=50K.:11360   
## Philippines : 283 >50K : 7508   
## Germany : 193 >50K. : 3700   
## Puerto-Rico : 175   
## Canada : 163   
## (Other) : 2213

#### Set categorical features to dummy variables

In this data, many categorical variables are needed to be converted to numeric and make a Dummy variable for the response variable.

#Dosbol  
### 1. Class (<=50K or 50K>)  
data$class = ifelse(data$class==" <=50K",0,1) # Our Response Variable. I made 0 for below 50K and 1 for above 50K  
data$class = as.factor(data$class)  
### 8. Native Country  
data$native.country = ifelse(data$native.country==" United-States",0,1)  
# So most people are from the United States, so I assigned them to 0 and other countries to 1  
data$native.country = as.factor(data$native.country)  
# Omit NA rows  
data = na.omit(data)

#### Split data into training and testing set

Below, we are now splitting the data into training and testing sets. With the training set containing 80% of the data and the testing set containing 20%.

#Adil  
sample <- sample(c(TRUE, FALSE), nrow(data), replace=TRUE, prob=c(0.8,0.2))  
train <- data[sample, ]  
test <- data[!sample, ]  
  
print(nrow(train))

## [1] 36212

print(nrow(test))

## [1] 9010

Checking for correlations

#Angelita  
#checking correlations  
pairs(adult[sapply(adult, is.numeric)]) A picture containing graphical user interface

Description automatically generated

cor(adult[sapply(adult, is.numeric)])

## age fnlwgt education.num capital.gain capital.loss  
## age 1.00000000 -0.076628079 0.03094038 0.077229022 0.05694383  
## fnlwgt -0.07662808 1.000000000 -0.03876068 -0.003706389 -0.00436615  
## education.num 0.03094038 -0.038760684 1.00000000 0.125146459 0.08097194  
## capital.gain 0.07722902 -0.003706389 0.12514646 1.000000000 -0.03144077  
## capital.loss 0.05694383 -0.004366150 0.08097194 -0.031440771 1.00000000  
## hours.per.week 0.07155834 -0.013518715 0.14368891 0.082157278 0.05446722  
## hours.per.week  
## age 0.07155834  
## fnlwgt -0.01351871  
## education.num 0.14368891  
## capital.gain 0.08215728  
## capital.loss 0.05446722  
## hours.per.week 1.00000000

From the output it doesn’t appear that from the original model there is much correlation between the data, but after making the occupation dummy values there may be some correlation between those.

Below, we will run our methodology on the pair of train/test set generated above. But will run our methods 10 times and deliver the average misclassification errors at the end of the experiment.

#### Model 1: Logistic regression

We will be using logistic regression, because the response variable is a categorical variable that takes 1 if their salary is >=50K. The advantages of using logistic regression is it is easy to implement and interpret, makes no assumptions about classes within the predictors, provides a measure of which predictors are most important in the predictions, and is less likely to overfit. The disadvantages of using logistic regression is non-linear problems cannot be used with this, requires little to no multicollinearity between independent variables. Lastly, with high dimensional datasets you can overfit/lead to less accurate results. The model formula for logistic regression is:

#Angelita  
lr.glm = glm(class ~., data = train, family = "binomial")  
glm.pred = predict.glm(lr.glm, test, type = "response")  
yHat = glm.pred > 0.5  
x1 <- table(test$class,yHat)

log\_reg\_misclass <- function(table) {  
 return((table[1,2]+table[2,1])/(table[1,1]+table[1,2]+table[2,1]+table[2,2]))   
}   
x2 <- log\_reg\_misclass(x1)  
x2

## [1] 0.3486127

The test error rate is kind of large sitting at about 35%. This means our model only predicted about 65% of the data correctly. Seeing this error rate, there may be some predictors that are less important/not needed in predicting the response, therefore there could be some predictors we can leave out of the model. So, the next step is to look at the summary to see which of those varibles are significant.

#Angelita  
summary(lr.glm)

(We are not going to be showing the summary here, run the base code to see it.) Based on the summary of the logistic regression model we can see that not all the variables are important in predicting, so we will try again but this time with the variables deemed important.

#Angelita   
lr.glm2 = glm(class ~ age+education+occupation+relationship+sex+capital.gain+capital.loss+hours.per.week, data = train, family = "binomial")  
do\_logistic\_regression <- function(train, test) {  
 lr.glm2 = glm(class ~ age+education+occupation+relationship+sex+capital.gain+capital.loss+hours.per.week, data = train, family = "binomial")  
 pred = predict.glm(lr.glm2, test, type = "response")  
 yHat2 = pred > 0.5  
 return(table(test$class,yHat2))  
}  
x2 <- do\_logistic\_regression(train, test)  
summary(lr.glm2)

log\_reg\_misclass(x2)

## [1] 0.3473918

The results when removing the less important variables come to be about the same as when leaving them in there.

#### Model 2: Decision Tree Model

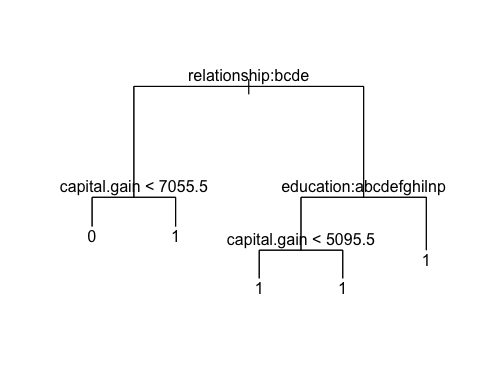
We will be using decision tree as our second model, because the dimensions of our dataset are pretty large and this should be able to manage it better than the logistic regression did. We are also using it, because our goal is to classify someone 1 if their salary is >=50K or 0 if below 50K. Some advantages of decision trees are: it is easy to understand and interpret, it takes less data preparation, is a non-parametric algorithm therefore needs little to no assumptions for classification, and you can use it for non-linear problems. Some disadvantages include that it can be prone to overfitting the data, it cannot handle too many features so you must do feature reduction, and lastly using this model can produce high variances and changes to the data can have a large affect in the predictions.

Here we create the tree model , where response variable is “class” and xi to xn are the variables used in prediction, and plot it.

#Udochukwu Amaefule  
library(tree)  
plot.new()  
do\_decision\_tree = function(train) {  
 class.tree = tree(class ~ ., data=train)  
 return(class.tree)  
}  
baseTree = do\_decision\_tree(train)  
summary(baseTree)

##   
## Classification tree:  
## tree(formula = class ~ ., data = train)  
## Variables actually used in tree construction:  
## [1] "relationship" "capital.gain" "education"   
## Number of terminal nodes: 5   
## Residual mean deviance: 1.254 = 45390 / 36210   
## Misclassification error rate: 0.3627 = 13133 / 36212

plot(baseTree)  
text(baseTree)

 As seen above, the tree model makes use of only 3 of the several variables available, “relationship”, “capital.gain”, “education”. The model has 5 terminal nodes and a misclassification error rate of 0.3627 or 36.3%

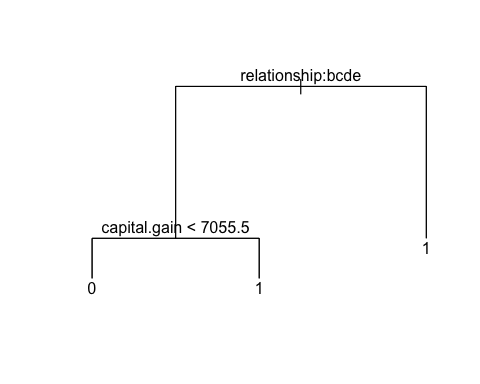
Usually after growing single decision tree, it has many leaf/terminal nodes, some of which might be important, whereas others are redundant and unnecessary. To further improve our results and accuracy of prediction, we perform cross validation on the tree to find out the best number of leaf/terminal nodes. Once we have obtained this, we can then proceed to prune the tree to obtain our final model. As shown above, the original tree has 5 terminal nodes. On the right half of the tree, most of those nodes result in the same decision, thus removing them should not change the error rate of the model. By cross validation we find that the best number of terminal nodes is 3, so our final tree after pruning will have 3 terminal nodes.

#### Cross Validation of tree

find\_best\_pruning\_param = function(class\_cv) {  
 mn = Inf;   
 mn\_idx = -1  
 for (x in length(class\_cv$dev):1) {   
 if (class\_cv$dev[x] < mn) { mn = class\_cv$dev[x]; mn\_idx = x }  
 }  
 return(class\_cv$size[mn\_idx])  
}

#Udochukwu Amaefule  
cv\_decision\_tree = function(baseTree) {  
 class.cv = cv.tree(baseTree,FUN = prune.misclass)  
 min\_index = find\_best\_pruning\_param(class.cv)  
 plot(class.cv$size,class.cv$dev,type = "b")  
 class.prune = prune.misclass(baseTree,best = min\_index)  
 summary(class.prune)  
 plot(class.prune)  
 text(class.prune)  
 x <- summary(baseTree)  
 return(x$misclass[1]/x$misclass[2])  
}  
errorRate = cv\_decision\_tree(baseTree) Chart, line chart

Description automatically generated



errorRate

## [1] 0.3626698

After pruning the tree model remains with a misclassification rate of 0.3627 or 36.3% as expected.

#### Average of 10 Trials

Now that we have built our models, we will evaluate them over 10 trials to determine the average misclassification rate.

# Redefined to avoid plots.  
cv\_decision\_tree = function(baseTree) {  
 class.cv = cv.tree(baseTree,FUN = prune.misclass)  
 min\_index = find\_best\_pruning\_param(class.cv)  
 class.prune = prune.misclass(baseTree,best = min\_index)  
 x <- summary(baseTree)  
 return(x$misclass[1]/x$misclass[2])  
}

iterations = 10  
lr\_err\_total = 0  
dtree\_err\_total = 0  
for (x in 1:iterations) {  
 set.seed(x)  
 sample <- sample(c(TRUE, FALSE), nrow(data), replace=TRUE, prob=c(0.8,0.2))  
 train <- data[sample, ]  
 test <- data[!sample, ]  
 table = do\_logistic\_regression(train, test)  
 lr\_err = log\_reg\_misclass(table)  
 lr\_err\_total = lr\_err\_total + lr\_err  
 baseTree = do\_decision\_tree(train)  
 dtree\_err = cv\_decision\_tree(baseTree)  
 dtree\_err\_total = dtree\_err\_total + dtree\_err  
 print(lr\_err)  
 print(dtree\_err)  
 print('--')  
}  
print('--- AVERAGE ERR ---')  
print(lr\_err\_total / iterations)  
print(dtree\_err\_total / iterations)

We can see from the above experiment, that the average misclassification rate for our linear regression model over 10 trials is 35.051% and our average misclassification rate for our decision tree is 36.187%

#### Best Model

Based on our models and their respective test error rates, the logistic regression model performed better showing to be the more accurate model in predicting whether or not an individual earns more than $50,000 USD per year. With a misclassification error rate of about 35.051% for the logistic regression model and 36.187% for the decision tree model. It is clear to see that logistic regression model does in fact perform slightly more accurate than the decision tree model. For the decision tree model, it is important to note that this test error rate comes after applying the pruning method to our model.

#### Conclusion

When using the logistic regression model we were able to achieve the smallest misclassificaiton error rate, and therefore is the best approach to our problem. Our initial error rate was 34.73%, after determining which predictors were not needed in our model we were able to improve our test error rate, however the improvement made only a slight difference, less than a 1%. The predictors found to be important to our model are the following: age, education, occupation, relationship, sex, capital.gain, capital.loss, hours.per.week. We would like to predict whether an individual will make $50,000 based on these important predictors. We do believe that these variables are significant in our prediction as those with a higher level of education such as a bachelors and higher, do in fact on average earn a higher salary then those with less education. One’s occupation also has a direct affect on one’s salary as well and generally higher paying occupations can be obtained with a higher level of education. Individuals that work more hours are often times compensated for their hours worked, through hourly pay or even overtime based compensation which has an effect on how much their yearly income is.

#### Bibliography

Satyam, Kumar. “Decision Tree in R Programming.” GeeksforGeeks, 3 Dec. 2021, <https://www.geeksforgeeks.org/decision-tree-in-r-programming/>.

“Home.” OARC Stats, <https://stats.oarc.ucla.edu/r/dae/logit-regression/>. Mishra, Rishu. “Cross-Validation in R Programming.” GeeksforGeeks, 15 Sept. 2021, <https://www.geeksforgeeks.org/cross-validation-in-r-programming/>.