CS X460 Project

Qianyi Guo

November 08, 2015

Contents

1 Introduction

This project aims at making predictions about people's annual income using demographic data. Understanding how other characteristics influence people's income is important for government, enterprises, and employers. The United States Census is a decennial census mandated by the United States Constitution, and provides a wide range of demographic data from US population. In this project, we want to use the US Census data about a person to predict how much the person earns – more specifically, whether the person earns more than 50,000 US dollars per year. In this project, we will analyze the dataset and build three types of models: logistic regression, classification and regression trees (CART), and random forests. Each model will be constructed using default parameters and improved later.

2 Data

The data comes from the **UCI Machine Learning Repository**, which can be downloaded from http://archive.ics.uci.edu/ml/datasets/Adult. Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: `((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))`.

The dataset includes the following 13 variables:

- `age` = the age of the individual in years
- `workclass` = the classification of the individual's working status (does the person work for the federal government, work for the local government, work without pay, and so on)
- `education` = the level of education of the individual (e.g., 5th-6th grade, high school graduate, PhD, so on)
- `maritalstatus` = the marital status of the individual
- `occupation` = the type of work the individual does (e.g., administrative/clerical work, farming/fishing, sales and so on)
- `relationship` = relationship of individual to his/her household
- `race` = the individual's race
- `sex` = the individual's sex
- `capitalgain` = the capital gains of the individual in 1994 (from selling an asset such as a stock or bond for more than the original purchase price)
- `capitalloss` = the capital losses of the individual in 1994 (from selling an asset such as a stock or bond for less than the original purchase price)
- `hoursperweek` = the number of hours the individual works per week
- `nativecountry` = the native country of the individual
- `over50k` = whether or not the individual earned more than $50,000 in 1994

First load the dataset from `census.csv`. Then split it into training and test set as usual.

```r
census = read.csv("census.csv")
str(census)
```

```
## 'data.frame':    31978 obs. of  13 variables:
##  $ age          : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass    : Factor w/ 9 levels " ?"," Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...
##  $ education    : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 13 7 12 13 10 ...
##  $ maritalstatus: Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5
##  $ occupation   : Factor w/ 15 levels " ?"," Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...
##  $ relationship : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ..
##  $ race         : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
##  $ sex          : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 2 1 2 ...
##  $ capitalgain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ capitalloss  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hoursperweek : int  40 13 40 40 40 40 16 45 50 40 ...
##  $ nativecountry: Factor w/ 41 levels " Cambodia"," Canada",..: 39 39 39 39 5 39 23 39 39 39
##  $ over50k      : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...
```

```r
summary(census)
```

```
##       age                       workclass             education
##  Min.   :17.00    Private           :22286    HS-grad     :10368
##  1st Qu.:28.00    Self-emp-not-inc: 2499    Some-college: 7187
##  Median :37.00    Local-gov         : 2067    Bachelors   : 5210
##  Mean   :38.58    ?                 : 1809    Masters     : 1674
##  3rd Qu.:48.00    State-gov         : 1279    Assoc-voc   : 1366
##  Max.   :90.00    Self-emp-inc      : 1074    11th        : 1167
##                  (Other)            :  964   (Other)      : 5006
##                  maritalstatus              occupation
##   Divorced              : 4394    Prof-specialty :4038
##   Married-AF-spouse     :   23    Craft-repair   :4030
##   Married-civ-spouse    :14692    Exec-managerial:3992
##   Married-spouse-absent:  397    Adm-clerical   :3721
##   Never-married         :10488    Sales          :3584
##   Separated             : 1005    Other-service  :3212
##   Widowed               :  979   (Other)         :9401
##           relationship                    race              sex
##   Husband       :12947    Amer-Indian-Eskimo:  311    Female:10608
##   Not-in-family : 8156    Asian-Pac-Islander:  956    Male  :21370
##   Other-relative:  952    Black             : 3028
##   Own-child     : 5005    Other             :  253
##   Unmarried     : 3384    White             :27430
##   Wife          : 1534
##
##   capitalgain     capitalloss       hoursperweek         nativecountry
##  Min.   :    0   Min.   :   0.00   Min.   : 1.00   United-States:29170
##  1st Qu.:    0   1st Qu.:   0.00   1st Qu.:40.00   Mexico       :  643
##  Median :    0   Median :   0.00   Median :40.00   Philippines  :  198
##  Mean   : 1064   Mean   :  86.74   Mean   :40.42   Germany      :  137
##  3rd Qu.:    0   3rd Qu.:   0.00   3rd Qu.:45.00   Canada       :  121
##  Max.   :99999   Max.   :4356.00   Max.   :99.00   Puerto-Rico  :  114
##                                                   (Other)       : 1595
##    over50k
##   <=50K:24283
##   >50K : 7695
##
##
##
##
##
```

```r
set.seed(1234)
library("caTools")
spl = sample.split(census$over50k, SplitRatio = 0.7)
```

```
train = subset(census, spl == TRUE)
test = subset(census, spl == FALSE)
```

3 Model Construction

3.1 Logistic Regression

First, build a most straightforward logistic regression model. The model simply fits the `over50k` response variable using all available predictors.

```
logModel = glm(over50k ~ . , data = train, family = "binomial")
summary(logModel)
```

```
##
## Call:
## glm(formula = over50k ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3666  -0.5115  -0.1849  -0.0216   3.6398
##
## Coefficients: (1 not defined because of singularities)
##                                             Estimate Std. Error z value
## (Intercept)                               -7.402e+00  8.884e-01  -8.333
## age                                        2.455e-02  1.975e-03  12.432
## workclass Federal-gov                      9.499e-01  1.865e-01   5.094
## workclass Local-gov                        2.504e-01  1.708e-01   1.466
## workclass Never-worked                    -1.211e+01  5.907e+02  -0.021
## workclass Private                          5.232e-01  1.523e-01   3.437
## workclass Self-emp-inc                     6.673e-01  1.816e-01   3.675
## workclass Self-emp-not-inc                 8.726e-02  1.664e-01   0.524
## workclass State-gov                        1.451e-01  1.851e-01   0.784
## workclass Without-pay                     -1.318e+01  4.947e+02  -0.027
## education 11th                            -2.940e-02  2.513e-01  -0.117
## education 12th                             4.113e-01  3.217e-01   1.279
## education 1st-4th                         -1.028e+00  7.148e-01  -1.438
## education 5th-6th                         -3.328e-01  3.974e-01  -0.837
## education 7th-8th                         -6.467e-01  2.841e-01  -2.276
## education 9th                             -4.172e-01  3.213e-01  -1.299
## education Assoc-acdm                       1.075e+00  2.135e-01   5.033
## education Assoc-voc                        1.264e+00  2.041e-01   6.192
## education Bachelors                        1.828e+00  1.906e-01   9.592
## education Doctorate                        2.914e+00  2.633e-01  11.065
## education HS-grad                          7.191e-01  1.855e-01   3.877
## education Masters                          2.250e+00  2.039e-01  11.035
```

```
## education Preschool                      -1.980e+01  1.681e+02  -0.118
## education Prof-school                     2.651e+00  2.396e-01  11.065
## education Some-college                    1.011e+00  1.882e-01   5.373
## maritalstatus Married-AF-spouse           3.505e+00  6.801e-01   5.153
## maritalstatus Married-civ-spouse          2.100e+00  3.152e-01   6.662
## maritalstatus Married-spouse-absent      -2.108e-01  2.911e-01  -0.724
## maritalstatus Never-married              -5.323e-01  1.032e-01  -5.159
## maritalstatus Separated                  -1.486e-01  1.997e-01  -0.744
## maritalstatus Widowed                     3.786e-02  1.839e-01   0.206
## occupation Adm-clerical                   1.560e-01  1.190e-01   1.311
## occupation Armed-Forces                  -9.086e-01  1.511e+00  -0.601
## occupation Craft-repair                   2.077e-01  1.022e-01   2.031
## occupation Exec-managerial                9.453e-01  1.049e-01   9.008
## occupation Farming-fishing               -8.931e-01  1.694e-01  -5.271
## occupation Handlers-cleaners             -4.389e-01  1.725e-01  -2.544
## occupation Machine-op-inspct             -8.498e-02  1.263e-01  -0.673
## occupation Other-service                 -6.829e-01  1.512e-01  -4.515
## occupation Priv-house-serv               -3.458e+00  2.052e+00  -1.685
## occupation Prof-specialty                 6.411e-01  1.128e-01   5.681
## occupation Protective-serv                8.074e-01  1.573e-01   5.133
## occupation Sales                          4.477e-01  1.084e-01   4.130
## occupation Tech-support                   8.014e-01  1.433e-01   5.593
## occupation Transport-moving                      NA         NA      NA
## relationship Not-in-family                5.445e-01  3.117e-01   1.747
## relationship Other-relative              -1.858e-01  2.899e-01  -0.641
## relationship Own-child                   -8.099e-01  3.100e-01  -2.612
## relationship Unmarried                    3.687e-01  3.313e-01   1.113
## relationship Wife                         1.358e+00  1.230e-01  11.044
## race Asian-Pac-Islander                   1.129e+00  3.423e-01   3.297
## race Black                                7.461e-01  2.923e-01   2.552
## race Other                                5.743e-01  4.536e-01   1.266
## race White                                8.577e-01  2.799e-01   3.064
## sex Male                                  8.378e-01  9.372e-02   8.940
## capitalgain                               3.208e-04  1.234e-05  25.987
## capitalloss                               6.131e-04  4.458e-05  13.755
## hoursperweek                              3.016e-02  1.958e-03  15.401
## nativecountry Canada                     -1.189e+00  7.997e-01  -1.486
## nativecountry China                      -1.964e+00  8.087e-01  -2.428
## nativecountry Columbia                   -3.333e+00  1.279e+00  -2.605
## nativecountry Cuba                       -1.178e+00  8.301e-01  -1.420
## nativecountry Dominican-Republic         -1.437e+01  1.891e+02  -0.076
## nativecountry Ecuador                    -1.398e+00  1.119e+00  -1.249
## nativecountry El-Salvador                -1.937e+00  9.669e-01  -2.003
## nativecountry England                    -1.382e+00  8.275e-01  -1.670
## nativecountry France                     -8.393e-01  9.425e-01  -0.891
## nativecountry Germany                    -9.936e-01  7.870e-01  -1.263
## nativecountry Greece                     -2.356e+00  1.014e+00  -2.323
## nativecountry Guatemala                  -1.194e+00  1.088e+00  -1.098
```

```
## nativecountry Haiti                      -1.482e+00  1.102e+00  -1.345
## nativecountry Holand-Netherlands         -1.310e+01  1.455e+03  -0.009
## nativecountry Honduras                   -1.323e+01  4.222e+02  -0.031
## nativecountry Hong                       -1.585e+00  1.064e+00  -1.489
## nativecountry Hungary                    -1.386e+00  1.136e+00  -1.220
## nativecountry India                      -1.823e+00  7.797e-01  -2.338
## nativecountry Iran                       -1.495e+00  8.933e-01  -1.674
## nativecountry Ireland                    -5.079e-03  1.015e+00  -0.005
## nativecountry Italy                      -5.697e-01  8.262e-01  -0.690
## nativecountry Jamaica                    -2.152e+00  9.864e-01  -2.182
## nativecountry Japan                      -1.162e+00  8.391e-01  -1.385
## nativecountry Laos                       -2.007e+00  1.098e+00  -1.828
## nativecountry Mexico                     -1.928e+00  7.756e-01  -2.486
## nativecountry Nicaragua                  -1.415e+01  2.510e+02  -0.056
## nativecountry Outlying-US(Guam-USVI-etc) -1.503e+01  5.638e+02  -0.027
## nativecountry Peru                       -1.756e+00  1.153e+00  -1.523
## nativecountry Philippines                -1.228e+00  7.539e-01  -1.629
## nativecountry Poland                     -1.051e+00  8.529e-01  -1.232
## nativecountry Portugal                   -1.558e+00  1.181e+00  -1.319
## nativecountry Puerto-Rico                -1.670e+00  8.635e-01  -1.934
## nativecountry Scotland                   -2.227e+00  1.377e+00  -1.617
## nativecountry South                      -2.444e+00  8.422e-01  -2.902
## nativecountry Taiwan                     -1.275e+00  8.743e-01  -1.458
## nativecountry Thailand                   -1.718e+00  1.166e+00  -1.474
## nativecountry Trinadad&Tobago            -1.471e+01  3.295e+02  -0.045
## nativecountry United-States              -1.152e+00  7.314e-01  -1.575
## nativecountry Vietnam                    -2.254e+00  9.241e-01  -2.439
## nativecountry Yugoslavia                 -1.666e-01  1.241e+00  -0.134
##                                          Pr(>|z|)
## (Intercept)                               < 2e-16 ***
## age                                       < 2e-16 ***
## workclass Federal-gov                    3.51e-07 ***
## workclass Local-gov                      0.142624
## workclass Never-worked                   0.983641
## workclass Private                        0.000589 ***
## workclass Self-emp-inc                   0.000238 ***
## workclass Self-emp-not-inc               0.600047
## workclass State-gov                      0.433060
## workclass Without-pay                    0.978739
## education 11th                           0.906890
## education 12th                           0.200964
## education 1st-4th                        0.150332
## education 5th-6th                        0.402314
## education 7th-8th                        0.022825 *
## education 9th                            0.194057
## education Assoc-acdm                     4.82e-07 ***
## education Assoc-voc                      5.93e-10 ***
## education Bachelors                       < 2e-16 ***
```

```
## education Doctorate                     < 2e-16 ***
## education HS-grad                      0.000106 ***
## education Masters                       < 2e-16 ***
## education Preschool                    0.906254
## education Prof-school                   < 2e-16 ***
## education Some-college                 7.76e-08 ***
## maritalstatus Married-AF-spouse        2.56e-07 ***
## maritalstatus Married-civ-spouse       2.70e-11 ***
## maritalstatus Married-spouse-absent    0.468966
## maritalstatus Never-married            2.48e-07 ***
## maritalstatus Separated                0.456991
## maritalstatus Widowed                  0.836849
## occupation Adm-clerical                0.189936
## occupation Armed-Forces                0.547609
## occupation Craft-repair                0.042209 *
## occupation Exec-managerial              < 2e-16 ***
## occupation Farming-fishing             1.36e-07 ***
## occupation Handlers-cleaners           0.010957 *
## occupation Machine-op-inspct           0.501159
## occupation Other-service               6.32e-06 ***
## occupation Priv-house-serv             0.091992 .
## occupation Prof-specialty              1.34e-08 ***
## occupation Protective-serv             2.86e-07 ***
## occupation Sales                       3.63e-05 ***
## occupation Tech-support                2.24e-08 ***
## occupation Transport-moving                  NA
## relationship Not-in-family             0.080678 .
## relationship Other-relative            0.521688
## relationship Own-child                 0.008994 **
## relationship Unmarried                 0.265797
## relationship Wife                       < 2e-16 ***
## race Asian-Pac-Islander                0.000977 ***
## race Black                             0.010697 *
## race Other                             0.205491
## race White                             0.002183 **
## sex Male                                < 2e-16 ***
## capitalgain                             < 2e-16 ***
## capitalloss                             < 2e-16 ***
## hoursperweek                            < 2e-16 ***
## nativecountry Canada                   0.137156
## nativecountry China                    0.015166 *
## nativecountry Columbia                 0.009188 **
## nativecountry Cuba                     0.155707
## nativecountry Dominican-Republic       0.939442
## nativecountry Ecuador                  0.211556
## nativecountry El-Salvador              0.045191 *
## nativecountry England                  0.094890 .
## nativecountry France                   0.373179
```

```
## nativecountry Germany                    0.206763
## nativecountry Greece                     0.020179 *
## nativecountry Guatemala                  0.272220
## nativecountry Haiti                      0.178702
## nativecountry Holand-Netherlands         0.992820
## nativecountry Honduras                   0.975006
## nativecountry Hong                       0.136377
## nativecountry Hungary                    0.222309
## nativecountry India                      0.019364 *
## nativecountry Iran                       0.094228 .
## nativecountry Ireland                    0.996006
## nativecountry Italy                      0.490508
## nativecountry Jamaica                    0.029140 *
## nativecountry Japan                      0.166092
## nativecountry Laos                       0.067614 .
## nativecountry Mexico                     0.012931 *
## nativecountry Nicaragua                  0.955040
## nativecountry Outlying-US(Guam-USVI-etc) 0.978738
## nativecountry Peru                       0.127677
## nativecountry Philippines                0.103404
## nativecountry Poland                     0.218066
## nativecountry Portugal                   0.187195
## nativecountry Puerto-Rico                0.053152 .
## nativecountry Scotland                   0.105829
## nativecountry South                      0.003714 **
## nativecountry Taiwan                     0.144861
## nativecountry Thailand                   0.140522
## nativecountry Trinadad&Tobago            0.964393
## nativecountry United-States              0.115232
## nativecountry Vietnam                    0.014711 *
## nativecountry Yugoslavia                 0.893271
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24703  on 22383  degrees of freedom
## Residual deviance: 14214  on 22287  degrees of freedom
## AIC: 14408
##
## Number of Fisher Scoring iterations: 14
```

```
logPred = predict(logModel, newdata = test, type = "response")
```

Now we look into the predictions of this model. The accuracy of this model is given by

```r
logTable = table(test$over50k, logPred > 0.5)
(logTable[[1]] + logTable[[4]]) / nrow(test)
```

```
## [1] 0.8528247
```

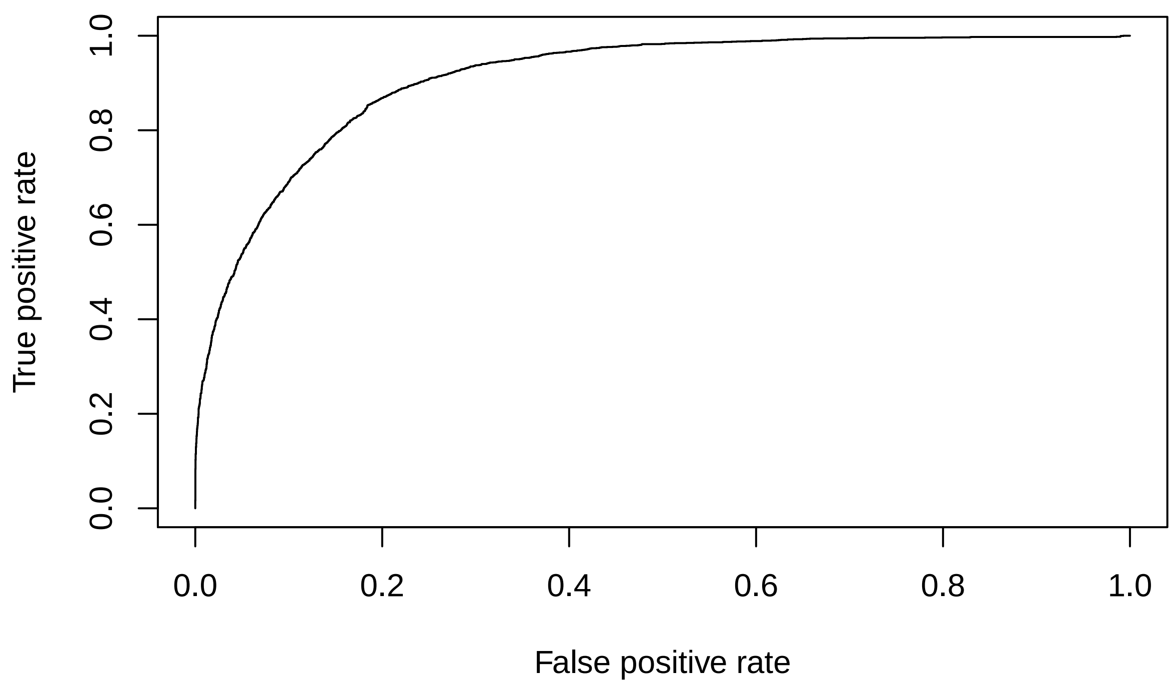
Comparing with the "naive model" – predicting annual income <= 50K for any people, the logistic regression model improves accuracy by about 10%. It is a decent progress, but we obviously want to do something more accurate.

```r
testTable = summary(test$over50k)
testTable[[1]] / nrow(test)
```

```
## [1] 0.7593287
```

Finally, visualize the relationship between false positive rate and true positive rate:

```r
library(ROCR)
ROCRpred = prediction(logPred, test$over50k)
ROCRperf = performance(ROCRpred, "tpr", "fpr")
plot(ROCRperf)
```



The area under the ROC curve is

```r
auc = as.numeric(performance(ROCRpred, "auc")@y.values)
auc
```

```
## [1] 0.9096057
```

3.2 Classification and Regression Trees

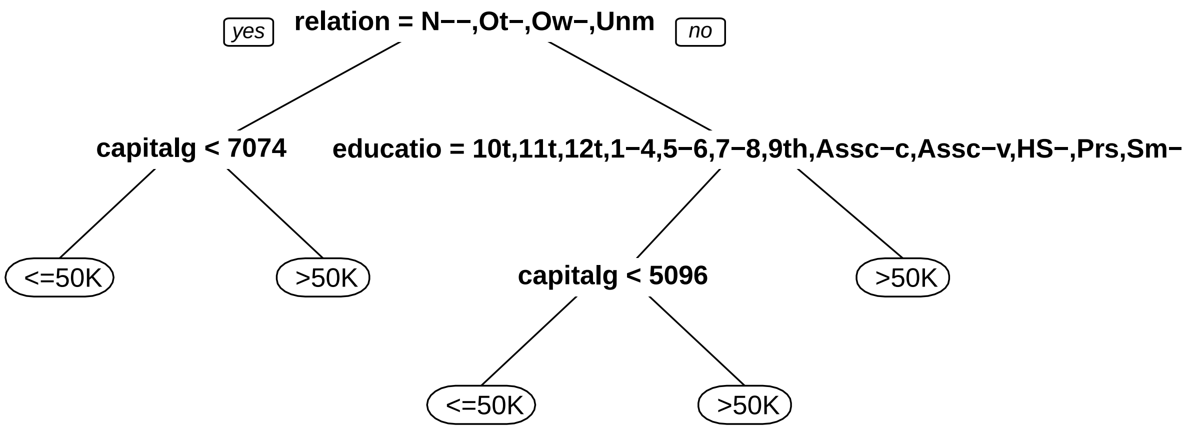
Since the dataset includes lots of categorical variables, some advanced methods like classification and regression trees may produce a more accurate model.

```r
library("rpart")
library("rpart.plot")
CARTmodel = rpart(over50k ~ ., data = train, method = "class")
CARTpred = predict(CARTmodel, newdata = test, type = "class")
CARTtable = table(CARTpred, test$over50k)
(CARTtable[[1]] + CARTtable[[4]]) / nrow(test)
```

```
## [1] 0.8475089
```

The CART model is slightly worse than the logistic regression model in terms of accuracy. But it is much easier to interpret since only a few predictors are involved in the model, which can be demonstrated using a graph:

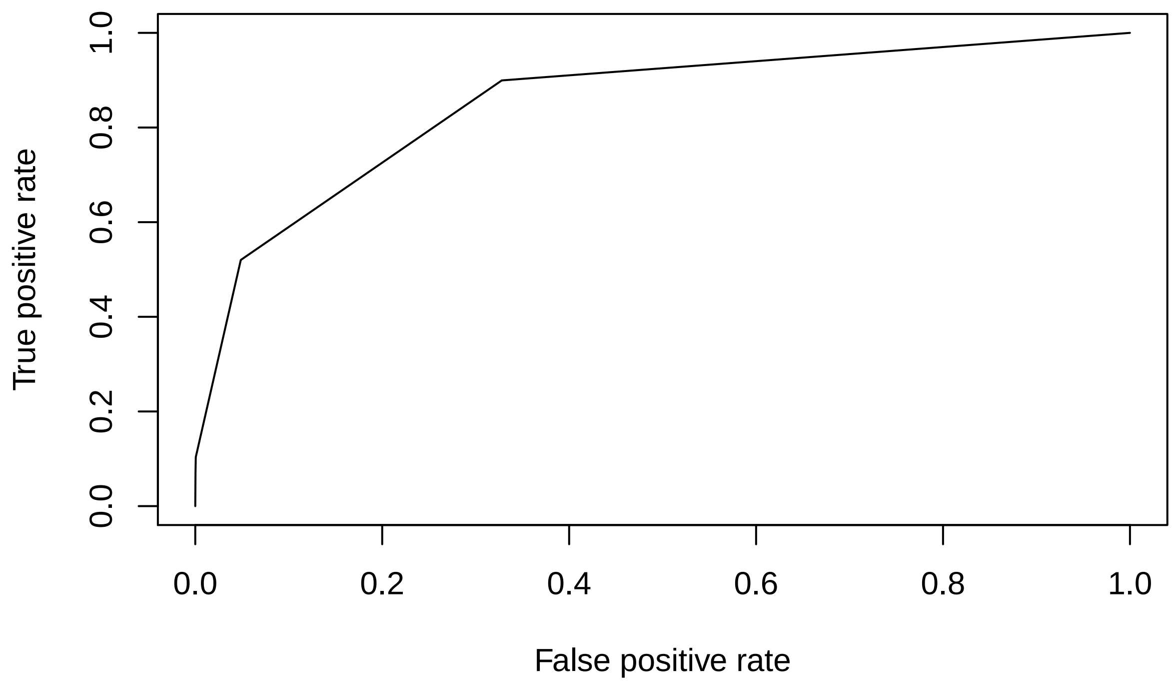
```r
prp(CARTmodel)
```



It looks like that `relationship`, `captialgain` and `education` are most important predictors in the CART model. But we can do better by appropriately tune the depth and number of splits in the tree, which will be discussed later.

Finally, visualize the relationship between false positive rate and true positive rate:

```
CARTpredVal = predict(CARTmodel, newdata=test)[,2]
ROCRpred = prediction(CARTpredVal, test$over50k)
ROCRperf = performance(ROCRpred, "tpr", "fpr")
plot(ROCRperf)
```



And the area under the ROC curve is

```
auc = as.numeric(performance(ROCRpred, "auc")@y.values)
auc
```

```
## [1] 0.8515478
```

3.3 Random Forests

Since the performance of the CART model isn't that satisfactory, we will use a more sophisticated method known as random forests. It can correct the habit of decision trees to overfit the training set.

```
library("randomForest")
set.seed(3333)
rfModel = randomForest(over50k ~., data = train)
rfPred = predict(rfModel, newdata = test)
```

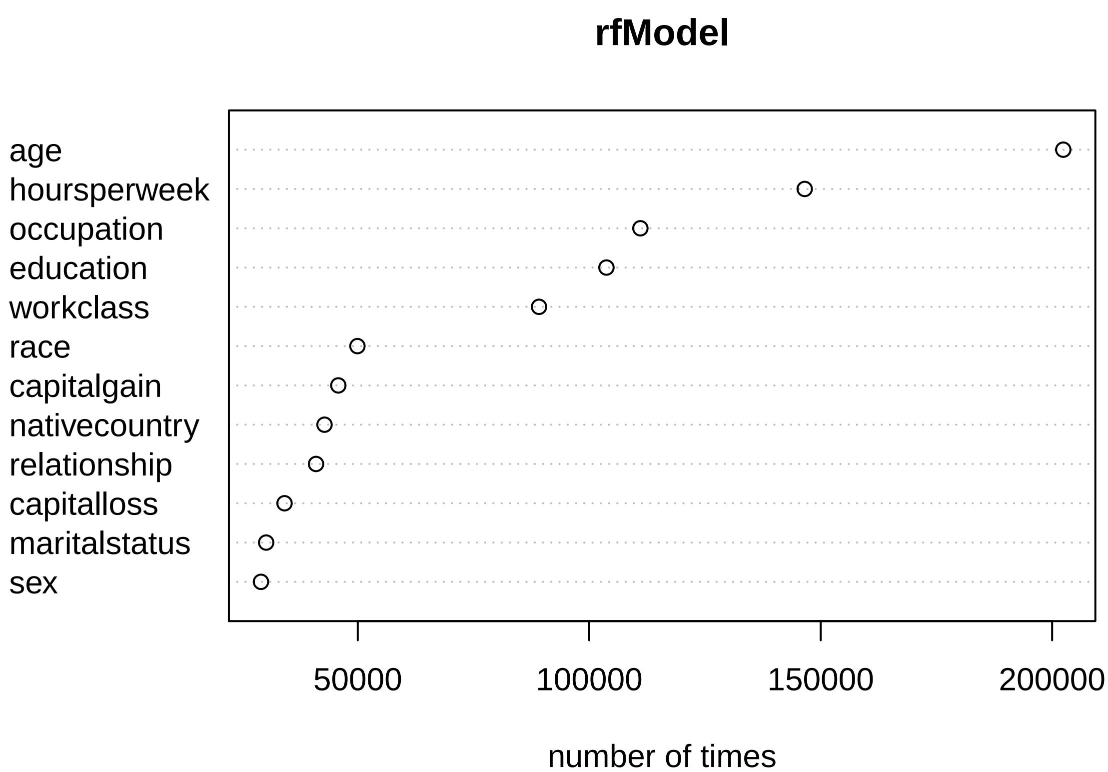
Building such a model is a bit time-consuming, but it can still be constructed within reasonable time on a personal computer. However, the random forests method actually decreases the accuracy.

```r
rfTable = table(test$over50k, rfPred)
(rfTable[[1]] + rfTable[[4]]) / nrow(test)
```

```
## [1] 0.8245779
```

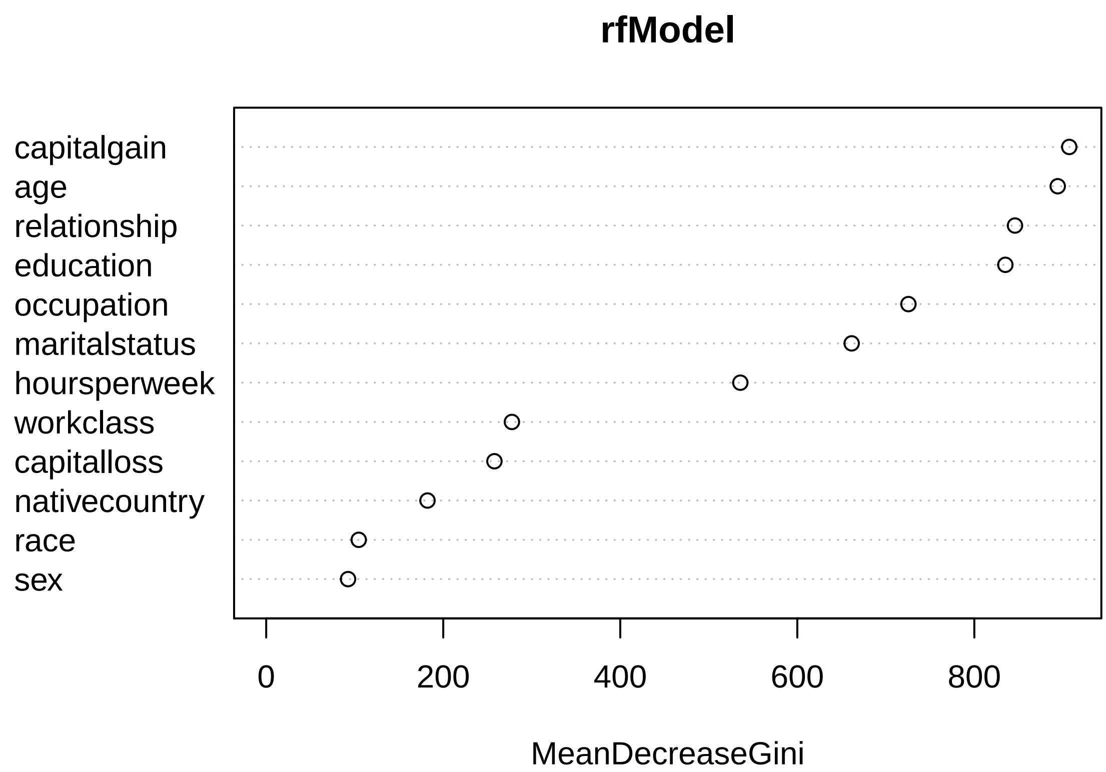
It is even worse that we lose some of the interpretability that comes with CART – seeing how predictions are made and which variables are important. Therefore, we need to calculate some metrics to tell us which variables are more important. For instance, we consider the number of times that a certain variable is selected for a split in all trees used in the random forest model.

```r
var_used = varUsed(rfModel, count=TRUE)
var_used_sorted = sort(var_used, decreasing = FALSE, index.return = TRUE)
dotchart(var_used_sorted$x, names(rfModel$forest$xlevels[var_used_sorted$ix]),
         xlab="number of times", main="rfModel")
```



It looks like `age` is the most important predictor of a person's employment status, which is consistent with common sense. Other important predictors include `hoursperweek`, `occupation`, `education` and `workclass`, which are somewhat different from the CART model. Then, we use another metric for comparison – mean decrease in inhomogeneity. In each tree in the forest, whenever we select a variable and perform a split, the inhomogeneity is reduced.

```r
varImpPlot(rfModel)
```



Four most important predictors are `captialgain`, `age`, `relationship` and `education`. We may use other metrics as well since there are no intuitive methods like the CART plot.

4 Model Improvements

4.1 Logistic Regression

```r
logModelImp = glm(over50k ~ .-nativecountry, data = train, family = "binomial")
summary(logModelImp)
```

```
##
## Call:
## glm(formula = over50k ~ . - nativecountry, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3656  -0.5158  -0.1882  -0.0289   3.6484
```

```
##
## Coefficients: (1 not defined because of singularities)
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          -8.557e+00  5.035e-01 -16.996  < 2e-16
## age                                   2.509e-02  1.963e-03  12.782  < 2e-16
## workclass Federal-gov                 9.798e-01  1.862e-01   5.263 1.42e-07
## workclass Local-gov                   2.746e-01  1.704e-01   1.612 0.107017
## workclass Never-worked               -1.104e+01  3.587e+02  -0.031 0.975437
## workclass Private                     5.415e-01  1.520e-01   3.562 0.000367
## workclass Self-emp-inc                6.729e-01  1.812e-01   3.713 0.000205
## workclass Self-emp-not-inc            9.837e-02  1.661e-01   0.592 0.553790
## workclass State-gov                   1.605e-01  1.848e-01   0.869 0.385069
## workclass Without-pay                -1.215e+01  3.002e+02  -0.040 0.967701
## education 11th                       -3.950e-02  2.508e-01  -0.157 0.874859
## education 12th                        4.098e-01  3.186e-01   1.286 0.198324
## education 1st-4th                    -1.201e+00  6.750e-01  -1.779 0.075176
## education 5th-6th                    -5.834e-01  3.822e-01  -1.527 0.126849
## education 7th-8th                    -6.977e-01  2.830e-01  -2.466 0.013671
## education 9th                        -5.011e-01  3.205e-01  -1.563 0.117987
## education Assoc-acdm                  1.096e+00  2.128e-01   5.150 2.60e-07
## education Assoc-voc                   1.275e+00  2.037e-01   6.258 3.89e-10
## education Bachelors                   1.834e+00  1.902e-01   9.643  < 2e-16
## education Doctorate                   2.905e+00  2.620e-01  11.086  < 2e-16
## education HS-grad                     7.270e-01  1.851e-01   3.926 8.63e-05
## education Masters                     2.233e+00  2.032e-01  10.985  < 2e-16
## education Preschool                  -2.004e+01  1.331e+02  -0.151 0.880352
## education Prof-school                 2.632e+00  2.387e-01  11.028  < 2e-16
## education Some-college                1.017e+00  1.878e-01   5.415 6.14e-08
## maritalstatus Married-AF-spouse       3.505e+00  6.792e-01   5.160 2.47e-07
## maritalstatus Married-civ-spouse      2.073e+00  3.121e-01   6.641 3.11e-11
## maritalstatus Married-spouse-absent -2.542e-01  2.893e-01  -0.879 0.379544
## maritalstatus Never-married          -5.271e-01  1.028e-01  -5.128 2.93e-07
## maritalstatus Separated              -1.582e-01  1.994e-01  -0.794 0.427371
## maritalstatus Widowed                 1.950e-02  1.840e-01   0.106 0.915581
## occupation Adm-clerical               1.485e-01  1.187e-01   1.251 0.210919
## occupation Armed-Forces              -9.020e-01  1.510e+00  -0.597 0.550270
## occupation Craft-repair               1.992e-01  1.020e-01   1.952 0.050909
## occupation Exec-managerial            9.322e-01  1.047e-01   8.904  < 2e-16
## occupation Farming-fishing           -8.924e-01  1.688e-01  -5.287 1.25e-07
## occupation Handlers-cleaners         -4.585e-01  1.717e-01  -2.671 0.007571
## occupation Machine-op-inspct         -1.082e-01  1.259e-01  -0.859 0.390130
## occupation Other-service             -7.038e-01  1.503e-01  -4.684 2.81e-06
## occupation Priv-house-serv           -3.462e+00  2.043e+00  -1.695 0.090094
## occupation Prof-specialty             6.385e-01  1.126e-01   5.670 1.43e-08
## occupation Protective-serv            7.934e-01  1.568e-01   5.061 4.16e-07
## occupation Sales                      4.328e-01  1.082e-01   3.999 6.36e-05
## occupation Tech-support               7.902e-01  1.429e-01   5.531 3.18e-08
## occupation Transport-moving                  NA         NA      NA       NA
```

```
## relationship Not-in-family           5.214e-01  3.086e-01   1.690 0.091035
## relationship Other-relative         -2.391e-01  2.865e-01  -0.835 0.403843
## relationship Own-child              -8.234e-01  3.079e-01  -2.674 0.007501
## relationship Unmarried               3.380e-01  3.282e-01   1.030 0.303087
## relationship Wife                    1.324e+00  1.223e-01  10.823  < 2e-16
## race Asian-Pac-Islander              8.061e-01  3.046e-01   2.646 0.008136
## race Black                           7.104e-01  2.918e-01   2.435 0.014905
## race Other                           2.635e-01  4.432e-01   0.594 0.552210
## race White                           8.495e-01  2.798e-01   3.037 0.002393
## sex Male                             8.251e-01  9.346e-02   8.828  < 2e-16
## capitalgain                          3.208e-04  1.230e-05  26.091  < 2e-16
## capitalloss                          6.137e-04  4.446e-05  13.803  < 2e-16
## hoursperweek                         3.025e-02  1.950e-03  15.515  < 2e-16
##
## (Intercept)                         ***
## age                                 ***
## workclass Federal-gov               ***
## workclass Local-gov
## workclass Never-worked
## workclass Private                   ***
## workclass Self-emp-inc              ***
## workclass Self-emp-not-inc
## workclass State-gov
## workclass Without-pay
## education 11th
## education 12th
## education 1st-4th                   .
## education 5th-6th
## education 7th-8th                   *
## education 9th
## education Assoc-acdm                ***
## education Assoc-voc                 ***
## education Bachelors                 ***
## education Doctorate                 ***
## education HS-grad                   ***
## education Masters                   ***
## education Preschool
## education Prof-school               ***
## education Some-college              ***
## maritalstatus Married-AF-spouse     ***
## maritalstatus Married-civ-spouse    ***
## maritalstatus Married-spouse-absent
## maritalstatus Never-married         ***
## maritalstatus Separated
## maritalstatus Widowed
## occupation Adm-clerical
## occupation Armed-Forces
## occupation Craft-repair             .
```

```
## occupation Exec-managerial          ***
## occupation Farming-fishing          ***
## occupation Handlers-cleaners        **
## occupation Machine-op-inspct
## occupation Other-service            ***
## occupation Priv-house-serv          .
## occupation Prof-specialty           ***
## occupation Protective-serv          ***
## occupation Sales                    ***
## occupation Tech-support             ***
## occupation Transport-moving
## relationship Not-in-family          .
## relationship Other-relative
## relationship Own-child              **
## relationship Unmarried
## relationship Wife                   ***
## race Asian-Pac-Islander             **
## race Black                          *
## race Other
## race White                          **
## sex Male                            ***
## capitalgain                         ***
## capitalloss                         ***
## hoursperweek                        ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24703  on 22383  degrees of freedom
## Residual deviance: 14286  on 22327  degrees of freedom
## AIC: 14400
##
## Number of Fisher Scoring iterations: 13
```

```
logPredImp = predict(logModelImp, newdata = test, type = "response")
logTableImp = table(test$over50k, logPredImp > 0.5)
(logTableImp[[1]] + logTableImp[[4]]) / nrow(test)
```

```
## [1] 0.8518866
```

We attempt to remove some predictors seem to be less significant, expecting it may reduce overfitting of the training set. However, it doesn't work and only decreases the accuracy.

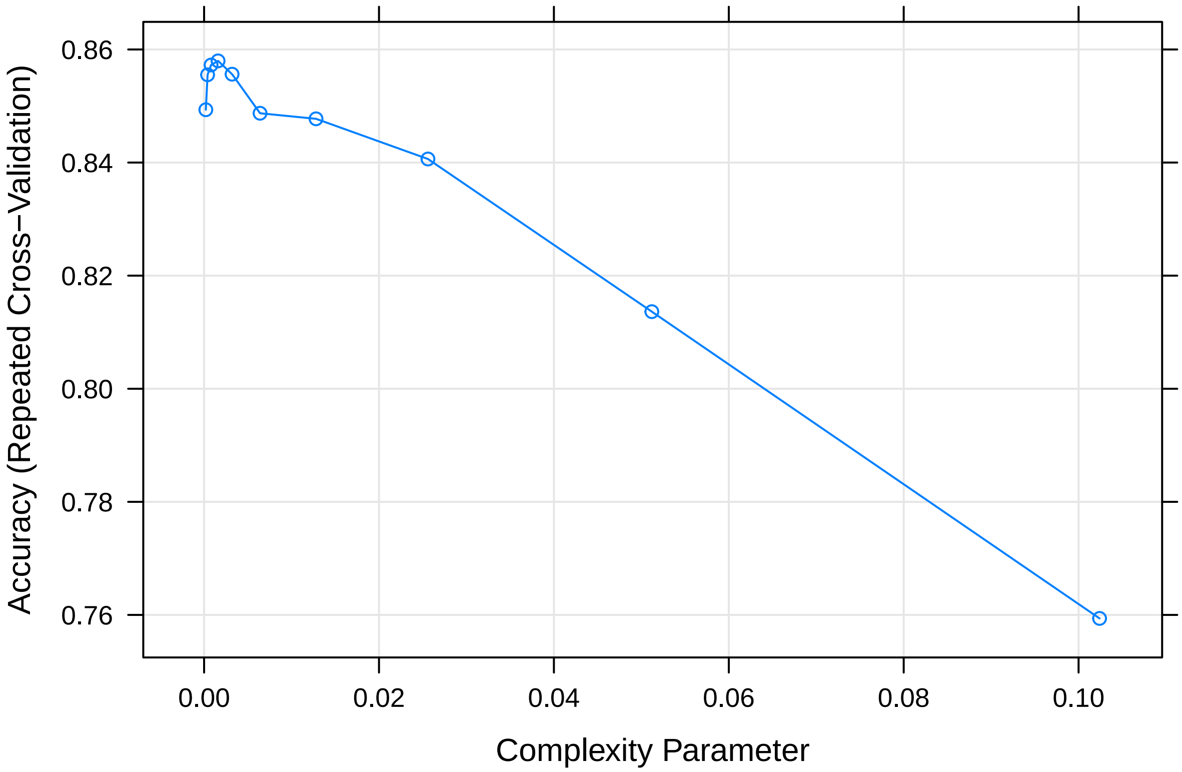
4.2 Classification and Regression Trees

As discussed above, simply applying more advanced methods to dataset can't guarantee improvement of model performance. Therefore, we need to carefully tune some parameters to make our models better than simple logistic regression. Now, we use k-fold cross-validation ($k = 10$) to find a optimal complexity parameter (`cp`) value for the CART model.

```
library(caret)
library(e1071)
set.seed(1111)
cp.grid = expand.grid( .cp = 2^seq(1, 10) * 0.0001)
tr.control = trainControl(method = "repeatedcv", number = 10, repeats = 3)
CARTCV = train(over50k ~ ., data = train, method = "rpart",
               trControl = tr.control, tuneGrid = cp.grid)
CARTCV
```

```
## CART
##
## 22384 samples
##    12 predictor
##     2 classes: ' <=50K', ' >50K'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 20146, 20146, 20146, 20145, 20145, 20146, ...
## Resampling results across tuning parameters:
##
##   cp      Accuracy   Kappa      Accuracy SD   Kappa SD
##   0.0002  0.8493419  0.5643198  0.0056554763  0.01626467
##   0.0004  0.8555217  0.5786250  0.0052746709  0.01438452
##   0.0008  0.8572491  0.5709688  0.0050204268  0.01494491
##   0.0016  0.8579789  0.5740501  0.0054882017  0.01649817
##   0.0032  0.8556262  0.5688212  0.0059999792  0.01805419
##   0.0064  0.8487165  0.5437220  0.0055745749  0.01799951
##   0.0128  0.8477336  0.5418085  0.0060279328  0.01837313
##   0.0256  0.8406303  0.5076450  0.0062506692  0.01895452
##   0.0512  0.8136320  0.3492579  0.0070599750  0.06140672
##   0.1024  0.7593817  0.0000000  0.0001911321  0.00000000
##
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was cp = 0.0016.
```

```
plot(CARTCV)
```



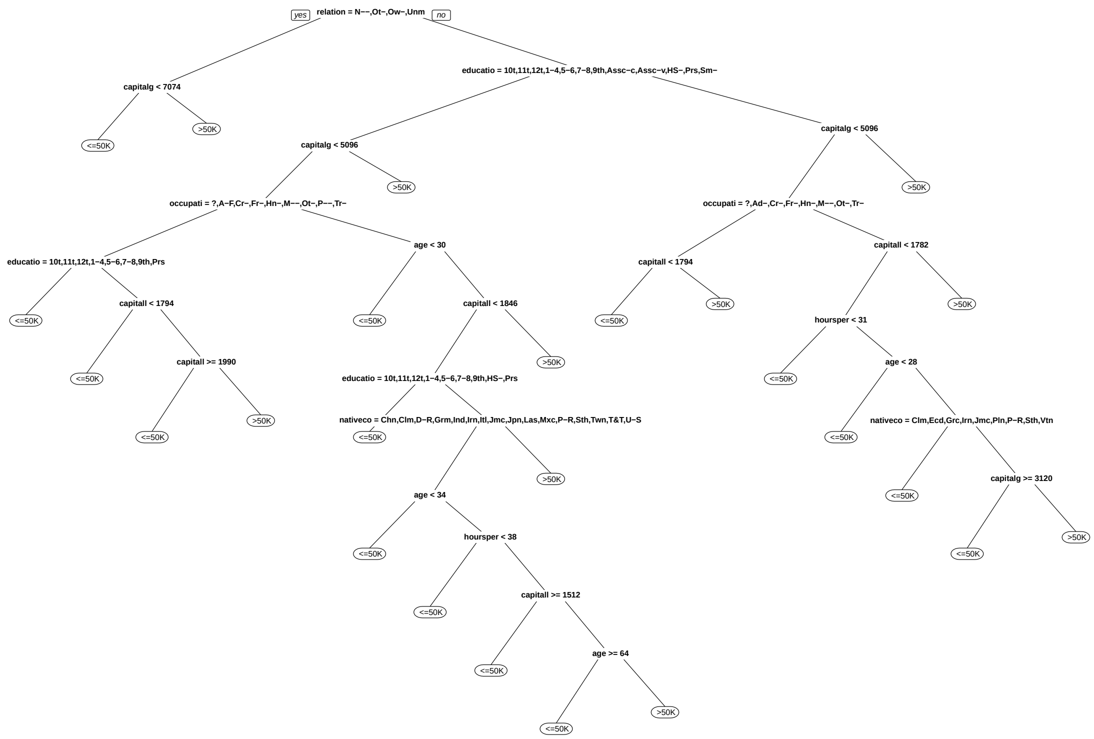
From the plot, We found 0.0016 seems to be the best `cp` value. Then we use this `cp` value to build a CART model and make predictions:

```
CARTmodelCV = rpart(over50k ~ ., data = train, method = "class", cp = 0.0016)
CARTpredCV = predict(CARTmodelCV, newdata = test, type = "class")
CARTtableCV = table(test$over50k, CARTpredCV)
(CARTtableCV[[1]] + CARTtableCV[[4]]) / nrow(test)
```

```
## [1] 0.8622055
```

After tuning the complex parameter, the CART model has been improved by nearly 2% in accuracy, and becomes 1% better than the logistic regression model. However, it comes with a price – the complexity of the tree increases significantly and become harder to interpret. It means we may still prefer the less accurate but simpler and more interpretable model.

```
prp(CARTmodelCV)
```



4.3 Random forests

Then, we attempt to improve the random forests model. Unfortunately, cross-validation for random forests on the entire training set takes impratically long time on a personal computer. As a result, we have to pick a random sample from the training set for our cross-validation purpose. For random forests model, we want to find a optimal number of randomly selected predictors (`mtry`). Unfortunately, it means the optimal value of `mtry` may not be generalized to the entire dataset, and we have to modify it again later.

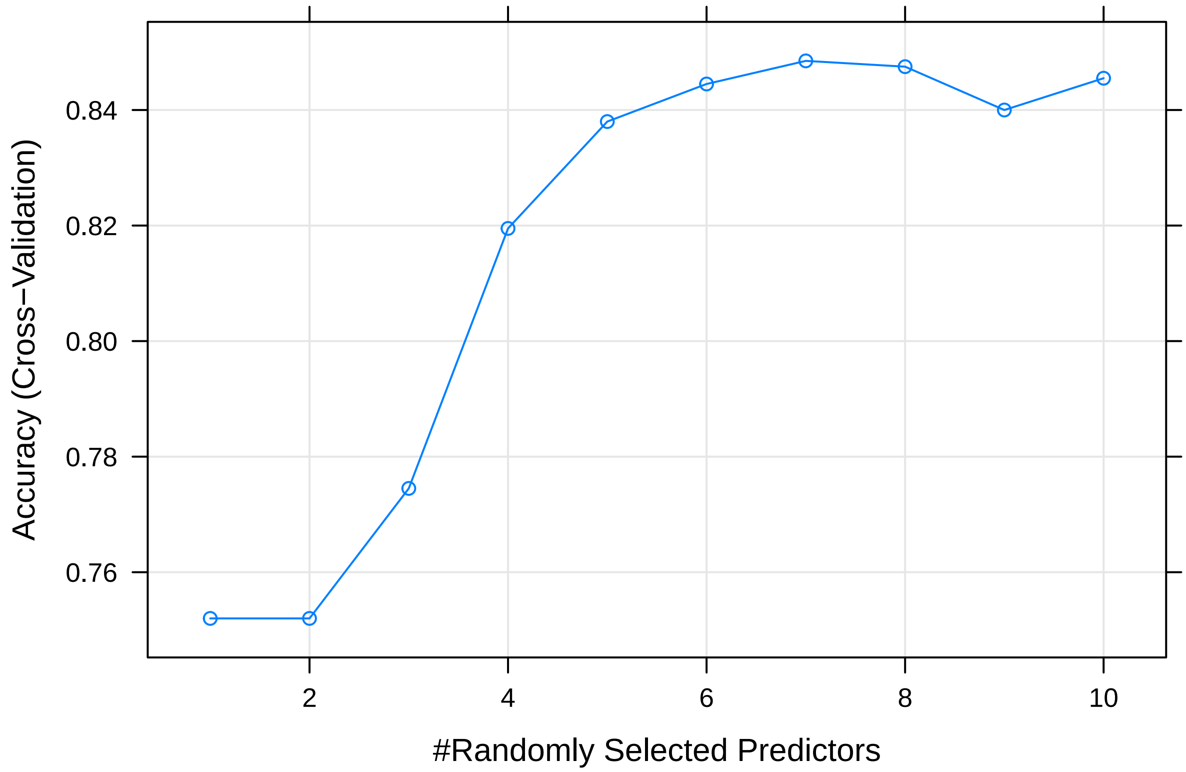
```
set.seed(3333)
train2000 = train[sample(nrow(train), 2000), ]

mtry.grid = expand.grid( .mtry = seq(1, 10))
tr.control = trainControl(method = "cv", number = 2)
set.seed(3333)
rfCV = train(over50k ~ ., data = train2000, method = "rf",
               trControl = tr.control, tuneGrid = mtry.grid)
rfCV
```

```
## Random Forest
##
## 2000 samples
```

```
##   12 predictor
##    2 classes: ' <=50K', ' >50K'
##
## No pre-processing
## Resampling: Cross-Validated (2 fold)
## Summary of sample sizes: 1000, 1000
## Resampling results across tuning parameters:
##
##   mtry  Accuracy  Kappa      Accuracy SD   Kappa SD
##    1    0.7520    0.0000000  0.0000000000  0.000000000
##    2    0.7520    0.0000000  0.0000000000  0.000000000
##    3    0.7745    0.1388048  0.0007071068  0.007848171
##    4    0.8195    0.4001201  0.0035355339  0.032324840
##    5    0.8380    0.4895181  0.0000000000  0.018332892
##    6    0.8445    0.5243431  0.0049497475  0.006326510
##    7    0.8485    0.5470980  0.0063639610  0.008893883
##    8    0.8475    0.5455202  0.0077781746  0.009185117
##    9    0.8400    0.5258814  0.0000000000  0.018011859
##   10    0.8455    0.5454694  0.0035355339  0.005323897
##
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was mtry = 7.
```

```r
plot(rfCV)
```



From the plot, We found 7 seems to be the best `mtry` value. Then we use this `mtry` value to build a random forests model and make predictions.

```
set.seed(3333)
rfModelCV = randomForest(over50k ~ ., data = train, mtry = 7)
rfPredCV = predict(rfModelCV, newdata = test)
rfTableCV = table(test$over50k, rfPredCV)
(rfTableCV[[1]] + rfTableCV[[4]]) / nrow(test)
```

```
## [1] 0.8234313
```

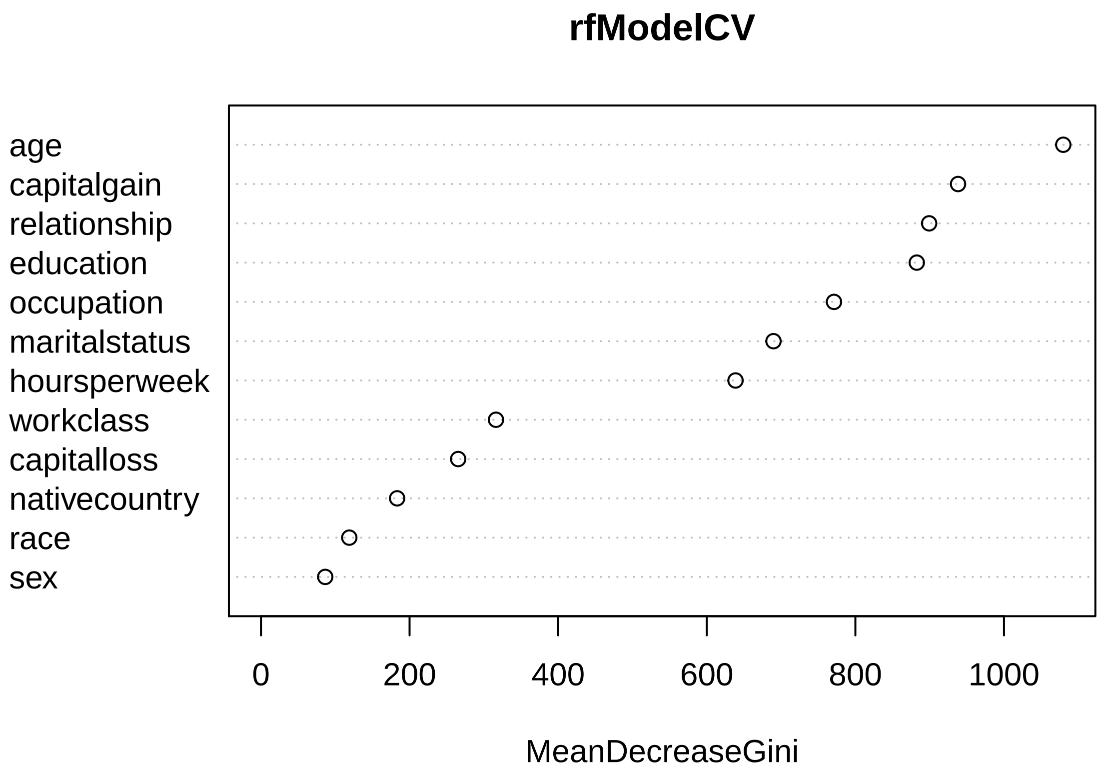
However, the result is worse than default (`mtry` = 3). The problem comes from the fact that we only use a small fraction of data from training set to tune our model. Therefore, we try to build models with `mtry` = 1-10 to check the model's performance. The result is `mtry` = 4 or 5 will improve the perfomance, while 4 is the optimal value.

```
set.seed(3333)
rfModelCV = randomForest(over50k ~ ., data = train, mtry = 4)
rfPredCV = predict(rfModelCV, newdata = test)
rfTableCV = table(test$over50k, rfPredCV)
(rfTableCV[[1]] + rfTableCV[[4]]) / nrow(test)
```

```
## [1] 0.8249948
```

This is only slightly better than the original random forests model before tuning, but still worse than the logistic regression model. The random forests model is actually good at finding appropriate parameters as the default values. Now look at the predictors involved in this model.

```r
varImpPlot(rfModelCV)
```



The most significant predictors are still `captialgain`, `age`, `relationship` and `education`, although the sort has changed a bit.

5 Conclusion

The following table summarizes the major results of this project:

Model	Initial Accuracy	Improved Accuracy
Logistic Regression	0.8528247	N/A
Classification and Regression Trees	0.8475089	0.8622055
Random Forests	0.8245779	0.8249948

Three models have quite similar accuracy, and it is extremely difficult to further improve them when the accuracy is already quite high. When choosing suitable models for a specific problem, we actually not only consider accuracy but also other aspects such as complexity and interpretability.

6 References

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. New York: Springer.

7 Appendix

The form and structure of this dataset are shown below:

```
head(census, 20)
```

```
##    age        workclass    education        maritalstatus
## 1   39        State-gov    Bachelors        Never-married
## 2   50 Self-emp-not-inc    Bachelors   Married-civ-spouse
## 3   38          Private      HS-grad             Divorced
## 4   53          Private         11th   Married-civ-spouse
## 5   28          Private    Bachelors   Married-civ-spouse
## 6   37          Private      Masters   Married-civ-spouse
## 7   49          Private          9th Married-spouse-absent
## 8   52 Self-emp-not-inc      HS-grad   Married-civ-spouse
## 9   31          Private      Masters        Never-married
## 10  42          Private    Bachelors   Married-civ-spouse
## 11  37          Private Some-college   Married-civ-spouse
## 12  30        State-gov    Bachelors   Married-civ-spouse
## 13  23          Private    Bachelors        Never-married
## 14  32          Private   Assoc-acdm        Never-married
## 15  34          Private      7th-8th   Married-civ-spouse
## 16  25 Self-emp-not-inc      HS-grad        Never-married
## 17  32          Private      HS-grad        Never-married
## 18  38          Private         11th   Married-civ-spouse
## 19  43 Self-emp-not-inc      Masters             Divorced
## 20  40          Private    Doctorate   Married-civ-spouse
##            occupation   relationship                race    sex
## 1        Adm-clerical  Not-in-family               White   Male
## 2     Exec-managerial        Husband               White   Male
## 3   Handlers-cleaners  Not-in-family               White   Male
## 4   Handlers-cleaners        Husband               Black   Male
## 5      Prof-specialty           Wife               Black Female
## 6     Exec-managerial           Wife               White Female
## 7       Other-service  Not-in-family               Black Female
## 8     Exec-managerial        Husband               White   Male
## 9      Prof-specialty  Not-in-family               White Female
## 10    Exec-managerial        Husband               White   Male
## 11    Exec-managerial        Husband               Black   Male
```

```
## 12     Prof-specialty        Husband  Asian-Pac-Islander    Male
## 13       Adm-clerical      Own-child               White  Female
## 14              Sales  Not-in-family               Black    Male
## 15   Transport-moving        Husband  Amer-Indian-Eskimo    Male
## 16    Farming-fishing      Own-child               White    Male
## 17  Machine-op-inspct      Unmarried               White    Male
## 18              Sales        Husband               White    Male
## 19    Exec-managerial      Unmarried               White  Female
## 20     Prof-specialty        Husband               White    Male
##    capitalgain capitalloss hoursperweek  nativecountry over50k
## 1         2174           0           40  United-States   <=50K
## 2            0           0           13  United-States   <=50K
## 3            0           0           40  United-States   <=50K
## 4            0           0           40  United-States   <=50K
## 5            0           0           40           Cuba   <=50K
## 6            0           0           40  United-States   <=50K
## 7            0           0           16        Jamaica   <=50K
## 8            0           0           45  United-States    >50K
## 9        14084           0           50  United-States    >50K
## 10        5178           0           40  United-States    >50K
## 11           0           0           80  United-States    >50K
## 12           0           0           40          India    >50K
## 13           0           0           30  United-States   <=50K
## 14           0           0           50  United-States   <=50K
## 15           0           0           45         Mexico   <=50K
## 16           0           0           35  United-States   <=50K
## 17           0           0           40  United-States   <=50K
## 18           0           0           50  United-States   <=50K
## 19           0           0           45  United-States    >50K
## 20           0           0           60  United-States    >50K
```