HW2

Sheng Zhang

February 23, 2017

```
# Read in data
bank_full <- read.csv("./bank-additional-full.csv",header = TRUE,sep=";",stringsAsFactors = TRUE)
head(bank_full)
```

```
##   age       job marital   education default housing loan   contact month
## 1  56 housemaid married    basic.4y      no      no   no telephone   may
## 2  57  services married high.school unknown      no   no telephone   may
## 3  37  services married high.school      no     yes   no telephone   may
## 4  40    admin. married    basic.6y      no      no   no telephone   may
## 5  56  services married high.school      no      no  yes telephone   may
## 6  45  services married    basic.9y unknown      no   no telephone   may
##   day_of_week duration campaign pdays previous    poutcome emp.var.rate
## 1         mon      261        1   999        0 nonexistent          1.1
## 2         mon      149        1   999        0 nonexistent          1.1
## 3         mon      226        1   999        0 nonexistent          1.1
## 4         mon      151        1   999        0 nonexistent          1.1
## 5         mon      307        1   999        0 nonexistent          1.1
## 6         mon      198        1   999        0 nonexistent          1.1
##   cons.price.idx cons.conf.idx euribor3m nr.employed  y
## 1         93.994         -36.4     4.857        5191 no
## 2         93.994         -36.4     4.857        5191 no
## 3         93.994         -36.4     4.857        5191 no
## 4         93.994         -36.4     4.857        5191 no
## 5         93.994         -36.4     4.857        5191 no
## 6         93.994         -36.4     4.857        5191 no
```

```
# Delete columns that will not used
bank_full$duration <- NULL
bank_full$day_of_week <- NULL
bank_full$month <- NULL
bank_full$nr.employed <- NULL
summary(bank_full)
```

```
##       age                 job           marital
##  Min.   :17.00   admin.     :10422   divorced: 4612
##  1st Qu.:32.00   blue-collar: 9254   married :24928
##  Median :38.00   technician : 6743   single  :11568
##  Mean   :40.02   services   : 3969   unknown :   80
##  3rd Qu.:47.00   management : 2924
##  Max.   :98.00   retired    : 1720
##                  (Other)    : 6156
##                education        default         housing
##  university.degree  :12168   no     :32588   no     :18622
##  high.school        : 9515   unknown: 8597   unknown:  990
##  basic.9y           : 6045   yes    :    3   yes    :21576
##  professional.course: 5243
```

```
##  basic.4y           : 4176
##  basic.6y           : 2292
##  (Other)            : 1749
##       loan            contact         campaign          pdays
##  no     :33950   cellular :26144   Min.   : 1.000   Min.   :  0.0
##  unknown:  990   telephone:15044   1st Qu.: 1.000   1st Qu.:999.0
##  yes    : 6248                     Median : 2.000   Median :999.0
##                                    Mean   : 2.568   Mean   :962.5
##                                    3rd Qu.: 3.000   3rd Qu.:999.0
##                                    Max.   :56.000   Max.   :999.0
##
##     previous            poutcome      emp.var.rate      cons.price.idx
##  Min.   :0.000   failure    : 4252   Min.   :-3.40000   Min.   :92.20
##  1st Qu.:0.000   nonexistent:35563   1st Qu.:-1.80000   1st Qu.:93.08
##  Median :0.000   success    : 1373   Median : 1.10000   Median :93.75
##  Mean   :0.173                       Mean   : 0.08189   Mean   :93.58
##  3rd Qu.:0.000                       3rd Qu.: 1.40000   3rd Qu.:93.99
##  Max.   :7.000                       Max.   : 1.40000   Max.   :94.77
##
##  cons.conf.idx     euribor3m       y
##  Min.   :-50.8   Min.   :0.634   no :36548
##  1st Qu.:-42.7   1st Qu.:1.344   yes: 4640
##  Median :-41.8   Median :4.857
##  Mean   :-40.5   Mean   :3.621
##  3rd Qu.:-36.4   3rd Qu.:4.961
##  Max.   :-26.9   Max.   :5.045
##
```

1)

Removing duration makes sense because duration can be used to predict y "deterministically". specifically, when duration is 0, y is no. Thus, duration should not be included in a realistic predictive model for y.

Removing day of the week and month of the year might make sense if seasonality and weekday vs weekend distinction do not matter for prediction term deposit prediction.

Removing nr.employed might make sense because the number of employees in the economy may just be an indicator of economic performance, which is probably already captured by the other social and economic context variables.

There are some unknowns in the data which we might have to remove. In addition, there are multiple unordered categorical predictors which might not be ideal for tree methods, so we may consider transform those variables as well if we were to use tree methods for predicting y.

```
# Remove unknowns
bank_full[bank_full=="unknown"] <- NA
bank_full <- na.omit(bank_full)

# Substitute values for certain columns
summary(bank_full$job)
```

```
##        admin.   blue-collar  entrepreneur     housemaid    management
##          8737          5675          1089           690          2311
##       retired self-employed      services       student    technician
##          1216          1092          2857           610          5473
```

```
##    unemployed       unknown
##           738             0
```

```
# bank_full$job[bank_full$job!="unemployed" & bank_full$job!="retired"] <- "employed"
# bank_full$job[bank_full$job=="unemployed" | bank_full$job=="retired"] <- "unemployed"
bank_full$job <- as.factor(ifelse(bank_full$job=="unemployed" | bank_full$job=="retired" | bank_full$jo
summary(bank_full$job)
```

```
##   employed unemployed
##      27924       2564
```

```
summary(bank_full$marital)
```

```
## divorced  married   single  unknown
##     3553    17492     9443        0
```

```
bank_full$marital <- as.factor(ifelse(bank_full$marital=="married","married","single"))
summary(bank_full$marital)
```

```
## married  single
##   17492   12996
```

```
summary(bank_full$education)
```

```
##            basic.4y            basic.6y            basic.9y
##                2380                1389                4276
##         high.school          illiterate professional.course
##                7699                  11                4321
##   university.degree             unknown
##               10412                   0
```

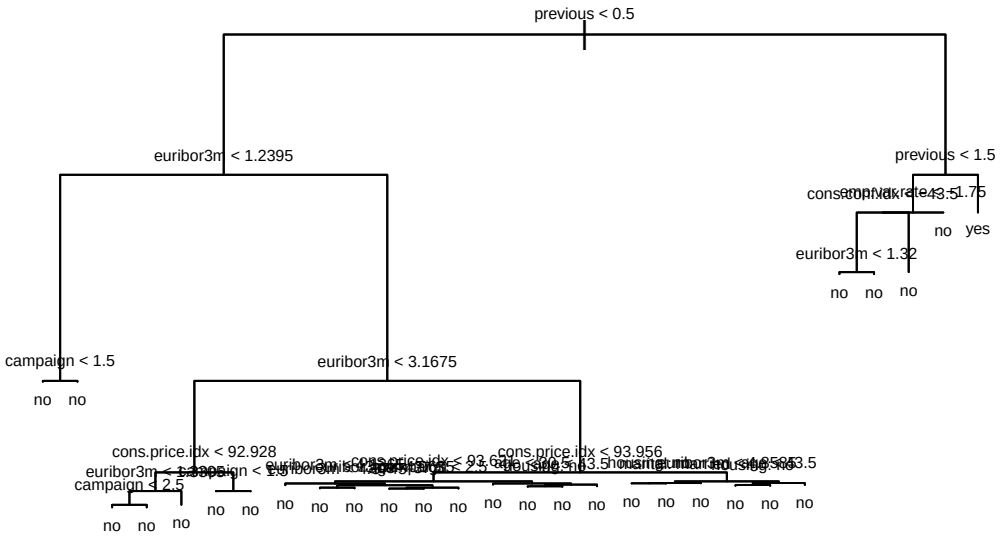
```
bank_full$education <- as.character(bank_full$education)
bank_full$education[bank_full$education=="illiterate"] <- "0"
bank_full$education[bank_full$education=="basic.4y"] <- "1"
bank_full$education[bank_full$education=="basic.6y"] <- "2"
bank_full$education[bank_full$education=="basic.9y"] <- "3"
bank_full$education[bank_full$education=="high.school"] <- "4"
bank_full$education[bank_full$education=="professional.course"] <- "5"
bank_full$education[bank_full$education=="university.degree"] <- "6"
bank_full$education <- as.factor(bank_full$education)
bank_full$education <- as.numeric(bank_full$education)
summary(bank_full$education)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   5.000   5.358   7.000   7.000
```

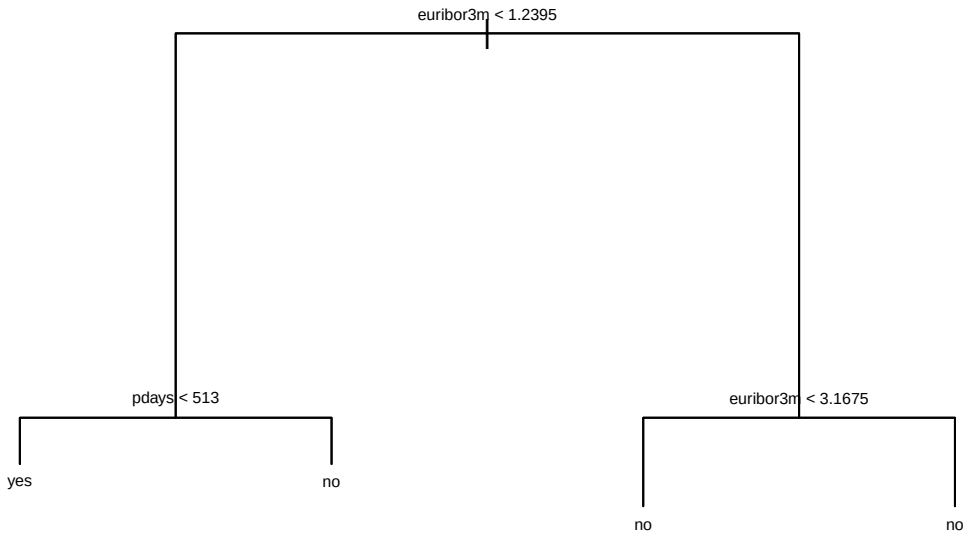
```
# Split into train and test
set.seed(1)
bank.train <- sample(1:nrow(bank_full), 0.5*nrow(bank_full))
bank.test <- bank_full[-bank.train, ]
```

```r
# Simple classification tree
library(tree)

# Gini
tree.bank.gini <- tree(y ~., data = bank_full, subset = bank.train, control = tree.control(nrow(bank_ful
tree.pred.gini <- predict(tree.bank.gini, bank.test, type="class")
gini.table <- table(tree.pred.gini, bank.test$y)
gini.accuracy <- (gini.table[1,1]+gini.table[2,2])/sum(gini.table)
plot(tree.bank.gini)
text(tree.bank.gini, pretty = 0, cex = .5)
```




```r
# Deviance
tree.bank.deviance <- tree(y ~., data = bank_full, subset = bank.train, split = "deviance")
tree.pred.deviance <- predict(tree.bank.deviance, bank.test, type="class")
deviance.table <- table(tree.pred.deviance, bank.test$y)
deviance.accuracy <- (deviance.table[1,1]+deviance.table[2,2])/sum(deviance.table)
plot(tree.bank.deviance)
text(tree.bank.deviance, pretty = 0, cex = .5)
```



4)

The tree I got using Gini has so many more terminal nodes than the tree I got using deviance.

```
library(randomForest)
```

```
## randomForest 4.6-12
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
library(MASS)
set.seed(2)
rf.bank <- randomForest(y~., data = bank_full, subset = bank.train, mtry = 4, importance = TRUE)
importance(rf.bank)
```

```
##                     no        yes MeanDecreaseAccuracy MeanDecreaseGini
## age           29.320546  -5.425068            26.019980     4.057715e+02
## job            9.651203   2.076427            10.482592     4.408761e+01
## marital       10.691601  -7.105241             6.405454     6.410288e+01
## education      8.734801   3.192652             9.311163     1.755783e+02
## default        0.000000   0.000000             0.000000     1.696527e-03
## housing        2.940522  -2.328176             1.459925     7.128982e+01
## loan           1.767770   1.727175             2.389939     5.475520e+01
## contact        9.206794  24.915906            12.024646     4.827992e+01
## campaign       9.481725   5.790106            11.499479     1.849972e+02
## pdays         12.737321  29.218846            26.095728     1.909901e+02
```

```
## previous        8.184693  -2.790429             7.515905     6.143526e+01
## poutcome       12.452690  10.806103            16.349812     1.200436e+02
## emp.var.rate   28.070872  10.409878            30.125241     1.429812e+02
## cons.price.idx 27.399117 -12.260905            27.846100     1.270561e+02
## cons.conf.idx  26.763138  -6.632324            27.751627     1.481308e+02
## euribor3m      42.050462  10.322752            47.669089     5.800451e+02
```

```
rf.pred <- predict(rf.bank, bank.test, type="class")
rf.table <- table(rf.pred, bank.test$y)
rf.accuracy <- (rf.table[1,1]+rf.table[2,2])/sum(rf.table)
```

```
# install.packages("adabag")
# install.packages("colorspace")
library(adabag)
```

```
## Loading required package: rpart
```

```
## Loading required package: mlbench
```

```
## Loading required package: caret
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:randomForest':
##
##     margin
```

```
boost.bank <- boosting(y ~., bank_full[bank.train,])
boost.bank$importance
```

```
##           age      campaign  cons.conf.idx cons.price.idx       contact
##    0.369749529    0.406105389   12.684632057    1.744921161    0.918883212
##        default      education   emp.var.rate      euribor3m        housing
##    0.000000000    0.089103240    1.674281750   74.148847082    0.060075483
##            job           loan        marital          pdays       poutcome
##    0.077167508    0.005823230    0.003168133    6.975456659    0.727641038
##       previous
##    0.114144530
```

```
boost.pred <- predict(boost.bank, bank.test)
boost.accuracy <- (boost.pred$confusion[1,1]+boost.pred$confusion[2,2])/sum(boost.pred$confusion)
```

```
# Check accuracy
gini.accuracy
```

```
## [1] 0.8752952
```

```
deviance.accuracy
```

```
## [1] 0.8868407
```

```
rf.accuracy
```

```
## [1] 0.8825768
```

```
boost.accuracy
```

```
## [1] 0.8883495
```

7)

It seems that the prediction accuracy is ranked as follows: Boosting > Deviance > Random Forest > Gini

The importance graphs of random forest and boosting both suggest that the most important independent variable is the euro libor rate, which is an indication of the interest rate in the economy. This indicates that interest rate is probably the most important determinant of term deposit subscription decisions.

However, since the y in our dataset contains predominantly "no"s, the prediction accuracy for "yes" is actually really poor. Moreover, we might consider accounting for heterogeneity in our dataset in future models.