# STAT 652 - Assignment 2

Dhruv Patel, 301471961

## Question 3

3. Use boosting to model the relationship between Ozone and all ONLY THE THREE ORIGINAL VARIABLES. Tune on an initial grid of shrink = 0.001, 0.005, 0.025, 0.125 and d = 2, 4, 6, and select trees optimally using twice the number suggested by OOB error. Use two reps of 5-fold CV (refer to the lecture note and R code to understand how to do this).

```
# Helper Functions
get.MSPE = function(Y, Y.hat){
  return(mean((Y - Y.hat)^2))
}

# Create k CV folds for a AQset of size n
get.folds = function(n, K) {
  ### Get the appropriate number of fold labels
  n.fold = ceiling(n / K) # Number of observations per fold (rounded up)
  fold.ids.raw = rep(1:K, times = n.fold) # Generate extra labels
  fold.ids = fold.ids.raw[1:n] # Keep only the correct number of labels
  # Shuffle the fold labels
  folds.rand = fold.ids[sample.int(n)]
  return(folds.rand)
}
```

3. (a) Report the mean root-MSPE for each combination of shrink and depth?
Answer: Root-MSPEs for each combination of shrink and depth are:
2|0.001 2|0.005 2|0.025 2|0.125 4|0.001 4|0.005 4|0.025 4|0.125 6|0.001 6|0.005 6|0.025 6|0.125
30.30477 23.41988 18.68153 18.84860 30.03603 22.63194 18.76087 18.93884 30.03713 22.61434 18.59552 18.94697
and mean of root-MSPEs for each combination of shrink and depth is 22.65137.

```
library(gbm)

set.seed(301471961)
AQ = airquality[1:4]

# Removing Null values
AQ = na.omit(AQ)

# Setting parameter values
trees = 200
shrink = c(0.001,0.005,0.025,0.125)
depth = c(2,4,6)
```

```
V=5
R=2
n = nrow(AQ)

# Create the folds and save in a matrix
folds = matrix(NA, nrow = n, ncol = R)
for(r in 1:R){
  folds[,r]=floor((sample.int(n)-1)*V/n) + 1
}

NS = length(shrink)
ND = length(depth)

gb.cv = matrix(NA, nrow=ND*NS, ncol=V*R)
opt.tree = matrix(NA, nrow=ND*NS, ncol=V*R)

qq = 1
for(r in 1:R){

  for(v in 1:V){

    AQ.train = AQ[folds[,r]!=v,]
    AQ.test = AQ[folds[,r]==v,]

    counter=1

    for(d in depth){

      for(s in shrink){

        AQ.gbm <- gbm(data = AQ.train, Ozone~., distribution = "gaussian",n.trees = trees,
                      interaction.depth = d, shrinkage = s,bag.fraction=0.8)

        treenum = min(trees, 2*gbm.perf(AQ.gbm, method = "OOB", plot.it = FALSE))

        opt.tree[counter,qq] = treenum

        preds = predict(AQ.gbm, newdata = AQ.test, n.trees=treenum)

        gb.cv[counter,qq] = mean((preds - AQ.test$Ozone)^2)

        counter=counter+1
      }
    }
    qq = qq+1
  }
}

parms = expand.grid(shrink,depth)
row.names(gb.cv) = paste(parms[,2], parms[,1], sep="|")
row.names(opt.tree) = paste(parms[,2], parms[,1], sep="|")

(mean.tree = apply(opt.tree, 1, mean))
```

```
## 2|0.001 2|0.005 2|0.025 2|0.125 4|0.001 4|0.005 4|0.025 4|0.125 6|0.001 6|0.005
##   200.0   200.0   172.8    32.6   200.0   200.0   149.8    29.6   200.0   200.0
## 6|0.025 6|0.125
##   158.8    29.8
```

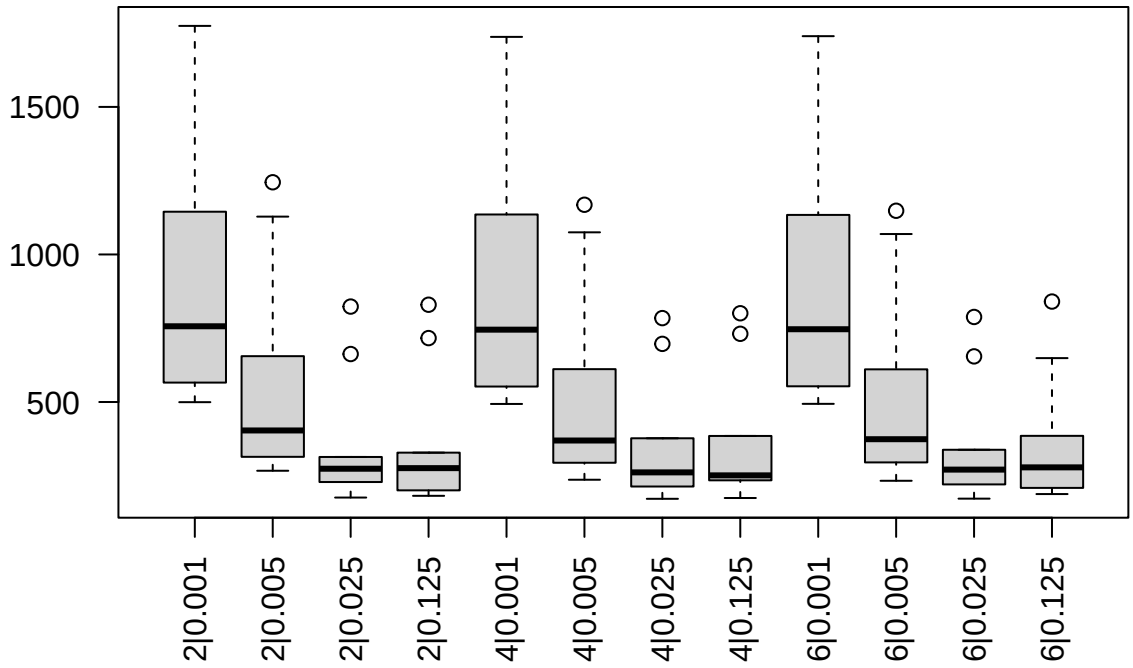
```
(sq.mean.cv = sqrt(apply(gb.cv, 1, mean)))
```

```
##  2|0.001  2|0.005  2|0.025  2|0.125  4|0.001  4|0.005  4|0.025  4|0.125
## 30.30477 23.41988 18.68153 18.84860 30.03603 22.63194 18.76087 18.93884
##  6|0.001  6|0.005  6|0.025  6|0.125
## 30.03713 22.61434 18.59552 18.94697
```

```
sq.mean.cv
```

```
##  2|0.001  2|0.005  2|0.025  2|0.125  4|0.001  4|0.005  4|0.025  4|0.125
## 30.30477 23.41988 18.68153 18.84860 30.03603 22.63194 18.76087 18.93884
##  6|0.001  6|0.005  6|0.025  6|0.125
## 30.03713 22.61434 18.59552 18.94697
```

```
boxplot(gb.cv, use.cols=FALSE, las=2)
```

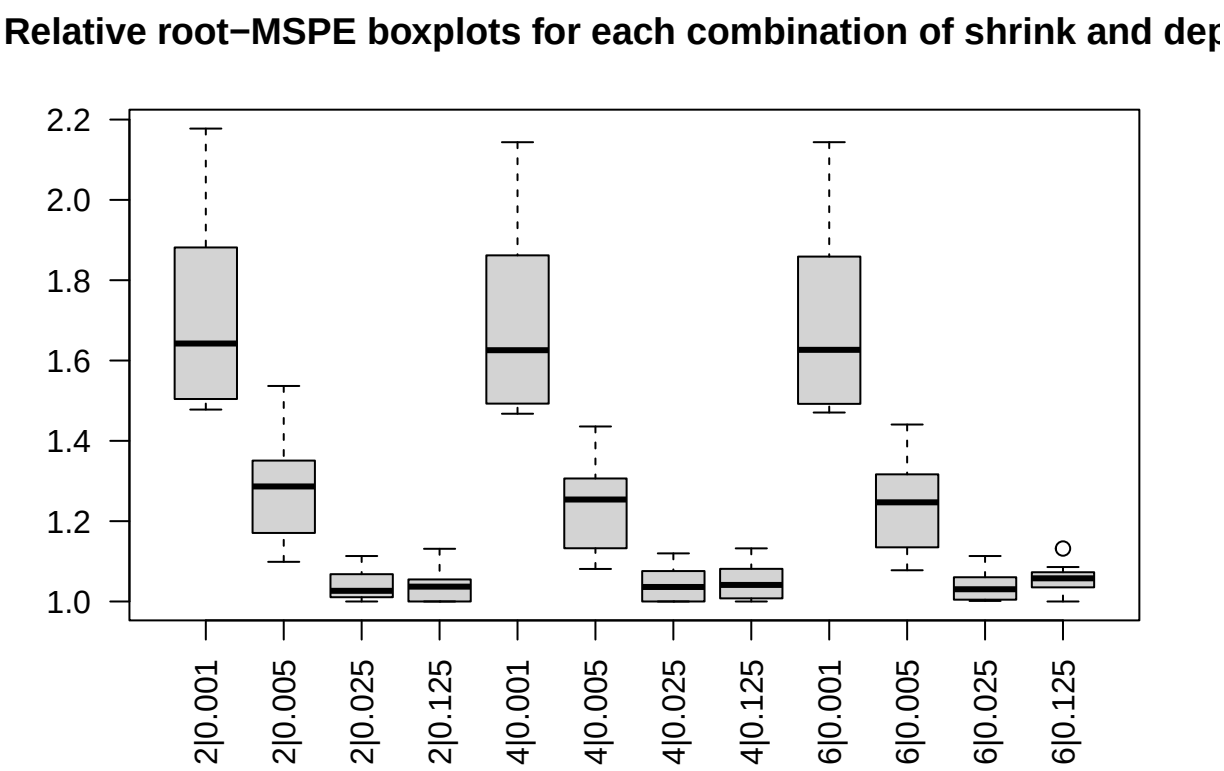



```
mean(sq.mean.cv)
```

```
## [1] 22.65137
```

3. (b) Show relative root-MSPE boxplots for each combination of shrink and depth?

```
# Get relative MSPEs and make boxplot
min.cv = apply(gb.cv, 2, min)
cv = sqrt(t(gb.cv)/min.cv)
boxplot(cv, use.cols=TRUE, las=2,
        main="Relative root-MSPE boxplots for each combination of shrink and depth")
```

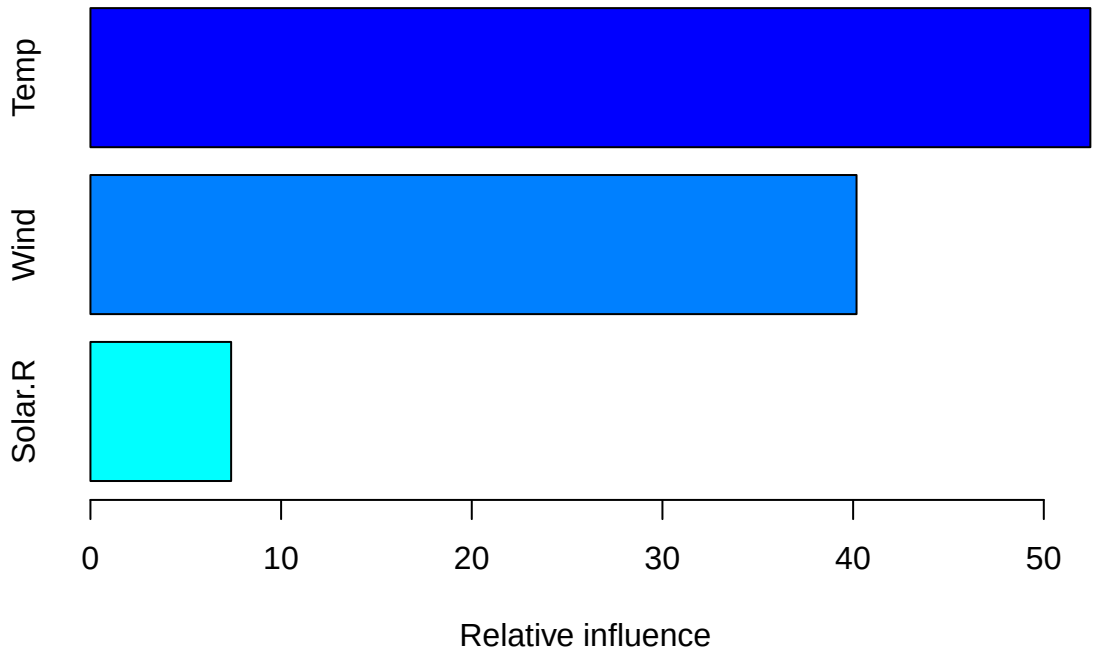



3. (c) What combination of shrink and depth do you prefer?
Answer: Based on the relative MSPE plot, shrinkage = 0.025 and depth = 6 looks like the best fitted model. And relatively, Temp is the most important among Wind and Solar.R.

```
set.seed(301471961)

fit.gbm.best = gbm(Ozone ~ ., data = AQ, distribution = "gaussian", n.trees = 125, interaction.depth = 

n.trees.best = gbm.perf(fit.gbm.best, plot.it = F) * 2

pred.gbm.best = predict(fit.gbm.best, AQ.test, n.trees.best)
MSPE.gbm = get.MSPE(AQ.test$Ozone, pred.gbm.best)
MSPE.gbm
```

```
## [1] 86.23719
```

```
summary(fit.gbm.best)
```




```
##             var   rel.inf
## Temp       Temp 52.442337
## Wind       Wind 40.179455
## Solar.R Solar.R  7.378207
```