Data\_setup

Rachel Schattman

February 19, 2019

# load libraries

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

## -------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.  
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:  
## library(plyr); library(dplyr)

## -------------------------------------------------------------------------

##   
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

## Warning: package 'randomForest' was built under R version 3.5.3

## randomForest 4.6-14

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

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':  
##   
## combine

## Warning: package 'Metrics' was built under R version 3.5.3

# Helpful resources:

## <https://www.r-bloggers.com/how-to-implement-random-forests-in-r/>

## <https://cran.r-project.org/web/packages/randomForest/randomForest.pdf>

## <https://rpubs.com/mbaumer/randomForest>

script\_path <- "C:/Users/rschattman/Documents/Research/RandomForestRMA/data"  
in\_dir <- "C:/Users/rschattman/Documents/Research/RandomForestRMA/data"  
out\_dir <- "C:/Users/rschattman/Documents/Research/RandomForestRMA/output/data"

# Read in data and combine into single dataframe

PAcip <- read.csv(file = "C:/Users/rschattman/Documents/Research/RandomForestRMA/data/monthly\_prcp\_PA.csv", header = TRUE, sep = ",")  
PAloss <- read.csv(file = "C:/Users/rschattman/Documents/Research/RandomForestRMA/data/PAannuallosses.csv", header = TRUE, sep = ",")  
PAbeta <-merge(PAcip, PAloss)

# Create new data frames with one dependent variable

head(PAbeta)

## Year month StateCollege\_PRCP Lebanon\_PRCP Selinsgrove\_PRCP State  
## 1 2001 1 3.39 6.69 3.71 PA  
## 2 2001 2 1.87 4.08 2.85 PA  
## 3 2001 3 10.70 8.51 12.10 PA  
## 4 2001 4 6.75 6.93 5.65 PA  
## 5 2001 5 3.51 9.61 4.78 PA  
## 6 2001 6 12.76 6.79 16.53 PA  
## WetDollars WetAcres nWet DryDollars DryAcres nDry  
## 1 129986 1700.12 57 14721696 216408 478  
## 2 129986 1700.12 57 14721696 216408 478  
## 3 129986 1700.12 57 14721696 216408 478  
## 4 129986 1700.12 57 14721696 216408 478  
## 5 129986 1700.12 57 14721696 216408 478  
## 6 129986 1700.12 57 14721696 216408 478

WetAcres <- PAbeta[,c("Year", "month", "StateCollege\_PRCP", "Lebanon\_PRCP", "Selinsgrove\_PRCP", "WetAcres")]  
WetDollars <- PAbeta[,c("Year", "month", "StateCollege\_PRCP", "Lebanon\_PRCP", "Selinsgrove\_PRCP", "WetDollars")]  
DryDollars <- PAbeta[,c("Year", "month", "StateCollege\_PRCP", "Lebanon\_PRCP", "Selinsgrove\_PRCP", "DryDollars")]  
DryAcres <- PAbeta[,c("Year", "month", "StateCollege\_PRCP", "Lebanon\_PRCP", "Selinsgrove\_PRCP", "DryAcres")]

# Review data

head(WetAcres)

## Year month StateCollege\_PRCP Lebanon\_PRCP Selinsgrove\_PRCP WetAcres  
## 1 2001 1 3.39 6.69 3.71 1700.12  
## 2 2001 2 1.87 4.08 2.85 1700.12  
## 3 2001 3 10.70 8.51 12.10 1700.12  
## 4 2001 4 6.75 6.93 5.65 1700.12  
## 5 2001 5 3.51 9.61 4.78 1700.12  
## 6 2001 6 12.76 6.79 16.53 1700.12

str(WetAcres)

## 'data.frame': 216 obs. of 6 variables:  
## $ Year : int 2001 2001 2001 2001 2001 2001 2001 2001 2001 2001 ...  
## $ month : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ StateCollege\_PRCP: num 3.39 1.87 10.7 6.75 3.51 ...  
## $ Lebanon\_PRCP : num 6.69 4.08 8.51 6.93 9.61 ...  
## $ Selinsgrove\_PRCP : num 3.71 2.85 12.1 5.65 4.78 ...  
## $ WetAcres : num 1700 1700 1700 1700 1700 ...

summary(WetAcres)

## Year month StateCollege\_PRCP Lebanon\_PRCP   
## Min. :2001 Min. : 1.00 Min. : 1.560 Min. : 0.910   
## 1st Qu.:2005 1st Qu.: 3.75 1st Qu.: 5.495 1st Qu.: 6.295   
## Median :2010 Median : 6.50 Median : 8.065 Median : 9.555   
## Mean :2010 Mean : 6.50 Mean : 8.953 Mean :10.134   
## 3rd Qu.:2014 3rd Qu.: 9.25 3rd Qu.:11.445 3rd Qu.:12.307   
## Max. :2018 Max. :12.00 Max. :26.840 Max. :45.690   
## Selinsgrove\_PRCP WetAcres   
## Min. : 0.000 Min. : 1700   
## 1st Qu.: 5.420 1st Qu.: 8658   
## Median : 7.985 Median : 54857   
## Mean : 9.064 Mean : 61888   
## 3rd Qu.:11.810 3rd Qu.: 98898   
## Max. :48.100 Max. :221234

# Split into trainning, validation, and test sets

set.seed(25)  
assignment <- sample(1:3, size = nrow(WetAcres), prob = c(0.7, 0.15, 0.15), replace = TRUE)  
  
Wettrain <- WetAcres[assignment == 1,]  
Wetvalid <- WetAcres[assignment == 2,]  
Wettest <- WetAcres[assignment == 3,]  
  
summary(Wettrain)

## Year month StateCollege\_PRCP Lebanon\_PRCP   
## Min. :2001 Min. : 1.000 Min. : 1.620 Min. : 0.910   
## 1st Qu.:2005 1st Qu.: 4.000 1st Qu.: 5.270 1st Qu.: 6.335   
## Median :2009 Median : 7.000 Median : 8.050 Median : 9.330   
## Mean :2009 Mean : 6.638 Mean : 8.831 Mean : 9.718   
## 3rd Qu.:2014 3rd Qu.:10.000 3rd Qu.:11.370 3rd Qu.:12.145   
## Max. :2018 Max. :12.000 Max. :26.840 Max. :39.150   
## Selinsgrove\_PRCP WetAcres   
## Min. : 2.060 Min. : 1700   
## 1st Qu.: 5.400 1st Qu.: 8658   
## Median : 7.920 Median : 53338   
## Mean : 8.842 Mean : 60893   
## 3rd Qu.:11.660 3rd Qu.: 98898   
## Max. :32.740 Max. :221234

summary(Wetvalid)

## Year month StateCollege\_PRCP Lebanon\_PRCP   
## Min. :2001 Min. : 1.000 Min. : 1.56 Min. : 1.190   
## 1st Qu.:2007 1st Qu.: 3.000 1st Qu.: 5.99 1st Qu.: 5.692   
## Median :2012 Median : 6.000 Median : 7.82 Median : 9.090   
## Mean :2011 Mean : 6.077 Mean : 9.17 Mean :11.616   
## 3rd Qu.:2015 3rd Qu.: 9.000 3rd Qu.:12.43 3rd Qu.:15.155   
## Max. :2018 Max. :12.000 Max. :22.57 Max. :45.690   
## Selinsgrove\_PRCP WetAcres   
## Min. : 0.000 Min. : 1700   
## 1st Qu.: 5.675 1st Qu.: 12630   
## Median : 8.850 Median : 54857   
## Mean :10.449 Mean : 61611   
## 3rd Qu.:12.215 3rd Qu.:100669   
## Max. :48.100 Max. :159070

summary(Wettest)

## Year month StateCollege\_PRCP Lebanon\_PRCP   
## Min. :2002 Min. : 1.000 Min. : 2.940 Min. : 2.31   
## 1st Qu.:2004 1st Qu.: 4.000 1st Qu.: 6.495 1st Qu.: 6.39   
## Median :2008 Median : 6.000 Median : 9.260 Median :10.04   
## Mean :2009 Mean : 6.074 Mean : 9.480 Mean :11.22   
## 3rd Qu.:2014 3rd Qu.: 8.000 3rd Qu.:10.905 3rd Qu.:12.82   
## Max. :2018 Max. :12.000 Max. :19.730 Max. :33.07   
## Selinsgrove\_PRCP WetAcres   
## Min. : 0.200 Min. : 5455   
## 1st Qu.: 5.165 1st Qu.: 12639   
## Median : 8.960 Median : 56377   
## Mean : 9.075 Mean : 68160   
## 3rd Qu.:11.850 3rd Qu.: 98898   
## Max. :26.700 Max. :221234

# Create Random Forest Model and test performance metrics

## Wet Acres

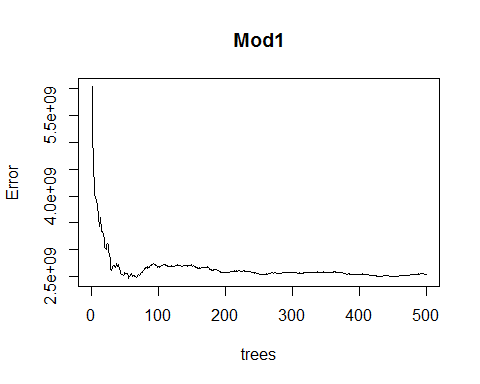
Mod1 <- randomForest(WetAcres ~ .,   
 data = Wettrain,   
 ntree = 500,   
 #method = "anova",   
 importance = TRUE)  
  
print(Mod1) # % of variance expalined is low. Tuning needed

##   
## Call:  
## randomForest(formula = WetAcres ~ ., data = Wettrain, ntree = 500, importance = TRUE)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 1  
##   
## Mean of squared residuals: 2537660723  
## % Var explained: 30.07

summary(Mod1)

## Length Class Mode   
## call 5 -none- call   
## type 1 -none- character  
## predicted 163 -none- numeric   
## mse 500 -none- numeric   
## rsq 500 -none- numeric   
## oob.times 163 -none- numeric   
## importance 10 -none- numeric   
## importanceSD 5 -none- numeric   
## localImportance 0 -none- NULL   
## proximity 0 -none- NULL   
## ntree 1 -none- numeric   
## mtry 1 -none- numeric   
## forest 11 -none- list   
## coefs 0 -none- NULL   
## y 163 -none- numeric   
## test 0 -none- NULL   
## inbag 0 -none- NULL   
## terms 3 terms call

plot(Mod1)

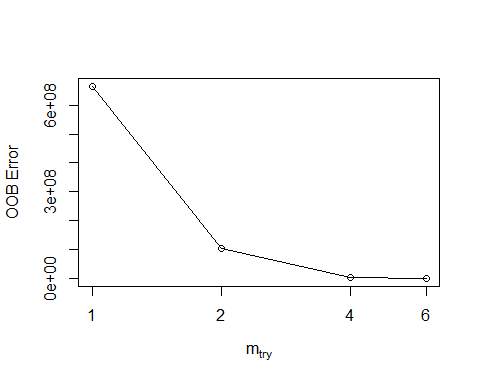


pred <- predict(object = Mod1, newdata = Wettest)  
RMSE\_Mod1 <- rmse(actual = Wettest$WetAcres, #actual values  
 predicted = pred) #predicted values  
print(RMSE\_Mod1/mean(Wettest$WetAcres)) #tells us the %of the mean represented by RMSE. AKA "coefficient of variation"

## [1] 0.7512441

# Tune mtry using OOB error  
set.seed(25)  
#train\_pred <- predict(object = Mod1, newdata = PAtrain)  
res <- tuneRF(x = Wettrain,  
 y = Wettrain$WetAcre,  
 ntree = 500,  
 stepfactor = 0.5,  
 doBest=TRUE, # Returns a random forest model with optimal mtry value  
 importance = TRUE)

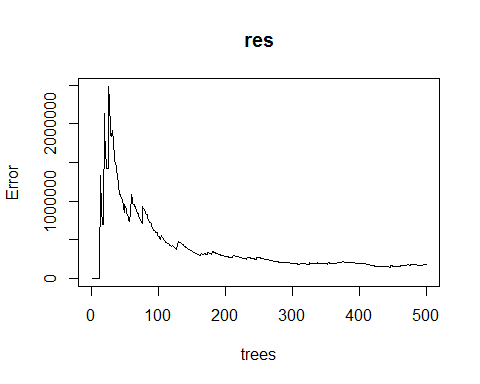
## mtry = 2 OOB error = 104055217   
## Searching left ...  
## mtry = 1 OOB error = 665119338   
## -5.391985 0.05   
## Searching right ...  
## mtry = 4 OOB error = 2698725   
## 0.9740645 0.05   
## mtry = 6 OOB error = 220350   
## 0.9183503 0.05



#localImp = TRUE)  
print(res)

##   
## Call:  
## randomForest(x = x, y = y, mtry = res[which.min(res[, 2]), 1], importance = TRUE, stepfactor = 0.5)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 6  
##   
## Mean of squared residuals: 182323  
## % Var explained: 99.99

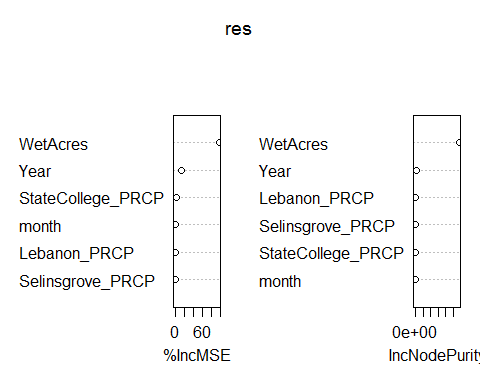
plot(res)



res$importance

## %IncMSE IncNodePurity  
## Year 234384347.4 1.705639e+10  
## month 128183.8 5.548648e+04  
## StateCollege\_PRCP 3052427.8 7.795892e+06  
## Lebanon\_PRCP 115510.7 7.563779e+07  
## Selinsgrove\_PRCP -184268.7 1.015508e+07  
## WetAcres 6960489329.8 5.724630e+11

varImpPlot(res)



## Wet Dollars

# Split into trainning, validation, and test sets  
set.seed(25)  
assignment <- sample(1:3, size = nrow(WetDollars), prob = c(0.7, 0.15, 0.15), replace = TRUE)  
  
Wettrain2 <- WetDollars[assignment == 1,]  
Wetvalid2 <- WetDollars[assignment == 2,]  
Wettest2 <- WetDollars[assignment == 3,]  
  
summary(Wettrain2)

## Year month StateCollege\_PRCP Lebanon\_PRCP   
## Min. :2001 Min. : 1.000 Min. : 1.620 Min. : 0.910   
## 1st Qu.:2005 1st Qu.: 4.000 1st Qu.: 5.270 1st Qu.: 6.335   
## Median :2009 Median : 7.000 Median : 8.050 Median : 9.330   
## Mean :2009 Mean : 6.638 Mean : 8.831 Mean : 9.718   
## 3rd Qu.:2014 3rd Qu.:10.000 3rd Qu.:11.370 3rd Qu.:12.145   
## Max. :2018 Max. :12.000 Max. :26.840 Max. :39.150   
## Selinsgrove\_PRCP WetDollars   
## Min. : 2.060 Min. : 129986   
## 1st Qu.: 5.400 1st Qu.: 1653080   
## Median : 7.920 Median : 7411996   
## Mean : 8.842 Mean : 8099715   
## 3rd Qu.:11.660 3rd Qu.:14673803   
## Max. :32.740 Max. :20514682

summary(Wetvalid2)

## Year month StateCollege\_PRCP Lebanon\_PRCP   
## Min. :2001 Min. : 1.000 Min. : 1.56 Min. : 1.190   
## 1st Qu.:2007 1st Qu.: 3.000 1st Qu.: 5.99 1st Qu.: 5.692   
## Median :2012 Median : 6.000 Median : 7.82 Median : 9.090   
## Mean :2011 Mean : 6.077 Mean : 9.17 Mean :11.616   
## 3rd Qu.:2015 3rd Qu.: 9.000 3rd Qu.:12.43 3rd Qu.:15.155   
## Max. :2018 Max. :12.000 Max. :22.57 Max. :45.690   
## Selinsgrove\_PRCP WetDollars   
## Min. : 0.000 Min. : 129986   
## 1st Qu.: 5.675 1st Qu.: 1668643   
## Median : 8.850 Median : 6842246   
## Mean :10.449 Mean : 8398955   
## 3rd Qu.:12.215 3rd Qu.:14673803   
## Max. :48.100 Max. :20011028

summary(Wettest2)

## Year month StateCollege\_PRCP Lebanon\_PRCP   
## Min. :2002 Min. : 1.000 Min. : 2.940 Min. : 2.31   
## 1st Qu.:2004 1st Qu.: 4.000 1st Qu.: 6.495 1st Qu.: 6.39   
## Median :2008 Median : 6.000 Median : 9.260 Median :10.04   
## Mean :2009 Mean : 6.074 Mean : 9.480 Mean :11.22   
## 3rd Qu.:2014 3rd Qu.: 8.000 3rd Qu.:10.905 3rd Qu.:12.82   
## Max. :2018 Max. :12.000 Max. :19.730 Max. :33.07   
## Selinsgrove\_PRCP WetDollars   
## Min. : 0.200 Min. : 419619   
## 1st Qu.: 5.165 1st Qu.: 1262458   
## Median : 8.960 Median : 7456851   
## Mean : 9.075 Mean : 8239412   
## 3rd Qu.:11.850 3rd Qu.:11783443   
## Max. :26.700 Max. :20514682

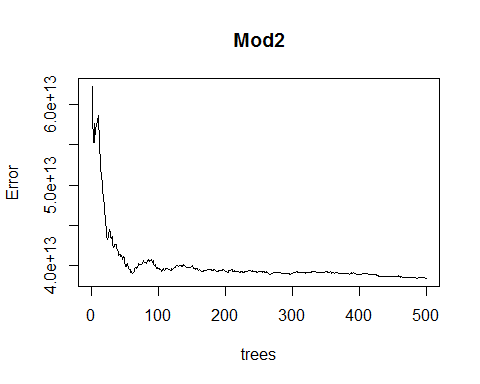
Mod2 <- randomForest(WetDollars ~ .,   
 data = Wettrain2,   
 ntree = 500,   
 #method = "anova",   
 importance = TRUE)  
  
print(Mod2) # % of variance expalined is low. Tuning needed

##   
## Call:  
## randomForest(formula = WetDollars ~ ., data = Wettrain2, ntree = 500, importance = TRUE)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 1  
##   
## Mean of squared residuals: 3.843481e+13  
## % Var explained: 20.23

summary(Mod2)

## Length Class Mode   
## call 5 -none- call   
## type 1 -none- character  
## predicted 163 -none- numeric   
## mse 500 -none- numeric   
## rsq 500 -none- numeric   
## oob.times 163 -none- numeric   
## importance 10 -none- numeric   
## importanceSD 5 -none- numeric   
## localImportance 0 -none- NULL   
## proximity 0 -none- NULL   
## ntree 1 -none- numeric   
## mtry 1 -none- numeric   
## forest 11 -none- list   
## coefs 0 -none- NULL   
## y 163 -none- numeric   
## test 0 -none- NULL   
## inbag 0 -none- NULL   
## terms 3 terms call

plot(Mod2)

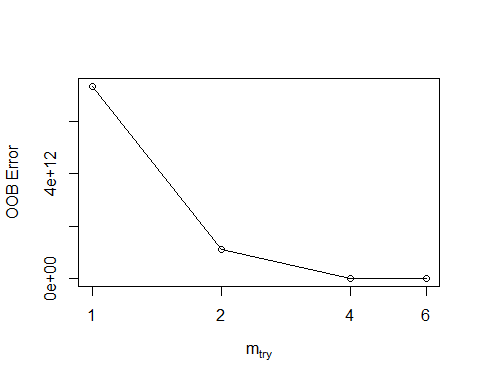


pred2 <- predict(object = Mod2, newdata = Wettest2)  
RMSE\_Mod2 <- rmse(actual = Wettest2$WetDollars, #actual values  
 predicted = pred2) #predicted values  
print(RMSE\_Mod2/mean(Wettest2$WetDollars)) #tells us the %of the mean represented by RMSE. AKA "coefficient of variation"

## [1] 0.6656554

# Tune mtry using OOB error  
set.seed(25)  
#train\_pred <- predict(object = Mod1, newdata = PAtrain)  
res2 <- tuneRF(x = Wettrain2,  
 y = Wettrain2$WetDollars,  
 ntree = 500,  
 stepfactor = 0.5,  
 doBest=TRUE, # Returns a random forest model with optimal mtry value  
 importance = TRUE)

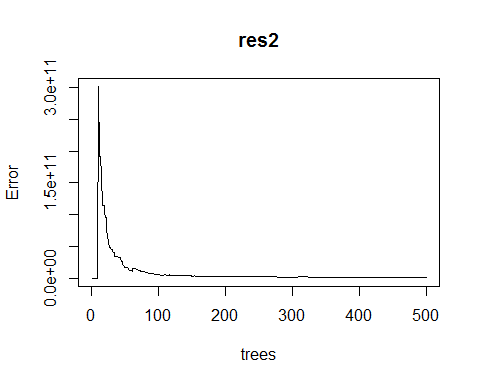
## mtry = 2 OOB error = 1.099613e+12   
## Searching left ...  
## mtry = 1 OOB error = 7.33856e+12   
## -5.673764 0.05   
## Searching right ...  
## mtry = 4 OOB error = 16241289456   
## 0.98523 0.05   
## mtry = 6 OOB error = 1715233239   
## 0.8943906 0.05



#localImp = TRUE)  
print(res2)

##   
## Call:  
## randomForest(x = x, y = y, mtry = res[which.min(res[, 2]), 1], importance = TRUE, stepfactor = 0.5)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 6  
##   
## Mean of squared residuals: 1355209009  
## % Var explained: 100

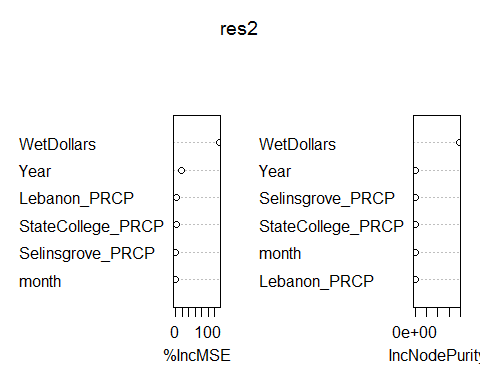
plot(res2)



res2$importance

## %IncMSE IncNodePurity  
## Year 8.579820e+11 6.849186e+13  
## month -3.274544e+10 1.883829e+10  
## StateCollege\_PRCP 4.678427e+09 2.393336e+10  
## Lebanon\_PRCP 3.499059e+10 1.863356e+10  
## Selinsgrove\_PRCP -8.042994e+09 1.105070e+11  
## WetDollars 9.075158e+13 7.776690e+15

varImpPlot(res2)

 ## Dry Acres

# Split into trainning, validation, and test sets  
set.seed(25)  
assignment <- sample(1:3, size = nrow(DryAcres), prob = c(0.7, 0.15, 0.15), replace = TRUE)  
  
Drytrain <- DryAcres[assignment == 1,]  
Dryvalid <- DryAcres[assignment == 2,]  
Drytest <- DryAcres[assignment == 3,]  
  
summary(Drytrain)

## Year month StateCollege\_PRCP Lebanon\_PRCP   
## Min. :2001 Min. : 1.000 Min. : 1.620 Min. : 0.910   
## 1st Qu.:2005 1st Qu.: 4.000 1st Qu.: 5.270 1st Qu.: 6.335   
## Median :2009 Median : 7.000 Median : 8.050 Median : 9.330   
## Mean :2009 Mean : 6.638 Mean : 8.831 Mean : 9.718   
## 3rd Qu.:2014 3rd Qu.:10.000 3rd Qu.:11.370 3rd Qu.:12.145   
## Max. :2018 Max. :12.000 Max. :26.840 Max. :39.150   
## Selinsgrove\_PRCP DryAcres   
## Min. : 2.060 Min. : 557.8   
## 1st Qu.: 5.400 1st Qu.: 16050.8   
## Median : 7.920 Median : 80698.0   
## Mean : 8.842 Mean :117531.2   
## 3rd Qu.:11.660 3rd Qu.:174858.0   
## Max. :32.740 Max. :613635.8

summary(Dryvalid)

## Year month StateCollege\_PRCP Lebanon\_PRCP   
## Min. :2001 Min. : 1.000 Min. : 1.56 Min. : 1.190   
## 1st Qu.:2007 1st Qu.: 3.000 1st Qu.: 5.99 1st Qu.: 5.692   
## Median :2012 Median : 6.000 Median : 7.82 Median : 9.090   
## Mean :2011 Mean : 6.077 Mean : 9.17 Mean :11.616   
## 3rd Qu.:2015 3rd Qu.: 9.000 3rd Qu.:12.43 3rd Qu.:15.155   
## Max. :2018 Max. :12.000 Max. :22.57 Max. :45.690   
## Selinsgrove\_PRCP DryAcres   
## Min. : 0.000 Min. : 557.8   
## 1st Qu.: 5.675 1st Qu.: 19810.8   
## Median : 8.850 Median : 77513.5   
## Mean :10.449 Mean :115677.5   
## 3rd Qu.:12.215 3rd Qu.:216408.0   
## Max. :48.100 Max. :613635.8

summary(Drytest)

## Year month StateCollege\_PRCP Lebanon\_PRCP   
## Min. :2002 Min. : 1.000 Min. : 2.940 Min. : 2.31   
## 1st Qu.:2004 1st Qu.: 4.000 1st Qu.: 6.495 1st Qu.: 6.39   
## Median :2008 Median : 6.000 Median : 9.260 Median :10.04   
## Mean :2009 Mean : 6.074 Mean : 9.480 Mean :11.22   
## 3rd Qu.:2014 3rd Qu.: 8.000 3rd Qu.:10.905 3rd Qu.:12.82   
## Max. :2018 Max. :12.000 Max. :19.730 Max. :33.07   
## Selinsgrove\_PRCP DryAcres   
## Min. : 0.200 Min. : 557.8   
## 1st Qu.: 5.165 1st Qu.: 10143.4   
## Median : 8.960 Median : 74329.0   
## Mean : 9.075 Mean :106350.5   
## 3rd Qu.:11.850 3rd Qu.:136717.0   
## Max. :26.700 Max. :613635.8

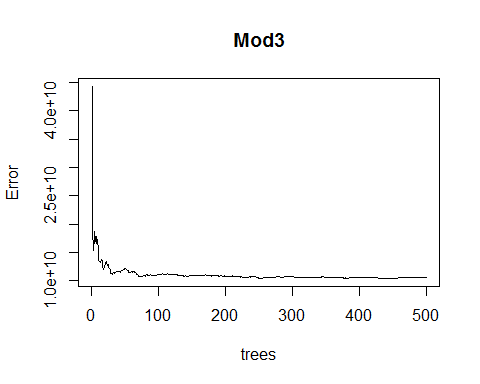
Mod3 <- randomForest(DryAcres ~ .,   
 data = Drytrain,   
 ntree = 500,   
 #method = "anova",   
 importance = TRUE)  
  
print(Mod3) # % of variance expalined is low. Tuning needed

##   
## Call:  
## randomForest(formula = DryAcres ~ ., data = Drytrain, ntree = 500, importance = TRUE)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 1  
##   
## Mean of squared residuals: 10588143143  
## % Var explained: 48.68

summary(Mod3)

## Length Class Mode   
## call 5 -none- call   
## type 1 -none- character  
## predicted 163 -none- numeric   
## mse 500 -none- numeric   
## rsq 500 -none- numeric   
## oob.times 163 -none- numeric   
## importance 10 -none- numeric   
## importanceSD 5 -none- numeric   
## localImportance 0 -none- NULL   
## proximity 0 -none- NULL   
## ntree 1 -none- numeric   
## mtry 1 -none- numeric   
## forest 11 -none- list   
## coefs 0 -none- NULL   
## y 163 -none- numeric   
## test 0 -none- NULL   
## inbag 0 -none- NULL   
## terms 3 terms call

plot(Mod3)

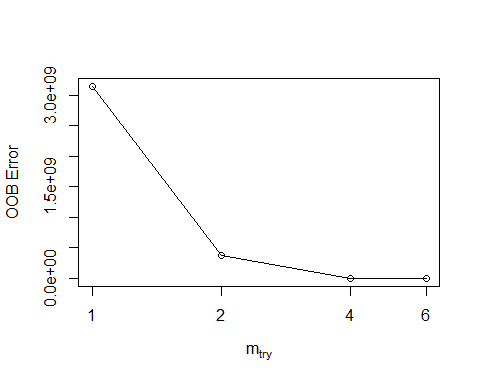


pred3 <- predict(object = Mod3, newdata = Drytest)  
RMSE\_Mod3 <- rmse(actual = Drytest$DryAcres, #actual values  
 predicted = pred3) #predicted values  
print(RMSE\_Mod3/mean(Drytest$DryAcres)) #tells us the %of the mean represented by RMSE. AKA "coefficient of variation"

## [1] 1.034596

# Tune mtry using OOB error  
set.seed(25)  
#train\_pred <- predict(object = Mod1, newdata = PAtrain)  
res3 <- tuneRF(x = Drytrain,  
 y = Drytrain$DryAcres,  
 ntree = 500,  
 stepfactor = 0.5,  
 doBest=TRUE, # Returns a random forest model with optimal mtry value  
 importance = TRUE)

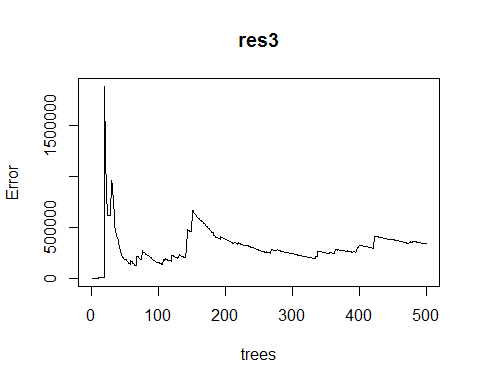
## mtry = 2 OOB error = 375443571   
## Searching left ...  
## mtry = 1 OOB error = 3148461497   
## -7.385978 0.05   
## Searching right ...  
## mtry = 4 OOB error = 7126065   
## 0.9810196 0.05   
## mtry = 6 OOB error = 1283593   
## 0.8198736 0.05



#localImp = TRUE)  
print(res3)

##   
## Call:  
## randomForest(x = x, y = y, mtry = res[which.min(res[, 2]), 1], importance = TRUE, stepfactor = 0.5)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 6  
##   
## Mean of squared residuals: 344393.2  
## % Var explained: 100

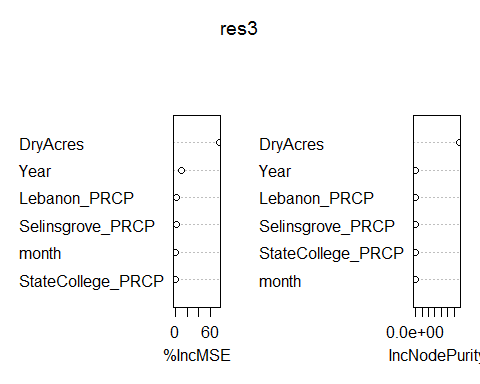
plot(res3) # looks pretty choppy?



res3$importance

## %IncMSE IncNodePurity  
## Year 3.408075e+07 2.555319e+09  
## month 3.531430e+05 2.042610e+04  
## StateCollege\_PRCP -5.826061e+04 3.869144e+05  
## Lebanon\_PRCP 4.122111e+07 1.712351e+07  
## Selinsgrove\_PRCP 1.101792e+07 1.132820e+06  
## DryAcres 4.068888e+10 3.363626e+12

varImpPlot(res3)



## Dry Dollars

# Split into trainning, validation, and test sets  
set.seed(25)  
assignment <- sample(1:3, size = nrow(DryDollars), prob = c(0.7, 0.15, 0.15), replace = TRUE)  
  
Drytrain2 <- DryDollars[assignment == 1,]  
Dryvalid2 <- DryDollars[assignment == 2,]  
Drytest2 <- DryDollars[assignment == 3,]  
  
summary(Drytrain2)

## Year month StateCollege\_PRCP Lebanon\_PRCP   
## Min. :2001 Min. : 1.000 Min. : 1.620 Min. : 0.910   
## 1st Qu.:2005 1st Qu.: 4.000 1st Qu.: 5.270 1st Qu.: 6.335   
## Median :2009 Median : 7.000 Median : 8.050 Median : 9.330   
## Mean :2009 Mean : 6.638 Mean : 8.831 Mean : 9.718   
## 3rd Qu.:2014 3rd Qu.:10.000 3rd Qu.:11.370 3rd Qu.:12.145   
## Max. :2018 Max. :12.000 Max. :26.840 Max. :39.150   
## Selinsgrove\_PRCP DryDollars   
## Min. : 2.060 Min. : 77650   
## 1st Qu.: 5.400 1st Qu.: 1592471   
## Median : 7.920 Median :11030151   
## Mean : 8.842 Mean :12920912   
## 3rd Qu.:11.660 3rd Qu.:17790895   
## Max. :32.740 Max. :54954463

summary(Dryvalid2)

## Year month StateCollege\_PRCP Lebanon\_PRCP   
## Min. :2001 Min. : 1.000 Min. : 1.56 Min. : 1.190   
## 1st Qu.:2007 1st Qu.: 3.000 1st Qu.: 5.99 1st Qu.: 5.692   
## Median :2012 Median : 6.000 Median : 7.82 Median : 9.090   
## Mean :2011 Mean : 6.077 Mean : 9.17 Mean :11.616   
## 3rd Qu.:2015 3rd Qu.: 9.000 3rd Qu.:12.43 3rd Qu.:15.155   
## Max. :2018 Max. :12.000 Max. :22.57 Max. :45.690   
## Selinsgrove\_PRCP DryDollars   
## Min. : 0.000 Min. : 77650   
## 1st Qu.: 5.675 1st Qu.: 2282997   
## Median : 8.850 Median : 8585168   
## Mean :10.449 Mean :13427212   
## 3rd Qu.:12.215 3rd Qu.:17790895   
## Max. :48.100 Max. :54954463

summary(Drytest2)

## Year month StateCollege\_PRCP Lebanon\_PRCP   
## Min. :2002 Min. : 1.000 Min. : 2.940 Min. : 2.31   
## 1st Qu.:2004 1st Qu.: 4.000 1st Qu.: 6.495 1st Qu.: 6.39   
## Median :2008 Median : 6.000 Median : 9.260 Median :10.04   
## Mean :2009 Mean : 6.074 Mean : 9.480 Mean :11.22   
## 3rd Qu.:2014 3rd Qu.: 8.000 3rd Qu.:10.905 3rd Qu.:12.82   
## Max. :2018 Max. :12.000 Max. :19.730 Max. :33.07   
## Selinsgrove\_PRCP DryDollars   
## Min. : 0.200 Min. : 77650   
## 1st Qu.: 5.165 1st Qu.: 986961   
## Median : 8.960 Median : 5463886   
## Mean : 9.075 Mean :11417461   
## 3rd Qu.:11.850 3rd Qu.:12737743   
## Max. :26.700 Max. :54954463

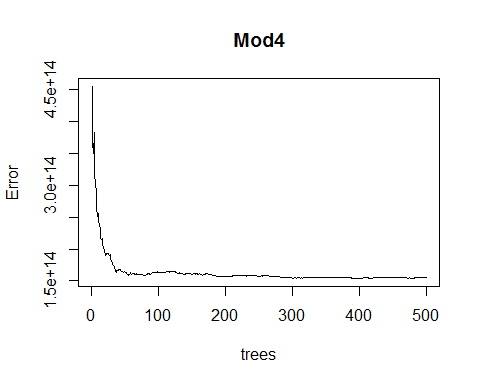
Mod4 <- randomForest(DryDollars ~ .,   
 data = Drytrain2,   
 ntree = 500,   
 #method = "anova",   
 importance = TRUE)  
  
print(Mod4) # % of variance expalined is low. Tuning needed

##   
## Call:  
## randomForest(formula = DryDollars ~ ., data = Drytrain2, ntree = 500, importance = TRUE)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 1  
##   
## Mean of squared residuals: 1.54949e+14  
## % Var explained: 29.39

summary(Mod4)

## Length Class Mode   
## call 5 -none- call   
## type 1 -none- character  
## predicted 163 -none- numeric   
## mse 500 -none- numeric   
## rsq 500 -none- numeric   
## oob.times 163 -none- numeric   
## importance 10 -none- numeric   
## importanceSD 5 -none- numeric   
## localImportance 0 -none- NULL   
## proximity 0 -none- NULL   
## ntree 1 -none- numeric   
## mtry 1 -none- numeric   
## forest 11 -none- list   
## coefs 0 -none- NULL   
## y 163 -none- numeric   
## test 0 -none- NULL   
## inbag 0 -none- NULL   
## terms 3 terms call

plot(Mod4)

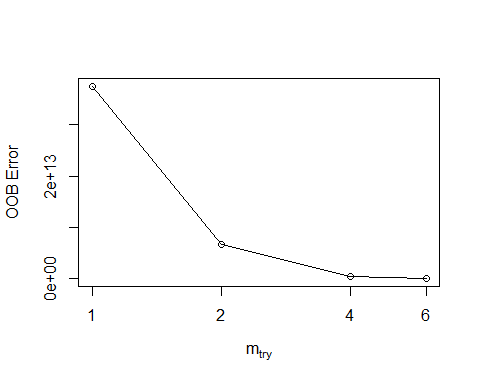


pred4 <- predict(object = Mod4, newdata = Drytest2)  
RMSE\_Mod4 <- rmse(actual = Drytest2$DryDollars, #actual values  
 predicted = pred4) #predicted values  
print(RMSE\_Mod4/mean(Drytest2$DryDollars)) #tells us the %of the mean represented by RMSE. AKA "coefficient of variation"

## [1] 1.092932

# Tune mtry using OOB error  
set.seed(25)  
#train\_pred <- predict(object = Mod1, newdata = PAtrain)  
res4 <- tuneRF(x = Drytrain2,  
 y = Drytrain2$DryDollars,  
 ntree = 500,  
 stepfactor = 0.5,  
 doBest=TRUE, # Returns a random forest model with optimal mtry value  
 importance = TRUE)

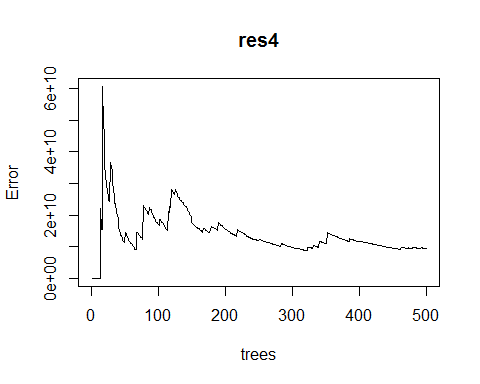
## mtry = 2 OOB error = 6.736511e+12   
## Searching left ...  
## mtry = 1 OOB error = 3.753115e+13   
## -4.571304 0.05   
## Searching right ...  
## mtry = 4 OOB error = 297907682694   
## 0.9557772 0.05   
## mtry = 6 OOB error = 17967396216   
## 0.939688 0.05



#localImp = TRUE)  
print(res4)

##   
## Call:  
## randomForest(x = x, y = y, mtry = res[which.min(res[, 2]), 1], importance = TRUE, stepfactor = 0.5)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 6  
##   
## Mean of squared residuals: 9454798682  
## % Var explained: 100

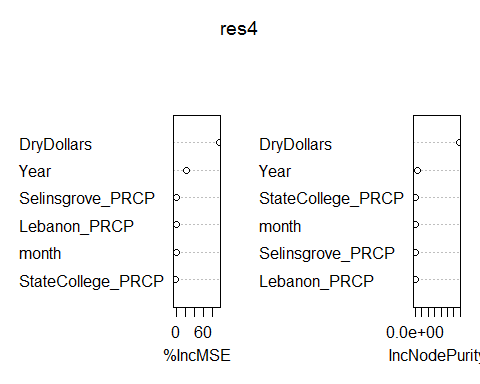
plot(res4) # looks pretty choppy?



res4$importance

## %IncMSE IncNodePurity  
## Year 1.713391e+13 1.581166e+15  
## month 1.378753e+10 8.809838e+11  
## StateCollege\_PRCP -1.767504e+11 1.531783e+12  
## Lebanon\_PRCP 8.771209e+10 2.434308e+10  
## Selinsgrove\_PRCP 1.089086e+11 1.199370e+11  
## DryDollars 4.401359e+14 3.398425e+16

varImpPlot(res4)



# Other optinos for fine tuning the model using control function

Mod2 <- randomForest(WetAcres ~ .,   
 data = Wettrain,   
 ntree = 500,   
 mtry = 6, # based on tuneRF function results above  
 importance = TRUE,  
 control = rpart.control(minsplit = 20, # default is 20  
 cp = 0.01, # default is 0.01  
 maxdepth = 30)) # default is 30

## Warning in randomForest.default(m, y, ...): invalid mtry: reset to within  
## valid range

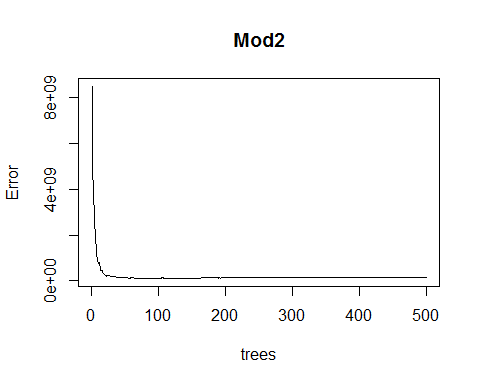
print(Mod2)

##   
## Call:  
## randomForest(formula = WetAcres ~ ., data = Wettrain, ntree = 500, mtry = 6, importance = TRUE, control = rpart.control(minsplit = 20, cp = 0.01, maxdepth = 30))   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 5  
##   
## Mean of squared residuals: 131875755  
## % Var explained: 96.37

summary(Mod2)

## Length Class Mode   
## call 7 -none- call   
## type 1 -none- character  
## predicted 163 -none- numeric   
## mse 500 -none- numeric   
## rsq 500 -none- numeric   
## oob.times 163 -none- numeric   
## importance 10 -none- numeric   
## importanceSD 5 -none- numeric   
## localImportance 0 -none- NULL   
## proximity 0 -none- NULL   
## ntree 1 -none- numeric   
## mtry 1 -none- numeric   
## forest 11 -none- list   
## coefs 0 -none- NULL   
## y 163 -none- numeric   
## test 0 -none- NULL   
## inbag 0 -none- NULL   
## terms 3 terms call

plot(Mod2)



pred <- predict(object = Mod2, newdata = Wettest)  
RMSE\_Mod2 <- rmse(actual = Wettest$WetAcres, #actual values  
 predicted = pred) #predicted values  
print(RMSE\_Mod2/mean(Wettest$WetAcres)) #tells us the %of the mean represented by RMSE. AKA "coefficient of variation"

## [1] 0.1633907

#start here in AM  
#plotcp(Mod2)  
#summary(rpart(Mod2))

# Predicting on train set and checking classification accuracy

#predTrain <- predict(Mod2, PAtrain, type = "class")  
#table(predTrain, PAtrain$DryAcres)