homework_8

2024-11-26

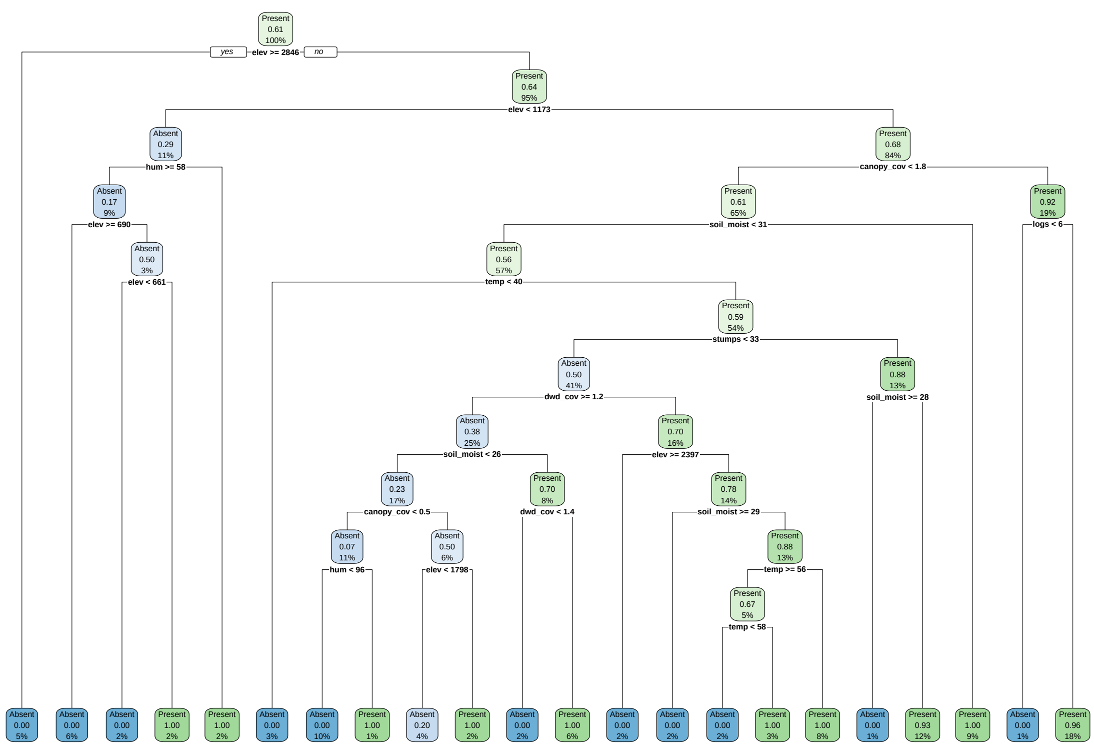
Homework 8

CART and Random Forest

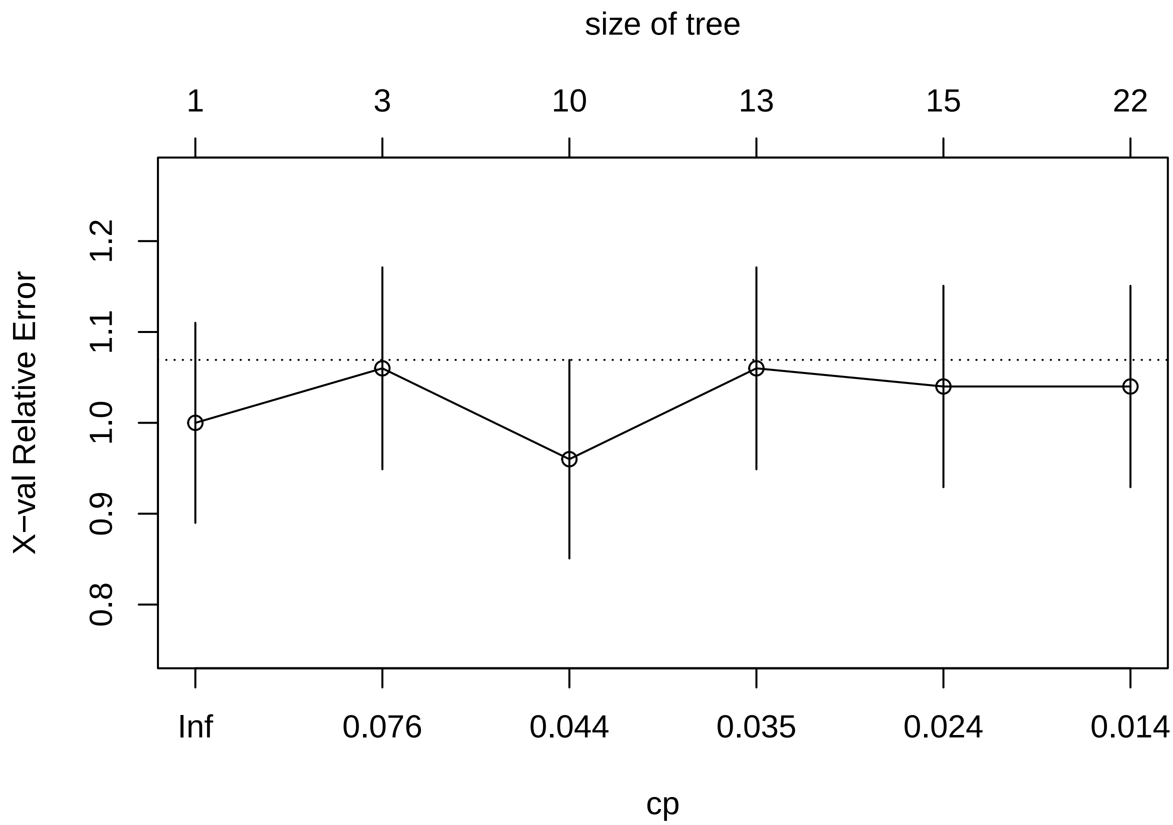
Question 1: Using your dataset, create a classification or regression tree to predict a chosen response variable from a set of predictors. Then, build a Random Forest using the same data.

1a. Use cross-validation to summarize the accuracy or MSE for each model. Classification Tree

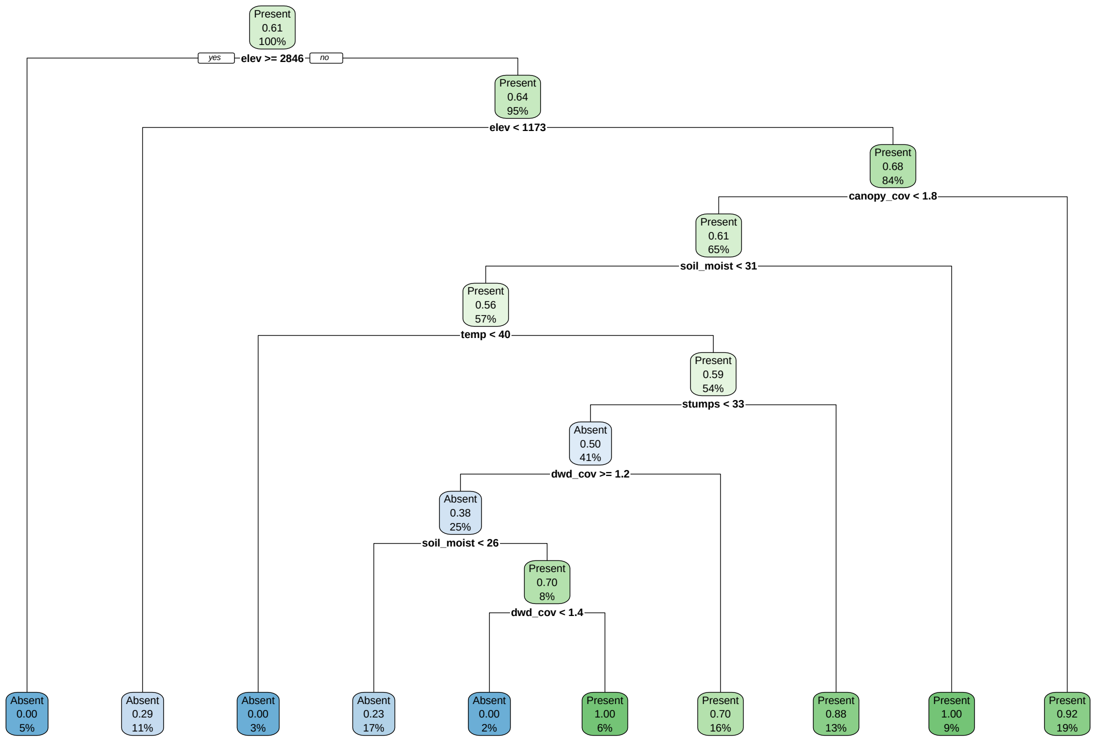
```
oss_PA <- ifelse(sals$oss > 0, "Present", "Absent")

oss.tree <- rpart(oss_PA ~ ., data=env_subset, minsplit=2, xval=5)
rpart.plot(oss.tree)
```



```r
plotcp(oss.tree)
```




```r
oss.tree.prune <- prune(oss.tree, 0.044)
rpart.plot(oss.tree.prune)
```



Random Forest

```
oss.forest <- randomForest(as.factor(oss_PA) ~ ., data=env_subset, ntree = 5000, mtry = 5, importance=TI

oss.forest
```

```
##
## Call:
##  randomForest(formula = as.factor(oss_PA) ~ ., data = env_subset,      ntree = 5000, mtry = 5, impor
##                Type of random forest: classification
##                      Number of trees: 5000
## No. of variables tried at each split: 5
##
##         OOB estimate of  error rate: 37.01%
## Confusion matrix:
##         Absent Present class.error
## Absent      18      32   0.6400000
## Present     15      62   0.1948052
```

```


oss.forest
```

1b. Examine the confusion matrix output. Explain how the Random Forest approach affects predictive accuracy compared to the single tree.

```
##
## Call:
##  randomForest(formula = as.factor(oss_PA) ~ ., data = env_subset,      ntree = 5000, mtry = 5, impor
##                Type of random forest: classification
##                      Number of trees: 5000
## No. of variables tried at each split: 5
##
##         OOB estimate of  error rate: 37.01%
## Confusion matrix:
##         Absent Present class.error
## Absent      18      32   0.6400000
## Present     15      62   0.1948052
```

```r
table(oss_PA)
```

```
## oss_PA
##  Absent Present
##      50      77
```

The error rate is pretty bad (35.43%), and most of that is due to the absences (62%). I tried tuning the mtry and ntree to improve the matrix but I havent figured out a way to make it better yet. RF generally improves accuracy compared to single tree because it averages over several trees.

1c. Discuss any improvements you observe with Random Forests over single trees, and why this might be the case. The confusion matrix shows better overall accuracy and lower error for "Present" compared to a single tree, though "Absent" still has higher misclassification. This suggests RF performs well with the majority class (present) but may need adjustments for imbalance.

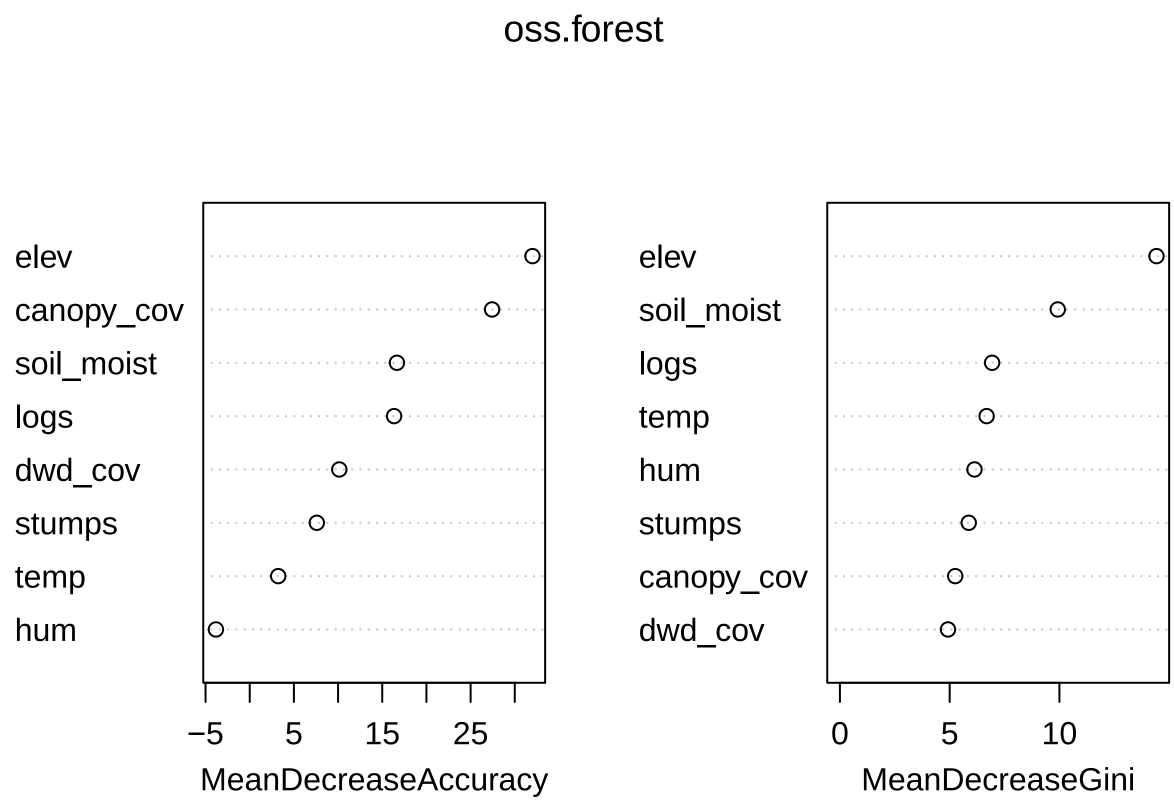
Question 2: Using the Random Forest output from Question 1, evaluate the importance of individual predictors in the model.

```r
importance(oss.forest)
```

```
##                Absent   Present MeanDecreaseAccuracy MeanDecreaseGini
## elev        14.7657796 30.999725            32.004100        14.421786
## temp        -0.5487004  4.560265             3.224890         6.684052
## hum        -10.6940947  4.587271            -3.825727         6.130093
## canopy_cov   6.4274066 29.219640            27.431739         5.252533
## dwd_cov      3.0134649 10.911550            10.138762         4.923518
## soil_moist   2.9047114 20.342670            16.661326         9.924071
## stumps     -22.7949467 25.187690             7.591146         5.869146
## logs         9.4480739 13.988874            16.340870         6.933188
```

```r
varImpPlot(oss.forest)
```

2a. Report and visualize the predictor importance scores. What are the top predictors in each model?

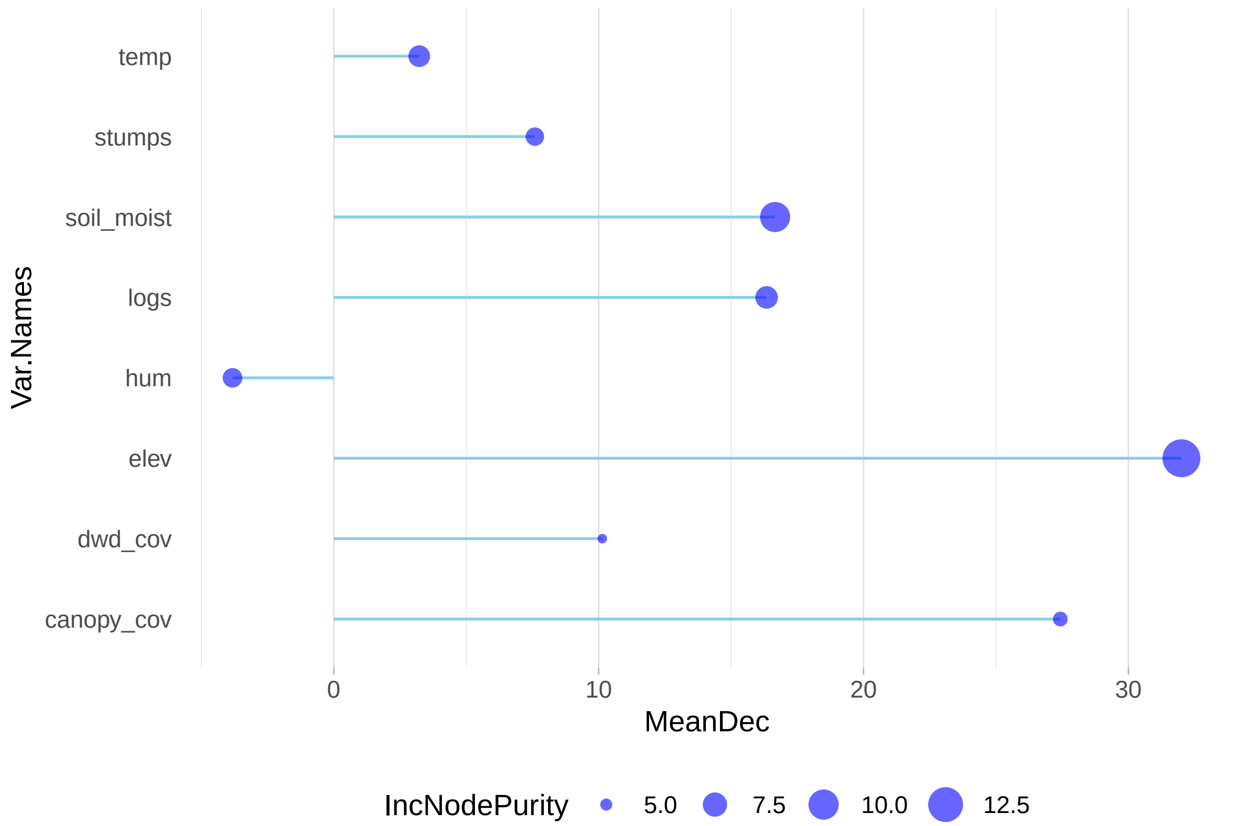



```r
ForestData <- as.data.frame(importance(oss.forest))
        ForestData <- ForestData[order(ForestData[,1]),]
        ForestData$Var.Names <- row.names(ForestData)
        colnames(ForestData) <- c("Absent","Present","MeanDec","IncNodePurity","Var.Names")
        ForestData
```

```
##                 Absent   Present   MeanDec IncNodePurity  Var.Names
## stumps     -22.7949467 25.187690  7.591146      5.869146     stumps
## hum        -10.6940947  4.587271 -3.825727      6.130093        hum
## temp        -0.5487004  4.560265  3.224890      6.684052       temp
## soil_moist   2.9047114 20.342670 16.661326      9.924071 soil_moist
## dwd_cov      3.0134649 10.911550 10.138762      4.923518    dwd_cov
## canopy_cov   6.4274066 29.219640 27.431739      5.252533 canopy_cov
## logs         9.4480739 13.988874 16.340870      6.933188       logs
## elev        14.7657796 30.999725 32.004100     14.421786       elev
```

```r
ggplot(ForestData, aes(x=Var.Names, y=MeanDec)) +
    geom_segment( aes(x=Var.Names, xend=Var.Names, y=0, yend=MeanDec), color="skyblue") +
        geom_point(aes(size = IncNodePurity), color="blue", alpha=0.6) +
        theme_light() +
        coord_flip() +
        theme(
                legend.position="bottom",
                panel.grid.major.y = element_blank(),
```

```r
                panel.border = element_blank(),
                axis.ticks.y = element_blank()
        )
```



The top predictors here are elevation, canopy cover, soil moisture, and logs. Elevation makes sense, because we know that they only exist in a certain elevational band, but I didn't expect it to be the most important driver because it isnt directly related to the disturbances we are investigating. Maybe I need to look at how the fires or harvest are spread across different elevations and see if there is an obvious spatial pattern. Canopy cover is an obvious one, and I'm happy to see soil moisture and logs on there because we hypothesized that those would be important.

2b. Are the most important features in the Random Forest model the same as those in the single decision tree? Explain any differences. Elevation, canopy cover, and soil moisture are also the top predictors for the classification tree, but logs dont show up.

Question 3: How does the ensemble nature of Random Forests affect model interpretation, and why might it present challenges or advantages for understanding predictor influence?

Random Forests combine multiple decision trees, making predictions based on the majority vote. They are more robust and accurate because of this, but they cant give us a straightforward decision tree or pinpoint specific relationships. This requires more thought and speculation about the top predictor variables.