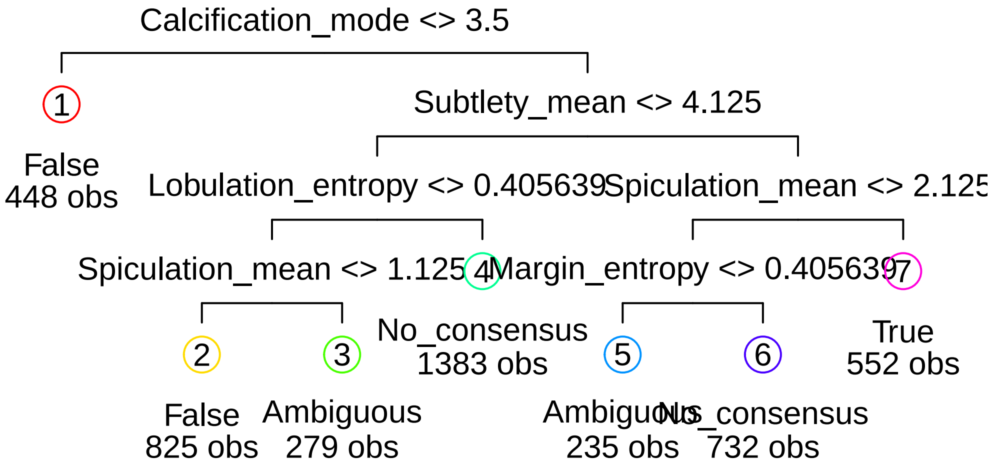
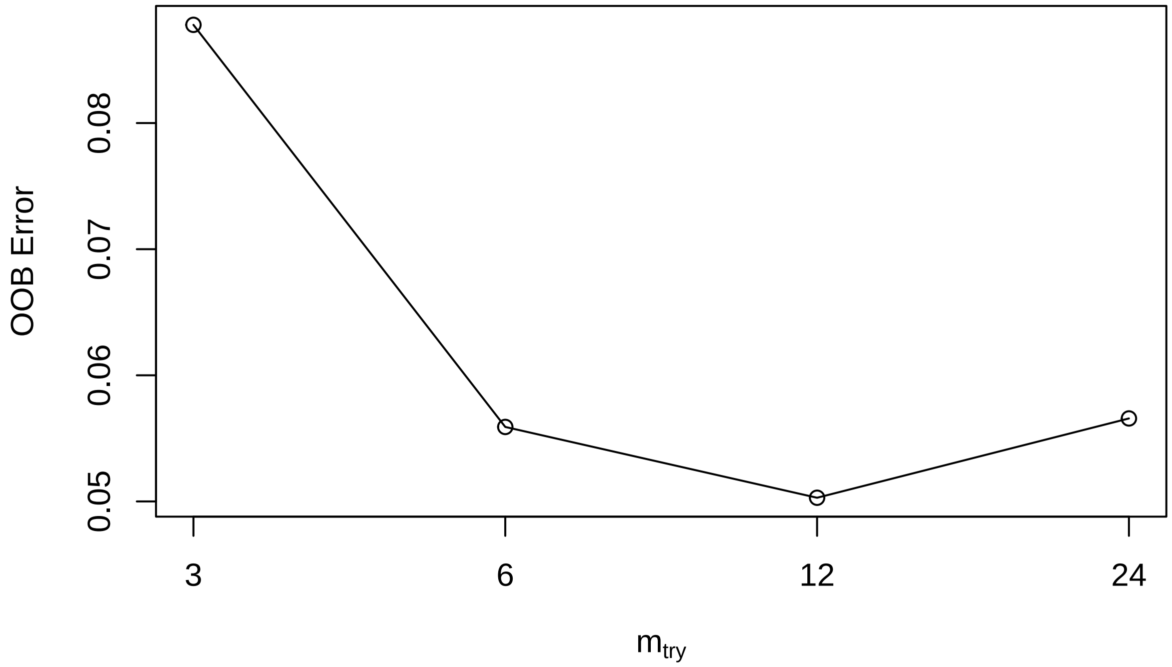
STAT 447 project

Ken Mawer

2023-04-05



```
## mtry = 6  OOB error = 5.59%
## Searching left ...
## mtry = 3     OOB error = 8.78%
## -0.5702811 0.05
## Searching right ...
## mtry = 12    OOB error = 5.03%
## 0.1004016 0.05
## mtry = 24    OOB error = 5.66%
## -0.125 0.05
```



```
##         mtry   OOBError
## 3.OOB      3 0.08778626
## 6.OOB      6 0.05590480
## 12.OOB    12 0.05029187
## 24.OOB    24 0.05657836
```

```
## Warning in confusionMatrix.default(predict(randomForest(Is_cancer ~ ., data =
## t), : Levels are not in the same order for reference and data. Refactoring data
## to match.
```

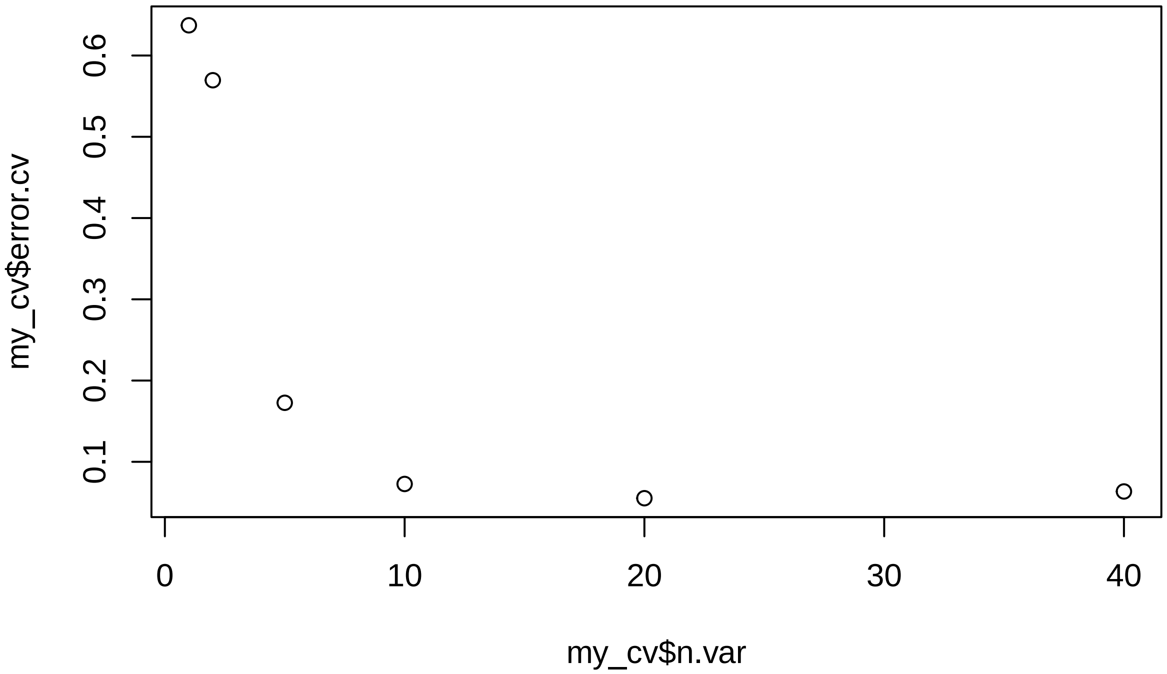
```
## Confusion Matrix and Statistics
##
##                Reference
## Prediction     True Ambiguous False No_consensus
##   True          166        25    13           76
##   Ambiguous      46       288   120          127
##   False          37       172   562           72
##   No_consensus   91       136   101          303
##
## Overall Statistics
##
##                Accuracy : 0.5649
##                  95% CI : (0.5445, 0.5851)
##     No Information Rate : 0.3409
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.401
```

```
##
##  Mcnemar's Test P-Value : 8.37e-06
##
## Statistics by Class:
##
##                      Class: True Class: Ambiguous Class: False
## Sensitivity              0.48824           0.4638       0.7060
## Specificity              0.94286           0.8291       0.8174
## Pos Pred Value           0.59286           0.4957       0.6667
## Neg Pred Value           0.91533           0.8101       0.8432
## Prevalence               0.14561           0.2660       0.3409
## Detection Rate           0.07109           0.1233       0.2407
## Detection Prevalence     0.11991           0.2488       0.3610
## Balanced Accuracy        0.71555           0.6464       0.7617
##                      Class: No_consensus
## Sensitivity                       0.5242
## Specificity                       0.8133
## Pos Pred Value                    0.4802
## Neg Pred Value                    0.8386
## Prevalence                        0.2475
## Detection Rate                    0.1298
## Detection Prevalence              0.2702
## Balanced Accuracy                 0.6688
```

```
## Warning in confusionMatrix.default(predict(randomForest(Is_cancer ~ ., data =
## t, : Levels are not in the same order for reference and data. Refactoring data
## to match.
```

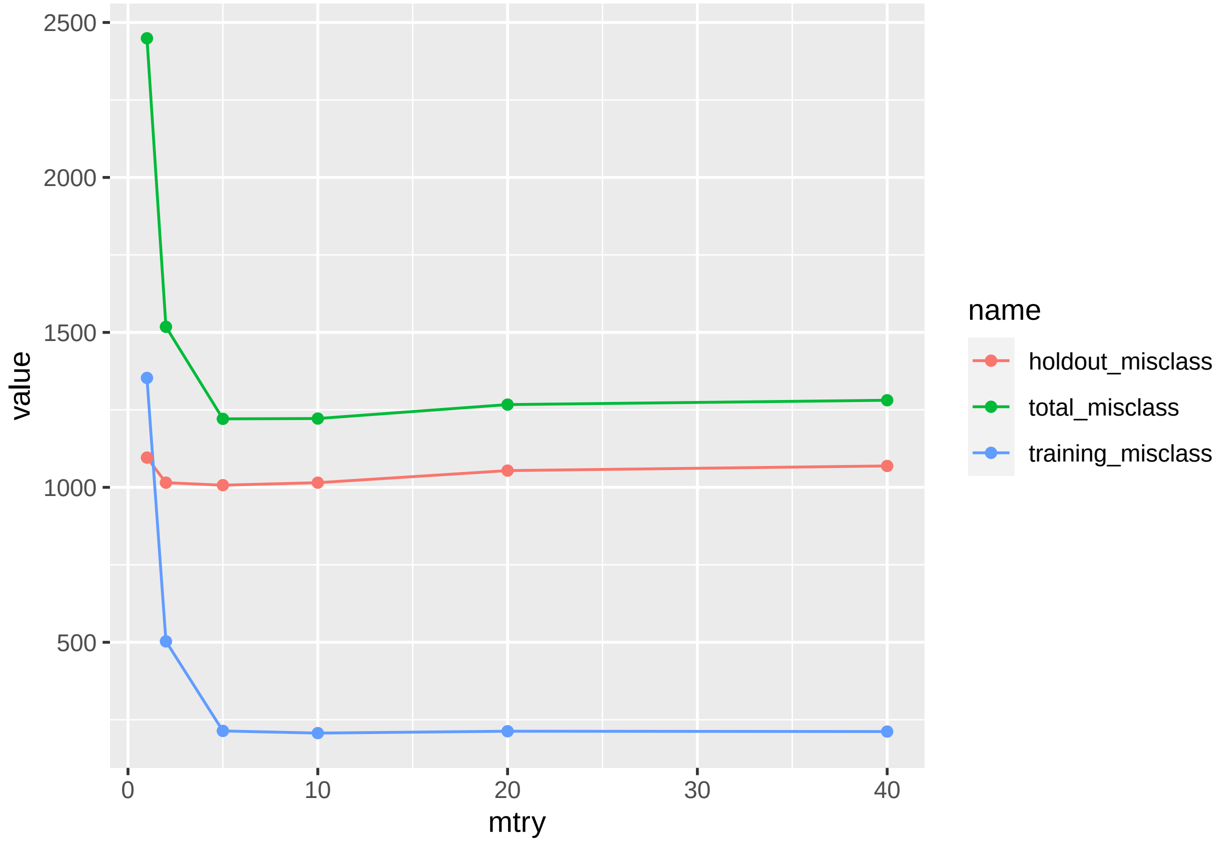
```
## Confusion Matrix and Statistics
##
##               Reference
## Prediction     True Ambiguous False No_consensus
##   True          169        29    16           80
##   Ambiguous      49       286   132          130
##   False          32       171   548           66
##   No_consensus   90       135   100          302
##
## Overall Statistics
##
##                Accuracy : 0.5589
##                  95% CI : (0.5385, 0.5792)
##     No Information Rate : 0.3409
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.394
##
##  Mcnemar's Test P-Value : 0.0007548
##
## Statistics by Class:
##
##                      Class: True Class: Ambiguous Class: False
## Sensitivity              0.49706           0.4605       0.6884
## Specificity              0.93734           0.8186       0.8252
## Pos Pred Value           0.57483           0.4791       0.6707
## Neg Pred Value           0.91622           0.8072       0.8366
## Prevalence                 0.14561          0.2660       0.3409
## Detection Rate             0.07238          0.1225       0.2347
## Detection Prevalence       0.12591          0.2557       0.3499
## Balanced Accuracy          0.71720          0.6396       0.7568
##                      Class: No_consensus
## Sensitivity                      0.5225
## Specificity                      0.8150
## Pos Pred Value                   0.4817
## Neg Pred Value                   0.8384
## Prevalence                       0.2475
## Detection Rate                   0.1293
## Detection Prevalence             0.2685
## Balanced Accuracy                0.6688
```

```
plot(my_cv$n.var, my_cv$error.cv)
```



```
# Number of training set misclassifications
t_error_ct <- function(rf) {
  cm <- rf$confusion
  sum(cm[1:4,1:4]) - sum(diag(cm))
}

# Number of holdout set misclassifications
h_error_ct <- function(rf) {
  pr <- predict(rf,h) == h$Is_cancer
  length(pr[!pr])
}
```

```r
mt <- c(40,20,10,5,2,1)

te <- c()
he <- c()
for (i in mt) {
  rf <- randomForest(Is_cancer~.,data=t,mtry = i)
  tr_error <- t_error_ct(rf)
  ho_error <- h_error_ct(rf)
  te <- c(te,tr_error)
  he <- c(he,ho_error)
}

info <- tibble(mtry = mt,
               training_misclass = te,
               holdout_misclass = he,
               total_misclass = te + he) %>%
  pivot_longer(-mtry)

ggplot(info,aes(x=mtry,y=value,group=name,color=name)) +
  geom_line() +
  geom_point()
```


value
mtry
name
holdout_misclass
total_misclass
training_misclass