Technical Report

Kanyin Olagbegi and Will Lonnquist

12/8/2018

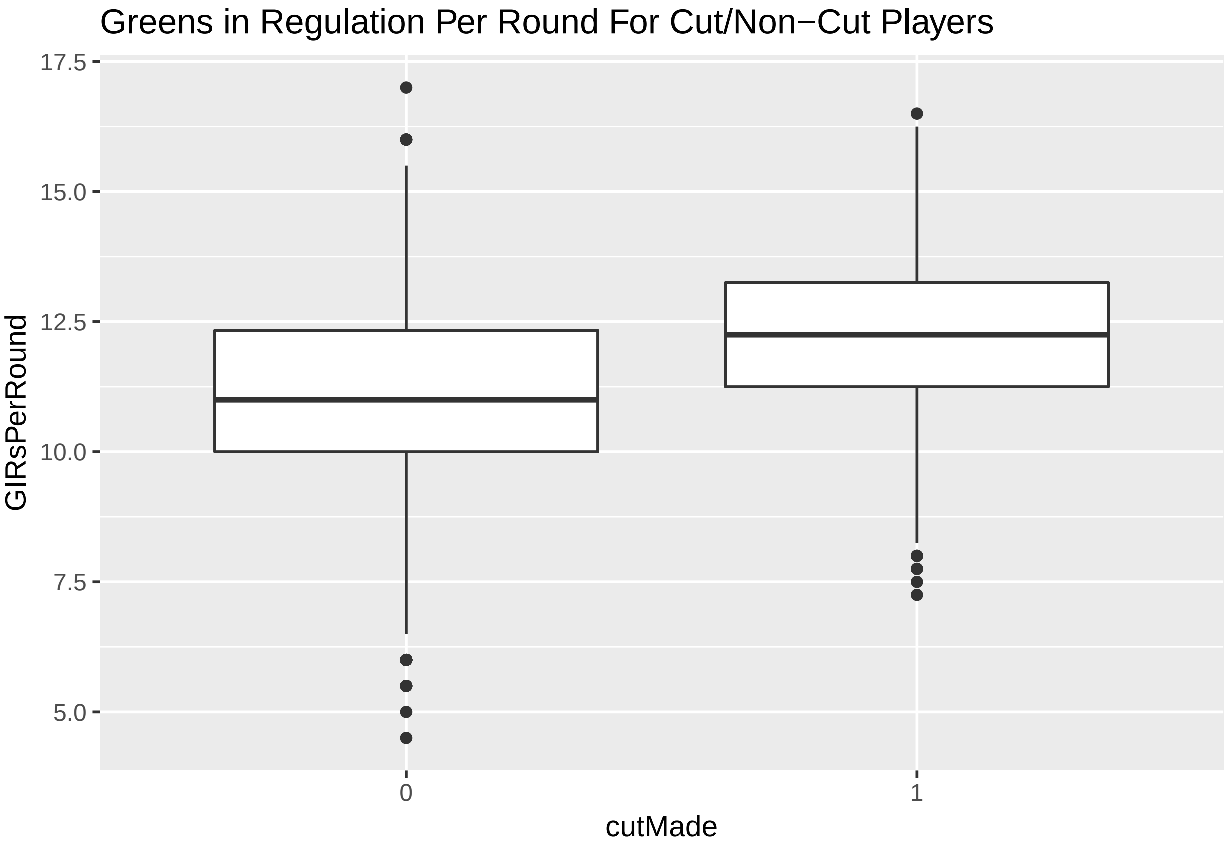
```r
library(readr)
library(dplyr)

data <- read_delim('revent.TXT', delim = ";")

dataFinal <- data %>%

  # Filtering for PGA TOUR Stroke Play Events
  filter(`Official Event(Y/N)` == "Y") %>%

  # Changing variables to the proper types
  mutate(`Total Rounds` = as.numeric(`Total Rounds)`),
         `Finish Position(numeric)` = as.numeric(`Finish Position(numeric)`),
         `Birdies` = as.numeric(`Birdies`),
         `Total Holes Over Par` = as.numeric(`Total Holes Over Par`),
         `Drives Over 300 Yards (# of Drives)` = as.numeric(`Drives Over 300 Yards (# of Drives)`),
         `3-Putt Avoid(Total 3 Putts)` = as.numeric(`3-Putt Avoid(Total 3 Putts)`),
         `Avg Distance of Putts Made(Total Distance of Putts)` = as.numeric(`Avg Distance of Putts Made
         `Total Holes Played` = as.numeric(`Total Holes Played`),
         `Total Greens in Regulation` = as.numeric(`Total Greens in Regulation`),
         `App. 50-125 Yards(ft)` = as.numeric(`App. 50-125 Yards(ft)`),
         `App.  50-125 Yards(attempts)` = as.numeric(`App.  50-125 Yards(attempts)`),

         #Creating our desired variables
         cutMade = as.factor(ifelse(`Finish Position(numeric)` < 999, 1, 0)),
         birdiesPerRound = `Birdies` / `Total Rounds`,
         GIRsPerRound = `Total Greens in Regulation` / `Total Rounds`,
         # GIRsPerRound = `Total Greens in Regulation` / `Total Holes Played`,
         overParHolesPerRound = `Total Holes Over Par` / `Total Rounds`,
         ThreePuttsPerRound = `3-Putt Avoid(Total 3 Putts)` / `Total Rounds`,
         over300DrivesPerRound = `Drives Over 300 Yards (# of Drives)` /
           `Total Rounds`,
         distPuttsMadePerRound = `Avg Distance of Putts Made(Total Distance of Putts)` / `Total Rounds`
         proxToHoleApproach = `App. 50-125 Yards(ft)` / `App.  50-125 Yards(attempts)`)  %>%

  #Selecting our desired columns
  select(`Player Name`,
        `Event Name`,
        cutMade,
        birdiesPerRound,
        GIRsPerRound,
        overParHolesPerRound,
        ThreePuttsPerRound,
        over300DrivesPerRound,
        distPuttsMadePerRound,
        proxToHoleApproach)
```

```
# Removing observations with missing values
# Source for code: https://stackoverflow.com/questions/4862178/remove-rows-with-all-or-some-nas-missing
dataFinal <- dataFinal[complete.cases(dataFinal), ]

head(dataFinal)
```

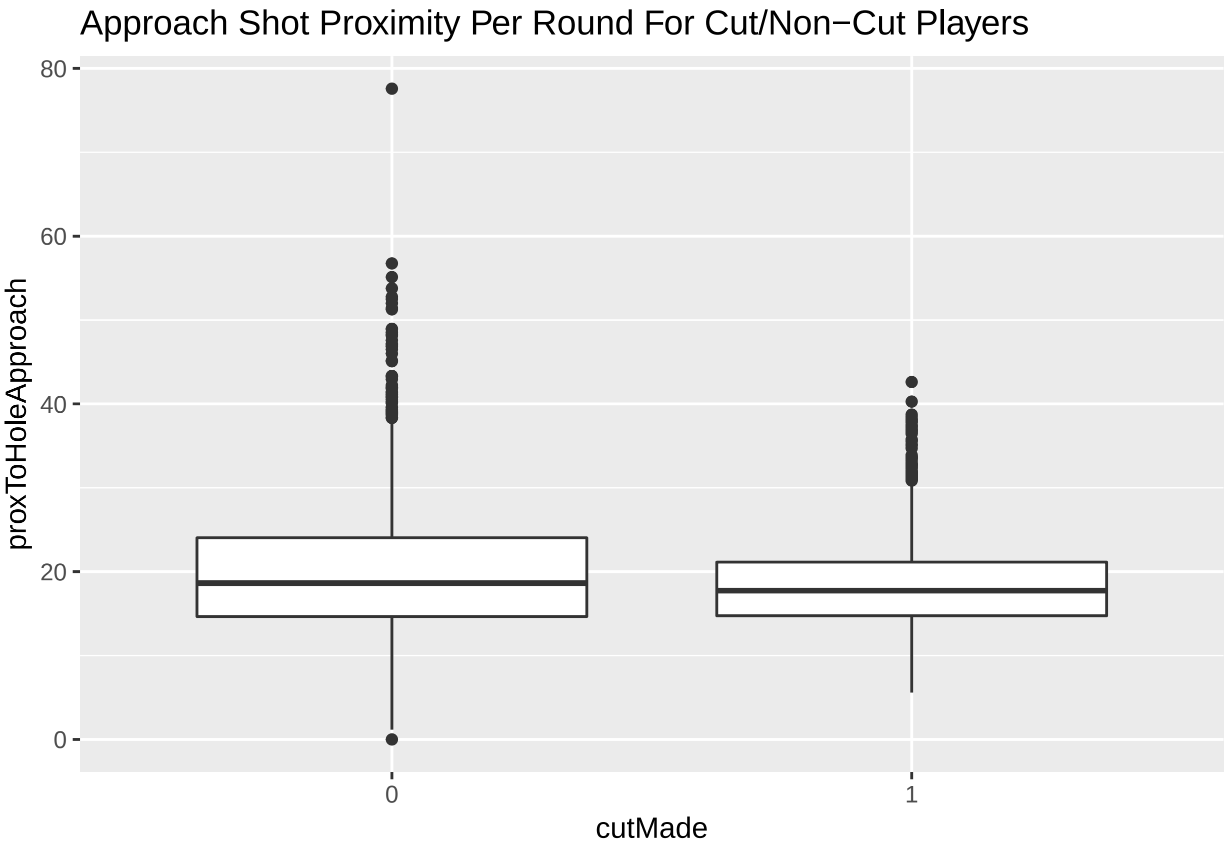
```
## # A tibble: 6 x 10
##   `Player Name` `Event Name` cutMade birdiesPerRound GIRsPerRound
##   <chr>         <chr>        <fct>             <dbl>        <dbl>
## 1 Allan, Steve  Safeway Open 0                  1.5          13
## 2 Ancer, Abrah~ Safeway Open 1                  3.25         11.8
## 3 Armour, Ryan  Safeway Open 0                  3            12.5
## 4 Atkins, Matt  Safeway Open 0                  3             9
## 5 Axley, Eric   Safeway Open 0                  2.5          10
## 6 Baddeley, Aa~ Safeway Open 0                  3.5          11.5
## # ... with 5 more variables: overParHolesPerRound <dbl>,
## #   ThreePuttsPerRound <dbl>, over300DrivesPerRound <dbl>,
## #   distPuttsMadePerRound <dbl>, proxToHoleApproach <dbl>
```

```
# nrow(dataFinal)
```

```
# Exploratory Data Analyis
library(ggplot2)

ggplot(dataFinal, aes(x=cutMade, y=GIRsPerRound)) +
  geom_boxplot() +
  ggtitle("Greens in Regulation Per Round For Cut/Non-Cut Players")
```



```r
ggplot(dataFinal, aes(x=cutMade, y=proxToHoleApproach)) +
  geom_boxplot() +
  ggtitle("Approach Shot Proximity Per Round For Cut/Non-Cut Players")
```



```r
# Splitting data into train and test subsets

# split 80/20 --------------------------
set.seed(123)
n <- nrow(dataFinal)
train_id <- sample(1:n, size=round(n*0.8)) # select approx 80% of the row numbers between 1 and n
train1 <- dataFinal[train_id,] # the data set we'll train the model on
test1 <- dataFinal[-train_id,] # the data set we'll test the model on
```

```r
# Building Random Forest
library(randomForest)
library(pROC)

# Setting formula for random forest
f2 <- as.formula(cutMade ~ GIRsPerRound + ThreePuttsPerRound + over300DrivesPerRound + distPuttsMadePer

# Training forest
set.seed(500)
mod_forest2 <- randomForest(f2, data = train1, ntree = 300, mtry = 2)
mod_forest2
```

```
##
## Call:
##  randomForest(formula = f2, data = train1, ntree = 300, mtry = 2)
##                Type of random forest: classification
##                      Number of trees: 300
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 16.5%
## Confusion matrix:
##      0    1 class.error
## 0 1349  286   0.1749235
## 1  345 1845   0.1575342
```

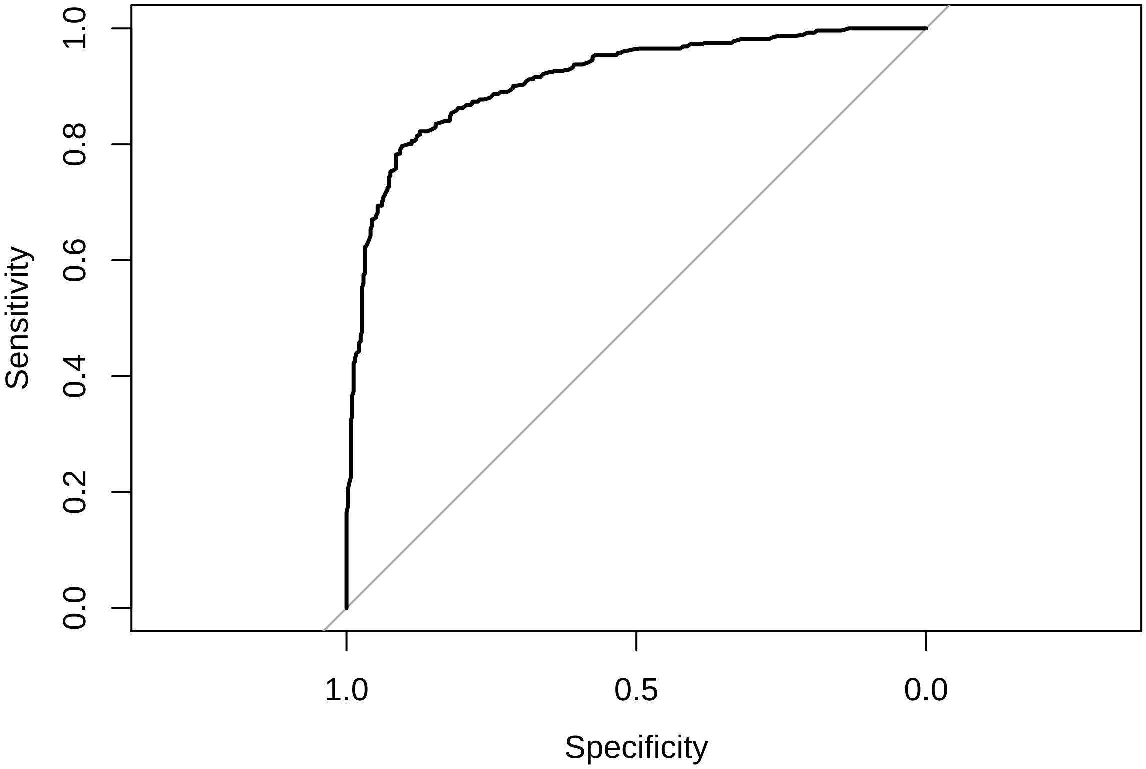
```
sum(diag(mod_forest2$confusion)) / nrow(train1)
```

```
## [1] 0.8350327
```

```
# predict on test and evaluate the model on test using auc-----------------------
pred_AUC2 <- predict(mod_forest2, test1, type = "prob")[,1]

test1 <- test1 %>% mutate(prediction2 = pred_AUC2)

roc_obj <- roc(test1$cutMade, test1$prediction2)
auc(roc_obj)
```

```
## Area under the curve: 0.9127
```

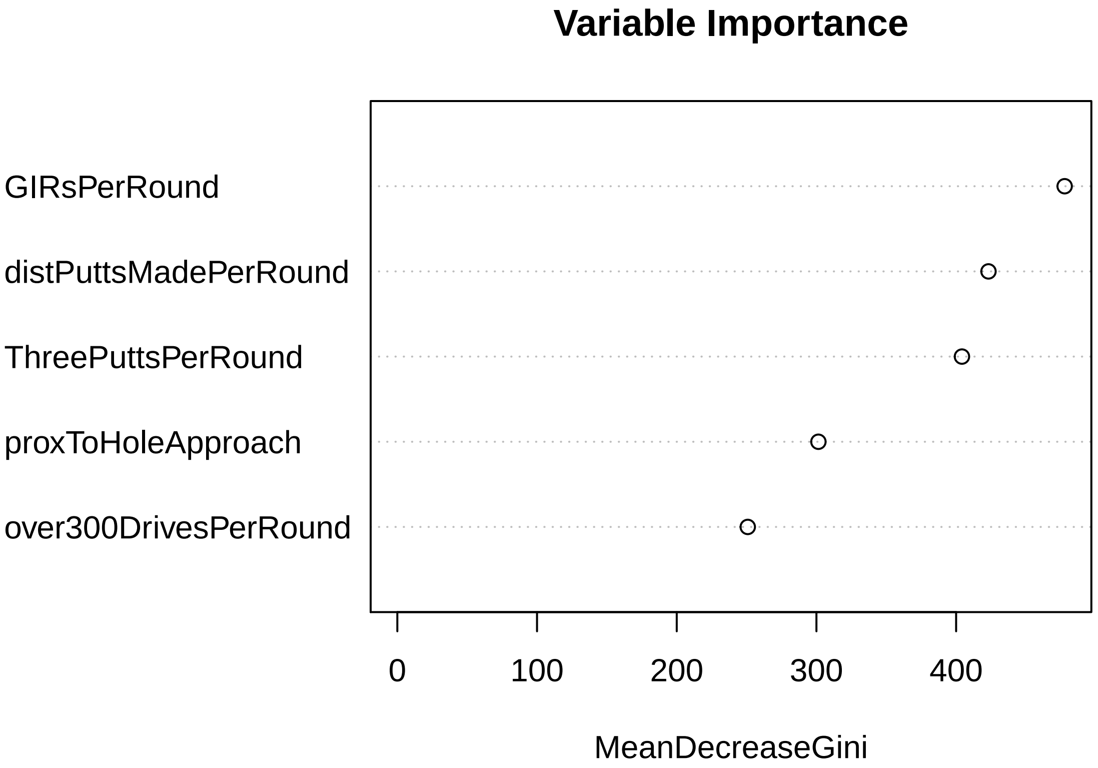
```
plot(roc_obj)
```



```
# Variable Importance

# Get variable importance, code from textbook
library(tibble)
importance(mod_forest2) %>%
  as.data.frame() %>%
  rownames_to_column() %>%
  arrange(desc(MeanDecreaseGini))
```

```
##                  rowname MeanDecreaseGini
## 1            GIRsPerRound         477.6903
## 2 distPuttsMadePerRound         423.1606
## 3     ThreePuttsPerRound         404.1596
## 4     proxToHoleApproach         301.4034
## 5 over300DrivesPerRound         250.8412
```

```
# Create Variable Importance Plot
varImpPlot(mod_forest2, main = "Variable Importance")
```


Variable Importance
GIRsPerRound
distPuttsMadePerRound
ThreePuttsPerRound
proxToHoleApproach
over300DrivesPerRound
MeanDecreaseGini