Technical Report

Kanyin Olagbegi and Will Lonnquist

12/15/2018

Project Title:

Factors Contributing to Success on the PGA TOUR

Abstract:

Our project analyzed player-based event statistics from all players who competed in 2018 PGA TOUR stroke play events to uncover characteristics which led to event success. We have defined success as making the cut in an event since players who make the cut and complete the event receive payment for their performance while players who miss the cut receive no payment. Using a random forest method to establish statistics associated with making, or missing, the cut, we discovered that players who maximize their number of greens in regulation per round and distance of putts made per round while minimizing their number of three-putts per round are more likely to have success in the event than players not excelling in these categories. These were the variables from our formula of all predictors that had the highest *variable importance* in the random forest. From these findings, we believe PGA Tour players should focus on improving their ball-striking and putting in order to maximize their greens-in-regulation and distance of putts made, respectively, in every event they compete in, thus increasing their likelihood of making the cut and earning money.

Project Introduction:

Our group approached this project with an interest in uncovering what leads professional golfers to success. Our goal for the investigation carried out below was to use raw player-based event statistics from all 2018 PGA (Professional Golfers Association) TOUR stroke play events to determine what characteristics of a player's performance contributed to their success. We defined success for a golfer in a PGA TOUR stroke play event as making the cut since players who make the cut and complete the event receive payment for their performance while players who miss the cut receive no payment. The cut in an event occurs after the first two rounds of play. Typically, the top 60 golfers will make the cut and compete in another two rounds of play for a total of four rounds in the event. Our research into this topic may be important for PGA TOUR players as they can better understand what statistics they should focus on maximizing (or minimizing) during their rounds so they can improve their chances of making the cut and making money. Also, our research may be valuable to fans of the PGA TOUR so they can understand why certain players are making more cuts than others.

We created a model, specifically a random forest, that predicts the outcome for a player of making or missing the cut in a probabilistic manner based on several predictor variables taken from their play in the event. Random forests are a collection of decision trees, which take in variables in the dataset and create cutoff levels for those variables in order to make a prediction about the response variable. In our case the response variable was whether or not a cut was made (binary) and our five predictors variables were average number of greens in regulation per round (quantitative), average number of three-putts per round (quantitative), average total distance of putts made per round (quantitative), average number of drivers over 300 yards per round (quantitative), and average distance to the hole on approach shots from 50-125 yards (quantitative).

We chose to use a random forest for two reasons. First, random forests generated a metric known as variable importance which associates a value with each predictor variable telling of that predictor variable's influence on the response variable. The higher the variable importance value the more influential that variable was on limiting the error in the model's predictions. Variable importance was useful for our investigation to determine and rank which statistics influenced whether or not a cut was made the most. Second, using a random forest classification method was beyond on the scope of the course and forced us to explore topics not covered in class, as was a requirement for the project.

After fitting our model we examined the differences in variable importance of our predictor variables. The differences in variable importance between our five predictor variables showed some trends that we believe may be meaningful to the desired audiences of our project, PGA TOUR players and fans. The specifics of our discoveries are outlined below.

```r
# Loading all packages required for analysis
library(readr)
library(dplyr)
library(ggplot2)
library(randomForest)
library(pROC)
library(tibble)
library(mosaic)
```

Data:

We gained access to the PGA TOUR's Shotlink Intelligence Program which contains data about PGA TOUR players, courses, and events. We exported the 2018 Event Detail file which contains player-based event statistics for all PGA TOUR sponsored events of the 2018 wrap-around season (some events played at the end of 2017 were included as they count towards the 2018 season). The dataset was downloaded as a semicolon-delimited text file and uploaded into R. This dataset had 6173 observations of 203 variables covering all aspects of a golfer's play in an event. An observation in this dataset represented a specific player's statistics from an event. For example, statistics about Tiger Woods' performance at the 2018 Tour Championship could be found in one row of the dataset.

We modified this large dataset to make it more suitable for our needs. First, we filtered out events that were not official PGA TOUR stroke play events such as the Ryder Cup, an international match play event that the PGA TOUR still sponsors. We then modified the variable types of the dataset. When we uploaded the original dataset into R all variables were designated as characters, even though many of them were meant to be numerics. Then, we transformed the raw statistics present in the dataset, typically from a total occurence count into a per-round average occurence count. For example, we took the total number of three-putts a player had in an event and divided that count by the total number of rounds played in that event by that player. This allowed us to account for increases to certain counts that may occur from players competing in more rounds than others. Next, we selected our newly transformed variable columns. Finally, we took out observations from the dataset that had missing values since we read online that random forests work better without any missing values. After wrangling the original dataset, we were left with a dataset of 4786 observations and 8 variables.

Codebook for our variables:

Response variable:
cutMade - a binary variable with value 0 for a player who missed the cut and 1 for players who made the cut. Players strive to make the cut.

Predictor variables:
1. GIRsPerRound - A quantitative variable detailing the average number of greens in regulation hit by a player per round in an event. A green in regulation occurs when a player's golf ball is on the green in two strokes less than the par for the hole. If a player's ball was on the green of a par 4 in 2 strokes they would have hit that green in regulation. Players strive to hit the green in regulation on every hole they play.
2. ThreePuttsPerRound - A quantitative variable detailing the average number of three-putts by a player per round in an event. A three-putt occurs when a player takes three strokes on the green to get their ball into the hole. Players strive to avoid three-putts.
3. over300DrivesPerRound - A quantitative variable detailing the average number of drives over 300-yards hit by a player per round in an event. Hitting the ball far on a shot means the next shot will begin closer to the hole and will thus be easier. 300-yards is an accepted threshold for what classifies as a far or long drive. Players strive to hit their drives over 300-yards when possible.
4. distPuttsMadePerRound - a quantitative variable detailing the average total distance in feet of all putts

holed by a player per round in an event. Making, or holing, a putt becomes exponentially harder as a player moves farther away from the hole so making long putts is an impressive task. Players strive to make every putt they have, but just making a few more long putts over the course of an event will add up.
5. proxToHoleApproach - a quantitative variable detailing the average distance in feet that a player's ball winds up from the hole on approach shots from 50-125 yards per round in a event. Hitting the ball closer to the hole will give the player a better chance of making their upcoming putt. Players strive to hit the ball as close as possible to the hole from this approach shot distance range.

Misc. variables:
1. `Player Name` - The name of the PGA TOUR player to whom the accompanying variables belong.
2. `Event Name` - The name of the PGA TOUR event from which the accompanying statistics were collected.

```
# Reading in dataset accessed from pgatourhq.com
data <- read_delim('revent.TXT', delim = ";")
# Dimensions of original dataset
nrow(data)
```

```
## [1] 6173
```

```
ncol(data)
```

```
## [1] 203
```

```
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)`),
         `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)),
         GIRsPerRound = `Total Greens in Regulation` / `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,
        GIRsPerRound,
        ThreePuttsPerRound,
        over300DrivesPerRound,
        distPuttsMadePerRound,
        proxToHoleApproach)
```

```r
# 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), ]

# First six rows of modified dataset
head(dataFinal)
```

```
## # A tibble: 6 x 8
##   `Player Name`   `Event Name` cutMade GIRsPerRound ThreePuttsPerRound
##   <chr>           <chr>        <fct>          <dbl>              <dbl>
## 1 Allan, Steve    Safeway Open 0              13.0               0.500
## 2 Ancer, Abraham  Safeway Open 1              11.8               0.
## 3 Armour, Ryan    Safeway Open 0              12.5               1.00
## 4 Atkins, Matt    Safeway Open 0               9.00              0.
## 5 Axley, Eric     Safeway Open 0              10.0               0.500
## 6 Baddeley, Aaron Safeway Open 0              11.5               2.00
## # ... with 3 more variables: over300DrivesPerRound <dbl>,
## #   distPuttsMadePerRound <dbl>, proxToHoleApproach <dbl>
```

```r
# Dimensions of modified dataset
nrow(dataFinal)
```

```
## [1] 4786
```

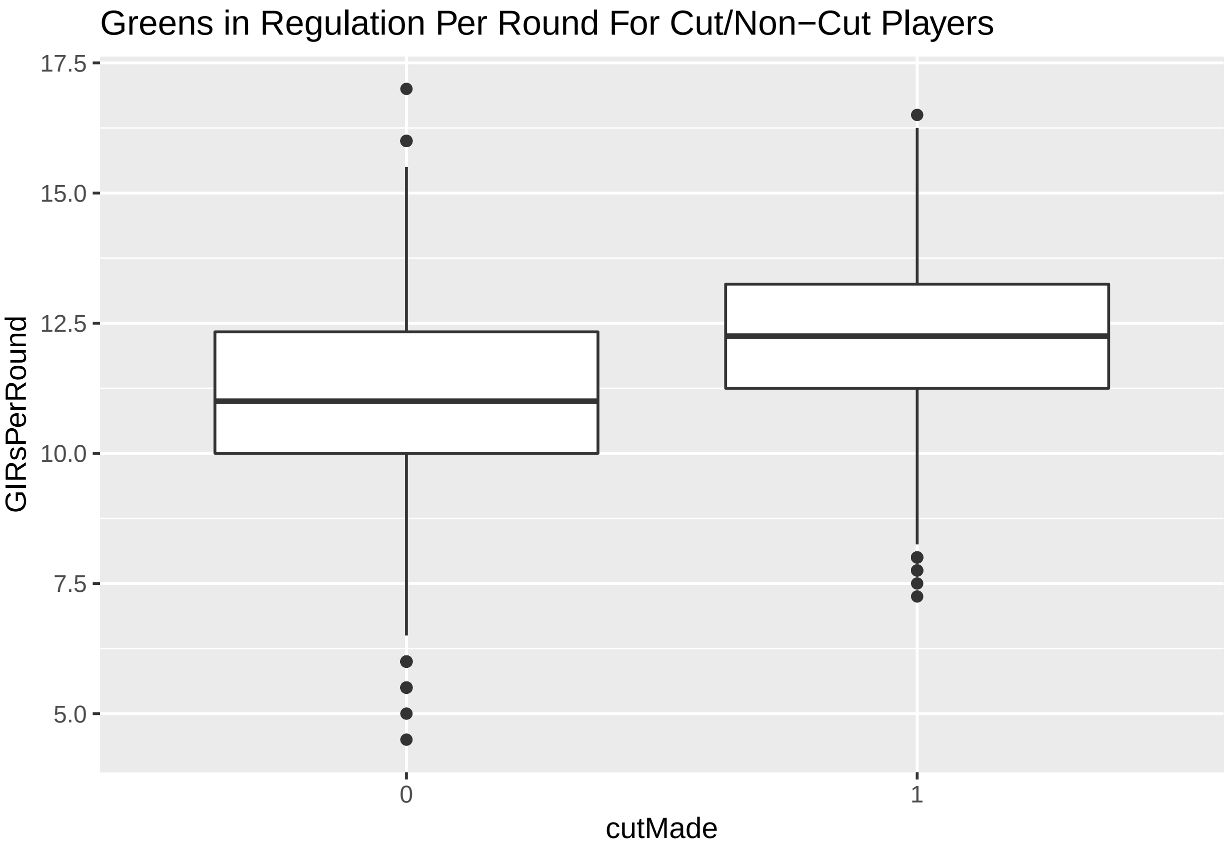
```r
ncol(dataFinal)
```

```
## [1] 8
```

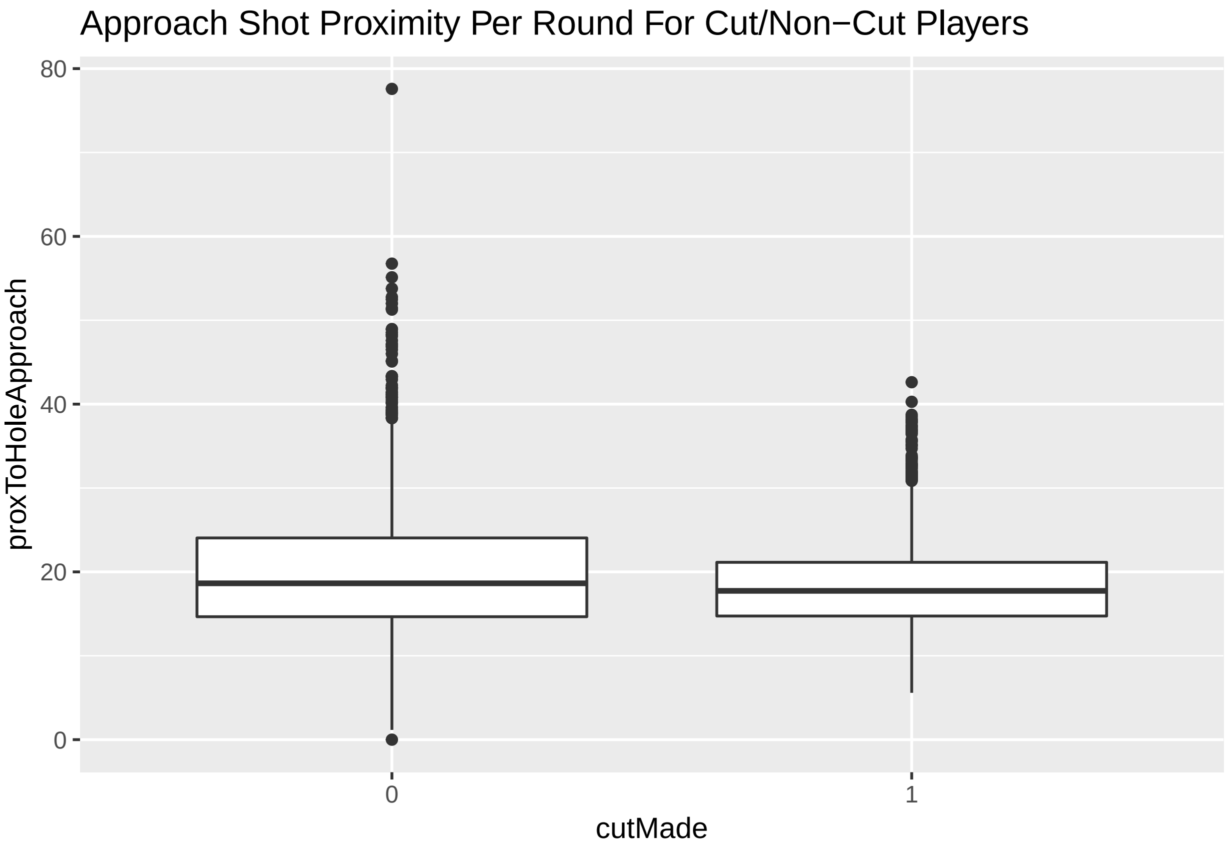
```r
# Exploratory Data Analysis

# Total players who missed the cut and made the cut in the 2018 season
tally(~ cutMade, data= dataFinal)
```

```
## cutMade
##    0    1
## 2050 2736
```

```r
# Looking for relationships between predictor variables and cutMade
ggplot(dataFinal, aes(x=cutMade, y=GIRsPerRound)) +
  geom_boxplot() +
  ggtitle("Greens in Regulation Per Round For Cut/Non-Cut Players")
```



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



After completing some exporatory data analysis on our variable, we plotted some of the interesting initial trends that we noticed. Above are the distributions of two different predictor variables split between groups where players missed or made the cut.

The distributions on the plot with cutMade and GIRsPerRound seem to be much different. While the two distributions generally seem to cover the same value range, roughly 7-16, they do not appear to be identical as players making the cut have higher 25th percentile, median, and 75th percentile values. Based on the visuals of the plot, the distribution for players missing the cut appears to be roughly 1 green in regulation lower per round than the distribution for players making the cut at the 25th percentile, median, and 75th percentile.

The distributions on the plot with cutMade and proxToHoleApproach do not appear very different. There is less spread on the distribution of players who made the cut especially on the upper end compared to players missing the cut, but the quartile and median values are all very close. Players who make the cut seem to hit the ball slightly closer to the hole from 50-125 yards on average than players who don't, yet this difference isn't stark.

These plots leads us to believe that going forward GIRsPerRound will be a useful predictor variable since the plot shows a clear difference in distributions between players who make or miss the cut in the event. Comparatively, we expect proxToHoleApproach will be a less useful predictor variable since the distribution between players who made or missed the cut are less obvious. From this exploratory data analysis we believe GIRsPerRound will have a higher variable importance value in our model than proxToHoleApproach.

```
# 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
```

```r
test1 <- dataFinal[-train_id,] # the data set we'll test the model on
```

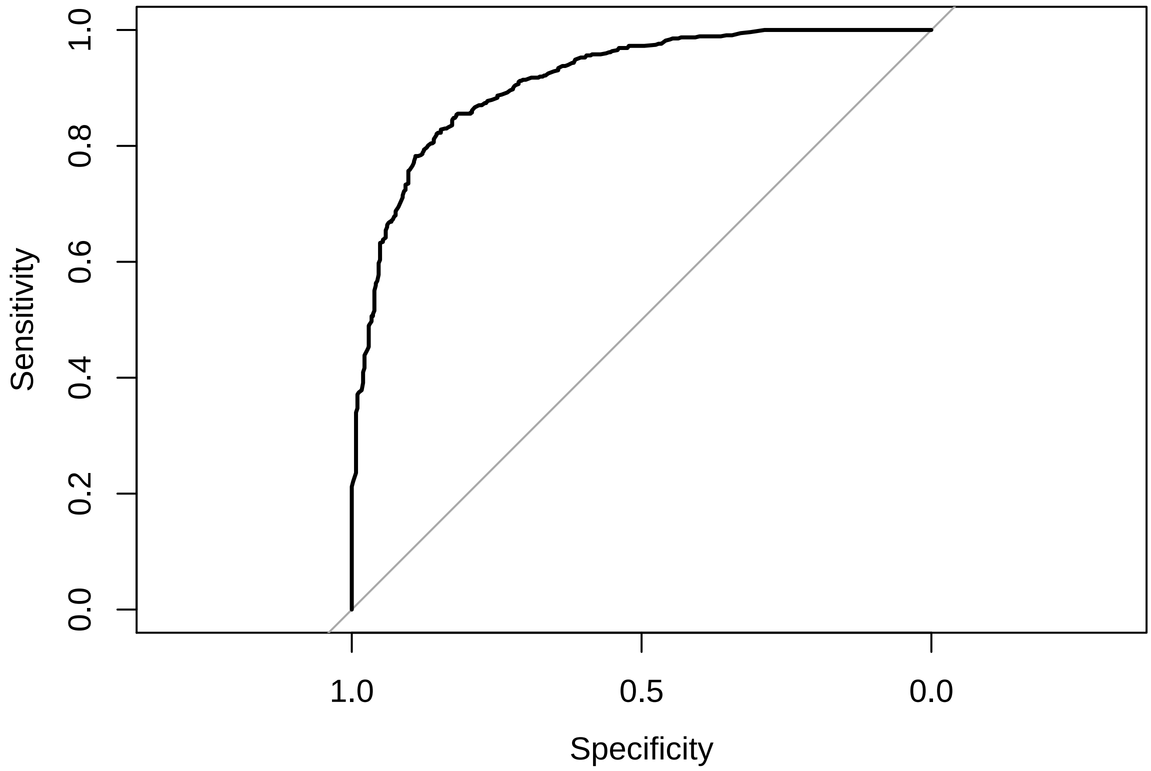
```r
# Building Random Forest

# 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: 16.9%
## Confusion matrix:
##      0    1 class.error
## 0 1343  297   0.1810976
## 1  350 1839   0.1598904
```

```r
# 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: 0.9131
```

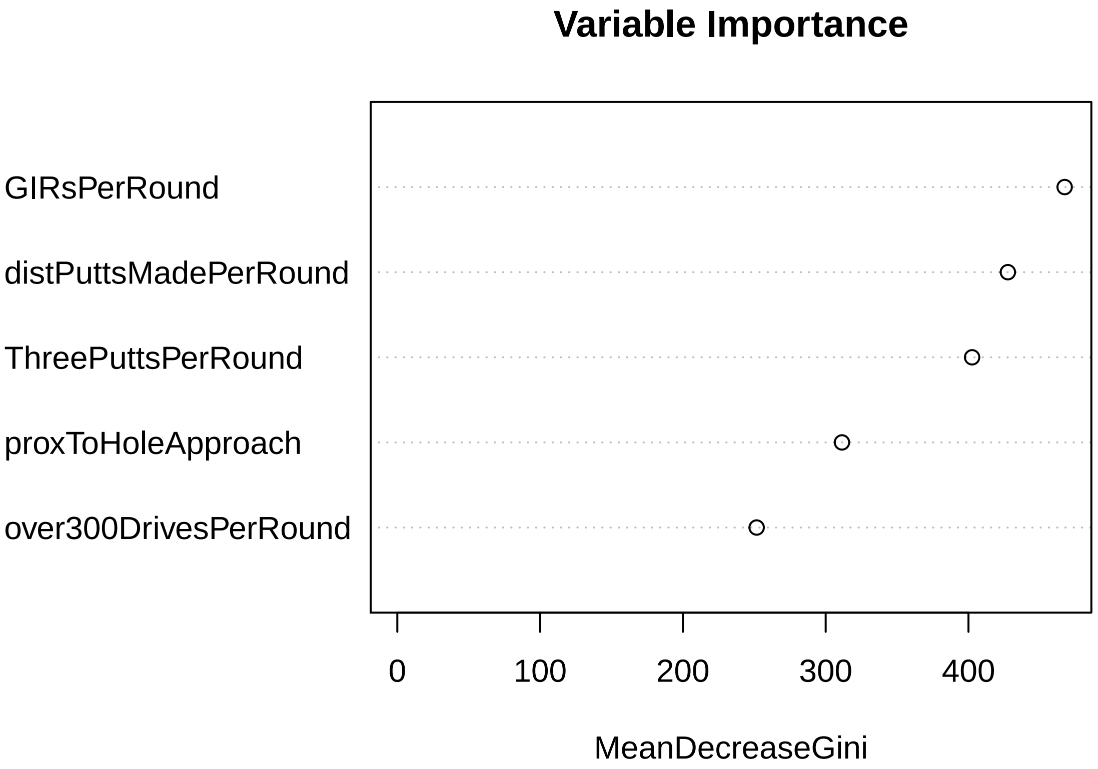
```r
plot(roc_obj)
```



```r
# Variable Importance

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

```
##                  rowname MeanDecreaseGini
## 1           GIRsPerRound         467.3577
## 2 distPuttsMadePerRound          427.5810
## 3    ThreePuttsPerRound          402.5273
## 4    proxToHoleApproach          311.3818
## 5 over300DrivesPerRound          251.6449
```

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



Results:

We built a random forest on the randomly selected training subset from our dataset. After creating our random forest we investigated the variable importance of our different predictors. Calculating the variable importance involves a complex formula accounting for changes in the gini coefficients of the predictor variable. Essentially, variable importance measures how much better the decision trees in the random forest perform with and without each predictor variable in the tree. Higher MeanDecreaseGini values, the name of the metric representing variable importance, imply that the predictor variable is more influential on the outcome predicted by the random forest. The MeanDecreaseGini value itself for a variable matters less than how the value compares to those of the other predictors so that a relative importance can be established. For more information about variable importance please look here: variable importance calculation.

The variable importance of our five predictor variables in order from highest to lowest are as follows: GIRsPerRound (467.36), distPuttsMadePerRound (427.58), ThreePuttsPerRound (402.53), proxToHoleApproach (311.38), over300DrivesPerRound(251.64). While all five of our predictor variables have some influence on the outcome from the random forest based on their nonzero values, there appears to be two tiers of importance levels. The variables of average greens in regulation per round, average total distance of putts made per round, and average number of three putts per round all have mildly similar and large variable importance values. There is a dropoff of roughly 100 units until the last two variables of average proximity to the hole on approach shots from 50-125 yards and average number drives over 300-yards in terms of variable importance. We have labeled these tiers as “more important” and “less important” and focused our interpretions of the random forest on the “more important” variables.

The most important single predictor in our random forest was average number of greens in regulation per round. In our random forest, trends in the average greens in regulation hit by players was the most useful predictor variable in explaining whether or not a player was cut. This matches what we saw in our exploratory data analysis where there were clear differences in the distrubutions of cut and not-cut players in terms of their average greens in regulation per round. Players who are looking to make the cut in an event should focus on hitting as many greens in regulation as possible to maximize their chances of making the cut.

The average proximity to the hole on approach shots from 50-125 yards variable had a relatively low variable importance in our model. Again, this matches with what we saw in the exploratory data analysis as the two distributions for cut and non-cut players were not extremely different.

Looking at the variable importance of our predictors gave us the statistical background to proceed with interpretations about what PGA TOUR players should be doing on the course in order to maximize their chance of making the cut. We know that not everyone will dedicate the time to understand how our random forest model compiled its results, but by putting those results in a non-statistical, golfing context using the ordering found in the variable importance of our predictors we can help our target audience understand our results.

Diagnostics:

Random forests are complicated models. They extend upon the decision tree method by creating many decision trees and evaluating the predictive trends that exist within the collective of trees by a majority rule. Each tree in the random forest has the same number of random variables, but obviously the specific variables differ between trees. Additionally, each tree sample data from the input set with replacement, this means that some observations may be used multiple times in creating a tree. When using a random forest, a prediction for a new observation is made by passing in the values of the new observation and using the forest's majority rule to establish a predicted outcome, in our case the chances that a player made the cut ranging between 0 and 1.

We fit several different models before settling upon the final iteration which was used for this analysis. Each of our different model drafts had either a different numbers of trees (ntree argument) in the random forest or a different number of random variables to be selected from each tree (mtry argument). We settled upon a model with 300 trees and 2 random variables for each tree. We chose this model because of the different models we fit it had the best combination of a high AUC value and a low out-of-bag estimate. The AUC value (Area Under the Receiver Operating Curve) represents how well any classifier method places its predictions for every cutoff threshold between 0 and 1. Values for AUC range between 0.5 and 1 with 1 being the best possible value and meaning the classifier method did not make a single incorrect prediction on any input at any threshold. Our random forest model achieved an AUC value of 0.9131. Since this value is much closer to 1 than it is to 0.5 we were pleased with how our model was performing. The out-of-bag estimate in a random forest is a measure of error within the model. Since samples for each tree are taken from the training subset of the dataset *with replacement*, in every decision tree within the random forest there will be observations in the training subset not used to create the tree. To calculate the out-of-bag estimate these observations which were not included in the tree, known as out-of-bag, are run through the tree and the error between their probabilistic outcome from the tree and their actual value is calculated. These errors from each tree are then aggregated to get the out-of-bag estimate. Our random forest model achieved an out-of-bag estimate of 16.9%. We felt this value was slightly high, but given that most of our other models had even higher out-ofbag estimates we stuck with our 300-tree, 2 random variable model.

While our model was not perfect at predicting whether or not a player made the cut, we were satisfied with its performance as a useful predictive tool on 2018 PGA TOUR data.

Conclusions:

In this project, we hoped to determine the most important attributes needed to be successful as a professional golfer. Success was defined as making the cut in an event as making the cut allows players to make money. We used data from all PGA TOUR stroke play events in the 2018 wrap-around season. From this data we selected five predictor variables that we felt might impact whether or not a player made the cut. Those variables were: average number of greens in regulation per round (quantitative), average number of three-putts per round (quantitative), average total distance of putts made per round (quantitative), average number of drives over 300 yards per round (quantitative), and average distance to the hole on approach shots from 50-125 yards (quantitative).

After creating a random forest with these five predictors to model whether or not a cut was made by a player, we obtained the variable importance for each of the variables. The predictor with the highest variable importance was tied to players hitting the ball efficiently from the tee to the green (average number of greens in regulation per round) and the second and third predictors in terms of variable importance both had to do with players putting the ball efficiently on the green (average total distance of putts made per round, average number of three-putts per round). The variable importance rankings of our predictors tell us that for PGA TOUR players to make the cut during the 2018 season they needed to be hitting the ball efficiently from tee-to-green and also putting efficiently when they are on the green. Players need to have a sharp, well-rounded game to make the cut.

The remaining two predictor variables of average distance to the hole on approach shots from 50-125 yards and average number of drives over 300 yards per round still have some explanatory power regarding whether or not a player made the cut since their MeanDecreaseGini values are nonzero. Yet, since these two variables were the lowest in terms of rankings their variable importance we feel players should not focus their attention on hitting the ball close to the hole on approach shots or blasting their driver over 300 yards on every tee shot. Instead, players should make smart, calculated decisions that allow them to hit greens in regulation since that variable had more influence on our random forest.

Based on our analysis, we recommend that PGA TOUR players practice their ball striking, or full swings, so that they can hit the golf ball well from tee-to-green and increase their number of greens in regulations. We also suggest that PGA TOUR players practice putting to eliminate their three-putts and increase their total distance of putts made per round. The PGA TOUR is a very competitive professional golf tour and making the cut in an event is an impressive feat. To make the cut consistently PGA TOUR players in 2018 needed consistent ball-striking and putting. Ball striking and putting are to of the main aspects of golf. While our findings are not ground-breaking, they reaffirm what is generally thought of as two important aspects of golf. Although we only used data from 2018, since we feel there are no differences to how golf on the PGA TOUR functioned last season to the current season, we recommend that PGA TOUR players competing in 2019 focus on maximizing their number of greens in regulation per round and distance of putts made per round while minimizing their number of three putts per round in order to give themselves the best chance to make the cut.

Unfortunately, there are several limitations to our analysis. First, we manually chose the five predictor variables that would explain whether or not a player made the cut. We may have selected predictors that are hardly associated with making the cut and we would not know since the variable importance values depend on what predictors are being used. We tried to create predictor variables from the inital raw dataset that would account for different aspects of a player's game and we believe we did a good job, but we may have failed in this regard. Second, greens in regulation is a combination statistic of multiple shots on certain holes. For example, on a typical par 4 a green in regulation would be hit only after a tee shot *and* an approach shot. All of our other variables are centered around singular shots during a round (tee shots, approach shots, putts), but this is not the case greens in regulation. We used greens in regulation in our model because we speculated the differences in values for players would have a strong association with whether or not the player made the cut, which it did based on it having the highest variable importance, but it's a statistic that may belong in a different category than the rest of our predictors. Third, we only trained and tested our random forest on data from the 2018 PGA TOUR wrap-around season. We would have liked to test our models on data from other past seasons or the current season thus far, but we were unable to obtain the necessary data. We should be very cautious extrapolating our results to other PGA TOUR seasons or other professional golf tours such as the European Tour, for those results may be much different. With that being said, we would expect to find very similar results in terms of the order of variable importance that our predictors held due to the nature of professional golf. Also, Our analysis did not include any predictor variables on chipping or pitching (short shots from just off the green) and this would be something we would likely include in further analysis. This is another important aspect of golf that we cannot comment on since none of our predictor variables are associated with chipping or pitching. Finally, the backbone of model is that success on the PGA TOUR is defined as making the cut. Making the cut is only one way of defining success in an event. Many players, particularly the superstars like Jordan Spieth or Rory McIlroy, would not consider an event successful is they merely made the cut. They would only consider it successful if they finished in the top 10, top 5, or even won the event. Our model does not let us determine what separates the players finishing at the top of tournament leaderboards over the rest of the field. Our analysis only lets understand what is needed to make the cut in an event, a very basic definition of success in a tournament. For this reason our analysis will likely not be helpful for the top players on the PGA TOUR. Still, for a struggling player attempting to make a living by making cuts our information will be relevant.

Sources:

Baumer, Benjamin, et al. Modern Data Science with R. Taylor & Francis CRC Press, 2017.

Breiman, Leo, and Adele Cutler. “Random Forests.” Statistics at UC Berkeley | Department of Statistics, www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

Louppe, Gilles, et al. Understanding Variable Importances in Forests of Randomized Trees. Dept. of EE & CS, University of Liege, Belgium, papers.nips.cc/paper/4928-understanding-variable-importances-in-forests-of-randomized-trees.pdf.