How Can I Get Better at Golf?

Will Lavey

2024-05-20

How Can I Get Better at Golf?

While brainstorming ideas for this final project, I thought to myself: How can I leverage the computer to create something practical and valuable for myself? What aspects in my life do I want to learn more about and improve upon?

And after some consideration, I landed on golf. **How can I get better at golf?**

Golf is a seemingly infinitely difficult sport filled with so many complex physics concepts to disturb your perfect strike and yet so many simple mental tricks to help you get better. What does the computer have to say about what makes one a great golfer? Is it your ability to hit a baby draw 200 yards with your 7 iron? Is it your ability to minimize three putting? Data may not be able to perfectly capture what determines the perfect golf game, but hopefully can point us in the right direction.

The data to be used are two PGA Tour data sets. One is data from each tournament in the seasons between 2015 and 2022. The other is all players' aggregate statistics over the entirety of the 2017 season. One of the most important metrics in golf is called Strokes Gained. The number relies on the average number of shots a player takes to get the ball in the cup from one's current position on the hole, which is data that has been accumulated over time. If a player hits an unusually good shot from a given position, their Strokes Gained index will be positive, and vice versa.

In golf, par is the average, or recommended number of strokes to get the ball in the hole. A course will typically have a par of 72, which consists of a variety of par 4's, par 3's and par 5's. Birdie is one under par (eg. 3 strokes on a par 4) and bogey is one over par (eg. 6 strokes on a par 5). The scoring in golf relates to how far over or under par (average) you are. The typical scores depend on the skill of golfer and the difficulty of course. PGA Tour professionals can typically shoot around 12 under par on a good day. Many metrics in the data sets relate to these scoring values.

```
# Tournament level data PGA Tour 2015-2022
tourney_level_2015_2022 <- read.csv("/Users/Will/Documents/College/Junior/Spring 2024/STAT218/Datasets/

# Player level data PGA Tour 2017
player_level_2017 <- read.csv("/Users/Will/Documents/College/Junior/Spring 2024/STAT218/Datasets/PGATOU
```

```
# Remove unnused columns
tourney_clean <- tourney_level_2015_2022
tourney_clean <- subset(tourney_clean, select = -c(Unnamed..2, Unnamed..3, Unnamed..4,
                                                   hole_DKP, hole_SDP, hole_FDP,
                                                   streak_DKP, streak_SDP, streak_FDP,
                                                   finish_DKP, finish_SDP, finish_FDP,
                                                   total_DKP, total_SDP, total_FDP))
```

```r
# Make pos 157 if missed cut. There are a max of 156 players allowed in a PGA tournament
tourney_clean$pos[is.na(tourney_clean$pos)] <- 157

# Remove NAs
tourney_clean <- na.omit(tourney_clean)
```

```r
player <- player_level_2017

# Remove NAs
player <- na.omit(player)

# Create 'won' column for if they have won a tournament
player <- player |>
  mutate(won = case_when(NUMBER_OF_WINS == 0 ~ 0,
                         NUMBER_OF_WINS > 0 ~ 1))


player$FAIRWAYS_HIT = as.numeric(player$FAIRWAY_HIT_.)
player$TOTAL_DRIVES <- as.numeric(gsub(",", "", player$TOTAL_DRIVES))

# Scaled data used for SVM
player_scale <- player |>
  mutate(across(-all_of(c("won", "Player")), scale))
```

```r
summary(tourney_clean)
```

```
##  Player_initial_last tournament.id         player.id          hole_par
##  Length:29180        Min.   :     2232   Min.   :      5   Min.   : 70.0
##  Class :character    1st Qu.:     2707   1st Qu.:   1185   1st Qu.:142.0
##  Mode  :character    Median :401056515   Median :   3950   Median :280.0
##                      Mean   :249366480   Mean   :  80107   Mean   :222.9
##                      3rd Qu.:401219800   3rd Qu.:   6689   3rd Qu.:284.0
##                      Max.   :401353273   Max.   :4845309   Max.   :292.0
##     strokes         n_rounds        made_cut           pos
##  Min.   : 66.0   Min.   :1.000   Min.   :0.0000   Min.   :  1.00
##  1st Qu.:146.0   1st Qu.:2.000   1st Qu.:0.0000   1st Qu.: 29.00
##  Median :271.0   Median :4.000   Median :1.0000   Median : 64.00
##  Mean   :221.5   Mean   :3.135   Mean   :0.5842   Mean   : 88.55
##  3rd Qu.:281.0   3rd Qu.:4.000   3rd Qu.:1.0000   3rd Qu.:157.00
##  Max.   :313.0   Max.   :4.000   Max.   :1.0000   Max.   :999.00
##     player          tournament.name       course              date
##  Length:29180       Length:29180       Length:29180       Length:29180
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      purse            season         no_cut           Finish
##  Min.   : 3.500   Min.   :2015   Min.   :0.00000   Length:29180
##  1st Qu.: 6.400   1st Qu.:2017   1st Qu.:0.00000   Class :character
##  Median : 7.100   Median :2019   Median :0.00000   Mode  :character
##  Mean   : 7.601   Mean   :2019   Mean   :0.05768
##  3rd Qu.: 8.400   3rd Qu.:2021   3rd Qu.:0.00000
##  Max.   :20.000   Max.   :2022   Max.   :1.00000
```

```
##     sg_putt           sg_arg             sg_app           sg_ott
##  Min.   :-5.990   Min.   :-6.43000   Min.   :-9.2500   Min.   :-7.7400
##  1st Qu.:-0.770   1st Qu.:-0.45000   1st Qu.:-0.7400   1st Qu.:-0.4500
##  Median :-0.040   Median : 0.00000   Median : 0.0000   Median : 0.0500
##  Mean   :-0.121   Mean   :-0.04074   Mean   :-0.1018   Mean   :-0.0459
##  3rd Qu.: 0.630   3rd Qu.: 0.42000   3rd Qu.: 0.6400   3rd Qu.: 0.4800
##  Max.   : 4.430   Max.   : 3.17000   Max.   : 4.6700   Max.   : 2.7700
##      sg_t2g            sg_total
##  Min.   :-13.9500   Min.   :-13.6700
##  1st Qu.: -1.0800   1st Qu.: -1.3700
##  Median : -0.0100   Median : -0.1600
##  Mean   : -0.1883   Mean   : -0.3056
##  3rd Qu.:  0.9200   3rd Qu.:  1.0600
##  Max.   :  6.3000   Max.   :  8.5200
```

```
summary(player)
```

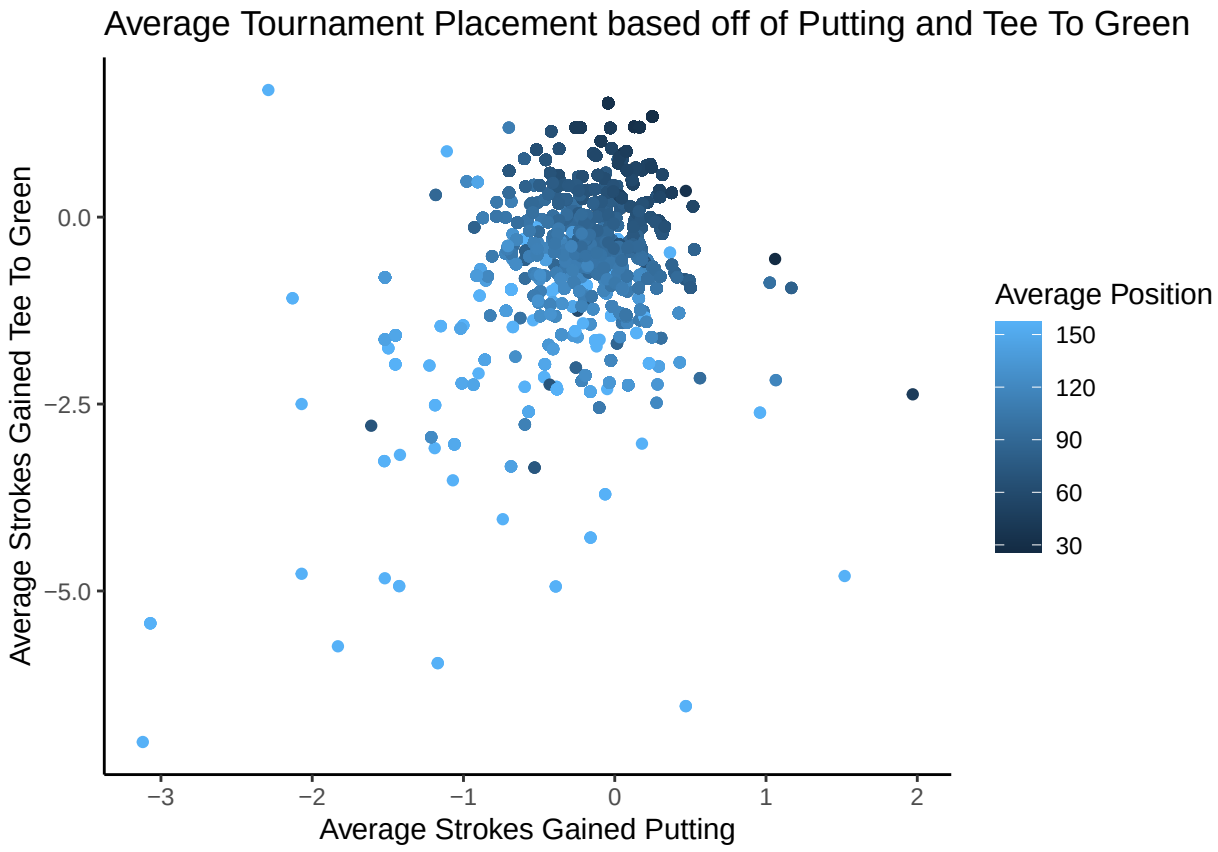
```
##     Player           EVENTS_PLAYED       POINTS       NUMBER_OF_WINS
##  Length:194         Min.   :15.00   Min.   :  10.0   Min.   :0.0000
##  Class :character   1st Qu.:21.00   1st Qu.: 267.2   1st Qu.:0.0000
##  Mode  :character   Median :24.50   Median : 580.0   Median :0.0000
##                     Mean   :24.18   Mean   : 720.9   Mean   :0.2268
##                     3rd Qu.:27.00   3rd Qu.:1024.2   3rd Qu.:0.0000
##                     Max.   :32.00   Max.   :3289.0   Max.   :3.0000
##  NUMBER_OF_TOP_Tens POINTS_BEHIND_LEAD ROUNDS_PLAYED    SG_PUTTING_PER_ROUND
##  Min.   : 0.000     Min.   :2328       Min.   : 46.00   Min.   :-0.75000
##  1st Qu.: 1.000     1st Qu.:4593       1st Qu.: 67.00   1st Qu.:-0.14775
##  Median : 2.000     Median :5037       Median : 80.00   Median : 0.05600
##  Mean   : 2.619     Mean   :4896       Mean   : 78.32   Mean   : 0.05029
##  3rd Qu.: 4.000     3rd Qu.:5350       3rd Qu.: 88.75   3rd Qu.: 0.23425
##  Max.   :11.000     Max.   :5607       Max.   :110.00   Max.   : 0.86200
##  TOTAL_SG.PUTTING  MEASURED_ROUNDS AVG_Driving_DISTANCE UP_AND_DOWN_.
##  Min.   :-42.673   Min.   :30.0    Min.   :278.4        Min.   :44.01
##  1st Qu.: -8.080   1st Qu.:49.0    1st Qu.:291.0        1st Qu.:56.47
##  Median :  3.913   Median :60.0    Median :296.2        Median :58.46
##  Mean   :  3.017   Mean   :59.9    Mean   :296.7        Mean   :58.58
##  3rd Qu.: 14.033   3rd Qu.:70.0    3rd Qu.:301.6        3rd Qu.:60.91
##  Max.   : 60.061   Max.   :92.0    Max.   :320.2        Max.   :66.59
##  PAR_OR_BETTER     MISSED_GIR    FAIRWAY_HIT_.    FAIRWAYS_HIT
##  Min.   :136.0   Min.   :227.0   Min.   :40.85   Min.   :40.85
##  1st Qu.:214.2   1st Qu.:373.2   1st Qu.:57.23   1st Qu.:57.23
##  Median :258.5   Median :445.0   Median :61.05   Median :61.05
##  Mean   :256.5   Mean   :437.9   Mean   :61.05   Mean   :61.05
##  3rd Qu.:295.0   3rd Qu.:492.0   3rd Qu.:64.44   3rd Qu.:64.44
##  Max.   :392.0   Max.   :635.0   Max.   :74.33   Max.   :74.33
##  POSSIBLE_FAIRWAYS    GIR_RANK      GOING_FOR_GREEN_IN_2.  ATTEMPTS_GFG
##  Min.   : 531      Min.   :  1.00   Min.   :28.36         Min.   : 38.0
##  1st Qu.: 856      1st Qu.: 50.25   1st Qu.:50.44         1st Qu.:100.2
##  Median :1031      Median : 98.50   Median :56.65         Median :123.5
##  Mean   :1014      Mean   : 98.29   Mean   :56.17         Mean   :123.9
##  3rd Qu.:1165      3rd Qu.:146.75   3rd Qu.:62.35         3rd Qu.:148.2
##  Max.   :1454      Max.   :195.00   Max.   :78.44         Max.   :227.0
##  NON.ATTEMPTS_GFG RTP.GOING_FOR_THE_GREEN RTP.NOT_GOING_FOR_THE_GRN
##  Min.   : 35.00   Min.   :-128.00         Min.   :-51.000
##  1st Qu.: 74.25   1st Qu.: -79.75         1st Qu.:-12.000
```

```
##  Median : 96.50   Median : -66.50         Median : -7.000
##  Mean   : 96.74   Mean   : -67.51         Mean   : -6.809
##  3rd Qu.:114.00   3rd Qu.: -55.00         3rd Qu.: -1.000
##  Max.   :228.00   Max.   : -22.00         Max.   : 27.000
##    HOLE_OUTS       SAND_SAVE.    NUMBER_OF_SAVES  NUMBER_OF_BUNKERS
##  Min.   : 1.00   Min.   :33.80   Min.   : 23.00   Min.   : 57.0
##  1st Qu.: 8.00   1st Qu.:45.96   1st Qu.: 50.25   1st Qu.:103.0
##  Median :10.00   Median :49.83   Median : 62.00   Median :125.0
##  Mean   :10.99   Mean   :49.94   Mean   : 61.98   Mean   :124.1
##  3rd Qu.:14.00   3rd Qu.:53.40   3rd Qu.: 73.00   3rd Qu.:143.8
##  Max.   :25.00   Max.   :66.00   Max.   :108.00   Max.   :193.0
##  TOTAL_O.U_PAR   Three_PUTT.    TOTAL_3_PUTTS    SG_PER_ROUND
##  Min.   :15.0   Min.   :1.480   Min.   :12.00   Min.   :-3.5860
##  1st Qu.:31.0   1st Qu.:2.402   1st Qu.:31.25   1st Qu.:-0.2208
##  Median :38.0   Median :2.840   Median :37.50   Median : 0.1425
##  Mean   :40.2   Mean   :2.864   Mean   :38.73   Mean   : 0.1298
##  3rd Qu.:49.0   3rd Qu.:3.190   3rd Qu.:45.00   3rd Qu.: 0.5543
##  Max.   :74.0   Max.   :5.090   Max.   :73.00   Max.   : 1.9870
##      SG.OTT             SG.APR            SG.ARG          DRIVES_320..
##  Min.   :-1.58500   Min.   :-1.58600   Min.   :-0.92400   Min.   : 1.240
##  1st Qu.:-0.14575   1st Qu.:-0.15125   1st Qu.:-0.12400   1st Qu.: 5.747
##  Median : 0.07150   Median : 0.05350   Median : 0.01950   Median : 8.715
##  Mean   : 0.03954   Mean   : 0.05873   Mean   : 0.03149   Mean   :10.773
##  3rd Qu.: 0.28400   3rd Qu.: 0.31500   3rd Qu.: 0.19825   3rd Qu.:13.810
##  Max.   : 1.00600   Max.   : 0.99000   Max.   : 0.63200   Max.   :33.920
##  TOTAL_DRIVES_FOR_320.  TOTAL_DRIVES    ROUGH_TENDNECY.  TOTAL_ROUGH
##  Min.   :  8.00        Min.   : 420.0   Min.   :16.67   Min.   : 93.0
##  1st Qu.: 48.00        1st Qu.: 701.0   1st Qu.:26.23   1st Qu.:198.5
##  Median : 74.50        Median : 862.0   Median :28.77   Median :242.5
##  Mean   : 91.73        Mean   : 858.9   Mean   :28.79   Mean   :243.9
##  3rd Qu.:119.75        3rd Qu.:1001.5   3rd Qu.:31.23   3rd Qu.:281.8
##  Max.   :307.00        Max.   :1344.0   Max.   :42.64   Max.   :428.0
##  FAIRWAY_BUNKER. TOTAL_FAIRWAY_BUNKERS AVG_CLUB_HEAD_SPEED FASTEST_CH_SPEED
##  Min.   :3.100   Min.   :18.00         Min.   :105.3       Min.   :109.2
##  1st Qu.:5.400   1st Qu.:41.00         1st Qu.:111.3       1st Qu.:115.5
##  Median :6.100   Median :52.00         Median :113.6       Median :118.0
##  Mean   :6.119   Mean   :52.18         Mean   :114.0       Mean   :118.3
##  3rd Qu.:6.800   3rd Qu.:62.00         3rd Qu.:116.8       3rd Qu.:120.8
##  Max.   :9.500   Max.   :94.00         Max.   :124.7       Max.   :129.2
##  SLOWEST_CH_SPEED AVG_BALL_SPEED  FASTEST_BALL_SPEED SLOWEST_BALL_SPEED
##  Min.   :102.2    Min.   :157.2   Min.   :160.3      Min.   :146.5
##  1st Qu.:107.3    1st Qu.:165.7   1st Qu.:170.5      1st Qu.:159.4
##  Median :109.9    Median :169.1   Median :174.0      Median :163.2
##  Mean   :110.1    Mean   :169.6   Mean   :174.2      Mean   :163.6
##  3rd Qu.:112.6    3rd Qu.:173.4   3rd Qu.:178.0      3rd Qu.:167.5
##  Max.   :118.7    Max.   :182.2   Max.   :187.5      Max.   :177.3
##  AVG_SMASH_FACTOR   HIGHEST_SF      LOWEST_SF     AVG_LAUNCH_ANGLE
##  Min.   :1.423    Min.   :1.473   Min.   :1.337   Min.   : 7.14
##  1st Qu.:1.483    1st Qu.:1.514   1st Qu.:1.417   1st Qu.:10.20
##  Median :1.490    Median :1.517   Median :1.432   Median :11.11
##  Mean   :1.488    Mean   :1.516   Mean   :1.432   Mean   :11.05
##  3rd Qu.:1.498    3rd Qu.:1.519   3rd Qu.:1.450   3rd Qu.:12.09
##  Max.   :1.507    Max.   :1.539   Max.   :1.492   Max.   :14.71
##  LOWEST_LAUNCH_ANGLE STEEPEST_LAUNCH_ANGLE AVG_SPIN_RATE  HIGHEST_SPIN_RATE
```

```
##  Min.   : 1.250      Min.   :10.70         Min.   :2127   Min.   :2819
##  1st Qu.: 5.880      1st Qu.:13.72         1st Qu.:2508   1st Qu.:4122
##  Median : 6.955      Median :14.84         Median :2628   Median :4935
##  Mean   : 6.993      Mean   :14.88         Mean   :2634   Mean   :5087
##  3rd Qu.: 8.070      3rd Qu.:16.02         3rd Qu.:2758   3rd Qu.:5873
##  Max.   :11.320      Max.   :18.62         Max.   :3346   Max.   :9640
##  LOWEST_SPIN_RATE AVG_HANG_TIME   LONGEST_ACT.HANG_TIME SHORTEST_ACT.HANG_TIME
##  Min.   :1400     Min.   :5.500   Min.   :6.800         Min.   :0.500
##  1st Qu.:1541     1st Qu.:6.200   1st Qu.:7.500         1st Qu.:2.400
##  Median :1714     Median :6.350   Median :7.600         Median :3.400
##  Mean   :1737     Mean   :6.346   Mean   :7.638         Mean   :3.209
##  3rd Qu.:1900     3rd Qu.:6.500   3rd Qu.:7.800         3rd Qu.:4.100
##  Max.   :2314     Max.   :6.900   Max.   :8.700         Max.   :5.100
##  AVG_CARRY_DISTANCE LONGEST_CARRY_DISTANCE SHORTEST_CARRY_DISTANCE
##  Min.   :249.8      Min.   :271.6          Min.   :192.8
##  1st Qu.:270.9      1st Qu.:295.4          1st Qu.:237.0
##  Median :278.2      Median :304.4          Median :245.8
##  Mean   :278.1      Mean   :304.9          Mean   :244.8
##  3rd Qu.:283.7      3rd Qu.:313.6          3rd Qu.:253.2
##  Max.   :302.6      Max.   :337.7          Max.   :275.7
##    AVG_SCORE     TOTAL_STROKES   TOTAL_ROUNDS     MAKES_BOGEY.
##  Min.   :68.70   Min.   :3261   Min.   : 45.00   Min.   :12.20
##  1st Qu.:70.44   1st Qu.:4527   1st Qu.: 64.00   1st Qu.:15.23
##  Median :70.85   Median :5366   Median : 76.00   Median :16.29
##  Mean   :70.90   Mean   :5324   Mean   : 75.43   Mean   :16.49
##  3rd Qu.:71.32   3rd Qu.:6100   3rd Qu.: 86.75   3rd Qu.:17.42
##  Max.   :74.89   Max.   :7515   Max.   :107.00   Max.   :28.25
##   BOGEYS_MADE     HOLES_PLAYED       AGE             won
##  Min.   :123.0   Min.   : 810   Min.   :21.00   Min.   :0.0000
##  1st Qu.:193.2   1st Qu.:1152   1st Qu.:29.00   1st Qu.:0.0000
##  Median :223.0   Median :1368   Median :33.00   Median :0.0000
##  Mean   :222.1   Mean   :1358   Mean   :32.93   Mean   :0.1701
##  3rd Qu.:251.0   3rd Qu.:1562   3rd Qu.:36.00   3rd Qu.:0.0000
##  Max.   :330.0   Max.   :1926   Max.   :49.00   Max.   :1.0000
```

These two data sets have a variety of features, however both contain information on Strokes Gained. The player level data has some interesting metrics on how well a player got out of the sand, if they made risky moves and went for the green in two shots instead of three, and whether they made it into the hole from off the green in two strokes. The tournament level data may be more useful to look at what helps a player on a given day, whereas the aggregated player level data may better pick up on season-long trends.

```
tourney_clean |>
  group_by(player.id) |>
  mutate(avg_pos = mean(pos),
         avg_sg_putt = mean(sg_putt),
         avg_sg_t2g = mean(sg_t2g)) |>
  ggplot() +
  geom_point(aes(x = avg_sg_putt, y = avg_sg_t2g, color = avg_pos)) +
  labs(x = "Average Strokes Gained Putting",
       y = "Average Strokes Gained Tee To Green",
       title = "Average Tournament Placement based off of Putting and Tee To Green",
       color = "Average Position") +
  theme_classic()
```



It is without a doubt that better putting and golfing to the putting surface correlate with better scores, and therefore tournament placement.

Let's now look at what best indicates good tournament placement. I have elected to use a random forest, as I am curious about variable importance.

```
tourney_train <- tourney_clean |>
  filter(pos != 157) |>
  sample_n(1000)

rf1 <- train(pos ~ sg_putt + sg_t2g + sg_arg + sg_ott + sg_app,
             data = tourney_train,
             importance = TRUE,
             method = "rf")

rf1
```

```
## Random Forest
##
## 1000 samples
##    5 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 1000, 1000, 1000, 1000, 1000, 1000, ...
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared   MAE
##   2     10.41947  0.7682945  7.695275
```

```
##   3     10.35459  0.7694673  7.463238
##   5     10.60040  0.7590828  7.505311
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 3.
```

```
varImp(rf1)
```

```
## rf variable importance
##
##          Overall
## sg_putt 100.0000
## sg_t2g   66.0579
## sg_app    6.1104
## sg_arg    0.4368
## sg_ott    0.0000
```

As we can see from the model's results, we received a mean absolute error of around 10. In the context of a golf tournament, this is not so bad as one stroke can sometimes be the difference between 12th place and 40th place.

What is interesting, is that the most important variable by far is Strokes Gained Putting. This tells me crucial information about what is most important for me to practice! While it may not be as glorious as going to the range and smashing my driver as far as possible, it is proven effective. It's advice that everyone hears all the time, but no one likes to do it. Now, obviously, I am not a PGA Tour professional, but for them to place the best, they have to score the best, and to place the best requires excellent putting.

Now let's take a look at player level data, aggregate across the entire 2017 season. There are quite a few more metrics here which could prove important. The two models used predict if they have won or not and the number of top ten finishes. The reason for separating the models these ways is to look at what indicates someone winning a given weekend, versus what indicates a more consistent player who places in the top ten more often.

```
rf_has_won <- train(factor(won) ~ . -NUMBER_OF_TOP_Tens -NUMBER_OF_WINS -Player -POINTS -POINTS_BEHIND_
                    data = player,
                    importance = TRUE,
                    classification = TRUE,
                    method = "rf")

rf_toptens <- train(NUMBER_OF_TOP_Tens ~ . -NUMBER_OF_WINS -won -Player -POINTS -POINTS_BEHIND_LEAD -GI
                       data = player,
                       importance = TRUE,
                       method = "rf")


rf_has_won
```

```
## Random Forest
##
## 194 samples
##  69 predictor
##   2 classes: '0', '1'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 194, 194, 194, 194, 194, 194, ...
```

```
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.8555028  0.1521886
##   27    0.8397520  0.2325519
##   52    0.8228259  0.2103449
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2.
```

```
varImp(rf_has_won)
```

```
## rf variable importance
##
##   only 20 most important variables shown (out of 52)
##
##                            Importance
## SG.OTT                         100.00
## AVG_Driving_DISTANCE            82.47
## DRIVES_320..                    74.43
## AVG_CARRY_DISTANCE              72.40
## TOTAL_3_PUTTS                   62.58
## TOTAL_O.U_PAR                   61.83
## STEEPEST_LAUNCH_ANGLE           61.53
## SG.APR                          60.54
## POSSIBLE_FAIRWAYS               60.01
## TOTAL_DRIVES_FOR_320.           59.50
## GOING_FOR_GREEN_IN_2.           57.88
## NON.ATTEMPTS_GFG                56.59
## RTP.NOT_GOING_FOR_THE_GRN       56.44
## FASTEST_BALL_SPEED              56.24
## AVG_CLUB_HEAD_SPEED             54.30
## ROUGH_TENDNECY.                 51.28
## RTP.GOING_FOR_THE_GREEN         49.95
## SLOWEST_CH_SPEED                49.63
## AVG_SPIN_RATE                   49.42
## NUMBER_OF_BUNKERS               49.27
```

```
rf_toptens
```

```
## Random Forest
##
## 194 samples
##  69 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 194, 194, 194, 194, 194, 194, ...
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared   MAE
##    2    1.647663  0.4169363  1.283246
##   27    1.568132  0.4465329  1.211513
##   52    1.585291  0.4297929  1.224413
##
```

```
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 27.
```

```
varImp(rf_toptens)
```

```
## rf variable importance
##
##   only 20 most important variables shown (out of 52)
##
##                             Overall
## SG.APR                       100.00
## SG.OTT                        57.08
## NUMBER_OF_SAVES               55.58
## UP_AND_DOWN_.                 53.30
## SHORTEST_CARRY_DISTANCE       51.50
## POSSIBLE_FAIRWAYS             47.61
## TOTAL_DRIVES_FOR_320.         47.45
## NUMBER_OF_BUNKERS             41.35
## TOTAL_SG.PUTTING              38.42
## GOING_FOR_GREEN_IN_2.         35.67
## RTP.GOING_FOR_THE_GREEN       35.66
## DRIVES_320..                  35.14
## RTP.NOT_GOING_FOR_THE_GRN     35.13
## SG_PUTTING_PER_ROUND          33.75
## PAR_OR_BETTER                 30.02
## TOTAL_O.U_PAR                 29.07
## SG.ARG                        28.87
## AVG_Driving_DISTANCE          28.30
## STEEPEST_LAUNCH_ANGLE         26.81
## FASTEST_BALL_SPEED            26.28
```

I opted to remove many of the key features of the data set as I felt they were unfairly indicative of the results we see in tournaments.

The performance of the models is okay. The Kappa value of the classification of if a player has won or not indicates some slight agreement among the predictions. Alternatively, the Mean Absolute Rrror of the number of top tens a player had that season is rather good, with it being able to predict within about 1 or two

What these two models reveal is that on Tour, among other factors, your ability to get off the tee well (SG.OTT) and hit the ball far (AVG_Driving_DISTANCE) are crucial. These two metrics go more or less hand in hand, as SG.OTT is in part calculated by how far you hit the ball with your driver.

It's interesting to note the difference between these models. A more consistent golfer that places in the top ten often is a player whose approach game (SG.APR) shines. Approach shots are typically between 80 and 200 yards out, and are a defining factor on a hole's score. Additionally, we see importance among variables like how well they get themselves out of tricky situations (NUMBER_OF_SAVES) and whether they can chip the ball close when just off the green (UP_AND_DOWN_.).

On the other hand, a player that has won on tour can be categorized by a riskier player. The model favors players that focus on distance. Players that drive the ball far (DRIVES_320..) and go for the green in tough situations (RTP.GOING_FOR_THE_GREEN) are players that have won a tournament.

That said, I am not a professional golfer, and I know for a fact I cannot hit the golf ball with my driver 320 yards, nor hit every fairway. While these things are important to consider, it's

even more important to take into consideration that these are what help you win. On the PGA Tour. The most prestigious of golf leagues.

I am not quite trying to go out and win a PGA Tournament, I'm more interested in a few more pars and the occasional birdie, but this information is important to keep in mind regardless.

Let's split our data into the players that have won, and the players that have not won on tour to see if we can make a prediction based on their season metrics. To do this, we will use a SVM model (Support Vector Machine). Although this won't necessarily aid me in improving my golf game because we are unable to extract variable importance from an SVM, it could be useful if you were to be looking at this data mid-season who has not won, but likely will. According to this data, about $\frac{1}{3}$ of golfers on the PGA tour won once or more in the 2017 season.

Talk about SVM and how it works in context

A Support Vector Machine works by first, plotting each player on a high dimensional graph, where each different axis represents one of the features within the data set that we have selected to use for our model.

The algorithm then attempts to split the data points into two different groups with a hyper plane; think of this as a line in 2-Dimensional graphs, or a plane in a 3-Dimensional graph. It is essentially

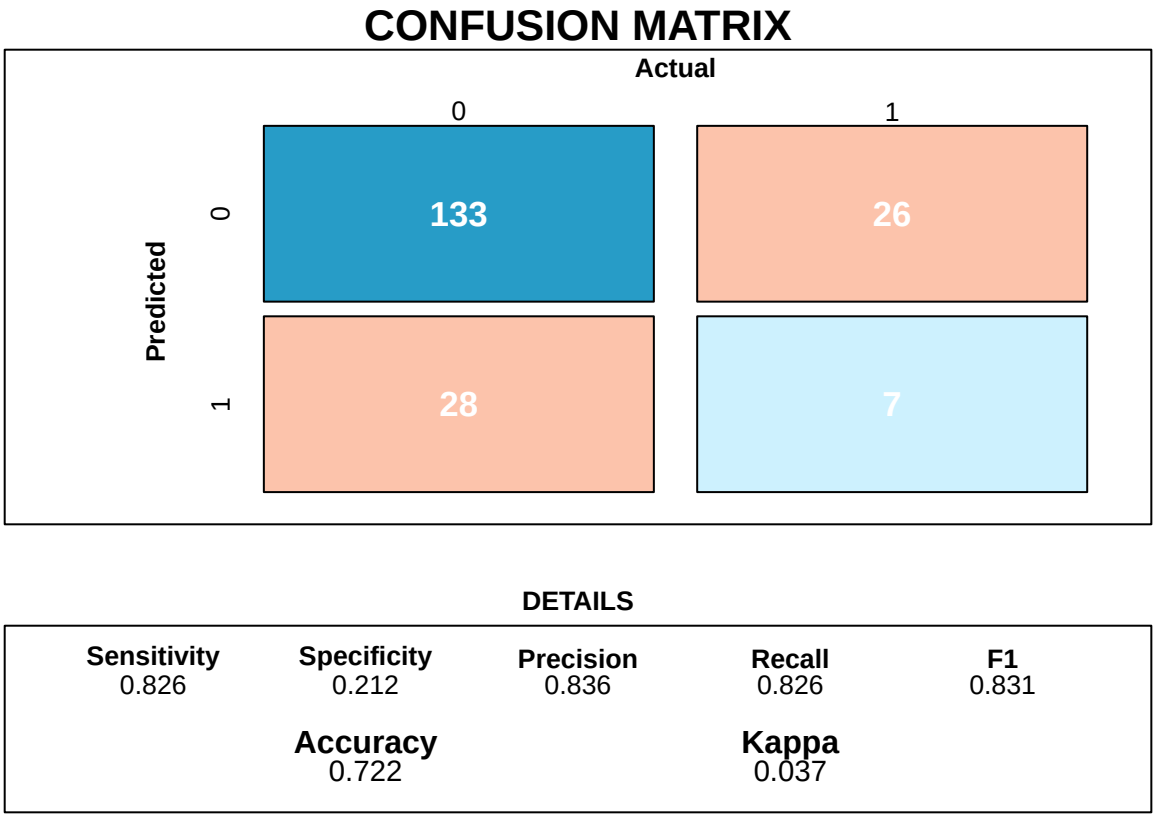
The golfers that are closest to this hyper plane according to their statistics are called the Support Vectors, and the algorithm aims to maximize the distance between these closest golfers plotted points and the hyper plane. This is achieved through mathematical processes.

```
train_control<- trainControl(method = "cv", number = 5, savePredictions = TRUE)

svm1 <- train(factor(won) ~ . -Player -NUMBER_OF_WINS -NUMBER_OF_TOP_Tens -POINTS -POINTS_BEHIND_LEAD -
              data = player_scale,
              trControl = train_control,
              method = "svmLinear")

preds <- svm1$pred
cm <- confusionMatrix(data = preds$pred, reference = preds$obs)
```

```
draw_confusion_matrix(cm)
```



Talk about results

The final confusion matrix of this model shows us similar results to the Kappa value of the random forest above. Although boasting a solid accuracy of around 0.75, the Kappa value tells us there is slight to little agreement among the prediction variables and the prediction.

It makes sense that we cannot make great predictions about whether a player has won a tournament or not. Two players could play almost identically up to the very final putt, and one will always come out on top. One stroke separates so much in the professional scene.

Although these results are rather lacking in their ability to predict which golfers have won tournaments, they do tell me a lot about what's important in golf. On the surface it may appear to be a numbers game, but in reality it's so much more. The fact that we cannot accurately separate the good from the great based on a multitude of numbers reveals to me so much about what is important for success. It's beyond the raw statistics.

A couple key parts of a player's game that cannot be expressed in these raw statistics are 1. their mentality, and 2. their course management. A player's course management does display itself within some Strokes Gained numbers, but that is only half of the result, as execution is another factor. The fact that we cannot accurately predict if someone has won emphasizes the importance of these two components of the game. This aligns with what makes the best golfers the best. Players that remain level headed, remain consistent and players that are able to reset after a bad shot typically score better.

Final findings

Taking into account all of these results (and their contexts) tells me that I need to focus on putting, accuracy off the tee, and possibly most importantly mentality.