Kelsey Schroeder and Roberto Argüello
June 3, 2016
MCS 100
Final Project Paper

Predicting the Winner of The Masters

Abstract

This paper presents a new way of predicting who will finish at the top of the leaderboard at the conclusion of The Masters. We used year-long statistics such as average driving distance, one putt percentage, par five scoring average, etc. from 2010 through 2014 to train our model – using year-long statistics from one year to predict the next year's Masters winner. We then used our trained model to predict the results of this year's 2016 Masters and compared the results which will be discussed in detail later in this paper. We also used our model to predict the top ten for The Masters in 2017, and we found some surprises in each of our results.

Background

The Masters is the Mecca of golf. Each year, thousands of golf fans travel from all over the world on their pilgrimage to golf's most prestigious golf course – Augusta National. Sure, St. Andrews was the first course and occasionally hosts The Open, but Augusta plays host to The Masters each year. The tradition of The Masters is unparalleled by any other golf tournament. The winner not only receives a handsome payout, but also the most prized piece of laundry in the sports world – the illustrious green jacket. The PGA Tour has several tournaments before The Masters, but for most golf fans, the Tour is not officially in season until the azaleas are in full bloom at Augusta. The Masters consistently receives higher ratings than the other three majors – making this a popular event in Vegas. Many people often bet on a single player to win the tournament. Another popular gambling game involving The Masters involves choosing one player from several sets of "groups." In each "group," there are fifteen to twenty golfers with similar odds. For example, there may be one group with all of the top favorites, then the next group might have the next wave of players with the best odds of winning and so on. After selecting players from each "group," the person who has picked his team can win the game if his team of players amasses the greatest sum of winnings from the tournament. Obviously, the person who picks the winner of The Masters in this game will have the greatest advantage since the winner is paid the most money – in excess of one million dollars. However, only players who make the cut get paid, so in this game there is an emphasis on having one's players make the cut in order for them to contribute to the winnings total. No matter which gambling game one is interested in, our model will prove useful in choosing which players to place your money on.

Related Work

Many predictions precede The Masters each year, though many of them rely on qualitative assessment. Sports writers, commentators, and golf fans around the world develop their own criteria for choosing the next winner. However, there are few approaches that derive from a statistical approach: a Google search for "best masters prediction analytics" yielded many results for Masters degree programs in predictive analytics, but next to none for predictions of the golf Masters. The most statistics-driven project we could find online involved minimal statistical analysis[1]. This is likely, in part, due to the inaccessibility of golf data to the public: the PGA, CBS, and other sites offer animations, live score tracking, and games that rely on comprehensive data, but the data is not available in an accessible format for analysis.

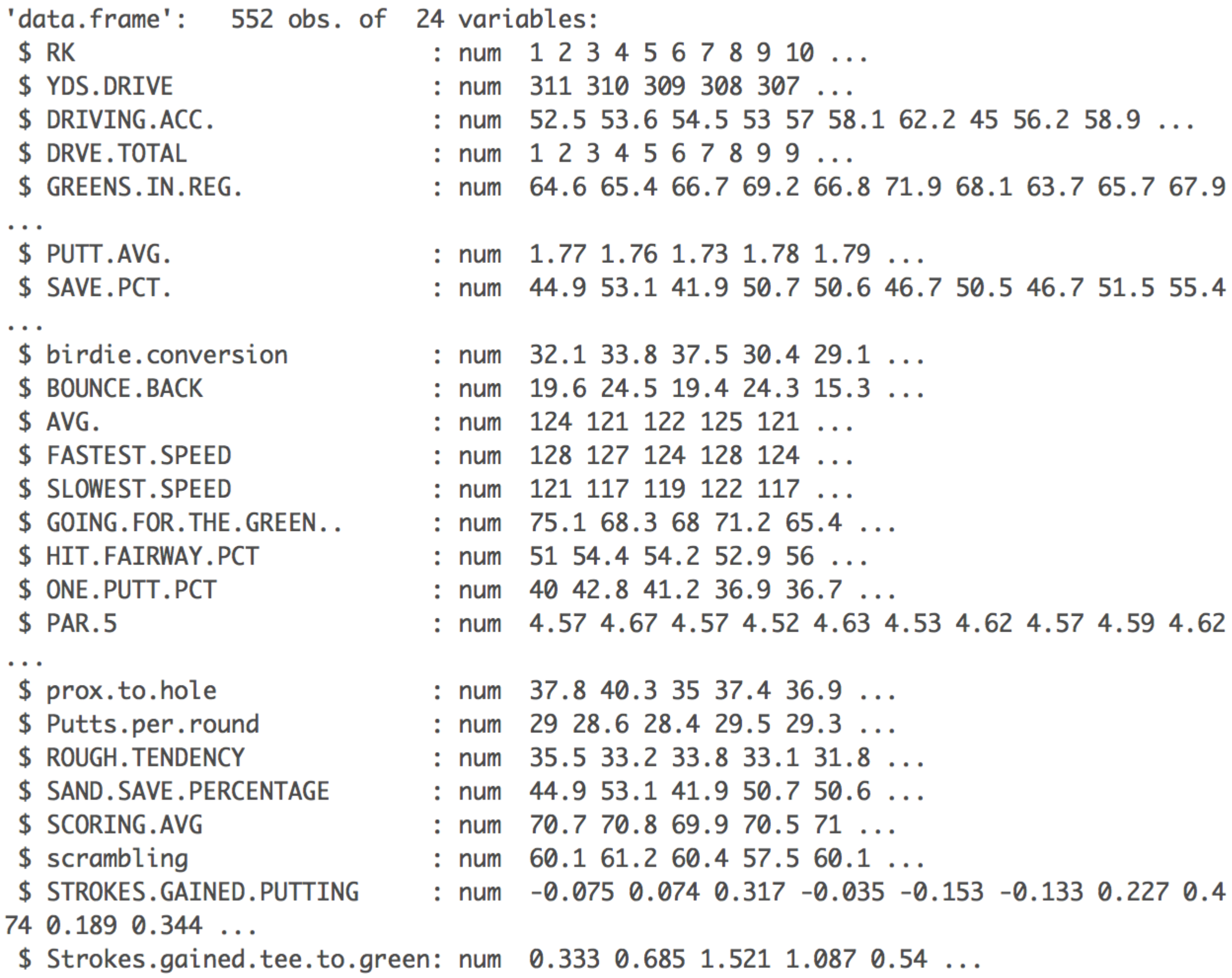
Methodology

Working with the glmnet package in R, we used a lasso regression to predict golfers' rankings in The Masters. We compiled statistics from the PGA tour and ESPN websites[2] to form a data set containing Masters' golfers and their statistics for 2010-2016. We used a given year's statistics to predict the Masters rankings of the following year (e.g. the 2010 tour data corresponds to the 2011 Masters results). Golfers who did not have adequate tour data from any given year were omitted from our analysis. We trained our model on 2010-2014 and used the 2015 data to predict the 2016 Masters as our test.

Data Set

We view our data set as our largest contribution to sports statistics from this project. We scraped data from dozens of web pages to generate a comprehensive set of tour data. Our data set contains the variables shown below (Figure 1).

Our data set is currently available on gitHub (/kelsey17), and we are looking into creating an R package that will allow others to load the data directly in R. We hope that more people will gain interest in golf statistics now that the data is publicly available. We look forward to seeing how others can build off our analysis and use statistics to analyze different aspects of the golf tour and game in general.



```
'data.frame':   552 obs. of  24 variables:
 $ RK                       : num  1 2 3 4 5 6 7 8 9 10 ...
 $ YDS.DRIVE                : num  311 310 309 308 307 ...
 $ DRIVING.ACC.             : num  52.5 53.6 54.5 53 57 58.1 62.2 45 56.2 58.9 ...
 $ DRVE.TOTAL               : num  1 2 3 4 5 6 7 8 9 9 ...
 $ GREENS.IN.REG.           : num  64.6 65.4 66.7 69.2 66.8 71.9 68.1 63.7 65.7 67.9
...
 $ PUTT.AVG.                : num  1.77 1.76 1.73 1.78 1.79 ...
 $ SAVE.PCT.                : num  44.9 53.1 41.9 50.7 50.6 46.7 50.5 46.7 51.5 55.4
...
 $ birdie.conversion        : num  32.1 33.8 37.5 30.4 29.1 ...
 $ BOUNCE.BACK              : num  19.6 24.5 19.4 24.3 15.3 ...
 $ AVG.                     : num  124 121 122 125 121 ...
 $ FASTEST.SPEED            : num  128 127 124 128 124 ...
 $ SLOWEST.SPEED            : num  121 117 119 122 117 ...
 $ GOING.FOR.THE.GREEN..    : num  75.1 68.3 68 71.2 65.4 ...
 $ HIT.FAIRWAY.PCT          : num  51 54.4 54.2 52.9 56 ...
 $ ONE.PUTT.PCT             : num  40 42.8 41.2 36.9 36.7 ...
 $ PAR.5                    : num  4.57 4.67 4.57 4.52 4.63 4.53 4.62 4.57 4.59 4.62
...
 $ prox.to.hole             : num  37.8 40.3 35 37.4 36.9 ...
 $ Putts.per.round          : num  29 28.6 28.4 29.5 29.3 ...
 $ ROUGH.TENDENCY           : num  35.5 33.2 33.8 33.1 31.8 ...
 $ SAND.SAVE.PERCENTAGE     : num  44.9 53.1 41.9 50.7 50.6 ...
 $ SCORING.AVG              : num  70.7 70.8 69.9 70.5 71 ...
 $ scrambling               : num  60.1 61.2 60.4 57.5 60.1 ...
 $ STROKES.GAINED.PUTTING   : num  -0.075 0.074 0.317 -0.035 -0.153 -0.133 0.227 0.4
74 0.189 0.344 ...
 $ Strokes.gained.tee.to.green: num  0.333 0.685 1.521 1.087 0.54 ...
```

Figure 1: Data Set Summary

Omitted Variables

We ran a preliminary regression with a subset of our data set and found that some statistics hindered our model more than they helped it or had near-zero coefficients, indicating they did not play a meaningful role in determining the output of our model. These are some of the year long statistics we investigated but did not use in our final model.

1. Rank
2. Age
3. Putting Average
4. Total Putting
5. Scoring Average on Par 3s and 4s

6. GIR Rank
7. Total Putts Gained
8. Greens in Regulation Percentage

Variables (Relevant Statistics for Regression)

The following statistics were used to build our final model:

1. **Scoring Average:** A player's average score per round of 18 holes.

2. **Proximity to the Hole:** The average distance from a player's ball to the hole after his approach shot

3. **Strokes Gained: Putting:** Strokes gained on the green per hole (average)

4. **Strokes Gained: Tee to Green:** Average strokes gained per hole from the tee through the green

5. **Sand Save Percentage:** Percentage of up and downs from bunkers (both fairway and greenside bunkers are included in this statistic

6. **One Putt Percentage:** Percentage of holes where a player takes one putt on the green

7. **Going for the Green:** Percentage of green reached under regulation when a play goes for the green

8. **Rough Tendency:** the percentage of a player's shots that miss the fairway and land in the rough

9. **Hit Fairway Percentage:** Percentage of fairways hit off the tee

10. **Putts Per Round:** number of putts per round

11. **Scrambling:** percentage of up and downs for a player when he misses a green in regulation

12. **Birdie Conversions:** percentage of holes where a player makes birdie or better when on the green in regulation

13. **Club Head Speed:** a player's average club head speed off the tee

14. **Bounce Back Percentage:** percentage of times a player makes birdie or better on a hole after making a bogey or worse on the previous hole

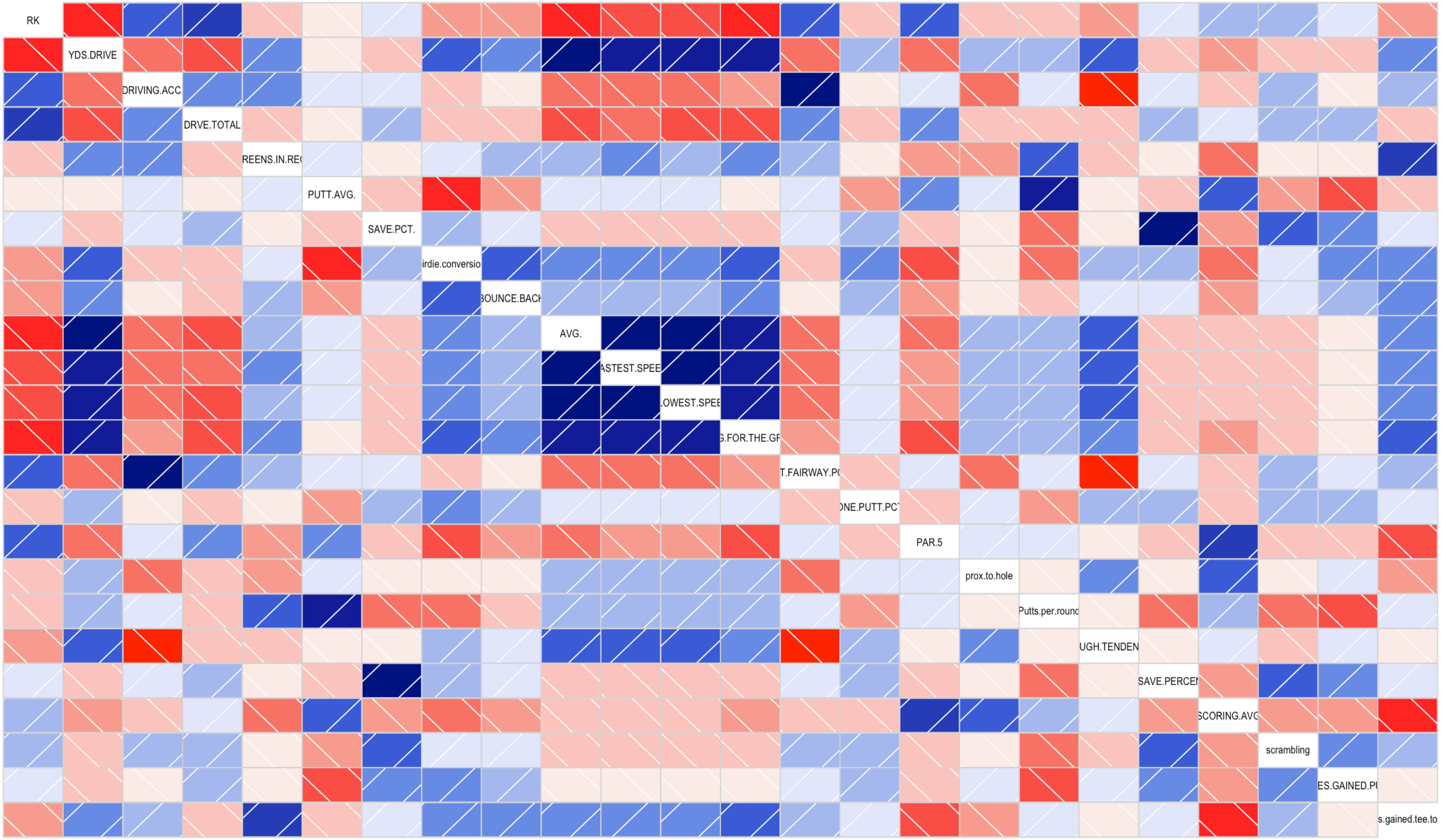
15. **Par 5 Scoring Average:** average number of strokes per par 5

How Did Our Variables Relate to Each Other?

We performed nonlinear matrix factorization and examined correlation values to explore the relationship between our variables. On the following page is a correlogram of our variables: blue indicates positive correlation and red indicates negative correlation. The vibrancy of the color corresponds to the strength of the correlation. Many of the results of this analysis are not surprising; for instance, driving distance is strongly correlated with club head speed, and hit fairway percentage is negatively correlated with rough tendency.

However, some results are intriguing. We can see that higher club head speed and greater driving distance is correlated with a higher rough tendency (and lower hit fairway percentage), so extra distance comes at the cost of accuracy. Curiously, rank is most dependent on driving distance rather than accuracy: higher driving accuracy corresponds to a higher (worse) rank, while players with high club head speed and large driving distances have more favorable ranks. Putting statistics contribute little to a player's rank. Putting and driving statistics in general seem largely unrelated, suggesting that top players must have a mastery of both skills individually to perform well.

Figure 2: Correlogram



Results

These were the players predicted by our model to finish in the top ten at The Masters in 2016 based off their statistics from their performances on the PGA Tour in 2015. We have listed each player's actual finish in parentheses.

1.	Bubba Watson	(T37)
2.	Justin Rose	(T10)
3.	Jason Day	(T10)
4.	Henrik Stenson	(T24)
5.	Brooks Koepka	(T21)
6.	Adam Scott	(T21)
7.	Dustin Johnson	(T42)
8.	Sergio Garcia	(T34)
9.	Hideki Matsuyama	(T7)
10.	Justin Thomas	(T39)

While our model failed to predict the winner of The Masters, it was still surprisingly accurate. All ten of the players whom we predicted would finish in the top ten made the cut. One reason why our model failed to predict the winner of the 2016 Masters was because the winner (Danny Willett) was not included in our data set because he did not become a full member of the PGA Tour until after he won at Augusta this April. We were very surprised that Jordan Spieth was not included in our top ten projections – especially because he won the previous year (which our model was trained on) and ranks so high in statistics such as putts per round and scoring percentage.

Predictions for 2017

1. Rory McIlroy
2. Adam Scott
3. Justin Rose
4. Jason Day
5. Dustin Johnson
6. Brenden Grace
7. Jordan Spieth
8. Brooks Koepka
9. Phil Mickelson
10. Matt Kuchar

We were surprised by some of the predictions that our model yielded for 2017's Masters. Specifically, we thought that Jason Day would have been predicted to finish higher than just fourth because he already has four wins on this year's tour (no one else has more than two). Not only was Day projected to finish fourth, but he was predicted to finish one spot worse than he was predicted to finish in our 2016 projections (third). Even though he had a strong year in 2015, it was nothing near the dominance he has displayed thus far in 2016. Additionally, our model for 2017 does include someone whom we expected to see in our 2016 projections but did not – Jordan Spieth. This was somewhat surprising because Spieth has not played nearly as well this year on tour as he did last year in 2015, but now he has jumped into our projected top ten for 2017.

Improvements

We would love to expand our analysis to include more detailed data. This could include hole-by-hole and stroke-by-stroke data, as well as data on weather, course conditions, etc. We could eliminate variables that are closely related, or add dimensions to variables we already have (e.g. distinguish between left rough percentage and right rough percentage instead of using both together). Instead of using data from a given calendar year, we'd like to use data from the previous year's Masters to the current year's Masters for a more appropriate range of data (this is not currently possible, as only summary statistics are available). Additionally, it would be interesting to include other tournament results in our data set, as golfers who have already won tournaments in a given year may be more likely to win The Masters.

With more detailed data we could perform more narrowed analyses. For instance, with stroke-by-stroke data, we could break down certain holes at Augusta and determine which strategies are optimal (e.g. going for the green in two on holes eight and thirteen). This type of analysis could benefit not only gamblers and golf fans, but also the golfers and caddies themselves.

Limitations of Golf Statistics

Over the course of the project, we discovered several limitations in golf statistics as a whole. First is the lack of accessible data: we would have liked to include hole-by-hole or even stroke-by-stroke data in our analysis, but this data is not available in any form to the public. Furthermore, we found minute differences between players in many of our variables, especially percentage: Aaron Baddeley's 45.76 putting percentage is hardly different from Steve Stricker's 45.40%. Player performance can also vary year to year, as a result of injury, new swing, etc. Additionally, tournaments like The Masters are often decided by one stroke, so the difference in performance is slight between top players, making it difficult to predict a winner. The course also changes slightly each year, which can add error to predictions. Finally, tournaments are inherently unpredictable: for example, in the 2013 Masters, Tiger Woods hit the pin, sending his very accurate shot directly into the water. Anomalies like these are beyond the scope of golf predictive statistics.

[1] http://espn.go.com/golf/masters16/story/_/id/15131730/eliminator-uses-statistics-predict-masters-winner

[2] http://www.pgatour.com/stats.html, http://espn.go.com/golf/statistics/_/type/expanded