Drive for Show, Putt for Dough

BRAD COOLEY

Trying to determine what makes someone a better golfer is a reasonably simple task. Hitting more fairways, consistently making par or under par, putting in two strokes or less are all parts of one's golf game that makes them an overall better golfer. However, trying to predict which of these factors will help them win more becomes increasingly more difficult. Taking raw PGA Tour data from 2010 through 2018 helps give clues as to how to solve this question but does not answer it definitively. More data and analysis beyond the scope of this course would be needed to give definitive answers.

ACM Reference Format:
Brad Cooley. 2019. Drive for Show, Putt for Dough. 1, 1 (December 2019), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

There's a famous saying in the golf world, "drive for show, putt for dough." This means that one who can drive the ball far will get a lot of looks because people somehow correlate how far a ball is driven to the golfer's strength. While that is neither here nor there, the latter part of the saying states that the real money maker is a golfer's putting game and how well they can putt. As an avid golfer, I wanted to take a data-focused approach to the game and see what I could do to improve my game. To do that, I decided to analyze the pros and get a general idea of what factors affect wins.

The Professional Golf Association (PGA) has long collected different data and metrics on their players. However, only in the past decade has it become as in-depth as the average driving distance on a course that was primarily Kentucky Blue Grass versus Poa Annua. As the past decade has progressed, those stats and metrics have become even more and more refined as players start realizing the true value of this data.

The data set analyzed for this project is from Kaggle and originally had 2312 observations, each with 18 variables. The data ranges from years 2010 to 2018 and includes simple metrics like Average Drive Distance and Wins to very advanced metrics such as Average Strokes Gained From Putting.

The goal for this data set and analysis was to determine what factors in a golfer's game make the most significant impact on their wins.

2 LITERARY REVIEW

The goal of this analysis was to find out what aspects of golf can influence whether a player wins or not. Believe it or not, this question has been asked since the beginning of the century. Back in 2001, Thomas Dorsel and Rob Rotunda asked how could they use scores, top 10 finishes, and the amount of money won to predict who would win a particular tournament on the PGA Tour [4]. While this question is fundamentally similar to mine, the data that they had available to them was far more limited. They tried using only three predictors, and one of them is known, in golf, to be a flawed predictor (amount of money won). This is because each event has a different "purse" (total amount of money available to win), and one event in particular (The Masters) is much larger than the rest. So, a golfer could win The Masters, come in last in every other tournament, and still

Author's address: Brad Cooley, brad.cooley@marquette.edu.

XXXX-XXXX/2019/12-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

win more money than a golfer who won three smaller tour events. Dorsel and Rotunda raise this concern in their work, but having such limited data, this was the best he could do with his analysis.

Dorsel and Rotunda, while being one of the first to publish a scholarly article about using statistics to predict a winner, were not the first to look into how one could use stats to predict various outcomes. In 1994, David Belkin, Bruce Gansneder, Morris Pickens, Robert Rotella, and David Striegel published their work titled "Predictability and Stability of Professional Golf Association Tour Statistics," which looked at the validity of individual statistics the PGA gathered [3]. The goal of this team's research was to show that certain statistics collected by the PGA were valid predictors in the successfulness of a golfer. The article uses data collected by the PGA about golfers who competed on the PGA Tour from 1986-1988. Being even earlier than Dorsel and Rotunda, this data was elementary and dealt with information like "Average Drive Distance" and "Average Score." More advanced metrics created in the past decade do not show up in their findings. However, they do establish, through a stepwise multiple regression, that "hitting greens in regulation is the most important factor in maintaining a low scoring average"[3]. With my data set, I chose to analyze that aspect as well to see how the game has changed in the past 25 years.

One thing that has stayed the same since the article by Belkin and his team was written in 1994 is statisticians' curiosity with the correlation between different statistics. In 2006, Robert Quinn published an article that looked at how different stats, such as a golfer's driving distance and how many strokes they take to putt, relate to each other and if one seems to affect the other [8]. While this approach is different from the last two articles, it still gives insight into what makes a better golfer overall. His findings were pretty consistent with his hypothesis that driving a ball further doesn't necessarily make you a better putter. However, if a golfer tended to hit the fairway more often, they had a high probability of making a putt in one stroke versus two or more strokes [8]. This insight, while still pretty simple, does allow a relationship to be established between two different aspects of the golf game, therefore, giving insight into what makes a better overall golfer.

Moving into recent years, technology has advanced so much that the game of golf can be predicted down to the shot level. What does this mean? It means that each shot of a round can be predicted (type of club used, yardage, chance that the golfer will end up with a good lie, etc.). Christian Drappi and Lance Co Ting Keh's article "Predicting golf scores at the shot level" give insight to this wave of metrics in the golf community[5]. They explain that the purpose of predicting at the shot level is because it provides more granularity to their metrics and helps with more precise and accurate models. Drappi and Co Ting Keh also present their model in the paper, which is a neural network that takes in predictors similar to the ones found in my data set. It just goes to show that the analysis I want to do is validated. Many other statisticians and data scientists are showing interest in the same things in this more modern era.

While this is not specific to golf, Rory Bunker and Fadi Thabtah published a paper discussing how they used machine learning as a framework for sports results prediction[1]. They state that machine learning is one of the best and most accurate methodologies for predicting outcomes in sports. Again, while this is not specifically about golf, they mention applications for their model in the game of golf. They believe it could help golfer's lower their average stroke by half of a stroke, which is a big deal in the PGA where a single stroke usually separates the top 10 golfers.

Even outside of academia, plenty of people are rushing to find how they can predict sports outcomes so they can have better odds betting. Specifically to golf, Stephen Hennessey of GolfDigest.com writes that experts at Golf Digest have been able to predict almost every winner of the 2019 PGA Tour season [6]. They claim their model is one of the best-trained models out there because of the amount of data they collect and feed it. Kyle Porter of CBSSports.com says that predicting golf is one of the hardest things in all of sports because of the amount of variance that can happen. However, he claims that betting wizard Mike McClure of Sportsline has the best model out there and has, also, predicted every winner in the 2019 PGA Tour season[7]. While neither of these authors shared much about how the models were trained or created, the Courchene brothers from Pinnacle.com did, and they revealed the top five factors in a such a complicated model. They even said that their model was elementary by comparison. However, it doesn't take long to realize that their model is incredibly complex and involves lots of predictors and other various factors[2].

3 DATA ORGANIZATION AND CLEANING

The data set is an uncleaned, raw-formatted CSV file that I downloaded from Kaggle.com. It contains various golfing statistics collected about golfers on the PGA tour from 2010-2018. The original data set was a little over 2700 rows by 18 columns. A lot of those rows were full of empty values, and because filling those empty values with 0's would lead to very skewed data, the decision was made to remove them. That was the first step in the cleaning process; any rows with more than three NA's or empty values were removed from the data set, and the resultant data set was given a new name for future cleaning. From there, it was essential to decide what data was going to be useful for the question at hand: what makes a better golfer (i.e., a lower score)?

The original data set, as mentioned, had 18 columns. After removing advanced metrics that were not essential to analyzing the core concepts of the game of golf, the data set shrunk down to 12 columns. Another motivation behind removing those six columns was because they were advanced metrics that came about as a result of some calculation or other parameters being combined into it. It left a lot of uncertainty when analyzing the data set; therefore, they were left out. Next on the list was reorganizing the columns into a more logical format. It also made for a better representation of the data, with metrics that matter the most showing up ahead of less valuable metrics. In the process of reordering the columns, they were also renamed to have more cohesive and clear titles.

One of the upsides of this data set was the way the data was entered left little to be done in the form of combining data and making the data set tall instead of long. The next item on the list was to deal with missing or empty values in the `Wins` and `Top.10` columns. They were replaced with 0's as we can assume that if they had won an event or placed in the top 10, that stat would have been recorded. This left only four rows with some empty or missing value.

Those final four rows were missing a value for the amount of money the player had won. Because of the nature of golf and how money won doesn't necessarily correlate with the number of wins a player has, the choice was made to remove these four rows. After that, the data set was ready to go for some analysis.

4 VISUALIZATIONS RESULTS

4.1 First Visualizations Findings

As stated before, the main question I had when I decided to look at my data set was, "what makes a better golfer?" I wanted to know how I could improve my game based on what pros were doing. So, based on this straightforward question, a graph was produced looking at how the first shot of a hole (the drive) affects the number of wins a golfer has. Figure 1 is the result of that curiosity.

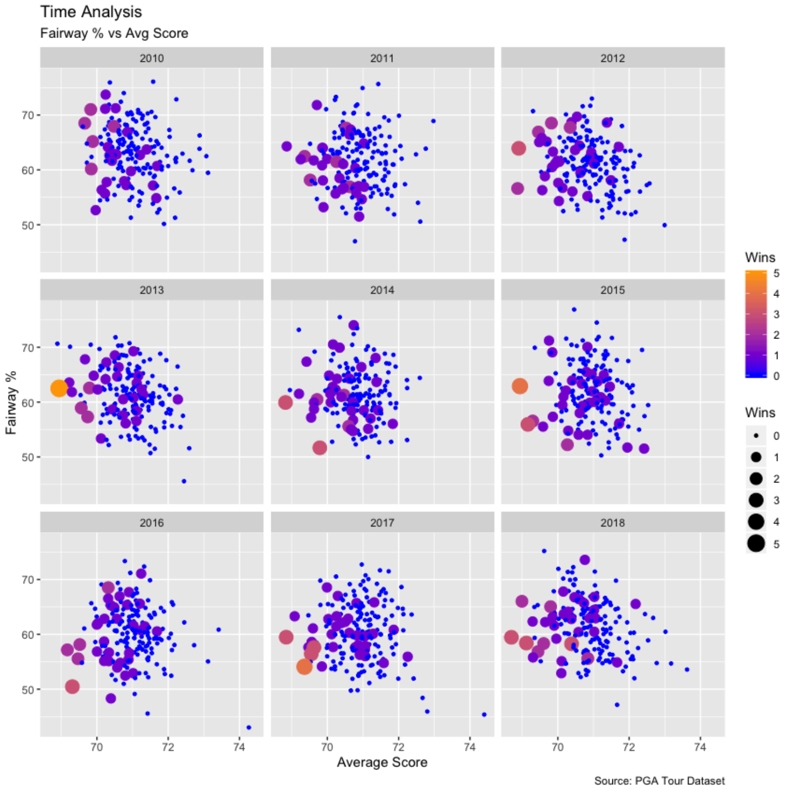
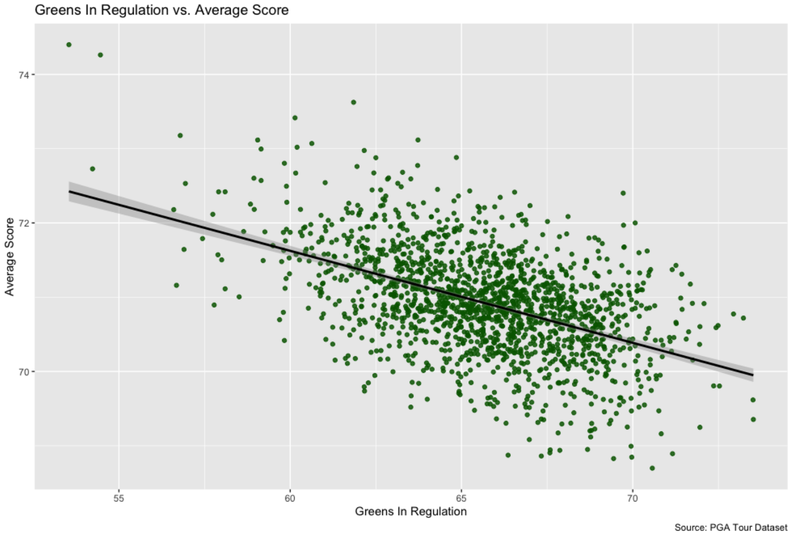


Fig. 1. Small multiples of Fairway Percentage versus Average Score over the 9 years in the data set.

Figure 1 was an early graph, and purely there to show the relationship between the two variables. I chose to graph it in a small multiple format because I wanted to see if there was any correlation to time. If there was, this might suggest technology or a realization within the golfing community happened and caused a change. However, it does not seem that there was any change; in general, the higher the fairway percentage, the lower the score.

This visualization guided my next question; what factors are the most significant predictors in a golfer's success? In my case, I determined success by the lowest average score.



Source: PGA Tour Dataset

Fig. 2. Scatter plot that displays the correlation between Greens In Regulation and Average Score. It contains a line of best fit (linear regression).

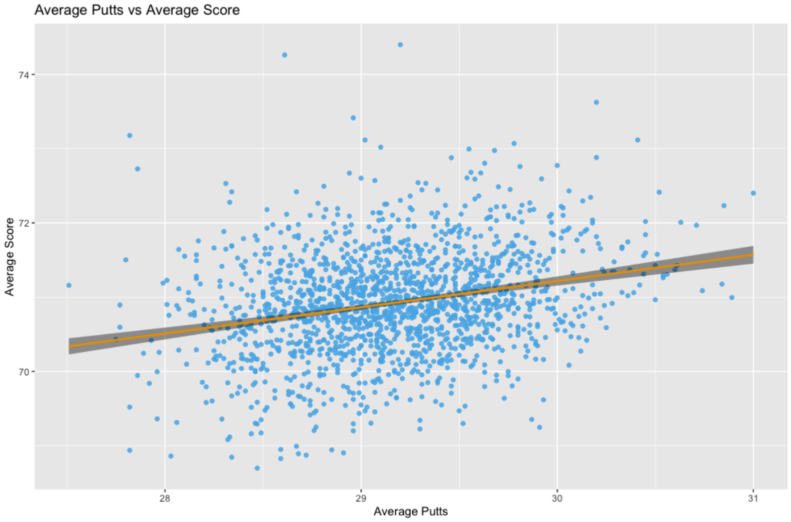
Figure 2 looks at how correlated hitting greens in regulation is to predicting one's score. The reasoning behind a scatter plot here was to see the distribution of data as well as the linear regression.

Predictor	Std. Error	T Value	Pr(>|t|)
Intercept	0.358290	220.65	$< 2 \cdot 10^{-16}$ * * *
Greens In Reg.	0.005451	-22.72	$< 2 \cdot 10^{-16}$ * * *
Signif. Codes:	0 ’***’	0.001 ’**’	0.05 ’*’

Residual Standard Error:	0.611 on 1668 DF
Multiple R^2 :	0.2364
Adjusted R^2 :	0.2359
F-statistic:	516.3 on 1 and 1668 DF
P-value:	$< 2.2 \cdot 10^{-16}$

Fig. 3. Linear regression results from the analysis of predicting the Average Score based on Greens In Regulation.

The R^2 value here is 0.2364, which shows that there is a decent amount of correlation between Greens In Regulation and Average Score. My next question based on this was if this is the best predictor, what is the next greatest? Could I use that in a multiple regression to see what the main factors in having a better golf game are?



Source: PGA Tour Dataset

Fig. 4. Scatter plot that displays the correlation between Average Putts and Average Score. It contains a line of best fit (linear regression).

Figure 4 tells a lot. While it is still only a simple scatter plot, the confidence interval in incredibly small, which means this might be a pretty good predictor in determining the average score. A scatter plot was chosen again as I wanted to see the variance in the data and see what the clustering looked like. Indeed, here, the visualization shows that the lower the number of putts a golfer has, the lower their score will be. This makes sense because if a golfer putts less, that means fewer strokes, which means a lower score.

Predictor	Std. Error	T Value	Pr(>|t|)
Intercept	0.92778	65.34	$< 2 \cdot 10^{-16} ***$
Avg. Putts	0.35326	11.11	$< 2 \cdot 10^{-16} ***$
Signif. Codes:	0 ’***’	0.001 ’**’	0.05 ’*’

Residual Standard Error:	0.6747 on 1668 DF
Multiple R^2 :	0.06885
Adjusted R^2 :	0.06829
F-statistic:	123.3 on 1 and 1668 DF
P-value:	$< 2.2 \cdot 10^{-16}$

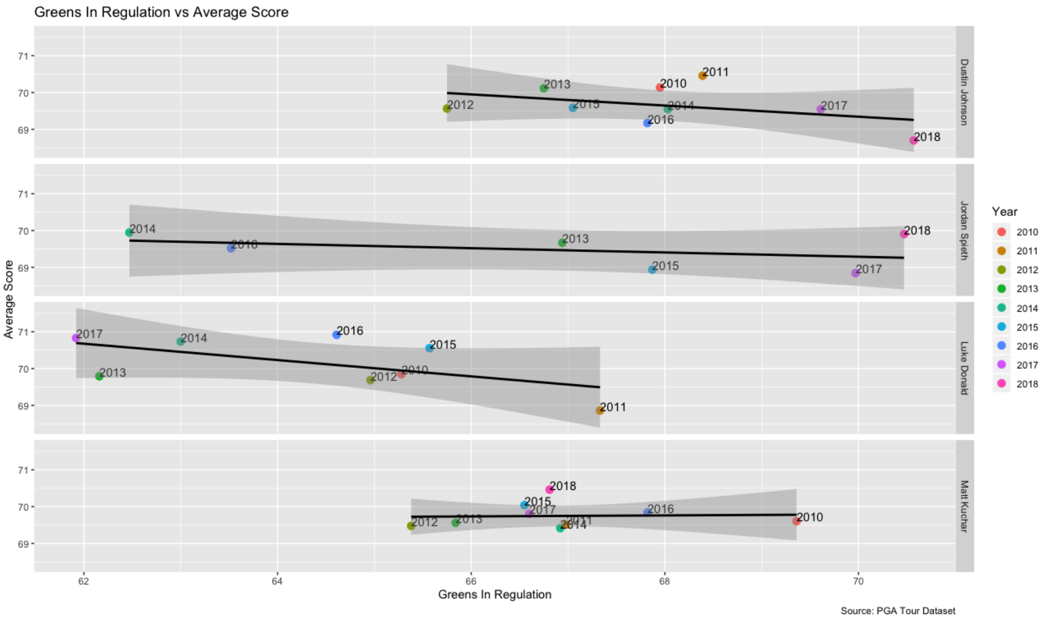
Fig. 5. Linear regression results from the analysis of predicting the Average Score based on Average Number of Putts.

The R^2 value here is 0.06885 which is much lower than the last. However, it is still the next biggest value amongst all predictors in the data set. This raises the question of, "Is there only one thing that affects how well a golfer performs, or are there multiple?" I would say, based on the low R^2 values found in the last two linear regressions, that no one factor has such a big effect on the score.

4.2 Elite vs. the Rest

Based on the findings already, it makes some sense that a multiple regression should be performed looking at various factors in achieving a low average score. One curiosity that I had when performing these tests was how do the elite golfers in the data set stacked up against the rest of the bunch? To find this out, some data sorting had to take place.

The data set was first ordered by top 10 finishes because wins aren't the best indicator of a good golfer. Sure, a player can't average a 78 and expect to win, but that doesn't mean that a player doesn't have one good tournament and then sucks. It also doesn't mean that a player wasn't elite and got hurt; take Tiger Woods, for example. He is one of the greatest golfers in the history of the sport, yet this data set only has two years of his data as he was hurt for the majority of the other time. After the data set was sorted, the top 5 data points based on top 10 finishes were grabbed and put into a new data frame where some unique comparative analysis could be done.



Source: PGA Tour Dataset

Fig. 6. Small Multiples scatter plots that display the correlation between Greens in Regulation and Average Score for the top four golfers in the data set. It contains a line of best fit (linear regression).

Figure 6 is another small multiples visualization. The goal of this one was to show how good of a predictor Greens In Regulation is for the most elite golfers and compare them against each other. In this case, it's a phenomenal predictor for Matt Kuchar as all of his data points except one are inside the confidence interval. The other key here is that his confidence interval is relatively small as compared to the other three golfers chosen.

However, don't trust everything. To prove that this is the case, we can do a linear regression on just these four players.

Predictor	Std. Error	T Value	Pr(>|t|)
Intercept	5.32596	14.770	0.000673 ***
Greens In Reg.	0.07819	-1.811	0.167765
Signif. Codes:	0 ’***’	0.001 ’**’	0.05 ’*’

Residual Standard Error:	0.2241 on 3 DF
Multiple R^2 :	0.5224
Adjusted R^2 :	0.3632
F-statistic:	3.281 on 1 and 3 DF
P-value:	0.1678

Fig. 7. Linear regression results from the analysis of predicting the Average Score based on Greens In Regulation of the top golfers.

Figure 7 turned up a lot of surprising results. Indeed, the R^2 value was lower, which means that there is less correlation for the most elite golfers in the group. This must mean that there are other aspects of their game that matter and make them the best of the best.

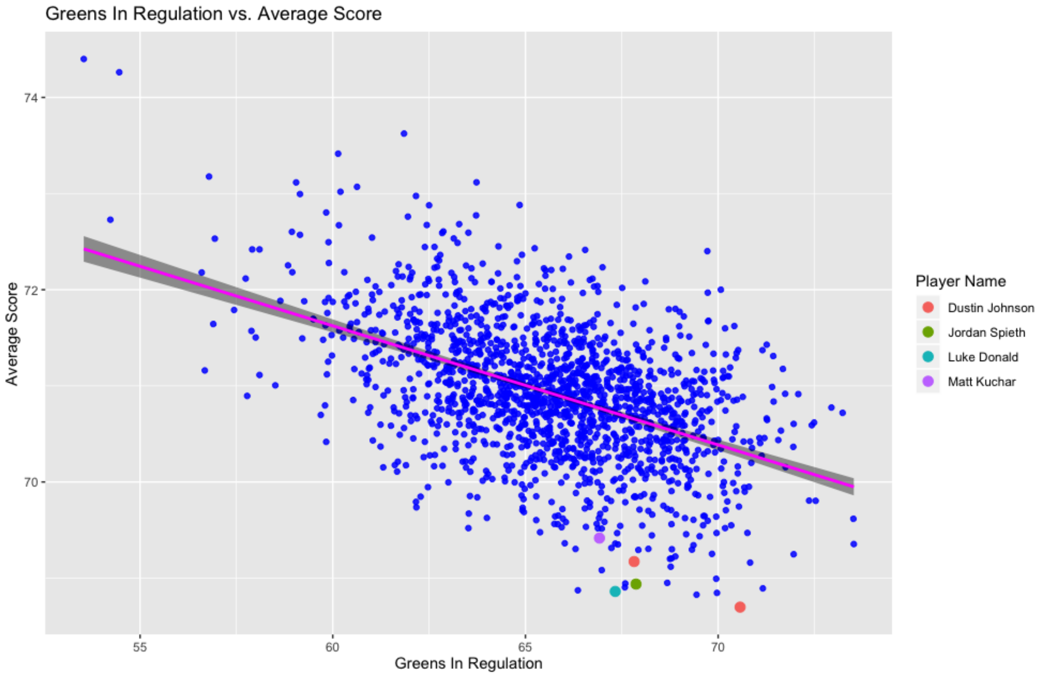
Running a simple multiple regression trying to predict the average score of the elite based on Greens In Regulation and the other top predictor, Average Putts, yielded a lot of exciting results.

Predictor	Std. Error	T Value	Pr(>|t|)
Intercept	5.70062	11.646	0.00729 ***
Greens In Reg.	0.04619	-3.265	0.08237 *
Avg Putts	0.17671	2.579	0.12321
Signif. Codes:	0.001 ’***’	0.01 ’**’	0.05 ’*’

Residual Standard Error:	0.132 on 2 DF
Multiple R^2 :	0.8896
Adjusted R^2 :	0.7791
F-statistic:	8.054 on 2 and 2 DF
P-value:	0.1104

Fig. 8. Multiple regression results from the analysis of predicting the Average Score based on Greens In Regulation and Average Putts of the top golfers.

Figure 8 proved that there is a lot of correlation between Greens In Regulation and Average Putts, when it comes to predicting the Average Score of the elite golfers. While correlation does not equal causation, these aspects of the golf game do have an effect on the average score. But just how good are the elite?



**Dustin Johnson has two data points as he had 2 years in the top 5 of the dataset.
Source: PGA Tour Dataset

Fig. 9. Scatter plots that display the correlation between Greens in Regulation and Average Score for the data set. The top five data points based on top 10 finishes are highlighted. It contains a line of best fit (linear regression).

Looking at Figure 9 is the same visualization from Figure 1, just with the elite players outlined. It indeed shows how big of a factor Greens In Regulation is. We can compare the multiple regression results from the elite to the entire data set to confirm this.

Predictor	Std. Error	T Value	Pr(>|t|)
Intercept	0.58229	96.40	$< 2\cdot10^{-16} * * *$
Greens In Reg.	0.04619	-3.265	$< 2\cdot10^{-16} * * *$
Avg Putts	0.17671	2.579	$< 2\cdot10^{-16} * * *$
Signif. Codes:	0 ’***’	0.001 ’**’	0.05 ’*’

Residual Standard Error:	0.4187 on 1667 DF
Multiple R^2 :	0.6416
Adjusted R^2 :	0.6412
F-statistic:	1492 on 2 and 1667 DF
P-value:	$< 2.2\cdot10^{-16}$

Fig. 10. Multiple regression results from the analysis of predicting the Average Score based on Greens In Regulation and Average Putts of the top golfers.

Figure 10 comes in with a smaller R^2 value than when we just analyzed it for the elite golfers. One explanation for this would be because Average Putts is not as correlated to Average Score for less skilled golfers as it is for the elite.

5 DISCUSSION

5.1 Information Overload

With over 1,000 data points, this data set can be a little bit of information overload. There aren't a lot of columns, but there still is a lot of data to analyze and visualize.

5.2 Data Tells a Story

Data always tells a story. In this case, there is a lot to learn from the data set. The main takeaway is that there isn't one answer to "how to become better at golf." It depends on the type of player one is. For someone like Jordan Spieth or Dustin Johnson, hitting more greens in regulation is not the biggest factor in their game. Matter of fact, there isn't even just one factor for them; it's multiple.

5.3 What's Next

This data set answered a lot of questions, and the clearest next step is for the golfers at the bottom of the top 10 ordering; hit more greens in regulation. While it was determined that there isn't one specific factor that affects a golfer's success, there is a clear place to start. If one of the less successful golfers were to start hitting more greens in regulation and eventually hit a new average score that they hover around, it might be time to look into multiple factors. That's what the data told us with Jordan Spieth and the other elite golfers from this data set.

6 CONCLUSION

After evaluating PGA Tour data from 2010 through 2018, it was evident that there is not one thing that will make someone a better golfer. There are places to start improving, but once an area has been worked on, other areas need attention. This was proven with linear and multiple regressions, as well as a comparative analysis of the best golfers in the data set with the rest of the group. As a golfer gets better, more and more factors affect success for them; it isn't a simple linear relationship between one aspect of the game and their success.

REFERENCES

[1] Rory P. Bunker and Fadi Thabtah. 2019. A machine learning framework for sport result prediction. *Applied Computing and Informatics* 15, 1 (Jan. 2019), 27–33. https://doi.org/10.1016/j.aci.2017.09.005

[2] Matt Courchene and Will Courchene. 2019. *Predicting golf is tough, but one model spits out winners and contenders on the regular.* Retrieved December 6, 2019 from https://www.pinnacle.com/en/betting-articles/Golf/golf-predictions-data-golf-model/GBX2H2EDPRB8MH2Y

[3] Morris Pickens Robert Rotella David Belkin, Bruce Gansneder and David Striegel. 1994. Predictability and Stability of Professional Golf Association Tour Statistics. *Perceptual and Motor Skills* 78, 3 (June 1994), 1275–1280. https://doi.org/10.2466/pms.1994.78.3c.1275

[4] Thomas N. Dorsel and Rob J. Rotunda. 2001. Low Scores, Top 10 Finishes, and Big Money: An Analysis of Professional Golf Association Tour Statistics and How These Relate to Overall Performance. *Perceptual and Motor Skills* 92, 2 (April 2001), 575–585. https://doi.org/10.2466/pms.2001.92.2.575

[5] Christian Drappi and Lance Co Ting Keh. 2019. Predicting golf scores at the shot level. *Journal of Sports Analytics* 5, 2 (April 2019), 65–73. https://doi.org/10.3233/JSA-170273

[6] Stephen Hennessey. 2019. *Betting on golf: How our experts have correctly predicted nearly every winner this season.* Retrieved December 4, 2019 from https://www.golfdigest.com/story/betting-on-golf-how-our-experts-have-correctly-predicted-nearly-every-winner-this-season

[7] Kyle Porter. 2018. *Predicting golf is tough, but one model spits out winners and contenders on the regular.* Retrieved December 6, 2019 from https://www.cbssports.com/golf/news/predicting-golf-is-tough-but-one-model-spits-out-winners-and-contenders-on-the-regular/

[8] Robert J. Quinn. 2006. Exploring Correlation Coefficients with Golf Statistics. *Teaching Statistics An International Journal for Teachers* 28, 1 (Jan. 2006), 10–13. https://doi.org/10.1111/j.1467-9639.2006.00229.x