**Modeling and Visualizing Tournament Performance on the PGA Tour**

Sports Analytics Final Project 2024

Josh Garzaniti

**Introduction:**

Arnold Palmer once said, “Golf is deceptively simple and endlessly complicated.” Within the game of golf, there are varying levels of detail one can focus on when analyzing a player’s performance. On a round-by-round basis, golf can seem incredibly simple to the naked eye. A player goes through 18 unique holes on a course, marks down how many shots it takes them get them from the tee box to underneath the green flag pin for each hole and then total them up to see how far above or below par (a usually implied 72 score for each round), they were. Yet, on an individual hole-by-hole basis, there are far more considerations. The par of each unique hole, the length, the shape, the hazards in play etc. All these factors scrutinize a player’s accuracy down to a stroke specific level of detail and beg the question, how do we analyze player performance not just over the course of any given round, but down to individual strokes that either positively or negatively impact a golfer’s performance.

Developed in 1999, *ShotLink* was the PGA Tour’s solution to answering that question. *ShotLink* enabled golf analysts to capture raw stroke data via the use of digitally scanned course maps, alongside an assortment of radar, camera, and laser sensors that would track ball flight and the final starting/resting position of every shot ever taken on the Tour. (PGA TOUR) The problem this project aims to take on is how to visualize player performance with *ShotLink* data using both par for each hole and overall player performance within a tournament as target variables. While the concept of tracking player performance with stroke data itself is not a novel idea, especially when it comes to sports analytics *GitHub* users (Flaska), the heat mapping and hole-specific components of data visualization, remain a meaningful scientific pursuit. With 38 events, nearly $400 million in prize money, and postseason bonuses totaling an additional $75 million in 2024(Kelly), stroke analysis and visualization on the Tour has become more valuable than ever to players, coaches, analysts, and fans wanting to breakdown how professional athletes will perform during their respective careers.

**Background:**

*ShotLink* data is no longer readily available to the public as of 2018, at least not in its public form. The PGA Tour sold the rights to data obtained through the transformational system to IMG Arena, who then brokered it to a variety of sportsbooks (Carp). So, for the purposes of this research project, analysis is composed of previously acquired data from a *GitHub* repository. With that said, while the data itself is based around the 2011 *PGA Tour Championship*, it is still thorough enough for analysis. The following research will compromise of breaking down the tournament data into stroke analysis with par (by hole), total par (for each round), and overall player scoring as outcome variables. The end goal will be to both project player performance based on the X, Y, and Z coordinates of their shot locations, and then also visualize how ending shot location impacts a player’s ability to make par both on a hole-by-hole level.

**Related Work:**

As previously mentioned, the *Github* repository that this project is based on currently exists. In it, the user plotted out shot maps by their coordinates on a grid. Those points were then colored by hole and overlayed on top of a *Google Maps* image to help users understand the context of each clustering in relation to the courses specific golf holes. Points were then further broken down into specific color shades based on their laser locations, which naturally segmented shots into fairways, bunkers, roughs, out of bounds/water, and greens. It looks like two maps were created for that project: one which shows the shots themselves colored by location, and another that takes the exact same points and colors them based off how many remaining strokes are left from each location. The way that user also calculated the remaining shots for each point was different than this project. In theirs, they took the score each player had for each specific hole minus that player’s “rolling strokes” and then added one.

They also had a separate project in their *GitHub* for exploring strokes gained. In that project, the user built a linear model alongside an *XGBoost* model which analyzed the log transformation of yards out and strokes remaining to attempt mapping out predicted player performances built on shot locations, but never actually got around to the strokes gained/lost part of their analysis.

**Data Description:**

As previously mentioned, the *ShotLink* data obtained for this research is primarily compromised of one singular tournament: *The 2011 Coca Cola PGA Tour Championship*. The data itself contains 8422 observations of 37 variables, so it’s suitable for data analysis both with the depth of the unique player shot values and in terms of the different factors and variables that can be analyzed impacting performance. Since it’s only one tournament, there’s a few variables that will remain constant throughout the dataset such as the number of players, the golf course name, tournament ID number, course number, tour code, tour type, year, and “Shot.Type.S.P.D”. The data itself also contains no naturally occurring N/A values, which is incredibly lucky for a raw *ShotLink* dataset. With that said there are manual entries made for the variable “Lie” in which N/A was assigned to a player’s lie that was not available. For cleaning and analysis, there are also values within the data frame that are set to “unknown”. This is most likely because the *ShotLink* course tracking systems couldn’t specifically pick up the ball’s lie, position, width, shot location or laser location. Coordinates that are set to “0” also represent shots that ended up in the hole, so that needs to be taken into consideration.

The dataset was then transformed through a few means to increase the scope of analysis and make it suitable for the purposes of this research. First, column names were cleaned up to make variable interpretation easier. Then, player names, which were first given as two different variables (First and Last name) were combined to make one player name column. The data frame was then organized and grouped by player name, and round to summarize the number of total shots each unique player per round. That was then ungrouped and pivoted by “Round” so that shots could be interpreted from a round-to-round basis. A “Total Score” variable was created to sum up the number of shots a player took over the course of the entire tournament. Then a binary “Par Made” outcome variable was created, which utilized an if statement comparing each player’s “Hole Score” to the “Hole Par” value for each player on each hole over each round. In this way, player performance can be analyzed down to a shot level by referencing if the ball placement coordinates of each shot location impacted their ability to make par.

**Methods:**

There were five main methods that were involved in the exploration of this PGA tournament data. Firstly, to help initially visualize shots taken on the course, a scatter graph based off a random forest model and clustering were used to visualize and group shots by their X, Y, and Z coordinates over the range of the entire golf course. Then heatmaps were used to display “Strokes Remaining” by shot location coordinates mapped out over a grid. After attempts to overlay heatmaps with images of the golf hole, it was suggested to display strokes remaining via scatter plots and density graphs. As such, one was made for each unique hole on the course. An initial general additive model was used to first derive expected shots remaining from the coordinates supplied, but was then replaced by a better random forest model that would do the same job but process at a higher level. Random forest models and an *XGBoost* model were both used when examining shots/strokes gained, and finally the last XGBoost model was used to make predictions on total strokes gained in correlation with strokes gained over each specific area of the course.

**Results:**

The results of this project displayed that with the right processing power and analytical models in place, it is possible for one to take raw shot locations combined with historical player scores and turn them into meaningful analysis, predictions and data visualizations. Initial shot mapping attempts (Figures 1, 6, and 7) represent a clear evolution in using machine learning to model shots initially from coordinates within a grid system to using derived variables such as Expected Shots Remaining (Shots\_Left\_Hole) to model what course shot models could look like based on historical player stroke level data. Advanced scatter point and density graphs (Figures 8 and 9) took that to a higher level, displaying Expected Shots Remaining alongside detailed landing zones to give players and analysts alike a better visual representation of hole-level shot analysis (where are clusters of common landing zones for each unique hole, which can give insight into where players are aiming/targeting). Variable analysis in figure 5 displayed that in terms of modeling Total Strokes Gained, the highest key indicators were strokes gained driving and strokes gained putting, at least for the 2011 PGA Tour Championship. Graphing out players by their individual Strokes Gained performances along the championship can also be broken out into a hole level of detail alongside a round level of detail (Figures 10 and 11), where analysts can pick apart high/low performing outliers such as Bubba Watson losing almost 10 shots alone in the 3rd round as a whole or Dustin Johnson finding a way to gain 4 strokes just on hole 9 par 3. Finally, the XgBoost model used (Figure 12), while slightly undervaluing player performance in terms of strokes gained, does display the ability to closely predict gains/losses over the course of a tournament.

**Discussion:**

Based on the results of this project, shot location data obtained through *ShotLink* can be transformed into graphics giving new insight into projecting a player’s remaining shots for unique holes during PGA tournaments. While heatmaps being overlayed on top of course maps can be difficult themselves to align properly in terms of the shading values for individual holes, density maps help to display the same logic in a different format. So given a tournament has raw shot coordinate tracking systems in place, it’s possible for one to take that data, wrap the “Expected Strokes Remaining” and “Shots Gained” models developed during this project as functions, and apply said functions to future data given it can be pre-processed in the same way. In terms of player analysis, through modeling strokes gained/lost on a tournament level, *XGBoost* models built on strokes remaining both before and after current stroke data can accurately predict gains/losses to a degree, since the model used in this part of *PGA* analysis closely followed the reference line of given historical data. That said, the model used also tended to undervalue player strokes gained, so it might be a bit harsher than reality. Many insights displayed show both an exciting new way to visualize and interpret golf from an analytical background, and yet still leave more to be explored within the game.

**Conclusion and Further Work:**

In conclusion the analysis done of golf players’ performance for this project is incredibly fascinating, and yet at the same time there is more that can be done going forward. It was already touched on just how many levels shot and player performance can be broken down to. Here, there was more of an emphasis on the higher levels of shot detail and an exploratory analysis into strokes gained/lost, but one can only imagine how many times this can be broken down over each unique round, hole, and even shot location. One thing that would be beneficial to explore would be individual hole analysis and seeing where on specific holes can players either add to their gains or even start to make up shots if they’re behind. Say a player was in the rough after one of their drives, it could be worthwhile to understand whether chancing a draw around a tree or a simple punch out to the fairway sets them up for better odds in terms of fractional strokes gained and, in that case, a lower hole/tournament score. That is the next step of these analytics in terms of delivering new insights. Not just projecting player performances but prescribing them with solutions once they fall off pace or enabling them with the same information to keep them consistently ahead of others.

Appendix

A group of colorful spots

Description automatically generated

**Figure 1.** Containing the initial mapping of each shot location via their coordinates.

A screenshot of a computer program

Description automatically generated

**Figure 2.** The initial random forest model results used to predict shots remaining for each hole.

A graph with numbers and a line

Description automatically generated

**Figure 3.** The variable analysis graphic produced from Fig 2’s model.

A computer screen with white text

Description automatically generated

**Figure 4.** The random forest model used to derive Total Strokes Gained.

A graph with numbers and lines

Description automatically generated with medium confidence

**Figure 5.** The variable importance graph of the Strokes Gained random forest model.

A screenshot of a computer screen

Description automatically generated

**Figure 6.** Heatmap of the first hole without background context.

A screenshot of a computer generated image

Description automatically generated

**Figure 7:** Representing the Heatmap of Strokes Remaining for Hole 1 overlaid against the Course map of Hole 1.

A screenshot of a computer screen

Description automatically generated

**Figure 8.** Scatterplot of the first hole over the entire tournament by ESR.

A screenshot of a computer screen

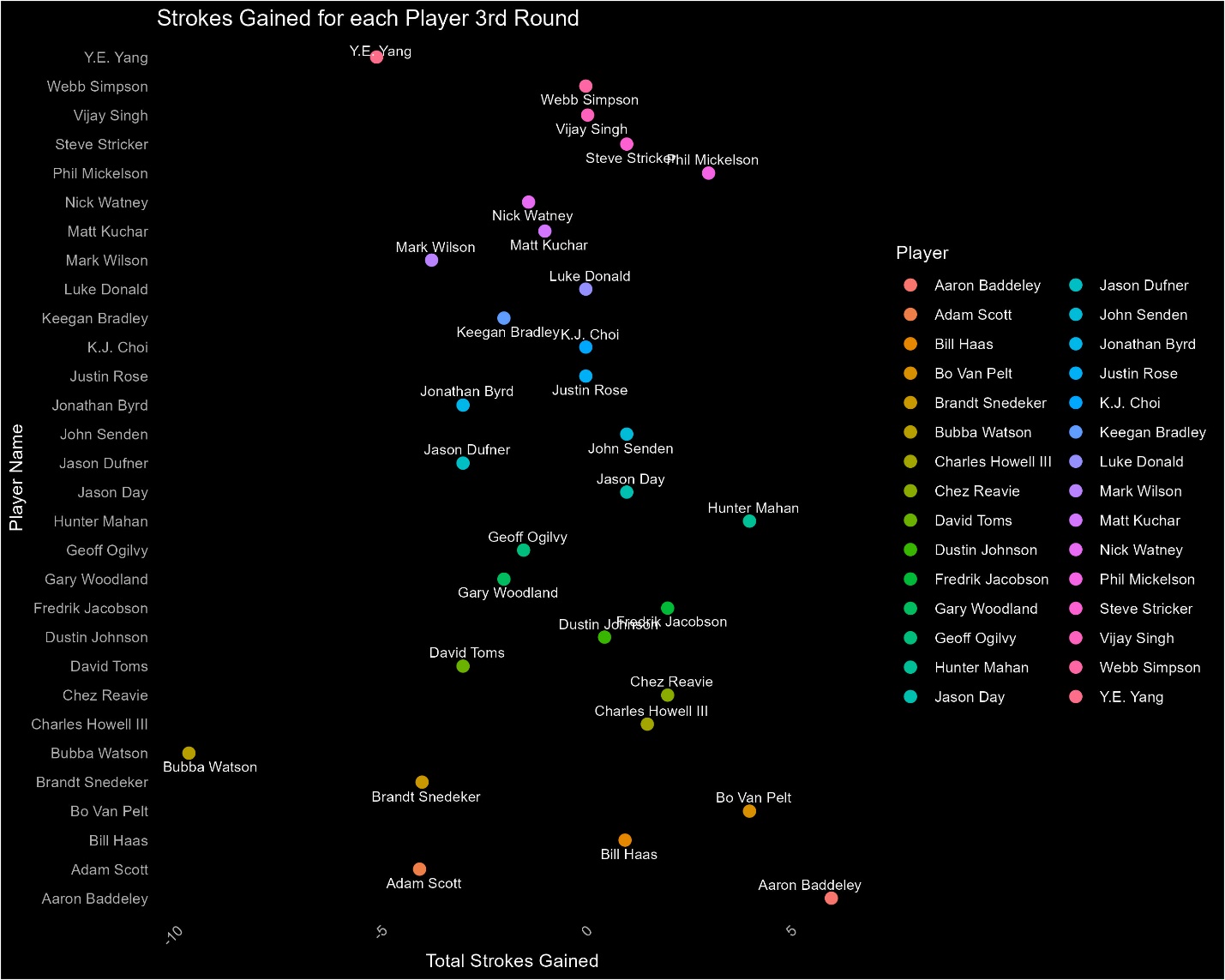
Description automatically generated

**Figure 9.** The density graph of shots on the first hole over the tournament.

A screenshot of a computer screen

Description automatically generated

**Figure 10:** The player SG chart representing how individuals did in terms of gained and lost strokes over the 9th hole for the entire tournament.



**Figure 11:** The chart displaying how players performed in terms of total strokes gained over the entire 3rd round.

A graph with blue dots and red line

Description automatically generated

**Figure 12:** Displaying Strokes Gained predictions made by the XgBoost model relative to their actual values for the tournament.

Works Cited

Carp, Sam."PGA Tour Expands Sports Betting Data Partnership with IMG Arena." *SportsPro Media*, SportsPro Media Ltd., 19 Nov. 2018, <https://www.sportspromedia.com/news/pga-tour-sports-betting-data-img-arena/>.

Flaska, Scott. “02\_shotlink\_explore.” *Scott Flaska’s Blog*, GitHub, 7 Mar. 2022, scottflaska.github.io/blog/posts/02\_shotlink\_explore/.

Kelly, Todd. "Prize Money Payouts for PGA Tour Tournaments in the 2024 Season." *Golfweek*, USA Today, 7 Nov. 2023, <https://golfweek.usatoday.com/lists/prize-money-payouts-pga-tour-tournaments-2024-season/>.

PGA TOUR. "ShotLink History." *ShotLink*, PGA TOUR, Inc., 2016, <https://shotlink.com/about/history>.