Predictive Modeling of World Golf Rankings

James Tyler
james.tyler@yale.edu

Advisor: Stephen Slade
stephen.slade@yale.edu

Mathematics Reader: Tamunonye Cheetham-West
tamunonye.cheetham-west@yale.edu

A Senior Thesis as a partial fulfillment of requirements for the Bachelor of Science in Computer Science and Mathematics

Department of Computer Science
Yale University
May 2, 2024

Acknowledgments

I would like to thank my advisor, Stephen Slade, for helping encourage me through the course of the semester, and for advising me to complete a project which I found interesting. I would also like to thank my math reader, Tam, for helping me think through the modeling aspects of the project, and for his support during the semester.

I would also like to thank my family, my Mom, and my Dad for their support of me. And to my partner, Aliaksandra, thank you for all you do for me.

Contents

Predictive Modeling of World Golf Rankings

James Tyler

Abstract

This paper sets out to develop and implement a world golf ranking. The inspiration for the project came from news stories surrounding the development of LIV Golf, and whether statistically driven rankings, such as that of datagolf.com, are better indicators of golfer performance than the Official Golf World Ranking (OWGR), which uses a points based approach. The datagolf ranking utilizes regression, so we also set out to test the effectiveness of machine learning algorithms on this problem. We wrote web scrapers to gather score level professional golf data, and fit a regression to determine adjusted strokes gained (independent skill level of round) across our dataset. We then fit models to predict each player's adjusted strokes gained, which becomes their Performance Rating, and ultimately the ranking. We found that Gradient Boosted Trees did not exhibit increased predictive power over regression, when fit with a limited feature set. Further work remains to be seen if with additional feature engineering and data collection if these models could be useful. The result of this project is the testing of these models, and the world ranking it develops.

1 Introduction

1.1 The Official World Golf Ranking

Men's professional golf developed its first ranking system in 1986, called the Official World Golf Rankings (OWGR) [1]. As of the past few years, the ranking has been the unquestionable source of who are the best professional golfers in the world. The ranking works by assigning points values to tournaments, based on some heuristics (such as the four major championships - The Masters Tournament, The US Open, The British Open, and The PGA Championship - assign 100 points to their winners). The number of points a player earns is divided by the number of events they play (minimum of 40 divisor), with a half-life applied to points earned in older tournaments. For non major championships, the points total would be determined by the strength of the field. Each player participating in the tournament would contribute points to the strength of the field (SOF) based on their OWGR ranking. The top ranked player would contribute more than the second ranked player, etc.

In recent years, almost all of the top players have played primarily on the PGA Tour. This means that given this SOF calculation, the PGA Tour events award more points to the top finishers than any other tour. This had the effect that in order to be a top ranked player, you must play on the PGA Tour. One effect of this fact is that top players from other tours, in particular the European Tour, started to leave for the US to play on the PGA Tour, as there were more opportunities to earn ranking points, which are used as the primary metric for determining who qualifies for the major championships. As the top players left the European Tour, it had fewer financial resources, and has decreased in relevance in recent years. However, in 2022, a new tour started that began to rival the financial resources of the PGA Tour, called LIV Golf. LIV attracted many top PGA Tour players under guaranteed contracts to play on their tour, including top players Jon Rahm, Cam Smith, and Bryson DeChambeau, all of whom were ranked in the top 10 in the OWGR when they left for LIV. The OWGR chose not to award ranking points to LIV tournaments. Over the past two seasons, many of the top LIV players have remained some of the best golfers in the world, however their OWGR rankings have plummeted since they can only receive ranking points in the majors.

1.2 Modeling Challenge

The challenge presented is that it is necessary to compare the skill level of players who mostly do not compete against each other, but rather in two separate settings, with a limited overlap (specifically the four major championships). One common approach to this problem is to estimate the strokes gained by a given player. Strokes gained is defined as the performance in a given round (or even a given shot, as detailed by Mark Broadie in his seminal work on the subject in 2011) against a baseline player (normally defined as an average PGA Tour player [2]. This approach allows us to compare rounds on different days and in different locations, as you can say how many strokes better a given player is than the average player.

One consideration to be made is that strokes gained is only against the field of players competing in a given tournament. Thus, some tournaments have stronger fields than others, and gaining one shot in the two tournaments has different implications on golfer skill level. To address this, it is necessary to calculate an adjusted strokes gained metric, which normalizes for field strength. This is precisely the approach taken by the main competitor to the OWGR, datagolf.com [3]. It is possible to fit a regression, as described by Connolly and Rendleman in their 2008 paper on PGA Tour player skill level, luck, and streaky play [4]. In this regression, we consider each golfer's ability as fixed over a given year long period. Then, the golfer skill level will be estimated based on the tournaments they played in and their raw strokes gained, as well as the difficulty of round, relative to all other rounds played. While the estimate of golfer skill level could be directly used to compute a ranking, it is not generally the case that golfer skill level is fixed over a year long period. Hence, the adjusted strokes gained approach of datagolf is to use the estimate of the round difficulty only (which depends greatly on the fixed player skill level estimates), to calculate adjusted strokes gained as raw strokes gained - round difficulty rating. Then, it is possible to treat a player's sequence of adjusted strokes gained inputs as a time series, which we can then build models to predict. This allows for variable golfer performance based on inputs to the model, which include exponentially weighted moving averages of adjusted strokes gained over varying half lives, number of rounds played, and time off since most recent tournament. The result of this model for predicted golfer performance in their next tournament round is hence the skill rating of a player used for the ranking, instead of the fixed estimate of their skill.

The goal our study set out to achieve is to gather the dataset of professional tournament rounds, recreate the adjusted strokes gained time series as described above, and then fit to this time series with many types of machine learning models, to compare the performance to the already existing models. The model datagolf uses is a regression based model, which while effective in this setting, can be limited in predictive power against gradient boosted trees, support vector machines, or neural nets. Our goal was to test these models, to see if they have better out of sample predictions of golfer ability than that of the regression model.

In order to do this, we first built web scrapers to gather the necessary data. Once we had the data, we calculate the fixed effects regression in order to attain the adjusted strokes gained time series. Then, we divided our data into training, testing, and validation sets, in order to properly evaluate model performance. We then proceeded to fit and fine tune our models, and then compared performance in the test set against the baseline regression model.

2 Background

2.1 Structure of Professional Golf Landscape

For those unfamiliar or vaguely familiar with professional golf, here is a high level view of how professional golf is structured. This is important to understanding how to construct a ranking. The OWGR officially recognizes 24 separate golf tours around the world. These tours are eligible to receive ranking points for their events, which must span either 3 or 4 rounds. Generally, a professional golfer will be a member of one or more of these tours. In addition to the 24 recognized professional golf tours, there are the four major championships: the Masters Tournament, the US Open Championship, the Open Championship (British Open), and the PGA Championship, which are not affiliated with any specific tour, but instead set invitation and qualification criteria to their events. The fields of the majors range from about 100 players in the masters, to 156 for the PGA Championship. One of the qualification standards is usually either a top 50 or top 60 ranking in the OWGR as of a fixed date before the championship. This means that the accuracy of the ranking affects who can participate in these events, which also receive 100 OWGR ranking points to the winner (all other events, regardless of tour, are capped at maximum 80 points to the winner, with the majority of events of most of the 24 eligible tours awarding less than 10 points to the winner for nearly every event).

Most of the top ranked golfers belong to the PGA Tour, which apart from the major championships, hosts the events with the consistently strongest strengths of fields (normally average 40-60 OWGR points to the winner). Top ranked players may also belong to the European Tour, having dual membership. While some players belong exclusively to the European tour. There are also development golf tours, which award top performing players membership to the top golf tours at the end of the season (in the US this is the Korn Ferry Tour, and in Europe this is the Challenge Tour), which are also eligible for ranking points. There are many other tours, such as the Asian Golf Tour, Australian Tour, and Sunshine Tour (South Africa), which host events through the year, but do not consistently draw large numbers of players from the highest tier of the PGA Tour and European Tour, apart from a select number of marquee events, which may draw wider participation.

The other recently formed tour, which is not eligible for OWGR ranking points is the LIV Golf Tour, which recruited players from the PGA Tour and European Tour to become members by paying them guaranteed contracts, but in doing so they lost membership of their original tour (now due to litigation LIV members can pay a fine to regain European Tour membership). LIV events do not receive ranking points because they do not meet the OWGR eligibility requirements, which are designed to foster competition. These include having a system whereby players can earn status in the tour through exceptional performance, while worse performing players lose status on the tour (this has since changed some). Also, the majority of events are required to have a cut structure, where before the final round or rounds of the tournament, the bottom performers in the field are removed from the tournament, and are no longer eligible for ranking points or prize money for that event.

There are other classes of events which also do not receive OWGR points. Elite amateur level competition, especially in the US with the NCAA golf structure, often is a pipeline for developing players who go on to be top ranked golfers. For instance Tiger Woods played at Stanford, Arnold Palmer at Wake Forest, and Phil Mickelson at Arizona State. More recently, stars like Jordan Spieth (Texas), Justin Thomas (Alabama), and Rickie Fowler (Oklahoma State), have quickly risen from amateur level college golf to professional golf. There are amateur events around the world, including elite level competitions like the US Amateur and British Amateur. In addition, there are professional mini golf tours, which do not receive ranking points, but are often used by players without tour membership to hone their game, in hopes of qualifying for membership of a professional tour. However this paper does not assess golfer performance in these events (as the data were not collected).

One of the main challenges of constructing a world golf ranking is the siloed nature of the professional golf landscape. Since players compete on separate tours, while it is possible to order players based on their relative performances against their peers on their tour, it is less obvious how to order them relative to members of a separate tour, when the two tours rarely have members compete directly with each other. There are the major championships, as well as cases when players players take a week or two to play abroad, which provide some opportunity to make construct a network where the players can be ordered across tours. Also, the top tours award more prize money, and hence top players are incentivized to play their events, as they are better compensated. This creates a feedback loop whereby the top players strive to play the PGA Tour, and usually play lower ranked tours if they do not earn membership, hoping to make their way up the ranks. Thus, it is generally accepted that the PGA Tour has the best golfers. However, in constructing a ranking it is important to quantify the amount by which we believe the average PGA Tour player is better than the average player of any other event. This is where the fixed effects regression discussed in the introduction (and detailed in the methodology section) becomes important.

2.2 Ways to Construct a World Ranking

Let us start with an analogy to constructing a golf world ranking. Suppose we wanted to construct a ranking of the top football clubs in the world. While English clubs compete against each other (although there are different levels, such as the Premier League and the Championship, which would only compete in the FA or Leagues Cup), as do Spanish and Brazilian clubs in their national leagues, they do not directly compete against each other. The top Spanish and the English clubs may compete with each other if they make the Champions League. Meanwhile, the Brazilian clubs may never compete with the European teams, except for friendlies, or an occasional match in the Club World Cup. It is widely accepted that the European clubs are consistently the best, while South American clubs are second best, followed by North American, Asian, and African clubs in some order. This is known from the limited intercontinental data points, as well as eye tests of the teams. One objective way to do this would be to construct an ELO style ranking, where teams win or lose points based on head to head performance in matches, as well as winning / losing more points for higher goal differential, or for playing a team with a more distant rank. However, there are also other data. For instance, the European clubs pay the highest salaries and transfer fees, so naturally the most talented players are recruited to play there, as the incentives are better. If we wanted to estimate how many goals Manchester United would defeat LA Galaxy by on average, it would be necessary to analyze all of these factors to fairly price this matchup.

For golf, the head to head matchup structure of football does not translate as easily. Instead of two golfers competing directly against each other, tournaments consist of 50 - 150 players, all of whom play the same course on the same week over a span of 3 - 4 days. Golfers play in groups of 2 - 3 players, starting in waves in the morning and afternoon, on the front and back nine, in about 10 minute intervals. Thus, different players may face different weather conditions (increased wind speed has a large negative impact on golf score). Ultimately, by the final round, the players with the best total scores up to that point start last, with the winner (and final positions) determined by the total score at the completion of all rounds.

The OWGR ranking system is to calculate the number of rating points every player in the field will contribute to the assigned Field Rating. If the number one player participates, he will contribute the most possible points. This number can vary, but let's say he contributes 15 points. If the number 1 player does not participate, these points do not get added to the field rating. If instead the 300th ranked player takes his spot, this player may only add 0.1 points to the field rating, and as a result all players at the end of the tournament will get less ranking points for finishing in the same position. This structure makes sense, because in expectation the top ranked player will finish close to the top of the leaderboard, pushing all other players down. As players get a larger percentage of the total points allocated for finishing in higher positions (first place may receive 15-25 percent of all the points, decreasing from there), finishing behind the top ranked player is thus less penalized as more total points were available. For every event a golfer plays, they will get points for their finish, and their ranking is defined as the average of the points they earned in the events they played (with a minimum divisor of 40 events), with a half life applied so that points fall off over a two year period.

This is a system that is recognized by the top tours and the major championships, and has been in place since 1986. However, it is susceptible to a few issues. First, not all players get the same number of opportunities to play. Depending on their tour membership priority ranking, they may not be eligible for all of the events. As fields are limited in some cases to 70 or 100 players, the players in positions 100 - 150 do not get into the top events, which have the strongest fields, and hence the top point earning potential. Moreover, as the LIV events are not ranked, players of that tour only earn points from playing the majors, and hence have very limited point earning opportunities). Additionally, this model does not take into account the scores which a player shoots. A player winning by 10 shots receives the same number of points as a player who wins in a playoff (like overtime, when two or more players are tied, they play one hole at a time until one wins). Also, we can model golf scores for a given player as coming from a (relatively) normal distribution (this distribution is usually left skewed, as very bad rounds are more likely than very good rounds). While you can argue that the player has some control over their performance, the score a player shoots has some mean and variance, and where their score is drawn from in their scoring distribution often has elements of randomness and luck. This is inherent to golf, as many factors are outside of player control. For instance, the probability of making a putt from 8 feet away is exactly 50 percent on the PGA Tour, which has the best golfers in the world. While some players may be better or worse, ultimately putting has high variance, as in one round a player will make three in a row, while in another he will miss three in a row, even though his putting ability in the two rounds was roughly the same.

Therefore, another way to construct a golf ranking is to estimate the mean of their scoring distribution. The better players have a higher mean, while the worse players have a lower mean. This is the approach taken in this paper.

3 Methodology

To start, using the Python Beautiful Soup library, we wrote web scrapers for owgr.com and livgolf.com. The choice of these two websites was for the following reasons:

- owgr.com contains the round level scoring data for all of the tournaments of the 24 OWGR sanctioned tours. As a result, a working owgr.com web scraper effectively means saving the work of writing 24 separate web scrapers. This is especially important given that the professional golf tours frequently update their websites, and it is more practical to rewrite 1 web scraper than it is to monitor and rewrite up to 24 web scrapers.
- One downside of this approach is the lack of webscraping of pgatour.com. The PGA Tour includes on their website shot level location data, which is the foundation for the category strokes gained databases on the PGA Tour. While not every tournament includes shot level data, the majority do, going back at least 15 years. We attempted to gather this data, but it proved to be impractical, as the data was not aggregated in any one place, and was contained over many player round web pages under the pgatour.com domain, which was impractical to gather.
- It became necessary to scrape the LIV Golf website, as this is not an OWGR sanctioned tour, and part of the motivation for the project was to include the data from LIV events.

The owgr.com domain contains two main categories of pages: player level pages, and event level pages. Player pages contain just that player's score, where the url contains the playerid of that player, while the event pages contain all the scores for a given event, with the tournamentid in the url. We wrote a web scraper for the event level page, as then we can a single url for every tournament, and get all round level scores by parsing the html with Beautiful Soup. For the livgolf.com domain, we also accessed round level webpages, but liv has dedicated developer webpages containing the tournament scores in jsons, which was even easier to access. We built these tools ourselves, as while there was an owgr web scraper published to github, it was outdated (from about 4 years ago), and did not attempt to access score level data. The one major source of data that is missing from this project is amateur level data, as from wagr.com. While this would be accessible, we did not think to add it until in the later phases of the project, at which point modeling had taken over as the main focus. One result is that for certain amateur players with astounding results, they may not appear in our rankings (as the threshold to be ranked is a minimum of 20 rounds played in the previous 2 years to the ranking date). Another side effect is that for specific amateurs, who have more than 20 professional rounds, such as Luke Clanton, who performed exceptionally on the PGA Tour as an amateur in summer/fall 2024, their rating will be based only off their professional rounds, and not amateur level events too. This can cause a discrepancy between our rating at datagolf.com's ranking, although it is limited in scope to only a few players.

Once we collected our dataset, we moved to working with it in python in a Jupyter Notebook. Our primary method for manipulating the data was the pandas and numpy libraries. The first step is to aggregate the data from the owgr and liv together, making sure that each liv player gets matchued to their results from the owgr database, in order to maintain their scoring record. Then, it is necessary to melt the dataframe, so that each row contains the information of a single tournament, player, round triple. When the data is collected, it is in tournament, player pairs, so if the tournament lasts 4 rounds, there would be 4 columns R1, R2, R3, and R4, containing the scores. Instead, we want four rows, where Round now becomes a column of the data set, and Score is now the score for the given round on that row. The pandas.melt function achieves this transformation.

The next step is to convert from raw score to strokes gained. The first measure of strokes gained is strokes gained against the field on a given round. To get this measure, we first must group by tournament round, and then take the mean of the Score column, this is the round scoring average. Then, for every player, we take their score and subtract the round scoring average, and this is now raw strokes gained. We call it raw strokes gained, as it depends only on the scoring average of the field, and does not adjust for the skill level of the players in the field. To see why this is not our desired metric, let us consider the following. Let's say we have two rounds, A and B, where round A has +3 adjusted strokes gained, while round B has -1 adjusted strokes gained. Naturally, we may ask which round was more impressive? We might be tempted to say that round A is more impressive, by exactly 4 strokes. Now, if they came from the exact same day on the same course, then this naive answer is indeed the correct one. However, what if round A came from a local high school golf state championship, while round B came from the US Open? While it is certainly impressive to beat the high school state championship field by 3 shots, it is very likely that the average player in this field is somewhere between 10-20 shots worse on average than the average player in the US Open, which hosts the 150 best golfers in the world each year. In this example, it is likely the -1 raw strokes gained round is at least 10 strokes more impressive than the +3 raw strokes gained round.

This example gets at the idea of adjusted strokes gained. Since raw strokes gained depends on the field scoring average, we need some notion of a baseline golfer, where we know exactly what their scoring average would be relative to any field. Then, we can adjust the scoring average of the field by the offset to this baseline golfer, and now we have an adjusted field scoring average. If the tournament is the US Open, it is likely this adjustment would be positive. That is, if the field averages 72 strokes, but the average golfer in the field is 1 stroke better than this baseline golfer, who would shoot in expectation 73, then each player in the field effectively gains an additional stroke from their raw strokes gained, to this new adjusted strokes gained metric.

Now, it becomes necessary to specify this notion of the baseline golfer. Unfortunately for statisticians, there isn't actually a golfer in the world who can play golf with 0 variance in his scoring distribution, while trying simultaneously trying to minimize his score (presumably some golfer who can reliably shoot 75, with a standard deviation of 5 shots, could be guaranteed to always shoot 125, with their method to be to play to the best of their ability for the first 17 holes, and then on the last hole, get very close to the hole, before intentionally missing enough times to get their score to 125 exactly. But even this misses the fact that on some courses, it is easier to shoot lower scores, so even shooting the same score every round wouldn't be sufficient). Since in practice it is impossible to know exactly how difficult every single course plays in every single round (as this depends on the weather, the length of tees played, the locations of the holes, the condition of the course, and many other factors), we must estimate this baseline golfer off the scores we observe players shoot. We use linear regression to solve this problem.

For this linear regression to be feasible, we first make some assumptions. First, we will assume that every single golfer is of some fixed skill level over the regression window (we use two year windows). In practice, we can imagine a golfer's scores as drawn from a personal scoring distribution, with some mean and variance. We could model this as a normal distribution (in practice it would close to normal, but left skewed, as very bad rounds are more likely than very good rounds). Then every golfer has their score for a given round drawn from their distribution (it's not really drawn, they play the round and shoot what they shoot, but there is certainly a random component to their score, combined with some inherent skill component (their average score / expected score). Thus, to estimate each golfer's distribution, it is useful to assume that their distribution is unchanged over a given interval. Then, we observe their distribution, and we can what that golfer would shoot in expectation, instead of simply relying on the score we observe.

In this model (developed by Connolly Rendleman, and which we learned from datagolf.com) [4], our target variable is raw strokes gained. We can say that a player's raw strokes gained is composed of three components: their underlying skill component, the random component, and the field/course difficulty component. This final component is the one we are most interested in estimating, as this is the offset we describe in the adjusted strokes gained methodology. Our model is:

$$RawStorkesGained_{i,j} = Skill_i + Random_{i,j} + FieldDifficulty_j$$

Note, i indexes players, and j indexes tournament rounds. We can fit this regression with our dataset. First, we need to generate two dummy matrices, one player dummy matrix, and one tournament round dummy matrix. Thus, for every row of the dataset, we will have a single 1 in the player dummy matrix, for the player who played the round, and also a single row in the tournament round dummy matrix, for the tournament round the entry occurred in. We build these dummy matrices with OneHotEncoder's fit-transform method, generating sparse matrices. Then, we fit a linear regression with sklearn. The coefficients of this lienar regression match up directly to the estimate player skill levels, and the estimated field difficulties for the tournament rounds. Populating these values into our dataset, it is then possible to get adjusted strokes gained by subtracting the round difficulty rating from the raw strokes gained value. An additional adjustment we make is to also subtract the average round difficulty rating of the PGA Tour rounds in the regression. This way the average PGA Tour adjusted strokes gained is 0. This is nice, as it makes the values more interpretable, but in practice the choice of this number is arbitrary, and we could subtract any constant, and it does not affect the results or relative values.

One thing discussed by datagolf [3] about this regression, is an obvious question that arises: why don't we just use the estimated skill rating as our golf ranking? If this was the case, we would be done right now! One obvious issue is that the model assumes that player skill is fixed over a two year period, so our ranking would also be estimating the best player over a fixed two year period, which is not necessarily the same as the best player today. Thus, a different approach is to instead utilize this model for the round difficulty ratings in order to obtain adjusted strokes gained data, and then to predict the adjusted strokes gained time series, the results of which will be the ranking.

Before we can fit a model, we first need to construct features. The principal type of feature we found useful was a player's average score. We wanted a causal feature, in the sense that any feature we feed the model could be calculated from known values prior to the round occurring. That means that the scoring average for all round prior to the given round could be used. In this setting, instead of a raw scoring average (of adjusted strokes gained), we instead opted for a exponentially weighted moving average (EWMA) of a player's adjusted strokes gained. The EWMA is calculated by applying a halflife to a player's rounds based on how long ago the round was from the current date. This decay factor can be adjusted, with the code written so that the units were in days. The values then fall off exponentially. Two additional constraints we imposed were the maximum time a round can be in the window, which we set to 2 years (after which point a round has weight 0 in the EWMA), and the minimum number of counting rounds needed in a player's scoring history before we calculate the EWMA (we used 20 rounds as the minimum). Using this, we calculated the EWMA for four different halflives: 1 year, 3 months, 1 month, and 1 week. The idea of the different halflives is that the longer halflives capture play over a longer period (still weighted towards recent play, but less dramatically), while the shorter halflives capture recent play more. These features are correlated, but when when we fit a linear regression with these four features, we found that all four regression coefficients were significant, with t-statistics between 8 and 60. When we tried additional EWMAs, including a two year EWMA, the t-statistics would decrease, with some having t-stats less than 2. Thus, we settled on these as our predictors. The regression we fit was also produced sensible coefficients, as all EWMA coefficients were positive, and their sum close to, but less than 1. This means that all four intervals of play contribute some to a player's ability according to the model, in a significant way, where their total contribution (if we suppose all EWMAs are close to the same value), is roughly that of the player's scoring average, although a little bit less than 1, to imply some mean reversion.

Some other candidate features would be the player skill ratings and field difficulty ratings obtained from the regression described before, but we elected not to use these (even though when supplying them, there was an increase in r-squared, and the regression coefficients were significant), because their calculation had look ahead (as the regression was fit on two year windows). Also, as the adjusted strokes gained depends directly on the round difficulty, as that is the offset used in calculating adjusted strokes gained, it feels like supplying this value to the regression isn't the best idea, as it directly is used in the calculation of the target variable, so has some look ahead in a sense. We could fix the look ahead by fitting the fixed regression at every date, but this is impractical, and the benefit to the predictions of the model would be on the order of at most a couple r squared points.

The other modeling technique that we set out to attempt, and did do, was to fit other models besides a linear regression. Our choice of model was a gradient boosted tree (GBT), which we fit with both XGBoost and CatBoost (fits GBTs with categorical data features). We were excited about trying different models, but with our set of predictors, it became clear that these models were not improving over the linear model, as we were achieving out of sample r-squared values equal to the third or fourth decimal point, indicating that the GBT was only learning the same linear relationships, rather than the hope, which is that it could learn non-linear relationships between the features. We found that if the GBT was not showing performance enhancements over the linear model, than it was best to stick with the linear model, rather than progress to neural nets, as this seemed unlikely to produce significant results, and we would not trust them given the GBT performance.

Thus, taking the results of the model, it is now possible to make a ranking. For this, the necessary step is to determine the day that the ranking will occur on, for instance September 29th, 2024. Then, filter the dataset for the last round of every player before this date, with the EWMA features calculated. We allow any player to be ranked who has at least 1 data point within 1 year of the ranking date that has non-null EWMAs (where the EWMAs require minimum 20 rounds in the prior 2 years). Then, while it would be possible to recalculate the EWMAS as if they were on the ranking date, we just used the value from this last round, and used the model to predict the adjusted strokes gained of that player. Then, for every player we have the adjusted strokes gained prediction, which we can sort, and get an ordering of the players by decreasing adjusted strokes gained. This is the ranking for that day. We fit the ranking for every 7 days, starting from mid 2014, as the round data starts in 2010, so that way there was at least 3 years of data for the EWMAs and the 1 year period. We merged the model predictions and ordering back onto the dataset, and then for any given player, we have their ranking history and model prediction history. We use these to make visualizations, which we present in the results section.

4 Results

For results, we will present the findings of the rankings, as described in the methodology section. The ranking is the result of the linear regression model, where first the adjusted strokes gained is calculated by fitting the fixed skill level regression to get estimated field difficulty ratings for each round. The features given to the model are exponentially weighted moving averages over different halflives of the player's adjusted strokes gained, for rounds prior to the ranking date. For a player to qualify, they must have at least 20 rounds within a two year period in the dataset, and their most recent round must be within 1 year of the ranking date. Using this model, the top-25 ranking for the last day in our dataset, September 29th, 2024 is the following:

Model Ranking, September 29th, 2024

model_ranking	name	model_prediction
1	Scottie Scheffler	2.954273
2	Xander Schauffele	2.564298
3	Jon Rahm	2.510045
4	Rory McIlroy	2.122786
5	Collin Morikawa	1.987199
6	Bryson DeChambeau	1.683906
7	Ludvig Aberg	1.667257
8	Hideki Matsuyama	1.646751
9	Tyrrell Hatton	1.642483
10	Patrick Cantlay	1.640291
11	Joaquin Niemann	1.618062
12	Viktor Hovland	1.586564
13	Sam Burns	1.574731
14	Russell Henley	1.548671
15	Tommy Fleetwood	1.545444
16	Corey Conners	1.463645
17	Sergio Garcia	1.405363
18	Luke Clanton(Am)	1.370812
19	Brooks Koepka	1.344556
20	Sungjae Im	1.337427
21	Tony Finau	1.334832
22	Wyndham Clark	1.323765
23	Adam Scott	1.233468
24	Billy Horschel	1.230687
25	Sahith Theegala	1.209157

Figure 4.1: The world ranking for a given day, based off model predictions

A few things to note from this result are that there are LIV players in the top 25, which is consistent with widely held belief that some of the best players in the world play on LIV, but which is not reflected in the OWGR, since those events do not receive points. Additionally, the rankings for this day match very closely with the rankings for a similar day on datagolf.com, which makes sense, since much of the methodology here follows their methodology. Moreover, at this time, Scottie Scheffler was undisputably the best golfer in the world, which this ranking corroborates, giving him a nearly 4 tenths of a shot advantage over the second best player, Xander Schauffele.

Two other visualizations that our code produces and that are useful for viewing the ranking of a player are: a history of the player's rounds as a scatter plot, with the player's model prediction overlayed as a time series, and a ranking history of a player according to the model. The results of these plots for world number 1, Scottie Scheffler are shown below:

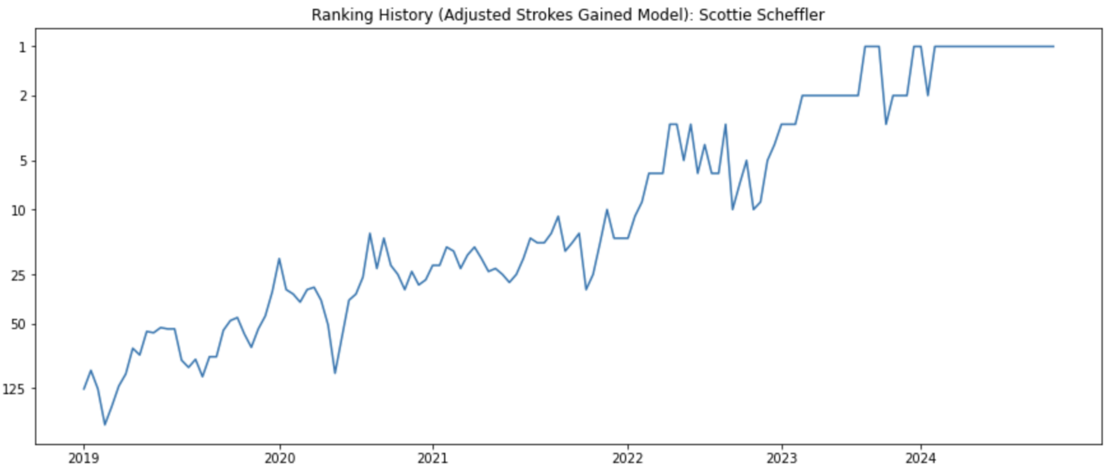

Figure 4.2: Example Ranking History for a given player

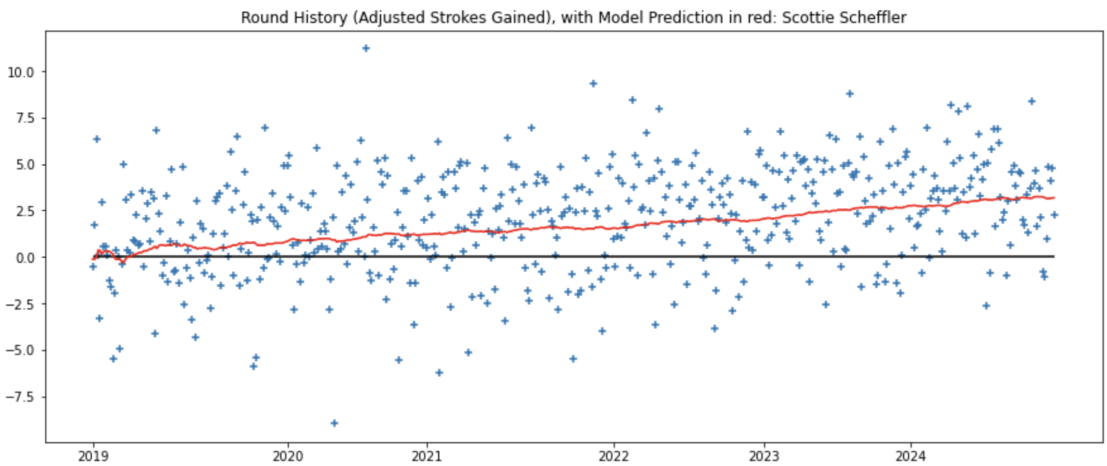

Figure 4.3: Example of Scoring and Model History for a given player

Our paper presents only these visualizations, but with the dataset and code base we developed, it is possible to produce these same visualizations for any player who meets the ranking criterion. We can also display the ranking for any of the ranking days.

5 Related Work

Much has already been discussed in the background section about datagolf.com, the OWGR, and the Connolly and Rendleman paper. Thus, we will dedicate this section to a brief discussion of an interesting result in the mathematical modeling of golf scores, due to the mathematician G.H. Hardy (known widely, partly due to the Hardy-Littlewood Theorem, among other results of pure mathematics). Hardy had an interest in golf, and in 1945 he published one of the first papers in the literature about modeling golf scores, titled *A Mathematical Theorem about Golf* [5].

In this paper, Hardy presents the following model for a golfer's score on a hole:

- Denote the par of the hole by P. P is either 3, 4, or 5.
- A golfer may make a normal shot, N, which advances him 1 towards the par of the hole
- A golfer may make an exception shot, E, which advances him 2 towards the par of the hole
- A golfer may make a bad shot, B, which does not advance him towards the par of the hole

Note that this is a simplification of Broadie's strokes gained [2], which allows for the number of strokes advanced towards the hole's par to be any value, between all of the strokes (for a hole in one), to even negative strokes (for a ball entering a recovery area).

With Hardy's model, on a par 4, one way to complete the hole is with four normal shots, denoted by the sequence NNNN. Another way to complete the hole in 4 shots would be BENN, as the first bad shot contributes, none, the exceptional shot contributes 2, and the normal shots each 1. The important case with Hardy's model is one like NNNE, in which the golfer finishes with an exceptional shot, but only profits 1 shot from this, as they finished the hole, instead of the usual 2. Hardy calculated that if a golfer makes a bad shot with the same probability x that he makes an exceptional shot, he will lose more often than he will win against a player who only makes normal shots for most values of x (this only changes at large values of x, close to 0.5, which are unrealistic). This is an interesting result, as it shows that the consistent golfer will outperform the streaky golfer.

This can also be attributed to golf rankings, as often the best players are the ones who consistently perform well, instead of those who have some remarkable results, but an equal number of poor results. Scottie Scheffler is the perfect example of this, as he has not missed a cut in any tournament in the 2024 season, and has the most top 10 finishes and wins of any player on the PGA Tour during that time, even though he is known for his consistent play, rather than any one exceptional ability.

6 Conclusions

This paper represents the development of a viable world golf ranking. This ranking largely draws on the work of datagolf.com and the OWGR database, and is fully functional, from data collection, to processing, to modeling, and finally to output and visualizations. The paper also tested additional modeling techniques beyond those which had been widely published in the literature, although it found that these techniques were not significant improvements over the current methods, given the set of features present in the this dataset. It remains to be seen if with further work on feature development, more advanced models, such as GBTs or neural nets, could yield increased predictive power over regression.

7 Future Work

The work done up to this point has been largely to get a functioning golf world ranking. This work has been successful, with the following points as continuing points of research:

- There is room for additional research into the modeling behind the ranking. This ranking ended up being limited by the number of features extracted from the dataset. The acquisition of additional data, such as text data from internet sources to gauge sentiment on player performance, could be an interesting avenue, although likely many golfers are discussed extensively, with the majority not discussed at all.

- We reserved the use of 20 percent of the data, which has not been trained or validated on up to this point. The reason for this was that during model selection, we wanted to test out of sample fit on one dataset with many models, and then test again on a final section of the data a very limited number of times. This option is still open to us, as with further work on the modeling it would be important to have this segment of data untouched.

- It would be interesting to compare the performance of the OWGR system versus this model's predicted rankings. A potential research question, which this paper did not attempt to answer is: Is the OWGR a worse predictor of golfer performance than Model X? Should Model X replace the OWGR as the official method for ranking professional golfers? To do this, additional questions would need to be addressed, such as if this is a more or less equitable way of ranking golfers, and which golfers would stand to benefit from such a switch.

- One area of research that datagolf.com has gone into after developing this sort of model is that of sports betting, and simulating tournament probabilities. It would be interesting to build out this sort of system, as this is a growing area of interest in the sports world.

Bibliography

[1] Official World Golf Ranking. *OWGR*. Accessed 6 Dec. 2024. URL: https://www.owgr.com.

[2] Mark Broadie. "Assessing Golfer Performance on the PGA TOUR". In: *Interfaces* 42.2 (Apr. 2012), pp. 146–165.

[3] DataGolf. *DataGolf*. Accessed 6 Dec. 2024. URL: https://www.datagolf.com.

[4] Robert A. Connolly and Jr. Rendleman Richard J. "Skill, Luck, and Streaky Play on the PGA Tour". In: *The Journal of the American Statistical Association* 103 (2008), pp. 74–88.

[5] G. H. Hardy. "A Mathematical Theorem about Golf". In: *The Mathematical Gazette* 29.287 (1945), p. 227. DOI: 10.2307/3609265.