Forecasting the PGA Tour

Brett Mele

MSDS 498 Capstone

Northwestern University

August 26th, 2023

1. Introduction

Analytics has changed the landscape and increased business opportunities across most major sports. Not only have *Moneyball*-like revolutions taken place within every major American team sport, but private companies have begun to reap the benefits of increasing the quality and quantity of data available. Driveline baseball has revolutionized private player development in baseball, offering a suite of data-driven products and services to amateur and professional athletes. TruMedia helps enterprise clients across five professional sports conduct in-depth analysis with their offerings of proprietary metrics and visualization tools. Pro Football Focus provides its clients with both charted in-game data and advanced metrics to better analyze American football.

Relative to other sports, professional golf has lagged behind the rest of the industry. Golf analytics website Data Golf sells products primarily geared towards gaming, but otherwise enterprise analytical tools are hard to come by given the lack of available data and market opportunities for similar products. While the former may still be an obstacle, the latter should no longer be of concern. The PGA tour generated $1.59 billion in revenue in 2021, a 31% year-to-year increase (Byers 2022), and has increased its prize money pool to a record $415 million in the 2022-2023 season (Goldberg 2022). These increases are driven by growing interest in the sport, with viewership up 3-4% on average as of April 2022, including a 19% increase in viewers for the 2022 Masters (Carpenter 2023).

The growing revenue and interest in the game of golf presents a market opportunity for an enterprise golf analytics product. While existing products exist for analysis and monitoring of player performance, forecasting and inference surrounding those forecasts are largely untapped areas of opportunity. Recognizing this, we developed a suite of models and tools geared towards forecasting skill, tournament wins and earnings of golfers on the professional (PGA) tour. This includes PGA Tour projections for players currently playing on the US amateur (Korn Ferry) tour. In addition to modeling Tour performance, we present an end-to-end data pipeline complete with a working prototype that could eventually be offered to prospective clients, including individual golfers seeking an objective perspective on their performance, enterprise clients looking to maximize the value of their sponsorships, or media partners looking to supplement on-air content.

2. Literature Review

Courchene and Courchene (2017) developed probabilistic forecasts for individual PGA tour events. They first adjust round scores for individual golfers using a fixed effects regression that controls for each unique golfer and estimates field strength and course difficulty. They then take actual round-by-round data, make model predictions and calculate adjusted scoring average as the model residual. To estimate player skill going into a tournament, they use adjusted scoring averages for three different time horizons along with the number of rounds a golfer has played in an OLS regression. To simulate tournament outcomes, predictions from the latter model are used as skill estimates, while a standard deviation of round-by-round scoring is estimated empirically for each golfer. The modeling approach was later updated to include detailed strokes-gained categories and updated the adjusted score calculation (Courchene & Courchene 2018).

Baumer et. al. (2013) introduced a model for player value attribution in professional baseball. A key component of the methodology was adjusting the values of batting, baserunning, and defensive events by controlling for factors that were agnostic of player skill. Similar to Chourchene and Courchene, the researchers control for these outside factors by creating linear regressions with the desired environmental variables, fitting the regressions, and using the residuals as the adjusted values.

Existing research on player earnings is scarce relative to player skill. However, Scully (2002) explored the distribution of earnings on the PGA tour and the link between skill and earnings. While scoring averages were found to be normally distributed, the prize structure was found to be nonlinear and highly inequitable, with large prizes earned by a select few players.

3. Data

The majority of data for this project were obtained via the Data Golf API. This includes round-by-round scoring, player demographics, current rankings, and event metadata. Round scoring data dates back to the 2017 season, and includes both the PGA and Korn Ferry tours. Event earnings were scraped from pgatour.com and player biographical info, including age, were scraped from espn.com.

It should be noted that we exclude any data for the LIV tour, which was established in 2022 and is an alternative to the PGA tour. Our methodology should easily allow for the inclusion of LIV data, but accounting for two different professional tours would have been too complicated to handle within the timeframe for this project.

All data were ingested and stored in a Postgres data warehouse hosted on Google Cloud Platform (GCP). Data available via REST API was ingested with Go, while data scraped from the web was ingested with Python. We utilized an ELT framework for our data architecture, with raw data stored in a staging (silver) database, and cleaned, transformed and merged data stored in a production (gold) database. The ELT pipeline is deployed via docker containers and services are run via GCP's compute engine.

4. Methodology

At a high level, the methodology for the project is as follows. First, we calculate adjusted strokes gained, where strokes gained is defined as the number of strokes below average in a round, and adjustments are made to account for course difficulty and field strength. Next, we compute an in-sample estimate of latent skill at every point in time available for each golfer, measured in adjusted strokes gained relative to PGA Tour average. To predict skill in future years, we construct an aging curve using all golfer performances within our sample. We then create a model to predict next season's ($year_{i+1}$) average adjusted strokes gained. This model includes the most recent latent skill estimate aged forward one season, as well as a number of moving averages of adjusted strokes gained to account for the degree of improvement (or regression) over time. Upon generating predictions for $year_{i+1}$, skill in subsequent seasons ($year_{i+2,\ldots,n}$) is estimated by simply applying the aforementioned aging curve. In addition to skill estimates, we obtain uncertainty estimates for latent skill and round-by-round scoring for each golfer. Once we have modeled golfer skill (and variance in skill) in future seasons, we can simulate major tournament outcomes and

predict season level earnings. We will discuss this process in more detail throughout the rest of this section.

All modeling was done in R in order to leverage mixed-effects and generalized additive models (GAMs), which were built with the lme4 and mgcv packages respectively. A full list of packages used can be found in the github repository.

4.1 Adjusted Strokes Gained

To calculate adjusted strokes gained, we use a linear-mixed-effects model with random effects for each golfer and fixed effects for event-rounds and the tour that each event was played on (PGA or KFT). The goal of this model is to control for golfer skill while estimating the difficulty of the round that was played, as courses and weather conditions can drastically affect performance. The target variable for the regression is the number of strokes below par for a golfer in a given round (i.e. a score of 67 on par 72 is -5). Separate models are fit for each year in the sample, which is a choice made primarily for computational efficiency. It is possible we could achieve greater accuracy by including the course each event was played at and using one model fit across all years, but we decided it was not worth the increase in computational complexity.

Example code for the model can be found in figure 1 in the appendix. Once the model is fit, we extract only the fixed-effect coefficients and calculate the residual between the actual and predicted score, which we define as adjusted strokes gained.

Although we do not use strokes gained categories (strokes gained putting, on approach, around-the-green and off-the-tee) in forecasting skill due to time constraints, we do calculate adjustments for each category in order to better contextualize historical performance in the prototype application. These are only done for the PGA tour, as detailed strokes gained categories are not available for the KFT tour.

4.2 In-Sample Latent Skill

In order to forecast skill into the future we need to estimate historical skill over time. Existing research has relied on aggregations of adjusted strokes gained across different yearly timeframes (career, last two years, etc.). We opt for a mixed-modeling approach, where our goal is to estimate the “latent skill” of each golfer using golfer random effects. By modeling skill as a random effect, we can leverage the implicit regression-to-the-mean that is applied in mixed-effects models, and account for changing performance over time simultaneously. In order to do so, we define a random intercept term for each golfer in the sample, along with a random slope for time (defined as number of days from the most golfer’s recent round). The random slope term is modeled using a natural cubic spline to account for potential non-linearities (i.e. changes in skill over time). The target variable is once again the adjusted strokes gained for a golfer in a given round.

Another benefit of using a mixed-effects approach is that we can extract the variance of random effects to account for uncertainty in the latent skill estimates. Golfers who play fewer rounds have more uncertainty associated with their skill estimates than others. This is useful when forecasting future skill, as it allows us

to track the uncertainty throughout the modeling process and generate more accurate confidence intervals for projections.

Model code is shown in figure 2, with an example of how the skill estimate changes over time in figure 3. Experiments were conducted to determine the optimal model specification, and the best model was chosen by AIC. Once the best model was identified we fit the model over the entire sample of data in order to obtain estimates for each golfer at every available point in time. These estimates are only used for inference; separate models used for prediction will leverage these values.

4.3 Skill Aging Curve

In addition to the latent skill model, a separate model is fit to define a skill aging curve. Aging curves have a long history in sports analytics. They are typically used to forecast player performance into the future, as athletes in any sport will experience an improvement and eventual decline in skill with age. To do so, we fit a similar model to 4.2, but instead use only a random intercept for golfers and a cubic-spline fixed effect for age. Since ages were not available for all golfers within the sample, missing ages are imputed with the average age for the golfer's primary tour played in each season.

Figure 4 shows the resulting skill aging curve from ages 20-60. This can be interpreted as the average skill (in terms of adjusted strokes gained) relative to tour average, by age. We see skill increases throughout the early 20's, slows down around age 28, decreases at a decreasing rate until the mid-30's, and then increases at an increasing rate from thereon. The mid-30s slope of the curve can be interpreted as a decline in physical traits like club speed, but potential improvements in less-physically demanding skills like putting and around-the-green play.

4.4 Future Latent Skill

To predict skill in $year_{i+1}$, we construct a GAM to predict average adjusted strokes gained in the following season. Features for this model include the most recent latent skill estimate in a given year aged forward one year, and moving averages of adjusted strokes gained over the previous 15, 30, and 60 rounds a golfer has played (adjusted strokes gained are regressed slightly toward a golfer's estimated skill level prior to calculating the moving average). Including moving averages helps account for the rate of increase in golfer skill level in the past year (the average golfer plays about 75 rounds in a season). The regression is weighted by the number of rounds each golfer played in $year_i$.

Models were trained using all seasons prior to 2022, with 2022 held out as testing data. A baseline linear regression yielded an R^2 value of .77 with RMSE of 0.946. GAM models with different variations of smoothing terms were then built and tested alongside the baseline. The best model yielded an R2 of .78 and RMSE of 0.924. This model specification is shown in figure 5. Once tested we fit over the entire sample of training data and use it to generate skill predictions in the following season.

Since our dataset only contains data from 2017-2023, and 2023 is not yet complete, we do not have a large enough sample of data to build models for more than one season into the future. So to project seasons beyond one year we simply apply the aging curve to the prediction for $year_{i+1}$. This is at least a

reasonable approximation given the correlations between skill and strokes gained in the following season (0.81), and strokes gained in the current season and strokes gained in the following season (0.58) are sufficiently high.

In addition to skill estimates, we also track uncertainty into the future. For $year_{i+1}$, we assume variance in latent skill is the same as the variance in $year_i$, which comes directly from model 4.2. For all future years, we assume uncertainty increases at a constant rate.

4.5 Tournament Simulation & Earnings

We take two approaches to modeling Tour outcomes in future seasons. For major tournament performance (The Masters, The Open Championship, U.S. Open and PGA Championship), we simulate each of the four tournaments over the next three seasons 10,000 times and calculate the percentage of wins, top 5 finishes and top 10 finishes in each tournament. For earnings, we create a model to predict earnings based on skill within the same season. This approach was implemented to avoid simulations of an entire PGA tour season, which would require significant logic to account for playoff points, event participation and more (FedexCup 2022).

In order to simulate tournaments and accurately estimate earnings, it is important to know which golfers will actually play on the PGA tour in a season. Players will both fall off the PGA tour and get promoted from the KFT tour. Assuming all golfers remain on their primary tour will vastly underestimate the earnings and tournament wins of young, up-and-coming golfers and overestimate those of older players. To account for this, we build models for the probability of promotion to the PGA tour (from KFT) and probability of attrition from the PGA tour. Both models are logistic GAMs with smooth terms for in-season latent skill and age. Model selection was done iteratively by tuning the number of knots for smooth terms and comparing the K-index of smooth terms and overall AIC values. Test accuracy for the final models was 88.6% for promotion probability and 72.4% for attrition probability.

To simulate tournaments in a future season, we first generate 10,000 possible sets of PGA tour eligible players using the promotion and attrition probability models. Then, for each major, we determine the players who will make up the tournament field. Each major has a maximum field size (90 for the Masters, 156 otherwise) and a number of automatic qualifiers based on previous tournament finishes. For example, each Masters field will automatically include: previous major winners the past three seasons, top 12 finishers at the previous year's tournament, top four finishers at any major the previous season, and any winner of a PGA tour event in the previous season. These rules vary slightly for each major but we account for those differences in our coding logic. To get the tournament field for each simulation, we take the automatic qualifiers and assume the rest of the field will be filled by the best remaining players by overall skill level. Skill level for each player within a simulation is drawn from a distribution with mean equal to a player's projected latent skill and standard deviation equal to the residual standard deviation extracted from model 4.2.

It is important to note that there are two levels of uncertainty that must be captured in the simulation. The first is the uncertainty regarding the predicted skill level, which we touched on above. The second is the uncertainty in a golfer's round-by-round scoring average. That is, in any given round a golfer's score will

fall somewhere along a distribution centered around his underlying skill level. For simplicity we calculate this uncertainty empirically, with bayesian regression to the mean, for each golfer based on their historical round-by-round variance in scoring,

Once the field for the tournament is set, we simulate each round of the tournament by taking a draw from a distribution for each participant with mean equal to projected latent skill and standard deviation equal to a golfer's empirically estimated round-by-round standard deviation in strokes gained. After two rounds, we cut players whose ranking in total strokes gained is below the cut line (top 50 plus ties for the Masters, top 70 plus ties otherwise), and simulate the next two rounds. This process is repeated for each of the 10,000 sets of players, or season variants. We do this process for each projected season up to 3 years in the future.

Earnings for each PGA tour season are modeled as a function of player skill. However, the relationship is non-linear. The total pool for a tournament is typically only distributed among players who make the cut, with the majority of money going to the top finishers. The distribution of real (inflation adjusted) earnings for all PGA or KFT players since 2017 can be visualized in figure 6. Given the over-inflation of zero earners we choose to use the Tweedie distribution to model earnings, which is a special case of an exponential distribution commonly used in the insurance industry that can account for a point mass at zero.

The earnings model uses inflation adjusted earnings as the target variable. Inflation is calculated as the growth rate from year to year in the total season pool. We use 2018 as the base year, but the choice is arbitrary. Features of the model include a non-linear spline for average latent skill within a season, standard deviation of the latent skill estimate, and golfer age. To generate predictions, we fit the model and then normalize the response to ensure that total predicted earnings do not exceed the total pool for any season. The test R^2 value for the final earnings model was 0.76 with a RMSE of $556,304. Model specification is shown in figure 7.

For future seasons, and within the application, we present nominal earnings. In order to do so, for historical seasons we re-apply the required inflation rate to predicted earnings. For future seasons, we calculate the average growth rate in pool size from season to season (~10.1%), and assume this to be the expected inflation rate going forward. We then apply the expected inflation rate to the product of predicted earnings and the probability of playing on the PGA tour in a given season (generated from the promotion and attrition models). Finally, we normalize predicted earnings in each season by assuming the total pool in a future season will be equal to the total pool in the most recent actual season times the expected inflation rate.

5. Results

Modeling was conducted using the 2023 season as the first projection year, as the 2023 PGA Tour season is still ongoing at the time of this writing. It would be a straightforward extension to produce rest-of-season projections for the current season, but that was beyond the scope of this project.

In the prototype application, the first page shows the current Official World Golf Rankings (OWGR) and DataGolf rankings compared to our predicted ranking for the 2023 season.

Current Rankings

updated 2023-07-14

Show 25 entries Search:

PLAYER	AGE	PRIMARY TOUR	COUNTRY	DG RANK	OWGR RANK	PROJ RANK
Scheffler, Scottie	26.8	PGA	USA	1	1	5
McIlroy, Rory	33.9	PGA	NIR	2	3	2
Rahm, Jon	28.4	PGA	ESP	3	2	1
Cantlay, Patrick	31.0	PGA	USA	4	4	3
Schauffele, Xander	29.4	PGA	USA	5	6	6
Fowler, Rickie	34.3	PGA	USA	6	21	96
Hovland, Viktor	25.5	PGA	NOR	7	5	4
Hatton, Tyrrell	31.5	PGA	ENG	8	16	20
Smith, Cameron	29.6	LIV	AUS	9	7	10
Morikawa, Collin	26.1	PGA	USA	10	19	8
Clark, Wyndham	29.3	PGA	USA	11	11	152
Koepka, Brooks	32.9	LIV	USA	12	12	72
McCarthy, Denny	30.1	PGA	USA	13	32	64
Fleetwood, Tommy	32.2	PGA	ENG	14	22	36
Henley, Russell	34.0	PGA	USA	15	35	22
Johnson, Dustin	38.8	LIV	USA	16	77	17
Fitzpatrick, Matthew	28.6	PGA	ENG	17	9	13
Finau, Tony	33.5	PGA	USA	18	14	15
Spieth, Jordan	29.7	PGA	USA	19	10	31

We see some large discrepancies in predicted and actual rankings with Rickie Fowler and Wyndham Clark. Both golfers have experienced somewhat unexpected rises this year, climbing from OWGR ranks of over 100 at the start of 2023 to the top 25 at the three-quarter mark of the season. Otherwise the predictions have seemed to align with actual results, with most of the projected top 10 golfers actually appearing in the OWGR top 10.

The application also allows us to examine golfer predictions over the next three seasons.

Three Year Projections

Show 25 entries Search:

PLAYER	CURRENT AGE	COUNTRY	THCP 2023	THCP 2024	THCP 2025	MAJOR WINS	MAJOR T5	MAJOR T10	EARNINGS
Rahm, Jon	27.4	ESP	-3.48	-3.44	-3.37	0.59	2.16	3.54	53.34
McIlroy, Rory	32.9	NIR	-3.46	-3.4	-3.37	0.61	2.19	3.52	72.47
Cantlay, Patrick	30	USA	-3.22	-3.1	-3.07	0.41	1.66	2.85	49.35
Hovland, Viktor	24.5	NOR	-3.18	-3.3	-3.32	0.46	1.71	2.83	42.59
Scheffler, Scottie	25.8	USA	-3.14	-3.2	-3.18	0.34	1.46	2.61	41.15
Schauffele, Xander	28.4	USA	-3.08	-2.97	-2.91	0.27	1.27	2.35	32.32
Thomas, Justin	28.9	USA	-2.94	-2.82	-2.77	0.3	1.29	2.27	27.05
Morikawa, Collin	25.1	USA	-2.94	-3.03	-3.03	0.32	1.33	2.31	30.5
Zalatoris, Will	25.6	USA	-2.89	-2.95	-2.94	0.25	1.13	2.09	32.97
Smith, Cameron	28.6	AUS	-2.87	-2.76	-2.7	0.39	1.38	2.31	27.51
Im, Sungjae	24	KOR	-2.84	-2.99	-3.02	0.18	0.92	1.8	29.41
Kim, Tom	19.8	KOR	-2.81	-3.11	-3.24	0.18	0.94	1.81	26.58
Fitzpatrick, Matthew	27.6	ENG	-2.71	-2.65	-2.58	0.13	0.73	1.52	19.29
Niemann, Joaquin	23.4	CHI	-2.67	-2.84	-2.9	0.18	0.86	1.66	25.7
Finau, Tony	32.5	USA	-2.66	-2.6	-2.57	0.17	0.86	1.64	25.9
Burns, Sam	25.7	USA	-2.65	-2.71	-2.7	0.25	1.04	1.84	20.66
Johnson, Dustin	37.8	USA	-2.48	-2.34	-2.25	0.17	0.77	1.45	20.13
Casey, Paul	44.7	ENG	-2.47	-2.21	-2.07	0.18	0.81	1.49	16.85
Wise, Aaron	25.8	USA	-2.45	-2.5	-2.48	0.1	0.55	1.15	17.54

Results are arranged by projected tour handicap (THCP), or the predicted skill estimate, for the next projection season. We include the number of major wins, top 5 finishes, top 10 finishes and total expected earnings over the next three years. Notably, Sunjae Im and Tom Kim seem to be potential up-and-comers, while Rory McIlroy and John Rahm are expected to continue their dominance.

Next, we can explore historical performances in terms of adjusted strokes gained.

Historical Adjusted Performance

Show 25 entries Search:

PLAYER	COUNTRY	SEASON	ROUNDS	SG	PUTT	ARG	APP	OTT	SG N	PUTT N	ARG N	APP N	OTT N
Koepka, Brooks	USA	2023	12	-4.45	-0.92	-0.61	-1.43	-0.98					
Scheffler, Scottie	USA	2023	72	-3.86	0.14	-0.53	-1.48	-1.26	100	35	98	100	100
Johnson, Dustin	USA	2018	78	-3.83	-0.4	-0.21	-1.11	-1.17	100	86	75	98	100
McIlroy, Rory	NIR	2022	64	-3.73	-0.56	-0.42	-0.94	-1.22	100	91	92	96	100
Rose, Justin	ENG	2018	70	-3.71	-0.53	-0.46	-0.71	-0.86	100	91	95	90	96
Smith, Jordan	ENG	2017	4	-3.64	-0.76	0.53	-0.92	-1.79					
Rahm, Jon	ESP	2023	59	-3.62	-0.53	-0.25	-1.26	-0.81	100	91	82	98	96
Macintyre, Robert	SCO	2019	4	-3.61									
McIlroy, Rory	NIR	2019	68	-3.61	-0.49	-0.37	-0.9	-1.46	100	90	90	96	100
Rahm, Jon	ESP	2021	77	-3.59	-0.4	-0.32	-0.94	-1.19	100	81	85	97	99
Waring, Paul	ENG	2020	4	-3.54									
Spieth, Jordan	USA	2017	78	-3.53	-0.38	-0.44	-1.19	-0.51	100	84	95	100	79
Wilson, Andrew	ENG	2022	1	-3.53	0.09	-2.27	-0.26	-1.23					
Thomas, Justin	USA	2018	86	-3.38	-0.33	-0.37	-1.17	-0.69	100	80	90	99	90
Smith, Cameron	AUS	2022	62	-3.37	-0.89	-0.37	-1.05	-0.23	100	97	90	97	54

Values in this table are the average adjusted strokes gained for each category and corresponding percentiles for players with at least 15 rounds played. Scottie Scheffler has performed at an elite level off-the-tee, on approach and around the green, but has been a below average putter in 2023. Putting is less predictive of future performance than other categories, so it is probably safe to assume Scheffler continues to compete at a high level going forward (Courchene and Courchene 2018).

From a historical perspective, we can compare a golfer's actual earnings to expected earnings based on their skill level and identify those who benefited from luck, or lack thereof, in previous seasons.

Historical Earnings

Show 25 entries Search:

PLAYER	COUNTRY	SEASON	EARNINGS	X EARNINGS	+/-
Na, Kevin	USA	2021	18.58	3.39	15.19
Schauffele, Xander	USA	2020	18.78	7.85	10.93
McIlroy, Rory	NIR	2019	22.57	13.82	8.75
Koepka, Brooks	USA	2019	12.89	4.22	8.67
Straka, Sepp	AUT	2022	8.42	0.95	7.47
Homa, Max	USA	2022	10.86	5.9	4.96
Im, Sungjae	KOR	2022	11.15	6.58	4.57
DeChambeau, Bryson	USA	2018	6.31	1.97	4.34
Schauffele, Xander	USA	2019	9.76	5.95	3.81
Johnson, Dustin	USA	2020	11.01	7.3	3.71
Scheffler, Scottie	USA	2020	7.26	3.57	3.69
Casey, Paul	ENG	2019	7.53	3.99	3.54
DeChambeau, Bryson	USA	2021	7.81	4.28	3.53

Historical Earnings

Show 25 entries Search:

PLAYER	COUNTRY	SEASON	EARNINGS	X EARNINGS	+/-
Rahm, Jon	ESP	2022	5.66	15.05	-9.39
Cantlay, Patrick	USA	2022	9.24	17.71	-8.47
McIlroy, Rory	NIR	2021	4.89	12.29	-7.4
Casey, Paul	ENG	2022	1.66	8.85	-7.19
Finau, Tony	USA	2022	7.27	13.72	-6.45
Johnson, Dustin	USA	2022	1.45	7.79	-6.34
Johnson, Dustin	USA	2021	5.95	11.67	-5.72
Cantlay, Patrick	USA	2020	2.09	7.49	-5.4
McIlroy, Rory	NIR	2020	5.6	10.76	-5.16
Garcia, Sergio	ESP	2022	0.77	5.92	-5.15
Schauffele, Xander	USA	2022	7.84	12.86	-5.02

In 2021, Kevin Na earned about $15 million more than expected thanks to an unexpected 3rd place finish in the FedexCup playoffs, one win, and two runner up finishes. On the other hand, John Rahm earned significantly less than expected in 2022, but has already surpassed $15 million in earnings in 2023 at the time of this writing.

Finally, the application contains player pages that offer a comprehensive view of past and projected future performance. At the top of the page we summarize projected ranking, skill, major performance and earnings over each of the next three seasons. Below that, we plot expected skill level for each of the next 5 seasons, along with a 95% confidence interval to quantify uncertainty. The page also includes a historical strokes gained profile, year-by-year breakdown of performance and summary of the player's last 10 events.

An example of this page is shown below. Additional examples can be found in the appendix, or within the live application.

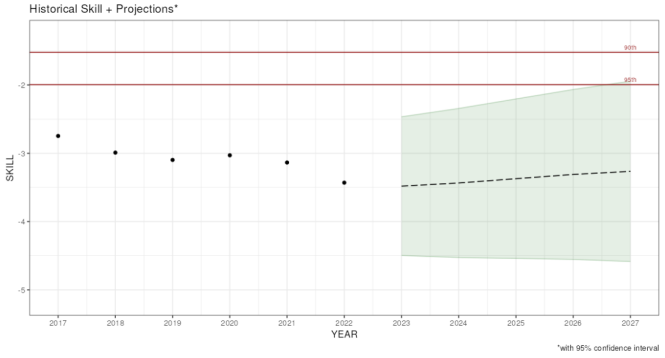
Jon Rahm

Birthdate: 1994-11-10 (28.4)

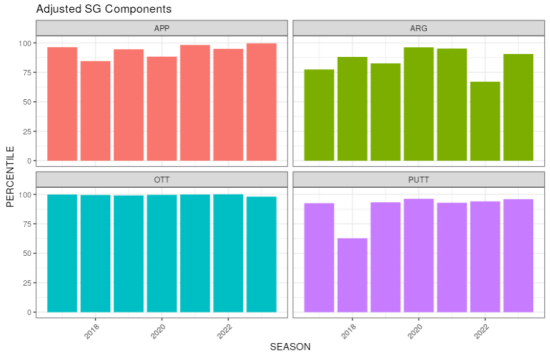
Country: ESP

	PROJECTION YEAR	PROJ RANK	SKILL	MAJOR WINS	MAJOR TSS	MAJOR T10S	EARNINGS
	2023	1	-3.482	0.193	0.717	1.172	15.794
	2024	1	-3.435	0.197	0.718	1.181	18.126
	2025	1	-3.372	0.196	0.726	1.192	19.418
TOTAL	–	–	–	0.586	2.161	3.545	53.338

Historical Skill + Projections*



Adjusted SG Components



Performance + Projections

SEASON	ROUNDS	SG	PUTT	ARG	APP	OTT	MAJOR WINS	MAJOR TSS	MAJOR T10S	EARNINGS	PRED EARNINGS
2025	**–**	**-3.37**	**–**	**–**	**–**	**–**	**0.20**	**0.73**	**1.19**	**–**	**19.42**
2024	**–**	**-3.44**	**–**	**–**	**–**	**–**	**0.20**	**0.72**	**1.18**	**–**	**18.13**
2023	**–**	**-3.48**	**–**	**–**	**–**	**–**	**0.19**	**0.72**	**1.17**	**–**	**15.79**
2022	70	-3.10	-0.37	0.01	-0.59	-1.26	0.00	0.00	0.00	5.66	15.05
2021	77	-3.59	-0.40	-0.32	-0.94	-1.19	1.00	3.00	4.00	12.11	15.13
2020	63	-3.22	-0.60	-0.35	-0.44	-0.96	0.00	0.00	1.00	6.98	9.89
2019	72	-3.18	-0.39	-0.14	-0.66	-0.93	0.00	1.00	2.00	4.28	7.99
2018	66	-2.79	0.02	-0.22	-0.40	-1.03	0.00	2.00	2.00	3.51	4.90
2017	71	-3.02	-0.38	-0.08	-0.72	-1.13	0.00	0.00	0.00	4.04	3.28

Last 10 Events

YEAR	EVENT NAME	START DATE	FIN	ADJ SG	ADJ PUTT	ADJ ARG	ADJ APP	ADJ OTT	SG RANK	PUTT RANK	ARG RANK	APP RANK	OTT RANK
2023	Travelers Championship	2023-06-25	CUT	-1.4	0.2	0.2	0.0	-0.8	78	86	83	86	75
2023	U.S. Open	2023-06-18	T10	-14.7	-2.9	-1.6	-5.6	-2.5	10	28	32	16	27
2023	the Memorial Tournament presented by Workday	2023-06-04	T16	-7.7	4.8	-0.3	-10.7	-2.8	18	112	61	2	24
2023	PGA Championship	2023-05-21	T50	-7.1	-1.0	1.1	-2.6	-2.8	50	62	114	44	27
2023	Mexico Open at Vidanta	2023-04-30	2	-18.6	-8.1	-0.3	-2.3	-6.3	2	2	63	34	1
2023	RBC Heritage	2023-04-16	T15	-12.2	0.4	3.0	-9.7	-3.8	15	84	133	1	12
2023	Masters Tournament	2023-04-09	1	-21.6	-2.4	-3.4	-5.9	-7.6	1	23	7	6	2
2023	THE PLAYERS Championship	2023-03-12	WD	-3.8	2.1	-1.3	-2.3	-0.7	72	112	36	47	70
2023	Arnold Palmer Invitational presented by Mastercard	2023-03-05	T39	-6.6	-4.0	-2.7	-2.8	5.2	39	17	16	38	118
2023	The Genesis Invitational	2023-02-19	1	-21.7	-4.9	0.2	-13.0	-1.6	1	12	65	1	35

6. Discussion

At the outset of this project we laid out a number of objectives, including (1) develop an enterprise architecture to collect and store golf data, (2) project golfer skill in terms of strokes gained over a four year time horizon for all active golfers, (3) project golfer performance on the PGA tour and in major

championships over a four year time horizon, (4) summarize current and project future earnings on the PGA tour and (5) surface model outputs and model derivatives in a prototype dashboard. Each of these objectives were accomplished with some slight adjustments to the scope of the project along the way. Notably, we decided not to project detailed strokes gained categories or incorporate those metrics into the overall projections. We also modeled earnings and simulated only major tournaments, as opposed to simulating the entire PGA tour season. Both decisions were made due to time constraints and, to a lesser extent, data availability.

The results of this project demonstrate the ability to develop reasonable projections of golfer skill in the future. This has a number of implications from a product perspective. The application can be sold as a subscription based service to individual golfers, fans or prospective sponsors. Partnerships can be sought with gaming companies who may wish to use our data in their products. Similarly, we can explore relationships with media companies who may wish to use our data products as a supplement to their broadcasts or commentary, similar to those seen in the National Football League and Major League Baseball with providers like Amazon Web Services.

Areas for future research are clear and potentially abundant. First, incorporating strokes gained categories into projections should have a meaningful impact on forecast accuracy. Coullehene and Coullehene (2018) have demonstrated an approach for doing so that should be easily extended to our approach. Expanding the dataset to include more professional tours than just the PGA and KFT is also a reasonable next step, which would cover a significantly larger population of professional golfers including those who recently left the PGA for the LIV tour. Building out a full season simulation in addition to major tournaments is another extension of this work that may produce better predictions for season-level outcomes like earnings.

7. Conclusion

This project presents an end-to-end data pipeline for PGA Tour forecasting. Data is ingested via API and web scrapers and stored in a cloud data warehouse hosted on GCP. A ELT architecture is used with containerized microservices deployed to ingest, load, clean and transform data to use in production. A novel methodology was introduced to predict the skill of amateur and professional golfers in future seasons. Using those skill predictions, season performance, including major tournament outcomes and earnings, was projected using simulation and regression approaches. Data and modeling results are surfaced in a prototype application deployed via Docker container on GCP.

Results of this project present an opportunity to introduce a unique data product into the golf analytics landscape. Existing work primarily centers around upcoming tournament prediction, with an eye towards gaming. This project deviates from existing approaches by forecasting future seasons. The prototype application can be developed and offered as a subscription service to golfers or enterprise customers who may be interested in contextualized performance metrics and expected future performance, while the data can be offered to media and gaming partners for use in their products and services.

REFERENCES

Baumer, Benjamin S., Shane T. Jensen, and Gregory J. Matthews. "openWAR: An open source system for evaluating overall player performance in major league baseball." Journal of Quantitative Analysis in Sports 11, no. 2 (2015): 69-84. https://arxiv.org/abs/1312.7158

Byers, Justin, and Justin Byers. “PGA Tour Reports 37% Increase in Revenue During 2021.” Front Office Sports, December 20, 2022. https://frontofficesports.com/pga-tour-reports-37-increase-in-revenue-during-2021/#:~:text=The%20North%20American%20pro%20golf,the%20controversial%20LIV%20Golf%20League.

Carpenter, Josh. “PGA Tour Seeing Viewership Increases Midway through 2023 Season.” Sports Business Journal, April 12, 2023. https://www.sportsbusinessjournal.com/Daily/Issues/2023/04/12/Media/pga-tour-viewership-cbs-nbc-golf-channel-espn.aspx.

Courchene, Matt, and Will Courchene. “A Predictive Model of Tournament Outcomes on the PGA Tour.” data golf blogs, February 14, 2017. https://datagolfblogs.ca/a-predictive-model-of-tournament-outcomes-on-the-pga-tour/.

Courchene, Matt, and Will Courchene. “Data Golf Predictive Model: Methodology.” Data Golf, November 22, 2018. https://datagolf.com/predictive-model-methodology/.

“Data Golf,” n.d. https://datagolf.com/.

Driveline Baseball. “About Us | Driveline Baseball,” August 3, 2022. https://www.drivelinebaseball.com/about/.

"FedExCup Overview." The Official 2022-23 PGA Tour Media Guide. PGA Tour, 2022. Accessed August 24, 2023. https://www.pgatourmediaguide.com/intro/fedex-cup-overview.

Goldberg, Rob. “PGA Tour Increases Prize Money to Record $415m for 44-Tournament 2022-23 Schedule.” Bleacher Report, August 1, 2022. https://bleacherreport.com/articles/10044029-pga-tour-increases-prize-money-to-record-415m-for-44-tournament-2022-23-schedule.

Scully, Gerald W. "The distribution of performance and earnings in a prize economy." Journal of Sports Economics 3, no. 3 (2002): 235-245. https://journals-sagepub-com.turing.library.northwestern.edu/doi/epdf/10.1177/152700250200300 3001

TruMedia. “TruMedia,” n.d. https://www.trumedianetworks.com/.

PFF. “NFL, Fantasy Football, and NFL Draft | PFF,” n.d. https://www.pff.com/.

APPENDIX

Github: https://github.com/melebrett/MSDS-Capstone

Application: https://pgaapp-jb4yato7fq-ue.a.run.app/

Figure 1: Adjusted strokes gained model

```
  mod <- lmer(
    data = rounds %>% filter(year == yr),
    round_score_ou ~ (1|dg_id) + round_num*event_year + tour
  )
```

Figure 2: Latent skill model

```
mod_ls <- lmer(
  data = train %>%
    group_by(dg_id) %>%
    mutate(
      day = lubridate::time_length(interval(date, max(date)), unit ="days")
    ) %>%
    ungroup(),
  adj_sg_total ~ (1 + ns(day, df = 4) | dg_id),
  control = lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 5000))
)
```

Figure 3: Latent skill example

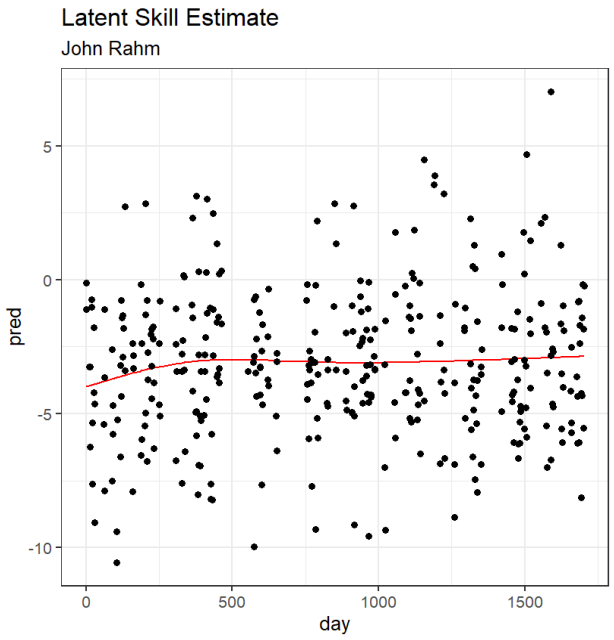


Figure 4: Latent skill aging curve

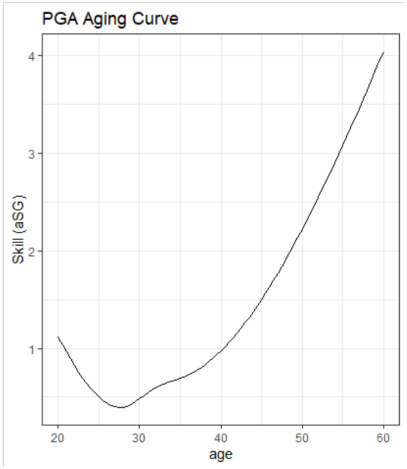


Figure 5: Future latent skill model

```
407 mod_ls_next_final <- mgcv::gam(
408   data = model_df,
409   mean_adj_sg_next ~ s(mean_latent_skill, ma_3) + ma_2,
410   weights = rounds
411 )
```

Figure 6: Tour Earnings

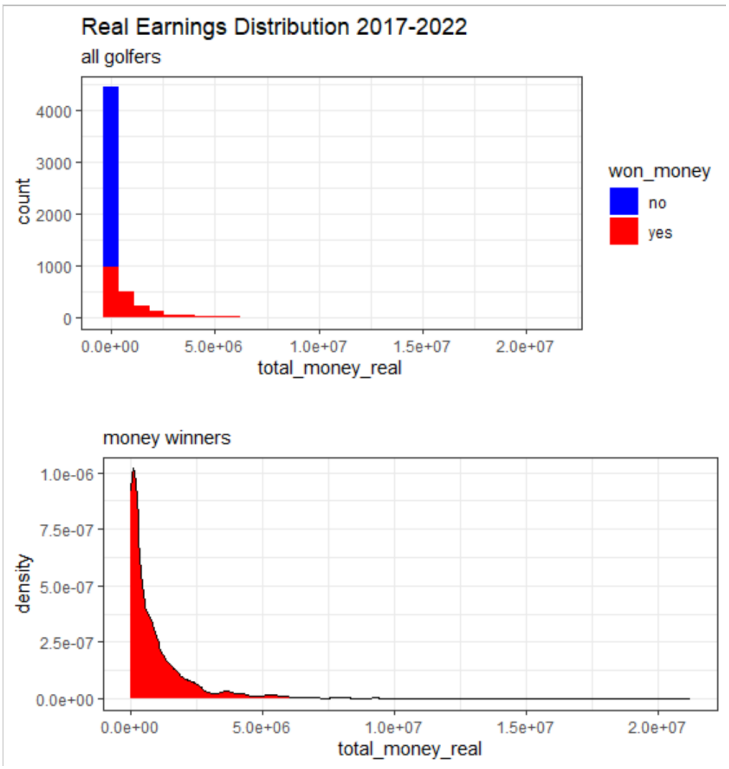


Figure 7: Earnings Model

```
475 mod_earnings <- mgcv::gam(
476   data = train,
477   total_money_real ~ bs(mean_latent_skill) + sd_latent_skill + age,
478   family = tw(),
479   select = T,
480   method = 'REML'
481 )
```

Figure 8: Sungjae Im Player Profile

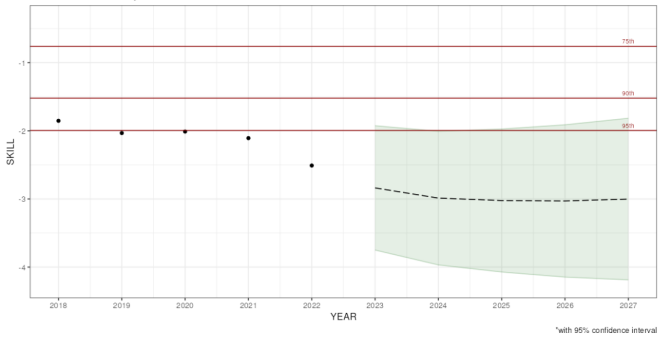
Sungjae Im

Birthdate: 1998-03-30 (25)

Country: KOR

	PROJECTION YEAR	PROJ RANK	SKILL	MAJOR WINS	MAJOR T5S	MAJOR T10S	EARNINGS
	2023	11	-2.838	0.056	0.300	0.594	6.773
	2024	8	-2.987	0.064	0.312	0.609	10.169
	2025	8	-3.023	0.058	0.311	0.599	12.465
TOTAL	–	–	–	0.178	0.923	1.802	29.407

Historical Skill + Projections*



Adjusted SG Components

Performance + Projections

SEASON	ROUNDS	SG	PUTT	ARG	APP	OTT	MAJOR WINS	MAJOR T5S	MAJOR T10S	EARNINGS	PRED EARNINGS
2025	**–**	**-3.02**	**–**	**–**	**–**	**–**	**0.06**	**0.31**	**0.60**	**–**	**12.46**
2024	**–**	**-2.99**	**–**	**–**	**–**	**–**	**0.06**	**0.31**	**0.61**	**–**	**10.17**
2023	**–**	**-2.84**	**–**	**–**	**–**	**–**	**0.06**	**0.30**	**0.59**	**–**	**6.77**
2022	88	-2.95	-0.27	-0.41	-0.53	-0.92	0.00	0.00	1.00	11.15	6.58
2021	126	-2.19	-0.32	0.00	-0.36	-0.74	0.00	0.00	0.00	4.38	3.72
2020	95	-2.09	-0.30	-0.05	-0.23	-0.61	0.00	1.00	1.00	5.39	2.24
2019	118	-2.02	-0.35	-0.31	-0.24	-0.51	0.00	0.00	0.00	2.88	1.78
2018	95	-1.89	-0.01	-0.73	-0.03	-0.60	0.00	0.00	0.00	0.02	1.02

Last 10 Events

YEAR	EVENT NAME	START DATE	FIN	ADJ SG	ADJ PUTT	ADJ ARG	ADJ APP	ADJ OTT	SG RANK	PUTT RANK	ARG RANK	APP RANK	OTT RANK
2023	Rocket Mortgage Classic	2023-07-02	T24	-8.0	-4.8	0.9	-1.5	-0.9	24	11	102	51	60
2023	Travelers Championship	2023-06-25	T29	-8.9	-2.6	-1.6	-1.8	-1.1	29	33	35	43	66
2023	U.S. Open	2023-06-18	CUT	0.8	-0.3	1.7	0.5	-0.4	104	77	128	95	82
2023	the Memorial Tournament presented by Workday	2023-06-04	T41	-3.7	-2.0	0.3	-3.6	0.2	43	33	76	30	78
2023	Charles Schwab Challenge	2023-05-28	CUT	-0.6	-1.0	-1.8	4.7	-1.5	69	44	22	113	38
2023	PGA Championship	2023-05-21	CUT	4.6	3.5	1.6	2.8	-2.4	134	138	130	138	35
2023	Wells Fargo Championship	2023-05-07	T8	-13.7	-5.9	-1.0	-1.7	-3.1	8	7	51	52	22
2023	Zurich Classic of New Orleans	2023-04-23	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2023	RBC Heritage	2023-04-16	T7	-14.2	0.9	-4.1	-3.7	-5.3	7	96	7	29	2
2023	Masters Tournament	2023-04-09	T16	-11.6	-0.6	-2.9	-4.5	-1.4	16	42	12	14	44

Figure 9: Jordan Spieth Player Profile

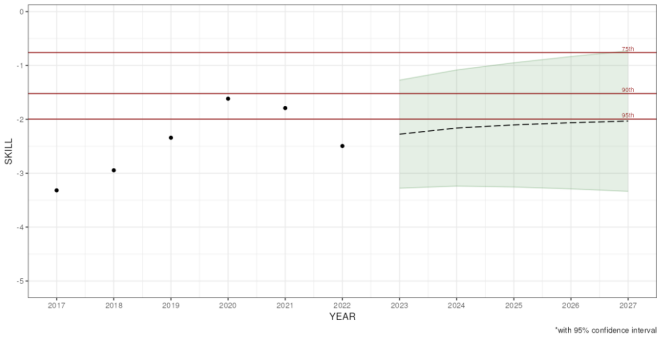
Jordan Spieth

Birthdate: 1993-07-27 (29.7)

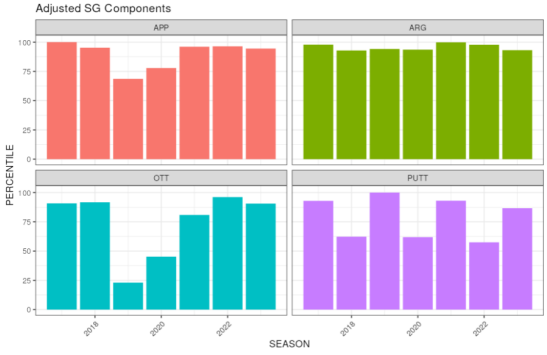
Country: USA

	PROJECTION YEAR	PROJ RANK	SKILL	MAJOR WINS	MAJOR T5S	MAJOR T10S	EARNINGS
	2023	31	-2.275	0.033	0.179	0.365	3.334
	2024	33	-2.161	0.035	0.192	0.380	3.416
	2025	32	-2.103	0.037	0.187	0.382	3.622
TOTAL	–	–	–	0.105	0.558	1.127	10.372

Historical Skill + Projections*



Adjusted SG Components



Performance + Projections

SEASON	ROUNDS	SG	PUTT	ARG	APP	OTT	MAJOR WINS	MAJOR T5S	MAJOR T10S	EARNINGS	PRED EARNINGS
2025	**–**	**-2.10**	**–**	**–**	**–**	**–**	**0.04**	**0.19**	**0.38**	**–**	**3.62**
2024	**–**	**-2.16**	**–**	**–**	**–**	**–**	**0.03**	**0.19**	**0.38**	**–**	**3.42**
2023	**–**	**-2.28**	**–**	**–**	**–**	**–**	**0.03**	**0.18**	**0.37**	**–**	**3.33**
2022	78	-2.22	0.08	-0.38	-0.71	-0.65	0.00	0.00	1.00	5.70	3.92
2021	88	-2.61	-0.38	-0.50	-0.70	-0.22	0.00	2.00	2.00	6.68	3.76
2020	67	-1.12	0.02	-0.31	-0.25	0.06	0.00	0.00	0.00	1.00	1.21
2019	79	-1.86	-0.87	-0.29	-0.04	0.22	0.00	1.00	1.00	1.86	1.68
2018	80	-2.73	0.04	-0.26	-0.68	-0.49	0.00	1.00	2.00	2.24	3.08
2017	78	-3.53	-0.38	-0.44	-1.19	-0.51	1.00	1.00	1.00	6.52	4.87

Last 10 Events

YEAR	EVENT NAME	START DATE	FIN	ADJ SG	ADJ PUTT	ADJ ARG	ADJ APP	ADJ OTT	SG RANK	PUTT RANK	ARG RANK	APP RANK	OTT RANK
2023	U.S. Open	2023-06-18	CUT	-2.2	-0.7	-1.3	1.2	-0.8	59	71	38	109	71
2023	the Memorial Tournament presented by Workday	2023-06-04	T5	-11.7	-0.3	-3.3	-4.1	-5.3	5	62	10	21	6
2023	Charles Schwab Challenge	2023-05-28	CUT	1.4	0.3	3.5	-1.3	0.0	90	63	117	50	65
2023	PGA Championship	2023-05-21	T29	-9.1	0.9	2.2	-3.5	-6.8	29	103	138	30	4
2023	Wells Fargo Championship	2023-05-07	CUT	4.6	0.4	1.2	3.2	0.6	135	87	115	137	105
2023	RBC Heritage	2023-04-16	2	-18.2	-5.4	-1.2	-7.7	-1.9	1	8	42	4	44
2023	Masters Tournament	2023-04-09	T4	-16.6	-3.8	-3.4	-3.8	-3.4	6	10	8	19	20
2023	Valspar Championship	2023-03-19	T3	-14.0	-4.8	-1.6	-4.5	-1.6	3	10	33	10	35
2023	THE PLAYERS Championship	2023-03-12	T19	-14.8	-1.2	-1.7	-5.6	-0.4	19	52	26	11	79
2023	Arnold Palmer Invitational presented by Mastercard	2023-03-05	T4	-14.6	-1.8	-6.2	-1.8	-2.4	4	39	1	50	33

Figure 10: Matt Kuchar Player Profile

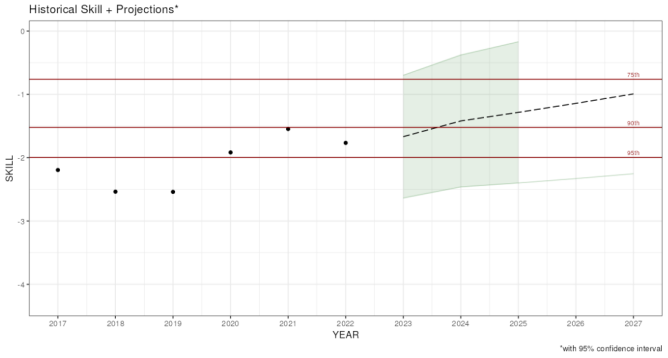
Matt Kuchar

Birthdate: 1978-06-21 (44.8)

Country: USA

	PROJECTION YEAR	PROJ RANK	SKILL	MAJOR WINS	MAJOR T5S	MAJOR T10S	EARNINGS
	2023	76	-1.668	0.002	0.028	0.082	2.800
	2024	105	-1.420	0.003	0.027	0.076	2.286
	2025	121	-1.284	0.003	0.027	0.082	2.115
TOTAL	–	–	–	0.008	0.082	0.24	7.201

Historical Skill + Projections*



Adjusted SG Components

Performance + Projections

SEASON	ROUNDS	SG	PUTT	ARG	APP	OTT	MAJOR WINS	MAJOR T5S	MAJOR T10S	EARNINGS	PRED EARNINGS
2025	**–**	**-1.28**	**–**	**–**	**–**	**–**	**0**	**0.03**	**0.08**	**–**	**2.12**
2024	**–**	**-1.42**	**–**	**–**	**–**	**–**	**0**	**0.03**	**0.08**	**–**	**2.29**
2023	**–**	**-1.67**	**–**	**–**	**–**	**–**	**0**	**0.03**	**0.08**	**–**	**2.80**
2022	71	-2.10	-0.59	-0.51	-0.24	0.04	0	0.00	0.00	2.02	4.81
2021	68	-1.00	-0.15	-0.34	-0.18	0.17	0	0.00	0.00	0.60	3.23
2020	61	-1.92	-0.72	-0.06	-0.17	-0.18	0	0.00	0.00	1.50	2.69
2019	80	-2.56	-0.36	-0.17	-0.82	-0.32	0	0.00	1.00	5.64	5.57
2018	84	-2.25	-0.40	-0.25	-0.44	-0.01	0	0.00	1.00	1.22	5.98
2017	84	-2.73	-0.35	-0.48	-0.43	-0.45	0	2.00	3.00	4.30	4.07

Last 10 Events

YEAR	EVENT NAME	START DATE	FIN	ADJ SG	ADJ PUTT	ADJ ARG	ADJ APP	ADJ OTT	SG RANK	PUTT RANK	ARG RANK	APP RANK	OTT RANK
2023	John Deere Classic	2023-07-09	67	3.8	1.1	-3.4	6.4	1.2	133	105	8	154	126
2023	Travelers Championship	2023-06-25	CUT	5.6	4.8	0.9	0.5	0.4	142	147	114	96	104
2023	U.S. Open	2023-06-18	CUT	0.8	2.9	0.3	-0.1	-1.7	104	131	89	80	47
2023	RBC Canadian Open	2023-06-11	T20	-8.2	-4.1	-0.7	-1.6	-0.2	20	16	55	43	75
2023	the Memorial Tournament presented by Workday	2023-06-04	62	3.3	-4.1	-0.2	6.4	-0.3	85	14	64	117	69
2023	PGA Championship	2023-05-21	CUT	-1.4	1.7	-1.5	-1.5	0.9	80	117	43	59	105
2023	AT&T Byron Nelson	2023-05-14	T43	-4.7	0.5	-2.2	-2.0	0.7	43	86	20	39	98
2023	Wells Fargo Championship	2023-05-07	T23	-9.7	-6.3	-5.2	3.8	0.0	23	5	3	143	96
2023	RBC Heritage	2023-04-16	T19	-11.2	-4.2	-2.5	0.8	-3.3	19	12	21	89	17
2023	Valero Texas Open	2023-04-02	T3	-15.8	-1.6	-4.5	-8.0	0.0	3	40	6	3	80