Forecasting the PGA Tour

Brett Mele

MSDS 498 Summer 2023

Overview

- Golf analytics primarily geared towards gaming (gambling & fantasy)
- Record revenue and interest in sport
  - $1.59B revenue in 2021
  - Record pool sizes
  - Viewership up 4% in 2022, including +19% for the Masters
- Opportunity:
  - Develop product(s) for forecasting and inference, similar to what exists for other pro sports
  - Market includes: pro golfers, prospective sponsors, media & gaming companies
  - Subscription service or data products

Objectives

1. Develop an enterprise architecture to ingest and store professional golf data
2. Project golfer skill over a specified time horizon for all active golfers
3. Project performance on the PGA tour and in major championships over the same time horizon
4. Summarize current and project future earnings on the PGA tour
5. Surface historical data, model outputs and model derivatives in a prototype application

Terminology

PGA Tour - Professional Golf Association Tour, the primary professional golf tour

KFT Tour - Korn Ferry Tour, the developmental tour for the PGA tour

FedEx Cup - PGA Tour playoffs; FedexCup points earned throughout season determine standings

Majors - Most prestigious tournaments in golf (The Masters, U.S. Open, OPEN Championship, PGA Championship)

Strokes Gained (SG) - Number of strokes better than the average golfer (score of 64 with field average of 71 = -7 SG)

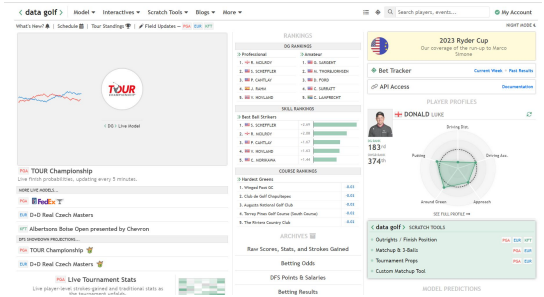
SG Categories - Skills that contribute to overall strokes gained: putting, around-the-green, approach and off-the-tee. The sum of these categories equals total strokes gained in a round.

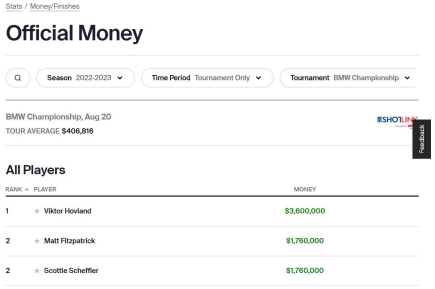
Adjusted Strokes Gained - SG with debit or credit for course/round/field difficulty

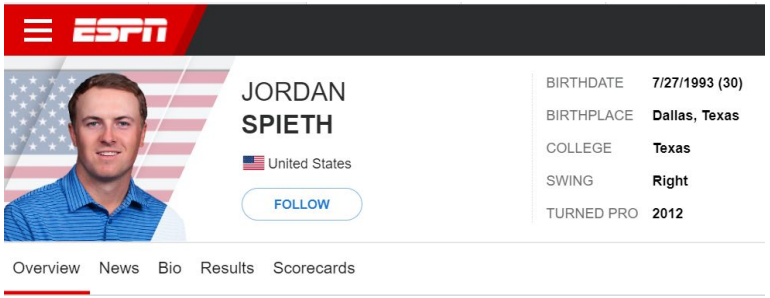
Latent Skill - The underlying/inherent ability of a golfer at a point in time, expressed in SG

Data

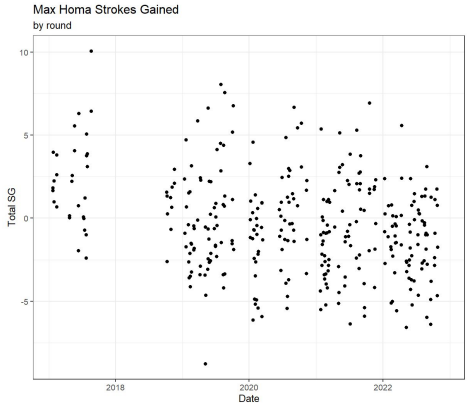
- Data Golf
  - Round-by-round scoring
  - Players
  - Current Rankings
  - Events
  - Unique ID system
- ESPN.com
  - Player Biographical Information
- PGATour.com
  - Event winnings

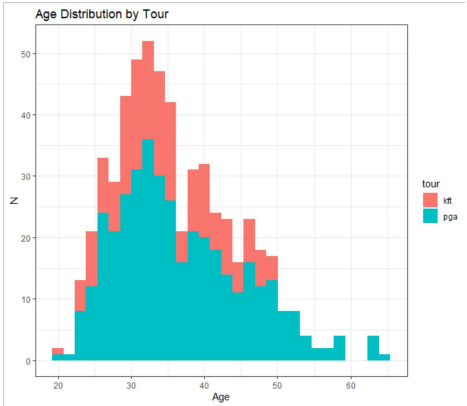


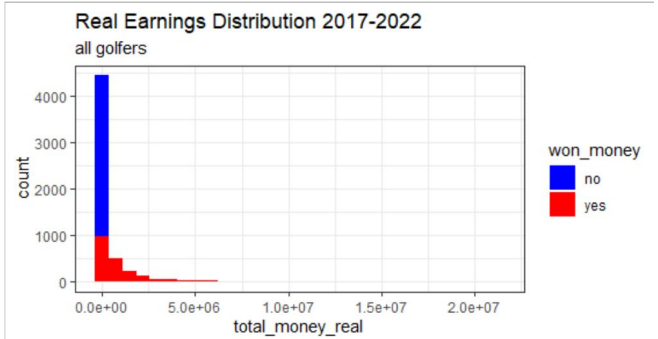


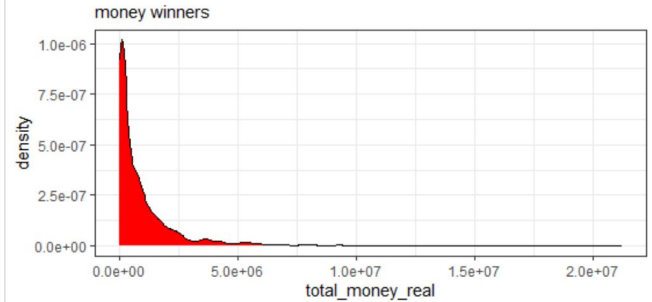


Data





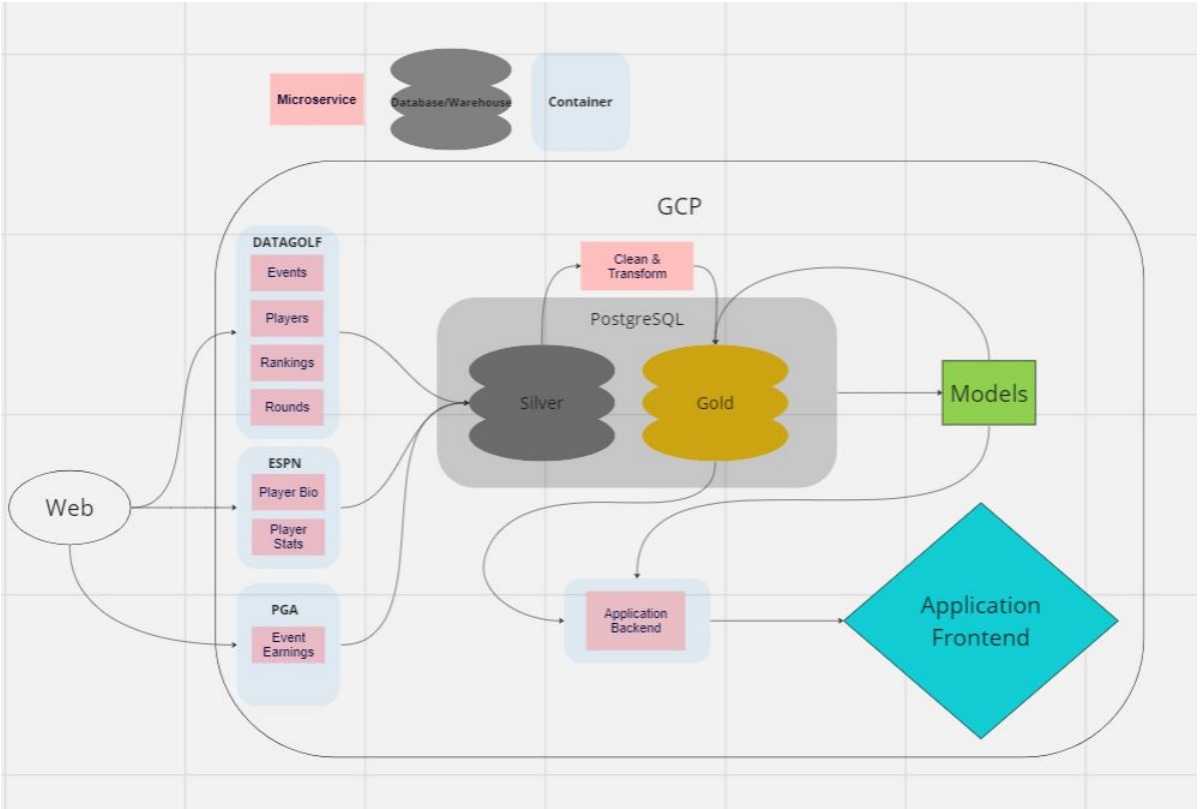




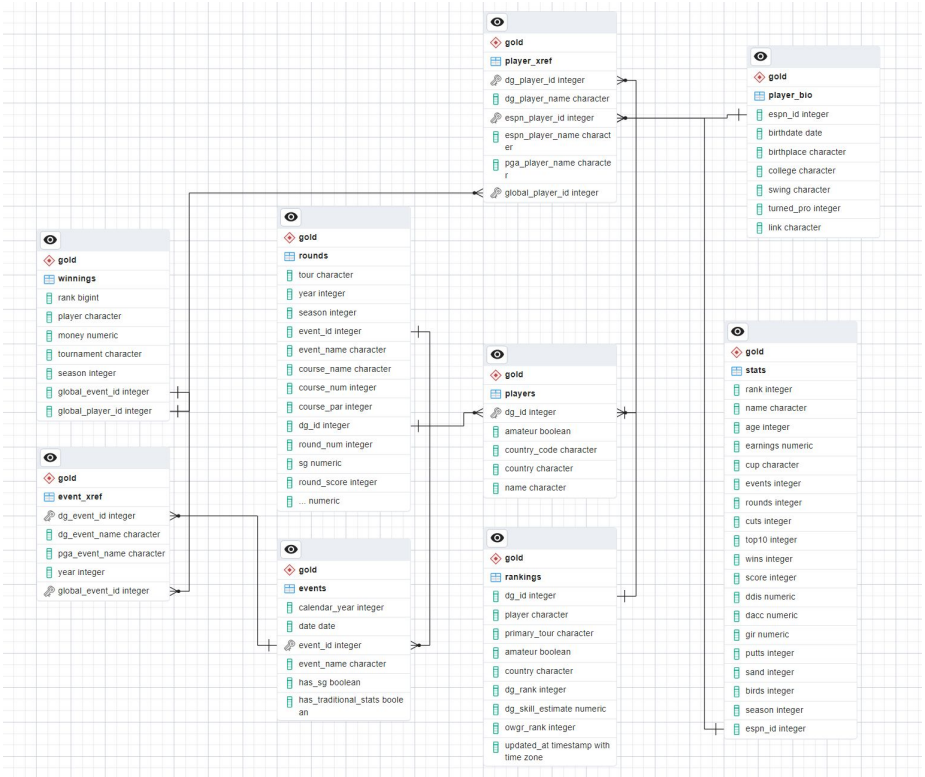
SEASON	TRADITIONAL STATS	STROKES GAINED
KFT		
2017	0.00%	0.00%
2018	0.00%	0.00%
2019	0.00%	0.00%
2020	0.00%	0.00%
2021	0.00%	0.00%
2022	0.00%	0.00%
PGA		
2017	78.97%	78.97%
2018	72.30%	72.30%
2019	69.64%	71.91%
2020	67.34%	68.33%
2021	72.70%	73.95%
2022	73.91%	76.29%
2023	55.65%	55.65%

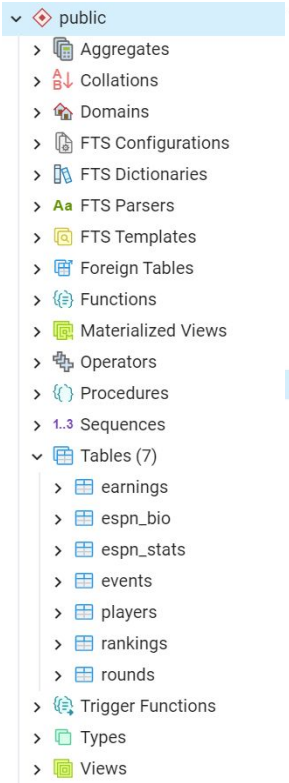
Data Architecture

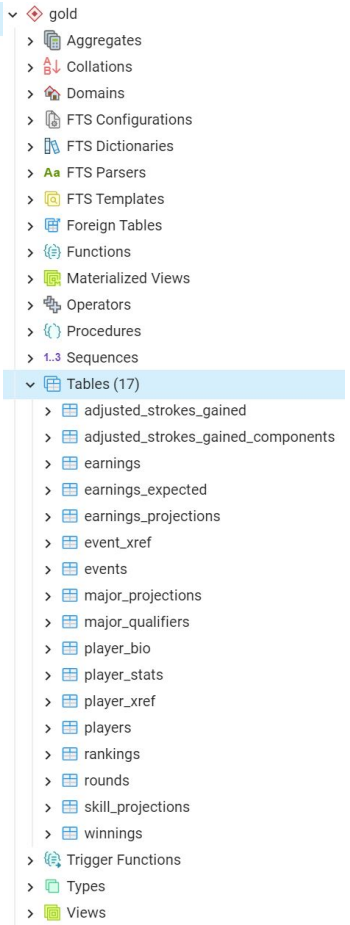
- ELT
  - Go, Python, Docker
- Database
  - PostgreSQL
- Modeling
  - R
- Web App
  - RShiny
- Cloud Computing
  - GCP



Data Architecture

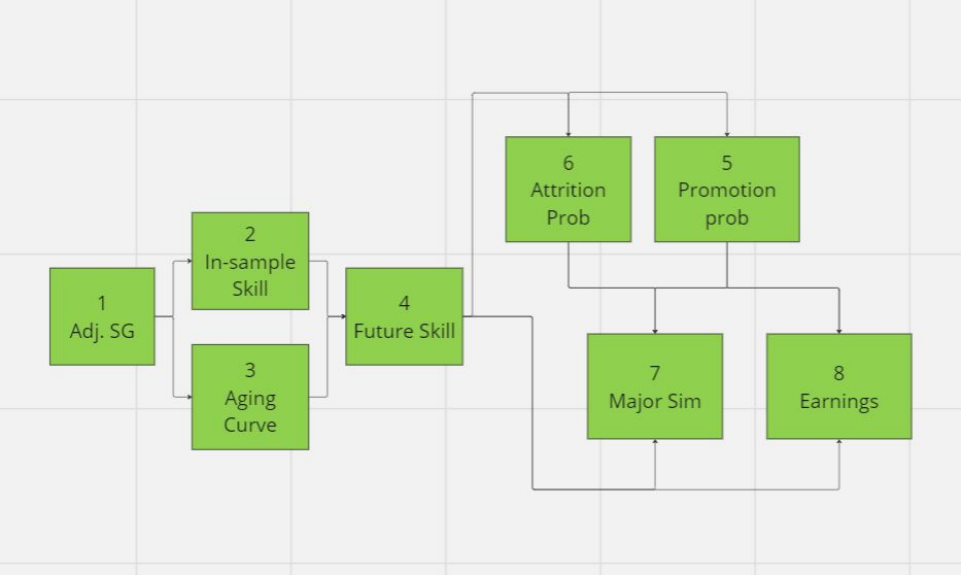






Modeling

1. Adjusted strokes gained
2. In-sample latent skill
3. Skill aging curve
4. Future skill
5. Promotion probability
6. Attrition probability
7. Tournament simulation
8. Season-level earnings



Modeling

Adjusted Strokes Gained

- Linear mixed-effects model
- Golfer random effects
- Event-round and tour fixed effects
- Control for golfer skill and adjust for course/round difficulty
- Also done for SG categories
- Best fit determined by AIC

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

Modeling

In-sample latent skill

- Linear mixed-effects model with splines
- Golfer Random Intercept
- Random slope for day within sample
- Account for changes in skill over time, rather than an average
- Best fit determined by AIC

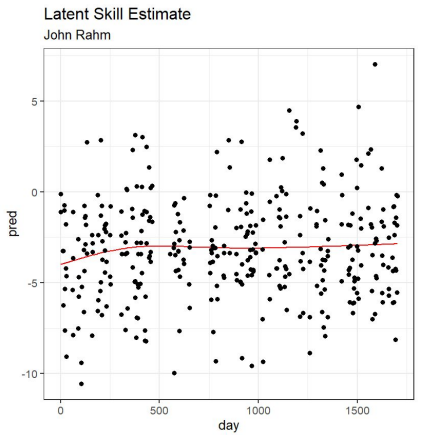
Skill Aging Curve

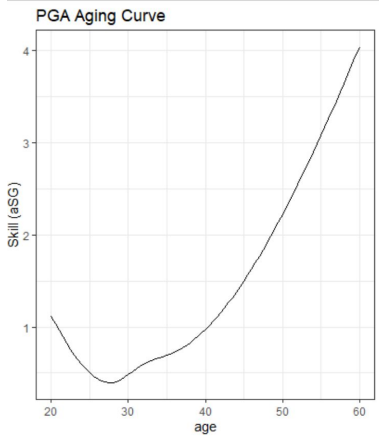
- Generalized additive model (GAM)
- Estimate average change in latent skill over time
- Best fit determined by AIC

Future Skill

- GAM
- Predict next season's adjusted strokes gained
- In-sample skill + moving averages + aging adjustment
- $R^2 = 0.78$; RMSE = 0.924

```
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))
)
```





Modeling

Promotion Probability

- Logistic GAM
- Probability a player is promoted from the KFT to PGA tour
- Best fit determined by AIC and K-Index for smooth terms
- 88.6% Accuracy

```
mod_promotion_prob <- gam(
  data = kft_players %>% filter(year < 2022),
  promotion ~ s(mean_latent_skill, k = 8) + age,
  family = binomial(link = "logit"),
  select = T
)
```

Attrition Probability

- Logistic GAM
- Probability a player drops off the PGA tour
- Best fit determined by AIC and K-Index for smooth terms
- 72.4% Accuracy

```
mod_attrition_prob <- gam(
  data = pga_players %>% filter(year < 2022),
  attrition ~ s(mean_latent_skill, k= 8) + s(age, k = 5),
  family = binomial(link = "logit"),
  weights = rounds,
  select = T
)
```

Modeling

Major Simulation

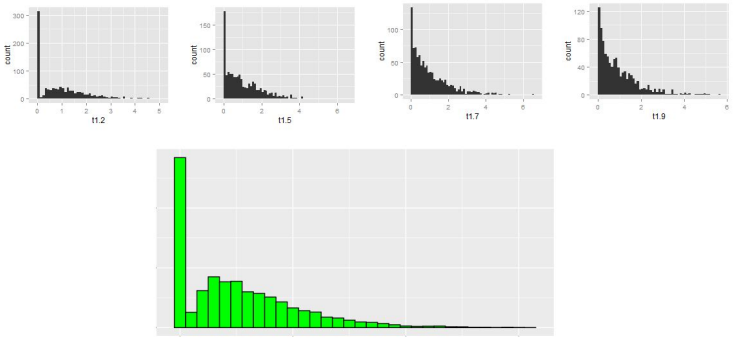
- Monte carlo simulation
- 10,000 simulations per season
  - PGA tour participants drawn from promotion and attrition model outputs
  - 70% promotion => on tour in 7,000 out of 10,00 sims
- Determine automatic qualifiers for each tournament
- Event field = automatic qualifiers plus n next-best golfers
  - n = field size - automatic qualifiers
  - Golfers ranked by predicted skill
- Two sources of uncertainty: latent skill prediction and round-by-round variance
- Within each simulation
  - Draw latent skill from distribution (mu = predicted skill, sd = predicted skill standard deviation)
  - Draw SG in each round from distribution (mu = latent skill, sd = round-by-round standard deviation)
  - Make a cut after first two rounds (ranking <= cut line)
  - Obtain score after four rounds and rank remaining golfers

Modeling

Earnings

- GAM with Tweedie link function
- Predict season-level real (inflation adjusted) earnings
  - Inflation rate determined by growth rate in total pool size
- $R^2 = 0.76$; RMSE = $556,304
- To calculate earnings in a future season:
  - multiply expected earnings by probability of playing on the PGA Tour
  - apply expected rate of inflation
  - normalize relative to expected pool size

Example Tweedie Distributions




```
mod_earnings <- mgcv::gam(
  data = train,
  total_money_real ~ bs(mean_latent_skill) + sd_latent_skill + age,
  family = tw(),
  select = T,
  method = 'REML'
)
```

Results

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

Results

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

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

Results

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

Results

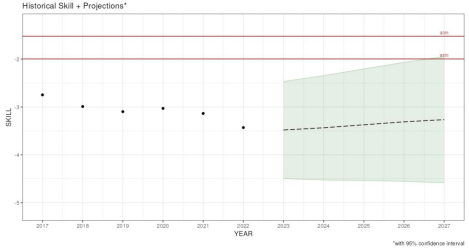
Jon Rahm

Birthdate: 1994-11-10 (28.4)

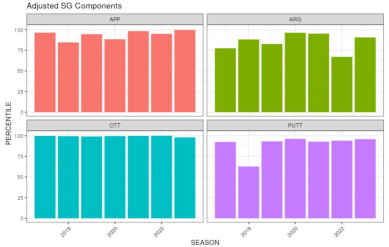
Country: ESP

	PROJECTION YEAR	PROJ RANK	SKILL	MAJOR WINS	MAJOR T5S	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 T5S	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

Future Research

- Project SG categories
- Incorporate projected SG categories in skill projection
- Simulate entire PGA tour season instead of majors only
- Include more professional tours

Summary

- End-to-end data pipeline for PGA Tour forecasting
  - ELT architecture with containerized microservices
  - Data stored in cloud Postgres database
  - Deployed via GCP
- Novel methodology for projecting skill of professional golfers
- Predictions for major tournament outcomes and future earnings
- Prototype application with historical data and model outputs

References

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/.

"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/10.1177/1527002502003003001