SDS 375: Group 01 Final Project Written Report

Project Title and Group Information

Title: Predicting and Explaining FedEx Cup Standings: Comparing Traditional Rate Statistics to Newer Strokes Gained (SG) Metrics

Group Members: Jess Erben, Quinn Hungerford, Allison Nguyen

Research Question and Objective:

Research Question:

Are traditional rate metrics or Strokes Gained metrics better predictors of PGA Tour players' total FedEx Cup points? And which individual variables within each group most strongly explain player performance across seasons?

Objective:

This project focuses on evaluating whether traditional rate statistics or more modern Strokes Gained metrics better explain and predict PGA Tour player performance, measured using FedEx Cup points. Because these points serve as the Tour's main measure of consistent success (with points allocated based on field strength, event importance, and finishing position), they offer a clear and comparable outcome variable for assessing performance.

This research provides insight into the skills that matter most for competitive success on the PGA tour by identifying which variables are most strongly associated with points and using 2024 data to forecast 2025 outcomes.

By comparing traditional golf analytics with more modern methods, our study builds on the ongoing debate about whether Strokes Gained is truly better than older rate statistics. Furthermore, contributing to the broader statistical question of whether context-adjusted performance metrics provide more performance metrics provide more reliable predictive power than raw rate statistics in high-variance competitive environments.

Data Collection and Preparation:

Source:

All data for this project was collected from the official PGA Tour statistics website, which provides downloadable tables for each performance metric. For every category, we selected the relevant year and filtered for the TOUR Championship (year-to-date) totals to capture each player's full-season performance.

Rate Statistics: Driving Accuracy Percentage, Green in Regulation Percentage, Scrambling, Overall Putting Average

Strokes Gained Statistics: Off-The-Tee, Approach the Green, Around the Green, Putting

Outcome Variable: FedEx Cup Points

Merging Process and Handling Missing Values:

We created two merged datasets, one for 2024 and one for 2025. The 2024 dataset serves as the training data, while the 2025 dataset is used for testing, but we also explored what explained point variation within each. Each dataset contains player-level season summaries through the TOUR Championship (September 1, 2024, and August 24, 2025, respectively). Merging the data was done in this Jupyter Notebook (.ipynb). We filtered both datasets to

remove players (about 50 for each year) with missing values in the key variables listed below. This step removed 55 incomplete records from 2024 and 66 from 2025, leaving only players with fully observed data for modeling and visualization.

Key Variables (Both):

Total Points, Player, Player ID, %_FIR, %_GIR, %_Scrambling, Avg_Putting, AVG_SG_Tee, AVG_SG_APP, AVG_SG_ARG, AVG_SG_PUTTING

2024 Master Dataset

Time period: 2024 PGA Tour Season through the TOUR Championship, Sept 1

Size: 184 players with 29 variables.

2025 Master Dataset

Time Period: 2025 PGA Tour Season through the TOUR Championship, Aug 24

Size: 171 players with 29 variables.

Exploratory Data Analysis:

Summary Statistics Tables:

2025 Summary Statistics Table of Key Variables

	POINTS PER EVENT	TOTAL POINTS	%_FIR	%_GIR	%_Scrambling	AVG_Putting	AVG_SG_Tee	AVG_SG_APP	AVG_SG_ARG	AVG_SG_PUTTING
count	237.000000	237.000000	171.000000	171.000000	171.000000	171.000000	171.000000	171.000000	171.000000	171.000000
mean	27.868629	518.212460	0.596384	0.664716	0.597416	1.605526	0.028222	0.065924	0.017053	0.028819
std	35.831329	621.877468	0.051722	0.026328	0.032891	0.027421	0.357949	0.365618	0.221561	0.320303
min	0.249000	1.493000	0.418987	0.584795	0.500000	1.521000	-1.469000	-0.984000	-0.570000	-1.068000
25%	6.850000	82.612000	0.557953	0.649263	0.577981	1.587000	-0.185000	-0.158000	-0.119000	-0.168500
50%	16.115000	313.717000	0.596059	0.663248	0.597473	1.603000	0.076000	0.079000	0.013000	0.055000
75%	33.913000	689.953000	0.629775	0.683451	0.618183	1.622500	0.249000	0.305000	0.151500	0.213500
max	300.373000	4805.967000	0.738523	0.724570	0.704225	1.673000	0.748000	1.291000	0.610000	0.983000

2024 Summary Statistics Table of Key Variables

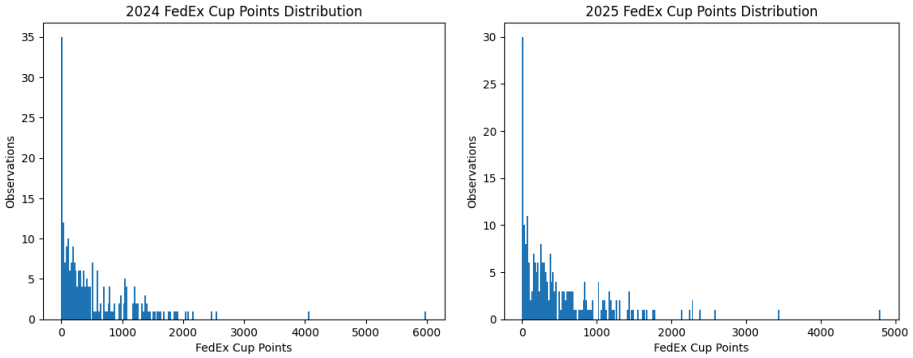
	POINTS PER EVENT	TOTAL POINTS	%_FIR	%_GIR	%_Scrambling	AVG_Putting	AVG_SG_Tee	AVG_SG_APP	AVG_SG_ARG	AVG_SG_PUTTING
count	239.000000	239.000000	184.000000	184.000000	184.000000	184.000000	184.000000	184.000000	184.000000	184.000000
mean	29.588749	546.288276	0.612287	0.672248	0.595860	1.610989	0.022821	0.034467	0.015408	0.020549
std	39.160116	687.231830	0.045846	0.028284	0.036000	0.030285	0.362693	0.372632	0.241351	0.332104
min	0.372000	1.804000	0.490642	0.573286	0.498462	1.543000	-1.352000	-1.047000	-0.932000	-0.954000
25%	6.599500	91.799500	0.583434	0.654018	0.573504	1.590750	-0.180250	-0.206250	-0.127750	-0.207500
50%	15.399000	307.976000	0.609762	0.674363	0.594695	1.612000	0.065000	0.098500	0.048000	0.038500
75%	40.513000	802.427500	0.644493	0.694083	0.621355	1.630000	0.259500	0.268500	0.171250	0.258500
max	374.547000	5992.750000	0.720168	0.742063	0.707260	1.713000	0.883000	1.269000	0.580000	0.866000

Description and Reflection:

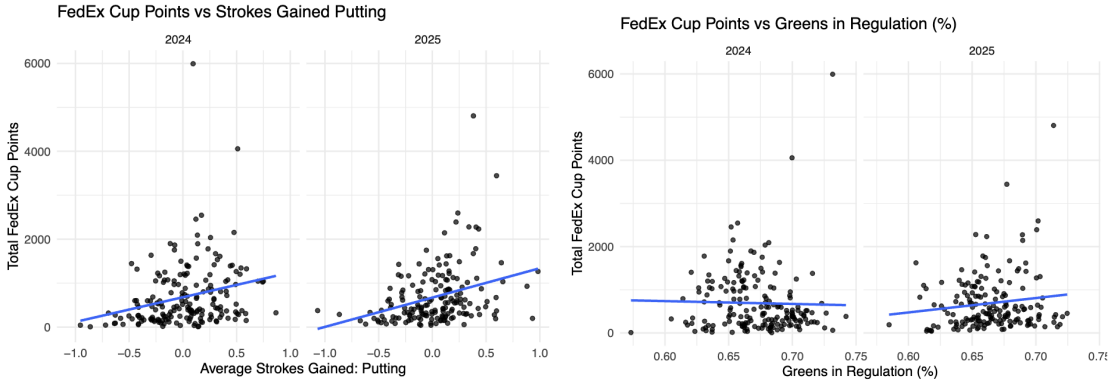
The summary statistics in Tables 1 and 2 show consistent variable distributions across both seasons. Average Greens in Regulation (%_GIR) and Scrambling rates are stable from 2024 to 2025, suggesting little variation from one year to another in these rate statistics. Comparatively, the Strokes Gained metrics (especially SG: Approach and SG: Putting) show greater variability, reflecting larger performance differences among top players. Mean FedEx Cup points are slightly lower in 2025, which seems consistent with a smaller sample of measured players or just general differences in participation in tournaments.

Visualizations:

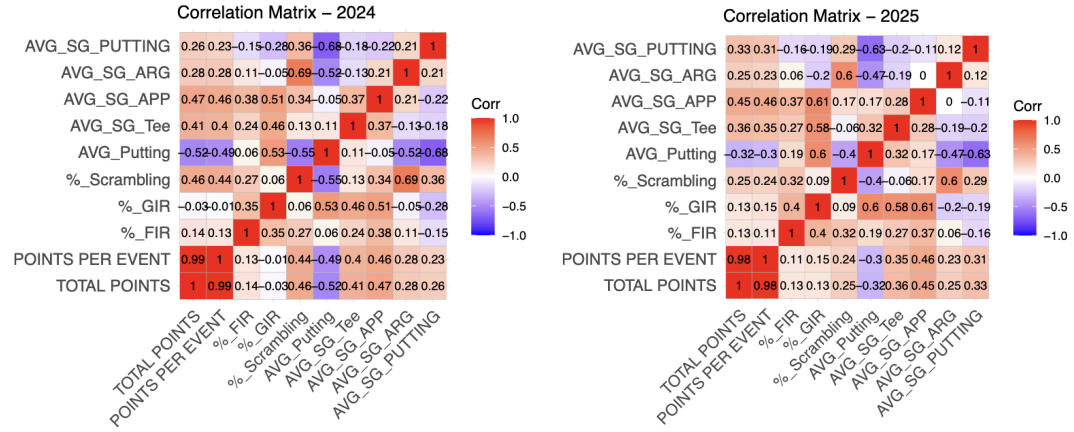
Histograms of Total FedEx Cup Points for 2024 and 2025



Scatterplots: FedEx Cup Points vs Strokes Gained Putting, FedEx Cup Points vs Greens in Regulation (%)



Correlation Heatmap Matrix: 2024 and 2025



Description and Reflection:

The first two visualizations of total FedEx cup point distributions for the 2024 and 2025 seasons show that the distributions of points are very right skewed and not normal. We will discuss this more later in the report, but this was a key observation for us to ensure this variable meets the linear model assumptions of normality by exploring different approaches to transforming the distribution.

The last four visualizations highlight clear differences in how traditional and advanced golf performance metrics relate to player success. In both 2024 and 2025, Strokes Gained Putting shows a moderately positive relationship with total FedEx Cup points, while Greens in Regulation (%) displays little to no association, suggesting here that putting performance may be a stronger driver of success than the more basic accuracy measures. The correlation heatmaps reinforce this observation, as Strokes Gained metrics, particularly SG: Approach and SG: Tee, show stronger correlations with total points compared to the traditional rate statistics like %_GIR and %_FIR. Overall, the plots suggest that advanced Strokes Gained measures better capture performance differences among players and are more consistent across seasons, so they likely provide greater explanatory and predictive value for understanding FedEx Cup outcomes.

Methodology

Our analysis consisted of two main components: an explanatory analysis to understand which performance metrics best explain variation in FedEx Cup points, and a predictive analysis that uses 2024 data to forecast 2025 outcomes.

1. **Data Collection & Preparation**

   To identify relevant variables and shape our research question, we first conducted background research on the use of traditional rate statistics and Strokes Gained metrics in evaluating PGA Tour performance. We then downloaded the 2024 and 2025 player statistics from the official PGA Tour website and merged the many tables for each season using Player ID. Also, before analysis, we removed players with missing values in key predictors.

2. **Exploratory Analysis and Normalization**

   We began by computing summary statistics and correlation matrices for all variables in the 2024 and 2025 datasets. We found that FedEx Cup points were highly right-skewed, with many players earning zero points. Because linear regression relies on approximately normal residuals, we applied a Box–Cox power transformation, shifting points by +1 to accommodate zero values. The optimal Box–Cox parameter (0.26) was estimated using 2024 data, and the transformation was then applied to both years. We also explored the approach of transforming the points within the glm() function by setting the family equal to Gamma(link = log) (due to the points mirroring a Gamma distribution).

3. **Model Building (Training on 2024)**

   Rate Stats GLM:

   - Box-Cox FedEx Points ~ FIR% + GIR% + Scrambling% + Putting Avg.
   - FedEx Points ~ FIR% + GIR% + Scrambling% + Putting Avg, family = Gamma(link = "log")

   Strokes Gained GLM:

   - Box-Cox FedEx Points ~ SG Off-the-Tee + SG Approach + SG Around-the-Green + SG Putting.
   - FedEx Points ~ SG Off-the-Tee + SG Approach + SG Around-the-Green + SG Putting. family = Gamma(link = "log")

These models were chosen to directly compare traditional rate statistic performance measures with modern Strokes Gained ones.For each model, we examined the explanatory power of both the models and specific factors within using R squared and statistical significance of predictors.

4. **Evaluate Predictive Performance (Testing on 2025)**

Using the fitted 2024 models, we generated predictions for the 2025 dataset on the transformed scale and then inverted the Box–Cox transformation to return predictions to the original FedEx Cup points scale. We evaluated predictive accuracy using multiple metrics:

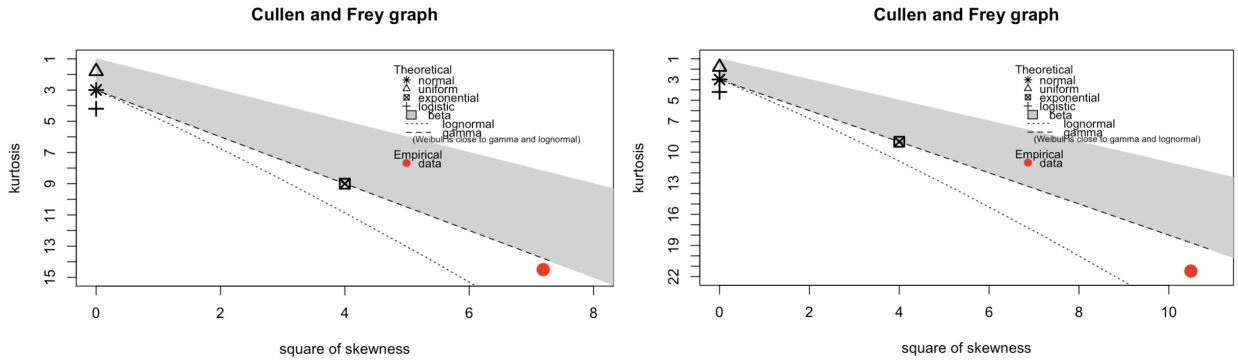
- Predictive correlation between predicted and actual points
- RMSE (root mean squared error)
- MAE (mean absolute error)
- R squared on test data
- MAPE (mean absolute percentage error)

We also visualized model performance by plotting actual vs. predicted FedEx Cup points and comparing predicted rankings vs. actual rankings for both models.

Results

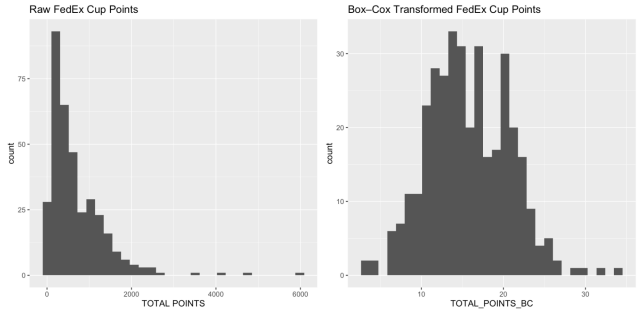
Before building our models, we needed to normalize FedEx Cup points because they were highly right skewed, with many players earning 0 points. This violates linear model assumptions by introducing non-normal residuals and heteroscedasticity, which would result in an overall bad model fit without transformation. We explored two approaches to transforming the FedEx Cup points. The first was a Box-Cox Power Transformation, which shifted points by +1 to handle zeros and then estimated λ to be 0.26 using 2024 data. The transformed distribution of FedEx Cup points from this approach is shown below. The second method was to use a Gamma GLM with a log link in which the lm() took a family = Gamma(link = “log”) argument. Both produced similar results.

Confirming the Gamma Distribution



While we knew our data was right-skewed, we created a Cullen and Frey plot for the 2025 (left) and 2024 (right) seasons to confirm that they resembled a Gamma distribution and that the method would be valid. 2025 fell more close to the Gamma line than 2024. This ambiguity in determining if our data actually follows a Gamma distribution may explain why the Gamma method was not as effective as the Box-Cox method in our analysis

Normalizing Outcome Variable (FedEx Cup Points)



Explanatory Analysis (2024 Season Discussed Here)

We first fit separate linear models using traditional rate statistics and Strokes Gained metrics to explain variation in the transformed FedEx Cup points during the 2024 season.

The Box-Cox model using %FIR, %GIR, Scrambling, and Putting Average explained 48.8% of the variation in transformed FedEx Cup points, while the Gamma explained 48.56% of the variation of FedEx Cup points. Both methods had only two statistically significant predictors. The first was Greens in Regulation (%), where players who hit more greens scored substantially more FedEx Cup points. The second was Putting Average, where more putts per hole was strongly associated with worse performance. Fairway-in-Regulation (%) and Scrambling were not significant. Overall, rate-based measures only captured a small amount of the skill variation relevant to scoring FedEx Cup points.

The models using SG: Off-the-Tee, Approach, Around-the-Green, and Putting provided a much stronger explanatory fit, where the Box-Cox method had an R squared of 0.6039 in transformed FedEx Cup points, and the Gamma method had an R squared of 0.7339696. In both methods, all four SG variables were statistically significant and positively associated with FedEx Cup points. This suggests that the more modern Strokes Gained performance metrics better capture the different components of player skill that the more traditional rate statistics overlook.

Predictive Analysis (Forecasting 2025)

We also assessed the models' ability to forecast 2025 FedEx Cup points using parameters trained on 2024 data. The Gamma method did not perform as well as the Box-Cox, where the SG model had a R-Squared of 0.3747858 and an RMSE of 513.9553, and the Rate model had an R-Squared of .02699162 and an RMSE of 641.1642. For the Box-Cox method, we generated predictions on the Box-Cox scale and then transformed back to the original point scale.

- Predictive Correlation
    - Rate Stats Model: r = 0.49
    - Strokes Gained Model: r = 0.81. This model more than doubled the predictive accuracy of the rate statistics model.
- Full Predictive Metrics

Model	RMSE	MAE	R2	MAPE
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
Rate Stats	583.5611	385.3556	0.1939709	87.10230
Strokes Gained	386.2946	279.6004	0.6468043	52.38984

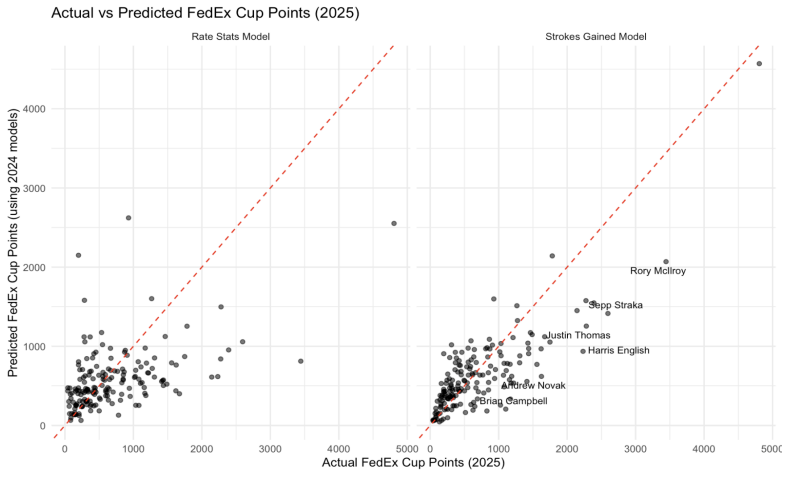
There are some key takeaways from this table of the full evaluation metrics we looked at to compare model performance on predicting 2025 points. First, the RMSE (Root Mean Squared Error) is about 34% lower for the SG model, which means that, when the model makes big mistakes for high scoring players, the SG model is more accurate and makes fewer extreme errors for this end of the distribution.

Second, the MAE (Mean Absolute Error) is the average size of the model's prediction error. On average, the SG model's predictions are about 105 FedEx Cup points closer to the truth than the rate stats model's predictions. In context, 100 points is about equal to a top-15 finish in a regular PGA event so this is a big improvement.

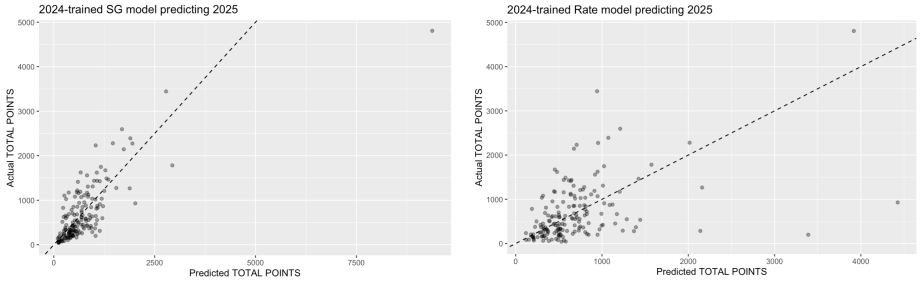
Third, R squared (coefficient of determination) for the SG model is 0.65, so it captures 65% of how/why players scored the way they did in 2025, meaning it finds the real skill relationships across seasons. The rate stats model explains only 19%, so it can't generalize well or capture the true performance variation. This is very low; only a bit better than guessing.

Finally, MAPE reflects overall percentage error. So, for the SG model, predictions are by 52% off on average which is a much smaller relative error than the rate stats model of 87%.

- Actual vs Predicted Scatterplots
  - Box Cox:



  - Gamma:



The two Box-Cox scatterplots here show that the SG predictions cluster closely to the 45 degree reference line, especially for the higher-performing players. The rate

statistics model shows much more underestimation for elite players and just general noisy predictions across the plot. The two Gamma scatterplots show trend lines rather than a 45 degree line, but also show a tighter trend and higher slope for the SG predictions than the rate stat predictions.

- Predicted 2025 Top-20 Ranking Accuracy (SG Model)

Pred_Rank	Player	Pred_SG	Actual_Rank	Actual
1	Scottie Scheffler	4570.535	1	4805.967
2	Tommy Fleetwood	2141.669	9	1782.791
3	Rory McIlroy	2068.740	2	3444.328
4	Harry Hall	1597.722	44	929.309
5	Ben Griffin	1576.298	6	2274.913
6	Russell Henley	1547.981	4	2390.884
7	Sam Burns	1512.059	24	1266.332
8	J.J. Spaun	1450.446	8	2144.131
9	Sepp Straka	1414.461	3	2595.307
10	Patrick Cantlay	1324.449	23	1275.202
11	Justin Thomas	1254.637	5	2279.907
12	Cameron Young	1172.146	16	1464.474
13	Robert MacIntyre	1147.246	15	1488.235
14	Maverick McNealy	1120.694	11	1672.214
15	Viktor Hovland	1109.681	26	1209.606
16	Jordan Spieth	1088.466	48	864.638
17	Nicolai Højgaard	1068.871	73	595.975
18	Keegan Bradley	1053.075	10	1748.967
19	Collin Morikawa	1039.827	19	1427.158
20	Denny McCarthy	1035.139	39	1033.723

As shown here, the Strokes Gained model made player point rankings that closely mirrored actual 2025 FedEx Cup standings, particularly for the top players. Some examples of this are Scottie Scheffler (predicted #1, actual #1), Rory McIlroy (predicted top 3), and Russell Henley (predicted top 6, actual top 5). Most of the top 10 players were placed somewhat close to their true rankings. There were a couple big misses like Harry Hall and Jordan Spieth, which could have come from limited participation in tournaments or inconsistent play.

Limitations

There are a couple limitations to consider when interpreting our findings. For one, we did remove a fairly large number of players due to missing data, which could introduce bias towards players who participated more consistently throughout the season. Also, we cannot capture short-term performance changes very well since our models are based on season averages. Another limitation is that we are assuming stable relationships between skills and final FedEx Cup points across seasons, while they can shift across years. Finally, the models don't directly account for contextual factors like weather, number of events played, or injuries, which can introduce uncertainty into individual predictions. Overall, these limitations should be kept in mind when generalizing our conclusions.

Discussion and Conclusion

Our research clearly shows that Strokes Gained metrics outperformed traditional rate statistics in explaining and predicting PGA Tour performance. Across both the 2024 and 2025 models, SG metrics consistently captured more variation in FedEX Cup points and produced more accurate predictions.

In the explanatory analysis, the rate statistics model explained 48.8% of the variation in the transformed FedEx Cup points and identified only two significant predictors (GIR% and Putting Average). However, the Strokes Gained model explained 60.39% of the variation and identified that all four SG metrics(Off-the-Tee, Approach, Putting, and Around-the-Green) were significant. SG metrics better capture the underlying skills that drive scoring, likely because they account for context such as shot difficulty and lie.

The results of the predictive model also support that the SG approach is better. When applied to 2025 data, the SG model achieved a predictive correlation of 0.81, far higher than the rate statistics model (0.49). The RMSE, MAE, R^2, and MAPE also consistently supported the SG model, with predictions that averaged about 105 FedEx Cup points closer to the truth. In its final ranking predictions, the SG model's closely matched the actual 2025 standings, especially for top players like Scottie Scheffler and Rory McIlroy.

While we should keep in mind the limitations listed above, our overall conclusion is that Strokes Gained metrics provide more reliable and informative measures of PGA golf player performance than traditional rate statistics. Approach and putting skill emerged as the most influential drivers of FedEx Cup success, highlighting where player performance matters most across PGA Tour seasons.

Works Cited

Broadie, M. (2012). Assessing golfer performance on the PGA TOUR. Interfaces, 42(2), 146–165. https://doi.org/10.1287/inte.1120.0626

Dealing with Right-Skewed Data. (n.d.). CRAN. Retrieved November 9, 2025, from https://cran.r-project.org/web/packages/GlmSimulatoR/vignettes/dealing_with_right_skewed_data.html

PGA Tour Statistics. (n.d.). PGA Tour. Retrieved October 16, 2025, from https://www.pgatour.com/stats

Transformations. (n.d.). Cornell CSS. Retrieved November 12, 2025, from https://www.css.cornell.edu/faculty/dgr2/_static/files/R_html/Transformations.html

Validating Strokes-Gained Method: Measuring PGA Tour Player Success. NYC Data Science. Retrieved October 29, 2025, from https://nycdatascience.com/blog/r/validating-strokes-gained-method-measuring-pga-tour-player-success/