Professional Golf Analytics: Predicting and Improving Players' Performance

Theo Ginting, Nathan Ebikwo, Matthew Kusno, Dawson McMahon, Anirudh Suresh, Matthew A. Lanham

Purdue University Krannert School of Management

tginting@purdue.edu; nebikwo@purdue.edu; mkusno@purdue.edu; mcmaho19@purdue.edu; suresh14@purdue.edu; lanhamm@purdue.edu

Abstract

This project attempts to conduct an analysis of historical data on extensive statistics on professional golf tour, to show what statistics make a good proxy of better relative performance in golf. Our analysis will provide the ability to turn raw golf statistics into valuable insights to help players better understand what aspects of their performance leads to better overall performance. To gain the insights we will create a regression model that takes the statistics on different performance on different part of the game as predictor to predict relative performance.

Introduction

Golf is a sports that grants player to win up $2 million dollar in a single tournament and $ 10 million prize to the best golfer in a season. Having the ability to turn raw statistics into meaningful game improving information can be the difference between winning million of dollars and missing the tournament cut line.

Insert Video Here

Research Question:

- In professional golf, what physical attributes lead to the best performance relative to the average field?
- In professional golf, what aspect of the golf game contribute the most to superior performance relative to the rest?

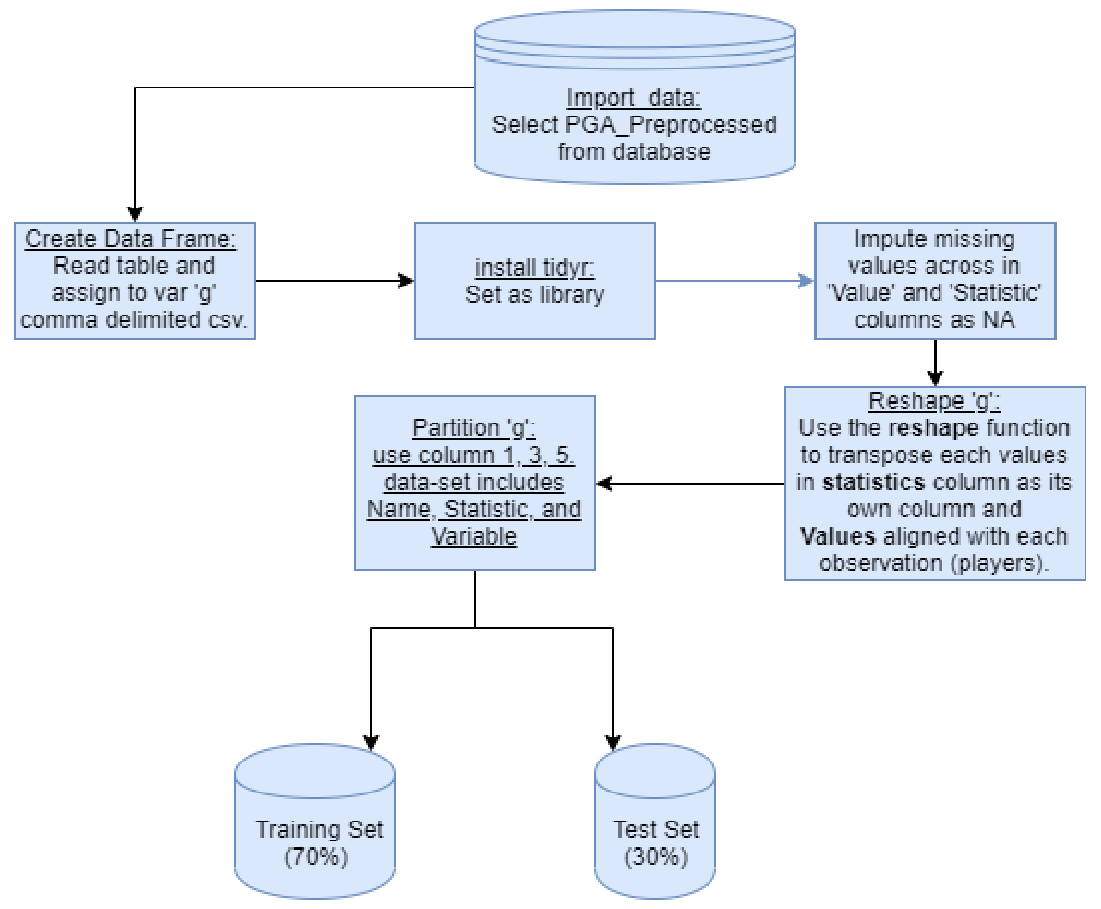
Literature Review

Assessing Golfer Performance on the PGA TOUR (Broadie, 2012) first coined the novel statistic, Strokes Gained (SG). Strokes Gained is a measure of quality of a golf shot, adjusting for variances that a golf shot are subject to, it shows how much better or worse a shot is to a benchmark function derived from calculation of eight million shots.

Strokes gained is created to better understand of the contribution of individual shot to the overall score.

This study is novel because we attempt to break-down different aspect of the game and predict by how much each aspect contributes to better Strokes Gained figure.

Methodology


Import data:
Select PGA_Preprocessed from database
Create Data Frame:
Read table and assign to var 'g' comma delimited csv.
install tidyr:
Set as library
Impute missing values across in 'Value' and 'Statistic' columns as NA
Reshape 'g':
Use the reshape function to transpose each values in statistics column as its own column and Values aligned with each observation (players).
Partition 'g':
use column 1, 3, 5.
data-set includes Name, Statistic, and Variable
Training Set (70%)
Test Set (30%)

Data

The data for this project is a collection of all statistics over the span of a year on a player in the PGA TOUR from Kaggle.com. The data has 8 main categories relating to different aspect of a player's performance on course.

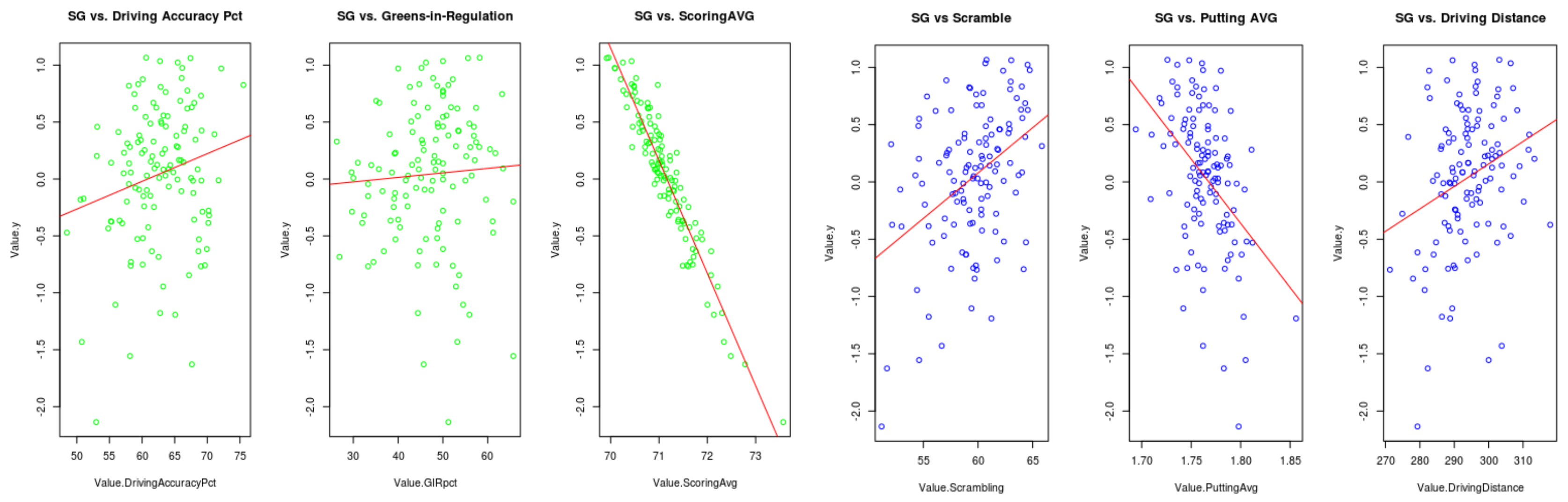
Data Cleaning & Pre-Processing

To clean the data we used **tidyr** to impute all NAs. We then pre-process the data from long format into wide format using spread function. Finally to make the data set into a format that the model can ingest, we used the reshape function so every feature would be assigned to one name, eliminating duplicates.

In order to partition the data we used the createdatapartition() function with 70% of the data allocated to the training and 30% allocated to the test set.

Feature Selection

We selected few attributes that seems interesting and will not raise issue of correlation with out predicted values, Stokes Gained Total.



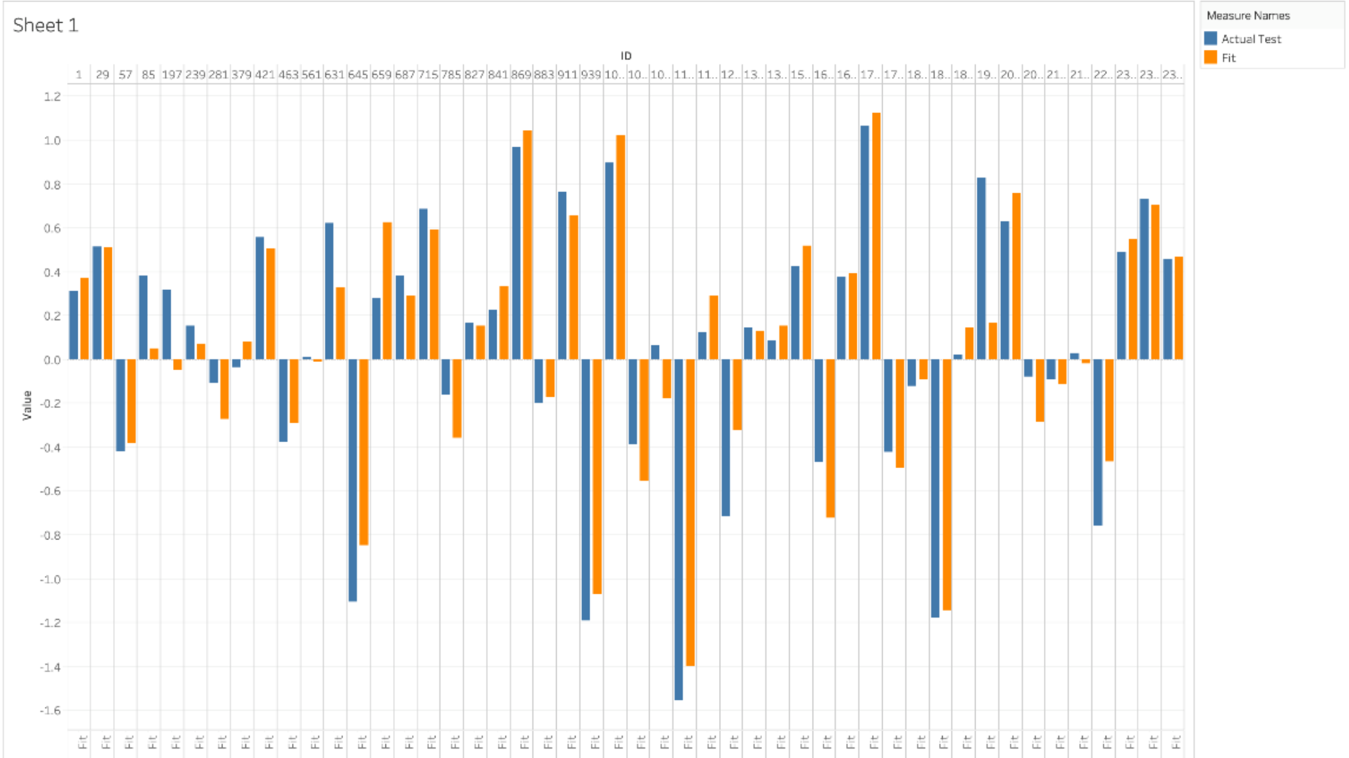
Model

We specify a linear regression model on a 70/30 split train/test data with a 3-fold cross-validation design because we find that this number of fold leaves us with a desirable error rate estimation and computing time.

Model Evaluation

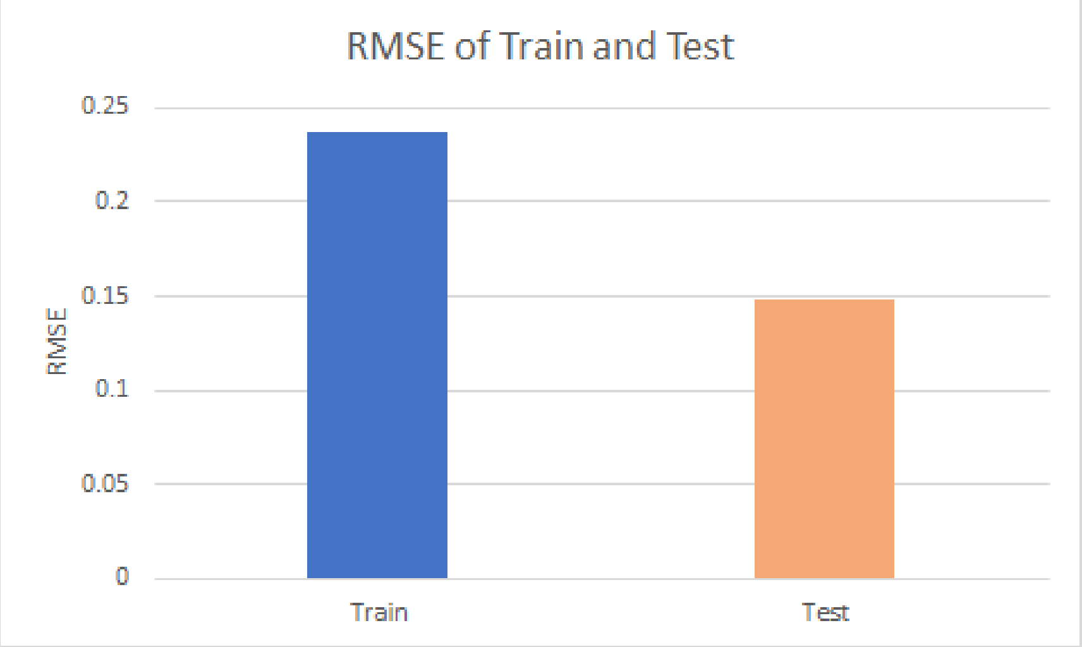
The predictive models were evaluated on overall accuracy, Adjusted R-Squared, and by looking at this metric, our model explained 92.6% variability in our Predicted Value.

Results



This shows the result of our model predicting the Strokes-Gained. The orange line represent the predicted value by our model and the blue line represent the actual value from our test set.

Our model explained 92.6% variability in our Predicted Value.



Furthermore, looking at the RMSE of our model, we got an RMSE that is lower in our test set, indicating a better performance

Conclusions

In conclusion, we are able to create a model that takes several statistical parameters from the raw data and estimates their effect on SG.

Although height and weight resulted in a weak positive correlation, we know Scrambling positively affects SG with a correlation of .50 and other variables negatively affect Strokes Gained such as Scoring Average which had a strong negative correlation of -.953. Additionally, Driving Accuracy Percentage and Greens In Regulation Percentage showed weak positive correlations with total SG while Putting Average showed a moderate negative correlation to Strokes Gained.

By knowing how greatly these endogenous variables affect SG, players can fine-tune their practice to be better at parts of the game that improve SG the most.

Acknowledgements

We thank Professor Matthew Lanham for constant guidance on this project.