Using Cumulative PGA Tour Statistics to Predict Master's Performance

Timothy Morales

Abstract

Predicting sports outcomes has proven a large topic of interest after the legalization and emergence of sports books around the world. Statistical regression methods are commonly used to model sports outcomes in a pursuit to monetize off common sporting events. Regression models are interesting as they allow for both prediction and inferential discovery. Throughout this analysis, cumulative PGA Tour statistics from the 2019 season dating up to April 7th, 2019, are fit to various regression models to predict the results of the 2019 Master's golf tournament. The paper looks at two commonly used regression and dimension reductions techniques, partial least squares regression (PLSR) and principle component analysis (PCA) paired with LASSO penalized regression. Both methods involved using pair-wise correlation cutoff tuning in the variable selection process. The analysis shows that PLSR with a correlation cutoff of 0.8 during the variable section process performs best in terms of root mean squared error (RMSE) and R squared values of resamples of the original data as well as prediction accuracy metrics.

Introduction

Routinely played on the first full week of April each year, the Master's tournament is one of golf's most coveted and prestigious events. The tournament consists of four separate rounds of 18 holes where the lowest score wins. With mass fan popularity, the Master's finds itself as one of the most commonly bet sporting events of the year. From "Master's pools" to individual prop bets, there are a variety of ways to monetize on the results, but all are centered around individual performance and overall finish.

Conceived in 1934, the Master's has a long-standing history in the game of golf with a very unique feature as it is continually played at the same venue every year. Augusta National Golf Club in Augusta, Georgia routinely hosts the tournament and although there have been modifications to the course in terms of tee box location and bunker placement, the course has largely remained the same over the years. This unique feature of consistent location poses an interesting proposition to bettors everywhere as results between years may be correlated and certain statistics may be telling of the results year in and year out.

The PGA Tour is the America's most well-known golf association. Established in 1929, it is the main organizer of many of the American based golf tournaments each year. Players compete to be allowed on the PGA Tour and get exemptions to compete in certain tournaments in order to qualify for others. Readily available on the PGA Tour website, are over 1000 different player statistics for those who compete in PGA Tour sponsored tournaments during the year. These statistics are cumulative and updated weekly after the culmination of rounds in events. Of the 87 players in the 2019

Master's, 65 of those had prior recorded statistics on the PGA Tour in 2019.

This report looks into
1.) The correlation between these 2019 PGA Tour statistics dating up to the week prior of the Master's tournament, and the individual's outcomes of the 2019 Master's.
2.) Comparing Regression-based models and dimension reduction techniques, when using these PGA Tour statistics to fit the individual outcomes in the 2019 Master's.
3.) Variable importance in the most predictive model in the 2019 iteration of the tournament.

Keywords: The Master's, Partial Least Squares, Dimension Reduction, Penalized Regression, Principle Component Analysis.

Methods

Data Preprocessing

The analysis started with two separate datasets, both of which were derived from the PGA Tour's statistical database. The first was the 2019 Master's results consisting of 87 observations and overall outcomes such as earnings, finish to par, and ranked final finish as well as individual rounds scores. Only the overall tournament outcomes were used, and individual rounds were not analyzed. The second dataset was longitudinal data in long format, consisting of 77 separate players and 1479 variable metrics updated weekly across the 2019 season. Only the cumulative data dating up to the week of April 7th, 2019 was used and only individuals with information in both datasets were kept. This resulted in 65 individuals with 1484 total variables when the two datasets were merged.

Within this final dataset, there were many missing observations. These missing observations could be split into two separate groups. The first group was where NA denoted a value of zero. The second group was where certain variables were not applicable to certain players or where a certain player did not have a listed value. Missing values in the first group were replaced with zero. Variables in the second group were dropped, as imputation in a sports context may have large consequences and with over 1400 variables, dropping these variables did not result in a massive loss of information.

Also, within the dataset were variables that held the exact same information as another variable such as number of holes played in a variety of different contexts. In this case, only the first of such variables was kept in the dataset. There were also variables that indirectly told the same story as other variables in the dataset. For example, the variable, holes played, and the variable, number of rounds played, give the same information. In that case, only the variable that gave more overall information or more unique information was kept. Counting variables that had a complimentary averaged variable accounting for the same metric in the dataset were dropped, as there are already variables such as number of holes played within the dataset to account for these differences. All character strings were also removed from variables using the stringr package in R (R Core Team. 2019).

The response variable used for the regression model was the log of the earnings variables. The earnings variable denotes the total dollar winnings for a player in the tournament. The log of earnings was taken because earnings increases at an exponential rate relative to overall finish in the tournament. Earnings was chosen over finish to par for many reasons. Those who get "cut" and therefore tie for last place, do not advance to the third round. Using finish to par would heavily skew the results, while those who are cut prior to the third round are paid the default amount of $10,000. Using the earnings variable also gave higher finishes greater influence in the regression models. The higher influence helps the model in assuring those in the top 10 and top 5 positions have a greater weighting (Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. 2013).

In order to further reduce the dimensions of the dataset, the findCorrelation function from the caret package in R was used on the PGA Tour statistics. The findCorrelation function looks at the absolute value of each pairwise correlation in the dataset and compares it to a manually set threshold. If a pairwise correlation is greater than the set threshold, it returns the index of the variable within the pair with the largest mean of absolute correlation and that variable was subsequently dropped (Kuhn. M. 2008). This correlation cutoff was used as a tuning parameter and four separate datasets were created from varying cutoff values. Those four datasets consisted of the original dataset with no cutoff, and datasets produced from a correlation cutoff of 0.7, 0.8, and 0.85. Any near zero variance variable were dropped again through the caret package.

Model Analysis

For each of the four datasets created using the correlation cutoffs, both a LASSO regression using all components from PCA as well as a PLSR were fit to the data. A LASSO regression without PCA was also fit to the dataset with no correlation cutoff. These models were fit using modifications of the train function in the caret package with leave one out cross validation (LOOCV) as the resampling technique for all models (Kuhn. M. 2008). LOOCV was used because in small sample size datasets, it ranks highly in mirroring future observation error rates (Molinaro AM, Simon R, Pfeiffer RM. 2005).

For the PLSR, the optimal number of components was chosen using the root mean squared error of predicted earnings values (RMSEP) of the LOOCV validation sets. The same procedure was used to optimize the shrinkage coefficient in the LASSO model. For the LASSO shrinkage coefficient, 100 evenly spaced potential coefficient values from 0.01 to 10_{10} were used.

LASSO was chosen over other penalized regression methods due to its dimension reduction capabilities and strong performance in other sports-based multidimensional penalized regression models (South, Charles & Egros, Edward. 2020). Optimization of an elastic net model may have potentially led to overfit models with little future predictive power. The LASSO has the unique ability to reduce certain coefficients to zero through its coefficient shrinkage and thus aid in avoiding overfitting the data (Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. 2013).

Two commonly known metrics, RMSE and R squared, were used to compare the models. The RMSE and R squared for each were calculated from 25 random resamples using the resamples function in the caret package (Kuhn. M. 2008). Median RMSE and R squared values were used as the determining factor. Two custom metrics were also used to assess model fit, they were Top 10 accuracy as well as Top 20 accuracy. These variables denote the proportion of the 65 observations used to fit the models that were correctly predicted to be within the top 10 and top 20 in Master's finish of the 65 original observations. Projected and observed finish were denoted by the order of either the projected or the actual earnings, with 1 denoting the highest and 65 the lowest. Ties were given the highest tied rank as is common practice in the Master's. RMSE of these finish values was also calculated for each model.

The features within the models were centered and scaled prior to the analysis in order to standardize units across all features. These steps are both essential in properly weighting the features and helps give a more interpretable and comparable explanation of coefficient values (Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. 2013).

Finally, for the selected optimal model, variable importance was reported on a 0-100 scale, with 100 denoting the highest importance, using the varImp function in the caret package. The varImp function looks at the sum of the absolute weights of coefficients for each variable and scales them appropriately for partial least squares models (Kuhn. M. 2008). Variable importance for the ideal model was plotted and used to tell a comprehensive story of which type of player succeeded at the Master's in 2019.

Results

Table 1. Number of Components Selected/Retained in Model Fit.

Partial Least Squares Optimization	
Correlation Cutoff	Number of Components
None	1
0.7	1
0.8	1
0.85	1
LASSO Optimization	
Correlation Cutoff	Number of Components
None	1
0.7	3
0.8	6
0.85	5
None No PCA	9

Table 2. Optimal Penalty Coefficient in LASSO Models.

LASSO Optimization	
Correlation Cutoff	Lambda Value
None	0.3764936
0.7	0.3764936
0.8	0.2848036
0.85	0.2848036
None No PCA	0.3764936

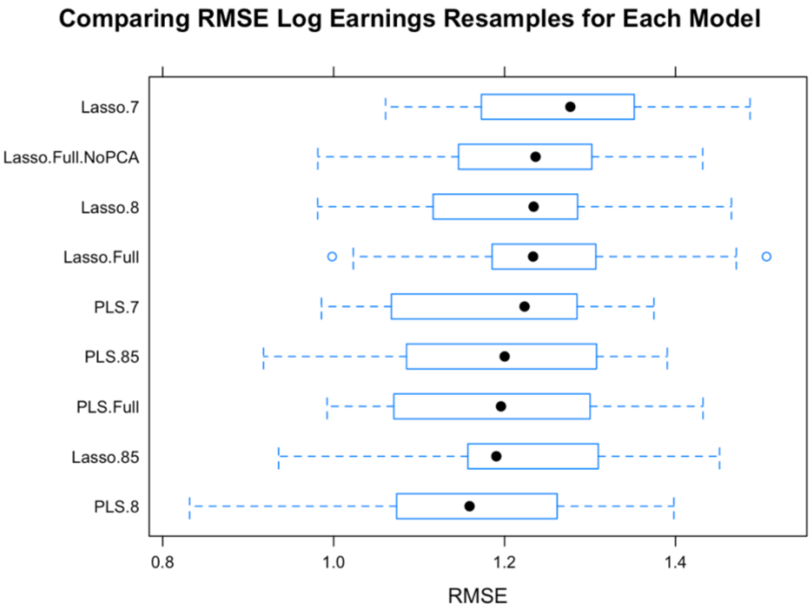

Figure 1. Distribution RMSE for each model of the 25 resamples. PLSR with pairwise correlation cutoff of 0.8 best performance.

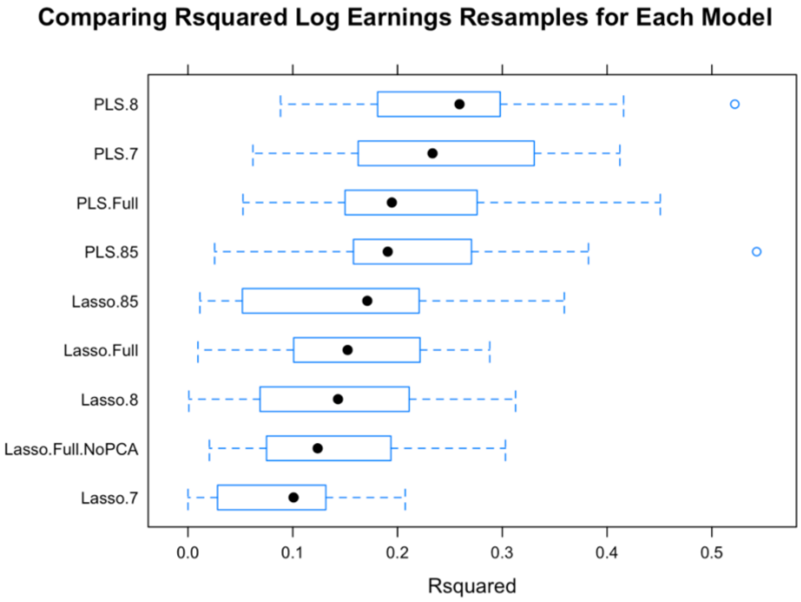

Figure 2. Distribution R squared values for each model of the 25 resamples. PLSR with 0.8 as correlation cutoff performs best.

Table 3. Top 10 and Top 20 Accuracy Across models. PLSR with correlation 0.8 performs best in all metrics other than RMSE of Finish, which PLSR (0.7) performs best.

Predictive Accuracy Across Models			
Model (Correlation Cutoff)	Top 10 Accuracy	Top 20 Accuracy	RMSE of Finish
PLSR (None)	0.6	0.7	16.28024
PLSR (0.7)	0.6	0.7	**14.57712**
PLSR (0.8)	**0.7**	**0.8**	14.70008
PLSR (0.85)	0.6	0.75	15.26686
LASSO(None)	0.4	0.6	18.13284
LASSO(0.7)	0.5	0.6	17.83601
LASSO(0.8)	0.5	0.7	16.19022
LASSO(0.85)	0.6	0.75	16.048
LASSO(None No PCA)	0.6	0.65	14.94606

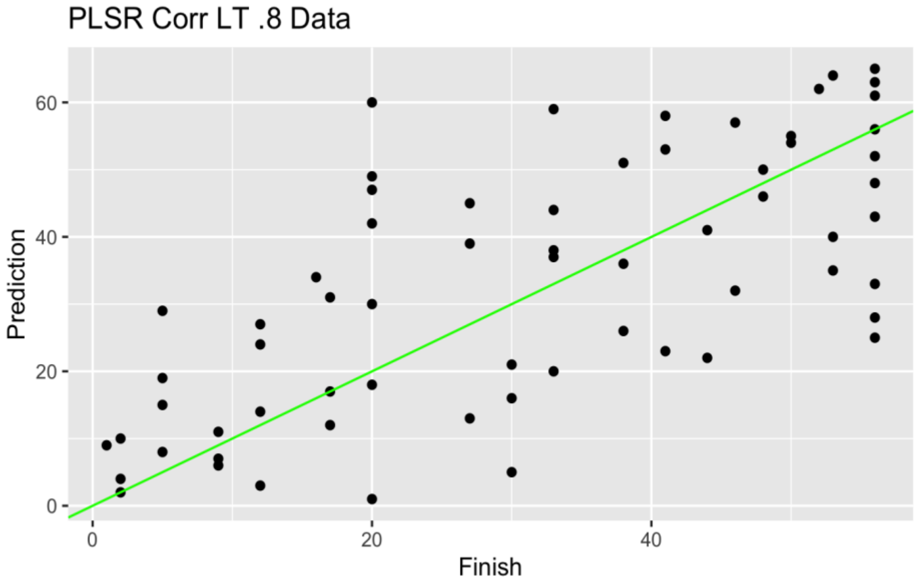

Figure 3. Observed and Predicted Finish for the 65 observations when fitting with the PLSR with a correlation cutoff of 0.8. y=x line denoting perfect prediction.

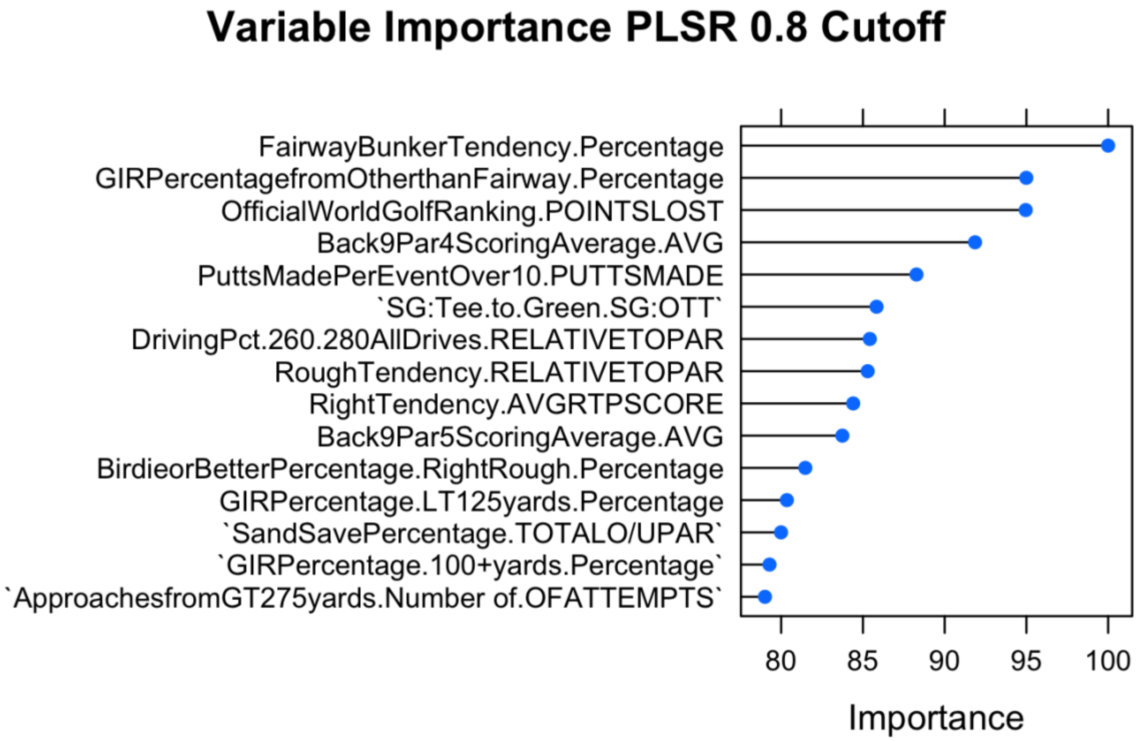

Figure 4. Top 20 variables by importance in PLSR with cutoff of 0.8

Discussion

Model Prediction

PLSR consistently outperformed the LASSO with PCA models in both the RMSE and R squared values as well as in terms of the in the finish accuracy metrics. This may be due to PLSR's ability to prioritize features with higher correlation to the response variable and thus handle noise throughout the dataset.

In terms of correlation cutoff, there is no clear optimal value between the models, but a correlation cutoff of 0.8 with PLSR gives the optimal result. The PLSR with the correlation cutoff of 0.8 performed best in every metric used to compare models other than RMSE of Finish, which it produced the second-best results (0.12296 larger than PLSR .7). The PLSR (0.8) successfully predicted 7 of the top 10 finishes as well as 16 of the top 20. With a median R squared value just above 0.25, the variation in the one PLS component only accounts for roughly 25% of the variation in the log of earnings from the tournament. In its prediction of the 2019 Master's it was able to predict finish to within 15 places, with most deviation coming from the average finished values as seen in figure 3. The only major drawbacks for the PLSR with correlation cutoff 0.8 were that its projected winner took 20th in the actual 2019 Master's tournament and it predicted the actual winner of the 2019 Master's to take 9th.

Variable Importance

When looking at the scaled variable importance of the PLSR model with 0.8 correlation cutoff in figure 4, the ability to avoid fairway bunkers proved the most important feature, followed by greens in regulation from locations other than the fairway. Other variables of interest would be back nine scoring average as the back nine holes at Augusta National are considered to be the most volatile on the course, offering high risk and high reward plays. This also follows suit with the back nine par 5 scoring average variable showing high importance.

In general, when looking at the 20 most important variables in a wholistic sense, those who have been able to consistently avoid fairway bunkers and drive the ball efficiently tend to do the best. For the times when players do miss the fairway, it is extremely important that they recover and reach the green in regulation. With only one variable relating to the "short game" of golf, sand saves relative to par, players with strong ball striking ability from the tee and from locations other than the fairway seem to do better than those with strong short games. Finally, those who have been able to consistently perform well on the final nine holes of events, especially the par fives of those final holes, tended to have more success the following week at Augusta National in 2019.

Limitations

Findings and prediction results may be overly optimistic as there is no test dataset. The cumulative golf statistics from the week leading up to the Master's are not readily available for prior years as they are usually updated weekly. Thus, the analysis was not able to use prior tournaments as a testing procedure. Dimension reduction techniques were also a challenge with this dataset. Some variables were dropped on a subjective basis, stating that they related to other variables, but this method of reduction may have led to bias results. There is also variance in golf that cannot be explained by statistics itself. Some players may simply have better days than others and that chance variation plays a huge role in the tournament's outcomes each year.

Finally, although PLSR with a correlation cutoff of 0.8 was the best model in most of the resampling and custom-made metrics, the lack of test set and lack of statistically proven metrics for analyzing top 10 and top 20 accuracy hinders the validity of the findings. In the future, using this model to fit and predict the 2020 Master's will be a way of assuring the validity and findings of this research.

References

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). An introduction to statistical learning : with applications in R. New York :Springer,

Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics (Oxford, England). 2005;21(15):3301–7. Epub 2005/05/21. 10.1093/bioinformatics/bti499 .

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1 - 26. doi:http://dx.doi.org/10.18637/jss.v028.i05

South, Charles & Egros, Edward. (2020). Forecasting college football game outcomes using modern modeling techniques. Journal of Sports Analytics. 6. 1-9. 10.3233/JSA-190314.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Data sources
https://www.pgatour.com/tournaments/masters-tournament/past-results.2019.html

https://www.kaggle.com/bradklassen/pga-tour-20102018-data