Drive for Show, Putt for Dough?

Unveiling Golf's Winning Formula

Aidan Booher, Esha Rao, Breana Valentovish, John Wang

Introduction

In the world of professional golf, the debate over which aspect of the game holds the most significant influence on a player's success has long been a topic of discussion. The age-old adage, "drive for show, putt for dough," encapsulates this debate, suggesting that while driving distance may garner attention for its spectacle, putting ultimately determines a player's success on the leaderboard and, consequently, their earnings. In this report, we delve into this debate using statistical analysis to determine which statistics in golf contribute most significantly to a player's overall earnings.

Problem Statement

There is a growing desire for professional golfers to gain speed and distance off the tee, resulting in less time for training on the green, posing a problem for player's if the old adage truly follows. Therefore, the primary aim of this project is to investigate the relationship between various golf statistics and a player's earnings in golf tournaments on the PGA Tour. We seek to determine whether there is a significant correlation between statistics related to driving distance, accuracy, putting, and other facets of the game and the amount of money won by players.

Importance and Motivation

Understanding the factors that contribute most significantly to a player's earnings in professional golf tournaments interest players, fans, coaches, analysts, and sponsors. Identifying key statistical metrics associated with higher earnings can tailor players' training regimes and on-course strategies accordingly. Coaches can use this information to provide targeted guidance to their players, while analysts can develop more accurate predictive models for player performance. Sponsors can also make more informed decisions about which players to support based on their statistical profiles.

Brief Summary of Results

Our analysis using various regression models revealed insights into the relationship between specific golf statistics and a player's finishing position in tournaments. More specifically, it appears driving distance and average birdies made have significant influence on earnings, even more so than putting.

The subsequent sections of this report will delve deeper into the methodology employed, the results obtained, and the implications of our findings in the professional golfing world. Through this analysis, we hope to shed light on the age-old debate surrounding the importance of driving versus putting in determining a player's success in the earnings column.

Data

Our analysis relies on two distinct datasets to explore the factors influencing golf performance and earnings. Dataset One is made up of varying performance metrics for professional golfers and their season earnings. Dataset Two offers a broader perspective by integrating player profiles, weather data, and golf course information spanning multiple seasons. These datasets provide a foundation for our analysis, allowing us to delve into the intricacies of professional golf and uncover valuable insights into the drivers of success on the PGA Tour.

An important aspect of each dataset is their use of “Strokes Gained” statistics. This variable compares a player's individual performance on each shot to a baseline of their competitors. Positive strokes gained means a player performed better than the average or baseline for that type of shot. This can be used to see how well a player drives, putts or hits approach shots compared to the field in each individual tournament.

PGA Tour Top 200 Player Data

This dataset consists of 1674 rows and 18 columns, each representing a golfer's performance for a given year. The dataset includes various performance metrics such as fairway percentage, average distance of tee-shots, greens in regulation (GIR), average number of putts, scrambling rate, average score, number of wins, strokes gained statistics, and prize money earned.

Source: The dataset was obtained from Kaggle, a popular data science company.

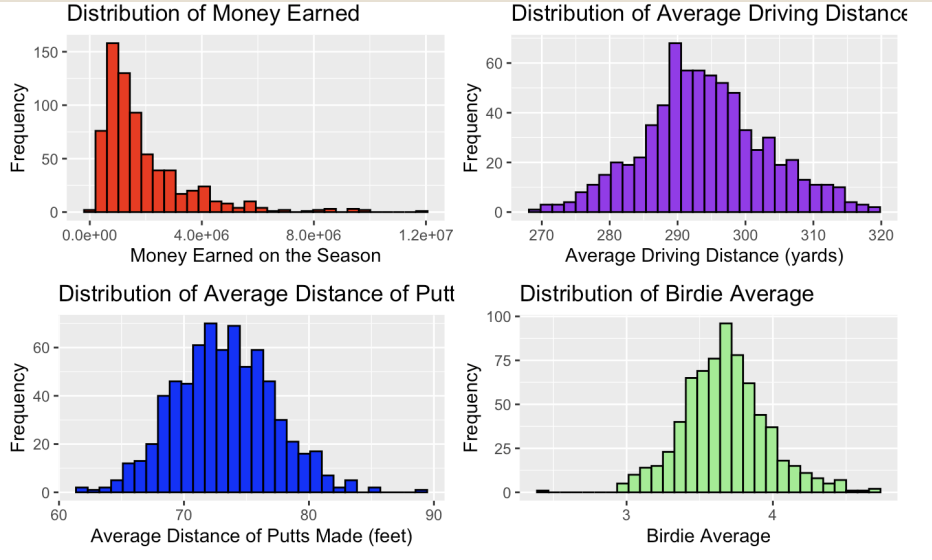
PGA Tour Dataset

The PGA Tour Results dataset spans the 2018 to 2021 seasons and incorporates player profile data, weather data, and golf course data. It contains all of the strokes gained statistics for each individual tournament, as well as tournament results, course characteristics, and weather conditions. It also contains a binary variable “Major” detailing whether or not the tournament was a major championship. We used this variable to create a subset with data on just golf’s most prestigious events, as they are also the highest paying events. Finally, other pre-processing steps included filtering the dataset to include just players who made the cut and adding in a new column called “Earnings” where we added the Players’ winnings from these individual events.

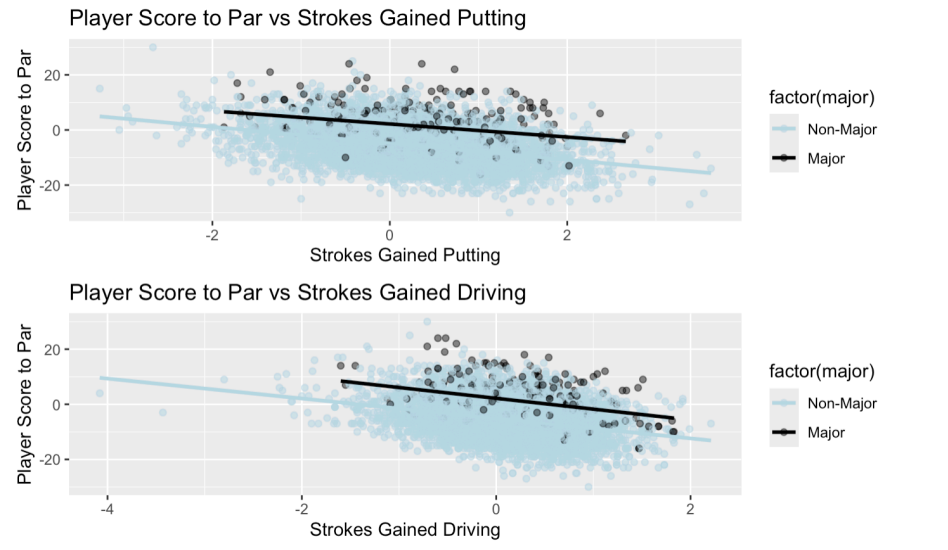
Source: The dataset was sourced from Zenodo, a reputable repository for research datasets.

EDA

I. Histograms of Key Variables



II. Scatterplot of Key SG Statistics



Notes:

From our univariate analysis, we see that 'money earned' is skewed to the right with a peak at an x value of roughly 1.0e+06. The variable 'average driving distance' looks like a normal distribution, as it is unimodal with no skew and centered around an x value of 295 with a peak around an x value of 290. The variable 'average distance of putts made' also looks relatively normal as it is centered with a slight right skew and has a peak at an x value of roughly 73. Finally, the variable 'birdie average', has its center and peak at an x value of roughly 3.8. This distribution has a left skew, however, due to the outlier at an x value of 1. Finally, we can see in the scatterplots that both SG putting and driving seem to have a negative relationship on a player's score to par.

Methods

In this section, we outline the statistical modeling techniques employed to address the problem of identifying key predictors of earnings and success in golf. We selected a variety of modeling techniques to capture different aspects of the relationship between golf performance metrics and earnings.

Linear Regression:

Linear regression was chosen as a baseline model to assess the linear relationship between various golf performance metrics and earnings. Linear regression allows us to quantify the association between each predictor variable and earnings, providing straightforward interpretations of coefficients.

Generalized Additive Models (GAM):

GAMs were employed to capture potential nonlinear relationships between predictors and earnings, while controlling for potential confounding factors. GAMs relax the linearity assumption of traditional linear regression by allowing for smooth, nonlinear relationships between predictors and the response variable, enabling us to capture potential nonlinear effects of golf performance metrics on earnings.

Lasso Regression:

Lasso regression was utilized to perform variable selection and regularization, aiming to identify the most influential predictors of earnings while mitigating the risk of overfitting, assuming only a subset of predictors has a nonzero coefficient. It automatically selects relevant predictors and penalizes less important ones, providing a model that highlights whether driving or putting, if either, has a more significant impact on earnings.

Comparison and Evaluation Approach:

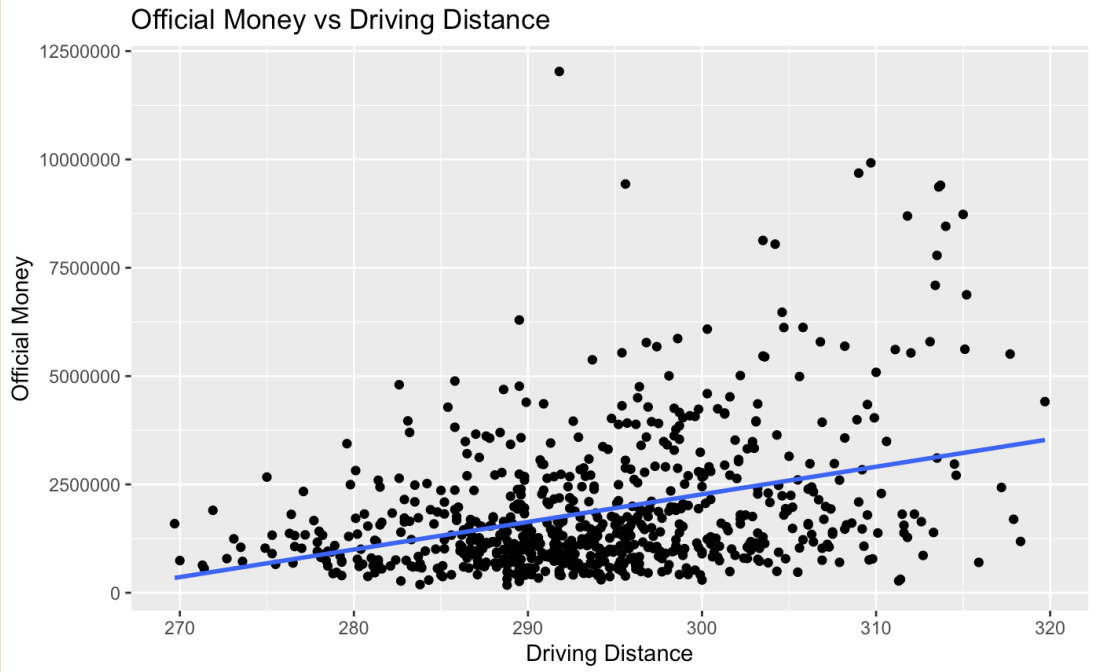
We plan to compare the performance of the different models using the root mean squared error (RMSE) metric. RMSE provides error measurements in the same units as the target variable (Dollars), so we can interpret the model's performance directly in terms of earnings. Golfer earnings can vary significantly with potential outliers. RMSE, being sensitive to outliers, would appropriately penalize models that make large errors on these extreme cases. Uncertainty in our estimates will be quantified through bootstrapping, which involves resampling from the dataset to obtain a distribution of parameter estimates.

Results

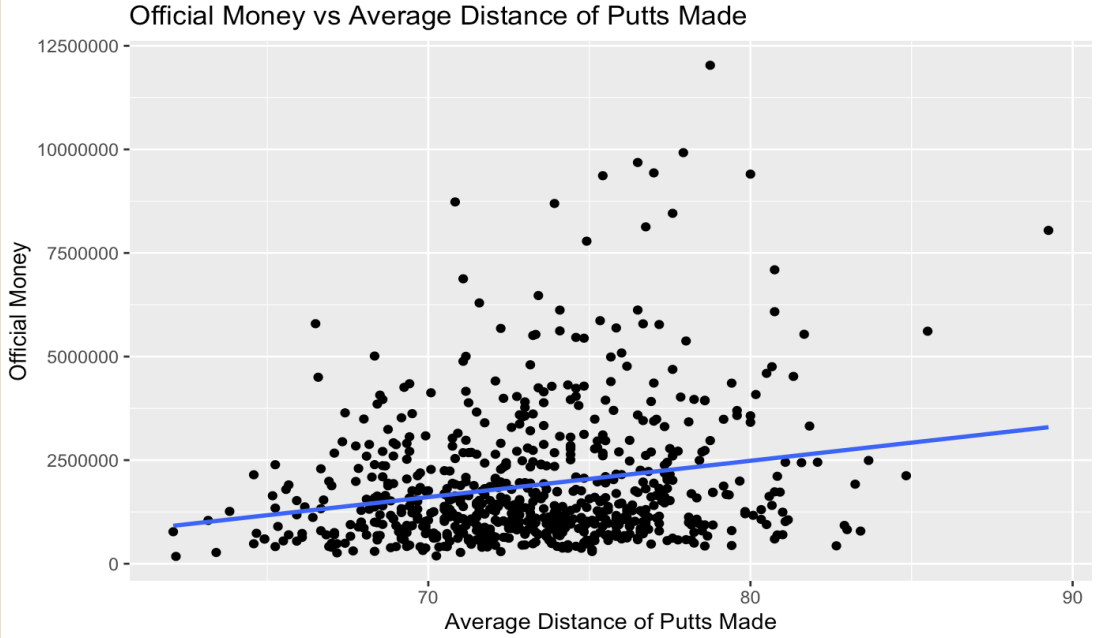
In this section, we present the results of our analysis from each model and their respective interpretations.

Linear Regression (LM):

III. Scatter Plot of Earnings vs Average Driving Distance



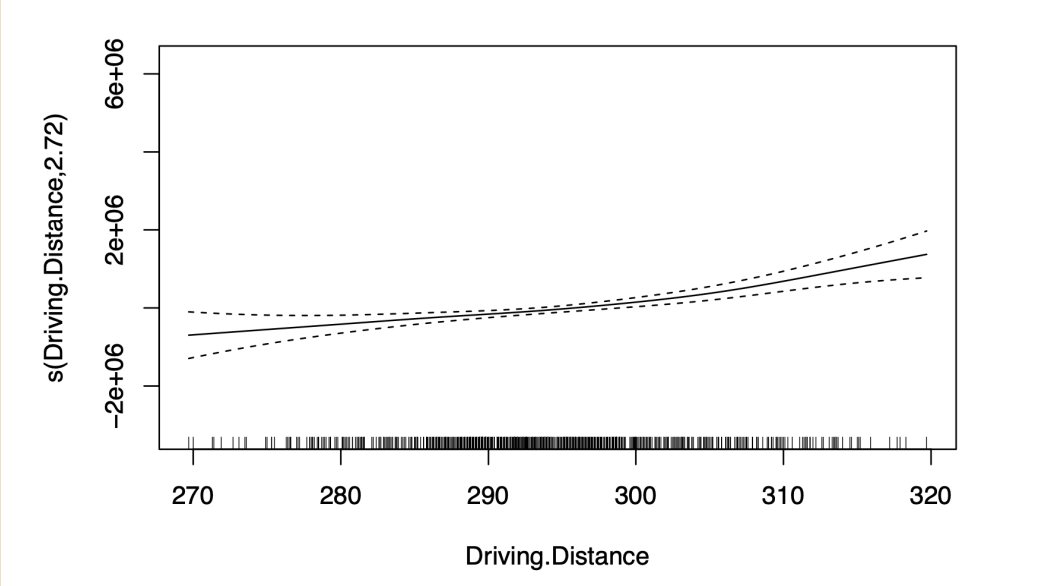
IV. Scatter Plot of Earnings vs Average Distance of Putts Made



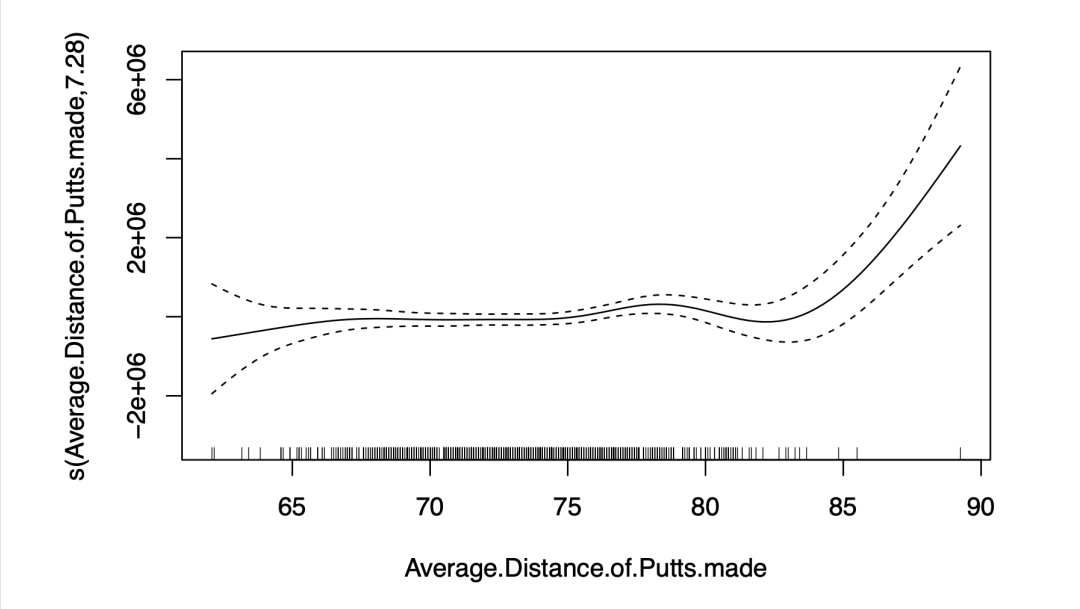
The linear regression model suggests that driving distance, birdie average, and bogey average significantly influence earnings in professional golf. Specifically, for every additional yard in driving distance, a player's earnings are estimated to increase by approximately $37,880. While driving distance appears significantly related to earnings, driving accuracy is not, indicating that being able to have shorter distances and clubs into approach shots matters more than being in the fairway. We believe this is because professionals are so good out of the rough that having a scoring club in their hands from any surface is beneficial. Additionally, as we can see in our first scatterplot (III), driving distance has a fairly strong positive linear relationship with official money earned. In comparison, the second scatterplot (IV) illustrates that while there does appear to be a positive relationship between the average distance of putts made and official money earned, it is not as strongly linearly correlated. This is supported by the fact that it was not deemed statistically significant in the linear regression model. Using bootstrapped resampling, we calculated the 95% confidence intervals. The bounds of the average distance of putts made coefficient is [-4369.01, 49891.00] and driving distance is [22087.49, 53218.17]. Since the putts confidence interval includes zero, it suggests that the coefficient may not be significantly different from zero, or in other words, there is a possibility that the predictor variable has no effect on the response variable. We will further test this statistic using a GAM model.

Generalized Additive Model (GAM):

V. Partial Response Function of Money on Average Driving Distance



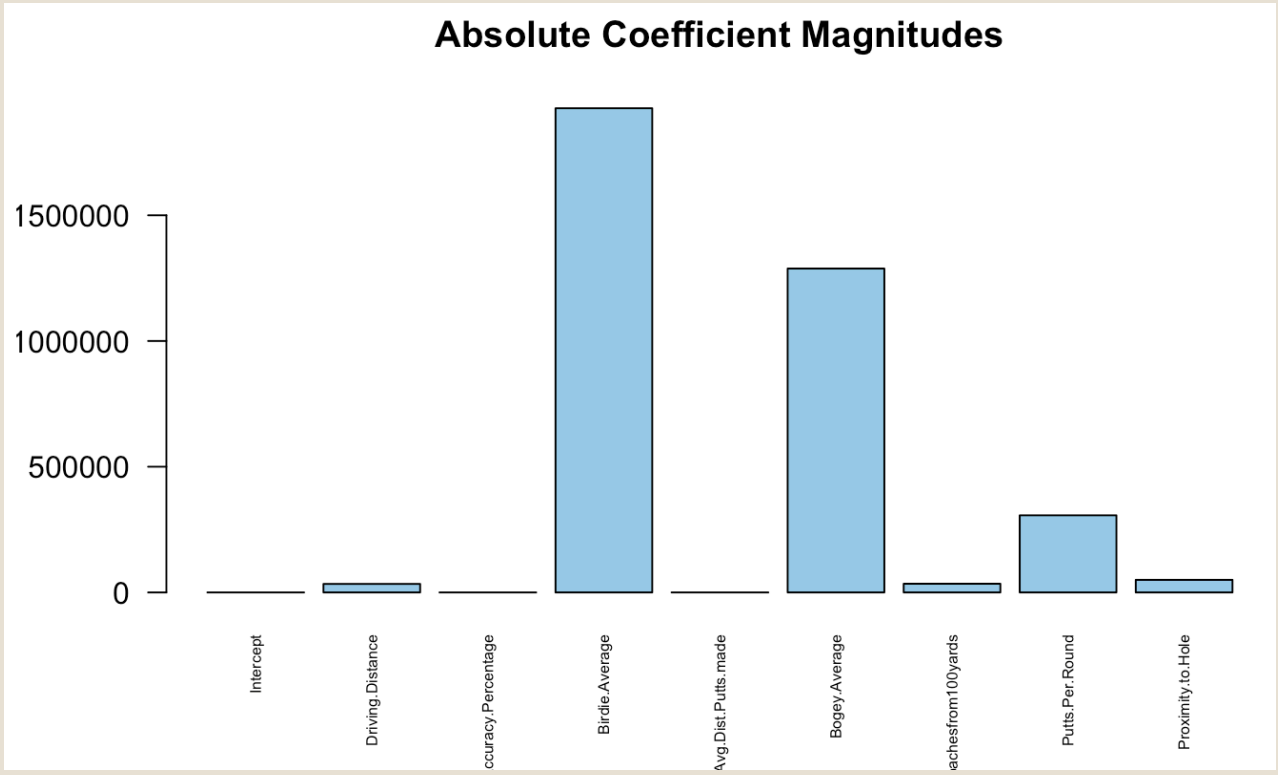
VI. Partial Response Function of Money on Average Distance of Putts Made



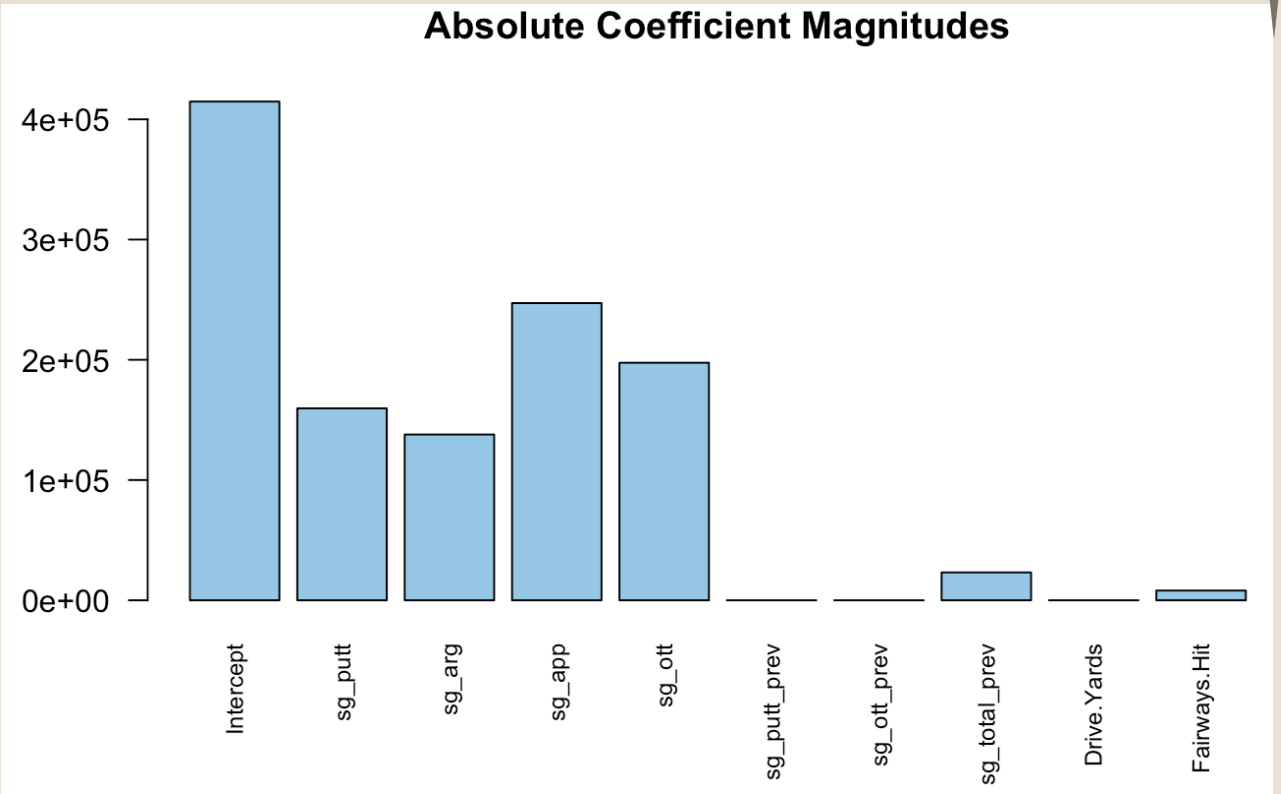
The GAM reveals nonlinear relationships between certain golf performance metrics and earnings. In this model, it does indicate that both driving distance and average distance of putts made, as well as average birdies per round, have statistically significant relationships with official money earned. The partial response plot for driving distance in figure V suggests a consistent positive association with earnings, indicating that greater distances correlate with higher earnings. Conversely, the partial response plot for average distance of putts made shows there isn't as obvious of a positive relationship, as the earnings stays relatively similar until about 83 feet of putts made. There is a large spike in earnings observed once golfers do get past this mark of 83 feet of putts made, however only .9% of golfers in our dataset reach this amount. Therefore, we conclude that increases in average distance of putts made does not necessarily lead to increased success in tournaments unless you reach the top tier of putters, whereas any increase in driving distance correlates directly with increases in success. The other significant plot was the average birdies plot, which also showed an increase in earnings as the average birdies increased, especially once you get past four birdies per round, where we see a large spike in earnings. We also used bootstrapped 95% CI's that did not reveal anything significant in quantifying our uncertainties.

Lasso Regression:

VII. Extracted coefficients from first Lasso model



VIII. Extracted coefficients from second Lasso model

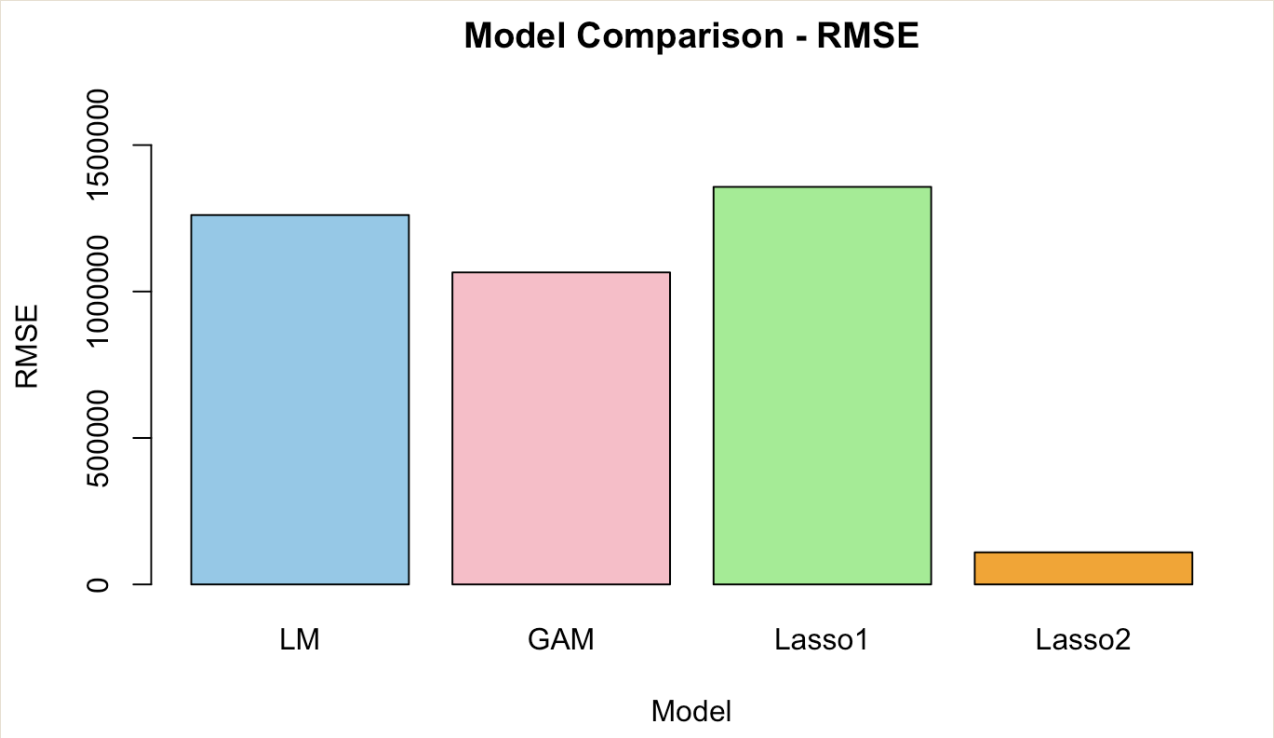


The first lasso regression identifies significant predictors of earnings, with driving distance and birdie average among the influential variables. Average distance of putts made is among the variables that have coefficients shrunk to zero, indicating their negligible impact on earnings. The extracted coefficients in figure V reveal which of the golfing metrics were identified as significant predictors. It should be noted, however, that just because some of the coefficients appear much greater than others, this does not mean that we can say they have greater effects on earnings. For example, average number of birdies is a much smaller number than average driving distance, so inevitably, an increase in one unit of birdies made will have a much larger effect on earnings than one yard of driving distance. Therefore, we must use a standardized metric if we want to compare exact effects on earnings. So, for our second lasso, we performed a lasso regression using strokes gained statistics as the predictors. In an attempt to further standardize this regression, we filtered out data down to include just the strokes gained statistics from Majors, as these are golf's most prestigious and highest paying events. As we can see in figure VIII, in the lasso regression focused on major tournaments, strokes gained metrics such as putting (sg_putt), approach shots (sg_app), and off-the-tee performance (sg_ott) emerged as significant predictors of earnings. While both driving and putting appear to be significant predictors here, we can see that our model predicts **$37,916.80** more per stroke gained for driving than putting, which again shows there is evidence that driving may actually be more influential on money made over the course of a season. Finally, we used bootstrapped standard errors on both models, which analyze the variability of coefficients across the bootstrapped samples. Since the standard errors were relatively small in the first lasso model, we have fairly high confidence in the coefficient estimates. The standard errors were very small in the second lasso, indicating great precision and low uncertainty.

Model Comparison and Evaluation:

Now that we have discussed the results of each model and how we interpret them in terms of our research question, we want to compare the RMSE of each model to analyze the performance of each model. Although we saw similar results from each model, understanding which model performs the best will allow us to further conclude which results, if any, we can be most confident in.

IX. Comparing RMSE's of Each Model



As seen in figure IX, across all models, our second Lasso regression model has a significantly lower RMSE than the other models. This could be due to the fact that since the second Lasso regression used data only from the Majors, there might be less variability compared to the other three models, since they all used data from multiple tournaments including lesser known regular season events. Of the first three models (which use the same data and variables), our GAM model has the lowest RMSE. Since a lower RMSE is synonymous with a model that fits the data well and has precise predictive performance, we conclude that the GAM model performs the best of those three.

Discussion

Our analysis provides valuable insights into the relationship between various golf statistics and professional golfers' earnings on the PGA Tour. Across all models, driving distance and birdie average consistently emerge as significant predictors of earnings, highlighting their importance in professional golf. We can interpret the driving distance and number of birdies made being significant as meaning playing a more aggressive golfing style pays off. Contrary to the popular adage, our analysis indicates that putting might not have quite as significant of an impact on earnings than it once did. Our second Lasso model has the best performance metrics, though, so it should be viewed more confidently. It reveals putting still is one of the most significant metrics, but also shows it is less impactful than driving distance.

It is important to acknowledge the limitations of our analyses. One limitation of our analysis is the reliance on aggregated tournament data, which may not account for the fact that a player could make a ton of money in one tournament because of a certain statistic that they perform poorly in the rest of the year. For example, having one good putting week could win a player a major, but it would not be correlated with high earnings if they struggle putting during the rest of the season. Moreover, while our models provide insights into statistical associations, they may not fully capture the complexity of player performance dynamics on the course.

Finally, future research could focus on incorporating additional data sources, such as player demographics and tournament conditions, to further refine the analysis. Exploring more advanced modeling techniques, such as machine learning algorithms, could also provide deeper insights into the nuances of player performance and earnings. Additionally, conducting longitudinal studies to track player performance over time could explain trends and patterns in professional golf.

In conclusion, while improving at any golf statistic will improve a golfer’s chances at success on the PGA tour, each of our models reveals a similar pattern that driving distance and birdie average are especially important. This has major implications for players and coaches when forming a training program or on-course strategy.