PGA Earnings Predictions

A Linear Regression and Web Scraping Project

Chris Byrnes
April 16, 2021

Challenge

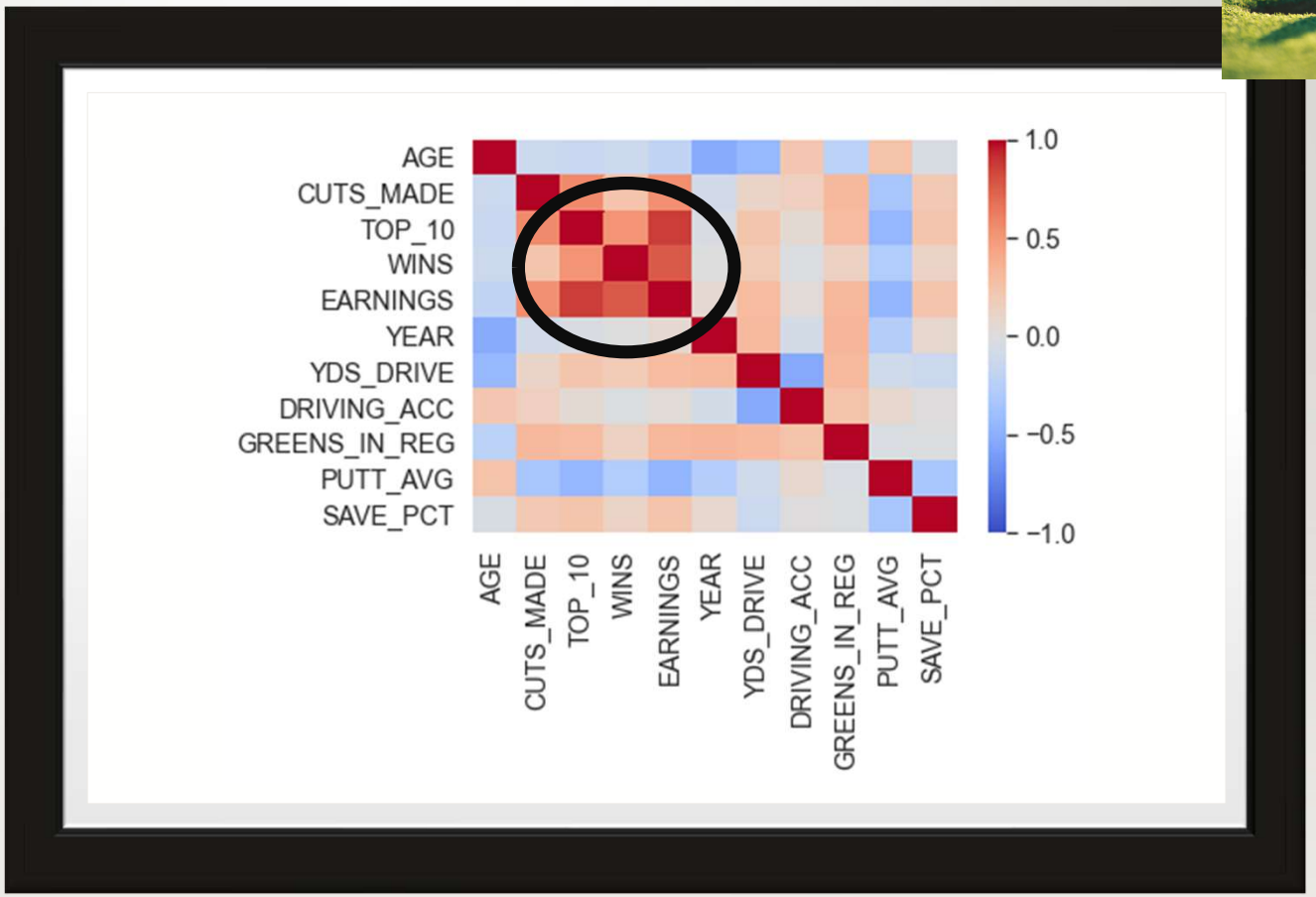
- Each year the PGA tour awards millions of dollars in prize money (over $400M in 2020)
- Telecasts are filled with statistics such as players' driving distance, greens in regulation, etc.
- **How closely can we predict a tour professional's annual earnings using linear regression based on their performance expressed through these statistics?**
- To make it a little more difficult we won't use highly correlated earnings features like tournament wins and top 10 finishes

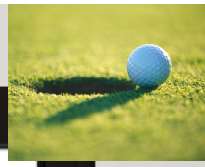
Data

- 22 data fields were scraped from ESPN.com using Beautiful Soup
- Years considered were 2010 - 2020
- Final data frame included 1,670 individual records
- Fields used for initial linear regression:

  - Age
  - Year
  - Cuts made
  - Driving accuracy
  - Driving distance
  - Greens in regulation
  - Putts per green
  - Save percent
  - Earnings

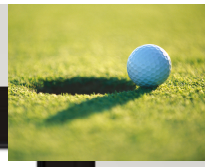
Correlation of Variables





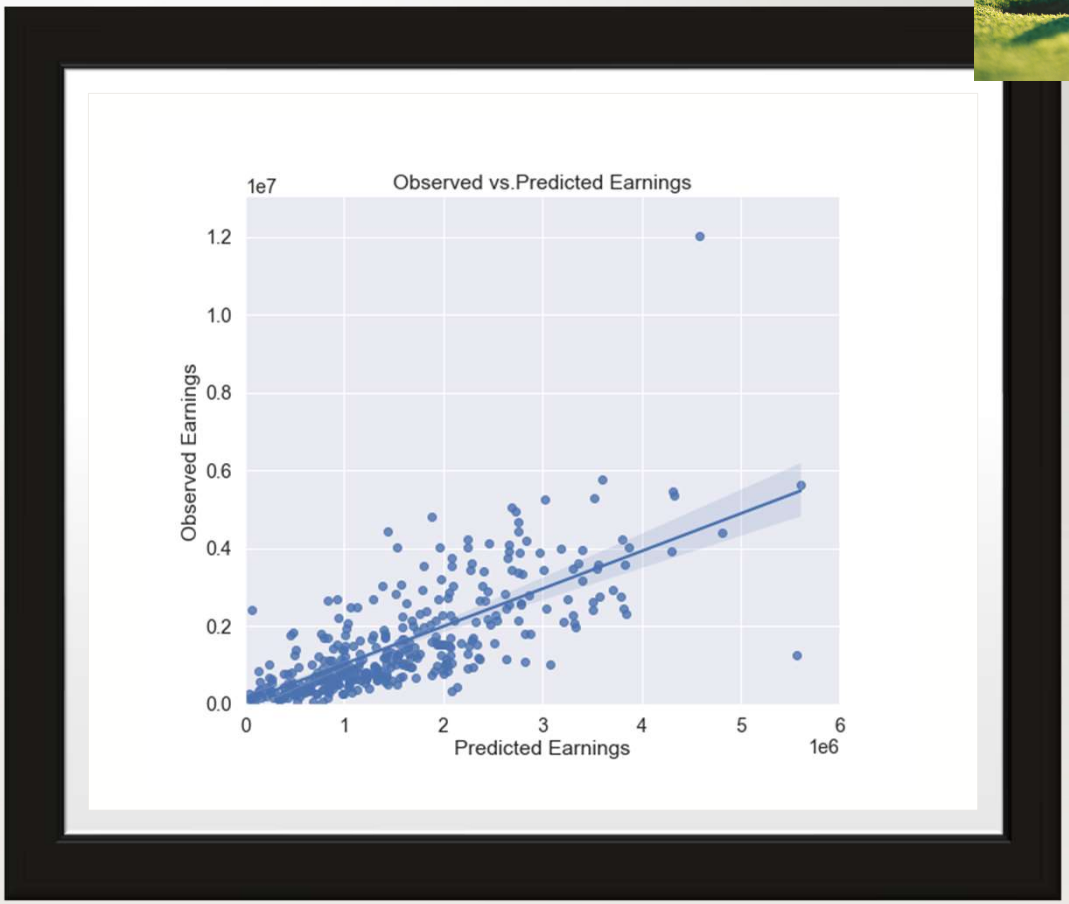
Simple Linear Regression

- Graph based on test data
- Initial results using basic linear regression include R^2 of .499 on the training data and .456 on the test data
- Suggests over fitting
- Visually can see
  - Model struggling to deal with higher earning outliers
  - Creating more error with high earners
  - Including Wins or Top 10's would greatly eliminate issue

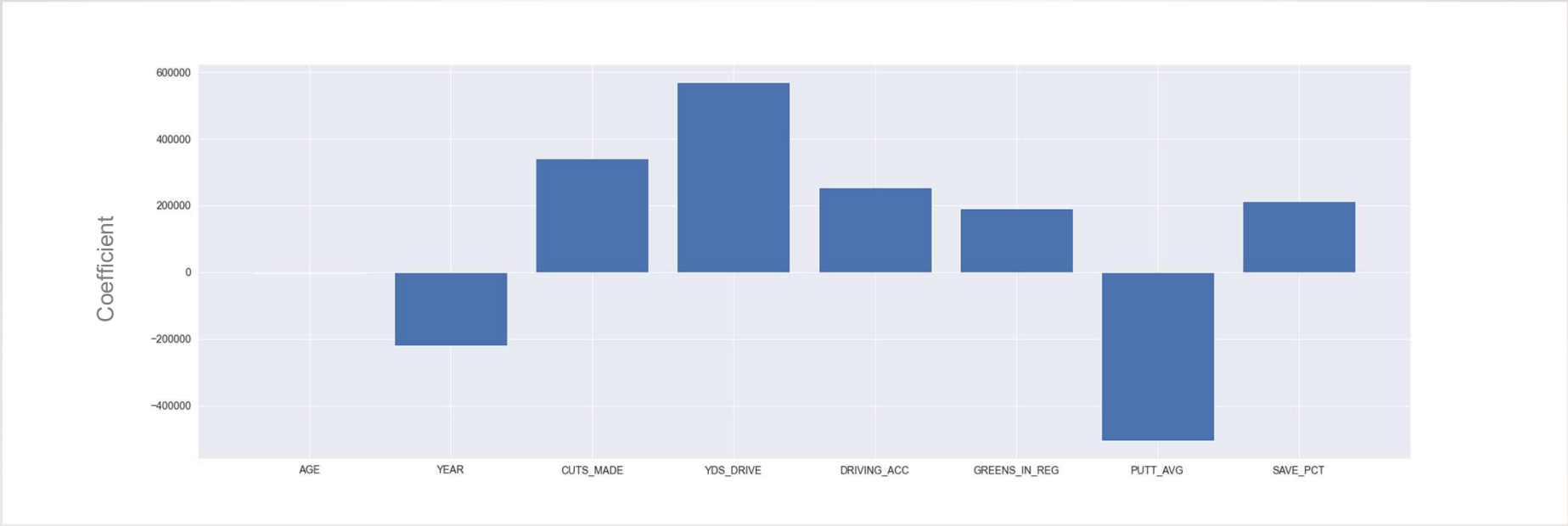


Process and Results

- Additionally worked with LASSO, Ridge, and Polynomial models
- Ultimately chose LASSO on 2nd degree polynomial features
- Was able to raise R^2 to 0.540 (from original 0.456)
- Models remains a bit overfit and would benefit from more data and addition data features to work with



Feature Strength



Appendix – Additional comments on data used

- Full data was not available for 2016 and information from this year was excluded from the analysis
- Age for 45 golfers was not available in the scraped data. The average age of 39 was used