Dyson Thacker

Predictive Analytics II

06 December 2020

PGA Tour Predictive Analytics

This project started in the beginning of the semester when Dr. Brandenburger accused me of being “copy-paste guy” because I was copy and pasting data from the PGA TOUR website into excel, and then loading it into excel. It took forever. He said I could make a script in R that would do it way faster. Just to prove that I could, I spent most of a Saturday writing an R script that scraped the PGA Tour website for statistics on golfers. It didn’t exactly save me any time, but it was really cool to know that I had the capability of doing it.

For the end of the semester project, I took that dataset and did my best to predict what attributes the best golfers on the PGA TOUR carried.

The dataset has 3221 observations and 17 columns. Each observation is one persons cumulative end of the year statistics for that respective year. The data ranges from 2004 to 2020. I created two flag variables. One called “good\_golfer” that signifies if a golfer was in the top ten percent of money earners in their respective year. The other called “good\_putter” that signifies if a golfer was in the top ten percent of Strokes Gained putting in their respective year.

I decided I wanted to make a set of models that predicted “good\_golfer” and a different set that predicted “good putter”. Each set of models has three different models; Logistic, Decision Tree, and Extreme Gradient (XG) Boost.

Below are the ROC results for my models predicting “good\_golfer”

Chart, scatter chart

Description automatically generated Chart

Description automatically generated Chart, line chart

Description automatically generated

The XG Boost model didn’t perform as well as I was hoping, but the logistic and the decision tree did.

Below are the ROC results predicting “good\_putter”. Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Similar to before, XG Boost didn’t perform super well, but logistic and decision tree did.

Below are some cool summary plots that I made with the dataset.

Table

Description automatically generated

Table

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

In conclusion, it was cool to build some models with a dataset that I made completely by myself. The world of XG Boost is huge, and I definitely need to learn more about it in order to make some better predictions.