COS424: FINAL PROJECT PROPOSAL

ROB WHITAKER (RMW2@PRINCETON.EDU) AND ERIC MITCHELL (EAM6@PRINCETON.EDU)

Summary. Our final project will use statistics available on the PGA Tour website to analyze the characteristics of the best professional golfers in the world. The PGA Tour keeps extremely detailed statistics about every player, course, and tournament throughout each season. In order to minimize the amount of work we will need to to acquiring our data, we expect to be able to use a pre-existing Python library which is built for scraping pgatour.com. We have two primary goals for this project: First, to determine the primary characteristics that determine how successful a player is during the course of the season. Second, we would like to be able to predict, for a given player, which courses (tournaments) are the best for him or her to play during the course of a season in order to maximize their success.

Statistics in Golf. Golf is game that is perhaps unique in how many statistics are available for a particular player. The game is so easily stratified into categories, that an analyst can easily examine only a very specific part of someones performance. Most simply, a players game is separated into the long game (long shots, generally longer than 100 yards) and the short game (inside 100 yards and putting). A players performance can be further assessed in terms of either power or accuracy, which generally (but not always) are inversely related (players with more power are generally less accurate). One of the hardest parts about the game of golf is understanding what one needs to do in order to improve; professional players spend very long amounts of time analyzing their own performance in order to determine what they need to refine in order to shoot better scores.

We hope to be able to automate this process by accomplishing the first goal stated in the summary. By analyzing the entirety of the tour, we can decide on which statistics a) are the greatest determinants of success on a year-to-year basis and b) which statistics produce the greatest change in success when they are improved or get worse.

The second stated goal will involve using much more in-depth data than the first, so it is possible that the scope of a three-week project will not be sufficient to complete it. However, if we are able to, it would be fabulously useful. For players who are not near the top of the earnings list, each year on the PGA Tour is a constant struggle to earn enough money to be able to come back the next year. Therefore, for these players, optimizing their schedules is vitally important. Currently, players do this by instinct, by talking with their coaches and going with a gut instinct. However, we believe that we can do better.

Data. On the largest professional tours in the world, extensive statistics are kept on every player that ever competes during the course of the year. More importantly this data is

available to the public. A Python library for scraping pgatour.com exists, and we hope that it will make the process of organizing our data. The most well-defined model we will use is on a very small scale: given a player with a set of statistics, and a hole with a set of features (distance to the green, curvature, sand traps, etc.), how can we expect the player to score on that hole. This is a complicated regression task, made more difficult by the fact that there is intuitively quite a bit of co-dependence between the course and player features. We plan to experiment with different feature kernels to model this dependence, and different regression models: comparing our success with linear approaches like Ridge or Lasso, and non-linear approaches, such as a gaussian proccess. Additionally, we want to make the project fairly interactive, creating an API (or even, time permitting, a UI) that can be used to visualize and perform additional exploration of the data. Unsupervised approaches can be used to classify players or courses into different groups or styles, while the main regression task can be used to predict the outcome of any hypothetical tournament given a course and a set of players involved.

Evaulation and Interpretation. Finally, all this facilitates evaluation of our models, as historical contests can be withheld from training and real results can be compared with our modeled output. We can even quantify error down to the mean square deviation at each hole. The most variation will probably result from changing the way we quantify or combine the features for the courses, and an interpretable model may require that we make some intuitive choices for which features ought to relate to each other. Training a model with many features and many observations may be time consuming, but it's likely we can perform the training and optimization in a separate step, allowing an interactive prediction engine to use a pre-fit model and return results in reasonable times. With a well-interpretable model, we can draw conclusions about which aspects of the game are the most volatile scoring wise, and how much variation there is across players on different types of shots. Are there certain areas where all the pros have identically stellar performance? Which skills are the most valuable for a given course? What types of holes would be required in order for one player to beat another? Is there a course that would allow it to happen?