CSE 482: Project Intermediate Report

Tyler Rozwadowski | CSE 482 | 3/24/19

For my project, I'm continuing with my idea to do data analysis on PGA Tour golf statistics. After meeting with the instructor, I realized that the data set I originally planned on using from Kaggle did not represent enough years or seasons of data to make accurate predictions or show correlations between certain attributes. After searching around the internet, I realized I couldn't find the downloadable data that I was looking for. Therefore, I started a new data collection process.

I wanted to use data from the official PGA Tour website, so I tried to see if it was possible to download a CSV file of the dataset. However, that feature is only available to corporate partners. Because of this I began writing a web-scraper to get the information from their website. This web crawler used some of the techniques shown in class earlier this year, along with other topics I've learned from the internet and a previous internship. I eventually had a python web crawler that created the data set that I'm going to be using for my analysis project. This web crawler grabs many different statistics from the PGA Tour website. For my dataset, I tried to grab the attributes that I thought would be most helpful or interesting when analysis or predictions. In fact, I am still deciding whether it would be good to include more attributes than the ones I already have.

Once I had scraped the data, I cleaned it by removing unnecessary attributes by deleting columns that I didn't need. It was also important to ensure that each attribute is the proper type (ie. int). I have also begun the process of determining how to impute values that are missing or don't exist. Sometimes it's as easy as just filling in zero, other times it may make sense to compute the average and fill that in.

The goals of my project have changed to now find correlations between different attributes. For example, if a player hits the ball much farther on average, I would like to know if that correlates with higher money earnings. There are many other statistics that may have correlations that I would like to analyze. However, I would like to further discuss with the instructor as to meaningful things that I can find using this set of data. It would also be feasible to do attribute predictions now that I have many seasons worth of data. The biggest challenges I'll face are determining how to use the data that I now have.

For now, my working project title is "The Most Important Statistics on the PGA Tour." Because most of my analysis will be based on finding correlations, I can find how one statistic may be more "important" to golf tournament success.

I plan on meeting with the instructor to discuss where to go with my data analysis or predictions. In the meantime, I will continue to preprocess the data. This should all be completed by the end of March. I will then be able to spend all of April doing the analysis and creating the final report.