The US Open tennis tournament is fast approaching. Starting August 28th, and continuing for 2 weeks, the top tennis players in the world will come to New York City to compete on the hard courts for the final Grand Slam tournament of the year. The US Open is the highest paid tennis event in the world, with total prizes totaling $50,400,000. It is also one of the only tournaments to evenly pay both the men and women champions. Because of the higher stakes, match predictability in Grand Slams are higher than other tournaments. I will predict the outcome of each match, and ultimately the winner of this year’s US Open. The data model is intended for recreation, and not professional gambling situations. I will personally use this model for bragging rights to family and friends.

To build the model I will use the following features:

* ATP tour points
* World Rank
* Tournament Seed
* Surface type
* Player height
* Player age
* Dominant hand of player
* Nationality of player

During data wrangling I learned a good lesson on not using generic titles for headers. After I had loaded all the data and began exploring with basic statistics, I found a problem with the field ‘Round’ that wouldn’t allow me to use it. I read through the code for a couple days looking for spelling and syntax errors before a conversation with my mentor gave me the answer. It turns out ‘Round’ is a reserved name in Python and the compiler would error on calculations with this field.

In analysis of the model I found that many of the features did not have an affect, or a negative effect, on the predictive result. By eliminating 1 feature at a time, and comparing the outputs, I found the optimal model with four features – ‘age\_delta’, ‘seed\_delta’,’hand\_delta’,’min\_age’. Surprisingly, the ATP tour points (both career points, and current points) were negatively correlated. I am not sure why this happened. Of all the features, tournament seed difference had the highest correlation to the match winner. This was expected since the higher seed was the favorite to win the match.

Though the model was intended for recreational use, at roughly 65% accuracy, the model is not remarkably successful. I think bringing in more match statistics of the players (i.e. head to head results, number of aces, number of winners, etc.) would raise the quality of the model. It would also be interesting to use weather conditions.