Matthew Zlotnik

Mcz334

Chen T/TH 2:00-3:30

12/3/18

Final Project: Tennis Data

**INTRODUCTION**

As machine learning and data science methods have expanded, so too have the fields of their applications. Initially used for business insights, machine learning has rapidly expanded into the field of sports. In attempts to predict erratic human behavior, professional teams and solo athletes alike have employed the help of trained data scientists to give them a competitive edge. While used most effectively in team sports such as baseball and basketball, machine learning has recently been introduced into tennis to allow athletes to focus their training on specific points of person weakness, or to focus their in-game strategy on exploiting the weaknesses of their opponent. This project will focus on two main aspects of machine learning as it relates to tennis. Primarily, what specific measurable statistics (net points, break points, first serve accuracy, etc.) should athletes focus on maximizing in general, regardless of opponent. Then, I will narrow the dataset to look into the weaknesses of specific recognizable names to see if an underdog facing one of the sport’s great talents can focus on one aspect of his game to pull off an improbable upset.

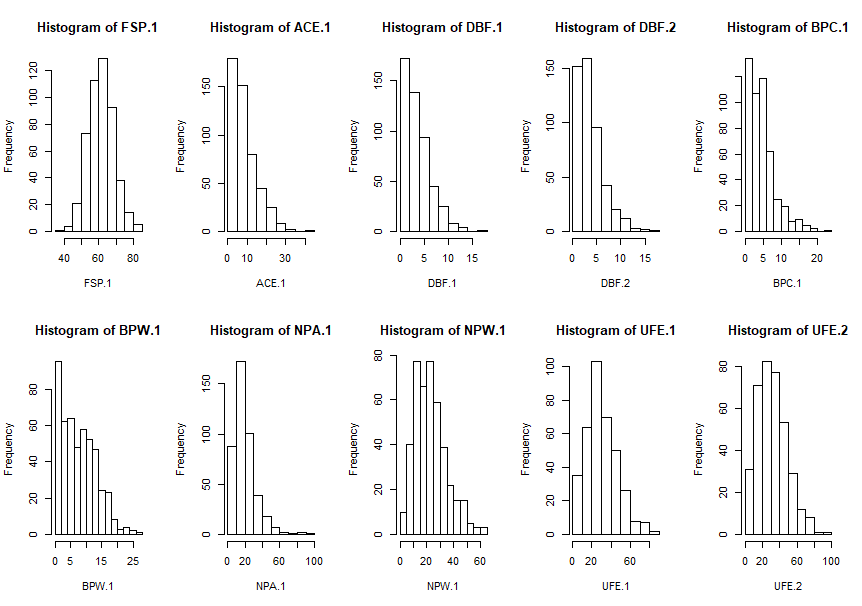
**DATA SOURCE AND CLEANING**

My project is based on the “Tennis Major Tournament Match Statistics” data set from the UCI Machine Learning Repository. The data frame includes 42 different attributes consisting of the players’ names, several vital match statistics, and the results of every set, round, and match of four different events in 2013. For my analysis, I limited the data to only men’s matches, as the underlying success factors in men’s and women’s tennis could be drastically different. The data set includes a wide array of different match statistics, including break points forced by each player, net points won by each player, etc. The data came split into four separate data frames, each containing one of the events from that year. The first step in my data cleaning and organizing was to compile these four datasets into one frame. This proved to be simple as each csv had the same column headers in the same order, allowing for easy concatenation. Next, to speed up analysis, I removed a few columns that I deemed unusable for analysis, such as points scored per round, as these would surely confound the result. Finally, as the response variable (“Result” in the dataset) exists relative to player 1’s perspective (That is, the variable is 1 if player 1 is victorious and 0 if player 2 wins the match), I decided to primarily use player 1’s statistics in my analysis.

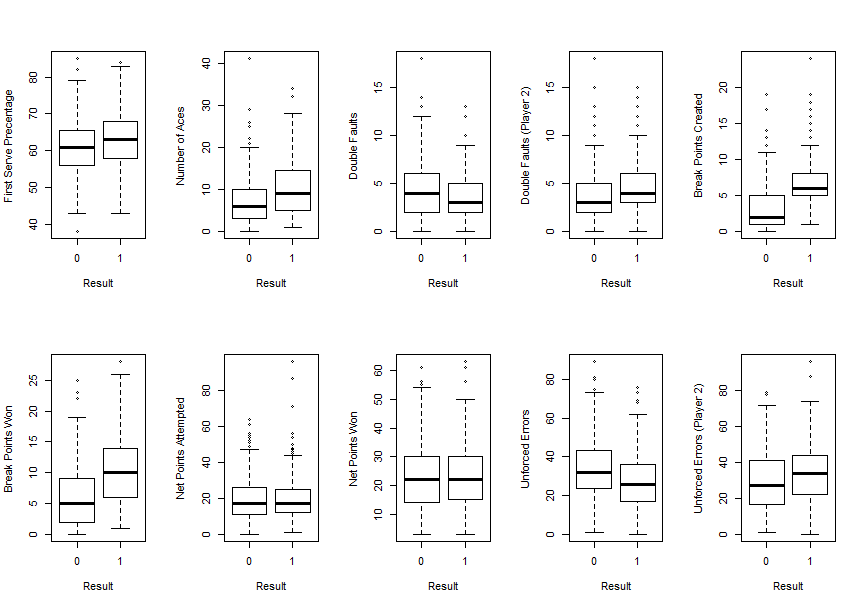
**METHODOLOGY**

As the response variable in this analysis is binary, the statistical learning methods will primarily be classification models. Because of the nature of tennis statistics, it can be reasonably assumed that most of the predictor variables used in this analysis are independent of each other. To ensure a thorough and complete analysis, I tried almost all classification models made available during the course of this class. To choose the most effective model, however, required an intensive process of comparing test error rates, analysis of variances, and interpretability of each model. For the sake of this analysis, most “black box” models ultimately were scrapped, as even though they provided low test error rates, they did not truly answer the problem of which statistics to focus on when trying to win a tennis match. To further this, only models that could be interpretable for a typical tennis player were included, as simply generating a prediction for the winner of a game given all of the stats from the match would be an entirely useless analysis from the start. As such, only logistic regression and decision trees were included, as all other models were determined to be too abstract in interpretation.

Before I could begin with model fitting, however, some exploratory data analysis had to be done. I selected ten predictor variables that I thought could possibly predict the outcome of a tennis match, opting for percentage-based statistics (like first serve percentage) rather than strictly additive integer-based statistics (like first serves won) when possible, as the latter could have a tendency to predict how long the match lasted rather than what factors led to one athlete’s victory in the match. The first bit of EDA I conducted was to construct simple histograms of the predictors I was interested. All of the histograms produced are shown below. Unsurprisingly, most predictors measured in percentages are relatively normally distributed, while most predictors measured additively tend to be right-skewed. This is to be expected, as some tennis matches can take upwards of 4-5 hours, and thus will include more occurrences of almost any statistic than an average match. As such, I do not believe that these skews will invalidate my analysis.

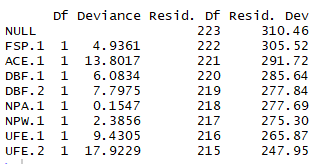


Next, I created bar-and-whisker graphs comparing each predictor’s effects on the response variable in a vacuum. Shown below are the results of these simple boxplot calls. Most of the results were to be expected, such as the implication that the player with more unforced errors tends to lose the match, and that the player with less double faults in the match tends to win. Some aspects of this chart were slightly more intriguing and will be looked further into, such as the relationship between break points and match victory. A cursory look at the differences between the break points created chart and the break points won chart may lead to the conclusion that it is more important to simply create the break points than it is to win them. As these two predictors are clearly correlated and may not tell the full story, I created a new predictor, BPP.1 and BPP.2 (Break point percentage for player 1 and 2) to see if capitalizing on

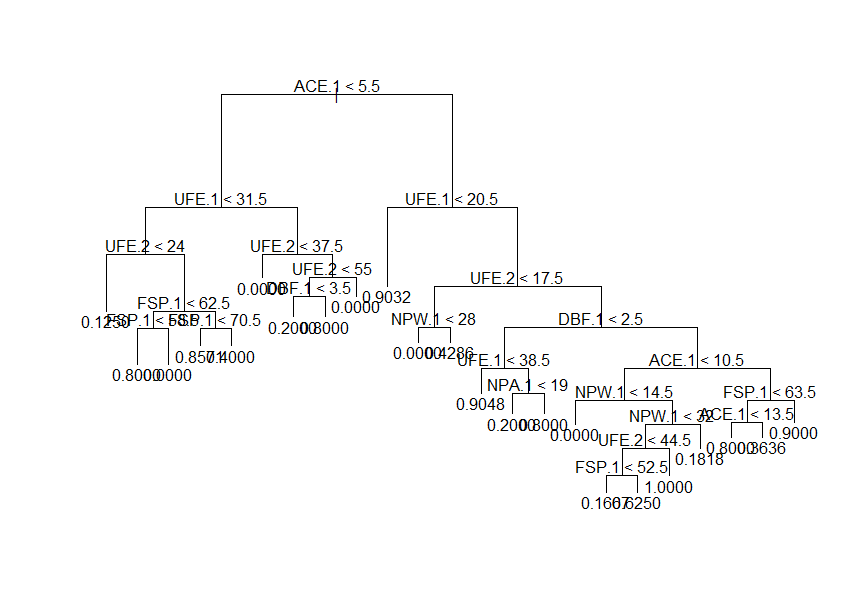
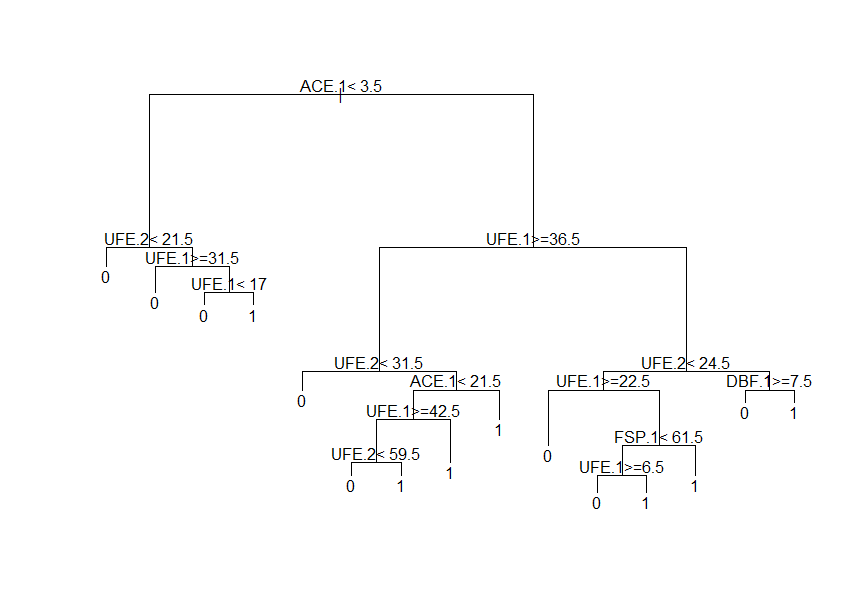


break point opportunities may be more significant than the raw number of opportunities created or the raw number of opportunities won. Upon further inspection of this data, it appeared as though in some matches, the break point statistics were miscalculated, as many matches had considerably more total break points won that break points forced. Additionally, many matches had considerably more break points forced than break points won. As such, I was forced to scrap break points from my analysis entirely.

**RESULTS**

 Moving on to the actual models tested, first I tested a logistic regression model. At first, I had thought that logistic regression would produce a meaningful summary of the effects of each predictor I chose on the probability of winning a game. To start with, when trained on one set of data and tested on another, the logistic regression model produced a test error rate of approximately 27%, which is not particularly good. But again, test error in prediction was not the purpose of this analysis. Looking into the coefficients, as coefficients in a logistic regression can often be difficult to interpret, the value here was in comparing coefficients of similarly-measured variable. One particular comparison of interest was that between ACE.1, or aces by player one (an ace is when one player serves the ball in such a way that the other player is unable to strike the ball in any way with their own racquet) and DBF.1, or double-faults by player one (a double-fault is where a player serves the ball, either out of bounds or against the net, twice in a row, resulting in a lost point). Typically in tennis, a player will serve aggressively on their first serve, in hopes of an ace, and if the aggressiveness leads to a fault, then the player will be less aggressive on his second serve, in hopes of avoiding a double-fault. The coefficients of the model are shown above. As you can see, ACE.1 has a coefficient of greater magnitude than that of DBF.1. This could be interpreted to mean that the reward of serving aggressively could outweigh the risks, assuming that if a player gets one fault, they will then serve conservatively, and that the probability of an ace on the first serve is at least close to equal to the probability of a double fault on the second serve. Another interesting comparison from the coefficients above is that between the effects of unforced errors by each player on the probability of player 1 winning the match. While it would not be particularly helpful in training for a match, these coefficients seem to imply that it is better to be good than it is to be lucky, as each unforced error by player one affects his chances of winning with greater magnitude than each unforced error by player 2. Finally, using the analysis of variance (ANOVA) function in R, I quantified the variance in winning probability accounted for by each of these predictors, and the results are shown below.

As you can see, once every predictor is taken into account, around 79% of deviance remains in the data. Despite producing a few interesting nuggets of information to back up a few commonly known facts about tennis, the logistic regression model did not turn out to be a particularly insightful way of analyzing this data set.

 Moving on to the decision tree model, my initial belief was that this model would shed the most light on the keys to victory for a player. The first model I made, however, I neglected to set a minimum objects-per-node parameter, and the tree ultimately looked like a jumbled and complex mess that definitely is overfitting the data set used to train it. To combat this, I decided to use cross-validation to prune the tree based on the misclassification error rate. After several hours of trying to debug the “prune.misclass” function in R, I finally decided to simply use a different tree-generating function. The tree produced below is much easier to interpret and understand. It maintains that Aces are the biggest key to victory, as having less than 3.5 aces in a game will almost always lead this tree to predict a loss. Interestingly enough, it seems as though the player can redeem himself from a lack of aces if he keeps his unforced errors below 17. This makes logical sense, as an ace is essentially a “free point” as the opponent does not make contact with his racquet. As such, if one is unable to force enough of these points, then they must in turn minimize his own mistakes and play a much tighter game. Alternatively, on the right side of the tree, you can see the final split at almost all points is either unforced errors or double faults. Thus, this analysis concludes that the most important factor in winning a tennis match is to minimize your own unforced errors while making sure to maximize the potential to get aces.

**DISCUSSION**

The biggest way this analysis could be improved would be if the statistics were inputted properly into the sheet. I had initially wanted to deep-dive into the importance of break points and net points, but the impossibility of not having equal total numbers of break points forced and won forced me to exclude this from the analysis. Additionally, as many of the matches had null inputs for relevant statistics, many of the already few data points had to be ignored in analysis. If this dataset were compiled over several years of tennis matches, then the models would have more data to train on, and thus could be more reliably accurate.