Final Project Memo

Jai Uparkar

October 02, 2022

```
## corrplot 0.92 loaded
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Predicting Professional WTA Tennis Match Outcomes with Machine Learning

Dataset Overview

The dataset I have chosen to predict professional tour tennis match outcomes with machine learning comes from an open source dataset from Jeff Sackman on GitHub. I will be downloading these files onto my local computer.

His open source datasets contains all the professional tennis match statistics and information from 1968 on the WTA. The information that is included in the dataset includes statistics on the match (percentage of points won, 1st points won, number of aces, double faults etc), information relevant to the specific matchup (each player's ranking and head to head), surface, player's physical attributes, and much more information. In total there are 48 features from every match obervsation and from this dataset I plan on extrating more information like whether or not the player had home course advantage, their head to head record with the opposing player, and much more information. There are approxmiately over 3000 records for each year and there are over 52 years of singles professional tour match information in the dataset. I plan on extracting more features from the dataset so that I will have more than 48 features from the dataset. Within this dataset I will be working with categorical and numerical variables.

The dataset contain missing information in 1992, some in 1985, and intermittently between 1973-1984. To offset this problem, I most likely will only analyze players within the the 22 years because I have the most familiarity with the players an formats of professional tennis from this time period. There is also some missingness for matches but since this dataset was used heavily in many other tennis machine learning projects, I know that other researchers found this problem to be sparing and not many records were eliminated. I do not believe that missingness will be an issue for my analysis based on existing research papers but in the event that there is a lot of missingess I may choose to focus on a different chunk of years, imputate, or decide which variables are revelant to my analysis and remove observations that don't have that information.

Research Questions

I am interested in predicting whether or not the player won the tennis match. That is my response variables and it is a yes or no question which makes it a classification question. Some other questions I am interested in answering are:

Can I predict Grand Slam or other big tournament winners? How does current win/loss streaks and accoloades affect player outcomes? Which players have extreme affinity for certain surfaces? How much does the surface contribute to the match outcome? Which surfaces tend to have the most upsetting (unexpected tennis outcomes)? How do ranking differences effect match outcomes? Were there certain periods of women's tennis in which outcomes were more volatiles and less predictable?

These questions are only the tip of the iceburg and I'm sure that once I begin the EDA process and become more familiar with the data I will have more questions to ask.

These questions will best be answered in a classification approach since I am interested in match outcomes (win or loss) but could be viewed in a regression perspective if I wanted to look at the percentage of winning vs losing.

The predictors I think that will be especifically useful are: player rankings, recent playing history, surface winning percentage, unforced errors, and also the player's age.

The goal of my model is predictive and inferential but mostly predictive since I want to best predict the match outcomes based on player statistics and current playing style.

Timeline

I plan on downloading the dataset early this week so I can figure out how to extract and derive other necessary information for my analysis. I plan on using the next 2-4 weeks to perform EDA and get comfortable with the data tha I am using. This will give me time to understand what information I am missing but will also give me time to reflect and see what questions I am really interested in solving. I will create models for the next 3-4 weeks after that. I then will use the last couple of weeks to compare the models and come to conclusions about my project. Throughout the process I will be working and writing my project so that my work is incremental and I won't feel overwhelmed at the end of the quarter.

Questions

How do you brainstorm EDA questions? What would be your proposed timeline for the project? What do you expect to be the greatest challenges of this project and what should I look out for? I'm a little worried as to how I am going to derive information like current win/loss streak and previous history at the tournament. I'm not exactly sure how to derive that information.