Predicting men’s tennis wining with ATP previous match information

Problem

Tennis is one of the most popular sports in the world, currently tennis sits at number 4 among all the sports in popularity with one billion worldwide fans. With the US open that held in New York city every year, tennis is also attracting a lot of fans in USA. The total prize money for the tournament is 62.5 million, with and unprecedented $4.1 million going to the winner. In the last 18 years, the total prize money has increase a whopping 351%.

With the growing popularity of tennis and increasing of the prize money, there is a huge market for sports analyst who can understand what the leading factors to affect the winning probability and who can predict the winner according to the previous performance. The information will also help people who are interesting in playing tennis betting game while watching games and try to maximum their profit.

The questions that I try to answer are:

* There are four different surfaces for tennis game, they are grass, hard, clay and hard courts. Does player perform differently on different surfaces?
* Who won the most titles from 2000 to 2018 among the four major “Grand Slam”? The four major “Grand Slam” are Australia Open, Wimbledon, French open and US open?
* Head to head analyzes between the top players, such as Roger Federer, Rafael Nadal and Novak Djokovic?
* What are the factors (ranking, age, number of Aces, points won from 1st serve and so on) that affect the winning probabilities?

Client

As mentioned earlier, I have two types of client clients.

1. Sports analyst

The information will help sports analyst to dig deep for each match and calculate the winning probability for each player, and also provide useful head to head information before the match.

1. People who are doing betting on the tennis

The information will also provide useful information for the people who are interested in playing the betting game and help them to put the betting and increase their chance to win the betting.

Dataset

I will build this model using the ATP match data 2000-2017(or 2018) dataset that is available on Kaggle([https://www.kaggle.com/gmadevs/atp-matches-dataset)](https://www.kaggle.com/gmadevs/atp-matches-dataset)%20and)  and GitHub at(<https://github.com/JeffSackmann/tennis_atp>).  This dataset contains various categories of data, including tournament name, tournament date, tournament level and surface type. It also includes the information about the winner and loser information, such as player ranking, seed number, player age and statistic results for each game, such as number of Aces, number of double faults, total serve points, 1st serve in and points won on 1st serve and on 2nd serve, break points saved and faced.  Each year’s data has more than 7000 rows and 49 columns.

Deliverables

I plan to deliver the code using Jupyter notebook file with 3-4 lines summary at the end of each code block, and also written a report (in PowerPoint slides) to summarize my findings.