STA531 Final Project-Preliminary Report

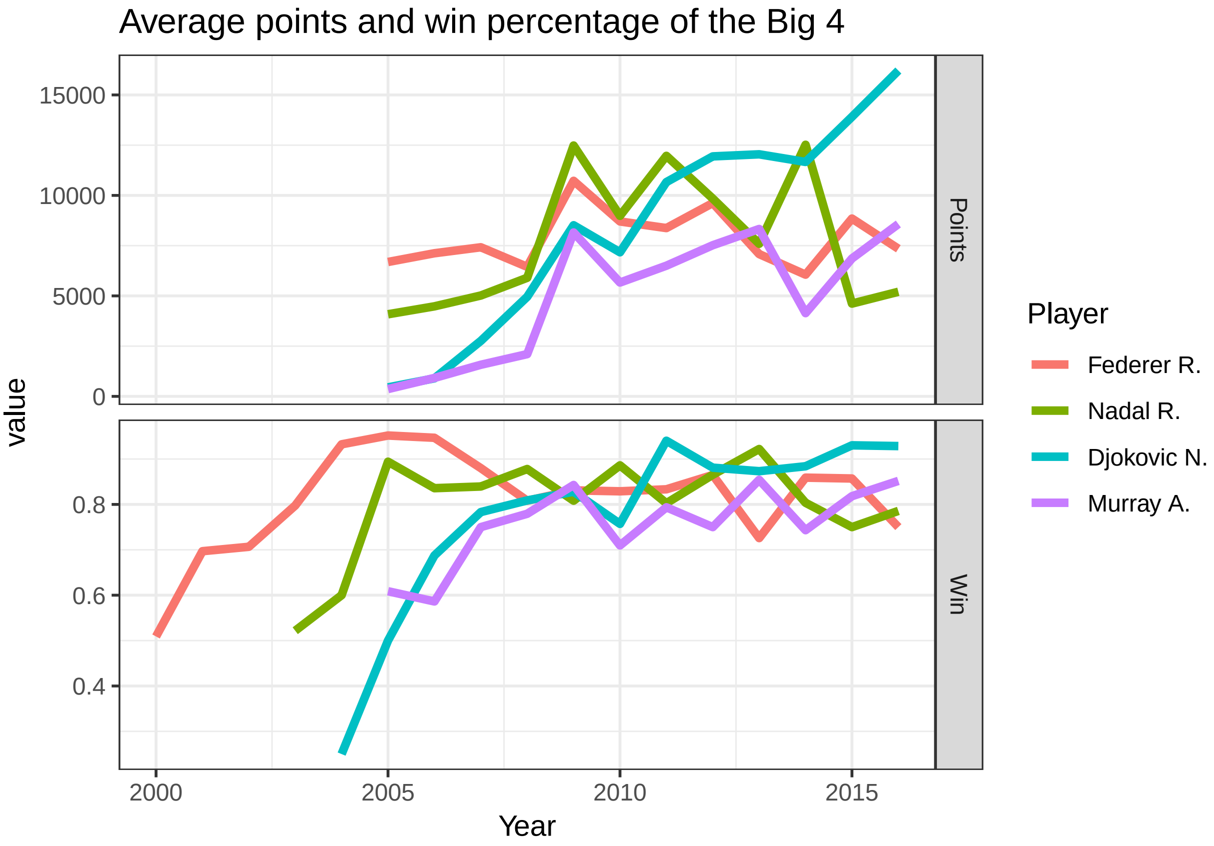
Eric Su

2019-04-10

Exploratory Data Analysis

The main purpose of this project is to quantify performance levels of professional tennis players. Currently, the rankings of players are determined by the amount of points they earned. Different levels of tournaments award different amount of points. For example, an ATP 500 series gives the champion 500 points, a Masters 1000 tournament gives the champion 1000 points, while each Grand Slam (the highest tournament level) awards the champion 2000 points.

The main goal of this project is to reliably estimate the true performance level of a player, which should be a more accurate metric for a player's ability than points or ranking. However, since performance levels are not directly observable, we cannot make EDA plots on them. Alternatively, we would make exploratory plots on points and win percentage and illustrate why they might not be suitable for reliably inferring a player's performance. In particular, we would focus on the four dominant players in the past 10 years, the so-called "Big 4" of men's professional tennis, namely *Roger Federer, Rafael Nadal, Novak Djokovic* and *Andy Murray*. Below we show the points and win percentage of these four players across the years.



It's clear that points and win percentage, while positively correlated, are not necessarily consistent with one another. Since the amount of points a player gains heavily depends on outcomes of certain matches (Grand slam matches in particular), differences in points may be the result of random variations. Win percentage also has the problem of not taking the performance levels of opponents into account. Obviously, winning against a tougher opponent is less likely than winning over a weaker opponents, but both points and win percentage are unable to reflect this.

To round up our exploratory analysis, we will take a look at some aspects of our data. First, we show the proportion of each court and surface that appears in the dataset.

Table 1: Proportion of Matches for each Court-Surface combinations

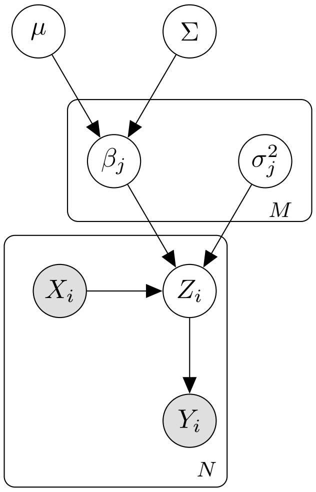
Court	Surface	Prop
Indoor	Carpet	0.036
Indoor	Clay	0.003
Indoor	Hard	0.139
Outdoor	Clay	0.327
Outdoor	Grass	0.111
Outdoor	Hard	0.383

The proportion of different combinations is clearly imbalanced. From the table, we notice that only a small fraction of matches are played on carpets. This is because this type of surface was removed from professional men's tennis tournament in 2009. One would also notice that the majority of indoor matches are essentially hard court matches, as there are only a handful of indoor clay court matches and no indoor grass court match. As a result, matches played on carpet would not be used and indoor/outdoor would not be included as a variable in this project.

Addtionally, since we want to estimate individual player specific coefficients, we have to deal with the issue of some players having played too few matches. To resolve this issue, we would group all players who have not played 300 or more matches as a seperate group: "Other".

Model

The graphical model this project uses can be expressed using the graph below.



where matches are represented using the index $i = 1, ..., N$ and players are represented using the index $j = 1, ..., M$ with variables

$$\begin{aligned}
X_i &: \text{Vector of external conditions (Series, Court, Surface, Round, Best of 3/4, Opponent rank)} \\
Y_i &: \text{Match outcome} \\
Z_{i,1:2} &: \text{Performance level of the two players} \\
\beta_j &: \text{Vector of regression coefficients} \\
\sigma_j^2 &: \text{Variance parameter for performance} \\
\mu &: \text{Mean hyperparameter for } \beta_j \\
\Sigma &: \text{Variance hyperparameter for } \beta_j
\end{aligned}$$

The outcome of each match will be modelled using the distribution:

$$Y_i = \begin{cases} 1 & \text{if } Z_{i,1} \geq Z_{i,2} \\ 0 & \text{if } Z_{i,1} < Z_{i,2} \end{cases}$$

For player j, his performance level in match i will be a linear combination of X_i(external factors) as shown below.

$$Z_{i,1 \text{ or } 2}^{(j)} = X_i^T \beta_j + \epsilon_{i,j} \quad \epsilon_{i,j} \sim N(0, \sigma_j^2)$$

with

$$\beta_j \sim N(\mu, \Sigma)$$

The prior distributions for this model will be

$$\begin{aligned}
\mu &\sim N(\mu_0, \Lambda_0) \\
\Sigma &\sim \text{inverse-Wishart}(\eta_0, S_0^{-1}) \\
\sigma_j^2 &\sim \text{inverse-gamma}(\nu_0/2, \nu_0\sigma_0^2/2)
\end{aligned}$$

The parameters of our model would be estimated using a Gibbs sampler with full conditionals as follows. We will use Z_j and X_j to indicate all (matrix) performace levels and external conditions for player j and n_j to represent the number of matches player j is involved.

$$\begin{aligned}
p(Z_{i,1}^{(j)} \mid X_i, Y_i, Z_{i,2}, \beta_j, \sigma_j^2) &= \begin{cases} \text{Truncated Normal}(X_i^T\beta_j, \sigma_j^2, Z_{i,2}, \infty) & \text{if } Y_i = 1 \\ \text{Truncated Normal}(X_i^T\beta_j, \sigma_j^2, -\infty, Z_{i,2}) & \text{if } Y_i = 0 \end{cases} \\
p(Z_{i,2}^{(j)} \mid X_i, Y_i, Z_{i,1}, \beta_j, \sigma_j^2) &= \begin{cases} \text{Truncated Normal}(X_i^T\beta_j, \sigma_j^2, -\infty, Z_{i,1}) & \text{if } Y_i = 1 \\ \text{Truncated Normal}(X_i^T\beta_j, \sigma_j^2, Z_{i,1}, \infty) & \text{if } Y_i = 0 \end{cases} \\
p(\beta_j \mid X_j, Z_j, \mu, \Sigma) &= N((\Sigma^{-1} + X_j^T X_j/\sigma_j^2)^{-1}(\Sigma^{-1}\mu + X_j^T Z_j/\sigma_j^2), (\Sigma^{-1} + X_j^T X_j/\sigma_j^2)^{-1}) \\
p(\mu \mid \beta_{1:M}, \Sigma) &= N((\Lambda_0^{-1} + M\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + M\Sigma^{-1}\bar{\beta}), (\Lambda_0^{-1} + M\Sigma^{-1})^{-1}) \\
p(\Sigma \mid \beta_{1:M}, \mu) &= \text{inverse-Wishart}(\eta_0 + M, (S_0 + \sum_{j=1}^{M}(\beta_j - \mu)(\beta_j - \mu)^T)^{-1}) \\
p(\sigma_j^2 \mid \beta_j, X_j) &= \text{inverse-gamma}((\nu_0 + n_j)/2, [\nu_0\sigma_0^2 + \sum_{i=1}^{n_j}(Z_{i,j} - \beta_j^T X_{i,j})^2]/2)
\end{aligned}$$