Initial List of Features

A. Player Features

These features will be computed for both Player A and Player B. In red is the name they will be given as a shorthand. A single 'W' or 'L' will be added to the end of each name to distinguish winner and loser. For example, 'RankW' and 'RankL'.

1) Rank (Rank)
2) Age (Age)
3) Weight (Weight)
4) Career Winning Percentage (CWP)
    a. Before start of match
5) Winning percentage on Surface that player is playing on (*POS*)
    a. Before that match
6) Winning Percentage in the tournament they are playing in (excluding results from this current year) (PIN)
    a. Tennis tournaments are usually annual affairs. Thus, we use their previous history at the same tournament in years past.
7) Increase or decrease in rank from their last 15 matches (slope of best fit line) (RTrend)
    a. We take their last 15 ranks, run a best fit line through it, and take the slope of this best fit line as our feature
8) Dominant Set Wins Percentage in the last 15 matches (*DSW*)
    a. We classify any set with a score of 6-2 in favor of the player and better as a dominant win (double break or higher). We count the total number of these "dominant set wins" in the last 15 matches. We then divide this number by the number of sets played in the last 15 matches.
9) Close Set Wins Percentage in the last 15 matches (*CSW)*
    a. We classify any set with a score of 7-6, 6-5, 6-4, or 6-3 in favor of the player as a close win. We count the total number of these "close set wins" in the last 15 matches. We then divide this number by the number of sets played in the last 15 matches.
10) Close Set Loss Percentage in the last 15 matches (*CSW)*
    a. We classify any set with a score of 7-6, 6-5, 6-4, or 6-3 against the player as a close loss. We count the total number of these "close set losses" in the last 15 matches. We then divide this number by the number of sets played in the last 15 matches.
11) Dominant Set Loss Percentage in the last 15 matches (*DSW*)
    a. We classify any set with a score of 6-2 against the player and better as a dominant loss (double break or higher). We count the total number of these "dominant set losses" in the last 15 matches. We then divide this number by the number of sets played in the last 15 matches.

Note: Our idea behind features 8-11 is to try and quantify certain factors that may cause a player to be due for regression to the mean. The record of a team in close wins is used by analysts in other sports (NFL for example) as a heavily weighted factor to try and predict regression to the mean when the season starts. A team that had a unusually high number of close wins in one season typically does worse overall the following year, and vice versa. We hope to see a similar result here. "

12) Aces Per Game in the last 15 matches (APG)
    a. Note we choose to divide by games here (and not matches or minutes) because games correspond closest with opportunities to serve and will give a better indicator of strength of serve
13) 1st Serve in / Game in the last 15 matches (FSI)
14) 1st Serve Won / 1st Serve In in the last 15 matches (FSW)
15) Break Points Faced / Game in the last 15 matches (BPFG)
    a. We wanted this to be indicative of how well they can return serves and set themselves up for success. A high number of break points faced (and low number won) could be indicative of future success (the player is currently getting "unlucky") and vice versa.
16) Break Points Won / Break Points Faced in the last 15 matches (BPWG)
17) Break Points Faced in career (BPFC)
    a. This serves as a balance to our previous thinking. Perhaps the player is just not good in clutch spots. So, we include career averages to hopefully see the difference in recent performance and career history.
18) Break Points Won / Break Points Faced in career (BPWC)
19) Record of similarly ranked players and Record against specific player (COPP)
    a. This are the same two features we have already computed
    b. We use the Bayesian update method to combine them into one useful feature.
20) Winning Percentage in a similar spot in the tournament (TSS)
    a. The stage of the tournament will be broken down into two stages:
        i. Round of 16 or earlier ("Early Stage)
        ii. Quarterfinals or Later ("Late Stage")
    b. Then, we will use the odds to determine an "underdog" and a "favorite"
    c. For each of the four possible options for each player, "underdog early stage", "underdog late stage", "favorite early stage", and "favorite late stage", we will determine the spot they are currently in and use their previous winning percentage in the same spots over their career as our feature
21) Number of Games Played in the last seven days (GP)

B. Game Features

22) Surface (Surf)
23) Binary variable to classify majors form non-majors (Major)
    a. We indicate the four biggest tournaments as "majors":
        i. Australian Open
        ii. French Open

iii. Wimbledon
iv. U.S. Open

24) Number of Sets (Sets)

25) Binary variable to classify stage of tournament (SOT)

a. The stage of the tournament will be broken down into two stages:
    i. Round of 16 or earlier (“Early Stage)
    ii. Quarterfinals or Later (“Late Stage”)