Predicting Tennis Matches & Tournaments

Team 19: Nathan Dennis, John Breedis, Yiming Chen

Problem & Interest

- Predicting tennis matches are difficult
  - Surfaces
  - Player ages
  - Experience levels
  - Past matchups
- Predicting tournaments based off these factors
  - Many different surfaces and set sizes based on tournament being played
- Tennis has ATP rankings based off a points system. We plan to use ELO scores
- Similar problem in other sports, hope to generalize

Scraping the Data

- Used requests library to read data from github
- For loop through the years 2000-2024 to be used in our dataset
    - https://github.com/JeffSackmann/tennis_atp
    - Includes data for all tennis matches within these years, except 2024.
    - Using 2000-2023 to train our models, then using 2024 to assess its accuracy.
    - Saved only columns we deemed necessary for analysis, including winner/loser names, surfaces, draw sizes, tournaments, ages, etc
- Players who have played little matches have base ELO of 1500, but we have adjusted the probability they will win a given match based off their experience
    - Explained more in mathematical function to predict games

Calculating ELO Scores

```python
# Adjusts ELO calculation rating based off tournament level.
if row['tourney_level'] == 'G':
    K = K * 6 # Worth double ATP 1000 matches, so multipled by 6
elif (row['tourney_level'] == 'A' or row['tourney_level'] == 'M'):
    K = K * 3 # Worth half grand slams.
elif row['tourney_level'] == 'F':
    K = K
elif row['tourney_level'] == 'D':
    K = K * 0.5 # Davis Cup has little effect on ELO scores.
```

- Began by utilizing base ELO calculation metric as given in class
- Different ELO scores for different surfaces of a tennis match
  - In tennis: Clay, Grass, Hard
  - Other sports, such as basketball (NBA): Home or Away (Maybe even international)
- Adjust ELO calculation scaled off tournament level
  - Grand Slams weighted more, similar to ATP ranking
  - Smaller level tournaments weighted less
- Adjusted ELO calculation scaled off year
  - Years closer to present day weighted more
  - Years further in the past weighted less based off a linear function
- Created new column for player age and games played
  - Age is calculated from last game played in the dataset, then adds a number dependent on the last match played (If last match played in 2019 at age 34, the new age is 39)
  - May not be 100% accurate

Predicting Games

- Predicting tennis matches set by set (Best of 3 or 5 sets)
- Individual games will be simulated based off factors
  - Implemented a decay function to have players who are older have their winning probability decrease as the match goes on for longer
  - Also implemented decay function for players who have less experience, where the probability of winning a match decreases
- Simulating tournament still in progress, hope to get results finalized soon

```
def compute_prob_in_sets(winning_prob, age, age_threshold1, age_threshold2, games_played):
    if age <= age_threshold1:
        return [winning_prob * ( 2/3 * np.exp(-1/(1+(games_played/20)**2)) +1/3) for i in range(5)]
    else:
        return [winning_prob * ( 2/3 * np.exp(-1/(1+(games_played/20)**2)) +1/3) *
                (1 - (age - age_threshold1)**2 / ((age_threshold2 - age_threshold1)**2 + (age - age_threshold1)**2) * i/5) for i in range(5)]
```

Comparing Results (Next steps)

- Using ATP rankings to compare our ELO scores
  - Currently, our ELO rankings and the ATP rankings for the end of 2023 have the same top 5 players, with our ranking having 8 of their top 10 in our top 10
- Predicting 2024 tournaments such as Australian Open, Wimbledon, French Open
  - Extracting draw from these tournaments, going to now predict winners and compare those results to who actually won
- Comparing results to betting odds

Results ELO scores

Elo results based off average across surfaces

```
Player_Name
Rafael Nadal            1732.529952
Stefanos Tsitsipas      1764.557890
Holger Rune             1771.833802
Grigor Dimitrov         1816.338822
Alexander Zverev        1834.402987
Andrey Rublev           1840.683475
Daniil Medvedev         1866.455613
Jannik Sinner           1894.798327
Carlos Alcaraz          2018.249570
Novak Djokovic          2165.507267
dtype: float64
```

ATP rankings, end of 2023

Rank ^	Player ^		Official Points ^	Next Best ^
1		N. Djokovic	11,245	-
2		C. Alcaraz	8,855	-
3		D. Medvedev	7,600	45
4		J. Sinner	6,490	-
5		A. Rublev	4,805	-
6		S. Tsitsipas	4,235	10
7		A. Zverev	3,985	-
8		H. Rune	3,660	-
9		H. Hurkacz	3,245	10
10		T. Fritz	3,100	90