Evaluating Tennis Match Models: Baseline vs Head-to-Head

Authors: Nathan Dennis, John Breedis, Yiming Chen

Abstract

Predicting tennis match outcomes is challenging due to factors such as variability in player performance and surface conditions. This paper compares two modeling approaches to solve this problem, one using general performance factors including surface specific ELO ratings and age, and another incorporating head-to-head match data. We evaluate four models with varying head-to-head weight adjustments to assess its impact on prediction accuracy, as well as the overall predictive performance of the models. Our results show that adding head-to-head match history reduces prediction accuracy compared to the general model, addressing debate over its reliability in predicting match outcomes.

1 Introduction

Predicting tennis matches is a difficult challenge that many struggle to perfect. While many models have been developed which attempt to predict sporting events such as tennis tournaments using features such as player statistics and historical data, reliable predictions remain difficult. This problem is of interest to not only analysts and researchers, but also to the millions of tennis fans and bettors who aim to accurately forecast match results and beat the oddsmakers.

This paper compares two different approaches to predict tennis match outcomes. The first approach uses general performance factors, including an enhanced model taking into account surface specific ELO ratings and player age to predict whether one player beats another. The second approach builds on this model, but incorporates a head-to-head (H2H) history between players, adjusting win probabilities based on past matchups. We utilize tennis match data between the years 2014-2022, all professional tournaments within this time period to calculate ELO scores where more recent matches played have a greater impact on player ratings. We make predictions for 3 major Grand Slam tournaments in 2023, Australian Open, Wimbledon, and Roland Garros using these ELO scores and corresponding age and head-to-head factors.

We examine how impactful implementing a model using H2H data improves prediction accuracy, as some argue that H2H records can be misleading[1]. We develop 3 different models with different parameters combinations with the H2H model. We evaluate prediction accuracy by comparing our models' predicted champions to betting odds[5]. By evaluating these models, we hope to contribute to the ongoing research in sports analytics and offer valuable insights for both bettors and analysts looking to refine their prediction models for tennis.

2 Description of Model

We utilized an ELO scoring metric to assign players a specific ELO rating. Using match data between 2014 and 2022, we first calculate players ELO rating across the 3 tennis surfaces (Clay, Grass, and Hard) using the formula in equation 1 with a scaling factor K and $p_{i,j}$ probability player i beats j on a given surface in equation 2, adjusting the new surface ELO given by R_i'. The K factor is a variable in the ELO rating system that controls how quickly and impactful players' ratings change based on a game, a higher K meaning more impactful.

$$\begin{aligned} R_i' &= R_i + K(1 - p_{i,j}) \quad \text{i wins} \\ R_i' &= R_i + K(0 - p_{i,j}) \quad \text{i loses} \end{aligned} \tag{1}$$

$$p_{i,j} = \frac{1}{1 + 10^{(R_i - R_j)/800}} \tag{2}$$

Before adjusting this formula, we researched what factors highly impacted tennis rankings and performance. First, we utilized a similar approach to the ATP rankings where higher level tournaments are weighted more highly, where we adjusted the K factor based on the tournament level when calculating ELO scores. We multiply the K factor by 4 for Grand Slams, 2 for masters/ATP tournaments, don't adjust K for all other tournaments and halve the K factor for the Davis Cup. This mirrors the ATP ranking system; the higher level tournaments contribute to more rapid changes in ratings, similar to our ELO score system.

Next, we adjusted the K factor in this formula based on the year a given match was played in. Matches closer to the present day are weighted more, as these matches are a more impactful indicator of how good a player is today. To calculate the new K factor, we used the formula $K \cdot e^{-0.3 \cdot x}$ where 0.3 is the decay factor and x is the current year minus the year the match was played. We chose a decay factor of 0.3 since it was a good balance between weighing the past 3 years highly with a K scaling adjustment of above 0.5 for recent matches, and reducing the impact of matches further in the past much less. In this approach, an exponential decay factor is applied to ensure the impact of older matches diminishes rapidly.

Furthermore, we utilized a new approach to adjust the ELO rating on specific surfaces, considering matches played on other surfaces. We realized that matches played should impact ELO ratings across all surfaces, but be more impactful for the surface a match was played on. It is not realistic to only create single surface ratings without accounting for success on other surfaces. When adjusting the ELO rating after a match is played, the ELO for the surface played on is adjusted using the scaling factor K. However, the other 2 surfaces also have their surface ELO rating adjusted by a scaling factor of 0.8, as given by equation 3. This allows players to gain ELO rating on other surfaces for any match played, making the model more realistic.

$$\begin{aligned} R_i' &= R_i + 0.8 \cdot K(1 - p_{i,j}) \quad \text{i wins} \\ R_i' &= R_i + 0.8 \cdot K(0 - p_{i,j}) \quad \text{i loses} \end{aligned} \tag{3}$$

$$\frac{P_{\text{Win Adjusted, n, i}}}{P_{\text{Win Adjusted, n, i}} + P_{\text{Win Adjusted, n, j}}} \tag{4}$$

After calculating ELO scores, we develop a model to simulate tennis matches, based on best of 3 to 5 sets, for Grand Slams 5 sets. First, we used the typical formula to compute the probability that player i beats player j based on their ELO scores using the formula given in equation 2. Now that there is a probability the players will beat one another we adjust these probabilities based on an age factor. Research has shown that tennis players usually peak when they are 25 years old[2,4], and as tennis players grow older fatigue becomes a major factor in their performance. As a match goes on, we decrease the probability a player will win a set using an exponential decay, where decay rate is controlled by λ, which determines how quickly player performance diminishes with age, only for players above 25 years old. For each set played the impact of the decay increases so the older player experiences a greater decline in performance, the formula being: $P_{win} \cdot (e^{-\lambda \cdot (age-25)})^{Set-1}$. So, P_{win} represents the probability of a player beating another, and this is multiplied by the exponential decay factor with parameter, λ, for players over 25 for each set in a match. This formula makes it so that as the number of sets increases, the probability an older player above 25 years old wins decreases. We set the decay factor to be different for Clay, where the decay factor for Clay was 0.015 and for Hard/Grass the decay factor was 0.0075. We did this as in analysis, we discovered Clay was a more unpredictable surface and is considered a more difficult court to play on with longer rallies[3], hence older players would be even more fatigued due to the long rallies on clay courts with a higher decay factor. We choose these decay factors both due to the fact the exponential distribution mirrored that of the research, and when optimizing this function through predicting past years tournaments

we discovered these parameters yield the most accurate results. To compute the probability a player i beats player j in a set, we use equation 4 with the new adjusted probability for each set, n.

This will be the first baseline model for analysis. Next, we incorporate a head-to-head factor, adjusting winning probabilities prior to incorporating the age factor. We utilized the winning percentage players had against each other and the number of games played between the players. We utilized a sigmoid function to decide on an adjustment factor based on the number of games played, with more games played between two players causing a greater win probability adjustment. The sigmoid function is given by:

$$\frac{0.5}{1 + e^{-k \cdot (x_{gamesplayed} - 10)}}$$

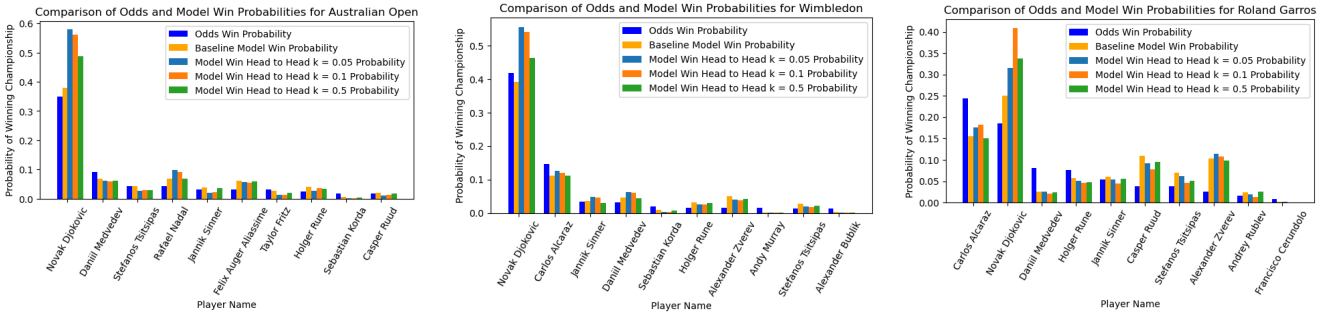
This sigmoid function calculates an adjustment factor, which determines the effect of the head-to-head winning probability on the match prediction. The logic behind this is that players who have more games played should have a higher weight on their head-to-head winning percentages. The sigmoid function uses rate parameter k, where a lower k indicates a smaller adjustment factor, and a higher k implies a larger factor. The 10 represents the transition of the sigmoid function to equal 0.5 when the number of games played between 2 opponents is 10. We decided to set this as 10 since when calculating the average number of games played between opponents, most players had a low number of games played between opponents such as 1 or 2, however some had a higher number of previous matchups, many as much as 20. We went with a value exactly half of this for the midpoint of the sigmoid function. Finally, we added a rate factor of 0.5 on the numerator to control for how impactful the number of games played is in the calculation. Without this rate factor, the sigmoid function had too high of an influence which wouldn't be as realistic which we found during parameter tuning. Using this adjustment factor, we adjust the probability that player i beats player j by using this equation:

$$\text{Adjusted } P_{i,j} = P_{i,j} + a \cdot (P_{hth} - 0.5)$$

Where $P_{i,j}$ is the original probability player i beats player j and a is the adjustment factor calculated from the sigmoid function above. Next, P_{hth} equals the head-to-head winning probability for player i beating player j, subtracted by 0.5 since 0.5 represents an equal number of times both players have won. This is the new adjusted probability that played i beats player j, where the probability player j beats player i would be $1 - \text{Adjusted}P_{i,j}$. Using these adjusted probabilities, the age factor decay function is taken into account to adjust the winning probabilities of both players. We experiment using 3 different k-factors in the H2H adjustment model, 0.05, 0.1, and 0.5, for analysis.

3 Analysis

Figure: Plots of 3 Grand Slam Tournaments and Predicted Probabilities



We observe the predicted probabilities given by each model in the 3 graphs above across the Australian Open, Wimbledon, and Roland Garros, alongside the predicting winning probability based on the odds, plotting the top 10 players. We ran tournament simulations 5000 times to compute the model's

winning probabilities. The 3 different H2H models are represented with 3 different k-factors, colored in the plot by blue, orange, and green, where the baseline model is colored yellow. We can make some key observations immediately. First, we see that in each plot the model incorporating head-to-head winning probabilities and games played consistently overrated Novak Djokovic's odds of winning each tournament, higher than both the baseline model and betting odds probabilities. This is most likely due to the fact he has a winning record against almost every player in the data. For any decay factor k, it seems like Novak Djokovic is heavily predicted to win each tournament from the H2H models.

Regarding the tournaments specifically, it seems like the Roland Garros, played on clay, has the most inaccurate predictions. Even using specific age decay functions for clay, it was not able to handle this complex behavior in the Roland Garros across all 4 models. It seems that in general across all the models, the baseline model without the H2H factor had the most consistent predictions, especially in the Roland Garros as the H2H models are heavily inaccurate with Novak Djokovic. Though, we do point out that Roland Garros is played later in the year, so having more match data before this tournament is played could've improved prediction accuracy.

The predicted winning probabilities for the baseline model look more accurate and consistent than the H2H models. One could claim these new models with this H2H factor could overfit the data and prioritize past matches too much, as they may not be indicative of the present day. Across the models, it doesn't seem like adding the H2H factor heavily improved the predictions, as it seems like in each tournament the baseline model was more accurate for most players.

3.1 Calculated Differences

We use 2 error metrics to compare the models results against each other and the betting odds. Across all 4 models, we calculate the root mean squared error (RMSE) and average of absolute differences (L_1) between the models predicted odds and true betting odds, observing this in the table below:

	Original Model	**H2H k = 0.05**	**H2H k = 0.1**	**H2H k = 0.5**
Australian Open RMSE	0.0094	0.023	0.021	0.015
Australian Open L_1	0.0035	0.0056	0.0053	0.0042
Wimbledon RMSE	0.0075	0.0139	0.0122	0.0076
Wimbledon L_1	0.0037	0.0041	0.0039	0.0039
Roland Garros RMSE	0.0159	0.0181	0.0244	0.02
Roland Garros L_1	0.0054	0.0056	0.0062	0.0058

We observe the calculated error metrics for all 4 models across each tournament. We can see that the RMSE for the original model is smaller than the head-to-head models across all k-factors and tournaments. We also see that the L_1 across all models and tournaments is also lower for the original model compared to any of the head-to-head models. This implies the average of the differences between the betting odds and predicted odds for the original model are closer compared to the H2H models. This helps to answer the question if the head-to-head factor improves the predictions, the error metrics show that the original model performs better as opposed to the model incorporating head-to-head data.

We notice that Wimbledon, played on grass, is consistently the most accurately predicted tournament, with the lowest RMSE scores across all 4 models. Roland Garros had the most inaccurate predictions, which indicates that even with a different age decay factor and head-to-head records the predictions for Roland Garros still fail to improve. While we have provided evidence that the head-to-head models did not improve predictions, we can conclude the predictive performance of the original model is very strong overall. The model's accuracy is high in comparison to the betting odds for both the Wimbledon

and Australian Open, which could suggest not only are these tournaments easier to model, but also the original model is closely aligned with that of the oddsmakers.

3.2 Parameter Sensitivity

While in this paper we only scaled the decay rate for the impact of the head-to-head factor, there are many parameters in this model that could change results. The first parameter used in this model was the decay factor for year when calculating ELO scores, this parameter could be adjusted to weight older years much less and solely prioritize the most recent years. While this approach seems interesting, when applying this idea it usually performed much worse as the more recent years too heavily influenced results.

Furthermore, when we adjusted the K factor for matches played on different surfaces, we chose a factor of 0.8. We experimented with different factors to calculate the ELO ratings, but using a very low factor made predictions more inaccurate since it assigned too high of a surface rating for some players. We made this factor larger due to the improved prediction accuracy across all models when attempting to optimize by predicting the 2023 tournaments based on different years match data.

The age decay factor was also a parameter of interest we changed in our approach. A higher decay factor forced older players to have greater fatigue as the match went on, severely decreasing their winning probability longer in a match. We decided to use a lower decay factor, which differed between Clay and other surfaces since Clay rallies usually take longer. We observed that with higher decay factors results greatly degraded prediction accuracy, most likely due to the fact these are still world class athletes. We utilized sensitivity analysis to determine the exact decay factors when adjusting the winning probabilities for those over 25 years old. We tested different decay factors across different years in the dataset in an attempt to optimize the exact formula, however none achieved significantly better results compared to the parameters we used in this paper.

4 Discussion

In this model, we utilized many different parameters and functions to make predictions for tennis tournaments. Starting with the elo calculation, we used a decay factor based on the year to weigh recent matches higher. This decay function used a decay rate parameter, which we adjusted based on our own analysis. We specifically chose a factor that weighted the past 3 years highly, with a K factor adjustment greater than 0.5, then the older years less. As stated in the analysis, choosing different parameters often harmed results, such as using a smaller decay factor which made past matches impact present day ELO calculations too highly. On the other hand, using a larger decay rate weighted the present day matches too highly, essentially giving higher ELO ratings to only players who won major tournaments the year prior. This makes sense, and we chose a good middle ground between these two extremes.

Furthermore, the age decay factor for older players, those above 25 years old, also made sense, but still could’ve been done differently. We realize these tennis players are gifted athletes, many of whom can play an entire tennis match for hours and not be severely impacted by fatigue. Because of this, we did not place an extremely high age adjustment factor for older players as we understand how much they train and do conditioning to play long tennis matches. This approach made more sense as we adjusted the parameters and conducted some sensitivity analysis, as an extremely high rate factor would harm predictions for older players such as Novak Djokovic. However, one could consider not using an age decay factor at all.

For the head-to-head factor, there were a variety of approaches one could use to implement this. Our approach utilized an adjustment factor on the winning probability based on the winning percentage between

the two players and the number of games they have played. This approach made sense since it is not only the winning percentage that matters, but also the amount of games played between players. We found in analysis that an average of 10 games played against each other to be a good middle ground, so we used this value in the sigmoid function as the halfway point of 0.5. In our analysis we experimented with different decay factors in the sigmoid function, which affected how much the adjustment factor changes the winning probability for players. However, this approach is flexible and could be adjusted further depending on future analysis and assumptions.

We learned a lot through this model building process. We took a very interesting approach where we utilized several decay factors across different functions. We learned how impactful the scaling of recent matches is on predictions, as weighing years closer to the present day too highly would skew results and favor those who have had recent success. We also learned that the age factor may not play as large a role in this analysis. While we kept it in the model, when experimenting by removing it from the model it did not heavily worsen results. This is probably because as stated, tennis players are world class athletes who train nearly daily for these tournaments, age may not be as big of an issue for these players.

For the main question on this project, it seems that the H2H factor did not improve model predictions, backing up prior claims that it is an unnecessary statistic to include in predictions. However, the use of only one approach is a major shortcoming of this analysis. There are many other approaches one could use to include this factor, which could change results. Even though the baseline model had the best predictions, we note that predictions were not uniformly accurate as the errors for the Roland Garros were much higher. There must be more factors that affect predictions on the Clay surface which contribute to its high unpredictability. Another shortcoming in this part analysis is not having more factors in our dataset that could influence match outcomes, such as player health, and the fact that we only had one source of betting odds as the Odds API had no data for our major tournaments.

5 Summary

In this project, we presented multiple approaches to build a model that predicts tennis match outcomes. We presented an approach which utilized a general ELO calculation scoring metric with factors such as year the match was played in and age of the player and another approach using head-to-head match data. Through our results and parameter analysis for the head-to-head models, we observed a declining accuracy compared to the baseline model. However, this approach could be done differently, which poses the future question about which approach may be the best at modeling head-to-head matches. In regards to our baseline model, we were pleased with its overall performance and accuracy compared to the betting odds. Future analysis could change the number of factors or formulas used in these models, as tennis is a sport with many factors that influence the outcome.

6 Attribution of Effort

We worked on the project together and decided on the topic to pursue as a group. We started by figuring out the proposal and what project we wanted to do, where we decided to expand on our project 1. We then discussed the approach we wanted to take for this project and communicated with each other regarding the details of the project and what code we will end up writing. Since we worked on it essentially as a team, it is harder to give team-member specific contributions. However, we do note that Nathan primarily worked on updating the new code and adding the new head to head adjustment factor. Yiming and John also helped with this, and also helped write the test cases for all of the code. Overall though we worked together to finish the project and write the paper.

7 References

[1] J.S. “The Cliché That Tennis Is a Sport of Matchups Is Probably Right.” *The Economist*, The Economist Newspaper, www.economist.com/game-theory/2017/07/12/the-cliche-that-tennis-is-a-sport-of-matchups-is-probably-right. Accessed 15 Nov. 2024.

[2] “When Do Players Peak? (The Definitive Answer - with Charts!).” *Tennis Frontier Forums*, Tennis Frontier Forums, 21 Apr. 2023, www.tennisfrontier.com/threads/when-do-players-peak-the-definitive-answer-with-charts.7631/.

[3] Wallace, Ava. *Order in the Court: An Animated Look at How Tennis Surfaces Change the Game*, The Washington post, 3 July 2024, www.washingtonpost.com/sports/interactive/2024/what-makes-tennis-surfaces-different/.

[4] Ferrer, David. “Home.” *Data Action Lab*, 5 Feb. 2022, www.data-action-lab.com/2021/11/29/analysis-of-age-in-tennis-data-2/.

[5] “Archived Tennis Futures Odds.” *SportsOddsHistory.Com | Archived Futures Lines of the Super Bowl, World Series & More*, 8 Apr. 2020, www.sportsoddshistory.com/tennis-odds/.