Data Science Research Final Report

Minh Trung (James) Vo
STUDENT NO. a1869086

April 19, 2024

Report submitted for Data Science Research Project at the School of Mathematical Sciences, University of Adelaide

THE UNIVERSITY of ADELAIDE

Project Area: **Predicting the outcome of tennis matches**
Project Supervisor: **Dr. Andrew Black**

In submitting this work I am indicating that I have read the University's Academic Integrity Policy. I declare that all material in this assessment is my own work except where there is clear acknowledgement and reference to the work of others.

I give permission for this work to be reproduced and submitted to other academic staff for educational purposes.

I give permission this work to be reproduced and provided to future students as an exemplar report.

Abstract

This study investigates the prediction of tennis match outcomes using various statistical and machine learning methods, with a focus on the Elo rating system and its variations, the Bookmaker Consensus Model, and the Glicko system. The primary objective is to accurately forecast the winner or loser of a tennis match, using available match statistics data. The performance of each prediction method is evaluated and compared, examining the trade-offs between model complexity, efficiency, and usability. The results offer an understanding of how each method performs in predicting match outcomes and highlight the relative strengths and limitations of each approach.

1 Introduction

Sport has always been a part of human culture for centuries, and for many audiences, the unpredictability of many of these physical and mental contests have arguably one of the beauties of it. However, it never stops people from trying to forecast what will happen in a sport match, it starts from just who might win, to score prediction, to individual players' statistic and events that can happen in the game. In current years, predictive analysis plays an ever-growing role in the realm of sports, serving various purposes such as coaching, enhancing fan engagement and informing betting decisions. The abundant of accessible data, combines with the advance in technology continuously pushes the accuracy of these forecast to a higher level.

In this project, we aimed to construct and critically evaluate a range of methods for predicting the outcomes of tennis matches. We built and tested several models, starting with simpler approaches such as naïve and logistic regression models as our baseline, and progressing to more advanced methods, including different implementations of the Elo rating system, the Glicko system, and the Bookmaker Consensus Model (BCM).

Most of our models were built and validated using ATP match statistics spanning from 1968 to 2019, except for the BCM, which was based on a distinct dataset that included bookmaker odds data from 2001 to 2019. Models built from previous match results, such as Elo and Glicko system, showed improvement as their complexity increased. However, the BCM outperformed every other model by a significant margin, likely due to the high level of complexity inherent in the bookmaker odds.

2 Background

Sport prediction plays a significant role in various areas, including sports betting, sports analytics, team strategy planning, and fan engagement. Accurate predictions can enhance the experience of fans and bettors by providing insights into potential outcomes, while also helping coaches and teams refine their strategies and gain a competitive edge. Additionally, predictions contribute to a deeper understanding of the sport itself, leading to more informed discussions and decision-making at multiple levels.

Tennis, like any other sport, can greatly benefit from predictions, and this is particularly true for single matches where factors such as team dynamics are minimised. This focus on individual players and matches allows for more precise forecasting, which can have significant implications for players, coaches, and analysts alike.

In the first trimester of our study, we explored and evaluated several models for predicting tennis match outcomes. These included naïve and logistic regression models, as well as the original Elo rating system with a constant k-factor and some of its variations. This phase provided a foundation for understanding basic modeling approaches and their limitations in predicting match outcomes.

During the second trimester, we shifted our focus to more advanced models, such as FiveThirtyEight (FTE) Elo rating system, the Bookmaker Consensus Model (BCM), and the Glicko system. These models offer different perspectives and techniques for assessing player performance and match outcomes. We also gained additional insights into how each model performs within specific brackets of matches, which allowed us to evaluate their effectiveness across different levels of competition.

Through this study, we aimed to assess the strengths and weaknesses of each model, ultimately contributing to a better understanding of which approaches are most effective for predicting tennis match outcomes.

3 Methods

Overall, we used two different datasets for our study. The main dataset was utilised for the naïve model, logistic regression model, the Elo rating system and its variations, as well as the Glicko system.

The secondary dataset was used specifically for the BCM and merged with the main dataset to enable a fair comparison between BCM and other models. All models were evaluated using several metrics, including accuracy, log loss, and calibration, to determine their predictive performance and reliability.

3.1 The data

Our main dataset, comprising 191,300 man single matches recorded between 1968 and 2024, was compiled and is maintained by Jeff Sackmann (Sackmann 2024) [7]. This comprehensive repository includes essential match details such as the tournament name, date, and level, along with information on the match surface, set scores, various serve statistics, players' information, and the ATP rankings and ranking points of both the winner and loser at the time.

Players are awarded ranking points based on how far they progress in an ATP tournament and the tournament's prestige, with the highest points available in the four Grand Slam events. Their rankings are then updated according to their current points. However, player rankings only become available from 1973 onward, with ranking points and serve statistics introduced in 1990. Our model will mainly be constructed using data from 1968 to 2018, with data from 2019 reserved for testing and validation. Figure 1 highlights the ranking for some of the top players from 1990 to 2019.

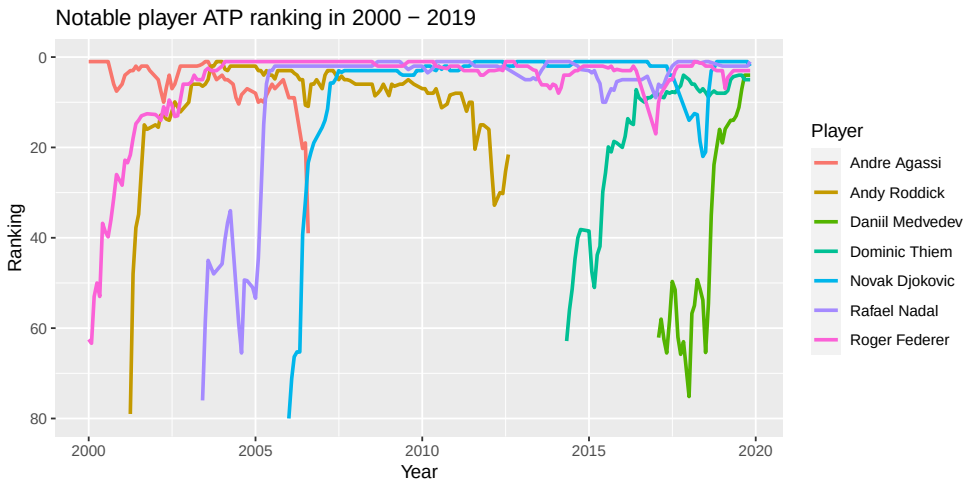

Figure 1: Notable player ranking from 2000. Federer, Nadal and Djokovic dominating in this period.

In the figure, we can observe the rise of the top 3 players in Federer, Nadal and Djokovic and their sustained domination throughout the years. On the other hand, Agassi concluded his career in 2006, while the rise and fall of Roddick were also captured within the period of early 2000s to the year 2012. Younger players such as Medvedev and Thiem also made their appearance in the mid-to-late 2010s.

Our secondary data set was sourced from tennis-data.co.uk and includes match statistics from 2001 to 2024 (Tennis-Data 2024) [8]. The primary focus of this data set is bookmaker odds, which are essential for constructing the BCM.

To fairly compare the performance of the BCM, we merged the betting data with our main ATP data set using match date and player ranking as common data points. The rationale behind this approach is that player rankings do not change frequently, so a ranking on a specific date can be used to identify the players participating in a given match.

While there were some discrepancies in match dates, as our ATP data only provided the tournament start date and tournaments typically last up to two weeks, we cross-checked the actual match date with the tournament date within the preceding two weeks to merge the data sets. The final data set for BCM analysis includes over 50,000 matches from 2001 to 2019.

3.2 Validation

All our models will be assessed based on their performance on the testing data from 2019, using three different metrics: accuracy, calibration, and log loss. While the naïve, logistic regression, and BCM models do not require tuning, we will tune the Elo and Glicko systems using the same three metrics.

3.2.1 Prediction accuracy

Being one of the simplest way to validate a model performance is accuracy, which is calculated by:

$$\alpha_1 = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\{f(i=y_i)\}}$$

In this formula, N is the number of games in our validation data, and y_i is an indicator variable such that:

$$y_i = \begin{cases} 1 & \text{if higher ranked player won game i} \\ 0 & \text{if higher ranked player lost game i} \end{cases}$$

with function $f(i)$ returns the prediction of game i winner. The function $\mathbf{1}_{\{A\}}$ is known as an indicator function that is expressed by:

$$\mathbf{1}_{\{A\}} = \begin{cases} 1 & \text{if condition A is satistifed} \\ 0 & \text{otherwise} \end{cases}$$

For models that return a probability, we need to define a cutoff $\eta \in [0, 1]$, as the threshold where the probability return a success outcome prediction. For our models (and also in most cases), we take $\eta = 0.5$.

For the naive model, since the probability of higher ranked player winning $\pi_{naive} > 0.5$, we have $f(i) = 1 \; \forall i = 1, 2, 3, ..., N$. For logistic regression model and Elo system model we adjust function f as follow:

$$f(i) = \mathbf{1}_{\{\pi_i > 0.5\}}$$

where π_i is the prediction probabilities from either models.

Accuracy is easy to calculate and gives a clear measure of how well a model performs overall. However, it doesn't consider data distribution. With an imbalance data set, and a model predicting the majority class may achieve high accuracy but lack the reliability, the metric proves to be insufficient. For example, if the dataset contains a large proportion of matches involving top players, a poorly performing model might consistently predict these top players to win and achieve high accuracy within the dataset. However, this model would likely lack the ability to transfer its predictive power to other datasets.

3.2.2 Calibration

The calibration, denoted as C, is defined by the equation:

$$C = \frac{1}{W} \sum_{i=1}^{N} \pi_i$$

where W is the number of games won by the higher ranked player, and π_i is the probability of the higher ranked player winning in the i'th game. We usually aim for a well calibrated model where $C \approx 1$. If $C > 1$, the model tends to overestimate the wins of the highest-ranked player, while $C < 1$ suggests an underestimation in this regard.

With these properties, calibration provides good insight in evaluating bias within our models by indicating whether predicted probabilities align with actual outcomes. However, it cannot be used alone to interpret a model's predictive power and needs to be utilised in conjunction with other metrics.

3.2.3 Log loss

The log-loss L, also known as the cross entropy, is defined by the equation:

$$L = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)$$

Log loss penalises incorrect decision mode severely, especially for those made with high confidence, which is why it play a crucial role in validating our models, as we always want to minimise overconfidence for a sport prediction model that might be used for betting.

However, similar to calibration, log loss cannot be easily used to interpret predictive capability on it own. Besides that, it may be overly sensitive to outliers and imbalanced data.

In summary, when assessing the models, an inclusive evaluation using accuracy, calibration, and log loss in conjunction provided us with a comprehensive assessment of their performances. This approach takes into account their predictive capability, level of bias, and whether the model tends to be overconfident in its predictions.

Despite our attempt to integrate other metrics such as sensitivity, specificity, and F1 scores, our testing indicated that these metrics did not offer additional insights compared to the combined evaluation of accuracy, calibration, and log loss.

3.3 The models

3.3.1 Naive Model

The naive model is quite straightforward, it predicts the player with the higher ATP ranking points to win the match. In this case, the prediction probability for higher player winning is always $\pi_i = 1$ if $A_i,_1 > A_i,_2$, with $A_i,_1$ as the higher ranked player points and $A_i,_2$ as the lower ranked player points. To implement the model, we create a new Boolean variable to show whether the winner of the match has higher ranking points.

3.3.2 Logistic Regression Model

The logistic regression model utilise the difference in ATP ranking points between 2 players, with the mathematical function as follows:

$$\text{logit}(\pi_i) = \log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 D_i$$

Inverting the function will then give us the probabilities as:

$$\pi_i = \frac{1}{1+\exp(-(\beta_0 + \beta_1 D_i))}$$

In our model, D_i represents the difference in ATP ranking points between the two players, hence $D_i = A_i,_1 - A_i,_2$. Additionally, we fitted the model without the intercept term β_0, implying that when the difference in points is zero, the probability of winning the match is set at 50% for both players. This implies that the model is simplified to:

$$\pi_i = \frac{1}{1+\exp(-\beta_1 D_i)}$$

A new variable was created to show the difference between the players' points. We then employed the glm() function from the stats package in R to conducted model fitting using our training data spanning from 1968 to 2018. The outcome yielded a coefficient $\beta_1 = 5.768e^{-04}$.

3.3.3 Standard Elo system

Let $E_i(t)$ and $E_j(t)$ denote the Elo scores of players i and j at time t, where t represents the t'th match played. In our model, we assumed that all players started with an Elo ranking of $E(1) = 1500$. Let $S_{i,j}$ denotes whether player i defeat player j, then the match can be expressed as:

$$E[S_{i,j}] = P(S_{i,j} = 1) \times 1 + P(S_{i,j} = 0) \times 0 = P(S_{i,j} = 1)$$

Let $\pi_{i,j} = P(S_{i,j} = 1)$ then the probability of player i winning is determined by the logistic function:

$$\pi_{i,j}(t) = \left[1 + 10^{\frac{E_j(t) - E_i(t)}{400}}\right]^{-1}$$

Using this predicted probabilities, player i rating would be updated as follow:

$$E_i(t+1) = E_i(t) + K_i(t)(W_i(t) - \pi_{i,j}(t))$$

With $W_i(t)$ as an indicator variable denoting whether player i won their t'th match. Thus, player i rating will increase by $K_i(t)(1 - \pi_{i,j}(t))$ for a win, and decrease by $K_i(t)(-\pi_{i,j}(t))$ for a loss.

Function $K_i(t)$ specify the rate of change in the Elo rating for player i. The original Elo system utilises K-factor model, in which the function remains a constant $K_i(t) = k$ for all i.

Our model was built in R by creating functions to calculate players' Elo ratings for each match, generate predictions, and update players' ratings according to the match outcome. The model was tuned using a range of k values from 1 to 100 on our training data and then refitted with the optimised metrics on our testing data. As depicted in Figure 2, both accuracy and log loss reach optimal levels for various k values, while calibration kept increasing in tandem with k value.

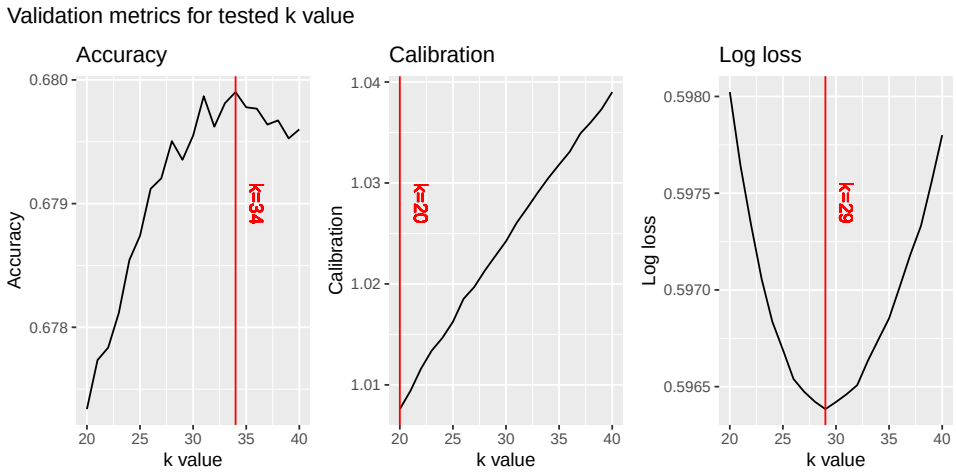

Figure 2: validation metrics in training for each k value. Optimal value for accuracy and log loss is 34 and 29 respectively, while calibration increase with k.

Our further analysis determined that $k = 31$ yielded optimal results for all three validation metrics. While the accuracy of 0.6799 was second only to the optimal value at $k = 34$, it retained excellent calibration at 1.0260 and log loss at 0.5965.

3.3.4 Elo for each surface type

This Elo system is built on the premise that players may excel on a specific court type, with the most notable example being Rafael Nadal, who won 11 out of his 17 Grand Slam titles (as of 2019) on clay courts, compared to just 1 out of 20 for Federer (Gorgi, Koopman & Lit 2019)[2].

For this Elo system, we partitioned our data into independent sets based on each court type and ran the Elo model for each data set separately. We followed the same tuning process, using a range of k values from 1 to 100 for each court type. The optimal k values for each court type are outlined below:

Table 1: Optimal k value for each court type Elo system, hard court projects the lowest adjustment per match with k = 35, while grass court has the highest adjustment with k = 56.

Court type	Best k value
Clay	40
Grass	56
Hard	35

3.3.5 Combine Elo system

This approach was employed by Leighton (2021)[9] in their paper and referred to as Adjusted Elo. For this system, we applied both the Original Elo and the Surface Elo models to our data set, generating two separate Elo ratings for each player. These ratings were then combined using weighted contributions to create the final Combined Elo E_C rating used for prediction.

The contribution from each Elo rating to create the final E_C rating for players is represented by the function:

$$E_C = \lambda * E_{Sd} + (1 - \lambda) * E_{Sf}$$

where E_{Sd} is the player Standard Elo, and E_{Sf} is the player Surface Elo.

To optimise this Elo system, we tuned our model by running λ from 0 to 1 with steps of 0.1. As shows in Figure 3, the optimal λ values were found to be 0.4 and 0.6 for different metrics.

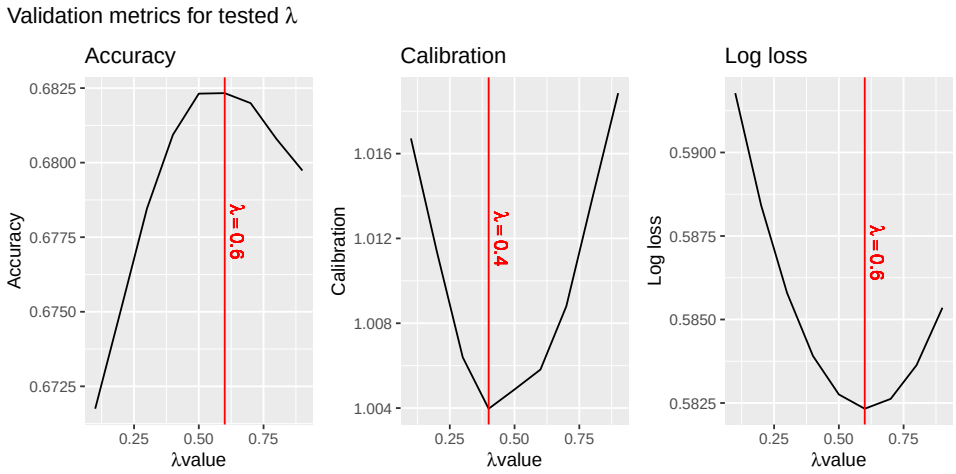

Figure 3: Tested λ for each metric. Optimal value for calibration and log loss is at 0.4, while maximum accuracy achived at 0.6.

In this case, we selected $\lambda = 0.5$ for our model as a compromise between accuracy and bias, meaning both Elo ratings contribute equally to our predictions.

3.3.6 FiveThirtyEight Elo rating system

Unlike the standard Elo system, the FTE system uses a dynamic $K(t)$ function that is based on the number of games a player has played. This function is defined as follows:

$$K_i(t) = \frac{\delta}{(m_i(t) + \nu)^\sigma}$$

In this function, $m_i(t)$ represents the number of games played by the player up to time t, while δ, ν and σ are tuning parameters. Overall, δ and ν determine the starting point of the K function, while σ controls the rate of decline in the k value across the player's matches.

In their article on forecasting US Open matches, Morris, Bialik, and Boice (2016)[6] refer to δ as the constant, ν as the offset, and σ as the shape. They used $\delta = 250$, $\nu = 5$ and $\sigma = 0.4$ for their predictions.

Using these parameters as a baseline, we explored and tuned our model with δ ranging from 150 to 300, ν from 1 to 50, and σ from 0.25 to 0.5.

Our preliminary results showed that $\nu = 5$ was already optimal, so we focused our main tuning efforts on δ and σ. As shown in Figure 4, an increase in δ often needs to be combined with a higher σ for optimal results.

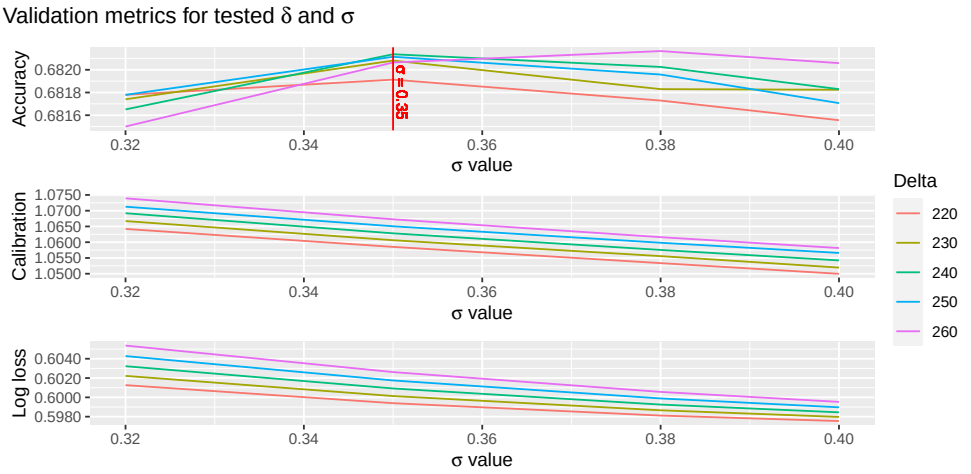

Figure 4: Tested λ and σ for each metric. Optimal value for accuracy achieved with $\lambda = 260$ and $\sigma = 0.4$, while calibration and log loss improved when σ increase for each model.

In conclusion, we selected $\delta = 240$ and $\sigma = 0.35$ which has the second highest accuracy with decent log loss and calibration to achieve a balance between accuracy and bias.

3.3.7 Bookmaker Consensus Mode

The BCM uses odds from different betting companies to arrive at a consensus estimate of a given player's probability of winning. This model was first introduced by Leiner, Zeileis, and Hornik (2009)[4] to predict a player's probability of winning a tournament, but it can also be adjusted for match predictions.

Before we proceed with our implementation, we need to address the issue that odds from bookmakers typically have overround, which is the built-in margin that bookmakers use to ensure profit regardless of the outcome. Overround occurs when the sum of implied probabilities based on the odds exceeds 1. This overestimation allows bookmakers to collect more money in bets than they pay out in winnings. To maintain accuracy in models using bookmaker odds, it is crucial to correct for overround by normalising the probabilities.

Consider a single match between two players. First, we convert the odds from a single company to probabilities. Let α and β represent the odds of players i and j winning, respectively. Then the implied probabilities of winning are given by $\tilde{p_i} = \frac{1}{\alpha}$ and $\tilde{p_j} = \frac{1}{\beta}$. The normalised

probability for player i is

$$p_i = \frac{\tilde{p_i}}{\tilde{p_i} + \tilde{p_j}}$$

substituting the odds in we get

$$p_i = \frac{1/\alpha}{1/\alpha + 1/\beta} = \frac{\beta}{\alpha + \beta}$$

Assume we have odds offered by N different companies, with odds offered by the kth company for player i and j winning being α_k and β_k respectively. Using the same caculation, the probabilities for player i winning is:

$$p_{k,i} = \frac{\beta_k}{\alpha_k + \beta_k}$$

The logit of player i's winning probability can then be calculated as the average logit probability from each company as follows:

$$\text{logit}(p_i) = \frac{1}{N} \sum_{k=1}^{N} \text{logit}(p_{k,i})$$

The function can then be inverted, and since $y = \text{logit}(p_i)$ is a number, and by definition

$$\text{logit}(p_i) = \log(\frac{p_i}{1 - p_i})$$

we get

$$(p_i) = \frac{e^y}{1 + e^y}$$

as the estimate of player i winning probability.

Since the BCM does not require tuning, we only needed to run the calculations on the bookmaker odds for all matches in our secondary merged data set.

3.3.8 The Glicko system

The Glicko system, developed by Mark Glickman (2016)[1], is an extension of the Elo system that introduces a measure quantifying the uncertainty of a player's skill level, known as the rating deviation (RD). This shift changes a player's rating from a fixed number to a confidence interval, where a player with rating G has their 95% confidence interval defined as:

$$(G - 1.96 \times RD, G + 1.96 \times RD)$$

The Glicko algorithm typically updates a player's rating during each rating period by treating games within the same period as if they occurred

simultaneously. Although a rating period can last for months and include many matches, it can also be applied on a match-by-match basis with shorter rating periods, such as a single day, since players typically play just one match per day in a tournament.

For longer rating periods, players' ratings and RDs are acquired at the start of the period, and the outcomes of games during the period are used to update their ratings and RDs at the end of the period.

The process for calculating a player's rating through each period is as follows:

Step 1: Determine the rating and RD for each player:

- For unrated players, set their rating and RD to default values.
- For players who participated in the previous period, obtain their existing rating and RD, and then calculate the new RD using the formula:

$$RD = \min(\sqrt{RD_{old}^2 + c^2}, 350)$$

where constant c determines how quickly uncertainty increases between each period, with a maximum limit of 350 to ensure the player's RD does not exceed this threshold.

Step 2: Update the rating for each player using these following instructions:

- Assume that the player's pre-period rating is G, with their RD from step 1.
- Let the pre-period ratings of the m opponents, also from step 1 be G_1, G_2,..., G_m and their ratings deviations be RD_1, RD_2,..., RD_m.
- Let s_1, s_2,..., s_m be the outcome against each opponent, in our case 1 or a win and 0 for a loss, as a tennis match cannot be draw.
- Let G' and RD' be the post period rating and ratings deviation for the player. The updating formulas are given by

$$G' = G + \frac{q}{1/RD^2 + 1/d^2} \sum_{n=1}^{m} g(RD_n)(s_n - E(s|r, r_n, RD_n))$$

$$RD' = \sqrt{\left(\frac{1}{RD^2} + \frac{1}{d^2}\right)^{-1}}$$

where

$$q = \frac{\ln}{400} = 0.0057565$$

$$g(RD) = \frac{1}{\sqrt{1 + 3q^2(RD^2)/\pi^2}}$$

$$E(s|r, r_n, RD_n) = \frac{1}{1 + 10^{-g(RD_n)(r-r_n)/400}}$$

$$d^2 = \left(q^2 \sum_{n=1}^{m} ((g(RD_n))^2 E(s|r, r_n, RD_n)(1 - E(s|r, r_n, RD_n)) \right)^{-1}$$

These calculations are performed for each player participating in the rating period.

For predicting match outcomes, the expected outcome formula for player i competing against player j with rating of r_i and r_j, RDs of RD_i and RD_j is as follows:

$$p_i = \frac{1}{1 + 10^{-g(\sqrt{RD_i^2 + RD_j^2})(G_i - G_j)/400}}$$

In general, the Glicko model can be tuned via the initial rating, initial RD, the c value and the upper and lower limit for RD. Due to the time constrain, we were only able to tune several model using only c constant. The tuned results is shown in Figure 5.

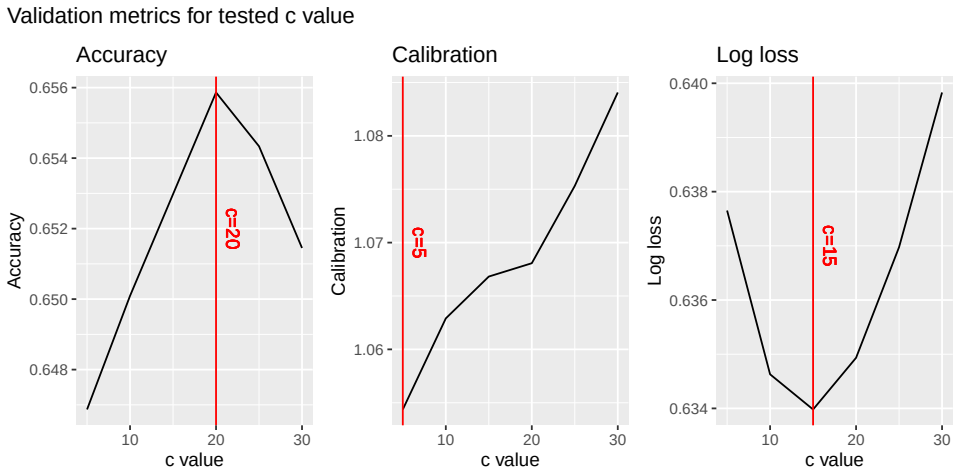

Figure 5: Tested c for each validation metric, with optimal c for accuracy at 0.2, and optimal c for log loss at 0.15. Calibration increase as c increase.

In this scenario, we selected the constant c value of 0.15 to maintain a balance between predictive power and bias.

4 Results

4.1 Elo systems behaviour

Overall, our standard Elo system behaved quite comparable to the ATP ranking. Figure 6 shows the Elo rating of the same players in Figure 1.

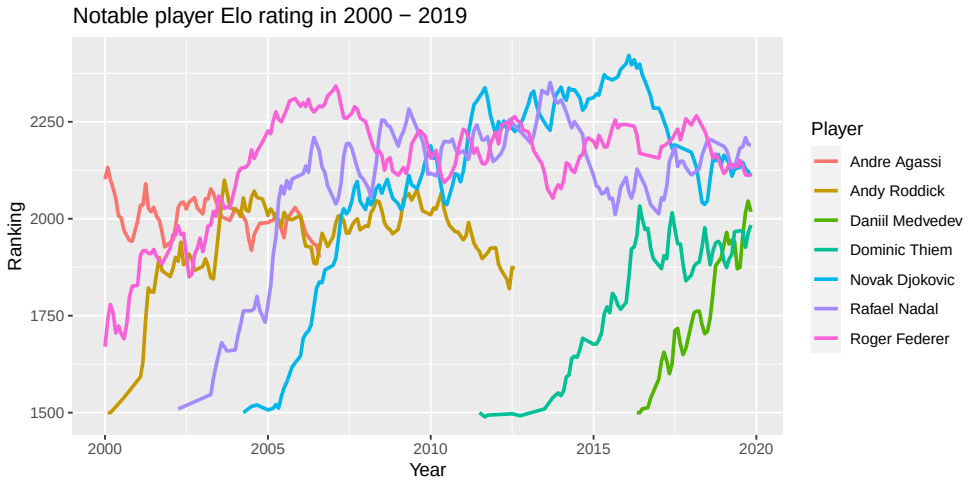

Figure 6: Notable player Elo rating from 2000. While Federer, Nadal and Djokovic is still dominating, their ratings are all dropped from their respective peaks.

While Figure 6, in a lot of sense, depicted a similar story in Figure 1, with the trio of Federer, Nadal, and Djokovic maintaining their dominance while others phased in and out, it clearly showed a decline for all three from their peak performances. This phenomenon has already been observed in Women's tennis, as reported by Morris and Bialik (2015)[5], where Serena Williams retained retained her top ranking despite a decline in her performance Elo. Morris and Bialik suggested that it might be due to the "weaker competition" within that period. Interestingly, by the end of 2019, Nadal had the highest Elo rating in our system, even though he was in second place in the ATP ranking.

The most notable difference between FTE and the standard Elo is the rate of change. Figure 7 shows that while Federer's rating followed a similar trajectory from 2005 onward in both systems, his FTE rating exhibited more volatility in the first couple of years. Nevertheless, his FTE rating still increased more rapidly compared to his standard Elo rating.

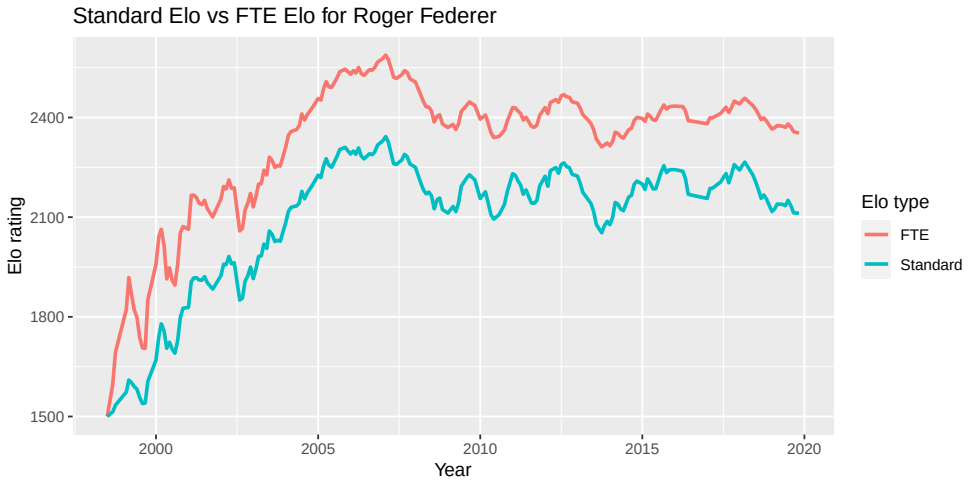

Figure 7: Roger Federer's Standard and FTE Elo ratings throughout his career, highlighting the faster rate of increment in FTE.

4.2 Models validation metrics

The model performance in our testing dataset (matches in 2019) shows a clear trend where the more advanced models offer additional predictive capability. The simple naive model turned to be the least effective, and while the logistic regression model had the lowest calibration, it only offered a minor improvement in term of accuracy over the naive model. Table 2 presents the validation metrics of our model on the main testing dataset, excluding the BCM model, which we will discuss separately.

Table 2: Validation metrics on the main testing data, showing that models with higher complexity provide better predictive power, except for the not fully optimised Glicko

Model	Accuracy	Log loss	Calibration
Naive Model	0.6134328	0.6711751	1.069772
Logistic Regression	0.6138060	0.6605081	1.034199
Standard Elo	0.6442484	0.6369879	1.057456
Surface Elo	0.6471885	0.6375742	1.069357
Combine Elo	0.6504961	0.6284328	1.047144
FTE Elo	0.6515987	0.6464454	1.087260
Glicko	0.6467379	0.6403769	1.075399

The Standard Elo system provided a respectable improvement in both accuracy and log loss compared to the naive and logistic regression mod-

els, but with a calibration higher than the logistic regression model. Besides that, the Surface Elo system performed comparably to the Standard Elo, only showed a slight edge in accuracy albeit a higher log loss and calibration. The Combine Elo came out as the best performing model among k-factor Elo model, with improvements across all metrics over the other 2 Elo systems.

The FTE model, despite high calibration and log loss, still offered significant improvement in predictive efficiency. The not fully optimised Glicko also performed comparatively to the standard Elo.

Since the BCM was built using a different data set, we used the merge data for our next comparison. The Elo models was built using the merged data from 2001 - 2018 for this section. Table 3 shows the performance of the Standard Elo, FTE and BCM in 2019.

Table 3: Validation metrics on merged data with bookmakers' odds, showcasing the significant improvement across all metrics from the BCM

Model	Accuracy	Log loss	Calibration
Standard Elo	0.6337806	0.6372470	1.061262
FTE Elo	0.6369762	0.640154	1.073260
BCM	0.6673244	0.5970870	1.002466

In general, the BCM model demonstrated superior performance compared to the Elo models, achieving high accuracy, with low bias indicated by the log loss, and near-perfect calibration.

4.3 Discussion

Overall, our models performed as expected in relation to one another, where more complicated models would provide some improvement over the simpler one.

The naive model, a simple and easy to implement method, served well as the base line for our comparison. Its simplicity allows for quick implementation and minimal computational resources. However, the simplicity also limited its predictive power.

Despite being a more complex model, the logistic regression model did not yield significant improvements over the naive approach. Although there were slight enhancements in log loss and calibration, which indicating reduced bias in the model, its accuracy only marginally improved to 61.38%, compared to the naive approach's 61.34%. This limited improvement may be due to the fact that a player with higher ranking

points typically has a higher ATP ranking, making the predictions of both models quite similar.

Standard Elo brings the first step up in predictive capability, with an accuracy of 64.42% in testing, combined with reduced log loss, indicating a reduction in overconfident predictions. However, as a more complex model, the Elo system demands a substantial amount of historical data and high computational resources, especially during the tuning process. Nevertheless, the model can be used effectively over an extended period before needing to be re-tuned.

While the Surface Elo only brought in marginal improvements, the Combine Elo approach was promising, with improvement across all metrics, yielding better predictive power at over 65% with lower overconfident issue and biases. However, being a combined approach from two other models, it required higher level of computational power for 2 different tuning and fitting processes.

The Glicko model performed somewhat similar to the Standard Elo. However, this might be due to the fact that initial and minimum / maximum RD has not been optimised for this model. Further optimisation for these parameters can be consider for future work.

The FTE model represents another significant leap in predictive power for Elo system variations with 65.15% accuracy, but it also presents challenges. While it enhances predictive accuracy, the model shows increased bias toward winning matches, as indicated by higher log loss and calibration values. This model's tuning requirements are also more extensive, as it requires adjustments across several metrics, and the number of tuning runs required grows exponentially with the addition of each new metric, demanding more computing resources and time for optimisation.

Overall, our results suggest that to achieve better predictive power, the Elo model often tends to exhibit a bias toward winning matches. For a deeper understanding of the main Elo models (Standard and FTE), we divided the predictions based on player Elo rating quartiles, with the results shown in Table 4.

Table 4: Standard and FTE Elo comparison by Elo rating Quartile, both the models performed best at the first and forth quartile.

Model	Accuracy		Log loss		Calibration	
Quartile	Standard	FTE	Standard	FTE	Standard	FTE
Q1	0.6982	0.7064	0.5837	0.5905	1.0797	1.0857
Q2/3	0.6505	0.6549	0.6348	0.6464	1.0532	1.0897
Q4	0.6756	0.6976	0.6292	0.6634	1.0642	1.1493

The results reveal that our Elo models demonstrated superior predictive power in the high and low quartiles, while struggled in the midsection. Both models performed comparably across the top three quartiles, with the FTE model offering less than a 1% improvement in accuracy.

The main competitive edge of the FTE model likely comes from its performance in lower quartile matches, which we hypothesise is attributed to its rapid adjustment during a player's initial matches. The log loss of 0.66 and calibration at over 1.14 in this quartile also indicate that in order to achieve this advantage, the FTE Elo model had to be significantly biased toward winning matches.

Apart from the model constructed from match result data, the BCM model stands in a league of its own, outperforming all Elo models across various metrics. Despite its simplicity in terms of calculation, the data needed for this model may not always be readily available, particularly for predicting a fantasy matchup. Additionally, the actual computations bookmakers undertake to provide odds can be quite intense. They might use their own rating systems and consider various factors such as players' recent performance (known as the 'hot hand'), live injury updates, and other relevant statistics.

4.4 Future goal

Our first future goals for improvement include fully optimising the Glicko model to enhance its predictive capabilities.

We can also consider incorporating covariates, with the idea of matches in more prestigious tournament such as the Grand Slam carry more value and have a greater impact on players' Elo ratings. The function $K_i(t)$ would be updated as:

$$K_i(t) = \tilde{K}_i(t)C(t)$$

This approach has actually been used in conjunction with the FTE model, where Morris, Bialik, and Boice (2016) used a covariate of 1.1 for the four Grand Slam events.

Another approach involves incorporating in-game set win/loss statistics, with one possible method being the use of margin of victory as an extension to the Elo rating system (Kovalchik 2020) [3].

One of the first margin of error approach is linear, wherein a match outcome $M_{ij}(t)$ for player i and j at time t does not simply remain at 1 or 0, but follow the actual set result, for example it may be 0.66 if player i won the game 2-1.

In this instance, the predicted match outcome $\hat{M}_{ij}(t)$ is given by:

$$\hat{M}_{ij}(t) = \frac{E_i(t) - E_j(t)}{\omega}$$

where ω is the tuning parameter, and the players' rating after the match would be update as

$$E_i(t+1) = E_i(t) + K_i(t)[(M_{ij}(t) - \hat{M}_{ij}(t)]$$

This extension enhances predictions by considering the margin of victory or defeat, offering a closer understanding of match outcomes. Since it captures the degree of one player's outperformance over another, it hopefully can help us distinguishing between closely contested matches and one-sided victories and adjust the players' rating accordingly.

This discussion marks the conclusion of our project report. In summary, the discussed models show a spectrum of simplicity to complexity, each with its unique strengths and limitations. Beside improving the existing models and exploring new ones, we might consider adding another goal of finding the balance between predictive capability and practicality for real-world implementation of these models.

5 Conclusion

In conclusion, our study has examined a range of models for predicting tennis match outcomes, spanning from simple to complex methodologies. For our Elo models, while the complex version tend to offer greater predictive power, we observed a tendency for these more accurate models to show bias toward winning matches. Additionally, complex Elo system such as the FTE may require significant computing resources to implement effectively.

The BCE stood out as a superior performer among the models examined. However, it relies on information provided by bookmakers, essentially using predictions derived from their own systems. This dependency on external data might limit its accessibility and applicability, as it necessitates access to bookmaker odds data for accurate predictions.

Moving forward, there are many interesting path for further exploration. Refinements to existing models, particularly in terms of addressing biases and improving computational efficiency, could enhance their predictive capabilities. Additionally, exploring novel approaches and integrating additional factors, such as margin of victory and event prestige, could enhance the predictive process and lead to more accurate forecasts.

Acknowledgements

We would like to extend our sincere appreciation to Dr. Andrew Black for his invaluable guidance and support throughout this project. His expertise and insights have been instrumental in shaping our research efforts. Additionally, we gratefully acknowledge the School of Mathematical Sciences for providing access to some of the necessary materials for this project. Their support has been essential to the completion of our work.

References

[1] Mark E. Glickman. The Glicko System. *Mark Glickman's World*, 2016.

[2] P. Gorgi, S. J. Koopman, and R. Lit. The Analysis and Forecasting of Tennis Matches by using a High Dimensional Dynamic Model. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(4):1393–1409, 04 2019.

[3] Stephanie Kovalchik. Extension of the Elo rating system to margin of victory. *International Journal of Forecasting*, 36(4):1329–1341, October 2020.

[4] Christoph Leitner, Achim Zeileis, and Kurt Hornik. Is Federer Stronger in a TournamentWithout Nadal? An Evaluation of Odds and Seedings for Wimbledon 2009. *Austrian Journal of Statistics*, 38(4), April 2016.

[5] Benjamin Morris. Serena Williams And The Difference Between All-Time Great And Greatest Of All Time, August 2015.

[6] Benjamin Morris. How We're Forecasting The 2016 U.S. Open, August 2016.

[7] Jeff Sackmann. ATP Tennis Rankings, Results, and Stats, September 2023.

[8] Tennis-Data. Tennis Betting, Tennis Results & Tennis Live Scores.

[9] Leighton Vaughan Williams, Chunping Liu, Lerato Dixon, and Hannah Gerrard. How well do Elo-based ratings predict professional tennis matches? *Journal of Quantitative Analysis in Sports*, 17(2):91–105, June 2021.