Data Science Research Project A Report

Minh Trung (James) Vo | Student No. a1869086

2023-11-25

Report submitted for Data Science Research Project A at the School of Mathematical Sciences, 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.

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 will try to build and critically evaluate different approaches to predict tennis matches outcome. The data being used is maintained by Jeff Sackmann, which contains players' information and ATP matches statistics from 1968 to 2023 (Sackmann 2023). We will use the match data from 1968 to 2018 to build 3 probabilistic models: a naive model and a logistic regression model base on players' ATP ranking, and an Elo system based on match results. These 3 models will then be used to predict the outcomes of tennis matches in 2019, and they will be evaluated based on their prediction performance compared to actual results.

Method

The data

Our dataset, comprising 191,300 matches recorded between 1968 and 2023, was compiled and is maintained by Jeff Sackmann. 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 rank points of both the winner and loser at the time. However, player rankings only become available from 1973 onward, with rank points and serve statistics introduced in 1990. Our model will be constructed using data from 1968 to 2018, and data from 2019 to 2023 will be used for testing and validation. Figure 1 highlights the ranking for some of the top players from 1990 to 2019.

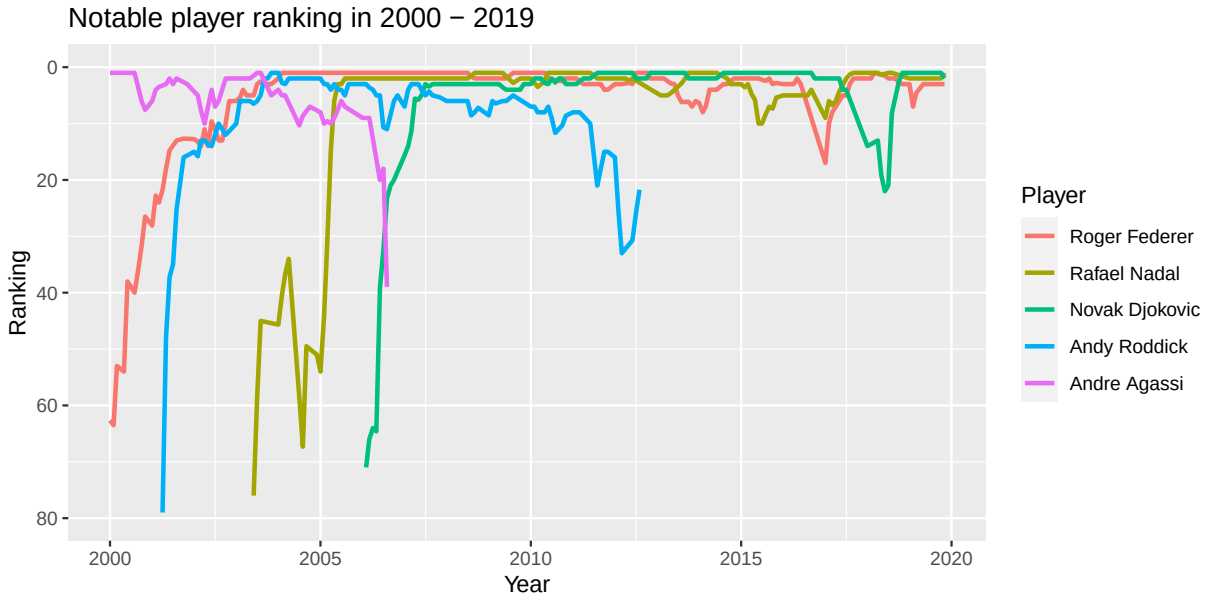

Figure 1: Notable player ranking from 2000. Federer, Nadal and Djokovic dominating in this period.

In the figure, we can observe the rise of the top 3 players in Federer, Nadal and Djokovic and their sustained domination throughout the year. 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.

The models

We deployed three distinct models to forecast match outcomes: a naive model based on ATP ranking points, a logistic regression model incorporating the difference in ranking points between players, and an Elo system constructed from players' match histories. We will discuss the mathematical properties of each model, their parameters and how we tuned each models here, while further discussion on models metrics and effectiveness will be presented in the Result section.

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 validate the model, we create a new Boolean variable to show whether the winner of the match has higher rank points.

Logistic Regression Model:

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

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

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 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 $\beta_1 = 5.768e^{-04}$.

Elo system:

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 relationship can be expressed as:

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

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

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

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 $K_i(t)$ defined the Elo class of models, where is specify the change in the Elo rating for player i. For our model, we have only explored K-factor model, in which the function remains a constant $K_i(t) = k$ for all i.

Our model was build in R, using the EloRating package (Neumann & Kulik 2015) and was tuned using a range of k from 1-100 on our training 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.

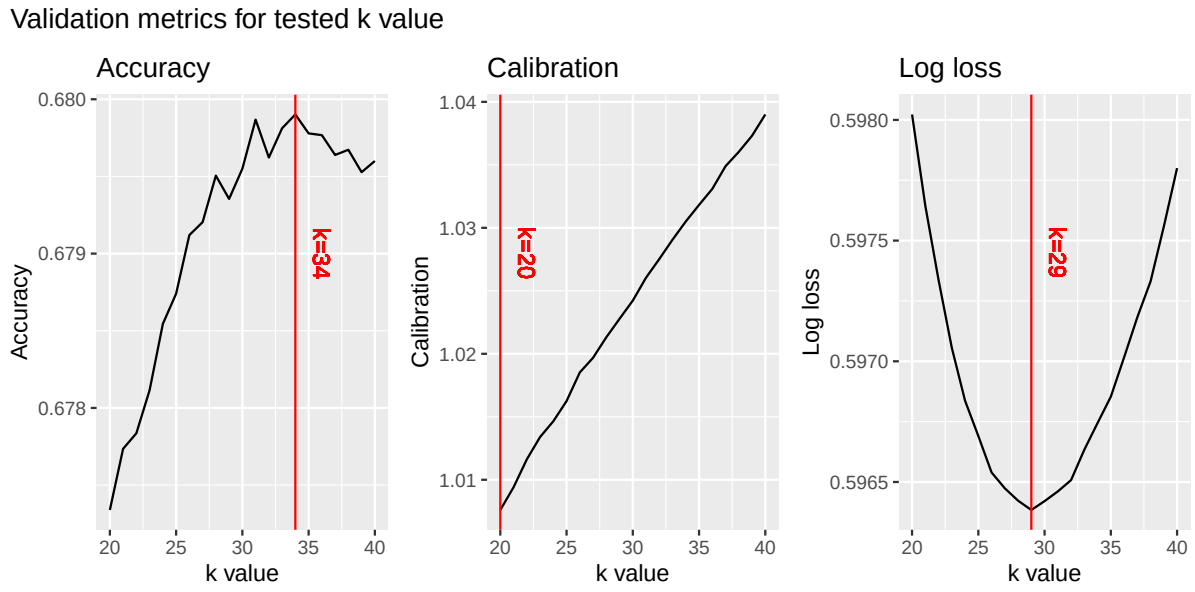

Figure 2: k value for each metric in training. Optimal value for accuracy and log loss is 34 and 29 respectively, while calibration increase with k

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.

Elo for each surface type:
This Elo system is built on the premise that player may excel on a specific court type, with the most notable example of Rafael Nadal's remarkable win rate of over 91% on clay courts (Wikipedia, n.d.).

For this Elo system, we partitioned our data based on each court type and run the model for each data set, exploring a range of k values from 1 to 100 for each court type. The optimal k for each court type is 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
Carpet	42
Clay	40
Grass	56
Hard	35

Combine Elo system:
This approach was employed by Leighton (2021) in their paper and referred to as Adjusted Elo. The Elo rating for players is represented by the function:

$$CombineElo = \lambda * StandardElo + (1 - \lambda) * SurfaceElo$$

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.

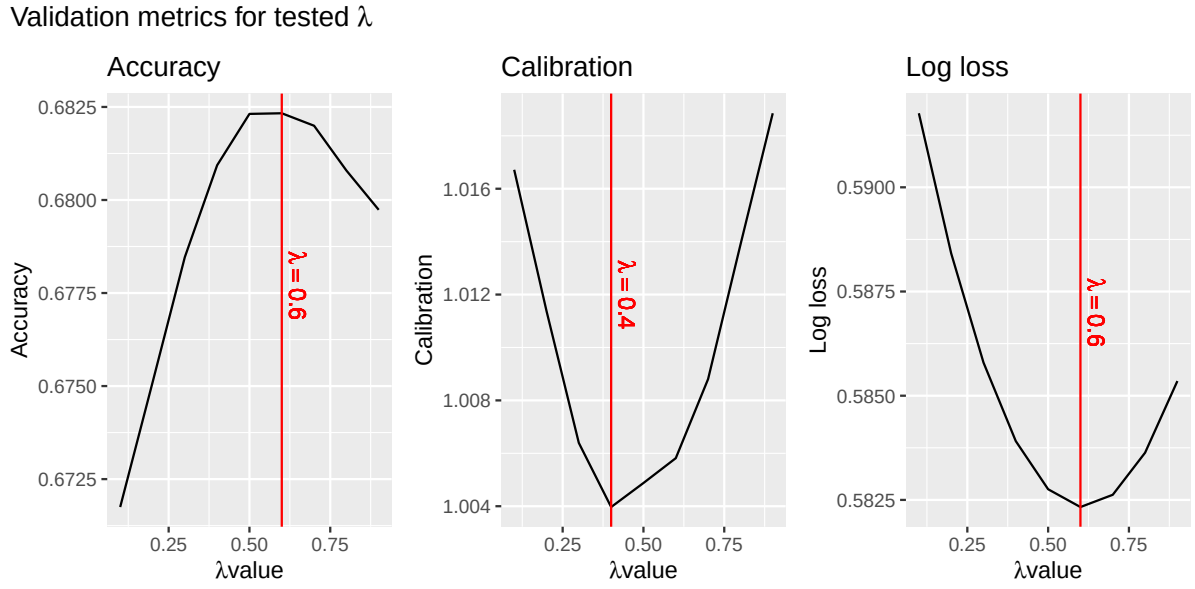

Figure 3: Tested λ for each metric. Optimal value for calibration and log loss is at 0.4, while maximum accuracy achived at 0.6

In this case, we selected $\lambda = 0.5$ for our model as a compromise between accuracy and bias.

Validation

All 3 models will be assessed based on their performance on the testing data in 2019, using 3 different metrics: accuracy, calibration, and log loss. While the naive and logistic regression model does not need to be tuned, we will tune the Elo system k factor using the same 3 metrics.

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 model 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 (an 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.

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

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

where is the total number of game where the higher ranked won, and π_i is the probability of the higher ranked player winning. For a well calibrated model, $C \approx 1$. When $C > 1$, the model over

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.

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

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

with $y_i = 1$ if the higher ranked player won game i'th, and $y_i = 0$ otherwise.

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.

Results

Elo systems behaviour:

Overall, our Elo system behaved quite comparable to the ATP ranking. Figure 4 shows the Elo rating of the same players in Figure 1.

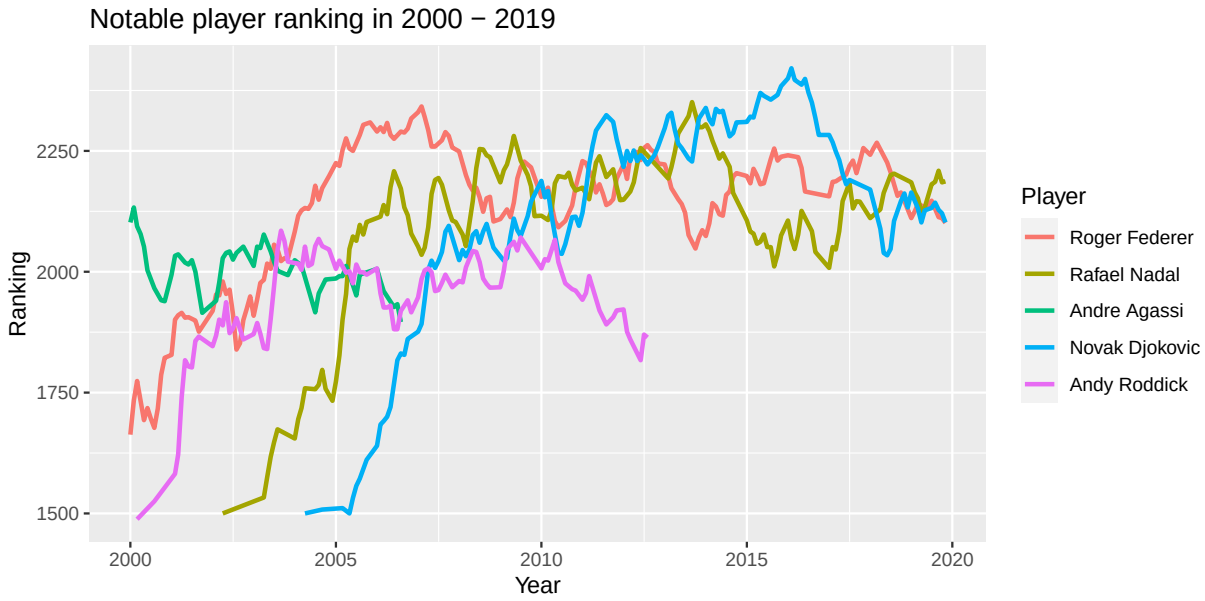

Figure 4: Notable player Elo rating from 2000. While Federer, Nadal and Djokovic is still dominating, their ratings are all dropped from their respective peaks.

While Figure 4, 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), 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.

Models validation metrics:

While training data metrics are not as important as testing metrics, there are some notable results that can be highlighted. First of all, the logistics regression model lags in performance across most metrics compared to other models, only marginally outperformed the naive model in log loss. Among the advanced models, the Standard Elo system showed a significant improvement over the naive and logistic regression models in all metrics. On the other hand, the Surface Elo system did not offer any significant advantage over the Standard Elo. However, the Combine Elo showed promise with a modest improvement across all 3 metrics. Metrics for all models in the training data set are detailed in Table 2.

Table 2: Validation metrics on training data. The logistic regression performed poorly in training, while all 3 Elo systems showcased favorable metrics, with improvement in accuracy and a more balanced calibration and log loss.

Model	Accuracy	Log loss	Calibration
Naive Model	0.6562330	0.6435025	1.0000000
Logistic Regression	0.6465307	0.6365532	0.9273310
Standard Elo	0.6798680	0.5964609	1.0260467
Surface Elo	0.6707217	0.5958786	0.9934444
Combine Elo	0.6823133	0.5820621	1.0042556

Turning to the focus of our project, which is the models' performance in our testing data set (matches in 2019), a clear trend emerges 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.

In testing, the Standard Elo system maintains a respectable improvement in both accuracy and log loss compared to the naive and logistic regression models, 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. Finally, the Combine Elo came out as the best performing model with improvements across all metrics over the other 2 Elo systems.

Table 3: Validation metrics on testing data. As expected, the performance of all 5 models declined, however the Elo systems still retains a significant lead across all metrics, except for calibration where logistic regression model have a small edge.

Model	Accuracy	Log loss	Calibration
Naive Model	0.6134328	0.6711751	1.069772
Logistic Regression Model	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

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%.

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. This limitation may hinder the model's practicality for real-time predictions or applications with resource constraints.

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

Overall, all 3 Elo system showed reasonable accuracy, without being overconfident, as reflected by the low log loss. However, all 3 systems have higher calibration compared to the logistic regression model, which suggest a bias toward matches won by the higher ranked player.

In this case, our goal moving forward is to improve our model by integrating the FiveThirtyEight Elo model with dynamic $K_i(t)$ defined as:

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

with ($m_i(t)$ is the number of game played until time t, δ, ν and σ are tuning parameters.

Another approach involves utilising covariates, with the idea of matches in higher tournament like the Grand Slam carry more value and have a greater impact on players' Elo ratings. The function $K_i(t)$ will then be updated as:

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

The $C(t)$ covariate can then be set to > 1.0 for top tier tournaments, or reduced to < 1.0 for less reputable ones.

Figure 4 shows how k value is adjusted through time for different $K_i(t)$ function.

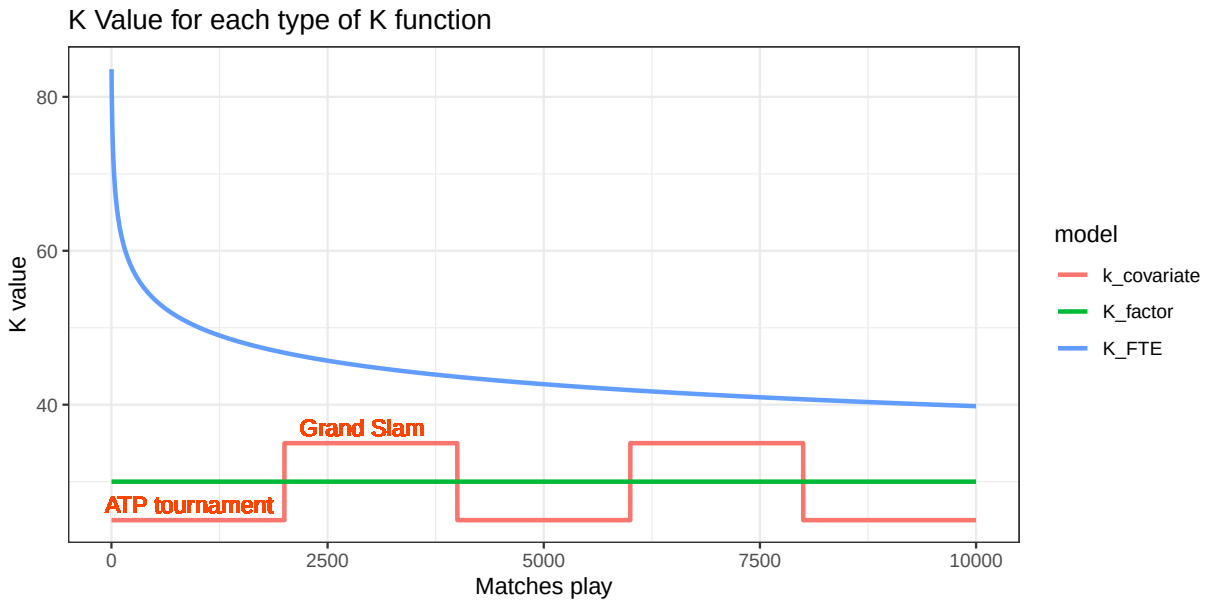

Figure 5: k value through time, with k for FiveThirtyEight model decreasing over time and k for covariate model changing depends on tournament level.

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.

Reference:

Morris, B. & Bialik, C. 2015, "Serena Williams And The Difference Between All-Time Great And Greatest Of All Time", FiveThirtyEight, viewed 15 Oct 2023 https://fivethirtyeight.com/features/serena-williams-and-the-difference-between-all-time-great-and-greatest-of-all-time/

Neumann, C. & Kulik, L. 2015, Package 'EloRating', viewed 28 Oct 2023 http://cran.nexr.com/web/packages/EloRating/EloRating.pdf

Sackmann, J. 2023, ATP Tennis Rankings, Results, and Stats, GitHub, viewed 28 Sep 2023 https://github.com/JeffSackmann/tennis_atp

Leighton V.W. 2021, "How well do Elo-based ratings predict professional tennis matches?", Journal of Quantitative Analysis in Sports, 17(2), 91-105, viewed 15 Oct 2023 https://econpapers.repec.org/article/bpjjqsprt/v_3a17_3ay_3a2021_3ai_3a2_3ap_3a91-105_3an_3a6.htm

Wikipedia (no date), List of career achievements by Rafael Nadal, Wikipedia, viewed 28 Sep 2023 https://en.wikipedia.org/wiki/List_of_career_achievements_by_Rafael_Nadal