Predicting the Outcome of Tennis Matches
Final Reports

Thi Kim Oanh Nguyen
STUDENT NO. 1879781

August 10, 2024

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

Project Area: **PREDICTING THE OUTCOME OF TENNIS MATCHES**
Project Supervisor: **DYLAN MORRIS**

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.

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

Abstract

This study aimed to predict tennis match outcomes, a critical concern for coaches, players, fans, and betting companies. Our primary goal was to develop predictive models based on players' past performances, addressing the limitations of traditional ranking systems, such as those relying on Association of Tennis Professionals (ATP) points. We evaluated the performance of a naive model, logistic regression, Elo models (including a modified version), and Glicko models against a benchmark Bookmakers Consensus Model. Through this comparison, we optimised log loss, a key metric for betting companies to minimise their losses, and achieved the lowest log loss with the Elo K-factor model incorporating margin of victory. While this score did not match the benchmark model as closely as desired, the inclusion of margin of victory led to a significant improvement in the standard Elo models' performance. The findings of this study suggest future work in parameter optimisation and the development of more advanced Elo models, contributing to the enhancement of tennis match analysis and prediction methods.

1 Introduction

Tennis is a complex and captivating sport, where predicting match outcomes holds significant value for various stakeholders. Coaches and players can use these predictions to refine strategies and improve their chances of winning. For enthusiasts, predictions add an extra layer of excitement, while betting companies can use this information to set more accurate odds and minimise potential losses.

The Association of Tennis Professionals (ATP) ranking system serves as a baseline for assessing player performance. However, traditional ranking points may not fully capture the nuances of player matchups and recent form, highlighting the need for more sophisticated prediction models.

This project aims to predict tennis match outcomes based on players' past performances, focusing on whether higher-ranked players are more likely to win. We use open-source betting market data to apply various probabilistic models, including a naive model based on the win proportion of higher-ranked players, a logistic regression model using the difference in ATP points, and more advanced systems like Elo and Glicko, which generate their own rating scores.

To validate our models, we will benchmark them against a Bookmakers Consensus Model, an expert model that uses odds from betting companies [1]. Our goal is to optimise our models to achieve the lowest log loss, along with accuracy and calibration that closely match the benchmark model. This comprehensive approach contributes to the field of sports analytics and offers practical applications for enhancing decision-making processes in professional tennis.

2 Background

In our research, we evaluate the performance of various probabilistic prediction models for forecasting tennis match outcomes. Our focus is on prototyping and comparing five distinct models that vary in complexity and data utilisation.

The naive model uses simplified data, essentially measuring the proportion of matches won by higher-ranked players. In contrast, the logistic regression model incorporates more detailed data by using the difference in ATP ranking points between players. The Elo model, originally developed for chess, has been adapted for various competitive settings, including tennis. It creates a dynamic rating system that adjusts based on match outcomes and the relative strength of opponents. Modified Elo models further incorporate tennis-specific factors, such as the margin of victory (sets won), while the advanced Glicko rating system builds upon the Elo model by introducing additional parameters to account for rating reliability and performance variability over time.

The key difference between these models lies in their complexity and approach to data utilisation. While simpler models rely on straightforward metrics, more complex models create new rating systems to better capture the intricacies of tennis performance.

Understanding the dynamics of tennis is crucial for developing effective prediction models. Factors such as the margin of victory can provide valuable information about a player's dominance in a match [3]. The Glicko model, in particular, addresses some limitations of the standard Elo system by incorporating rating reliability and performance variability, aiming for a more accurate representation of player performance [6].

3 Methods

3.1 Data

We utilised open-source ATP men's tour betting market data from tennis-data.co.uk to run our models [5]. We developed the Bookmakers Consensus Model (BCM) using the same data sources as our models to establish a benchmark for comparison. We randomly selected data from 2013 to 2023 and used odds from Bet365 and Pinnacle Sports, as our dataset lacked odds data from other bookmakers from 2019 onwards.

To address missing player ranks, we assigned a value of 100,000, assuming that players with no ranks can be considered very low ranked. We calculated the point difference between higher and lower-ranked players, regardless of match outcomes, and used this difference in our logistic regression model. Missing data in the point difference column was removed to ensure data integrity. Additionally, we created a binary outcome variable to indicate whether the higher-ranked player won the match.

Matches classified as walkovers or ending due to retirement were excluded, as they do not accurately reflect the relative strength of the opponents. This exclusion particularly benefits the Elo and Glicko models. We used data from the year 2022 as test data, while the remaining data was used for training the models to prevent overfitting.

3.2 Models

3.2.1 Naive Model

We start with a naive model to establish a baseline. This model calculates the proportion of matches won by higher-ranked players across the dataset. Its simplicity assumes that higher-ranked players always have a better chance of winning, overlooking factors like lower-ranked player improvements or situational contexts. This makes it less accurate than more advanced models that consider specific match data.

The probability of the higher-ranked player winning is calculated as:

$$P(\text{win}) = \frac{\sum_{i=1}^{N} \text{higher\_rank\_won}_i}{N}$$

Where $\sum \text{higher\_rank\_won}_i$ is the sum of matches where the higher-ranked player won, and N is the total number of matches.

3.2.2 Logistic Regression Model

Building on the naive model, we implement a logistic regression model that incorporates the difference in ATP ranking points between players.

This model provides probability estimates for match outcomes, differentiating between matches with varying degrees of ranking disparity. While it still tends to favor higher-ranked players, it offers more nuanced predictions by considering the point difference between players.

The probability of the higher-ranked player winning is calculated as:

$$P(\text{win}) = \frac{1}{1 + \exp(-\beta \cdot \text{point\_difference})}$$

Where β is a coefficient determining the sensitivity of the winning probability to the difference in scores.

3.2.3 Elo Model

The Elo model represents a significant advancement in our prediction methodology compared to traditional ranking systems and our previous models. It creates a new and dynamic ranking system for players, estimating the win probability using the difference between the Elo ratings of the two players[3]. This difference is scaled by a factor of 400 to ensure the probability falls within a meaningful range (between 0 and 1):

$$P(\text{win}) = \frac{1}{1 + 10^{(\text{ELO}_{\text{loser}} - \text{ELO}_{\text{winner}})/400}}$$

Player ratings are then updated based on the outcome and the opponent's strength. The update formulas are as follows:

$$\text{ELO}'_{\text{winner}} = \text{ELO}_{\text{winner}} + K \cdot (1 - P(\text{win}))$$

$$\text{ELO}'_{\text{loser}} = \text{ELO}_{\text{loser}} + K \cdot (0 - (1 - P(\text{win})))$$

Initial ratings are assigned to all players, and following each match, ratings are updated according to the pre-match ratings of both players and the match outcome, along with a k coefficient, which determines the magnitude of rating changes. Winners' scores increase by k multiplied by the probability that they would have lost, while losers' scores decrease by k multiplied by the probability that they would have won.

The Elo model demonstrates that wins against higher-rated opponents result in larger rating increases than victories over lower-rated players. This allows player ratings to evolve based on recent performance, providing a more current representation of skill. Moreover, this model considers the strength of the opposition in each match, not just the outcome. The standard Elo model is referred to as the Elo K-factor model.

The Elo 538 model is an enhanced version of the traditional Elo system, introducing a dynamic k coefficient to address some limitations of the original model:

$$K(\text{matches\_played}) = \frac{\delta}{(\text{matches\_played} + \nu)^{\sigma}}$$

Where the parameters δ (delta), ν (nu), and σ (sigma) play crucial roles. Delta controls the overall magnitude of the k coefficient. Higher delta values will lead to larger changes in Elo scores after each match. Nu is a stabilising parameter that prevents k from becoming too large when the number of matches is small. A higher value of nu causes k to decrease more slowly as the number of matches increases. Sigma controls the rate at which k decreases with the number of matches played. A higher value of sigma causes k to decrease faster as the number of matches increases.

The ratings are updated as follows:

$$\text{ELO}'_{\text{winner}} = \text{ELO}_{\text{winner}} + K(\text{winner\_matches}) \cdot (1 - P(\text{win}))$$

$$\text{ELO}'_{\text{loser}} = \text{ELO}_{\text{loser}} + K(\text{loser\_matches}) \cdot (0 - (1 - P(\text{win})))$$

The Elo 538 model's k coefficient varies based on the number of matches played by each player. This reflects increased rating stability as players accumulate more matches, allowing the model to better handle players with varying amounts of historical data. Each player's rating changes based on their own match history rather than a fixed global parameter. However, the model's performance depends on the appropriate selection of delta, nu, and sigma values and increases computational requirements compared to the standard Elo model.

3.2.4 Margin of Victory

Incorporating the margin of victory enhances the Elo system by reflecting the dominance of wins in tennis matches[3]. The margin of victory in tennis is defined as the difference in the number of sets won between the winner and the loser:

$$\text{margin} = W_{sets} - L_{sets}$$

Here, W_{sets} and L_{sets} represent the number of sets won by the winner and the loser, respectively. To account for this, the k factor is adjusted based on the margin of victory. This can be expressed in the Elo K-factor and Elo 538 models, where m is a margin factor, as follows:

$$k = k + m \cdot \text{margin}$$

$$k = \frac{\delta \cdot m \cdot \text{margin}}{(\text{matches\_played} + \nu)^{\sigma}}$$

A larger difference in the number of sets won indicates a more dominant win. Winning with a larger margin results in a more significant Elo change, offering a detailed representation of player performance. This method recognises close wins and overwhelming victories distinctly, providing a more accurate reflection of player skill and match dynamics. Thus, it accounts not only for win/loss results but also for the degree of victory, enhancing the model's ability to predict match outcomes. However, the effectiveness of this approach depends on the appropriate selection of the parameters m, delta, nu, and sigma.

3.2.5 Glicko Model

The Glicko model builds upon the Elo system by introducing a measure of rating reliability to address uncertainty in player performance[6]. In the Glicko model, the probability of winning, is calculated similarly to that in Elo, but with adjustments through the g(RD) function:

$$P(\text{win}) = E = \frac{1}{1 + 10^{g(r - r_{opponent})/(-400)}}$$

The player's new rating is calculated as follows:

$$r' = r + \frac{q}{1/RD^2 + 1/d2} \cdot g \cdot (outcome - E)$$

Where:

$$q = \frac{\ln(10)}{400}$$

$$d_2 = \frac{1}{q^2 \cdot g^2 \cdot E \cdot (1 - E)}$$

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

In these equations, r' and r are the new and old ratings of the player. RD is the rating deviation, which measures the reliability of the score. A lower RD indicates a more reliable score. *outcome* represents the match result (1 for win, 0.5 for draw, 0 for loss). $g(RD)$ is a function that reduces the impact of matches against opponents with high RD (uncertain scores). q is a logarithmic constant that converts between Elo and Glicko scales. E is the expected probability of the player winning over their opponent.

Glicko updates the RD after each review period:

$$RD' = \sqrt{\frac{1}{1/RD^2 + 1/d2}}$$

Additionally, Glicko adjusts RD over time:

$$RD_{new} = \sqrt{(RD')^2 + c^2 t}$$

Where c is the volatility constant that controls the rate at which RD increases over time, and t is the time (usually in rating periods) since the last RD update.

Compared to the Elo model, Glicko adds the concept of RD, which allows for the evaluation of the reliability of the rating. RD increases over time when the player is inactive, reflecting increased uncertainty. This accounts for periods of inactivity due to injuries or breaks. Glicko uses the function g(RD) to reduce the impact of matches against opponents with uncertain ratings. However, it also increases complexity and requires careful parameter tuning for c and t.

3.2.6 Bookmaker Consensus Model (BCM)

The Bookmaker Consensus Model (BCM) serves as a benchmark for our predictive models by leveraging expert knowledge embedded in betting odds[4]. The BCM is defined as follows:

$$P_W = \frac{1}{N} \sum_{i=1}^{N} \frac{\beta_i}{\alpha_i + \beta_i}$$

$$P_L = \frac{1}{N} \sum_{i=1}^{N} \frac{\alpha_i}{\alpha_i + \beta_i}$$

Here, P_W represents the average implicit probability of winning, and P_L represents the average implicit probability of losing. The terms α_i and β_i denote the odds for the winner and loser from the i^{th} bookmaker, respectively, and N is the number of bookmakers (in this case, B365 and PS).

$$P_{W,norm} = \frac{P_W}{P_W + P_L}$$

$$P_{L,norm} = \frac{P_L}{P_W + P_L}$$

The probabilities $P_{W,norm}$ and $P_{L,norm}$ are the normalised probabilities of winning and losing.

$$P_{BCM} = P_{W,norm}$$

The final probability according to the BCM is denoted as P_{BCM}.

The BCM is used as a benchmark rather than for prediction purposes because we do not have access to betting odds until after the games have been played. The BCM relies on data from bookmakers and requires odds for each match, which may not always be available. Furthermore, the exact methods used by betting companies to calculate these odds are proprietary and not accessible in advance.

3.3 Validation

The following metrics are used to evaluate the performance of predictive models:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(y_i = \hat{y}_i)$$

Accuracy measures the proportion of correct predictions out of the total number of samples. Accuracy ranges from 0 to 1, with higher values indicating better performance. However, it does not account for the confidence of the predictions.

$$\text{Log Loss} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]$$

Log loss quantifies the performance by penalising false predictions, especially those made with high confidence. Lower log loss values indicate better alignment between predicted probabilities and actual outcomes, with 0 being ideal.

$$\text{Calibration} = \frac{\frac{1}{n}\sum_{i=1}^{n} p_i}{\frac{1}{n}\sum_{i=1}^{n} y_i}$$

Calibration assesses how well the predicted probabilities reflect the actual event rates. A calibration value of 1 indicates perfect alignment. Values greater than 1 suggest overconfidence, while values less than 1 indicate underconfidence.

3.4 Optimisation

Our models were optimised with a primary focus on minimising log loss, a metric likely used by bookmakers to maximise their profits. To ensure a thorough evaluation, we also assessed accuracy and calibration in relation to this minimised log loss, which may result in a trade-off with achieving the highest possible accuracy.

We employed Bayesian optimisation to find the optimal parameter values for minimising log loss. For the Elo K-factor model, we optimised the K factor within a range of 1 to 50. In the Elo 538 model, we optimised the delta parameter between 100 and 500, the nu parameter from 5 to 50, and the sigma parameter from 0.1 to 1.0. For the Elo K-factor model with a margin of victory, we adjusted the k factor within the same range as the standard model and optimised the margin factor from 5 to 50. The Glicko model optimisation involved adjusting the volatility parameter from 1 to 100 and the rating periods from 0.01 to 10.

In the case of the Elo 538 model with a margin of victory using a complex k function, we constrained the parameter ranges to avoid extreme values that could lead to infinite log loss. Specifically, delta was optimised between 10 and 200, nu from 1 to 30, sigma from 0.01 to 0.5, and the margin factor from 0.5 to 10. To explore the search space more thoroughly and achieve better optimisation for this complex model, we increased the number of function evaluations from 50 to 100 and set the number of random initialisations to 20. This approach was intended to diversify starting points and avoid local minima.

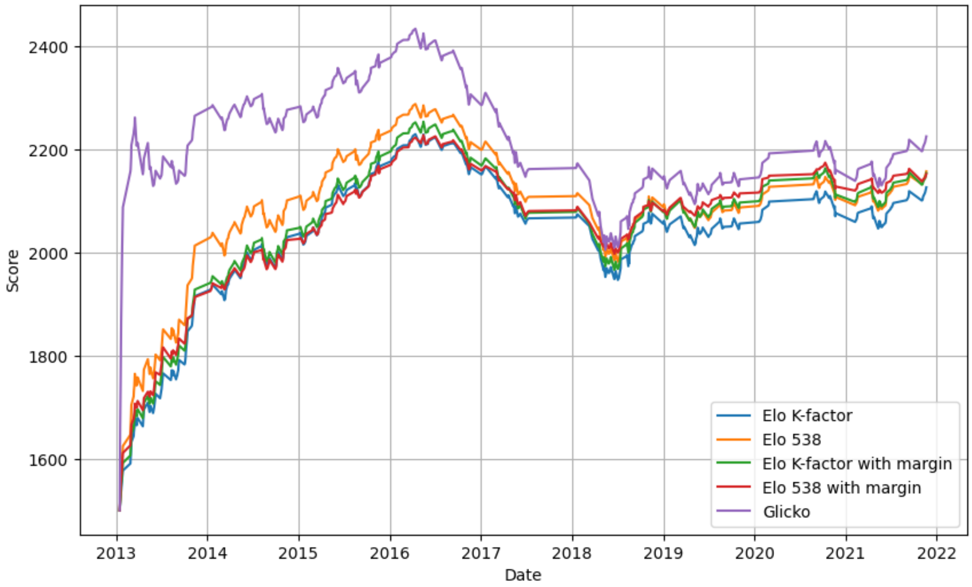

Figure 1: Novak Djokovic's rating scores from 2013 to 2022 using Elo and Glicko models.

4 Results

4.1 Behaviours

Figure 1 illustrates the change in Novak Djokovic's rating scores from 2013 to 2022 using Elo and Glicko models. The overall trends in player rankings are similar across the models, though there are noticeable differences in the detailed movement, especially between the Glicko scores and all the Elo-based scores. All models capture Djokovic's period of dominance from 2015 to 2016, his decline due to injury, and his subsequent comeback after 2018, accurately reflecting his career trajectory.

The Glicko model shows significantly higher scores than the Elo models from 2013 to around 2018, peaking above 2400, compared to over 2200 for the Elo models. After 2018, Glicko scores become more aligned with those of the Elo models. The scores from the various Elo models are very close to each other, which explains the similar log loss and accuracy results among these models (as shown in the validation results).

4.2 Validation Results

Figures 2 and 3 present the final results for all models with optimised parameters. The optimal parameter values are as follows: for the Elo K-factor model, the best k value is 25.10; for the Elo 538 model, the

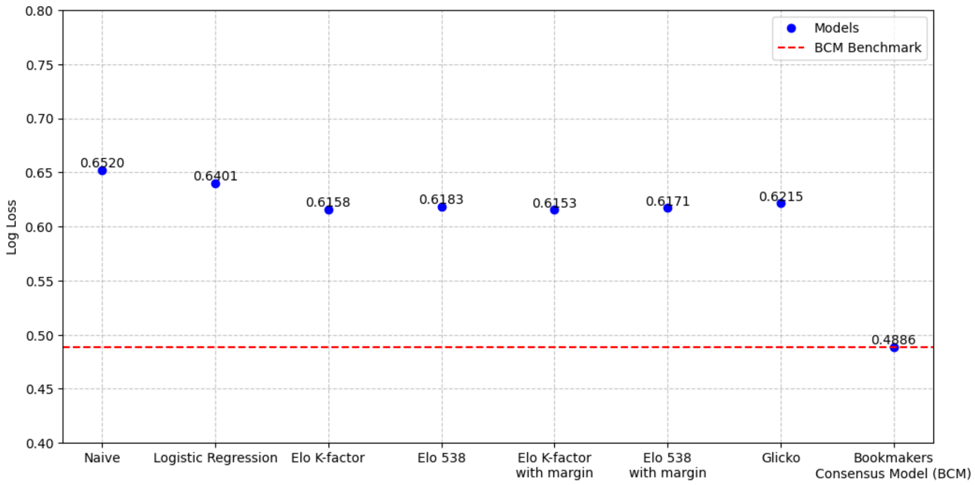

Figure 2: Log loss of different models. The blue dot indicates the log loss of each model, the red dashed line is the benchmark.

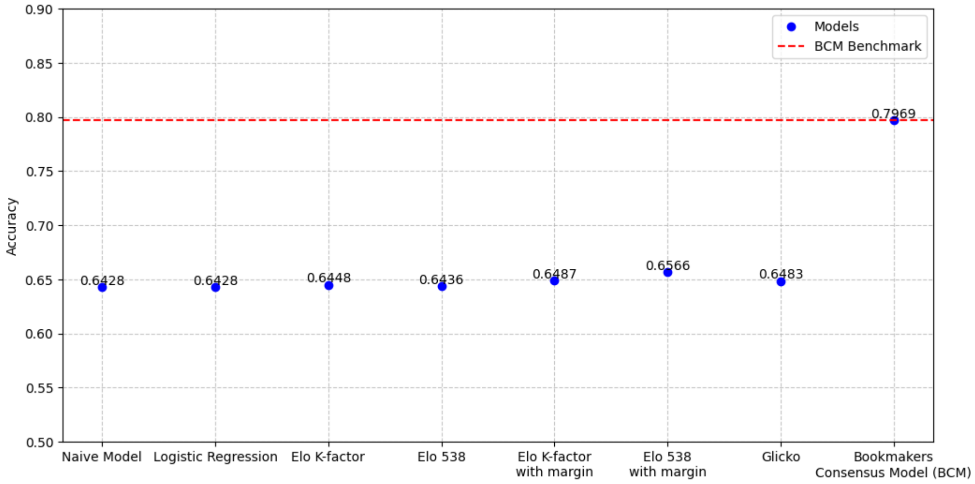

Figure 3: Accuracy of different models. The blue dot indicates the accuracy of each model, the red dashed line is the benchmark.

optimal parameters are a delta of 100, nu of 50, and sigma of 0.23. The Elo K-factor model with the margin of victory has an optimal k of 14 and a margin factor of 6.35. The Elo 538 model with the margin of victory achieves the best delta of 10, nu of 1, sigma of 0.06, and a margin factor of 1.55. The Glicko model has the best volatility parameter set to 100 and the optimal rating period at 0.01.

Overall, all models achieve relatively close log loss and accuracy. The Elo K-factor model with the margin of victory has the lowest log loss at 0.6153, followed closely by the Elo 538 model with margin of victory at 0.6171. In terms of accuracy, the Elo 538 model with margin of victory achieves the highest result of 0.6566, followed by the Elo K-factor model with margin of victory at 0.6487. This indicates that incorporating the margin of victory has enhanced both log loss and accuracy for the Elo models.

The models also demonstrate good calibration, with scores close to 1: naive model achieves 1.0172, logistic regression achieves 0.9533, Elo K-factor achieves 0.9835, Elo 538 achieves 0.9967, Elo K-factor with margin of victory achieves 0.9911, Elo 538 with margin of victory achieves 0.9885, and Glicko achieves 0.9923.

However, the Bookmakers Consensus Model (BCM) serves as a benchmark with substantially better performance, showing a log loss of 0.4886 and an accuracy of 0.7969. While our models did not match the BCM's performance, incorporating the margin of victory significantly improved the Elo models. Specifically, the log loss decreased from 0.6158 to 0.6153 for the Elo K-factor model and from 0.6183 to 0.6171 for the Elo 538 model. Accuracy also improved significantly, with the margin of victory models achieving the highest results of 0.6487 and 0.6566, respectively.

5 Conclusion

The Elo K-factor model with the margin of victory achieves the lowest log loss, indicating its strong performance in this metric. Meanwhile, the Elo 538 model with the margin of victory has a slightly higher log loss but demonstrates significantly better accuracy. This suggests that expanding the search space for parameter optimisation is particularly beneficial for the more complex Elo 538 model. Future work could involve further optimising parameters and exploring additional effective factors, such as surface[2], to enhance both log loss and accuracy.

The Glicko model, while simpler to optimise due to its straightforward implementation without a complex k function, faces challenges due to its reliance on rating deviation and rating period. Selecting an appropriate rating period is critical; if it is too large or too small, it can adversely affect the results[6]. The Glicko model may not be performing optimally due to this sensitivity. Future improvements could involve more precise optimisation of these parameters or investigating new approaches to enhance the model's performance. There may be limitations in our methods that could cause the models to converge on local minima rather than achieving the global minimum. To address this, more robust parameter tuning and exploration techniques could be necessary.

In summary, this study compared the performance of tennis match outcome prediction models, including a naive model, logistic regression, and Elo and Glicko models. We used the Bookmakers Consensus Model (BCM) as a benchmark. The models were optimised using Bayesian optimisation. The results show that the Elo K-factor model with the margin of victory factor has the lowest log loss, demonstrating that adding the margin of victory factor significantly improves the performance of the Elo models. Future studies should focus on further parameter optimisation and exploring new factors to improve the prediction performance.

Acknowledgements

I would like to express my appreciation to my supervisor, Dylan Morris, who provided valuable insights, instructions, code advice, and moral support for this work.

A Appendices

The complete code for this project is available at: https://github.com/oanhnguyenolivia/TennisOutcomePrediction/blob/main/Final%20Codes%20for%20Project%20B.ipynb.

References

[1] Buhamra, N, Groll, A Brunner, S 2024, 'Modeling and prediction of tennis matches at Grand Slam tournaments', *Journal of Sports Analytics*, vol. 10, no. 1, pp. 17–33. 2

[2] 'How well do Elo-based ratings predict professional tennis matches?' 2021, *Journal of Quantitative Analysis in Sports*. 14

[3] Kovalchik, S 2020, 'Extension of the Elo rating system to margin of victory', *International Journal of Forecasting*, vol. 36, no. 4, pp. 1329–1341. 3, 5, 6

[4] Leitner, C, Zeileis, A Hornik, K 2016, 'Is Federer Stronger in a Tournament Without Nadal? An Evaluation of Odds and Seedings for Wimbledon 2009', *Osterreichische Zeitschrift Fur Statistik*, vol. 38, no. 4. 8

[5] tennis-data.co.uk (2024) *n.d.* Available at: http://www.tennis-data.co.uk/alldata.php (Accessed: 10 August 2024). 4

[6] Yue, JC, Chou, EP, Hsieh, M-H Hsiao, L-C 2022, 'A study of forecasting tennis matches via the Glicko model', *PloS One*, vol. 17, no. 4, pp. e0266838–e0266838. 3, 7, 14