Modelling the Outcome of Tennis Matches
Interim Report

S. Gilbert
A1737770.

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: **Data Science - Modelling**
Project Supervisor: **Dylan Morris**

Abstract

Tennis is a multi-billion dollar industry with the outcome of both tournaments and individual matches having both enormous monetary ramifications' and impacts for player legacy. Accurately predicting the outcome of tennis matches underpin the tennis bookmaking industry. This leads to profit incentives both on the side of the bookmakers and on betters looking to "beat the odds". This report investigates various mathematical modelling methods to evaluate what is the best performing model when built leveraging past performance to determine future outcomes.

By leveraging open-source datasets of historical tennis data, mathematical models can be built to predict the outcome of tennis matches. These models will use a subset of the available data initially to form a base performance for these models. Future work will be outlined to incorporate more features of the dataset to improve the overall performance of the model. This performance was measured using accuracy, which is the proportion of guesses that were correct and log loss, which aims to measure how confident the model is in its prediction.

The models investigated ranged from a simple logistics model through to several Elo based models. The simpler logistic model use more traditional statistical analysis methods for deriving model parameters whereas the Elo models rely on predefined hyperparameters to influence the performance of the model. It can be observed that the simpler models, ie logistic model or simply predicting the higher ranked player outperforms the Elo models when they use unoptimised hyperparameters. The log loss however tells a different story, although these simpler models have a higher accuracy currently, the Elo models have a substantially lower log loss value implying that the future work of tuning the hyperparameters of the models should result in greatly improved prediction accuracy. The simpler models however cannot be meaningfully modified to improve accuracy. Potential improvements to the various models are explored including expanding the parameters used for the logistic model and hyperparameter tuning for the various Elo models.

1 Introduction

Using historical data to predict future events is extremely important across many areas/fields. A prime instance of this is predicting winners of tennis matches at a professional level. With sponsorships, player legacy and a multi-billion-dollar gambling industry relying on the outcomes of each game, being able to predict the outcome of a game more accurately is of extreme importance. Currently, the Association of Tennis Professionals (ATP) and the Women's Tennis Association (WTA) are responsible for ranking players based on performance at applicable grand slams [6]. Theoretically the highest ranked player should beat the lower ranked player. However, it can be observed that between 1968 and 2023 in the men's singles competition the higher ranked player wins approximately 66 percent of the time as seen in Figure 1.

Field	Value
Games Played	143860
Higher Rank	94896
Lower Rank	48964
Percentage	65.9641

Figure 1: Model Accuracy Using WTA Rankings

Implementing a mathematical model that can perform better translates into higher returns when matches are bet on or more certainty for companies when trying to predict forward for offering sponsorship opportunities. There are many approaches to mathematically model the outcome of tennis matches based on statistical analysis ranging from simplistic linear/logistical models to complex machine learning based models. This report investigated a subset of these models and identified which of these performed the best and what future work will be conducted to improve the overall performance of these models. Using the naive approach of always predicting the higher ranked player to win as a baseline the performance will be compared. Log loss, which measures how close the calculated probability of an event happening when compared to the actual outcome was used to identify how confident the various models are in its predictions. This log loss shows that despite a lower accuracy (how many correct guesses), when it does make correct predictions it may be more confident than a naive model which will allow for further tuning to occur to increase the accuracy.
Based on the results outlined in this report future work will be outlined to

improve the implemented models in order to achieve higher accuracy. As a part of this future work comparison of the implemented and optimised models against bookmaker predictions rating will be conducted.

2 Background

2.1 Tennis

Tennis is a two player head-to-head game in which players use a racquet to hit a ball repeatedly until one of them either hits the ball out of bounds or misses an in bound shot and the ball subsequently goes out of bounds. Professional tennis is played within the structure of a tournament which has players playing in rounds, getting knocked out until two players reach the final. Not all tournaments are the same, with some leveraging losers brackets and a soccer like point scoring however these nuances and tournaments are out of scope of this report.

Traditionally these tournaments have a two-sided draw which is designed to have an even split of the highest ranked players on either side. The intent of this is to have the higher ranked players play each other in later rounds of the tournament. The ultimate goal of the tournament organisers is to have the two highest seeds play each other in the final.

This report focuses on professional men's singles tennis which is overseen by the ATP. The ATP uses performance at sanctioned events in a rolling 52 week window to calculate a point value assigned to each player [8]. The rolling window is to remove the chance that a players rank will stay indefinitely if they stop playing for an extended period of time thus removing "stale" ranks. These point values are used to determine overall rankings for each professional player with the player with the highest total points being the "Number 1 Ranked" player in the world. These rankings are used to allocate seeding at tournaments and incentivise players to seek as high a seed as they can manage as it means an easier path to the later rounds. The players best twenty performances at tournaments contribute to their rating.

2.2 Existing Models

2.2.1 Logistic Model

The logistic model utilises a transform on predictors/features of the data set to determine the probability of the higher ranked player winning. For the initial implementation the difference in rank points will be simply the difference between the points of the higher ranked player and the lower ranked player. This model is defined as;

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

with β_0 and β_1 being fixed parameters. The probability can be found by inverting the equation as follows

$$\pi_i = \frac{1}{(1 + \exp(-(\beta_0 + \beta_1 D_i)))} \tag{2}$$

This model provides a probability based on the difference in rank points and as such the intercept term, or β_0, needs to be dropped otherwise it will be added to each probability which is not appropriate for this problem space.

2.2.2 Elo Model

The Elo Rating Algorithm is a commonly used method of ranking players in competitive games. Some examples of this include official Chess rankings and online Player vs Player (PvP) video games. Elo measures the probability of a higher ranked player beating a lower ranked player. Players are assigned an initial rating value which is then updated after each contest. This value will notionally be set to 1500. A larger difference in score increases the expected probability of the higher ranked player winning. This probability can be defined as follows

$$\pi_{i,j}(t) = (1 + 10^{\frac{E_j(t) - E_i(t)}{400}})^{-1} \tag{3}$$

and updating the players ranking is done as follows

$$E_i(t+1) = E_i(t) + K_i(W_i(t) - \pi_{i,j}(t)) \tag{4}$$

Where $W_i(t)$ is an indicator variable for whether the i'th individual won their t'th match. From the above it can be observed that if the higher ranked player beats the lower ranked player the rate of change will be lower than if the lower ranked player beats the higher ranked player. The K value impacts the rate of change of a player's ranking after a match. Different approaches for setting this K value will be explored next as it can have a major impact on the models outputs and therefore its performance.

2.2.3 K Factor Model

The most naive approach is to set the K value as a constant and scale the ranking update based on this constant this K value is defined as follows

$$K_i(t) = k \quad \forall i \tag{5}$$

2.2.4 Five Thirty Eight Model

The Five Thirty Eight model introduces some new factors for determining the K value. Calulating the K value for the Five Thirty Eight model is done as follows.

$$K_i(t) = \frac{\delta}{(m_i(t) + \nu)^{\sigma}} \tag{6}$$

δ, ν and σ are parameters that need to be tuned to find the optimal values. This optimisation is out of scope of this initial paper however will be investigated as future work.

2.2.5 Elo MMR Model

A novel approach to calculate a players ranking is a bayesian system resulting in an Elo-MMR (Massive, Monotonic and Robust) metric [2]. This model uses a gaussian diffusion to underpin its predictions whilst also provides capability for a logistic performance model which limits the impact of outlier performances. It differs from other similar ranking systems such as Microsofts TrueSkill [3] by exposing internal properties and provides mathematical proofs for its optimisation. Models such as this tend to be vulnerable to volatility farming which is the act of a player intentionally tanking their rating such that later improved performance will improve their overall rating due to the models perceived increased player volatility rating. This models high level implementation can be seen in Figure 2.

Algorithm 1 Elo-MMR($\rho, \beta, \gamma, \mu_{init}, \sigma_{init}$)

for all rounds t **do**
  $\mathcal{P}, \preceq, \succeq \leftarrow$ outcome of round t
  for all players $i \in \mathcal{P}$ in parallel **do**
    if i has never competed before **then**
      $\mu_i, \sigma_i \leftarrow \mu_{init}, \sigma_{init}$
      $p_i, w_i \leftarrow [\mu_i], [1/\sigma_i^2]$
    diffuse(i)
    $\mu_i^\pi, \delta_i \leftarrow \mu_i, \sqrt{\sigma_i^2 + \beta^2}$
  for all players $i \in \mathcal{P}$ in parallel **do**
    update(i)

Algorithm 2 diffuse(i)

$\kappa \leftarrow (1 + \gamma^2/\sigma_i^2)^{-1}$
$w_G, w_L \leftarrow \kappa^\rho w_{i,0}, (1 - \kappa^\rho) \sum_{k \geq 0} w_{i,k}$
$p_{i,0} \leftarrow (w_G p_{i,0} + w_L \mu_i)/(w_G + w_L)$
$w_{i,0} \leftarrow \kappa(w_G + w_L)$
for all $k > 0$ **do**
  $w_{i,k} \leftarrow \kappa^{1+\rho} w_{i,k}$
$\sigma_i \leftarrow \sigma_i/\sqrt{\kappa}$

Algorithm 3 update(i)

$p \leftarrow \underset{x \in \mathbb{R}}{\text{zero of}} \sum_{j \preceq i} \frac{1}{\delta_j}\left(\tanh \frac{x - \mu_j^\pi}{2\delta_j} - 1\right) + \sum_{j \succeq i} \frac{1}{\delta_j}\left(\tanh \frac{x - \mu_j^\pi}{2\delta_j} + 1\right)$
p_i.push(p)
w_i.push($1/\beta^2$)
$\mu_i \leftarrow \underset{x \in \mathbb{R}}{\text{zero of}}\ w_{i,0}(x - p_{i,0}) + \sum_{k>0} \frac{w_{i,k}\beta^2}{\beta} \tanh \frac{x - p_{i,k}}{2\beta}$

Figure 2: Elo-MMR Definition

2.3 Performance Metrics

Measuring the performance of the different mathematical models will be done using two metrics, Accuracy and Log Loss. Accuracy is how many of the matches played did the model accurately predict the correct outcome. Notionally this metric is defined as follows.

$$Accuracy = \frac{count(R)}{count(R) + count(W)} \tag{7}$$

where $count(R)$ is the number of correct guesses and $count(W)$ is the number of incorrect guesses.

Log Loss is a measure of how close the prediction probability is to the actual true/false value. It provides an indication of how confident the output of the model is in its prediction. Log loss punishes predictions where the predicted probability is further from the actual outcome more than if the prediction is closer. Log loss on an individual outcome is defined as follows.

$$Logloss = -[y \ln p + (1 - y)ln(1 - p)] \tag{8}$$

where y is the actual outcome of the match and p is the prediction probability. The overall log loss of the model is defined as.

$$Logloss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \ln p_i + (1 - y_i)ln(1 - p_i)] \tag{9}$$

which calculates the average log loss value of all the individual match outcomes and is the final value used to evaluate the log loss performance of the model.

3 Method

Evaluating the previously mentioned models is a multistep process. The first step is to decide how to "train" the models. There are two approaches that were considered, splitting the data into a training and testing set. Running the model over the training set to tune parameters and then test its performance over the test set to evaluate performance. Traditionally data is split 80-20 into training and test sets. The other approach is to, where possible, iterate over each match make a prediction then update the model. This works well with the Elo models as updating the point values occurs after the outcome of any given match. For the logistic model, however, the parameters will be calculated utilising the R studio logistic model package over the entire dataset. This model will then be used over each match in the dataset to calculate probabilities of the winner winning. This approach runs the risk of overfitting as the entire dataset is being used for both training and testing, however due to the size of the dataset and the minimal impact an individual game has on the overall model the potential impact of this is relatively minimal.

3.1 Dataset

The data used for this investigation is the Tennis Rankings, Results and Stats open-source Github repository[7] provided by Jeff Sackman. This dataset contains statistical information for men's and women's professional tennis including singles and doubles competitions. Not all information contained in this dataset will be used for all the models however fields such as playing surface, player hand and tournament location can be used to test for statistical significance when it comes to predicting outcomes of tennis matches. Investigating the impact of these features will be investigated in future work. The dataset is split up based on year and competition type, e.g. men's singles. For this investigation the focus will be on the men's singles competition. Dealing with missing information within the dataset will be handled by omitting matches which don't have all information required for the model to run. Future work will include alternative approaches to handling models when required information is missing.

3.2 Logistic Model

As mentioned above, the slopes for the logistic model are calculated using R Studio's general linear model (GLM) package. Plotting the outputs of predicted values against this model are.

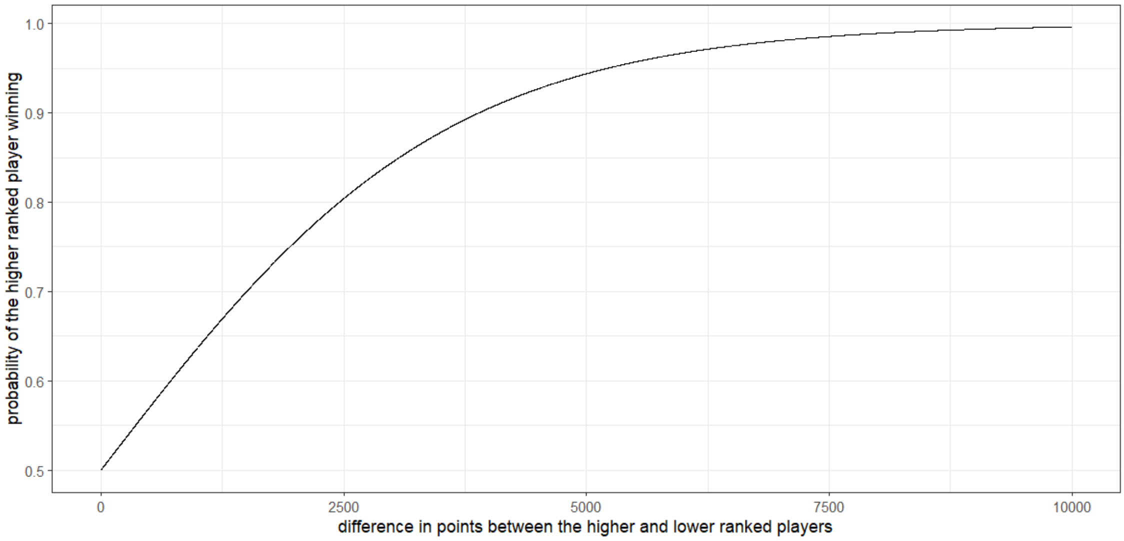

Figure 3: Point Difference v Probability

This parameter was found to be $\theta = 0.000565$. As a result of this the Logistic Model can be defined as follows.

$$P = -\frac{1}{1 + e^{-0.000565x}} \tag{10}$$

where x is the point difference between the two players. Applying this model to each match where both players have rank points will generate the results required to calculate both accuracy and log loss. As well as this, altering the threshold to only provide a prediction if the calculated probability is over a certain threshold will be implemented. Testing a range of threshold values will inform what value provides the best balance of providing predictions to as many matches as possible whilst maintaining a high level of confidence in the predictions. This work will be expanded in the future work section of this report.

3.3 K Factor Model

Adjusting the value of the K value in the K Factor model will involve iterating over various values of K to identify any differences in accuracy and log loss that come with shifting these K values. Initially the value will be set to 5, once this baseline is established K values between 5 and 50 will be tested in intervals of 5.

3.4 Five Thirty Eight Model

Similar to the K Factor model various values for δ, ν and σ will be iterated over. Since there are three parameters that can be tuned the search space will need to be reduced when compared to the K Factor model. Unlike the K Factor model, a brute force approach is most likely inappropriate and more targeted hyperparameter optimisation will need to be done to identify the optimal values. The initial values will be set to $(\delta, \nu, \sigma) = (100, 5, 0.1)$. Future work will be to utilise hyperparameter tuning methods to find the optimal values.

3.5 Elo MMR Model

Due to the complexity of the Elo MMR model the aim for this report is to analyse the default implementations effectiveness when applied to the tennis prediction problem. Similar to the Five Thirty Eight model future work will be conducted to find optimal hyperparameters and improve performance if applicable. Currently the implementation being used for this model is a three step process. First, parsing the data to generate a separate file for each tournament and ranking the performance of each player that competed. This ranking will be determined by the total number of games won within that individual tournament. The number of games played is a good measure of performance for a player as the more games played can be directly tied to how far in a tournament they progressed. The players will be given a ranking from one to n where n is the number of players who competed in a specific tournament. Once these separate files have been generated, these files will be considered a single round in the context of the Elo-MMR model [2]. The model will then iterate over each tournament and provide a set of rankings for the players. These rankings will be output as a map of player names to player ranks at a specific date. Finally the prediction work will be completed which involves iterating over the initial dataset, using the player ranks at the date of the match to predict the outcome. If either of the players dont have a rating, ie it is their first match, a baseline rank of 1500 will be used. Since these results will be loaded into a map of tournament dates and player rankings at that time, the model will be capable of using a players current rating value to determine the probability of them winning the match.

4 Results

4.1 Logistic Model

Running the logistic model provides the below output

Correct	Incorrect
68084	38486

For an overall accuracy value of 0.639. The corresponding log loss for this model is 0.481. With an accuracy lower than simply using the player ATP rank to predict the outcome of a match the model in this form has definite room for improvement. Since the predictions of this model in its current form is based solely on the difference between the players rank points it can be observed that the accuracy of its predictions are similar to using the ATP rank. In the future work section, capturing more features of the dataset to include in the logistic model will be done which is an extension not possible with simply using the players ATP rank.

4.2 K Factor

The K Factor model provides one parameter that can be adjusted to alter the models' performance. This K value impacts both the accuracy and the log loss produced from the model. The table below shows the accuracy and log loss for each value of K.

K Value	Accuracy	Log Loss
5	0.606	0.466
10	0.601	0.383
15	0.598	0.333
20	0.594	0.298
25	0.592	0.272
30	0.590	0.251
35	0.589	0.234
40	0.587	0.220
45	0.585	0.208
50	0.584	0.197

From the results above it can be observed that as the K value increases, the log loss decreases. This, however, corresponds with a decline in the models overall accuracy. As this model does not consider the past performance of the players when updating their corresponding ranks potential improvements to this model are minimal. The following plot tracks the performance of Roger Federer across his long career.

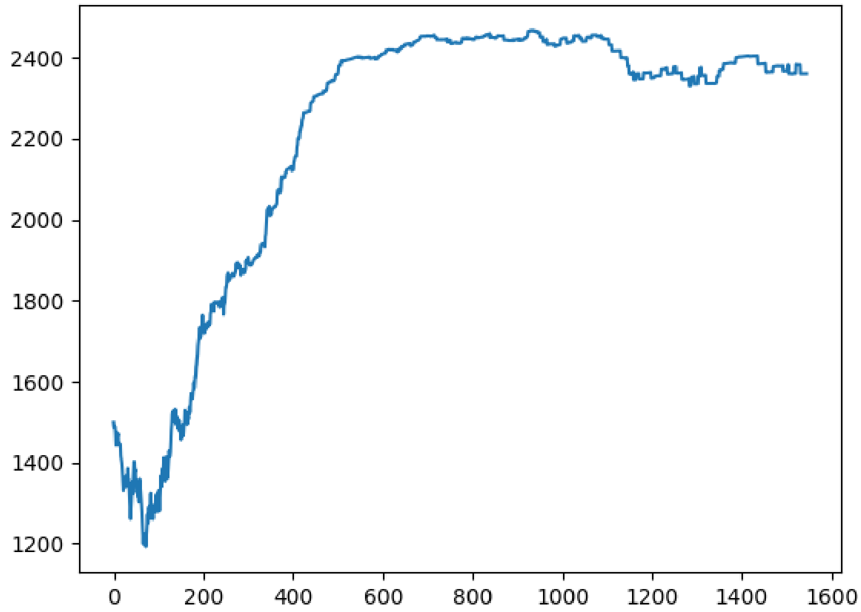

Figure 4: K Factor Federer Rank Over Time

It can be observed the model captures the trajectory of his career quite well with a slow start corresponding to his poor performances at the start of his career, an extended peak during the middle of his career coinciding with his "prime" and then a slight drop off towards the end of his career before his retirement.

4.3 Five Thirty Eight Model

As discussed in the Method section of this report the Five Thirty Eight model relies on three hyperparameters to define the model. The values $(\delta, \nu, \sigma) = (100, 5, 0.1)$ are known to be suboptimal, and the following results are just to demonstrate in initial implementation. Running the Five Thirty Eight model with these parameters provides an accuracy value of 0.577 and an overall log loss of 0.167. This shows similar performance to the K Factor model with a lower corresponding log loss value implying a higher confidence in the models predictions. Hyperparameter optimisation should improve the performance of this model further and this will be expanded further in the future work section of this report. Similar to the K Factor result, the career ranking of Roger Federer has been plotted below.

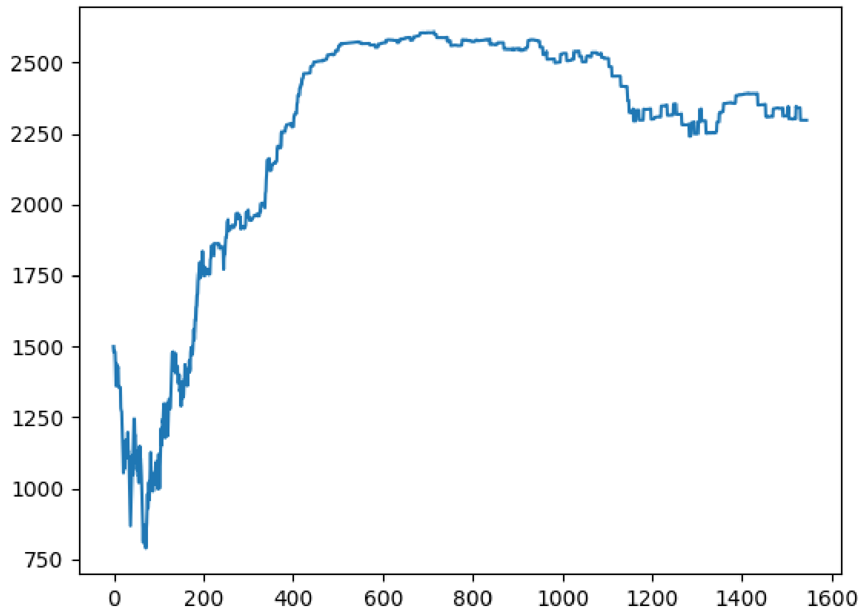

Figure 5: Five Thirty Eight Federer Rank Over Time

This plot looks almost identical to the K Factor ranking for Federer which explains why accuracy of the models is quite similar. The Five Thirty Eight model has Federer having a higher peak rating with a marginally higher rate of change when outlier performances occur.

4.4 Elo-MMR Model

A baseline implementation of the Elo-MMR model has been implemented to set a baseline for future work to build on. The model produced an accuracy of 0.635 with a log loss of 0.533. This model outperformed the other Elo models in accuracy with a significantly worse log loss. Due to the complexity of the model there are many reasons why the model lacks confidence in its predictions. By plotting the performance of a known high performing player, Roger Federer, deficiencies in the model can be observed.

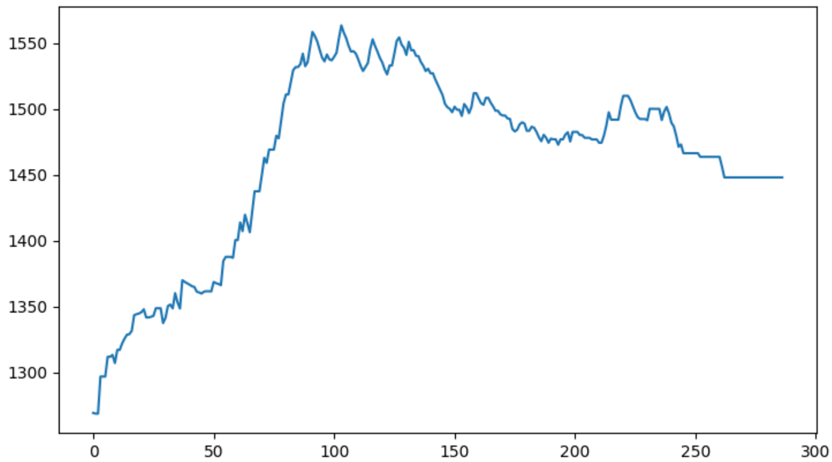

Figure 6: Elo-MMR Federer Rank Over Time

From the above plot, it can be observed that the model is extremely punitive for a player if their performance early in their career is poor and doesn't appropriately rate higher performances to compensate. Future work will include investigating methods of considering the difficulty in reaching higher rounds of a tournament and updating the players rank to more accurately reflect that.

5 Discussion

From the results of the initial implementations of these models there is some key takeaways. The basic logistic model currently has the highest accuracy meaning that it makes the most correct predictions. This is because the model in its current form is very similar to simply using the ATP rank for the prediction as that value is based on the players current rank point totals. A higher point total corresponds to a higher ATP rank resulting in essentially the same prediction. In its current state this logistic model currently only considers a single parameter, the players rank points, however there is opportunity to extend this model to include other parameters present in the dataset

The current implementations of the various Elo models captures a similar trajectory in a players career. This can be observed by overlaying the two Elo models which are updated on a game by game basis, K Factor and Five Thirty Eight. The Elo-MMR model is calculated on a tournament basis so cannot be overlaid simply on the same plot.

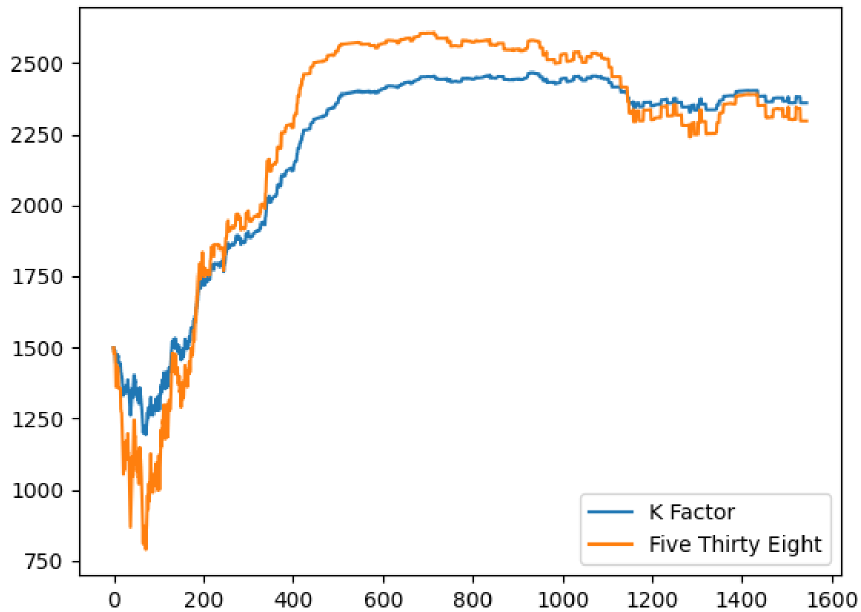

Figure 7: K Factor vs Five Thirty Eight

From this plot it can be observed that the Five Thirty Eight Model varies more than the K Factor model as a result of individual matches. This is because it considers the players history when determining a K value for the rank update. Outlier performances have a greater impact on an individual update which results in a lower rating being assigned early in the career with poorer performances occurring. Conversely, as Federers performance increases, the improvement is picked up earlier in the Five Thirty Eight model when compared to the K Factor model. This results in the model reaching a steady state that accurately reflects Federers skill earlier than the K Factor model which when optimal parameters are identified should result in a higher overall accuracy.

The Elo-MMR model accurately captures the trajectory of Federers career as shown in Figure 6. This model seems to not appropriately reward a player for excellent performance in a tournament and is extremely confident in its initial rating of a player. From the models calculated ratings, this results in a large percentage of the players having ratings that is around the same values. As a result the models predictions is often based on very similar player ratings. By more accurately capturing performance in tournaments and incorporating that in to rating updates there will be more variance in player ratings resulting in more accurate predictions. The current approach for rating performance for a specific tournament is to assign a rank based on the round reached within the tournament. This results in an accurate rating for the first and second best performing players in the tournament as they are the ones who made it to the final. The issue arises when ranking players who get eliminated in earlier rounds. The losers of the semi-finals will be tied for the third-best performance in the tournament. This becomes a problem when considering earlier rounds as the number of players with the same performance increases exponentially. Alternative approaches for more accurately capturing a players performance at a tournament will be expanded in future work.

Currently, all the models in their current states have significantly lowered prediction accuracy than the simple approach of making a prediction solely based on the players ATP rank. The true value in the above work is that these models have much more potential for extension or modification to improve their accuracy. As mentioned, the logistic model can be extended with more parameters, the Elo models can have hyperparameters tuned and since these models are probability based the threshold at which a prediction is made can be changed.

5.1 Future Work

Extending the investigated models will be the focus for the next part of this research project. Starting with the logistics model the natural extension of the current model is to investigate additional parameters to include. Utilising the R logistics model package, both additional parameters and interaction terms will be tested for statistical significance to the match outcomes. Incorporating playing surface and playing hand will be investigated first as there is some evidence that it can have an impact on match outcome [5].

The basic K Factor Elo model will not be investigated any further as the various values of K and its impact on model performance have already been investigated above. The Five Thirty Eight model leverages three independent hyperparameters to calculate the change in ranking after each contest. The search space for brute forcing the optimal values of these hyperparameters is infeasible. There are many methods of optimisation for hyperparameter search [1] and investigating the most appropriate one for optimising the Elo model parameters will form the basis of future work for improving the Five Thirty Eight model.

From the results of the Elo-MMR model there are multiple opportunities for optimisation. By plotting the score of a single player known to have a higher than average performance across a long period of time, Roger Federer, deficiencies in the model performance can be identified. Initial observations indicate that the model becomes heavily confident in its predictions based on player performance early in their career. This results in reducing the rate of change of a player ranking for matches later in their career. The problem with this is player performance early in a career is most likely lower than once they get more experienced in professional play. Currently, the Elo-MMR model is updated for each player after each tournament. To combat the issue of overconfidence in the model based on early career performance, the players initial ranking will be produced from a set of their first n matches or the first n tournaments. Various values of this initial bucket size will be investigated to optimise between model confidence and maximising the number of matches with a player ranking. An investigation will also be conducted to determine the models' performance if instead of updating the ratings on a tournament basis instead it is done on a game by game basis like the other Elo models. The risk of implementing the model in this way is that the identified issue of over-confidence based on early career performance could be increased. The current implementation of the model defaults to the ATP rankings of the players if both players dont have a valid Elo-MMR ranking. By increasing the required number of games played by a player before calculating a Elo-MMR rank the number of games that will default to the ATP rankings will increase. This potentially reduces the accuracy of the model and needs to be minimised. The other optimisation that will be investigated is instead of updating the Elo-MMR after every tournament, the players rank will be updated after n amount of tournaments. This will result in a higher confidence in the update of the model ranking.

The current method of determining whether any of these Elo models have produced a correct prediction is by determining if the probability of player i beating player j is greater than 0.5. Logically this is a good starting point as it implies that theres is a greater than 50% chance of a player winning. Setting this probability threshold to different values will be investigated further to identify any improvements in the models prediction accuracy. With a higher threshold, it is expected that the model will make predictions on less matches as when players have similar Elo scores the model will not make a prediction. Balancing the improvement to model accuracy versus the reduced number of matches in which a prediction is made will be completed and discussed.

Comparing model performance against open source metrics to rate performance of the various models will be included in future work. Investigations into these various metrics will be conducted. The Five Thirty Eight model with optimal parameters has been found to provide predictions with approximately 75% accuracy [4]. This is just one of the metrics that will be compared and will be used to set a baseline of which to

6 Conclusion

Performance of these models have been quantified by measuring both model accuracy and the log loss of the models predictions. There are many mathematical models that can be applied to the tennis prediction problem. These models range in both complexity and performance as such robust performance testing has been conducted to identify which models perform the best. Performance of these models have been quantified by measuring both model accuracy and the log loss of the models predictions. An extensive open-source dataset of historical tennis data was leveraged to both build and test these models. Currently a small subset of the available data is being used to build these models. Future work will include incorporating a greater number of features within the dataset. Capturing more features of the available data and conducting statistical analysis on its impact on match outcome will result in more accurate predictions with a higher confidence as measured by its corresponding log loss value.

The logistic model performs well with just a basic classifier based on a players ATP rank. The log loss of the logistic model, is higher than other models investigated in this report. This implies that the model, confidence in its predictions is lower than alternatives such as the Elo models. Attempting to address this issue will be conducted as a part of future work by extending the parameters used to build the logistic model to increase the models confidence in its predictions.

Various Elo models were implemented and their respective performances compared. These Elo models have similar performance to the basic logistics model however they demonstrate a lower log loss value. This demonstrates that these models are more confident in their predictions so if they can be modified to increase the accuracy the model will outperform the basic logistic model. Differences in the ratings' of an individual player generated by the individual models shows that they all succeed in capturing the overall trajectory of a players career. The different Elo models update the player ratings at different rates based on performance. Having the model accurately update these ratings is the key to maximising accuracy whilst minimising log loss. The more complex Elo models such as the Five Thirty Eight and the Elo-MMR provide more mechanisms for tuning than the simpler ones such as the K Factor model. This comes at the cost of model complexity as incorrectly configured models can have a massive impact on its performance. Balancing the need to optimise the models to solve the given problem with model complexity will be investigated in future work.

References

[1] Marc Claesen and Bart De Moor. Hyperparameter Search in Machine Learning, April 2015. arXiv:1502.02127 [cs, stat].

[2] Aram Ebtekar and Paul Liu. Elo-mmr: A rating system for massive multiplayer competitions, 2021.

[3] Ralf Herbrich, Tom Minka, and Thore Graepel. TrueSkill: A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

[4] Stephanie Ann Kovalchik. Searching for the GOAT of tennis win prediction. *Journal of Quantitative Analysis in Sports*, 12(3):127–138, September 2016. Publisher: De Gruyter.

[5] Florian Loffing, Norbert Hagemann, and Bernd Strauss. Left-Handedness in Professional and Amateur Tennis. *PLoS ONE*, 7(11):e49325, November 2012.

[6] Utathya Nag. Tennis rankings: Everything you need to know, April 2022.

[7] Jeff Sackmann. JeffSackmann/tennis_atp, April 2024. original-date: 2015-03-13T12:15:27Z.

[8] ATP Tour. Faq about pif atp rankings: Atp tour: Tennis, 2023.