Modelling the Outcome of Tennis Matches
Interim Report

S. Gilbert
A1737770.

July 20, 2024

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

Project Area: **Data Science - Modelling**
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.

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

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 under pin 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. Alongside the data used to build and train these models, open source historical gambling odds were used to benchmark the various models performances.

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. Comparing and contrasting these different approaches and their impact on prediction accuracy and log loss was investigated in this report.

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 [5]. 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

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. To effectively measure the performance of a given model a benchmark for what is considered a "good" result must be used. For this report the Bookmakers Consensus model (BCM) was implemented which combines betting odds from a number of bookmakers to provide a probability of winning for each player in a given match. For reasons that will be discussed later, this model is unsuitable for the future match prediction problem however is useful as a benchmark validation tool. 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.

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 [7]. 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;

π_i = probability of player winning
D_i = difference in rank points between player i and player j

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

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

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

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

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

and updating the players ranking is done as follows

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

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

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

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 Bookmakers Consensus Model

The BCM takes the odds from different bookmakers in order to converge on a quorum estimate of match probabilities [?]. In the United Kingdom where the data is from, bookmakers represent the odds of a given outcome as a float value. This float value represents the amount won if the predicted outcome occurs. For example if a player has odds of winning of 1.3 then for every one hundred dollars bet, the better would receive one hundred and thirty dollars if the outcome occurs. The BCM begins by calculating the probabilities of player α and β winning the given match. This probability is found to be $\pi_1 = 1/\alpha$ and $\pi_2 = 1/\beta$. Intuitively it would be expected that adding these probabilities would result in a total value of 1.0. This is not the case due to over round which is a method used by the companies to maximise their profits. To address this the probabilities must be normalised. Normalising for player 1 is

$$\pi_1 = \frac{\pi_1}{\pi_1 + \pi_2} \tag{7}$$

then substituting the odds

$$\pi_1 = \frac{1/\alpha}{1/\alpha + 1/\beta} \tag{8}$$

The same is done for player two is done and is

$$\pi_2 = \frac{1/\beta}{1/\alpha + 1/\beta} \tag{9}$$

Since this is a BCM a set of bookmaker odds for any given match need to be considered. For each k company in the set of length N, the odds of player 1 and player 2 winning a given match according to the kth company, defined as $\pi_{k,1}$ and $\pi_{k,2}$ can be found as

$$\pi_{k,1} = \frac{\beta_k}{\alpha_k + \beta_k} \ \ and \ \ \pi_{k,2} = \frac{\alpha_k}{\alpha_k + \beta_k} \tag{10}$$

The BCM finds a type of mean probabilities across all the companies with published odds and this is defined as

$$logit(\pi_1) = \frac{1}{N}\sum_{k=1}^{N} logit(\pi_{k,1}) \tag{11}$$

This equation can be inverted using the definition of the *logit* function

$$logit(\pi_1) = log(\frac{\pi_1}{1-\pi_1}) \tag{12}$$

which can be used to obtain

$$\pi_1 = \frac{e^y}{1+e^y} \tag{13}$$

where $y = logit(\pi_1)$.

2.3 Performance Metrics

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

$$Accuracy = \frac{count(R)}{count(R) + count(W)} \tag{14}$$

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.

$$Logloss = -[y \ln p + (1-y)ln(1-p)] \tag{15}$$

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.

$$Logloss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \ln p_i + (1-y_i)ln(1-p_i)] \tag{16}$$

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.

Finally, model calibration, C, is defined as follows

$$C = \frac{1}{W} \sum_{i=1}^{N} \pi_i \tag{17}$$

with W being the number of games won by the higher ranked player and pi_i is the probability of the higher ranked player winning. Calibration aims to measure the rate at which the higher ranked player wins. If the model is well calibrated then $C \approx 1$, if $C > 1$ then the model overestimates the wins of the highest ranked player and if $C < 1$ then the model underestimates the wins.

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

Two key datasets were used for this investigation. The first is the Tennis Rankings, Results and Stats open-source Github repository[6] 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.

The other dataset is the Tennis Data UK [8] dataset which contains a smaller subset of usable fields then the Jeff Sackmann data, however contains betting odds for all professional tennis matches between the years of 2001 to the present. This project was able to use features from both datasets by first merging them. This process involved selecting common fields from both datasets and doing a join based on the selected fields. Winner rank points, loser rank points and location were the fields chosen due to the differences in recording methods used for player names and the differences in the match played dates which was potentially due to the differences in time zones of the recorders. Some modifications had to be done to align the location names of various tournaments due to sponsor changes and shifting tournament names that were captured in one of the datasets but not the other.

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.

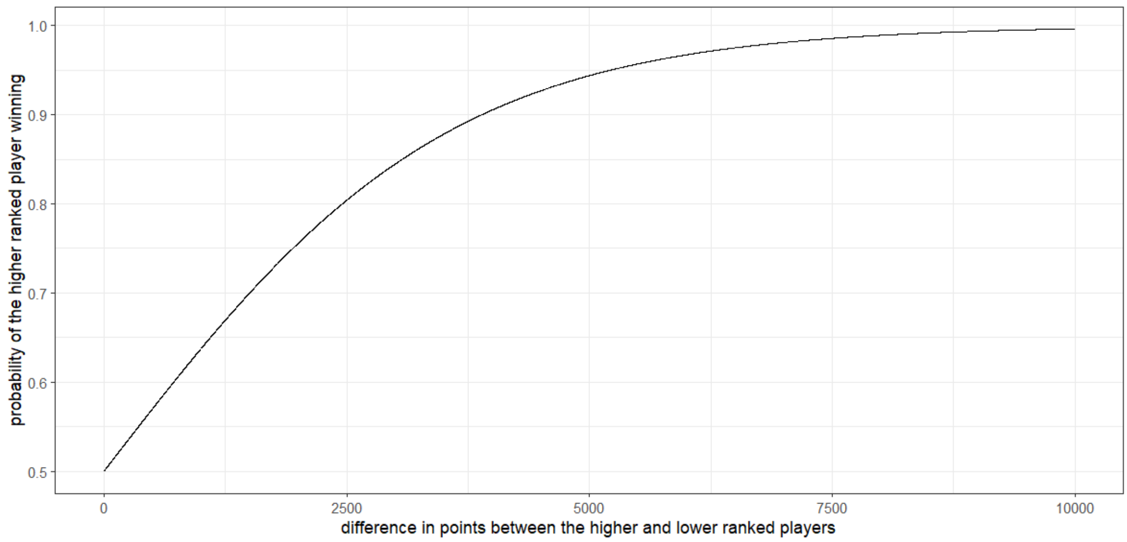

Figure 1: Point Difference v Probability

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

$$\pi_i = -\frac{1}{1 + e^{-0.000565 D_i}} \tag{18}$$

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 Surface Augmented Five Thirty Eight Model

An extension to the Five Thirty Eight model was implemented which calculates a set of Elo scores for each player on each surface. These scores are calculated using the Five Thirty Eight model by calculating a probability that each player would win on the given surface. This probability is then used to update the Elo rating fore each player on the given surface. As a standalone model, these surface Elo ratings capture such a small amount of information that prediction ability is minimal. These surface Elo's were combined with the base Five Thirty Eight model as an augmentation. The optimal method of combining these surface models was investigated. A defined weighting value γ was used to influence what proportion of the final model output was the Five Thirty Eight model and what was the Surface model. The first approach of combining these models was taking the calculated probability of player 1 winning, p_1, from each models and combining them to generate a new probability $p_{1,combined}$ based on the weighting factor

$$\pi_{1,combined} = \gamma * \pi_{1,fivethirtheight} + (1 - \gamma) * \pi_{1,surface} \tag{19}$$

The other approach was to combine the Elo scores produced by each model for each match using the γ weighting factor with the Five Thirty Eight Elo rating at match T defined as $E_{fivethirtyeight}(T)$ and the surface Elo at match T, $E_{surface}(T)$, the combined Elo rating, $E_{combined}(T)$, can be calculated using the γ value As

$$E_{combined}(T) = \gamma * E_{fivethirtyeight}(T) + (1 - \gamma) * E_{surface}(T) \tag{20}$$

Then using the same formula from (3) to calculate π_{ij} the probability of player 1 beating player 2 can be found

4 Results

After building and testing the different models, the metrics for each have been displayed in Table 1. In the following section the methods for optimising the resulting metrics will be explored. From Table 1 it can be observed that the best performing model was the Bookmakers Consensus model with the highest accuracy and lowest log loss. Although it did have a marginally worse calibration. Out of the models appropriate for predicting match outcomes, the Surface Combined Elo models performed the best with the different methods for including the surface information resulting in a negligible difference in calculated metrics.

Model	Accuracy	Log Loss	Calibration
Higher Rank	0.660	NA	NA
Logistic Regression	0.639	0.634	0.867
K Factor	0.658	0.613	0.889
Five Thirty Eight	0.662	0.614	0.897
Surface Elo	0.598	0.722	0.944
Five Thirty Eight - Surface Combined Elo	0.663	0.614	0.893
Five Thirty Eight - Surface Combined Prob	0.663	0.612	0.892
Bookmakers Consensus	0.695	0.575	0.863

Table 1: Metrics for all investigated models showing that the best performing model was the Bookmakers Consensus and the best overall predictive model being either of the Surface Combined Elo models

4.1 Logistic Model

Running the logistic model provides the below output

Accuracy	Log Loss	Calibration	Correct	Incorrect
0.639	0.634	0.867	68084	38486

Table 2: Generated metrics for the Logistic Regression Model

For an overall accuracy value of 0.639. The corresponding log loss for this model is 0.634. 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	0.639	0.631
10	0.648	0.619
15	0.653	0.614
20	0.655	0.613
25	0.657	0.612
30	0.658	0.613
35	0.657	0.615
40	0.658	0.620
45	0.659	0.623
50	0.659	0.627

Table 3: K value tuning to find the optimal value based on Accuracy and Log Loss, It can be observed that the optimal value is K=30

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.

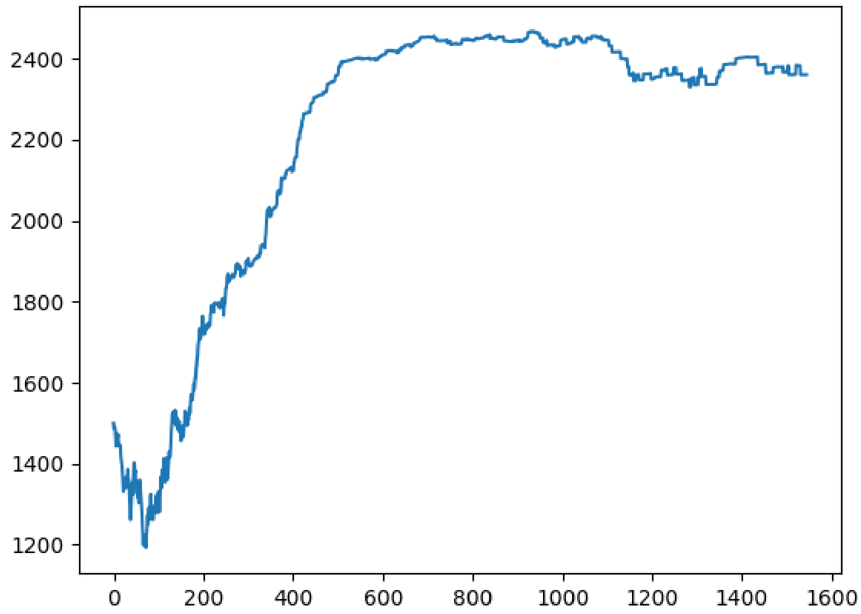

Figure 2: K Factor Federer Rank Over Time

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 $(\delta, \nu, \sigma) = (240, 5, 0.4)$ were determined after completing a broad grid search across all permutations for *delta* between the values of 100 and 300 with a step increment of 20, ν between the values of 1 and 10 with a step increment of 1 and σ between the values of 0.1 and 1 with a step increment of 0.1. This involved building and testing a thousand different models and generating summary statistics for all of them. Although more advanced methods of hyperparameter tuning could have been implemented the results of this broad grid search showed that the optimal values for accuracy and log loss were maximised at the given values and align closely with values published in existing literature [3]

Running the Five Thirty Eight model with these parameters provides an accuracy value of 0.622 and an overall log loss of 0.614. This shows higher accuracy when compared to the K Factor model with a lower corresponding log loss value implying a higher confidence in the models predictions. Similar to the K Factor result, the career ranking of Roger Federer has been plotted below.

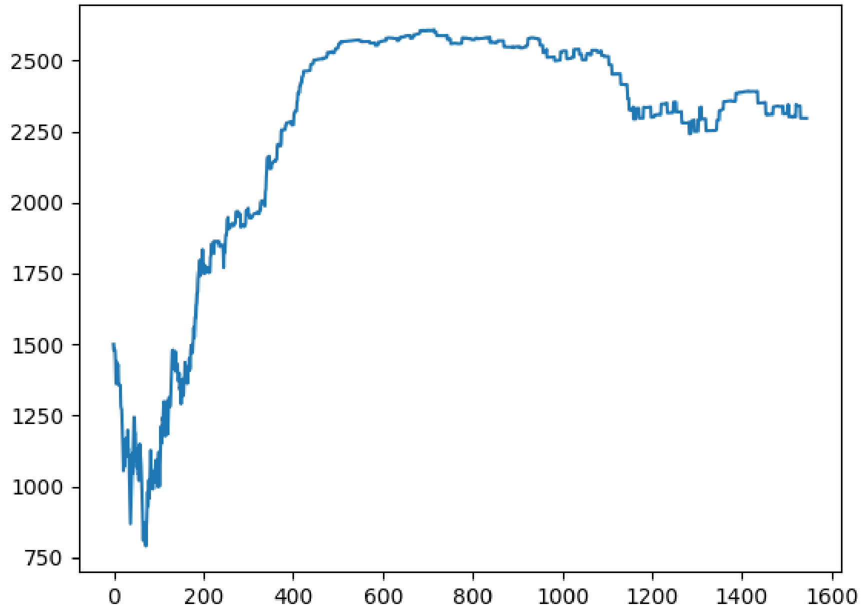

Figure 3: Five Thirty Eight Federer Rank Over Time

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 Surface Combined Five Thirty Eight Model

As discussed in the method section, the existing Five Thirty Eight model was augmented with a separate Elo rating generated for each player on each surface type. From Table 1 it can be observed that the Accuracy was identical for both methods of combining the models sitting at 0.663 with Log Loss and calibration both having a negligible difference ≈ 0.002. When looking at the Surface Elo model that hasnt been combined with an optimised Five Thirty Eight model, it can be observed that its performance is significantly worse than any of the other models in all metrics excluding calibration. Discussion into why this is the case will be done later in this report.

4.5 Bookmakers Consensus Model

The Bookmakers Consensus model outperformed all the other models significantly in accuracy and log loss with values of 0.695 and 0.575 respectively. The natural question that follows from the results of this model is why is this not the model selected as the best for predicting tennis match outcomes? For reasons that will be explored on in the discussion section of this report, due to how the information for each match played is reported, this model cant be used until after a match is played making it unsuitable as a solution for the prediction problem.

4.6 Match Subset Investigation

To dive deeper into the behaviour of the individual Elo models, a random subset of a single players matches were selected from the data and the Elo scores for each of those matches were plotted. Roger Federer was the player selected and his 300th to 330th games were used. The below plot was generated.

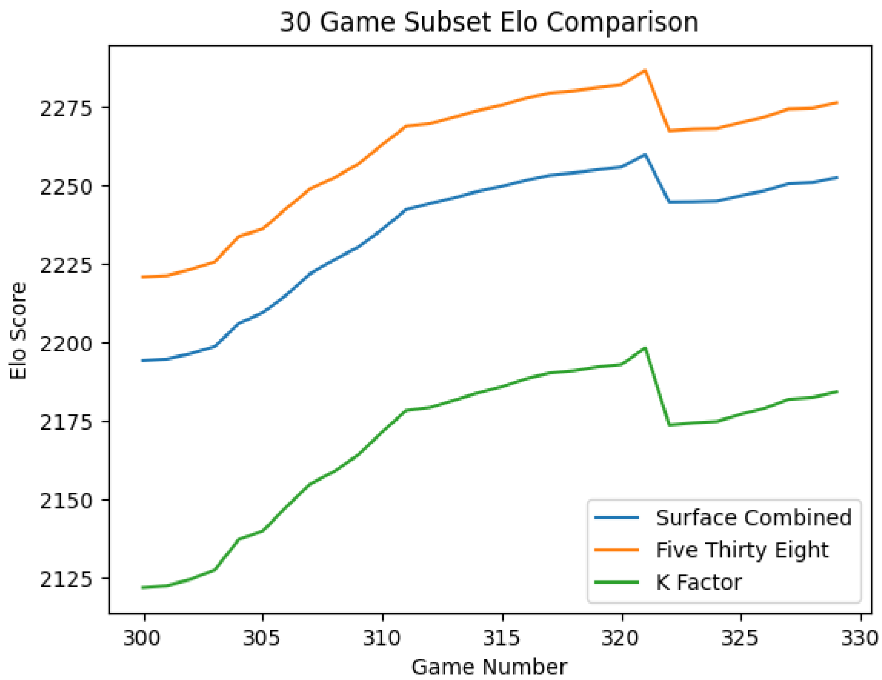

Figure 4: Federer 30 game subset

From Figure 4 it can be observed that the models have the same trajectories across the various Elo models which is to be expected. If a player wins a game the score should go up and if a player loses the game the score should go down. Where the models differ is in the magnitude of change. It can be observed that the models are in lockstep whilst Federer is winning, however when Federer loses game 321 the K model heavily punishes his score as it does not take into account his history of winning and solely focuses on the difference in player scores and K value to determine the amount to reduce the score. The Five Thirty Eight based models still punishes his score due to the loss, however at a lower rate. The Surface Combined Elo has the smallest drop of all the models which is most likely due to Federer having a lower Surface Elo for that match resulting in a reduction of his combined elo score. This would correlate to a smaller drop as the difference in rating is smaller.

Table 4 s

4.7 Federer Career Investigation

The current implementations of the various Elo models captures a similar trajectory in a players career. This can be observed by overlaying the

Winner	Loser	Surface W	Surface L	K W	K L	538 W	538 L
Safin M	Federer R	1982	2244	2030	2267	1937	2173

Table 4: Elo scores for all elo models between Federer and Safin where Federer lost. This figure highlights the impact the differences between the models when calculating Elo score changes

two Elo models which are updated on a game by game basis, K Factor and Five Thirty Eight.

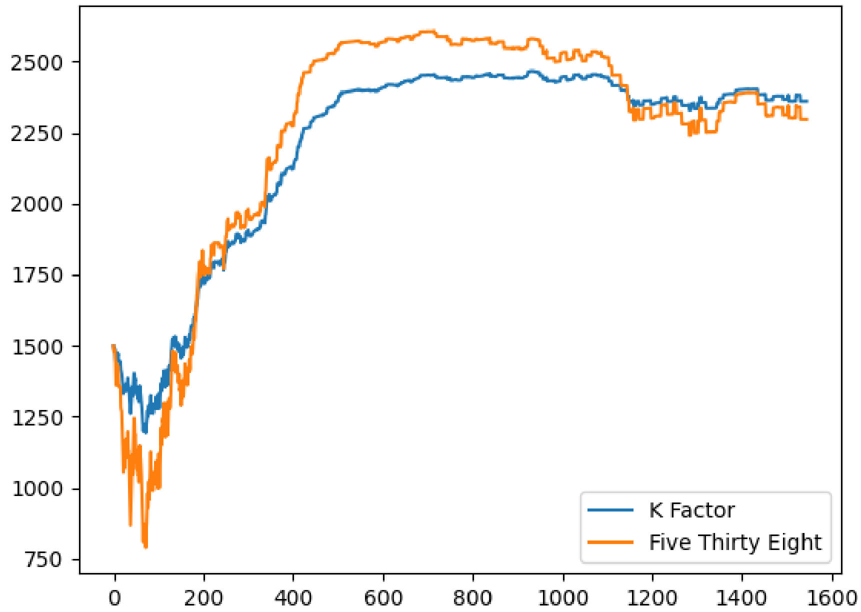

Figure 5: K Factor vs Five Thirty Eight

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.

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

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 [4].

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.

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 [3]. 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.

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] 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.

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

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

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

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

[8] Tennis Data UK. Tennis Data UK: All results, June 2024.