Tennis Match Prediction Project Report

Rohan Vasant Padaya
A1872217

August 13, 2024

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

THE UNIVERSITY
of ADELAIDE

Project Area: **Data Science**
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 report aims to predict the tennis matches outcome by exploring various machine learning models, thereby emphasizing the relevance of accurate sports prediction. The report explored models such as the Naive model and Logistic regression model, which are relatively simpler to implement. The analysis in the report further explored models like the ELO model with different K values and a benchmark model called the Bookmakers Consensus Model (BCM). The models were evaluated using metrics like accuracy, calibration, and log-loss on a dataset of men's tennis matches spanning two decades. The analysis revealed that the ELO model with a dynamic K-factor achieved the most effective balance between accuracy and log-loss, whereas the other models struggled to consistently maintain this balance across all evaluated metrics. The report also identified areas for future research, including the incorporation of additional predictor variables and exploration of complex models like Glicko and TrueSkill.

1 Introduction

The field of sports analytics has grown remarkably in recent years, largely due to the abundance of data available for analysis and the advancement of various machine learning techniques. This growth has enabled the analysis of the vast datasets to identify underlying patterns and predict outcomes. of the sport games with unprecedented accuracy [1]. A popular sport like tennis also involves rich trove of data available for exploration and analysis. Tennis is a sport played on a rectangular court. with different surface types such as grass, clay, or hardcourt. The game is either played between two players in a singles game or four players in a doubles game, separated by a net in the middle of the court. The primary objective is to score points by hitting the ball across the net, separating the players. and ensuring that the opposition player is unable to return the ball within the set boundaries of the game [2].

Before machine learning came into existence, the outcome of the tennis matches was predicted by simply comparing the ranks of the players or by looking at their head-to-head records [3]. Since the complexity of the sport has grown and the players are more competitive, it has become essential to use more advanced statistical techniques to predict the result of the matches. Our project uses historical tennis match data from 2000 to 2019 to predict the outcome of the tennis matches. The data includes a wide range of tournaments, such as Grand Slams, Masters, and other tour-level tournaments. The primary objective is to use probabilistic models to predict the outcome of the tennis matches.

2 Methods

2.1 The Data

In our analysis, we use historical tennis match data obtained from a reliable betting website, a comprehensive source that provides match results along with the betting odds [4]. The data includes matches of both men and women, but we are solely focusing on the men's tennis matches for our analysis. Furthermore, we specifically chose to focus on matches played between 2005 and 2019. This time period was selected due to the completeness and consistency of the data available. Data from years prior to 2005 were found to have significant missing information, particularly some key columns necessary for our analysis. Including such incomplete data would have compromised the consistency and reliability of our models. By selecting data from 2005 onwards, we ensured that our analysis would be based on a more robust and comprehensive dataset, allowing for more accurate and consistent predictions.

The dataset comprises a total of 40390 entries and 48 columns, each representing individual men's tennis matches from various tournaments. Table 1 shows the columns we have selected for the analysis. The selected columns offer a comprehensive array of data for analysing the tennis matches and the player rankings, which are vital for developing and refining our predictive models. For some of our initial models, we are mainly focusing on columns such as the ATP rankings and the Ranking points of the players. The Association Of Tennis Professionals (ATP) uses a merit based system to assign ranks to the players based on their ranking points. The ATP rankings are updated weekly and are used to determine entry and seeding in tournaments, reflecting a player's current standing and performance in the professional tennis world. These ranks are calculated over a rolling period of 52 weeks and then refreshed at the end of the period. For example, Figure 1 shows the comparison of the ATP rankings of two players, Roger Federer and Tommy Haas between the time period 2000 to 2020.

Using the ATP ranks and ranking points of the players, we have added two more columns, "higher ranked won" and "diff" through feature engineering. These columns act as the predictor variables for our initial few models.

For the development of our models, the dataset is partitioned to facilitate robust training and effective validation. The data spanning from 2005 to 2018 is used as the training dataset. For testing, we exclusively use the data from the year 2019, which helps us effectively assess the performance of our models. The performance of the models can be computed using this dataset to identify the most efficient model among the

Column Name	Description
Date	The date of the match.
Tournament	The name of the tournament where the match took place.
Surface	The type of surface (e.g., clay, hard, grass) on which the match was played.
Winner	The name of the player who won the match.
Loser	The name of the player who lost the match.
WRank	The ATP ranking of the winner.
WPts	The ATP ranking points of the winner.
LRank	The ATP ranking of the loser.
LPts	The ATP ranking points of the loser.
B365W	Bet365 odds of the winner.
B365L	Bet365 odds of the loser.
PSW	Pinnacle Sports odds of the match winner.
PSL	Pinnacle Sports odds of the match loser.
higher_rank_won	A boolean indicator showing whether the higher-ranked player won the match.
diff	The difference in ATP ranking points between the two players.

Table 1: Columns selected from the dataset for the analysis

ones developed.

2.2 The models

In this analysis, we explored Five different types of probabilistic models for predicting the outcome of the tennis matches. The Five models explored are: Naive, Logistic regression model, ELO model with constant K, ELO model with dynamic K and Bookmakers Consensus Model (BCM). We will go through each model one by one.

2.2.1 Naive Model

In our naive model implementation, we calculate the probability that a higher-ranked player wins by dividing the total number of matches won by a higher-ranked player by the total number of matches in the dataset. The Naive model assumes that the higher-ranked player is more likely to win. The formula to calculate the probability that the higher-ranked player wins a match is given by:

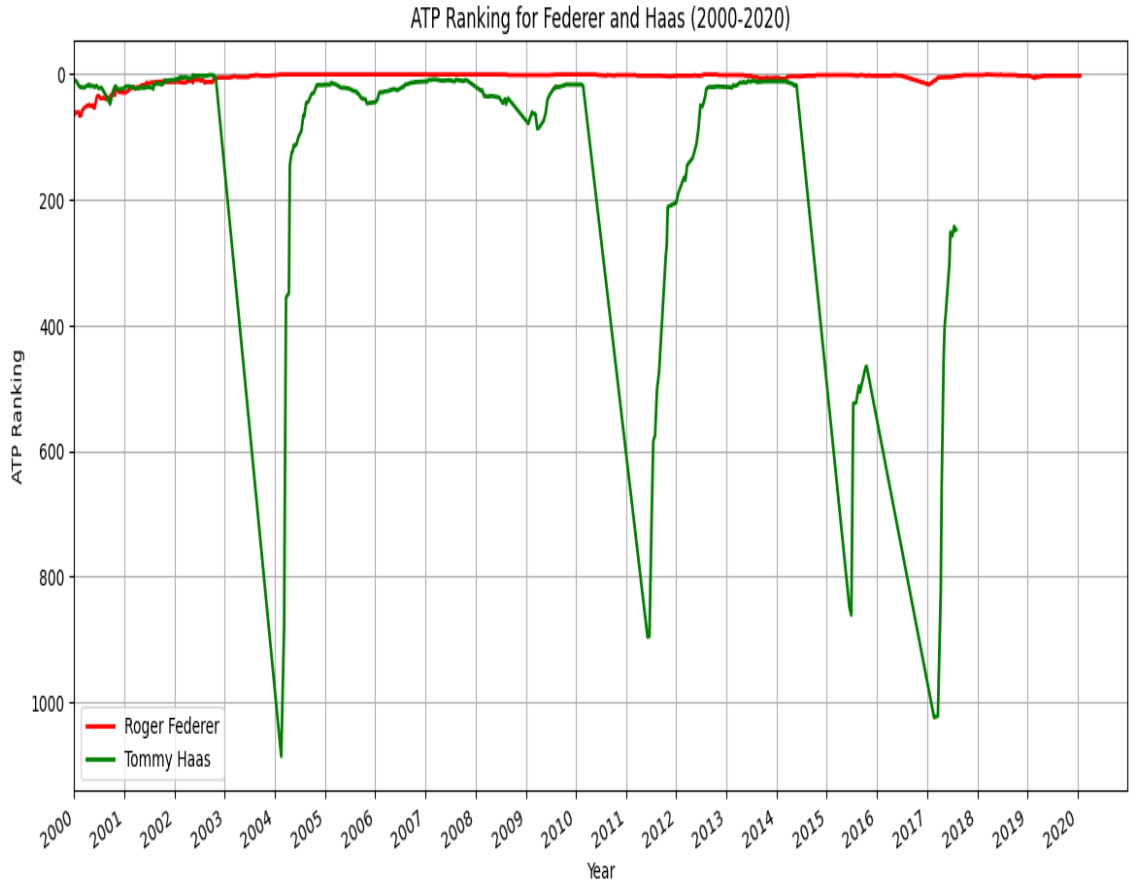

Figure 1: ATP Rankings of players

$$P(\text{higher-ranked wins}) = \frac{\text{Number of matches won by the higher-ranked player}}{\text{Total number of matches}} \tag{1}$$

This simple probabilistic model is used as the baseline model for the rest of our analysis to compare and contrast with other models.

2.2.2 Logistic Regression Model

The Logistic regression model is primarily used in scenarios where the outcome is a binary value (0 or 1). In our analysis, the logistic regression model is used to predict the probability that the higher-ranked player wins based on some transformed predictor. It is different from the previous model, where we were simply calculating the average number of matches won by the higher-ranked player to compute the probability.

In this model, our primary transformed predictor is the difference

in ranking points between the higher-ranked player and the lower-ranked player. Let D_i be the difference in the ranking points between the higher-ranked player and the lower-ranked player, and let $A_{i,1}$ and $A_{i,2}$ be the ranking points of higher-ranked player and lower-ranked player, respectively, for the i-th match. The difference D_i will be as follows:

$$D_i = A_{i,1} - A_{i,2}.$$

The probability that a higher-ranked player wins is defined in terms of log-odds. The logistic model is defined by the following logistic function:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta D_i,$$

where π_i is the probability of the higher-ranked player winning, D_i represents the difference in ranking points between the two players in a match, β is the model parameter to be estimated from the data.

The probabilities are calculated by inverting the log-odds back to probability scale using the following function:

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

where π_i is the probability of the higher-ranked player winning, D_i represents the difference in ranking points between the two players in a match, β is the model parameter estimated from the dataset.

2.2.3 ELO model

The ELO rating system is a system used to dynamically rate a player based on their previous performance, particularly in a zero-sum game. A zero sum game is a sport that does not have the possibility of a draw. The player either loses or wins the game. Tennis, for example, is a zero-sum game that uses ELO rating to gauge the relative skill level of the players [5]. The ELO rating system has an advantage over the traditional ATP rankings as it offers more understanding of a player's relative strength and is more dynamic than the ATP rankings. Using these ratings, we compute the expected probability of the higher-ranked player winning over a lower-ranked player. This model uses ELO rating as its primary predictor for fitting the model.

The ELO rating of a player changes after every match based on the outcome of the match. The Expected score of a match is computed based on the difference in the ELO ratings of the players using the following equation:

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

where $E_i(t)$ and $E_j(t)$ are the Elo scores of players i and j at time t, respectively. And $\pi_{i,j}(t)$ is the expected score of the match.

This expected score is then used to update the ELO scores of both the players after the match using the following equation:

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

where $E_i(t)$ is the Elo rating of player i at time t, $K_i(t)$ is the update function at time t, $W_i(t)$ is an indicator variable denoting whether player i won the match (0/1), $\pi_{i,j}(t)$ is the predicted probability that player i wins over player j at time t.

As we can see in the equation, if the player i wins the match, $W_i(t)$ will be 1, which means the updated ELO rating will be higher tham the previous rating. Similarly, if $W_i(t)$ is zero, the updated ELO rating will be lower than the previous ELO rating of the player $W_i(t)$.

$K_i(t)$ is a multiplicative factor which is decided based on the model that we are using. There are essentially two types of ELO models that we have explored in our analysis.

ELO model with Constant K: In this model, a constant K value is used, particularly for our analysis we have assumed initial value of K as 25. This model does not consider the impact of the number of matches played by the player.

$$K_i(t) = k,$$

where k could be any constant value.

ELO model with Dynamic K: ELO model with Dynamic K Model uses a dynamic K value which changes based on the number of matches played by the player. This model assumes that as the player becomes more experienced, the impact of K value on the player's ELO rating decreases as the player's skill level stabilizes. The K value is computed using the following equation:

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

where $m_i(t)$ is the number of matches played by player i up to time t, δ, ν, σ are model parameters which can be tuned based on the requirements of the model

This equation allows to dynamically update the K factor which in turn helps to update the ELO rating of the player keeping the experience level in mind.

2.3 Bookmakers Consensus Model (BCM)

The Bookmakers Consensus Model is a model which utilizes the betting odds from multiple bookmakers or companies to compute the probability that a higher-ranked player will win. This model acts as a benchmark to evaluate the performance of our other models, as it reflects the collective predictions of all the betting companies. These companies invest in highly robust and efficient algorithms to generate the odds [6]. These odds then can be used to predict the match outcome with high accuracy. Hence, this model is ideal to use as a benchmark for our analysis.

To calculate the probabilities in the BCM, the betting odds given by various bookmakers for a match are first converted into implied probabilities. For a single bookmaker, if the odds for player 1 winning are denoted by α and the odds for player 2 winning are denoted by β, the implied probabilities for player 1 and player 2 are:

$$p_1 = \frac{1/\alpha}{1/\alpha + 1/\beta} = \frac{\beta}{\alpha + \beta}$$

.

$$p_2 = \frac{1/\beta}{1/\alpha + 1/\beta} = \frac{\alpha}{\alpha + \beta}$$

.

These implied probabilities should ideally sum up to 1. However, bookmakers include a margin, known as the "overround" to ensure profitability. This margin causes the sum of the probabilities to exceed 1. To avoid this, the BCM normalizes the probabilties to account for this overround. When multiple bookmakers are involved, the consensus probability for player 1 winning is calculated as the average of the logit-transformed probabilities from each bookmaker. If there are N bookmakers, the consensus probability p_1 is computed as follows:

$$\text{logit}(p_1) = \frac{1}{N} \sum_{k=1}^{N} \text{logit}(p_{k,1}).$$

After averaging the logit values, the consensus probabilities p_1 are obtained back again by inverting the logit function.

$$p_1 = \frac{e^{\text{logit}(p_1)}}{1 + e^{\text{logit}(p_1)}}.$$

3 Validation

Validation is one of the most critical steps in the context of modeling. It ensures that the model not only performs well on the training data but also generalizes effectively to the unseen testing data.
For the validation of our models, we are using the following metrics:

3.1 Accuracy

Accuracy is used to measure how well the model predicts the outcome. It essentially calculates the proportion of times the model's prediction matches the actual prediction. This gives us a straightforward measure of the effectiveness of the model. It is defined mathematically as:

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

where:

- N is the total number of games in the validation set.
- y_i is an indicator variable defined as:

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

- The function $f(i)$ returns the prediction (either 0 or 1) of who will win game i.
- The indicator function $\mathbf{1}\{A\}$ is defined as:

$$\mathbf{1}\{A\} = \begin{cases} 1 & \text{if condition } A \text{ is satisfied}, \\ 0 & \text{otherwise.} \end{cases}$$

Without this metric, it will be difficult to measure the effectiveness of the model. Hence, it is one of the vital metrics used in the validation process.

3.2 Calibration

The calibration metric measures the reliability of the probabilistic output of the model. It is calculated using the following equation:

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

where W is the number of games won by the higher ranked player and π_i is the probability of the higher ranked player winning. If the Calibration value of the model is approximately equal to 1, then the model is said to be well-calibrated. If it is greater than 1, then the model is said to overestimate the probability values and if it is less than 1, then the model is said to underestimate the probability values. Our analysis warrants the need for reliability since it involves real-time prediction of a sports game. Hence, it is essential to use this metric for our model validation.

3.3 Log-loss

Log loss, also known as Cross Entropy, is a validation metric that penalizes the incorrect predictions made by the model in high confidence.

Its mathematical equation is given by:

$$L = -\frac{1}{N} \sum_{j=1}^{N} \log(\pi)$$

where π is the probability of higher-ranked player of the j_th game. The focus is to minimize the overconfidence on incorrect predictions which is why this metric is used for our validation process.

4 Results

In this section of the report, we will compare all our models based on the validation metrics such as Accuracy, Calibration, and Log-loss values to find the most efficient model for our tennis dataset.

The table 2 shows the validation metrics' values for all our implemented models.

Model	Accuracy	Calibration	Log-Loss
Naive Model	0.6599	1.0000	0.6411
Logistic Model	0.6146	1.0384	0.6512
Elo (Dynamic K)	0.6378	1.0575	0.6418
Elo (Constant K)	0.6347	1.0410	0.6347
BCM Model	0.7867	0.9563	0.5032

Table 2: Performance metrics of all the models

As we can see in the table above, the ELO Model with dynamic K has the best accuracy among all the models. This model slightly outperforms the ELO model with a constant K factor, suggesting that dynamically adjusting the K factor based on the number of matches played provides some improvement in predictive power. However, the Naive model does have a higher value than ELO model with dynamic K but since it is our base model, it will not be under consideration for the selection of the best model. On the other hand, the Logistic model shows the lowest accuracy, indicating it struggles to correctly predict match outcomes as effectively as the other models.

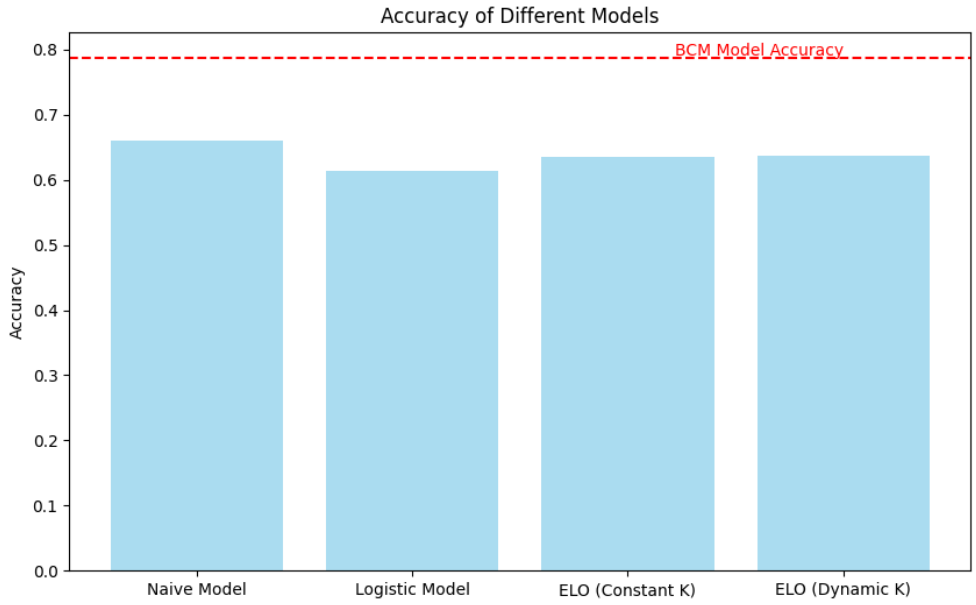

Figure 2: Accuracy comparison of different models, with the BCM model accuracy represented by the red dashed line.

In terms of Calibration, Logistic regression model shows the best calibration as it is closest to 1. This indicates that, although the Logistic model has lower accuracy, it does considerably well in predicting outcomes in a manner that closely aligns with the actual probabilities. Naive model has the perfect calibration but this needs to carefully interpreted as a perfect calibration can suggest overfitting. The simplistic nature of the Naive model can also be a factor for such a calibration value. Both the ELO models show a tendency to slightly overestimate the winning probabilities despite their accuracies being the strongest of all.

For log-loss, the ELO model with constant K performs the best, suggesting that it effectively balances the prediction accuracies with good

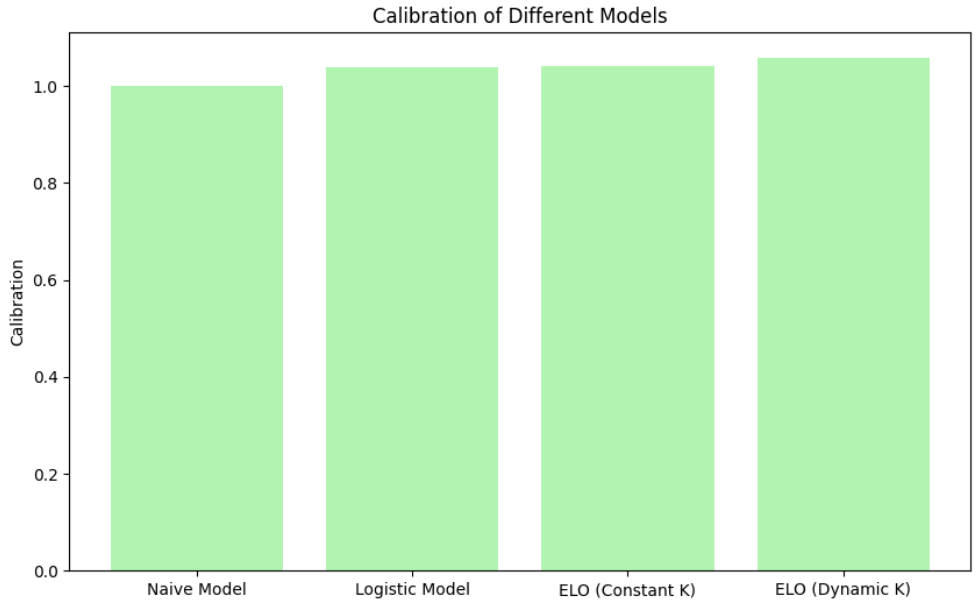

Figure 3: Calibration comparison of different models.

confidence. The Logistic regression model has the highest log-loss which shows its difficulty in balancing prediction confidence with accuracy. ELo Model with dynamic K has slightly higher log-loss than the constant K model but owing to its good accuracy it is still doing great. The naive model also has a good log loss but still not the best.

4.1 Model Performance Summary

The ELO Model with Dynamic K is the best model overall as it provides predictions with higher accuracy and good performance in terms of log-loss. Its ability to adapt based on the experience of the player makes it the most lucrative model for our dataset. For scenarios where confidence on the prediction is more critical, logistic regression can be chosen as a best model. The other ELO model with constant K wasn't too far eitherin terms of its metric values, but the extra factor of number of matches played improves the validation metrics a bit.

5 Discussion

5.1 Naive Model

The simplicity of the Naive model is the primary strength of the model. Since the complexity is less, it requires less computational resources and

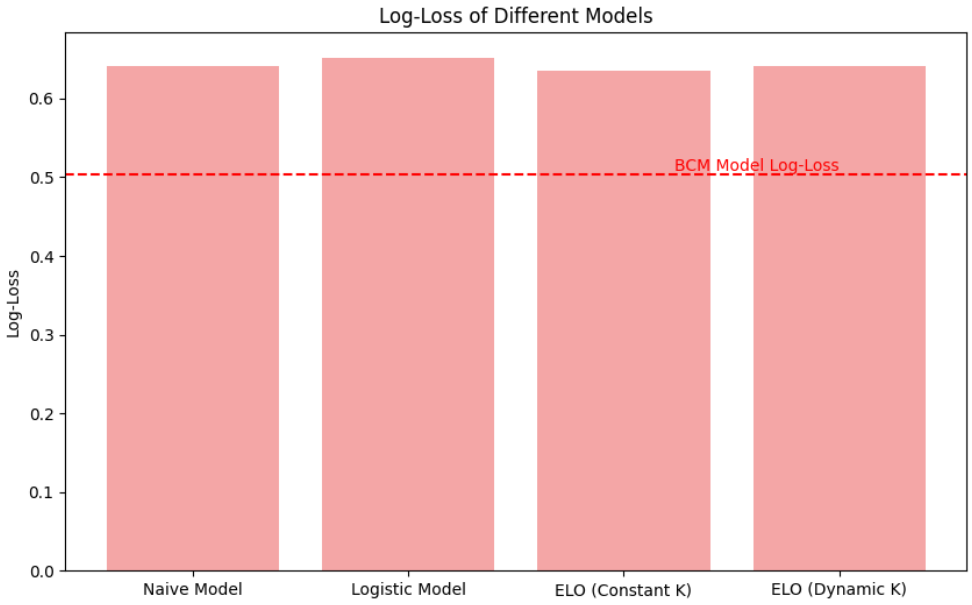

Figure 4: Log-loss comparison of different models, with the BCM model log-loss represented by the red dashed line.

it is easy to understand and implement. This simplicity can be beneficial in a scenario where quick and rough calculations are required. The simplicity of this model can also be its weakness as it fails to capture the intrinsic patterns of the data. It does not take into account other contextual variables such as player's form, injuries, surface type, or the tournament level. Hence the model is not the ideal choice for the prediction of the tennis matches outcome.

5.2 Logistic Model

The Logistic model offers a more sophisticated approach by incorporating the difference in the ranking points as the predictor variable as compared to the Naive model. It also gives flexibility in choosing a different predictor variable like ratio of ranking points which can help in identifying underlying patterns in different types of data. However, a notable limitation of the logistic regression model is its assumption of linearity in the log odds of the outcomes. This assumption may not hold true in cases where the relationship between the predictors and the target variable is non-linear, potentially resulting into less accurate predictions which was the case for our dataset.

5.3 ELO model (Constant and Dynamic K)

The major strength of both the ELO models is the usage of the ELO ratings instead of the ATP ratings as the predictor variable as it allows the model to be more sophisticated and realistic in computing the probability of the outcome of the match. This dynamic rating system captures the relative strength of the players which can potentially lead to better accuracy in prediction which was the case with our dataset. The only limitation with these models could be the computation cost for optimizing them, especially the ELO model with dynamic K as it involves multiple parameters that can take range of values. For larger datasets, the computational cost and the time it takes to compute can be vital factors that might make them not suitable for such datasets.

5.4 Conclusion:

In this analysis, our primary goal was to predict the tennis matches outcome for the year 2019 using the dataset spanning from years 2005-2018. We performed an in-depth analysis of various models to predict the outcomes of tennis matches, with a particular focus on men's tennis data. We explored models such as Naive, Logistic Regression, ELO models with constant and dynamic K and the Bookmakers Consensus Model (BCM), which was our benchmark model for comparison. We were able to compare our model on validation metrics such as Accuracy, Calibration and Log-Loss.

The result of our analysis showed that the ELO model with dynamic K outperformed the other models overall with the validation metrics and that it is the most suitable model that can be used for our usecase. The model's dynamic K-factor, which adjusts according to a number of matches played, enabled it to deliver more precise and accurate predictions. This adaptability highlights the model's effectiveness in capturing the subtle variations in player performance over time.

This analysis has also offered a glimpse into the betting world of the sports industry, emphasizing the advantages of using sophisticated betting odds from renowned bookmakers to make predictions. The analysis has also pointed out areas where enhancements are needed, especially in integrating additional predictor variables like surface type and investigating more complex models such as Glicko and TrueSkill.

Overall, this project has laid a strong foundation for further exploration and refinement in sports prediction modeling, offering a comprehensive comparison of traditional and advanced methods within the realm of tennis.

References

[1] S. J. Ahmed, "Machine learning in sports analytics and performance prediction — dataduniya," Medium, 07 2023. [Online]. Available: https://medium.com/dataduniya/machine-learning-in-sports-analytics-and-performance-prediction-d7f50799f684

[2] W. Contributors, "Tennis," Wikipedia, 02 2019. [Online]. Available: https://en.wikipedia.org/wiki/Tennis

[3] A. De Seranno, "Predicting tennis matches using machine learning," 2019. [Online]. Available: https://libstore.ugent.be/fulltxt/RUG01/002/945/727/RUG01-002945727_2021_0001_AC.pdf

[4] "Tennis betting — tennis results — tennis odds," Tennis-data.co.uk, 2019. [Online]. Available: http://www.tennis-data.co.uk/alldata.php

[5] "Elo rating system," Wikipedia, 02 2024. [Online]. Available: https://en.wikipedia.org/wiki/Elo_rating_system#:~:text=A%20player%27s%20Elo%20rating%20is

[6] J. Stübinger, B. Mangold, and J. Knoll, "Machine learning in football betting: Prediction of match results based on player characteristics," *Applied Sciences*, vol. 10, p. 46, 12 2019.