Study of tennis athletes performance on different surfaces and factors that affect wins

2024-12-15

Alejandro Paredes La Torre, Liangcheng (Jay) Liu

Nzarama Michaella Desire Kouadio, Jahnavi Maddhuri

Abstract

Modeling the key factors that influence professional tennis match outcomes is crucial for athletes, coaches, and researchers seeking to optimize performance and enhance player development. This study utilizes ATP data to examine how player rankings and aces impact match duration and outcomes. The analysis reveals that larger ranking disparities are associated with shorter match durations, suggesting that players with higher rankings are more likely to win quickly. In addition, the study investigates the relationship between the number of aces a player hits and their probability of winning, finding that more aces significantly improve a player's chances of victory. The effect of aces varies across different court surfaces, with clay courts showing the strongest influence. This research provides valuable insights into the complex dynamics of tennis matches and highlights the importance of surface type, player ranking, and performance metrics like aces in shaping match outcomes.

Introduction

Predicting outcomes in sports has long been a central focus for athletes, teams, and industries alike. Statistical science, when applied to sports, plays a pivotal role in optimizing an athlete's performance, making it a subject of significant interest for both researchers and commercial enterprises. In professional tennis, analyzing the factors that influence match outcomes offers insights that extend beyond simple predictions, encompassing strategic advancements and broader applications.

Such insights have practical implications for refining rankings (Klaassen and Magnus, 2003)[16], ratings (Kovalchik, 2016)[17], and seedings(Boulier et. al 1999)[15], serving as a foundation for performance analysis and strategy optimization. Predictive models are pivotal in industries like sports betting, where they drive decision-making and risk management (Foley-Train, 2014)[18]. . In academic research, these models contribute to methodological improvements, as seen in the work of Štrumbelj & Vračar (2012)[6], who explored probabilistic approaches to predicting match outcomes.

Analyzing professional tennis data bridges practical applications and theoretical advancements. Studies like those by Boulier and Stekler (1999)[7], Lasek, Szlávik, and Bhulai (2013) [8], and McHale and Davies (2007)[9] highlight how data-driven approaches can uncover nuanced patterns in sports, offering a deeper understanding of the game and its predictive dynamics

A growing body of literature underscores the role of data-driven approaches in uncovering intricate patterns within tennis. For instance, Boulier and Stekler (1999)[7] demonstrated how statistical models could be used to assess player performance, while Lasek, Szlávik, and Bhulai (2013)[8] examined predictive methods for ranking systems. McHale and Davies (2007)[9] explored the influence of match dynamics on outcomes, further bridging theoretical advancements with practical applications.

Early foundational studies, such as those by Newton and Keller (2005)[2], O'Malley (2008)[3], and Riddle (1988)[4], established that under the assumption of independent and identically distributed (iid) point outcomes on a player's serve, the probability of winning a match could be derived from serve-point probabilities. These studies laid the groundwork for subsequent research in probabilistic modeling of tennis outcomes.

Kovalchik (2016) [5] conducted a comparison of 11 published tennis prediction models, categorizing them into three classes: point-based models relying on the iid assumption, regression-based models, and paired comparison models. The study found that point-based models had lower accuracy and higher log loss, regression and paired comparison models generally outperformed them.

Furthermore, given the importance of player rankings and performance metrics in tennis prediction models, it is essential to understand the context provided by the ATP (Association of Tennis Professionals). This institution is the principal governing body of men's professional tennis. It oversees the ATP Tour, which features the highest level of men's tennis tournaments worldwide, including Grand Slams, Masters 1000 events, and other competitive circuits. The ATP rankings system, introduced in 1973, is used to evaluate and rank players based on their performance in sanctioned tournaments over a rolling 52-week period, serving as a critical metric for seedings and qualifications (ATP Tour, n.d.)[19].

Building on these studies, this research uses the Tennis ATP dataset (Sackmann, 2021), a collection of information for professional tennis matches and professional players, to investigate two specific questions:

1. How does the difference in player rankings influence the duration of a tennis match?
2. How do the number of aces and the court surface type affect a player's odds of winning a match?

Methods

Data and preprocessing

The dataset used in this study is the Tennis ATP Dataset curated by Jeff Sackmann (2021) [1]. This dataset serves as a comprehensive repository of professional tennis data, encompassing a wide range of player information, historical rankings, match outcomes, and statistical metrics. This dataset serves as a valuable resource for analyzing trends and performance in professional tennis, forming the basis for addressing the research questions in this study.

The time frame selected includes ATP match data from 2014-2024, the subsets chosen are challenger matches and professional and tournament class A matches such as Davis Cup, Roland Garros and others. The records from this period consist in 116,103 matches where each match has 49 variables.

Pertaining the first research question, to assess the effect of factors on the match length, the difference in the number of aces, rankings, and ranking points between the winner and loser were calculated. These perfomance metrics aim to provide more tangible and interpretable predictors for the model by focusing on measurable aspects of the players performance. Grouping variables into performance metrics also helps to frame the analysis in a way that aligns with the context of the sport and makes the regressors more meaningful and insightful. Based on a VIF analysis, highly correlated variables were addressed

by introducing derived metrics, such as the first break point win ratio, to mitigate multicollinearity and improve model stability.

To address the issue of missing values, a multiple imputation approach was applied using the mice package. This method generates several imputed datasets to account for the uncertainty inherent in estimating missing values, thereby enhancing the robustness of subsequent analyses. Predictive mean matching (PMM) was employed as the imputation technique, combining regression-based prediction with donor-based imputation. This ensures that imputed values align with the observed data distribution, remaining realistic and within plausible ranges.

In respect to the second research question, which focuses on analyzing match outcomes, additional data processing steps were taken. The original dataset, which contains match-level features with information on both the winner and loser in a single record, was restructured. This transformation reorganized the data at the player level, ensuring that each record represents either a win or a loss for a given player. The dimensions for this transformed dataset are 218321 records and 30 variables. This adjustment allows for a more focused analysis of the factors influencing match outcomes. Furthermore, in order to deal with multicolinearity new derived variables such as the first serve win ratio, calculated from two underlying variables, were used to merge highly correlated variables, the procedure being explained in detail in the results section.

For the second research question, 4,739 records (2%) with missing values in the aces column were excluded. This decision was based on the observation that missing data in aces systematically coincided with missing values for all other match statistics. A similar approach was applied to the serving games variable 4,740 records (2%), as its missing data also indicated the absence of other critical match details, further supporting exclusion. For the remaining variables without systematic missingness, a multiple imputation technique was applied, informed by the results of the imputation method used in the first research question.

Variable selection

Taking as reference previous research regarding the most relevant features involved in the outcome of a tennis match (Newton et al., 2005[1]; O'Malley 2008[2]; Kovalchik, 2016) a pre selection was made observing the limitation of the available data. In order to further refine the process of feature selection exploratory data analysis was conducted using correlation plots, box plots and scatterplots.

Model fitting and evaluation

In regard to the first research question, Multiple linear regression (MLR) was used to analyze the factors influencing match duration. Variance Inflation Factor (VIF) revealed collinearity among some variables, prompting the creation of ratios for breakpoints saved and faced to address this issue. Assumptions of linear regression, such as normality, homoscedasticity, and linearity, were evaluated using diagnostic plots. To address deviations from normality, a log transformation was applied to the target variable, match duration, which significantly improved the distribution of residuals. Residual versus fitted plots and Q-Q plots were used to assess model assumptions and ensure the validity of results. R-squared was used to evaluate the overall performance of the linear regression model. All statistical analyses were performed using R programming language (version 4.3.1)

Regarding the second research question a logistic regression model was developed to predict the outcome of a tennis match, with the dependent variable being a binary outcome representing a win or loss. A series of predictor variables were selected based on relevant literature and expert knowledge, including player

attributes, tournament characteristics, and performance metrics. The initial model included a variety of raw variables such as player height, age, rank, double faults, aces, serve statistics, and tournament-level indicators. However, multicollinearity among these predictors was detected using Variance Inflation Factor (VIF) scores, which were notably high for several variables related to serve and break-point statistics. To mitigate multicollinearity, new derived variables such as the first serve win ratio, second serve win ratio, and break-point save ratio were created. Multiple imputation was applied to handle missing data, followed by the use of pooled regression coefficients to account for uncertainty in the imputed data. The receiver operating characteristic (ROC) curve and area under the curve (AUC) were computed to evaluate the classification performance. Finally, binary predictions were generated, and a confusion matrix was produced, along with several classification metrics including accuracy, precision, recall, F1 score, specificity, and sensitivity.

Results

Overview of key variables of interest

The dataset shows a fairly balanced distribution of players who win (101,746) versus those who lose (111,621), indicating minimal information gain from either outcome. Regarding surface types, most matches are played on Hard (53.5%) or Clay (41.1%) courts, with Grass accounting for 5% and Carpet only 0.2%. The distribution of key continuous variables is examined below. Missing values are present in 9% of the minutes variable and 2% of the aces variable, primarily corresponding to canceled or rescheduled matches.

Table 1: Summary Statistics for Variables

Variable	Min	Mean	Median	Std..Deviation	Max
minutes	1	100.65	94	39.79	4756
rank	0	270.95	202	275.01	2257
aces	0	4.76	4	4.38	75

Research question 1: Effects of the difference in ranking over the length in minutes for a tennis match

Match duration in tennis is influenced by a variety of factors, including tournament level, player performance, court surface, and physical characteristics. During model development, we assessed multicollinearity among predictors using the Variance Inflation Factor (VIF). High collinearity was identified between the breakpoints saved and breakpoints faced for winners and losers. To address this issue and improve model stability, these variables were combined into ratios of breakpoints saved and ratios of breakpoints faced. This transformation not only reduced redundancy but also provided a clearer representation of the relative performance between winners and losers during matches.

Tournament level alone significantly affects match duration, with a clear trend emerging when comparing lower and higher levels to the baseline (level A). Matches at level D (Beginner) and level C (Intermediate) tend to be shorter, with level C showing a 3.1% decrease in duration compared to level A (95% CI: -3.7%, -2.5%, $p < 0.001$). This indicates that lower-level tournaments are associated with shorter match durations. However, as tournament levels rise above C, match duration increases significantly. For instance, matches at level G extend by 21.9% (95% CI: 20.9%, 22.8%, $p < 0.001$).

The interaction between rank difference and tournament level adds further nuance to the findings. When considering the interaction between rank difference and tournament level, the effect varies depending on competitiveness. At level C (Intermediate), the interaction term is significant, with match duration increasing slightly as rank disparity grows, approximately 0.01%. In contrast, at level G (Advanced Recreational), the interaction term is also significant but negative, showing that match duration decreases by 0.01% as rank disparity increases. These results indicate that at intermediate levels, skill differences can prolong matches due to competitive balance, while at higher levels like G, greater rank disparities lead to quicker resolutions due to clearer dominance. For levels D (Beginner), F (Competitive), and M (Professional), the interaction terms are not significant, suggesting that rank disparity does not meaningfully influence match duration at these levels.

Differences in player performance, such as aces, also influence match duration. For every additional ace difference between players, match duration increases by 1% (95% CI: 0.9%, 1.1%, $p < 0.001$). This indicates that strong serving performances often lead to prolonged matches, as players must engage in more games and sets. Our study also highlights how service errors extend gameplay, likely due to increased rally lengths or additional points. Each additional double fault by the winner increases match length by approximately 5.1% (95% CI: 5.0%, 5.2%, $p < 0.001$).

Furthermore, the ratio of breakpoints saved by players shows contrasting effects. Winners who save more breakpoints reduce match duration by 12.6% (95% CI: -13.3%, -11.9%, $p < 0.001$), likely due to their ability to close out critical points effectively. On the other hand, losers who save more breakpoints increase match length by 48.1% (95% CI: 47.2%, 49.1%, $p < 0.001$). This demonstrates how defensive efforts by losers prolong gameplay, as they manage to stave off losing points but fail to secure victory. The court surface plays a pivotal role in match duration. Matches on clay courts last significantly longer, with duration increasing by 14% (95% CI: 10.1%, 17.8%, $p < 0.001$). However other types of surfaces have shown no significant impact on length of the match.

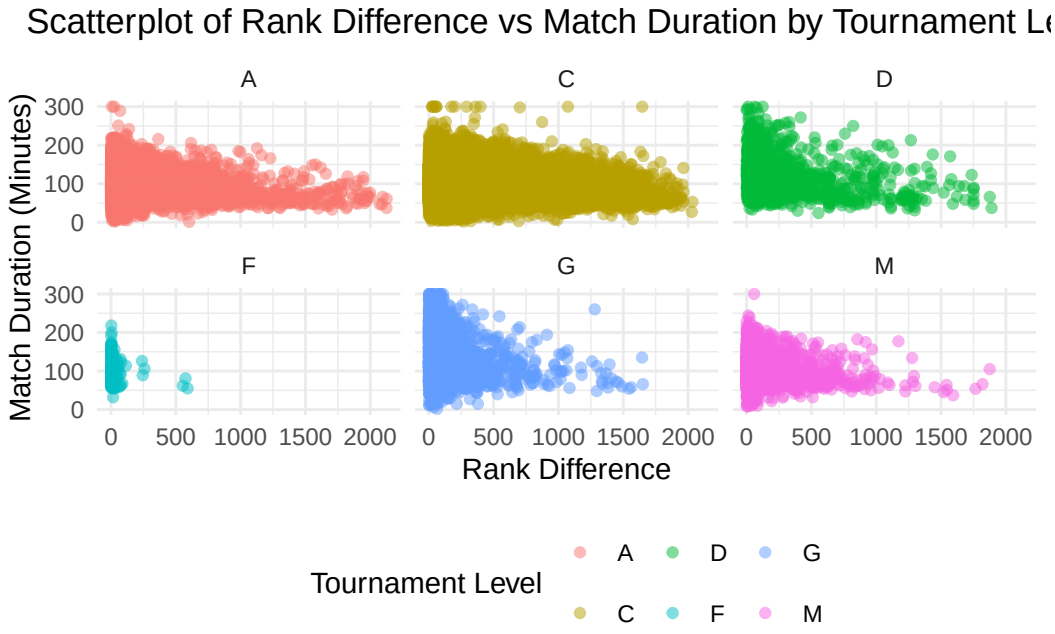
The overall model adjusted R-squared is 0.31.

Table 2: Linear Regression Model Results

Predictors	Coefficient	Standard Error	95% CI	P-Value
diff_rank	-0.0002	0.00	0, 0	<0.001
tourney_levelC	-0.0309	0.00	-0.04, -0.03	<0.001
tourney_levelD	0.1442	0.01	0.12, 0.17	<0.001
tourney_levelF	0.0792	0.02	0.03, 0.13	0.001
tourney_levelG	0.2185	0.00	0.21, 0.23	<0.001
tourney_levelM	0.0249	0.00	0.02, 0.03	<0.001
diff_aces	0.0100	0.00	0.01, 0.01	<0.001
winner_ht	-0.0031	0.00	0, 0	<0.001
loser_ht	-0.0006	0.00	0, 0	0.006
surfaceClay	0.1396	0.02	0.1, 0.18	<0.001
surfaceGrass	0.0144	0.02	-0.03, 0.05	0.477
surfaceHard	0.0606	0.02	0.02, 0.1	0.002
winner_age	0.0014	0.00	0, 0	<0.001
loser_age	0.0003	0.00	0, 0	0.118
w_df	0.0509	0.00	0.05, 0.05	<0.001
l_df	0.0167	0.00	0.02, 0.02	<0.001
w_bpSave_ratio	-0.1259	0.00	-0.13, -0.12	<0.001
l_bpSave_ratio	0.4813	0.00	0.47, 0.49	<0.001
diff_rank:tourney_levelC	0.0001	0.00	0, 0	<0.001

Predictors	Coefficient	Standard Error	95% CI	P-Value
diff_rank:tourney_levelD	0.0000	0.00	0, 0	0.25
diff_rank:tourney_levelF	-0.0004	0.00	0, 0	0.116
diff_rank:tourney_levelG	-0.0001	0.00	0, 0	<0.001
diff_rank:tourney_levelM	0.0000	0.00	0, 0	0.624

It is important to examine the underlying assumptions of our model to ensure that the regression estimates are both reliable and valid, providing meaningful insights into the data. The model assumptions is assessed by residual vs. fitted plot. The log transformation was applied to our target variable (minutes) in order to reduce skewness, stabilize variance, and improve the normality of the residuals, ensuring the assumptions of linear regression are better met. Even though, we observe slight deviations in the normality curve and a pattern that suggests non linearity, the assumptions are not severely violated, which is acceptable for reliable linear regression estimates.



This scatterplot shows the relationship between Rank Difference (x-axis) and Match Duration in Minutes (y-axis), faceted by Tournament Level. In each tournament level (A, C, D, F, G, M), Rank Difference is skewed towards lower values (concentrated near 0), indicating that matches are typically played between players with similar ranks. Plus, Matches with small rank differences (competitive matches) tend to have longer durations.

Research question 2: Aces and court surface type influence in match outcome

To structure this model the relevant variables in related literature as well as the variables of interest along with the interaction term of the type of surface were included. To evaluate multicollinearity, the VIF score was used on an initial logistic regression model including all the variables selected. The raw variables representing a player's total serve points, number of first serve points made, number of first serve points won, number of second serve points won, number of break points faced, number of break points saved, total draw size in the tournament and the tournament level were all highly correlated. It

was found that total number of serves attempted, first serves attempted, first serve points won and second serve points one were highly correlated with VIF scores ranging from 9 to 72.

To continue to capture this information, but limit multicollinearity, we combine these into the total first serve points won ratio which is a ratio of the total first serve points won to the total first serves attempted. Similarly, the total second serve points won ratio is the ratio of the total second serve points won to the difference between the total serves attempted and the total first serves attempted. The new VIF scores for these two ratios were less than 1.3 Similarly, the break points saved ratio is used in place of the overall counts. Lastly, between draw size and tournament level, draw size is only used as there is more granular information derived from draw size than tournament level.

Mentioned in the methodology section of this document, a multiple imputation technique with the mice package was applied to handle non systematic missing data correspondent to the variable height. Presented below is a summary of the final logistic regression model.

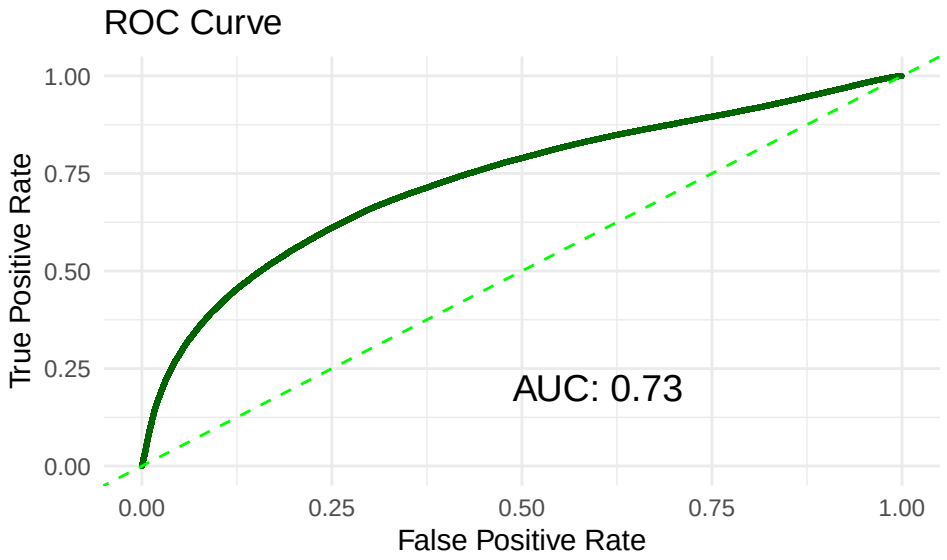
Table 3: Logistic Regression Coefficients, Odds Ratios, Confidence Intervals, Standard Errors, and P-values

	Variable	Coefficient	Standard Error	Odds Ratio	95% CI	P-value
2	draw_size	0.00	0.00	1.00	1,1	<0.001
3	player_handL	0.24	0.02	1.27	1.22,1.33	<0.001
4	player_handR	0.25	0.02	1.29	1.24,1.33	<0.001
5	player_height	-0.02	0.00	0.98	0.98,0.99	<0.001
6	player_age	-0.02	0.00	0.98	0.98,0.99	<0.001
7	rank	0.00	0.00	1.00	1,1	<0.001
8	rank_points	0.00	0.00	1.00	1,1	<0.001
9	aces	0.08	0.02	1.09	1.05,1.13	<0.001
10	surfaceClay	0.36	0.18	1.43	1,2.05	0.053
11	surfaceGrass	0.02	0.19	1.02	0.7,1.47	0.927
12	surfaceHard	0.22	0.18	1.24	0.87,1.79	0.238
13	double_faults	-0.17	0.00	0.85	0.84,0.85	<0.001
14	break_pt_save_ratio	0.02	0.00	1.02	1.02,1.03	<0.001
15	aces:surfaceClay	0.02	0.02	1.02	0.98,1.06	0.255
16	aces:surfaceGrass	-0.01	0.02	0.99	0.96,1.03	0.659
17	aces:surfaceHard	0.00	0.02	1.00	0.96,1.03	0.905

Analyzing the variables of interest aces and surface, considering that the variable aces is statistically significant while the interation terms are not, we have a combined effect of 1.11 times increase in odds of winning for every extra ace point when the match is disputed on a clay surface, with a combined confidence interval between 1.03 and 1.19. Similarly, for matches played on Grass the model has a combined effect increasing the odds of winning by 1.07 times. The impact of every extra ace point while playing on hard surfaces represent a combine effect of 1.09 increase in the odds of winning the match.

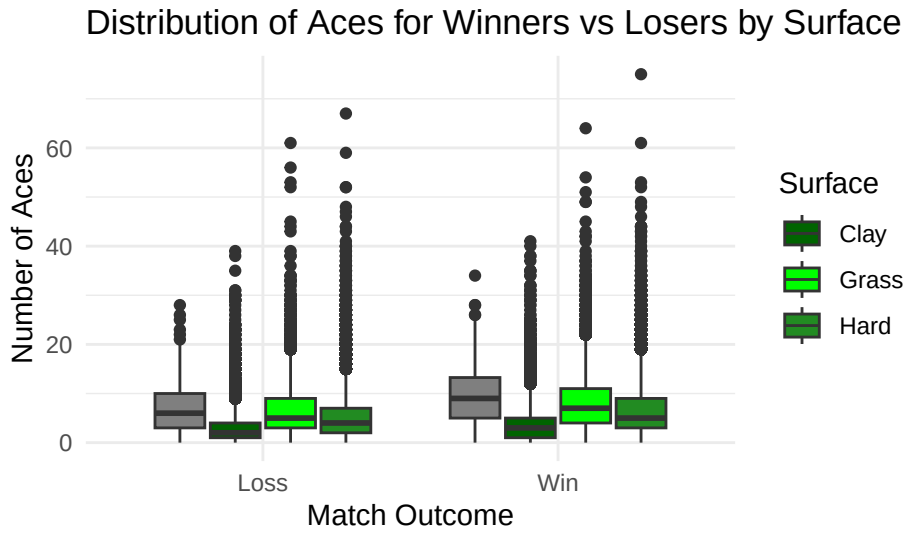
Another statistically significant variable impacting the odds of winning is the referred to right vs left-handed player. Specifically being Right or Left Handed (as opposed to ambidextrous, the base level) is associated with approximately 1.27 times and 1.29 times increase in odds of winning respectively. Furthermore, the break point serve ration explains an effect of 1.03 times the increase in odds of winning, holding the other variables constant. Finally, statistically significant variables related to the player

attributes showed that an increase of a year for a player's age reduces the odds of winning by 2%, while for every extra inch for player the odds of winning reduce by 2%.



The performance of the logistic regression model was evaluated using standard classification metrics. The model achieved an accuracy of 0.68. The precision of the model was 0.68, reflecting its ability to correctly identify positive cases while minimizing false positives. The recall was measured at 0.70, demonstrating the model's capability to correctly identify a high proportion of actual positive cases. Finally, the F1 score, a harmonic mean of precision and recall, was calculated to be 0.8192, indicating a balanced performance between these two metrics. Together, these results suggest the model performs reliably in predicting match outcomes based on the given features.

Finally, shown below, a plot showing the average of aces achieved in a game is slightly higher in respect to the average of aces by Losers. This supports the model estimates and the slight increase in odds of winning (by 1.1 times) for every additional ace played.



Conclusion

This study identified key factors influencing tennis match durations and outcome, shedding light on how player characteristics, tournament conditions, and performance metrics contribute to match length. Notably, tournament levels emerged as a significant determinant, with higher-level tournaments, such as G and M, associated with substantially longer matches, reflecting greater competitiveness and intensity. While tournament level alone drives significant differences in match duration, particularly at higher tiers, the interaction with rank difference reveals that player ranking disparities are only influential under specific conditions. In intermediate and advanced tournaments (levels C and G), rank disparity slightly amplifies match duration, while at other levels, its effect remains negligible. Other significant predictors included differences in player performance, such as the number of aces and breakpoints saved, both of which revealed nuanced relationships with match duration. Physical characteristics, such as player height, also contributed, albeit modestly, suggesting a strategic role in gameplay.

For match outcomes, the logistic regression model identified aces, court surface, player handedness, draw size, and breakpoint save ratio as significant predictors. An additional ace increased the odds of winning across surfaces, with the strongest effects observed on clay (1.11 times), followed by hard (1.09 times) and grass (1.07 times). The breakpoint serve ratio also contributed, with an incremental effect of 1.03 times on winning odds. The player's attributes such as height or age showed that every extra unit of increase of the respective variables reduce the odds of winning. The logistic regression model demonstrated robust predictive performance, achieving an accuracy of 68%, a precision of 68%, and a recall of 70%. The F1 score of 0.82 reflects a balanced capacity to minimize false positives and false negatives.

Despite these insights, the study faced limitations. The reliance on imputed data may have introduced bias, while the linear regression model's assumptions, such as normality, were not perfectly met despite transformations. Plus, for the multiple linear model the R squared explained 31% of the variance in match duration ($R^2 = 0.31$), leaving room for unaccounted influences such as weather, fatigue or external conditions, potentially leaving room for unexplored contributors to match durations. These limitations may have influenced the precision of the model's estimates

Future research should address these limitations by refining data collection, especially for lower-level tournaments, and incorporating additional variables. By exploring these dimensions, future analyses could provide an even more comprehensive understanding of the factors shaping match durations, aiding in tournament planning and player preparation.

References

[1] Sackmann, J. (n.d.). Tennis databases, files, and algorithms [Data set]. Tennis Abstract. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. Based on a work at https://github.com/JeffSackmann.

[2] Newton, P. K., & Keller, J. B. (2005). Probability of winning at tennis I. Theory and data. Studies in applied Mathematics, 114(3), 241-269.

[3] O'Malley, A. J. (2008). Probability formulas and statistical analysis in tennis. Journal of Quantitative Analysis in Sports, 4(2).

[4] Riddle, L. H. (1988). Probability models for tennis scoring systems. Journal of the Royal Statistical Society Series C: Applied Statistics, 37(1), 63-75.

[5] Kovalchik, S. A. (2016). Searching for the GOAT of tennis win prediction. Journal of Quantitative Analysis in Sports, 12(3), 127-138.

[6] Štrumbelj, E., & Vračar, P. (2012). Simulating a basketball match with a homogeneous Markov model and forecasting the outcome. International Journal of Forecasting, 28(2), 532-542.

[7] Boulier, B. L., & Stekler, H. O. (1999). Are sports seedings good predictors?: an evaluation. International Journal of Forecasting, 15(1), 83-91.

[8] Lasek, J., Szlávik, Z., & Bhulai, S. (2013). The predictive power of ranking systems in association football. International Journal of Applied Pattern Recognition, 1(1), 27-46.

[9] Sklenička, J. (2024). Predicting the outcomes of tennis matches. How important is the factor of different surfaces?.

[15] Boulier, B. L., & Stekler, H. O. (1999). Are sports seedings good predictors?: an evaluation. International Journal of Forecasting, 15(1), 83-91.

[16] Klaassen, F. J., & Magnus, J. R. (2003). Forecasting the winner of a tennis match. European Journal of Operational Research, 148(2), 257-267.

[17] Kovalchik, S. A. (2016). Searching for the GOAT of tennis win prediction. Journal of Quantitative Analysis in Sports, 12(3), 127-138.

[18] Foley-Train, J. (2014). Sports betting: Commercial and integrity issues. Report prepared for the Association of British Bookmakers, European Gaming and Betting Association, European Sport Security Association and Remote Gambling Association. Retrieved January, 21, 2015.

[19] ATP Tour. (n.d.). About the ATP. Retrieved from https://www.atptour.com