Study of tennis 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

This study explores how player rankings and aces affect tennis match outcomes using Association of Tennis Professionals (ATP) data. The first research topic examines the impact of ranking differences on match duration. The research then evolves to investigate the relationship between the number of aces by a tennis player and that player's odds of winning.

Methodology to the study includes exploratory data analysis combined with linear and logistic regression models, supported by visualizations. These findings highlight the connection between player rankings, aces, and match outcomes, while emphasizing the interactive impact of surface type.

Results show that larger ranking gaps lead to shorter matches, though surface type and match conditions also influence duration. Additionally, the findings indicate that hitting more aces improves the odds of winning, with surface types like clay and hard courts playing a significant role in this relationship.

Introduction

A substantial body of research has explored the prediction of tennis match outcomes using statistical models, highlighting the importance of player attributes and match statistics. Early studies, such as those by Newton and Keller (2005)[2], O'Malley (2008)[3], and Riddle (1988)[4], demonstrate that under the assumption of independent and identically distributed (iid) point outcomes on a player's serve, the probability of winning a match can be derived from the probabilities of winning points on serve.

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 while point-based models had lower accuracy and higher log loss, regression and paired comparison models generally outperformed them.

Methods

Data and preprocessing

The dataset utilized in this study is the Tennis ATP Dataset curated by Jeff Sackmann (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. Specifically, it includes a player file containing detailed biographical data, such as unique player identifiers, names, handedness, birth dates, nationalities, and physical attributes like height. Additionally, ranking files provide a historical record of ATP rankings over time, while the results file covers match outcomes across tour-level, challenger, and futures events. This dataset forms a robust foundation for exploring various aspects of professional tennis performance and trends.

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

The initial collection of data contains features at the match level, therefore it has information from the winner player and the loser player. In order to analyze the effect of match win this structured has been modified to portrait the results at the player level, it has been added the feature win which represents the outcome of the match in respect to the player (either win or loose).

In order to improve the quality of the data, those players that do not have a rank and rank points have been set to zero as they explain unranked players new to these tournaments. An inputing technique has been used for the player height using the average of the country of birth of the player.

Furthermore, analyzing that matches whose number of aces for the winner nor the loser are missing lack all the statistical information from the the other aspects of the match therefore those records have been filtered. The same rationale applies to the serving games as it signals records that have missing information overall therefore those records have been filtered outside of the analysis data.

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.

Modeling and evaluation

The present study focuses on the effects of the duration of a match using linear regression to evaluate inference capabilities and determining the principal factors for a match win using logistic regression to evaluate the probability of a win. Variance Inflation Factor (VIF) was used to test multi-colinearity. For the linear regression task the assumptions for the model are tested via residual vs fitted plots and normal q-q plots, furthermore, the performance of the model is evaluated using the adjusted r squared metric. In terms of logistic regression, accuracy, recall sensitivity and specificity are

Results

Overview of exploratory data analysis

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

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

The results for the final fitted model for win prediction are in the Anex I since the table is large. The selected variables along with the interaction term of the type of surface was included. Multiple iterations to find the best model was performed and multicolinearity evaluations were used to assess the model.

$$\begin{aligned}\log\left(\frac{P(\text{win})}{1-P(\text{win})}\right) = \beta_0 \\ &+ \beta_1 \cdot \text{draw size} \\ &+ \beta_2 \cdot \text{tourney level} \\ &+ \beta_3 \cdot \text{match num} \\ &+ \beta_4 \cdot \text{player hand} \\ &+ \beta_5 \cdot \text{player height} \\ &+ \beta_6 \cdot \text{player age} \\ &+ \beta_7 \cdot \text{rank} \\ &+ \beta_8 \cdot \text{rank points} \\ &+ \beta_9 \cdot (\text{aces} \cdot \text{surface}) \\ &+ \beta_{10} \cdot \text{double faults} \\ &+ \beta_{11} \cdot \text{serve points} \\ &+ \beta_{12} \cdot \text{first serves} \\ &+ \beta_{13} \cdot \text{first serves points won} \\ &+ \beta_{14} \cdot \text{second serves points won} \\ &+ \beta_{15} \cdot \text{break points saved} \\ &+ \beta_{16} \cdot \text{break points faced} \\ &+ \beta_{17} \cdot \text{match month}\end{aligned}$$

Table 1: Logistic Regression Model Summary

	Term	Estimate	Standard.Error	z.value	P.value
(Intercept)	(Intercept)	2.639	0.756	3.489	0.000
draw_size	draw_size	0.001	0.000	3.015	0.003
tourney_levelC	tourney_levelC	0.473	0.017	27.565	0.000
tourney_levelD	tourney_levelD	0.210	0.065	3.229	0.001
tourney_levelF	tourney_levelF	-0.820	0.142	-5.767	0.000
tourney_levelG	tourney_levelG	0.036	0.053	0.675	0.500
tourney_levelM	tourney_levelM	-0.304	0.037	-8.273	0.000
match_num	match_num	0.000	0.000	8.472	0.000
player_handL	player_handL	2.129	0.699	3.044	0.002

	Term	Estimate	Standard.Error	z.value	P.value
player_handR	player_handR	2.206	0.699	3.155	0.002
player_handU	player_handU	2.017	0.699	2.885	0.004
player_height	player_height	-0.027	0.001	-26.844	0.000
player_age	player_age	-0.019	0.001	-13.631	0.000
rank	rank	-0.001	0.000	-17.859	0.000
rank_points	rank_points	0.000	0.000	21.934	0.000
aces	aces	-0.061	0.021	-2.951	0.003
surfaceClay	surfaceClay	1.510	0.217	6.970	0.000
surfaceGrass	surfaceGrass	0.449	0.222	2.020	0.043
surfaceHard	surfaceHard	0.952	0.217	4.399	0.000
double_faults	double_faults	0.026	0.003	8.501	0.000
serve_points	serve_points	-0.133	0.002	-63.659	0.000
first_serves	first_serves	0.006	0.002	3.117	0.002
first_serves_points_won	first_serves_points_won	0.220	0.003	77.097	0.000
second_serves_points_won	second_serves_points_won	0.229	0.003	74.769	0.000
break_points_saved	break_points_saved	0.603	0.007	84.608	0.000
break_points_faced	break_points_faced	-0.559	0.007	-76.005	0.000
match_month02	match_month02	-0.062	0.027	-2.273	0.023
match_month03	match_month03	0.142	0.030	4.725	0.000
match_month04	match_month04	0.106	0.029	3.615	0.000
match_month05	match_month05	-0.015	0.030	-0.488	0.626
match_month06	match_month06	-0.067	0.035	-1.916	0.055
match_month07	match_month07	-0.088	0.028	-3.107	0.002
match_month08	match_month08	0.069	0.029	2.420	0.016
match_month09	match_month09	-0.007	0.028	-0.248	0.804
match_month10	match_month10	-0.039	0.027	-1.467	0.142
match_month11	match_month11	-0.190	0.032	-6.025	0.000
match_month12	match_month12	0.200	0.074	2.689	0.007
aces:surfaceClay	aces:surfaceClay	-0.007	0.021	-0.333	0.739
aces:surfaceGrass	aces:surfaceGrass	-0.003	0.021	-0.130	0.897
aces:surfaceHard	aces:surfaceHard	0.000	0.021	-0.017	0.987

Analyzing the variables of interest aces and surface, holding every other variable constant, applying exponential we have an effect of 0.94 times of odds increase for every extra ace point. On the other hand surface Clay and surface hard showed to be statistically significant and their effects is an increase of 4.52 times the odds of winning for clay surface and an increase of 2.5 times the odds of winning for Hard surface.

Metric	Value
Accuracy	0.8117290
Precision	0.8229832
Recall	0.8155292
F1 Score	0.8192393
Specificity	0.8075600
Sensitivity	0.8155292
Positive Predictive Value	0.8229832
Negative Predictive Value	0.7996146

The performance of the logistic regression model was evaluated using standard classification metrics, including accuracy, precision, recall, and F1 score. The model achieved an accuracy of 0.8117, indicating that approximately 81.17% of predictions matched the true outcomes. The precision of the model was 0.8230, reflecting its ability to correctly identify positive cases while minimizing false positives. The recall was measured at 0.8155, 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.

Conclusion

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.