Breaking Serve, Not Breaking Rank: Match Statistics as a Leading Indicator of ATP Ranking Movement

William Mattox

Abstract

The ATP ranking is a 52-week running total of tournament points: transparent, dispute-proof, and slow to react to injury, decline, and breakout form. This study asks whether the statistics recorded inside a match — match length, the ratio of winners to unforced errors, and service breaks — carry information about a player's true level that the points table misses. Using 6,372 tour-level matches involving top-250 players from 2020 through 2024, we test three framings. First, match-level classification identifies the winner with over 90% accuracy against a 63% rank-only baseline, but this result is descriptive rather than predictive. Second, predicting the ranking gap between two players from a single box score effectively fails ($R^2 \approx .02$). Third, aggregating a full season of statistics per player recovers end-of-year rank well (Ridge, Spearman $\rho = .73$), and the result holds after win percentage is removed ($\rho = .71$). Frozen at the end of 2024, the stats model does not track future ranking better than ATP's own ranking ($\rho = .56$ versus .62), but it does add significant incremental signal ($R^2 = .511$ to .551), and that signal is concentrated where the two systems disagree most: among the top quarter of disagreements the model calls the direction of a player's future move correctly 66% of the time (19 of 29). Regarded on out-of-sample match outcomes rather than on future ATP rank, an ATP-plus-statistics blend beats ATP points outright, although a plain Elo rating beats every system tested and statistics add nothing once Elo is present. Match statistics are best used not as a replacement ranking but as an early-warning diagnostic that flags players whose level has drifted away from their points.

Keywords: tennis analytics, ATP ranking, match statistics, Elo rating, machine learning, sports forecasting

Background:

The Association of Tennis Professionals (ATP) ranking system is based on a series of points allocated for placing in each professional tournament. Each placement gives a different number of points based on how far the player reached in the tournament bracket (Quarterfinals, Semifinals) and the level of the tournament (e.g., Grand Slam, Masters 1000). Points are held for 52 weeks and dropped after the tournament in the following year. This ranking determines tournament entry, player seeding, and overall success in brand deals and earnings. It is deliberately basic and dispute-proof, which is the reason for its weakness.

For the game of tennis, many different pieces of data can be taken from each match. Tennis is scored based on points, games, sets, and matches. Break Points are when a player wins a game while the opponent is serving. Break point conversion is the most direct reference to who is winning or losing. We refer to breaks made and breaks faced/saved. Aces are unreturnable serves where the returner does not touch the ball at all, indicating serve dominance. A winner is a shot during the rally that the opponent does not touch, and an unforced error is a mistake by a player without pressure. Match length is the total time a match lasts, which can indicate skill similarity across matches.

Because points must be earned and then defended, the rankings are slow to react. A player coming back from injury has lost all previous points, even if they are still playing at a high level. However, a declining top player may still keep their high rankings until their points are

reset throughout the year. The ranking system also does not account for a player's archetype, whether they dominate with serve plus one or grind out each point, and the rankings cannot determine the difference between a blowout win (e.g. 6-0) and a narrow victory. These gaps are what this study addresses.

Introduction:

The ATP ranking is a transparent but basic ranking system. This raises the question: what if we look directly at the matches themself? Then we can create a system that captures the true ability of each player in a match instead of only caring about the outcome. This study follows the question drawn from our methodology. Can we predict the ATP ranking of tennis players by analyzing match length, ratio of winners to unforced errors, and service breaks? Our claim is not that statistics should completely replace the ranking system, but that they carry more information that can create a better ranking. The phrase "predict rankings" can have two different meanings. At the match level, where we predict the winner or the difference in ranking, and at the player level, where we use match statistics to output their end-of-season rank.

By using statistics from professional matches, the ranking system can better adapt to new players, declining players, and players returning from injury. Unlike the ATP ranking system, implementing statistics will allow for quicker growth for players, giving those who deserve spots in high-level tournaments a chance. Furthermore, high-level players such as Nick Kyrgios can reclaim their previous ranking more quickly after injury.

The main issue that we looked at first was which statistics to include in the calculation of the ranking. The main statistics we looked at are the winners-to-unforced-errors ratio, service aces, break points saved and won, and match length. By testing these statistics, we can find the ones that correlate the most with wins and which statistics do not matter as much in the outcome of a match. These stats are what we considered most important when it came to deciding the better player in a match.

Methods

We extracted data from Tennis Abstract, which contains almost every professional match played by professional players since the beginning of the ATP, as well as statistics on each player. For this project, we focused on data from 2020 to 2024. This includes concrete statistics, such as aces, break points faced/saved, match length, and player rankings, but also includes more biased stats, such as winners and unforced errors, where top players and higher-level tournaments are covered more. First, we limited the data to the top 250 ATP players, then found each match from those players that was covered by Tennis Abstract and included all the stats we needed. This resulted in a total of 6,779 matches covered throughout 2020-2024. To make sure the data in each match was correct, we then went through each one and dropped matches that did not include all we needed. This left us with a total of 6,372 matches. Then, for each match, we assigned all 250 players either player 1 (P1) or player 2 (P2) and found the difference in each statistic (P1 - P2).

The next part that we tested is whether match statistics could be used to predict the ranking gap between two players. To do this, we removed the rank difference stat used in the previous test and changed what we are asking the models to predict: rank differential. This test will show us if the winner of the match always shows better stats, which would mean the rank differential won't be calculated, or if the better-ranked player always has the better stats, giving us high accuracy. For this test, we used Ridge, Random Forest, and XGBoost. Ridge is best for

data with a clear, straight-line relationship between input features and the target. Random Forest is best for data with complex, non-linear patterns that straight lines cannot capture. XGBoost is best for clean, structured tabular data where maximum prediction accuracy is needed.

For the model, we used Logistic Regression, KNN, Random Forest, XGBoost, and SVM for classification, as well as Ridge, Random Forest, and XGBoost for regression. The rationale for the latter is to understand the cause of each outcome. A logistic model establishes a baseline and exposes coefficient signs, then tree/boosted models test for nonlinearity and interactions, and KNN/SVM find whether local or margin-based structure helps.

Now, we test if the statistics from each match throughout the year can predict the end rankings of each player, showing us whether or not using statistics can produce a valid ranking list. For this, we created a row for each player and year to create an average of statistics throughout the season and see if we can then predict the end-of-year ranking. Each player had a chart that looked just like the one below.

Field	Row 0	Row 1	Row 2	Row 3	Row 4
slug	adam_walton	adrian_mannarino	adrian_mannarino	adrian_mannarino	adrian_mannarino
year	2024	2020	2021	2022	2023
player	Adam Walton	Adrian Mannarino	Adrian Mannarino	Adrian Mannarino	Adrian Mannarino
matches_played	10	19	22	33	46
wins	2	9	6	16	27
avg_minutes	127.000000	125.789474	97.954545	114.242424	120.239130
avg_aces	6.200000	5.421053	3.909091	4.727273	4.239130
avg_bp_saved	3.900000	5.105263	3.500000	3.333333	3.956522
avg_bp_faced	6.200000	8.263158	6.545455	5.909091	6.500000
avg_breaks_made	1.700000	3.210526	1.590909	2.454545	3.065217
charted_matches	1	0	1	14	14
avg_winners	17.000000	NaN	17.000000	19.642857	22.142857
avg_unforced	23.000000	NaN	31.000000	26.357143	30.357143
win_pct	0.200000	0.473684	0.272727	0.484848	0.586957
bp_save_pct	0.629032	0.617834	0.534722	0.564103	0.608696

Table 1. Format of the player-season record. The table is transposed: each column is one example player-season row.

Then we assigned each player a ranking based on the ATP ranking system to give us a baseline that our models should try to predict. With both of these data points, we then graphed

them onto a chart to see if there is any correlation between each statistic and a player's end-of-year ranking.

Now, utilizing these statistics that we created, we can see how each model does with actually predicting the season ranking of each player. To do this, we need to remove the ranking statistic given in the previous test because now that is the stat we are trying to predict.

Because of the high accuracy, we decided to further improve the data used. The first step to this is removing the win percentage from the data because the win percentage is directly related to ATP points and ranking: the higher the win percentage, the more ATP points. If we still see similar results, then that means the statistics from a match are a good way to calculate ranking.

To further improve this finding, we wanted to see if winners vs. unforced errors still affect these predictions. To do this, we run the same test, but now we remove winners vs. unforced errors from the dataset and compare accuracy. To run this test, we had to shrink the dataset. Winners and unforced errors only exist in the charted subset of matches, which covers about 30% of our data, leaning towards top players and big tournaments. Because of this, we rebuilt the player-season table on only players who had at least five charted matches in a season (221 players), and compared three versions of the model side by side: the full lean model, the same lean model on a smaller charted subset, and the charted subset with the winners/unforced-error feature added back.

Next, we wanted to check where our model disagreed with the ATP and whether we were correct in the end or not. To test this, we froze the stats-only model's verdict at the end of 2024 and looked at where those same players ended up later. If the model thinks a player is better than their 2024 ranking says, and that player then climbs in 2025-2026, the model sees something the ranking wasn't able to catch. If the player drops instead, the model is incorrect.

After establishing what the model predicts, we ran the planned tuning experiments to see whether we could push it into being a sharper ranking than ATP. First, we tried building a blend, combining the ATP ranking and the stats prediction into a single ranking.

Finally, we tried to close the ranking gap of injury. Using the full match calendar, we built simple availability proxies for each player-season, number of mid-match retirements, longest in-season layoff, and number of long gaps, without any scraping or outside data.

So for the final test, we changed the target to something that the ATP formula does not produce: actual match results. We froze each ranking system at the end of the year and used it to predict who would win every match, then scored it based on who actually won. We compared five systems: ATP, Elo (a chess-style rating that updates after every match), our stats model, an ATP and stats blend, and an Elo and stats blend, across two seasons.

Results

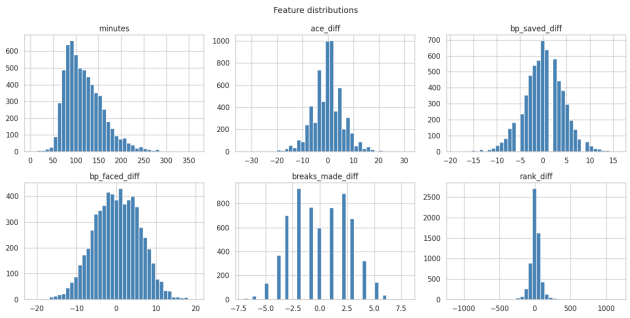

Figure 1. Match Stat Distribution Graph

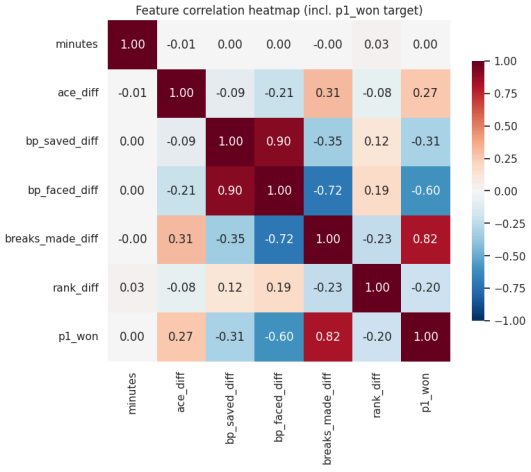

Figure 2. Heat Map of Stats

The rank-only baseline had an accuracy of 0.6332, which is expected because upsets do happen frequently. Now the goal is to improve on this accuracy by including each statistic.

	model	accuracy	logloss
0	Logistic Regression	0.917933	0.194365
1	Random Forest	0.916920	0.206636
2	SVM (RBF)	0.916920	0.195379
3	XGBoost	0.913374	0.195043
4	KNN (k=25)	0.912867	0.209233
5	Rank-only baseline	0.633232	NaN

Table 2. Accuracy of each Model

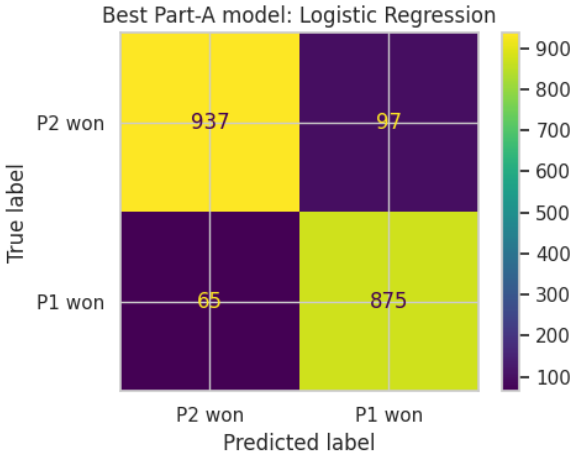

Figure 3. Best Model

As seen from Table 2, each model shows an accuracy above 0.9, which is very uncommon. This tells us that the statistics we chose almost directly contributed to the result of the match, and that by looking only at the statistics, you could better predict the winner than with rankings alone. Although a high accuracy is what you are looking for, we decided to dig deeper into what is directly causing this high number.

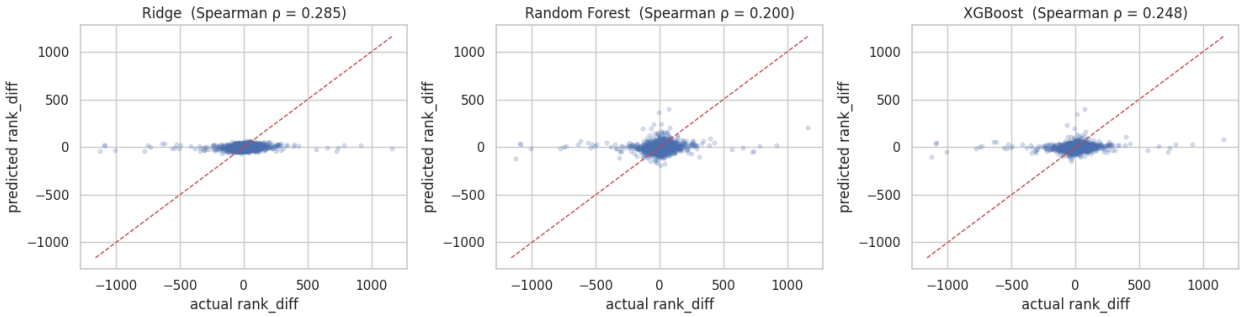

Figure 4. Predicted Rank vs Actual Rank, where the red line shows perfect accuracy and blue dots show the actual results of each player

Table 3. Accuracy of each model

Model	MSE	R^2	Spearman ρ
Ridge	12046.025202	0.022428	0.284919
Mean baseline	12330.090621	−0.000625	NaN
XGBoost	12341.970780	−0.001589	0.248120
Random Forest	12938.097986	−0.049967	0.200142

The outcome of this test shows almost no correlation between the ranking difference and stats, which is another interesting finding. After testing this, we can see that match statistics can predict the outcome of the match almost perfectly, as seen from Table 2, but cannot predict the difference in ranking or the skill gap between the two players, as seen in Table 3.

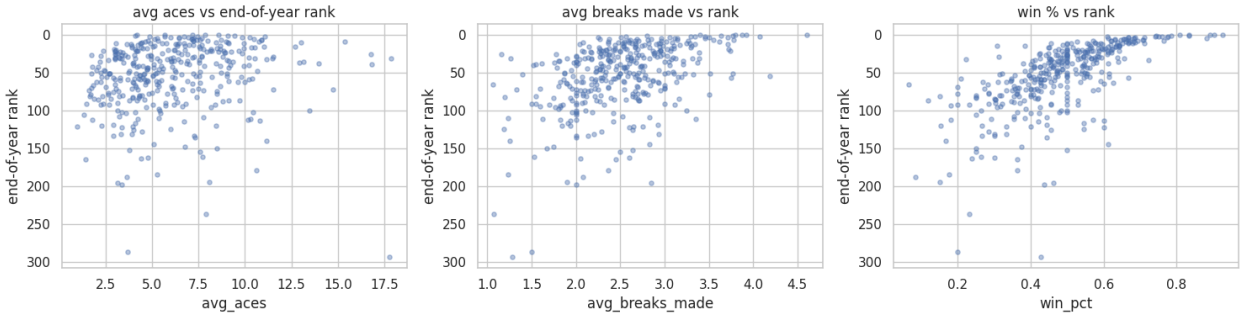

Figure 5. Ranking vs. Data

As you can see from Figure 5, there is some correlation found. For aces vs ranking, we can see that the majority of players are below 10 average aces, and the ones with more are around the higher rankings. However, aces are not very consistent, as a majority of top-level players are still in the back half of average aces. For the average number of breaks, we can see a better trend where the highest level players are in the upper half, while lower players are seen way at the bottom of the ranking. For break points, there are not many outliers because winning break points is directly correlated to winning a match. For win percentage, this is a predictable outcome where the better the win percentage, the higher the ranking. This one especially has

almost no outliers, and those who are outliers may have a greater chance of ranking in the future.

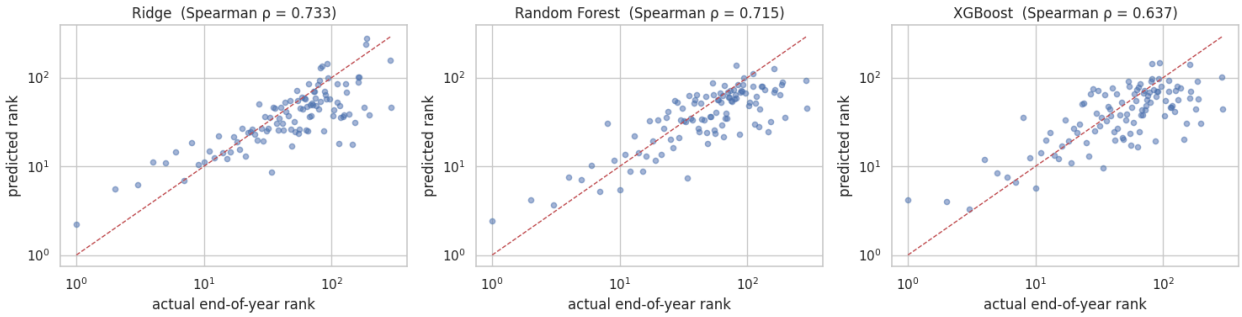

Figure 6. Predicted Rankings vs Actual Rankings

From Figure 6, we can see that Ridge was the best predictor, followed by Random Forest, then XGBoost, as seen by the increasing p-value, representing accuracy. An accuracy of 0.733 is high because it shows that there is a correlation between statistics and the end-of-year ranking.

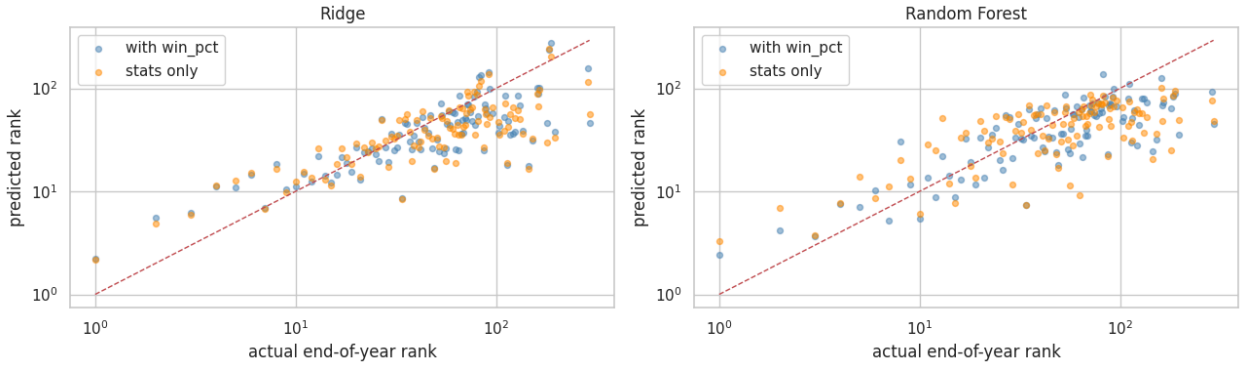

Figure 7. Predicted Rankings without Win Percentage

As we can see from Figure 7, removing the win percentage stat barely changed the data, showing a strong correlation between stats and ranking. Ridge still reaches a p of 0.710. The box-score statistics are what is carrying this model.

Model	R^2 (log)			Spearman rank		
	Full, lean	Subset, +W/UE	Subset, lean	Full, lean	Subset, +W/UE	Subset, lean
Random Forest	0.460	0.496	0.528	0.592	0.705	0.702
Ridge	0.547	0.634	0.640	0.710	0.768	0.773
XGBoost	0.412	0.373	0.403	0.567	0.657	0.650

Table 4. Accuracy with and without W/UE

The result was another surprise. Adding winners and unforced errors did almost nothing. On the charter subset, the Ridge model went from a Spearman of 0.773 to 0.768, and the R^2 dropped slightly from 0.640 to 0.634. The other two models moved by a similar amount in both directions. This is important because the winner/unforced-error ratio is one of the top three statistics in our original question and is constantly talked about in broadcasts as an indication of a top player. The feature that dominates every version of this model is the average number of breaks made, showing that breaking serve is where almost all season-level signals occur.

Match statistics can reproduce the ATP ranking fairly well, and serve breaks do almost all of the work, but reproducing the current ranking is not our goal. The reason for checking if our model can predict the rankings is to find if there is a correlation between ranking and match statistics. However, the whole motivation of the project is that the ATP ranking is slow, so we further our question: when our stats model disagrees with ATP about where a player belongs, who turns out to be right?

The first thing we checked was the blunt version of the question: does the model's predicted ranking line up with the future ranking better than ATP's own ranking? Here, the answer is no. In predicting the live-2026 ranking, ATP's 2024 ranking scores a Spearman of .622, while our model scores .562. The 52-week points total is a strong summary of a player's level, but that is not the whole story. When we controlled for the 2024 ranking and asked whether the model's prediction adds anything on top of it, it does. Adding the model's prediction to a simple regression of the 2024 rank raised the R^2 from .511 to .551, and the model term was statistically significant.

Across all 108 players in the test group, the model called the direction of the future move correctly 54% of the time, basically a coin flip. But when we narrowed it to the quarter of players where the model disagreed with the ATP the most, the hit rate jumped to 66% (19 of 29 correct). The named cases tell the story well. The model correctly moved players who were being held down by injury or slow points climb: Marin Cilic (179 to 48), Reilly Opelka (293 to 74), and Joao Fonseca (145 to 30). It also correctly marked players who were about to fade: Nicolas Jarry (35 to 165) and Sebastian Baez (27 to 65). It's one clean miss: Alexander Bublik, who the model demoted from 33, which Bublik climbed to around 10. Bublik is a high-variance server whose ceiling does not show in his season averages, so we added match-to-match variance to give the model a sense of a player's spread. It made little to no difference, however, so the blind spot remains.

Because the two carry nearly the same information, the blend did not beat the ATP alone: ATP scored about .62 against the future ranking, and so did every blend we tried. The extra signal we found earlier is concentrated in a handful of divergent players, not spread across the whole ranking list, so averaging the two rankings just dilutes it. This told us that a synthesis model is not going to produce a better ranking, and the stats are better used as a diagnostic flag.

We also tested the surface of tournaments and whether players who specialize in certain surfaces become more prominent. Splitting the rank-gap regression by surface gave essentially no improvement; every surface landed around R^2 of .02, with breaks making the top feature on all three. Clustering players by style of play produced three clean archetypes: a big server type (high aces, middle of the pack rank ~53), a returner type (high breaks made, best average rank ~38), and an all-court baseliner (low on everything, worst rank ~76). That the returner archetype

is the best-ranked group is another vote for the same conclusion that keeps appearing: returning and breaking serve is the skill that separates the top players.

For injury, as a description, they worked, but as a predictor, they failed. Adding them slightly lowered the model's accuracy, and the layoff feature is confounded with status: elite players can take time off while lower-ranked players grind every week, so a long layoff weakly associates with a better ranking, not a worse one.

At this point, we had tension. Every test above graded our rankings on how well they predict the future ATP ranking, and on that scale, nothing beats ATP. But that scale is circular; we are grading rankings against the ATP formula's own future output, which hands ATP a built-in advantage and punishes any ranking that gets ahead of the points exactly when it is right.

Split	ATP	Elo	Model	Blend (ATP)	Blend (Elo)
end-2023 → 2024 (n = 1320)	0.621	**0.655**	0.638	0.636	0.642
end-2024 → 2025 (n = 1681)	0.624	**0.640**	0.627	0.637	0.637

Table 5. Accuracy of Each Model

The results, in Table 5, finally show statistics beating ATP, as seen in the difference between ATP and the model. However, the best model we could create was a blend between ATP and statistics. Furthermore, nothing was able to flat out beat Elo ranking, which makes sense because that ranking is what we are basing the results on.

Discussion

The goal of this project was never to just build a model that predicts tennis matches. The goal was to answer a specific question: can the statistics from inside a match, match length, the winners-to-unforced-errors ratio, and service breaks say something about a player's true level that the ATP ranking misses? Working through the three framings gave us a clearer answer than we expected, and a couple of results surprised us.

The first thing we learned is that match statistics describe a match almost perfectly, but that this is much less impressive than it looks. Every classifier we tried landed above 90% accuracy at picking the match winner, against a rank-only baseline of 63%. At first glance, this looks like a high win for statistics over rankings. But the statistics we fed the model- which made more breaks and who saved more break points- are recorded during the match, so of course they line up with who won. Breaking your opponent's serve and holding your own is, almost by definition, how you win a tennis match. This part of the study is explanatory, not predictive.

The second finding was the first real surprise. When we asked the models to predict the ranking gap between the two players from their match statistics, the correlation almost completely disappeared: Ridge managed an R^2 of just .02 and the tree models did no better. So the same statistics that nail the winner of a match tell you almost nothing about how far apart

the two players sit in the standings. One thing to consider as well is the different archetypes of each player. Those with bigger serves and not much else may see more aces, but struggle against those who have better returns. In hindsight, this makes sense: a single match is just a snapshot, and a world No. 5 can beat a No. 55 by the same scoreline that a No. 50 used to beat a No. 55. The box score knows who won that day, but it does not know the season-long ordering behind it.

The third framing, aggregating a whole season of statistics per player and predicting their ranking, is the one that answers the actual question, and the statistics did recover the ranking well: Ridge reached a Spearman of .73. The important test came when we started stripping features out. Removing the win percentage barely moved the result. That was reassuring: it meant the box-score statistics, not just win-loss record, were doing the work. Then came the second real surprise. When we added the winners to the unforced errors ratio back in, something broadcasters talk about constantly as a mark of a great player, it changed almost nothing. The feature that dominated every version of the model was the average number of service breaks made.

At this point, we had shown that statistics can reproduce the current ranking, but reproducing a ranking that already exists is not useful. Our goal is to prove that ATP is not the best way of ranking players. Freezing the model's verdict at the end of 2024 and checking where those players stood over a year later gave a two-sided answer. The model does not predict future rankings better than the ATP's own rankings. The 52-week points total is a strong summary of a player's level. But when we controlled for the 2023 ranking and asked whether the model's prediction adds anything on top of it, it did: the fit improved from .51 to .55, and the statistics know something the points table does not.

The extra signal turned out to be concentrated exactly where we hoped. Across all 108 players in the test group, the model called the direction of a player's future move correctly only 54% of the time. But among the quarter of players the model disagreed with most strongly on the ATP, the hit rate jumped to 66%. The only big one was Alexander Bublik, who is a blind spot and worth stating plainly.

. Every test above graded our rankings on how well they predicted the future ATP ranking, pretty much putting ATP against ATP. So for the final experiment, we created an Elo system to see who actually won.

Graded this way, the ATP plus stats blend beats the ATP points. When we graded on real outcomes instead of rankings, the conclusion that statistics add nothing flips. Secondly, a plain Elo rating beats everything. Thirdly, once Elo is in the picture, adding statistics on top of it does nothing. The base accuracies here sit around 62-66% rather than the ~72% usually quoted for tennis.

The deepest takeaway, though, is not any single number. The same ATP and stats blend looked like a failure, but became a success when we changed what we based the accuracy on. The model never changed; only the yardstick did. That chosen evaluation metric is a real finding in its own right. If you benchmark a model against an existing ranking system using that system's own future output, you can certify an improvement as a failure. Graded honestly against what actually happens on court, our statistics improve on ATP points, which leaves the stats model's real, durable use as the diagnostic one we found in the disagreement analysis: an early warning that flags the injured comebacks, the breakouts, and the quiet declines months before the ranking admits them. Overall, the results met our broad expectation that rankings and rating systems already compress most of the predictable signal, exactly as the prior work

warned, but the two surprises (the uselessness of the winners/unforced ratio and the way the metric choice reversed the verdict) are the parts we did not see coming and the parts most worth carrying forward.

Limitations

Several things were out of scope for this project, and a few of them limited how far the conclusion could reach. The biggest is that we only ever evaluated on strong, closely ranked fields. The whole dataset is built from matches involving top-250 players, and the final match prediction test is player vs player in that same group. This was a deliberate choice because data is not well tracked for players below this ranking, but it means our numbers should not be read as tour-wide accuracy.

The second limitation is the winner/unforced errors data itself. Those two statistics only exist in the volunteer-charted subset, which covers roughly 30% of our matches, which is heavily biased towards higher players. So the findings that the winner/unforced ratio adds almost nothing are established on this slide of data, not the full set.

Third, the leading indicator analysis, the most interesting result, rests on a small sample. The test contains only 108 players, and while the effect is statistically significant, the significance is modest. The final experiment replicates across two seasons, but both are single-season test windows. We did not test longer or rolling seasons, so we cannot say how stable the effect is over time.

Finally, our attempt to capture injury and availability was limited to what we could derive from the match calendar. Part of the reason is that our “at least 10 matches” filter already removes the most badly injured players. So the injury gap is real, but we could not turn it into a predictive power from match data alone. Real injury information, such as actual withdrawals, medical timeouts, and physical condition, was out of scope because it would have required going through each match and checking whether or not each of these was recorded. If that gap could be closed with reliable availability data, it is the one place we think the model’s weakest results could be most improved.

Conclusion

The reason this project matters is not that we built a slightly better match predictor. It is what the exercise says about how we judge a player’s level in the first place. We set out to test whether the story inside a match could produce a sharper picture of a player than a 52-week tally of tournament points. Our methods were good enough to answer that hypothesis clearly: yes, match statistics carry real information about a player’s level, but they do not improve on a simple Elo-based rating, which already captures that information efficiently. The strong version of our hypothesis, that statistics could replace ranking, is not supported. A weaker but well-supported version is that they are an early-warning signal; the points system is too slow to register.

That distinction is where the value for the sport lies. The ATP ranking will and should remain as the official system, precisely because it is simple and impossible to dispute. But our results point to a complementary use for match statistics: a layer that flags players whose true level has drifted away from their points months before the ranking admits it. That is genuinely actionable for the people who have to make forward-looking decisions, such as wild card committees deciding who deserves a shot and players’own teams planning around rising or falling opponents.

Just as importantly, the project carries a warning to anyone doing this kind of work. We nearly concluded that statistics could not beat the ATP, and we would have been wrong: not because the model was weak, but because we had been grading it against the ATP itself. Only when we changed to grading against what actually happened on court did the real improvement appear. The lesson is simple and general: benchmark against reality, not against the system you are trying to beat.

Works Cited

"ATP Rankings FAQ." *ATP Tour*, 2026, www.atptour.com/en/rankings/rankings-faq.

Bozděch, M., et al. "A Detailed Analysis of Game Statistics of Professional Tennis Players: An Inferential and Machine Learning Approach." *PLOS ONE*, vol. 19, no. 11, 2024, e0309085, https://doi.org/10.1371/journal.pone.0309085.

Chen, T., and C. Guestrin. "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, 2016, pp. 785–94, https://doi.org/10.1145/2939672.2939785.

Cornman, A., et al. *Machine Learning for Professional Tennis Match Prediction and Betting*. CS229 Final Project Report, Stanford University, 2017, cs229.stanford.edu/proj2017/final-reports/5242116.pdf.

Elo, A. E. *The Rating of Chessplayers, Past and Present*. Arco Publishing, 1978.

Gao, Z., and A. Kowalczyk. "Random Forest Model Identifies Serve Strength as a Key Predictor of Tennis Match Outcome." *arXiv*, 2019, arXiv:1910.03203, arxiv.org/abs/1910.03203.

Glickman, M. E. "Parameter Estimation in Large Dynamic Paired Comparison Experiments." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 48, no. 3, 1999, pp. 377–94, https://doi.org/10.1111/1467-9876.00159.

Harris, C. R., et al. "Array Programming with NumPy." *Nature*, vol. 585, 2020, pp. 357–62, https://doi.org/10.1038/s41586-020-2649-2.

McKinney, W. "Data Structures for Statistical Computing in Python." *Proceedings of the 9th Python in Science Conference*, 2010, pp. 56–61, https://doi.org/10.25080/Majora-92bf1922-00a.

Pedregosa, F., et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825–30, www.jmlr.org/papers/v12/pedregosa11a.html.

Sackmann, J. *The Match Charting Project*. GitHub, 2026, github.com/JeffSackmann/tennis_MatchChartingProject. Dataset.

---. *tennis_atp: ATP Tennis Rankings, Results, and Stats*. GitHub, 2026, github.com/JeffSackmann/tennis_atp. Dataset.

Seabold, S., and J. Perktold. "Statsmodels: Econometric and Statistical Modeling with Python." *Proceedings of the 9th Python in Science Conference*, 2010, pp. 92–96, https://doi.org/10.25080/Majora-92bf1922-011.

Sipko, M., and W. Knottenbelt. *Machine Learning for the Prediction of Professional Tennis Matches*. Master's thesis, Imperial College London, 2015, www.doc.ic.ac.uk/teaching/distinguished-projects/2015/m.sipko.pdf.

"Tennis Abstract: Match Results, Statistics, and Analysis." *Tennis Abstract*, 2026, www.tennisabstract.com/.

Yue, J. C., et al. "A Study of Forecasting Tennis Matches via the Glicko Model." *PLOS ONE*, vol. 17, no. 4, 2022, e0266838, https://doi.org/10.1371/journal.pone.0266838.