Acing Predictions: Logistic Regression in the 2024 US Open Men's Tennis Championship

Krithin Visvesh

Abstract

Predicting match outcomes in tennis poses a significant challenge due to the sport's unpredictable nature and the influence of numerous factors on player performance. This study seeks to forecast the top 10 ranked athletes and their respective winning probabilities to win the 2024 US Open Men's Championship using Logistic Regression. The research analyzes data from US Opens from 2016 to 2023. The primary variables selected for the analysis are the winner's rank and the opponent's rank, applied in a logistic regression model using an 80/20 train-test split. The test accuracy was 68%. The probability of winning the US Open was also calculated for the top 10 ranked players, finding that the No.1 ranked player's probability of winning the US Open was 3.1%. As No.1 seeds have won 28.9% of the men's US Open singles tournaments, this suggests that using the player's and opponent's rank is insufficient to determine the probability of an individual winning the US Open.

Keywords: tennis; US Open; sports forecasting; logistic regression

Introduction

Predicting wins is an interesting aspect of sports that is very difficult to do, especially for tennis. Tennis has a rich history; in fact, the origins of the sport trace back all the way to the 12th century. It used to be an indoor sport, but the outdoor version gained more popularity only in the 19th century when Wimbledon, the oldest tennis tournament in existence, was introduced. Wimbledon was played on grass, but tennis offered other surfaces to play on like hard-court and clay which made the sport so diverse. Each game can produce an unexpected outcome and shock fans from all over the world. The four most prestigious tournaments in tennis are known as the Grand Slams: the US Open, French Open, Wimbledon, and Australian Open. Winning a Grand Slam finals match puts you in the extraordinary category of tennis athletes. It is a feat that very few can achieve as it requires utmost dedication and hard work. Specifically, the US Open is one of the Grand Slams that are more accessible for tennis fans to watch because of the location in Queens, New York. Yes, it is the last Slam in the calendar year, but it arguably contains the most excitement with top-ranked athletes fighting to secure another accolade. Winning a Grand Slam can be a significant milestone in an athlete's career and can add to their legacy in the sport. Tennis, and any sport in general, is a great opportunity to witness rising stars and their talent in action on the biggest stages. Many people love to predict matches before they occur. For instance, coaches can provide insights for the players, and bettors can use prediction models to make informed decisions. That being said, whether it is on betting sites or sports channels, there is a strong desire to figure out who will be the champion this year.

Literature Review

In the Journal of Quantitative Analytics in Sports, Stephanie A. Kovalchik wrote an article, "Searching for the GOAT of tennis win prediction." Kovalchik examined the performance of 11 different forecasting models for tennis matches and categorized it into 3 groups: regression-based, point-based, and paired comparison models. The study showed the GiveThirtyEight Elo model as particularly effective, reaching a 75% accuracy rate for matches involving top-ranked players. However, it showed a reduced accuracy (59-64%) for lower-ranked players [1]. In another study, Franc J.G.M. Klaassen and Jan R. Magnus both proposed a method to forecast the winner of a tennis match, not just at the beginning, but also during the match. They use a computer program called TENNISPROB and data from Wimbledon singles matches from the years of 1992-1995. TENNISPROB depends on two parameters, the probability pa that player A wins a point on service, and the probability pb that Player B wins a point with a serve [2]. The study doesn't have a quantifiable accuracy and therefore, is unreliable.

Betting markets are also a big beneficiary of statistical models. An analysis has been conducted on the US Open betting odds by VSiN emphasizing the importance of betting data with statistical models that improve model accuracy. The study concluded that betting odds reflect a combination of public sentiment and expert research, providing a respectable metric for prediction [3]. Also, an overview by Sportsbook Review compared various models to set up odds for the 2024 US Open. The comparison highlighted combining historical performance data with the current player's record. It also took into account factors like surface type and weather conditions to make the odds as accurate as possible. In fact, models incorporating these types of variables offer more accurate predictions, especially in the quarterfinals, semifinals, or finals of tournaments where players' energy start to deteriorate and mental factors become more significant [4]. Although multiple forecasting models exist for tennis, they use smaller datasets or the source data doesn't span over multiple years. Statistical models have not been tested for full efficiency, but they should be to ensure accuracy. However, none of the past studies have taken into account the player's and opponent's rank to determine probable match winners or found percentage probabilities for each athlete to win a given Grand Slam tournament. The purpose of the present paper is to come to the conclusion of the top 10 players most likely to win the 2024 US Open and their respective percentage probabilities by using Logistic Regression.

Methods

I'll be using the 'Huge Tennis Database' from Kaggle to analyze tennis data from 2016 to 2023 [5]. To focus on relevant matches, I filtered the data to include only those played at the US Open during these years. The main variables I'll be using are the winner's rank and opponent's rank to determine the probability of that player winning their matches. I chose these factors together because they resulted in higher model accuracy when predicting match outcomes compared to using factors like number of aces and number of break points won in a match. A Logistic Regression model is a statistical model that models the log-odds of a particular situation as a combination of one or more independent variables. In mathematical terms, it models the probability P(Y = 1) as a function of the independent variables X1, X2,...Xk. The model can be shown as an equation where P(Y=1|X) is the probability that Y equals 1 given the independent variables (X1, X2). I will be using an 80-20 split where 20% of the sample is reserved for testing. Afterwards, I checked the accuracy of the statistical model and it ended up as 68%.

Results

The probabilities of winning the US Open can be seen in Table 1. As seen here, a player ranked Number 1 in the world (in men's singles) prior to the start of the tournament has the highest chance of winning the US Open. They also have accumulated the highest number of Rank Points. The probability of winning the tournament also decreases as the rank increases.

Table 1: Prediction Results based on Player Ranks Prior to 2024 US Open

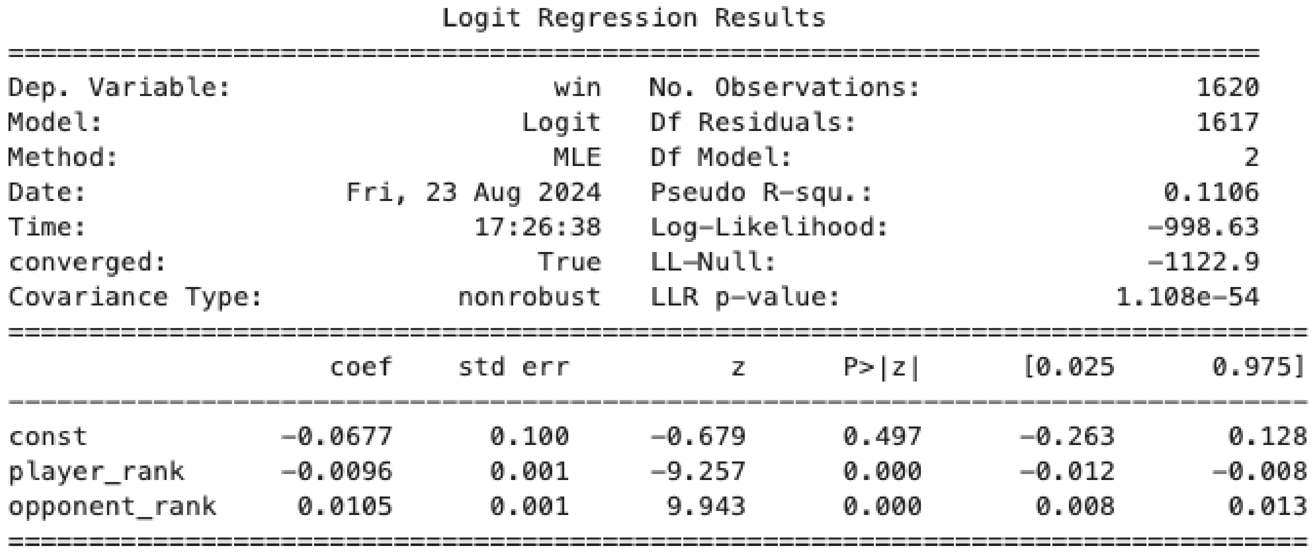
Player	Player Rank	Player Rank Points	Odds of Victory (%)
Jannik Sinner	1	9,360	3.07
Novak Djokovic	2	7,460	2.99
Carlos Alcaraz	3	7,360	2.91
Alexander Zverev	4	7,035	2.84
Daniil Medvedev	5	6,275	2.76
Andrey Rublev	6	4,805	2.69
Hubert Hurkacz	7	4,055	2.62
Casper Ruud	8	3,855	2.55
Grigor Dimitrov	9	3,655	2.48
Alex de Minaur	10	3,435	2.42

Table 2 shows the average opponent rank faced by the top 10 each round. These values were used in the logistic regression model to determine the opponent rank for each player in the top 10.

Table 2: Average Opponent Ranks by Round	
Round	Rank
1	132
2	89
3	59
4	50
5	28
6	17
7	5

The final accuracy of the Logistic Model for the test set was 68%, meaning that the model using only player ranks and opponent ranks was able to accurately predict the winner of a tennis match 68% of the time. Both player's rank and opponent's rank were statistically significant to the model.

Table 3: Logistic Model used to calculate the probabilities



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                    win   No. Observations:                 1620
Model:                          Logit   Df Residuals:                     1617
Method:                           MLE   Df Model:                            2
Date:                Fri, 23 Aug 2024   Pseudo R-squ.:                  0.1106
Time:                        17:26:38   Log-Likelihood:                -998.63
converged:                       True   LL-Null:                       -1122.9
Covariance Type:            nonrobust   LLR p-value:                 1.108e-54
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -0.0677      0.100     -0.679      0.497      -0.263       0.128
player_rank     -0.0096      0.001     -9.257      0.000      -0.012      -0.008
opponent_rank    0.0105      0.001      9.943      0.000       0.008       0.013
================================================================================
```

Discussion

Why are the probabilities so low? According to Bermejo and Ruano [6], a tennis athlete ending the year as the world #1 has approximately a 46% chance to win a Grand Slam and the rest of athletes in the top 10 normally have percentage probabilities between 10-30%. The model accuracy ended up being 68%, which is comparable to model accuracy shown by Kovalchick, which was around 59-64% [1]. I tried many things like increasing the dataset size to all hard court matches instead of just US Open matches. I also tried including more variables like the player's rank points, opponent's rank points, their respective ages, but all of them led to a lower model accuracy and lower correlation coefficient.

In short, it seems that player rank and opponent rank are not enough to calculate a top 10 ranked player's probability to win the US Open. One of the assumptions made for calculating the probabilities is that the result of each round was assumed to be independent of each other. This might not be a valid assumption because the results of a previous round may influence the next round, such as fatigue, facing an up and coming player who does not yet have a good rank, or others.

Another issue is player form. As the US Open is the last Grand Slam, it might make more sense to predict probabilities using the results of the current year's Grand Slam tournaments (the Australian Open, the French Open, and Wimbledon), as player form may be a factor in predicting the probability of winning the US Open.

Conclusion

In conclusion, a player's rank and the opponent's rank are not enough to determine a tennis athlete's probability to win a Grand Slam tournament. Other factors are most likely in play and it is difficult to pinpoint the exact factors due to the unpredictable nature of the sport.

So, how could I improve my study? I could include more factors, specifically that being UTR. UTR stands for Universal Tennis Rating, and it is a good way to quantify a tennis athlete's skills as it is said to be the world's most accurate tennis rating [7]. A way I could include UTR without too much of a hassle is to start with a new dataset to extract data. If that new dataset had UTR as an independent variable, it could help increase the model’s accuracy and increase the players’ percentage probabilities. There could be other factors that include win outcomes like head-to-head statistics, injury status, and recent performance in past Grand Slams. This improved model can also serve as a useful tool for coaches, analysts, and bettors in making strategic decisions.

References

[1] Kovalchik, S. A. (2016). Searching for the GOAT of tennis win prediction. *Journal of Quantitative Analysis in Sports, 12(3)*. https://doi.org/10.1515/jqas-2015-0059

[2] Klaassen, F. J., & Magnus, J. R. (2003). Forecasting the winner of a tennis match. *European Journal of Operational Research, 148(2),* 257–267. https://doi.org/10.1016/s0377-2217(02)00682-3

[3] Cohen, Z. (2024, July 22). 2024 US Open Betting Odds: Early analysis, predictions and players to watch. VSiN. https://vsin.com/tennis/2024-us-open-betting-odds-early-analysis-predictions-and-players-to-watch/

[4] Pearson, G. (2024, July 30). 2024 US Open Odds: Djokovic, Alcaraz, Swiatek, & Sabalenka favorites for Final Grand Slam. *Sportsbook Review*. https://www.sportsbookreview.com/picks/more-sports/tennis-us-open-odds/

[5] Servera, G. (2024). *Huge tennis database*. (2024, June 4). Kaggle. https://www.kaggle.com/datasets/guillemservera/tennis

[6] Bermejo, J. P., & Ruano, M. Á. G. (2016). *Entering tennis men's Grand Slams within the top-10 and its relationship with the fact of winning the tournament.* https://www.redalyc.org/journal/710/71047482006/html/#:~:text=In%20particular%2C%2045.65%25%20(.,1%20ranking%20in%20the%20world.

[7] UTR Sports. (2024). *Tennis & Pickleball ratings and events Platform | UTR Sports.* https://www.utrsports.net/#:~:text=The%20UTR%20Rating%20is%20the%20world's%20most%20accurate%20tennis%20rating,based%20on%20daily%20match%20results.