Predicting Success Probability in Professional Tennis Tournaments Using a Logistic Regression Model

Saurabh Srivastava

Abstract With a global audience of over 1 billion, professional tennis is the most widely followed individual sports in the world. The present study attempts to model the probability of success for a tennis player in a men's singles tournament of a given type (ATP 250, ATP 500, ATP Masters and Grand Slams) so as to enable his management team to take better decisions with respect to his calendar planning. The model in this study tries to arrive at the probability of success in a given category of the tournament by modelling the success of an athlete in that tournament (measured by his ability to reach the quarterfinals), using the logistic regression method. The scorecard that is built uses five variable categories to arrive at the probability of success, which can be used to rank order the tournaments in a given category for a player, and can be subsequently augmented through a linear programming method to help a player arrive at the most optimum selection of tournaments.

Keywords Sports analytics · Logistic regression · Tennis

1 Introduction

With a global audience of over 1 billion, professional tennis is the most widely followed individual sports in the world. The total prize money in 2016 for its biggest events, called the Grand Slams, ranged from 31.1 million USD (Australian Open) to 46.3 million USD (US Open), making it the most lucrative individual sports for athletes. Tennis is also one of the most geographically spread out sports in the world, with its venues present in more nations than any other sports except for football, and its top 100 ranked players coming from over 45 different countries.

Professional Tennis tournaments are usually organized based on gender and number of players. The most common configuration includes men's singles, women's singles, men's and women's doubles (where two players play on each side of the

S. Srivastava (✉)
EXL Analytics, Gurugram, India
e-mail: saurabh.dse@gmail.com

A. K. Laha (ed.), *Advances in Analytics and Applications*, Springer Proceedings
in Business and Economics, https://doi.org/10.1007/978-981-13-1208-3_6

net), and the mixed doubles (in which teams are formed of two players of opposite genders). The present study is based on the men's singles professional tournaments, which are managed by the Association of Tennis Professionals (ATP) and are divided into four categories (ATP 250, ATP 500, ATP Masters, and Grand Slams). Within each category, the prize money and the ranking points remain more or less comparable, but from one category to another the prize money and ranking points on offer differ significantly (the lowest being in ATP 250 and the highest being in the Grand Slams).

Therefore, winning the more advanced category tournaments is more rewarding. However, the probability of losing in an early stage is also high in the advanced category tournaments as they are usually intensely competitive. In case of an early exit, an athlete usually faces a triple blow—(a) time loss (player has to wait till the next week before he can participate in another tournament), (b) financial loss (due to cost of travel and fees incurred for participation), and (c) opportunity loss (points and prize money which he/she could have otherwise earned had they played a more suitable tournament where he could have advanced to later rounds).

This chapter begins by looking at the existing studies on the subject, reviewing some of the research on the similar issue of finding success/victory probability in tennis tournaments. It then goes on to discuss the model it proposes—the data sources, variables, and the appropriateness of the underlying modelling methodology. It finally reviews the results of the model and suggests ways on how this model can be further improved.

2 Literature Review

Although professional tennis is one of the most competitive and lucrative sport, limited research in the realm of sports analytics has been done on it. Three research works, however, are important. First is that of Klaassen and Magnus (2003), wherein the authors have attempted to predict the winning athlete during an ongoing tennis match, using a fast and flexible statistical model based on a fuzzy logic algorithm. However, this study, as acknowledged by the authors themselves, had a limitation that it would be able to predict victory in an ongoing match only and could not predict match victory beforehand. The second is the study by Boulier and Stekler (1999) who have tried to check, through Brier scores and using data related to two sports—US collegiate basketball and professional tennis—whether or not the seeding (ranking) of a player is a good predictor of his/her victory. Their study validated the hypothesis that rankings are good victory predictors for these sports, but did not talk about any other factor that might also be responsible for winning the match.

The third important study is again from Magnus and Klaassen (1999) in which they test the hypothesis that the person serving first in the match has a higher probability to win it. Their study, using a simple Bayesian error rate to study misclassification, found that this hypothesis could only be accepted for the first set of the match while in the subsequent sets, the player serving first had a higher probability of losing that set instead of winning it. This study again was based on one variable, which is also decided during a match on the basis of a toss, and thus could not be used for prior planning.

It was found that while the existing studies in the area are novel in their approaches, they use a limited number of variables and are not able to point to a comprehensive range of determinants of match victory. Moreover, the variables they use cannot be used for prior planning. The present study attempts to overcome these limitations.

3 Model, Data and Results

The present model is the first layer of a two-stage model, which attempts to provide an optimum solution to an athlete or his/her management team in choosing his events. The first stage (subject of the present paper) attempts to predict the probability of success of a tennis athlete in the event, while the second stage would be performing the final selection based on four factors—cost of travel, time taken to reach the tournament city, participation fees, and benefits from tournaments (prize money, ranking points, etc.). The second stage of the model would be based on a linear programming-based optimization algorithm.

The present study attempts to model the probability of success for a tennis player in a men's singles tournament so as to enable his management team to take better decisions with respect to his calendar planning. If a player chooses a tournament of a type higher than his calibre, he risks having an early exit from the tournament, leading to a significant loss of points and prize money. On the other hand, if he chooses a tournament of a type lower than his calibre, he may not be able to move ahead fast in the rankings and lose on both the prize money in the current tournament and more favourable draws in future ones. Therefore, an accurate knowledge of his chance of succeeding in a tournament of a particular category can be very helpful for an athlete in obtaining the best possible results. This study attempts to develop a model to aid decision-making precisely in this area.

In this study, we try to arrive at the probability of success by modelling the success of an athlete in a tournament (measured by his ability to reach the quarterfinals). This variable x takes a value of 1 when a player reaches the quarterfinals and 0 when he does not. The choice of quarterfinals as the cut off stage for determining a player's success in the tournament is arbitrary, though it is guided by the logic that quarterfinals usually arrive in the latter half of the tournament and are widely considered as a mark of accomplishment. The model uses publicly available ATP tournament data for training and validating the model, as detailed (Table 1).

Sanity checks to attest Quality and Integrity of Data on the following parameters: (1) Completeness of Information/Missing Value, (2) Outlier/Extreme Value Study, (3) Duplicate Record and (4) Distribution were done before proceeding with model development. Match-level statistics were pooled and adjusted to provide tournament-level information. Variables used only for indexing purposes were dropped, along with other match-related variables which clearly seemed to have no direct or indirect

Table 1 Data usage for training and validation

Purpose	Vintage	% of data used
Training	2015	70
Out of sample (OOS) validation	2015	30
Out of time (OOT) validation	2016	100

impact on match outcomes. Data, thus, obtained was used for model development using principles of logistic regression.

Event was defined as ‘Qualification to the Quarter Finals of a Tournament’. The hypothesis to be tested was whether factors such as Type of Tournament, Points, Physiological Characteristics (such as Age, Height and Weight), Player Rank, Surface of Play and Roof Characteristics (Indoor/Outdoor) impact the probability of an athlete in reaching the Quarter Final stage of a Tournament. Signs in Table 2 represent the hypothesized relation of the variable with the outcome (reaching Quarterfinals of the

Table 2 Variable description

#	Variable	Description	Category	Expected sign	Validated	Significant
1	Ranking	ATP ranking of the player	Numeric	Negative	Yes	Yes
2	Points	ATP points of the player	Numeric	Positive	Yes	Yes
3	Surface	Grass, clay, hard, or carpet	Categorical	NA-categorical		Yes
4	Tournament type	ATP250, ATP500, ATP masters, or Grand Slam	Categorical	NA-categorical		Yes
5	Roof characteristics	Indoor or outdoor	Categorical	NA-categorical		Yes
6	Age	Age of the player	Numeric	Negative	No	No
7	Height	Height of the player	Numeric	Positive	No	No
8	Weight	Weight of the player	Numeric	Negative	No	No
9	Draw size	No of opponents	Numeric	Negative	Yes	No

Tournament). Variables with p-value of less than 0.05 were considered significant in the model. Only the variables that were significant were allowed to stay in the model.

The modelling methodology is based on logistic regression. Logistic regression measures the statistical relationship between the dependent variable and one or more independent variables by estimating probabilities through a logistic function. Logistic regression is generally thought of as a more suitable method for developing models, where there is a binary response variable and the predicted values are probabilities and are restricted to (0, 1).

Assuming $p(\bar{x})$ as the probability of win, the model in this study constructs the following equation:

$$\log\left(\frac{p(\bar{x})}{1-p(\bar{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

Here, β_0 represents the intercept, β_p represent the coefficients of the various variables (x_p), and ε the error term. The probability of success in reaching quarterfinals can thereafter be obtained using the following equation:

$$p(\bar{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}$$

After training our model, the estimates for the various parameters are calculated to arrive at the final equation for $p(\bar{x})$ (please refer to Table 3). The values thus obtained for the various observations of the model are rank ordered based on their $p(\bar{x})$ values and grouped into deciles for evaluating the model parameters (Table 4).

The Variance Inflation Factor (VIF) represents the severity of multi-collinearity in a regression analysis. It is an index which measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient has increased because of collinearity. Cut Off of VIF in this study was kept at 10 to avoid multi-collinearity. However, all variables which entered the model (based on significance test) had a very desirable VIF value of less than 2 (Table 4).

The C-statistic (or the 'concordance' statistic or C-index) is a statistic that is used to measure the goodness of fit for binary outcomes in a logistic regression model. The model had a very desirable c statistic of 0.886, indicating good model strength (Table 5).

In order to validate the results, the model's lift was studied across the Training, Out of Sample (OOS), and Out of Time (OOT) datasets (refer to Table 6). The model not only performed well on these parameters for all three datasets, but also showed stability in lift results.

Table 3 Model estimates and significance

#	Variable	Description	Estimate	Prob Chi Sq	Wald Chi Sq
1		Intercept the expected mean value of Y, given X = 0	−3.5453	<0.0001	468.2977
2	ATP250	Indicator—when tournament type is ATP250	2.6564	<0.0001	437.1616
3	ATP500	Indicator—when tournament type is ATP500	1.5221	<0.0001	117.1053
4	Grand Slam	Indicator—when tournament type is Grand Slam	−0.4433	0.0205	5.3667
5	Points	ATP points of the athlete	0.0006	<0.0001	374.5926
6	Rank	ATP rank of the athlete	−0.0017	<0.0001	165.2119
7	Indi_Clay	Indicator—when playing surface is clay	0.4108	<0.0001	25.1634
8	Indi_Indoor	Indicator—when the match is played indoor	0.2224	0.0157	5.841

Table 4 Variable correlation, contribution, and variance inflation factor

#	Variable	Correlation	Variable contribution	Variance inflation factor (VIF)
1	Intercept	–	–	0
2	ATP250	0.14956	0.23799	1.87703
3	ATP500	0.03406	0.09896	1.48064
4	Grand Slam	−0.14901	0.03164	1.55394
5	Points	0.47848	0.22901	1.15711
6	Rank	−0.21658	0.35254	1.14118
7	Indi_Clay	0.0378	0.03447	1.12884
8	Indi_Indoor	0.09764	0.01539	1.22021

Table 5 Model performance statistics

#	Concordance	Value	Other stats	Performance
1	Percent concordant	88.5	Somers' D	0.772
2	Percent discordant	11.3	Gamma	0.774
3	Percent tied	0.2	Tau-a	0.222
4	Pairs	10,362,632	c	0.886

Table 6 Model lift

Sample	Lift at decile 1	Lift at decile 2
Train	0.448	0.666
Out of sample validation	0.449	0.659
Out of time validation	0.462	0.69

4 Conclusion

This research provides the most insightful variables deciding the success probability of a player in the match, and attempts to throw light on their relative importance. While the model confirmed the previous studies that players' competency indicators (such as Ranking and Points) continue to be the leading factors in deciding their performance, it showed that other factors such as playing condition and tournament category play an important role as well. Physiological and Demographic characteristics of athletes, such as Age, Height, and Weight were found to be statistically less significant in determining their ability to reach quarterfinals of the event, while player performance attributes (rank and points) were found to be most significant, followed by tournament characteristics (indoor or outdoor, surface, and tournament category). In order to develop a robust calendar management tool for an athlete, the current model can be complemented by other factors such as cost of travel, time taken to reach the tournament city, participation fees, and benefits from tournaments (prize money, ranking points, etc.).

References

Boulier, B. L., & Stekler, H. O. (1999). Are sports seedings good predictors? An evaluation. *International Journal of Forecasting, 15,* 83–91.

Hosmer, D. W., Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). Wiley.

Klaassen, F., & Magnus, J. (2003). Forecasting the winner of a tennis match. *European Journal of Operational Research, 148,* 257–267.

Magnus, J. R., Klaassen, F. J. G. M. (1999). On the advantage of serving first in a tennis set: Four years at Wimbledon. *Journal of the Royal Statistical Society, Series D (The Statistician), 48,* 247–256.