Tennis Player Classification

Johnattan Ontiveros

May 2021

Abstract

This study investigates the effect of tennis playstyle on match outcomes. Using a combination of dimensionality reduction techniques and K-means clustering, the top 100 ranked ATP players of 2017 were classified from Grandlsam point-by-point tournament data. With this labeled data, a logistic model was trained upon years of match outcomes from the ATP Masters series. A baseline logistic model was fit upon this data using the Elo rating difference between players as a predictor to control for player skill level. This baseline model was then compared to a logistic model that included Elo and classified playstyles via a likelihood ratio test.

The model including playstyle significantly improved the deviance and showed significant predicted advantages of some playstyles over others. While this was interpreted as evidence that playstyle has a significant role in match outcomes, the model's interaction coefficients produce some counterintuitive results. This error, however, likely stems from the unevenness in samples from certain classification interactions.

Background

A game of singles tennis is one between two players who alternate serving and sides of the court. A point in tennis begins with one player serving the ball, which is usually a high-speed overhead shot. If the serve lands in the service-portion of the court, the "rally", or back-and-forth hits between the players, begins. A successful serve that the returner fails to make contact with is called an 'Ace', which is usually an indicator of a well placed serve. Once in a rally, the point can end in one of two ways (in regards to the data used): a player returns a ball so well that the opposing player cannot reasonably return the ball and fails to do so (categorized as a winner) or a player makes an error such as hitting a ball out of bounds or into the net (categorized as a unforced error). Because of the discrete point-by-point nature of the sport, this study uses point-by-point data to classify player playstyles.

There are a number of distinct playstyles in professional tennis, and one longstanding question is whether a player's playstyle makes a significant difference on the match outcome, and more specifically, which playstyles work best against others? A classic example of playstyle-differences arising in tennis is the Roger Federer v. Rafael Nadal rivalry. Despite the two players being relatively equal in skill at the top of rankings, Nadal has a win tally against Federer of 24 to 16. Although no definitive cause of this has been determined, the popular belief is that Nadal's playstyle simply works well against Federer's playstyle. This example of how playstyles interact will serve as the "sanity check" throughout the study. With a given set of tennis playstlyes, this study seeks to determine if playstyle is a significant factor in match outcomes, and which playstyles have an advantage over other playstyles.

Data

The data used for playstyle classification is sourced from the four GrandSlam tournaments from the years 2011 to 2017. For each of the top 100 ranked players in the 2017 ATP season, the average ***Aces, Serve Speed, Net Points, Net Wins, Unforced Errors*** were calculated. Furthermore, a supplementary dataset detailing the player's handedness was used as well. Below are the compiled metadata for the top four players of the 2017 season. *Note that "Net" in this case refers to the points taken at the net, not the total.*

Player	Aces	ServSpeed	Net Pts	Net Wins	UnfErr
Roger Federer	0.098	104mph	0.0051	0.73	0.0044
Novak Djokovic	0.072	103mph	0.0039	0.73	0.0048
Andy Murray	0.077	98mph	0.0035	0.76	0.0045
Rafael Nadal	0.051	101mph	0.0023	0.82	0.0031

Player	Handedness	Backhand
Roger Federer	R	One-Handed Backhand
Novak Djokovic	R	Two-Handed Backhand
Andy Murray	R	Two-Handed Backhand
Rafael Nadal	L	Two-Handed Backhand

Table 1: Sample of point-by-point data

The training data used in win probability modelling was sourced from ATP Masters Tournament datasets. A supplementary dataset listing the Elo rankings of players during the 2017 season was used as a predictor. In tennis the Elo rating of a player is a tool used to compare relative skill levels between players, the higher the Elo the higher the skill level of the player. For reference, in the 2017 season Roger Federer had a peak Elo rating of 2444 while the average Elo rating in the Mens top 100 was 1968. Therefore the data used consisted of:

- The Elo difference betwen the higher Elo player and lower Elo player

- The classification of the higher Elo player
- The classification of the lower Elo player
- The outcome: 1 if higher Elo player wins, 0 if lower Elo player wins

Methodology

The experiment consists of two parts: playstyle classification and win probability modelling. The classification portion of the experiment is necessary both because a labeled dataset of players is not publicly available, and because there is no concrete number or criteria for tennis playstyles. The classification of playstyles serves as a necessary step towards win probability modelling, where the effectiveness of adding player classifications to a model is evaluated.

Player Classification

As mentioned, the set of players used for this study are the top 100 ranked mens players of 2017. The compiled metadata for each player is used to cluster the players into playstyle types. In order to do this, the numerical and categorical data must first be consolidated via Factor Analysis of Mixed Data (FAMD) and then clustered using K-means clustering. The reason for this approach is both for dimensionality reduction and for creating appropriate standardized numerical data for the K-means clustering approach to work well.

FAMD is a form of dimensionality reduction that can be thought of as a mixture of Principle Component Analysis (PCA) for the numerical data and multiple correspondence analysis (MCA) [1]. The FAMD algorithm essentially reduces the dimensionality of the mixed data by optimizing the explained variance for both sets of data once appending the categorical data to the numerical data after conversion to euclidian space [1]. Similar to PCA and MCA, the outcome is a set of orthogonal vectors representative of the data up to a certain explained variance.

In order to choose the number of vectors to include for clustering, a minimum explained variance criterion is set from the dimensionality reduction technique. From informal experimentation, a minimum explained variance of 80 percent was set to decide the minimum number of produced dimensions needed. With this reduced data, the next step in classification is the actual clustering of the data.

The K-means algorithm is a non-parametric approach to classification [2]. The variant used in this experiment proceeds by randomly initializing k centroids in the data space and iteratively reduces the in-cluster variance by shifting the location of the centroids until convergence [2]. The in-cluster variance, sum of squared error, or 'inertia', that is minimized is defined as:

$$SSE = \sum_{j=1}^{P} \sum_{k=1}^{K} \sum_{i \in C_k} (x_{ij} - \bar{x}_j^{(k)})^2$$

Where P is the dimensions of the data, K is each cluster, and C_k represents the membership of each point in a cluster [2].

The random initialization of centroids can lead to sub-optimal results, therefore the best of 10 random initializations is chosen as the final model. In order to determine which number of clusters (in this context meaning which number of tennis playstyes) to use, a plot of the inertias against the k number of clusters will be made, known as an elbow plot, in order to qualitatively choose the value of k that minimizes inertia without overfitting. The values of k tested are from 1 to 20, and the optimal K-means model of k clusters will be used as the playstyle classifier for the second portion of the experiment.

Win Probability Modelling

The goal of this project is to determine whether some tennis playstyles perform better against other tennis playstyles. The first question of whether playstyles play an important role in match outcomes is determined by first fitting a baseline predictive model that does not include player classifications as a predictor. This model serves as a reference and is verified for goodness of fit to continue the model fitting process. Next, the classifications of players is added to the baseline model and tested for significant improvement, which serves as an indicator for playstyle having a significant role in match outcome. The second part of the question: which playstyles perform better against others, is determined by statistical inference on the model including playstyle.

The baseline model is a generalized Bernoulli linear model with a logit link function (logistic regression). The model uses the absolute Elo difference between players as a predictor, making the win probability of a higher ranked player in matchup i:

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 * elo\_diff$$

The absolute Elo difference is included as a control for differences in player abilities, making it a good comparison to any model that adds predictors.

The logistic model including playstyle classification does so by adding the interaction between the lower ranked player's playstyle and the higher ranked player's playstyle. This makes the probability of a higher ranked play in matchup i as:

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_1 elo\_diff + \gamma_1 x_{11,i} x_{21,i} + \gamma_2 x_{12,i} x_{21,i} + ... + \gamma_J x_{1K,i} x_{2K,i}$$

Where $x_{1k,i}$ and $x_{2k,i}$ represent indicators for one of the total K playstyle classifications of the players. The exclusion of the non-interacted terms and intercept is done with the intention of simplifying inference and analysis of the model. This setup allows for the direct inference of interactions between playstyles.

The comparison of the baseline model to the playstyle model will then be done through a likelihood ratio chi-squared test. A significant reduction in the deviance by the added degrees of freedom from the playstyle interactions on the $\alpha = 0.05$ level will be interpreted as playstyle playing a significant role in match outcomes, and the coefficient estimates of the playstyle model will be analyzed.

Analysis of Results

Player Classification

With the minimum of 80 percent criterion, the FAMD dimensionality reduction method left us with four vectors of data. Below is a sample of the reduced dataset used for clustering.

Player	0	1	2	3
Roger Federer	3.47	0.74	1.56	-0.37
Novak Djokovic	2.11	-1.19	1.59	-1.01
Andy Murray	1.90	-1.12	1.37	-1.20
Rafael Nadal	-0.19	0.19	2.53	-2.00

Table 2: Sample of Reduced Data

Multiple k-means with k values ranging from $1-20$ were tested. Fig.1 is the elbow plot of the number of clusters to the inertia of the fitted model.

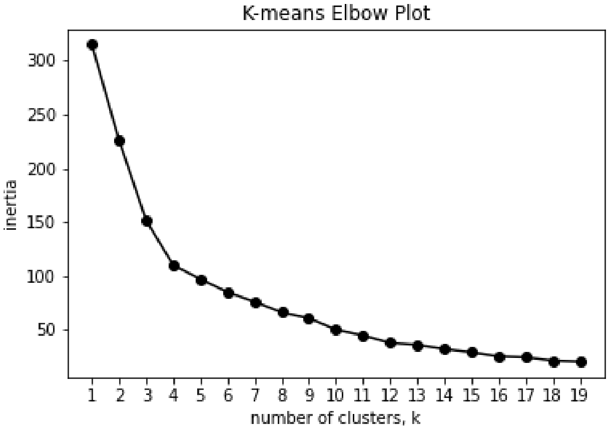

Figure 1: K-means elbow plot

From the plot it became clear that four clusters/playstyles should be the value of k chosen because it was where the steep drops in inertia end, making it a good trade off point between the bias and variance of the classifications. Having chosen the number of clusters, the players in the ATP top 100 were assigned a playstyle classification. Table 3 is a sample of the classification results, the generated classes range from 0-3. From a qualitative look at the classification results, the descriptions of the playstyles are:

- 0: Characterized by consistency, low unforced error rates and left-handedness, no particular backhand.
- 1: Distinctly right handed, two handed backhands, low net point success rate but consistent points overall.
- 2: Distinctly One-Handed backhands, second highest serve speed playstyle.
- 3: Distinctly fast serves and high average net approaches, overall an aggressive playstyle.

Player	Classification
Roger Federer	3
Novak Djokovic	3
Andy Murray	3
Rafael Nadal	0

Table 3: Sample of Classified Playstyles

As our sanity check, Federer and Nadal being in different playstyle groups was a reassuring result.

Win Probability Modelling

Starting with the baseline model, below are its coefficient estimates:

Pred	Estimate	$\Pr(> \|z\|)$
intercept	-0.035	0.311
elo_diff	0.0046	<2e-16

Table 4: Baseline Model Coefficient Estimates

From the Elo estimate of 0.0046, we can see that an increase of 100 Elo point difference corresponds to a multiplicative increase of the odds of the higher Elo player winning by 1.57. This makes intuitive sense, a greater difference in Elo between players corresponds to the higher Elo player having a higher probability of winning the matchup.

Multiple diagnostics were checked on the baseline model, the first being the ”calibration” of the model, which is done by plotting the predicted outcomes against the actual outcomes.

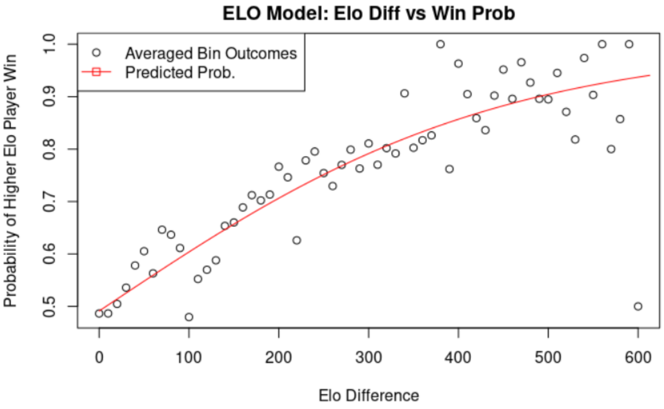

Figure 2: Calibration Plot

Fig.2 verified that the model was a good enough fit to the data, having a relatively even spread and little to no concerning outliers. Next, checking for overdispersion, the residual deviance (11160) to residual degrees of freedom (9196) ratio was 1.21, indicating that the model did not have a concerning level of overdispersion. And as a last check for any overinfluential points, the Cook's distances were checked, with every point being well below 0.01. Having been determined to be a good fit on the data, the classifications of the players were then added to the baseline model for comparison.

With four playstyles, the baseline model added an additional 15 variables through the interaction between lower ranked classes and higher ranked classes. Performing the Likelihood Ratio Test, the playstyle model reduced its deviance by 62.72 in comparison to the baseline model with an additional 15 degrees of freedom. This was a very significant reduction, with a p-value well below $\alpha = 0.05$. This outcome was taken as evidence that the added information of player playstyles does improve the model and therefore likely plays a role in match outcomes. Table 5 contains the estimated coefficients of the model.

Looking at the coefficients that are significant, we have:

- Lower Elo Playstyle 3 has a **disadvantage** over Higher Elo Playstyle 0.
- Lower Elo Playstyle 0 has a **disadvantage** over Higher Elo Playstyle 1.
- Lower Elo Playstyle 2 has an **advantage** over Higher Elo Playstyle 1.
- Lower Elo Playstyle 3 has an **advantage** over Higher Elo Playstyle 2.
- Lower Elo Playstyle 1 has a **disadvantage** over Higher Elo Playstyle 3.

Something concerning with these results, however, is that none of the significant interactions are also significant in their reverse case. For example, although the interaction between a lower ranked playstyle 1 and a higher ranked playstyle 3 very significantly favors playstyle category 3, the reverse case (lower ranked playstyle 3 interacted with higher ranked playstyle 3) is not significant, with a

Predictor	Estimate	pval
elo_diff	0.0040055	< 2e-16 ***
lower.elo.class0:higher.elo.class0	0.1011451	0.534020
lower.elo.class1:higher.elo.class0	-0.0152538	0.867427
lower.elo.class2:higher.elo.class0	-0.1290263	0.296972
lower.elo.class3:higher.elo.class0	0.6741977	0.008905 **
lower.elo.class0:higher.elo.class1	0.1942881	0.034752 *
lower.elo.class1:higher.elo.class1	-0.0810320	0.124765
lower.elo.class2:higher.elo.class1	-0.1728118	0.022127 *
lower.elo.class3:higher.elo.class1	-0.2686146	0.125559
lower.elo.class0:higher.elo.class2	0.1039646	0.439634
lower.elo.class1:higher.elo.class2	0.0284064	0.728964
lower.elo.class2:higher.elo.class2	-0.0438833	0.710786
lower.elo.class3:higher.elo.class2	-0.9156011	0.000489 ***
lower.elo.class0:higher.elo.class3	0.0535382	0.601701
lower.elo.class1:higher.elo.class3	0.3421606	2.01e-05 ***
lower.elo.class2:higher.elo.class3	0.1460858	0.163838
lower.elo.class3:higher.elo.class3	0.1342229	0.165896

Table 5: Playstyle Model Coefficients

p-value of 0.12. It is important to note, however, that the signs of the coefficients still match up, but that is not always the case.

Taking the Federer and Nadal case as an example again, we note that Federer was classified as playstyle 3, and Nadal playstyle 0. In 2017, where the data ends, Federer was at a peak Elo ranking of 2444, 94 Elo points above Nadal with a ranking of 2350. The two tables below examine the differences in prediction from the baseline Elo model and the Classification model.

Baseline (Just Elo) Model

Higher Elo Player	Lower Elo Player	Elo Diff	Prob. High Elo Wins
Federer	Nadal	94	0.60
Federer	Nadal	0	0.49
Nadal	Federer	94	0.60
Nadal	Federer	0	0.49

The baseline model predicts matchups as expected, regardless of player playstyle the higher Elo ranked player has a 0.6 probability of winning a matchup when ahead by 94 Elo points, and about 0.5 when there is no difference in Elo rankings.

The playstyle model, however, shows some important differences in predicted outcomes. The interaction between Lower Elo Playstyle 3 and Higher Elo Playstyle 0 was significant, while the reverse case was not. The first two rows comparing Federer (3) as the higher Elo player against Nadal (0) corre-

Playstyle (Elo + Classification) Model

Higher Elo Player	Lower Elo Player	Elo Diff	Prob. High Elo Wins
Federer	Nadal	94	0.61
Federer	Nadal	0	0.51
Nadal	Federer	94	0.74
Nadal	Federer	0	0.66

Table 6: Model Prediction Comparison

spond to the insignificant interaction. Accordingly, we see that the playstyles do not make much of an impact on prediction, counterintuitive to tennis knowledge where an evenly matched Rafael Nadal and Roger Federer leans towards Nadal winning.

The results from the last two rows correspond more to the tennis expectation. Nadal's playstyle has a much higher probability of winning a matchup over Federer's playstyle, regardless of the differences in Elo ranking.

Conclusion and Improvements

While the significance of the likelihood ratio test with the addition of playstyle classification shows promise of playstyle analysis in predicted outcomes, the coefficients of the interactions themselves are questionable.

The problems identified from the Federer-Nadal predictions comparison could come from a variety of factors, most of which could be remedied in a further study. The model is set up to where there are different amounts of data for each interaction, which could be amplified by the second possible flaw in this model: data leakage. When clustering playstyles the predictors used were carefully chosen not to be related to actual player skill level for the reason of clustering on playstyle not skill level. Having a significant classifier that correlates playstyle and skill level would not only make the significance of the addition of playstyle in a predictive model misleading, but would also make the disparity of amounts of data for each interaction worse.

In defense of this project, however, the results of this approach are mostly intuitive, there are four generally recognized playstlyes in tennis, and playstyle was expected to be a significant factor, which this study has provided compelling evidence in agreement.

Data Sources

I would not have been able to do this project without the selfless individuals who collect and maintain these open data sources:

- Jeff Sackman at: https://github.com/JeffSackmann/tennis_slam_pointbypoint.git
- The maintainers of Ultimate Tennis Statistics: https://www.ultimatetennisstatistics.com/

References

[1] Gilbert Saporta. Simultaneous analysis of qualitative and quantitative data. *ocieta Italiana diStatistica*, 1990.

[2] Douglas Steinley. K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 2006.