

Using Multiple Regression to Predict the Outcome of Tennis Matches

By Samuel Alkire, Ryan Sun, and Mithun Rameshkumar

Link to Colab:

<https://colab.research.google.com/drive/1kuJ4kC5wF-sahZqhS7T7ainSkECeN2Eu?usp=sharing>

A substantial amount of research has been conducted on using analytics to predict the outcomes of tennis matches. For ATP matches, statistics like first serve percentage, double faults, aces, and break points saved are tracked. Since a tennis match includes many measurable statistics, this lends itself well to studying the data in a regression framework. One foundational paper in the space, “Forecasting the Winner of a Tennis Match” by Klaasen and Magnus, was one of the first attempts at predicting tennis matches using regression analysis. The paper achieved significantly improved forecasting accuracy by using in-match statistics to create regression models for Elo ratings, logistic regression, and point-by-point probability estimates (Klaasen & Magnus, 2002).

For our project, we examined the paper “A framework for applying the Logistic Regression model to obtain predictive analytics for tennis matches” written by Georgios Friligkos et al (Friglikos & Papaioannou, 2023). This paper applies Logistic Regression in order to predict the outcome of tennis matches, particularly finding a winning probability for each of the competing players. To maximise the accuracy of predictions, the researchers begin by testing their model under many different configurations to identify which independent variables are most effective in predicting the outcome. At a high level, this connects directly to what we learned in class with our multiple predictor regression models and ties it into a real-world application. To gather data for their project, the researchers used a dataset which stretched decades of tennis matches, capturing all aspects of every ATP match played, with factors ranging from serve percentage to the number of double faults on a specific court type. From here, the researchers use logistic regression to best predict the outcome of these tennis matches and identify what key features need to be used.

While we may not have discussed logistic regression directly in class, the literature provides a great example that builds off of what we learned. In short, logistic regression is a method that is used for classifying binary tasks (like who will win/lose a tennis match). The paper begins with the model we saw in class for our regression with multiple predictors, modelling their value:

$$z = w_0 + w_1 \cdot x_1 + \dots + w_n \cdot x_n, \text{ where the } w_i \text{ are the } \beta_i \text{ from class}$$

However, unlike the regression we have seen before, we need to convert the z-value into a value between 0 and 1 to get our probability of winning the match. This is done through a Sigmoid function, which takes our z and maps it to a value in the range 0 to 1:

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

Now that the researchers have set up their regression model, the next key step is determining the best set of w coefficients to predict the outcome most accurately. This is done by attempting to minimize the error between our predicted probability from the sigmoid function (which is between 0 and 1) with the binary outcome of the match (0 for losing or 1 for winning). This error is measured using cross-entropy loss, which takes the difference between y and the predicted probabilities $\hat{y}$ as follows:

$$L(y, \hat{y}) = - \sum_{i=1}^N y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)$$

Using this formula in addition to an algorithm similar to gradient descent (which is out of the scope of this class), the researchers identified the optimal parameters in order to predict the outcomes with maximum accuracy.

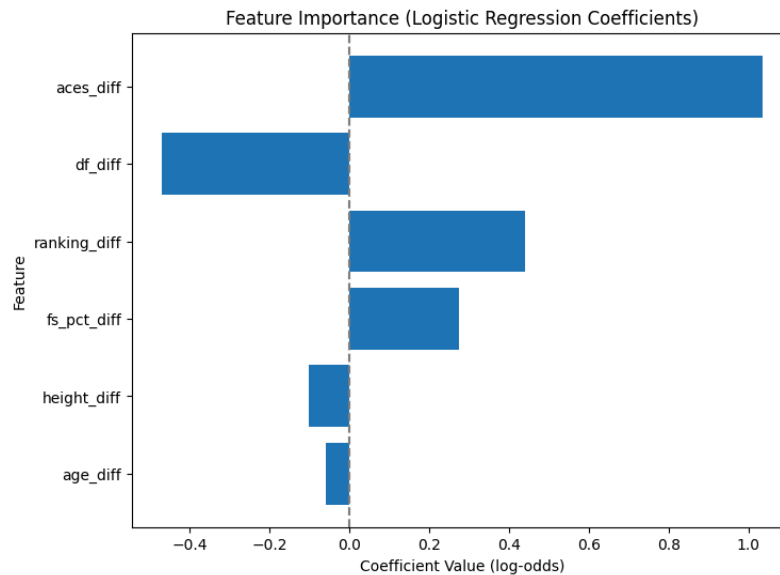
Following this analysis, the paper provides detailed results of the researchers' findings. Ultimately, they chose 17 features to be used in testing their Logistic Regression model, the majority of which are focused on serving and return accuracy, based on a set of 630 tennis matches in 2023. They were able to then achieve a maximum accuracy of ~72% using this. In terms of applying this model, the researchers ended up creating a site called *Tennis Crystal Ball*, where users could select an upcoming tennis match and the site would pull data on the players to make a live prediction. This site demonstrates how useful statistics can be within sports and the real world applications of this class.

For our project, we began by selecting the 2024 ATP dataset. Although we initially considered using data spanning multiple years, we realised that players' performances can vary greatly by season and that the literature we studied similarly only used one year's worth of data. This dataset contains every men's ATP match result from the year, with stats recorded from each match. Once we gathered this data, we split the first 80% of matches in 2024 into training data and the remaining matches into testing data. We then used the training data set to create our multi predictor regression model. The predictors we were interested in using were first serve percentage, ace percentage, double fault percentage, ranking, height, and age. Then, we measured the difference between the winner and the loser for every feature and used this to train our logistic regression model.

Then, to test the model, we created a function that uses the average data from players to predict the outcome of matches. For each player in the testing set, we take the average of their historical matches for each feature (other than ranking, height, and age). Then, to predict the outcome of a match, we calculate a new difference using these historical averages and feed this into the trained model.

We used a number of metrics to measure the effectiveness of our logistic regression model. Looking at the classification report, we find that the average of the precision, recall, and f-1 scores is 0.7. This indicates that we are able to predict match outcomes with about 70% accuracy. It also shows that the model is not biased toward favoring favored players or underdogs and is balanced in correctly predicting winners and losers.

Looking at the coefficient chart, we see that the most significant predictors are the differences in ace percentage and double fault percentage. We were quite surprised to see that these predictors had a higher importance than the ranking difference between the two players. We were also surprised by height having a negative weight, meaning that shorter players had a greater chance of winning the match when controlling for serve strength. The feature with the least significant coefficient was the age difference, which makes sense given that older players like Djokovic are still able to dominate the game.



One limitation of our model, as described later, is failing to account for interaction effects between predictors. To account for this, we created a correlation heatmap of the predictors to get a sense for which predictors were most closely associated. Two features that were closely related were the ace percentage and height. This makes sense, as a taller player can strike the ball from a higher vertical position, resulting in a much steeper downward angle. In addition, players with strong first-serve consistency generally commit fewer double faults.

A second limitation that is also not addressed in the original paper is whether the results of recent matches should hold greater weight than those in the past. Tennis is an especially dynamic sport and a player's performance can drastically change over the course of a year. For instance, in 2024, Jannik Sinner won his first two Grand Slam titles and finished with a 73-6 record, whereas his results in the prior year were significantly worse. Averaging his performance uniformly across years would dilute his current form and underestimate his true strength. A solution to this would be introducing time-decay weights, which place more importance on matches that have occurred recently. This would meaningfully improve predictive accuracy and track dynamic changes in player strength.

Finally, our model does not account for the effect of the playing surface. ATP matches occur on hard courts, clay, and grass, each of which have fundamentally different properties. For example, serve-dominant players can generate far more aces on faster surfaces like grass, while shorter players have an advantage on clay, which favors baseline rallies. By fitting a single model to all surfaces, we assume that the predictor effects behave identically across all environments. Implementing adjustments for each surface would likely improve our predictive accuracy by aligning the model more closely with the structure of professional tennis.

Works Cited

- Friligkos, Georgios, and Evi Papaioannou. *A Framework for Applying the Logistic Regression Model To*, Oct. 2023,
www.researchgate.net/publication/374732544_A_framework_for_applying_the_Logistic_Regression_model_to_obtain_predictive_analytics_for_tennis_matches.
- Klaassen, F. J. G. M., and J. R. Magnus. *Forecasting the Winner of a Tennis Match*. *European Journal of Operational Research*, vol. 148, no. 2, 2003, pp. 257-267.
<https://www.hse.ru/data/695/551/1234/Magnus.pdf>.