Predictors for Winning in Men's Professional Tennis

Abstract

In this project, we use logistic regression, combined with AIC and BIC criteria, to find an optimal model in R for predicting the outcome of a professional tennis match. The data for this analysis comes from the Men's 2014 Grand Slam tournaments, which are combined into one data set, randomized, and broken into training and validation sets. The optimal model, using information from both players, under the BIC criterion using variables concerning both players accurately predicts 93.2% of the cases in the training set and 93.4% of the cases in our validation set using a 50% probability as a cutoff for win or loss for player one.

Introduction

The Grand Slam tournaments consist of the four most important tennis events for any tennis player. Even people who are not avid tennis fans may still recognize recurring competitors such as Nadal, Djokavic, or Sharapova. The top players across the world come to play for ranking, notoriety, fame, and prize money. Each of the four tournaments, the Australian Open, French Open, US Open, and Wimbledon, bring hundreds of thousands of fans eager to watch the best tennis players fight to be on top. Because the competition in this tournament is often fierce, and players' performances are often based on many factors, we would like to know which specific variables are most important in determining the outcome of a match.

This research question is incredibly relevant in the world of sports. Coaches, players, and fans would want to know the factors most important to winning matches. Thousands of dollars are bet on matches by zealous fans. Creating a model to determine likely predictors of success could allow players to target their training to focus on the aspects deemed most critical to success. In terms of the both players model this would give an overall, more general prediction depending on the performance of both players, while a single player model would deal only with an individual player, which the player has more control over.

After some attempts at creating predictive logistic regression models for the outcome of a match, it became clear that the data in the Australian Open would not be enough for fitting a model. Due to the large number of predictors and fairly small number of observations (126), it was possible to fit a perfect model. As a result, although the model could predict all results in the Australian Open correctly, it could not be used for any other data set. Thus, it became necessary to combine data sets in order to train the model with more observations. We combine the Men's Australian Open, French Open, US Open, and Wimbledon into one data set with 491 observations initially. The data was randomized and 365 observations were used to train a model while the remaining 126 for validation.

It is possible that some tennis matches in the different tournaments involve the same two players. However, the conditions for playing in each tournament are not identical. For instance, changes in weather, surface played on, preparation taken by the players for that particular tournament, and additional practice in between tournaments all mean that two players facing each other in a second match are not playing the same game as before. Therefore, we consider such repeated cases with the same two players in a match to be independent and we keep them in the data set.

Trimming

In trimming the data, we decide to exclude numerous variables in the data analysis process. The original data set contains 42 variables but many of these are not meaningful. We remove identifier variables pertaining to the players and rounds in addition to variables with many missing values and variables that would clearly be very highly or perfectly correlated with a victory (such as matches or sets won) without any quantitative measurements.

Finally, we consider the potential issue of multicollinearity. As a cutoff, we consider any two variables with a correlation coefficient greater than or equal to 0.85 to be "highly correlated." As a result, we remove multiple variables, which are described in greater detail in the Appendix. We also check to see if multicollinearity is an issue after fitting our models.

Modeling

The first question we ask, given our trimmed data, is: What is the optimal model for predicting the outcome of a match, given the performance of both players?

We know that a logistic regression model is appropriate since our response variable, "Result," is binary. Thus, we first consider a full model (using all 14 available predictors in the trimmed data set across both players) and use a stepwise procedure to find a potentially "best" model. In order to consider a couple candidates, we use both the BIC and AIC criteria for the

stepwise procedure. Additionally, when fitting the models, we randomly assign 365 of the 486 observations in the trimmed data set to train the model while the remaining 121 observations are used to validate that model. We consider the accuracy of each model on both the training and validation sets later.

After running the stepwise procedure, we find that our best model under the BIC criterion is:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 0.9091 + 0.2195\,FSW.1 + 0.1865\,SSW.1 + 0.3421\,BPC.1 + 0.2691\,BPW.1 - 0.2362\,FSW.2 - 0.1873\,SSW.2 - 0.3521\,BPC.2 - 0.2870\,BPW.2$$

where $\hat{p}$ indicates the predicted probability of player one winning the entire match, FSW refers to first serves won, SSW to second serves won, BPC to break points created, BPW to break points won, and the "1" or "2" appended at the end of each variable refers to whether player one or two made that action. It is interesting that, according to the coefficients of each variable, a change in break points created for either player would have the largest impact on the probability of player one winning the match when compared to the other variables, including break points won, given that all the other variables in the model remain constant. We also note that all those variables for which a unit increase would be intuitively better for player one, do in fact have positive coefficients while the opposite is true for those variables applying to player two. This is fitting given that we had defined that $\hat{p}$ indicates the predicted probability of player one winning.

We can check for potential issues with our logistic model using plot diagnostics (please refer to Figure 1 in the Appendix). It is clear from these plots that case 53 stands out as a possibly influential point. However, this case's Cook Distance is well below the 50th percentile for its appropriate F distribution, according to the fourth plot of residuals vs. leverage. Therefore, although case 53 has a large residual based on the predictive model, we will not consider it an influential point. We can say similarly for case 96 which also appear to be potentially (but not truly) influential points. We can bear these cases in mind as we continue with our analysis but we will not remove them from the data set.

To check for any potential issue of multicollinearity in this model, we use the variance inflation factor (VIF) from the package "car." Using this function, we find that the VIF values for the regressors are considerably below the cutoff of 10 that usually indicates serious multicollinearity:

FSW.1	SSW.1	BPC.1	BPW.1	FSW.2	SSW.2	BPC.2	BPW.2
7.8525	2.6602	2.1012	2.3324	8.9699	2.3972	2.1631	2.2852

Thus, we do not have multicollinearity in this model.

When we consider the accuracy of our model, we find that the BPC.1 and BPC.2 model accurately predicts 93.2% of the cases in our original data set and 93.4% of the cases in our validation data set using a 50% probability as a cutoff for Result. Most of the probabilities of a win or loss obtained through predictions are very close to either 0 or 1. Thus, using cutoffs that are moderately different from 50% (such as 40% and 60% or even 30% and 70%) do not make much of a difference, if any, in the accuracy of this model. This is also true for other models to come.

We now want to consider our model under the AIC criterion in the stepwise procedure. We find that our best model under this condition is slightly more complicated than that under the BIC criterion:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 6.6438 + 0.1054\,FSP.1 + 0.2343\,FSW.1 + 0.2972\,SSW.1 - 0.1565\,DBF.1 + 0.3349\,BPC.1 + 0.2840\,BPW.1 - 0.1809\,FSP.2 - 0.2310\,FSW.2 - 0.3494\,SSW.2 - 0.3836\,BPC.2 - 0.3089\,BPW.2$$

Our model now additionally includes DBF.1 and DBF.2 (double faults by each player), BPW.1 (break points won by player one), and FSP.1 and FSP.2 (first serve percentage by each player). We could interpret our coefficients similarly to as we did previously with the BIC model.

We again check for model diagnostics using plots (please refer to Figure 2 in the Appendix). We obtain plots very similar to those we obtained earlier. Again, cases 53 and 195 stand out as potentially influential points but both have Cook's Distances falling well below the 50th percentile of the appropriate F distribution, indicating that these observations are not influential. In this case we other potentially influential observations – cases 5 and 79 - but they are even less severe than cases 53 and 195. Thus, we can again bear these cases in mind when continuing with analysis but we will not adjust for them.

Checking for multicollinearity with the variance inflation factor (VIF), we find that the VIF values are:

FSP.1	FSW.1	SSW.1	DBF.1	BPC.1	BPW.1
4.2255	10.6242	7.4091	1.3838	2.025	2.3314

FSP.2	FSW.2	SSW.2	BPC.2	BPW.2
4.1828	10.9715	6.2887	2.0852	2.4272

We note that the VIF values for FSW.1 and FSW.2 both exceed 10 which indicates that this model may be subject to multicollinearity in addition to being more complicated than the BIC model given.

When considering the accuracy of this AIC model, we find that it accurately predicts 94.8% of the cases in our original data set, making it very slightly more accurate in these terms than our BIC model, but only 89.3% of the cases in our validation set. Thus, although the AIC model is a slightly more accurate for our training data set, it is worse for the validation set. Additionally, this AIC model is much more complicated than the BIC model we obtained earlier which involved fewer predictors. Based on these conditions, we would choose the BIC model as the better predictive model of the two, for the sake of modestly easier interpretation and accurate prediction.

Conclusion

It seems that the BIC model:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 0.9091 + 0.2195\,FSW.1 + 0.1865\,SSW.1 + 0.3421\,BPC.1 + 0.2691\,BPW.1 - 0.2362\,FSW.2 - 0.1873\,SSW.2 - 0.3521\,BPC.2 - 0.2870\,BPW.2$$

is the most accurate among those models considered, even though it has only two predictors. This indicates that among the many different predictors initially included in the data set, break points created by each player are very useful predictors for determining who wins a match.

We may also want to consider the possibilities for future research questions and models. For instance, instead of using data solely gathered from a match being played, we could collect data regarding different characteristics of each player, such as time spent practicing per day, the type of court that a player is used to playing on (grass, clay, synthetic materials), whether a player practices indoors or outdoors, and so on. It is very possible that such attributes make a difference in the outcome of matches. The only drawback to this would be that the data required to run these analyses may not be readily available to us or may depend on interviews and player comments rather than actual data. Overall, there are many interesting possible extensions and questions that looking into this data set brings up that can be answered with further data collection and research.

Appendix

Trimming

In trimming the data, we decide to exclude numerous variables in the data analysis process. The original data set contains 42 variables but many of these are not meaningful. We remove the names of the players in a match and the round in which they play because these variables are not pertinent to predicting victory. Similarly, we decide to take out ST1.1, ST2.1, ST3.1, ST4.1, ST5.1, ST1.2, ST2.2, ST3.2, ST4.2 and ST5.2 (individual set results), TPW.1 and TPW.2 (total points won), and FNL.1 and FNL.2 (final number of games won). The set total variables are left out because by definition, the more sets a player wins, the more likely he will win the match. So, these are not meaningful variables to predict a player's chance of winning. The same is true of total points won and the final number of games won.

We exclude other variables due to a large number of missing values: NPA.1 and NPA.2 (net points attempted), NPW.1 and NPW.2 (net points won), WNR.1 and WNR.2 (winners earned), and UFE.1 and UFE.2 (unforced errors).

Finally, we consider the potential issue of multicollinearity. As a cutoff, we consider any two variables with a correlation coefficient greater than or equal to 0.85 to be "highly correlated." As a result, we remove SSP.1 and SSP.2 (second serve percentage) because these are equal to 1 - FSP.1 and 1 - FSP.2 (first serve percentage) respectively. We can also check to see if multicollinearity is an issue after fitting our models.

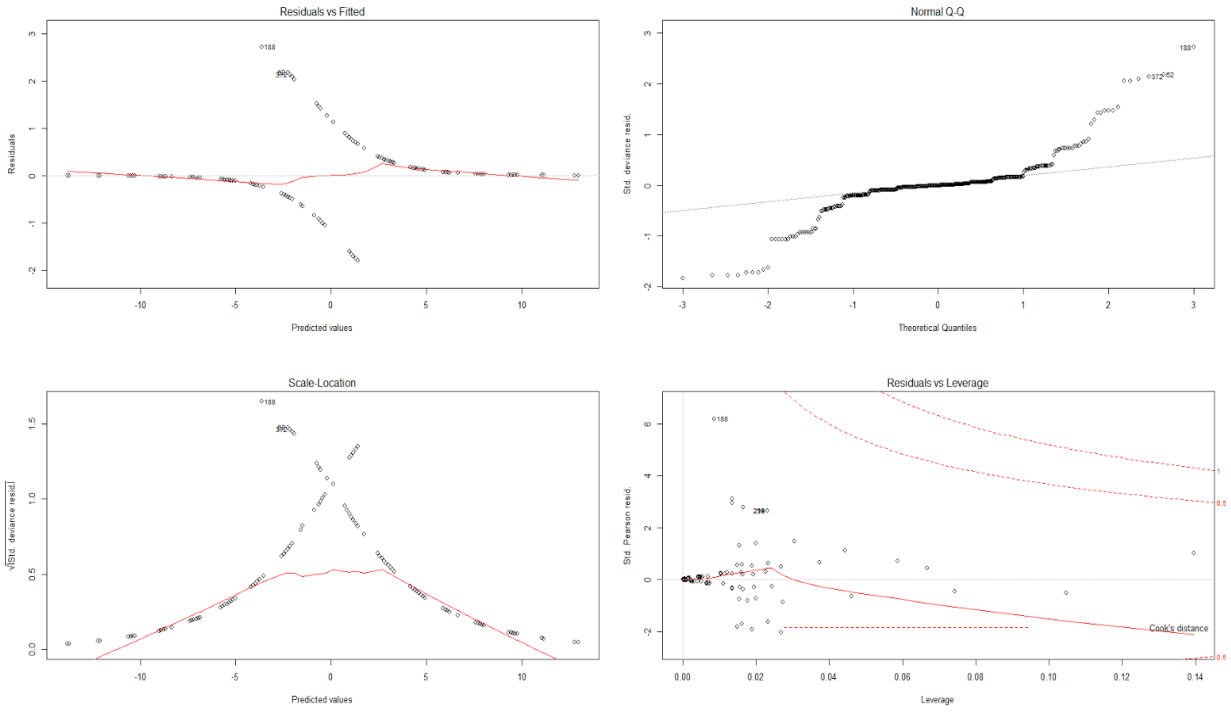

Figure 1

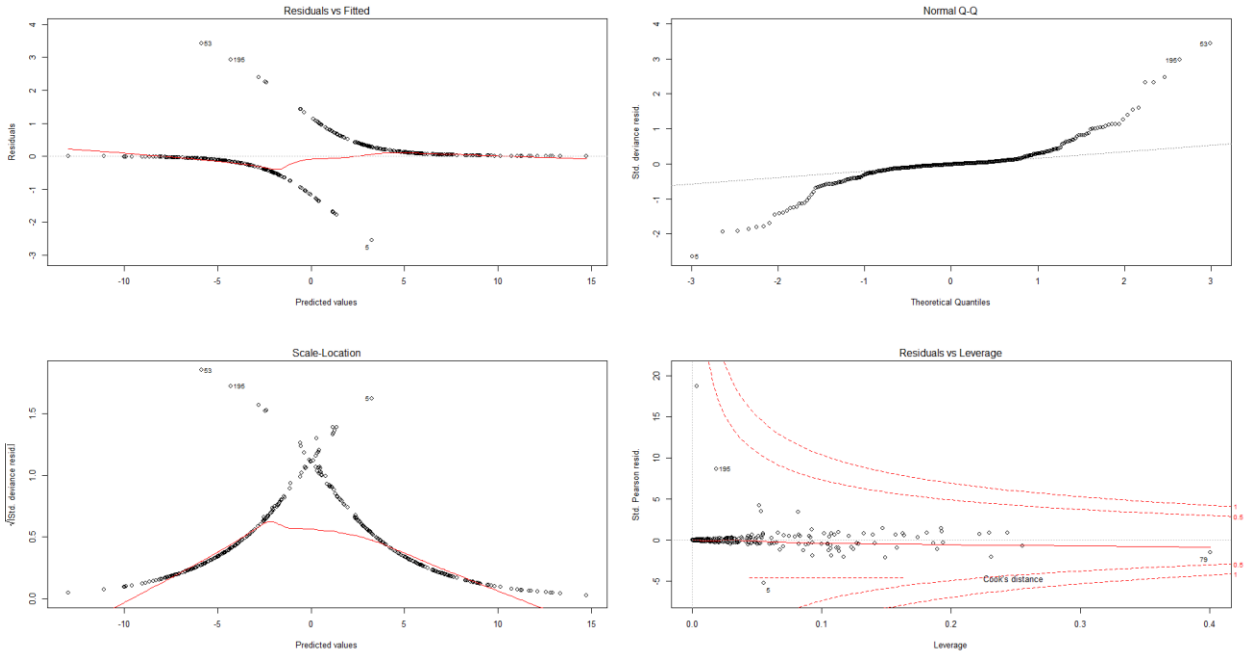
Residuals vs Fitted
Residuals
Predicted values
Normal Q-Q
Std. deviance resid.
Theoretical Quantiles
Scale-Location
√|Std. deviance resid.|
Predicted values
Residuals vs Leverage
Std. Pearson resid.
Cook's distance
Leverage

Figure 2

References

Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.