CAN STATISTICAL LEARNING OUTPERFORM BOOKMAKERS?

DSE - Statistical Learning Project
University of Milan

Author
Davide Matta

Contents

1 Introduction

The aim of this project is to find a statistical learning model that can outperform the odds in correctly predicting the outcome of tennis matches.

In order to achieve this goal, supervised learning algorithms will be used to make the prediction, while unsupervised algorithms will be used as data analysis tools.

In this document only the main points will be reported, but additional analyses that have been performed can be found in the code (see the last section).

2 Methodology and Data

2.1 Methodology

In order to simplify the problem, the analysis will be conducted focusing on a single player. This allows to reduce the number of necessary variables and the computation times.

The analysis has been conducted first on ATP #1 Novak Djokovic, and then has been replicated on #3 Daniil Medvedev and #4 Casper Ruud. Since ATP #2 Carlos Alcaraz has so far taken part to a too low amount of matches, Ruud has been chosen instead.

As it could be expected, models and their results change for each player, reflecting their different characteristics.

It is also important to specify that the main goal is to obtain models with a high accuracy without giving too much importance to other metrics like the R-squared or the mean square error, which however was computed. The main reason for this choice is that we cannot compute the R-squared for the odds, and it is difficult to compute the mean square error as well (odds could in principle be converted to probabilities, but bookmakers adjust them to match betting flows and keep the activity profitable regardless of the match outcome). What we can actually do with precision is to look whether the favorite player according to the odds wins or not.

2.2 Data

The dataset is available at this link: Kaggle - ATP Tennis 2000 - 2023 Daily update. Note that since the dataset is updated at the end of every tournament, the currently available version is different to the one used in this work. In particular, the latest available tournament was the 2023 Italian Open (Internazionali BNL d'Italia), whereas in the meanwhile other tournaments have been played.

The dataset contains information for all the ATP matches played since 2000. Matches ended by retirement are excluded as well as walkovers. Matches from minor tours (Challenger and ITF) are excluded as well.

The variables useful for this project are the date of the match, the players, their rank and odds, who won the match, the kind of court and surface, the tournament series and round, and the maximum number of sets in the match. Starting from these variable, some manipulation has been done, which will now be explained. Different datasets have been created for the three analysed players. From now, unless differntly specified, the focus will be on Djokovic's example.

2.2.1 Data cleaning

After some exploratory analysis there were principally two problems. The first one is that in some few cases odds were negative or incredibly high, which is inconsistent: the involved rows have been dropped. The second one is that Casper Ruud's father Christian also played tennis until the 2000's, and they are both labeled as "Ruud C.": Christian's entries has been relabeled as "Ruud Cr.".

2.2.2 Change of perspective

Instead of generic "Player 1" and "Player 2", now all the columns refer to either Djokovic or Djokovic's opponent. The column "Winner" has been replaced by "Win", a binary variable equal to 1 in case Djokovic won the match, 0 otherwise.

2.2.3 Court and Surface

Courts can be indoor or outdoor. In addition, there are four different surfaces in the dataset: Hardcourts, Clay, Grass and Carpet. Since carpet courts are no longer used in the ATP Tour since 2009, matches played on carpet courts have been removed. Indoor clay matches have been removed as well, since they are extremely rare, and in addition there are not indoor grass court. This leaves only four different kinds of courts: Grass (outdoor), Clay (outdoor), Hard (both indoor and outdoor). Four binary variables have been created: Clay, Grass, Hard (outdoor hardcourts), Indoor (indoor hardcourts).

2.2.4 Number of set

Matches can be played at best of either 5 (Grand Slams, and until 2007 some Master 1000 finals as well) or 3 sets (all the other matches). A binary variable "Five_Set" has been generated instead of the "Best of" variable.

2.2.5 Series and Round

Instead of using the tournament tier (from ATP250 to Grand Slam) and the round (from 1st Round to The Final) a score for the importance of the match has been created. It closely resembles the number of ranking points the player obtains it that match is the last one he wins in that tournament. The minimum is 45 (first matches of ATP250 tournaments), the maximum 2000 (Grand Slams' finals).

2.2.6 Odds

A binary variable "Odd_Pred", equal to 1 if the predicted winner is Djokovic and 0 otherwise, and a binary variable "Odd_Acc" equal to 1 if the odds prediction is correct and 0 otherwise have been created.

2.2.7 Opponent Characteristics

Since players perform differently according to circumstances, two variables have been created to address this fact. "Opp_Court_Perc" shows the percentage of matches the opponent won on that particular kind of court and surface up to the date of that match. In case he has not played any match in that setting, the default value is 0. "Opp_Set_Perc" does the same thing but for the maximum number of set.

2.2.8 H2H

A variable called "H2H_Djo" that shows the percentage of matches won by Djokovic against that particular opponent up to the date of that match has been created. In case there was not any previous H2H, the default value is 50.

2.2.9 Previous Matches

In order to consider the latest performance of the players, the variabes "Pre_Djo" and "Pre_5Djo" have been created. Pre_Djo is equal to 1 if Djokovic won the last match he played, 0 otherwise. Pre5_Djo goes from 0 to 1 according to how many of the previous five match played up to that date he won (for instance, if he won 3 matches over 5, Pre5_Djo is equal to 0.6). Equivalent "Pre_Opp" and "Pre5_Opp" have been created for the opponents.

2.3 Potential issues

2.3.1 Time

A first issue is that the dataset could be interpreted from some angles as a time series, though some of the time series characteristics are not present. However:

- Data have been manipulated in order to take into account what happened previously, in particular the form of the players.
- Any improvements of the players is generally reflected by an improvement of their ranking.
- The fact that we computed the new variables using data only up to the values before the referred rows, prevents the models from using future data to influence past observation.
- The fact that we use updated statistics instead of just a variable with the opponent name allows us to not worry about the evolution of that specific opponent.
- The predisposition of a player to do well on a particular surface or rule can be reasonably assumed to stay constant over time, or at least this is what is considered to be true for the analysed players. For example Ruud was always known to be particularly strong on Clay courts and Medvedev on Hard courts, and this is not changing over time, while at the same way Djokovic has always been considered particularly tough in best of 5 matches.
- Other than using a classic random train-test sample division, the models will be tested also on the latest tournament, the 2023 French Open (whose data were not available when the dataset was downloaded).

2.3.2 Data completeness

Other issues involve directly the data:

- Hardcourts are a quite heterogeneous category of courts, probably a deeper categorization among slow, medium and fast hardcourts would have been appropriate but, other than not being available, it is probably negligible since many of the hardourts features (i.e. uniform and clean bouncing, difficulty to slide compared to clay courts) remain the same.
- Statistics are computed only from the year 2000, so for older players, their first career years are excluded from the computations. However, this does not same a large extent problem.

- Data are about only ATP Tour, so performance in matches from lower tours are not reflected by statistics. However, only low-ranking players usually take part to those matches.
- Other variables such as the weather or the time past since player's last injury should probably be taken into account, but there is no easily available data.
- There are variables that cannot be measured or even known at all and that may influence the outcome, for instance how the players are feeling in the match day or the crowd behavior.

3 Descriptive Analysis

3.1 Variables Analysis

In this section the data relative to Djokovic's example will be described. Only the key points will be reported, additional insights can be found by executing the code.

- Rankings
  For the largest part of his career Djokovic has occupied one of the first positions in the ranking standings, while the ranking of his opponents is as expected much more variable. However, in both cases there are some outliers (for instance, at his first ATP match Djokovic was ranked 368). In order to limit the outliers' influence, we transform the ranking variables to logarithms: for Rank_Djo changes are minimal, but the logarithmic transformation of Rank_Opp is two times more correlated to the Win variable than it is Rank_Opp.
  Win is negatively correlated to Rank_Djo. This might seem counterintuitive, but actually more ranking points mean a numerically lower ranking. On the other hand, for the same reason the correlation with Rank_Opp is positive.

- Five Set
  As it was predictable, there is a positive correlation between the outcome of the match and the match being played at best of five sets.

- Court
  Chances of Djokovic's win are higher when playing on Hard and Grass courts, and lower on Indoor and Clay.

- Previous match
  Curiously, the correlation between the victory in this match and victory in the previous one is negative. Though it is difficult to explain why, this may be due to the fact that is extremely unlikely that Djokovic loses two matches in a row.

For all the other variables there is not much to report. A graph with the correlation index for each pair of variables is shown below.

Some of the variables that are most correlated to each others are the court dummies, and this is reasonable since each one of them is a linear function of the other three.

The ranking of the opponent is also highly correlated with the other variables that involve the opponent. It is also negatively correlated with the match importance: in fact, higher the importance of the match, higher the probability to find a strong opponent (then numerically lower ranking).

Now some unsupervised learning techniques will be used to see if there are relevant trends over the dataset.

3.2 Principal Component Analysis

PCA has been performed on the dataset. At first all the variable previously shown in the correlation graph have been used, except Win. Of course they have been scaled.

However, the approach was not particularly successful, and the first two principal component only explained less than 40% of the variance.

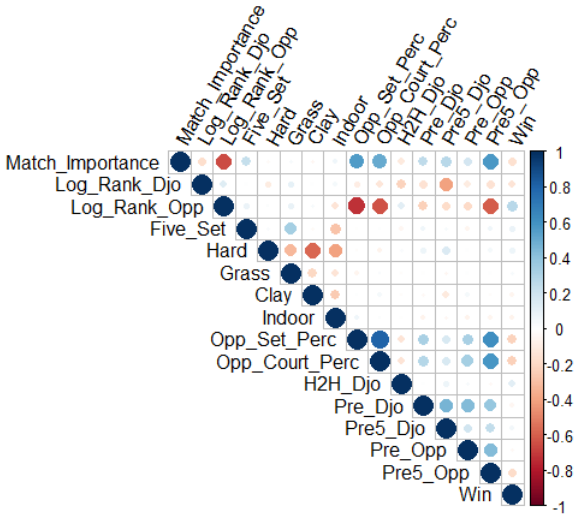

Figure 1: Djokovic - Correlation between each pair of variables

Despite this fact, there are still some facts to comment. The picture below shows the two most important principal components and how they are related to each feature.

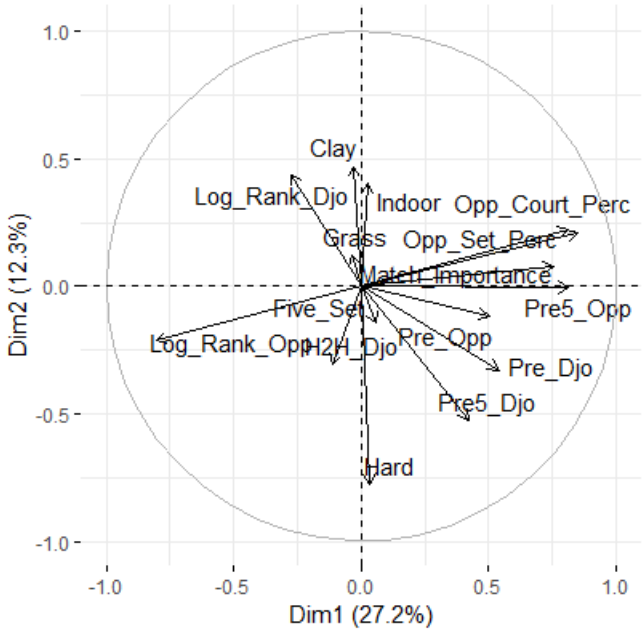

Figure 2: Djokovic - PCA on all variables

The first PC is mainly correlated with both Djokovic's and the opponents' statistics, negatively with the rankings, positively with the others. We can also see that the courts dummies' vectors lie very close to vertical axis, and that is because they are highly correlated with the second PC (except Grass).

3.3 2-Means Clustering

After the PCA, trying to apply a clustering algorithm may be a good choice. 2-Means Clustering may be appropriate since we can see if the clusters resemble in some way the Win - Not Win division. We use all the variables again, excluding Win. The results are shown below.

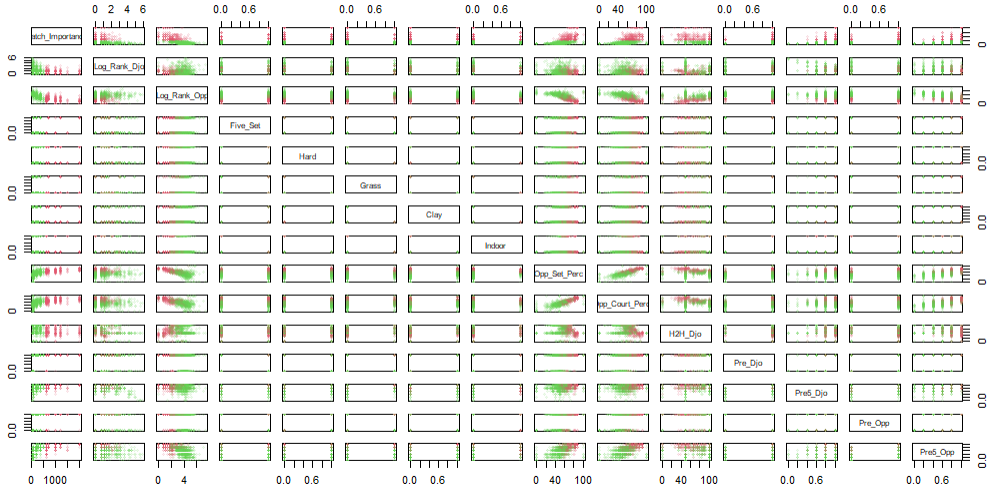

Figure 3: Djokovic - 2-Means Clustering

This is probably more interesting than what we obtained from PCA. Despite the algorithm does not attribute any particular meaning to the clusters, it seems possible to associate them to what we hope.

In fact, Cluster 1 is larger than Cluster 2, as wins proportion is larger than losses'. Furthermore, if we look at the same proportion dividing between wins and losses, we note that they are quite different: in matches won by Djokovic Cluster 1 covers around 76% of the observations, whereas in lost matches only 56% of the observations belong to Cluster 1.

Although there is not an explicit link between the clusters and the matches outcome, we might try to use the clusters as a first trivial predictor. If we consider belonging to Cluster 1 as a prediction for Win = 1 and to Cluster 2 for Win = 0, we get a 70.87% accuracy. However, other than not having any statistical value, this is still far from the 85.79% achieved by the bookmakers on the same data.

4 Models

4.1 Overview

Concluded the descriptive analysis, it is the moment to see whether it is possible to construct reliable predictors for the outcome of a tennis match or not. The odds accuracy, which is our benchmark, for the whole Djokovic dataset is 85.79

The sample, that contains 1126 observations, has been randomly splitted in a training (70%) and a test (30%) set. The classification algorithms have been performed on the training set and then tested on the test set. The odds accuracy in the test set is 86.35

The starting point is the linear regression model. The model has then been modified to address the emerged issues. When it was not possible to do it in the standard linear regression framework, robust regression and logistic regression have been used.

4.2 Linear Regression

As a first try, we performed a multiple linear regression model. All the variables shown in previous figures have been included, except "Hard" in order to avoid perfect multicollinearity. Output is shown below.

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       7.652e-01  1.390e-01   5.504 5.06e-08 ***
Match_Importance -3.176e-05  4.259e-05  -0.746 0.456010
Log_Rank_Djo     -4.646e-02  1.279e-02  -3.632 0.000299 ***
Log_Rank_Opp      4.225e-02  1.673e-02   2.525 0.011766 *
Five_Set          7.978e-02  3.165e-02   2.521 0.011898 *
Grass             1.852e-02  4.364e-02   0.424 0.671437
Clay             -5.597e-02  3.006e-02  -1.862 0.062974 .
Indoor           -4.159e-02  3.815e-02  -1.090 0.275980
Opp_Set_Perc      8.712e-04  1.402e-03   0.621 0.534568
Opp_Court_Perc   -2.505e-03  1.215e-03  -2.062 0.039567 *
H2H_Djo           6.078e-04  4.920e-04   1.235 0.217034
Pre_Djo          -3.520e-02  4.137e-02  -0.851 0.395168
Pre5_Djo          1.160e-01  9.140e-02   1.269 0.204762
Pre_Opp           8.332e-02  3.760e-02   2.216 0.026973 *
Pre5_Opp         -1.439e-01  7.850e-02  -1.832 0.067263 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3407 on 774 degrees of freedom
Multiple R-squared:  0.1233,        Adjusted R-squared:  0.1074
F-statistic: 7.772 on 14 and 774 DF,  p-value: 1.511e-15
```

Despite the low adjusted R-squared, the model actually has a high test accuracy: 84.27%. This is already close to what bookmakers achieve. However, we can do better.

4.2.1 Best subset selection

It is immediate to note that not all the coefficient are significant, and probably a better result can be achieved with another combination of variables. Since we do not have a too large number of features we can afford to perform a best subset selection algorithm.

The algorithm has been set to minimize Mallows Cp, which penalize adding new regressors. The new model has seven attributes, and the results are displayed below.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.882672   0.096970   9.103  < 2e-16 ***
Log_Rank_Djo   -0.054752   0.011436  -4.788 2.02e-06 ***
Log_Rank_Opp    0.047540   0.013771   3.452 0.000586 ***
Five_Set        0.083395   0.026525   3.144 0.001729 **
Clay           -0.053495   0.027540  -1.942 0.052436 .
Opp_Court_Perc -0.002258   0.001000  -2.258 0.024223 *
Pre_Opp         0.082748   0.035259   2.347 0.019182 *
Pre5_Opp       -0.141763   0.074686  -1.898 0.058050 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3404 on 781 degrees of freedom
Multiple R-squared:  0.1167,        Adjusted R-squared:  0.1088
F-statistic: 14.74 on 7 and 781 DF,  p-value: < 2.2e-16
```

Now all the coefficients are significant, but the adjusted R-squared did non changed in a relevant way. However, we have a small increase in accuracy, that reaches 84.56%. Since the performances are not affected, it can be reasonable to keep only these seven predictors and simplify the model.

4.2.2 OLS Assumptions

Now we will go through the OLS assumptions to see if we are breaking any of them, and what the possible solutions are.

- Omoscedasticity
  Although this assumption is not relevant for us, since we do not aim to perform test or build confidence intervals, it may be worth to run the model with heteroscedastic errors. The result is that the significance of coefficients has not been affected.

- I.I.D. sample
  Training and test sets have been chosen randomly from the population. The autocorrelation issue have already been addressed in subsection 2.3.1..

- Multicollinearity
  The variance inflation factor (VIF) can be used to detect if multicollinearity is an issue in our model. As Figure 4 shows, all VIFs are low enough to allows us to state that multicollinearity is not an issue here.

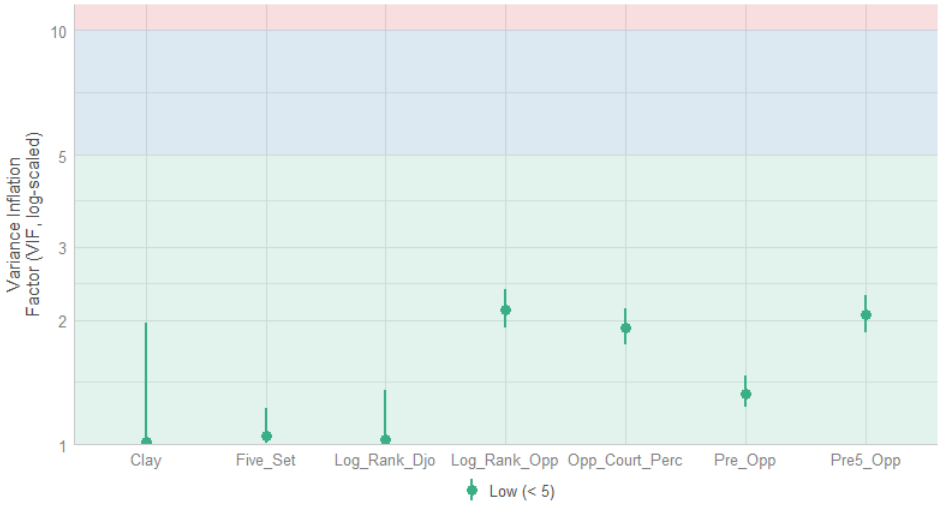

Figure 4: Djokovic - VIF for 7-variables linear model

- Outliers
  The dataset appeared to be clean. However, a robust regression using the M estimator has been run to see if results change. This model will be commented later.

- Linearity
  Since the response variable is binary, this is probably the weakest assumption in our model. Logistic regression seems more appropriate for the task. A logit model will be discussed later.

- Other assumptions
  There are other two assumptions: Normality of residuals and Exogeneity. Since we use the model for predictions and not inference, these assumptions are not so relevant. However, we can go through them.

  Exogeneity. As previously stated, there are variables that we cannot include and that are probably correlated with the outcome of the match. However, it is unlikely that they are correlated with the other regressors, thus the error would still be exogeneous.

  Normality. Since the sample is large, by asymptotic theory even though errors' distribution is not normal, predictions resemble a normal distribution. However, the performed logit model seems to be close to respect this assumption since its errors approximately follow a normal distribution.

4.3 Robust OLS

As anticipated in the last paragraph, a robust method has been performed using M estimators. The main goal is to reduce eventual outliers' influence. It is still worth noting that M estimators are particularly effective when outliers are in the response variable, but their power with respect to outliers in the features, which is what interest us, is more limited. However, below are the estimates for the model that includes all the variables.

```
Coefficients:
                 Value      Std. Error t value
(Intercept)         1.0000     0.0000 72322.4899
Match_Importance    0.0000     0.0000    -0.4390
Log_Rank_Djo        0.0000     0.0000    -4.1483
Log_Rank_Opp        0.0000     0.0000     3.3705
Five_Set            0.0000     0.0000     1.9970
Grass               0.0000     0.0000     0.5434
Clay                0.0000     0.0000    -1.9282
Indoor              0.0000     0.0000    -0.7677
Opp_Set_Perc        0.0000     0.0000     0.6613
Opp_Court_Perc      0.0000     0.0000    -2.3127
H2H_Djo             0.0000     0.0000     1.7549
Pre_Djo             0.0000     0.0000    -0.7532
Pre5_Djo            0.0000     0.0000     1.7291
Pre_Opp             0.0000     0.0000     2.5050
Pre5_Opp            0.0000     0.0000    -1.4825
```

It seems that the robust method always predict a Djokovic win. Since he has won the most of the matches he played in his career, this provides some accuracy, but it is not wise to rely on such model.

If we try to estimate the model keeping only the seven most relevant variables, coefficients still remain zeros (actually in both models they are not exactly zero, but they are small enough to be approximated as zeros).

The two models have a 83.98% accuracy, which is still surprisingly high for models of this kind, but is a natural consequence of Djokovic's career.

4.4 Logistic Regression

The last algorithm is logistic regression. It is a natural choice, since it is more suited for tasks that involve a binary response variable. In fact, each prediction is by construction bounded between 0 and 1, and the logistic function allows for more flexibility compared to a linear relation.

About the assumptions, what held previously still holds, but this times the residuals fit better the model expectations. In addition, as already stated, logistic regression does not require the linearity assumption.

Below are reported the results of the model with all the variables. Some variables seem more relevant than they were in the linear model, while others have lost significance.

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.5402085  1.3466816   1.886 0.059258 .
Match_Importance -0.0003917  0.0003431  -1.142 0.253627
Log_Rank_Djo     -0.4079283  0.1076745  -3.789 0.000152 ***
Log_Rank_Opp      0.2466530  0.1558347   1.583 0.113470
Five_Set          0.9713003  0.3337326   2.910 0.003609 **
Grass             0.2714949  0.4732036   0.574 0.566145
Clay             -0.4707050  0.2652761  -1.774 0.075998 .
Indoor           -0.3832652  0.3034049  -1.263 0.206512
Opp_Set_Perc      0.0083239  0.0161518   0.515 0.606304
Opp_Court_Perc   -0.0250139  0.0131837  -1.897 0.057784 .
H2H_Djo           0.0028550  0.0044168   0.646 0.518032
Pre_Djo          -0.2750968  0.4125233  -0.667 0.504859
Pre5_Djo          0.8870597  0.7828058   1.133 0.257139
Pre_Opp           0.7402168  0.3520336   2.103 0.035493 *
Pre5_Opp         -1.9142001  0.7622026  -2.511 0.012025 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The model with seven variables has been estimated as well. As it is shown below, all variables are somewhat significant again as in the linear regression case.

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.95315    0.96320   3.066  0.00217 **
Log_Rank_Djo   -0.44830    0.09310  -4.816 1.47e-06 ***
Log_Rank_Opp    0.33241    0.12372   2.687  0.00722 **
Five_Set        0.97584    0.26721   3.652  0.00026 ***
Clay           -0.42682    0.23585  -1.810  0.07034 .
Opp_Court_Perc -0.02057    0.01040  -1.978  0.04797 *
Pre_Opp         0.74296    0.32527   2.284  0.02236 *
Pre5_Opp       -1.83252    0.72785  -2.518  0.01181 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From these two models we can expect a better performance than those of the linear models. In fact, that is what happened: the all variables model has shown a test accuracy equal to 86.94%, while the seven variables model scored 86.35%. Since the first model does not show any sign of multicollinearity, we keep it because of its higher accuracy.

But what really matters is that the project's goal has been reached. The logit model outperformed the bookmakers by slightly more than half a percentage point.

5 Odds - Models Comparison

We now look in dept at how our best models performed with regards to the odds. We already talked about the overall accuracy, so what is interesting to see is the behavior of the models when the odds were wrong.

Limited to the test set, odds predictions were wrong 46 times. In most cases the three models were wrong as well, but the linear model and the robust model correctly predicted 8 of these matches, while the logit did it 10 times.

Odds' inaccuracies involve matches where Djokovic both won and lost. For all of those 46 matches, the linear and robust models always predicted a win, whereas the logistic regression sometimes predicted losses as well.

Probably the best quality of the logit model is that, differently from the other two, it is able to provide loss predictions and get them right also when the odds predicted a win. In the test set this happened three times, which add up to the seven matches that were correctly predicted as wins by the logit but as losses by the odds.

On the other hand, the linear model predicts losses only in some cases when also the odds does, while the robust model always predict wins.

Also for this reason there is no doubt that for our purpose the logit is the best model among the ones we estimated.

6 Replicating the analysis on other players

The goal of this section is to understand whether the methods that work with Djokovic work with other top players as well.

It is worth noting that both Medvedev and Ruud are much younger than Djokovic, and thus we have a lower number of matches to train and test the models. If before we had 1126 matches, now we only have 385 for Medvedev and 236 for Ruud.

Now the results on the two players will be commented separately, highlighting only the most relevant results and the differences with Djokovic.

6.1 Daniil Medvedev

6.1.1 Player characteristics

Contrary to Djokovic's, Medvedev's chances to win a match are positively correlated with playing on an indoor court and negatively with grass courts. There is also a positive correlation with the outcome of Medvedev's previous match (the index was negative with Djokovic). Other than that, there is not much to highlight about data.

Medvedev is also been more difficult to predict for the bookmakers compared to Djokovic. The test set odds accuracy for Djokovic was 86.35%, but for Medvedev it drops to 73.91%, so we probably should expect our models to work worse as well.

6.1.2 Models and Results

- Assumptions
  The first important aspect to comment is that the OLS assumptions seems to hold with Medvedev too. In particular, the logistic regression residuals clearly resemble a normal distribution, even better than with Djokovic.

- Test vs Training
  An interesting difference is that in this case, performance on the test set are significantly worse compared to those on the training set, while with Djokovic they were similar and in some cases even slightly better.

- Subset selection
  Since Medvedev style and characteristics are differeìnt from Djokovic's, we could also expect that the best subset selection algorithm selects different variables, and this is what happened. Five variables are selected instead of seven: with respect to Djokovic, we do not have Five_Set, Pre_Opp and Pre5_Opp, but we have Pre_Med.

- Robust regression
  This time the robust method is able to predict losses as well, since now many coefficients are different from zero.

- Accuracy
  But now here comes what matters more: none of the models outperformed the odds. The most accurate standard linear model achieves a 71.30% accuracy, the best robust 71.30% as well, and the best logit 73.04%. The logit is still the best models but, despite being close, it does not do better than the odds. It is also interesting to note

that the only case where reducing the number of variables increased the accuracy is the logit, while with Djokovic it was the opposite.

- R-squared
  Moreover, despite the accuracy is lower, the linear model's adjusted R-squared is slightly higher compared to Djokovic's.

- Odds errors and comparison with the models
  If we look at what the models did when the odd were wrong, we note that the three models provided the same predictions (except the linear that predicted a win -like the odds- while the other two correctly predicted a loss). However, this time all the models predict both wins and losses and there are cases where they were able to predict the correct outcome when odds were wrong in both directions. Over 30 wrong odds predictions, 8 were correctly predicted by the logit and robust regressions, while only 7 by the standard linear regression.

6.2 Casper Ruud

6.2.1 Player characteristics

This time, the most relevant difference it that win chances are positively (and strongly) correlated with the match being played on clay, which is the only case among the three players, and slightly positive with Indoor. There is also a significant negative correlation with Grass, and a smaller negative correlation with Five_Set. Other than this, data are similar to Medvedev's and Djokovic's.

For the bookmakers, predicting Ruud was even more difficult than with Medvedev. Odds predicted the right result only 70% of times.

6.2.2 Models and Results

- Assumptions
  Assumptions still hold. This time, relation seems actually more linear than with the other players and the linear model fits well, despite still not as well as the logit.

- Subset selection
  This time we have six variables. Compared to Djokovic we do not have Pre_Opp, Pre5_Opp and Five_Set, but we have instead Indoor and H2H_Ruud. However, the latest two variables are not significant.

- Similarities with Medvedev
  Predictions on the test set are less accurate than on the training set as it was with Medvedev.
  This time the adjusted R-squared is even higher (twice as much as Djokovic's).
  Robust regression has non-zero coefficients again.

- Accuracy
  This time reducing the number of regressors to six increased the accuracy of all the models, except the robust that kept the same level of accuracy. The peculiar result

is that all the models, in their reduced versions, have a 70% accuracy, that is exactly equal to the odds.

- Odds errors and comparison with the models
  This time, odds were wrong 21 times in the test set. Although our models had the same level of accuracy, predictions were different. Standard OLS and logit predicted right 8 of those 21 matches, while robust OLS only 7. All models where able to provide opposite predictions in case of both wins and losses as well as they did with Medvedev.

7 Accuracy on future data

7.1 New data

After the models were trained, data for three new tournament were made available: Lyon (ATP 250), Geneva (ATP 250) and the French Open / Roland Garros (Grand Slam).

The three players all took part to the French Open, but none of them was present in Lyon and only Ruud played at Geneva. In addition, for Lyon and Geneva data on the ranking of the defeated players are for some reason corrupted (the number of points is shown instead of the standings position) and it is impossible to predict values for those matches.

However, those matches can still be used for updating the statistics (form, head to head and win percentage according to court type and number of sets). Then they are kept in the dataset despite we will test the models only on the French Open data. We will use the logit since it has been the one that performed better on test set.

Djokovic and Ruud performed well, since they both reached the Final, where Djokovic beated Ruud. On the other hand, Medvedev was upsettingly defeated by ATP #131 Thiago Seyboth Wild in the first round.

7.2 Predictions - Djokovic

We have seven match to check, from the first round to the Final. Since Djokovic won the tournament, he obviously also won all the matches. The odds predicted correctly all his wins except the semifinal. Our logit model showed instead 100% accuracy.

7.2.1 Djokovic vs Alcaraz

Since we produced a different prediction for this match, it deserves to be seen in depth.

This was a very awaited match. Djokovic is the last one of the "Big Three" (the other two are Rafael Nadal, who is suffering a serious injury, and Roger Federer, who retired last year) that is still competitive nowadays, while 20 year old Alcaraz is considered as the most accredited player to dominate the sport in the next following years and he entered the French Open as ATP #1 (Djokovic regained the head of the ranking after winning the tournament). Djokovic showed some signs of decline in the previous months but, for his whole career, Grand Slams have been the venues where he performed at his best. In addition, winning the Roland Garros would have (and has) made him the player with most Slams victories in the tennis history, overtaking his rival Rafael Nadal.

The public expected the match to be particularly balanced. Djokovic odds were 2.63, which means that according to the bookmakers, he approximately had a 36% chance to win, while our logit model gave him a 54% chance.

The match actually was balanced: Djokovic won the first set, Alcaraz the second, but at the beginning of the third set, the Spanish started suffering some cramps, which is fairly rare at this level, but it may have been the consequence of Djokovic's strategy, that in the first sets seemed to aim at making his opponent run as much as possible, which could eventually cause cramps. Despite Alcaraz managed to continue the match, Djokovic easily won the third and fourth sets, achieving a 3-1 victory in line with what the logit predicted.

7.3 Predictions - Medvedev

There is not much to say here: no one would have predicted Medvedev'loss. His odds were 1.05, which is close to a 91% probability of victory, while our logit predicted 95%.

7.4 Predictions - Ruud

Ruud played seven matches like Djokovic, and has won each one of them except the final against Djokovic. Both the odds and our logit model predited correctly the outcome of all the matches. For the final, Ruud's logit predicted a 42% chance for a Ruud win (from Djokovic's logit model perspective, Ruud had 36%, so the two models are somewhat close to each other), while the odds only gave him 20

8 Conclusions

To sum up, we can be at least satisfied with what we achieved. The models' assumption are satisfied or at least not relevant for our purpose, and most of the models seem to work well, in some cases even better than the odds, which was the initial goal. In addition, their predictive power is good with both the new available data and the randomly extracted old data test set.

The accuracy of the models is not the same for every player, but neither the odds accuracy is. That may be because of the different sample sizes, or maybe just because some players are by characteristics harder to predict than others.

9 Code

The R code can be found at this repository: GitHub - DavideMatta/TennisticalLearning

- The folder Data contains the original dataset (both the old one and the update) and its transformations.
- The folder Djokovic contains four R files:
  - rdatacleaning: it contains the code for transforming the dataset.
  - rdataanalysis: it contains indexes and charts useful for understanding the data.
  - runsupervised: in this file are performed the unsupervised techniques.
  - rsupervised: in this file are performed the supervised techniques.

  This folder also contains the images attached to this pdf.
- The folders Medvedev and Ruud have the same structure of Djokovic, but their files need to be executed after Djokovic's rdatacleaning.
- The file rRG has been used to create the test datasets with the Roland Garros data.
- The file rfunctions contains the functions used for data cleaning.