University of Liège

PROJ0016 - Big Data Project

Predicting the 2019 French Open's winner

Master 1 in Data Science & Engineering

Authors:
Antoine Louis
Tom Crasset
Maxime Lamborelle

Professors :
G. Louppe
P. Geurts
B. Cornelusse

Academic year 2018-2019

Contents

1 Introduction

According to many surveys, such as the one investigated by `Bloomberg` [1], tennis is the fourth most popular sport in the world. Each year, a multitude of tournaments are held by the Association of Tennis Professionals (ATP) in 30 different countries. The matches attract a huge amount of spectators, in the stadiums and on TV. This results in a tennis betting market that is constantly growing. Being able to predict which player has the biggest chances to win a match becomes interesting, on one side for the sport betting sites that obviously want to maximise their profits, and on the other side for all the gamblers that look to multiply their gains.

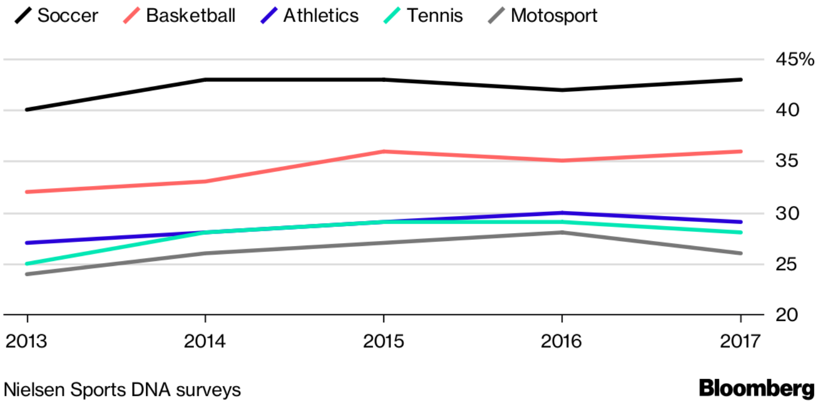


Figure 1: Percent of people who say they are "very interested" or "interested" in some popular sports in 18 markets across Americas, Europe, the Middle East and Asia.

As a consequence, extensive research have been pursued during the last years in order to predict the winner of tennis matches. Many of them use machine learning on past performance data of both players to predict their percentage of winning the match. For example, Barnett and Clarke [2] shows in their highly mentioned research paper how the standard statistics published by the ATP can be combined to compute statistics which can then be used to predict match outcomes. More recent researchers, such as Sipko [3], go even further by using real-time in-game statistics to predict the turn of a match.

In this project, we extend the outcome prediction of a match to the prediction of the winner of an entire tournament, namely the 2019 French Open that will be held in late May in the iconic city of Paris. To do so, we first propose a supervised machine learning approach that uses pre-processed historical data on past tennis matches to predict future match outcomes. Then, using this match-predictor model, we define a method that uses Monte-Carlo simulations on real possible draws to estimate the percentage for each participating player to win the tournament. As a result, our final tournament-predictor model was tested on the three last French Opens and predicted the actual winners for two of them.

2 Dataset and features

2.1 Dataset

To begin, we need to collect data. In order to train a tennis prediction model, two types of data are required : historical data about previous tennis matches and statistical data about the players and their performances. The former would reflect the date of the match, the surface on which it has been played, the type of tournament (Grand Slam, Master, ATP World Tour Finals or ATP 500/250), the considered round, etc. The latter would summarize the average winning percentage of a player, his average first serve percentage, double faults, aces, etc.

Historical data about past matches is widely available online. Tennis websites such as `tennis-data.co.uk` provides it in a structured form (CSV or Excel files). Collecting all their open source data results in a dataset of approximately 50000 tennis matches of ATP tournaments from 2000 to 2018, taking a total of about 10 MB.

Statistical data about players is a bit more challenging to obtain. In fact, there doesn't exist a dataset that includes all players' statistics. Instead, there rather are a multitude of tennis websites, such as `ultimatetennisstatistics.com`, that compute players' statistics from many matches. Our initial idea was to scrap this website using a small python script, and this is what we did by jumping in head first. At first sight, this seemed straight forward, one just needs to extract the HTML from the site. However, a big hurdle arose: nearly everything on the site was dynamically generated using a Javascript code. The solution was to use a browser driver to emulate a browser that executes the Javascript from the site and then extracts the HTML code with all the statistics it included. This was done using the `selenium` python package. After that, it was only a matter of using the `BeautifoulSoup` package to extract the relevant information from the site, iterating through all the *playerIDs*. At the end, we had collected statistical data about 1000 different tennis players.

That way, merging the statistical data of the players with the corresponding matches of our historical dataset about past matches could have been our final dataset. However, after having collected these statistical data, we realised an important point that we missed before doing the scraping phase : `ultimatetennisstatistics.com` only gives us statistical data about a player for a given period of time, that is for example, the statistics of a player for the year 2017, the year 2018 or for his entire career. These player statistics could obviously not be linked the same way to all matches involving these players during the 2000-2018 period, as a players performance during his career does not specially reflect his performance at a given time. Thus, we had to find another way to collect representative statistics about players.

The solution was found thanks to Jeff Sackmann, an author and software developer who has worked in the fields of sports statistics and test preparation. In 2015, he published an extensive database of tennis results, rankings, and stats from 1968 to 2018 [4]. A lot of data was missing before 1990 though, so we only considered matches from 1993 to 2018, totalling over 70000 ATP matches between more than 1500 different professional players. We noticed that, in addition to the player statistics given per match, this dataset also contained relevant information about match details such as the date, the round, the surface, the draw size, etc, data that we previously collected through the `tennis-data.co.uk` website. Therefore, for reasons of simplicity, we considered Jeff Sackmann's dataset as our unique dataset, as it contained all the data needed for our future computations. The idea is to use the features of this dataset as a starting point for computing new statistics for the players. The features present in this dataset are shown in Table 1.

In the next section, we will discuss how we can use these basic features to compute more representative ones, that will themselves later be used to compute estimations of player performances.

Players details	Match details	Player statistics for the match
Name	Tournament name	Number of service points
Age	Tournament type	Number of first serves in
Nationality	Date	Number of first serve points won
Handedness	Draw size	Number of second serve points won
ATP rank	Best of	Number of aces
ATP points	Surface	Number of double faults
	Round	Number of break points faced
	Score	Number of break points saved
	Duration	

Table 1: Features of Jeff Sackman's dataset

2.2 Feature construction

From the previously described dataset, new features can be extracted concerning the stats of a player for a particular match. Indeed, instead of talking about "number of", which obviously depends on the total number of points played during the match (and thus on the match duration), more representative features can be computed in terms of percentage. We define :

1. Service points $\% = \frac{\text{Number of service points of Player A}}{\text{Number of service points of Player A + Number of service points of Player B}}$
2. 1^{st} serve $\% = \frac{\text{Number of first serves in}}{\text{Number of service points}}$

3. 1^{st} serve points won % $= \frac{\text{Number of first serve points won}}{\text{Number of first serves in}}$

4. 2^{nd} serve points won % $= \frac{\text{Number of second serve points won}}{\text{Number of service points - Number of first serves in}}$

5. Aces % $= \frac{\text{Number of aces}}{\text{Number of service points + Number of aces + Number of double faults}}$

6. Double faults % $= \frac{\text{Number of double faults}}{\text{Number of service points + Number of aces + Number of double faults}}$

7. Break points faced % $= \frac{\text{Number of break points faced}}{\text{Number of service points + Number of aces + Number of double faults}}$

8. Break points saved % $= \frac{\text{Number of break points saved}}{\text{Number of break points faced}}$

This way, these eight new features replace the ones expressing player statistics in the initial Jeff Sackmann Dataset [4] (see Table 1). Recall that these statistics only concern the performances of players during one given match, but what we would like as data to train a machine learning model is an estimation, for each given match, of the recent performances of both players before competing. The next section explains how to compute such performance estimations using our new dataset.

2.3 Weighted average statistics

At this stage, our dataset is composed of thousand of matches, each one being described by player details, match details and player statistics for that match. Ideally, to train a machine learning model, we would want, in addition to player details, to reflect the current performance of both players before a given match. It turns out that a fair estimation of these players performance can be achieved by averaging, for both players, the new constructed features described in the previous section over all the matches preceding the time we want to estimate their performances.

However, when estimating the current performance of a player, it seems obvious that more recent matches better reflect the form of the player than older ones, and thus need more attention. This first aspect will be considered in the "time discounting" method, described in Section 2.3.1. Moreover, it has been proven that player performance is affected by the court surface [5]. Indeed, tennis is played on four different surfaces : clay, hard, grass and carpet. Each surface has a different impact on the bounce of the ball and, more precisely, on its speed. For example, the grass court is the fastest of all the tennis court surfaces, due to its slippery surface. By contrast, clay courts reduces the speed of the ball, making it easier for an opponent to return a hard shot. Therefore, a player is likely to perform differently depending on the surface he is playing on. This second aspect is taken into consideration in the "surface weighting" method, described in Section 2.3.2.

2.3.1 Time discounting

Estimating the current performance of a player isn't an easy thing to do. Indeed, many factors can come into play, such as injuries, fatigue, family problems or joys which impact, in favour or disfavour, the player's mind. However, all these factors are reflected in the player's scores during the season, telling us if he is currently in a good or bad shape. That being said, one could think to only consider the player's statistics of the last season to compute an estimate of his performances. However, doing so would not take into consideration that some great champions sometimes face very bad seasons. This has been the case for Roger Federer in 2013 for example. Ranked number 1 at the ATP ranking at the time, having won 6 titles including Wimbledon, a silver medal at the Olympics, and the record of weeks spent as #1 in 2012, he faced one of his worst season in 2013 by losing 17 matches and winning only one title at the ATP 250 of Halle Open in Germany.

Therefore, a better approach is to consider all the matches of a player during his career, but weight each one them in order to give more importance to the most recent ones. Formally, each past match $i = 1, ..., n$ of a player will be weighted according to the following formula :

$$w_{i,time} = \eta^{\Delta t} \tag{1}$$

where

$$\begin{cases} \eta \in [0,1] & \text{is the discount factor} \\ \Delta t = t_{now} - t_{past} & \text{is the elapsed time between the considered matches (in years)} \end{cases}$$

Note that the parameter η is a hyperparameter of our model and needs to be optimised. The smaller η is, the less significance the older matches will have. After having tested several values of η and analysed the results on the predictions, we chose a discount factor of $\eta = 0.6$. Figure 2 shows the weights given to matches depending on Δt when $\eta = 0.6$. Notice that we also chose to give the same maximal weight to all matches played in the past year of the considered match.

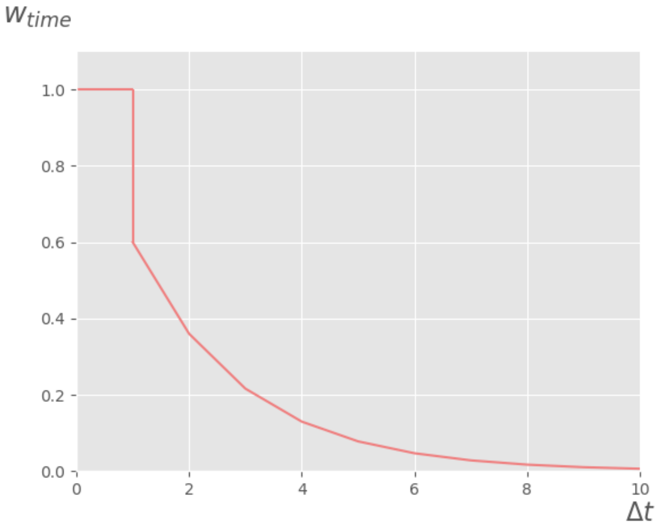


Figure 2: Time discounting function when $\eta = 0.6$

2.3.2 Surface weighting

According to a detailed analysis of Barnett [5], some fundamental relationships do exist between player performance across different court surfaces. For example, he deduces that if a player's optimal surface is grass, he is likely to perform better on a hard court than on a clay one.

As our goal is to predict the outcomes of tennis matches happening on clay, one could decide to only consider the matches played on clay in our dataset. However, this would be problematic for two reasons. First, doing so would considerably decrease the number of data used to train our model, as matches on clay only represent 30% of the matches of our dataset. Second, this decision would affect the accuracy of the estimated player performances as a lot of matches played on other surfaces would not be considered. A smarter approach, as Sipko pointed out in [3], would be to find a correlation coefficient between clay and the other surfaces, in order to attribute a weight to each single match of our dataset, depending on the relation between the match surface and clay. Such coefficients can be estimated using our dataset. First, for each player, we can compute the percentage of matches won across the different surfaces during their career. From that, we can deduce the mean and the standard deviation for each surface, parameters that will be used when computing the correlation coefficients. Finally, for each pair of surfaces (A, B), we can calculate their correlation by applying the following formula:

$$\rho_{A,B} = \frac{cov(A,B)}{\sigma_A \sigma_B} = \frac{1}{N} \frac{\sum_{k=0}^{N} (x_{A,k} - \mu_A)(x_{B,k} - \mu_B)}{\sigma_A \sigma_B} \tag{2}$$

where

$$\begin{cases} N & \text{is the number of considered players} \\ x_{S,k} & \text{is the percentage of matches won by player } k \text{ on surface } S \\ \mu_S & \text{is the mean percentage of matches won on surface } S \\ \sigma_S & \text{is the standard deviation of percentage of matches won on surface } S \end{cases}$$

Computing Equation (2) for all possible pairs of surfaces, using our dataset of more than 70000 matches between 1993 and 2018, yields the coefficients presented in Table 2. We notice that all coefficients are positive, which means that a player that tends to win on clay also tends to win on another surface, but in smaller proportions. In our case, the coefficients of the first column, concerning the correlations between clay and the other surfaces, will be used as the surface weights $w_{i,surface}$ involved in the weighting of each past match i of a player when computing his weighted means on his statistics.

	Clay	**Carpet**	**Grass**	**Hard**
Clay	1	0.257	0.277	0.491
Carpet	0.257	1	0.451	0.565
Grass	0.277	0.451	1	0.574
Hard	0.491	0.565	0.574	1

Table 2: Correlation coefficients between court surfaces

2.3.3 Combining weights

Now, the problem is knowing how to weight each one of our computed weights, namely the timing weight and the surface weight, in a final weight that will be attributed to the matches. Upon reflection, it seems more logical to give more importance to the timing weight than to the surface weight, as what we really want to put forward is the current performance of a player. The surface weighting makes our model more accurate for the specific problem of predicting outcomes of matches on clay, but is clearly less significant than time discounting. After having tested different combinations and analysed the results on the further predictions, described in Section 3, the right compromise seems to give each past match $i = 1, ..., n$ of a player a total weight described by the following equation :

$$W_i = 0.95 * w_{i,time} + 0.05 * w_{i,surface} \tag{3}$$

Then, by looping over each match of our dataset, we compute for both players new statistics representing a weighted average of the players' match statistics, described in Section 2.2, over all their past matches before the time of their meeting. Formally, for a given statistical feature x, we compute the new feature:

$$\overline{x} = \frac{\sum_{i=1}^{n} W_i x_i}{\sum_{i=1}^{n} W_i} \tag{4}$$

where x_i is the value of the feature x in the i^{th} match, and W_i represent the weight given to the i^{th} match enrolled in the average and is given by Equation (3).

This way, by computing these new weighted statistics for each player of each match of our dataset, we create a whole new dataset having the same size of the previous one, but where all the statistical features of each match are replaced by the ones computed with Equation (4).

2.4 Symmetric feature representation

From now on, our dataset could be used to train a machine learning model. However, doing so would result in a model that would consider the characteristics of both players independently of each other. As a consequence, we would have two different values representing the same variable for each statistical feature. The problem with this representation is that the model may assign, due to noise in the data, more importance to a feature of Player A than to the same feature of Player B. This is problematic in the sense that this difference of significance for the same feature could produce a different outcome prediction when the labels of the players are swapped (i.e., if Player A becomes Player B and vice-versa). Obviously, this is not what we expect. We would like our model to be symmetric in its outcome predictions.

A solution to this problem consists in building new features resulting in the difference of the two values representing the same variable. For example, considering the features $Rank_A$ and $Rank_B$ representing respectively the ATP ranking of Player A and Player B, a new feature can be built as the following difference :

$$Rank_{diff} = Rank_A - Rank_B$$

This idea was inspired by Clarke and Dyte [6], who precisely used the rank difference as the sole feature in their logistic regression model. Another benefit of using variable differences is that it halves the number of features, which reduces the variance of the model and thus helps to prevent overfitting.

2.5 Final features

Before diving into the training step, let us make a summary of the final features. These are shown in Table 3. Most of the features do not require a further clarification, as they simply are differences between weighted means over a set of matches of the few statistics defined in Section 2.2. Note that, in addition to them, we also considered the differences in ages of the players, in ATP rankings, in ATP points, as well as four new features that were computed during the generation of the weighted means. The first one expresses the fact that the two players of a match have the same handedness (1) or not (0). The second one represents the difference in the average winning percentage of the players, reflecting the difference in their ability to win most of their matches. The third one is the difference in the average match duration of the players, whereas the last new feature is the difference in the best-of average of the players, reflecting the difference in their trends to play more Grand Slams (best-of 5) or ATP Series and Masters (best-of 3). Note that these three last features were computed as a weighted mean over a set of previous matches, exactly as the other statistics.

Finally, all the features are standardized in order to have them all in the same order of magnitude.

Final Features
Difference in ages
Difference in ATP rankings
Difference in ATP points
Same handedness
Difference in average match duration
Difference in average winning percentage
Difference in best-of average
Difference in average percentage of aces
Difference in average percentage of double faults
Difference in average percentage of service points
Difference in average percentage of first serves
Difference in average percentage of first serves won
Difference in average percentage of second serves won
Difference in average percentage of break points faced
Difference in average percentage of break points saved

Table 3: Final features used to train our model

3 Match prediction

In order to be able to predict the winner of a tournament, we must be able to predict the winners of the progressive matches. To do so, we will train multiple supervised learning algorithms with our new dataset and keep the one that give the most satisfactory results. The supervised machine learning algorithm requires a set of labelled examples for training. In the context of tennis prediction, each training example corresponds to a single historical tennis match played between two players, and is composed of two elements :

- A vector of input features $\mathbf{x}$, composed of our 15 features described in Table 3.
- An output value y, corresponding to the outcome of the match. Considering a match between two players, labelled as Player A and Player B, the output value will be equal to 1 if Player A won, and 0 if Player A lost.

This problem of predicting a winner can thus be seen as a classification problem, in which we attempt to classify a given player of a particular match as either the "winner" or the "loser" of the match.

In order to get the best model as possible, multiple supervised learning techniques were tested, namely the Random Forest algorithm, Logistic Regression, Support Vector Machine (SVM) and Neural Networks. When testing different models, making an analysis reproducible is crucial to get consistent results. There are many causes that can lead to non-determinism, such as shuffling of the dataset, changes in machine learning frameworks or randomness in the machine learning model itself. To combat these, we chose to fix the size of our dataset and sort it before beginning to train, so that if our model uses a fixed range of the dataset, the contents of this set will be consistent across runs. Additionally, we didn't want to predict matches in the past with information from the future, so our testing set was always composed of matches that were played at a later time than matches in the training set. Moreover, we chose to use the `scikit-learn v0.20.1` and not deviate from this version. Our final dataset contains approximately 53100 matches, of which 20% define our testing set (about 10600 matches), and the rest our training set (about 42500 matches).

Finally, to select the best parameters of each model, we used `scikit-learn`'s `RandomizedSearchCV` method. By defining a grid of hyperparameter ranges, it randomly samples from the grid, performing 5-Fold cross-validation with each combination of values. The results given by the tuned algorithms are described in Section 3.2.

3.1 Supervised learning techniques

3.1.1 Random Forest

As a starting point, we applied the Random Forest algorithm on our data. This choice was motivated due to the fact that it is a flexible and easy-to-use machine learning algorithm that is widely used for classification problems. As a reminder, Random Forest is an ensemble learning method mostly used for classification and regression, that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (when talking about classification) or the mean prediction (when talking about regression) of the individual trees. Random Forest has the well-known property of correcting the decision trees' habit of overfitting their training set.

Another advantage of this method is that it allows to compute feature importance, providing us with an overview of the most important features to keep for other models, and conversely those to omit in order to reduce overfitting. Figure 3 shows the importance of each one of our 15 features. Without great surprise, the three most important features are the difference in average winning percentage, the difference in ATP points and the difference in rankings.

In practice, we used the RandomForestClassifier function from `scikit-learn`.

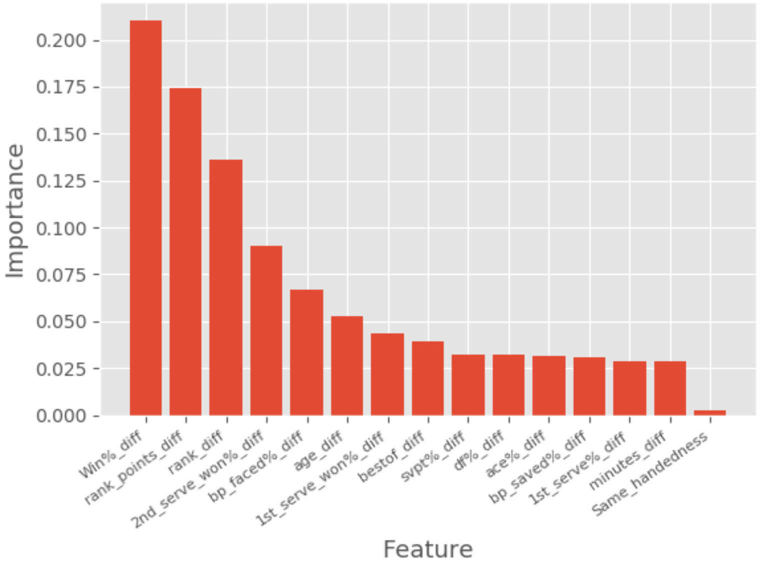


Figure 3: Feature importance according to the Random Forest algorithm

3.1.2 Logistic Regression

Next, we tried to apply a Logistic Regression on our data as it is another popular choice for classification problems. A Logistic Regression model takes as input the vector of features and attributes a weight to each one of them. It then outputs a real value between $-\infty$ and $+\infty$, that is a linear combination of these weighted features, and passes it into the logistic function which, as a reminder, maps real-valued inputs to values between 0 and 1.

Logistic Regression is attractive in the sense that its training is pretty quick compared to other machine learning algorithms, it has a strong resistance to overfitting and furthermore, its outputs can be interpreted as probabilities, which can be seen as match-winning probabilities in our case. It further has been quite popular in the field of tennis matches predictions. For example, Clarke and Dyte [6] used a Logistic Regression model on one single feature, being the difference in ATP ranking points of two players, to predict the outcome of a given set.

For the hyperparameters part, we chose to let the default parameters of the LogisticRegression function from `scikit-learn`, that is L2-regularization with the *liblinear* algorithm.

3.1.3 Support Vector Machine

Moreover, we used a Support Vector Classifier to predict tennis match outcomes, as it appeared to have given pretty good results for some related work [7]. The goal of such a supervised learning technique is to find a linear classifier that maximises the hyperplane margin separating the data into the categories with which they are labelled. Once the optimised classifier found, an unseen example, such as a new tennis match, can then be classified according to which side of the margin it falls on, once mapped to the same space.

In general, SVMs are known to be very efficient classifiers. However, the choice of a good kernel is difficult in practice and critical to achieve good results. Here, we explored multiple options such as linear, RBF (Gaussian) and polynomial kernels using the SVC classifier from `scikit-learn`.

3.1.4 Neural Networks

The motivation behind the use of a Neural Network (NN) was brought by Somboonphokkaphan [8] that used a three-layer feed-forward NN for match prediction with the backpropagation algorithm, and had an average accuracy of about 75% in predicting the outcomes of matches in the 2007 and 2008 Grand Slam tournaments.

Although Neural Networks can lead to very efficient models in practice due to their ability to detect complex relationships between the input features, it also comes with some issues. First, defining the adapted architecture for the NN, that is the number of hidden layers to use in the network as well as the number of neurons to use in each hidden layer, is very complex and requires a trial and error approach that can take a long time. Furthermore, the training of such a model can result in a local minimum and ones need to find a way of avoiding such scenario. Then, Neural Networks usually require much more data than traditional machine learning algorithms, as training on smaller datasets often leads to overfitting. Lastly, converging to an optimal solution can sometimes be a hard thing to do in a reasonable amount of time.

Here, we used the MLPClassifier function with stochastic gradient descent from the `scikit-learn` library. The other parameters were tuned and the combination achieving the best results is reported in Table 4.

3.2 Results

The detailed performance comparison of each machine learning algorithm on test scores can be seen in Table 4. Note that each algorithm was trained by considering all the possible sets of top features, that is by considering only the best feature, then the two best features, and so on until considering all features together. This is finally this last option that gave us the best results for all the tested methods.

At the end, we decided to keep the Random Forest model as this was the one achieving the best prediction outcomes among all the models. Moreover, this classification algorithm also gives some measure of the certainty of an instance belonging to a class, which can be used as the match-winning probability.

Model	Hyperparameters	Considered features	Test accuracy

Random Forest	n_estimators = *2000*
	max_depth = *10*
	min_samples_split = *5*
	min_samples_leaf = *1*
		All 15 features	66.84%

Logistic Regression	solver = *liblinear*
	penalty = *l2*
		All 15 features	66.56%

MLP Classifier	solver = *sgd*
	hidden_layer_sizes = *(20,)*
	activation = *tanh*
	learning_rate = *0.05*
	momentum = *0.6*
		All 15 features	66.47%

SVC	kernel = *rbf*
	kernel = *linear*
	kernel = *poly*
		All 15 features	66.38%
			66.36%
			65.38%

Table 4: Test accuracy comparison between different machine learning algorithms

4 Tournament prediction

From now, we are able to predict the outcome of a given match with an accuracy of a bit less than 70%. This means that, using our match-predictor model on the appropriate players' statistics, we can now predict the outcomes between all possible pairs of players participating in a tournament. Given the fact that a Grand Slam tournament begins with 128 players, the number of different matches that can possibly happen during this type of tournament is obtained by computing the number of combinations without repetition of two players among 128, which results in $C_{128}^{2} = 8128$ possible different matches.

Once the outcomes of all these matches are predicted, that is once we have, for each of the 8128 possible matches, the predicted winning percentage of both players, we can predict the final winner of the entire tournament by generating a large number of possible actual draws, and simulating, for each one of them, the tournament according to our predictions. By doing that, we can then compute a winning probability for each player, as the number of times he has won one of the simulation on the total number of simulations. This method of Monte-Carlo simulations is further detailed in Section 4.2.

4.1 Draw generation

As previously mentioned, our idea is to estimate the probability for each player to win the tournament by making simulations on a large number of generated draws. One could play the simulations by considering that the draws of a Grand Slam tournament are randomly computer-generated. However, a crucial point in the generation of Grand Slam draws would then not be considered : the seeds distribution. To understand this concept, a little explanation about the players selection and distribution in a Grand Slam tournament is necessary.

In a Grand Slam, 128 tennis players are taken into account. Among them, there are:

- The world's best 104 players according to the official ATP ranking, which is stopped by the organisers six weeks before the tournament.
- 16 players that come from the qualifications. The qualifications are a pre-tournament that begins five days before the real tournament and also welcomes 128 players, ranked outside the TOP 104, among which 16 will access the real tournament.
- 8 players that receive a "wild-card", an exclusive invitation to participate in the real tournament without having to pass the qualifications.

From the 128 players, the 32 best players according to the ATP ranking are called "seeds". In a Grand Slam, these seeds enjoy a privilege in terms of their initial position in the draw. Indeed, when generating the draw, the organisers ensure that the seeds do not meet before the third round. The goal of this ruling is to "protect" a minimum the best players in the first rounds of the tournament in order to show the public the most awaited matches later in the tournament. Furthermore, among these 32 seeds, specific meeting rules apply depending on their seed numbers :

- The seeds 1 and 2 can only meet in Final, and they theoretically face the seeds 3 or 4 (according to the draw) in the Semifinals.
- In the Eighth-finals (4th round), the seeds 1 to 4 face one of the seeds 13 to 16, drawn by lot, and the seeds 5 to 8 face one of the seeds 9 to 12, drawn by lot.
- In the round of 32 (3rd round), the seeds 1 to 8 face one of the seeds 25 to 32, drawn by lot, and the seeds 9 to 16 face one of the seeds 17 to 24, drawn by lot.

Therefore, by knowing that, we can significantly reduce the number of possible draws that can come out in practice, and that way, improve the accuracy of our model.

Now, the next step consists in doing Monte-Carlo simulations on a large number of draws in order to estimate the winning probability of each player according to our predictions. This step is described further in the next section.

4.2 Monte-Carlo simulations

As a reminder, Monte-Carlo methods are a subset of computational algorithms that use the process of repeated random sampling to make numerical estimations of unknown parameters. In our case, it is a question of sampling randomly a certain number of possible draws, as described in Section 4.1, and simulating them according to our predictions. More precisely, for each given match of each sampled draw, let us define by α the predicted probability for Player A to win the match against Player B. Then, the resulting outcome of the match will be :

$$\begin{cases} \text{Player A wins with probability } \alpha \\ \text{Player B wins with probability } 1-\alpha \end{cases} \tag{5}$$

By following this scheme, we are able to compute, at each round of a simulation, the different winners that will meet in the next round, according to their repartition in the considered draw. Eventually, each simulation will output a unique final winner. By computing, for each player of the tournament, the number of times that he comes out as a winner from some simulations on the total number of simulations, we can estimate which players will have the highest probabilities to win the tournament. Formally, we have :

$$\text{Player's probability of winning} = \frac{\text{Number of simulations won by player}}{\text{Total number of simulations}}$$

To be even more precise, a similar process can be applied in order to compute the probability for each player to reach a certain round. Indeed, we will have :

$$\text{Player's probability of reaching a round} = \frac{\text{Number of simulations where the player reached the round}}{\text{Total number of simulations}}$$

These precise probabilities will allow to analyse in depth the accuracy of our model when tested on different past tournaments, as it is the case in Section 4.3.

4.3 Model testing

In order to evaluate the accuracy of our model, it was tested on the three last French Opens. Prior to that, to have an idea of the best players on clay nowadays, we computed the winning percentage on clay of the current best ranked players using our dataset. The results are shown in Figure 4. As one can notice, Rafael Nadal, the current number 2 in the world, outperforms all other players on that surface.

The following sections describe the results obtained when applying our tournament-model to the 2018, 2017 and 2016 French Opens. For each tournament, the process is the following : having as inputs the names of the 128 players having participated at the past tournament, the first step consists in computing the stats of each player by applying time discounting and surface weighting, as explained in Section 2.3.1, 2.3.2 and 2.3.3. Then, we generate the 8128 possible matches of that tournament and compute, for each one of them, the final features as described in Section 2.5. After that, using our match-predictor model, we predict the outcomes, in terms of winning probabilities, of the 8128 matches. The last step consists in doing Monte-Carlo simulations over a large number of possible draws (one million), as described in Section 4.2, in order to output the probabilities for each player to win the tournament.

For ease of reading, most graphs are placed in the Appendix.

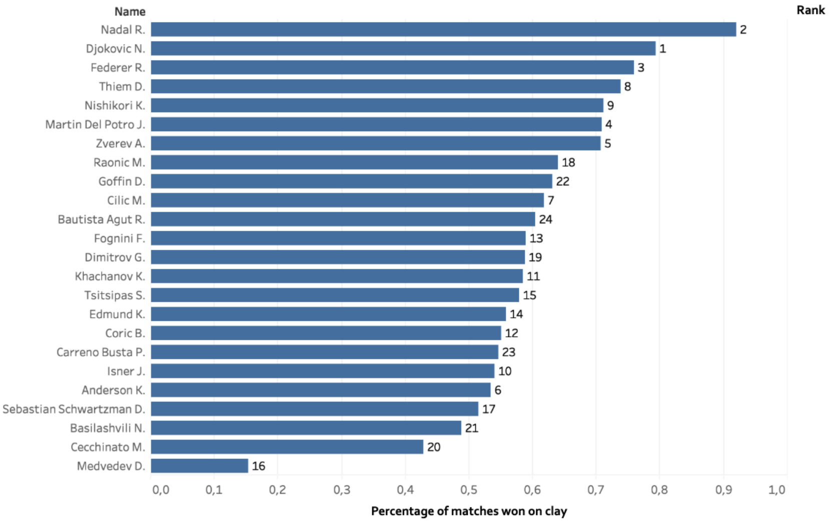


Figure 4: Percentage of matches won on clay for the 24 current best players during their career

4.3.1 Testing on the 2018 French Open

By applying our final model on the 2018 French Open, we obtained very satisfactory results. First, as can be seen in Figure 11 in Appendix A.1, we predict Rafael Nadal, the actual winner of the tournament, as the player who had the highest probability to win the Grand Slam. Moreover, when looking at the actual quarterfinals draw in Figure 10, and comparing it to the eight players that we predicted to have the highest probabilities to reach the quarterfinals (QF), shown in Figure 12c, we notice that six out of the eight players actually did reach the QF. The names of these players are given in Table 5.

Predicted players of the Quarterfinals	**Actual players of the Quarterfinals**
Rafael Nadal (2)	Rafael Nadal (2)
Juan Martin Del Potro (6)	Juan Martin Del Potro (6)
Alexander Zverev (3)	Alexander Zverev (3)
Marin Cilic (5)	Marin Cilic (5)
Novak Djokovic (18)	Novak Djokovic (18)
Dominic Thiem (8)	Dominic Thiem (8)
Grigor Dimitrov (4)	Diego Schwartzman (15)
Kevin Anderson (7)	Marco Cecchinato (72)

Table 5: Comparison between our predictions and reality about the players reaching the Quarterfinals of the 2018 French Open. The number in parenthesis corresponds to the player rank.

An interesting point to notice is the presence of the player Peter Gojowczyk, ranked 49th at the time, in the top 10 players having the highest probabilities to win the tournament according to our predictions. In general, we are not used to see such a "low-ranked" player win a tournament as significant as the French Open. However, when analysing a bit more in depth the recent performances of that player, we noticed that this prediction totally makes sense in practice. Indeed, in 2017, Gojowczyk was ranked 141st. He then made an incredible season that made him go from 141st to 49th in only one year. Unfortunately for him, during his first match of the French Open in May 2018, he retired due to a hip injury. This misfortune factor is obviously something that is difficult to control and can sometimes distort a lot the predictions, even with the best prediction model in the world in hand. But who knows how far Gojowczyk could have gone without this injury ?

4.3.2 Testing on the 2017 French Open

In order to ensure that our model is really conclusive and that our predictions on the 2018 French Open weren't just a good coincidence, we also applied it to the 2017 French Open. Figure 14 in Appendix A.2 shows the winning probabilities of the 10 most-likely-to-win players. As can be seen, we predicted Rafael Nadal, once again actual winner of the tournament, as the player with the second highest probability to win the Grand Slam. Before him, Novak Djokovic, eliminated in quarterfinals by the Austrian Dominic Thiem, had 2% more chance of winning than Nadal, according to our model. However, this difference in winning probability between the two best predicted players is quite small compared to the one of 2018. Indeed, even though Novak Djokovic has here 2% more chance of winning the tournament than Rafael Nadal, this difference in winning probability between Nadal and Djokovic amounted to more than 10% in 2018, which dips the gap more clearly here.

Then, by analysing the quarterfinals draw, shown in Figure 13, and comparing it with our predicted players having the most chances to reach the QF, five out of eight predicted players actually reached the QF. The names of these players are given in Table 6. Notice that the two next players that had the biggest probabilities to win were Dominic Thiem and Marin Cilic, who both reached the quarterfinals.

Predicted players of the Quarterfinals	**Actual players of the Quarterfinals**
Novak Djokovic (2)	Novak Djokovic (2)
Rafael Nadal (4)	Rafael Nadal (4)
Andy Murray (1)	Andy Murray (1)
Stanislas Wawrinka (3)	Stanislas Wawrinka (3)
Key Nishikori (9)	Key Nishikori (9)
Milos Raonic (3)	Marin Cilic (8)
Jo-Wilfried Tsonga (13)	Dominic Thiem (7)
Nick Kyrgios (19)	Pablo Carreno-Busta (21)

Table 6: Comparison between our predictions and reality about the players reaching the Quarterfinals of the 2017 French Open. The number in parenthesis corresponds to the player rank.

4.3.3 Testing on the 2016 French Open

Finally, we made one last test on the 2016 French Open. As can be observed in Figure 17 in Appendix A.3, we predicted Novak Djokovic, the actual winner of the Grand Slam, as the player with the highest probability to win, outperforming the others with a difference in winning probability of about 8% on the second player with highest probability, Rafael Nadal. Moreover, what is interesting to notice in Figure 18a is that, even though Nadal has the second highest probability to win the tournament, the two players that have the highest probabilities to reach the final are Novak Djokovic, obviously, and Andy Murray, which actually represents the real final. This means that, on all the simulation runs, Murray reached the final more often than Nadal did, though the latter won the final a greater number of times than Murray.

Also note that Rafael Nadal, who appears in the top 3 of all our predicted probabilities in Figure 18, retired during the third round due to a wrist injury. This abandonment obviously affects the progress of the tournament a lot, as it actually concerns the best tennis player on clay, winner of 11 different French Opens. That is, once again, difficult to consider in our model and can as a consequence potentially distort the results of our predictions.

5 Predicting the 2019 French Open's winner

Now comes the time to finally predict the winner of the 2019 French Open. To do so, two intermediary steps are however required. First, one needs to determine the 128 players that will participate in the future tournament. Second, recent data about tennis matches of the 2019 season are required to compute the weighted statistics of these players. These two steps are described in details in the following sections.

5.1 Players selection

When testing our tournament-model in Section 4.3, we always made the hypothesis that we knew the 128 players participating in the Grand Slam. However, in practice, predicting which players will be selected to play the incoming tournament is not as easy as it seems, as many factors come into place. The first difficulty concerns the wild-cards, eight special invitations offered to eight players currently ranked outside the TOP 104, and allowing them to access the Grand Slam without playing qualifying matches. Talking about these, it represents the second difficulty : predicting the sixteen players who will pass the qualifications. As explained in Section 4.1, the qualifications are a pre-tournament that begins five days before the real tournament and also welcomes 128 players, ranked outside the TOP 104. The sixteen players who will win the final round of the qualifications, that is the round of 32, will access the final tournament. Lastly, the 104 remaining players are the easiest to consider, as they simply are the TOP 104 players according to the ATP ranking at a specific date. When finally having our 128 players, one needs to consider the ones that could possibly not be able to play the French Open due to an injury. All these considerations are developed in the following sections.

5.1.1 Wild-cards

Attribution process For each Grand Slam, eight so-called "wild-cards", representing special invitations for accessing the final draw without having to pass the qualifications, are given to eight nominated players whose ranking is too low to directly enter the final draw. To select the lucky ones, many factors come into consideration such as the results of the player, his sporting success, his current form, his performances on the tournament surface, etc. However, young players are often advantaged as there exists a desire to give a boost to new talents who could reveal themselves during the tournament. Furthermore, when a well-known champion has tumbled in the ATP ranking for some reason, we can almost be sure that he will be granted a wild-card to avoid making him go through the perilous qualifications. Finally, there is a national preference, each Grand Slam promotes his players.

Note that these invitations are also the subject of exchanges between national federations. In the particular case of the French Open, the French Tennis Federation (FTF) grants each year a wild card to an Australian player and to an American player. In return, the Australian Tennis Federation and the American Tennis Federation respectively gives each year a wild-card to a French player at the Australian Open and the US Open. The players are selected by their home Tennis Federation.

Since this year, a new process set up by the FTF will allow two French players to receive wild-cards for the French Open. The first card will be offered to the French player who has won the most ATP points after a "Race", comprising ten specific French tournaments, if this player is not already selected for the final draw. Otherwise, the wild-card will be given to the first player in the "Race" ranking who can not access it. The objective of this new process, according to Pierre Cherret, DTN of the French Tennis Federation, "is to encourage French players to participate in national tournaments, to promote them, [...] and to help young French players to progress on the circuit" [14]. The second invitation will be awarded to the highest ranked French player of the ATP ranking Race to London, stopped on May 13. The ATP Race is another ranking established by the ATP which, unlike the technical ATP ranking which is updated every week by taking into account the points earned over the previous 52 weeks, only takes into account the points earned during the current season.

Estimations Let us first determine which Australian and American players will be granted a wild-card this year. On May 2, the American player Tommy Paul, 21-year-old and ranked 143rd, earned his first berth in the main draw at the French Open by winning the U.S. Tennis Association's wild-card challenge, a small circuit of three tournaments on clay [16]. Concerning the Australian player, the Australian Tennis Federation has released no information about its choice. We will then make the simple hypothesis that the player that is the most likely to be chosen is probably the best ranked Australian player outside the TOP 104. This concerns the Australian Alexei Popyrin, 19 years old and currently ranked 120th.

Then, the French player who is the most likely to take the first wild-card offered by the FTF is Gregoire Barrere, as he is currently 100 points ahead of the second best player in the "Race". The second French invitation, attributed according to the ATP ranking Race to London, will probably be given to Corentin Moutet, 19 years old, the highest ranked French player of the ATP ranking Race to London, when omitting the already selected French players, with more than 40 points ahead of the next French player in the ranking.

With some exceptions, the remaining wild-cards are most of the time attributed to French players. By considering that the FTF gives a lot of importance on very young players demonstrating good performances, the players who will receive these cards are probably among the twelve ones presented in Table 10 of Appendix B.1.1. In particular, the first four players seem to correspond to the ranking-age ratio to which the FTF values so much. Note that these predictions are personal estimations from the authors resulting from the analysis of the previous selected players on past French Opens. The ideal solution would be to train a machine learning model using data about the players who have received a wild-card in the past French Opens, in order to predict which players will be nominated this year. However, the amount of training samples would be extremely limited, as there have only been 46 French Open since the creation of the Association of Tennis Professionals (ATP) in 1972, and very limited data was available before the 2000's. For these reasons, intuition seems to be the only viable option.

To summarise, Table 7 shows the eight players that we estimate to receive a wild-card for this edition of the French Open.

Players	**Nationality**
Gregoire Barrere	FR
Corentin Moutet	FR
Antoine Hoang	FR
Quentin Halys	FR
Maxime Janvier	FR
Elliot Benchetrit	FR
Alexei Popyrin	AUS
Tommy Paul	US

Table 7: Players prone to receive a wild-card for the 2019 French Open according to our estimations.

5.1.2 Qualifications

For most people, the French Open starts the last Sunday of May. However, five days before begin the qualifications, a tournament in the tournament where 128 players ranked outside the TOP 104 compete in order to grab one of the sixteen places allowing to access the final tournament. As in the Grand Slam, there are also "seeds" in the qualifications, where the same distribution applies when creating the draw.

Our goal is thus to predict the sixteen players coming from the qualifications that are the most likely to participate in the final tournament. To do so, we can use our tournament-predictor model, described in Section 4, on the 128 best players outside the TOP 104. We deliberately choose not to exclude the players that we estimate to receive a wild-card, shown in Table 7, if they are part of the TOP 104-232. This way, we ensure to take them into account if our estimations on the wild-card attributions turn out to be wrong. After applying our model on the considered players, we were able to predict the sixteen players having the highest probabilities to pass the qualifications. These players are shown in Figure 5. One can notice that Alexei Popyrin stands among them, which strengthens our choice to consider him, whether through the attribution of a wild-card or through the qualifications. Hence, we will take into account the next player having the highest probability to reach the final draw, namely Marcos Baghdatis.

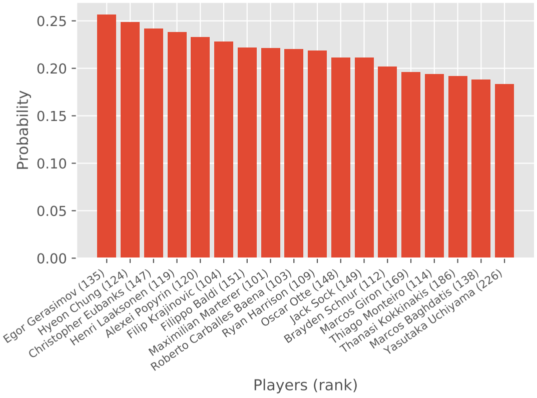


Figure 5: Top 18 players with highest predicted probabilities to pass the 2019 Qualifications.

5.1.3 TOP 104 players

Six weeks before each French Open, the organisers freeze the ATP ranking and the TOP 104 players are guaranteed to have their place in the final draw. Note that the ranking may change after this date, but this will not be taken into account for the player selection. Therefore, the 2019 French Open starting on Sunday, May 26, the first 104 selected players will be the best 104 players according to the ATP ranking of the week of April 14. The list of participating players is publicly available on the official website of the French Open [11]. Notice that in this list, some players denoted by an asterisk, have used their protected ranking, which is a favour granted on a case by case basis by the ATP, allowing a player to benefit, during a transitional period, from his former technical ranking upon his return from a long period of absence due to injury. The player can thus enter the draw directly without going through the qualifications thanks to this status. This is for example the case of Jozef Kovalik, currently ranked 121st, but benefiting for this tournament from his ranking before interruption due to injury. In addition to him, three other players benefit from this privilege this year.

5.1.4 Injured players

Among the 128 final players, some might not play the tournament, because of injuries that could happen during the tournaments preceding the French Open. From early April to the end of May, there are a total of 12 different ATP tournaments, all happening on clay, including the Monte-Carlo Masters, the Barcelona Open, the Madrid Open and the Italian Open. These are four ATP Masters 1000 attracting a lot of good players due to the high prize money involved. Therefore, we must consider the fact that some of the selected players may suffer from an injury in the meantime. The website `tennisexplorer.com` [10] helps doing so by giving a complete list of all the tennis players who recently withdraw because of an injury. The website also gives another list of the players who returned from an injury by starting competition again. Thanks to these lists, we were able to identify players that recently suffered from an injury and didn't play yet since there. By not reappearing in recent competitions, it could mean that these players could possibly not be able to play the incoming French Open. Among the 128 considered players, three of theme are currently concerned by this scenario, namely Gilles Simon (26th), Pablo Carreno Busta (27th) and Matthew Ebden (53rd).

If some of these players were to withdraw before the first round of the tournament, their substitutes would be picked among the "Lucky losers". This is the status given to a player coming from the qualifications who lost just before accessing the final draw. Among the 16 losers, the top eight players must stay in Paris and a draw is made to determine the order in which they would have to replace the forfeit player. Hence, determining which players might replace the injured ones comes down to determining which players will lose the final round of the qualifications. This can be retrieved from our predicted results concerning the qualifications, discussed in Section 5.1.2. Indeed, we only have to consider the next sixteen players having the highest probabilities to win the final round of the qualifications and, among them, the eight best ranked ones will represent the possible substitutes. The names of these predicted substitutes are given in Table 8, and were retrieved from Figure 6.

At this writing, there are three weeks left until the beginning of the tournament, which makes it completely possible for a player to recover from a minor injury. Therefore, we will consider two scenarios : one where all the selected participants are able to play the tournament, and the other where the currently injured players are replaced by their predicted substitutes.

Substitute players
Denis Istomin (105)
Marcel Granollers (107)
Tennys Sandgren (111)
Sergiy Stakhovsky (115)
Matthias Bachinger (127)
Bjorn Fratangelo (132)
Marco Trungelliti (139)
Stefano Travaglia (154)

Table 8: Predicted substitute players. The number in parenthesis corresponds to the player rank.

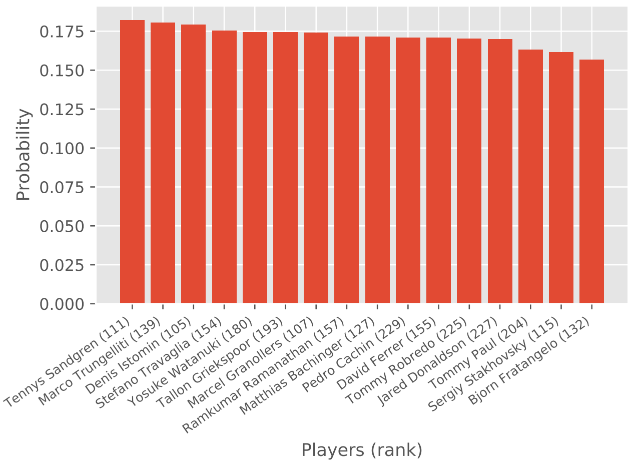


Figure 6: Those players represent the predicted losers of the final round of the Qualifications of the 2019 French Open. The number in parenthesis corresponds to the player rank.

5.2 Data collection

To compute recent statistics for the 128 players of this edition, we had to collect new data about recent tennis matches. Indeed, the Jeff Sackman Dataset [4] that we used until here only lists matches from 1993 to 2018. It neither contained the end of year matches in 2018 nor the 2019 matches. Thus, we scraped the official ATP website using the Python scripts from [9] with some modifications to fit our personal needs, and were able to complete our dataset with the most recent tennis matches.

5.3 Predictions

5.3.1 Predicted winner

Now, after considering the 128 players that are the most likely to play this edition of the French Open, we can finally use our tournament-predictor model on these particular players to predict the winner of the 2019 French Open. And the predicted winner is... Rafael Nadal ! He would then pick up his 12th title for the Grand Slam in Paris. However, this year, competition seems fierce. Indeed, Novak Djokovic, the current best player in the world, has the second highest winning probability by just 0.0013% beneath Nadal's one, as shown in Figure 7. When analysing their probabilities to progress in the tournament, shown in Figure 20 of Appendix B.2, we would be tempted to say that the most likely Final will probably concerns these two players.

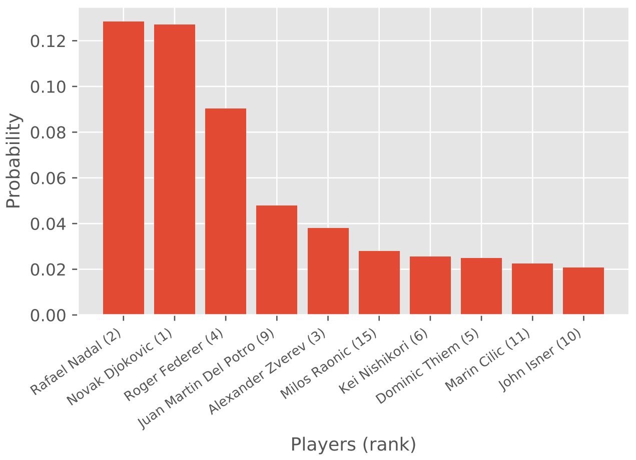


Figure 7: Top 10 players having the highest predicted probabilities to win the 2019 French Open. The number in parenthesis corresponds to the player rank.

5.3.2 Predicted Quarterfinals

By analysing our predicted results for the different rounds of the tournament and considering the distribution of the seeds in a Grand Slam tournament, we can create a possible draw from the Quarterfinals.

The Semifinals would concern Novak Djokovic and Rafael Nadal facing either Roger Federer or Juan Martin Del Potro, depending on the draw, as shown in Figure 20b of Appendix B.2. Note that Alexander Zverev stands very close to Juan Martin Del Potro in terms of probability to reach the Semifinals.

Concerning the Quarterfinals, the eight players that have the most chances to reach them according to our predictions are N. Djokovic, R. Nadal, R. Federer, J.M. Del Potro, A. Zverev, K. Nishikori, D. Thiem and M. Cilic, as shown in Figure 20c of Appendix B.2. Hence, a predicted draw from the Quarterfinals is presented in Figure 8.

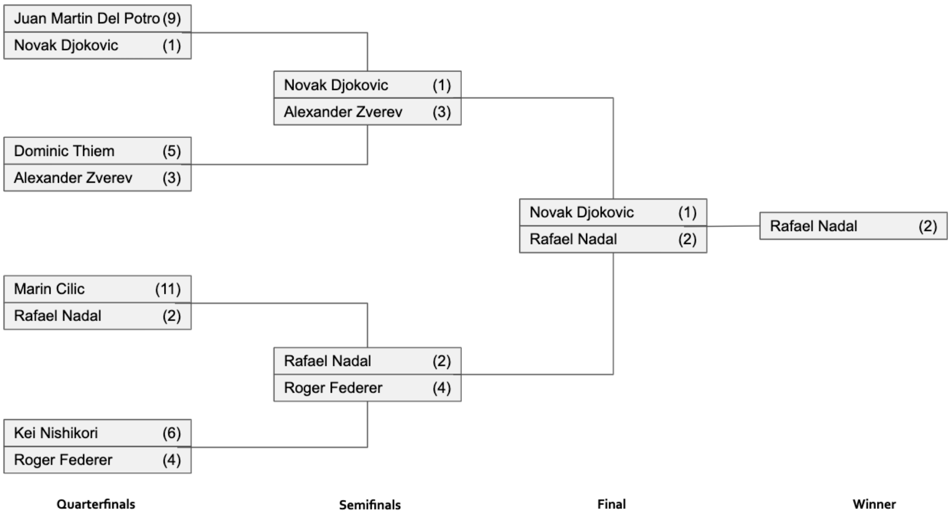


Figure 8: Predicted draw from the Quarterfinals of the 2019 French Open. The number in parenthesis corresponds to the player rank.

5.3.3 Successful outsiders

The outsiders refer to the players thought to have little chance of success in the tournament. Here, we can cite some names that our model predicts to be quite successful. We will consider the outsiders that have the best chances to reach Eight-finals according to our predictions. Table 9 shows some of these successful outsiders. We can find the ex-champion Tomas Berdych, once ranked 4th, but also more surprising players coming from the qualifications such as Egor Gerasimov, the player that had the highest probability to pass the qualifications when computing our simulations in Section 5.1.2, but also Filippo Baldi and Marcos Giron. One very young player, Miomir Kecmanovic, 19 years old, seems to have a good potential too.

Outsiders	Rank	Age	Predicted probability
Tomas Berdych	98	33	20.8% (top 13)
Juan Ignacio Londero	79	25	19.6% (top 15)
Daniel Evans	89	28	18.2% (top 17)
Miomir Kecmanovic	91	19	16.9% (top 19)
Egor Gerasimov	135	26	16.5% (top 22)
Lorenzo Sonego	96	23	15.9% (top 24)
Filippo Baldi	151	23	15.6% (top 26)
Marcos Giron	169	25	15.5% (top 27)

Table 9: Outsiders having the highest probabilities to reach the Eight-finals of the 2019 French Open.

5.3.4 Favourites

Three weeks before the start of the Grand Slam, a lot of sports journalists begin to give their opinions on the favourites of this edition. Reviewing these opinions can give us a hint on the accuracy of our predictions. According to many sports experts, the two big favourites are Rafael Nadal and Novak Djokovic. Amarjeet Nayak, analyst at Sportskeeda, affirms that "at the moment, it is Rafael Nadal who is the clear favorite to win a record-extending twelfth French Open. [...] However, if there is one player who can seriously challenge him at the French Open, it has to be Novak." [17]. Arnaud di Pasquale, sports consultant at Eurosport, reminds the threat that the Austrian player Dominic Thiem could represent for Nadal. However, he claims that "if Nadal regains his form, he is unbeatable, and no matter the level of Thiem." [18]. In brief, many analysts agree to say that the top five 2019 French Open favourites are Rafael Nadal, Novak Djokovic, Dominic Thiem, Alexander Zverev and Roger Federer [19]. Note that these favourites perfectly match the players that we predict to go quite far in the tournament, which is a good point for proving that our predictions made sense.

Another way of confirming our predictions is to look at the odds for the current top contenders. The website `sportsbettingdime.com` allows to track the evolution of the player's odds to win the men's singles at the 2019 French Open. Figure 9 shows the evolution of these odds for the top four contenders. It can be seen that Nadal has the lowest odd (+100), which means that he is the player that most people think will win the French Open. He is then followed by Djokovic (+200), Thiem (+700) and finally Zverev (+1500). Once again, this strengthens our motivation to believe that our predictions are quite accurate, as Rafael Nadal is our predicted winner, but is followed closely by Novak Djokovic.

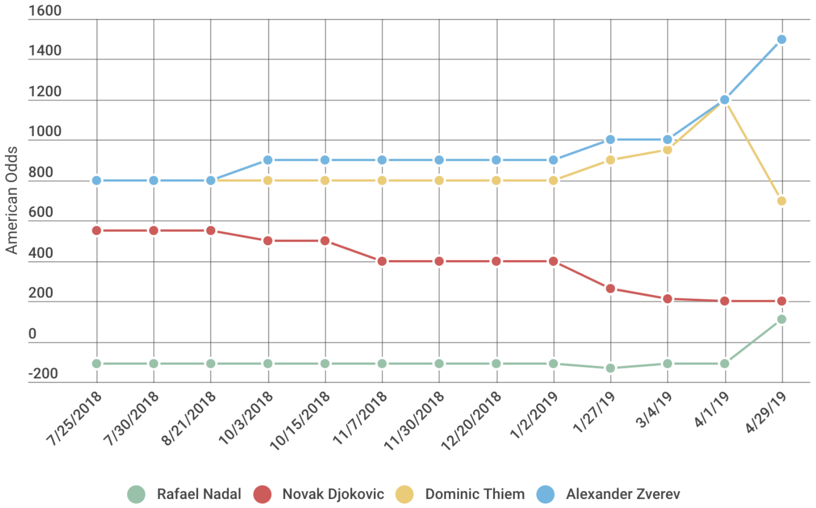


Figure 9: Evolution of French Open odds for the top four contenders according to a number of top online sportsbooks.

6 Conclusion

6.1 Innovation

Extensive research has been conducted on the prediction of tennis matches. However, no work, to the best of our knowledge, has been made about predicting the winner of an entire tennis tournament. We thus have developed a novel method to predict the winner of a Grand Slam tournament, in particular the French Open happening in late May every year in Paris.

First, we developed a method to extract tennis statistics features from raw historical data. Indeed, we were able to compute average player performances by weighting historical tennis matches with surface correlation coefficients and time discounting coefficients. Then, we explored the application of machine learning methods to predict the outcomes of tennis matches. Finally, we implemented a method that is able to give an accurate probability of the winning percentage of each player participating in the tournament, by running Monte-Carlo simulations on real possible draws.

With our tournament-predictor model, we were able to predict correctly six out of the eight players that reached the Quarterfinals of the 2018 French Open, as well as the winner of that year. We also tested it on the 2017 French Open where we were able to predict five out of the eight players that reached the Quarterfinals, and on the 2016 French Open where we predicted the actual Final as well as the winner of that edition. Concerning the tournament of this year, we compared our predictions to the opinions of sports experts and analysed betting odds to give us an idea of their accuracy. They turned out to be pretty consistent with what is expected in this edition.

6.2 Difficulties

The main difficulty of this project was without a doubt the preprocessing of the data. We initially started by feeding our machine learning algorithms with simple data features and achieved poor results. We then had to think how to construct representative data reflecting player performances by taking into consideration the fact that these performances evolve with time and depends on the surface for some players. The difficulty was thus to transform our initial dataset [4] into representative data to help us gain in accuracy.

Another point that involved some deep reflection was the selection of the 128 players for the 2019 French Open edition. We first had to understand the whole selection process, and then find a way to estimate which players to take into consideration for both the qualifications and the wild-cards.

6.3 Future work

Additional Features In addition to the features we constructed, professional bettors suggest other factors that could also play a role in the outcome of a tennis match. Most of them are psychological factors, such as the player motivation or the home bias. Others concern the conditions of a match, such as the time it is played at or the weather conditions (wind, temperature) that may favour a particular playing style.

Other ML Algorithms Here, we focused our efforts on four common machine learning algorithms : Random Forest, Logistic Regression, Support Vector Machine and Multilayer Perceptron. More sophisticated approaches may produce better results. In particular Deep Neural Networks that are currently trending because of their ability to outperform a lot of well-known machine learning techniques.

Women's Tennis In this project, we limited the scope of our investigation to ATP matches, for the single reason that we knew men tennis players better. Nonetheless, all our code is generic enough to accommodate predictions for WTA matches. The adaptation task would be quite easy as Jeff Sackmann [4] also provides a dataset about WTA matches in his Github repository, containing approximately the same features as for men.

Tournament Generalisation This entire project was dedicated to predicting the winner of one particular tournament : the French Open. However, our model was implemented such that it could easily be extended to predict the winner of any tennis tournament. The only factor that has to be modified is the surface coefficient involved in the weighted averages, that would be adapted according to the corresponding tournament's surface (see Table 2).

References

[1] Boudway I. (2008, June 12). Soccer Is the World's Most Popular Sport and Still Growing. Retrieved from https://www.bloomberg.com/news/articles/2018-06-12/soccer-is-the-world-s-most-popular-sport-and-still-growing.

[2] Barnett T. and Clarke S.R. (2005). Combining player statistics to predict outcomes of tennis matches. *IMA Journal of Management Mathematics*, *16(2)*, 113-120.

[3] Sipko M. (2015, June). Machine learning for the prediction of professional tennis matches. *MEng computing - Final Year Project, Imperial College London.*

[4] Sackmann J. (2015-2018). tennis_atp, *GitHub repository.* Retrieved from https://github.com/JeffSackmann/tennis_atp.

[5] Barnett T. and Pollard G. (2007). How the tennis court surface affects player performance and injuries. *Medicine Science Tennis*, 34–37.

[6] Clarke S.R. and Dyte D. (2000). Using official ratings to simulate major tennis tournaments. *International Transactions in Operational Research*, 585–594.

[7] Chen Y., Tian Y. and Zhong Y. (2017). Real Time Tennis Match Prediction Using Machine Learning. *Stanford CS 229 - Final Project.*

[8] Somboonphokkaphan A., Phimoltares S. and Lursinsap C. (2009). Tennis Winner Prediction based on Time-Series History with Neural Modeling. *IMECS 2009: International Multi-Conference of Engineers and Computer Scientists*, *Vols I and II*, I:127–132.

[9] Lin K. (2017). atp-world-tour-tennis-data, *GitHub repository.* Retrieved from https://github.com/serve-and-volley/atp-world-tour-tennis-data.

[10] Recently injured players. Retrieved from https://www.tennisexplorer.com/list-players/injured/.

[11] Roland-Garros 2019: The list of participants. (2019, April 17). Retrieved from https://www.rolandgarros.com/en-us/article/roland-garros-2019-list-participants-singles-men-women.

[12] Tennis : mode d'emploi des qualifications de Roland-Garros. (2011, May 18). Retrieved from https://www.lemonde.fr/sport/article/2011/05/18/tennis-mode-d-emploi-des-qualifications-de-roland-garros_1523432_3242.html.

[13] Baheux R. and Nouaux M. (2013, May 14). Tennis: Comment attribue-t-on les wild cards de Roland-Garros? Retrieved from https://www.20minutes.fr/sport/1155091-20130514-20130514-tennis-comment-attribue-t-on-wild-cards-roland-garros.

[14] Roland-Garros : nouveau processus d'attribution des wild-cards pour les Français. (2019, January 21). Retrived from http://www.fft.fr/actualites/fil-d-infos/roland-garros-nouveau-processus-dattribution-des-wild-cards-pour-les-francais.

[15] Ghazouani-Durand J. (2019, April 9). Roland Garros: Barrère en tête pour la wild-card. Retrieved from https://www.welovetennis.fr/roland-garros/145803-barrere-en-tete-pour-la-wild-card.

[16] The Associated Press. (2019, May 2). Tommy Paul Earns USTA's Wild-Card Entry for French Open. Retrieved from https://www.nytimes.com/aponline/2019/05/02/sports/tennis/ap-ten-french-open-usta-wild-card.html.

[17] Amarjeet Nayak. (2019, January 28). Djokovic or Nadal: Who is the favourite at the 2019 French Open? Retrieved from https://www.sportskeeda.com/tennis/djokovic-or-nadal-who-is-the-favourite-at-the-2019-french-open.

[18] DIP IMPACT. (2019, April 30). VIDEO - Le favori à Roland-Garros ? "Ça va dépendre de Nadal et non pas de Thiem". Retrieved from https://video.eurosport.fr/tennis/roland-garros/2019/video-le-favori-a-roland-garros-ca-va-dependre-de-nadal-et-non-pas-de-thiem$_vid1191717/video.shtml$.

[19] Harry Floyd. (2019, April). Top Five 2019 French Open Favorites and Predictions for Men's Singles. Retrieved from https://lobandsmash.com/2019/04/02/top-five-french-open-favorites/.

Appendices

A Model testing

A.1 Testing on the 2018 French Open

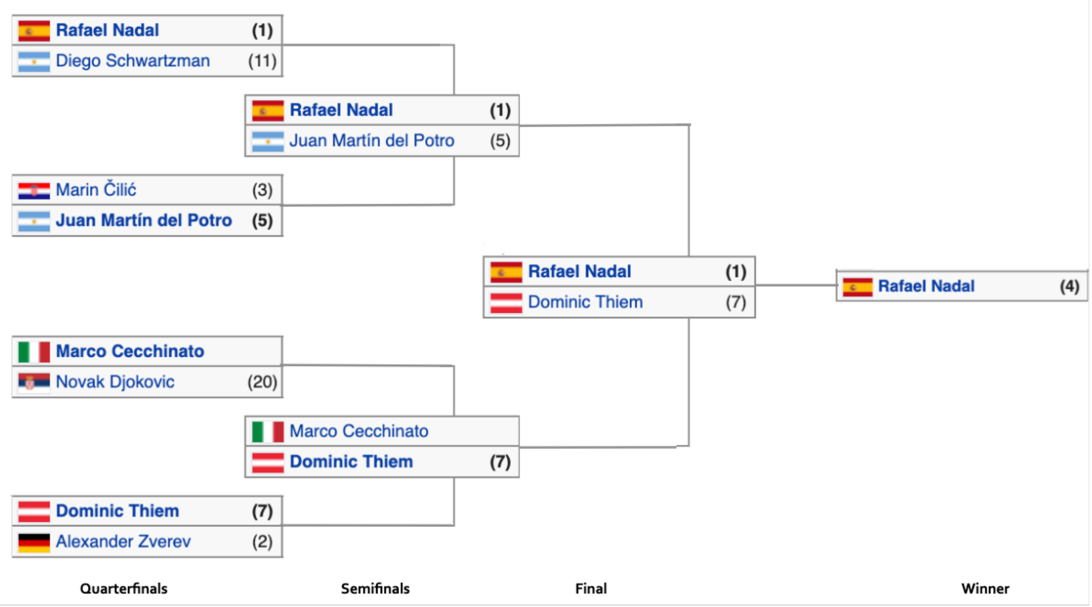


Figure 10: Draw from the Quarterfinals of the 2018 French Open. The number in parenthesis corresponds to the seed number.

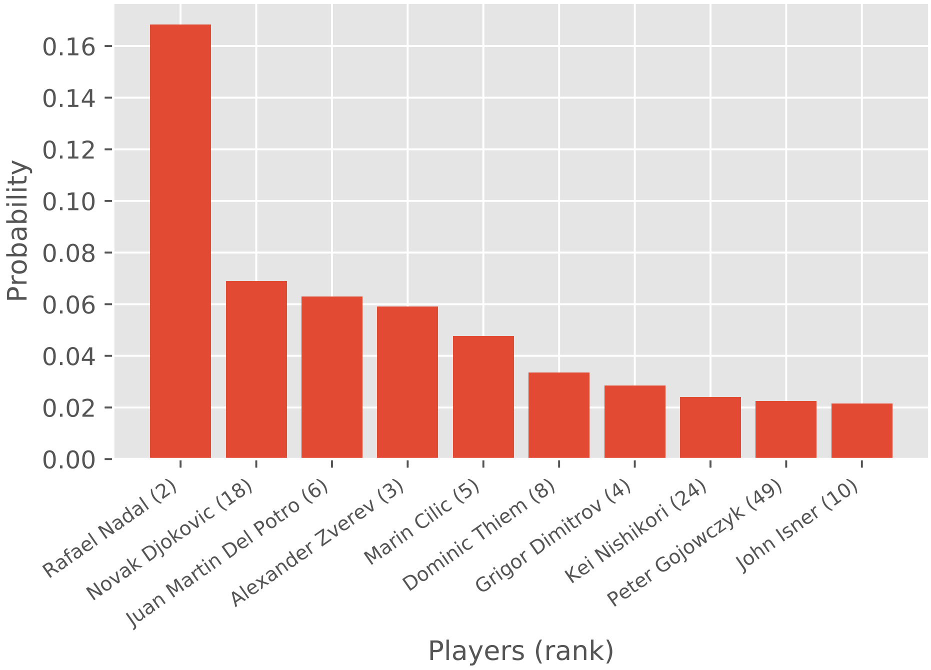
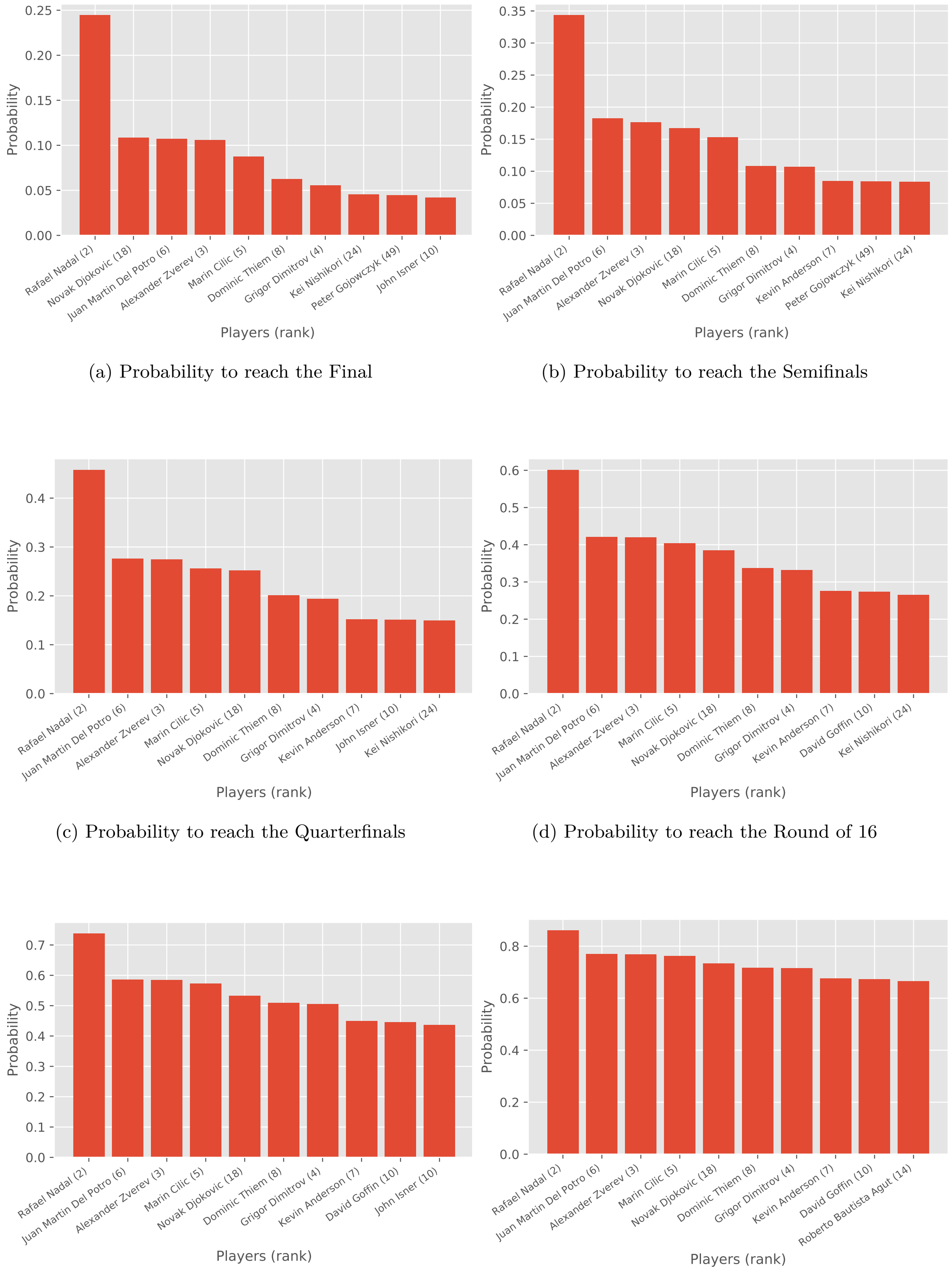


Figure 11: Top 10 players having the highest predicted probabilities to win the 2018 French Open. The number in parenthesis corresponds to the player rank.



(a) Probability to reach the Final

(b) Probability to reach the Semifinals

(c) Probability to reach the Quarterfinals

(d) Probability to reach the Round of 16

(e) Probability to reach the Round of 32

(f) Probability to reach the Round of 64

Figure 12: Top 10 players having the highest predicted probabilities to reach each round of the 2018 French Open. The number in parenthesis corresponds to the player rank.

A.2 Testing on the 2017 French Open

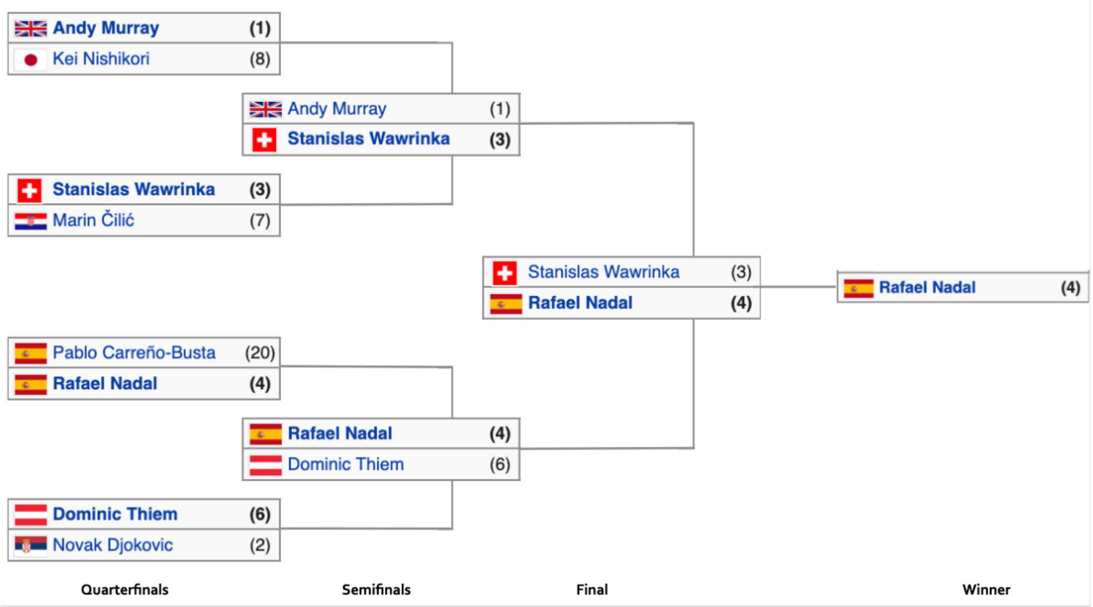


Figure 13: Draw from the Quarterfinals of the 2017 French Open. The number in parenthesis corresponds to the seed number.

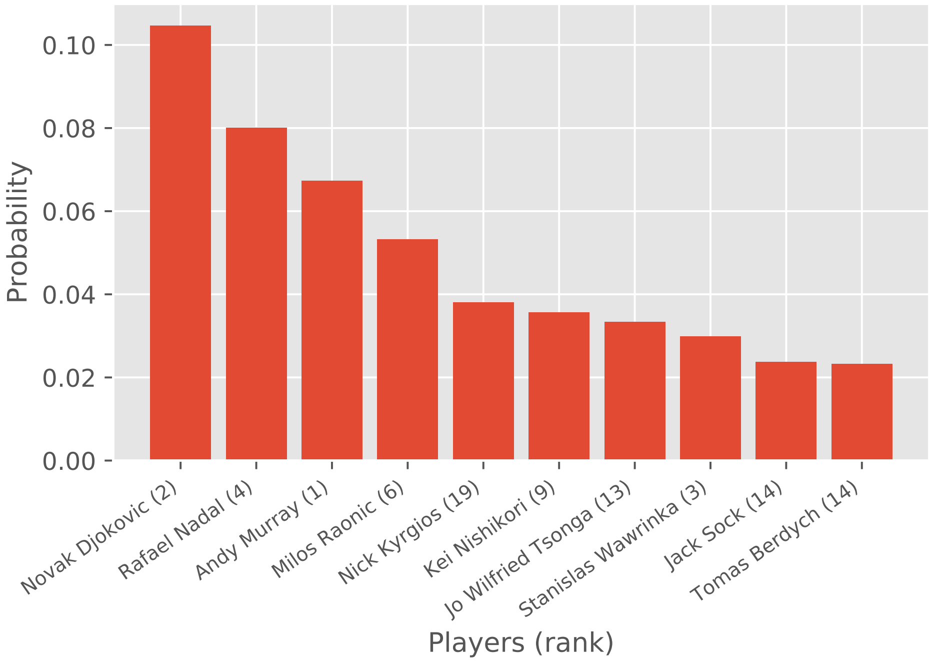
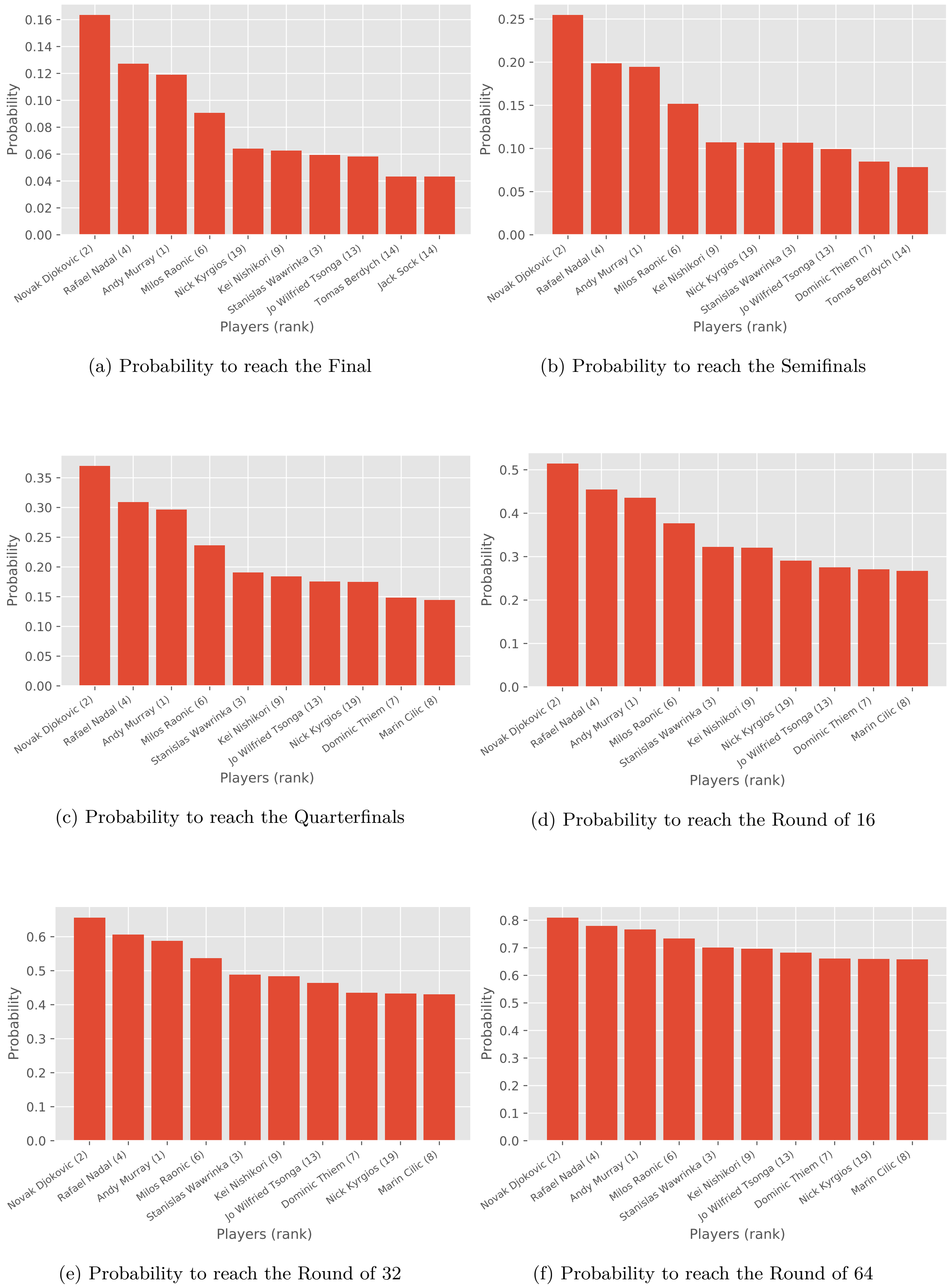


Figure 14: Top 10 players having the highest predicted probabilities to win the 2017 French Open. The number in parenthesis corresponds to the player rank.



(a) Probability to reach the Final

(b) Probability to reach the Semifinals

(c) Probability to reach the Quarterfinals

(d) Probability to reach the Round of 16

(e) Probability to reach the Round of 32

(f) Probability to reach the Round of 64

Figure 15: Top 10 players having the highest predicted probabilities to reach each round of the 2017 French Open. The number in parenthesis corresponds to the player rank.

A.3 Testing on the 2016 French Open

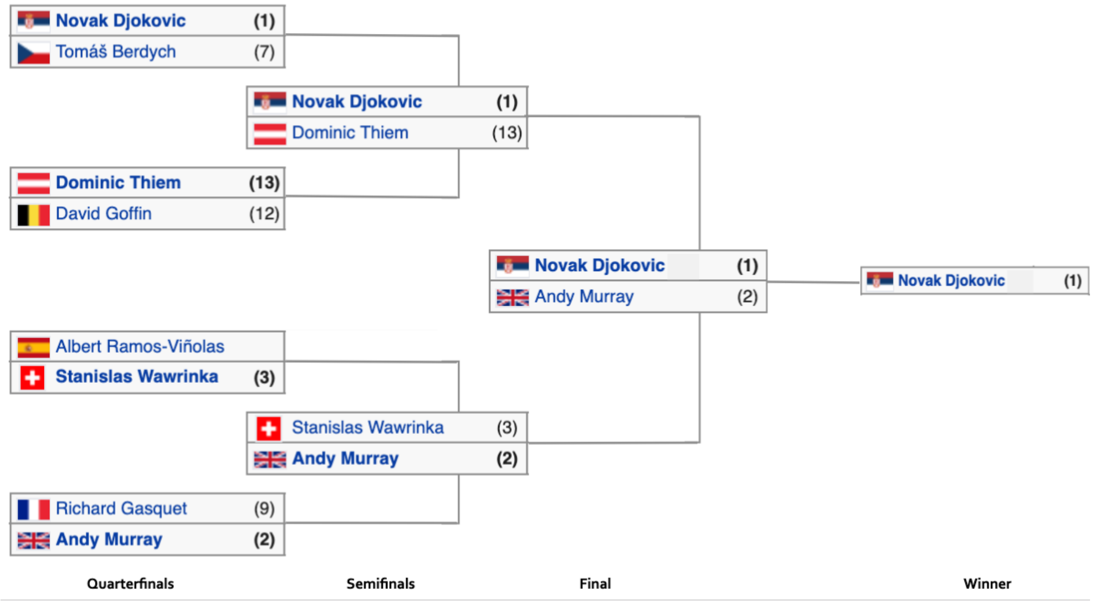


Figure 16: Draw from the Quarterfinals of the 2016 French Open. The number in parenthesis corresponds to the seed number.

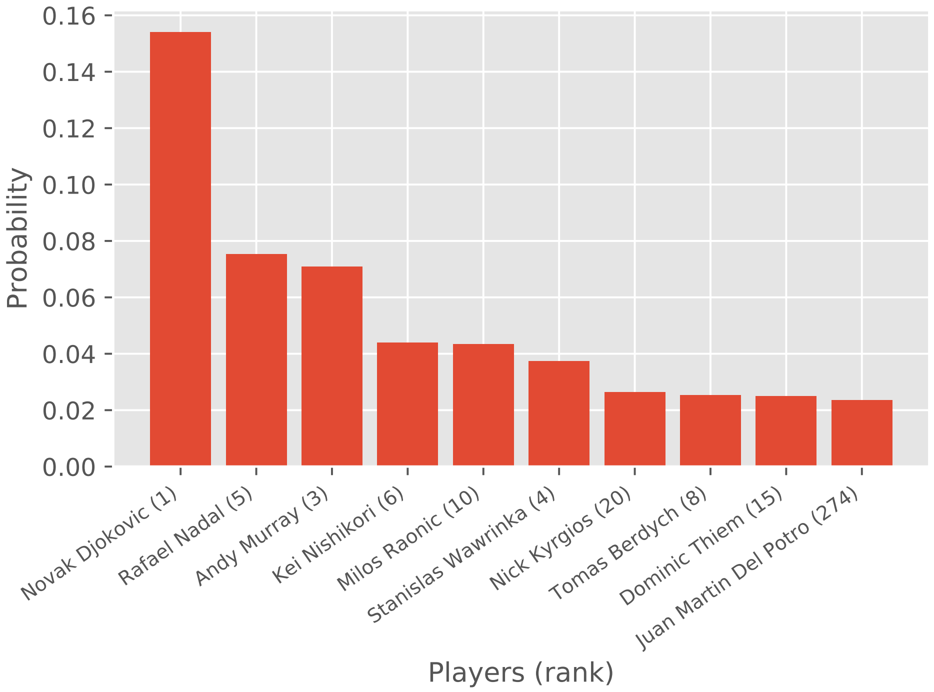
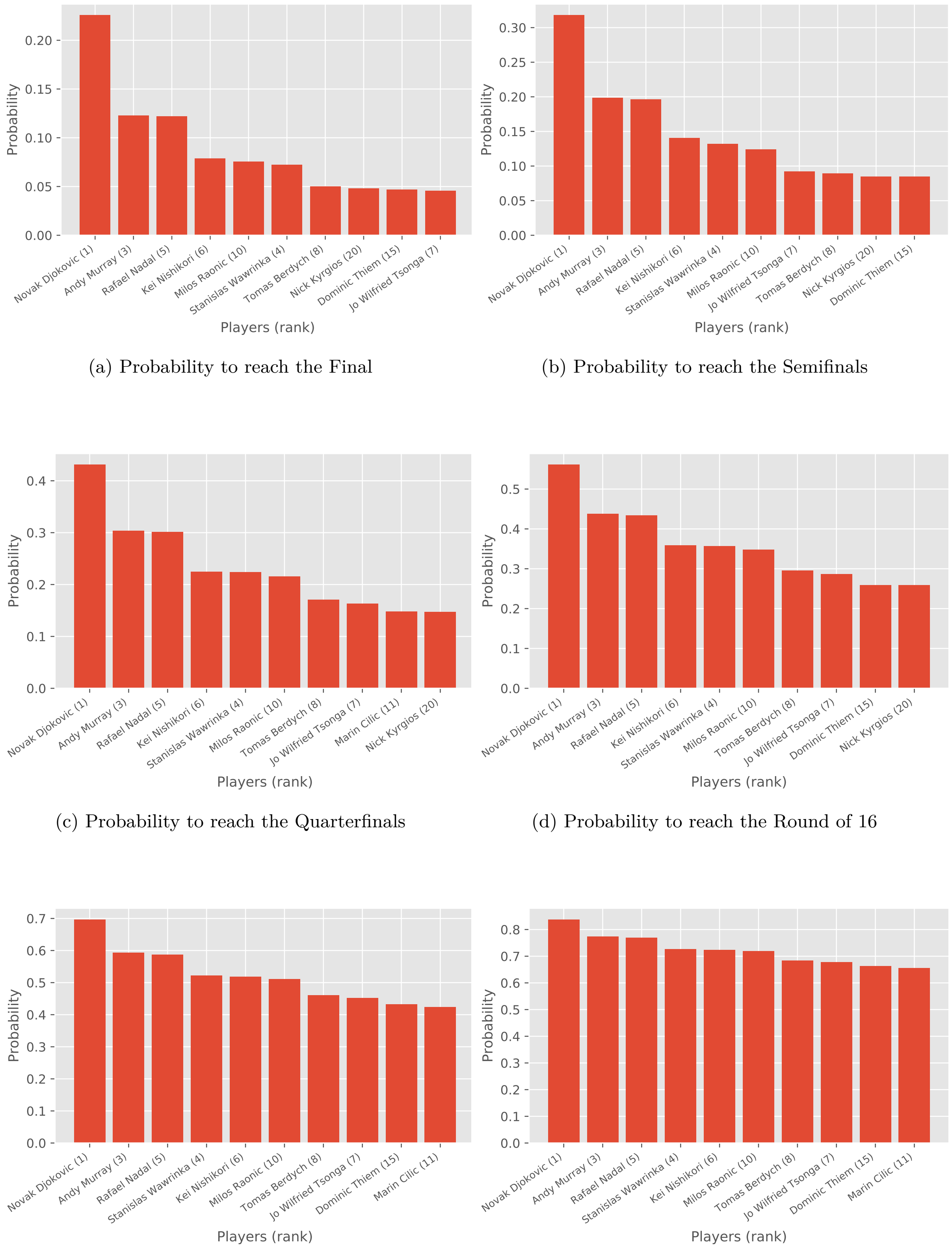


Figure 17: Top 10 players having the highest predicted probabilities to win the 2016 French Open. The number in parenthesis corresponds to the player rank.



(a) Probability to reach the Final

(b) Probability to reach the Semifinals

(c) Probability to reach the Quarterfinals

(d) Probability to reach the Round of 16

(e) Probability to reach the Round of 32

(f) Probability to reach the Round of 64

Figure 18: Top 10 players having the highest predicted probabilities to reach each round of the 2016 French Open. The number in parenthesis corresponds to the player rank.

B The 2019 French Open

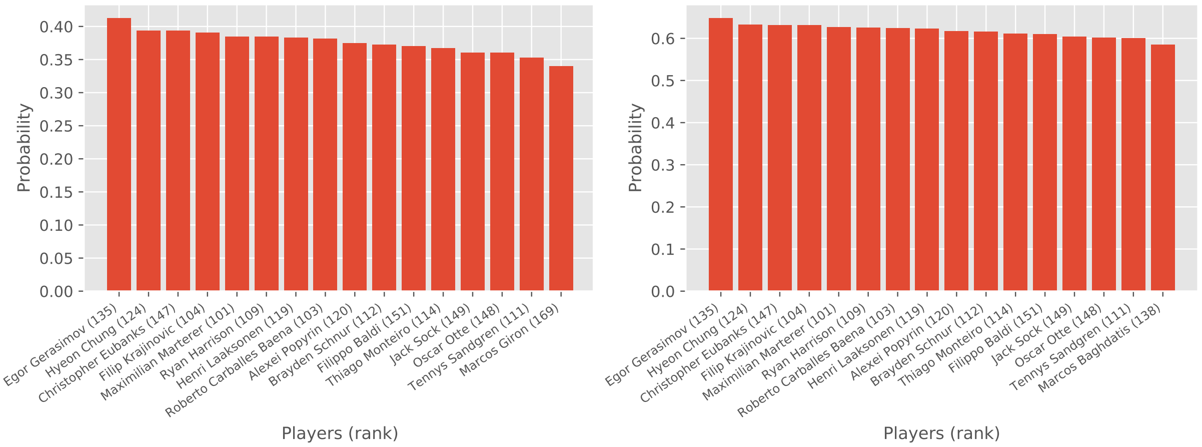
B.1 Players selection

B.1.1 Wild-cards

Players	Age	ATP Ranking	ATP Race to London
Antoine Hoang	23	136	147
Quentin Halys	22	156	142
Maxime Janvier	22	200	154
Elliot Benchetrit	20	263	203
Enzo Couacaud	24	232	202
Constant Lestienne	26	197	214
Mathias Bourgue	25	216	157
Johan Tatlot	23	260	315
Calvin Hemery	24	275	307
Benjamin Bonzi	22	290	323
Alexandre Muller	22	330	309

Table 10: French players prone to receive a wild-card for the 2019 French Open.

B.1.2 Qualifications

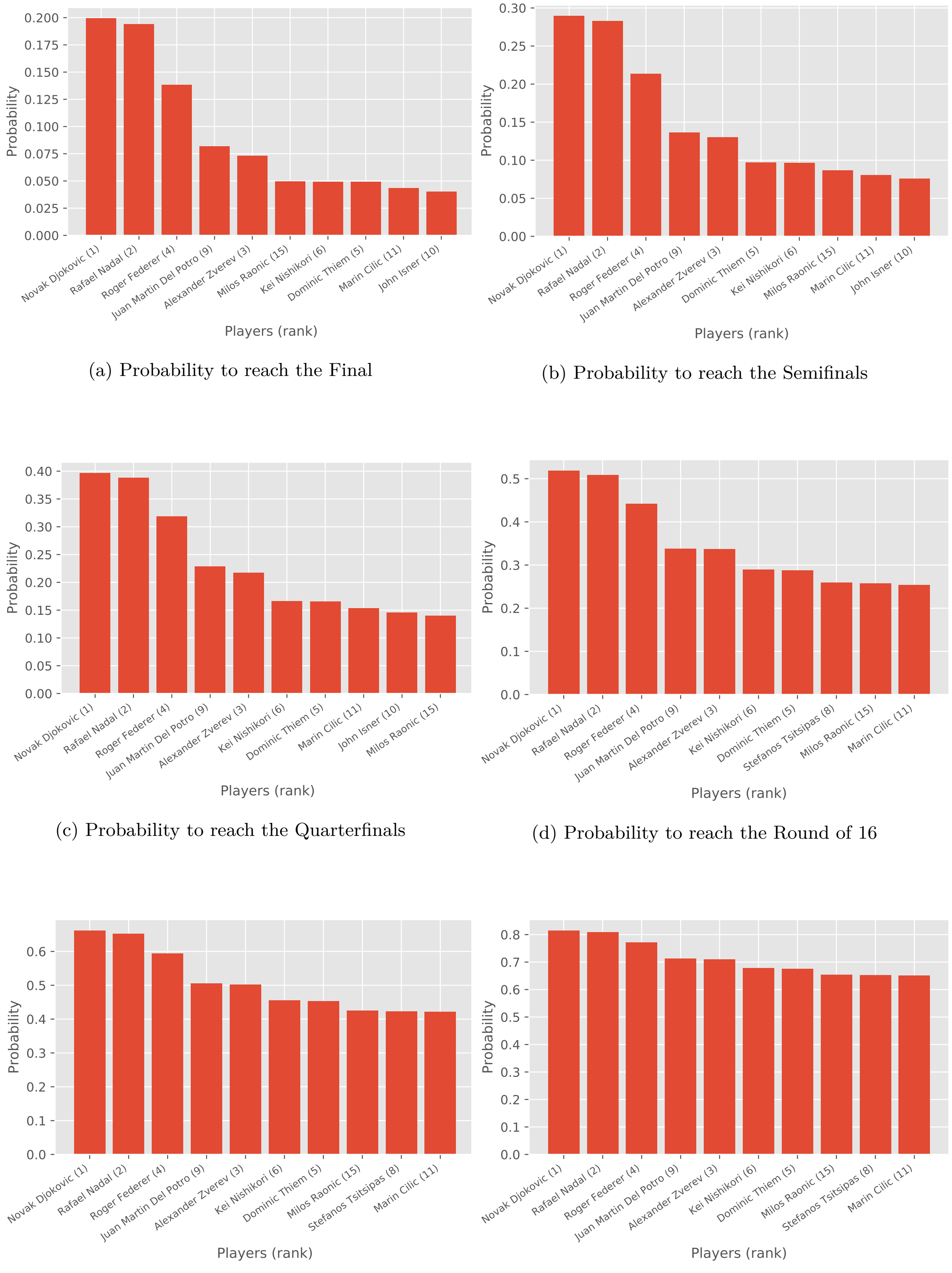


(a) Probability to reach the Round of 32

(b) Probability to reach the Round of 64

Figure 19: Top 16 players having the highest predicted probabilities to reach each round of the Qualifications in the 2019 French Open. The number in parenthesis corresponds to the player rank.

B.2 Predictions



(a) Probability to reach the Final

(b) Probability to reach the Semifinals

(c) Probability to reach the Quarterfinals

(d) Probability to reach the Round of 16

(e) Probability to reach the Round of 32

(f) Probability to reach the Round of 64

Figure 20: Top 10 players having the highest predicted probabilities to reach each round of the 2019 French Open. The number in parenthesis corresponds to the player rank.