Department of Electrical Engineering and Computer Science
University of Liege, Belgium

Big Data Project

Predicting the 2019 French Open's winner

Antoine Louis, Maxime Lamborelle and Tom Crasset

May, 2019

Outline

Part I : Introduction

Sports popularity

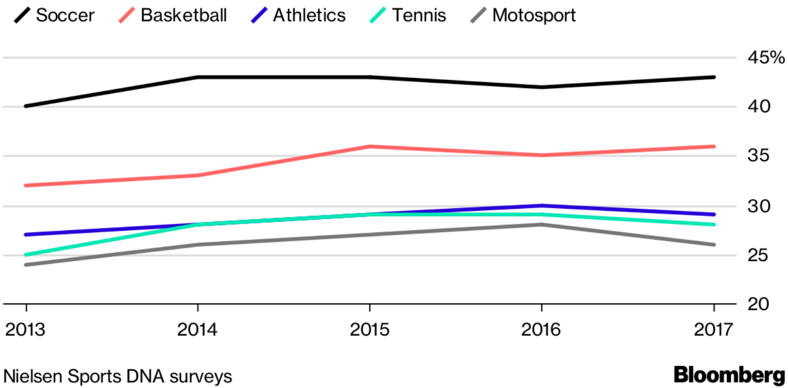


Figure: Percent of people who say they are "very interested" or "interested" in some popular sports in 18 markets across Americas, Europe, the Middle East and Asia.

Tennis Tournaments

The ATP World Tour includes 68 tournaments in total:

- **4** Grand Slam events: winner receives 2000 ranking points.
- **9** Masters 1000 events: winner receives 1000 ranking points.
- **13** ATP World Tour 500 events: winner receives 500 ranking points.
- **40** ATP World Tour 250 events: winner receives 250 ranking points.
- **1** ATP Finals: winner receives 1500 ranking points but only top 8 players in the world eligible.
- **1** ATP NextGen Finals: no ATP rankings awarded for this tournament and only top 8 players in the world aged 21 or younger eligible.

Grand Slams

Winner	$3.2m	$2.7m	$2.91m	$3.8m
Total (all players)	$42.85m	$48m	$45m	$53m

Figure: Grand Slams Prize Money Comparison in 2019.

Challenging question

" Who will win the 2019 French Open ? "

Part II : Dataset and Features

Dataset

Jeff Sackmann

- Author
- Software developer
- Sports statistics expert

The Jeff Sackmann Dataset

- More than 70 000 ATP matches
- Containing:
  - Players details
  - Match details
  - Players statistics
- Concerning more than 1500 players
- From 1968 to 2018

Jeff Sackmann Dataset

- Each match of the dataset contains the following features:

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	

J.Sackmann, *tennis_ATP*, Github Repository. Retrieved from https://github.com/JeffSackmann/tennis_atp

Features Extraction

- New features extracted from Jeff Sackmann Dataset:

$$\text{Service points \%} = \frac{\text{Number of service points of Player A}}{\text{Number of service points of Player A} + \text{Number of service points of Player B}}$$

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

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

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

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

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

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

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

Weighted average statistics

- To train a machine learning model, we would want, in addition to player details, to compute statistics reflecting the current performances of both players before a given match.
- These stats can be computed by **averaging**, for both players, the new constructed features over all the matches preceding their meeting.
- Considerations:
  1. More recent matches better reflect the current form of the player than older ones.
  2. Some players perform better on a surface than on another.

Time discounting

- Aims to estimate a player current performance.
- Also takes into consideration the fact that some great players sometimes face very bad seasons.
- Considers **all** matches of a player during his career and weight them so that the most recent ones have a bigger importance.

Formally, for each past match $i = 1, \ldots, n$ of a player, weight it so that:

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

where

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

Time discounting

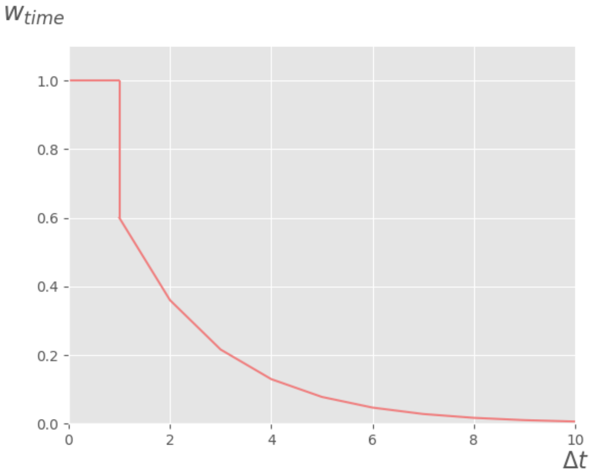


Figure: Time discounting function when $\eta = 0.6$

Surface weighting

- Aims to estimate a player current performance on a **given surface**.
- Instead of considering only matches on that surface, find **correlation coefficients** between surfaces to consider all matches.

Formally, for each pair of surfaces (A, B), we compute:

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

where

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

Surface weighting

	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: Correlation coefficients between court surfaces.

Combining weights

Final weight

For each past match $i = 1, \ldots, n$ of a player, give it a final weight:

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

Weighted mean

For a given statistical feature x, we compute the new feature $\bar{x}$ such that:

$$\bar{x} = \frac{\sum_{i=1}^{n} W_i x_i}{\sum_{i=1}^{n} W_i}$$

where

$$\begin{cases} x_i & \text{is the value of the feature } x \text{ in the } i^{\text{th}} \text{ match.} \\ W_i & \text{represent the weight given to the } i^{\text{th}} \text{ match enrolled in the average.} \end{cases}$$

Symmetric feature representation

- Considering the two values of the same feature independently of each other when training a model could lead to an **unfair consideration** of these same-feature values.
- Problematic because this difference of significance for the same feature could produce a different outcome prediction when the labels of the players are swapped.
- Solution: building new features resulting in the **difference** of the two values representing the same variable. For example:

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

Final features

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

Part III : Match prediction

Match prediction

- 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 trained multiple supervised learning algorithms where each training example was composed of:
  - A vector of input features **x**, composed of our 15 features.
  - An output value y, corresponding to the outcome of the match.
- This problem 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.

Reproducible analysis

- When testing different models, making an analysis reproducible is crucial to get consistent results.
- To do so, we:
  - fixed the size of our dataset,
  - sorted it before training to stay consistent across runs,
  - stuck with `scikit-learn v0.20.1`.
- The final dataset contains approximately **53100** matches, where:
  - 20% was used as testing set (about 10600 matches),
  - 80% was used as training set (about 42500 matches).

Machine Learning techniques

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: Test accuracy comparison between different machine learning algorithms.

Feature importance

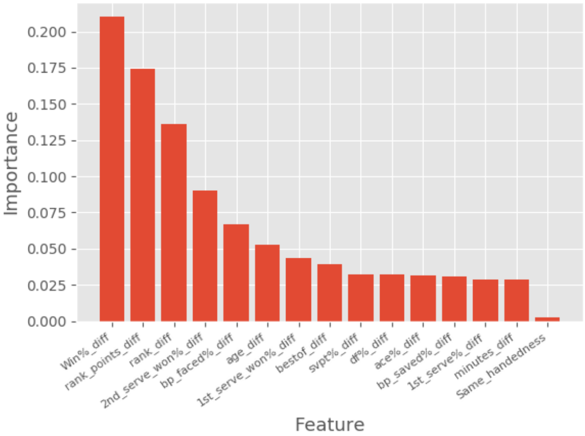


Figure: Feature importance according to the Random Forest algorithm.

Part IV : Tournament prediction

Tournament prediction

General idea

1. Taking the 128 players participating to the Grand Slam, predict the outcomes of each possible match:

$$C_{128}^{2} = 8128 \text{ possible different matches}$$

2. Simulate a large number of tournaments.
3. Compute for each player the probability of winning the tournament as a winning percentage over these simulations.

Draw generation

- **Naive way**: generate completely random draws.
  - Drawback: some generated draws would never exist in practice, as organisers always ensure that seeds meet as far as possible in the draw.
- **More accurate way**: generate draws according to the well-known seeds distribution in a Grand Slam tournament.
  - Advantage: significantly reduces the number of possible draws to consider (gain in accuracy).

Seed

From the 128 players participating to a Grand Slam, the 32 best players (according to the ATP ranking) are called "seeds".

Seeds distribution in a Grand Slam tournament

- 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** (4^{th} 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** (3^{rd} 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.

Monte-Carlo simulations

- Process: Sample randomly a certain number of possible draws and simulate them according to our predictions.

Probability to win a match

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 :

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

Probability to win a tournament

Given the total number of simulations n_{tot} and the number $n_{won,i}$ of simulations won by player i, the probability that player i win the tournament is given by:

$$P_i = \frac{n_{won,i}}{n_{tot}}$$

Part V : Model evaluation

Testing on the 2018 French Open

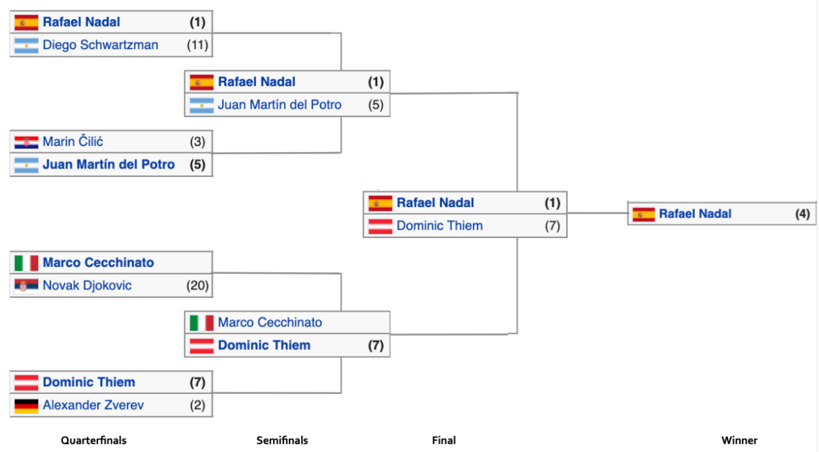
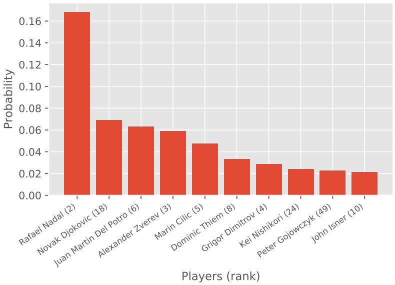
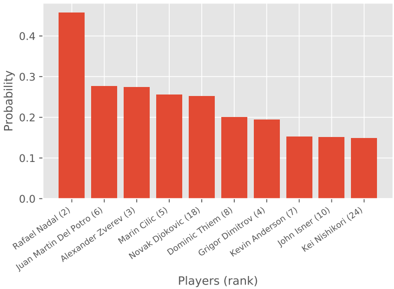


Figure: Draw from the Quarterfinals of the French Open 2018.

Testing on the 2018 French Open



(a) Probability to win the Grand Slam



(b) Probability to reach the Quarterfinals

Testing on the 2017 French Open

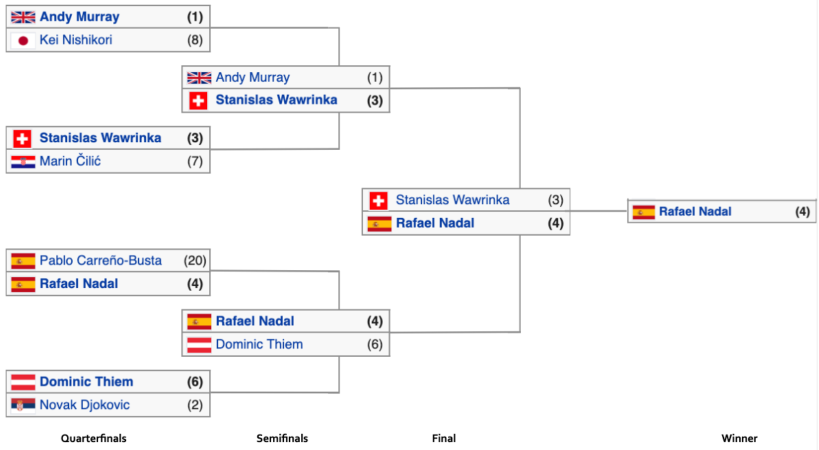
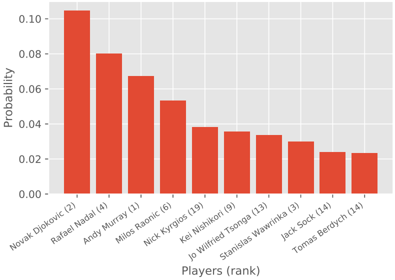
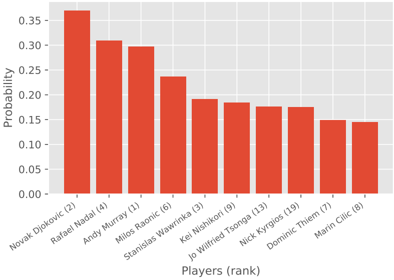


Figure: Draw from the Quarterfinals of the French Open 2017.

Testing on the 2017 French Open



(a) Probability to win the Grand Slam



(b) Probability to reach the Quarterfinals

Testing on the 2016 French Open

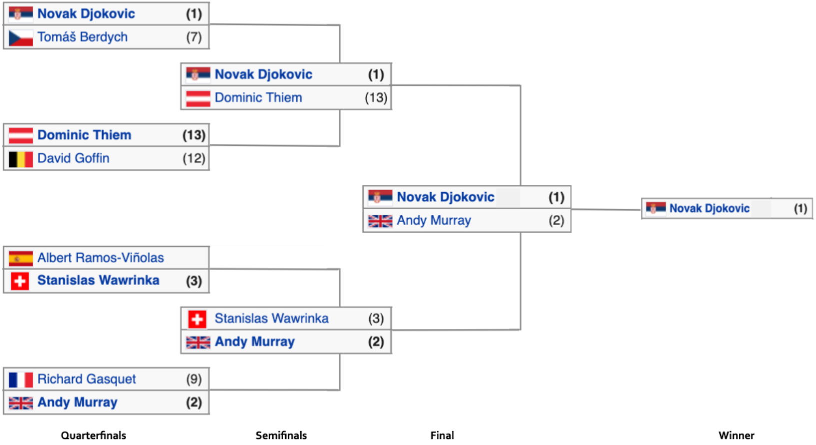
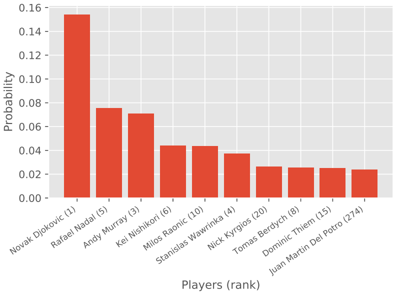
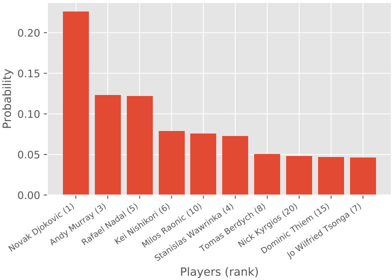


Figure: Draw from the Quarterfinals of the French Open 2016.

Testing on the 2016 French Open



(a) Probability to win the Grand Slam



(b) Probability to reach the Final

Part VI : Predicting the 2019 French Open’s winner

Players selection

In a Grand Slam, **128** tennis players are taken into account:

- **8** players receiving a “wild-card”.
- **16** players coming from the Qualifications.
- The world’s best **104** players according to the official ATP ranking, which is stopped by the organisers six weeks before the tournament.

Wild-cards

Definition

Special invitations offered to 8 players currently ranked outside the TOP 104, and allowing them to access the final draw without playing the Qualifications.

Attribution process

- **2** invitations are reserved for two French players:
  - The first card is offered to the French player who has won the most ATP points after a specific "Race", comprising ten French tournaments.
  - The second card is awarded to the highest ranked French player according to the ATP ranking Race to London, stopped two weeks before the tournament.
- **2** invitations result from agreements between National Tennis Federations:
  - One card is offered to an Australian player.
  - Another card is offered to an American player.
- **4** invitations given to young deserving players.

Wild-cards

- **2** invitations reserved for two French players:

NOM	PRÉNOM	TOTAL DES POINTS GAGNÉS
1 BARRERE	Grégoire	233
2 POUILLE	Lucas	110
3 HUMBERT	Ugo	109
4 HALYS	Quentin	83
5 JANVIER	Maxime	77
6 HOANG	Antoine	75
7 MOUTET	Corentin	62
8 BOURGUE	Mathias	37
9 MAHUT	Nicolas	30
10 BENCHETRIT	Elliot	21

(a) Ranking of "Wild-card Race" (9/05/2019)

ATP RACE TO LONDON

Ranking	Move	Country	Player	Age	Points
104	↑ 2		Corentin Moutet	20	213
143	↑ 16		Quentin Halys	22	146
154	↑ 15		Adrian Mannarino	30	135
160	↓ 3		Maxime Janvier	22	133
165	↓ 12		Antoine Hoang	23	128
175	↓ 4		Mathias Bourgue	25	112

(b) Ranking of ATP Race To London (11/05/2019)

Wild-cards

- **2** invitations resulting from agreements between National Tennis Federations:

Tommy Paul (136th, 21yo)

Alexei Popyrin (120th, 19yo)

Wild-cards

- 4 invitations given to young deserving players, most of the time French players:

Players	Age	ATP Ranking	ATP Race to London
Antoine Hoang	23	148	165
Quentin Halys	22	176	143
Maxime Janvier	22	197	160
Elliot Benchetrit	20	261	216
Enzo Couacaud	24	238	224
Constant Lestienne	26	193	240
Mathias Bourgue	25	229	175
Johan Tatlot	23	251	302
Calvin Hemery	24	311	318
Benjamin Bonzi	22	294	341
Alexandre Muller	22	300	264

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

Qualifications

Definition

The Qualifications are a pre-tournament that begins five days before the real tournament and welcomes 128 players, ranked outside the TOP 104. The sixteen players who win the final round of the Qualifications, that is the Round of 32, access the final tournament.

- As in the final Grand Slam, there are also "seeds" in the Qualifications, where the same distribution applies when creating the draw.
- To predict the 16 players that will pass the Qualifications, we can apply our model on the 128 best players outside the TOP 104 (ranked between 104th - 232th).
- The 16 players that we predict to pass the Round of 32 will be the considered ones.

Qualifications

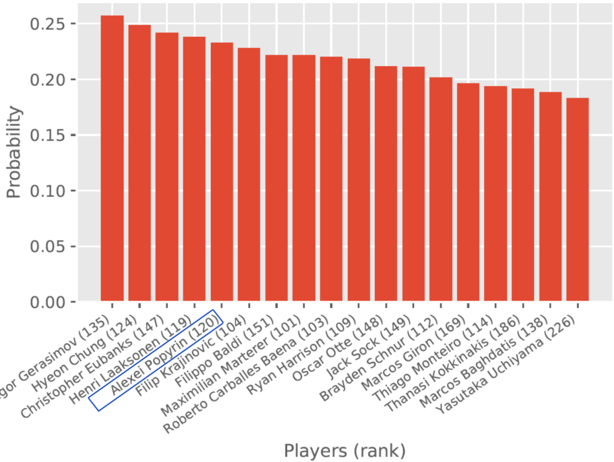


Figure: Top 18 players having the highest predicted probabilities to pass the Qualifications of the French Open 2019

TOP 104 players

ROLAND-GARROS 2019
26th May - 9th June

Men's singles
Official player list
Ranking on 15/04/2019

1	DJOKOVIC Novak (SRB)	1	65	HUMBERT Ugo (FRA)	64
2	NADAL Rafael (ESP)	2	66	FRITZ Taylor (USA)	65
3	ZVEREV Alexander (GER)	3	67	QUERREY Sam (USA)	66
4	FEDERER Roger (SUI)	4	68	THOMPSON Jordan (AUS)	67
5	THIEM Dominic (AUT)	5	69	RUUD Casper (NOR)	68
6	NISHIKORI Kei (JPN)	6	70	NISHIOKA Yoshihito (JPN)	69
7	ANDERSON Kevin (RSA)	7	71	BEDENE Aljaz (SLO)	70
8	TSITSIPAS Stefanos (GRE)	8	72	JAZIRI Malek (TUN)	71
9	DEL POTRO Juan Martin (ARG)	9	73	DANIEL Taro (JPN)	72
10	ISNER John (USA)	10	74	TOMIC Bernard (AUS)	73
11	CILIC Marin (CRO)	11	75	DELBONIS Federico (ARG)	74
12	KHACHANOV Karen (RUS)	12	76	DELLIEN Hugo (BOL)	75
13	CORIC Borna (CRO)	13	77	GULBIS Ernests (LAT)	76
14	MEDVEDEV Daniil (RUS)	14	78	ZVEREV Mischa (GER)	77
15	RAONIC Milos (CAN)	15	79	COPIL Marius (ROU)	78
16	CECCHINATO Marco (ITA)	16	80	LONDERO Juan Ignacio (ARG)	79
17	BASILASHVILI Nikoloz (GEO)	17	81	GUNNESWARAN Prajnesh (IND)	80
18	FOGNINI Fabio (ITA)	18	82	CUEVAS Pablo (URU)	81
19	MONFILS Gael (FRA)	19	83	RAMOS-VINOLAS Albert (ESP)	82
20	SHAPOVALOV Denis (CAN)	20	84	JARRY Nicolas (CHI)	83
21	GOFFIN David (BEL)	21	85	ANDREOZZI Guido (ARG)	84
22	BAUTISTA AGUT Roberto (ESP)	22	86	KUDLA Denis (USA)	85
23	EDMUND Kyle (GBR)	23	87	KOVALIK Jozef (SVK)	85*
24	SCHWARTZMAN Diego (ARG)	24	88	ANDUJAR Pablo (ESP)	86
25	DE MINAUR Alex (AUS)	25	89	GOJOWCZYK Peter (GER)	87
26	SIMON Gilles (FRA)	26	90	LOPEZ Feliciano (ESP)	88
27	CARRENO BUSTA Pablo (ESP)	27	91	TIPSAREVIC Janko (SRB)	88*
28	DIMITROV Grigor (BUL)	28	92	EVANS Daniel (GBR)	89
29	TIAFOE Frances (USA)	29	93	RUBLEV Andrey (RUS)	90
30	GASQUET Richard (FRA)	30	94	KECMANOVIC Miomir (SRB)	91
31	POUILLE Lucas (FRA)	31	95	KLAHN Bradley (USA)	92
32	DJERE Laslo (SRB)	32	96	FABBIANO Thomas (ITA)	93
33	AUGER-ALIASSIME Felix (CAN)	33	97	HARRIS Lloyd (RSA)	94
34	KYRGIOS Nick (AUS)	34	98	VESELY Jiri (CZE)	95
35	TSONGA Jo-Wilfried (FRA)	34*	99	STEBE Cedrik-Marcel (GER)	95*
36	PELLA Guido (ARG)	35	100	SONEGO Lorenzo (ITA)	96
37	WAWRINKA Stan (SUI)	36	101	KARLOVIC Ivo (CRO)	97
38	FUCSOVICS Marton (HUN)	37	102	BERDYCH Tomas (CZE)	98
39	VERDASCO Fernando (ESP)	38	103	BERANKIS Ricardas (LTU)	99
40	MILLMAN John (AUS)	39	104	BUBLIK Alexander (KAZ)	100
41	KOHLSCHREIBER Philipp (GER)	40			
42	CHARDY Jeremy (FRA)	41			
43	KUKUSHKIN Mikhail (KAZ)	42			
44	PAIRE Benoit (FRA)	43			
45	STRUFF Jan-Lennard (GER)	44			
46	ALBOT Radu (MDA)	45			
47	SEPPI Andreas (ITA)	46			
48	GARIN Cristian (CHI)	47			
49	LAJOVIC Dusan (SRB)	48			
50	HERBERT Pierre-Hugues (FRA)	49			
51	KLIZAN Martin (SVK)	50			
52	SOUSA Joao (POR)	51			
53	HURKACZ Hubert (POL)	52			
54	EBDEN Matthew (AUS)	53			
55	DZUMHUR Damir (BIH)	54			
56	BERRETTINI Matteo (ITA)	55			
57	NORRIE Cameron (GBR)	56			
58	OPELKA Reilly (USA)	57			
59	MANNARINO Adrian (FRA)	58			
60	JOHNSON Steve (USA)	59			
61	MCDONALD Mackenzie (USA)	60			
62	MUNAR Jaume (ESP)	61			
63	MAYER Leonardo (ARG)	62			
64	HAASE Robin (NED)	63			

16/04/2019 10:01   * : Classement protégé après interruption pour cause de blessure   Page 1/1

- The 104 remaining players are the TOP 104 players according to the ATP ranking 6 weeks before the beginning of the French Open.
- List officially available on Roland-Garros website since April 17, 2019.

Injuries

- Among the 128 final players, some might not play the tournament, because of injuries that could happen during the tournaments preceding the French Open.
- The website `tennisexplorer.com` gives:
  - a list of all the tennis players who recently withdraw because of an injury,
  - a list of the players who returned from an injury by starting competition again.
- By analysing these two lists, we are able to identify players that recently suffered from an injury and didn't play since there.
- Among the 128 considered players, three of them are currently concerned by this scenario:
  - Gilles Simon (26th),
  - Pablo Carreno Busta (27th),
  - Matthew Ebden (53rd).

Substitute players

- If some players were to withdraw before the first round of the Grand Slam, their substitutes would be picked among the **"Lucky losers"**.

Lucky Loser

Status given to the top 8 players who lost the final round of the Qualifications.

- These 8 players have to stay in Paris and a random draw is made to determine the order in which they would have to replace the forfeit players.
- Determining the Lucky Losers comes down to determining which players will lose the final round of the Qualifications and taking the eight top-ranked.

Substitute players

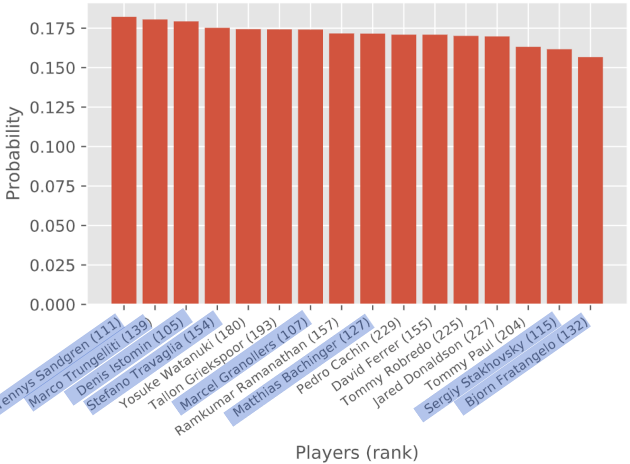


Figure: Second Top 16 players having the highest predicted probabilities to reach the final round of the Qualifications in the French Open 2019.

Predictions

- To compute recent statistics for the 128 players of this edition, we had to collect new data about recent tennis matches.
- We scraped the official ATP website and were able to complete our dataset with the most recent tennis matches.
- After computing the stats of the 128 selected players, we used our model to finally predict the winner of the French Open 2019.

And the predicted winner is...

Predicted winner

...Rafael Nadal !

Predicted winner

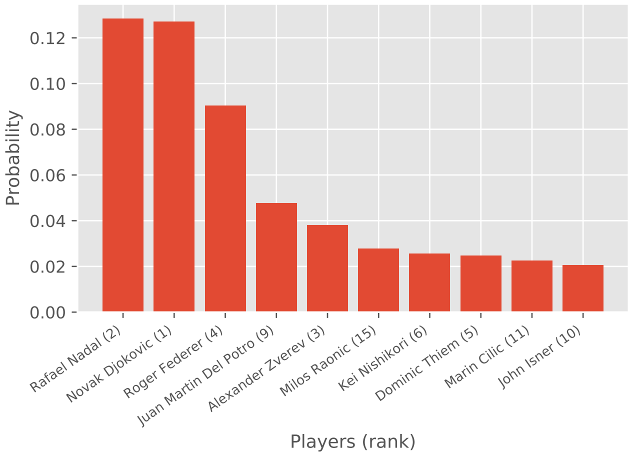
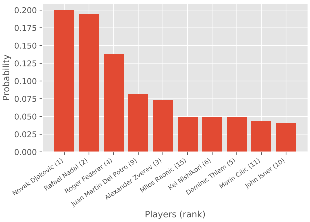
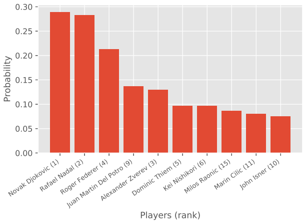


Figure: Top 10 players having the highest predicted probabilities to win the French Open 2019.

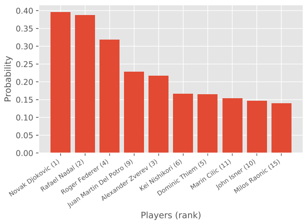
Predicted rounds



(a) Probability to reach the Final

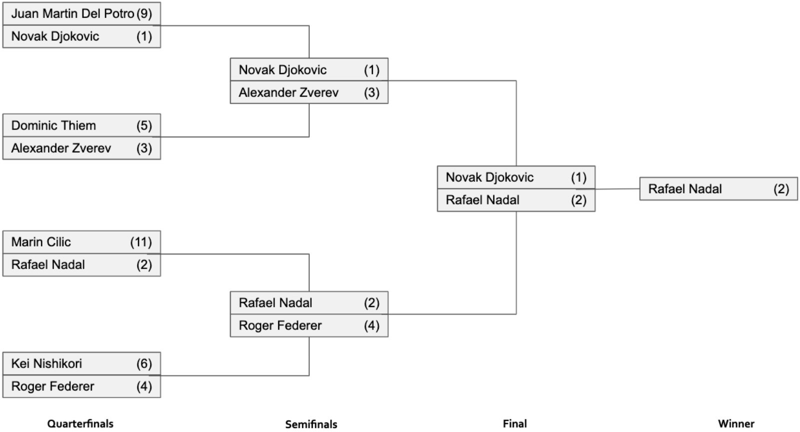


(b) Probability to reach the Semifinals



(c) Probability to reach the Quarterfinals

Predicted Quarterfinals draw



Successful outsiders

Outsider

Player thought to have little chance of success in the tournament.

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

Favourites

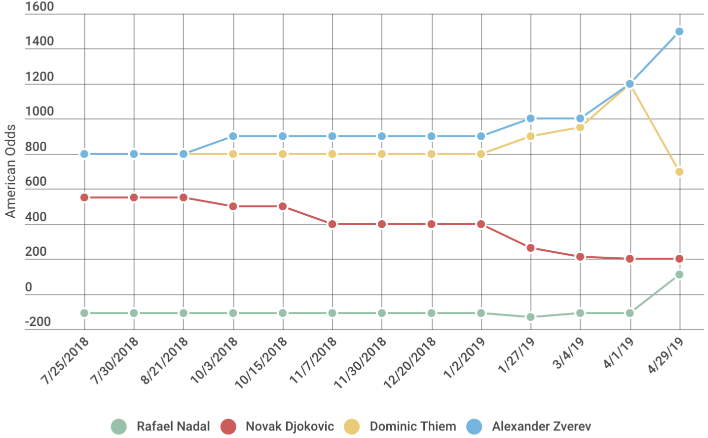


Figure: Evolution of French Open odds for the top four contenders according to a number of top online sportsbooks.

Part VII : Conclusion

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 developed a novel method to predict the winner of a Grand Slam tournament, and applied it to the 2019 French Open:
  - First, we developed a method to extract tennis statistics features from raw historical data.
  - 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.
- With our model, we were able to predict correctly the majority of the players that reached the Quarterfinals of three past French Opens, as well as the winners of two past editions.

Difficulties

- The main difficulty of this project was the **preprocessing of the data**. We had to transform in a clever way our initial Jeff Sackmann dataset into representative data to help us gain in accuracy when predicting the outcomes of single matches.
- Another difficulty appears when **selecting the 128 players** of the 2019 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.

Future work

- **Additional Features**
- **Other ML Algorithms**
- **Women's Tennis**
- **Tournament Generalisation**

Thanks !