Tennis Prediction: Who is the Winner of the 2020 US Open™ ?

Introduction

Data science and predictive analytics are becoming more mainstream in tennis. Players and coaches use existing data extensively to predict odds of winning and identify areas for potential improvement.

Given historical data, we aim to predict the likelihood of winning future hard surface tournaments, specifically the upcoming 2020 U.S. Open, of current Top 30 ATP male singles tennis players.

Data Collection

We gathered all historical match statistics from the Match Charting Project, a public crowdsourcing GitHub repository. Upon cleaning and processing, we obtained 394 observations concerning 24 out of the Top 30 players with 28 features.

We also scraped the respective players' overall, serve, return and under pressure statistics from the ATP website.

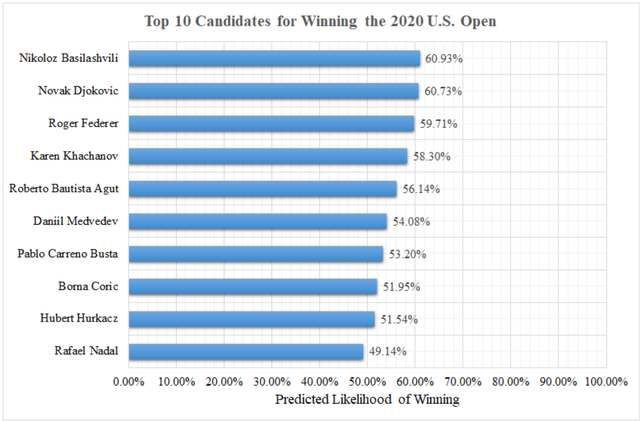
Methodology

We applied the following analyses and associated evaluation criteria:

- Logistic Regression (Training and Testing Accuracy)
- K-Means Clustering (Qualitative)
- Comparison with ATP (Kendall's Tau Coefficient and P-values)

Logistic Regression

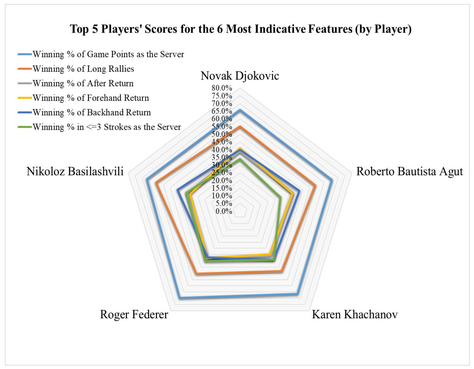
Predictions from our best-performing model on all 28 features with an accuracy of 84.22%:



Logistic Regression Model	# of Features	Training Accuracy	Testing Accuracy
Our Full Model (best performing)	**28**	**0.8656**	**0.8422**
Only Return-Related Features	6	0.8095	0.8030
Only Under Pressure-Related Features	3	0.6649	0.6614
Only Serve-Related Features	9	0.6632	0.6597
Our Full Model (held out Basilashvili)	28	0.8656	0.8419
Our Full Model (held out Djokovic)	28	0.8523	0.8226
Our Full Model (held out Federer)	28	0.8502	0.8268
Our Full Model (held out Khachanov)	28	0.8647	0.8414
Our Full Model (held out Agut)	28	0.8653	0.8387
Baseline Model (always guessing win)	**0**	**0.4984**	**0.5063**

6 Most Predictive Features ranked by RFE:

Rank	Feature	Mean of Top 5 Players	Mean of Last 5 Players	Difference in Mean
1	Winning % of Game Points as the Server	67.0%	58.1%	8.8%
2	Winning % of Long Rallies	53.5%	41.3%	12.2%
3	Winning % of After Return	37.6%	28.8%	8.8%
4	Winning % of Forehand Return	37.3%	28.6%	8.7%
5	Winning % of Backhand Return	40.7%	29.8%	10.9%
6	Winning % in <=3 Strokes as the Server	36.2%	37.6%	**-1.5%**



***Note*: John Isner**, though among the least likely winners, is the best performer in **winning % in <=3 strokes as server.**

What a Powerful Server!!!

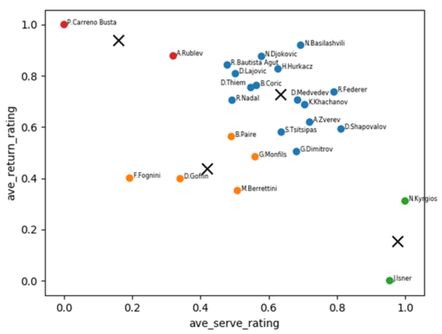
GoldenBagel vs. ATP

GB K-means Clustering (using 6 most predictive features by RFE):

serve rating: average of **2** serve features

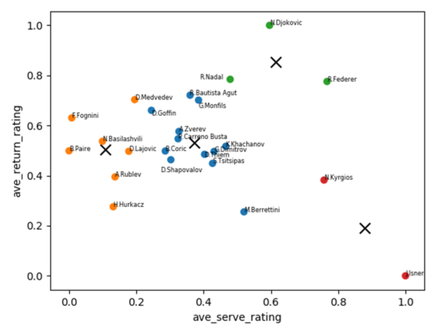
return rating: average of **4** return features

(rescaled to 0-1 range by min-max scaling)



ATP K-means Clustering (using serve and return ratings scraped from "Stats Leaderboards" section on ATP website):

(rescaled to 0-1 range by min-max scaling)



Stats LeaderBoards												
Overall				Serve				Return				Under Pressure			
Player Name	GB Ranking	ATP Ranking	ATP Relative Ranking	Player Name	GB Ranking	ATP Ranking	ATP Relative Ranking	Player Name	GB Ranking	ATP Ranking	ATP Relative Ranking	Player Name	GB Ranking	ATP Ranking	ATP Relative Ranking
Nikoloz Basilashvili	1	27	21	John Isner	1	2	1	Pablo Carreno Busta	1	52	19	John Isner	1	34	14
Novak Djokovic	2	1	1	Nick Kyrgios	2	4	2	Nikoloz Basilashvili	2	25	10	Karen Khachanov	2	47	18
Roger Federer	3	4	4	Roger Federer	3	5	3	Novak Djokovic	3	2	1	Borna Coric	3	59	20
Karen Khachanov	4	15	14	Denis Shapovalov	4	25	12	Hubert Hurkacz	4	48	18	Roger Federer	4	15	5
Roberto Bautista Agut	5	12	12	Nikoloz Basilashvili	5	98	24	Roberto Bautista Agut	5	15	6	Nick Kyrgios	5	18	6
Daniil Medvedev	6	5	5	Alexander Zverev	6	37	15	Roger Federer	6	13	5	Denis Shapovalov	6	28	11
Pablo Carreno Busta	7	25	20	Daniil Medvedev	7	10	7	Rafael Nadal	7	3	2	Alexander Zverev	7	46	17
Borna Coric	8	33	23	Stefanos Tsitsipas	8	6	4	Daniil Medvedev	8	4	3	Stefanos Tsitsipas	8	23	9
Hubert Hurkacz	9	29	22	Karen Khachanov	9	20	11	Dominic Thiem	9	21	9	Pablo Carreno Busta	9	40	16
Rafael Nadal	10	2	2	Grigor Dimitrov	10	53	18	Dusan Lajovic	10	30	14	Novak Djokovic	10	5	2
Denis Shapovalov	11	16	15	Hubert Hurkacz	11	48	17	Karen Khachanov	11	12	4	Nikoloz Basilashvili	11	37	15
Alexander Zverev	12	7	7	Novak Djokovic	12	8	6	Andrey Rublev	12	28	13	Rafael Nadal	12	3	1
Stefanos Tsitsipas	13	6	6	Gael Monfils	13	39	16	Borna Coric	13	62	21	Daniil Medvedev	13	27	10
Dominic Thiem	14	3	3	Borna Coric	14	72	21	Denis Shapovalov	14	74	22	Roberto Bautista Agut	14	94	24
Dusan Lajovic	15	23	19	Dusan Lajovic	15	75	23	Alexander Zverev	15	26	11	Grigor Dimitrov	15	61	21
Andrey Rublev	16	14	13	Dominic Thiem	16	13	8	Stefanos Tsitsipas	16	59	20	Dusan Lajovic	16	86	23
Grigor Dimitrov	17	19	16	Roberto Bautista Agut	17	36	14	Gael Monfils	17	40	16	David Goffin	17	30	12
Nick Kyrgios	18	40	24	Rafael Nadal	18	7	5	Grigor Dimitrov	18	16	7	Matteo Berrettini	18	10	4
Gael Monfils	19	9	9	Matteo Berrettini	19	17	9	Benoit Paire	19	27	12	Andrey Rublev	19	19	7
Benoit Paire	20	22	18	Benoit Paire	20	73	22	Fabio Fognini	20	45	17	Dominic Thiem	20	6	3
John Isner	21	21	17	David Goffin	21	71	20	Matteo Berrettini	21	34	15	Gael Monfils	21	32	13
David Goffin	22	10	10	Andrey Rublev	22	32	13	David Goffin	22	19	8	Hubert Hurkacz	22	22	8
Matteo Berrettini	23	8	8	Fabio Fognini	23	66	19	Nick Kyrgios	23	75	23	Benoit Paire	23	53	19
Fabio Fognini	24	11	11	Pablo Carreno Busta	24	18	10	John Isner	24	105	24	Fabio Fognini	24	68	22

GB ranking: predictions from logistic regression on all, serve-, return-, and under pressure-related features

ATP ranking: average of selected criteria

The results below suggest weak correlation between GB & ATP rankings.

Ranking	Kendall's Tau	P-value
Overall	0.0652	0.6765
Serve	0.2754	0.0623
Return	0.3188	0.0298
Under Pressure	0.0217	0.9024

Challenges

- Understand the raw data in the context of tennis, and come up with meaningful features about different components of the game
- Handle null values when certain statistics for a match are missing

Significance & Limitations

***Significance*:**

- Our model can be used to predict the probability of winning for unfamiliar or emerging male tennis players if we have access to their match statistics.

Limitations:

- When we selected features for logistic regression, those that contained more than half of null values were disregarded. This might diminish some predictive power of our regression model.
- Omitted variables to consider: age, health conditions, injuries, career length at time of the match

Extension

- We tried to factor in career length (i.e. more recent matches are more indicative of player's current and near-future performance) and calculated time-weighted prediction inputs. The results turned out to be different. The top 5 players are: Djokovic, Federer, Basilashvili, Shapovalov and Khachanov.

BROWN

Team GoldenBagel:

Yuxiao Cui, Mindy Li, Yuexin Li, Thomas Sze