Predicting the Winner of the 2020 U.S. Open

Team GoldenBagel: yli153, tsze, ycui10, sli90

Goal

Men's tennis has a new player on the court: data science! Predictive analytics has become more mainstream in tennis where players are now using existing data to predict the odds of winning to gain a competitive advantage. Our project can be used by tennis fans, coaches, sponsors and players to predict the likelihood of winning on hard surfaces and help identify areas for potential improvement. Therefore, our project focuses on the following task: given historical performance data on matches between current (ATP ranked) Top 30 male singles tennis players, predict their likelihood of winning future hard surface tournaments (i.e. the 2020 U.S. Open).

Data

We used two primary data sources for this project: The Match Charting Project, a collaborative, crowdsourcing GitHub repository which documents detailed match statistics by player, and the ATP website which serves as an official reference for male tennis.

From the Match Charting Project, we collected data from 1000+ historical matches (hard surfaces only), filtered them by keeping only the 197 matches between Top 30 male singles players ranked by ATP by 2020/02/24, and extracted 60+ different features encompassing serve, return, under pressure, shot directions, shot types, shot depths and many other aspects of a match. After processing the features and throwing away the ones that contain over half null values, we were left with 28 "good" features, which all had values between 0 and 1. The selected features strike a balance between wide coverage of different components of a match and the limitations of too many null values. Our final data table has a total of 394 observations concerning 24 out of the Top 30 players, where each row consists of the player's name, a binary match outcome (1 if he is the winner else 0), followed by his scores for the 28 features.

From the ATP official website, we scraped data including the overall male singles rankings as of 2020/02/24, as well as serve, return and under pressure leaders rankings and ratings from the "Stats LeaderBoard" section filtered with fields "versus Top 50 Players", "52 weeks" and "Hard Surface".

Model and Evaluation Setup

Our primary goal is to optimize the accuracy in predicting the probability of winning for each player given historical match data. Since our dependent variable is binary (i.e. 1 if win and 0 if lose), we have chosen to use logistic regression for the prediction task. We randomly split our data into 80% and 20% for training and testing respectively and report the average training and testing accuracies from 100 iterations. The probability of winning for each player is reported as the average of the predicted outcomes aggregated from 100 iterations. This performs cross validation and ensures that our model produces stabilized results that converge to a steady level. We also held out data from the predicted most likely winners to check whether our model is memorizing the label for one certain player. Testing accuracy is compared to training accuracy to make sure that our model does not overfit.

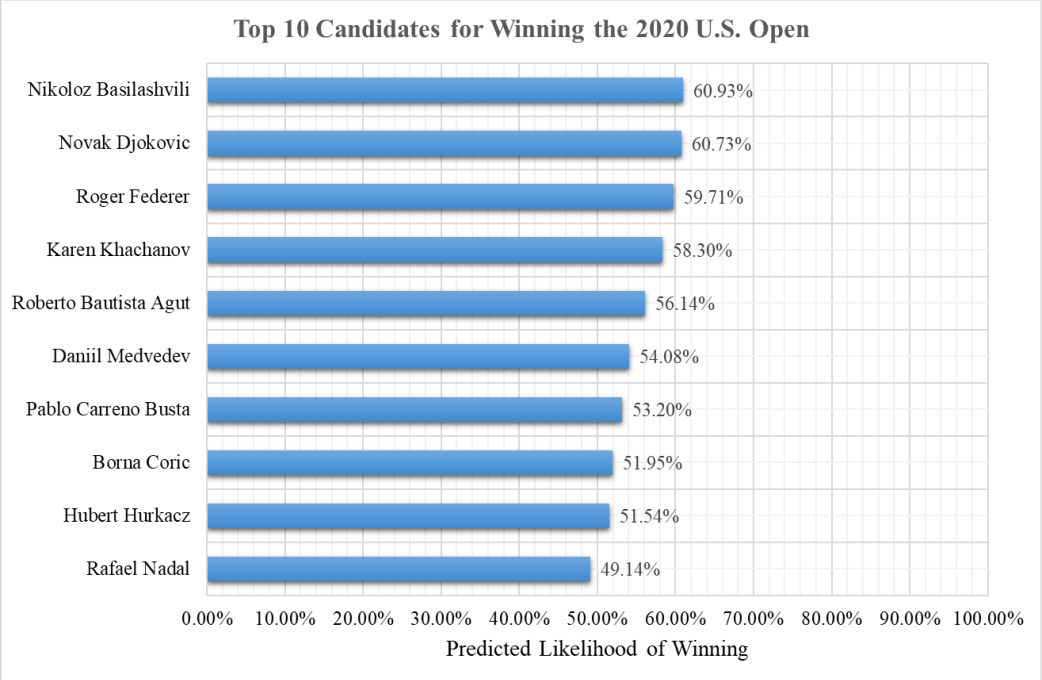
Next, we use two statistical analyses to compare the results from our logistic regression model to ATP which serves as a mainstream reference. We first run K-Means clustering with our six most predictive features derived from Recursive Feature Elimination (RFE) and then with ATP-selected features. We also compare our rankings of the 24 players to the ATP rankings under overall, serve, return, and under

pressure categories. The former is evaluated qualitatively with respect to visualization, while the latter is evaluated with respect to Kendall's Tau Correlation Coefficient.

Results and Analysis

Claim #1: Our full logistic regression model on all 28 features predicts the probability of winning with an accuracy of 84.22%, outperforming the baseline model which always guesses win by a significant margin (i.e., baseline accuracy of 50.63%), and is not memorizing the label for a certain player.

Prediction Results: The following figure shows the top 10 most likely winners of the 2020 U.S. Open with their predicted likelihood of winning in the descending order.

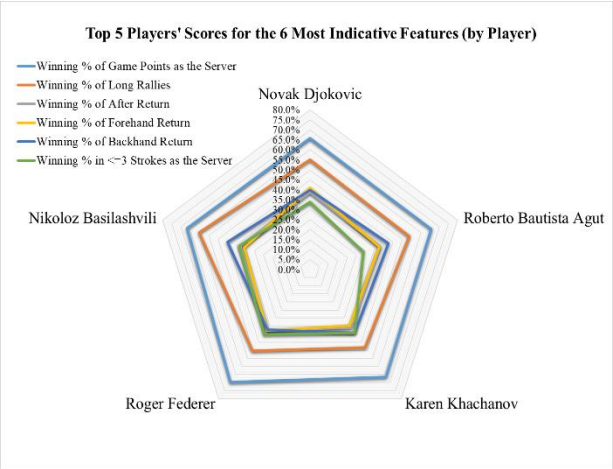


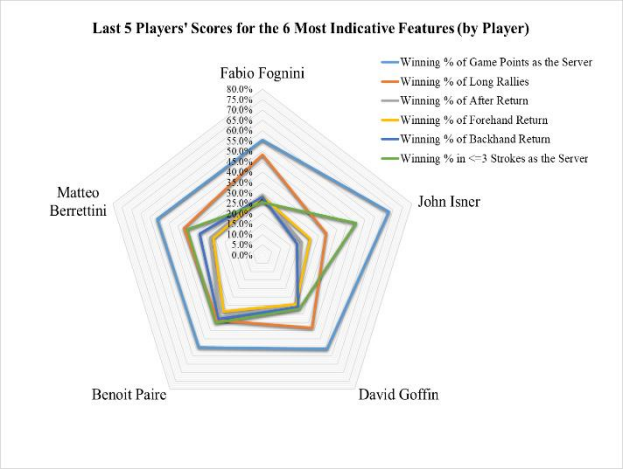
Support for Claim #1: The table below shows the training and testing accuracies of our best-performing model compared to the baseline model, which always guesses win and models regressing on different subsets of features (i.e., only return, under pressure and serve-related features). In addition, when the predicted most likely winners were held out from our regression, the testing accuracy remains almost unaffected, suggesting that our model is not memorizing the label for a certain player.

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**

Claim #2: Winning percentages of game points as the server, long rallies, after return, forehand return, backhand return and in fewer than or equal to 3 strokes as the server are the 6 most predictive features of match outcome.

Support for Claim #2: The aforementioned 6 features are the ones with the highest rank determined by Recursive Feature Elimination (RFE). In addition, on average, the most likely winners outperform the least likely winners in all but one. The two radar charts below compare the top 5 players' scores for the 6 most indicate features to the last 5 players' scores. The discrepancies can be visually estimated from sizes of the "stars" in the charts below.





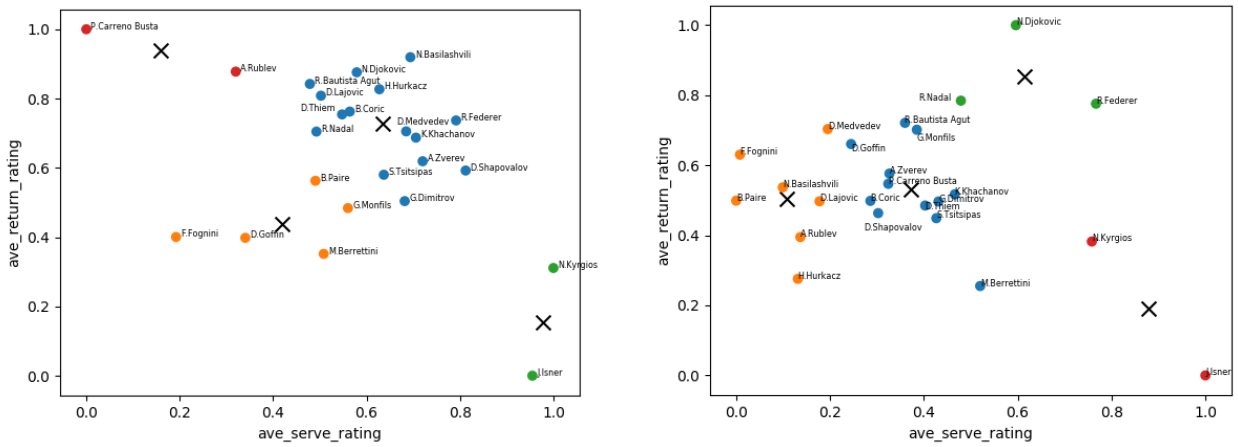
The following table shows the means and standard deviations of each feature for top and last 5 players. On average, the top 5 players outperform the last 5 players by a margin of 8.7% to 12.2% in the first five features but do slightly worse in the feature "winning % in fewer than or equal to 3 strokes as the server" (i.e., the difference is -1.5%). Among the least likely winners, John Isner has an outstanding score of 50.1% for the last feature, which is actually the highest among all 24 players. While the results are mostly in line with our expectation that the most likely winners show stronger performance in the most predictive features than the least likely winners, the outperformance of John Isner in the last feature skews the average for that particular case.

Rank	Feature	Mean of Top 5 Players	Std Dev. of Top 5 Players	Mean of Last 5 Players	Std Dev. of Last 5 Players	Difference in Mean
1	Winning % of Game Points as the Server	67.0%	1.4%	58.1%	4.9%	8.8%
2	Winning % of Long Rallies	53.5%	4.1%	41.3%	4.7%	12.2%
3	Winning % of After Return	37.6%	0.9%	28.8%	5.1%	8.8%
4	Winning % of Forehand Return	37.3%	2.1%	28.6%	2.8%	8.7%
5	Winning % of Backhand Return	40.7%	2.6%	29.8%	6.6%	10.9%
6	Winning % in <=3 Strokes as the Server	36.2%	4.2%	37.6%	8.4%	**-1.5%**

Claim #3: K-means clustering using our six most predictive features shares limited similarities with that using ATP-selected features.

Support for Claim #3: The graph below on the left shows K-means clustering using our 6 most predictive features, whereas the graph on the right shows that using ATP-selected features. Our serve rating is the average of winning % of game points as the server and winning % in <=3 strokes as the server, and return rating is the average of the remaining 4 features. Note that the ratings are then re-scaled to 0-1 range through min-max scaling in both cases.

If we use the optimal K = 4, the two K-means clustering results are quite different and show only a few similarities: John Isner and Nick Kyrgios belong to the same cluster in both for their powerful serves but relatively weak returns; Novak Djokovic, Roger Federer and Rafael Nadal are assigned to the same centroid in both, but the compositions of the cluster to which they belong are different.



Claim #4: There is no evident correlation between our rankings of overall, serve, return and under pressure leaders and the ATP rankings. Our choice of features and ranking methodology evaluate match performances differently from ATP and provide an alternative.

Support for Claim #4: The table below compares our rankings to the ATP rankings with respect to players' overall, serve, return and under pressure performances. For each category, we ranked the players based on the predicted probabilities of winning from logistic regression on all, serve-, return-, and under pressure-related features, respectively. In contrast, ATP uses a different set of criteria for each category and ranks players by averaging the selected criteria.

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

The following table shows the Kendall's Tau correlation coefficients and the associated p-values for each category. The strongest agreement lies in return rankings while the weakest agreement is found with under pressure rankings. The p-values of 0.0623 and 0.0298 suggest that there is statistically significant evidence at 10% level that there is some concordance between our rankings and ATP rankings of serve and return leaders, while the p-values of 0.6765 and 0.9024 suggest that the correlations between our rankings and ATP rankings of overall and under pressure leaders are statistically insignificant.

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