Eric Browne
6/2/2020
Probability and Statistics 2
Final Project Essay

Upset Prediction in Tennis

Abstract:

Sports are no stranger to the Data Science world, whether it be predicting winners between teams, calculating contribution margins per player for their team, or deciding who is the absolute best, in each league. However, Tennis is not usually among the most popular sports to model from, partly due to its lack of quantifiable statistics on players, and other match features that can easily be recorded. In order to bridge that gap, we created a model to accurately predict an upset in a given tennis match; an upset being the lower ranking player beating the higher-ranking player. Taking multiple explanatory variables into account, the most significant variables were the Winner's and Loser's rank, and the amount of games the Winner and Loser won in each set of the match. We also found that as a match progresses into more sets pass the minimum in a match, the chances of an upset occurring increases, the surface type of the court was insignificant, as well as that upsets occur more often in the Top 100 Players bracket. To better improve our model in the future, having more match features such as age of players, match history between players, and average speed of serves between players would greatly benefit its accuracy and overall completeness.

Main Body:

Predicting an upset in any sport can be of vital importance, considering that betting in favor of an upset has dramatically larger payouts than betting in favor of a non-upset. However, because Tennis is not a team sport, it can be more difficult to accurately predict a winner and loser if not enough sufficient data is present.

The data that we gathered for this project came from a GitHub user: Jacob Gollub, containing 40,000 rounds of data from ATP Men's Single's Main Events; spanning the years 2000-2016. The data was recorded in relation to each round, meaning that we did not have individual data points for each game of each set, rather just the cumulation of all sets for that round. Our initial questions that we wanted to explore were: What are the most significant predictors of an upset, Do Upsets occur more often within a certain rank size, Do upsets occur more often in Best of 3 set matches or Best of 5 set matches? For more clarification, a Tennis match is comprised of either a Best of 3, or a Best of 5 matches, meaning that the winner will need to win 2 or 3 sets to win the match, respectively. A set is when a player wins 6 games with at least a lead of 2, with a game consisting of 4 points, having the winner win by at least a 2-point advantage.

Directly after omitting N/A and blank cells in each row, we decided to plot some graphs to look at the spread on the Winner's Rank of each match, as well as the Winner's Rank when an upset occurs. From Figure A-C, we can see that most of the time, the higher (lower number) ranked players are winning more often than higher ranked players; which stands true even after sub-setting the top 100 and 200 ranked players. In Figure D, we can see that when an upset

occurs, the winner tends to be higher ranked as well, however the outliers of upsets for players ranked above 500 should be noted. If we zoom in on the top 200 player's ranking of the winner when an upset occurs in Figure E, we can see that the distribution is more evenly spread, almost modeling a normal distribution. Taking an even closer look in Figure F, the spread of the winner's during an upset is very even across the entire subset. The distribution of the difference in ranking of the winner and loser is almost identical to that of the winner's ranking themselves, which would indicate that players only play others of similar rank; which can be seen in Figures H-J. From looking at these charts, we can see that the majority of the data is occurring in the ranks that are lower than 200; so, we decided to divide into three subsets: Top 200 players, Top 100 players, and Top 25 players.

After sub-setting, we calculated the percentage of upsets for each set. From Figure J we can note that all three subsets have equal proportions of sub setting, just over 1/3 of the time. If we can break down the data even further, we can see the average rank difference and percentage of an upset occurring in each match of the Top 100 and Top 25 subset, that went 2 through 5 Sets. From this chart we can see that as the match progresses in each subset, the proportion of upsets increases; as well as the average rank difference in the Top 100 subset. This indicates that as the match prolongs past the minimum amount of sets, the chance of an upset increases.

Finally, we ran a Logistic regression on each subset to determine which predictors were most significant. We used an additional integration of Bias Reduction through the BRGML2 package, because we did not have some vital match predictors/features, as well as players showing up multiple times for repeated entries in the data set. Using a bias reduced logistic regression will give us better predictors. Our initial predictors included: The Surface type: Hard, Clay, Grass, a Factor of the Best of 3 or Best of 5, The amount of Aces the of the Winner and Loser, the ranks of both players, the amount of sets the loser and win won in the match, as well as the amount of games each player won in each set. For the first subset of all ranked players, after backward selection, the remaining significant variables were: Winner's Rank, Loser's Rank, and the amount of games each player won in sets 1-3, as well as the amount of sets the Winner son. For the Top 100 players subset, after backward selection, the remaining significant variables were just the ranks of each player. Finally, for the Top 25 players subset, the remaining variables after backward selection were: the ranks of each player, as well as the amount of games the Winner won in set 1. These can all be noted in Figures: K-M. After dividing each subset into a 50%, 25%, 25% split for training, testing and validation, the accuracies for each subset are displayed in Figure N. The set for all players was only 66% accurate, whilst the subsets for the Top 100 and 25 players were: 96.83% and 97.14% accurate respectively. This could in part be due to those subsets containing much fewer data points.

From our findings, we could conclude that the Rank of a Player was always significant in predicting an upset of a match. This was expected, as the ranking for each player changes weekly based on immediate performance after tournaments. Additionally, the longer the match progresses past the minimum amount of sets, the higher chance of an upset occurring, and that the surface type and factor of the Best of 3 or 5 were both insignificant for all three subsets.

Eric Browne
6/2/2020
Probability and Statistics 2
Final Project Essay

Although our prediction rate was extremely high in the smaller subsets, there are some improvements to be made. These include having individual game data for each set of each match with much more quantifiable data on the players, such as: Age, Height, APE index, Average speed of serve, Average RPM of the ball on a serve, and Match history/records between each player. This would allow you to predict upsets of players as a given tournament progresses, which would in turn improve your overall winnings if placing bets.

Eric Browne
6/2/2020
Probability and Statistics 2
Final Project Essay

APPENDIX

x

Figure A

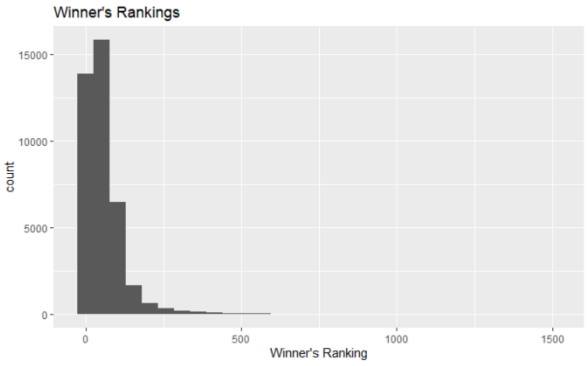


Figure B

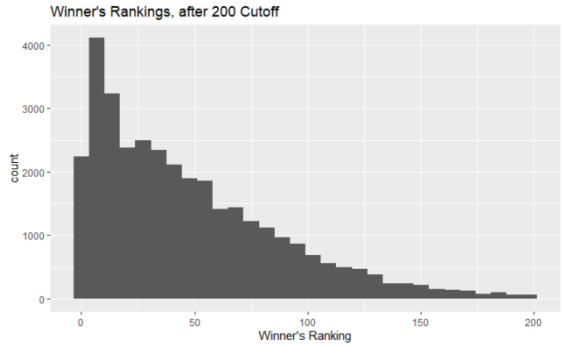


Figure C

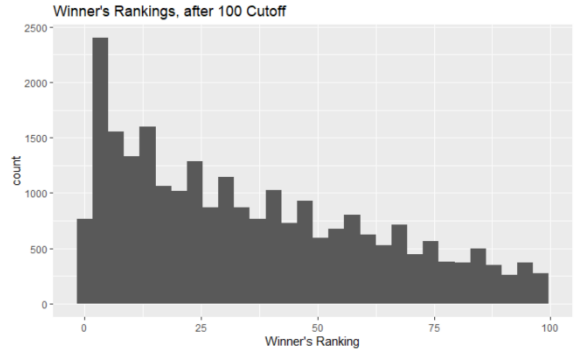
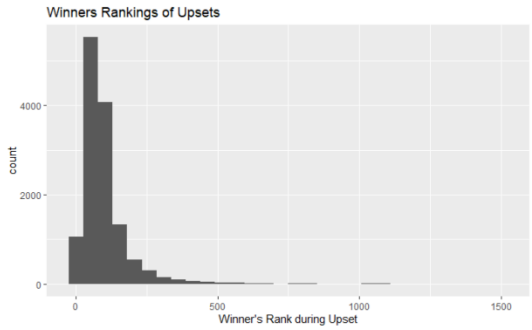


Figure D



Eric Browne
6/2/2020
Probability and Statistics 2
Final Project Essay

Figure E

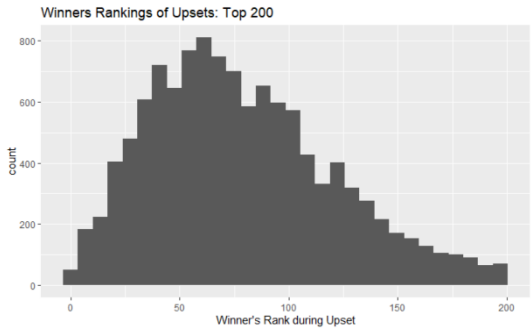


Figure F

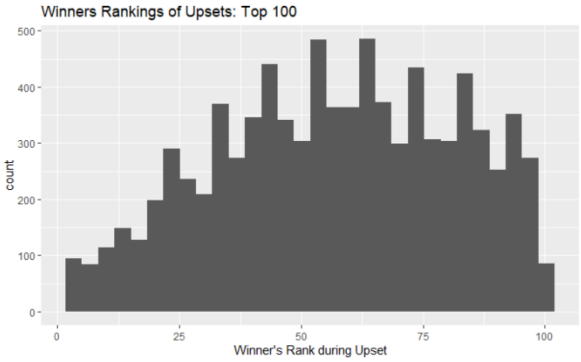


Figure G

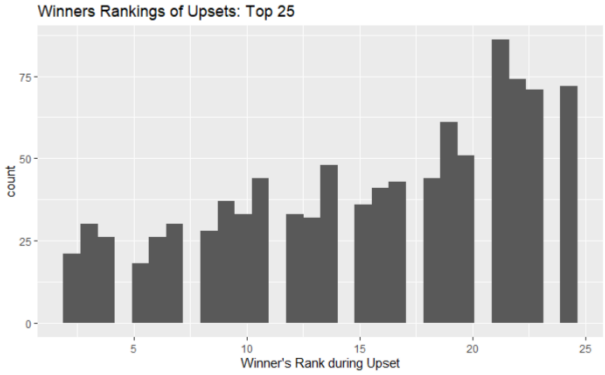


Figure H

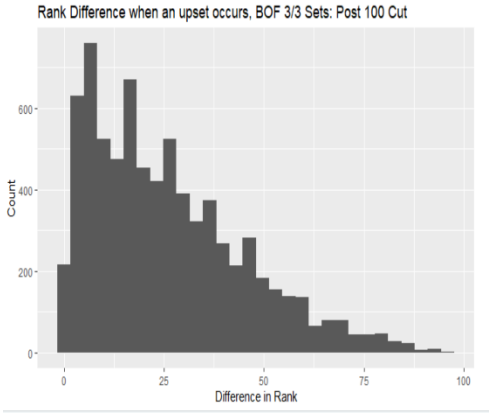


Figure I

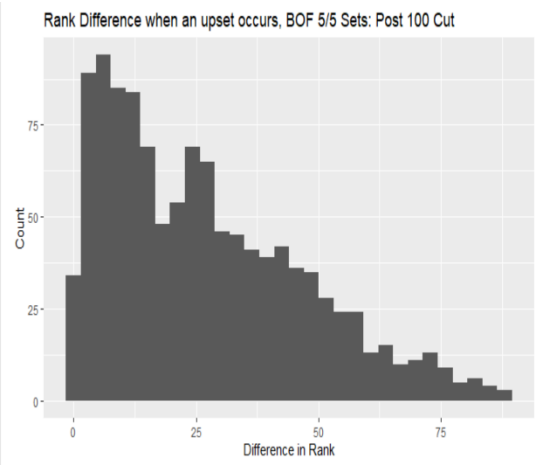


Figure J

All

	Percent of Upsets
All Ranks	0.340
Top 100 Ranks	0.351
Top 25 Ranks	0.343

Top 100

	Best of 3	Best of 5
2 Sets	24.843	---
3 Sets	26.975	25.047
4 Sets	---	25.819
5 Sets	---	29.712

	Best of 3	Best of 5
2 Sets	0.338	--
3 Sets	0.41	0.195
4 Sets	--	0.299
5 Sets	--	0.424

Top 25

	Best of 3	Best of 5
2 Sets	7.373	---
3 Sets	7.831	7.172
4 Sets	---	6.671
5 Sets	---	6.609

	Best of 3	Best of 5
2 Sets	0.316	--
3 Sets	0.426	0.194
4 Sets	--	0.35
5 Sets	--	0.387

Eric Browne
6/2/2020
Probability and Statistics 2
Final Project Essay

Figure(s) K.1,K.2

```
Call:
glm(formula = wnlse ~ factor(Surface) + factor(BestOf) + WAce +
    LAce + WRank + LRank + W1 + W2 + W3 + W4 + L1 + L2 + L3 +
    L4 + Wsets + Lsets, data = t.dat_train, method = "brglmFit")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7865  -0.4300   0.1875   0.2933   3.1541

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.599e+00  2.943e-02  54.336  < 2e-16 ***
factor(Surface)2 -1.002e-02  6.477e-03  -1.547 0.121795
factor(BestOf)5  -5.687e-03  1.929e-02  -0.295 0.768172
WAce             -1.307e-03  1.571e-03  -0.832 0.405473
LAce             -2.502e-03  3.835e-03  -0.652 0.514207
WRank            -3.122e-03  4.518e-05 -69.108  < 2e-16 ***
LRank             1.334e-03  2.671e-05  49.942  < 2e-16 ***
W1                8.338e-03  3.426e-03   2.434 0.014946 *
W2                1.292e-02  3.317e-03   3.894 9.85e-05 ***
W3                1.313e-02  3.194e-03   4.113 3.91e-05 ***
W4                6.369e-03  4.581e-03   1.390 0.164474
L1               -1.410e-02  2.086e-03  -6.758 1.40e-11 ***
L2               -1.018e-02  2.011e-03  -5.061 4.18e-07 ***
L3               -1.115e-02  2.870e-03  -3.884 0.000103 ***
L4               -6.439e-03  5.868e-03  -1.097 0.272509
Wsets             4.207e-02  1.546e-02   2.721 0.006499 **
Lsets            -1.866e-02  1.675e-02  -1.114 0.265248
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1542577)

    Null deviance: 3533.0  on 15689  degrees of freedom
Residual deviance: 2417.6  on 15673  degrees of freedom
  (72 observations deleted due to missingness)
AIC: 15218

Number of Fisher Scoring iterations: 1
```

```
Call:
glm(formula = wnlse ~ WRank + LRank + W1 + W2 + W3 + L1 + L2 +
    L3 + Wsets, data = t.dat_train, method = "brglmFit")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7875  -0.4320   0.1889   0.2927   3.1544

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.575e+00  2.475e-02  63.640  < 2e-16 ***
WRank       -3.128e-03  4.508e-05 -69.382  < 2e-16 ***
LRank        1.334e-03  2.669e-05  49.965  < 2e-16 ***
W1           9.843e-03  3.046e-03   3.231  0.00123 **
W2           1.392e-02  2.809e-03   4.954 7.25e-07 ***
W3           1.161e-02  2.197e-03   5.282 1.28e-07 ***
L1          -1.471e-02  1.933e-03  -7.612 2.71e-14 ***
L2          -1.078e-02  1.832e-03  -5.885 3.97e-09 ***
L3          -1.194e-02  2.684e-03  -4.448 8.67e-06 ***
Wsets        4.341e-02  9.390e-03   4.623 3.77e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1542929)

    Null deviance: 3533.5  on 15690  degrees of freedom
Residual deviance: 2419.4  on 15681  degrees of freedom
  (71 observations deleted due to missingness)
AIC: 15216

Number of Fisher Scoring iterations: 1
```

Figure(s) L.1, L.2

```
Call:
glm(formula = wnlse ~ factor(Surface) + factor(BestOf) + WAce +
    LAce + WRank + LRank + W1 + W2 + W3 + W4 + L1 + L2 + L3 +
    L4 + Wsets + Lsets, data = t.dat100l_train, method = "brglmFit")

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-0.60347  -0.24114   0.01672   0.24642   1.14442

Coefficients:
                  Estimate Std. Error  z value Pr(>|z|)
(Intercept)       6.124e-01  2.827e-02   21.662  < 2e-16 ***
factor(Surface)2  1.267e-04  6.047e-03    0.021  0.98329
factor(BestOf)5  -1.103e-03  1.826e-02   -0.060  0.95182
WAce              2.811e-06  2.166e-04    0.013  0.98964
LAce              1.041e-04  1.969e-04    0.529  0.59714
WRank            -1.164e-02  1.109e-04 -104.985  < 2e-16 ***
LRank             9.841e-03  1.100e-04   89.429  < 2e-16 ***
W1                3.664e-03  3.187e-03    1.150  0.25018
W2                7.513e-03  3.082e-03    2.438  0.01478 *
W3                7.553e-03  2.949e-03    2.561  0.01043 *
W4                1.162e-02  4.129e-03    2.815  0.00488 **
L1               -2.423e-03  1.912e-03   -1.267  0.20502
L2                7.480e-04  1.858e-03    0.403  0.68721
L3               -2.932e-03  2.632e-03   -1.114  0.26539
L4                8.280e-04  5.394e-03    0.153  0.87801
Wsets            -4.405e-02  1.482e-02   -2.972  0.00295 **
Lsets            -2.035e-02  1.552e-02   -1.311  0.18974
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.08248795)

    Null deviance: 2258.23  on 9916  degrees of freedom
Residual deviance:  816.58  on 9900  degrees of freedom
  (57 observations deleted due to missingness)
AIC: 3417.6

Number of Fisher Scoring iterations: 1
```

```
Call:
glm(formula = wnlse ~ WRank + LRank, data = t.dat100l_train,
    method = "brglmFit")

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-0.57758  -0.24076   0.01734   0.24655   1.18065

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.5910211  0.0066826   88.44   <2e-16 ***
WRank       -0.0116511  0.0001085 -107.39   <2e-16 ***
LRank        0.0098582  0.0001081   91.21   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.08253407)

    Null deviance: 2273.00  on 9973  degrees of freedom
Residual deviance:  822.89  on 9971  degrees of freedom
AIC: 3428.7

Number of Fisher Scoring iterations: 1
```

Eric Browne
6/2/2020
Probability and Statistics 2
Final Project Essay

Figure(s) M.1, M.2

```
Call:
glm(formula = wnlse ~ factor(Surface) + WAce + LAce + WRank +
    LRank + W1 + W2 + W3 + W4 + L1 + L2 + L3 + L4 + Wsets + Lsets,
    data = top25_train, method = "brglmFit")

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.60973  -0.20777   0.00938   0.21721   0.59495

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.4816964  0.0657267   7.329 2.32e-13 ***
factor(Surface)2 -0.0015668  0.0172516  -0.091   0.9276
WAce             -0.0006450  0.0006030  -1.070   0.2847
LAce              0.0002926  0.0005472   0.535   0.5929
WRank            -0.0449443  0.0011523 -39.003  < 2e-16 ***
LRank             0.0385103  0.0011786  32.675  < 2e-16 ***
W1                0.0184643  0.0087711   2.105   0.0353 *
W2               -0.0061053  0.0081599  -0.748   0.4543
W3                0.0096812  0.0065519   1.478   0.1395
W4               -0.0017431  0.0095652  -0.182   0.8554
L1               -0.0010983  0.0054240  -0.202   0.8395
L2                0.0003329  0.0052662   0.063   0.9496
L3               -0.0091249  0.0067166  -1.359   0.1743
L4               -0.0002999  0.0123843  -0.024   0.9807
Wsets             0.0073582  0.0338816   0.217   0.8281
Lsets             0.0337255  0.0320261   1.053   0.2923
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.06972692)

    Null deviance: 255.167  on 1122  degrees of freedom
Residual deviance:  77.141  on 1107  degrees of freedom
  (17 observations deleted due to missingness)
AIC: 213.41

Number of Fisher Scoring iterations: 2
```

```
Call:
glm(formula = wnlse ~ WRank + LRank + W1, data = top25_train,
    method = "brglmFit")

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.55320  -0.20393   0.00087   0.22396   0.57428

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.522419   0.036965  14.133   <2e-16 ***
WRank       -0.044920   0.001110 -40.467   <2e-16 ***
LRank        0.037946   0.001156  32.830   <2e-16 ***
W1           0.011810   0.005610   2.105   0.0353 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.07001515)

    Null deviance: 258.746  on 1139  degrees of freedom
Residual deviance:  79.491  on 1136  degrees of freedom
AIC: 209.2

Number of Fisher Scoring iterations: 1
```

Figure N

ALL PLAYERS	TOP 100 RANKED PLAYERS	TOP 25 RANKED PLAYERS
prediction FALSE TRUE 0 82 5214 1 0 10372	prediction FALSE TRUE 0 3262 223 1 90 6330	prediction FALSE TRUE 0 364 27 1 6 760
66.72% Accurate	**96.83%** Accurate	**97.14%** Accurate