March Madness Predictions

Author: Hayden McAlpin

December 13th, 2020

Introduction

The NCAA organizes one of the biggest sporting events in the country every year known as March Madness, the college basketball tournament to declare a national champion. It is the NCAA's more profitable event. As reported by MSN, in the 2018-2019 season the organization brought in $867.5 million from just the television and marketing rights and another $177.9 million from ticket sales (McGee 2020). One of the most popular challenges that coincide with the tournament is the bracket challenge. Participants are asked to select winners in each game and the person with the most correct bracket wins. Organizations like ESPN and CBS allow people to create their own competition groups to compete against their friends and families. In 2019, the challenges hosted by ESPN had over 17.2 million entries (Ota 2019). The popularity and extreme unfavorable odds of selecting a perfect bracket (nearly 1 in 9.2 quintillion) lead to some extravagant prize possibilities. In the same year, Warren Buffet offered a grand prize of $1 Billion dollars to any fan who could predict and generate the perfect bracket (Bloomenthal 2020). Now the casual fan or participant may just select winners based on their gut feeling or the seed ranking set by the NCAA., but what if there was a way to statistically analyze games from the past then there could be some substance behind the decisions made. Kaggle has hosted their own competition for this event, challenging contestants to create a statistical model to calculate and predict win probabilities as well as select winners of each possible game. Additionally, it asks them to use this model to generate a bracket of predicted outcomes. The website supplies contestants with data files filled with regular season and post season statistics from games of years past as well as the seeds for teams in the upcoming tournament. This report will dive into the competition, data supplied by Kaggle and the processes used to solve this problem.

Competition Overview

The competition is split into two parts:

1) Use the stats from years 2012 and prior to predict the known outcomes in years 2013 – 2016. Giving the participants the chance to build and test their model.

2) Use the model and all previous data to predict the outcomes of the 2017 tournament.

The submissions are lists of possible tournament matchups along with the probability that team 1 (team with the lower ID number) will win the game. The submissions will be evaluated using log loss:

$$\text{LogLoss} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right],$$

where

- n is the number of games played
- $\hat{y}_i$ is the predicted probability of team 1 beating team 2
- y_i is 1 if team 1 wins, 0 if team 2 wins
- $log()$ is the natural (base e) logarithm

Data

The competition allows for the use of several different data files but only those used to create the model are detailed here:

RegularSeasonDetailedResults.csv :

This file contains recorded statistics from years 2003-2017 of regular season games for teams in the tournament those years

COLUMNS

Season – year of season
Daynum – day number of season the game occured
(W/L)team – id of designated team
(W/L)score – score of designated team
Wloc – location of game relative to winning team (H = home, A = away, N = neutral)
Numot – number of overtimes in the game
(W/L)fgm – field goals made by designated team
(W/L)fga – field goals attempted by designated team
(W/L)fgm3 – 3pt field goals made by designated team
(W/L)fga3 – 3pt field goals attempted by designated team
(W/L)ftm – free throws made by designated team
(W/L)fta – free throws attempted by designated team
(W/L)or – offensive rebounds by designated team
(W/L)dr – defensive rebounds by designated team
(W/L)ast – assists by designated team
(W/L)to – turnovers by designated team
(W/L)stl – steals by designated team
(W/L)blk – blocks by designated team
(W/L)pf – personal fouls by designated team

TourneyDetailedResults.csv:

This file contains the same information as the *RegularSeasonDetailedResults.csv* file except the stats are from the march madness tournament games. (Same columns)

TourneyCompactResults.csv :

This file contains the basic information from the tournament games played

COLUMNS

Season – year of season

Daynum – day number of season the game occured

(W/L)team – id of designated team

(W/L)score – score of designated team

Wloc – location of game relative to winning team (H = home, A = away, N = neutral)

Numot – number of overtimes in the game

TourneySeeds.csv :

This file contains the assigned seeds and region to the teams in the tournament

COLUMNS

Season – year of season

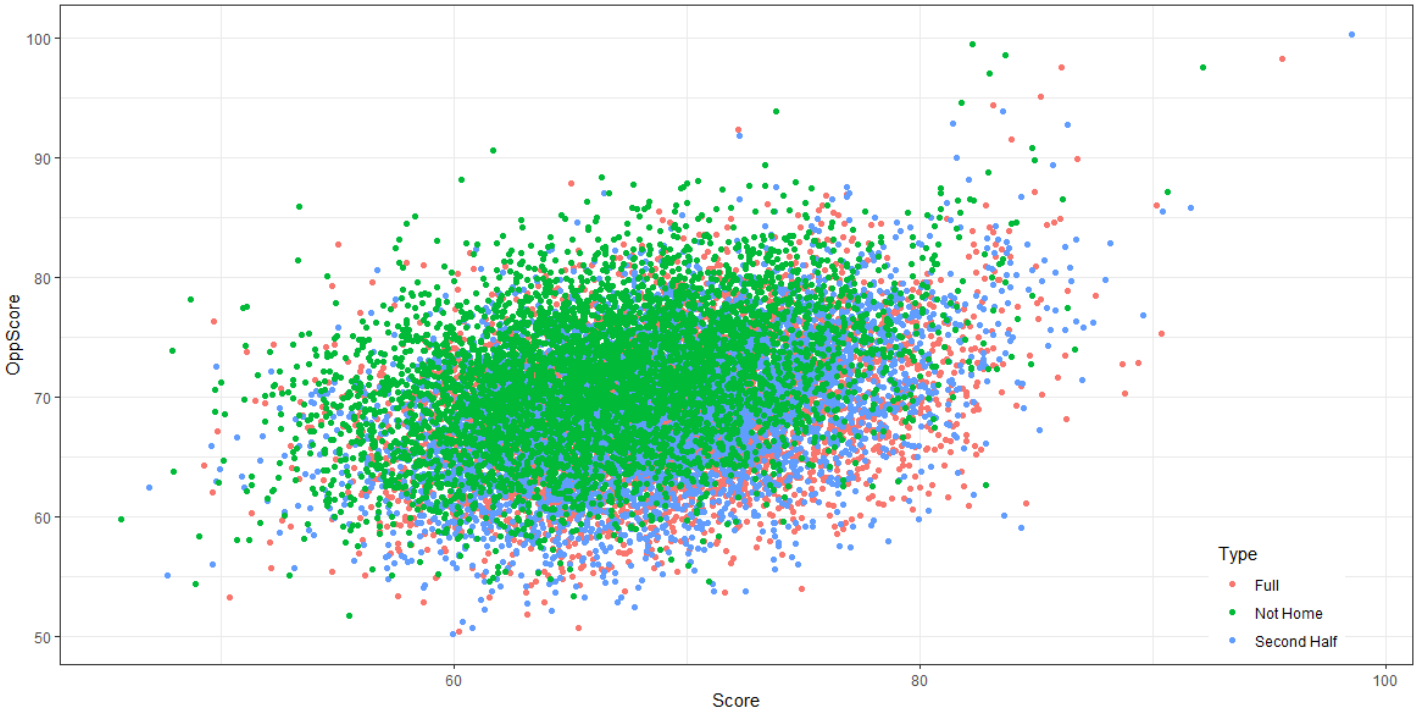
Seed – region and seed number

Team – team id

Data Preparation

Firstly, the file containing the regular season stats was split into two data frames representing winner stats and loser stats. The only column that represented wrong info was the Wloc column in the loser stats. This represented the location of the game relative to the winning team so the entries must be switched to reflect the location relative to the losing team. A win/loss variable was added to both data frames with a 1 representing a win and 0 a loss. The data was then combined based on the season and team id. New columns representing field goal percentage and 3-pt percentage were added to calculate the team's efficiencies in those areas for each game. Season averages on all stats were then calculated to help evaluate team success throughout the year (This includes opponent's points scored, average shot percentages and the team's win percentage on the year. The opponent's points scored give a good indicator on the level of defense is team can play). Columns were then renamed to more commonly used abbreviations for the stats.

College basketball regular seasons typically begin in early November and run until early March. Teams can have anywhere around 30 games for their season. With this stretch of time and number of games, teams typically go through some development and improvement as the season goes on. How teams compete in the beginning of the season could end up being drastically different than how they compete towards the end of the season. For example, in the 2015-2016 season, the Oklahoma Sooners and the Villanova Wildcats met early in the season in November where the Sooners won the matchup 78 – 55. They would then go on to meet again in the Final Four of the March Madness tournament where Villanova got their revenge and won that meeting with a score of 95 – 51. To better evaluate how a team may compete in the tournament, the data was also averaged into two separate other data frames. The second data frame represented the team's averages for the second half of the season. The third data frame was used to store averages of the teams away or neutral site games.



The above graph depicts the relationship between two of the more important variables (Points Scored and Opponent Points Scored) and how they change between the datasets. As the graph shows there does not seem to be much change in value. The following tables depict a more direct numerical comparison. The average points scored, opponent points scored, and win percentage were calculated for the various data frames.

	Full Season	**Second Half of Season**	**Not Home Games**
Points Scored	68.911	68.748	66.591
Opponents Points Scored	69.127	68.962	71.301
Win Percentage	49.399 %	49.288 %	36.660 %

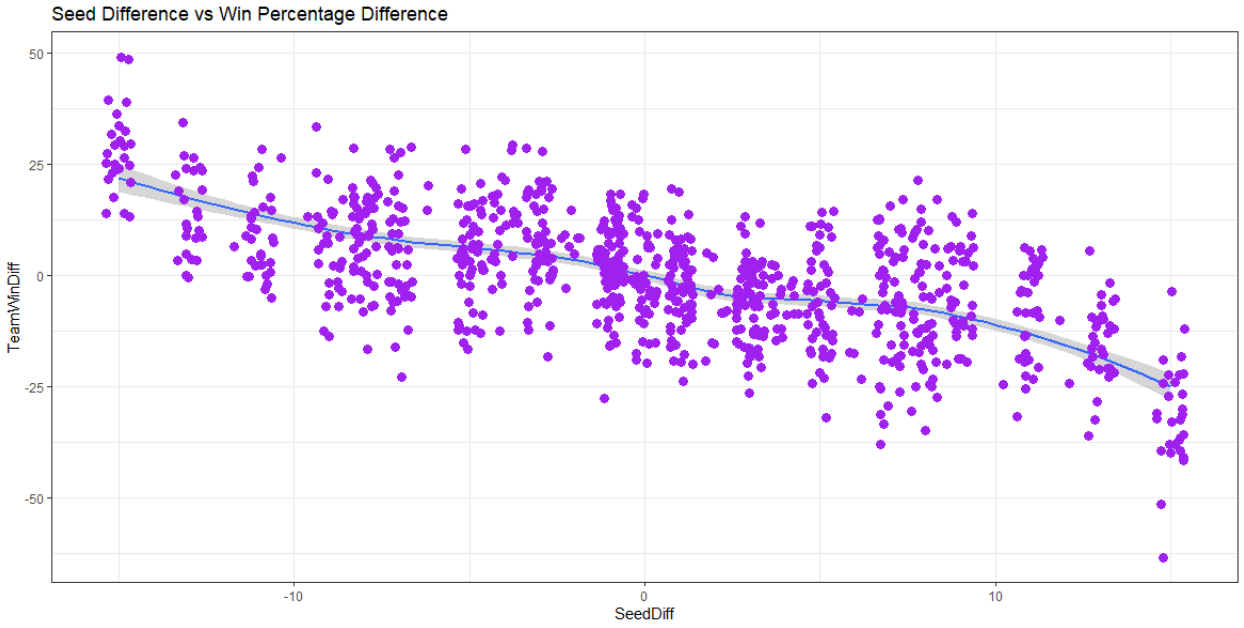
When comparing the full season statistics to the second half of the season there is not much difference. The only major change is the win percentage for away from home games vs winning percentage for all games. This may be worth looking into further in testing. Although there does not seem to be much difference between the 3 data frames, they will still be used for modeling purposes to see if they result in any significantly different models.

Another point to consider is how seeds typically perform against one another. For example, until UMBC upset Virginia in the first round of the 2018 tournament a 16 seed had never beaten a 1 seed. To assign the associated seeds and stats, the seed value in the tournament seeds file was split into two variables region and seed so that the actual seed value could be used. As the region only indicates the locations of the games but the actual seed values are important for comparing the teams. The train data frames used for testing were then created by merging the compact tournament stats for that season team id's, the score for the games, and the win\loss column with each team's averages for that season from the previously formed train data frames. The statistical averages on how individual seeds perform in the tournament were also calculated and merged into the data frame. The individual team statistics would not give too much inference as to who would win the game but rather how the two teams compare in those areas. The statistical differences between the two teams as well as the statistical differences between their respective seeds (Note: the percentage stats were converted to actual percentages as opposed to decimals to increase importance of their differences) were calculated as well.

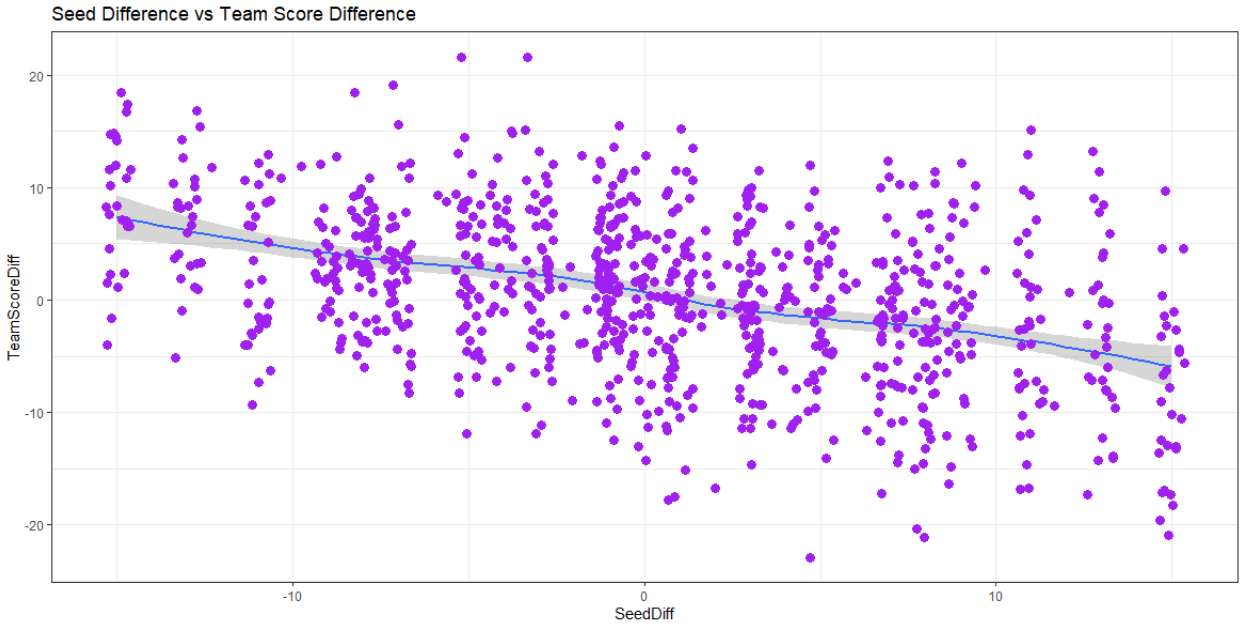
Modeling

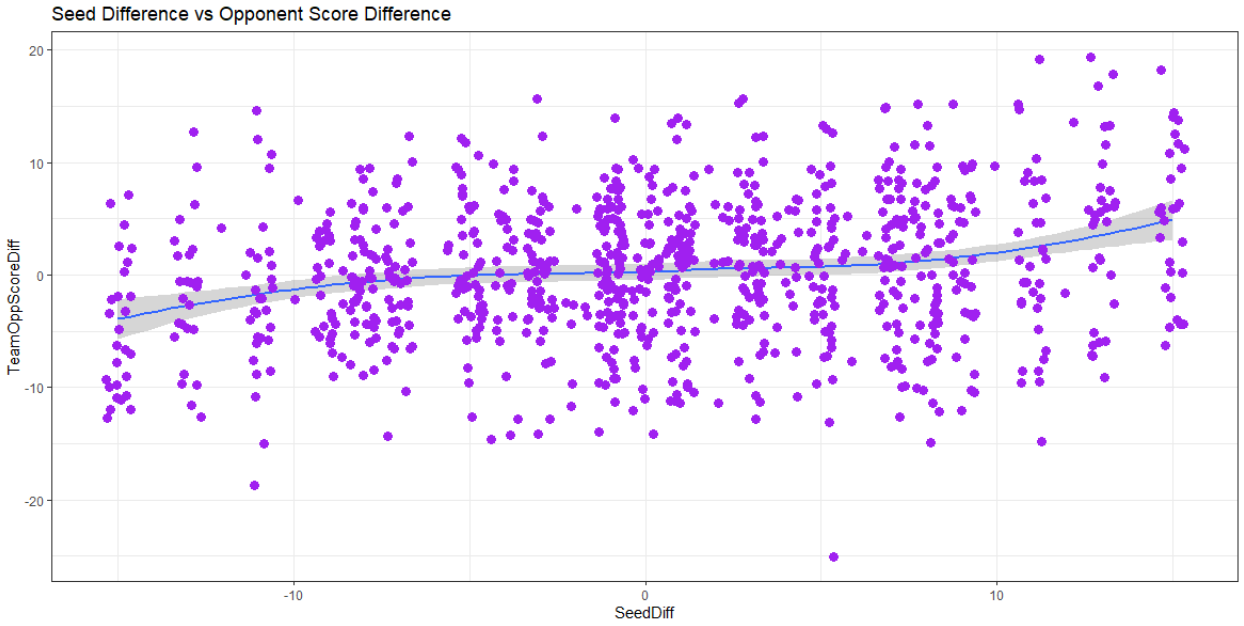
The overall success of the models will be determined in how well they predicted the outcome of the games in the last four seasons for part one and the successful prediction for the 2017 tournament (Note: the Log Loss will also be tracked for the competition purposes). The modeling will be done using 3 different models: generalized linear modeling from the MASS library in R, earth model from the caret library in R, and a neural network in R. The results will be compared by their log loss and actual accuracy to determine the best one.

Selecting the right combination of variables will be important in developing the best model to predict game outcomes. It would be anticipated that seed difference and win percentage difference would be the best indicators as to who would win the matchup. The following graphs show the relationships between these two variables and a few more:

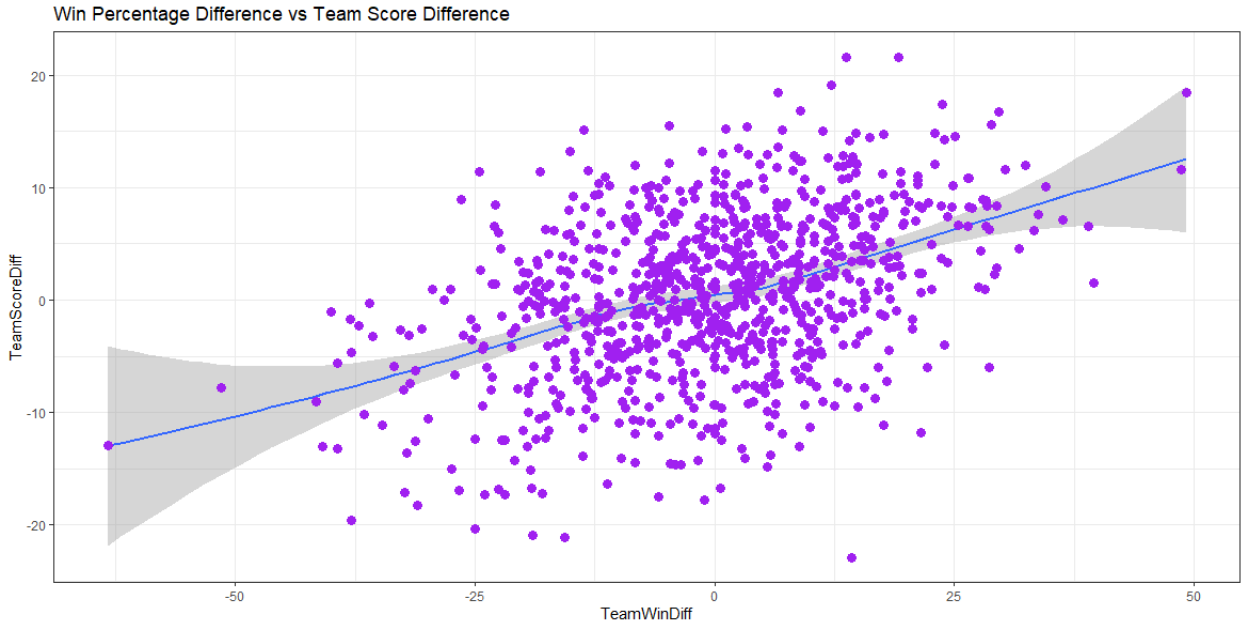


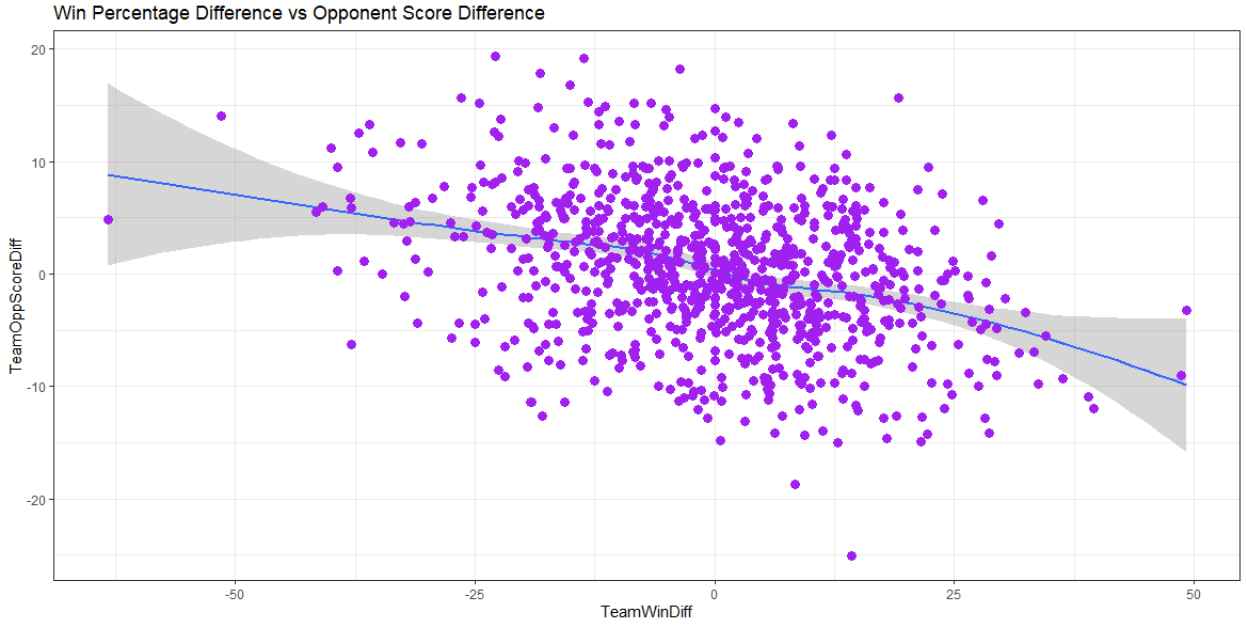
This shows that there is some correlation to how the teams are seeded and how the teams had performed in terms of win percentage in the regular season. This would make sense since ESPN does their own analysis and rankings to determine their seeds for the tournament. A negative seed difference would indicate that team 1 is the higher seeded team in the matchup and therefore it would be expected that their win percentage would be higher than their counterparts.





These two graphs show the relationship between the difference in the team's seeds against how many points they score and how many points they allow their opponents to score. There seems to be a very slight negative correlation in seed difference versus points scored but there does not seem to be much correlation with hoe much they allow their opponents to score.





These two graphs display the relationship between win percentage difference against points scored and opponents points scored. Team's with a higher win percentage (positive difference) tend to score more points (positive difference) and hold their opponents to fewer points (negative difference).

The biggest question when predicting the victor for a given game is how probable the upset is. This is where filling out a perfect bracket becomes so difficult. Just like the UMBC

victory over Virginia mentioned earlier, upsets can happen at any time and against any team. When these upsets are likely to happen could be a helpful predictor.

This graph shows when upsets occur with the size of each point indicating the frequency of how many times that upset has occurred. Most of the upsets are more likely to happen in the early rounds of the tournament and less likely in the later rounds. Which makes sense as there are more opportunities for the upset to happen in the earlier rounds with more teams. This shows that to predict the upset it is important to consider the day or round the game occurs.

Results

To get a baseline on how successful the model and predictions need to be, it is important to look at how correct a model would be if in every game the team with the higher seed won the game. If this were the case, the model would be 72.757 % correct. This is the minimum success rate the models need to be to show improvement.

The following table shows the results from various tests using every difference variable at disposal and their specific function used.

Model (Hyperparameters)	Data Frame Used	Log Loss	Actual Accuracy
Generalized Linear Modeling (NA)	Full season	1.385099	76.866 %
Generalized Linear Modeling (NA)	2nd Half Season	1.889052	76.866 %
Generalized Linear Modeling (NA)	Away From Home Games	2.039939	76.866 %
Earth (nprune = 3, degree = 1)	Full season	0.6928497	67.910 %
Earth (nprune = 3, degree = 2)	2nd Half Season	0.5500565	67.910 %
Earth (nprune = 2, degree = 2)	Away From Home Games	0.5483479	68.656 %
Neural Network (size = 3, decay = 0.1)	Full season	0.6193695	66.418 %
Neural Network (size = 3, decay = 0.1)	2nd Half Season	0.6077092	66.791 %
Neural Network (size = 3, decay = 0.1)	Away From Home Games	0.6264596	72.761 %

In terms of overall accuracy, the GLM model performed the best and gave the most accurate results. In terms of log loss, the Earth model was the best. As expected in most models, the data frame used did not make much of a difference in terms of actual accuracy except for the neural network model. In this model, the data from neutral or away site games gave more accurate predictions.

The GLM model already has a more accurate prediction than just trusting the seeds so this will be the model to focus on. The following is the summary of the most accurate GLM model from the previous table:

```
Call:
NULL
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7871  -0.8458   0.0000   0.8901   2.2676

Coefficients: (6 not defined because of singularities)
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -8.228e+00  3.103e+00  -2.651  0.00802 **
SeedDiff         6.720e+00  9.294e+05   0.000  0.99999
TeamScoreDiff    2.517e-01  1.081e-01   2.327  0.01996 *
TeamOppScoreDiff -1.254e-01  4.890e-02  -2.564  0.01036 *
TeamOTDiff      -2.627e-02  1.478e-02  -1.777  0.07565 .
TeamFgmDiff     -4.671e-01  6.316e-01  -0.739  0.45961
TeamFgaDiff      1.354e-01  2.967e-01   0.457  0.64799
Team3pmDiff      6.650e-01  5.435e-01   1.223  0.22115
Team3paDiff     -2.928e-01  1.920e-01  -1.525  0.12728
TeamFtmDiff             NA         NA      NA       NA
TeamFtaDiff     -1.199e-01  8.496e-02  -1.411  0.15826
TeamOrebDiff     3.826e-02  9.983e-02   0.383  0.70154
TeamDrebDiff    -8.399e-02  7.571e-02  -1.109  0.26727
TeamAstDiff     -9.935e-02  6.240e-02  -1.592  0.11135
TeamTODiff       1.028e-02  9.317e-02   0.110  0.91215
TeamStlDiff     -3.900e-03  9.387e-02  -0.042  0.96686
TeamBlkDiff     -2.748e-03  6.825e-02  -0.040  0.96788
TeamPFDiff      -2.821e-02  5.423e-02  -0.520  0.60292
TeamWinDiff     -2.008e-02  1.392e-02  -1.443  0.14903
TeamFGperbDiff   1.426e-01  3.281e-01   0.435  0.66383
Team3PTperDiff  -1.847e-01  9.310e-02  -1.984  0.04730 *
SeedScoreDiff    2.434e+01  3.332e+06   0.000  0.99999
SeedOppScoreDiff -1.280e+01  1.693e+06   0.000  0.99999
SeedOTDiff      -3.417e+00  4.568e+05   0.000  0.99999
SeedFgmDiff     -4.521e+01  6.155e+06   0.000  0.99999
SeedFgaDiff     -1.868e+01  2.567e+06   0.000  0.99999
Seed3pmDiff     -1.616e+01  2.205e+06   0.000  0.99999
Seed3paDiff      1.962e+01  2.645e+06   0.000  0.99999
SeedFtmDiff             NA         NA      NA       NA
SeedFtaDiff     -1.241e+01  1.737e+06   0.000  0.99999
SeedOrebDiff     4.219e+01  5.706e+06   0.000  0.99999
SeedDrebDiff    -2.275e+01  3.100e+06   0.000  0.99999
SeedAstDiff     -3.804e+01  5.097e+06   0.000  0.99999
```

```
SeedTODiff       -2.650e+01  3.585e+06    0.000  0.99999
SeedStlDiff       9.563e+00  1.431e+06    0.000  0.99999
SeedBlkDiff       7.432e+01  1.010e+07    0.000  0.99999
SeedPFDiff               NA         NA       NA       NA
SeedWinDiff              NA         NA       NA       NA
SeedFGperbDiff           NA         NA       NA       NA
Seed3PTperDiff           NA         NA       NA       NA
Daynum            5.894e-02  2.224e-02    2.650  0.00805 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 895.15  on 645  degrees of freedom
Residual deviance: 641.17  on 611  degrees of freedom
AIC: 711.17

Number of Fisher Scoring iterations: 19
```

As expected, there was some over fitting issues and some variables (with NA values) had irrelevant information for the model. Therefore, the variables need to be trimmed down to find the best model for predictions. After further testing the optimal combination of variables for the GLM model were obtained and ran for every model. The following table displays those results:

Model (Hyperparameters)	**Data Frame Used**	**Log Loss**	**Actual Accuracy**
Generalized Linear Modeling (NA)	Full season	0.5605082	77.612 %
Generalized Linear Modeling (NA)	2nd Half Season	0.5555359	77.612 %
Generalized Linear Modeling (NA)	Away From Home Games	0.5529897	77.612 %
Earth (nprune = 2, degree = 1)	Full season	0.5607149	68.657 %
Earth (nprune = 4, degree = 1)	2nd Half Season	0.5581215	69.403 %
Earth (nprune = 4, degree = 1)	Away From Home Games	0.5584997	69.403 %
Neural Network (size = 1, decay = 0.1)	Full season	0.6218740	67.910 %
Neural Network (size = 1, decay = 0.1)	2nd Half Season	0.6254738	66.045 %
Neural Network (size = 3, decay = 0.1)	Away From Home Games	0.6166107	63.806 %

Again, the GLM model gave the most accurate results and gave the best log loss value. With the right combination of variables, the log loss value for the GLM model was greatly improved. Since the actual accuracy was the same for all three data frames, the best log loss value will determine the best results for the problem. Therefore, by the slightest of margins, the data from neutral site or away games turned out to be the best predictor for March Madness victories. The following is the summary of the best GLM model:

```
Call:
NULL
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.78876  -0.85186  -0.00437   0.89054   2.27230

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -8.20748    3.09686  -2.650  0.00804 **
SeedDiff         0.60841    0.69207   0.879  0.37934
TeamScoreDiff    0.25014    0.10488   2.385  0.01708 *
TeamOppScoreDiff -0.12265   0.04038  -3.037  0.00239 **
TeamOTDiff      -0.02639    0.01472  -1.793  0.07298 .
TeamFgmDiff     -0.45488    0.62189  -0.731  0.46450
TeamFgaDiff      0.12563    0.28058   0.448  0.65432
Team3pmDiff      0.66307    0.54290   1.221  0.22196
```

```
Team3paDiff       -0.29182    0.19179  -1.522  0.12812
TeamFtaDiff       -0.12078    0.08072  -1.496  0.13460
TeamOrebDiff       0.04424    0.07853   0.563  0.57314
TeamDrebDiff      -0.07993    0.05055  -1.581  0.11383
TeamAstDiff       -0.09881    0.06172  -1.601  0.10940
TeamPFDiff        -0.02471    0.04669  -0.529  0.59657
TeamWinDiff       -0.01988    0.01366  -1.455  0.14571
TeamFGperbDiff     0.13873    0.32510   0.427  0.66958
Team3PTperDiff    -0.18462    0.09302  -1.985  0.04718 *
SeedScoreDiff      2.18822    2.54921   0.858  0.39068
SeedOppScoreDiff -1.54574    1.32273  -1.169  0.24257
SeedOTDiff        -0.42253    0.40585  -1.041  0.29783
SeedFgmDiff       -4.52925    5.35144  -0.846  0.39735
SeedFgaDiff       -1.49001    1.25605  -1.186  0.23552
Seed3paDiff        1.70022    1.63030   1.043  0.29700
SeedFtaDiff       -0.80892    1.00096  -0.808  0.41901
SeedOrebDiff       4.14568    3.41694   1.213  0.22503
SeedDrebDiff      -1.82453    1.94443  -0.938  0.34807
SeedAstDiff       -4.61522    4.10516  -1.124  0.26091
SeedTODiff        -3.02267    2.54001  -1.190  0.23404
SeedBlkDiff        7.49692    7.67589   0.977  0.32873
Daynum             0.05880    0.02220   2.649  0.00807 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 895.15  on 645  degrees of freedom
Residual deviance: 641.20  on 616  degrees of freedom
AIC: 701.2

Number of Fisher Scoring iterations: 9
```

The variables with the best predictive ability are the team’s score difference, team’s opponent’s score difference, team 3-pt percentage difference, and the day the game occurs. It is somewhat surprising that the seed difference was not one of the best variables but three of the 4 were as expected. The interesting variable is the 3-pt percentage difference. This does make sense though as 3-pt shots are worth points and if a team can make them more efficiently than their opponent than it would give them an advantage. The points scored and opponent’s points lead directly into how a game is won so they would be two of the best variables or the model.

Using the GLM model and the stats for away from home games, the resulting predictions for part 2 of the competition are presented in the form of a bracket in the Appendix. The results were 70.149 % correct. Due to the nature of the competition the results of the games were unknown at the time of submission so the model could only predict the outcome of matchups it had previously predicted. The model produced ¾ of the Final Four teams correctly but did not predict the correct winners to go to the national championship game.

Conclusion:

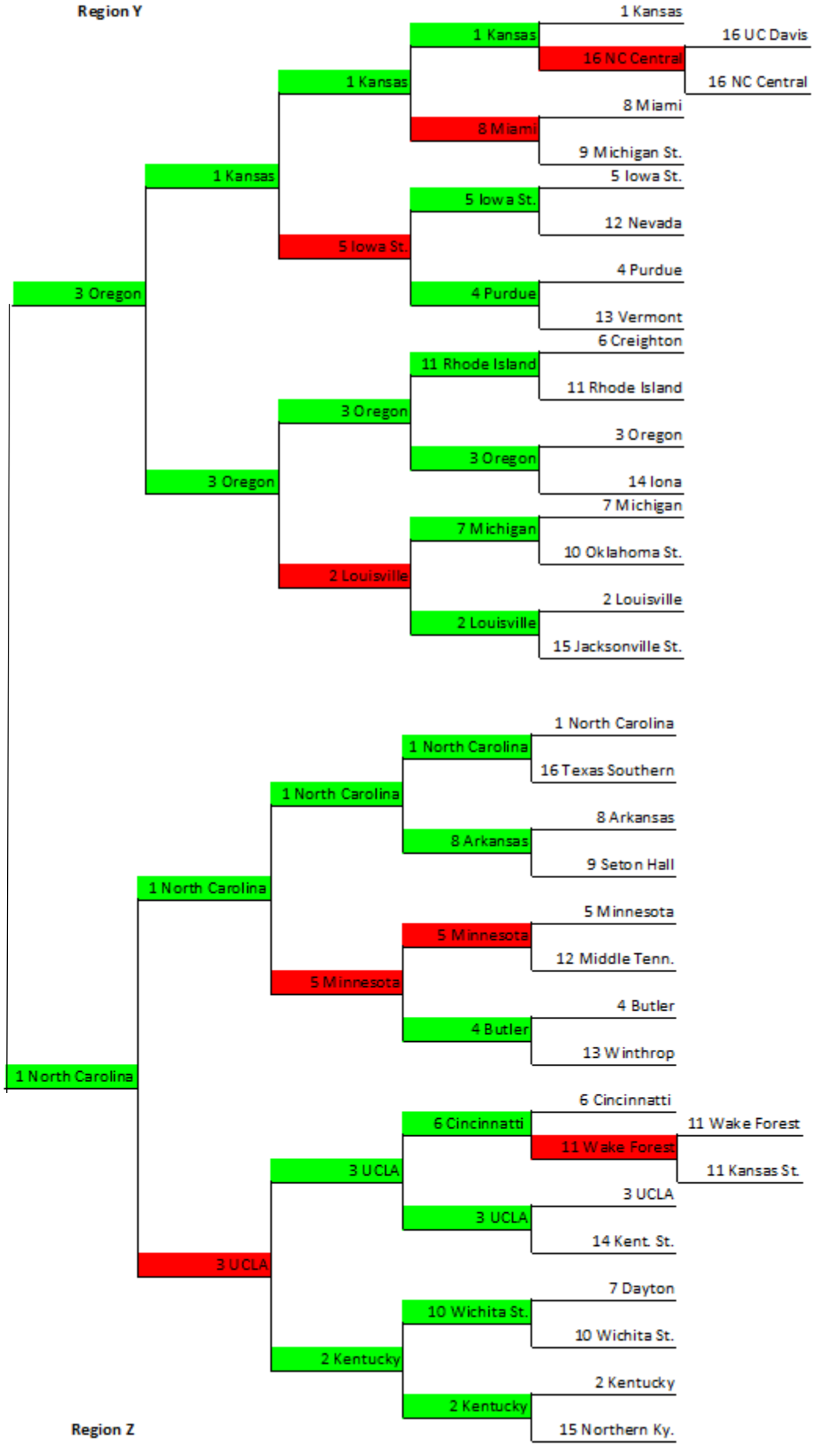
Filling out brackets for March Madness is a fun activity to get fans and family involved in the tournament. It can be hard to predict the perfect bracket but with the help of some statical analysis and modeling the decisions can be made with some evidence to support them. The generalized linear modeling model proved to produce the most accurate results from part one of the competition. This was then used for part two. The overall accuracy of the model was 70.149 %. To get a more accurate predictor more stats could be useful. Looking into subjects like team experience, level of talent on teams, and coaching experience could serve to further understand and predict a team's performance in the tournament. Unfortunately, there was not any readily available data on these subjects. In sports, things are not always straightly analytical or predictable. The games are played by humans and they have inconsistencies that can be difficult to predict. Some games were decided by 1 point and some are decided by a last second shot to beat the buzzer. The analytics can give a good insight to predicting winners in the March Madness games, but they should not be the only factor in determining winners.

Appendix

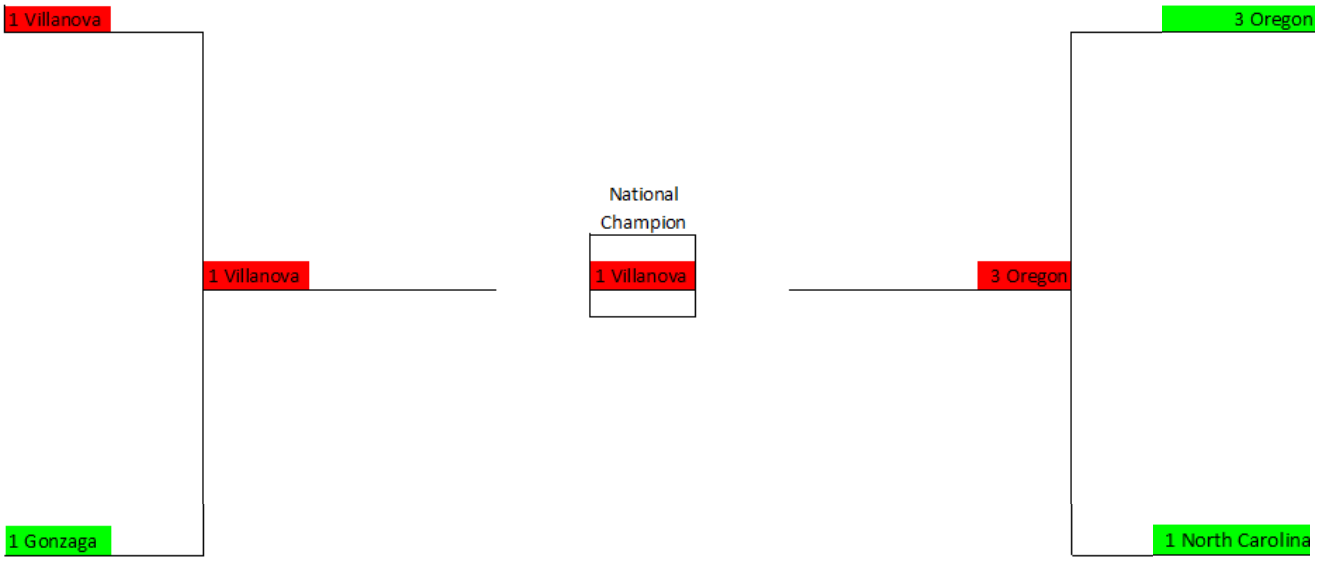
Part 1 Results

Team1	Team2	Win
1181	1458	3.578659e-01
1301	1437	1.857653e-01
1157	1438	1.361197e-04
1112	1458	4.110460e-01
1277	1438	2.898460e-01
1393	1438	1.671358e-01
1107	1196	1.775681e-04
1181	1257	3.666070e-01
1328	1332	4.372389e-01
1231	1314	3.561972e-01
1246	1455	1.968010e-01
1214	1246	3.518561e-05
1214	1438	4.989481e-05
1235	1438	2.056891e-01
1163	1242	2.920704e-01
1173	1196	1.719296e-01
1272	1438	1.633506e-01
1246	1458	5.216603e-01
1157	1458	6.278502e-05
1195	1314	4.545032e-05
1153	1246	9.979504e-02
1221	1332	4.481094e-05
1332	1458	6.149698e-02
1314	1458	1.462816e-01
1139	1438	1.807913e-01
1163	1196	3.515356e-01
1142	1455	1.274986e-04
1248	1437	8.770228e-05
1161	1257	1.706623e-01
1181	1332	3.294717e-01
1122	1242	7.875832e-05
1167	1328	5.517656e-02
1166	1181	3.726046e-01
1228	1274	3.873988e-01
1112	1458	5.855141e-01
1246	1458	2.381827e-01
1124	1458	2.072222e-01
1314	1437	5.851207e-01
1163	1437	5.123729e-01
1110	1458	4.293459e-02
1266	1274	4.231837e-01
1234	1437	3.246786e-01
1328	1437	4.096910e-01
1458	1462	3.739131e-01
1421	1437	4.980222e-02
1332	1458	2.741426e-01
1181	1211	5.961251e-01
1277	1438	1.453848e-01
1233	1326	3.492679e-02
1107	1181	4.811279e-02
1451	1462	2.966040e-02
1125	1438	2.705746e-02
1274	1437	4.131391e-01
1195	1207	8.570416e-02
1112	1326	2.908549e-01
1246	1276	2.134539e-01
1242	1437	3.562448e-01
1235	1326	2.054259e-01
1184	1242	7.216859e-02
1328	1401	5.168248e-01
1372	1452	8.529797e-02
1139	1266	2.518809e-01
1272	1277	1.787384e-01
1124	1166	1.760714e-01
1277	1328	1.629212e-01
1173	1393	1.443603e-01
1211	1428	5.308780e-01
1181	1277	4.526939e-01
1201	1428	7.442241e-02
1318	1323	2.028973e-02
1173	1328	1.517259e-01
1172	1266	1.386631e-01
1139	1323	2.701214e-01
1195	1196	1.450282e-02
1107	1328	6.269854e-02
1163	1235	5.858603e-01
1217	1307	9.121864e-02
1246	1323	6.779189e-01
1320	1401	2.451628e-01
1138	1274	4.505012e-02
1246	1257	1.234649e-01
1276	1393	5.378936e-0
1233	1235	2.010115e-01
1196	1276	8.097776e-01
1163	1277	6.046633e-01
1242	1276	5.579365e-01
1217	1314	1.419408e-01
1308	1361	1.196856e-01
1112	1361	6.620287e-01
1217	1277	3.960374e-01
1143	1393	3.455604e-01
1257	1276	6.839127e-01
1196	1417	7.589796e-01
1332	1387	4.216728e-01
1231	1393	7.253557e-01
1308	1387	2.443873e-01
1285	1393	7.451223e-02
1295	1361	3.993100e-01
1114	1235	4.367620e-01
1409	1417	1.996611e-01
1231	1246	6.439179e-01
1186	1207	2.554150e-01
1116	1314	5.535047e-01
1266	1393	5.467613e-01
1372	1417	5.076781e-01
1174	1277	1.851321e-01
1257	1387	8.350617e-01
1181	1428	6.819841e-01
1218	1268	1.292253e-01
1114	1345	2.907136e-01
1103	1433	1.439308e-01
1207	1428	3.031373e-01
1301	1387	3.573617e-01
1372	1428	3.406836e-01
1151	1231	1.614958e-01
1279	1458	4.064006e-01
1372	1433	3.035425e-01
1268	1452	3.072800e-01
1143	1424	5.742580e-01
1242	1268	5.506527e-01
1246	1452	8.371529e-01
1276	1433	3.895316e-01
1295	1328	5.192540e-01
1138	1452	3.787908e-01
1257	1320	6.392260e-01
1173	1326	2.318522e-01
1320	1400	4.443578e-01
1211	1371	6.701111e-01
1279	1462	5.178339e-01
1314	1323	7.887680e-01
1173	1344	3.027571e-01
1209	1462	3.180458e-01
1137	1139	5.453032e-01
1235	1314	5.335909e-01
1276	1323	4.220051e-01
1278	1417	5.513139e-01
1112	1462	8.545387e-01
1211	1234	6.241497e-01
1338	1458	4.895594e-01
1153	1166	3.690425e-01
1195	1361	7.667601e-02
1113	1400	2.486159e-01
1140	1332	3.344365e-01
1323	1458	5.164602e-01
1276	1400	7.180010e-01
1328	1361	4.268172e-01
1181	1277	8.481112e-01
1231	1455	3.243040e-01
1235	1323	5.818279e-01
1326	1433	6.318307e-01
1323	1455	5.793249e-01
1242	1455	4.657292e-01
1208	1277	3.276705e-01
1172	1234	4.151692e-01
1160	1228	2.497732e-01
1257	1277	7.341081e-01
1139	1403	6.091834e-01
1332	1386	8.387110e-01
1257	1301	6.965240e-01
1242	1314	7.486288e-01
1181	1361	7.887653e-01
1203	1272	3.923409e-01
1261	1301	2.719506e-01
1243	1246	2.771496e-01
1153	1386	4.864511e-01
1329	1332	5.108226e-01
1163	1246	6.548749e-01
1344	1425	3.576729e-01
1112	1211	7.994951e-01
1231	1396	9.168733e-01
1153	1345	6.426302e-01
1211	1329	5.928655e-01
1338	1455	7.048625e-01
1161	1281	5.754919e-01
1196	1338	8.221059e-01
1160	1163	3.542055e-01
1211	1455	8.989001e-01
1314	1437	8.066905e-01
1361	1385	6.835232e-01
1247	1455	4.055287e-01
1257	1455	8.813101e-01
1301	1396	5.557948e-01
1326	1455	8.384337e-01
1160	1338	4.547675e-01
1314	1344	9.392762e-01
1307	1390	5.103769e-01
1333	1433	3.059285e-01
1173	1390	4.469708e-01
1173	1393	4.522551e-01
1242	1390	7.334186e-01
1292	1393	9.556229e-02
1314	1393	8.794337e-01
1211	1393	7.039554e-01
1163	1386	7.760074e-01
1234	1396	7.890835e-01
1112	1326	7.016220e-01
1328	1433	6.122697e-01
1272	1388	3.854578e-01
1273	1397	7.437801e-02
1374	1417	6.214292e-01
1314	1344	6.837783e-01
1196	1278	9.066098e-01
1211	1417	7.957135e-01
1124	1304	4.997835e-01
1274	1455	6.327788e-01
1234	1397	4.518525e-01
1269	1397	3.167457e-01
1435	1455	3.242628e-01
1112	1455	4.344789e-01
1129	1173	5.788202e-01
1276	1397	6.324535e-01
1139	1400	5.753205e-01
1412	1417	1.157739e-01
1112	1125	5.010376e-01
1292	1388	4.882741e-01
1140	1279	4.655723e-01
1276	1409	6.547294e-01
1153	1217	7.676551e-01
1268	1355	6.979044e-01
1301	1462	6.193408e-01
1320	1461	7.756595e-01
1329	1332	7.142276e-01
1257	1332	8.773704e-01
1181	1463	6.245024e-01
1247	1279	3.044499e-01
1124	1463	5.303468e-01
1116	1459	7.283141e-01
1129	1247	4.060221e-01
1243	1247	7.210109e-01
1246	1392	8.543317e-01
1181	1423	8.892943e-01
1257	1414	9.242545e-01
1143	1218	7.460905e-01
1257	1264	9.828789e-01
1268	1434	7.185373e-01
1276	1355	9.037084e-01
1166	1418	9.684893e-01
1124	1209	8.761519e-01
1235	1412	9.628908e-01
1277	1434	9.341399e-01
1181	1273	9.686085e-01
1323	1372	7.446001e-01
1196	1322	9.760135e-01
1401	1453	9.492018e-01
1393	1444	9.671310e-01
1112	1217	9.216185e-01
1235	1300	7.707736e-01
1276	1459	9.609614e-01
1437	1454	9.607885e-01
1274	1334	9.793245e-01
1242	1308	9.119078e-01
1277	1292	9.729523e-01
1211	1295	9.646168e-01
1112	1411	9.845795e-01
1242	1443	9.998913e-01
1211	1380	9.999034e-01
1316	1352	6.211835e-01
1142	1411	6.470647e-01
1112	1451	9.999450e-01
1192	1195	4.768619e-01
1107	1291	6.414800e-01
1214	1264	5.452941e-01
1251	1299	3.677402e-01
1231	1241	9.999459e-01
1257	1299	9.999441e-01
1181	1352	9.999324e-01
1241	1254	5.508402e-01
1221	1380	2.765441e-01

Part 2 Results


Region Y
1 Kansas
1 Kansas
16 UC Davis
16 NC Central
16 NC Central
8 Miami
8 Miami
9 Michigan St.
1 Kansas
5 Iowa St.
5 Iowa St.
12 Nevada
4 Purdue
4 Purdue
13 Vermont
3 Oregon
6 Creighton
11 Rhode Island
11 Rhode Island
3 Oregon
3 Oregon
3 Oregon
14 Iona
7 Michigan
7 Michigan
10 Oklahoma St.
2 Louisville
2 Louisville
2 Louisville
15 Jacksonville St.
1 North Carolina
1 North Carolina
16 Texas Southern
8 Arkansas
8 Arkansas
9 Seton Hall
1 North Carolina
5 Minnesota
5 Minnesota
12 Middle Tenn.
4 Butler
4 Butler
13 Winthrop
1 North Carolina
6 Cincinnatti
6 Cincinnatti
11 Wake Forest
11 Wake Forest
11 Kansas St.
3 UCLA
3 UCLA
3 UCLA
14 Kent. St.
7 Dayton
10 Wichita St.
10 Wichita St.
2 Kentucky
2 Kentucky
2 Kentucky
15 Northern Ky.
Region Z

Final Four


1 Villanova
1 Gonzaga
1 Villanova
National Champion
1 Villanova
3 Oregon
3 Oregon
1 North Carolina

Sources

Bloomenthal, Andrew "Warren Buffet's March Madness Bracket Challenge: What Are the Odds?" *Investopedia.com. March 12, 2020.* *https://www.investopedia.com/ask/answers/082714/what-are-odds-getting-perfect-bracket-warren-buffetts-1-billion-march-madness-bracket-challenge.asp*

McGee, Noah "How much money does the NCAA make from March Madness". *MSN.com.* *https://www.msn.com/en-us/money/markets/how-much-money-does-the-ncaa-make-from-march-madness/ar-BBZIYDV#:~:text=The%20NCAA%20made%20%24867.5%20million,the%202018%2D19%20fiscal%20year. 2020*

Ota, Kevin "Not So Perfect: Only 0.25% of 17.2 Million Brackets Remain Perfect in ESPN's Tournament Challenge Game". *ESPN.com. March 22, 2019.* *https://espnpressroom.com/us/press-releases/2019/03/not-so-perfect-only-0-25-of-17-2-million-brackets-remain-perfect-in-espns-tournament-challenge-game/*

Kaggle Competition Link

https://www.kaggle.com/c/march-machine-learning-mania-2017