Winning It All: A Model to Predict the Winner of NCAA® March Madness

By Marigrace Seaton, Chase Wang, Luke Wheeler, and Matthew Wheeler

The Problem

It is hard to not be a basketball fan at the University of North Carolina. Basketball culture is all around us. Thus, it was an easy decision for us to create a model to predict, given regular-season game statistics from NCAA® men's basketball, who the overall winner of the postseason tournament would be.

Related Works

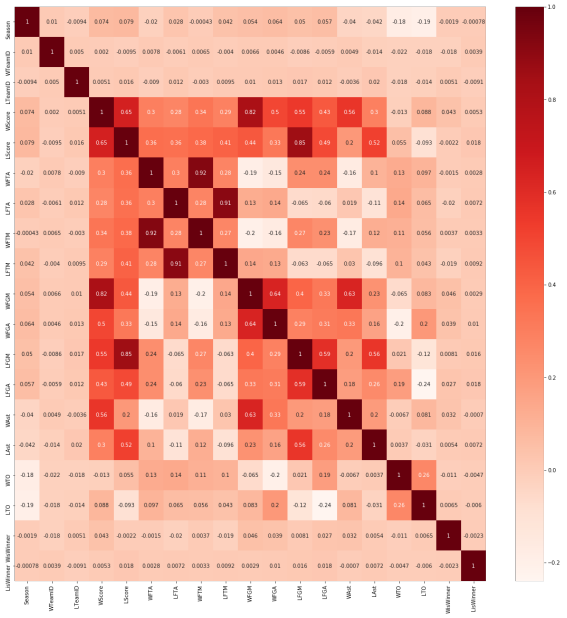
Works such as "Logistic Regression on Tournament seeds" [1] and "The Tale of Kaglerella" [2] take on a similar problem. However, we found that "Logistic Regression on Tournament seeds" left too much room for error, because higher-seeded teams often lose out in the early stages of the tournament, and upsets are a common occurrence. Similarly, "The Tale of Kaglerella" failed to take several ratios into account which we found to be important in determining team strength.

Feature Selection

We derived our training datasets from the data provided for the 2019 Kaggle Google Cloud & NCAA® ML Competition 2019 for men's basketball[3]. These datasets include information on regular-season statistics, tournament statistics, home or away statistics, seed information, and more. Using pandas, we merged seed information and tournament outcome into the regular season statics to create one single data frame from which we would select our features. The challenge with preparing the data to become features was picking what we thought to be the most important statistics from those provided and manipulating them so that they might be more useful and simultaneously more compact.

We saw through a correlation heatmap (following page, left) that there was a strong correlation between a given team's number of assists, turnovers, field goals scored, free throws scored, and whether either of the teams was the winner of that season's tournament (indicated by a "1" under either "WisWinner" or "LisWinner" for each season).

[1] Logistic Regression on Tournament seeds
[2] The Tale of Kagglerella
[3] Google Cloud & NCAA® ML Competition 2019-Men's



	0	1	2	3	4
Season	2003	2003	2003	2003	2003
WTeamID	1104	1272	1266	1296	1400
LTeamID	1328	1393	1437	1457	1208
WScore	68	70	73	56	77
LScore	62	63	61	50	71
WLoc	N	N	N	N	N
WisWinner	0	0	0	0	0
LisWinner	0	1	0	0	0
WSeeds	Y10	Z07	Y03	0	X01
LSeeds	W01	W03	0	0	0
WAstTORatios	0.562771	1.22137	1.48515	0.909091	0.851064
LAstTORatios	0.441989	0.578512	0.743802	0.471204	1.18812
WFTRatios	0.611111	0.526316	0.586207	0.548387	0.846154
LFTRatios	0.727273	0.45	0.608696	0.533333	0.62963
WFGRatios	0.465517	0.419355	0.413793	0.473684	0.491803
LFGRatios	0.415094	0.358209	0.30137	0.367347	0.387097

According to Breakthrough Basketball[4], the most important statistics that affect the strength of a college basketball team and the outcome of their future games are the assist-to-turnover ratio, free-throw accuracy percentage, and field goal accuracy percentage. We chose to use all three of these as features in our model to predict the strongest teams going into the postseason tournament.

Because these statistics were not provided to us, we manipulated our dataset using pandas dataframe to include the following for each regular season game: the winner's and loser's assist-to-turnover ratios, the winner's and loser's free throw accuracy percentage, and the winner's and loser's field goal accuracy percentage. We also reduced the dataset (above, right) to include only the aforementioned statistics, along with the following: season year (indicated by "Season"), the ID number for the winning and losing teams (indicated by "WTeamID" and "LTeamID," respectively), the winner's and loser's score (indicated by "WScore" and "LScore," respectively), the location of the game (indicated by "WLoc," defined as "N" if the game was played on a neutral court, "A" if the game was played on the losing team's court, and "H" if the game was played on the winning team's court), whether the winner or loser went on to win that season's tournament (0 if the team in question did not win the tournament and 1 if it did), and the designated seeds of the winning and losing teams ("W," "X," "Y," and "Z" before the seed number referring to the team's position in the tournament bracket: East, Midwest, South, and West, respectively). If seeds value is 0, that indicates the team did not make it to the March Madness playoff.

Furthermore, in order for our model to more easily predict which team would emerge the winner, we transformed the data such that each column of our dataset would correspond with only one team's performance in a single game, and changed the meaning of the "isWinner" label to correspond to a "1" or a "0," depending on if the team in question had won the

[4] The Most Important Stats To Track For Your Basketball Team - Marcus Hagness

game for which their stats were being shown. In other words, we essentially separated the losing team's game data and the winning team's game data (below).

	0	1	2	3
Season	2003	2003	2003	2003
TeamID	1104	1328	1272	1393
Score	68	62	70	63
Loc	N	N	N	N
IsWinner	1	0	1	0
Seeds	Y10	W01	Z07	W03
AstTORatios	0.562771	0.441989	1.22137	0.578512
FTRatios	0.611111	0.727273	0.526316	0.45
FGRatios	0.465517	0.415094	0.419355	0.358209

Method

We used a logistic regression model to classify teams as either a winner or a loser based on their stats for a game. After that, we trained the model on one season's games, and then validated using the rest of the season data (excluding the 2019-2020 season). We repeated this for every season, and checked which training model had the least amount of misclassified games. In addition, we merged all the data for each team for the 2019-2020 season[5] and calculated the average scores, assist-to-turnover ratios, free throw ratios, and field goal ratios for each team in teh regular season. We used our training model on the averaged data to predict which team had the highest chance of winning the tournament. The model that ended up having the lowest error was the one based on the 2007-2008 season, which was around 26.07% error. The other models had higher error but no more than 2% higher. While high, due to the somewhat random and spontaneous nature of basketball, a perfect system is not necessarily feasible.

Results

From the 2019-2020 season data, the model gave the teams listed in the chart on the following page the highest probabilities of being classified as a "winner" of the final game of the tournament based on their average stats for the season.

[5] Google Cloud & NCAA® ML Competition 2020-NCAAM

Team	Team ID	Probability
Gonzaga	1211	0.936
Belmont	1125	0.892
Hofstra	1220	0.885
Tennessee	1397	0.854
South Dakota State	1355	0.837
Murray State	1293	0.817
Yale	1463	0.807
Michigan State	1277	0.804
North Carolina	1314	0.795
Buffalo	1138	0.794

The team that it predicted to win was Gonzaga, which makes sense as it was one of the best teams in the regular season. Many of the teams listed (such as Hofstra and Yale) were among the best in their conferences, and almost all of them had winning records. North Carolina, on the other hand, had a losing record, which shows some of the flaws in the model - despite doing very well in many statistics that appear to be associated with winning games, North Carolina did not win many games.

GitHub Repository

Our work and data can be found here.

Contributors

Marigrace Seaton, Chase Wang - Feature Selection and Data transformation
Luke Wheeler, Matthew Wheeler - Model Construction and Results Analysis

References

[1] Kplauritzen. *Logistic Regression on Tournament seeds.* Feb. 2017. https://www.kaggle.com/kplauritzen/notebookde27b18258

[2] Iamleonie. *The Tale of Kaglerella.* April 2020. https://www.kaggle.com/iamleonie/the-tale-of-kagglerella

[3] Kaggle. *Google Cloud NCAA® ML Competition 2019-Men's.* https://www.kaggle.com/c/mens-machine-learning-competition-2019/data

[4] Hagness, Marcus. *The Most Important Stats To Track For Your Basketball Team.* Breakthrough Basketball. https://www.breakthroughbasketball.com/stats/how-we-use-stats-Hagness.html

[5] Kaggle. *Google Cloud NCAA® ML Competition 2020-NCAAM.* https://www.kaggle.com/c/google-cloud-ncaa-march-madness-2020-division-1-mens-tournament