Predicting Games in "March Madness" - NCAA Men's Basketball Tournament

Karlo Borovčak[1]

[1] Faculty of electrical engineering and computing, University of Zagreb

January 12, 2023

Mentor: doc. dr. sc. Marina Bagić Babac

Abstract

March Madness is the biggest annual tournament in the US. It is also known for bracket prediction and millions of people every year try to predict the impossible perfect bracket. That's why it is of interest of teams, players, coaches and fans to know which statistics are important in deciding a winner of a March Madness tournament game. After studying multiple similar papers we find which statistics are the most significant for making a model and what kind of results we can expect from a model trying to predict March Madness games. Model that is most commonly used to predict March Madness games is Logistic regression so it was the obvious choice. The model was developed based on SRS (simple rating system) and other common statistics like: assists, three pointers, offensive rebounds, steals, blocks... It was trained and tested on March Madness tournament data from 2014. to 2021. (excluding 2020.). Fitting the model was done on tournament games only which resulted in good prediction scores compared to similar work which mostly fitted their models on both the tournament and regular season games.

Keywords: *NCAA Men's Basketball; March Madness; Logistic regression; Prediction*

1. Introduction

With 358 teams playing across 49 states (all but Alaska), NCAA men's college basketball is one of the most popular and widespread sports in the United States. During the 2021-2022 basketball season, a total of 23,789,492 people attended 5,516 total Division I men's basketball games.

The most important part of the NCAA men's basketball season is the annual tournament, also known as March Madness. This basketball tournament has become one of the most popular and famous sporting tournaments in the United States. Millions of people around the country participate in contests in which the participants make their best guesses on who will win each game throughout the tournament. These types of contests have become so popular that more than 17.3 million brackets were filled out on ESPN.com in 2022. One major hindrance to a human's ability to make accurate predictions is the presence of bias. Everyone is biased, given the reality is that there is no clear-cut answer to the question of what factors, or features, contribute to the result of a game. With the use of data, machine learning algorithms can mathematically and algorithmically attempt to learn which statistics correlate the most with the result of a game. The emergence of more accurate machine learning techniques has led to increased prediction accuracy using algorithms powered by historical data.

To make these algorithms efficient one of the most important things to do is feature selection, it's also one of the hardest things to do because some of the features might be important, and vice versa, even if we intuitively thought different. That is why a good exploratory analysis is an important step in making a good model. This paper discovers some of the previous work that has been done on the topic of predicting college basketball games and explores which methods worked best and which features are important in making a solid prediction in college basketball.

2. Related work

Since March Madness is one of the most popular tournaments in the US and around millions of people try to predict the outcome of the tournament each year, there is also a lot of scientific research on it. In the Table 1 is a summary of papers with their models, features they used in their models and results of trying to predict the outcome of games and the whole March Madness tournament.

Authors	**Model type**	**Features Selected**	**Results**
(Lopez, Matthews, 2014)[1]	Logistic regression	Las Vegas point spread, team efficiency ratings	Won 2014 Kaggle March Madness contest
(Stekler, Klein, 2012)[2]	Probit model	Consensus forecasts, team seeds	73.6% (game outcomes)
(Forsyth, Wilde, 2014)[3]	kNN	94 features from ESPN team stats that were reduced to 5 based on SVM weights	73.62% (game outcomes)
(Magel, Unruh, 2013)[4]	Logistic regression	Assists, free throw attempts, defensive rebounds, turnovers	66.67% (game outcomes)
(Shen, Hua, et al, 2015)[5]	Binomial generalized linear regression model with Cauchy link	Field goals made, seed, defensive rebounds, average scoring margin, strength of schedule	71.43% (game outcomes)
(Brown, 2019)[6]	Logistic regression	Square root of the cumulative average of field goals made, square root of the moving average of steals, cumulative average of score, moving average of personal fouls, and square root of the moving average of turnovers	70.2% (game outcomes)
(Kocher, Hoblin, 2017)[7]	Logistic regression and Decision trees	Multiple models including defensive stats model, offensive stats model etc.	50th percentile of all brackets on ESPN in 2018
(Lobo, Levandoski, 2017)[8]	kNN, regression, SVM, neural network	Basic offensive and defensive team stats per game and Home or Away binary flag	Bracket score 900 (over 50th percentile on ESPN in 2017)
(Unrruh, 2013)[9]	Linear and logistic regression	Assists, free throw attempts, defensive rebounds, and turnovers	64% and 68% (game outcomes)
(Kvam, Sokol, 2006)[10]	Logistic regression/Markov chain (LRMC)	Home/Away/Neutral court, point differential	73.28% (game outcomes)

Table 1. Summary of March Madness prediction papers

As you can see from Table 1 the most popular model for predicting outcomes is Logistic regression which makes sense given the binary nature that gives us 1 if a team wins or 0 if a team loses. Most used features in seem to be defensive rebounds, Home/Away/Neutral court, assists and other basic team stats.

Most of the models predict around 70% of the games outcomes but when it comes to predicting a good bracket it's a lot harder to make a better prediction than just going off of team seeds or other similar strength rankings. That's why when we are trying to make a bracket prediction we must consider a different model for each round of the tournament or a similar approach.

3. Methodology

After researching through some similar work on predicting the outcome of a March Madness tournament we can try to make our own Logistic regression model. The data was scraped from ***Sports Reference*** using Selenium and Beautiful Soup Python libraries. Scraper, preprocessing code and the dataset can be found on ***GitHub***.

3.1. Data set Preprocessing

To make our scraped and raw data set ready for making predictions, we have to do some data set preprocessing. I used Pythons Pandas library to clean the data. Most of the cleaning was removing useless columns, removing rows with missing values and renaming and rearranging columns. After cleaning the data set I did some feature engineering so I created some new columns from existing ones like TRB (Total rebounds) - ORB (Offensive rebounds) = DRB (defensive rebounds) and divided all of the stats by the number of games each team has played since they played different amount of games. It was also important to map some values to numeric values like POSTSEASON which contains the round of the tournament reached by a team to be mapped like: 'R68' -> 1, 'R64' -> 2, 'R32' -> 3,..., so it can be used in feature selection which I will discuss in a later subsection.

3.2. Dataset description

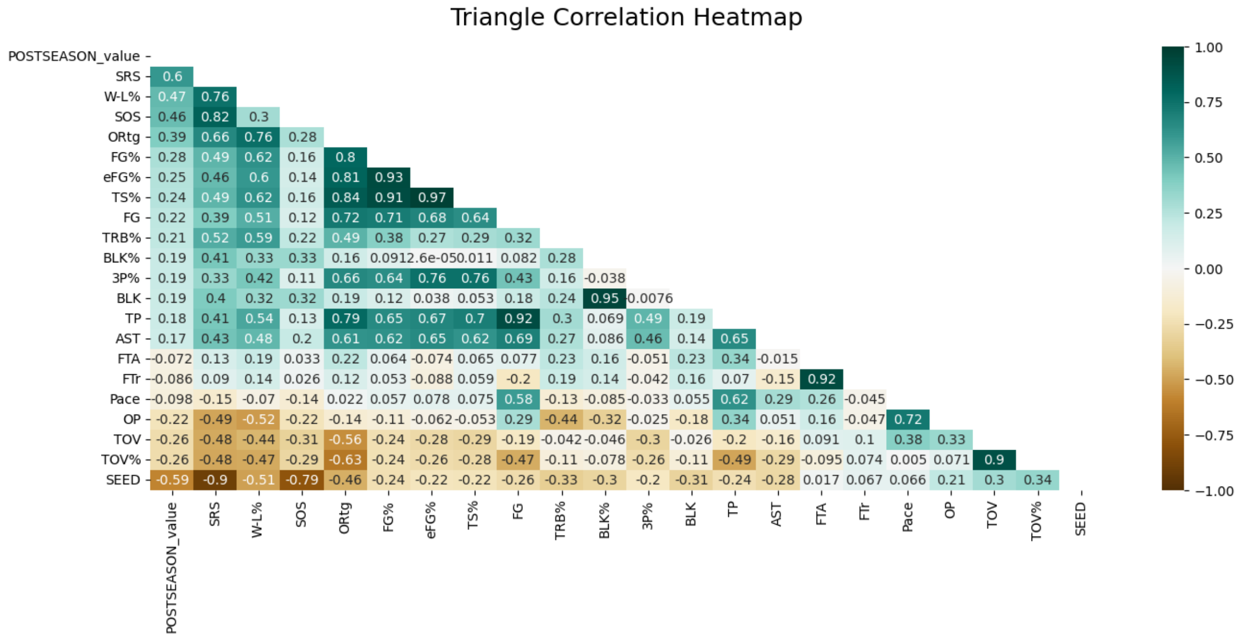
The dataset consist of team stats for 2014 to 2021 NCAA Division I college basketball seasons for 355 different teams. Each team has the following features in the Table 2.

Feature label	Feature description
G	Number of games
Overall.W	Total wins
Overall.L	Total loses
W-L%	Total win-loss percentage
SRS	(Simple Rating System) A rating that takes into account average point differential and strength of schedule. The rating is denominated in points above/below average, where zero is average. Non-Division I games are excluded from the ratings.
SOS	(Strength of Schedule) A rating of strength of schedule. The rating is denominated in points above/below average, where zero is average. Non-Division I games are excluded from the ratings.
Conf.W	Number of conference wins
Conf.L	Number of conference loses
Home.W	Number of home wins
Home.L	Number of home loses
Away.W	Number of away wins
Away.L	Number of away loses
TP	Team points scored per game
OP	Opponnent points scored per game
MP	Total minutes played
FG	Field goals made per game
FGA	Field goal attempts per game
FG%	Field goal percentage
3P	Three pointers made per game
3PA	Three point attempts per game
3P%	Three point percentage
FT	Free throws made per game
FTA	Free throw attempts per game
FT%	Free throw percentage
TRB	Total rrebounds per game
ORB	Offensive rebounds per game
DRB	Defensive rebounds per game
AST	Assists per game
STL	Steals per game
BLK	Blocks per game
TOV	Turnovers per game
PF	Personal fouls per game
Pace	(Pace Factor) An estimate of school possessions per 40 minutes
ORtg	(Offensive Rating) An estimate of points scored per 100 possessions
FTr	(Free Throw Attempt Rate) Number of FT Attempts Per FG Attempt
3PAr	(3-Point Attempt Rate) Percentage of FG Attempts from 3-Point Range
FT/FGA	Free Throws Per Field Goal Attempt
TS%	(True Shooting Percentage) A measure of shooting efficiency that takes into account 2-point field goals, 3-point field goals, and free throws
TRB%	(Total Rebound Percentage) An estimate of the percentage of available rebounds a player grabbed while he was on the floor
AST%	(Assist Percentage) An estimate of the percentage of teammate field goals a player assisted while he was on the floor
STL%	(Steal Percentage) An estimate of the percentage of opponent possessions that end with a steal by the player while he was on the floor
BLK%	(Block Percentage) An estimate of the percentage of opponent two-point field goal attempts blocked by the player while he was on the floor
eFG%	(Effective Field Goal Percentage) this statistic adjusts for the fact that a 3-point field goal is worth one more point than a 2-point field goal
TOV%	(Turnover Percentage) an estimate of turnovers per 100 plays
ORB%	(Offensive Rebound Percentage) an estimate of the percentage of available offensive rebounds a player grabbed while he was on the floor
SEED	Team seed
POSTSEASON	Round of tournament reached
REGION	Region of the tournament

Table 2. Dataset features

3.3. Feature selection

Most of the features in our data set won't be useful for making a prediction. To choose features I did some exploratory analysis with Pythons Seaborn and Matplotlib libraries. After exploring the data here is a correlation heatmap of features that are correlated to the feature POSTSEASON_value (mapped value of round of tournament reached by a team) the most.



As you can see from the Figure 3.3 above, features: SRS, SOS, ORtg, FG%, SEED, TOV% are correlated to POSTSEASON_value the most. However to avoid overfitting we must be careful not to choose too many features depending on the number of observations [11]. Having that in mind and combining it with features used in Related Work 2, I chose the following features for our Logistic regression model: SRS, SEED, AST, 3P, ORB, STL, BLK, PF, FT% and TS%. (Look at Table 2 for detailed description of each feature)

3.4. Logistic regression model

From the data set I was able to recreate a tournament bracket for each March Madness tournament from 2014. to 2021. (excluding 2020. because of COVID) which gave me a 469 games sample. Each tournament has 67 games so I split the training data and test data to 402 games of 6 tournaments and 67 games of a single tournament. To fit a Logistic regression model I used Pythons Scikit-learn library. For each of the 469 games in the sample, there was a team that was randomly selected to be the “team of interest”, and the value of all regressors were equal to the feature value of the “team of interest” minus the value for the “opposing team” (Look at Table 3 for an example of training data). The model was fitted using Pythons Scikit-learn library and it outputs the value 1 if the team is predicted to win or 0 if it predicts a loss for the "team of interest".

TeamOfInterest	Opponnent	Won	SRS	ORB	PF	3P	STL	AST	BLK	FT%	TS%	SEED
Virginia	Memphis	1	5.35	-2.31	-1.94	0.21	-3.4	-4.3	-0.62	0.029	-0.005	-7.0
Harvard	Michigan State	0	-8.56	0.074	-1.28	-2.32	0.77	-2.73	0.26	0.017	-0.009	8.0
Iowa State	North Carolina	1	3.27	-3.78	-2.98	4.06	-1.37	3.03	-1.79	0.06	0.04	-3.0
Villanova	Connecticut	0	1.73	1.47	1.78	1.76	-0.08	3.18	-1.60	-0.061	0.01	-5.0
Saint Louis	Louisville	0	-13.06	-3.44	-1.98	-2.19	-2.55	-1.17	-0.61	0.035	-0.02	1.0

Table 3. Example of training data

4. Results

After studying the related work, scraping the data, cleaning and preprocessing the data and of course fitting our Logistic regression model we finally got some results. To test the model I split the data by year and tried predicting games for each tournament by year while training the model on the games from all the other tournaments in the data set. I mesaured the results by a percantage of correct game outcome predictions by year.

Year of the tournament	Percentage of correct game outcome predictions
2014	71.64%
2015	76.12%
2016	70.15%
2017	74.63%
2018	76.12%
2019	82.09%
2021	77.61%
Average model accuracy	75.48%

Table 4. Logistic model prediction results

As you can see in the Table 4 above the results are pretty decent and I think the most significant feature in the model is SRS which gives us a good idea of how strong a team is but the other features give us an edge in picking some weaker teams to win. It seems that the model performs better for later years because basketball has been changing pretty rapidly for the past 20 years and even a 5 year interval brings a lot of novelty into the game and its statistics.

5. Discussion

If we compare our model to others in Table 2 we can see that the model did pretty well. Most of the other models have a score of around 70% correct game outcomes and ours on average had 75.48%. This model worked well for tournament games and it wasn't tested or fitted on regular season games which a lot of the models from Table 2 did. So fitting the model based only on tournament games seemed to work well in predicting other tournament game outcomes. In future work, other models and maybe even some machine learning models could be explored with other features and newer data. Also, models based on trying to predict the tournament outcome (getting a good bracket score) could be explored which would have a whole different approach in trying to identify "Cinderellas" (team that is no. 8 seed or below that makes a deep run into the tournament) and not always picking the team with a better chance to win.

References

[1] M. Lopez and G. Matthews, "Building an ncaa mens basketball predictive model and quantifying its success," *Journal of Quantitative Analysis in Sports*, vol. 11, 11 2014.

[2] H. Stekler and A. Klein, "Predicting the outcomes of ncaa basketball championship games," *Journal of Quantitative Analysis in Sports*, vol. 8, pp. 3–3, 01 2012.

[3] J. Forsyth and A. Wilde, "A machine learning approach to march madness," *Department of Computer Science, Brigham Young University*, Winter 2014.

[4] R. Magel and S. Unruh, "Determining factors influencing the outcome of college basketball games," *Open Journal of Statistics*, vol. 03, pp. 225–230, 01 2013.

[5] G. Shen, S. Hua, X. Zhang, Y. Mu, and R. Magel, "Predicting results of march madness using the probability self-consistent method," *International Journal of Sports Science*, vol. 05, pp. 139–144, 2015.

[6] B. Brown, "Predictive analytics for college basketball: Using logistic regression for determining the outcome of a game," Master's thesis, University of New Hampshire, Spring 2019.

[7] C. Kocher and T. Hoblin, "Predictive model for the ncaa men's basketball tournament," Master's thesis, Ball State University, 04 2017.

[8] J. Lobo and A. Levandoski, "Predicting the ncaa men's basketball tournament with machine learning," *Department of Computer Science, University of Pittsburgh*, 04 2017.

[9] S. P. Unruh, "Analysis of significant factors in divison i men's college basketball and development of a predictive model," Master's thesis, North Dakota State University, 04 2013.

[10] P. Kvam and J. S. Sokol, "A logistic regression/markov chain model for ncaa basketball," *Naval research Logistics (NrL)*, pp. 788–803, 2006.

[11] J. Frost, "Overfitting regression models: Problems, detection, and avoidance," *Statistics By Jim*. [Online]. Available: https://statisticsbyjim.com/regression/overfitting-regression-models/