Using a Multi Layer Perceptron to Predict the Final Placement of Teams in March Madness

Jorden Anfinson

April 2023

Abstract

Due to its 6-round tournament format with 64 teams, March Madness is notoriously difficult to predict accurately. Typically, forecasting March Madness entails using a model to predict the outcome of each game, in the hopes that the model will predict each one correctly. Building a model capable of doing so is considered next to impossible, and no model to this point has been able to do so. However, a more manageable task is to forecast the round in which each team will be eliminated. In this paper, we explore the effectiveness of using a multi-layer perceptron to classify college basketball teams based on the round of March Madness in which they are eliminated. In doing so, we create a network that can accurately forecast the tournament outcome of March Madness teams nearly 50 percent of the time, and find that Kansas, UCLA, Arizona, Alabama and USC are the most likely teams to become the 2023 March Madness Tournament Champion.

1 March Madness Background

March Madness is the annual end-of-year college basketball tournament consisting of 68 qualifying teams. 32 of those teams automatically qualify by winning their respective conference tournament, while the remaining 36 are chosen by a selection committee made up of senior NCAA officials and athletic directors. Each team is given a seed (a numeric value) ranging from 1 to 16 by the committee based on their performance throughout the season. The teams that are considered to be better are assigned seeds closer to 1, whereas those that barely qualify are assigned seeds closer to 16. Each seed is assigned to 4 teams, for example, the 4 teams considered to be the best in the country are assigned 1 seeds. This year, those teams were Kansas, Alabama, Purdue, and Houston. Each round of the tournament is also given a name, which corresponds to the number of teams remaining in the tournament. For example the semi-finals are called the "Final 4", while the quarter finals are called the "Elite 8".

2 Introduction

Supervised learning algorithms attempt to use labeled training data, data for which a desired value is already known, to predict or classify the values of new data, called testing data. Neural networks belong to the category of supervised learning and these networks are composed of neurons, weights, and activation functions, which all work together to solve complex problems. They are often leveraged to solve problems related to sports, and they are used by sports organizations and betting companies to predict outcomes.

A multi-layer perceptron is a type of neural network that consists of multiple layers of interconnected neurons, each performing a transformation on the data passed to the neuron according to its activation function.

In this paper we utilize a 4-layer perceptron neural network to classify teams based on the round of March Madness in which they are expected to be eliminated, or each team's "tournament outcome". We evaluate the accuracy of the network, and use it to identify the 5 teams that are most likely to emerge as winners in the upcoming 2023 March Madness tournament.

3 Data

The data used to develop this paper was originally scraped from the advanced college basketball statistics website "Barttorvik".

Our main dataset consists of a selection of stats from every college basketball team from 2013 to 2021, including both teams that ended up making the tournament and teams that did not. It consists of 24 features in total, the most important of them being the team's tournament outcome: a categorical value representing whether they finished as the champion, runner up, in the final four, etc. Some additional features in the data include the team's name, home conference, resulting tournament seed and the year the team played. It additionally consists of more advanced statistics such as BARTHAG, which is a numeric representation of the percent chance that a given team would beat the average D1 basketball team. We use this data to construct the neural network we use to predict the tournament outcome of 2023 College Basketball teams in March Madness.

TEAM	CONF	G	W	ADJOE	ADJDE	BARTHAG	EFG O	SEED	...	SEASON
Duke	ACC	39	35	125.2	90.6	0.9764	56.6	1	...	2015
North Carolina	ACC	40	33	123.3	94.9	0.9531	1	48.1	...	2016
...	...	...	...	...	...	...	...	...	...	...

Table 1: Preview of Dataset

While we later split the main dataset mentioned above into training, testing and validation sets to evaluate the network created by this data. After evaluating the network, we pass our final testing dataset through it to predict the

tournament outcomes of team's competing in the 2023 March Madness Tournament. This dataset consists of the stats of 2023 Tournament Teams. These stats were scraped from Barttorvik right after teams had been assigned seeds for the upcoming tournament, but right before the tournament started. From this point on we refer to this set of teams as "2023 Tournament Teams."

TEAM	CONF	G	W	ADJOE	ADJDE	BARTHAG	EFG O	SEED	...	SEASON
Houston	Amer	34	31	117.1	88.0	0.9638	52.7	1	...	2023
Alabama	SEC	34	29	115.4	88.3	0.9557	52.7	1	...	2023
...	...	...	...	...	...	...	...	...	...	...

Table 2: Preview of Data for 2023 Tournament Teams

4 Methodology

In order to make the data cleaning process more efficient, we first combine our main training data (team data from 2013-2021) with our final testing data (2023 Tournament Teams).

We then clean the data by first filtering out teams that did not make the Tournament in their respective season. We do so because our final testing data only consists of teams we know to be in the 2023 March Madness tournament, and because we are not interested in predicting whether or not a team will make the postseason or not.

We additionally represent the round in which a team was eliminated (their final placement) as a numeric value. The table below shows which numbers correspond with every possible tournament result for a tournament team.

After cleaning the data as described above, we re-split the data into two dataframes: "currentTournamentTeams" containing the 2023 Tournament Teams, and "data" which contains Tournament Teams from 2013-2021. Re-splitting the data in this manner is necessary because we must apply the same transformations used to normalize the training data to the testing data as well before

Value	Tournament Outcome
0	Champion
1	Runner Up
2	Final 4
3	Elite 8
4	Sweet 16
5	Round of 32
6	Round of 64
7	Round of 68

Table 3: Possible Tournament Outcomes

feeding the training data into the network. Splitting the data in this fashion makes that process easier.

As stated above, we then normalize our training data, and apply the same transformation to our final testing data. A preview of the result of this transformation can be seen below.

We then allocate a portion of our main training data to be used for a validation set and a testing set. (Note : this testing set is different than our final testing set which is made up of 2023 Tournament Teams, this testing set will be used to evaluate the effectiveness of the network). To do so we take 20 percent of our original training data to create both our new testing and validation sets. This means that 60 percent of the original training data is left over to be fed to the network.

5 Structure of Network

We create a 4-layer perceptron to classify teams based on their tournament outcome. The input layer of the perceptron contains 18 neurons, one for each remaining feature in the normalized training data. The first hidden layer of the perceptron contains 36 neurons, while the second contains 18. Both utilize a "relu" activation function. Our output layer contains 8 neurons, one for each possible March Madness outcome (these outcomes are visible in Figure 1). The output layer's neurons use a "soft-max" activation function to compute the chance that an observation belongs to any of the possible outcomes, and the outcome with the highest chance will be the predicted value for this observation.

As visualized in figure 1, through overfeeding the network we find the optimal number of training steps to be around 120. At this point the validation loss begins to increase, and training the network any further would risk overfitting it to the training data.

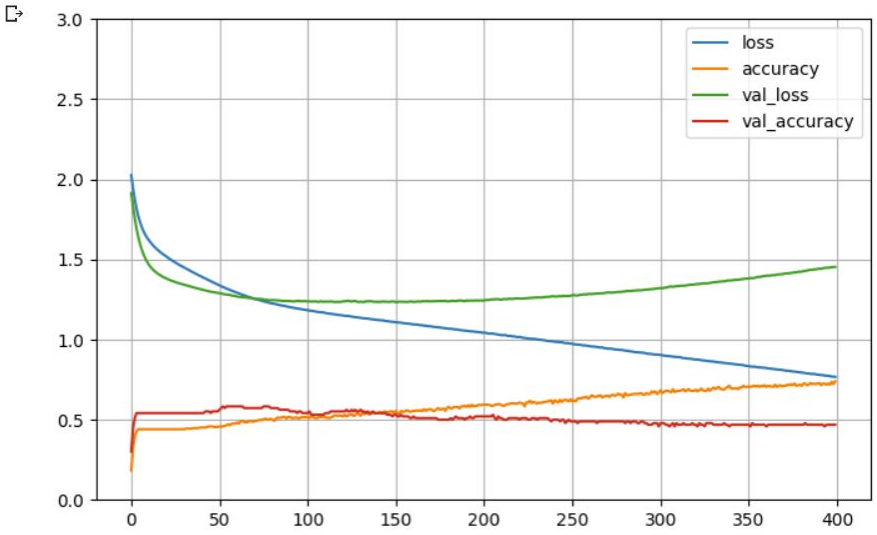
Thus, we retrain the network, and this time training it only 120 times to avoid overfitting.

6 Results

By running our testing data through the network, we generate the probability each team obtains a certain tournament outcome. When tested on the testing data allocated earlier, we find that the network correctly classifies team's tournament outcome around 50 percent of the time. While this is an improvement over the mere 1/8 chance of correctly predicting a team's tournament outcome without any knowledge of the team, the network's accuracy falls short of being sufficient to make accurate tournament predictions.

Additionally, after feeding the network our final testing data, the 2023 Tournament Teams, we find that Kansas, UCLA, Arizona, and USC have the best chance of winning the tournament. We base this on the fact that these teams

Figure 1: Graph of Training Process



were given the highest probability of belonging to the class "Champion" by the network. This output is visualized below.

7 Conclusion

While the network yielded predictions much more accurate than a random guess, it still was unable to predict at a high enough accuracy to warrant using in the future.

8 Appendix

Table 4: Predicted Tournament Outcomes for 2023 Tournament Teams

Team	Predicted Final Placement
Houston	Round of 32
Alabama	Round of 32
Creighton	Round of 32
UConn	Round of 32
San Diego St.	Round of 32
Purdue	Round of 32
Arizona	Sweet 16
West Virginia	Round of 64
FAU	Round of 64
Princeton	Round of 64
Miami FL	Round of 64
Virginia	Round of 32
TCU	Round of 64
Kansas St.	Round of 32
Marquette	Sweet 16
Gonzaga	Round of 64
Texas	Round of 32
UCLA	Round of 32
Tennessee	Round of 32
Saint Mary's	Round of 32
Kansas	Round of 32
Baylor	Round of 64
Memphis	Sweet 16
Arkansas	Round of 64
Xavier	Round of 64
Auburn	Round of 64
Iowa St.	Round of 32
Duke	Round of 32
Texas A&M	Round of 64
Utah St.	Round of 64

Table 5: Predicted Tournament Outcomes for 2023 Tournament Teams

Michigan St.	Round of 64
Maryland	Round of 32
Indiana	Round of 32
Kentucky	Round of 64
Boise St.	Round of 64
Northwestern	Round of 64
Iowa	Round of 64
NC State	Round of 64
Penn St.	Round of 64
Illinois	Round of 64
Providence	Round of 64
USC	Round of 32
Arizona St.	Round of 64
Missouri	Round of 64
Miss St.	Round of 32
Drake	Round of 64
Iona	Round of 64
VCU	Round of 64
C of Charleston	Round of 64
Kent St.	Round of 64
Nevada	Round of 64
Pittsburgh	Round of 64
Oral Roberts	Round of 64
Louisiana Lafayette	Round of 64
Furman	Round of 64
Montana St.	Round of 64
Vermont	Round of 64
Colgate	Round of 64
Grand Canyon	Round of 64
Kennesaw St.	Round of 64
UC Santa Barbara	Round of 64
NKU	Round of 64
UNC Asheville	Round of 64
Texas A&M Corpus Chris.	Round of 64
Howard	Round of 64
Southeast Missouri St.	Round of 64
Fairleigh Dickinson	Round of 64