Predicting March Madness Using Neural Networks

Photo of UConn hoisting the trophy

March Madness is one of the most watched televised events in the US. Every year, an average of over 9 million viewers tune in to watch this chaotic and exciting basketball tournament. Some tune in for the love of the game, others root for their team or alma mater, but there is one motivation to watch that rises above all others - the desire to have a perfect bracket.

Achieving a perfect bracket is no small feat, however. The odds of picking every single matchup correctly is 1 in 9,223,372,036,854,775,808 (approximately 9.2 quintillion)! That's a 1 in 120.2 billion chance (via (NCAA). To put this into perspective, a group of researchers at the University of Hawaii estimated that there are 7.5 quintillion grains of sand on Earth. If you selected one at random, the odds that someone guessed the exact grain you selected would still be 23% better than picking a perfect bracket. After almost 100 years of the tournament's existence, no one has ever been close to a perfect bracket.

It is evident that there is an extremely high level of unpredictability in this tournament, which is why it is so beloved. There will ALWAYS be upsets. Take this year's 11th seeded NC State for example. Who would've thought that they would make it all the way to the Final Four? And what about 14th seeded Oakland taking down 3rd seeded Kentucky in the first round behind a player that no one knew before the game (Jack Gohlke)? These types of games are what keep the tournament fresh and continue attracting an audience. But what do the top teams do to stay ahead and avoid embarrassment?

The answer to that question lies in analytics. In today's sports world, data drives everything. At an increasing rate, advanced metrics are being developed and are able to dramatically help teams with their play calls or help viewers with their decisions, both on and off the court. In-game analytics have allowed teams to make adjustments and data-informed decisions to get a leg up on their competitors.

In our Artificial Neural Networks course, our project team dove into the world of neural networks and discovered what they are capable of. And in the month of March, we couldn't help but wonder: What would it look like if we fed these advanced metrics into a neural network to make predictions for March Madness?

The first step to solving this question was to figure out which metrics and analytics we would feed our neural network. It was important to find data that covered a sufficiently large amount of games and tournament results for the network to learn from. At this point, we discovered that there was a DIS group that had attempted to do this assignment in years past, but

were ultimately unsuccessful. This group used raw team statistics, such as Free Throw percentage and 3-point percentage, as their training data. Our group believed that there were other metrics available that could help to create a more accurate and distinct model. For this reason, we landed on using the Pomeroy College Basketball Ratings (better known as KenPom).

The KenPom rankings were an ideal pick for our training data because they account for these raw team statistics, but combine them into more holistic categories that better capture a team's body of work. This data set mainly focuses on a teams offensive, defensive, and overall efficiencies. Essentially, these categories calculated how many points each team scored per 100 possessions (adjusted offensive efficiency), and how many points were scored on them per 100 possessions (adjusted defensive efficiency), with the difference between the two being the overall adjusted efficiency margin. This dataset also factored teams' strength of schedules into their rankings, averaging the above efficiencies of every opponent. The results are then compared and listed in ranking form each year. The KenPom rankings were invented in 2013, so we had 10 years of regular season and tournament data to go off of, before we fed our network the 2024 data we wanted it to predict. After cleaning up the data and organizing it into a CSV file, we were now ready to move onto our next step - deciding which model we wanted to use.

Ultimately, we decided on developing two models to accomplish this. The first one would be used to predict which round of the tournament a team would be eliminated in. The second would process game by game to simulate the bracket. Each team would be randomly matched up with another until the model narrowed it down to 1 winner.

Using the historical KenPom data from 2013-2023, we wanted the first model to predict how far it thought the team would go based on the historical data from 2013-2023. One model's training data would have the team names, and the other would not. We used embedding to prepare the team names for the model, which worked quite well. The main point of splitting the models was to see whether or not there was much of a difference and if the model would pick up a bias based off of the team's performance in past tournaments.

We began developing this model by loading and preprocessing our data. Historical team data, vital for training our prediction model, was imported from CSV files. Each team was assigned a unique index, including a category for 'unknown' to manage missing or new team entries.

Moving on to feature engineering, we employed various techniques to prepare our data for machine learning. Categorical variables, like conference affiliations, were transformed into a binary format using one-hot encoding. Additionally, the target variable representing postseason results underwent label encoding followed by one-hot encoding to facilitate classification tasks.

Features were normalized using MinMaxScaler to ensure uniform contributions to model training.

Our neural network model was designed with two inputs: a categorical input representing team indices processed through an Embedding layer, and numerical features derived from each team's historical data. These inputs were concatenated and passed through several dense layers with ReLU activation, culminating in a softmax layer for classification. The model was compiled using the Adam optimizer and categorical cross-entropy loss, then trained on the historical data.

For prediction and optimization, upcoming season data was preprocessed similarly to historical data, ensuring compatibility with the trained model. Predictions were generated using the trained model. Additionally, we employed Integer Linear Programming (ILP) techniques. Binary decision variables were defined for each postseason stage and team, with the objective function aiming to maximize the sum of probabilities of teams reaching each stage, weighted by binary variables. Constraints were set to ensure a valid tournament structure, and the ILP model provided the most probable postseason configuration.

In post-processing, ILP results were used to assign postseason predictions to teams in the upcoming season dataset. Predictions were then saved to a new CSV file.

Throughout this process, our feature engineering insights emphasized the importance of indexing teams, enabling the model to capture subtle interactions specific to each team through embeddings. Normalization and encoding steps standardized input data, facilitating efficient learning for our neural network by presenting inputs in a consistent format.

With the second model, we continued to use the KenPom data set, we started along the same lines by predicting which round a team would be eliminated. However, we hit a roadblock when the model initially produced either too many losers of one round or not enough winners of another round. We then decided to pivot and create a legitimate bracket prediction. Using a sequential neural network, there were initially some problems with overfitting, but after implementing early stopping, L2 regularization, and dropout layers, we were able to get a better result. More satisfied with these numbers, we started simulating matchups. Ultimately, the results were somewhat successful, as the model picked teams with a higher ranking and better statistics to win. However, there were no upsets chosen, which are very common in March Madness. From this year’s data, the model most often predicted Connecticut or Purdue to win, as a result of their high-end talent and elite KenPom statistics. In certain iterations, it would also occasionally feature a Florida or Duke champion. Below is an example of what the simulated winners would look like:

Round of 64: ['UAB ', 'Connecticut ', 'Akron ', 'Creighton ', 'Utah St. ', 'Alabama ', 'Florida ', 'Illinois ', 'Dayton ', 'Kentucky ', 'Marquette ', 'Arizona ', 'Duke ', 'N.C. State ', 'Grand Canyon ', 'Colorado St. ', 'Tennessee ', 'Florida Atlantic ', 'Drake ', 'Purdue ', 'Baylor ', 'Nevada ', 'Kansas ', 'Auburn ', 'Iowa St. ', 'South Dakota St. ', 'Vermont ', 'Houston ', 'TCU ', 'Clemson ', 'Nebraska ', 'Northwestern ']

Round of 32: ['Creighton ', 'Baylor ', 'Nebraska ', 'Purdue ', 'Dayton ', 'Connecticut ', 'Marquette ', 'Auburn ', 'Colorado St. ', 'Florida ', 'Iowa St. ', 'Clemson ', 'Duke ', 'Tennessee ', 'Alabama ', 'Houston ']

Sweet 16: ['Tennessee ', 'Connecticut ', 'Iowa St. ', 'Florida ', 'Purdue ', 'Auburn ', 'Creighton ', 'Houston ']

Elite 8: ['Houston ', 'Florida ', 'Connecticut ', 'Purdue ']

Final Four: ['Connecticut ', 'Purdue']

Champion: ['Connecticut ']

While it is certainly not perfect, the model correctly predicted this year's champion (UConn) to win it all. Other notable correct predictions include Purdue to finish 2nd, Tennessee to make the Elite 8, and correctly choosing 11 of the Sweet 16 teams. Our model was also not afraid of picking upsets, as it regularly had 7th seeded Florida making it far into the tournament (even though they lost in the round of 64 this year). The model had a particular liking to 1 seeds, and based on historical data, that assumption makes perfect sense. 1 seeds historically do excel in the tournament, having won 25 of the last 39 national championships (64.1%), including 13 of the last 17 (76.5%) via Forbes.

Overall, we were pleased with the results of both models. Choosing KenPom data proved to be a critical addition to our models, as the set provided a good measurement on which teams might advance farther and which have the ability to create an upset. Practically speaking, these models can continue to be used in future years as an information tool before filling out your March Madness bracket, as they can provide potential upset picks and predict which teams have the best chances to make it all the way.

Some improvements we could make to our models would be to include even more specific and tailored datasets, and to incorporate actual team matchups. Although KenPom did provide a fairly accurate representation of the strengths of each team, there could always be more things to train the model on. With more time and resources, utilizing statistics such as individual player performances and how well teams are playing headed into March could provide more

accurate insights into the tournament. A lot of upsets were caused by individual players that flew under the analytical radar, showing some glimpses of that potential in the regular season. Using player data from the regular season can help to identify which teams might have this advantage, and which players have the opportunity to breakout and be the hero for their Cinderella team in March. Additionally, matching up each team in their correct bracket position gives our model the ability to judge teams against one another, instead of the entire field of 68. While the margin for error is much more thin, this can provide a more proper prediction and result in actual outcomes from the tournament. This could give our models a greater use when predicting the tournament as a whole.

This project was fascinating and we are grateful for the opportunity to combine an interest with what we learned in neural networks. Working on these models gave us insights into the practicality of neural networks, and the potential roadblocks and limitations you can run into when gathering data and developing models from the ground up. We look forward to implementing future data into this model and using it to fill out our brackets for years to come. Maybe one day this will help us finally predict the perfect bracket…

UpcomingSeasonPredictions

Year	KenPom Rank	Team	CONF	Seed Number	AdjEM	AdjO	AdjD	AdjT	SOS AdjEM	SOS OppO	SOS OppD	NCSOS AdjEM	Team_idx	Predicted_POSTSEASON
2024	1	Connecticut	BE	1	36.43	127.5	91.1	64.6	12.42	113.2	100.8	-3.4	76	Champion
2024	2	Houston	B12	1	31.17	118.9	87.7	63.5	11.57	111.9	100.3	-1.02	168	none
2024	3	Purdue	B10	1	30.62	125.2	94.6	67.0	14.65	114.4	99.8	10.58	117	Sweet Sixteen
2024	4	Auburn	SEC	4	27.99	120.4	92.4	70.0	9.49	111.9	102.4	1.47	164	Round of 32
2024	5	Tennessee	SEC	2	26.61	116.8	90.2	69.3	13.35	114.6	101.2	8.97	96	Sweet Sixteen
2024	6	Arizona	P12	2	26.55	120.2	93.7	72.2	11.12	112.2	101.1	10.47	3	Elite Eight
2024	7	Duke	ACC	4	26.47	121.6	95.2	66.4	10.07	111.1	101.1	-0.04	14	Round of 32
2024	8	Iowa St.	B12	2	26.47	113.9	87.5	67.2	10.43	111.3	100.8	-7.05	23	Runner Up
2024	9	North Carolina	ACC	1	26.19	119.7	93.5	70.6	12.17	112.6	100.5	6.99	43	none
2024	10	Illinois	B10	3	24.53	125.5	101.0	69.8	11.92	111.8	99.9	-2.33	20	Round of 32
2024	11	Creighton	BE	3	24.22	120.9	96.7	66.8	11.96	112.3	100.4	4.99	12	Elite Eight
2024	12	Gonzaga	WCC	5	23.17	122.6	99.4	68.9	5.21	109.3	104.1	7.7	18	Elite Eight
2024	13	Marquette	BE	2	23.02	118.2	95.2	69.1	13.6	113.4	99.8	8.21	31	Round of 32
2024	14	Alabama	SEC	4	22.96	126.0	103.0	72.6	14.71	115.1	100.4	9.46	163	Sweet Sixteen
2024	15	Baylor	B12	3	21.9	122.4	100.5	65.8	13.44	112.4	99.0	3.36	71	Round of 32
2024	16	Michigan St.	B10	9	20.58	114.3	93.7	65.3	13.03	114.2	101.1	4.62	35	Round of 32
2024	17	Wisconsin	B10	5	20.06	119.2	99.2	65.5	14.04	113.2	99.2	6.22	68	none
2024	18	BYU	B12	6	19.96	119.8	99.9	69.2	9.21	110.1	100.9	-5.67	72	none
2024	19	Clemson	ACC	6	19.44	117.7	98.3	66.4	12.09	113.5	101.4	4.91	166	Sweet Sixteen
2024	20	Saint Mary's	WCC	5	19.43	114.5	95.0	62.5	3.54	108.5	104.9	5.2	55	Round of 32
2024	22	San Diego St.	MWC	5	19.36	113.4	94.0	66.2	11.16	112.8	101.7	10.56	56	none
2024	23	Kentucky	SEC	3	19.29	122.2	102.9	72.7	8.94	111.4	102.5	-0.17	82	Sweet Sixteen
2024	24	Colorado	P12	10	19.03	118.4	99.3	67.6	9.34	110.9	101.6	-1.47	10	none
2024	25	Texas	B12	7	18.77	116.5	97.7	67.4	10.91	111.4	100.5	-3.13	97	Round of 32
2024	26	Florida	SEC	7	18.19	120.0	101.8	72.0	11.0	112.5	101.5	2.02	15	Round of 32
2024	27	Kansas	B12	4	17.94	113.5	95.6	69.2	13.61	113.0	99.3	5.2	25	Round of 32
2024	29	New Mexico	MWC	11	17.8	114.4	96.6	72.8	7.04	110.2	103.2	-0.5	41	none
2024	30	Nebraska	B10	8	17.55	116.2	98.7	69.0	9.39	110.9	101.6	-5.41	89	none
2024	31	Texas Tech	B12	6	17.32	117.4	100.1	66.8	10.24	111.1	100.9	-2.63	139	Round of 32
2024	32	Dayton	A10	7	17.3	117.7	100.4	64.5	6.88	109.3	102.4	6.97	77	none

UpcomingSeasonPredictions(no-names)

Year	KenPom Rank	CONF	Seed Number	AdjEM	AdjO	AdjD	AdjT	SOS AdjEM	SOS OppO	SOS OppD	NCSOS AdjEM	Predicted_POSTSEASON
2024	1	BE	1	36.43	127.5	91.1	64.6	12.42	113.2	100.8	-3.4	Champion
2024	2	B12	1	31.17	118.9	87.7	63.5	11.57	111.9	100.3	-1.02	Final Four
2024	3	B10	1	30.62	125.2	94.6	67.0	14.65	114.4	99.8	10.58	None
2024	4	SEC	4	27.99	120.4	92.4	70.0	9.49	111.9	102.4	1.47	Elite Eight
2024	5	SEC	2	26.61	116.8	90.2	69.3	13.35	114.6	101.2	8.97	Final Four
2024	6	P12	2	26.55	120.2	93.7	72.2	11.12	112.2	101.1	10.47	Elite Eight
2024	7	ACC	4	26.47	121.6	95.2	66.4	10.07	111.1	101.1	-0.04	Elite Eight
2024	8	B12	2	26.47	113.9	87.5	67.2	10.43	111.3	100.8	-7.05	Final Four
2024	9	ACC	1	26.19	119.7	93.5	70.6	12.17	112.6	100.5	6.99	Elite Eight
2024	10	B10	3	24.53	125.5	101.0	69.8	11.92	111.8	99.9	-2.33	Sweet Sixteen
2024	11	BE	3	24.22	120.9	96.7	66.8	11.96	112.3	100.4	4.99	Sweet Sixteen
2024	12	WCC	5	23.17	122.6	99.4	68.9	5.21	109.3	104.1	7.7	None
2024	13	BE	2	23.02	118.2	95.2	69.1	13.6	113.4	99.8	8.21	None
2024	14	SEC	4	22.96	126.0	103.0	72.6	14.71	115.1	100.4	9.46	Sweet Sixteen
2024	15	B12	3	21.9	122.4	100.5	65.8	13.44	112.4	99.0	3.36	Round of 32
2024	16	B10	9	20.58	114.3	93.7	65.3	13.03	114.2	101.1	4.62	Sweet Sixteen
2024	17	B10	5	20.06	119.2	99.2	65.5	14.04	113.2	99.2	6.22	Sweet Sixteen
2024	18	B12	6	19.96	119.8	99.9	69.2	9.21	110.1	100.9	-5.67	Round of 32
2024	19	ACC	6	19.44	117.7	98.3	66.4	12.09	113.5	101.4	4.91	None
2024	20	WCC	5	19.43	114.5	95.0	62.5	3.54	108.5	104.9	5.2	None
2024	22	MWC	5	19.36	113.4	94.0	66.2	11.16	112.8	101.7	10.56	None
2024	23	SEC	3	19.29	122.2	102.9	72.7	8.94	111.4	102.5	-0.17	Sweet Sixteen
2024	24	P12	10	19.03	118.4	99.3	67.6	9.34	110.9	101.6	-1.47	None
2024	25	B12	7	18.77	116.5	97.7	67.4	10.91	111.4	100.5	-3.13	Round of 32
2024	26	SEC	7	18.19	120.0	101.8	72.0	11.0	112.5	101.5	2.02	Round of 32
2024	27	B12	4	17.94	113.5	95.6	69.2	13.61	113.0	99.3	5.2	Round of 32
2024	29	MWC	11	17.8	114.4	96.6	72.8	7.04	110.2	103.2	-0.5	None
2024	30	B10	8	17.55	116.2	98.7	69.0	9.39	110.9	101.6	-5.41	Round of 32
2024	31	B12	6	17.32	117.4	100.1	66.8	10.24	111.1	100.9	-2.63	Round of 32
2024	32	A10	7	17.3	117.7	100.4	64.5	6.88	109.3	102.4	6.97	Round of 32