**2022 March Madness Prediction**

Created by:

Annie Titus, Jiebin (Alex) Zhu, Matthew Laken, Zachary Chipman, Farahin Choudhury​

*Based on the XGBoost machine learning model’s prediction, the predicted champion for the NCAA 2022 March Madness Tournament is Gonzaga University. The accuracy that’s provided by this model is ~76% at predicting that Gonzaga would have the highest possibility to win the tournament. There are other ways the model can be improved further as well. However, for the purpose of bragging rights with co-workers, friends, and families, Gonzaga is the recommendation our group would think is going to win everything and take home the trophy in New Orleans.*

**Introduction**

Each year in the month of March, there’s a sports competition that is highly covered by the media and favored by fans all over the United States. This is the famous NCAA March Madness tournament where 64 college basketball teams play against each other in a series of knockout matches until a victor is determined. This tournament also calls out to all fans to create a bracket that predicts the outcome of these matches leading up to determine the winner. Some fans do their research and try to fill out the best bracket, some will just purely guess, or some will utilize their knowledge in data science to predict the winner. It is worth noting that there are tens of millions of brackets that have been filled out throughout the history of the competition, yet no one has produced the *perfect* one.

***SPECIFICATION***

**Problem Statement**

The objective of this analysis was to determine the winner of the March Madness tournament. To do that, historical data from past tournaments was applied to the models created for this analysis. Tournament matchups outputted from the data helped to arrive at a final winner. The matchups pertain to brackets that typically outline which teams are playing against each other; the matchups are denoted by the teams’ ID numbers.

**Assumptions**

Throughout the course of building the models for predicting the March Madness champion, there were some assumptions that we had to take:

* The main goal was to predict winners with the necessary information but to obtain an even better result we included some of the round 1 match results in the model for reference to see how our original submission is holding up.
* Simulations of the probability of teams winning in each match were returned. We assumed as long as the probability for a team is greater than 50% that team would advance to the next round.
* Our brackets will not include the “human touch” and are solely reliant on the model itself
* The initial bracket showcase to the class on March 16, 2022, wasn’t done with a specific “seed”, thus not replicable. However, any brackets created afterward were done with a “seed” so results can be reproduced.

**Hypothesis**

Without any additional context, the best piece of information we have for teams is their seed number. Typically, the lower the seed number the better the school. We believe that seed 1 or seed 2 schools would make it to the final four with a seed 1 school winning. There’s no specific preference of the team that would win the tournament so we didn’t want to branch out that far. In the case where there’s only one seed 1 school in the final four in our predicted bracket, we believe that seed 1 school will become the champion.

**Data Collection**

The main source of data used for creating models to predict the champion of the March Madness tournament was retrieved from Kaggle, which can be found [here](https://www.kaggle.com/c/mens-march-mania-2022/data). For this project, data for the tournaments from the years 2016 to 2021 were collected, except for 2020 due to the start of the COVID-19 pandemic. The data was provided in a structured format as .csv files, which made it easy to import into Google Colab.

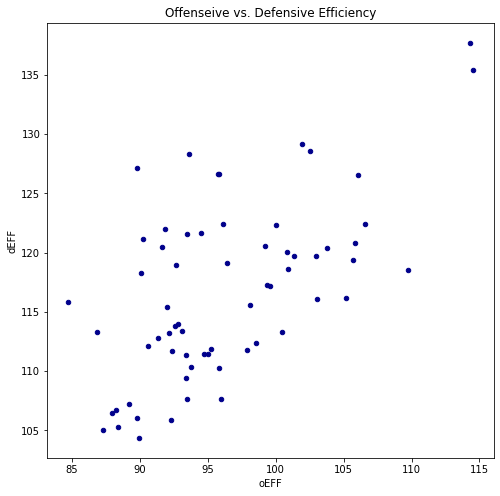
**Data Preparation and Preprocessing**

Our data obtained from the Kaggle tournament was clean for the most part, and it did not have missing values which help us have a more robust analysis. However, we needed to create new data frames to implement into our models. This was achieved by creating various functions and applying the panda's library and corresponding modules to subset, calculate values and merge data frames. The aggregated data frames (**megareg\_df** and **megatourney\_df**) were the finalized data frames to be used for the train and test data in the models. Ultimately, this meant this process was rather smooth and didn’t encounter any absurd outliers or blockers preventing us from moving along with the analysis.

***OBSERVATION***

**Data Exploration**

Before we start making the models, it would be a good idea to look at the statistics of each team in the tournament.



*Figure 1-1*

Figure 1-1 shows the Offensive versus Defensive Efficiency (oEFF vs dEFF) for every team in the tournament. For the sake of clarity, here is how these metrics are calculated:

oEFF = Points Made + Assists + Offensive Rebounds

dEFF = Points Missed (by Opponent) + Steals + Blocks + Defensive Rebounds + Turnovers (by Opponent) - Personal Fouls

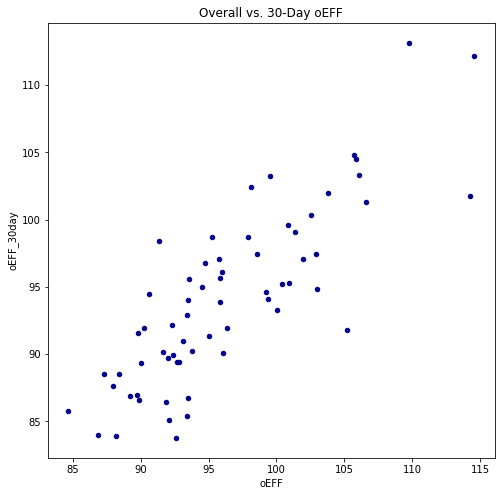
We can see some correlation between the two metrics, which means that teams with good offense have a good defense as well. But two teams outperform everyone else in the field, Gonzaga and Arizona (Figures 1-2).

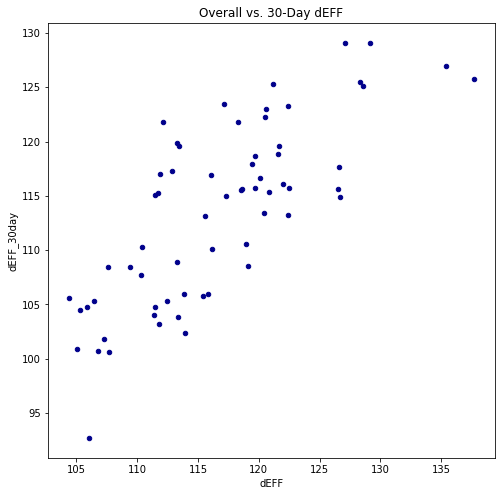




*Figures 1-2*

Another interesting thing to point out is who had the highest 30-day offensive and defensive efficiency. Please read below.



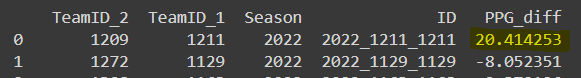


*Figures 1-3*

This metric was designed to quantify momentum heading into the tournament. Both of these graphs (Figure 1-3) indicate that good teams all season tended to carry that momentum into the tournament.

Iowa had the highest 30-day offensive efficiency (113.15) and Auburn had the highest 30-day defensive efficiency (129.1). As of the writing of this paper, both teams are out of the tournament, being upset by much weaker teams. Conversely, Iowa St. was one of the worst teams in both offensive and defensive efficiency but still made it to the Sweet 16. Michigan is another overachiever, making it to the Sweet 16 while having the second-worst 30-day defensive efficiency (only behind Texas). This may suggest that momentum heading into the tournament is not an accurate predictor of tournament success.

Our initial exploration also consisted of calculating the differences in variables between the actual teams competing in this year’s tournament. We compiled a table to show us a combined statistic which we then will use in generating our prediction. For example, in Figure 1-4, we show our comparison of the matchup between **Gonzaga (1211)** and **Georgia State (1209)**:



*Figure 1-4*

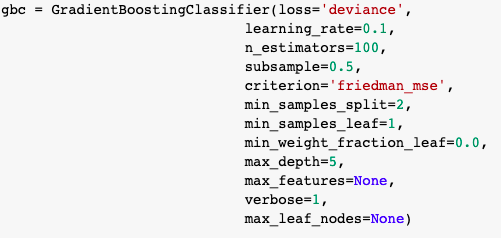
We see that the point differential between the two teams shows Gonzaga with +20.4 points. This generally shows how well Gonzaga would do when playing against Georgia State. At the time of writing this report, Gonzaga has already beaten Georgia State in round 1 and advanced into round 2.

***ANALYSIS***

**Model Building**

The two models created and implemented in this analysis were XGBoost and convolutional neural networks. XGBoost, or eXtreme Gradient Boosting, is a library meant for computing gradient boosting at scale. The premise of gradient boosting is to provide a prediction from classification tasks defined in the model. Since XGBoost requires a classification metric for this model, the output of the training model provides the ‘logloss’ error as the evaluation. XGBoost was selected due to its accuracy and time performance. A model that doesn’t take too long to run, but still has a rather accurate result is what our team was looking for.

The following screenshot in Figure 2-1 shows how the XGBoost model was set up for replication:

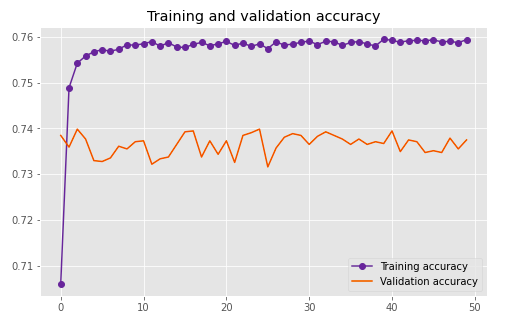


*Figure 2-1*

In addition to setting up the XGBoost model in this manner, we used np.random.seed(**5**) before running this to ensure the same set of randomized generated numbers are the same to produce matching results.

Our team decided to explore another model for better reference. The other model, using a convolutional neural network (CNN), did not produce the ideal outcome but it is still worth exploring and including as part of the overall analysis. This deep learning model implemented the activation function ‘rectified linear unit’, also known as ReLU (activation = ‘relu’). This made the most sense to use and seemed more efficient than the sigmoid activation. In general, neural networks are quite powerful and can provide insights other algorithms might not be able to.

As the title of this graphic in Figure 2-2 suggests, the two plots show the accuracies of the training and validation results for the CNN model:

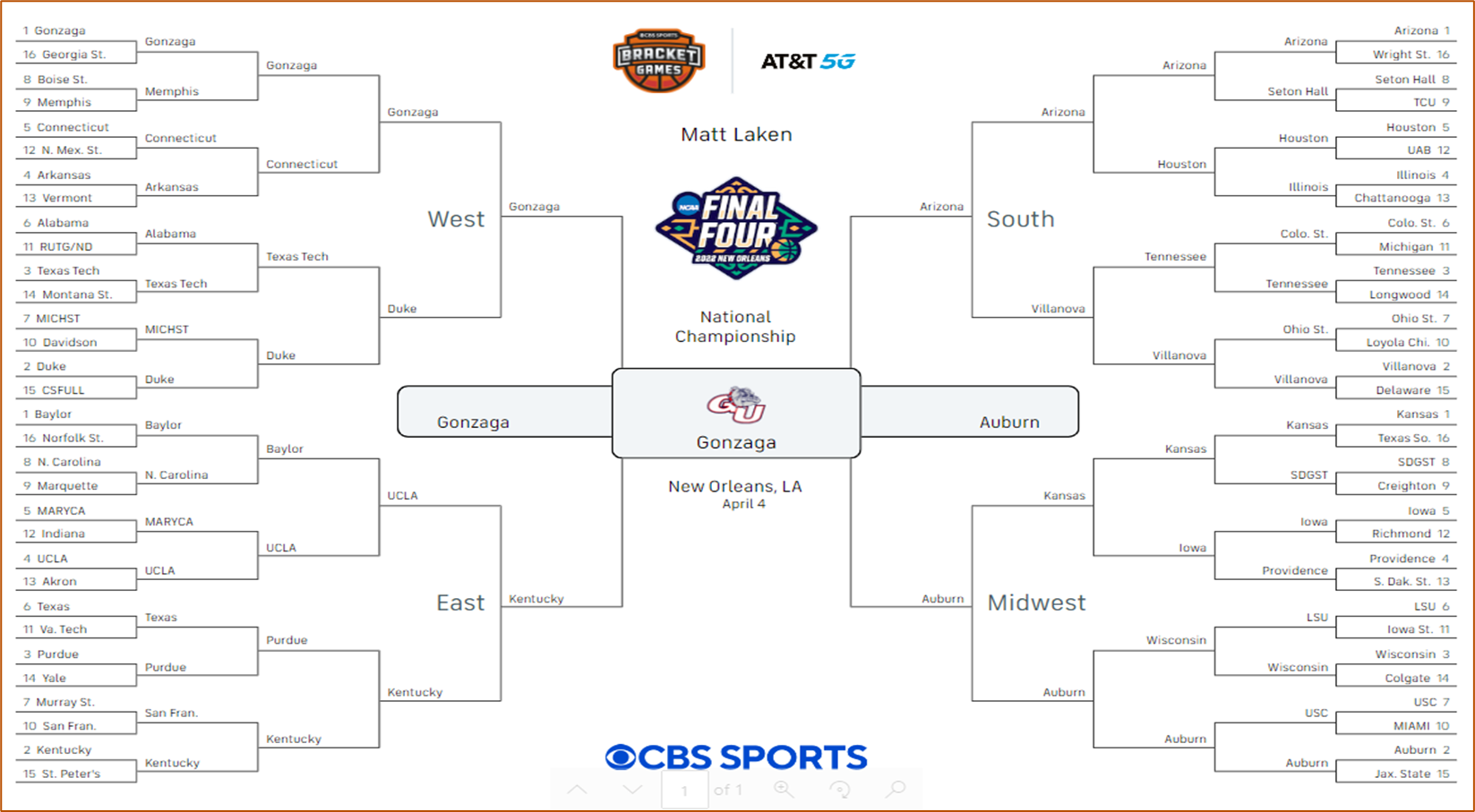


*Figure 2-2*

A key thing to point out here is that Figure 2-2 is zoomed in and hard to convey the scale of the training and validation accuracy. If zoomed out, the lines would look a lot closer to each other than they are currently.

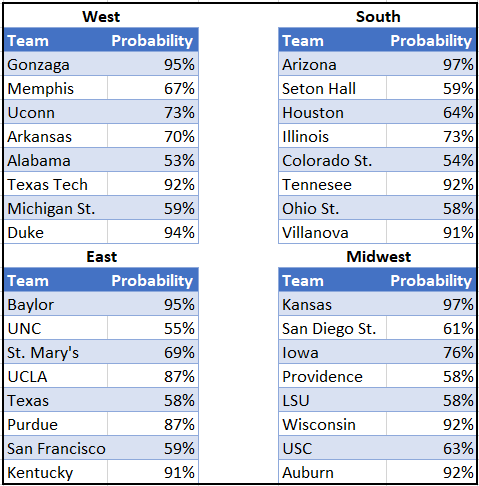
After understanding the results produced by the two different models, we decided to use the XGBoost model to continue with our prediction in selecting the champion of the 2022 March Madness Tournament since it was the main model we created before submitting the brackets. The CNN model was built afterward and used as a reference only since the predictions were similar with Gonzaga winning the tournament and Baylor, Arizona, and Auburn being the other teams in the final four.

***RECOMMENDATION***

****

*Figure 3-1*

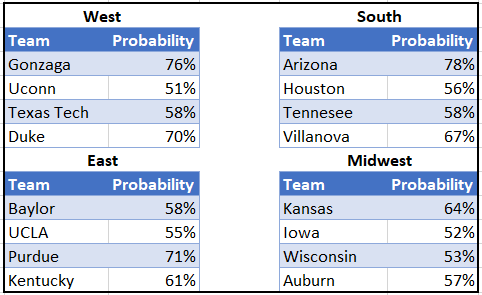
We filled out our recommended bracket for the 2022 March Madness tournament based on analyzing the results of both the XGBoost and the CNN models seen in Figure 3-1. Below is a table (Figure 3-2) containing the probability of the round of 64 winners from the XGBoost Model which is supported by the CNN results (used as reference) from our code. To reiterate an assumption made earlier, the selection to advance was made solely on teams who had a higher probability of winning.



*Figure 3-2*

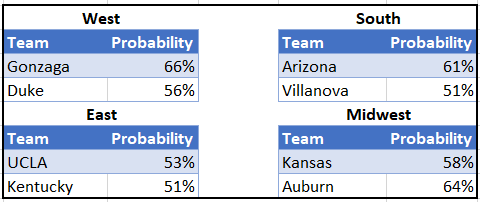
After the initial round, we made more predictions solely on the teams’ probability of winning for future rounds leading up to determine the winner.

The recommendations for the round of 32 based on probability of winning are as follows:



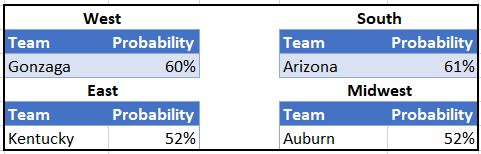
*Figure 3-3*

The recommendations for the Sweet 16 based on probability of winning are as follows:



*Figure 3-4*

The recommendations for the Elite 8 based on probability of winning are as follows:



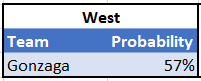
*Figure 3-5*

The recommendations for the Final 4 based on probability of winning are as follows:



*Figure 3-6*

The recommendation for the Final round based on the probability of winning is as follows:



*Figure 3-7*

As a part of the assumptions listed earlier, teams that had a probability of greater than 50% were selected as winners for that round and can advance to the next one. From this analysis here, it can be seen that some matchups are close to the 50-50 probability. This means that there could be some upsets within the final results, but this is the best our model could predict here and therefore the method for predicting a winner. We tried to eliminate the “human intuition” element from this when selecting the winners and chose based on our assumption. Ultimately we had Gonzaga playing against Auburn in the finals based on Figure 3-6 and Figure 3-7, with Gonzaga coming out as the champion of the 2022 March Madness Tournament. This is also the selection we’d hope you pick because those bragging rights are at stake here!

**Wrapping Up**

The prediction obtained from the model was aligned with our group’s hypothesis at the beginning with how seed 1 or seed 2 teams would appear in the final four rounds. At the time of writing this report, there have been some upsets to our bracket, but our selected winner, Gonzaga, is still in the tournament.

There are many ways to tackle this specific objective of creating the “perfect” bracket in the hopes of predicting the final winner of the March Madness tournament, such as using different data sets and trying new modeling algorithms. Our approach here was just one of many different paths to take. We stuck with the Kaggle data set since it was detailed and provided a lot of information that we could tailor into an XGBoost and CNN model. For future reference to improve predicting the winner, the [Kenpom dataset](https://kenpom.com/) would be a great addition to a model since it includes a “luck” factor attribute within it. Another way of doing so is to try and factor in some human intuition when making the final bracket selection on games that have rather close probabilities such as 53% to 47%. Furthermore, different modeling techniques can be used, perhaps even just reconfiguring the XGBoost model setup here would work as well!

Ultimately, there’s no right or wrong way to try to figure out who’s going to win the tournament. Since people have different preferences to the models that they want to use and depending on your skillset and familiarity with certain models, that’s what makes your prediction and selection so unique. Lastly, just remember to have fun when filling your brackets out and be prepared for upsets as that tends to happen! Thanks for reading!

**REFERENCES**

*March machine learning mania 2022 - NCAAM*. Kaggle. (2022). Retrieved March 19, 2022, from https://www.kaggle.com/c/mens-march-mania-2022/data

Pomeroy, K. (2022). 2022 Pomeroy College Basketball Ratings. Retrieved March 19, 2022, from https://kenpom.com/

reHOOPerate. (2020, January 15). *New and improved March madness neural network for 2020*. Medium. Retrieved March 19, 2022, from https://medium.com/re-hoop-per-rate/new-and-improved-march-madness-neural-network-for-2020-c154aa1041b7

Sblendorio, D. (2021, March 19). *Fantasy march madness: How to predict winners*. ActiveState. Retrieved March 19, 2022, from https://www.activestate.com/blog/fantasy-march-madness-how-to-predict-winners/