# **Investigating the Most Effective Metrics in Predicting March Madness Success**

*Group members: Dom Fenoglio: djf49, Cameron Morgan: cwm36, Ricardo Ureña: ru15, Chris Johnson: cgj17*

### **Section 1: Introduction and Research Questions**

This very weekend, Duke men’s basketball will compete in the Sweet 16 of the NCAA tournament, looking to compete for its sixth national championship. However, no one truly knows how the Blue Devils will do. March Madness, the colloquial name for the NCAA men’s basketball tournament each spring, is known for its massive upsets and unpredictability. However, what if we could try to find a method to the madness? In other words, what factors best predict a team’s postseason success, ranging from offensive and defensive efficiency to tempo and free-throw shooting? Our research questions are as followed:

1. What factors are most correlated with success in March Madness?
2. What factors may cause a team to be overvalued heading into March Madness?
3. To what degree is it possible to predict a team’s success in March Madness?

Our group wants to examine team statistics in college basketball in order to answer this question and learn more about why certain teams are successful. While this question is not necessarily novel, there have been many people who have tried (and failed) to construct a perfect bracket. We believe that our research is **substantial** because we can use our newfound knowledge of data science to get a better grasp on what drives success in college basketball. Instead of simply looking at a team's point differential or their net rating (a measure of how many points they score and allow per 100 possessions), we also want to look at other factors. Specifically from our data sets, we want to look at adjusted tempo (the number of possessions a team gets a game), 3-point %, BARTHAG (probability of beating an average college team), and turnover rate.

We were able to pull data from the past 5 years and filter it so the data we used is from before March Madness began — so that in-tournament performance does not skew our data. This means that we can widen our scope to see why teams make the NCAA tournament in the first place. We also found a dataset that contains results of each team’s performance in the tournament, including how many games they won. The plethora of college basketball data as well as the statistic-based nature of the sport make this project **feasible**.

We believe our research will be **relevant** because we aim to discover what makes a college basketball team successful, which can have important consequences in the realms of sports betting, coaching, and player and team development.

For the prototype, we took the suggestion to look closely at specific variables and see how they are related to a team’s performance in March Madness. In our proposal, we did not explicitly write out our research questions, so we did so here. We will primarily focus on research question #1 at first so we can be better equipped to answer research question #3. Research question #2 may come as a result of our deeper look into question #1. In the prototype, we have primarily focused on question #1 using our box plots.

### **Section 2: Data Sources**

The bulk of our data was taken from Barttorvik.com. Bart Torvik contains a variety of downloadable CSV files, ranging from individual game stats to team performance over the course of the season. One data set we will look at is the T-Rank table which gives many statistics about every team in division one college basketball, including adjusted offensive efficiency, adjusted defensive efficiency, adjusted tempo, field goal percentage, turnover percentage, and many more. We took five T-Rank table csv files from the last five years of college basketball (not including 2020 as there was no tournament that year). We filtered by Regular Season on Barttorvik.com to get data that was at the end of the season, but before the start of the tournament. This way, how a team performs during the tournament does not affect any of the statistics, and one could use our analysis to predict how a team will perform before the tournament begins. We merged this data together with 538’s tournament ratings, to give us a complete understanding of how the teams performed in the tournament.

Link to data: [https://www.barttorvik.com/trank.php](https://www.barttorvik.com/trank.php#)

### **Section 3: What modules are we using?**

* Data Wrangling

We used data wrangling to clean up information across different data sets. We dropped unneeded columns and fit all the data so that it could easily be concatenated and merged later. We also used data wrangling to create two new columns based on the finish of the team in our final dataset. The first column showed if the team made the Final Four. The second column showed if the team was a Champion of March Madness in a given year.

* Combining Data

After wrangling all of our data, we used pd.concat and pd.merge to finish polishing up our data and merge it into one final dataset.

* Visualization

We then utilized seaborn to create catplots for a visualization about how certain statistical categories affect how far a team makes it into March Madness. The three statistics we chose to model are adjusted offensive efficiency, defensive offensive efficiency, and three-point percentage. We used these statistics because they are typically a good indicator of how a team performs in March Madness. We used seaborns relplot to examine adjusted offensive efficiency vs defensive efficiency, two-point shot percentage vs three-point shot percentage, and free throw rate vs free throw percentage, concerning whether a team won March Madness or not.

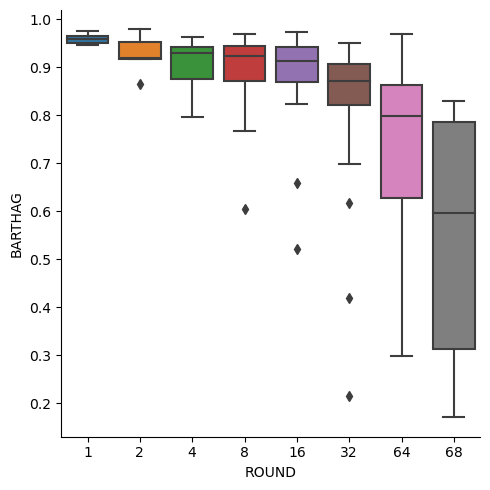
* Prediction and Supervised Machine Learning

We then utilized sklearn to create, fix, and predict a logistic model. We also utilized an accuracy score to see how our model did, and a confusion matrix to get a visualization of our predictions.

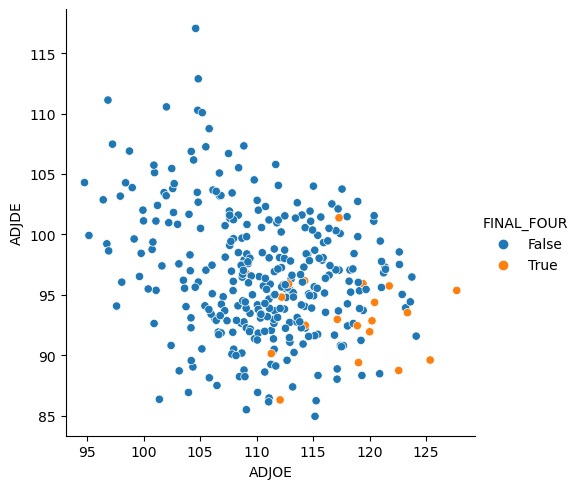
### **Section 4: Preliminary research and methods**

First, we needed to join our data sets together, since we both had data from before the tournament and tournament performance. Our first 5 data sets (each corresponding to one of the past five years) contained all of the same variables for pre-tournament metrics, so we were able to join them together using concat. A few important notes from this step is that we excluded data from 2020 since there was no NCAA tournament played that year because of COVID-19, and each data set contained multiple empty columns produced from exporting them from our primary source. Therefore, we dropped those columns before moving forward. Next, we joined our full five-year data with the tournament performance dataset, using an inner merge and joining by year and team to ensure that we finished with our final correct data.

Our next step was to create a few new columns that identified categorically whether a team won a championship and made the final four. This was done using the ROUND column, which has numeric values corresponding to how far a team advanced (e.g. 4 ~ final four, 32 ~ round of 32, etc.), so the data is actually categorical. We used this ROUND variable to look at some different columns and how they vary based on round. As we move forward, we are thinking of also creating a column related to this that simply counts the number of games a team has won, because that might be an easier thing to predict. <https://duke.box.com/s/wdshw22i3l0rqc2bgls0amge5pa3n098>



**We made a variety of boxplots examining the distribution of different numerical variables based on ROUND (again, 16 ~ Sweet 16, 4 ~ Final Four, etc.). This boxplot looks at BARTHAG, a measure of a team’s probability of beating an average team, calculated by averaging offensive and defensive adjusted efficiencies and putting it into a log formula. There appears to be a correlation between the two variables. That is, as teams advance farther in the tournament, their BARTHAG tends to increase. Additionally, as team advance farther, the variance in the BARTHAG tends to decreases drastically. An important note, however, is that the sample sizes for each group decreases in size as teams advance. This was generated using Seaborn’s catplot function.**

****

**This scatter plot is a representation of each team’s offensive and defensive efficiency (points scored and allowed per 100 possessions, respectively). An important note is that while a higher offensive efficiency is desired, a lower defensive efficiency is desired. We used Seaborn’s relplot to create this graph, and colored points by whether or not teams reached the Final Four. We can see that teams who reached the Final Four tend to have higher offensive efficiencies and lower defensive efficiencies, but the difference on offense is more pronounced.**

### **Section 5: Reflection and next steps**

We **have successfully** pulled, cleaned, and merged our data. This process was not too complicated, as we generally just removed empty columns and grouped entries by year and team. Next, we have started visualizing our data (boxplots and scatterplots). Finally, we have developed a rough logistic model along with confusion matrices to predict whether a team will make it to the Final Four, but it is still a work in progress.

We are **facing challenges** when it comes to making and editing our predictive models. Originally, we wanted to predict whether or not a team would make it to the Final Four. However, we found that the baseline model does almost as well as our logistic model because so few teams make the Final Four. Additionally, there are several upsets every year, which makes prediction difficult. Thus, we want to broaden the scope of our prediction (Elite Eight vs Final Four), and we want to try creating a KNN model.

**In regards to collaboration,** our team has strong communication, collaboration, and work dynamics. We have been meeting after Friday lecture, which has been a good framework for us. During this time, we work on the project, ask for help, and discuss potential next steps. One area that we can improve on is sharing our work with other members of the team. It can be difficult to follow the logic behind someone else’s code, whether it be for data cleaning, visualizing, or exploration. From now on, we plan to add comments with a brief explanation of what the code is doing.

Our **next steps** involve exploring new machine learning models (KNN) and improving our existing logistic model. Doing this will allow us to better predict Final Four teams and potentially find valuable, new insights regarding our data. Furthermore, we want to explore hypothesis testing on our data. For example, we could create confidence intervals for team values like EFGD% to characterize what Final Four team statistics are like. For this, I might want to include data from more March Madness years. Finally, we intend to create a data dictionary to explain what our column names in our data tables mean.