Julian Savini: savini.j@northeastern.edu

Benoit Cambournac: cambournac.b@northeastern.edu

Beckett Sanderson: sanderson.b@northeastern.edu

NCAA Tournament Team Predictions

Problem Statement and Background

At the same time this project was released, the teams which had made the tournament were established and the NCAA March Madness Tournament had just begun. It is a particularly intense time of the season, when teams are locked in final matchups that could either guarantee them a spot in the madness or leave on the outside looking in. Given the excitement of this part of the season, we were curious if there was any way to use data to get a deeper understanding of what it takes to make the coveted March Madness. As such, we decided to ask if it is possible to predict if a team will make it into the NCAA tournament, and if so what factors contributed to their advancement. Obviously, a team's win record was a significant factor, but we wanted to explore more unique variables such as shooting percentages, basic counting stats, and advanced metrics among many others.

There are many individuals, among both the basketball world and outside of it, that would find this information important. For one, this information could be critical to the coaching staff of a particular team. Coaching staff work to improve their teams statistics using various methods and knowing which statistics to focus on is imperative. They also create a scheme for their team outlining how they prefer their players play that involves variables such as pace factors and shooting preferences. Using this information, perhaps there are variables they should emphasize more to increase their team's chances of making the tournament. Additionally,

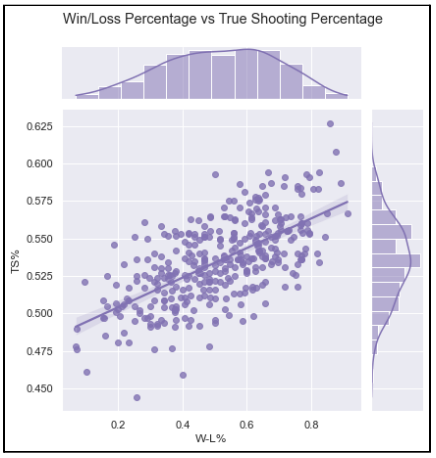
networks which display basketball games could make use of these statistics. Over the years, sports statistics are being displayed more and more frequently during games. Shot charts, which display how a player's field goal percentage varies on different zones of the court, is one specific example out of the variety of ways stats have become more integrated into the entertainment side of the game. If our findings show that there are surprising correlations between statistics and advancing to the tournament, the frequency of statistics a network may display could change. Finally, these statistics could simply be helpful for spectators and fans of the NCAA. There are so many elements of a game that it can be hard to know what to pay attention to. Understanding of the statistics could help fans and spectators make better predictions for their brackets or simply understand their team better.

Introduction to your Data

Our dataset is basketball team statistics sorted by team for all 358 NCAA Men's College Basketball teams during the 2021-22 NCAA season. We obtained the data from Sports Reference, specifically the College Basketball site. Sports Reference is a database and website that collects and stores data on various sports teams and players from box scores of individual matchups and then calculates more complicated metrics from there. The matchups for our particular data are all the games between the different Division 1 NCAA basketball teams.

The dataset contains 358 rows corresponding with the 358 different NCAA Division 1 college basketball teams and each row has 45 columns worth of data about the team. We incorporated each column of the data for our project since we used a logistic regression. The data had several different general categories including win and loss counts, shooting percentages and tracking, per game stats, and advanced metrics. The win and loss counts included the number of wins and losses for total games, home and away splits, and conference wins and losses as well as

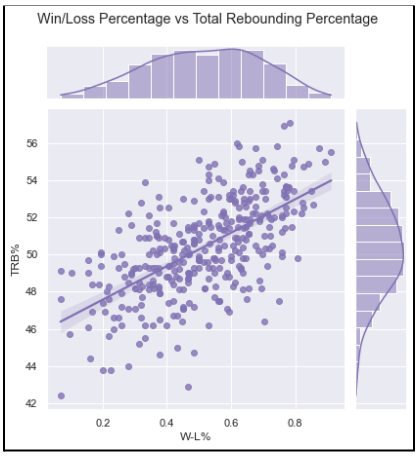
a helpful win versus loss percentage. The shooting percentages and tracking looked at the number of shots taken from the three main areas of the court — two point shots, three point shots, and free throws. The dataset kept the number of shots taken and made per game as well as the shooting percentage of the team from each area. The per game stats are the more standard basketball box score stats such as points, rebounds, assists, blocks, among others and the data values contain the average number of each stat per game for the team. Finally the advanced stats relate to more nuanced calculations such as a weighted shooting percentage (TS%), the percentage of rebounds or blocks available to the team that are taken advantage of (TRB% or BLK%), and catch all metrics to measure overall offensive efficiency (ORtg) to name a few.



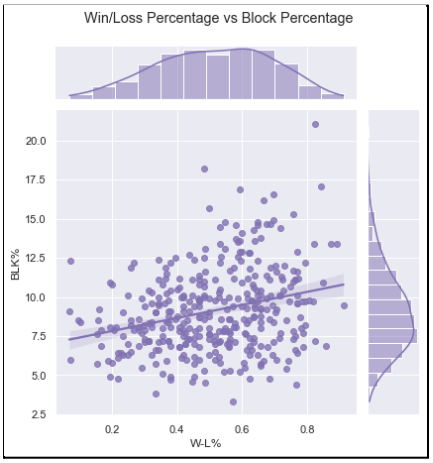
However, the most important data we had at our disposal was the teams that made the NCAA tournament. For those teams that made it, the string of their names ended in “NCAA”, and we took advantage of that to make a separate column for all teams denoting if they had reached that achievement.

Data Science Approaches

Going into the project we had completed a few initial calculations on the teams that made the tournament and came out with the understanding that teams with a high win percentage seemed much more likely to make the tournament than their counterparts. As such we decided to begin our approach with a nuanced scatterplot called a joinplot which provides information regarding the spread of the data, a



line of best fit with built in error calculated, and marginal changes along the side and top of the graph to demonstrate how the data adjusted as values increased. Plotting between all stats for this plot would have been tedious with the amount of stats we had so instead we focused on win and loss percentage and picked out a few choice plots to get an idea of how our data looked.

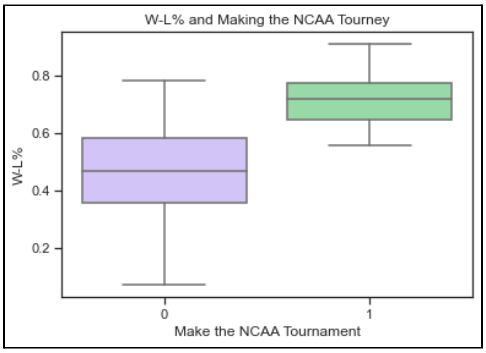


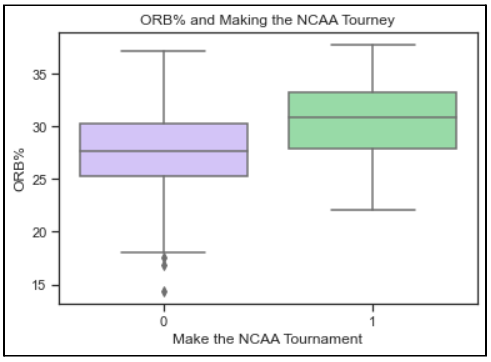
Next we wanted to understand how certain stats compared to making the tournament since that was the end goal of our project. For this purpose we used seaborn boxplots which allowed us to compare spread and center for teams that did and did not make the NCAA tournament (where the binary one represents a team that made it and a zero represents a team that did not). Since we were only comparing stats with whether a team made the tournament, we decided to plot a boxplot for every stat and see how it affected a team's chances of moving on.

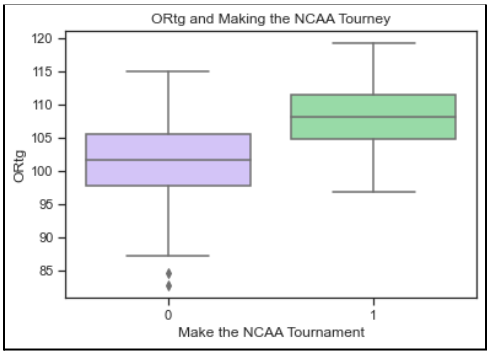
Finally, now that we had a good idea of how the data looked and how certain statistics might influence a team's chances of making the bracket, we decided to make a logistic regression. We used the regression to predict the teams that would make the tournament and we paired that with a confusion matrix to display the effectiveness of our model. To create this we split the data into a train and test split where there were 286 teams (80%) we used to acclimate our model and 72 teams (20%) we used to test it. Within both the train and test splits the data about the team was put into the prediction side of the model while whether a team made the tournament were the results. Finally, to give our model the best chance to be accurate, we then weighted the stats based on how much they affect making the tournament.

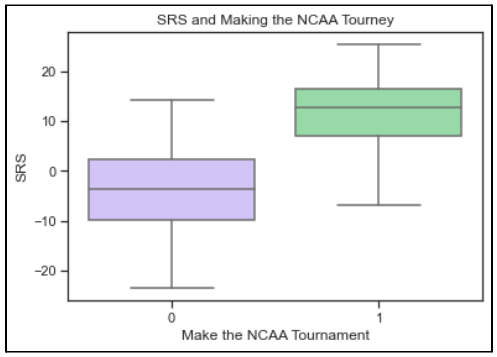
Results and Conclusions

When analyzing the results from our program, the first thing that stuck out was the distributions of the various statistics compared to whether a team made the tournament. These distributions showed us which statistics are most highly correlated to making the NCAA Tournament, as seen in the following boxplots that depict a few statistics' distributions split into the teams that made the tournament and those that didn't. The uppermost plot showcases a team's win and loss percentage distribution with the green box plot demonstrating that teams with a clear higher percentage had a better chance to make the tournament. The same applies to offensive rating (ORtg), offensive rebounding percentage (ORB%), and the simple rating system (SRS) with both having a cluster of overall higher values for the teams that did make the tournament as compared to those that did not. This combination of statistics that have clear shapes and boundaries when sorted led us to see the difference certain statistics had in a team's chances of making the tournament. This led us to balance the data for our logistic regression model to take into account these differences and balance the stats' weights based on the correlation and distribution of these statistics.

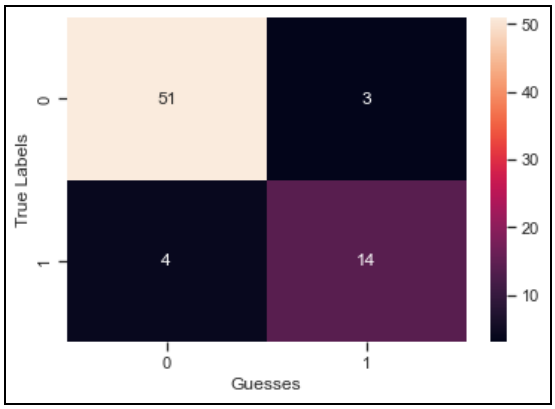








With all these statistics taken into consideration we were able to run a logistic regression binary classifier with our balanced weights as mentioned where we used 80% of the teams in the data set for the training set and the remaining 20% for the testing set. This meant we had a solid training base with 286 teams from the current season that we used to predict whether the other 72 teams from the same season made the NCAA Tournament or not. After completing its predictions, our model had a 91.6% accuracy when predicting whether the 20% testing set did or did not make the tournament, with only three false positives and four false negatives among the testing batch. In addition, our model had an average of 92% in precision and recall, along with a 91% f1 score when predicting the 72 team testing batch.



Overall, our model was very accurate in categorizing and predicting which teams made the NCAA Tournament and which ones did not using the myriad of regular and advanced statistics that represented each team. This success would enable us to eventually train this model with data from all the previous seasons and use it to predict which teams would make the NCAA Tournament before the committee actually decides on the final tournament field. This also enables us to dive deeper into the numbers and statistics ingrained in basketball and see how they influence team's successes, and in this case their chances of making the tournament.

Future Work

One aspect we would like to explore further is how grouping could impact a team's chances of making the tournament. Obviously, there are the teams such as Duke, North Carolina, Kansas, and others which more consistently make the top sixty-four, but are there any other trends to predict a team's chances aside from their stats. We would consider using Geopandas to

determine if there are regions in the U.S. from which teams often advance to the tournament. It would be fairly simple to find data containing the coordinates of the schools in the NCAA, and to approach this question from a geographical standpoint. Another grouping option is to group the teams by the conference they play for and see if certain conferences might give weight to a team's chances.

Another aspect we would like to explore are a team's stats on a year to year basis, and how they may contribute to a team's chances of making the tournament. In addition, we would be interested in comparing how a team's stats have changed over the past decade and determine how these changes influenced their inclusion or removal from the tournament. We could focus this research on teams which have a percentage of making the tournament between 20 and 80 percent. Teams which consistently never make the tournament or consistently are included could skew a theoretical project's ability to understand changes over time. The website we obtained our data from does have basketball statistics from the early 1990s to the 2021-2022 season, so it is possible to perform this analysis, however it would require many hours of data cleaning. Finally, there are many schools that have joined and left the NCAA since the 1990s, we would want to only select the teams which have played for a majority of the seasons.