Luke McMeans, Nithil Suresh, Bay Wiggins

**General Question**

- How far is each team predicted to go in the March Madness tournament?
- *Similar/sub questions*
    - Can we predict the tournament accurately?
    - Which teams are likely to make a ‘Cinderella’ run?
    - Who is most likely to win the national championship?

**What is an observation in your study?**

- With this dataset, an observation would be a team in a given year’s March Madness tournament. Each team has a set of ratings from the KenPom and Bart Torvik sites, indicating their offensive and defensive efficiencies, tempo, statistical rankings, and more. Along with this, our data also contains the farthest round that each team went that year, which is what we will be predicting for this year's tournament.

**Are you doing supervised or unsupervised learning? Classification or regression?**

- We’re dealing with supervised learning here. This dataset also contains information from each tournament since 2008, detailing how far each team went in each tournament. We can use this data to train our models. Since we’re trying to predict how far a team goes, the predicted values are categorical (even though their values are numbers). Given this, we will be working with a classification approach. We only have a discrete amount of values for their ‘round’ value, and a strict number of teams that can have each value.

**What models or algorithms do you plan to use in your analysis? How?**

- Given our desires to predict the tournament, we will approach this from multiple angles.
    - We’ll establish a baseline by doing a logistic regression. This feels like the most simple approach, and has good prevention of overfitting, if the correct regularization is applied. Along with this, logistic regression works better than most with smaller datasets. Though we have a lot of data to work with, we only have 16 tournaments overall, so this information could prove to be beneficial. It also can help determine which statistics (features) are most important in determining how far a team goes.
    - We can also use a Random Forest Classifier to help with non-linear relationships in our data. We are also able to determine the important features in determining prediction with this model. An important aspect of this model is that it can help the randomness of upsets. At the same time, this model is good at handling outliers, ensuring nothing is too crazy in our predictions. These two pieces of information in tandem ensure we can make reasonable predictions and upsets.
    - We may also find it beneficial to utilize a neural network. This type of model is very good at finding hidden trends and complex patterns in our data, which could be very helpful in our predictions given the variability in results year after year. This is another piece to help up find out what it takes to win in certain rounds, and even pull off some upsets. Hopefully this can build off the Random Forest and create reasonable ‘Cinderella’ teams and underwhelming performances.

**How will you know if your approach “works”? What does success mean?**

- Our model works if we are able to predict reasonable results for a March Madness tournament given the team’s KenPom and Bart Torvik ratings. Our dataset has the information for the ongoing tournament, with no data in the column that determines their furthest round. This is the column we will predict given the data from previous tournaments. For this to be accurate, it must follow the true structure of a single elimination tournament, so we must have one champion, one runner up, two semi-final losers, four quarter-final losers, etc. With this prediction, the results must be considered reasonable. For example, a 16-seed has never made it past the second round, so we should be extremely suspicious if one makes a deep run in our findings. As long as the general trends of a tournament, with appropriate winners and upsets, we should be in good shape.
- Along with this, by the time we will have finished our model, this year’s tournament will have been completed. We can directly compare the results of the two to see how accurate our model was. We can also hide the round values of previous tournaments to test and compare these as well.
- We can use a confusion matrix to see what rounds or years provided the most accurate predictions. This also provides a metric if the model overestimated or underestimated their predictions.
- Finally, as our classes are imbalanced, utilizing a metric such as an F-1 Score could provide more insight into performance beyond just accuracy.

**What are weaknesses that you anticipate being an issue? How will you deal with them if they come up? If your approach fails, what might you learn from this unfortunate outcome?**

- Building off of making a reasonable structure, we have to make sure there are an appropriate amount of upsets (a significantly lower seeded team beating a higher seed). Given the nature of seeding, the better teams typically get a higher seed, and thus will have better overall stats/ratings from these sources. We fear there’s a possibility that our predictions could be ‘chalk’, meaning the higher seed will win significantly more than average (if not every game). This tournament is called ‘March Madness’ for a reason. Sometimes, the worse teams will make surprises against the better ones. We intend to pull information from relevant ‘Cinderella’ teams, along with finding how often upsets happen, to implement upsets where reasonable. Our Random Forest and Neural Network models could be beneficial in minimizing this, but you never know what can happen.
- Another weakness related to the use of a neural network is that neural networks traditionally require large amounts of data to generalize well. This is a limitation that we will keep in mind while training, but we will attempt to use this approach just in case it turns out to be effective.
- Lastly, we may very well end up finding it practically impossible to figure this tournament out. The odds of predicting a bracket with perfect accuracy are one in 9.2 quintillion (assuming all games have 50/50 odds). With this, our goal is to be as accurate as possible.