ML Madness: The Difficulty of Making the Perfect Bracket

Michelangelo Pagan, Sabrina Matsui, Jared Maksoud

I. Abstract

In order to predict the success of Division I college basketball teams in the March Madness bracket system, a random forest and neural network was used. The success of teams were split into categories of which teams made it to the tournament, Round of 64 and 68 (R64 or R68), Round of 32 and Sweet 16 (R32 or S16), Elite Eight and Final Four (E8 or F4), and won the championship and were runner-ups (2nd or Champions). These machine learning prediction methods showed that the higher the bracket (the closer the teams are to winning), the model performs worse because the skill and performance of the teams become more similar and the margin decreases. We constructed a random forest model which predicted on data from the teams who made it into the competition. This model performed decently with a weighted-average f1-score of 0.695. To further, and more accurately predict the bracket, we implemented a multilayer perceptron (MLP) neural network classifier, again, without including the teams that did not make it to the tournament. This model performed well, and very similarly to the random forest, and slightly better, with a testing weighted-average f1-score of 0.70.

Work Statement: For this project, the original dataset scrubbing and exploratory analysis was done by Sabrina. The coding and analysis with the Random Forest model was done by Michelangelo, and the coding and analysis for the MLP neural network model was done by Jared. All three students collaborated for the final writeup and conclusions.

II. The Dataset

The data to produce these models was retrieved from a publicly available dataset on Kaggle named *College Basketball Dataset*, created by Andrew Sunberg in 2023. The data includes team statistics from all 353 Division I basketball teams from the years of 2013 to 2023, not including 2020 due to the March Madness tournament being canceled. With the teams that will participate in the 2024 tournament about to be revealed on March 17th, this historical data of tournament history could lend itself to making an informed prediction for this years' tournament and fuel the obsession with the creation of a "perfect bracket." The dataset includes 19 advanced statistics for each team, as well as their historical tournament success. An example of these statistics includes Adjusted Offensive and Adjusted Defensive Efficiency, which are measures of how many points a team scores or gives up per 100 possessions. Simpler statistics in the dataset include things like Games Played, Wins, and shooting percentages from various areas of the court (three-pointers, free-throws, and two-pointers). We examined each of these factors for feature selection, which will be discussed in the exploratory analysis section.

III. Literature Review

Due to the difficulty in predicting a "perfect bracket", there have been many attempts at using college basketball data to train a model for the March Madness Tournament and accurately predict the winner. In 2018, a writer from the basketball blog "reHOOPerate" trained a deep neural network model on a very similar dataframe of team statistics, achieving over 99% accuracy on their testing dataframe. Although we cannot recreate such a computationally-complex model, it is worth noting that despite the accuracy of this model, the writer predicted the 2018 champion to be UVA, which ultimately was incorrect as Villanova won the tournament. Another article written by Hunter Kempf in 2022 revealed the difficulty in predicting the men's tournament by comparison to the women's tournament, which had tournament seeding as the most significant predictor of tournament success. As a result, Kempf used a variety of hand-created features and power rankings but still could not overcome the barrier of predicting upsets, placing his bracket in the top 16% of all brackets for that year. With our analysis, we aim to explore how prediction models change as the rounds in the tournament get closer to the final, and what features are most important in predicting the success of a team.

IV. Exploratory Data Analysis

After splitting the data into teams who made the Sweet 16 and teams who did not, we created a scatter plot of the offensive (x-axis) versus defensive (y-axis) efficiency metrics and found that there were negative linear relationships every year and that teams who made Sweet 16 and champions had a high ratio of ADJOE to ADJDE (Fig. 1). As such, we created a new variable called "Win Efficiency" based

on the results of the scatter plot of ADJOE/ADJDE and dropped those individual columns in the dataset. (In context this would make sense because a high offense rate and a low defense rate correspond to a better performing team).

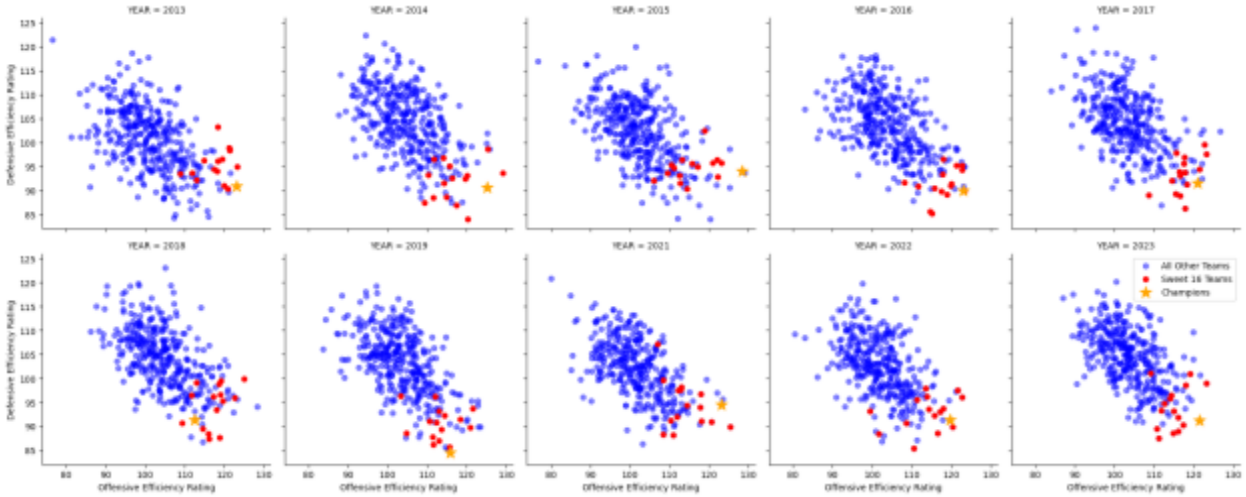


Figure 1 Scatter plot of adjusted offensive efficiency (x-axis) versus adjusted defensive efficiency (y-axis) by year. Sweet 16 teams and championship teams highlighted separately.

Following these scatterplots, many alterations were made to the data before fitting our prediction models. First, a correlation heatmap was created to examine potential relationships between our features (Figure 2). It can be seen that certain metrics such as "BARTHAG" (a measure of wins against an average team) and "WAB" (a measure of wins above the cutoff for making the tournament) have the potential for multicollinearity due to their high correlation with each other. As such, we removed one of two features that had a high correlation with each other, above r = 0.85. In addition, all numerical columns that were used in the dataframe were standardized since they were measured on different scales, and tournament results were separated on a 1-4 scale: 1 representing a team reaching the Round of 68 or 64, 2 representing a team reaching the Round of 32 or 16, 3 representing a team reaching the elite 8 or final four, and 4 representing the top two teams in the tournament.

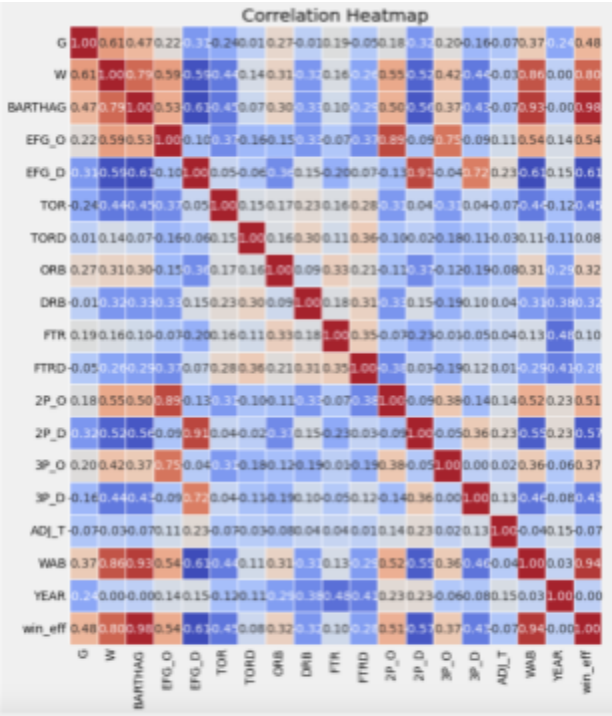


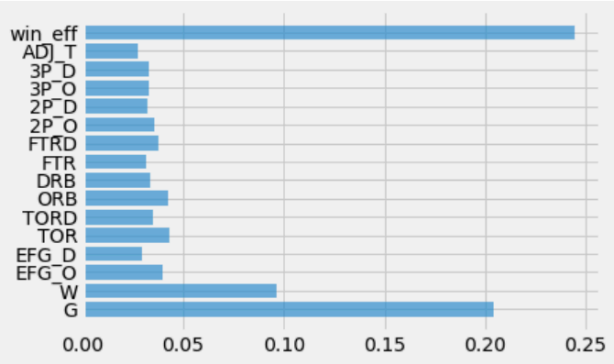
Figure 2 Correlation heatmap of the dataset to avoid multicollinearity.

V. Prediction Models

To predict which "category" a team will fall under – or how far they will make it into the tournament – we applied two models to predict team success. Although we initially used the dataframe including all 353 teams to train the dataset, we found that accuracy scores were inflated due to a large majority of the teams not making the tournament, which were generally easier to predict than the specific tournament results. Initially, the teams that did not make the tournament at all were grouped as 0 where the NaN values were replaced with 0 for the post-season ranking. Therefore, our newer models were trained specifically using an 80-20 split with randomly selected data from teams that specifically made the tournament.

A. Random Forest Model

Before moving on to the neural network model, we first examined the prediction power of a random forest prediction model. This was done in order to simplify the model before moving on to further analysis, as random forests reduce the potential for overfitting given a dataframe. Using k-fold cross validation and RandomizedSearchCV to fit multiple models and compare f1-scores, it was found that the best parameters for the random forest model was a model with a max depth of 6 layers and 331 estimators.



F1 score for the *training* dataset = 0.893
F1 score for the *testing* dataset = 0.695

	precision	recall	f1-score	support
1.0	0.78	0.82	0.80	71
2.0	0.62	0.70	0.66	47
3.0	0.67	0.27	0.38	15
4.0	0.33	0.33	0.33	3
accuracy			0.71	136
macro avg	0.60	0.53	0.54	136
weighted avg	0.71	0.71	0.70	136

Figure 3 A bar plot of feature importance (left) for the forest prediction model, as well as evaluation metrics including f1-score for each category of tournament success.

Based on the Random Forest Model, it can be seen that our feature of "win efficiency" seems to be a very strong predictor in the model, and helps the model predict tournament success overall with an f1-score of 0.695, very close to a "good" model (generally 0.7 or higher). However, it can also be seen that as the prediction categories get smaller, the prediction power of the model significantly decreases. With the last two groups (elite eight and beyond) having f1 scores both below 0.4.

B. MLP (Multi-Layer Perceptron Neural Network) Classifier Mode

A MLP Neural Network model was used because it allows for a more complex model using back propagation and can be more accurate because it contains multiple layers of nodes, hence more updated weights. The testing accuracy here is 0.71 which is close to a good model and shows to be higher than the accuracy of the random forest model. Similar to the random forest model, the precision and f1-scores of the groups as the number of teams get smaller and as it predicts into higher brackets becomes lower. The groups making it into the tournament showed an f1-score of 0.81, while Sweet 16 and below was 0.60, and the Elite 8 and Final Four was 0.55. The confusion matrix also shows that as the groups go higher, the accuracy of predicted success decreases.

```
Training Accuracy: 0.97
Testing Accuracy: 0.71
              precision    recall  f1-score   support

         1.0       0.84      0.78      0.81        79
         2.0       0.53      0.68      0.60        41
         3.0       0.67      0.46      0.55        13
         4.0       0.00      0.00      0.00         3

    accuracy                           0.71       136
   macro avg       0.51      0.48      0.49       136
weighted avg       0.71      0.71      0.70       136
```

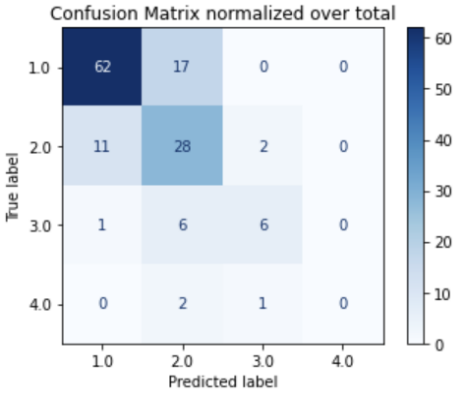


Figure 4: Evaluation metrics including f1-score for each category of tournament success (left), confusion matrix of results (right).

VI. Model Evaluation

The prediction models that only included teams that made it to the March Madness tournament proved to have a higher testing accuracy and f1-scores compared to the models that included all college Division I teams. Overall, each model showed that f1-scores were lower as the brackets went up. This can be explained by the fact that as teams move up, the closer the performances of each team are and the harder it is to statistically differentiate between their performances. (The model however performed with less accuracy when the training data included all years before 2023 and the testing data was 2023. To improve the model to be able to create more accurate predictions for specific years, we would need to include factors like funding for the team and other situational factors.) The random forest model and the neural network models showed similar overall accuracies, however the random forest model was able to have a 0.33 f1-score for the winner and runner-up whereas the neural network had a f1-score of 0. This could be due to adding too much of a penalty or weight in the last bracket because there are only two teams and because the team skills are much more similar. Further, we predicted that Gonzaga would win in the missed out 2020 (Covid) year which ESPN predicted as their third choice.

VII. Works Cited

Sunberg, Andrew. (2023). *College Basketball Dataset*, Version 5. https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset/data.

Kempf, H. (2022, April 6). Predicting the 2022 NCAA Basketball Tournament Using Data Science. *Medium*. https://medium.com/@HunterKempf/predicting-the-2022-ncaa-basketball-tournament-using-data-science-3400a2c84098

reHOOPerate. (2020, February 22). Training a Neural Network to fill out my March Madness Bracket. *Medium*. https://medium.com/re-hoop-per-rate/training-a-neural-network-to-fill-out-my-march-madness-bracket-2e5ee562eab1

Walder, Seth. (nd) "BPI's NCAA Tournament Bracket Simulation Will Surprise You." *ESPN*, ESPN Internet Ventures, www.espn.com/mens-college-basketball/story/_/id/28915817/bpi-projects-how-2020-ncaa-tournament-played-out