Marching Towards Madness:
Developing a ML Model for NCAA Tournament Predictions

Joshua Kraus

Abstract— The NCAA basketball tournament is a single-elimination, 64-team tournament where teams compete to win the national championship. Known as "March Madness" it has become popular to try to pick which team will advance to each round in the tournament, which is known to be a near impossible task. In recent years, utilizing analytical techniques has become popular for use in filing out brackets, often manifesting in predictive models provided by third-party sources. FiveThirtyEight, a website known for their statistical analysis of national news, is an example of this as each year they provide well-regarded NCAA tournament predictions using a proprietary predictive model. This begs the question, however, is there a strategy for filling out brackets that consistently performs better than current approaches? The purpose of this study was to develop a superior strategy for filling out brackets using machine learning classification, and then compare its performance to another high-performing predictive model. A bracket selection strategy was developed by identifying which team has the highest probability of winning the entire tournament and recursively filling out a bracket. An ensemble of classification models was developed, and results were tested using cross-validation. The results of this study demonstrated superior performance of the strategy, and suggested which predictor variables were impactful to its success.

Introduction

The NCAA basketball tournament is a single-elimination, 64-team tournament where teams compete to win the national championship (Wilco, 2023). Once the teams are selected for the tournament, each one is given a seed one through 16 and placed into one of four regions (Wilco, 2023). In this seeding, a one seed is considered the best team in the region while a sixteen seed is considered the worst (Leonhart, 2015). Additionally, each of the four one seeds are ranked so that the best overall team competes in the top left region against the worst one seed in the bottom left region (Leonhart, 2015). Colloquially known as "March Madness" it has become popular to try to pick which team will advance to each round in the tournament to beat other users' brackets (Wilco, 2023). Users fill out brackets on websites such as CBS or ESPN where the point totals for their brackets are calculated in real-time and ranked against other users' brackets (Wilco, 2023). Therefore, the goal of filling out a bracket becomes to make predictions of which teams will advance to each round to gain the most points possible and beat other users.

The NCAA tournament has become known as "March Madness" due to unexpected outcomes that often occur during the tournament (Wilco, 2023). There is a roughly 1 in 9.2 quintillion chance of selecting a perfect bracket by randomly deciding each matchup, or roughly a 1 in 120 billion chance using some basic strategies to do so (Wilco, 2023). These nearly impossible odds and seemingly random outcomes that occur during the tournament are part of what makes filling out NCAA tournament brackets exciting for users. Not only is there a competitive aspect from beating other users, but excitement also comes from attempting to make correct picks that are difficult to predict. This has led to the development of strategies to attempt to correctly make bracket picks in any given year to strive for perfection (Mass, 2023; Ota, 2023; Boice & Silver, 2018; Dutta et al., 2017; Niemi et al., 2008). While devising a strategy to create a perfect bracket may be a futile task, creating one which accurately models the seeming randomness of this tournament could give users an edge.

Recently, utilizing analytics has become popular for use in filing out brackets with predictive models being provided by third-party sources. FiveThirtyEight, a website known for statistical analyses, provides well-regarded NCAA tournament predictions using a proprietary predictive model (Boice & Silver, 2018). They provide probabilistic forecasts for each team to advance to each round using a number of externally calculated power ratings, which is complex for FiveThirtyEight to model and predict but is simple for users to their predictions and make bracket selections (FiveThirtyEight, 2024). Given that this approach utilizes advanced analytics, it is possible that it could be effective for predicting the behavior of teams in the tournament.

Problem Statement

While FiveThirtyEight's method is rigorous and may give users an edge when filling out brackets it contains one major flaw, which is that correct picks in later rounds of the tournament are worth more points than correct picks in earlier rounds. Using ESPN's scoring system, the points awarded for a single correct pick in each round of the tournament can be seen in Table 1.

Round	Points
Round of 32	10
Sweet 16	20
Elite 8	40
Final 4	80
Championship	160
Winner	320

Table 1 - Point Values per Round (ESPN Scoring Rules)

From this, it is clear that picking the correct winner of the tournament is worth 32 times more points than a correct pick in the Round of 32. This suggests that when filling out bracket, starting by picking the winner of the tournament and recursively filling out a bracket could be an approach that maximizes the points gained. This is where FiveThirtyEight's method is flawed: users could employ the recursive selection approach previously mentioned, but the predictions provided utilize a forwards selection approach to compute probabilities. That is, FiveThirtyEight's method considers each possible matchup in the order of rounds and determines the probability each team will win that game, not which team is most likely to advance to each round regardless of matchups. The key difference is that a forwards selection approach assumes that team success is comprised of intrinsic qualities as well as external factors from the teams they are competing against, but places equal weight upon correctly determining the outcome of each game. In contrast, a recursive selection approach assumes that team success is more dependent on the intrinsic qualities of a team, and that determining the outcome of the most valuable games should be prioritized. In this manner, the traditional method promotes overall accuracy while a recursive selection approach prioritizes total points gained, which is the true metric used to determine a bracket's success. This provides the need for a predictive model utilizing recursive bracket selection, which can then be compared to traditional methods to determine the superior approach based on points gained.

Data Sources

Two data sets are required for this study to collect predictor and response variables for classification models. Each data set ranged from the 2013-2023 tournaments, as 2013 was as far as historical data was feasibly available.

The first data set used was KenPom's team ratings data set, which was collected using an internally developed web scraper (KenPom, 2024). Ken Pomeroy (KenPom) Ratings, developed by college basketball statistician Ken Pomeroy, are team power ratings which are considered to be highly accurate in predicting game outcomes (Dutta et al., 2017). If KenPom ratings are truly more accurate than other rating systems and a reasonable probabilistic model can be fit using this data, then this data set could provide the benefit of harnessing advanced analytics to make possible accurate predictions. These ratings consist of metrics quantifying the performance of a team, which have been summarized in Table 2 (KenPom, 2012).

Metric	Description
Adjusted Offensive Efficiency	Points scored per 100 possessions a team would have against the average D-I defense.
Adjust Defensive Efficiency	Points allowed per 100 possessions a team would have against the average D-I offense.
Adjusted Tempo	Possessions per 40 minutes a team would have against the team that wants to play at an average D-I tempo.
Luck	A measure of the deviation between a team's actual winning percentage and what one would expect from its game-by-game efficiencies.
Adj. Offensive Efficiency of Opposing Offenses	Points scored per 100 possessions opposing teams would have against the average D-I defense.
Adj. Defensive Efficiency of Opposing Defenses	Points allowed per 100 possessions opposing teams would have against the average D-I offense.

Table 2 - KenPom Rating Definitions

These variables all served as possible predictors and are anticipated to be highly predictive given their characterization of team performance.

The second data set required was historical tournament records, which were collected using an internally developed web scraper using the website SportsReference (SportsReference, 2024). This data set consisted historical tournament results which were used to determine the performance of models, as well as the response variable for this study and other possible predictors. Definitions of each variable collected can be seen in Table 3.

Variable	Description
Year	The year a team competed in the tournament.
Seed	Which seed they were in the tournament. Seeds range from 1 to 16, with 1 seeds being considered the best teams in the tournament.
Name	Name of the team.
Conference	Which conference the team is a member of.
Region	Which region of the tournament they were in, given that there are 4 quadrants to the bracket.
Wins	How many games a team won that season.
Conference Championship	Whether a team won their conference championship, or not.
Round	Which round in the tournament a team made it to.

Table 3 – SportReference Rating Variables

A team's round was used to construct the response variable for each model, and all other variables would serve as possible predictors.

Methodology

Research Questions

The purpose of this study was to determine a strategy for filling out brackets that consistently performs well. A strategy was developed, its performance on historical data was tested using cross-validation, and the approach was compared to FiveThirtyEight's predictive model. Specific research questions included:

1. How did the strategy perform? On average, how accurate was it in every round, and how many points did it score overall?
2. Which predictors were chosen for each classification model?
3. How does this approach perform again another well-known predictive model?

Analytical Approach

The strategy that was developed utilized non-intuitive assumptions about bracket selection strategies that involve predicting which team is most likely to win the last round, filling this team in throughout the bracket, and recursively filling out the bracket through each round. This involved fitting six binary classification models to determine whether a team reached a given round (1) or not (0) for all 6 rounds in the tournament.

Four individual classification models were fit for each round: a logistic regression, random forest, gradient boosting, and multi-layer perceptron model. Each of these models were chosen based on their differing assumptions, hopefully promoting various trends in the data being captured. Logistic regression is a linear model that estimates

the probability of a binary outcome based on a set of predictor variables and is the simplest model utilized in the study, but still an effective one when decision boundaries are fairly linearly separable. Random forest models utilize an ensemble learning method that constructs multiple decision trees during training and utilizes majority voting from individual trees to predict the class of a data point. Random forest models provide the benefit of being less prone to being overfit due to the ensemble learning method used. Similarly, gradient boosting models also utilize an ensemble learning approach that build models sequentially by fitting new models to the residual errors of the preceding model, aiming to minimize a loss function. Gradient boosting models provide the benefit of potentially robust, high predictive accuracy through its use of multiple learners and sequentially improving upon its past weaknesses. Lastly, Multi-layer perceptron (MLP) models are a type of artificial neural network with multiple layers of nodes or neurons including an input layer, one or more hidden layers, and an output layer providing probabilistic estimates of the membership of a data point to a certain class. MLP models provide the benefit of being able to capture complex, non-linear relationships in data which could potentially provide high predictive accuracy. For each model hyperparameters were tuned which can be seen in Table 4.

Model	Hyperparameters
Logistic Regression	**penalty**: norm of regularization penalty **C:** regularization strength **solver**: optimization algorithm
Random Forest	**criterion**: function to measure the quality of a split **max_depth**: maximum depth of a tree **min_samples_split:** # of samples required to split a node **min_samples_leaf:** # of samples required to be a leaf node
Gradient Boosting	**max_depth**: maximum depth of a tree **min_samples_split**: # of samples required to split a node **min_samples_leaf:** # of samples required to be a leaf node
Multi-Layer Perceptron	**hidden_layer_size**: # of neurons, # of hidden layers **activation**: activation function of hidden layer **alpha**: regularization term (L2 regularization) **learning_rate:** rate of learning for weight updates **solver**: weight optimization algorithm

Table 4 - Hyperparameter Definitions

After training each individual model, an ensemble of learners was created for each round of the tournament, totaling 6 ensemble models comprised of 24 sub-models. Each learner was weighted in the ensemble model based on its cross-validated performance during hyperparameter tuning, to more strongly weight individual learners which performed better. The result would be an ensemble method combined from each of the learners into a "soft" voting classifier which provide probabilistic outputs, rather than a binary classification. This was required to employ the bracket selection method outlined previously, which selected the team with the highest probability to advance to each

round utilizing the recursive selection method. Utilizing an ensemble method for each round should lead to an improved, robust predictive accuracy through the ensemble capturing various tends in the data from the multiple individual trained models used.

When considering binary classification for each round of the tournament, data was highly imbalanced for most rounds due to the fact that there are 64 teams in the tournament, yet only 1 team wins, 2 make the national championship, and so on. Data imbalances can be challenging for classification models as many machine learning models ignore the minority class therefore resulting in poor performance on it. Given that the minority class were teams who advanced to a certain round, they were the primary focus of the models being fit and resampling techniques would need to be employed to address data imbalances. To do so, two separate resampling techniques were combined. First, Tomek Links was used to under-sample the majority class. This is a modification of a Nearest Neighbors model and involves finding data points of the majority class which are closest to the minority class and then removing them to emphasize the decision boundary between classes (Tomek, 1976). Next, Synthetic Minority Oversampling Technique (SMOTE) was utilized to over-sample the minority class. SMOTE, a common method to solve class imbalances, selects data points that are nearby in the feature space and are close to the decision boundary and then creates new, synthetic samples of the minority class by connecting data points with these new samples (Shawla et al., 2002). When SMOTE and Tomek Links are utilized together the combination of under-sampling the majority class and over-sampling the minority class can improve classification performance better than if only one method was used to solve class imbalances (Batista et al, 2004). This process was used when tuning hyperparameters for each learner, as well as when fitting the ensemble model.

To assess the performance of each model, leave-one-out cross-validation (LOOCV) was used to tune hyperparameters for individual models and examine the overall performance on the strategy. This assumed that each year's tournament is considered an "observation" and by using LOOCV reliable and unbiased results for any given tournament were promoted. With this understanding, this method can also be considered 10-fold cross validation given that there are 10 years of data being utilized. To assess each individual model, precision for the minority class was used to evaluate classification performance. This metric was chosen as opposed to overall accuracy, to quantify each models' ability to correctly identify team's that will reach a given round without misclassifying too many teams who will not advance. This is given by the following formula:

$$\frac{(\hat{y} = 1,\ y = 1)}{(\hat{y} = 1,\ y = 1) + (\hat{y} = 1,\ y = 0)} \text{ or } \frac{True\ Positives}{True\ Positives + Flase\ Positives}$$

Given that correctly predicting the minority class of imbalanced data is the focus of these models, utilizing precision for the minority class promoted model performance for the notable data points. Additionally, to assess the performance of the overall strategy the total points gained from each bracket was calculated to provide an estimate of the average bracket points with a standard deviation measure. Bracket scoring was done using the ESPN bracket scoring rules described previously, and in this manner the total possible points for a bracket was 1920. These results provided support for how many points can be expected from a bracket when utilizing this strategy which was compared to FiveThirtyEight's model utilizing cross-validation.

In addition to tuning hyperparameters in each learner, the predictors chosen for each by-round ensemble model were tuned. This was done to account for differences in data trends across rounds, as it was seen that the effects of certain variables may not be consistent for each binary classification task. To account for this, backwards feature elimination was utilized for each model. In this manner, all variables were included in the initial models, hyperparameters were tuned, then features were removed based on the largest improvement to the precision for the minority class in each classification model until no further improvements were made. By utilizing this method, model performance was promoted through reducing overfitting to random effects in the data and retaining a simpler model.

Results

Exploratory Data Analysis

To explore the effects of possible predictors for each by-round model, correlation matrices were created for numerical predictors, which can be seen in Figure 1 and 2.

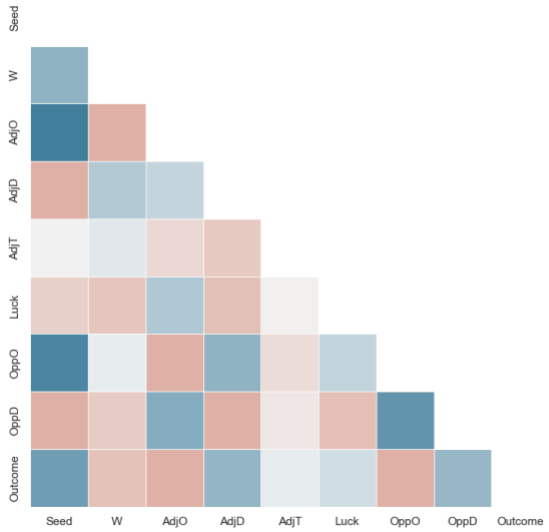

Figure 1 - Correlation Matrix for Round of 32

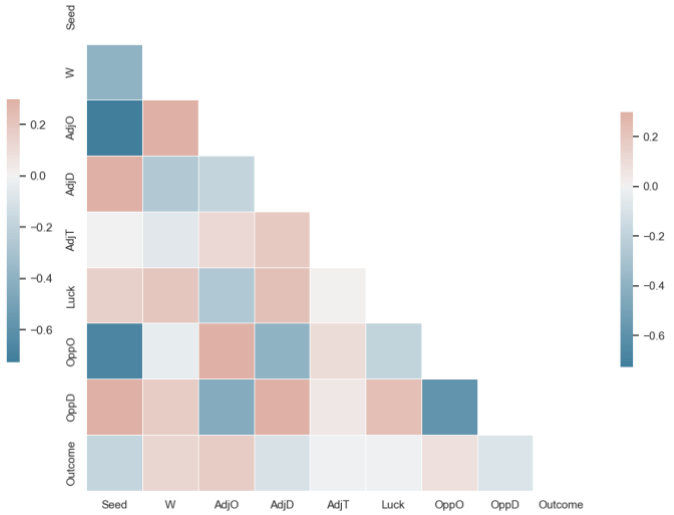

Figure 1 - Correlation Matrix for Winner

From this, it can be seen that seed, number of wins, offensive efficiency, and defensive efficiency appeared to have the strong correlations with the response variable in each data set. This suggested that these variables may be more commonly chosen in the models, due to their perceived importance. Additionally, to determine if differences in the possible effects of predictor variables exist across data sets, the correlations were compared for the first and last round of the tournament. This demonstrated that the correlations for the best predictors were stronger in the Round of 32 than for the Winner, suggesting that splitting models by-round may have some benefit to being able to accurately classify the response in each respective round. Furthermore, a number of predictors appeared strongly correlated with each other such as offensive and defensive efficiency for opposing offenses, as well as both of these variables with a team's seed. This suggested that multiples of these correlated variables may not be chosen together in a model, given that issues of multicollinearity would case and that including multiple correlated variables in a model may not provide additional benefit.

Categorical variables were also examined in terms of their strength with the response, but due to the large number of categories within them visualizing their effects was complex. There were 32 different conferences present in the data set, suggesting that there was a good chance some subset of conferences may possess differences in the response variable for each round. Whether a team won their conference tournament only contained two categories, yes or no, but there did not appear to be much difference in these categories based on the response variable. This suggested that conference may have some predictive power in the models, but that whether a team won their conference championship or not may not.

Model Performance

Before fitting an ensemble model for each round of the tournament, individual learners needed to be trained for each round. Given that logistic regression, random forest, gradient boosting, and multi-layer perceptron models were trained for each round, this consisted of 24 sub-models that were tuned. A summary of the tuned hyperparameters for each model in each round can be seen in Table 5 and 6.

Model	Round of 32	Sweet 16	Elite 8
Logistic Regression	**penalty**: L1 **C:** 10 **solver**: liblinear	**penalty**: L2 **C:** 10 **solver**: liblinear	**penalty**: L1 **C:** 10 **solver**: liblinear
Random Forest	**criterion**: log_loss **depth**: 10 **samples_split:** 2 **samples_leaf:** 2	**criterion**: gini **depth**: 10 **samples_split:** 2 **samples_leaf:** 2	**criterion**: log_loss **depth**: 10 **samples_split:** 5 **samples_leaf:** 2
Gradient Boosting	**depth**: 1 **samples_split**: 5 **samples_leaf:** 2	**depth**: 10 **samples_split**: 5 **samples_leaf:** 5	**depth**: 5 **samples_split**: 10 **samples_leaf:** 5

Mutli-Layer Perceptron	**hidden_layer**: (100,100) **activation**: relu **alpha**: 0.0001 **learning_rate:** invscaling **solver**: adam	**hidden_layer**: (100,100) **activation**: tanh **alpha**: 0.1 **learning_rate:** invscaling **solver**: adam	**hidden_layer**: (100,100) **activation**: relu **alpha**: 0.1 **learning_rate:** constant **solver**: adam

Table 5 - Tuned Hyperparameters (First Three Rounds)

Model	Final 4	Championship	Winner
Logistic Regression	**penalty**: L2 **C:** 0.1 **solver**: newton-cholesky	**penalty**: L1 **C:** 10 **solver**: liblinear	**penalty**: L1 **C:** 10 **solver**: saga
Random Forest	**criterion**: entropy **depth**: 10 **samples_split:** 10 **samples_leaf:** 2	**criterion**: log_loss **depth**: 10 **samples_split:** 2 **samples_leaf:** 2	**criterion**: gini **depth**: 5 **samples_split:** 5 **samples_leaf:** 2
Gradient Boosting	**depth**: 10 **samples_split**: 2 **samples_leaf:** 2	**depth**: 1 **samples_split**: 10 **samples_leaf:** 10	**depth**: 1 **samples_split**: 5 **samples_leaf:** 10
Multi-Layer Perceptron	**hidden_layer**: (100,100) **activation**: tanh **alpha**: 0.001 **learning_rate:** adaptive **solver**: adam	**hidden_layer**: (100,) **activation**: tanh **alpha**: 0.01 **learning_rate:** adaptive **solver**: adam	**hidden_layer**: (100,) **activation**: logistic **alpha**: 0.001 **learning_rate:** invscaling **solver**: adam

Table 6 - Tuned Hyperparameters (Last Three Rounds)

From this, many similarities can be seen in the chosen hyperparameters for each model across each round, though notably difference do exist across rounds demonstrating the need to build and tune separate models for each round. Next, the cross-validated precision for the minority class for each individual model was examined.

Model	Round 32	Sweet 16	Elite 8	Final 4	Championship	Winner
Logistic Regression	73%	55%	36%	21%	29%	15%
Random Forest	71%	54%	38%	32%	31%	23%
Gradient Boosting	71%	57%	35%	29%	40%	22%
MLP	74%	59%	44%	21%	31%	32%

Table 7 - Model Performance By-Round using Cross Validated Precision for the Minority Class

This demonstrated that performance tended to decrease as rounds advanced in the tournament. This is to be expected, however, as increased class imbalance with each subsequent round could create difficulties in each model identifying unique members of the minority class. An exception to this is the championship round and the winner, , where accuracy increased from subsequent rounds. This result was promising given that the goal of this strategy was to prioritize correct picks in the later, more valuable rounds. Additionally, this suggested that teams who reach the championship game and win the tournament may be uniquely characterized better than teams in earlier rounds, leading to improved classification performance. Furthermore, the MLP model tended to perform better than any of the other classification models tested, with the exception of the Final 4 and Championship rounds. This suggested that the neural network's ability to capture complex, non-linear trends in the data may have been advantageous and

resulted in increased performance compared to the other models. It should be noted, however, that all the models perform adequately relative to each other and would all be beneficial in an ensemble method. After combining each trained learner and weighting them to their respective performance in each round, the by-round performance of each ensemble model was examined.

Model	Round 32	Sweet 16	Elite 8	Final 4	Championship	Winner
Ensemble	72%	55%	38%	25%	37%	45%

Table 8 - Ensemble Performance By-Round using Cross Validated Precision for the Minority Class

The performance of the ensemble models varied by round with the Round of 32, Sweet 16, and Final 4 not seeing much improvement from the individual learners while the Elite 8, Championship, and Winner saw more improvement. Notably, the precision of choosing the Winner increased by 13% from the best performing individual learner, suggesting a significant benefit from the ensemble method. While all rounds may have not seen notably improvement, it can be assumed that all predictions would be more robust as a result of utilizing the ensemble method.

Next, the predictions of all six models were combined into the recursive selection bracket strategy. This involved making predictions for the Winner, selecting the team with the highest probability throughout the bracket, and working recursively through the rounds of the tournament. This approach would yield the benefit of maximizing the points gained from later rounds but would also allowing the model to be more precise. When utilizing binary classification in previous experiments it was possible for the model to predict more members of the minority class than possible in a specific round of the tournament. When utilizing probabilistic predictions, however, the correct number of members of the minority class can be selected in each round and therefore may lead to more precise modeling. The by-round performance of the recursive selection strategy can be seen in Table 9.

Round 32	Sweet 16	Elite 8	Final 4	Championship	Winner
72%	67%	49%	40%	55%	60%

Table 9 – Recursive Selection Performance using Cross Validated Precision for the Minority Class

From this, it can be seen that the performance of the recursive selection strategy was significantly better once utilizing probabilistic predictions to limit the number of members that could be selected in the minority class. Furthermore, the same trend in performance as seen previously was demonstrated, with precision decreasing as rounds increased until the Championship and Winner which demonstrated improved performance. When scoring the brackets created using this method, the point totals by year can be seen in Table 9.

2013	2014	2015	2016	2017	2018	2019	2021	2022	2023
1280	470	1000	1080	1030	1100	1270	1430	1190	1100

Table 10 - Bracket Performance by Year

This method averaged 1096 points with a standard deviation of 243, as calculated from the cross-validated point totals seen above. Notably, the brackets created from this method would have scored in the 99th percentile of brackets submitted to ESPN from 2019 to 2023, which are the only years historical brackets are still available. Overall, this appears to provide strong evidence that this method performs well for maximizing points gained.

Predictor Variables

After determining the performance of the overall method, examining the predictors utilized in each model provided evidence of the main drivers of team success in each round. Predictors were chosen for each by-round model using backwards feature elimination, where predictors were removed based on the largest improvement to the precision for the minority class in each round until no further improvements were made. The predictors selected for the ensemble model in each round can be seen in Table 11.

Predictors	Round 32	Sweet 16	Elite 8	Final 4	Championship	Winner
Offensive Efficiency	■	■	■	■	■	■
Defensive Efficiency	■	■	■	■	■	■
Tempo						■
Opp. Team Off. Eff.		■			■	
Opp. Team Def. Eff.	■	■				■
# of Wins	■	■	■	■	■	■
Seed	■	■	■	■	■	
Conference	■	■	■		■	■
Conference Tourney		■				

Table 11 - Model Predictors Selected in Each Round

From this, it can be seen that offensive efficiency, defensive efficiency, number of wins, seed and conference were most commonly chosen. This aligned with previous findings, as it was determined that seed, number of wins, offensive efficiency, and defensive efficiency all had the strongest correlation with the response variable. Number of wins

and offensive efficiency had positive correlations suggesting that as the number of wins a team has in a season and the number of points they score per 100 possessions increases, so does their chance to advance to a given round. Conversely, seed and defensive efficiency had negative correlations suggesting that as a team's seed increases and the number of point they allow per 100 possessions increases, their chance to advance to a given round decreases. A team's conference was also commonly chosen and demonstrated that the large number of categories possible in this variable did provide some significant difference in the response variable for most models. Tempo, offensive and defensive efficiency of the opposing team, and whether a team won their conference tourney were not commonly selected, which aligns with previous findings as these variables that the weakest correlations with the response. Notably, offensive and defensive efficiency of the opposing team were mostly not both utilized for a specific model, which supported the finding that these variables were strongly correlated with each other. Furthermore, whether a team won their conference tournament or not was shown to be not a good metric of team success for most models, which confirmed previous findings. Overall, these findings provide evidence for what the main drivers of team performance are and their effects on team success.

Comparing Methods

Thus far, there appears to be evidence that the bracket selection method developed performs well on a sufficiently large data set, but how does it compare to other predictive models? To quantify the success of this method it was compared to the performance of FiveThirtyEight's predictive model, which only dates back to the 2016 tournament. If the approach suggested in this study performs better, this could provide evidence that it is superior to readily available predictive models for bracket selection. The results of each model can be seen in Table 12.

Model	2016	2017	2018	2019	2021	2022	2023
Study	1090	960	1060	1270	1420	1180	1100
538	1010	680	1140	950	820	790	510

Table 12 - Bracket Performance of Study Method vs. FiveThirtyEight

Notably, the method utilized in this study averaged 1153 points over the 7-year period tested, with a standard deviation of 152 points. In comparison, FiveThirtyEight's method averages 843 points with a standard deviation of 211 points. This demonstrated how developed method outperformed FiveThirtyEight in 6 out 7 years while scoring over 300 points more on average with a smaller standard deviation. When considering the performance of both methods, this provided overwhelming support that the method utilized in this study was superior to that of FiveThirtyEight.

Conclusions

Discussion

The presented results supported the understanding that the recursive method for building NCAA tournament brackets provided superior performance to other readily available predictive models. Notably, by developing a unique approach for selecting teams that prioritized correct picks in the most valuable rounds, this strategy was successful at maximizing the points gained for a given tournament. This recursive selection method was unintuitive compared to how many users fill out brackets, but through the theoretical justification and support from quantifiable results it was proven to be quite powerful for filling out brackets. This was most clearly quantified by model performance in the Championship and Winner rounds, which saw significant upticks in accuracy compared to the trend of decreasing performance in previous rounds. This allowed for the method to promote accuracy of the most valuable picks, which in return led to large gains when scoring the brackets.

Throughout the analytical implementation of the by-round classification models needed for this method, the careful consideration of model performance at each step led to an overall product which was accurate for the given task. Initially, by determining to build separate models for each round based on the finding that predictor variables may behave different for different rounds of the tournament, unique trends in each data set were able to be captured. Additionally, by addressing class imbalances through under-sampling and over-sampling techniques, model performance in each round was enhanced by the ability to focus on notable data trends closest to the decision boundary. Furthermore, by tuning the hyperparameters and predictors utilized in each model, high predictive accuracy as well as generalization to new data was ensured. These trained models were then combined into ensemble models for each round, which provided a significant increase in performance for most rounds. Through being able to capture various trends in data by combining multiple models, this ensemble approach led to models which were even more accurate and robust than their individual counterparts. Overall, this led to a modelling approach which was highly tuned to the individual, unique effects present in the NCAA tournament data while also developing robust models capable of generalizing well to unseen data.

The combination of thoughtful strategy development and careful modeling led to an approach which was highly successful on numerous years of data, and this performance was accurately quantified through the cross-validation procedure developed to assess overall performance. By implementing LOOCV, or 10-fold CV in this case, unbiased results were presented to give an accurate representation of model performance. This was beneficial when evaluating model performance for tuning hyperparameters and

selecting predictors, which allowed for the best subset of each to be chosen while not being overfit to any certain time period of data. Moreover, this provided an accurate representation of how this strategy may perform for future years of data, and how it compared to similar methods. In total, this cross-validation procedure allowed for reliable results to be utilized when both developing models and reporting their performance.

Through both exploratory data analysis of the effects of predictors in each model and by implementing backwards selection to retain only the most salient features, findings about which factors mainly drive team success in each round were determined. Notably, the number of wins a team has, their offensive, and defensive efficiency were predictors chosen for models in all rounds. When considering the correlation matrices examined previously, this provided support for what these variables may explain about team success. Namely, teams who win more games are more likely to advance in each round of the tournament. This may seem like an obvious finding, but nonetheless it can be useful to consider when determining which teams to select in a field with countless options. Similarly, teams who score more points and allow less points per 100 possessions are more likely to advance in each round. This is again a straightforward result, but it also proved that the use of KenPom Ratings was indeed successful for this approach. These ratings are often highly regarded for their ability to quantify team success, and their use in these models further proved that belief. A team's seed and the conference they are a member of were also chosen in a large majority of models. In regard to a team's seed, one of the simplest bracket selection strategies is to choose the team with the better seed in each matchup as these teams are expected to perform well, so this predictor being utilized commonly is not surprising. This does provide evidence that even though the NCAA tournament is known for its "madness", in general, better seeded teams tend to perform well. A team's conference is a more difficult predictor to infer findings from, as there are 32 conferences that have been observed over the 10 year period tested, but not all conferences are present in every tournament nor is there equal representation from each conference. For example, some conferences are much larger than others containing many of the name-brand schools commonly known, and these conferences may provide more teams in every tournament compared to smaller conferences. Given this, garnering generalized results for how this variable was implemented is difficult, but nonetheless it appears that there is a significant difference in the frequency at which teams from certain conferences advance to most rounds. Overall, these most salient predictors provide evidence for the main drivers of team success in the NCAA tournament, which can be used when determining the success of a team in any given tournament.

When considering the relative benefit of utilizing this method, its comparison to other readily available predictive models using similar advanced analytics needed to be considered. Given the popularity, performance, and accessibility of FiveThirtyEight's bracket forecast their predictions served as an adequate comparison point for the strategy developed in this study. In short, the proposed method outperformed FiveThirtyEight's forecasts in every facet, scoring more points on average with less variability, leading to this strategy beating FiveThirtyEight in 6 out of 7 years tested. Comparing the analytical implementation of each approach is difficult, as FiveThirtyEight does not public disclose every detail of their predictive model. They state their use of team power ratings, similar to those utilized in this study, but do not discuss feature engineering, model tuning, or testing methods utilized. Given this, focusing on the differences in assumptions between these two approaches may provide more context into the relative success of each method. Notably, the proposed method's recursive bracket approach is the main difference to highlight. FiveThirtyEight's model considers each individual matchup a team may face in the tournament to forecast the chance they advance to each round. In comparison, the approach utilized in this study does not consider matchups and instead assumes there are unique qualities possessed by the teams who advance to each round of the tournament. Both assumptions have their advantages and drawbacks, but when considering the goal of filling out brackets is to score more points than other users, it can be seen why prioritizing the select few picks worth the most points may lead to increased performance. Overall, this comparison not only provides validity to the proposed method, but also shows the effects assumptions can have when implementing game theory.

Recommendations for Future Work

While the success of the developed method has been seen thus far, there are areas for improvement which could lead to increased results. Most notably, this strategy does not consider possible matchups between teams, such as how FiveThirtyEight's forecasts are built. While the recursive selection strategy appears to have some success, assuming that possible matchups between teams in each tournament has no effect on individual team success seems farfetched. Given this, it is possible that the method utilized in this study could have missed some trends in historical NCAA tournaments that could have been gained from assessing the differences between teams in each matchup. Developing a more traditional model for filling out brackets, similar to FiveThirtyEight, could provide more context into this phenomenon. Alternatively, it is possible that a hybrid approach which considers recursive selection and potential matchups could provide a balance between these differing assumptions and should be examined as well.

Additionally, collecting additional data not only could improve current models, but could also provide more reliable evidence of the potential benefit of this approach. While utilizing 10 years of data appeared to provide relatively stable predictions and results, the NCAA tournament dates back to 1983 which is another 30 years of data that could be analyzed. Additional data was not utilized for this study due to complications in web scraping which would have taken considerably more time to address than was allotted for this study, but nonetheless capturing and testing this data could provide additional benefit and support to the models developed. Similarly, sourcing additional predictor variables such as historical matchups or individual player data could provide additional benefit to the models developed. The current approach provides a good balance between intrinsic metrics to quantify team performance as well as external factors regarding the tournament teams are competing in, but none of these consider more nuanced details such as the individual players on each team who could have significant impacts in any given game. Given this, sourcing additional features for modeling could provide additional benefit to models or further confirm the importance of the predictors already being used.

Lessons Learned

Embarking on this project taught me a number of crucial modeling techniques and provided evidence for the importance of methods that I have previous experience with. Most notably, I have never dealt with issues of significant class imbalance for classification models before this project. When first fitting models it was clear that this issue was negatively affecting the results of models, which required me to do extensive research into methods for addressing class imbalance and how to best implement them for the purposes of my project. This led to me learning about the mechanics of SMOTE and Tomek Links, how they could be beneficial for my models, and how to implement them. This skill was invaluable not only for the success of my project but also for future project where I may deal with similar issues. Similarly, I have never ensembled multiple models for the purposes of improved model performance. When determining what models to use for this study, I settled on logistic regression, random forest, gradient boosting, and multi-layer perceptron models with the belief that one model may perform significantly better than the rest. Once it became clear that each model had its benefits, this led me to consider how the advantages of each learner could be leveraged for my overall method. After discovering how multiple learners could be ensembled for classification I learned how to combine these models effectively using weightings for each model, and eventually saw the power utilizing this technique could have. Through learning this skill, creating ensemble models is a technique I am able to consider and implement effectively now which could provide benefit for future projects.

Furthermore, I have previous experience with other analytical techniques such as hyperparameter tuning, feature selection, cross-validation, and choosing classification metrics, but utilizing these skills in this project not only refined my abilities but also further proved the importance each of these factors can have on model performance. There were multiple methods I considered for tuning hyperparameters and selecting predictors, as well as various data splits I could have utilized to do so. By considering the purpose of my models which was to develop reliable, highly accurate predictions, I was able to choose the best technique for each step of my process. This allowed me to refine what hyperparameters to tune, how to select features, how to best develop cross-validation procedures, and which classification metrics would best quantify the performance of my models, all of which are valuable skills I can utilize for future projects. Overall, this study taught me valuable new skills while further developing pre-existing abilities I have, both of which will benefit me as an analytics professional when embarking on future projects.

References

1. Batista, Gustavo & Prati, Ronaldo & Monard, Maria-Carolina. (2004). A Study of the Behavior of Several Methods for Balancing machine Learning Training Data. SIGKDD Explorations. 6. 20-29. 10.1145/1007730.1007735.
2. Boice, J., & Silver, N. (2018, March 11). *How our march madness predictions work*. FiveThirtyEight. https://fivethirtyeight.com/features/how-our-march- madness-predictions-work/
3. Chawla, N. V. and Bowyer, K. W. and Hall, L. O. and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 16. 321–357. http://dx.doi.org/10.1613/jair.953
4. Dutta, S., Jacobson, S. & Sauppe, J. (2017). Identifying NCAA tournament upsets using Balance Optimization Subset Selection. Journal of Quantitative Analysis in Sports, 13(2), 79-93. https://doi.org/10.1515/jqas-2016-0062
5. FiveThirtyEight. (2024). *2023 march madness predictions*. FiveThirtyEight. https://projects.fivethirtyeight.com/2023-march-madness-predictions/
6. Katz, J. (2015, March 15). *Here's how our N.C.A.A. bracket works*. The New York Times. https://www.nytimes.com/2015/03/16/upshot/heres-how-our-ncaa-bracket-works.html
7. KenPom. (2024). *2023 Pomeroy Ratings*. KenPom. https://kenpom.com/
8. Leonhart, D. (2015, March 15). *March madness 2015: Welcome to a difference kind of bracket*. The New York Times. https://www.nytimes.com/2015/03/16/upshot/march-madness-welcome-to- a-different-kind-of-bracket.html
9. Mass, A. J. (2023, March 2). *How to fill out a march madness tournament bracket: all the basics so you can join the madness.* ESPN. https://www.espn.com/fantasy/basketball/story/_/id/26103343/how-fill-tournament-bracket-all-basics-join-madness
10. Niemi J. B., Carlin B. P., Alexander J. M. (2008). Contrarian strategies for NCAA tournament pools: A cure for march madness? *CHANCE 21*, 35-42. https://doi.org/10.1007/s00144-008-0009-3
11. Ota, K. (2023, March 16). *ESPN tournament challenge sets new all-time record: 20 million brackets*. ESPN Press Room. https://espnpressroom.com/us/press-releases/2023/03/espn-tournament-challenge-sets-new-all-time-record-20-million-brackets/
12. Pomeroy, Ken. (2012, June 8). *Ratings Glossary*. KenPom. https://kenpom.com/blog/page/50/
13. Sport Reference. (2024). *College basketball stats and history.* Sports Reference. https://www.sports-reference.com/cbb/

14. Tomek, I. (1976) Two Modifications of CNN. IEEE Transactions on Systems Man and Communications, 6, 769-772. http://dx.doi.org/10.1109/TSMC.1976.4309452
15. Wilco, D. (2023, March 15). *What is march madness: The NCAA tournament explained.* NCAA. https://www.ncaa.com/news/basketball- men/bracketiq/2023-03-15/what-march-madness-ncaa-tournament-explained#:~:text=The%20NCAA%20Division%20I%20men's%20basketball%20tournament%20is%20a%20single,only%20four%20teams%20are%20le ft.