Group 21 – Selection Consultants

Data Mining Final Project Report

Executive Summary:

Every March, the NCAA Tournament Selection committee faces a dilemma of who to put into the field, as well as what the seeding of the field should be. We have created a model that can assist the selection committee to make seeding decisions and which bubble teams are in and out for future NCAA tournaments. A dataset sourced from Kaggle.com provides many useful performance metrics for every single team starting in 2014, excluding the 2020 non-tournament season. Notable features from this dataset include wins, strength of record, shooting percentage, rebound percentage, and turnover rate. The response feature for this model was labeled "MakeTourney", a binary variable used to determine if a team qualified for the tournament that season. Logistic Regression became the clear model of choice for this project. The chosen model had the strongest AUC and the best confusion matrix distribution when compared to other model types.

To put the model to the test, we removed the response variable from the 2023 dataset and had our model predict which teams would make the field, using the past 10 years as training data. Unsurprisingly given the training data AUC, the model performed very well. The top team our model outputted was UConn, who were the eventual tournament champions. All four final four teams had equal to or better seeding in our model than the actual tournament. Overall our model produced accurate results compared to the 2023 bracket, including what bubble teams were left in and out. A notable limitation of the model is the omission of conference tournaments. It treats every team like they need an at large bid and does not consider that conference

tournament champions receive automatic bids into the tournament. Regardless of certain limitations, our final recommendation is for the committee to use the model to assist in the selection and seeding process of future tournaments. The following report describes the processes of the model in great detail, as well as a closer look into multiple results produced by the model.

Problem Description:

The NCAA Selection Committee plays a pivotal role in the thrilling and competitive landscape of college basketball, particularly in the high-stakes NCAA Men's Basketball Tournament, widely known as "March Madness." This tournament is a focal point for over 350 Division 1 teams annually, each seeking a bid into the bracket. While some teams secure an automatic bid by winning their conference tournaments, a significant number of slots are filled by non-conference champions based on their performance throughout the regular season.

For these at-large bids, the problem at hand is driven by taking a data-centric approach aimed at simplifying the selection process. The primary goal is to develop a model that not only identifies the critical metrics contributing to a team's eligibility for the tournament but also predicts a team's ability to secure a spot. The purpose for this model is to aid the NCAA Selection Committee in their decision-making process and provide valuable insights to coaches and college teams, especially those seeking an at-large bid.

Furthermore, the proposal extends beyond just deciding which teams are in and out. It aspires to create predictive models that can estimate a team's potential seeding in the tournament and forecast their performance based on regular-season statistics. This approach is designed to provide a transparent and understandable explanation for teams and coaches on their

qualification status. By pinpointing the areas of strength and those needing improvement, coaches can strategize more effectively, focusing on aspects that enhance their chances of not just participating in March Madness but also contending for the NCAA title. Simply put, this proposal is about using data to create a model that brings clarity and reasoning to the complex and debatable process of NCAA tournament team selection, seeding, and performance prediction.

Data Description:

Our data contains information on college basketball teams across all NCAA conferences from 2013-2019, and 2021-2022. Data from the 2020 season was kept separate due to the postseason being canceled because of the coronavirus. The data used was sourced from Kaggle. After cleaning the data the final CSV file contained 3,160 rows and 25 categorical attributes. The target variable for our model is whether a team will make the March Madness tournament. To predict this outcome there were 13 variables used. Common ones were the number of games played, wins/losses, power rating, and offensive/defensive efficiency. Variables that aren't as straightforward were also used, such as "wins above bubble." WAB refers to the cutoff value between making the NCAA March Madness tournament and not. Others include 3PO (3-point percentage), Tempo (how many possessions a given team should have in 40 minutes), and others.

	Team	Conference	Games	Wins	Losses	ffensive Efficienc	Defense Efficiency	Power Rating
18	Iowa	B10	36	26	10	121.1	97.2	0.9265
363	Iowa	B10	29	21	8	123.5	95.7	0.9491
1000	Iowa	B10	35	23	12	116.1	100.6	0.8385

The variables in our data set are discrete, continuous, and categorical. The dataset from Kaggle had 25 variables but only 13 were used to build the model. The target variable for our model is

"MakeTourney" which is the categorical feature, 1 meaning they did make the tournament and 0 for not.

Data Mining Solution:

In our attempt to find the best model to predict "MakeTourney" our binary target variable, we used three methods, logistic regression, gradient boosting, and random forest methods. We built the logistic regression model to give every meaningful feature a value to represent how each feature affects the log odds of making the tournament. We found that the conference a team played in had the largest impact on the log odds of that team making the tournament, typically if the team was in a smaller conference this would benefit the team, increasing the log odds by about a handful of percent. For the random forest, we used 200 trees and considered 3 attributes at every split, we also used replicable training for repeatable model building, and we finally limited the individual tree depth to 5 and did not split subsets smaller than 5. We used the typical sickit-learn method, as well as 40 trees and a 0.035 learning rate. Like the random forest, we used replicable training and did not split subsets smaller than 5, we did limit the depth of individual trees to 3, which was the main hyperparameter difference between our random forest and gradient boosting models.

All three models performed very similarly and very well, accuracy was no less than 0.958 on the training set, and no less than 0.946 on the testing set. We would recommend the logistic regression model as the most accurate and the model we would recommend to the NCAA Selection Committee to use to help determine tournament teams versus non-tournament teams. One problem we have with the models is that we do not lock the teams making the

tournament at 68 teams, so we would have to compare the results of the model versus the teams in question and use a combination of the two to determine the teams that will make the tournament.

Recommendation and Conclusion:

Our recommendation is for the selection committee to utilize this model to assist in future NCAA tournaments, beginning in 2024. This recommendation is backed by the model's notable accuracy in projecting the 2023 tournament field and its superior performance relative to the actual bracket. When comparing last year's bubble teams in the actual bracket to our model's predictions, compelling evidence emerges indicating that our model excelled in predicting the rightful inclusions. In the 2023 bracket, Nevada and Mississippi State were the last two teams in the tournament, and neither team won a game. Our model did not have either team qualify for the tournament. On the flip side, a team the model included that the committee left out was North Texas. UNT would go on to win the National Invitational Tournament, and the entire final four of the NIT was made up of teams that our model qualified for the NCAA tournament.

Not only should the committee use our model to determine who gets in, we also recommend that the model can assist in seeding teams. The probability output for the predictions of future data serves as a power ranking for our model. The top four teams in the output should be given a one seed, and so on. Once again, the model performed remarkably well when looking at the 2023 tournament. In the output, the team with the highest probability and what the model predicted was the best team going in was the UConn Huskies. For the real bracket, UConn was

given a four seed, and they went on to win the championship. The eventual final four also performed well in the model. San Diego State, Miami, and FAU were ranked 17, 20, and 33, respectively. Our model had SDSU at 11, Miami at 21, and FAU at 15 heading into the bracket. Using this model can help gain better insight into who the best teams truly are.

The model does come with a couple limitations. For one, it does not take into account conference tournaments. The winners of each conference tournament are awarded an automatic berth into the NCAA tournament. While typically most conference tournament winners are already tournament worthy teams, there are times where that is not always the case. For example, Fairleigh Dickinson is well known for upsetting Purdue in the 2023 tournament. While our model did not qualify them for the tournament, they got in for winning the Northeast Conference tournament. Another limitation is the seeding of bubble teams based on conferences. Since the model uses the output probability as a power ranking, it simply gives the last four teams a 16 seed, regardless of conference. Traditionally, power conference bubble teams are given an 11 seed and tasked with playing each other in the first round.

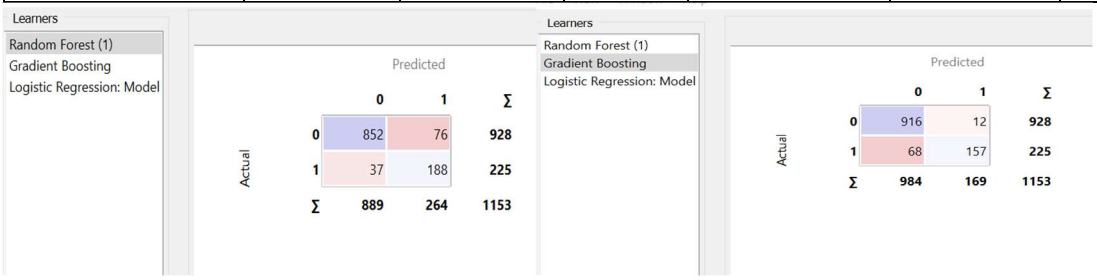
To reiterate, we recommend using our model to assist the NCAA selection committee for future work, beginning in 2024. Backed by comparisons to last year's tournament, our model has shown to be successful and accurate regarding bubble teams and seeding decisions. While it is impossible to predict everything that happens during March madness, using this model backed by data can help the selection committee make informed decisions.

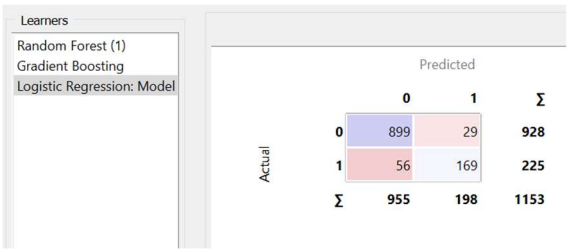
Appendix:

Training Error	AUC	CA	F1	Precision	Recall	MCC
Random Forest	0.965	0.898	0.902	0.909	0.898	0.697
Gradient Boosting	0.966	0.93	0.925	0.93	0.93	0.757
Logistic Regression	0.958	0.924	0.922	0.922	0.924	0.742

Testing Error	AUC	CA	F1	Precision	Recall	MCC
Random Forest	0.946	0.902	0.905	0.91	0.902	0.711
Gradient Boosting	0.96	0.931	0.927	0.931	0.931	0.767
Logistic Regression	0.961	0.926	0.924	0.922	0.926	0.756

Cross Validation	AUC	CA	F1	Precision	Recall	MCC
Random Forest	0.948	0.894	0.897	0.903	0.894	0.696
Gradient Boosting	0.954	0.926	0.921	0.928	0.926	0.759
Logistic Regression	0.946	0.913	0.911	0.91	0.913	0.719





Works Cited

Norlander, Matt. "2023 NCAA Tournament Bracket: Ranking Every Team Playing in March Madness from No. 1 to 68." *CBSSports.com*, CBS, 13 Mar. 2023, www.cbssports.com/college-basketball/news/2023-ncaa-tournament-bracket-ranking-every-team-playing-in-march-madness-from-no-1-to-68/. Accessed 10 Dec. 2023.

Selbe, Nick. "Selection Committee Reveals Official 1–68 Rankings for 2023 Men's NCAA Tournament." *Sports Illustrated*, 12 Mar. 2023, www.si.com/college/2023/03/12/selection-committee-rankings-2023-mens-ncaa-tournament-march-madness. Accessed 10 Dec. 2023.