Proposal

Owen Fiore and Sam Hughes

10/26/2022

a

Sam Hughes and Owen Fiore

b

March Madness is an annual college basketball team that features 64 to 68 of the best teams in the country competing in a single elimination tournament. The basketball tournament is highly watched and followed by sports fans who fill out brackets and try to predict the winners of all 63 to 67 games (The rules have changed over the years allowing for either 64 or 68 teams). According to boydsbets.com, 70 million brackets are created per year by fans attempting to fill out a perfect bracket. However, there are 9.2 quintillion possible combinations for brackets making this feat extremely challenging.

c

The data comes from: https://www.kaggle.com/competitions/mens-march-mania-2022. While there is data from as far back as 1985, for the purposes of this project, only data from 2005 and onward will be used. This is because we want to keep the data relevant to modern basketball strategy, and because certain metrics such as polling rankings were not tracked in most tournaments prior to 2005. Note that any variables in which we are taking the difference, we are measuring team1 minus team0 for the given metric. Additionally, it is worth noting that all the data is focused on a single game in a single season, however some teams have appeared multiple times in the tournament so their team IDs will appear multiple times. Here are all of the variables:

Variable	Description	Role	Type
winner	The team that won the game	Response	Binary
round	The round of the tournament	Predictor	Discrete
team0	The id of one of the two teams that played in the game	Predictor	Categorical
team1	The id of one of the two teams that played in the game	Predictor	Categorical
team0_coach	The head coach of team0	Predictor	Categorical
team1_coach	The head coach of team1	Predictor	Categorical
team0_conf	The conference of team0	Predictor	Categorical
team1_conf	The conference of team1	Predictor	Categorical

Variable	Description	Role	Type
team0_conf_standing	The placement of team0 within their conference tournament	Predictor	Continuous (could be a fraction if team finishes tied between 3rd and 4th, for example)
team1_conf_standing	The placement of team1 within their conference tournament	Predictor	Continuous (could be a fraction if team finishes tied between 3rd and 4th, for example)
prev_matchups	The win percentage of team0 against team1 in previous matches during the season, if any	Predictor	Continuous
seed_diff	The difference in seeding	Predictor	Discrete
strength_diff	The difference in strength; strength being the average ranking among numerous polling sources	Predictor	Continuous
win_pct_diff	The difference in win percentage	Predictor	Continuous
sos_loss_diff	The difference in the median strength of opponents in regular season losses	Predictor	Continuous
sos_win_diff	The difference in the median strength of opponents in regular season wins	Predictor	Continuous
best_win_diff	The difference in strength of best beaten opponent	Predictor	Continuous
worst_loss_diff	The difference in strength of opponent in worst defeat	Predictor	Continuous
common_opps_diff	The difference in win percentage among opponents that both teams played	Predictor	Continuous

There are 1,053 observations of tournament games from 2005 to 2021. However, it is possible to double the number of observations by flipping team0 and team1 to generate an additional 1,053 observations for a total of 2,106 observations. Doubling the data will also allow the model to better learn the symmetry between the observations (as long as we keep them within the same cross validation fold to prevent data leakage).

d

This is supervised learning because we are interested in predicting whether a team will win or lose based on the 18 variables listed above. This is also binary classification as there are only two outcomes that we want to predict: win or loss (games in the tournament cannot end in tie and will play consecutive 5 minute overtime periods until a team wins). We will likely use advanced methods such as XGBoost because we are mainly concerned in finding the most accurate predictions, and due to the complexity and multi-dimensionality of our predictors in relation to our response.

e

We do not have any comments or concerns.