Asher Erickson
Colin Truran
Elise Longenecker
Yvan Longin

Applied Analytics Project - Week 2 Assignment

Target Variable:

The target variable for this project is game outcome for the NCAA March Madness Men's Basketball tournament. This variable is defined as whether a given team wins or loses a specific matchup. For our models, this outcome will be represented as a binary variable, where win is equal to 1 and loss is equal to 0.

Using this type of target variable, we aim to estimate the probability of a team winning each game. This method is well-suited for March Madness, since the tournament has a reputation as having high uncertainty and variability in game outcomes. The probabilities represented in the target variable will also be useful for distinguishing between closely matched games and those with higher upset potential.

By defining the target variable at the game level, each game will be treated as a separate observation. This will allow us to have a large number of data points and compare outcomes across seasons. This structure will be best for predictive accuracy and interpretability. It will allow us to effectively evaluate model performance, using metrics like accuracy. We will then be able to assess whether or not predicted probabilities can actually be useful in selecting brackets, specifically in a sports betting context.

Predictor Variables:

Since the goal of our target variable is to predict the outcome of an NCAA tournament matchup, we rely on a dataset that contains season-level statistics and metrics for each participating team. All predictors are based on team performance prior to the tournament to avoid any data leakage. These metrics are used to capture each team's overall strength, playing style, and matchup tendencies, which help highlight strengths and weaknesses against teams with different styles of play.

Given that the dataset contains 165 total columns, we initially focus on the predictors that best represent these core characteristics. Because our dataset is large, we may increase or decrease the number of predictors used as we evaluate model performance and refine our approach.

Overall Strength

Key predictors: Adjusted Offensive Efficiency, Adjusted Defensive Efficiency, Adjusted Efficiency Margin, Seed

These variables capture how strong a team is on a possession-by-possession basis and how that strength translates against high-level competition. Adjusted efficiency metrics are important because they take into account the opponent's strength and pace rather than only focusing on raw scoring totals. These metrics give us a summary of how well a team scores, defends, and performs overall.

Playing Style

Key predictors: Tempo and adjusted tempo, average possession length, shooting efficiency metrics (eFG%, FG2%, FG3%), turnover rate, free-throw rate

Playing style helps explain how teams operate their offense and control games. Tempo and possession length show whether a team prefers fast or slow-paced games. Shooting efficiencies, turnover rates, and free-throw rates further describe the preferences of an offense and their tendencies. These factors matter because different styles can either strengthen or cancel out a team's advantages, especially when teams with different approaches play each other.

Matchup Tendencies

Key predictors: Opponent shooting percentages, defensive metrics (blocks, steals), rebounding rates, Team Height, Experience

Matchup tendencies focus on how a team performs against specific types of opponents. Defensive metrics and opponent shooting percentages indicate how well a team can disrupt opposing offenses while. Together, these variables help identify situations where a team's strengths align well, or poorly, against an opponent's style.

Dataset Exploration:

The dataset used for this project is sourced from Kaggle, but will likely use another helper dataset to ensure a seamless transition into the format we need to model accurately. The dataset consists of the NCAA Division I Men's basketball teams that participated in the March Madness tournament from 2002 to 2025. Each tournament consists of 64 teams playing a total of 63 games. Across 22 years, this means there are 1,408 rows in the original dataset. In the format used for modeling, the final row total will be 1,353 (33 earlier games do not have complete data). The reason for this is each row in our modeling dataset will be a specific game of the tournament.

The initial count of columns in the dataset is 165. This includes many categorical and numerical variables for each team including, but not limited to: year, conference, team name, coach, efficiency ratings, shooting percentages, rebounding totals, turnover rates, coaching tenure, etc. The categorical variables will mainly be removed due to the strict categorizing nature of each of them. A game is not decided by the name of a coach, but it could be affected by the tenure/experience of the coach. The numerical variables will be the main predictors in the models used later in the project.

Overall, the dataset provides a thorough statistical analysis of each team which should give the best chance at producing results in a further model. With 22 years of data, any trends that are present in today's brand of college basketball should be able to be found with powerful models. Steps of feature engineering, cleaning, and final dataset formulation will be vital for the remaining objectives of this project.