Perfecting the March Madness Bracket

Alan Qiao, Angelina Huang, Julia Healy-Parera, John Lee, Daniel Zhu

Introduction and Research Questions

Each student's college experience is uniquely shaped by the culture of the school which they attend. At schools like Alabama and Michigan, school spirit manifests itself in the form of enthusiastic and jeering football stadiums. For others, like UConn and Duke, college basketball shapes school reputation and student connection in similar ways. But beyond the grasp each college athletics team has on its respective campus, college teams represent a multi billion dollar industry that captivates those across the nation (and the world!). Consequently, much research and thought goes into the construction and strategy of each team and game. Inspired by this, we are seeking to learn more about the college sport that has most impactfully shaped our college experience: basketball. Beyond simple investigation, we hope to dive deeper into the model we create to investigate its efficacy after removing the three most influential predictors.

Specifically, we aim to answer these two research questions:

1. What indicators are the best predictors of success or games won in NCAA basketball and in March Madness (2014 - 2024)?
2. Specifically, are historical win percentages and averages or game-specific statistics more influential as predictors of success? How effective are the resulting models?

Because factors influencing success in basketball are likely to interact with each other, we have attempted to account for multicollinearity or confounding factors to overshadow the results of our model and data exploration. In addition, because external factors are not included in datasets exclusively relating to NCAA basketball statistics, we have cleaned and merged multiple of them in a sensible manner. Despite the possible difficulties created by term interactions and dataset merging, the widespread availability of NCAA statistics will allow us to explore and extrapolate various questions that we seek to reason through using this project. A multitude of games are played during both pre- and regular-season, allowing for an abundance of datasets available for analysis alongside further data surrounding each institution.

In addition, due to the relevancy of this project discussed earlier in the introduction, college basketball prediction will be interesting both within and outside of the Duke student population. Moreover, coaches and sports fanatics will always want to know more about predicting the game outcomes and relevant metrics impacting performance. By investigating external factors like team funding, we may also reveal findings about the effects of commercializing sports.

Access the code, data, and supplemental materials at: https://github.com/JuliaHealeyParera/cs216

Data Sources

1. March Madness Dataset

   Source: Kaggle [Nishaan Amin. (March 2024). March Madness Data]

   - This dataset provides data related to the NCAA Men's Basketball Tournament (March Madness), including information on team performance and game results. The dataset's relevant predictors included their conference results, seed results, and team results. This information was incorporated into our initial predictive model. The data wrangling we used for the data set involved transforming the data frame such that each row represented one game, with both team's statistics represented in that row. Originally, the dataframe was structured such that every 2 rows represented one game, with team A's statistics being in one row and team B's statistics being in the other. We also wrangled the data such that there were no duplicate games. From there, the data was joined with our larger dataframe.

2. Scoring Datasets
   Source: Team Rankings [NCAA Basketball Team Win Trends - All Games. (2024). TeamRankings.com]
   - This source includes a multitude of datasets from Team Rankings reporting variables like a given school's home game win rate, away win rate, underdog win rate (i.e. games won against teams of higher ranking), after-loss win rate (game-winning rate after losing the last game), and more. Because these datasets were in tabular form on the Team Rankings website, in order to download them, we used Python's Beautiful Soup module to web scrape the data into multiple CSV files. Since the data was structured in the same format as our larger data frame, conjoining them was rather straightforward.
3. Team Data
   Source: sports-reference.com [NCAA Seasons Index. (2024). College Basketball at Sports-Reference.com]
   - The following website provides data on each relevant team that the project is inquiring about. Relevant information includes data sets pertaining to each team's roster, whether certain players were ranked nationally in the top 100 in highschool, points per 40 minutes, stats per 100 losses, as well as more advanced statistics pertaining to each individual player's overall attributes. Using this website, we scraped data from the 21-22, 22-23, and 23-24 seasons and included all teams that had a March Madness appearance which was denoted with a 'NCAA' marking besides their name. From each individual team's roster, a 'top 7 ranking' was created based on each player's points scored—from there, we counted the number of upperclassmen (juniors or seniors) in each. This data was appended into a separate data frame denoted by a column counting the number of upperclassmen for each school. We pulled team specific statistics into this separate data frame as well (free throw percentage, rebounds, etc).

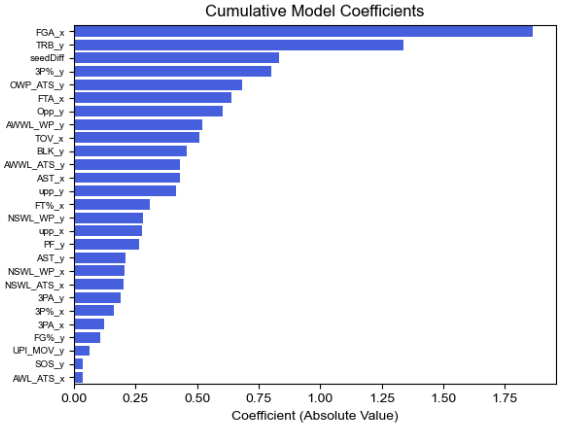
Results and Methods

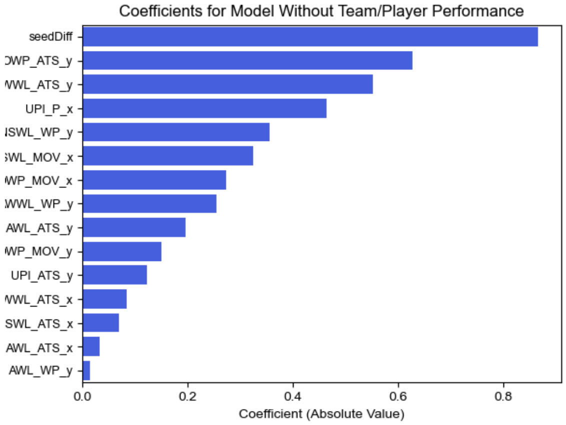
A main concern of all model fitting is the possibility of multicollinearity confounding model results. Multicollinearity occurs when multiple variables are correlated. In smaller models, this issue can be dealt with using pair plots, practical reasoning, and, for linear regression, the use of an F-statistic. Logistic regression allows for specialized ways to deal with multicollinearity, which we employed for this project. Lasso regression, which is a special regression technique that algorithmically eliminates irrelevant or multicollinear variables, was the method used for the three models created in this project.

Lasso regression works by penalizing variable coefficients if they are multicollinear or uninfluential, thereby shrinking irrelevant or harmful features to 0. We chose this technique because it is effective in highlighting relevant features in high-dimension data (i.e. datasets with many predictors, like ours) and effectively dealing with the multicollinearity that would otherwise likely be present.

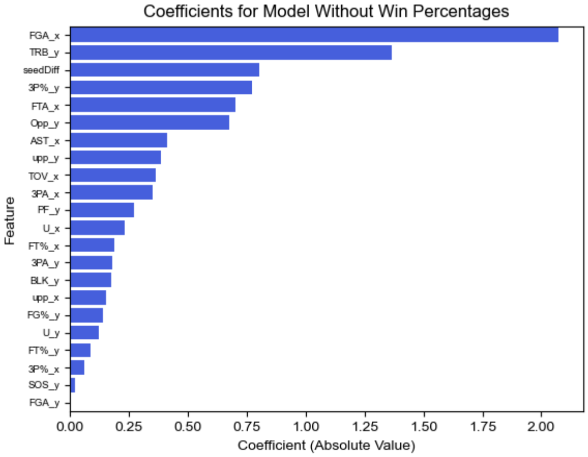
In our model creation, we incorporated win percentages and ratios for a variety of scenarios (ex. away vs. home game), aggregate team performance statistics (ex. total rebound percentage), averaged individual player statistics (ex. field goals attempted), and an interaction term considering a team's win rate if they are/are not considered the underdog. Our first model incorporated all of these predictors, our second only used the interaction term and the win percentage, and the third only used the aggregate team performance and the individual player statistics. We hypothesized that either the first or second model would be most effective in predicting which team would win a given match, since both of these models would be using literal win-rates from differing scenarios. The interpretation of all feature acronyms can be found in the appendix.

Our first model, displayed in the feature plot on the right, had an accuracy of 0.75. Its most influential predictor (which is also the predictor with the highest coefficient absolute value, because of the nature of lasso regression) was the total number of field goals attempted by the team. The majority of the predictors deemed to be influential and non multi-collinear by the lasso regression technique were not win percentages.





Our second model had an accuracy of 0.54. Its most influential predictor was, noticeably, not a win percentage. Instead, it was seedDiff, or the difference in rankings between the two teams. An additional interesting observation is that the most influential win percentages in this model are not the same as the most influential win percentages in Model 1. For example, although the underdog win percentage interaction term was very influential in Model 2, it is not an influential predictor at all in Model 1.



Our third model eliminated all win percentages and our interaction term (which was dependent on a win percentage) and evaluated whether a given team would win a match solely based on individual and team characteristics. The accuracy of this model was 0.75, the same as Model 1. Notably, however, the two models have different influential predictors since Model 3 does not incorporate any win percentages. Differences in seed (rankings) between Team X and Y is a relevant predictor here as well.

This is an interesting observation because seedDiff was calculated as an absolute value, positive regardless of whether X or Y was ranked higher (because the team assignments were arbitrary). Since this is the case, it is interesting that a variable that is not necessarily informative of whether X or Y is considered to be a more successful team would be such an influential predictor (especially because it was not used as an interaction term).

As noted in the previous paragraph, the accuracy of models 1 and 3 were the same. Not only is this surprising because the two models are very different in their under-the-hood makeup, but also because the accuracy of Model 2 is significantly lower. This is contrary to our hypothesis, which stated that either Model 1 or Model 2 would be most effective in predicting a game's outcome given that it incorporates what we considered to be the most straightforward predictors. However, with Model 2's accuracy of 0.54 and Model 1 and Model 3's accuracy of 0.75, the opposite was true. This provides interesting insight on the value of specific basketball players and team statistics as opposed to overall rates on how often a team wins. Field goal attempts and total rebounds, specifically, seem to be the most indicative indicators of a team's success. Such interpretations should be done with caution, however, since these variables are only considered to be the most influential within the context of this specific regression technique and feature combination.

Limitations and Future Work

Limitations:

This analysis has several limitations. First, the data wrangling methodology for upperclassmen focused only on the top seven players by points scored from each team, which may exclude important contributors. For example, star or important players on a team may contribute in numerous other ways including but not limited to assists, rebounds, steals, team morale, etc. If free point throws are significantly influential for Team X but not for Team Y, it may bias the results because the labels of Team X and Y are arbitrary. And beyond that, due to injuries are extenuating circumstances, player statistics may not necessarily reflect full potential.

Additionally, team statistics, such as total rebounds of a team that season reflects the number of games played that season instead of the rebound statistic. We could, in the future, analyze statistics per game instead of per season. Additionally, the dataset is restricted to the 2021–2024 seasons, providing a narrow time frame that does not fully capture long-term trends, especially as college basketball programs tend to be variable from year to year depending on recruiting class. Another limitation is that not all teams in the dataset participated in the March Madness playoffs during these years, which skews the analysis only towards teams that win their individual conferences. Unequal weighting is also a concern, as teams that advance further in the playoffs within the same year appear multiple times with the same attributes, despite playing against different opponents. This creates a bias towards teams with higher win rates, favoring consistently successful programs and potentially skewing results towards better-performing teams.

Future Work:

Future work that could be implemented moving forward for this project could involve testing for more interaction terms and adjusting for more accurate statistics. In addition, we could analyze how teams fare in high-stakes settings with other tournaments, such as the Conference tournament. With sports and basketball specifically, the data sets that we've collected contain a plethora of variables beyond the ones we've analyzed in depth in this report. Further, potentially changing some of the methodologies we used to wrangle the data, as described in the limitations portion above, could provide further insights. For example, accounting for more overall statistics when aggregating our 'top players' for each college program may provide a more well-rounded insight. Additionally, simply accounting for more data across more than just three seasons would innately provide more holistic insights.

Conclusion

Our analysis demonstrates that the most influential predictors of NCAA basketball success (measured by games won) are team-specific statistics. However, Model 1 and 3's equal performance with the highest accuracy, recall, and precision scores indicates that team-specific statistics hold significant predictive value, and are the features that the model relies on. More specifically, answering the first research question, the top features of both models are: average attempted field goals, total rebound percentage, seed difference between two teams, average successful 3-pointers, and free-throw attempts.

Thus, our analysis demonstrates that the most influential predictors of NCAA basketball success (measured by games won) are mostly dependent on team-specific statistics and not the historical win percentage of team A over team B. These predictors collectively align with existing basketball analytics theories, underscoring the importance of technical playing performance and player experience, among other factors, in driving performance. Our findings also highlight opportunities for future refinement in basketball analytics, especially through broader datasets and more nuanced statistical metrics, to further enhance accuracy and provide deeper insights into the dynamics of college basketball performance.

Appendix:

stat_x = one team's statistic, stat_y = another team's statistic

Predictor	**Meaning**
afterWin_winLoss_Win_percent (AWWL_WP)	Team's win percentage after winning a game
afterWin_winLoss_MOV (AWWL_MOV)	Team's average margin of victory after winning a game
afterWin_winLoss_ATS (AWWL_ATS)	Team's against the spread percentage after winning a game
underdogPerc_interaction_percent (UPI_P)	Interaction term with a team's win percent when underdog and their overall win percentage
underdogPerc_interaction_MOV (UPI_MOV)	Interaction term with a team's win percent when underdog and their overall margin of victory
underdogPerc_interaction_ATS (UPI_ATS)	Interaction term with a team's win percent when underdog and their overall against the spread
away_winLoss_MOV (AWL_MOV)	Team's average margin of victory playing an away game

away_winLoss_ATS (AWL_ATS)	Team’s against the spread percentage playing an away game
neutralSite_winLoss_Win_percent (NSWL_WP)	Team’s win percentage at a neutral site
neutralSite_winLoss_MOV (NSWL_MOV)	Team’s average margin of victory playing at a neutral site
neutralSite_winLoss_ATS (NSWL_ATS)	Team’s against the spread percentage at a neutral site
overallWinPercentage_Win_percent (OWP_WP)	A team’s overall win percentage
overallWinPercentage_MOV (OWP_MOV)	Team’s overall margin of victory
overallWinPercentage_ATS (OWP_ATS)	Team’s overall against the spread percentage
Upperclassmen (upp)	Number of upperclassmen in the team’s top 7 players
SRS	Simple Rating System = Point Differential + Strength of Schedule
SOS	A team’s strength of schedule
FG	Total number of a team’s field goals
FGA	Total number of a team’s field goal attempts
FG%	Team’s field goal percentage
3P	Team’s total 3-point field goals
3PA	Total number of a team’s 3-point field goal attempts
3P%	Team’s 3-point field goal percentage
FT	Total number of a team’s free throws
FTA	Total number of a team’s free throw attempts
FT%	Team’s free throw percentage
ORB	Total number of a team’s offensive rebounds
TRB	Total number of a team’s rebounds
AST	Total number of a team’s assists
STL	Total number of a team’s steals

BLK	Total number of a team’s blocks
TOV	Total number of a team’s turnovers
PF	Total number of a team’s personal fouls