Asmita Nayak, Brooke Pauly, Shashwat Prabhakar, Frank Thaingtham, Devan Patel, Devarshi Dave (Group 9)

Professor Thomas

STOR 320: Methods and Models of Data Science

28 April 2025

A Deep Investigation Into The Statistics Behind College Basketball

Introduction

It's that time of year again. The time when millions of Americans meticulously fill out forms, hoping every selection is just right. But this isn't tax season...it's March Madness. Filled with buzzer-beaters and Cinderella stories, the NCAA Men's Basketball Tournament is known for its unpredictability. Each year fans fill out a bracket, trying to correctly predict the outcome of each game in the tournament. This tradition spans many decades and is one of the most bet-on events in American sports. For that reason, we felt that it would be important to study this sport using statistics, as we can search for patterns that might lead to an advantage. In our project, we explore a decade of data taken on all NCAA D1 teams from 2013-2023. This data focuses on metrics such as free throw rate (FTR and FTRD), offensive and defensive turnover rate (TOR and TORD), offensive rebounding percentage (ORB), defensive rebounding percentage (DRB), effective field goal percentage (EFG_O and EFG_D), pace of play (AdjT), and several efficiency-based measures (BARTHAG, ADJOE, ADJDE, and WAB). Our goal is to investigate how these performance indicators relate to both regular season success, and how we can use it to predict postseason outcomes. Using statistical principles and machine learning techniques, we hope to uncover patterns and predictability in this sport.

Winning is the mark of success in basketball. We wondered if some key statistics can be used to predict win percentage. We asked, “Can we predict win percentage based on offensive and defensive free throw rate (FTR/FTRD), offensive three-point shooting percentage (3P_O), offensive and defensive turnovers (TOR/TORD), offensive and defensive rebounds (ORB/DRB), and luck (difference between predicted win percent based on efficiency and actual win percent)?” Win percentage serves as a fundamental measure of a team’s performance throughout the regular season. To answer this question, we applied regression-based modeling approaches. Metrics like free throw rate, three-point shooting percentage, and turnover rate reflect how well teams capitalize on possessions and minimize mistakes, while luck accounts for variance in expected wins based on efficiency. By quantifying the relationships between these specific features and winning, we aim to highlight the critical drivers of regular season success.

The main purpose of our analysis was to see if we could guide March Madness predictions. We asked, “Can we classify what postseason outcome a team will have (Early, Mid, Late, Final) based on their pace of play (AdjT), turnover rate (TOR), effective field goal percentages for both offense and defense (EFG_O, EFG_D), team seed (SEED), and several advanced efficiency metrics (BARTHAG, ADJOE, ADJDE, WAB)?” Tournament advancement is challenging to predict as the single-elimination structure elevates variability. However, classifying postseason performance into stages provides a broader and more interpretable view of success. We approach this task using classification modeling techniques. By incorporating a mix of pace of play, shooting efficiency, team seeding, and advanced overall performance measures, our analysis captures both playing style and underlying team quality. This provides a better understanding of what differentiates early tournament exits from deep postseason runs.

Previous studies, such as “Predicting Results of March Madness Using Three Different Methods,” have shown that using a wide range of team performance metrics alongside machine learning models such as Random Forests and Support Vector Machines can improve NCAA tournament prediction accuracy (Shen et al., 2016). However, their primary focus was on predicting individual game outcomes and constructing full tournament brackets. In contrast, our project investigates how specific regular-season and team-level performance metrics predict both a team's season-long win percentage and its categorical postseason outcome (Early, Mid, Late, and Final). By modeling broader seasonal success rather than individual games, our analysis offers a new perspective on the statistical factors that drive both regular and postseason performance. Our analysis can benefit sports analysts, fans, bettors, and anyone interested in understanding the true drivers of college basketball success.

Data

Our final dataset brings together information about NCAA Division I men’s basketball teams from 2013 to 2023, excluding 2020, which we removed because the season was cut short by COVID-19. We combined data from two different sources. The main dataset came from Kaggle, but the original data was collected by scraping team stats from BartTorvik.com. This dataset included season-long team performance metrics like shooting percentages, win-loss records, and tempo. We also scraped data from KenPom.com, which added advanced statistics like team rankings, adjusted offensive/defensive efficiencies, pace of play, and luck rating. After combining the two datasets by matching team name and year, we removed 35 teams that were missing data due to them either joining or disbanding between 2013 and 2023.

The final dataset contains 3,400 rows and 37 columns. Each row represents one team during one season (for example, “North Carolina 2015”). This gives us a full view of how each team performed in that year. Since our focus is on team-level performance, each observation is a sample of a team’s season, not an individual player. This allows us to look at trends and see which team qualities might lead to winning more games or going further in the NCAA tournament.

We didn’t use all 31 columns in our analysis. Instead, we focused on variables that are most useful for answering our research questions. For example, ADJOE (Adjusted Offensive Efficiency) shows how many points a team scores per 100 possessions, and ADJDE (Adjusted Defensive Efficiency) shows how many points they allow. EFG_O and EFG_D are shooting percentages that give extra weight to three-point shots. We also used FTR (Free Throw Rate), which tells us how often teams get to the free-throw line, and 3P_O (Three-Point Percentage), which shows how well teams shoot from beyond the arc. AdjT tells us how fast a team plays, measured in possessions per game. BARTHAG is a power rating that estimates how likely a team is to beat an average team. Finally, we included POSTSEASON to track how far a team made it in the tournament, and SEED to see their ranking going into March Madness.

To clean the data, we made sure all the column names and data types matched, especially for key variables like YEAR, TEAM, and SEED. We removed the 2020 season entirely because the tournament was canceled and the season was incomplete. These steps gave us a clean dataset without gaps or inconsistencies.

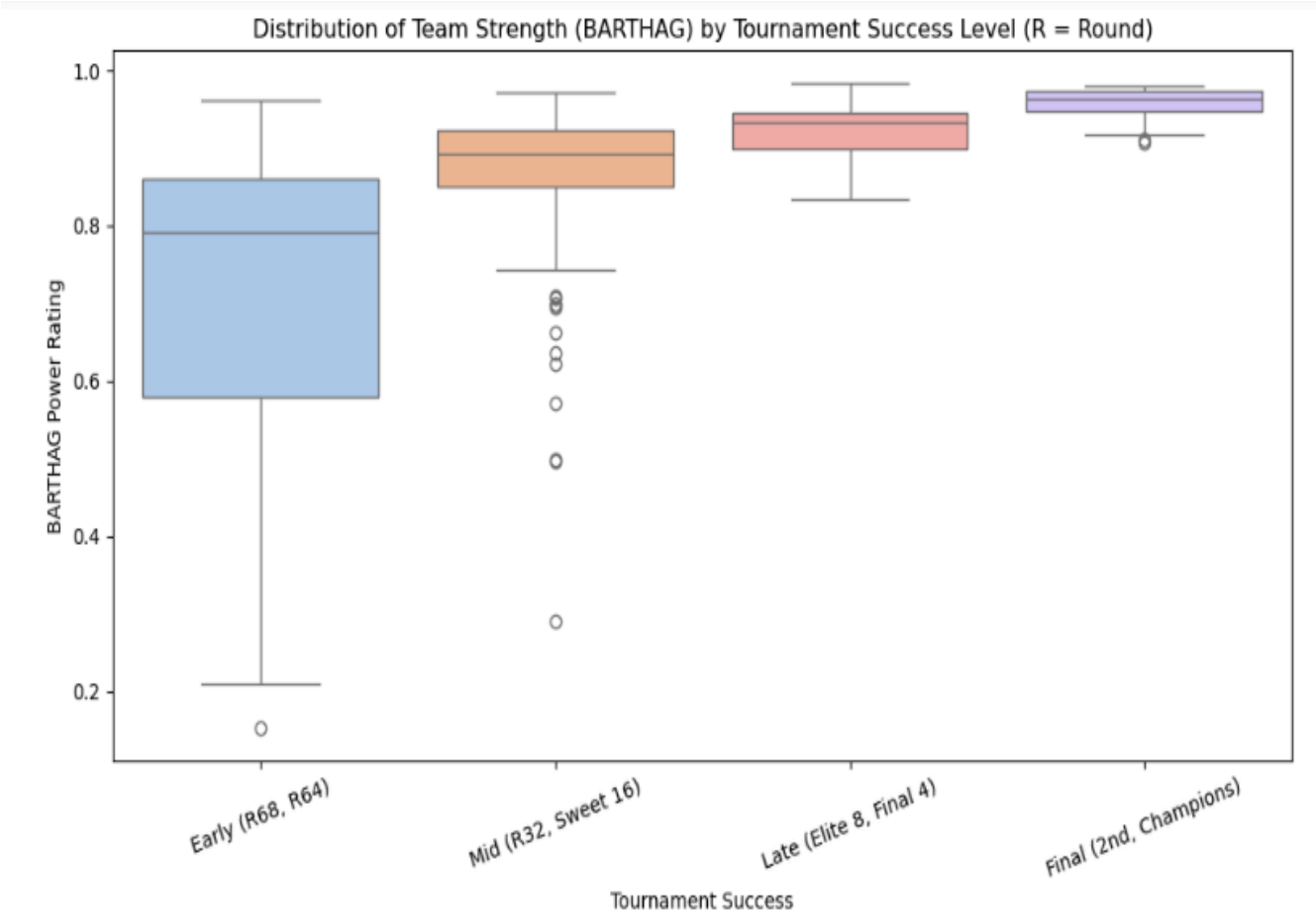
This dataset helps us explore important questions like: are offensive or defensive stats more important in March Madness? Does shooting better from the free-throw line or three-point

range make a bigger difference? Does a team's ranking or pace of play affect how far they go? The data is rich and detailed enough to help us answer these questions with real evidence.

We've also included a summary table that briefly explains the most important columns used in our analysis, such as ADJOE, ADJDE, BARTHAG, EFG_O, TOR, and others. These variables were key for answering our research questions about what drives team success in the NCAA tournament. In addition, we included a boxplot comparing BARTHAG (BartTorvik's team strength rating which compares team performance to the average team) across four levels of tournament success: Early (Round of 68, Round of 64), Mid (Round of 32, Sweet 16), Late (Elite 8, Final 4), and Final (2nd, Champions). The boxplot clearly shows that teams with higher BARTHAG scores tend to advance further in the tournament. After seeing this trend in the visualization, we conducted further analysis to examine which advanced metrics had the strongest correlations with tournament success, allowing us to better identify which team characteristics are most predictive of deeper postseason runs.

Variable	Description
TEAM	Team name
YEAR	Season year
Conf	Teams athletic conference
W-L	Win-Loss record
Luck	Difference between expected and actual wins
ADJOE	Adjusted offensive efficiency
ADJDE	Adjusted defensive efficiency
AdjT	Adjusted tempo – number of possessions per 40 minutes
EFG_O	Effective field goal percentage on offense (weights 3-point shots more heavily)
EFG_D	Effective field goal percentage allowed on defense
TOR	Offensive turnover rate (turnovers per 100 possessions)
TORD	Defensive turnover rate (turnovers forced per 100 possessions)

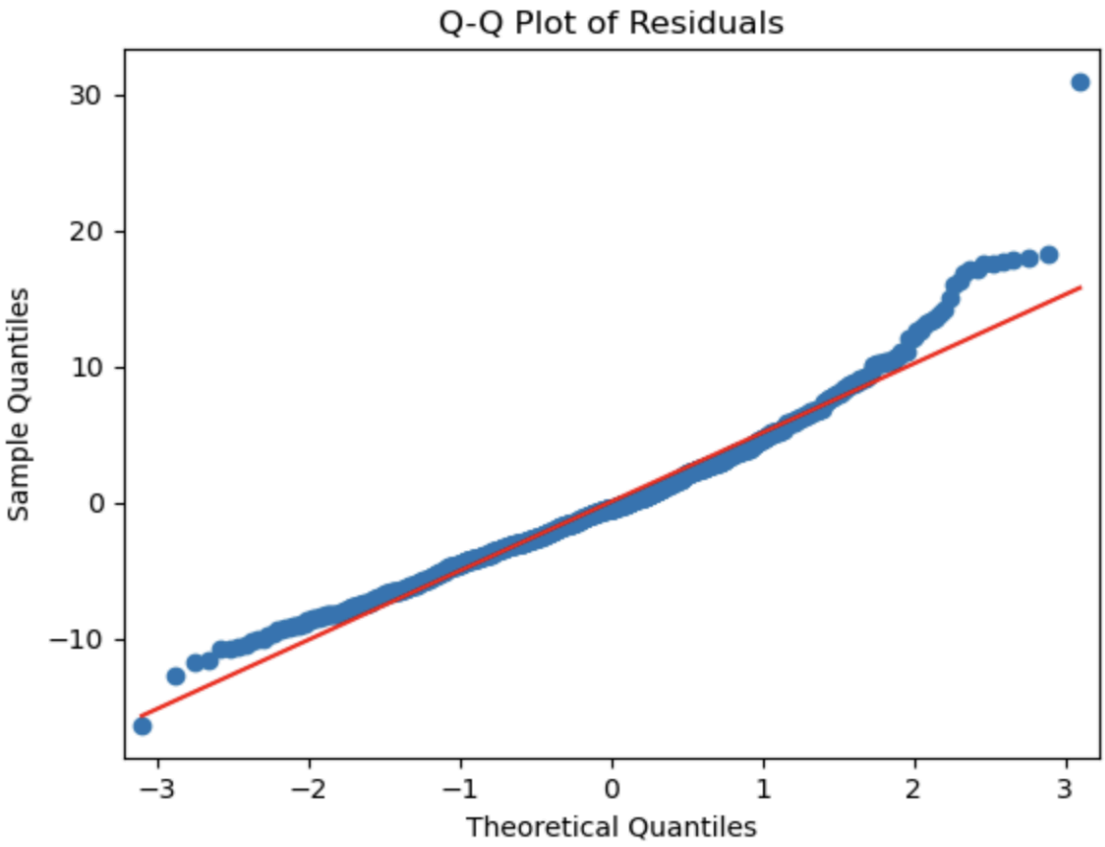
SEED	NCAA tournament seed assigned to the team
POSTSEASON	Furthest round reached in the NCAA tournament (e.g., R64, S16, Champ)
BARTHAG	How good a team plays against an average rated team
FTR	The ratio of free throw attempts to field goal attempts for a team
FTRD	The ratio of free-throw attempts allowed to field goal attempts faced
WAB	How many more wins a team has compared to an average “bubble team” (teams on the edge of NCAA Tournament selection)
3P_O	Offensive 3-point shooting percentage
ORB	Offensive rebound percentage
DRB	Defensive rebound percentage

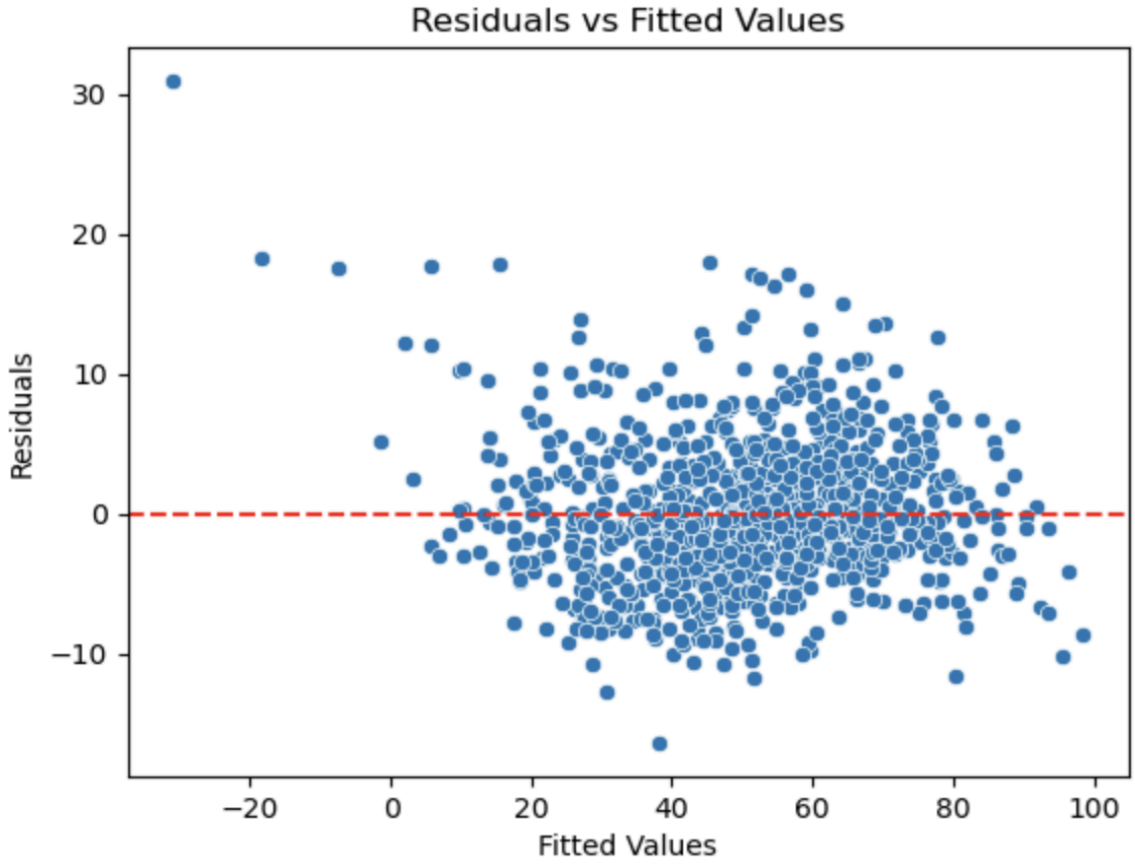


Results

Question One

We tried to answer: can we predict win percentage utilizing offensive and defensive free throw rate (FTR/FTRD), offensive three-point shooting percentage (3P_O), offensive and defensive turnovers (TOR/TORD), offensive and defensive rebounds (ORB/DRB), and luck (the deviation between expected win percent based off of efficiency and actual win percent)? Since we are attempting to predict a continuous variable, we implemented regression models. A regression model requires several assumptions to be met: normality, homoscedasticity, linearity, mean of residuals equal to zero, and error independence. Normality requires there to be a normal distribution of the residuals (how far off the predicted value is from the actual value), which we made sure to check through a Q-Q plot. We checked for homoscedasticity, mean zero residuals, and error independence through a residual plot, to ensure the residuals are varied evenly and the model is precisely capturing the data. Linearity was checked through linear plots on all feature variables to ensure they are all a predictor for the target variable win percent, which was created by dividing wins by games and multiplying by 100. Finally, our independent variables were selected based on VIF (variance inflation factor), a measurement of collinearity between multiple variables.


Q-Q Plot of Residuals
Sample Quantiles
Theoretical Quantiles
30
20
10
0
−10
−3
−2
−1
0
1
2
3


Residuals vs Fitted Values
Residuals
Fitted Values
30
20
10
0
−10
−20
0
20
40
60
80
100

There were three regressions we decided to test to see which one was the most accurate at predicting win percent. First, a baseline linear regression model was created, which utilized average win percent from training data to compare against testing data. This baseline model helped create a comparison that could be used against all testing and training data for the model, with an R squared (how well the independent variables account for the variability or deviations seen within the explanatory variable) of -.003, and a root mean squared error (how far off the predicted values are) of 18.25. These results create a good baseline to work with. This baseline is good because if our data has a lack of explanation for variability within the data and the predictions are close to 20 percentage points off, we know our model is not explaining more than what the average win percent would tell us. Thus, we tested the baseline against three regressions to find the best model: ridge regression, lasso regression, and linear regression.

For the ridge regression and lasso regression, we had to set up our model hyperparameters – inputs into the model that tell it how much penalty to give the variables. Through a net elastic grid search cross-validation, we found that the best model would be a lasso regression with an L1 of 1, and an alpha of .001. Then, we ran a lasso and ridge regression utilizing an alpha of .001 (essentially turning the ridge into a lasso), and got an R squared of .9221, and a root mean squared error of 5.086. We ran a linear regression model to further compare and got an R squared of .92206 and a root mean squared error of 5.085. The results were very similar, but it seems the lasso regression has a slightly better R squared and root mean squared error. Finally, we also derived a linear regression root mean squared error using training and testing data separately, to double-check for overfitting due to a high R squared score, and found that the root mean squared error was approximately the same for both – indicating it is not overfit.

From our results, we have found that we can indeed predict win percentage within five percentage points utilizing a lasso regression based on the variables: offensive and defensive free throw rate (FTR/FTRD), luck (the deviation between expected win percent based on efficiency and actual win percent), offensive three-point shooting percentage (3P_O), offensive and defensive turnovers (TOR/TORD), as well as offensive and defensive rebounds (ORB/DRB). The high R squared of .92 indicates that only about 8 percent of variability is not explained within our model, and our root mean squared error tells us our model consistently gets the win percent within 5 percentage points. These results are promising for the applicability of the model in the future. There are however, a few limitations with the model. One limitation, despite its longitudinal nature, is that the model does not have mid-season statistics to train on, it is all cumulative, so it would not be very good at predicting early on in the season, which in the context of win percent may not be as useful. Secondly, it utilizes luck, which is a statistic based on the deviation between win percentage and predicted win percentage based on efficiency. Removing luck shifts the R squared from .92 to .84, meaning it would still have strength utilizing other variables, but if you'd like to purely predict win percentage while being early into the season just based on performance, you may need to exclude this variable and reduce some of its predictive power. This may be a useful trade-off however in the context of win percent.

The results that we garnered from the lasso regression do not quite fit that of other literature on win percent. Aside from luck, which makes up a large portion of explanatory power, our model utilized field goals and turnovers the most in determining the win percentage. Offensive and defensive field goal lasso regression coefficients hovered around 2.6 and -2.6, whereas offensive and defensive turnovers hovered around 2.38 and -2.11 respectively. These held a higher absolute value than other coefficients, even when luck was removed from the

model. However, according to scientific literature, field goals as well as offensive and defensive rebounding accounted the most for a larger difference in win percentage, with each making up 13.6% and 14.2% of the explained variance (Dimitrije et al. 1). This means our results matched the findings – offensive and defensive field goals are one of the largest determinants of win percent, but differ in that we found turnovers to hold greater power than rebounding in determining win percent.

Question Two

The question we're trying to answer is: can we classify what postseason outcome a team will have based on different season statistics? To answer this, we used three models: a dummy baseline, decision trees, and a random forest classifier. Our predictors included pace of play (AdjT), turnover rate (TOR), effective field goal percentages for both offense and defense (EFG_O, EFG_D), team seed (SEED), and several advanced efficiency metrics explained above (BARTHAG, ADJOE, ADJDE, WAB). We started with a dummy classifier using the "most frequent" strategy, which always predicts the most common outcome. This gave us a baseline accuracy to compare the more complex models against.

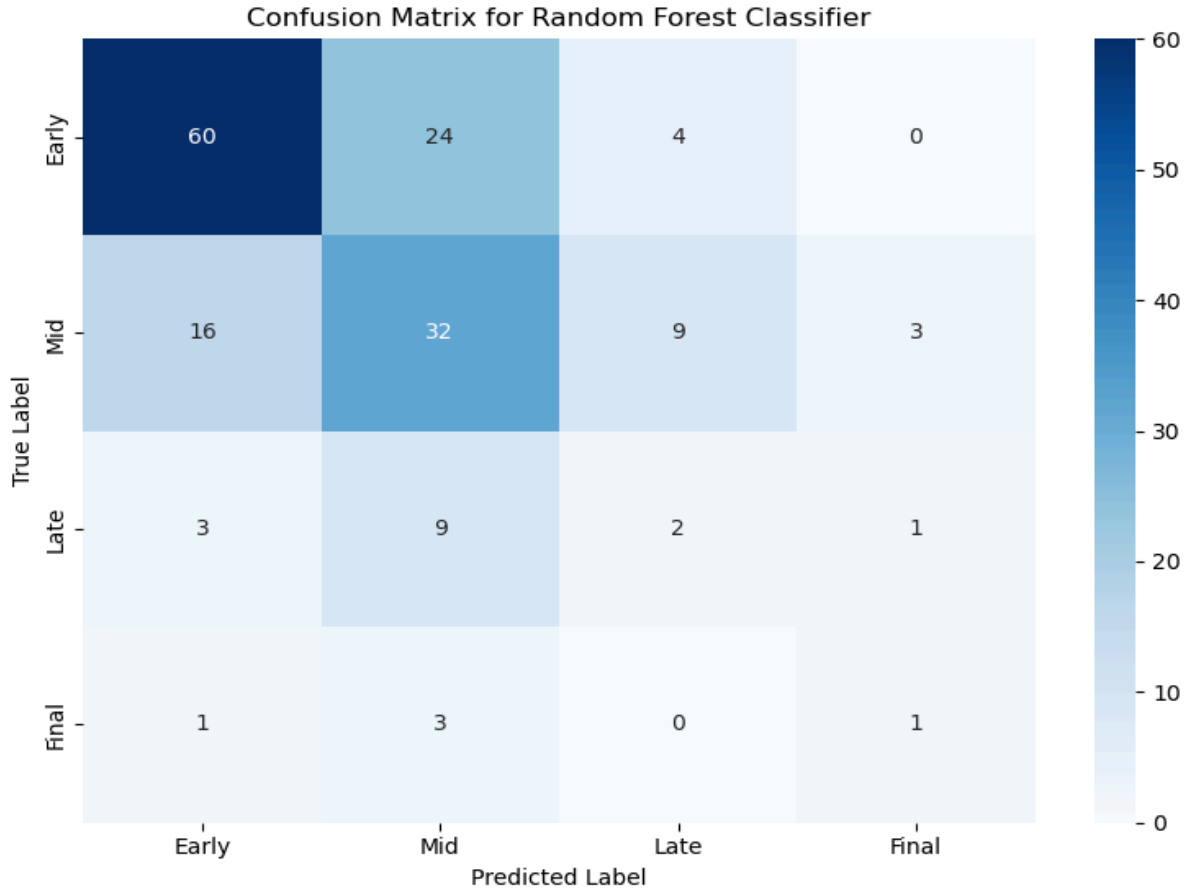
After that, we trained a decision tree classifier to see how well basic splits on these features could separate postseason outcomes. Then, we used a random forest model to improve performance by averaging across multiple trees, which helps reduce overfitting and capture more complex patterns in the data. We used a stratified train-test split to keep the class distribution balanced and added a balanced class weight argument to both models to help with class imbalance. To find the best setup for the random forest, we used grid search with 5-fold cross-validation to tune the number of trees and the depth of the trees. We initially tested a small range but later expanded it after seeing improvements with slightly deeper trees. The

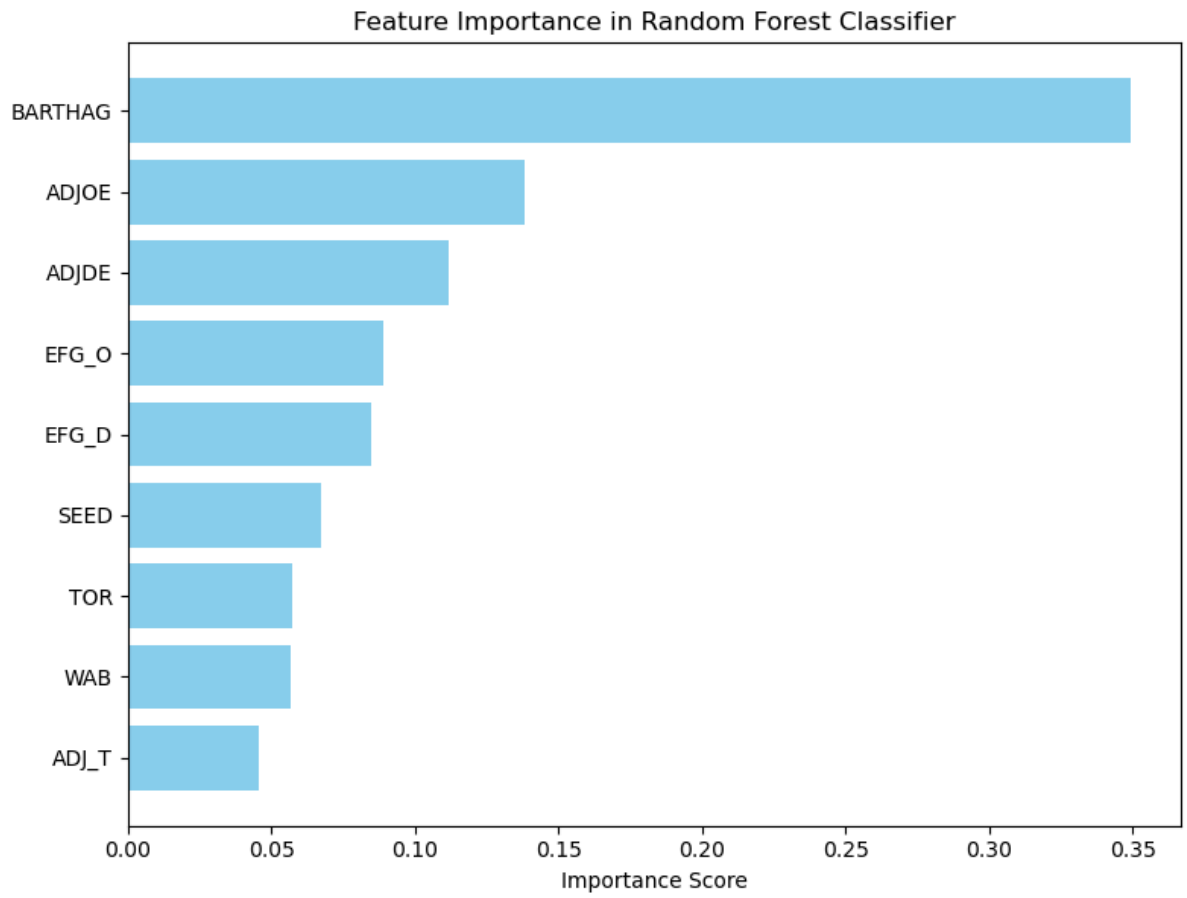
best-performing model used 100 trees with a maximum depth of 10, achieving the highest average accuracy across the folds. These steps helped ensure that the final model was both accurate and generalizable.

To evaluate model performance, we started with a dummy classifier using the most frequent strategy, which always predicts the most common class. This model achieved 52% accuracy, which is the proportion of teams in the training set that fall into the most common category — in this case, early round exits. The most common result constituting the entire accuracy score means the model failed to identify any teams in the final, late, or mid categories, with f1-scores of zero for all three. This baseline clearly highlighted the presence of class imbalance and showed the need for a more complex model that can better capture postseason variety.

The decision tree classifier improved on the dummy baseline, achieving 57% accuracy. It was able to identify more mid and early outcomes, but predictions for final and late were still very limited. The best performance came from the random forest model, which achieved 65% accuracy and showed better balance across all outcome categories. It started to correctly identify some teams in the final and late rounds, and its overall macro F1-score and weighted metrics reflected improved generalization. According to the feature importance rankings generated by the random forest, BARTHAG (how well a team does against an average team) was the most influential predictor of postseason success, suggesting that a team's relationship to the average, alongside win probability, play a major role in determining how far they go in the tournament. While none of the models perfectly captured the less frequent outcomes, the random forest clearly performed best and showed the most potential for real-world application. For example, albeit only having an accuracy score of 50% for final rounds, a decision tree that can

predict the final round teams 50 percent of the time is promising in real-world contexts. Thus, despite it not always being accurate, the model holds some predictive power in determining postseason outcomes based on the aforementioned metrics, which can be incredibly useful for NCAA postseason basketball.





Conclusion

Our project aimed to answer two key questions: (1) Can we predict a team's win percentage based on performance metrics like free throw rate, turnovers, rebounds, three-point shooting, and luck? (2) Can we classify a team's postseason outcome (Early, Mid, Late, Final) based on pace of play, shooting efficiency, team seed, and advanced efficiency metrics? For the first question, we found that a lasso regression model best predicted win percentage, achieving an R^2 of 0.9221 and an RMSE of approximately 5 percentage points. Offensive and defensive effective field goal percentage and turnover rates emerged as the most critical predictors. For the second question, a random forest classifier outperformed both the dummy and decision tree models, achieving 65% accuracy and identifying BARTHAG as the most influential feature for

postseason success. These findings confirm the strong predictive power of efficiency metrics seen in prior studies but challenge some past literature that emphasized rebounding as a more dominant factor in regular-season success.

Our results suggest that win percentage can be reliably predicted using a small set of season statistics and that postseason outcomes, though harder to classify due to variability, can still be meaningfully predicted with ensemble methods like random forests. In a broader context, this demonstrates how statistical modeling can cut through the "madness" of college basketball, offering real insights into team quality and tournament potential. One possible future extension would be to explore game-level data instead of season averages to predict individual tournament games or to update models dynamically over the course of a season. Limitations of our work include reliance on end-of-season statistics and the influence of the "Luck" metric, which reduces early-season predictive utility.

Future studies could address these limitations by incorporating player-level data, injury reports, or in-season trends, making models more adaptable and precise. Additionally, exploring other machine learning techniques, such as deep learning, could improve predictive performance. Our project adds new insight by emphasizing the predictive strength of BARTHAG and team efficiency ratings for broader postseason success categories, complementing past research focused solely on individual game outcomes.

Works Cited

Cabarkapa, Dimitrije et al. “Game statistics that discriminate winning and losing at the NBA level of basketball competition.” PloS one vol. 17,8 e0273427. 19 Aug. 2022, doi:10.1371/journal.pone.0273427

Shen, Gang, et al. "Predicting Results of March Madness Using Three Different Methods." *Journal of Sports Research*, vol. 3, no. 1, Jan. 2016, pp. 10-17. doi:10.18488/journal.90/2016.3.1/90.1.10.17.