College Basketball Postseason Predictor (2013-2023)

By Andrew Wolfe

September 2024

Abstract

Machine Learning is one of the most important concepts of artificial intelligence. It can be used to make predictions and decisions based on data using advanced algorithms. This project report presents a prediction model for how far college basketball teams playing in the National Collegiate Athletic Association (NCAA) will advance in the postseason tournament. Keeping the objective of predicting postseason success along with discovering the most important variables for this prediction will be the main purpose of this report. Many different models, such as Random Forest, Gradient Boosting, Logistic Regression, and Support Vector Machines were tested along with a feature importance test to find the most important variables. The predicted variable is called 'POSTSEASON' and I found that a Logistic Regression model was the most accurate in predicting this outcome. The variables, "ADJDE", "ADJOE", "WAB", "ORB", and "ThreePO" (3P_O) were found to have the most importance for determining how far a team will advance in postseason play. These findings will provide valuable insights that can lay the groundwork for future research in the field of predicting postseason performance and determining the most important statistics for this outcome.

Introduction

For this Machine Learning Project, I chose the dataset cbb.csv from kaggle. This dataset has various metrics of how college basketball teams performed and many variables including their record (wins vs losses), field goal percentage, rebounding rates, opponents shooting percentages, seeding in the NCAA tournament, and tournament finish. I intend to fit multiple models, such as logistic regression, random forest, gradient boosting, and a support vector machine, and choose the best model based on accuracy for predicting postseason performance.

I want to discover the most important variables for determining how far a team will advance in postseason play and discuss which models work well in this prediction. I plan on keeping the following variables below for my model. I will explain some exploratory data analysis of the dataset, including boxplots and barplots, walk through the production of the models, choose the best-performing model, and then finally explain the use of it in the future.

- **"ADJOE"**: Adjusted Offensive Efficiency (An estimate of the offensive efficiency (points scored per 100 possessions) a team would have against the average Division I defense)
- **"ADJDE":** Adjusted Defensive Efficiency (An estimate of the defensive efficiency (points allowed per 100 possessions) a team would have against the average Division I offense)
- **"BARTHAG":** Power Rating (Chance of beating an average Division I team)
- "**EFG_O"**: Effective Field Goal Percentage Shot
- **"EFG_D":** Effective Field Goal Percentage Allowed
- **"TOR"**: Turnover Percentage Allowed (Turnover Rate)
- **"TORD"**: Turnover Percentage Committed (Steal Rate)
- **"ORB"**: Offensive Rebound Rate
- **"DRB"**: Offensive Rebound Rate Allowed
- **"FTR"**: Free Throw Rate (How often the given team shoots Free Throws)
- **"FTRD"**: Free Throw Rate Allowed
- **"2P_0":** Two-Point Shooting Percentage
- **"2P_D":** Two-Point Shooting Percentage Allowed
- **"3P_O":** Three-Point Shooting Percentage
- **"3P_D":** Three-Point Shooting Percentage Allowed

- **"ADJ_T":** Adjusted Tempo (An estimate of the tempo (possessions per 40 minutes) a team would have against the team that wants to play at an average Division I tempo)
- **"WAB":** Wins Above Bubble (The bubble refers to the cut off between making the NCAA March Madness Tournament and not making it)
- **"SEED":** Seed in the NCAA March Madness Tournament
- **"POSTSEASON"**: Round where the given team was eliminated or where their season ended (R64 = Round of 64 or 68, R32 = Round of 32, S16 = Sweet Sixteen, E8 = Elite Eight, F4 = Final Four, 2ND = Runner-up, Champion = Winner of the NCAA March Madness Tournament for that given year)

I chose to use this dataset because I wanted to see if there is a way to better predict postseason performance and help coaches in the off-season determine which statistics are the most important when evaluating players in the transfer portal and in high school. As I do more research and create models, I will determine which stats have the highest influence on the postseason advancement.

EDA

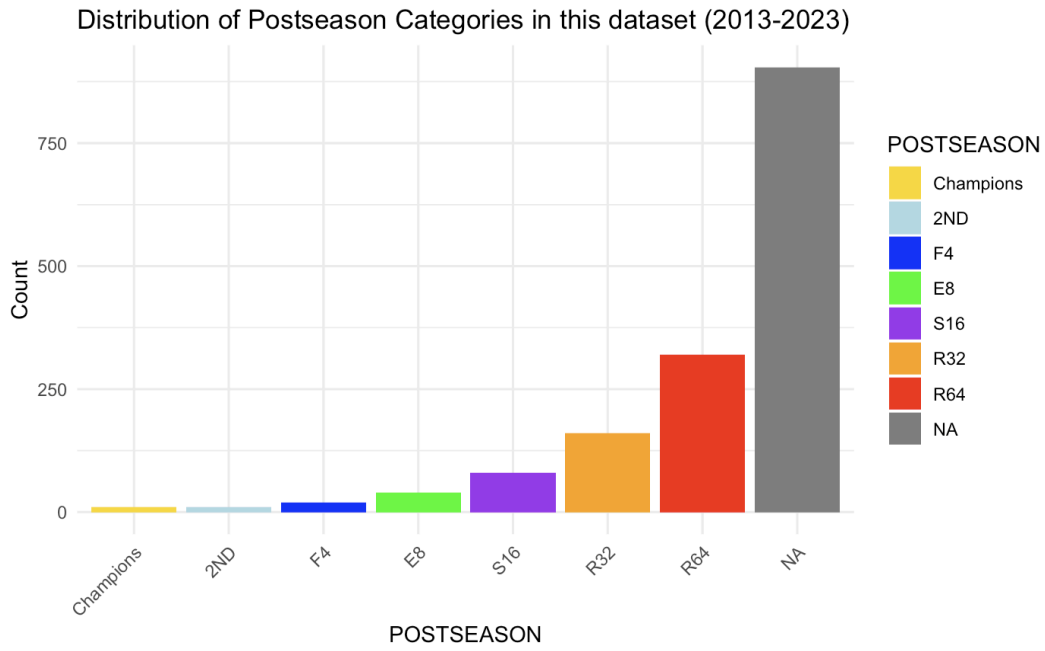


Figure 1: *Distribution of Postseason Categories* (2013-2023)

The barplot in **Figure 1** of the postseason categories in our dataset reveals that all "N/A" values must be removed. This takes up the most space in our dataset and because I want to only examine teams

that qualified for the NCAA Tournament, it must be removed. All of the other values, “R64”, “R32”, “S16”, “E8”, “F4”, “2ND”, and “Champions” follow in a descending order, which is correct as fewer teams advance to each round as the NCAA Tournament progresses.

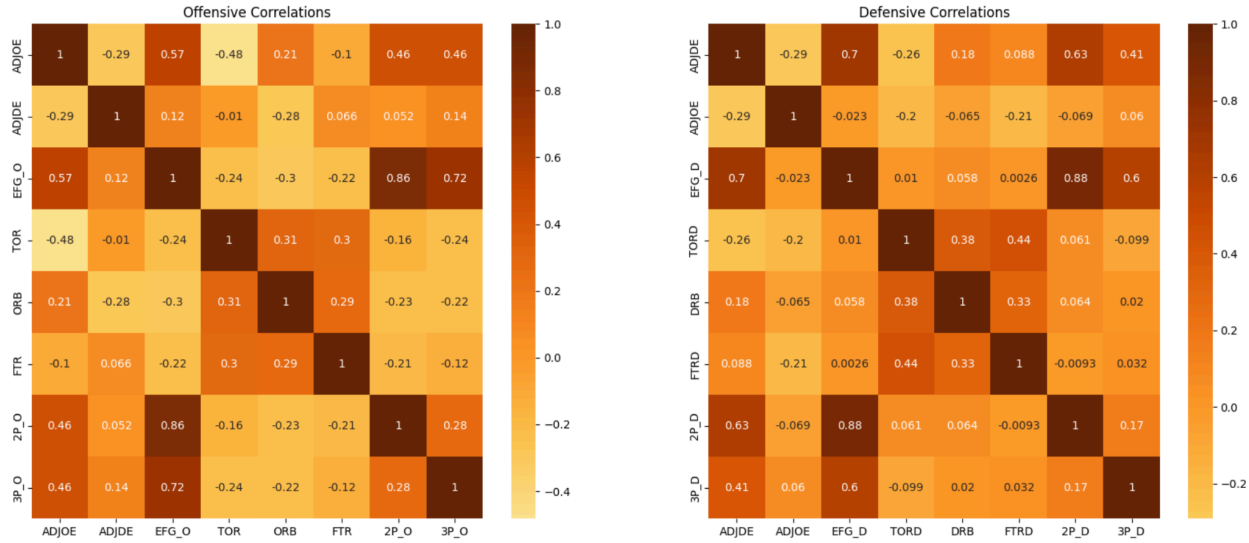


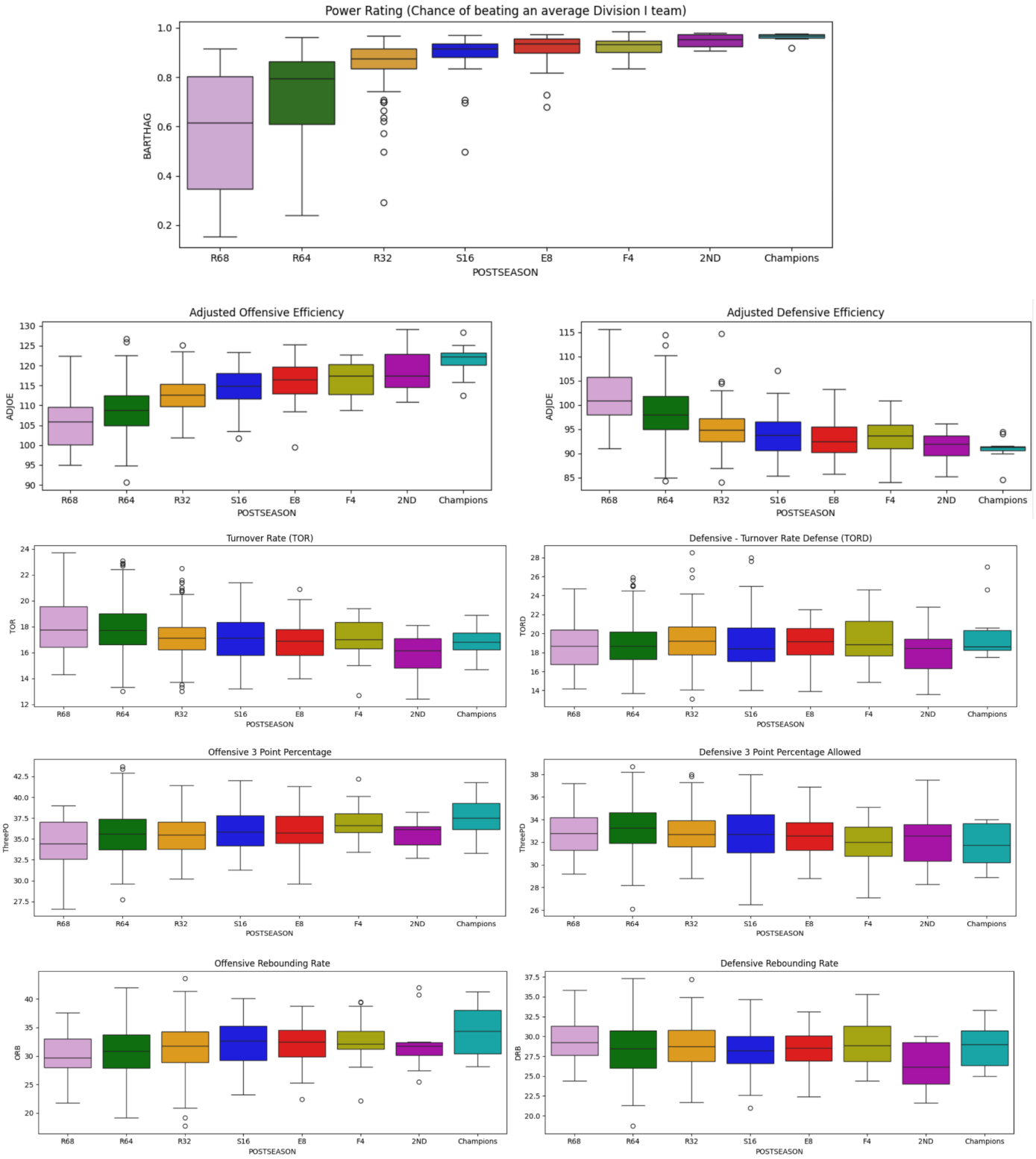
Figure 2: *Heatmap of Gameday Variables*

From the correlation matrix in **Figure 2**, we can identify some initial relationships in our dataset that appear to exist. These appear to be:

- “ADJOE” and “EFG_O”
- “EFG_O” and “2P_O”/ “3P_O”
- “TOR” and “ADJOE”
- “ADJDE” and “2P_D”
- “TORD” and “FTRD”

Initially, these variables look like they will play a significant role in my model of predicting the postseason performance of a team. More exploratory data analysis is required to get a deeper understanding of our data. Because of this, I created boxplots of the following variables, grouped by postseason finished: “BARTHAG”, “ADJOE”, “ADJDE”, “TOR”, “TORD”, “ORB”, “DRB”, “3P_O”, and“3P_D”.

Figure 3: *Boxplots Grouped by Postseason Finish*



By looking at **Figure 3**, I see that the Adjusted Offensive Efficiency and Defensively Ratings were much higher for teams that advanced further in the NCAA Tournament. Moreover, as expected the Barthag Rating and 3PT Percentage were much higher for teams that advanced further in the NCAA Tournament as well. Also, Turnover Rate appears to be a small percentage lower for teams that advanced further in the NCAA Tournament. There doesn't seem to be a large difference in the defensive rebounding rate for postseason advancement, but it seems teams that had a higher offensive rebounding rate advanced further in the NCAA Tournament. However, it is important to remember that correlation does not mean causation. There must be further testing and analysis to discover the true relationships between these variables. This is what I will be doing in the remainder of the report in the next sections.

Modeling Selection

Because I am predicting the 'POSTSEASON' variables, a categorical variable that can have the values: 'Champions', '2nd', 'F4', 'E8', 'S16', 'R32', 'R64', I decided to use logistic regression, support vector machines, gradient boosting, and random forest models. I ran each model and cross-validated it to determine the model with the highest accuracy score. It is good to keep in mind that if someone was making a random guest about predicting the postseason finish for a team, he or she would have a 1 in 7 chance (14%) of making the correct choice. With this in mind, I strived for an accuracy score that is significantly above this number and close to 50%. Listed below you will see each model that was used and a description of the model as well, which will highlight the differences between them.

- ***Logistic Regression***: Statistical method used for binary classification and estimating probability for a given input belonging to a specific category.
- ***Gradient Boosting***: Ensemble technique that uses decision trees in a sequential matter to make decisions. These models learn from its mistakes from earlier trees and improve on them to make predictions.
- ***Random Forest*:** Ensemble technique that uses previous decision trees predictions and averages them out to make its own prediction. This method is great for reducing the effects of over fittings.
- ***Support Vector Machines*:** Machine learning model technique that finds the best margin or boundary between the closest points of different classes to make predictions. This model will find the best margin or boundary for the different classes that are produced.

Below is a table of the accuracy and cross-validation accuracy score for each of the models above:

Model	Accuracy Score	Cross-Validation Accuracy Score
Logistic Regression	**0.523**	**0.533**
Gradient Boosting	0.453	0.467
Random Forest	0.492	0.490
Support Vector Machines	0.484	0.504

Table 1: Accuracy Scores of Each Model

In Table 1, we see that the Logistic Regression produced our highest accuracy score. The Logistic Regression Model produced a 0.523 accuracy score and when using cross-validation, a 0.533 accuracy score for predicting the 'POSTSEASON' results of a specific college basketball team.

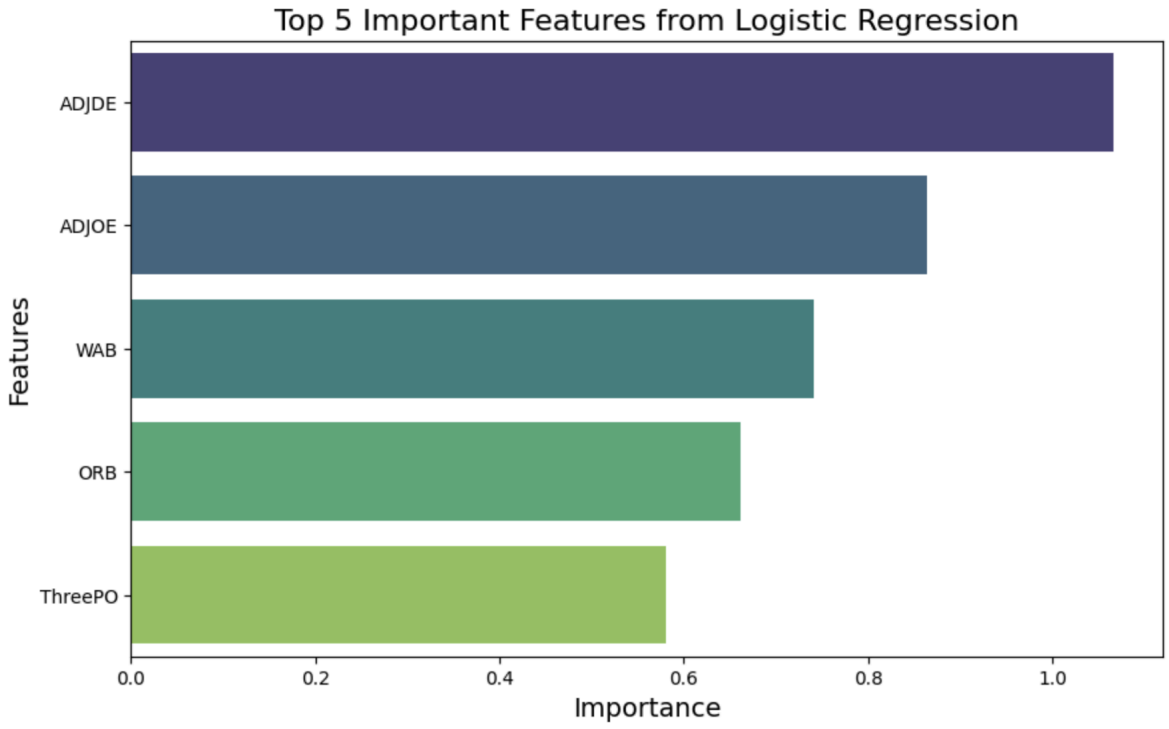
Modeling

When beginning the model, it was necessary to define the following variables as the features: "ADJOE", "ADJDE", "BARTHAG", "EFG_O", "EFG_D", "TOR", "TORD", "ORB", "DRB", "FTR", "FTRD", "2P_O", "2P_D", "3P_O", "3P_D", "ADJ_T", "WAB", AND "SEED". The target variable was: "POSTSEASON". Moreover, missing values were handled by an imputer using the mean and the SimpleImputer Function. Then, the features needed to be normalized to ensure that the features were scaled to ensure equal weighting and improved convergence. All rows with NA values in the 'POSTSEASON' label were removed (i.e. teams who did not make the NCAA Tournament were not included in my model).

In addition to the accuracy score, I also used the 'best_estimator_' function to calculate the five most important features in my model. Feature Importance is crucial because it determines the individual significance of each variable and can reveal to fans and coaches what statistics are the most important for a deep run in the Men's NCAA Tournament.

Looking at **Figure 4**, we see that the variables "ADJDE", "ADJOE", "WAB", "ORB", and "ThreePO" (3P_O) were the five most important variables in our model for predicting postseason performance. This put a heavy importance on defense, shooting, and rebounding the ball.

Figure 4: Top 5 Important Features from Logistic Regression Model



Modeling Assumptions

I did need to check for model assumptions in my machine learning model. These were different for each of the four models:

- ***Logistic Regression Assumptions***:
    - *Linearity of Log-Odds*: Relationship between independent variables and log-odds of the outcome is linear.
    - *Independence of Errors*
    - *No multicollinearity*: Independent Variables are not highly correlated
- ***Gradient Boosting Assumptions***:
    - Independence of Observations
- ***Random Forest Assumptions:***
    - Independence of Observations
- ***Support Vector Machines Assumptions:***
    - Independence of Observations
    - Linear Separable Data: Data needs to be linearly separable

Figure 5: Support Vector Machines Assumptions of Independence and Linear Separability

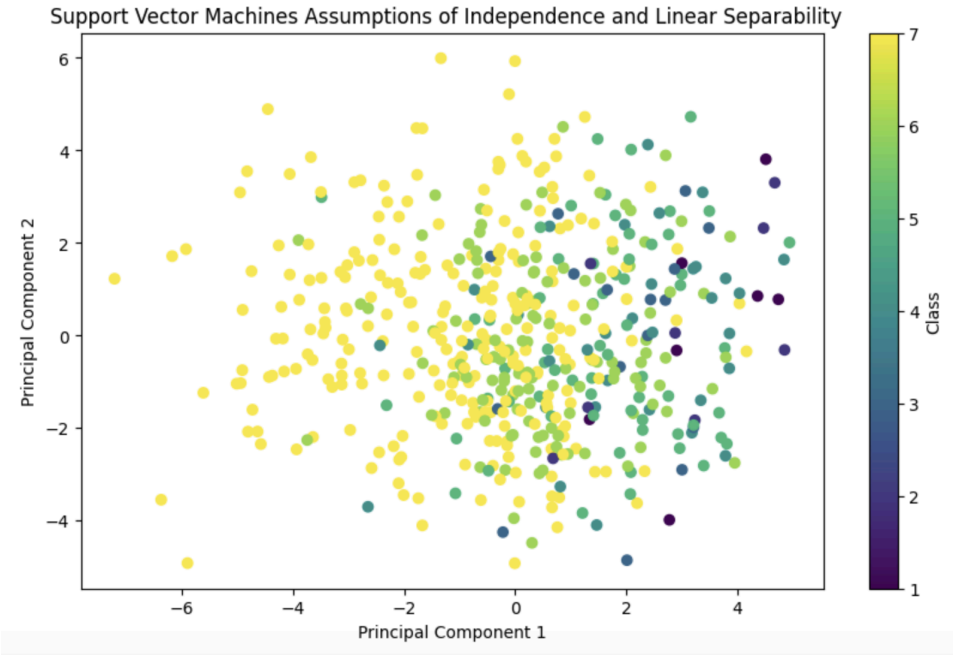
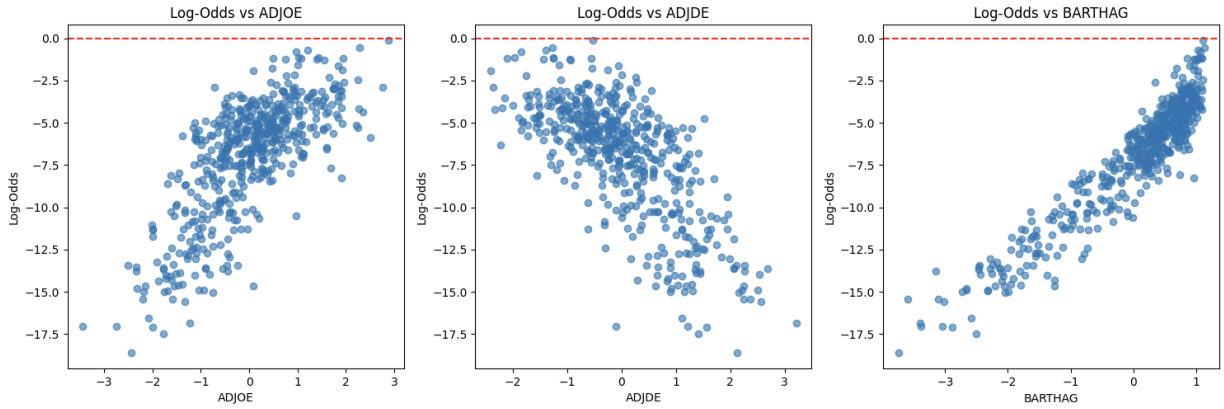


Figure 6: Example of Linearity of Log-Odds Test for Logistic Regression (3 Variables)



By looking at **Figure 5** and **Figure 6** above, we see that all of the assumptions are met.

Conclusions

In conclusion, I predicted how far college basketball teams would advance in the NCAA Tournament using data from 2013-2023. I used logistic regression, random forest, support vector machines, and gradient boosting models and after all of the modeling was finished, the logistic regression model produced the highest accuracy score. Model assumptions, such as no multicollinearity, linearity of log odds, independence of observations, and linear separability, were tested and met. The highest accuracy score I received after cross-validating was 0.533. Initially, this score may seem a little low, but when seeing that the random chance of predicting the correct postseason finish for a college basketball team is 1 in 7 (14% chance), one notices that the accuracy I received is quite high. The purpose of this model was to help college basketball fans, bettors, and coaches know what statistics are the most important and contribute the most to a team's championship chances. If this is known, it can change the complexion of betting odds, rooting interest, lineup analysis, and roster management (including the transfer portal and high school recruiting).

In the future, something that I could look into would be calculating the p-value scores for the model performances to enhance and understand the significance of each model more than I do at the moment. I also could look into including teams in a model that did not make the postseason and see if I could create a model that could predict this.

I am very much looking forward to exploring these opportunities in the future and expanding on my machine learning skills. I hope that this report can be a starting point for college basketball fans, bettors, and coaches to determine what statistics to value and emphasize more for their team to win the ultimate prize in college basketball.