College Basketball Game Prediction

By: Sierra Sikorski

Summary

Schools are interested in the most important factors to winning a college basketball game. I found that the three keys were aggressive defense, not allowing the other team to get easy shots, and offensive efficiency.

Research Objective:

With college basketball season underway, fans nationwide are starting to think about one of the most coveted sports events: March Madness. As a college basketball fan, I have always been fascinated with using data to predict sports outcomes. The research objective of this project is to predict game winners of an individual game in a college basketball season. Though the analysis focuses on sports outcomes, there are severe implications in higher education policy. There is a phenomenon called the "Flutie" effect, named after Doug Flutie, who led Boston College to a win over the University of Miami by throwing a Hail Mary pass in the final seconds of their game (Romboy, 2023). After this, applications to Boston College rose by 30%. A researcher later found that similar results could only be achieved by decreasing tuition or recruiting higher-quality faculty (Chung, 2013). Another effect was that campus diversity increased due to a larger applicant pool (Mayes & Gaimbalvo, 2018). Currently, schools are looking to find new ways to encourage diversity after the reversal of Affirmative Action, and an unlikely solution is focusing on winning athletics.

Data:

I use Google's 2022 March Madness Competition data (Sonas et al., 2023). Since Kaggle hosts this competition, that is where I downloaded the data. There were many different datasets, but I focused on the BoxScore data called MRegularSeasonDetailedResults.csv. The data initially contained 34 variables and over 100,000 observations. Each observation was a single game, and the data went back to the 2002-2003 season. I decided to filter the data only to use the data from the last complete season, 2022. I did this because, over the years, the sport has progressed, and the dynamic has changed, so using data from earlier seasons to train a model might not be beneficial. After doing this, 5,345 observations remained.

The initial variables included for both teams were the standard box score variables: points, rebounds, assists, turnovers, blocks, field goals attempted/made, three points attempted/made, free throws attempted/made, turnovers, and personal fouls. They also included the ID for the winning and losing team, the year, days into the season, if the win was home/away/neutral location, and if the game went into overtime. These variables provided a fantastic starting point to calculate a few more variables that may be important in the outcome of a game: number of possessions, points per possession, offensive and defensive efficiency, and true shooting percentage. To calculate the number of possessions, I utilized Dean Oliver's method (Cappe, 2020), which is:

$$possessions = FGA + .44 * FTA - 1.07 \frac{OR}{OR + ODR} * (FGA - FGM) + TO$$

Where FGA is field goals attempted, FGM is field goals made, FTA is free throws attempted, and TO is turnovers. After this, I used possessions to calculate offensive efficiency by dividing points scored by possessions and multiplying that by 100. Defensive efficiency is the other team's offensive efficiency since it shows how many points a team gives up per 100 possessions. Points per possession is also straightforward forward, being points scored over possession. Finally, the true shooting percentage is calculated by:

$$true\ shooting = \frac{points}{(2 * (FGA + .44 * FTA))}$$

Where FGA is field goals attempted, and FTA is free throws attempted. After calculating these, I looked at the summary statistics that I felt were the most important in winning games. I split the results by winning teams and losing teams. The summary statistics for winning teams are listed below.

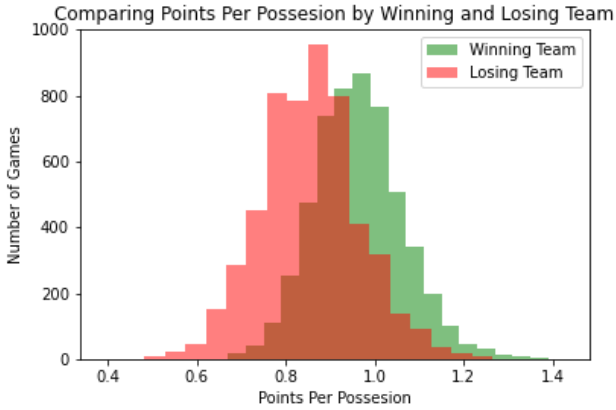
Winning teams	POINTS PER POSSESSION	DEFENSIVE EFFICIENCY	TRUE SHOOTING PERCENTAGE	DEFENSIVE REBOUNDS	STEALS
COUNT	5,344	5,344	5,345	5,345	5,345
MEAN	0.95	85.36	58%	25.36	6.82
STANDARD DEVIATION	0.10	11.17	7%	4.73	3.01
MINIMUM	0.63	39.24	37%	11	0
25%	0.89	77.95	53%	22	5
50%	0.95	85.06	58%	25	6
75%	1.02	92.41	63%	28	9
MAXIMUM	1.43	131.24	84%	44	19

Here, we see only one missing value for points per possession and defensive efficiency. There is a significant difference in the minimum and maximum value for each variable, which could be when teams win against an easy team compared to a more challenging team. The summary statistics for the losing team are shown below.

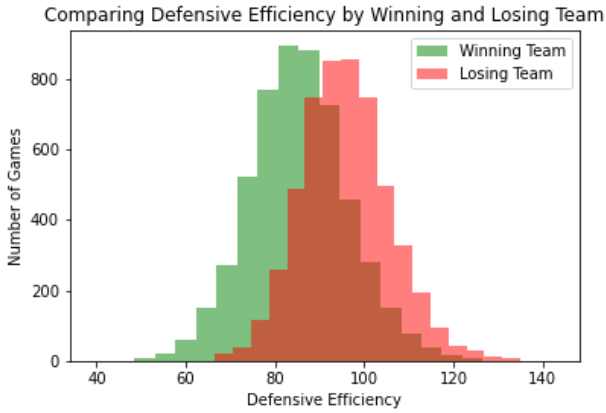
Losing Team	POINTS PER POSSESSION	DEFENSIVE EFFICIENCY	TRUE SHOOTING PERCENTAGE	DEFENSIVE REBOUNDS	STEALS
COUNT	5,344	5,344	5,345	5,345	5,345
MEAN	0.85	95.88	50%	21.46	6.03
STANDARD DEVIATION	0.11	10.03	0.07	4.38	2.69
MINIMUM	0.39	62.59	24%	6	0
25%	0.78	89.07	45%	18	4
50%	0.85	95.38	50%	21	6
75%	0.93	102.10	54%	24	8
MAXIMUM	1.31	143.05	77%	44	19

Similar to the statistics for the winning team, here we also see a significant difference in minimum and maximum values. Interestingly, there is only a .1 difference between the average points per possession of the winning and losing team.

Looking into specific variables more, I wanted to compare the winning and losing teams. I looked at the offensive measure of points per possession and the defensive efficiency. Points per possession are shown below.



As expected, we can see that both shapes are normal, and the average for the winning team is higher than the average for the losing team. The same is shown below for defensive efficiency.



Here, we can see that the winning team holds the losing team to fewer points per possession. Both of these results are in line with expectations.

One huge limitation of this dataset pertains specifically to predicting March Madness outcomes. Though this project aims to predict the winner of a college basketball game, the dataset does contain data from the March Madness tournament. Since March Madness games are played every other day with little breaks, when a team makes a deep run, they must have a solid bench to rest their starters while staying active in the game. Bench points are an essential metric unique to March Madness, but no variable in the data captures that.

Techniques:

I use two techniques on the game data. First, I will take the team's average in each variable up to the game they are about to play. The team's averages are a good indicator of their skill level. This way, each row (game match-up) is converted from that game's data to the averages of the team's games leading up to the current one. For the second technique, I will use a format similar to the averages, but instead of the averages of all games, it will be the past 10. Sometimes, teams get momentum from a win and start performing well, so looking at the ten most recent games is an attempt to catch the effects of their recent performances.

For my techniques, I used principal component analysis (PCA) for dimension reduction and then compared the logistic regression and random forest. PCA reduces the dimensions of a dataset and is useful when there are many variables. PCA is a good option for this project because many variables could be used in the analysis. Many variables are correlated like points scored and points scored per possession. PCA will help determine which variables are the best in explaining the variation in the data so I can effectively reduce the dimension.

Moving on to the models, I used logistic regression and random forest models. Logistic regression uses maximum likelihood to predict a binary outcome. I picked logistic regression because it is a simpler model to predict outcomes. It is a good starting point since I predict if a team wins. After this, I want to compare the results with the results from the random forest. Random forest is a method that combines the output of multiple decision trees for the model. I am using random forest since the dataset is large, and there is a risk for overfitting, so random forest eliminates that by using multiple trees to make final predictions.

I ran each model four times. I will run it with the PCA data frame for the full average and ten-game window. I also run it with the average and 10-game window datasets. Once I run the models, I compare each model's top 5 most important features. The extraction of the most important features is straightforward for the models without PCA, but I have to start with getting the most important eigenvectors for the models with PCA. Then, I look at the loadings for the important eigenvectors to determine the variables most likely to impact the model.

Findings:

Firstly, the full average models performed better than the 10-game window models, and within that, the logistic model performed better than the random forest model. A summary is shown in the table below.

Model	Score
Average Logistic	0.693
Average Logistic with PCA	0.687
Average Random Forest	0.662
Average Random Forest with PCA	0.661
10 Game Window Logistic	0.686
10 Game Window Logistic with PCA	0.652
10 Game Window Random Forest	0.629
10 Game Window Random Forest with PCA	0.625

The models using the average data performing better than the 10-game window models make sense since there are likely wins and losses outside the 10-game window that show the team's potential. Additionally, the PCA models usually were within .05 of the non-PCA models' accuracy score. Since the

full team average was consistently the best, I used those four models to compare the most important variables for each model. I also included the top variables for the PCA models since the accuracy scores were close. The results are shown below.

	First Variable	Second Variable	Third Variable	Fourth Variable	Fifth Variable
Logistic	T2 Three points made	T2 Field goals attempted	T2 Turnovers	T2 Free throws attempted	T1 Blocks
Logistic with PCA	T1 Score	T1 Field goals made	T2 Personal Fouls	T2 Three points made	T1 Offensive Rebounds
Random Forest	T1 Defensive Efficiency	T1 Offensive Efficiency	T2 Offensive Efficiency	T2 Defensive Efficiency	T2 Score
Random Forest with PCA	T1 Score	T1 Field goals made	T1 Offensive Efficiency	T2 Three points made	T2 Free throws made

Since the model predicts if team one wins, team two represents the opposing team. Out of the four models, the opposing teams' three points made is a top-five variable in three models. A reason for this could be that as a team makes more three-point shots, they will also likely miss a decent amount, which the other team can capitalize on. Another variable that occurs more frequently is team one's score. This concept is intuitive since a team is more likely to win the more points they score. One surprising variable is team one's defensive efficiency. Though it is the most influential variable in the random forest model, it does not appear in any of the other models. There is a saying that "defense wins championships," so the variable explicitly capturing that not being included in most models was surprising. The final surprising aspect was the model's reliance on what the opposing team was doing instead of team one. The most important variables in the models without using PCA tended to be the opposing team's metrics. However, in the models with PCA, they were a majority of team one metrics. This lack of opposing team metrics leads me to believe that some dimension reduction excluded the opposing team's variables when using PCA. Overall, the trend to win games seems to be offensive efficiency (as seen in the chart), aggressive defense (hence the opposing team's free throws), and not allowing the other team to get easy shots (hence the opposing team's three points made). Based on this, teams must prioritize defense to win and capitalize on the opposing team's mistakes, like turnovers. When doing this, teams will win games and shock people by making deep runs in March Madness. This deep run will bring more attention to the schools and increase applications, increasing campus diversity.

One important thing to note is that in March Madness, teams rely on their bench significantly more than in the regular season due to playing back-to-back games. In this data, no data captures that despite the data including tournament games. A future consideration could be to scrape similar data from ESPN's website to get box office scores and individual player statistics. This way, we can calculate bench points for each team so the model has better predictive power for March Madness.

References:

Cappe. (2020, September 3). *Learn a stat: Possessions and pace*. Hack a Stat. https://hackastat.eu/en/learn-a-stat-possessions-and-pace/

Chung, D. (2013). The dynamic advertising effect of collegiate athletics. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2345220

Mayes, B. R., & Giambalvo, E. (2018, December 6). *Does sports glory create a spike in college applications? it's not a slam dunk.* The Washington Post. https://www.washingtonpost.com/graphics/2018/sports/ncaa-applicants/

Romboy, D. (2023, October 21). *Is there too much emphasis (or not enough) on college sports?*. Deseret News. https://www.deseret.com/sports/2023/10/20/23880615/college-football-basketball-emphasis-winning-increase-admissions-enrollment-flutie-effect#:~:text=In%20his%20study%20of%20the,see%20applications%20increase%20by%2017.7%25

Sonas, J., Maggie, & Cukierski, W. (2023). March Machine Learning Mania 2023. Kaggle. https://kaggle.com/competitions/march-machine-learning-mania-2023