Issues in Information Systems
Volume 26, Issue 1, pp. 486-501, 2025

DOI: https://doi.org/10.48009/1_iis_2025_136

Predicting the Unpredictable: Predicting the March Madness Champion using Statistical Modeling

Jack Sweeney, *Bryant University, jsweeney7@bryant.edu*
Suhong Li, *Bryant University, sli@bryant.edu*

Abstract

One of the more exciting and hardest parts of the men's college basketball postseason tournament, named March Madness, is to pick who wins each game and determine the overall champion of the tournament. The data analysis conducted will help determine the overall tournament winner. A machine learning model is implemented using college basketball metrics from previous years to determine the overall winner of this prestigious tournament in 2025. The model was able to pick up the winner in 2024. The result also finds that the most important features in determining the winner of the tournament are wins above bubble, shooting guard height, small forward height, center height, and defensive rebound percentage. Being able to understand the most important variables in determining a winner of the tournament can help individuals in picking the winner in their March Madness bracket groups, which is where the results can lead to implementation or future study.

Keywords: March Madness, Sports Analytics, Machine Learning

Introduction

Data can be utilized in many different industries. In this case, data is used in sports for many different reasons. Since the introduction of data analytics into sports, different sports have been revolutionized. Specifically in basketball, teams utilize data analytics to determine different trends and revolutionize the sport and game. Today, teams use data to determine health and longevity tendencies and changing in-game decision making from coaches and players. College basketball is part of this revolution. Databases that contain college basketball data are utilized for many various reasons. KenPom.com and Barttorvik.com, college basketball databases that contain advanced metrics on different teams, provide insights on teams that can't be determined watching the games. With these advanced metrics, teams can understand how to improve their team and playstyle. Advanced metrics collected for each team are used in the March Madness tournament selection process. Teams that didn't receive a bid from the tournament from winning their conference championship go through a process in which a NCAA committee determines which teams make the tournament and which teams are left out. The systematic process is where the NCAA committee looks at a "resume", which includes record and wins against teams that are considered "harder" to beat. After the selection process, data analysis is used in college basketball to seed the teams in the tournament. (Is it all based on analysis or is there an opportunity to improve seeding as well?). From these specific reasons, there is no doubt that the implementation of data analytics into sports such as basketball has revolutionized the sport.

Literature Review

This section presents a literature review on sports analytics, with a focus that progresses from general sports analytics to basketball-specific research, and ultimately to college basketball and March Madness. This review lays the foundation for our analysis of March Madness predictions.

Sports Analytics

Sports analytics is a field that continues to grow. With the new advancements in technology, professional sports franchises can collect stats to be analyzed to discover and determine different trends. In this case, different stats are created and used to analyze player and team performance, influence in-game decision making from coaches, and determine business decisions. In terms of player evaluation and team performance, "Predictive analytics includes different types of econometric and statistical techniques that range from machine learning, classification, multivariate regression, and other related techniques. Such techniques attempt to analyze historical data to make predictions, which could be about an event in the future or an outcome about the rank of a player" (Sinh, 2020, p. 64). With the data collected, teams can make certain decisions that maximize their team's probability of success. In terms of the business side of sports, analytics can be used in many different aspects. Specifically, "The business applications in different functional areas of management include strategy, operations, leadership, finance and marketing. Business issues deliberated upon are leadership selection, fan-base marketing, social media analytics, match outcome, bookmaker odds, team composition, online marketing, match strategy and others" (Sinh, 2020, p. 67). The business side and decision making of sports continue to evolve. Teams and professional franchises now have the knowledge to run technological applications that help determine better business decisions to maximize the team and player performance. With the increase in player and team performance, fans of the professional sports franchises will be influenced to become more involved with the team.

The increases in the uses of AI can have an immense impact on different areas in sports. In sports, health is a variable that can affect how a player or team performs. Minor injuries can have a negative effect on players. However, major injuries can have a significant impact on a sports team's performance. If a more skilled player that has a positive impact on a team was to suffer from a major injury, the team's performance will decrease. However, with the new innovations in AI, there are possibilities that injuries can be prevented. Injuries can potentially be prevented because, "AI technology has become a key tool in the sports team's medical toolkit. It can not only be used to arrange regular physical examinations of athletes and analyze health parameters but can also be combined with wearable devices to track physical conditions to avoid serious health problems and sports injuries" (Li & Xu, 2021, p. 55). These wearable devices allow analysts to create motion charts that track the player's movements. Teams can collect in-game data at real time based off movement patterns. In this case, the data collected can be analyzed to determine which certain movements can potentially lead to minor or major injuries. Injuries at all levels will continue to decrease with the help of this innovation. If players that positively impact the team's performance are injured less, the team's performance will increase tremendously.

Basketball Analysis

With the evolution of sports analytics, basketball is the latest sport to be revolutionized by its practices. Collecting stats had been started with the use of box scores. First introduced in the 1900s by Henry Chadwick, the data presented in box scores were lists of certain data from the games played that described each player and team performance. In basketball, some of the box score stats include points, assists, rebounds, field goals made/field goals attempted, and turnovers. With the introduction of box scores, many basketball websites have been created. One, named "Basketball Reference", contains preliminary box score information on the NBA and its precursors, the American Basketball Association, Basketball Association of America, and National Basketball League, dating back to the 1946–1947 season; rebounds first appear for every player in the 1959–1960 NBA season. There are also options for variants on traditional box score data, including statistics on a per 100-possession, per game, or per 36-minute basis, as well as an option for advanced box score statistics" (Terner & Franks, 2021, p.3). However, data collection changed with the introduction of AI technology.

With the different applications of AI technology, teams can evaluate the performance of different players and their impact on team success. Different areas of data that can be accumulated include game data, teaching data, and training data. Data is collected in a specific process using AI technology. Data from basketball is collected and organized, leading to the creation of a data warehouse used by teams to make better in-game decisions (Liu, 2019, p. 4). Although, the AI technology system can be organized into three different sections. Each section has a particular function to create an efficient process in which teams can collect and analyze data. The three sections are labeled as the processing layer, the display layer, and the database layer. The processing layer is the area in which an analysis is conducted from the data that creates and determines differing strategies. The display layer is the visualization of the decisions and tactics based on the data. Finally, the database layer is the groundwork in which the data is stored. Professional sports franchises can use the process to conduct data analysis and discover trends in the data.

Data Mining/Machine Learning in Basketball

Cleaning and processing data can be defined as the term "data mining" where data is cleaned and processed for the purpose of discovering different trends (Jackson, 2002, p. 267). Data mining has different technologies that "can perform data classification and prediction, cluster analysis and association analysis, etc., and can conduct deep data mining, which is an important research field to improve analysis and decision-making capabilities" (OuYang et al., 2022, p. 2). When teams can determine different trends, general managers and coaches can make certain decisions to maximize the team's performance. For example, in terms of shot selection, basketball players will have a distinct shot selection and scoring probability influenced by a different number of variables. Trends in data can be discovered with the help of machine learning, which "is significantly related to Computational Statistics whose main aim is to focus on making predictions via computers" (Alzubi et al., 2018, p. 2). Some of the machine learning models used to make predictions include logistic regression, k-nearest neighbors, and random forest (Kim et al., 2022, 2230). Although, the machine learning models listed are not the only ones used to make predictions.

Machine Learning can be conducted in Apache Spark, which is "a fast and general engine for large-scale data processing. It has tremendous speed as compared to Hadoop MapReduce and is up to 100 times faster in memory, and 10 times on disk" (Bhattacharya & Bhatnagar, 2016, p. 208). A data analysis, conducted by Zuccolotto et al., had determined that the scoring probability is impacted the most by the expiration of the shot clock and when a player is shooting a free throw after previously missing the one before. However, there is another variable that affects scoring probability. In this case, scoring probability can be influenced when the shot clock resets.

The data analysis had been conducted by collecting data from the "Series 2A", a professional basketball league in Italy, where a model had been created to determine the most impactful variables associated with scoring probability. After creating the model, validation was conducted by collecting data from the 2016 Olympics held in Rio De Janeiro. The validation process occurred due to the question of whether differences in skills for different leagues influences the variables. Knowing how scoring probability can be affected, coaches can strategize and game plan how their players can get highly efficient shots that aren't affected by the shot clock. With the innovation in the team's shot selection and timing, the team will have a potential increase in wins from taking shots in games that are not "high pressured".

In basketball, there are multiple variables that can determine a team's success. Dean Olliver, a statistician who worked with Dean Smith, determined four main factors that contribute to team success. The four most important factors in determining a team's success are shooting, turnovers, rebounding, and free throws (Kim et al., 2022, 2234). If a team is very efficient shooting the basketball and can limit the opposing team's makes, teams will have a much higher success rate. For turnovers, if a team can limit the number of turnovers they commit and cause the other team to commit more turnovers, that team will have more opportunities to score the basketball, leading to potentially more points than the other team. Like turnovers, if a team can limit the number of shots a team takes by getting defensive rebounds and create more opportunities to score points through offensive rebounding, the team will potentially have more success. For free throws, if teams can generate more free throw attempts and limit free throw attempts from the other team, those teams will have more opportunities to generate more points. Overall, those four factors can generate success at both the professional and college level.

Data Applications and Analysis for College Basketball and March Madness

With the introduction and evolution of different technological applications into sports, college basketball has found a way to use data analytics and math for different uses. One such use revolves around the ranking of college basketball teams that can determine the seeds for teams in March Madness. There are 68 teams in March Madness where 32 teams make the tournament from winning their conference championship and 36 teams are picked from a committee as an at-large bid. Today, a committee selects at-large bid teams based on a "resume" created throughout their season. This resume includes the total number of wins the team had, their strength of schedule, and what type of teams they beat or lost to. In terms of "type of teams", that refers to the difference between the conferences in college basketball. Teams that are in more-talented conferences such as the ACC and the Big East will be considered better than the teams in mid-major conferences such as the America East and the NEC. The committee that selects the college basketball teams to make the tournament that didn't win their conference also looks at advanced statistics of each team. However, a new ranking system, the College Basketball Rating (CBR), can be implemented. The CBR provides a number value for each team that can be determined with the use of statistics from the box score. This certain number value can be used to compare and rank teams to determine which teams should make March Madness. The CBR is based on a game score, which "is a continuous number between 0 and 1 that represents a team's performance in a game; 0 and 1 represent total domination on both the offensive and defensive ends of the court (0 if loser; 1 if winner), and 0.5 represents a close game in which teams were evenly matched" (Stocks-Smith, 2021, p. 48). A logistic regression model is implemented to determine a fitted value for the team. The logistic regression model uses statistics such as assists, steals, defensive rebounding, offensive rebounding, turnovers, blocks, free throws attempted, personal fouls, and field goals attempted. In determining which teams are selected to compete in March Madness, strength of schedule is a major indicator that the committee looks at.

College Basketball Strength of Schedule Evaluation

For a long time, the NCAA tournament committee had calculated a team's strength of schedule based on a calculated weighted percentage. This calculation considers the winning percentages related to the team, their opponent, and their opponent's opponent. The winning percentages are given certain rates used to calculate the team's strength of schedule. This rating process is known as the rating percentage index (RPI). There are two underlying issues to the RPI. The two issues include not having the capability to compare the records of two different teams and the team's abilities directly impact the value of the different schedules. However, with the use of data analytics, there were different techniques introduced to calculate a team's strength of schedule. One of these ranking methods is calculated a lot differently than past methods.

For this new method, a team's strength of schedule is determined by "calculating a team's strength of schedule in terms of the expected number of wins a team on the borderline of receiving an at-large bid would get if they played that schedule" (Fearnhead & Taylor, 2010, p. 109). The method's process revolves around creating a linear model based on the score of two teams that can be used to create a schedule for each team. The strength of the schedule for each team is then determined by calculating the expected win total if they were to play that schedule. From that, teams that deserve to make the tournament at an at-large bid will have more wins than the calculated strength of schedule that was used as a baseline for comparison.

March Madness Evaluation

With new innovations in calculating the strength of schedule and ranking the teams, determining seeds for the tournament is another story. In March Madness, there are 4 regions where there are 16 teams in each region after the first 4 play-in games end. Teams are assigned seeds based on their entire "resume". The 10-person committee can use linear programming to determine and seed all 68 teams. The process of seeding teams has multiple steps, which is detailed as focusing on the rankings of the 2012-2016 tournament teams. Next, factors like team metrics that are used by the NCAA selection committee to select and rank teams and opinions from coaches and reporters are determined. Then, factors that shouldn't be considered by the NCAA guidelines are not used. Following that, the model "determines a set of nonnegative weights wi to assign to the factors, which produces a linear evaluation Vk of each team (Equation (B.1)). The first step requires that we identify the set of nonredundant dominant relationships among the 68 teams that will be seeded for the tournament. The total number of possible dominant relationships among 68 teams is given by 68 × (68−1)/2=2,278. The actual number of dominant relationships in our historical data ranges from a low of 888 in 2012 to a high of 1,154 in 2013. We then identify all redundant dominant relationships (rules) and eliminate them from the set" (Reinig & Horowitz, 2017, p. 183). The described would allow the NCAA committee to make effective decisions in terms of seeding with no controversy involved.

Today, bracketology is a big event in college basketball. People are interested in predicting the outcome of the tournament. Many individuals create groups where the individual who earns the most points from making correct predictions wins their group. In some groups, betting is involved where individuals provide a certain amount of money before the tournament begins. As the tournament advances into further rounds, the points for each game increase as the total amount of points you can earn from making correct picks is 320 for every round. By using data analytics and statistics, one can determine how far a team will advance based on the seeds they were assigned. By using geometric distribution, one can determine which seed will most likely win based on their matchup. In a study conducted by Jacobson et al., it had been determined that the "distribution appears to be most valid for the Elite Eight, the National Semifinals, and the National Championship Game rounds. There is also some indication that this distribution may also be reasonable in the Sweet Sixteen (Jacobson et al., 2021, p. 724). With this knowledge, individuals can make accurate decisions and picks based on different seed matchups. Like the NBA and other professional basketball leagues, data analytics can be applied to college basketball. In this case, college basketball programs can determine how calling a timeout at different occurrences of a game can have differing effects on how well a team scores. In an analysis conducted by Lloveras and Vollmer, they were able to determine conducting an analysis that "on average, the team that did not call the timeout was outshooting the team that called the timeout during the 2-min interval before said timeout" (2021 p. 553). With this knowledge, coaches can better understand when to call a timeout at certain points of the game to give their team a competitive advantage. From the increase interest in March Madness, there has been a focus on whether March Madness can be predicted. A tournament that has an abundance of potential outcomes, multiple methodologies are implemented to determine the tournament's outcomes. One example highlights the implementation of a "purely objective classification model to determine the outcome of March Madness tournaments" (Avalos et al., 2025, p. 7). With the increased development of AI, there can be new discoveries and more accurate predictions for March Madness.

Methodology

For this project, a data analysis will be conducted from men's college basketball data that will provide valuable insights into the men's college basketball post-season tournament (March Madness). The process involves collecting, cleaning, and processing data from multiple verified college basketball statistical websites from the past 11 years. These websites include KenPom.com and Barttorvik.com. Some of the stats collected are advanced stats such as defensive rebound percentage, adjusted offensive efficiency, and free throw rate as well as experience and continuity. Python is used to collect the data from the past 11 years (excluding 2020 due to the cancellation of the tournament) for the website Barttorvik. For KenPom, Microsoft Excel was used to copy the data into a file. After collecting the data, Python had been used to clean and process the data collected.

Discussion of the Results

After the data had been cleaned and processed, Jupyter Notebook and Python were used to visualize the data to find different insights from the information. Different graphs were created to look at the differences between multiple variables based on where that team ended up finishing in the tournament and are shown in Figure 1-5. Figure 1 highlights a key statistic that supposedly determines March Madness success. Figures 2-5 visualize the key statistics basketball statistician Dean Oliver highlighted as the most important aspects of team performance.

Adjusted Offensive Efficiency vs Adjusted Defensive Efficiency

Adjusted Offensive Efficiency measures how many points the team would score per 100 possessions vs. an average Division I team; and Adjusted Defensive Efficiency measures how many points the team would give up per 100 possessions vs. an average Division I offense.

Figure 1 shows that teams that tend to go farther in the tournament have a higher Adjusted Offensive Efficiency and a lower Adjusted Defensive Efficiency. This signals that teams who make it farther in the tournament have more efficient offenses and defenses.

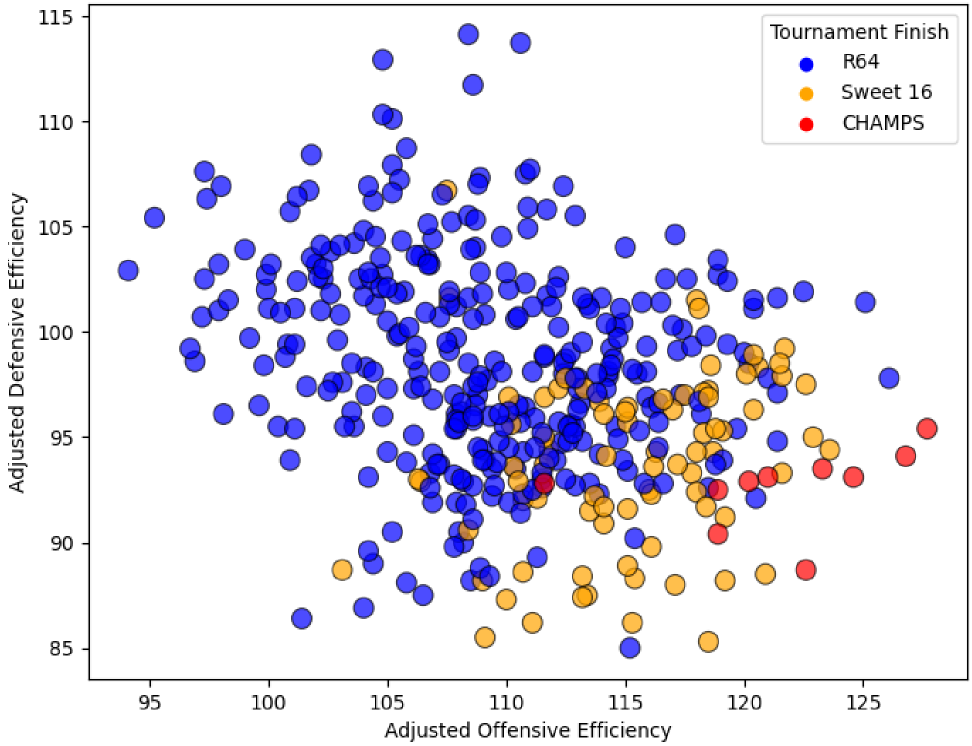

Figure 1. Adjusted Offensive Efficiency vs Adjusted Defensive Efficiency

Free Throw Rate vs Opponent Free Throw Rate

Free Throw Rate measures how often the team gets to the free throw line and Opponent Free Throw Rate measures how often the team limits the opposing team from getting to the free throw line. Figure 2 shows For Free Throw Rate and Opponent Free Throw Rate, there is a minimal difference between teams that have lost in the Round of 64, Sweet 16, and teams who have won the championship.

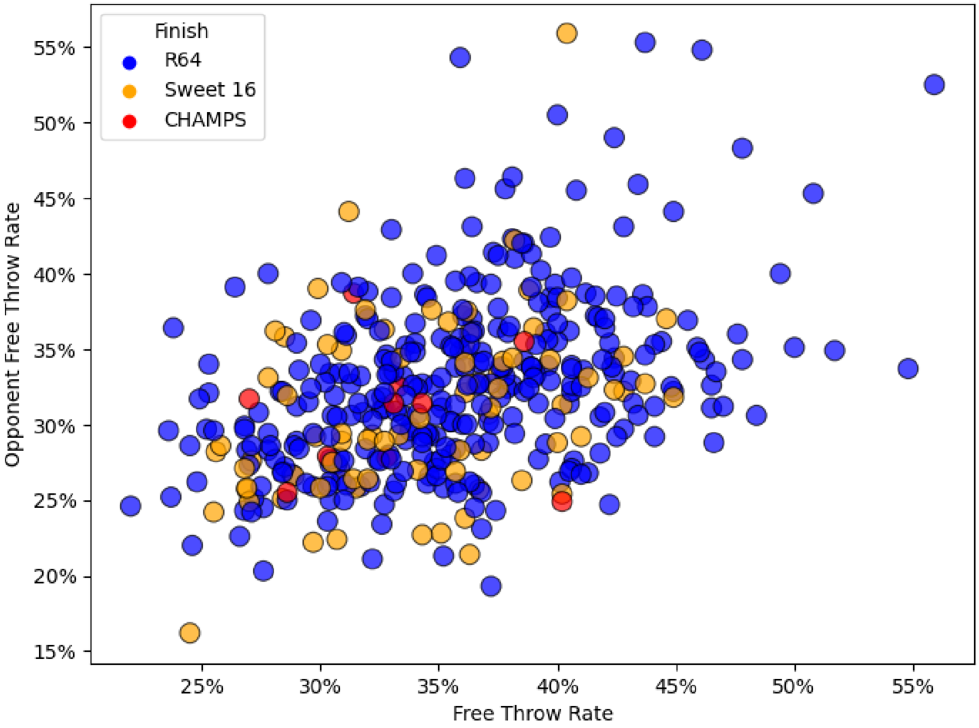

Figure 2. Free Throw Rate vs Opponent Free Throw Rate

Turnover Rate vs. Opponent Turnover Rate

Turnover Rate measures how often a team turns the ball over and Opponent Turnover Rate measures how often a team makes the opposing team turn the ball over. From Figure 3, for Turnover Rate and Opponent Turnover Rate, there appears to be minimal differences between teams that lost in the Round of 64, Sweet 16, and teams who have won the championship. This indicates that the Turnover Rate and Opponent Turnover Rate have a minimal impact on tournament success.

Offensive Rebound Percentage vs. Defensive Rebound Percentage

Offensive Rebound Percentage measures how often a team secures a rebound on the offensive end of the game and Defensive Rebound Percentage measures how often a team secures a rebound on the defensive end of the game.

For Offensive Rebound Percentage and Defensive Rebounding Percentage, Figure 4 shows the distribution of the data has very minimal differences between teams that lost in the Round of 64, Sweet 16, and teams who have won the championship. This indicates that Offensive Rebounding Percentage and Defensive Rebounding Percentage have a minimal impact on tournament success.

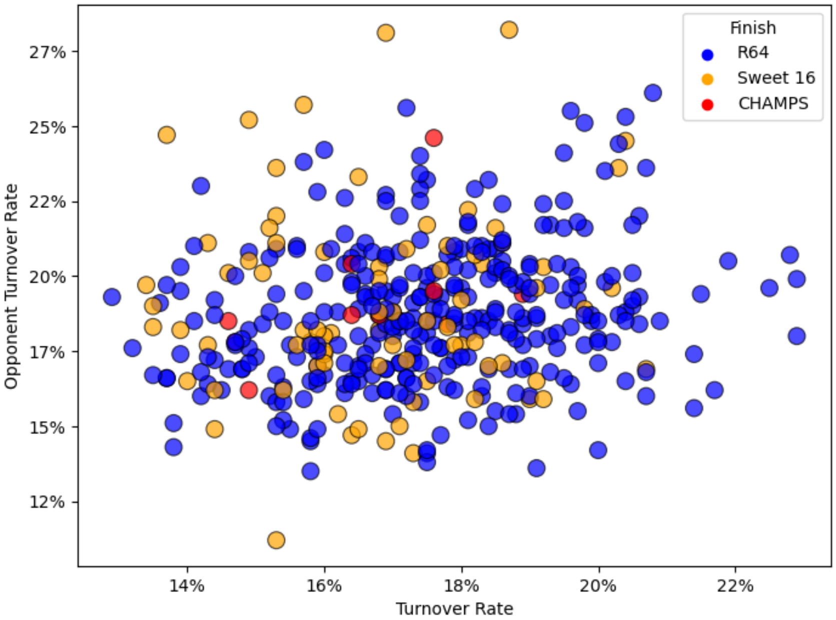

Figure 3. Turnover Rate vs Opponent Turnover Rate

Effective Field Goal Percentage vs Opponent Effective Field Goal

Effective Field Goal Percentage and Opponent Effective Field Goal Percentage measure a team's ability to make shots within a game and limit a team's ability to make shots. Like Field Goal Percentage, Effective Field Goal Percentage values three-point shots more compared to standard Field Goal Percentage. From Figure 5, there appears to be minimal difference between teams that lost in the Round of 64, Sweet 16, and who won the championship.

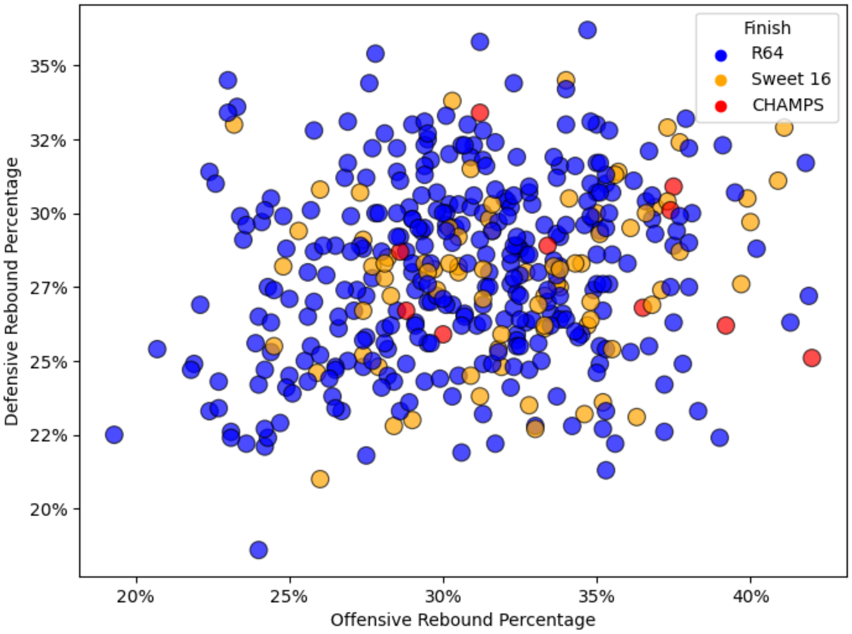

Figure 4. Offensive Rebound Percentage vs Defensive Rebound Percentage

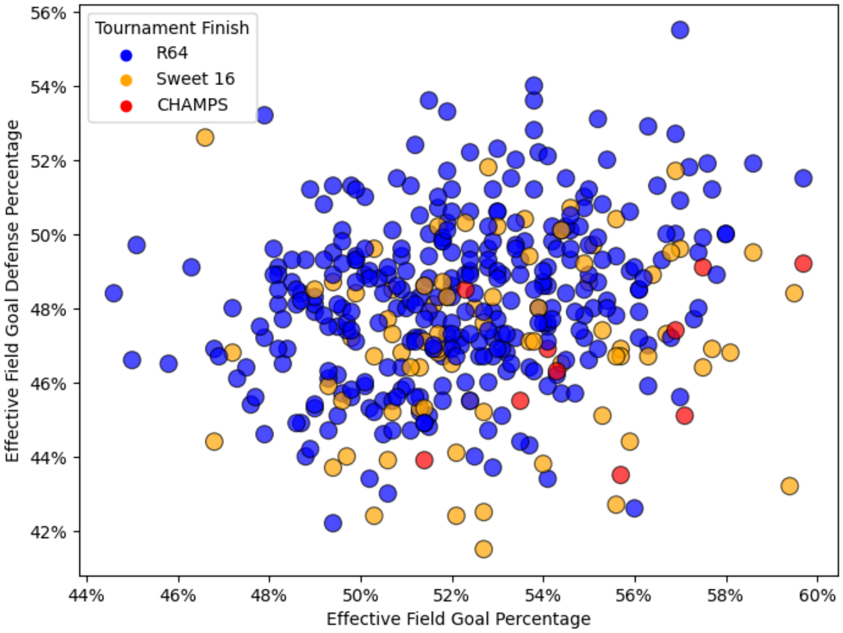

Figure 5. Effective Field Goal Percentage vs Opponent Effective Field Goal Percentage

Experience

Experience measures the average college basketball year experience of a team. First, Figure 6 shows that teams with higher experience, meaning Division I basketball teams with the players that have played the most college basketball, advance further within the tournament. In terms of the average experience by seed as shown in Figure 7, mid to low seeded teams have the most experience. 1 and 2 seeded teams don't have as much experience compared to other teams because Power-Four Conference teams that are very talented and ranked high recruit very talented high school players. The five-star high school players are very talented, able to lead teams to higher seeds, but lack the experience as it is their first year in college basketball.

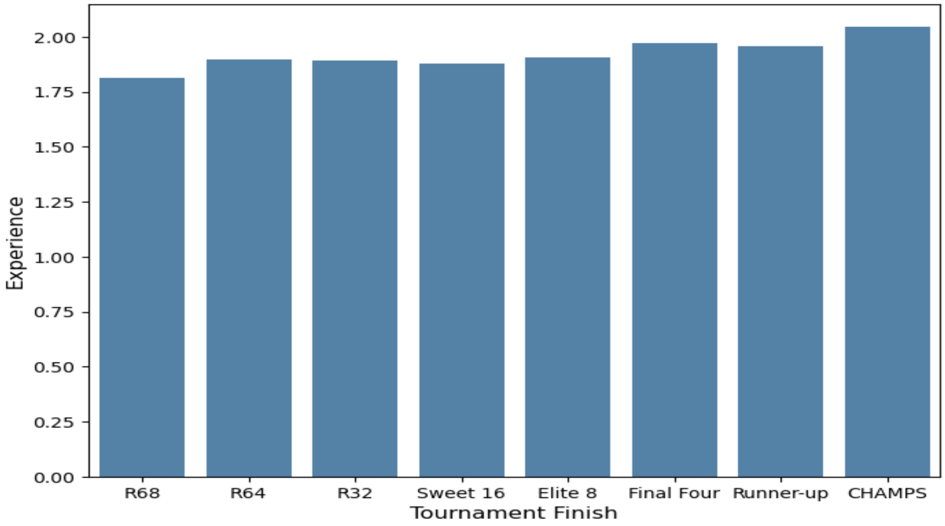

Figure 6. Experience by Tournament Finish

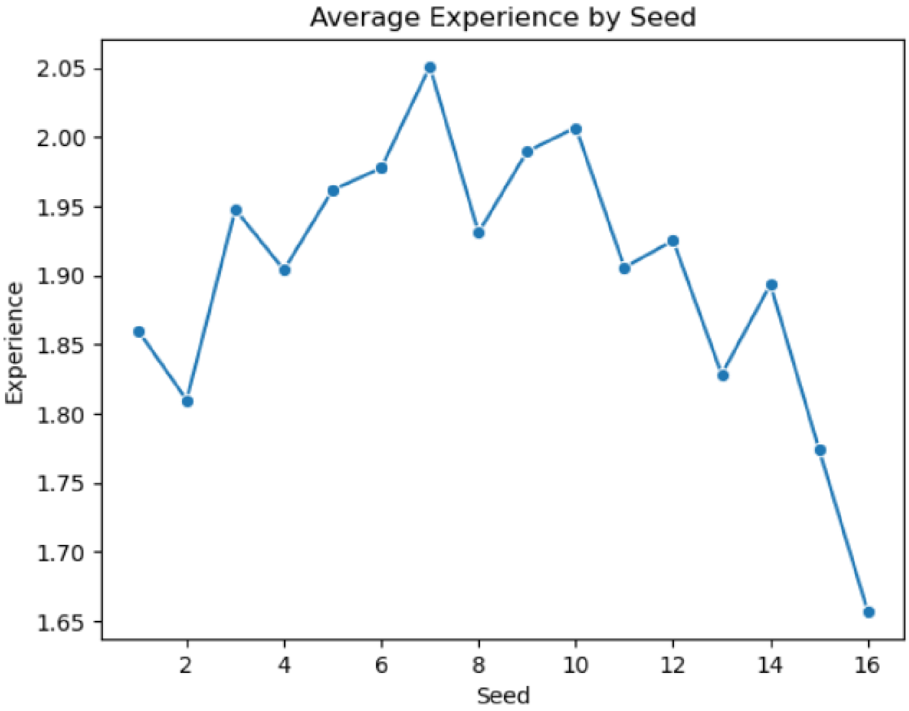

Figure 7. Average Experience by Seed

Continuity

Continuity refers to the percentage of the minutes played by returning players. Figure 8 shows that for continuity, the percentage for each tournament round is peaked at teams who lost in the round of 64 and have won the championship. This indicates that continuity matters in winning the championship. However, continuity also matters in smaller conferences. This is from Figure 9, which indicates that lower seeds, such as 12, 13, 14, and 15, are teams from conferences that only have one tournament bid. This is based off the tournament in which smaller conference teams that don't have a comparable "resume" don't have the capability to make the tournament unless that team were to win its conference. In terms of continuity based off the seeds, continuity matters in smaller conference tournaments.

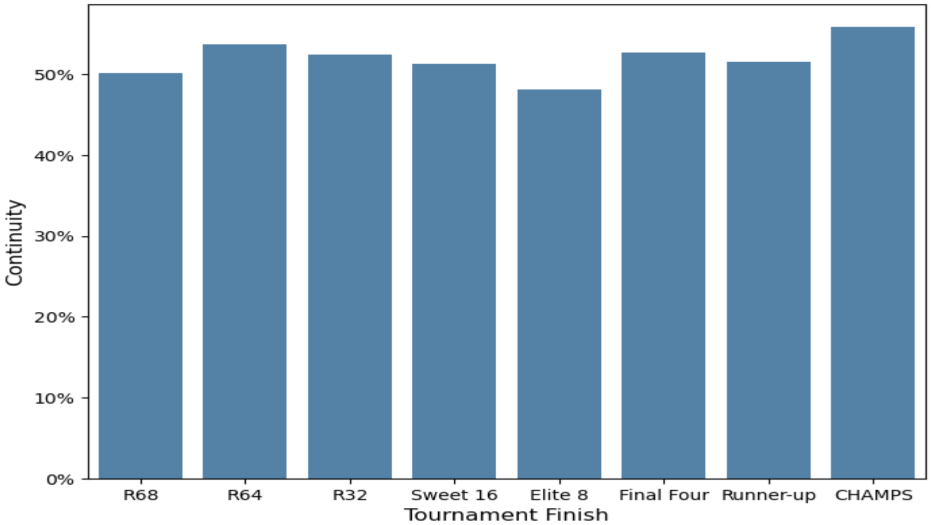

Figure 8. Average Continuity by Tournament Finish

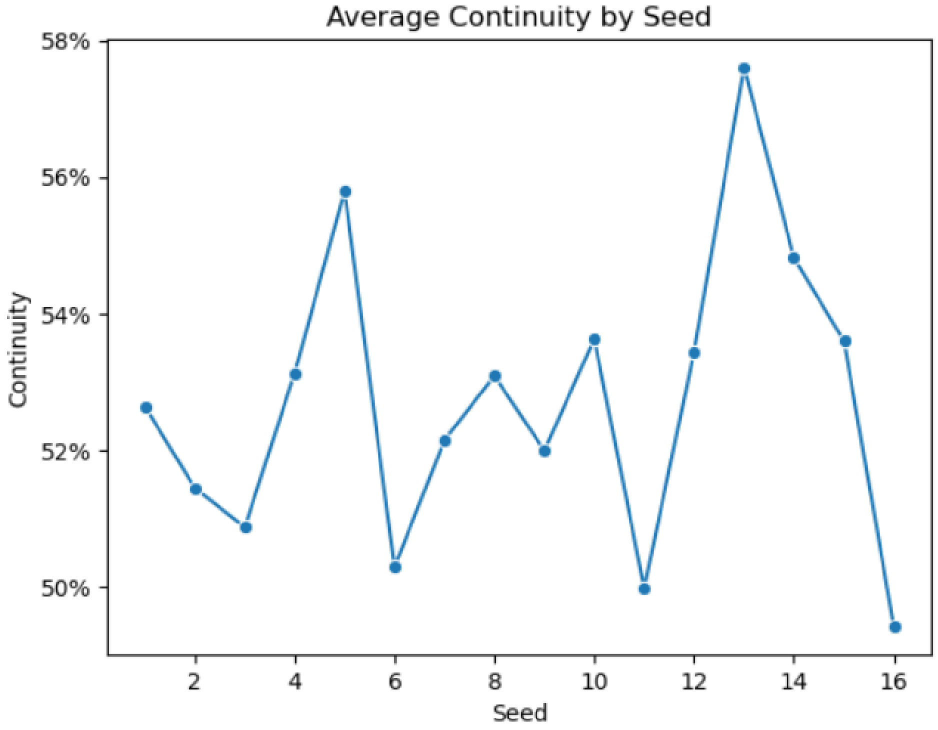

Figure 9. Average Continuity by Seed

Results of Machine Learning Models

Following that, a Logistic Regression model, Random Forest model, and Decision Tree model will be created to provide the probability a team would win the national champion based on the stats. With these models, it can be determined which of the stats collected had the biggest influence on determining the winner of March Madness. The features utilized for the model were: Adjusted Offensive Efficiency, Adjusted Defensive Efficiency, Effective Field Goal %, Opponent Effective Field Goal %, Turnover Rate, Opponent Turnover Rate, Offensive Rebound %, Defensive Rebound %, Free Throw Rate, Opponent Free Throw Rate, Three Point %, Opponents Three Point %, Adjusted Tempo, PG Height, SG Height, SF Height, PF Height, C Height, Average Height, Wins Above Bubble, Experience, Bench, Continuity, Rank, and Games Played.

From Figure 10, features such as Effective Height, Seed, Two Point %, Opponent Two Point %, Barthag, number of wins, number of losses, number of conference wins, and number of conference losses were not included, as their high correlation with other features would lead to multicollinearity, which would disrupt the model's performance.

After finishing the three models, valuable information can be used to determine the winner. The model's training data was the data from 2014-2023, excluding 2020 due to the tournament cancellation. The testing data was the 2023-2024 season, which would be used as a baseline to determine each model's performance.

Even though Random Forest had the highest accuracy score, the model was unable to predict the 2024 champion. The Logistic Regression model predicted the winner correctly, which was UConn. In comparison, Random Forest predicted Arizona as the 2024 tournament champion, who lost in the Sweet 16 and Auburn, who lost in the Round of 64. For Decision Tree, the model highlighted Auburn, who lost in the Round of 64, Arizona, who lost in the Sweet 16, North Carolina, who lost in the Sweet 16, Houston, who lost in the Sweet 16, and Tennessee, who lost in the Elite 8. Based on its 100% recall, which captures the True Positive Rate (the champion), its higher precision and F1-Score, the Logistic Regression model was the model chosen to predict 2025.

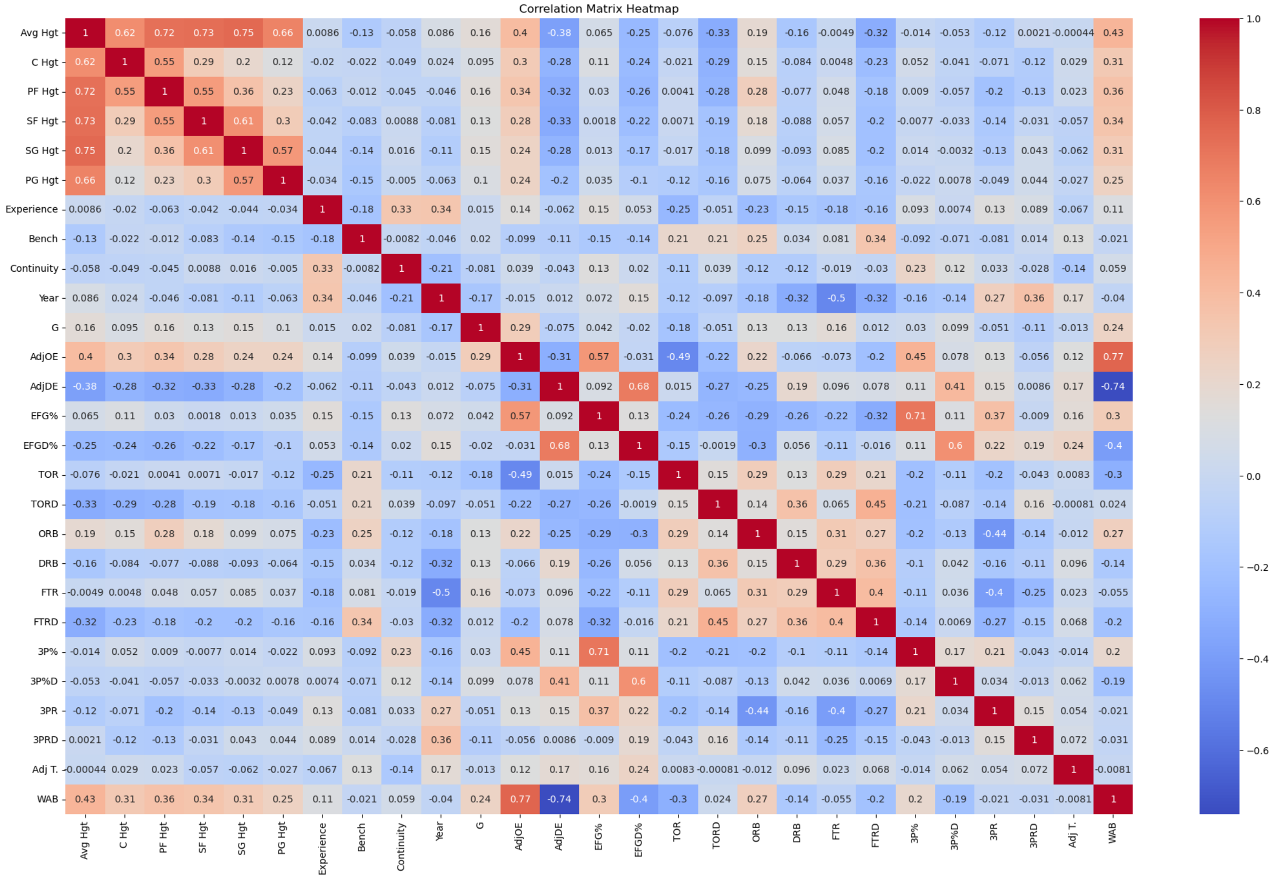

Figure 10. Correlation Matrix of Features

For 2025, the Logistic Regression Model predicted Auburn as the champion. Some of the other teams highlighted were Florida, Houston, Alabama, and Texas Tech

Figure 10 shows the top five most important features, including Wins Above Bubble, Shooting Guard Height, Small Forward Height, Center Height, and Defensive Rebound Percentage. The most important feature is wins above bubble (WAB). This metric compares a team's performance against tournament bubble teams (teams who possibly make or miss the tournament). The higher the WAB, the better the team would perform against bubble teams. In terms of tournament success, teams that are taller at the shooting guard, small forward, and center and can effectively defensive rebound tend to impact a team's success and is a major factor in whether a team becomes the NCAA Division I Men's Basketball Champion. Teams that have more size tend to be successful.

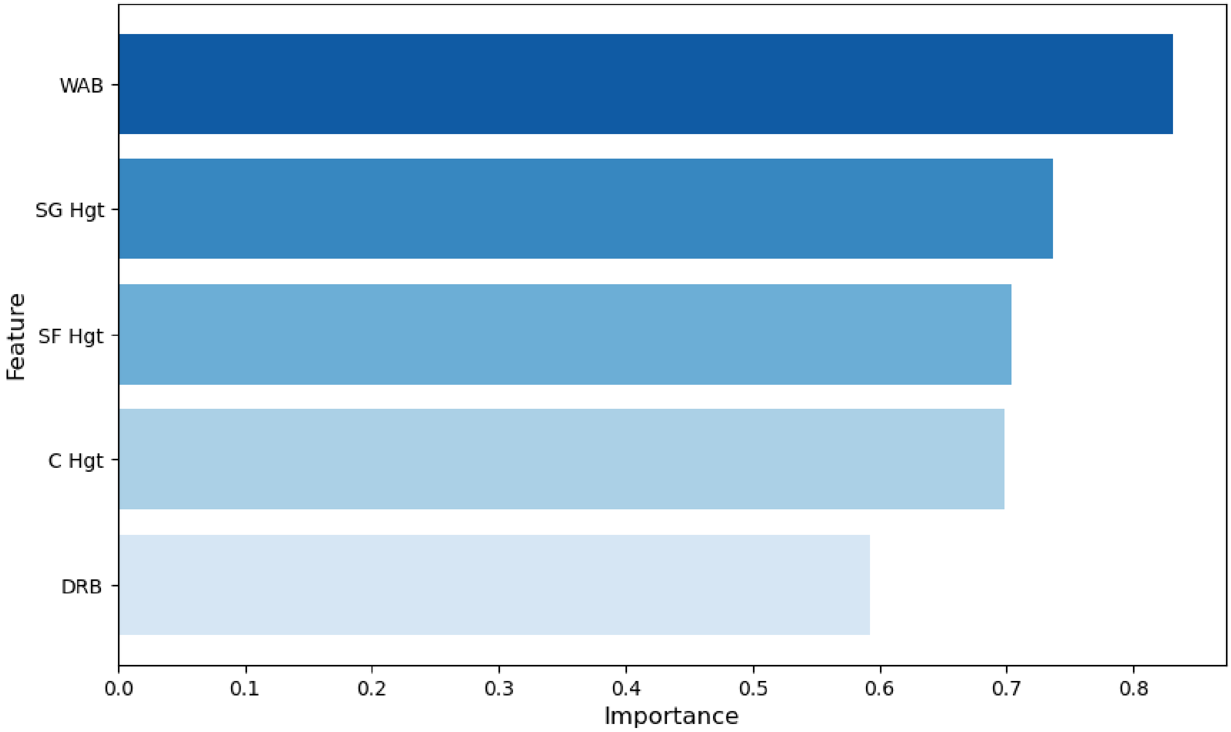

Figure 11. Top Five Most Important Features

Conclusion and Future Research

In the final analysis, the Division I Men's Postseason Tournament Champion can moderately statistically be predicted. From utilizing a Logistic Regression model and visualizing advanced metrics, the tournament champion can be predicted. For 2024, it was able to predict UConn as the champion while listing other potential winners such as Houston, Purdue, and Creighton. For 2025, it predicted Auburn as the champion and highlighted teams, such as Florida, Houston, Alabama, and Texas Tech. The most important variables that contribute to crowning the national champion is wins above bubble (WAB), which compares a team's performance against tournament bubble teams (teams who possibly make or miss the tournament), size at the shooting guard, small forward, and center positions, and being able to rebound the basketball effectively on defense. Even though the 2025 prediction of Auburn was incorrect, the Logistic Regression Model was able to highlight Florida, who won the national championship, and was correct in predicting the 2024 March Madness Champion, which was UConn.

From the analysis, coaches can change their roster constructor and playstyle. Coaches can focus on recruiting high school/college transfer players that are taller at the shooting guard, small forward, and center positions. For example, coaches at top teams such as Duke should prioritize a 6'5'' shooting guard compared to a 6'2" shooting guard with very similar talent. In terms of play style, coaches should implement a bigger focus on defensive rebounding. If coaches were to empathize defensive rebounding, the opposing team will have more limited chances on offense. With recruiting taller players at the shooting guard, small forward, and center positions, teams can increase their chances at success in the NCAA Division I Tournament.

There were multiple limitations within this data analysis. One limitation was the inability to capture in-game adjustments from coaches. In-game adjustments measure a coach's ability to adapt to an opponent's playstyle and game plan. In-game adjustments are important in March Madness, as it's a single-elimination tournament where every game matters. Another limitation is the inability to capture the impact of regional seeding. Teams that are placed in regions farther from their school could potentially be a disadvantage compared to teams that are placed in a nearby region. For example, an east coast team could be at a disadvantage against a West Coast team placed in the West Region, due to regional issues leading to a potential "home court advantage" for the west coast team.

A future study can focus on the generalization of the model. Because the implemented Logistic Regression model predicted 2024 correctly but 2025 incorrectly, the 2024 prediction could potentially be incidental. To determine the model's overall success, a five-year rolling window can be implemented to determine whether the accurate predictions are accurate and not coincidental. These predictions can then be compared to models utilizing a seed-based approach that would be the baseline to determine the model's overall effectiveness.

Another future study can focus on seeing the overall impact of "Name Image Likeness" (NIL) on college basketball. NIL allows student athletes to receive compensation for their accomplishments on the basketball court. Higher NCAA Division I Men's Basketball Teams can offer contracts to players on opposing teams that are in the transfer portal, almost as free agency in professional sports. In the future, I plan on conducting an analysis on how big an impact NIL has on college basketball teams and players. As NIL continues to grow, it's important to understand its overall impact on team success. Understanding how important NIL is can determine decisions made by college basketball players and their overall impact on their team.

With the evolution of AI and data analytics in basketball, another future study that can be focused on is the improvement of seeding within March Madness. The committee has the weakness of making mistakes by overseeding and underseeding teams based on their "resume". A future study could be to find a similar but more accurate baseline to seed teams that disputes overseeding and underseeding. The baseline could make the tournament seeding more accurate, leading to a more accurate tournament. However, increased accuracy could lead to the reduction of potential upsets within the tournament, as that could lead to potentially reducing fan excitement of the tournament.

References

Avalos, K., McIver, C., & Nayak, N. (2025). March Madness Tournament Predictions Model: A Mathematical Modeling Approach. *Proceedings of the Intercollegiate Mathematical Modeling Challenge*. Harvard University, MA.

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series.* 1142. 012012.

Bhattacharya, A., & Bhatnagar, S. (2016). Big Data and Apache Spark: A review. *International Journal of Engineering Research & Science*. *2*(5). pp.206-210.

Fearnhead, P., & Taylor, B. M. (2010). Calculating Strength of Schedule, and Choosing Teams for March Madness. *The American Statistician*. *64*(2).pp.108-115.

Jackson, J. (2002). Data Mining; a Conceptual Overview. *Communications of the Association for Information Systems*. *8*. pp. https://doi.org/10.17705/1CAIS.00819

Jacobson, S. H., Nikolaev, A. G., King, D. M., & Lee, A. J. (2011). Seed Distributions for the NCAA Men's Basketball Tournament. *Omega*. 39(6). pp.719-724.

Kim, J. W., Magnusen, M., & Jeong, S. (2023). March Madness Prediction: Different Machine

Learning Approaches with Non-box Score Statistics. *Managerial and Decision Economics*. 44(4). pp. 2223-2236.

Li, B., & Xu, X. (2021). Application of Artificial Intelligence in Basketball Sport. *Journal of Education, Health and Sport*. 11(7). pp.54-67.

Liu, Z. (2020). Application of artificial intelligence technology in basketball games. *IOP Conference Series: Materials Science and Engineering.* 750. 012093

Lloveras, L. A., & Vollmer, T. R. (2021). An Analysis of Timeout Calling in College Basketball. *The Psychological Record*, 72. pp.551-559.

OuYang, Y., Zhang, Y., & Li, Y. (2022). The Application Method of Big Data of Data Mining Algorithm in College Basketball Teaching. *Scientific Programming*. https://doi.org/10.1155/2022/2352676

Reinig, B. A., & Horowitz, I. (2018). Using Mathematical Programming to Select Seed teams for the NCAA Tournament. *Interfaces*. *48*(3). pp.181-188.

Sarlis, V., & Tjortjis, C. (2020). Sports Analytics—Evaluation of Basketball Players and Team Performance. *Information Systems*. 93.101562.

Singh, N. (2020). Sport Analytics: A Review. Learning. 9(11). DOI:10.2991/itmr.k.200831.001

Stocks-Smith, J. (2021). College Basketball Rating (CBR): A New Body-of-Work Metric for NCAA Tournament Selection. *Journal of Sports Analytics. 7*(1). pp.47-55.

Terner, Z., & Franks, A. (2021). Modeling Player and Team Performance in Basketball. *Annual Review of Statistics and Its Application*. 8. pp. 1-23.

Zuccolotto, P., Manisera, M., & Sandri, M. (2018). Big Data Analytics for Modeling Scoring Probability in Basketball: The Effect of Shooting under High-pressure Conditions. *International journal of sports science & coaching*. 13(4). pp. 569-589.