Sports Analytics: Using Null Hypothesis Testing and Logistic Regression to Determine Contributing Factors to Team Success within the National Collegiate Athletic Association Division I Men's Basketball Tournament

Elena Adame

Bellevue University

DSC 680: Applied Data Science

Dr. Brett Werner

May 7, 2023

Contents

Background

The National Collegiate Athletic Association (NCAA) Division I men's basketball tournament, better known to the majority as March Madness, has been around since 1939. Originally played with only eight teams, the tournament has expanded to include 68 teams from 32 Division I conferences spread across the United States (NCAA Division I men's basketball tournament, n.d.). The 84th edition of the tournament saw several upsets, with lower ranked teams overcoming those that had been favored to win. However, despite these upsets, the University of Connecticut took the trophy home, garnering them their fifth overall win as national champions. The school remains amongst the ranks of the few other schools who have won the tournament multiple times, but what makes a team successful in this tournament? This project examines what statistics make a team into champions and aims to provide a guide for improvement for teams that struggle in the NCAA tournament.

Data Explanation

The data used in this project was collected from Sports-Reference.com, a site dedicated to collecting current and historical data on college basketball. The data was pulled from the NCAA Seasons Index; information dating back to the inception of the tournament was not available. Data was only available dating back to 1993. Due to COVID-19 restrictions, the NCAA tournament was not played during 2020 and no data was collected. As such, only 29 years' worth of data was able to be gathered. Each year's corresponding index included the complete list of teams eligible for selection into the tournament along with their associated statistics. These statistics are comprised of each team's overall games, wins, losses, win-loss percentage, and their in-game statistics, which includes features such as assists, steals, blocks, etc. As the NCAA Tournament outcome was the specific focus of this project, the data was split further into teams that were eligible for the tournament and teams that were not. Teams that did not make the tournament were discarded from consideration.

Methods and Analysis

To begin analysis of the data, a Shapiro-Wilk test was conducted to test the data for normality. This test operates under a null hypothesis that the data is a normal distribution. The result produced from this test was a p-value of 0.0, indicating that the data had a non-normal distribution. As such, the data did not meet the conditions to perform a T-Test or ANOVA test. As such, a Kruskal-Wallis Test was performed. This test does not assume there is a normal distribution within the data and operates with a null hypothesis that there is no statistical difference between the two data groups being tested. The data was separated into teams that had won the NCAA and teams that had not. For this test, the p-value was determined to be equal to 0.0, indicating there was a significant statistical difference between the two groups.

Exploratory data analysis was begun by first creating a correlation matrix to determine if an initial determination could be made as to what statistics are closely related to a team's success.

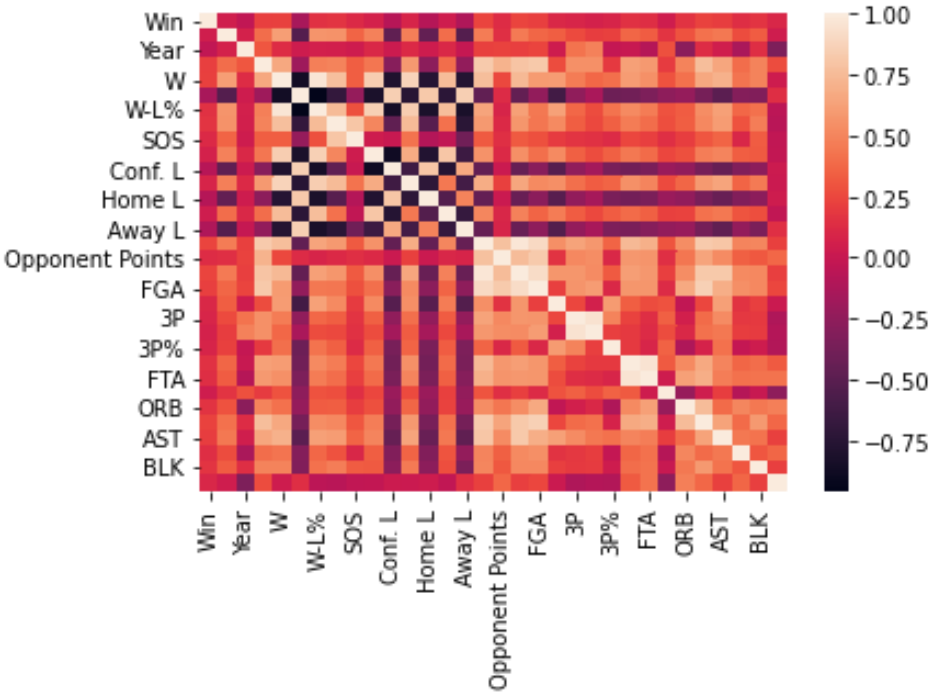

Figure 1: Correlation Matrix for Basketball Team Statistics as Related to Team Success (Win)

The matrix above shows that there is no one statistic that is closely correlated to that of team success. The five statistics with the highest correlation were Team Points (0.25), Field Goals (0.24), Field Goal Attempts (0.22), Total Rebounds (0.23), and Assists (0.21). The Team Points statistic was not closely looked at for this project as it stands to reason that teams that score more points win more often. Figure 2a – 2d below show the distribution of for winning and losing teams for the other four statistics identified above.

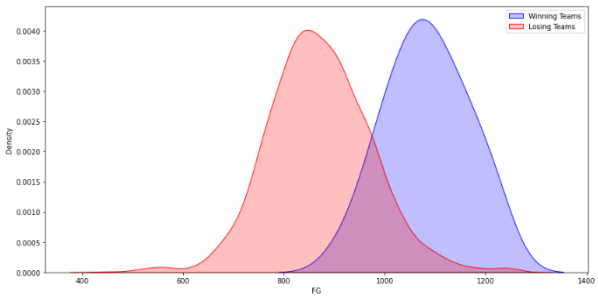

Figure 2a: Field Goals for Winning and Losing Teams

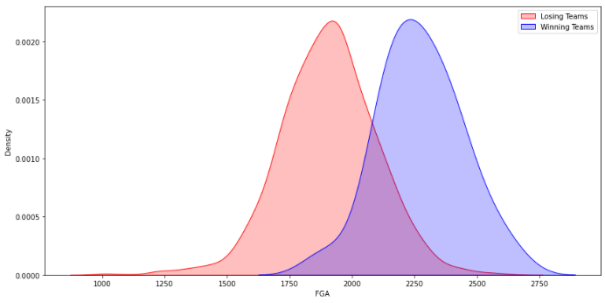

Figure 2b: Field Goal Attempts for Winning and Losing Teams

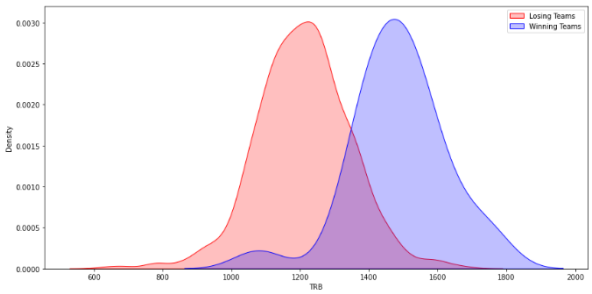

Figure 2c: Total Rebounds for Winning and Losing Teams

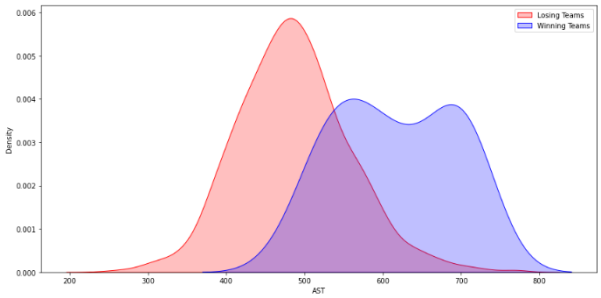

Figure 2d: Assists for Winning and Losing Teams

Each of the figures highlights there is a clear difference in performance between the teams that win the NCAA tournament and those that do not. There is, on average, a difference of at least 500 between statistics for winning and losing teams, the exception being for Assists. However, this does suggest that winning teams are either training their players on achieving these types of statistics or are recruiting players with higher-than-average scores for these statistics. There is a wider distribution with the assist statistic, possibly indicating that team dynamics are at play here and winning teams place different weight on development for this.

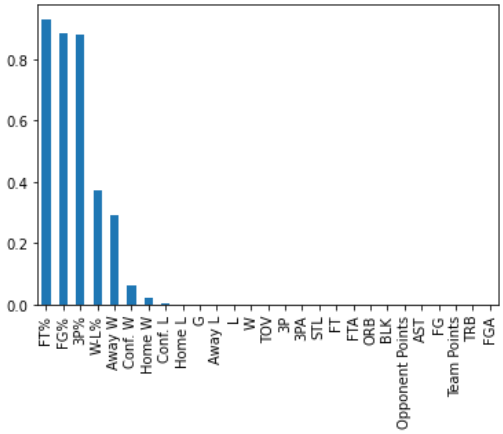

Figure 3: Result of Chi-Square Test for Team Statistics

To determine if these features truly were the best features by which to judge a team's success, a Chi-Square test was conducted. This test produced the output shown in Figure 3 and corroborated the selection of the five features identified above as being significant to a team's success within the NCAA tournament. The statistics that returned p-values higher than the null hypothesis (0.05) were the Free Throw Percentage (FT%), Field Goal Percentage (FG%), 3-Point Field Goal Percentage (3P%), Win-Loss Percentage (W-L%), Away Wins (Away W), Conference Wins (Conf. W), Home Wins (Home W), and Conference Losses (Conf. L).

A Logistic Regression model was created for this project to determine if the model could correctly assess which teams would win based on the provided features. Those team statistics that the Chi-Square test showed a high p-value were removed from the dataset and the remaining statistics were loaded as the 'Features' variable. The chosen Target variable was the 'Win' variable. Unfortunately, this model, while producing an accuracy of 98.17%, was entirely unable to predict team success. This model only correctly predicted 2 of the 11 winning teams included in this set. The classification report for this model showed that the recall value was equal to 0.18, indicating that it could only correctly identify 18% of true positive cases and would benefit from an increased number of positive cases in the training set. Adding the removed features back into the model produced the same accuracy. However, this model was entirely unable to provide a correct prediction for the eight included winning teams. Finally, a third model was

created with only the top four features selected: Field Goal Attempts, Total Rebounds, Field Goals, and Assists. While this model produced an accuracy of 99.19%, it was truly only capable of correctly identifying losing teams. The model failed to correctly identify any of the winning teams. Of the three models, only the second model, which excluded the features with p-values higher than 0.05, was able to produce any type of values for winning teams with regard to Precision, Recall, and F1-Scores. This model despite not having the overall best accuracy score, was the best performing model due to its ability to identify winning teams.

Conclusion

All three models' inability to correctly recognize true positive wins leads to the conclusion that there is not enough data regarding wins for the model to make an accurate assessment of patterns. As such, the result of these models was the opposite of this project's intention, they are accurate at predicting losing teams but fail more than half of the time to identify winning teams. This also leads to the conclusion that these models are potentially overfit to the data.

Assumptions and Limitations

With regard to assumptions made during the course of this project, outliers were not considered. This was an oversight made on the part of the researcher. As outliers were not checked for, it can be assumed that the model is potentially overfitted to the data which could also contribute to its inability identify winning cases. There may be similarities between losing team statistics containing outliers and winning team statistics. This model is also limited by the data available. Initial assumptions were made as to availability of data. However, the data only dates to 1993. Prior to this year, data either does not exist or contains some but not all statistics contained within data dated after 1993.

Challenges

One challenge that may have inhibited this project is the lack of qualitative data being used. Team interviews and dynamics are very important to understanding how the team operates. By operating solely on facts and statistics of each team, this project is only able to provide a limited scope of the situation. A team may present lower statistics regarding shooting averages, rebounds, assists, etc., but what is unable to be examined is how the team dynamic played into creating a situation that allowed or promoted these lower statistics. Additionally, this project lacks information about coaches for each college being examined. With over 200 teams able to be eligible for selection into this tournament, the data is simply too large and not consolidated to pull it together in a timely and coherent manner. Working on a limited time schedule for this project does not allow for this data to be collected and examined. However, what can be done to address this challenge is to identify if losing teams do suffer lower team statistics. Schools that display losing qualities can be looked at to determine if coaching or team dynamics are at fault.

Future Uses and Applications

This model could be used by school sports programs to assess their team's performance during the regular season to determine if they have a chance at winning the NCAA tournament once selected. As this model is currently very accurate at predicting team losses, this could help coaches and directors identify where their team is struggling by changing different statistics for their team. Additionally, this model could be used by sports analysts during the pre-season to build out bracket predictions.

Recommendations

One recommendation for improvement of this model would be to identify and remove any outliers from the data and reconstruct the model absent these samples. Additionally, this could help in determining if the model is overfit to the data. Furthermore, this model would benefit from conducting tests to determine if it truly is overfit to the data.

Implementation Plan

Once perfected and able to correctly identify winning team cases, this model could be set up as an interactive web GUI with various inputs for team statistics. This would allow sports analysts and team coaches to input their chosen team's information and be returned a prediction and assessment of their team's performance.

Ethical Assessment

For ethical concerns, the use of player data is not of concern for this project as all data has been sanitized and consolidated to represent the team as a single entity rather than focused on individual players. In this way, this project avoids directly using single player data and instead looks at each team as a whole. One consideration that must be noted is the use of statistical analysis to attempt to build a better team. This project is focused toward how to improve teams that have historically been on the losing side and does not investigate how to improve winning or average teams. This may be considered unfair to some members of the basketball community as it could potentially provide an unfair advantage to some teams over others. The only way this ethical concern may be addressed is by providing this analysis to the public and ensuring transparency with regard to how future decisions based on this project are being made.

Appendix A: Questions

1. What features are most important to this model and/or team success within the NCAA tournament?
    a. Based on the Chi-Square Test that was performed, the most important features for this model are Field Goal Attempts, Total Rebounds, Team Points, Field Goals, and Assists. These features also correlated the highest to the Win Target Variable as shown in the correlation matrix (Figure 1).
2. How would the model change if budgetary information was included in this dataset?
    a. It is unclear how this model would change. My hypothesis is that the model would be biased toward colleges that have higher budgets for their basketball program as many of the winning teams come from larger schools that are able to divert more money to supporting their sports programs.
3. How would you incorporate qualitative data such as team interviews pre- and post-game?
    a. I would separate out the interviews using NLP to determine if the players had a positive or negative feeling about the game both pre- and post-game. This could also help the model determine if negative feelings heading into the game affect the outcome and whether that needs to be addressed by the coaches to see team improvement.
4. How do you plan to address the model overfitting to the data?
    a. Because the model is overfit to the data, I would take the dataset containing all losing teams and aggregate the school information so that each school is represented once. I would do this by taking the average of each schools statistics and assigning it to that school This would help slim down the overall dataset.
5. Would including team information for those that were not eligible for the NCAA be beneficial?

    a. No, I do not believe so. I believe that including data from schools that were not eligible for the NCAA would further skew the model and decrease it's ability to predict winning teams. These schools were not eligible for the tournament for specific reasons and to include that data in the overall model would increase the ratio of losing teams to winning teams, which is currently an issue.

6. Do you think this could skew future NCAA tournaments were this model to be implemented?
    a. If the model could properly predict winning teams, it could potentially skew future NCAA tournaments. However, coaches would have to use the model to improve their team for this to happen. What this model would most likely affect in the near future is the gambling aspect that surrounds the NCAA with bracket building competitions being a very popular thing for fans to participate in.
7. The default decision threshold for Logistic Regression models is 0.5, have you considered adjusting your threshold to improve the model's performance?
    a. I have considered raising the decision threshold for the logistic regression model, however, I would not want to rely solely on this to improve the model's performance. While raising the decision threshold could help, it may not be enough to fix the model overall. I would rather explore continued feature selection and slimming down the dataset.
8. Would scaling down the data in terms of losing team samples be beneficial to this model's performance?
    a. While the model is currently working off of an already small dataset, I believe slimming down the Losing Team dataset will help with the model being overfit to that specific data.

9. After checking for and removing outliers, how else would you adjust the model to improve performance?
    a. I would also work to determine if hyperparameter tuning for the model would improve overall performance, however, with Logistic Regression, there is not much to adjust. Additionally, as stated above, I would work to aggregate the data of the Losing teams to slim down the datapoints.
10. You said that data prior to 1993 was not available or did not contain all of the same statistics, would you consider adjusting your dataset to include the years (prior to 1993) that do have some of statistics?
    a. Initially, I did consider this approach. However, I don't believe this would be beneficial for two reasons: 1) I am already seeing overfitting in the model. Introducing more data to the model in terms of Losing Teams does not seem like the right way to go about fixing the overfitting problem. 2) These statistics are not all inclusive and potentially hurt the model to slim down the number of features.

Appendix B: References and Supporting Documentation

McKeon, K. (2020, April 17). *Fun with Finances: Basketball Budgets*. Retrieved from Three Man Weave: https://www.three-man-weave.com/3mw/college-basketball-budgets-2020

NCAA Division I men's basketball tournament. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/NCAA_Division_I_men's_basketball_tournament

NCAA Seasons Index. (n.d.). Retrieved from SRCBB: https://www.sports-reference.com/cbb/seasons/

Schuster, B. (2023, March 25). *Kansas State's Elite Eight loss keeps bizarre men's NCAA tournament streak alive for a 76th year*. Retrieved from FTW USA Today: https://ftw.usatoday.com/2023/03/kansas-states-elite-eight-loss-bizarre-ncaa-tournament-streaks-purple-teams

U.S. Department of Education. (n.d.). *Comparing Data from Multiple Schools*. Retrieved from Equity in Athletics Data Analytics: https://ope.ed.gov/athletics/#/compare/search

Wilco, D. (2023). *What is March Madness: The NCAA tournament explained*. Retrieved from NCAA Men's Basketball: https://www.ncaa.com/news/basketball-men/bracketiq/2023-03-15/what-march-madness-ncaa-tournament-explained