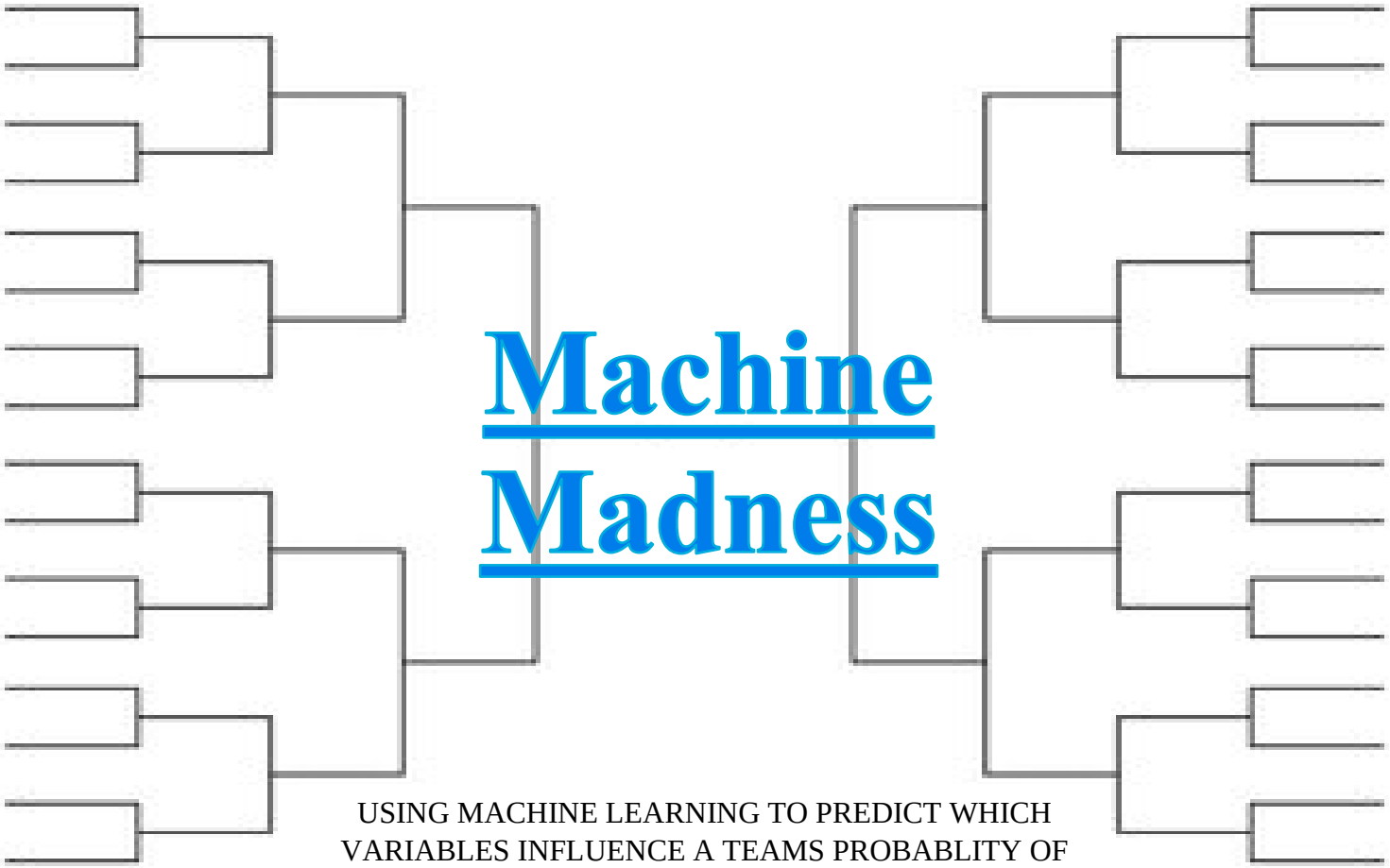

Machine Madness

USING MACHINE LEARNING TO PREDICT WHICH VARIABLES INFLUENCE A TEAMS PROBABLITY OF ADVACING TO THE ELITE 8 OF THE NCAA MEN'S BASKETBALL TOURNAMENT

GROUP 5

CHASE BAKER, GABRIEL LAMMERS, JOEY POWELL AND ANDREW SEBAHAR

Contents

Executive Summary

The aim of this project is to help coaches, players and sport bettors improve their predictions on a team's probability of advancing to the Elite 8 in the Men's NCAA March Madness Tournament. For players and coaches, the tournament is a prime-time event to be scouted and an opportunity for fame. For sports bettors, the tournament presents the opportunity to win millions of dollars. As avid college basketball fans, our team explored different statistics relevant to a team's success and determined which features have the greatest impact on determining whether a team makes the 'Elite 8'. Our data shows that working hard in the regular season pays great dividends in the tournament. Basic fundamentals such as limiting turnovers and maximizing rebounds also lead to success. In our exploratory analysis we go over similarities between Elite 8 teams and examine the importance of being in a power conference. Teams in power conferences are historically frequent in the last rounds of the tournament and see tougher competition during the regular season. We built three models to calculate the most accurate results. Linear regression turned out to be our most effective model, however perfectly predicting Elite 8 appearances is difficult for any machine considering there are so many teams and so few of them make it past the Sweet 16.

Problem Description

Business Goal

In 1985 the NCAA expanded its national basketball tournament to 64 teams. Since then, over 60 million March Madness brackets are filled out each year with the miniscule probability of having a perfect bracket. The odds that just one of these brackets are perfectly completed are 1 in 9.2 quintillion! Billions of dollars in prizes are awarded to the most successful brackets through sweepstakes and competitions. Warren Buffet has even laid down a $1 billion wager of his own money for anyone who conquers a perfect bracket. In addition to sports-betting, the tournament offers a massive platform for players and coaches to showcase their skills on national television for NBA scouts and fans. So, what makes a perfect bracket so difficult to predict? And how can we use machine learning methods to predict the winners and reap the benefits of these mega-competitions?

Data Mining Goal

For our project, we examined the odds that a team makes it as far as the Elite 8 round in March Madness. Of more than 300 Division 1 men's college basketball teams, only 64 make the tournament. Half of these teams are eliminated each round, so a team must win three games before being one of the final 8 teams remaining. Our data mining goal was illustrated by a classification problem with the target variable being whether a team made it to the Elite 8 in a given year (Yes - 1/No - 0). We wanted to figure out which of our features were the best predictors of each team's success in the tournament and determine

which variables are the most influential. These answers can help not only bracket-predictions but also coaches and players who want to dominate the game and scouts who want to build a winning roster.

Data Description

Data Overview

Our dataset was recovered from Kaggle.com. It is a college basketball dataset covering the last 9 years of Division I Men's Basketball. We decided to include only seasons 2013-2019 and the 2021 season, while omitting 2020 because it was not completed due to the onset of COVID-19 regulations. The collection spans over 300 teams for each year and includes a range of statistics based on offensive and defensive efficiency and other relevant information.

Exploratory Analysis

Our exploratory analysis incorporated various distribution methods to easily visualize our data and make relevant assumptions. Methods we used included histograms, box plots, and clustering.

```
       G                W               ADJOE            ADJDE          PowerRating       FGPCT_O         FGPCT_D         TURNPCT          TORD      
 Min.   : 9.00   Min.   : 0.00   Min.   : 76.6   Min.   : 84.0   Min.   :0.0050   Min.   :39.2   Min.   :39.60   Min.   :11.90   Min.   :12.0  
 1st Qu.:29.00   1st Qu.:11.00   1st Qu.: 98.0   1st Qu.: 98.5   1st Qu.:0.2683   1st Qu.:47.8   1st Qu.:48.10   1st Qu.:17.30   1st Qu.:17.2  
 Median :31.00   Median :15.00   Median :102.7   Median :103.4   Median :0.4744   Median :49.7   Median :50.10   Median :18.70   Median :18.6  
 Mean   :30.33   Mean   :15.68   Mean   :103.1   Mean   :103.3   Mean   :0.4894   Mean   :49.8   Mean   :50.08   Mean   :18.79   Mean   :18.7  
 3rd Qu.:33.00   3rd Qu.:20.00   3rd Qu.:107.9   3rd Qu.:107.8   3rd Qu.:0.7106   3rd Qu.:51.9   3rd Qu.:52.10   3rd Qu.:20.20   3rd Qu.:20.1  
 Max.   :40.00   Max.   :38.00   Max.   :129.1   Max.   :124.0   Max.   :0.9842   Max.   :61.0   Max.   :60.10   Max.   :27.10   Max.   :27.6  

      ORB              DRB             FTR             FTRD            X2P_O           X2P_D           X3P_O           X3P_D           ADJ_T      
 Min.   :15.00   Min.   :18.40   Min.   :19.60   Min.   :19.70   Min.   :37.70   Min.   :37.70   Min.   :24.90   Min.   :26.10   Min.   :57.20  
 1st Qu.:26.80   1st Qu.:27.60   1st Qu.:31.60   1st Qu.:31.30   1st Qu.:46.60   1st Qu.:46.80   1st Qu.:32.40   1st Qu.:32.90   1st Qu.:65.80  
 Median :29.60   Median :29.80   Median :35.20   Median :35.30   Median :48.80   Median :49.20   Median :34.30   Median :34.50   Median :67.90  
 Mean   :29.59   Mean   :29.82   Mean   :35.38   Mean   :35.71   Mean   :48.87   Mean   :49.14   Mean   :34.31   Mean   :34.55   Mean   :67.86  
 3rd Qu.:32.30   3rd Qu.:31.90   3rd Qu.:38.90   3rd Qu.:39.70   3rd Qu.:51.10   3rd Qu.:51.40   3rd Qu.:36.20   3rd Qu.:36.20   3rd Qu.:70.00  
 Max.   :43.60   Max.   :40.40   Max.   :55.50   Max.   :60.70   Max.   :64.00   Max.   :61.20   Max.   :43.40   Max.   :43.10   Max.   :81.70  

      WAB             POSTSEASON       SEED          tournament.      power_conf     TFC_elite8.
 Min.   :-25.20   R64    : 208   Min.   : 1.00   Min.   :0.0000   Min.   :0.0000   [0,0]:2189  
 1st Qu.:-12.90   R32    : 102   1st Qu.:17.00   1st Qu.:0.0000   1st Qu.:0.0000   (0,1]:  52  
 Median : -8.00   S16    :  56   Median :17.00   Median :0.0000   Median :0.0000               
 Mean   : -7.66   E8     :  26   Mean   :15.38   Mean   :0.1981   Mean   :0.2129               
 3rd Qu.: -3.00   R68    :  26   3rd Qu.:17.00   3rd Qu.:0.0000   3rd Qu.:0.0000               
 Max.   : 13.10   (Other):  26   Max.   :17.00   Max.   :1.0000   Max.   :1.0000               
                  NA's   :1797                                                                
```

Figure 1

When first looking at our data, we noticed how lopsided it was in terms of our target variable. Out of all the observations, our target variable only accounted for a little over 2%. Furthermore, only around 19% of our data had teams that made the tournament while most of the teams were schools that rarely, if ever make the tournament. Lastly, we noticed that the range for our statistics was quite large between the best and the worst performers. Therefore, it became important to create a model that could weed out the unnecessary data.

To get a better picture at how our categorical variables fit with our target variable, we built a bar graph illustrating the number of Elite 8 appearances for each conference.

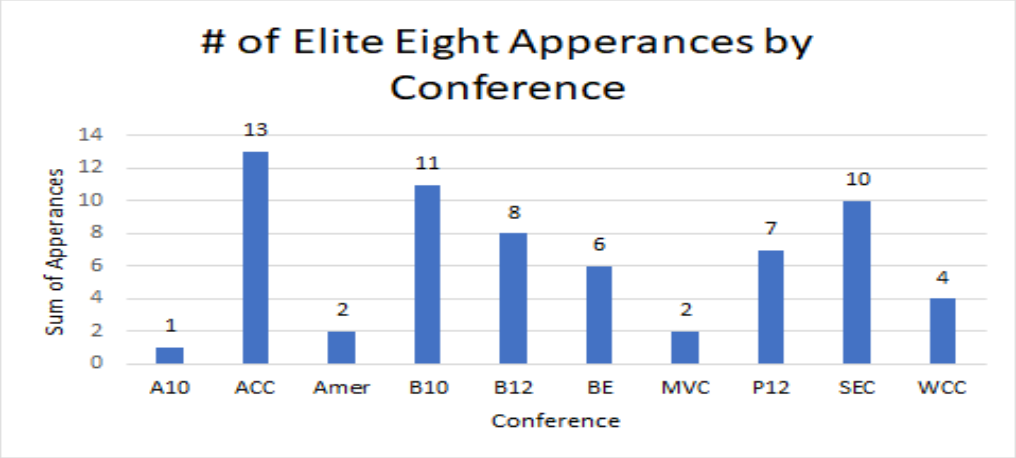

Figure 2

Figure 2 shows that a team's conference has an impact on the likelihood of that team making the Elite 8. Out of the thirty-five conferences, only ten have housed teams that made an Elite 8 appearance since 2013. Of those ten, the five largest conferences, commonly referred to as the "Power 5", have seen their teams make the most appearances. Therefore, we concluded that there is a positive correlation between a team belonging to a power 5 conference and their chance of making the Elite 8.

Two numeric variables that had clear differences between Elite 8 teams and the rest of the field are their adjusted offensive efficiency and adjusted defensive efficiency. See Figure 3 below.

.

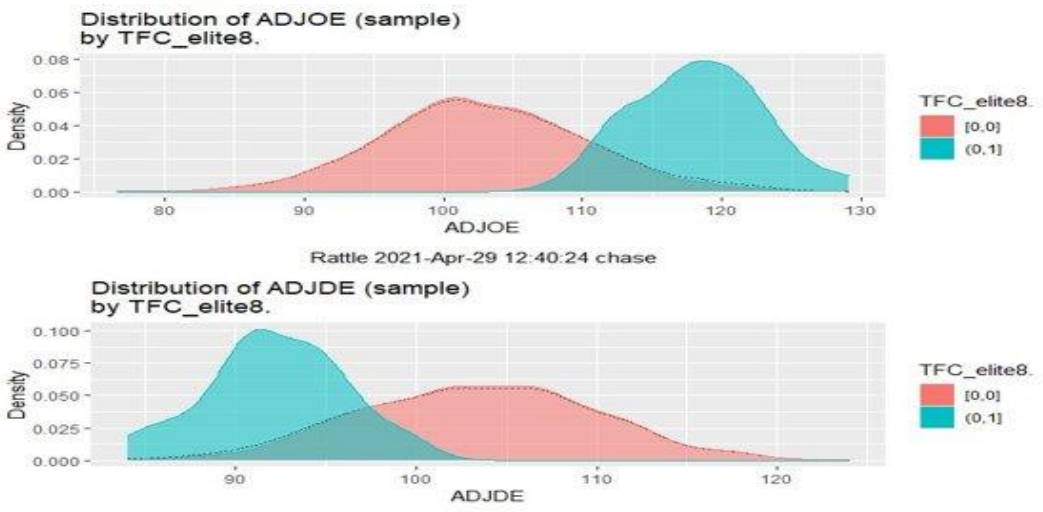

Figure 3

From Figure 3, we can see that being efficient on both ends of the court pays off. Scoring more than 110 points per 100 possessions and allowing under 95 points per 100 possessions greatly increases a team's chance of making the Elite 8.
Not surprisingly, the greatest numeric predictor of making the elite 8 was wins. The number of regular season wins translates well into March Madness success, as elite 8 teams win around 30 times in the regular season, which equates to around an 80%-win percentage. The average number of wins for teams that do not make the Elite 8 is only about 20.

The final step we conducted in our exploratory analysis was to cluster the teams into different groups. This gave us insights on the similarities and differences across observations and showed which features made these clusters unique. After reaping the highest marginal benefit and maintaining as much simplicity as possible, we decided on 5 clusters. The variables that we used for clustering were number of wins, whether a team was in a Power 5 conference, whether they made the tournament, power ranking, and their adjusted offensive and defensive efficiencies. The results of the 5 clusters are below.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Power 5 Teams	Non-Power 5 Teams	Non-Power 5 Teams	Both Power 5 and Non-Power 5 Teams	Non-Power 5 Teams
Sometimes make the tournament	Very rarely make the tournament	Do not make the tournament	Make the tournament	Sometimes make the tournament
Average win rate	Do not win many games	Around 50%-win rate	Win most of their games	Win less than average Power 5 teams
Average to good stats	Awful stats	Good stats	The best stats	Below average stats
Above average power rankings	Extremely poor power rankings	Average power rankings	The best power rankings	Below average power rankings

Data Pre-Processing

Before we could begin training our models, our team needed to make sure our data was processed in a way that allowed these models to be as accurate as possible. We started by removing features from the raw dataset that would cause extrapolation. The variable "postseason", which indicates which round of the tournament a team made it to, is an example of one of these factors. Once we removed these features, there were no missing values in our dataset, and we continued by transforming features. Categoric features were recoded to indicator variables and numeric variables were scaled from 0-1. We kept our

target variable as categoric because the data mining problem we are interested in solving utilizes classification methods. Once we felt that our dataset was ready to go, we partitioned it to 70/15/15 for the implementation and subsequent evaluation of our models.

Data Mining Solution

Models

The classification models we analyzed were neural net, random forest, and logistic regression. Each of these models required an involved parameter-tuning process before we arrived at the results that we felt were the most accurate. Our random forest was tuned to have 135 trees and 6 variables, and our neural net had 10 hidden layer nodes, while logistic regression did not require tuning. The next section talks about how we evaluated each of these models before we were able to conclude which one was the best.

Performance Evaluation

Our team decided to evaluate each model based on AUC scores. Random Forest had an AUC of 0.77, Neural Net 0.93 and Logistic Regression 0.96. We concluded that Logistic Regression would be the best model because it had the most accurate predictions with an AUC at 0.96. Figure 5 shows the confusion matrix that summarizes the results for our logistic regression model.

Confusion Matrix	Predicted (0,0)	Predicted (0,1)
Actual (0,0)	TN 412	FP 0
Actual (0,1)	FN 5	TP 4

Figure 5

Along with AUC we looked at recall and precision using a confusion matrix. We calculated recall = 0.44 and precision = 1.0. Recall is poor because our data is skewed with so little teams making the elite 8. One way to improve recall would be to lower the threshold for predicting positives. Recall is low because precision is high, they have an inverse relationship.

Conclusion

Recommendations

In this study, we applied different data mining methods to analyze which variables influence a team's probability of advancing to the Elite 8. After evaluating our solutions, we determined that logistic regression concluded the best results. Moving forward we believe coaches, players, scouts, and sports bettors can learn from our models.

Recommendations for Coaches/Players:

1. Focus on variables you can control, practice the fundamentals of basketball. Turnover percentage, rebound percentage and offensive and defensive efficiency are factors that coaches/players can control. If they can improve all of these features, they can increase their odds of making the Elite 8.
2. Play every game like it is for the National Championship. By working hard in the regular season, a team can improve their power rating thus increasing the likelihood of advancing to the Elite 8.

Recommendations for Sport's Bettors:

1. Be cautious of using seeding to make brackets and bets. Seeds 1-4 have a history of success, seeds 5-9 have underperformed and 10 seeds have had surprising runs.
2. Instead of focusing on a team's record, analyze a team's wins above the bubble before filling out brackets or making bets. Wins above the bubble is the difference in the number of wins a team has compared to their expected number of wins against an average tournament team. WAB is a better predictor than a team's record when seeing if a team makes the Elite 8.
3. We recommend looking at power conferences too. As stated earlier, power conferences have an overwhelming presence in the tournament. Bettors should take teams in power conferences more than teams that are not.

Limitations

Our models were limited in a few ways. Our dataset included data from the past decade of college basketball. If our data ranged back to 1985 when the 64-team tournament started, our model could have been more efficient. We could have used more variables to give even more interesting insights such as margin of victory, hot streaks, type of offense and defense, and player class (i.e., Freshman/Sophomore/Junior/Senior).