Predicting NBA Playoffs Teams

Angela Chen, Michelle Li, Vera Wang

Abstract

We picked the basketball dataset, which contained information about players, players' stats, team stats, team standings, etc. Our overall goal was to predict which teams would make the playoffs based on the information we had. We explored different statistics such as player demographics, how well players did the NCAA/NBA, team statistics, etc. in order to build a model that can accurately predict which teams will make it into the playoffs. We built a logistic regression model in order to make the predictions, and also used a decision tree.

Introduction

The first thing we did was clean the data (which will be explained in the section below). After cleaning the data, we wanted to explore the data a little bit more to see if we could find any interesting trends or patterns. We thought of the data as having two main parts: individual player statistics and team statistics. We wanted to see which one of these parts (or possibly a combination of these parts) could bring us more information about if a team would eventually make it to the playoffs or not. Within the individual player statistics section, we also wanted to explore a little bit about the difference between college level basketball and professional level basketball. For example: do good NCAA players become "better" NBA players, which ultimately leads them to make the playoffs? Is there a significant difference between NBA players who are in some of the top scoring teams and the lower scoring teams? What specific gameplay patterns are there amongst teams who make the playoffs (i.e, do they have a possession time in games played)? These are some of the questions we framed our EDA and model around.

Cleaning and EDA

The first thing that we did was clean our datasets. This is the way we approached cleaning each dataset:

1. College

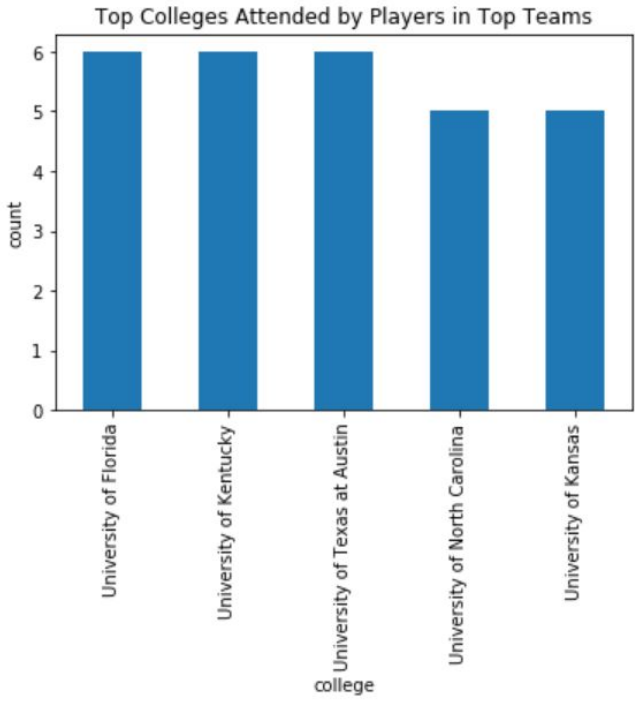
-We removed the Unnamed column because it was the same as the index.

-We wanted to keep the years consistent, so we only used information from 2012-2018.

-We renamed all the inconsistent column titles between NCAA and NBA. For example, NCAA games played was labeled "NCAA_games" and NBA games played was labeled "NBA_g_played", so I renamed it to "NBA_games" to better easily understand the data.

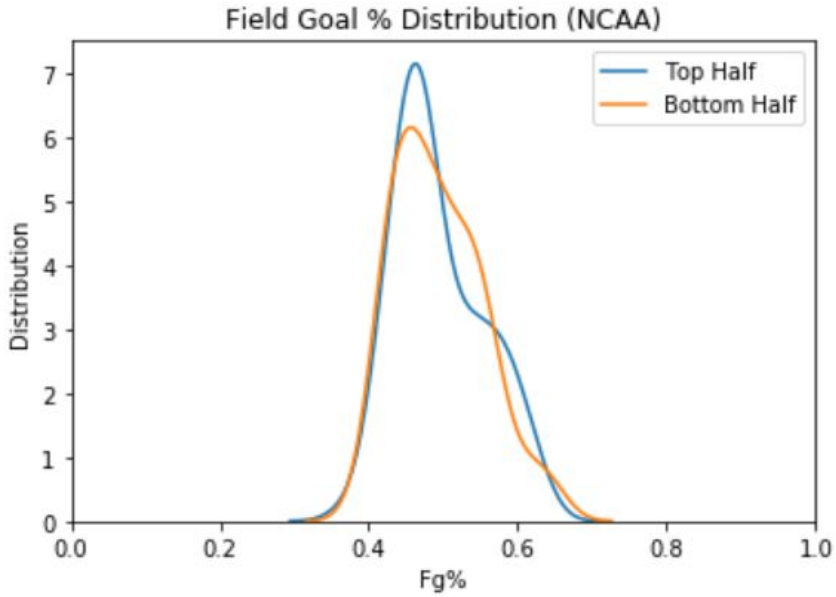
2. Standings
-We removed rankOrd because it was a repetition of rank.
3. Player box score
-We renamed the columns for clarity
We did not change the box score dataset or team box score dataset.

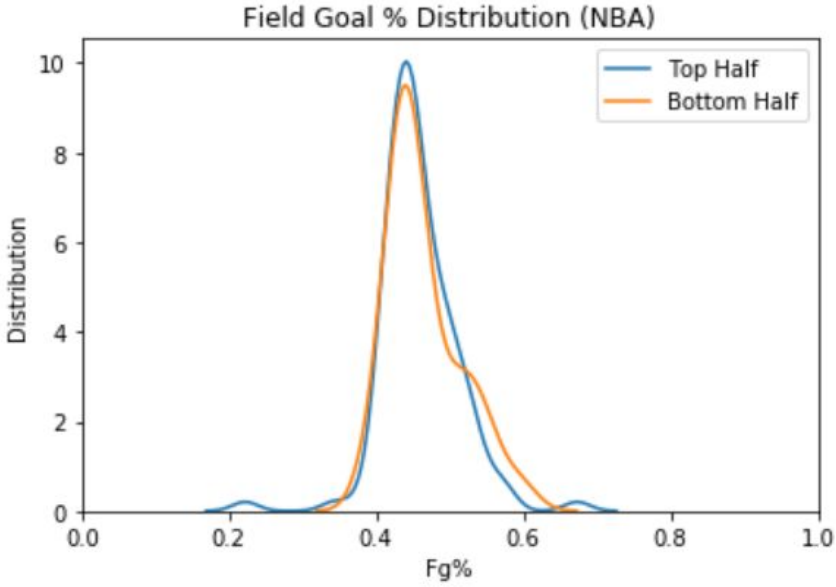
After we cleaned our datasets, we wanted to see if there was a major difference between players in the top teams vs. other teams. We split the dataset into two groups based on the total number of wins each team had from all six seasons (30 teams, so top 15 with most wins from all six seasons and bottom 15 with least wins from all six seasons). To get the total number of wins, we took the data from the last day of each season from the Standings dataset and added up the wins from each season for each team. Number of wins does not necessarily equate to a "successful" team, but we thought this was a relatively fair assessment because playoff qualifications are based on the number of wins within a conference. Since each team played the same amount of games per season, we just compared the raw amount of wins. Then, we merged the player box score and college datasets so we could see each player's information, how they did in college and the NBA, and which team they ended up on. We first looked at the most common colleges that NBA players in the top teams went to.



This was just out of interest and to better understand our data. These colleges are all D1, which makes sense in the context of our dataset.

Next, we looked at the field goal percentage distribution to see if it was higher for players in the top teams or bottom teams. We looked at the field goal percentage both within the NCAA and NBA. Perhaps there isn't a significant difference when the players are in college, but once they enter the NBA, players from top teams will have a significantly higher percentage.

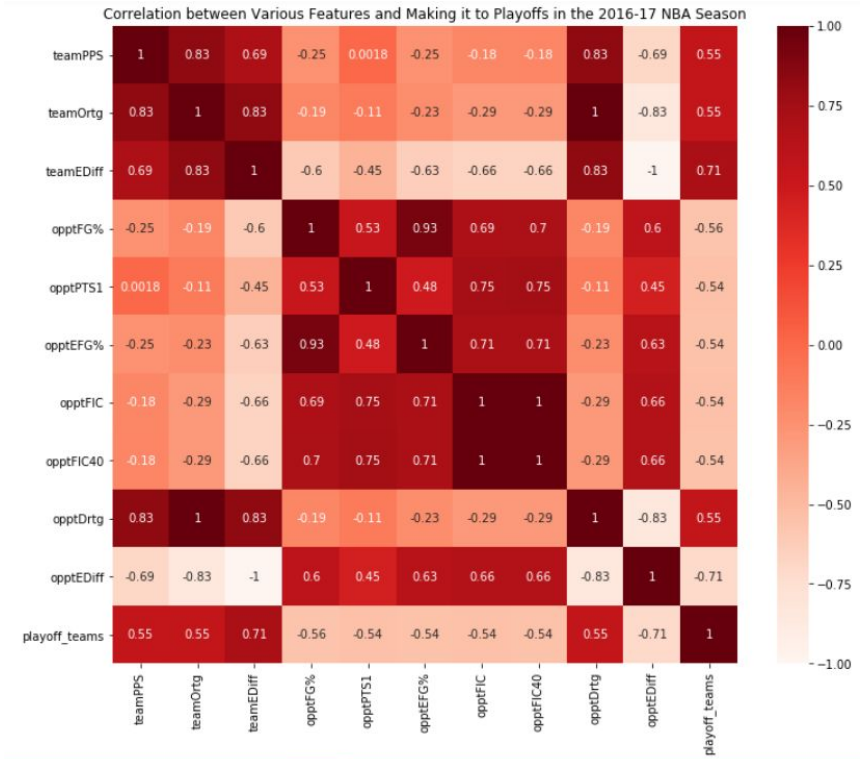




We found that there was not really that much of a difference in the player statistics in general. From there, we decided to focus more on overall team statistics instead of individual player statistics.

Description of Methods

We wanted to determine which features correlated with if a team made it into the playoff for each season. First, we filtered the data based on season, standardized it to get rid of any potential bias that could affect the correlation between a feature and making it to playoffs, created a one hot encoding for if each team made it into playoffs, and then found the features most highly correlated with making playoffs. While analyzing features, we used the 2016-2017 season to find correlations between the features and making playoffs and plotted the features with the highest correlations in a heatmap.



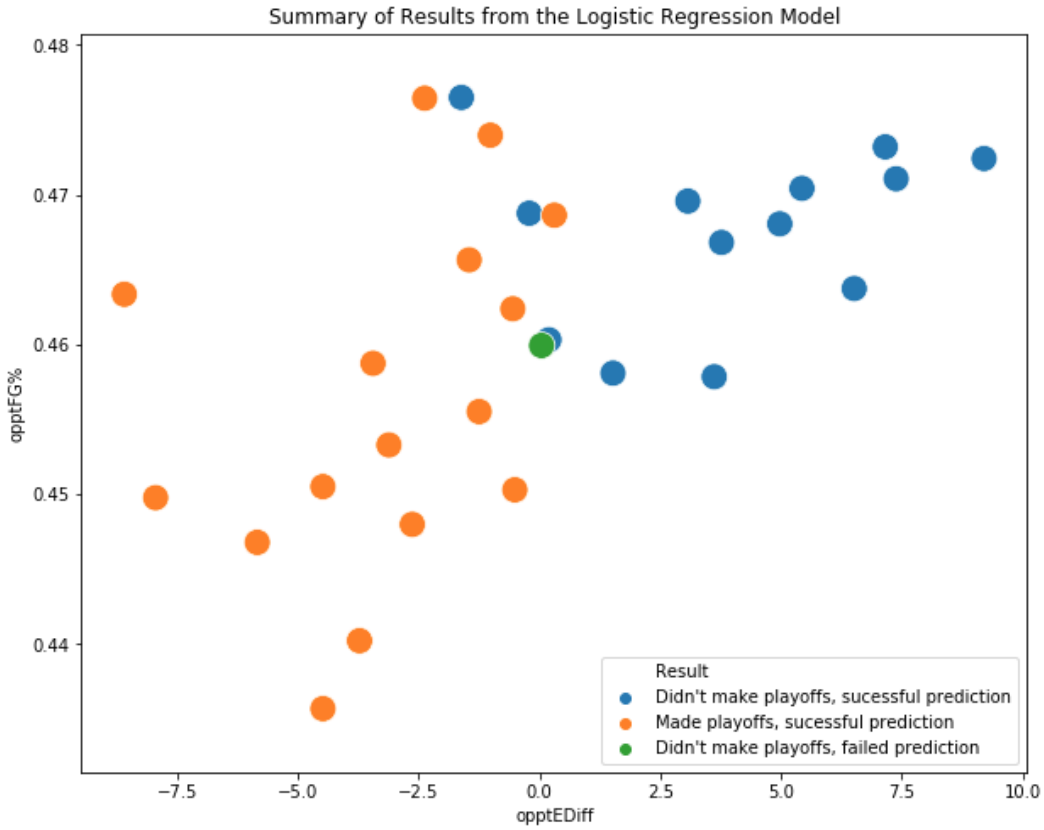
Description of Model

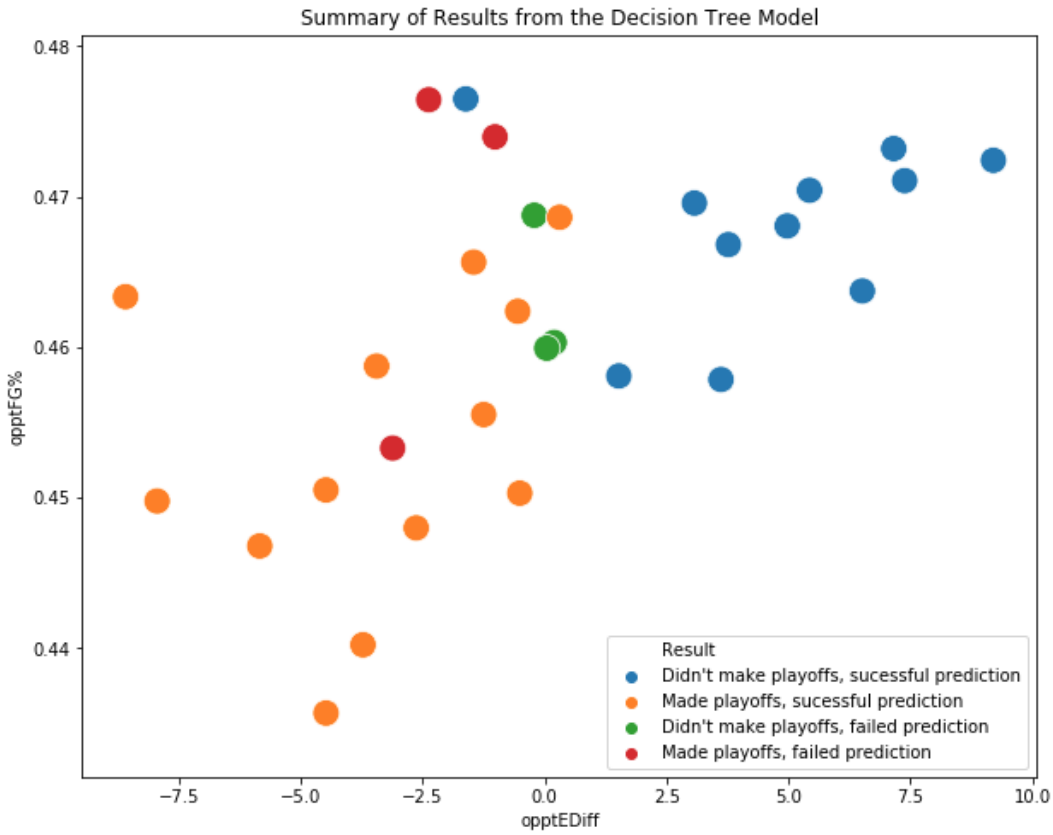
Our goal was to determine the common stats of playoff achieving teams from games over the season. We attempted predicting the playoff teams for each season with two different types of models: logistic regression and decision trees. We used a logistic regression model because we want to classify the teams as either making it to playoffs or not so we thought a logistic regression model would be good for this kind of classification. We used a decision tree model because we wanted to use another model to test our classifications and see if the features we chose were actually good indicators of making it to playoffs. We used seasons from 2012 to

2017 as training data to fit our models, and validated our accuracy on predicting the 2017-18 season.

Summary

Here are diagrams for how accurate our models were in predicting if a team made playoffs. Both models were fairly successful in predicting teams. These graphs show that the logistic regression model might've been more accurate than the decision tree model because of overfitting.





Followup Questions

(i) What were two or three of the most interesting features you came across for your particular question?

The two features with the highest correlation on if a team made playoffs were field goal percentage, and efficiency difference between the teams. The efficiency difference is the difference between a team's offensive and defensive ratings which are based on the number of points generated by a team when they have possession of the ball.

(ii) Describe one feature you thought would be useful, but turned out to be ineffective.

We thought looking at individual player stats could be useful in predicting team success within the season. However, we found that in general, there wasn't a significant difference in performance among individual players in better and worse performing teams, so we looked at overall team statistics instead to see if there were any common factors among better performing teams.

(iii) What challenges did you find with your data? Where did you get stuck?

One of the first challenges we had with our data was understanding the Standings dataset because we knew using this dataset would be necessary to not only see which teams performed the best season to season, but also to know which teams actually made it to the playoffs each

season. At first, we were confused and had trouble using the Standings dataset to collect the number of wins per team to figure out the best performing teams throughout the six seasons in the dataset. We didn't realize that the data in Standings got carried over to each day even if there wasn't any change to the data (probably due to a team not playing a game on that particular day but the previous day's data carried over to that day). We then realized that to get the data that we wanted, we had to take the data from the last day of the season for each season.

(iv) What are some limitations of the analysis that you did? What assumptions did you make that could prove to be incorrect?
One of the limitations is that we generalized a few of the findings. For example, when looking at the differences between individual players in better and worse performing teams, we did not use every single statistic to see if there was a significant difference, and instead just looked at a few statistics. This definitely could have been more concretely done. It is possible that if we combined all the statistics, there would be a significant difference in individual players in better and worse performing teams. This does not ruin our model, however, we could have possibly used some of these statistics to better our model's accuracy.

(v) What ethical dilemmas did you face with this data?
There were not any ethical dilemmas that we explicitly faced, but some possible ethical dilemmas with basketball datasets is that there might be bias in NBA drafting, exploitation of college-level athletes, or gender equity within the sport.

(vi) What additional data, if available, would strengthen your analysis, or allow you to test some other hypotheses?
Since we focused on a small subset of years, it would have been great to have additional data from multiple years. Having more data could probably make our model more accurate. Also, it would be great to have more playoff/championship information about who actually won and made it to the championships each year, because we currently have to manually find that information for each year.

(vii) What ethical concerns might you encounter in studying this problem? How might you address those concerns?
An ethical concern when studying NBA data is bias in drafting players. One way we can address this concern is making note of players who performed poorly in the NCAA or were not college players to see if there are special circumstances for their recruitment.

Evaluation of Approach and Limitations

Starting with the EDA, there were a few issues that we had in our approach. One issue is that when we merged the tables for all the player information, we had to drop duplicates. Some players were on multiple NBA teams, so their name showed up multiple times. The reason we dropped duplicates was because having duplicates would place more weight on certain people. For example, for our college attended chart, if we did not drop duplicates, a player who attended Duke would show up twice, although they should only be counted for once. This presented an issue later on, because when we were categorizing players into top teams and bottom teams, they might have been in both groups when we only counted them for one. However, this seems to not present a huge issue, because there was not much variation between any players in general. Another limitation is that we only showed one statistic (field goal percentage). We did look at other statistics through distribution plots as well, but they followed a similar pattern (very little difference between top players and bottom players). We did not include them because we did not want to bulk up the notebook and we thought the pattern would be repetitive.

For the models that we chose to use, we think that the logistic regression model was appropriate, but was not tested on enough data. As for the decision tree, the decision tree had a high training accuracy but had a lower test accuracy. This is because the decision tree slightly overfits the training data. We still believe that these were good models, but we probably could have made our models a bit more complex (i.e., include non-numerical features, since we only used numerical).