Ian Murphy
BrainStation, Capstone Report
12/17/20

NBA: What Does it Take to Make the Playoffs?

In the modern era of professional sports, analytics and data science have become an integral piece in the formula for success. Owners, general managers, executives, and coaches are using advanced statistics and analytics to make decisions that have massive impact their team's performance. In my project I aimed to determine what statistics are most predictive of a team making the playoffs. What are the key factors that influence a team's success during the grueling 82 game NBA season? This question can be answered with data science techniques by creating a machine learning model with advanced metrics and then determining which of those metrics have the most impact on if a team made the playoffs.

Executives and coaches want to know where they need to focus their attention. They want to know what they can do during the off season to improve their chances next year. By pinpointing the factors that heavily influence success, executives will have a clear roadmap of what moves need to be made to better their team.

For this project I pulled data from basketball-reference.com. Basketball Reference contains data from every team and player in the NBA dating back to the 1940s. My data set will consist of player and team statistics. Each data point will represent a team and year and the target variable will be if they made the playoffs or not. If the value of the 'Playoffs' column is 0 that team did not make the playoffs that year. If the value is 1, that indicates the team did make the playoffs. I will have a 'Year' and 'Team' columns so I can identify what team and year each row represents. The player specific stats are Player Efficiency Rating (PER) and True Shooting Percentage (TSP) for each of the top 12 players on every roster. The team level stats are Simple Rating System, Offensive Rating, Defensive Rating, and Pace (a full data dictionary can be found in the attached Jupyter notebooks).

A large part of this project was getting the data ready for modeling. One challenge I faced was needing a small amount of data from over 1000 different URLs. To work through this, I created a for loop using lists of team abbreviations and years. I imported 'requests' and used 'pd.read_html' to get the tables. Some index slicing was necessary to get the data that I was looking for. After transposing and concatenating various data frames I had the information, and it was time to start cleaning the data. Cleaning involved trimming and splitting the URL to create the 'Team' and 'Year' columns. There were also some extra years that ended up in the data set that were throwing off the results when I concatenated various data frames. I investigated this and dropped the necessary rows. I feature engineered 5 new features: 'Division', 'Conference', 'AVG_PER', 'AVG_TSP', and 'Location'. Division', 'Conference', and 'Location' were engineered by building a dictionary and then using the map method. Conference did not have an impact on the outcome because the same number of teams (8) make the playoffs from each conference. Division is much different because the division a team is in affects their strength of schedule and there is not a specified number of teams from each division that can make the playoffs (in theory it could be all or none). 'AVG_TSP' and 'AVG_PER' are the overall averages of the teams TSP and PER. This will allow the model to investigate what level of impact the overall PER and TSP have on making the playoffs.

When doing exploratory data analysis (EDA) I used Tableau to create some insightful visuals. The comparison between PER for teams that made the playoffs and teams that did not make the playoffs was incredibly insightful. When looking at team average PER, there was little difference, but when comparing the average player 1 PER for teams that did and did not make the playoffs the difference was much greater. This indicates that the star player for each team is a huge factor on if they make the playoffs or not. The AVG player 1 PER of a team that made the playoffs was 20.2. This is actionable in that if a team is looking for a star player in free agency, they should target players who had a PER of over 20 in the previous season. Division also seemed to have an impact on making the playoffs. Teams in the Southwest Division made the playoffs 61% of the time and teams in the Northwest Division made the playoffs only 47% of the time.

Modeling using the data I collected proved extraordinarily successful. I used GridSearchCV to determine the best model and hyper parameters. The model with the best performance was Logistic Regression, with C = 1, using MinMaxScaler, and penalty = 'l2'. To evaluate the model, I produced a confusion matrix and a classification report. Precision and Recall were both around 90% as was the F1 score. These results made me feel extremely comfortable that the coefficients produced by the model would be great indicators of what features impact a team's success.

The coefficients produced gave some key insights to pass along to NBA executives. First and foremost, offensive efficiency is the biggest driver of success. In my final model ORtg and Relative ORtg were the highest positive coefficients. This tells me that during practice the bulk of time should be spent on offensive sets and plays. Building chemistry on offense should be the number one goal. Individual coaches were present in both the top 10 positive and negative coefficients. It was clear that having the right coach was a key to making the playoffs. For executives searching for a new coach, I suggest targeting coaches like Eddie Jordan, Larry Drew, or Vinny Del Negro. These 3 coaches have a great track record, and had high coefficients indicating they will help a team make the playoffs. The top three players on each team are vital to success. Many teams have two great players but having a third player that has a PER above league average will propel a team to the playoffs. I wrote about the importance of player 1 in my EDA, but the coefficients indicate that player 1, 2, and 3 each have a significant impact on success. In conclusion, I found three major insights that could benefit NBA executives. Offense efficiency rules, coaches' matter, and focus on getting three-star players.