NBA Game Outcome Prediction

Capstone Report

The National Basketball Association, more commonly known as the NBA, is a professional basketball league in North America that comprises 30 teams that represent different cities in the United States and Canada. Every regular season, which spans from months October to April, each team plays 82 games with the purpose of developing a ranking for who will get the chance to compete for that season's championship during the months May and June. The ultimate metric for success in this sport is winning as it is required to have more points than the opposing team to win. The catch is that there are an enormous amount of variables that come into play that can make predicting which team will win any given game a great challenge. The purpose of this analysis is to investigate the abilities to predict the winners and losers of each game during the regular season based on team and player statistical attributes with machine learning.

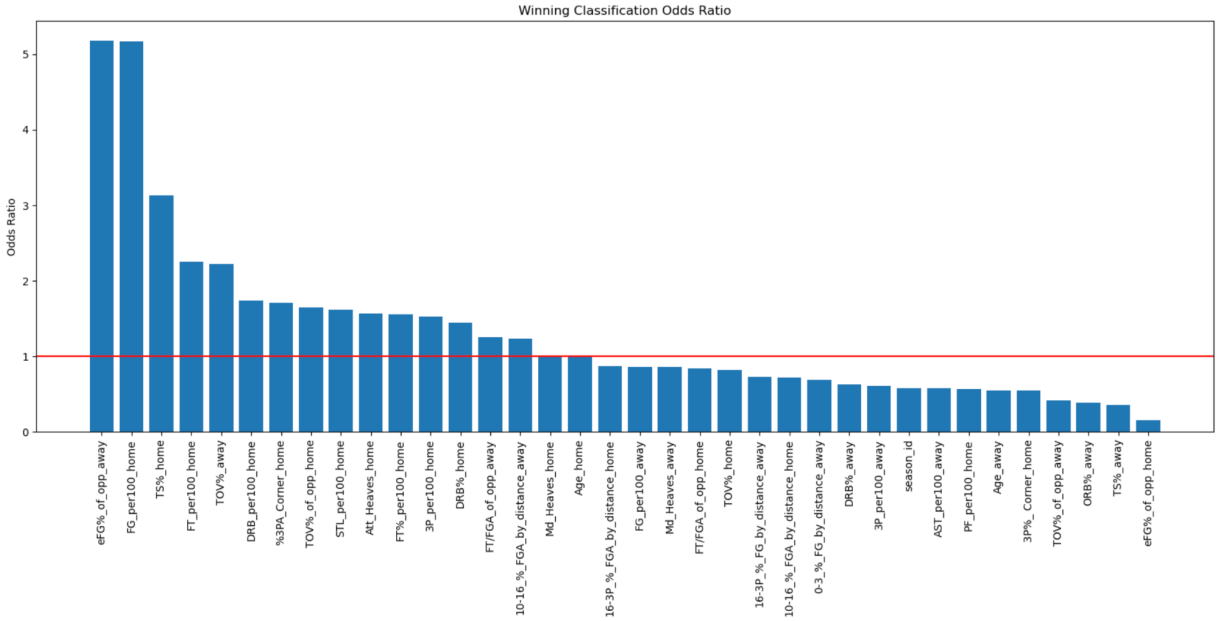
The ability to predict wins with accuracy can provide insight on what attributes lead to success in the NBA. Implications to this can be to assist in team and player development because they will be able to focus on improving parts of their team that more directly lead to improving success. In the past, insights and strategies in the NBA were mostly developed using the “eye test”, meaning things were always done based on intuition and visual observations of the game. As technology has advanced creating more ways to track players, more data has become available to describe the game. Since everything in basketball is quantified using statistical metrics, data science can leverage these metrics to develop insights and summaries that are not very visually obvious.

The data necessary for this analysis came in the form of several organized tables from basketball-reference.com and ESPN.com. After every game, data is uploaded to the site so that it always has the most up to data statistics from every game. I took a two sided approach to try to achieve the goal of predicting games. The first approach was to use team level statistics to try to predict games, and the second approach was to use individual player level statistics.

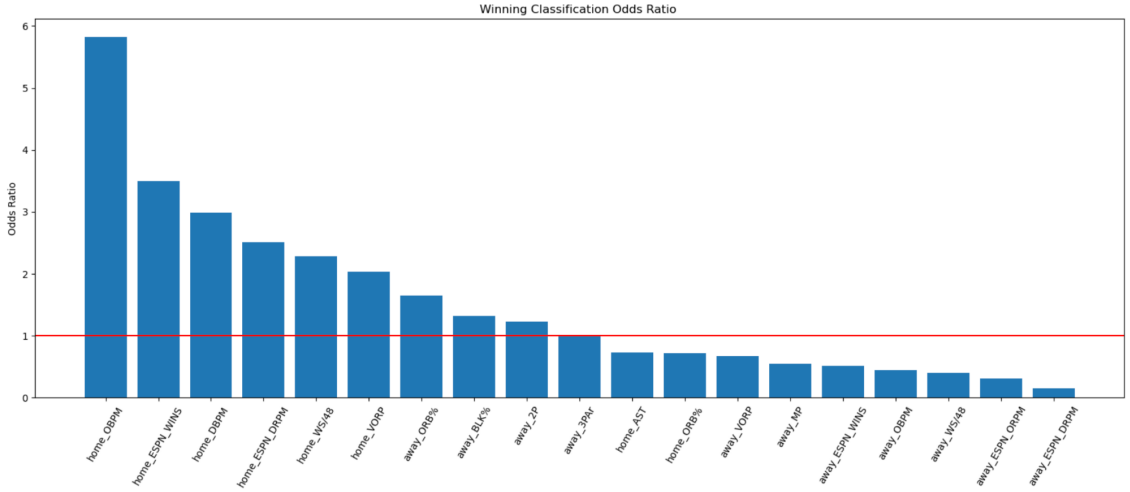
The team level statistics used three data sets that were scraped and imported and provided “advanced” statistics about each team, how each team performs per possession, and statistics on how each team shoots the ball. The data sets did not contain any nulls or require much cleaning, but some elements of the data needed to be wrangled for purposes of the analysis. Each data set had to be processed separately because each data frame contained different types of information. Overall, the names of each player were treated to eliminate any special characters, and highly correlated columns were removed to reduce multicollinearity in each individual data set. Once this was done, all three data frames were then merged together, making a new bigger data set for modeling. The player level statistics data sets were combined in a similar fashion to the team data sets, and collinear columns were also removed from each data set and then the data sets were combined together.

Apart from the team and player data sets, another data set was required that had the schedule for every game and the result of that game, which is the target variable. Focusing on the results of the home team, it was found that the classes were not balanced. 59% of the outcomes were wins for the home team and 41% were losses. This imbalance was treated by upsampling the minority class, which were losses, using SMOTE.

Once the data sets were acquired for modeling, a strategy was put in motion for how predictions would be done. First the team data was modeled using a logistic regression using all the features to get a baseline performance. This achieved an accuracy of 67% using 97 features. From here lasso regression was used to try to reduce the dimensions while minimizing model performance loss. This led to a reduction in features from 97 to 34 with one percent lost in testing accuracy making it 66%. The odds ratio detailed that the most predictive feature for winning comes from how good your opponent is able to play defense, and how many baskets your team can make per 100 possessions.



Since players make up teams, this drove the motivation to take things a step further to try and build predictions from a player level. This approach required more domain knowledge but I suspected that it could potentially provide better results. The idea was to look at a match up, and instead of comparing the teams stats, to go a level deeper and to extract the players that played in the game and find a way to aggregate the stats of the players on each team to try and predict a winner. This process ended up successfully improving training and testing accuracy compared to using the team data. The performance was boosted to 70% training and 69% testing accuracy. The top performing features are outlined below in the odds ratio bar graph.



Overall this analysis was a success because the achieved results are on par with results in the industry and the most important features for winning have been outlined. Future steps would be to explore different combinations or play stats aggregations to try to achieve better results, as well as exploring different models. Another step I would also like to try is to try to generate predictions on final game scores on top of the win/lose prediction.