Introduction

For my final project, I wanted to construct a model that can accurately predict wins for NBA teams using advanced stat metrics. By identifying which metrics are positively correlated to wins, NBA team owners will be able to identify which areas of their teams they should focus on with a goal of maximizing their win total.

The goal of every NBA owner is to make the playoffs, win a championship and essentially increase the value of their brand. Making the playoffs is a lucrative proposition for team owners due to the media exposure these additional games provide and it qualifies them to be included in a playoff revenue sharing pool. This extra pool of money is split between all the playoff participants, so just making the playoffs is quite the reward.

The NBA is made up of 30 teams, 15 from the Eastern conference and 15 from the Western conference. In a typical NBA season, these teams play 82 games in the regular season and the top 8 teams from each conference at the end of the regular season get to advance to the NBA playoffs. Going back to the 2014-2015 season, the average amount of wins a team in the Western conference would need to advance to the playoffs is 44.4 wins. A team in the Eastern conference has needed to have at least 41.4 wins on average. I excluded the 2019-2021 seasons in this average due to shortened seasons because of the coronavirus.

|  |  |  |
| --- | --- | --- |
| **NBA Season** | **Western Conference** | **Eastern Conference** |
| 2014-2015 | 45 | 38 |
| 2015-2016 | 41 | 44 |
| 2016-2017 | 41 | 41 |
| 2017-2018 | 47 | 43 |
| 2018-2019 | 48 | 41 |
| Average | 44.4 | 41.4 |

*Figure 1.1: Average wins needed per conference to qualify for post-season.*

Data Source

The data for this modeling was pulled from <https://www.basketball-reference.com/>. That was determined to be the best source for the data due to the amount of data they contained and the ability to filter the data we need for our analysis. Basketball Reference allowed me to export the data into a CSV file that could easily be read into Jupyter. After cleaning up the data set and removing variables that did not have anything to do with the on-court product (attendance/venue), I was then able to select the features I was going to include.

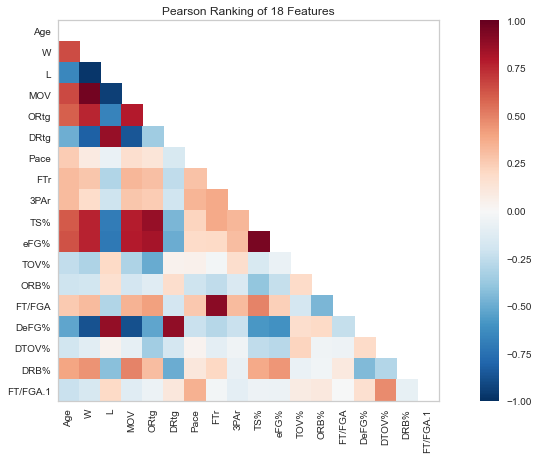
The following are the categories used in the modeling:

* MOV: Margin of Victory
* ORtg: Offensive Rating
* DRtg: Defensive Rating
* Pace: Pace of Play
* FTr: Free Throw Rate
* 3PAr: Three Point Attempt Rate
* TS%: True Shooting Percentage. Shooting efficiency metric that factors in 2pt, 3pt, and Free Throws.
* eFG%: Effective Field Goal Percentage. Factors in that 3pt field goals are worth more than 2pt.
* TOV%: Turnover Percentage.
* ORB%: Offensive Rebound Percentage.
* FT/FGA: Free Throws per Field Goal Attempt.
* DeFG%: Opponent Effective Field Goal Percentage.
* DTOV%: Opponent Turnover Percentage.
* DRB%: Defensive Rebound Percentage.
* FT/FGA.1: Defensive Free Throws per Field Goal Attempt.

Model Creation

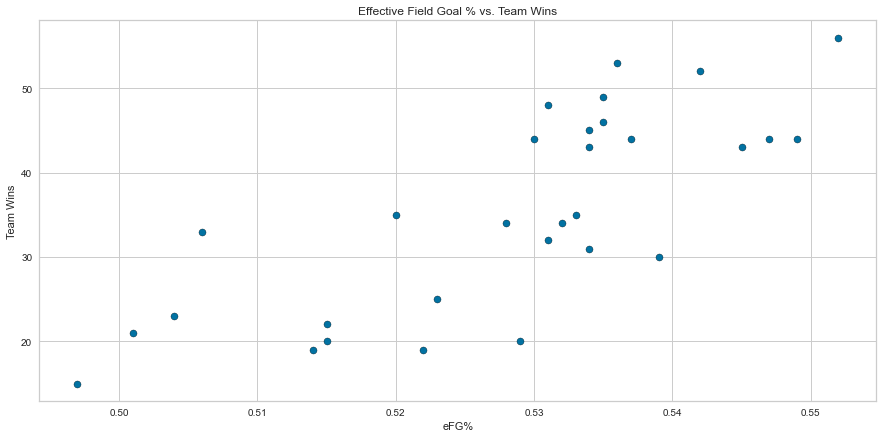
I wanted to include defensive metrics as well due to the previously established notion that defense wins championships. By including that data, we will be able to identify how closely defense correlates to actual wins in the win column. After identifying the features that were going to be used, I utilized the Yellow Brick library to create Pearson’s Ranking Visualizer to see what variables had a strong and positive correlation with Wins.

Pearson’s Correlation Visualizer:



*Fig 2.1: Pearson’s Ranking of 18 Features.*

As can be seen from Pearson’s Correlation Visualizer above, the defensive categories didn’t have as much of an impact on wins as offensive categories do. On the other hand, offensive categories such as offensive rating, effective field goal percentage and a team’s true shooting percentage were shown to play the biggest role leading to wins. This change in philosophy from defense to offense can be due to a couple of different reasons. The NBA has initiated a hand check rule that has made it harder for defenders to guard offensive players on the perimeter as close as they would like. The rise of analytics has also changed the game of basketball by exemplifying how much better value a 3-point shot is than a 2-point shot. That point is validated in Fig. 2.1 and 3.1, with how positively correlated a team’s eFG% is to the Wins metric.



*Fig 3.1: Scatterplot of Effective Field Goal Percentage vs. Team Wins.*

Based on the linear relationship evident in the various plots that were completed and because I am attempting to predict a continuous variable, it made sense to use a linear regression model. The data was split into an 80/20 ratio and then was fitted to the training data. Once that was completed, I was able to obtain a coefficient of determination score of 98.8% on the training set and a score of 92.3% on the testing set. I then found the mean absolute error to be 2.7, the mean squared error to be 10.4 and the root mean squared error to be 3.2.

Challenges

While the data is clear about what has contributed to Wins the most with the features used, there are some items that need further consideration. The model was trained using data from the 2020-2021 NBA season, which was shortened to 72 games from the usual 82 games. Those 10 games might not have a drastic impact on the data, however accumulating more data would have been more ideal and can validate our conclusions even more. Another item that hasn’t been taken into account is injuries. Injuries are a part of sports and can’t be predicted. If a star player on one of the top teams were to go down for a long period of time, that would skew the data and lead to statistical anomalies. Ways to improve the model would be to include data from more than one season to develop a larger training and testing data set to pull from. I would also be interested in constructing similar models with just playoff statistics to see if the features we identified to be significant to Wins would hold up in the playoffs.

Conclusion

Based on the coefficient of determination of the testing set being close to the coefficient of determination of the training set, I feel confident in recommending this model as one to use to help business owners predict Wins based on how their team is trending. That validation is backed up with the mean absolute error, mean squared error and the root mean squared error being relatively low. With this model, team owners will know the barometer of wins that they need to reach, depending on their conference, in order to make the playoffs. They will then be able to scour the market to see what individuals can be signed or traded for that will boost those statistics, which in turn should lead to more team wins. Having any sort of model that can target a company’s weakness is essential because it exemplifies ways the business can continue to improve.

References

“2020-21 NBA Season Summary.” *Basketball*, [www.basketball-reference.com/leagues/NBA\_2021.html](http://www.basketball-reference.com/leagues/NBA_2021.html).