Report Notebook

July 1, 2020

Key Takeaways

Our analysis aimed to generate and interpret an efficient model of the summary season statistics for 36 NBA teams ranging from the 2003-2004 season to 2018-2019. There were two kinds of variables, play frequencies and success percentages. We modified the play frequencies to account for seasons with shorter game play by scaling our variables down to an individual game level. After extensive exploratory data analysis, we discovered that the season statistics indicated a close game score on average. This led us to try the average points per game metric as the response variable in our initial models. Our final model included all of the play frequency variables scaled to the average game level. We discovered that less direct metrics, such as assists and rebounds, strongly influence our model predictions. Future research would explore the possibility of including more direct variables in our model and the implications of our coefficients.

put graph here

Introduction

Necessary Code Components and Data Description

To generate this model, we used 7 R packages. These included *GGally*, *ggcorrplot*, *tidyverse*, *glmnet*, *broom*, *coefplot*, and *nbastatR*. The National Basketball Association (NBA) released our data set of 479 observations via the NBA stats website. It is a comprehensive summary of regular season performance across 36 NBA teams ranging from the 2003-04 season to the 2018-19 season. Originally, it included 28 variables consisting of stats such as free throw percentage, win percentage, point accumulation, attempted field goals, and other in-game stats. As we progressed, we mutated 14 more variables to represent statistics on a game level rather than across the season to improve interpretability. The list of variables included in our consideration may be found in the table below.

Explanation of Variables

put table here

Exploratory Data Analysis

The first step in our research was to identify candidate response variables. We decided upon average score differential and average points per game. Next, we viewed the distribution of our intended response variables. The histograms resulted in the distributions being normal enough for us to feel confident in using either of them as our final response variable. We checked the distributions of our other variables as well which produced relatively normal results. After discovering multiple variable distributions with somewhat concerning tail width, we followed up with outlier testing. An outlier would have been classified as a point outside of the bounding created by extending the benchmarks of the first and third quartiles by the quantity of one and a

half times the interquartile range towards the extremes in each direction. We created a scatter plot to arrive at a better understanding of the types of games within the data set through the lens of average points per game and average score differential. Three categories of games appeared: those with a drastic win or loss, an average win or loss, and close games. A game that was won or lost by 9 or more points on average was considered a Big Loss/Win, between 3 and 9 points on average was considered a Regular Loss/Win, and a difference of under 3 points was a Close Game.

`scatterplot goes here`

Modeling

When selecting the subset of variables for our final model, we ultimately chose the average points per game value over the average score differential as our response variable. The number of points in a game is easily interpretable for coaches, and the rules of basketball dictate that success requires a team to maximize the number of points they score. Our beginning assumption was that variables measuring percentage of successfully scoring plays such as field goal percentage and free throw percentage would be positively correlated with average points per game. We used correlation matrices to confirm our assumption and found very high linear correlation coefficients between those statistics. To uncover more interesting trends, we considered less direct variables such as the counts for assists, fouls, rebounds, steals, blocks, and turnovers.

insert error and diagnostics here

Results

All of the variables used in our model are on a frequency scale, so we can consider the magnitude of the coefficients relative to each other when interpreting our results. The coefficients of our linear model appear in the table below.

insert coefficient table here

Turnovers, blocks, and blocks have a negative relationship with the number of points per game. This negative correlation may be attributed to the fact that an instance of a turnover, block, or block against a player would indicate that the team failed to score a basket on a possession or shot attempt. In our model, assists and rebounds increase points per game the most drastically, with steals not far behind. By definition, an assist automatically precedes a basket. Thus, the mere definition of the play supports its prominence as a key predictor of the average points scored in a game. Defensive rebounds and steals force a switch in possession, which allows the team an attempt to shoot and score more points. Offensive rebounds, in contrast, allow for another shot attempt on a current offensive possession. Fouls have the least effect on our model, because it is uncertain that they will lead to a basket. If they do, it is only worth 1 point.

Conclusion

Our final model helped us develop three important conclusions. We observed that all of the frequency metrics in our data set contribute to the accuracy of point predictions when scaled down to the individual game level. In order to produce the most consistent model, we rejected models that took solely offensive, defensive, or percentage statistics into account. This approach aligns with the tactic of combining offensive and defensive strategies in gameplay. We also recognized that collinearity is a prominent hazard in basketball data. Collinearity concerns arose when looking at field goal percentage, two point percentage and three point percentage and again with offensive rebounds, defensive rebounds, and total rebounds. Finally, we noted that close games are fairly common in the NBA. Extensive exploratory data analysis supported this assertion and motivated the choice of average points per game as a response variable. The perspective of average points

per game over the course of a season benefits players and coaches, as the insight that one play could be the difference between a win and a loss can shape gameplay. In the future, we hope to consider new models that include more direct game statistics such as field goal percentage, free throw percentage, and win percentage. We would also like to further our understanding of the implications of some of our coefficients on the current model, such as the negative value for average blocks against players on the team.