Mike Fratoni

Adam Stocker

Yubo Diwu

**CE263 Final Project: NBA Season 2014/2015 Standings Prediction**

***Introduction/Methodology:***

For our final project, we decided to attempt to predict the rankings at the end of this current 2014/2015 NBA season based on past player and team statistics. We went about this by ultimately predicting number of wins over the season and did this by executing the following steps:

**Step 1:** Run a correlation and linear regression on cumulative team statistics to determine what statistics were significant to wins and what weight each significant statistic had on number of wins.

**Step 2:** Predict number of minutes played for each individual player based on current season minutes played (roughly 22 out of 82 games played so far).

**Step 3:** Run a regression on individual player statistics from (up to) the last 5 seasons to determine what their expected 2014/2015 season output will be for their projected minutes played over the season. Then aggregate all individual player outputs by team, and total these team statistics.

**Step 4:** Plug these totaled team statistics into our regression model in step 1 to determine what the win output will be for each team in the 2014/2015 season and compile predicted standings based off these win counts.

***Analysis Details:***

**Step 1:**

In our first analysis step, we took a look at which statistics were most correlated with wins for the past season (2013/2014). We only looked at the last season here since it’s the season likely to have the most similar type of play and therefore the most similar significant statistics as the current season. From the correlation matrix, we took the 5 statistics most highly correlated with wins. The correlations for these 5 statistics are shown in the table below.

*Correlation results*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Statistic** | **FG%** | **2P%** | **3P%** | **PTS** | **DRB** |
| **Correlation** | 0.67 | 0.60 | 0.53 | 0.52 | 0.49 |

We see that FG% is most highly correlated with wins, while the next is 2pt FG% down to the lowest of the five which is defensive rebounds. This makes intuitive sense, shooting (related to points) is well known to be the most important statistics in basketball, and defensive rebounds are also highly correlated since regaining possession after an opposing team’s shooting attempts keeps points against lower and allows you to shoot more yourself.

After deciding these five statistics to use, we set up a linear regression model (done using R) to predict number of wins given these five statistics. Again, we used data from the 2013/2014 season to remain consistent. Our multiple regression model is shown in the table below, along with the standard errors associated with each coefficient.

*Team linear regression model*

|  |  |  |
| --- | --- | --- |
|  | **Coefficient** | **Std. Error** |
| **Intercept** | -313.45 | 56.04 |
| **FG%** | 397.29 | 379.55 |
| **2P%** | 136.43 | 117.74 |
| **3P%** | 67.17 | 264.23 |
| **PTS** | 0.047 | 0.013 |
| **DRB** | 0.0038 | 0.0066 |

We see that the shooting percentage statistics have coefficients higher than 1, whereas the point and defensive rebound totals are lower than 1. This is because season totals for points and rebounds are both in the 1000’s, whereas percentages are obviously between 0 and 1, so these coefficients reflect this difference. As expected from the correlation, FG% has the highest coefficient, and the coefficient for defensive rebounds is actually negative. This seems off at first, but there are many factors to consider here as to why this is. First, the standard error for DRB is larger than the absolute value of the coefficient itself, meaning that there is enough variation that some teams likely have a positive association with DRB, and others a negative one. This makes sense when you think about what a defensive rebound constitutes. A team that has a bad overall defense is going to get shot on a lot, which means that regardless of how “good” their defensive rebounding is, they’ll still get a lot of DRB over a season solely because they are allowing a lot of shots to be taken. In this sense, a high DRB could lead to less wins, as it’s more a reflection of a poor defense than good rebounding. On the other hand, we’d expect a team that allows a low or normal amount of shots (decent to good overall defense) but that has a high number of DRB has a positively correlated effect for DRB. This is why our DRB coefficient and error look the way that they do. Overall, the effect of DRB on wins is somewhat low and the shooting percentages reflect very highly on wins.

**Step 2:**

In our second step, we simply predicted number of minutes played for the 2014/2015 season for each player. There are many ways to do this, each with their own respective chances of error, but we decided to use the current season minutes played (MP) statistics and extrapolate up to the full 82 game season to predict minutes played. To do this we took the current season MP for each player and divided that by the total team minutes played, then multiplied this fraction by the total average minutes played over a season to find the predicted minutes played for that player. This estimate may be a little off, as it doesn’t take into account cases such as players that are injured now but are normally starters, or rookies that haven’t played thus far 20 games into the season but will get more minutes as the season goes on. In this sense, there’s an inevitable amount of error we must accept when trying to “predict the future”, but this is the most reasonable way to extrapolate minutes played.

**Step 3:**

We next took all player statistics for the current season and up to the last 5 seasons and ran a decision tree regressor with inputs of MP and Age in order to predict the 5 significant statistics for the current season given current age and projected total minutes played. Age represented our time component here, as we wanted to capture how well each given player did as a function of their age as well as how much they were expected to play. We didn’t specify a max depth in order to allow the regressor to expand to an accuracy suitable for the specific player, as some players have played more seasons and have more data points than others. The limitation of this model is that since we’re only going 5 seasons back, no one person has a whole lot of data points to fit, and therefore we may overfit some of the data depending on the deviation in statistics for the given player.

The statistics we predicted were the raw number of shots and shot attempts (FG, FGA, etc.), points, and defensive rebounds, and the shot statistics were totaled at the end per team and then divided to get the predicted team FG%, 3P%, and 2P%.

**Step 4:**

Our final step was simply to plug in these aggregate statistics for each team back into our linear regression equation found in Step 1, and obtain the total predicted number of wins for each team in the 2014/2015 season. Doing so we predicted the overall NBA standings for 2014/2015 and the number of wins/losses expected for each team, which is shown in the standing table at the end of this document in the appendix.

***Conclusion/Analysis of Final Results:***

Overall, other than a few outliers, our rankings match up fairly well to what the current season and past couple of season’s rankings look like. Based on this current season’s standings top 15 teams (only about 1/4th of the way through the season so far), we see that our model has 11 of these top 15 teams in our top 15 in the table.

There are a few outliers however, which I think should be noted on a case to case basis in order to demonstrate the limitations of our model.

First, we see that the Cleveland Cavaliers are at the top of our modeled standings by a whopping 11 wins. Looking at the actual standings thus far, they are only in the 12th spot. This is due to the fact that arguably the league’s best player, Lebron James, moved back to the Cavs this season. His past statistical outputs, combined with newfound projected playing time at the Cavs is heavily skewing their win output to a (probably) unrealistic level. Kevin Love also moved to the Cavs from Minnesota last season, who is also an impact player and may have a skewed output given his past statistics with a very bad team that he probably had a disproportionate amount of positive statistics for compared to his new role on a better team. The fact of the matter is that basketball, unlike other sports, is influenced by a small number of players at a time therefore one player can have a large impact on performance, and our model captures this (probably over captures it though) in this sense.

Another ranking to note is the place of the Oklahoma City Thunder. In our model, they are projected to place 26th this season, which is very low considering they finished in the top 5 the last 3 seasons in a row. This is due to our projected minutes played algorithm. Two of their star players, Kevin Durant and Russell Westbrook, have been injured for most of the season so far and are just now becoming healthy to play again. Since our model factors MP as a function of the proportion of MP thus far in the season, these two players MP are being grossly underestimated and their high outputs are not coming through as much as they realistically should be, which is why OKC is ranked so low in our model.

A last observation to note is the high placement of the Detroit Pistons and low placement of the Miami Heat in our model, even though these two teams have recently been very bad and very good, respectively. The Detroit Pistons’ roster from last season to this season changed dramatically, almost half of the players on the team this year were not there last year. Same goes for Miami, who in addition to losing Lebron James made a lot of changes of their own. These relatively high number of roster changes are what account for the skew of these two teams’ rankings, since the players on the team now are bringing statistics with them from other teams, which may not accurately reflect how the team is utilizing those players in an accurate manner.

Overall though, we see that our model at least somewhat accurately predicts the standings for the 2014/2015 NBA season, and uses past statistics and regression models to do so.

**Appendix**

*Predicted 2014/2015 NBA Standings*

|  |  |  |
| --- | --- | --- |
| **Team** | **Wins** | **Losses** |
| Cleveland Cavaliers | 69 | 13 |
| San Antonio Spurs\* | 58 | 24 |
| Dallas Mavericks\* | 54 | 28 |
| Toronto Raptors\* | 53 | 29 |
| Golden State Warriors\* | 53 | 29 |
| Detroit Pistons | 51 | 31 |
| New Orleans Pelicans | 49 | 33 |
| New York Knicks | 48 | 34 |
| Phoenix Suns | 47 | 35 |
| Portland Trail Blazers\* | 46 | 36 |
| Atlanta Hawks\* | 45 | 37 |
| Denver Nuggets | 45 | 37 |
| Washington Wizards\* | 43 | 39 |
| Sacramento Kings | 40 | 42 |
| Memphis Grizzlies\* | 40 | 42 |
| Los Angeles Clippers\* | 39 | 43 |
| Houston Rockets\* | 39 | 43 |
| Brooklyn Nets\* | 38 | 44 |
| Utah Jazz | 37 | 45 |
| Chicago Bulls\* | 36 | 46 |
| Charlotte Bobcats\* | 36 | 46 |
| Milwaukee Bucks | 35 | 47 |
| Los Angeles Lakers | 34 | 48 |
| Indiana Pacers\* | 34 | 48 |
| Boston Celtics | 34 | 48 |
| Oklahoma City Thunder\* | 33 | 49 |
| Orlando Magic | 32 | 50 |
| Miami Heat\* | 29 | 53 |
| Minnesota Timberwolves | 19 | 63 |
| Philadelphia 76ers | 15 | 67 |