EE-4563 Final Project: Predicting NBA Playoff Average Scores

Group 20:
Jian Nan (Andy) Huang,
Mohammed Hasan

Introduction

The NBA playoffs are incredibly difficult to predict as upsets can happen at any time. Just because a team is good in the regular season does not mean that they will win NBA finals. Our objective of this project was to predict the team that would win the NBA finals. To do this we chose to predict the average points scored by the teams because usually the team that scored the most points wins the games. We attempted to approach this problem using multiple linear regression two times. The first time we tried to predict the players stats. And then the second time we used the players stats to predict the outcome of their teams points. The NBA data was taken from https://www.basketball-reference.com/ and was written in csv format. All of the code and data used in this project is available at https://github.com/AndySoftware653/Machine-Learning-NBA-Project.

Machine Learning Approach

The model chosen to predict the NBA average playoff scores was multiple linear regression. Multiple linear regression is an exhausted version of simple linear regression. It has multiple coefficients and multiple independent variables, in this project the independent variables are referred to as predictors and the dependent variables are target variables. Figure 1 shows the difference between a simple linear regression and multiple linear regression.

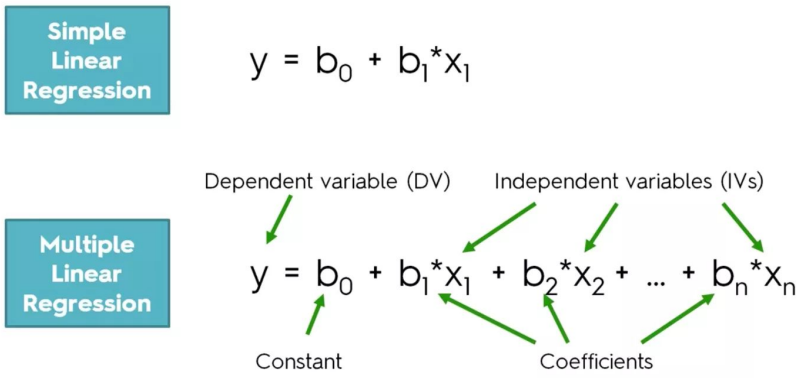

Figure1: Simple vs. Multiple Linear Regression

This technique was applied twice: once for predicting NBA players' seasonal stats, and once more for predicting average playoff scores for each playoff team respectively. The chosen predictors for predicting players' seasonal stats are: field goals, field goal attempts, field goal percentage, three pointers, three point attempts, three point percentage, two pointers, two point attempts, two point percentage, effective field goal percentage, free throws, free throw attempts, free throw percentage, steals, minutes played, turn overs, offensive rebounds, defensive rebounds, blocks, and personal fouls. Our target variables are points per game, steals, and total rebounds. After performing the first regression, every class in our target variables were summed

with other players' target variable. Finally, the summed target variables and finals average score per team were trained in another multiple linear regression.

Datasets

We got our data from https://www.basketball-reference.com/ and we used the 2014-2017 data for all the players that played in the NBA. We created this csv file by manually copying and pasting the data from 2014-2017. We then used the players data for the 2017-2018 season as our test data. One of the problems we found was that some players did not actually play much due to injuries or not getting enough playing time. The next dataset that was used was to find the teams that played in the 2018 playoffs and also which players played in the playoffs. Another problem we found when doing this was that the datasets contained a new team labeled TOT which was for traded players. Since TOT was not an actual team we had to go through the file and manually change anybody that made the playoffs and assign them to the correct team. Also some player categories were left blank in the csv files because they did not perform that action so we just filled those spaces with 0s.

Shortcomings of our Approach

Some of the shortcomings of our approach is that we did not factor in defense at all. We just assumed that if a team scored a lot of points that would mean they won the NBA finals. Since our code uses player stats to predict the outcome of the NBA finals, this means that teams that did good in the regular season would have a higher chance of winning with our code. This is not a necessarily accurate way of predicting because in any sport there are can be upsets where lower ranked teams defeat higher ranked teams.

Results

```
Residual sum of squares for training data: 0.0096
Residual sum of squares for test data: 0.0097
```

Figure 2:RSS values for training and test data

Our RSS values for the training and test data was really low and that means that the predictors we used had a really high correlation here. This could be due to the fact that some of our factors are directly correlated to one another, such as FG% and FGA correlating with points, since the more attempts you make the more likely you would be to have more points.

```
Residual sum of squares for test data: 0.8886
```

Figure 3: RSS values for NBA finals scores

Since our residual sum of squares was really high here it means the training predictors had little relation with the score outcome

Team	Actual Average Score	Predicted Average Score
Boston Celtics	101.4	101.758620
Cleveland Cavaliers	101.1	104.636596
Golden State Warriors	110.4	106.273070
Houston Rockets	104.9	106.104655
Indiana Pacers	100.6	103.844485
Miami Heat	103.4	103.411679
Milwaukee Bucks	101.9	105.655731
Minnesota Timberwolves	101.6	104.845624
New Orleans Pelicans	110.0	103.349910
Oklahoma City Thunder	101.2	102.685347
Philadelphia 76ers	108.8	105.138399
Portland Trail Blazers	105.5	103.200641
San Antonio Spurs	96.8	103.016891
Toronto Raptors	107.3	103.150457
Utah Jazz	102.4	103.449751
Washington Wizards	107.2	103.978144

Figure 4: Actual team average in the playoffs vs. our predicted score

In figure 4 above we can see that the scores for most of the teams were close to the predicted results.

```
Actual Team that won the Playoff:  Golden State Warriors
Predicted Team that won the Playoff:  Golden State Warriors
```

Figure 5: Actual and predicted winners

Since our code choose the highest predicted score as the winner we determined that the Golden State Warriors would win the 2018 playoffs, which they did. However, our code almost picked the Houston Rockets to win the 2018 playoffs as they happened to do really well in the regular season which made their predicted score higher than the actual.

Team	Error Percentage
Boston Celtics	0.353669
Cleveland Cavaliers	3.498117
Golden State Warriors	3.738161
Houston Rockets	1.148384
Indiana Pacers	3.225134
Miami Heat	0.011295
Milwaukee Bucks	3.685703
Minnesota Timberwolves	3.194512
New Orleans Pelicans	6.045536
Oklahoma City Thunder	1.467734
Philadelphia 76ers	3.365442
Portland Trail Blazers	2.179487
San Antonio Spurs	6.422408
Toronto Raptors	3.867235
Utah Jazz	1.025147
Washington Wizards	3.005463

Figure 6: Error Percentage per team

In figure 6 we can see what was our error percentage for each team. The highest error was for the San Antonio Spurs who did rather poorly in the playoffs and only managed to average 96.8 points each game. This is most likely because they got eliminated early on, since they lost in the first round. Most of our errors were around 3% so the overall accuracy was pretty good.

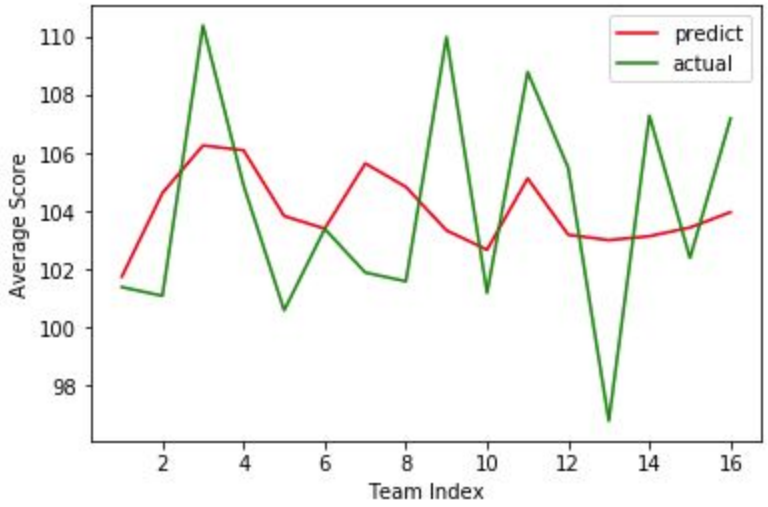

Figure 7: Actual vs predicted average playoff scores

In figure 7 the our predicted and actual results were graphed. The graph shows the winning team as Golden State Warriors which matches with the actual result. The shape of the graph doesn't match with the actual one. This could mean that not enough training data was given, or the training predictors had little relation with the score outcome.

Conclusion

Since the goal was to predict the winner of the NBA finals our code worked relatively well because it picked the right team. However according to our RSS value our predictors were not a good fit to predict the target variable. A more interesting approach we could have taken was to predict the outcome of each separate game and see who would be the winner and see if our code could accurately predict each game winner but unfortunately we did not have the time to do that. We could have also used different methods to predict our results such as SVM classifiers or using neural networks to see which approach would result in better accuracy. But we decided that the linear regression approach was able to predict pretty well since it only had a 3% error.