Nicholas Komarovsky

Professor Carter

MA346-SN1

12/08/2020

NBA Data Final Project

From as early as I can remember, I have been in love with the game of basketball. Through my early childhood, I remember watching Paul Pierce carry an otherwise uninspiring Boston Celtics roster for most of the 2000s until the current day with the young stars like Jaylen Brown and Jayson Tatum. As I grew through the years with the sport, I developed a thirst for finding and interpreting some of the data. In this project, I am looking to figure out some of the most important variables in predicting good basketball teams. Upon some brief research, I found some data scientists saying that effective field goal percentage, turnover percentage, free throw factor, offensive rebounding percentage, and defensive rebounding percentage are the factors that determine good teams. I am putting that to the test and seeing if perhaps other factors could be more significant.

I start off by creating two CSV files from the NBA Reference website (<https://www.basketball-reference.com/leagues/NBA_2020.html>). The data comes with having asterisks next to the playoff teams, so I must remove that. Additionally, the last row needs to be removed because it shows league averages of the statistics. I must also create the effective field goal percentage (the sum field goals made and 0.5 \* 3 pointers made all divided by field goal attempts), turnover percentage (turnovers divided by the sum of turnovers, field goal attempts, 0.44 \* free throw attempts, and assists), and free throw factor (free throws made divided by field goal attempts) columns. I then use this data to build a logistic regression model. I select 7 predictors from the CSV files that are significant, including the 5 mentioned above. The additional 2 are points per game and net rating, two variables that I would intuitively find to be important in predicting good teams. In order to make sure that these variables show a trend, I create a dashboard with 7 graphs of the teams on the x axis and the predictor variable in question on the y axis, sorted in ascending order of win percentage. I also keep the asterisks on the playoff teams in order to highlight those. All the predictors seem to have a trend as the win percentage of the teams increases, so a model on these 7 predictors looks valid. In order to make the logistic model, another column is created to make the response variable. This column will have a value of 1 if the team made the playoffs in 2020, and 0 if the team did not.

Now, a model can be fitted to the predictor variables. The output of the models will be the F1 score, which uses the precision and recall. Precision takes the number of true positive divided by the sum of true positives and false positives. Recall is the true positive divided by the sum of true positives and false negatives. True positive relates to situations where the predicted and actual value are both 1 (playoff team). False positive is when the predicted value is 1 (playoff team), but the actual value is 0 (not a playoff team. False negative is when the predicted value is 0, but the actual value is 1. The first model, using all 7 of the predictor variables, has an F1 score of around 0.9. This signifies a very effective model, as results closer to 1 signify more accurate models.

Despite such a high F1 score, there could be some overfitting going on. In order to make sure, the dataset is split up 70% into training and 30% into validation. The validation data in this situation will show how accurately the model can be applied to a larger dataset. This split leads to an F1 score of around 0.92 for the training data and 0.89 for the validation data. This suggests that the data may be overfitting, so I will go back to just the 5 variables that data scientists find to be the best predictors of good teams. I find that the F1 score for the testing data is similar when running the new model, but the validation data decreases to around 0.75. Since the validation data decreases even more in this situation, there is more evidence for overfitting with these 5 predictors in the model. I will also check to see how much each variable affects the model by printing out the standardized coefficient values. It is important to standardize because otherwise the values all have different scales and it would be difficult to make interpretations. Based on this, the offensive rebounding percentage looks to be the least important variable in the model, but the new model created with removing that variable leads to similar F1 scores.

Overall, the model with the 7 predictor variables is highly accurate in predicting playoff teams. However, there is reason to believe that this model is overfitting the data. This makes sense since the size of the dataset is small and most of the predictor variables have a strong correlation with the response. In order to make the sample size bigger, I could add data from other NBA seasons in order to get a more accurate representation.