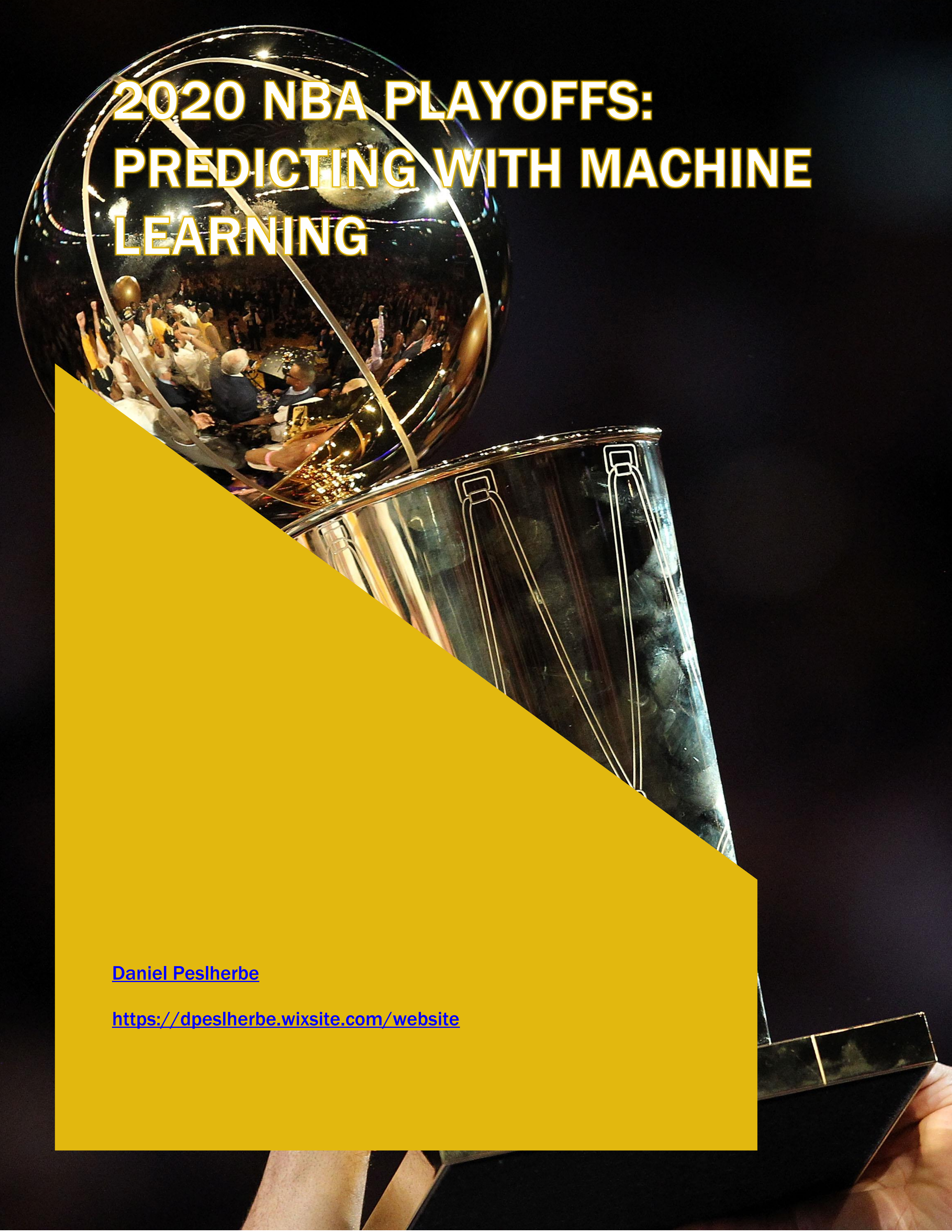
2020 NBA PLAYOFFS: PREDICTING WITH MACHINE LEARNING



Daniel Peslherbe

https://dpeslherbe.wixsite.com/website

Introduction

In 2020, due to the COVID-19 pandemic, the NBA Season was indefinitely suspended before reaching the playoffs. While discussions are underway about eventually and/or possibly resuming the season, many have wondered about the effects this suspension would have on the rest of the season and its playoff implications. Some have argued that teams with a weaker strength of schedule for the rest of the season are at a disadvantage, given they have played most of their hardest games already, while others are in support of a play-in tournament for the lower seeds of the playoffs. Thus, we shall attempt to classify our 2019-2020 NBA season teams into specific categories, as follows:

- Playoff Teams
- Second Round Teams
- Conference Finals Teams
- Finals Teams
- NBA Champions

Models & Data

The models we shall use here are Logistic Regression, Linear Discriminant Analysis and K-Nearest Neighbors.

The Data used are a combination of a modified version of an NBA Financial Dataset (found here: data.world/makeovermonday/2018w29-historical-nba-team-spending-against-the-cap) and data scraped from basketball-reference.com through the nbastatR package from Alex Bresler.

Methodology

We start by taking our overall training data (from the 1990-1991 NBA season to the 2018-2019 NBA season) and separate it into 4 folds. These 4 folds will then be used to test our models before using them to predict on our actual 2019-2020 data.

Note however, the high number of predictor variables in our data is problematic (See the curse of dimensionality). To alleviate this issue, we use Recursive Feature Elimination to find our best set of predictors for each category (this is computed through the rfe() and rfeControl() functions of the caret package). Note that since K-Nearest Neighbors is not a true linear model, instead, the best set of predictors used for each category will be the set of predictors with variable importance > 20% (this is computed through train() and varImp() functions of the caret package).

We use these predictors to train our models on 3 of our 4 folds, then record multiple model metrics against the fold that was left out and apply this again for each fold. These metrics will give us an idea of our model precisions. The metrics we use are the Mean Squared Error (MSE, which is the average error of the model's predictions compared to actual data), the Accuracy (which is the proportion of the model's predictions which are correct), the Precision (which is the proportion of the model's true correct observations over its true observations), the Recall (which is the proportion of the model's true correct observations on the total true observations), the F1 Score (which is an harmonic mean of our Precision and Recall), and the Area Under the Receiver Operating Characteristic curve (AUROC, which is the area under a curve that shows the trade-off between true positive rate and false positive rate).

Note that Accuracy, Precision, Recall and F1 Score are values between 0 and 1.The AUROC is a value between 0.5 and 1. The higher the result of these metrics, the better the model's fit to the data. Note that we also want a small MSE for a good model.

We also make sure to check the Confusion Matrix for each fold (see link for explanation), to make sure that the models are not systematically making the same mistakes (like a strong proportion of false positives or false negatives compared to true positives or true negatives).

Note that Nearest Neighbor Model is an 11-to-13-Nearest Neighbor Model (we use a loop for every K-Nearest Neighbor Model for k ranging from 1 to 100; in this case, the model with the smallest training and out-of-sample error varies between 11 and 13 neighbors). We must acknowledge that this means the model is prone to overfitting and can have a high variance. Luckily, we also have our two other models to help compare.

Playoff Team Classification

We start by creating a finish variable based on where a team finished in the postseason (missed it, lost in the first round, lost in the second round, lost in the conference finals, lost in the finals, or won the championship), and then convert it into a numeric value (5, 4, 3, 2, 1, and 0 respectively). Since

we want to start by classifying only playoff level teams, we convert all the playoff teams into the numeric 0 (success in reaching the playoffs), and the teams who missed out on the playoffs at the numeric 1.
We then train our models through 4-fold cross-validation, and record our aggregated metrics, which we then average over the 4 folds.

Note for our Logistic Regression Model, we get the following metrics:

MSE	0.08155236
Accuracy	0.9184476
Precision	0.9081982
Recall	0.9316275
F1 Score	0.9194962
AUROC	0.9186381

We note the Confusion Matrix seem well balanced.
The predictors for our Logistic Regression Model were :
Conference Standing, BLKPG, STLPG, Wins, EFG%, FG%, TS%, NRTG, ORTG, DRTG, Margin of Victory, Opponent's EFG%, Opponent's 2FGAPG, & Opponent's 2FGMPG.
Then, when applied to 2019-2020 NBA season metrics, our Logistic Regression Model classifies the Boston Celtics, Brooklyn Nets, Denver Nuggets, Houston Rockets, Indiana Pacers, LA Clippers, Los Angeles Lakers, Miami Heat, Milwaukee Bucks, Oklahoma City Thunder, Philadelphia 76ers, Toronto Raptors and Utah Jazz as Playoff level teams.

Note for our Linear Discriminant Model, we get the following metrics:

MSE	0.0779755
Accuracy	0.9220245
Precision	0.9121764
Recall	0.9342391
F1 Score	0.9228092
AUROC	0.9224942

We note the Confusion Matrix seem well balanced.

The predictors for our Linear Discriminant Model were :
Conference Standing, Wins, Margin of Victory, NRTG, ORTG, DRTG, EFG%, TS%, FG%, FG2%, Opponent's EFG%, Opponent's FG%, Opponent's FG2%, & Opponent's PTSPG.
Then, when applied to 2019-2020 NBA season metrics, our Linear Discriminant Model classifies the Boston Celtics, Brooklyn Nets, Denver Nuggets, Houston Rockets, Indiana Pacers, LA Clippers, Los Angeles Lakers, Miami Heat, Milwaukee Bucks, Oklahoma City Thunder, Orlando Magic, Philadelphia 76ers, Toronto Raptors and Utah Jazz as Playoff level teams. Note most teams were already classified as such by our Logistic Regression Model, with the exception of the Orlando Magic.

Note for our Nearest Neighbors Model, we get the following metrics:

MSE	**0.1111173**
Accuracy	**0.8888827**
Precision	**0.868416**
Recall	**0.9143444**
F1 Score	**0.8908118**
AUROC	**0.889209**

We note the Confusion Matrix seem well balanced.
The predictors for our Nearest Neighbors Model were :
Conference Standing, Wins, Margin of Victory, NRTG, ORTG, DRTG, EFG%, TS%, FG2%, Average Team Age, DRBPG, Opponent's FG%, Opponent's EFG%, Opponent's FG2%, Opponent's FG3%, Opponent's ASTPG, & Opponent's FGMPG.
Then, when applied to 2019-2020 NBA season metrics, our Nearest Neighbors Model classifies the Boston Celtics, Brooklyn Nets, Dallas Mavericks, Denver Nuggets, Houston Rockets, Indiana Pacers, LA Clippers, Los Angeles Lakers, Miami Heat, Milwaukee Bucks, Oklahoma City Thunder, Philadelphia 76ers, Toronto Raptors and Utah Jazz as Playoff Level Teams.

Second Round Team Classification

Now, we define success as being able to reach the second round of the playoffs. Using the same methods as for our Playoffs Classification, we observe the following.

Note for our Logistic Regression Model, we get the following metrics:

MSE	0.1146383
Accuracy	0.8853617
Precision	0.807378
Recall	0.8105199
F1 Score	0.8084898
AUROC	0.8639303

We note the Confusion Matrix seem well balanced, even though the model tends to output false positives more than false negatives.
The predictors for our Logistic Regression Model were :
Conference Standing, Wins, STLPG, ORTG, FG3MPG, FG2MPG, FG2APG, FT%, FG%, Opponent's FG3MPG, Opponent's FG3APG, Opponent's STLPG, & Opponent's TOV%.
Then, when applied to 2019-2020 NBA season metrics, our Logistic Regression Model classifies the Boston Celtics, Denver Nuggets, Indiana Pacers, LA Clippers, Los Angeles Lakers, Miami Heat, Milwaukee Bucks, Oklahoma City Thunder, Toronto Raptors and Utah Jazz as Second Round level teams.

Note for our Linear Discriminant Model, we get the following metrics:

MSE	0.124089
Accuracy	0.875911
Precision	0.7624871
Recall	0.849895
F1 Score	0.8036193
AUROC	0.8684665

We note the Confusion Matrix seem well balanced, even though the model also tends to output false positives more than false negatives.

The predictors for our Linear Discriminant Model were :
Conference Standing, Wins, Margin of Victory, NRTG, ORTG, DRTG, EFG%, TS%, FG%, FG2%, Opponent's EFG%, Opponent's FG2%, & Opponent's PTSPG.
Then, when applied to 2019-2020 NBA season metrics, our Linear Discriminant Model classifies the Boston Celtics, Denver Nuggets, Indiana Pacers, LA Clippers, Los Angeles Lakers, Miami Heat, Milwaukee Bucks, Oklahoma City Thunder, Toronto Raptors and Utah Jazz as Second Round level teams.

Note for our Nearest Neighbor Model, we get the following metrics:

MSE	**0.1335397**
Accuracy	**0.8664603**
Precision	**0.7864766**
Recall	**0.7590596**
F1 Score	**0.7721646**
AUROC	**0.8357098**

We note the Confusion Matrix seem well balanced.
The predictors for our Nearest Neighbors Model were :
Conference Standing, Wins, Margin of Victory, NRTG, ORTG, DRTG, EFG%, TS%, FG%, FG2%, Average Team Age, DRBPG, BLKPG, Opponent's FG%, Opponent's EFG%, Opponent's FG2%, Opponent's FG3%, Opponent's PTSPG, Opponent's ASTPG, & Opponent's BLKPG.
Then, when applied to 2019-2020 NBA season metrics, our Nearest Neighbors Model classifies the Boston Celtics, Dallas Mavericks, Indiana Pacers, LA Clippers, Los Angeles Lakers, Miami Heat, Milwaukee Bucks, Oklahoma City Thunder, Toronto Raptors, and Utah Jazz as Second Round Level Teams.

Conference Finals Team Classification

Now, we define success as being able to reach the Conference Finals of the playoffs. Using the same methods as for our Playoffs Classification, we observe the following.

Note for our Logistic Regression Model, we get the following metrics:

MSE	0.09102544
Accuracy	0.9089746
Precision	0.7238839
Recall	0.6893939
F1 Score	0.7029568
AUROC	0.8194906

We note the Confusion Matrix seem well balanced, even though the model tends to output false positives more than false negatives.
The predictors for our Logistic Regression Model were :
Conference Standing, FG3MPG, FGAPG, FG3APG, FG3%, FG2%, Opponent's FG2%, Opponent's FG3%, & Opponent's FT-to-FGA Ratio.
Then, when applied to 2019-2020 NBA season metrics, our Logistic Regression Model classifies the LA Clippers, Los Angeles Lakers, Milwaukee Bucks and Toronto Raptors as Conference Finals level teams.

Note for our Linear Discriminant Model, we get the following metrics:

MSE	0.1028514
Accuracy	0.8971486
Precision	0.7185066
Recall	0.5606061
F1 Score	0.6277216
AUROC	0.7599928

We note the Confusion Matrix seem well balanced, even though the model also tends to output false negatives more than false positives.
The predictors for our Linear Discriminant Model were :
Conference Standing, Wins, Margin of Victory, NRTG, & Opponent's EFG%.

Then, when applied to 2019-2020 NBA season metrics, our Linear Discriminant Model classifies the LA Clippers, Los Angeles Lakers, Milwaukee Bucks, and Toronto Raptors as Conference Finals level teams.

Note for our Nearest Neighbor Model, we get the following metrics:

MSE	**0.1075572**
Accuracy	**0.8924428**
Precision	**0.7058937**
Recall	**0.530303**
F1 Score	**0.6050676**
AUROC	**0.7448413**

We note the Confusion Matrix seem well balanced.
The predictors for our Nearest Neighbors Model were :
Conference Standing, Wins, Margin of Victory, NRTG, ORTG, DRTG, EFG%, TS%, FG%, FG2%, Opponent's FG%, Opponent's EFG%, Opponent's FG2%, & Opponent's FG3%.
Then, when applied to 2019-2020 NBA season metrics, our Nearest Neighbors Model classifies the LA Clippers and Los Angeles Lakers as Conference Finals Level Teams.

Finals Team Classification

Now, we define success as being able to reach the Finals of the playoffs. Using the same methods as for our Playoffs Classification, we observe the following.

Note for our Logistic Regression Model, we get the following metrics:

MSE	0.08864459
Accuracy	0.9113554
Precision	0.6197917
Recall	0.3702899
F1 Score	0.4597718
AUROC	0.6719864

We note the Confusion Matrix seem well balanced.
The predictors for our Logistic Regression Model were :
Conference Standing, FG3MPG, Opponent's BLKPG, Opponent's PTSPG, Opponent's FTMPG, Opponent's TOVPG, & Opponent's TOV%.
Then, when applied to 2019-2020 NBA season metrics, our Logistic Regression Model classifies the Los Angeles Lakers and Milwaukee Bucks and Finals level teams.

Note for our Linear Discriminant Model, we get the following metrics:

MSE	0.08154677
Accuracy	0.9184532
Precision	0.7944444
Recall	0.3031703
F1 Score	0.4289152
AUROC	0.6463357

We note the Confusion Matrix seem well balanced, even though the model also tends to output false negatives more than false positives.
The predictors for our Linear Discriminant Model were :
Conference Standing, Wins, Margin of Victory, NRTG, ORTG, DRTG, FG%, FG2%, EFG%, Opponent's FG%, Opponent's FG2%, Opponent's EFG%, & Opponent's BLKPG.

Then, when applied to 2019-2020 NBA season metrics, our Linear Discriminant Model classifies the Los Angeles Lakers and Milwaukee Bucks as Finals level teams.

Note for our Nearest Neighbor Model, we get the following metrics:

MSE	**0.07800344**
Accuracy	**0.9219966**
Precision	**0.7528409**
Recall	**0.3707428**
F1 Score	**0.4914747**
AUROC	**0.6781412**

We note the Confusion Matrix seem well balanced.
The predictors for our Nearest Neighbors Model were :
Conference Standing, Wins, Margin of Victory, NRTG, ORTG, DRTG, FG%, FG2%, EFG%, TS%, ASTPG, Opponent's BLKPG, Opponent's FG%, Opponent's FG3%, & Opponent's EFG%.
Then, when applied to 2019-2020 NBA season metrics, our Nearest Neighbors Model classifies the Dallas Mavericks and Los Angeles Lakers as Finals Level Teams.

Championship Team Classification

Now, we define success as being able to reach the Finals of the playoffs. Using the same methods as for our Playoffs Classification, we observe the following.

Note for our Logistic Regression Model, we get the following metrics:

MSE	**0.04254225**
Accuracy	**0.9574577**
Precision	**0.0**
Recall	**0.0**
F1 Score	**0.0**
AUROC	**0.5**

We note the Confusion Matrix are unbalanced, as the model is not capable of predicting any champions in the training data.
The predictors for our Logistic Regression Model were :
Conference Standing, FG3MPG, Opponent's PTSPG, Opponent's FTMPG, Opponent's BLKPG, Opponent's TOVPG, & Opponent's TOV%.
Then, when applied to 2019-2020 NBA season metrics, our Logistic Regression Model classifies the Los Angeles Lakers as a Championship Level Team.

Note for our Linear Discriminant Model, we get the following metrics:

MSE	**0.03664603**
Accuracy	**0.963354**
Precision	**0.75**
Recall	**0.1666667**
F1 Score	**0.2714286**
AUROC	**0.5821256**

We note the Confusion Matrix seem well balanced, even though the model also tends to output false negatives more than false positives.
The predictors for our Linear Discriminant Model were :
Conference Standing, Wins, Margin of Victory, NRTG, ORTG, DRTG, FG%, FG2%, EFG%, Opponent's FG%, Opponent's FG2%, Opponent's EFG%, & Opponent's BLKPG.

Then, when applied to 2019-2020 NBA season metrics, our Linear Discriminant Model classifies no team as Championship Level Team.

Note for our Nearest Neighbor Model, we get the following metrics:

MSE	**0.04018935**
Accuracy	**0.9598107**
Precision	**0.4375**
Recall	**0.2083333**
F1 Score	**0.2669526**
AUROC	**0.6004981**

We note the Confusion Matrix seem well balanced.
The predictors for our Nearest Neighbors Model were :
Conference Standing, Wins, Margin of Victory, NRTG, ORTG, DRTG, FG%, FG2%, EFG%, TS%, ASTPG, Opponent's FG%, Opponent's FG3%, Opponent's EFG%, & Opponent's BLKPG.
Then, when applied to 2019-2020 NBA season metrics, our Nearest Neighbors Model classifies no team as Championship Level Teams.

Table of Team Playoff Levels (from Machine Learning Analysis)

Playoff Team	2nd Round Teams	Conference Finals Teams	Finals Teams	Championship Teams
Boston Celtics[1,2,3]	Boston Celtics[1,2,3]			
Brooklyn Nets[1,2,3]				
Dallas Mavericks[3]	Dallas Mavericks[3]		Dallas Mavericks[3]	
Denver Nuggets[1,2,3]	Denver Nuggets[1,2]			
Houston Rockets[1,2,3]				
Indiana Pacers[1,2,3]	Indiana Pacers[1,2,3]			
LA Clippers[1,2,3]	LA Clippers[1,2,3]	LA Clippers[1,2,3]		
Los Angeles Lakers[1,2,3]	Los Angeles Lakers[1,2,3]	Los Angeles Lakers[1,2,3]	Los Angeles Lakers[1,2,3]	Los Angeles Lakers[1]
Miami Heat[1,2,3]	Miami Heat[1,2,3]			
Milwaukee Bucks[1,2,3]	Milwaukee Bucks[1,2,3]	Milwaukee Bucks[1,2]	Milwaukee Bucks[1,2]	
Oklahoma City Thunder[1,2,3]	Oklahoma City Thunder[1,2,3]			
Orlando Magic[2]				
Philadelphia 76ers[1,2,3]				
Toronto Raptors[1,2,3]	Toronto Raptors[1,2,3]	Toronto Raptors[1,2]		
Utah Jazz[1,2,3]	Utah Jazz[1,2,3]			

[1] Logistic Regression, [2] Linear Discriminant, [3] Nearest Neighbors

Final Considerations

There are some notes to take into consideration with this analysis.

- We cannot use the predictors of the models to judge true significance of a team's performance, since the predictors used are tailored to each model. This means that, over the next couple seasons, there significance for the model could diminish depending on the state and advancement of League-wide Offense and/ or Defense. This is also observable for the Dallas Mavericks; our Finals level Nearest Neighbor Model classifies them as Finalists, but our Conference Finals Nearest Neighbors Model does not classify them as Conference Finalist. This is because, the models are joint, they classify separately based on variables of interest specific to the Playoff Level desired to reach.
- To make sure that the effect and dispersion of MoreyBall principles (increased 3-point shooting and decreased mid-range shooting) are not overly strong, every season's data is scaled separately, then scaled again all together. This means that our predictors rely on performance based against the rest of the league.
- The level reached by teams is not meant to be a true reflection of team strength based against the rest of the league; what we have done is look at statistics from other teams who have reached that level based against the rest of the league, and identified where those teams are similar. This doesn't mean the Clippers cannot reach the Finals, but that based on historical data, the Clippers are not likely to reach the Finals. It is also why the Thunder is classified as a Playoff level team in 2 of the three models, but as a Second-Round level team in all three models. This means that the Thunder's model predictors share more similarity with Second Round achieving teams predictors than Playoff achieving team predictors.
- The models take into account the statistics for the entire season; this can lead to underestimation or overestimation for teams who made significant changes to their roster mid-season (acquisitions of Marcus Morris Sr., Jae Crowder) or who have had significant player injuries for a part of the season (Zion Williamson, Raptors having multiple injured players (albeit never all at the same time)). In a world where the season continues, it can be possible the Pelicans come back to make the Playoffs. This model does not consider these situations, or game winning momentum and/or streaks.

- Penultimately, we must finally acknowledge that the models are much more reliable for lower levels of expectation. This is due to the fact that there is more homogeneous training data for those lower expectation levels (For example, there is around half the league in the Playoffs every year (16 teams on 27~30 teams) compared to only 4 teams in the Conference Finals, and 1 Champion per season). Most notably, while the Los Angeles Lakers are crowned as a Championship Level team by the Logistic Regression model, the model is quite unreliable. This is normal, given the sparsity of championship team data, but also because nothing is set in stone in the playoffs (upsets can happen! Injuries can derail championship dreams!), and thus, the variance of the outcomes is increasing.
- On a final note, what can be observed is that team strength is a combination of Conference Standing, but also of different team metrics; while teams can have pretty different playing styles (for example, the Indiana Pacers compared to the Milwaukee Bucks for shot distribution) and, thus, different team metrics for Offense and/or Defense, multiple teams can be classified into the same category independently of said playing styles. Another example that can be seen is the Dallas Mavericks (once again) who stand out in the Nearest Neighbors Model, even while being the West's 7th seed, but are still classified as a Finalist Level team by the model (which is likely due to the very efficient Offense they play this season).

What Next? Prediction improvements and suggestions

- To get a true sense of predictors for playoff success, it would be better to build a classification model that tries to predict the best possible playoff level reached by a team based on its statistics (as opposed to the Playoffs – yes or no, Second Round – yes or no, Conference Finals – yes or no, etc… models that we have explored here). It might also be interesting to attempt Regression rather Classification (using our yes or no models would give us probabilities of reaching or failing to reach a certain playoff level rather than the playoff level class). Finally, there is also the fact that the predictors vary with every model built and playoff level to be attained; it would be informative to know if there are variations between models and/or results if we use the same set of predictors for all models (Dean Oliver's Four Factors of Basketball for example).