Determining the Greatest NBA Teams of All Time and Current Postseason Predictions Using Machine Learning

Fred Jones

December 14, 2023

1 Introduction

Arguments will always be made about which NBA team deserves the crown of the greatest team of all time. Is it the '96 Chicago Bulls? The '86 Boston Celtics? The '16 Golden State Warriors? At the end of the day, there will never be one objective answer to this debate.

Every fan of the NBA holds their own opinions and biases. My goal is to narrow this argument down to several teams that are undeniably worthy of being in this conversation using seasonal data starting from the 1984-1985 season. (Note: this is not me delegitimizing teams before this season. The game has just changed so much since then.) I will also predict who has the best chance of becoming this season's NBA champion, and what teams have the best chances of making the conference finals. NBA analytics can reveal a lot about a team's offensive and defensive prowess, and they allow us to compare current NBA teams with those from the past like never before. It's crucial to take NBA data into account when comparing any team against another; Quite frankly, it's the only objective way to go about it.

The main algorithms used in this project are agglomerative clustering, linear regression, XGBoost, Lasso regression, and Ridge regression. Through agglomerative clustering, my goal was to find teams with similar season-long statistics that are representative of their defensive greatness, offensive prowess, and shot-making ability. But, regular season stats rarely tell the whole story of a team. Championships shift the balance between a good team and a great one, and championship-winning teams will stand out using this method and data. Through the predictive techniques listed, I will present results from various models using historical data to predict both confidence scores of teams making the conference finals and winning the NBA finals for the current NBA season.

2 Prior Work/References

Sports betting is growing at a rapid pace as sports books are opening across the country. We are in a golden age in terms of having access to similar data to the odds makers. Predicting the outcomes of certain sporting events has become a popular topic in the realm of machine learning as a result. Several notable research projects in this field have inspired me to complete one of my own.

- **"Predicting NBA Playoffs using Machine Learning"**
  Authors: Sean Liu
  URL: https://www.researchgate.net/publication/349646430_Predicting_NBA_Playoffs_Using_Machine_Learning

  This paper leverages the use of various machine learning techniques to predict the NBA playoff picture and round-by-round outcomes, all the way up to predicting the Finals Champion. Methods he used include Logistic Regression and K-nearest neighbors, among others. My project draws inspiration from this paper, as I will be using Agglomerative Clustering to group similar teams across different eras of the game, and predict NBA Champions using Linear Regression and other regression models. The topic of this paper motivated me to create my own NBA data project.

- **”Predictive Analysis and Modelling Football Results using Machine Learning Approach for English Premier League”**
  Authors: Rahul Baboota, Harleen Kaur
  URL:https://www.sciencedirect.com/science/article/pii/S0169207018300116

  I was very intrigued with the idea of modeling premier league soccer as well, although my interest in basketball slightly trumps it. This project includes lots of work in feature engineering and selection, which is not included to this extent in my project. Either way, I still drew a lot of inspiration from this paper. The algorithms that the authors chose to use included SVM, Gaussian naive Bayes, Random Forest, and Gradient Boosting; I chose to implement XGBoost regression due to its noted success in this project.

3 Model/Algorithm/Method

3.1 Data Pre-processing, Scraping, and Exploratory Data Analysis

As in every data science project, data pre-processing is the first task to conquer. In this project, I needed to remove the likes of empty columns, rows with NaN values, and duplicate columns created after merging season data with advanced season analytics. I added columns for the season year, the NBA champion, and the teams who made the conference finals in a given season. This was done for every dataset. The final result was a singular dataset containing basic and advanced seasonal data from every team in the association, starting from the 1984-1985 season and stretching to the 2022-2023 season. The test set in use is the current season data that is available as the season continues to progress. The same data pre-processing and cleaning techniques were used on this data as well to ensure compatibility.

Included in the *predicting.py* file is a function for scraping tables from https://www.basketball-reference.com/leagues/NBA_2024.html. With this implemented, the data frame used for the test set will be updated throughout the season which will make for better informed predictions.

Included in the *clustering.py* file is a visualization of different shooting and efficiency trends over each season. Both plots show interesting offensive trends that show the evolution of the league quite concisely.

3.2 Agglomerative Clustering with PCA

The first algorithm I implemented was Agglomerative Clustering, one of the most popular methods of hierarchical clustering. I chose stats that I believe are indicative of a particular team’s offensive and defensive performance across an entire season, as well as efficiency on both sides of the ball, and whether a team won the finals that year or not. The stats used were:

[FG%, 3P%, FT%, ORB, DRB, AST, STL, BLK, TOV, PTS, W, L, NRtg, Champion, TS%, eFG%, opponent eFG%]

Principal Component Analysis (PCA) was used to transform the features listed above into 2 uncorrelated principal components, while also making cluster visualizations far more effective. The results from this algorithm provide a clear basis for my objective of singling out the greatest NBA teams of all time, as you will see in the results section of this paper.

3.3 Linear Regression

Linear regression proved to be a great addition to this project. I ran two different predictions using a linear regression model - one using Champion as the target variable, and one using Conference Finals as the target variable. These columns are binary, which usually calls for some classification method. I chose to implement regression models instead for several reasons, one of which includes class imbalance in both target columns, which heavily restricts classification algorithms like Logistic Regression and Random Forests. The output from these linear regression models can be considered as confidence scores; a higher float value indicates more confidence in that team potentially becoming the current NBA champion or conference finalist, respectively.

3.4 Lasso Regression

Lasso regression was the second regression model I chose to train and use for predictions, although it proved not as effective as I initially thought. Lasso models tend to shrink some variable coefficients down to 0, which essentially acts as built-in feature selection. It works great with high-dimensional datasets containing far more features than the data used in this project, while this project requires hand-selected features.

3.5 Ridge Regression

I chose to implement Ridge regression as well after seeing the results from the Lasso models. Ridge regression adds an L2-regularization penalty term to the regression coefficients meaning the coefficient terms will shrink, but never reduce to 0. No selected features were left out while training the model. Results stemming from the ridge regression models were comparable to those of linear regression, and in most cases seem more realistic.

3.6 XGBoost Regressor

I decided to implement an XGBoost regression model after seeing promising results from another project. Over-fitting was the biggest issue I came across when training these models; It took lots of trial and error with hyper-parameter tuning. The initial XGB regression models I trained were far too complex for the prediction task at hand. Eventually, I settled for a model that only allows trees with a maximum depth of four in the ensemble, simplifying each tree enough to avoid over-fitting to the training data.

4 Results

4.1 Shooting Tendencies and Efficiency Trends

Below are two different graphs that show the league average in multiple shooting categories by year. Figure 1 includes field-goal percentage, three-point percentage, true shooting percentage, effective field goal percentage, and free throw percentage. Figure 2 includes the volume of such shots, average pace, and average points

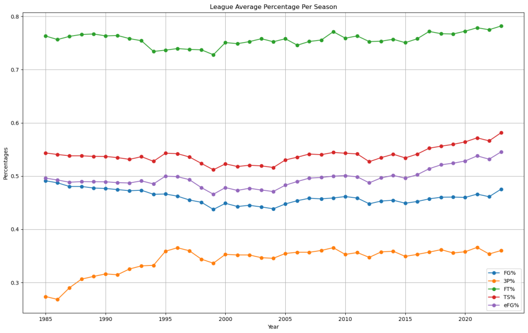


Figure 1: Shooting and Efficiency Percentages

One of the clearest trends in this data shows the rise of the three-point shot and the fall of the two-point shot. The league average of three-point attempts has risen by over ten attempts per game since 2015. Teams were averaging just 2.7 three-pointers per game in 1984-1985 - the league average is now 33.7 attempts. The volume of the two-point shot has reduced drastically since the mid-1980s

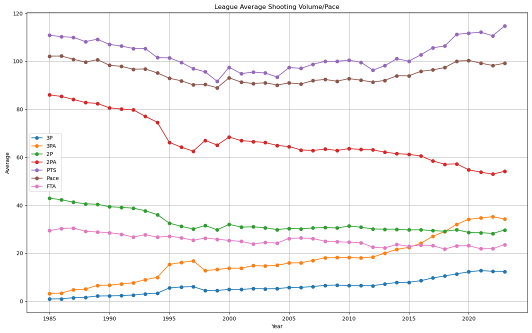


Figure 2: Shooting Volume, Points, Pace

as well, with teams attempting 30 fewer two-point field goals per game last season than they were in. The three-ball has completely changed the game.

4.2 Clustering

My goal with clustering was to single out some of the greatest teams to ever play, taking into account championships, offensive firepower, and defensive prowess. I chose to reduce the dimensionality of my clustering data down to two principal components for interpretation and visualization purposes. Each principal captures the varied influences of the original 17 attributes used in clustering. Figure 3 shows the mass clustering of every team, and Figure 4 shows a portion of the clustering with clear statistical advantages.

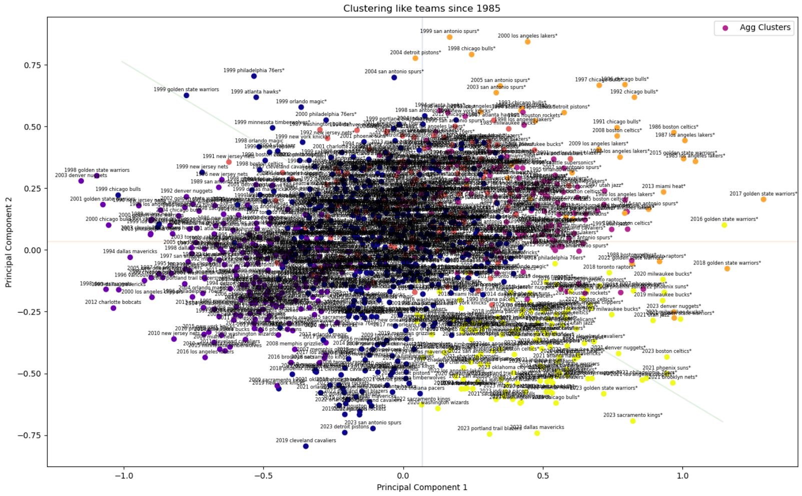


Figure 3: Every NBA team since 1984-1985 season in 4 clusters

The section of this clustering to focus on is in Figure 4. Every team in the orange cluster was the NBA champion in that given season while also dominating the league in the regular season. Now let's look at the teams above the line in Figure 4 - the teams that stand out more than any in the cluster.

- 1985, 1987, and 2000 Los Angeles Lakers
- 1996, 1997, and 1992 Chicago Bulls
- 1986 Boston Celtics

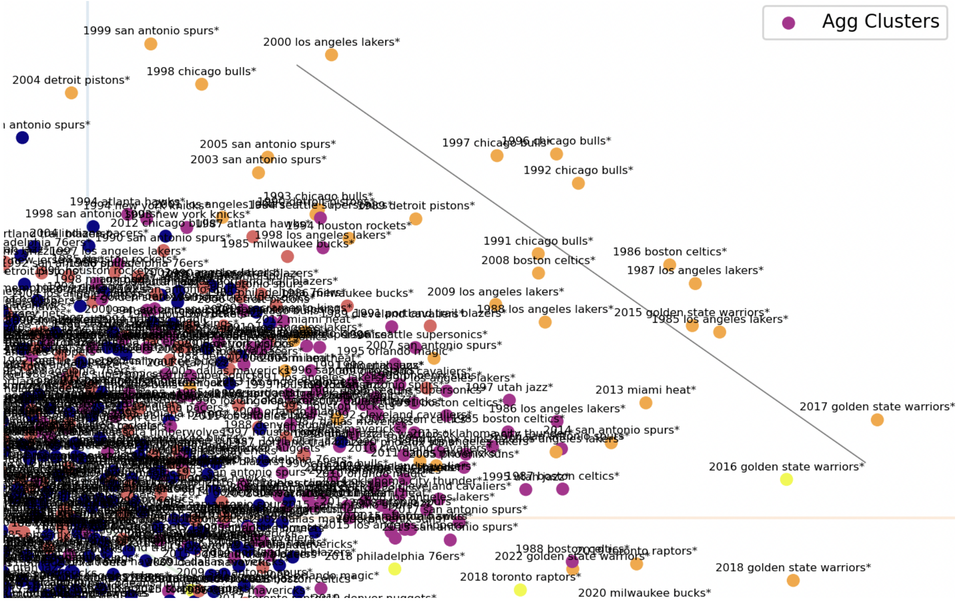


Figure 4: Top right quadrant of Figure 3

- 2015 and 2017 Golden State Warriors

Statistically speaking, these teams are in an echelon of their own. The discussion of the greatest NBA Team of all time starts with this group of nine teams. Now it's a question of what you value more - offense or defense.

At the top left of the line, we find teams like the 2000 Lakers grouped near the '98 Bulls, '99 Spurs, and the '04 Pistons. This portion of the graph certainly favors defensive ability with those teams defining the space. Now let's go to the opposite end of the line, where we find the '16 and '17 Warriors in their own space. They won 67 and 73 games respectively in those back-to-back seasons while shooting the ball like no team ever has; it's clear this space in the graph is defined by offensive prowess.

The teams between these poles are the perfect storm of both worlds and can be looked at as the most well-formed, balanced, and successful teams in the history of the NBA. I'll leave it for the reader to decide, but it's nearly impossible to make a valid argument for any team not above or close to the line in Figure 4.

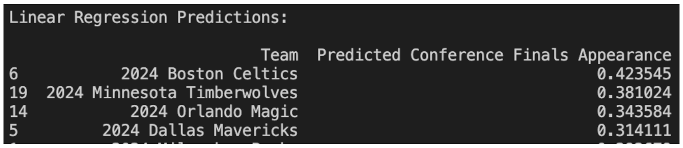
On the opposite side of the clustering, you'll find some of the worst teams in NBA history. I won't get into this side of the plot, but the 2012 Charlotte Bobcats stand out immediately. For context, they had a -15.2 net rating and went 7-59 on the season, the lowest winning percentage of all time (2011-2012 started with a lockout, so the season was shortened).

4.3 Predictions

4.3.1 Conference Finals

The final leg of this project involved predicting which current NBA teams are most likely to not only win the NBA finals this year, but also which teams have the best shot at making the conference finals. I used four models to make predictions for every team. Let's look at some predicted confidence scores for the conference finals.

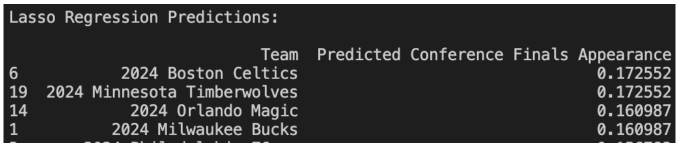
Standard linear regression favors the Celtics and Magic to face off in the Eastern Conference finals while the Timberwolves and Mavericks play in the West. I chose regression models because the current test set is being built as the season progresses. It is rare to see a confidence score above 0.500 this early in the season, and I decided setting a threshold to a value lower than that would lead to some very premature classifications, most of them not being able to classify a single champion.



```
Linear Regression Predictions:

                          Team  Predicted Conference Finals Appearance
6           2024 Boston Celtics                                0.423545
19  2024 Minnesota Timberwolves                                0.381024
14            2024 Orlando Magic                               0.343584
5         2024 Dallas Mavericks                                0.314111
```

Figure 5: Linear Regression confidence scores prediction

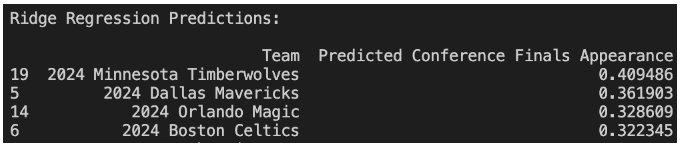
The Lasso regression results highlight the issues addressed in the previous section of this paper. It appears many of the regression coefficients for attributes were struck down to zero. The teams predicted are similar to the linear regression predictions, but the values are far closer together.



```
Lasso Regression Predictions:

                          Team  Predicted Conference Finals Appearance
6           2024 Boston Celtics                                0.172552
19  2024 Minnesota Timberwolves                                0.172552
14            2024 Orlando Magic                               0.160987
1          2024 Milwaukee Bucks                                0.160987
```

Figure 6: Lasso Regression confidence scores prediction

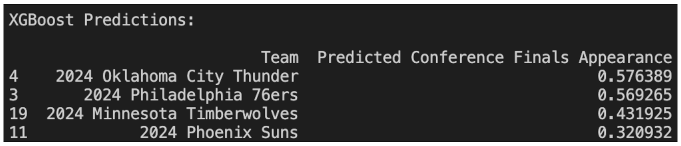
The ridge regression model is most confident in the same teams squaring off in the conference finals as our standard linear regression model, and the scores are extremely similar. There is some variance outside of the top 4, but in general, these results corroborate with our linear model nicely.



```
Ridge Regression Predictions:

                          Team  Predicted Conference Finals Appearance
19  2024 Minnesota Timberwolves                                0.409486
5         2024 Dallas Mavericks                                0.361903
14            2024 Orlando Magic                               0.328609
6           2024 Boston Celtics                                0.322345
```

Figure 7: Ridge Regression confidence scores prediction

Lastly, we have the XGBoost model. These scores are marginally higher than the previous models, and the teams predicted are as well. If I were to classify whether a team does or does not reach the conference finals, this would currently be the best model to use.



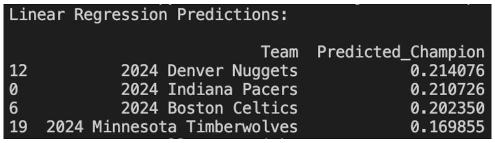
```
XGBoost Predictions:

                          Team  Predicted Conference Finals Appearance
4    2024 Oklahoma City Thunder                                0.576389
3       2024 Philadelphia 76ers                                0.569265
19  2024 Minnesota Timberwolves                                0.431925
11            2024 Phoenix Suns                                0.320932
```

Figure 8: XGBoost Regression confidence scores prediction

I highly recommend going in and running the *predicting.py* script yourself. These predictions change as the season progresses, and you'll see confidence scores for every team rather than just the top 4. Confidence scores that are nearly the same mean the model is equally confident that those teams will reach the conference finals.

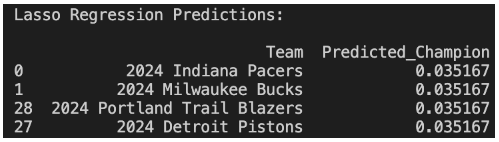
4.3.2 NBA Champion

Below are the same algorithms trained for predicting confidence scores for teams to be crowned the NBA champions this year. The top 4 teams are displayed, but I encourage running the *predicting.py* script to see values for every team. Note that these scores are lower than the conference final prediction scores. There can only be one champion per season compared to four conference finalists, and the test data is not a full season's worth of data.



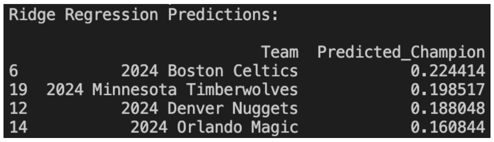
```
Linear Regression Predictions:

                           Team  Predicted_Champion
12         2024 Denver Nuggets            0.214076
0          2024 Indiana Pacers            0.210726
6          2024 Boston Celtics            0.202350
19  2024 Minnesota Timberwolves           0.169855
```

Figure 9: Linear Regression Champion Predictions



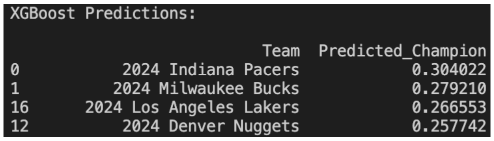
```
Lasso Regression Predictions:

                           Team  Predicted_Champion
0          2024 Indiana Pacers            0.035167
1         2024 Milwaukee Bucks            0.035167
28  2024 Portland Trail Blazers           0.035167
27         2024 Detroit Pistons           0.035167
```

Figure 10: Lasso Regression Champion Predictions



```
Ridge Regression Predictions:

                           Team  Predicted_Champion
6          2024 Boston Celtics            0.224414
19  2024 Minnesota Timberwolves           0.198517
12         2024 Denver Nuggets            0.188048
14          2024 Orlando Magic            0.160844
```

Figure 11: Ridge Regression Champion Predictions



```
XGBoost Predictions:

                           Team  Predicted_Champion
0          2024 Indiana Pacers            0.304022
1         2024 Milwaukee Bucks            0.279210
16     2024 Los Angeles Lakers            0.266553
12         2024 Denver Nuggets            0.257742
```

Figure 12: XGBoost Regression Champion Predictions

The linear regression model is equally confident that the Nuggets, Pacers, and Celtics can win the NBA finals based on the current season data accumulated. The Lasso regression model is interesting; it predicts the same value for every single team, 0.035167. This is due to the lack of variability for our target variable in the training data. Only one team per season can win the finals, which correlates to a 1 in the training data. Every other team from that season has a 0 in the target column.

Ridge regression and XGBoost regression have clearer favorites. Ridge predicts the Celtics to be this year's title holders with the most confidence, while the Timberwolves, Nuggets, and Magic are not far behind. XGBoost currently has the most confidence in the Pacers to hoist a banner, with similar confidence in the Bucks, Lakers, and Nuggets.

4.3.3 How "Good" Are These Results

The clustering results provide statistical evidence for the greatest teams of all time. There is no objective answer to this question, but this narrows it down to a handful of teams deemed worthy of the conversation. It's impossible to legitimately discredit any of the standout teams in this clustering model.

As for the prediction results, time will tell how accurate these confidence scores are. It's about a quarter of the way through the NBA season, and as mentioned earlier, these predictions are dynamic.

Let's look at the predictions from the figures above. The ridge regression model has the highest confidence in the Boston Celtics to win the finals and has the highest confidence score for the Timberwolves, Mavericks, Magic, and Celtics to reach the conference finals this year. If these turn out to be the exact Eastern and Western Conference matchups this year, with the Celtics beating either the Timberwolves or Mavericks in the final, then the model was perfect.

Now let's suppose only two of those teams even make the conference finals, the Celtics and the Mavericks, while two teams with similar confidence scores, say the 76ers and the Suns, are the other two teams to make it. The model should still be considered "good" if the discrepancy between predictions is small (ie. confidence score for the Magic: 0.3645 and confidence score for the 76ers: 0.3429). If there is a large discrepancy between a predicted team and the actual team, that's when the model is underperforming. The same can be said for predicting the NBA Champion. Say a team with the fifth-lowest (or worse) confidence score in multiple models ends up winning the title. Unless a bang-average team went on a miracle playoff run, this scenario would likely happen because due to the models overlooking said team.

5 Conclusion

Through agglomerative clustering, there is now statistical evidence backing a handful of teams as the greatest to ever play in the NBA. The final choice is left up to the reader. Through various regression models, each team from this season has a confidence measure for them to make the conference finals, and for them to be crowned champion. While this project has produced my desired results, a major constraint was the amount of data used. Statistics like injury information, star player analytics, matchup histories, and more would take these predictions to the next level and will be something I look into soon. I encourage the use of the prediction code throughout the season to see how the predictions change as the landscape of the league becomes more clear.