Predicting NBA Games

Using Machine Learning to predict which team will win in an NBA matchup

---

Nahom Debela
Washington State University

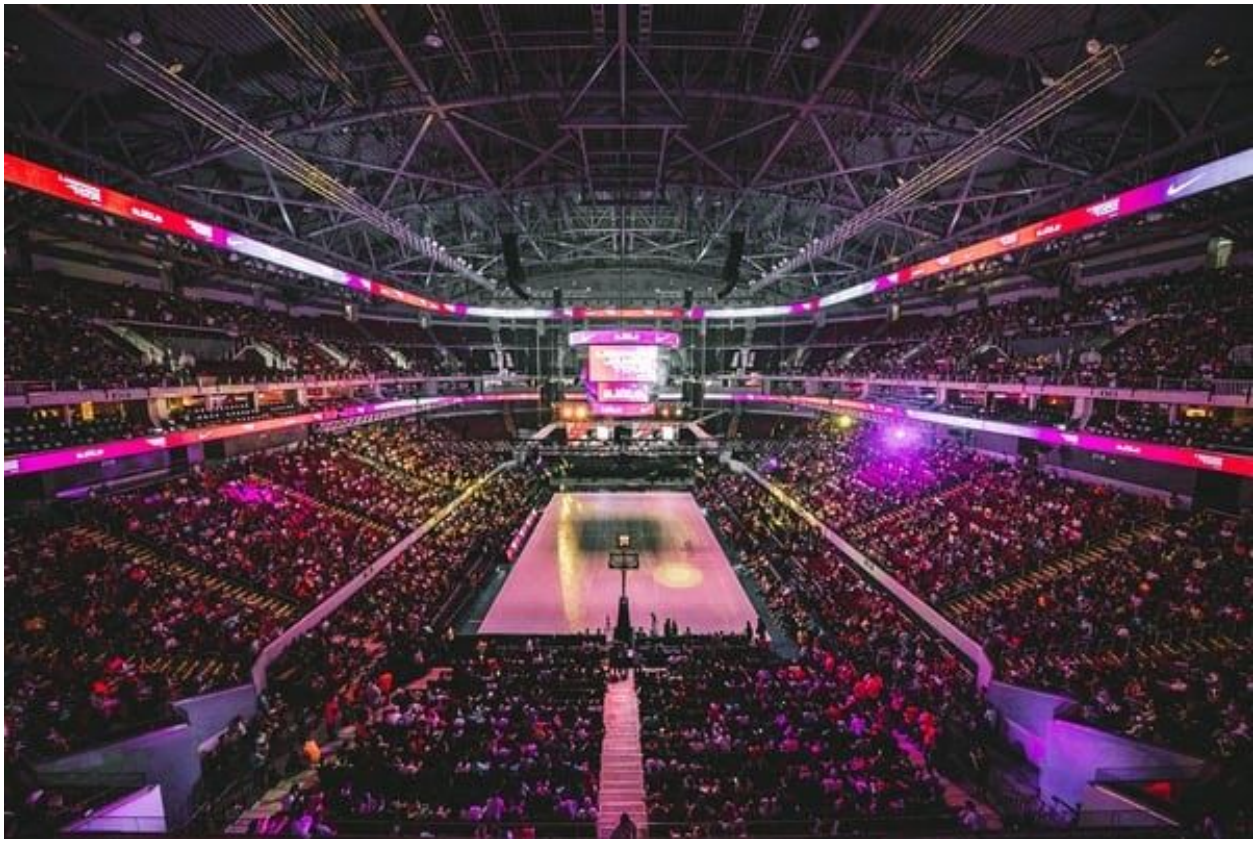


TABLE OF CONTENTS

1 Introduction

Data Analysis in the NBA has grown tremendously over the last couple of decades. Today every NBA team records their data at a high level and has an assigned team of data scientists. The growth of data analytics and my love for basketball has driven me to find out for myself just how effective data analysis is in the NBA.

NBA data analysis is conducted professionally not only by NBA data scientists but by sports betting companies to gain an edge and a profit.

“In reality, to maximize profit, bookmakers employ teams of data scientists to analyze decades of sports data and develop highly accurate models for predicting the outcome of sports events and giving odds to their advantage.” (Nguyen, 2020)

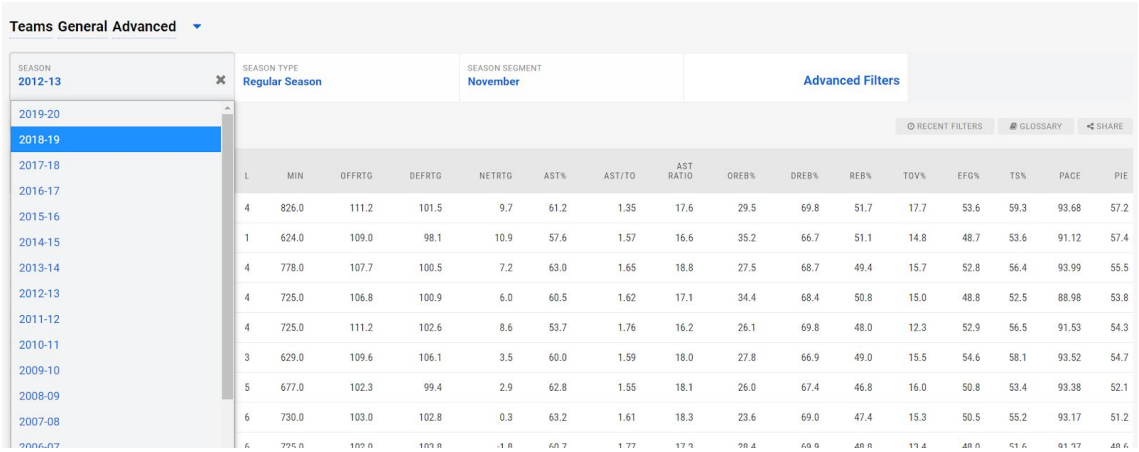
2 Data Mining Task

I will attempt to predict the outcomes of NBA matchups. In my exploration I will answer the following questions:

1. What NBA statistics have the biggest impact on the game's outcome?
2. What model is the most accurate model for predicting wins?
3. Can I predict NBA matchup outcomes with satisfactory accuracy?

3 Data Preparation

To begin I scraped NBA data from the official site www.NBA.com and www.basketball-reference.com



L	MIN	OFFRTG	DEFRTG	NETRTG	AST%	AST/TO	AST RATIO	OREB%	DREB%	REB%	TOV%	EFG%	TS%	PACE	PIE
4	826.0	111.2	101.5	9.7	61.2	1.35	17.6	29.5	69.8	51.7	17.7	53.6	59.3	93.68	57.2
1	624.0	109.0	98.1	10.9	57.6	1.57	16.6	35.2	66.7	51.1	14.8	48.7	53.6	91.12	57.4
4	778.0	107.7	100.5	7.2	63.0	1.65	18.8	27.5	68.7	49.4	15.7	52.8	56.4	93.99	55.5
4	725.0	106.8	100.9	6.0	60.5	1.62	17.1	34.4	68.4	50.8	15.0	48.8	52.5	88.98	53.8
4	725.0	111.2	102.6	8.6	53.7	1.76	16.2	26.1	69.8	48.0	12.3	52.9	56.5	91.53	54.3
3	629.0	109.6	106.1	3.5	60.0	1.59	18.0	27.8	66.9	49.0	15.5	54.6	58.1	93.52	54.7
5	677.0	102.3	99.4	2.9	62.8	1.55	18.1	26.0	67.4	46.8	16.0	50.8	53.4	93.38	52.1
6	730.0	103.0	102.8	0.3	63.2	1.61	18.3	23.6	69.0	47.4	15.3	50.5	55.2	93.17	51.2

I organized my data in a specific format. So for every individual matchup, I recorded which team won and both of those respective teams average statistics for the month they played.

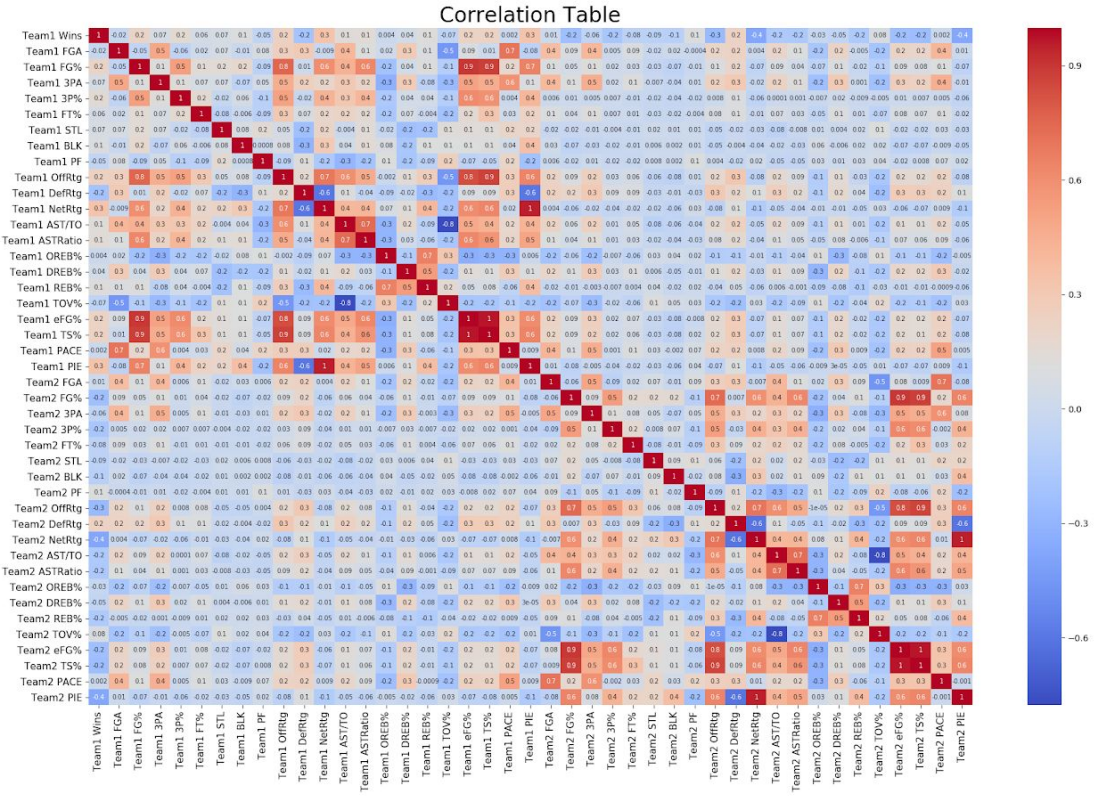
Data Format:

Team 1	Team 2	Which Team Won	Team 1 Monthly Stats	Team 2 Monthly Stats

Once I organized my data, I had a total of 7566 NBA matchups over the span of the 2012-2013 season to the 2018-2019 season!

4 Evaluation Methodology

My next step was to understand which features determine wins. I then conducted exploratory data analysis to figure out what factors were significant to teams winning.

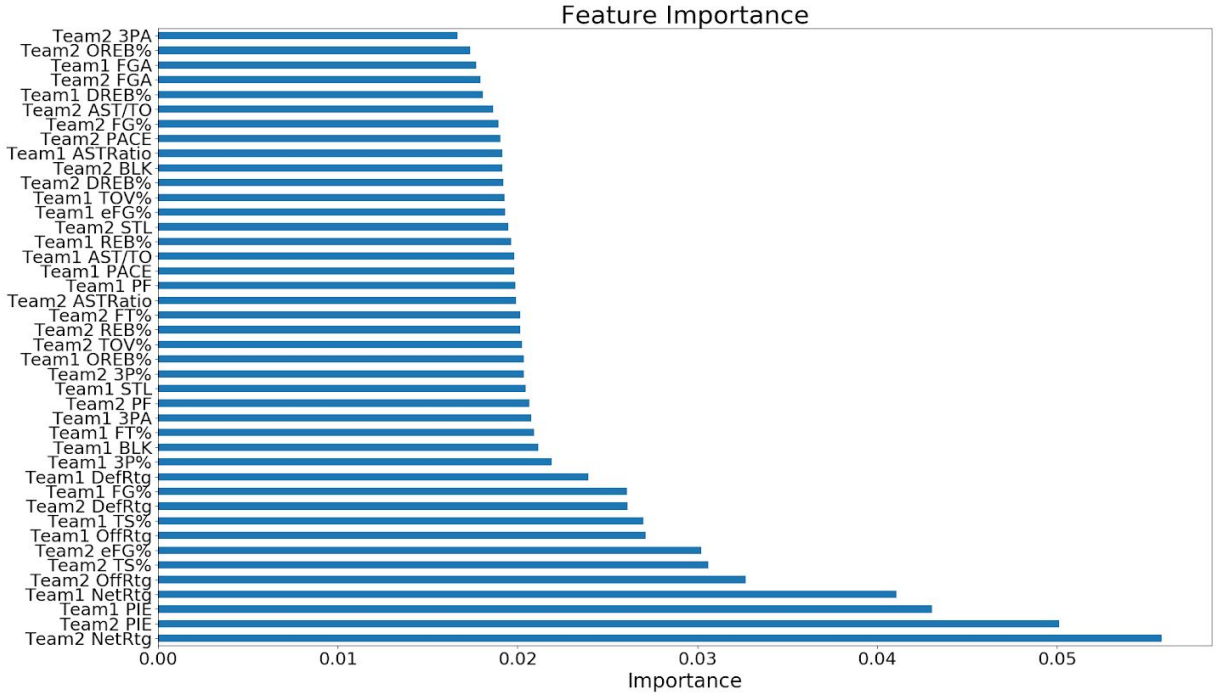


The NBA statistics that were most correlated with winning were:

PIE
NetRtg
OffRtg

From these variables, correlation was at least: ± 0.3

4.1 Evaluation Methodology



Furthermore, I have created a chart which displays the most important features to predicting wins in NBA matchups. From my stats the chart suggests that NetRtg, PIE, OffRtg, TS%, eFG%, and FG% provide the most importance in determining NBA matchup outcomes.

From this chart I also didn't find any reason to remove any features and deemed all features in this chart to be significant. All features in this chart were used in my models and predictions.

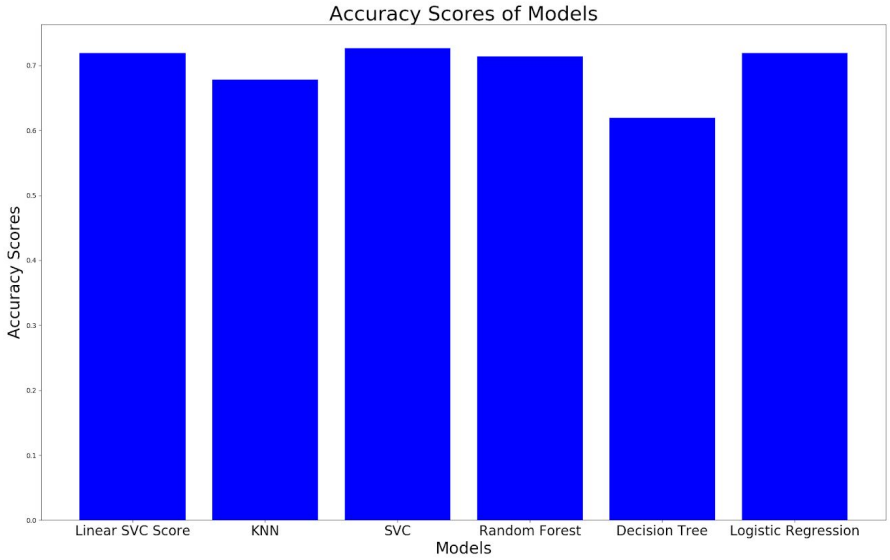
5 Technical Approach

I decided to try out these models to answer my research questions.

1. Linear SVC
2. KNN
3. SVC
4. Decision Tree
5. Random Forest
6. Logistic Regression

I trained each model with my data and computed the accuracy scores for each model with my test data. Linear SVC, Logistic Regression, SVC, and Random Forest all had similar accuracy scores. The accuracy is nowhere near perfect, however they are satisfactory and I am satisfied with the accuracy of my models. Any model I choose from those four should give me just about the same results.

Model	Accuracy
Linear SVC	0.7195
KNN	0.6780
SVC	0.7269
Decision Tree	0.6198
Random Forest	0.7142
Logistic Regression	0.7189



6 Results and Discussion

To test the success of my models, I simply needed to see just how effective my model is in predicting NBA matchups. I made sure to keep 100 samples that I did not use in my training or testing of models. I decided to use my Linear SVC model to predict the outcome of 100 games.

My results showed me that the model is in fact a success. I was able to successfully predict the outcome of 78 out of 100 games.

	Team 1	Team 2	Predicted Outcome	Actual Outcome
0	Brooklyn Nets	Detroit Pistons	Team 1 Lost	Team 1 Lost
1	Memphis Grizzlies	Indiana Pacers	Team 1 Lost	Team 1 Lost
2	Miami Heat	Orlando Magic	Team 1 Won	Team 1 Lost
3	Atlanta Hawks	New York Knicks	Team 1 Lost	Team 1 Lost
4	Cleveland Cavaliers	Toronto Raptors	Team 1 Lost	Team 1 Lost
5	New Orleans Pelicans	Houston Rockets	Team 1 Won	Team 1 Won
6	Minnesota Timberwolves	San Antonio Spurs	Team 1 Lost	Team 1 Lost
7	Utah Jazz	Sacramento Kings	Team 1 Won	Team 1 Won
8	Dallas Mavericks	Phoenix Suns	Team 1 Won	Team 1 Lost
9	Chicago Bulls	Philadelphia 76ers	Team 1 Lost	Team 1 Lost
10	Miami Heat	Washington Wizards	Team 1 Won	Team 1 Won
11	Los Angeles Lakers	Portland Trail Blazers	Team 1 Lost	Team 1 Lost
12	Charlotte Hornets	Orlando Magic	Team 1 Won	Team 1 Won
13	New York Knicks	Brooklyn Nets	Team 1 Lost	Team 1 Lost
14	Atlanta Hawks	Memphis Grizzlies	Team 1 Lost	Team 1 Lost
15	Cleveland Cavaliers	Minnesota Timberwolves	Team 1 Lost	Team 1 Lost
16	Sacramento Kings	New Orleans Pelicans	Team 1 Lost	Team 1 Lost
17	Boston Celtics	Toronto Raptors	Team 1 Lost	Team 1 Lost
18	Indiana Pacers	Milwaukee Bucks	Team 1 Lost	Team 1 Lost

My model had an underdog win % of about 22% after testing 100 matches. So in my sample of 100 games played, 22 games had an outcome that was not expected. In other words the underdog won in 22 of the games. The underdog win % in the NBA is 32.1% (Osborne, 2020). My model is off by about 10% of the real underdog win %. All in all, I am satisfied with the results in this project.

7 Lessons Learned

From the relationships between my team's statistics and wins I learned that some statistics did not have much significance to winning basketball games. I was surprised to see little to no correlation between 3PT% and FT%. In my opinion as a viewer, I would expect teams that shoot higher 3 point percentages to win more games but that as we can see there are many aspects to winning NBA games. The combination of these NBA statistics were used to predict match results and in reality of NBA games simply knowing one or two stats of a team does not constitute that you know the result of the match. Furthermore, I can suggest that with more samples to train and test, my model's accuracy should increase.

Additionally, a more accurate model can be obtained by simply changing the format in which NBA data is gathered. My format consisted of the average statistics of a team throughout a month in the season. To increase the models accuracy the data can be broken down to be more time appropriate. For example instead of recording the average for the month, I can have the average teams statistics for say, the last 3 games. This should lead to more accurate data being used and in return, models would increase from my current Linear SVC score of .72.

8 References

2 Data Mining Task

Nguyen, Tuan Doan. "'Making Big Bucks' with a Data-Driven Sports Betting Strategy." *Medium*, Towards Data Science, 4 Nov. 2020, towardsdatascience.com/making-big-bucks-with-a-data-driven-sports-betting-strategy-6c21a6869171.

6 Results and Discussion

Osborne, Joe. "In Which Sport Do Underdogs Win Most Often?" Odds Shark, 30 Mar. 2020, www.oddsshark.com/sports-betting/which-sport-do-betting-underdogs-win-most-often.

9 Appendices

Variables Explained

FGA	Total number of field goals attempted
FG%	Total number of field goals made / Total number of field goals attempted
3PA	Total number of 3 pointers attempted
3P%	Total number of 3 pointers made / Total number of 3 pointers attempted
FT%	Total number of free throws made / Total number of free throws attempted
STL	Total number of Steals
BLK	Total number of Blocks
PF	Total number of Fouls Committed
OffRtg	Amount of points produced by a team per 100 possessions.
DefRtg	Amount of points a team allows per 100 possessions.
NetRtg	Measure of a team's point differential per 100 possessions
AST/TO	Assist to Turnover ratio
ASTRatio	Percentage of team's field goals a player assists on.
OREB%	Teams Offensive rebound rate
DREB%	Teams Defensive rebound rate
REB%	Teams total rebound rate
TOV%	Estimate of turnovers per 100 possessions
eFG%	Effective Field Goal Percentage - adjusts FG% to account for 3 Pointers
TS%	Measure of shooting efficiency - accounts for all facets of scoring
PACE	Number of possessions a team uses per game
PIE	What % of game events did the team achieve