Who Would Have Made the 2020 NBA Playoffs?

Predicting NBA Playoff Teams with Machine Learning

1 Overview

In this analysis, I wanted to predict which NBA teams were going to make the playoffs in the 2019-2020 NBA season that was abruptly cut short due to the COVID-19 pandemic. Fans across the country were bummed and looking forward to a competitive, high-octane playoff season. The two top-seeded teams, the Los Angeles Lakers backed by LeBron James, and the Milwaukee Bucks lead by Giannis Antetokounmpo, were slated to duel for the championship title.

At the time of cancellation (March 11, 2020), a few teams clinched their playoff berths, others were jostling for their seeding positions, and some were fighting for the coveted 8th seed. Usually, playoffs begin in the middle of April, and the top 8 teams of each conference (Eastern and Western) are eligible. I wanted to know which teams were going to make the playoffs, had the season continued.

I used several classification machine learning models that predicted playoff teams based on aggregate team statistics over the course of the regular season. I trained the models on data from 15 NBA seasons (2004-2005 through 2018-2019) and predicted outcomes for the 2019-2020 season.

1.1 Results

The final model predicted 17 teams to be in the 2020 NBA Playoffs (10 teams from the West, 7 in the East).

- The model predicted two extra teams to qualify out of the West, who were not in the top 8 at the time.
- The model failed to predict one team to qualify out of the East, which was in the top 8 at the time.
- Out of all the predicted playoff teams, 15 of them were in the top 8 seeds in their respective conferences at the time of cancellation.

2 Data Collection

The data were scraped[1] and cleaned from the ESPN website and consist of information for 30 teams over 15 seasons. The data contain traditional basketball statistics (aggregated on a per-game basis), and an indicator for whether or not that team made the playoffs in that season. The goal here is to find if any of these variables are indicative of a team's chances of being a playoff team.

Below are a sample of some of the metrics in the data and the first few records of the dataset:

- `"PTS" - Total points scored.`
- `"FGA" - Total field goal attempts (shots).`
- `"X3P" - Total three-point attempts.`
- `"FTM" - Total free-throws made.`
- `"REB" - Total rebounds collected.`
- `"AST" - Total assists.`
- `"STL" - Total steals.`
- `"TO" - Total turnovers (loss of posession of the ball).`
- `"PF" - Total personal fouls committed.`
- `"playoffs" - Indicator for playoffs (1 = in playoffs).`

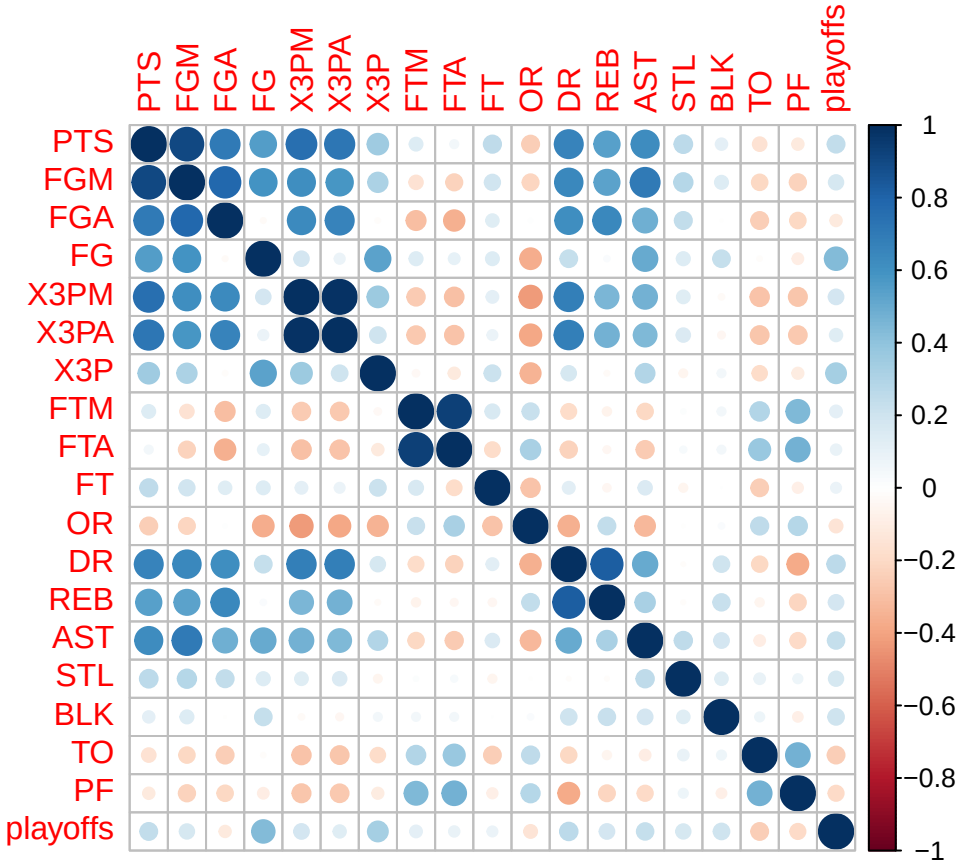
[1] The dataset and script for scraping and cleaning can be found in the "Data" and "Scripts" folder, respectively, on the GitHub repository

teams	PTS	FGM	FGA	STL	BLK	TO	PF	season	playoffs
Phoenix Suns	110.4	40.9	85.6	7.0	5.5	13.7	19.1	2005	1
Sacramento Kings	103.7	39.1	85.1	8.2	3.9	13.1	20.5	2005	1
Dallas Mavericks	102.5	37.3	81.6	8.6	5.6	13.4	22.3	2005	1
Miami Heat	101.5	37.8	77.7	6.4	5.8	13.7	22.1	2005	1
Boston Celtics	101.3	37.1	79.4	8.1	5.2	15.8	24.4	2005	1

3 Data Exploration

3.1 Correlation Plot

First let's explore any variables significantly related to the outcome variable. The correlation plot below shows that most of variables are strongly correlated with each other, and have moderate correlations to the target variable, `playoffs`.



4 Modeling the Data

Since the goal of this analysis is to classify a team as a playoff team or not, I decided to use classification models that specialize in predicting categorical (binary) outcomes. The data was split into a training and test set (70%/30% split) and the models were fit on the training data using repeated 10-fold cross-validation.

4.1 Variables Included in the Model

The following variables were selected to be included in the models based on stepwise-AIC logistic regression:

`PTS, FGM , X3PA, FTM ,FTA, FT, OR, DR, AST, STL, TO`

The models and their respective performances on the training data are shown below:

Model	Accuracy
Logistic Regression	0.8798137
Random Forest	0.7080153
Naive Bayes	0.6630190
SVM Linear	0.8589815

4.2 Brief aside on the Logistic Regression Model

In the Logistic Regression, we model the Bernoulli data-generating process of the outcome variable `"playoffs"` $P(Playoffs = 1) = p$ by assuming a linear relationship between predictor variables and the log-odds of the event that $P(Playoffs = 1)$.

This model takes the form:

$$log(\frac{p}{1-p}) = \beta_0 + \sum_{i=1}^{n} \beta_i X_i$$

where p = probability of being in playoffs and X_i = predictor i

Below, the output of the model shows that all predictors are statistically significant ($p < 0.05$). It is interesting to note that the predictors `STLS` and `TO` have coefficients of `2.148` and `-1.597`, respectively. This means:

- On average, a higher amount of turnovers translate to a smaller log-odds (and subsequently probability) of being in the playoffs, holding all other variables constant.
- Similarly, if a team has a high amount of steals, the probability is much greater.

This seems to confirm the idea posited by most basketball gurus that defense is the best offense, and that sticking to fundamentals of the game most often wins championships.

	Coefficient	P.value
(Intercept)	-202.640	0.000
PTS	2.546	0.000
FGM	-7.795	0.000
FG	2.898	0.000
X3PA	-0.697	0.010
FTM	-7.846	0.003
FTA	4.065	0.029
FT	1.418	0.016
OR	1.685	0.000
DR	1.204	0.000
AST	0.426	0.007
STL	2.148	0.000
TO	-1.597	0.000

5 Model Evaluation

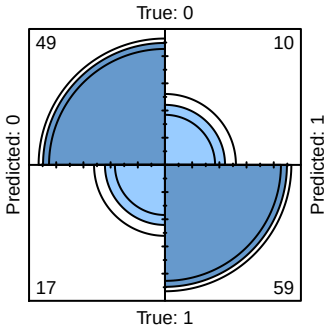
5.1 Confusion Matricies

Now that the models are trained on the training data, we can evaluate their performance on the test sets and see how well each can distinguish between a playoff team and a non-playoff team.

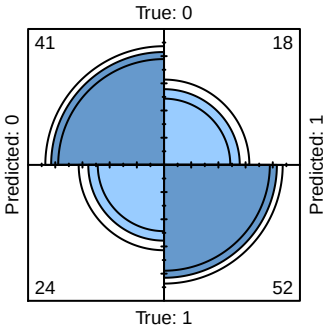
> **For this analysis, I wanted to choose a model that is able to detect a playoff team well, but also limits the amount of playoff teams that it misses (i.e a balance between false positives and false negatives).**

Based on the confusion matrix plots below, the SVM model appears to be the best at predicting out-of-sample data, since it has the lowest False Positive Rate (FPR) and False Negative Rates (FNR) (rows represent true values, and columns represent the predicted values).

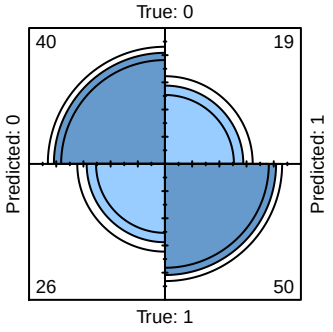
Logistic Model



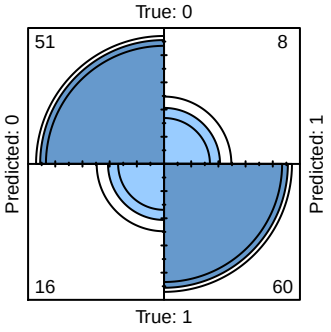
Random Forest Model



Naive Bayes Model

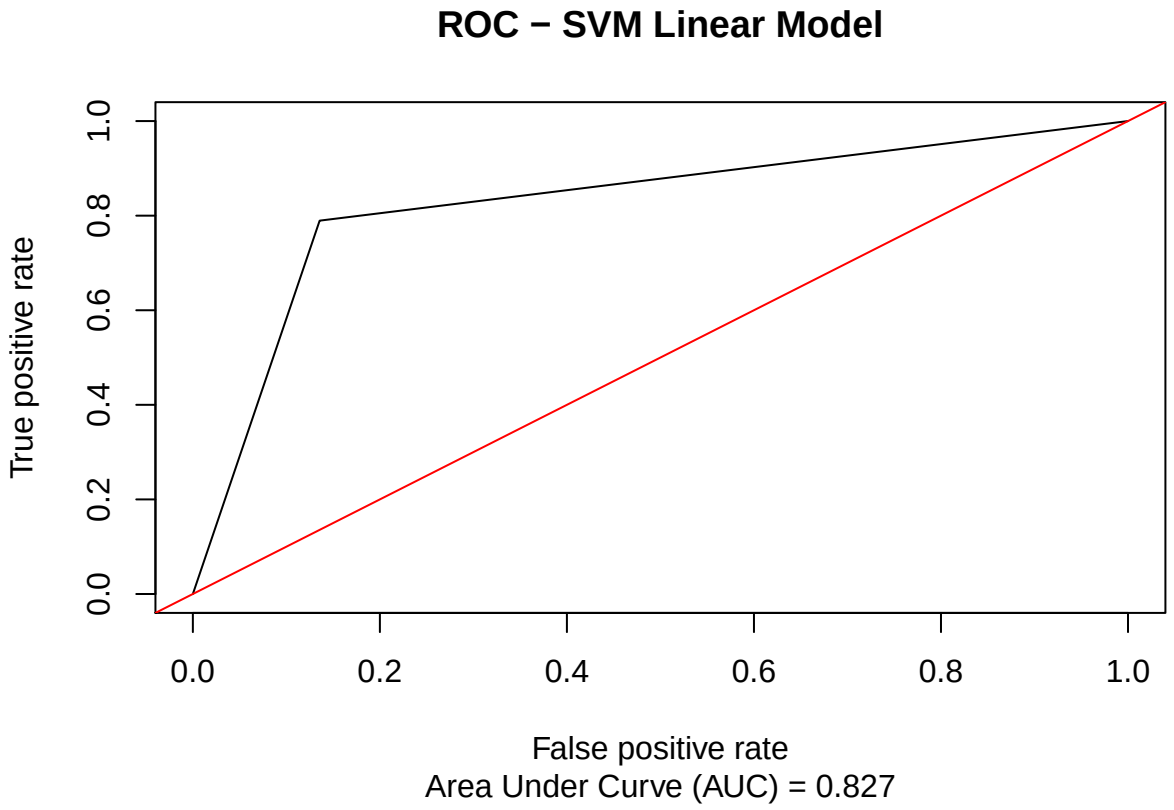


SVM Linear Model



5.2 ROC and AUC

To confirm the selection of the SVM model, we can also look at the ROC curve and the associated AUC metric (area under the curve).



The AUC's for the other models are shown below:

Model	AUC Metric
SVM Linear Model	0.83
Logistic Model	0.8
Random Forest Model	0.69
Naive Bayes Model	0.67

6 Predicting 2020 NBA Playoff Teams

According to the metrics, the SVM model performed best at classifying playoff teams. Let's see how it performs on predicting the 2020 NBA season.

The predictions are shown below along with the teams that were among the top 16 in the league at the time the season was cancelled:

	Predicted Playoff Teams	Actual Top 16 Teams
1	Milwaukee Bucks	Milwaukee Bucks
2	Houston Rockets	Houston Rockets
3	Dallas Mavericks	Dallas Mavericks
4	LA Clippers	LA Clippers
5	Los Angeles Lakers	Los Angeles Lakers
6	San Antonio Spurs	Brooklyn Nets
7	Boston Celtics	Boston Celtics
8	Toronto Raptors	Toronto Raptors
9	Memphis Grizzlies	Memphis Grizzlies
10	Phoenix Suns	Miami Heat
11	Miami Heat	Utah Jazz
12	Utah Jazz	Oklahoma City Thunder
13	Oklahoma City Thunder	Denver Nuggets
14	Denver Nuggets	Philadelphia 76ers
15	Philadelphia 76ers	Indiana Pacers
16	Indiana Pacers	Orlando Magic
17	Orlando Magic	-

This model performed pretty well, as it predicted that most of the teams that were in the top 16 in the league would eventually make it to the playoffs. This seems reasonable because usually around March, teams begin to solidify their playoff berths, and the teams who are top 8 in their respective conferences at that time are the ones that will be in the playoffs. There are obvious exceptions, for instance, a team could go on a significant winning or losing streak, or, there is fierce competition and only a few games separate the last few seeds of the conference.

- The model predicted two teams in the West as playoff teams (San Antonio Spurs and Phoenix Suns) that were not in the top 8 of the West at the time.
- The model also failed to classify the Brooklyn Nets as a playoff team in the East, even though it was in the top 8 at the time.

Let's examine the average statistics of these predicted playoff teams and compare them to the non-playoff teams:

Variable	Playoff Teams	Non-Playoff Teams	Difference (Playoff v. Non-Playoff)
PTS	112.806	109.708	3.098
FGM	41.300	40.223	1.077
X3PA	33.629	34.346	-0.717
FTM	18.035	17.177	0.858
FTA	23.118	22.615	0.502
FT	77.953	75.931	2.022
OR	9.841	10.485	-0.643
DR	35.612	33.615	1.996
AST	24.759	23.815	0.943
STL	7.694	7.623	0.071
TO	13.459	14.415	-0.957

On average, it seems that there are clear differences between playoff and non-playoff teams, namely in PTS, FGM, ASTS, DR.

We can conduct a difference in means T-test for each variable to see if these differences are statistically significant, and verify which aspects of the game that playoff teams generally perform better at.

Variable	T-Test P-Value
PTS	0.029
FGM	0.031
X3PA	0.627
FTM	0.147
FTA	0.457
FT	0.053
OR	0.043
DR	0.009
AST	0.152
STL	0.833
TO	0.011

The results show that on average, playoff teams have statistically significant differences in the following metrics (at a significance level 0.05):

`PTS, FGM, FT, OR, DR, TO`

This means that generally, teams who score more points, make more field goals, and handle the ball well are the teams ending up in the playoffs.

7 Key Takeaways

This analysis showed that traditional basketball statistics are useful in predicting NBA playoff eligibility. It provided insight into specific factors of a team that can lead to better performance.

The results suggest that if teams understand their metrics early in the season, they can focus on improving in specific areas of their play style (for example, limiting turnovers, or ensuring excellent shot selection and rebounding) to better their chances of qualifying for the playoffs. This also allows coaches and basketball management to make better informed decisions in terms of trades, and coaching styles.