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 (including me) 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. I wanted to know which teams were going to make the playoffs, had the season continued.

I used several supervised machine learning models that predicted playoff teams by using 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. The model incorrectly predicted one extra team, and out of all the teams predicted to be in the playoffs, 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 and cleaned from the ESPN website and consist of information for 30 teams over 15 seasons. The data contain several metrics for each team, 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 aggregate statistics are indicative of a team's chances of being a playoff team. Below are a sample of some of the metrics included in the data:

- `"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).`

The script for scraping, cleaning, and formatting the data from the ESPN website can be found in the "scripts" folder.

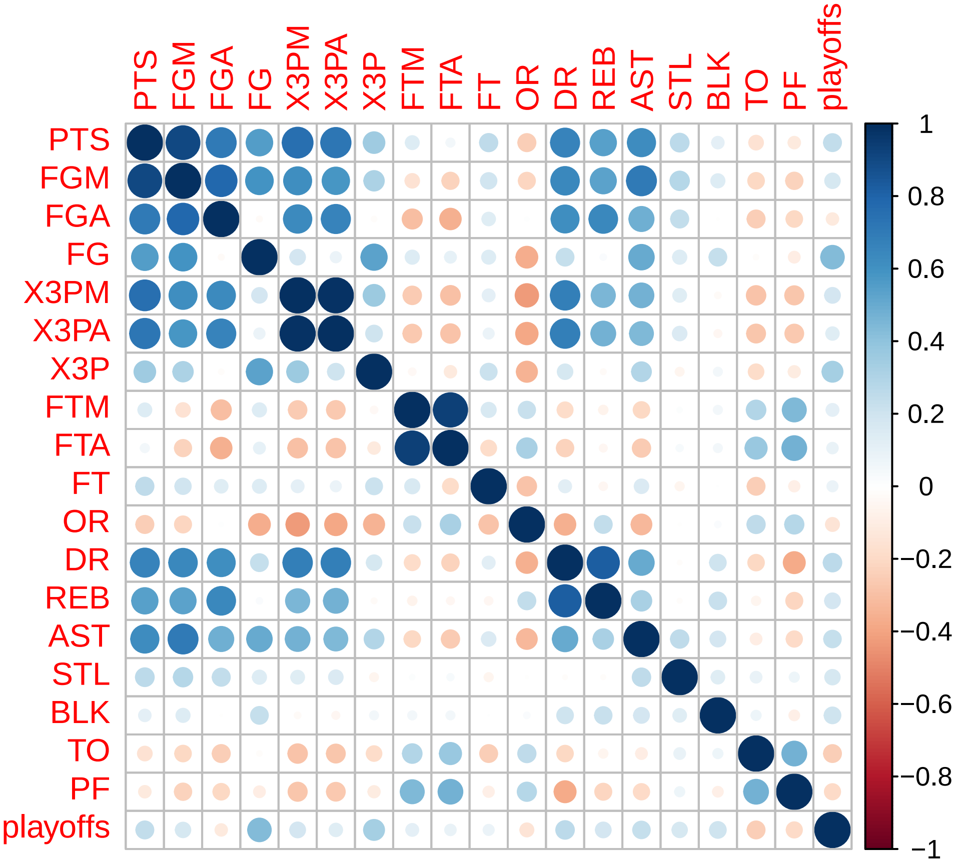
Below is a sample of the cleaned dataset:

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, but have weaker 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 several 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.

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

4.1 Brief aside for the Logistic Regression Model

In the Logistic Regression model, 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 `"TLS` and `TO` have coefficients of `2.148` and `-1.597`, respectively.

- 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.
- 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

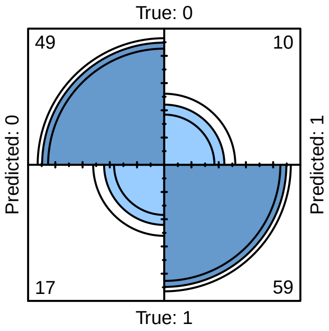
5.1 Confusion Matricies

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.

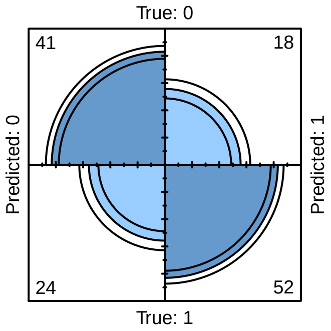
Based on the confusion matrix plots below, the SVM model appears to be the best at predicting out-of-sample data, since it has the lowest False Positive Rate (FPR) and False Negative Rates (FNR).

- **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).**

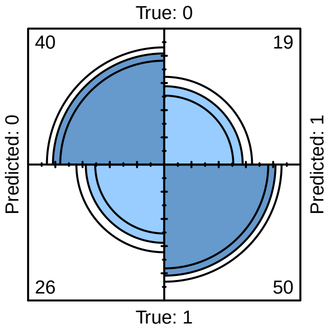
Logistic Model



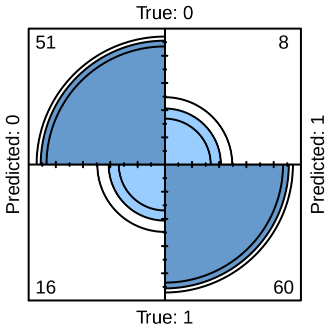
Random Forest Model



Naive Bayes Model

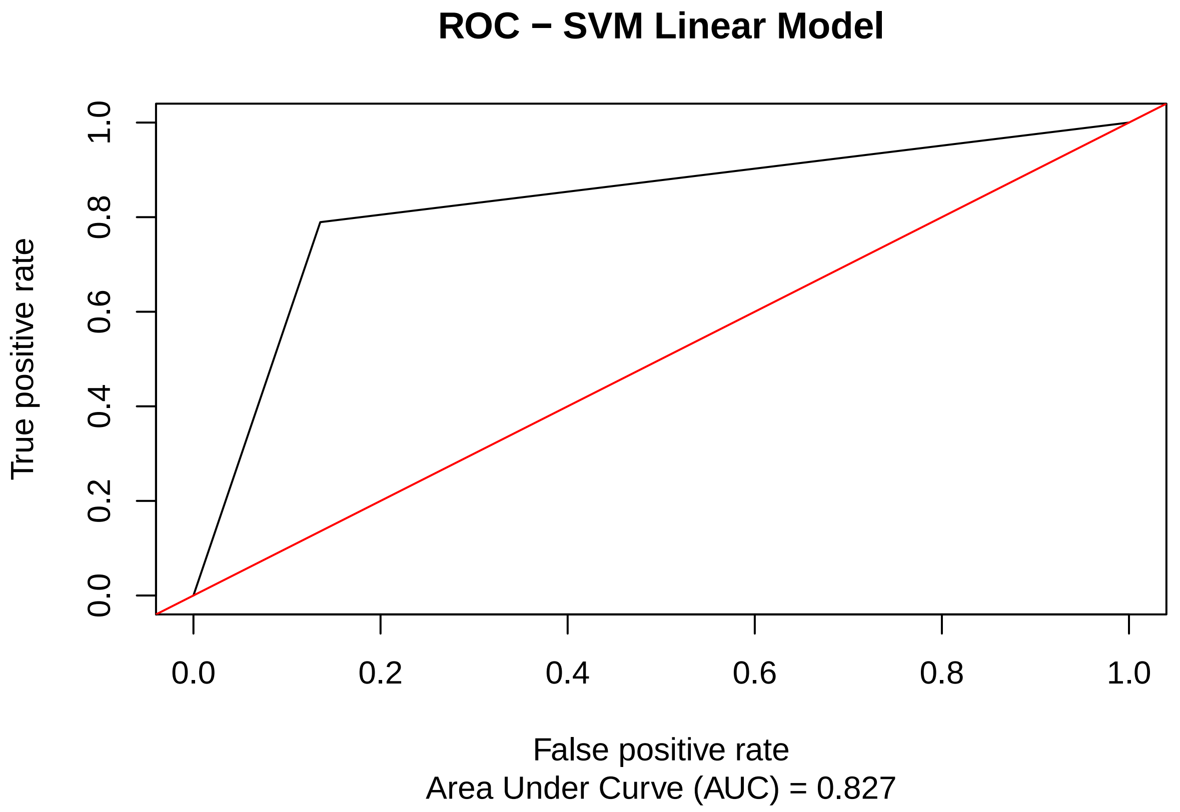


SVM Linear Model



5.2 ROC and AUC

To confirm the selection of the SVM model, we can also look at the ROC curve and calculate the AUC (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

From these metrics, it is clear that SVM performed best at classifying playoff teams. Let's see how it performs on 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 would eventually make it to the playoffs. This makes sense, because usually around March, teams begin to solidify their playoff berths and the top 16 teams are the ones that will be in the playoffs.

- The model incorrectly predicted two teams as playoff teams (San Antonio Spurs and Phoenix Suns): they were not in the top 16.
- The model also failed to classify the Brooklyn Nets as a playoff team, even though it was in the top 16 at the time of cancellation.

The other model predictions are shown below for reference:

	Logistic Model	Random Forest Model	Naive Bayes Model
1	Milwaukee Bucks	Houston Rockets	Milwaukee Bucks
2	Dallas Mavericks	Dallas Mavericks	Houston Rockets
3	LA Clippers	LA Clippers	Dallas Mavericks
4	Los Angeles Lakers	Washington Wizards	New Orleans Pelicans
5	San Antonio Spurs	Los Angeles Lakers	LA Clippers
6	Boston Celtics	Portland Trail Blazers	Washington Wizards
7	Toronto Raptors	San Antonio Spurs	Los Angeles Lakers
8	Miami Heat	Boston Celtics	Portland Trail Blazers
9	Utah Jazz	Toronto Raptors	Minnesota Timberwolves
10	Oklahoma City Thunder	Memphis Grizzlies	San Antonio Spurs
11	Denver Nuggets	Phoenix Suns	Boston Celtics
12	Philadelphia 76ers	Miami Heat	Toronto Raptors
13	Orlando Magic	Utah Jazz	Memphis Grizzlies
14	NA	Oklahoma City Thunder	Phoenix Suns
15	NA	Denver Nuggets	Miami Heat

	Logistic Model	Random Forest Model	Naive Bayes Model
16	NA	Philadelphia 76ers	Atlanta Hawks
17	NA	Indiana Pacers	Utah Jazz
18	NA	Sacramento Kings	Brooklyn Nets
19	NA	Chicago Bulls	Oklahoma City Thunder
20	NA	Orlando Magic	Denver Nuggets
21	NA	NA	Philadelphia 76ers
22	NA	NA	Indiana Pacers
23	NA	NA	Sacramento Kings
24	NA	NA	Detroit Pistons
25	NA	NA	Cleveland Cavaliers
26	NA	NA	Chicago Bulls
27	NA	NA	Orlando Magic
28	NA	NA	Golden State Warriors
29	NA	NA	Charlotte Hornets

7 Conclusion