Predicting Probability of Scoring In The NBA

Jiang, Han h_jiang@uncg.edu
Martin, David drmartin@uncg.edu
Miller, Lee ljmille3@uncg.edu
Snow, Michael mrsnow@uncg.edu

Introduction

The idea of leveraging data in the context of professional and amatuer athletics has been around for a long time, but with advancements in data collection tools, machine learning, and other technologies, the field of sports analytics has greatly expanded during the last decade. Teams are turning to analytics to select players based on their predicted performance and then using data to put those players in situations that will maximize their output, perform at a high efficiency, and ultimately have the highest probability of winning games and championships. Sports broadcast companies are also showing a great interest in sports analytics as they continue to search for ways to incorporate situational data into their broadcasts of events.

For this study, the focus is on NBA shooting statistics. We examined data from shot attempts made during the 2015 NBA season. Our goal was to develop a probabilistic model for a successful shot attempt using several indicators related to shot characteristics and the player's activity during the shot attempt.

In addition to creating a model for the probability of a made shot based on multiple features, we also proposed five questions that we wanted to answer to give additional insight into shot performance of teams and players for the 2015 NBA season. The questions were:

1. What was the average number of shots made per game for winning teams versus losing teams?

2. What was the overall field goal percentage of teams who made the playoffs in 2015 versus teams that did not make the playoffs?

3. Which three players had the highest 3-point shooting percentage among all players that attempted at least ten 3-point shots?

4. What was the average number of dribbles taken before attempting a made shot versus attempting a missed shot? And was the number of dribbles correlated with the likelihood of making a shot?

5. How does the distance from which the shot was taken or how far away the closest defender was change the likelihood of the shot being made?

By examining and aggregating the data appropriately, we were able to answer these questions as outlined in the analysis section below as well as develop a classification model using logistic regression.

Data

The primary dataset used originated from the NBA's REST API and was later published on www.kaggle.com/dansbecker/nba-shot-logs. The structure of the data involves a single row per shot attempt during the 2015 NBA season (128,069 total shots attempted). The attributes measured for each shot attempt which we will include in the analysis were player, matchup, win/loss, how many shots per game by player, time remaining, number of dribbles, amount of time in possession of the ball, shot distance, two vs. three point attempt, defender distance, and shot outcome (make/miss).

The secondary dataset gives information on whether each team made the playoffs in 2015 which we joined with our primary dataset to help answer our research question pertaining to differences in shooting percentage based on playoff appearance. This dataset was constructed using information from www.basketball-reference.com/playoffs/NBA_2015.html.

Methods

To prepare the data for analysis, there were several cleaning measures that were required. First, we checked for missing data and found over 5,500 missing values for the "shot clock" variable. These values were most likely not missing at random, as the shot clock is turned off when a team takes possession of the ball with less than 24 seconds remaining in the period. Therefore, missing shot clock values were replaced with game clock values as an equivalent proxy. In addition, the "touch time" measure included approximately 3,300 values that were less than or equal to zero. As touch time must be a positive value, the validity of these observations could not be determined and were handled by listwise deletion to remove the potential bias of these values.

Our original dataset did not contain a variable that explicitly listed the team each player was on. However, it did contain a "Matchup" variable where each value was in the form " 'Date' - 'Team' @/vs. 'Opponent' ". To extract the "Team" component from this value, we introduced code to split each value on the '-' and '@/vs' characters, stripped out any remaining whitespace, created a new variable called "Team" from the result, and dropped the other unnecessary columns.

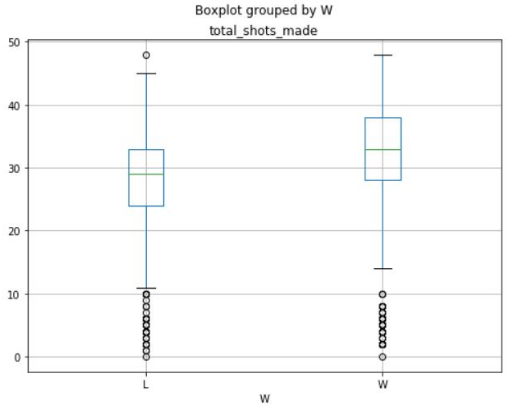
To explore the shape of each numeric variables distribution, we created a histogram (see code template) of values for each numeric feature. We discovered positive skewness in several variables (shot number, number of dribbles, and closest defender distance), which could require a logarithmic transformation for certain future analysis, and a distinct bimodal distribution in shot distance.

The two datasets were stored in SQL tables and an integrated table was built by performing an inner join on the team identifier. The resulting combined table was then queried to return the information needed to answer our five questions above.

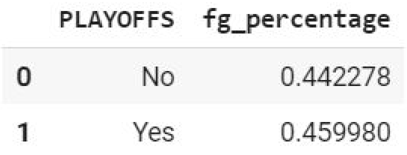
Analysis

Database queries and visualizations were created to explore the summary questions proposed.

First, what was the average number of field goals per game for winning teams versus losing teams? In order to answer this question, a query was written that summed field goals made and grouped the results by game ID and win/loss. Creating a box plot allowed us to determine that the dataset cannot be used to understand total game characteristics. The box plot shows that the dataset does not have complete shot data for each game after cleaning. The dataset is still appropriate for individual shot analysis.

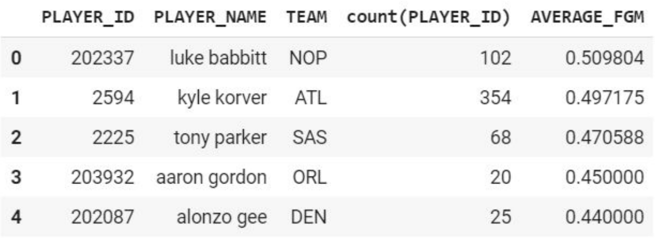


Next, we compared the overall field goal percentage of teams who made the playoffs in 2015 versus teams that did not make the playoffs. This involved writing a query that joined the shot information with the team playoff information and averaging the field goal made variable and grouping it by whether the team made the playoffs or not.



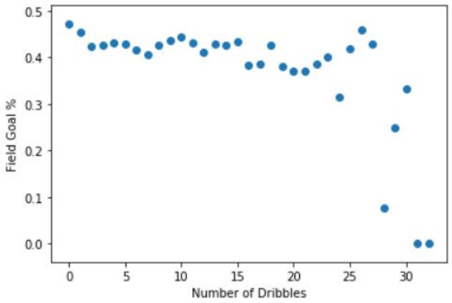
	PLAYOFFS	fg_percentage
0	No	0.442278
1	Yes	0.459980

In order to identify the players with the highest field goal percentage (with a minimum of 10 attempts) a query was written to include only 3-point shots, averaging field goal made by player ID and including only results with a player ID count greater than or equal to 10. Luke Babbit led the league in field goal percentage on 3-point shots.

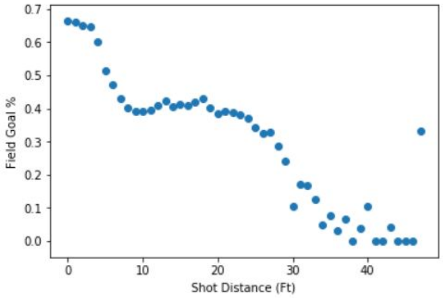


	PLAYER_ID	PLAYER_NAME	TEAM	count(PLAYER_ID)	AVERAGE_FGM
0	202337	luke babbitt	NOP	102	0.509804
1	2594	kyle korver	ATL	354	0.497175
2	2225	tony parker	SAS	68	0.470588
3	203932	aaron gordon	ORL	20	0.450000
4	202087	alonzo gee	DEN	25	0.440000

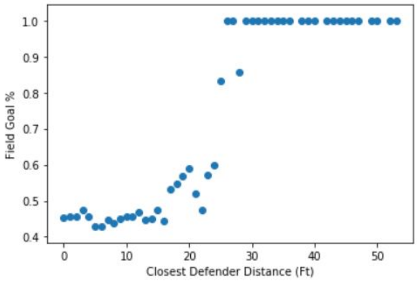
The relationship between dribbles and likelihood of making a shot (field goal percentage) was estimated by plotting the data returned from a query grouping average field goals made by the number of dribbles.



To determine the relationship between shot distance and field goal percentage, the shot distance was rounded to the nearest integer and used as the grouping variable in the query.



The relationship between distance of closest defender and field goal percentage was determined in a similar way to the relationship between shot distance and field goal percentage. The closest defender distance was rounded to the nearest foot and used as the group by variable.



Classification Model:

The probabilistic model is to demonstrate the predicting power of selected features: shot number, dribbles, touch time, shot distance, and closest defender distance. To predict whether or not a shot is made (0: missed, 1: made), PySpark uses Binary Logistic regression algorithm. The various steps took to build the model are listed below:

1. Import the new file (shot_new.csv) with the chosen features in the local 'sample_data' folder
2. Initialize PySpark to read the file
3. Import libraries (e.g. VectorAssembler) to prepare data
4. Train, test split
5. Logistics Regression and ROC Curve

All six selected features are cluttered into the "Attribute" feature column to predict the "FGM" response variable in the label column. The whole dataset is split into 7:3 train versus test proportions. The maximum iteration is set to 10 for logistic algorithms to find the maximum likelihood. The ROC curve is performed to test the model performance.

Results

We found that on average the team that won made four more shots than the team that lost. Furthermore, the average field goal percentage for teams that made the playoffs and those that did not were very similar, with the teams making the playoffs shooting about 1.7% better.

The 3 players with the highest 3-point shooting percentage among players that attempted at least ten 3-point shots were Luke Babbit (NOP/51.0%), Kyle Korver (ATL/49.7%), and Tony Parker (SAS/47.1%).

In addition, we found there is not a significant correlation between the number of dribbles and field goal percentage. However, there is a negative correlation between shot distance and field goal percentage; as the distance increases field goal percentage decreases. Interestingly, there was very little change in field goal percentage in shots taken between 10 and 20 feet. This shows that all mid-range shots have a similar likelihood of being made. There is also a correlation between how close the closest defender is and field goal percentage. Field goal percentage increases the farther away the closest defender is. When the closest defender is more than 25 feet away, the field goal percentage is almost 100%.

According to model evaluation results, it is evident that the majority of selected variables are minimally correlated (positively and negatively) to the response variable. To some extent, one variable (shot distance) has a strong positive correlation to the response variable.
Classification Model:

The model coefficient vector is listed below:

[-0.165, 0.0055, -0.1363, -0.1289, -0.1576, 16.3492]

The AUC (Area Under Curve) of the model is:

0.999957393070663

For our logistic regression model, the model coefficient vector suggests a unit increase in shot distance variable can result in a 16.3492 increase in the log-odds of ‘FGM’ variable, holding the rest of independent variables constant. The AUC (0.999) of ROC suggests an approximate 100% sensitivity and specificity in terms of model performance.

Conclusion

Our goal for this study was to develop a probabilistic model for a successful shot attempt using several indicators as well as answer these five questions:

1. What was the average number of shots made per game for winning teams versus losing teams?

2. What was the overall field goal percentage of teams who made the playoffs in 2015 versus teams that did not make the playoffs?

3. Which three players had the highest 3-point shooting percentage among all players that attempted at least ten 3-point shots?

4. What was the average number of dribbles taken before attempting a made shot versus attempting a missed shot? And was the number of dribbles correlated with the likelihood of making a shot.

5. How does the distance from which the shot was taken or how far away the closest defender was change the likelihood of the shot being made.

The extraordinarily high AUC might be influenced heavily by the single variable ‘shot distance’. Without the comparison to another model, the accuracy of the model may vary depending on cross-validation as well as many other potential explanatory variables. However, as a result, the Pyspark logistic regression is a suitable classification model to our dataset.

The main takeaway for our project is that there are many hidden analytical results that can be determined through different SQL commands and visualizations. In terms of sports analytics, teams that wish to recruit and evaluate players’ true performance in-depth can adapt logistic regression models in order to find potential factors that are correlated to the player performance as well as to provide a reference to support the decision making process.