

# An Analysis of NBA Spatio-Temporal Data

by

Megan Robertson

Department of Statistical Science  
Duke University

Date: \_\_\_\_\_

Approved:

---

Sayan Mukherjee, Supervisor

---

Vikas Bhandawat

---

Scott Schmidler

Thesis submitted in partial fulfillment of the requirements for the degree of  
Master of Science in the Department of Statistical Science  
in the Graduate School of Duke University  
2017

# ABSTRACT

## An Analysis of NBA Spatio-Temporal Data

by

Megan Robertson

Department of Statistical Science  
Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_  
Sayan Mukherjee, Supervisor

\_\_\_\_\_  
Vikas Bhandawat

\_\_\_\_\_  
Scott Schmidler

An abstract of a thesis submitted in partial fulfillment of the requirements for  
the degree of Masters of Science in the Department of Statistical Science  
in the Graduate School of Duke University  
2017

Copyright © 2017 by Megan Robertson  
All rights reserved except the rights granted by the  
Creative Commons Attribution-Noncommercial Licence

# Abstract

This project examines the utility of spatio-temporal tracking data from professional basketball games by fitting models predicting whether a player will make a shot. The first part of the project involved the exploration of the data, evaluated its issues, and generated features to use as co-variates in the models. The second part fit various classification models and evaluated their predictive performance. The paper concludes with a discussion of methods to improve the models and future work.

# Contents

<b>Abstract</b>	<b>iv</b>
<b>List of Tables</b>	<b>viii</b>
<b>List of Figures</b>	<b>ix</b>
<b>List of Abbreviations and Symbols</b>	<b>x</b>
<b>Acknowledgements</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 The Data</b>	<b>4</b>
<b>3 Feature Generation</b>	<b>8</b>
3.1 Defensive Information . . . . .	8
3.1.1 Distance to the nearest defender . . . . .	9
3.1.2 Angle between shooter and closest defender . . . . .	9
3.1.3 Side of defender . . . . .	10
3.1.4 Number of close defenders . . . . .	11
3.2 Teammate Information . . . . .	11
3.2.1 Distance to nearest teammate . . . . .	12
3.2.2 Angle between shooter and closest teammate . . . . .	12
3.2.3 Side of teammate . . . . .	12
3.2.4 Number of close teammates . . . . .	13
3.3 Location . . . . .	13

3.3.1	Distance from the basket . . . . .	13
3.3.2	Court zone . . . . .	13
3.4	Player Information . . . . .	13
3.4.1	Distance traveled . . . . .	13
3.4.2	Shooter velocity at time of shot . . . . .	14
3.4.3	Average velocity of shooter . . . . .	15
3.4.4	Changes in acceleration . . . . .	15
3.5	Game Information . . . . .	15
3.5.1	Game clock . . . . .	15
3.5.2	Shot clock . . . . .	15
3.5.3	Quarter . . . . .	16
3.6	Other Variables . . . . .	16
3.6.1	Number of Posesions . . . . .	16
3.6.2	Other Shot . . . . .	17
<b>4</b>	<b>Modeling</b>	<b>18</b>
4.1	Ridge Regression . . . . .	19
4.1.1	Description . . . . .	19
4.1.2	Performance . . . . .	20
4.2	Decision Tree . . . . .	21
4.2.1	Description . . . . .	22
4.2.2	Performance . . . . .	22
4.3	Random Forests . . . . .	23
4.3.1	Description . . . . .	23
4.3.2	Performance . . . . .	23
4.4	Support Vector Machine . . . . .	24

4.4.1	Description . . . . .	24
4.4.2	Performance . . . . .	25
<b>5</b>	<b>Conclusion and Future Work</b>	<b>26</b>
5.1	Data Issues and Missing Information . . . . .	26
5.2	Model Improvement . . . . .	28
<b>A</b>	<b>Variables Used in Model Fitting</b>	<b>30</b>
<b>B</b>	<b>Random Forests Model Performance</b>	<b>32</b>
	<b>Bibliography</b>	<b>33</b>

# List of Tables

4.1	Testing Data Set Breakdown . . . . .	19
4.2	$\lambda$ Values . . . . .	20
4.3	Ridge Regression Performance . . . . .	20
4.4	Decision Tree Performance . . . . .	22
4.5	Decision Tree Model Parameters . . . . .	23
4.6	Random Forests Model 1 Performance . . . . .	24
4.7	SVM Performance . . . . .	25
B.1	Random Forests Model 2 Performance . . . . .	32
B.2	Random Forests Model 3 Performance . . . . .	32



# List of Figures

2.1	Moments data . . . . .	5
2.2	ESPN Play by Play Data . . . . .	6
3.1	Angle between shooter and closest defender . . . . .	10
3.2	Court Zones . . . . .	14
4.1	Data Used for Modeling . . . . .	18
4.2	Lambda Parameters CV . . . . .	20
4.3	Ridge Regression Model Probabilities . . . . .	21

# List of Abbreviations and Symbols

## Abbreviations

ESPN	Entertainment and Sports Programming Networks
NBA	National Basketball Association

# Acknowledgements

I would like to extend a thank you to all of my committee members who assisted me with the project throughout the year. I would also like to thank Lorin Crawford for his help as well as the Center for Genomic and Computational Biology. Finally, I extend thanks to the Charlotte Hornets organization, particularly Alex Lee and David Kaplan.

# 1

## Introduction

Sports analytics has grown in popularity over the past couple decades. Sabermetrics, the analysis of baseball data using statistical methods entered pop culture with the publication of Bill James' book *Moneyball*. Since the rise of sabermetrics, statistical analysis has expanded to other sports, including basketball, hockey, tennis and more. Dean Oliver published *Basketball on Paper* in 2004 and is considered by many to be the father of basketball analytics.

Baseball analytics has developed more than its basketball counterpart since basketball is a more complicated game to analyze. The actions of a baseball game can be summarized by discrete events such as a thrown strike, a pop fly that is caught or a player running to first. Defensive players also tend to remain in their own section of the field. The outfielders typically do not head into the infield during games, and you rarely see the third baseman travel toward the first base side of the field. Offensive players, the baserunners, have established paths to follow between the bases. On the other hand, basketball players are only restricted by the boundaries of the court and do not spend the whole game in one area. Guards can cut toward the basket

on a driving lay-up and centers can test their outside shooting range from behind the three-point line. A player cutting through the key can cause defenders to shift and open up a teammate for a scoring opportunity. As a result, the development of basketball analytics has followed a different path than baseball requiring larger and more complicated data sets.

NBA teams employ statistical analysts in their front office in order to explore data for different purposes. Analysts use data to inform game strategy, evaluate players, research trades and more. Advances in technology over the past few years have provided more detailed data that has the ability to garner insights beyond what is possible with traditional box score statistics. Cameras in every NBA arena record tracking data for all games and provide information on the location of the players as well as the ball throughout a contest. This data can be used to extract information such as a player's average speed, the number of touches near the elbow or corner and more.<sup>1</sup>

Using the tracking data, it is possible to recreate the movement of all the players and the ball throughout the game using the details of the data. There are many articles and websites describing projects done with the tracking data. However, teams do not publish the research and projects using tracking data given the competitive nature of the league. The proprietary nature and cost of the tracking data also prevents the development of research.

This thesis explores the potential value of the NBA player tracking data. Information on player and ball location allows one to consider details such as the distances a player travels, the speed at which they are traveling, acceleration, and relative lo-

---

<sup>1</sup> NBA (2017)

cation to other players on the court. The project focuses on examining the time before a player takes a shot during a game, and exploring factors that could be relevant in predicting whether the player is going to make a shot. These models have the potential to identify the key factors that affect the probability of a shot going in. Successful models could be used to inform defensive or offensive strategy. For example, if the model indicates a player is more likely to make a shot when they are on the right side of the basket, an opposing team should force the player toward the left side of the basket. Offensive teams could exploit the characteristics that the model determines to be important.

These models could also evaluate the diversity of a player's skill set and value their skills. A model classifying whether a shot will be made determines the variables that are important in predicting shot outcomes. If a player is skilled in more of the important variables, they are more valuable than a player who does not possess as many of the important characteristics. Such information could inform the evaluations of trades and comparing the relative values of players.

# 2

## The Data

Every NBA arena has had the technology to collect player tracking data since the 2013-2014 season, and the technology has even been implemented in some arenas at the collegiate level, including Cameron Indoor Stadium.<sup>1</sup> The camera system used to collect the data, known as SportVU, is owned by STATS, LLC. There are six cameras in the rafters of the teams' arenas that are used to record information throughout a game. The cameras take 25 images per second (every 0.04 seconds). A computer then plots the locations of the ball and the ten players throughout the game. The players do not wear a tracking device, instead the location of the player is measured by the location of their torso.<sup>2</sup>

All NBA teams have access to the SportVU tracking data as well as reports that are generated from the data. The NBA team data contains information on the times of events such as shots, turnovers, rebounds and more. The tracking data used by the teams is proprietary and thus was unavailable for this project. Therefore a less

---

<sup>1</sup> Cohen (2013)

<sup>2</sup> Partnow (2015)

detailed version known as the moments data was used. Until around January 2016, it was possible to scrape the moments data from the NBA statistics website. The data used for this project is the moments data for over four hundred games from the 2015-2016 NBA regular season.<sup>3</sup> A moment in the data is defined by a set of eleven observations taken from the same image. A location for each player as well as the ball at a certain time in the game composes a moment.

team_id	player_id	x_loc	y_loc	radius	moment	game_clock	shot_clock	quarter	player_name	player_jersey
-1	-1	45.85088	8.21207	2.48915	700	48778	20.64	1	ball	NA
1610612737	2594	67.84156	5.49996	0.00000	700	48778	20.64	1	Kyle Korver	26
1610612737	200794	66.62832	31.49737	0.00000	700	48778	20.64	1	Paul Millsap	4
1610612737	201143	15.36756	23.05129	0.00000	700	48778	20.64	1	Al Horford	15
1610612737	201952	43.59962	7.68843	0.00000	700	48778	20.64	1	Jeff Teague	0
1610612737	203145	27.60430	47.34347	0.00000	700	48778	20.64	1	Kent Bazemore	24
1610612765	101141	72.10339	28.20127	0.00000	700	48778	20.64	1	Ersan Ilyasova	23
1610612765	202704	62.87406	15.28413	0.00000	700	48778	20.64	1	Reggie Jackson	1
1610612765	202694	57.40299	28.45878	0.00000	700	48778	20.64	1	Marcus Morris	13
1610612765	203484	71.72965	8.52779	0.00000	700	48778	20.64	1	Kentavious Caldwell-Pope	5
1610612765	203083	52.87533	24.58322	0.00000	700	48778	20.64	1	Andre Drummond	0

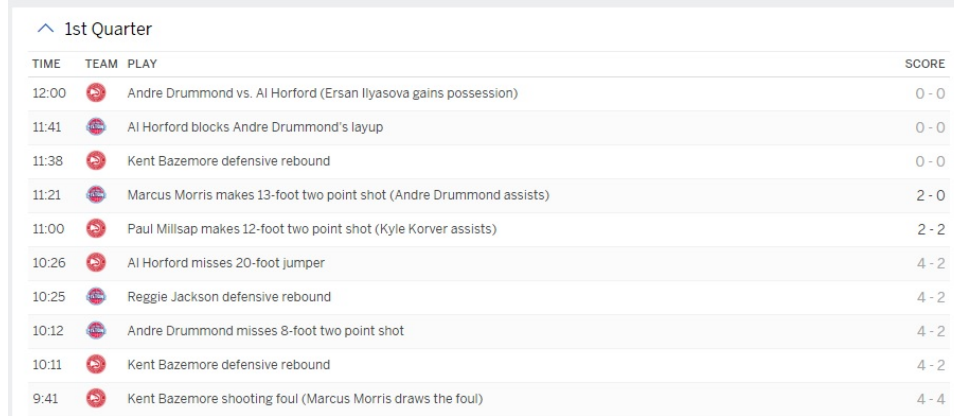
FIGURE 2.1: A single moment in the data consists of eleven rows with player information and coordinate locations.

The moments data does not contain the times or information of events such as a shot being taken. Thus it was necessary to supplement the moments data using an additional data source. ESPN provides play by play accounts of all NBA games on their website. The site provides the time of game events including free throws, turnovers, shots and timeouts. The play by play data was scraped from the ESPN data in order to supplement the moments data. Shot times were identified using the ESPN data by searching the strings of play descriptions. These times were used to filter the moments data to ten second sections of the games prior to a shot being taken.

The biggest challenge of the project was the investigation and organization of the data. More than 90% of the time spent on the project was dedicated to data clean-

<sup>3</sup> neilmj (2016)















1st Quarter			
TIME	TEAM	PLAY	SCORE
12:00		Andre Drummond vs. Al Horford (Ersan Ilyasova gains possession)	0 - 0
11:41		Al Horford blocks Andre Drummond's layup	0 - 0
11:38		Kent Bazemore defensive rebound	0 - 0
11:21		Marcus Morris makes 13-foot two point shot (Andre Drummond assists)	2 - 0
11:00		Paul Millsap makes 12-foot two point shot (Kyle Korver assists)	2 - 2
10:26		Al Horford misses 20-foot jumper	4 - 2
10:25		Reggie Jackson defensive rebound	4 - 2
10:12		Andre Drummond misses 8-foot two point shot	4 - 2
10:11		Kent Bazemore defensive rebound	4 - 2
9:41		Kent Bazemore shooting foul (Marcus Morris draws the foul)	4 - 4

FIGURE 2.2: The ESPN website provides a play by play account for all NBA games, <http://www.espn.com/nba/playbyplay?gameId=400827888>

ing, data management, and the feature generation described in the next chapter. A comprehensive codebook does not exist for the moments data since it is an unofficial version of the data. There are some sites <sup>4</sup> that provide information about the names of variables, but no official documentation exists.

The moments data was littered with inconsistencies and issues that made the data processing challenging. To begin with, the SportVU cameras did not always stop recording during a time out or a stoppage in play. The cameras capture 250 images in ten seconds, and each image is comprised of eleven rows, ten players and the ball. Thus there should be 2,750 observations for a given shot. Each value of game clock should appear only eleven times in this subset since the time on the game clock decreases between subsequent images. However, if there was a stoppage of play during those ten seconds, the cameras kept recording and there would be more than eleven observations for each value of game clock. Shots that had more than eleven observations of a particular game clock value were removed from the data. The project only used shots that had an uninterrupted ten seconds of game

<sup>4</sup> Tjortoglou (2015)

play prior to the shot. Therefore free throws were removed from the data as well as shots that occur less than ten seconds after a timeout or turnover.

In addition to stoppages in play creating issues with the data, some of the moments did not have eleven observations. Spot checking some of these individual moments revealed that they tended to be missing the location of the ball or had multiple locations for the ball. Shots with these erroneous moments were discarded since it was not possible to match the missing ball locations with the appropriate moments given the magnitude of the data and the timeline of the project.

There also were discrepancies between the shot times from the ESPN play by play data and the locations of the players and ball in the moment data. For example, there were times in the data where a player was supposedly shooting the ball, but the distance between the player and the ball was more than thirty feet. This is obviously a physical impossibility and reflects some discrepancies between the ESPN and the moments data. The shots were further reduced to only those where the shooter was less than six and a half feet from the ball at the time that they were supposedly shooting.

The data management stage also required an infrastructure to handle the size of the data. A portion of the data management was also dedicated to learning how to work with the server infrastructure using tools such as a Linux operating environment, slurm, and vim.

# 3

## Feature Generation

Models were constructed to predict the outcomes of shots, but each shot was comprised of 2,750 rows of data. Therefore it was necessary to reduce the data for each shot to a single row with relevant features to use in the model-fitting process. A major part of the data processing stage was feature generation. This chapter examines the calculations of these features and the aspects of an NBA game they have the potential to shed light on. There are many factors that affect whether a shot will be made by an NBA player. The player may be tired, they could have multiple defenders in their face, or they might be far away from the basket with the shot clock winding down. The tracking data was used to generate features to attempt to capture some of these characteristics. The feature generation required an application of basketball knowledge to the data. The variables described in this chapter as well as the player location were used to fit models.

### 3.1 Defensive Information

It is obvious that defense plays a huge role in whether a player will sink a shot. It requires more skill to hit a jumper over two defenders than to hit the same shot

without any defenders. Defenders can apply pressure or step back if their player is not a strong shooter. Defenders can also influence the direction an offensive player can travel. This section describes the features related to the defenders that were generated from the moments data in order to capture such information.

### *3.1.1 Distance to the nearest defender*

The distance to the nearest defender is the Euclidean distance between the shooter and closest defender to the shooter at the time of the shot. Many different elements of defensive strategy will influence this variable. Players with particular skills are guarded in different ways. For example, a team wants to prevent a player like Steph Curry from shooting a lot of three pointers. A coach might instruct their player to play very close to Curry to prevent him from shooting. On the other hand, a coach could tell a defender to play a few steps off of a player that is known for their ability to drive to the basket and finish near the rim. The distance between the defender and the shooter could also provide evidence of whether the defense was out of position. In a fast break situation, the nearest defender could be at the other end of the court and more than twenty-five feet from the shooter.

### *3.1.2 Angle between shooter and closest defender*

This feature is the acute angle made between two vectors measured in radians. Figure 3.1 demonstrates the angle referred to by this variable. One vector is defined by the coordinates of the basket and the shooter, and the other vector is defined by the coordinates of the basket and the defender. The angle between the shooter and closest defender is then calculated using the relationship  $\cos(\theta) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|}$ .

Positive values of the angle indicate that the defender is between the player

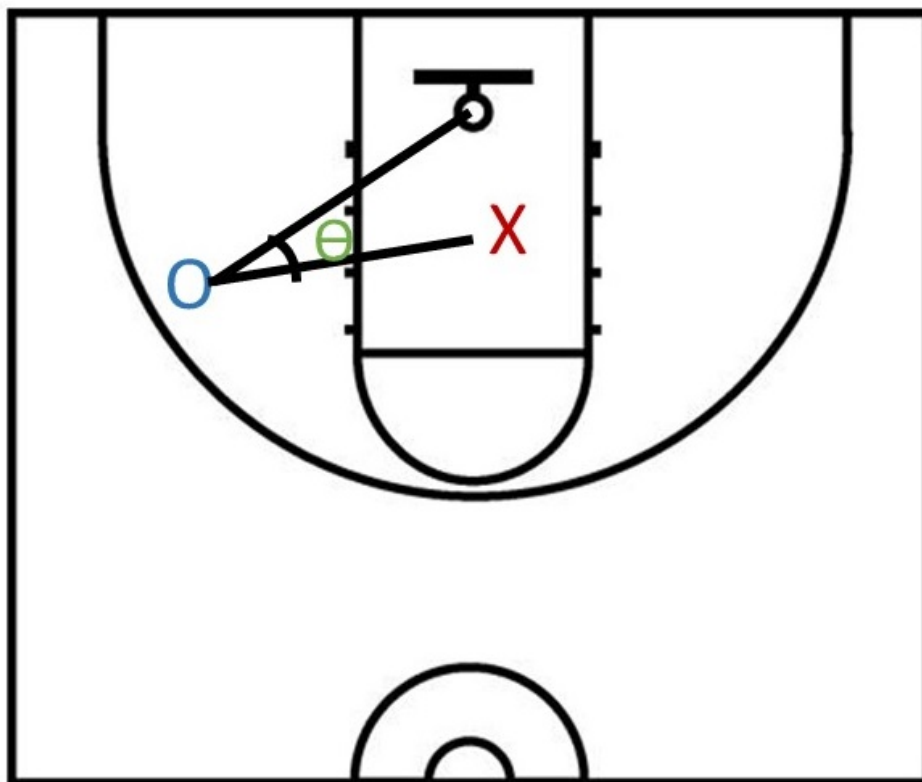


FIGURE 3.1: Angle between shooter and closest defender. Court Image from Winnetka Bullets.

and the basket and negative values indicate that the closest defender is behind the shooter. Defenders being behind the shooter could be indicative of a fast break or a situation when the defense is out of position. This measure provides more information about the position of the defender relative to the shooter.

### 3.1.3 Side of defender

The side of the defender refers to what side of the shooter the defender is on. The left side means that the y-coordinate<sup>1</sup> of the defender is more than three feet to the left of the y-coordinate of the shooter when the shooter is facing the baseline.

<sup>1</sup> The y-coordinate measures the position of a player relative to the baseline. The center of the court is at  $y=25$

The defender being on the right side means that the defender's y-coordinate is more than three feet to the right than the y-coordinate of the shooter. If a defender's y-coordinate is within three feet of the shooter's y-coordinate, they are defined to be in front of the shooter. This co-variate provides additional evidence of whether the defender is attempting to force the shooter to travel in a particular direction. The side of the defender together with the angle might be indicative of whether a defender is trying to force a shooter to a certain location of the court. There could be help defense in that direction or the defense forces the shooter to their weaker side. The defense could also be preventing the shooter from being able to move to an area of the court they shoot well from.

#### *3.1.4 Number of close defenders*

The number of close defenders is calculated as the number of defenders that are within five feet of the shooter at the time of the shot. Defenders being close to a shooter can prevent them from moving closer to the basket or to a space on the court where they are better at shooting from. Defenders also double or triple team more skilled players to prevent them from scoring. The difficulty of making a shot increases as the number of defenders increases.

### 3.2 Teammate Information

The features representing the teammate information are the same as those listed in the defender section above. The calculations are the same except for the closest defender is replaced with the closest teammate. This section does not redefine the calculations, but instead reviews the utility of the features.

### *3.2.1 Distance to nearest teammate*

Even though a teammate would not intentionally block a shot like a defender would, the distance to the closest teammate provides information about the spread of players on the floor. If a teammate is very close to the shooter, it is possible that their defender is also close to the shooter and influencing the shooter's behavior. This could crowd the area near the shooter and make it more difficult to make a shot or move to a different area of the court.

### *3.2.2 Angle between shooter and closest teammate*

As with the angle between the shooter and the closest defender, this feature provides information on the location of other players relative to the shooter. These can impact a shooter's performance if they are being crowded and prevented from traveling to a certain area of the court. The angle to the closest teammate provides information about the game situation at the time of the shot. The teammate being in front of the shooter could be indicative of a set play whereas the teammate behind might be a fast break.

### *3.2.3 Side of teammate*

When teammates are close to one another, a defender can more easily guard both players at the same time. Thus the shooter may not be able to move in the direction of their nearest teammate. The shooter might want to prevent overcrowding by avoiding moving near their teammates and take a shot if their teammate is preventing them from moving to a more preferred location.

#### *3.2.4 Number of close teammates*

Teammates being close to a shooter could result in more defenders being close to the shooter. It could also mean that a teammate was setting a screen or blocking defenders from reaching the shooter. All of these factors and situations influence whether a player will make a shot.

### **3.3 Location**

#### *3.3.1 Distance from the basket*

The Euclidean distance between the shooter and the basket at the time of shot is calculated for each shot. It is much easier and requires less skill for a player to make a lay-up than a three-pointer near half court. The relationship between the probability of making a shot and the distance from the basket is not as simple in a game situation. Being closer to the basket means that the defenders tend to be taller and have longer arms. Therefore it is necessary to include information about the defenders together with the distance measure.

#### *3.3.2 Court zone*

Three different court zones were defined in order to reduce the information contained in the location of the shooter. The three court zones are outside the three point line, close to the basket, and in between. These can be seen in Figure 3.2 below.

### **3.4 Player Information**

#### *3.4.1 Distance traveled*

The distance traveled by the shooter between consecutive images was calculated in the moments data using the Euclidean distance formula. The total distance traveled



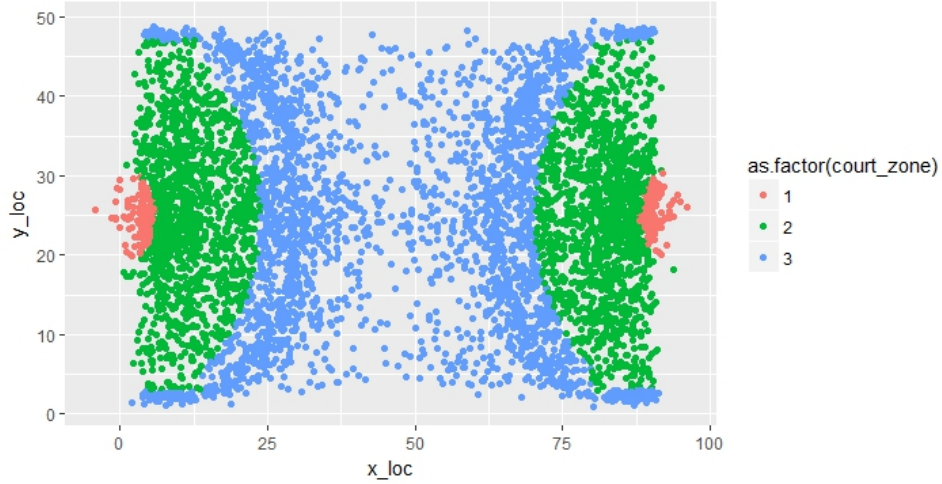


FIGURE 3.2: The court zones.

by the shooter prior to the shot is the sum of all of these distances. If a player travels a large distance prior to a shot they could be more tired and it will be more difficult to make the shot. There could be a series of fast breaks leading up to the shot that caused the player to run a lot. On the other hand, the shot attempt could be the result of a set play where the shooter does not travel much before shooting and their total distance traveled will be smaller.

#### 3.4.2 Shooter velocity at time of shot

The velocity of the shooter at each moment was found by dividing the distance traveled by 0.04 seconds. This variable is the velocity of the shooter at the time that the shot is taken. This provides insight into the movement of the player at the time of the shot. There are different skills required to hit a jumper when stationary and the same shot when traveling at a high velocity. A shooter traveling very quickly could indicate a lay-up in a fast break situation.

### *3.4.3 Average velocity of shooter*

The velocity of the shooters over the course of the ten seconds was averaged and included in the models. If the shooter has a higher average velocity, it means that they are traveling at a higher speed. It could be more challenging for the shooter to hit a shot if they are tired from sprinting prior to their shot.

### *3.4.4 Changes in acceleration*

The number of changes in acceleration is calculated by counting the number of times that a player's acceleration has a magnitude of at least ten in the ten seconds prior to the shot. The acceleration at every point during the period before the shot was calculated by dividing the velocity of the shooter by 0.04. This variable aims to capture information about how a player might be using hesitations or cuts in their movement.

## **3.5 Game Information**

### *3.5.1 Game clock*

The moments data includes the time on the game clock for each image. An NBA quarter lasts for twelve minutes. Therefore, the game clock variable ranges from 0 to 720 seconds. The variable counts down from twelve minutes, so a larger value of game clock means that there is more time remaining in the quarter. Game clock is important because it can influence when a shooter decides to take a shot. If there are only a few seconds left in the quarter, a player will be more likely to take a shot that they would not normally attempt given the time limitation.

### *3.5.2 Shot clock*

The moments data also includes the value on the shot clock. During the 2015-2016 season, the shot clock lasted 24 seconds. The shot clock plays an important role in

a player’s decision to shoot the ball. If a team has not attempted a shot that hits the rim of the basket before the shot clock expires, a turnover occurs and the ball changes possession to the other team. A player must pay attention to the time on the shot clock throughout the possession. If a player shoots with a small amount of time on the shot clock, they may have been forced to shoot a poor shot in order to prevent a turnover.

### *3.5.3 Quarter*

The quarter of the game is also included in the moments data and was used in the model. The quarter provides a measurement of the overall time that has passed in the game. An athlete will not be as tired in the first quarter as they are in the fourth quarter. Fatigue plays a big role in whether a shot is made or not. It is more difficult to hit shots later in the game, especially if the player has been playing a lot.

## 3.6 Other Variables

### *3.6.1 Number of Possessions*

The possession of the ball at each image is assigned to the player who is closest to the ball. If a player is the closest player for at least two seconds, they are determined to have held the ball long enough to count for a possession. This variable is the number of changes in possession that occur in the ten seconds of interest. The number of possessions would be informative of whether there was many passes and rapid ball movement or if the shooter held the ball for the ten seconds before the shot.

### *3.6.2 Other Shot*

This variable indicates whether there is another shot occurring in the ten seconds before the shot. If shots occur within ten seconds of one another, information will overlap between the features for the two shots. This could also provide information if a shot is put back off of an offensive rebound.

# 4

## Modeling

This project explores the binary classification problem of predicting the outcome of a shot. The models were fit using a training set and a test set was used for evaluation. The modeling data set consisted of 3,284 made shots and 2,322 missed shots. The training set consisted of sixty five percent of the data. There are twenty-three co-variates used as predictors in the data and these are listed in Appendix A.

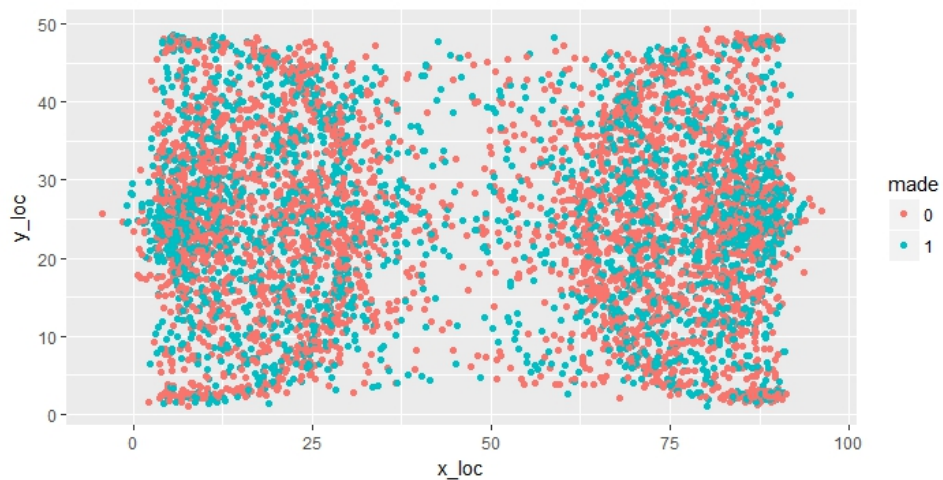


FIGURE 4.1: The modeling data set color coded by made and missed shots.

The models were evaluated on a test set that comprises 45% of the modeling data.

Table 4.1: In the test data set, 41% of the shots were made and 59% were missed.

Shot Outcome	Count
Made	2322
Missed	3284

## 4.1 Ridge Regression

### 4.1.1 Description

Penalized logistic regression was the first modeling approach used. This approach explores the feasibility of linear methods. The ridge regression model takes the form:

$$\hat{y} = \beta_0 + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \dots + \beta_p \mathbf{X}_p + \epsilon$$

Ridge regression is fit by choosing the  $\beta_i$ ,  $0 < i < p$  that minimizes the residual sum of squares defined by <sup>1</sup>

$$\sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p \beta_j^2$$

.

The  $\lambda$  parameter for the model was chosen using ten-fold cross-validation. The value that minimized the deviance was selected as the  $\lambda$  to use. Cross validation was used five different times, and the resulting values of  $\lambda$  are displayed in the Table 4.2.

These values are all close or the same in value. The  $\lambda$  value selected to use in the model is  $\lambda = 0.01274487$ .

---

<sup>1</sup> (Gareth James and Tibshirani, 2013, pg. 215)

Table 4.2: Cross-validation was used five times in order to explore different values of  $\lambda$ .

CV Fit	$\lambda$
1	0.01398748
2	0.01274487
3	0.01274487
4	0.01398748
5	0.01274487

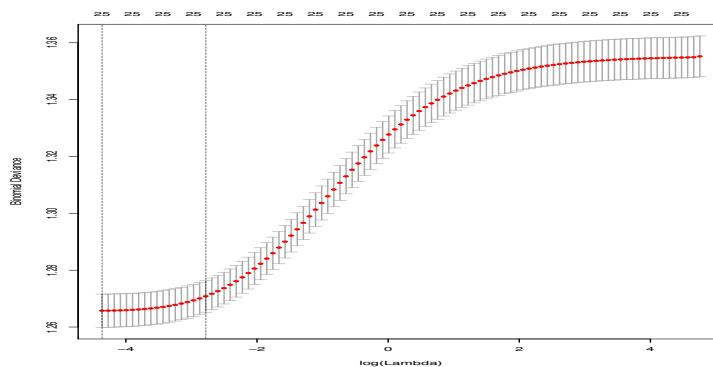


FIGURE 4.2: The deviance calculated for various values of  $\lambda$  in one instance of cross-validation.

#### 4.1.2 Performance

The predictions from the ridge regression did not perform better than random guessing. Figure 4.3 displays the confusion matrix for this matrix.

Table 4.3: The ridge regression did not perform better than random guessing.

	Truth	
	Missed	Made
Missed	862	276
Made	468	357

The ridge regression model correctly predicted more missed shots than made shots. Figure 4.2 below displays the probabilities predicted for whether each shot is

made. Probabilities greater than 0.5 were classified as made baskets and less than 0.5 were missed baskets. The distribution of probabilities has a mode around 0.5. The largest predicted probability is just over 0.7. The model does not predict that many shots have a high probability of being made.

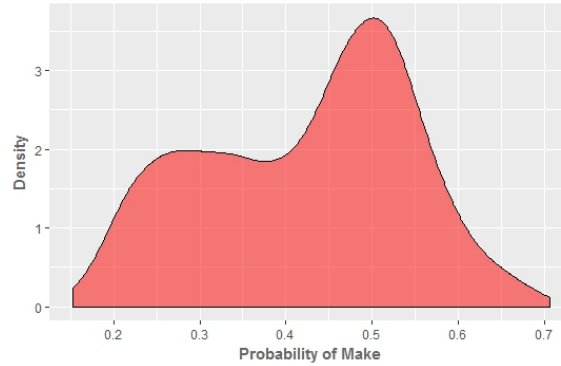


FIGURE 4.3: The predicted probabilities of from the test set.

## 4.2 Decision Tree

Due to the poor performance of the penalized logistic regression, fitting a linear model did not seem appropriate. In addition to the penalized logistic regression, a decision tree model was fit to the training data. It was fit since a decision tree is a machine learning technique that is easy to interpret and also has a nice graphical representation.<sup>2</sup>

---

<sup>2</sup> (Gareth James and Tibshirani, 2013, pg. 315)



#### 4.2.1 Description

The decision tree model was fit using the **sklearn** package in Python.<sup>3</sup> The Gini index was used as the measure of node purity and is defined as<sup>4</sup>

$$G = \sum_{k=1}^K \hat{p}_{mk}(1 - \hat{p}_{mk}).$$

Classification error is another method to measure node purity, but this measure is not sensitive enough for the creation of a decision tree.<sup>5</sup> At the time of this paper, the **sklearn** package does not support pruning, so there is a risk that the model overfits.

#### 4.2.2 Performance

Like the ridge regression, the decision tree model also did not perform well when used to make predictions from the test set.

Table 4.4: The decision tree model did not perform better than random guessing.

	Truth	
	Missed	Made
Missed	707	431
Made	429	396

The decision tree model correctly classified only 56% of the data and predicted that more shots were missed than made.

---

<sup>3</sup> scikitlearn (2017)

<sup>4</sup> (Gareth James and Tibshirani, 2013, pg. 312)

<sup>5</sup> (Gareth James and Tibshirani, 2013, pg. 312)

### 4.3 Random Forests

A random forests model was also fit in order to determine if a combination of decision trees performed better than the ridge regression or the decision tree. Random forests are a complex tool<sup>6</sup> and thus might be able to capture the complicated relationships that exist in the data. Random forests are another nonlinear method.

#### 4.3.1 Description

The Python package **sklearn** was also used to fit the random forests model. The number of features to consider at each split and the number of trees were chosen using cross-validation. Different initializations of cross-validation resulted in different values for both of these parameters. This is not surprising as it can be difficult to tune the parameters in random forests.<sup>7</sup> Thus, multiple random forests models were fit. The parameters of the random forests models fit are defined in Table 4.5.

Table 4.5: Three random forests were fit using the parameters in the table

Model	Max. Features	Number of Estimators
1	6	150
2	4	200
3	14	200

#### 4.3.2 Performance

The three random forests also did not perform well. The results for model 1 are displayed below and the results for the remaining models can be found in Appendix B.

---

<sup>6</sup> Rudin (2016a)

<sup>7</sup> Rudin (2016a)

Table 4.6: The random forests model did not predict better than the other models.

	Truth	
	Missed	Made
Missed	837	301
Made	449	376

## 4.4 Support Vector Machine

The support vector machine is another classification method that can produce non-linear boundaries. Support vector machines can also be generalized to instances where the data is non-separable. Thus a support vector machine was also fit.

### 4.4.1 Description

The support vector machine boundary is defined by the function <sup>8</sup>

$$f(x) = \sum_{j=1}^p \lambda_j x_j + \lambda_0$$

It can be shown that calculating the support vector machine is the optimization problem

$$\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,k=1}^n \alpha_i \alpha_k y_i y_k \mathbf{x}_i^T \mathbf{x}_k$$

where  $0 \leq \alpha_i \leq C$ ,  $i = 1, \dots, n$  and  $\sum_{i=1}^n \alpha_i y_i = 0$ . The inner product,  $\mathbf{x}_i^T \mathbf{x}_k$  can be replaced with a kernel. These details and derivations can be found in various sources.<sup>9</sup>. A linear kernel did not seem appropriate given the poor performance of the linear method, so the radial basis kernel,  $K(x, x') = \exp(-\gamma \|x - x'\|^2)$  <sup>10</sup> was used.

---

<sup>8</sup> (Rudin, 2016c, pg. 2)

<sup>9</sup> (Rudin, 2016b, pg. 2)

<sup>10</sup> Trevor Hastie and Friedman (2009)

#### 4.4.2 Performance

The support vector machine did not beat the predictive performance of the other models fit.

Table 4.7: The support vector machine did not perform better than the other models

	Missed	Make
Missed	836	202
Made	465	360

The support vector machine predicted a miss about 70% of the time and as a result correctly predicts more misses than makes.

Overall, all of the models fit throughout the course of the project did not perform well. The models all tend to predict more missed shots than made shots. The models do not appropriately capture the relationships that exist between the features and whether a shot is made. The final chapter explores some issues and potential next steps to improve upon the results.

## Conclusion and Future Work

The different models fit throughout the course of the project were not successful. None of the models performed substantially better than random guessing. There are numerous possibilities to explain the poor performance of the models and some are explored below.

### 5.1 Data Issues and Missing Information

Basketball is a complicated sport that involves many different players and the ball constantly moving throughout the game. The movement of the ball and the players are not independent of one another, and the trajectories of the players and the ball influence one another. These relationships affect the decisions that players make when they have the ball and are considering taking a shot. The features generated from the data account for the relative positions of players by including information such as distance to the nearest defender and distance to the closest teammate at the time of the attempted basket. However, these variables do not account for all the variation introduced by the shooter's relationship with all of the players on the

court. The features currently only account for the closest teammates and defenders, but the location of other players on the court have the ability to influence a shot as well.

In addition to the influence of the other players' and ball's position, many other factors influence whether a shot goes in that are not captured in the data. One of the most debated issues in basketball is the existence of the "hot hand". This theory states that if a shooter is "hot", hitting a bunch of shots, they are more likely to keep scoring. As with all sports, psychology can play a huge role in the performance of a player. If a player has missed a lot of shots, committed a turnover or made a poor defensive decision they might not have their best mental focus. A boisterous crowd could also affect a player's ability to score. It is impossible for the tracking data to account for this.

The data used to fit the models contain some issues that need to be addressed. To begin with, shooters appear multiple times in the data set as NBA players shoot more than once during the season. Thus the observations are related in that shooters are the same players. In addition, if there are shots that are less than ten seconds apart their summary features share some information. Finally, some of the variables used in the model fitting process are related and result in collinearity issues. Due to the amount of time establishing the infrastructure to deal with the data, these are issues that need to be addressed in future work.

The data issues presented in this paper resulted in many observations being dropped from the data set. As a result there were not many players with enough observations to fit a model using their shot data alone. The model fit for one player was not successful and had results similar to the models presented in Chapter 4, and

thus was not included in the paper. At the moment it is not possible to compare players based on their models with the current data.

## 5.2 Model Improvement

The project of predicting shot outcomes using the moments data is a very challenging problem. The models presented in this paper were not successful and did not have the information necessary to provide insight into the problem. The moments data captures the change in movement for players over time. It is necessary to incorporate the spatial and temporal aspects of the data. At the moment, the models do not account for the movement of all the players and the ball in the time before the shot is taken. These could be improved by identifying additional variables to include in the model. For example, determining the type of defense being played by the opposing team would be very helpful in predicting shot outcome. It can be easier to shoot threes against a zone defense, a team might execute a box and one in order to apply more defensive pressure to a skilled player. This information influences shot outcomes and is not accounted for in the features currently generated from the data. In addition, a better method to determine the player possessing the ball and to detect passes could provide more accurate information to feed into the models. The next step is to explore spatio-temporal models that can be applied to this problem and to use spatial aspects to extract better predictors from the data.

Overall, the project exposed the many challenges of working with the moments data. The majority of the project time was spent learning the infrastructure necessary to handle the data and exploring the inconsistencies between the tracking data and the ESPN data. The project could be improved by using the SportVU tracking data that is available to teams. This data contains information on game events such

as shots being taken, turnovers, rebounds and more. The time of the events will be more reliable as the times are coming from the SportVU camera tracking system. Even though the project did not produce successful models, it demonstrates the amount of information available in the NBA tracking data and its potential utility.



# Appendix A

## Variables Used in Model Fitting

- `accel.changes` - Number of times the shooter's acceleration changes by a magnitude of at least  $10ft/s^2$
- `closest.def.loc` `closest.teammate.loc` - Side of the shooter that the closest defender (teammate) is defined as left, right, or front
- `closest.defender.angle` , `closest.teammate.angle` - Angle between the vectors defined by the location of the basket and the shooter, and the location of the shooter and the defender
- `closest.defender.dist`, `closest.teammate.dist` - Euclidean distance between the shooter and the closest defender (teammate)
- `court.zone` - One of three locations on the court, defined as near the basket or inside/outside the three-point line
- `game.clock` - Time remaining on the game clock

- num.close.def, num.close.team - Number of defenders (teammates) within five feet of the shooter at the time of the shot
- num.poss - Number of times that the ball changes possession
- other.shot - Indicator of whether or not another shot was taken in the ten seconds before the shot
- quarter - Quarter of the game
- shooter.avg.vel - Average velocity of the shooter in ten seconds before the shot
- shooter.dist.traveled - Distance traveled by the shooter in the ten seconds before the shot
- shooter.vel.at.shot - Velocity of the shooter at the time of the shot
- shot.clock - Time on the shot clock at the time of the shot
- shot.val - Whether the shot is worth 2 or 3 points
- x.loc - x-coordinate of the shooter
- y.loc - y-coordinate of the shooter

# Appendix B

## Random Forests Model Performance

Table B.1: Model 2 Confusion Matrix

	Truth	
	Missed	Made
Missed	894	244
Made	496	329

Table B.2: Model 3 Confusion Matrix.

	Truth	
	Missed	Made
Missed	821	317
Made	445	380

# Bibliography

- Cohen, B. (2013), “Player Tracking Technology Goes to Duke,” <https://www.wsj.com/articles/SB10001424052702304520704579127281774259324>.
- Gareth James, Daniela Witten, T. H. and Tibshirani, R. (2013), *An Introduction to Statistical Learning with Applications with R*, Springer.
- NBA (2017)[stats.nba.com](http://stats.nba.com).
- neilmj (2016), “Basketball Data,” <https://github.com/neilmj/BasketballData/tree/master/2016.NBA.Raw.SportVU.Game.Logs>.
- Partnow, S. (2015), “Nylon Calculus 101: Intro to SportVU,” <http://nyloncalculus.com/2015/08/13/nylon-calculus-101-intro-to-sportvu/>.
- Rudin, C. (2016a), *Decision Forests*, Duke University STA 561 Course Notes.
- Rudin, C. (2016b), *Kernels*, Duke University STA 561 Course Notes.
- Rudin, C. (2016c), *Support Vector Machines*, Duke University STA 561 Course Notes.
- scikitlearn (2017), “sklearn.tree.DecisionTreeClassifier,” <http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.htm>.
- Tjortoglou, S. (2015), “How to Create NBA Shot Charts in Python,” <http://savvastjortjoglou.com/nba-shot-sharts.html>.
- Trevor Hastie, R. T. and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.