CIS 490: Big Data Analytics
Final Project
Sports Science

Matt Hixon

Kansas State University

December 14, 2016

Contents

List of Figures

1 Introduction

Live sports generate some of the most diverse, populous, and unpredictable datasets. New technologies like motion tracking and RFID have helped increase the capacity in which we can gather data, but this surge has not been matched with an ability to readily extract meaningful conclusions from the data. Sport science is an evolving field focusing on gaining a competitive advantage. Two major areas of research in sports science are statistical macroscopic trends and analyzing spatial data. Analyzing statistical macroscopic trends among players and teams is commonly used to predict player's value, identify a team weakness, or even predict the outcome of a game. Spatial data is used to classify in-game decisions and find movement tendencies that could be exploited. This short paper will be an exploration of the data generated by the National Basketball Association. It will focus on spatial aspects and leverage big data analysis techniques in order to look for interesting insights.

2 The Dataset

The dataset consists of over 128,000 rows detailing information about shots taken by NBA players in the 2014-2015 season. Each row in the dataset is representative of a shot and contains information about the game (Teams, Score, etc.), players (Closest Defender, Shot Taker, etc.), and spatial locality (Shot Distance, Defender Distance, etc.). The dataset was made public on Kaggle by Dans Becker who acquired the data by scraping the NBA's REST API.

3 Problem Statement

An integral part of sport science is enabling personnel with visualizations and insights so they can better accomplish their job. In this circumstance, we have 128,000 different shots taken by different players on different days. **Our goal is to see if there is any trend between the shot situation and the shots players successfully make.** Should there be a high correlation for a shot situation, coaches could gain an advantage by devising plays that put players in that situation. After all, the main part of a leader's job is to put people in a position to succeed.

4 The Method

The first step is to define a shot situation. On first glance, there are infinitely many shot situations that could occur. From a thundering break away dunk by Russell Westbrook to a contested 30 footer from Steph Curry, these seemingly opposite situations need to be represented so analysis can be performed. This situational representation is achieved by utilizing a partitioning system. A situation is required to consist of Shot Type, Shot Quality, Game Time, Shot Clock, Shot Distance, Dribbles, and Touch Time. This covers a large majority of factors strictly relating to any given spatial shot situation. For example, classifying a Russell Westbrook break away dunk might look something like the following:

Table 1: Westbrook's Dunk

Shot Type	Drive
Shot Quality	Wide Open
Game Time	End
Shot Clock	Very Early
Shot Distance	Dunk
Dribbles	4
Touch Time	4.2

One would think this combination of partitions, given to any player, would have a highly successful shot attempt. Thus, we have successfully created a reasonable partition of situational factors.

With this set of partitions, we look to find a combination that has a high field goal percentage. We also look to predict whether a given shot will be successful based on the situational partitions. Given the approach we are taking, we will use a random forest model for our prediction. The random forest model creates many decision trees, which use sets of factors to determine output, and calculates an importance metric for each partition. Once the importance metrics have been calculated, we can feed our model new partitioned situations and let it predict!

4.1 Feature Engineering

The different partitions were picked using my own knowledge of basketball. Here is an overview of the partitions and my reasoning behind their thresholds:

Table 2: Shot Type Partition

Shot Type	**Dribbles**	**Shot Distance**	**Point Type**
Catch and Shoot	<= 1	>4	2
Cut	<= 1	<= 4	2
Drive	>1	<= 4	2
ISO/Post Up	>4	any	2 or 3
Long ISO	>20	any	2 or 3
Spot Up Three	<= 1		3

Notes: A spot up 3 is more selective than a catch and shoot as it requires the shot attempt to be a 3 pointer. This is to help break up the large catch and shoot partition. Many shots were left unknown as it was difficult to determine exactly what type of shot occurred. These unknown shot types were treated as a feature and they expressed similar dribble and shot distance characteristics upon further analysis. This feature was posted by user StefanLangenbord on Kaggle.

Table 3: Shot Quality Partition

Shot Quality	**Defender Distance**
Tightly Contested	<= 2.0
Contested	2.0 <–>3.0
Open	3.0 <–>5.0
Wide Open	>= 5.0

Notes: These values were chosen because they divide the data into reasonable partitions and relate to the openness of a shot in real life. This feature was posted by user StefanLangenbord on Kaggle.

Table 4: Game Time Partition

Game Time	**Game Clock**
Beginning	>= 9:00
Early Middle	6:00 <–>9:00
Late Middle	3:00 <–>6:00
End	<= 3:00

Notes: This partition simply divides each period into 4 separate times

Table 5: Shot Clock Partition

Shot Clock	**Shot Clock**
Beginning	>= 24
Very Early	20 <–>24
Early	15 <–>20
Middle	10 <–>15
Late	5 <–>10
Very Late	2 <–>5
Forced	<= 2

Notes: The original data had many NA's for the shot clock column. Operating under the assumption that an NA was present when there was not enough time left in a period for a full shot clock, the NA shot clock columns were filled with the corresponding game clock time. The very late and forced partition were separated in hopes of differentiating a situation in which a player was forced to shoot to avoid a violation and a long possession that resulted in a quality shot. Further exploration helped validate this partition, as only 27% of shots were successful in the last two seconds of the shot clock.

Table 6: Shot Distance Partition

Shot Distance	**Shot Distance**
Prayer	>= 30
Deep 3	25 <–>30
3 Pointer	22 <–>25
Deep 2	15 <–>22
Mid Range	8 <–>15
Close	3 <–>8
Dunk	<= 3

Notes: These partitions are solely based off of shot distance and help frame the spatial whereabouts for the model. The prayer partition had significantly less shots than the others, but this will be effective as the shots in this partition have a very low execution rate.

4.2 Exploration and Visualizations

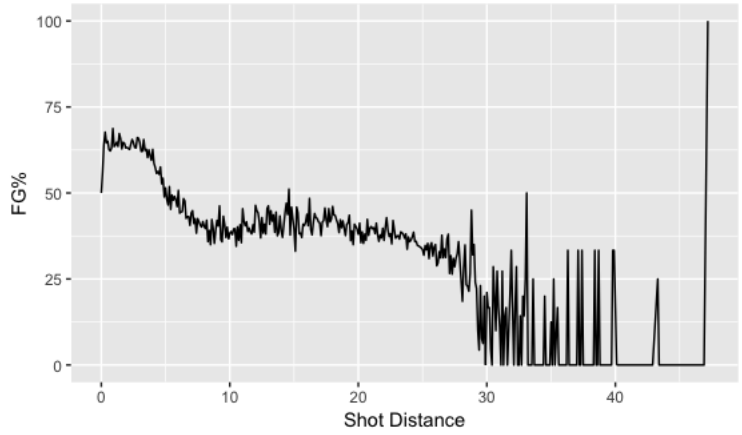

Figure 1: Shot Distance vs. FG Percentage

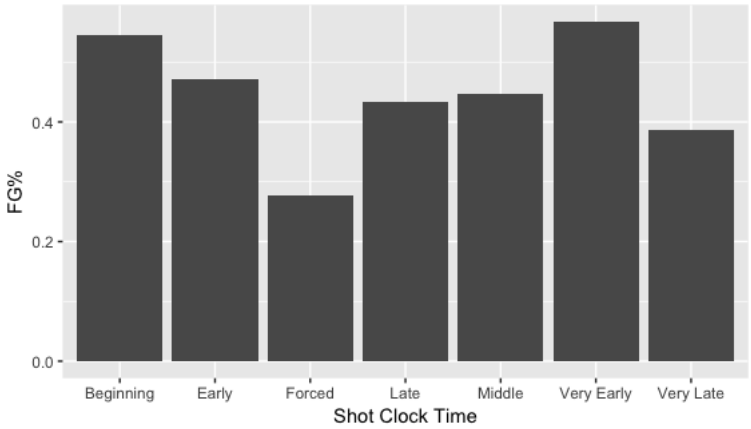

Figure 2: Shot Clock vs. FG Percentage

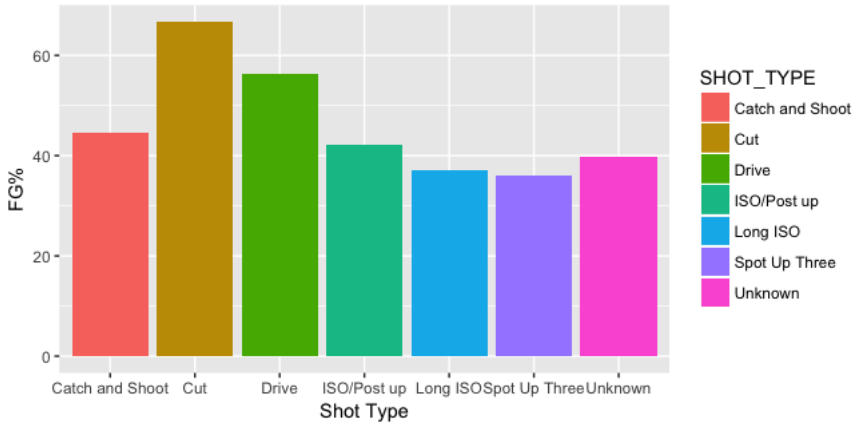

Figure 3: Shot Type vs. FG Percentage

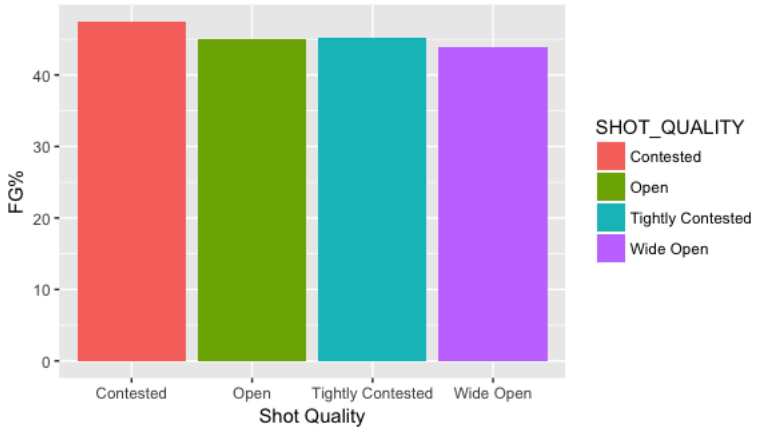

Figure 4: Shot Quality vs. FG Percentage

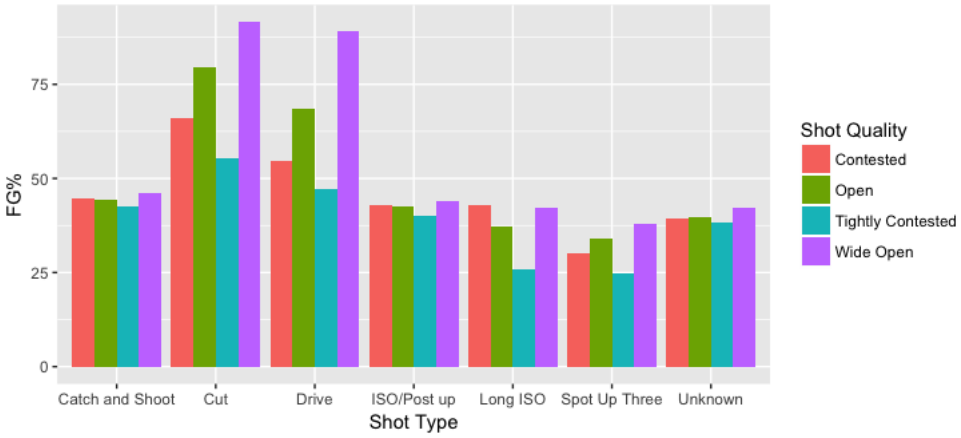

Figure 5: Shot Type vs. FG Percentage (with Shot Quality)

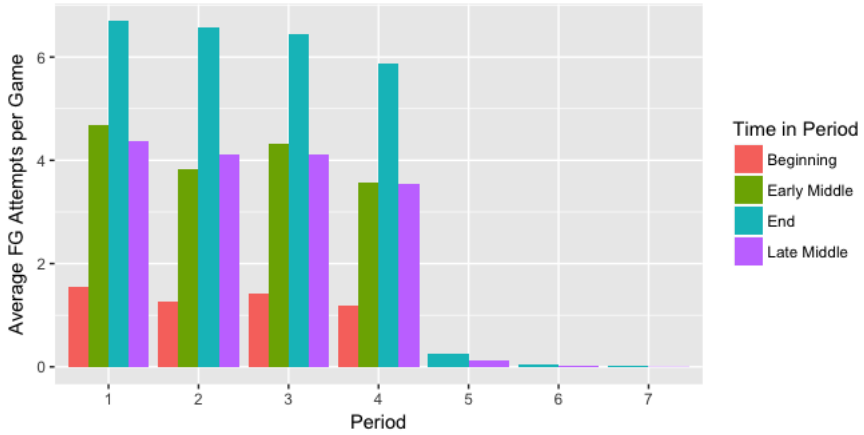

Figure 6: Period vs. Average FG Attempt (with Time in Period)

	SHOT_CLOCK_PARTITION	SHOT_DIST_PARTITION	SHOT_TYPE	SHOT_QUALITY	GAME_CLOCK_PARTITION	FGM.x	FGM.y	SORT_FACTOR
1599	Middle	Deep 2	Catch and Shoot	Wide Open	End	538	0.44026187	236.86088380
1486	Middle	3 Pointer	Spot Up Three	Wide Open	End	568	0.39887640	226.56179775
1709	Middle	Dunk	Cut	Open	End	241	0.84561404	203.79298246
362	Early	3 Pointer	Spot Up Three	Wide Open	End	497	0.40275527	200.16936791
1091	Late	3 Pointer	Spot Up Three	Wide Open	End	471	0.39983022	188.32003396
1598	Middle	Deep 2	Catch and Shoot	Wide Open	Early Middle	398	0.43027027	171.24756757
1591	Middle	Deep 2	Catch and Shoot	Open	End	386	0.42278204	163.19386637
580	Early	Dunk	Cut	Open	End	194	0.82553191	160.15319149
1705	Middle	Dunk	Cut	Contested	End	215	0.70491803	151.55737705
361	Early	3 Pointer	Spot Up Three	Wide Open	Early Middle	375	0.39851222	149.44208289
1485	Middle	3 Pointer	Spot Up Three	Wide Open	Early Middle	391	0.38109162	149.00682261
363	Early	3 Pointer	Spot Up Three	Wide Open	Late Middle	365	0.40376106	147.37278761
1600	Middle	Deep 2	Catch and Shoot	Wide Open	Late Middle	334	0.43832021	146.39895013
223	Beginning	Dunk	Cut	Tightly Contested	End	263	0.55252101	145.31302521
1713	Middle	Dunk	Cut	Tightly Contested	End	256	0.56263736	144.03516484
1213	Late	Deep 2	Catch and Shoot	Wide Open	End	334	0.42385787	141.56852792
1708	Middle	Dunk	Cut	Open	Early Middle	160	0.84656085	135.44973545
1487	Middle	3 Pointer	Spot Up Three	Wide Open	Late Middle	362	0.36938776	133.71836735
576	Early	Dunk	Cut	Contested	End	178	0.70634921	125.73015873
1329	Late	Dunk	Cut	Open	End	163	0.76525822	124.73708920
1478	Middle	3 Pointer	Spot Up Three	Open	End	312	0.39294710	122.59949622
1710	Middle	Dunk	Cut	Open	Late Middle	144	0.84705882	121.97647059
1205	Late	Deep 2	Catch and Shoot	Open	End	297	0.40408163	120.01224490
579	Early	Dunk	Cut	Open	Early Middle	145	0.82386364	119.46022727
1214	Late	Deep 2	Catch and Shoot	Wide Open	Late Middle	243	0.48502994	117.86227545
1092	Late	3 Pointer	Spot Up Three	Wide Open	Late Middle	291	0.39972527	116.32005495
1706	Middle	Dunk	Cut	Contested	Late Middle	158	0.73148148	115.57407407
1714	Middle	Dunk	Cut	Tightly Contested	Late Middle	196	0.58160237	113.99406528
222	Beginning	Dunk	Cut	Tightly Contested	Early Middle	205	0.55555556	113.88888889
215	Beginning	Dunk	Cut	Contested	End	177	0.62544170	110.70318021
224	Beginning	Dunk	Cut	Tightly Contested	Late Middle	208	0.53196931	110.64961637

Figure 7: Best Factors for Shots

4.3 Model

The dataset is split into two samples, one for training and the other for testing. The training dataset consists of 75% of the data and the test dataset contains the other 25%. Then the training data is assembled by appending the factors we would like to run through our random forest model. For the purpose of this experiment, the default random forest was run with 300 trees. It appears the model actually converges around 100 trees.

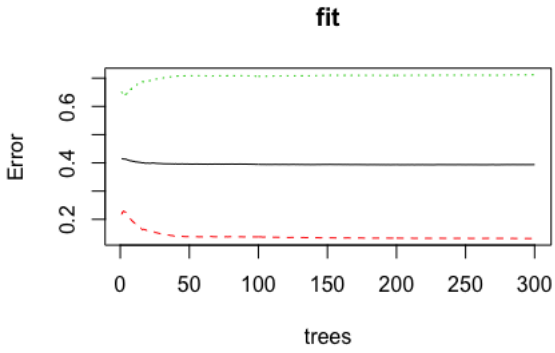

Figure 8: Number of Trees vs. Error

The model predicted on the test set with an accuracy of 60.83% which is better than the baseline prediction of around 54% (all misses).

```
        OOB estimate of  error rate: 39.38%
Confusion matrix:
      0     1 class.error
0 45743  6954    0.131962
1 30870 12484    0.712045
                Test set error rate: 39.17%
Confusion matrix:
      0    1 class.error
0 15197 2270   0.1299594
1 10273 4278   0.7059996
```

Figure 9: Confusion Matrix

The model found that shot type was the most prominent feature.

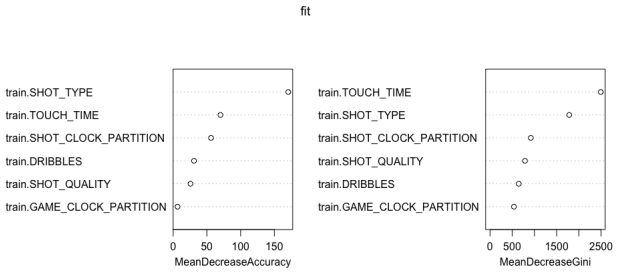

Figure 10: Importance Factor

5 Conclusion

The purpose of this paper was to find insightful information that players and coaches could use to their advantage. Although the predictive model was less than stellar at only 60.83% accuracy rate, the exploratory analysis could still be useful.

Why is the predictive model only 60.83%? It could be for a variety of factors, but my intuition suggests that while the factors we had seemed to prove useful, they could not be generalized for every player in the league. Since players in the NBA come in all shapes and sizes, a high quality shot for Anthony Morrow (average height) may not be a high quality shot for Steven Adams (very tall). This model failed to take into account any information about players and only considered spatial situations.

Future work with this project include appending physical information (height, weight, etc.) and previous statistics (FG%, PPG, etc.) about players to the dataset. The machine learning model could benefit as this will provide more details about the situation a shot was taken from. Another extension of this project would be the incorporation of play-by-play data. Again, this would further refine the specific situation a player experiences to include ideas such as game pace and momentum.

As the dataset grows in size and the computations become more intense, spark would be a great place to start processing data. Feature engineering and other data manipulation tasks could be completed using mapreduce. Additionally, a few other models should also be considered, such as naive bayes, as they may lead to a better situational classification.

This project combined my interest of sports and computer science in a unique way. I learned about R, feature engineering, random forest modeling, dealing with NAs, scaling issues, visualizations, and technical writing.