COLUMBIA UNIVERSITY
School of Professional Studies

NBA Shot Probability Analysis

Project Deliverable #1

October 22nd, 2017

Eric Pridgen, Fan Yang, Kun Yang, Wangyi Zhang, Zixin Jia

Table of Contents

1. PART ONE

1.1 Dataset Description

Dataset 1 Name: NBA Shot Logs
Observations: 128069 **Variables:** 21

This dataset includes 128069 shot observations taken during the NBA 2014-2015 season. There are 21 variables including who took the shot, where on the floor was the shot taken from, who was the nearest defender, how far away was the nearest defender, time on the shot clock, and much more.

Source: Kaggle https://www.kaggle.com/dansbecker/nba-shot-logs

The uploader scraped this dataset from NBA's REST API

Brief introduction on REST API:

REST stands for Representational state transfer which essentially refers to a style of web architecture that has many underlying characteristics and governs the behavior of clients and servers. API is an acronym for Application Programming Interface. An API is a set of routines, protocols, and tools for building web-enabled and mobile-based applications. The API specifies how you can authenticate (optional), request, and receive data from the API server.

A REST API defines a set of functions which developers can perform requests and receive responses via HTTP protocol such as GET and POST. Because REST API's use HTTP, they can be used by practically any programming language and it is easy to test (it's a requirement of a REST API that the client and server are independent of each other allowing either to be coded in any language and improved upon supporting longevity and evolution).

The World Wide Web (WWW) is an example of a distributed system that uses REST protocol architecture to provide a hypermedia driven interface for websites.

Work plans that support the activity summary may be attached, and may be referenced to support the methodology and schedule summary.

1.2 Variables Explanation:

GAME_ID (numeric): The ID number of each game. Example: 21400899
MATCHUP (character): Explain the two opposing teams and game time. Example: MAR 04, 2015 - CHA @ BKN (Charlotte Bobcats VS Brooklyn Nets in Mar 04)
LOCATION (category): "A" stands for "Away", "H" stands for "Home".
W(category): "W" stands for "Win", "L" stands for "Lose".
FINAL_MARGIN (Integer): The final margin of the game. If it is positive, the team win the game and vice versa. Example: 24. (This team win the opponent by 24 points)
Player name (character): The name of the player who made this shot.
Player ID (numeric): The ID number of this player.
SHOT_NUMBER (Integer): The number of shots the player took in this game. Example: 3. (This is the third shot of this player in this game.)
PERIOD (category): Which period was the shot happened in (there are four periods:1-4). Example: 3. (This shot happened in the third period of the game.)
GAME_CLOCK (date time): This variable tells us when did the shot happen in the game. But it is a count down. Example: 1:09. (This shot happened when "1:09" was shown on the count down time board.)

SHOT_CLOCK (date time): When did the shot happen in the offensive time limit. There is a 24-second count down with every offense. Each team needs to score a point offense within 24 seconds. Example: 10.8. (This shot was made at 10.8 on the 24-second count down, which means the shot was made after 13.2 seconds past in this offense round. (24-10.8) before this shot was made).
DRIBBLES (numeric): How many dribbles did the player make before this shot. Example: 3. (This player made 3 dribbles before this shot.)
TOUCH_TIME (numeric): How long has the player held the ball before the shot. Example: 1.9. (This player has held the ball for 1.9 seconds before he made the shot).
SHOT_DIST (numeric): How far did this shot was made from the basket. Example: 7.7. (This shot was made 7.7 feet away from the basket)
PTS_TYPE(category): "2" stands for "2 points", "3" stands for "3 points"
SHOT_RESULT (category): "made" means the shot is successful, "missed" means the shot is missed.
CLOSEST_DEFENDER (character): The name of the closest defender.
CLOSE_DEF_DIST (numeric): How far was the closest defender away from the player on offense. Example: 1.3. (The closest defender was 1.3 feet away from the player).
FGM (category): FGM stands for field goal made, which means whether the shot was made. "1" means this shot made the points successfully. "0" means this shot was not made.
PTS (numeric): The points made of this particular shot. "0" stands for "0 points", "2" stands for "2 points" and 3 stands for "3 points".

1.3 Research Questions

What are the factors contributing to the NBA shot results?

How these factors influence the shot results?
How does offenders affect shot results?
How does defenders affect shot results?

1.4 Diagnosis of Dataset 1 NBA Shot Logs

1.4.1 Wrong Data Types / Too Many Levels

After we imported the dataset into R, the variable "GAME_CLOCK" is defined as factor with 719 levels. R cannot identify the form of this variable correctly so we need to transfer this variable into manageable form.
The variables GAME_ID, GAME_CLOCK, CLOSEST_DEFENDER, CLOSEST_DEFENDER_PLAYER_ID, PLAYER_NAME, and PLAYER_ID should be a character.
The variables PERIOD and PTS_TYPE should be factors.

1.4.2 Missing Value

There are missing values for the SHOT_CLOCK column and no other missing values in the rest of the dataset. The reason why some of these shot clock values are missing is because people turn off the shot clock when when game_clock reaches 0:24.

1.4.3 Invalid Data

There are some negative values in the variable "TOUCH_TIME" which are meaningless.

1.4.4 Mismatching Data

External resources: https://www.nba.com/media/dleague/1314-nba-rule-book.pdf
According the rule book, there are some mismatchings in the dataset. In the NBA games, the shortest-possible 3-pointer is 23.75 feet from the hoop (from either corner). However, the dataset has many observations where the SHOT_DIST is much smaller than 23.75 and the shot is still listed as a "3" in the PTS_TYPE category.
More precisely, we see that there are 9627 observations where the shot distance is less than 23.75 and the shot is still considered a 3, as demonstrated by executing in R:
Therefore, we need some modifications about the “PTS_TYPE” variable before we build the model.

1.4.5 Typos/Nonstandard Spelling

When we tried to compare the names in two datasets, we found some typos and nonstandard name spellings in the variable “Player_Name” of “NBA shot logs” dataset.

1.4.6 Strange Distribution

After we plotted the distributions of numeric variables, we found that the distribution of "TOUCH_TIME" is highly skewed.

1.4.7 Outliers

We used a boxplot method to identify the outliers in our datasets.

1.4.8 Variables Unrelated to Our Questions

Some variables such as “GAME_ID”, “MATCHUP”, “W”, “FINAL_MARGIN”, “Player ID”, and “CLOSEST_DEFENDER” are unrelated to our questions. We may drop these variables when building our models.

1.4.9 Useful Variables That Are Not found in The Dataset

Some other variables concerning the defenders are not in the dataset. For example, data of the height, position, and arm span of the defenders are also important for our analysis. We need to find some extra data to consummate our dataset.

1.4.10 Mis-Formatted Data

In the second dataset, the “Experience” column has mis-formatted data.
According to the data structure, the variable type of the “Experience” column was found to be a factor instead of a numeric. When we select data in the “Experience” column, we found that there is character “R” in some shells and these “R” are mis-formatted data.

2. PART TWO

Note: The specific R code and outputs are in the Appendix.

2.1 Dataset 1 NBA shot logs Cleaning

2.1.1 Wrong data types / Too many levels

We corrected the data types in R.

2.1.2 Missing Value

There were some missing values in “SHOT_CLOCK” and we deleted these missing values in R.

2.1.3 Invalid Data

There were some negative values in the variable “TOUCH_TIME”. We filtered the TOUCH_TIME between 0 and 24, since the touch time of an offender cannot exceed the offensive time limit, which is 24 seconds.

2.1.4 Mismatching Data

We relabeled the variable “PTS_TYPE” according to the “SHOT_DIST”. If the “SHOT_DIST” is greater than or equal to 23.75, the “PTS_TYPE” is labeled as “3”. Otherwise the “PTS_TYPE” is labeled as “2”.

2.1.5 Typos/Nonstandard Spelling

We verified the names according to the external resources: https://www.basketball-reference.com/players/
Some of the typos resulted from spelling mistakes, while some of the typos resulted from mis-format.
Here are some examples:
JON INGLES should be JOE INGLES
JIMMER DREDETTE should be JIMMER FREDETTE
MNTA ELLI should be MONTA ELLIS
DIRK NOWTIZSKI should be DIRK NOWITZKI

2.1.6 Strange Distributions

Since we are more interested in short "TOUCH_TIME", and we would like to analyze performance in some short key moments. We decided to filter the dataset where “TOUCH_TIME” is below 3. After filtering, the distribution of “TOUCH_TIME” is quite natural.

2.1.7 Outliers

We identified outliers by using the boxplot methond and then we removed these outliers. We wrote a new function to transfer the outliers in the datasets into NA, and then we removed these missing values.
As the range or distributions of variables would change after we deleted some of the observations, we decided to run the removing steps repeatedly to remove all the outliers.

2.1.8 Variables Unrelated to Our Questions

We dropped these unrelated variables to shrink our dataset from 21 variables to 13 variables.

2.1.9 Useful Variables That Are Not Found in The Dataset

We found a supplementrary dataset "NBA Players Stats", and we merged this dataset to acquire other useful variables that we needed.

2.1.9.1 Introduce Supplementary Data

Supplementary Dataset: NBA Players Stats - 2014-2015 (players_stats)
Observations: 490 Variables: 33
This dataset includes 490 Players Stats during the 2014-2015 season.
Source: Kaggle https://www.kaggle.com/drgilermo/nba-players-stats-20142015
The uploader scraped this dataset from
http://stats.nba.com/leaders#!?Season=2014-15&SeasonType=Regular%20Season&StatCategory=MIN&CF=MIN*G*2&PerMode=Totals
We verified the information of this dataset by referring to the external resources:
https://www.basketball-reference.com/players/.

Variables Explanation:
Games Played (numeric): the total number of games in which a player has participated
MIN (numeric): Minutes Played, the total number of minutes in which a player has played
PTS (numeric): Points, the total number of minutes in which a player has earned
FGM(numeric): Field Goals Made, total number of baskets in which a player has scored on any shot or tap other than a free throw—worth two or three points depending on the distance of the attempt from the basket.
FGA (numeric): Field Goals Attempted, total number of baskets in which a player has attempted on any shot or tap other than a free throw
FG% (numeric): Field Goal Percentage, ratio of field goals made to field goals attempted
3PM (numeric): 3 Points Field Goals Made, total number of baskets in which a player has scored on any shot or tap that worth three points
3PA (numeric): 3 Point Field Goals Attempted, total number of baskets in which a player has attempted on any shot or tap that worth three points
3P% (numeric): 3 Points Field Goals Percentage, ratio of three points field goals made to three points field goals attempted
FTM (numeric): Free Throws Made, total number of baskets in which a player has scored on free throw attempts
FTA (numeric): Free Throws Attempted, total number of baskets in which a player has attempted on free throws
FT% (numeric): Free Throw Percentage, ratio of free throws made to free throws attempted

OREB (numeric): Offensive Rebounds, total number of offensive rebounds for this player
DREB (numeric): Defensive Rebounds, total number of defensive rebounds for this player
REB (numeric): Rebounds**,** total number of rebounds for player
AST (numeric): Assists, total number of assists for this player
STL (numeric): Steals, total number of steals for this player
BLK (numeric): Blocks, total number of blocks for this player
TOV (numeric): Turnovers, total number of turnovers for this player
PF (numeric): personal fouls, total number of personal fouls for this player
EFF (numeric): Efficiency, (PTS + REB + AST + STL + BLK − ((FGA − FGM) + (FTA − FTM) + TO))
AST/TOV (numeric): Assists / Turnovers, player's number of assists in context with the number of turnovers the player has caused
STL/TOV (numeric): Steals / Turnovers, player's number of steals in context with the number of turnovers the player has caused
Age (numeric): the age of this player
Birth_Place (character): the place that the player was born
Birthdate (date time): the time that the player was born
Collage (character): the college where the player attended
Experience (category): Number of years the player has been in the NBA. (Should be numeric, needs cleaning).
Height (numeric): the height of this player
Pos(category): position, the player is in the one of five basketball positions—point guard (PG), the shooting guard (SG), the small forward (SF), the power forward (PF), and the center (C)
Team (category): the team which this player belongs to
Weight (numeric): the weight of this player
BMI (numeric): body mass index, units of kg/m2, of this player

2.1.9.2 Supplementary Dataset Cleaning

i. Missing Values

We diagnosed that there were over 60 records of missing values. The missing values that relevant to our study are: Age, Experience, Height, Weight, and, BMI. Therefore, we used external sources such as NBA.com and ESPN.com to find the data and imputed them manually.

ii. **Select Variables**

We selected defensive related variables to analyze the defensive side, and deleted useless variables. We removed a total of 27 out of 34 variables.

iii. **Process Mis-Formatted Data in "Experience" Variable**

We identified that value "R" in the Experience column represents Rookie. Therefore, we replaced "R" with "0" which means zero experience to the value, and then changed the variable type into numeric.

iv. Format "Name" Variable for Further Analysis to Merge Datasets

We identified that in order for us to merge two datasets, we must reformat the variable "Name" in players_stats dataset. In the first dataset, "shot_logs", the variable "closest_defender" is formatted as "Last Name, First Name."

2.1.9.3 Merge Two Datasets

We wanted to use the information in supplementary data to see how defenders could affect the shot results. We merged the two datasets by "CLOSEST_DEFENDER" (the name of defender) in dataset1 and the "Names" in supplementary data.

2.2 Key Dataset Issues Summary and Suggested Solutions

KEY PROBLEMS	DESCRIPTION OF PROBLEMS	SOLUTIONS
Wrong Data Type	R cannot identify the form of some variables correctly	Corrected the data types in R
Missing Value	There are some missing values for the SHOT_CLOCK variable	Deleted the missing value in R
Invalid Data	Negative values found in the variable "TOUCH_TIME"	Filtered the TOUCH_TIME between 0 and 24
Mismatching Data	Mismatching data found compared with real stats	Relabeled the variable "PTS_TYPE"
Typos	Typos and nonstandard name spelling found in the variable Player_Name	Verified the names and Corrected typos
Strange Distribution	The distribution of "TOUCH_TIME" is highly skew	Filtered the dataset where "TOUCH_TIME" is below 3
Outliers	Boxplot identifies the outliers in our datasets	Removed outliers
Unrelated Variables	Some variables are unrelated	Dropped these unrelated variables
Extra Variables Needed	Some other variables concerning the defenders are not in the dataset	Merged the other dataset NBA Players Stats
Mis-Formatted Data	The variable "Experience" has mis-formatted data	Replaced "R" with valve "0"

2.3 Dataset Exploration

Note: the specific R code and outputs is in the Appendix.

As the size of our dataset is quite big. We sampled 1000 observations from our cleaned and merged dataset. And we used scatterplot matrix to see the relationship between two variables.

2.4 What We Have Done

CONTENTS	PLACE
Identified key data issues in your dataset.	1.4
Diagnosed the data issue using tools in R.	1.4 and Appendix1
Demonstrated how to resolve the data issue using tools in R.	2.1 and Appendix
Showed a before- and after-picture of each data issue, confirming that the issue has been fixed.	Appendix1
Exercised good judgment about issues that are not worth fixing.	2.1
Consulted external sources as needed to confirm issues with the data, or to fix these issues.	1.4 and 2.1
Data cleaning/processing effort is aligned with the question you have outlined for the project.	2.1

3. APPENDIXES

i. Appendix1 Rscript and Output: 5200Appendix1_CodeandOutput.pdf
ii. Appendix2 Analytical Dataset: shotmerge.RData
iii. Appendix3 Data Dictionary: Data Dictionary.pdf