PREDICTIVE ANALYSIS OF SUCCESSFUL BASKETBALL SHOTS: THE EUROLEAGUE CASE

by
CEM YIĞMAN

Submitted to the Graduate School of Management
in partial fulfillment of the requirements for the degree of
Master of Science in Business Analytics

Sabancı University
July 2018

PREDICTIVE ANALYSIS OF
SUCCESSFUL BASKETBALL SHOTS: THE EUROLEAGUE CASE

Approved by:

Assoc. Prof. Dr. Raha Akhavan-Tabatabaei .....
(Thesis Supervisor)

Prof. Dr. Nihat Kasap .....

Asst. Prof. Dr. Mustafa Hayri Tongarlak .....

Date of approval: .....

ABSTRACT

PREDICTIVE ANALYSIS OF
SUCCESSFUL BASKETBALL SHOTS: THE EUROLEAGUE CASE

CEM YIĞMAN

Business Analytics, Master of Science Thesis, July 2018

Thesis Supervisors: Assoc. Prof. Dr. Raha Akhavan-Tabatabaei,
Assoc. Prof. Dr. Abdullah Daşcı

Keywords: Basketball Analytics, Predictive Modeling, Binary Classification

Basketball industry creates vast amounts of data from which the organizations benefit to improve their business processes like revenue management, roster selection, fan engagement, and on-field decision making. Sophisticated data collection systems are being developed in order to get the maximum benefit from analysis of the movements and actions of all elements in the field. Since elite teams do not have huge differences when compared to each other in terms of advanced fundamental, physical capacity, and motivation, this valuable information is helping them to develop data driven decision making applications which give them a significant advantage on winning. In this thesis, we analyze ten seasons of the Euroleague professional basketball data consisting of spatiotemporal, player based, and situational variables such as score difference, shot type, and home or away team. Using these variables, we build predictive models for the accurate prediction of successful shots. We develop binary classification methods such as logistic regression, random forest, naive bayes, support vector machines, and artificial neural networks. We compare these models to evaluate the best approach for classification problems of successful basketball shots. Among all models we applied, random forest is the most accurate and logistic regression is computationally the most efficient model.

ÖZET

BAŞARILI BASKETBOL ŞUTLARININ
TAHMİNSEL ANALİZİ: EUROLEAGUE ÖRNEĞİ

CEM YIĞMAN

İş Analitiği, Yüksek Lisans Tezi, Temmuz 2018

Tez Danışmanları: Doç. Dr. Raha Akhavan-Tabatabaei, Doç. Dr. Abdullah Daşcı

Anahtar Kelimeler: Basketbol Analitiği, Tahminsel Analiz, İkili Sınıflandırma

Basketbol sporu, kulüplerin gelir yönetimi, kadro planlama, taraftar etkileşim yönetimi ve saha içi karar verme sistemlerinde yararlandıkları yüksek miktarlarda veri üretiyor. Gelişmiş veri toplama sistemlerinden elde edilen veriler, oyuncuların ve topun sahadaki hareketinden en yüksek faydayı sağlayacak analitik uygulamaların geliştirilmesinde kullanılıyor. Üst düzey takımların arasında temel basketbol tekniği, fiziksel kapasite ve motivasyon parametreleri açısından çok büyük büyük farklar olmadığı göz önünde bulundurulursa, bu değeri bilgi, takımların karar verme süreçlerinde destekliyor ve başarılı olma konusunda kayda değer bir etkide bulunuyor. Bu tezde on sezounluk Eurloegue profesyonel basketbol ligi verisini, oyun içi değişkenleri ile inceleyerek, başarılı basketbol şutları ile ilgili tahmin modelleri yaratıyoruz. İkili sınıflandırma metodlarından, lojistik regresyon, random forest, naive Bayes, support vector makinaları ve yapay sinir ağları gibi modelleri kullanıyoruz. Bu farklı modellerin basketbol alanındaki verimlerini değerlendirerek, bu alan için en iyi modeli tespit ediyoruz. Çalıştırdığımız modeller arsında random forest en iyi tahmin sonucunu verirken, lojistik regresyon gerektirdiği bilgisayar kapasitesi açısından en verimli model oldu.

To my beautiful and beloved wife Deniz

ACKNOWLEDGMENTS

I consider myself as a very lucky graduate student who worked under the supervision of Dr. Raha Akhavan-Tabatabaei and Dr. Abdullah Daşcı at Sabancı University School of Management. During the course of this research, their continuous support and academic wisdom helped me in various ways.

I give my deep thanks to office staff at Sabancı University School of Management, especially to Ms. Ekin Başat, for her continuous support and great body of knowledge on academic processes which helped me during the past two years.

Furthermore, I must express my gratitude to Dr. Elif Ayiter for not only accepting me to her classes but also for sharing her invaluable artistic and cultural intelligence.

Completing this research would have been more difficult without the support provided by the other members of the Collaboration Space in Sabancı University.

Finally, I also thank my wife who encouraged and supported me throughout the time of my research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACB - Asociación de Clubes de Baloncesto (Spanish Basketball League).

API - Application Programming Interface

CRAN - The Comprehensive R Archive Network

ETM - End-of-game Tactics Metric

FIBA - Fédération Internationale de Basket-Ball (The International Basketball Federation)

GLM - Generalized Linear Model

MLB - Major League Baseball

NBA - National Basketball Association

SSE - Sum of Squared Errors

XG BOOST - Extreme Gradient Boosting

XML - Extensible Markup Language

VAR - Video Assistant Referee

2FG - 2-point field goal

3FG - 3-point field goal

CHAPTER 1

INTRODUCTION

Analytics has been one of the most important areas that industries and corporations benefit from, since the technological advances in computational power and the increased ability to collect and analyze large quantities of data. Sophisticated on-line and off-line data collection methods supported by economically sustainable data storage systems as well as cloud storage systems led analytics to become the major decision support system in the past decades. According to LaValle et al. (2011), those organizations that strongly believe in the use of business information systems and analytics were twice as likely to be top performers as compared to other organizations.

Alamar and Mehrotra (2011) describe the term "sport analytics" as "the management of structured historical data, the application of predictive analytic models that utilize that data, and the use of information systems to inform decision makers and enable them to help their organizations in gaining a competitive advantage on the field of play."

According to Alamar (2013), the essence of sport analytics includes managing data, using predictive analytics, and informing decision makers to provide competitive advantage. This reveals a continuous analysis process in sport settings wherein situation-specific information informs analytical algorithms, which in turn guide data collection and analysis (Baker and Kwartler, 2015).

The main focus of sports analytics is to find better ways to increase the efficiency on decision making in sports, whether it is an individual sport like golf and the players are trying to predict their opponents' putt counts per holes, or a team sport like volleyball and the coaches are trying to optimize their attack styles and minimize their attack errors.

Like all other industries which benefit from analytical approaches, there are also many opportunities for improvement to apply analytics to sports. Figure 1.1 is an adaptation of sports analytics taxonomy designed by Cokins et al. (2016) which can be examined in 8 major fields. Based on this taxonomy, the focus of this thesis can be classified as "Game Win Strategy" which is dependent on both individual and team based analytical applications.

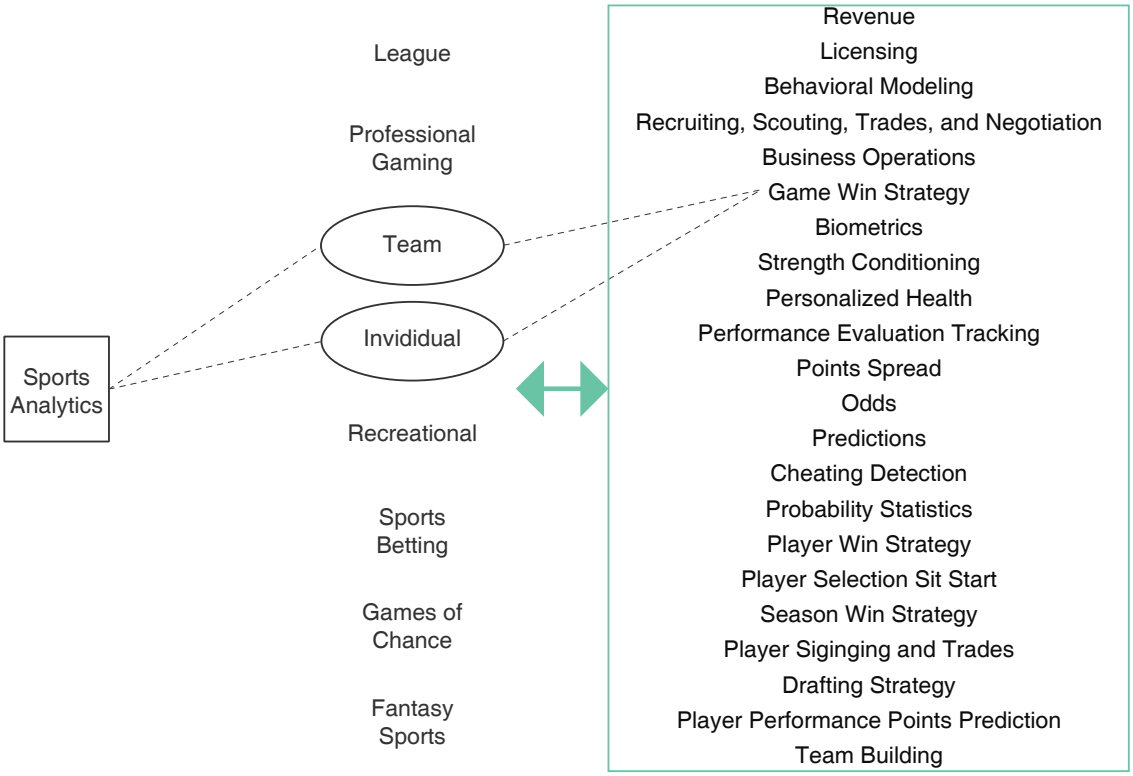

Figure 1.1: Adaptation of Cokins et al. (2016) Sports Analytics Taxonomy

The opportunities that come with the efficient analytics applications evoke organizations to develop sophisticated high-tech applications to collect more accurate data. As the pioneer sport that started to apply analytics, baseball has the most detailed and precise data collection system among all sports. Statcast, which was introduced to all 30 The United States' Major League Baseball (MLB) stadiums in 2015, is a high-speed, high-accuracy, automated tool developed to analyze player movements and athletic abilities as well as movement of the baseball. It captures data using high-resolution optical cameras with radar equipment (Figure 1.2). As of 2018, Statcast is able to collect 30 variables about pitching, hitting, base-running, and fielding.

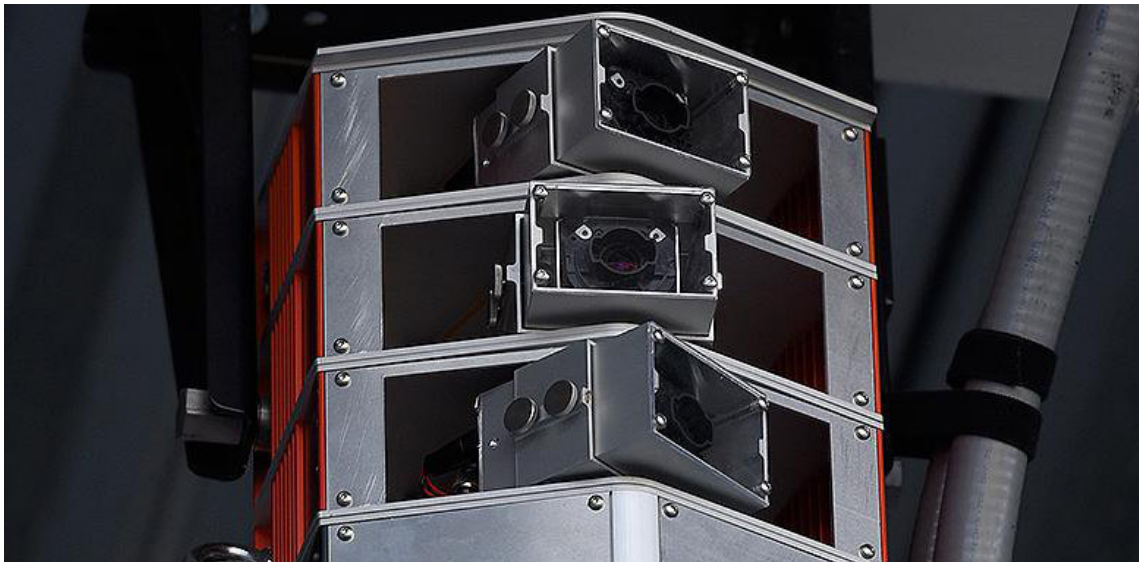

Figure 1.2: Wass N. (Photographer)(2015). Statcast high-speed cameras

The United States' National Basketball Association (NBA) also introduced its player tracking technology starting from season 2013-2014, to increase the amount of data being collected hence to improve analytical applications. All twenty nine NBA arenes are equiped with cameras which are able to collect data in real time at a rate of twenty five frames per second. The system collects X and Y positions for the players and X, Y, and Z positions for the ball. The majority of the data collected is open to public.

1.1 Sports Analytics History

Analytics has been used since the middle of the nineteenth century to improve various aspects of sports including revenue maximization, merchandising, roster selection and on-field decision making. The first application of analytics in the field of sports was developed by Henry Chadwick, a sportswriter, who first used statistics in the field of baseball at 1858. He made an analysis using box scores of baseball games which led the development of the evaluation method called "Sabermetrics" in the middle of the 20th century ("Sabr" stands for the Society for American Baseball Research). Starting from the early 1980s, Major League Baseball (MLB) teams started to employ sabermetricians who worked on advanced metrics on team statistics.

Towards the end of the 20th century, teams started to use sabermetrics to obtain relatively undervalued players. During 1990s Oakland Athletics' Billy Beane used a more

quantitative approach on building a roster, including on-base performance of the players, which led Athletics to win 20 games in a row in 2002 while other teams were building their rosters with physical player attributes. The system Beane developed not only revolutionized the way baseball benefited from the field of statistics, it also started a new era in professional sports.

The boom in sports analytics happened at the beginning of the twenty-first century with the increased usage of analytical applications in the NBA. Arthur and Watt (2016) analyzed the number of front-office analysts in NBA teams using the data they gathered from https://basketball.realgm.com. Their findings showed that there is an increasing number of front-office analysts in NBA teams since the starting year of their analysis 2008. They split the years from 2008 to 2012 as "Early Analytics Adoption Period" and from 2013 to 2016 as "Late Analytics Adoption Period". Figure 1.3 represents the total number of analytics personnel in NBA teams. Their statistical study showed that the teams employed at least one full-time analytics personnel by the season 2012 averaged about 7 more wins per season from 2008 to 2012 than the teams which had not. The study also showed that, the early adopters which employed analytics personnel before 2012, averaged about 8 more wins than their late-adopting opponents even for the seasons from 2013 to 2016.

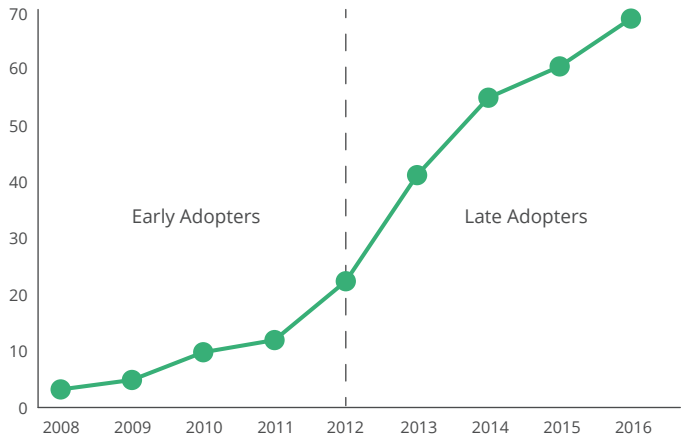

Figure 1.3: Number of analytics personnel per season in the NBA

(Adapted from https://basketball.realgm.com)

1.2 Basketball: A Growing Industry

Basketball is a highly competitive sport, which has evolved to an industry from being “just a sport”. There are two major elite basketball organizations in the world. The National Basketball Association (NBA) is the largest indoor sports organization in the world, in revenue with $7.4 billion reported in 2018 according to Fortune Magazine. Established in 2000, The Euroleague® organization has become the second largest basketball organization in the world after the NBA. It boasts an average of 8,864 attendees per game and a 2.25 million accumulated audience (2017 - 2018). In the 2017 - 2018 season, 16 teams from 9 countries played 260 games to achieve the Euroleague championship. It is a hard task to distinguish only the Euroleague revenues and budgets of the Euroleague teams, since they are not based only on the Euroleague but also on their local leagues unlike the NBA teams. It is clear that due to financial improvements since its foundation, Euroleague has become the most important indoor team sports organization in Europe with teams having a total budget of €256 million (2016 - 2017).

1.3 Basketball Analytics and Data Collection in Basketball

Data collection systems and analytical applications are also very common in basketball. NBA teams have high-speed cameras installed in their venues in order to track players and the ball which gives the opportunity to better evaluate the players and create set plays according to the opponent’s defensive system. Euroleague has also introduced its advanced player tracking system at the 2017 Final Four. SportsVU was the first player tracking system used in Euroleague history which is able to track the positions of the players, referees and the basketball. The system will provide accurate and efficient data to organizations which will result in an increase of data driven decision support applications in the Euroleague basketball. The organizations will benefit from analytical approaches more on their on-field applications including shot success optimization and offensive and defensive strategy planning. Off-field applications like revenue management, roster selection, and optimizing fan engagement will also benefit from this data and analytical applications to a lesser extent.

The descriptive analysis for points per 100 shot attempts in the Table 1.1 clearly indicates that the team levels are very close to each other for the 59 teams which participated to the Euroleague from 2007 - 2008 season to 2016 - 2017 season. While the most efficient team, in terms of points per 100 attempts, has 114 points, the worst has 92, and the mean points for all teams is 102. With only 10% improvement, the worst team becomes the average in the Euroleague organization. There is a significant opportunity for the teams to benefit from analytical applications to improve their shooting efficiency and hence, their revenues with the increased fan engagement using data driven decision making systems.

Table 1.1: Team based descriptive statistics for points per 100 shot attempts

Metric	Value
Mean	102
Standard Error	0.55
Median	103
Mode	107
Standard Deviation	4.25
Sample Variance	18.08
Kurtosis	0.29
Skewness	-0.25
Range	22
Minimum	92
Maximum	114

1.4 Euroleague Season Structure

According to Euroleague Bylaws (2018), a Euroleague competition is played in three different phases; Regular season, Playoffs, and Final Four. The regular season starts on July 1st and the season ends on June 30th of the following year. Sixteen teams plays in a round-robin format (each team against all others, both at home and away). Top eight teams at the end of the regular season qualify for the playoffs which is held in a best of five games format. The winners of the playoff stage advance to the final four stage. The winners of the semi-finals at the final four, play the championship game.

1.5 Thesis Objectives

In this thesis, we build predictive models for the accurate prediction of successful basketball shots using spatiotemporal, player based and conditional variables. These models can be used by the coaches to make better decisions in their offensive organizations. We also compare these predictive models to evaluate the best model for a binary classification problem in this domain. This thesis will first, in Chapter 2, introduce the concepts and analysis methods available in the basketball analytics literature. In Chapter 3, the data source, data gathering, and variable transformations will be described. Chapter 4 will present the methods that are used in predictive analysis of basketball shots. Finally, Chapter 5 will conclude the dissertation and summarize its contribution to the field of basketball analytics.

CHAPTER 2

LITERATURE REVIEW

There is a growing body of literature that recognizes the importance of analytics in the field of sports. This chapter presents the current literature on the key aspects of basketball analytics including data collection, variable selection, variable creation, and analysis methods. The goal of this chapter is to establish the significance of the field of study and then to identify the main contribution that could be made to the literature. The literature is ordered according to themes of the research.

Reich et al. (2006) analyzed the conventional shot charts, which have been used by the NBA coaches as a descriptive analysis tool, using hierarchical spatial models and Markov Chain Monte Carlo (MCMC) methods to assist the coaches with the best possible shot location information. The dataset was downloaded from www.espn.com, which contains 1,139 shots Minnesota Timberwolves' veteran point guard Sam Cassell took during 2003 - 2004 season. For each shot, they have game clock time elapsed since Cassell's last shot (excluding time on bench), location in polar coordinates, shot result, and 10 binary variables which are shown in Table 2.1. They also have the interaction term NOKG x NOLS because Cassell's performance may be affected by both Kevin Garnett's and Latrell Sprewell's absence. The article developed a statistical model for analyzing basketball shot-charts and found a meaningful relationship between spatial-temporal variables and a successful shot.

Table 2.1: Binary variables used by Reich et al. (2006)

Variable Name	Variable equals “1” when
NOKG	Kevin Garnett is not in the game
NOLS	Latrell Sprewell is not in the game
HOME	The game is played in Minnesota
NOREST	The Timberwolves had less than 2 days of since their last game
2HALF/OT	Second half or overtime
BEHIND	The Timberwolves are losing
BLOCK	The opponent averages more than 4.8 blocks per game
FGPALL	The opponent allowed a field goal percentage under 44 %
MISSLAST	Cassell missed his previous shot
TEAMFGA	The Timerwolves took more than 80 shots in the current game

The most common shooting evaluation factor is measured using “Effective Field Goal Percentage” (eFG%) which is used to adjust the data for the fact that a 3-point field goal is worth one more point than a 2-point field goal. Chang et al. (2014) introduced and proposed methodologies for deriving and evaluating two new metrics: (1) Effective Shot Quality (ESQ) and (2) EFG+, which is EFG minus ESQ, a measure of shooting ability above expectation. They discussed how this definition re-characterized performance for teams and players and how this type of analysis can affect analysis beyond shooting. With the help of player tracking data with the variable “Defender Distance”, they applied several models that are generally applied in machine learning, like decision trees and logistic regression. With their proposed new variables, they were able to quantify (1) the quality of shots that a team or player is generating and (2) their skill in hitting those shots, which were previously confounded in EFG.

Shortridge et al. (2014) introduced several measures of relative field goal effectiveness that explicitly account for spatial variability in scoring. These measures identify an expected point total for locations across the court and contrast this expected point total with the actual points scored by a particular player. Critically, these metrics can be either calculated locally for each position on the floor or they can be aggregated into a single global measures of relative shooting effectiveness. They used a database recording every shot taken during the 2011–2012 NBA regular season. Data were obtained from www.espn.com which includes cartesian coordinates for over 141,000 field goal attempts and

detailed attribute information including who took the shot and whether or not the attempt resulted in a made basket. To construct a robust estimate of local field goal probability, they used Empirical Bayes (EB) rate estimation. It appears that EB surface is much more smoother than the raw field goal rate. The noise present in the raw rates has been removed as shown in the Figure 2.1. They also introduced Spatial Shooting Effectiveness (SSE) and Points Above League Average (POLA), which compare and contrast shooters using a spatially explicit perspective that accounts for the individual shot constellation and the relative performance of all other shooters from those spatial locations.

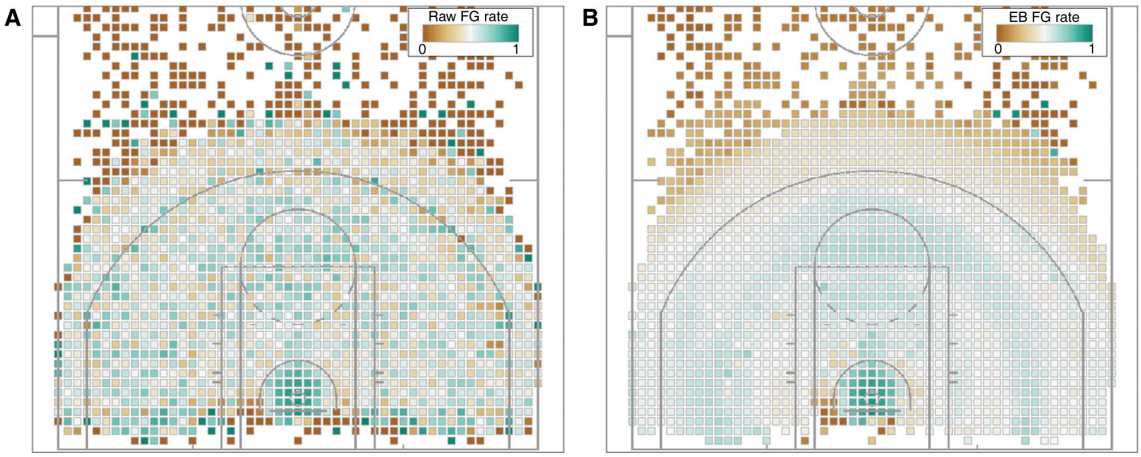

Figure 2.1: (A) Raw field goal rate surface; (B) Empirical Bayesian smoothed rate

Spatial aspects of basketball have been widely analyzed by the researchers. Goldberry (2012) investigates spatial and visual analytics as means to enhance basketball expertise. This article uses game data for every NBA game played between 2006 and 2011 and the authors compiled a spatial field goal database that included Cartesian coordinates (x,y) for every field goal attempt in this 5-year period. This data set includes player name, shot location, and shot outcome for over 700,000 field goal attempts. The authors mapped the shot data atop a base map of a NBA basketball court according to "Range" metric, which shows the effective shooting range of a player across all scoring cells as shown in the Figure 2.2.

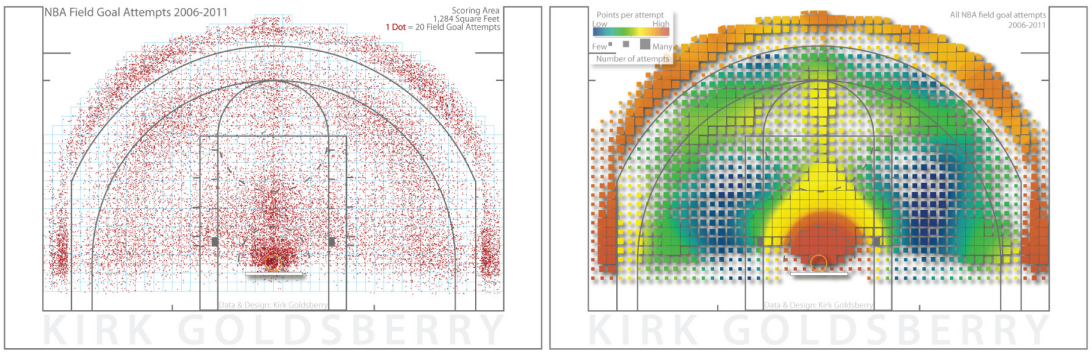

Figure 2.2: Composite shot map from 2006-2011 NBA season

After the rise of the big data and analytics applications, sports organizations needed to develop online applications to assist coaches with the information "who should take the shot?" For this purpose, Wright et al. (2016) focused on a shot recommender system for NBA coaches. They used 2015 -2016 season NBA data which was downloaded using the API provided by the www.nba.com website. They stated that predicting shot success in basketball is a challenging task because the existing data is sparse, which means that the outcomes of all combinations are not represented in the datasets evenly. They compared factorization machine model performance with logistic regression and support vector machines. They concluded that the factorization machine which uses k=25 and optimized using stochastic gradient descent performs better for latent variables.

The temporal properties of basketball have also been the focus of researchers. Garcia et al. (2013) created a model to identify basketball game performance indicators which best discriminate winners and losers in regular season and playoffs. They used a sample consisting of 323 games of Spanish Basketball League (ACB) which is collected manually by professional technicians on 306 regular season and 17 playoff games. They split the dataset according to end-of-game score differences. *Balanced* games are discriminated as the score is equal or difference is below 12 points, *unbalanced* games as the difference is between 13 and 28 points and *very unbalanced* games as the difference is above 28 points. Their study showed that "while in regular season assists, defensive rebounds, and field goal percentages were significant, the winning teams' superiority was only in defensive rebounding in the playoff games."

The in-game temporal aspects of a successful basketball shot is also examined by the researchers. McFarlane (2018) introduced a method for end-of-game decision making model based on a probabilistic method which evaluates the situation for the last three minutes of NBA games. The data consists of NBA regular season games from 2011 to 2015 season and gathered from stats.nba.com. The point spread data was collected from sportsdatabase.com. McFarlane (2018) modeled the state of a game "as a function of possession, time remaining, and score differential" which led the problem to be able to handled as a "Markov Chain with a transition probabilities based on team statistics." Logistic regression was used to build the win probability model to introduce End-of-game Tactics Metric (ETM) for helping decision makers on whether offensively to attempt 2 point or 3 point field goal or defensively make an intentional foul.

While there are numerous studies on spatial-temporal properties that influence the success of a basketball shot, behavioral properties are also used in models to describe the outcomes of basketball shots. Avugos et al. (2013) reconsidered the "hot hand" phenomenon which in basketball could be described as the tendency to reject the randomness effect due to the belief that after a streak of successful shots, subsequent success becomes more likely. Reviewing the results of the meta-analysis, Avugos et al. (2013) concluded that the "results provided sufficient evidence that argues against the existence of the hot hand" in the field of basketball. On the other hand, Arkes (2010) found evidence using 64,698 free throws taken in the NBA 2005 - 2006 season, supporting the "hot hand" in free throws. He shows that when the first throw is successful, there is a significantly higher probability that the second free throw will be successful. The variables in his fixed-effects logit model are;

• The sequence of a set of free throws (Players take 2 free throws for fouls. This variable shows whether that shot was the first or the second).

• The game quarter of the game when the shot took place.

• Number of free throws the player made in the prior 1, 2, 3, 4, and 5 attempts.

• Number of free throws the player made in the prior 1, 2, and 3 attempts.

• How many of the past 10 free throws the team attempted were made.

In contrast to “hot hand”, the significance of “momentum effect”, which could be described as the increase in the probability of a team winning a game if that team has been playing well in the last few games, is proved by Arkes (2011) using the data from NBA in seasons from 2007 to 2009.” The variables for the analysis were home vs. away team and resting days each team had before the given game. Arkes (2011) then constructed a variable on how the teams did in their previous 3 and 5 games”. The study used Bradley-Terry (Bradley and Terry, 1952) model concluding that they “find evidence for a positive momentum effect, in that stronger performance over the past 3 or 5 games is associated with a higher probability of winning the next game, with the estimated effect being stronger for home teams.”

Since it shares its data publicly, the most common data set that has been used in basketball analytics is the NBA data. NBA provides player tracking data that is a valuable information for predictive modeling. The literature cited in this thesis also mostly uses NBA data. While seven of the papers are working with NBA data, two of them are focused on the Spanish Basketball League (ACB). We on the other hand, focus on the most important basketball tournament in Europe, the Euroleague, to build our predictive models.

The existing literature shows the importance of analytics in the field of sports. The most common area that has been in the analytical focus is the spatiotemporal properties of basketball. Three of the cited papers on this thesis use radially designed zones and two of them use Cartesian coordinates / rectangular zones to examine the shot location’s effect on the outcomes. We also designed radial spatial zones to understand shot location’s effect on a successful basketball shot. But since we use the shot type variable that distinguishes two-point and three-point field goals, we do not use radial zones based on the three-point line.

To the best of our knowledge, there are only a handful of models in the literature that focus on the influential variables for successful basketball shots which are summarized in Table 2.2. Five of the cited papers in this thesis focus on the variables that affect winning or losing a game while the others focus on the variables that influence individual shots. Four of them use logistic regression to estimate the outcomes while the other ones

use wide-spread techniques.

Our research investigates whether different models give significantly different outcomes in predicting successful basketball shots. For this purpose, we apply logistic regression, random forest, Naive Bayes, support vector machines, and artificial neural networks models in the software R® to understand the variables' effects on a successful basketball shot.

Table 2.2: Literature review summary table

Author(s) (Year)	Title	Topic	Data Source	Analysis Method
Chang et al. (2014)	Quantifying shot quality in the NBA	Maximizing the expected value of each shot opportunity	NBA player tracking data for 2013-2014 season	Decision trees, logistic regression, gaussian process regression
Reich et al. (2006)	A spatial analysis of basketball shot chart data	Optimizing shots using polar zones	NBA Minnesota Timberwolves Sam Casell 2003-2004 season	Markov chain monte carlo
Shortridge et al. (2014)	Quantifying spatial relative field goal efficiency in basketball	Maximize spatial shooting effectiveness	NBA 2011-2012 season	Empirical bayes rate estimator
Gómez et al. (2013)	Possession effectiveness in elite basketball according to situational variables in different game periods	Identifying performance indicators in predicting effectiveness of ball possessions	Spanish basketball league 2006 - 2007 season	Logistic regression
Wright et al. (2016)	Shot Recommender System for NBA Coaches	Determining the odds of success of a shot	NBA season 2015 - 2016 season	Factorization machine
García et al. (2013)	Identifying Basketball Performance Indicators	Discriminating winners and losers	Spanish basketball league 2007 - 2008 season	Discriminant analysis
McFarlane (2018)	Evaluating NBA end-of-game Decision Making	Evaluating the tactical decisions	NBA season from 2011 to 2015	Logistic regression
Arkes (2011)	Finally, Evidence for a Momentum Effect in the NBA	Evaluating the `momentum effect`	NBA seasons from 2007 to 2009	Econometric model (Bradley-Terry Model)
Arkes (2010)	Revisiting the Hot Hand Theory with Free Throw Data in a Multivariate Framework	The “hot hand” effect in free throws	NBA season 2005 - 2006 free throws	Fixed effects logit model

CHAPTER 3

DATA GATHERING AND PREPROCESSING

Perhaps, the most important part for an analytics research in the field of sports or any other field is data gathering. In this chapter, firstly the data source will be presented followed by the data download process. After describing the variables that come with the dataset, we explain our method to create some new variables as well. We will conclude this chapter with the descriptive statistics.

The data for this thesis was obtained from the Euroleague website www.euroleague.net "Game Center" section. This section includes individual and team statistics starting from the 2000 - 2001 Euroleague season. After 2007 - 2008 season "Play by Play", "Graphic Stats" and "Shooting Chart" tabs were added. In the "Play by Play" tab the game evolution is represented as a possession history of each game (Figure 3.1)

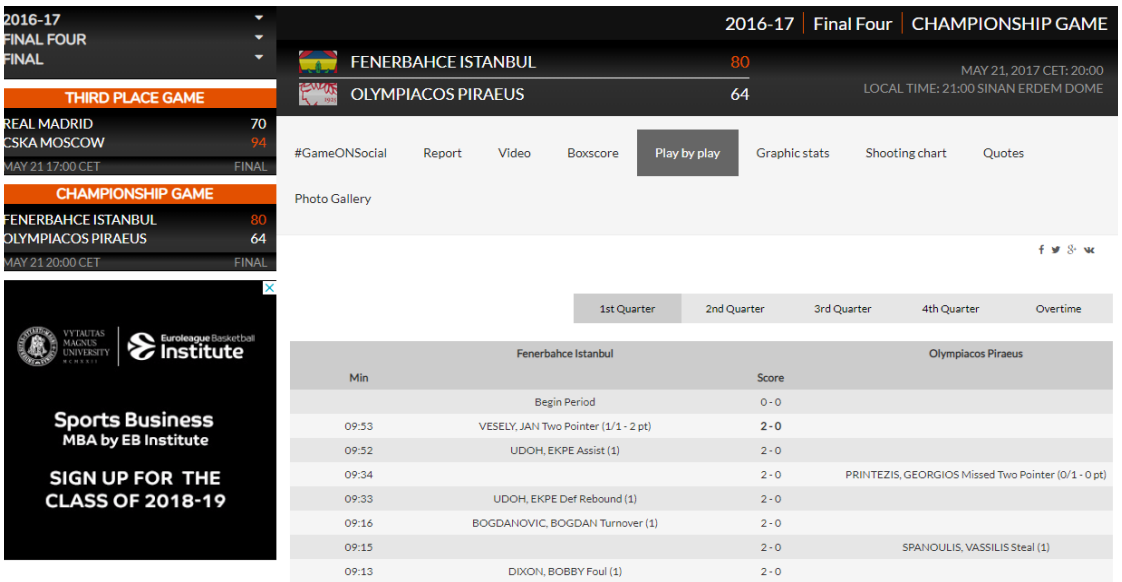

Figure 3.1: "Play by Play" tab of the Euroleague website

As depicted in Figure 3.2, “Graphic Stats” tab includes descriptive statistics which visually presents the game evolution in different metrics.

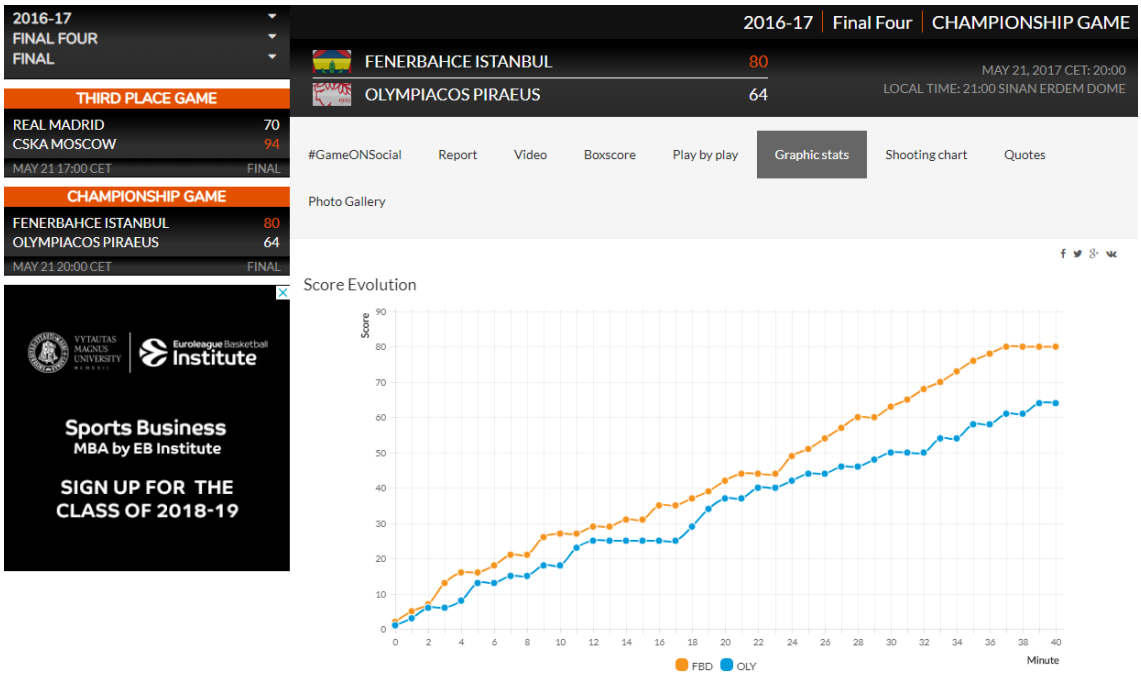

Figure 3.2: “Graphic Stats” tab of the Euroleague website

The game data was scraped from the “Shooting Chart” tab of the website which is depicted in Figure 3.3. This tab shows the successful shots with a solid circle and unsuccessful shot attempts with an empty circle on a virtual basketball court with the dimensions 800 x 400 pixels. There are also filters available for players, quarters, and shot types.

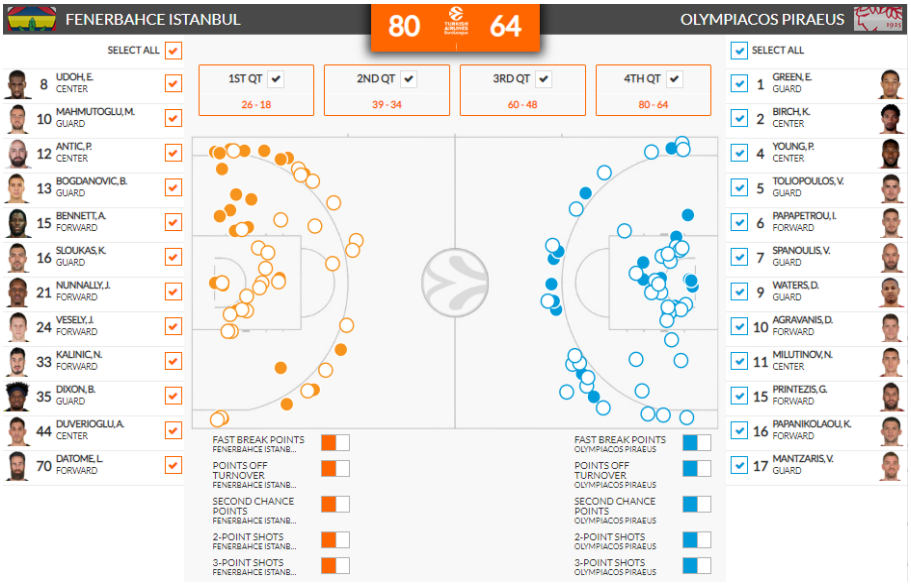

Figure 3.3: “Shooting Chart” tab of the Euroleague website

Euroleague website provides a javascript API which can be used on scraping game data using the software R. The library “jsonlite” developed by Jeroen Ooms (2014) is used to call the API. The scraped game data includes the variables shown in Table 3.1.

Table 3.1: Game data variables

Variable	Variable Description
Shot ID	Unique shot ID for every game
Team	Player’s team
Player ID & Palyer Name	Player Information
Action Type	2FG, 3FG, Layup or Dunk
Points	If the attempt is successful 2 or 3
X and Y coordinates for the virtual field	Cartesian coordinates
Minute	Shot timing
Season	Season
Game ID	Game

Static game information was scraped from the “Game Information” banner shown on Figure 3.4. Firstly the ”htmlparse” function from the library “XML”developed by Duncan Temple Lang and the CRAN Team (2017) was used to parse the HTML code from the Eurolegue web site. After selecting the necessary parts of the html code with “xpathSApply” function, “GET” function from the library “httr” developed by Hadley Wickham (2017) was used to get the response from the web site. R data frame was created with “xpathSApply” and the data were stored in the local hard drive with the “write.csv” function called to store the data as a comma separated value file. Game variables are given in Table 3.2.

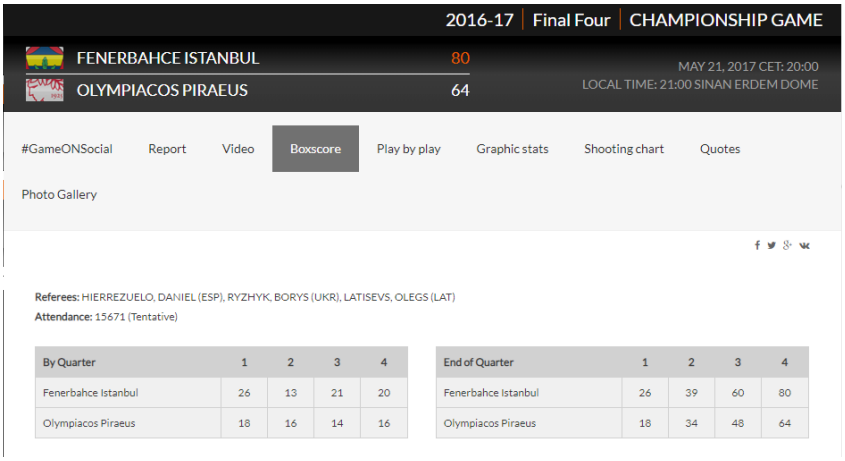

Figure 3.4: “Game Information” banner of the Euroleague website

Table 3.2: Game information variables

Variable
Game Date
Game Hour
Game Venue
Venue Capacity
Home Team
Away Team
Game Score
Period Scores
Game Phase
Referees
Referee Nationalities

Player-based variables were scraped from the “Players” section of the Euroleague website http://www.euroleague.net/competition/players (Figure 3.5). For gathering this data same libraries and technique were used as in game information variables. Player variables can be examined in the Table 3.3.

UDOH, EKPE

FENERBAHCE ISTANBUL | 8 | CENTER
HEIGHT: 2.08 | BORN: 20 MAY, 1987 | NATIONALITY: UNITED STATES OF AMERICA |
@ekpeudoh

Current Stats | Rankings | Career Highs | Biography | All-Time Stats

EuroLeague 2016-17 STATISTICS

								Rebounds						Blocks		Fouls		
	G	GS	Min	Pts	2FG	3FG	FT	O	D	T	As	St	To	Fv	Ag	Cm	Rv	PIR
Totals	31	22	992:51	376	150/256	0/1	76/118	76	165	241	68	30	38	68	7	64	116	641
Averages	31	22	32:01	12.1	58.6%	0%	64.4%	2.5	5.3	7.8	2.2	1	1.2	2.2	0.2	2.1	3.7	20.7

Figure 3.5: “Player Information” section of the Euroleague website

Table 3.3: Player information variables

Variable
Player ID
Player Name
Player Height
Birthday
Nationality
Position (Guard, Forward, Center)

The raw database was created using game variables, game information, and player information. The dataset consists of 10 seasons between 2007 - 2008 and 2016 - 2017. The free throws and dunks were excluded due to their unique nature and non-spatial properties. Shortridge et al. (2014) propose to only include those players attempting at least 250 shots in the NBA season they were analyzing. Since Euroleague teams play less games in each season, in our analysis only the individual players who attempted at least 100 shots for each season are included. Cartesian coordinates provided by the website were converted to polar coordinates, with angle and distance information, to be able to create radial zones. The remaining dataset contains 169,691 shot attempts of total 2,233 games and 473 unique players.

3.1 Variable Definitions

In this section the collected and newly generated variables will be introduced and their properties or calculation methods will be described.

Player Position: We identifiy player roles as “Guard”, “Forward”, and “Center” as Abdelkrim et al. (2010).

Action: The shot attempts were defined as “2FG” (2-point field goal attempt), “3FG” (3-point field goal attempt), and “Layup” (Layup attempt).

Radial Zone: The effect of shot location on the success of a shot attempt was analyzed using a radial grid. For this purpose we use k-means clustering analysis in the software R. Firstly we standardize the shot distances for comparability. Secondly, we create fifteen different models with number of clusters starting from two to fifteen. Considering a lower sum of squared distances within groups and model performances, we apply 3-cluster solution to the data (Appendix A-1). As a result we have three distance groups as described below.

• First group with distance lower or equal to 4.60 meters

• Second group with distance higher than 4.60 meters and lower or equal to 7.50 meters

• Third group with distance higher than 7.50 meters

Shot angle information is also used in the radial grid design. The half-court was divided into 7 different angular-based zones and this zones was blended with distance information. Since it is clear that the success rates for layups, which are the shot attempts from very near to the hoop, would have very high success rates, the angular zones in the first distance cluster were merged. The resulting radial grid is presented in Figure 3.6.

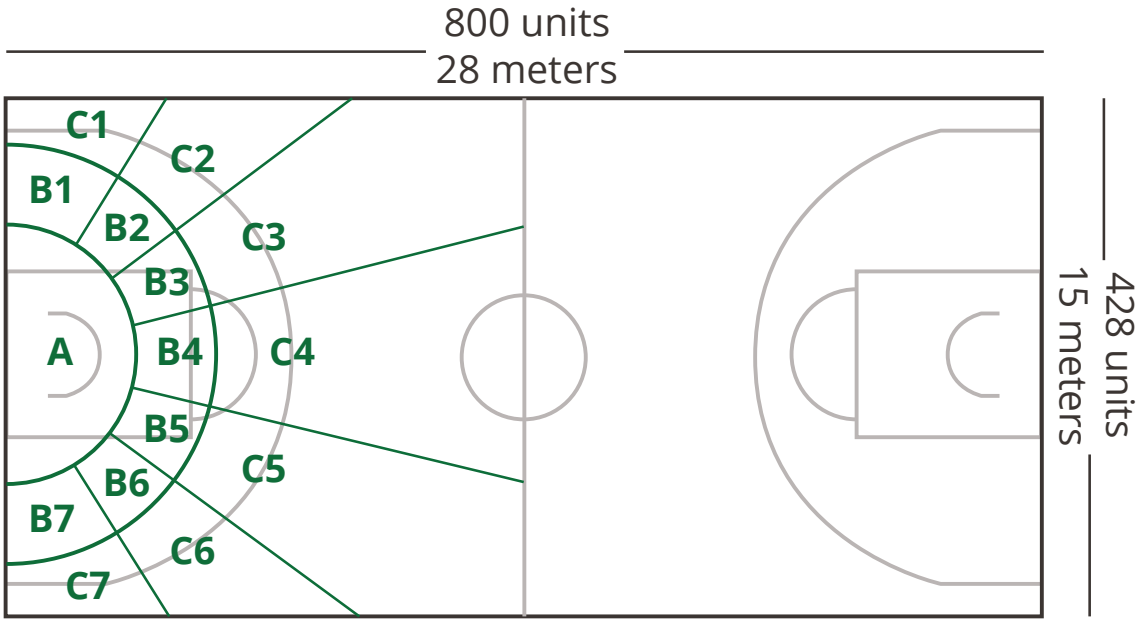

Figure 3.6: Radial grid

Shot Tension: We use score difference clusters in our predictive models to include the effect of score difference when the shot was taken. It shows the psychological tension of the players if their team is losing or relaxation if their team is winning. K-means clustering analysis is done in the software R on the score difference variable and a shot tension variable was created with this information (Appendix A-2). According to the results of our clustering analysis shot tension variable is identified with 7 levels which is presented in Table 3.4.

Table 3.4: Score Difference Clusters

Cluster	Score Difference
High Advantage	Leads more than or equal to 21 points
Moderate Advantage	Leads more than or equal to 11, less than 21 points
Low Advantage	Leads more than or equal to 3, less than 10 points
Balanced	Leads less than 2 points or back less than 5 points
Low Disadvantage	Back less than or equal to 13 points, more than 6 points
Moderate Disadvantage	Back less than or equal to 14 points, more than 23 points
High Disadvantage	Back more than or equal to 24 points

Shot Timing: To understand the behavioral effect of the timing on a success of a shot, three binary variables were introduced to the dataset. First one indicates if the shot was taken during the first five minutes, second one indicates if the shot was taken during the last five minutes, and the third one indicates if the shot was taken during the middle thirty minutes of the game. Overtime periods were included in the variable showing the last five minutes (Sampaio et al., 2010a; Sampaio et al., 2010b).

Home or Away Shot: This variable takes "H" if the shot was taken by the home team and "A" if the shot was taken by the away team.

Game Phase: The games are distinguished as "Regular Season", and "Postseason".

Player Season Index: To examine the effect of prior individual performance on a successful shot, player index rating variable was introduced to the data set. For 2007 - 2008 season, end-of-season general success rate was used for all players. For the other seasons, individual success rates of the previous season were used. If a player was excluded from the previous year, due to the minimum 100 shot limit, the general success rate for the previous year was used.

Successful Shot: The dependent binary variable which takes "1" if the shot is successful and "0" if the shot is unsuccessful.

3.3 Descriptive Analysis

One of the most analyzed features of basketball is the spatial success rate. For this purpose the radial zones described in the data source section are used. Spatial decisions are made according to offensive strategies of the teams as well as the defensive strategies of the opponents. Figure 3.7 (right) clearly shows that the shots with low angles to the hoop have lower success rates but players tend to attempt more from the lower angles as represented in Figure 3.7 (left). Figure 3.8 represents the spatial grid with "points per 100 attempts" based on the radial zones. According to the Euroleague organization's findings, the percentage of possessions derived from pick-and-roll increased from 23% in 2009 to 42.2% in 2018. Teams are generally starting pick-and-roll offenses from the low angle zones and try to find a match-up which the guard can be matched with a center. When the match-up does not work, or the center defends the guard effectively, then they try to find

shooters from the high angle zones. That's why the high angle zones have higher success rates.

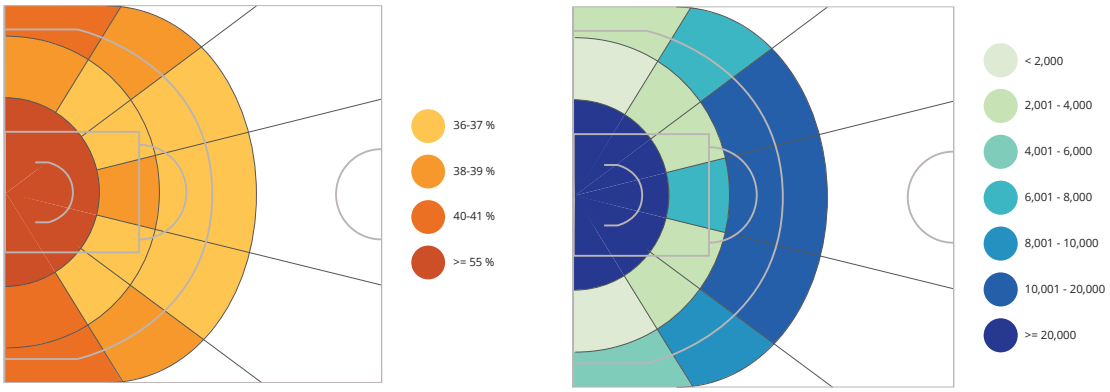

Figure 3.7: Spatial grid (Left: Shot success rates, Right: Total shot attempts)

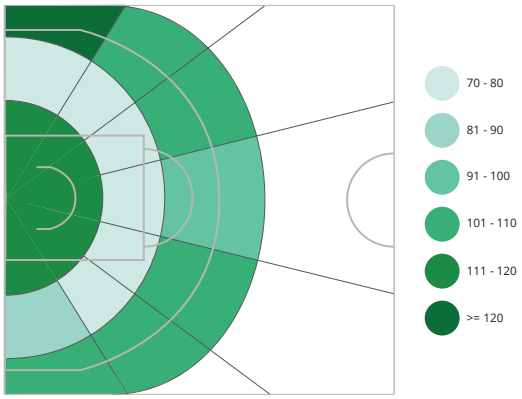

Figure 3.8: Spatial grid (Points per 100 shot attempt)

When the "Points per 100 attempts" information shown in Table 3.4 is detailed with "Home/Away" and "Shot Tension" variables, it can clearly be seen that having a score advantage increases the points for the home team (Figure 3.9). This shows the importance of coaching and psychological strength. If the coaches takes the time-outs according to this information, they will have the chance to motivate their teams and close the score difference.

Table 3.5 represents the “points per 100 attempts”, “success rate” and “total shot attempts” for each player position, i.e. guard, forward and center. Although it is very clear that the center players are more efficient, teams are using guards and forwards more. This is expected because the Euroleague is defensively a tough organization. Especially the “big guys” playing as centers are strong and have good defensive fundamentals. On the other hand, rise of the pick-and-roll offenses again create more chances for guards and forwards.

Table 3.5: Player position based shot attempt statistics

Player Position	Total Shot Attempts	Success Rate	Points Per 100 Attempts
Guard	81,372	43.0 %	102.6
Forward	63,573	44.8 %	103.1
Center	24,745	52.8 %	109.7

Figure 3.9 and Figure 3.10 represent the frequency for point per 100 points, team based and player based, respectively. Only one team in the Euroleague organization exceeds 110 points per 100 shot attempts in the data scope of this research. We run a Shapiro-Wilk test on R for points per 100 attempts data (AppendixA-3). For players, W = 0.9915 and p-value = 0.008303, and for teams W = 0.93878 and p-value = 0.01097. According to these values we can say that the points per 100 attempts data are normally distributed for the teams but not for the players with 95% confidence.

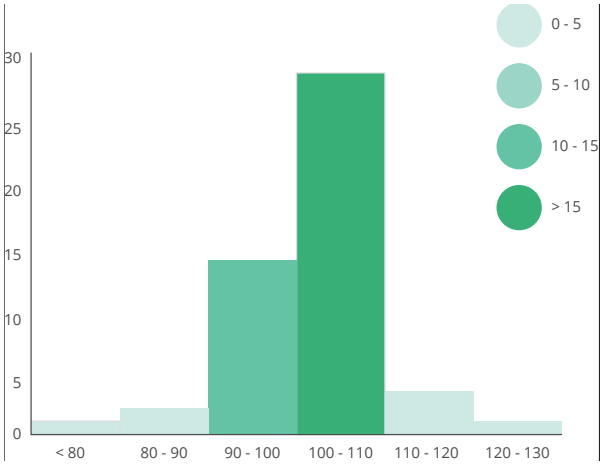

Figure 3.9: Points per 100 attempts for teams

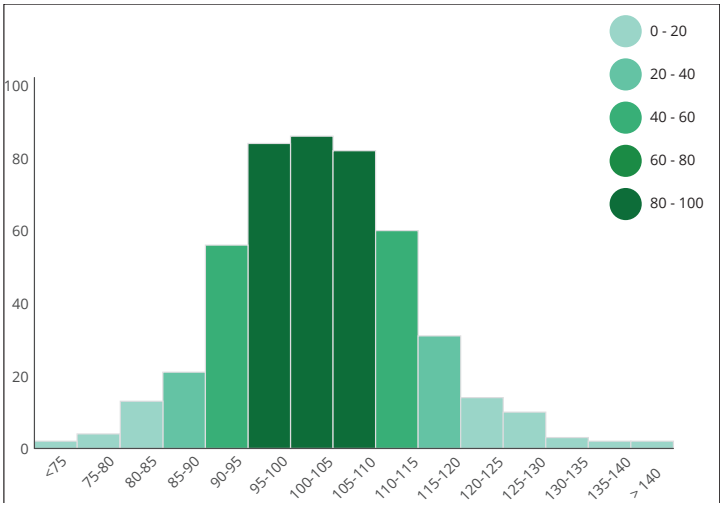

Figure 3.10: Point per 100 attempts for players

CHAPTER 4

ANALYSIS AND METHODS

In this chapter, the predictive models, their parameters, and results will be presented. Since we focus on the success probability of a basketball shot, the dependent variable for all models is a binary variable which makes our problem a binary classification problem that can be analyzed by supervised learning techniques. Since all data gathered from the Euroleague website is preprocessed before (as explained in the Chapter 3), only the necessary transformations that R libraries demand were included in this chapter. We randomly split the data into an 80% "train set" with 135,752 observations and a 20% "test set" with 33,938 observations (Appendix A-4). Both datasets have 45% success rate. We train models on the train set and predict on the test set. We apply logistic regression, random forest, Naive Bayes, support vector machines, and artificial neural networks algorithms and compare their accuracies. All analyses were carried out using R. At the end of this chapter the summary table for all models is presented. The independent variables and their data types are presented in Table 4.1.

Table 4.1: Independent variables details

Variable	Data Type	Data Type Detail
Player Season Index	Numerical	Normalized
Radial Zone	Categorical	15 Levels
Player Position	Categorical	3 Levels
Action	Categorical	3 Levels
Home / Away	Binary	2 Levels
First or Last Five Minutes	Categorical	3 Levels
Game Phase	Binary	2 Levels
Shot Tension	Categorical	7 Levels

The classification outcomes of the models will be presented as confusion matrices shown in Table 4.2. The “Positive Predictive Value” is described as “Σ True Positive / Σ Test Outcome Positive”. The “Negative Predictive Value” is described as “Σ True Negative / Σ Test Outcome Negative”.

Table 4.2: Confusion matrix

		Prediction	
		1	**0**
Actual	**1**	True Positive	False Negative
	0	False Positive	True Negative

4.1 Logistic Regression

The most widely used classification technique for binary classification problems since 19th century is the logistic regression. The probabilistic outcomes of the logistic regression range between 0 and 1 which the prediction is dependent using a threshold value, If the predicted success probability is above the threshold value than the shot is assigned as successful. In our analysis, we firstly use the threshold value of 0.50.

The “glm” (Generalized linear model) library was used for prediction on the test set and “confusionMatrix” function from the library “caret” developed by Kuhn et al. (2017) was used to create the confusion matrix which is presented in Table 4.3. The prediction on the test data has positive predictive value of 56.62 %, the negative predictive value of 66.61 % and the overall prediction accuracy value of 60.13 %. All the variables were statistically significant (Appendix A-5).

Table 4.3: Confusion matrix for logistic regression

		Prediction	
		1	**0**
Actual	**1**	7,944	7,443
	0	6,086	12,465

We also apply threshold values ranging from 0.5 to 0.7 with an increment of 0.01 to see the effect of the threshold value in logistic regression model with the test predictions. (Table 4.4). The results show that the test accuracy with threshold value of 0.50 is 60.13% and 54.66% with the threshold value 0.70.

Table 4.4: Accuracies for Different Threshold Values

Threshold Value	Test Accuracy	Threshold Value	Test Accuracy
0.50	0.60136	**0.61**	0.55165
0.51	0.60136	**0.62**	0.54897
0.52	0.60239	**0.63**	0.54770
0.53	0.60065	**0.64**	0.54729
0.54	0.60095	**0.65**	0.54700
0.55	0.59461	**0.66**	0.54664
0.56	0.58787	**0.67**	0.54667
0.57	0.57861	**0.68**	0.54661
0.58	0.56868	**0.69**	0.54661
0.59	0.56052	**0.70**	0.54661
0.60	0.55513		

Since the radial zones and player positions could have a correlation, the interaction term was introduced to the model and the results are shown in the Table 4.5. Although it is obvious that the centers are shooting mostly from closer ranges to the hoop, the interaction term did not change the accuracy level significantly. The prediction on the test data has a positive predictive value of 56.72 %, the negative predictive value of 62.57 % and the overall prediction accuracy value of 60.17 %.

Table 4.5: Confusion matrix for logistic regression including interaction term

		Prediction	
		1	**0**
Actual	**1**	7,892	7,495
	0	6,023	12,528

4.2 Random Forest

The random forest model is an ensemble learning method used for classification and regression. It creates multiple decision trees with randomly chosen variable subsets. The random forest model was run with the "randomForest" function from the R library "randomForest" which was developed by Liaw and Wiener (2002). We use number of trees starting from 500 to 3000 with 500 increments to see if this parameter will increase the accuracy. The results for random forest models with different tree sizes are presented in Table 4.6. According to this information, increasing the tree size of a random forest model, which needs more computational power and time, does not significantly affect the accuracy (Appendix A-6).

Table 4.6: Accuracies for Random Forest Models with Different Tree Sizes

Number of Trees	Test Accuracy (%)	Train Accuracy (%)
500	60.68	61.95
1,000	60.72	61.94
1,500	60.69	61.98
2,000	60.69	61.93
2,500	60.65	61.94
3,000	60.68	61.95

The results with the random forest model with 1,000 trees are detailed in Table 4.7. The positive predictive value is 48.46 %, the negative predictive value is 70.89 % and the overall prediction accuracy value is 60.72 %. The interesting outcome of the random forest model is that the negative predictive value is much higher than the positive predictive value.

Table 4.7: Confusion matrix for random forest with 1000 trees

		Prediction	
		1	**0**
Actual	**1**	7,456	7,931
	0	5,401	13,150

4.3 Naive Bayes Estimator

Naive bayes is a classification method from the "probabilistic classifiers" family. The algorithm is based on the prior probabilities of the outcomes. Different from the Bayes algorithm, Naive Bayes assumes that the independent variables are conditionally independent from each other. For running Naive Bayes on the software R, library "klaR" developed by Weihs et al. (2005) is used. The data was sampled as 80% train and 20% test splits. 10-fold cross validation was applied through the Naive Bayes classifier. The results are represented in Table 4.8. According to this table the positive predictive value is 51.92 %, the negative predictive value is 66.81 % and the overall prediction accuracy value is 60.06 % (Appendix A-7).

Table 4.8: Confusion matrix for Naive Bayes

		Prediction	
		1	**0**
Actual	**1**	7,988	7,399
	0	6,157	12,394

The common “mlr” extension was applied to the Naive Bayes model to examine if a linear extension of the model will improve the accuracy. The results were represented in Table 4.9. The positive predictive value is 51.92 %, the negative predictive value is 67.07 % and the overall prediction accuracy value is 60.20 % (Appendix A-8).

Table 4.9: Confusion matrix for Naive Bayes MLR extension

		Prediction	
		1	**0**
Actual	**1**	7,989	7,398
	0	6,108	12,443

4.4 Support Vector Machine (SVM)

For support vector machines model on the software R library “e1071” which was developed by Meyer et al. (2017). The data is sampled as 2/3 train and 1/3 test splits and 10-fold cross validation was applied. Linear kernel “Vanilladot” from the library “kernlab” which was developed by Karatzoglou (2004) was used. Linear kernels are used to find the largest possible linear margin that separates two regions. The results are represented in Table 4.10. According to the table the positive predictive value is 51.79 %, the negative predictive value is 67.10 % and the overall prediction accuracy value is 60.16 % (Appendix A-9).

Table 4.10: Confusion matrix for Support Vector Machine

		Prediction	
		1	**0**
Actual	**1**	7,969	7,418
	0	6,103	12,448

4.5 Artificial Neural Networks (Deep Learning)

Deep learning, which is a field of artificial neural networks, is one of the most common machine learning techniques applied in classification problems, robotics, and artificial intelligence. It uses “Multi Layer Perceptron” (MLP) to build the models. To create a neural network model in R, library “keras” developed by Allaire (2017) was used and the steps presented below were executed (Appendix A-10).

1 - The data was sampled as 80% training and 20% test

2 - Recipe object was created. Recipe objects are the data preprocessing instruments for the neural network models. The recipe object in our model checks if the data types provided for the model are appropriate or not. It also applies one-hot-encoding method to convert categorical variables into dummy variables to be able to use as a matrix. The library “recipes” is used to apply the recipes to the data set (Kuhn and Wickham, 2018).

3 - The data was baked with the recipe. Baking, in neural networks, is the process of applying the recipe objects to the datasets.

4 - The dependent variable data frame was converted into a vector.

5 - The neural network model was built with two hidden layers. On both layers dropout rate was 0.1 to prevent over-fitting. The first layer was created with 16 units, “uniform” kernel and “relu” activation. The second layer was created with 16 units, “uniform” kernel and “sigmoid” activation.

6 - The complier was set with “adam” optimizer and “binary cross-entropy” loss function.

7 - The model was fitted with a batch size 100, epochs 100 and validation split of 20%.

The confusion matrix for the artificial neural network model is presented in Table 4.11. According to this table, the positive predictive value is 57.49 %, the negative predictive value is 61.78 % and the overall prediction accuracy value is 60.20 %. Figure 4.1 represents the variable importance plot on which the positive correlations contribute to success and negative correlations prevent success. According to this, radial zone, shot type, player season index, and home or away team variables play a significant role in

predicting successful shots. Three point field goals prevent shots from being successful the most.

Table 4.11: Confusion matrix for Artificial Neural Network

		Prediction	
		1	**0**
Actual	**1**	7,211	8,176
	0	5,331	13,220

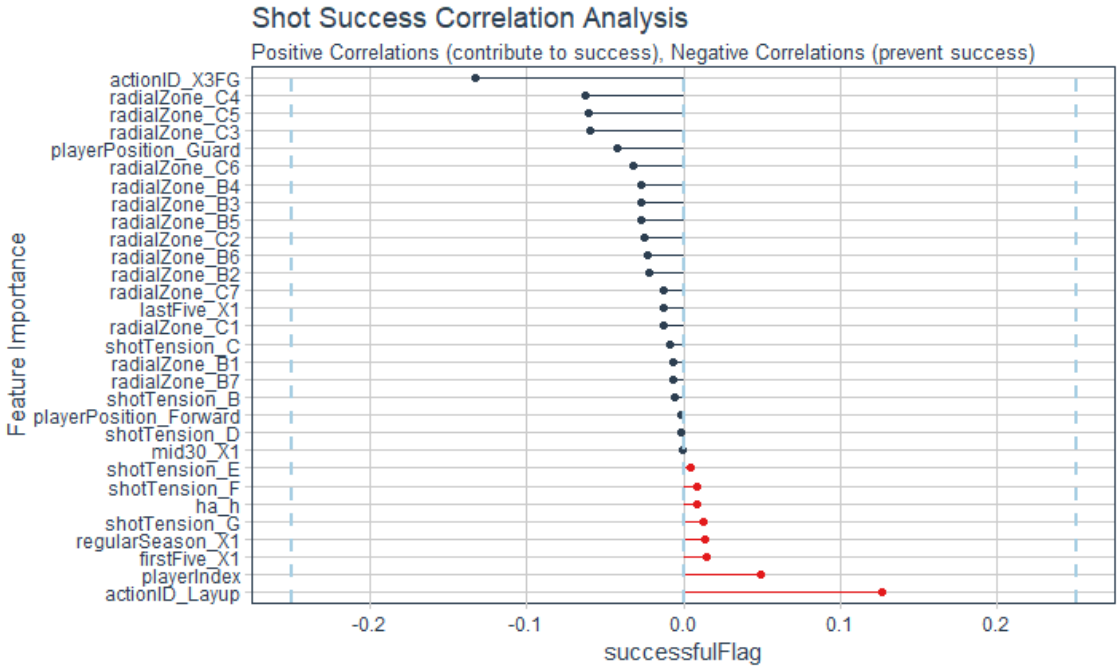

Figure 4.1: Neural network variable importance plot

4.5 Summary of Predictive Models

Chang et al. (2014). suggest that getting a model with predictability above 65% seems difficult, because the success rate ranges between 35% and 65% in the NBA. Our study showed that Euroleague success rate per game also ranges between 30% and 65%. In summary, these results also show that predicting a successful basketball shot with more than 65% accuracy is a challenging task and needs more sophisticated and detailed

data source. The summary accuracy information of the models used in the thesis is represented in Table 4.12.

Table 4.12: Predictive models summary table

Model	Positive Predictive Value (%)	Negative Predictive Value (%)	Overall Test Accuracy (%)	Overall Train Accuracy (%)
Logistic Regression	56.62	66.61	60.14	59.99
Logistic Regression Interaction Term	56.72	62.57	60.17	60.04
Random Forest	48.46	70.89	60.72	61.94
Naive Bayes	51.91	66.81	60.06	59.89
Naive Bayes (MLR Extension)	51.92	67.07	60.20	59.99
Support Vector Machines	51.79	67.10	60.16	59.99
Artificial Neural Networks	57.49	61.78	60.20	60.27

Table 4.12 is quite revealing in several ways. First, the overall accuracy range is very small. Furthermore more the random forest model has the highest accuracy among all models. There are also interesting differences on positive and negative predictive values among all models. While the random forest model performs the best with the negative predictive value, artificial neural networks model has the highest positive predictive value.

CHAPTER 5

CONCLUSION

Compared with the conventional statistical methods that have been used for many years, analytical approaches offer a wide range of analysis methods that help the decision makers in the field of basketball. The aim of the present research is to determine the significant aspects of basketball games and players which have an influence on the success of basketball shots. The second aim of this study is to investigate the effects of different models on prediction accuracy to evaluate the best approach for predicting successful basketball shots.

The relevance of the relationship between spatial-temporal aspects of shots and the success rate is clearly supported by the current findings. Shooting zones and game phases show a statistically significant effect on the shot success. The second major finding is that different predictive models used in this research did not show a significant difference on the accuracy of results.

The scope of this study was limited in terms of the collected data. The data used in this research consists of offense teams' shot attempts. More information on ball possessions, defense strategies, and more importantly, player tracking data with variables like nearest defense player, mean distances between players and ball velocity would help us to establish a greater degree of accuracy. Further work needs to be done to establish an on-line decision making system, which will collect the game data simultaneously and support coaches on the decisions for clutch shooting situations.

BIBLIOGRAPHY

Abdelkrim, N. B., Chaouachi, A., Chamari, K., Chtara, M., & Castagna, C. (2010). Positional Role and Competitive Level Differences in Elite Level Men's Basketball Players. The Journal of Strength & Conditioning Research, 24(5), 1346-1355.

Alamar, B., & Mehrotra, V. (2011). Beyond 'Moneyball': The Rapidly Evolving World of Sports Analytics, Part 2. Analytics Magazine.

Alamar, B. (2013). Sport analytics: A Guide for Coaches, Managers, and Other Decision Makers. New York, NY: Columbia University Press.

Arkes, J. (2010). Revisiting the Hot Hand Theory with Free Throw Data in a Multivariate Framework. Journal of Quantitative Analysis in Sports, 6 (1), Article 2.

Arkes, J., & Martinez, J. (2011). Finally, Evidence for a Momentum Effect in the NBA. Journal of Quantitative Analysis in Sports, 7(3).

Arthur, Rob, and Rian Watt. "How the Front-Office Analyst Took Over the NBA." Sports, Sports, 24 Oct. 2016, sports.vice.com/en_us/article/53xdbb/how-the-front-office-analyst-took-over-the-nba.

Baker R. E. and Kwartler T. (2015). Using Open Source Logistic Regression Software to Classify Upcoming Play Type in the NFL. Journal of Applied Sport Management, Vol. 7, No. 2, Summer 2015

Bradley R.A., Terry M.E. (1952). Rank Analysis of Incomplete Block Designs I: The method of paired comparisons. Biometrika, 39, 324–45.

Chang, Y. H., Maheswaran, R., Su, J., Kwok, S., Levy, T., Wexler, A., & Squire, K. (2014, February). Quantifying Shot Quality in the NBA. In Proceedings of the 8th Annual MIT Sloan Sports Analytics Conference. MIT, Boston, MA.

Cokins G., DeGrange W., Chambal S. and Walker R (2016) https://www.informs.org/ORMS-Today/Public-Articles/June-Volume-43-Number-3/Sports-analytics-taxonomy-V1.0

Eurloeague Bylaws (2018). http://www.euroleague.net/rs/89rc6t9mc63fbgq-t/84bd1f8d-134d-42a0-a8ee-cd688d29aaa2/562/filename/201718taebylaws.pdf

FIBA Official Basketball Rules (2017). http://www.fiba.basketball/OBR2017/Final.pdf

García, J., Ibáñez, S. J., De Santos, R. M., Leite, N., & Sampaio, J. (2013). Identifying Basketball Performance Indicators in Regular Season and Playoff Games. Journal of human kinetics, 36(1), 161-168.

Gómez, M. A., Lorenzo, A., Ibañez, S. J., & Sampaio, J. (2013). Ball Possession Effectiveness in Men's and Women's Elite Basketball According to Situational Variables in Different Game Periods. Journal of sports sciences, 31(14), 1578-1587.

Karatzoglou A., Smola A., Hornik K., Zeileis A. (2004). kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software 11(9), 1-20. URL http://www.jstatsoft.org/v11/i09/

Kuhn M. Contributions from Wing J., Weston S., Williams A., Keefer C., Engelhardt A., Cooper T., Mayer Z., Kenkel B., the R Core Team, Benesty M., Lescarbeau R., Ziem A., Scrucca L., Tang Y., Candan C. and Hunt T. (2017). caret: Classification and Regression Training. R package version 6.0-78. https://CRAN.R-project.org/package=caret

Kuhn M. and Wickham H. (2018). recipes: Preprocessing Tools to Create Design Matrices. R package version 0.1.3. https://CRAN.R-project.org/package=recipes

Lang D. T. and the CRAN Team (2017). XML: Tools for Parsing and Generating XML Within R and S-Plus. R package version 3.98-1.9. https://CRAN.R-project.org/package=XML

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big Data, Analytics and the Path from Insights to Value. MIT sloan management review, 52(2), 21.

Liaw A. and Wiener M. (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Marcelino, R., Mesquita, I., & Sampaio, J. (2011). Effects of Quality of Opposition and Match Status on Technical and Tactical Performances in Elite Volleyball. Journal of Sport Sciences, 29, 733–741.

McFarlane, P. Evaluating NBA End-of-Game Decision Making. Journal of Sports Analytics, (Preprint), 1-6.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. and Leisch, F. (2017). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-8. https://CRAN.R-project.org/package=e1071

Morris, D. Z. "The NFL vs. The NBA: Which Will Be America's Biggest Sport 10 Years From Now?" Fortune, Fortune, 26 May 2018. Web. 31 July 2018. <fortune.com>

Ooms J. (2014). The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. arXiv:1403.2805 [stat.CO] URL https://arxiv.org/abs/1403.2805.

Ostojic, S.M., Mazic, S. & Dikic, N. ()2006). Profiling in Basketball: Physical and Physiological Characteristics of Elite Players. Journal of Strength and Conditioning Research, 2006, 20(4), 740-744

Reich, B. J., Hodges, J. S., Carlin, B. P., & Reich, A. M. (2006). A Spatial Analysis of Basketball Shot Chart Data. The American Statistician, 60(1), 3-12.

Sampaio, J., Lago, C., Casais, L., & Leite, N. (2010a). Effects of Starting Score Line, Game Location and Quality of Opposition in Basketball Quarter Score. European Journal of Sport Sciences, 10, 391–396.

Sampaio, J., Lago, C., & Drinkwater, E. J. (2010b). Explanations for the United States of America's Dominance in Basketball at the Beijing Olympic Games (2008). Journal of Sports Sciences, 28, 147–152.

Shortridge, A., Goldsberry, K., & Adams, M. (2014). Creating Space to Shoot: Quantifying Spatial Relative Field Goal Efficiency in Basketball. Journal of Quantitative Analysis in Sports, 10(3), 303-313.

Weihs, C., Ligges, U., Luebke, K. and Raabe, N. (2005). klaR Analyzing German Business Cycles. In Baier, D., Decker, R. and Schmidt-Thieme, L. (eds.). Data Analysis and Decision Support, 335-343, Springer-Verlag, Berlin.

Wickham H. (2017). httr: Tools for Working with URLs and HTTP. R Package Version 1.3.1. https://CRAN.R-project.org/package=httr

Wright, R. E., Silva, J., & Kaynar-Kabul, I. (2016). Shot Recommender System for NBA Coaches. In KDD Workshop on Large-Scale Sports Analytics.

APPENDIX A

R CODES

1 - K-means clustering for creating radial shotting zones

Before running k-means clustering algorithm, shot distances were normalized. This process is followed by determining number of clusters. We use three-cluster solution for shooting zones in our analysis.

```
clusteringDistance$scaledDistance = scale(clusteringDistance$virtualDistance)

wss = (nrow(clusteringDistance)-1) * sum(apply(clusteringDistance$scaledDistance,2,-
var))

for (i in 1:15) wss[i] = sum(kmeans(clusteringDistance$scaledDistance,
                             centers=i)$withinss)

plot(1:15, wss, type="b", xlab="Number of Clusters",
    ylab="Within groups sum of squares")

fit = kmeans(clusteringDistance$scaledDistance, 3)

aggregate(clusteringDistance$scaledDistance,by=list(fit$cluster),FUN=mean)

clusteringDistance = data.frame(clusteringDistance, fit$cluster)
```

2 - k-means Clustering for Creating Shot Tension Variable Based on Score Difference

We use seven-cluster solution for score difference variable.

```
clusteringScoreDif$scaledScoreDifference = scale(clusteringScoreDif$scaledScoreDifference)

wss = (nrow(clusteringScoreDif)-1)*sum(apply(clusteringScoreDif$scaledScoreDifference,2,var))

for (i in 2:15) wss[i] = sum(kmeans(clusteringScoreDif$scaledScoreDifference,
                                    centers=i)$withinss)

plot(1:15, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares")

fit = kmeans(clusteringDistance$scaledDistance, 7)

aggregate(clusteringDistance$scaledDistance,by=list(fit$cluster),FUN=mean)

clusteringDistance = data.frame(clusteringDistance, fit$cluster)
```

3 - Shapiro-Wild Test for Determining Normal Distribution of "Points per 100 Attempts"

```
shapiro.test(PP100$teamsPP100)

shapiro.test(PP100$playersPP100)
```

4 - Train and Test Data Splitting

```
set.seed(100)

smp_size = floor(0.80 * nrow(data))

index = sample(seq_len(nrow(data)),size=smp_size)

train = data[index,]

test = data[-index,]
```

5 - Logistic Regression

```
library(caret)
fpr = NULL
fnr = NULL
acc = NULL
model = glm(successfulFlag ~
              playerPosition + actionID + ha + regularSeason + radialZone + firstFive
            + lastFive + mid30 + shotTension + playerIndex + radialZone * playerPosition
            ,family=binomial,data=train)
results_prob = predict(model,subset(test,select=c(2:11)),type='response')
results = ifelse(results_prob > 0.5 ,1,0)
answers = test$successfulFlag
misClasificError = mean(answers != results)
acc = 1-misClasificError
cm = confusionMatrix(data=results, reference=answers)
```

6 - Random Forest

```
library(randomForest)
rf_model = randomForest(successfulFlag ~ playerPosition + actionID + ha
                + regularSeason + radialZone + firstFive + lastFive + mid30
                +shotTension + playerIndex , data = train, importance = TRUE, ntree=1000)
rf_pred = predict(rf_model, test)
table(observed = test$successfulFlag, predicted = rf_pred)
```

7 - Naive Bayes

library(ElemStatLearn)

library(klaR)

library(caret)

xTrain = train[,-1] # removing y-outcome variable.

yTrain = as.factor(train$successfulFlag) # only y.

xTest = test[,-1]

yTest = as.factor(test$successfulFlag)

model = train(xTrain,yTrain,'nb',trControl=trainControl(method='cv',number=10))

(table(predict(model$finalModel,xTest)$class,yTest))

8 - Naive Bayes MLR Extension

task = makeClassifTask(data = data, target = "successfulFlag")

selected_model = makeLearner("classif.naiveBayes")

NB_mlr = train(selected_model, task)

NB_mlr$learner.model

predictions_mlr = as.data.frame(predict(NB_mlr, newdata = data[,2:8]))

table(predictions_mlr[,1],data$successfulFlag)

9 - Support Vector Machines

library(kernlab)

letter_classifier = ksvm(as.factor(successfulFlag) ~ ., data = train, kernel = "vanilladot", cross = 10)

svm_pred = predict(letter_classifier, test[,-1])

table(svm_pred, test$successfulFlag)

10 - Artificial Neural Networks

```
library(keras)

library(recipes)

rec_obj = recipe(successfulFlag ~ ., data = train) %>%

  step_dummy(all_nominal(), - all_outcomes()) %>%

  step_center(all_predictors(), - all_outcomes()) %>%

  step_scale(all_predictors(), -all_outcomes()) %>%

  prep(data = train)

x_train = bake(rec_obj, newdata = train %>% select(-successfulFlag))

x_test = bake(rec_obj, newdata = test %>% select(-successfulFlag))

y_train = ifelse(pull(train,successfulFlag) == 1,1,0)

y_test = ifelse(pull(test,successfulFlag) == 1,1,0)

model_keras = keras_model_sequential()


model_keras %>%

layer_dense(units = 16, kernel_initializer = "uniform", activation = "relu",

input_shape = ncol(x_train)) %>%

layer_dropout(rate = 0.1) %>%


layer_dense(units = 16, kernel_initializer = "uniform", activation = "sigmoid") %>%

layer_dropout(rate = 0.1) %>%

layer_dense(units = 1, kernel_initializer = "uniform", activation = "sigmoid") %>%

compile(optimizer = 'adam', loss = 'binary_crossentropy', metrics = 'accuracy')
```