A FEATURE ENGINEERING APPROACH TO PREDICTING PLAYER PERFORMANCE IN BASKETBALL

FEYZULLAH ALİM KALYONCU

BOĞAZİÇİ UNIVERSITY

2019

A FEATURE ENGINEERING APPROACH TO PREDICTING PLAYER PERFORMANCE IN BASKETBALL

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Master of Arts

in

Management

by

Feyzullah Alim Kalyoncu

Boğaziçi University

2019

DECLARATION OF ORIGINALITY

I, Feyzullah Alim Kalyoncu, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature.....

Date .....28.06.2019.....

ABSTRACT

A Feature Engineering Approach to Predicting

Player Performance in Basketball

Recent advancements in sports analytics have found many fields of applications in basketball. Player performance prediction is one of the main goals of basketball analytics because of the potential implications for both teams and fans. This study aims to create a predictive modeling approach that is designed for accurately estimating the performances of basketball players while addressing the main issues in basketball statistics. Euroleague, the highest level of European basketball club competition, is selected to conduct our study. The data set used in this study contains 720 regular- season games from 2016-2017, 2017-2018, 2018-2019 Euroleague seasons. During these seasons, a total of 15368 records obtained from performances of 464 individual athletes. In order to create models for predicting performances of basketball players, we followed a structured data mining process. Most predictive models in the literature have relied on offensive statistics because of the scarcity of statistics that are related to defense. However, this study addresses the need for defensive metrics in player performance prediction, so far lacking in the literature. We developed a methodology and proposed a feature engineering approach to create data-driven defensive metrics. Our results demonstrate that the most significant boost in both R-squared and rmse values have been achieved after adding position-based defensive metrics.

ÖZET

Basketbol Oyuncu Performansı Tahminlemesi İçin Bir Özellik Mühendisliği Yaklaşımı

Spor analitiğindeki son gelişmeler basketbolda birçok uygulama alanı bulmuştur. Oyuncu performansı tahminlemesi, basketbol analitiğinin temel hedeflerinden biridir çünkü hem takımlar hem de taraftarlar için çeşitli potansiyel faydaları vardır. Bu çalışmanın amacı, basketbol istatistiklerinde yer alan temel problemleri göz önünde bulunduran ve basketbol oyuncularının performanslarını doğru bir şekilde tahminleyen öngörücü bir modelleme yaklaşımı oluşturmaktır. Araştırmamızda Avrupa basketbolunun kulüp düzeyindeki en üst seviyesi olan Euroleague verisi kullanıldı. Çalışmada kullanılan bu veriler 2016-2017, 2017-2018, 2018-2019 Euroleague sezonlarında oynanan toplam 720 normal sezon maçını içermektedir ve bu sezonlarda 464 bireysel sporcunun performansından toplam 15368 kayıt elde edilmiştir. Çalışmamızda basketbol oyuncularının performansını öngörebilecek modeller oluşturmak için özel olarak yapılandırılmış bir veri madenciliği sürecini takip ettik. Literatürdeki çoğu öngörücü model, savunma ile ilgili istatistiklerin azlığı nedeniyle ofansif istatistiklere dayanmaktadır. Ancak, bu çalışma literatürde şu ana kadar eksik olan oyuncu performans tahmininde savunma metriklerine duyulan ihtiyacı da ele almaktadır. Bu bağlamda, veriye dayalı savunma metrikleri oluşturmak için yeni bir metodoloji geliştirdik ve bir özellik mühendisliği yaklaşımı önerdik. Sonuçlarımız, hem R-squared hem de rmse değerlerinde en önemli geliştirmenin pozisyon bazlı savunma metrikleri ekledikten sonra elde edildiğini göstermektedir.

ACKNOWLEDGMENTS

I would like to show my greatest appreciation to my thesis advisor Assist. Prof. Hüseyin Sami Karaca whose enormous support and insightful comments were invaluable during the writing of my thesis.

Besides my advisor, I would like to thank the committee members Prof. Gökhan Özertan and Assist. Prof. S. Mehmet Özsoy for accepting to take part in my jury.

Additionally, the financial support of the Scientific and Technological Research Council of Turkey (TÜBİTAK) BİDEB through the 2210-A Master's Scholarship Program is also gratefully acknowledged.

Finally, my sincere thanks goes to my fellows Berkay Bulut and Ahmet Yıldırım for their continuous encouragement throughout the development of this thesis. I would also like to express my gratitude to my parents, my sister and my brother for their moral support and warm encouragements. They have always trusted me and supported me. I could not have written this thesis without their love, patience, and support.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The use of historical game records and their combination with game-related information can help teams and players to achieve better performance. Prior to developments in data mining, sports institutions were heavily dependent on the human experience of coaches, managers, and players. These human experts were believed to transform the historical game records of their teams into valuable information. However, as the complexity of data collected increased over time, sports institutions looked for more convenient methods to process the data they already had (Cao, 2012).

According to a definition provided by Alamar and Mehrotra (2011), sports analytics is "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" (p. 33). The adoption of analytics in sports varies considerably by kind. The applications and research of sports analytics focus primarily on competitive team sports because of the availability of data and the potential monetary return on investment.

Basketball is one of the most popular team sports in the world. Analytics is not particularly new to basketball and it has been used for many years. There are various potential analytics applications for the sport of basketball, but the problem of predicting performances of the basketball players has received substantial interest

from both statisticians and researchers. Performance prediction in basketball has two main practical areas depending on the user of the predictions.

Almost every year, the salary budgets of basketball teams are increasing in parallel with the people's interest in the sport of basketball and teams are spending more and more money to their players. It could be said that teams risk their future to the performances of the players they select. The primary motivation behind the need for predicting player performances in terms of teams is that decision-makers such as coaches and general managers shape their decisions according to the potential performances of players they consider. Some of these decisions include recruiting and scouting for new players and in-game winning strategy designs. For example, coaches can use player performance predictions to optimize their team rosters during the game. Moreover, general managers can decide on which players to transfer in/out according to player performance predictions.

Other than coaches and general managers, basketball fans are also potential users of the player performance predictions mostly because of the fantasy sports industry and its growth trend in different branches of sport such as basketball, baseball, and football. Today, most of the professional basketball leagues, including the NBA and Euroleague have their online fantasy basketball game which helps these leagues to increase fan engagement. While playing a fantasy basketball game to compete with friends and family, fans can create virtual teams of real players. The performances of these real players in actual games are used in the fantasy game by transforming the actual statistics into fantasy points. In this respect, fans who want to play fantasy basketball and win their leagues might have a competitive advantage by accurately predicting player performances. The first signs of the emerging need for player performance predictions are the websites that are dedicated to giving the latest

information and projections about player performances in the upcoming games. Additionally, there are also bet- ting opportunities on player performances in basketball. Over/Under betting odds on basic statistics including points scored, assists, and rebounds are offered by betting markets. People can bet on individual player statistics and make money if they can predict the performances of the players correctly.

Although predicting player performance in basketball has many implications, a very limited number of approaches have been made in the literature. The studies that focus on player performance prediction have failed to include most of the factors that have an influence on the performance of players. A new approach is, therefore, needed for the problem of player performance prediction. In this study, we propose a feature engineering approach to predict player performance in basketball.

CHAPTER 2

LITERATURE REVIEW

In today's world, a substantial amount of statistical information is generated about every team, player, and game. The first users of available data were coaches, team managers, and statisticians. They have realized that there exists a vast potential for their team if they can extract meaningful insights from the data they collect. In this sense, they have become more encouraged to invest in analytics. Researchers also paid attention to this emerging opportunity in sports analytics. There are academic conferences and journals that are solely devoted to sports analytics such as MIT Sloan Sports Analytics Conference and Journal of Sports Analytics. The sport of basketball was one of the first areas for conducting academic research because of the available information online and the easily quantifiable nature of basketball.

In this chapter, the existing literature that paves the way for this study is reviewed. The chapter starts with the literature on the prediction of game results which was the pioneering topic that leads to research conducted on player performance prediction. Afterward, the section continues with the main problems with basketball statistics and player positions. Addressing these problems that exist in player performance prediction is the foundation for this study.

2.1 Predicting game results

The prediction of game results has become widely popular among sports fans around the world, in particular, football and basketball fans (Haghighat et al., 2013). The primary motivation behind the popularization of game result prediction was gam-

bling opportunities. Different probabilistic models were proposed for the prediction of the results of football matches. In the literature, there exists a line of research that only focuses on market efficiency. Feng et al. (2016) proposed a methodology for beating English Premier League betting odds by modeling game results as a Skellam distribution. On the other hand, a great number of existing studies in the broader literature have been using various features derived from game, team, and player information as an alternative way for predicting game outcomes.

The literature review shows that many researchers have focused on trying different machine learning algorithms to create a system that predicts basketball game results. Even though Zdravevski and Kulakov (2010) included many meaningful features such as "number of injured players", "winning streak", and "fatigue" on two seasons of NBA data, they limited their modeling stage with algorithms in WEKA tool. Besides, Miljković et al. (2010) also created various features from basketball statistics data collected on 2009-2010 NBA and applied Bayes method and multivariate linear regression to predict game outcomes. However, the accuracy of their models is low compared due to the simplicity of the algorithms used.

In the last few years, due to the exponential growth in computing power, researchers are able to apply neural networks to many real-life problems. Loeffelholz et al. (2009) investigated the use of the neural networks as a tool to estimate the success of basketball teams in the NBA. They have collected data for 650 NBA games and constructed four different neural networks, including radial basis, feed-forward, probabilistic and generalized regression neural networks. After training created neural networks with 620 games, Loeffelholz et al. (2009) tested their performance on the 30 game validation set. It was reported that neural networks are

able to beat previous models and betting experts in their experiment. However, it is stated that only the main statistics are used in the neural networks and the overall accuracy of the models was 74.33%.

Although a range of data mining techniques, including neural networks, decision trees, Bayes method, logistic regression and support vector machines have been used to predict the game results, the current research still remains insufficient in terms of accuracy (Haghighat et al., 2013). Lack of a comprehensive set of statistics is stated as one of the main reasons for the low forecast accuracy. It was also reported in the literature that including features related to player performance may help to obtain more accurate predictions. Wheeler (2012) emphasized that predicting player performances and summing them can be used in the models that are designed to predict game outcomes. Therefore, the application of machine learning algorithms to predict each athlete's performance also naturally extends to predict game results. This chapter has evaluated the methods used in-game result prediction and demonstrated its limitations. The chapter that follows moves on to consider the player performance prediction in basketball.

2.2 Predicting player performances

Predicting the performance of their players is essential for the success of professional basketball teams. Coaches, managers, and scouts consistently evaluate the performance of the players in order to use in both short-term and long-term purposes. Examples of short-term purposes include pre-game roster formation decisions, in-game player substitution decisions, and in-game tactic design decisions. On the other hand, the decisions about transferring new players and renewing contracts with

existing players are the primary examples of long-term purposes. The sport of basketball is one of the leading sports that coaches and general managers can easily interfere with the team using their short-term and long-term decisions. An extreme example of this fact can be stated as there are only three substitutions allowed during a football game, whereas basketball coaches have an unlimited chance for substituting their players. In the literature, both short-term and long-term player performance prediction in basketball has been assessed to some extent.

Before designing prediction models on basketball players' performances, researchers mainly focused on the factors that have an effect on performance. The initial findings have shown that aging (Berri et al., 2006) and home court advantage (Arkes and Martinez, 2011) influence the performance of basketball players. The literature review also shows that most of the time, playing at home court increases the performance (Hwang, 2012). Additionally, after some point, aging has a negative effect on overall performance because athletic ability slowly decreases with age (Hwang, 2012).

There have been numerous studies to investigate the effects of main basketball statistics (minutes, points, rebounds, assists, steals, blocks, turnovers and shot attempts) on player performance. Casals and Martinez (2013) tried to understand and determine variables that are affecting performance. They have created a statistical model to study relative contributions of different variables on the performance. In their analysis, performance is quantified as the variability in points scored and win share by different players. The study was conducted using the main basketball statistics data of 27 NBA players during a single 82 game regular season. LMM and GLMM models were applied in order to predict points scored and win share. In this study, momentum effect which is defined as positive or negative trends

in previous game results may have a positive or negative effect on the outcome of a subsequent game for that player were also included by creating features from the last five games. The final findings of this study reported that the performance of players mostly affected by minutes played, the usage rate and the quality difference between teams. Although the work of Casals and Martinez (2013) is comprehensive and explanatory, we argue that the predictive accuracy of this study suffers from certain weaknesses in terms of feature selection and model generation. Applicability of this study is also limited because of the usage of a small sample of players with only 27 from the NBA.

In the literature, there are also studies that focus on data sets from the leagues other than the NBA. For instance, Sindik (2015) collected a data set that was deliberately sampled from the top Croatian basketball players (47 Guards, 27 Centers/Forwards) who played in nine different professional teams. This study aimed to identify the differences in the performance of players by looking at several independent variables for top male basketball players. Position in the team, total situation related efficiency, age, experience, and the playing time were the independent variables used in this work. Multivariate analysis of variance (MANOVA) was the method utilized in this study for assessing the performance of players. The findings showed that the most critical factors affecting performance were positions, overall situation efficiency and total time spent on the court in a game. However, as asserted by Sindik (2015) the significance of these elements varies from team to team. Limitations of this work can be stated as the Croatian Men’s Basketball League is not an elite league in Europe and only 74 players were included in the study.

Hwang (2012) raised an objection on the usage of end of season advanced basket- ball metrics invented by Oliver (2004) in basketball players' performance evaluation. It is stated that it would be valuable to assess players' value to the team by projecting future performance since NBA franchises are taking risks for individual players in many ways. Examples of these risks include financial commitments to players, salary cap restrictions, overall franchise value, but also team marketability. In this sense, projections on multiple derived metrics have been conducted. However, it was reported that predicting points scored in the next years could be a simple and valuable metric for assessing future performances for a player. The utilization of the Weibull-Gamma model is important for the success of this study because the Weibull-Gamma model can handle the time-dependent nature of player performance by making it possible to establish a statistical prediction of how a player will perform during the next years based on the trends after they first entered the NBA. The main purpose of the development of this model is stated as predicting future performances of the players to estimate contract value and aging effect accurately. In the test set of this experiment, only seven free agent players from the 2010 NBA season are evaluated and career projections for these players are provided. Even if this study brings a different perspective in terms of modeling, it is still limited because the conceptual framework of this study is based on the evaluation of NBA free agents.

The problem of predicting player performances becomes more challenging and complex when we try to predict the performances of newcomers to a basketball league. Based on the 2018-2019 season-opening night rosters, 108 players from 42 countries which constitutes 24.5% of all players in the entire league are in the roster of different NBA teams. Before international players entered the NBA, teams could

only focus on scouting and assessing amateur college basketball players in the United States. Afterward, general managers in the NBA have been challenged to evaluate qualities and to predict future performances of international players. In recent literature, only a limited number of authors have focused on predicting performances' of newcomers in a league. Salador (2011) conducted research to address this issue by trying to forecast the performances of international players that are coming to the NBA for the first time. To measure the performance of international players in the NBA, Salador (2011) used player efficiency rating (PER). Correlations between the statistics collected from NBA and international leagues are inspected to see which basketball skills are reflected in the NBA game. The main basketball statistics that are used to quantify success in the NBA were height, weight, field goals attempted (FGA), field goals made (FGM), field goal percentage (FG%), three-point field goals made (3PM), three-point field goals at- tempted (3PA), three-point field goal percentage (3P %), free throws made(FTM), free throws attempted (FTA), free throw percentage (FT %), years played, games played, games started, minutes played, rebounds, assists, steals, and points scored. On the other hand, nationality, position, draft rankings, years played in the international stage, and the number of tournaments that each player attended was an additional variable about international players. It has been found that shooting percentage and per-game statistics on assists, rebounds, blocks, steals are positively correlated factors meaning that these features of the international players are translated to NBA. Even though regression-based models in this analysis have low accuracy on predicting inter-national players' performances in NBA, the research by Salador (2011) is one of the pioneering works that handles one of the pitfalls of previous studies.

Basketball is a highly competitive team sport where two teams that consist of five players play on a court. Individual players can not be thought of as a single unit apart from their teammates. Therefore, when evaluating and predicting the performances of the players, team effects should also be considered. It is reported in the literature that the main barrier for analysts and researchers to predict player performances in basketball is to quantify the interaction effects arising from teamwork (Piette et al., 2011). Instead of traditional regression methods for prediction player performance, Piette et al. (2011) applied a network analysis technique to understand the importance of players relative to other teammates and the ability of players to perform their role. Latent Pathway Identification Analysis methodology is used in the network designed with players as nodes and interaction between players in the same team as edges. Centrality scores are calculated for each player in the network to find out individual performances. This study establishes a quantitative framework and shows that it is possible to deduct team effects from player performances to find out the purified effects of individual athletes.

Although different studies have been conducted by many authors, player performance prediction in basketball is still insufficiently explored. The literature review shows that trying to predict player performance still has significance and possible future applications. Prediction models in most of the previous works use only currently available metrics and statistics. In the following two chapters, the main problems related to current statistics are examined.

2.3 The deficit in basketball statistics

The most common statistics in various kinds of sports such as football, hockey, and basketball are divided into two main categories: offensive statistics and defensive statistics. Many different metrics are defined under these categories and these metrics are used in player performance evaluation (Brown, 2017). It was highlighted in the literature that the only way to evaluate players accurately is to extract all the available information properly (Franks et al., 2016).

Generally speaking, any person making management or coaching has a wide range of metrics at their fingertips, but sometimes it can be challenging to find the right metrics to support their decisions. Franks et al. (2016) inspected the basketball and hockey metrics and brought many questions regarding the uniqueness and reliability of the most common metrics used at that time. First, they checked whether metrics properly differentiate between players. Second, they tried to assess the stability of metrics over time. Lastly, they measured the information gain from different metrics. The findings of this study show that there is an undeniable redundancy across basketball metrics. It was also highlighted that defensive metrics carry more information about players compared to offensive metrics.

Although, defensive metrics are more valuable for differentiating players, when we look at a simple stat sheet from a basketball game, we can see many indicators that are related to the offense; however, only steals, blocks and rebounds give some information about defense. Unfortunately, until now, the vast majority of the analytics of the basketball tend to analyze offensive performance and almost entirely neglect the performance at defensive end (Franks et al., 2015). On the offensive side, most of the statisticians, analysts, and even fans have an opinion

about who the top performers are, whereas, on the defensive end, the definition of an elite defender is still debatable (Safir, 2015).

Statisticians and analytics experts have a constant motivation to evaluate and improve current metrics used for the sport of basketball. In the literature, some researchers have tried to come up with new metrics to evaluate player performance. For instance, Brown (2017) used Google's PageRank algorithm and play-by-play data to give a rating as a single variable that contains information about both offensive and defensive performance of a player for a game. The main purpose for creating this variable was including game difficulty to current metrics so as to have a comparable variable be- tween different games. Although new metrics similar to this example provide valuable insights for the game, they still lack information solely on defensive performance.

Basketball games are continually changing in time and space as players interact regularly with their teammates, their opponents and even with the ball. However, current basketball statistics have a low level of resolution because aggregate statistics can not capture high-resolution motifs that characterize basketball strategy (Cervone et al., 2016). Thanks to advancements in technology, every event in the sport of basketball is becoming more and more measurable. In recent years, player tracking technology which collects data with six cameras that records the coordinates of the ball and all players on the court is installed in all arenas of the NBA. Using software that utilizes these coordinates and movements, NBA teams are provided with more advanced and detailed statistics that they can not access previously. In recent studies, researchers have also realized the significance of player tracking data. Franks et al. (2015) presented a new set of metrics designed to enhance defensive measurement in advance basketball analytics. First, they utilized player

tracking data and generated a model that estimates defensive matchups for every single position. These matchup estimations enabled them to process who is responsible for the points allowed. Five different defensive metrics were defined in this study:

1. “Volume Score” to measure the total attempts on an individual defender
2. “Disruption Score” to quantify the efficiency of defending players on their opponents
3. “Defensive Shot Chart” to visually show the coordinates of shots being taken on the defender
4. “Shots Against” to measure shots attempted on the defender per 100 possessions
5. “Counterpoints” to measure points scored on a defender per 100 possessions

All of these metrics are applicable in the sport of basketball and they give extremely insightful information about the defensive performances of players. The work of Franks et al. (2015) took the role of initiator and showed that it is possible to create enhanced defensive metrics using scientific approaches.

2.4 The problem with player positions

Traditionally, there are five player positions defined for basketball: Point Guard, Shooting Guard, Small Forward, Power Forward, and Center. These positions somewhat describe the role of basketball players on the court. In the first place, the physical size of the players determines these positions. If a player has a small body and quickness, he is considered as a guard. On the other hand, if a player is bigger

and stronger compared to other players, he can be either a center or forward. In recent years, both academicians and practitioners criticized the current assumptions regarding player positions.

The decisions of coaches and general managers are impacted by the positions of players. For example, coaches tend to arrange in-game playing times according to these positions, and general managers acquire new players for their teams by looking at positions of the players. Lutz (2012) asserted that most of the time, players are assigned with positions in a non-scientific manner. The methodology of player positioning has two main issues: oversimplification and incorrect classification (Alagappan, 2012). To overcome these problems, researchers used data mining techniques and tried to learn the positions of the players from their data.

According to the similarity between data points, grouping data into subgroups is called as clustering in the scope of machine learning (Singh and Ahmad, 2015). To find out complex patterns in players' statistics and group them into distinct positions, the cluster analysis method is utilized in the literature. Alagappan (2012) proposed a methodology that uses a K-Means Clustering algorithm to explore different player positions that reflect unique playing styles. It is reported that minute-wise normalization of various statistics such as points scored, assists, and rebounds are required and data should be pre-processed. After data preparation, a total of 452 players from the NBA are clustered into separate groups. In this analysis, it is asserted that there can be 13 different positions that are deductible from player statistics. These positions are Offensive Ball Handler, Defensive Ball Handler, Combo Ball Handler, Shooting Ball Handler, Roll Playing Ball Handler, Three-Point Rebounder, Scoring Rebounder, Paint Protector, Scoring Paint Protector, Role Player, NBA 1st Team, NBA 2nd Team, and one-of-a-kind. The positions defined in

this study reflect different playing styles and handle the oversimplification problem with position definitions. Even though it provides valuable information, the 13 positions defined in this study can not be considered as an exact solution to player positioning problem. Other works in the literature showed that using different data set and different features may change the number of positions observed from data (Singh and Ahmad, 2015).

Positions deducted from data may have many applications. For example, it was reported that algorithm-based positions could be used in determining which types of players are more crucial for the success of basketball teams (Lutz, 2012).

CHAPTER 3

METHODOLOGY

3.1 Data mining process

Data mining can be defined as a general process of extracting knowledge and insights from raw data to predict outcomes using various machine learning algorithms. The use of an organized experimental approach to the problem of prediction is useful in order to achieve the best outcome from a collected data set. In the literature, there have been numerous efforts to standardize the data mining process. Shearer (2000) proposed a methodology that is designed to be used in data mining tasks. Figure 1 shows the general architecture in this study and it was named as Cross Industry Standard Process for Data Mining also known as CRISP-DM.

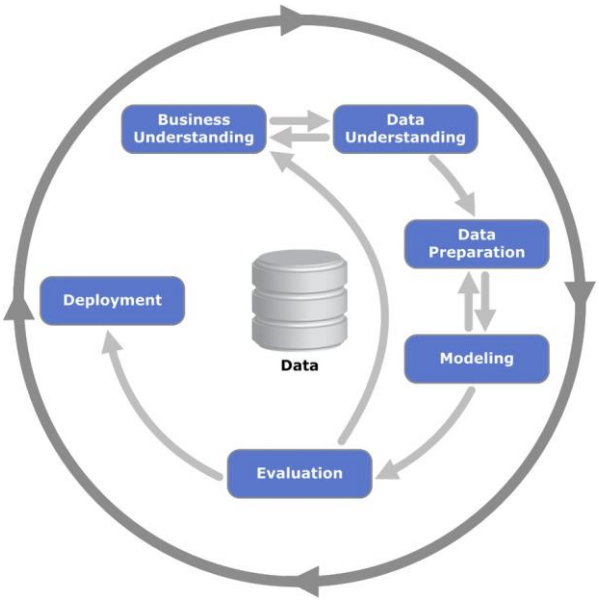

Figure 1. CRISP-DM architecture of Shearer (2000)

The process starts with business understanding phase followed by data understanding phase. Afterward, data preparation is essential before applying any predictive model. Model results are evaluated and if there is any possibility of an improvement because of the missing points in business understanding, the process goes back to the first step. This cycle is repeated until it reaches a stable state. As the final step, the deployment of models is completed in order to use regularly in business tasks.

CRISP-DM is a generalized process that can be used in any data mining task and this methodology is both robust and widely used in various machine learning projects. Bunker and Thabtah (2019) extended CRISP-DM architecture and proposed a method- ology that is specifically designed for the complex problem of sports prediction. They named their framework as Sport Result Prediction Cross Industry Standard Process for Data Mining, "SRP-CRISP-DM". The steps are similar to the base framework but the contents of each step are modified according to the dynamics of sports prediction. The steps for SRP-CRISP-DM can be summarized as in the Table 1.

Table 1. Stages and Steps for SRP-CRISP-DM

Stage	Steps
Domain Understanding	Understand objective
	Understand the main characteristics of the sports being modeled
Data Understanding	Connect the data source and automate data collection if possible
	Decide on the granularity of data to be used in modeling
	Decide on the target variable
Data Preparation and Feature Extraction	Preprocess collected data set and merge with external data sets
	Split features into different groups according to their information content
	Select features using feature selection algorithms
Modeling	Review literature and select models to apply
	Try out different models using different features in preprocessed data
Model Evaluation	Select measurement metric to evaluate model performance
	Retain the order of games played in order not to allow forward-looking
	Decide on the splitting method of train and test sets
Model Deployment	Select best model
	Automate data collection and preprocessing steps if necessary

In this study, the framework of SRP-CRISP-DM is followed step-by-step to structure an effective machine learning pipeline.

3.2 Dataset selection

Selecting a proper data set is crucial for measuring the success and applicability of this study. In the literature, there are many works that use the NBA data and propose predictive models for player performance estimation for the NBA players. However, for many years, basketball is professionally played all around the world. Additional studies to propose comprehensive predictive models and understand the dynamics influencing player performance more completely are required.

After comparing data sets from different basketball leagues in terms of maturity, availability, and cleanliness, Euroleague is selected as the best candidate to conduct our experiments. Euroleague is the highest level of European basketball club competition (Salador, 2011) and it has the essence of European basketball. The champion team is considered to be the biggest team in European basketball. Euroleague is the second most-watched basketball league after NBA, and it has been broadcast on TV screens of more than 200 countries.

The Euroleague system was played with 24 teams until it changed in 2015. Starting from the 2016-2017 season, regular-season games are played for 30 weeks with 16 teams. According to 2018-2019 Euroleague Bylaws, each team has a total of 30 matches with 15 of their opponents. There are 15 home games and 15 away games for a team. At the end of the regular season, the top 8 teams qualify to the playoff stage. Matches are between the 1st and 8th, 2nd and 7th, 3rd and 6th and 4th and 5th with respect to their rankings in the regular season. After the best of 5

playoff series in which the first four seed teams in the regular season have the home-court advantage, the teams that reach three wins qualify to the final stage. The final stage, also known as the Final Four, takes place in a predetermined city between the four finalist teams. The four final matches are played according to the single match elimination system. The winning teams in the semi-finals play for the championship game and teams eliminated in the semi-finals play in the third-place game.

3.3 Player performance evaluation metrics

In order to evaluate the performance of basketball players, there are many advanced metrics available today. These metrics are designed in a way that they summarize the stat sheet and provide a single number representing the performance. Efficiency (EFF), Player Efficiency Rating (PER) and Performance Index Rating (PIR) are widely used examples of the performance evaluation metrics.

Martin Manley who is a former statistician and sportswriter was the inventor of Efficiency formula. Efficiency is the first player performance evaluation metric and it is officially used by the NBA. Efficiency is calculated by using the formula in Equation 1. The formula is a linear combination of all the basic statistics in the stat sheet.

Equation 1. Equation of Efficiency

$$\begin{aligned} \mathrm{EFF} = {} & \mathrm{PTS} + \mathrm{REB} + \mathrm{AST} + \mathrm{STL} + \mathrm{BLK} \\ & - (\mathrm{FGA} - \mathrm{FGM}) - (\mathrm{FTA} - \mathrm{FTM}) - \mathrm{TO} \end{aligned}$$

There are more advanced metrics for player performance evaluation. Player Efficiency Rating also known as PER is developed by John Hollinger who is also a former analyst and sportswriter. PER has a more complex formula compared to Efficiency. The intention of Hollinger (2002) for designing such a complex formula was to come up with a single number that considers different playing styles of the teams and variations of minutes played by each player.

PIR is the metric that is mainly used in Europe. PIR is initially used by the Spanish ACB League. Other European basketball leagues have some version of PIR to evaluate the performances of the players. The Euroleague also uses PIR to determine the Most Valuable Player of the Week, the Most Valuable Player of the Month and all Euroleague first and second team. The formula of PIR is also similar to EFF and it summarizes basic statistics for a player. PIR can be calculated as in the Equation 2. Basically, fouls committed and blocks against are subtracted and fouls received added into the calculation of Efficiency formula to come up with PIR.

Equation 2. Equation of Performance Index Rating

$$\mathrm{PIR} = (\mathrm{PTS} + \mathrm{REB} + \mathrm{AST} + \mathrm{STL} + \mathrm{BLK} + \mathrm{RF}) - (\mathrm{FGA} - \mathrm{FGM}) - (\mathrm{FTA} - \mathrm{FTM}) - \mathrm{TO} - \mathrm{PF} - \mathrm{BA}$$

Performance evaluation metric in the scope of this study will be PIR. There are a couple of reasons why we select PIR over other metrics. First, PIR is a more comprehensive metric than Efficiency and a more interpretable and understandable metric than Player Efficiency Rating. Second, PIR is a widely used metric for the Euroleague Basketball and other European Leagues. In the Euroleague, most of the player rewards are given by looking at PIR. Additionally, news and game reports

also use PIR. In this study, experiments are conducted using the data collected from Euroleague. Therefore, when considering model outputs, it would be easier to track the reasons for overperforming and underperforming players.

3.4 Model selection

Model selection is one of the most critical stages for the data mining process. In this study, a fast and accurate model is needed to try out many different combinations of features for predicting basketball player performance. PIR which is the main metric for performance evaluation in this study is a continuous variable; therefore, the machine learning task for this study is a regression which is a supervised learning algorithm that uses previous outcomes for model training.

The availability of computing power made machine learning algorithms applicable to many real-life problems. In recent years, there have been online data science competitions that challenge machine learning enthusiasts to solve different problems in many fields. One of the most used algorithms for regression tasks in these competitions is the Gradient Boosting Machines algorithm (Friedman, 2001) because of its high accuracy and speed. The machine-learning algorithm of Gradient Boosting Machines uses weak learners (decision trees) and ensembles them sequentially to create a strong learner. Weak learners are sequentially created because after each iteration, a new weak learner that learns from the errors of previous weak learners is added into the model. Learning from the errors of subsequent weak learners enables Gradient Boosting Machines to reach a low level of errors in a short time. Gradient Boosting Machines algorithm is prone to overfit the data if they iterate for too many rounds. Stopping criteria should be defined for

the algorithm of Gradient Boosting Machines before it memorizes the data and becomes ungeneralizable.

In the literature, there are number of implementations of Gradient Boosting Machines including XGBoost (Chen and Guestrin, 2016), lightgbm (Ke et al., 2017), and CatBoost (Prokhorenkova et al., 2018). These different implementations address different issues regarding performance, scalability and, handling of categorical variables. The comparisons of these algorithms on different data sets show that lightgbm is significantly better than XGboost in terms of computational speed and memory consumption (Ke et al., 2017). When the number of categorical variables is limited, lightgbm also outperforms CatBoost.

Additionally, McNamara (2018) provides detailed comparison of three different implementations of Gradient Boosting Machines for a regression task. As it can be observed from Table 2, when early stopping parameter is used, lightgbm is better than XGboost and CatBoost in terms of both speed and accuracy.

Table 2. Implementation Comparisons of Gradient Boosting Machines Algorithm for a Regression Task

	GridSearch CV		Early stopping	
Implementation	time (seconds)	rmse	time (seconds)	rmse
XGboost	2064 s	478.03	205 s	480.58
lightgbm	240 s	774.08	74 s	475.219
CatBoost	1908 s	2993.34	453 s	582.82

In this study, the focus is on trying and deriving various features to accurately predict player performances in Euroleague. Therefore, in addition to accuracy, speed is also crucial for the aim of this work. After considering all the pros and cons, lightgbm model is selected for the application of feature engineering.

CHAPTER 4

DATA AND PREPROCESSING

In this chapter, the source of the data sets and the methods for collecting these data sets are provided. Afterward, the steps for cleaning collected data sets are mentioned and finally, exploratory data analysis is conducted to gain insights about the collected data sets.

4.1 Data collection

In order to achieve the aim of this study, the historical performances of basketball players in the Euroleague should be collected. Other than historical performances, any metadata related to players or teams has a potential to be useful for feature engineering.

basketball-reference.com, euroleague.net, mackolik.com, gigabasket.org could be considered as reliable online sources for Euroleague Basketball data. euroleague.net is the official website of Euroleague Basketball and it is the most reliable and cleanest data source that can be found for Euroleague. euroleague.net website not only serves news, articles and game results but also provides the main statistical data for Euroleague Basketball. Coaches, fans, statisticians or anyone interested in Euroleague Basketball data can openly access euroleague.net and investigate different statistics on players, teams or games.

euroleague.net provides an Application Programming Interface (API) for accessing the database through software written in different programming languages

such as PHP, javascript, R, etc. In this study, python 2.7 is utilized for collecting and maintaining the required data sets.

Three main data sets are scraped from euroleague.net website for the purpose of this study: Play-by-play, boxscore and player information. API requires Game Id and season information to collect play-by-play and boxscore data. Additionally, Player ID and season information are required to collect player information data. The data set used in this study contains 720 regular-season games from 2016-2017, 2017-2018, 2018-2019 Euroleague seasons. During these seasons, a total of 15368 records obtained from performances of 464 individual athletes.

Play-by-play data consist of sequential event logs during a game. The first few lines of an example play-by-play data are shown in Figure 2. In the sport of basketball, an event refers to a low-level play action that is recorded in the statistics sheet.

1st Quarter | 2nd Quarter | 3rd Quarter | 4th Quarter | Overtime

Min	Anadolu Efes Istanbul	Score	FC Barcelona Lassa
	Begin Period	0 - 0	
09:59	DUNSTON, BRYANT	0 - 0	
09:59		0 - 0	TOMIC, ANTE
09:55		**0 - 2**	BLAZIC, JAKA Two Pointer (1/1 - 2 pt)
09:33	MICIC, VASILIJE Turnover (1)	0 - 2	
09:31		0 - 2	CLAVER, VICTOR Steal (1)
09:27		**0 - 4**	CLAVER, VICTOR Two Pointer (1/1 - 2 pt)
09:26		0 - 4	BLAZIC, JAKA Assist (1)
09:15		0 - 4	HEURTEL, THOMAS Foul (1)
09:15	MOERMAN, ADRIEN Foul Drawn (1)	0 - 4	
09:03	SIMON, KRUNOSLAV Three Pointer (1/1 - 3 pt)	**3 - 4**	
08:45		3 - 4	HEURTEL, THOMAS Missed Two Pointer (0/1 - 0 pt)
08:44	DUNSTON, BRYANT Def Rebound (1)	3 - 4	
08:33	SIMON, KRUNOSLAV Missed Two Pointer (0/1 - 3 pt)	3 - 4	
08:33		3 - 4	CLAVER, VICTOR Def Rebound (1)
08:04		3 - 4	CLAVER, VICTOR Turnover (1)
08:04	SIMON, KRUNOSLAV Steal (1)	3 - 4	
08:04	ANDERSON, JAMES Two Pointer (1/1 - 2 pt)	**5 - 4**	

Figure 2. Example play-by-play data for a game

Main events that are included in this data set consist of Missed Two Pointer, Two Pointer, Missed Three Pointer, Three Pointer, Shot Rejected, Assist, Bench Foul, Begin Period, Coach Foul, Foul, Technical Foul, Unsportsmanlike Foul, Def Rebound, End Game, End Period, Missed Free Throw, Free Throw, Block, Sub In, Off Rebound, Offensive Foul, Sub Out, Foul Drawn, Steal, Turnover, Time Out and TV Time Out.

Boxscore data shows game, team and player statistics in a tabular format. An example of a boxscore data for a game is shown in Figure 3 Boxscore is derived from play-by-play event logs that are aggregated and structured in a way that it can be used to observe game summary for players and teams. Summary statis- tics included in Boxscore data consists of Minutes Played, Points, 2-Point Field Goals (Made-Attempted), 3-Point Field Goals (Made-Attempted), Free Throws (Made-Attempted), Rebounds (Offensive, Defensive and Total), Assists, Steals, Turnovers, Blocks (In Favor and Against), Fouls (Committed and Received), PIR.

							Rebounds						Blocks		Fouls		
#	Player	Min	Pts	2FG	3FG	FT	O	D	T	As	St	To	Fv	Ag	Cm	Rv	PIR
0	LARKIN, SHANE	24:29	18	2/5	3/6	5/7		2	2	2	2	1			1	4	18
1	BEAUBOIS, RODRIGUE	9:45		0/1	0/2			1	1						1	1	-2
3	SAYBIR, YIGITCAN	DNP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	BALBAY, DOGUS	6:57						2	2	1			1		2		2
15	SANLI, SERTAC	DNP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
18	**MOERMAN, ADRIEN**	40:00	11	3/6	1/3	2/2	1	5	6	2	1	1			1	4	17
19	TUNCER, BUGRAHAN	DNP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
21	PLEISS, TIBOR	10:01	4	2/2				5	5						2	1	8
22	**MICIC, VASILIJE**	35:37	15	2/7	2/8	5/5		1	1	4		3		1	2	5	8
23	**ANDERSON, JAMES**	15:46	2	1/1	0/2			4	4						1		3
42	**DUNSTON, BRYANT**	29:56	8	3/4	0/1	2/3	2	2	4				1		2	4	12
44	**SIMON, KRUNOSLAV**	27:29	14	2/5	3/5	1/2	1	1	2	3	3	1		1	4	2	12
	Team						2	1	3								3
	Totals	200:00	72	15/31	9/27	15/19	6	24	30	12	6	6	2	2	16	21	81
				48.4%	33.3%	78.9%											

Head coach: ATAMAN, ERGIN

Figure 3. Example boxscore data for a game

Player Information data is the metadata about every player. As shown in Figure 4, name, team, jersey number, height, birth date and nationality information available for all players that are played in Euroleague.

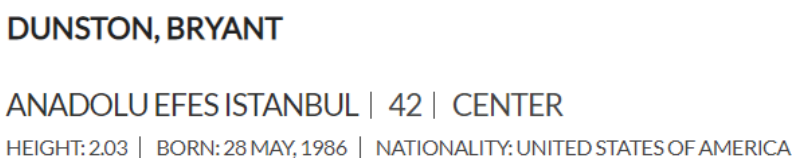

Figure 4. Example player information data for a player

4.2 Data cleaning

Inconsistent and inaccurate records can mislead the results while doing data analysis and modeling. In order to prevent misinterpretations and false calculations, the cleaned data set should be prepared before starting feature extraction. A step by step cleaning process is designed for detecting and correcting inaccurate records from collected data sets (play-by-play, boxscore and player information).

1. Data type constraints are handled by extracting and converting the related fields. For example, in the player information data, height is given in string format as “2.03” for a player having 203 cm height. This field is converted to float after successfully extracting meters and centimeters part and converting all to centimeters.
2. Range constraints are checked according to the maximum and minimum values of each field in collected data sets. For instance, in the boxscore data received fouls for a player in a game has the maximum value of 5 and a minimum value of 0. Any other value exceeding these limits are corrected using play-by-play data.

3. Mandatory fields such as birth date, height are checked. For example, the height and birth date fields in the records of players who do not have any height or birth date information on euroleague.net are imputed manually by searching birth dates and heights through the search engines.
4. Set-Membership constraints are created. Multiple values that should be equal to the same value are reduced to a single value. For example, the nationality field in Player Information data has different values for the players from the same nationality ("USA" and "United States of America" converted to "United States").
5. Foreign-key constraints are checked. Player Id is a foreign key to merge all collected data sets into a single enriched data frame. Players that are not available in the Player Information data are omitted from boxscore and play-by-play data.

4.3 Data visualisation

In statistics, Exploratory Data Analysis (EDA) is a visual analysis methodology for summarizing and observing the main characteristics of a given data set. In this section, an extensive Exploratory Data Analysis is performed to investigate collected data sets in order to discover patterns and anomalies. Additionally, the main assumptions are checked with the help of summary statistics and graphical representations. To begin with, there is a total of 464 players who played in the Euroleague in 2016-2017, 2017-2018 and 2018-2019 seasons. From those who played in these seasons, 59.91% of them played for a single season, 22.19% of them played for two different sea- sons and only remaining 17.88% of them played in all

three seasons. These proportions indicate that a significant amount of the players have observations for a single season only.

During the Regular Season, 30 rounds of games are played between 16 teams and these rounds are called as weeks. Figure 5 displays that during their respective season's, 43.79% of the players played more than 27 weeks. It can also be noted that 39 of the players have played less than three games meaning that the sample size for those players' performances is limited for analysis. According to Berri et al. (2006), aging has an influence on player performance. In Figure 5, age distributions for all players can be observed. Using their birth date data, the age of the players is calculated for each season separately in order to include an aging effect to the predictive models to be developed. In Euroleague, players have an age in the range of 15 to 39 having a mean of 27.07.

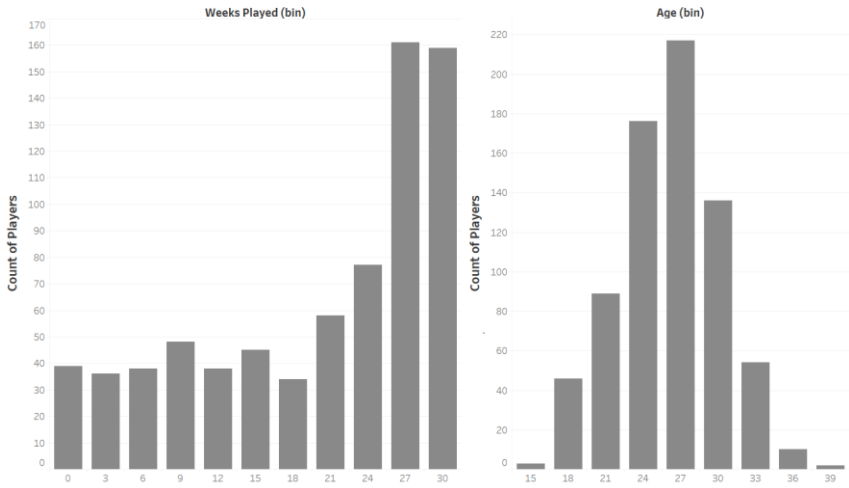

Figure 5. Weeks Played and Age histograms for Euroleague players

Salador (2011) asserted that country of origin should be considered while developing models for player performance prediction. In Figure 6, the map shows the geographical representation of nationalities for all 464 players. As can be seen from

the map, Euroleague is a relatively multinational competition attracting players from all over the world. Therefore, nationality factor may have an influence on player performances and it should be inspected. Most of the players are from the United States followed by European countries including Serbia, Greece, Spain, Russia, Turkey, and France.

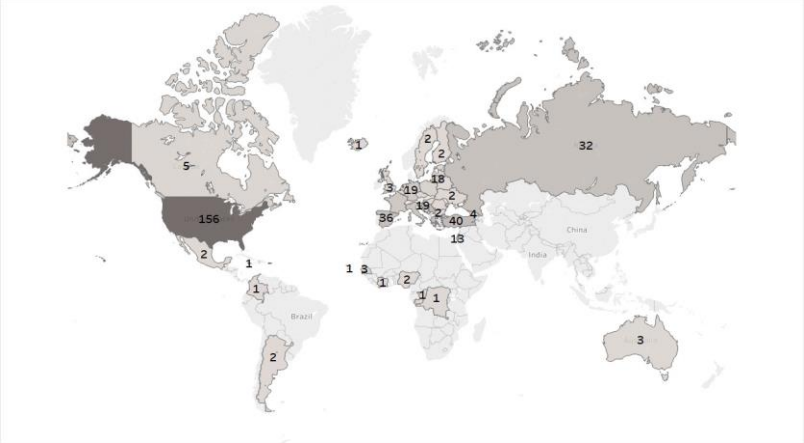

Figure 6. Nationality map of Euroleague players

Salador (2011) argues that physical characteristics such as height may be decisive for player performance. Distributions of height for different positions are shown in Figure 7. Height is an important factor for defining positions, that's why distributions widely differ from each other per position. On average, Centers are 7 cm longer than Forwards and 18 cm longer than Guards. Before assessing the performance of the players, height advantages should be considered for different positions.

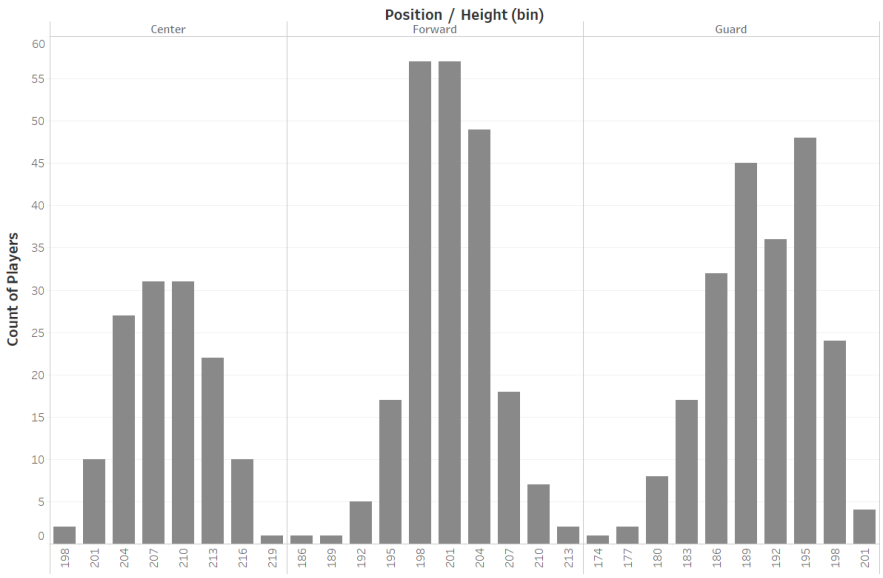

Figure 7. Height histograms per position for Euroleague players

Figure 8 allows us to see both distributions of four main variables in the collected dataset (PIR, Points Scored, Rebounds, Assists) and pairwise relationships between them. The distribution of PIR has a standard deviation of 8.002, and it is positively skewed with a mean of 7.929 and a median of 7.000. During the considered seasons, the maximum value for PIR is 44, whereas the minimum value is −13. As can be expected, the correlation between PIR and Points Scored is positive and statistically significant at the level of 0.872. Points Scored is not the only parameter that had a statistically significant correlation with PIR. Rebounds and Assists are also positively correlated with PIR. These statistically significant correlations should be taken into account during the feature extraction part of the analysis.

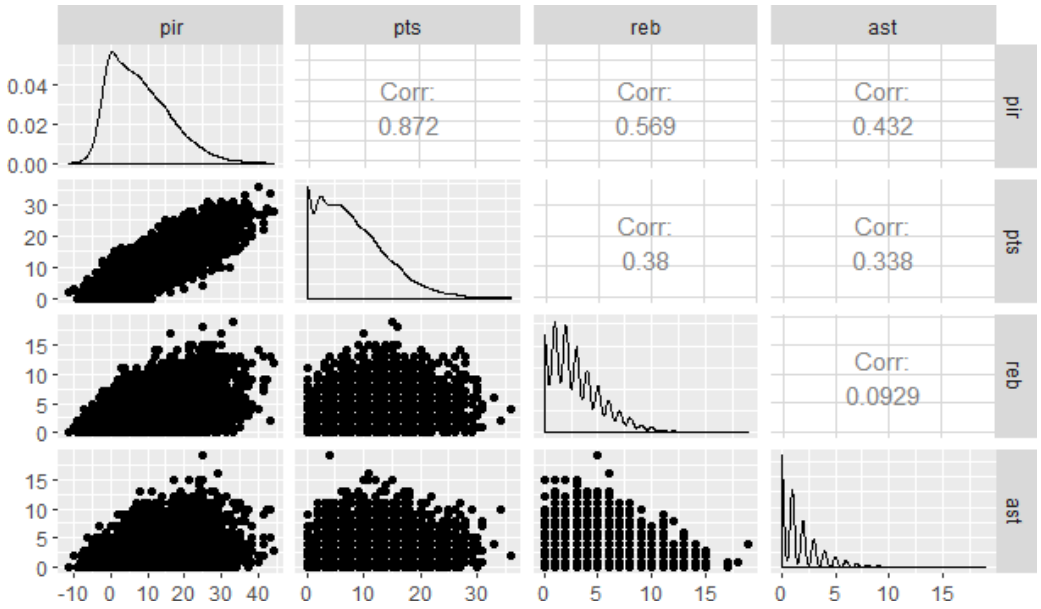

Figure 8. Pairplot of Performance Index Rating with points, rebounds and assists

Even though rebounds and assists have the same coefficient in the formula of PIR, the Performance Index Rating's correlations with these two variables are significantly smaller compared to the correlation with Points Scored. Generally speaking, it is easier to score points than collecting rebounds or providing assists to teammates. However, this general assumption differs when positions and other specifications of the players are taken into account. For a center player with 220 cm height, it may be easier to collect rebounds than scoring points, whereas, for a guard with a pass-first mind, it could be easier to provide assists. Other than Points scored, Minutes Played also has a high correlation with PIR at the level of 0.65. Certainly, players have to be on the court to perform well and fill the stat sheet.

Figure 9 shows us the mean PIR for players during the stated seasons. It is observed that upper and lower whiskers for PIR not only change from season to season but also differ by position.

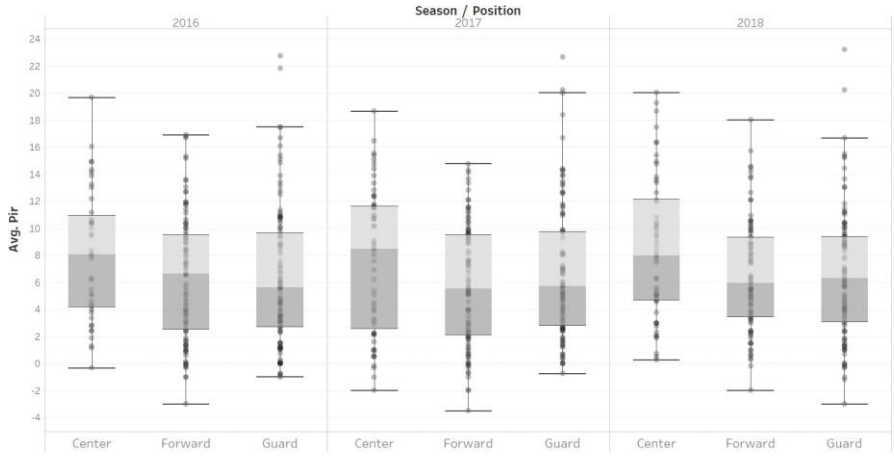

Figure 9. Position-wise Performance Index Rating boxplots

Although guards have a lower mean value compared to forwards and centers, they have multiple high performing outliers for observed in different years. For example, Luka Doncic from Real Madrid Basketball had a mean PIR of 22.67 which was honored with Most Valuable Player prize at the end of the 2017-2018 season. Outlier performances are mostly observed among guards, possibly because guards are the main ball handlers in the court and this gives them an upper hand for using more shots resulting in more points scored. In general, centers have a greater mean value of PIR compared to guards and forwards. Centers are assigned with some specific roles that guards and forwards are limited to do. These roles include protecting the paint area, offensive rebounding and post up playing. As a general rule of thumb for the sport of basketball, teams prefer to play with two guards, two forwards, and one center. Therefore, guards and forwards have a backup player on the court for their duties. Considering all, being a center player gives an advantage in terms of PIR.

CHAPTER 5

FEATURE ENGINEERING AND MODEL GENERATION

In the scope of machine learning, a feature is a measurable variable that is used to explain some part of individual data objects (Dong and Liu, 2018). For example, sepal length and petal length are some of the features that are used to describe species of iris flower in the Iris Data Set (Dua and Graff, 2017).

In order to design effective machine learning models, comprehensive and independent features that explain the underlying information on the target variable should be presented. Feature engineering is the process of transforming, generating, and selecting features on the collected data sets. Even with the recent developments in the data analytics and machine learning area, most of the designed algorithms are not fully capable of understanding the reasoning behind the target variables only being applied on a collected data set. Machine learning experts are needed for generating features in order to extract useful information for machine learning models to work.

Extracting meaningful features requires extensive domain knowledge. The process of feature engineering is not a simple line rather a cycle of learning that goes back and forth between the feature engineering stage and model development stage. In this section, step by step feature engineering is applied so as to improve models by inserting more explanatory variables. After each step, the most erroneous examples are inspected and the reasonings behind the model outputs are tried to be understood with the help of our domain knowledge.

Before applying any machine learning algorithm, collected datasets are usually split into training, validation and test sets to prevent overfitting and underfitting. Overfitting occurs when the model memorizes all the examples in the dataset. Complex model selection or running the machine learning algorithm with too many features may lead to overfitting. On the other hand, underfitting is observed when the model is insufficient to explain the relationship between model input and model output. Simple model selection or lack of explanatory input variables may be the reasons for underfitting. A training data set is used to learn different patterns and underlying trends of the dataset. A validation data set is a representative sample of the data that is utilized for tuning hyperparameters of the model and understanding whether the model is generalizable or not. The test data set is to measure the performance of the developed model on unseen data. The evaluation of the model on the test set gives information about the applicability and predictive capacity of the model in real life.

In this study, the player performances in the 2018-2019 Euroleague season are used in the test set. Prediction accuracies of models with different features are measured on this unseen data. The player performances in 2016-2017 and 2017-2018 seasons are separated and randomly divided into training and validation sets. Two-thirds of the performances in these seasons are in the training set, whereas the remaining one-third of the performances are in the validation set.

5.1 Historical time series performance

As the beginning point for the feature engineering stage, the most basic features that are spotted in the literature review are prepared as the input for the lightgbm model.

These features include both historical game-related features and player information features. In our model, we are trying to predict PIRs of players at a given week w which is denoted as PIR_w in our preprocessed data set. Firstly, historical game-related features that are used in this stage are PIRs for a player from previous games denoted as PIR_w-t where t is the count of the weeks from the current week. Other historical game-related features which are denoted as home_away_w-t show whether the PIR observed at week w-t was in a home game or away game. Additionally, height (Salador, 2011), age (Berri et al., 2006), position (Berri and Schmidt, 2010), and nationality (Salador, 2011) of the players are used as player information features. The important thing to note is that for each player, position denoted as Euroleague Position is defined by Euroleague. In the collected data set from Euroleague, there are three positions available: Guard, Forward, and Center.

In Figure 10, a plot of a weak learner tree built by the lightgbm model is presented. Because the leaf-wise tree growth algorithm is utilized for lightgbm, the depth of each branch might be different from others.

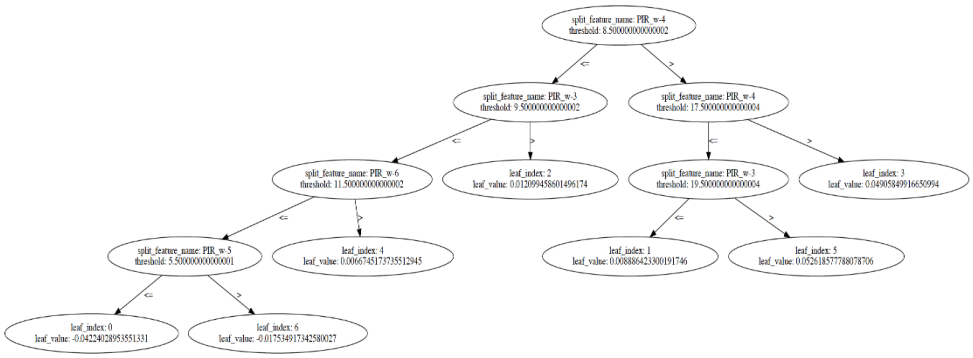

Figure 10. Example of weak learner tree built by lightgbm

The example tree has a maximum depth of four and it uses PIR_w-3, PIR_w-4, PIR_w-5, and PIR_w-6. In this model, there are a number of weak learner trees that use other available features to cover different aspects of player performance. However, by looking at the features used in this tree, it can be concluded that noncurrent performances are tried to be handled in this tree.

Equation 3 shows the calculation of Root Mean Square Error (rmse) which basically measures the average magnitude of the error.

Equation 3. Root Mean Square Error Calculation

$$\text{rmse} = \sqrt{\frac{\sum_{i=1}^{N}(\text{ Predicted }_i - \text{Actual}_i)^2}{N}}$$

The graph on the left in the Figure 11 shows the number of iterations and the model error which is rmse. The model iterated around 400 times before stopping.

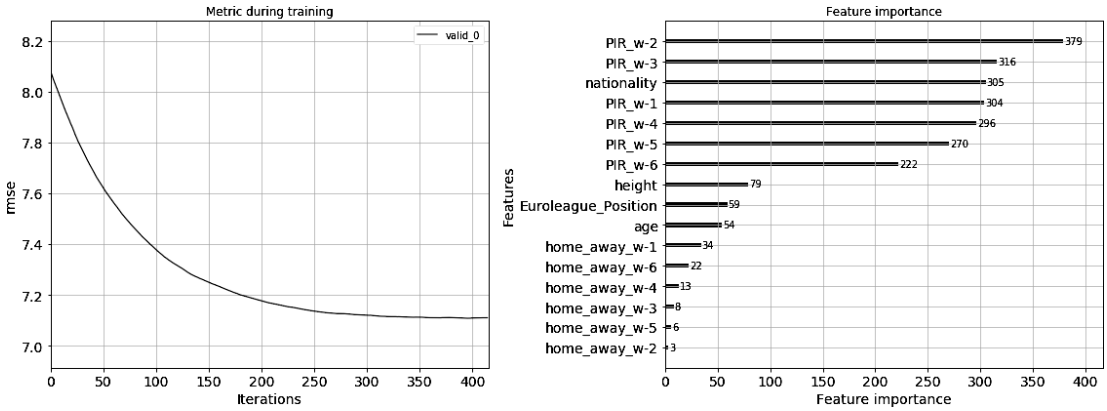

Figure 11. Training metrics and feature importances of the lightgbm model with historical time series performance and player information features

To avoid overfitting, we have regularized our model by setting an early stopping parameter. An early stopping parameter forces the model to stop when the model accuracy on the validation set does not improve for the given number of rounds.

The model started training with 8.05 rmse and at the end of the training, rmse reduced to 7.12 after 416 iterations. The rmse value on the test set is 7.07 which is close to what we've observed during training. R-squared of the initial model on the test set is 19.08% meaning that the variations in the PIR for the week w are not sufficiently explained. Feature importance in lightgbm is measured by the number of times that the considered feature is used to split a tree. From the feature importances chart in Figure 11, it can be observed that PIR_w-2, PIR_w-3, and nationality are the most important features at the end of model training. The reason why PIR_w-1 comes after PIR_w-2 and PIR_w-3 could be the fact that coaches in basketball tend to take action against the previous week's top-performing players of the opponent team. Height, age and, position information of the players are used in the model but these features are not as utilized as historical performances. As expected, the most important home/away feature is home away w-1 probably because players' performances can be affected whether they played their last game at their home court or their opponents' court. At the end of the first model trial, it can be said that more features are required to give better estimations on the performance of players.

5.2 Advanced basketball statistics

In the early 1990s, Dean Oliver, who is an American Statistician, introduced new statistical metrics that are derived from boxscore statistics to evaluate player's and

team's performances. Dean Oliver is known as a pioneer in analytics that helps to rule many basketball teams, from NBA to high schools. One of the key metrics that Oliver came up with was the possession metric which is used to quantify the team and player possessions' Oliver (2004). Using the possession metric, several other offensive and defensive metrics were created afterward such as True Rebounding Rate, Assist to Turnover Ratio, etc. (Kubatko et al., 2007). Following advanced basketball metrics with the given equations are introduced to use in the modeling stage in our analysis. Field Goals Attempted (FGA) in Equation 4 is the sum of two points attempted and three points attempted.

Equation 4. Field Goals Attempted

$$\mathrm{FGA} = 2\mathrm{PA} + 3\mathrm{PA}$$

Possessions (POS) is the estimation of the possessions a player has in a game. The calculation for POS is provided in Equation 5.

Equation 5. Possessions

$$POS = FGA + 0.44 \times FTA - OFFREB + TO$$

As it can be seen from the Equation 6 Assist to Turnover ratio (ast/to) is the number of assists for a player compared to the number of turnovers committed.

Equation 6. Assists to Turnover Ratio

$$ast/to = \frac{AST}{TO}$$

Offensive Efficiency Rating (OER) measures a team's points scored per 100 possessions and the formula for OER is given in Equation 7.

Equation 7. Offensive Efficiency Rating

$$OER = 100 \times \frac{PTS}{POS}$$

Defensive Efficiency Rating(DER) is the number of points allowed per 100 possessions by a team and it is calculated as in Equation 8.

Equation 8. Defensive Efficiency Rating

$$DER = 100 \times \frac{allowedPTS}{POS}$$

Equation 9 shows the calculation for Effective Field Goal Percentage (eFG) which is the percentage that weights three points made 1.5 times higher than two points made so as to get an weighted field goal rate.

Equation 9. Effective Field Goal Percentage

$$eFG = \frac{2PM + 3PM \times 1.5}{FGA}$$

The formula for True Shooting Percentage (TS) is provided in Equation 10 and it is defined as shooting percentage that takes not only two-point field goals but also three point field goals and free throws according to predefined weights.

Equation 10. True Shooting Percentage

$$TS = \frac{PTS}{2 \times (FGA + 0.44 \times FTA)}$$

True Rebounding rate (TR) measures the percentage of the available rebounds a player grabs at the offensive and defensive ends of the floor.

Equation 11. True Rebounding Rate

$$TR = 100 \times \frac{\big(REB \times (TmMP \times 0.2)\big)}{\big(MP \times (TmREB + OppREB)\big)}$$

Usage Percentage is the usage of a player compared to his teammates while he is on the floor. As it can be seen from the Equation 12, usage percentage has a more complex formula compared to other advanced statistics.

Equation 12. Usage Percentage

$$USG = 100 \times \frac{\big((FGA + 0.44 \times FTA + TO) \times (TmMP \times 0.2)\big)}{\big(MP \times (TmFGA + 0.44 \times TmFTA + TmTO)\big)}$$

In addition to historical game-related features and player information features, new features using the formulas of advanced basketball statistics are generated and added into our modeling as input variables. At the end of the training, rmse value reduces to 7.05 meaning that advanced basketball statistics slightly affected the overall accuracy. On the test set, 7.03 rmse is observed and R-squared is 20.20%. Possessions and field goals attempted are the most important advanced

basketball statistics for prediction player performance. Even though, almost all the features from advanced basketball statistics seem essential in the feature importance chart of Figure 12, the predictive power of our model is still not improved as expected.

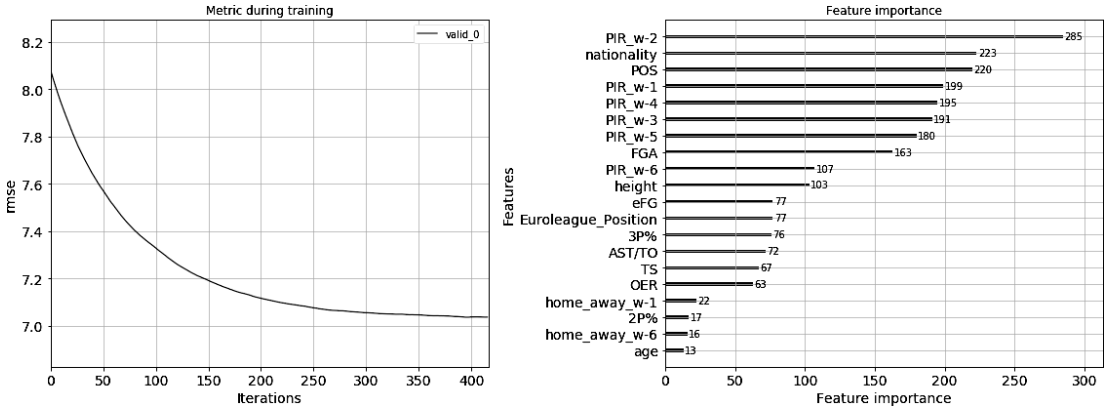

Figure 12. Training metrics and feature importances of the lightgbm model with advanced basketball statistics

5.3 Basic statistics for the players and the opponent team

In the first two iterations with different features, PIRs from previous rounds are excessively used in trees that are constructed with lightgbm. However, as it was stated in the methodology chapter, PIR is composed of multiple other basic player statistics including Points, Rebounds, Assists, Steals, Blocks, Received Fouls, Committed Fouls, Blocks Against, Turnovers, Missed Field Goals and Missed Free Throws. In the third iteration, the historical values of these basic player statistics are added into our model as new input variables in order to make the underlying effects of basic player statistics that lead to resultant player performance more apparent for the model. In addition, basic team statistics for the opponent in the next week also

included in this iteration. The purpose of adding basic team statistics is to cover opponent team effects in the player performance. Opponent team statistics are denoted similar to player statistics, but an indicator of ‘o’ is included after each statistics’ abbreviation. The performance of the players is highly dependent on the minutes they played in the court as it was shown in the previous chapter. Therefore, features related to minutes played in the previous rounds are also created in this step.

As in the second iteration, rmse value at the end of training is slightly improved to 7.01. On the other hand, the rmse value on the test set is 7.00, and the explanatory power of our model is also improved with 20.78% R-squared. From the feature importance chart in Figure 13, it can be observed that field goals attempted becomes the most critical feature. Generally, field goals attempted represents the offensive potential of a player by measuring total shots taken by the considered player. Additionally, minutes played in the last week is the second most important feature probably because the recent player rotation preferences of the coaches are mostly reflected in the minutes played by each player in the last week. Two points attempted and received fouls are two main opponent team statistics that are found as important for predicting player performance. When the opponent team has players who can easily receive fouls, the probability of getting into foul trouble may increase, resulting in low performances for the players of the other team.

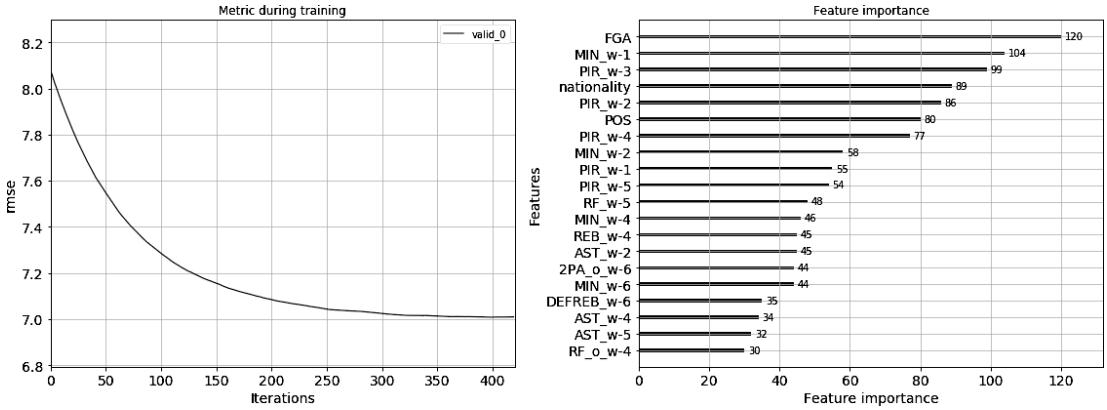

Figure 13. Training metrics and feature importances of the lightgbm model with basic statistics for the players and the opponent team

5.4 Defensive metrics from Euroleague defined positions

In the literature review chapter, it was stated that basketball statistics are mostly focused on the offensive accomplishments of the players and the statistics related to defense are mostly omitted. In the sport of basketball, the offensive skills and statistics of a player are important to estimate his performance; however, a player's performance also heavily depends on the defensive skills of his matchup player. In today's basketball, most of the coaches prefer man-to-man defensive strategy over zone defense. As a result, player matchups on defense are decided according to players' positions. The main reason behind the position-based matchups is that characteristics of different positions such as quickness and strength differ between positions and a player can suffer from these differences while defending players in other positions. For example, most of the time, a guard is responsible for defending guards in the opponent team. However, when a guard tries to defend one of the forwards in the opponent team, the size mismatch happens, and the forward can take advantage of this mismatch to score an easy basket.

Until now, the features that we have created does not capture the defensive ability of the players in the opponent team. When we review the previous studies which propose a method to predict player performance in basketball, it is observed that defensive features about the opponent team are not considered. To fill this literature gap, position-based defensive metrics are defined by calculating allowed points, rebounds, assists, free throws, two pointers, and three-pointers by each position.

In order to create position based defensive metrics, we follow the approach in Figure 14 and first calculate position-wise average statistics of opponents for each team. Secondly, we add the information of the opponent team in the next week for each player. Calculated position-wise average statistics are merged with players' data according to players' position and next week opponent. In this way, we have six newly created features for each player that represent the opponent team's ability to defend the position of the player. These new features are denoted as EuroleagueP_a_PTS for points allowed per position, EuroleagueP_a_AST for assists allowed per position, EuroleagueP_a_REB for rebounds allowed per position, EuroleagueP_a_3PM for three-pointers allowed per position, EuroleagueP_a_2PM for two pointers allowed per position, and EuroleagueP_a_FTM for free throws allowed per position.

Algorithm Extraction of Position-Based Defensive Metrics

procedure GET DEFENSIVE METRICS(Boxscore)
  for each position $\in \mathcal{C}$ **do**
    for each opponent $\in \mathcal{O}$ **do**
      Group based on week
      Calculate mean of PTS, AST, REB, 2PM, 3PM, and FTM
      Insert calculated record into Dataframe **D**
    end for
  end for
  for each player $\in \mathcal{P}$ **do**
    Find next week opponent $\mathbf{O}_p$
    Filter record of $\mathbf{O}_p$ and position $\mathbf{C}_p$ of player in Dataframe **D**
    Merge filtered record with Boxscore Data of $\mathcal{P}$
  end for
  Return joined Boxscore Data
end procedure

Figure 14. Position-based defensive metrics extraction algorithm

Unlike the first three iterations, the model accuracy during the training improved to the rmse value of 5.62. Metric during the training graph in figure 5.5 shows that almost 300 more iterations are required for training with new features. On the test set, critical progress is achieved by reducing the rmse value from 7.00 to 5.57, and the value of R-squared is increased to 49.75%. Our model now has a significant boost in its explanatory power to explain the variations in the player performances. As it can be seen from the feature importance chart in Figure 15, all of the position based defensive metrics are considered as important by the lightgbm model. In addition to field goals attempted, PIRs and minutes played in the previous rounds are again considered in the top features. The nationality of the players is also used by the model for splitting trees.

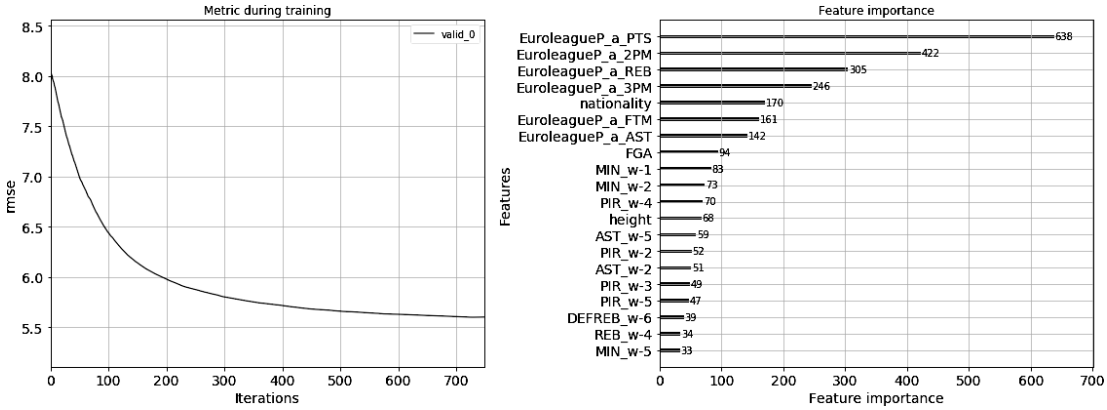

Figure 15. Training metrics and feature importances of the lightgbm model with defensive metrics from Euroleague defined positions

5.5 Defensive metrics from clustering based positions

After evaluating the results in the fourth iteration, we can conclude that the defensive ability of the opponent team is now included in our model. However, for calculating position based defensive metrics, we heavily rely on the positions defined by Euroleague. In the literature review, we found that positions may have oversimplification and incorrect classification problems. In order to overcome these problems, we tried to define our own positions that are derived from the player's own data. Clustering algorithms are used for extracting similar groups in an unsupervised fashion. According to Ding and He (2004), the application of Principal Component Analysis (PCA) enables clustering algorithms to perform better.

PCA is a technique that reduces the dimension of the data set without missing any of the significant information (Zhang, 2000). Before applying PCA, per minute main statistics for the players are calculated, and these calculated statistics are normalized according to a minimum and maximum values of each per minute statistics. Figure 16 shows that a cumulative variance of 61.86% is explained by the

principal components pca_0 and pca_1. Adding more components increases the cumulative variance; however, after pca_1, additional explained variance from extra principal components significantly reduces. Therefore, pca_0 and pca_1 are utilized for the principal component analysis.

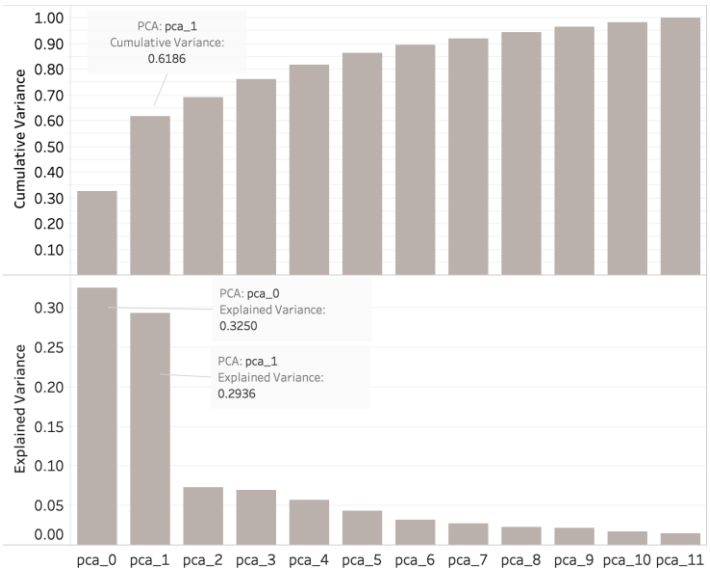

Figure 16. Explained and cumulative variance plot of principal component analysis on Euroleague data

In Figure 17, the factor loadings heatmap from the result of PCA is shown. After investigating Figure 17, it can be observed that height normalized, offreb_normalized, deffreb_normalized, and 2pm_normalized are the main contributing features to pca_0. These features are mainly required for being an inside scorer and a good rebounder. On the other hand, ast_normalized, stl_normalized, to_normalized, and 3pm_normalized are the key components for pca_1. A good play-making guard should possess these features to be successful in the game of basketball.

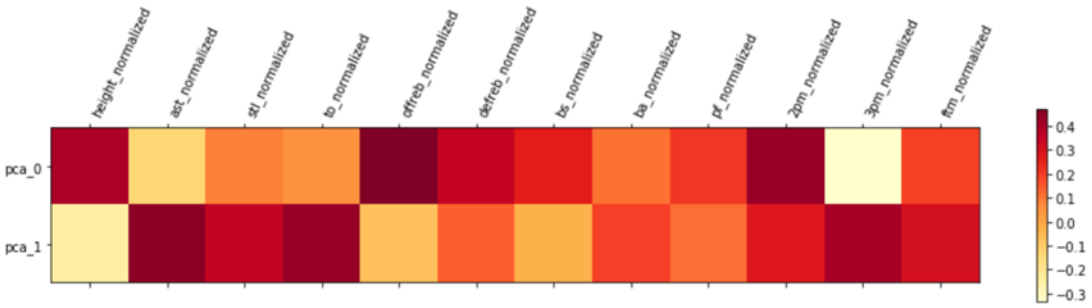

Figure 17. PCA decomposition of normalized boxscore statistics

By looking at the result of the PCA, the components can be named as follows:

1. pca_0: Big men skills component
2. pca_1: Playmaking and shooting skills component

Using the result of PCA, we run a K-Means clustering algorithm to find out our new positions. K-Means is an unsupervised learning algorithm that is used for clustering data points into a predefined number of subgroups according to their similarity (Singh and Ahmad, 2015). pca_0 and pca_1 are the inputs of the clustering algorithm, and cluster centers are decided to minimize intra-cluster variance. The optimal value of k is decided based on the silhouette method. The method of silhouette analysis is created by Rousseeuw (1987), and it is used to measure the separation between clusters. For our case, the silhouette score for k = 4 has the maximum value with 0.46 and it decreases starting from k = 5 meaning that creating four cluster can be considered as the optimum for our data set. In Figure 5.8, four new positions that are created from the data can be observed.

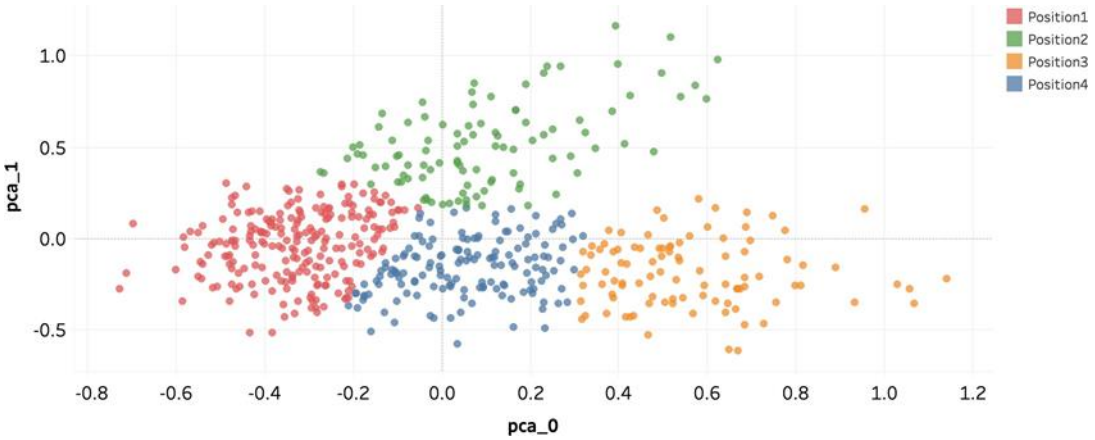

Figure 18. Position clusters for Euroleague players

The players inside the Position1 cluster have low pca_0 and low pca_1, meaning that these players do not have dominant skills in their repertoire. Position2 cluster is separated from the other 3 clusters with high pca_1 values. Distinctively, pca_0 values in this cluster vary from -0.25 to +0.60. When Position2 cluster is inspected, it is observed that players with multiple dominant basketball skills are in this cluster. Players in the Position3 cluster have high pca_0 and moderate pca_1. Closer inspection of the Position3 cluster shows that center and forward players with outstanding rebounding and inside scoring skills can be found in this cluster. The cluster of Position4 is made up with above the average pca_0 and moderate pca_1 values. Players in this cluster are guards and forwards that are mostly capable of collecting rebounds and shooting three-pointers. According to the review of clusters, our four new positions are labeled as below:

1. Position1: Role Players
2. Position2: All-arounders
3. Position3: Rebounding Big Men
4. Position4: Shooting Big Men

Using the clustering-based positions, the defensive metrics in the previous iteration are recreated. In this time, these features are labeled as ClusterP_a_PTS, ClusterP_a_AST, ClusterP_a_REB, ClusterP_a_3PM, ClusterP_a_2PM, and ClusterP_a_FTM.

The results are substantially better than the previous models. From Figure 19, it can be observed that at the end of training with 700 iterations, the rmse value is now reached 4.98. On the test set, the accuracy is even better than accuracy on the validation set with the rmse value of 4.79. R-squared has increased to 62.87%, which is almost 13% higher compared to the model that uses Euroleague defined positions to create defensive metrics. All of the six created defensive metrics are more important than other features. As shown in the feature importance chart in Figure 19, nationality, offensive efficiency rating, and possessions are still have an importance for the lightgbm model.

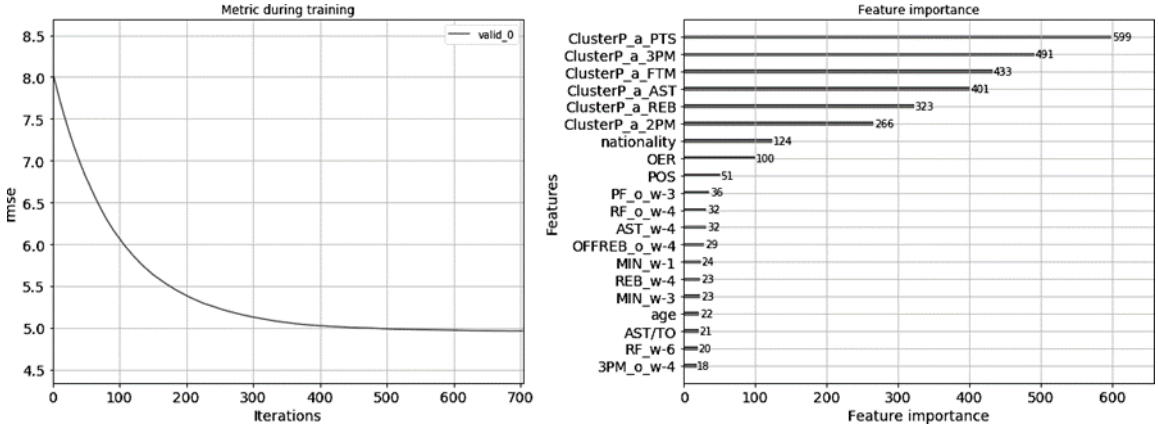

Figure 19. Training metrics and feature importances of the lightgbm model with defensive metrics from clustering-based positions

Figure 20 provides three examples for the performance predictions of the best model on the out-of-sample data. Bryant Dunston plays as a center, Adrien Moerman is a forward and Vasilije Micic is a guard. Most of the time, forwards and centers have more stable performances compared to guards. It can be observed from the Figure 20 that the performance predictions for Bryant Dunston and Adrien Moerman are more accurate than predictions for Vasilije Micic.

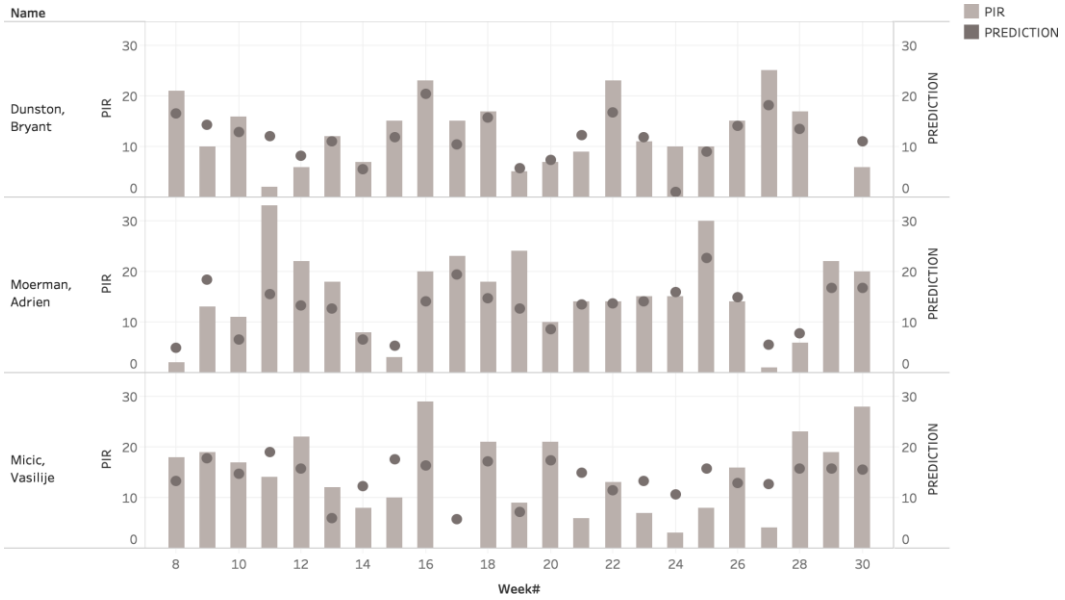

Figure 20. Predicted vs actual plot of Performance Index Rating on test data

CHAPTER 6

CONCLUSION

6.1 Results

In this study, a novel feature engineering approach to the player performance prediction in the basketball problem is given. We have followed the Sports Result Prediction Cross Industry Standard Process for Data Mining framework to come up with better estimations for player performances. In Table 3, step-by-step improvements on the test set accuracy are summarized. After each feature engineering iteration, we have observed different levels of improvements. The most significant boost in both R-squared and rmse values are achieved after adding position-based defensive metrics. We have found that the player positions provided by Euroleague have oversimplification and incorrect classification problems. New positions for each player are created using the clustering methods in the literature. Deducting the positions of the players from the data enabled our models to perform even better. As a result, the R-squared value of our model increased from 19.08% to 62.87% and the rmse value of our model is reduced from 7.07 to 4.79. The empirical findings in this study prove the importance of data-driven defensive metrics for player performance prediction in basketball.

Table 3. Test Set Accuracy Improvements with Feature Engineering

Features Added	Test Set Accuracy	
	R2	rmse
Historical Performance and Player Information	19.08%	7.07
Advanced Statistics	20.20%	7.03
Basic Statistics for the Players and the Opponent Team	20.78%	7.00
Defensive Metrics from Euroleague Defined Positions	49.75%	5.57
Defensive Metrics from Clustering Based Positions	62.87%	4.79

6.2 Managerial implications

Wheeler (2012) suggested that using machine learning algorithms to predict player performance in basketball has value and many future implications. However, in the literature, the predictive models that were developed to project player performance had inadequate accuracy levels, which made them not reliable and usable for most of the potential applications. On the other hand, our results demonstrated that when data-driven defensive metrics included in the predictive models, a significant improvement for predicting player performances in basketball can be achieved. In this sense, our models can be more reliably applied in many fields that identify player performance as an essential component.

In Sports Analytics Taxonomy V1.0, Cokins et al. (2016) proposed a technique for the classification of sports analytics applications. They created eight major branches for analytical applications in sports. These branches can be grouped into three main categories as follows:

1. Competition based branches
2. Athletic health related branches
3. Betting related branches

Sports Analytics Taxonomy provides detailed assessments for each category and shows various current analytical implications as subcategories. When we carefully inspect these subcategories, we can conclude that our models have many fields of applications especially under competition-based branches and betting-related branches.

To start with subcategories under competition-based branches, using the projections provided by our predictive models, coaches might be able to design a

more effective game/win strategy. Coaches may evaluate their players according to projections and decide on the formations for their next game. In addition, potential best performers can be prioritized based on performance projections to get the most optimal result to win the next game. In this way, the usage rate and timing of the potential best performers might be selected carefully in order not to hinder the effectiveness of them. The player performance predictions of our models are not only applicable for offensive strategy design but also provide meaningful insights for the defensive strategy design. For instance, coaches might evaluate projections of the players in the opponent team and determine their defensive strategy to take precautions for how to stop the best performers in their opponent. As a result, helping coaches for evaluating both the offensive and defensive side of their players to shape their game/win strategy can be considered as a practical in-game implication of our study.

Turning now to subcategories under betting related branches, it might be asserted that accurate performance predictions of basketball players can be used in fantasy sports and sports betting. According to several industry reports, the market size of fantasy sports is currently around $13.9 billion and is expected to reach $33 billion by 2025 (QYResearchGroup, 2019). There are many key platforms that provide an interface for playing fantasy sports online. One of the most popular categories in these fantasy sports platforms can be stated as fantasy basketball. As it was previously mentioned, fans can create virtual teams of actual players and collect fantasy points from the performances of these players to compete with both friends and family, and other basketball fans all around the world. Even though there are some platforms that fans play fantasy sports just for fun, in some other platforms such as “DraftKings.com” and “FanDuel.com”, fans can make money when they win

their fantasy basketball competition. Our predictions of player performances in basketball might be used in fantasy basketball to create better virtual teams. In fantasy basketball, contestants have a budget constraint to create virtual teams of a predefined number of players. If our predictions are combined with an optimization model, it can be useful as a decision support tool for maximizing fantasy points and observing the best possible team combinations. Furthermore, there also exists betting markets that give opportunities for gamblers to make money by betting on performances of single players in various basketball leagues such as the Euroleague and the NBA. The predictions of our models can directly be used to select the best values from these types of betting odds. Last but not least, if we compare the summation of performances of all players in two competing teams, it is possible to get insights about the game result (Wheeler, 2012). Therefore, our feature engineering approach and predictive models can also be used in game result betting which constitutes a substantial percentage of betting odds provided by betting markets.

Additionally, there are now YouTube channels that are dedicated only to the player performances in basketball. These YouTube channels have a considerable number of subscribers from all around the world. For example, “FreeDawkins” and “House of Highlights” are the most popular YouTube channels that offer daily highlights and performances for the NBA players. Both of these channels have more than 1 million subscribers and get an average of 30 million views per month. The demand for channels that are providing highlights of the basketball players is increasing due to the fact that people do not have time to watch all of the games during a season. However, most of the basketball fans are interested in watching the games in which their favorite player performs well. Predicting performances of the

players and providing these predictions to the basketball fans before the games started have the possibility to increase the number of viewers of the live games. Every basketball league in the world is seeking new ways to increase their TV ratings and viewership in order to make more profit from broadcasting. Therefore, our feature engineering approach to generate more accurate estimates for player performances in basketball might help leagues and broadcasters to increase their viewership when they advertise the games through potential best performers.

All in all, coaches, gamblers, and basketball fans are the main stakeholders who can benefit from player performance predictions. In this respect, our feature engineering approach and resulting high-accuracy predictive models have many mentioned and unmentioned applications in terms of these stakeholders.

6.3 Limitations and future work

The findings reported in this study shed new light on player performance prediction in basketball. However, there are several weaknesses in this study that should be addressed in future research.

First of all, this study is conducted on the features that are derived from play-by-play, boxscore and player information data sets. Franks et al. (2015) demonstrated that the usage of player tracking data could provide more meaningful information regarding the defensive strengths and weaknesses of both teams and players. However, in Europe, optical tracking systems are still not used for collecting the spatiotemporal data of players. Therefore, our predictive models are limited by the lack of features from the potential information that can be derived from player tracking data. Future research should be conducted using player tracking data

because it would help predictive models to establish a higher degree of accuracy on this matter.

Secondly, on a regular game night, coaches tend to stick to their player rotations meaning that minutes played for a player does not tremendously fluctuate game to game. However, when one of the players in the team has to leave the game because of foul trouble or an injury, his playing minutes automatically decreases, whereas one of his teammates can have an opportunity to play more minutes than expected. Additionally, long-term injuries of the players in the main rotation may affect the playing times of all players in the team because of the inevitable changes in coaching strategies. Further research should be undertaken to understand the resulting effects of injuries and foul troubles in player performance.

Notwithstanding these limitations, the present study has been one of the first attempts to include data-driven defensive metrics for player performance prediction in basketball. Even though our feature engineering approach has brought new insights and understandings on player performance, our predictive models still have much more room for improvement. Future research on this topic should consider the main limitations of this study carefully.

REFERENCES

Alagappan, M. (2012). *From 5 to 13: Redefining the positions in basketball*. Paper presented at MIT Sloan Sports Analytics Conference, Boston, MA, USA. Retrieved from http://www.sloansportsconference.com/?p=5431

Alamar, B., & Mehrotra, V. (2011). Beyond 'Moneyball': The rapidly evolving world of sports analytics, Part I. *Analytics Magazine*. Retrieved from http://analytics-magazine.org/beyond-moneyball-the-rapidly-evolving-world-of-sports-analytics-part-i/

Arkes, J., & Martinez, J. (2011, 1). Finally, evidence for a momentum effect in the NBA. *Journal of Quantitative Analysis in Sports, 7(3)*, 13-13. doi:10.2202/1559-0410.1304

Berri, D. (2012, 1). Measuring performance in the National Basketball Association. *The Oxford Handbook of Sports Economics, 2*. doi:10.1093/oxfordhb/9780195387780.013.0006

Berri, D. J., & Schmidt, M. B. (2010). *Stumbling on wins: Two economists expose the pitfalls on the road to victory in professional sports.* Upper Saddle River, NJ: FT Press.

Berri, D., Schmidt, M., & Brook, S. (2006, 1). *The wages of wins: Taking measure of the many myths in modern sport.* Chicago: University of Chicago Press.

Brown, S. (2017, 3 31). *A pagerank model for player performance assessment in basketball, soccer and hockey*. Paper presented at MIT Sloan Sports Analytics Conference, Boston, MA, USA. Retrieved from https://arxiv.org/abs/1704.00583

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics, 15(1)*, 27-33. doi:https://doi.org/10.1016/j.aci.2017.09.005

Cao, C. (2012). Sports data mining technology used in basketball outcome prediction. Retrieved from https://arrow.dit.ie/scschcomdis/39/

Casals, M., & Martinez, A. J. (2013). Modelling player performance in basketball through mixed models. *International Journal of Performance Analysis in Sport, 13(1)*, 64-82.

Cervone, D., D'Amour, A., Bornn, L., & Goldsberry, K. (2016). A multiresolution stochastic process model for predicting basketball possession outcomes. *Journal of the American Statistical Association, 111(514)*, 585-599. doi:10.1080/01621459.2016.11416857

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system.* Paper presented at the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA. doi:10.1145/2939672.2939785

Cokins, G., DeGrange, W., Chambal, S., & Walker, R. (2016, 6). Sports analytics taxonomy, V1.0. Retrieved from https://www. informs.org/ORMS-Today/Public-Articles/June-Volume-43-Number-3/ Sports-analytics-taxonomy-V1.0.

Ding, C., & He, X. (2004). *K-means clustering via principal component analysis.* Paper presented at the 21st International Conference on Machine Learning (pp. 29). New York, NY, USA: ACM. doi:10.1145/1015330.1015408.

Dong, G., & Liu, H. (2018). *Feature engineering for machine learning and data analytics.* Boca Raton, FL: CRC Press.

Dua, D., & Graff, C. (2017). UCI machine learning repository. Retrieved from http://archive.ics.uci.edu/ml

Feng, G., G. Polson, N., & Xu, J. (2016, 4). The market for English Premier League (EPL) odds. *Journal of Quantitative Analysis in Sports, 12(4).* doi:10.1515/jqas-2016-0039

Franks, A., D'Amour, A., Cervone, D., & Bornn, L. (2016, 9 30). *Meta-Analytics: Tools for understanding the statistical properties of sports metrics.* Paper presented at MIT Sloan Sports Analytics Conference, Boston, MA, USA. Retrieved from https://arxiv.org/abs/1609.09830

Franks, A., Miller, A., Bornn, L., & Goldsberry, K. (2015). *Counterpoints: Advanced defensive metrics for NBA basketball.* Paper presented at MIT Sloan Sports Analytics Conference, Boston, MA, USA. Retrieved from http://www.sloansportsconference.com/content/counterpoints-advanced-defensive-metrics-for-nba-basketball/

Friedman, J. (2001, 10). Greedy function approximation: A gradient boosting Machine. *The Annals of Statistics, 29(5),* 1189-1232. doi:10.2307/2699986

Haghighat, M., Rastegari, H., & Nourafza, N. (2013). A review of data mining techniques for result prediction in sports. *Advances in Computer Science: An International Journal, 2(5),* 7-12.

Hollinger, J. (2002). *Pro basketball prospectus, 2002*. Dulles, VA: Brassey's Inc.

Hughes, M., & Franks, I. M. (2004). *Notational analysis of sport: Systems for better coaching and performance in sport.* London: Routledge.

Hwang, D. (2012). *Forecasting NBA player performance using a Weibull-Gamma statistical timing model.* Paper presented at MIT Sloan Sports Analytics Conference, Boston, MA, USA. Retrieved from http://www.sloansportsconference.com/wp-content/uploads/2012/02/46-Forecasting-NBA-Player-Performance_DouglasHwang.pdf

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 3149-3157)*. USA: Curran Associates Inc. Retrieved from http://dl.acm.org/citation.cfm?id=3294996.3295074

Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D. T. (2007). A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sport s, 3(3),* 1-24. doi:10.2202/1559-0410.1070

Kumar, S., Kumar, D., & Ali, R. (2012). Factor analysis using two stages neural network architecture. *International Journal of Machine Learning and Computing, 2(6),* 860–863. doi:10.7763/ijmlc.2012.v2.253

Loeffelholz, B., Bednar, E., & Bauer, K. W. (2009). Predicting NBA games using neural networks. *Journal of Quantitative Analysis in Sports, 5(1),* 1-17.

Lutz, D. (2012). *A cluster analysis of NBA players.* Paper presented at MIT Sloan Sports Analytics Conference, Boston, MA, USA. Retrieved from http://www.sloansportsconference.com/wp-content/uploads/2012/02/44-Lutz_cluster_analysis_NBA.pdf

McNamara, C. (2018, 5). Xgboost vs Catboost vs Lightgbm: Which is best for price prediction? Retrieved from https://blog.griddynamics.com/ xgboost-vs-catboost-vs-lightgbm-which-is-best-for-price-prediction/.

Miljković, D., Gajić, L., Kovačević, A., & Konjović, Z. (2010, 9). The use of data mining for basketball matches outcomes prediction. *Proceedings of the IEEE 8th International Symposium on Intelligent Systems and Informatics, (pp. 309-312).* doi:10.1109/SISY.2010.5647440

Oliver, D. (2004). *Basketball on paper: rules and tools for performance analysis.* Washington, D.C: Potomac Books.

Piette, J., Pham, L., & Anand, S. (2011). *Evaluating basketball player performance via statistical network modeling.* Paper presented at MIT Sloan Sports Analytics Conference, Boston, MA, USA. Retrieved from http://www.sloansportsconference.com/wp-content/uploads/2011/08/Evaluating-Basketball-Player-Performance-via-Statistical-Network-Modeling.pdf

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Proceedings of the 32nd International Conference on Neural Information Processing Systems (pp. 6639-6649).* USA: Curran Associates Inc. Retrieved from http://dl.acm.org/citation.cfm?id=3327757.3327770

QYResearchGroup. (2019). Global fantasy sports market size, status and forecast 2019-2025. Retrieved from https://www.marketwatch.com.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20(1),* 53-65.

Safir, J. (2015). How analytics, big data, and technology have impacted basketball's quest to maximize efficiency and optimization. *Senior Capstone Projects (Paper 390)* Retrieved from https://digitalwindow.vassar.edu/senior_capstone/.

Salador, K. (2011). *Forecasting performance of international players in the NBA.* Paper presented at MIT Sloan Sports Analytics Conference, Boston, MA, USA. Retrieved from http://www.sloansportsconference.com/wp-content/uploads/2011/08/Forecasting-Performance-of-International-Players-in-the-NBA.pdf

Shearer, C. (2000, 1). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing, 5(4),* 13-22.

Sindik, J. (2015). Performance indicators of the top basketball players: relations with several variables. *Collegium antropologicum, 39(3),* 617-624.

Singh, P. K., & Ahmad, M. (2015). Performance prediction of players in sports league matches. *International Journal of Science and Research (ISJR), 4(4),* 2207-2213.

Wang, K.-C., & Zemel, R. (2016). *Classifying NBA offensive plays using neural networks.* Paper presented at MIT Sloan Sports Analytics Conference, Boston, MA, USA Retrieved from http://www.sloansportsconference.com/wp-content/uploads/2016/02/1536-Classifying-NBA-Offensive-Plays-Using-Neural-Networks.pdf

Wheeler, K. (2012). Predicting NBA player performance. Retrieved from http://cs229.stanford.edu/proj2012/Wheeler-PredictingNBAPlayerPerformance.pdf.

Zdravevski, E., & Kulakov, A. (2010). System for Prediction of the Winner in a Sports Game. In D. Davcev, & J. M. Gómez (Ed.), *ICT Innovations 2009* (pp. 55-63). Berlin: Springer Berlin Heidelberg.

Zhang, G. P. (2000, 11). Neural networks for classification: a survey. *Part C (Applications and Reviews) IEEE Transactions on Systems, Man, and Cybernetics, 30*, 451-462. doi:10.1109/5326.897072