INTERNATIONAL
HELLENIC
UNIVERSITY

Basketball Analytics for Prediction of Performance during the last minutes of a game

Dimitris Gerakas

SID: 3308200009

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

SPRING 2023

THESSALONIKI – GREECE

Basketball Analytics for last minutes Performance

Dimitris Gerakas

SID: 3308200009

Supervisor:	Assoc. Prof. Christos Tjortjis
Supervising Committee Members:	Assoc. Prof. Koukaras Assist. Prof. Berberidis

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

SPRING 2023

THESSALONIKI – GREECE

Abstract

This dissertation was written as a part of the MSc in Data Science at the International Hellenic University.

Statistics and traditional data analysis were always a big part of the sports industry. Due to the nature of the field, analysts could always define some clear (distinct) numbers that would determine various relevant questions (whether a team performed well or if a particular player had an exemplary performance etc.). With the data revolution of the last decade, the use of sports data skyrocketed, while the arrival of Big Data and the appearance of the term data analytics have led to a definite change in regards with how the information gain is achieved and how the decision making process is carried out. It is extremely difficult (if even possible) to find a sector in sports that has not been influenced by the abundance of data during the last years. Training models that had been around for decades have been challenged as the performance of the athletes can now be monitored and measured in high detail; coaching principles were thrown, while new patterns have emerged.

The main objective of this dissertation is to examine the last-minute statistics in the NBA, that is to explore which choices lead to better results, which decisions have the greatest impact during the last critical plays of a game and to explore which features of the individual performance have the highest impact to the winning chances of a team during these moments. A variety of models and algorithms were utilized and the results were thoroughly compared and analyzed.

Acknowledgements

I would like to thank professor Christos Tjortjis for his guidance, support and patience throughout the duration of my studies.

I would also like to thank Vaggelis Sarlis for his useful insights and recommendations regarding my dissertation.

Dimitris Gerakas
11/06/2023

Contents

1 Introduction

In this introductory chapter, a general description of the problem and the motivation to engage it will be presented, as well as a detailed structure of the dissertation will be provided. We will offer basic information about the task and scope of our project, and establish the foundation of our thesis. The reasons that incentivized us to work on this study will be analyzed, and previous work that inspired us will be mentioned.

1.1 Problem – Motivation

In the game of Basketball and specifically in the NBA, two teams of 5 (active) players compete against each other for 48 minutes to determine the winner. Each offensive possession can last no more than 24 seconds, a characteristic that gives a fast-paced rhythm to the game. On average, an NBA team has around 100 possessions of the ball per game [1], during which the offensive team can make baskets and thus increase the overall team score. Despite the fact that the every basket counts the same regardless the time it was made, there is a certain increased weight for the possessions during the last minutes of the game, when the score is close and the outcome uncertain. In basketball, the term “clutch” is used to refer a player’s ability to perform well under pressure, when a game is on the line and the result undetermined.

Instances of such situations occur during the final minutes of a close match, characteristically in the 4th quarter or overtime, and include scoring a game-winning basket, making a successful defensive play or more generally making correct decisions under pressure.

Theoretically, a clutch player demonstrates more than sheer skill; it requires mental qualities such as confidence, resilience, experience and/or responsibility. Confidence that he can make the proper decisions in tough situations, mental resilience means that the player can keep his composure under adverse circumstances, experience in dealing with critical situations can only improve the choice selection and finally that the player is not afraid to take responsibility for losing or making a crucial play. Naturally, these mental capacities cannot be measured. How can we count the composure of a player in critical moments, or how to measure his confidence?

The purpose of this thesis is twofold: 1) to use data mining techniques to determine the most important factors that contribute to a winning result in a close game and 2) To use the results of the first question as well as relevant literature and to define new metrics in order to rank the best clutch performers in NBA during the last two decades. For the purpose of this study relevant statistics of NBA player during the seasons 1997 to 2018 were obtained.

One of the main reasons that incentivized us to tackle this analysis is the relative shortage of scientific literature regarding the clutch performance of basketball players. A fact that naturally is not unfounded: the scarcity of useful data in this field is the underlying factor. Even popular and widely used player efficiency ratings (PER) are difficult to implement given the scope of the last minute statistics, while more specific and complicated stats (like Value Over Replacement Player (VORP) and Usage Percentage (USG%)) have not or cannot be measured during the relevant time. Nevertheless, we tried to establish our work on the foundations of a proper statistical approach and derive results that were not only extremely relevant to the questions at hand but also logical according to the corresponding analytical NBA environment.

1.2 Thesis Structure

The dissertation is structured into seven chapters.

- Chapter 1 is the introduction,
- Chapter 2 contains the literature review, with numerous mentions of previous work in relative fields in the domain of sports analytics,
- Chapter 3 provides a theoretical background in general data mining techniques, as well as more specific methods that were implemented in this particular study,
- Chapter 4 includes the procedures directly related to data acquisition, data cleaning, preprocessing and processing. In this part there is also detailed mention of the feature extraction techniques that constitute an important part of the study questions,
- Chapter 5 concentrates in the feature selection techniques as well as modeling with tuning of weights, hyper-parameters, datasets and algorithms,

- Chapter 6 implements the EoCC (Estimation of Clutch Competency), which is a newly defined performance metric adapted entirely in the research question, and statistically grounded on machine learning insights from chapter 5,
- Chapter 7 determines the conclusion of the project, as well as what could be done differently and recommendations for future work.

2 Literature Review

In this section we will engage ourselves in a literature review concerning various articles and studies that revolve around the general topic of sports analytics. We will discover and analyze results and conclusions of scientific papers, algorithms that have been used with or without success in order to meaningfully translate sports data, and even compare some similar researches and their deductions. In the first part, we will discuss how sports analytics have affected a variety of sports, while in the second part we will focus on the game of basketball and the NBA and its relevant literature.

2.1 Sports Analytics

In this segment, scientific literature will be presented concerning the general field of sports analytics and how researchers utilized data to scrutinize their respective domain. We will examine several scientific papers from a variety of sports that have used data analytics to investigate popular questions and draw meaningful conclusions.

2.1.1 Data Analytics in anthropometric measurements for sports

One of the most popular topics in sports during the latest years is the comparison between genetics and training. Although the notion that there are more factors than training and sports intelligence for determining the general capability of a player existed for a long time, no one could answer with scientific certainty why e.g. black people dominate in NBA, what makes them on average more suitable for this sport, or e.g. what is the most important quality for a football player, what is that specific quality that can pre-determine that this individual has a really strong potential to become a great football star. With the emergence of analytics, researchers were given a powerful tool to tackle issues such as these. There are databases with the physical characteristics of players and data mining can assist in deciding whether the anthropometric measurements can distinguish the potential of players in a variety of sports. Although these metrics was initially just a tool for coaches and managers to scout effectively young talent (which means that everything was still reliant on the human proficiency [2]), as the volume of the data grew and more data-based techniques appeared, the algorithms evolved order to be able to predict the potential of young athletes (or at least provide useful information) on their own – solely based on the related data.

Researchers from Sri Lanka [3] conducted a study on boys aged from 18 to 28 years, where they implemented the Spherical Associated Keyword Space (SASKS) algorithm which would assign them for a suitable sport according to their anthropometric measurements. The relevant sports were Tennis, Kabadi and Koko while there were 30 attributes that were measured including Height, Weight, Flexibility, Chest Girth, Wrist Girth, etc. The popular min-max Normalization technique was applied on the values of the attributes and the SASKS algorithm was chosen as the clustering approach. Briefly, the SASKS algorithm plots the athletes onto a 2-Dimensional spherical surface [3] with an affinity matrix as input, which is built and computed based on the Euclidean distance between the measurements of the players:

$$ED(P_i - P_j) = \sqrt{(\sum A_r(P_i) - A_r(P_j)))}$$

P_i is the i^{th} player
$A_r(P_i)$ is the r^{th} attribute of the P_i player

Using the above technique the researchers represented how similar are the athletes in 2D space or in other words created clusters of athletes of similar measurements, which would indicate that the athletes are suitable for the same sport. As we can observe in Figure 1, there seems to be three main clusters, every one of which corresponds to one of the relevant sports, while, naturally, players closer in this surface have more similarities in their measurements.

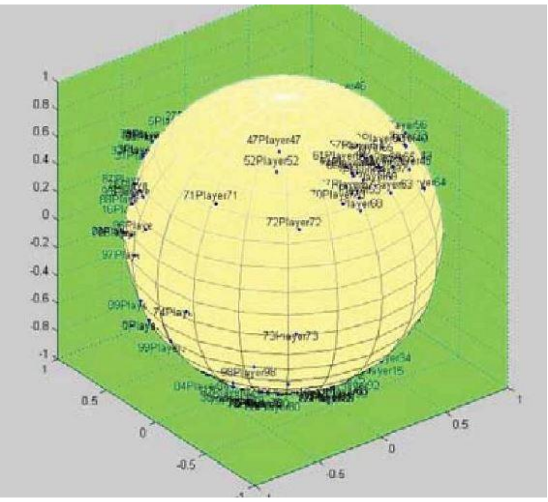


Figure 1

After the clustering process, manual categorization and anthropometric measurements were compared for each individual cluster in order to identify possible invalid members. Conclusively, there were indeed a couple of athletes placed incorrectly in each group, with the authors [3] mentioning that the accuracy might increase by reducing the number of features to only those more fitting for each of the corresponding sports.

As competitive sports evolve and teams always search for the players with the greatest potential, studies such as these can assist scouts and managers in their work, while at the same time giving also the actual athletes a hint about which sport they would really excel on. A statistical study [4] in University of California investigated the correlation between body proportions and elite athletic success on four groups of individuals: NBA players, MMA fighters and army recruits (both female and male). This study looked for a relationship or a pattern between their achievements and their limb measurements using the Linear Regression algorithm as well as the ANOVA variations to pinpoint correlations among the variables and differences between the groups.

Linear Regression finds that Arm Span to Height Ratio can significantly predict the athletic success in both NBA players and MMA fighters (Figure 2). As we can observe, the NBA players with higher Arm Span to Height ratio are more probable to achieve a better draft pick (lower is better) and respectively the MMA fighters have better Win to Loss ratio (in Figure 2 the values are negative Loss to Win).

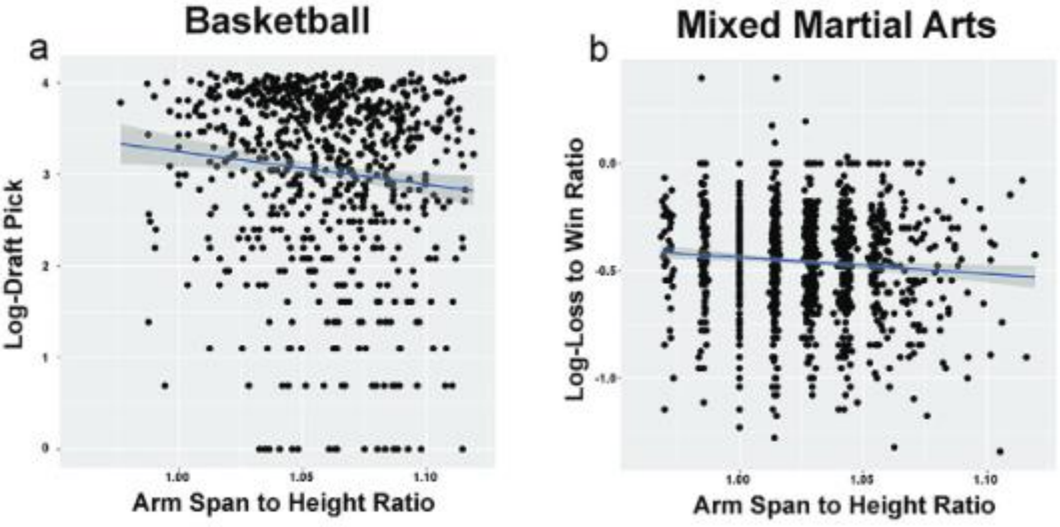


Figure 2

Another interesting point of the study [4] is: where does “the cream of the crop” stand in this classification? Can we relate Height to Arm Span Ratio to the most successful players of our generation? The following figure (Figure 3) provides a hint:

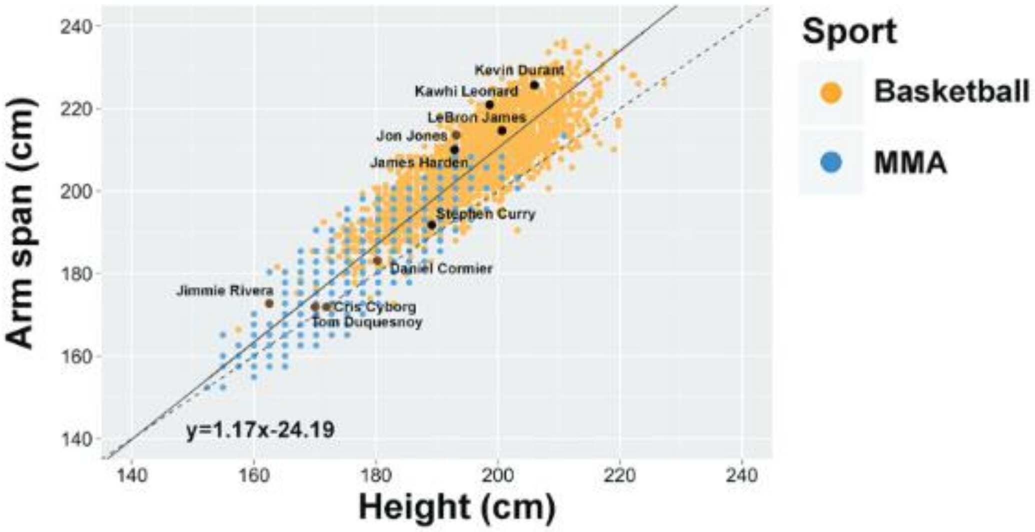


Figure 3

We can see that four out of the five top NBA players sit above the regression line; only Stephen Curry lies below, while regarding the MMA players the ratio is more balanced as two of the fighters are above and three below the regression line. Other than that, there is a clear correlation between arm span and height (R^2=0.85, p<0.001), which is expected.

For the most part, successful athletes tend to have higher arm span to height ratio, within the expected limits of variation of the human body though. That being said, this is just one of a multitude of factors (developmental, behavioral etc.) that dictate athletic success but studies like these seem to suggest that genetics play a bigger role in professional sports than previously thought.

2.1.2 Golf

An interesting idea, in which data mining has been used extensively, is clustering groups of similar items (athletes/movements/athletic actions) together. For instance in golf [5] there was an attempt to create diverse competitions in which players of different levels and playing styles would take place. Due to the nature of the sport though, this was not a straightforward process. To determine how to level both the golfers and identify specific styles of play cluster analysis was used. Eventually, the research [5] provided useful information to the coaches, mainly because it identified: five different movement phases in the hammer throw and three different swing styles based on the level of players. This was a valuable input for the coaches as they can use this feedback to improve the technique of their athletes in specific motions depending on the individual weaknesses.

2.1.3 American Football

K-means clustering was also used on a study [6] in Mississippi State University, where the researchers attempted to find a way to optimize student football training by pinpointing position-specific drills for the athletes. The data on which the authors worked was obtained by NCAA Division 1 American football team and it was used in order to create a set of the average demands of each position and subsequently create training programs suitable for these specific demands. Figure 4 illustrates the groupings for K ranging from 2 to 7.

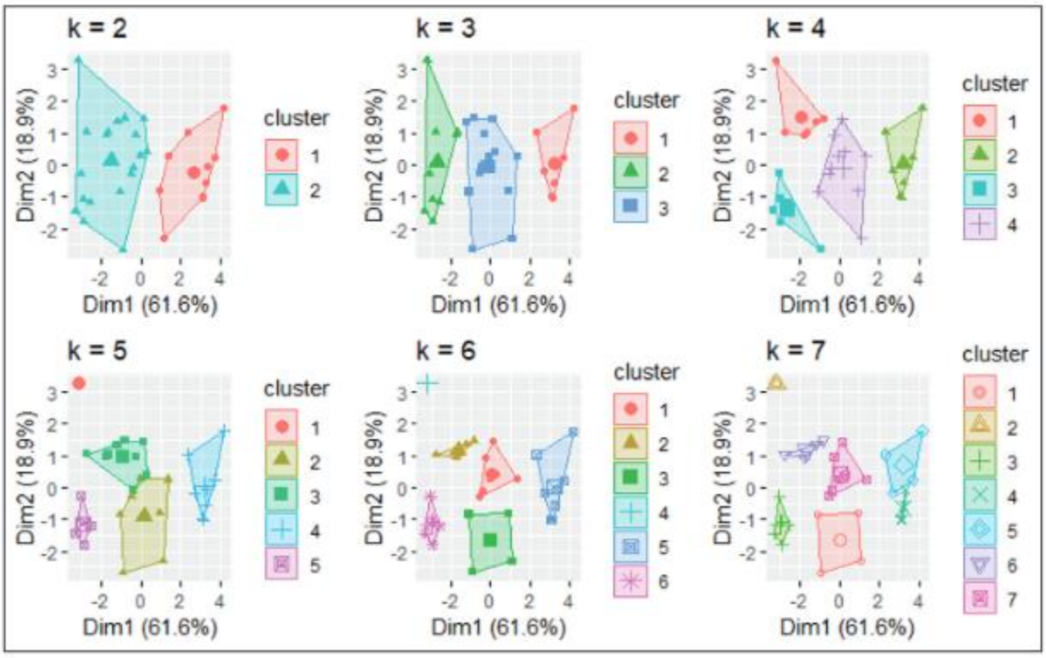


Figure 4

Nevertheless, the authors [6] wrote that k=3 provide the most efficient number of training groups. These clusters do not represent specific player positions but rather a combination of the demands of each position and the specific characteristics of a player. For instance, a strong linebacker that needs improvement in acceleration should be in a different training group than a fast linebacker, despite the fact that they play in the same position. Studies such as this can assist the coaches with targeted training drills for the players to address their needs without jeopardizing teamwork. As the authors mentioned, there is great potential for more detailed analysis, associating specific playstyles, roles and characteristics with the athletes. One major drawback of this technique was that while the data covered two whole seasons, the results for each season were quite different. Although this is not unexpected (a college team is not a stable constant, it changes year by year), it forces the interested parties to have a new model with new data every year. And since this data needs some time to be gathered, there is clearly lost time before the model is able to produce meaningful results. That being said, a more versatile system could be able to address this issue, but of course that requires further research.

2.1.4 Soccer

One of the most common uses of classification algorithms in sports is trying to predict the role of a player using as input his attributes. An experiment [7] conducted in Greece to test the effectiveness of a couple of algorithms used data from the game FIFA 18 and tried to classify the soccer players in one of the main 4 roles on the field (goalkeeper, defender, midfielder, forward). Two algorithms were applied for this task, namely the Random Forest and the Sequential Minimal Optimization (SMO). The results were encouraging as they had a success rate of around 80%. In this context it would be interesting to try to classify the players according only to their physical characteristics in order to provide some insight about the needs of each position. Afterwards, we could be able to predict the potential of young talents and how well they would fare in each position individually. An interesting attempt was made by the same author [7] to predict the number of goals that specific players will score based on historic data. The experiment was done for two of Barcelona's FC most popular stars, L. Messi and Neymar. With the help of data from previous years the researcher applied four algorithms – Random Forest, Logistic Regression, MLP classifier, Linear SVC (Support Vector Classifier) to predict how many times the aforementioned players will score during the year 2017-2018. For the first player (Figure 5) the random forest algorithm and MLP classifier predicted exactly the actual number (34). For the second player, only Random Forest and Logistic Regression fell close, (31 to 25), the other techniques were highly inaccurate. Conclusively, the Random Forest algorithm displayed the best results for this particular prediction. An interesting argument in the paper is that by predicting each player's stats individually and aggregating them, we can achieve better predictions for the team stats than by using historical team data. Unfortunately, it gets increasingly difficult to measure successfully the impact of new players in the teams and so the models need a lot of work to achieve a sufficient accuracy.

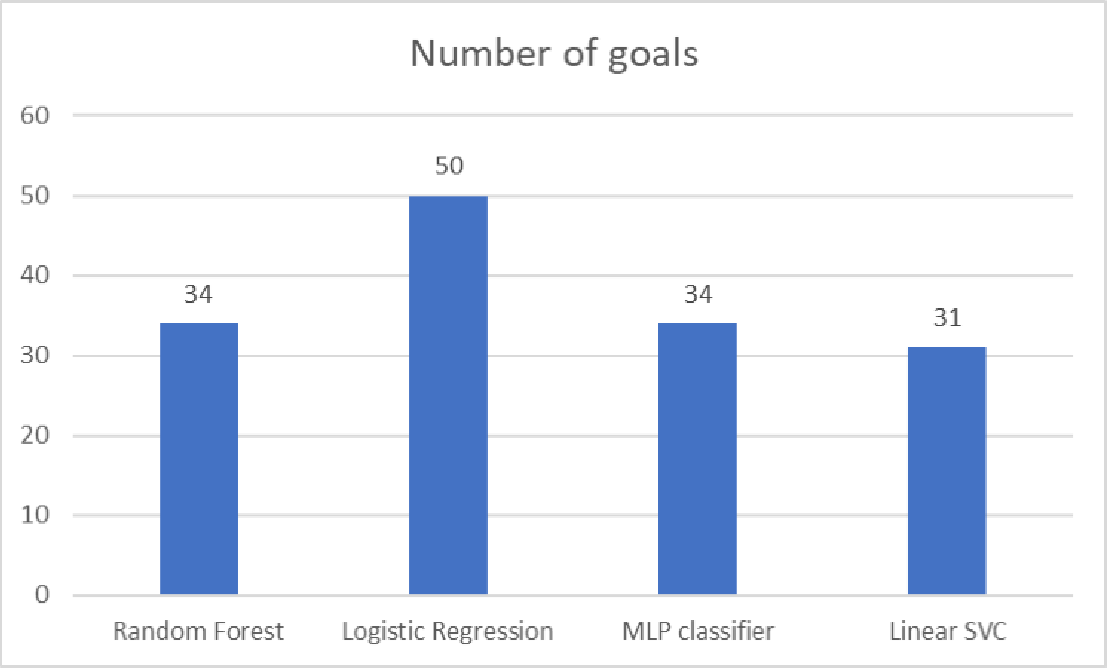
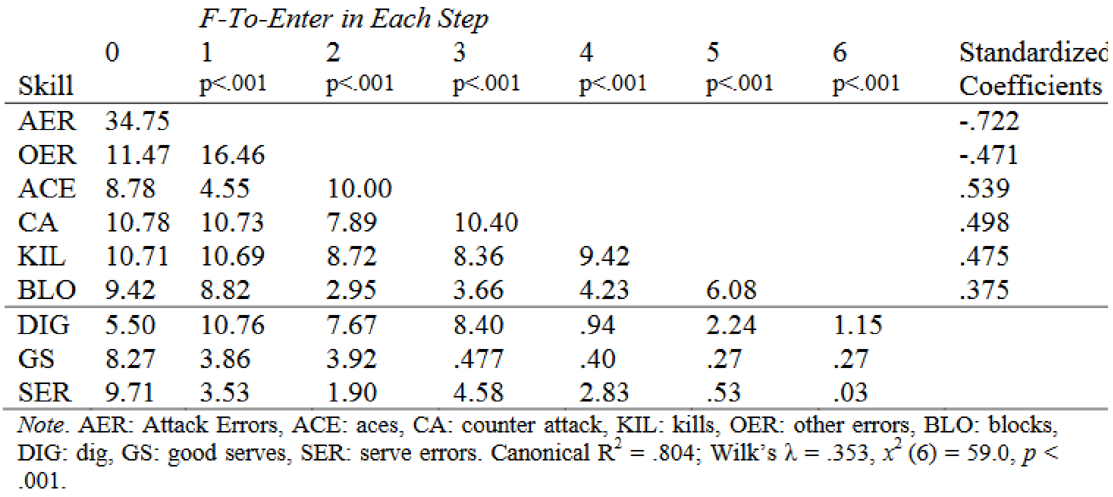


Figure 5

2.1.5 Volleyball

Although not as popular as other sports, volleyball had its share of researchers investigating data in order to cluster characteristics between winning and losing teams [8] or predict the outcome of specific games [9]. Researchers from AUTH [8] in Greece explored the most significant factors that contribute to a team winning or losing by analyzing data in FIVB men’s Beach Volleyball (BV) World Tour tournament. The relevant attributes included serve, attack, block and dig. They separated the data based on the final set score (2-0 or 2-1) and utilized independent t-tests to find correlations between the statistics and the final scores, as well as to define proper descriptive models. The general conclusion was that in cases of a 2-0 result, the winning teams surpassed the losing teams in all categories, while in cases of a 2-1 result, there was no clear pattern that could describe if there were specific statistics that could on average indicate a victory. Interestingly (and relevant to our own work) the authors continued their study with a discriminant analysis to determine the most significant parameters that contributed to the 2-0 results and as figure 6 illustrates, these were error attacks, other errors, aces, counters etc. in descending order as indicated by the magnitude of the standardized coefficients.



		F-To-Enter in Each Step						
	0	1	2	3	4	5	6	Standardized
Skill		p<.001	p<.001	p<.001	p<.001	p<.001	p<.001	Coefficients
AER	34.75							-.722
OER	11.47	16.46						-.471
ACE	8.78	4.55	10.00					.539
CA	10.78	10.73	7.89	10.40				.498
KIL	10.71	10.69	8.72	8.36	9.42			.475
BLO	9.42	8.82	2.95	3.66	4.23	6.08		.375
DIG	5.50	10.76	7.67	8.40	.94	2.24	1.15	
GS	8.27	3.86	3.92	.477	.40	.27	.27	
SER	9.71	3.53	1.90	4.58	2.83	.53	.03	

Note. AER: Attack Errors, ACE: aces, CA: counter attack, KIL: kills, OER: other errors, BLO: blocks, DIG: dig, GS: good serves, SER: serve errors. Canonical R^2 = .804; Wilk's λ = .353, x^2 (6) = 59.0, p < .001.

Figure 6

In another study [9], the researchers specified two different models that predicted the rankings of the women's volleyball Italian Serie A1 2017-2018 season. Their models were based on Bayesian hierarchical model that tried to estimate the outcomes of the individual games. The relevant attributes for the analysis were the teams, points, sets, serves, defense, attack and blocks. Their models initially categorized the teams into four clusters based on their stats (with greater significance on attack and defense) and subsequently tuned the parameters in order to achieve the best results on the given data. Conclusively, their Bayesian models were quite accurate with only small discrepancies regarding the individual match result. There were only two teams for which the model overestimated their performance in comparison with the actual outcomes. As far as the overall ratings of the teams in the championship, their models agreed primarily on the general placing of each team with variations that did not exceed 7% [9].

2.2 Basketball Analytics

In this part of the thesis, we will discuss literature directly related to basketball and analytics. We will investigate techniques that have been already used to determine performance of basketball players and teams, we will analyze different rating methods that were proposed and finally debate whether or not they are applicable for our own project.

This section is further divided into player analytics, team analytics and betting. In the player analytics subsection we will focus on literature directly related to the personal statistics, while in the team subsection, the literature includes cases where the researchers employed analytics on team performance to investigate their topics.

2.2.1 SportsVU Equipment

Before starting with actual Basketball Analytics, a special mention has to take place concerning the SportVU equipment. It is a tracking system based on cameras that is able to capture data many times per second and has revolutionized the statistical side of the game. The information that can be gathered from the system contains player speed, distance covered, acceleration, shot trajectories, passing and even more advanced statistics. It was one of the most important factors that have led the NBA teams to adopt the data-driven decision making strategy.

2.2.2 Performance Evaluation and Performance Prediction

In the first section of this part we will engage literature relevant to individual player statistics, while in the second we will discuss papers that have to do with team statistics. Performance evaluation contains the combination of statistics and metrics to describe how impactful a player is for their team, and how his stats have assisted his team in winning, while performance prediction is a machine learning task, which using historical data tries to predict the outcome of certain events in a game under specific circumstances.

2.2.2.1 Player Statistics

NBA is known to be one of the most illustrious domains of sports and due to the abundance of data there was always interest in analytical processes that could help franchises improve their decision making either on or off field. Regarding the on-field performance there are more than 20 metrics [10] that can evaluate different performance indicators and judge how well a player or a team is doing. These metrics extend from the old, traditional Tendex value (one of the first models to calculate player efficiency) to the relatively fresh and popular APM-Adjusted Plus-Minus value, which calculates the impact of a player's presence or absence during the game taking into consideration both teammates and opponents [11]. The authors [10] did a thorough research evaluating dozens of metrics and often combining them to produce meaningful results. The following figure [Figure 7] shows a radial chart with the performance metrics (with logarithmic normalization) of five of the most efficient players for the year 2018 – 2019:

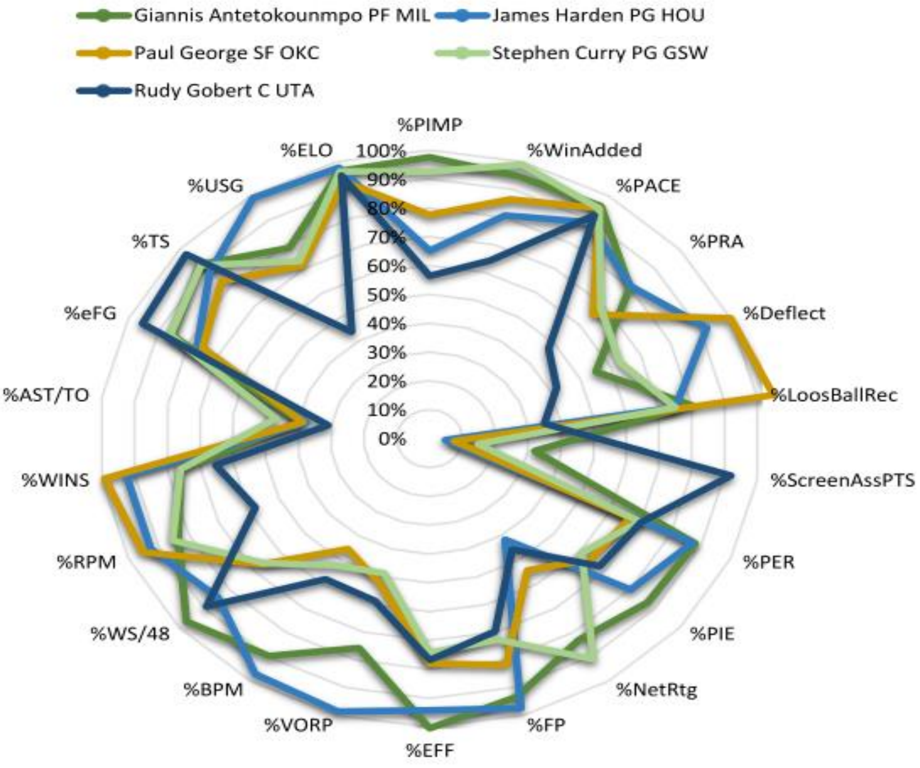
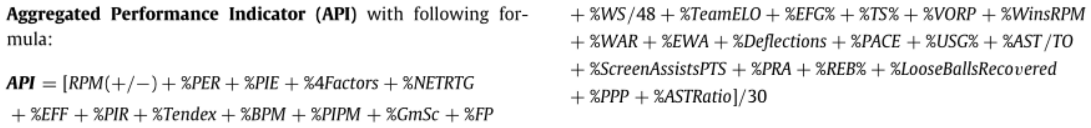


Figure 7

Some of the most notable observations are the defensive ratings of Paul George (%Loose Ball Recovery and %Deflection), which are considered also "really big plays" [10], as they have the potential to give a psychological boost to the team, and the overall values of %TrueShooting and %effectiveFieldGoal of all the players which indicate the offensive efficiency of the NBA stars.

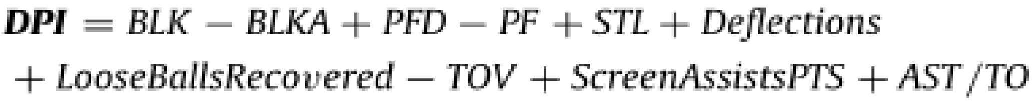


Aggregated Performance Indicator (API) with following formula:

$$API = [RPM(+/-) + \%PER + \%PIE + \%4Factors + \%NETRTG + \%EFF + \%PIR + \%Tendex + \%BPM + \%PIPM + \%GmSc + \%FP + \%WS/48 + \%TeamELO + \%EFG\% + \%TS\% + \%VORP + \%WinsRPM + \%WAR + \%EWA + \%Deflections + \%PACE + \%USG\% + \%AST/TO + \%ScreenAssistsPTS + \%PRA + \%REB\% + \%LooseBallsRecovered + \%PPP + \%ASTRatio]/30$$

Figure 8

Moreover, in the same study there was a forecasting scenario implemented in order to predict the MVP and DPOY (Defensive Player Of the Year) according to a number of formulas. The one responsible for finding the next MVP was named API (Aggregated Performance Indicator) and is illustrated in Figure 8, while the other was named DPI (Defensive Performance Indicator) and was built to predict the DPOY and is shown in Figure 9:



$$DPI = BLK - BLKA + PFD - PF + STL + Deflections + LooseBallsRecovered - TOV + ScreenAssistsPTS + AST/TO$$

Figure 9

All in all, the formulas proved to be semi - successful as they managed to predict accurately the MVP of the season 2019-2020 but failed to do so for the DPOY (G. Antetokoumpo won both awards while DPI formula suggested that R. Gobert will be nominated DPOY). Other formulas were also tested for accuracy on predicting the MVP but either they failed to do so, or they were dependent on historical data while API needed only current stats. Of course, basketball is a complicated team sport where even finding the most valuable player in the league may be dependent in external factors like the current social sympathy, the worth of a franchise, the market shares or the betting companies.

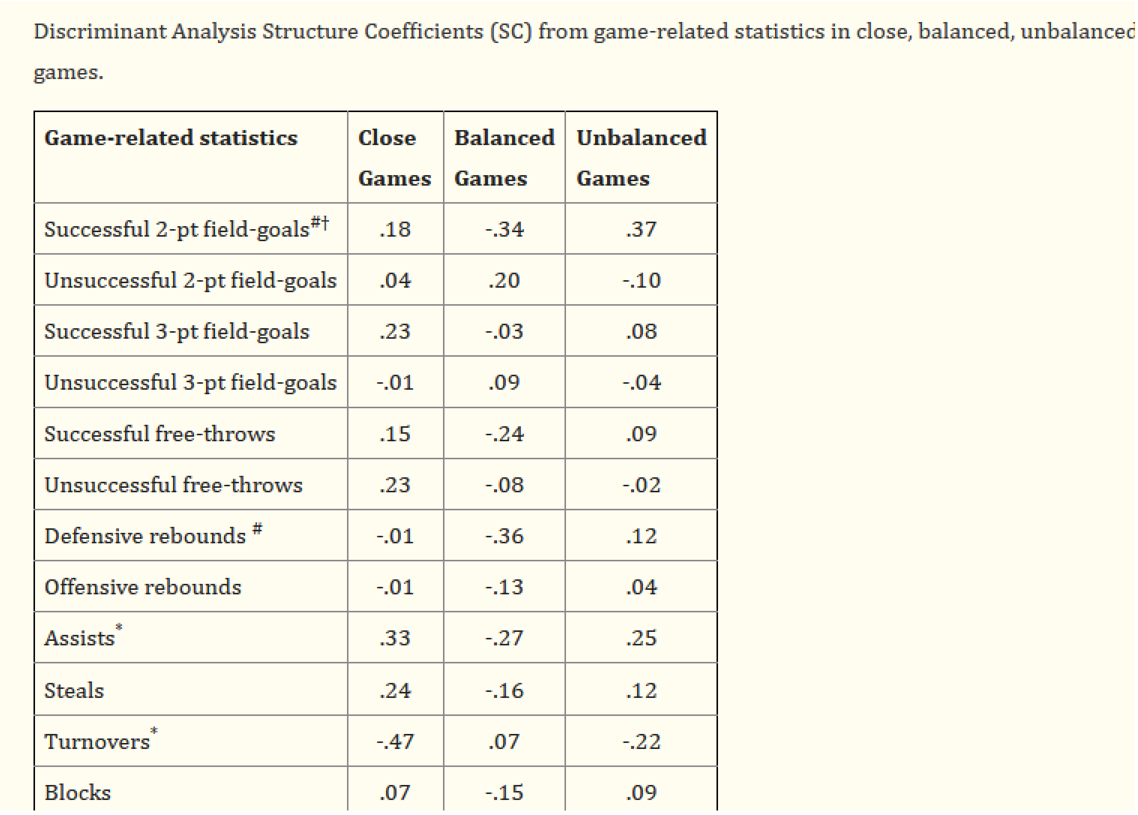
Another characteristic property of the game of Basketball is that despite the previously mentioned SportVU system and the technological breakthroughs of the last decade it is still quite hard to quantify certain aspects of the game, especially on defense [16]. This difficulty was noticed by researchers in MacEwan University in Edmonton [13], who tried to tackle it using data mining and complex metrics. J. Hollinger, a former NBA analyst and current Vice President of Operations for Memphis Grizzlies developed a rating to determine the players and team's efficiency on the court, called the Player Efficiency Rating (PER) [13]. It basically adds all the positive stats (field goals, steals, blocks) of a player on the court while it subtracts the negative ones (turnovers, fouls). There were several observed drawbacks of this rating; some players are known to artificially increase their stats without some clear benefit for their team, while in some other cases the defensive hustle is hard to quantify. A characteristic example for cases like these comes from K. Goldsberry [14], who named the phenomenon the Dwight Effect (from NBA player Dwight Howard) and mentioned in his paper that while some other defenders may have had better defensive stats in paper (blocks etc.), the offensive players when guarded by Howard were 15% less probable to shoot from within 5 feet of the basket, which is translated into a big difference in field goal accuracy. This is one quite difficult statistic to find and even harder to quantify. Even more ideas were proposed in order to tackle this problem, like computing the percentage of contests in shots by the defenders [15]. This was an approach to measure the on-game impact of the defenders even when it does not translate into a steal or a block.

In the same context, Franks et al. [12] used spatial data from the SportVU system to try and quantify the defensive performance of individual players. For their study, they introduced a new model that is primarily dependent on 2 factors: the shot frequency and shot efficiency of the opposite players. For instance, a competent defender may either discourage his 1v1 matchup to shoot the ball (frequency reduction), or he may force more misses if his opponent do make the shot attempt (efficiency reduction). The experimental results found that e.g. Roy Hibbert is very good at challenging shots and the opponents tend to miss more shots inside the paint when he is on the defending team, while D. Howard (as other studies demonstrated too [14]) is top in not allowing opponents to shoot around the space he is defending but when they do shoot, they tend to have better accuracy than against other defenders. Interestingly, the results uncovered even other patterns, not initially obtainable from the data. In individual cases, the experiment detected couples of 1v1 matchups of particular interest. For example, the data showed that Lebron James was expected to score characteristically fewer points when defended by K. Leonard. All in all, studies like this investigate performance statistics that are very hard to measure and although they should always be considered on the relevant context, they can definitely assist the NBA managers and coaches in the decision making procedure.

2.2.2.2 Team Statistics

Another very popular topic among sports scientists is to distinguish the most important factors that contribute to a team's success.

In an analysis that was based on data from Under-16 European male Championship teams [17], the researchers employed discriminant analysis to determine the most significant attributes that dictated the final result. To explore the topic even wider, they firstly separated the individual matches into 3 clusters based on the corresponding difference in the final score. One cluster included the close games (under 9 points difference), another included the balanced games (10-29) and a final one the blowouts (difference larger than 29 points). Interestingly, the results of each cluster were different; in the close games the most significant attributes were the assists and the turnovers, while in balanced games the defensive rebounds and the 2-point field goals were the defining factors. Finally, in unbalanced games, only the accurate 2-point field goals discriminated between the winning and the losing team. As it turns out, there was no common dominant statistic that dictated the results among all of the three clusters [Figure 10].

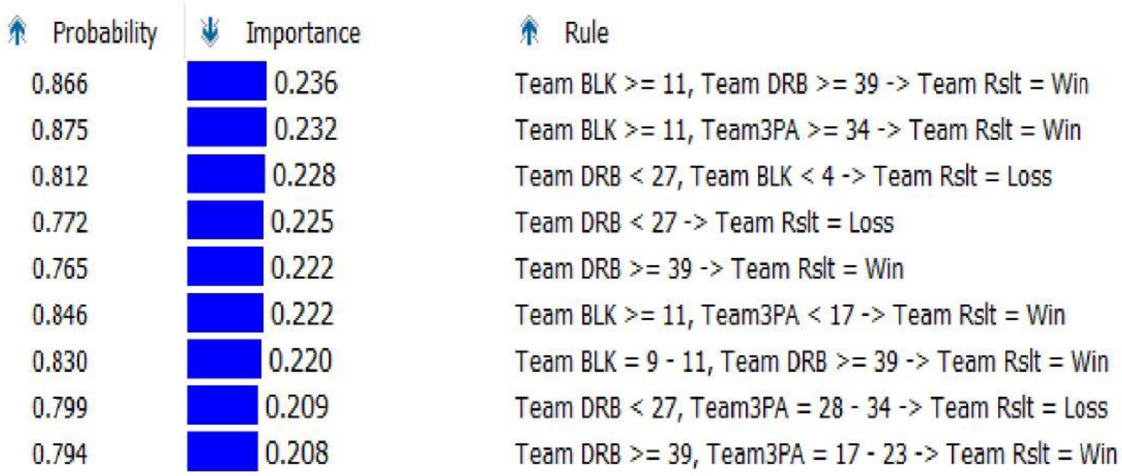


Discriminant Analysis Structure Coefficients (SC) from game-related statistics in close, balanced, unbalanced games.

Game-related statistics	Close Games	Balanced Games	Unbalanced Games
Successful 2-pt field-goals#†	.18	-.34	.37
Unsuccessful 2-pt field-goals	.04	.20	-.10
Successful 3-pt field-goals	.23	-.03	.08
Unsuccessful 3-pt field-goals	-.01	.09	-.04
Successful free-throws	.15	-.24	.09
Unsuccessful free-throws	.23	-.08	-.02
Defensive rebounds #	-.01	-.36	.12
Offensive rebounds	-.01	-.13	.04
Assists*	.33	-.27	.25
Steals	.24	-.16	.12
Turnovers*	-.47	.07	-.22
Blocks	.07	-.15	.09

Figure 10

In similar context, the researchers from Edmonton [13] tried to pinpoint the most important factors that can determine the victory in a game. They applied diverse analysis techniques: Association Rules Algorithm, Decision Trees and Neural Networks on a database that included the basic stats of all the games in NBA from 2012-2018. Also, a variety of software options was used like Python, Microsoft SQL Server Data Tools and Visual Studio either for cleaning or for processing the data. The Association Rules Algorithm showed that the most important factors that dictate the winner are the defensive rebounds, the blocks and the 3-pt attempted, as shown in Figure 11:



Probability	Importance	Rule
0.866	0.236	Team BLK >= 11, Team DRB >= 39 -> Team Rslt = Win
0.875	0.232	Team BLK >= 11, Team3PA >= 34 -> Team Rslt = Win
0.812	0.228	Team DRB < 27, Team BLK < 4 -> Team Rslt = Loss
0.772	0.225	Team DRB < 27 -> Team Rslt = Loss
0.765	0.222	Team DRB >= 39 -> Team Rslt = Win
0.846	0.222	Team BLK >= 11, Team3PA < 17 -> Team Rslt = Win
0.830	0.220	Team BLK = 9 - 11, Team DRB >= 39 -> Team Rslt = Win
0.799	0.209	Team DRB < 27, Team3PA = 28 - 34 -> Team Rslt = Loss
0.794	0.208	Team DRB >= 39, Team3PA = 17 - 23 -> Team Rslt = Win

Figure 11

As we can see, with high confidence and importance, if a team gathered around 40 defensive rebounds and managed to score more than 11 blocks, it was a definite favorite for the win. Furthermore, a team with high number of attempted 3-points had more probabilities to get the W, as long as there was at least one good defensive rating (either in blocks or rebounds). In general, of all the correlation statistics the study showed that the number of defensive rebounds has a direct relationship with the probabilities of victory. Of course even this point has many interpretations; a team with high DREB rating should have a good defensive system as to force the offence to miss their attempts, as well as, having good "box out" mechanisms to keep offence away from the offensive rebound.

Another clear trend that was investigated in [13] was the increase in 3-pt attempts overall in the NBA. Figure 12 shows the rise of 3-pt shots league wise during the years 2012-2018.

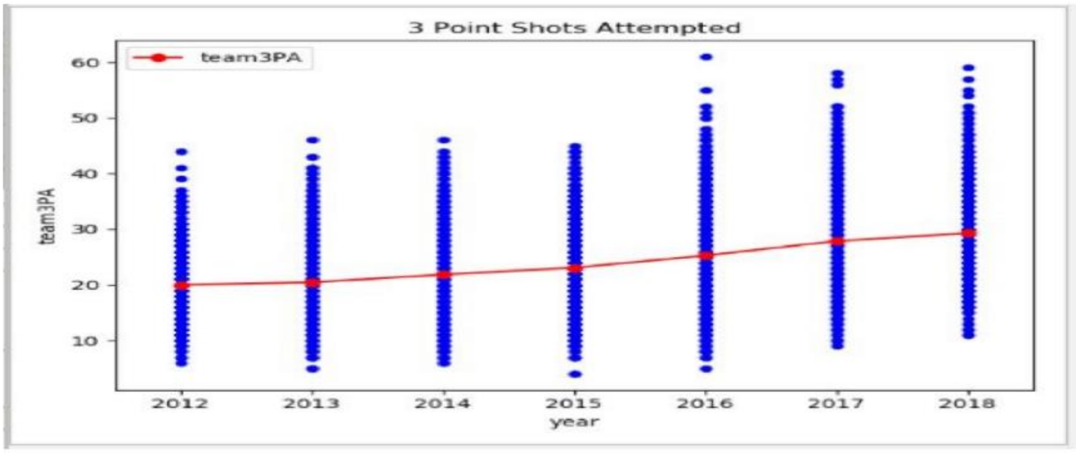
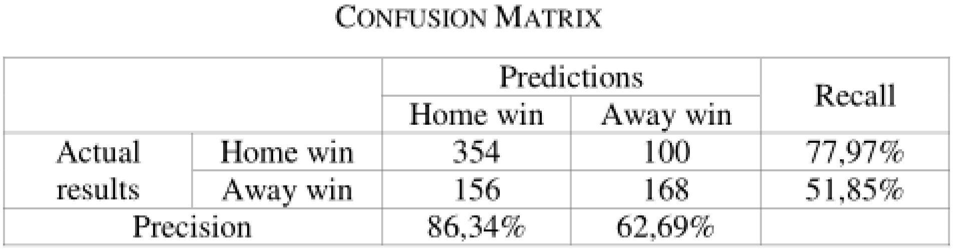


Figure 12

We can observe the clear ascending trend in the average numbers of 3-pt shooting throughout the years. The results of the study have shown that the teams prefer to take shots either from the 3-pt line or within 5 feet of the basket; these have been proved to be the most efficient plays. Generally, a lot of arguments can be made about the meaning of all the difference metrics in basketball, some of which will be analyzed in further sections of this paper.

2.2.3 Betting

Predicting scorers, champions and league leaders is the epicenter of the huge gambling community in sports. The betting industry is a large part of the sports industry. Millions of revenue is exchanging hands every day through the bets of fans and gamblers alike. Researchers [18] from Serbia tried to forecast the winner of NBA games by implementing the Naïve Bayes algorithm (since it produced the best results), combined with multivariate Linear Regression for the spread (a value that is added to the score of one of the teams in order to equalize the chances of victory for betting purposes). The researchers used in game stats for the model, as well as the corresponding standings of the year with the total sum of attributes to reach as high as 141. They used a variety of software options including Rapidminer, MySQL and JDBC. Naïve Bayes algorithm with 10-fold cross validation achieved a satisfactory 67% accuracy. On the contrary, the spread was predicted successfully only in 78 out of 778 matches (10% accuracy), but the authors notes that it was expected as it is very difficult to find the exact difference at which a game will end. Figure 13 shows the general results of the model:



CONFUSION MATRIX

		Predictions		Recall
		Home win	Away win	
Actual results	Home win	354	100	77,97%
	Away win	156	168	51,85%
Precision		86,34%	62,69%	

Figure 13

This type of forecast is quite difficult because during the regular season in NBA it is not always the case that the best team wins (or that it will use its stronger assets in every game). The teams are not always so focused on winning every time, as the advantage in the playoffs is trivial. Consequently, a model that tries to predict the whole season will not be very accurate. It would be probably quite easier to build an algorithm that predicts only specific games (e.g. a specific matchup or a specific team) as it will have higher accuracy and as a result higher gains for the betting community.

2.3 Clutch Performance in the NBA

Our focus in this study is to examine the performance of the players during the last critical moments of the game. This is called clutch performance. Although the existing bibliography is not numerous, and the term is not yet officially defined, several scientists attempted to unravel the mysteries behind this topic.

One of the most important characteristic of successful professional athletes is their ability to perform under pressure. Whether a sprinter is in the Olympics final in front of thousands of people, or making preliminary rounds in a local insignificant event, being a true professional he should be able to perform equally well. This is easier said than done: it requires experience, concentration, willpower and other mental capacities in order to be able to focus on the task at hand. Avoiding thoughts and distractions that may hinder the way to success, limiting fears and hopes that can negatively affect the performance [50] are skills that can make the difference between the first and the second place, especially in sports that are being decided in mere seconds (100 meters sprint/last shot in a close basketball match etc.).

In NBA the discussions about the clutch performances and "clutch gene" is almost as old as the game itself. Jerry West, the player in the Logo of the NBA was the first one called "Mr. Clutch", due to his uncanny ability to make strong plays in critical situations during the games, in conjunction with him being able to hold his composure even in the most intense moments. Michael Jordan is probably the most widely approved clutch player by the fans, while more recently Damian Lillard has been given the nickname "Dame Time" due to a series of game winning shots he made, Kawhi Leonard has been praised for his overall performance in critical games and Lebron James, Kevin Durand and Stephen Curry have all acquired reputation of being extremely qualified to make the baskets when it matters most.

In scientific literature, the notion of clutch performance is much more recent. As it generally incorporates a strong mental factor, some scientists consider it mainly psychological in origin [51], while others [54] argue that it has not been clearly defined and needs more interpretations and specifications as a sports definition. In another paper [52] Papatheodorou et al. discuss about the general notions of "clutch" and "choking" and suggest a variety of training programs to help the young athletes perform better under pressure. Although the recommendations on the paper are diverse and even reach the social impact of these training advices, the foundations lie on the principle of experience. They encourage the trainers to practice in situations similar to games with their young trainees, to teach them self-reflection and how to control their emotions and make better decisions in pressuring circumstances. This perspective analyzes clutch performance as a skill that can be trained and improved upon, in comparison with other studies [53] that advocate that clutch is not an ability but a very good performance at a critical time. A one-time event that resembles more an episodic incident and the corresponding players do not necessarily elevate their game at those moments by choice but rather circumstantial. Schweikle et al. [54] in their systematic review tried to build a framework by separating these distinctions (ability vs episodic performance) and analyzing both cases based on the corresponding literature, although conclusively they decided that the definition of clutch remains problematic. Traditionally, the focus in scientific literature was mainly on ‘chocking’ – when there can be observed a clear performance drop during the crucial moments of a game - especially when the stakes are high. Schwann [56] defined a psychological term as ‘clutch state’, which exist alongside the flow state, and under the certain circumstances of a game on the line, activate functions in the brain that can lead athletes to experience anxiety, technique abruption or focus difficulties.

From a more statistical point of view, Berry and Eshker [57] in a relatively old paper compared the playoff versus the regular season performances of the same players looking for differences that could indicate clutch or choking behavior. Their results showed no hint of clutch gene: almost every player had worse numbers in the playoffs than in the regular season. That being said, the playoff games are notoriously tougher especially from a defensive standpoint [58], and so this result hardly comes as a surprise. It would be interesting in such a study to rank the corresponding players according to their performance both during playoffs and during the season, and then compare these two ranking lists to pinpoint potential differences. Moreover, the authors [57] argued for confirmation bias, where NBA fans have the tendency to remember stunning achievements (especially clutch moments) and forget about poor performances. This idea is further supported by empirical study conducted by Wallace et al. [59], who clearly agree with the notion that fans like to overweight crucial successful plays, when they are achieved by their favorite players. The researchers compared the player statistics of the 4th quarter versus the corresponding box scores from 1st to 3rd quarters and discovered no such result as elevated or clutch performance. In similar lines, in another study from Israel [60], data from 222 close games during the 2005-2006 season was analyzed to determine potential clutch factor in the performance of the players. More specifically, the researchers used different versions of ANOVA (Analysis Of Variance Method) on the scores and percentages of the players during the last 5 minutes of the games versus the last 5 minutes of the half-time. The result showed that the conceived star (clutch) players did indeed achieved higher scores, but this did not reflect a higher shooting percentage but rather increased number of shot attempts. In Figure 14 the performances of the players are illustrated and further categorized as clutch and non-clutch. As we can observe, the averages per minute are characteristically higher for the clutch players, a fact which, as the authors suggest, can be explained by a variety of reasons (from being just better players to fans expectations to feeling the responsibility to justify their increased salary).

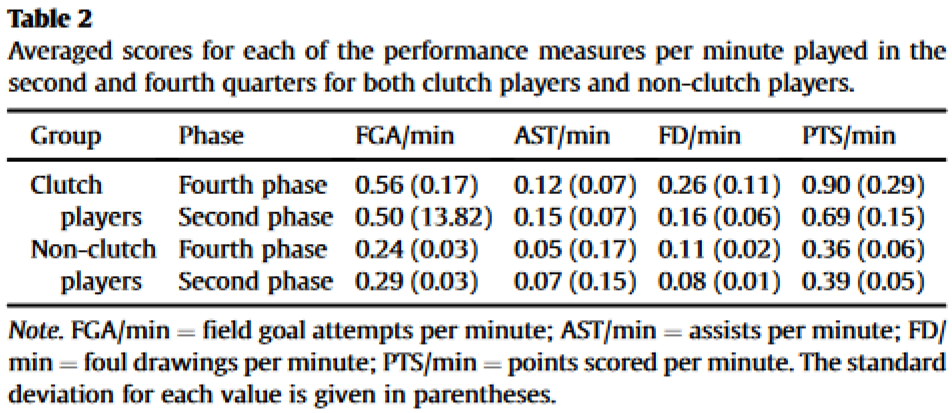


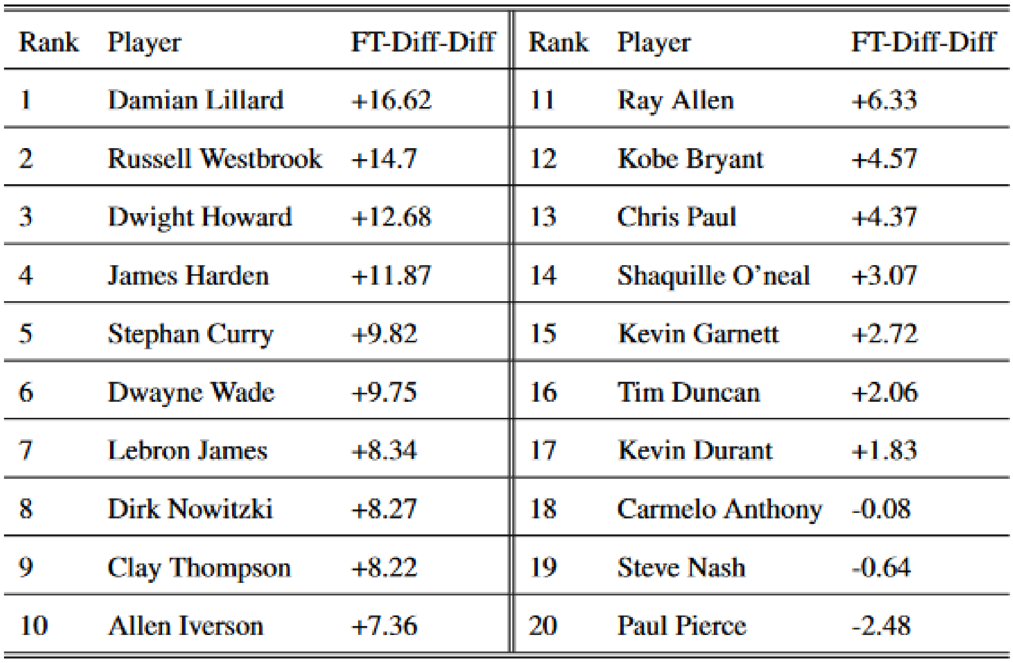
Table 2
Averaged scores for each of the performance measures per minute played in the second and fourth quarters for both clutch players and non-clutch players.

Group	Phase	FGA/min	AST/min	FD/min	PTS/min
Clutch players	Fourth phase	0.56 (0.17)	0.12 (0.07)	0.26 (0.11)	0.90 (0.29)
	Second phase	0.50 (13.82)	0.15 (0.07)	0.16 (0.06)	0.69 (0.15)
Non-clutch players	Fourth phase	0.24 (0.03)	0.05 (0.17)	0.11 (0.02)	0.36 (0.06)
	Second phase	0.29 (0.03)	0.07 (0.15)	0.08 (0.01)	0.39 (0.05)

Note. FGA/min = field goal attempts per minute; AST/min = assists per minute; FD/min = foul drawings per minute; PTS/min = points scored per minute. The standard deviation for each value is given in parentheses.

Figure 14

While most studies at this level tried to incorporate all sources of scoring during the critical moments of a game, Price et al. [61] investigated specifically the free throw component. With play-by-play data from 2003 to 2010 in NBA games and almost 500,000 data elements the empirical analysis gave a multitude of interesting results. Firstly, there was clear evidence that the NBA players do choke at the free throw line at critical moments; the situation is even worse, when the team of the shooter is losing, and when the shooter is inherently bad at free throws – they tend to shoot even worse when the game is on the line. Another interesting finding of this particular experiment was that the free throw is of great importance when the game is tied, but of almost negligible significance if the team who is winning is also shooting the free throw. In these cases, the team who is shooting the free throw has the same percentage of wins regardless of the outcome of the free throw. Finally, the authors [61] did not find any correlation between the shot percentages and the home/away court advantage, which although somewhat surprising, shows that the players are not particularly distracted from the crowd. Similar deductions were observed in another case [62] where the researchers explored specifically free throws and offensive rebounds, two totally different statistics, in combination with home and away court advantage parameter. The results showed no effect for the visiting team either in rebounds or in free throw percentage, but in comparison revealed an asymmetric impact of the pressure to the home team. There were cases that the home advantage could inspire the corresponding team, and other cases that the home pressure had detrimental effect on the home team leading them to a loss. The authors could not establish a pattern in the data to answer why or when each instance happened and so they called it “asymmetric impact”. Free throws in the clutch were the focus of another project [63] that tackled the issue from another angle. The author used traditional statistics in order to rank how well some of the top players in the league perform at the free throw line during the last moments of the game, by comparing them both to the league average and their own performance throughout the nonclutch game time. Figure 15 shows the results:



Rank	Player	FT-Diff-Diff	Rank	Player	FT-Diff-Diff
1	Damian Lillard	+16.62	11	Ray Allen	+6.33
2	Russell Westbrook	+14.7	12	Kobe Bryant	+4.57
3	Dwight Howard	+12.68	13	Chris Paul	+4.37
4	James Harden	+11.87	14	Shaquille O’neal	+3.07
5	Stephan Curry	+9.82	15	Kevin Garnett	+2.72
6	Dwayne Wade	+9.75	16	Tim Duncan	+2.06
7	Lebron James	+8.34	17	Kevin Durant	+1.83
8	Dirk Nowitzki	+8.27	18	Carmelo Anthony	-0.08
9	Clay Thompson	+8.22	19	Steve Nash	-0.64
10	Allen Iverson	+7.36	20	Paul Pierce	-2.48

Figure 15

The paradox of a free throw metric that ranks higher D. Howard than S. Nash is justified because Howard outperformed his own shooting percentage while in the clutch. Despite being a terrible free throw shooter on average, he managed to improve his accuracy considerably when the game was on the line. Steve Nash, one of the best free throw shooters in the history of the league, had such a high percentage, that his results during clutch time were slightly inferior (although by far better to Howard’s percentage) and that is why he ranks so low in the list. In spite of the fact that this is not a perfect metric at any rate, a study like this can offer other insights like which players manage to outperform themselves when their team needs them to.

More recently, Sarioz [53] used 11 years of data (2009-2019) to compare clutch versus non-clutch shooting percentages. He used t-test and binomial test to determine the existence of correlation in the distributions between the two variables. With generally strong p-values the results found no indication of clutch performance among the players. On the contrary he observed almost universal chocking, as the players' percentages tend to fall during these stages of the game. As we have already mentioned there is limited evidence both statistically and experience-wise that the players can actually improve their shooting accuracy during these critical moments. Except from all the mental factors that we have thoroughly analyzed in the literature, the players are generally tired at that point, while the defense most probably will challenge every shot when the game is on the line. To our understanding, a clutch performance ranking should incorporate both offensive and defensive stats, as well as individual and team modifiers.

3 General Terms

In this part, important definitions about the topics of data mining and machine learning will be given. In the third subpart, some specific techniques that were directly or indirectly used in this project will be analyzed, both theoretically and in the context of sports analytics.

3.1 Data Mining

Data mining is the process of discovering patterns, relationships and other information from large sets of data [64]. It incorporates extracting knowledge and insights that can assist the user answering a scientific question or making a calculated business decision.

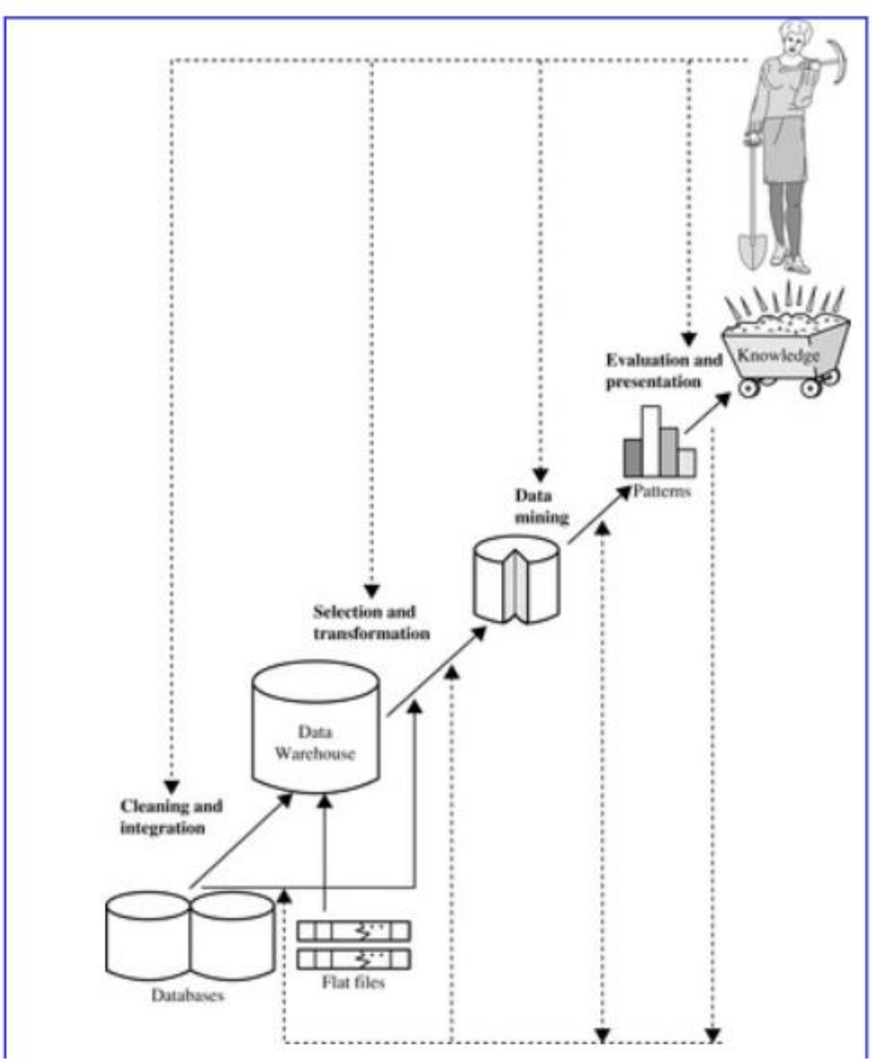


Figure 16 [65, pp 17]

Starting from the datasets and following the process to acquire useful information, data mining can use mathematical, statistical and computational techniques to uncover interesting insights by analyzing the structure of the observations. Data mining is consistently utilized in business to identify specific trends, anomalies and patterns [65] that are not inherently obvious to the user. Specialized methods often used in DM are association rules mining, anomaly detection and clustering.

3.2 Machine Learning

Machine learning is a field of Artificial Intelligence that engages the development of algorithms and statistical models that allow computational systems to learn and self - improve while training without being explicitly programmed. Their task generally revolves around creating a relation (function) that maps data points from input set X to output set Y (f: X → Y) [66].

Machine learning algorithms are divided into 3 main categories: Supervised learning, unsupervised learning and reinforcement learning.

- Supervised Learning uses labeled datasets to train the corresponding algorithms. In other words, it takes as input both data features and the target variable and adjusts the weights of the model accordingly until the model is optimized. Supervised learning is further split into 2 types of algorithms, namely classification and regression. In short, the focus of classification algorithm is to predict a discrete target variable, while regression works for continuous values.
- Unsupervised Learning algorithms find patterns based on similarity or density, but without apriori knowledge of what exactly are those common characteristics.
- Reinforcement Learning is based on the trial and error principle. The algorithm works through a system of punishment and reward, where it tries to maximize reward by trying different approaches to the problem.

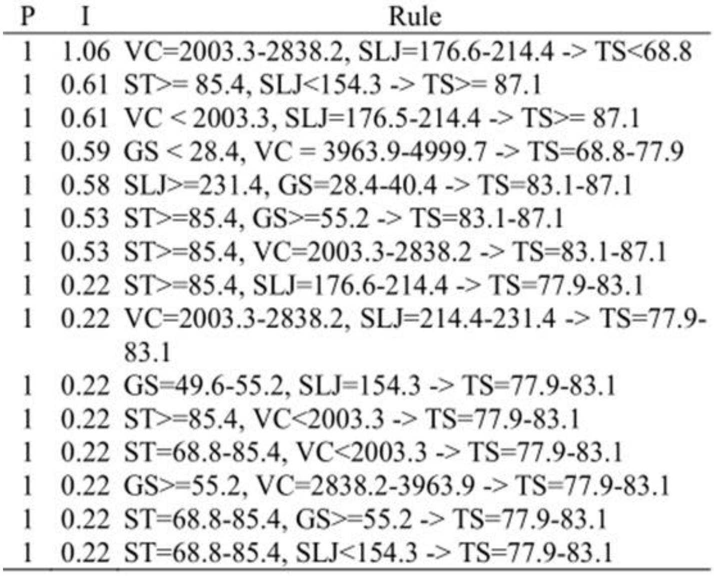
3.3 Machine Learning Techniques

In this subsection, we will mention and explain specific machine learning methods that were tested in this experiment.

3.3.1 Apriori Algorithm

Apriori algorithm is used to determine relationships between items in a database. In brief, it is comprised of two stages: During the first stage, the algorithm calculates the number of appearances of items and consequently the frequency of all the possible item sets and keeps track of those combinations that exceed the minimum support percentage. During the second stage it takes all of the item sets that advanced through the previous phase and generates the association rules that exist between them, based on a given level of confidence.

In a relevant study, Lan Yu [19] implemented a variation of the apriori algorithm on a database of students according to the PHS (Physical Health Standard) and their physical test scores, where he used the grades of Grip Strength, Vital Capacity, Standing Long Jump and Step Test as input and the grades of Total Score for prediction. Figure 17 illustrates the results of the algorithm sorted by probability and importance.



P	I	Rule
1	1.06	VC=2003.3-2838.2, SLJ=176.6-214.4 -> TS<68.8
1	0.61	ST>= 85.4, SLJ<154.3 -> TS>= 87.1
1	0.61	VC < 2003.3, SLJ=176.5-214.4 -> TS>= 87.1
1	0.59	GS < 28.4, VC = 3963.9-4999.7 -> TS=68.8-77.9
1	0.58	SLJ>=231.4, GS=28.4-40.4 -> TS=83.1-87.1
1	0.53	ST>=85.4, GS>=55.2 -> TS=83.1-87.1
1	0.53	ST>=85.4, VC=2003.3-2838.2 -> TS=83.1-87.1
1	0.22	ST>=85.4, SLJ=176.6-214.4 -> TS=77.9-83.1
1	0.22	VC=2003.3-2838.2, SLJ=214.4-231.4 -> TS=77.9-83.1
1	0.22	GS=49.6-55.2, SLJ=154.3 -> TS=77.9-83.1
1	0.22	ST>=85.4, VC<2003.3 -> TS=77.9-83.1
1	0.22	ST=68.8-85.4, VC<2003.3 -> TS=77.9-83.1
1	0.22	GS>=55.2, VC=2838.2-3963.9 -> TS=77.9-83.1
1	0.22	ST=68.8-85.4, GS>=55.2 -> TS=77.9-83.1
1	0.22	ST=68.8-85.4, SLJ<154.3 -> TS=77.9-83.1

Figure 17

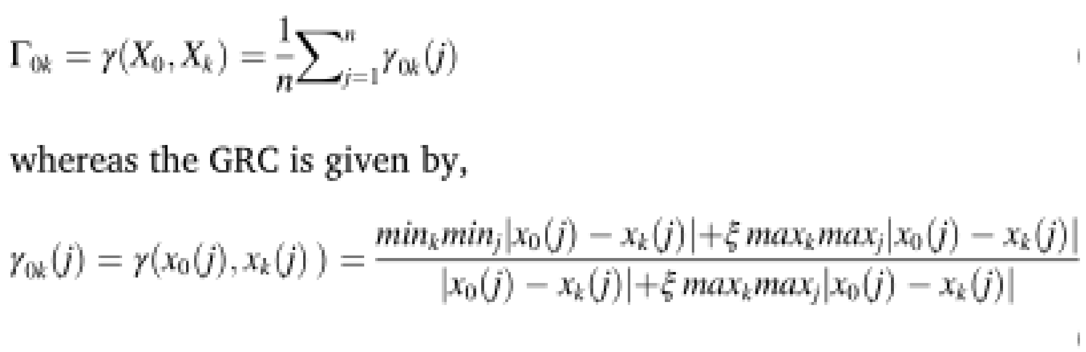
The results dictate that the two strongest dependencies are

$SLJ \geq 231.4 \Rightarrow TS \geq 87.1$ and $SLJ \geq 231.4 \Rightarrow TS \geq 83.1 - 87.1$, which means that the Standing Long Jump (SLJ) seems to be the most influential factor on the total score (TS) of the students. This kind of analysis takes the physical attributes and physical scores of a number of individuals and tries to predict the most important factor considering the general athleticism of a student. Studies such as these can encourage universities to staff their teams with players that have higher SLJ for the competitions as they are more probable to achieve better results, or trainers may be advised to help the students improve their SLJ, which would lead to a better performance overall. Nevertheless, we have to mention here that these are possibilities that have to be researched to be proved or rejected and not something we can take as granted. Even the author [2] notes that other algorithms have to be tested (Decision Trees, NN etc) in order to evaluate the consistency of the results.

3.3.2 Gray Relational Analysis

Gray Relational Analysis (GRA) is used in systems where there is lack of relevant information in the variables, called Grey Systems. The architecture [20] depicts the values as layers of white (known) and black (unknown) information, with the algorithm taking into consideration both types (grey information) to ultimately measure the importance of factors and assign relational degrees of importance to them that will determine their corresponding weight in the model. The technique calculates the Gray Relational Grade (GRG), based on the relation of the actual dataset compared to an ideal dataset, and uses this grade to rate the influence of the factors in a system.

Mathematically, it creates a correlation matrix between a reference dataset and the actual observed values [21]. Given that $X_0 = (x_0(1),\ldots,x_0(n))$ contains the reference data, and $X_k = (x_k(1),\ldots x_k(n))$ contain the data that needs to be compared with X_0, the algorithm calculates the GRG (Γ_{0k}) based on the following formula [20]:



$$\Gamma_{0k} = \gamma(X_0, X_k) = \frac{1}{n}\sum_{j=1}^{n} \gamma_{0k}(j)$$

whereas the GRC is given by,

$$\gamma_{0k}(j) = \gamma(x_0(j), x_k(j)\,) = \frac{min_k min_j |x_0(j) - x_k(j)| + \xi\, max_k max_j |x_0(j) - x_k(j)|}{|x_0(j) - x_k(j)| + \xi\, max_k max_j |x_0(j) - x_k(j)|}$$

Figure 18

GRC (γ_{0k}), is called the Gray Relational Coefficient and measures the degree of correlations between two specific factors, when GRG aggregates all the coefficients

│x_0(j) – x_k(j)│ is the absolute difference between the values

minmin│x_0(j) – x_k(j)│ is the minimum observed difference

maxmax│x_0(j) – x_k(j)│ is the maximum observed difference

ξ, is the Distinguishing Coefficient

The Distinguishing Coefficient cannot be determined precisely before the experiment and that comprises the only unresolved element of the formula. Nevertheless, studies [22] have indicated optimal values in the range (0,1), an estimation that is also proposed by the founder of Grey System Theory, Dr Deng [20]. Throughout the years it became a tradition for the researchers to use the value $\xi = 0.5$ [23].

NBA studies that are based on Grey System Theory are not numerous due to the plethora of statistical information that is available since the introduction of SportsVu capture infrastructure. They had been used at the dawn of sports analytics, when there was the need to deal with missing information. For example, Kuo et al [24] used GRA (Grey Relational Analysis) to compare the fighting abilities of teams in the NBA from both Eastern and Western Divisions based on in-game performance statistics (field goals, assist, rebounds etc.) of the season 2003/2004 in order to predict who will make the playoff. The study achieved a good accuracy of 13 out of 16 correctly classified teams, which displayed the viability of GRA in cases of basketball statistics. More recently, Dr. Pradhan [25] applied GRA on regular season individual player statistics to rank them based on performance metrics. The study overall took into consideration 45 performance metrics, 25 traditional (points, assist, fg, etc.) and 20 advanced (true shooting %, free-throw attempt %, etc.) and for simplicity reasons regarded only the top 100 single player seasons. Interestingly, the overall and traditional statistics [25] ranked S. Curry's 2015-16 season first and Jordan's 1987-88 and 1988-89 seasons as second and third, while the advanced statistics had Jordan's seasons first and Curry's as third. All in all, GRA proved to be a reliable technique to approach this type of ranking challenges, especially due to its capabilities in cases of limited data.

3.3.3 Linear Regression Analysis

One of the most popular families of modeling techniques is the regression modeling. In regression, the models try to explain a response variable, with regards to a number of explanatory (independent) variables [26]. The target is to detect the best coefficients (weights) for the explanatory values that describe the response variable with the minimum error (residuals). Since the models examine in detail the dependent variables, they assign corresponding weights that generally depend on the significance of each variable, but they can also pinpoint interesting correlations between them. The above characteristics establish regression techniques as a suitable tool for predicting dynamic values and changes in real world cases.

A special case of the regression family is the Linear Regression: In this model the dependent variable has a linear relation with the independent ones and the general formula is the following:

$$\mathbf{Y = w_1x_1 + \ldots + w_2x_2 + b,}$$

Y, the independent variable

$X_1,\ldots,X_n$, the dependent variables

$W_{1,\ldots}$, the corresponding weights of the dependent variables

B, is a constant that represents the error

Due to the amount of on-court stats in the game of basketball, linear regression modeling can be used to rate their importance in regards to a particular response variable (e.g. win, scoring, offensive performance). Chang et al [27] implemented linear regression on the basic statistical categories of 30 NBA teams during the seasons 2014-2019 to predict the expected score of every team in the relevant matches and thus, the consequent winner. The results were not discouraging, as the model achieved a descent 83% average accuracy. Characteristically, in games between teams of the same division, the results were even better, although the model did not perform as well in cases of teams of different conferences. The main explanation for this behavior is the uneven number of games between teams in the league; teams in the same division play 4 times against each other per year, while teams in the same conference but different division play 3 times and teams in different conference play only 2 times. The researchers tried to surpass their initial results by employing forward and backward stepwise variable selection, but despite their efforts, the model did not improve any further.

In another case [28], the author used regression to correlate player statistics with team results. Unsurprisingly, teams with more efficient players tent to perform better, although in such cases the interesting part lies with the outliers. Yang uncovered which teams under-perform or over-perform and argued about the reasons behind these irregularities based on his analysis. This example demonstrates that linear regression can statistically assist in a variety of ways. Along the same lines, this study [29] examines if there is a linear relation between the on-court performance of a player, his salary and his social media status. Although in the majority the assumption holds, there are several cases that were considered anomalies; for instance players with very weak performance rating but with more than average salary and even better social media reports. The linear models can easily pinpoint them out in order to study these cases further to determine the underlying factors.

3.3.4 Neural Networks

Probably the most popular algorithm of the recent years is the Artificial Neural Networks. ANNs are exceptionally capable of dealing with large amounts of data, and they can cope efficiently with dimensionality issues that may hinder other algorithms. The architecture (as well the name of the algorithm) was conceived based on the biological neurons of the human and animal brain. The basic elements of the ANN are the neurons, which are structured in layers, and in the simplest case, they comprise an input layer, a hidden layer and an output layer [30]. The input layer receives the data, and it transmits it to the hidden layer – with every piece of information travelling to every neuron in the hidden layer. Most of the work is being done there, and the input data transforms according to the specialized weights of the system, until the information is transmitted to the output layer through an activation function. This activation function provides a clear non-linear connection between the input and output values. Regarding the hidden layers, there is no limit on their number. There could be as many hidden layers as one wants (Deep Neural Networks), but that could increase the computation costs exponentially without any actual gain for the model. The number of hidden layers is generally dictated by the type of the data in question and the amount of data. Trying to build a deep NN with only a handful of input would most probably lead to overfitting problems.

An ANN with one hidden layer with 4 nodes (neurons) is illustrated in Figure 19. The input values are multiplied with the corresponding weights upon entering the nodes. These weights are trained inside the model to achieve the smallest possible error based on the specific data and task.

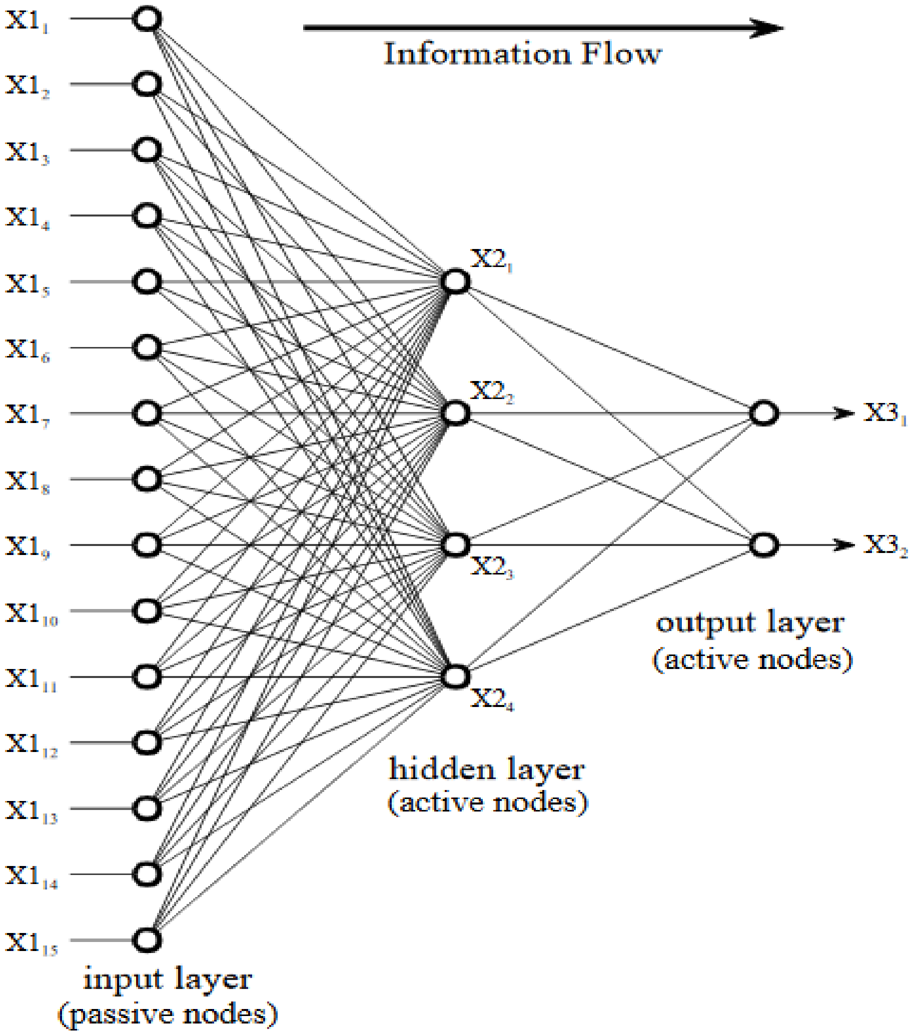


Figure 19 [30, page 459]

Consequently, the values inside the node are aggregated and passed to a sigmoid function before they depart for the next layer (Figure 20).

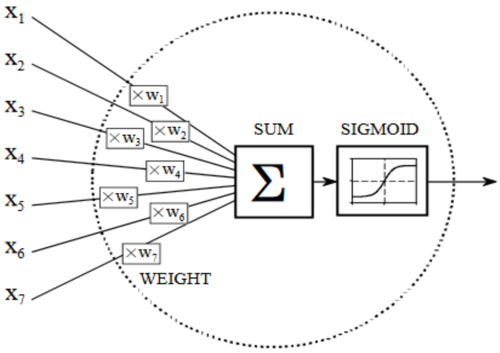


Figure 20 [30, page 461]

ANNs have proved to be quite effective in a variety of real world problems, especially in cases that the input data format is difficult to comprehend or edit (audio, image processing, signal processing, pattern recognition etc). Especially CNNs (Convolutional Neural Networks) were built for above tasks [31] and perform much better than the traditional algorithms.

In spite of the mentioned advantages, a trait of NNs that is still discouraging scientists to use them is that their inner workings are like a “black-box” [32]: no one really understands why a particular model works and why these specific trained weights can achieve such high accuracy in a prediction or what is their meaning. That means that it is difficult for researchers to answer qualitative questions based on NN. For instance, if a NN model predicts correctly the specific hotel that a customer chooses in booking.com, it will be unable to provide more information about the reasons of that particular choice or why it did not recommend another hotel. This kind of information is crucial in a lot of scientific fields, which means that there is still room for the more traditional machine learning approaches.

In the basketball universe, an application [33] was built implementing the Neural Networks algorithms that predicts the players that are qualified to be inducted into the Hall of Fame (HoF) with an accuracy rating that reaches 0.93 (93%). The study showed that ANN achieved much better results than older attempts to predict hall-of-famers (like linear or logistic regression). The researcher also tested the ANN against a CNN with the former achieving higher accuracy (0.93 vs 0.91). Some interesting points of discussion derived from the report [33] regard the choice of variables that were inserted in the model. For instance, the number of championships a player has won was deducted due to the fact that it could skew the results without having the appropriate meaning. The variable could not separate important players from rotation players if they had the same amount of rings. So in effect it renders Robert Horry with 7 championships much more probable Hall-of-Famer than Charles Barkley (0 championships), which is not true. There is also the case of duration of one’s career. The researchers noticed some outliers that would be very difficult for a model to predict (e.g. John Thompson with only two years in the league made the Hall-of-Fame). Nevertheless, the program continues to predict HOF inductees with high accuracy based on their stats and accomplishments.

Another report that demonstrates the capabilities of NN was conducted in the University of Toronto [34] in Canada, where Wang and Zemel cooperated with the Toronto Raptors to analyze the offensive plays with NNs. In this particular study, the variables were actually location information being available through the SportVU tracking system as depicted in Figure 21:

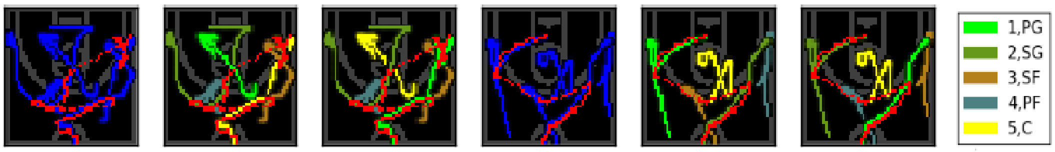


Figure 21

The red trajectory follows the ball, while the blue is the single-representation of all the players. The rest of the colors are explained in figure.

The input data for the NN model was in such pictorial format that represented the location of the offensive players in the court and was captured with the VU equipment at 25 frames per second. The particular model that provided the best results was the RNN (Recurrent Neural Network), a type of NN that can keep the algorithm stationary along the process, without accumulating useless noise (in this case by the irrelevant locations that the players would take in the court). After the data cleaning and preprocessing, there were selected 11 classes for the target variable representing 11 distinct offensive plays, while the size of the set of offensive sequences was 1435. After training and testing, the advanced RNN model achieved 66% top-1 accuracy and 80% top-3. This was definitely an encouraging result given the type data and the complication of the study. Although this experiment may not have yielded any groundbreaking information, it was a clear indication of the special capabilities of NNs regarding pictorial data in pattern recognition achieved inside a complicated environment such as the basketball court.

In similar philosophy, there was another study [36] that utilized RNN (Recurrent Neural Networks) in order to predict the accuracy of 3-point shots given the trajectories of more than 20,000 three pointers from NBA SportVU data. As the researchers discussed, the model had no prior knowledge of basketball notions or feature engineering; the only actual data was the trajectories of the ball as sequence movement. The final model managed to achieve accuracy of more than 84% when the ball was 8-feet away from the basket. All in all, the RNNs clearly outperformed the other tested models (Linear, and Gradient Boosted Machines (GBM)) in this particular experiment.

In other cases, the NNs were used to predict the MVP of the league for the years 2011-2019 [35] with relative accuracy. Nevertheless, as we have already discussed, the authors [35] agree that the MVP voting tends to depend on various off-court factors (popularity, social status, commercials, team approval and many others) that are not directly connected with in-game statistics. This kind of predictions is the hardest to accomplish without taking into consideration the team of the corresponding players, as it clearly impacts the voting results.

4 Methodology

This part explains the steps taken to complete the data mining task. It includes Data Acquision and Collection, where there will be thorough analysis of the terms and the procedures that were executed in order to complete these processes. In Data Acquisition section there are many important information about the challenges that modern data scientists face when trying to obtain the data, while in Data Collection we cover our own undertakings in order to acquire the relevant datasets.

In the following sections, we continue with data preprocessing steps, which is data cleansing, variable filtering and feature transformation. Data preprocessing is a critical and time-consuming step in the data analysis process, as it ensures that the data is in suitable format and quality for analysis. The final quality of the data and how well it has been cleaned and prepared, directly impact the accuracy and reliability of the analysis and models.

4.1 Data Acquisition

Obtaining the relevant data is a vital part of every data mining procedure. The quality of the processed data is of paramount importance; it can make the difference between a well calculated decision and a wrong recommendation. Moreover, the format of the data determines its eligibility to a variety of platforms, operations or algorithms. Data with a lot of noise – that is, irrelevant content, empty values, or in inconsistent formats – is difficult to process and translate into useful information.

Especially with the expansion of big data services, the amount of sports data being generated has increased dramatically. Real time statistics, exercise performance [37], health metrics, training schedules and recommendations are only a handful of the affected categories. Performance data can now be tracked in real time through wireless and mobile technologies [38]. Sensors are able to transmit signals from the athletes during the actual activity, which can lead to more effective monitoring and predictive (or even prescriptive) adjustments to be applied.

Nevertheless, these advances do not come without equally significant challenges [39]. Firstly, the data is obtained by so many diverse sources and in so many different formats, which creates problems of heterogeneity and algorithmic biases. Moreover, as Big Data is characterized by high dimensionality and huge sample sizes, it demands heavy computational resources and cause instability issues [39] in algorithmic procedures (especially in case of outliers). These challenges require a change in statistical methods and computational techniques in order to be addressed successfully. And although there are already new frameworks, which are designed with high data complexity and data security in mind [40], there is still a lot of work to be done, mainly in aggregating the different formats of data into a single homogenous unit that can be accessed and processed by data mining algorithms.

4.2 Data Collection

In this particular project we used clearly defined basketball statistics from reliable online sources, namely basketball-reference.com [44] and nba.com [45]. To obtain the datasets we used two methods: 1) we downloaded freely accessible excel workbooks when possible, 2) we used data scraping techniques with the programming language Python [46] and its relevant libraries to download the necessary data.

4.2.1 Definition and Properties

Data collection involves the process of gathering information on variables of interest from a variety of sources [41] in order to answer a data-based question or to use in data-based projects. The data may consist of different data types (e.g. numerical, categorical) from different data sources (observations, books, questionnaires, webpages) [42] and formats (.csv files, text files, video files etc.) for a single project; it is the responsibility of the data scientist to find ways to integrate these data in order to provide useful information for the particular question.

Even more importantly, the researcher should ensure that the data satisfies a set of basic qualities [43]:

- **Accuracy:** The data points should be accurate, without errors and describe the true values and properties of the measured variables.

- **Completeness:** The data should not have empty values and should include all the mandatory information.
- **Validity:** Data should correctly represent the concepts it is intended to measure, avoiding biases and distortions.
- **Consistency:** Data should be relatively homogenous and following certain standards in its entirety.
- **Relevance:** Data should not contain information unrelated to the topic, which just add noise to the problem.
- **Security:** Data should screen private or sensitive information (according to the latest GDPR standard in Europe and the state laws in the USA).

4.2.2 Web Scraping

It is advisable at this point to reference some points concerning the legality of web scraping. Scraping is a method for data extraction in order to secure information from online sources automatically. It is used through software platforms like Python or R and the goal is to download the selected information in a concise and structured format [47]. Due to its recent emergence and peculiar nature, data scraping has not yet been defined under a universal legal framework [48]. Disputes over Web Data that have been brought to the court of justice were based on illegal access of a computer, illegal access of data, copyright issues and breach of contract but several courts remained divided and had difficulty reaching a verdict [48]. These cases tend to get resolved based on the particular website's "terms of service". Still, in some instances even the terms of service were not enough to obtain guilty verdict, as, in a widely discussed case [49], the court of appeals ruled initially that scraping publicly available data does not violate the Computer Fraud and Abuse Act (CFAA). As a general rule, until solid legal guidelines around web scraping have been established, the users must be aware of the legal implication of data scraping and always check the permissions provided by the website and the terms of use.

4.2.3 Our Sources

As we already mentioned, we downloaded traditional basketball statistics from the NBA season 1996-1997 to season 2017-2018. The data included two (2) .csv files for each year: one for the regular season and one for the playoffs, so we ended up with 44 files.

The relevant code for scraping the data lies at appendix 1.

Regarding the initial filters from the [44] source, we chose to include only games where the point differential was 5 or less – which means only very close games, and we limited the stats only to the last 3 minutes. 3 minutes seemed a reliable choice as there is still enough time to separate random phenomena from a consistent behavior and it is not far away from the end that the teams would not get their best players to chase the victory. The following table (Table 1) illustrates the format of our initial variables:

Table 1

The name of the Variable	Short Description of the Variable
Player	"The name of the Player"
Team	"The team for which he played in that particular season"
Age	"The age of the Player"
GP	"How many games he played in that period"
W	"How many wins he got in the aforementioned games"
L	"How many loses"
Min	"How many minutes he played on average"
PTS	"How many points he scored on average"
FGM	"How many field goals he made on average"
FGA	"How many field goals he attempted on average"

FG%	“The field goal percentage”
3PM	“How many 3-pointers he made on average”
3PA	“How many 3-pointers he attempted on average”
3P%	“The 3-pointer percentage”
FTM	“How many free throws he made on average”
FTA	“How many free throws he attempted on average”
FT%	“The free throw percentage”
OREB	“How many offensive rebounds he averaged”
DREB	“How many defensive rebounds he averaged”
REB	“How many total rebounds he averaged”
AST	“How many assists he averaged”
TOV	“How many turnovers he averaged”
STL	“How many steals he averaged”
BLK	“How many blocks he averaged”
PF	“How many personal fouls he averaged”
FP	“His average Fantasy Points”
DD2	“Number of double doubles”
TD3	“Number of triple doubles”
+/-	“Plus – Minus”

*The numbers are specific for the studied time period (last 3 minutes with the optional addition of overtime)

The FP statistic (fantasy points) is a made-up metric to calculate the weekly performance of the players. In NBA, it is calculated based on the following formula:

FP = 1*Points + 1.2*Rebounds + 1.5*Assists

Although an interesting statistic to discuss and further analyze, it will not affect our work, as we will define our own metrics that are directly related to the topic of interest.

The +/- (plus minus) statistic signify the score difference of the team while the particular player is on the court. If a team is losing 5 points when a player enters the court and finishes winning by 3 when he is substituted, that means that for this specific period in the game, the corresponding plus/minus score is +8. Again, an engaging variable when examined under the proper context, it can definitely offer special insights about the impact of the player on a team. Nonetheless, in a case like ours where we do not necessarily compare players of the same team or against their own previous performances, it should give diminishing rewards.

The following table (Table 2) shows a part of the preliminary dataset of the regular season 1996-1997 that we obtained from the webpages:

Table 2 : Players are ordered based on their average points

	Player	Team	Age	GP	W	L	Min	PTS	FGM	FGA	FG%
1	Melvin Booker	GSW	24	1	0	1	6.2	4	2	4	50
2	Eldridge Recasner	ATL	29	3	3	0	2.3	3.7	0.7	1.3	50
3	Michael Jordan	CHI	34	28	19	9	2.6	3.1	0.7	1.8	38.8
4	Hakeem Olajuwon	HOU	34	37	25	12	2.6	2.9	1.1	1.8	59.1
5	Mitch Richmond	SAC	32	36	19	17	2.8	2.6	0.7	1.7	40.3
6	Grant Hill	DET	24	32	20	12	2.6	2.5	0.8	1.5	50
7	Cedric Ceballos	PHX	27	10	6	4	2.4	2.5	1	2.1	47.6
7	David Robinson	SAS	31	2	1	1	2.1	2.5	1	2	50
9	Terrell Brandon	CLE	27	34	12	22	2.8	2.4	0.7	1.5	47.1
10	Dana Barros	BOS	30	3	1	2	1.6	2.3	1	1.3	75

We can already notice some irregularities: Melvin Booker is depicted as the best clutch scorer for this year with 4 points per game but if we look closer we can observe that he has only a single relevant match in which he managed to score 4 points and not only that but the minutes average is at 6.2 while we are examining a period of 3 minutes.

Obviously, this is not a mistake, but an indication that Booker played in a single close game, which was decided in the overtime and he managed to score 2 field goals.

If we check further down the list we can notice that half of the first 10 players have an extremely low amount of games and should be considered outliers, which will be dealt with in the data cleaning phase of the project.

4.3 Data Cleaning

Data cleaning, otherwise called cleansing or scrubbing [67] is the process of finding and correcting errors, irregularities and anomalies. It ensures that the data possesses the qualities mentioned in paragraph 4.2.1 and assists in the analytical process of the data mining project by making the data easier to comprehend, process and scrutinize. The scope of data cleaning is to improve the general quality of the data sets and deal with any form of collection errors, missing values, outliers or irrelevant information that may hinder the whole process. In real world problems this step is usually time-consuming as the unprocessed data tends to hold large amounts of noise that obstruct the efficient application of the data mining algorithms. In some cases, certain machine learning techniques may be used in this part – like clustering, to distinguish observations that have similar properties and are of interest to the particular research question from the rest of the useless observations that are considered as noise.

In our case, the data sets we downloaded are in very good condition, already processed and cleared of errors and missing values. So in this part, we will concentrate on three operations: 1) dealing with outliers, 2) filtering pointless and impractical information, 3) create and transform variables to make them more relevant to our research question.

4.3.1 Outliers

The main issue with our dataset is the one we mentioned in the previous section: the existence of players with very few games that may have had exemplary performances but with a little or no reproducibility. To alleviate this fact we either needed to establish a lower limit at the number of games of the eligible players or use the number of games as a variable with negative weight at our formulas. We easily chose the former approach as the latter would complicate our algorithms without any tangible benefit. The question then was: how are we going to define this limit? It definitely cannot be arbitrary; it should have some statistical justification.

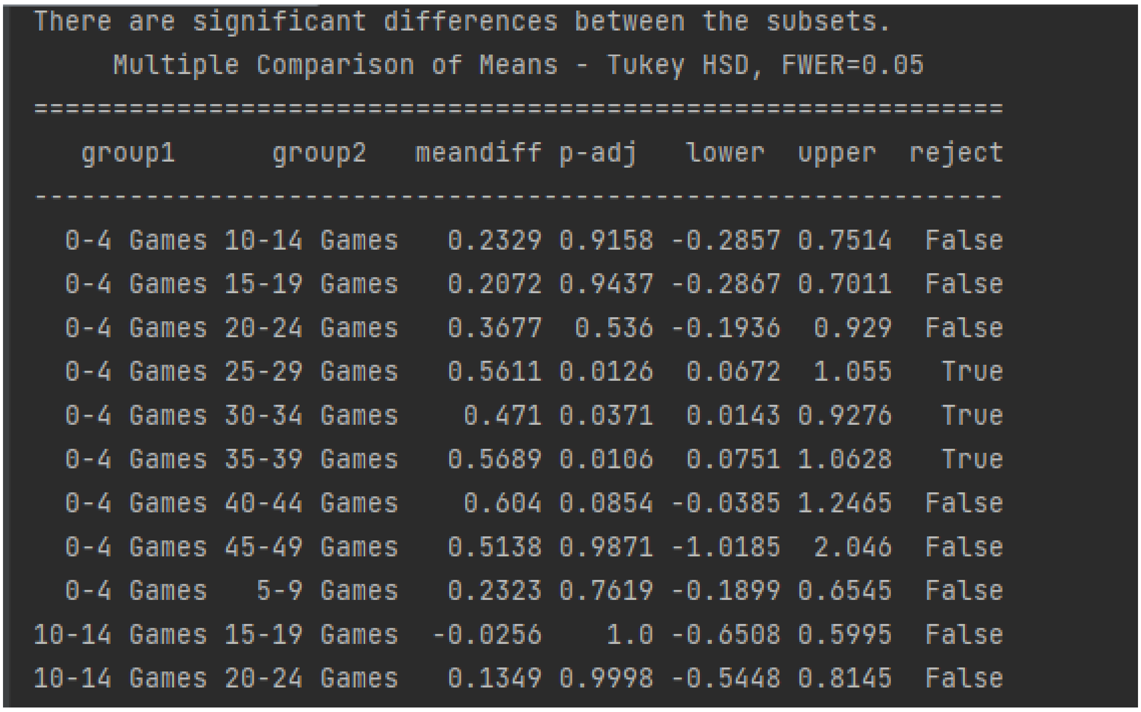
The main goals are for the sample to be representative of the clutch as ability and not as random incidents of performance. Also, as a general guideline, a larger sample size (more games) tends to provide more reliable and stable statistical estimates. If a player has played a significant number of games, their performance statistics are likely to be more representative of their true abilities. In order to manage to achieve a relevant statistical stability we decided to run ANOVA tests in a small sample of our datasets to find an optimized threshold for the "GP" ("Games Played") attribute.

The procedure was as follows: First we divided the datasets into smaller groups based on the GP. We used the regular season datasets because the samples were characteristically larger and they would be much more helpful. We separated the groups by 5 games:

- Group 1 : 0-4 Games
- Group 2 : 5-9 Games
- Group 3 : 10-14 Games
- Group 4 : 15-19 Games
- ….
- Group 10 : 45-49 Games

We should notice here that the 10th group in all cases had only just a couple of players, e.g. for 1996-97 the only players in the group were P. Ewing and C. Oakley from the New York Knicks. Groups like that would definitely behave strangely but for the purpose of our research that did not bother us at all; the larger number of close games these players took part in, the better. In this part of our work we focused on the groups with the least amount of games.

After separating the groups we ran ANOVA (Analysis Of Variance) to determine differences in the distribution of the groups, with variable of interest a Clutch Performance Estimator (CPE) that we will define later in this paper. It generally resembles Tendex values but with a bit more emphasis in specific statistics related to clutch performance. The ANOVA showed in all cases significant differences between the sets, so we chose the Tuckey's HCD test as a post-hoc test to compare all possible pairs of groups and determine where the differences are significant. The results of the HCD test are depicted in the following figure [Figure 22]:



```
There are significant differences between the subsets.
     Multiple Comparison of Means - Tukey HSD, FWER=0.05
==============================================================
   group1      group2   meandiff p-adj   lower  upper  reject
--------------------------------------------------------------
  0-4 Games 10-14 Games   0.2329 0.9158 -0.2857 0.7514  False
  0-4 Games 15-19 Games   0.2072 0.9437 -0.2867 0.7011  False
  0-4 Games 20-24 Games   0.3677  0.536 -0.1936  0.929  False
  0-4 Games 25-29 Games   0.5611 0.0126  0.0672  1.055   True
  0-4 Games 30-34 Games    0.471 0.0371  0.0143 0.9276   True
  0-4 Games 35-39 Games   0.5689 0.0106  0.0751 1.0628   True
  0-4 Games 40-44 Games    0.604 0.0854 -0.0385 1.2465  False
  0-4 Games 45-49 Games   0.5138 0.9871 -1.0185  2.046  False
  0-4 Games   5-9 Games   0.2323 0.7619 -0.1899 0.6545  False
10-14 Games 15-19 Games  -0.0256    1.0 -0.6508 0.5995  False
10-14 Games 20-24 Games   0.1349 0.9998 -0.5448 0.8145  False
```

Figure 22

We did not expect a definitive answer, given the nature of the problem and the limited game-time period that we are examining, we had our doubts if this method will actually help, but as it turns out it gave us useful insights.

As we can observe the first group with the fewer games (0-4) – which we would definitely omit, seems to have not important statistical differences with the next 4 groups, while the p-value would get much lower from the group 25-29 and onwards. This meant that the distribution of the players with less than 25 games had different characteristics than for the players that had more close games in a season. We also run more of the same tests ANOVA and Tukey’s HCD for the different groups and seasons and finally we decided to define the threshold at 20 games for each season. In similar logic we defined the equivalent threshold for the playoffs at 5 games.

The only problem would be in cases where a clutch performer played for a team that got consistently in the playoffs but was eliminated in the first round, and the player did not make our playoff lists (because he might list less than 5 close games). Still, if his performances were that good, he should appear in the regular season’s sets, which will definitely include him in our analysis. All in all, we felt that the 5 games limit in the playoff, and the 20 games limit in the regular season made a lot of sense both from statistical and empirical standpoint.

4.4 Variable Filtering

At this stage of our project we needed to further improve our dataset in order to make it even more suitable to our specific goals. We examined theoretically and from our own knowledge and experience which of our initial attributes had little or no useful information to offer in order to decrease the number of variables for our problem.

So let's offer a summary of the variables we decided to omit [Appendix]:

- **TEAM** – the team variable in a time span of 22 years should have no effect in the analysis. The teams change every couple of years, as does the coaching staff and even the directors. Also we are having a player comparison in this study not a team comparison, which makes it irrelevant
- **AGE –** although we would like very much to include if the age of a player does play a role through his experience in his crucial decision in clutch moments, this particular study will not include this parameter. We would have to include players as a time series and observe their performances as the years go by. Nevertheless, we will absolutely keep this as a future work project.
- **REB/DREB –** this was a very curious debate, because despite the fact that rebounds are one of the most important statistic in basketball, the defensive rebound in clutch moments is not what makes the difference, and it is not a big play – teams are expected to gather the defensive rebounds. From this perspective offensive rebounds are much more interesting to examine at these moments. That being said, we could not delete the rebound category without using statistical methods first, so we kept it for the time being.
- **FP –** as we mentioned fantasy points are not the focus of this research and offer no benefit to keep
- **DD2/TD3 –** double-doubles and triple-doubles have no usefulness because they do not represent the clutch period but the whole game.

- **Plus/Minus** – although an interesting statistic on its own, that definitely can show the impact of a player on the court, we could not think of a way to utilize it in our project.

One last action was to cut all the regular season datasets to 50 players. This was a minor change; all our data sets were at around 50 already after the previous data cleaning steps. The playoff datasets had of course even less players but that is expected as the teams are fewer. The main reason was that having the same amount of observations in each set would further enhance homogeneity in the data and it would make our later work with the algorithms much smoother. The filters to decide which players would be cut were points, games and performance indicator.

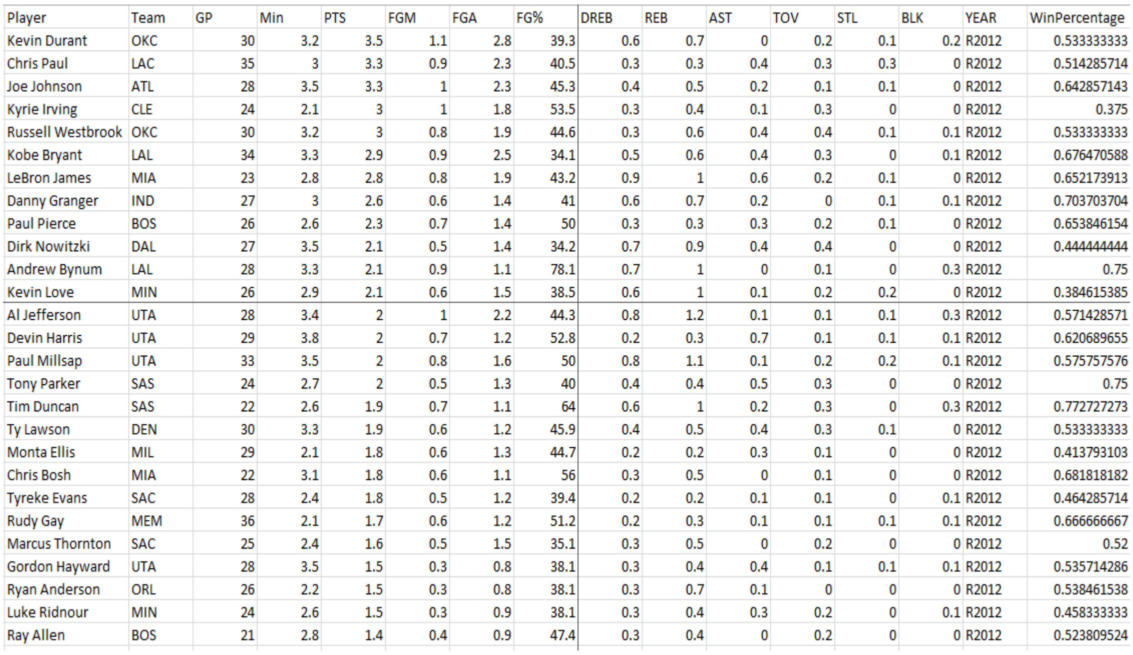
4.5 Data Transformation

In this part of the project we will make some slight changed to our dataset in order to make it more relevant to the particular domain that we are examining and assist us in the algorithmic solutions that we will test in the next chapter.

Our first step was to add a new attribute in the datasets that declares the relevant year of the specific statistics of the player. Since we are at some point going to compare and aggregate many of the datasets, we will need more details about the performance of a specific player. For instance, we will stumble upon the name Kobe Bryant and his stats. He always played for the Los Angeles Lakers. How are we going to separate if we are talking about 2001 Kobe, or 2008, or any of the other years. So we added this variable that will inform us about the year and the stage (regular season vs playoffs).

Furthermore, we had to define a target statistic. As we explained earlier, supervised learning requires a target variable: a dependent variable that we are trying to comprehend and predict. For this reason, we decided that the W/GP (Wins divided by Games Played) was a good enough statistic. It is not perfect, but since there was no way for us to distinguish the result of every game individually, it was the next best choice. Most of our data points are on average, and so this average win/loss fraction will be a descent target variable as we will witness in the next chapter. We also subtracted the “W” (win) and “L” (lose) columns as their information were passed to the new attribute “WinPercentage”.

So, Figure 23 illustrates a sample of the final form of our 44 datasets after the data Pre-processing part.



Player	Team	GP	Min	PTS	FGM	FGA	FG%	DREB	REB	AST	TOV	STL	BLK	YEAR	WinPercentage
Kevin Durant	OKC	30	3.2	3.5	1.1	2.8	39.3	0.6	0.7	0	0.2	0.1	0.2	R2012	0.533333333
Chris Paul	LAC	35	3	3.3	0.9	2.3	40.5	0.3	0.3	0.4	0.3	0.3	0	R2012	0.514285714
Joe Johnson	ATL	28	3.5	3.3	1	2.3	45.3	0.4	0.5	0.2	0.1	0.1	0	R2012	0.642857143
Kyrie Irving	CLE	24	2.1	3	1	1.8	53.5	0.3	0.4	0.1	0.3	0	0	R2012	0.375
Russell Westbrook	OKC	30	3.2	3	0.8	1.9	44.6	0.3	0.6	0.4	0.4	0.1	0.1	R2012	0.533333333
Kobe Bryant	LAL	34	3.3	2.9	0.9	2.5	34.1	0.5	0.6	0.4	0.3	0	0.1	R2012	0.676470588
LeBron James	MIA	23	2.8	2.8	0.8	1.9	43.2	0.9	1	0.6	0.2	0.1	0	R2012	0.652173913
Danny Granger	IND	27	3	2.6	0.6	1.4	41	0.6	0.7	0.2	0	0.1	0.1	R2012	0.703703704
Paul Pierce	BOS	26	2.6	2.3	0.7	1.4	50	0.3	0.3	0.3	0.2	0.1	0	R2012	0.653846154
Dirk Nowitzki	DAL	27	3.5	2.1	0.5	1.4	34.2	0.7	0.9	0.4	0.4	0	0	R2012	0.444444444
Andrew Bynum	LAL	28	3.3	2.1	0.9	1.1	78.1	0.7	1	0	0.1	0	0.3	R2012	0.75
Kevin Love	MIN	26	2.9	2.1	0.6	1.5	38.5	0.6	1	0.1	0.2	0.2	0	R2012	0.384615385
Al Jefferson	UTA	28	3.4	2	1	2.2	44.3	0.8	1.2	0.1	0.1	0.1	0.3	R2012	0.571428571
Devin Harris	UTA	29	3.8	2	0.7	1.2	52.8	0.2	0.3	0.7	0.1	0.1	0.1	R2012	0.620689655
Paul Millsap	UTA	33	3.5	2	0.8	1.6	50	0.8	1.1	0.1	0.2	0.2	0.1	R2012	0.575757576
Tony Parker	SAS	24	2.7	2	0.5	1.3	40	0.4	0.4	0.5	0.3	0	0	R2012	0.75
Tim Duncan	SAS	22	2.6	1.9	0.7	1.1	64	0.6	1	0.2	0.3	0	0.3	R2012	0.772727273
Ty Lawson	DEN	30	3.3	1.9	0.6	1.2	45.9	0.4	0.5	0.4	0.3	0.1	0	R2012	0.533333333
Monta Ellis	MIL	29	2.1	1.8	0.6	1.3	44.7	0.2	0.2	0.3	0.1	0	0	R2012	0.413793103
Chris Bosh	MIA	22	3.1	1.8	0.6	1.1	56	0.3	0.5	0	0.1	0	0	R2012	0.681818182
Tyreke Evans	SAC	28	2.4	1.8	0.5	1.2	39.4	0.2	0.2	0.1	0.1	0	0.1	R2012	0.464285714
Rudy Gay	MEM	36	2.1	1.7	0.6	1.2	51.2	0.2	0.3	0.1	0.1	0.1	0.1	R2012	0.666666667
Marcus Thornton	SAC	25	2.4	1.6	0.5	1.5	35.1	0.3	0.5	0	0.2	0	0	R2012	0.52
Gordon Hayward	UTA	28	3.5	1.5	0.3	0.8	38.1	0.3	0.4	0.4	0.1	0.1	0.1	R2012	0.535714286
Ryan Anderson	ORL	26	2.2	1.5	0.3	0.8	38.1	0.3	0.7	0.1	0	0	0	R2012	0.538461538
Luke Ridnour	MIN	24	2.6	1.5	0.3	0.9	38.1	0.3	0.4	0.3	0.2	0	0.1	R2012	0.458333333
Ray Allen	BOS	21	2.8	1.4	0.4	0.9	47.4	0.3	0.4	0	0.2	0	0	R2012	0.523809524

Figure 23

As we explained, all the relevant stats are here plus the “YEAR” (R2012 means Regular Season of 2011-2012, as opposed to P2012 which would indicate Play Off of the same season) and the “WinPercentage” means the Win/Games_Played fraction.

5 Algorithms and Modelling

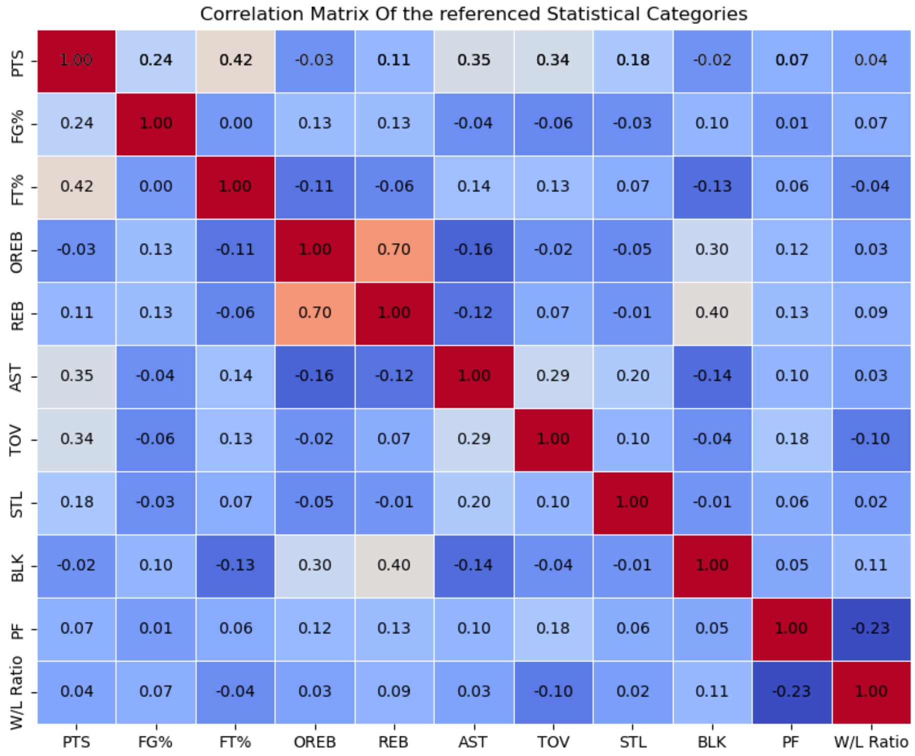
In this section we are going to use a variety of algorithms to rank feature importance and then test diverse models that adequately evaluate the performance of the players in the dataset. So the first part consists of the Feature Selection section, where are goal is to determine the most significant attributes that contribute to a successful clutch player. Afterwards, we will utilize our findings in order to build a model that can approximate the observations in the dataset concerning our target variable.

5.1 Feature Selection

Firstly, to decrease the number of variables we decided to omit the “FGM” (field goals made) and “FGA” (field goals attempted) as they are sufficiently described in FG% (field goal percentage). The same applies for the free throws, as the only relevant statistic is the percentage. We were also able to exclude the 3-Point statistics as they are inherently represented in both the “PTS” (points) and “FG%” (field goal percentage), both the volume of the 3 point shot and the accuracy were calculated in the above metrics.

The next variables that we wanted to address were the “REB” (Rebounds) and the “PF” (Personal Fouls).

Rebounding is a complicated statistic category; it is imperative to gather the rebounds as a team effort to win the match but as a personal performance indicator does not always correlates with wining conditions. Getting a lot of defensive rebounds is important, but also dependent on the other team missing their shots, which in turn is mainly contingent on the overall team defense. As in our dataset we could not obtain advanced defensive statistics for this little game period that we are analyzing, we did not want to have the results skewed by high defensive rebounding numbers that are not considered “personal big plays” by the literature, and are better explained with overall team defense performance. The original correlation matrix of all the statistics of the aggregated dataset depicted in Figure 24 gave some information to continue. By far the highest correlation between our variables was that between the offensive and overall rebounds. That was expected as the one category is part of the other, but we had little other information about how the categories impact the overall performance.



Correlation Matrix Of the referenced Statistical Categories

	PTS	FG%	FT%	OREB	REB	AST	TOV	STL	BLK	PF	W/L Ratio
PTS	1.00	0.24	0.42	-0.03	0.11	0.35	0.34	0.18	-0.02	0.07	0.04
FG%	0.24	1.00	0.00	0.13	0.13	-0.04	-0.06	-0.03	0.10	0.01	0.07
FT%	0.42	0.00	1.00	-0.11	-0.06	0.14	0.13	0.07	-0.13	0.06	-0.04
OREB	-0.03	0.13	-0.11	1.00	0.70	-0.16	-0.02	-0.05	0.30	0.12	0.03
REB	0.11	0.13	-0.06	0.70	1.00	-0.12	0.07	-0.01	0.40	0.13	0.09
AST	0.35	-0.04	0.14	-0.16	-0.12	1.00	0.29	0.20	-0.14	0.10	0.03
TOV	0.34	-0.06	0.13	-0.02	0.07	0.29	1.00	0.10	-0.04	0.18	-0.10
STL	0.18	-0.03	0.07	-0.05	-0.01	0.20	0.10	1.00	-0.01	0.06	0.02
BLK	-0.02	0.10	-0.13	0.30	0.40	-0.14	-0.04	-0.01	1.00	0.05	0.11
PF	0.07	0.01	0.06	0.12	0.13	0.10	0.18	0.06	0.05	1.00	-0.23
W/L Ratio	0.04	0.07	-0.04	0.03	0.09	0.03	-0.10	0.02	0.11	-0.23	1.00

Figure 24

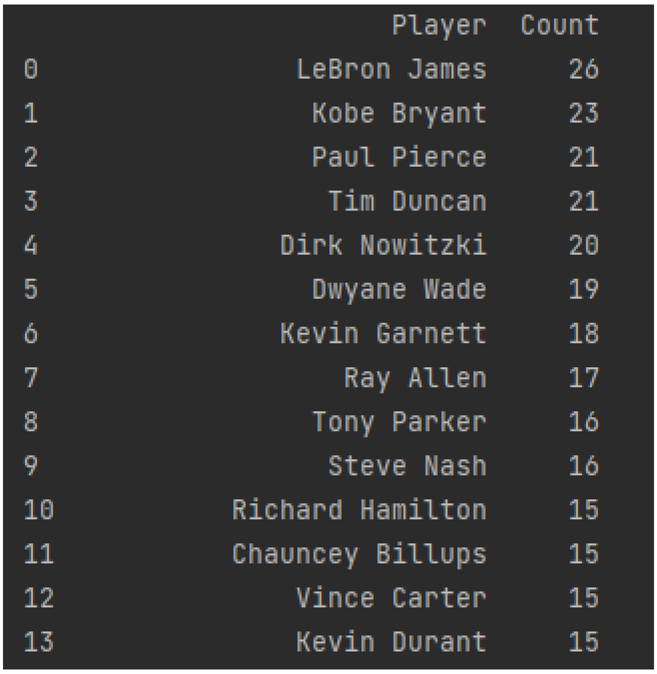
We also applied the mutual information (MI) metric, the Variance Inflation (VIF) measure, the Pearson Correlation Test and more, in order to start obtaining patterns or similarities in the data but the results initially were underwhelming: The actual problem was the lack of uniformity in the dataset.

5.1.1 The Best of the most consistent - approach

We initially started by trying to analyze the dataset in order to recognize important features that contribute in achieving a descent Win-Loss Ratio but the data contained more than 1500 records of players that were part of the last minutes of close games: The data was full of players that appeared in the games –but without regarding their individual performance, which means that as far as the statistical methods are concerned, the contribution of M. Jordan and L. Longley in the W/L Ratio of the Bulls games during the years 1997,1998 is the same, even if the latter rarely touched the ball during these critical moments of the last minutes. This had a devastating impact in our earliest models as they could not properly explain how players with low stats managed a much better W/L Ratio than other players with much superior clutch performance. In order to properly model the performance of the players (that also leads to a high win rate), we needed to find a way to use a compound of features that achieves a strong statistical presence and simultaneously leads to better W/L Ratio.

The first step was to filter the best performances and we achieved that through a combination of traditional Tendex metric with True Shooting(%), that we called “Initial Metric”. At this point, we would like to note that this metric was not meant for specific individual evaluation of the players. We tested different formulas that returned similar results in 90% of the cases, which means that the actual purpose of the formula was to give a vague but reliable list of the best performances along these 22 years.

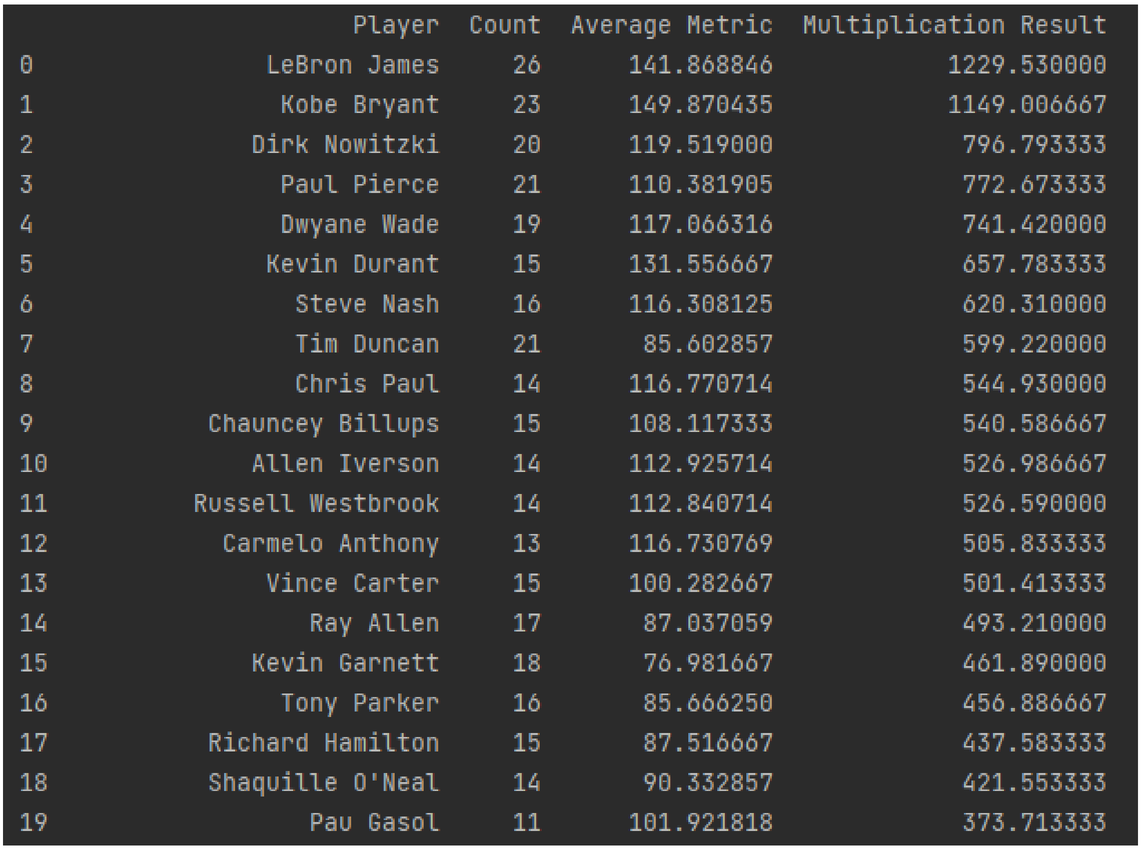
Then we added a count variable to the dataframe that indicated how many times a player appears in the aggregated list. As we have already mentioned, the data consists of 22 years of regular season and play-offs, so a player could at maximum appear 44 times (of course this is impossible as nobody in the NBA has played more than 21 consecutive years as of yet). The count list is illustrated in figure 25.



```
                  Player  Count
0           LeBron James     26
1            Kobe Bryant     23
2            Paul Pierce     21
3             Tim Duncan     21
4          Dirk Nowitzki     20
5            Dwyane Wade     19
6          Kevin Garnett     18
7              Ray Allen     17
8            Tony Parker     16
9             Steve Nash     16
10      Richard Hamilton     15
11      Chauncey Billups     15
12          Vince Carter     15
13          Kevin Durant     15
```

Figure 25

Subsequently, we took the average of the Initial Metric for every player and multiplied it by the number “count/3”. This was done in order to let the consistency of the players in the list play its role in the final results. So a player who had appeared more times in this period of the NBA would have higher rating than another with fewer appearances given their Initial Ratings would be equal. So our dataframe at this point was as follows in Figure 26:



```
               Player  Count  Average Metric  Multiplication Result
0        LeBron James     26      141.868846            1229.530000
1         Kobe Bryant     23      149.870435            1149.006667
2       Dirk Nowitzki     20      119.519000             796.793333
3         Paul Pierce     21      110.381905             772.673333
4         Dwyane Wade     19      117.066316             741.420000
5        Kevin Durant     15      131.556667             657.783333
6          Steve Nash     16      116.308125             620.310000
7          Tim Duncan     21       85.602857             599.220000
8          Chris Paul     14      116.770714             544.930000
9    Chauncey Billups     15      108.117333             540.586667
10      Allen Iverson     14      112.925714             526.986667
11  Russell Westbrook     14      112.840714             526.590000
12    Carmelo Anthony     13      116.730769             505.833333
13       Vince Carter     15      100.282667             501.413333
14          Ray Allen     17       87.037059             493.210000
15      Kevin Garnett     18       76.981667             461.890000
16        Tony Parker     16       85.666250             456.886667
17   Richard Hamilton     15       87.516667             437.583333
18   Shaquille O'Neal     14       90.332857             421.553333
19          Pau Gasol     11      101.921818             373.713333
```

Figure 26

Next we chose the leading 50 players based on the Multiplication Result and we returned in the original dataset and we filtered all the records to include only our desired list of top and consistent performers. The names of the players that made it to the list is the following:

['LeBron James', 'Kobe Bryant', 'Dirk Nowitzki', 'Paul Pierce', 'Dwyane Wade', 'Kevin Durant', 'Steve Nash', 'Tim Duncan', 'Chris Paul', 'Chauncey Billups', 'Allen Iverson', 'Russell Westbrook', 'Carmelo Anthony', 'Vince Carter', 'Ray Allen', 'Kevin Garnett', 'Tony Parker', 'Richard Hamilton', "Shaquille O'Neal", 'Pau Gasol', 'Chris Bosh', 'Joe Johnson', 'Jason Kidd', 'Kyrie Irving', 'Stephon

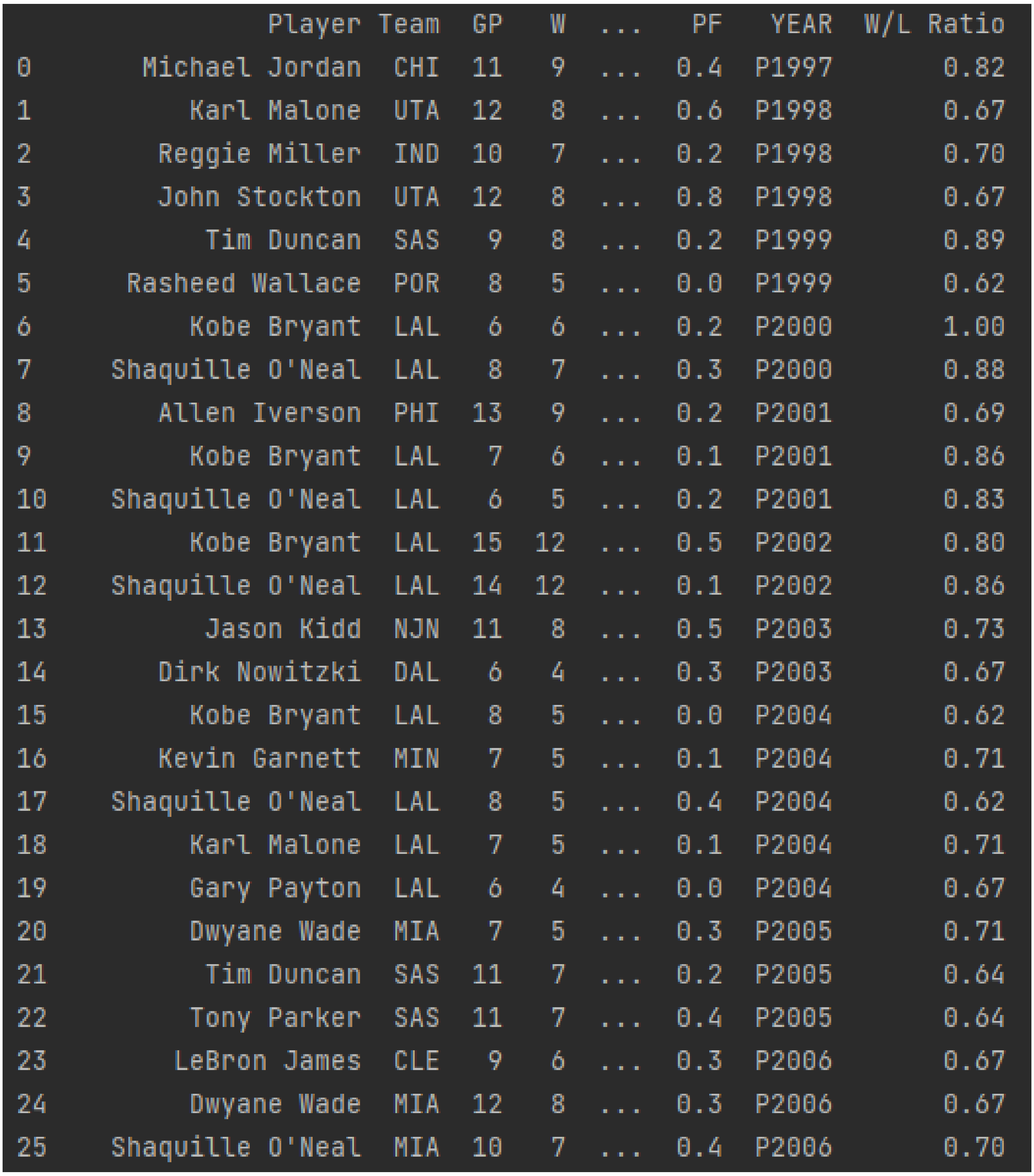
Marbury', 'Reggie Miller', 'Karl Malone', 'Stephen Curry', 'John Stockton', 'DeMar DeRozan', 'Monta Ellis', 'David West', 'Michael Jordan', 'Deron Williams', 'Kyle Lowry', 'Rasheed Wallace', 'Carlos Boozer', 'Al Jefferson', 'Damian Lillard', 'Chris Webber', 'Zach Randolph', 'Jason Terry', 'James Harden', 'Derrick Rose', 'Sam Cassell', "Amar'e Stoudemire", 'Andre Miller', 'John Wall', 'Gary Payton', 'Paul George']

We were satisfied with the result, as the list had a nice variety. Both younger (John Wall, Paul George) and older players (John Stockton, Michael Jordan) made it to the list through our algorithm (mainly based on their performances as they did not have too many appearances in this era), while less famous but consistent players like Carlos Boozer and Al Jefferson were not missing.

There was one final issue with the dataframe: It depicted the performances of the best and most consistent players of the period without regarding their teams' results. Since this dataset was built in order to use data analytics to investigate the major factors that contribute into winning during the 'clutch' period of the game and given the fact that players like Lebron or Kobe, who have 26 and 23 appearances would definitely have some good and some bad runs, we used the variable "W/L Ratio" in order to delete their less than average performances. So, in general, after some testing, we found that a limit of 0.6 "W/L Ratio" was good enough for our analysis.

To provide an example: Shaquille O'neal's 1998 playoff stats would be filtered out because he played 6 close games from which his team won only 1 (17%), while his stats in the 2000 playoff run would be added to the dataset, since from the 8 close games that he took part in, LAL managed to emerge victorious in 7 (88%).

What we achieved was for this specific dataset to include the individual performances of the best and most consistent 'clutch' players of the era, but only the cases where they managed to actually lead their teams in a descent ratio of Wins. A part of the final filtered dataset is shown in Figure 27.



```
              Player Team  GP   W  ...   PF   YEAR  W/L Ratio
0     Michael Jordan  CHI  11   9  ...  0.4  P1997       0.82
1        Karl Malone  UTA  12   8  ...  0.6  P1998       0.67
2      Reggie Miller  IND  10   7  ...  0.2  P1998       0.70
3      John Stockton  UTA  12   8  ...  0.8  P1998       0.67
4         Tim Duncan  SAS   9   8  ...  0.2  P1999       0.89
5    Rasheed Wallace  POR   8   5  ...  0.0  P1999       0.62
6        Kobe Bryant  LAL   6   6  ...  0.2  P2000       1.00
7   Shaquille O'Neal  LAL   8   7  ...  0.3  P2000       0.88
8      Allen Iverson  PHI  13   9  ...  0.2  P2001       0.69
9        Kobe Bryant  LAL   7   6  ...  0.1  P2001       0.86
10  Shaquille O'Neal  LAL   6   5  ...  0.2  P2001       0.83
11       Kobe Bryant  LAL  15  12  ...  0.5  P2002       0.80
12  Shaquille O'Neal  LAL  14  12  ...  0.1  P2002       0.86
13        Jason Kidd  NJN  11   8  ...  0.5  P2003       0.73
14     Dirk Nowitzki  DAL   6   4  ...  0.3  P2003       0.67
15       Kobe Bryant  LAL   8   5  ...  0.0  P2004       0.62
16     Kevin Garnett  MIN   7   5  ...  0.1  P2004       0.71
17  Shaquille O'Neal  LAL   8   5  ...  0.4  P2004       0.62
18       Karl Malone  LAL   7   5  ...  0.1  P2004       0.71
19       Gary Payton  LAL   6   4  ...  0.0  P2004       0.67
20       Dwyane Wade  MIA   7   5  ...  0.3  P2005       0.71
21        Tim Duncan  SAS  11   7  ...  0.2  P2005       0.64
22       Tony Parker  SAS  11   7  ...  0.4  P2005       0.64
23      LeBron James  CLE   9   6  ...  0.3  P2006       0.67
24       Dwyane Wade  MIA  12   8  ...  0.3  P2006       0.67
25  Shaquille O'Neal  MIA  10   7  ...  0.4  P2006       0.70
```

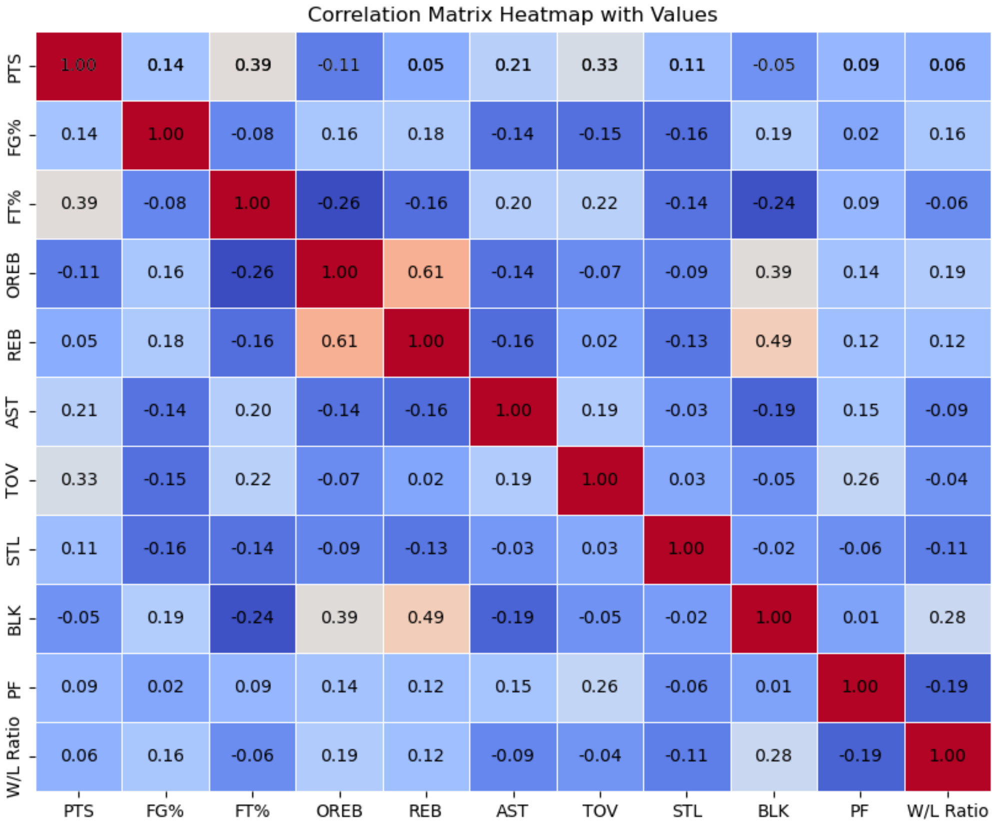
Figure 27

As we can observe in Figure 27, the dataset that we will work to build our performance model consists of some of the best individual clutch performances compared with great team results. Jordan in P1998 champion, Malone and Stockton in P1998 made it to the finals, Kidd in P2003 made it to the finals, of course Kobe and Shaq from 2000-2004 had a great couple of years. All in all, this dataset is made from the greatest and the most consistent winners. A last note, the final dataset consisted of 212 records.

5.1.2 Variable Analysis

- **Correlation**

With the new dataset the **correlation** table looks much more insightful (figure 28).



	PTS	FG%	FT%	OREB	REB	AST	TOV	STL	BLK	PF	W/L Ratio
PTS	1.00	0.14	0.39	-0.11	0.05	0.21	0.33	0.11	-0.05	0.09	0.06
FG%	0.14	1.00	-0.08	0.16	0.18	-0.14	-0.15	-0.16	0.19	0.02	0.16
FT%	0.39	-0.08	1.00	-0.26	-0.16	0.20	0.22	-0.14	-0.24	0.09	-0.06
OREB	-0.11	0.16	-0.26	1.00	0.61	-0.14	-0.07	-0.09	0.39	0.14	0.19
REB	0.05	0.18	-0.16	0.61	1.00	-0.16	0.02	-0.13	0.49	0.12	0.12
AST	0.21	-0.14	0.20	-0.14	-0.16	1.00	0.19	-0.03	-0.19	0.15	-0.09
TOV	0.33	-0.15	0.22	-0.07	0.02	0.19	1.00	0.03	-0.05	0.26	-0.04
STL	0.11	-0.16	-0.14	-0.09	-0.13	-0.03	0.03	1.00	-0.02	-0.06	-0.11
BLK	-0.05	0.19	-0.24	0.39	0.49	-0.19	-0.05	-0.02	1.00	0.01	0.28
PF	0.09	0.02	0.09	0.14	0.12	0.15	0.26	-0.06	0.01	1.00	-0.19
W/L Ratio	0.06	0.16	-0.06	0.19	0.12	-0.09	-0.04	-0.11	0.28	-0.19	1.00

Figure 28

Except the obvious relations between REB and OREB (and BLK as they are commonly achieved by big players), we can also observe that the calculated correlation between OREB and W/L Ratio is quite larger than that between REB and W/L Ratio, which means that our dataset has managed to uncover the importance of offensive rebounding in clutch situations. Moreover, the FG% statistic seems to be exponentially more important than the PTS category as far as W/L Ratio is concerned, which comes as no surprise since even players that lose many shots can achieve descent scoring numbers, but their lack of accuracy could lead their team to defeat. Another remark would be the high correlation between BLK and W/L Ratio; this tendency has been appearing throughout all of our tests and in almost all our datasets and was something not entirely expected. It seems that these really big defensive plays are very important (especially) at the closing of the game, something that we will explore even more. Finally, we would like to mention the negative relations of TOV (turnovers) and PF (personal fouls) with W/L Ratio. Although turnovers are self-explanatory, the actual significance of personal fouls and the impact they have on the performance of the players or the outcome of the game is quite more complicated. Actually, we can assume theoretical causation between these variables, in the context that the teams that are chasing in score, tend to make more fouls to stop the clock. These are intentional fouls which have nothing to do with performance. A smart foul –or a bad foul could be a good indicator as a clutch play, but unfortunately there is no way to distinguish between these fouls and the intentional ones. So we decided to drop this variable from our analysis, to avoid wrong implications.

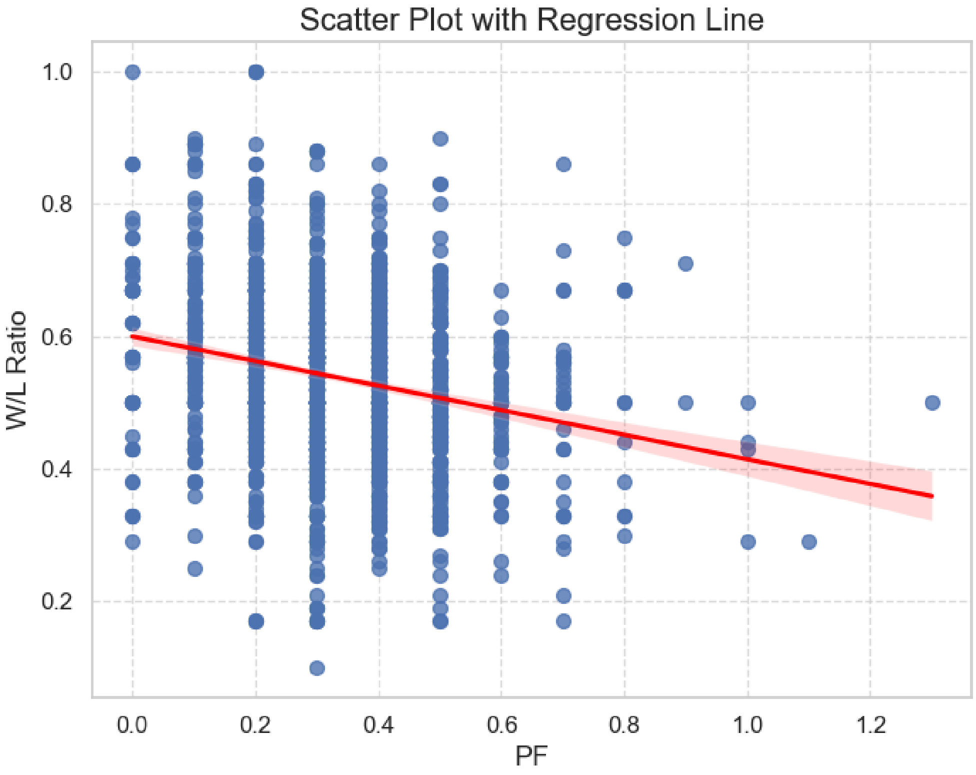
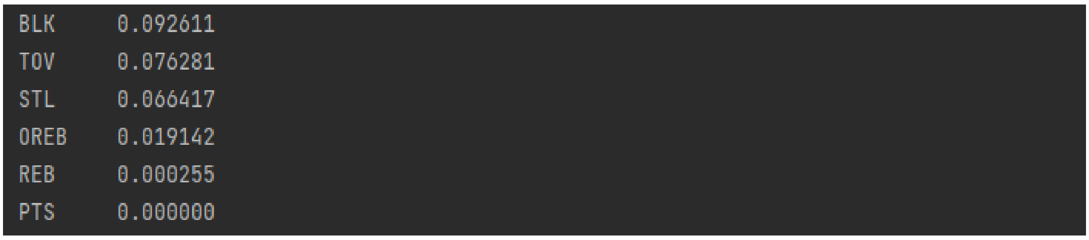


Figure 29

Figure 29 hints at a linear relationship between fouls commited and W/L ratio. Unfortunately, this can be explained by the intentional fouls tactics to stop the clock.

- **Mutual Information**

Another interesting experiment was through the use of the Mutual Information Measure. Mutual Information (MI) calculates the amount of information shared between two variables. In more detail, it tries to quantify how much our knowledge of one variable (e.g. an independent variable) reduces the uncertainty about the other variable (e.g. the target variable). So in machine learning it can be applied to assess the dependency between a feature and the target variable. Figure 30 depicts the results of the use of the non-parametric version of MI in our dataset concerning the W/L Ratio variable:



```
BLK     0.092611
TOV     0.076281
STL     0.066417
OREB    0.019142
REB     0.000255
PTS     0.000000
```

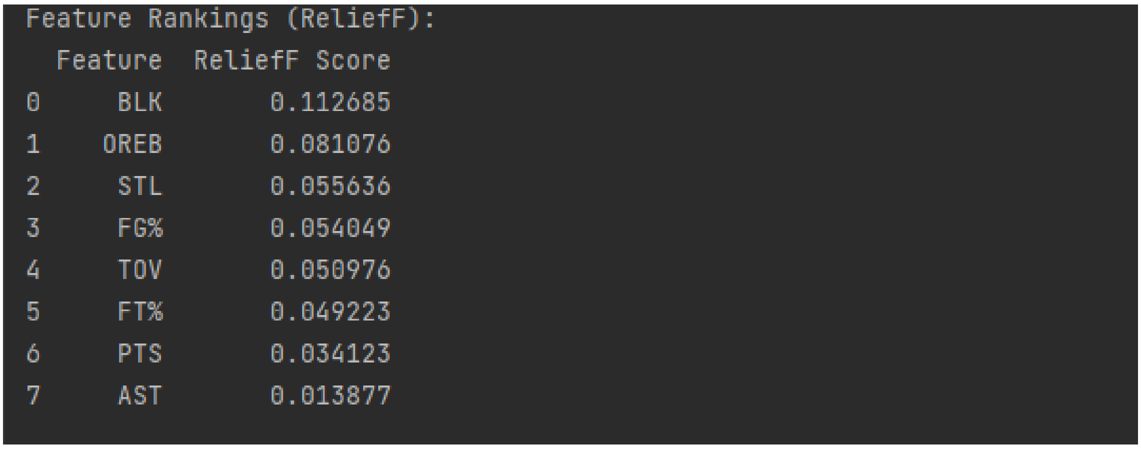
Figure 30

As we can observe the results again show that the W/L Ratio has a much stronger relationship with offensive rebounding than overall (and defensive) rebounding during the last moment of the game. Other than that, blocks again show up as a dominant factor, while turnovers and steals follow.

So following the above test, we decided to keep OREB and drop REB for the purpose of the feature ranking calculations. Both from empirical standpoint and now from statistical one, we have shown that REB is less significant than OREB given their dependency, and as we have already mentioned we consider offensive rebounding a much stronger late game personal statistic, than defensive rebounding which is mainly a team effort (contest shots, block-outs etc.).

- **ReliefF Algorithm**

ReliefF is method that attempts to evaluate the importance of features in the dataset by comparing the values of instances with their neighbors. It is underlying assumption entails that attributes that are contributing towards the diversity of the instances in the feature space are more probable to be significant for the regression task. In general, ReliefF can be applied for feature ranking, assisting in identifying the most prominent features for predicting a continuous target variable.



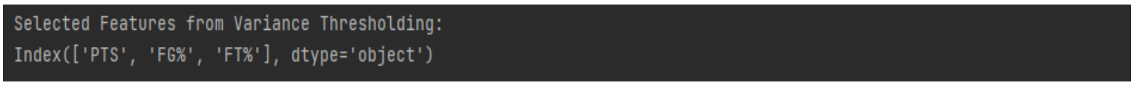
```
Feature Rankings (ReliefF):
  Feature  ReliefF Score
0     BLK       0.112685
1    OREB       0.081076
2     STL       0.055636
3     FG%       0.054049
4     TOV       0.050976
5     FT%       0.049223
6     PTS       0.034123
7     AST       0.013877
```

Figure 31

Figure 31 shows the results of the ReliefF Algorithm: We notice that blocking is still one of the most prominent features that contribute to better W/L Ratio, and generally the results have not been much different than the previous methods. We started discerning a pattern in the data.

- **Variance Thresholding**

Variance Thresholding is simple metric in variable selection that is used to filter the variables based on their variance to determine which of them surpass a certain limit (threshold). The primary notion is that attributes with low variance contribute less information towards the target variable, which in turn may be used to filter them out or adjust the relevant weights in the model. In our case given a threshold variance of **0.05**, the results were the following (figure 32):



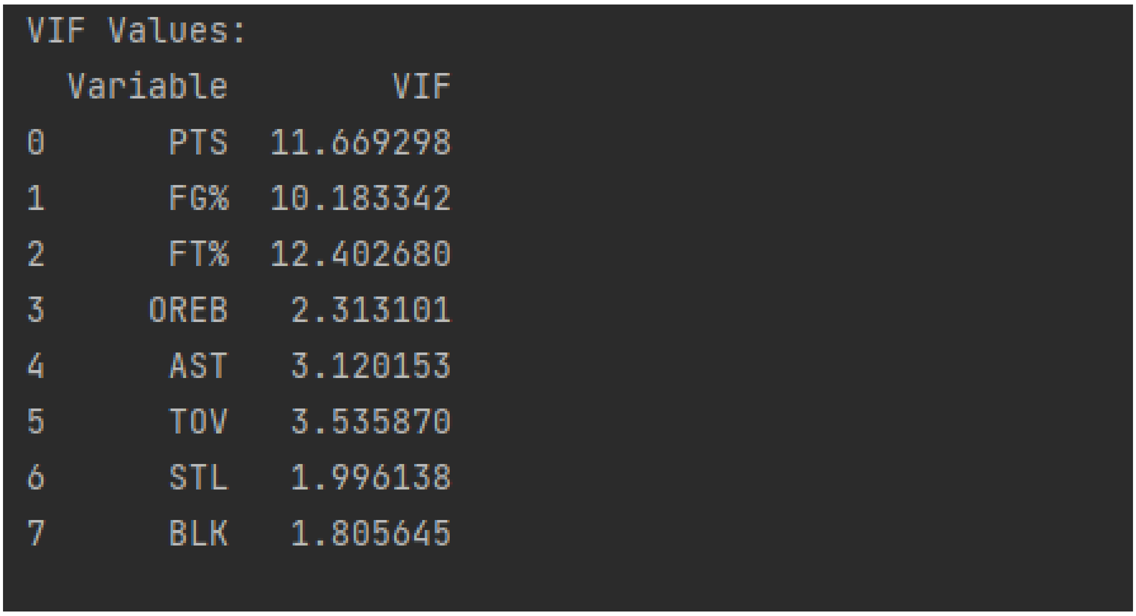
```
Selected Features from Variance Thresholding:
Index(['PTS', 'FG%', 'FT%'], dtype='object')
```

Figure 32

So we can see that the features PTS, FG% and FT% have the highest variability in the dataset, and despite the fact that according to our previous tests they had less impact on win percentage, they still important information for the model and they should not be omitted.

- **Variance Inflation Factor (VIF)**

Variable Inflation (VIF) is a statistical measure that calculates the level of collinearity between independent variables in regression analysis. Multicollinearity is defined by two or more features being highly correlated, which makes it more difficult to comprehend the individual impact of each of these features towards the target variable. In statistics, a value of 1 indicates no collinearity while a value of 5 or 10 or higher is considered indicative of problematic levels of collinearity that may increase standard error, and cause lower precision in coefficient estimation in the models. Let's see the result in our experiment in figure 33:

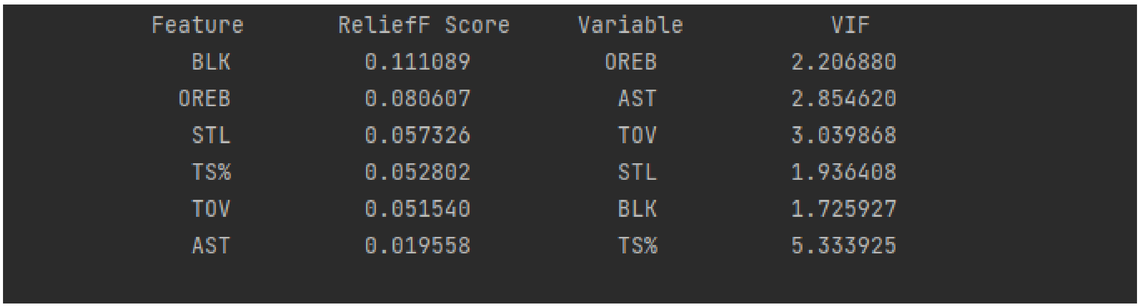


```
VIF Values:
  Variable        VIF
0      PTS  11.669298
1      FG%  10.183342
2      FT%  12.402680
3     OREB   2.313101
4      AST   3.120153
5      TOV   3.535870
6      STL   1.996138
7      BLK   1.805645
```

Figure 33

Expectedly, the levels of collinearity between PTS, FG%, FT% in the datasets were very high. Although this has a logical basis, as a player is expected to score points when he has higher shooting percentages, it was amplified by our filtered dataset since we only took the best performances from the top players.

To deal with this, we decided to merge these features into one that encompasses them all sufficiently: True Shooting Percentage (TS%). We analyzed this statistic thoroughly in the literature review [53] and we tried using it for our own purpose. Let's see the results (figure 34):



```
Feature       ReliefF Score     Variable          VIF
   BLK          0.111089          OREB         2.206880
  OREB          0.080607           AST         2.854620
   STL          0.057326           TOV         3.039868
   TS%          0.052802           STL         1.936408
   TOV          0.051540           BLK         1.725927
   AST          0.019558           TS%         5.333925
```

Figure 34

So conclusively the new set of VIF values along with the other measures indicate a significant improvement in terms of collinearity and correlation scores. The inclusion of the new variable, True Shooting Percentage (TS%) seems to have contributed to a more balanced set of VIF values, which generally lead to more stable coefficient estimates in the regression models.

Finally, with this modification, our variable selection is finalized and we can proceed to model testing.

5.2 Modelling

In this section we are going to test different machine learning models in order to find the better fits and use the relevant coefficients to define our own performance metric.

5.2.1 LASSO (Least Absolute Shrinkage and Selection Operator)

Lasso algorithm is useful for variable selection, as well as regularization. In general, the method appends a penalty to the linear regression function, and attempts to minimize the squared residuals under the restrictions that the sum of the absolute coefficient values does not supersedes a particular constant. This penalty term is defined by the user and is known as parameter Alpha (a). All in all, through the use of the parameter its purpose is to reduce the significance of irrelevant features and helps with overfitting by limiting the coefficients. The assignment of the parameter value is crucial for the task and should be tuned according to the relevant dataset.

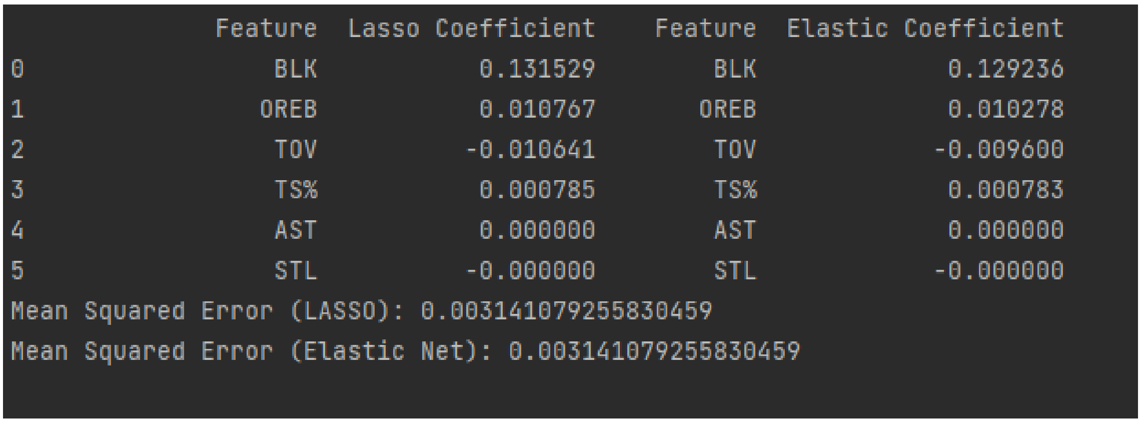
To calculate the optimal Alpha for our dataset, we used 10-fold Cross Validation for the values from 0.0001 to 100. Eventually the optimal value was found Alpha = 0.0006.

Using this value we separated the data into training and test data in proportions 80%-20% and applied the model. The results of the coefficients are shown in figure 35. We can see that blocks were by far the most influential factor while offensive rebounds and turnovers followed.

5.2.2 Elastic Net

Elastic Net is a regression technique similar to Lasso that is also used for variable selection and regularization. It actually uses a twofold L1/L2 regularization, where L1 is the Lasso technique that attempts to nullify the impact of certain variables by assigning their coefficients to 0 and L2 is derived by the Ridge technique and its purpose is to deal with possible collinearity issues by penalizing the sum of the square of the coefficients. In Elastic Net there are two (2) hyper parameters: Alpha (a) which defines the overall strength of the penalty and L1_ratio which regulates the balance between L1 and L2 values. So if L1_ratio is close to 1 then L2 is almost 0 and the Elastic Net gives similar results with the Lasso algorithm. In case L1 is 0 then L2 is 1 and the Elastic Net resembles the Ridge algorithm. And for all the values in between there is a combination of the two.

In our case we again used 10-fold Cross Validation to determine the best values for the hyper-parameters. The optimal alpha value turned out to be Alpha = 0.0007, while the best L1_ratio was found to be L1_ratio = 0.9. This actually meant that the best Elastic model would very much approximate the Lasso model, as illustrated in figure 35.

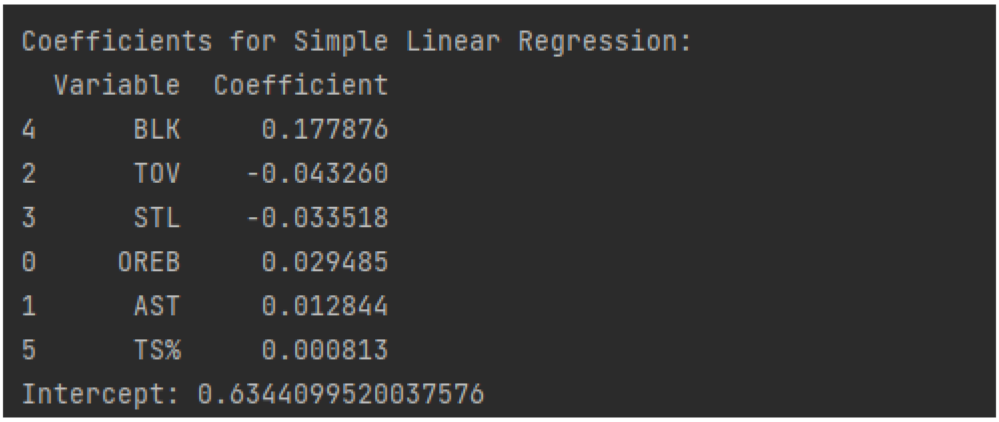


```
             Feature  Lasso Coefficient    Feature  Elastic Coefficient
0                BLK           0.131529        BLK             0.129236
1               OREB           0.010767       OREB             0.010278
2                TOV          -0.010641        TOV            -0.009600
3                TS%           0.000785        TS%             0.000783
4                AST           0.000000        AST             0.000000
5                STL          -0.000000        STL            -0.000000
Mean Squared Error (LASSO): 0.003141079255830459
Mean Squared Error (Elastic Net): 0.003141079255830459
```

Figure 35

5.2.3 Simple Linear Regression

Both Lasso and Elastic Net are actually based in linear regression, so we did not expect much different results from simple linear regression. We again followed the 80-20 split of the data, and the outcome was not surprising (figure 36) although the (negative) coefficient for the TOV (turnover) category was much higher, which actually hints at how important are the mistakes during the final moments of the game.



```
Coefficients for Simple Linear Regression:
  Variable  Coefficient
4      BLK     0.177876
2      TOV    -0.043260
3      STL    -0.033518
0     OREB     0.029485
1      AST     0.012844
5      TS%     0.000813
Intercept: 0.6344099520037576
```

Figure 36

5.2.4 Decision Trees and Random Forest

After the linear parametric methods we saw above, we decided to test the non-parametric decision tree and random forest techniques. These models make limited or, no assumptions at all, about the underlying distribution and they are not defined by a fixed number of parameters. Decision trees perform splits based on the data characteristics alone, while random forests consist of an ensemble of Decision Trees, with each tree in the forest built upon a specific subset of the data, a characteristic which makes Random Forest quite robust against overfitting.

In our test we continue to use the 80-20 split of the data, and in the random forest after a few tries, we settled at the hyper-parameter: n=100 estimators. The following figure (figure 37) represents the model results concerning the importance of the variables.

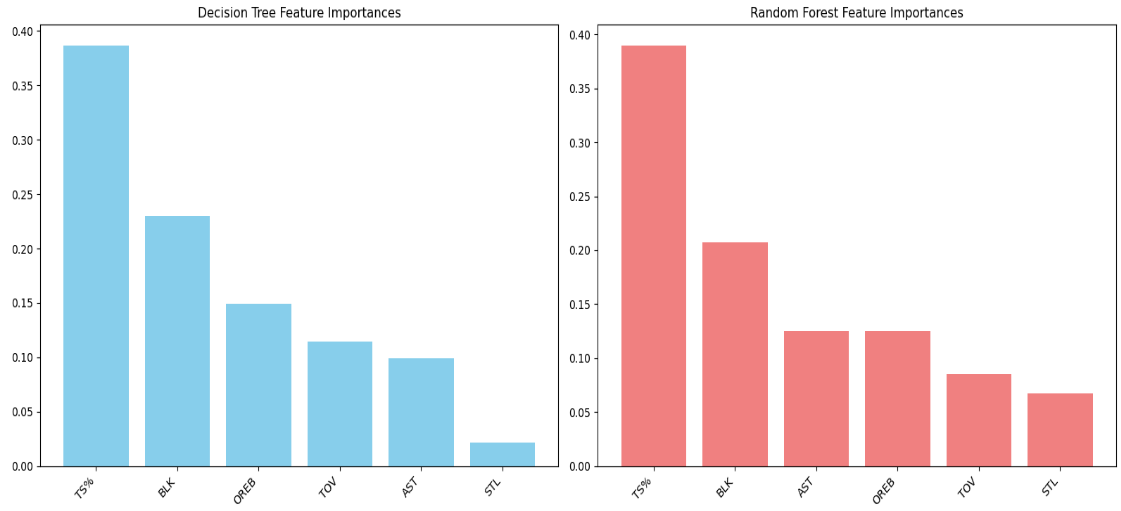


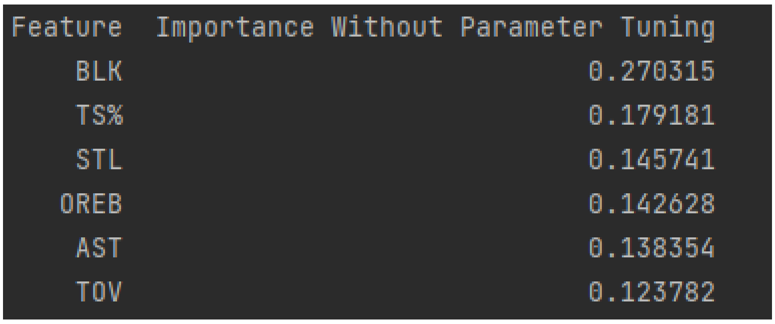
Figure 37

Evidently, these methods have captured different peculiarities inside the data compared to the previous algorithms; it is the first time that the TS% shows stronger bond with the W/L Ratio than every other variable (and with a large margin). Taking into consideration the inner workings of the models, we can reasonably claim that these models have caught the pattern that high scoring and percentages tend to lead to better results, totally ignoring the cases that they do not – in comparison with the parametric models, where the coefficient of the TS% variable was heavily mitigated by the opposing cases. That being said, the BLK feature is following in significance establishing the notion that clutch defensive plays can dictate the outcome of the games.

5.2.5 XGBoost (Extreme Gradient Boosting)

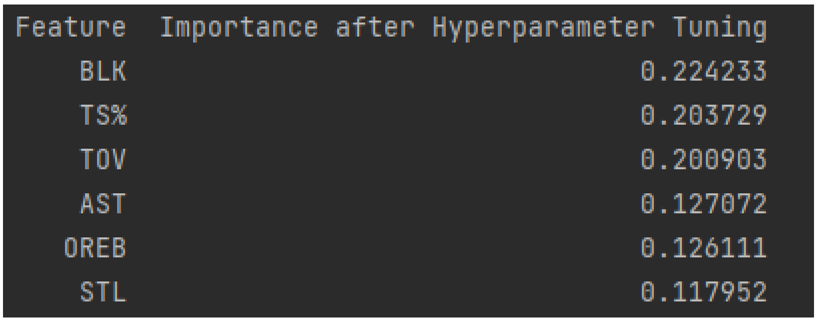
XGBoost algorithm belongs to the family of gradient boosting, a technique that constructs a model by using an ensemble of weak learners in sequence (e.g. decision trees), forming a new and improved model step by step by correcting the mistakes of the previous versions of the model. It contains a loss function, the L1 and L2 regularization terms that we already met at Elastic Net to penalize large coefficients. Another important property for our case is that it provides a clear measure of feature importance based on the number of times that a specific feature is used to split the dataset across all trees in the model. In effect, XGBoost is considered a non-parametric technique, though there are a handful of hyper-parameters that actually dictate the amount of trees, their depth, as well as other characteristics.

In our dataset we tested the XGBoost Algorithm extensively, since it provided quite interesting results. Figures 38 and 39 portray the feature importances of two versions of the model: the first one without any hyper-parameter tuning, while the second one after running 10-fold cross validation parameter tuning on the data.



```
Feature  Importance Without Parameter Tuning
    BLK                             0.270315
    TS%                             0.179181
    STL                             0.145741
   OREB                             0.142628
    AST                             0.138354
    TOV                             0.123782
```

Figure 38



```
Feature  Importance after Hyperparameter Tuning
    BLK                                0.224233
    TS%                                0.203729
    TOV                                0.200903
    AST                                0.127072
   OREB                                0.126111
    STL                                0.117952
```

Figure 39

We noticed a significant improvement in the feature importances after the proper hyper parameter tuning. We can see that the significance of BLK was heavily decreased (almost by 25%) while the importance of TS% and TOV increased considerably. This was one of the most successful models that we tested, it described the data reasonably well and the conclusions converged with our testing and domain knowledge.

5.2.6 Other Algorithms

We also tested a couple of more techniques but the results were underwhelming and there is no reason to incorporate them in this study. These algorithms included the discriminant stepwise regression method, the Gaussian processes method, and even a couple of neural network attempts.

5.2.7 Conclusions

The purpose of this extensive testing and modeling was to investigate the most dominant individual performance factors that contribute in higher win rate for the team in close games. The models could not (and were not expected to) be able to predict the outcome of games with high precision; basketball is a team sport and personal performance could never substitute or surpass the importance of team effort.

That being said, we uncovered characteristics in the personal stats of a player that can improve the chances of a team to win a clutch game. We managed to discover the most significant factors that are highly correlated with higher win-rate in close games and we found out which algorithms comprise a good fit for the data in question, and which could not provide relevant results or sufficient conclusions.

In the final chapter, we are going to define our own clutch performance metric based on the observations we made so far, and apply it to the original sets, to rank and evaluate the best clutch performances in the NBA from 1997 to 2018.

6 Player Evaluation

In this part, we will present our own clutch performance formula, which is grounded both upon our domain knowledge and basketball sense, as well as the machine learning deductions we have established so far.

6.1 Estimation of Clutch Competency (EoCC)

To define a meaningful statistic that can adequately describe the ability to perform well under pressure in the basketball court, we have to start from scoring. It is by definition the most important category of the game: if the players cannot score, the team cannot win. The fans celebrate the heroic players who manage to score the winning bucket, the press praises or condemns the corresponding shot, and statistically having good scoring with descent accuracy highly improves the chances of success. So the first term would be the Points [PTS] statistic, mediated by True Shooting (TS%).

Subsequently, to limit opponent's scoring the team has to present strong defensive capabilities. In personal terms, and given the scope of our dataset that would mean Blocks [BLK] (as suggested by our previous work) and Steals [STL] with relevant coefficients based on the machine learning algorithm results.

Next we have offensive rebounds, and assists: [OREB] can offer a second chance for a team offensive, which cannot be underestimated while [AST] is less obvious, since we managed to find no clear indication of its impact on the winning percentage or if team effort in these critical moments is actually more effective than personal attempts.

Finally, we will multiply the formula with a Turnover [TOV] coefficient, since last minute mistakes are detrimental to the chances of winning: Every algorithm, every test we tried showed a clear negative relation between winning percentage and turnovers. There is also the psychological element of a clutch turnover, a factor that has led many teams to lose even double figure advantages in a few moments.

The formula we ended up using our notes and after some tuning was the following:

Estimation of Clutch Competency (EoCC):

$$([PTS]^2 * \frac{[TS\%]}{100} * 1.2 + 1.4*(1.75[BLK]+[STL]) + 2[OREB] + [AST])*(0.625-[TOV])$$

Some notes on the formula:

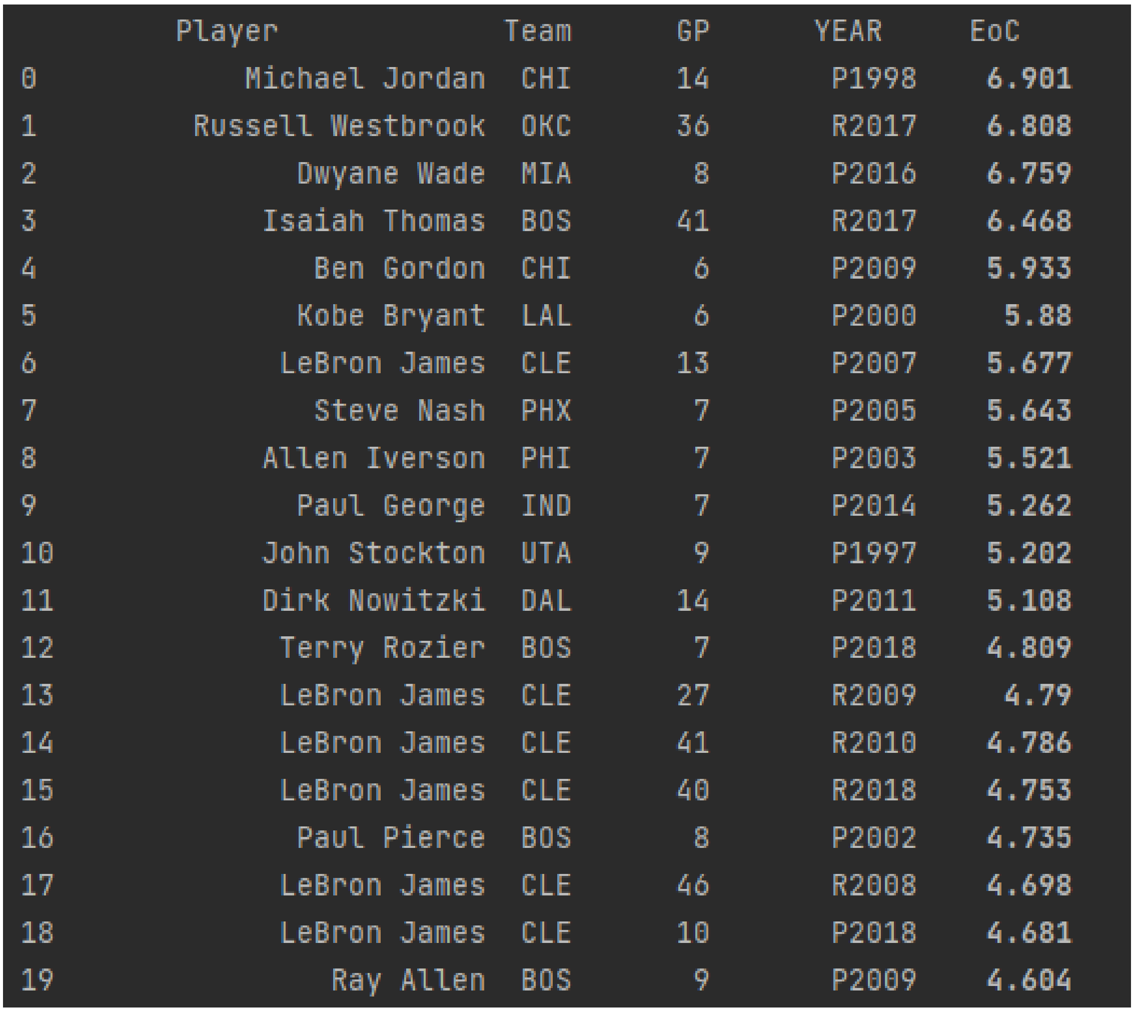
The most important variable is [PTS] with [BLK] and [TS%] next, as suggested by the modelling. The final 3 in order of significance are [OREB], [STL] and [AST].

The most intriguing variable is the [TOV]. The turnover statistic plays the role of the modifier in the formula: a high value is detrimental for the performance rating (as the average turnovers of the player approximate the value 0.625* the rating will converge to effectively 0) while a low [TOV] will barely affect the rating.

*99.5% of the players have lower [TOV] value than 0.625, while the average throughout the records is only 0.16. That being said it was imperative to penalize a high turnover count.

6.2 Overall Player Evaluation

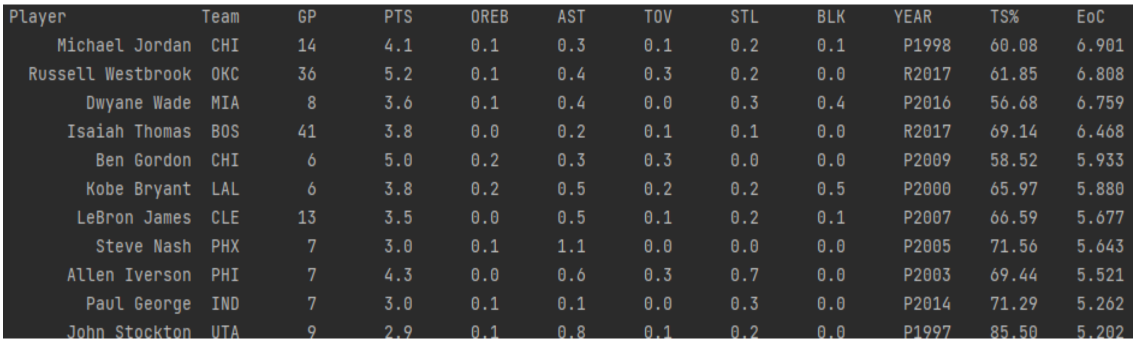
We started with an overall evaluation of the performance of the players along all datasets (1997-2018). The results (figure 40) uncovered some unexpected records: After M. Jordan's exemplary 1998 Playoff performance, in the second place ranks R. Westbrook at the 2017 regular season. In 36 close games, Westbrook achieved (top of the whole dataset) 5.2 average points, 0.4 assists and even 61% True Shooting. His only misplay was 0.3 average turnovers, which mediated his score considerably. In other notes, Isaiah Thomas managed great numbers throughout the whole 2017-18 regular season, while in a typical Lebron fashion, 6 of the 20 best performances during these 22 years belong to him.



```
             Player          Team     GP     YEAR     EoC
0       Michael Jordan  CHI     14     P1998   6.901
1    Russell Westbrook  OKC     36     R2017   6.808
2          Dwyane Wade  MIA      8     P2016   6.759
3       Isaiah Thomas  BOS     41     R2017   6.468
4           Ben Gordon  CHI      6     P2009   5.933
5          Kobe Bryant  LAL      6     P2000    5.88
6         LeBron James  CLE     13     P2007   5.677
7           Steve Nash  PHX      7     P2005   5.643
8       Allen Iverson  PHI      7     P2003   5.521
9          Paul George  IND      7     P2014   5.262
10       John Stockton  UTA      9     P1997   5.202
11      Dirk Nowitzki  DAL     14     P2011   5.108
12        Terry Rozier  BOS      7     P2018   4.809
13        LeBron James  CLE     27     R2009    4.79
14        LeBron James  CLE     41     R2010   4.786
15        LeBron James  CLE     40     R2018   4.753
16         Paul Pierce  BOS      8     P2002   4.735
17        LeBron James  CLE     46     R2008   4.698
18        LeBron James  CLE     10     P2018   4.681
19           Ray Allen  BOS      9     P2009   4.604
```

Figure 40

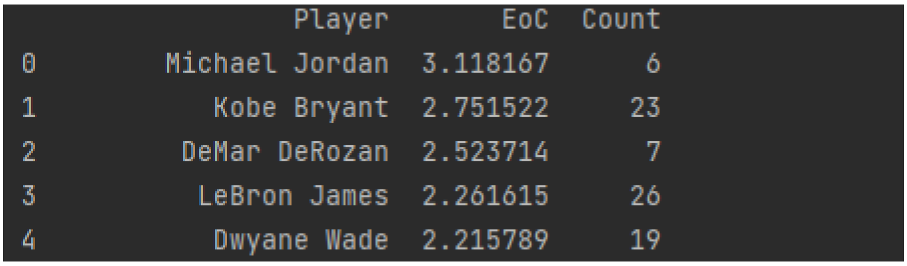
A very interesting point that we can raise from the list, is that the majority of the strongest performances were achieved in cases where the team had a "lone superstar" in the roster, who was the go-to player during the last moments of the games. And that is exactly where our models had difficulties adjusting, because these teams rarely managed to be championship winners. Let's elaborate: Lebron's appearances in the top-20 list are entirely with the Cleveland team (2007-08-09-10-18) and specifically never when he was accompanied by a top-player like Kyrie Irving. Allen Iverson was leading Philadelphia 76ers for years without major help, Westbrook achieved extreme stats (averaged a triple-double) following Durand's departure, Dirk's second best player was J. Kidd at age 38 and even Jordan was almost always the universal choice for the Bulls when the game was on the line. Notable exceptions are Stockton 1997 playoff stats and K. Bryants' 2000 playoff run, where they had help from Malone and Shaquille O Neal respectively. Interstingly, Kobe's ranking was elevated especially due to the defensive factor (figure 41), while Stockton averaged a top 0.8 assists with 84% True Shooting, which hint that they were not the exclusive choice in their teams.



```
Player              Team     GP      PTS     OREB    AST     TOV     STL     BLK     YEAR    TS%     EoC
    Michael Jordan  CHI      14      4.1     0.1     0.3     0.1     0.2     0.1     P1998   60.08   6.901
 Russell Westbrook  OKC      36      5.2     0.1     0.4     0.3     0.2     0.0     R2017   61.85   6.808
       Dwyane Wade  MIA       8      3.6     0.1     0.4     0.0     0.3     0.4     P2016   56.68   6.759
     Isaiah Thomas  BOS      41      3.8     0.0     0.2     0.1     0.1     0.0     R2017   69.14   6.468
        Ben Gordon  CHI       6      5.0     0.2     0.3     0.3     0.0     0.0     P2009   58.52   5.933
       Kobe Bryant  LAL       6      3.8     0.2     0.5     0.2     0.2     0.5     P2000   65.97   5.880
      LeBron James  CLE      13      3.5     0.0     0.5     0.1     0.2     0.1     P2007   66.59   5.677
        Steve Nash  PHX       7      3.0     0.1     1.1     0.0     0.0     0.0     P2005   71.56   5.643
     Allen Iverson  PHI       7      4.3     0.0     0.6     0.3     0.7     0.0     P2003   69.44   5.521
       Paul George  IND       7      3.0     0.1     0.1     0.0     0.3     0.0     P2014   71.29   5.262
     John Stockton  UTA       9      2.9     0.1     0.8     0.1     0.2     0.0     P1997   85.50   5.202
```

Figure 41

Another query that we had was to explore the most consistent top clutch performers throughout the analyzed seasons. To achieve this, we created a new dataframe with the count of the appearances of every player in the list and their average EoC score. To preserve the significance of consistency we only included in our set players that had at least 6 records. Figure 42 provides some new insights:



```
             Player       EoC  Count
0    Michael Jordan  3.118167      6
1       Kobe Bryant  2.751522     23
2     DeMar DeRozan  2.523714      7
3      LeBron James  2.261615     26
4       Dwyane Wade  2.215789     19
```

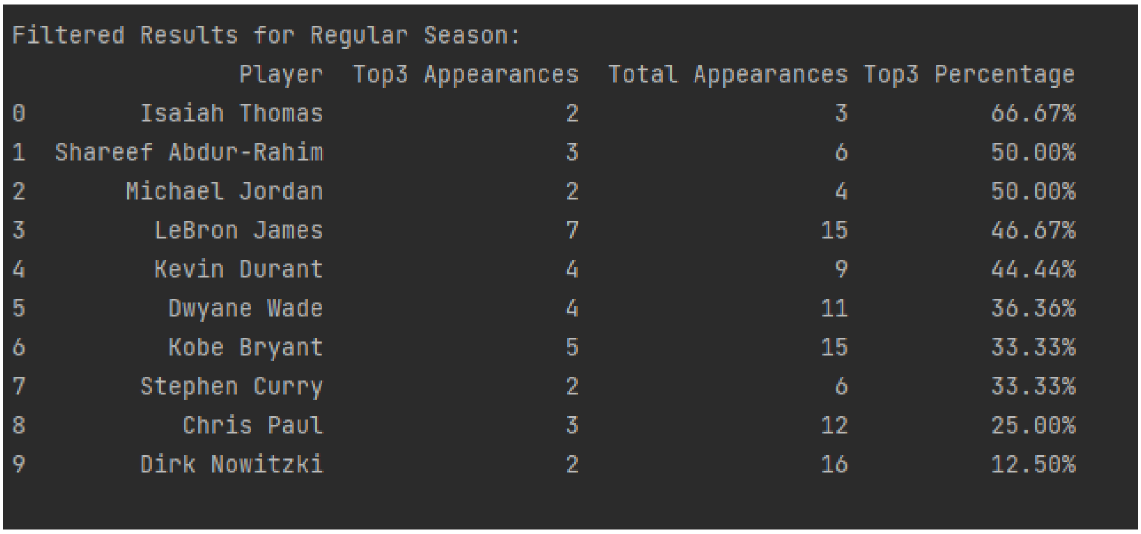
Figure 42

As we can see Jordan's average is the highest (4 runs with Bulls, 2 regular seasons with the Wizards), while Kobe is ranked close second, with DeRozan, Lebron and Wade following. DeRozan's numbers, although a bit of a surprise, were not entirely unexpected. DeMar DeRozan was the go-to player in Toronto between 2014 and 2018 and had very reliable clutch statistics. Having the reputation of a modern mid-range specialist, DeRozan's jumper could provide a valuable scoring option when the three-point opportunities may be limited or when a higher percentage shot is needed to secure crucial points. Taking into consideration the top-5 list, it raises the question whether the mid-range is the better choice during those late game scenarios. All of the mentioned players excel in mid-range and despite the fact that Lebron and Kobe relied also in consistent 3-point scoring, it was never their specialization. Investigating the correlations between the choice of shot in late game situations, and how this affects the outcome of a game will be a future goal to continue on this research.

6.3 Yearly Ratings, Top 3 Evaluation

In this section, we wanted to analyze in more detail the consistency of the top clutch players and how often they appeared in our list. For this task, we made lists with the Top 3 clutch rating based on EoCC of every occasion. We ended up with 44 lists of 3 players: 22 for the playoff years and 22 for the regular season. Then we calculated the count of the respective player appearances in our Top-3 lists divided by the count of their appearances in the whole dataset. This way we could calculate a percentage of how often these star players managed to achieve top performance in order to evaluate their consistency.

In figure 43 we can see the result based on our Regular Season data, while in figure 44 are the respective results from the Playoff data.

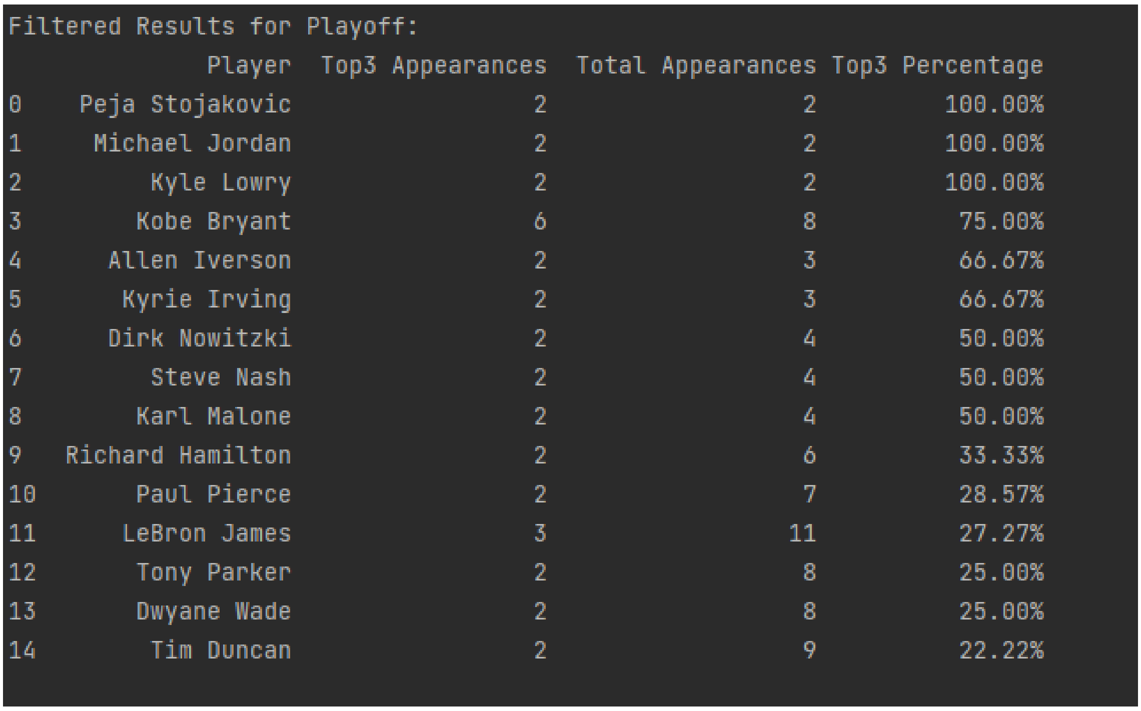


```
Filtered Results for Regular Season:
                Player  Top3 Appearances  Total Appearances Top3 Percentage
0        Isaiah Thomas                 2                  3          66.67%
1  Shareef Abdur-Rahim                 3                  6          50.00%
2       Michael Jordan                 2                  4          50.00%
3         LeBron James                 7                 15          46.67%
4         Kevin Durant                 4                  9          44.44%
5          Dwyane Wade                 4                 11          36.36%
6          Kobe Bryant                 5                 15          33.33%
7        Stephen Curry                 2                  6          33.33%
8           Chris Paul                 3                 12          25.00%
9        Dirk Nowitzki                 2                 16          12.50%
```

Figure 43

We can notice that the highest percentages are associated with lower number of appearances as far as the Regular Season dataset is concerned – but the results of Lebron James, Kevin Durant, Dwyane Wade, Kobe and Stephen Curry are remarkable. Lebron James managed to rank among the best 3 clutch performers throughout half of his long career. Another note for the regular season leaders was that the list of the Top-3 performers was looking like a private club for the most part. The same names appeared again and again. That can also be argued about: While in the relatively small scope of the playoffs, some rotation players may manage to surpass their usual stats and achieve average numbers comparable to the numbers of the superstars, in the long regular season, the ball will end up to the hands of the best player in the vast majority of cases. Furthermore, the defense in the playoffs in notoriously harder, especially on the stars of the opposing team, which gives the opportunity to rotation players have their “moment in the sun”. For example, in the 1997 finals, Jordan got double-teamed in the last moments of the last game, as J. Stockton rushed towards him, leaving Kerr alone. Jordan passed the ball, Steve Kerr took the shot and the rest is history.

The relevant data from the playoffs provides some other insights though. There are more players in the Top-3 lists, and especially lower profile players like Peja Stojakovic and Kyle Lowry* seemed to have a knack for this kind of close games. Nevertheless, the most intriguing performance of this table is Kobe Bryant who managed to achieve great numbers 6 out of his 8 playoff runs, which makes him by far the most dominant clutch playoff player of this era. Although, this comes at little or no surprise, it turns out Kobe consistently outperformed himself in the playoffs, when compared to the regular season. To present an analogy, since here we have two of the greatest players of all time, our data indicate that Kobe (over)doubled his percentage of being in the Top3 from 33% to 75% when in playoffs while Lebron almost halved his own from 46% to 27%. That is really quite impressive given that Kobe has 3 times more appearances in the Top-3 lists from almost all his competitors and speaks volumes about his resolve, determination and winning mentality.



```
Filtered Results for Playoff:
              Player  Top3 Appearances  Total Appearances Top3 Percentage
0    Peja Stojakovic                 2                  2         100.00%
1     Michael Jordan                 2                  2         100.00%
2         Kyle Lowry                 2                  2         100.00%
3        Kobe Bryant                 6                  8          75.00%
4      Allen Iverson                 2                  3          66.67%
5       Kyrie Irving                 2                  3          66.67%
6      Dirk Nowitzki                 2                  4          50.00%
7         Steve Nash                 2                  4          50.00%
8        Karl Malone                 2                  4          50.00%
9   Richard Hamilton                 2                  6          33.33%
10       Paul Pierce                 2                  7          28.57%
11      LeBron James                 3                 11          27.27%
12       Tony Parker                 2                  8          25.00%
13       Dwyane Wade                 2                  8          25.00%
14        Tim Duncan                 2                  9          22.22%
```

Figure 44

A couple of notes: We subtracted players with a single record in either category; Even if their performance were admirable, this test was supposed to assess consistency. Also, the percentage calculates how many times a player appeared in the top 3 category divided by how many times it appeared in the filtered clutch player datasets. So of course all of the players in the tables actually played many more seasons, they just did not participate in enough close games to qualify for our datasets.

6.4 Trends And Inflation

Another property we scrutinized based on our knowledge and the literature is how these ratings were affected throughout the years. Our formula was hard-grounded in original traditional basketball statistics, so trends or changes in the game (especially considering the extensive duration we are studying) could affect the statistical objectivity. To explore this issue we averaged the EoCC ranking based on the respective year (e.g. average 1997 ranking would encompass all the player ratings from both the regular season and the playoffs of the year 1997). Figure 45 graphically illustrates the results:

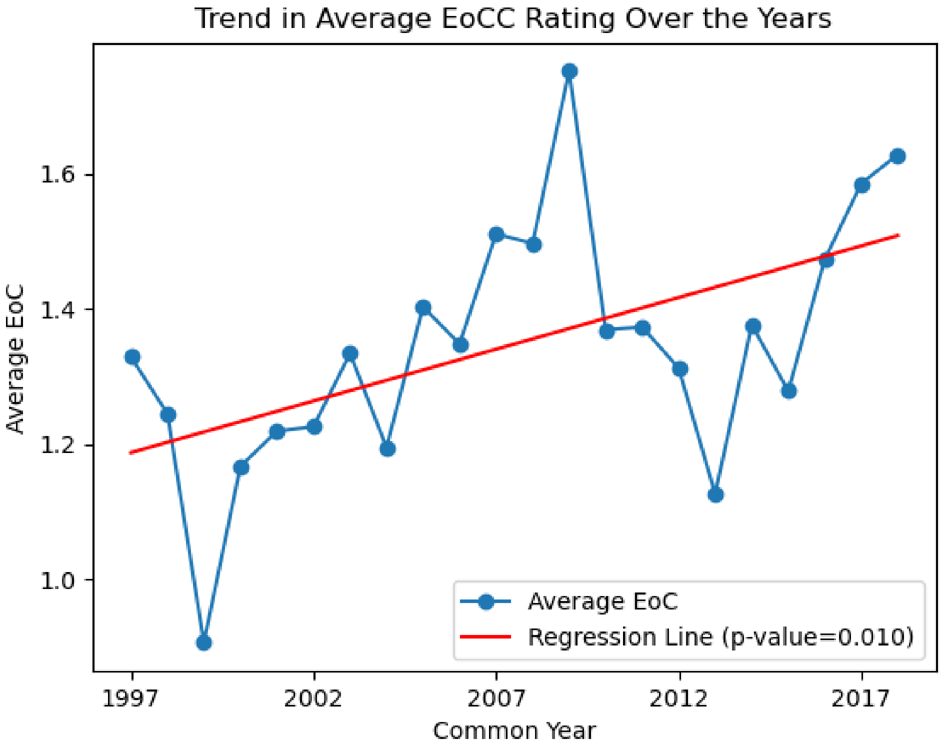
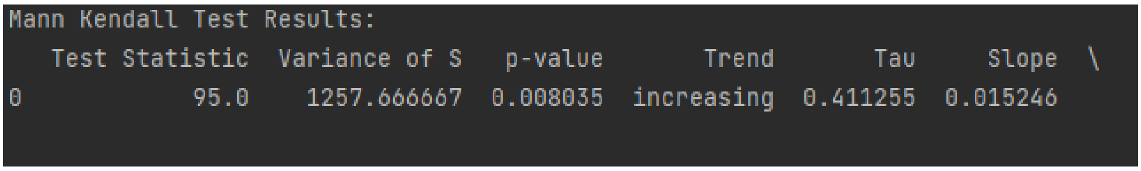


Figure 45

We can detect a slight ascending trend in the data, which means that the average ratings of the players tend to increase throughout the years.

Nevertheless, to statistically examine this hypothesis, we applied the Mann-Kendall Trend Test. The Mann-Kendall test is a non-parametric test that evaluates the existence of a monotonic trend in a time series.

In the following picture (figure 46), we have the Mann-Kendall Test results:



```
Mann Kendall Test Results:
   Test Statistic  Variance of S   p-value       Trend       Tau     Slope  \
0            95.0    1257.666667  0.008035  increasing  0.411255  0.015246
```

Figure 46

Evidently, there is significant (p<0.05) increasing trend in our data. Tau is also an indication parameter of a strong monotonic trend in the data.

In our context, this means that the statistics of the players are increasing during the period we are examining. This could be dependent on a variety of factors: The statistical patterns may change over time due to shifts in playing styles and strategies (e.g. modern 3-point focus), the coaching mentality and the team dynamics rarely remain constant, and even the phenomenon (often referred to as “stat-padding) that players actively pursue certain statistics to enhance their individual records. Although, interesting by nature, we do not have the data and we are not going to speculate about the causes of this pattern in this paper.

6.5 Observations

Clearly, the concept of a “lone superstar” or primary option in crunch-time situations is evident in many successful teams, where a star player shoulders the responsibility of scoring or playmaking when the game is on the line. However, our study also raises an interesting point about the correlation between individual performances and team championship success. While having a clutch performer is undoubtedly an asset, winning championships requires a combination of factors, including team chemistry, depth coaching strategies, and overall team performance. In some cases, a team’s success in clutch situations may not directly translate to championships due to various factors, such as the competitiveness of the league, the strength of opposing teams, or specific matchups. Additionally, playoff success often involves multiple players stepping up at different points in a series, not just in clutch moments.

7 Discussion / Future Work

7.1 Discussion

In this paper we investigated the performance of NBA players during the last critical moments of close games. The term 'clutch' appeared a lot throughout our literature review and our own work and not without a reason. As we thoroughly analyzed in section 2 of this paper, 'clutchness' has not yet been clearly defined within the sports universe. Some researchers claim it to be the ability to perform better than normal under special circumstances in critical moments. Others consider it a spontaneous exemplary performance doubting its reproducibility. We explored the corresponding final minutes' stats from the NBA and compared the players to rank the best of the best in this category. The focus of this study was to distinguish the most important factors that contribute towards better winning chances during the last moments of the games. We defined our own clutch performance metric (EoCC) and used it to compare the performances of different players under similar circumstances. Something that was argued in the literature but was out of the scope of this experiment, is comparing the players' clutch statistics with their own average numbers during the rest of the game. Firstly, this has been done in previous work discussed in section 2 of this article, and secondly, from our own intuition and basketball knowledge, aggregating and averaging the statistics of the last critical plays and comparing them with the averages of the rest 45 minutes would create certain loopholes. Teams tend to behave differently when the game (or even a championship!) is on the line. Some coaches can even have special plays for these moments [55] that have never revealed just to catch the other teams off-guard. On that account, our research totally ignored the rest of the player stats and we only worked on the last 3 minutes period.

Our results provided interesting insights about the player efficiency during the last minutes and how this affects the probabilities of a positive outcome. For instance, we found out that the best clutch players were mid-range specialists, while the big-men appeared sparsely. That is not out of common sense: the perimeter players will firstly initialize the offence and then try to pass the ball inside. That being said, we found high correlation between blocks, offensive rebounds and winning percentages, which means that the centers and power forwards of the teams, still contribute towards the victory conditions, although their role is somewhat outshined, since it is usually the perimeter players that dictate the offence during these moments.

Another puzzling observation of our study was the relation between top clutch performance and seasonal team success. We found out that although the best personal stats were achieved by players in teams who relied on a single superstar for their clutch scoring, the better team results and the higher winning percentages were obtained by teams who had more than one go-to clutch players. For example, while Lebron’s best individual performances were regarded with the team of Cleveland in the years 2007-2009, his most successful runs were with Miami and the co-superstar Dwayne Wade, where they also won 2 national championships. So in general while individual performance may lead to victories, it cannot guarantee overall team success.

7.2 Future Work

The main problem we encountered when trying to materialize our ideas was the lack of specific data. Thankfully we managed to find some basic basketball player statistics and filter them according to our purposes but there could be so much more done if the required data was publicly available. Due to the nature of these statistics computational power is not an issue. There are not complex or ambiguous attributes; the difficulty lies in accessing the particular data values. Thanks to the SportsVU equipment, advanced information in the basketball court can be captured in real-time and even processed later but it is not freely available for the public or the student data scientist. Nevertheless, we would like to share some ideas for expanding our work in the future, in case we are given the opportunity.

A proposition that we explored but eventually did not manage to implement, was to find a way to introduce weights into our performance formula based on the timing of the in court statistics. A shot during the last 20 seconds of the game has probably more significance from a shot 2 minutes earlier and so on and so forth. So it was intriguing to inspect ways to realize this notion into a mathematical term.

Moreover, a query that occurred during our research was to determine whether mid-range game is still prevalent during late game situations when the need for secure decisions may discourage the teams from taking the currently popular but risky 3-point shot. Our work indicated that the best clutch players were experts of mid-range game so a following up study could investigate further.

Also, there is the presumed term of “stat-padding”: Since almost all performance ratings are based on certain stats, there is all-around discussion about players trying to artificially increase their numbers (forced assists, free rebounds etc.). In our study we encountered clear increase in the numbers throughout the years, although there could be a variety of explanations except stat-padding, and this assumption has to be examined in detail.

One interesting topic that we would like very much to investigate is to clearly differentiate and compare team plays vs isolation plays in the last critical moments. A study like this could also provide hints to the importance of having strong clutch performers in the team; the results may direct coaches towards having a more balanced roster that can collectively deal with critical plays than having a superstar that is bound to get the ball during the last seconds. Unfortunately, advanced statistics are required as opposed to using only the assist as a team effort. An assist is counted only when the play is successful, but there is no way to count unsuccessful team plays, for instance. Still, this kind of information is not too difficult to obtain with the right equipment, and it definitely makes a nice research question.

Another idea is the individual plus-minus, especially between certain couples of offensive-defensive players. As we have already discussed [12], there is data to work on this project, so it would be very intriguing to explore whether some players can force others to a bad decision that may decide the game.

Moreover, a discussion among basketball enthusiasts is whether the ball should go to the player that is having the best performance in that particular game, even if he is not the first option in general. Although not backed by any data, this is an argument that instinctively makes sense; if a player has a strong game overall, why would not he be able to perform during the final stages?

All in all, there is a lot of space that can be explored in the particular domain. From psychological factors to pure basketball statistics the researchers have not even yet decided that clutch performance is an ability that can be honed and not an incidental event. Still, with the complex data that can be obtained with the new systems, and given the required resources, eager sports analysts can definitely assist in answering this kind of questions, giving important information to the coaches and teams and expanding the knowledge and understanding of the basketball universe.

References

[1] Kubatko, Justin, Oliver, Dean, Pelton, Kevin and Rosenbaum, Dan T. "A Starting Point for Analyzing Basketball Statistics" *Journal of Quantitative Analysis in Sports* 3, no. 3 (2007). https://doi.org/10.2202/1559-0410.1070

[2] Sun Ya Xue, "Research on the Influence of Big Data on School Physical Education Development", *Sports Frontier*, 2019

[3] P.T. Amarasena, B.T.G.S. Kumara, "Data Mining Approach for Identifying Suitable Sport for Beginners", 2019

[4] Tesla A. Monson, Marianne F. Brasil, Leslea J. Hlusko, "Allometric Variation in Modern Humans and the Relationship Between Body Proportions and Elite Athletic Success", 2018

[5] S. Dowlan, K Ball, "Applications for Cluster Analysis in sport Biomechanics", Victoria University, 2007

[6] Z. Shelly, R. F. Burch, W. Tian, L. Strawderman, A. Piroli, C. Bichey , "Using K-means Clustering to Create Training Groups for Elite American Football student-athletes based on Game Demands", Mississippi State University, 2020

[7] K. Apostolou, C. Tjortjis, "Sports Analytics algorithms for performance prediction", IHU, 2018

[8] G. George, Z. Panagiotis, "Statistical analysis of men's fivb beach volleyball team performance", International Journal of Performance Analysis in Sport 8 (2008), pp. 31-43

[9] Gabrio A, "Bayesian Hierarchical Models for the Prediction of Volleyball Results", University College London, 2019

[10] Sarlis V, Tjortjis C, "Sports analytics - Evaluation of basketball players and team performance", Information Systems, Volume 93, (2020) 101562, ISSN 0306-4379, https://doi.org/10.1016/j.is.2020.101562

[11] J. Sill, "Improved NBA adjusted +/- using regularization and out-of-sample testing", MIT Sloan Sports Analytics Conference, 2010, pp. 1-7

[12] A. Franks, A. Miller, L. Bornn, K. Goldsberry, "Characterizing the spatial structure of defensive skill in professional basketball," The Annals of Applied Statistics, Ann. Appl. Stat. 9(1), 94-121, (2015)

[13] J. Hewko, R. Sullivan, M. El-Hajj, S. Reige, “Data Mining in the NBA: An applied approach”, 2019

[14] K. Goldsberry, E. Weiss, “The Dwight Effect: A New Ensemble of Interior Defense Analytics for the NBA”, MIT Sloan Sports Analytics Conference, 2013

[15] A. Franks, A. Miller, L. Bornin, K. Goldsberry, “Characterizing the spatial structure of defensive skill in professional basketball, 2015

[16] Terner Z., A. Franks, “Modeling Player and Team Performance in Basketball”, Annual Review of Statistics and Its Application, 1-23 (2021)

[17] Lorenzo A, Gómez MÁ, Ortega E, Ibáñez SJ, Sampaio J. “Game related statistics which discriminate between winning and losing under-16 male basketball games”, J Sports Sci Med. (2010)

[18] D. Miljkovic, L. Gajic, A. Kovacevic, Z. Konjovic, “The Use of Data Mining for Basketball Matches Outcomes Prediction”, Novi Sad, Republic of Serbia, 2010

[19] Lan Yu, “Association Rules Based Data Mining on Test Data of Physical Health Standard” (2009)

[20] J. L. Deng “Introduction to Grey system theory” J. Grey Syst. 1, 1 (1989), 1–24

[21] S. A. Javed, A. Gunasekaran, A. Mahmoudi, “DGRA: Multi-sourcing and supplier classification through Dynamic Grey Relational Analysis method”, Computers & Industrial Engineering, Vol. 173 (2022), ISSN 0360-8352

[22] Kuo, Yiyo & Yang, Taho & Huang, Guan-Wei, “The use of Grey Relational Analysis in solving multiple attribute decision-making problems”, (2006), Computers & Industrial Engineering. 55. 80-93. 10.1016/j.cie.2007.12.002.

[23] Mahmoudi, A & Javed, S & Liu, S & Deng, X. , “Distinguishing Coefficient Driven Sensitivity Analysis of GRA Model for Intelligent Decisions: Application in Project Management”, (2020), Technological and Economic Development of Economy. 26. 621–641

[24] C. Ker-Chang & H. Li-Fei & L. Hsin-Yi & L. Shu-Chen & K. Ju-Chun, “Fight evaluation of NBA teams — application of grey relational analysis”, (2006), Journal of Information & Optimization Sciences

[25] Pradhan, S. “Ranking regular seasons in the NBA’s Modern Era using grey relational analysis”, (2017), Journal of Sports Analytics. 4. 1-33

[26] Draper, N. R., Smith, H. “Applied regression analysis”, (2014), (Vol. 326). John Wiley & Sons

[27] Chang H., He W, Chiang Y., “Predicting the NBA Winning Percentage Base on the Linear Regression Model”, (2020), 34th Conference of Japanese Society for AI

[28] Yang Y., “Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Basketball Statistics”, (2015), University of California

[29] "Analysis of NBA Player Performance, Popularity and Salary." ukdiss.com, (2018), Business Bliss Consultants FZE

[30] S. W. Smith, "Neural Networks (and more!)," in The Scientist and Engineer's Guide to

Digital Signal Processing, California Technical Publishing, 1997

[31] Ghosh, A., Sufian, A., Sultana, F., Chakrabarti, A., De, D. “Fundamental Concepts of Convolutional Neural Network”, (2020), Balas, V., Kumar, R., Srivastava, R. (eds) Recent Trends and Advances in Artificial Intelligence and Internet of Things. Intelligent Systems Reference Library, vol 172. Springer, Cham. https://doi.org/10.1007/978-3-030-32644-9_36

[32] Castelvecchi, D. "Can we open the black box of AI?" Nature, vol. 538, no. 7623, 6 Oct. 2016

[33] Chou, P & Chien, S & Yang, T & Yeh, Y & Chou, W & Yeh, C., “Predicting Active NBA Players Most Likely to Be Inducted into the Basketball Hall of Famers Using Artificial Neural Networks in Microsoft Excel: Development and Usability Study”, (2021), International Journal of Environmental Research and Public Health

[34] Wang, Kuan-Chieh Jackson and R. S. Zemel. “Classifying NBA Offensive Plays Using Neural Networks.” (2016)

[35] Hu J, Zhang H, Qiu J, "Prediction of MVP Attribution in NBA Regular Match Based on BP Neural Network Model", (2022), National Institute of Technology Karnataka

[36] Shah R & Romijnders R, “Applying Deep Learning to Basketball Trajectories”, (2016)

[37] Bai Z & Bai X, “Sports Big Data: Management, Analysis, Applications, and Challenges”, Complexity, (2021)

[38] Novatchkov H. & Bichler S. & Tampier M. & Kornfeind P. & Baca A., “Real-Time Data Acquisition and Performance Analysis in Sports”, (2011), 76-80

[39] Fan J. & Han F. & Liu H., “Challenges of Big Data Analysis”, (2013), National Science Review

[40] C. H. Liu, Q. Lin and S. Wen, "Blockchain-Enabled Data Collection and Sharing for Industrial IoT With Deep Reinforcement Learning," IEEE Transactions on Industrial Informatics, vol. 15, no. 6, pp. 3516-3526

[41] Kabir, Syed Muhammad, “METHODS OF DATA COLLECTION”, (2016)

[42] Ajayi V, “A Review on Primary Sources of Data and Secondary Sources of Data”, (2023)

[43] Scannapieco M. & Missier P. & Batini C. , “Data Quality at a Glance”, (2005), Datenbank-Spektrum. 14. 6-14.

[44] Basketball-reference.com, [Online]. Available: https://www.basketball-reference.com/

[45] NBA.com, [Online], Available: https://stats.nba.com

[46] Van Rossum, G., & Drake, F. L., “Python 3 Reference Manual”, Scotts Valley, CA: CreateSpace

[47] Khder M., “Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application”, International Journal of Advances in Soft Computing and its Applications. 13. 145-168. 10.15849/IJASCA.211128.11

[48] Krotov, Vlad & Silva, Leiser, “Legality and Ethics of Web Scraping”, (2018)

[49] Heimes R., “Data scraping and the implications of the latest LinkedIN-hiQ court ruling”, (September 2019), iapp.org, International Association of Privacy Professionals

[50] Vast R. & Young R & Thomas P, “Emotions in sport: Perceived effects on attention, concentration, and performance”, (2010), Australian Psychologist

[51] Otten M. & Barrett M. “Pitching and clutch hitting in Major League Baseball: What 109 years of statistics reveal”, (2013), Psychology of Sport and Exercise. 14. 531–537

[52] Seifried C. and Papatheodorou M., "The Concepts of Clutch and Choking: Recommendations for Improving Performance under Pressure", Journal of Coaching Education 3, 1 (2010): 90-98

[53] Sarioz C., “The ‘Clutch’ gene myth: An analysis of Late Game Shooting Performance in the NBA”, (2021), University of California, Berkeley

[54] Schweickle, M., Swann, C., Jackman, P. & Vella, S, “Clutch performance in sport and exercise: a systematic review”, (2020), International Review of Sport and Exercise Psychology

[55] Jackson P. and Hugh D., “Eleven Rings: The Soul of Success”, (2013), Penguin Press

[56] Swann C, Crust L, Vella SA, “New directions in the psychology of optimal performance in sport: flow and clutch states”, Curr Opin Psychol. 2017 Aug; 16:48-53

[57] Berri D. & Eschker E, “Performance When It Counts - The Myth of the Prime Time Performer in Professional Basketball”, (2005), Journal of Economic Issues 39

[58] Teramoto M. and Cross C., "Relative Importance of Performance Factors in Winning NBA Games in Regular Season versus Playoffs", (2010), Journal of Quantitative Analysis in Sports 6, no. 3

[59] Scott W., Steven B. C. & Franklin G. Mixon Jr., “*Homo certus* in professional basketball? Empirical evidence from the 2011 NBA Playoffs”, (2013), Applied Economics Letters, 20:7, 642-648, DOI: 10.1080/13504851.2012.727965

[60] Solomonov, Y. & Avugos S. & Bar-Eli M., “Do Clutch Players Win the Game - Testing the Validity of the Clutch Player's Reputation in Basketball”, (2014), Psychology of Sport and Exercise

[61] Cao Z. & Price J & Stone D., “Performance Under Pressure in the NBA”, (2010), Journal of Sports Economics. 12. 231-252

[62] Goldman, M. and Justin M. Rao. “Effort vs. Concentration: The Asymmetric Impact of Pressure on NBA Performance 1.” (2012)

[63] Barkazian A., “Clutchness in Basketball”, California State Polytechnic University, Pomona, 2019

[64] Chakrabarti, S., Ester, M., Fayyad, U., Gehrke, J., Han, J., Morishita, S., Piatetsky-Shapiro, G. and Wang, W., “Data mining curriculum: A proposal” (Version 1.0), Intensive Working Group of ACM SIGKDD Curriculum Committee, 140, pp.1-10

[65] Han J., Kamber M. and Pei J., "Data mining concepts and techniques, third edition.", (2012)

[66] Gkenios G., “Processing EEG signals using Deep Learning methods for medical diagnosis”, 2022

[67] Rahm E. & Do H., “Data Cleaning: Problems and Current Approaches”, (2000), IEEE Data Eng. Bull. 23. 3-13.

[68] Sarlis V, Papageorgiou G, Tjortjis C, “Sports Analytics and Text Mining NBA Data to Assess Recovery from Injuries and Their Economic Impact”, Computers 12(12), 261

[69] Sarlis V, Chatziilias V, Tjortjis C, Mandalidis D, “A data science approach analyzing the impact of injuries on basketball player and team performance”, Information Systems 99, 101750

[70] Eppel Y, Kaspi M, Painsky A., “Decision Making for Basketball clutch shots: A data driven approach, Journal of Sports Analytics (2023)