INTERNATIONAL
HELLENIC
UNIVERSITY

Sports Analytics Algorithm for NBA Champion Prediction

Kaimakamis Christos

SID: 3308190009

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc)

JANUARY 2012

THESSALONIKI – GREECE

INTERNATIONAL
HELLENIC
UNIVERSITY

Sports Analytics Algorithm for NBA Champion Prediction

Kaimakamis Christos

SID: 3308190009

Supervisor:	Prof. Christos Tjortjis
Supervising Committee Members :	Dimitrios Karapiperis Ioannis Magnisalis

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc)

JANUARY 2012

THESSALONIKI – GREECE

Abstract

Machine learning is a rapidly grown term which has application in various fields of our life one of them is in sports analytics.

In this paper we made use of data, which were extracted from 19 seasons of NBA games. The goal of the thesis is to exploit the data we had trying to measure every team's game performance and predict their final position after NBA Playoffs.

Extracted data concerns not only the most fundamental team statistical categories but also some miscellaneous features regarding team performance. In this thesis firstly we attempt to depict the development of the NBA industry within the years along with the change of the game's nature itself. Furthermore, with the analytics tools we can extract valuable information regarding which are the key factors that can lead a team in the top of NBA championship.

We conducted several experiments using a variety of classifiers aiming to predict the number of wins of every team participating in Playoffs. Despite of the new format at playoffs results were accurate and very interesting assumptions were made.

Kaimakamis Christos

04 - 01 - 2020

Acknowledgments

At this point I would like to thank my supervisor professor Christos Tjortjis for his valuable advices during this process, he was always eager to assist, aiming to solve any issue which occurred. Moreover, I would like to thank my family and friends for their support during those difficult and stressful months, being there supporting me with every mean they had

Kaimakamis Christos

04 - 01 - 2020

Contents

List of Figures

List of Tables

1 Chapter Introduction

1 Introduction

Basketball is one of the two most popular team sports, influencing billions of people globally. It is generally accepted that NBA is the best basketball league around the world, that is why we decided to base our project on this league.30 teams participate in NBA including teams from USA and Canada, it also divides into two Conferences East and West. With the heavily growing impact in commercial and influence level, NBA commissioner established a very tight match schedule that means if we include summer league, preseason, friendly matches, post-season and playoff matches, we are dealing with the immense number of 200 matches per season for a team [1]. Therefore, it goes without saying that this league is highly demanding and with too many matches and too many important factors like injuries, suspensions, it is too difficult to predict the champion every year.

In this thesis we aim to predict the champion for every season based on game-level statistics in various categories using machine learning classifiers. A significant amount of people already tried to predict in the most optimal way the results of basketball matches using a variety of systems and techniques based on data collected from people opinions, but the main obstacle was that most of the people as being already fans were extremely affected by their feeling and cannot think straight in order to make a prediction [2]. Thus, the data collected were not optimal for the prediction.

The sport analytics and performance prediction field affects more and more people daily. Many teams already have hired a data analyst in order to gather all the data extracted from matches and training sessions. If all this gathered info preprocessed correctly it will benefit the team and the coaching staff in a variety of fields. Another application of sports analytics is for betting purposes. The betting industry is growing rapidly subsequently ,the need of gambling is also increased among people. In order to set the betting odds, correct and avoid a loss, betting companies relied on data analyzed and offer the most advantageous odds for them [5].

For all the above reasons we attempted to predict the outcome of 2020 NBA playoffs, using machine learning models and applying various correlation techniques aiming to achieve most accurate predictions regarding the final NBA standings. Although this NBA season was different than the others, we managed in our thesis, to made long-term predictions with a high accuracy score. Against the odds from betting companies we predicted the final standings on the Playoff tree using supervised training models that can be used for both regression and classification problems, with high accuracy.

Since we observed the impact of sport analytics, we decided to use machine learning algorithms to achieve optimal prediction score. The major elements to success are to build the most appropriate machine learning algorithm and accumulate a very descriptive dataset suitable for our project. In this research we have transformed the prediction problem to a classification problem which is based on data in many different categories, per season in every NBA team [1].

2 Chapter Background

2.1 Related Work

As we stated above sport analytics is that not only grows rapidly but also has application in a variety of different kind of sports. Before the term sport analytics was known, coaches, players, fans, people who make a living out of sports and sport journalists acknowledged the beneficial results of collecting and analyzing data extracted from matches and training sessions. In order to predict the outcome of the match we needed to determine which statistical categories need improvement aiming to assist them in achieving their goal, to win the championship and improve their team.

2.1.1 Tennis

Clarke and Dyte were the first researchers who tried to fit a logistic regression algorithm in order to depict the variation in ATP ratings of the two players with the intention to foresee the result of the match. Moreover, Sipko stated that the result in the previous matches between two players affect the prediction of the new match. Last but no least Scavincky was the first to use ELO rating as an technique to foresee the results of a Tennis match [3].

2.1.2 Cricket

Cricket is considered an international sport which gained more attention than football regarding studies and literature. Clarke, based on the papers that Reep, Pollard, Elderton and Benjamin wrote, discovered that the rate of occurrence of very small and very big scores were far less than the geometric distribution implied. Furthermore, Clarke combined all the essential variables in order to predict the outcome of the game, taking into consideration the condition of the field, the time restrictions, personal milestones or scores from past series are able to affect the outcome of the match [4].

2.1.3 Baseball

Baseball is a very popular in the USA, consequently baseball have a vast application into sport analytics. B. James was the first to pinpoint the need of new methods in order to analyze the data extracted from this sport and used the term Sabermetrics. Sabermetrics is the empirical analysis of Baseball, especially Baseball statistics that measure in-game activity. Established on this term James invented another statistical measurement called “Runs Created” to calculate the number of the runs a hitter provides to his team hence he optimized and gain a major advantage by extracting those data benefiting the team to win the championship [5].

2.1.4 Soccer

Soccer is the “king of sports” for many people around the globe, but due to its complexity it’s not easy to extract data and analyze them afterwards. The primal steps of soccer analytics occurred when Reep and Benjamin published a statistical analysis including various patterns of each team play style extracting data from 3300 matches in a timespan of 15 seasons [6]. Those results originated the first attempt of soccer analytics, because they conclude that most of the goals are scored when you have a “possession play style” and also that in order to score a goal you have to make 8 attempts. The impact of this research was huge even some of the best British coaches have adopted and learned from those studies, many of them hire an assistant with the intention to have him occupied by the task of searching and extracting data based on Reep’s research [7]. Many authors in next years aimed to predict the outcome of a football game, V.Chazan predicted with high accuracy the final standings for Spanish La Liga (year 2018-19) and predicted the teams are finished to Champions League and Europa league qualification spot with 64% and 75% accuracy[5].

2.1.5 Volleyball

In volleyball we do not have a considerable history regarding data analysis. Although in 2017 Tumer and Kocer predicted the final league positions with a success rate over 90% to achieve that they used a multi-layer perceptron. Moreover, aiming to enhance the training procedure they used lazy algorithms like k-NN and extracted data from jumping drills (in-game and post-game activity), analyzing those data, helped them to optimize the ferocity of the jumps, indicating that in blocks an athlete should try to jump more intensely. Additional to Tumer and Kocer, Wang Zao got the analytics techniques one step beyond installing sensors into players “good” hand with the intention of taking measurements of the spike power and how the motion of the hand affects the power of the spike. Exploiting those data categorized player into mediocre, bad, good, superstar with a high level of accuracy. Finally, Van Haaren et al. (2016) invented a technique which can recognize and categorize different in-game patterns and compare them afterwards for the purpose of choosing the most optimal and beneficial offensive and defensive play for each team [8].

2.1.6 NFL

American Football is hard and difficult sport but has many fans around the globe. Of course, it has a high level of application from the perspective of sport analytics. Many people tried to predict the outcome of the NFL championship, Warner and Shau for example predicted NFL champions with a percentage near 70% using data extracted from last seasons and analyzing those data with random forest, support vector machines and neural networks [9].

2.2 Literature Review

Basketball could not be unaffected from the sport analytics era, as a matter of fact basketball has a huge application of sport analytics because due to its complexity anyone can extract many different data which can be exploited in various fields. Teams are counting on those extracted data to change their defensive or offensive strategies or even sometimes to see which players should be offered a new contract or not [10].

In 1977 Stefani R.T. proposed a method of foreseeing the outcome of NCAA basketball games, the method used to gather all the necessary data was ordinary least squares. It is worth mentioning that Stefani 3 years later published a new research with an enhanced least squares method aiming to predict the basketball games result [11]. T. Zak publicized a paper called “Production Efficiency: The Case of Professional Basketball”, in this research Zak concluded that categories like shooting percentage , Assists , steals and rebounding had a huge impact on the games result. Combing those categories Zak was able to classify the participating teams into 2 groups: Defensive and Offensive teams. In 1984 Shanahan collecting data from IOWA University for both men and women from 1981-1983 found out that field goal, rebounds, steals, blocks and turnovers are the categories that matter most for predicting the games’ result for women. For men now the most crucial statistics are fouls, field goals, rebounds and forced errors. To conclude to these results Shanahan used a logistic regression model that predicted with

60% accuracy for women and 80% for men [12]. Stern in 1994 tried to foresee the final result of a basketball game if a team was leading by x points, but the Brownian motion model was not able to work smoothly due to a major issue regarding the points per second distribution due to variation of every team performance Stern could not come up with solid results [13].

The first groundbreaking and innovative research was published by Berri a sports economist from Utah in 1999 trying to depict how much impact in the game's outcome has an individual player's performance. In his study he collected data from 4 seasons 1994-98 he found that every position has different statistical categories that value, but he did not answer properly the question "can an individual performance change the outcome of the game?". This question was hard to answer since many different aspects and variables can change during the game like playing time, injury, different coach decisions [14]. Hu and Zidek used data from 1997/1998 season in order to predict the playoff winner. They used weighted likelihood technique by using in the equation some categories which have bigger impact in the outcome of the game than the others, to achieve their goal [15]. It is worth mentioning that Melnick (2001) in his paper combined data from 5 NBA seasons aiming to prove the correlation between assists per team and win/lose outcome. Also proved that the total numbers of assists per team values more than the assist from the five starter players, concluding that the way a team scores their baskets is very crucial and affects to either winning or losing the game [16]. Kvam and Sokol in 2004 with their research paper "A logistic regression/Markov chain model for NCAA basketball" attempted to foresee the winner of the NBA pre-season. In their paper they created first a Markov chain model that applied to one team per state, the differential part than past research is that they used only data extracted from each team score and took into account the "home" and "away" variable (Figure 1.), not any individual or team associated data [17]. Shirley in 2007 tried to apply the same method but used in-game data and was not able to calculate properly the variable of transition so his method was not optimal [18]. Trawinski (2010) used 10 fuzzy classification algorithms also included major voting in his research intending to guess the result of a match [19]. Strumbelj and Vracar used a Markov model with in-game statistics as a direct value and team statistics as an indirect input [20].

Cao in 2012, managed to achieve 69% accuracy score. He collected data from 5 NBA seasons and used logistic regression model achieving 69% accuracy score in forecasting game’s outcome [21]. Magel and Unruch (2013) achieved 68% accuracy using logistic regression and 64% by using least squares. They utilized different data for offense and different for defense aiming to determine if one team will win or lose the game [22]. DeLong et al. (2013) using data extracted from in-game statistics proved that team cohesion is a strong factor defining winning or losing. Additionally, according to Delong’s paper another significant point that designates the game’s outcome is the miss matches (the times that in defense or offense a team have a player shorter than the player who is guarding him) inside the game [23]. Omidiran (2013) choose as input 10 post-game statistical categories (Figure 2.) He applied several classification techniques in order, to foresee which team will get more wins in the regular season. Moreover, he cross-checked all his findings with all betting odds from Las Vegas betting companies and found out that in some times he could even guess with more accuracy than the bookers [24]. Another method of logistic regression algorithm was implemented from Lopez, M. J. and G. J. Matthews (2015) who got as an input one NCAA team and they guessed the probabilities of winning or losing the game. Their experiments were made after taking into consideration betting handicaps(for example to win the selected team with +5.5 points) they found that luck is a strong factor deciding the outcome of either winning or losing with the handicap [25]. In addition to previous researchers Jones (2016) [10] used logistic regression to estimate the probability of winning a game. He tried 3 different methods to make the prediction. First, he used the model of point spread for all playoff games, secondly he took point differential average between two rival teams and finally the weighted average was calculated including also the 0.30 times the average point differential from method 2. The last method had the best results on playoff predictions. Finally V. Sarlis (2020) predicted correctly the MVP of this NBA championship using machine learning techniques using the not only team statistics but also individual as an input [38].

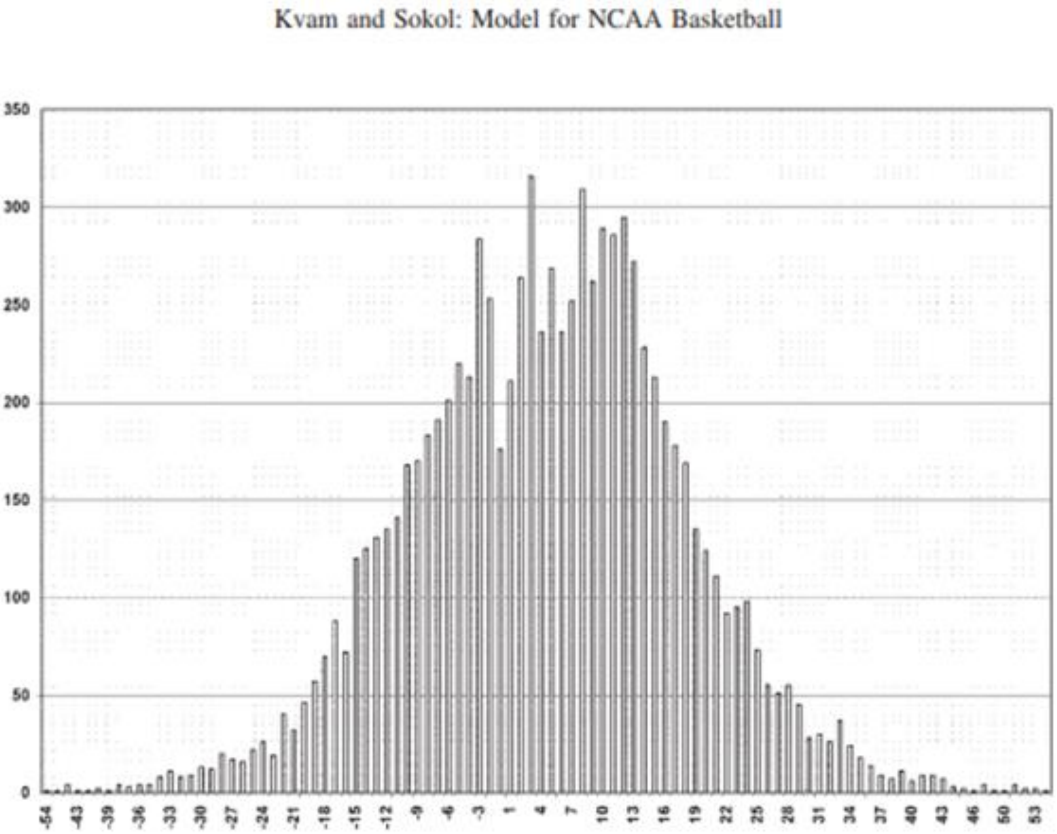
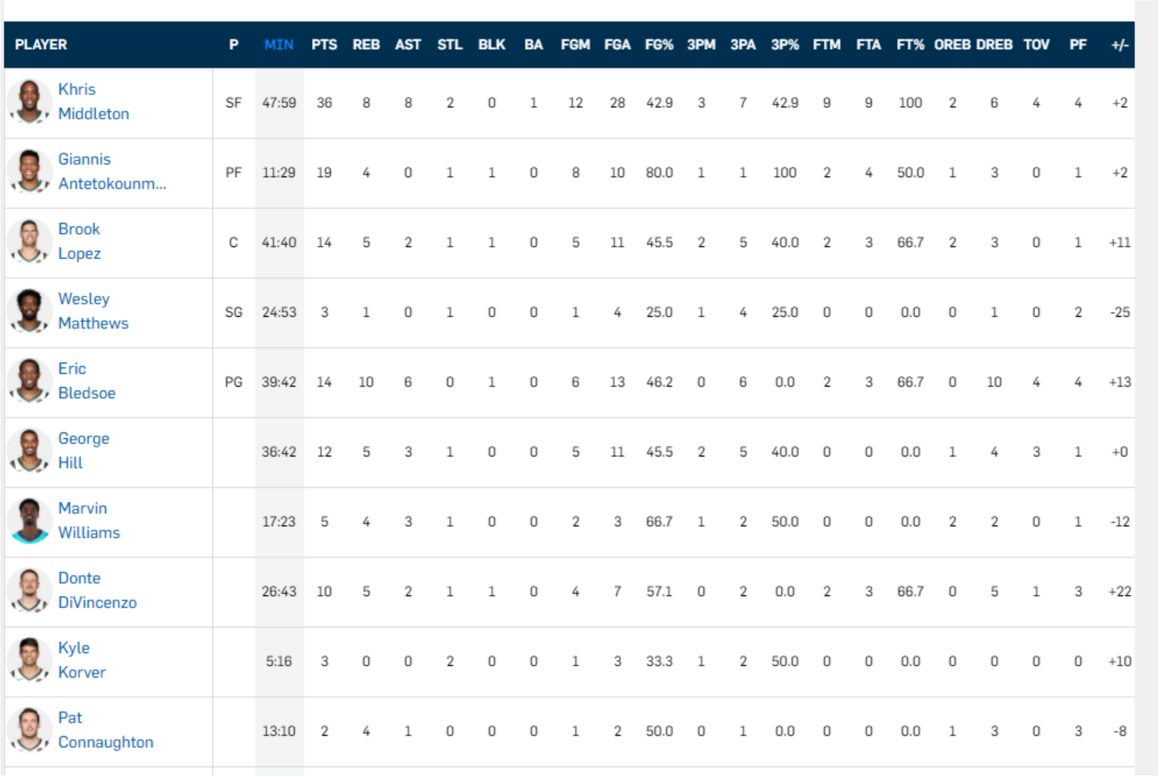


Figure 1. Number of Home and Away game by home team victory margin



PLAYER	P	MIN	PTS	REB	AST	STL	BLK	BA	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	TOV	PF	+/-
Khris Middleton	SF	47:59	36	8	8	2	0	1	12	28	42.9	3	7	42.9	9	9	100	2	6	4	4	+2
Giannis Antetokounm...	PF	11:29	19	4	0	1	1	0	8	10	80.0	1	1	100	2	4	50.0	1	3	0	1	+2
Brook Lopez	C	41:40	14	5	2	1	1	0	5	11	45.5	2	5	40.0	2	3	66.7	2	3	0	1	+11
Wesley Matthews	SG	24:53	3	1	0	1	0	0	1	4	25.0	1	4	25.0	0	0	0.0	0	1	0	2	-25
Eric Bledsoe	PG	39:42	14	10	6	0	1	0	6	13	46.2	0	6	0.0	2	3	66.7	0	10	4	4	+13
George Hill		36:42	12	5	3	1	0	0	5	11	45.5	2	5	40.0	0	0	0.0	1	4	3	1	+0
Marvin Williams		17:23	5	4	3	1	0	0	2	3	66.7	1	2	50.0	0	0	0.0	2	2	0	1	-12
Donte DiVincenzo		26:43	10	5	2	1	1	0	4	7	57.1	0	2	0.0	2	3	66.7	0	5	1	3	+22
Kyle Korver		5:16	3	0	0	2	0	0	1	3	33.3	1	2	50.0	0	0	0.0	0	0	0	0	+10
Pat Connaughton		13:10	2	4	1	0	0	0	1	2	50.0	0	1	0.0	0	0	0.0	1	3	0	3	-8

Figure2. Box score for post-game statistics

3 Chapter Goals and Techniques

3.1 Expected outcome

In this paper our object is to predict the winner of NBA championship season 2019/20. In the NBA format the first 8 teams from each division (East and West) complete each other into a new set of games called "The Playoffs" (figure 3). The winner gets the "ring" and wins the Championship. In order to win the first place in the championship, the winner team must do at least 16 wins. Our collected data took as input the number of wins from every champion between the 2000/18, in order to have more solid results we took the number of 16 wins. Our goal was very difficult because for the first time of NBA franchise history all teams must travel to Orlando, for the purpose of playing all the games in a neutral stadium. Due to the corona virus outbreak, regular season was terminated one month earlier than normal and all teams which were qualified into playoff spots travel to Orlando. In Orlando was created a “bubble” which cost the amount of 190 million. Every team which was invited had a unique training facility and place to stay and, in the matches, neither crowd nor journalists were prohibited. As you can see this year is a completely new process that no-one can predict how it will end, from sport journalists to betting companies and finally to simple sport fans. That incident highly motivated us to attempt find the winner of NBA, after we take into consideration all the new variables and conditions for this year.

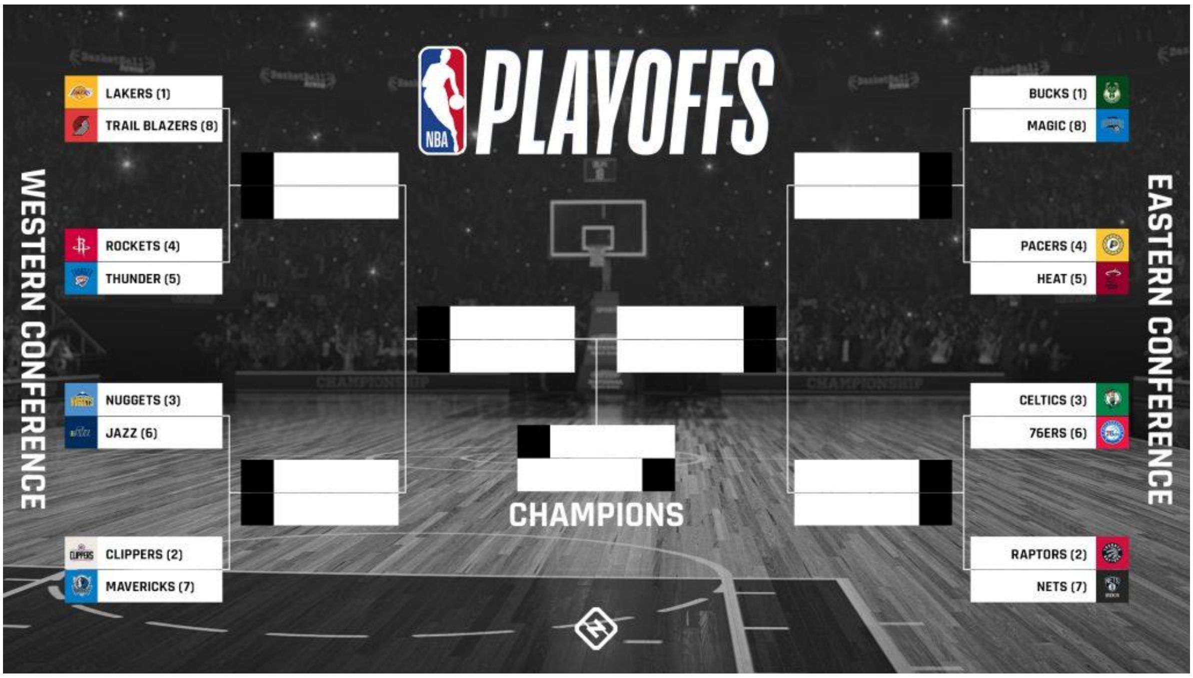


Figure3. Playoff Bracket (2019/20)

3.2 Terms and Definitions

3.2.1 Hyper-parameter tuning

Hyper-parameter optimization is a crucial part of data science projects (Figure1).The first part as we see in the graph below (Figure 4) is to collect the essential data the second step is to choose the most suitable classification algorithm and finally the last part is to find the most optimal parameters. Every classifier has various parameters which if we do not attempt to optimize them the algorithm will choose the default ones, negatively affecting our algorithm prediction [26]. Consequently, hyper-parameter tuning is an important factor determining if our project will be successful or not.

3.2.2 Grid Search

It is worth mentioning that choosing the parameter on our own it is not the optimal solution, since every classification problem needs different variables. We implemented Grid Search which is a tuning technique that can calculate the most optimal combination of parameters, aiming to improve our prediction score. The only drawback on this technique is the high computation time needed to have our results [26]. Most of the times we use four or more nested loops in order to find the most suitable combination, that procedure needed enough time and made our code debugging difficult since we needed to wait for example 20 minutes – 1 hour per classifier (figure 5).

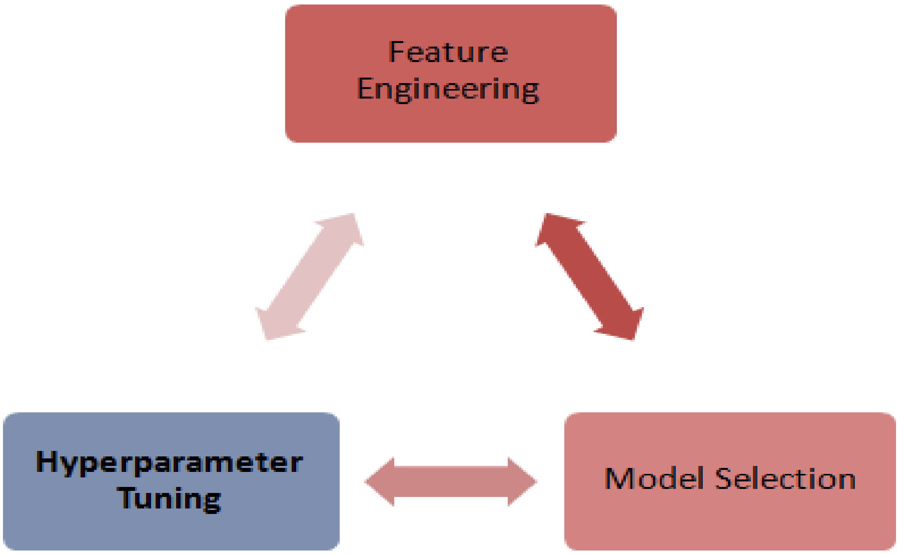


Figure 4. 3 steps of classification

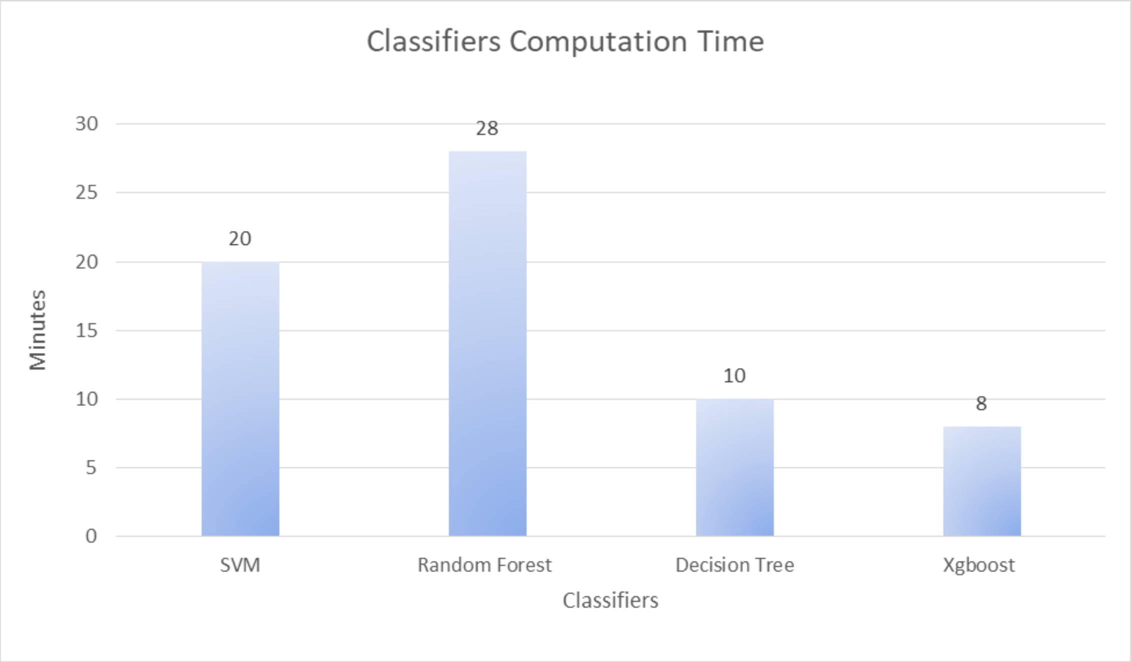


Figure 5. Computation time of Grid search

3.2.3 Cross-Validation

After we trained our model, we had to be certain that our model is accurate, with this propose we proceeded to our classifier validation. Cross validation is a method which is implemented by us to measure the efficiency of our machine learning algorithms. Furthermore, it also can be used as re-sampling technique to assist to a further model evaluation [27].

3.2.4 K- Fold cross-validation

This procedure is used, as we stated before, to validate and re-sample our model. This technique has only one parameter (k) which needs to be configured. This parameter indicates how much iteration will take place. We usually choose 5 to 10 that depend on data-set's size (Figure 6). The higher value of k increases the risk of over-fitting. We kept the score from every iteration and finally when we completed all iterations, we took the average of our scores [28]. In our project we used 5-fold Cross Validation because we had a small dataset.

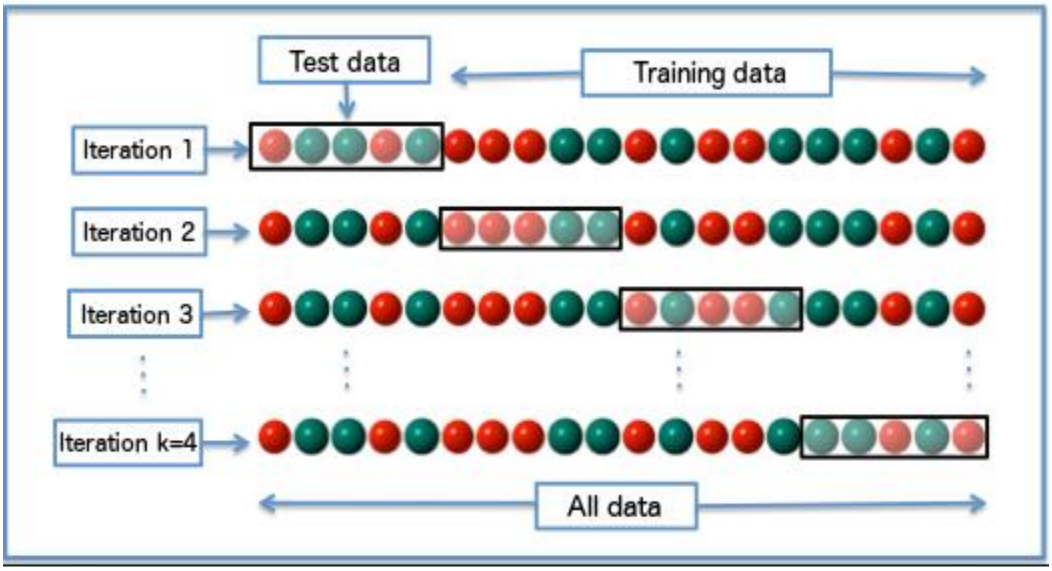


Figure 6.K folds cross validation

3.2.5 Classifiers and Techniques

3.3.5.1 Random Forest

Random forest is supervised learning algorithm. The forest creates an ensemble multiple decision trees, this method is suitable for both regression and classification use this fact is helpful because it fits on the most of machine learning problems. Furthermore, this classifier searches for the most optimal feature between a variety of features that's provides us with much better results. As we mentioned above random forest generates an ensemble of trees, each tree depends on an attribute sample feature. In classification problems the class with the most votes is chosen and in regression projects we choose the average scores of all trees (figure 7). Additionally, another significant function of random forest is that we can measure the importance of its features, with the method called feature importance. This procedure is measuring the score of each feature in every tree node after training and scales the findings so if you sum all the variables the importance will be equal to 1 [29].

3.3.5.2 Support Vector Machines

Support Vector Machines is a simple algorithm which, like Random Forest, can be used for linear and classification problems. The linear classifier that we used is a SVM model and called SVC (Support vector classifier). Support vector machine (SVM) is a formulation of pattern recognition problems that has many advantages over other approaches. An SVM finds the best isolating (maximal margin) hyper-plane between the two classes of training sets in the feature space. A linear function has the following form: f(x) = + c which relates to a hyper-plane diving the feature space. If, for given pattern mapped in the feature space to x, the value of f(x) is a positive number then the pattern belongs to a class labeled by the value 1, otherwise, it belongs to the class with the value -1 [30]. In our classifier we try to calculate the margin, which is the distance between support vectors and the line. In any classification project there are a variety of hyper-planes which can be used in order to separate the values. Our goal is to find the solution which gives us the maximum margin [31].

SVM has 3 basic parameters that needed to be tuned: kernel, C, gamma. Firstly, kernel is mapping our key points into a space, in order to get separated ideally from the line. Choosing the correct kernel value (linear, polynomial, radial) is affecting the way that our features will be separated [32]. In our project (Table 2) linear kernel was the most optimal way to separate our classes. The C parameter indicates the amount of miss-classification between our features. In our paper C value was 0.1 after tuning, meaning that our classifier wanted a larger margin separating hyper-plane. Finally, gamma is informing us how much curve we need on our decision boundary, for a high gamma value we had high curve for a small we have low curve.

Parameters	**Pool of Values**	**Selected Values after tuning**
Gamma	0.1,1,10,100	0.1
C	0.1,1,10,100	0.1
Kernel	linear , rbf, polynomial	linear

Table 2.SVM Hyper-parameter tuning

3.3.5.3 Extreme Gradient Booster

XGboost is an ensemble of gradient boosted decision trees which are built to provide speed and high efficiency levels. The model’s outcome is a very accurate prediction score after the combination of the weak learners. The boosting method generates a number of models that their purpose is to correct the mistakes of the models which are before of them in the created sequence. Therefore, the first created model is the training one the second generated model attempts to fix the first and the third attempts to correct the second and so on, the exact number of iterations needed can be tuned via hyper-parameter tuning (figure 8). Moreover, XGboost has, like random forest, feature importance technique which as we mentioned before can be very helpful for our effort to improve our models results [33].

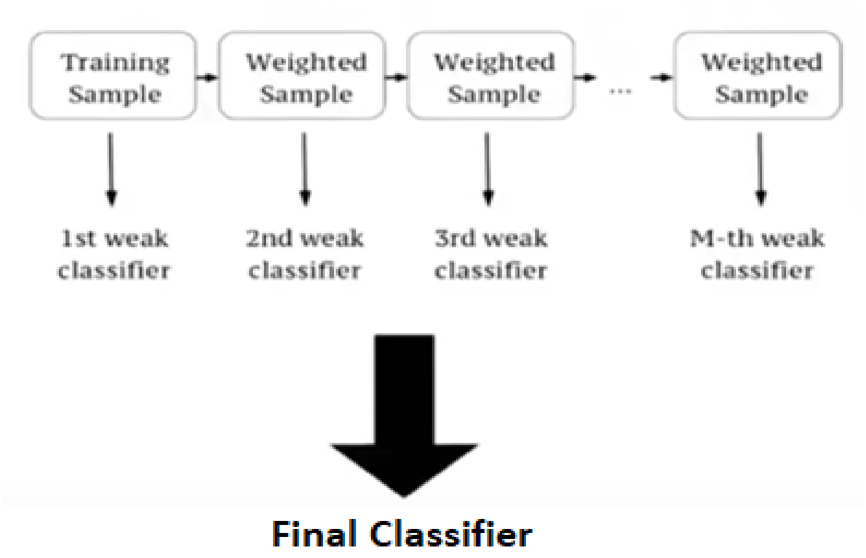


Figure 8. How XGboost works

This classifier has multiple parameters; thus, the hyper-parameter tuning is the trickiest section of our project. As you see in the table below (Table 3) we optimized our parameters using Grid Search. We have some parameters which are already being discussed by us before (gamma, max depth) and some new ones like min_child_weight , subsample, colsample_bytree, learning rate. Min_child_weight is defining the sum of the weights of all observations, it also manages over-fitting, high values may result over-fit so in our project we used cross validation, in order to tune it. Subsample we found that 0.8 was our optimal value meaning that we will prevent our model from over-fitting [34].

Parameters	**Pool of Values**	**Selected Values after tuning**
max_depth	4, 5, 6	4
min_child weight	4, 5, 6	4
gamma	0, 1, 2, 3, 4, 5	0
subsample	0.1 , 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8	0.8
colsample_bytree	0.1 , 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8	0.8
Learning rate	0.1 , 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 , 0.9, 1	0.1

Table 3. XGboost Hyper-Parameter tuning

3.3.5.4 Decision Tree

Decision Tree is a supervised learning method which can be used for Classification and Regression projects. There are two nodes the Decision node is used to take a decision, for that way we have various nodes. Finally, the second is being used as an output for Decision Node is called Leaf node. The process of decision tree, starts from root node by comparing the feature of the train dataset and, depending on the comparison, go to the next node. On the next node we repeat the process of comparing the values with the other node, until leaf node is reached [35].

As you see in the below table (Table 4) we have the results of our Hyper-parameter tuning after Grid Search. Decision Tree uses similar parameters like Random Forest and Xgboost, additionally we on max_features parameter after tuning process we saw that the most optimal value is auto [36].

Parameters	Pool of Values	Selected Values after tuning
max_features	auto', 'sqrt', 'log2'	auto
min_samples_split	2,3,4,5,6,7,8,9,10,11,12,13,14,15	4
min_samples_leaf	1,2,3,4,5,6,7,8,9,10,11	9
random_state	123	123

Table 4. Decision Tree Hyper-parameter

4 Chapter Features Extraction

4.1 Data Acquisition

The most essential part of our thesis was to collect the data, we tried to extract all the essential features which can provide us all the information needed to understand the game and proceed to our experimental phase.

The dataset we used found it online [37] depicts initially every team's league rank from season 2000-19, aiming to have a more accurate result teams are ranked from 1 to 30 (1 the best 30 the worst) based on their performance and wins percentage in 18 NBA seasons. Moreover this dataset includes multiple team stats categories not only the typical ones, like (assist, rebounds, 3p etc.) but also some more sophisticated like margin of victory, Pythagorean Loss and Win and some others, those features assisted us on extracting many results that guide us to a more solid and accurate prediction.

4.2 Columns and Contents analyze

In the below table (Table 5) you can see all the features we used in with the aim to implement our machine learning model and receive the more accurate results we can. We mostly used team statistical categories depicting the majority of stats which play significant factor on winning or losing a game. The most important feature is Playoff wins which is our target column on our model training and testing phase. Moreover, besides the usual categories like assists, points, field goal etc. we also have included the same categories but for opponent team statistics aiming to measure how much affects the others team performance on the outcome of the game.

Furthermore, the most interesting part of our dataset is the miscellaneous team categories that are not so common to us. Let us explain the terms and analyze them, so you can have a better understanding of our project procedure. We exploited data like Pythagorean Win and Loss which is the expected wins/losses based on how many points scored or allocated. Another feature is Margin of Victory which is the ability of each team to reach the projected wins in each season. A strong factor regarding team's performance is the Simple Rating System, this feature indicates a team rating that takes into account the average point differential and the strength of schedule (SOS), this rating denominated in points above/below average, while zero is the average value. Additionally, we took into consideration point allowed and scored per 100 possessions (Offensive and Defensive Ratings) for every team, with Pace as possession estimator per 48 minutes. Finally, we have the correlation between the Free Throws attempts with Field goals shows us how many fouls you get on your shot attempts.

Features Abbreviation	Description

Rk	Rank
Team	Team
Year	Year
Playoff Wins	Playoff Wins
MP	Minutes Played
FG	Field Goal
FGA%	Field Goal Attempts
3P	3 Point Field Goal
3PA	3 Point Field Goal Attempts
3P%	3 Point Field Goal Percentage
2P	2 Point Field Goal
2PA	2 Point Field Goal Attempts
2P%	2 Point Field Goal Percentage
FT	Free Throw
FTA	Free Throw Attempt
FT%	Free Throw Attempt Percentage
ORB	Offensive Rebounds
DRB	Defensive Rebounds
TRB	Total Rebounds
AST	Assists per game
STL	Steals per game
BLK	Blocks
TOV	Turnovers
PF	Personal Fouls
PTS	Points scored
O_MP	Opponent Minutes Played
O_FG	Opponent Field Goal
O_FGA	Opponent Field Goal Attempts
O_FG%	Opponent Field Goal Percentage
O_3P	Opponent 3 Point Field Goal
O_3PA	Opponent 3 Point Field Goal Attempts
O_3P%	Opponent 3 Point Field Goal Percentage
O_2P	Opponent 2 Point Field Goal
O_2PA	Opponent 2 Point Field Goal Attempts
O_2P%	Opponent 2 Point Field Goal Percentage
O_FT	Opponent Free Throw
O_FTA	Opponent Free Throw Attempt
O_FT%	Opponent Free Throw Attempt Percentage
O_ORB	Opponent Offensive Rebounds
O_DRB	Opponent Defensive Rebounds
O_TRB	Opponent Total Rebounds

O_AST	Opponent Assists per game
O_STL	Opponent Steals per game
O_BLK	Opponent Blocks
O_TOV	Opponent Turnovers
O_PF	Opponent Personal Fouls
O_PTS	Opponent Points scored
W	Wins
L	Losses
PW	Pythagorean Wins
PL	Pythagorean Losses
MOV	Margin Of Victory
SOS	Strength Of Schedule
SRS	Simple Rating System
ORtg	Offensive Rating
DRtg	Defensive Rating
Pace	Pace Factor
FTr	Free Throw Attempt Rate
3PAr	3 Point Attempt Rate
eFG%	Effective Field Goal Percentage
TOV%	Turnover Percentage
ORB%	Offensive Rebounds Percentage
FT/FGA	Free Throws per Field Goal Attempt
eFG%.1	Opponent Effective Field Goal Percentage
TOV%.1	Opponent Turnover Percentage
DRB%	Defensive Rebound Percentage
FT/FGA.1	Opponent Free Throws per Field Goal Attempt
Arena	Stadium
Attendance	Attendance

Table 5. Dataset Contents

4.3 Methodology

4.3.1 Data Preprocessing

Our initial action was to examine our dataset with the aim to find any inconsistencies or issues that could generate obstacles in implementing our machine learning techniques. Consequently, we checked for any missing or zero values on our dataset, but our dataset was clean in this sector since it was created using data scraping technique. Secondly, we proceed on dropping duplicate columns Rank and Opponent Minutes Played and we renamed columns relevant with opponent’s statistical categories, in order to be more convenient on reading and extracting results (Table 6). Moreover, the variables: “Arena”, “Attendance” were dropped since those categories will not play a role on the outcome of the playoffs because all playoff game will be placed in Orlando in neutral stadium, without crowd attending.

.

Features	Renamed
eFG%.1	O_eFG%
TOV%.1	O_TOV%
FT/FGA.1	O_FT/FGA'

Table 6. Renamed Features

4.3.2 Features Correlation

Aiming to find the parameters that affect the most on our result prediction we try to depict in our paper how different components interact with each other. Firstly, using a correlation matrix, we filtered the variables that will play a major part in our result. In the below graph you can see the correlation of our features with our target variable (Playoff Wins)

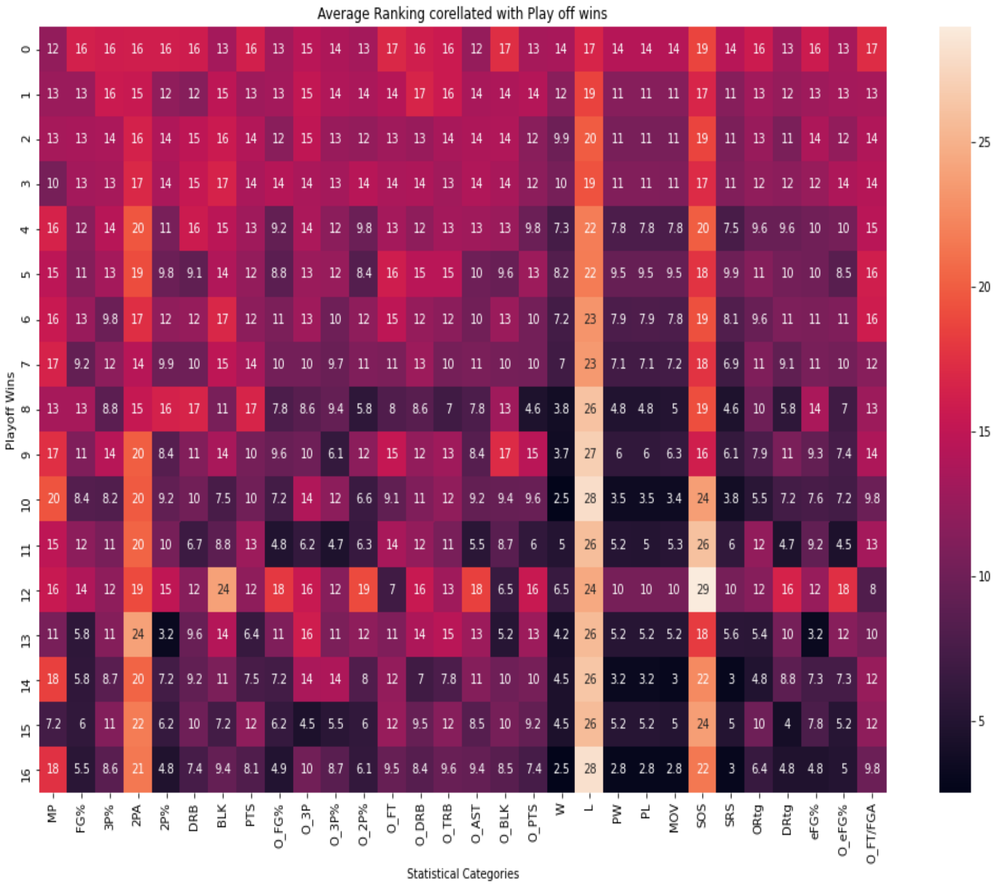


Figure 8. Average Ranking correlated with Playoff wins

Moreover, we made a correlation matrix between the filtered variable to how the variables affect each other. In the below graph (Figure 9) is depicted that “PL”, “PW”, “MOV” and “SRS” are highly correlated since they have the value 1 on the matrix. Although those variables are strongly related to each other we cannot drop them at once therefore we made a linear correlation plot to demonstrate better the correlation and assist us on making the correct decision to drop the not essential feature.

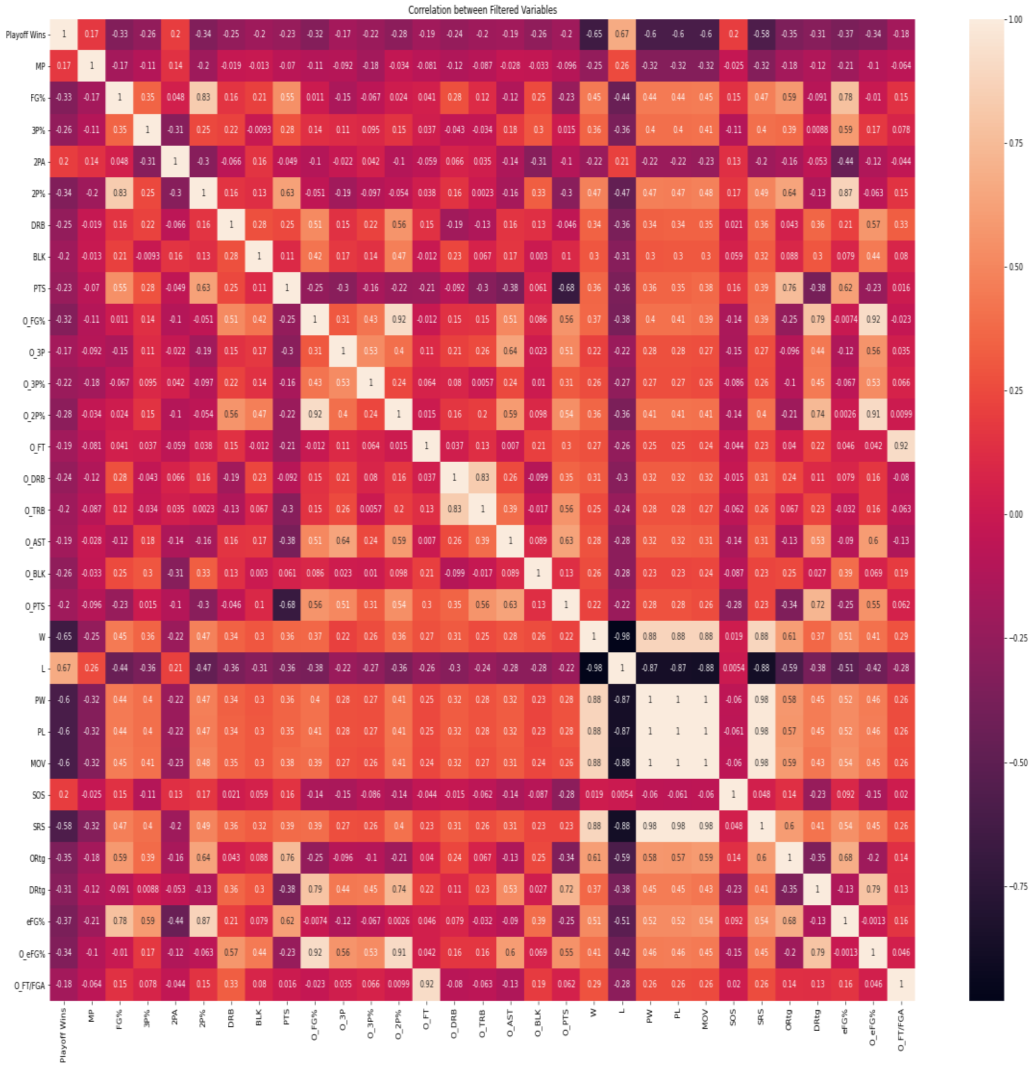


Figure 9. Filtered Variables Correlation Matrix

So in the below graph (Figure 10) we checked more specifically the connection between MOV ,SRS ,PL and PW and in the second graph (Figure 11) we depicted the link between the shooting stats (O_FG%, O_2P%, FG%, eFG%). We noticed that on the first graph (Figure 10) all variables are highly correlated since the lines are almost identical, we decided to drop MOV because SRS is a major factor that all GM's and coaches taking into consideration in weekly basis, not only in matches but at training drills as well and PW and PL which can be replaced from the variables W and L. In the second graph (Figure 11) we see that O_2P%, FG%, eFG% are strongly connected to each other we conclude on dropping that O_2P%, FG%, since eFg% can substitute both of them, on the contrary O_FG% it is not related at all meaning that we will keep it.

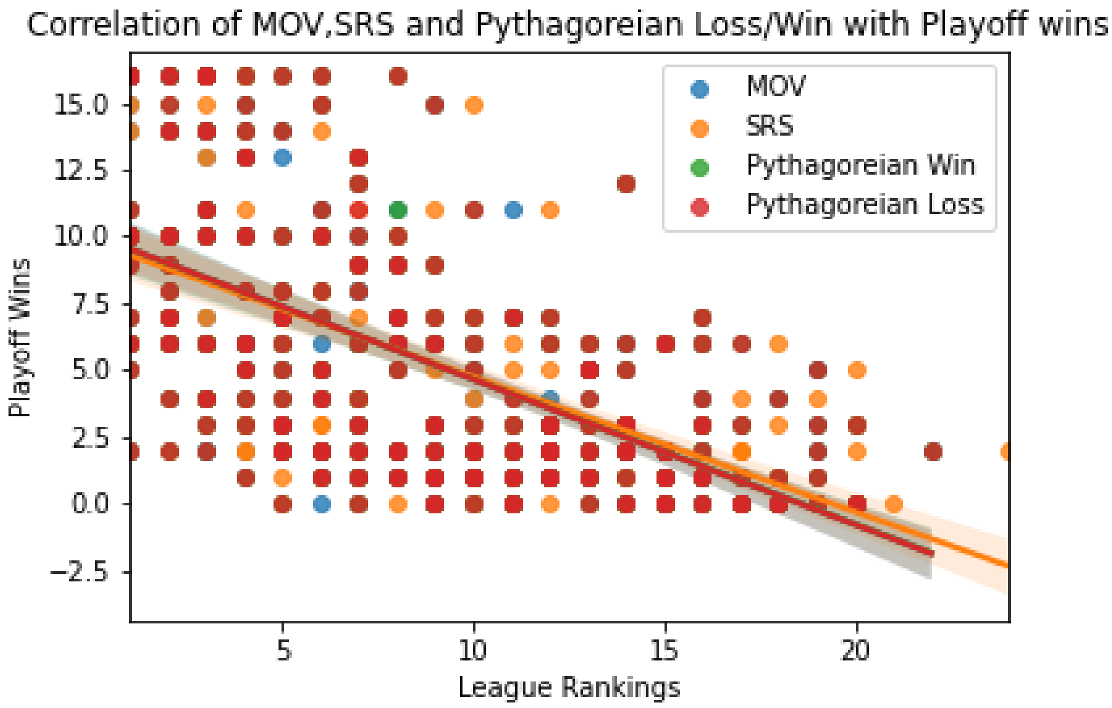


Figure 10. Correlation of MOV,SRS and Pythagorean Loss/Win with Playoff wins"

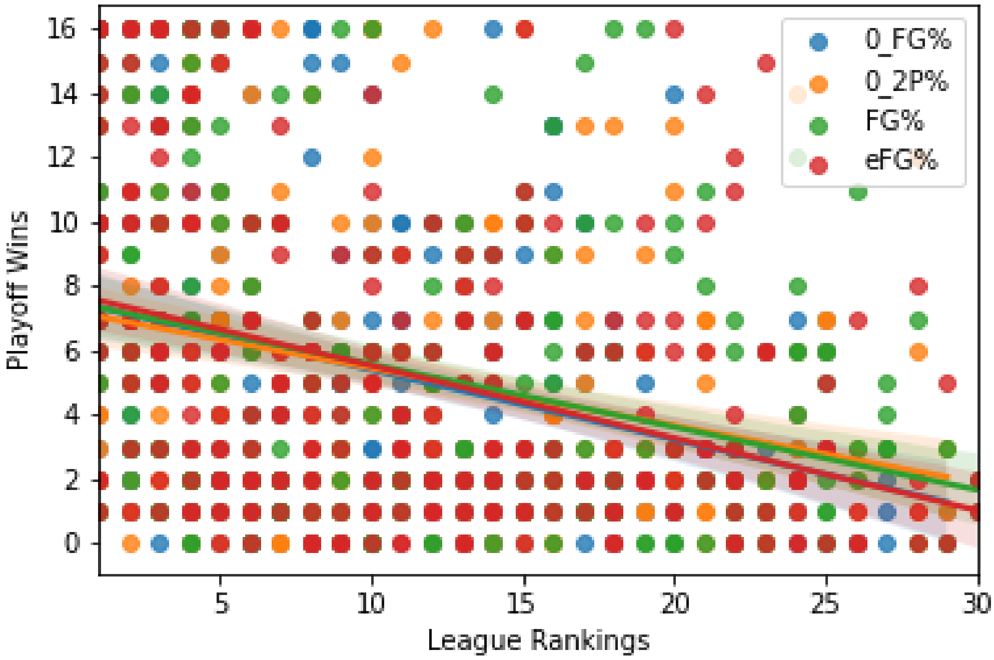


Figure 11. Shooting Stats Correlation

Another one result that is worth to be mentioned is the one is showed in the graph below (Figure 12). Due to my personal experience with Basketball I figured out that a strong factor on winning a game was to make shoots unguarded, so I decided to test my accusation by comparing Opponent Block percentage with Playoff wins. We see on the graph (Figure 12) teams with small Opponent Block percentage tend to have higher number of wins in Playoff, that is why all coaches in the last year try to put higher player in their frontline , so they can block or deflect any Field Goal attempts.

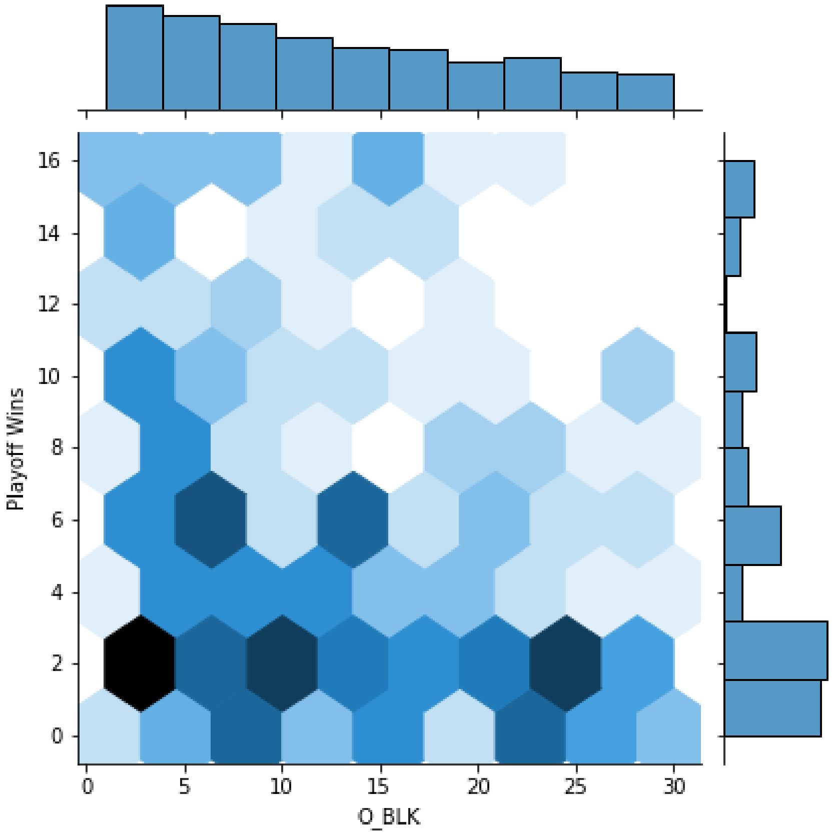


Figure 12. Opponent Block to Playoff wins

Basketball is a highly complicated sport that has a big variety of offensive and defensive strategies those are changing year by year along with the nature of the game. We tried to understand the change of the game’s nature within the 19 season, we took as our dataset, by comparing basic offensive and defensive statistical categories. Let see what results we can extract from the first graph (Figure 13). We observed that the in the early 00’s the most crucial factor for offensive stats was EFG% ,2P% , OR this outcome shows us that the game in the first seasons of 2000-2005 was played mostly from the "bigger" players, most of teams had in their roster the most talented frontline players of NBA franchise history (Garnett, Duncan, Nowitzki, Stoudamire, O'Neal, Ben Wallace), meaning that the coaches offensive strategy was to get the ball near to the basket aiming to get the majority of their offensive possession by their tall guys (Power Forward, Center positions). Therefore, we had a high percentage of 2P%, eFG and offensive rating indicator but a low percentage of 3P. On the season 2009-13 we observed a rapid grow of 3P percentage in the league, so the offensive plays were based on 3-point attempts, again the stats were correct because a team Golden State Warriors and their star player S. Curry (broke a record with 272 3points made within a regular season) along with his teammate K. Thopson they broke the record on 2012-13 they set the record for combined three-pointers in an NBA season with 484. This trend was continued until 2016 but the percentage is low because only warriors followed this revolutionized way of playing with their coach Steve Kerr, the rest of the teams tried to follow a bit more organized method of offensive possessions.

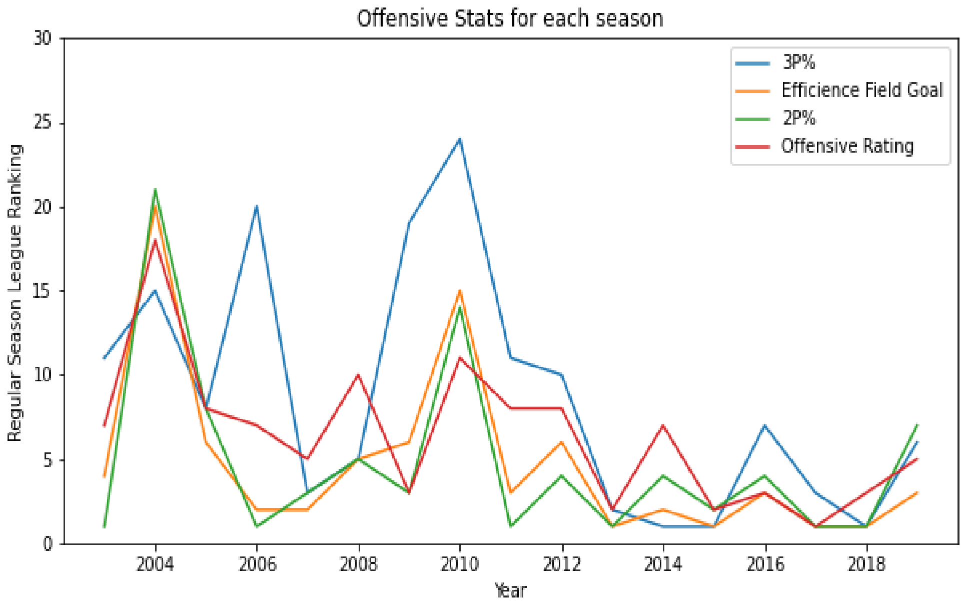


Figure 13. Offensive stats for each Season

Following, we implemented the same technique for the Defensive stats for each season. According to the below graph (Figure 14) the dominant category is Steal, this category was the most important for defensive plays the peak of it was in 2006, 2011 and 2016. Moreover, we can see that offense is far more important after 2012 defensive ratings are steadily reduced and offensive stats are increasing over the years. One explanation is that NBA commission applauds this strategy for marketing reasons, higher offensive ratings means higher scores, more dunks and highlights which help to increase the fans of the sports worldwide. It goes without saying that commission had nothing to do with this change of strategy, teams and coaches are responsible for that. As a matter of fact the start of defensive stats decline was at 2008's when Phoenix Suns with their coach Mike D'Antoni decided to go in 7-sec or less offensive possessions (meaning that they have to make their shoot attempt within the first 7 seconds), due to this fact going to a game with more possessions and faster executed attacks, the inside game was reduced and subsequently all shots were outside 6 meters. This strategy is depicted from our stats in 2008-10 watching the defensive rebounds stat to be increased and the opponent efficiency FG will be reduced. The highest raise on defensive rebound was from 2011-13, this strongly correlates with the lift on 3p percentage (figure me off stats) since taking many shots increases the miss shots per game and the rebounds are increased as well. Finally, Defensive Rating and Opponent Efficiency Field Goal are highly correlated, this is a gain completely normal considering, that if the defense is tighter the opponent will have most off the time a not so efficient offensive attempt (2-point or 3- point field goal) leading to a drop on the Efficiency Field Goal.

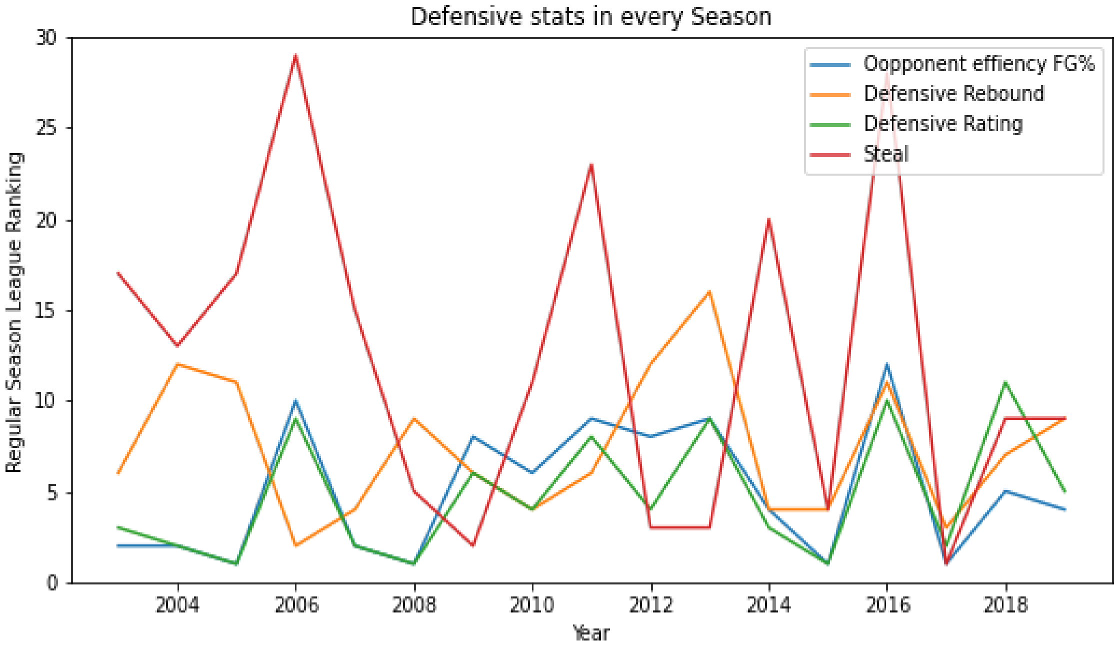
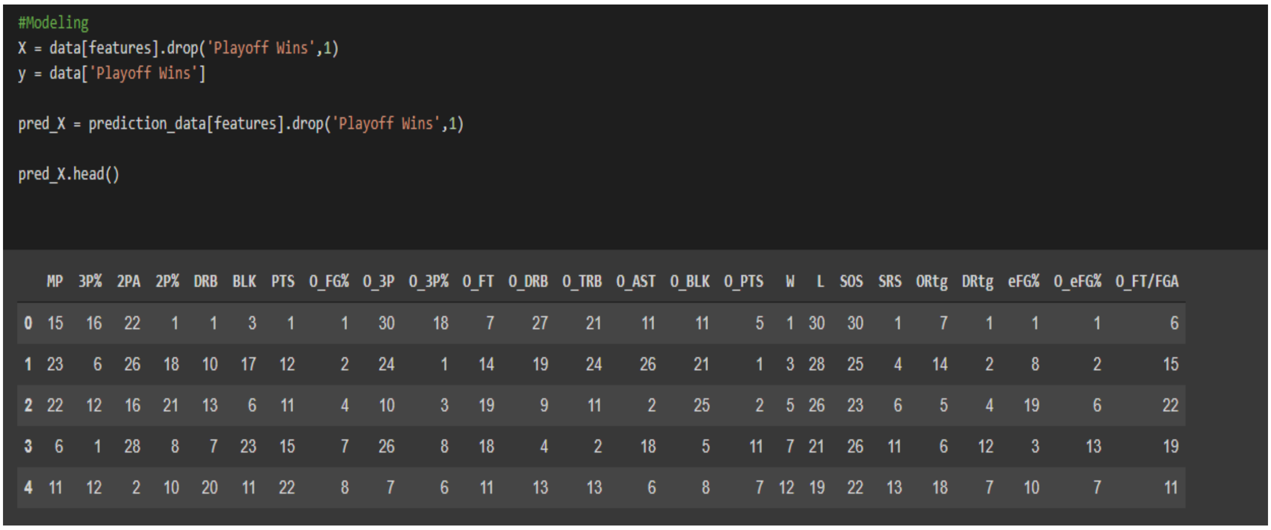


Figure 14. Defensive Stats for Each Season

4.3.3 Modeling

Initially we modeled our dataset (Figure 15) by dropping our desired feature (Playoff wins) from our test dataset. After that we proceed on the dataset split it was decided to make the split until season 2015. So, our train dataset has seasons from 2000-2015 and the test dataset has from 2015-2019.



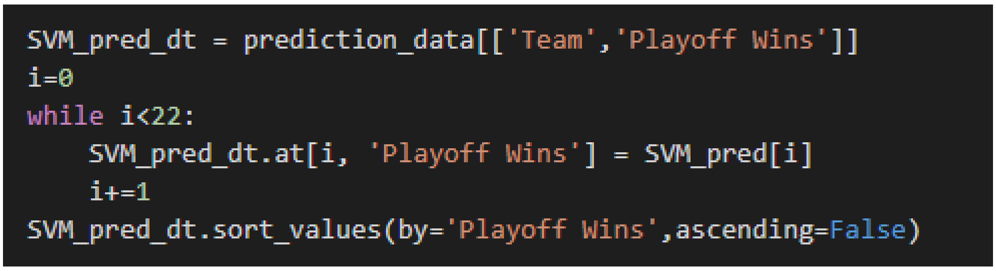
```
#Modeling
X = data[features].drop('Playoff Wins',1)
y = data['Playoff Wins']

pred_X = prediction_data[features].drop('Playoff Wins',1)

pred_X.head()
```

	MP	3P%	2PA	2P%	DRB	BLK	PTS	O_FG%	O_3P	O_3P%	O_FT	O_DRB	O_TRB	O_AST	O_BLK	O_PTS	W	L	SOS	SRS	ORtg	DRtg	eFG%	O_eFG%	O_FT/FGA
0	15	16	22	1	1	3	1	1	30	18	7	27	21	11	11	5	1	30	30	1	7	1	1	1	6
1	23	6	26	18	10	17	12	2	24	1	14	19	24	26	21	1	3	28	25	4	14	2	8	2	15
2	22	12	16	21	13	6	11	4	10	3	19	9	11	2	25	2	5	26	23	6	5	4	19	6	22
3	6	1	28	8	7	23	15	7	26	8	18	4	2	18	5	11	7	21	26	11	6	12	3	13	19
4	11	12	2	10	20	11	22	8	7	6	11	13	13	6	8	7	12	19	22	13	18	7	10	7	11

Figure 15. Modeling Code

4.3.4 Classifiers Training / Feature Importance

The classifiers we used were SVM, Decision Tree, Random Forrest and XGboost. All algorithms were trained with the optimal parameters since, as we already stated before, we used Grid Search in order to tune the parameters.



```
SVM_pred_dt = prediction_data[['Team','Playoff Wins']]
i=0
while i<22:
    SVM_pred_dt.at[i, 'Playoff Wins'] = SVM_pred[i]
    i+=1
SVM_pred_dt.sort_values(by='Playoff Wins',ascending=False)
```

Figure 16.Results Presentation

We found the Mean Absolute Error for each classifier and then we made our predictions in order to combine our predictions with the correct team and present them clearly, we used the below lines of code for each algorithm (figure 16). According to MSE value the most efficient algorithm in our first results was XGboost (Figure 17).

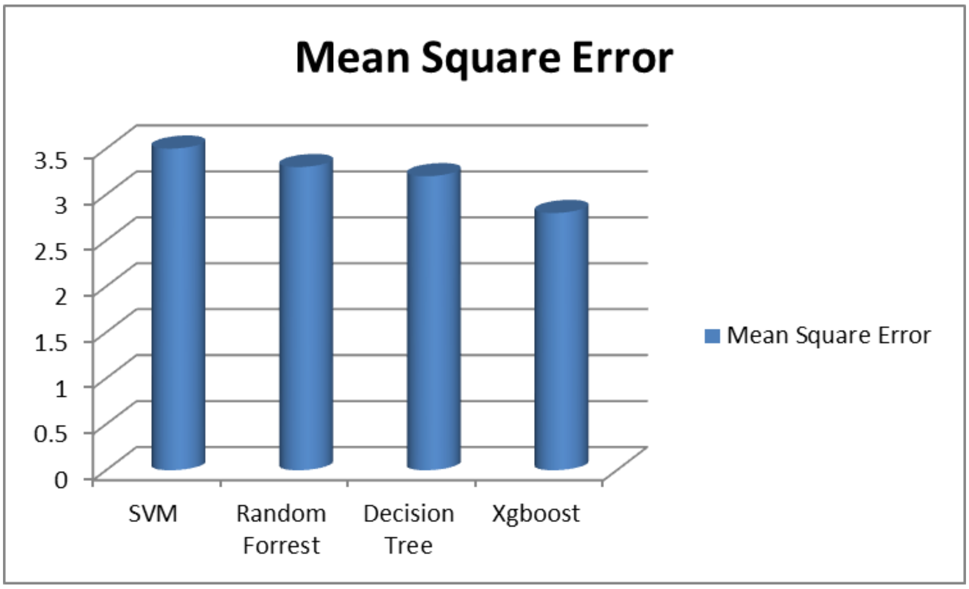


Figure 17.MSE per Classifier

Afterwards Feature Importance (Figure 18) was implemented on Random Forrest results intending to, discover which features value the most in our classifiers outcome. An interesting stat we observed is that the most important factor are the personal fouls for each player, the higher amount of fouls a player have the less efficient defense he can play , for this reason this category is so crucial for the game's outcome. Moreover, the second most valuable category is Points scored per game that is totally understandable and nobody can dispute that because the nature and the quintessence of the game is to score more points that your opponents. The rest of the significant features are also normal as we have Field Goal Percentage, OFGA and FGA. The less important categories are 3PA, Total Rebounds and Offensive Rebounds.

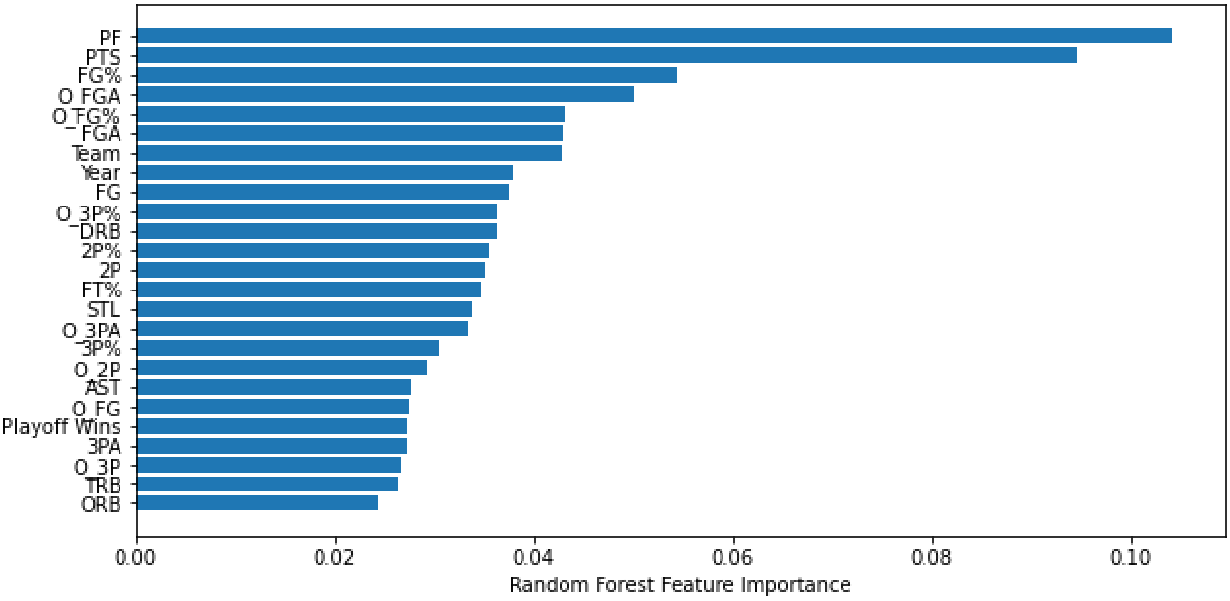
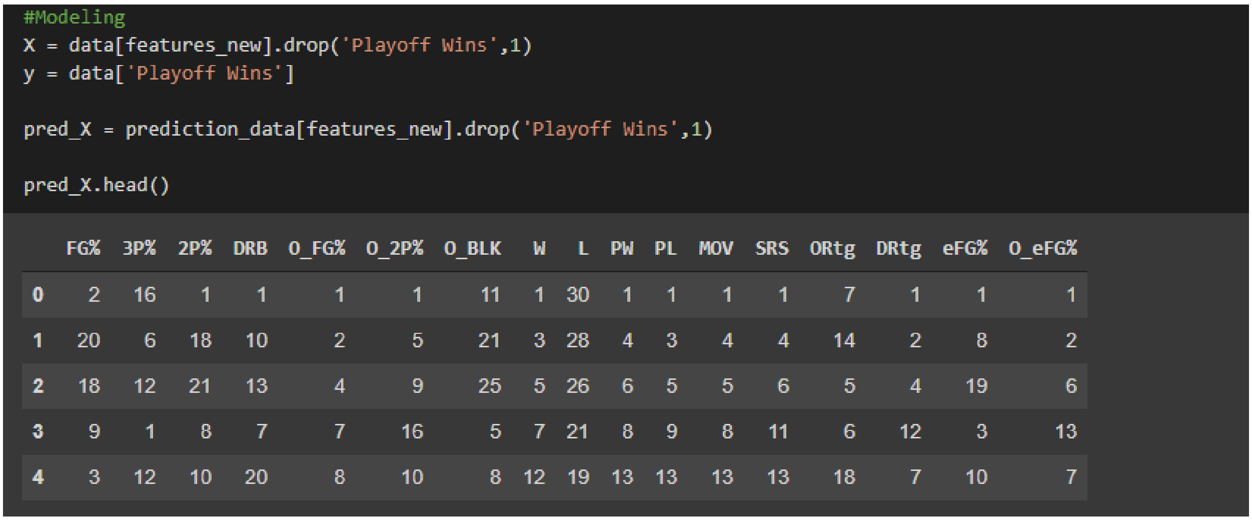


Figure 18. Feature importance

Following feature importance, we decided to drop again the not so significant categories and move forward to re-model our classifiers again keeping only the features with the most importance on our project's outcome (figure 19).



```
#Modeling
X = data[features_new].drop('Playoff Wins',1)
y = data['Playoff Wins']

pred_X = prediction_data[features_new].drop('Playoff Wins',1)

pred_X.head()
```

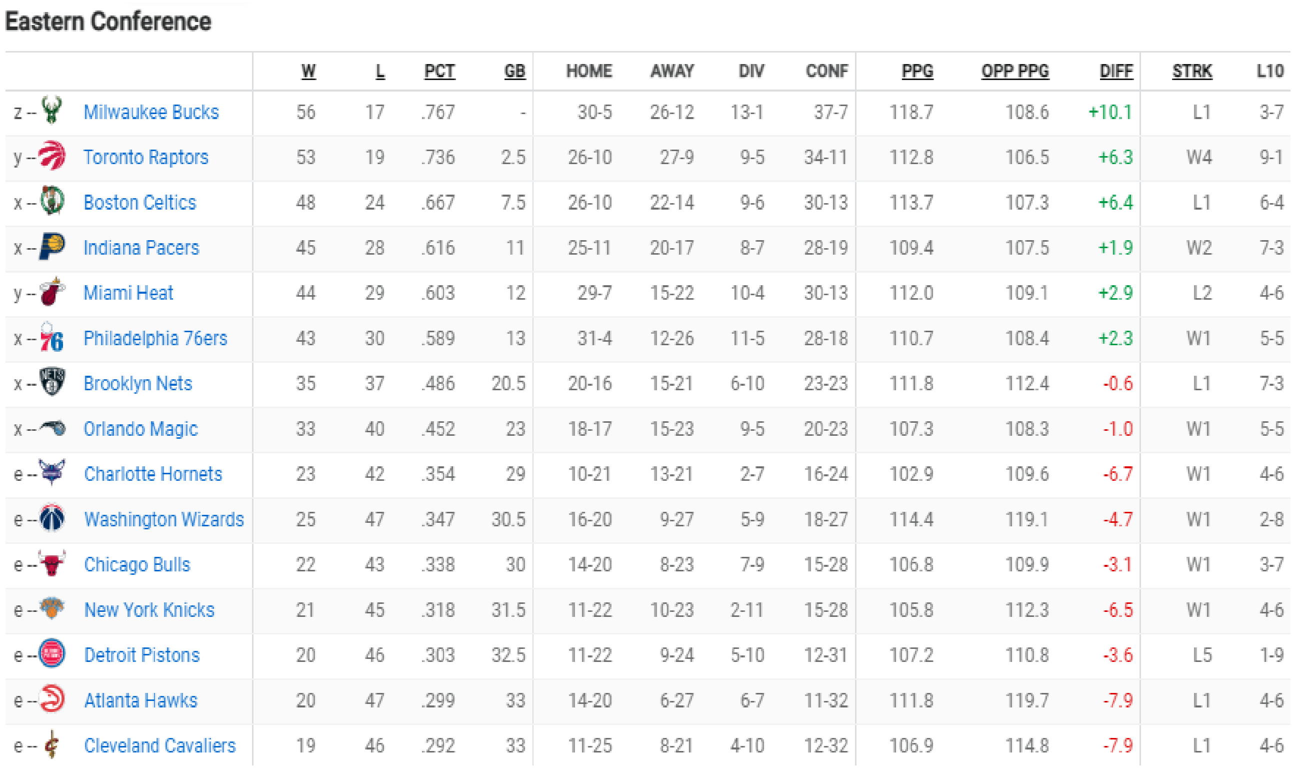
```
   FG%  3P%  2P%  DRB  O_FG%  O_2P%  O_BLK   W   L  PW  PL  MOV  SRS  ORtg  DRtg  eFG%  O_eFG%
0    2   16    1    1      1      1     11   1  30   1   1    1    1     7     1     1       1
1   20    6   18   10      2      5     21   3  28   4   3    4    4    14     2     8       2
2   18   12   21   13      4      9     25   5  26   6   5    5    6     5     4    19       6
3    9    1    8    7      7     16      5   7  21   8   9    8   11     6    12     3      13
4    3   12   10   20      8     10      8  12  19  13  13   13   13    18     7    10       7
```

Figure 19. Re-modeling after Feature Importance

5 Chapter Experimental Reviews

5.1 Results before Re-Modeling

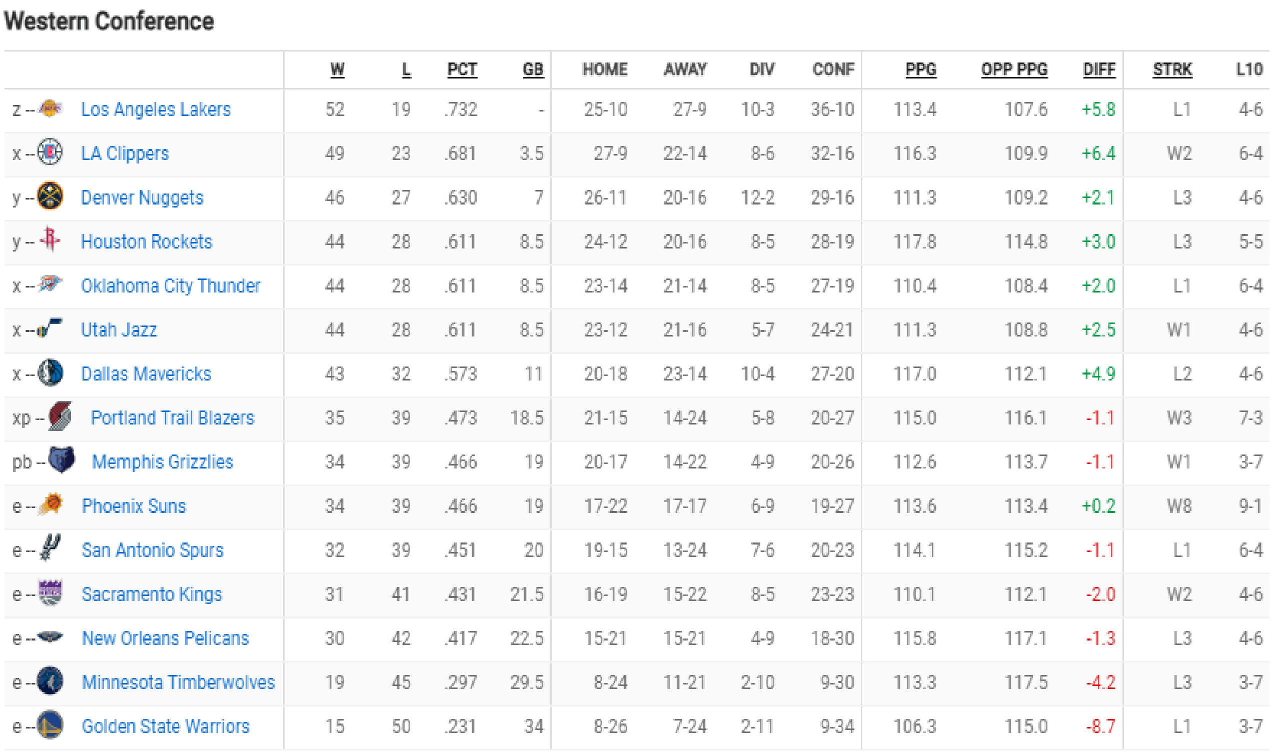
Results from our first Classifier (SVM) are displayed below (Figure 22). Observing our initial results Dallas Mavericks are in top position with 13 wins and Milwaukee Bucks on the second following Nuggets and Jazz. Those results seem not so accurate. Only Bucks obtain a high spot on our table (Bucks were being considered strong favorites of winning the Championship) (Figure 20). Mavericks and Utah Jazz were good franchise teams, but odds makers do not consider them as favorites for Title winners. Moreover, Lakers and Clippers two of the best teams in the league are extremely low on wins prediction. Lakers had the best record in West Division with the Clippers following them (Figure 21). Besides that, Celtics, and Raptors the second and third best record in East Division (Figure 20) but this superiority is not depicted in our SVM results. At first, our algorithm appears to be inaccurate, but let us wait for the playoff results in order to proceed in the results evaluation.



Eastern Conference

	W	L	PCT	GB	HOME	AWAY	DIV	CONF	PPG	OPP PPG	DIFF	STRK	L10
z -- Milwaukee Bucks	56	17	.767	-	30-5	26-12	13-1	37-7	118.7	108.6	+10.1	L1	3-7
y -- Toronto Raptors	53	19	.736	2.5	26-10	27-9	9-5	34-11	112.8	106.5	+6.3	W4	9-1
x -- Boston Celtics	48	24	.667	7.5	26-10	22-14	9-6	30-13	113.7	107.3	+6.4	L1	6-4
x -- Indiana Pacers	45	28	.616	11	25-11	20-17	8-7	28-19	109.4	107.5	+1.9	W2	7-3
y -- Miami Heat	44	29	.603	12	29-7	15-22	10-4	30-13	112.0	109.1	+2.9	L2	4-6
x -- Philadelphia 76ers	43	30	.589	13	31-4	12-26	11-5	28-18	110.7	108.4	+2.3	W1	5-5
x -- Brooklyn Nets	35	37	.486	20.5	20-16	15-21	6-10	23-23	111.8	112.4	-0.6	L1	7-3
x -- Orlando Magic	33	40	.452	23	18-17	15-23	9-5	20-23	107.3	108.3	-1.0	W1	5-5
e -- Charlotte Hornets	23	42	.354	29	10-21	13-21	2-7	16-24	102.9	109.6	-6.7	W1	4-6
e -- Washington Wizards	25	47	.347	30.5	16-20	9-27	5-9	18-27	114.4	119.1	-4.7	W1	2-8
e -- Chicago Bulls	22	43	.338	30	14-20	8-23	7-9	15-28	106.8	109.9	-3.1	W1	3-7
e -- New York Knicks	21	45	.318	31.5	11-22	10-23	2-11	15-28	105.8	112.3	-6.5	W1	4-6
e -- Detroit Pistons	20	46	.303	32.5	11-22	9-24	5-10	12-31	107.2	110.8	-3.6	L5	1-9
e -- Atlanta Hawks	20	47	.299	33	14-20	6-27	6-7	11-32	111.8	119.7	-7.9	L1	4-6
e -- Cleveland Cavaliers	19	46	.292	33	11-25	8-21	4-10	12-32	106.9	114.8	-7.9	L1	4-6

Figure. 20 East Rankings

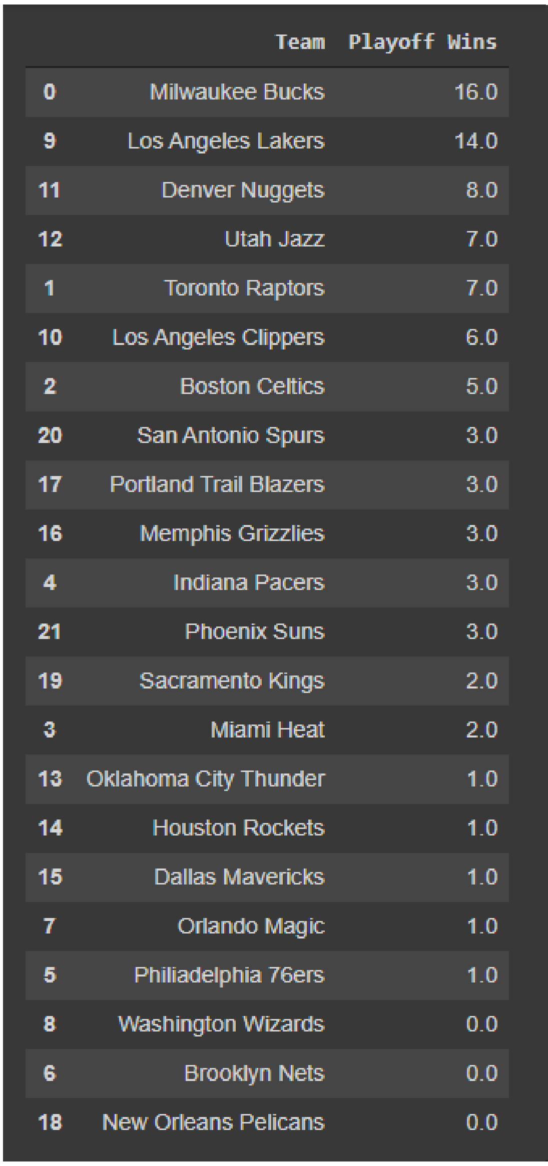


Western Conference

	W	L	PCT	GB	HOME	AWAY	DIV	CONF	PPG	OPP PPG	DIFF	STRK	L10
z -- Los Angeles Lakers	52	19	.732	-	25-10	27-9	10-3	36-10	113.4	107.6	+5.8	L1	4-6
x -- LA Clippers	49	23	.681	3.5	27-9	22-14	8-6	32-16	116.3	109.9	+6.4	W2	6-4
y -- Denver Nuggets	46	27	.630	7	26-11	20-16	12-2	29-16	111.3	109.2	+2.1	L3	4-6
y -- Houston Rockets	44	28	.611	8.5	24-12	20-16	8-5	28-19	117.8	114.8	+3.0	L3	5-5
x -- Oklahoma City Thunder	44	28	.611	8.5	23-14	21-14	8-5	27-19	110.4	108.4	+2.0	L1	6-4
x -- Utah Jazz	44	28	.611	8.5	23-12	21-16	5-7	24-21	111.3	108.8	+2.5	W1	4-6
x -- Dallas Mavericks	43	32	.573	11	20-18	23-14	10-4	27-20	117.0	112.1	+4.9	L2	4-6
xp -- Portland Trail Blazers	35	39	.473	18.5	21-15	14-24	5-8	20-27	115.0	116.1	-1.1	W3	7-3
pb -- Memphis Grizzlies	34	39	.466	19	20-17	14-22	4-9	20-26	112.6	113.7	-1.1	W1	3-7
e -- Phoenix Suns	34	39	.466	19	17-22	17-17	6-9	19-27	113.6	113.4	+0.2	W8	9-1
e -- San Antonio Spurs	32	39	.451	20	19-15	13-24	7-6	20-23	114.1	115.2	-1.1	L1	6-4
e -- Sacramento Kings	31	41	.431	21.5	16-19	15-22	8-5	23-23	110.1	112.1	-2.0	W2	4-6
e -- New Orleans Pelicans	30	42	.417	22.5	15-21	15-21	4-9	18-30	115.8	117.1	-1.3	L3	4-6
e -- Minnesota Timberwolves	19	45	.297	29.5	8-24	11-21	2-10	9-30	113.3	117.5	-4.2	L3	3-7
e -- Golden State Warriors	15	50	.231	34	8-26	7-24	2-11	9-34	106.3	115.0	-8.7	L1	3-7

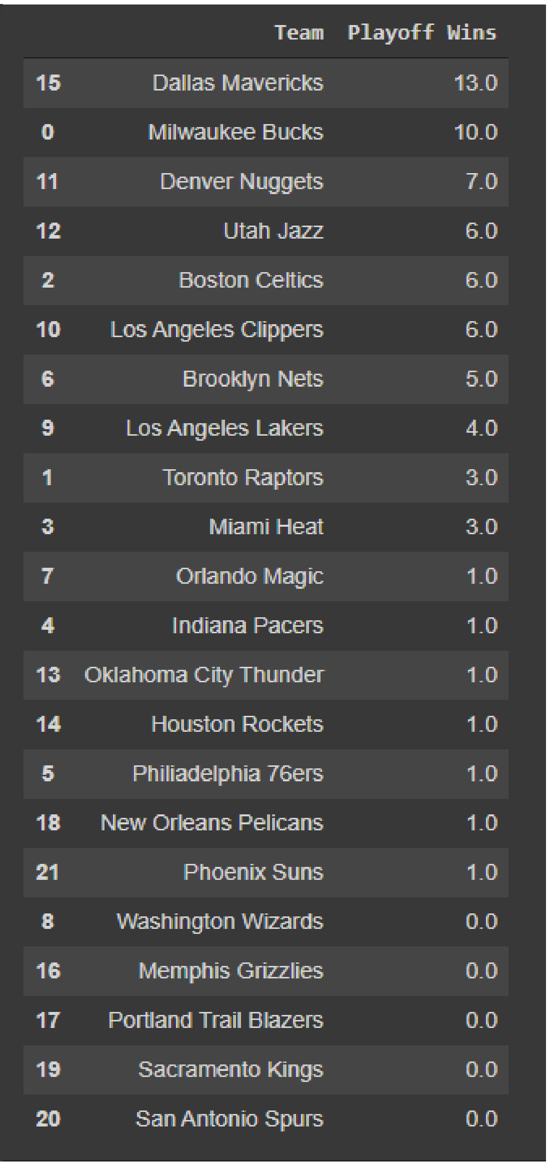
Figure 21. West Rankings

Decision tree classifier predictions (Figure 23) give the impression of being more precise than SVM results. Strong favorites like Bucks and Lakers projected to have a big number of wins. We noticed that again Nuggets and Jazz are expected to make more wins than Celtics, Raptors and Clippers. The last three teams were higher in the ranks according to odd makers.



	Team	Playoff Wins
0	Milwaukee Bucks	16.0
9	Los Angeles Lakers	14.0
11	Denver Nuggets	8.0
12	Utah Jazz	7.0
1	Toronto Raptors	7.0
10	Los Angeles Clippers	6.0
2	Boston Celtics	5.0
20	San Antonio Spurs	3.0
17	Portland Trail Blazers	3.0
16	Memphis Grizzlies	3.0
4	Indiana Pacers	3.0
21	Phoenix Suns	3.0
19	Sacramento Kings	2.0
3	Miami Heat	2.0
13	Oklahoma City Thunder	1.0
14	Houston Rockets	1.0
15	Dallas Mavericks	1.0
7	Orlando Magic	1.0
5	Philiadelphia 76ers	1.0
8	Washington Wizards	0.0
6	Brooklyn Nets	0.0
18	New Orleans Pelicans	0.0

Figure 23. Decision Tree Results

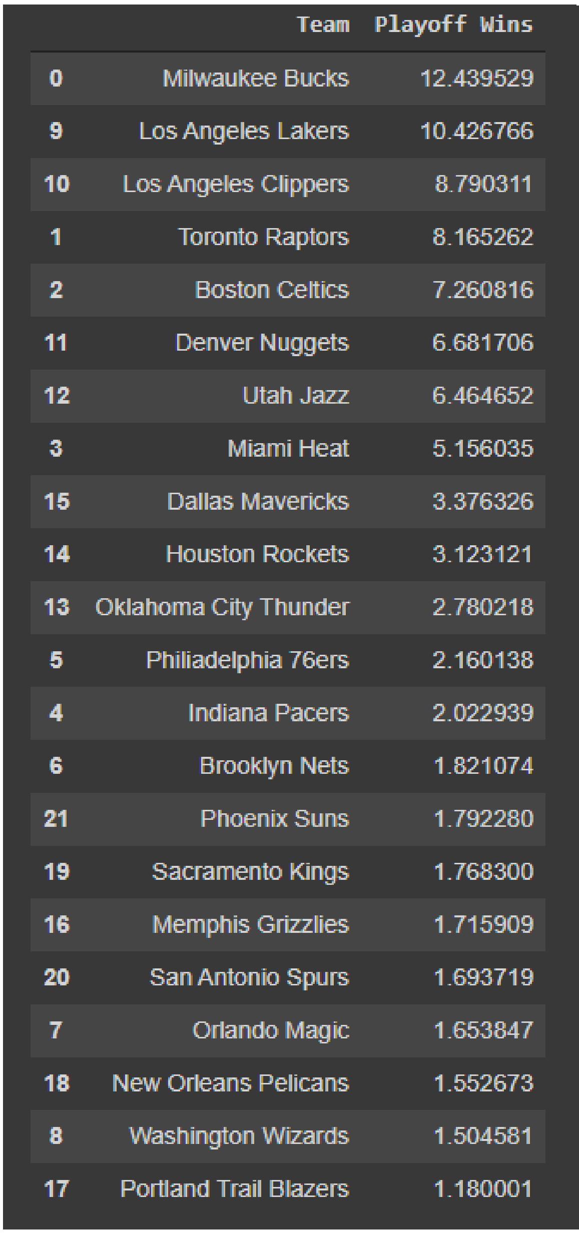


	Team	Playoff Wins
15	Dallas Mavericks	13.0
0	Milwaukee Bucks	10.0
11	Denver Nuggets	7.0
12	Utah Jazz	6.0
2	Boston Celtics	6.0
10	Los Angeles Clippers	6.0
6	Brooklyn Nets	5.0
9	Los Angeles Lakers	4.0
1	Toronto Raptors	3.0
3	Miami Heat	3.0
7	Orlando Magic	1.0
4	Indiana Pacers	1.0
13	Oklahoma City Thunder	1.0
14	Houston Rockets	1.0
5	Philiadelphia 76ers	1.0
18	New Orleans Pelicans	1.0
21	Phoenix Suns	1.0
8	Washington Wizards	0.0
16	Memphis Grizzlies	0.0
17	Portland Trail Blazers	0.0
19	Sacramento Kings	0.0
20	San Antonio Spurs	0.0

Figure 22.SVM Results

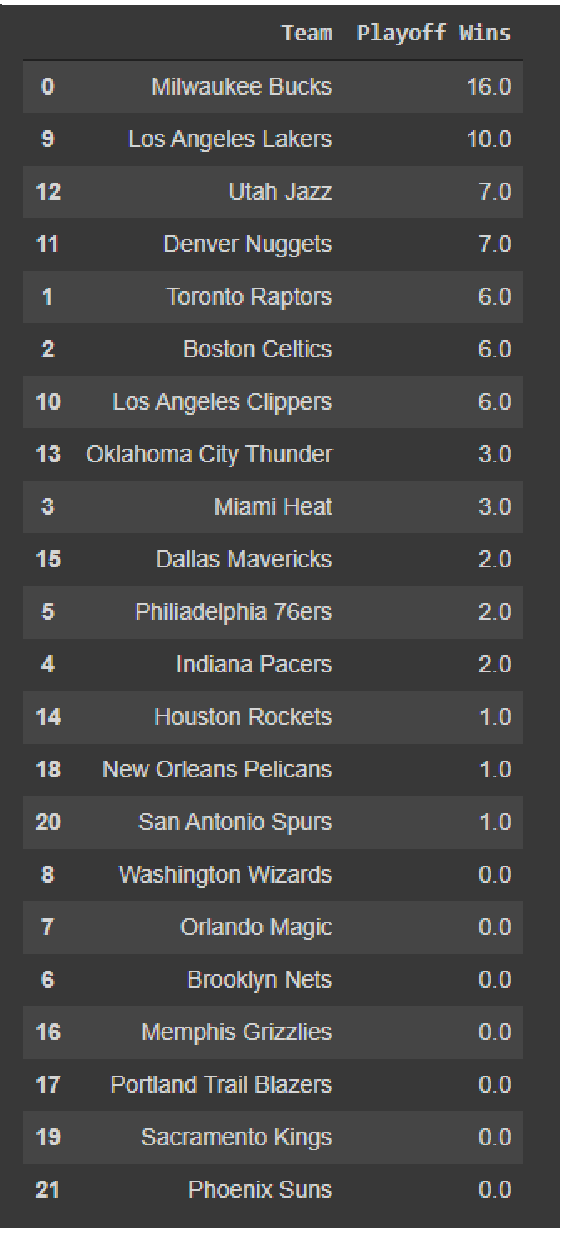
Third classifier XGboost seem to have the most precise results (Figure 24) compared to bookers predictions. Bucks, Lakers and Clippers and first, second and third spot. Raptors, Celtics and Nuggets are following. So, in general we have the top tier teams in the first 6 positions.

Our last classifier was Random Forest, we observe similar results (Figure 25) with Decision tree again Bucks and Lakers(Lakers projected to have 4 less wins than before) take the first 2 positions followed by Nuggets and Jazz. Again, three strong contender teams (Celtics) are low on our predictions.



	Team	Playoff Wins
0	Milwaukee Bucks	12.439529
9	Los Angeles Lakers	10.426766
10	Los Angeles Clippers	8.790311
1	Toronto Raptors	8.165262
2	Boston Celtics	7.260816
11	Denver Nuggets	6.681706
12	Utah Jazz	6.464652
3	Miami Heat	5.156035
15	Dallas Mavericks	3.376326
14	Houston Rockets	3.123121
13	Oklahoma City Thunder	2.780218
5	Philiadelphia 76ers	2.160138
4	Indiana Pacers	2.022939
6	Brooklyn Nets	1.821074
21	Phoenix Suns	1.792280
19	Sacramento Kings	1.768300
16	Memphis Grizzlies	1.715909
20	San Antonio Spurs	1.693719
7	Orlando Magic	1.653847
18	New Orleans Pelicans	1.552673
8	Washington Wizards	1.504581
17	Portland Trail Blazers	1.180001

Figure 24. XGBoost Results

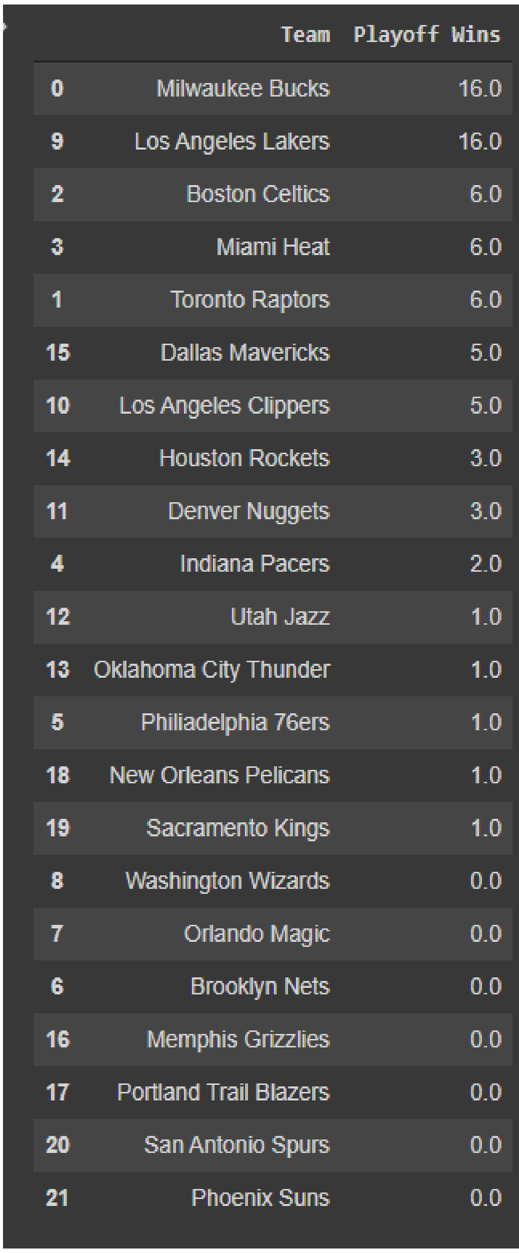


	Team	Playoff Wins
0	Milwaukee Bucks	16.0
9	Los Angeles Lakers	10.0
12	Utah Jazz	7.0
11	Denver Nuggets	7.0
1	Toronto Raptors	6.0
2	Boston Celtics	6.0
10	Los Angeles Clippers	6.0
13	Oklahoma City Thunder	3.0
3	Miami Heat	3.0
15	Dallas Mavericks	2.0
5	Philiadelphia 76ers	2.0
4	Indiana Pacers	2.0
14	Houston Rockets	1.0
18	New Orleans Pelicans	1.0
20	San Antonio Spurs	1.0
8	Washington Wizards	0.0
7	Orlando Magic	0.0
6	Brooklyn Nets	0.0
16	Memphis Grizzlies	0.0
17	Portland Trail Blazers	0.0
19	Sacramento Kings	0.0
21	Phoenix Suns	0.0

Figure 25. Random Forest Results

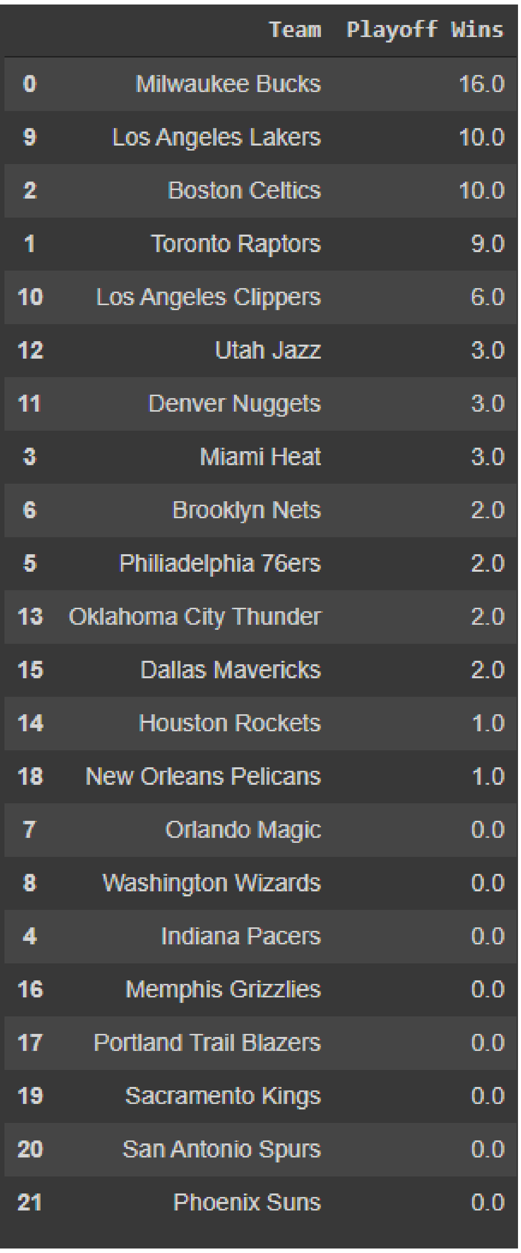
5.2 Results after Re-modeling

Results from our first Classifier (SVM), after we dropped the less essential features, are displayed below (Figure 26). At first glance results seem natural having the two best teams in the league are projected to achieve the maximum number of wins. Also, as we observed later Celtics, Raptors are predicted to get high place but not so many wins. Heat, Mavericks and Clippers again forecasted to get small number of wins. In Decision Tree the results (Figure 27) indicate that Bucks will be the undisputed champion and Lakers will battle with Celtics for a spot in a final. As we analyzed the results we perceived that Raptors are estimated to get 9 wins a big amount compared to other classifiers. Moreover, LA Clippers are expected to have more wins than the outcomes from other classifiers.



	Team	Playoff Wins
0	Milwaukee Bucks	16.0
9	Los Angeles Lakers	16.0
2	Boston Celtics	6.0
3	Miami Heat	6.0
1	Toronto Raptors	6.0
15	Dallas Mavericks	5.0
10	Los Angeles Clippers	5.0
14	Houston Rockets	3.0
11	Denver Nuggets	3.0
4	Indiana Pacers	2.0
12	Utah Jazz	1.0
13	Oklahoma City Thunder	1.0
5	Philiadelphia 76ers	1.0
18	New Orleans Pelicans	1.0
19	Sacramento Kings	1.0
8	Washington Wizards	0.0
7	Orlando Magic	0.0
6	Brooklyn Nets	0.0
16	Memphis Grizzlies	0.0
17	Portland Trail Blazers	0.0
20	San Antonio Spurs	0.0
21	Phoenix Suns	0.0

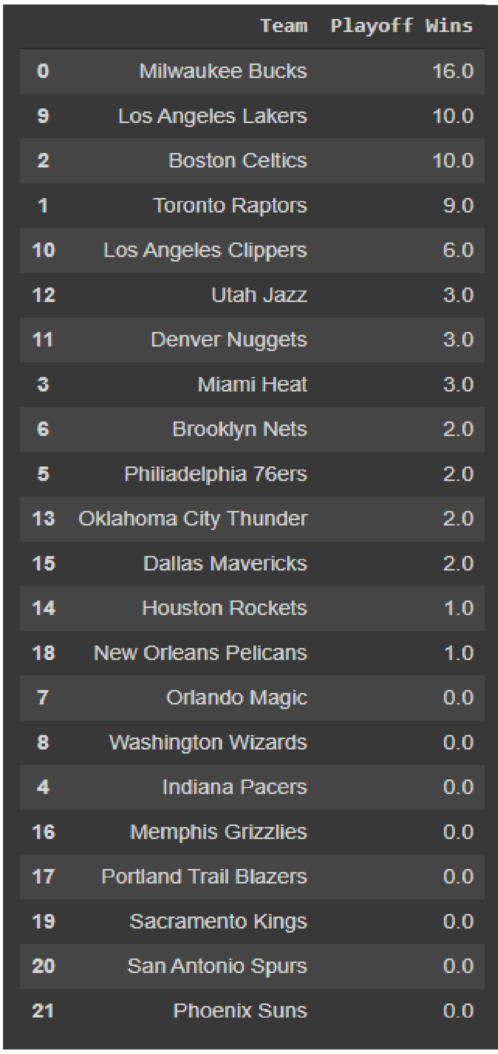
Figure 26. SVM Results



	Team	Playoff Wins
0	Milwaukee Bucks	16.0
9	Los Angeles Lakers	10.0
2	Boston Celtics	10.0
1	Toronto Raptors	9.0
10	Los Angeles Clippers	6.0
12	Utah Jazz	3.0
11	Denver Nuggets	3.0
3	Miami Heat	3.0
6	Brooklyn Nets	2.0
5	Philiadelphia 76ers	2.0
13	Oklahoma City Thunder	2.0
15	Dallas Mavericks	2.0
14	Houston Rockets	1.0
18	New Orleans Pelicans	1.0
7	Orlando Magic	0.0
8	Washington Wizards	0.0
4	Indiana Pacers	0.0
16	Memphis Grizzlies	0.0
17	Portland Trail Blazers	0.0
19	Sacramento Kings	0.0
20	San Antonio Spurs	0.0
21	Phoenix Suns	0.0

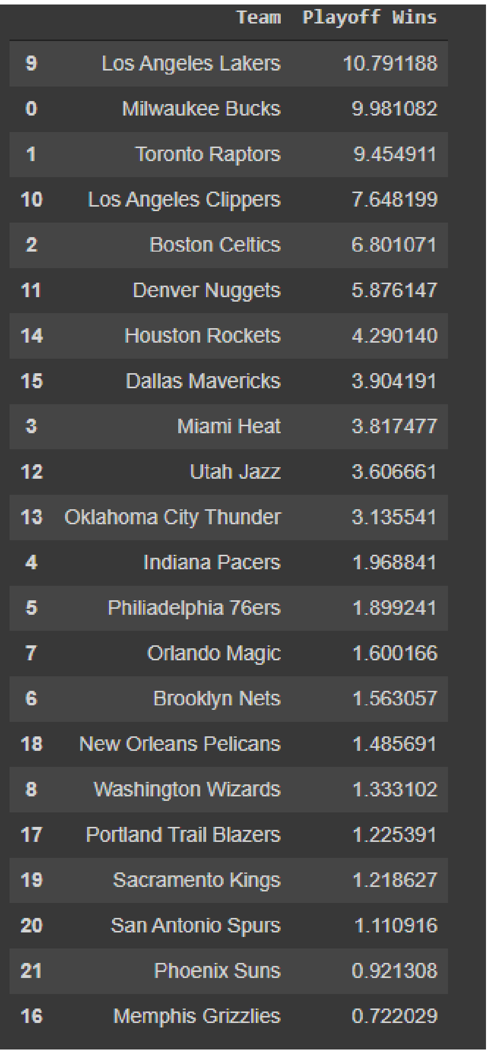
Figure 27. Decision Tree Results

In Random Forest results (Figure 29) we noticed a great change since the reigning Champions Raptors are expected to achieve 16 Wins along with the Bucks, leaving Lakers and Clippers in the next spots with 10 wins each. Nuggets again are estimated to finish in a high position with 7 wins and Heat with 5. Ultimately, the outcomes (Figure 30) from our last algorithm XGboost, showed us a new set of assumptions since Lakers are estimated to win the ring of the Champion with Milwaukee Bucks finishing in the second a spot leaving behind Raptors with a small difference. Furthermore, Clippers predicted to finish in the fourth position and Celtics, Nuggets in the following spots. One assumption we can make is that we a big drop for the Celtics and Miami Heat as well in contrast with the other results was monitored.



	Team	Playoff Wins
0	Milwaukee Bucks	16.0
9	Los Angeles Lakers	10.0
2	Boston Celtics	10.0
1	Toronto Raptors	9.0
10	Los Angeles Clippers	6.0
12	Utah Jazz	3.0
11	Denver Nuggets	3.0
3	Miami Heat	3.0
6	Brooklyn Nets	2.0
5	Philiadelphia 76ers	2.0
13	Oklahoma City Thunder	2.0
15	Dallas Mavericks	2.0
14	Houston Rockets	1.0
18	New Orleans Pelicans	1.0
7	Orlando Magic	0.0
8	Washington Wizards	0.0
4	Indiana Pacers	0.0
16	Memphis Grizzlies	0.0
17	Portland Trail Blazers	0.0
19	Sacramento Kings	0.0
20	San Antonio Spurs	0.0
21	Phoenix Suns	0.0

Figure 29. Random Forest Results

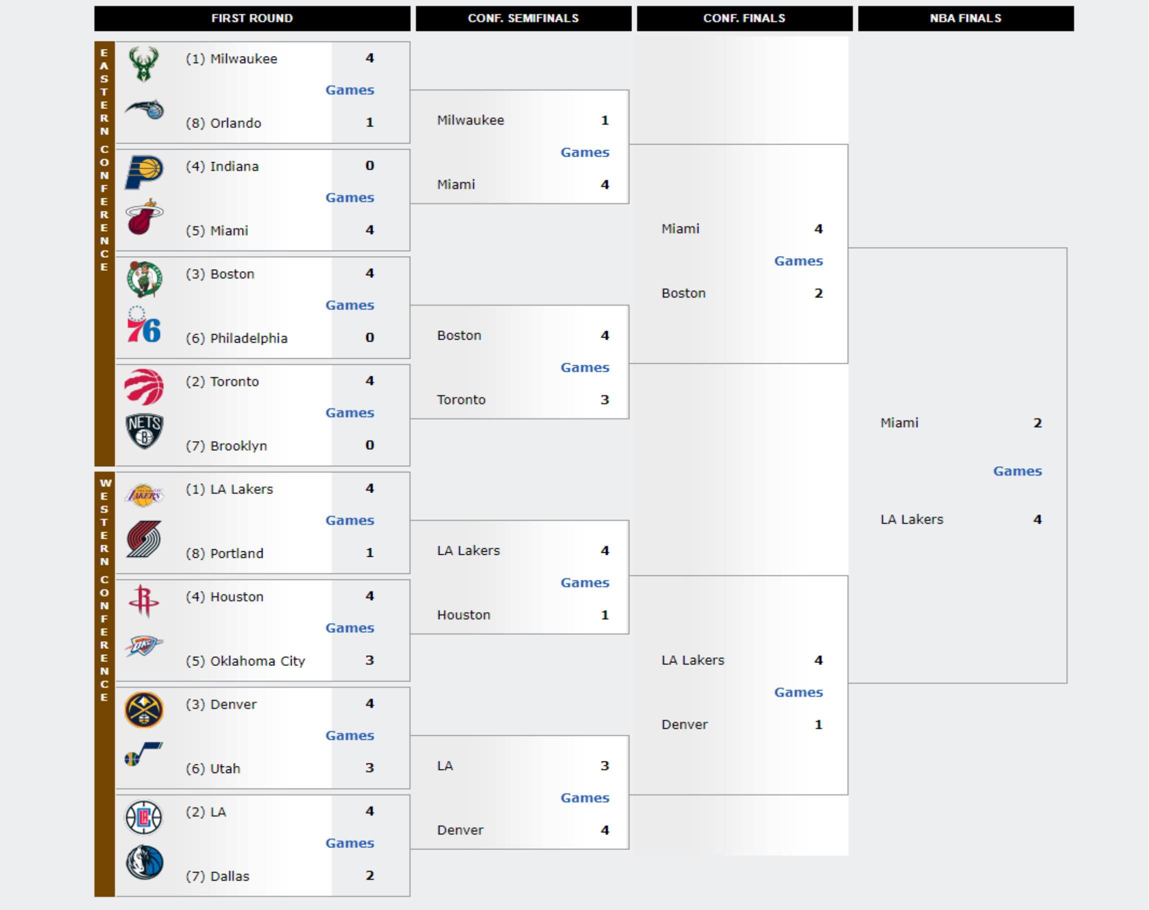


	Team	Playoff Wins
9	Los Angeles Lakers	10.791188
0	Milwaukee Bucks	9.981082
1	Toronto Raptors	9.454911
10	Los Angeles Clippers	7.648199
2	Boston Celtics	6.801071
11	Denver Nuggets	5.876147
14	Houston Rockets	4.290140
15	Dallas Mavericks	3.904191
3	Miami Heat	3.817477
12	Utah Jazz	3.606661
13	Oklahoma City Thunder	3.135541
4	Indiana Pacers	1.968841
5	Philiadelphia 76ers	1.899241
7	Orlando Magic	1.600166
6	Brooklyn Nets	1.563057
18	New Orleans Pelicans	1.485691
8	Washington Wizards	1.333102
17	Portland Trail Blazers	1.225391
19	Sacramento Kings	1.218627
20	San Antonio Spurs	1.110916
21	Phoenix Suns	0.921308
16	Memphis Grizzlies	0.722029

Figure 30. XGBoost Results

5.3 Projections Evaluation

As we mentioned before Playoffs for season 2019-20 took place in Orlando’s bubble. Let us see the results on the table below (Figure 31) as we see Lakers was the Champions as they beat Miami Heat in the finals. Milwaukee’s results were a really unpleasant surprise for their fans as they disqualified in the second playoff round from the biggest surprise of the tournament Miami Heat as they marched until the NBA Finals. Heat’s final position was a huge surprise. They shocked the NBA analysts and sport journalists with their presence to the NBA finals since their regular season results and stats were not foreboding this Payoff’s spot. Clippers was disqualified from Denver that also consists a surprise since Clippers were the third favorite, after Lakers and Bucks in Championship winners at the start of the season (Figure 31).Nuggets also move forward against the odds which they presented them as the last favorite team, until they disqualified from the, eventually Champions Lakers. Boston and Rockets match their expectations since they move to the second and third round each. Furthermore, 76ers did not even achieve a victory they eliminated in the first round, when odds placed them in the second round. Finally Nets and Mavericks reached their expectations (Figure 31) thus; they got eliminated in the first playoff round.



FIRST ROUND

EASTERN CONFERENCE

(1) Milwaukee 4
Games
(8) Orlando 1

(4) Indiana 0
Games
(5) Miami 4

(3) Boston 4
Games
(6) Philadelphia 0

(2) Toronto 4
Games
(7) Brooklyn 0

WESTERN CONFERENCE

(1) LA Lakers 4
Games
(8) Portland 1

(4) Houston 4
Games
(5) Oklahoma City 3

(3) Denver 4
Games
(6) Utah 3

(2) LA 4
Games
(7) Dallas 2

CONF. SEMIFINALS

Milwaukee 1
Games
Miami 4

Boston 4
Games
Toronto 3

LA Lakers 4
Games
Houston 1

LA 3
Games
Denver 4

CONF. FINALS

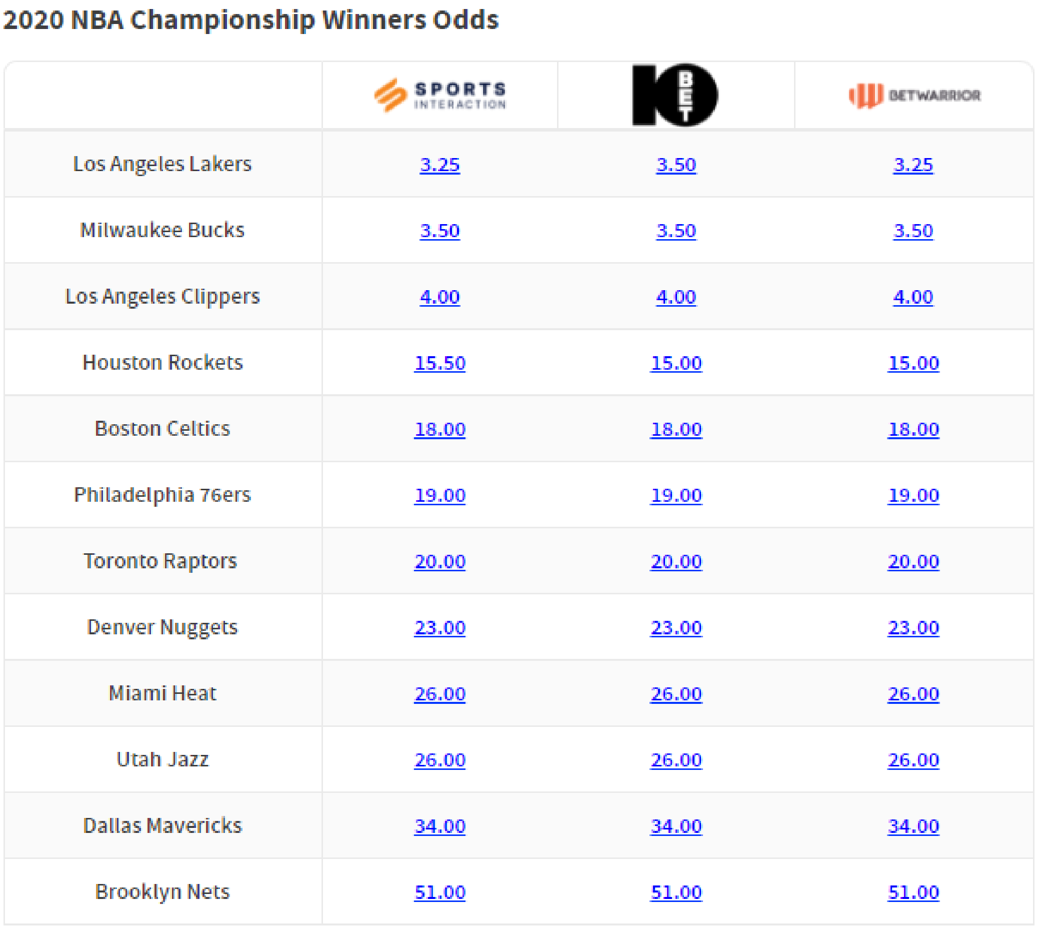
Miami 4
Games
Boston 2

LA Lakers 4
Games
Denver 1

NBA FINALS

Miami 2
Games
LA Lakers 4

Figure 30. Playoff bracket 2020



2020 NBA Championship Winners Odds

	SPORTS INTERACTION	10 BET	BETWARRIOR
Los Angeles Lakers	3.25	3.50	3.25
Milwaukee Bucks	3.50	3.50	3.50
Los Angeles Clippers	4.00	4.00	4.00
Houston Rockets	15.50	15.00	15.00
Boston Celtics	18.00	18.00	18.00
Philadelphia 76ers	19.00	19.00	19.00
Toronto Raptors	20.00	20.00	20.00
Denver Nuggets	23.00	23.00	23.00
Miami Heat	26.00	26.00	26.00
Utah Jazz	26.00	26.00	26.00
Dallas Mavericks	34.00	34.00	34.00
Brooklyn Nets	51.00	51.00	51.00

Figure 31. Odds for Playoff Winner

Now that we finally know, the results from the playoff matches we can compare them with our projections aiming to extract conclusions and test our algorithms efficiency. Initially in SVM classifier we predicted that Bucks along with Lakers will achieve the best percentage of wins. Partly we guessed right Lakers won the Championship, but Bucks did not move as far as they could, since their team affected the most from the game absence, consequently the lose their game rhythm and they play very low from their standards. Heat with Raptors and Celtics projected a fair number of wins and in this situation our algorithm was considered successful we can easily see that Celtics reached their number of wins and Heat they exceed them, raptors also failed to reach the number of wins. Finally, we predicted correct that Nets Indiana and Orlando will not proceed after the first round.

Later regarding Decision Tree classifier results, our estimated winner is Milwaukee, but as we mentioned earlier got eliminated after the first round from Miami. Lakers won the championship despite our predictions and Boston reached our projections wins and finished in the third spot. One more time we can observe that Miami was low on our results that is one more clue pointing that Miami made a big surprise on the Championship and were the biggest over achievers of the season. Furthermore, Clippers and Nuggets did not match our predictions since Clippers finished lower than our expectations and Nuggets higher than our estimations. We successfully guessed that Orlando, Indiana and Houston will not move forward at the second round.

The most efficient classifier was XGboost; thus, it predicted that Lakers would win the championship. One more time Bucks were falsely projected to finish second, but their expectations were not matched. Celtics and Raptors are correctly projected that their will finish high in the playoff tree. Again, we failed to forecast the big surprise of Miami since it was low on our results.

Finally, Random Forest falsely predicted that Bucks and Raptors will achieve the max amount of wins but correct estimated that Boston will finish in the fourth spot. The champions Lakers predicted to make only 10 wins. Additionally, we noticed that Nuggets first time are high on our results, very close on their actual position. According to our findings Heat predicted to achieve more wins than the other classifiers but again we could not guess their true amount of victory.

6 Chapter Conclusions and Further Work

6.1 Conclusions

In this research paper, we intended to predict the champion of NBA based on individual and team performance as well. Aiming to make the most accurate prediction, we used team statistical categories and some miscellaneous team categories from 18 seasons of

NBA, including all pre-season and playoff games stats except the playoff of this year (2020) that we aim to predict the winner.

The methodology which followed was to rank all teams according to their previous year's positions and then we filtered our data, keeping only the most important data that can be proven helpful on our paper's expected outcome. We trained our models with those data and we made our predictions. We classified the results be setting a number of desired wins, that a team need to achieve in order to win the championship. The amount of 16 wins was the max number but the issue was that 16 wins might not be sufficient to win the title. Therefore, the predicted wins number was scaled in purpose of seeing which team has the most predicted values as a potential champion.

We implemented various correlation graphs trying to depict how the features affect each other. Moreover, we tried to show the correlation that all our features have with the category "Play-off wins" which was our target feature. In that way we managed to filter not important features and achieve better prediction from our 4 algorithms. Next step in our experiment was to model our dataset and split by 70%-30% ratio. Our models were trained with their parameters tuned. Later, feature importance was applied and again we dropped not important columns according to the results which were extracted from feature importance technique. Afterwards, we re-modeled and trained again our classifiers and extract the final predictions.

Our findings were very interesting, we can say that the results as we predicted them matched the bookmaker's odds. The champions Lakers were predicted in all algorithms in the first position so we can say that we have a solid guess there. Moreover Celtics and Raptors were correctly estimated by our classifiers. Considering that bookers projected to be the eighth team in the ranking, we can claim that Nuggets was a successful prediction since we predict that that they can make a fair amount of wins, enough to secure a spot in the western final. Finally another solid guess was the low spot that eventually finished Brooklyn, Philadelphia, Utah and Dallas. Those were the accurate predictions now let's take a look on the in-accurate ones. It goes without saying that Bucks was the most imprecise projection, thus they hardly move until the first playoff round and according to our algorithms should have won the championship. Furthermore, Miami expected to finish on a low spot as stated in our classifiers results but they made a big surprise not only to us but to the bookers as well and they got to the NBA finals. Summarizing, this year was unprecedented hence, the pandemic changed the odds and the NBA format so nobody could project the outcome of this year's playoffs. Milwaukee was one of those teams which were mentally affected according to their players and coach statements. Moreover, many teams affected from the long absence of sport activities and failed to adapt to the new conditions. Taking into consideration all those parameters we can say that our predictions were accurate enough but of course we can enhance our experiments and achieve even more accurate results.

6.2 Future Work

The results from our experiments proved that we can predict the outcome of NBA games having data extracted from team statistical categories. It goes without saying that those results can be improved by adding more features or applying different techniques.

Since we live in a highly technological era, we can use data extracted from camera for various projects. Firstly, we can collect granular data on players movements, shooting habits and defensive positions. After that, those data can be analyzed aiming to suggest the best offensive and defensive strategy in order to win the game. With those video footage coaching teams can focus not only to the basic statistics like points assists but some very important information like, from where a player likes to shoot, or which foot did he use to make a layup or how defense reacts in a certain offensive system. So, having all this info available helped coaches to plan better strategies guiding their players to shoot from the week side of defense and take better positions in order to defend better according to their opponents' offensive habits.

Injury management is another field that can be exploited. Teams are collecting data by wearable magnetic jackets that can provide data for players sleep, heart condition and vital organs functionality. They even extract saliva samples in order to measure the fatigue level. Feeding all those data to machine learning models can be proven helpful to coaching stuff indicating them when to rest a player (a tired player is more injury prone) or when to use a player with minute restriction.

Finally, data extracted from NCAA college both team and players stats can be used with various machine learning models in order to indicate to club owners which player suits better on their team. This might be very useful on the draft night, having a suggestion extracted from machine learning combined with the potential ability evaluation of the possible draft player can guide the coaches to draft the player that fits the best on their team.

Bibliography

1. Lin, Rongyu. "Mason: Real-time NBA Matches Outcome Prediction." (2017).
2. Miljković, Dragan, et al. "The use of data mining for basketball matches outcomes prediction." IEEE 8th International Symposium on Intelligent Systems and Informatics. IEEE, 2010.
3. De Deyn, Bram. "Predicting Sport Results by using Recommendation." (2018).
4. Bailey, Michael J. Predicting sporting outcomes: A statistical approach. Diss. Faculty of Life and Social Sciences, Swinburne University of Technology, 2005.
5. Chazan-Pantzalis, Victor. "Sports Analytics Algorithms for Performance Prediction." (2020).
6. Kumar, Gunjan. "Machine learning for soccer analytics." University of Leuven (2013).
7. https://en.wikipedia.org/wiki/Charles_Reep
8. Wenninger, Sebastian, Daniel Link, and Martin Lames. "Performance of machine learning models in application to beach volleyball data." International Journal of Computer Science in Sport 19.1 (2020): 24-36.
9. Bosch, Pablo. "Predicting the winner of NFL-games using Machine and Deep Learning." (2018).
10. Jones, Eric Scot. "Predicting outcomes of NBA basketball games." (2016).

11. Stefani, Raymond T. "Football and basketball predictions using least squares." IEEE Transactions on systems, man, and cybernetics 7.2 (1977): 117-21.

12. Zak, Thomas A., Cliff J. Huang, and John J. Siegfried. "Production efficiency: the case of professional basketball." Journal of Business (1979): 379-392.

13. Stern, Hal S. "A Brownian motion model for the progress of sports scores." Journal of the American Statistical Association 89.427 (1994): 1128-1134.

14. Berri, D. J. 1999. Who is'most valuable'? Measuring the player's production of wins in the National Basketball Association.Managerial and Decision Economics20 (8):411–27

15. Hu, F. and J. V. Zidek, "Forecasting nba basketball playoff outcomes using the weighted likelihood", Lecture Notes-Monograph Series pp. 385–395 (2004).

16. Melnick, M. J., "Relationship between team assists and win-loss record in the national basketball association", Perceptual and Motor Skills 92, 2, 595–602 (2001).

17. Kvam, P. and J. S. Sokol, "A logistic regression/markov chain model for ncaa basketball", Naval Research Logistics (NrL) 53, 8, 788–803 (2006).

18. Shirley, K. 2007. Markov model for basketball. In New England Symposium for Statistics in Sports.

19. Trawinski, K., "A fuzzy classification system for prediction of the results of the basketball games", in "Fuzzy Systems (FUZZ), 2010 IEEE International Conference on", pp. 1–7 (IEEE, 2010).

20. Strumbelj, E., and P. Vra car. 2012. Simulating a basketball match with a homogeneous Markov model and forecasting the outcome.International Journal of Forecasting28 (2):532–42.

21. Cao, C., "Sports data mining technology used in basketball outcome prediction", (2012).

22. Magel, Rhonda, and Samuel Unruh. "Determining factors influencing the outcome of college basketball games." Open Journal of Statistics 3, no. 04 (2013): 225.

23. DeLong, C., N. Pathak, K. Erickson, E. Perrino, K. Shim and J. Srivastava, "Teamskill: modeling team chemistry in online multi-player games", in "Pacific-Asia Conference on Knowledge Discovery and Data Mining", pp. 519–531 (Springer, 2011).

24. Omidiran, D. 2013. Penalized regression models for the NBA. arXiv preprint, arXiv:1301.3523.

25. Lopez, Michael J., and Gregory J. Matthews. "Building an NCAA men's basketball predictive model and quantifying its success." Journal of Quantitative Analysis in Sports 11.1 (2015): 5-12.

26. https://medium.com/fintechexplained/what-is-grid-search-c01fe886ef0a

27. https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f

28. https://machinelearningmastery.com/k-fold-cross-validation/

29. https://builtin.com/data-science/random-forest-algorithm

30. Mavroforakis, M. E., & Theodoridis, S. (2006). A geometric approach to support vector machine (SVM) classification. IEEE transactions on neural networks, 17(3), 671-682.

31. https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989

32. https://philipppro.github.io/Hyperparameters_svm_/

33. https://blog.quantinsti.com/xgboost-python/

34. https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/

35. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

36. https://www.tutorialspoint.com/scikit_learn/scikit_learn_decision_trees.htm

37. https://github.com/trustinyoon/2020-NBA-Chip-Predictor/blob/master/README.md

38. V. Sarlis, C. Tjortjis, "Sports Analytics – Evaluation of Basketball Players and Team Performance", Information Systems, 2020