The Methods and Techniques used in Different Professional Sports Analytics

Emmanuel Oziegbe
Yahia Elrayes

Abstract

Businesses have been using analytics, since the late 19th century, to investigate and explore past and current data to understand their performance and predict future challenges in order to develop new business planning, improve that performance and overcome the challenges; therefore, businesses are far ahead of professional sports in this game. However, the majority of professional sports started using analytics for different purposes including: measuring the performance of the team as well as the individual player, predicting the likelihood of winning and losing, predicting the next moves of the opponent team, preventing injury, and many other purposes that could help boost the performance of the team and increase the chances of winning. The use of analytics in sports is becoming increasingly important as it changed how the game in each sport is played because decisions in all aspects of the game are now based on analyzed facts and predictions.

However, there are different analytical techniques and methods that can be in different sports for different purposes. In this paper, we review 25 papers on sports analytics published in academic journals and we review the analytical techniques and methods that are used by almost 10 different professional sports, the purpose of using those techniques, the analytical methods used in each sport, and the outcome of using the technique and how it might have helped the team manager take the right decisions. We conclude the paper with ranking the analytical techniques by the most widely used ones across different sports to see what methods work best for sports analytics in general.

Introduction

Sports analytics is an extensive study that assist in examining production on and off each sport's respective grounds, to gain insight or in many situations a competitive advantage. Sports analytics help a front office to better identify the intricate strategies their opponents utilize in exhibition. In addition, it can be used to generate the best possible team and playbook which generally corresponds to optimal production (a la 2003's "Moneyball" by Michael Lewis). The extensive study directs the attention to specified player(s) game, and places values based on data-supported identifiable skills. The proper utilization of analytics can create athletic success on the field, court, rink and even ticket sales.

Although much evidence supports the growth of sports analytics, it is not immune to flaws and challenges. There is much drawback and criticism for data-driven decision making. *Harvard Business Review* supports the notion that computers and algorithms may ignore variables when anticipating an athlete's behavior (Davenport, 2014). It also cannot be understated the established culture in the realm of sports. In many sports athletes thrive on intuition and experience to make key decisions during fast-paced gameplay. As a result, regardless of the vast amount of data and analytics at their disposal it is unnatural for it place precedent over lifelong instinct. However, while it is expressed analytics will not perform to act as a substitute for skilled play and great coaching, it certainly has established itself as an important enhancement for those basic key success factors.

We will dive into scholarly articles that detail plenty of research exploring methods and techniques within the realm of sports analytics. These methods explore forecasting performance and outcomes, identifying value drivers, draw inferences, etc. The prevalent techniques dissected by the researchers are the Naïve Bayesian model, Logistical Regression, Clustering, K-Nearest Neighbor, Bootstrap resampling, Random forest, Multitask Regression, Pruned Regression Tree, Patterning Mining, Support Vector Machines, and Maximum Entropy Model with K-Means Clustering. The techniques were applied in research to a broad spectrum of sports, ranging from basketball, soccer, football, hockey, cricket, volleyball, golf,

and rugby.

Literature Review

Data Analytics, whether it was for businesses or sports, can be classified into three different types; Descriptive Analytics, Predictive Analytics, and Prescriptive Analytics.

Descriptive Analytics is the most common and basic form of data analytics; the majority of organizations, including business and sports organizations, use descriptive analytics today. This type of analytics involves gathering historical data, organizing it, and then provides details in the form of useful information, or in other words facts, regarding the quality of this data. It simply answers the question "What has happened?" However, Descriptive Analytics does not provide the reasons for why something happened. Past data can be a year ago, or a minute ago and therefore, when a system says it provides a real-time analytics, it basically falls under Descriptive Analytics.

The second type is Predictive Analytics; this type incorporates analyzing past data trends and patterns to, potentially, predict what is likely going to happen in the future; although the vast majority of data analytics, used by commercial businesses, falls under descriptive analytics, most of sports researched in this paper, as we are going to see, are focusing on methods of predictive analytics because predicting the future performance of players and opponent's next move is crucial to winning. However, predictive analytics does not provide any course of action based on those predictions and that's why we have a third type of analytics.

The third and least common type of analytics is prescriptive analytics; this one includes methods for optimization by offering suggestions of solutions to some problems and ways to tackle future challenges; prescriptive analytics considered to be the next phase of analytics after predictive analytics and it is certainly more advanced than the other two as it actually uses the other two types of analytics to advice the best course of action.

In the following sections of this paper, we will present all analytical methods and techniques used by different sports in the researched papers and classify them per type of data analytics; we will first describe the method and briefly explain how it is used and then talk about the

papers that used this technique as well as the sport that it was used for and how it helped or can help achieve the desired results.

1- Regression Analysis

Among the 25 researched papers, Regression analysis is the most commonly used in different sports as a form of predictive modeling technique that basically looks into the relationship between the target, which is the dependent variable that we are trying to predict, and the predictor(s), which is/are the dependent variable(s) that might affect the results of our target. For example, if we are trying to predict whether a player is going to score in the next game (this would be our target/ dependent variable), we might use the number of goals he scored in the past 5 games, the number assists, and the number of shots on goal (these would be our predictors/independent variables); the regression analysis will allow us to know the significance of the relationship between the dependent variable and the independent variables. Out of 25 papers researched, Regression Analysis was used in **9** papers (sometimes with other techniques).

However, there are many types of regression that can be performed depends on the type of sport as well as the type of variable we are trying to explain or predict. Nevertheless, we are only going to list those types used in the researched 25 papers if sports analytics.

1.1- Linear Regression / ordinary least squares

The ordinary least squares is a technique, used in statistics, to estimate a parameter in a linear regression model or linear function to minimize the squares of the difference between the actual numbers of what we are trying to predict and the numbers we have predicted by our model. This can also be graphed as a scatter-plot with a straight line, representing the regression, and data points, representing the actual values of the variable being predicted; the closer the points to line the better the model.

Out of 25 papers on sports analytics, the linear regression or ordinary least squares technique was used in 2 papers to predict variables in the game of Hockey.

The first paper uses linear regression to analyze team and player performance in the National Hockey League to predict player's future performance and contribution to his team; this is translated into the expected goals the team is will score based on data from the last NHL season. In this model, they initially used multiple independent variables to predict goals scoring rate including: goals, shots, missed shots, blocked shots, hits, Shots plus missed shots, zone starts, turnovers, goal shooting percentage, faceoffs, and Fenwick and Corsi ratings.[1] However, they start removing variables until the model yielded a good fit. The mean squared error (between actual and predicted) was used to measure the performance of the model itself.
Also for the game of Hockey, the second paper uses linear regression in combination with k-means clustering to first, define distinct player types, using the clustering, for different positions in a NHL team; and then second, use the linear regression to determine a quantitative relationship between the player type and his performance and effect on the team; the results were then used to determine two things: Accounting for their salaries based on their contribution to the performance of the team and determine whether a certain type of players is undervalued or overvalued and also evaluate the trades of each player type and his performance and how that can affect the performance of the team. The different types of players used as the independent variables in a multiple linear regression equation; while the number of points obtained by a team in a regular NHL season (which is the performance they are trying to estimate based on the player type) was used as the dependent variable in the same regression equation[2].

1.2 - Logistical Regression

The logistical regression is a method used in data analytics where a data set has one or more independent variable that can determine an outcome. The method seeks to find the probability of an occurrence depending on the values of the independent variables. These independent variables can be categorical or numerical, and are identified as the influencer to the dependant. Logistical regression discovers the optimal predictive model to describe the relationship between the independent and dependent variables. In logistical regression

dependant variables can only be categorical, meaning the values must be binary. It estimates the probability that an event occurs for a randomly selected observation versus the probability that the event does not occur. In the format of numerical values 0 or 1; 0 representing non happening of event and 1 represent happening of event. In the realm of sports analytics, Logistic regression can be best utilized to predict the probability of success in an exhibition match, based on past statistical measurements.

Researchers Paul Britton and Carl R. Yerger dissected the coaching strategy of well-known Bob McKillop; basketball coach at Davidson University. McKillop divided the game of basketball into 10 rounds. Before each game coach McKillop would provide his team with "round goals"to assist in measurement of success in certain aspects of the game. Reaching these goals within each round hopes to result in a win by McKillop's standards. The research sought to discover information on whether these rounds affected the outcome of the game and if so, which rounds are more important. These factors are then used to create a predictive model, Logistic regression is incorporated. McKillop felt the strategy of dividing the game of college basketball into "boxing style" rounds would help narrow the focus and motivate his players. McKillop's desire for the team was too win at least 7 rounds of the 10. The binary variables identified were wins and losses (1,0). Each round examined was tested for level of significance and correlation. The results of the study showed that the first round had the highest Z-score in logistic regression, however all 10 rounds showed to be significant predictors of the outcome of a college basketball matchup (.0001 significance level). The regression model was also an effective predictor for margin of victory (Britton & Yerger,2015).

In a similar fashion researchers Robert E Baker and Ted Kwartler used logistic regression to better understand the effectiveness of offensive play calling in the National Football League. The study focused on data from two teams over 13 seasons of play (2000-2012); Pittsburgh Steelers and Cleveland Browns. That is about 442 games of football between both teams. The main objective of this study to demonstrate logistic regression as an optimal methodology to assist NFL defensive coaches assess the chances of a rush or pass by an opposing team's offense.

The predictive model was based on 3 downs of play (1st down, 2nd down, and 3rd down) regulated on a data point system which totaled 26,310. The regression developed by researchers sought to identify the driving factors in play selections. Logistic regression proved to be effective in game coaching strategy and determining opposing team game plan. The algorithm (created through logistical regression) used for offensive play calling was correct for 66.4% of plays (Cleveland Browns) and 66.9% of plays (Pittsburgh Steelers). Further proving this method of sports analytics can benefit in game decisions (Baker & Kwartler, 2015).

Logistic regression was also used in analyzing shooting style of NBA players using their body pose and then see if various types of movements and body poses correlate with shots being missed or made. To do that, different movement types and poses needed to be classified with a focus on three point shot attempts. In identifying between open shots versus tough shots in the research, linear and nonlinear classifiers were trained in the model including: Random Forest, Support Vector, and Logistic Regression. In the model, the Logistic Regression outperformed the other two methods. The paper evaluated and identified attributes and body poses that might be important for a successful shot outcome[1].

And because, as mentioned, logistic regression is used when the predicting results that are not continues or the outcome has a limited number of possible values, the method was also used, in combination with other model, to Identify fast bowlers likely to play test cricket based on age-group performances. The model helped determining individuals with great prosperity to play test cricket for New Zealand[2].

1.1- Poisson Regression

Poisson Regression is used in modeling only count and numerical data so the dependent variable can only be continuous number; it can be positive or zero but it cannot be negative. This model was used in only one study out of the paper reached for this review; the objective of the study was to measure the performance of Golf players as the Model helped in describing the number of strokes needed to finish a hole. The regression function consisted of shots,

strokes assessed before the current shot, yardage of the hole being played, and the distance (in inches) to Pin[3].

1.2- Support Vector Machines

Support Vector Machines is a classification technique that is used with classification and regression analysis; it helps splitting the data in the best possible way by designing a hyperplane in the widest road that separate into two classes. Support Vector Machines was used, with other techniques, in 5 different papers out of the total of 25 research papers studied.

The method was used to help predict the outcome of ODI (One Day International), in the game of Cricket, based on four factors or variables. ODI is an internationally recognized format of Cricket matches.

The results showed that, on average, Support Vector Machines classifier outperformed Naïve Bayesian and Random Forest classification techniques in the study. Another study used Support Vector Machines is in analyzing shooting style of NBA players using their body pose and then see if various types of movements and body poses correlate with shots being missed or made (referenced above). The method was used here with logistic regression and Random Forest. However, the precision of Support Vector Machine was not as good as logistic regression or Random Forest in this paper. So, we conclude that it really depends on the sport, the outcome we are trying to predict, and the data itself when it comes to evaluate methods and techniques.

In a different study, SVM was used to estimate individual player fitness and link that to the performance of the player. The approach was suggested to be generally applicable for any sports, however, the data used and the case study in the paper was for a professional rugby club; The model performed generally well according to the study. However, the accuracy of the model was lower than the linear regression which was used in the same study; the reason is that the linear regression is using the entire data set to develop the model while SVM only uses a subset of the data. Finally, the model was used again in a classification task to illustrate some insights about passes and attempt passes in the game of soccer and categorize teams by where

in the pitch they attempt to pass, then use passing locations to predict successful shots, the results of this predictive model is then used to rank individual players. The data used was from the 2012–2013 La Liga season. the study used two models, SVM and K-nearest neighbor and achieved accuracy of 87%.

2- Random Forest

Random forest is an algorithm used as a method for classification, regression, and other tasks. The algorithm chooses a random sample from a large set of data and a subset of the initial variable to build a CART model; then repeats the same process a number of times then make a final prediction decision that based on a function of each prediction (that can be the average of each prediction). Furthermore, Random forest can overcome the issues of decision trees that are sometime overfitting to their training dataset.

Random Forest was used in, two studies among the researched ones, as a classifier, along with Naïve Bayesian and Support Vector Machines, to help predict the outcome of ODI (One Day International), in the game of Cricket, based on four factors or variables. (study is already mentioned above). The method was used in another study, with two other classifiers SVM and logistic regression, to analyze the shooting style of NBA players using their body pose and then see if various types of movements and body poses correlate with shots being missed or made (paper mentioned above). The method was not the best here and it was not the worst either. With mean

Average Precision 57.1%, the Logistic Regression classifier outperforms Random Forest (54.7%) and Support Vector Machine (51.5%) classifiers. All classifiers outperform random chance (50.0%)[4].

3- Naïve Bayesian

Naïve Bayesian is a probabilistic classifier and like decision trees, Naïve Bayesian is based on frequency tables. The technique is known to be used primarily for solving text classification problems and it's simple, models can be built fast and enables for quick prediction. It is called

naïve because it assumes that the occurrence of some features is completely independent of the occurrence of other features.

This, however, makes it a good choice for many classification problems and it was the reason behind choosing Naïve Bayesian to help predict the outcome of ODI (One Day International), in the game of Cricket, based on four factors or variables since those variables were independent; and because it employs Bayes' theorem to compute the probability of winning or losing, which was the outcome the study was trying to predict.

In addition to this study, Naïve Bayesian was employed in other studies among the researched 25 paper.

We see a high relevance of the Naive Bayesian model utilized in the sport of basketball. Various researchers took on the task of detecting the effectiveness of the Bayesian model when it comes to forecasting outcomes in professional and amateur sports. "Basketball predictions in the NCAAB and NBA: Similarities and differences" is a research paper published by Albretch Zimmerman which explores the differences in predicting the outcomes of NBA and NCAAB games. The purpose of this experiment is to gain an understanding on which type of data information is most effective for achieving high probability predictions. Zimmerman evaluated different kinds of NBA and NCAAB descriptive stats, trained the classifier on available data, and utilize prior season statistics. The four factors used to assist in creating a Naïve Bayes model were Effective field goal percentage, Turnover percentage, Offensive Rebound percentage, and Free throw percentage. The results of the studied show that it would be easier to predict season success in correlation to post-season success in the pros but not so much on an amateur level (Zimmerman, 2016).

"Luck" will always be in argument against data analytics in sports, the idea of probability welcomes such an argument. Regardless of the likelihood for a certain situation to occur, there's a small chance that a different action will take form, many refer to that as "luck". Sports analytics acts a deterrent to such, removing luck from the equation and replacing it with irrefutable evidence through data and methodology solidifying in game decision making. One

of the research papers in reference, highlights the difficulties of sport prediction and ultimately defeating "luck". The researchers examined a variety of sports to study the roles of skill and luck in competitive sports using the Bayesian Model along with other analytical techniques. Data was gathered from 198 sport leagues, from 84 countries, and composed of 4 different sports: basketball, handball, soccer, and volleyball. The researchers created a Naïve Bayesian probabilistic graphical model to understand skills needed in each sport and measure the relative input of luck and skill spotted in each game. The Bayesian model was mainly used on the sport of basketball, to find the correlation between estimated team skill and wins on the season. Result from data testing determined that in typical NBA season 35% of the time an underdog wins a match, this would be considered luck. Also the home advantage, adds .18 to the probability as opposed to .27 when away (Aoki, Assuncao, & Melo, 2017)

3.1- K-Nearest Neighbor

k-nearest neighbor is another algorithm used for classification and regression analysis. The input is the number of the closest data points (K) and the outcome, in the case the method was used for classification, will be classifying an object by the majority class of its neighbor data points (in other words, if an object is similar to its neighbors, then it's actually one of them).

This classification technique was used in two studies out of the researched 25 papers. k-nearest neighbor was used in the game of soccer to categorize teams by where in the pitch they pass or attempt to pass, then use the passing locations to predict successful shots. The study used heatmaps to represent the teams' passing style and then K-nearest neighbor was used as a classifier on data from 20 teams in the 2012–2013 La Liga season.

The experiment was repeated 2000 times until an accuracy of 87.3% was achieved, which suggest that, first: K-nearest neighbor was a good choice for the classification task, and second: a teams' passing style is highly characteristic.

A second study we explored A second study attempted to predict soccer results using simple

classification algorithms – K- nearest neighbor. The researchers plan was to compare different classification algorithm and choose which was best for forecasting in professional and amateur sports. K-Nearest Neighbor method is built on the idea that new unclassified data points may be considered classified depending on nearest data points in spectrum. To measure the skills of the players and their respective teams an algorithm was developed based on 2 years' worth of past data. The data types were - successful attempt ratio, per game ratio, and both ratios multiplied by one another. They compared the player/team performance for different K values. K-NN did not prove to be the better method between the two (Linear regression) but it did help prove as expected, that home wins go up as the difference in score goes up and the same is applied to away team victories (Stockholm, 2016).

4- Clustering

Clustering is used in data mining and statistical analysis to group a set of objects, that have similar features, together in the same group or cluster. Clustering technique has no specific algorithm or parameter setting because that depends on the data we want to cluster and the desired use of the results; and because clustering is an optimization problem, we usually will need to do trial and error and modify the data processing and data parameters until we achieve the desired results. However, there are many algorithm and cluster model used for clustering and we will present one or more of these models in this paper.

4.1- K-means Clustering

k-means clustering is based on centroid clustering algorithm, where clusters are represented by variables that may or may not be part of the used dataset. This algorithm is known work well for large dataset and to be computationally tractable. The input of the algorithm will be K (which is the number of clusters we would like to find) and a set n number of points (n is also the number of observation in our model). After choosing the number of clusters (for example two clusters) we want form, we then use the same number of centroids (data points); these centroids will initially be assigned randomly by the algorithm because the model still does not

know where is the center of each cluster; then it calculates the distance between the centroids then a perpendicular bisector (boundary line) divides this line between the centroid by half to mark the regions of the two clusters; the next step in the process is to repeat the same steps, and so the centroid will be moving slightly, to minimize the objective function and get closer to the desired results with every iteration.

In our researched papers, K-means clustering was used in two studies.

In Quantifying the Contribution of NHL Player Types to Team Performance, K-means clustering was used to create a player cluster and to define distinct player types for each of the three positions on a National Hockey League (NHL) team and then a regression technique was used on the results to help determine a relationship between player types identified in the clustering and overall team performance[1].

K-means clustering was also used to predict the results of NBA playoffs.

The study uses Maximum Entropy Model which requires discretization of feature space to train the model and that's where K-means cluttering comes into play and clustering software package was used here to cluster the data of each feature. Iteration was used multiple times to get closer to the desired results and to train the NBA Maximum Entropy model to make the prediction for the NBA playoffs[2].

5 Bootstrap Sampling

---

[1] Chan, T. C., Cho, J. A., & Novati, D. C. (2012). Quantifying the Contribution of NHL Player Types to Team Performance. Interfaces, 42(2), 131-145. doi:10.1287/inte.1110.0612

[2] Cheng, Ge, et al. "Predicting the Outcome of NBA Playoffs Based on the Maximum Entropy Principle." Entropy 18.12 (2016): 450.

[1] Felsen, Panna, and Patrick Lucey. "“Body Shots”: Analyzing Shooting Styles in the NBA using Body Pose."

[2] Patel, Ankit K., et al. "Identifying fast bowlers likely to play test cricket based on age-group performances." International Journal of Sports Science & Coaching 12.3 (2017): 328-338. APA

[3] Chimka, Justin R., and Thomas P. Talafuse. "Poisson regression analysis of additional strokes assessed at golf." International journal of Sports Science & Coaching 11.4 (2016): 619-622.

[4] Felsen, Panna, and Patrick Lucey. "“Body Shots”: Analyzing Shooting Styles in the NBA using Body Pose."

---

[1] Patel, Ankit K., et al. "Identifying fast bowlers likely to play test cricket based on age-group performances." International Journal of Sports Science & Coaching 12.3 (2017): 328-338. APA

[1] Macdonald, B. (2012). Adjusted Plus-Minus for NHL Players using Ridge Regression with Goals, Shots, Fenwick, and Corsi. Journal of Quantitative Analysis in Sports, 8(3). doi:10.1515/1559-0410.1447

[2] Chan, T. C., Cho, J. A., & Novati, D. C. (2012). Quantifying the Contribution of NHL Player Types to Team Performance. Interfaces, 42(2), 131-145. doi:10.1287/inte.1110.0612

[3] Felsen, Panna, and Patrick Lucey. "“Body Shots”: Analyzing Shooting Styles in the NBA using Body Pose."