# Mathematical model approach for draft picking in basketball

Lawrence Thanakumar Rajappa
IDA Linköping University
lawra776@student.liu.se

## 1 INTRODUCTION

Analytics is being used in all fields such as healthcare, manufacturing, banking and etc. for decision-making. Likewise, analytics is also playing a major role in Sports industry such as football, baseball, basketball and etc. to predict player's next move, injury analysis, position analysis and etc. Sports analytics is being spoken as a concept for many years which could be used to improve team performance, as a result, the revenue generation is very much improved for the team. For this project, we will mostly focus on basketball. In order to have a good prediction and analysis report, we need to have proper dataset and most importantly, the data must also have most important attributes that could provide insights for a given problem.

There are many ways that the data can be used by the team for various purposes. The kind of data that the team would use includes average stats of the players, per game stats, and etc. These data can be used to understand a player in terms of strengths and weaknesses, emotional stability and etc. These attributes could be used to assess the team's performance. Moreover, there are other attributes such as weather conditions, the condition of the field and even psychological factors such as the fans support should be included along with player's data to determine the team's performance. This document speaks about how the data are being used by a basketball team to select players using mathematical models.

## 2 AIM

In this project, the aim is to create a draft picking system for a basketball team by using 3 mathematical models namely, model 1 : model to predict whether a player will stay in the team for five years or not, model 2: model to determine the position of players based on their previous experiences and model 3: model to cluster or group players based on previous performances. These models would facilitate team managers and coaches to select players and make best out of them.

## 3 MOTIVATION

Before the advent of analytics, the selection of players or draft formation was done manually which was a time consuming and huge workload. The emergence of analytics and computing resources has paved a new way in recruiting best players based on their previous performances in a short period of time with minimal workload. However, Sports industry has restricted for the complete adoption of analytics into their respective teams because teams spend three-fourth of their revenue for paying salaries to the players and to cover other expenditures. Hence, the teams cannot afford to invest huge sum of money in technology, data and analytical tools [1]. This project would remove the above mentioned bottleneck and facilitate the teams to use analytical tools with much lower cost and at ease.

## 4 EARLIER SYSTEM

Earlier to 2005, the data was collected by a person watching the game using either a notepad and pen or black boards with chalks. This data was prone to human errors. As a result, the analysis carried out on this data and results were misleading. In 2005, two Isareli scientists, Gal Oz and Miky Tamir, created a system called *SportsVU* (see in figure 1) [5] [12]. This system captures the ball movement as well as athletes movement, all these data are combined together for statistical analysis using the statistical algorithms that the company has created [9]. Based on the statistial analysis inference, the players were chosen for a team, but this method was manual. Moreover, other data such as Rebounds, TurnOver and etc were calculated from this system as well as from manually gathered data.

**Figure 1: SportsVU in basketball court.**

## 5 BACKGROUND

In this part, some general terms that are relative to the topic are going to be discussed. It is important to understand them for further reading.

The terms that will be discussed are:

- Draft picking and its process.
- Machine learning.
- Sports Analytics.

### 5.1 Draft picking and its process

NBA draft is an annual event where basketball teams select players from american colleges and from international professional league for their rosters. Once a team selects a player, then the team has right to sign a NBA contract with the player.

In draft picking process, teams select eligible players in turns. There are two rounds in the draft where all 30 teams participate to select a player in turns, meaning every year 60 eligible players are drafted, but teams that did not reach the playoffs in the previous regular season or teams with worst performs selects a player by undergoing a process called *NBA Draft Lottery*. This process determines the selection order of the team or provides an opportunity for the team which wins the lottery to pick the first draft followed by other worst performing teams. The team with best records receives the 30th pick. During the second round in the draft, there is no lottery system, but teams pick the draft in the reverse order based on the previous regular season's standings. Moreover, the teams can exchange their draft picks with each other, for example, in 2019 the Minnesota Timberwolves traded the No. 11 pick and forward Dario Saric to the Phoenix Suns in exchange for the No. 6 pick. But, there are some restrictions based on *The Stepien Rule*, that is, this rule prevents the team from trading their first-round draft pick in consecutive years [3].

## 5.2 Machine learning

Machine learning is a general concept and broader area which consists of many definitions provided by recognized and reliable universities, institutions, professors and organizations and they are as follows,

- "Machine learning is based on algorithms that can learn from data without relying on rules-based programming." [8]
- "The field of Machine Learning seeks to answer the question “How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?" [6]
- and etc.

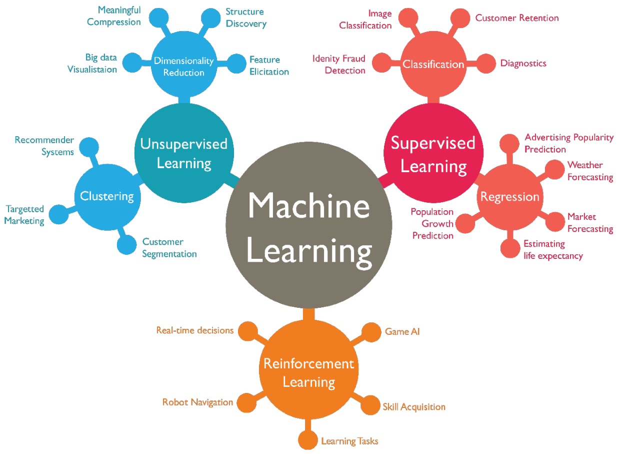

**Figure 2: Machine Learning in an eagle view**

Machine learning can be categorized in three types,

- Supervised Learning.
- Unsupervised Learning.
- Reinforcement Learning.

The definitions for the above terms are:

- "Supervised learning algorithms generate a function that maps inputs to desired outputs, based on a set of examples with known output (labeled examples)" [11].
- "Unsupervised learning algorithms find patterns and relationships over a given set of inputs (unlabeled examples)" [11].
- "Reinforcement learning, where an algorithm learns a policy of how to act given an observation of the world" [11].

In this project, we will mostly focus on Supervised and Unsupervised learning algorithms. The different types of algorithms in both supervised and unsupervised learning are given below,

Some algorithms of supervised learning:

- Nearest Neighbor
- Naive Bayes
- Support Vector Machine (SVM)
- Logistic Regression
- Linear Regression
- and etc.

Some algorithms of unsupervised learning:

- k-means clustering
- Association Rules [2]
- and etc.

## 5.3 Sports Analytics

Sports analytics is the application of above mentioned algorithms to sport in order to draw useful insights which could help an individual athlete's performance, or a team's performance for a season. It can also help teams to perform injury analysis and steps to mitigate them, salary of a player based on his previous performances and etc. Nowadays, many teams, coaches and even players are adopting sports analytics for decision making.

"The analytics split nicely between the front-office and back-office. Front-office analytics include topics like analyzing fan behavior, ranging from predictive models for season ticket renewals and regular ticket sales, to scoring tweets by fans regarding the team, athletes, coaches, and owners. This is very similar to traditional customer relationship management. Financial analysis is also a key area, especially for the pros where salary caps or scholarship limits are part of the equation. Back-office uses include analysis of both individual athletes as well as team play. For individual players, there is a focus on recruitment models and scouting analytics, analytics for strength and fitness as well as development, and predictive models for avoiding overtraining and injuries. Concussion research is a hot field. Team analytics include strategies and tactics, competitive assessments, and optimal roster choices under various onfield or on-court situations.” [10]

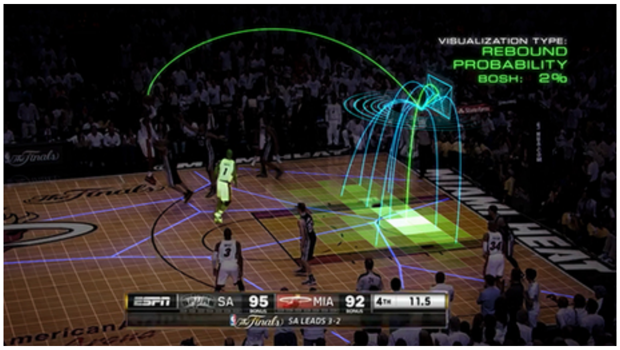

**Figure 3: Sports Analytics**

However, the analytical methods and data has to be kept safe and should be extremely careful because the data and methodology could lead to numerous problems such as issues with betting companies, non-ethical training of atheletes leading to injuries and etc. [7]

## 6 METHODOLOGY

This section describes the framework used for draft picking in basketball using Machine learning. The proposed framework consists of 3 mathematical models namely career longevity model, providing positions to selected players and grouping players based on their performance into two groups, either 1 or 0. Initially, the user or the coach or the team manager loads the players data to **model 1**, which provides list of players who will stay with the league for more than 5 years and list of players who don't stay for more than 5 years. This output is provided as an input to **model 2** which predicts the positions of the players based on previous performance. Finally, the output of $2^{nd}$ model is provided as an input to the **model 3** where players are grouped as 1 and 0, that is, either good or bad pick. This proposed framework is time efficiency, environmental friendly (less paperwork) and user friendly.

Subsection **6.1** describes the libraries or functions used in the proposed framework. Subsection **6.2** descibes about the source of data, structure of data, and pre-processing of data. Subsection **6.3** describe both supervised and unsupervised learning algorithms which are used to create 3 mathematical models for draft picking. The framework applies many algorithms and chooses one algorithm which performs better than the other algorithms for each model. This section concludes with Subsection **6.4**, which describes the evaluation metrics used.

### 6.1 Libraries

The following Python programming packages or libraries have been used by the proposed framework for data pre-processing and model building,

- NumPy for numerical computations.
- Pandas for data loading.
- MatplotLib and Seaborn for visualization.
- sklearn for evaluation metrics and model building (Supervised and Unsupervised learning).
- Pickle for loading and saving of data objects, mathematical model objects and etc.
- statsmodel for checking multicollinearity in data.

### 6.2 Dataset

*6.2.1 Source of data.*
The two main sources for dataset used by the proposed framework are,

- basketball-reference.com
- Data.world

The data provided by Basketball-reference.com contains the NBA players overall performance such as rebounds, turnovers, points, games played and etc. for a regular season which can be either web scrapped or downloaded in comma-separated format file (CSV). Data.world provides NBA players data such as players demographic data, players salary data and etc. which are less in number. By combining data from both sources, yields a large dataset required for model building.

*6.2.2 Structure of data.*
The dataset contains players performance with the following main attributes,

| Attributes | Descriptions |
|---|---|
| GP | Games Played |
| POS | Position |
| MP | Minutes Played |
| PTS | Points |
| FG | Field Goals (both 2 point field goals and 3 point field goals) |
| FGA | Field Goal Attempts (includes both 2-point field goal attempts and 3-point field goal attempts) |
| FG% | Field Goal Percentage; the formula is FG / FGA. |
| 3P | 3-Point Field Goals (available since the 1979-80 season in the NBA) |
| 3PA | 3-Point Field Goal Attempts (available since the 1979-80 season in the NBA) |
| 3P% | 3-Point Field Goal Percentage (available since the 1979-80 season in the NBA); the formula is 3P / 3PA. |
| FT | Free Throws |
| FTA | Free Throw Attempts |
| FT% | Free Throw Percentage; the formula is FT / FTA |
| ORB | Offensive Rebounds |
| DRB | Defensive Rebounds |
| TRB | Total Rebounds |
| AST | Assist |
| AST% | Assist percentage<br>the formula is 100 * AST / (((MP / (Tm MP / 5)) * Tm FG) - FG). Assist percentage is an estimate of the percentage of teammate field goals a player assisted while he was on the floor. |
| STL | Steal |
| BLK | Block |
| BLK% | Block Percentage<br>the formula is 100 * (BLK * (Tm MP / 5)) / (MP * (Opp FGA - Opp 3PA))<br>Block percentage is an estimate of the percentage of opponent two-point field goal attempts blocked by the player while he was on the floor |
| TOV | Turnover |
| TOV% | Turnover percentage<br>the formula is 100 * TOV / (FGA + 0.44 * FTA + TOV)<br>Turnover percentage is an estimate of turnovers per 100 plays. |
|  | and etc. |

**Table 1: Data attributes and its description**

There are other attributes such as VORP, WS, WS/48 and etc. which can be known further from the following page **basketball-reference.com** glossary. These attributes contains both continuous and categorical data type.

### 6.2.3 *Pre-processing.*

The pre-processing process consists of following steps,

- Missing value imputation - Missing values in the data are imputed by taking median of a column.
- Converting target variable or variable to be predicted to category data type.
- Filtering data using a constraint for getting useful insights from data.
- In case of regression, highly correlated variables with each other are removed using correlation plot to avoid multi-collinearity problem.
- Sampling, if the classes (response variable data) are imbalanced.
- Checking for linear relationship for best separaion in case of classification and best line fitting in case of regression.
- Scaling, that is, normalizing the data to have normal distribution.

## 6.3 Machine learning algorithms

This section describes about machine learning algorithms that were chosen for each model and reasons for chosing them.

### 6.3.1 *Model 1 and Model 2.*

The model 1 performs predicting whether a players will stay in the league for more than 5 years or not. The dataset used for this task is taken from *data.world* which contains 21 attributes where 20 attributes are independent variables and 1 attribute is response variable and has 1340 rows. After performing the above mentioned pre-processing steps, the separability between independent variables and dependent variable was found using scatter plot between games played and points earned coloured by response variable, see figure 4.

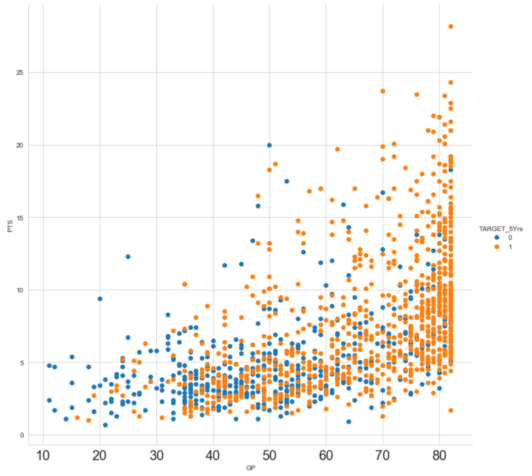

**Figure 4: Model 1 scatter plot between independent and dependent variables**

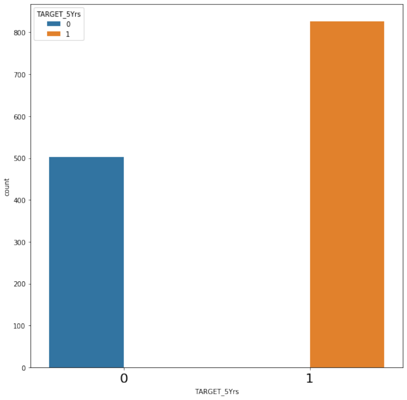

**Figure 5: Model 1 count plot for classes in response variable**

The model 2 performs predicting position in the basketball court based on the player's previous performance in the earlier regular seasons. The dataset used for this task is taken from *basketball-reference.com* which consists of 47 attributes where 46 attributes are independent variable and 1 variable is response variable and has over 10,000 rows of data. After performing the above mentioned pre-processing steps, the separability between independent variables and dependent variable was found using scatter plot between games played and points earned coloured by response variable, see figure 6.

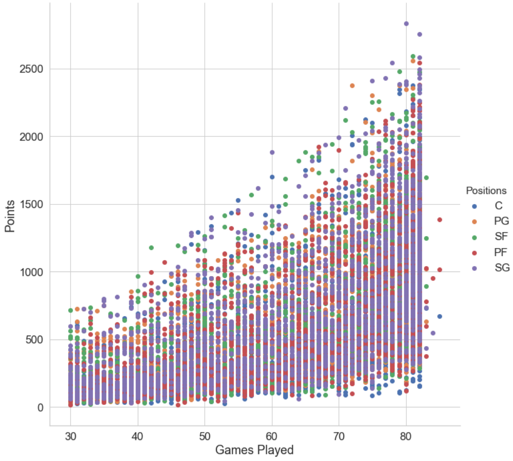

**Figure 6: Model 2 scatter plot between independent and dependent variables**

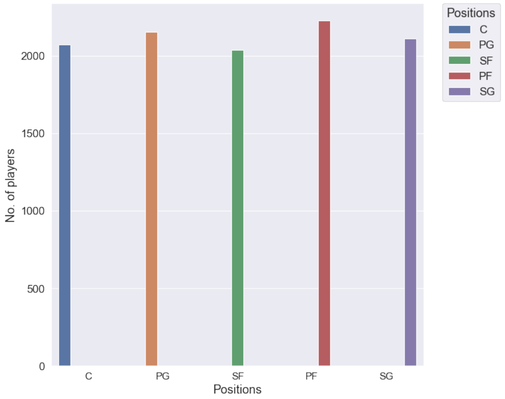

**Figure 7: Model 2 count plot for classes in response variable**

From the scatter plots, it can be observed that the data is not linearly separable. Moreover, it is not possible to create a classification boundary to separate data points that belong to differenct classes of the response variable. Hence, non-linear classification algorithms such as Support Vector Machine (SVM), K-Nearest Neighbor, and etc. has to be applied. Given the non-linear separability constraint and imbalanced dataset (model 1), see figure 5, it is best to choose Support Vector Machine (SVM) algorithm for both model 1 and model 2 because they train very well and yield good prediction results than other supervised learning algorithms [4]. Before getting into detailed execution and discussion of test results, let us first understand about Support Vector Machine (SVM) algorithm.

# REFERENCES

[1] Thomas H Davenport. 2014. Analytics in sports: The new science of winning. *International Institute for Analytics* 2 (2014), 1–28.

[2] David Fumo. [n.d.]. *Types of Machine Learning Algorithms You Should Know.* https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861 Accessed: 2020-12-24.

[3] Jack Green. [n.d.]. *How does the NBA draft work?* https://blog.betway.com/basketball/how-does-the-nba-draft-work-nba-draft-explained/ Accessed: 2020-12-24.

[4] Tasadduq Imam, Kai Ming Ting, and Joarder Kamruzzaman. 2006. z-SVM: An SVM for improved classification of imbalanced data. In *Australasian Joint Conference on Artificial Intelligence*. Springer, 264–273.

[5] Zach McCann. 2012. Player tracking transforming NBA analytics. *Tech-ESPN Playbook, May* 19 (2012).

[6] Tom Michael Mitchell. 2006. *The discipline of machine learning.* Vol. 9. Carnegie Mellon University, School of Computer Science, Machine Learning ....

[7] Munni. [n.d.]. *Sports analytics have changed the way sports are played.* https://tanukamandal.com/2017/12/12/sports-analytics-changed-play/ Accessed: 2020-12-25.

[8] Dorian Pyle and Cristina San José. [n.d.]. *An executive's guide to machine learning.* https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/an-executives-guide-to-machine-learning Accessed: 2020-12-24.

[9] Rusty Simmons. [n.d.]. *Competitive fire helps Kirk Lacob make his own name with Warriors.* https://www.sfgate.com/warriors/article/Competitive-fire-helps-Kirk-Lacob-make-his-own-6339796.php Accessed: 2020-12-23.

[10] Walter Tichy. 2016. Changing the Game: Dr. Dave Schrader on sports analytics. *Ubiquity* 2016, May (2016), 1–10.

[11] George Tzanis, Christos Berberidis, and Ioannis Vlahavas. 2009. Machine learning and data mining in bioinformatics. In *Handbook of research on innovations in database technologies and applications: Current and future trends*. IGI Global, 612–621.

[12] wsbc.uoregon.edu. 2015. Industry Insights: The Evolution of Basketball through Technology | Warsaw Sports Business Club.