

**Project Proposal Write Up**

**Predicting shot success in the NBA using recurrent neural networks**

By: Thomas Little and Joseph Walbeck

## Overview

The game of basketball is centred around teamwork, where the primary goal of a team is to outscore their opponents by the conclusion of the game. Each team may field a maximum of 5 players on the court at any given time, and there is no limit on the number of rolling substitutions that may occur during a game. The objective of a basketball game, which consists of four quarters each of twelve minutes in length, is to score more points than the opposition. Players score by successfully shooting the basketball through the opposition's hoop, and depending on the circumstances, may either earn one, two, or three points. In the case that two teams are tied in points at the end of a game, five-minute overtime periods are played up until there is a winner by the end of an overtime period. A shot clock exists to limit the amount of time a team may possess the ball before attempting to score a field goal and therefore affects shot decisions in basketball. In the NBA a shot clock violation forcing a team to give up possession is called if a shot is not attempted within 24 seconds of a given possession. The value of a basketball shot is decided depending on where the shot was taken and whether the shooter is fouled. The following figures detail how the points of shot are decided.

**Figure 1**

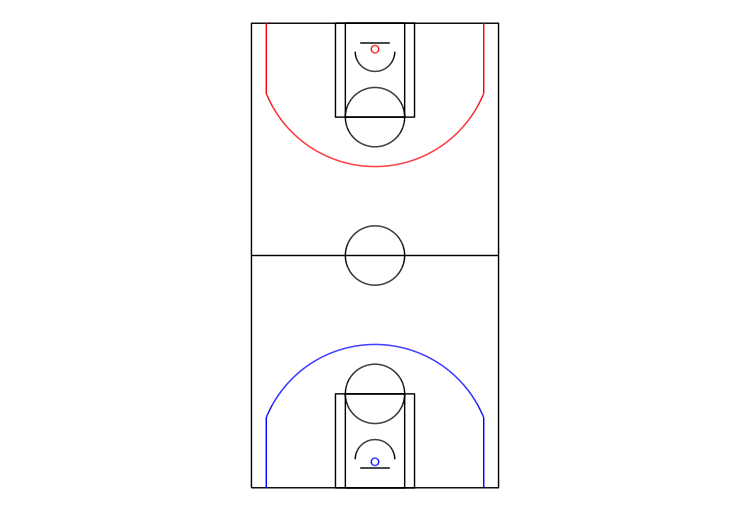


Figure 1 illustrates an ordinary basketball court, 15.2 by 28.7 metres in dimension. The blue and red circles represent the baskets of each respective team. The objective of the red team is to score in the blue team's basket, and vice versa. Any shot made within the opposing team's arc, outlined in colour, is awarded 2 points. On the other hand, any shot made outside of the opposing team's arc is awarded 3 points.

**Figure 2**

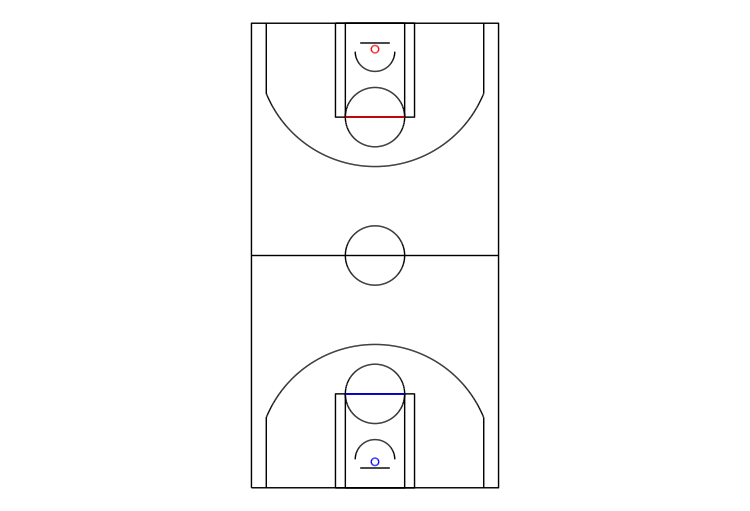


Figure 2 depicts the exception. If a player is fouled by an opposition player while attempting a shot, a penalty shot, termed a “free throw” is awarded. The number of free throws awarded is dependent on the location of the foul, fouls that occur inside the arc are awarded two, and fouls outside the arc are awarded three. Positionally, a free throw is taken behind the coloured horizontal lines depicted in figure 2 and is uncontested by any opposing players. A single point is awarded for a successfully made free throw.

Given the scoring-oriented nature of the sport, basketball necessitates intricate player movement and complex decision-making. As such, deriving success as a team is heavily reliant on the creation of high-percentage scoring chances. Shot success is dependent on several sequential events that occur during the build-up to a shot, most of which are linked to the spatial position of the players on the court. Previous research into the topic was conducted by Harmon et al (2021) predicting shot-making in the NBA. They combined a convolutional neural network and a feedforward neural network to make their predictions. They fed a multichannel image of the offensive players, defensive players, and the basketball’s position in the five seconds preceding a shot attempt into the convolutional neural network to generate predictions. For our research, we have retrieved both the spatial and event data from 277 games during the 2015-2016 NBA season. Given the sequential nature of this data, we aim to build on the research of Harmon et al (2021) by constructing a deep-learning model to predict shot-making in the NBA using recurrent neural networks (RNN). RNNs are designed in such a way that they can take sequential data as inputs. This makes RNNs suited to predicting shot success as sequential locational data will be inputted into the model. The nodes in the RNN can connect with each other and create a cycle that makes them able to have internal memory. This internal memory allows sequential data to be taken as inputs.

## Objectives

### Primary Objective

Predicting if a shot taken in the NBA will be successful using neural networks given location data of the basketball and all players at and before a shot is taken.

### Secondary Objective

Apply the results and model of the primary objective to evaluate which players and teams make the best shot-taking decisions. This creates a new metric that allows an intangible attribute of basketball players - their decision-making - to be evaluated.

## Data

The data used in this thesis was sourced from a GitHub repository that contains copies of the datasets when they were open to the public for free. It comprises two separate datasets - one containing location data of all the players and ball and one detailing events. These two datasets will be merged to be used to predict shot-making. The datasets contain 277 NBA games that took place in December 2015, January 2016, and February 2016.

### Location Data

The data was collected through SportVU by Stats Perform (formerly STATS LLC). The data contains x and y coordinates for the 10 players on the court. X, Y, and Z coordinates are provided for the ball. The data specifies which player the coordinates are for and which team the player is part of (see Appendix A). SportVU is a tracking system that uses six cameras to track the location of the players and the ball (Harmon et al., 2021). The data is recorded at 25 hertz. The data is originally in a JSON file but is converted to an RData to be used in R.

### Events Data

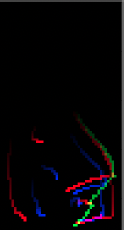
This events data provides us with details surrounding the play-by-play events of interest per NBA game on record. This data consists of a timestamped record of the precise time each event of interest occurs within a game. The events data has been collected in a CSV file and the events it records include but are not limited to:

* Successful shot attempt
* Unsuccessful shot attempt
* Miss shot
* Free throw result
* Rebound
* Turnover
* Foul
* Violation
* Substitution
* Timeout
* Jump Ball

## Methodology

To build a baseline we create three base models. The first base model is a logistic regression model. This model is the simplest model we will consider. To begin with, only the spatial coordinates of the players and the ball at the time of the event of a shot attempt will be fed into the logistic regression model. This will allow the model to be easily interpretable as using location data only at the time of the shot restricts the number of variables considered. For the first iteration of this model, 22 variables will be considered - the x and y coordinates of the player and the x and y coordinates of the basketball at the time of the attempted shot. The second base model that will be considered is a convolutional neural network. The same 22 variables as the logistic model - the x and y coordinates of the player and the x and y coordinates of the basketball at the time of the attempted shot – will initially be considered. Using our exploratory data analysis and literature review as a guide, we will create features from the data, which will be fed into the logistic and convolutional neural networks. These two baseline models will allow us to compare how much the addition of the temporal aspect of the data – the location data through time – affects the predictions.

In our third base model, we aim to replicate Harmon et al’s model. In their best model, they use a multiagent image containing x and y coordinates for the basketball and all the players on the court for the 5 seconds preceding the shot attempt (Harmon et al 2021). These images are fed into a convolutional neural network and is combined with a feed-forward model that has features from the data as its input, to predict whether the shots are successful (see Appendix A).

**Figure 3**

In figure 3 we see an example of the image that is created. The red lines represent the trajectories of the offensive players while the blue lines represent the trajectories of the defensive players. The green line represents the trajectory of the basketball. As the observations of the location move forward in time the pixels of the trajectories become brighter.

A recurrent neural network (RNN) will be the model of interest in our paper. RNNs have inherent internal memory due to their interconnecting nodes. This makes them suitable to be used on sequential data. A many-to-one network will be used as we will feed a sequence of location data into the model for every event, and a single-shot outcome will be predicted. Wang and Zemel (2016) experienced the vanishing gradient problem while developing their RNN to classify plays in the NBA and therefore we expect the vanishing gradient problem might occur in our RNN. If we encounter this, we plan to use a long short-term memory neural network to overcome the problem.

## Division of Labour and Timeline

**Figure 4**

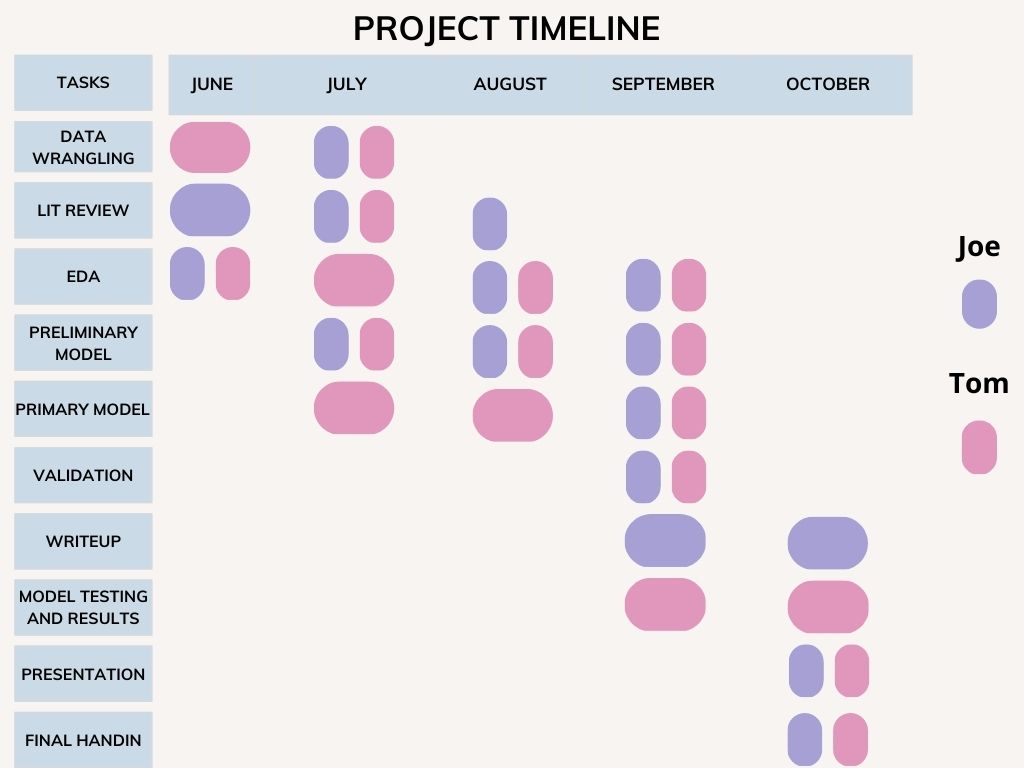


Figure 3 depicts the schedule to be followed to ensure all deadlines and deliverables are adequately met within sufficient time. Furthermore, the division of labour between the two researchers has been captured.

## Closing Remarks

The utility of advanced machine learning algorithms has a wide range of applicability across multiple industries, not the least of which is sports. Tools used to analyse strategy are becoming increasingly more sophisticated, and the landscape of data analytics is ever-changing. Hence, the objective of our research is to spearhead innovation in this regard specifically as it pertains to the sport of basketball. In constructing a statistically accurate and robust model we believe that valuable statistical insight could be offered to owners of basketball teams looking to utilise data-driven analyses to improve the likelihood of their teams’ success.

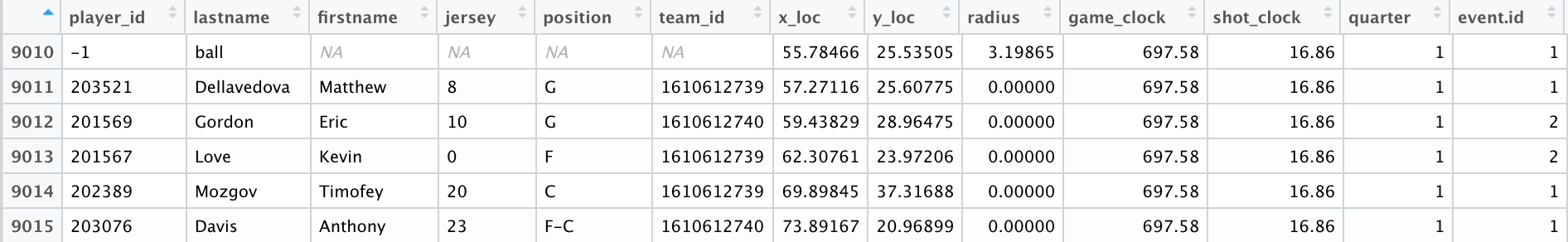
## Reference List

Harmon, M., Ebrahimi, A., Lucey, P. & Klabjan, D. 2021. Predicting Shot Making in Basketball Learnt from Adversarial Multiagent Trajectories. *arXiv:1609.04849 [cs, stat]*. (January, 15). Available: <https://arxiv.org/abs/1609.04849> [2023, April 16].

Wang, K.-C. & Zemel, R. 2016. *Classifying NBA Offensive Plays Using Neural Networks*. MIT Sloan Sports Analytics Conference.

## Appendices

*Appendix A – Screenshot of observations from the SportVU dataset*



*Appendix B – List of features created by Harmon et al (2021) for their feed-forward neural network*

* *Player and ball positions at the time of the shot*
* *Game time and quarter time left on the clock*
* *Player speeds over five seconds*
* *Speed of the ball*
* *Distances and angles (with respect to the hoop) between players*
* *Number of defenders in front of the shooter (300 angle of the shooter) and within six feet based upon the angles calculated between players*
* *Ball possession time for each offensive player*
* *Number of all individuals near the shooter (including teammates)*