**Optimizing the Best Play in Basketball using Deep Learning**

Basketball Track

Paper ID: 1548820

1. **Abstract**

In a close game of basketball, victory or defeat can depend on a last second shot. Being able to identify the best player and play scenario for a given opponent’s defense, can increase the likelihood of victory. Progress in technology has resulted in an increase in the popularity of sports analytics over the last two decades, where data can be used by teams and individuals to their advantage. A popular data analytic technique in sports is deep learning. Deep learning is a branch of machine learning that finds patterns within big data and can predict future decisions. The process relies on a raw dataset for training purposes. It can be utilized in sports by using deep learning to read the data and provide a better understanding of where players can be the most successful.

In this study we collected data from basketball games of a college women’s team and applied deep learning to optimize the best offensive play in a game scenario for given a set of features. The system is used to predict the play that would lead to the highest probability of a made shot.

1. **Introduction**

The increasing interest in sports analytics over the last two decades can be attributed to advances in technology, where data has been used by teams and individuals to gain a competitive advantage (1). Statistics have always played a role in sports but the use of predictive analysis has increase in recent years. The volume of data collected for each game, makes it a big data problem since it is not readily feasible to gain meaningful insights from raw data. Data driven decision making is being incorporated in different aspects of sports from gambling, fantasy leagues, to improving team dynamics and performance, decision making, preventing injuries, etc. Deep learning and machine learning techniques are critical techniques since data is unstructured and lacks context.

Deep learning is a branch of machine learning that identifies patterns within big data and can be used to predict future decisions. The process relies on a raw dataset for training purposes (3). Deep learning can be used to give teams better insights on which players to select, predict an opponent’s actions at each time, determine how to train players and prevent injuries, and provide management a better understanding on ways to enhance revenue and fan engagement.

In basketball, offensive possessions are arguably one of the most valued components of the game. In a close game, the last few possessions can usually determine if a team wins or loses (4). The last play of a game can completely alter the outlook of how a team played and/or impact a fan’s perceptions about a game (5). Turnovers often limit a team’s chance to score and gives the opposing team extra opportunities. Missed shots in a close game can be just as costly as turnovers. A successful possession results in a player taking a reasonably high percentage shot (6).

As a global sport, research on aspects of the game of basketball are valuable in its improvement. Traditional measurement of game performance has been based on statistics that are collected and displayed in a box score (7). More recently, data analytics in basketball have used machine learning to predict the outcome of an NBA game using Naive Bayes and Artificial Neural Networks (8), or visualization of made and missed shots via shot charts within a game (9). Shot charts demonstrate where a team has made a shot with a circle and missed a shot with an X with respect to the location of a shot. These charts can be useful in determining where most of the shots have been made for each team, however they do not provide additional insights beyond a location within the court. To extend deeper using data analytics, more information about the shot needs to be integrated, rather than location alone. In another study to help evaluate shooting within the NBA, Effective Shot Quality (ESQ) was created to help improve Effective Field Goal percentage (EGF) (10). ESQ takes into account factors such as the angle of the defender to the shooter, defender distance, shot angle etc.

The goal of this study is to develop a framework to predict the best offensive play based on a set of defensive features. Our deep learning model predicts the offensive features along with the best player to take the shot. The Expected Possession Value (EPV) of each player is used to identify the players to perform a selected play that will result in the greatest probability of a made basket.

The type of questions that our framework is able to answer are:

* During an offensive possession, given an opponent’s defensive scheme, what is the best play (i.e. the greatest probability of a made basket)?
* Who is the best player to take a shot?
* What is the likelihood of a made shot at different locations by different players?
* Assume 5 seconds remains on the shot clock. Timeout is requested. The developed system is able to predict: what is the best play? Who should take the shot? Who are the other players that should be in the game based on the opponent’s defensive scheme?

1. **Data**

The data was collected over the 2018-2019 season of a college women’s basketball team. The feature vector consists of a set of attributes: Player, Play, Defense type, Defender position, Screen, Quarter, Seconds on the shot clock, Number of defenders, Location, Hand, Shot type, Passes in half court, Time left in the quarter, and Difference in points. Table 1 lists the features used in our model.

Table 1 Feature description

|  |  |
| --- | --- |
| **Attributes** | **Description** |
| *Player* | Player who takes the shot |
| *Play* | The play that was run to get a player a shot |
| *Defense Type* | Whether the opponent’s defense is in a zone or man-to-man |
| *Defender Position* | The location of defender. |
| *Screen* | if a screen was used to get the player an open shot |
| *Quarter* | The quarter the shot was taken in |
| *Seconds on Shot Clock* | Number of seconds left on the shot clock |
| *Number of Defenders* | the number of defenders in the half court at the time of the shot |
| *Location* | The location shot was taken (out of 11 sports) |
| *Hand* | right or left |
| *Shot type* | Labeled as lay-up, dribble jumper, spot up, turn-around jumper (TAJ), floater, step back, or spin shot. |
| *Passes in half court* | Number of passes prior to the shot |
| *Mins left in quarter* | Minutes remaining in the quarter |
| *Result* | make or miss |

Based on these attributes the model is trained to predict the ‘Make’ or ‘Miss’ of a shot in a given game situation. Data preprocessing consists of cleaning and reduction. In the data cleaning step, we deal with inconsistent, noisy and missing data. After data cleaning, attribute correlation is used to find the most important attributes (figure 1).

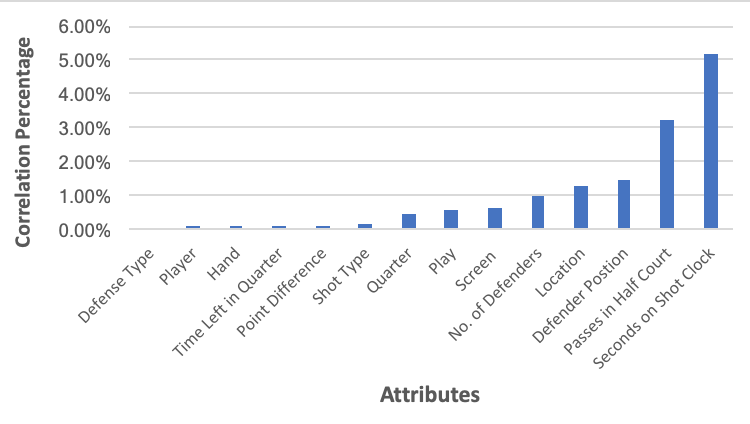


Figure 1. Attribute correlation to the result of a shot

1. **Methodology**

This paper introduces a new method for the field of basketball analytics by identifying the best game plan at time *t* given a specific game scenario. Our model is used to predict the best shooter, offensive play, and determining the top 4 players to support the play. The top 4 players is selected based on the maximum expected number of points given the location of the players.

* 1. **Deep learning – Who is the shooter and what is the play*?***

The goal of this study is to predict the best shooter and best play in a basketball game given the opponent’s defensive scheme.

There are numerous statistics tracked in a basketball game. One of the most significant is Field Goal Percentage. In many instances, someone with the highest Field Goal Percentage or the most Field Goals Made on the team is the person most likely to take a game winning shot in a close game. In the NBA, the average 3 point Field Goal Percentage has been 35% for the past 20 years (11). NCAA Division 1 women’s basketball had an average 3 point percentage around 31% and an average 2 point percentage of 40%. Previous research found that a player who made a shot the possession before is more likely to take the next shot (12). In Figure 2, the shot chart on the left shows the traditional probability distribution density that describes the likelihood of players making their shots for a given location on the court. The shot chart on the right (Figure 2) shows the optimized model which considers the correlation among the dataset and is calculated based on constant defensive attribute values.

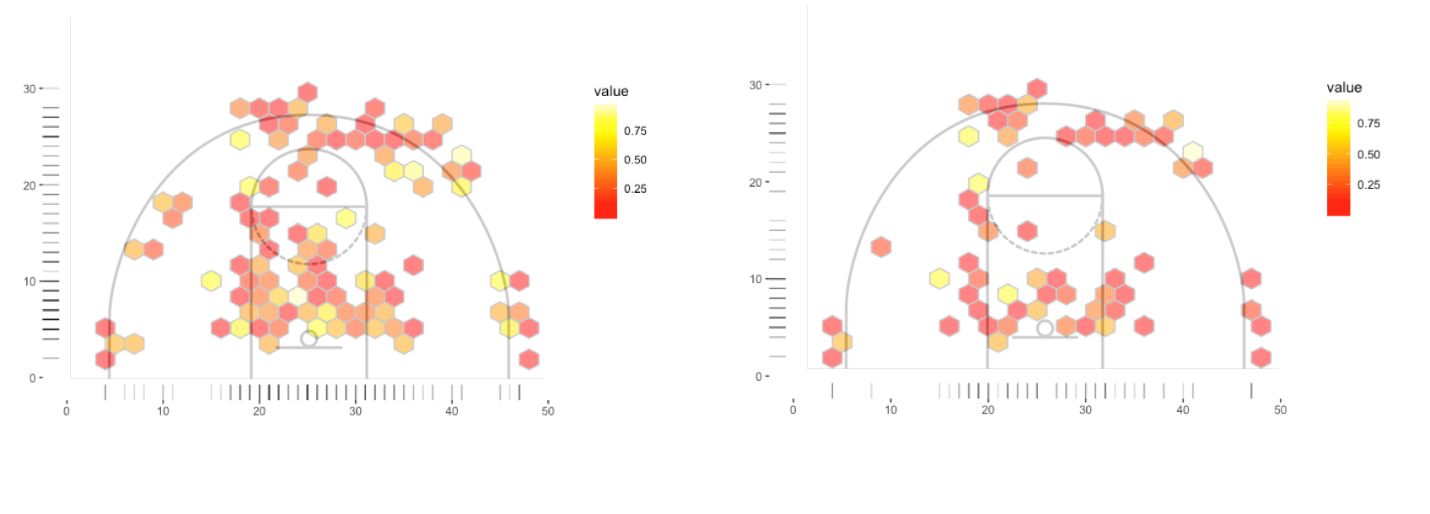
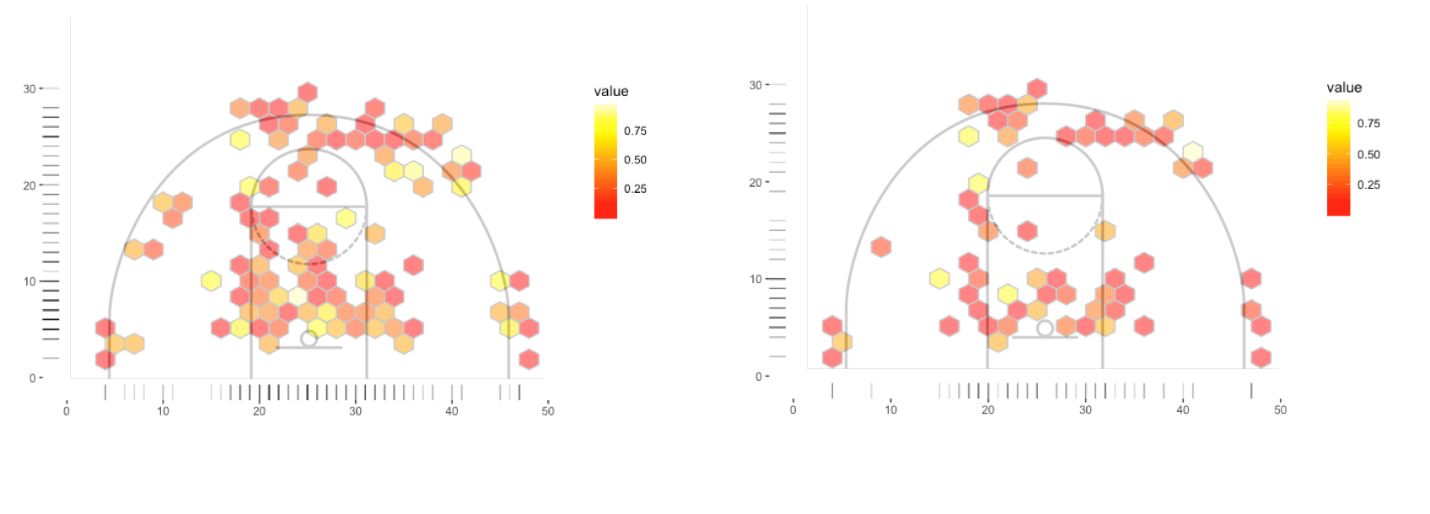


Figure 2. Graphical depiction of an offensive likelihood of making shot; Likelihood of making a shot based on the constraints on the model’s defensive

Previously, the best shooters were predicted by data scientists using a factorization machine model (13) and Adversarial Multiagent Trajectories (14). Deep learning can be applied to shots in basketball to predict the best shooter and the optimal play to run in specific game situations. With the use of these predictions, coaches can learn how to optimize each possession within a game, by choosing the best approach to taking a shot.

After generating the feature vector for each shot, the data was imported to RapidMiner, a data science software platform for machine learning and data mining experiments. The deep learning model was used to predict the best play based on our defined feature vector. An overview of the methodology is presented in Figure 3.

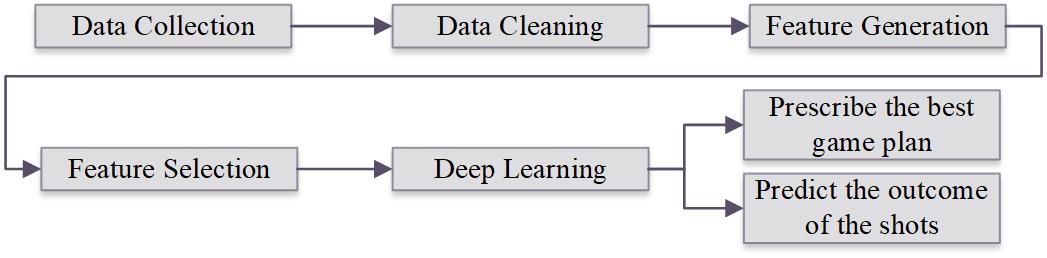


Figure 3. Proposed framework for predicting the best play using deep learning

The deep learning model predicts the actual result of the shot with an accuracy of 64.4%. The *f*-measure of the model is 73.78%.

* + 1. **Deep Learning Results**

Deep learning uses the imported data as a training tool to predict future outcomes of shots. The process of deep learning is as follows: first, the data is retrieved, then the data is processed where the target column is identified. After this, if the data has any missing values, those will be replaced with the average of the column. However, in our current model all data was cleaned prior to importing it into RapidMiner, so no missing values were included in the input. The data is then filtered and sent to deep learning to train the model, multiply training to copy the data, and multiply validation to copy validation data. Finally, the data is sent to the model simulator, where it can be applied and the result can be predicted.



Figure 4. Deep Learning Framework

Table 2 and Figure 5 show a sample feature vector as well as the result of deep learning model.

Table 2. Overview of the system

|  |  |
| --- | --- |
| ***Attributes*** | ***Results*** |
| Defense Type: Man  Defender Position: Contested  Passes in Half Court: 2  Quarter: 4  Seconds on Shot Clock: 20  Time left in quarter: 5 minutes  Location: 5  Number of Defenders in half court: 5 | Player: 51  Shot type: Spot up  Location: 8  Play: Picket fence |
| 73% chance of *MAKE*    Figure 5. System result | |

After collecting data and an explanatory data analysis, feature selection and extraction is performed. Using a prediction model, we built a machine learning model to predict the make/miss values based on the other features. We then utilized supervised learning to train the model. We evaluated the accuracy for each pattern recognition model described in Table 4 and we chose a multi-level neural network for learning non-linear relationships.

The model predicts the actual result of the shot with an accuracy of 64.4%. The model predicts the percentage of a shot going in based on the attributes provided. The *f*-measure of the model is 73.78%. This gives a predicted make class precision of 58.33% and predicted miss class precision of 62.5%.

Table 3. F-measure

|  |  |  |  |
| --- | --- | --- | --- |
|  | **True make** | **True miss** | **Class prediction** |
| **Predicted Make** | 7 | 5 | 58.33% |
| **Predicted Miss** | 30 | 50 | 62.50% |
| **Class recall** | 18.92% | 90.91% |  |

As shown in Table 4, the training time for the Deep Learning model is 826 milliseconds. Deep learning is the second most accurate model, behind Random Tree. However, Random Tree requires 4 seconds to create the model; however, 4 seconds is typically too long to create a model in a time sensitive game situation. A team has a limited amount of time during stoppage of play during a timeout and selecting attributes also requires time as well. That is why Deep Learning is likely the best model methods based on accuracy and total time.

Table 4. Different Model Methods with Accuracy and Total Time

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Standard Deviation** | **Total Time** | **Training Time** | | **Scoring Time** | |
| **Naive Bayes** | 56.7% | ± 6.1% | 165 ms | | 5 ms | ~0 ms |
| **Generalized Linear Model** | 59.4% | ± 5.5% | 198 ms | | 147 ms | 8 ms |
| **Logistic Regression** | 57.8% | ± 8.4% | 211 ms | | 147 ms | 8 ms |
| **Fast Large Margin** | 58.3% | ± 6.8% | 368 ms | | 5 ms | ~0 ms |
| **Deep Learning** | 64.4% | ± 8.4% | 826 ms | | 1 s | 31 ms |
| **Decision Tree** | 61.9% | ± 9.1% | 183 ms | | ~0 ms | ~0 ms |
| **Random Tree** | 66.3% | ± 7.3% | 4 s | | 131 ms | 78 ms |

* 1. **Expected Possession Value – Who should play to support the shooter?**

We also calculated the EPV (Expected Possession Value) of players for each successful shot. EPV reacts to every on-court movement and action and is the expected number of points the offense will score, given the spatial configuration of the players and ball at time during the possession (15).

(1)

To calculate the EPV, tracking game data is used. Each data point represents a time point and consists of values such as: time, quarter, shot clock, game clock, play, player number and player possession location (*x,y*). EPV is calculated using Monte Carlo simulation (15). Table 5 shows the EPV of players up to the made shot described in Table 2. EPV results are used to select the 4 players to be on the floor for a play selected for a specific shooter.

Table 5. EPV of players

|  |  |
| --- | --- |
| Player | EPV |
| 1 | 1.65 |
| 2 | 1.45 |
| 3 | 1.38 |
| 4 | 1.21 |
| 5 | 1.05 |
| 6 | 1.03 |
| 7 | 1.00 |
| 8 | 0.90 |
| 9 | 0.70 |

The results indicate that choosing the top players based on EPV, to perform the selected play increases the likelihood of the shot being made.

1. **Discussion**

The model is designed to be used for predicting the best play to perform in a given game situation. Besides defender position and number of defenders position, location, passes in the half court, and time remaining on the shot clock have a significant impact on the chances of a made shot. In general, the further the shot is away from the basket, the lower the chances of a made shot. Passes in the half court have an important impact on the likelihood of success since most basketball plays involve movement of the ball in order to get the defense to move. Getting the defense to move and shift can help change the spacing between players and create better opportunities for shots.

Time remaining on the shot clock has the largest correlation to a Make or Miss. Time remaining is significant in basketball since it adds pressure to the shooter, causing players to rush and potentially leads to poor shot selection. When time remaining on the shot clock is relatively high, the likelihood of a made shot tends to increase. Whereas, when time remaining on the shot clock is low, the likelihood of a made shot tends to decrease. In the model, when the shot clock is at 28 seconds for the scenario described in Table 2, the chances of a made shot are 52%. When the shot clock is at 6 seconds the chances of a made shot are 13%. This also demonstrates that the model understands the importance of the shot clock in basketball.

Although the model can be used to show the probability of a made shot in terms of percentage, the primary goal of the model is to use it in a game situation where a team is in need of a last second shot. For instance, if a team is behind by three points and have one more possession to take a shot, a coach can turn to this model to identify a play and the shooter with the best chances of making the last shot. The coach is also able to identify the 4 other players to perform the play by selecting the players with the highest real-time EPV based on in-game information as well as a player’s past performance in prior games.

1. **Future Work**

We made a number of assumptions to simplify the model. Further refinement of the model could define a more precise outcome for the coach’s playbook. For instance we didn’t consider transitional movement in calculating the EPV of players, we also didn’t take into account defensive attributes of players on the opposing team. In addition, for future research, it seems desirable to model the conditional probability based on the selected shooter and choose the other players not only based on individual EPVs but also take into account the chemistry of the shooter and other players.

**References**

[1] Goldsberry, K. (2012, March). Courtvision: New visual and spatial analytics for the NBA. In *2012 MIT Sloan sports analytics conference* (Vol. 9, pp. 12-15).

[2] Mondello, M., & Kamke, C. (2014). The introduction and application of sports analytics in professional sport organizations. *Journal of Applied Sport Management*, *6*(2).

[3] Rangel, W., Ugrinowitsch, C., & Lamas, L. (2019). Basketball players' versatility: Assessing the diversity of tactical roles. *International Journal of Sports Science & Coaching*, *14*(4), 552-561

[4] Christmann, J., Akamphuber, M., Müllenbach, A. L., & Güllich, A. (2018). Crunch time in the NBA–The effectiveness of different play types in the endgame of close matches in professional basketball. *International Journal of Sports Science & Coaching*, *13*(6), 1090-1099

[5] Bashuk, M. (2012, March). Using cumulative win probabilities to predict NCAA basketball performance. In *Proceedings of the MIT Sloan Sports Analytics Conference* (pp. 1-10).

[6] Skinner, B. (2012). The problem of shot selection in basketball. *PloS one*, *7*(1), e30776.

[7] Skinner, B., & Guy, S. J. (2015). A method for using player tracking data in basketball to learn player skills and predict team performance. *PloS one*, *10*(9), e0136393.

[8] Cao, C. (2012). Sports data mining technology used in basketball outcome prediction

[9] Reich, B. J., Hodges, J. S., Carlin, B. P., & Reich, A. M. (2006). A spatial analysis of basketball shot chart data. *The American Statistician*, *60*(1), 3-12.

[10] Chang, Y. H., Maheswaran, R., Su, J., Kwok, S., Levy, T., Wexler, A., & Squire, K. (2014, February). Quantifying shot quality in the NBA. In *Proceedings of the 8th Annual MIT Sloan Sports Analytics Conference. MIT, Boston, MA*.

[11] Marty, R., & Lucey, S. (2017). A data-driven method for understanding and increasing 3-point shooting percentage. In *Proceedings of the 2017 MIT Sloan Sports Analytics Conference*.

[12] Attali, Y. (2013). Perceived Hotness Affects Behavior of Basketball Players and Coaches.

[13] Wright, R. E., Silva, J., & Kaynar-Kabul, I. (2016). Shot Recommender System for NBA Coaches.

[14] Harmon, M., Lucey, P., & Klabjan, D. (2016). Predicting Shot Making in Basketball using Convolutional Neural Networks Learnt from Adversarial Multiagent Trajectories.

[15] Cervone, D., D’Amour, A., Bornn, L., & Goldsberry, K. (2014, February). POINTWISE: Predicting points and valuing decisions in real time with NBA optical tracking data. In Proceedings of the 8th MIT Sloan Sports Analytics Conference, Boston, MA, USA (Vol. 28, p. 3).