What factors influence success?: A study based on the NBA

Rebecca Jones

08 February 2021

Abstract

This report explores the different characteristics between successful and unsuccessful shot attempts within the context of the NBA. Characteristics that are individual to the player and external characteristics are considered. There is a conclusion that shooting players do have an impact on these characteristics. The player-controlled variables are explored in a logistic model and it is found that the time that the player directly holds the ball for has the greatest impact on the shot result. This model does have a good performance for predicting missed shots but not successful shot attempts.

Contents

1 Introduction

The report examines the difference between successful ('made') and unsuccessful ('missed') shot attempts. It is structured as follows:

Section 1 provides an overview of the data selected for use with summaries of the made and missed shot attempts provided.

Section 2 considers characteristics that the player has a certain amount of ability to control that could have an impact on the result of the shot attempt. There is a focus on the shot distance, how far the player is from the basket and the shot clock, how long does the player take to attempt the shot. These are then considered in the context of individual players.

Section 3 considers the team that the player plays for, a particular factor that the player does not have full control over but could still impact the result of a shot attempt.

Section 4 proposes a logistic regression model to attempt to predict whether a shot is made or missed depending on only the player-controlled variables. It will be considered which of these variables have the greatest impact on the type of shot attempted.

Section 5 concludes and provides the main results from this report.

Note: The following acronyms will be used: Golden State Warriors (GSW), Sacramento Kings (SAC). There is also an assumption that the significance level for any statistical tests will be set to 95% (alpha = 0.05). This has been chosen because it is a general default for statistical tests.

2 The Dataset – The 2014-15 National Basketball Association (NBA) Season

The National Basketball Association (NBA) is a popular basketball league primarily based in the USA (NBA, 2019).

The focus of this study is in the context of the 2014-2015 NBA season. The data contains metrics on every attempted shot during the regular season of the 2014-15 season.

After the regular season, there is a section called 'Playoffs' where the best teams in each group by geographical location (or 'Conferences') play against each other to determine the season winner (NBA, 2019). This season, won by the Golden State Warriors (GSW) was particularly popular as approximately 20 million people on average watched during the Playoffs (Brown, 2015).

There are 20 columns in this dataset with a significant number of metrics devoted to the shooting player including the number of times the player bounced the ball on the ground before making the shot (DRIBBLES), the distance from the basket (SHOT_DIST) and the time taken to attempt the shot (SHOT_CLOCK). There are also columns based on other factors, whether the shooting player was playing at the court their team was based at ('Home') or at the opposing teams court ('Away') and which team the shooting player is a member of.

Studies have been made into what contributes to success in the NBA. Zhang et al. (2017) considered different variables from the 2015-16 season but found that the team which the player plays for does have an impact on performance but that the location of the game does not have an impact on success.

Table 1: Summary of made shots

DRIBBLES	SHOT_DIST	SHOT_CLOCK
Min. : 0.000	Min. : 0.10	Min. : 0.00
1st Qu.: 0.000	1st Qu.: 3.60	1st Qu.: 8.60
Median : 0.000	Median : 8.65	Median :12.80
Mean : 1.853	Mean :11.63	Mean :12.96
3rd Qu.: 2.000	3rd Qu.:20.40	3rd Qu.:17.40
Max. :27.000	Max. :29.40	Max. :24.00

Table 2: Summary of missed shots

DRIBBLES	SHOT_DIST	SHOT_CLOCK
Min. : 0.000	Min. : 0.10	Min. : 0.00
1st Qu.: 0.000	1st Qu.: 6.30	1st Qu.: 7.90
Median : 1.000	Median :16.90	Median :11.90
Mean : 2.089	Mean :14.99	Mean :11.97
3rd Qu.: 3.000	3rd Qu.:23.10	3rd Qu.:15.90
Max. :32.000	Max. :39.90	Max. :24.00

3 Section I: Differences between made and missed shot attempts

A key alteration was made to the dataset before analysis was conducted by dropping rows with any missing values (5567 rows) as there was a relatively small proportion of rows with missing values compared to the dataset size. A sample of the dataset was taken to prevent any issues with the inferential tests due to the size of the dataset. A sample of 10% of the remaining dataset (12250 rows) was taken. The proportions of made to missed shots are almost the same as in the full dataset, meaning that the results should be reflective of the full dataset. In this short section, the basic summary statistics are presented, and it is considered whether there is a greater proportion of missed shots or not.

Given that the aim of this report is to establish if there are factors that influences the most and least successful shooting players, an overview of the dataset from the perspective of the result. The data was divided in to two groups: data for the shot attempts that were successful and data for the shot attempts that were not successful. The most important summary statistics for the made and missed shot attempts are provided in Tables 1 and 2, respectively.

There do appear to be differences between the summary statistics for the 'made' and 'missed' shot attempts. The sample mean for the missed shot distance is greater than the sample mean for the made shot distance. In addition to this, there appears to be less dribbles for the successful shots. There is more time left on the shot clock for the made shots meaning that the shot took less time to complete.

As the sample number of missed shots appeared to be larger than the number of made shots, a binomial test was carried out in order to ascertain the proportion of made and missed shots.

H0: Proportion Missed shots = 55%

H1: Proportion Missed shots > 55%

It is assumed that alpha is 0.05. The resulting p value was 0.5579. As the p value is larger than the alpha value of 0.05, H0 is not rejected. Therefore, it suggests that the proportion of unsuccessful shots is **not** greater than 55%. However, this test still suggests that the proportion of missed shots could be greater than 50% meaning that there would be more missed shots than made shots. Further tests will now be carried out to determine what differences there are between the missed and made shots and why shots could be unsuccessful.

4 Section II: The Individual Shooting Player

The reason for the shot result could be a result of specific metrics relating to the shooting player. This section determines if there is a difference between shot distances for made and missed attempts using a t-test. Finally, this section concludes by conducting ANOVA tests to find if the name of the shooting player has an impact on the distance from the basket and the time left on the shot clock for successful shots.

4.1 The distance from the basket- SHOT_DIST

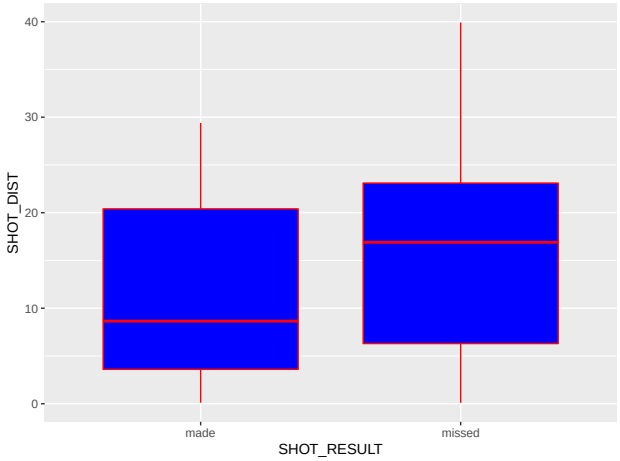

Figure 1: Boxplot showing the range of shot distances for made and missed shots

Figure 2 suggests that there is a difference in shot distance between the 'made' and 'missed' goals. There appears to be a significant difference in the median of the two types: with the shot distance median for 'made' goals under 10 ft and the 'missed' goals approx. 16 ft. This implies that a shot is more likely to be missed if it is attempted from a further distance.

The confidence intervals for these are:

- Made shots: 11.40416, 11.86576
- Missed Shots: 14.78263, 15.18896

A particularly interesting observation is that the minimum distance for a three-point shot, 23.9ft (NBA, 2021a) is not within the range of either of these confidence intervals. Also, the confidence intervals do not overlap in any way. There is an assumption of normality for this data because the sample size is very large.

Therefore, these results are reliable. The confidence intervals and sample means in the previous section suggest that the shot distance is greater for the missed shots. A one sample t-test was carried out to examine if the mean shot distance for the made shot attempts was greater than half of the distance for a 3 point shot, 11.5 ft.

H0: mu made shot distance = 11.5

H1: mu made shot distance > 11.5

The resulting p value is 0.1258 which is larger than 0.05 meaning that H0 should not be rejected. This implies that the made shot distance is not greater than 11.5 ft. This means that being closer to the basket could mean a greater chance of success.

After examining the shot distance for made shots, the key question is, do made and missed shot attempts have different shot distances? The sample means suggest that they do. To answer this question a two-sample hypothesis test was carried out to examine if this is the case. One sample is the made shots and the other the missed shots. It is assumed that alpha is 0.05. The null and alternative hypothesis are as follows:

H0: mu missed shot distance = mu made shot distance

H1: mu missed shot distance > mu made shot distance

The resulting p value was 2.2 e-16. As the p value is smaller than the significance level of 95% , H0 is rejected in favour of H1. Therefore, it can be suggested that the shot distance for missed shot attempts is greater than the shot distance for made shot attempts. This could be because the player may not be able to hold on to the ball and could attempt to shoot quickly from a far distance. This could occur when the shot clock time is up and this is the next variable which will be explored.

4.2 Time taken to attempt shot- SHOT_CLOCK

The next factor that will be considered is the amount of time that the player takes to make the shot attempt. This is represented by the shot clock column. Figure 2 shows the histogram for the made shots for this variable. The histogram appears to look relatively Gaussian. However, there is an assumption that the data follows a normal distribution because the number of observations in this sample is quite large at 12250 rows. Section I stated that the dataset suggested that for missed goals there is less time left on the shot clock, meaning that the player takes more time to attempt to make the shot.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

The first test to be conducted is an F test for equality of variance between the two samples. The successclock and missclock are the shot clock columns for made and missed shots respectively.

```
##
##  F test to compare two variances
##
## data:  successclock and missclock
## F = 1.055, num df = 5519, denom df = 6729, p-value = 0.03683
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.003279 1.109554
```

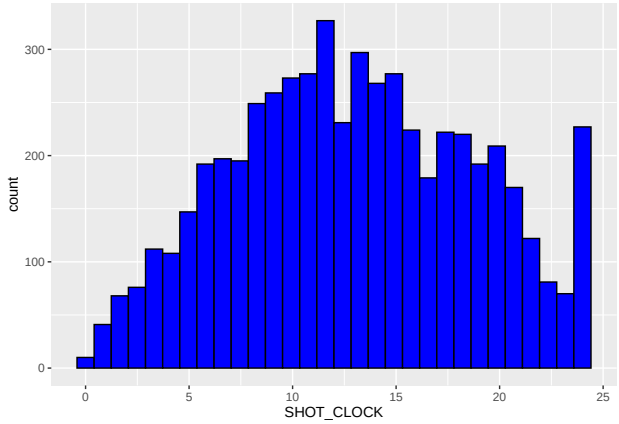

Figure 2: A histogram showing the distribution of the made shot attempts

```
## sample estimates:
## ratio of variances
##           1.055012
```

The p value is very small at the 95% significance level and therefore H0 (the variances are the same) would be rejected. However, it is statistically significant at the 99% level. Therefore, it can be suggested that the variances for made and missed shot clocks are not the same.

The next consideration is that the made shots appear to have more time left on the shot clock. A one sample t test on the successful shot attempts only was carried out to determine if more than 12.5 seconds remained on the shot clock.

```
##
##  One Sample t-test
##
## data:  successclock
## t = 5.9529, df = 5519, p-value = 1.398e-09
## alternative hypothesis: true mean is greater than 12.5
## 95 percent confidence interval:
##  12.83503      Inf
## sample estimates:
## mean of x
##  12.96297
```

The p value is less than 0.05. This suggests that it takes less time to complete a successful shot (more time left on the shot clock). This could be a result of where the player is positioned on the court. If the player is directly under the basket then a shot would not take a long time to complete.

4.3 The Individual Shooting Player

Finally, this section concludes by considering the impact of the individual shooting player on the variables. Figure 3 shows the 5 most frequent players in the sample (James Harden, Nikola Vucevic, Damian Lillard, Monta Ellis and Kyrie Irving) and the shot distance when the shot attempt is successful.

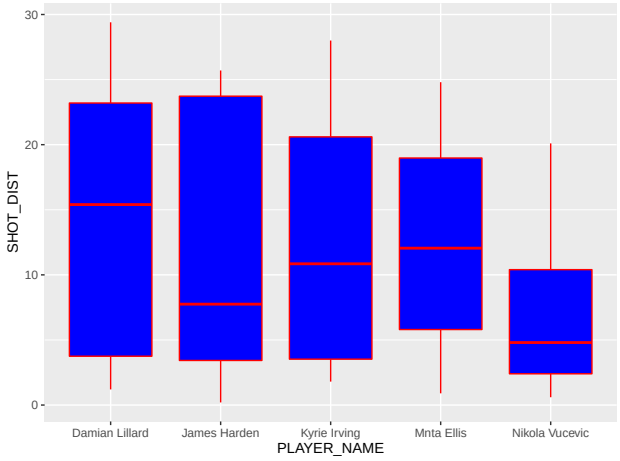

Figure 3: A boxplot showing the shot distance range for 5 NBA players

On a descriptive level, it does appear that there is a difference between the median shot distance. It is noticeable that Lillard has the highest median, so the greatest distance away from the basket. In contrast, Vucevic has the lowest median shot distance for successful shots. This suggests that players could have their specialist skills, at being successful from a short or longer shot distance. Also, it provides some evidence to the theory that the mean shot distances for successful shots is different for different players. A one-way ANOVA test is performed to determine if the mean shot distances for successful goals are different for different players.

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## PLAYER_NAME   280 109073   389.5   6.516 <2e-16 ***
## Residuals    5239 313194    59.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value resulting from this is very small suggesting that the player name does have an impact on the shot distance for a successful shot attempt. There could be different mean shot distances because of the position which the player normally plays on court. Different positions on court have different specialities. In the case of the five players covered in Figure 3, all play in the Guard position with the exception of Vucevic, who plays in a centre position (2021b). Also, there could be a matter of height, for example Lillard is the shortest player at 6ft 2 (2021c). Vucevic is the tallest at 6ft 11 (2021b) which means that it could be easier to shoot directly into the basket.

Another one way ANOVA test was conducted to find if the shot clock was affected by the player name.

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## PLAYER_NAME   280  15852   56.61   1.761 4.66e-13 ***
## Residuals    5239 168417   32.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results suggest that the player name does have an impact on the shot clock variable. This could be because some players specialise in shots from further away and they want to take the time to consider the shot before making it.

Finally, a two way ANOVA is used to consider if the type of shot (a 2 or 3 point shot) or the player name contribute to a successful shot.

```
##                          Df Sum Sq Mean Sq F value   Pr(>F)
## PTS_TYPE                  1    544   543.8  17.125 3.56e-05 ***
## PLAYER_NAME             280  15732    56.2   1.769 3.15e-13 ***
## PTS_TYPE:PLAYER_NAME    195   7858    40.3   1.269  0.00762 **
## Residuals              5043 160136    31.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All of the p values resulting from this test are lower than 0.05 meaning that both the type of shot and the individual player do have a statistically significant impact on the amount of time that the shot attempt takes.

To conclude this section it can be suggested that it is not only the individual factors that impact a players success in shots, but also the individual player themselves.

5 Section III: The Team

This section briefly considers the impact that the team could have on the shot result. It should be noted that this is for shots that were attempted by each team for home and away games. Two teams will be considered GSW and SAC.

Figures 4 and 5 show pie charts showing the proportion of shots made and missed for two teams. GSW and SAC. It shows that the proportion of made shots is greater for GSW than for SAC.

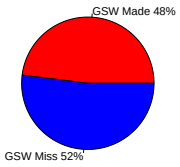

Figure 4: GSW

A proportion test will now be carried out to determine if the proportion of successful shots by GSW is greater than the proportion of successful shots by SAC.

H0: proportion made shots GSW = proportion made shots SAC

H1: proportion made shots GSW > proportion made shots SAC

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  made out of all
```

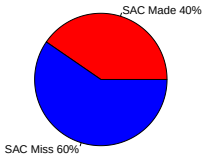

Figure 5: SAC

```
## X-squared = 4.984, df = 1, p-value = 0.01279
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.01964607 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.4802632 0.4040632
```

The results of this test suggest that GSW do have a higher proportion of made shots than SAC at 95% confidence interval because 0.012 is less than 0.05 so H0 is rejected in favour of H1. Consequently, it suggests that the proportion of made shots by GSW is greater than the proportion of made shots by SAC.

This could suggest that the team does have an impact on the likelihood of success when attempting shots. This could be because of different coaching methods or training regimes between teams. There was only a consideration between the top and lower team in the Western Conference. The result could be different in the Eastern Conference. Another factor that is outside the scope of this report but could impact these results is if the team is playing at home or away.

6 Section IV: A Regression Model

This final section of analysis considers a suitable model for the made and missed attempts. It will consider which numerical variables relating to the individual shooting player only which have an impact on the shot result. Feature selection will be used to attempt to find the best model. Logistic regression will be the technique used for this because the SHOT_RESULT variable is a qualitative variable, and it is influenced by numerical variables.

6.1 The Models

The variables that are being considered are variables that are controlled by the individual shooting player. Namely, shot distance, the number of shots the player has taken in the game, the touch time, the number of dribbles and the amount of time spent attempting the shot. Feature selection was carried out by adding each variable at a time to the model, calculating the AIC for each model and comparing them. It was found that a model with the SHOT_NUMBER variable actually made the model fit less well. Model 1 will not include the SHOT_NUMBER variable and Model 2 will contain the SHOT_NUMBER variable. These will both

be compared. The shot result took two values, made or missed and these were converted into binary values where 1 = ‘made’ and 0 = ‘missed’.

The resulting models are:

Model 1

Y = 0.363162 + -0.044207 SHOT DISTANCE + 0.037631 DRIBBLES +-0.086605 TOUCH TIME + 0.015157 SHOT CLOCK

Model 2

Y = 0.369685 + -0.044202 SHOT DISTANCE + 0.037683 DRIBBLES +-0.086434 TOUCHTIME +0.015139 SHOT CLOCK + -0.001081 SHOT NUMBER

The probabilities shown in the model summary for model 1 are statistically significant when alpha = 0.05. In addition to this, dribbles and shot clock both have positive although not very large coefficients.

The model also suggests that shot distance does have a negative impact on the success of the shot therefore reaffirming the earlier tests completed in section 2.

For model 2 the probability for the SHOT_Number is 0.788 suggesting that this variable is not statistically significant and therefore not very valuable to the model. The decision threshold has been set at 50%. An ROC curve was plotted for both models where Model 1 is Figure 6 and Model 2 is Figure 7.

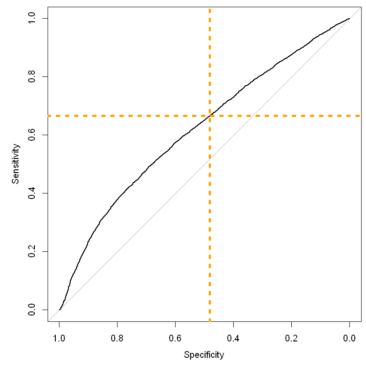

Figure 6: Model 1 ROC Curve

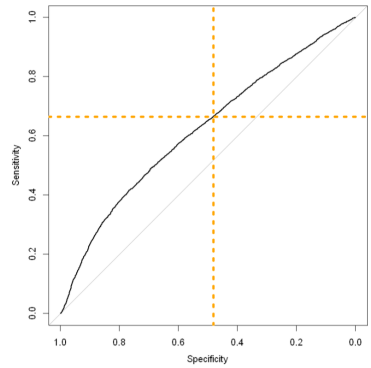

Figure 7: Model 2 ROC Curve

6.2 Discussion of the Results

Neither Model 1 or Model 2 is particularly satisfactory in actually predicting whether a shot is successful or not. The ROC curve for both models show that it is better than randomly guessing. The curves for both models are above the linear line which represents randomly guessing.

The confusion matrix for Model 1 shows that 27% of 'missed' shots were not classified correctly and 54% of 'made' shots were misclassified. The confusion matrix for Model 2 is the same in terms of percentage for classification errors. Therefore, further suggesting that this model is not very good for predicting what makes a successful shot attempt but it is better at predicting a missed shot attempt.

Therefore, this section suggests that individual player actions on the basketball court do have some impact into predicting the outcome of a shot. However, it can be suggested that the players that are around the shooting player whether they play on the same team or on the opposing team do have an impact on the probability of a successful shot attempt.

7 Section V: Conclusion

This report focussed on player-controlled variables. Proportions of the shot result was considered with a binomial test. The differences between the means of made and missed shots were examined using f and t tests. The impact of the individual shooting player on variables was examined using ANOVA. Finally, a logistic regression model was selected and evaluated in an attempt to predict which variables have the most significant impact on the shot result.

The main conclusions from this report are as follows:

1. The shot distance is generally greater for a missed shot and the player generally takes more time to attempt a shot when it is **not** successful.
2. Different players do have an impact on the distance from the basket which the shot is attempted.
3. There is a difference between teams, with some teams being more successful than others.
4. The models fitted containing variables only relating to an individual player are not very effective in predicting if a shot attempt is successful or not.

It is clear that other variables that are not player controlled such as the location the player is playing in and the coaching regime of the team need to be considered. Other variables which are not included in this dataset including the height and playing position of the shooting players could be considered when looking at the variables that could contribute to individual players and their success in the NBA.

8 References

Accenture LLP. *An MPV caliber fan experience*. Available at: https://www.accenture.com/gb-en/case-studies/communications-media/golden-state-warriors-fan-experience [Accessed: 04/02/2021]

Brown, M. 2015. *Inside the Numbers: 2015 NBA Finals Were Highest-Rated, Most-Viewed Ever For ABC*. Available at: https://www.forbes.com/sites/maurybrown/2015/06/17/inside-the-numbers-2015-nba-finals-were-highest-rated-ever-for-abc/?sh=4665e1816aef [Accessed: 04/02/2021]

Ciampolini, V. et al. 2017. *Factors associated with basketball field goals made in the 2014 NBA finals*. Motriz: Revista de Educaçã Física. 23(4).

Evans, D and Gauthier, B. 2021. *Excercise 8 Solutions*. MAT022. Cardiff University.

NBA.2019. *NBA Frequently Asked Questions*. Available at: https://www.nba.com/news/faq [Accessed: 07/02/2021]

NBA.2021a. *Rule No 1: Court Dimensions-Equipment*. Available at: https://official.nba.com/rule-no-1-court-dimensions-equipment/ [Accessed: 07/02/2021]

NBA.2021b. *Nikola Vucevic*. Available at: https://www.nba.com/player/202696/nikola_vucevic [Accessed: 07/02/2021]

NBA. 2021c. *Damian Lillard* . Available at: https://www.nba.com/player/203081/damian_lillard [Accessed: 07/02/2021]

Zhang, S. et al. 2017. *Players' technical and physical performance profiles and game-to-game variation in NBA*. International Journal of Performance Analysis in Sport. 17(4). pp.466-483.