Modeling Shot Efficiency in the NBA

Stat Guys: Lewis Eatherton, Chris Yang, Charlie Bonetti

11/17/2020

Introduction

For our STA 210 Final Project, our group is interested in investigating how certain factors may have an influence on a basketball players' shot during a game in the National Basketball Association (NBA). We are all avid NBA enthusiasts and express great curiosity in what makes a good shot versus a bad shot. Thus, conducting this research was an exciting opportunity to explore the presence or absence of certain variables that surprisingly, or unsurprisingly, have an effect on scoring a shot in the basket, also known as shot success.

Our interest in this subject matter was piqued by an 2015 article titled "Basketball Shot Types and Shot Success in Different Levels of Competitive Basketball" published by Frane Erčulj and Erik Štrumbel, which described the effect of different shot types on shot success across different levels of competition. It was insightful to learn that there were no discernable differences between situational variables, such as the type of player (e.g. Center) or where he shot the ball (e.g. in the paint) on shot success between levels (1). Since it was demonstrated that the effect of situation variables on shot success remained constant throughout all levels of competition, we were motivated to identify what exactly these situation variables might be, and to what degree of influence they had on player's shot success at the professional level.

As a result, we formulated our research question: "Do certain situational variables during an NBA game have an effect on shot success?" Based on our prior knowledge of the game of basketball, we recognize that both players and coaches deem certain shots as "good" and as "bad". Yet, what qualities characterizes them to fall under those two distinct categories? Furthermore, by our own intuition after watching countless games and playing the sport ourselves, we hypothesize that some situational variables (e.g. shot distance from the hoop or the proximity of the closest defender) will have a greater effect on an NBA player's shot success than other situational variables. It seems intuitive that, as the time left on the clock and distance to the defender decrease, the likelihood of making a shot increases. To what extent is this true? Seeking clarity in how these factors, amongst others, contribute to a successful shot are the key objectives of our analysis.

Data

To seek answers to our series of questions, we will explore data from the 2014-2015 National Basketball Association (NBA) season. The data was originally collected via the NBA API PHP Library, a software tool developed by Jason Roman which scrapes historical information of past games, including statistical data, from previous seasons off of the NBA website. The data from the NBA website was acquired via manual documentation by analysts at the scoring table during games as well as computer vision techniques that track numbers such as scores, assists, etc.

Each observation in our data set is a unique shot that was taken in an NBA basketball game during the 2014-2015 season. Each observations is characterized by 21 features, including factors such as the game matchup, player name, and shot result. In total, there are 128,069 observations within the data set.

The response variable we are interested in investigating is:

- `SHOT_SUCCESS`: Indicates whether or not the NBA player scored the shot. `SHOT_SUCCESS` takes a value of 1 if the shot was made, 0 otherwise.

The predictor variables that we expect to have an influence on the response variable are:

- `SHOT_CLOCK`: Number of seconds left on the shot clock when the player shot the ball.
- `SHOT_DIST`: Distance in feet from the hoop when the player shot the ball.
- `SHOT_RESULT`: Description of shot, either "made" or "missed".
- `TOUCH_TIME`: Number of seconds before the player shot the ball after being passed to.
- `CLOSEST_DEFENDER`: Who the closest defender was to the shooting player.
- `CLOSE_DEF_DIST`: Number of feet the defender was from the player when the ball was shot.

Data Cleansing

To adequately prepare the data for analysis, we removed and modified incomplete data. To consider the categorical variable of whether the shot was "made" or "missed" in our model, we created an indicator variable to represent these two categories, with "1" denoting the shot was made, "0" denoting the shot was made. Analysis of missing data reveals that only one variable, `SHOT_CLOCK`, contains unavailable values. This is likely because the shot clock is turned off when the game clock in the NBA is below 24 seconds. To accommodate for this, we removed 5567 missing values from the dataset since they accounted for only 4.3468755 percent of all observations, which is an inconsequential amount.

feature	null_count
GAME_ID	0
MATCHUP	0
LOCATION	0
W	0
FINAL_MARGIN	0
SHOT_NUMBER	0
PERIOD	0
GAME_CLOCK	0
SHOT_CLOCK	5567
DRIBBLES	0
TOUCH_TIME	0
SHOT_DIST	0
PTS_TYPE	0
SHOT_RESULT	0
CLOSEST_DEFENDER	0
CLOSEST_DEFENDER_PLAYER_ID	0
CLOSE_DEF_DIST	0
FGM	0
PTS	0
player_name	0
player_id	0
SHOT_SUCCESS	0

Exploratory Data Analysis

The next step in our project was conducting an Exploratory Data Analysis (EDA) of the variables of interest pertaining to our model. Here, we reveal intriguing relationships between several predictors and the response variable that demonstrate how they have an impactful or negligible effect on shot success.

The first EDA was on our response variable, 'SHOT_SUCCESS'. The summary statistics demonstrate that out of the total 128,069 taken during an NBA game, approximately 45.6 percent made the basket.

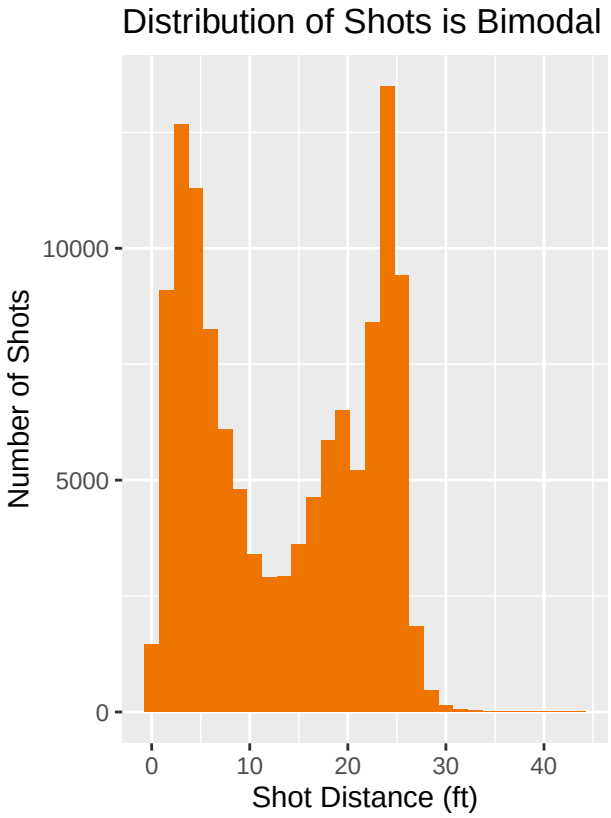
SHOT_SUCCESS	n	freq
0	66622	0.544
1	55880	0.456

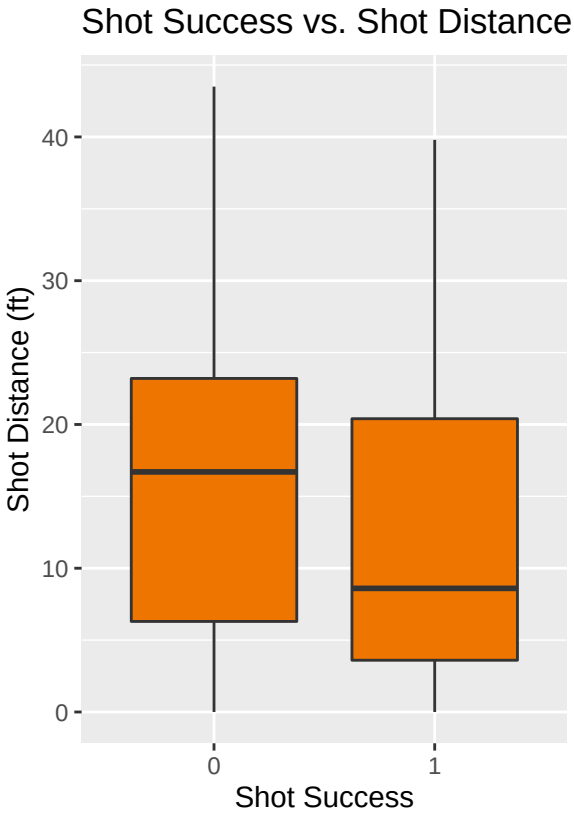
Now, we proceed to perform EDA on the predictor variables. Of particular interest are `SHOT_DIST` and `CLOSE_DEF_DIST`, which represent the distance in feet from the hoop when the player shot the ball and the number of feet the defender was from the player when the ball was shot, respectively.

Summary Statistics of `SHOT_DIST`:

SHOT_SUCCESS	min	max	mean	sd	median	IQR
0	0	43.5	14.920	8.520	16.7	16.9
1	0	39.8	11.683	8.754	8.6	16.8

For the predictor variable that measures the distance to the basket (`SHOT_DIST`), we observed that the median shot distance for a missed shot was nearly twice as large (16.7 feet) as the median for a made shot (8.6 feet). While the minimum values for a missed and made shot were both zero, the maximum value for a missed shot (43.5 feet) was greater than that of a made shot (39.8 feet). The spread across both groups was relatively large, at 16.9 feet and 16.8 feet respectively. This suggests that type of shots taken during the 2014-2015 NBA season were of a wide variety, including both close-up shots (e.g. lay-ups) as well as long-distance shots (e.g. 3-pointers).





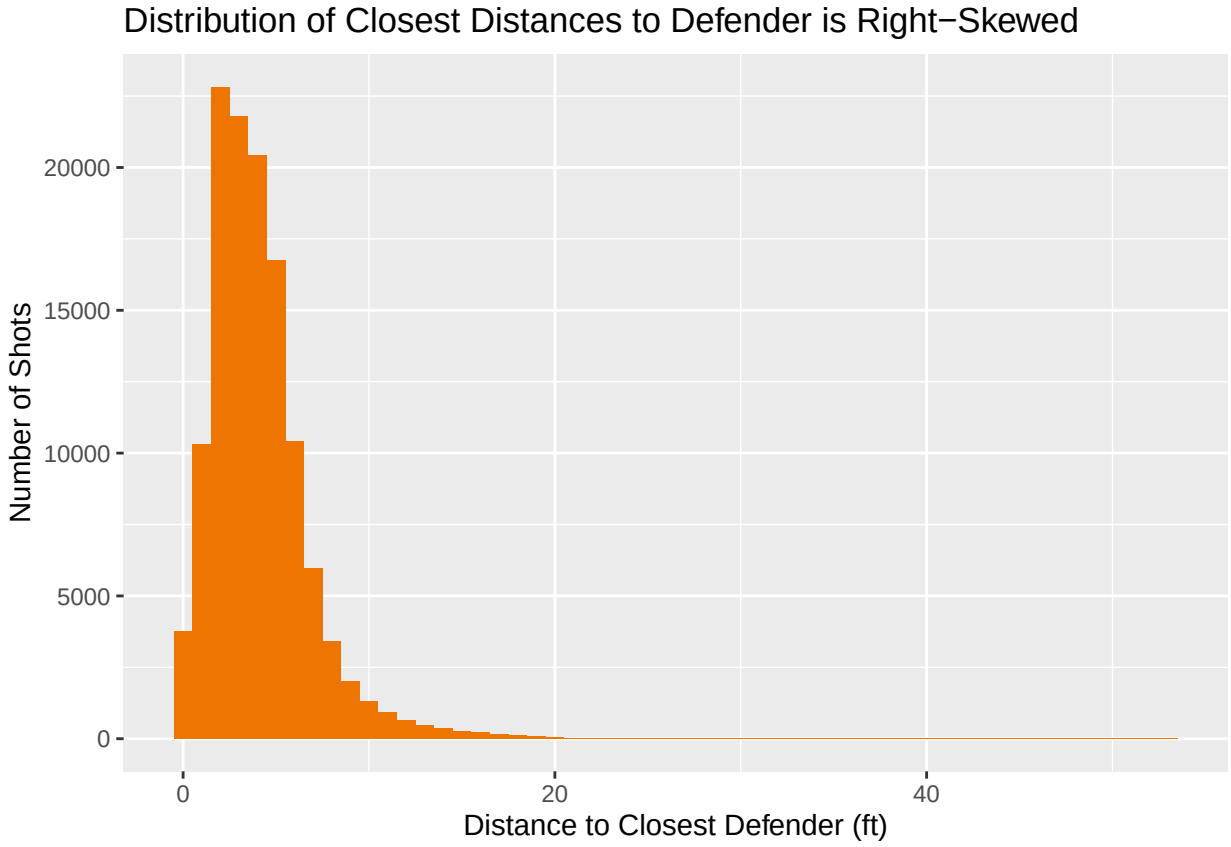
These graphs tell us interesting relationship about the distribution of shots in the NBA and the relationship between shot distance and success. The plot for Shot Distance versus Number of Shots shows us that the distribution of shots in the NBA is bimodal and aysmmetrical, with most shots either being right next to the basket (<5 feet) or pretty far from the basket (~25 ft). This is likely because the NBA three point line (the line in which shots shot behind it are worth three points instead of two) is 24ft away from the basket.

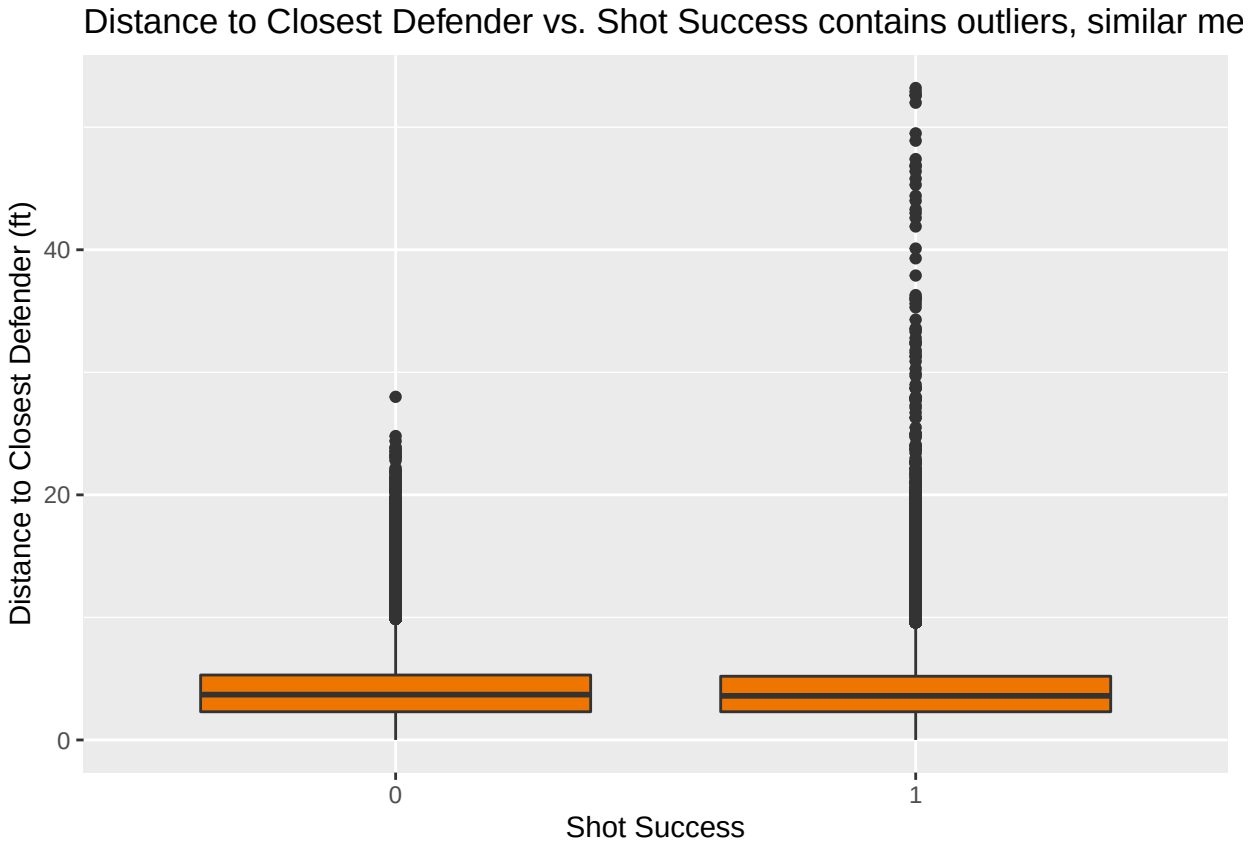
In the plot for Shot Success versus Shot Distance, there seems to be a very similar spread in shot distance between a shot made and and a shot missed. The medians appear to be noticeably different from one another as the median shot distance for missed shots is significantly greater than that of made shots. Lastly, there are no noticeable outliers presented in the boxplot.

Summary Statistics of `CLOSE_DEF_DIST`:

SHOT_SUCCESS	min	max	mean	sd	median	IQR
0	0	28.0	4.119	2.613	3.7	3.0
1	0	53.2	4.125	2.908	3.6	2.9

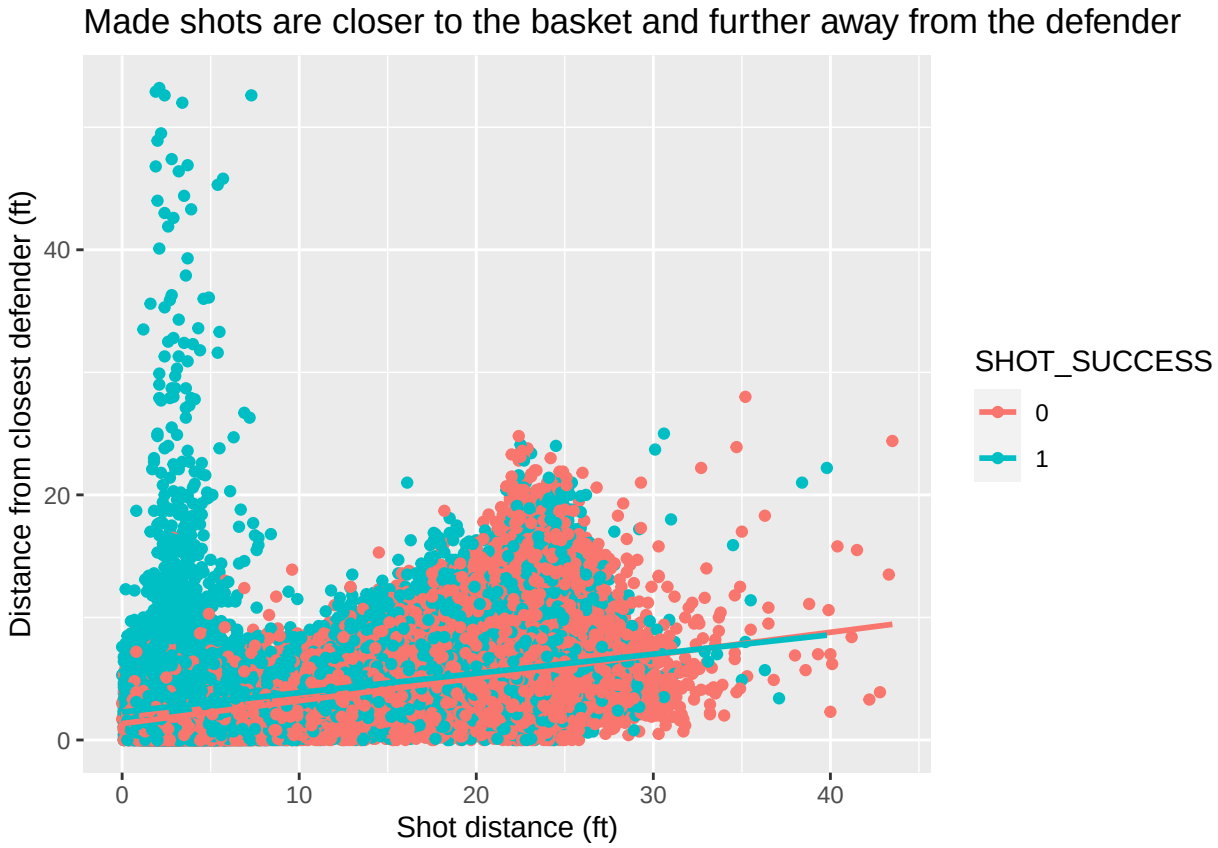
For the predictor variable that measures the distance to the closest defender (`CLOSE_DEF_DIST`), we observed that the median shot distance was roughly the same for both made and missed shots at 3.7 feet and 3.6 feet respectively. However, while the minimum distances between the closet defender to a player were both zero, the maximum distance for a made shot (53.2) was nearly 2 times than that of a missed shot (28.0 feet). This suggests that a decreased distance to the closest defender meant that a shot was more likely to be made. Additionally, the spread across both groups was relatively small, at 3.0 feet and 2.9 feet respectively.





In the histogram of Number of Shots versus the Distance to Closest Defender, the distribution is clearly right skewed and unimodal. An interesting relationship illustrated by these graphs is that at most times, there is a nearby defender guarding the player. However, there also exist circumstances in which the closest defender is significantly very far away, which is why the graph appears to be concentrated towards the left; this could occur in cases where there is a fast-break scenario, when the player quickly escapes the defender, or when the player decides to take a shot at the end of the quarter.

In the boxplot for Shot Success versus Shot Distance, there seems to be a very similar spread between a shot made and the distance to the closest defender. The medians appear to be noticeably similar from one another. Lastly, there are plenty of outliers across both shots that are made and missed, which informs us that many of the shots tended to differ in distance from the average, which is why the median was a better measurement for this predictor. An interesting relationship is that while the median distance of a made or missed shot was the same considering the distance to the closest defender, a greater number of made shots occurred at a distance further away from the defender.



According to this plot, there appears to be a relationship between shot distance and shot success, but not distance from defender and shot success. In the plot, there are clearly more made shots than not made shots when the shot is close to the basket, but as the shot distance goes further away there is a significant increase in proportion of missed shots as compared to made shots. On the other hand, the distance from closest defender does not appear to tell us much about the shot success when the shot is somewhat contested (nearest defender <15 feet away) as in some cases there are a higher proportion of made shots and in other cases there is a lower proportion of made shots. One interesting relationship is that shots closer to the basket have more observations where the closest defender was very far away (20-50ft). This is likely due to the fact that turnovers in basketball lead to transition layups, where the offensive player is attacking the basket uncontested. This graph also helps us explain a phenomenon from earlier, as teams are likely taking shots closer to the basket because these shots are wide open (no defender nearby) in some situations.

Methodology

After gathering meaningful insight into the predictor variables and response variable through EDA, we proceed to the next step of our analysis, which is the modeling process. The model diagnostics as well as conditions were evaluated to ensure that our model selection was appropriate.

Model Selection

We thoughtfully chose to utilize a linear regression model to fit the different predictor variables to our response variable, which indicates whether or not the NBA player scored the shot. Given that it is a binary response variable, using linear regression was most fitting for our project.

To obtain our final model, we chose to conduct a backward selection procedure. Therefore, we started out with the model that included all seven predictor variables that we assumed to have a significant effect on the response variable. The predictor variables we selected in our initial model were the time left on the shot clock (`SHOT_CLOCK`), the distance of the shot (`SHOT_DIST`), the amount of time the shooter touched/held the ball

for (`TOUCH_TIME`), and the distance of the closest defender (`CLOSE_DEF_DIST`), and whether the shot was a 2-pointer or 3-pointer (`PTS_TYPE`). It is important to note that the closest defender was a categorical variable with over 200 levels, so it wasn't feasible for us to use it in our model, which was why we excluded it from the model.

term	estimate	std.error	statistic	p.value
(Intercept)	-0.048	0.042	-1.155	0.248
SHOT_CLOCK	0.015	0.001	13.780	0.000
SHOT_DIST	-0.063	0.001	-55.793	0.000
DRIBBLES	0.026	0.005	5.440	0.000
TOUCH_TIME	-0.055	0.006	-10.041	0.000
PTS_TYPE	0.092	0.021	4.483	0.000
CLOSE_DEF_DIST	0.103	0.003	36.706	0.000

In terms of conducting the backward selection, we eliminated the variables based on which predictor variable resulted in the model having the lowest AIC value. The resulting model had time left on the shot clock, the distance of the shot, and the distance of the closest defender.

term	estimate	std.error	statistic	p.value
(Intercept)	-0.058	0.018	-3.171	0.002
SHOT_CLOCK	0.019	0.001	17.652	0.000
SHOT_DIST	-0.060	0.001	-69.588	0.000
CLOSE_DEF_DIST	0.108	0.003	38.807	0.000

Interaction Terms

After we had created an initial model, the next step towards creating a final model is exploring inclusion of possible interaction terms.

- `CLOSE_DEF_DIST*SHOT_DIST`: From our EDA, it was demonstrated that the vast majority of wide open shots (where the closest defender to the player was greater than 20 feet away) were lay-ups, or shots within a few feet of the basket. Therefore, the effect of shot distance on shot success may depend on the distance of the player to the closest defender.

Table 1: AIC & BIC of model with interaction term

AIC	BIC
161851.3	161899.8

Table 2: AIC & BIC of model without interaction term

AIC	BIC
162561.8	162600.7

Resid..Df	Resid..Dev	df	Deviance	p.value
122498	162553.8	NA	NA	NA
122497	161841.3	1	712.549	0

term	estimate	std.error	statistic	p.value
(Intercept)	-0.51874	0.02550	-20.34628	0
SHOT_CLOCK	0.02018	0.00105	19.13285	0
SHOT_DIST	-0.03029	0.00140	-21.57903	0
CLOSE_DEF_DIST	0.26408	0.00675	39.12785	0
SHOT_DIST:CLOSE_DEF_DIST	-0.00863	0.00033	-25.95053	0

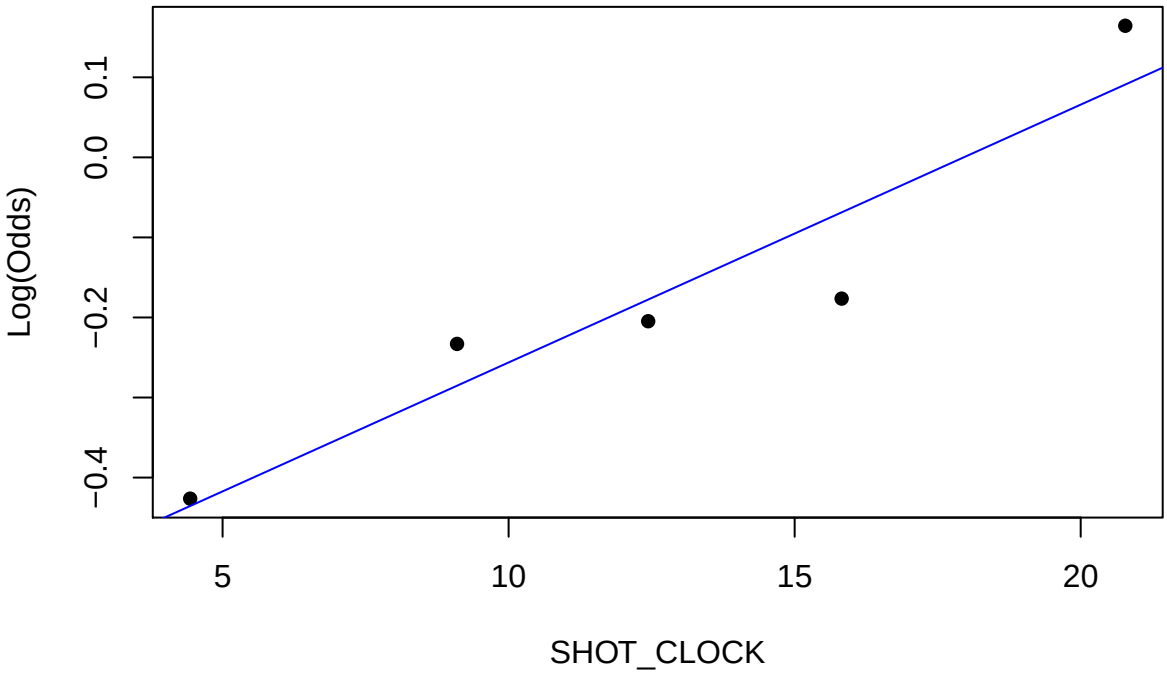
After doing a drop-in deviance test and looking at AIC and BIC values, we concluded that this interaction term did indeed help the predictive value of our model and added it as a predictor to the response variable.

Model Conditions

The next step in our analysis is checking the model conditions. Given that our model is a logistic regression model, we should check three conditions: linearity, randomness, and independence.

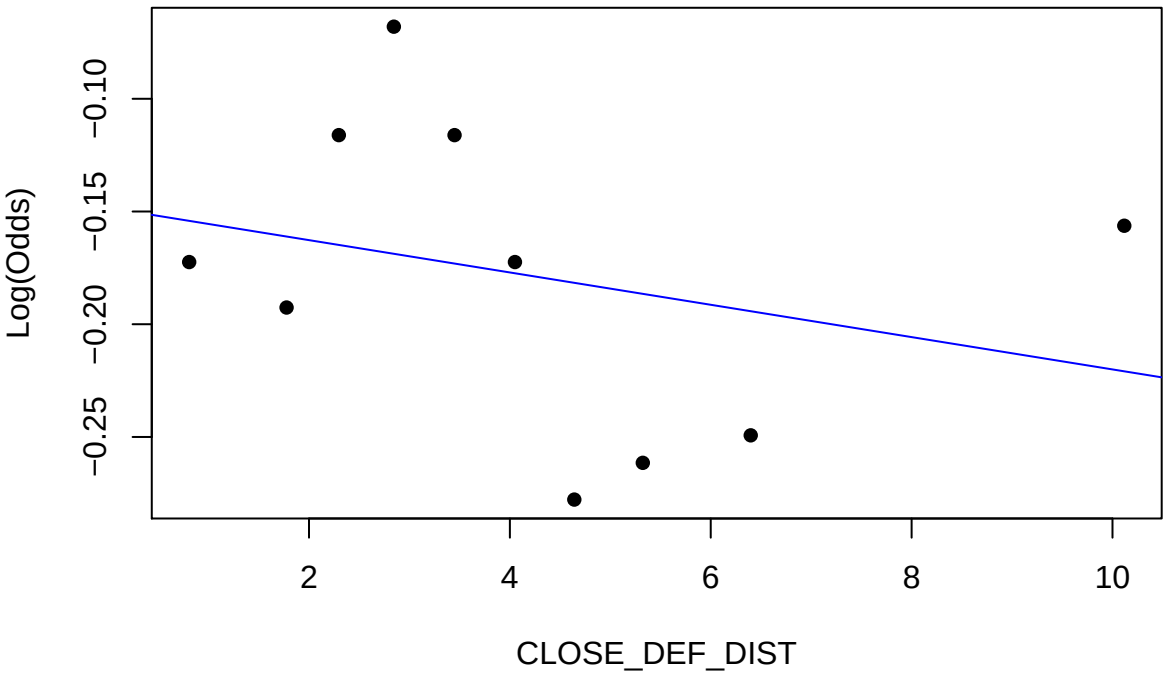
Linearity

We evaluate the linearity condition by calculating the empirical logit plots for all quantitative predictors from our model. Note, we excluded the variable which characterized each defender's name (`CLOSEST_DEFENDER`) as it was a categorical variable with over 200 levels, so it wasn't feasible for us to use it in our model, which was why we excluded it from the model and model conditions analysis.



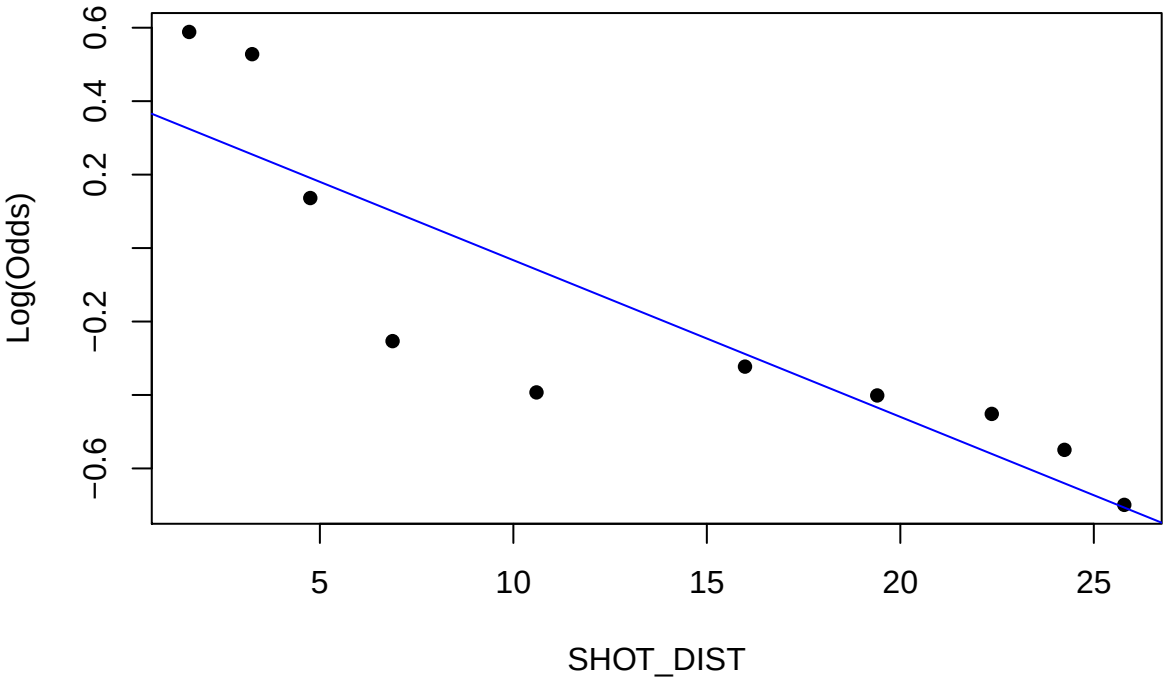

```
## NULL
```

Linearity conditions are satisfied for the time left on the shot clock (`SHOT_CLOCK`) because the log-odds have a linear relationship with the predictor.



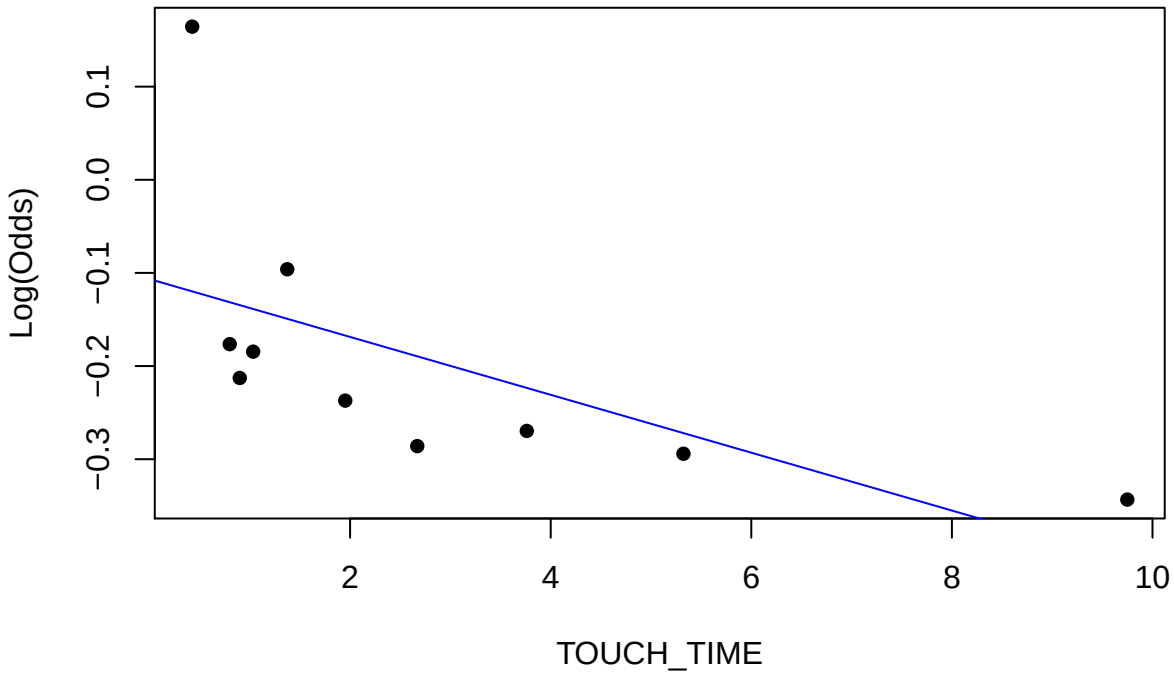

```
## NULL
```

Linearity conditions are satisfied for the distance to the closest defender (`CLOSE_DEF_DIST`) because the log-odds still have a negative linear relationship with the predictor, despite it being very weak.




```
## NULL
```

Linearity conditions are satisfied for the distance of the shot (`SHOT_DIST`) because the log-odds have a negative linear relationship with the predictor.




```
## NULL
```

Linearity conditions are satisfied for the distance of the shot (`SHOT_DIST`) because the log-odds have a negative linear relationship with the predictor.

Independence

The independence condition is satisfied because the shot of one NBA basketball player does not appear to have an affect on the shot of another NBA basketball player.

Randomness

The condition for randomness is satisfied because we have no reason to believe otherwise when considering how the data was collected, given the fact that all statistical game history was recorded from the 2014-2015 season.

Model Diagnostics

The next step in our analysis is evaluating the model diagonostics.

Multicollinearity

To adequately check model diagonistics, we interpreted the multicollinearity of the model.

	x
SHOT_CLOCK	1.040
SHOT_DIST	4.242
CLOSE_DEF_DIST	8.958
SHOT_DIST:CLOSE_DEF_DIST	16.244

While our interaction term did have a Variance Inflation Factor (VIF) value of 16.244, which is greater than 10 (indicating multicollinearity), we chose to keep the term in our model for several reasons. First, being that it is somewhat expected that an interaction term will correlate with the terms it is made up from. Second,

given that the multicollinearity value was only off by single digits, we felt that it was necessary to include in the model, especially considering it was significant according to the interaction procedure we had conducted using AIC as a criterion earlier.

Final Model and Results

The predictors in our final model were the amount of time left on the shot clock, the distance of the shot, the distance of the closest defender, and the interaction term between the distance of the closest defender and the distance of the shot. Our final model can be expressed as:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.519	0.025	-20.346	0	-0.569	-0.469
SHOT_CLOCK	0.020	0.001	19.133	0	0.018	0.022
SHOT_DIST	-0.030	0.001	-21.579	0	-0.033	-0.028
CLOSE_DEF_DIST	0.264	0.007	39.128	0	0.251	0.277
SHOT_DIST:CLOSE_DEF_DIST	-0.009	0.000	-25.951	0	-0.009	-0.008

The results show that in our final model, we have three predictors and one interaction term that were ultimately selected to predict shot success. All predictors, including the interaction term, were statistically significant as their p-value was nearly zero and below our established alpha threshold of 0.05. As the model predicts, if the distance between a player and the basket increases by 1 foot, we expect the shot success of a basketball player to multiply by a factor of exp(-0.030), holding all else constant. This aligns with our intuitive sense established earlier in our analysis; as the distance between player and the basket increases, it is generally more difficult for the player to make the shot. However, it is noticed that the model predicts that as the distance to the closest defender increases by one foot, we expect the shot success of a basketball player to multiply by a factor of exp(0.264). While this seems counterintuitive at first, further analysis of relationship between the defender and the player and points scored makes greater sense. The majority of shots taken in the paint are heavily guarded as the EDA between shot success and distance to closest defender demonstrated, despite there being successfully scored baskets. Consequently, it is understandable that the closer the defender, there is an increased liklihood that the shot is going to be scored. Thus, to answer our research question based on our analysis stated during our introduction, we can conclude that there are indeed situational variables in-game that can affect the outcome of a shot in an NBA game. While some variables may help facilitate scoring a basket, proven by the log-odds of the time left on the clock and the distance closest to the defender, there are also factors that decrease the chance of shot success. This was exemplified by the distance of the shot from the basket, and the interaction between the shot distance along with the distance to the closest defender.

Below illustrates the formula, including all variable coefficients, for our model:

Predicted log odds of SHOT SUCCESS ~ -0.51874 + 0.02018 * SHOTCLOCK - 0.03029 * SHOT DISTANCE + 0.26408 * CLOSE DEF DIST - 0.00863 * (CLOSE DEF DIST * SHOT DISTANCE)

or in terms of predicted probability

Predicted probability of SHOT_SUCCESS ~ exp(-0.51874 + 0.02018 * SHOT_CLOCK - 0.03029 * SHOT_DISTANCE + 0.26408 * CLOSE_DEF_DIST - 0.00863 * (CLOSE_DEF_DIST * SHOT_DISTANCE)) / (1 + (-0.51874 + 0.02018 * SHOT_CLOCK - 0.03029 * SHOT_DISTANCE + 0.26408 * CLOSE_DEF_DIST - 0.00863 * (CLOSE_DEF_DIST * SHOT_DISTANCE)))

Discussion

Limitations:

There a few important limitations with our analysis and model. Firstly, one variable had missing observations; there were 5,567 unavailable values for the variable which counted the time left on the shot clock when the player shot the ball. This may have been due to recording error or instances where the shot clock was turned

off. Given that 5,567 is marginal compared to the total number of observations (128,069), we chose to remove all rows containing NAs for this variable from the dataset. As a result, The dataset theoretically does not differ to a significant degree systematically compared to the larger population. Secondly, there appeared to be some inconsistencies/errors with the data. One example is the fact that there was negatives values for the amount of time the player touched the ball before shooting, which shouldn't be possible as a player can not possess the ball for a negative amount of time. There was also issues with the linearity assumption with the distance of the closest defender and the amount of time the shooter touches the ball for, which could affect the validity of our model (though we did not end up using the amount of time the shooter touched for in our final model). There was also issues with collinearity between our interaction term and the terms comprising the said interaction term (the distance of the closest defender and the distance of the shot). While this might inflate the variance of our coefficients, we chose to keep the interaction term in our model because it seemed to hold significant predictive value as seen by a lower AIC and BIC when included in the model. Lastly, while we felt like the categorical variable CLOSEST_DEF might have held predictive value in our model (as some players in the NBA are considered 'better' defenders than others), there were simply too many levels (or players) for us to feasibly include in our model.

Next steps:

To address the limitation of our model there are a few steps additional we would have liked to take. Firstly, to account for the fact that there were missing values for the time left on shot clock, we might look to replace time left on the shot clock with the time left in the quarter (if less than 24 seconds). It seems very likely that the missing values for the time left on the shot clock come from the fact that the shot clock is turned off when there's less than 24 seconds left in an NBA quarter. So, if we had data that included the amount of time left in the quarter, we could replace shot clock values for the amount of time left in the quarter in the instances when there was 24 seconds or less left in the quarter. In order to work around the fact that there are too many individual NBA players to include the closest defender in our model, it might make sense to replace that categorical variable with a continous one that is reflective of defense capabilities. We would have liked to add a continuous variable like defense rating (a quantative measure of a player's defense abilities) in order to see how that affects shot success.

Furthermore, there are additional avenues we'd like to explore in our analysis. One we would be interested in is using player height (or the difference in height of the shooting player and closest defender) affects shot success, as players sometimes have to change their release to shoot over taller players in the NBA. Alternatively, something like the difference in wingspan could be used. Another avenue we'd be interested in exploring how the situational variables of making a successful shot may differ or stay the same during "clutch" scenarios, where the outcome of a game is at stake. In this case, we could explore other predictors within our model and set thresholds, such as having less than 10 seconds left during the game, being in the 4th period of the game, and having a final margin of victory of less than or equal to 3 points. Furthermore, another avenue we would be interested in exploring is the effectiveness of different defenders on a player's shot success in NBA games. In the current stage of our project, we investigated the variables that played in a role in optimizing a successful shot. Out of curiosity, how would these factors change if we explored the data from the perspective of the defender, who is attempting to prevent a player from scoring a basket? Further analysis of the distance between contested shots as well as field goal percentage could yield interesting factors that may have some degree of influence on defending a shot. Additionally, based on these characteristics and others, perhaps we could identify which NBA players constitute the "elite" class of the defensive side of basketball.

References

(1) Erčulj F, Štrumbelj E (2015) Basketball Shot Types and Shot Success in Different Levels of Competitive Basketball. PLoS ONE 10(6): e0128885. https://doi.org/10.1371/journal.pone.0128885

Appendix

Backward Selection Output

```
## Single term deletions
```

```
##
## Model:
## SHOT_SUCCESS ~ SHOT_NUMBER + SHOT_CLOCK + SHOT_DIST + DRIBBLES +
##     TOUCH_TIME + PTS_TYPE + CLOSE_DEF_DIST
##                Df Deviance    AIC
## <none>              162298 162314
## SHOT_NUMBER     1   162304 162318
## SHOT_CLOCK      1   162489 162503
## SHOT_DIST       1   165538 165552
## DRIBBLES        1   162327 162341
## TOUCH_TIME      1   162404 162418
## PTS_TYPE        1   162318 162332
## CLOSE_DEF_DIST  1   163739 163753
## Single term deletions
##
## Model:
## SHOT_SUCCESS ~ SHOT_CLOCK + DRIBBLES + TOUCH_TIME + SHOT_DIST +
##     PTS_TYPE + CLOSE_DEF_DIST
##                Df Deviance    AIC
## <none>              162304 162318
## SHOT_CLOCK      1   162495 162507
## DRIBBLES        1   162334 162346
## TOUCH_TIME      1   162409 162421
## SHOT_DIST       1   165541 165553
## PTS_TYPE        1   162325 162337
## CLOSE_DEF_DIST  1   163741 163753
## Single term deletions
##
## Model:
## SHOT_SUCCESS ~ SHOT_CLOCK + DRIBBLES + TOUCH_TIME + SHOT_DIST +
##     CLOSE_DEF_DIST
##                Df Deviance    AIC
## <none>              162325 162337
## SHOT_CLOCK      1   162532 162542
## DRIBBLES        1   162354 162364
## TOUCH_TIME      1   162434 162444
## SHOT_DIST       1   167561 167571
## CLOSE_DEF_DIST  1   163764 163774
## Single term deletions
##
## Model:
## SHOT_SUCCESS ~ SHOT_CLOCK + TOUCH_TIME + SHOT_DIST + CLOSE_DEF_DIST
##                Df Deviance    AIC
## <none>              162354 162364
## SHOT_CLOCK      1   162590 162598
## TOUCH_TIME      1   162554 162562
## SHOT_DIST       1   167583 167591
## CLOSE_DEF_DIST  1   163792 163800
```