Factors Contributing to Shot-Making Probabilities in the NBA

Samuel Yarger, Yilin Luo, James Liang, Sathwika Manda

Introduction

Our project explores what factors influence a player's probability of making a shot, and how we can model those probabilities to better understand shot selection. Many fans, coaches, and players debate shot selection. Questions like *"Is a layup better than an open mid-range jumper?"* or *"Should teams avoid mid-range shots altogether?"* are often answered anecdotally. However, we wanted to analyze these questions empirically.

Understanding what makes a shot more or less likely to go in is important for multiple reasons. First, shot selection is directly tied to offensive efficiency, and therefore to winning games. Second, teams are increasingly relying on analytics to guide in-game decision-making, scouting reports, and player development. Finally, modeling shot probability can also help identify inefficiencies or tendencies in player behavior, which can be useful for both coaches and analysts in developing better strategies.

In this project, we aim to model the probability that a shot will be made using predictors such as shot distance, court location, defender proximity, quarter of the game, score differential (i.e., whether the team is leading or trailing), and shot type (e.g., 2-point vs 3-point attempts). We did Generalized Additive Modeling to flexibly capture nonlinear relationships and spatial effects while ensuring results are interpretable, and Multilevel Logistic Regression Modeling to account for team-level variation in shot-making tendencies.

With these models, we looked to analyze both general and team-specific questions, such as how shot probability varies with distance and location on the court, whether teams differ in their baseline shooting accuracy or in how distance impacts their shot success, and whether game context, such as the score or quarter, significantly affects shot probability.

Our analysis confirms that shot distance is the single strongest predictor of shot success, but court location and shot type also play key roles, with "hot" zones near the corners and around the basket.

Data

The data we used was from the 'NBA 2023 Player Shot Dataset from Kaggle. We used three shot charts from that dataset: one each for LeBron James (1,533 shots), James Harden (1,025 shots), and Stephen Curry (1,434 shots). Each dataset records detailed shot-level information from the 2022–2023 NBA season.

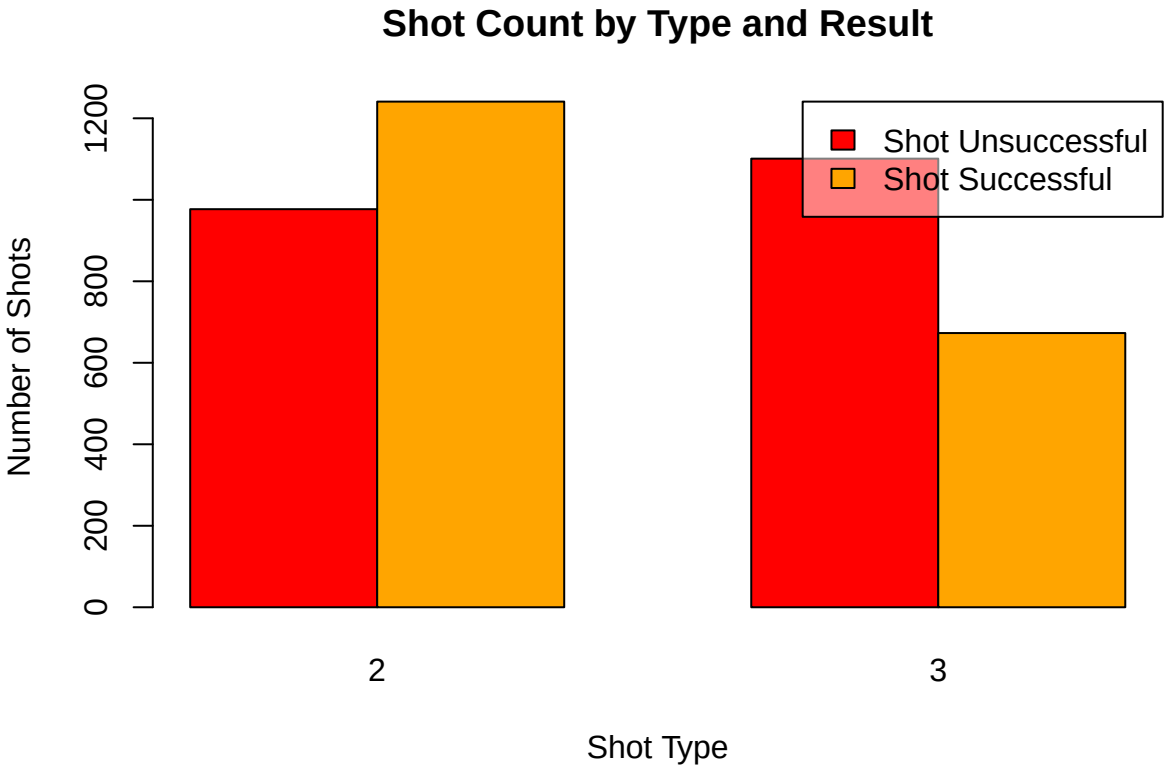
Each shot entry includes 'Court location' (top, left coordinates for shot position), 'Shot distance' (in feet), Shot result (a binary variable of 'made' or 'missed'), 'Game context' (quarter, whether the team was leading, and game time remaining), 'Shot type' (2-pointer or 3-pointer), Opponent and team info, Game date, and Scores.

To prepare the data for modeling, we merged all three player datasets into a single dataframe. We also created a binary response variable (‘result_binary’) to indicate whether a shot was made (‘1’) or missed (‘0’). We standardized logical variables like ‘lead’(whether the team was leading) and ‘result’ to ensure consistency across players. We also converted categorical variables like ‘qtr’(the quarter) and ‘shot_type’ (2-pointer or 3-pointer) into factor variables.

During the model setup, we also excluded columns unrelated to prediction (like the date of when the shot was taken), and we verified that spatial coordinates (top, left) were within the expected ranges.

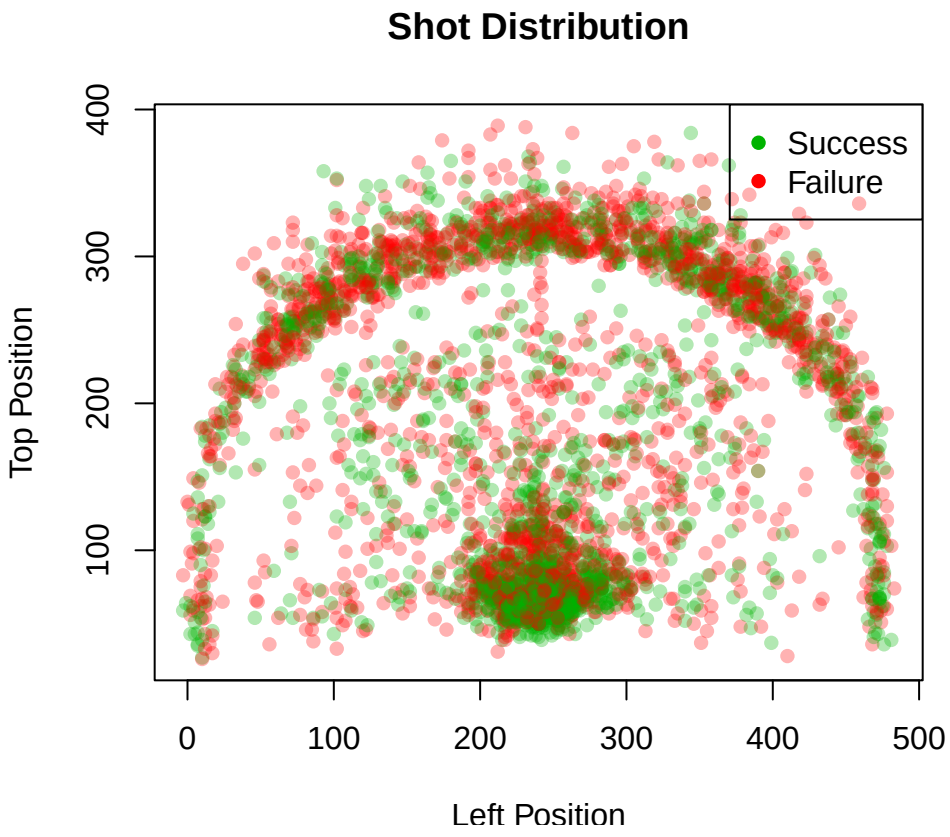
EDA

Before we do anything in an attempt to model the data, we first need to consider the types of shots being made

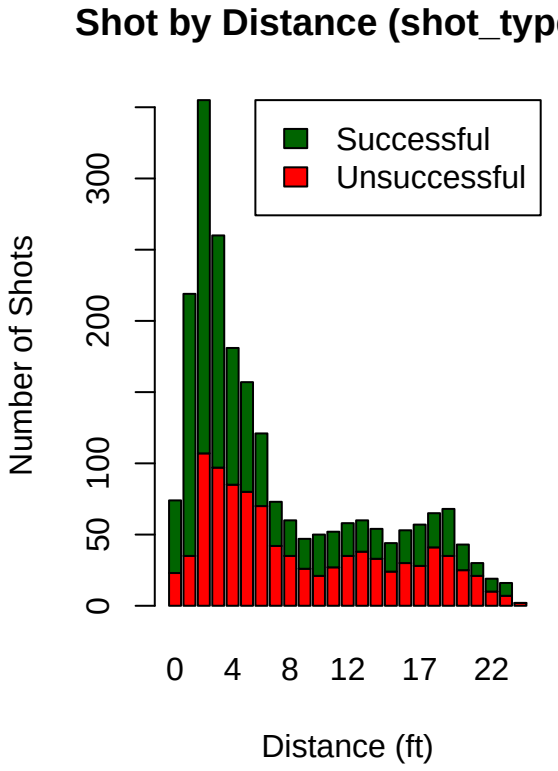


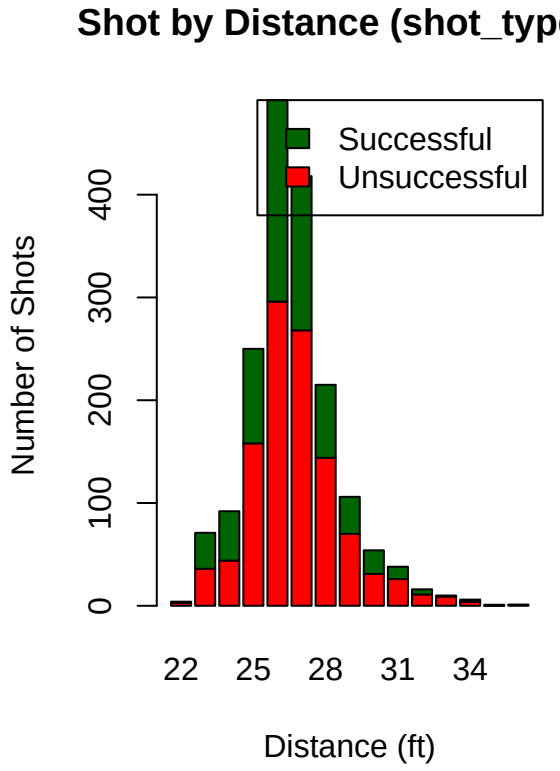
Here wee see that a similar amount of 2 point and 3 point shots are attempted, yet we see that 2 point shots are successful more often then they are missed, while 3 point shots are missed almost twice as much as they are successful.

Now lets take a look at where shots are being taken from.



As we can see, there is a lot of overlap. Here we can see that using position as a predictor by itself may not lead to the best results, and our models are going to require other predictors. The location around the hoop does seem to be a hotspot, with a ring around the the 3 point curve.





Finally we see that the most shots are being taken around 3 feet from the hoop for 2 pointers. For 3 pointers we see that most shots are taken around 26 feet away from the hoop. For both of these locations, we also see that the proportion of shots that are successful is higher around these hotshots as well (where wee see the green bars for successful shots get larger in comparison to the red bars for unsuccessful shots). This is showing that these hot spots could possibly exist as making shots from these spots does lead to more successful shots on average for players, and that players have learned to shoot from there as a response. We hope to demonstrate this later on through our modeling.

Methods & Results, Part One

What Will We Do? A Brief Abstract for "Methods & Results, Part One"

We will fit a generalized additive model (GAM) and a multilevel logistic regression model. We will explain and analyze both models, provide an example prediction, and conclude with a brief comparison.

Preparation

To begin with, the dataset includes three CSV files, each corresponding to a well-known player with distinct playing styles and positions – LeBron James (small forward / power forward), James Harden (guard), and Stephen Curry (point guard). We need to combine these three CSV files to obtain a comprehensive perspective. However, it is also worth noting that although these three players provide a relatively broad view, bias may still exist.

Additionally, data cleaning is necessary here and involves two main steps. First, we need to standardize the values in the `lead` and `result` columns by converting all logical values to uppercase, ensuring consistency; that is, converting `True` and `False` to `TRUE` and `FALSE`, respectively. Second, we will create a new variable, `result_binary`, which will be set to 1 when `result = TRUE`, and 0 when `result = FALSE`.

Here are the first five rows of the new dataset `shot`:

```
##   top left         date     qtr time_remaining result shot_type distance_ft
## 1 310  203 Oct 18, 2022 1st Qtr          09:26  FALSE         3          26
## 2 213  259 Oct 18, 2022 1st Qtr          08:38  FALSE         2          16
## 3 143  171 Oct 18, 2022 1st Qtr          08:10  FALSE         2          11
## 4  68  215 Oct 18, 2022 1st Qtr          05:24   TRUE         2           3
## 5  66  470 Oct 18, 2022 1st Qtr          01:02  FALSE         3          23
##    lead lebron_team_score opponent_team_score opponent team season color
## 1 FALSE                2                   2      GSW  LAL   2023   red
## 2 FALSE                4                   5      GSW  LAL   2023   red
## 3 FALSE                4                   7      GSW  LAL   2023   red
## 4 FALSE               12                  19      GSW  LAL   2023 green
## 5 FALSE               22                  23      GSW  LAL   2023   red
##   result_binary
## 1             0
## 2             0
## 3             0
## 4             1
## 5             0
```

Methods

Generalized Additive Model, Introduction

The first model I choose to fit is a generalized additive model (GAM).

First, a GAM is a good choice because it flexibly models the nonlinear relationships between predictors and the probability of a made shot. In other words, unlike a simple linear model, a GAM can capture complex patterns without manually specifying polynomial / interaction terms. Additionally, since we have a binary outcome (i.e. made vs. missed shots), the logistic link function in a GAM naturally extends logistic regression to more flexible shapes.

Second, a GAM can handle both categorical and continuous predictors smoothly, making it well-suited to incorporate predictors like quarter, shot type, and location within the same model.

Last, the interpretability of the partial response functions plot of GAM is very helpful for understanding how each predictor or combination of predictors influences the shot probability. In simpler terms, a GAM has a good balance between flexibility and interpretability.

Generalized Additive Model, Explanation

The formula of the GAM is:

```
result_binary ~
s(distance_ft) + s(top, left) +
factor(shot_type) + factor(qtr) + factor(lead).
```

That is, we predict the "Result" from "Distance", "Court Location", "Shot Type", "Quarter", and "Lead".

Why these predictors?

1. Distance (distance_ft)

Shot distance is one of the strongest predictors of shooting success. Longer shots are typically more difficult. The relationship need not be linear, and a smooth term `s(distance_ft)` can capture this nonlinearity.

2. Court Location [(top, left)]

Shots from different areas have different success rates. For example, "corner three-pointers" differ from "above the break three-pointers". A two-dimensional smooth term `s(top, left)` can capture these spatial patterns.

3. Shot Type (shot_type)

There is a fundamental difference in accuracy and intent between two-point attempts and three-point attempts, so including it as a factor variable `factor(shot_type)` is important.

That is, we factor `shot_type` because it represents two fundamentally different categories. There is a unique jump in difficulty between two-point and three-point attempts, and players have different intentions on two-point and three-point attempts.

Moreover, it is also worth noting that leaving `shot_type` as numeric may confound the effect of `distance_ft`.

4. Quarter (qtr)

Shooting percentages may differ by quarter (1st, 2nd, 3rd, 4th, OT). That is, late-game tension or fatigue could matter. Thus, including it as a factor variable `factor(qtr)` is crucial.

In other words, similarly to `shot_type`, `qtr` is best treated as a discrete category because each quarter can have unique conditions.

5. Lead (lead)

Players may face different defensive pressure and change their shot selection strategy, so including it as a factor variable `factor(lead)` is useful.

Why `s(top, left)`, not `s(top) + s(left)`?

`s(top, left)` creates a two-dimensional smooth surface that captures how shot probability varies across the combined values of top and left. In contrast, `s(top) + s(left)` only models separate effects for each coordinate and cannot represent interactions.

In other words, `s(top, left)` allows the model to learn how different combinations of top and left positions affect shot outcomes. This is very crucial since specific court locations often have unique shooting patterns that cannot be explained by simply adding the two coordinates.

Therefore, `s(top, left)` is more reasonable for modeling basketball shot data, and if we only used `s(top) + s(left)`, we may miss critical hot spots or cold spots where players shoot differently.

Additionally, the answer to this question can also explain why we can have both `s(distance_ft)` and `s(top, left)` in the model. Distance and `(top, left)` capture different aspects of shot location – distance only measures how far away the shooter is, while `(top, left)` provides the spatial context on the court – so the model can benefit from both simultaneously without redundancy.

Multilevel Logistic Regression Model, Introduction

The second model I choose to fit is a logistic regression model with varying intercepts and slopes.

Varying Intercepts: Different teams may have different baseline probabilities of making a shot. Allowing each team to have its own intercept captures variations in overall shooting skill.

Varying Slopes: Teams may also differ in how shot distance affects their probability of success. For example, some teams may be less impacted by taking longer shots, while others may have a sharp decline in accuracy. Hence, we allow the coefficient of `distance_ft` to vary by team.

And again, we include `shot_type` because two-point and three-point shots typically have different success probabilities, so it captures a key difference in shot difficulty and player intention; we include `lead` because whether a team is ahead or behind can influence shot selection and player performance, reflecting both situational and psychological factors.

However, it is worth noting that this model would face some limitations, as the `shot` dataset is derived from the performance of a few players (as previously mentioned) – each associated with a limited number of teams. And thus, to some extent, we may understand teams as roughly equivalent to players in this context.

Multilevel Logistic Regression Model, Explanation

The formula of the multilevel model is:

```
result_binary ~
factor(shot_type) + factor(lead) + distance_ft +
(1 | team) + (0 + distance_ft | team),
```

or:

```
result_binary ~
1 + factor(shot_type) + factor(lead) + distance_ft +
(1 | team) + (0 + distance_ft | team).
```

(The two specifications above are functionally the same in R since, by default, R includes an intercept in the model.)

Response of the Model & What Probability Distribution It Follows

Let Y_{ij} represent the outcome (shot i by team j):

$$Y_{ij} = \texttt{result\_binary} = \begin{cases} 1 & \text{(shot made)} \\ 0 & \text{(shot missed)} \end{cases}$$

We assume:

$$Y_{ij} \sim \text{Bernoulli}(p_{ij}).$$

All of the Relevant Levels

Level One:

$$\log \frac{p_{ij}}{1-p_{ij}} = \alpha_j + \beta_j \, \texttt{distance\_ft}_{ij} + \gamma_1 \, \texttt{shot\_type}_{ij} + \gamma_2 \, \texttt{lead}_{ij}.$$

Where:

α_j is team j's intercept (random intercept);

β_j is team j's slope on `distance_ft` (random slope);

γ_1 is the fixed effect for `shot_type`;

γ_2 is the fixed effect for `lead`.

Level Two:

$$\alpha_j = \alpha_0 + u_j$$

$$\beta_j = \beta_0 + v_j$$

$$\begin{bmatrix} u_j \\ v_j \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_v^2 \end{bmatrix} \right)$$

The Composite Model

$$\log \frac{p_{ij}}{1-p_{ij}} = \underbrace{\left(\alpha_0 + \beta_0 \, \texttt{distance\_ft}_{ij} + \gamma_1 \, \texttt{shot\_type}_{ij} + \gamma_2 \, \texttt{lead}_{ij}\right)}_{\text{fixed effects}} + \underbrace{\left(u_j + v_j \, \texttt{distance\_ft}_{ij}\right)}_{\text{random effects}}.$$

How many parameters are we estimating?

We need to estimate six parameters: α_0, β_0, σ_u^2, σ_v^2, γ_1, γ_2.

We do not estimate the covariance between the random intercept and slope terms since our model assumes it equals 0.

Results

Generalized Additive Model, Summary

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## result_binary ~ s(distance_ft) + s(top, left) + factor(shot_type) +
##     factor(qtr) + factor(lead)
##
## Parametric coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.90697    0.34051  -2.664  0.00773 **
## factor(shot_type)3  0.27020    0.32645   0.828  0.40784
## factor(qtr)1st Qtr  0.51898    0.31493   1.648  0.09937 .
## factor(qtr)2nd OT  -0.76866    1.13056  -0.680  0.49657
## factor(qtr)2nd Qtr  0.47598    0.31377   1.517  0.12927
## factor(qtr)3rd Qtr  0.47591    0.31435   1.514  0.13004
## factor(qtr)4th Qtr  0.42430    0.31396   1.351  0.17655
## factor(lead)TRUE    0.49034    0.06685   7.335 2.21e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                   edf Ref.df  Chi.sq p-value
## s(distance_ft) 5.251  6.405 113.147  <2e-16 ***
## s(top,left)    2.002  2.003   3.706   0.157
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

First, let's take a look at the summary. Here is the interpretation:

1. Parametric Coefficients:

– The "Intercept" is negative (-0.90697), indicating that, at the reference categories (`shot_type = 2`, `qtr = 1st OT`, and `lead = FALSE`), the log-odds of making a shot is below 0.

– `shot_type(3)` has a positive estimate (0.27020) but a high p-value (0.40784), so there is no strong evidence that three-point shots differ from two-point shots in odds of success, controlling for the other terms in the model. In other words, after accounting for the effects of other predictors, the adjusted odds suggest a non-significant advantage for three-pointers.

– The `qtr` coefficients are all positive except for `2nd OT` (-0.76866), but none are statistically significant (`p > 0.05`). This suggests that there is no distinct difference in shot-making probability by quarter.

– The `lead` factor shows that having a lead yields a significantly positive coefficient (0.49034 with `p = 2.21e-13 < 0.05`), suggesting higher odds of making a shot when the team is leading.

2. Smooth Terms:

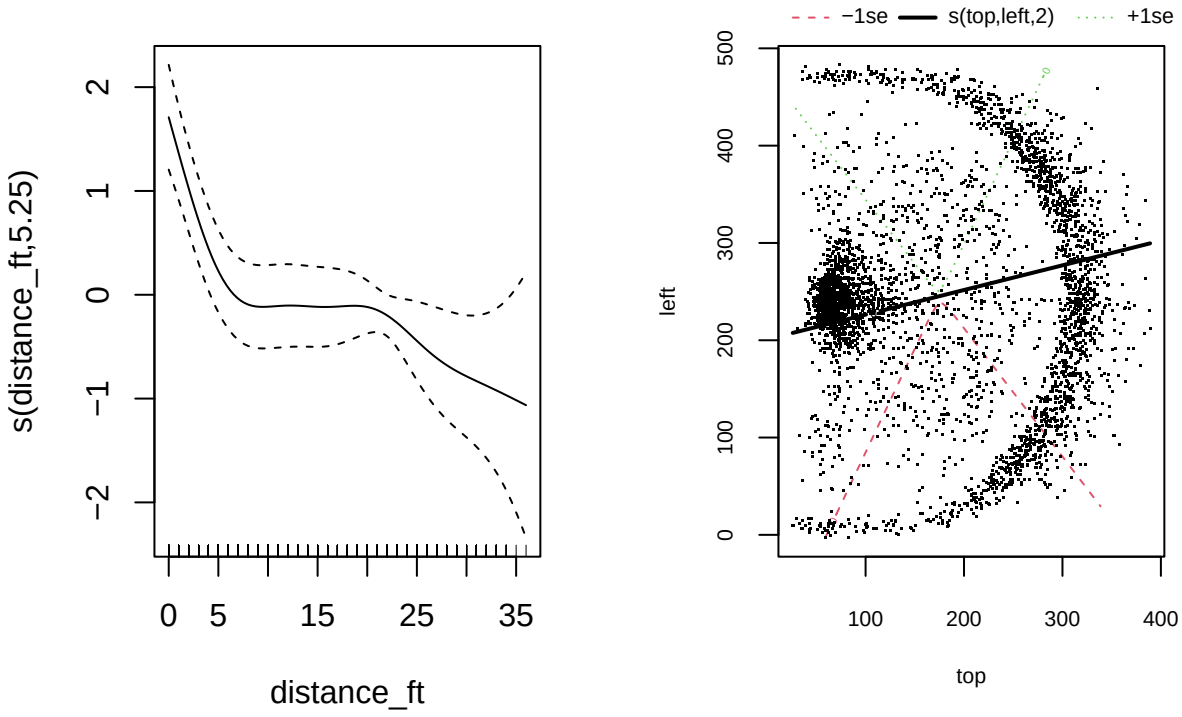
– `s(distance_ft)` has an EDF of 5.251, a highly significant p-value (`p < 2e-16 < 0.05`), and a large chi-square (113.147). This implies that distance from the hoop is a very important predictor of whether a shot goes in.

– `s(top, left)` has an EDF of 2.002 and a p-value of 0.157, indicating that the two-dimensional location on the court does not have a statistically strong effect beyond distance in this model.

3. Conclusion

Overall, the summary shows that distance from the hoop is by far the most crucial predictor, the effect of leading is significantly positive, and the 2D location does not appear strongly predictive once the distance is considered.

Generalized Additive Model, Visualization



Next, let's take a look at the partial response functions.

The panel on the left shows the partial effect of distance on the log-odds of making the shot, while the panel on the right is a visualization of `(top, left)`.

1. In the left plot:

The solid line is the estimated smooth effect of `distance_ft` on the log-odds of a made shot, and the dashed lines represent the approximate confidence interval.

The odds of making the shot decrease fairly rapidly with a distance from about 0 to 7 feet, then become stable until about 20 feet, and the decline becomes a bit steeper again to the end (though not as steep as the initial drop, and with wide error margins).

2. In the right plot:

This is a slice of the surface `s(top, left)` along with the raw shot locations. Since this term is not highly significant, the estimated partial effect is relatively flat; this also means that once the distance is accounted for, there is no strong additional pattern in the `(top, left)` coordinates.

Generalized Additive Model, Application & Uncertainty

Table 1: Predicted Probability of Making a Shot with 95% CI

Probability	Lower_95_CI	Upper_95_CI
0.4361842	0.3850354	0.487333

The last part accounts for the uncertainty estimates and the application of the GAM model.

We want to predict the probability of making a shot, with a 95% confidence interval, that:

1. Distance from the hoop = 25 feet;

2. Shot location = (top = 200, left = 100);

3. Three-pointer in the 4th quarter, leading when the shot was attempted.

Table 1 shows the results. The probability of making a shot is 0.4361842, the lower bound of the 95% CI is 0.3850354, and the upper bound is 0.487333.

Multilevel Logistic Regression Model, Summary

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: result_binary ~ factor(shot_type) + factor(lead) + distance_ft +
##     (1 | team) + (0 + distance_ft | team)
##    Data: shot
##
##      AIC      BIC   logLik deviance df.resid
##   5212.8   5250.6  -2600.4   5200.8     3986
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.8578 -0.8833 -0.5649  0.9472  2.2250
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  team   (Intercept) 0.098717 0.31419
##  team.1 distance_ft 0.000301 0.01735
## Number of obs: 3992, groups:  team, 3
##
## Fixed effects:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.45776    0.19644   2.330   0.0198 *
## factor(shot_type)3   0.59303    0.15081   3.932 8.41e-05 ***
## factor(lead)TRUE     0.50396    0.06659   7.568 3.79e-14 ***
## distance_ft         -0.06966    0.01226  -5.684 1.32e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. Random Effects:

– The random intercept for teams has a variance of 0.098717, corresponding to a standard deviation of 0.31419, which indicates that baseline log-odds of making a shot vary by 0.31419 across teams.

– The random slope for `distance_ft` shows a variance of 0.000301, corresponding to a standard deviation of 0.01735, suggesting that there is only a small difference across teams in how distance from the hoop affects the log-odds of making a shot.

2. Fixed Effects:

– The "Intercept" is 0.45776, implying that at zero distance, with a two-point shot and not in the lead, the log-odds of making a shot is 0.45776.

– For `shot_type`, the estimate for `factor(shot_type)3` is 0.59303, meaning given distance and other variables, three-point shots have 0.59303 higher log-odds of success compared to two-point shots. Moreover, this is statistically significant since `p-value = 8.41e-05 < 0.05`.

– Regarding `lead`, the estimate for `factor(lead)TRUE` is 0.50396, indicating that being ahead increases the log-odds of making a shot by 0.50396. Additionally, this is statistically significant since `p-value = 3.79e-14 < 0.05`.

– Finally, the `distance_ft` effect is -0.06966, meaning each additional foot from the hoop reduces the log-odds of success by 0.06966, which is a `1 - exp(-0.06966) = 0.06728911` decrease in odds per foot. This is also statistically significant since `p-value = 1.32e-08 < 0.05`.

3. Overall, from the summary, we find:

– Being in the lead increases the likelihood of a successful shot.

– After controlling for other factors in the model, three-point shots appear to be more accurate than two-point shots. Moreover, it is important to note that I state "controlling for other variables" when explaining the `shot_type` variable in both models. It is because in practice, a three-pointer is almost always taken farther from the hoop. If we compare predicted probabilities at typical real distances for each shot type, the overall probability for the three-pointer is likely to be lower.

– Increased distance lowers the odds of success.

– There is a measurable variation across teams in their baseline shooting (random intercept), and a very small variation in how distance affects each team (random slope).

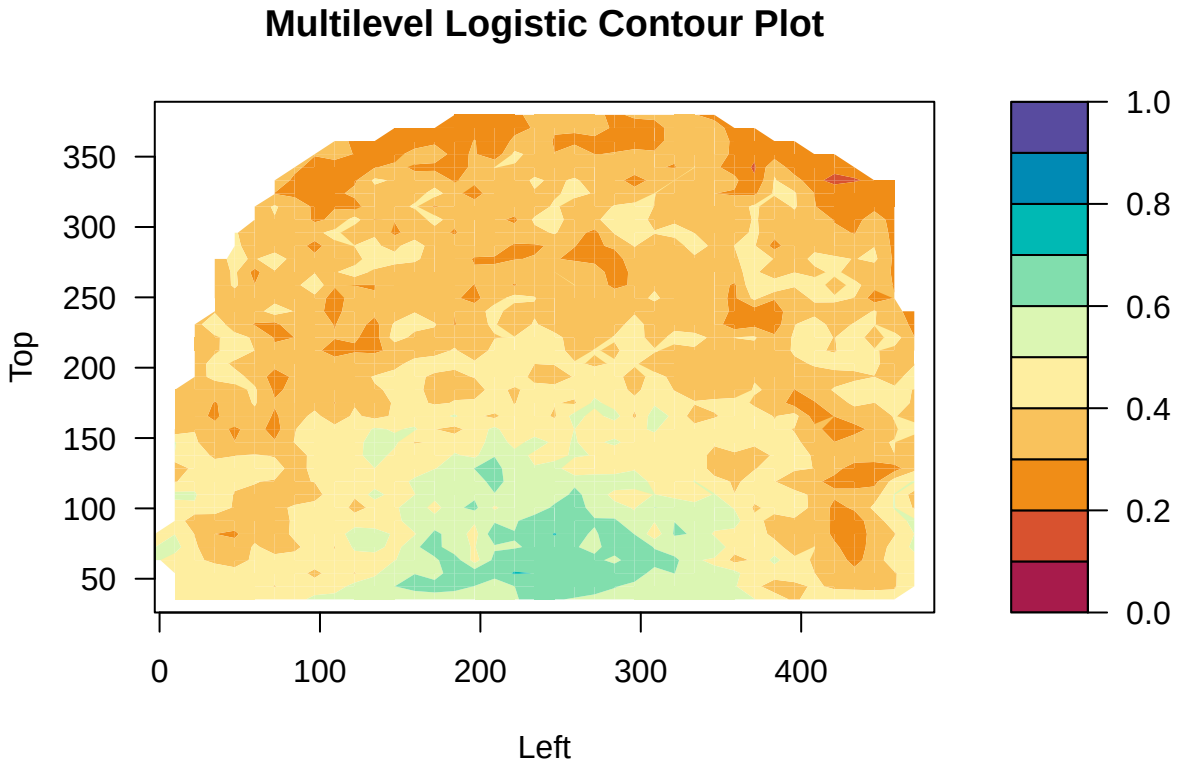
What Did We Learn? A Brief Conclusion for "Methods & Results, Part One"

We fitted two models, a generalized additive model (GAM) and a multilevel logistic regression model. Each of these models has its strengths and weaknesses. The generalized additive model is more broadly applicable but less accurate, primarily because many predictor estimates are not statistically significant. In contrast, the multilevel logistic regression model is more accurate, with statistically significant estimates for all predictors. However, as we mentioned before, this model faces limitations, since the data are derived from the performance of only a few players / teams.

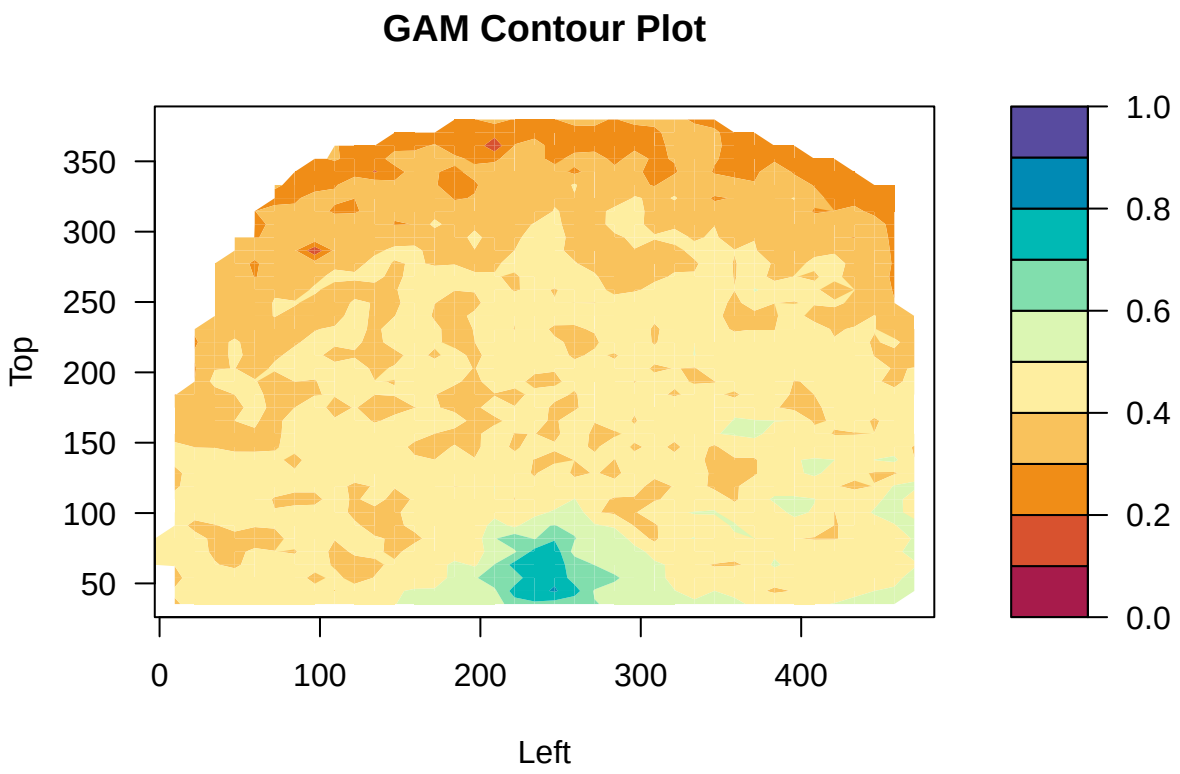
Moreover, both models indicate that "being in the lead" and "distance from the hoop" are critical factors influencing shot success, as these two variables are statistically significant in both models.

Visualizing Both Models

In order to better understand these models, lets make some contour plots. To make this contour plot, we simply got the predicted output of the models, and then interpolated them into a grid which we could plot.



For the Multilevel Logistic Regression Model, we can see there are not strict cut offs. There is much overlap for what areas are for each range of probabilities. We see that like earlier EDA results concluded, there is a hot spot around the hoop. We can also see the curve of the 3 point line represented with the curve near the top of the chart, where we couldn't interplate the data (due to there being insufficient observations there). We then continue and repeat this process with the GAM to get a better understanding of the model.



Here we see that once that once again there is much overlap with the levels, but less so than with the Logistic Regression Model. However, we see that around the hoop, our predictions seem to be much more uniform and consistent, while also being predicted as being more successful than with the Multilevel Logistic Regression Model. We also see that there seems to be a bit off a preference to the right side short corner, with a higher contour level being seen there.

Discussion

According to the statistical methods applied in this study, namely the generalized additive model (GAM) and the multilevel logistic regression model, we have constructed a framework to examine the factors that influence the probability of a successful basketball shot. These models allow for a detailed analysis of how various contextual variables, such as shot distance, spatial court location, game quarter, and whether the team was leading at the time of the shot, interact to shape outcomes.

The models were trained on shot-level data from three prominent National Basketball Association players, each with a distinct playing style and positional role: Stephen Curry, James Harden, and LeBron James. This sample captures a wide range of offensive strategies, including high-volume three-point shooting, midrange isolation play, and physical interior scoring. While this diversity in play style offers valuable variety in shot types and court zones, it also introduces several limitations in the generalizability of our conclusions.

We agree the dataset is inherently biased toward elite players whose shot selection and skill levels differ meaningfully from the average NBA player. The players represented are often tasked with taking high-difficulty or late-clock shots and they operate with a high degree of offense. As such, shot success probabilities derived from this dataset may be skewed toward scenarios involving elevated decision-making autonomy. This limitation poses a challenge in extending our findings to broader populations, such as defense-role players, developmental athletes, or amateur participants in training.

From a statistical perspective, the use of the GAM allowed us to model nonlinear relationships between predictors and the log odds of a made shot. The inclusion of smooth terms for shot distance and two-dimensional court location enabled the identification of curvature in the shot success function, which would not be detectable under a standard logistic model with linear effects. The estimated degrees of freedom for the distance-based smooth term was over five, indicating a relatively non-monotonic effect that aligns with basketball intuition: success probability falls sharply within the first few feet from the basket, levels off in the midrange, and decreases again with longer three-point attempts. However, the smooth term for spatial location did not yield statistical significance, suggesting that the information contained in exact court coordinates may be mostly captured by distance alone in the absence of defender data.

In parallel, the multilevel logistic regression model was deployed to incorporate group-level variability across teams. By introducing random intercepts and random slopes for shot distance at the team level, we aimed to account for systematic differences in team offensive systems and player-shot selection behavior. Although the model revealed modest variation in baseline shot success rates across teams, the random slope component for distance exhibited very low

variance. This suggests that across the few teams present in the dataset, the way distance impacts success probability is relatively uniform. The limited random effect observed also reinforces the importance of expanding the dataset in future analyses, as more variability may become apparent with a wider sample of teams.

To deepen our understanding of the dataset, EDA was also performed at the beginning. We examined shot types, distances, and positional trends visually through bar plots and scatterplots. This initial analysis reaffirmed several intuitive observations: two-point shots had a higher overall success rate compared to three-point attempts, and made shots clustered more densely near the basket. Binning the court into segments along the left and top dimensions also revealed that the raw location of shots, when not adjusted for distance, displays clear density patterns of made and missed shots.

In addition to fitting the two models, we used their outputs to generate shot prediction plots across the court. By mapping predicted probabilities spatially, we were able to visualize how the GAM and multilevel logistic regression models differ in their predictive focus. The GAM prediction map appeared smoother due to the nonlinear terms, while the multilevel model produced a slightly more rigid probability surface reflecting its reliance on fixed and random linear effects. Bar plots comparing predicted success by shot type further highlighted the models' practical implications, showing that both models captured the known difficulty gap between two-point and three-point shots.

However, several challenges were identified when attempting to build alternative models. The dataset lacked detailed play-by-play context, including identifiers for specific players involved during each play, which prevented the construction of more lineup-level effects. This limitation suggests that more detailed data sources would be necessary for future extensions involving player-specific evaluation.

Both models consistently confirmed the statistical significance of two key variables: shot distance and whether the player’s team was leading at the time of the attempt. The positive coefficient associated with the lead variable suggests that performance may be modulated by psychological or strategic conditions. Teams that are ahead may face less defensive pressure or may take better shots, while players may also be more confident. This finding could be explored more rigorously using causal inference frameworks in future work.

Despite the usefulness of the dataset, additional variables not currently present would greatly enhance the robustness and scope of the models. Defender proximity, player fatigue, assist origin, and shot clock context are all potential features that could refine the analysis substantially. These features could allow for hierarchical modeling not only at the team level but also nested within player-defender pairings, offering deeper insight into matchup-based efficiencies.

Moreover, resampling strategies such as the bootstrap could be implemented to assess model stability across random draws of shot attempts. For Bayesian extensions, incorporating prior distributions on team-level effects could help stabilize inference in small-sample or high-variance contexts. Hierarchical shrinkage estimators may also reduce overfitting in random slopes when the number of levels is limited. In addition, assessing model calibration through log-loss or Brier score could provide insight into predictive accuracy beyond traditional classification metrics.

Finally, an extension of this study could involve the application of time-varying coefficients to account for within-player correlation over time. Such frameworks are particularly appropriate if longitudinal data over multiple seasons becomes available. Generalized estimating equations would allow for population-averaged estimates while accommodating repeated measures, thereby enabling the examination of changes in player behavior or efficiency over a career.

Conclusion

This study has examined the factors contributing to the success of basketball shots through the application of statistical models capable of capturing both nonlinearity and hierarchical structure. The analysis successfully addressed the original question of interest by identifying key drivers of shot success and by quantifying the role of contextual factors.

The key findings of this research include the following listed below.

First, shot distance remains the most consistent and influential predictor of success probability. The decline in shooting efficiency is steepest within the first several feet of the basket and continues more gradually over longer distances. This confirms longstanding basketball heuristics regarding shot quality. Second, being in the lead at the time of the shot attempt is associated with increased shot success, a pattern that may reflect reduced defensive intensity, psychological confidence, or optimized shot selection when playing with a scoring cushion. Third, spatial court location, while intuitively important, does not demonstrate statistical significance in our generalized additive model once shot distance is included. This may be due to the high correlation between distance and location, or the absence of other variables such as defensive alignment.

Looking forward, there are numerous ways this analysis can be expanded. With more comprehensive data, researchers could examine how shot success varies across defender matchups, team tactics, or pace of play. Exploring temporal effects such as changes in player behavior during close fourth quarters or in transition versus half-court sets could provide further clarity. Additionally, modeling interdependencies between teammates on the floor could reveal synergistic effects that influence shot quality.

From a practical standpoint, the insights derived from this study are relevant for a wide range of basketball players, professional and recreational. Professional teams can use similar models to inform shot selection guidelines, simulate game scenarios, or develop targeted player development programs. Coaches at the collegiate and high school levels could implement efficiency-driven training practices informed by data on high-probability zones and situational shooting trends. For recreational players, adopting a data-informed perspective could translate into improved decision-making during competitive or casual play. Understanding which factors significantly increase shot success may allow non-professional athletes to adjust their shot selection and practice habits more intelligently.

Ultimately, this research report underscores the value of statistical modeling in sports analytics and highlights the potential for data to inform not only professional competition but also the broader basketball community. By quantifying performance drivers and embedding them into a flexible modeling framework, we open the door to ongoing innovation in how basketball gameplay is taught and understood.