# Exploring the Interplay Between Shot Outcomes and Shot Selection for the Denver Nuggets in the 2022-2023 NBA Season

Amrinder Sehmbi
Department of Statistical Science, University of Toronto
STA303H1: Methods of Data Analysis II
Mohammad K.A. Khan
April 7, 2024

## Introduction

### Background

The Denver Nuggets 2023 NBA championship win is a significant milestone for the franchise, showcasing their resilience, skill, and strategic prowess. Examining their success provides insights for enhancing player development, refining play style, and informing decisions regarding player acquisition and defensive strategies for opposing teams.

Shot selection is crucial for basketball success, as successful shots lead to increased points and improved chances of winning. However, evaluating shot selection beyond the outcome can be challenging due to the fast pace of the game and various influential factors. Conducting statistical analysis allows for a comprehensive assessment of the impact of different factors on shot outcomes for the Nuggets.

### Literature Review

In "Optimal Shot Selection Strategies for the NBA," Fichman and O'Brien observed that losing teams favor more 2-point shots, particularly in the right center box, whereas winning teams opt for fewer 2-point shots and more 3-point shots. However, their analysis evaluated shot metrics individually, while our research integrates all predictors into a unified predictive model, focusing specifically on the current Denver team.

In "Basketball Shot Types and Shot Success in Different Levels of Competitive Basketball," Erculj and Strumbelj found that dunks are more common, while hook shots are less frequent, and situational variables consistently influence shot selection. In contrast, our analysis concentrates on predicting shot success, differing from their model's output of probabilities using multinomial regression.

In "Performance Analysis of Game Dynamics During the 4th Game Quarter of NBA Close Games," Gomez, Gasperi, and Lupo highlighted the variability of home court advantage in close-margin games using linear regression. Our analysis seeks to predict a binary outcome rather than a quantitative one, while also considering quantitative predictors.

## Methods

### Data Set and Features

The dataset utilized in this study is the "NBA play-by-play and shot details data (1996-2022)" sourced from Kaggle. We focused on the shotdetail_2022.csv file and filtered for only records of the Denver Nuggets. We narrowed our analysis to variables supported by our literature review. For further information on the selected features, please refer to figure [6] in the appendix.

### Model

For our analysis, we will be using a logistic generalized linear mixed model, a statistical model used for analyzing binary outcome data while accounting for the presence of both fixed and random effects. The following are assumptions of a logistic GLMM model:

$$Y_{ij}|U_i \sim Bernoullie(\pi_{ij})$$
$$logit(\pi_{ij}) = X_{ij}\beta + U_i$$
$$[U_1, U_2, \ldots, U_n] \sim MVN(0, \Sigma)$$

where , $Y_{ij}$ is the binary response for $j^{th}$ observation from $i^{th}$ group, $X_{ij}\beta$ is the fixed effect and $U_i$ is the random effect group of the model.

**Variable Selection**

Employing a backward stepwise selection approach, we will start with a full model with all variables and identify a subset of predictor variables that effectively capture variation in the outcome variable while mitigating the risk of overfitting. We will iteratively pruned variables based on variable insignificance, small change in coefficient estimates and standard errors, and small decrease in goodness of model fit assessed by the Area Under the Curve (AUC).

**Diagnostics**

Model diagnostics will assess influential observations using standardized measures like DFBETAS to gauge individual observation impact on parameter estimates and Cook's distance to summarize collective influence of higher-level units. We evaluate potential impact and consider removal of influential points, considering AUC value and assessing model interpretability via parameter estimates and standard errors.

## Results

**Description of the Data**

***1.EVENT_TYPE***

Figure [1] : Distribution of Shots Made vs Missed.

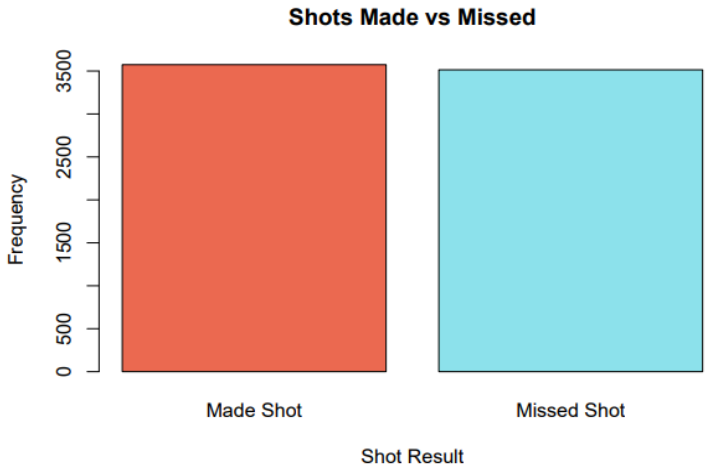


Figure [1] illustrates that the distribution of outcomes is almost evenly spread between shots made and shots missed and should not cause bias in our analysis. Hence, this satisfies the assumption of distribution for logistic GLMM.

***2. SHOT_ZONE_BASIC***

Figure [2]: Distribution of Shot Outcome by Zone

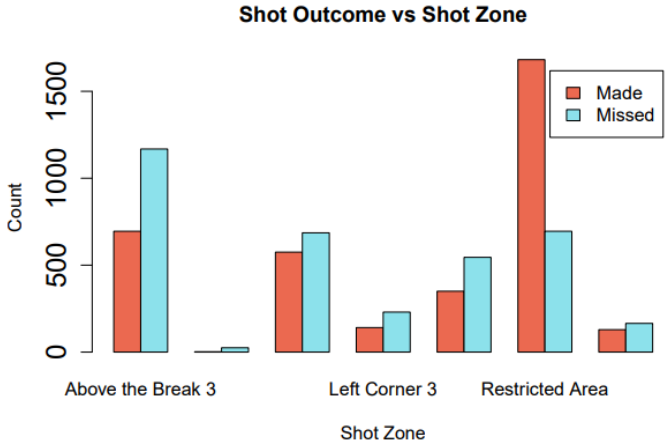


Figure [2] illustrates a notable disparity in shot outcomes across different shot zones. While most zones have approximately 40-50% of shots resulting in successful attempts, the restricted area presents an inverse pattern. Notably, the backcourt zone stands out due to its limited number of observations, and could be considered for exclusion from the analysis if deemed necessary.

***3. PLAYER_NAME and ACTION_TYPE***

Figure [3]: Mosaic Plot of Shots Made by Player and Action Type

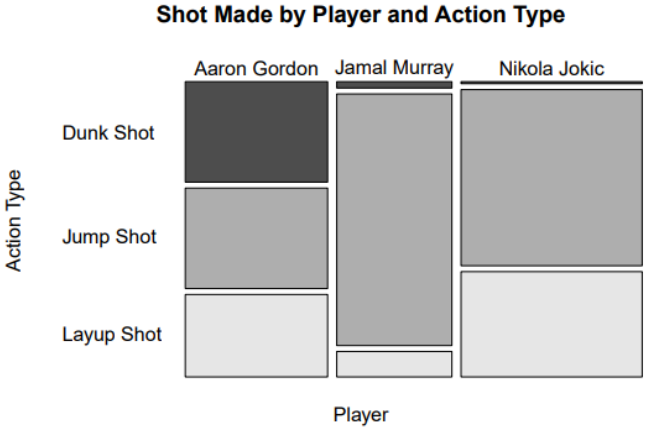

Figure [4]: Table of Summaries of Initial Model, Final Model and Final Model excluding Influential Observations.

| | Initial Model | | Final Model | | Model After Removing Influential Groups | |
|---|---|---|---|---|---|---|
| | Estimate | Pr(>\|z\|) | Estimate | Pr(>\|z\|) | Estimate | Pr(>\|z\|) |
| **Fixed Effect** | | | | | | |
| (Intercept) Above The Break 3 | 0.423348 | 0.23544 | -0.45375 | 0.00353 | -0.430134 | 0.02159 |
| SHOT_ZONE_BASICBackcourt | -1.02469 | 0.21915 | -2.01905 | 0.00669 | -0.670835 | 0.55653 |
| SHOT_ZONE_BASICIn The Paint (Non-RA) | -0.039717 | 0.87752 | 0.5447 | 0.0000282 | 0.527474 | 0.00114 |
| SHOT_ZONE_BASICLeft Corner 3 | -0.050119 | 0.69672 | 0.04923 | 0.68672 | 0.236563 | 0.55376 |
| SHOT_ZONE_BASICMid-Range | -0.202317 | 0.21384 | 0.12924 | 0.20464 | 0.028299 | 0.83264 |
| SHOT_ZONE_BASICRestricted Area | 0.857938 | 0.00733 | 1.59447 | < 2e-16 | 1.58687 | < 2e-16 |
| SHOT_ZONE_BASICRight Corner 3 | 0.224001 | 0.10449 | 0.33255 | 0.01207 | 0.007594 | 0.9851 |
| SHOT_DISTANCE | -0.037102 | 0.00877 | - | - | - | - |
| MINUTES_REMAINING | 0.001484 | 0.8438 | - | - | - | - |
| factor(HOME_AWAY)Home | 0.047821 | 0.35586 | - | - | - | - |
| factor(PERIOD)2 | -0.109496 | 0.12634 | - | - | - | - |
| factor(PERIOD)3 | -0.053919 | 0.45141 | - | - | - | - |
| factor(PERIOD)4 | -0.179013 | 0.01534 | - | - | - | - |
| factor(PERIOD)5 | 0.150878 | 0.72983 | - | - | - | - |
| **Random Effect** | | | | | | |
| ACTION_TYPE:PLAYER_NAME (Intercept) ICC | 0.023 | | 0.024 | | 0.03 | |
| ACTION_TYPE (Intercept) ICC | 0.09 | | 0.092 | | 0.099 | |
| **AUC** | 0.7102 | | 0.7093 | | 0.7342 | |

Figure [3] illustrates the variation in the distribution of shots made by player and action type. This observation suggests a dependency of our outcome of interest on these categorical variables. Therefore, it is reasonable to consider the inclusion of a random effect that incorporates these predictors.

## Analysis Process and Results

### *1. Variable Selection Results*

*Model 1: Initial Model*

The initial model consisted of all variables identified in the literature review, including SHOT_ZONE_BASIC, SHOT_DISTANCE, MINUTES_REMAINING, HOME_AWAY, and PERIOD as fixed effect predictors. Additionally, it incorporated a random effect of intercept, allowing for variability among ACTION_TYPE and PLAYER_NAME within ACTION_TYPE. This decision was made from our exploratory data analysis (EDA), revealing that certain players make shots with a certain action than others, while particular actions are more frequently successful overall.

The model's random effects gives intraclass correlation (ICC) values of 0.023 and 0.09, respectively, indicating the appropriateness of using a GLMM. Furthermore, the model demonstrated an AUC of 0.7102, indicating a satisfactory fit to the data.

*Model 2-4: Removing variables*

The initial model summary, as shown in Figure [4], led to the exclusion of the MINUTES_REMAINING predictor due to its high p-value of 0.8438, indicating insignificance. Coefficient and standard error estimates remained largely unchanged, suggesting minimal impact on other predictors' interpretation. The model's AUC remained stable at 0.7102, indicating no significant decline in goodness of fit. Similarly, in Model 3, HOME_AWAY was excluded due to its high p-value of 0.35457, with minimal changes in coefficient and standard error estimates, and a stable AUC of 0.7101. In Model 4, SHOT_DISTANCE was removed due to high correlation with SHOT_ZONE_BASIC. Despite this, the AUC remained relatively stable at 0.71. Finally, in the final Model 5, PERIOD was excluded due to all

coefficients having p-values less than 0.05, with no significant changes in coefficient estimates, standard errors, or AUC (0.7093). Further removal of variables was deemed inappropriate as it would substantially reduce AUC and compromise interpretability for answering the research question.

### *2. Diagnostics Results*

Figure [5]: Plot of Influential Groups Determined Using Cook's Distance.

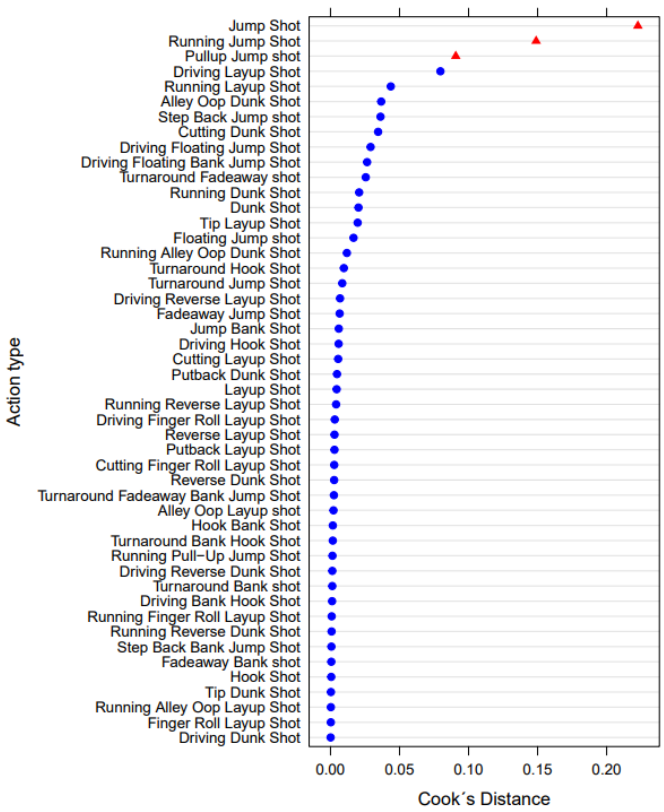


Figure [5] shows the result of the diagnostics using Cook's distance and indicates that "Jump Shot", "Running Jump Shot" and "Pullup Jump Shot" are significant influential categories of the ACTION_TYPE group predictor. After removing these categories from the data, we re-estimated our final model and recognized that the AUC significantly increased to 0.7342(Figure[4]). But, we noticed that this removes around half of the observations leading to distorted parameter estimates and inflated standard errors causing lack of interpretability in the context of our question.

## Discussion

### Final Model Interpretation and Importance

Figure [4] provides the coefficient estimates of the final model which can be used to determine odds of making a shot from a particular zone. Overall, the model suggests that a Nuggets player has higher odds of successfully making a shot taken in a particular way from the right corner compared to other zones, excluding the restricted area. For instance, holding the player shooting and shot type taken constant, the odds of the Nuggets making a shot from the right corner 3 are $e^{0.33255} = 1.4$ times the odds of making it from above the break 3. The odds of the Nuggets making a shot from right corner 3 are $e^{0.33255-0.04923} = 1.33$ times the odds of making a left corner 3 for a given player taking a certain type of shot. The odds of the Nuggets making a shot from right corner 3 are $e^{0.33255-0.12924} = 1.23$ times the odds of making a mid range for a given player taking a certain type of shot. These insights are crucial for developing game strategies that prioritize opportunities for players to take shots from the right corner 3. Furthermore, they provide opposing teams with valuable information to strategize defensively, ensuring their best defender is positioned effectively at the right corner.

### Limitations of the Analysis

One limitation of this analysis is the utilization of AUC for variable selection. While AUC serves as a measure of goodness of fit, it lacks the capacity to validate the model. Consequently, reliance on AUC alone could lead to overfitting the model on the training data but could fail to generalize to unseen data.

Furthermore, the decision to include influential observations could be another limitation. Although these observations constitute only three categories out of over 20, they collectively represent approximately half of the dataset. This imbalance could introduce bias into the coefficient estimates, compromising the reliability of the results.

# Appendix

Figure [6]: Table of Variables Incorporated in Analysis, Chosen From Literature Review.

| Feature | Description |
|---|---|
| ACTION_TYPE | Categorical data labeling the type of shot taken |
| PLAYER_NAME | Categorical data that gives the players full name. |
| SHOT_ZONE_BASIC | Categorical data describing the area on the court relative to the basket |
| MINUTES_REMAINING | Numerical data describing the minutes remaining in a period |
| PERIOD | Categorical data denoting game quarter the shot was attempted |
| SHOT_DISTANCE | Numerical data describing distance of shot from basket |
| HOME_AWAY | Categorical data for shot taken a home or away |
| EVENT_TYPE | Binary data indicating if a shot was successfull |


Figure [7]: Plot of Influential Groups Determined by Dfbetas, (Similar to Cook's Distance)


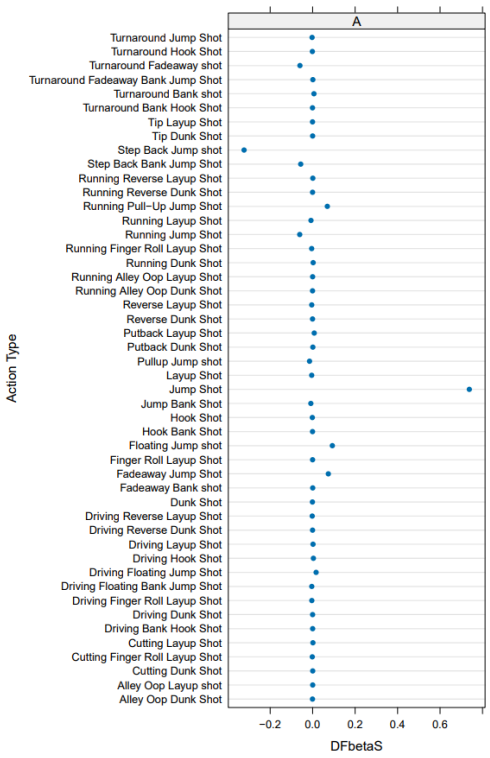

Figure [8]: Table of Summaries of Model 2,3,4, Discussed in the Paper

| | Model 2 | | Model 3 | | Model 4 | |
|---|---|---|---|---|---|---|
| | Estimate | Pr(>\|z\|) | Estimate | Pr(>\|z\|) | Estimate | Pr(>\|z\|) |
| **Fixed Effect** | | | | | | |
| (Intercept) Above The Break 3 | 0.4328 | 0.22072 | 0.44857 | 0.20407 | -0.36604 | 0.02334 |
| SHOT_ZONE_BASICBackcourt | -1.03038 | 0.2161 | -1.03986 | 0.21257 | -2.03572 | 0.00624 |
| SHOT_ZONE_BASICIn The Paint (Non-RA) | -0.04017 | 0.87607 | -0.03366 | 0.89603 | 0.5417 | 3.13E-05 |
| SHOT_ZONE_BASICLeft Corner 3 | -0.05008 | 0.69689 | -0.04748 | 0.71188 | 0.05663 | 0.64287 |
| SHOT_ZONE_BASICMid-Range | -0.20247 | 0.21332 | -0.20065 | 0.21762 | 0.12671 | 0.21376 |
| SHOT_ZONE_BASICRestricted Area | 0.85666 | 0.00737 | 0.86452 | 0.00685 | 1.59055 | < 20-16 |
| SHOT_ZONE_BASICRight Corner 3 | 0.22427 | 0.10404 | 0.22596 | 0.10144 | 0.32477 | 0.01431 |
| SHOT_DISTANCE | -0.03717 | 0.00859 | -0.03658 | 0.00966 | - | - |
| MINUTES_REMAINING | - | - | - | - | - | - |
| factor(HOME_AWAY)Home | 0.04795 | 0.35457 | - | - | - | - |
| factor(PERIOD)2 | -0.1097 | 0.1256 | -0.11025 | 0.1237 | -0.11175 | 0.11851 |
| factor(PERIOD)3 | -0.054 | 0.45071 | -0.05423 | 0.44873 | -0.05804 | 0.41722 |
| factor(PERIOD)4 | -0.17893 | 0.01539 | -0.17976 | 0.0149 | -0.183 | 0.01314 |
| factor(PERIOD)5 | 0.14578 | 0.73814 | 0.15534 | 0.72152 | 0.15683 | 0.71898 |
| **Random Effect** | | | | | | |
| ACTION_TYPE:PLAYER_NAME (Intercept) ICC | 0.023 | | 0.023 | | 0.023 | |
| ACTION_TYPE (Intercept) ICC | 0.09 | | 0.091 | | 0.093 | |
| **AUC** | 0.7102 | | 0.7101 | | 0.71 | |

## References


[1] Vladislav Shufinskiy. (2023). NBA play-by-play and shotdetails data (1996-2022). Retrieved Mar 11, 2023 from https://www.kaggle.com/datasets/brains14482/nba-playbyplay-and-shotdetails-data-19962021?select=shotdetail_po_2022.csv.

[2] Fichman, M., & O'Brien, J. R. (2019). Optimal shot selection strategies for the NBA. Journal of Quantitative Analysis in Sports, 15(3),203–211. https://doi.org/10.1515/jqas-2017-0113

[3] Erculj, F., & Strumbelj, E. (2015). Basketball shot types and shot success in different levels of competitive basketball. PloS One, 10(6), e0128885–e0128885. https://doi.org/10.1371/journal.pone.0128885

[4] Gomez, M. A., Gasperi, L., & Lupo, C. (2016). Performance analysis of game dynamics during the 4th game quarter of NBA close games. International Journal of Performance Analysis in Sport, 16(1), 249–263. https://doi.org/10.1080/24748668.2016.11868884