Report - Probabilistic ML Project

Samuel Rost

August 15, 2025

1 Overview and Motivation

The primary objective of this study is to model and predict the outcomes of individual shot attempts in the National Basketball Association (NBA) using detailed shot-level data from the 2014–15 season. The goal is to explore wether predicting shot-outcomes based on shot-specific contextual attributes is feasible. The analysis focuses on quantifying player-specific tendencies in shooting performance while simultaneously accounting for the influence of situational and contextual variables, such as shot distance, defender proximity, and time remaining on the shot clock.
Unlike conventional approaches that estimate a single set of effects for all players, this study employs a Bayesian hierarchical logistic regression framework in which both the baseline shooting ability and the sensitivity to contextual factors are modeled separately for each player. This allows the model to capture not only whether certain players are generally more accurate than others, but also how different players respond differently to game situations.
Formally, let $y_{i,j} \in \{0, 1\}$ denote the binary shot outcome (1 = made, 0 = missed) for the i-th shot taken by player j, and let $x_{i,j}$ represent a vector of contextual covariates. The probability of a made shot is modeled as:

$$P(y_{i,j} = 1 \mid \alpha_j, \boldsymbol{\beta}_j, \mathbf{x}_{i,j}) = \sigma\left(\alpha_j + \mathbf{x}_{i,j}^{\top}\boldsymbol{\beta}_j\right),$$

where $\boldsymbol{\sigma}\,(\boldsymbol{z}) = \frac{1}{1+e^{-z}}$ is the logistic function,
$\boldsymbol{\alpha_j}$ is the player-specific intercept (baseline ability), and
$\boldsymbol{\beta}_j$ is the player-specific coefficient vector capturing the influence of contextual features.

A hierarchical prior structure enables partial pooling across players:

$$\boldsymbol{\alpha}_j \sim \mathcal{N}(\mu_\alpha, \tau_\alpha^2), \quad \boldsymbol{\beta}_j \sim \mathcal{N}(\mu_\beta, \Sigma_\beta)$$

with hyperpriors placed on the population-level parameters (μ_α, τ_α^2, μ_β, Σ_β). This approach leverages information across all players, mitigating overfitting for those with fewer shot attempts while still allowing for substantial individual differences. To minimize the risks of overfitting To further reduce the risk of both overfitting and underfitting, and to ensure a sufficient sample size for reliable parameter estimation, the analysis is restricted to the 30 players with the highest number of shot attempts in the 2014–15 NBA season. Although the original dataset contains nearly all shots attempted during that season, complete with extensive contextual annotations, only the subset corresponding to these top 30 players is retained for model training. The distribution of shot attempts across all players, as well as the identification of the top 30 players by total attempts, is depicted in Figure 1.
The primary modeling goals are twofold:

1. **Predictive performance:** Accurately estimate the probability of a made shot, with explicit uncertainty quantification.
2. **Interpretability:** Characterize how individual players' performance changes as contextual conditions vary (e.g., defender proximity, shot clock pressure).

This framework allows the following guiding questions to be addressed:

- To what extent do players differ in their response to situational variables?
- Which contextual features exert the greatest influence on shooting probability, and do these effects vary systematically across players?

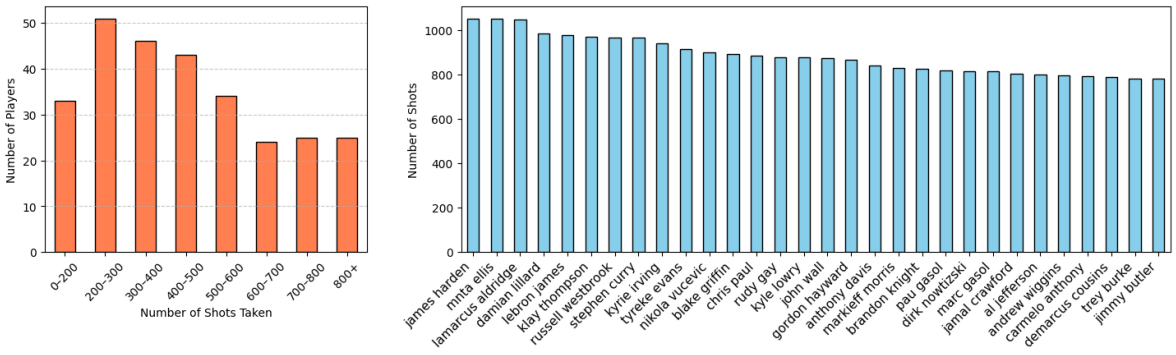

Figure 1: Distribution of Players by Total Number of Shots (left) and Top 30 Players by Number of Shots (right)

The potential contributions are both methodological and practical. From a methodological perspective, the hierarchical Bayesian structure illustrates the value of multilevel modeling for sparse, imbalanced sports datasets. From a practical perspective, the results could inform tactical decisions and quantifying context-specific player strengths and weaknesses.

2 Data Description

2.1 Dataset Overview

The dataset utilized for this study comprises shot-level data from the 2014–15 NBA season. It contains a total of 128.069 rows, each representing a single shot attempt, and includes 21 features capturing a wide range of contextual, situational, and performance-related information. Each record is annotated with identifiers such as `GAME_ID` and `player_id`, alongside shot-specific features `SHOT_DIST` (distance to basket), `SHOT_CLOCK` (remaining shot clock time), `TOUCH_TIME` (time the player had possession of the ball before shooting), `DRIBBLES` (number of dribbles before shooting) and `CLOSE_DEF_DIST` (distance to closest defender).
Additional contextual variables include `PERIOD` (game period 1-4), `GAME_CLOCK` (time remaining in period), and `FINAL_SCORE_MARGIN` (point differential at game end), all capturing the dynamic state of the game at shot time.
The target variable for the predictive model is the binary indicator `FGM` (Field Goal Made), which encodes whether the shot was successful.
For the full overview of features and their explaination/description see the `Notebook` section **1**.

2.2 Data Validation and Cleaning

A thorough data validation process was conducted to ensure data integrity and suitability for modeling. This included: Missing values: The dataset was largely complete, with the exception of the `SHOT_CLOCK` variable, which exhibited 5567 missing entries. These missing values were addressed by filling in the average shot-clock-value across all shot-attempts.
Format consistency: All temporal variables such as `GAME_CLOCK` adhered to the expected `mm:ss` format. Categorical variables like `LOCATION` and `PERIOD` were checked for valid entries and found to be consistent as well.
Logical consistency: Cross-validation between `FGM` and `SHOT_RESULT` confirmed agreement, ensuring that binary and categorical labels for shot outcomes were synchronized. Additionally, numerical sanity checks were performed to confirm non-negativity of distance measures and alignment of points scored (PTS) with shot type (`PTS_TYPE`).
For further detail see `Notebook` section **2.2.1**

2.3 Feature Engineering

To enhance the model's ability to capture the complexity of shot outcomes, several new features were engineered based on domain knowledge and preliminary data exploration (Notebook 2.2.2):
Clutch time indicator (`CLUTCH_TIME`): Binary variable indicating shots taken in high-pressure scenarios defined as fourth period or later, within 4 minutes remaining, and a score margin less than 10 points. Shooting zones (`SHOT_ZONE`) constitute a categorical variable obtained by discretizing `SHOT_DIST` into basketball-specific regions such as Restricted Area, Paint, Mid-Range, and beyond the Three-Point Line, thereby capturing spatial shooting tendencies (see Figure 2). For modeling purposes, this variable was one-hot encoded to enable its integration into regression models, which cannot directly process categorical predictors.

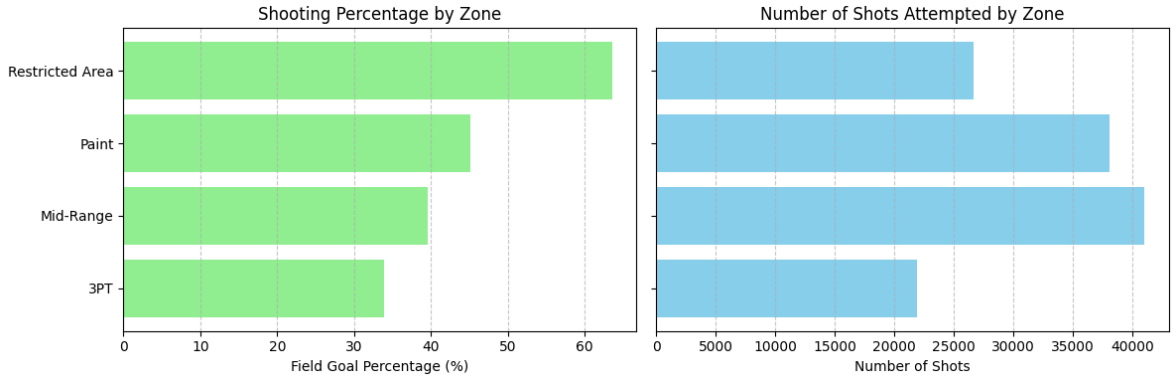

Figure 2: Shot Distribution and Efficiency by Zone (see Notebook section 3.3)

Defensive pressure (`DEFENDER_PROXIMITY`): Inverse of defender distance (1 / (`CLOSE_DEF_DIST` + 0.1)), serving as a continuous proxy for defensive intensity at the moment of the shot.
Time pressure (`TIME_PRESSURE`): Indicator of urgency derived from the shot clock, set to 1 if fewer than 3 seconds remained on the shot clock.

3 Patterns and key-insights into dataset

An examination of shooting performance across the distinct one-hot encoded shooting zones (restricted area, mid-range etc.) (Figure 2) highlights the pronounced impact of shot distance on field goal efficiency. The league-wide averages reveal a clear dichotomy: while 2-point shots convert at 48.8%, 3-point attempts succeed at just 35.2%, resulting in an overall field goal percentage of 45.2%. Attempts in the restricted area achieve the highest conversion rates (58.1%), reflecting the proximity to the basket and generally higher shot quality. Efficiency declines progressively with increasing distance, with mid-range shots (39.7%) exhibiting substantially lower percentages than the 2-point average, and three-point attempts demonstrating the steepest drop in accuracy. In contrast, the distribution of shot attempts reveals a concentration not only in the restricted area but also in the mid-range and paint zones, indicating that shot frequency is influenced by both tactical considerations and in-game circumstances. This dual perspective underscores the trade-off between shot value and shot difficulty inherent in different zones of the court.

In addition to shot distance, the most influential contextual factors affecting shooting performance is the distance to the closest defender. While greater defensive separation is generally associated with higher shooting efficiency, this relationship is confounded by the fact that longer-distance shots — often taken when defenders are farther away — tend to have inherently lower success rates. As these two effects operate in opposing directions, analyzing defender proximity in isolation risks misrepresenting its true impact. To disentangle these dynamics, field goal percentages are stratified simultaneously by defender distance and shot type (2-point vs. 3-point attempts), as illustrated in Figure 3. This

approach reveals that, for both shot types, increased defensive distance yields measurable gains in efficiency, yet the magnitude of these gains varies considerably between two-point and three-point attempts. Furthermore, the average 2-point field goal percentage is depicted as a dashed blue line, and the average 3-point field goal percentage is depicted as a dashed orange line in the figure, providing a league-wide benchmark against which the observed values can be compared.

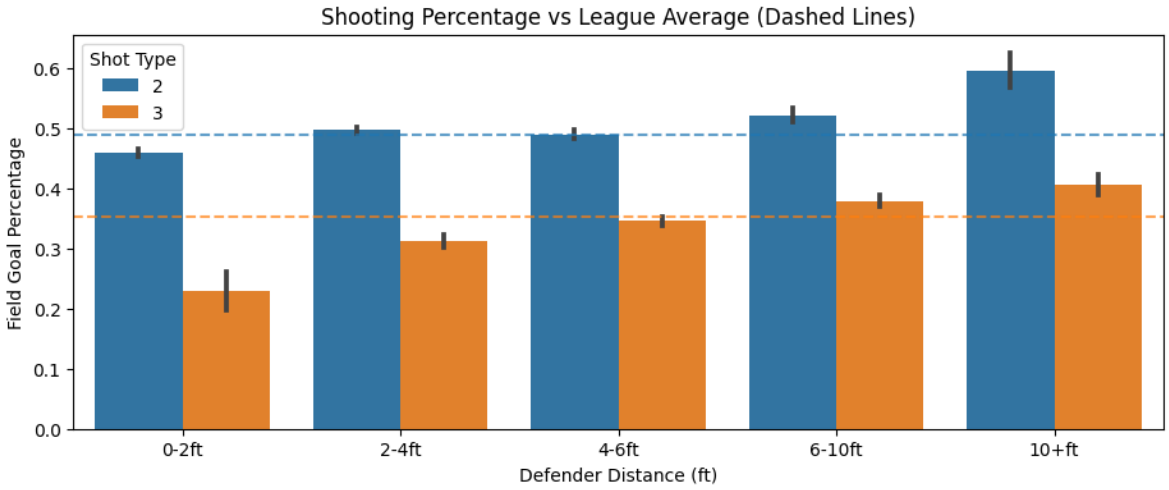

Figure 3: Shot-efficiency by defender distance (see Notebook section 3.4)

The correlation analysis (Notebook section 3.5) provides a quantitative perspective on the relationships between shooting performance and various game- and shot-specific features. When examining correlations with the target variable Field Goal Made (FGM), the strongest associations emerge for `SHOT_DIST` (–0.19) and the `RESTRICTED_AREA` indicator (+0.19), reflecting the well-established trade-off between shooting distance and accuracy. As expected, longer attempts, particularly from three-point range (–0.10), tend to reduce conversion likelihood, whereas attempts near the basket enhance it. Other temporal or contextual variables, such as `SHOT_CLOCK` (+0.10) and `TIME_PRESSURE` (–0.05), show weaker correlations, indicating a more modest direct influence on shot success. Overall, this correlation analysis reveals that no single feature exhibits a strong linear association with the target variable Field Goal Made (FGM), even the strongest relationships remain modest in magnitude.
When examining `CLOSE_DEF_DIST` (see Notebook), its strongest relationship is with `SHOT_DIST` (+0.52), confirming that open looks generally occur at longer ranges, as highlighted in the preceding section. Conversely, negative correlations with PAINT (–0.32) and `RESTRICTED_AREA` (–0.27) highlight that tightly contested situations are more common near the rim. Ball-handling metrics such as `TOUCH_TIME` (–0.16) and `DRIBBLES` (–0.15) display weak associations, Ball-handling metrics such as `TOUCH_TIME` (–0.16) and `DRIBBLES` (–0.15) show weak correlations with defender distance, that indicate that skilled players can create effective scoring opportunities through dribble penetration even in more tightly contested spaces near the rim.
The full correlation matrix (Figure 6) further indicates that, apart from the spatial link between `SHOT_DIST` and `CLOSE_DEF_DIST`, none of the remaining features display strong pairwise relationships. This low inter-feature correlation suggests that the variables included for modeling are largely complementary, reducing redundancy and supporting the stability of multivariate estimation.

To explore wether players exhibit a significant decline in shooting accuracy under high-pressure scenarios, I examined the clutch-performance of players to not only touch on spatial and statistical relationships but also on potential psychological factors. Clutch-time situations— defined here as the final four minutes of a game with a score margin of ten points or fewer—account for only 5.5% of all shots in the dataset (n = 7,008). In these high-pressure moments, league-wide shooting efficiency declines noticeably, with the overall clutch field-goal percentage (42.1%) falling 3.3 percentage points below the non-clutch average (45.4%). This performance drop is consistent across most shot zones, with the most pronounced reductions observed for three-point attempts (–5.0 percentage points) and mid-range shots (–3.2 percentage points), while efficiency in the restricted area remains relatively sta-

ble. The player-specific analysis (Figure 7) reveals considerable heterogeneity in clutch performance, even among high-volume shooters ($\geq$50 attempts), suggesting that psychological and situational factors may amplify individual differences under pressure.

4 Methodology

4.1 Bayesian Hierarchical Logistic Regression (Primary Model)

The primary modeling framework employed is a *Bayesian hierarchical logistic regression*, designed to capture both population-level shot-making tendencies and player-specific deviations. Logistic regression models the probability of a made shot $y_i \in \{0, 1\}$ via the log-odds transformation:

$$\log \frac{\pi_i}{1 - \pi_i} = \alpha_{j[i]} + \mathbf{x}_i^\top \boldsymbol{\beta}_{j[i]},$$

where $\pi_i = P(y_i = 1)$, $\mathbf{x}_i \in \mathbb{R}^p$ is the feature vector for shot i, and $j[i]$ denotes the player who took the shot.

To enable *partial pooling* across players and mitigate overfitting for those with fewer attempts, we impose hierarchical priors on the intercepts and coefficients:

$$\alpha_j \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2), \quad \beta_{j,k} \sim \mathcal{N}(\mu_{\beta_k}, \sigma_{\beta_k}^2),$$

with hyperpriors

$$\mu_\alpha \sim \mathcal{N}(0, 1), \quad \sigma_\alpha \sim \text{HalfNormal}(1), \quad \mu_{\beta_k} \sim \mathcal{N}(0, 1), \quad \sigma_{\beta_k} \sim \text{HalfNormal}(1).$$

Here, $\mu_\alpha, \mu_{\beta_k}$ represent *population-level means*, and $\sigma_\alpha, \sigma_{\beta_k}$ capture between-player variability. This structure balances *information sharing* across players with flexibility to model substantial individual differences.

The feature set includes shot-level spatial and contextual variables (e.g., `SHOT_DIST`, `CLOSE_DEF_DIST`, `DRIBBLES`, `TOUCH_TIME`, `PTS_TYPE`, `SHOT_NUMBER`, and situational indicators such as `HOME_GAME`, `LATE_GAME`, `CLUTCH_TIME`, `TIME_PRESSURE`). Continuous predictors are standardized, preserving interpretability of distance variables.

Bayesian inference proceeds via *Hamiltonian Monte Carlo* (HMC) with the No-U-Turn Sampler (NUTS), which augments parameters with auxiliary momentum variables and simulates Hamiltonian dynamics to efficiently explore the posterior. NUTS adaptively selects trajectory lengths and tunes step sizes, avoiding the inefficiency of random-walk sampling. We use 4 chains with 700 warm-up and 1,500 sampling iterations, targeting an acceptance rate of 0.9 for stable convergence.

The model implementation can be seen in Notebook 5.3.

On a held-out test set, the model achieves:

$$\text{Accuracy} = 59.18\%, \quad \text{ROC-AUC} = 0.606,$$

with a confusion matrix:

$$\begin{bmatrix} 2264 & 607 \\ 1561 & 879 \end{bmatrix}$$

From this, we obtain:

$$\text{Precision} = \frac{879}{879 + 607} \approx 59.2\%, \quad \text{Recall} = \frac{879}{879 + 1561} \approx 36.0\%, \quad \text{F1-Score} \approx 44.8\%$$

These metrics indicate a very limited predictive power. Most notably, the model's overall accuracy of 59.2% represents only marginal improvement over a naive baseline prediction where all shot attempts are classified as misses (which would yield approximately 55% accuracy given the dataset's class distribution). This modest performance enhancement suggests the model captures only minimal predictive signal beyond the inherent class imbalance. The model exhibits particular weakness in correctly identifying made shots (recall = 0.36), while being more accurate in identifying missed shots. This asymmetric performance pattern indicates the current feature set and modeling approach may

be insufficient for robust shot outcome prediction.

The additional feature diagnostics (**Notebook** section **3.7**) underscore the limited predictive capacity of the dataset. Variance Inflation Factors (VIF), displayed in table 2, reveal only moderate multicollinearity for `TOUCH_TIME` and `DRIBBLES` (VIF ≈ 7.4), while all other predictors remain well below the common threshold of 10, indicating that redundant linear dependencies are not the primary issue. Mutual information scores 3 are uniformly close to zero, with `SHOT_DIST` (0.015) being the only variable with even a marginal relationship to the outcome, confirming the absence of strongly informative single predictors.
Statistical significance tests seen in table 4 identify highly significant differences for spatial and temporal features (`SHOT_DIST`, `PTS_TYPE`, `TOUCH_TIME`, `TIME_PRESSURE`), but the small effect sizes and minimal mutual information indicate that these variables, while statistically distinct, are not practically decisive in classification. Contextual indicators such as `CLUTCH_TIME`, `SHOT_NUMBER`, and `LATE_GAME` fail to reach conventional significance, largely due to their rarity or weak aggregate impact.
These results suggest that the feature space lacks meaningful separation between made and missed shots (Figure 9), aligning with the modest performance of all tested models and reflecting the inherently stochastic nature of basketball shot outcomes.

4.2 Alternative Approaches

The limited predictive performance of the Bayesian hierarchical logistic regression suggests that the current feature set and modeling framework may not sufficiently capture the complexity of shot-making determinants in the NBA. To address this, two complementary strategies are pursued: (1) the development of more informative, domain-specific features through targeted feature engineering and selection, and (2) the evaluation of alternative modeling approaches as benchmarks to assess the attainable performance ceiling given the available data.

4.2.1 More Advanced Feature Engineering and Selection

To improve predictive performance beyond the baseline feature set, I implemented another feature engineering pipeline comprised two main steps: (i) creation of novel, context-aware predictors informed by both basketball domain knowledge and statistical considerations, and (ii) systematic feature selection via multiple complementary methods.

Feature Engineering We derived a range of new variables designed to capture spatial, temporal, and situational aspects of shot attempts that are not fully represented in the raw dataset. For instance, `SHOT_ANGLE` approximates the angular position of the shooter relative to the basket using court geometry, enabling the model to distinguish between central and corner shots. The `IS_HOT` indicator captures momentum effects by flagging situations in which a player has made multiple consecutive field goals within a game, reflecting potential confidence-driven performance boosts. Specifically `IS_HOT` is set to true if the last three shot attempts were successful.
To account for shooting form and consistency, we implemented `SMART_2P_PCT` and `SMART_3P_PCT`, which combine a player's career-average shooting percentage with their in-game performance. These features adaptively update once a minimum of three shots of the respective type have been taken in a game, thereby balancing early-game stability with responsiveness to current form. For more detail see Notebook 5.4.1.
Other engineered features included metrics such as `SHOT_RATIO` (shot distance relative to defensive proximity) and `DRIBBLE_SPEED` (touch time per dribble), `SHOT_DIFFICULTY` (a composite of distance, dribble count, and defender proximity), among many others. A complete list of all newly engineered variables and their precise definitions can be found in Notebook 5.4.1.

Feature Selection To identify the most informative variables, we applied four independent selection methods:

1. **SelectKBest** (univariate ANOVA F-test), selecting the ten highest-scoring predictors.
2. **Recursive Feature Elimination** (RFE) using logistic regression to iteratively remove the least relevant features.

3. **Random Forest** feature importance ranking, based on mean decrease in impurity.
4. **Lasso regularization** to shrink coefficients and select features exceeding a data-driven effect size threshold.

Results of Feature Selection While each feature selection method produced a slightly different top-10 list, certain predictors appeared repeatedly across multiple approaches. Variables such as `SHOT_DIST`, `RESTRICTED_AREA`, `3PT`, `LATE_CLOCK`, and `DEF_PRESSURE` were consistently selected by at least three methods. Interestingly, pretty much all of the newly engineered features were not often selected by the feature selection techniques. This suggests that, despite their theoretical relevance, these complex features did not provide significant additional explanatory power beyond the basic shot-related statistics! Consequently, the results show that these newly engineered features may be of low relevance for predicting shot outcomes in the current model, with the already established features still playing the dominant role in predictive accuracy.

4.2.2 Benchmark Models

To estimate the potential upper bound of predictive performance given the available features, three additional classification models were trained and evaluated (see Notebook section 5.4.2). These benchmarks serve as reference points for both linear and non-linear approaches, enabling comparison with the Bayesian hierarchical logistic regression.

Linear Discriminant Analysis (LDA) LDA assumes that observations from each class are drawn from a multivariate normal distribution with a class-specific mean vector $\boldsymbol{\mu}_k$ but a common covariance matrix Σ. Classification is based on maximising the discriminant function

$$\delta_k(\mathbf{x}) = \mathbf{x}^\top \Sigma^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log \pi_k,$$

where π_k denotes the prior probability of class k. This formulation yields linear decision boundaries in the feature space. The model achieved an accuracy, precision and recall of 60%, 37.4% and 58.4%, matching the Bayesian regression almost perfectly, indicating that the classes are not easily separable by linear means.

Random Forest Classifier The random forest classifier is an ensemble of B decision trees $\{T_b\}_{b=1}^{B}$, each trained on a bootstrap sample of the data and using random feature subsets at each split to decorrelate the trees. The final prediction is obtained via majority voting:

$$\hat{y} = \text{mode}\{T_1(\mathbf{x}), T_2(\mathbf{x}), \ldots, T_B(\mathbf{x})\}.$$

$$\hat{y} = \arg_k \max \left(\frac{1}{B} \sum_{b=1}^{B} \mathbb{I}(T_b(\mathbf{x}) = k) \right),$$

This method can capture non-linear interactions and high-order feature dependencies. However, despite its flexibility, the model performances worse reaching only 56% accuracy as well as a precision of 50.1% and recall of 44.5%, suggesting limited exploitable non-linear structure in the current feature set.

Neural Network The neural network benchmark consisted of a fully connected feed-forward architecture $d_{\text{in}} \to 128 \to 64 \to 32 \to 1$ implemented with dropout regularization:

$$\begin{aligned} f(\mathbf{x}) = \sigma_{\text{sigmoid}}(\mathbf{W}_4 \cdot \text{Dropout}_{0.2}\big(\sigma_{\text{ReLU}}(\\ \mathbf{W}_3 \cdot \text{Dropout}_{0.3}\big(\sigma_{\text{ReLU}}(\\ \mathbf{W}_2 \cdot \text{Dropout}_{0.4}\big(\sigma_{\text{ReLU}}(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1)\big) \\ + \mathbf{b}_2\big) + \mathbf{b}_3\big) + \mathbf{b}_4), \end{aligned}$$

where $\sigma_{\text{ReLU}}(z) = \max(0, z)$, $\sigma_{\text{sigmoid}}(z) = (1 + e^{-z})^{-1}$ and W_i and $\mathbf{b}_i$ denote the weight matrices and bias vectors for layer i. The model was trained with: Adam optimization ($\alpha = 0.0005$) - Early

stopping (patience=10) - Learning rate reduction (factor=0.5, patience=3) - Batch size=32 and 20% validation split.
The network achieved an accuracy of roughly 61% (precision=0.61, recall=0.32), indicating that additional representational (non-linear) capacity alone does not yield significant improvement.

4.2.3 Takeaway from the Alternative Approaches

The results from the benchmark models suggest that the performance of the Bayesian hierarchical logistic regression model, with an accuracy of 59.2%, is close to the maximum achievable with the available data. Despite extensive feature engineering and the use of other linear models like LDA or more advanced models such as Random Forest and Neural Networks, all models reached a similar performance, with accuracies around 60%. The precision and recall metrics remained similarly modest across all models, with no significant improvement despite the increased complexity.
This indicates that the feature set, even after incorporating more complex and context-aware variables, is insufficient to improve predictive performance beyond a certain threshold. There might be factors influencing shot outcomes that are not captured in the available features. As a result, improving model performance beyond this point appears unattainable with the current data.
Therefore, the key takeaway is that the performance ceiling for this problem is likely capped due to an inherent randomness in NBA shooting that leads to an intrinsic unpredictability of shot success regardless of model complexity.

5 Analysis of the Hierarchical Logistic Regression Model

5.1 Global and Player-Specific Effects of Contextual Variables

The Bayesian hierarchical logistic regression model estimates both global (population-level) effects and player-specific deviations. Figure 4 shows global average coefficients: Among all predictors, shot distance (SHOT_DIST) exhibits the largest magnitude, with a substantial negative coefficient, confirming the intuitive relationship that longer shots have a lower probability of success. Defender proximity (CLOSE_DEF_DIST) shows the second-strongest effect, positively associated with shot success—indicating that increased space from the nearest defender improves shooting outcomes. Smaller but still significant effects are observed for dribbles and touch time (TOUCH_TIME) as well as shot sequence position (SHOT_NUMBER), with the latter displaying a significant negative association. This suggests a possible influence of fatigue, as higher shot numbers generally occur later in games, when physical exertion may degrade shooting performance. Variables such as HOME_GAME, CLUTCH_TIME, and TIME_PRESSURE show minimal global influence, indicating substantial inter-player variability. The limited impact of clutch time at the population level is likely due to its rarity, as such situations occur only in closely contested games during the final minutes, reducing their overall statistical weight in determining shot success.
Figure 8 highlights this variability. While the SHOT_DIST effect is consistently negative, some players are less affected, likely due to long-range shooting skill. Similarly, CLOSE_DEF_DIST benefits most players, but the magnitude varies. Pressure-related variables (LATE_GAME, CLUTCH_TIME, TIME_PRESSURE) and HOME_GAME effects are centered near zero with wide distributions, indicating that some players thrive under these conditions while others decline.
These findings support the hypothesis that while certain contextual factors (distance, defensive pressure) have strong universal effects, many situational influences are player-specific. The heterogeneity in coefficients underscores the importance of individualized performance modeling and suggests that a significant portion of shot outcome variability is inherently stochastic, limiting predictive accuracy.

5.2 Player-Specific Coefficients: Damian Lillard vs. Marc Gasol

To illustrate the interpretive value of the hierarchical structure, Figure 5 compares player-specific coefficients for Damian Lillard and Marc Gasol—two players with contrasting profiles during the 2014–15 season. Lillard, an elite perimeter shooter and guard, shows a less negative effect for SHOT_DIST than Gasol, indicating greater efficiency from longer range. Both players benefit similarly from increased CLOSE_DEF_DIST, though Lillard's advantage is slightly smaller, suggesting that defensive pressure impacts them in comparable ways despite stylistic differences.

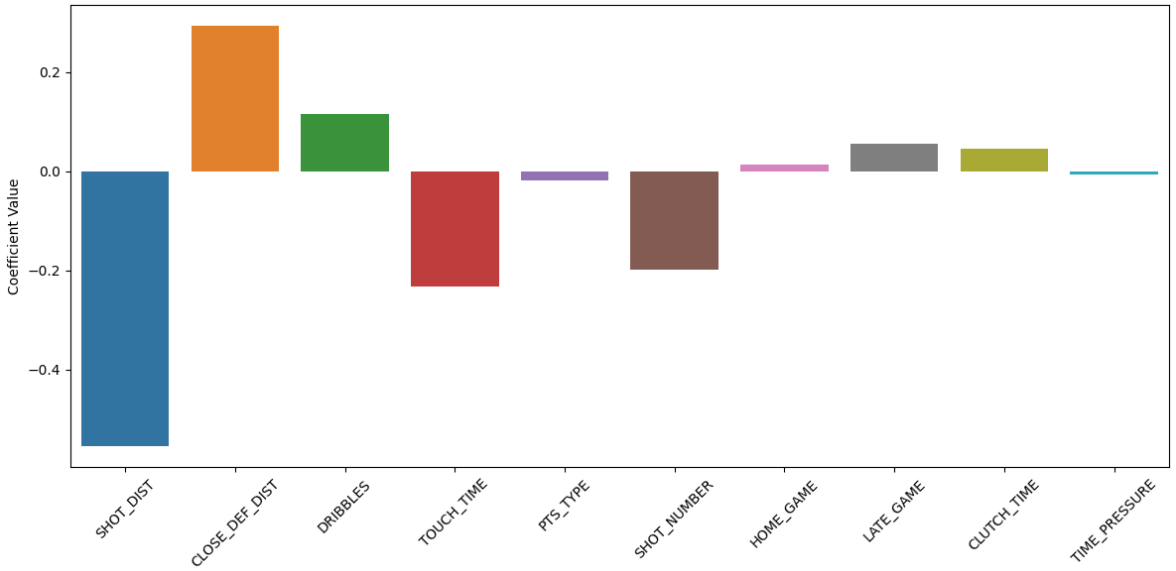

Figure 4: Global Feature Importance (Average Coefficient Values), Notebook 6.3

Lillard exhibits a more positive response to DRIBBLES, consistent with his role as a ball-dominant shot creator, while Gasol shows a stronger positive effect for HOME_GAME and LATE_GAME, possibly reflecting performance boosts in familiar environments or under certain game contexts. For TOUCH_TIME, both players display negative effects, but the magnitude is slightly larger for Lillard, implying potential efficiency loss when holding the ball longer.
Overall, these differences align with basketball-specific expectations: guards like Lillard maintain higher long-range efficiency and thrive with more dribble creation, while big men like Gasol rely less on perimeter play and may benefit more from situational or positional advantages. This case study highlights the model's capacity to capture player-specific sensitivities that would be masked in non-hierarchical approaches.

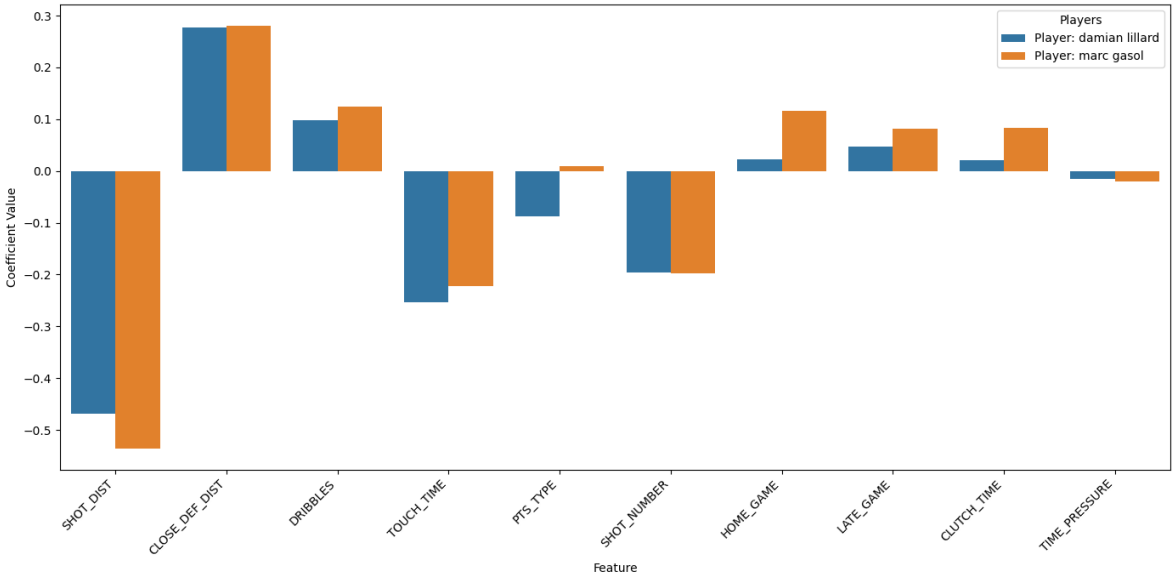

Figure 5: Player-Specific Feature Coefficients Comparison: Damian Lillard vs. Marc Gasol (see Notebook section 6.4)

5.3 Limitations of Player-Specific Coefficients: The Case of Stephen Curry

While player-specific coefficients can reveal meaningful differences in shot-making tendencies, they are not always aligned with on-court realities. A striking example is Stephen Curry—widely regarded as the greatest shooter in NBA history—whose coefficient for SHOT_DIST (-0.6393) is substantially more negative than that of Damian Lillard (-0.4690). From a basketball perspective, this is counterintuitive: Curry's three-point shooting percentage during the 2014–15 season far exceeded that of most players, including Lillard and especially Marc Gasol, yet the model suggests that increasing shot distance is more detrimental to his shot success probability than for these other players.

One plausible explanation lies in the interaction between Curry's shot selection and the regression framework. Curry attempts an exceptionally high volume of long-distance shots, including many from well beyond the three-point line. Even if he converts these attempts at a historically elite rate, their efficiency remains lower than that of shots taken near the rim. Since the model estimates coefficients relative to all distances, and rim attempts generally yield much higher field goal percentages, frequent deep shooting can inflate the apparent penalty of SHOT_DIST for Curry. In essence, the model captures the overall statistical trade-off between distance and shot success, but not the relative superiority of Curry's long-range shooting compared to other players.

Table 1: 3-Point Shooting Statistics for Selected Players

Player	Attempts per Game	FG%	Avg. Distance (ft)
Stephen Curry	7.9	41.7%	26.0
Marc Gasol	0.2	14.3%	24.4
League Avg.	3.3	36.8%	24.8

This example illustrates a broader limitation: the hierarchical logistic regression framework does not fully disentangle a player's exceptional skill in a specific shot type from the global efficiency patterns associated with shot location. Consequently, coefficients may reflect both player tendencies and the underlying baseline probabilities in a way that obscures true performance advantages, especially for outlier players with highly unorthodox but effective shooting profiles.

6 Conculsion

This study applied a Bayesian hierarchical logistic regression model to NBA shot data, capturing both global effects and player-specific deviations. Across all approaches—including LDA, Random Forests, and Neural Networks—accuracy plateaued around 60%, indicating a performance ceiling imposed by the available features and the inherent randomness of shot outcomes.

Globally, shot distance and defender proximity dominated as predictors, with distance strongly negative and spacing strongly positive. Other variables, such as dribbles, touch time, and shot sequence, showed smaller effects—particularly the latter, likely reflecting fatigue late in games. Factors like home court, clutch time, and time pressure had negligible average effects due to high player-to-player variability and rarity.

The hierarchical framework revealed some meaningful individual differences, as in the contrast between Damian Lillard and Marc Gasol. However, it also exposed strong limitations: Stephen Curry's unusually negative distance coefficient illustrates how the model can conflate exceptional skill with shot selection patterns, penalizing high-volume deep shooters despite superior efficiency.

Ultimately, while the model captures broad patterns and individual tendencies, predictive accuracy is bounded by the stochastic nature of basketball and the limits of the feature set. Richer spatial, temporal, and contextual data are needed to move beyond this ceiling.

A Appendix

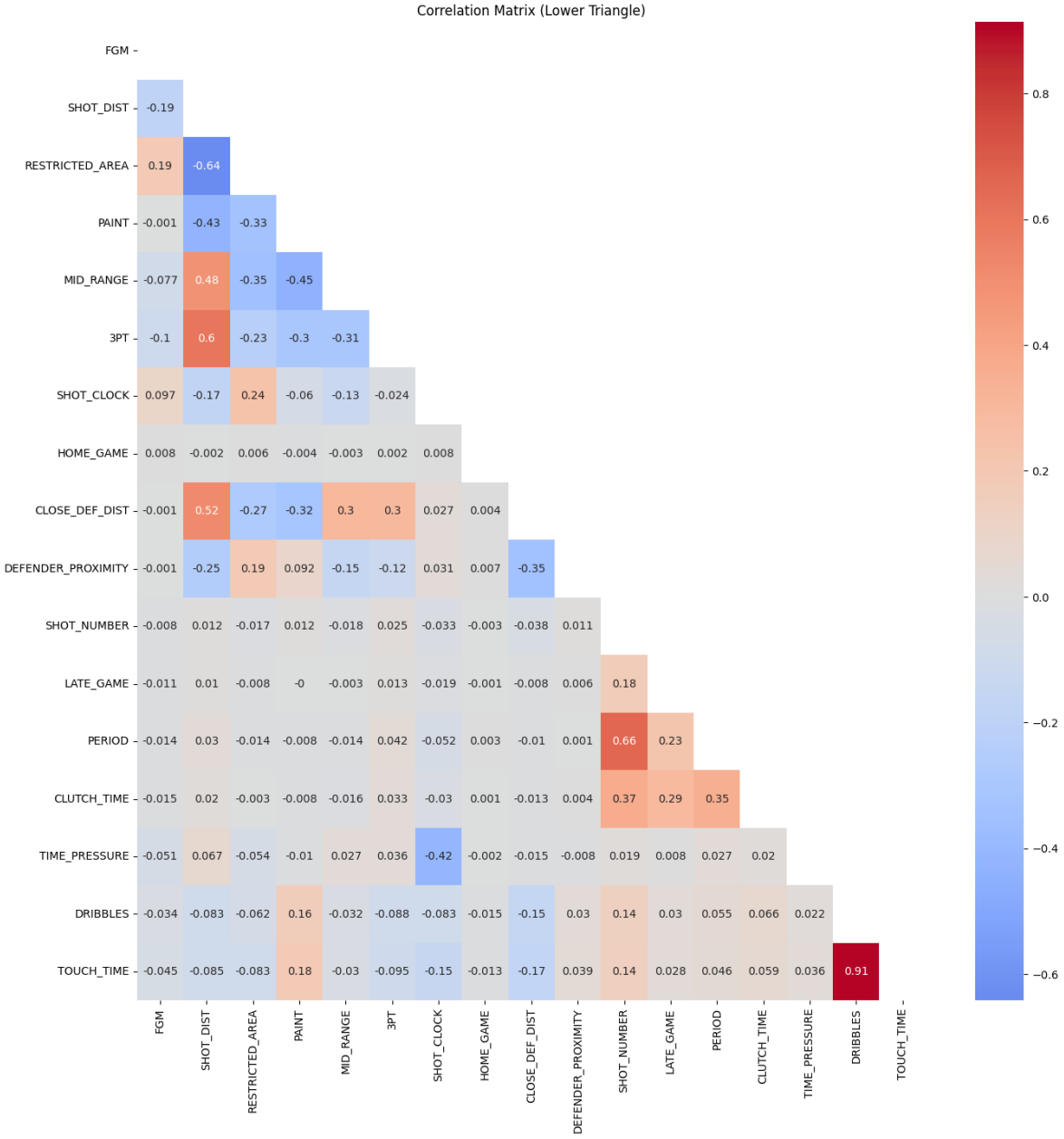

Figure 6: Correlation Matrix (only Lower Triangle), see Notebook section 3.5

Table 2: Variance Inflation Factors (VIF) for Selected Features

Feature	**VIF**
TOUCH_TIME	7.431
DRIBBLES	7.383
PTS_TYPE	2.131
HOME_GAME	1.963
CLUTCH_TIME	1.361
SHOT_DIST	1.319
CLOSE_DEF_DIST	1.310
SHOT_NUMBER	1.229
LATE_GAME	1.104
TIME_PRESSURE	1.063

Table 3: Mutual Information (MI) Scores for Selected Features

Feature	**MI_Score**
SHOT_DIST	0.015192
PTS_TYPE	0.008943
LATE_GAME	0.003582
SHOT_NUMBER	0.003449
DRIBBLES	0.002454
CLOSE_DEF_DIST	0.002342
TOUCH_TIME	0.002070
HOME_GAME	0.001803
CLUTCH_TIME	0.000411
TIME_PRESSURE	0.000000

Table 4: Statistical Significance Tests for Feature Differences Between Made and Missed Shots

Feature	**Mean_1**	**Mean_0**	**T-statistic**	**P-value**
SHOT_DIST	-0.1933	0.1643	-26.3693	9.72×10^{-151}
PTS_TYPE	2.1723	2.2623	-16.0429	1.40×10^{-57}
TOUCH_TIME	-0.0585	0.0497	-7.9130	2.64×10^{-15}
TIME_PRESSURE	0.0364	0.0557	-6.7657	1.36×10^{-11}
DRIBBLES	-0.0432	0.0367	-5.8325	5.54×10^{-9}
CLOSE_DEF_DIST	0.0194	-0.0165	2.5703	1.02×10^{-2}
HOME_GAME	0.5066	0.4920	2.1102	3.49×10^{-2}
CLUTCH_TIME	0.0702	0.0764	-1.7440	8.12×10^{-2}
SHOT_NUMBER	-0.0112	0.0096	-1.5129	1.30×10^{-1}
LATE_GAME	0.0096	0.0104	-0.5324	5.94×10^{-1}

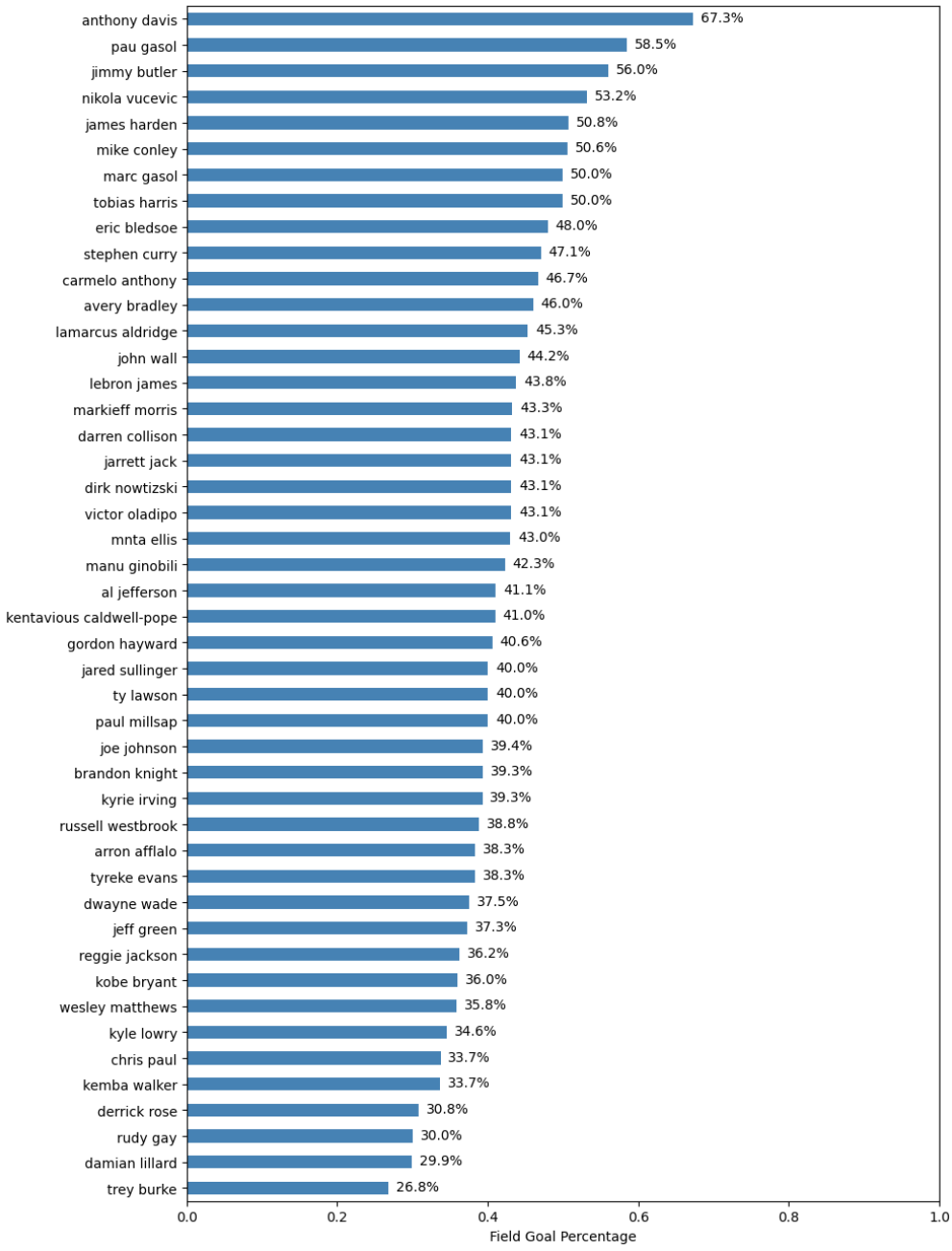

Figure 7: Sorted Clutch-Shooter (min. 50 attempts), see Notebook section 3.6

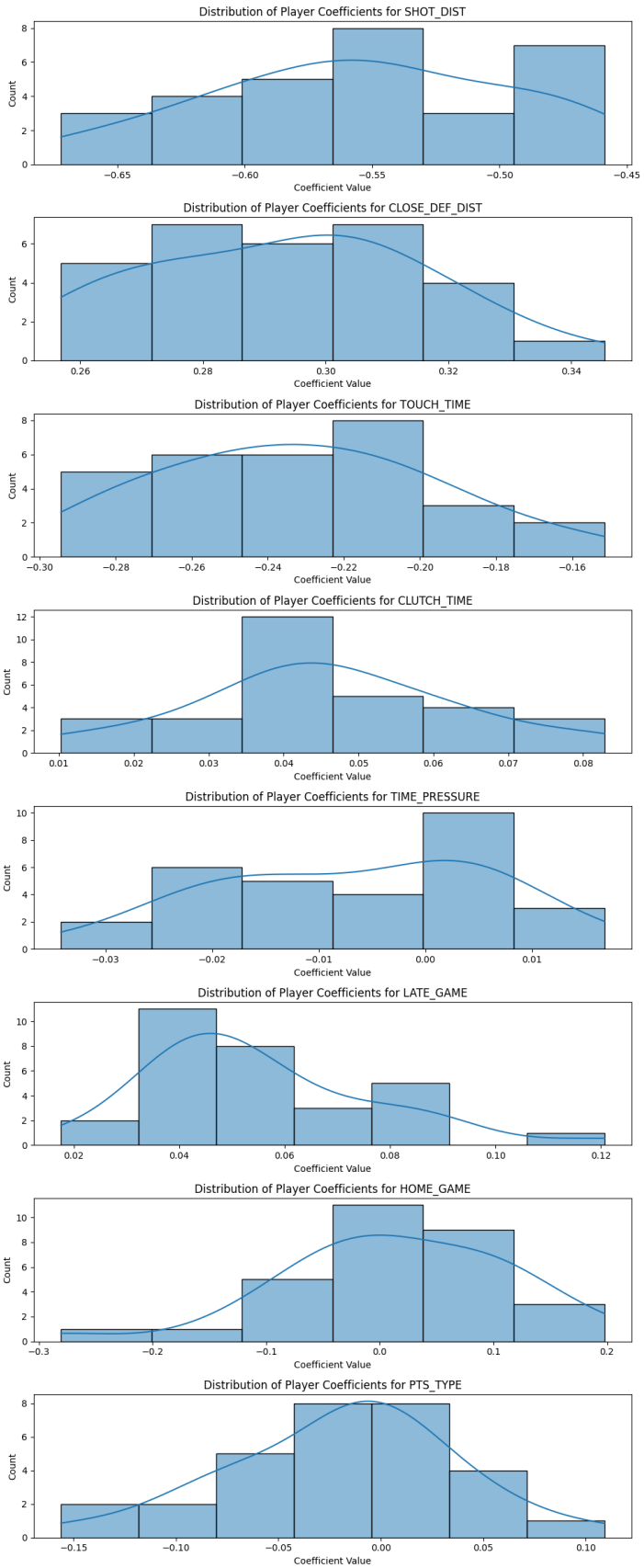

Figure 8: Plot distribution of all player-specific coefficients (features), see Notebook section 5.4.1

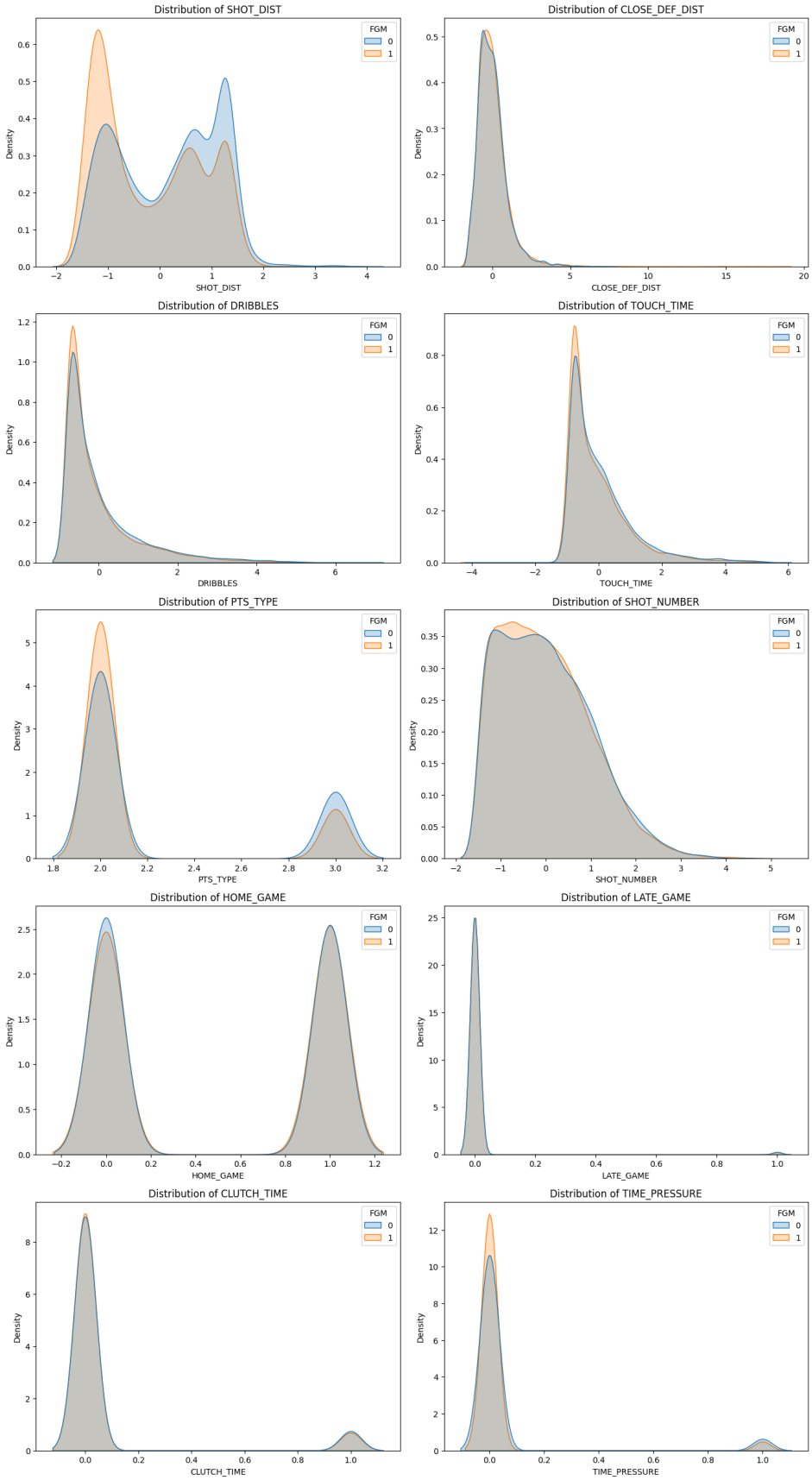

Figure 9: Feature Distribution Plots, see Notebook section 3.7