An Analysis of the Factors Influencing Stephen Curry's Shot Selection

CID: 06006350

May 13, 2025

Abstract

This study addresses a classification problem: identifying the factors that shape NBA legend Stephen Curry's shot selection, determining which variables govern his field-goal success, and, from a data-driven perspective, elucidating why he is regarded as one of the league's pre-eminent scorers. We collected Curry's complete regular-season shot-level data and analysed it with a suite of machine-learning and statistical techniques—including generalized linear models (GLM), decision trees and random forests, logistic regression, support vector machines (SVM), and neural networks (NN). The neural-network model emerged as the most accurate predictor of shot choice, revealing three primary determinants: distance to the basket, on-court location, and time remaining in the game. The findings culminate in a data-informed training framework aimed at basketball practitioners seeking to emulate or counter Curry's shot-selection tendencies.

Keywords: Binary Classification, Machine Learning, Data Science

1 Introduction

1.1 Formulation of the Research Question

Over the past decade of the NBA's "three-point revolution," Stephen Curry has redefined offensive spacing with his unprecedented shooting range. However, the decision-making logic underlying his shot selection remains largely unexplained systematically and quantitatively. This study investigates a central question: within a multidimensional framework encompassing spatial (e.g., distance, court zones), temporal (e.g., game quarter, time remaining), and contextual (e.g., shot type, home/away status) factors, which variables best explain and predict Curry's shot selection and success probability?

1.2 Data Acquisition and Data Preprocessing

This study is based on official NBA statistics, which are retrieved and processed programmatically using the `nba_api` interface. The raw dataset consists of $18,354$ rows and 25 columns, covering all of Stephen Curry's field goal attempts in regular season games from his rookie year in 2009 through the 2025 season. The data includes detailed information for each shot attempt in his career, such as timestamp, spatial coordinates, shot type, shot outcome, shooting team, opponent, and other multidimensional variables.

We proceed to clean the raw dataset. First, eight non-informative columns were removed from the raw dataset. The missing values were then checked in all fields and the data was found to be complete, which required no imputation.

To ensure logical consistency between `shot_type` and `shot_distance`, we verified three-point attempts based on the NBA three-point line (shots beyond 21.65-23.75 feet). An inconsistent three-point record was removed. Additionally, a new binary variable, `is_home`, was created by comparing team abbreviations to indicate whether each shot occurred at a home game.

Finally, several variables were renamed to improve clarity and interpretability. The cleaned dataset contains 18,353 observations and 18 variables.

2 Exploratory Data Analysis

2.1 Univariate Analysis

From the univariate analysis presented in Figure 1, several key insights can be drawn regarding Stephen Curry's shooting behavior.

- **Shot Outcome**

  The proportion of made shots (`made = 1`) and missed shots (`made = 0`) is nearly equal, indicating an overall shooting accuracy close to 50%. This balanced outcome reflects Curry's consistent shot performance across his career. This makes a well-balanced binary response variable.

- **Action Type**

  Most shots are classified as `Jump Shot`, overwhelmingly surpassing other action types, while less frequent actions such as `Tip Layup Shot`, `Layup Shot`, and `Alley Oop` appear only occasionally. This strong dominance by a single category indicates high cardinality but low diversity, which may reduce the predictive power of the original variable in modeling tasks.

Table 1 Descriptions of outcome and feature variables in Curry's shot dataset.

Variable Name	**Type**	**Description**
quarter	integer	Quarter of the game (1-4)
minutes_left	integer	Minutes remaining in the quarter
seconds_left	integer	Seconds remaining in the quarter
action_type	factor	Specific type of shot (e.g., Jump Shot, Layup)
shot_type	factor	General type of shot (2PT or 3PT Field Goal)
zone_basic	factor	Basic shot zone (e.g., Mid-Range, Restricted Area)
zone_area	factor	Area of the court (e.g., Left Side, Right Side)
zone_range	factor	Range of the shot (e.g., 8–16 ft., 24+ ft.)
distance_ft	integer	Distance from the basket in feet
x	integer	x-coordinate of the shot location
y	integer	y-coordinate of the shot location
made	binary	1 = shot made, 0 = missed
date	integer	Game date in YYYYMMDD format
season	factor	NBA season (e.g., 2009–10)
is_home	binary	1 = game played at home, 0 = away

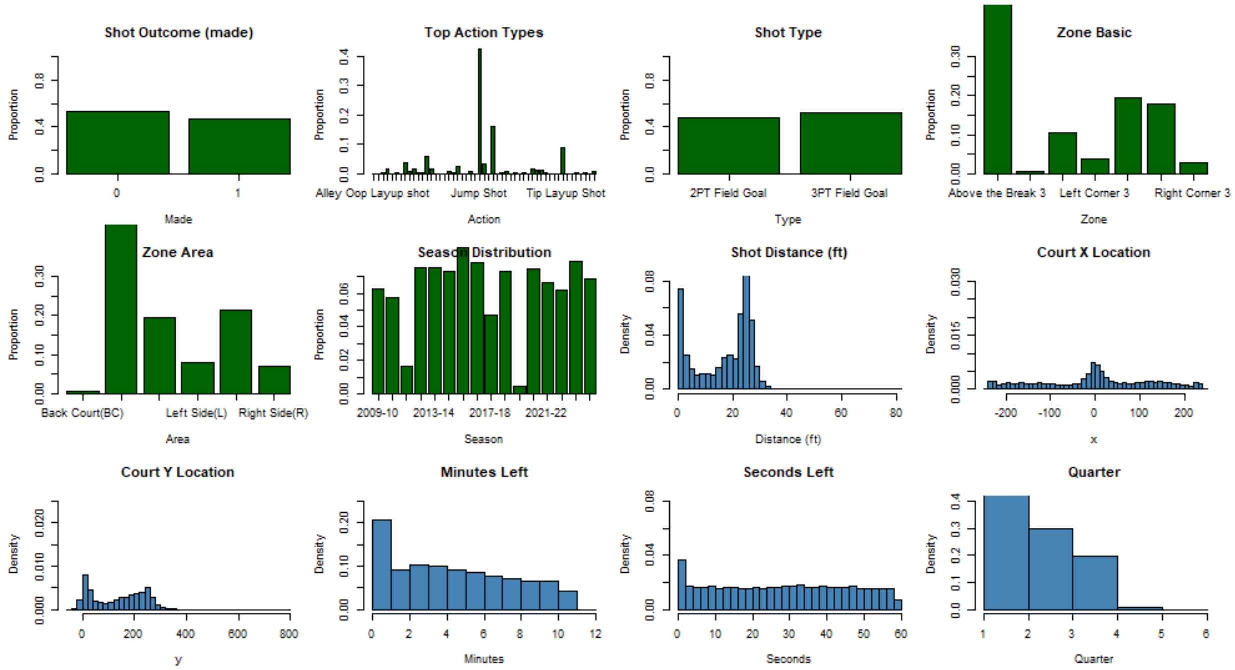

Figure 1 Univariate distributions of key variables.

To address this, action types will be consolidated into five commonly observed shooting styles—*Jump Shot, Hook Shot, Dunk, Driving Layup*, and *Bank Shot*—in order to simplify the structure, reduce noise from rare categories, and improve model interpretability and performance.

- **Shot Type, Shot Distance (ft) and Zone Area**

  Both `shot_type` and `zone_area` exhibit relatively balanced distributions, making them suitable for direct inclusion as categorical predictors in modeling. In contrast, `zone_basic` is highly imbalanced, and its treatment may require adjustment depending on model performance—such as category merging or re-encoding. The `distance_ft` variable is notably right-skewed, suggesting the potential benefit of applying a log transformation or introducing nonlinear terms (e.g., polynomial features) to better capture its relationship with shot success.

To examine Stephen Curry's shot selection across different court areas and shooting styles, I visualized his shot attempts using court-positioned scatter plots categorized by zone range, zone area, zone basic, and grouped action types in Figure 2. The primary shot types include jump shots, hook shots, dunks, driving layups, and bank shots. I manually grouped Curry's 52 original action types into these five representative categories to simplify the analysis and enhance interpretability.

Stephen Curry's shot location chart reveals a highly specialized and consistent shooting pattern centered on perimeter play. Across all zone-based groupings—Zone Range, Zone Area, and Zone Basic—his shot attempts are heavily concentrated in the "Above the Break 3" and "24+ ft." regions, confirming his identity as a deep-range shooter. The lateral distribution is balanced between the left and right sides of the court, with slightly more volume from the right-center area, indicating adaptability and spatial coverage. Corner threes and mid-range shots appear far less frequently, reflecting a modern, analytics-driven approach that favors high-efficiency zones. In terms of action type, jump shots dominate overwhelmingly, spanning nearly the entire half court, while actions like driving layups, dunks, bank shots, and hook shots are sparse and confined to the paint. This pattern reinforces Curry's role as a volume three-point scorer whose offensive threat is primarily generated through jump shooting from deep positions, with limited reliance on interior scoring.

- **Season Distribution**

  Stephen Curry's shot attempts are evenly distributed across seasons from 2009–10 to 2022–23, ensuring a comprehensive and temporally balanced dataset that captures his entire NBA career without seasonal bias.

- **Game Clock Context (Quarter, Minutes Left, Seconds Left)**

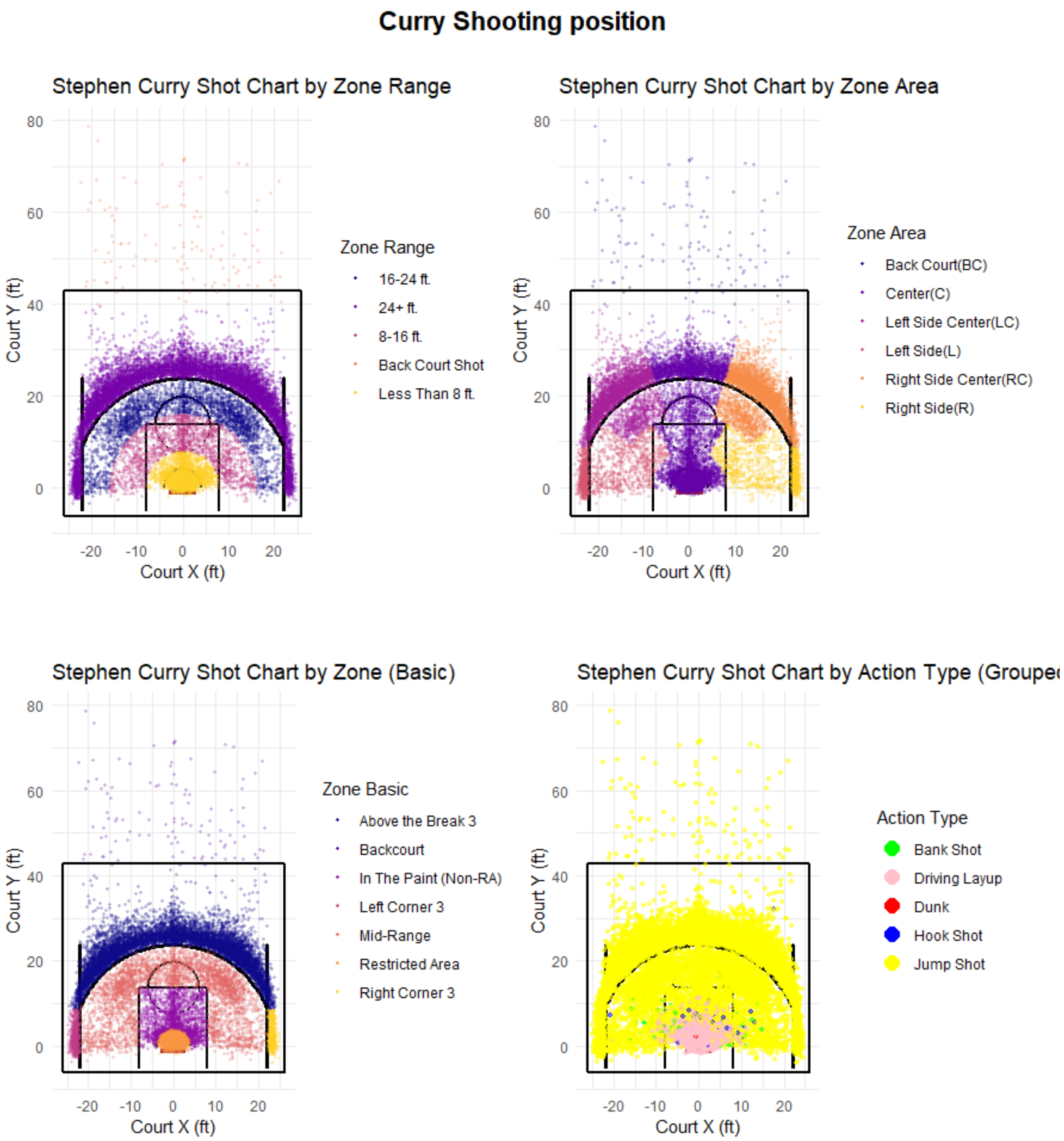

Figure 2 Curry Shooting position by zone range, zone area, zone basic, and grouped action types.

While shot frequency is highest in the first quarter and decreases over the course of the game, `minutes_left` and `seconds_left` exhibit relatively uniform distributions. This contrast highlights that variations in quarterly shot volume may be more reflective of rotation patterns or tactical decisions. In contrast, the consistent distribution of shots across the game clock suggests that Curry maintains an active offensive presence throughout the game. Notably, there is a slight increase in shot frequency near the end of each game, indicating that Curry is often entrusted with crucial, late-game shot opportunities. This pattern reinforces his role as the primary offensive leader and a trusted clutch performer for his team.

2.2 Bivariate Analysis

Figure 3 illustrates the univariate distributions of key variables in Curry's shot dataset. First, there is a very strong positive correlation between `distance_ft` and `x` ($r = 0.84$), reflecting the geometric relationship between horizontal court position and shot distance. This high multicollinearity suggests that one of these variables may be excluded in predictive modeling to avoid redundancy. In contrast, the outcome variable `made` shows minimal linear correlation with any individual numeric predictor—including shot distance, court location (`x`, `y`), game time (`minutes_left`, `seconds_left`), and `quarter`—with all coefficients close to zero. This indicates that no single variable strongly determines whether a shot is made. Nevertheless, small differences in marginal distributions between made and missed shots suggest potential nonlinear effects or interaction patterns.

We further employed visual analysis in Figure 4 to illustrate Stephen Curry's shooting accuracy across multiple variable dimensions, including seasonal trends, shot zones, distance ranges, and specific shot types.

Stephen Curry's shooting accuracy demonstrates remarkable consistency across his NBA career, with seasonal field goal percentages generally ranging between 45% and 50%. Despite a noticeable dip during the 2019–2021 period—due to injuries—his performance quickly rebounded, continuing his long-term shooting excellence.

Zone-wise, Curry achieves the highest accuracy in the restricted area, while his signature *Above the Break 3* zone maintains respectable efficiency despite the increased distance.

When categorized by distance, his accuracy declines as the distance increases, yet he remains effective even beyond 24 feet, underscoring his unparalleled long-range capability.

Analyzing shot types, close-range actions like dunks and hooks exhibit field goal percentages above 90%, while more complex jump shots—particularly turnaround and step-back variants—yield lower efficiencies around $30 - 40\%$. Notably, low-frequency shots such as push shots and pullup bank shots have the poorest success rates.

Overall, Curry's shot selection reflects a strategic balance: while his most frequent attempts (jump shots) are not the most efficient individually, their volume, spatial value, and his elite shot-making skill collectively sustain one of the most effective offensive profiles in modern basketball.

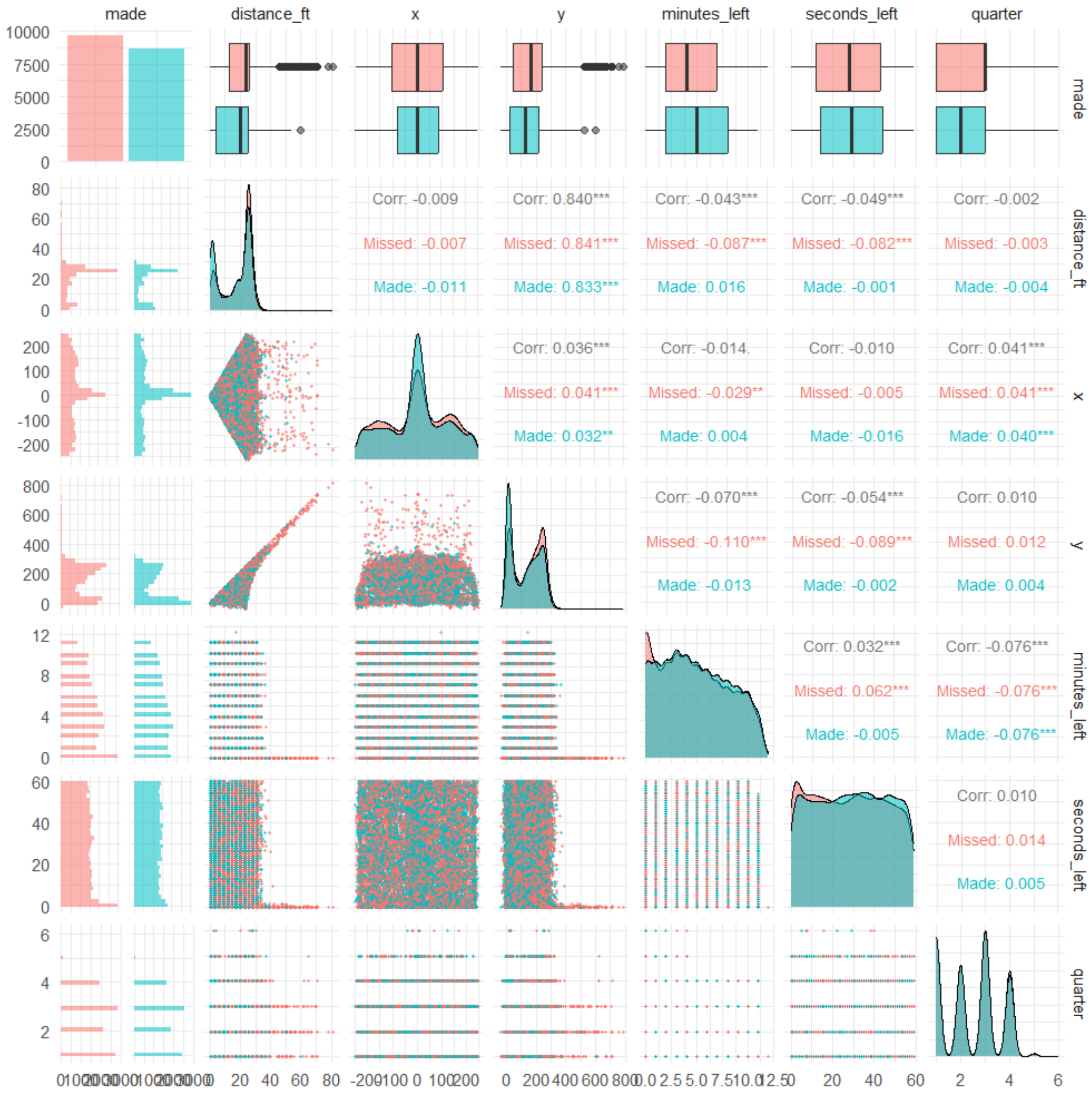

Figure 3 Bivariate distributions and conditional patterns.

2.3 Improving Linearity in GLMs Through Predictor Transformations

Based on the Figure 5, `distance_ft`, `x`, and `seconds_left` show clear nonlinear relationships with the estimated logit, suggesting violations of the linearity assumption in the logistic regression model. In contrast, `y` appears approximately linear, while `minutes_left` shows little variation and can be considered nearly flat.

To evaluate the linearity assumption in logistic regression, we created empirical logit plots for three continuous variables: `distance_ft`, `x`, and `seconds_left`, under four transformation types—original, square root, logarithmic, and inverse.

The results show that for `distance_ft`, both the logarithmic and square-root transformations

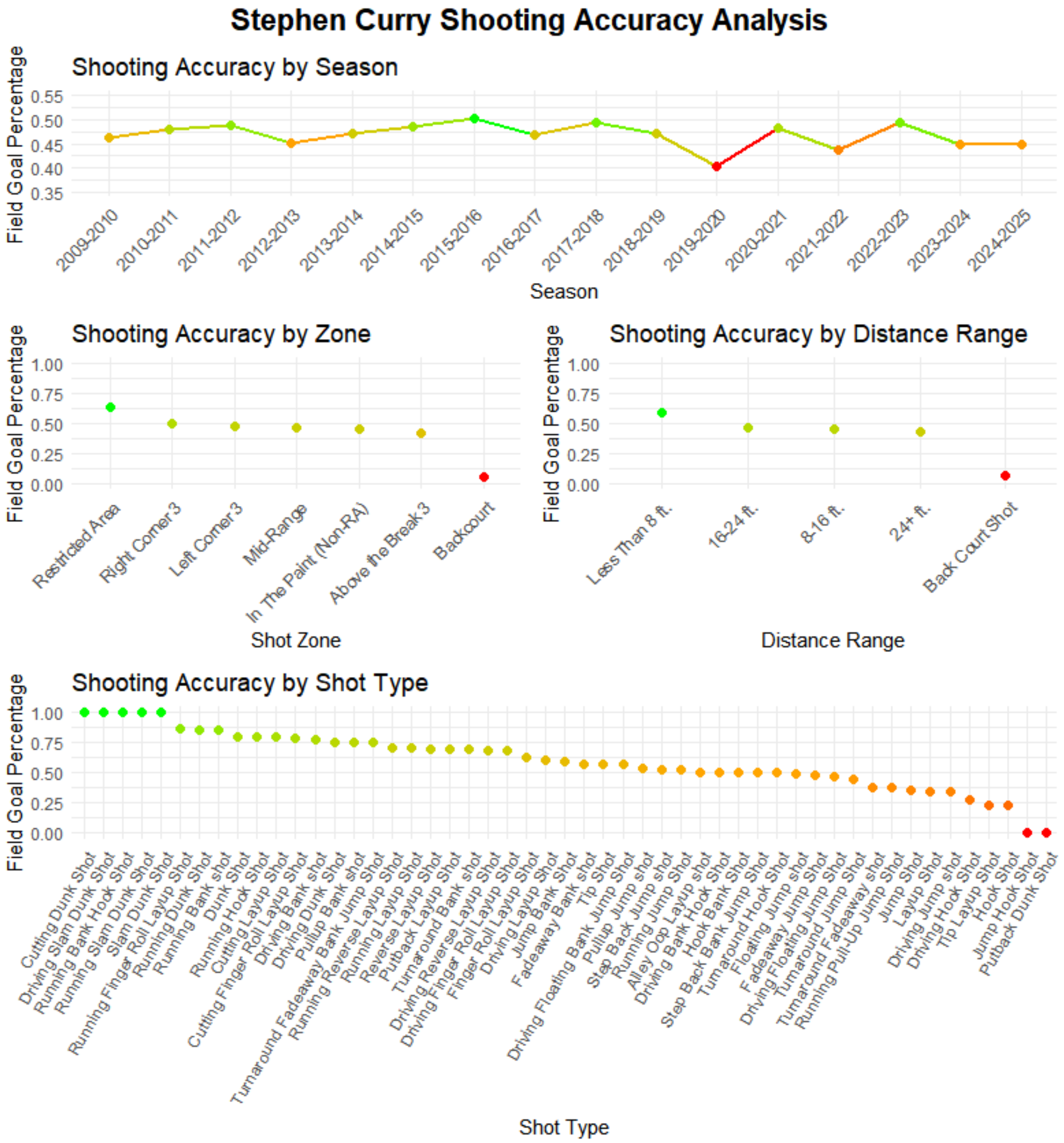

Figure 4 Shooting Accuracy across seasonal trends, shot zones, distance ranges, and specific shot types.

improve linearity, with the log transformation showing the most consistent trend. For `x`, the original and inverse forms show little structure, while the log transformation yields a clearer linear decrease. The results are illustrate in (Figure 6) For `seconds_left`, the square-root transformation slightly enhances the linear trend without distorting scale.

In conclusion, we selected the following transformations for modeling:

$$\log(\texttt{distance\_ft} + 1), \quad \sqrt{\texttt{seconds\_left}}$$

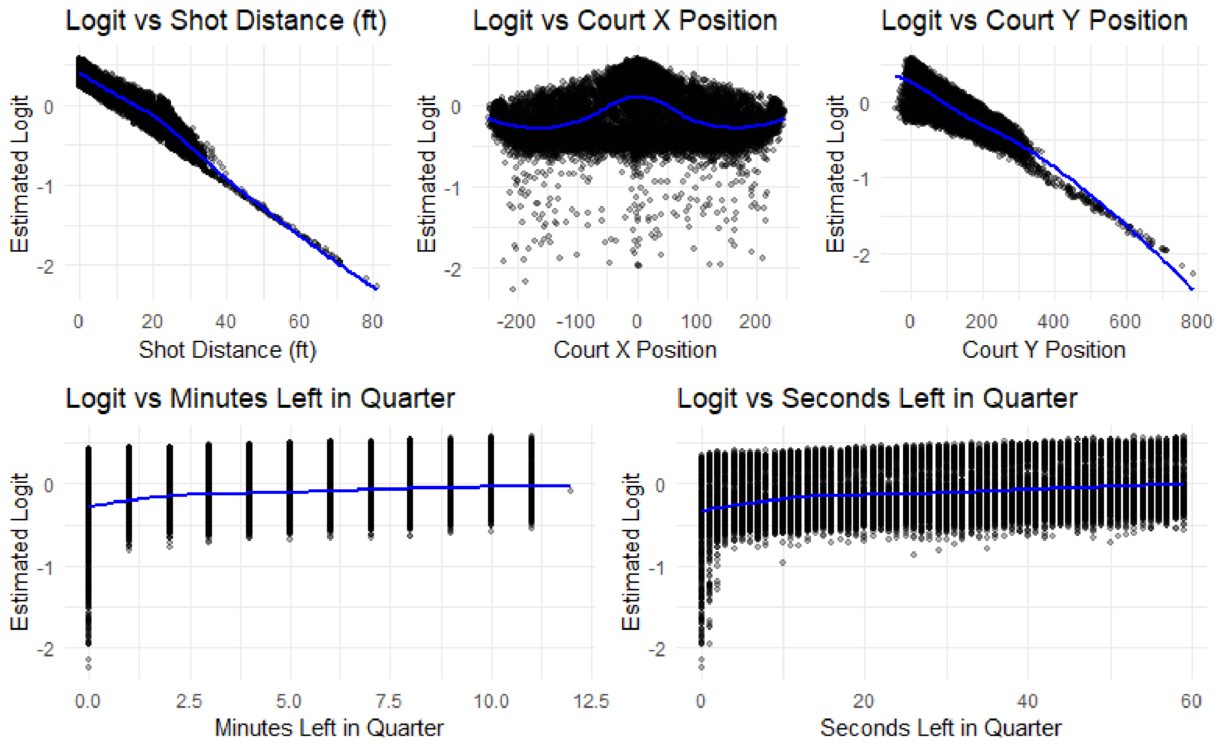

Figure 5 Empirical logit plots between log-odds of shot made and initial continuous game variables.

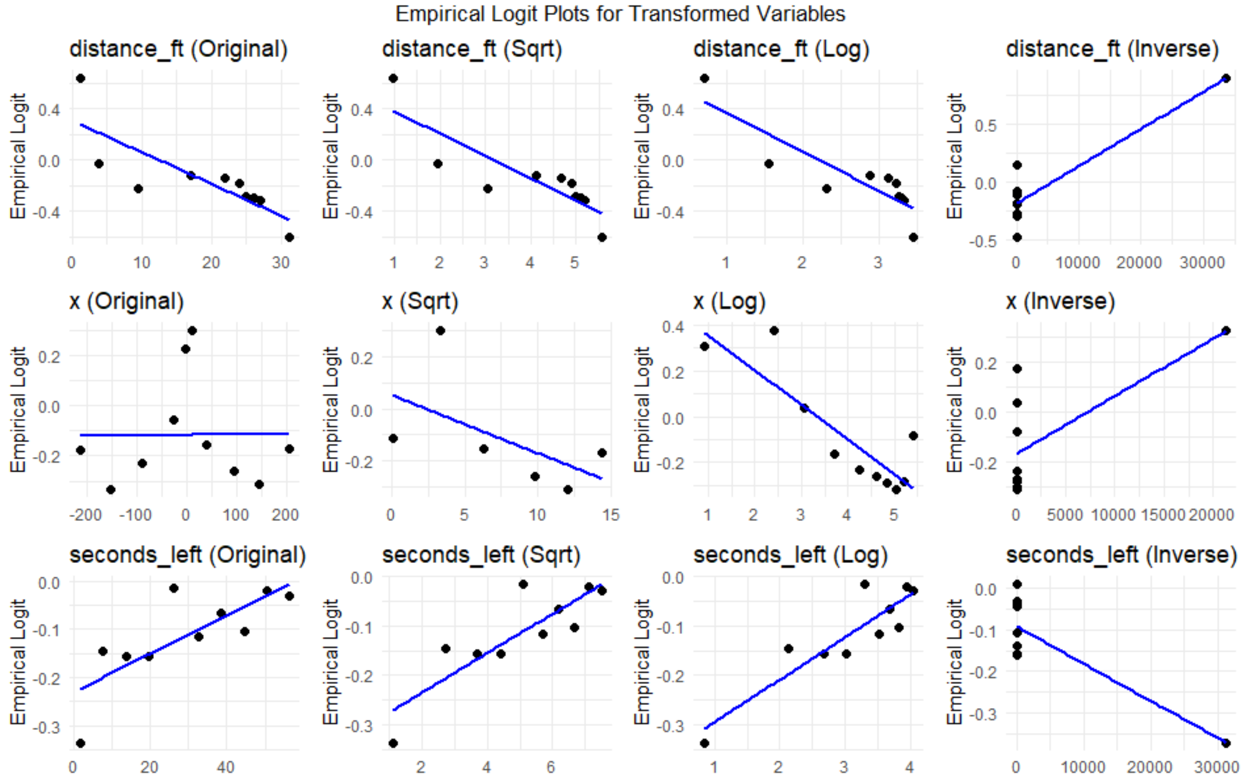

Figure 6 logit analysis of transformed continuous predictors

3 Model Fitting and Evaluation: Stephen Curry's Shot Accuracy Prediction Model

Prior to model fitting, we performed a series of preprocessing steps guided by univariate, bivariate, and GLM-based diagnostic analyses. First, the variable `shot_type` was consolidated into six types: Jump Shot, Hook Shot, Dunk, Driving Layup, Bank Shot, and Other, in order to reduce high-cardinality noise and improve interpretability. To eliminate unrepresentative and low-intentionality attempts, we filtered out all shots with a distance greater than 40 feet. Such attempts typically occur at the end of quarters or possessions and do not reflect genuine shot selection behavior.

All character-type variables were converted into factor variables to ensure proper handling within modeling functions. Additionally, empirical logit plots and Box-Tidwell transformations revealed several violations of the linearity assumption with respect to the logit link function. Therefore, we applied appropriate nonlinear transformations to selected predictors to improve linearity, including: $\log(\texttt{distance\_ft} + 1)$, $\sqrt{\texttt{seconds\_left}}$, $\log(\texttt{y} + 1)$.

The cleaned dataset was randomly partitioned into a training set (70%) and a test set (30%).

3.1 Fitting the Model with Alternative Methods

3.1.1 Stepwise Selection for Logistic Regression

In this stepwise logistic regression analysis, we compared model selection results based on Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The AIC-based model retained six main predictors—`distance_ft`, `seconds_left`, `minutes_left`, `action_group`, `zone_basic`, and `is_home`—with approximately 13 estimated coefficients. This model achieved a residual deviance of $17,227$ and an AIC value of $17,255$.

In contrast, the BIC-based model selected a more parsimonious specification, retaining only `minutes_left` and `action_group` (approximately 6 coefficients), with a slightly higher residual deviance of $17,281$ and a corresponding AIC of $17,338$.

3.1.2 Decision Tree

The resulting tree reveals interpretable hierarchical decision rules that influence the probability of a successful shot, as shown in Figure 7.

The most influential predictor was the categorical variable `action_group`, distinguishing between shot types such as *Jump Shot*, *Hook Shot*, and *Other*. This split forms the root of the tree, highlighting that certain shot types inherently differ in difficulty. Specifically, shots categorized as

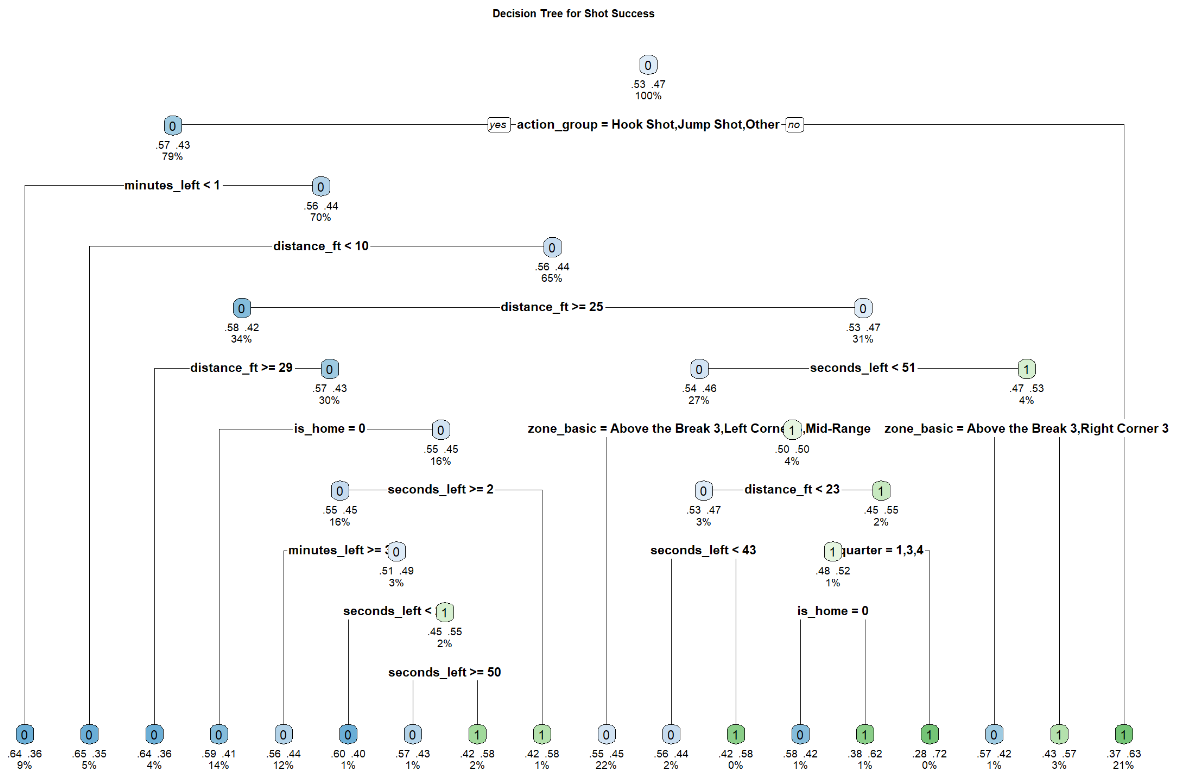

Figure 7 Classification decision tree modeling shot success

Jump Shot or *Hook Shot* generally exhibit lower success rates, prompting more refined splits based on game context.

Within the *Jump Shot* branch, the model identifies `minutes_left` and `distance_ft` as critical secondary predictors. For example, jump shots attempted with less than one minute remaining in the game had substantially lower success probabilities (approximately 21%), suggesting a strong negative impact of late-game pressure on shooting performance. Similarly, long-range jump shots (e.g., `distance_ft` $\geq$ 29) further decrease the likelihood of success, especially when taken in away games.

Conversely, non-jump shot actions, such as *Driving Layups* and *Dunks*, exhibit relatively higher baseline success rates. Within this group, spatial features like `zone_basic` proved to be highly informative. Shots taken from the *Right Corner 3* zone showed the highest estimated success rate (approximately 63%), confirming the known efficiency of corner threes. Moreover, shots from the *Mid-Range* zone with moderate distance (`distance_ft` $<$ 23) and sufficient time remaining (`seconds_left` $\geq$ 43) were also associated with increased success.

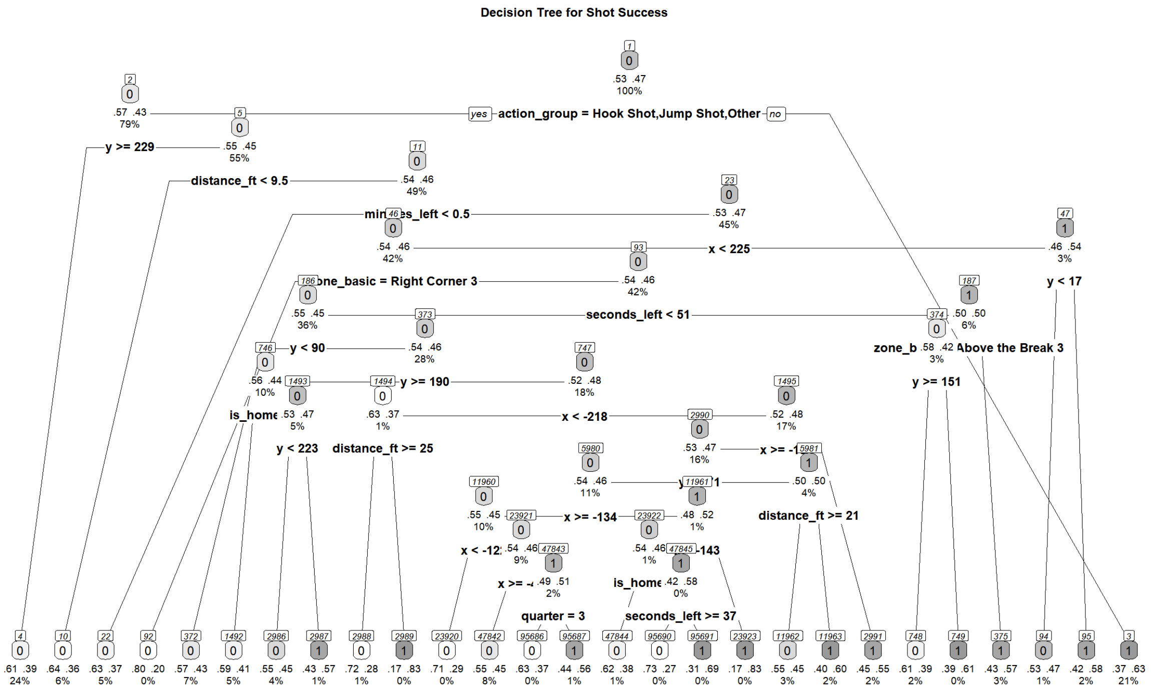

Figure 8 Classification random forest modeling shot success

3.1.3 Random Forest

The random-forest model ensembles 500 bootstrap decision trees to capture the complex, non-linear drivers of Curry's shot success in Figure 8

Partial dependence diagnostics reveal intuitive patterns: the probability of a make declines sharply beyond ∼25 ft, rises for attempts taken with more than ∼5 seconds remaining, and peaks for close-range lay-ups and dunks.

Interaction effects uncovered by the forest show that long jumpers taken in late-clock situations or from the deep right wing suffer the lowest success rates, whereas short-range attempts at home in early-clock scenarios are classified as high-probability makes (> 70%).

3.1.4 Logistic regression

We fitted a logistic regression model to analyze the factors affecting Stephen Curry's shot success. The model reveals that shot distance has a significant negative effect on the probability of success ($\hat{\beta} = -0.2715$, $p < 0.001$), indicating that longer shots are less likely to be made. Time-related features such as $\sqrt{\texttt{seconds\_left}}$ and `minutes_left` are also statistically significant and positively associated with shot success. Regarding shot technique, both "Jump Shot" and "Hook Shot" are significantly less effective compared to the baseline action group. Although being a "3PT Field

Goal" is not statistically significant after accounting for other variables, certain court zones like "Left Corner 3" and "Right Corner 3" show a significantly higher likelihood of success. Finally, playing at home is positively associated with shot success ($p = 0.0032$). Overall, the model provides interpretable insights into how spatial, temporal, and contextual features influence Curry's scoring probability.

3.1.5 Support Vector Machine(SVM)

Figure 9 illustrates the classification result of a SVM model trained to predict shot outcomes based on the basketball court's spatial coordinates (x and y). Each point represents a shot attempt, where `x` denotes a made shot and `o` denotes a missed shot. The background colors indicate the model's predicted classification regions: red areas correspond to zones where the model predicts a higher probability of a made shot, while yellow areas indicate predicted misses. The SVM model effectively captures the spatial structure of shot success, identifying regions near the basket (with higher y values and x values close to zero) as high-probability zones, consistent with empirical basketball knowledge. In contrast, areas farther from the basket or toward the sidelines are classified as lower-probability regions.

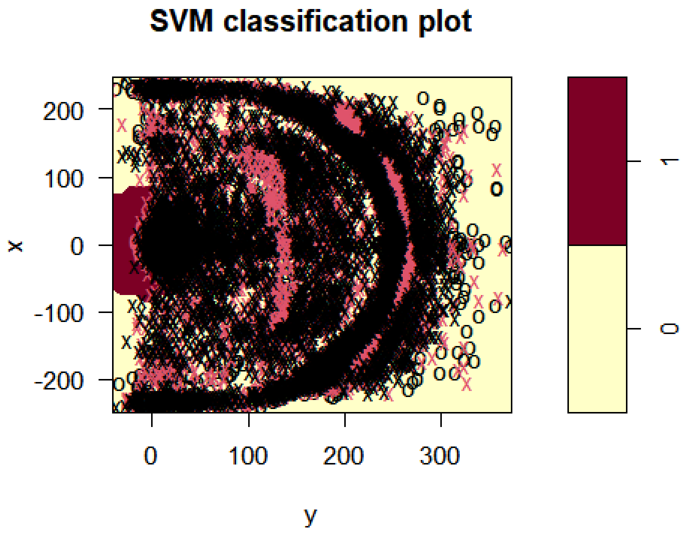

Figure 9 SVM classification regions for shot success.

3.1.6 Neural Network

A single–hidden–layer neural network with five hidden units was trained on the cleaned training set, which comprises the continuous predictors `log_distance_ft` and `minutes_left` together with four categorical variables. The network was optimised for 200 iterations with a weight–decay penalty of 10^{-4}. Adopting the formula `made` $\sim$ `.`, the model is designed to capture nonlinear interactions among shot distance, remaining time, and action/zone factors.

3.2 Model Evaluation

Based on the results in Table 2, the following key conclusions can be drawn:

- **Neural network** delivers the strongest overall discrimination: highest accuracy (0.643) and recall (0.760) with the top AUC (0.660). Its F_1 score (0.494) also exceeds all baselines, confirming that the single hidden layer captures non-linear interactions absent in simpler models. The cost is a lower precision (0.365), indicating a tendency to label more shots as "make".
- **BIC stepwise logistic regression** is the most conservative predictor. Precision peaks at 0.633, yet recall collapses to 0.276, showing that the model sacrifices many true makes in exchange for cleaner positive predictions. This behaviour may be preferable when false positives are costly.
- **AIC stepwise logistic regression** and the **decision tree** provide balanced recall (about 0.58–0.61) but modest precision (about 0.34–0.38). Their interpretable structure makes them valuable for tactical insight even if raw accuracy (≈ 0.57) lags behind the neural net.
- The **random forest** improves on the single tree in recall (0.597) and AUC (0.579) but still suffers from the lowest precision (0.321), suggesting that the ensemble favours sensitivity over specificity.
- The **SVM** attains the second-best accuracy (0.600) by maintaining the highest true-negative count, yet its precision (0.299) remains weak; its AUC (0.600) shows no clear edge over tree-based methods.
- **Best-threshold logistic regression** (threshold optimised for Youden's J) marginally lifts accuracy (0.577) and precision (0.617) relative to the default-threshold AIC model, but recall falls to 0.327, underscoring the inevitable precision–recall trade-off.

4 Conclusion and Further Discussions

This study builds a systematic model on 18353 field-goal attempts by Stephen Curry recorded in NBA regular seasons from 2009–10 through 2024–25. Benchmarking seven classification algorithms

Table 2 Comparison of model performance on the test set

Model	Confusion-matrix counts				Performance metrics				
	TN	FP	FN	TP	Accuracy	Precision	Recall	F1	AUC
AIC stepwise logit	2238	1733	590	906	0.575	0.343	0.606	0.438	0.582
BIC stepwise logit	2407	1912	421	727	0.573	0.633	0.276	0.384	0.576
Decision tree	2084	1628	744	1011	0.566	0.383	0.576	0.460	0.579
Random forest	2256	1791	572	848	0.568	0.321	0.597	0.418	0.579
Best-threshold logit	2292	1776	536	863	0.577	0.617	0.327	0.427	0.581
SVM	2532	1752	435	748	0.600	0.299	0.633	0.406	0.600
Neural network	2565	1652	300	950	0.643	0.365	0.760	0.494	0.660

demonstrates that a single–hidden-layer neural network achieves the best predictive performance (accuracy = 0.643, recall = 0.760, AUC = 0.660), showing that deep-learning architectures can capture the non-linear interplay among shot distance, temporal pressure, and shot type. By contrast, stepwise logistic regression selected via the Bayesian Information Criterion remains attractive in scenarios that privilege parsimony and interpretability while still delivering high precision. Among continuous covariates, distance is the decisive driver: once the attempt exceeds 25 ft, the make probability declines sharply. Corner three-pointers retain comparatively high efficiency owing to their favourable geometry, whereas jump shots attempted in the final minute of the fourth quarter suffer a pronounced decrease in success rate.

Several limitations merit attention. First, the data set covers only regular-season games and omits the heightened competitive context of the play-offs. Second, micro-level tracking features such as defender proximity, contact intensity, and screen quality are absent, potentially understating contextual complexity. Third, evaluation metrics are reported as overall means, leaving season-level or injury-phase heterogeneity unexplored.

Future work may evolve along two directions. **(i)** Integrate SportVU or Second Spectrum tracking feeds to embed defender pressure, screen efficacy, and other spatio-temporal micro-features. **(ii)** Adopt a Bayesian hierarchical framework to estimate season- and opponent-specific effects dynamically.

References

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1): 5–32.

Cervone, D., D'Amour, A., Bornn, L., and Goldsberry, K. (2016). A multiresolution stochastic process model for predicting basketball possession outcomes. *Journal of the American Statistical Association*, 111(514): 585–599.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3): 273–297.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer.

Goldsberry, K. (2019). *SprawlBall: A Visual Tour of the New Era of the NBA*. Houghton Mifflin Harcourt.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer, 2 edition.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

Miller, A. and Bornn, L. (2017). Possession sketches: Mapping nba strategies. In *MIT Sloan Sports Analytics Conference*.