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

06006350

Compiled: May 13, 2025

Github Repo: REPLACE-WITH-LINK-TO-YOUR-TAGGED-RELEASE

1 Project Description

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. Using a bespoke web-scraping pipeline, we collected Curry's complete regular-season shot-level data from the 2009–10 to 2024–25 seasons, totaling 18,353 observations. This dataset includes detailed spatial, temporal, and contextual features such as shot distance, game quarter, clock time, action type, and home-court status.

Prior to model fitting, we conducted an extensive exploratory data analysis (EDA), revealing intuitive and domain-consistent patterns: shots taken closer to the basket and during early quarters had higher success rates, while corner threes and breakaway layups exhibited distinctive spatial efficiencies. Empirical logit plots and partial dependence diagnostics further illustrated key nonlinearities in shooting behaviour.

We then applied a suite of machine learning and statistical models—including generalized linear models (GLM), decision trees, random forests, stepwise logistic regression (AIC and BIC), support vector machines (SVM), and neural networks (NN)—to predict both shot selection and outcome. Among them, the neural network emerged as the most accurate model, identifying distance to the basket, court location, and temporal pressure as the dominant predictive factors. The findings culminate in a data-driven framework that offers actionable insights for players and coaches seeking to emulate or strategically defend against Curry's historically unparalleled shot profile.

2 Assessment Criteria

Technical Competence:

- Constructed a custom web-scraping pipeline using the `nba_api` package to retrieve and consolidate all $18,353$ of Curry's regular-season field goal attempts, including detailed spatial, temporal, and contextual features.
- Employed R for full-cycle data analysis: preprocessing, transformation, empirical visualization, and predictive modeling using a diverse set of algorithms (GLM, decision trees, random forests, SVM, neural networks).

User Interface:

- Developed intuitive visualizations such as heatmaps and decision boundaries to display model predictions and spatial shooting tendencies.
- Report is clearly structured.

Analysis and Interpretation:

- Conducted extensive exploratory data analysis to uncover empirical shot patterns; applied transformations based on model diagnostics.
- Compared seven models using metrics including accuracy, recall, precision, AUC, and F1-score; highlighted the neural network's superiority in capturing non-linear decision factors.

Presentation and Communication:

- The report is well-structured with clear sections on background, data, methodology, model evaluation, and conclusion.
- Graphs and tables have descriptive captions, fonts and explanations.

Reproducibility and Documentation:

- All analysis was conducted in Latex.
- The project includes detailed comments in code chunks and appropriate justification for methodological choices.
- Wrote separate R scripts for each part of the workflow.
- Saved all plots in outputs.

Project Management:

- Scripts are modular and organized by task.
- Final results and report are published on GitHub with a tagged release.

3 Project Reflection

Learnings:

- Gained a foundational understanding of the full data science workflow, including how to organize project files systematically and apply consistent, meaningful file naming conventions.
- Learned how to publish projects on GitHub, including uploading files, writing clear README documentation, and managing version control effectively.
- Became familiar with writing a comprehensive data science report—from data exploration to model evaluation—and reinforced theoretical knowledge of data cleaning, regression, and machine learning techniques through practical application.

Challenges:

- Most predictive models performed suboptimally, indicating potential missing variables in the data collection process. This highlighted the importance of capturing key contextual features to improve model accuracy.
- Still encountered difficulties in the data preprocessing stage, especially with complex cleaning tasks such as handling inconsistent factor levels and engineering meaningful features from raw variables.

Further Development:

- Integrate SportVU or Second Spectrum tracking feeds to embed defender pressure, screen efficacy, and other spatio-temporal micro-features. Adopt a Bayesian hierarchical framework to estimate season- and opponent-specific effects dynamically.