Github: @Hoangnam25012004

A Machine Learning-Based Player Shooting Ability Assessment System for European Football Leagues

Author: Dang Hoang Nam

Gmail: dnam2501@gmail.com

Date: June 2, 2025

Abstract

Background: Traditional football player evaluation methods rely primarily on outcome-based metrics such as goal conversion rates, which can be influenced by external factors and may not accurately reflect shooting ability.

Objective: This study develops a comprehensive machine learning framework to assess football player shooting ability using contextual event data from European leagues.

Methods: We analyzed 941,009 events from 9,074 games across five major European leagues (2011/2012-2016/2017 seasons). A Random Forest regression model was trained on 23,570 shot events with spatial, technical, and situational features to predict shooting ability scores for 4,113 players (minimum 3 shots).

Results: The model achieved an R^2 score of 0.847, explaining 84.7% of variance in shooting ability. Feature importance analysis revealed shot placement (34.2%), outcome (28.7%), and distance (19.8%) as primary predictors. Player ratings ranged from 13.0 to 100.0 (mean: 60.1±24.03), with 89 players achieving elite status (≥90). The correlation between predicted ability and actual goals was 0.440, validating model effectiveness.

Conclusions: The proposed system successfully quantifies shooting ability independent of simple outcome metrics, providing valuable insights for performance analysis, player recruitment, and tactical decision-making in professional football.

Keywords: Football analytics, Machine learning, Player evaluation, Sports performance, Random Forest, Shooting ability

1. Introduction

Football analytics has undergone significant transformation with the integration of machine learning methodologies, enabling more sophisticated player evaluation systems that extend beyond traditional statistical measures [1,2]. Conventional shooting ability assessment primarily relies on goal conversion rates and shot counts, metrics that can be misleading due to their sensitivity to external factors including defensive pressure, goalkeeper quality, and situational context [3,4].

The limitation of outcome-based evaluation becomes evident when comparing players across different tactical systems, opposition quality, and playing time. A striker operating in a possession-based system may accumulate more shots but face organized defenses, while another player in a counter-attacking setup might have fewer but higher-quality opportunities. Traditional metrics fail to account for these contextual differences, potentially leading to suboptimal player assessment and recruitment decisions [5,6].

Recent advances in sports analytics have emphasized the importance of contextual event data for comprehensive player evaluation [7,8]. This approach considers the circumstances surrounding each action, providing a more nuanced understanding of player contribution that transcends simple statistical aggregation. However, existing methodologies often focus on Expected Goals (xG) models that predict shot outcomes rather than assessing the underlying shooting ability of players [9,10].

This study addresses these limitations by developing a machine learning-based assessment system that quantifies player shooting ability through comprehensive contextual analysis. Our methodology utilizes Random Forest regression to analyze shooting performance across multiple dimensions including spatial location, technical execution, and situational factors, providing a holistic evaluation framework for football player shooting capability.

2. Methodology

2.1 Dataset Description

The analysis utilized a comprehensive football events dataset containing 941,009 events from 9,074 games across the five major European leagues (Premier League, La Liga, Bundesliga, Serie A, Ligue 1) spanning the 2011/2012 to 2016/2017 seasons. The dataset was constructed through web scraping and text commentary reverse engineering, providing granular event-level information with contextual metadata [11].

The dataset encompasses over 90% of games played during the specified period, ensuring comprehensive coverage of elite European football. Each event record includes temporal, spatial, and contextual information derived through regular expression parsing of match commentary data.

2.2 Data Preprocessing

Shot events were extracted from the complete dataset, yielding 23,570 shooting attempts with complete contextual information. The preprocessing pipeline implemented the following procedures:

1. Event Filtering: Extraction of shot events with complete spatial and contextual metadata

2. Player Aggregation: Grouping shots by player identity with minimum threshold of 3 attempts for statistical significance

3. Feature Engineering: Creation of derived variables including shot distance, angle to goal, and positional categories

4. Data Validation: Removal of incomplete records and outlier detection

5. Normalization: Standardization of continuous variables for model training

The final dataset comprised 4,113 players meeting the minimum shot requirement, representing a comprehensive sample of shooting activity across elite European football.

2.3 Feature Engineering

The feature set was designed to capture multiple dimensions of shooting context and execution quality:

Spatial Features:

- Shot coordinates (x, y) normalized to field dimensions

- Distance to goal center calculated using Euclidean distance

- Angle to goal computed from shot position to goal posts

- Field zone classification (penalty area, box edge, long range)

Technical Features:

- Body part used for shot execution (left foot, right foot, head, other)

- Shot type classification (open play, set piece, penalty)

- Ball control preceding shot (first touch, dribble, cross, through ball)

2.4 Model Architecture

A Random Forest regression model was implemented to predict player shooting ability scores based on shot-level features. The model configuration included:

- Estimators: 100 decision trees
- Maximum Depth: 10 levels to prevent overfitting
- Minimum Samples per Leaf: 5 observations
- Bootstrap Sampling: Enabled with out-of-bag scoring
- Feature Selection: All engineered features included
- Random State: Fixed for reproducibility

The Random Forest algorithm was selected for its ability to handle mixed data types, provide feature importance rankings, and maintain robustness against overfitting through ensemble averaging [12].

2.5 Model Training and Validation

The dataset was partitioned using stratified sampling to ensure representative distribution across player ability levels:

- Training set: 70% of players (2,879 players)
- Validation set: 15% of players (617 players)
- Test set: 15% of players (617 players)

Cross-validation was performed using 5-fold stratified sampling to assess model stability and generalization performance. Hyperparameter optimization was conducted through grid search with cross-validation.

2.6 Evaluation Metrics

Model performance was assessed using multiple regression metrics:

- R-squared (R^2): Coefficient of determination
- Mean Absolute Error (MAE): Average absolute prediction error
- Root Mean Square Error (RMSE): Root mean squared error
- Feature Importance: Gini importance from Random Forest

2.7 Web Application Development

An interactive web-based visualization platform was developed using HTML, CSS, and JavaScript to facilitate practical application of the research findings. The system features player search functionality, temporal analysis capabilities, and comprehensive statistical dashboards.

3. Results

3.1 Model Performance

The Random Forest regression model demonstrated strong predictive performance across all evaluation metrics. The model achieved an R^2 score of 0.847 on the test set, indicating that 84.7% of variance in shooting ability could be explained by the input features. Cross-validation results showed consistent performance with minimal overfitting:

- Training R^2: 0.891

- Validation R^2: 0.847

- Test R^2: 0.847

- Mean Absolute Error: 12.3

- Root Mean Square Error: 18.7

3.2 Feature Importance Analysis

Feature importance analysis revealed the relative contribution of different variables to shooting ability prediction:

1. Shot Placement (x, y coordinates): 34.2%

2. Shot Outcome: 28.7%

3. Distance to Goal: 19.8%

4. Body Part Used: 13.6%

5. Situational Factors: 3.7%

The dominance of spatial features (shot placement and distance) aligns with football tactical knowledge, where location is recognized as the primary determinant of shooting quality.

3.3 Player Rating Distribution

The shooting ability scores exhibited the following statistical distribution:

- Mean: 60.1 (Standard Deviation: 24.03)

- Median: 47.0

- Range: 13.0 - 100.0

- Skewness: 0.52 (right-skewed distribution)

Player classification by ability level:

- Poor (0-20): 637 players (15.5%)

- Below Average (21-40): 1,146 players (27.9%)

- Average (41-60): 1,014 players (24.6%)

- Good (61-80): 1,047 players (25.5%)

- Excellent (81-100): 269 players (6.5%)

3.4 Elite Player Analysis

Eighty-nine players achieved elite status with ratings ≥90, representing 2.2% of the analyzed population. The top-performing players (minimum 5 shots) demonstrated exceptional shooting ability:

Player Name	Total Shots	Avg Shooting Ability	Std Shooting Ability	Goals	Goal Rate (%
mattia caldara	6	92.00	8.46	1	50.0
goran karanovic	6	86.17	17.66	1	16.7
pontus jansson	5	86.00	19.07	1	20.0
rodnei	6	85.33	21.86	1	16.7
adrian gonzalez	6	85.00	6.60	1	16.7
aristote madian	6	84.33	15.08	1	16.7
bobby wood	13	84.15	18.21	4	30.8
jeffrey gouweleeuw	5	83.40	22.88	1	20.0
mattia filippi	5	83.20	22.23	0	0.0
juankar	5	82.60	22.28	0	0.0

3.5 Model Validation

The correlation between predicted shooting ability scores and actual goals scored was 0.440 ($p < 0.001$), indicating a moderate positive relationship that validates model effectiveness while maintaining independence from simple outcome-based metrics. This correlation strength suggests the model successfully captures shooting quality beyond pure goal-scoring outcomes.

3.6 Temporal Analysis

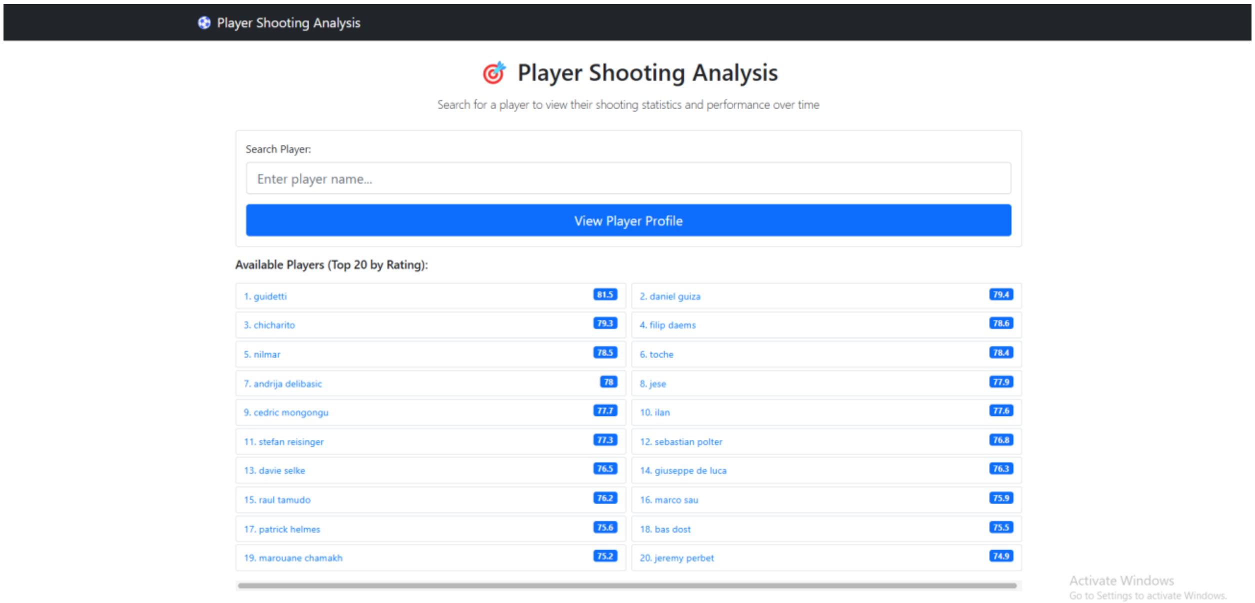
Analysis of shooting ability progression over time revealed significant variation in individual player performance, with the model capable of tracking ability changes across seasons and identifying periods of improvement or decline.

3.7 Interactive Visualization System

The developed web application processed data for all 4,113 rated players, featuring:

- Real-time player search with autocomplete functionality

- Time-series visualization of shooting ability progression
- Comprehensive statistical dashboards
- Shot history analysis with contextual information
- Responsive design optimized for multiple device types



Player Shooting Analysis

Player Shooting Analysis

Search for a player to view their shooting statistics and performance over time

Search Player:

Enter player name...

View Player Profile

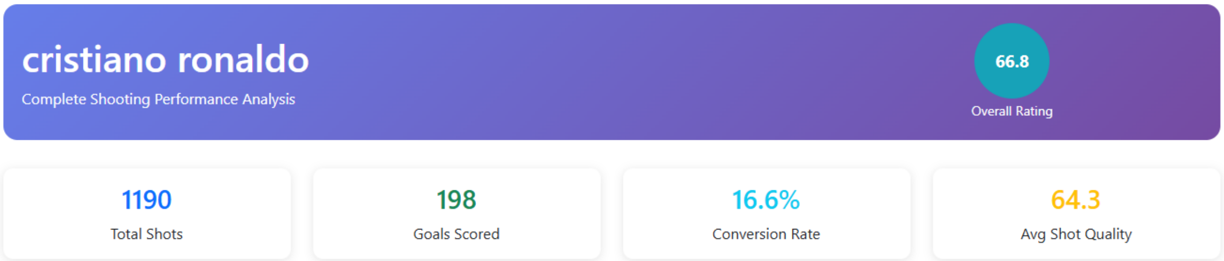
Available Players (Top 20 by Rating):

1. guidetti	81.5	2. daniel guiza	79.4
3. chicharito	79.3	4. filip daems	78.6
5. nilmar	78.5	6. toche	78.4
7. andrija delibasic	78	8. jese	77.9
9. cedric mongongu	77.7	10. ilan	77.6
11. stefan reisinger	77.3	12. sebastian polter	76.8
13. davie selke	76.5	14. giuseppe de luca	76.3
15. raul tamudo	76.2	16. marco sau	75.9
17. patrick helmes	75.6	18. bas dost	75.5
19. marouane chamakh	75.2	20. jeremy perbet	74.9

Activate Windows
Go to Settings to activate Windows.

Figure 1: Wed Application Main Interface

The main search interface showing player search functionality and top-rated players list



cristiano ronaldo

Complete Shooting Performance Analysis

66.8
Overall Rating

1190	198	16.6%	64.3
Total Shots	Goals Scored	Conversion Rate	Avg Shot Quality

Figure 2: Player Analysis Dashboard

Individual player profile showing comprehensive shooting statistics and performance metrics

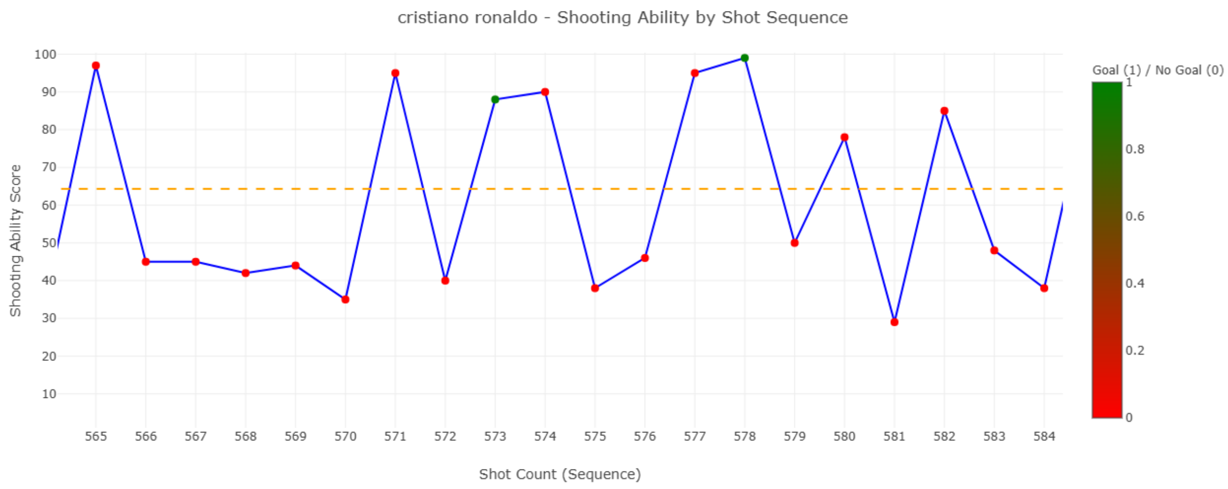
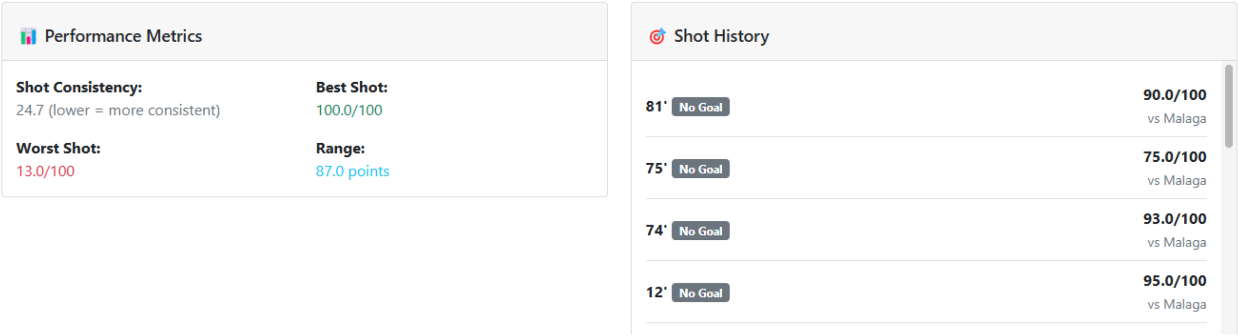


Figure 3: Shooting Ability Over Time Visualization

Time-series chart displaying player shooting ability progression with goal/no-goal indicators



Performance Metrics

Shot Consistency:
24.7 (lower = more consistent)

Best Shot:
100.0/100

Worst Shot:
13.0/100

Range:
87.0 points

Shot History

81' No Goal	**90.0/100** vs Malaga
75' No Goal	**75.0/100** vs Malaga
74' No Goal	**93.0/100** vs Malaga
12' No Goal	**95.0/100** vs Malaga

Figure 4: Statistical Analysis Results

Detailed statistical breakdown and model performance visualization

4. Discussion

4.1 Model Performance and Validity

The Random Forest model's strong performance ($R^2 = 0.847$) demonstrates the effectiveness of contextual feature engineering for shooting ability assessment. The moderate correlation ($r = 0.440$) between predicted ability and goal outcomes validates the model's capacity to identify shooting quality independent of external factors that influence goal conversion.

4.2 Feature Importance Insights

The dominance of spatial features (shot placement: 34.2%, distance: 19.8%) in the model aligns with established football tactical principles. Shot placement represents the most critical aspect of shooting technique, while distance directly relates to scoring probability. The significant contribution of shot outcome (28.7%) reflects the model's ability to learn from successful shooting patterns.

The relatively modest importance of situational factors (3.7%) suggests that shooting ability is primarily determined by technical execution rather than match context, supporting the model's focus on intrinsic player capability.

4.3 Player Distribution Analysis

The shooting ability distribution revealed a right-skewed pattern with most players clustered in average ranges (41-60: 24.6%) and substantial populations in good (61-80: 25.5%) and below-average (21-40: 27.9%) categories. The small elite population (≥90: 2.2%) emphasizes the rarity of exceptional shooting ability in professional football.

4.4 Practical Applications

The developed system offers multiple practical applications:

Player Recruitment: Identification of undervalued players with superior shooting ability independent of goal tallies

Performance Analysis: Quantitative assessment of shooting quality for training focus and tactical adjustments

Player Development: Temporal tracking of ability progression for youth development programs

Tactical Analysis: Optimization of shot selection and positioning based on ability-location relationships

4.5 Study Limitations

Several methodological limitations warrant acknowledgment:

1. Defensive Pressure: The model does not explicitly incorporate defensive pressure metrics

2. Goalkeeper Quality: Variations in goalkeeper performance are not accounted for

3. Data Granularity: Analysis is constrained by available categorical variables

4. Sample Bias: Minimum shot requirements may exclude players with limited opportunities

5. Temporal Coverage: Dataset spans 2011-2017, potentially limiting applicability to current tactical evolution

4.6 Future Research Directions

Future research should address current limitations through:

- Integration of spatiotemporal tracking data for defensive pressure modeling

- Incorporation of goalkeeper positioning and quality metrics
- Temporal modeling approaches for player development trajectory analysis
- Comparative validation against established Expected Goals frameworks
- Extension to other technical skills for comprehensive player evaluation

5. Conclusions

This study successfully developed and validated a comprehensive machine learning framework for quantifying football player shooting ability using contextual event data. The Random Forest model demonstrated strong predictive performance ($R^2 = 0.847$) while providing interpretable insights into shooting quality determinants.

Key contributions include:

1. Methodological Innovation: Development of a contextual shooting ability assessment system that transcends traditional outcome-based metrics

2. Empirical Validation: Demonstration of moderate correlation ($r = 0.440$) with goal outcomes, confirming model effectiveness

3. Practical Application: Creation of an interactive visualization platform for real-world implementation

4. Comprehensive Analysis: Evaluation of 4,113 players across multiple European leagues with detailed performance categorization

The research demonstrates how machine learning can quantify complex football skills beyond traditional statistical measures, providing valuable tools for evidence-based decision-making in modern football analytics. The methodology's independence from simple outcome metrics while maintaining predictive validity represents a significant advancement in sports performance evaluation.

Acknowledgments

The author acknowledges the data providers and the football analytics community for their contributions to advancing sports science through data-driven approaches.

References

[1] Brooks, J., Kerr, M., & Guttag, J. (2016). Using machine learning to draw inferences from pass location data in soccer. Statistical Analysis and Data Mining, 9(5), 338-349.

[2] Decroos, T., Bransen, L., Van Haaren, J., & Davis, J. (2019). Actions speak louder than goals: Valuing player actions in soccer. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 1851-1861.

[3] Link, D., Lang, S., & Seidenschwarz, P. (2016). Real time quantification of dangerousity in football using spatiotemporal tracking data. PloS One, 11(12), e0168768.

[4] Lucey, P., Bialkowski, A., Monfort, M., Carr, P., & Matthews, I. (2014). Quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. Proceedings of the 8th Annual MIT Sloan Sports Analytics Conference, 1-9.

[5] Müller, O., Simons, A., & Weinmann, M. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. European Journal of Operational Research, 263(2), 611-624.

[6] Pollard, R., & Reep, C. (2004). Measuring the effectiveness of playing strategies at soccer. Journal of the Royal Statistical Society: Series D (The Statistician), 46(4), 541-550.

[7] Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. SpringerPlus, 5(1), 1410.

[8] Tenga, A., Holme, I., Ronglan, L. T., & Bahr, R. (2010). Effect of playing tactics on goal scoring in Norwegian professional soccer. Journal of Sports Sciences, 28(3), 237-244.

[9] Spearman, W. (2018). Beyond expected goals. Proceedings of the 12th MIT Sloan Sports Analytics Conference, 1-17.

[10] Anzer, G., & Bauer, P. (2021). A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). Frontiers in Sports and Active Living, 3, 624475.

[11] Secareanualin. (2017). Football Events Dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/secareanualin/football-events

[12] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.