Predicting Soccer National Team Selection Using Club-Level Performance

by Md Nafisul Hasan Sami

Introduction

National team selection in professional soccer is often viewed as the pinnacle of achievement, influenced by a mix of performance, reputation, and team needs. While it is commonly assumed that players with strong club-level statistics are more likely to be selected, the relationship between these stats and actual selection remains unclear.

This project explores the question: Can club-level performance metrics reliably predict whether a player will be selected for the national team? Using data from top European leagues and World Cup squad rosters, I investigate this relationship through classification models.

We combined two datasets: one with individual performance statistics (e.g., goals, assists, minutes played) and another listing players selected for World Cup squads. After preprocessing and cleaning, we trained logistic regression and random forest models to classify players as selected or not.

Although both models achieved over 80% accuracy, they struggled to correctly identify selected players. The random forest model slightly outperformed logistic regression in identifying true positives, but the recall remained low.

This project demonstrates that club statistics alone offer limited predictive power for national team selection, emphasizing the complexity of selection decisions in professional soccer.

Methods

Datasets and Features

We used two main datasets:

1. Club Performance Data: 2,852 players from top 5 European leagues, with 37 numerical features including goals, assists, minutes played, expected goals (xG), and progressive passes.
2. National Team Selection: Historical FIFA World Cup squads from 1930 to recent years, listing players selected for each tournament.

Data Preparation

- Merged datasets by matching player names (normalized to lowercase).
- Created a binary target variable `Selected` (1 if player was in World Cup squad, 0 otherwise).
- Dropped irrelevant columns such as names, team, and competition.
- Removed rows with missing values.

Exploratory Analysis

- Computed correlation between features and `Selected`. Features such as age, xAG, and G+A showed weak positive correlation with selection.
- Detected class imbalance: only 17% of players were labeled as selected.

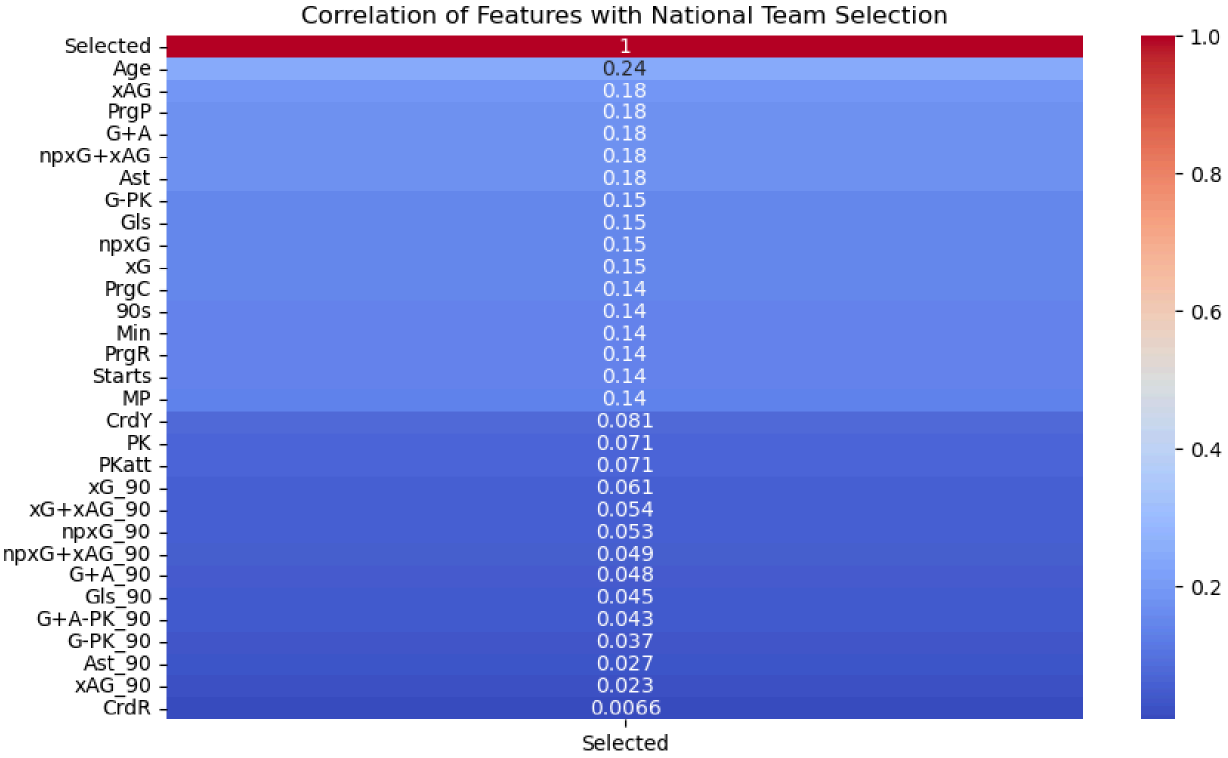

Figure 1: Correlation of Features

Correlation between numeric club performance features and national team selection status. Age has the highest positive correlation (0.24), followed by xAG, progressive passes, and goal contributions. All correlations are relatively weak.

Models Used

- **Logistic Regression**: Baseline model due to its interpretability.

- **Random Forest Classifier**: Ensemble model chosen for its ability to capture non-linear relationships and handle large feature sets.

Tools

Python (Pandas, Scikit-learn, Seaborn, Matplotlib) was used for data processing, modeling, and visualization. All models were evaluated using accuracy, precision, recall, and F1-score on a held-out test set.

We compared model performance and visualized results using feature importance plots and confusion matrices.

Results

The classification models were evaluated on a test set comprising 20% of the data.

Model	Accuracy	Precision (Selected)	Recall(Selected)	F1(Selected)
Logistic Regression	82%	0.32	0.06	0.10
Random Forest	82%	0.36	0.08	0.13

Figures

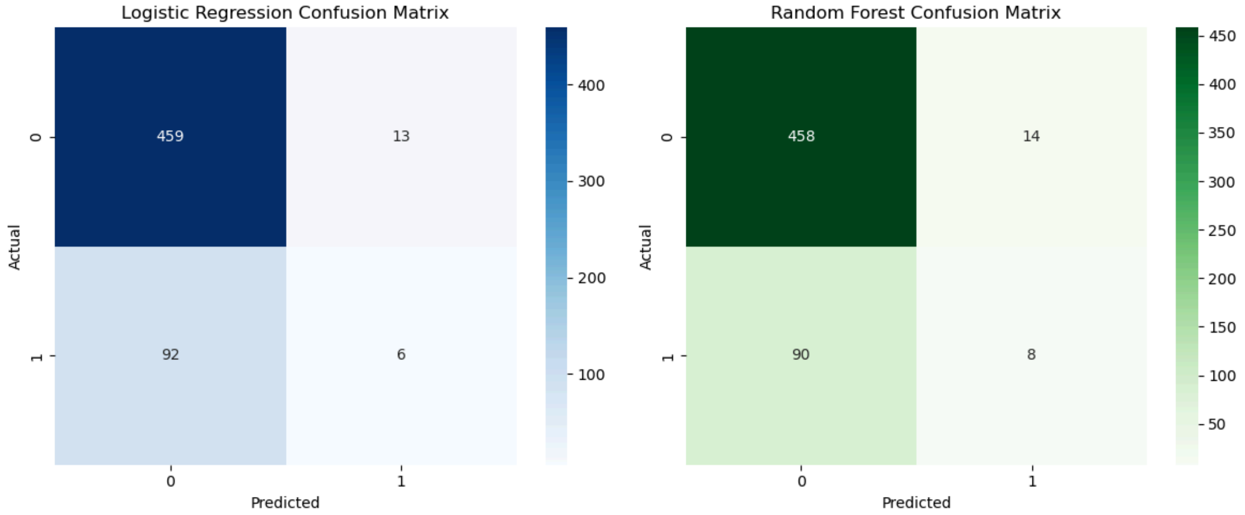

Figure 2: Confusion Matrix

Confusion matrices for Logistic Regression (left) and Random Forest (right). Both models accurately predicted the majority class (not selected), but struggled with recall for selected players. Random Forest correctly identified 8 selected players, while Logistic Regression identified 6.

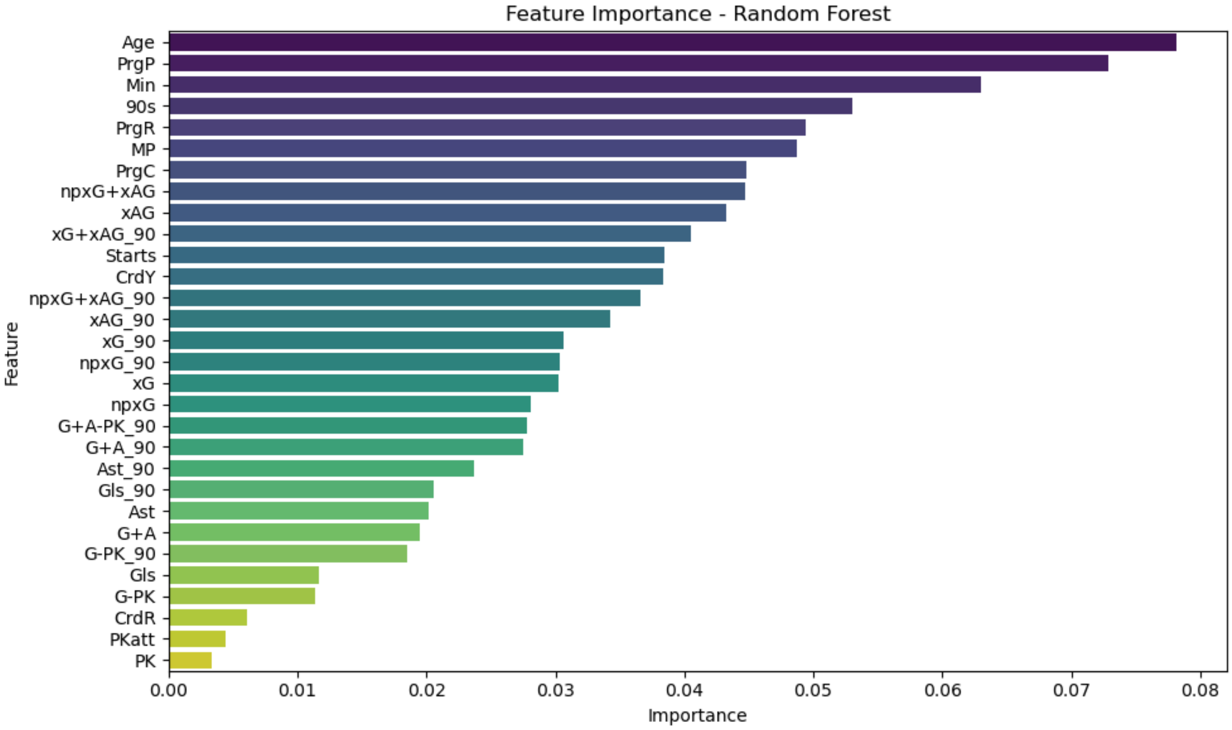

Figure 3: Feature Importance- Random Forest

Feature importance plots indicated that age, assists, xAG, and progressive passes were among the top features contributing to prediction. Despite good overall accuracy, the low recall for the selected class highlights the models' limitations in identifying national squad members based on club stats alone.

Discussion

The results suggest that while club-level performance metrics offer some insight, they are insufficient on their own to predict national team selection with high reliability. The models' poor recall for selected players indicates that many important aspects of selection—such as tactical fit, leadership, injury status, and reputation—are not present in the data.

One of the main hurdles was class imbalance: only 17% of players were selected, which led models to favor the majority class (not selected). As a result, even models with high accuracy did not perform meaningfully on the target class.

Random forest slightly outperformed logistic regression in terms of identifying selected players, suggesting some non-linear patterns exist. Feature importance visualization helped in interpreting which features the model relied on, though most had weak correlations individually.

If continued, the project could be improved by incorporating more contextual data: player injuries, position, team success, and perhaps even media presence. Additionally, training on more comprehensive national team rosters (not just World Cup squads) could provide a better-balanced dataset.

Overall, this project demonstrates the importance of using machine learning to test real-world assumptions and shows the value—and limits—of data-driven decision making in sports.

Sources and AI Usage:

- Aktas, O. (2024). *All Football Players Stats in Top 5 Leagues 23/24*. Kaggle. Retrieved from https://www.kaggle.com/datasets/orkunaktas/all-football-players-stats-in-top-5-leagues-2324
- World Cup squad data: https://github.com/jfjelstul/worldcup
- ChatGPT by OpenAI was used to assist in editing, formatting, structuring, and proofreading the written content.