VISHAL CHAUDHARY · DATA ANALYST · DUBLIN, IRELAND

Sports Analytics & Machine Learning · Academic Project

Football Match Outcome Prediction

XGBoost with SHAP/LIME interpretability predicting outcomes across 18,894 European football matches.

18,894	64%	150+	10
Matches Analysed	Prediction Accuracy	Engineered Features	Leagues Covered

TOOLS & TECHNOLOGIES

Python	XGBoost	SHAP	LIME	Pandas	API-Football
Scikit-learn	Matplotlib				

Email vishal.ch1401@gmail.com	LinkedIn linkedin.com/in/vishal111	GitHub github.com/chaudhary521	Location Dublin, Ireland

PROBLEM STATEMENT

Predicting football match outcomes is one of the hardest problems in sports analytics because of the sport's high inherent variance. Most commercial prediction models use simplistic features without accounting for team form trajectories, head-to-head history, or momentum shifts within a season. The goal was to build an interpretable model that not only predicts outcomes but explains individual predictions in human-readable terms.

DATASET

18,894 football matches across 10 European leagues including the Premier League, La Liga, Bundesliga, Serie A, and Ligue 1. Raw results data was sourced via the API-Football service covering five-plus seasons. From these raw results, 150+ engineered features were constructed including rolling win rates, goal differential momentum, home and away form trajectories, head-to-head records, and league position movement.

APPROACH & METHODOLOGY

Feature engineering was the central focus of this project — 150+ match-level predictors were constructed from raw results data. XGBoost was selected after benchmarking against Logistic Regression, Random Forest, and a shallow Neural Network on tabular data. SHAP values were used for global feature importance analysis; LIME was used to generate individual match prediction explanations. Seasonal cross-validation was used to simulate realistic forward-prediction scenarios and prevent data leakage.

KEY TECHNICAL HIGHLIGHTS

- Engineered 150+ features: rolling 5-match form, goal difference momentum, venue performance splits, H2H records.
- XGBoost outperformed all benchmarked models including a Neural Network on tabular football data.
- SHAP analysis identified home advantage, recent form, and head-to-head record as the top three predictors.
- LIME provided per-match plain-language explanations for individual prediction outputs.
- Seasonal cross-validation prevented data leakage and simulated real-world forecasting conditions.
- Draw prediction identified as hardest class — consistent with statistical literature on football variance.

KEY INSIGHTS & RESULTS

The model achieved 64% accuracy — substantially above the 33% random baseline for three-class prediction (Home Win / Draw / Away Win). Home advantage carried the highest SHAP value (0.31). Draw prediction remained the weakest class due to football's inherent variance — no model in the literature solves this cleanly. XGBoost's advantage over neural networks on tabular data was confirmed by this experiment.

BUSINESS IMPACT

The SHAP/LIME interpretability layer makes this model suitable for contexts where explaining a prediction matters as much as the prediction itself — sports media, tactical scouting, and betting analysis. The project also validates the academic finding that gradient boosting consistently outperforms deep learning on structured tabular data, adding empirical weight to that literature through a large-scale practical experiment.
