## Advanced Programming 2025

# Predicting FIFA World Cup Tournament Progression Using Machine Learning

Final Project Report

Alexis Gavillet
alexis.gavillet@unil.ch
Student ID: 21814595

January 10, 2026

**Abstract**

Predicting the performance of national football teams in major international tournaments is a challenging task due to limited match observations, strong uncertainty, and structural class imbalance. This project addresses the problem of predicting the stage reached by a national team in the FIFA World Cup using historical and contextual data.

The task is formulated as an ordinal classification problem, where each team is assigned one of six ordered categories ranging from group stage elimination to tournament winner. Multiple publicly available datasets were combined, including World Cup match results, squad information, and historical ranking data. Extensive data preprocessing was performed to harmonize team identifiers, handle missing values, and construct meaningful performance indicators.

Two machine learning models were implemented and compared: a multinomial logistic regression baseline and a gradient boosting model based on XGBoost. Model performance was evaluated using accuracy and macro-averaged F1-score to account for strong class imbalance across tournament stages.

The results show that non-linear models substantially outperform linear baselines, primarily by improving prediction performance on intermediate tournament stages. Group-stage performance variables capture most of the predictive signal, while adding pre-tournament ranking-based and contextual features does not consistently improve generalization. These findings highlight both the potential and the limitations of machine learning approaches for modeling tournament progression in international football.

**Keywords:** data science, machine learning, football analytics, ordinal classification, XGBoost, Python

# Contents

# 1 Introduction

International football tournaments such as the FIFA World Cup attract global attention and represent the highest level of competition between national teams. Predicting team performance in such tournaments is of interest not only to fans and analysts, but also to researchers studying decision-making, uncertainty, and performance modeling in complex systems. Unlike domestic leagues, World Cups are characterized by short time horizons, limited matches, and high variability, making outcome prediction particularly challenging.

The increasing availability of structured sports data has enabled the application of data science and machine learning techniques to football analytics. Previous research has explored match-level outcome prediction, team strength estimation, and player performance analysis [6, 10]. However, fewer studies focus on predicting the overall stage reached by a team in a knockout-style tournament, where progression depends on both group-stage consistency and elimination-round performance.

This project aims to predict the stage reached by each national team participating in a FIFA World Cup. The problem is formulated as an ordinal classification task, where teams are assigned one of six ordered categories ranging from group stage elimination to tournament winner. The objective is to assess whether historical performance indicators, squad characteristics, and ranking-based features can meaningfully explain and predict tournament progression.

To achieve this goal, multiple publicly available datasets were combined, including World Cup match data, squad and player information, and historical FIFA and Elo rankings. Two machine learning models were implemented: a logistic regression model as a baseline and an XGBoost classifier to capture non-linear relationships. Model performance is evaluated using accuracy and macro F1-score to account for class imbalance across tournament stages.

The remainder of this report is structured as follows. Section 2 reviews related work in football analytics and tournament prediction. Section 3 describes the datasets, feature engineering process, and modeling approach. Section 4 presents the experimental results, followed by a discussion of limitations and insights in Section 5. Finally, Section 6 concludes the report and outlines directions for future work.

# 2 Literature Review / Related Work

The use of data science and machine learning in football analytics has expanded significantly over the last two decades, largely driven by the growing availability of structured match, team, and player data. Early contributions in this field mainly focused on match-level outcome prediction, aiming to forecast wins, draws, or losses based on historical results, team strength, and contextual factors. Common methodological approaches include logistic regression models, Poisson-based goal models, and rating systems such as Elo rankings [6].

Rating-based indicators have received particular attention in the literature as proxies for latent team strength. Elo ratings, originally developed for chess, have been widely adapted to football and shown to outperform simple win–loss statistics in international contexts [6]. FIFA rankings, despite methodological limitations and occasional criticism, remain a widely used benchmark and have been incorporated into predictive models for major international tournaments. However, prior research suggests that ranking systems alone may have limited predictive power, especially when applied to tournament-level outcomes rather than individual matches [10].

More recent studies have explored machine learning techniques capable of capturing non-linear relationships and complex interactions between variables. Tree-based models, including

random forests and gradient boosting methods, have demonstrated strong performance in football prediction tasks due to their flexibility and ability to handle heterogeneous feature sets. In particular, gradient boosting models such as XGBoost have been shown to outperform linear baselines in many structured data settings [1, 2].

While the majority of existing research focuses on predicting individual match outcomes, fewer studies address tournament-level performance prediction. Predicting the stage reached by a team in a knockout-style competition introduces additional challenges, including limited sample sizes, strong class imbalance, and dependence between successive matches. Some authors have proposed framing tournament progression as a multi-class or ordinal classification problem, explicitly recognizing the ordered nature of competitive stages [10].

An important distinction in the literature lies between match-level prediction and tournament-level modeling. Match prediction typically considers short-term outcomes under relatively stable conditions, whereas tournament prediction involves longer horizons, structural constraints imposed by competition formats, and cumulative effects across matches. These characteristics substantially increase uncertainty and reduce the amount of available training data.

Building on this literature, the present project focuses on predicting World Cup tournament progression rather than individual match outcomes. By formulating the task as an ordinal classification problem and explicitly accounting for historical variations in tournament formats, this study contributes to the growing body of work on interpretable and reproducible football analytics. In particular, it investigates whether pre-tournament ranking-based indicators provide additional predictive value when combined with observed in-tournament performance variables, an open question in existing research.

# 3 Methodology

## 3.1 Data Description

This project relies on multiple publicly available datasets covering FIFA World Cup tournaments, team compositions, match results, and international football rankings. All datasets were merged to construct a unified modeling dataset with one observation per team per World Cup edition.

World Cup match, squad, and player information were obtained from the *World Cup Match, Squad and Player Data* GitHub repository [4]. The `squads.csv` dataset provides team-level information for each tournament, while `players.csv` contains individual player data that was aggregated at the team level. Match-level results were extracted from `matches.csv` and used to derive indicators related to tournament progression, such as points, goal difference, and win ratios.

To capture team strength beyond observed World Cup performance, multiple ranking datasets were incorporated. Historical Elo ratings spanning from 1901 to 2023 were sourced from the *World Football Elo Ratings* GitHub repository [5]. In addition, FIFA rankings were obtained from two Kaggle datasets covering the periods 1993–2018 and 2022 [9, 7]. These ranking systems provide complementary perspectives on pre-tournament team strength and competitive standing.

Additional tournament-level information, including final rankings (winner, runner-up, third and fourth place), was extracted from the `WorldCups.csv` file available on Kaggle [8]. Missing entries for the 2018 and 2022 tournaments were manually completed using official FIFA World Cup records to ensure data completeness and consistency across editions [3].

After harmonizing team names across datasets, all sources were merged using standardized country identifiers and tournament year. The final dataset includes historical performance in-

dicators, ranking-based contextual features, and the ordinal target variable representing the maximum stage reached by each team during a given World Cup.

### 3.2 Target Variable Construction

The target variable represents the stage reached by a national team during a given FIFA World Cup. Rather than predicting individual match outcomes, the task is formulated as an ordinal classification problem reflecting overall tournament progression.

Each team is assigned an ordinal label ranging from 0 to 5, defined as follows: teams eliminated during the group stage are labeled 0; teams reaching the Round of 16 are labeled 1; teams reaching the quarterfinals are labeled 2; teams reaching the semifinals are labeled 3; losing finalists are labeled 4; and tournament winners are labeled 5. This encoding captures the ordered nature of competitive success while allowing the use of standard multi-class classification models.

Special care was taken to harmonize historical World Cup formats. Several tournaments, notably the 1974, 1978, and 1982 editions, did not include a Round of 16 and instead featured multiple group stages. In these cases, teams qualifying for the second group stage were mapped to a quarterfinal-equivalent category (label 2), ensuring consistency of the ordinal target variable across all tournaments.

Final placements for the top four teams were obtained directly from official World Cup records, guaranteeing correct assignment of the highest ordinal categories. This approach ensures that the target variable reflects comparable levels of achievement across different tournament structures.

### 3.3 Approach

Two distinct feature sets were considered in this study in order to assess the incremental value of contextual information. The baseline feature set consists exclusively of in-tournament performance indicators derived from the group stage, including points, wins, losses, goal difference, average goals scored and conceded, and win ratio. These variables summarize a team's observed performance during the early phase of the tournament.

The enriched feature set extends the baseline features by incorporating pre-tournament and contextual variables, namely historical FIFA and Elo-based strength indicators, host status, and average team age. These features are commonly used in the literature as proxies for intrinsic team quality prior to tournament play.

Comparing baseline and enriched models allows for a direct evaluation of whether contextual and ranking-based information provides additional predictive value beyond observed tournament performance.

Two machine learning models were implemented and compared in this study. A multinomial logistic regression model was used as a baseline due to its interpretability and widespread use in classification tasks. This model assumes linear relationships between the input features and the log-odds of each class.

To capture non-linear interactions and complex feature relationships, a gradient boosting classifier using XGBoost was also implemented. Gradient boosting methods iteratively build ensembles of decision trees and have demonstrated strong performance in structured tabular data problems, particularly in the presence of heterogeneous features.

Prior to model training, data preprocessing steps included handling missing values, scaling numerical features when required, and encoding categorical variables. The dataset was split into training and test sets using a predefined train-test split designed to preserve the distribution of ordinal classes.

Model performance was evaluated using accuracy and macro-averaged F1-score. The macro F1-score was selected to account for class imbalance, particularly the relatively small number of teams reaching the final stages of the tournament.

### 3.4 Implementation

All data processing, modeling, and evaluation steps were implemented in Python. The main libraries used include `pandas` and `numpy` for data manipulation, `scikit-learn` for preprocessing, model evaluation, and logistic regression, and `xgboost` for gradient boosting classification.

The project was structured following a modular design, separating data loading, preprocessing, feature engineering, model training, and evaluation into distinct components. This organization improves code readability and reproducibility.

Experiments were conducted using consistent random seeds to ensure replicable results. Hyperparameters for the XGBoost model were selected based on preliminary experimentation and kept fixed across runs to avoid overfitting to the test set.

The XGBoost model was trained using a fixed set of hyperparameters selected through preliminary manual experimentation on the training data. Key parameters include a limited tree depth to control model complexity, a moderate number of boosting iterations, a low learning rate, and subsampling of both observations and features. This conservative configuration was chosen to reduce the risk of overfitting, given the relatively small dataset and the strong class imbalance.

This code snippet illustrates the construction of the ordinal target variable by mapping heterogeneous tournament stages to a common ordered scale and assigning each team the maximum stage reached during a given World Cup. Final tournament winners were assigned the highest ordinal category based on official World Cup records, ensuring consistency with the ordinal scale.

```
# Map tournament stages to ordinal values
stage_to_yord = {
    "group stage": 0,
    "second group stage": 2,
    "round of 16": 1,
    "quarter-finals": 2,
    "semi-finals": 3,
    "third-place match": 3,
    "final": 4,
    "winner":5
}

# Normalize stage names
stages_long["stage_name"] = (
    stages_long["stage_name"]
    .str.strip()
    .str.lower()
)

# Map stages to ordinal values
stages_long["y_ord_tmp"] = stages_long["stage_name"].map(stage_to_yord)

# Assign each team the maximum stage reached during the tournament
y_from_matches = (
    stages_long
    .groupby(["tournament_year", "team_code"], as_index=False)["y_ord_tmp"]
    .max()
    .rename(columns={"y_ord_tmp": "y_ord"})
)
```

Listing 1: Construction of the ordinal target variable from tournament stages

# 4 Results

## 4.1 Experimental Setup

All experiments were conducted using Python on a standard personal computing environment. The implementation relies on widely used open-source libraries, including `pandas` and `numpy` for data manipulation, `scikit-learn` for preprocessing, model training, and evaluation, and `xgboost` for gradient boosting classification.

The dataset was split into training and test sets based on tournament editions, ensuring that entire World Cups were held out during evaluation. This setup prevents information leakage between training and test data and reflects a realistic prediction scenario. The resulting split contains 304 observations in the training set and 64 observations in the test set.

The training and test split was performed at the tournament edition level, ensuring that entire World Cups were held out during evaluation. This strategy prevents information leakage between training and test sets and reflects a realistic prediction scenario in which future tournaments are predicted based on historical data.

Alternative validation strategies, such as leave-one-tournament-out cross-validation, were considered. However, given the limited number of World Cup editions and the resulting instability of performance estimates, a single train-test split was preferred for clarity and robustness. As a consequence, the reported results should be interpreted with caution, as performance remains sensitive to the choice of split, an unavoidable limitation when working with small historical sports datasets.

Two types of models were evaluated: a multinomial logistic regression baseline and an XGBoost classifier. The logistic regression model was trained using standardized input features and balanced class weights. The XGBoost model was trained with a fixed set of hyperparameters, including a limited tree depth and a moderate number of estimators, selected through preliminary experimentation.

Model performance was assessed using accuracy and macro-averaged F1-score. Accuracy provides a global measure of prediction correctness, while the macro F1-score accounts for class imbalance by weighting each tournament stage equally.

## 4.2 Performance Evaluation

Table 1 summarizes the predictive performance of the evaluated models on the held-out test set. Results are reported in terms of accuracy, macro-averaged F1-score, and weighted F1-score.

Table 1: Model Performance on the Test Set

| **Model** | **Accuracy** | **Macro F1** | **Weighted F1** |
|---|---|---|---|
| Logistic Regression (baseline) | 0.625 | 0.345 | 0.645 |
| Logistic Regression (enriched) | 0.594 | 0.312 | 0.636 |
| XGBoost (baseline) | **0.641** | **0.449** | 0.664 |
| XGBoost (enriched) | 0.625 | 0.307 | 0.637 |

## 4.3 Visualizations

Figure 1 illustrates the distribution of the ordinal target variable across all observations. As expected, advanced tournament stages are underrepresented due to the knockout structure of the competition.

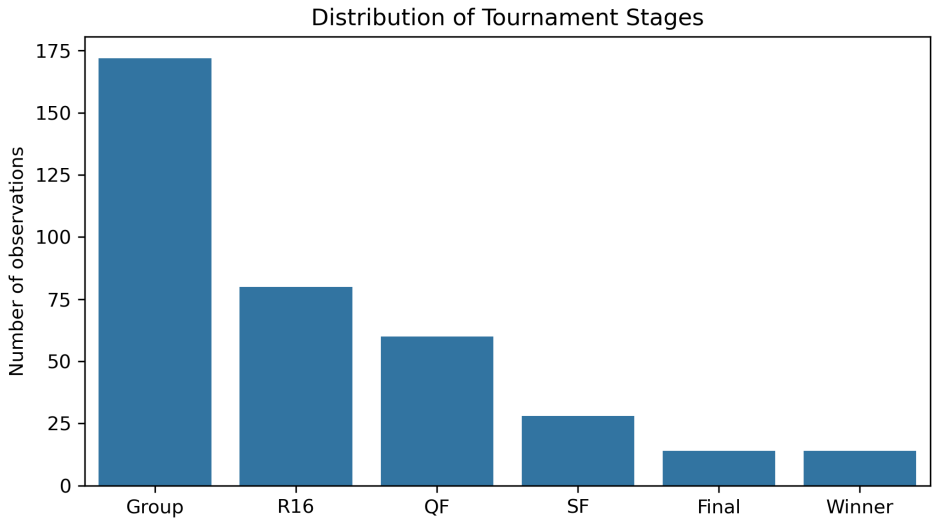


Figure 1: Distribution of tournament stages (ordinal target variable).

Figure 2 compares model performance in terms of accuracy and macro-averaged F1-score. While accuracy differences are limited, substantial variations are observed in macro F1-score, highlighting the importance of balanced evaluation metrics.

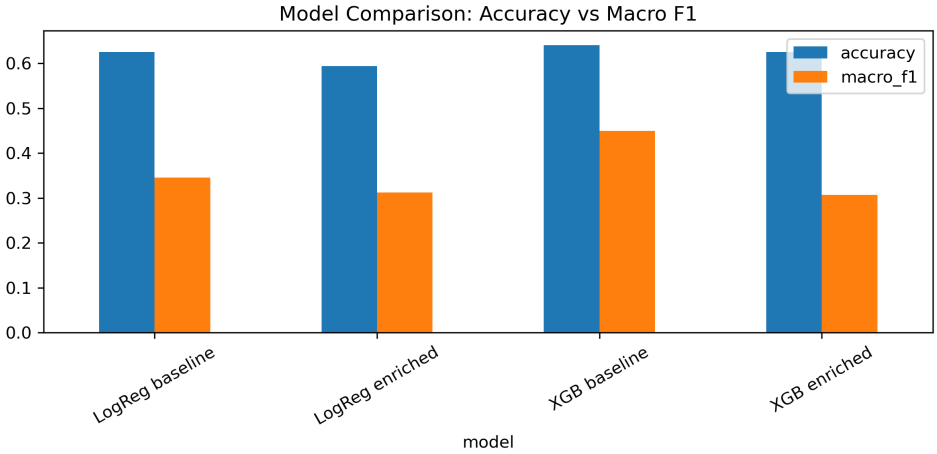


Figure 2: Comparison of model performance using accuracy and macro F1-score.

Figure 3 presents the row-normalized confusion matrix for the best-performing model (XGBoost baseline). The model achieves improved recall for intermediate stages, which explains its superior macro-averaged F1-score.

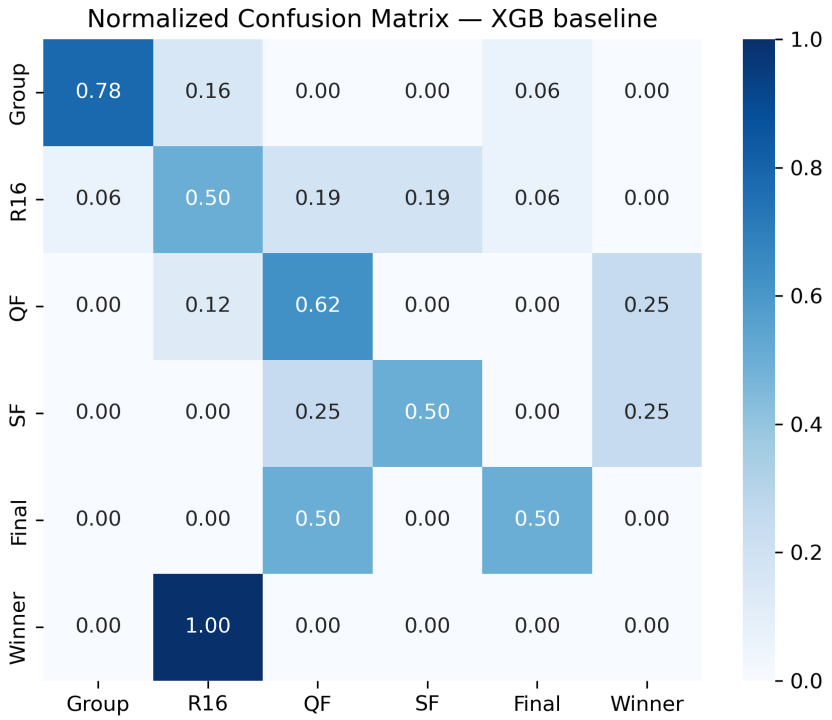


Figure 3: Normalized confusion matrix for the XGBoost baseline model.

## 5 Discussion

The results of this study show that predicting World Cup tournament progression as an ordinal classification problem is feasible, despite the strong structural constraints of international tournaments. Among the evaluated models, the XGBoost baseline achieved the best overall performance, particularly in terms of macro-averaged F1-score. This indicates that non-linear models are better suited to capture the complex relationships between group-stage performance indicators and subsequent tournament progression.

A key finding is that enriched feature sets, including FIFA and Elo rankings, host status, and average team age, do not consistently improve predictive performance. While these variables are commonly used as proxies for pre-tournament team strength, the results suggest that group-stage performance already captures most of the relevant information needed to predict how far a team advances. This observation is consistent with the strong correlations observed between early tournament outcomes and final progression.

An analysis of the confusion matrices further shows that most prediction errors occur between adjacent tournament stages, such as between the Round of 16 and the quarterfinals or between the quarterfinals and semifinals. Severe misclassifications across distant stages are rare. This error structure aligns well with the ordinal nature of the target variable and suggests that the model learns a meaningful ordering of team performance, even when exact stage prediction remains difficult.

Nevertheless, several limitations must be acknowledged. The dataset is inherently small, as World Cups occur only every four years and involve a limited number of teams. This results in strong class imbalance, particularly for finalists and tournament winners, which limits predictive performance for the highest stages. Moreover, the use of standard multi-class classifiers does not explicitly exploit the ordered nature of tournament stages, potentially penalizing small and large errors equally.

Finally, historical variations in tournament formats introduce additional uncertainty. Although careful harmonization was performed, differences in competition structure may still affect model stability. These limitations highlight the need for cautious interpretation of the results and motivate the exploration of ordinal-specific models and alternative validation strategies in future work.

# 6 Conclusion and Future Work

## 6.1 Summary

This project investigated the problem of predicting national team performance in FIFA World Cups by modeling tournament progression as an ordinal classification task. By combining historical match results, team-level performance indicators, and ranking-based contextual features, a unified dataset was constructed with one observation per team per tournament.

Two machine learning approaches were implemented and compared: a multinomial logistic regression baseline and a gradient boosting model using XGBoost. The results demonstrate that non-linear models provide improved performance, particularly when evaluated using macro-averaged F1-score, which better reflects predictive quality under strong class imbalance. Overall, the XGBoost baseline model achieved the best balance between accuracy and class-level performance.

Beyond predictive performance, this project highlights the importance of careful target variable construction and historical harmonization when working with long-term sports data. By explicitly addressing variations in World Cup formats and leveraging an ordinal outcome representation, the study provides a reproducible framework for analyzing tournament-level success in international football.

## 6.2 Future Directions

From a methodological perspective, future work could more explicitly exploit the ordinal nature of the target variable. In this project, tournament progression was modeled using standard multiclass classification algorithms, which treat all misclassifications equally, even though predicting a semifinalist as a finalist is less severe than predicting a group-stage elimination as a tournament winner. Ordinal-specific approaches such as ordinal logistic regression, cumulative link models, or ordinal-aware loss functions could address this limitation by penalizing errors according to their distance on the ordinal scale. These methods explicitly incorporate the ordering between classes and may lead to more meaningful performance evaluation in tournament prediction tasks. Such models were not implemented in the present study due to practical constraints, including limited library support in standard machine learning frameworks, the relatively small sample size, and the scope of the course, but they represent a promising direction for improving both predictive performance and interpretability.

Additional experiments could also focus on probabilistic and simulation-based approaches. For instance, combining match-level win probabilities with Monte Carlo simulations of entire tournaments could provide more granular predictions and allow the estimation of uncertainty around progression outcomes. Incorporating temporal dynamics, such as recent form or changes in team composition shortly before a tournament, may further enhance predictive performance.

Finally, the proposed framework has potential real-world applications beyond retrospective analysis. Similar methods could be applied to upcoming international tournaments to support scenario analysis, media forecasting, or decision-making in sports analytics contexts. More generally, the approach illustrates how structured historical data can be used to model progression in elimination-based competitions across different domains.

## References


[1] Rohit Baboota and Harleen Kaur. "Predictive analysis and modelling football results using machine learning approaches". In: *International Journal of Forecasting* 34.4 (2018), pp. 741–755.

[2] Tianqi Chen and Carlos Guestrin. "XGBoost: A scalable tree boosting system". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 785–794.

[3] Fédération Internationale de Football Association. *FIFA World Cup Archive*. `https://www.fifa.com/tournaments/mens/worldcup`. Used to manually complete missing final rankings for the 2018 and 2022 tournaments. 2024.

[4] Joshua Fjelstul. *World Cup match, squad and player data*. `https://github.com/jfjelstul/worldcup/tree/master/data-csv`. 2023.

[5] Julien Gravier. *World football Elo ratings (1901–2023)*. `https://github.com/JGravier/soccer-elo/blob/main/csv/ranking_soccer_1901-2023.csv`. 2023.

[6] Lars Magnus Hvattum and Henrik Arntzen. "Using Elo ratings for match result prediction in association football". In: *International Journal of Forecasting* 26.3 (2010), pp. 460–470.

[7] Kaggle. *FIFA Football World Cup rankings dataset*. `https://www.kaggle.com/datasets/piterfm/fifa-football-world-cup`. 2022.

[8] Kaggle. *FIFA World Cup dataset*. `https://www.kaggle.com/datasets/abecklas/fifa-world-cup`. 2022.

[9] Kaggle. *FIFA world rankings (1993–2018)*. `https://www.kaggle.com/datasets/espsiyam/fifa-ranking-19932018`. 2018.

[10] Jakub Lasek, Zoltán Szlávik, and Sandjai Bhulai. "The predictive power of ranking systems in association football". In: *International Journal of Applied Pattern Recognition* 1.1 (2013), pp. 27–46.

[11] Simon Scheidegger and Anna Smirnova. *Data Science and Advanced Programming 2025*. `https://ap-unil-2025.github.io/course-materials/`. Course materials for Master's students in Economics and Finance. 2025.

# A Additional Figures

This appendix presents supplementary figures that provide additional insights into the data and modeling process but are not essential to the main narrative of the report. These figures support the exploratory data analysis and help illustrate intermediate results discussed in earlier sections.

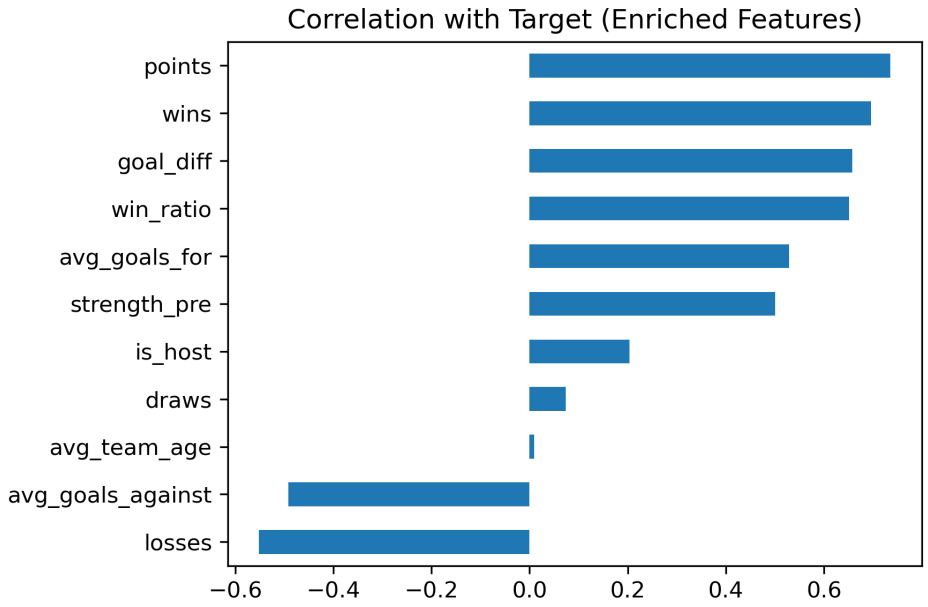


Figure 4: Correlation between enriched input features and the ordinal target variable.

# B AI Tools Used

AI-assisted tools (ChatGPT and GitHub Copilot) were used for limited support during development and writing. The project was developed within the framework of the Data Science and Advanced Programming course at HEC Lausanne, following the methodological and reproducibility guidelines provided in the course materials [11].

# C Code Repository

**GitHub Repository:** `https://github.com/Alexg285/Road-to-glory`

The repository contains the full source code, datasets, and notebooks required to reproduce the results presented in this report. The project is organized as follows:

- `data/`: raw and cleaned datasets used in the analysis
- `notebooks/`: data cleaning and exploratory data analysis notebooks
- `src/`: modular Python source code for data loading, modeling, and evaluation
- `results/`: saved metrics and generated figures used in the report
- `main.py`: entry point for model training and evaluation
- `environment.yml`: Conda environment specification
- `README.md`: project overview, installation instructions, and usage guidelines

- `AI_USAGE.md`: documentation of how AI-assisted tools were used during the project

**Reproducibility:** The numerical results (model training and evaluation metrics) can be reproduced by running the main script from the repository root after creating and activating the Conda environment `worldcup-project`:

```
conda env create -f environment.yml
conda activate worldcup-project
python main.py
```

This script loads the final cleaned dataset, trains the baseline and enriched models, evaluates them on the held-out test set, and saves the performance metrics to `results/metrics.csv`.

The figures included in the report (plots in `results/plots/`) are generated by running the notebook `notebooks/EDA_Results.ipynb`.