The Impact of Formations on Football Matches Using Double Machine Learning. Is it worth parking the bus?

Genís Ruiz Menàrguez[1*],Llorenç Badiella Busquets[2],

[1]Graduate in Data Science, Universitat Autònoma de Barcelona. **Genis.Ruiz@autonoma.cat**
[2]Department of Statistics Professor,Universitat Autònoma de Barcelona. **llorenc.badiella@uab.cat**

ABSTRACT

[ENGLISH] This study addresses a central tactical dilemma in modern football: Is it worth parking the bus? Using an advanced Double Machine Learning (DML) framework, the project estimates the causal impact of defensive versus offensive formations on key match outcomes such as goal difference, possession, corners, and disciplinary actions. Leveraging a dataset of over 22,000 matches from top European leagues, formations were categorized into six representative types based on tactical structure and expert consultation. A major methodological contribution lies in the adaptation of DML to handle categorical treatments, specifically formation combinations, through a novel matrix-based residualization process—allowing for a detailed estimation of formation-versus-formation effects. Results show that while offensive formations like 4-3-3 and 4-2-3-1 offer modest statistical advantages in possession and corners, the impact on goals is limited, and no evidence supports the idea that defensive formations—commonly associated with parking the bus—increase a team's scoring potential. Additionally, red cards appear unaffected by formation choice, suggesting other behavioral factors dominate. Complementary SHAP value analysis confirms that match context, particularly home advantage, possession, and corners, are more influential than formation alone. Although this approach does not fully capture all aspects of playing style or team strength, it provides a robust and interpretable framework for formation analysis and sets a valuable precedent for future research in sports analytics.

ABSTRACT

[CAT] Aquest estudi aborda un dels grans dilemes tàctics del futbol modern: val la pena aparcar l'autobús? Fent ús del Double Machine Learning (DML), el projecte estima l'impacte causal de les formacions defensives enfront de les ofensives sobre diferents variables resposta, com ara la diferència de gols, la possessió, els córners i les targetes grogues o vermelles. A partir d'un conjunt de dades de més de 22.000 partits de les principals lligues europees, les formacions es van agrupar en sis diferents basant-se en criteris tàctics i assessorament d'experts. En l'estudi s'ha aplicat una modificació metodològica rellevant, que consisteix en l'adaptació del DML per gestionar variables de tractament categòriques. Concretament, es tracten combinacions de formacions a través d'un procés innovador de residualització matricial, que permet estimar detalladament els efectes de cada enfrontament tàctic. Els resultats mostren que, tot i que les formacions ofensives com el 4-3-3 i el 4-2-3-1 ofereixen certs avantatges estadístics en possessió i córners, l'impacte sobre els gols és limitat, i no hi ha evidència que les formacions defensives, com les habitualment associades a aparcar l'autobús, millorin la capacitat anotadora d'un equip. A més, les targetes vermelles no semblen ser influïdes per l'elecció de la formació, cosa que suggereix que hi intervenen factors més aleatoris o conductuals. L'anàlisi complementària mitjançant valors SHAP confirma que el camp de joc del partit, especialment el fet de jugar a casa, la possessió i els córners acumulats, tenen més pes que la formació per si sola. Tot i que aquest enfocament no captura totalment l'estil del joc ni la força dels equips (*team strength*), proporciona un marc sòlid i interpretable per a l'anàlisi de formacions i estableix un precedent valuós per a futures investigacions en anàlisi esportiva.

ABSTRACT

[ESP] Este estudio aborda uno de los grandes dilemas tácticos del fútbol moderno: ¿vale la pena aparcar el autobús? Utilizando un marco avanzado de Double Machine Learning (DML), el proyecto estima el impacto causal de las formaciones defensivas frente a las ofensivas sobre distintos variables respuesta, como la diferencia de goles, la posesión, los córners y las acciones disciplinarias. Con un conjunto de datos que abarca más de 22.000 partidos de las principales ligas europeas, las formaciones se agruparon en seis, según criterios tácticos y asesoramiento de expertos. Una contribución metodológica destacada del estudio es la adaptación del DML para tratar variables categóricas. Específicamente, se tratan combinaciones de formaciones a través de un novedoso proceso de residualización matricial, lo que permite estimar de forma detallada los efectos de cada enfrentamiento táctico. Los resultados muestran que, aunque las formaciones ofensivas como el 4-3-3 y el 4-2-3-1 ofrecen ciertas ventajas estadísticas en posesión y córners, el impacto sobre los goles es limitado, y no hay evidencia de que las formaciones defensivas, comúnmente asociadas con aparcar el autobús, mejoren la capacidad anotadora de un equipo. Además, las tarjetas rojas no parecen estar influenciadas por la elección de la formación, lo que sugiere que predominan factores más aleatorios o conductuales. El análisis complementario mediante valores SHAP confirma que el campo de juego del partido, en particular el hecho de jugar en casa, la posesión y los córners, tienen más peso que la formación en sí misma. Aunque este enfoque no captura por completo todos los aspectos del estilo de juego ni de la fortaleza (*team strength*) de los equipos, proporciona un marco sólido e interpretable para el análisis de formaciones y sienta un precedente valioso para futuras investigaciones en analítica deportiva.

Keywords: Causal Inference – Confounding – Double Machine Learning – Formation – SHAP Values – XGBoost

1. INTRODUCTION

Football coaches must decide which formation to implement in each game. The eleven starting players are typically organized according to specific structural patterns. This project aims to analyze whether certain defensive formations, commonly referred to as *parking the bus* (8), are truly effective in specific scenarios. José Mourinho, former Chelsea FC manager, popularized this term after a 0–0 draw against Tottenham Hotspur in 2004, using it to emphasize the opponent's excessively defensive approach during the fixture.

These defensive strategies are often employed against stronger opponents, yet their actual effectiveness remains debated among football analysts. Therefore, the purpose of this study is to examine the direct relationship between formations and several match outcome variables. After generalizing and grouping formations based on tactical similarity and expert criteria, a Double Machine Learning framework is applied to estimate the intrinsic impact of each formation combination (i.e., a team's formation against that of its rival) on various target variables, such as goal difference. Although formations may not be the most influential variables in predicting match statistics, there is strong interest in determining whether specific systems consistently enhance or hinder key performance outcomes.

As stated above, the objective is to assess the isolated impact of formations themselves, rather than focusing on the individual characteristics or quality of particular teams.

2. DATA SOURCES

To access the match data, the Sportmonks(12) API (https://www.sportmonks.com/) was used. This API contains a wide range of match features involving numerous teams from different countries and divisions. The Standard plan enabled access to a comprehensive set of football match data.

2.1. Match Data

Since the goal is to analyze formation data, some important decisions have been made. Firstly, following the COVID-19 pandemic, several regulatory changes were implemented that directly affected player and team performance. The most relevant change was the introduction of up to five substitutions per match in regular league fixtures, allowing coaches greater flexibility in squad rotation and enabling players to perform at a more demanding level. Thus, the majority of the selected seasons reflect this updated regulation. These seasons are 2021–2022, 2022–2023, 2023–2024, and 2024–2025. Additionally, to broaden the dataset and improve the statistical significance of the resulting models, seasons 2018–2019, 2019–2020 (excluding the pandemic period), and 2020–2021 have also been considered.

Secondly, several leagues were included to ensure the study is based on a robust and representative dataset. According to UEFA(14), the seven highest-ranked European men's football leagues were selected: the first divisions of England, Italy, Spain, Germany, France, the Netherlands, and Portugal. To expand the dataset further, other leagues were also included based on two criteria: having more than 32 regular matches (i.e., at least 16 teams per league) and having sufficient data availability in Sportmonks. Consequently, the top divisions of Turkey, Belgium, and Poland, as well as the second divisions of Spain, Italy, and England, were added.

The Sportmonks API is a rich source of variables and has provided the following data:

- **General Data:** Local and away teams, date, league, season, and fixture round.
- **Statistics:** Number of goals, corners, possession, and red and yellow cards for both teams.
- **Formations:** Formation system used by both teams (e.g., 4-3-3).
- **Lineups:** A list of players and their personal information for both teams.
- **Events:** A time-ordered, detailed list of match events (e.g., penalties or substitutions).
- **Weather:** Temperature and other weather variables at the match venue (e.g., pressure, humidity, or wind).

Since the objective is to assess the impact of each formation on various target variables (initially, goal difference), some variables have been excluded from the analysis because they are not considered confounders.

3. DATA CLEANING AND FEATURE ENGINEERING

3.1. Data Cleaning

Some Pro League (Belgian league) and Turkish Süper Lig seasons were excluded due to missing match data, preventing meaningful feature engineering. Additionally, play-offs, play-outs, and other non-regular rounds were removed, as their nature differs from standard league matches. The first two rounds of each season were also discarded due to the low statistical strength of team performance indicators at that stage. Furthermore,

the final four rounds were excluded, since teams with their season goals already achieved might experiment with alternative formations or give playing time to younger players, potentially altering match dynamics in a way unrelated to team strength.

3.2. Feature Engineering

Some variables were created to quantify team strength. However, odds data from betting houses were excluded, as such data may implicitly account for formation choices, potentially introducing bias. Examples of team strength variables include:

- **Accumulated Points:** Numerical variables representing the ratio of points obtained out of the possible total in past matches, separately for home and away performances.
- **Champions League Flag:** A binary variable indicating whether the home or away team is participating in the UEFA Champions League at the time of the fixture.
- **Team Ranking:** An integer variable reflecting the league position of each team. In the case of equal points, the same rank is assigned to both teams.

As stated in the Introduction section, the goal is to analyze the pure effect of formations on match outcomes. Therefore, variables likely to be structurally related to formations—such as accumulated goals or fouls—were not included, to avoid capturing formation effects indirectly.

4. DOUBLE MACHINE LEARNING (DML)

4.1. Formations Confusion

This project aims to assess the true impact of formations on various target variables, including goal difference. The matrix 1 shows the mean goal difference between home and away team formation combinations.

P-values indicating statistical significance are shown below each mean:

- (***) : p-value < 0.001
- (**) : p-value < 0.01
- (*) : p-value < 0.05
- (ns) : p-value > 0.05

Cells with a p-value below 0.05 are considered statistically significant, as the null hypothesis that the mean equals zero is rejected.

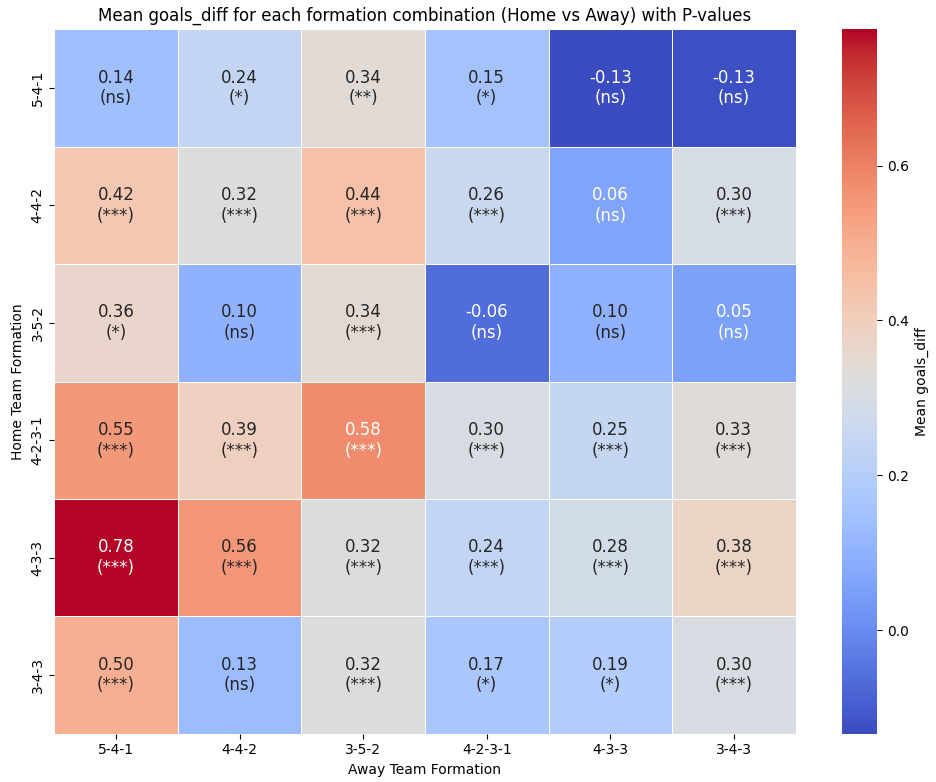


Figure 1. Mean goal difference by home and away team formation combinations

Some specific formation combinations clearly correlate with higher average goal differences. For instance, a home team using a 4-3-3 against an away team using a 5-4-1 yields an average goal difference of +0.78. This implies that, over five such matches, the home team would score approximately four more goals in total—a substantial margin. The most notable advantages are found in offensive formations (4-2-3-1, 4-3-3, and 3-4-3) against more defensive setups.

However, these raw means do not provide valid causal insights, as they conflate formation effects with contextual factors. Notably, home teams tend to win more frequently due to factors like home-field advantage, crowd support, and rest time.

Furthermore, some extreme results may be misleading. For example, top-ranked teams rarely use formations such as 5-4-1 or 3-5-2, which may partially explain the lower goal averages associated with those systems. To address these biases and identify genuine formation effects, a proper statistical method that accounts for confounding variables is required—Double Machine Learning serves this purpose.

4.2. Introduction to Double Machine Learning methods

By using adaptable machine learning models to handle confounding variables, Double Machine Learning (DML (5)) offers benefits for researching how football formation style affects goals difference. Based on confounders, it forecasts both formation and aims differences before examining the residual link to determine the causal effect. DML is a powerful tool for causal inference in this setting because it can handle complex relationships, high-dimensional data, and provides more reliable causal estimates than older methods. As it is an extension of Frisch-Waugh-Lowell (FWL) (7) procedure, it keeps its statistics properties and it adjust to a variety of real-world situations.

While estimating causal effects we may face two problems:

1. Confounding bias: There are some variables which influence both treatment variable (formation) and target variable (goals difference).
2. Overfitting: Some models are so simple to predict properly and others will not extend correctly to unseen data.

DML is a great approach that deals with the previous issues. For a detailed explanation, see Section 4.4.

4.3. Methodology

4.3.1. Original Model

We consider the following linear model:

$$Y = D\beta + X\gamma + \varepsilon$$

where,

- $Y \in \mathbb{R}^n$: outcome variable (e.g., goal difference, which is discrete but will be treated as continuous),
- $D \in \mathbb{R}^{n \times p_1}$: treatment variable (e.g., formation combination of both teams), D could also be $\mathbb{R}^n$
- $X \in \mathbb{R}^{n \times p_2}$: matrix of control variables (confounders),
- $\beta \in \mathbb{R}^{p_1}$: causal parameter of interest,
- $\gamma \in \mathbb{R}^{p_2}$: nuisance parameters,
- $\varepsilon \in \mathbb{R}^n$: error term.

Where n is the number of rows (number of fixtures $\cdot$ 2) and p_1 and p_2 are the number of variables of D and X respectively.

4.3.2. Three-Step Frisch-Waugh-Lowell (FWL) Procedure

Step 1: Regress outcome on controls

$$Y = \gamma X + \varepsilon_Y$$

And get the predicted $\hat{Y}$.

$$r_Y = Y - \hat{Y}$$

Where r_Y is the residual of the target variable prediction.

Step 2: Regress treatment on controls

$$D = \alpha X + \varepsilon_D$$

And get the predicted $\hat{D}$.

$$r_D = D - \hat{D}$$

Where r_D is the residual of the treatment variable D prediction.

Step 3: Regress residuals of outcome on residuals of treatment

$$r_Y = \beta r_D + \epsilon$$

The estimator for β is given by:

$$\hat{\beta} = \frac{r_D^\top r_Y}{r_D^\top r_D}$$

4.3.3. Double Machine Learning Extension

In DML, the projection step is generalized using Machine Learning (ML). Instead of linear projections, we estimate residuals as:

$$r_Y = Y - \hat{f}(X), \quad r_D = D - \hat{g}(X)$$

where:

- $\hat{f}(X)$: ML estimator of Y on X,
- $\hat{g}(X)$: ML estimator of D on X.

Both $\hat{f}(X)$ and $\hat{g}(X)$ can be adaptive to high dimensional spaces, and capture nonlinearities, interactions, and complex dependencies automatically.

The final stage remains:

$$r_Y = \beta r_D + \epsilon \quad \Rightarrow \quad \hat{\beta} = \frac{r_D^\top r_Y}{r_D^\top r_D}$$

Through sample splitting and cross-fitting, this modification maintains orthogonality and robustness while allowing the estimation of treatment effects in high-dimensional or nonlinear environments. Some examples of this extension include using boosting or decision trees models as a substitution of traditional linear models.

4.4. DML Properties

Some of the main properties are:

- **Orthogonalization**: In the DML procedure, the moment condition used to estimate β satisfies a Neyman orthogonality property (3). Formally:

$$\mathbb{E}\left[(r_Y - \beta r_D) \times r_D\right] = 0$$

where:

$$r_Y = Y - \hat{f}(X)$$

$$r_D = D - \hat{g}(X)$$

This implies that small estimation errors in $\hat{f}(X)$ and $\hat{g}(X)$ do not impact the consistency of $\hat{\beta}$ significantly, as their first-order influence cancels out.

- **Flexibility**: DML enables to use different types of ML models for both $\hat{f}(X)$ and $\hat{g}(X)$. Some examples range from simple linear models to decision trees, boosting methods or even neural networks.

• **Statistical Inference Validity**: Even after using complex ML methods in the first stage, the final estimator:

$$\hat{\beta} = \frac{r_D^\top r_Y}{r_D^\top r_D}$$

is still:

– $\sqrt{n}$-consistent, and
– Asymptotically normal.

Formally, under regular conditions:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, V)$$

where V is a consistently estimable variance.

Thus, standard statistical inference remains valid and hypothesis testing or confidence intervals can be calculated. This approach troubleshoots the black-box ML nature.

• **Double Robustness**: If $\hat{f}(X)$ or $\hat{g}(X)$ is estimated poorly, but the other function is accurate enough, then $\hat{\beta}$ will still be consistent. Therefore, just one of the two nuisance estimators needs to be "good enough" (with an estimation rate faster than $n^{-1/4}$) for β to be correctly estimated. This property is really helpful as in real-world situations is not common to have both nuisance models well estimated.

4.5. New Approach for DML with categorical treatment

4.5.1. Methodology

To consider each formation combination in the variable D, dummy variables for $k^2 - 1$ unique formation combinations will be generated, where k is the number of different formations and it is equal for both main and rival team. The goal is to identify which formations are prone to contribute to a higher value in the target variable (e.g. goals difference), meaning winning by a larger margin in the case of goals. To do so, each treatment variable vector $D_{i,j}$ correspond to the combination of the main team formation i and rival team formation j. Therefore, the the resulting models can be expressed in a two-dimensional matrix.

$$\begin{pmatrix} D_{1,1} & = \alpha_{1,1}X + \varepsilon_{D_{1,1}} & \cdots & D_{1,k-1} & = \alpha_{1,k-1}X + \varepsilon_{D_{1,k-1}} & D_{1,k} = \alpha_{1,k}X + \varepsilon_{D_{1,k}} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ D_{k,1} & = \alpha_{k,1}X + \varepsilon_{D_{k,1}} & \cdots & D_{k,k-1} & = \alpha_{k,k-1}X + \varepsilon_{D_{k,k-1}} & \end{pmatrix}$$

Where,

• $D_{i,j} \in \mathbb{R}^n : D_{i,j}$ is a dummy vector which contains 1 in the rows where exists the combination of main team formation i and rival team formation j, -1 if the row corresponds to the unused formation (D_{k^2}) and 0 otherwise.

• $\alpha_{i,j} \in \mathbb{R}^p$ is a vector of coefficients corresponding to each variable in the $X \in \mathbb{R}^{n \times p}$ matrix.

Note that the $D_{k,k}$ formation combination is ommited, as it is not used to avoid multicollinearity. Nevertheless, this last formation data can be deduced from the other ones.
This encoding scheme enables to gather the true effect of each formation combination $D_{i,j}$ on the target variable Y. (See effect coding (13)).

$$\begin{pmatrix} r_{1,1} & = D_{1,1} - \hat{D}_{1,1} & \cdots & r_{1,k-1} & = D_{1,k-1} - \hat{D}_{1,k-1} & r_{1,k} = D_{1,k} - \hat{D}_{1,k} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ r_{k,1} & = D_{k,1} - \hat{D}_{k,1} & \cdots & r_{k,k-1} & = D_{k,k-1} - \hat{D}_{k,k-1} & \end{pmatrix}$$

Where $r_{i,j}$ is the residual of the treatment variable $D_{i,j}$ prediction.

Consequently, the third model is computed:

$$r_Y = \beta_{1,1} r_{D_{1,1}} + \cdots + \beta_{1,k} r_{D_{1,k}} + \cdots + \beta_{k,k-1} r_{D_{k,k-1}} + \epsilon$$

Estimator of $\beta_{i,j}$:

$$\hat{\beta}_{i,j} = \frac{r_{D_{i,j}}^\top r_Y}{r_{D_{i,j}}^\top r_{D_{i,j}}} \quad \text{for } i, j = 1, 2, \ldots, k \text{ and not } i = j = k$$

The greater the $\beta_{i,j}$ is, the greater positive impact it will have on the target variable Y.

Notice how can the estimator $\hat{\beta}_{k,k}$ be obtained:

Since

$$\sum_{i=1}^{k} \sum_{j=1}^{k} \beta_{i,j} = 0$$

$$\hat{\beta}_{k,k} = -(\hat{\beta}_{1,1} + \cdots + \hat{\beta}_{1,k} + \cdots + \hat{\beta}_{k,k-1})$$

As $\beta_{k,k}$ belongs to the diagonal estimators, $\beta_{k,k} = 0$.

4.5.2. Categorical Treatment Properties

Applying this new approach, the goal is to orthogonalize the treatment D with respect to the confounders X. After regressing each dummy D_i, j on X and taking residuals $r_{D_i,j}$, each residual r_{D_i} is orthogonal to the controls X. As any correlation between D_i, j and Y that was due to X has been purged, the remaining correlation between $r_{D_i,j}$ and r_Y also captures the true causal effect.

However, this new approach leads to a different interpretation of the parameter β. Each coefficient β_i, j estimates the causal effect on goal difference of choosing the the main team formation i given the rival team formation j in comparison to the omitted combination $D_{k,k}$.

4.6. Project Methodology

To tackle the formation impact analysis, some changes have been made comparing to the original DML structure.

1. Firstly, the first regression model is an XGBoost Regressor, which builds an ensemble of decision trees sequentially, each one correcting the errors of the previous. (4)

XGBoost Regressor Definition (with MSE (6))

Let $X \in \mathbb{R}^{n \times p}$ be the matrix of confounders and $Y \in \mathbb{R}^n$ the target variable. The XGBoost regressor approximates the mapping $Y \approx f(X)$ using an additive model:

$$\hat{Y} = \sum_{t=1}^{T} f_t(X), \quad f_t \in \mathcal{F}$$

where each f_t is a regression tree.

The model minimizes the regularized mean squared error:

$$\mathcal{L} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{t=1}^{T} \Omega(f_t)$$

with the regularization term:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \|w_t\|^2$$

where w_t are the leaf weights of tree f_t, and γ, λ are regularization hyperparameters.

And get the predicted $\hat{Y}$.

$$r_Y = Y - \hat{Y}$$

Where r_Y is the residual of the target variable prediction.

2. Secondly, as stated before, regression model is computed for each formation combination (leaving one out).

3. Thirdly, the residuals from both previous models are used to compute a final regression model which will provide $\hat{\beta_{i,j}}$ estimations.

5. FORMATION ANALYSIS

This approach consists on analyzing the impact of different feature strategies on the match result. To do so, different formations from the data have been grouped into 6 different ones, sorting them from the most defensive to the most offensive one. This grouping has been carried out with the help of two football experts. Juan Jesús Rodríguez, an exprofessional football player and experienced coach, Juan Camilo Vázquez, a veteran football coach.

5.1. Formation Treatment

The formations are chains of numbers, which correspond to the number of players of each playing line sorted from the most defensive to the most offensive line (mainly defenders, midfielders and forwards). Notice that the goalkeeper is normally not counted, so the total of players must sum ten.

Initially, there were 28 different formations. There were too many in order to compute a regression model and some of them where variations of the main ones, so the most important and distinctive between each other have been selected. The remaining formations have been grouped into these Top 6, with the help of Juan Jesús Rodríguez and Juan Camilo Vázquez (see 7.1).

The different grouped formations sorted from the most defensive to the most offensive and their proportion of real cases in our dataset:

Formation	**5-4-1**	**4-4-2**	**3-5-2**	**4-2-3-1**	**4-3-3**	**3-4-3**
Local (%)	7.09	15.96	14.39	30.17	20.11	12.28
Away (%)	8.11	15.28	15.52	28.67	19.74	12.68

Notice that the proportion of formations used when playing as a local or away team are almost equal. Furthermore, the most used formations are the 4-2-3-1, 4-3-3 and 4-4-2 in that order.

Find below some fixture statistics groupped by formation and still ordered from most defensive to most offensive ones.

Formation	**5-4-1**	**4-4-2**	**3-5-2**	**4-2-3-1**	**4-3-3**	**3-4-3**
Goals Avg	1.236	1.249	1.268	1.392	1.437	1.341
Red Cards Avg	0.110	0.114	0.104	0.106	0.103	0.098
Yellow Cards Avg	2.169	2.246	2.215	2.110	2.075	2.188
Possession Avg	48.98%	48.62%	48.24%	50.72%	51.63%	49.80%
Corners Avg	4.751	4.719	4.754	5.004	5.107	4.854

Some formations such as 4-2-3-1 or 4-3-3 exhibit higher goals or possesion values on average compared to other formations. Nonetheless, this results are not enough to infer the impact of each formation, since there are several correlations and a high confusion beneath the results.

Finally, the resulting lineups have been graphically displayed below, from most defensive to most offensive one, using *lineup11.net* (9).

5-4-1, RCD Mallorca 23-24

4-4-2, Leipzig FC 23-24

3-5-2, FC Inter. Milano 23-24

4-2-3-1, Athletic Club 23-24

4-3-3, FC Barcelona 23-24

3-4-3, Bayern Leverkusen 23-24

Consider these 6 formations are just a simplification of in-game tactical structure. During the game, these formations can change due to several reasons and beneath any of these formations there are lots of playing styles considered.

5.2. Confounders Selected

The aim of this first approach is to focus on the intrinsic relationship between the rival and own formation regarding different target variables such as the goals difference. To do so, some of the variables selected for the first DML model just define each team strength at the moment of the fixture and do not explain their match statistics nature, which might correlate with the different formations. Thus, mediator variables such as the accumulated goals or possession of each team during the season are excluded from the model.

List of variables selected:

- The season and league of the fixture.
- The day of the week.
- The side of the main team (home or away).
- A rate of accumulated points for both teams before that match.
- Champions League (CL) flag, which defines whether each team is playing in CL or not.
- Teams ranking in the league.

- Weather data.

5.3. Model Results

5 different models have been computed each one testing on a different target variable. In the first step, the metric maximized to find the hyperparameters for the XGBoost Regressor is the negative MSE (-MSE):

MSE is a widely used metric in regression problems to quantify the average squared difference between predicted values ($\hat{Y}_i$) and actual true values (Y_i).

The formula for MSE is:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Where:

- n is the number of observations.
- Y_i is the actual value for the i-th observation.
- $\hat{Y}_i$ is the predicted value for the i-th observation.

The GridSearchCV (see (11)) aims to maximize the -MSE, and a value of 0 would correspond to a perfect fit. The final hyperparameters obtained from the model are the ones which maximize this metric.

In the GridSearchCV function, the upper limit for the hyperparameter *max_depth* is set to 5. This measure limits the depth of the tree, limiting the number of nodes and splits. Even though the learning capacity of each tree is restricted, the model would be less prone to overfitting and consequently would decrease the variance of the model (impeding large approximations on β estimators), which would help keep the symmetry in the resulting estimators.

To evaluate the prediction goodness of the 1st step, the R^2 is calculated. The R^2 estimator, quantifies the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model. It ranges from 0 to 1, with higher values indicating a better fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Where $\bar{Y}$ corresponds to target variable mean.

Statistic	Goals Model	Red Cards	Yellow Cards	Possession	Corners
MSE	2.70	0.19	3.18	343.35	7.98
R^2	0.118	0.003	0.030	0.298	0.781

Notice that each MSE value changes because the units differ (Possession is measured in % and red cards are normally small values between 0 and 1). The most accurate XGBoost model is undoubtedly the one which predicts the corners.

However, since the goal is to find the $\hat{\beta}_{i,j}$ estimations, there is not an interest on maximizing this metric by making small changes in the hyperparameters. When it comes to football predictions, the metrics like R^2 are normally too low as football is a highly unpredictable sport.

The 3rd model, a GLM, intends to minimize the MSE (the same as maximizing the -MSE).

5.4. Formation Matrix

This DML approach has enabled to computed several analysis regarding the combination of formations. Different target variables have been tested. To analyze the matrix of estimators, it is important to remember some properties stated before, such as:

- The diagonal of the matrix has to be 0.
- The matrix is symmetrical (Any small difference corresponds to XGBoost Regressor approximations).

Thus, $\hat{\beta}_{i,j} = -\hat{\beta}_{j,i}$

Consider the same pvalue significance indicators from 4.1. Therefore, those cells with a p-value lower than 0.05 will be considered statistically significant as the null hypothesis of their $\hat{\beta}_{i,j}$ estimator being 0 is rejected.

Since the following matrices just consider the pure effect between formations regarding a specific target variable, it is also useful to contextualize this estimations when playing as local or away team. To do so, the target variable effect on average is calculated and then added to the estimation, which jointly provides a close approximation of the true side effect.

Hence,

$$\hat{\beta}_{side_{i,j}} = \hat{\beta}_{i,j} + E(\hat{Y}_{home})$$

Where $E(\hat{Y}_{home})$ corresponds to target variable on average between the home and away team (e.g. goals difference).

5.4.1. Target variable 1: Goals difference

The resulting $\hat{\beta}_{i,j}$ estimators for goals difference as a target variable are:

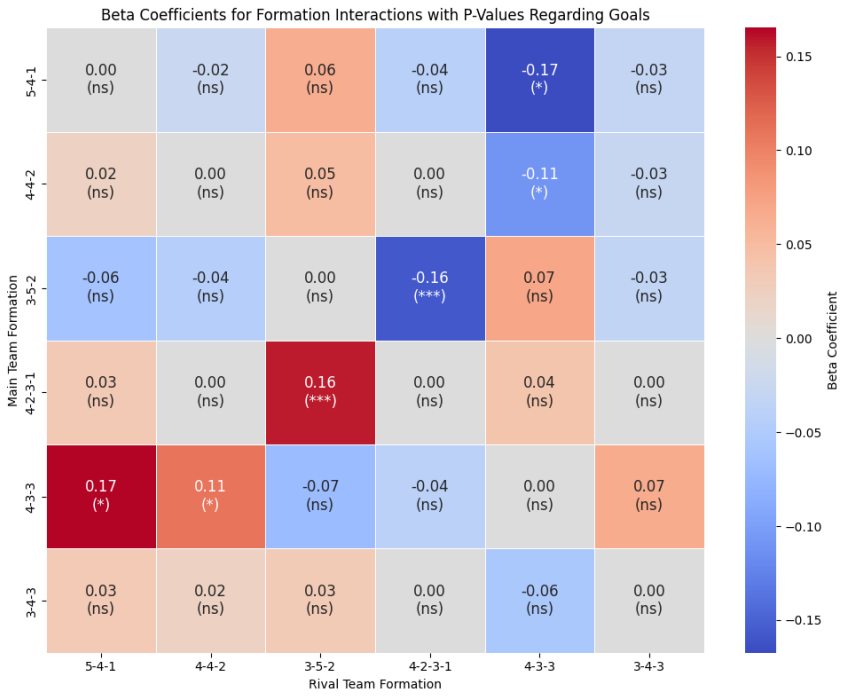


Figure 2. $\hat{\beta}_{i,j}$ coefficients and pvalues for goals difference

Notice that there are just 3 significant $\hat{\beta}_{i,j} \neq 0$ since their pvalues are <0.05. This estimators are:

- 4-2-3-1 against 3-5-2: Its advantage corresponds of 0.16 goals (1 advantage goal in every 6 matches) with a $pvalue < 0.001$, suggesting strong significance.
- 4-3-3 against 5-4-1: A 0.17 goals advantage with a $pvalue < 0.05$.
- 4-3-3 against 4-4-2: A 0.11 goals advantage with a $pvalue < 0.05$.

However, the rest of estimators are not significant and can be considered 0. It is widely known that playing at home increases the winning probabilities in comparison to playing away due to several motivational or environmental variables. Thus, adding the goals estimation when playing at home would adjust the final estimation. $E(\hat{Y}_{home})$ **= 0.285** when Y = goals difference. Therefore, the advantage of a 4-2-3-1 formation playing at home against a 3-5-2 would be around a **+0.445** in terms of goals.

At this point, it seems that formations do not have a great impact in the number of goals scored, since most of estimators are 0. Nonetheless, understanding the variables which truly affect the number of goals scored is crucial, so a complementary analysis regarding these impactful variables is carried out further on 6.

5.4.2. Target variable 2: Red Cards difference

The difference of the red cards in a match between main and rival team has also been computed. In this approach, the data has not been splitted by home or away team, so the estimators purely define the influence of each formation combination in the absolute red cards difference (considering both the second yellow card and the direct red card).

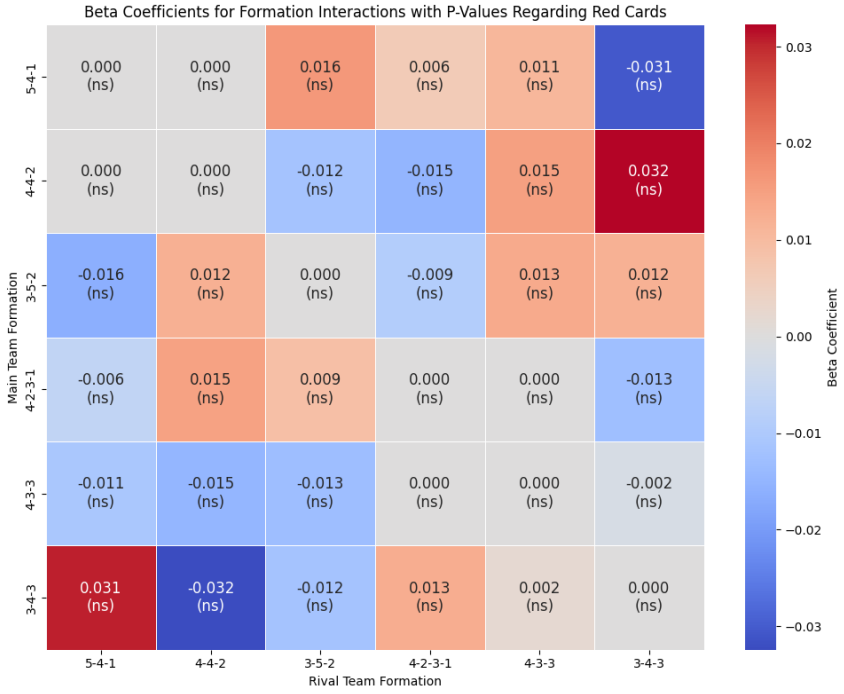


Figure 3. $\hat{\beta}_{i,j}$ coefficients and pvalues for red cards difference

It is completely perceivable that there is not a single statistically significant estimator regarding the red cards difference since all the pvalues are considered as *ns* (pvalue > 0.05). Therefore, $\hat{\beta}_{i,j} = 0 \, \forall i, j \in \{1, 2, \cdots, k\}$.

The main cause of not finding any significant estimator might that red cards are not influenced by team strength variables, and even less by general fixture variables or weather ones. Red cards can be influenced by some random and unpredictable match events. Additionally, both red and yellow cards, as stated *Llorenç Badiella Busquets* in (1) , are linked to the players' intensity, which is not trivial to quantify.

5.4.3. Target variable 3: Yellow Cards difference

The yellow cards difference in a match has also been analyzed, disregarding the split between home or away team.

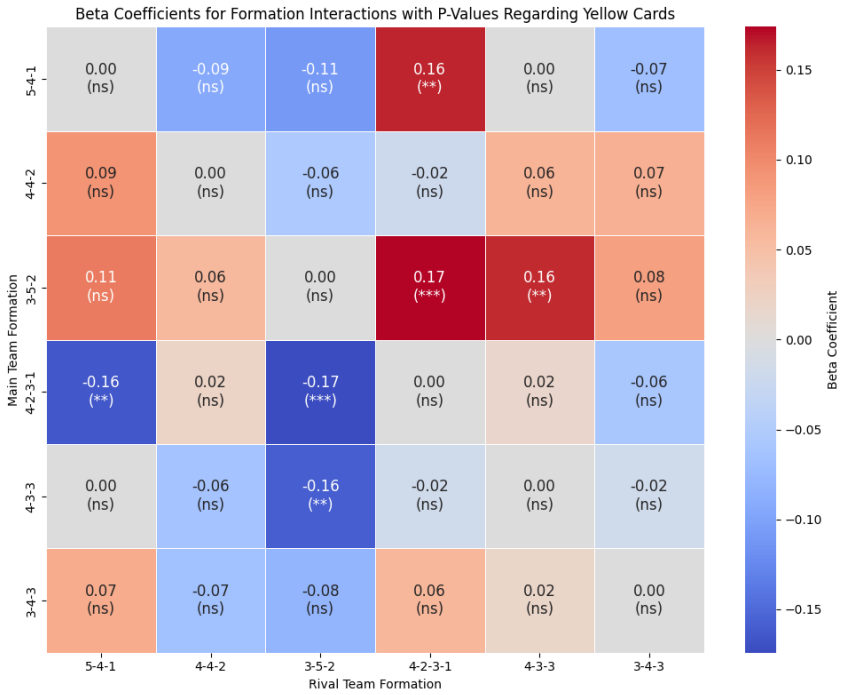


Figure 4. $\hat{\beta}_{i,j}$ coefficients and pvalues for yellow cards difference

In this case, it is observable that there are 3 combinations with a significant p-value (notice that even though there appear to be 6, the matrix is symmetrical and, in fact, there are 3 pairs of opposite coefficients).

- Firstly, the relationship between using a 5-4-1 against a 4-2-3-1 formation provides an estimator of +0.16 on the yellow cards difference, suggesting that the 5-4-1 team would be penalized on average with 0.16 yellow cards (approximately 1 yellow card extra per every 6 games).Its pvalue is lower than 0.01.
- Secondly, the combination 3-5-2 against the 4-2-3-1 also suggest a higher average on yellow cards for the 3-5-2 team (+ 0.17), with a pvalue < 0.001.
- Thirdly, the relationship between 3-5-2 and 4-3-3 provides a higher average on yellow cards of 0.16 for the 3-5-2 team, with also a pvalue < 0.01.

It is perceptible that defensive formations have a subtle increase on yellow cards comparing to the most offensive ones.

Nonetheless, there are clearly other variables which might have a higher impact on these different target variables.

5.4.4. Target variable 4: Possession

The same analysis for the possession is done, since it is hypothesized that each structure of players might influence the ball control over a match.

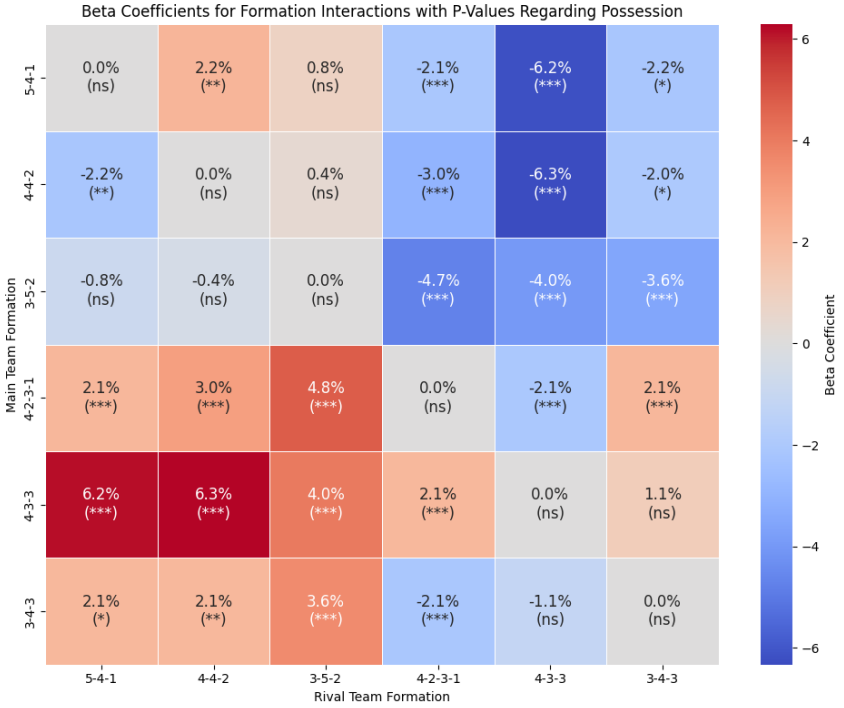


Figure 5. $\hat{\beta}_{i,j}$ coefficients and pvalues for possession difference

It is clearly visible that this matrix provides valuable insights of the formations pure behaviour regarding possession. There are 13/18 formation combinations with significant P-values, most of them < 0.001.

There are two differenced coloured regions, the top-right blue one, corresponding to the the 3 most defensive formation against the 3 most offensive, and the bottom-left one, corresponding to the most offensive formations against the defensive ones. The other two remaining 3x3 cells contain mostly neutral and non-significant values.

- Top-right blue region: It is highly perceivable that the defensive formations tend to have a less control over the possession during a football match. There are obvious differences in the estimated possession between defensive formations and offensive ones like 4-3-3, all three of them with a $pvalue < 0.001$, proving evident significance.
- Bottom-left red region. This region is completely the opposite as the blue one, since the amtrix is symmetric. Thus, it is proved that offensive formations have a greater control on the ball comparing to defensive ones, disregarding specific team profiles since the analysis is purely on the formations toward the target variable.

Generally, it seems that the formation 4-3-3 is the one with a better ball control since it has a positive estimator when comparing it with other offensive ones. Additionally, 3-4-3 formation, despite being the most offensive one it showcases a lower ball dominance in comparison to 4-2-3-1, possibly caused by some comings and goings on the ball transition and the impossibility of keeping calmly the ball.

5.4.5. Target variable 5: Corners

Finally, the influence towards corners has also been scrutinized.

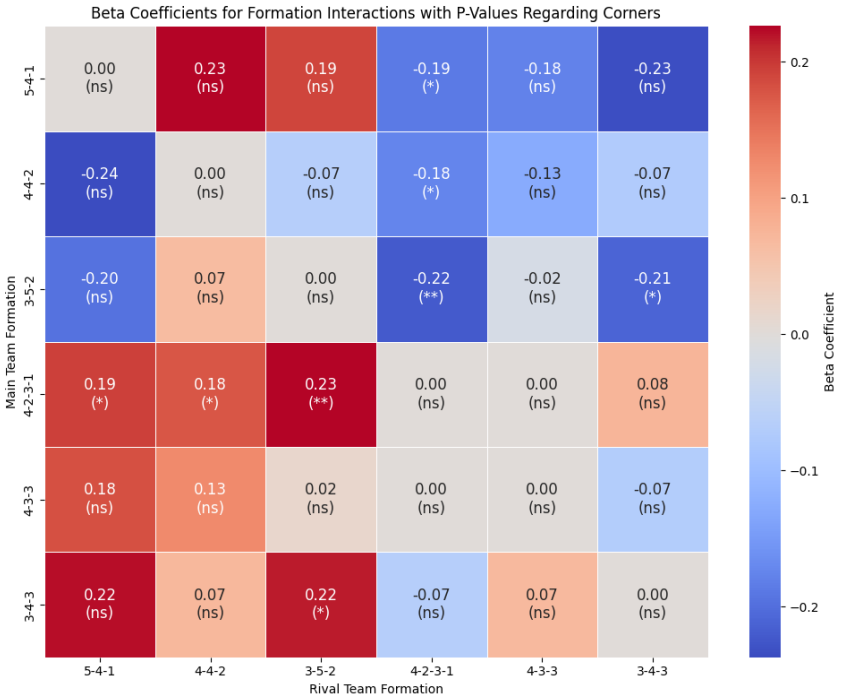


Figure 6. $\hat{\beta}_{i,j}$ coefficients and pvalues for corners difference

Notice that those red-coloured cells, the ones from the bottom-left matrix, are the ones corresponding to use an offensive formation against a defensive one. The corners estimators are aligned with the possession ones, in light of the fact that most offensive formations and with a higher ball possession have more time and opportunities to create scoring chances, which influences in the corner occasions. For instance, 4-2-3-1, the most common formation in the dataset, has an estimator of 0.19 over the 5-4-1, of 0.18 over the 4-4-2 and of 0.23 over the 3-5-2. Albeit these estimators are significant, their effects are not meaningful enough, seeing the number of corners on average in the dataset is aorund 5 corners per match per each team.

6. COMPLEMENTARY FEATURE ANALYSIS

Given the formation combination does not provide a useful insight on the match outcome, there is interest in understanding which features related to match statistics might be relevant. To do so, a Machine Learning model has been computed. This model, which keeps the goals difference as the target variable, aims to explore the feature importance of some statistical variables such as the corners or possession difference between both teams.

6.1. ML Model

The model consists of an XGBoost Regressor, the same model for the previous DML analysis. It keeps the same variables in the X matrix but also adding the *side* component, since there is also interest on knowing its potential effect. However, the formations of both teams have been removed so the analysis focuses on those variables which have a determining effect on the target variable (goals difference in this case).

6.2. Model Interpretability via SHAP Values

To quantify the influence of individual input variables on the predicted goal difference, we employ SHAP (SHapley Additive exPlanations) values (10), a model-agnostic explanation technique grounded in cooperative game theory. SHAP values decompose a prediction into additive contributions from each feature, enabling both global and local interpretability.

6.2.1. Mathematical Formulation

For a given prediction function $f : \mathbb{R}^p \to \mathbb{R}$, and an input vector $x \in \mathbb{R}^p$, the SHAP framework expresses the model output as:

$$\nu(x) = \phi_0 + \sum_{j=1}^{p} \phi_j(x) \tag{1}$$

- $\nu(x)$: total value predicted, attributed to the full set of features.
- $\phi_0 = \mathbb{E}[\nu(x)]$ is the **base value**, i.e., when no features are included, $\nu(\emptyset)$.
- $\phi_j(x) \in \mathbb{R}$ is the **SHAP value** corresponding to feature j, indicating how much that feature contributes to the deviation from the base value for the given instance.
- p: total number of features.

SHAP values for each feature are computed as the average marginal contribution of that feature over all possible feature subsets:

$$\phi_j(x) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!(p - |S| - 1)!}{p!} \left[\nu(S \cup \{j\}) - \nu(S)\right] \tag{2}$$

- S: any subset of features that does not include j.
- $|S|$: number of elements in the subset S.
- $\nu(S)$: value (e.g., R^2 function) of the model when using only the features in S.
- $\nu(S \cup \{j\})$: value of the model when feature j is added to subset S.

6.3. Model's feature importance

The variables belonging to the X matrix have been grouped by nature and color-coded for clarity.

- Yellow: Accumulated statistics difference between the main and rival team (e.g. possesion, corners or yellow or red cards difference).
- Blue: Accumulated points ratio reached of each team up to a certain round as Local or Away accross the whole season.
- Red: Rankings of each team up to a certain round in each season.
- Purple: Binary variable which determines if each team is still playing Champions League during the fixture period.
- Green: Weather variables (e.g. temperature or humidity).
- Grey: Other general variables. (e.g. side, round or season).

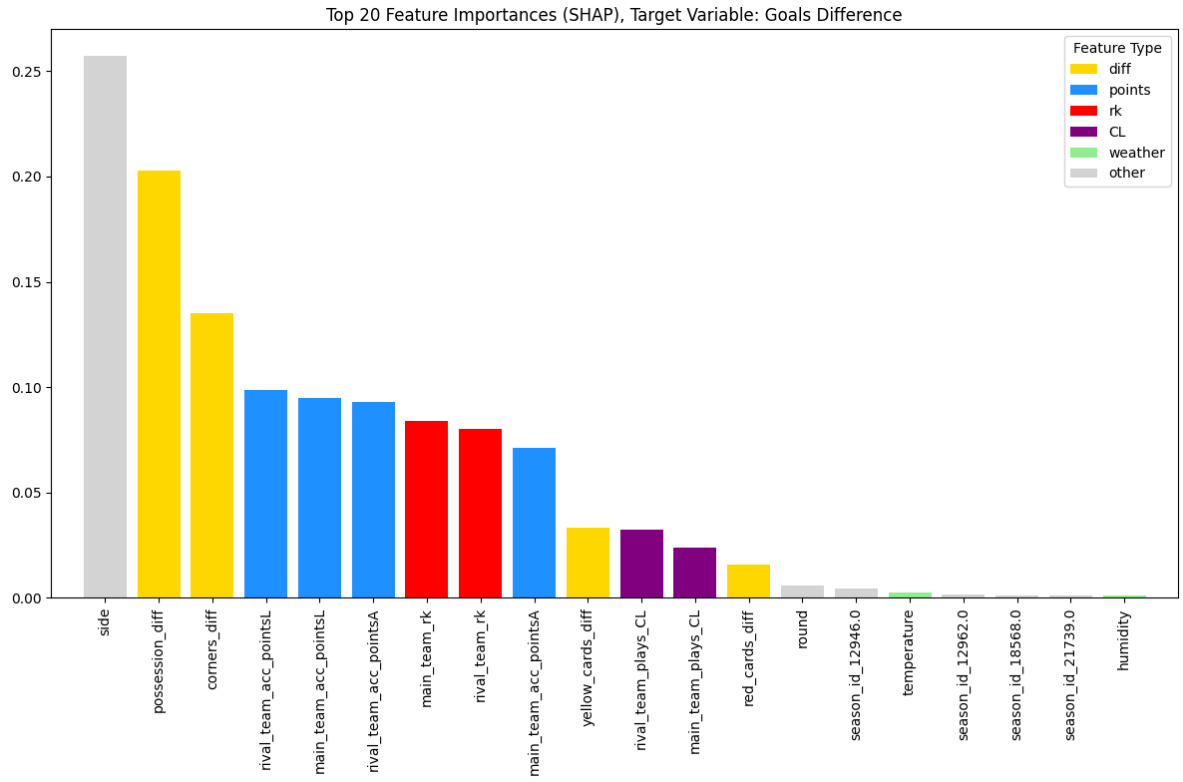


Figure 7. Top 20 SHAP values means

Each displayed SHAP values correspond to the mean of each ϕ_j in absolute value. We can notice that *side* has the highest SHAP value, corresponding to 0.26. That means that the prediction of the model differs on average ±0.26 depending whether the analyzed team is playing at home or not. Moreover, the difference in possession and corners have a value of 0.20 and 0.14 respectively. Summing the rest of yellow SHAP values, the total is a ±40 value on average just considering accumulated statistical differences between both teams. The final prediction for a single fixture row will depend on the sum of the different SHAP values. Hence,

$$\hat{Y} = \phi_0 + \phi_{side} + \phi_{possession} + \phi_{corners} + \cdots \tag{3}$$

Notice that the yellow-coloured variables were not considered in the DML models, since they are correlated with the formation. Notwithstanding, although these statistical variables seem to influence the goal difference prediction the most, the blue and red ones—related to teams' points and rankings—also contribute heavily to the final estimation and were indeed considered in the DML models.

7. CONCLUSION

This study reveals that while football formations provide a baseline tactical framework, they are not sufficient to define or predict a team's playing style or match outcome. Each nominal formation—such as 4-3-3 or 4-2-3-1—can embody a variety of tactical behaviors, and the assumption that the starting formation remains unchanged throughout the match overlooks the dynamic nature of in-game adaptations. Moreover, some teams apply a different formation when attacking in comparison to the one defending.

Through the application of Double Machine Learning—redefined in this project to deal with the categorical treatment through a novel matrix-based residualization of formation combinations—it becomes evident that

defensive formations, popularly known as *parking the bus*, generally serve to neutralize offensive threats rather than produce statistically advantageous outcomes in terms of goals, possession, corners, or disciplinary measures. However, *parking the bus* often implies other out-of-the-game strategies such as time wasting, committing faults, and slowing down the fixture pace, which have not been considered.

Conversely, offensive formations like 4-2-3-1, 4-3-3, and 3-4-3 show a modest but statistically significant positive impact on possession, corners, and yellow card differences when facing more defensive systems. However, formation choice alone has minimal impact on red cards, which appear to be driven by more random or behavioral factors. Importantly, the effect of formations on goal difference is limited, with only three formation combinations yielding significant estimators—underscoring that formations do not fully account for scoring dynamics. Complementary analysis using SHAP values confirms that contextual variables such as home advantage, possession, and corners exert a greater influence on match outcomes than formations themselves. Therefore, there is no evidence that *parking the bus* offers any scoring or winning advantage.

While the results provide insightful patterns, they come with limitations, including the team strength criteria and the simplification involved in formation grouping, which may not fully explain playing behavior. Nonetheless, this serves as a valuable starting point for deeper formation-based performance analysis in football.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Juan Jesús Rodríguez and Juan Camilo Vázquez for their invaluable support throughout this project. Their deep football knowledge, especially in helping group formations meaningfully and providing tactical insights, was essential to interpreting the data accurately and grounding the analysis in real-world football logic.

REFERENCES

Badiella, L., Puig, P., Lago-Peñas, C., & Casals, M. (2023). Influence of red and yellow cards on team performance in elite soccer. *Annals of Operations Research, 325*, 149–165. https://doi.org/10.1007/s10479-022-04733-0

Bhattacharya, S. (2023). *Understanding the Dummy Variable Trap in Regression Models*. Retrieved from https://www.researchgate.net/publication/375756030_Understanding_the_Dummy_Variable_Trap_in_Regression_Models

Bonhomme, S., Jochmans, K., & Weidner, M. (2024). A Neyman-orthogonalization approach to the incidental parameter problem. arXiv preprint arXiv:2412.10304v1. Retrieved from https://arxiv.org/abs/2412.10304

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. arXiv preprint arXiv:1603.02754. Retrieved from https://arxiv.org/abs/1603.02754

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal, 21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

Elsevier. (n.d.). *Mean Square Error*. ScienceDirect Topics. Retrieved from https://www.sciencedirect.com/topics/engineering/mean-square-error

Frisch, R., & Waugh, F. V. (2008). *The Frisch-Waugh theorem and generalized least squares*. Retrieved from https://www.researchgate.net/publication/4879513_The_Frisch-Waugh_theorem_and_generalized_least_squares

GiveMeSport. (2024, November 2024). What "park the bus" means in football. Retrieved from https://www.givemesport.com/football-park-the-bus-meaning/

Lineup11.net. (n.d.). Retrieved from https://www.lineup11.net/es [Accessed: April 17, 2025].

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. arXiv:1705.07874. https://doi.org/10.48550/arXiv.1705.07874

scikit-learn developers. (n.d.). *sklearn.model_selection.GridSearchCV*. Retrieved June 6, 2025, from https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Sportmonks. (n.d.). *Football and Cricket data API*. Retrieved during 2025, from https://www.sportmonks.com/

UCLA Statistical Consulting Group. (n.d.). *FAQ: What is effect coding?* Retrieved from https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faqwhat-is-effect-coding/

UEFA. (2025). *Country coefficients*. Retrieved in April 2025, from https://www.uefa.com/nationalassociations/uefarankings/country/?year=2025

APPENDIX

See below the list of Leagues and Seasons considered in the project. The discarded seasons did not contain full information about match results or statistics. For instance, some team strength variables such as the rate of points or other match-related target variables such as possession or yellow cards difference could not be calculated and consequently impeded a proper analysis.

League	18-19	19-20	20-21	21-22	22-23	23-24	24-25
Premier League	✓	✓	✓	✓	✓	✓	✓
La Liga	✓	✓	✓	✓	✓	✓	✓
Serie A	✓	✓	✓	✓	✓	✓	✓
Bundesliga	✓	✓	✓	✓	✓	✓	✓
Ligue 1	✓	×	✓	✓	✓	✓	✓
Primeira Liga	✓	✓	✓	✓	✓	✓	✓
Eredivisie	✓	×	✓	✓	✓	✓	✓
Pro League (Belgium)	✓	✓	✓	✓	✓	✓	✓
Turkish Süper Lig	✓	✓	×	✓	×	✓	×
Ekstraklasa (Poland)	×	×	✓	✓	✓	✓	✓
Championship (2nd UK)	×	✓	✓	✓	✓	✓	✓
La Liga 2	×	✓	✓	✓	✓	✓	✓
Serie B	×	×	×	✓	✓	✓	✓

7.1. Formation grouping

```
def map_formations(row_column):
    # 4-3-3 group
    if row_column in ["4-1-2-3", "4-2-1-3", "4-3-2-1", "4-5-1", "4-3-1-2"]:
        return "4-3-3"
    # 4-4-2 group
    elif row_column in ["4-1-2-1-2", "4-2-2-2", "4-1-3-2", "4-4-1-1", "4-2-4"]:
        return "4-4-2"
    # 5-4-1 group
    elif row_column in ["5-3-1-1", "3-3-3-1", "3-2-4-1", "4-1-4-1", "5-2-3"]:
        return "5-4-1"
    # 3-5-2 group
    elif row_column in ["3-2-3-2", "3-3-2-2", "3-1-4-2", "3-5-1-1", "5-3-2","3-4-1-2"]:
        return "3-5-2"
    # 3-4-3 group
    elif row_column in ["3-3-1-3", "3-4-2-1"]:
        return "3-4-3"
    elif(row_column=="4-2-1-2-1"):
        return "4-2-3-1"
    # leave as is for all other formations
    else:
        return row_column
```

Is it worth parking the bus?: Formations impact in football matches with Double Machine Learning

Is it worth parking the bus?

G. Ruiz-Menárguez [1]
genis.ruiz@autonoma.cat
L. Badiella [1,2]
llorenc.badiella@uab.cat
[1] *Department de Matemàtiques, Universitat Autònoma de Barcelona*
[2] *Servei de Consultoria Matemàtica, Universitat Autònoma de Barcelona*

Formations impact in football

Different events during a football match can influence the final outcome (1), but the impact of team formations on match results is still an open question. The aim of the current study is to analyze the effects of team formation on match outcome in elite soccer. The sample is composed of 22114 matches from the top seven European leagues including first and second divisions of the last 4 seasons, from 2021/22 to 2024/25.

Methods

Data processing

In order to access to the match data, Sportmonks(2) API (https://www.sportmonks.com/) has been used. This API contains multitude of match features involving a plenty of teams from different countries and divisions.

The seven best European male football leagues, according to UEFA(3) have been included in the present study: the first division from England, Italy, Spain, Germany, France, the Netherlands and Portugal.

The following variables are considered:

- Match Data: Home and away team, date, league, season and round of the fixture, weather conditions
- Team data: Team strength playing at home or away, champions league indicator
- Formations and player's data: Formation system of each team (e.g. 4-3-3), Lineups with player's information such as age, height, weight, position.
- Outcome: Result, number of goals, corners, possessions and red and yellow cards of both teams.

Formations have been classified into 6 categories:

5-4-1, RCD Mallorca 23-24 | 4-4-2, Leipzig FC 23-24 | 3-5-2, FC Inter. Milano 23-24

4-2-3-1, Athletic Club 23-24 | 4-3-3, FC Barcelona 23-24 | 3-4-3, Bayern Leverkusen 23-24

Statistical modelling

Since the relationship between team formations (D)and outcome (Y) is prone to confounding, we used a Double Machine Learning (DML) approach(4). This method allows to control for the effect of confounders (X) while estimating the causal effect of team formations on match outcomes.

DML methods typically consider continuous treatments, but in the present case, we are interested in a categorical treatment (i.e., combinations of team formations). To adapt the DML approach, we used the following steps:

- **First stage**: Model $Y = f(X)$ using a XGBoost model (5) and obtain the model residuals (r_Y).
- **Second stage**: Define multiple dummy indicators for each formation combination (D_i).
- **Third stage**: Model each dummy variable as a response variable considering the nuisance parameters using a XGBoost model $D_i = g(X)$ and obtain again the model residuals ($r_{D,i}$).
- **Final stage**: Regress r_Y with respect all dummy variable residuals $r_{D,i}$ to obtain the causal effect of each formation combination.

Results

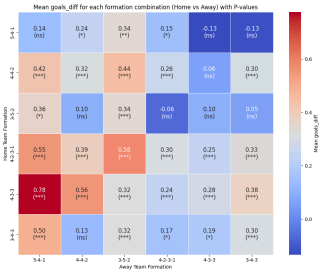
Descriptive summary

A preliminary analysis shows that some strategies may have a potential impact on match results:

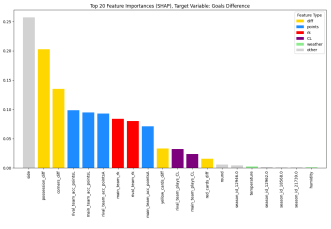
Table 1: Summary statistics for different team formations

	5-4-1	4-4-2	3-5-2	4-2-3-1	4-3-3	3-4-3
Goals	1.24	1.25	1.27	1.39	1.44	1.34
Red Cards	0.110	0.114	0.104	0.106	0.103	0.098
Yellow Cards	2.17	2.25	2.22	2.11	2.08	2.19
Possess.(%)	49.0	48.6	48.2	50.7	51.6	49.8
Corners	4.75	4.72	4.75	5.00	5.11	4.85

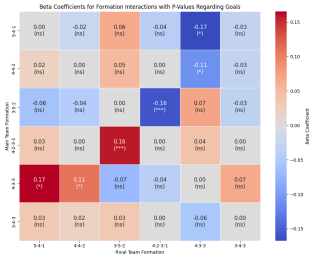
Unadjusted impact on Goal Difference of different team lineups

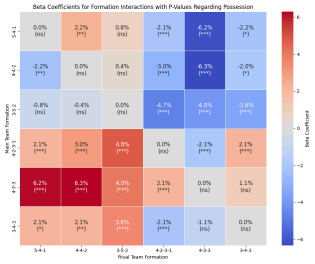


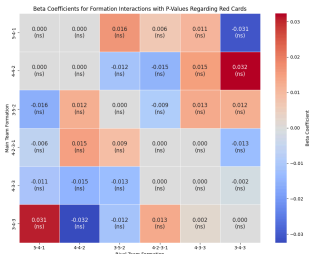
Confounding between team characteristics and goal difference



Adjusted impact on different outcomes







Conclusions

No evidences that parking the bus is a good strategy to score more goals or win matches.

References

1. Badiella, L., Puig, P., Lago-Peñas, C., & Casals, M. (2023). Influence of Red and Yellow cards on team performance in elite soccer. Annals of Operations Research, 325(1), 149-165.
2. Sportmonks. https://www.sportmonks.com/
3. UEFA. https://www.uefa.com/
4. Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1), C1–C68.
5. XGBoost: A Scalable Tree Boosting System. Chen, T., Guestrin, C. (2016). *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

UAB Universitat Autònoma de Barcelona

mathematical consulting service

Figure 8. Project Poster

7.2. Github Repository

The complete codebase, including data extraction, feature engineering, and Double Machine Learning (DML) model implementations, is publicly available at the following GitHub repository:

https://github.com/genissrruiz/Formation-Analysis-DML

A comprehensive explanation of the codebase, including the API extraction, feature engineering, and Double Machine Learning (DML) model implementations, is provided within the `README.md` file of the repository.