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*
[1] *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:

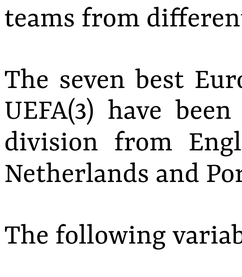
- Match Data: Home and away team, date, league, season and round of the fixture, weather conditions
- Team data: Team strength playing at home or away, champions league indicator
- Formations and player's data: Formation system of each team (e.g. 4-3-3), Lineups with player's information such as age, 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:

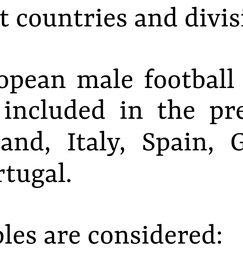
5-4-1, RCD Mallorca 23-24

4-4-2, Leipzig FC 23-24

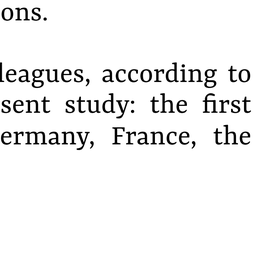
3-5-2, FC Inter. Milano 23-24



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

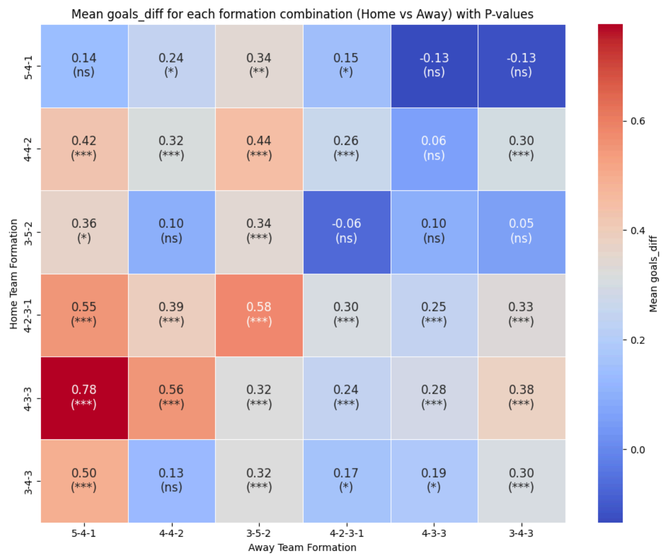
Descriptive summary

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

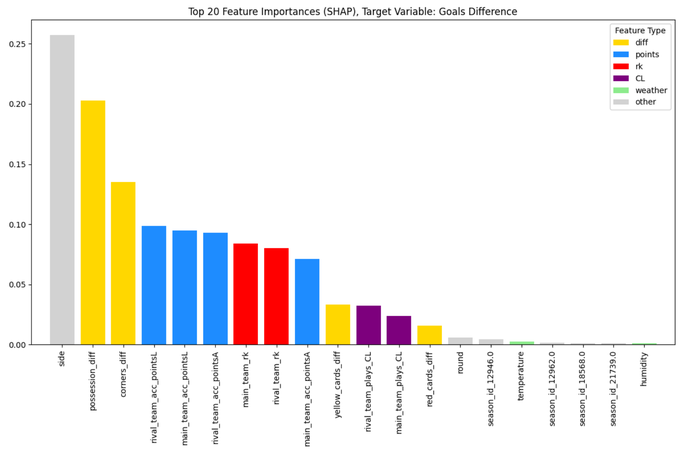
Table 1: Summary statistics for different team formations

	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

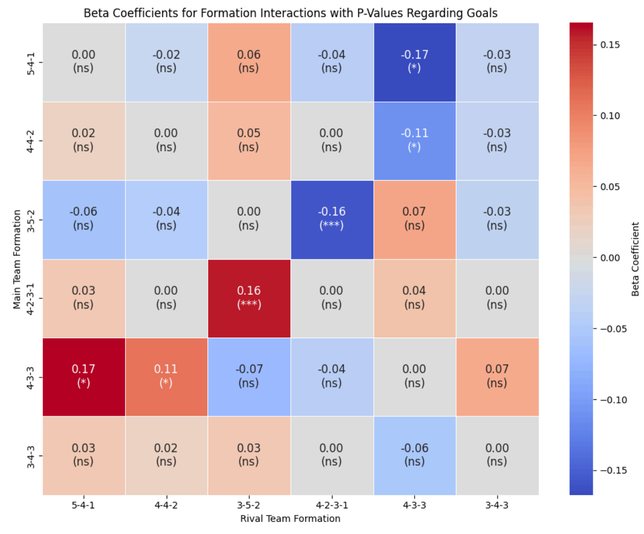
Unadjusted impact on Goal Difference of different team lineups

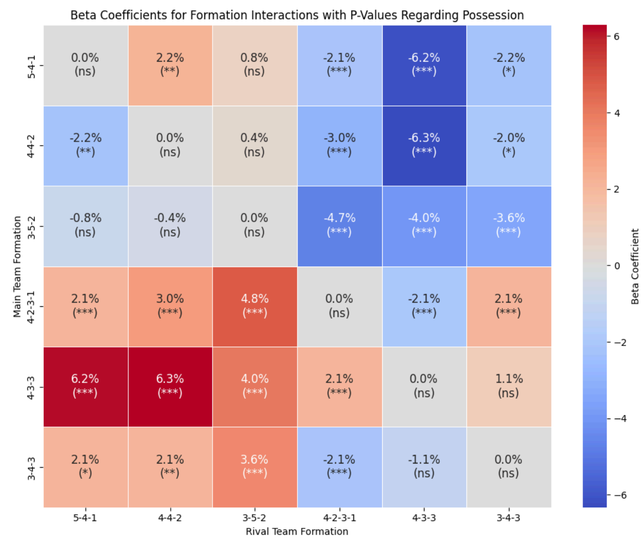


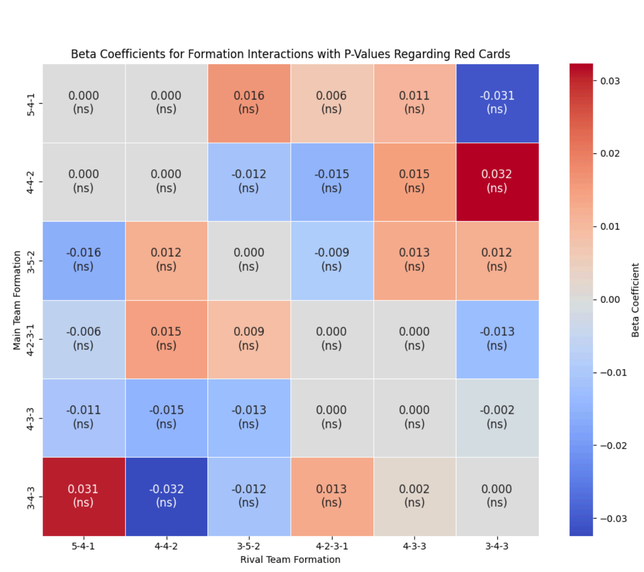
Confounding between team characteristics and goal difference



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