Microsimulation of Team Formation for Football Matches

Mahdi Munshi
Student Number: 018059726

The project can be found at this Github repository:
https://github.com/mahdimunsi/Football-Formation-Simulator

Introduction

Professional football is a highly competitive and dynamic sport, with teams employing various tactics and strategies to gain an advantage over their opponents. One key aspect of this competition is the choice of team formation, which can significantly influence the outcome of a match. In this project, we aim to develop a predictive model that can estimate the likelihood of a home win, away win, or draw based on the formations used by the teams and the competition they are playing in.

Background

The choice of team formation is a crucial tactical decision in professional football, as it can impact factors such as ball possession, defensive solidity, and attacking potency. Different formations can be more effective in different situations, and teams often adapt their formations to suit their opponents and the specific match circumstances. However, there is limited research on the relationship between team formations and match outcomes, and no existing models that can accurately predict the likelihood of a match outcome based on the formations used by the teams.

Methodology

Data

The data for this project was collected from Kaggle that has a comprehensive and up-to-date compilation of football data, including detailed match formations. The "games.csv" file was the only data frame (table) required for this project, and some data wrangling technique was used to get the formation data for the to 5 European football leagues: LaLiga, Ligue 1, Serie A, Premier League, and Bundesliga. The dataset included variables such as the home and away team names, the home and away team formations, the final score, and the match winner.

Method

To develop a predictive model that could estimate the probability of each match outcome (home win, away win, or draw) based on the home and away team formations, the model utilized multinomial logistic regression model which is well-suited for predicting categorical outcomes with more than two classes (in this case, the three possible match results). The model was trained on the "winner" variable, using the "home_club_formation" and "away_club_formation" variables as predictors.

The multinomial logistic regression model is a powerful tool for this type of analysis, as it allows to estimate the probability of each outcome (home win, away win, or draw) given the specific formation combination. This information can be invaluable for coaches, analysts, and bettors, as it can inform strategic decision-making and provide a more data-driven approach to predicting match results.

Workflow

For the initial attempt, we simulate the likely outcome of a match between any two teams, based solely on their chosen formations. As an example, we simulated a match between two teams both playing in the "4-3-3" formation. The simulation showed that the predicted probability of a home win was approximately 45%, the probability of an away win was around 32%, and the probability of a draw was roughly 23%. This type of simulation can provide valuable insights and inform strategic planning, as it allows stakeholders to anticipate the potential outcomes of a match based on the formations employed by the teams.

```
# Example simulation
simulate_match("4-3-3", "4-3-3") *100
```

Description: df [1 × 3]

away <dbl>	draw <dbl>	home <dbl>
31.91113	22.71999	45.36889

1 row

However, this is a very simple model, with no factor having any impact on the simulation. By focusing solely on team formations as the predictive variables, the model was overly simplified and overlooked other factors that may influence match outcomes, and have created a strong generalization across leagues.

However, for the next and final attempt, the model addressed the name of the competition as an important predictor, while keeping all the methodology exactly the same, resulting in a ‘better’ model.

```
simulate_match("3-1-4-2", "4-2-3-1", "LaLiga") *100
```

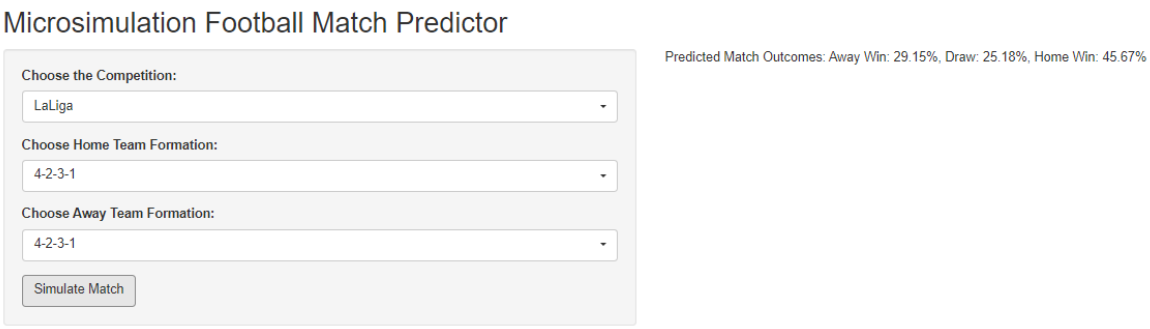
Description: df [1 × 3]

away <dbl>	draw <dbl>	home <dbl>
34.90035	24.39366	40.70599

1 row

We simulated a match between two teams playing in the "3-1-4-2" and "4-2-3-1" formations in LaLiga. The simulation showed that the predicted probability of a home win was approximately 41%, the probability of an away win was around 35%, and the probability of a draw was roughly 24%.

The final step of this project was to build a shiny app where one can input through dropdowns and do not need to run the function codes and type in the inputs manually.



Limitations and Future Directions

While this analysis provides valuable insights into the relationship between team formations, competitions, and match outcomes, there are several limitations to consider:

1. Simplification of the Model: By focusing solely on team formations and competitions as the predictive variables, we have simplified the model and overlooked other factors that may influence match outcomes, such as player quality, team morale, home advantage, and weather conditions. In reality, football is a complex game, and many variables can contribute to the final result.
2. Temporal Considerations: The dataset covers multiple seasons, but the analysis does not account for potential changes in team formations or tactical trends over time. The predictive power of the model may be influenced by the evolution of football tactics and strategies.

Moving forward, there are several avenues for further exploration and refinement of this analysis:

1. Incorporating Additional Variables: Expanding the model to include other relevant factors, such as player statistics, team performance metrics, and contextual information, could lead to more accurate and comprehensive predictions.
2. Investigating Temporal Dynamics: Incorporating time-series analysis techniques could shed light on the evolution of team formations and their impact on match outcomes over multiple seasons, providing a more dynamic and adaptive predictive model.

3. Validating the Model: Testing the model's performance on a holdout dataset or conducting cross-validation procedures would help assess its generalizability and robustness, ensuring the reliability of the predictions.

Conclusion

This project has demonstrated the potential of using team formations and competitions as key factors in predicting the outcomes of football matches. The insights gained from the visualization and the predictive model highlight the importance of tactical considerations in the game of football, and the value of data-driven approaches in supporting strategic decision-making.

As we continue to explore and refine this analysis, we aim to provide stakeholders, such as coaches, analysts, and fans, with a more robust and reliable tool for anticipating match results. By understanding the relationship between team formations, competitions, and winning probabilities, we can empower decision-makers to make more informed choices and enhance the overall competitiveness and excitement of the sport.

Appendix

Grading Matrix

Grading Factor	1	2	3	4	5
Report as a whole; layout and finishing	Report is uneasy to read and the text is not understandable. Layout and finishing are mostly incomplete.		Report is easy to read and the text is somewhat fluent. Layout and finishing are quite all right.		Report is very easy to read and the text is fluent. Layout and finishing are well completed.
Defining topic and making restrictions	The topic chosen is inadequately defined and/or the chosen topic needs to be narrowed down.		The topic chosen is carefully defined and properly restricted.		The topic chosen is particularly carefully defined and the chosen topic is particularly carefully delimited.
Using references and deepening the knowledge on the subject	There are no four references in the report. The background to the chosen topic is vague. The analysis of the topic is very brief.		The report has at least four references. The background to the chosen topic is based on previous knowledge (literature). The coverage of the topic is sufficient.		There are at least four sources in the report and they have been used to excellent effect throughout the report. The background to the chosen topic is based on previous knowledge (literature). The analysis of the topic is quite in-depth.

Simulation Accuracy	Simulation results are inaccurate; lacks complexity and detail.		Simulation is mostly accurate; represents the model reasonably well.		Simulation is highly accurate, very detailed, excellently representing the dynamics of football matches.
Simulation Complexity	Simulation lacks complexity and detail.		Simulation shows a moderate level of complexity		Simulation is highly complex.
Analysis and Interpretation of Results	Analysis is incorrect or irrelevant; fails to interpret results or draw conclusions.		Good analysis of results with clear and relevant interpretations; logical conclusions based on data.		Exceptional analysis and interpretation of results; provides deep insights of results and recommendations on the model.