Team Performance Analysis and Prediction in the 5 Big European Football Leagues (2023-2024 Season)

Introduction

The primary objective of this project is to analyze and predict team performance in the 2023-2024 season across the top 5 European football leagues: the Premier League (England), La Liga (Spain), Bundesliga (Germany), Serie A (Italy), and Ligue 1 (France). The dataset includes essential statistics such as goals for, goals against, goal difference, expected goals (xG), expected goals against (xGA), and points.

Part I: Visualizing the Data

Goals for and against teams are visualized in order to observe a trend if there is any.

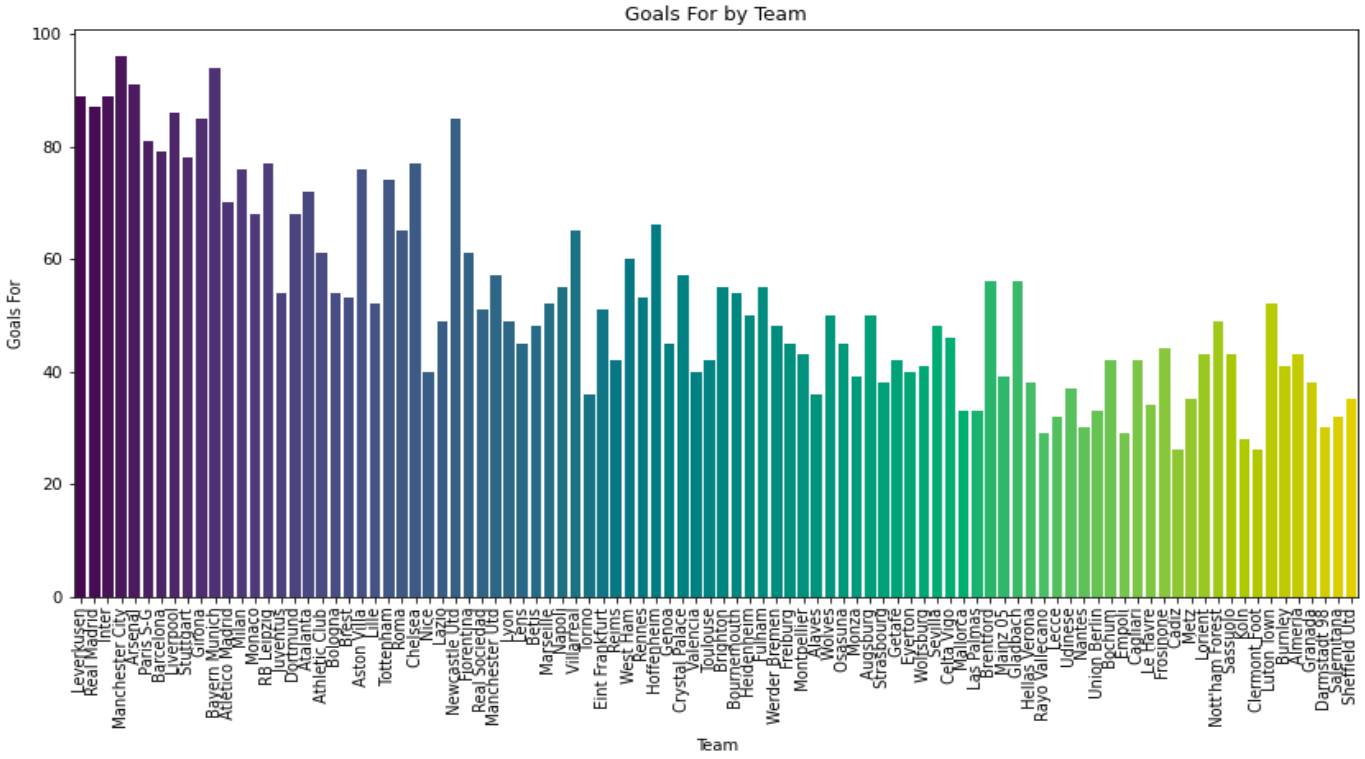

Table 1: Goals scored by team

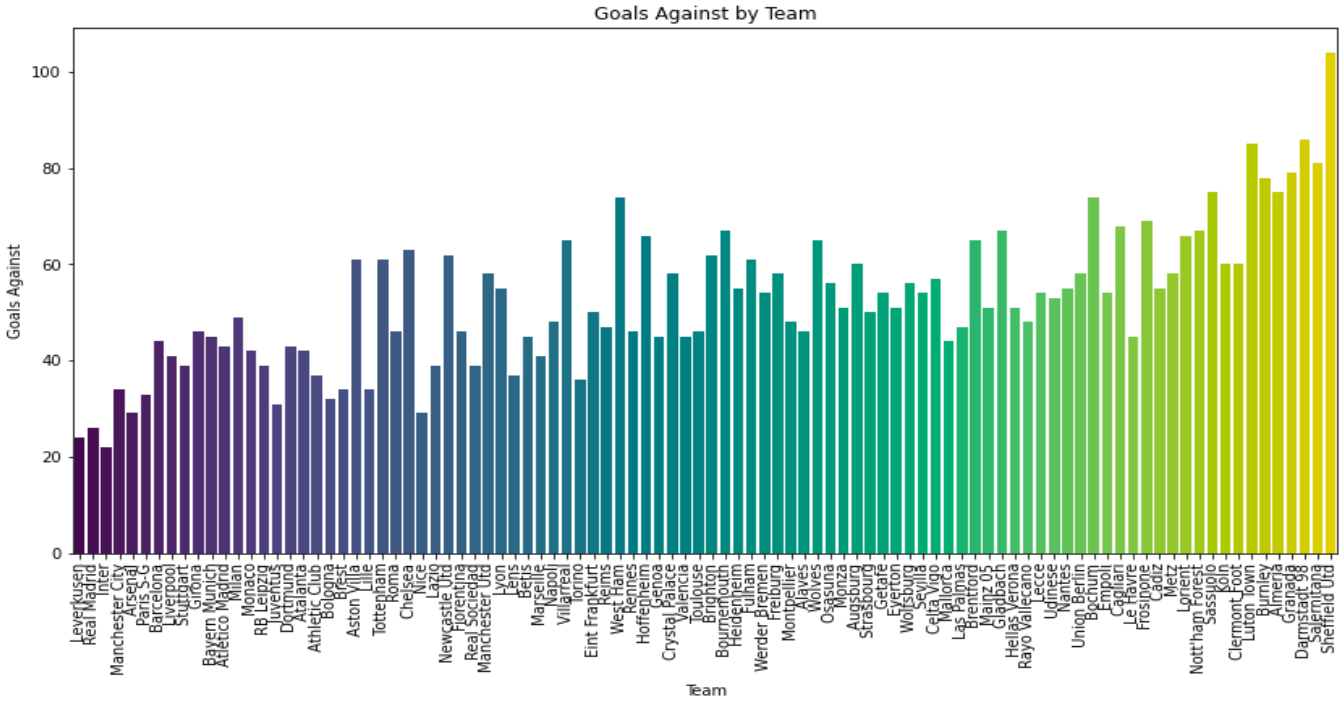

Table 2: Goals conceded by team

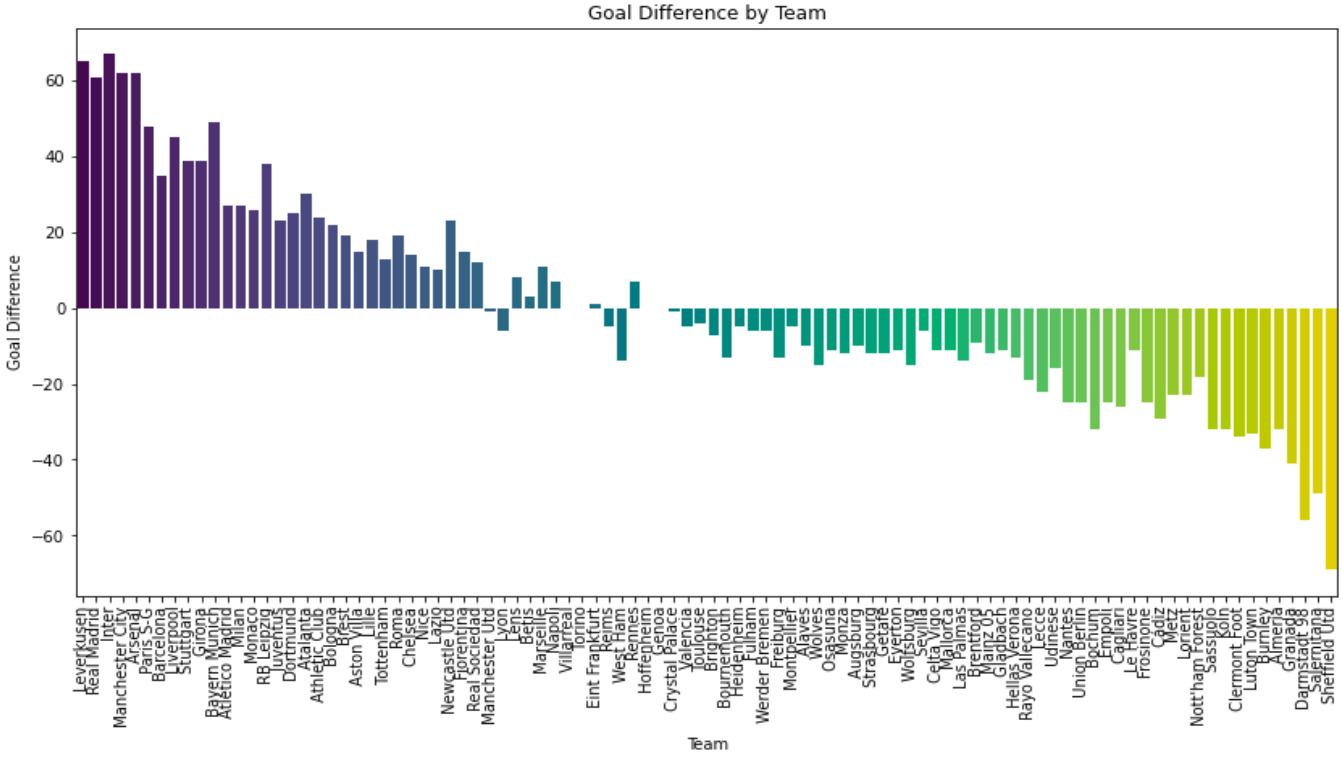

Table 3: Goal difference by team

As it can be seen from the charts above, there is a clear trend for goals scored and conceded by teams. Teams finished their respective leagues in a higher position scored more and conceded less as expected.

However, these charts are not adequate to claim the relationship between goals and league rankings. In order to observe the relationship clearly, a correlation analysis is needed. The correlation matrix can be found below:

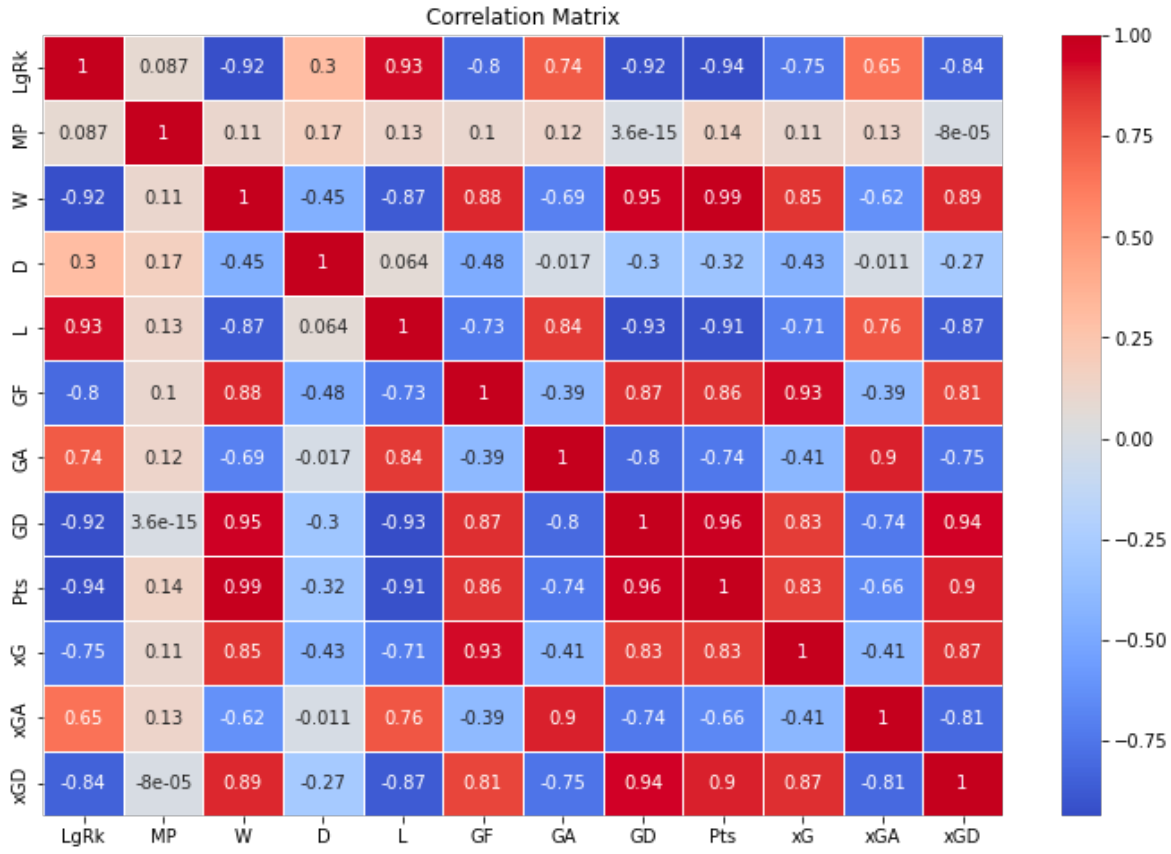

Table 4: Correlation Matrix

Now, it is fair to say that goals scored and conceded are highly correlated with league rankings. Still, correlation does not always indicate causation. In the next part, implementation of a regression model will be used to examine if goals scored and conceded have an impact on league rankings.

Part II: Fitting Regression Models on Dataset

Linear regression and XGBoost regressor models were implemented on the dataset. XGBoost regressor performed slightly better than linear regression. XGBoost regressor model has 2.0563 MSE, while linear regression has 3.6927 MSE. Because of these results, XGBoost regressor model is selected to explain the relationship between variables. Feature importance for the model can be seen below:

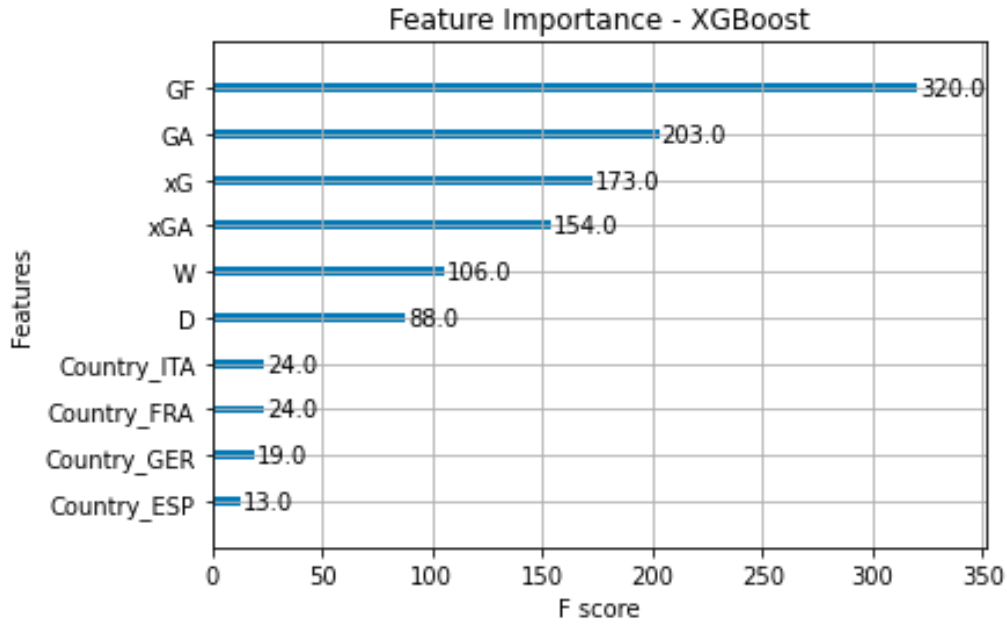

Table 5: Feature importance in XGBoost regressor model

As it can be seen from the chart, goals scored and conceded are the most important features to explain league rankings of the teams. Expected goals and expected goals against follow these two features. Since expected goals and goals scored are highly correlated, it is an expected result.

Conclusion

This project offered a comprehensive approach to analyzing and predicting team performance in the 2023-2024 European football season. By using regression models, the analysis revealed the correlation between features more clearly. The feature importance analysis further deepened the understanding of key factors that influence team rankings.