Predicting Premier League Match Outcomes

1. Introduction

Initially, I aimed to analyze bus delays in Trondheim based on weather conditions. However, I struggled to find an appropriate dataset or API with sufficient historical data. Given my strong interest in football, I pivoted towards predicting match outcomes in the Premier League.

Finding a suitable dataset was a challenge, but I eventually discovered historical Premier League data from this repository: https://github.com/datasets/football-datasets. I consolidated all seasons into a single CSV file, allowing for a structured and comprehensive analysis.

The goal of this project is to build a *classification model* to predict match outcomes: home win, draw, or away win. To evaluate the model's performance, I trained it on data from the first 10 seasons and tested it on the most recent season (2024/25).

2. Dataset Description

The dataset consists of 4,081 matches across 24 columns, including team names, match dates, goals scored, and other match-related statistics. The key variables used in the model include:

- **FTR (Full-Time Result):** The target variable with three classes (H = Home Win, D = Draw, A = Away Win)
- **Team Strength Indicators**: Historical Elo ratings for home and away teams
- **Recent Form:** Points earned in the last five matches
- **Offensive and Defensive Performance:** Goals scored and conceded over recent matches
- **Rest Days:** The number of rest days before a match

Given the moderate size of the dataset, I determined that a traditional machine learning model like Random Forest would be suitable. More complex models like XGBoost may require significantly larger datasets to provide an advantage.

3. Exploratory Data Analysis (EDA)

To understand the dataset, I conducted an EDA, uncovering the following insights:

- **Outcome Distribution:** Home teams have a slight advantage, as expected, but the distribution remains relatively balanced.
- **Recent Performance Matters:** Teams that performed well in the last five games tend to continue their positive form**.**

- **Goal Differences:** Higher average goals conceded by an away team correlate with a higher probability of a home win.
- **Rest Days Impact:** Teams with more rest days perform better, reinforcing the importance of player fatigue in match results.

These insights helped refine feature selection before training the model.

4. Model Selection and Training

Given the dataset's size (4,081 x 24), I selected Random Forest due to its ability to handle categorical and numerical data efficiently while preventing overfitting. The model was trained with **100 estimators** and evaluated on unseen test data.

The Random Forest classifier achieved an accuracy of 69%, significantly outperforming the baseline accuracy of 45% (which represents always predicting the most frequent outcome). While the model performed well in predicting home and away wins, it struggled with draws, likely due to the inherent difficulty in predicting evenly matched games.

Feature importance analysis revealed that the most significant predictors were:

- Home and Away Elo Ratings (team strength)
- Points earned in the last 5 matches
- Goals scored and conceded in the last 5 matches
- Rest days before the match

5. Future Improvements

Feature Engineering Enhancements:

To improve predictive performance, additional features could be integrated:

- **Current injuries:** The number of injured players per team before a match.
- **Recent average of expected Goals (xG):** This would be a more fair evaluation of a team's recent performance.

However, acquiring these features requires significant manual data collection as fewer open-source datasets provide them.

Real-World Deployment:

1. **Automated Betting Recommendations:** – Users receive probability-based predictions for upcoming matches.
2. **Fantasy Football Insights**: Predictions can assist fantasy players in team selection.
3. **Broadcast & Media Use:** Broadcasters could use the model for pre-match analysis.

6. Conclusion

This project successfully demonstrated how machine learning can be used to predict Premier League match outcomes. By leveraging historical data and a Random Forest classifier, the model achieved an accuracy of 69%, surpassing the baseline prediction approach.

Future improvements, such as integrating injury data and xG metrics, could significantly enhance model performance. Additionally, a real-time deployment could transform this model into a useful tool for sports analytics and betting advisory platforms.