University of Hertfordshire UH

Predicting Premier League Match Outcomes using Machine Learning.

Srihari Mohan - 23069726 - Github
M.Sc Data Science
Supervisor - Ralf Napiwotzki

Research Question & Objectives

Research Question: To what extent can Machine Learning predict match outcomes of English Premier League matches using historical match statistics? And which performance metrics are the most significant predictors of success?.

Objectives:

- Review of existing research in football analytics & match outcome prediction.
- Collect, Clean & derive features from the dataset.
- Develop, Train & Optimise the best of three models.
- Evaluate & compare models predictive performance using metrics beyond simple accuracy
- Interpret the best performing model to identify and analyse the key outcomes.

Dataset & Preprocessing

Dataset: “England CSV” (12,000+ matches). Kaggle source. Includes Goals, Shots, Fouls, Corners, etc data (1993-2025) of EPL.

Ethics: It follows all the university ethics and approved by the supervisor.

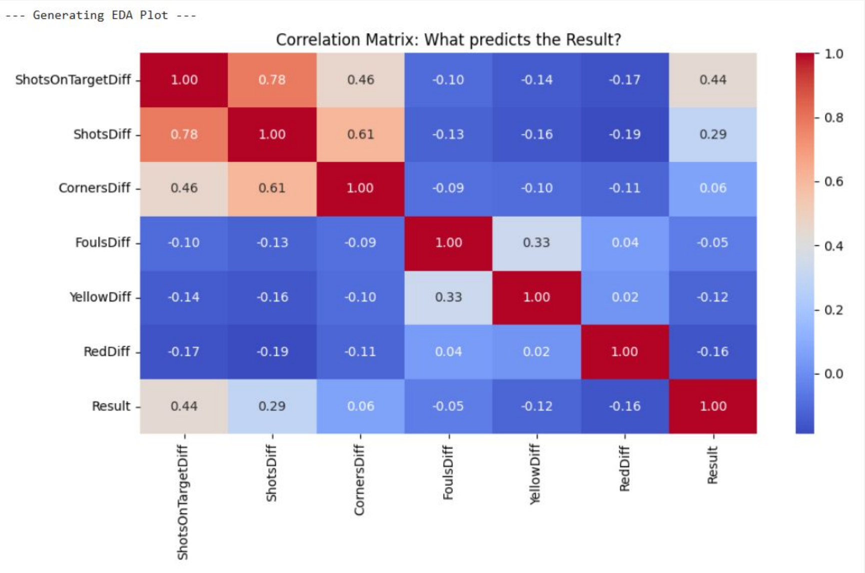
Preprocessing:

- Dropped rows with missing stats
- Excluded team names to prevent historical bias, included only required columns for the analysis.
- Created different stats as features (e.g., *ShotsonTargetDiff, RedDiff*, etc.)
- Converted Target variable *Results*(Home, Draw, Away) to numbers(0,1, 2).

Exploratory Data Analysis

- Exploratory Data Analysis to understand the data. In this correlation matrix, the feature **'*Shots on Target Difference*'** has the strongest correlation with the result.

- This scientifically justifies why it is the most important input for my model. Features like '*Fouls Difference*' had very low correlation, suggesting they are less predictive."



```
--- Generating EDA Plot ---
```

Correlation Matrix: What predicts the Result?

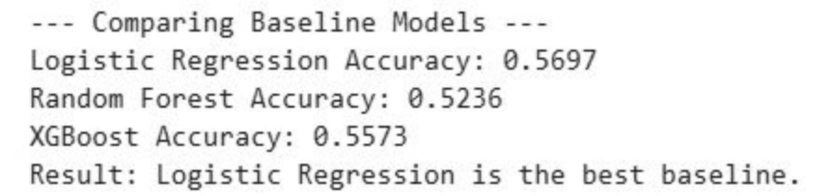
	ShotsOnTargetDiff	ShotsDiff	CornersDiff	FoulsDiff	YellowDiff	RedDiff	Result
ShotsOnTargetDiff	1.00	0.78	0.46	-0.10	-0.14	-0.17	0.44
ShotsDiff	0.78	1.00	0.61	-0.13	-0.16	-0.19	0.29
CornersDiff	0.46	0.61	1.00	-0.09	-0.10	-0.11	0.06
FoulsDiff	-0.10	-0.13	-0.09	1.00	0.33	0.04	-0.05
YellowDiff	-0.14	-0.16	-0.10	0.33	1.00	0.02	-0.12
RedDiff	-0.17	-0.19	-0.11	0.04	0.02	1.00	-0.16
Result	0.44	0.29	0.06	-0.05	-0.12	-0.16	1.00

Model Comparison

Models Tested: Logistic Regression, Random Forest and XGBoost.

Performance:

Model Name	Score
Logistic Regression	~56%
Random Forest	~52%
XGBoost	~55%



```
--- Comparing Baseline Models ---
Logistic Regression Accuracy: 0.5697
Random Forest Accuracy: 0.5236
XGBoost Accuracy: 0.5573
Result: Logistic Regression is the best baseline.
```

Key Takeaway: The simpler linear model outperformed complex tree-based models, suggesting linear relationships are dominant in this data.

Optimisation

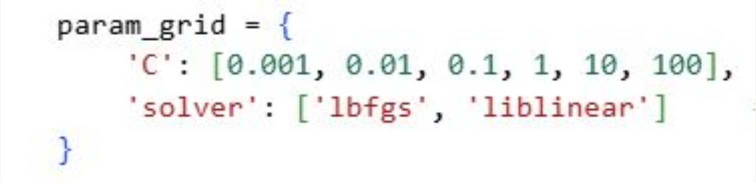
Process: Used GridSearchCV to tune the model.

Model Selected: Logistic Regression was selected as it outperformed complex tree-based models initially.

Parameters Tuned:

- **C (Regularization):** To control overfitting vs. underfitting.
- **solver:** Tested 'lbfgs' vs 'liblinear' for mathematical convergence.

Result: The grid search identified C=0.01 as the optimal setting, improving generalisation on unseen data.

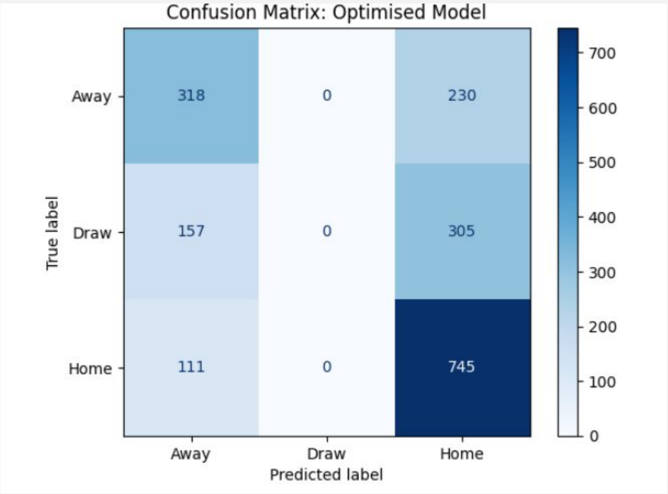


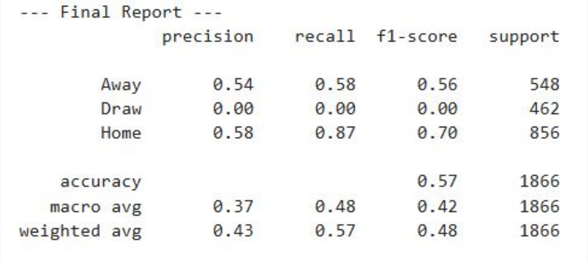
```
param_grid = {
    'C': [0.001, 0.01, 0.1, 1, 10, 100],
    'solver': ['lbfgs', 'liblinear']
}
```

Results

Key Metrics:

- **Final Accuracy: 56%** (vs. Random Chance of 33%).
- **Precision:** High for Home Wins, Low for Draws.
- **Observation:** The model effectively predicts wins/losses but struggles with Draws (the "Stochastic" element of football).



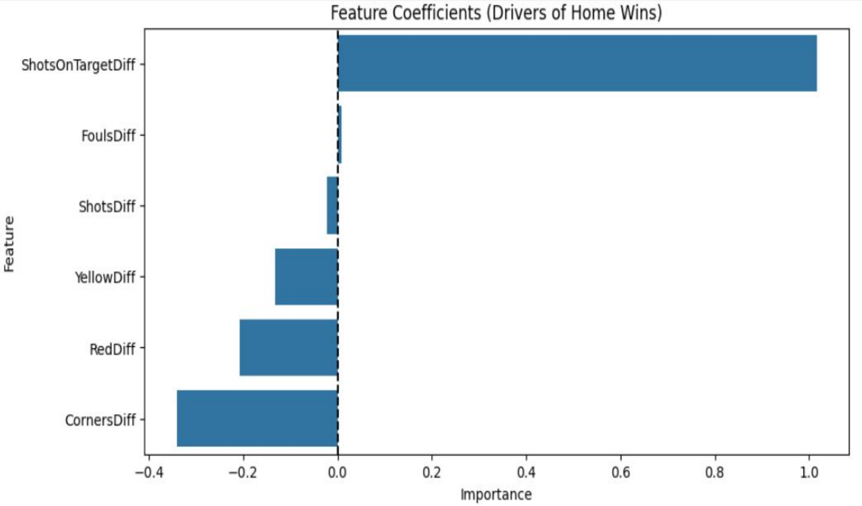


```
--- Final Report ---
              precision    recall  f1-score   support

        Away       0.54      0.58      0.56       548
        Draw       0.00      0.00      0.00       462
        Home       0.58      0.87      0.70       856

    accuracy                           0.57      1866
   macro avg       0.37      0.48      0.42      1866
weighted avg       0.43      0.57      0.48      1866
```

Analysis

Critical Analysis:

- **#1 Predictor:** *ShotsOnTargetDiff* (Positive Coefficient) – Proves that offensive efficiency is the biggest driver of wins.
- **Negative Driver:** *RedCardDiff* – Getting a red card significantly penalises chances.
- **Noise:** *FoulsDiff* had low importance, suggesting aggressive play does not guarantee wins.



Conclusion

Simple linear models (Logistic Regression) outperformed complex ones (XGBoost) for this dataset. Match outcomes are ~56% predictable based on stats; the rest is luck/human factors.

Limitations: Current model uses **post-match** statistics.

Future Work:

- Implement **Rolling Averages** (Team Form) to allow for pre-match forecasting.
- Integrate Player-level data (injuries/lineups).

References

- **Reference 1 (The Classic Statistical Model):**

  Maher, M.J., (1982), ‘Modelling association football scores’, *Statistica Neerlandica*, 36(3), pp. 109-118. (Available at: https://doi.org/10.1111/j.1467-9574.1982.tb00782.x)

- **Reference 2 (Machine Learning specific - use this for XGBoost/Random Forest):**

  Baboota, R. and Kaur, H., (2019), ‘Predictive analysis and modelling football results using machine learning approach for English Premier League’, *International Journal of Forecasting*, 35(2), pp. 741-755. (Available at: https://doi.org/10.1016/j.ijforecast.2018.01.003)

- **Reference 3 (Comparing Models):**

  Shin, J. and Gasparyan, R., (2014), ‘A novel way to soccer match prediction’, *International Journal of Contents*, 10(4), pp. 46-51. (Available at: https://doi.org/10.5392/IJoC.2014.10.4.046)