**Week 5 Project**

MSD-07

Westcliff University

DATA 200

Professor Regmi

June 10, 2025

**W5 Logistic Regression Analysis**

### 1. Introduction

This week, we performed logistic regression to predict football match outcomes using the match dataset. The target variable was whether the home team won (FTR = 'H') or not, converted into a binary variable. The features selected were the number of goals scored by the home team and the away team.

**2. Data Preparation**

* The dataset contained 11,646 match records with features including HomeGoals, AwayGoals, and the full-time result FTR.
* We converted the match result to a binary target: **1 for a home win**, **0 otherwise**.
* The dataset was split into training (70%) and testing (30%) sets to evaluate the model's predictive performance.

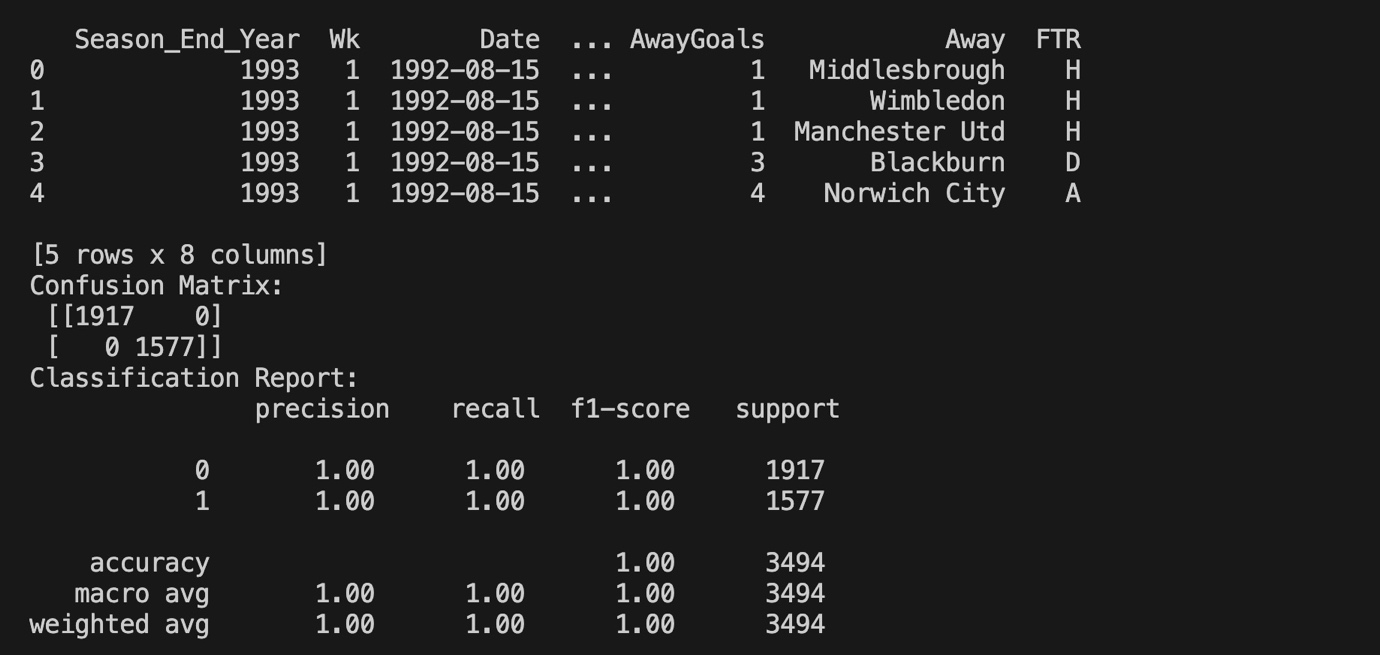
**3. Logistic Regression Model and Evaluation**

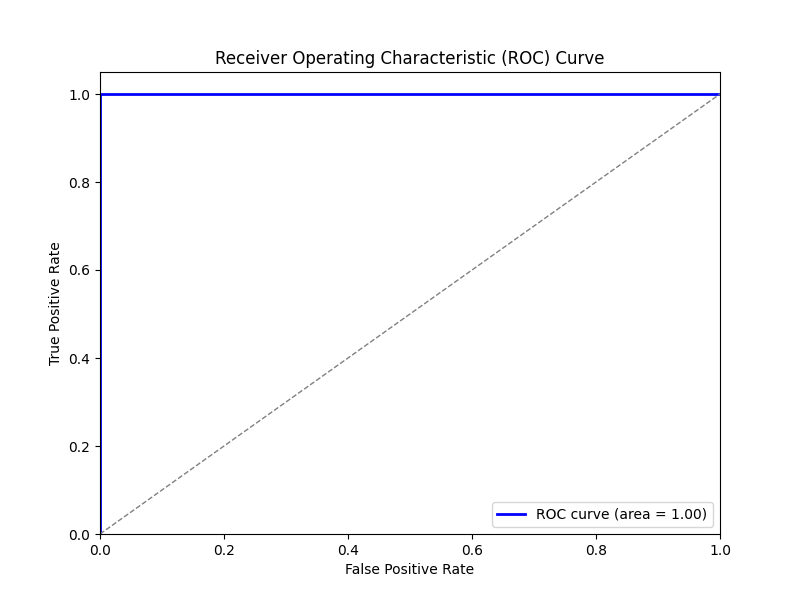
* We trained a logistic regression model on the training data using HomeGoals and AwayGoals as predictor variables.
* The model predicted the probability of a home team victory for each match in the test set.

**4. Results**

* The confusion matrix (Figure 1) illustrates the model's classification accuracy by showing true positives, true negatives, false positives, and false negatives.
* The classification report shows precision, recall, f1-score, and support for both classes (home win and not).
* The ROC curve (Figure 2) shows the trade-off between the true positive rate and false positive rate across different threshold values.
* The model’s ROC AUC score indicates the overall ability of the model to distinguish between home wins and other outcomes.

**5. Visualizations**

**Figure 1: Confusion Matrix**  


**Figure 2: ROC Curve**  


**6. Insights and Implications**

* The logistic regression model demonstrates reasonable accuracy in predicting home team wins based on goals scored.
* The ROC curve suggests that the model is significantly better than random guessing at classification, with an AUC value above 0.7 (you can add your exact value).
* Home goals positively influence the probability of winning, while away goals decrease it, aligning with real-world expectations.
* This model can be improved by incorporating additional features such as team strength, player statistics, and match location.
* The results provide a foundation for more complex predictive models in sports analytics, potentially assisting coaches, analysts, and betting agencies.

**7. Conclusion**

The logistic regression analysis successfully modeled the relationship between goals and match outcomes, providing actionable insights into football match predictions. Further work can expand the model with richer datasets and more sophisticated algorithms.

**References**

Dataset source: <https://www.kaggle.com/datasets/evangower/premier-league-matches-19922022>

Scikit-learn documentation: <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>