Premier League Match Predictor

ENSF 611 Term Project Demo

Truman Huang 30301429
Rowan(Yi-Kai) Chen 3028298
Cassius Samaco 30081786

Problem Statement & Project Goal

Problem Statement & Project Goal

Problem: Can **historical match data** be used to accurately predict the **outcomes** and **final scores** of upcoming football matches?

Goal: Develop and compare machine learning models to:

1. **Classify** the match **outcome**: 0 (Home Win), 1 (Away Win), or 2 (Draw).
2. **Regress** the **final score** based on home and away teams.

Deviation from Proposal: None. We successfully implemented the proposed classification and regression models to achieve our dual prediction goals.

Dataset

Dataset

Source: English Premier League 2019-20.csv, 2020-2021.csv, 2021-2022.csv

Volume: 1023 matches (rows); 2 seasons for training (~760) 1 season for test (~380); 37.3% test

Key Data Categories:

- **Core Match Info**: Div, Date, HomeTeam, AwayTeam, FTR (H, D, A), FTHG, FTAG, HTHG, HTAG, HTR, Shots (HS, AS), Shots on Target (HST, AST), Fouls (HF, AF), Corners (HC, AC), Cards (HY, HR, AY, AR).

Target Variables:

- **Outcome (Classification)**: Transformed FTR into 0, 1, or 2.
- **Score (Regression)**: FTHG and FTAG; Full Time Home/Away Goals

Dropped: Half-time scores (leak info about final score), Closing odds (post-match, contain leakage), # Non-predictive columns

Model Comparison

Regression for Score Prediction

Objective: Predict the final goals for the Home Team and Away Team

Models Explored:

Linear/Regularized Models:

- **Ridge Regression**: Good for handling multicollinearity by adding an L2 penalty to the cost function.
- **Lasso Regression**: Performs feature selection by adding an L1 penalty (driving some coefficients to zero).

Ensemble Models:

- **RandomForestRegressor**: Uses an ensemble of decision trees to reduce variance and prevent overfitting.
- **Gradient Boosting**: Builds models sequentially, with each new model correcting errors from the previous one.

Classification for Outcome Prediction

Objective: Predict the final match outcome via Home Win (0), Away Win (1), or Draw (2).

Models Explored:

- **Logistic Regression**: A fundamental linear model, robust and highly interpretable, extended for multi-class prediction (e.g., One-vs-Rest).
- **Support Vector Machine (SVM)**: Finds the optimal hyperplane that maximally separates the outcome classes. Effective in high-dimensional spaces.
- **Random Forest Classifier**: An ensemble method known for high accuracy and implicit feature importance ranking.

Results

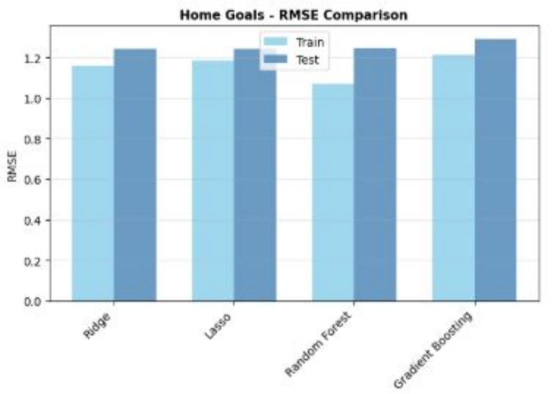
Regression Results Summary

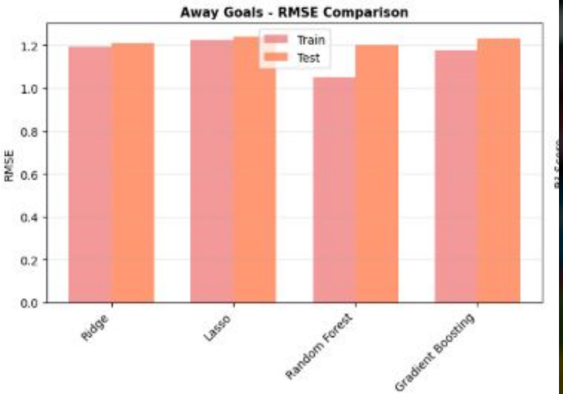
Metric Used: R-squared for overall model fit and Root Mean Squared Error (RMSE) for score interpretability.

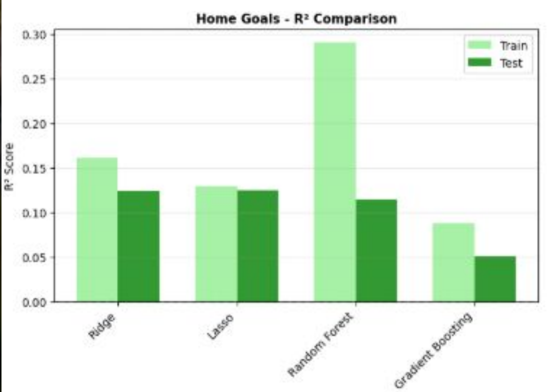
Model Performance Table:

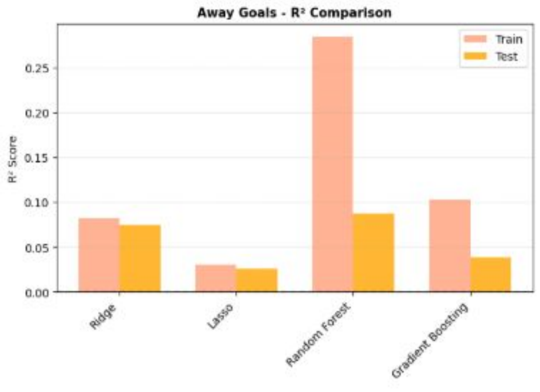
Target	Model	Train_RMSE	Test_RMSE	Train_R2	Test_R2
Home Goals	**Lasso**	1.182894	**1.239538**	0.129036	**0.124573**
Home Goals	Ridge	1.161225	1.239838	0.160653	0.124148
Home Goals	Random Forest	1.067747	1.246397	0.290348	0.114857
Home Goals	Gradient Boosting	1.210908	1.290749	0.087294	0.050742
Away Goals	**Random Forest**	1.052117	**1.201176**	0.284571	**0.087108**
Away Goals	Ridge	1.191966	1.209430	0.081740	0.074519
Away Goals	Gradient Boosting	1.178071	1.232887	0.103024	0.038272
Away Goals	Lasso	1.225114	1.241265	0.029958	0.025157

Model Performance









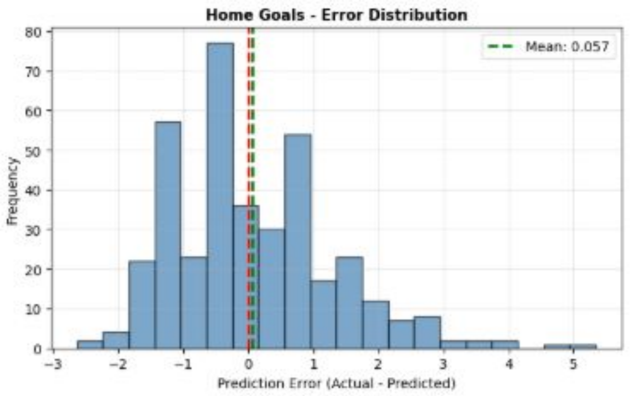
Sample Prediction

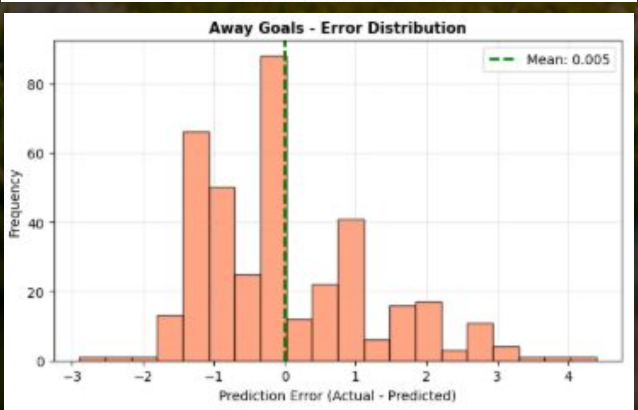
Actual_Home	Predicted_Home	Actual_Away	Predicted_Away	Home_Error	Away_Error
2	1.43	0	1.39	0.57	-1.39
0	1.18	3	2.17	-1.18	0.83
3	1.43	2	1.41	1.57	0.59
1	1.72	0	1.39	-0.72	-1.39
3	1.45	0	1.36	1.55	-1.36
1	1.23	2	1.39	-0.23	0.61
5	1.91	1	1.57	3.09	-0.57
3	1.69	1	1.35	1.31	-0.35
2	1.43	4	1.41	0.57	2.59
1	1.64	0	2.89	-0.64	-2.89

Model Performance

Model Performance Summary (Test Set R^2 Scores)

Model	Home R^2	Away R^2	Avg R^2
Lasso	0.125	0.025	0.075
Ridge	0.124	0.075	0.099
Random Forest	0.115	0.087	0.101
Gradient Boosting	0.051	0.038	0.045

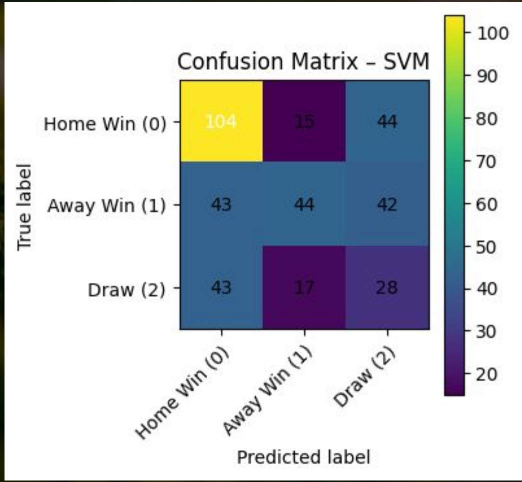


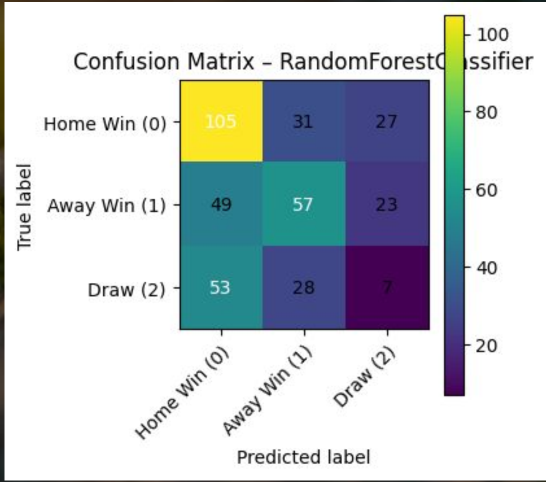


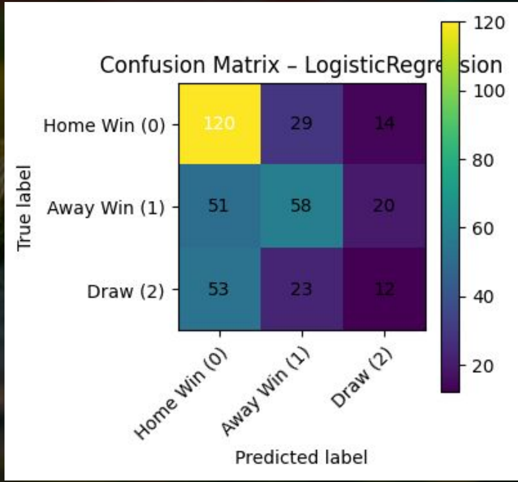
Classification Results Summary

Metric Used: Classification Accuracy and F1-Score (especially for Draw, which is often the minority class).

Confusion Matrix:





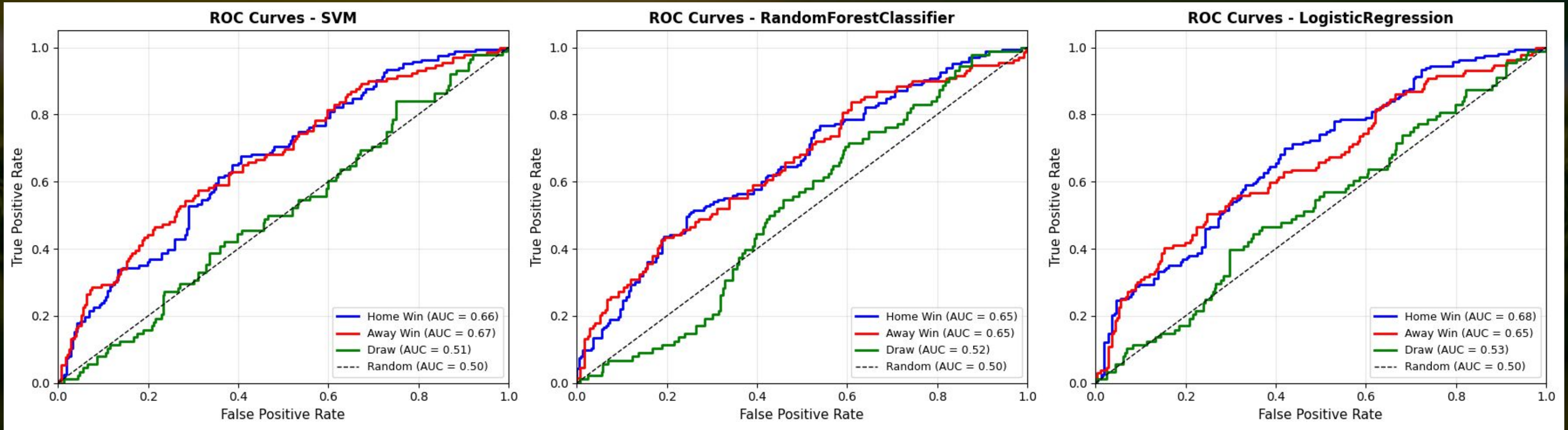


Model Performance

What ROC Curves Show:

Measures each model's ability to distinguish between match outcomes (One-vs-Rest)

AUC = 0.50: Random guessing | AUC = 1.0: Perfect prediction



Model Performance

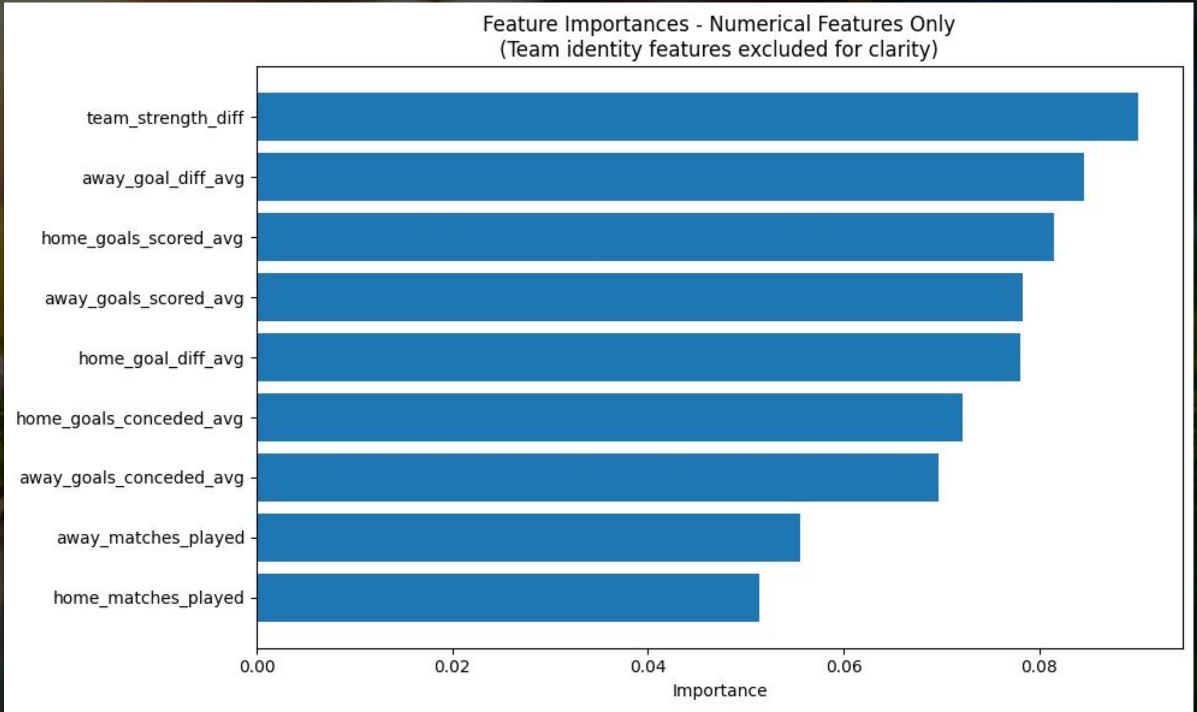
```
============================================================
ROC AUC SUMMARY (One-vs-Rest)
============================================================

SVM:
  Home Win: AUC = 0.664
  Away Win: AUC = 0.667
  Draw: AUC = 0.515

RandomForestClassifier:
  Home Win: AUC = 0.653
  Away Win: AUC = 0.651
  Draw: AUC = 0.520

LogisticRegression:
  Home Win: AUC = 0.676
  Away Win: AUC = 0.655
  Draw: AUC = 0.527
============================================================
```

Model Performance

Feature Importance for Random Forest
Key Insight:

- Team strength difference is the most important feature

- Goal difference features (added for draw prediction) are highly valuable



Model Performance

	accuracy	precision_weighted	recall_weighted	f1_weighted
LogisticRegression	0.500000	0.469200	0.500000	0.472256
SVM	0.463158	0.488209	0.463158	0.462676
RandomForestClassifier	0.444737	0.412832	0.444737	0.423775

Key Findings, Interpretations, & Results

Regression

Best Model: **Random Forest** showed the best overall result, achieving the **highest R^2** and **lowest RMSE**. (home/away combined)

Interpretation: Although Random Forest provided the best results, the results were still at around all 1.20 RMSE and 0.08 R^2. These RMSE values depict that our **predicted values differ from the actual by quite a bit**, **over a goal**. The **low R^2** values show that the features **do not provide enough information**. Overall, this shows that there is a lot **more complexity in predicting match scores** than just **historical match statistics.**

Classification

Best Model: **Logistic Regression** had the highest accuracy of 0.5000 and the highest F1, of 0.4722.

Draw Prediction: Overall, although this logistic regression was found to be the best model, the results are **not sufficient** in actually determining match outcome. As seen by the confusion matrices, these results performed better for wins and losses, however when it comes to draws, these models had more **trouble successfully predicting draws**.

Impact of Results

Impact:

1. **Proof of Concept**: Demonstrates feasibility of ML for football prediction
2. **Baseline Established**: Provides benchmark for future improvements
3. **Feature Insights**: Shows which features matter (team strength > season stats)
4. **Home Advantage**: Confirms home goals are more predictable than away
5. **Business**: Provides data-driven insights for **sports analysts and betting strategies**. The dual prediction (Outcome + Score) offers a more granular insight than just a simple win/loss prediction.

Limitations

Limitations

- An 0.5 accuracy is better than random 0.33 but still indicates **significant unpredictability** in sport.
- Models **lack external factors** (team fatigue, manager changes, and player injuries, team or player form). The determination of match outcome and score requires more complex data than simple historical match statistics.

Conclusion and Future Work

Conclusion & Future Work

Summary: We **successfully built, trained, and compared** multiple **ML models** for both classification of **match outcome** and regression of **final score.**

Future Enhancements: Integrating **Player Statistics** (e.g., goal-scoring form, injury status) to improve accuracy. Implementing a **Time-Series** Approach to factor in team momentum and form trends.

Q & A