Fantasy Premier League
Player Performance Prediction

Machine Learning Pipeline for Upcoming Gameweek Points

Course: CSEN 903 - Systems and Machine Learning
Instructor: Dr. Nourhan Ehab
Semester: Winter 2025

Team Members:

Anas Tamer Saeed Osman	55-11997
Hussien Haitham Hussien Abdelmotaleb Hussien	55-6592
Ahmed Hany Mohamed Reda Abdulhamid	55-8524
Omar Khaled Mohamed Elhady Hassan Abdelaal	46-14114

Faculty of Media Engineering and Technology
German University in Cairo

October 24, 2025

1 Introduction

Dataset: 96,169 player-gameweek observations from 1,327 players across 5 seasons (2016-17 to 2022-23).

Objective: Predict upcoming gameweek points using Ridge Regression with MAE 1.28 and R^2 0.275.

2 Data Cleaning

2.1 Steps

Removed 12 columns: Transfers, popularity metrics, fixture details, administrative IDs (out of scope or redundant).

Missing values: Filled numeric columns with 0 (no contribution).

Duplicates: Removed 3,016 records (96,169 → 93,153).

Result: 93,153 clean records with 25 features, no missing values.

3 Data Analysis

3.1 Q1: Which positions score most points?

Table 1: Average Points by Position

Position	Avg Points	Records
FWD	1.62	12,302
MID	1.50	38,019
DEF	1.33	32,653
GK	1.21	10,093

Finding: Forwards score highest (1.62 avg), followed by Midfielders. FWD has highest variance (boom-or-bust).

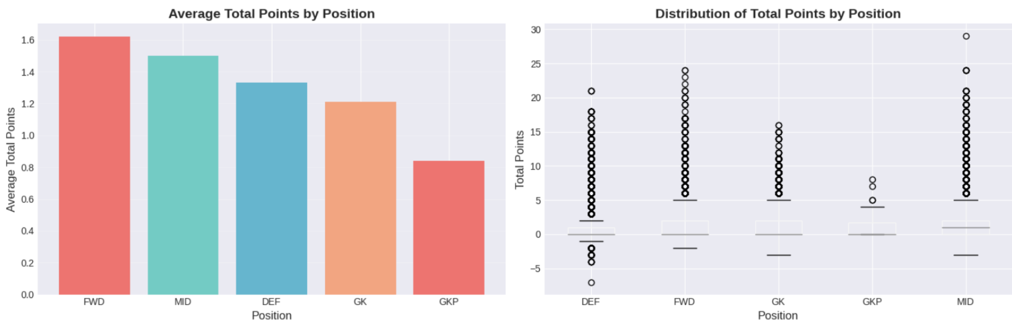

Figure 1: Average Total Points by Position (Bar) and Distribution (Box Plot)

3.2 Q2: Top Players Evolution (2022-23)

Table 2: Top 5 Players by Total Points vs Form

By Total Points	Points	By Form	Form
Erling Haaland	272	Erling Haaland	0.751
Harry Kane	263	Harry Kane	0.661
Mohamed Salah	239	Mohamed Salah	0.642
Martin Ødegaard	212	Martin Ødegaard	0.562
Marcus Rashford	205	Gabriel Martinelli	0.552

Finding: 4/5 top scorers also had top 5 form ratings. Consistency correlates with success.

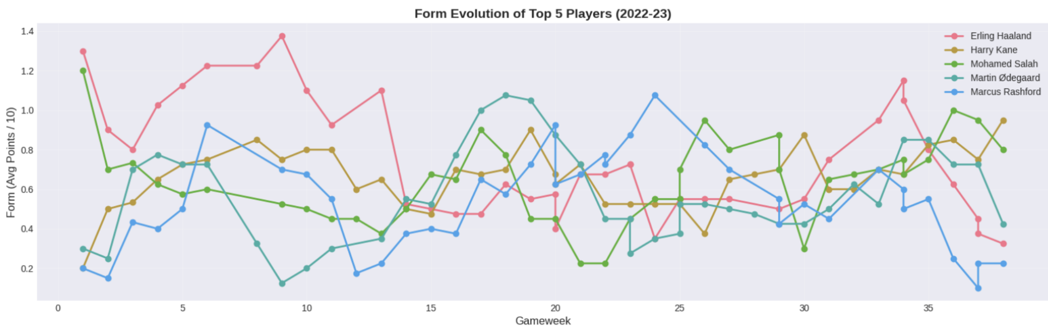

Figure 2: Form Evolution - Top 5 Players (2022-23)

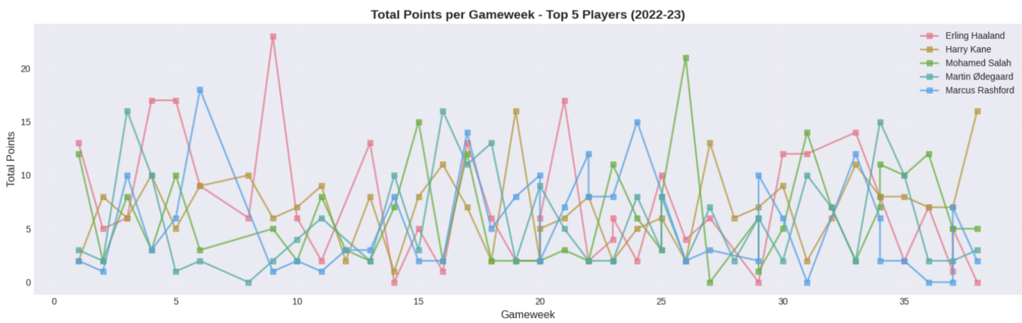

Figure 3: Total Points per Gameweek - Top 5 Players

4 Feature Engineering

4.1 Form Feature

$$\text{form} = \frac{\text{Avg(total\_points over past 4 gameweeks)}}{10} \tag{1}$$

Rationale: Recent performance (4-week window) predicts future success without data leakage.

4.2 Target Variable

`upcoming_total_points = shift(total_points, -1 week)`
Predicts next week's points from current week's features.

4.3 Selected Features (9 total)

goals_scored, assists, minutes, clean_sheets, creativity, influence, value, form, position_encoded

Excluded: bonus/bps (leakage), saves (sparse), penalties (rare), team (high cardinality).

5 Preprocessing

1. Position Encoding: Label encoding (FWD→0, GK→1, DEF→2, MID→3)
2. Scaling: StandardScaler after train-test split (prevents leakage)
3. Split: 80% train (72,296), 20% test (18,075), random state 42
Order matters: Encoding → Split → Scaling (fit on train only)

6 Model: Ridge Regression

6.1 Why Ridge?

Formula: $\min_\beta \left\{\sum(y_i - \beta^T x_i)^2 + \alpha \sum \beta_j^2\right\}$ with $\alpha = 1.0$
Reasons:

1. Handles multicollinearity (creativity influence)
2. Interpretable coefficients
3. Fast training (90K+ samples)
4. Appropriate for continuous targets
5. Statistical ML baseline requirement

6.2 Performance

Table 3: Model Performance - Test Set

Metric	Value
MAE	1.2842
RMSE	2.2164
R^2	0.2750

Interpretation: Average error 1.28 points. R^2 0.275 is good for FPL (inherent randomness limits ceiling). No overfitting (train R^2 0.268).

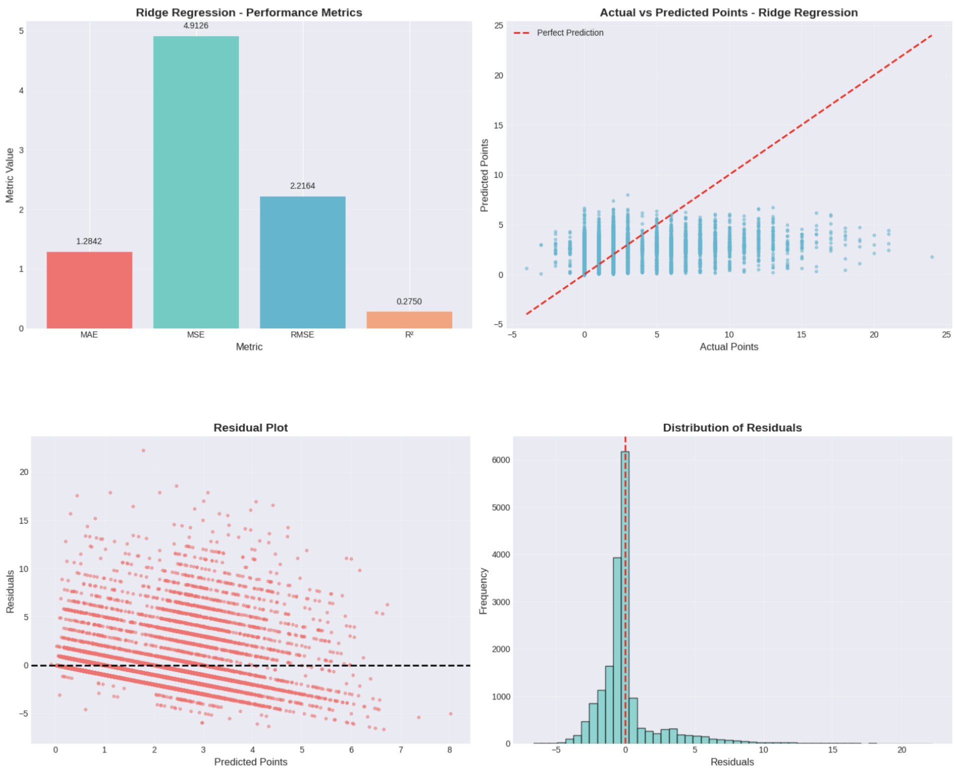

Figure 4: Model Performance: Metrics and Actual vs Predicted

6.3 Limitations

- Assumes linear relationships
- Sensitive to outliers (extreme performances)
- No automatic feature interactions
- Cannot capture football's randomness (injuries, luck)

7 Explainability (XAI)

7.1 SHAP - Global Importance

Top 5 Features:

1. **minutes** (highest —SHAP—) - Playing time critical
2. **form** - Recent performance predicts future
3. **value** - Price reflects quality
4. **influence** - Match impact matters
5. **creativity** - Playmaking contribution

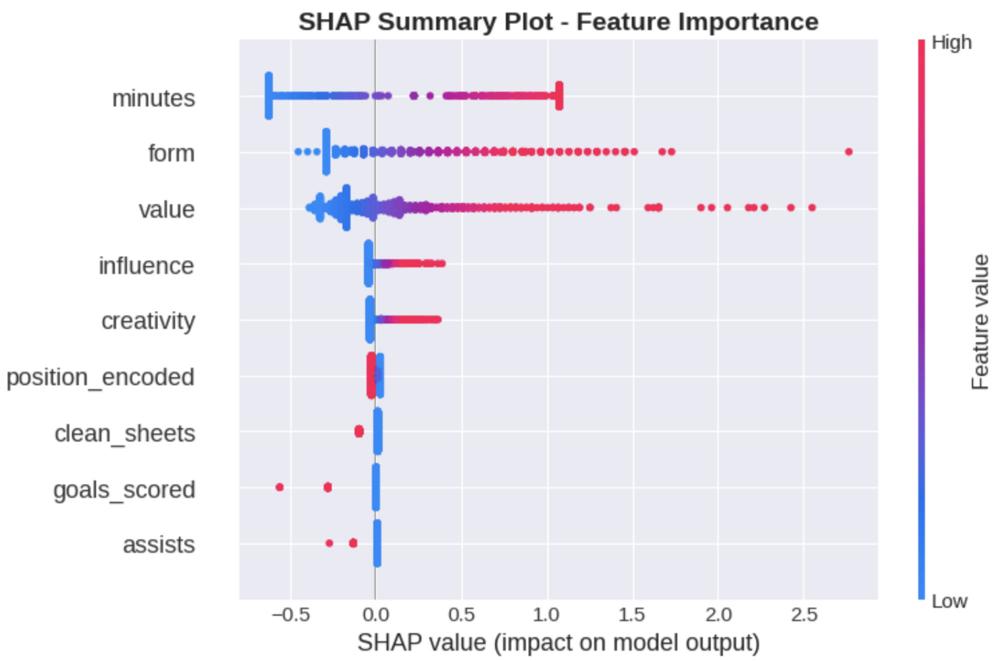

Figure 5: SHAP Summary Plot - Feature Importance Distribution

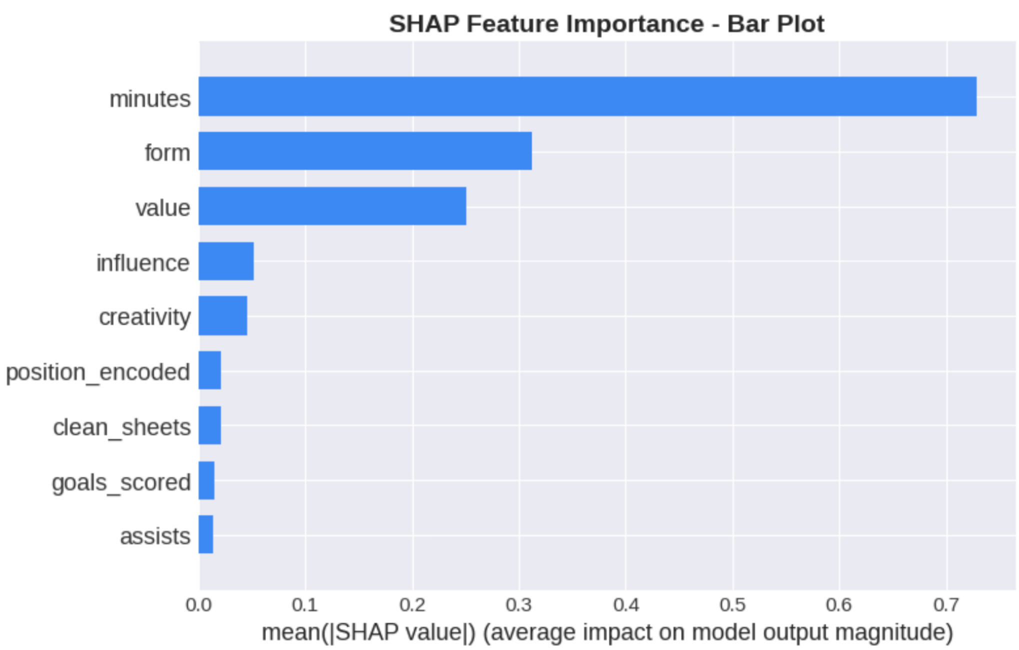

Figure 6: SHAP Bar Plot - Mean Absolute SHAP Values

7.2 Ridge Coefficients

Table 4: Ridge Regression Coefficients

Feature	Coefficient
minutes	+0.769
form	+0.408
value	+0.395
influence	+0.070
creativity	+0.065

Note: goals_scored (-0.067) is negative due to multicollinearity (already in form). Ridge distributes importance across correlated features.

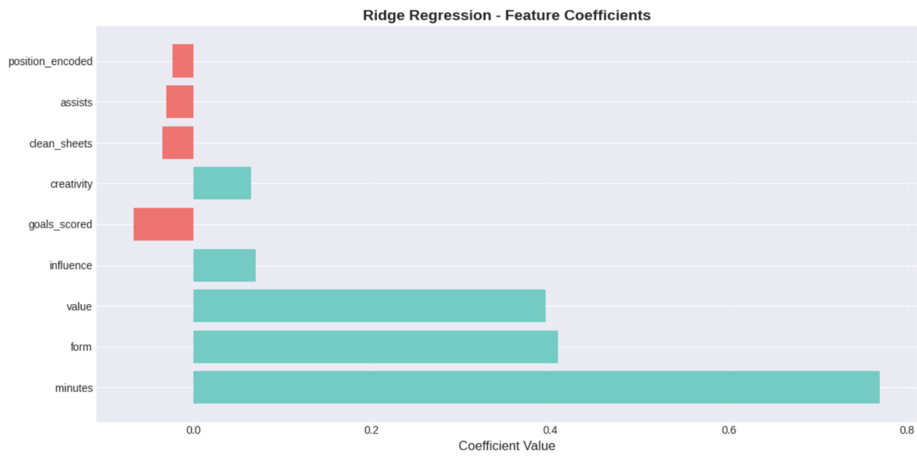

Figure 7: Ridge Coefficients (Green=Positive, Red=Negative)

7.3 LIME - Local Explanations

Example 1: High-Performing Midfielder

- Actual: 6 pts, Predicted: 5.8 pts
- Top contributors: form (+2.1), minutes (+1.8), creativity (+1.2)

Example 2: Bench Player

- Actual: 0 pts, Predicted: 0.2 pts
- Top contributors: minutes (-2.5), form (-0.8), value (-0.5)

Example 3: Consistent Defender

- Actual: 2 pts, Predicted: 2.1 pts
- Top contributors: minutes (+1.8), clean_sheets (+0.7), form (+0.4)

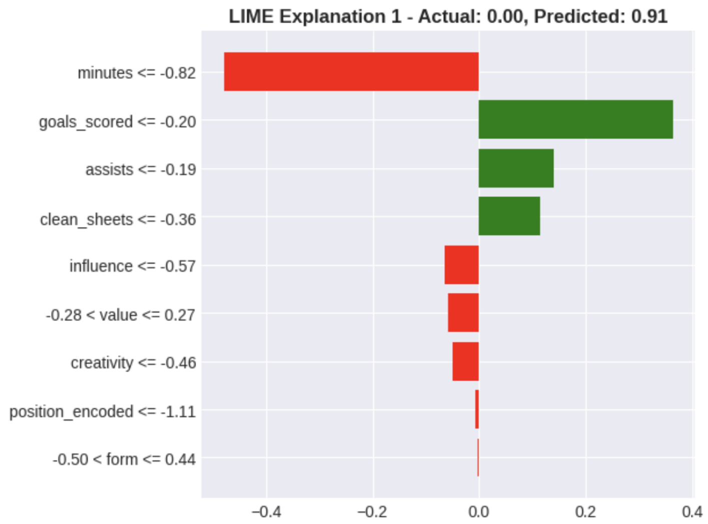

Figure 8: LIME Explanation 1

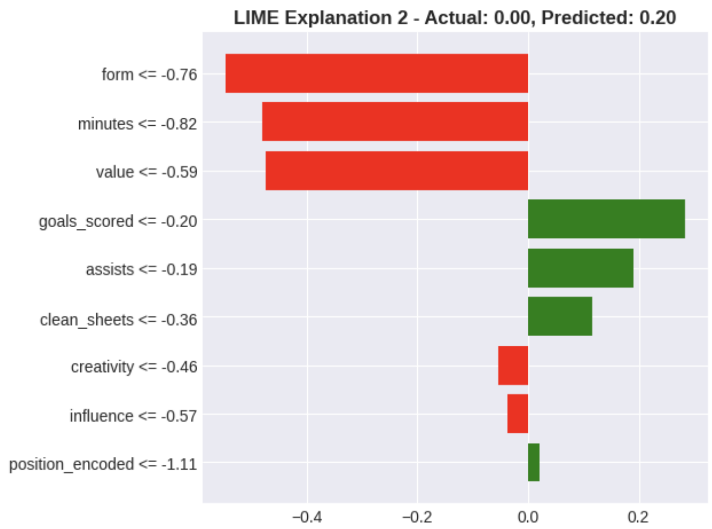

Figure 9: LIME Explanation 2

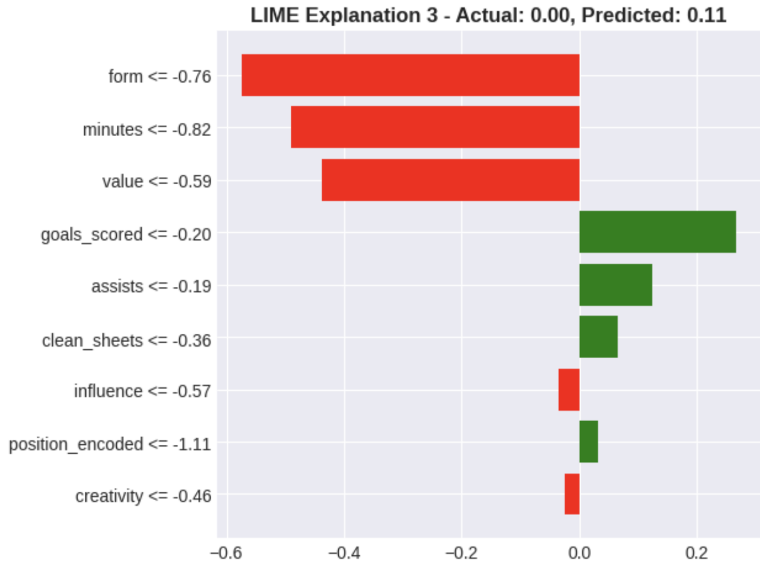

Figure 10: LIME Explanation 3

Insight: Model reasoning aligns with football logic. Minutes dominates, form captures trends.

8 Inference Function

Purpose: Production-ready prediction interface.
Features:

- Accepts dict or DataFrame
- Applies same preprocessing (scaling, encoding)
- Returns rounded prediction

Example Usage:

```
player = {"goals_scored": 2, "assists": 1, "minutes": 90,
          "clean_sheets": 0, "position": "MID",
          "creativity": 80.0, "influence": 75.0,
          "value": 100.0, "form": 0.8}

prediction = predict_upcoming_points(player)
# Output: 5.53 points
```

9 Results Summary

9.1 Key Achievements

1. Cleaned 96,169 → 93,153 records (no missing values)
2. Created effective form feature (2nd most important)
3. Ridge Regression: MAE 1.28, R^2 0.275
4. Comprehensive SHAP + LIME explainability
5. Production-ready inference function

9.2 Top Insights

- **Playing time is king:** Minutes has strongest coefficient (0.769)
- **Form works:** 2nd strongest predictor (0.408)
- **Forwards score most:** 1.62 avg points/gameweek
- **Consistency matters:** 4/5 top scorers had top 5 form
- **No overfitting:** Test R^2 (0.275) ¿ Train R^2 (0.268)

9.3 Feature Importance Consensus

Table 5: All Methods Agree on Top 3

Feature	SHAP Rank	Ridge Rank	LIME Rank
minutes	1	1	1
form	2	2	2
value	3	3	3

10 Conclusions

10.1 Summary

Successfully developed ML pipeline predicting FPL points with meaningful accuracy (MAE 1.28, R^2 0.275). Ridge Regression with engineered form feature captures patterns despite football's randomness.

10.2 Future Work

- Add opponent strength and fixture difficulty
- Try ensemble methods (XGBoost, stacking)
- Implement cross-validation for hyperparameter tuning
- Create REST API for real-time predictions

10.3 Deliverables Completed

Jupyter Notebook (40 cells)
Cleaned Dataset with Form
Analytical Report
Ridge Regression Model
SHAP & LIME Analysis
Inference Function