# Phân tích các feature cho vị trí Tiền đạo (FW) trong dự đoán giá trị chuyển nhượng (ETV)

Người làm: Nguyễn Hải Nam

Ngày: 3 tháng 5 năm 2025

Tài liệu này phân tích các feature được chọn để dự đoán giá trị chuyển nhượng ước tính (ETV) cho vị trí Tiền đạo trong bóng đá, sử dụng dữ liệu từ tệp result.csv.

# Mục lục

## 1. Tổng quan về feature của tiền đạo

Trong bài toán dự đoán giá trị chuyển nhượng ước tính (ETV) cho vị trí Tiền đạo (FW), các feature được chọn cần phản ánh các khía cạnh chính của vai trò tiền đạo, bao gồm:

- **Khả năng ghi bàn:** Tạo ra và tận dụng cơ hội ghi bàn thông qua bàn thắng, tỷ lệ sút trúng đích, và hiệu quả ghi bàn.
- **Khả năng sáng tạo:** Tạo cơ hội ghi bàn cho đồng đội thông qua kiến tạo và các hành động dẫn đến bàn thắng.
- **Khả năng tiến công:** Đưa bóng vào các khu vực nguy hiểm qua mang bóng hoặc tranh chấp bóng bổng.
- **Hiệu quả mỗi 90 phút:** Đo lường hiệu suất ổn định trong thời gian thi đấu.

Các feature được chia thành:

- **Feature liên quan đến ghi bàn:** Bàn thắng (Gls), Bàn thắng kỳ vọng mỗi 90 phút (xG per 90), Số bàn thắng trung bình mỗi 90 phút, Tỷ lệ sút trúng đích (SoT%), Tỷ lệ ghi bàn trên mỗi cú sút (G per sh).
- **Feature liên quan đến sáng tạo:** Số pha kiến tạo (Ast), Hành động tạo cơ hội sút mỗi 90 phút (SCA90), Hành động tạo ra bàn thắng mỗi 90 phút (GCA90).
- **Feature liên quan đến tiến công:** Số lần mang bóng tiến gần khung thành (PrgC), Số lần mang bóng vào 1/3 sân cuối (Carries 1/3), Tỷ lệ thắng tranh chấp trên không (Aerial Won%).

**Lý do chọn:** Các feature này bao quát vai trò của tiền đạo, phản ánh đúng giá trị thị trường, và có sẵn trong tệp result.csv. Chúng được chọn dựa trên:

- Liên quan đến vai trò tiền đạo: Tiền đạo cần ghi bàn, kiến tạo, và đưa bóng vào khu vực nguy hiểm.
- Tầm quan trọng trong thị trường chuyển nhượng: Các chỉ số này là yếu tố chính mà các CLB xem xét khi định giá tiền đạo.
- Dữ liệu sẵn có: Các feature được thu thập từ thống kê bóng đá tiêu chuẩn.

## 2. Phân tích chi tiết từng feature

### 2.1. Bàn thắng (Gls - Goals)

#### 2.1.1. Định nghĩa

Bàn thắng (Gls) là số bàn thắng mà tiền đạo ghi được trong mùa giải. Ví dụ: Một tiền đạo ghi 25 bàn thắng trong mùa giải.

#### 2.1.2. Lý do chọn

- Cốt lõi của vai trò tiền đạo: Ghi bàn là nhiệm vụ chính của tiền đạo, trực tiếp ảnh hưởng đến kết quả trận đấu.
- Tương quan với ETV: Tiền đạo có Gls cao (như Erling Haaland) thường có ETV cao hơn, vì họ mang lại giá trị bàn thắng.
- Phổ biến: Gls là chỉ số tiêu chuẩn, có sẵn trong result.csv.

#### 2.1.3. Cách xử lý trong mã

- Chuyển đổi và điền giá trị thiếu: Chuyển thành số, điền giá trị thiếu bằng trung vị (hoặc 0 nếu trung vị là NaN).
- Tăng trọng số: Là `important_features`, được nhân với 2.0 để nhấn mạnh vai trò ghi bàn.
- Chuẩn hóa: Chuẩn hóa bằng `StandardScaler` để đưa về thang đo tương đương.

#### 2.1.4. Ví dụ minh họa

Tiền đạo A có Gls = 30 và tiền đạo B có Gls = 15. Tiền đạo A có khả năng được định giá cao hơn, vì anh ta ghi nhiều bàn hơn.

### 2.2. Số pha kiến tạo (Ast - Assists)

#### 2.2.1. Định nghĩa

Số pha kiến tạo (Ast) là số lần tiền đạo chuyền bóng dẫn đến bàn thắng trực tiếp. Ví dụ: Tiền đạo có 10 pha kiến tạo trong mùa giải.

#### 2.2.2. Lý do chọn

- Khả năng sáng tạo: Ast đo lường khả năng tạo cơ hội ghi bàn cho đồng đội, bổ sung cho vai trò ghi bàn.

- Tương quan với ETV: Tiền đạo có Ast cao thường được định giá cao hơn, vì họ đa năng hơn.
- Phổ biến: Ast là chỉ số tiêu chuẩn, có sẵn trong result.csv.

#### 2.2.3. Cách xử lý trong mã

- Chuyển đổi và điền giá trị thiếu: Tương tự Gls.
- Biến đổi log: Áp dụng `np.log1p`.
- Tăng trọng số: Là `important_features`, được nhân với 2.0.
- Chuẩn hóa: Chuẩn hóa bằng `StandardScaler`.

#### 2.2.4. Ví dụ minh họa

Tiền đạo A có Ast = 12 và tiền đạo B có Ast = 5. Tiền đạo A có khả năng được định giá cao hơn, vì anh ta sáng tạo hơn.

### 2.3. Bàn thắng kỳ vọng mỗi 90 phút (xG per 90)

#### 2.3.1. Định nghĩa

Bàn thắng kỳ vọng mỗi 90 phút (xG per 90) là giá trị kỳ vọng của số bàn thắng mà tiền đạo ghi được trong 90 phút, dựa trên chất lượng cơ hội. Ví dụ: Tiền đạo có xG per 90 = 0.8.

#### 2.3.2. Lý do chọn

- Đo lường chất lượng cơ hội: xG per 90 phản ánh khả năng tận dụng cơ hội ghi bàn.
- Tương quan với ETV: Tiền đạo có xG per 90 cao thường được định giá cao hơn, vì họ thường xuyên ở vị trí ghi bàn.
- Phổ biến: xG per 90 là chỉ số nâng cao, có sẵn trong result.csv.

#### 2.3.3. Cách xử lý trong mã

- Chuyển đổi và điền giá trị thiếu: Tương tự Gls.
- Biến đổi log: Áp dụng `np.log1p`.
- Tăng trọng số: Là `important_features`, được nhân với 2.0.
- Chuẩn hóa: Chuẩn hóa bằng `StandardScaler`.

#### 2.3.4. Ví dụ minh họa

Tiền đạo A có xG per 90 = 0.9 và tiền đạo B có xG per 90 = 0.4. Tiền đạo A có khả năng được định giá cao hơn, vì anh ta có nhiều cơ hội chất lượng hơn.

### 2.4. Số bàn thắng trung bình mỗi 90 phút

#### 2.4.1. Định nghĩa

Số bàn thắng trung bình mỗi 90 phút là số bàn thắng mà tiền đạo ghi được trung bình trong 90 phút thi đấu:

$$\text{Gls per 90} = \frac{\text{Tổng số bàn thắng}}{\text{Tổng số phút thi đấu}} \times 90$$

Ví dụ: Tiền đạo ghi 20 bàn trong 1800 phút, thì Gls per 90 = 1.0.

#### 2.4.2. Lý do chọn

- Đo lường hiệu suất ghi bàn: Gls per 90 phản ánh hiệu quả ghi bàn theo thời gian thi đấu.
- Tương quan với ETV: Tiền đạo có Gls per 90 cao thường được định giá cao hơn.
- Phổ biến: Gls per 90 có thể tính từ Gls và số phút thi đấu, có sẵn trong result.csv.

#### 2.4.3. Cách xử lý trong mã

- Chuyển đổi và điền giá trị thiếu: Tương tự Gls.
- BiHAi đổi log: Áp dụng `np.log1p`.
- Tăng trọng số: Là `important_features`, được nhân với 2.0.
- Chuẩn hóa: Chuẩn hóa bằng `StandardScaler`.

#### 2.4.4. Ví dụ minh họa

Tiền đạo A có Gls per 90 = 1.2 và tiền đạo B có Gls per 90 = 0.6. Tiền đạo A có khả năng được định giá cao hơn, vì anh ta ghi bàn hiệu quả hơn.

### 2.5. Tỷ lệ sút trúng đích (SoT% - Shots on Target Percentage)

#### 2.5.1. Định nghĩa

Tỷ lệ sút trúng đích (SoT%) là tỷ lệ phần trăm các cú sút trúng khung thành so với tổng số cú sút:

$$\text{SoT\%} = \frac{\text{Số cú sút trúng đích}}{\text{Tổng số cú sút}} \times 100$$

Ví dụ: Tiền đạo có 50 cú sút trúng đích trong 100 cú sút, thì SoT% = 50%.

#### 2.5.2. Lý do chọn

- Đo lường độ chính xác: SoT% phản ánh khả năng sút bóng chính xác của tiền đạo.
- Tương quan với ETV: Tiền đạo có SoT% cao thường được định giá cao hơn, vì họ tạo áp lực lên thủ môn.
- Phổ biến: SoT% là chỉ số tiêu chuẩn, có sẵn trong result.csv.

#### 2.5.3. Cách xử lý trong mã

- Chuyển đổi và điền giá trị thiếu: Tương tự Gls.
- Biến đổi log: Áp dụng `np.log1p`.
- Không tăng trọng số: Không nằm trong `important_features`.
- Chuẩn hóa: Chuẩn hóa bằng `StandardScaler`.

#### 2.5.4. Ví dụ minh họa

Tiền đạo A có SoT% = 55% và tiền đạo B có SoT% = 30%. Tiền đạo A có khả năng được định giá cao hơn, vì anh ta sút chính xác hơn.

### 2.6. Tỷ lệ ghi bàn trên mỗi cú sút (G per sh - Goals per Shot)

#### 2.6.1. Định nghĩa

Tỷ lệ ghi bàn trên mỗi cú sút (G per sh) là số bàn thắng trung bình trên mỗi cú sút:

$$\text{G per sh} = \frac{\text{Số bàn thắng}}{\text{Tổng số cú sút}}$$

Ví dụ: Tiền đạo ghi 20 bàn trong 80 cú sút, thì G per sh = 0.25.

#### 2.6.2. Lý do chọn

- Đo lường hiệu quả ghi bàn: G per sh phản ánh khả năng chuyển hóa cú sút thành bàn thắng.
- Tương quan với ETV: Tiền đạo có G per sh cao thường được định giá cao hơn.
- Phổ biến: G per sh có thể tính từ Gls và số cú sút, có sẵn trong result.csv.

#### 2.6.3. Cách xử lý trong mã

- Chuyển đổi và điền giá trị thiếu: Tương tự Gls.
- Biến đổi log: Áp dụng `np.log1p`.

- Tăng trọng số: Là `important_features`, được nhân với 2.0.
- Chuẩn hóa: Chuẩn hóa bằng `StandardScaler`.

#### 2.6.4. Ví dụ minh họa

Tiền đạo A có G per sh = 0.3 và tiền đạo B có G per sh = 0.1. Tiền đạo A có khả năng được định giá cao hơn, vì anh ta hiệu quả hơn trong việc ghi bàn.

### 2.7. Số lần mang bóng tiến gần khung thành (PrgC - Progressive Carries)

#### 2.7.1. Định nghĩa

Số lần mang bóng tiến gần khung thành (PrgC) là số lần tiền đạo rê bóng hoặc mang bóng tiến gần hơn đến khung thành đối phương. Ví dụ: Tiền đạo thực hiện 60 lần mang bóng tiến gần khung thành trong mùa giải.

#### 2.7.2. Lý do chọn

- Tấn công trực tiếp: PrgC thể hiện khả năng tự mình tạo nguy hiểm.
- Tương quan với ETV: Tiền đạo có PrgC cao thường được định giá cao hơn, vì họ đe dọa hàng thủ đối phương.
- Phổ biến: PrgC là chỉ số nâng cao, có sẵn trong result.csv.

#### 2.7.3. Cách xử lý trong mã

- Chuyển đổi và điền giá trị thiếu: Tương tự Gls.
- Biến đổi log: Áp dụng `np.log1p`.
- Tăng trọng số: Là `important_features`, được nhân với 2.0.
- Chuẩn hóa: Chuẩn hóa bằng `StandardScaler`.

#### 2.7.4. Ví dụ minh họa

Tiền đạo A có PrgC = 70 và tiền đạo B có PrgC = 30. Tiền đạo A có khả năng được định giá cao hơn, vì anh ta tấn công trực tiếp hơn.

### 2.8. Số lần mang bóng vào 1/3 sân cuối (Carries 1/3)

#### 2.8.1. Định nghĩa

Số lần mang bóng vào 1/3 sân cuối (Carries 1/3) là số lần tiền đạo rê bóng hoặc mang bóng vào 1/3 sân gần khung thành đối phương. Ví dụ: Tiền đạo thực hiện 50 lần mang bóng vào 1/3 sân cuối trong mùa giải.

#### 2.8.2. Lý do chọn

- Tạo nguy hiểm: Carries 1/3 thể hiện khả năng đưa bóng vào khu vực tấn công nguy hiểm.
- Tương quan với ETV: Tiền đạo có Carries 1/3 cao thường được định giá cao hơn.
- Phổ biến: Carries 1/3 là chỉ số nâng cao, có sẵn trong result.csv.

#### 2.8.3. Cách xử lý trong mã

- Chuyển đổi và điền giá trị thiếu: Tương tự Gls.
- Biến đổi log: Áp dụng `np.log1p`.
- Tăng trọng số: Là `important_features`, được nhân với 2.0.
- Chuẩn hóa: Chuẩn hóa bằng `StandardScaler`.

#### 2.8.4. Ví dụ minh họa

Tiền đạo A có Carries 1/3 = 60 và tiền đạo B có Carries 1/3 = 25. Tiền đạo A có khả năng được định giá cao hơn, vì anh ta đưa bóng vào khu vực nguy hiểm nhiều hơn.

### 2.9. Tỷ lệ thắng tranh chấp trên không (Aerial Won% - Aerial Duels Won Percentage)

#### 2.9.1. Định nghĩa

Tỷ lệ thắng tranh chấp trên không (Aerial Won%) là tỷ lệ phần trăm các pha tranh chấp bóng bổng mà tiền đạo thắng:

$$\text{Aerial Won\%} = \frac{\text{Số lần thắng tranh chấp bóng bổng}}{\text{Tổng số pha tranh chấp bóng bổng}} \times 100$$

Ví dụ: Tiền đạo thắng 40/50 pha tranh chấp, thì Aerial Won% = 80%.

#### 2.9.2. Lý do chọn

- Khả năng không chiến: Aerial Won% quan trọng với tiền đạo cao lớn hoặc chơi ở vị trí mục tiêu.
- Tương quan với ETV: Tiền đạo có Aerial Won% cao thường được định giá cao hơn, đặc biệt trong các đội chơi bóng dài.
- Phổ biến: Aerial Won% là chỉ số tiêu chuẩn, có sẵn trong result.csv.

#### 2.9.3. Cách xử lý trong mã

- Chuyển đổi và điền giá trị thiếu: Tương tự Gls.

- Biến đổi log: Áp dụng `np.log1p`.
- Không tăng trọng số: Không nằm trong `important_features`.
- Chuẩn hóa: Chuẩn hóa bằng `StandardScaler`.

#### 2.9.4. Ví dụ minh họa

Tiền đạo A có Aerial Won% = 85% và tiền đạo B có Aerial Won% = 60%. Tiền đạo A có khả năng được định giá cao hơn, vì anh ta hiệu quả hơn trong không chiến.

### 2.10. Hành động tạo cơ hội sút mỗi 90 phút (SCA90 - Shot-Creating Actions per 90)

#### 2.10.1. Định nghĩa

Hành động tạo cơ hội sút mỗi 90 phút (SCA90) là số hành động dẫn đến cú sút (chuyền bóng, rê bóng, phạm lỗi dẫn đến sút) trung bình trong 90 phút. Ví dụ: Tiền đạo có SCA90 = 3.5.

#### 2.10.2. Lý do chọn

- Đo lường sự sáng tạo: SCA90 phản ánh khả năng tạo cơ hội sút, bổ sung cho vai trò ghi bàn.
- Tương quan với ETV: Tiền đạo có SCA90 cao thường được định giá cao hơn, vì họ đóng góp vào tấn công.
- Phổ biến: SCA90 là chỉ số nâng cao, có sẵn trong result.csv.

#### 2.10.3. Cách xử lý trong mã

- Chuyển đổi và điền giá trị thiếu: Tương tự Gls.
- Biến đổi log: Áp dụng `np.log1p`.
- Tăng trọng số: Là `important_features`, được nhân với 2.0.
- Chuẩn hóa: Chuẩn hóa bằng `StandardScaler`.

#### 2.10.4. Ví dụ minh họa

Tiền đạo A có SCA90 = 4.0 và tiền đạo B có SCA90 = 2.0. Tiền đạo A có khả năng được định giá cao hơn, vì anh ta sáng tạo hơn.

### 2.11. Hành động tạo ra bàn thắng mỗi 90 phút (GCA90 - Goal-Creating Actions per 90)

#### 2.11.1. Định nghĩa

Hành động tạo ra bàn thắng mỗi 90 phút (GCA90) là số hành động dẫn trực tiếp đến bàn thắng (chuyền bóng, rê bóng, phạm lỗi dẫn đến bàn) trung bình trong 90 phút. Ví dụ: Tiền đạo có GCA90 = 0.5.

#### 2.11.2. Lý do chọn

- Đo lường hiệu quả sáng tạo: GCA90 tập trung vào các hành động dẫn đến bàn thắng.
- Tương quan với ETV: Tiền đạo có GCA90 cao thường được định giá cao hơn, vì họ trực tiếp tạo ra bàn thắng.
- Phổ biến: GCA90 là chỉ số nâng cao, có sẵn trong result.csv.

#### 2.11.3. Cách xử lý trong mã

- Chuyển đổi và điền giá trị thiếu: Tương tự Gls.
- Biến đổi log: Áp dụng `np.log1p`.
- Tăng trọng số: Là `important_features`, được nhân với 2.0.
- Chuẩn hóa: Chuẩn hóa bằng `StandardScaler`.

#### 2.11.4. Ví dụ minh họa

Tiền đạo A có GCA90 = 0.6 và tiền đạo B có GCA90 = 0.2. Tiền đạo A có khả năng được định giá cao hơn, vì anh ta tạo ra nhiều bàn thắng hơn.

## 3. Lý do các feature này được chọn cùng nhau

- **Phản ánh toàn diện vai trò tiền đạo:**
  - *Ghi bàn:* Gls, xG per 90, Gls per 90, SoT%, G per sh đo lường khả năng ghi bàn và hiệu quả sút.
  - *Sáng tạo:* Ast, SCA90, GCA90 thể hiện khả năng tạo cơ hội và đóng góp vào bàn thắng.
  - *Tiến công:* PrgC, Carries 1/3, Aerial Won% đo lường khả năng đưa bóng vào khu vực nguy hiểm và không chiến.
- **Bổ sung lẫn nhau:**
  - Gls và xG per 90 đo lường kết quả và kỳ vọng ghi bàn.

  - SoT% và G per sh đo lường độ chính xác và hiệu quả sút.
  - SCA90 và GCA90 đo lường số lượng và chất lượng sáng tạo.
  - PrgC và Carries 1/3 cung cấp các khía cạnh khác nhau của tiến công.
- **Tương quan với thị trường chuyển nhượng:** Tiền đạo toàn diện (ghi bàn, sáng tạo, tấn công trực tiếp) như Erling Haaland có ETV cao nhờ các chỉ số này.
- **Phù hợp với dữ liệu:** Các feature đều có sẵn trong result.csv.

## 4. Cách các feature tương tác với mô hình

Mô hình hồi quy tuyến tính giả định ETV là tổ hợp tuyến tính:

$$\text{ETV} = \beta_0 + \beta_1 \cdot \text{Gls} + \beta_2 \cdot \text{Ast} + \beta_3 \cdot \text{xG per 90} + \cdots + \beta_{11} \cdot \text{SCA90} + \beta_{12} \cdot \text{GCA90}$$

Trong đó:

- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{11}, \beta_{12}$ (cho Gls, Ast, xG per 90, Gls per 90, G per sh, PrgC, Carries 1/3, SCA90, GCA90) thường là số dương và lớn, do được nhân trọng số 2.0.
- $\beta_5, \beta_{10}$ (cho SoT%, Aerial Won%) là số dương, nhưng nhỏ hơn.

**Tiền xử lý:**

- Biến đổi log và chuẩn hóa đảm bảo các feature có thang đo tương đương.
- Trọng số 2.0 cho các feature ghi bàn, sáng tạo, và tiến công nhấn mạnh vai trò cốt lõi.

## 5. Kết luận

Các feature Gls, Ast, xG per 90, Gls per 90, SoT%, G per sh, PrgC, Carries 1/3, Aerial Won%, SCA90, và GCA90 được chọn vì chúng bao quát vai trò ghi bàn, sáng tạo, và tiến công của tiền đạo. Chúng phản ánh đúng giá trị thị trường, được xử lý kỹ lưỡng để phù hợp với mô hình hồi quy tuyến tính, và cung cấp cơ sở mạnh mẽ để dự đoán ETV.