

# Phân tích các Features cho vị trí Hậu vệ (DF) 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 Hậu vệ trong bóng đá, sử dụng dữ liệu từ tệp result.csv.

# Mục lục

- 1 Tổng quan về feature của hậu vệ
- 2 Phân tích chi tiết từng feature
  - 2.1 Tắc bóng (Tkl - Tackles)
    - 2.1.1 Định nghĩa
    - 2.1.2 Lý do chọn
    - 2.1.3 Cách xử lý trong mã
    - 2.1.4 Ví dụ minh họa
  - 2.2 Tắc bóng giành lại quyền kiểm soát bóng (TklW - Tackles Won)
    - 2.2.1 Định nghĩa
    - 2.2.2 Lý do chọn
    - 2.2.3 Cách xử lý trong mã
    - 2.2.4 Ví dụ minh họa
  - 2.3 Đánh chặn (Int - Interceptions)
    - 2.3.1 Định nghĩa
    - 2.3.2 Lý do chọn
    - 2.3.3 Cách xử lý trong mã
    - 2.3.4 Ví dụ minh họa
  - 2.4 Chặn bóng (Blocks)
    - 2.4.1 Định nghĩa
    - 2.4.2 Lý do chọn
    - 2.4.3 Cách xử lý trong mã
    - 2.4.4 Ví dụ minh họa
  - 2.5 Số lần thắng trong tranh chấp bóng bổng (Aerl Won - Aerial Duels Won)
    - 2.5.1 Định nghĩa
    - 2.5.2 Lý do chọn
    - 2.5.3 Cách xử lý trong mã
    - 2.5.4 Ví dụ minh họa
  - 2.6 Tỷ lệ thắng tranh chấp trên không (Aerl Won% - Aerial Duels Won Percentage)
    - 2.6.1 Định nghĩa
    - 2.6.2 Lý do chọn
    - 2.6.3 Cách xử lý trong mã
    - 2.6.4 Ví dụ minh họa
  - 2.7 Số lần thu hồi bóng thành công (Recov - Recoveries)
    - 2.7.1 Định nghĩa
    - 2.7.2 Lý do chọn

- 2.7.3 Cách xử lý trong mã
    - 2.7.4 Ví dụ minh họa
  - 2.8 Số đường chuyền chính xác (Cmp - Completed Passes)
    - 2.8.1 Định nghĩa
    - 2.8.2 Lý do chọn
    - 2.8.3 Cách xử lý trong mã
    - 2.8.4 Ví dụ minh họa
  - 2.9 Tỷ lệ chuyền chính xác (Cmp% - Pass Completion Percentage)
    - 2.9.1 Định nghĩa
    - 2.9.2 Lý do chọn
    - 2.9.3 Cách xử lý trong mã
    - 2.9.4 Ví dụ minh họa
  - 2.10 Số đường chuyền tịnh tiến (PrgP - Progressive Passes)
    - 2.10.1 Định nghĩa
    - 2.10.2 Lý do chọn
    - 2.10.3 Cách xử lý trong mã
    - 2.10.4 Ví dụ minh họa
  - 2.11 Tỷ lệ chuyền dài thành công (LongCmp% - Long Pass Completion Percentage)
    - 2.11.1 Định nghĩa
    - 2.11.2 Lý do chọn
    - 2.11.3 Cách xử lý trong mã
    - 2.11.4 Ví dụ minh họa
  - 2.12 Số lần mất bóng do đối thủ đoạt được (Dis - Dispossessed)
    - 2.12.1 Định nghĩa
    - 2.12.2 Lý do chọn
    - 2.12.3 Cách xử lý trong mã
    - 2.12.4 Ví dụ minh họa
  - 2.13 Số lần xử lý bóng lỗi (Mis - Miscontrols)
    - 2.13.1 Định nghĩa
    - 2.13.2 Lý do chọn
    - 2.13.3 Cách xử lý trong mã
    - 2.13.4 Ví dụ minh họa
- 3 Tại sao các feature này được chọn cùng nhau?
  - 4 Cách các feature tương tác với mô hình
  - 5 Kết luận

## 1. Tổng quan về feature của hậu vệ

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

- Khả năng phòng ngự: Ngăn chặn đối phương ghi bàn thông qua tắc bóng, đánh chặn, chặn bóng, và thu hồi bóng.
- Khả năng phát triển bóng: Tham gia xây dựng lối chơi thông qua chuyền bóng chính xác, đặc biệt là các đường chuyền tịnh tiến hoặc chuyền dài.
- Kỹ năng tranh chấp: Đặc biệt là tranh chấp bóng bổng, quan trọng với hậu vệ trung tâm.
- Độ tin cậy kỹ thuật: Giảm thiểu lỗi mất bóng hoặc xử lý bóng kém.

Các feature được chia thành:

- Feature liên quan tới phòng ngự: Tắc bóng (Tkl), Tắc bóng giành lại quyền kiểm soát bóng (TklW), Đánh chặn (Int), Chặn bóng (Blocks), Số lần thắng trong tranh chấp bóng bổng (Aerl Won), Tỷ lệ thắng tranh chấp trên không (Aerl Won%), Số lần thu hồi bóng thành công (Recov).
- Feature liên quan tới phát triển bóng: Số đường chuyền chính xác (Cmp), Tỷ lệ chuyền chính xác (Cmp%), Số đường chuyền tịnh tiến (PrgP), Tỷ lệ chuyền dài thành công (LongCmp%).
- Feature liên quan tới kỹ thuật tiêu cực: Số lần mất bóng do đối thủ đoạt được (Dis), Số lần xử lý bóng lỗi (Mis).

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

- Liên quan đến vai trò hậu vệ: Hậu vệ cần phòng ngự hiệu quả, phát triển bóng tốt, và giảm thiểu sai lầm.
- Tầm quan trọng trong thị trường chuyển nhượng: Các chỉ số này là yếu tố chính mà các CLB xem xét khi định giá hậu vệ.
- 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. Tắc bóng (Tkl - Tackles)

#### 2.1.1. Định nghĩa

Tắc bóng (Tkl) là số lần hậu vệ thực hiện hành động tắc bóng để ngăn chặn đối thủ, bất kể có giành được bóng hay không. Ví dụ: Một hậu vệ thực hiện 50 lần tắc bóng trong mùa giải.

#### 2.1.2. Lý do chọn

- Cốt lõi của phòng ngự: Tắc bóng là kỹ năng phòng ngự cơ bản, thể hiện khả năng ngăn chặn đối thủ trong các tình huống 1v1 hoặc cướp bóng từ chân đối phương. Hậu vệ giỏi tắc bóng (như Aaron Wan-Bissaka) thường được đánh giá cao.
- Tương quan với ETV: Hậu vệ có số lần tắc bóng cao thường có ETV cao hơn, vì họ góp phần lớn vào việc ngăn chặn các đợt tấn công của đối phương.
- Phổ biến và sẵn có: Tkl là chỉ số tiêu chuẩn, được ghi nhận bởi các nền tảng như Opta, WhoScored, và có trong result.csv.
- Quan trọng với mọi loại hậu vệ: Dù là hậu vệ cánh hay trung tâm, khả năng tắc bóng đều quan trọng.

#### 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).
- Biến đổi log: Áp dụng `np.log1p` để giảm độ lệch, vì Tkl có phân phối lệch.
- Tăng trọng số: Là `important_features`, được nhân với 2.0 để nhấn mạnh vai trò cốt lõi.
- Chuẩn hóa: Chuẩn hóa bằng `StandardScaler` để đưa về thang đo tương đương.

#### 2.1.4. Ví dụ minh họa

Hậu vệ A có  $Tkl = 60$  và hậu vệ B có  $Tkl = 30$ . Hậu vệ A có khả năng được định giá cao hơn, vì anh ta phòng ngự tích cực hơn.

## 2.2. Tắc bóng giành lại quyền kiểm soát bóng (TklW - Tackles Won)

### 2.2.1. Định nghĩa

Tắc bóng giành lại quyền kiểm soát bóng (TklW) là số lần tắc bóng thành công, tức là hậu vệ không chỉ thực hiện tắc bóng mà còn giành được quyền kiểm soát bóng. Ví dụ: Trong 50 lần tắc bóng, hậu vệ giành lại bóng 35 lần, thì  $TklW = 35$ .

### 2.2.2. Lý do chọn

- Đo lường hiệu quả tắc bóng: TklW bổ sung cho Tkl bằng cách tập trung vào kết quả thành công, thể hiện kỹ năng và độ chính xác.
- Tương quan với ETV: Hậu vệ có TklW cao thường được định giá cao hơn, vì họ giúp đội bóng chuyển trạng thái từ phòng ngự sang tấn công.
- Bổ sung cho Tkl: Tkl đo lường tần suất, TklW đo lường chất lượng.
- Phổ biến: TklW 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 Tkl.
- Biến đổi log: Áp dụng  $\text{np.log1p}$ .
- Tăng trọng số: Là `important_features`, được nhân với 2.0.
- Chuẩn hóa: Chuẩn hóa bằng `StandardScaler`.

### 2.2.4. Ví dụ minh họa

Hậu vệ A có  $TklW = 40$  (40/50 lần tắc bóng thành công) và hậu vệ B có  $TklW = 20$  (20/50). Hậu vệ A có khả năng được định giá cao hơn, vì anh ta hiệu quả hơn trong việc giành bóng.

## 2.3. Đánh chặn (Int - Interceptions)

### 2.3.1. Định nghĩa

Đánh chặn (Int) là số lần hậu vệ chặn được đường chuyền của đối phương, ngăn chặn đợt tấn công trước khi nó trở nên nguy hiểm. Ví dụ: Hậu vệ thực hiện 45 lần đánh chặn trong mùa giải.

### 2.3.2. Lý do chọn

- Ngăn chặn sớm: Đánh chặn thể hiện khả năng đọc trận đấu và định vị tốt.

- Tương quan với ETV: Hậu vệ có Int cao thường có ETV cao hơn, vì họ giảm thiểu nguy cơ bàn thua.
- Phổ biến: Int là chỉ số tiêu chuẩn, có sẵn trong result.csv.

### 2.3.3. Cách xử lý trong mã

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

Hậu vệ A có  $\text{Int} = 50$  và hậu vệ B có  $\text{Int} = 20$ . Hậu vệ A có khả năng được định giá cao hơn, vì anh ta giỏi ngăn chặn sớm.

## 2.4. Chặn bóng (Blocks)

### 2.4.1. Định nghĩa

Chặn bóng (Blocks) là số lần hậu vệ chặn được cú sút, đường chuyền, hoặc pha bóng nguy hiểm của đối phương. Ví dụ: Hậu vệ chặn 30 cú sút hoặc đường chuyền trong mùa giải.

### 2.4.2. Lý do chọn

- Bảo vệ khung thành: Chặn bóng là hành động phòng ngự trực tiếp.
- Tương quan với ETV: Hậu vệ có Blocks cao thường được định giá cao hơn.
- Phổ biến: Blocks là chỉ số tiêu chuẩn, có sẵn trong result.csv.

### 2.4.3. Cách xử lý trong mã

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

#### 2.4.4. Ví dụ minh họa

Hậu vệ A có Blocks = 35 và hậu vệ B có Blocks = 15. Hậu vệ A có khả năng được định giá cao hơn, vì anh ta giỏi bảo vệ khung thành.

### 2.5. Số lần thắng trong tranh chấp bóng bổng (Aerl Won - Aerial Duels Won)

#### 2.5.1. Định nghĩa

Số lần thắng trong tranh chấp bóng bổng (Aerl Won) là số lần hậu vệ thắng trong các pha tranh chấp bóng bổng. Ví dụ: Hậu vệ thắng 80 pha tranh chấp bóng bổng trong mùa giải.

#### 2.5.2. Lý do chọn

- Quan trọng với hậu vệ trung tâm: Tranh chấp bóng bổng là kỹ năng cốt lõi trong các tình huống phòng ngự cố định.
- Tương quan với ETV: Hậu vệ có Aerl Won cao thường có ETV cao hơn.
- Phổ biến: Aerl Won 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 Tkl.
- Biến đổi log: Áp dụng np.log1p.
- Tăng trọng số: Là important\_features, được nhân với 2.0.
- Chuẩn hóa: Chuẩn hóa bằng StandardScaler.

#### 2.5.4. Ví dụ minh họa

Hậu vệ A có Aerl Won = 90 và hậu vệ B có Aerl Won = 40. Hậu vệ A có khả năng được định giá cao hơn, vì anh ta giỏi tranh chấp bóng bổng.

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

#### 2.6.1. Định nghĩa

Tỷ lệ thắng tranh chấp trên không (Aerl Won%) là tỷ lệ phần trăm các pha tranh chấp bóng bổng mà hậu vệ thắng:

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

Ví dụ: Hậu vệ thắng 80/100 pha tranh chấp, thì Aerl Won% = 80%.



### 2.6.2. Lý do chọn

- Đo lường hiệu quả tranh chấp: Aeri Won% bổ sung cho Aeri Won bằng cách tập trung vào hiệu quả.
- Tương quan với ETV: Hậu vệ có Aeri Won% cao thường được định giá cao hơn.
- Bổ sung cho Aeri Won: Aeri Won đo lường số lượng, Aeri Won% đo lường chất lượng.

### 2.6.3. Cách xử lý trong mã

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

Hậu vệ A có Aeri Won% = 85% và hậu vệ B có Aeri Won% = 60%. Hậu vệ A có khả năng được định giá cao hơn, vì anh ta hiệu quả hơn trong tranh chấp bóng bổng.

## 2.7. Số lần thu hồi bóng thành công (Recov - Recoveries)

### 2.7.1. Định nghĩa

Số lần thu hồi bóng thành công (Recov) là số lần hậu vệ thu hồi bóng lỏng hoặc bóng không có người kiểm soát. Ví dụ: Hậu vệ thu hồi bóng 200 lần trong mùa giải.

### 2.7.2. Lý do chọn

- Khả năng kiểm soát trận đấu: Thu hồi bóng thể hiện ý thức chiến thuật và định vị tốt.
- Tương quan với ETV: Hậu vệ có Recov cao thường được định giá cao hơn.
- Phổ biến: Recov là chỉ số tiêu chuẩn, có sẵn trong `result.csv`.

### 2.7.3. Cách xử lý trong mã

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

Hậu vệ A có Recov = 250 và hậu vệ B có Recov = 150. Hậu vệ A có khả năng được định giá cao hơn, vì anh ta kiểm soát trận đấu tốt hơn.

### 2.8. Số đường chuyền chính xác (Cmp - Completed Passes)

#### 2.8.1. Định nghĩa

Số đường chuyền chính xác (Cmp) là số đường chuyền mà hậu vệ thực hiện thành công. Ví dụ: Hậu vệ hoàn thành 1,000 đường chuyền trong mùa giải.

#### 2.8.2. Lý do chọn

- Khả năng phát triển bóng: Hậu vệ có Cmp cao giúp đội bóng triển khai tấn công từ tuyến dưới.
- Tương quan với ETV: Hậu vệ có Cmp cao thường được định giá cao hơn.
- Phổ biến: Cmp là chỉ số tiêu chuẩn, có sẵn trong result.csv.

#### 2.8.3. Cách xử lý trong mã

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

#### 2.8.4. Ví dụ minh họa

Hậu vệ A có Cmp = 1,200 và hậu vệ B có Cmp = 800. Hậu vệ A có khả năng được định giá cao hơn, vì anh ta tham gia nhiều vào lối chơi.

### 2.9. Tỷ lệ chuyền chính xác (Cmp% - Pass Completion Percentage)

#### 2.9.1. Định nghĩa

Tỷ lệ chuyền chính xác (Cmp%) là tỷ lệ phần trăm đường chuyền thành công:

$$\text{Cmp\%} = \frac{\text{Số đường chuyền hoàn thành}}{\text{Tổng số đường chuyền}} \times 100$$

Ví dụ: Hậu vệ hoàn thành 1,000/1,100 đường chuyền, thì Cmp% = 90.9%.

### 2.9.2. Lý do chọn

- Đo lường độ tin cậy khi chuyển bóng: Cmp% tập trung vào hiệu quả chuyển bóng.
- Tương quan với ETV: Hậu vệ có Cmp% cao thường được định giá cao hơn.
- Bổ sung cho Cmp: Cmp đo lường tần suất, Cmp% đo lường chất lượng.

### 2.9.3. Cách xử lý trong mã

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

Hậu vệ A có Cmp% = 92% và hậu vệ B có Cmp% = 80%. Hậu vệ A có khả năng được định giá cao hơn, vì anh ta đáng tin cậy hơn khi chuyển bóng.

## 2.10. Số đường chuyển tịnh tiến (PrgP - Progressive Passes)

### 2.10.1. Định nghĩa

Số đường chuyển tịnh tiến (PrgP) là số đường chuyển đưa bóng tiến gần hơn đến khung thành đối phương. Ví dụ: Hậu vệ thực hiện 150 đường chuyển tịnh tiến trong mùa giải.

### 2.10.2. Lý do chọn

- Tham gia tấn công: PrgP thể hiện khả năng khởi tạo tấn công.
- Tương quan với ETV: Hậu vệ có PrgP cao thường được định giá cao hơn.
- Phổ biến: PrgP 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 Tkl.
- Biến đổi log: Áp dụng `np.log1p`.
- Không tăng trọng số: Không nằm trong `important_features`.
- Chuẩn hóa: Chuẩn hóa bằng `StandardScaler`.

#### 2.10.4. Ví dụ minh họa

Hậu vệ A có  $\text{PrgP} = 200$  và hậu vệ B có  $\text{PrgP} = 100$ . Hậu vệ A có khả năng được định giá cao hơn, vì anh ta đóng góp nhiều vào tấn công.

#### 2.11. Tỷ lệ chuyền dài thành công (LongCmp% - Long Pass Completion Percentage)

##### 2.11.1. Định nghĩa

Tỷ lệ chuyền dài thành công (LongCmp%) là tỷ lệ phần trăm đường chuyền dài (thường trên 30 mét) thành công:

$$\text{LongCmp\%} = \frac{\text{Số đường chuyền dài hoàn thành}}{\text{Tổng số đường chuyền dài}} \times 100$$

Ví dụ: Hậu vệ hoàn thành 50/80 đường chuyền dài, thì  $\text{LongCmp\%} = 62.5\%$ .

##### 2.11.2. Lý do chọn

- Khả năng chuyền bóng chiến thuật: Chuyền dài giúp chuyển trạng thái nhanh hoặc tìm đồng đội ở xa.
- Tương quan với ETV: Hậu vệ có LongCmp% cao thường được định giá cao hơn.
- Bổ sung cho Cmp%: Cmp% đo lường tổng thể, LongCmp% tập trung vào chuyền dài.

##### 2.11.3. Cách xử lý trong mã

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

##### 2.11.4. Ví dụ minh họa

Hậu vệ A có  $\text{LongCmp\%} = 65\%$  và hậu vệ B có  $\text{LongCmp\%} = 45\%$ . Hậu vệ A có khả năng được định giá cao hơn, vì anh ta giỏi chuyền dài.

#### 2.12. Số lần mất bóng do đối thủ đoạt được (Dis - Dispossessed)

##### 2.12.1. Định nghĩa

Số lần mất bóng do đối thủ đoạt được (Dis) là số lần hậu vệ mất bóng do bị đối thủ cướp bóng. Ví dụ: Hậu vệ bị cướp bóng 20 lần trong mùa giải.

#### 2.12.2. Lý do chọn

- Đo lường sai lầm kỹ thuật: Dis phản ánh độ tin cậy của hậu vệ khi kiểm soát bóng.
- Tương quan với ETV: Hậu vệ có Dis thấp thường được định giá cao hơn.
- Phổ biến: Dis là chỉ số tiêu chuẩn, có sẵn trong result.csv.

#### 2.12.3. Cách xử lý trong mã

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

#### 2.12.4. Ví dụ minh họa

Hậu vệ A có Dis = 10 và hậu vệ B có Dis = 30. Hậu vệ A có khả năng được định giá cao hơn, vì anh ta ít bị cướp bóng.

### 2.13. Số lần xử lý bóng lỗi (Mis - Miscontrols)

#### 2.13.1. Định nghĩa

Số lần xử lý bóng lỗi (Mis) là số lần hậu vệ kiểm soát bóng không tốt, dẫn đến mất bóng mà không phải do đối thủ cướp. Ví dụ: Hậu vệ xử lý bóng lỗi 15 lần trong mùa giải.

#### 2.13.2. Lý do chọn

- Đo lường độ tin cậy kỹ thuật: Mis phản ánh kỹ năng kiểm soát bóng.
- Tương quan với ETV: Hậu vệ có Mis thấp thường có ETV cao hơn.
- Phổ biến: Mis là chỉ số tiêu chuẩn, có sẵn trong result.csv.

#### 2.13.3. Cách xử lý trong mã

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

#### 2.13.4. Ví dụ minh họa

Hậu vệ A có Mis = 10 và hậu vệ B có Mis = 25. Hậu vệ A có khả năng được định giá cao hơn, vì anh ta kiểm soát bóng tốt hơn.

### 3. Tại sao các feature này được chọn cùng nhau?

- Phản ánh toàn diện vai trò hậu vệ:
  - Phòng ngự: Tkl, TklW, Int, Blocks, Recov, Aeryl Won, Aeryl Won% đo lường khả năng ngăn chặn đối phương.
  - Phát triển bóng: Cmp, Cmp%, PrgP, LongCmp% thể hiện khả năng tham gia xây dựng lối chơi.
  - Độ tin cậy kỹ thuật: Dis, Mis đánh giá khả năng tránh sai lầm.
- Bổ sung lẫn nhau:
  - Tkl và TklW đo lường tần suất và hiệu quả tắc bóng.
  - Aeryl Won và Aeryl Won% đo lường số lượng và hiệu quả tranh chấp bóng bổng.
  - Cmp và Cmp% đo lường tần suất và hiệu quả chuyền bóng.
  - Dis và Mis cung cấp góc nhìn tiêu cực, bổ sung cho các chỉ số tích cực.
- Tương quan với thị trường chuyển nhượng: Hậu vệ toàn diện (phòng ngự tốt, chuyên bóng giỏi, ít sai lầm) như Virgil van Dijk 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:

$$ETV = \beta_0 + \beta_1 \cdot Tkl + \beta_2 \cdot TklW + \beta_3 \cdot Int + \dots + \beta_{12} \cdot Dis + \beta_{13} \cdot Mis$$

Trong đó:

- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$  (cho Tkl, TklW, Int, Blocks, Aeryl Won, Aeryl Won%, Recov) thường là số dương và lớn, do được nhân trọng số 2.0.
- $\beta_8, \beta_9, \beta_{10}, \beta_{11}$  (cho Cmp, Cmp%, PrgP, LongCmp%) là số dương, nhưng nhỏ hơn.
- $\beta_{12}, \beta_{13}$  (cho Dis, Mis) thường là số âm.

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 phòng ngự nhấn mạnh vai trò cốt lõi.

## 5. Kết luận

Các feature Tkl, TklW, Int, Blocks, Recov, Aeri Won, Aeri Won%, Cmp, Cmp%, PrgP, LongCmp%, Dis, và Mis được chọn vì chúng bao quát vai trò phòng ngự, phát triển bóng, và độ tin cậy kỹ thuật của hậu vệ. 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.