Phân tích các feature cho vị trí Thủ môn (GK) 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 Thủ môn trong bóng đá, sử dụng dữ liệu từ tệp result.csv.

Mục lục

1. Tổng quan về feature của thủ môn

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

- **Khả năng ngăn bàn thua:** Giảm thiểu số bàn thua thông qua cản phá và giữ sạch lưới.
- **Hiệu quả cản phá:** Tỷ lệ cản phá cú sút và đặc biệt là penalty, thể hiện kỹ năng phản xạ và tâm lý.
- **Hiệu suất mỗi 90 phút:** Đo lường hiệu quả phòng ngự ổn định trong thời gian thi đấu.

Các feature được chọn bao gồm:

- **Feature liên quan đến phòng ngự:** Số bàn thua trung bình mỗi 90 phút (GA90), Tỷ lệ giữ sạch lưới (CS%).
- **Feature liên quan đến cản phá:** Tỷ lệ cản phá (Save%), Tỷ lệ cản phá penalty (PK Save%).

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

- Liên quan đến vai trò thủ môn: Thủ môn cần ngăn bàn thua, giữ sạch lưới, và cản phá hiệu quả.
- Tầm quan trọng trong thị trường chuyển nhượng: Các chỉ số này là yếu tố chính mà các CLB xem xét khi định giá thủ môn.
- 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. Số bàn thua trung bình mỗi 90 phút (GA90 - Goals Against per 90)

2.1.1. Định nghĩa

Số bàn thua trung bình mỗi 90 phút (GA90) là số bàn thua trung bình mà thủ môn để lọt lưới trong 90 phút thi đấu:

$$\text{GA90} = \frac{\text{Tổng số bàn thua}}{\text{Tổng số phút thi đấu}} \times 90$$

Ví dụ: Thủ môn để lọt 30 bàn trong 2700 phút, thì GA90 = 1.0.

2.1.2. Lý do chọn

- Cốt lõi của vai trò thủ môn: GA90 đo lường khả năng ngăn bàn thua, là chỉ số chính đánh giá hiệu suất thủ môn.

- Tương quan với ETV: Thủ môn có GA90 thấp (như Alisson Becker) thường có ETV cao hơn, vì họ giảm thiểu bàn thua hiệu quả.
- Phổ biến: GA90 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 phòng ngự.
- Chuẩn hóa: Chuẩn hóa bằng `StandardScaler` để đưa về thang đo tương đương.
- Lưu ý đặc biệt: Vì GA90 thấp là tốt, giá trị có thể được đảo ngược (ví dụ: lấy nghịch đảo hoặc trừ đi giá trị tối đa) trước khi đưa vào mô hình.

2.1.4. Ví dụ minh họa

Thủ môn A có GA90 = 0.8 và thủ môn B có GA90 = 1.5. Thủ môn A có khả năng được định giá cao hơn, vì anh ta để lọt ít bàn hơn.

2.2. Tỷ lệ giữ sạch lưới (CS% - Clean Sheet Percentage)

2.2.1. Định nghĩa

Tỷ lệ giữ sạch lưới (CS%) là tỷ lệ phần trăm các trận đấu mà thủ môn không để lọt bàn:

$$\text{CS\%} = \frac{\text{Số trận giữ sạch lưới}}{\text{Tổng số trận đấu}} \times 100$$

Ví dụ: Thủ môn giữ sạch lưới trong 12/30 trận, thì CS% = 40%.

2.2.2. Lý do chọn

- Thành tích phòng ngự: CS% phản ánh khả năng giữ sạch lưới, là chỉ số quan trọng của thủ môn.
- Tương quan với ETV: Thủ môn có CS% cao thường được định giá cao hơn, vì họ mang lại sự ổn định cho hàng thủ.
- Phổ biến: CS% 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 GA90.

- 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

Thủ môn A có CS% = 45% và thủ môn B có CS% = 20%. Thủ môn A có khả năng được định giá cao hơn, vì anh ta giữ sạch lưới thường xuyên hơn.

2.3. Tỷ lệ cản phá (Save% - Save Percentage)

2.3.1. Định nghĩa

Tỷ lệ cản phá (Save%) là tỷ lệ phần trăm các cú sút trúng đích được thủ môn cản phá:

$$\text{Save\%} = \frac{\text{Số cú sút trúng đích được cản phá}}{\text{Tổng số cú sút trúng đích}} \times 100$$

Ví dụ: Thủ môn cản phá 70/100 cú sút trúng đích, thì Save% = 70%.

2.3.2. Lý do chọn

- Kỹ năng cản phá: Save% đo lường khả năng phản xạ và kỹ thuật cản phá của thủ môn.
- Tương quan với ETV: Thủ môn có Save% cao thường được định giá cao hơn, vì họ ngăn bàn thua hiệu quả.
- Phổ biến: Save% 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 GA90.
- 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

Thủ môn A có Save% = 75% và thủ môn B có Save% = 60%. Thủ môn A có khả năng được định giá cao hơn, vì anh ta cản phá hiệu quả hơn.

2.4. Tỷ lệ cản phá penalty (PK Save% - Penalty Kick Save Percentage)

2.4.1. Định nghĩa

Tỷ lệ cản phá penalty (PK Save%) là tỷ lệ phần trăm các quả penalty mà thủ môn cản phá:

$$\text{PK Save\%} = \frac{\text{Số quả penalty được cản phá}}{\text{Tổng số quả penalty đối mặt}} \times 100$$

Ví dụ: Thủ môn cản phá 3/10 quả penalty, thì PK Save% = 30%.

2.4.2. Lý do chọn

- Tâm lý và kỹ năng đặc biệt: PK Save% phản ánh khả năng cản phá trong tình huống áp lực cao.
- Tương quan với ETV: Thủ môn có PK Save% cao thường được định giá cao hơn, vì họ tạo ra sự khác biệt trong các khoảnh khắc quyết định.
- Phổ biến: PK Save% 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 GA90.
- 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

Thủ môn A có PK Save% = 40% và thủ môn B có PK Save% = 15%. Thủ môn A có khả năng được định giá cao hơn, vì anh ta cản phá penalty tốt hơn.

3. Lý do các feature này được chọn cùng nhau

- **Phản ánh toàn diện vai trò thủ môn:**
  - *Phòng ngự:* GA90 và CS% đo lường khả năng ngăn bàn thua và giữ sạch lưới.
  - *Cản phá:* Save% và PK Save% thể hiện kỹ năng phản xạ và cản phá trong các tình huống thông thường và áp lực cao.
- **Bổ sung lẫn nhau:**
  - GA90 và CS% cung cấp góc nhìn tổng thể về hiệu suất phòng ngự.

  - Save% và PK Save% tập trung vào kỹ năng cản phá ở các tình huống khác nhau (thông thường và penalty).

- **Tương quan với thị trường chuyển nhượng:** Thủ môn xuất sắc (ngăn bàn thua, giữ sạch lưới, cản phá tốt) như Alisson Becker 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:

$$\text{ETV} = \beta_0 + \beta_1 \cdot \text{GA90} + \beta_2 \cdot \text{CS\%} + \beta_3 \cdot \text{Save\%} + \beta_4 \cdot \text{PKSave\%}$$

Trong đó:

- $\beta_2, \beta_3, \beta_4$ (cho CS%, Save%, PK Save%) thường là số dương và lớn, do được nhân trọng số 2.0.
- β_1 (cho GA90) thường là số âm, vì GA90 thấp là tốt, và giá trị có thể được đảo ngược trước khi đưa vào mô hình.

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 nhấn mạnh vai trò cốt lõi của thủ môn.
- Đảo ngược GA90 (nếu cần) để đảm bảo tương quan đúng với ETV.

5. Kết luận

Các feature GA90, CS%, Save%, và PK Save% được chọn vì chúng bao quát vai trò phòng ngự và cản phá của thủ môn. 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.