# Dự đoán Giá trị Chuyển Nhượng của Cầu thủ Bóng đá

**Minh Tuan Huu Nguyen**[1,2], **Khang Trong Nguyen Huynh**[1,2], **Khanh Quoc Tran**[1,2], **Anh Tuan Gia Nguyen**[1,2]
[1] *University of Information Technology, Ho Chi Minh City, Vietnam*
[2] *Vietnam National University, Ho Chi Minh City, Vietnam*

## Abstract

Đồ án này được thực hiện nhằm tạo ra một mô hình dự đoán giá trị chuyển nhượng của các cầu thủ bóng đá. Chúng em sẽ sử dụng dữ liệu của các cầu thủ bóng đá và xây dựng một mô hình dự đoán phí chuyển nhượng dựa trên dữ liệu đó. Dữ liệu của các cầu thủ bao gồm thông tin cơ bản của (độ tuổi, chiều cao, vị trí thi đấu,...), dữ liệu chuyên môn (số bàn thắng, kiến tạo, chấn thương,...), các giải thưởng cá nhân cũng như tập thể của cầu thủ đó. Thuộc tính đích mà chúng em dự đoán là giá trị chuyển nhượng mà Transfermarkt dự đoán cho từng cầu thủ tính đến ngày 10/06/2023. Sau khi thu thập dữ liệu của tất cả cầu thủ, chúng em đã tiến hành làm sạch và tiền xử lý dữ liệu, sau đó áp dụng 3 mô hình học máy: Hồi quy Ridge, Rừng ngẫu nhiên và Tăng cường gradient để kiểm chứng bộ dữ liệu của mình. Kết quả thí nghiệm cho thấy mô hình XGBoost hoạt động tốt nhất với R-squared score trên tập dữ liệu kiểm tra là 0.597. Chúng em sẽ nâng cao chất lượng cho bộ dữ liệu và sử dụng thêm các mô hình học máy cũng như học sâu khác nhằm phát triển đề tài này trong tương lai.

## 1 Giới thiệu

Đồ án này được thực hiện nhằm tạo ra một mô hình dự đoán giá trị chuyển nhượng của các cầu thủ bóng đá. Bóng đá, còn gọi là túc cầu, là môn thể thao đồng đội chơi giữa hai đội, mỗi đội có 11 cầu thủ. Trò chơi dùng quả bóng chơi trên sân cỏ hình chữ nhật với hai khung thành ở hai đầu sân. Mục tiêu của trò chơi là ghi điểm bằng cách đưa bóng vào khung thành của đội đối địch. Ngoài thủ môn, các cầu thủ không được cố ý dùng tay hoặc cánh tay để chơi bóng. Đội chiến thắng là đội ghi được nhiều bàn hơn khi kết thúc trận đấu. Đây là một trong những môn thể thao được yêu thích nhất thế giới. Bóng đá được chơi ở cấp độ chuyên nghiệp trên khắp thế giới với hàng triệu người đến sân theo dõi các trận đấu cũng như hàng tỷ người theo dõi qua truyền hình. Theo một cuộc thăm dò do FIFA tiến hành năm 2001 [3], có trên 240 triệu người từ trên 200 quốc gia thường xuyên chơi bóng đá.

Ngoài là món ăn tinh thần của người hâm mộ, ở góc độ kinh tế, bống đá đã trở thành một thị trường được trị giá đến hơn 22 tỷ bảng Anh (năm 2015). Nền bóng đá đã có những đóng góp lớn vào kinh tế toàn cầu nói chung và nhiều quốc gia nói riêng. Các nước nỗ lực để được đăng cai World Cup, mong tận dụng cơ hội quảng bá văn hóa, du lịch và thúc đẩy kinh tế. Các giải đấu lớn như Cúp C1 Châu Âu hay Premier League (Ngoại hạng Anh) cũng mang về những khoản thu không hề nhỏ. Theo báo cáo của Deloitte năm 2018, thị trường bóng đá châu Âu có giá trị khoảng 25 tỷ bảng Anh [1].

Ngày nay, khi vấn đề kinh tế ngày càng được chú trọng, các đội bóng phải đối diện với một thách thức mới: phải nâng cao thành tích trên sân bóng nhưng đồng thời cũng phải tạo ra lợi nhuận về mặt kinh tế. Việc chuyển nhượng cầu thủ cũng trở thành vấn đề mà mọi đội bóng đều quan tâm. Các đội bóng ngày nay đều hướng tới việc tính toán chi phí cũng như nguồn thu từ việc chuyển nhượng cầu thủ nhằm mang lại hiệu quả tối đa cho mình. Việc dự đoán giá trị của một cầu thủ bây giờ trở thành một vấn đề cực kỳ quan trọng đối với các câu lạc bộ, các nhà đầu tư cùng các bên liên quan.

Dự đoán giá trị chuyển nhượng của cầu thủ là một nhiệm vụ phức tạp, đòi hỏi sự kết hợp giữa kiến thức về bóng đá và khả năng dự báo. Tuy nhiên, với sự phát triển của công nghệ và việc sử dụng các phương pháp máy học và học sâu, việc dự đoán giá trị chuyển nhượng của cầu thủ bóng đá đã trở nên khả thi và hứa hẹn hơn.

Đồ án chúng em thực hiện sẽ sử dụng dữ liệu của các cầu thủ bóng đá để xây dựng một mô hình dự đoán phí chuyển nhượng. Dữ liệu của các cầu thủ bao gồm thông tin cơ bản của (độ tuổi, chiều cao, vị trí thi đấu,...), dữ liệu chuyên môn (số bàn thắng, kiến tạo, chấn thương,...), các giải thưởng cá nhân cũng như tập thể của cầu thủ đó. Thuộc tính đích chúng em sẽ sử dụng là giá trị chuyển nhượng của

các cầu thủ dựa trên ước tính của Transfermarkt. Transfermarkt [2] là một trang web chuyên về dữ liệu và thông tin về cầu thủ, các câu lạc bộ và thị trường chuyển nhượng. Hệ thống định giá cầu thủ của Transfermarkt rất đáng tin cậy và đã được công nhận, tham khảo rộng rãi bởi giới chuyên môn và cả những người hâm mộ bóng đá trên toàn thế giới. Sau đó, chúng em sẽ sử dụng các kỹ thuật học máy nhắm đưa ra những dự đoán về phí chuyển nhượng của một cầu thủ và so sánh với thuộc tính đích để kiểm tra độ chính xác của mô hình.

Việc dự đoán tốt giá trị chuyển nhượng của các cầu thủ sẽ giúp các đội bóng có cái nhìn chi tiết, rõ ràng hơn khi đưa ra những quyết định về mặt chuyển nhượng. Kết quả dự đoán có thể được sử dụng để hỗ trợ quyết định đầu tư của các câu lạc bộ, giúp các đội bóng có thể chiêu mộ được những cầu thủ sáng giá với giá hời hay tránh việc mua phải những cầu thủ không đủ khả năng với giá đắt.

Trong phần tiếp theo (tức phần 2), chúng em sẽ giới thiệu một số công trình liên quan về chủ đề này. Trong phần 3, chúng em sẽ trình bày về quá trình bộ dữ liệu mà chúng em thu thập và quá trình thu thập bộ dữ liệu này. Quá trình làm sạch và tiền xử lý dữ liệu sẽ được trình bày ở phần 4. Các mô hình máy học mà chúng em áp dụng nằm ở phần 5 và cuối cùng, phần 6 sẽ bao gồm kết luận và hướng phát triển tiếp theo cho đồ án này.

## 2 Các Công trình Liên quan

Transfermarkt có một hệ thống định giá cầu thủ uy tín nên đã có nhiều nghiên cứu dự đoán giá trị chuyển nhượng cầu thủ dựa trên hệ thống này. Miao He và các cộng sự [4] đã sử dụng nhiều mô hình hồi quy để dự đoán giá trị chuyển nhượng của các cầu thủ dựa trên dữ liệu chuyên môn của họ trong bài nghiên cứu "Football Player's Performance and Market Value". Bài nghiên cứu này cho ra kết quả khá tốt, tuy nhiên bộ dữ liệu mà họ sử dụng không đảm bảo được độ đầy đủ: họ chỉ dự đoán giá trị chuyển nhượng của cầu thủ dựa trên dữ liệu chuyên môn của họ mà bỏ qua nhiều yếu tố khác như lịch sử chấn thương, giải thưởng cá nhân cũng như tập thể của các cầu thủ đó,...

Steffen Herm, giáo sư ngành marketing tại trường Hochschule für Technik und Wirtschaft Berlin, cùng các cộng sự [8] đã xuất bản bài nghiên cứu mang tên 'When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community'. Trong nghiên cứu này, họ đã dự đoán giá trị chuyển nhượng của cầu thủ dựa trên đánh giá của người hâm mộ. Họ đã sử dụng 5 thuộc tính để đánh giá tài năng của các cầu thủ là: độ tuổi, sự chính xác, sự thành công, sự quyết đoán, sự linh hoạt. Kết quả cho thấy chỉ có độ tuổi là thuộc tính duy nhất không có tương quan tuyến tính với giá trị chuyển nhượng của cầu thủ. Sử dụng đánh giá của người hâm mộ làm thước đo đánh giá, do đó hạn chế dễ thấy của nghiên cứu này là những đánh giá đó có thể bị sai lệch do sự thiếu kiến thức chuyên môn hoặc thiên vị của những người tham gia.

Một nghiên cứu khác có thể kể đến là 'Money Talks: Team Variables and Player Positions that Most Influence the Market Value of Professional Male Footballers in Europe', được thực hiện bởi Jose Luis Felipe cùng các đồng nghiệp [6]. Họ đã phân tích ảnh hưởng từ vị thế của đội bóng, độ tuổi và vị trí thi đấu lên giá trị chuyển nhượng của các cầu thủ. Họ đã sử dụng dữ liệu của các cầu thủ trong 5 giải vô địch quốc gia hàng đầu châu Âu (Anh, Pháp, Đức, Ý, Tây Ban Nha) trong mùa giải năm 2017 và 2018. Sử dụng mô hình hồi quy, nghiên cứu của họ đã cho thấy rằng thứ hạng của đội bóng, tháng sinh, vị trí thi đấu, độ tuổi và giải đấu đều có ảnh hưởng lên giá trị chuyển nhượng của cầu thủ. Hơn nữa, nghiên cứu này đã chỉ ra rằng những cầu thủ thi đấu ở giải Ngoại hạng Anh và Champions League, ở vị trí tiền vệ tấn công và sinh vào quý đầu tiên trong năm sẽ là những cầu thủ có giá trị chuyển nhượng cao nhất.

Majewski (2016) [7] đã nghiên cứu về tác động của nhiều yếu tố khác nhau lên giá trị của các tiền đạo để xác định những yếu tố quan trọng nhất. Trong nghiên cứu 'Identification of Factors Determining Market Value of the Most Valuable Football Players', ông đã sử dụng thông tin về 150 cầu thủ tiền đạo nổi tiếng từ trang Transfemarkt và áp dụng phương pháp GLS (generalized least squared) để tìm ra những yếu tố quan trọng. Dựa trên kết quả của ông, số lượng bàn thắng, số lượng kiến tạo, giá trị của toàn đội và điểm xếp hạng FIFA có tác động đến giá trị thị trường của các cầu thủ tiền đạo. Điểm hạn chế của nghiên cứu này là bộ dữ liệu có rất ít thực thể (150) và bị giới hạn, chỉ lấy dữ liệu ở những tiền đạo.

## 3 Bộ dữ liệu

### 3.1 Định nghĩa bài toán

Với mục tiêu dự đoán giá trị chuyển nhượng của mỗi cầu thủ dựa trên dữ liệu của cầu thủ đó, chúng em định nghĩa bài toán này như sau:

- Input: Dữ liệu của một cầu thủ (sẽ được trình bày chi tiết trong phần 3.2)
- Output: Giá trị chuyển nhượng của cầu thủ đó

### 3.2 Quá trình thu thập dữ liệu

Đầu tiên, chúng em sẽ xác định quy mô của bộ dữ liệu và các thuộc tính trong bộ dữ liệu mà chúng em sẽ thu thập. Chúng em sẽ thu thập thông tin của các cầu thủ đang chơi cho một số giải đấu hàng đầu thế giới. Cụ thể, chúng em sẽ thu thập thông tin của các cầu thủ chơi cho các giải đấu sau:

- 11 giải đấu ở châu Âu: Bao gồm giải Premier League và Championship ở nước Anh, giải Bundesliga ở Đức, giải La Liga ở Tây Ban Nha, giải Serie A ở Ý, giải Ligue 1 ở Pháp, giải Eredivisie ở Hà Lan, giải Liga NOS ở Bồ Đào Nha, giải Premier Liga ở Nga, giải Super Lig ở Thổ Nhĩ Kỳ, giải Bundesliga ở Áo.
- 4 giải đấu ở châu Mỹ: Bao gồm giải Brasileiro ở Brasil, giiar Major League Soccer ở Mỹ, giải Primera División ở Argentina và Liga MX ở Mexico.
- 1 giải đấu ở châu Phi là DStv Premiership ở Nam Phi.
- 4 giải đấu ở châu Á: Đó là giải J-League ở Nhật Bản, giải Saudi Pro League ở Ả Rập Xê Út, giải K-League 1 ở Hàn Quốc và giải A-League ở Úc.

Tiếp theo, chúng em sẽ xác định các thuộc tính trong bộ dữ liệu của mình. Nhằm dự đoán giá trị chuyển nhượng của mỗi cầu thủ, chúng em sẽ sử dụng những thông tin sau về cầu thủ đó:

- Thông tin cơ bản của cầu thủ: Chúng em sẽ thu thập thông tin về độ tuổi, chiều cao, vị trí thi đấu cũng như đội bóng hiện tại của mỗi cầu thủ.
- Dữ liệu chuyên môn: Số bàn thắng, kiến tạo, số thẻ vàng/đỏ, số phút thi đấu, số bàn thua phải nhận và số trận giữ sạch lưới (đối với thủ môn) sẽ được thu thập.
- Thành tích thi đấu: Bao gồm những giải thưởng cá nhân cũng như tập thể mà cầu thủ đó đã giành được.
- Chấn thương: Chúng em sẽ thu thập số ngày và số trận đấu mà mỗi cầu thủ đã bỏ lỡ vì chấn thương.

Khi đã xác định kích thước của bộ dữ liệu, chúng em sẽ tìm cách thu thập dữ liệu bằng các công cụ Beautiful Soup và Selenium.

Đầu tiên, chúng em sẽ thu thập danh sách các cầu thủ trong bộ dữ liệu. Với các giải đấu đã được xác định trên, chúng em sẽ thu thập đường dẫn đến trang web của mỗi giải đấu trên trang Transfermarkt. Link của từng giải đấu chứa thông tin về tên cũng như đường dẫn đến trang web của mọi đội bóng trong giải đấu đó. Trong link của từng đội bóng lại có thông tin về tên và link của từng cầu thủ tổng đội bóng đó. Từ link của các giải đấu, chúng em có thể thu thập link các đội bóng, và từ đó thu thập được link của các cầu thủ trong giải đấu đó. Tổng số lượng cầu thủ được thu thập dữ liệu trong đồ án này của chúng em là 10754.

Trong đường dẫn của mỗi cầu thủ sẽ có tên và một mã số (id) riêng mà Transfermarkt gán cho cầu thủ đó, và với thông tin này chúng em sẽ có thể thu thập các thuộc tính của từng cầu thủ. Ví dụ, với cầu thủ có đường dẫn là `www.transfermarkt.co.uk/lionel-messi/profil/spieler/28003`, ta có thể hiểu rằng cầu thủ này có tên là lionel-messi và id là 28003. Khi đó ta có thể thu thập các thông tin của cầu thủ này, cụ thể:

- Thông tin cơ bản của cầu thủ: có thể thu thập trực tiếp từ link của cầu thủ đó.
- Dữ liệu chuyên môn: Khi thay "profil" bằng "leistungsdaten" trong link của cầu thủ có thể thu thập được dữ liệu này. Ví dụ: `www.transfermarkt.co.uk/lionel-messi/leistungsdaten/spieler/28003`
- Thành tích thi đấu: Có thể thu thập bằng cách thay "profil" bằng "erfolge". Ví dụ: `www.transfermarkt.co.uk/lionel-messi/erfolge/spieler/28003`
- Chấn thương: Khi thay "profil" bằng "verletzungen" ta có thể thu thập được dữ liệu chấn thương của cầu thủ này. Ví dụ: `www.transfermarkt.co.uk/lionel-messi/verletzungen/spieler/28003`

Lúc này, ta đã có thể thu thập được thông tin cần thiết của từng cầu thủ.

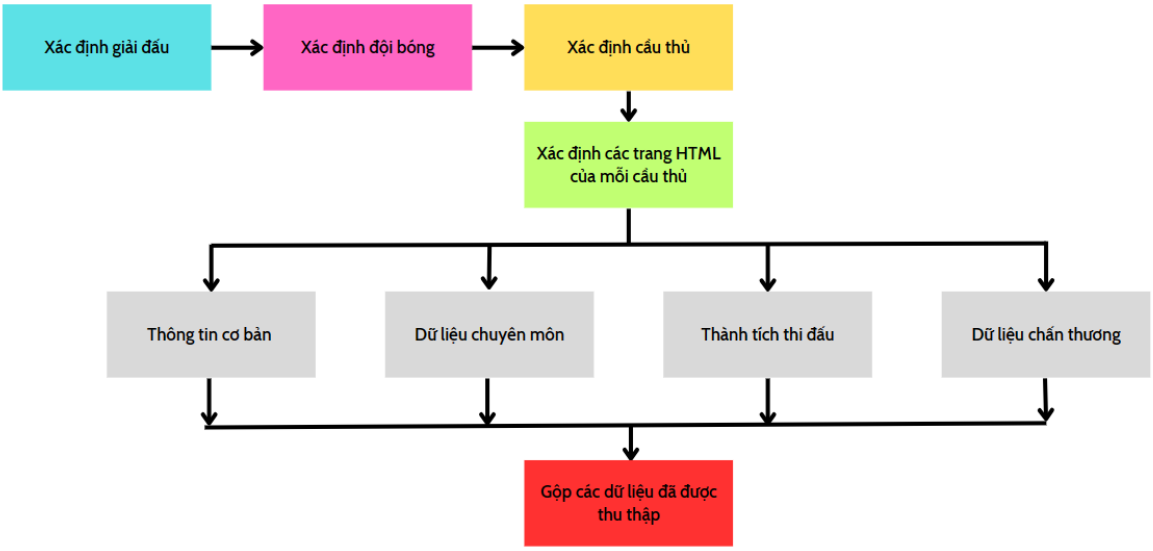


Figure 1: Quá trình thu thập dữ liệu

### 3.3 Gán nhãn dữ liệu

Thuộc tính đích mà chúng em sử dụng cho đồ án này là giá trị chuyển nhượng mà Transfermarkt dự đoán cho từng cầu thủ. Thông tin này có thể được thu thập trên link của từng cầu thủ đã được đề cập ở trên.

### 3.4 Quá trình tiền xử lý dữ liệu

Tình trạng trong bộ dữ liệu mà chúng em đã thu thập và cách chúng em xử lý như sau:

- Loại bỏ dữ liệu nhiễu: Trong bộ dữ liệu mà chúng em thu thập có những thông tin không cần thiết như số thẻ vàng, thẻ đỏ, vốn không ảnh hưởng nhiều đến giá trị chuyển nhượng của các cầu thủ. Chúng em sẽ tiến hành loại bỏ những cột không cần thiết này khỏi bộ dữ liệu của mình.

- Chuẩn hóa giá trị: Các giá trị trong bộ dữ liệu này đều được thu thập từ HTML của các trang web nên giá trị ban đầu của chúng là chuỗi (string). Chúng em sẽ xử lý tất cả giá trị này và đưa chúng về kiểu số (numeric) nhằm phục vụ tốt nhất cho các mô hình học máy.

- Bộ dữ liệu này có các giá trị bị bỏ trống (null). Khi đó, tùy thuộc vào tính chất của dữ liệu, chúng em sẽ thay giá trị bị bỏ trống này bằng giá trị 0, hoặc bằng giá trị trung bình của các giá trị không bị bỏ trống.

### 3.5 Đánh giá bộ dữ liệu

Chúng em sẽ đánh giá bộ dữ liệu này dựa trên 6 tiêu chí được liệt kê trong cuốn sách "Data Mining – Concept and Techniques 3rd edition" bởi J. Han, J. Pei và M. Kamber [5]. Cụ thể, 6 tiêu chí đó là: sự chính xác (accuracy), sự đầy đủ (completeness), sự nhất quán (consistency), sự kịp thời (timeliness), sự tin cậy (believability) và sự dễ hiểu (interpretability).

Chúng em thu thập bộ dữ liệu này dựa trên nguồn dữ liệu thực tế và được cập nhật liên tục bởi các tổ chức chuyên về dữ liệu thể thao nổi tiếng, uy tín trên thế giới. Chúng em đã thu thập dữ liệu của các cầu thủ đang thi đấu ở các giải đấu hàng đầu thế giới. Ở các giải đấu này, khía cạnh kinh tế của bóng đá được các câu lạc bộ rất chú trọng, vì thế giá trị chuyển nhượng của mỗi cầu thủ đều được quan tâm, phân tích rất kĩ. Điều này có thể chứng minh rằng bộ dữ liệu của chúng em đảm bảo tính chất chính xác, tin cậy và liên tục của dữ liệu.

Mỗi cầu thủ trong bộ dữ liệu của chúng em có một mã số riêng để phân biệt, và không có cầu thủ nào trùng mã số với nhau, cho thấy bộ dữ liệu này đảm bảo tính nhất quán của dữ liệu. Về sự đầy đủ của dữ liệu, tuy vẫn còn những giá trị trống nhưng số lượng lại rất ít, nên vẫn có thể nói đây là một bộ dữ liệu đầy đủ. Hơn nữa, các thuộc tính trong bộ dữ liệu này của chúng em đảm bảo tính chuyên môn nhưng vẫn dễ hiểu đối với những người xem bóng đá phổ thông. Có thể khẳng định rằng bộ dữ

liệu chúng em thu thập là một bộ dữ liệu tốt.

Ngoài ra, các cầu thủ trong bộ dữ liệu của chúng em đến từ các quốc gia khác nhau, thi đấu ở nhiều giải đấu khác nhau ở các châu lục khác nhau. Và với số lượng cầu thủ là trên 10000, có thể nói đây là một bộ dữ liệu đa dạng, phong phú.

Tuy nhiên, vì bộ dữ liệu mà chúng em xây dựng được lấy từ thực tế, nên nhãn trong bài toán này có giá trị chênh lệch: Có 25% số cầu thủ có giá trị chuyển nhượng được Transfermarkt dự đoán là dưới 300,000 Euro; 25% số cầu thủ được dự đoán giá trị chuyển nhượng trên 300,000 Euro và dưới 800,000 Euro; 25% số cầu thủ có giá trị chuyển nhượng được dự đoán trên 800,000 Euro và dưới 2,500,000 Euro; 25% còn lại có giá trị chuyển nhượng được dự đoán là trên 2,500,000 Euro. Cầu thủ được định giá chuyển nhượng cao nhất lên tới 180,000,000 Euro, trong khi cầu thủ được định giá thấp nhất là 0 Euro.

Lý do cho sự chênh lệch này là vì sự chênh lệch kinh tế giữa các đội bóng ở các quốc gia khác nhau. Biểu đồ cột sau thể hiện phân bố giá trị của các giá trị trong giá trị chuyển nhượng được Transfermarkt dự đoán cho các cầu thủ:

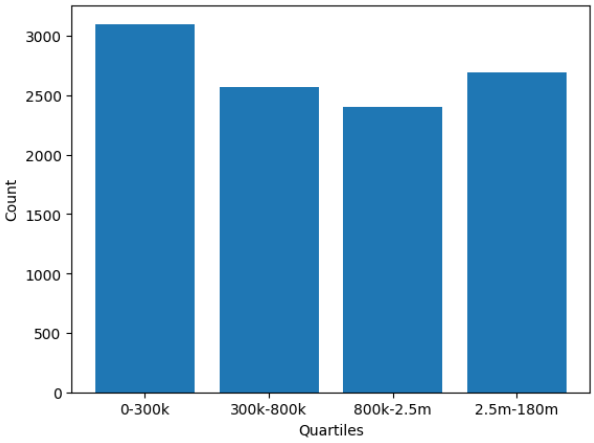


Figure 2: Bảng thể hiện tứ phân vị của thuộc tính đích (Đơn vị: Euro)

Một hạn chế nữa mà ta có thể nhận thấy là số thuộc tính vẫn chưa nhiều. Vì bị hạn chế về mặt thời gian nên chúng em mới chỉ thu thập được một số thuộc tính cơ bản thường dùng để đánh giá cầu thủ bóng đá. Đây là một hướng đi mà chúng em sẽ nhắm tới nhằm phát triển bộ dữ liệu này trong tương lai.

### 3.6 Phân chia dữ liệu

Chúng em chia tập dữ liệu của mình thành hai tập con là tập train và tập test với tỷ lệ là 8:2. Mục đích của việc phân chia này là để đánh giá hiệu suất của mô hình trên dữ liệu mới, không được sử dụng trong quá trình huấn luyện. Điều này sẽ giúp việc kiểm tra hiệu quả của mô hình trở nên tin cậy hơn.

## 4 Mô hình học máy áp dụng

Sau khi có được bộ dữ liệu đầy đủ và đã được tiền xử lý, chúng em sẽ tiến thành hành thử nghiệm các phương pháp bên dưới nhằm đánh giá hiệu quả của chúng trên bộ dữ liệu của mình. Quá trình này bao gồm việc chạy các thí nghiệm và đánh giá kết quả để có thể so sánh và lựa chọn phương pháp tốt nhất. Qua quá trình này, chúng em mong muốn có thể tìm ra được phương pháp tối ưu để giải quyết bài toán mà chúng em đang quan tâm, từ đó có thể đưa ra các giải pháp hiệu quả cũng như ứng dụng được vào thực tế.

Bài toán chúng em đang cố gắng giải quyết là một bài toán hồi quy với nhãn là giá trị liên tục. Từ đó, chúng em sẽ chọn những mô hình sau để giải quyết bài toán và so sánh hiệu suất của chúng với nhau:

### 4.1 Hồi quy Ridge

Bài toán hồi quy tuyến tính có những hạn chế như rất nhạy cảm với nhiễu. Khi hệ thống gặp nhiễu gây ra sự không tuyến tính giữa đầu vào và đầu ra. Điều này sẽ ảnh hưởng đến kết quả dự báo khi dùng phương pháp hồi quy tuyến tính. Tiếp theo, hồi quy tuyến tính không biểu diễn được các mô hình có độ phức tạp lớn như mô hình bao gồm nhiều đầu vào. Khi mô hình trở nên phức tạp dẫn đến hiện tượng quá khớp.

Mô hình hồi quy Ridge là một biến thể của hồi quy tuyến tính, nhằm giảm thiểu hiện tượng quá khớp bằng cách áp dụng phạt (penalty) vào các hệ số hồi quy. Phương pháp này sử dụng thêm một thành phần độ lớn của các hệ số trong quá trình tối thiểu hóa hàm mất mát. Điều này có tác dụng làm giảm độ lớn của các hệ số và ổn định mô hình.

Nghiệm của mô hình hồi quy Ridge được biểu diễn như sau:

$$\theta = (X^T X + \lambda I)^{-1} X^T y$$

Trong đó, ma trận $X$ chứa các biến ngõ vào và có ma trận chuyển vị là $X^T$. Ma trận đơn vị $I$ được thiết kế với kích thước phù hợp. Dễ dàng nhận ra nghiệm của mô hình hồi quy Ridge bao gồm thêm một lượng $\lambda I$ khác với hồi quy tuyến tính. được gọi là tham số điều chỉnh độ phức tạp của mô hình. Tham số này được dùng để kiểm soát độ lớn của

thành phần điều chỉnh tác động lên hàm mất mát. Hồi quy Ridge áp dụng kỹ thuật kiểm soát thành phần hiệu chỉnh này để giúp giải quyết các hiện tượng đa cộng tuyến hoàn hảo và quá khớp của dữ liệu. Trong trường hợp $\lambda$ rất lớn, hầu như tất cả các tham số mô hình suy giảm về 0 và được gọi là hiện tượng phù hợp dưới mức (underfitting). Khi $\lambda$ rất nhỏ, hồi quy Ridge trở thành hồi quy tuyến tính thông thường. Điều này dẫn đến hiện tượng quá khớp (overfitting). Mô hình huấn luyện tốt nhất là mô hình cần giảm thiểu hiện tượng quá khớp, khi đó giá trị tham số $\lambda$ tối ưu.

Mô hình hồi quy Ridge được áp dụng rộng rãi trong các bài toán có số lượng biến đầu vào lớn và có sự tương quan cao giữa các biến. Nó cung cấp một cách để tăng tính ổn định và khả năng dự báo của mô hình. Đồng thời, phương pháp này cũng giúp chúng ta xác định độ quan trọng của các biến đầu vào và loại bỏ những biến không cần thiết.

Trong bài toán này, chúng em sẽ sử dụng mô hình hồi quy Ridge đa thức. Chúng em sẽ biến đổi dữ liệu ban đầu thành các đặc trưng đa thức bậc hai (với giả định là giá trị chuyển nhượng phụ thuộc của các cầu thủ phụ thuộc vào các thuộc tính theo một hàm số bậc 2), sau đó áp dụng mô hình hồi quy Ridge lên dữ liệu đã được biến đổi.

### 4.2 Random Forest

Random Forest (Rừng Ngẫu Nhiên) là một thuật toán học máy mạnh mẽ và phổ biến được sử dụng trong việc xây dựng mô hình dự đoán và phân loại. Thuật toán này được xây dựng dựa trên ý tưởng kết hợp của nhiều cây quyết định (Decision Trees) để tạo thành một "rừng" các cây quyết định.

Mô hình Rừng Ngẫu Nhiên hoạt động bằng cách tạo ra một tập hợp các cây quyết định độc lập. Mỗi cây quyết định được xây dựng bằng cách chọn ngẫu nhiên một phần tử từ tập dữ liệu huấn luyện, cùng với việc chọn ngẫu nhiên một phần các biến đầu vào. Quá trình này được lặp lại nhiều lần để tạo ra các cây quyết định khác nhau.

Khi có một dữ liệu mới cần dự đoán, mỗi cây quyết định trong Rừng Ngẫu Nhiên sẽ đưa ra một dự đoán riêng. Cuối cùng, dự đoán của Rừng Ngẫu Nhiên sẽ là kết quả được đưa ra dựa trên sự biểu quyết (voting) hoặc trung bình của các dự đoán từ các cây quyết định.

Mô hình Rừng Ngẫu Nhiên có nhiều ưu điểm. Đầu tiên, nó có khả năng xử lý cả dữ liệu đầu vào có tính tương quan cao và dữ liệu có nhiễu. Thứ hai, nó có khả năng xác định độ quan trọng của các biến đầu vào, giúp chúng ta hiểu rõ hơn về tác động của từng biến đến kết quả dự đoán. Cuối cùng, mô hình này thường cho kết quả dự đoán chính xác và ổn định.

### 4.3 XGBoost

XGBoost (eXtreme Gradient Boosting) là một thuật toán học máy dựa trên ý tưởng gradient boosting. Gradient boosting là phương pháp kết hợp nhiều mô hình yếu để tạo ra một mô hình dự đoán mạnh mẽ hơn.

XGBoost là một thuật toán gradient boosting tối ưu hàm mục tiêu bằng cách thêm tuần tự các mô hình yếu vào tổ hợp. Hàm mục tiêu bao gồm hàm mất mát để đo lường sai số giữa các giá trị dự đoán và giá trị thực tế, cùng với thuật ngữ chính quy hóa để điều chỉnh độ phức tạp của mô hình.

Thuật toán bắt đầu với một mô hình yếu ban đầu, thường là một cây quyết định đơn giản với độ sâu hạn chế. Sau đó, nó tính toán gradient của hàm mất mát đối với các dự đoán của tổ hợp hiện tại. Gradient này cho biết hướng cần điều chỉnh các dự đoán để giảm thiểu sai số.

Trong mỗi vòng lặp, XGBoost huấn luyện một mô hình yếu mới dựa trên các gradient âm của hàm mất mát (còn gọi là "residuals") bằng quá trình gradient boosting. Mô hình yếu được huấn luyện để dự đoán các gradient âm, tức là học cách sửa các sai lầm của các mô hình trước đó.

Để ngăn chặn overfitting và cải thiện khả năng tổng quát hóa, XGBoost áp dụng các kỹ thuật chính quy hóa. Nó bao gồm cả thuật ngữ chính quy hóa L1 và L2 trong hàm mục tiêu, giảm thiểu độ phức tạp của mô hình bằng cách thêm tổng các trọng số của mô hình. Điều này giúp ngăn chặn mô hình trở nên quá phức tạp và overfitting dữ liệu huấn luyện.

Ngoài ra, XGBoost tích hợp các kỹ thuật tối ưu tiên tiến để cải thiện hiệu suất. Nó sử dụng sự kết hợp của việc xây dựng cây đồng thời, tối ưu bộ nhớ và mẫu truy cập cache để xử lý hiệu quả các tập dữ liệu quy mô lớn và giảm thời gian huấn luyện.

## 5 Kết quả thí nghiệm

Chúng em đã thử nghiệm các mô hình đã nêu với các tham số khác nhau và dùng chiến lược GridSearchCV kết hợp với cross-validation nhằm tìm ra bộ siêu tham số tốt nhất cho bài toán. Sau đó, chúng em sẽ dùng RMSE và R-squared score làm công cụ đánh giá cho các mô hình mà chúng em áp dụng. Kết quả thí nghiệm được ghi lại trong bảng 1:

| Mô hình | RMSE | Test score |
|---|---|---|
| Ridge | 6585778 | 0.52 |
| Random Forest | 6254486 | 0.567 |
| XGBoost | 6040649 | 0.597 |

Table 1: Kết quả thí nghiệm

# 6 Kết luận và hướng phát triển

## 6.1 Kết luận

Trong đồ án này, chúng em đã xây dựng một bộ dữ liệu nhằm phục vụ cho việc dự đoán giá trị chuyển nhượng của các cầu thủ bóng đá. Quá trình thu thập dữ liệu đã được chúng em theo dõi sát sao, cẩn thận nhằm mang lại bộ dữ liệu có chất lượng tốt nhất. Bộ dữ liệu cuối cùng mà chúng em công bố gồm 10754 dòng tương ứng với 10754 cầu thủ khác nhau trên thế giới và 20 cột tương ứng với thông tin cơ bản và dữ liệu chuyên môn của các cầu thủ.

Vì hạn chế về mặt thời gian nên chúng em chưa thể nghiên cứu, thêm vào các thuộc tính khác cũng như chưa tìm ra các phương án tối ưu hiệu suất nhằm tối đa hóa hiệu quả của các mô hình học máy.

## 6.2 Hướng phát triển

Như đã đề cập ở phần đánh giá bộ dữ liệu, chúng em sẽ thu thập thêm các thuộc tính khác của cầu thủ nhằm nâng cao chất lượng của bộ dữ liệu này trong tương lai. Một số thuộc tính mà chúng em dự định sẽ thu thập thêm là:

- Dữ liệu chuyên sâu: Hiện nay bộ dữ liệu của chúng em mới chỉ có khoảng 10 thuộc tính thể hiện chuyên môn của cầu thủ, đó đều là các thuộc tính cơ bản như số bàn thắng, kiến tạo, số thẻ phạt, số phút thi đấu,... Chúng em dự định sẽ thu thập thêm dữ liệu từ trang web `fbref.com`, khi đó mỗi cầu thủ sẽ có hơn 100 thuộc tính về chuyên môn.

- Dữ liệu cảm xúc (sentiment data) của người hâm mộ đối với mỗi cầu thủ: Chúng em đã thử thu thập dữ liệu này trên trang web `twitter.com`, tuy nhiên việc này yêu cầu một kinh phí rất lớn để có thể thực hiện do số lượng cầu thủ chúng em thu thập là khá lớn. Trong tương lai, kiểu dữ liệu này có thể được thu thập để nâng cao hiệu suất cho các mô hình học máy.

Ngoài ra, chúng em sẽ nghiên cứu thêm các phương pháp học máy hoặc học sâu khác nhằm nâng cao hiệu quả cho bài toán mà chúng em đang nghiên cứu.

# References


[1] cafebiz.vn/giai-ngoai-hang-manh-ghep-khong-nho-cua-nen-kinh-te-anh-nhung-cung-chang-thieu-san-20190531165049583.chn.

[2] Transfermarkt.

[3] FIFA. 2000. Fifa survey: approximately 250 million footballers worldwide.

[4] Miao He, Ricardo Cachucho, and Arno Knobbe. 2012. Football player's performance and market value.

[5] J. Pei J. Han and M. Kamber. 2012. *Data Mining – Concept and Techniques 3rd edition*.

[6] Pablo Burillo Luis Eduardo de la Riva Javier Sanchez Jose Luis Felipe, Alvaro Fernandez-Luna and Jorge Garcia-Unanue. 2020. Money talks: Team variables and player positions that most influence the market value of professional male footballers in europe.

[7] Sebastian Majewski. 2016. 'identification of factors determining market value of the most valuable football players.

[8] Henning Kreis Steffen Herm, Hans-Markus Callsen-Bracker. 2014. When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community.


| STT | Thành viên | Công việc | Mức độ hoàn thành |
|---|---|---|---|
| 1 | Nguyễn Hữu Tuấn Minh | Lên ý tưởng, thu thập dữ liệu, thiết kế bộ nhãn, làm sạch và tiền xử lý dữ liệu, đánh giá bộ dữ liệu, viết báo cáo | Đã hoàn thành 100% |
| 2 | Huỳnh Nguyễn Trọng Khang | Thu thập dữ liệu, thiết kế bộ nhãn, viết báo cáo, đánh giá bộ dữ liệu, chuẩn bị slide thuyết trình | Đã hoàn thành 100% |

Figure 3: Bảng phân công công việc