

BỘ GIÁO DỤC VÀ ĐÀO TẠO BỘ NÔNG NGHIỆP VÀ PTNT
TRƯỜNG ĐẠI HỌC THỦY LỢI



NGUYỄN THÀNH TRUNG

**ỨNG DỤNG CÁC MÔ HÌNH HỌC MÁY DỰ ĐOÁN GIÁ TRỊ
CHUYỂN NHƯỢNG CỦA CÀU THỦ BÓNG ĐÁ CHUYÊN
NGHIỆP**

ĐỒ ÁN TỐT NGHIỆP

HÀ NỘI, NĂM 2024

**BỘ GIÁO DỤC VÀ ĐÀO TẠO BỘ NÔNG NGHIỆP VÀ PTNT
TRƯỜNG ĐẠI HỌC THỦY LỢI**

NGUYỄN THÀNH TRUNG

**ỨNG DỤNG CÁC MÔ HÌNH HỌC MÁY DỰ ĐOÁN GIÁ
TRỊ CHUYỂN NHƯỢNG CỦA CẦU THỦ BÓNG ĐÁ
CHUYÊN NGHIỆP**

Ngành: Công nghệ thông tin

Mã số: 7480201

NGƯỜI HƯỚNG DẪN: ThS. Nguyễn Đức Hiếu

HÀ NỘI, NĂM 2024



CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

Độc lập - Tự do - Hạnh phúc



NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP

Họ tên sinh viên: Nguyễn Thành Trung

Hệ đào tạo: Đại học chính quy

Lớp: 61TH6

Ngành: Công nghệ thông tin

Khoa: Công nghệ thông tin

1 - TÊN ĐỀ TÀI:

ỨNG DỤNG CÁC MÔ HÌNH HỌC MÁY DỰ ĐOÁN GIÁ TRỊ CHUYÊN NHƯỢNG CỦA CÁC CẦU THỦ BÓNG ĐÁ CHUYÊN NGHIỆP.

2 - CÁC TÀI LIỆU

- [1] Al-Asadi, M. A. and S. Tasdemir, Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques. IEEE Access, 10, 22631–22645. <https://doi.org/10.1109/ACCESS.2022.3154767>, 2022.
- [2] Kirschstein and Liebscher, Assessing the market values of soccer players—a robust analysis of data from German 1. and 2. Bundesliga. Journal of Applied Statistics, 46(7), 1336–1349. <https://doi.org/10.1080/02664763.2018.1540689>, 2019.
- [3] Yiğit, A. T., Samak, B. and T. Kaya, "Football Player Value Assessment Using Machine Learning Techniques. In S. and C. O. S. and O. B. and T. A. C. and S. I. U. Kahraman Cengiz and Cebi (Ed.), Intelligent and Fuzzy Techniques in Big Data Analytics and Decision Making (pp. 289–297).," 2020. [Online]. Available: Springer International Publishing, <https://link-springer-com.tilburguniversity.idm.oclc.org/chapter/10.1007/978-3-030->.
- [4] Behravan, I., & Razavi and S. M., A novel machine learning method for estimating football players' value in the transfer market. Soft Computing, 25(3), 2499–2511. <https://doi.org/10.1007/s00500-020-05319-3>, 2021.
- [5] Barbuscak and L., What Makes a Soccer Player Expensive? Analyzing the Transfer Activity of the Richest Soccer. In Augsburg Honors Review (Vol. 11). https://idun.augsburg.edu/honors_review Available at: https://idun.augsburg.edu/honors_review/vol11/iss1/5, 2018.
- [6] Poli, R., R. Besson, Ravenel and L., Econometric Approach to Assessing the Transfer Fees and Values of Professional Football Players. Economies, 10(1). <https://doi.org/10.3390/economies10010004>, 2022.

3 - NỘI DUNG CÁC PHẦN THUYẾT MINH VÀ TÍNH TOÁN:

Nội dung các phần	Tỷ lệ
Chương 1 - Giới thiệu bài toán, giới thiệu tổng quan	20%
Chương 2 - Các công cụ và kỹ thuật sử dụng trong bài toán	20%
Chương 3 - Ứng dụng phương pháp và xây dựng mô hình	35%
Chương 4 - Kết quả và đánh giá mô hình	25%

4. GIÁO VIÊN HƯỚNG DẪN TỪNG PHẦN

Phần	Họ tên giáo viên hướng dẫn
Chương 1 - Giới thiệu bài toán, giới thiệu tổng quan	Ths. Nguyễn Đắc Hiếu
Chương 2 - Các công cụ và kỹ thuật sử dụng trong bài toán	
Chương 3 - Ứng dụng phương pháp và xây dựng mô hình	
Chương 4 - Kết quả và đánh giá mô hình	

5. NGÀY GIAO NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP

Ngày tháng năm 20..

Trưởng Bộ môn
(Ký và ghi rõ Họ tên)

Giáo viên hướng dẫn chính
(Ký và ghi rõ Họ tên)

Nhiệm vụ Đồ án tốt nghiệp đã được Hội đồng thi tốt nghiệp của Khoa thông qua

Ngày. . . . tháng. . . . năm 20..

Chủ tịch Hội đồng
(Ký và ghi rõ Họ tên)

Sinh viên đã hoàn thành và nộp bản Đồ án tốt nghiệp cho Hội đồng thi ngày... tháng... năm 20...

Sinh viên làm Đồ án tốt nghiệp
(Ký và ghi rõ Họ tên)



TRƯỜNG ĐẠI HỌC THỦY LỢI
KHOA CÔNG NGHỆ THÔNG TIN

BẢN TÓM TẮT ĐỀ CƯƠNG ĐỒ ÁN TỐT NGHIỆP

TÊN ĐỀ TÀI: ỨNG DỤNG CÁC MÔ HÌNH HỌC MÁY DỰ ĐOÁN GIÁ TRỊ CHUYỂN NHƯỢNG CỦA CẦU THỦ BÓNG ĐÁ CHUYÊN NGHIỆP

Sinh viên thực hiện: Nguyễn Thành Trung

Lớp: 61TH6

Mã sinh viên: 1951061075

Giáo viên hướng dẫn: Ths. Nguyễn Đắc Hiếu

TÓM TẮT ĐỀ TÀI

Trong làng bóng đá, chuyển nhượng cầu thủ không chỉ là một khía cạnh quan trọng mà còn là một thương vụ tài chính đáng kể. Việc đánh giá và định giá cầu thủ trước đây thường phụ thuộc vào sự hiểu biết cá nhân và đánh giá tay nghề. Tuy nhiên, sự phát triển của công nghệ, đặc biệt là trí tuệ nhân tạo (AI), đã thay đổi cách các đội bóng quản lý và phát triển đội hình.

Hiện nay, mô hình học máy và trí tuệ nhân tạo đang được sử dụng để dự đoán giá trị chuyển nhượng và giá thị trường của cầu thủ. Các mô hình này có khả năng phân tích đồng thời nhiều yếu tố về hiệu suất của cầu thủ, từ số bàn thắng đến kiến tạo và các chỉ số khác. Điều này giúp tăng cường tính chính xác và khách quan trong quá trình đánh giá cầu thủ. Sử dụng mô hình học máy và AI cũng mang lại khả năng dự báo tiềm năng phát triển của cầu thủ. Các thuật toán có thể theo dõi sự tiến bộ của họ theo thời gian, dự đoán khả năng thích nghi với môi trường mới, và cung cấp thông tin hữu ích cho các cuộc đàm phán hợp đồng và chuyển nhượng.

Với sự kết hợp giữa vấn đề thực tế và sức mạnh của công nghệ hiện đại, đề tài “**Ứng dụng các mô hình học máy dự đoán giá trị chuyển nhượng của cầu thủ bóng đá chuyên nghiệp**” của em đặt ra một quan điểm mới và hứa hẹn đối với phát triển trong lĩnh vực này.

CÁC MỤC TIÊU CHÍNH

- Nghiên cứu bài toán, tìm nguồn dữ liệu và thực hiện thu thập dữ liệu
- Thu thập và xử lý dữ liệu
- Tìm hiểu và nghiên cứu mô hình học máy
- Áp dụng mô hình học máy cho bài toán
- Đánh giá mô hình

KẾT QUẢ DỰ KIẾN

- Trình bày hoàn chỉnh về việc thu thập dữ liệu và xây dựng mô hình học máy cho bài toán dự đoán giá chuyển nhượng.
- Xây dựng một chương trình minh họa cho việc dự đoán.
- Báo cáo đồ án tốt nghiệp

LỜI CAM ĐOAN

Tác giả xin cam đoan đây là Đồ án tốt nghiệp của bản thân tác giả. Các kết quả trong Đồ án tốt nghiệp này là trung thực, và không sao chép từ bất kỳ một nguồn nào và dưới bất kỳ hình thức nào. Việc tham khảo các nguồn tài liệu (nếu có) đã được thực hiện trích dẫn và ghi nguồn tài liệu tham khảo đúng quy định.

Tác giả ĐATN

Nguyễn Thành Trung

LỜI CẢM ƠN

Trong suốt quá trình học tập tại khoa Công nghệ thông tin trường Đại học Thủy Lợi, em cảm thấy vô cùng may mắn khi đã nhận được sự quan tâm, giúp đỡ và sự chỉ bảo nhiệt tình của các thầy, cô giáo trong cả học tập và đời sống.

Em muốn gửi lời cảm ơn chân thành và sâu sắc nhất đến với các thầy, cô giáo trường Đại học Thủy Lợi, những người đã tận tâm truyền đạt các kiến thức và niềm cảm hứng học tập đến cho cá nhân em và các bạn sinh viên khác. Và đặc biệt hơn, trong quãng thời gian làm đồ án tốt nghiệp, em đã nhận được sự hướng dẫn tận tình của Thạc sĩ, thầy giáo Nguyễn Đắc Hiếu. Em xin gửi lời biết ơn sâu sắc đến thầy, người đã giúp đỡ, hỗ trợ và cho em những lời khuyên hữu ích để em có thể hoàn thành đồ án tốt nghiệp của mình hiệu quả và kịp tiến độ của học phần.

Trong thời gian thực hiện đồ án tốt nghiệp của mình với đề tài “ **Ứng dụng các mô hình học máy dự đoán giá trị chuyển nhượng của các cầu thủ bóng đá chuyên nghiệp**”, em đã cố gắng hết sức để xây dựng và hoàn thiện đồ án một cách tốt nhất, nhưng do kiến thức còn nhiều thiếu sót, thời gian hoàn thiện có hạn và thiếu kinh nghiệm thực tế nên không thể tránh khỏi các sai sót. Em rất mong nhận được sự góp ý của các thầy, cô để đồ án trở nên hoàn thiện hơn.

Em xin chân thành cảm ơn !

MỤC LỤC

CHƯƠNG 1: GIỚI THIỆU BÀI TOÁN, GIỚI THIỆU TỔNG QUAN	1
1.1 Lý do chọn đề tài:	1
1.2 Mục tiêu:.....	2
1.3 Đối tượng, phạm vi nghiên cứu:.....	2
1.3.1 Đối tượng nghiên cứu:.....	2
1.3.2 Phạm vi nghiên cứu:	3
CHƯƠNG 2 : CÁC CÔNG CỤ VÀ KỸ THUẬT SỬ DỤNG TRONG BÀI.....	4
2.1 Tổng quan lý thuyết về học máy:	4
2.2 Các thuật toán hồi quy:.....	5
2.2.1 Hồi quy rừng ngẫu nhiên (Random Forest Regression):.....	5
2.2.2 Hồi quy XGBoost Regression (Extreme Gradient Boosting Regression):	6
2.2.3 Hồi quy LightGBM (Light Gradient Boosting Machine Regression):	8
2.3 Các phương pháp đánh giá mô hình:.....	10
2.4 Các công cụ sử dụng trong bài:	12
2.5 Tối ưu siêu tham số:	13
CHƯƠNG 3 : ỨNG DỤNG PHƯƠNG PHÁP VÀ XÂY DỰNG MÔ HÌNH	15
3.1 Phân tích bài toán và lựa chọn dữ liệu:	15
3.1.1 Mô tả bài toán:.....	15
3.1.2 Quy trình tổng quan các bước thực hiện:	15
3.1.3 Dữ liệu:	17
3.2 Thu thập dữ liệu:.....	26
3.2.1 Thu thập dữ liệu cầu thủ từ trang web sofifa:	26
3.2.2 Thu thập dữ liệu giá chuyển nhượng từ trang web transfermarkt:	30
3.2.3 Hợp nhất dữ liệu:	35
3.3 Tiền xử lý dữ liệu:	37
3.3.1 Lựa chọn thuộc tính (Feature Selection):	37
3.3.2 Xử lý dữ liệu thiếu (Handle Missing Data):	38
3.3.3 Kỹ thuật thuộc tính (Feature Engineering):.....	38
3.3.4 Xử lý dữ liệu ngoại lai (Remove Outlier):	39
3.3.5 Chuẩn hóa và mã hóa dữ liệu:	42

3.4 Xây dựng mô hình:	42
3.4.1 Mô hình LightGBM:.....	42
3.4.2 Mô hình XGBoost:	44
3.4.3 Mô hình Random Forest Regression:	45
CHƯƠNG 4 : KẾT QUẢ VÀ ĐÁNH GIÁ MÔ HÌNH	47
4.1 Kết quả các độ đo của các mô hình:	47
4.2 Đánh giá mô hình:	50
KẾT LUẬN	55
TÀI LIỆU THAM KHẢO	56

DANH MỤC CÁC HÌNH ẢNH

Hình 2-1: Sơ đồ biểu diễn cách hoạt động của mô hình Random Forest Regression	6
Hình 2-2: Các tối ưu hóa thuật toán của XGBoost so với GBM tiêu chuẩn	7
Hình 2-3: Sơ đồ hoạt động của LightGBM	9
Hình 3-1: Sơ đồ quy trình tổng quan các bước thực hiện	16
Hình 3-2: Dữ liệu cầu thủ trên web SoFIFA	19
Hình 3-3: Các giải đấu có tổng số tiền chi cho thị trường chuyển nhượng cao nhất	23
Hình 3-4: Dữ liệu chuyển nhượng của Transfermarkt	25
Hình 3-5: Quy trình thu thập dữ liệu Sofifa	26
Hình 3-6: Giao diện chứa đường liên kết của cầu thủ	27
Hình 3-7: Dữ liệu thông tin cá nhân cầu thủ	28
Hình 3-8: Dữ liệu về đặc điểm và chỉ số kỹ năng của cầu thủ	29
Hình 3-9: Mã số từng năm của Sofifa	30
Hình 3-10: Kết quả thu được	30
Hình 3-11: Quy trình thu thập dữ liệu giá chuyển nhượng	31
Hình 3-12: Lựa chọn tham số để tạo đường dẫn	32
Hình 3-13: Các dữ liệu đặc biệt	34
Hình 3-14: Kết quả sau quá trình thu thập giá trị chuyển nhượng	35
Hình 3-15: Dữ liệu từng năm của giá chuyển nhượng	35
Hình 3-16: Kết quả sau khi lọc của bộ dữ liệu giá chuyển nhượng	36
Hình 3-17: Kết quả sau khi thêm trường vào 2 bộ dữ liệu	36
Hình 3-18: Phân bố dữ liệu của goalkeeping_reflexes và mentality_composure với các vị trí của cầu thủ	38
Hình 3-19: Phân bố dữ liệu theo nhóm các vị trí	39
Hình 3-20: Biểu đồ heatmap thể hiện sự tương quan giữa các thuộc tính	40
Hình 3-21: Biểu đồ trực quan mối quan hệ giữa các thuộc tính overall, potential, movement_reactions, mentality_composure với biến mục tiêu fee	40
Hình 3-22: Biểu đồ trực quan mối quan hệ giữa các thuộc tính overall, potential, movement_reactions, mentality_composure với biến mục tiêu fee sau khi loại bỏ các dữ liệu ngoại lai	41
Hình 4-1: Biểu đồ kết quả đánh giá mô hình với Logarit và chuẩn hóa biến ‘fee’	47
Hình 4-2: Biểu đồ kết quả đánh giá mô hình với Logarit (fee)	48
Hình 4-3: Biểu đồ kết quả đánh giá mô hình với chuẩn hóa (fee)	48
Hình 4-4: Biểu đồ so sánh kết quả giữa 3 trường hợp dữ liệu khác nhau	49
Hình 4-5: Giao diện dự đoán giá trị chuyển nhượng	50
Hình 4-6: Biểu đồ thể hiện giữa giá trị thực tế và giá trị dự đoán của mô hình LightGBM	51
Hình 4-7: Biểu đồ thể hiện giữa giá trị thực tế và giá trị dự đoán của mô hình Random Forest	51

Hình 4-8: Biểu đồ thể hiện giữa giá trị thực tế và giá trị dự đoán của mô hình XGBoost	52
Hình 4-9: Top 10 thuộc tính quan trọng nhất của mô hình LightGBM	53

DANH MỤC BẢNG BIỂU

Bảng 3-1: Mô tả các vị trí của cầu thủ.....	18
Bảng 3-2: Mô tả phạm vi giá trị của các thuộc tính kỹ năng	20
Bảng 3-3: Mô tả các thuộc tính trong bộ dữ liệu.....	21
Bảng 3-4: 20 giải đấu bóng đá với tổng số tiền chi cho việc chuyển nhượng cao nhất	24
Bảng 3-5: Mô tả các thuộc tính trong bộ dữ liệu giá chuyển nhượng.....	25
Bảng 4-1: Kết quả các độ đo của các mô hình học máy	47

DANH MỤC CÁC TỪ VIẾT TẮT VÀ GIẢI THÍCH

1	AI	Artificial Intelligence
2	CSV	Comma-Separated Values
3	CSS	Cascading Style Sheets
4	EFB	Exclusive Feature Bundling
5	FIFA	Fédération Internationale de Football Association
6	GBM	Gradient Boosting Model
7	GOSS	Gradient Based One Side Sampling
8	GUI	Graphical User Interface
9	HTML	HyperText Markup Language
10	MAE	Mean Absolute Error
11	MSE	Mean Squared Error
12	NaN	Not a Number
13	RMSE	Root Mean Squared Error
14	RFR	Random Forest Regression
15	R ²	Coefficient of Determination
16	SIGKDD	Special Interest Group on Knowledge Discovery and Data Mining
17	SQL	Structured Query Language
18	URL	Uniform Resource Locator
19	XML	eXtensible Markup Language

CHƯƠNG 1: GIỚI THIỆU BÀI TOÁN, GIỚI THIỆU TỔNG QUAN

1.1 Lý do chọn đề tài:

Với hơn 5 tỷ người hâm mộ trên toàn thế giới, bóng đá không chỉ là một môn thể thao phổ biến nhất trên thế giới mà còn là một hiện tượng văn hóa mạnh mẽ. Sự hấp dẫn của nó đã tạo ra một cộng đồng đa dạng và đông lòng, làm cho bóng đá trở thành một phần không thể tách rời từ cuộc sống hàng ngày của hàng tỷ người trên toàn thế giới.

Với sự yêu thích và đón nhận mạnh mẽ như vậy, bóng đá không chỉ là trò chơi mà còn là cầu nối văn hóa kết nối mọi ngóc ngách của thế giới. Các giải đấu hàng đầu như World Cup, Premier League,... và các sự kiện thể thao tầm cỡ thế giới đã trở thành những điểm nhấn quan trọng trong lịch sử của từng quốc gia và đội bóng. Và chính vì những giải đấu này là nơi diễn ra những cuộc cạnh tranh khốc liệt giữa các đội bóng, việc lựa chọn cầu thủ phù hợp đã trở thành yếu tố không thể thiếu, đóng góp quan trọng vào thành công và danh tiếng của mỗi đội. Sự thành công của bất kỳ trận đấu bóng đá nào đều nằm ở việc lựa chọn cầu thủ, đây là một quyết định khó khăn đối với các huấn luyện viên của mỗi đội bóng. Quá trình này không chỉ đòi hỏi sự nhạy bén với kỹ thuật và tình hình trận đấu mà còn đòi hỏi sự hiểu biết sâu sắc về các yếu tố như kỹ năng cá nhân, thống kê hiệu suất, sự kết hợp giữa các cầu thủ, thể lực, yếu tố tâm lý và chấn thương. Thị trường chuyển nhượng ngày càng trở nên phức tạp với sự tăng cường của nguồn lực tài chính, sự cạnh tranh giữa các đội bóng lớn, và sự tăng giá của các cầu thủ tiêu biểu. Quản lý chuyển nhượng không chỉ đòi hỏi sự hiểu biết vững về bóng đá mà còn đòi hỏi khả năng phân tích và đánh giá dữ liệu một cách chính xác. Giá trị chuyển nhượng của một cầu thủ ảnh hưởng trực tiếp đến cơ cấu đội hình và chiến lược của đội bóng. Việc đưa ra quyết định chuyển nhượng chính xác có thể tối ưu hóa hiệu suất trận đấu và đồng thời duy trì sự ổn định tài chính của đội.

Thị trường chuyển nhượng cầu thủ bóng đá chắc chắn đã trở thành một ngành kinh doanh lớn, và với số tiền lớn như vậy được chi cho một số lượng tài sản tương đối nhỏ sẽ có rủi ro cao. Sai lầm là chuyện thường tình và các câu lạc bộ thường rơi vào tình thế khó khăn khi phải trả hàng chục triệu euro cho một cầu thủ không đáp ứng được kỳ vọng. Những sai lầm như vậy có thể gây ra hậu quả thảm khốc cho câu lạc bộ và người

hâm mộ, bao gồm cả việc xuống hạng và phá sản. Chính vì lý do này, đồ án này sẽ sử dụng kỹ thuật học máy để xây dựng mô hình dự đoán giá chuyển nhượng của cầu thủ bóng đá chuyên nghiệp dựa vào các kỹ năng và đặc điểm của các cầu thủ. Với sự hỗ trợ của mô hình này không chỉ giúp nhà quản lý đưa ra quyết định thông minh và chiến lược mà còn giúp họ tối ưu hóa nguồn lực tài chính, tạo ra đội hình mạnh mẽ và cạnh tranh trên mọi mặt trận. Việc này không chỉ là quan trọng đối với sự thành công của đội bóng mà còn góp phần vào sự hấp dẫn và sức sống của môn thể thao vua này.

1.2 Mục tiêu:

Mục tiêu chính của đồ án là phát triển một mô hình dự đoán giá trị chuyển nhượng của cầu thủ bóng đá, tập trung vào việc khám phá mức độ có thể dự đoán của giá trị chuyển nhượng dựa trên kỹ năng và đặc điểm cá nhân của từng cầu thủ. Dữ liệu sẽ được thu thập bao gồm thông tin chi tiết về các kỹ năng và các yếu tố quan trọng khác từ trang web Sofifa, cùng với dữ liệu về giá trị chuyển nhượng từ trang web Transfermarkt. Sau đó chúng tôi sẽ sử dụng dữ liệu này để xây dựng một mô hình học máy, có khả năng dự đoán giá trị chuyển nhượng của cầu thủ.

Quá trình nghiên cứu sẽ tập trung vào việc xác định mức độ ảnh hưởng của từng yếu tố, như kỹ năng chi tiết, hiệu suất trong trận đấu, và các chỉ số tâm lý, đối với giá trị chuyển nhượng. Mục tiêu cuối cùng là cung cấp một công cụ dự đoán linh hoạt, giúp nhà quản lý đội bóng và đội ngũ tuyển trạch viên có cái nhìn chi tiết về giá trị của cầu thủ dựa trên những yếu tố cụ thể, từ đó hỗ trợ quyết định chuyển nhượng thông minh và tối ưu hóa nguồn lực tài chính của đội. Dựa vào đó cũng sẽ khám phá và đánh giá tiềm năng của các phương pháp học máy trong việc ước tính giá trị chuyển nhượng, tạo ra sự đóng góp mới và nâng cao hiệu suất so với các phương pháp truyền thống.

1.3 Đối tượng, phạm vi nghiên cứu:

1.3.1 Đối tượng nghiên cứu:

Nghiên cứu sẽ tập trung vào cầu thủ bóng đá chuyên nghiệp, đặc biệt là những cá nhân có tầm ảnh hưởng lớn đối với thị trường chuyển nhượng. Điều này bao gồm cả các cầu thủ nổi tiếng và những tài năng mới nổi có tiềm năng phát triển. Trọng tâm của nghiên cứu sẽ được đặt vào việc đánh giá các đặc điểm và kỹ năng cá nhân của cầu thủ, nhằm

mục đích hiểu rõ hơn về những yếu tố nào góp phần vào giá trị chuyển nhượng của họ. Phân tích sẽ được tiến hành trên cầu thủ từ các vị trí khác nhau trên sân, bao gồm tiền đạo, tiền vệ, và hậu vệ, để khám phá ảnh hưởng của từng vị trí đối với giá trị chuyển nhượng.

1.3.2 Phạm vi nghiên cứu:

Phạm vi nghiên cứu sẽ được giới hạn trong một khoảng thời gian cụ thể từ mùa giải 2014/2015 – 2022/2023, tập trung vào các mùa chuyển nhượng hoặc chu kỳ thương vụ quan trọng trong lịch sử giải đấu bóng đá lớn và đầy uy tín. Nghiên cứu sẽ chú trọng vào việc thu thập và phân tích dữ liệu liên quan đến kỹ năng cá nhân của cầu thủ, thông tin cá nhân và giá trị chuyển nhượng. Cụ thể, dữ liệu sẽ bao gồm thông tin chi tiết về kỹ năng của cầu thủ, các yếu tố tâm lý và dữ liệu về giá trị chuyển nhượng của 20 giải đấu hàng đầu, được xác định bởi tổng số tiền mà các đội bóng của giải đấu đó đầu tư vào thị trường chuyển nhượng trong khoảng thời gian xác định. Điều này sẽ giúp nghiên cứu theo dõi và đánh giá sự biến động của giá trị chuyển nhượng theo thời gian, cung cấp cái nhìn sâu sắc về các yếu tố ảnh hưởng đến thị trường chuyển nhượng trong ngữ cảnh của những giải đấu danh tiếng. Tiếp theo, sẽ triển khai các mô hình học máy để dự đoán giá trị chuyển nhượng dựa trên các biến số nghiên cứu. Điều này sẽ giúp xác định mối quan hệ giữa các yếu tố như kỹ năng, thông tin cá nhân và giá trị chuyển nhượng.

CHƯƠNG 2 : CÁC CÔNG CỤ VÀ KỸ THUẬT SỬ DỤNG TRONG BÀI

2.1 Tổng quan lý thuyết về học máy:

Học máy là một lĩnh vực trong trí tuệ nhân tạo (AI) mà nó tập trung vào việc phát triển các phương pháp và thuật toán để máy tính có thể học từ dữ liệu mà không cần phải được lập trình một cách tường minh. Mục tiêu của học máy là xây dựng mô hình dự đoán và phân loại có khả năng tự cải thiện thông qua trải nghiệm. Học máy có thể ứng dụng hầu hết các loại dữ liệu từ cấu trúc (bảng số liệu kinh tế, số liệu nông nghiệp,...) và cả dữ liệu phi cấu trúc (hình ảnh, văn bản, tín hiệu âm thanh,...). Học máy được chia thành 3 loại chính: học có giám sát, học không giám sát và học tăng cường.

Học có giám sát (Supervised Learning) là một dạng học máy trong đó mô hình được huấn luyện trên dữ liệu có nhãn. Mục tiêu là dự đoán hoặc phân loại dữ liệu mới dựa trên mối liên kết giữa đầu vào và đầu ra đã biết. Sử dụng khi bạn muốn nhận dự đoán một kết quả đầu ra từ dữ liệu đầu vào và bạn có các cặp dữ liệu (đầu vào/đầu ra) tương ứng. Để xây dựng mô hình học có giám sát thường phải có sự nỗ lực từ con người để gán nhãn cho tập dữ liệu.

Học không giám sát (Unsupervised Learning) không sử dụng dữ liệu có nhãn trong quá trình huấn luyện. Mô hình phải tự tìm ra cấu trúc hoặc mẫu trong dữ liệu mà không có sự hướng dẫn nào từ bên ngoài. Các nhiệm vụ thường gặp trong học không giám sát bao gồm phân loại tự nhiên (clustering) để nhóm các điểm dữ liệu có tính chất tương đồng và giảm chiều dữ liệu.

Học tăng cường (Reinforcement Learning) liên quan đến việc mô hình tương tác với một môi trường và tự cập nhật dự đoán và hành động dựa trên phản hồi từ môi trường. Mô hình học từ các tình huống trải nghiệm và tìm cách tối ưu hóa một chính sách hành động để đạt được mục tiêu cụ thể. Điều này thường được áp dụng trong các bài toán như điều khiển robot, chơi trò chơi, và quản lý tài chính.

Học máy có thể được áp dụng trong nhiều lĩnh vực khác nhau như y tế, tài chính, giáo dục, sản xuất, v.v. Học máy không chỉ là một công cụ mạnh mẽ để giải quyết các vấn đề thực tế mà còn mang lại những hiểu biết sâu rộng về mối quan hệ giữa dữ liệu và

kiến thức, đóng góp quan trọng vào sự tiến bộ của trí tuệ nhân tạo và công nghệ thông tin.

2.2 Các thuật toán hồi quy:

2.2.1 Hồi quy rừng ngẫu nhiên (*Random Forest Regression*):

Random Forest Regression (RFR) là một phương pháp mạnh mẽ trong học máy, kết hợp sức mạnh của nhiều cây quyết định để thực hiện nhiệm vụ hồi quy. Phương pháp này sử dụng một kỹ thuật gọi là Bootstrap và Aggregation, hay còn được biết đến là bagging, để tạo ra nhiều cây quyết định độc lập nhau. Bagging cho phép lựa chọn ngẫu nhiên một nhóm nhỏ các thuộc tính tại mỗi nút của cây để phân chia dữ liệu, giúp mô hình có khả năng phân loại một cách linh hoạt và nhanh chóng.

Một đặc điểm quan trọng của RFR là khả năng chọn ngẫu nhiên một số lượng đặc trưng để sử dụng trong quá trình xây dựng mỗi cây. Điều này tăng tính đa dạng và linh hoạt của mô hình, đồng thời giảm nguy cơ quá khớp. Kết quả của RFR được xác định thông qua quá trình biểu quyết, nơi kết quả của từng cây được tổng hợp theo nguyên tắc đa số hoặc lấy trung bình, tùy thuộc vào loại bài toán là phân loại hay hồi quy. Đối với bài toán hồi quy, kết quả cuối cùng của mô hình RFR sẽ là trung bình của tất cả các kết quả dự báo của các cây (Hình 2-1).

Cách thuật toán RFR hoạt động có thể được tóm tắt như sau:

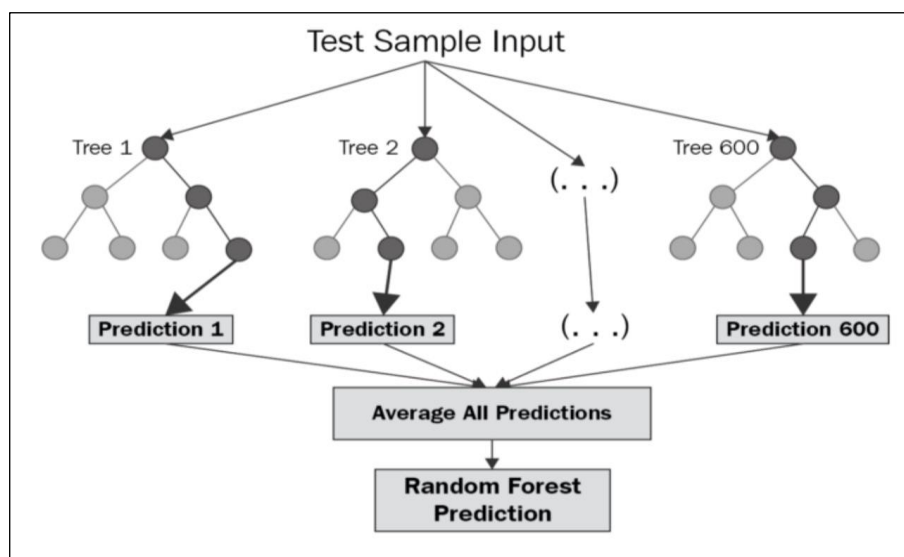
- Dùng phương pháp bootstrap để tạo ra nhiều tập dữ liệu con, mỗi tập con được tạo ngẫu nhiên từ tập dữ liệu huấn luyện với việc thay thế mẫu để tạo tính đa dạng.
- Mỗi tập dữ liệu con được sử dụng để xây dựng một cây hồi quy độc lập. Quá trình này tối ưu hóa tiêu chuẩn chia nút để tạo cây hiệu quả.
- Kết quả dự đoán cuối cùng được tính toán bằng cách biểu quyết kết quả từ tất cả các cây hồi quy. Đối với bài toán hồi quy, kết quả cuối cùng là giá trị trung bình của các dự đoán từ các cây như công thức sau:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N h_i(x)$$

(2-1)

Trong đó:

- \hat{y} là giá trị dự đoán cuối cùng.
- N là số lượng cây hồi quy.
- $h_i(x)$ là dự đoán của cây thứ i



Hình 2-1: Sơ đồ biểu diễn cách hoạt động của mô hình Random Forest Regression

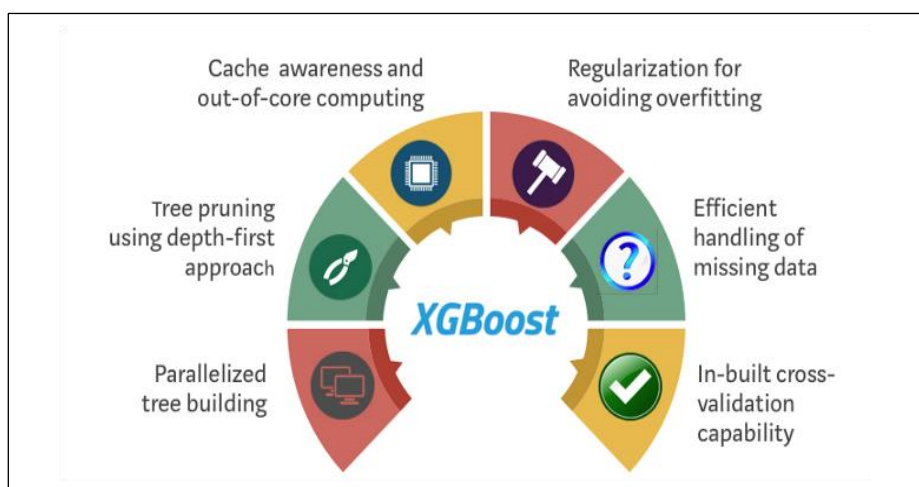
2.2.2 Hồi quy XGBoost Regression (Extreme Gradient Boosting Regression):

XGBoost (Extreme Gradient Boosting) là một thuật toán học máy hàng đầu trong lĩnh vực cây quyết định, xây dựng trên khung tăng cường độ dốc. Phát triển ban đầu bởi Tianqi Chen và Carlos Guestrin tại Đại học Washington, XGBoost đã đem lại sự đột phá trong lĩnh vực Machine Learning khi được giới thiệu tại Hội nghị SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining) vào năm 2016. Ban đầu nhắm đến giải quyết các bài toán phân loại và hồi quy, XGBoost đã nhanh chóng trở thành một trong những thuật toán phổ biến nhất.

XGBoost thuộc họ thuật toán "boosting," trong đó các mô hình được xây dựng theo kiểu tuần tự để cải thiện từng bước và khắc phục sai lệch của mô hình trước đó. Thuật toán chủ yếu tập trung vào việc tối ưu hóa hàm mất mát thông qua phương pháp Gradient Descent.

Các tối ưu và cải tiến của XGBoost dựa trên khung Gradient Boosting Model (GBM) bao gồm (Hình 2-2):

- Song song hóa: XGBoost sử dụng song song hóa để tiếp cận quá trình xây dựng cây một cách hiệu quả. Điều này thực hiện được bằng cách lồng các vòng lặp và sắp xếp chúng một cách song song để giảm thời gian chạy.
- Tỉa cây: Sử dụng tham số 'max_depth' để cắt tỉa cây từ phía sau, cải thiện độ hiệu quả tính toán.
- Tối ưu hóa phần cứng: XGBoost được tối ưu hóa cho việc sử dụng tài nguyên phần cứng thông qua việc nhận thức bộ nhớ cache và phân bổ bộ đệm nội bộ.
- Tránh overfitting: Thêm vào hàm mất mát một thành phần chặn để kiểm soát độ lớn của trọng số và giảm thiểu nguy cơ quá khớp, sử dụng 'alpha' và 'lambda' làm tham số kiểm soát.
- Xử lý dữ liệu thiếu: XGBoost tự động xử lý tính năng thiếu thốn bằng cách học giá trị tốt nhất từ sự mất mát trong quá trình đào tạo.
- Xác thực chéo: Phương pháp xác thực chéo được tích hợp sẵn giúp quá trình đào tạo được kiểm định một cách hiệu quả.



Hình 2-2: Các tối ưu hóa thuật toán của XGBoost so với GBM tiêu chuẩn

Công thức của XGBoost Regression được biểu diễn bằng công thức sau:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad (2-2)$$

Trong đó:

- \hat{y}_i là giá trị dự đoán cho mẫu thứ .
- K là số lượng cây hồi quy.

- $f_k(x_i)$ là dự đoán của cây thứ k cho mẫu thứ i .

Cách thuật toán hoạt động có thể tóm tắt như sau:

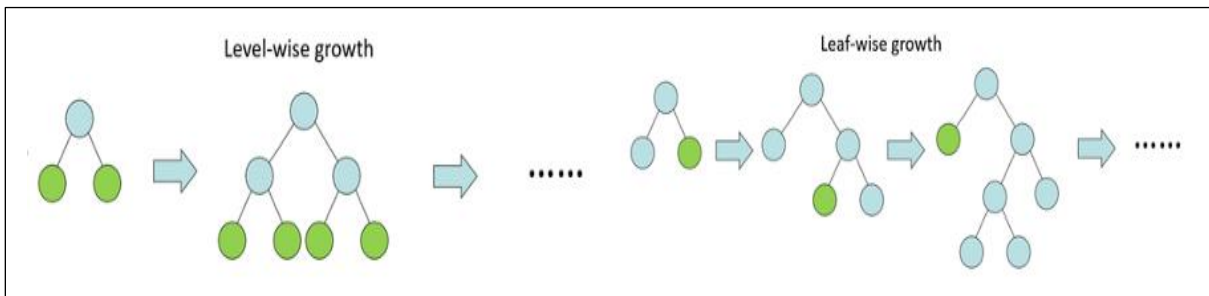
- Tạo cây quyết định (Decision Trees): Mỗi cây quyết định được xây dựng bằng cách chia dữ liệu thành các nhóm dựa trên các đặc trưng của dữ liệu. XGBoost sử dụng cây quyết định có cấp độ sâu thấp, thường chỉ vài chục nút lá, để tránh quá mức phức tạp và giữ cho mô hình tốt hơn trong việc tổng hợp thông tin.
- Tối ưu hóa hàm mất mát (Loss Function Optimization): Mục tiêu là tối thiểu hóa hàm mất mát, là sự chênh lệch giữa giá trị dự đoán và giá trị thực tế. Hàm mất mát thường là một hàm phi tuyến, và XGBoost sử dụng kỹ thuật Gradient Boosting để cập nhật dự đoán mô hình sao cho mất mát giảm thiểu.
- Regularization để tránh overfitting, XGBoost thêm các thành phần chặn vào hàm mất mát, bao gồm cả l1 (Lasso) và l2 (Ridge) regularization. Các thành phần này giúp kiểm soát trọng số của cây, ngăn chặn chúng trở nên quá phức tạp và giảm khả năng tổng hợp dữ liệu nhiễu.
- Quyết định cây tiếp theo dựa trên Gradient: XGBoost không xây dựng tất cả các cây cùng một lúc, mà là từng cây một. Sau khi xây dựng một cây, nó sẽ tính gradient của hàm mất mát đối với dự đoán hiện tại và sử dụng gradient này để xây dựng cây tiếp theo.
- Tích hợp các cây để tạo mô hình cuối cùng: Khi tất cả các cây đã được xây dựng, dự đoán cuối cùng của mô hình là tổng của các dự đoán từ tất cả các cây.
- Learning Rate (Tốc độ học): XGBoost có một tham số được gọi là 'learning rate' để kiểm soát mức độ cập nhật của dự đoán sau mỗi cây được thêm vào. Learning rate giúp kiểm soát tốc độ học và tránh quá mức điều chỉnh mô hình.
- Cross-validation và Early Stopping: XGBoost thường sử dụng kỹ thuật cross-validation để đánh giá hiệu suất của mô hình. Early stopping được sử dụng để dừng quá trình huấn luyện khi không có sự cải thiện đáng kể nữa, giúp giảm thời gian và nguồn lực.

2.2.3 Hồi quy *LightGBM* (*Light Gradient Boosting Machine Regression*):

LightGBM là một thuật toán học máy thuộc dòng Gradient Boosting Frameworks, đã đem lại đột phá đáng kể trong việc xử lý dữ liệu lớn và tăng cường hiệu suất của các mô

hình. Được Microsoft phát triển và cho ra mắt bản thử nghiệm vào tháng 1 năm 2016, và nhanh chóng trở thành thuật toán ensemble được ưa chuộng nhất.

LightGBM phát triển cây theo lá dựa trên ‘leaf-wise tree growth’, trong khi hầu hết các thuật toán ensemble khác dựa trên phương pháp tăng trưởng cây theo cấp ‘level (depth)-wise tree growth’. ‘Leaf-wise tree growth’ lựa chọn nút để phát triển cây dựa trên tối ưu toàn bộ tree, trong khi ‘level (depth)-wise tree growth’ tối ưu trên nhánh đang xét, do đó với số node nhỏ, các cây xây dựng từ ‘leaf-wise tree growth’ thường thực hiện tốt hơn ‘level (depth)-wise tree growth’ (Hình 2-3).



Hình 2-3: Sơ đồ hoạt động của LightGBM

Công thức của LightGBM Regression được biểu diễn bằng công thức sau:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i)$$

(2-3)

Trong đó:

- \hat{y}_i là giá trị dự đoán cho mẫu thứ i .
- K là số lượng cây hồi quy.
- $f_k(x_i)$ là dự đoán của cây thứ k cho mẫu thứ i .

Cách thuật toán hoạt động có thể tóm tắt như sau:

- LightGBM sử dụng phương pháp ‘leaf-wise tree growth’ làm tối ưu toàn bộ cây dựa trên các nút lá thay vì theo cấp ‘level (depth)-wise tree growth’ như nhiều thuật toán khác. Cách tiếp cận này giúp tối ưu hóa hiệu suất của cây với số lượng node ít hơn, đặc biệt là hiệu quả với dữ liệu có kích thước lớn.

- LightGBM sử dụng ‘histogram-based algorithms’ thay vì ‘pre-sort-based algorithms’ để tìm kiếm ‘split points’ khi xây dựng cây. Cải tiến này giúp tăng tốc quá trình đào tạo và giảm lượng bộ nhớ cần sử dụng bằng cách sử dụng các biểu đồ histogram để nhanh chóng tính toán các giá trị tối ưu.
- Sử dụng thuật toán GOSS (Gradient Based One Side Sampling) để tối ưu hóa quá trình lấy mẫu (sampling) bằng cách tập trung vào các mẫu có gradient lớn, giảm kích thước dữ liệu được sử dụng trong quá trình đào tạo.
- LightGBM sử dụng EFB (Exclusive Feature Bundling) để kết hợp các đặc trưng một cách độc quyền, giảm độ phức tạp của cây và tăng tốc độ tính toán. Việc này giúp giảm số lượng cây cần xây dựng và tăng hiệu suất.

2.3 Các phương pháp đánh giá mô hình:

Để đánh giá mức độ dự đoán chính xác của mô hình, bốn độ đo sau được sử dụng: MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), R2 (Coefficient of Determination).

MSE (Mean Squared Error): Đây là trung bình giá trị bình phương của sai số giữa giá trị dự đoán và giá trị thực tế. MSE được sử dụng để đo lường độ lỗi trung bình của mô hình. MSE càng thấp thì mô hình càng chính xác. MSE có thể bị ảnh hưởng bởi các giá trị ngoại lai.

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (2-4)$$

Trong đó:

- N là số lượng điểm dữ liệu.
- y_i là giá trị thực tế.
- \hat{y}_i là giá trị dự đoán.

MAE (Mean Absolute Error): Đo độ lớn trung bình các lỗi trong một tập hợp các dự đoán mà không cần xem xét hướng của chúng. Đó là giá trị trung bình trên tập mẫu kiểm tra về sự khác biệt tuyệt đối giữa dự đoán và quan sát thực tế, trong đó tất cả các khác biệt riêng lẻ có trọng số bằng nhau.

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

(2-5)

Trong đó:

- N là số lượng điểm dữ liệu.
- y_i là giá trị thực tế.
- \hat{y}_i là giá trị dự đoán.

MAE luôn không âm và giá trị 0 (hầu như không bao giờ đạt được trong thực tế) sẽ chỉ ra sự phù hợp hoàn hảo với dữ liệu. MAE càng thấp thì kết quả dự báo càng gần với thực tế.

R2 (Coefficient of Determination): Đây là độ đo đánh giá mức độ giải thích của mô hình hồi quy. R2 có giá trị từ 0 đến 1, trong đó 1 là một mô hình hoàn hảo và 0 là một mô hình không có khả năng giải thích. R2 càng gần 1 thì mô hình càng tốt.

$$R^2 = 1 - \frac{SSE}{SST}$$

(2-6)

Trong đó:

- $SSE = \sum_{i=1}^N (y_i - \hat{y}_i)^2$ là tổng bình phương sai số (Sum of Squared Errors), tức là tổng của bình phương hiệu giữa giá trị dự đoán và giá trị thực tế
- $SST = \sum_{i=1}^N (y_i - \bar{y}_l)^2$ là tổng bình phương độ lệch (Sum of Squared Total), tức là tổng của bình phương hiệu giữa giá trị thực tế và giá trị trung bình của biến phụ thuộc.

RMSE (Root Mean Squared Error): Đây là căn bậc hai của trung bình giá trị bình phương của sai số giữa giá trị dự đoán và giá trị thực tế. RMSE được sử dụng để đo lường độ lỗi trung bình của mô hình. RMSE càng thấp thì mô hình càng chính xác. RMSE có thể được sử dụng để so sánh các mô hình khác nhau.

$$RMSE = \sqrt{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

(2-7)

2.4 Các công cụ sử dụng trong bài:

Môi trường thực nghiệm:

- Ngôn ngữ: Python
- Nền tảng phát triển: Jupyter Notebook
- Công cụ: Visual Studio Code
- Window 11, Ram 16GB, CPU Intel® Core(TM) i5-13420H

Các thư viện sử dụng:

- CloudScraper là một thư viện Python dựa trên Python Requests, được thiết kế để vượt qua hệ thống bảo vệ chống bot của Cloudflare. Điều này cho phép thu thập dữ liệu từ các trang web triển khai trên nền tảng Cloudflare, nơi thường áp dụng các biện pháp bảo vệ bot. Thư viện này sử dụng các kỹ thuật mô phỏng hành vi trình duyệt, bao gồm chạy trình duyệt headless, thực hiện JavaScript, và giả mạo User-Agent để mô phỏng trình duyệt thực sự và tránh được các biện pháp bảo vệ bot.
- BeautifulSoup là một thư viện Python mạnh mẽ giúp phân tích cú pháp HTML và XML một cách linh hoạt. Nó tạo cây cú pháp từ nguồn HTML hoặc XML, cung cấp công cụ để điều hướng và trích xuất thông tin. Thư viện này cho phép tìm kiếm thẻ, thuộc tính, truy cập nội dung thẻ, và tìm kiếm đoạn văn bản cụ thể. BeautifulSoup thích hợp để trích xuất dữ liệu từ trang web và lưu trữ nó trong nhiều định dạng khác nhau như CSV, Excel, hoặc SQL để phân tích hoặc xây dựng ứng dụng. Điều đặc biệt là nó linh hoạt và có thể sử dụng cho nhiều ngôn ngữ đánh dấu, không chỉ riêng HTML hoặc XML.
- Pandas: là một thư viện Python cung cấp các cấu trúc dữ liệu và công cụ phân tích dữ liệu mạnh mẽ, đặc biệt là DataFrame. DataFrame là một bảng dữ liệu hai chiều với các hàng và cột được đặt tên, cho phép thực hiện nhanh chóng các thao tác như lọc, chọn, và xử lý dữ liệu. Pandas cũng hỗ trợ đọc và ghi dữ liệu từ nhiều định dạng như CSV, Excel, SQL databases.

- NumPy: là thư viện Python cho tính toán số học và thống kê, cung cấp đối tượng mảng nhiều chiều (ndarray) linh hoạt. NumPy giúp thực hiện các phép toán số học nhanh chóng trên dữ liệu mảng, làm cho việc xử lý và tính toán trên dữ liệu lớn trở nên hiệu quả. NumPy còn chứa các chức năng linh hoạt cho đại số tuyến tính, biến đổi Fourier, và nhiều tính năng khác.
- Matplotlib: là một thư viện vẽ đồ thị và biểu đồ 2D cho Python, cho phép hiển thị dữ liệu một cách trực quan. Nó cung cấp giao diện để tạo các biểu đồ, đồ thị đường, đồ thị phức tạp, và nhiều loại biểu đồ khác. Matplotlib giúp tạo ra đồ thị chất lượng cao, có thể tùy chỉnh đồ họa và chú thích để phản ánh thông tin một cách rõ ràng.
- Seaborn: là một thư viện trực quan hóa dữ liệu dựa trên Matplotlib, tối ưu hóa cho việc tạo ra các biểu đồ thống kê hấp dẫn và dễ hiểu. Seaborn cung cấp các giao diện đơn giản để vẽ các biểu đồ phức tạp như biểu đồ violin, biểu đồ hộp, và biểu đồ phân phối. Nó cũng có khả năng tương tác mạnh mẽ với Pandas DataFrame, giúp dễ dàng thực hiện phân tích thống kê và thăm dò dữ liệu.
- Tkinter là một thư viện GUI (Graphical User Interface) tích hợp sẵn trong Python, cung cấp các công cụ và widgets để xây dựng các giao diện người dùng đồ họa đơn giản. Tkinter chủ yếu dựa trên thư viện GUI Tk (Tcl/Tk) và là một phần của thư viện tiêu chuẩn của Python. Tkinter cung cấp cơ chế xử lý sự kiện thông qua việc sử dụng các hàm callback. Khi một sự kiện xảy ra, hàm được gọi để xử lý sự kiện đó.

2.5 Tối ưu siêu tham số:

Bayesian Search là một phương pháp tối ưu hóa siêu tham số (hyperparameter) trong quá trình đào tạo mô hình máy học. Nó dựa trên ý tưởng của Bayesian Optimization, một phương pháp tối ưu hóa toàn diện, linh hoạt và hiệu quả. Trong ngữ cảnh của việc tinh chỉnh hyperparameter, phương pháp này thường được gọi là Bayesian Hyperparameter Optimization hoặc Bayesian Search for Hyperparameter Tuning.

Cách Hoạt Động:

- Mô hình ước lượng số đối tượng (Surrogate Model): Bayesian Search sử dụng một mô hình ước lượng số đối tượng để xấp xỉ hàm mất mát (loss function) tại các điểm chưa biết của không gian siêu tham số. Gaussian Process là một trong những mô hình phổ biến được sử dụng cho mục đích này.

- Hàm đánh giá (Acquisition Function): Một hàm đánh giá được sử dụng để chọn ra điểm kế tiếp để đánh giá. Hàm này kết hợp giữa sự cân nhắc giữa việc khám phá các khu vực chưa biết và khai thác các khu vực mà chúng ta có đủ thông tin.
- Quá trình lựa chọn và cập nhật mô hình: Dựa vào hàm đánh giá, một điểm mới được chọn để đánh giá. Sau đó, mô hình ước lượng số đối tượng được cập nhật với thông tin mới. Quá trình này lặp đi lặp lại cho đến khi đạt được sự hội tụ hoặc đã đạt đến số lượng lần lặp quy định.

CHƯƠNG 3 : ỨNG DỤNG PHƯƠNG PHÁP VÀ XÂY DỰNG MÔ HÌNH

3.1 Phân tích bài toán và lựa chọn dữ liệu:

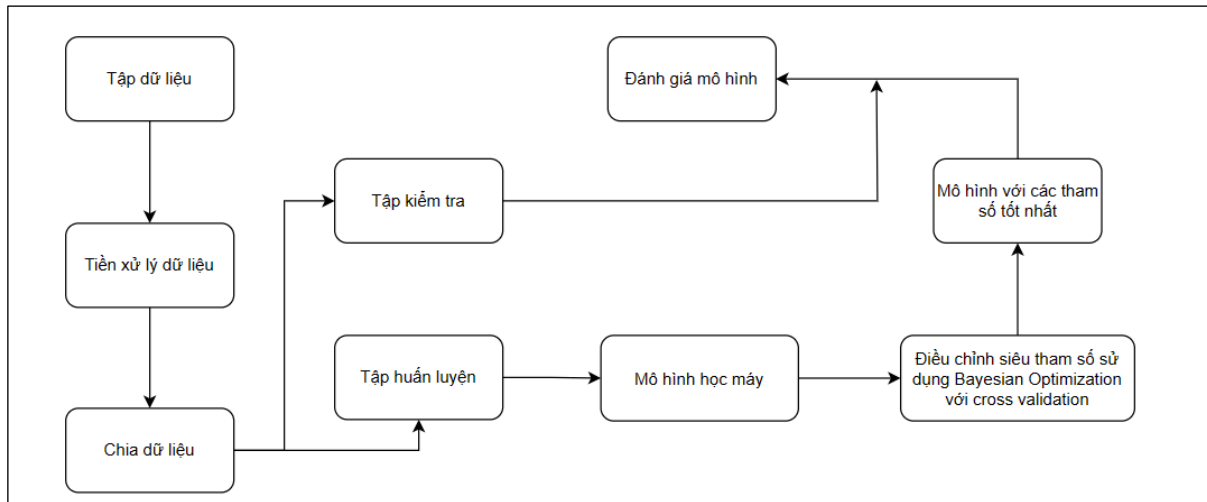
3.1.1 Mô tả bài toán:

Để đội bóng hoạt động tốt nhất, các đội bóng thường tìm kiếm những cầu thủ mới có nhiều tiềm năng, hứa hẹn có khả năng phát triển để đáp ứng mục tiêu và định hình tương lai của đội. Có ba lựa chọn khác nhau để ký hợp đồng với một cầu thủ mới, đó là đi mượn từ các đội bóng khác, ký hợp đồng với cầu thủ tự do hoặc mua một cầu thủ. Nghiên cứu này tập trung vào lựa chọn mua một cầu thủ, có nghĩa là một câu lạc bộ bóng đá trả tiền bồi thường cho một câu lạc bộ khác để có thể có cầu thủ đó chơi cho họ, khoản bồi thường này được gọi là phí chuyển nhượng.

Việc đạt được sự đồng thuận về mức phí chuyển nhượng giữa câu lạc bộ mua và bán là quan trọng để duy trì cân bằng hợp lý. Sự phù hợp giữa số tiền thực tế đã trả và giá trị được đánh giá của cầu thủ là quyết định quan trọng, giúp tránh rủi ro thoái vốn lớn khi một cầu thủ không đáp ứng kỳ vọng. Nghiên cứu này nhằm mục đích khám phá khả năng dự đoán giá trị chuyển nhượng của cầu thủ, một thông tin quan trọng mà câu lạc bộ có thể sử dụng trong quá trình đàm phán, cũng như mang lại những đánh giá về giá trị của một cầu thủ trong cộng đồng bóng đá.

Có rất nhiều yếu tố để có thể định giá một cầu thủ bóng đá có thể kể đến như hiệu suất trên sân, dựa vào sự cạnh tranh của thị trường chuyển nhượng, các quy tắc của hợp đồng, ảnh hưởng của truyền thông,... nhưng đề án này đặt ra mục tiêu là xác định mức độ có thể dự đoán giá trị chuyển nhượng của cầu thủ bóng đá dựa trên kỹ năng và đặc điểm cá nhân của họ. Câu hỏi nghiên cứu chính là: "Giá trị chuyển nhượng của các cầu thủ có thể được dự đoán ở mức độ nào dựa trên kỹ năng và đặc điểm cá nhân của họ?"

3.1.2 Quy trình tổng quan các bước thực hiện:



Hình 3-1: Sơ đồ quy trình tổng quan các bước thực hiện

Các bước thực hiện:

Bước 1: Tiền xử lý dữ liệu: Trong bước này ta sẽ thực hiện các công việc bao gồm: lựa chọn và xác định các thuộc tính phù hợp và cần thiết cho mô hình, xử lý các dữ liệu thiếu, thêm thuộc tính mới, xử lý dữ liệu ngoại lai và chuẩn hóa, mã hóa dữ liệu trước khi đưa vào huấn luyện mô hình.

Bước 2: Chia dữ liệu: Ở bước chia dữ liệu này, ta sẽ sử dụng hàm `train_test-split` của `scikit-learn` để chia dữ liệu thành 80% dữ liệu huấn luyện và 20% dữ liệu kiểm tra.

Bước 3: Xây dựng các mô hình dự đoán khác nhau, huấn luyện với nhiều tập tham số khác nhau

Bước 4: Điều chỉnh siêu tham số sử dụng Bayesian Optimization kết hợp với cross validation. Khi chia dữ liệu thành tập huấn luyện và tập kiểm tra, ta chia tập huấn luyện thành các folds sau đó tiến hành tìm siêu tham số trong khoảng giá trị đã định nghĩa sẵn. Khi chia các folds này, ví dụ chia thành 5 folds thì lần một fold 1 sẽ để kiểm tra và các fold từ 2-5 sẽ để huấn luyện, lần 2 fold 2 sẽ để kiểm tra và các fold còn lại sẽ để huấn luyện.

Bước 5: Với mô hình cùng các tham số đã tìm được ở bước 4 ta sẽ đánh giá trên tập kiểm tra ban đầu với các độ đo MAE, MSE, RMSE, R2.

3.1.3 Dữ liệu:

3.1.3.1 Lựa chọn dữ liệu cầu thủ:

Hệ thống trò chơi điện tử EA Sports FIFA được phát triển bởi Electronic Arts (EA). Nó cung cấp sẵn các thông tin cá nhân, về các kỹ năng chơi bóng và đặc điểm của hơn 17,000 cầu thủ bóng đá trên thế giới. Thông tin này có sẵn trên trang web chính thức của họ. EA Sports tích hợp một mạng lưới tuyển trạch viên ngoài đời thực rộng lớn, tham gia trực tiếp các sự kiện và xem trực tiếp các trận đấu để đánh giá và cập nhật điểm kỹ năng cho từng người chơi trong FIFA. Thực tế, điểm số của cầu thủ có thể biến động theo mùa giải thực tế. Ví dụ: nếu một cầu thủ trẻ thể hiện sự xuất sắc trong việc sút và rê bóng, điểm của anh ấy sẽ tăng trong phiên bản tiếp theo của FIFA. Ngược lại, nếu một cầu thủ lớn tuổi mất đi tốc độ hoặc thể lực, điều này cũng sẽ được phản ánh qua điểm số [1]. Quá trình này giúp đảm bảo tính mạch lạc và tính đại diện cho bộ dữ liệu.

Các nghiên cứu trước đây về giá trị của các cầu thủ bóng đá chuyên nghiệp, dữ liệu từ các trò chơi điện tử đã được sử dụng.

- Nghiên cứu của Kirschstein và Liebscher (2019) [2] tập trung vào việc sử dụng kỹ thuật học máy để dự đoán giá trị thị trường của cầu thủ bóng đá. Họ đã phát triển một mô hình dựa trên các biến số kỹ năng từ FIFA 16 bằng cách so sánh với giá trị thị trường thực tế từ transfermarkt.com, nghiên cứu này cung cấp thông tin quan trọng về khả năng dự đoán giá trị cầu thủ thông qua mô hình học máy và dữ liệu từ trò chơi FIFA.
- Trong nghiên cứu của Yiğit et al. (2020) [3], họ đã sử dụng mô hình học máy, bao gồm Random Forests, Gradient Boosting, và Ridge And Lasso Regression, dựa trên kỹ năng của cầu thủ trong trò chơi điện tử Football Manager để dự đoán giá trị thị trường. Kết quả chỉ ra rằng việc kết hợp dữ liệu từ Football Manager với giá trị thị trường từ transfermarkt.com cung cấp dự đoán chính xác hơn về giá trị cầu thủ.
- Behravan và Razavi (2021) [4] đã sử dụng dữ liệu FIFA 20 để dự đoán giá trị thị trường, với nhận định rằng vị trí của cầu thủ trên sân cần được xem xét. Họ cũng nhấn mạnh tầm quan trọng của xếp hạng tổng thể, cho biết nó phản ánh trình độ kỹ năng hiện tại của người chơi.

Dựa trên các dẫn chứng và các nghiên cứu đã nêu trên, có thể kết luận rằng việc sử dụng dữ liệu cầu thủ từ trò chơi điện tử EA Sports FIFA là một quyết định có lý. Các nghiên cứu này đã chứng minh rằng thông tin từ trò chơi này, đặc biệt là về kỹ năng và đánh giá của cầu thủ, có thể đóng vai trò quan trọng trong việc dự đoán giá trị thị trường của họ. Sự kết hợp giữa dữ liệu từ FIFA và các thông tin thị trường thực tế mang lại kết quả dự đoán chính xác hơn, cũng như sự cân nhắc đối với các yếu tố khác.

3.1.3.2 Mô tả dữ liệu cầu thủ:

Bộ dữ liệu này được thu thập từ trang web sofifa.com từ mùa giải 2014/2015 – 2021/2022 (tương ứng với FIFA15 – FIFA22). Bộ dữ liệu SOFIFA bao gồm các thuộc tính về khả năng, hồ sơ và vị trí như một loại thuộc tính trong trò chơi được đo lường định lượng dựa trên hiệu suất thực tế của cầu thủ. Thuộc tính kỹ năng được mô tả bằng dữ liệu hiển thị các số liệu thống kê liên quan đến bóng đá của cầu thủ từ 1 đến 99 (Bảng 3-2), với tổng cộng 34 thuộc tính về khả năng. Ngoài ra, đề án này phân loại 34 thuộc tính về kỹ năng thành bảy loại thuộc tính khả năng như ‘attacking’, ‘skill’, ‘movement’, ‘power’, ‘mentality’, ‘defending’, và ‘goalkeeping’ (Hình 3-2). Dữ liệu này cũng cung cấp các thuộc tính hồ sơ là dữ liệu thực tế của cầu thủ bóng đá bao gồm như: tên, tuổi, quốc tịch, chiều cao, cân nặng, câu lạc bộ đang chơi,... (Bảng 3-3). Bộ dữ liệu cung cấp một vị trí của thủ môn, và lên đến ba vị trí cho tiền đạo, hậu vệ và tiền vệ (Bảng 3-1).

Loại vị trí	Vị trí viết tắt	Tên vị trí	Tên tiếng việt
Attacker	ST	Striker	Tiền đạo
	CF	Centrer Forward	Tiền đạo trung tâm
	LW	Left Winger	Tiền đạo cánh trái
	RW	Right Winger	Tiền đạo cánh phải
Midfielder	CAM	Center Attacking Midfielder	Tiền vệ trung tâm tấn công
	CM	Center Midfielder	Tiền vệ trung tâm
	LM	Left Midfielder	Tiền vệ cánh trái
	RM	Right Midfielder	Tiền vệ cánh phải
	CDM	Center Defensive Midfielder	Tiền vệ trung tâm phòng ngự
Defender	CB	Center Back	Trung vệ trung tâm
	LB	Left Back	Hậu vệ cánh trái
	RB	Right Back	Hậu vệ cánh phải
Goalkeeper	GK	Goalkeeper	Thủ môn

Bảng 3-1: Mô tả các vị trí của cầu thủ

Kim Min Jae

FIFA 22

Aug 18, 2022

[In game](#)
[In real life](#)
[Change log](#)
[Related squads](#)
[Customized](#)
[Prime](#)
[Random](#)

Search player ...

Best position **CB**
Best overall **79**

김민재 金敏在

CB

24y.o. (Nov 15, 1996) 190cm / 6'3" 81kg / 179lbs

77

Overall rating

83

Potential

€15M

Value

€37K

Wage

Like (336)

Dislike (55)

Follow (847)

History version...

Customize

Calculator

Profile

Player specialties

Club

Preferred foot **Right**

3 ★ Skill moves

3 ★ Weak foot

1 ★ International reputation

Work rate **Medium/ High**

Body type **Normal (185+)**

Real face **No**

Release clause **€31.5M**

ID 237086

Fenerbahçe

Süper Lig

75 ★★★★★

Position **LCB**

Kit number **3**

Joined **Aug 13, 2021**

Contract valid until **2025**

Layout 1 2 3

Attacking

35 Crossing

23 Finishing

73 Heading accuracy

72 Short passing

24 Volleys

Skill

57 Dribbling

34 Curve

23 FK Accuracy

67 Long passing

66 Ball control

Movement

75 Acceleration

84 Sprint speed

64 Agility

74 Reactions

65 Balance

Power

50 Shot power

77 Jumping

75 Stamina

87 Strength

37 Long shots

Mentality

81 Aggression

77 Interceptions

48 Att. Position

57 Vision

30 Penalties

67 Composure

Defending

77 Defensive awareness

80 Standing tackle

75 Sliding tackle

Goalkeeping

13 GK Diving

7 GK Handling

11 GK Kicking

14 GK Positioning

11 GK Reflexes

Traits

Dives into tackles (AI)

Hình 3-2: Dữ liệu cầu thủ trên web SoFIFA

19

Loại kỹ năng	Thuộc tính chi tiết	Mô tả	Phạm vi giá trị
Attack	Crossing	Chuyền bóng	1-99
	Finishing	Kỹ thuật dứt điểm	1-99
	Heading Accuracy	Độ chính xác đánh đầu	1-99
	Short Passing	Chuyền ngắn	1-99
	Volleys	Kỹ thuật vô lê	1-99
Skill	Dribbling	Dẫn bóng	1-99
	Curve	Độ cong khi chuyền	1-99
	Free Kick Accuracy	Độ chính xác sút phạt	1-99
	Long Passing	Chuyền xa	1-99
	Ball Control	Kiểm soát bóng	1-99
Movement	Acceleration	Tăng tốc	1-99
	Agility	Linh hoạt	1-99
	Sprint Speed	Tốc độ nhanh	1-99
	Reactions	Phản ứng	1-99
	Balance	Cân bằng	1-99
Power	Shot Power	Lực sút	1-99
	Jumping	Nhảy cao	1-99
	Stamina	Thể lực	1-99
	Strength	Sức mạnh	1-99
	Long Shots	Sút xa	1-99
Mentality	Aggression	Quyết tâm	1-99
	Penalties	Phạt đền	1-99
	Positioning	Vị trí	1-99
	Interceptions	Chặn bóng	1-99
	Vision	Tầm nhìn	1-99
	Composure	Bình tĩnh	1-99
Defending	Marking	Theo kèm đối thủ	1-99
	Sliding Tackle	Pha vào bóng trượt	1-99
	Standing Tackle	Pha vào bóng đứng	1-99
Goalkeeping	Positioning	Chọn vị trí	1-99
	Diving	Ném mình	1-99
	Handling	Xử lý bóng	1-99
	Kicking	Đá bóng	1-99
	Reflexes	Phản xạ	1-99

Bảng 3-2: Mô tả phạm vi giá trị của các thuộc tính kỹ năng

Tên thuộc tính	Mô tả thuộc tính
sofifa_id	ID của sofifa.com
player_url	URL tới hồ sơ người chơi trên sofifa.com
short_name	Tên viết tắt của cầu thủ
long_name	Tên đầy đủ của cầu thủ
age	Tuổi của cầu thủ
dob	Ngày sinh của cầu thủ
height_cm	Chiều cao của cầu thủ tính bằng cm
weight_kg	Trọng lượng của cầu thủ tính bằng kg
nationality	Quốc tịch của cầu thủ
club_name	Tên câu lạc bộ của cầu thủ
league_name	Giải đấu mà câu lạc bộ tham gia
league_rank	Thứ hạng của giải đấu mà câu lạc bộ tham gia
overall	Đánh giá chung của cầu thủ
potential	Đánh giá tiềm năng, tương lai của cầu thủ
value_eur	Giá trị cầu thủ theo FIFA
wage_eur	Mức lương cầu thủ theo FIFA
player_positions	Vị trí của cầu thủ
preferred_foot	Chân thuận của cầu thủ
international_reputation	Danh tiếng quốc tế của cầu thủ
weak_foot	Mức độ chân yếu của cầu thủ (1-5)
skill_moves	Cấp độ di chuyển kỹ năng của cầu thủ (1-5)
work_rate	Tốc độ làm việc của cầu thủ
body_type	Kiểu cơ thể của cầu thủ
real_face	Khuôn mặt của cầu thủ
release_clause_eur	Điều khoản giải phóng cầu thủ bằng euro theo FIFA
player_tags	Thẻ cầu thủ trong FIFA
team_position	Vị trí với đội của họ trong FIFA
team_jersey_number	Số áo đấu với đội của họ trong FIFA
loaned_from	Cầu thủ mượn từ đâu
joined	Khi một cầu thủ gia nhập câu lạc bộ của họ
contract_valid_until	Hiệu lực hợp đồng của cầu thủ
nation_position	Vị trí của trên đội tuyển quốc gia
nation_jersey_number	Số áo đấu trên đội tuyển quốc gia

Bảng 3-3: Mô tả các thuộc tính trong bộ dữ liệu

3.1.3.3 Lựa chọn dữ liệu giá chuyển nhượng:

Transfermarkt là một trang web chuyên về thống kê và thông tin liên quan đến thị trường chuyển nhượng, cũng như thông tin về cầu thủ, câu lạc bộ, và các giải đấu bóng đá trên

khắp thế giới. Đây là một nguồn dữ liệu quan trọng mà nhiều chuyên gia bóng đá, đội quản lý, và người hâm mộ thường sử dụng để theo dõi và đánh giá giá trị chuyển nhượng của cầu thủ. Sự độc lập và liên tục của dữ liệu là điểm mạnh của Transfermarkt. Với hàng ngàn cầu thủ và câu lạc bộ được cập nhật mỗi ngày, trang web này giúp người hâm mộ và chuyên gia theo dõi những thay đổi trên thị trường chuyển nhượng một cách chính xác. Thông tin chi tiết về kỹ năng, hiệu suất, hợp đồng, và nhiều khía cạnh khác của cầu thủ cung cấp một cái nhìn toàn diện

Có nhiều bằng chứng chứng minh sự đáng tin cậy của dữ liệu từ Transfermarkt. Thương vụ chuyển nhượng thực tế thường được cập nhật nhanh chóng và chính xác trên trang web này. Sự đồng thuận và tin tưởng từ cộng đồng bóng đá là một lý do khác chứng minh sự uy tín của trang web. Sự cam kết với sự đổi mới và cập nhật liên tục cũng làm cho Transfermarkt trở thành một nguồn dữ liệu linh hoạt và phản ánh chính xác xu hướng thị trường.

Một số nghiên cứu đã sử dụng dữ liệu từ Transfermarkt như:

- Barbuscak (2018) [5] đã nghiên cứu giá trị thị trường của các cầu thủ bóng đá bằng cách sử dụng dữ liệu từ transfermarkt.com. Barbuscak đã chạy một phép hồi quy tuyến tính, cho thấy tác động đáng kể của nhiều biến số lên giá trị thị trường, chẳng hạn như số năm còn lại của hợp đồng.
- Poli và cộng sự (2022) [6] cố gắng dự đoán phí chuyển nhượng cầu thủ bằng mô hình hồi quy tuyến tính bội. Họ dùng transfermarkt.com và các nguồn dữ liệu uy tín trong thế giới bóng đá để xác thực phí chuyển nhượng trả cho cầu thủ.

3.1.3.4 Mô tả bộ dữ liệu giá chuyển nhượng:

Dữ liệu về giá chuyển nhượng của cầu thủ được thu thập từ trang web Transfermarkt trong khoảng thời gian từ mùa giải 2014/2015 – 2022/2023 để phù hợp với dữ liệu cầu thủ đã thu thập ở trang web sofifa.com.

Compact		Detailed					
Competition	Country	Clubs	Player ↑	Avg. age ↑	Foreigners ↑	Forum	Total value ↑
First Tier							
Premier League		20	570	26.5	68.2 %		€10.99bn
LaLiga		20	498	27.6	42.8 %		€4.90bn
Serie A		20	566	26.5	63.1 %		€4.62bn
Bundesliga		18	508	26.1	48.6 %		€4.41bn
Ligue 1		18	493	25.4	59.4 %		€3.70bn
Liga Portugal		18	501	26.1	60.1 %		€1.41bn
Eredivisie		18	489	24.6	46.8 %		€1.25bn
Süper Lig		20	593	26.8	49.1 %		€1.09bn
Jupiler Pro League		16	438	24.8	58.7 %		€960.83m
Premier Liga		16	424	26.2	35.8 %		€840.48m
Super League 1		14	419	27.5	59.7 %		€421.08m
Bundesliga		12	339	24.8	39.2 %		€418.90m
Premiership		12	315	26.5	63.2 %		€339.36m
Superliga		12	331	25.1	43.8 %		€312.13m
Super League		12	362	25.1	56.6 %		€310.85m
Premier Liga		16	423	25.6	16.8 %		€309.73m
Ekstraklasa		18	555	25.6	37.5 %		€280.13m
Super liga Srbije		16	498	24.7	21.7 %		€266.10m
Fortuna Liga		16	457	26.1	29.8 %		€250.12m
Allsvenskan		16	437	24.9	35.0 %		€240.57m
SuperSport HNL		10	299	25.0	35.5 %		€232.80m
Eliteserien		16	414	24.3	25.1 %		€221.10m
SuperLiga		16	495	25.9	36.8 %		€219.28m
efbet Liga		16	381	25.6	43.8 %		€171.77m
Protathlima Cyta		14	418	26.9	64.6 %		€160.28m

Hình 3-3: Các giải đấu có tổng số tiền chi cho thị trường chuyển nhượng cao nhất

Dựa vào (Hình 3-3) ta có thể thấy các giải đấu đã chi bao nhiêu tiền cho thị trường chuyển nhượng, giải đấu có cấp độ càng cao thì mức chi tiêu cho việc mua, bán các cầu thủ cũng tỷ lệ thuận theo. Do đó để có thể có được nhiều dữ liệu nhất tôi sẽ lựa chọn 20 giải đấu giảm dần theo số tiền, chi tiết các giải đấu như sau (Bảng 3-4)

Tên các giải đấu	Mô tả
Premier League	Giải bóng đá hàng đầu ở Anh
La Liga	Giải bóng đá hàng đầu ở Tây Ban Nha.
Bundesliga	Giải bóng đá hàng đầu ở Đức.
Serie A	Giải bóng đá hàng đầu ở Ý
Ligue 1	Giải bóng đá hàng đầu ở Pháp
POR Liga Portugal	Giải bóng đá hàng đầu ở Bồ Đào Nha.
TUR Süper Lig	Giải bóng đá hàng đầu ở Thổ Nhĩ Kỳ.
NLD Eredivisie	Giải bóng đá hàng đầu ở Hà Lan.
BEL Jupiler Pro League	Giải bóng đá hàng đầu ở Bỉ.
RUS Premier Liga	Giải bóng đá hàng đầu ở Nga.
GRE Super League 1	Giải bóng đá hàng đầu ở Hy Lạp.
AUS Bundesliga	Giải bóng đá hàng đầu ở Áo.
Scottish Premiership	Giải bóng đá hàng đầu ở Scotland.
Super League	Giải bóng đá hàng đầu ở Thụy Sĩ.
Premier Liga	Giải bóng đá hàng đầu ở Ukraine.
Ekstraklasa	Giải bóng đá hàng đầu ở Ba Lan.
Superliga	Giải bóng đá hàng đầu ở Đan Mạch.
Super liga Srbije	Giải bóng đá hàng đầu ở Serbia.
Allsvenskan	Giải bóng đá hàng đầu ở Thụy Điển.
SuperSport HNL	Giải bóng đá hàng đầu ở Croatia.

Bảng 3-4: 20 giải đấu bóng đá với tổng số tiền chi cho việc chuyển nhượng cao nhất

Như đã đề cập ở trên, đồ án này tập trung vào việc mua một cầu thủ, nghĩa là có tồn tại giao dịch giữa 2 câu lạc bộ. Do đó nên các dữ liệu chuyển nhượng tự do hay đi mượn sẽ không được thu thập trong bộ dữ liệu này (Hình 3-4)

 ARSENAL FC							
In	Age	Nat.	Position	Market values	Left	Fee	
Ben White	23		Right-Back	€55.00m	Brighton	€58.50m	
Martin Ødegaard	22		Attacking Midfield	€90.00m	Real Madrid	€35.00m	
Aaron Ramsdale	23		Goalkeeper	€28.00m	Sheff Utd	€28.00m	
Takehiro Tomiyasu	22		Right-Back	€30.00m	Bologna	€18.60m	
Albert Sambi Lokonga	21		Central Midfield	€12.00m	RSC Anderlecht	€17.50m	
Nuno Tavares	21		Left-Back	€16.00m	Benfica	€8.00m	
Auston Trusty	23		Centre-Back	€8.00m	Colorado	€1.80m	
Arthur Okonkwo	19		Goalkeeper	€1.80m	Arsenal U23	-	
Dejan Iliev	26		Goalkeeper	€300k	SKF Sered	End of loan Dec 31, 2021	
William Saliba	20		Centre-Back	€75.00m	OGC Nice	End of loan Jun 30, 2021	
Sead Kolasinac	28		Centre-Back	€8.00m	FC Schalke 04	End of loan Jun 30, 2021	
Lucas Torreira	25		Defensive Midfield	€16.00m	Atlético Madrid	End of loan Jun 30, 2021	
Mattéo Guendouzi	22		Central Midfield	€20.00m	Hertha BSC	End of loan Jun 30, 2021	
Average age of arrivals: 22.7				Total market value of arrivals: €360.10m		Expenditure: €167.40m	

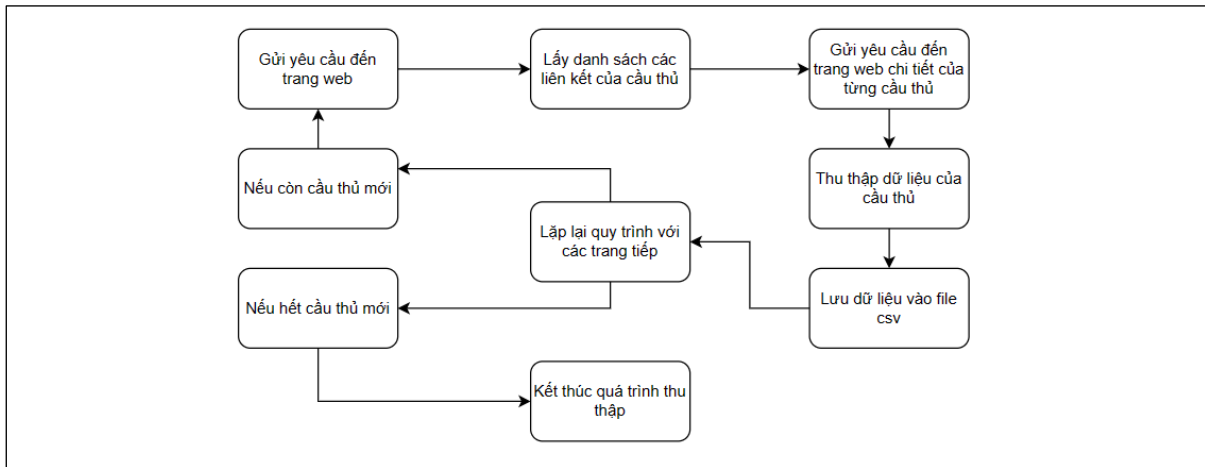
Hình 3-4: Dữ liệu chuyển nhượng của Transfermarkt

Tên các thuộc tính	Mô tả
Club name	Câu lạc bộ của cầu thủ
Player name	Tên của cầu thủ
Age	Tuổi của cầu thủ
Position	Vị trí của cầu thủ
Nationality	Quốc tịch của cầu thủ
Dealing_club	Câu lạc bộ mua
Fee	Phí chuyển nhượng trả cho cầu thủ
Movement	Chuyển khoản đến hoặc đi
Window	Kỳ chuyển nhượng
League name	Tên giải đấu mà cầu thủ thi đấu
Year	Năm chuyển nhượng
Season	Mùa chuyển nhượng

Bảng 3-5: Mô tả các thuộc tính trong bộ dữ liệu giá chuyển nhượng

3.2 Thu thập dữ liệu:

3.2.1 Thu thập dữ liệu cầu thủ từ trang web sofifa:



Hình 3-5: Quy trình thu thập dữ liệu Sofifa























Dựa vào mô tả dữ liệu đã nói ở trên, dữ liệu cầu thủ sẽ được thu thập từ FIFA15 – FIFA22 (tương ứng với mùa giải 2014/2015 – 2021/2022) tôi sẽ xây dựng các bước thu thập dữ liệu cầu thủ như sau:

Bước 1: Gửi yêu cầu đến trang web

Khi bắt đầu quá trình thu thập dữ liệu từ trang web Sofifa.com, ta đối mặt với thách thức là trang web này áp đặt các hạn chế và chặn yêu cầu, thường trả về mã lỗi 403 khi phát hiện các yêu cầu tự động. Để vượt qua giới hạn này, chúng ta sử dụng thư viện Cloudscraper. Thay vì sử dụng thư viện requests thông thường, Cloudscraper giúp giả mạo thông tin người dùng và thực hiện các bước phức tạp để tránh bị chặn bot.

Bước 2: Lấy danh sách liên kết các cầu thủ

Trên trang web Sofifa, thông tin chi tiết của từng cầu thủ không được hiển thị trực tiếp trên danh sách chính. Để có được dữ liệu chi tiết, quá trình crawl bao gồm việc lấy các đường liên kết của từng cầu thủ từ trang web. Mỗi đường liên kết này đại diện cho một cầu thủ cụ thể và chứa thông tin chi tiết về cầu thủ đó. Chúng ta sẽ lấy các đường liên kết này và sau đó tiếp tục truy cập từng đường liên kết để thu thập dữ liệu chi tiết.

Name	Age	O...	Po...	Team & Contract	Value	Wage	Total ...
 Z. Suzuki ● GK	20	66 +2	75 +2	 Sint-Truiden 2020 ~ 2027	€1.6M	€2K	1081
 Vitor Roque ST RW LW	18	76 +1	88 +1	 FC Barcelona 2024 ~ 2031	€17.5M	€44K	1824
 T. Almada CAM CM CF	22	80 +1	88 +1	 Atlanta United 2022 ~ 2025	€47.5M	€11K	2058
 A. Nusa LM LW RM	18	71 +1	87 +1	 Club Brugge 2021 ~ 2027	€4.8M	€5K	1836
 N. Irankunda RM LM	17	64 +1	85 +1	 Adelaide United 2021 ~ 2024	€1.8M	€500	1803
 V. Boniface ST	22	80 +2	86 +1	 Bayer 04 Leverkusen 2023 ~ 2028	€34M	€56K	1930
 L. Suárez ST	36	84 +1	84 +1	 Inter Miami 2024 ~ 2024	€14.5M	€14K	2212
 L. Bergvall CM	17	64 +1	84 +1	 Djurgården 2023 ~ 2025	€1.6M	€500	1719
 R. Lewis RB CDM	18	75 +1	88 +1	 Manchester City 2022 ~ 2028	€12.5M	€23K	1867
 W. Zaire-Emery CM CDM	17	79 +1	90 +1	 Paris Saint Germain 2022 ~ 2025	€36M	€10K	2102
 M. O'Riley CM CDM CAM	22	77 +1	85 +1	 Celtic 2022 ~ 2027	€23.5M	€45K	2082

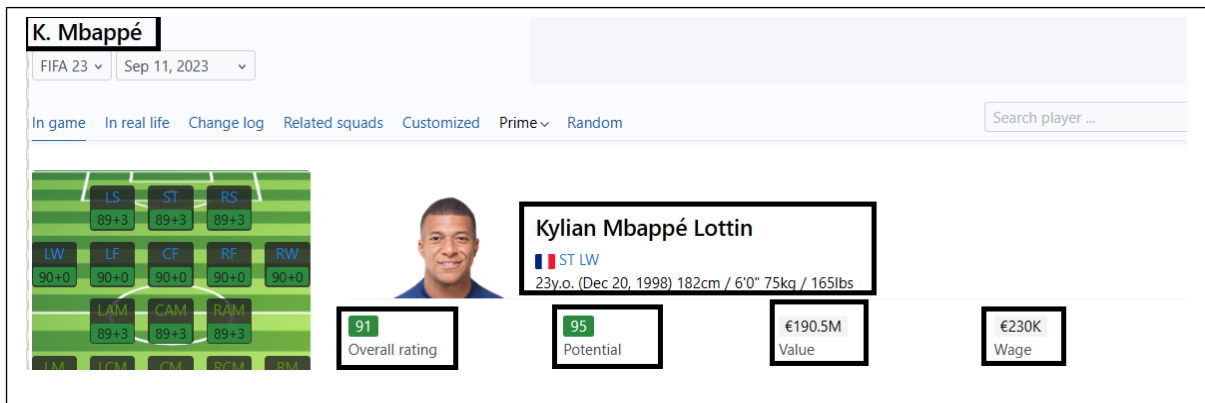
Hình 3-6: Giao diện chứa đường liên kết của cầu thủ

Bước 3: Lấy dữ liệu chi tiết của cầu thủ

Sau khi đã có được đường liên kết đến trang web chứa thông tin chi tiết cầu thủ ta sẽ tiến hành thu thập dữ liệu. Trong bước này ta sẽ sử dụng BeautifulSoup để phân tích và lấy ra giá trị của thuộc tính dựa trên cấu trúc HTML của trang web.

Với các thuộc tính như short_name, player_name, nationality, overall, potential, value và wage ta dùng phương thức 'find' để định vị thẻ và class chứa các dữ liệu sau đó loại bỏ các khoảng trắng sử dụng 'strip()'.

Với dữ liệu là một biểu thức văn bản, ta sử dụng Regular Expression (Regex) để có thể lấy được dữ liệu chính xác. Trong bài tôi đã sử dụng để lấy dữ liệu từ biểu thức có dạng như sau: "32y.o. (Aug 21, 1988) 185cm / 6'1" 81kg / 179lbs" để lấy ra các thông tin về tuổi, ngày tháng năm sinh, vị trí, chiều cao (cm), cân nặng (kg).



Hình 3-7: Dữ liệu thông tin cá nhân cầu thủ

Đối với các thuộc tính có cùng cấu trúc HTML giống nhau tôi sử dụng hàm `find_all` để tìm tất cả các thẻ chứa dữ liệu, sau đó tôi sẽ lặp qua từng phần tử của `find_all` để lấy dữ liệu. Các dữ liệu thuộc trường hợp này (Hình 3-8). Chúng ta sẽ ví dụ cách để thu thập các dữ liệu này như sau:

```
card_divs = soup.find_all('div', class_='card')
for card_div in card_divs:
    ul_element = card_div.find('ul', class_='pl')
    if ul_element:
        li_elements = ul_element.find_all('li')
        for li in li_elements:
            span_tag = li.find('span', class_='bp3-tag')
            attribute_name_span = li.find('span',
                                           role='tooltip')
            if span_tag and attribute_name_span:
                attribute_value =
                span_tag.text.strip()
                attribute_name =
                attribute_name_span.text.strip()
                player_data.append((attribute_name,
                                     attribute_value))
            else:
                pass
    else:
        pass
```

Mã nguồn trích xuất thông tin chi tiết về cầu thủ từ các phần tử `div` có class `'card'`. Đối với mỗi phần tử này, ta tìm thẻ `'ul'` với class `'pl'` để lấy thông tin cầu thủ.

Trong mỗi ‘ul’, ta tìm tất cả các phần tử ‘li’ để trích xuất dữ liệu. Trong mỗi phần tử ‘li’, ta kiểm tra thẻ ‘span’ và thẻ ‘span’ có vai trò ‘tooltip’. Nếu cả hai thẻ này tồn tại, ta lấy tên thuộc tính và giá trị tương ứng, nếu không ta bỏ qua phần tử đó. Làm tương tự với các dữ liệu khác.

Profile Preferred foot Right 5 ★ Skill moves 4 ★ Weak foot 4 ★ International reputation Work rate High/ Low Body type Unique Real face Yes Release clause €366.7M ID 231747	Player specialities #Speedster #Dribbler #Acrobat #Clinical finisher #Complete forward	Club Paris Saint Germain Ligue 1 84 ★★★★★ Position LS Kit number 7 Joined Jul 1, 2018 Contract valid until 2024	National team France Friendly International 83 ★★★★★ Position LW Kit number 10
---	--	--	---

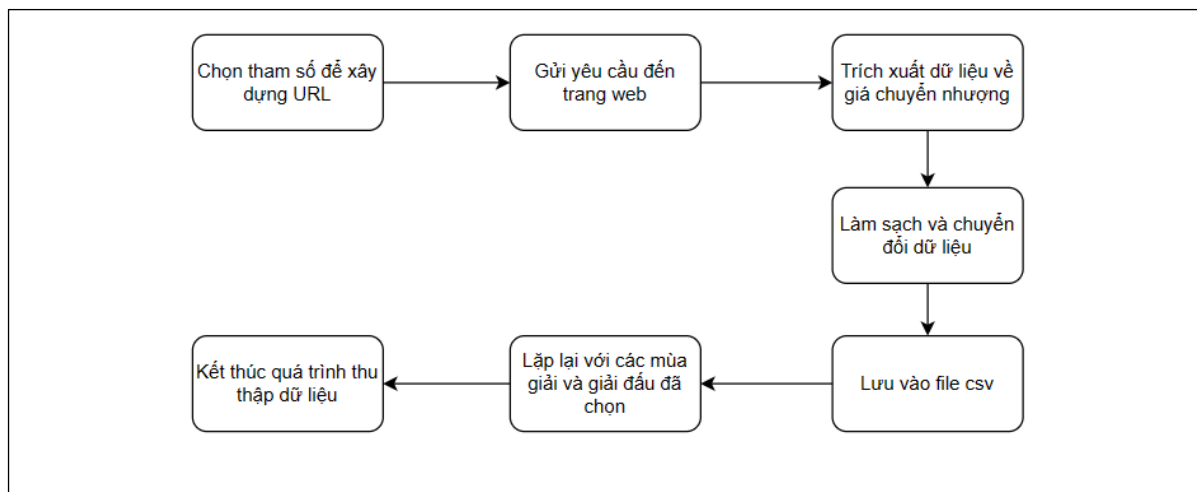
Layout 1	2	3
----------	---	---

Attacking	Skill	Movement	Power
78 Crossing	93 Dribbling	97 Acceleration	88 Shot power
93 Finishing	80 Curve	97 Sprint speed	77 Jumping
72 Heading accuracy	69 FK Accuracy	93 Agility	87 Stamina
85 Short passing	71 Long passing	93 Reactions	76 Strength
83 Volleys	91 Ball control	81 Balance	82 Long shots
Mentality	Defending	Goalkeeping	Traits
64 Aggression	26 Defensive awareness	13 GK Diving	Solid player
38 Interceptions	34 Standing tackle	5 GK Handling	Flair
92 Att. Position	32 Sliding tackle	7 GK Kicking	Speed dribbler (AI)
83 Vision		11 GK Positioning	Outside foot shot
84 Penalties		6 GK Reflexes	Technical dribbler (AI)
88 Composure			

Hình 3-8: Dữ liệu về đặc điểm và chỉ số kỹ năng của cầu thủ

Bước 4: Sau khi hoàn tất quá trình trích xuất dữ liệu từ trang web Sofifa, chúng ta chuyển sang bước xử lý theo năm. Đầu tiên, cần xác định các mã số được đánh cho từng năm trên trang web (Hình 3-9). Điều này sẽ giúp xây dựng đường dẫn đến trang danh sách cầu thủ của từng năm khi chạy mã nguồn.

Với mỗi năm, chúng ta sẽ lặp qua các trang, mỗi trang chứa thông tin về 60 cầu thủ. Bằng cách này, ta có thể lấy dữ liệu cho tất cả các cầu thủ trên trang và di chuyển đến trang tiếp theo cho đến khi không còn dữ liệu cầu thủ nào nữa. Điều này giúp đảm bảo việc thu thập toàn bộ dữ liệu từ Sofifa theo năm.



Hình 3-11: Quy trình thu thập dữ liệu giá chuyển nhượng

Dựa vào mô tả dữ liệu đã nói ở trên, dữ liệu cầu thủ sẽ được thu thập từ mùa giải 2014/2015 – 2021/2022, tôi sẽ xây dựng các bước thu thập dữ liệu giá chuyển nhượng như sau:

Bước 1: Để xây dựng đường liên kết để có thể gửi yêu cầu ta hãy phân tích một đường dẫn sau đây:

https://www.transfermarkt.com/premier_league/transfers/wettbewerb/GB1/plus/?saigon_id=2022&s_w=&leihe=1&intern=0&intern=1

Trong đó:

- https://www.transfermarkt.com/premier_league/transfers/wettbewerb/GB1/plus/.
Với các thông tin bao gồm giải đấu, mã giải đấu và thông tin đang là về giá trị chuyển nhượng.
- `saigon_id=2022`: Tham số này chỉ định mùa giải cụ thể, trong trường hợp này là mùa giải 2022/2023.
- `s_w=`: Tham số này đại diện cho thời điểm chuyển nhượng (summer - mùa hè, winter - mùa đông).
- `leihe=1`: Tham số này chỉ định cách xử lý chuyển nhượng liên quan đến việc cho mượn và nó có các giá trị sau: Giá trị = 0 hiển thị chuyển nhượng không bao gồm các khoản vay. Giá trị = 1 hiển thị tất cả các chuyển nhượng. Giá trị = 2 chỉ hiển thị các khoản vay. Giá trị = 3 hiển thị tất cả chuyển nhượng, nhưng loại trừ các cầu thủ quay về từ khoản cho mượn.

Ta sẽ lựa chọn các tham số để có thể tạo ra các đường dẫn phù hợp với dữ liệu muốn thu thập (Hình 3-12).

```
Select league(s), e.g. '1', '3 5', '6-10' (default is top 5):
[1] ENG Premier League
[2] ESP La Liga
[3] GER Bundesliga
[4] ITA Serie A
[5] FRA Ligue 1
[6] POR Liga Portugal
[7] TUR Süper Lig
[8] NLD Eredivisie
[9] BEL Jupiler Pro League
[10] RUS Premier Liga
[11] GRE Super League 1
[12] AUS Bundesliga
[13] SCO Scottish Premiership
[14] SWI Super League
[15] UKR Premier Liga
[16] POL Ekstraklasa
[17] DNK Superliga
[18] SER Super liga Srbije
[19] SWE Allsvenskan
[20] CRO SuperSport HNL

Enter desired seasons as years (default is current season).
Years should be input as the first calendar year in a season, e.g. '2015' for the 2015/16 season.
You can input both individual years and year ranges, e.g. '1992 2004-2007'.
Error: Seasons are limited to the range 1992-2023.

Select transfer window (default is both):
[1] Both
[2] Summer
[3] Winter

Select how to handle loan transfers (default is without players back from loan):
[1] Exclude loans
[2] Include loans
[3] Loans only
[4] Without players back from loan
```

Hình 3-12: Lựa chọn tham số để tạo đường dẫn

Bước 2: Xử lý dữ liệu đầu vào: Đối với các giá trị khác, chúng tôi sẽ chỉ giữ lại một giá trị duy nhất. Trong khi đó, đối với giải đấu và mùa giải, chúng tôi sẽ chọn theo phạm vi nhất định. Phạm vi mùa giải sẽ bắt đầu từ năm 1992 đến hiện tại, và chỉ giữ lại các giải đấu nằm trong danh sách top 20. Sử dụng hàm `_parse_hyphenated_string` để phân tích thông tin về mùa bóng đá nhập từ người dùng. Hàm này sẽ chuyển đổi thông tin mùa bóng đá từ định dạng "1992-2023" thành một danh sách các năm chạy từ 1992 đến 2023. Tương tự với giải đấu cũng vậy.

Bước 3: Sau khi đã xây dựng được URL, ta sẽ sử dụng Cloudscraper để gửi yêu cầu đến trang web và sau đó sẽ sử dụng BeautifulSoup để phân tích và trích xuất dữ liệu.

Bước 4: Dữ liệu chúng ta cần thu thập nằm trong bảng của hình 3-4 của từng câu lạc bộ vì vậy ta sẽ lấy tên câu lạc bộ trước. Sử dụng `soup.select(".content-box-headline--logo a")`, mã này chọn tất cả các thẻ `<a>` nằm trong lớp CSS

`.content-box-headline-logo`. Với mỗi thẻ được chọn sẽ trích xuất tên của các câu lạc bộ tham gia vào chuyển nhượng. Kết quả được lưu vào một list.

Ta sẽ sử dụng phương thức `find-all` để chọn tất cả các thẻ `<div>` có lớp CSS là `'responsive-table'` và lưu trữ chúng trong một list. Sau đó ta dùng cơ chế cắt list để phân biệt giữa bảng IN và OUT. Tiếp đến ta sẽ lặp qua từng câu lạc bộ cùng với bảng dữ liệu chuyển nhượng tương ứng. Thêm dữ liệu của bảng IN và OUT vào các dataframe tương ứng sau đó gộp chúng lại để thành một file kết quả.

Bước 5: Xử lý các dữ liệu đặc biệt (Hình 3-13)

Xử lý phí và cho mượn: Kiểm tra xem có một số từ khóa nhất định trong `fee` và thực hiện các bước xử lý tương ứng:

- Nếu bắt đầu bằng "end of loan", cập nhật `fee` thành "\$0" và thiết lập các trạng thái về cho mượn.
- Nếu bắt đầu bằng "loan fee", loại bỏ phần "loan fee:" và thiết lập các trạng thái.
- Nếu bắt đầu bằng "loan transfer", cập nhật `fee` thành "\$0" và thiết lập các trạng thái.
- Nếu bắt đầu bằng "free transfer", cập nhật `fee` thành "\$0".

















































Xử lý các giá trị để chuyển đổi chúng thành số: Áp dụng cho các cột `market_value` và `fee`.

- Nếu giá trị là NaN, hoặc là "-", hoặc là "?", trả về NaN.
- Nếu giá trị là "0", trả về 0.
- Nếu giá trị kết thúc bằng 'm', nhân với 1 triệu.
- Nếu giá trị kết thúc bằng 'th.', nhân với 1 nghìn.

Chỉnh sửa kiểu dữ liệu và xử lý NaN:

- Dùng `assign(is_loan=False, loan_status="")` để thêm cột `is_loan` với giá trị mặc định là False và `loan_status` với giá trị rỗng cho mỗi hàng.
- Sử dụng `pd.to_numeric` để chuyển đổi cột `age` thành kiểu dữ liệu số
- Chuyển cột `season` thành kiểu dữ liệu năm từ định dạng ngày tháng và thiết lập các giá trị NaN nếu cần.

- Đối với các cột nhất định (nationality, position, short_pos, dealing_club, dealing_country), sử dụng fillna("", inplace=True) để điền giá trị rỗng bằng chuỗi trống.
- Trong cột league, thay thế các dấu gạch ngang bằng khoảng trắng và chuyển đổi thành chữ in hoa.

BRENTFORD FC							
In	Age	Nat.	Position	Market value	Left	Fee	
Nathan Collins	22		Centre-Back	€25.00m	  Wolves	€26.85m	
Kevin Schade	21		Right Winger	€25.00m	  SC Freiburg	€25.00m	
Mark Flekken	30		Goalkeeper	€12.00m	  SC Freiburg	€13.00m	
Yunus Emre Konak	18		Defensive Midfield	€1.80m	  Sivasspor	€4.50m	
Neal Maupay	27		Centre-Forward	€10.00m	  Everton	loan transfer	
Sergio Reguilón	27		Left-Back	€10.00m	  Tottenham	loan transfer	
Yegor Yarmolyuk	19		Attacking Midfield	€2.50m	  Brentford B	-	
Ryan Trevitt	20		Central Midfield	€275k	  Brentford B	-	
Joel Valencia	28		Attacking Midfield	€250k	  De Graafschap	End of loan Jun 30, 2023	
Mads Bech Sørensen	24		Centre-Back	€2.50m	  FC Groningen	End of loan Jun 30, 2023	
Sergi Canós	26		Right Winger	€2.50m	  Olympiacos	End of loan Jun 30, 2023	
Ryan Trevitt	20		Central Midfield	€275k	  Exeter City	End of loan Jan 8, 2024	
Paris Maghoma	23		Central Midfield	€350k	  Bolton	End of loan May 31, 2024	
Matthew Cox	21		Goalkeeper	€800k	  Bristol Rovers	End of loan May 31, 2024	
Fin Stevens	21		Right-Back	€450k	  Oxford United	End of loan May 31, 2024	
Mads Bidstrup	22		Central Midfield	€9.00m	  Nordsjaelland	End of loan Jun 30, 2023	
Average age of arrivals: 23.1				Total market value of arrivals: €102.70m		Expenditure: €69.35m	

Hình 3-13: Các dữ liệu đặc biệt

Bước 6: Dữ liệu sau khi thu thập sẽ được xuất vào file csv ứng với mùa giải và giải đấu bóng đá. Lập lại quy trình thu thập qua các mùa giải và giải đấu dựa vào các tham số đầu vào.

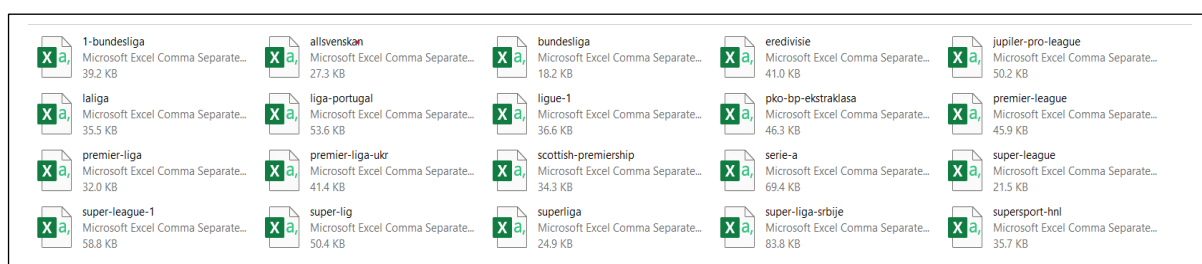
Kết quả sau quá trình thu thập như sau (Hình 3-14):

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	club	name	age	nationality	position	short_pos	market_val	dealing_cl	dealing_co	fee	movement	window	league	season	is_loan	loan_status
2	AFC Bourn	Marcos Se	25	Argentina	Centre-Bai	CB	22000000	Feyenoord	Netherland	15000000	in	summer	Premier Le	2022	FALSE	
3	AFC Bourn	Marcus Ta	23	England	Left Midfiel	LM	17000000	Middlesbrc	England	11900000	in	summer	Premier Le	2022	FALSE	
4	AFC Bourn	Joe Rothwe	27	England	Central Mi	CM	4000000	Blackburn	England	0	in	summer	Premier Le	2022	FALSE	
5	AFC Bourn	Ryan Frede	29	England	Right-Back	RB	1500000	West Ham	England	0	in	summer	Premier Le	2022	FALSE	
6	AFC Bourn	Neto	33	Brazil	Goalkeepe	GK	2500000	Barcelona	Spain	0	in	summer	Premier Le	2022	FALSE	
7	AFC Bourn	Robbie Bra	30	Ireland	Left Midfiel	LM		Preston	England	0	out	summer	Premier Le	2022	FALSE	
8	AFC Bourn	Zeno Ibsen	21	England	Centre-Bai	CB		Cambridge	England		out	summer	Premier Le	2022	FALSE	
9	AFC Bourn	Gary Cahill	36	England	Centre-Bai	CB		Without Club			out	summer	Premier Le	2022	FALSE	
10	AFC Bourn	Ilya Zabarn	20	Ukraine	Centre-Bai	CB	20000000	Dynamo Ky	Ukraine	22700000	in	winter	Premier Le	2022	FALSE	
11	AFC Bourn	Dango Oua	20	Burkina Fa	Right Wing	RW	25000000	FC Lorient	France	22500000	in	winter	Premier Le	2022	FALSE	
12	AFC Bourn	Antoine Se	23	Ghana	Centre-Foi	CF	9000000	Bristol City	England	10250000	in	winter	Premier Le	2022	FALSE	
13	AFC Bourn	Darren Rai	35	Ireland	Goalkeepe	GK		West Ham	England	0	in	winter	Premier Le	2022	FALSE	
14	Arsenal FC	Gabriel Jes	25	Brazil	Centre-Foi	CF	75000000	Man City	England	52200000	in	summer	Premier Le	2022	FALSE	
15	Arsenal FC	Fábio Vieir	22	Portugal	Attacking M	AM	25000000	FC Porto	Portugal	35000000	in	summer	Premier Le	2022	FALSE	
16	Arsenal FC	Oleksandr	25	Ukraine	Left-Back	LB	42000000	Man City	England	35000000	in	summer	Premier Le	2022	FALSE	
17	Arsenal FC	Matt Turne	28	United Stat	Goalkeepe	GK	10000000	New Engla	United Stat	5900000	in	summer	Premier Le	2022	FALSE	
18	Arsenal FC	Marquinho	19	Brazil	Right Wing	RW	10000000	São Paulo	Brazil	3500000	in	summer	Premier Le	2022	FALSE	
19	Arsenal FC	Mattéo Gu	23	France	Central Mi	CM	20000000	Marseille	France	11000000	out	summer	Premier Le	2022	FALSE	
20	Arsenal FC	Lucas Torr	26	Uruguay	Defensive	DM	15000000	Galatasar	Turkey	6000000	out	summer	Premier Le	2022	FALSE	
21	Arsenal FC	Bernd Lenc	30	Germany	Goalkeepe	GK	12000000	Fulham	England	3600000	out	summer	Premier Le	2022	FALSE	
22	Arsenal FC	Konstantin	24	Greece	Centre-Bai	CB	15000000	VfB Stuttga	Germany	3200000	out	summer	Premier Le	2022	FALSE	
23	Arsenal FC	Alexandre	31	France	Centre-Foi	CF	12000000	Olympique	France	0	out	summer	Premier Le	2022	FALSE	
24	Arsenal FC	Héctor Bel	27	Spain	Right-Back	RB	8000000	Barcelona	Spain	0	out	summer	Premier Le	2022	FALSE	
25	Arsenal FC	Jakub Kiwi	22	Poland	Centre-Bai	CB	25000000	Spezia Cal	Italy	25000000	in	winter	Premier Le	2022	FALSE	
26	Arsenal FC	Leandro Tr	28	Belgium	Left Winge	LW	35000000	Brighton	England	24000000	in	winter	Premier Le	2022	FALSE	
27	Arsenal FC	Jorginho	31	Italy	Defensive	DM	15000000	Chelsea	England	11300000	in	winter	Premier Le	2022	FALSE	

Hình 3-14: Kết quả sau quá trình thu thập giá trị chuyển nhượng

3.2.3 Hợp nhất dữ liệu:

Hai bộ dữ liệu được hợp nhất dựa trên tên người chơi và năm chuyển nhượng. Nếu một vụ chuyển nhượng diễn ra vào năm 2015 tôi sẽ sử dụng dữ liệu của người chơi đó vào năm 2015. Với dữ liệu giá chuyển nhượng sau khi thu thập, mỗi folder chứa dữ liệu theo năm hiện tại đang có 20 file csv theo 20 giải đấu (Hình 3-15) vì vậy nên tôi sẽ gộp tất cả dữ liệu 20 giải đấu lại theo từng năm.



Hình 3-15: Dữ liệu từng năm của giá chuyển nhượng

Trong tập dữ liệu giá chuyển nhượng vì chỉ tập trung vào những cầu thủ đã được trả phí nên các dữ liệu có fee = 0 hoặc fee = NaN sẽ được loại bỏ. Cả chuyển khoản đến và đi đều được thu thập, điều này dẫn đến một số giao dịch sẽ xuất hiện 2 lần, do đó nên chỉ chuyển khoản đến sẽ được giữ lại. Kết quả sau khi loại bỏ (Hình 3-16).

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	club	name	age	nationality	position	short_pos	dealing_cl	fee	movement	window	league	season	is_loan
2	Hamburge	Aaron Hun	28	Germany	Attacking M	AM	VfL Wolfsburg	3000000	in	summer	1 Bundesliga	2015	FALSE
3	Everton FC	Aaron Lennon	28	England	Right Wing	RW	Tottenham	6000000	in	summer	Premier League	2015	FALSE
4	US Palermo	Abdelhamid	25	Morocco	Centre-Back	CB	Montpellier	2000000	in	summer	Serie A	2015	FALSE
5	Club Brugge	Abdoulaye	24	Mali	Right Wing	RW	LOSC Lille	2000000	in	summer	Jupiler Pro	2015	FALSE
6	Watford FC	Abdoulaye	23	Mali	Central Mid	CM	Stade Renn	10600000	in	winter	Premier League	2015	FALSE
7	Chelsea FC	Abdul Rahr	21	Ghana	Left-Back	LB	FC Augsburg	26000000	in	summer	Premier League	2015	FALSE
8	Watford FC	Adalberto	18	Venezuela	Left Winger	LW	Udinese Calcio	10600000	in	winter	Premier League	2015	FALSE
9	Sunderland	Adam Matthews	23	Wales	Right-Back	RB	Celtic	2800000	in	summer	Premier League	2015	FALSE
10	AS Monaco	Adama Traor	20	Mali	Attacking M	AM	LOSC Lille	14000000	in	summer	Ligue 1	2015	FALSE
11	Aston Villa	Adama Traor	19	Spain	Right Wing	RW	Barcelona	10000000	in	summer	Premier League	2015	FALSE
12	Valencia CF	Adrian San	26	Brazil	Centre-Back	CB	SC Braga	9500000	in	summer	Laliga	2015	FALSE
13	Sevilla FC	Adil Rami	29	France	Centre-Back	CB	AC Milan	2500000	in	summer	Laliga	2015	FALSE
14	Bayer 04 L	Admir Meh	24	Switzerland	Second Str	SS	SC Freiburg	8000000	in	summer	1 Bundesliga	2015	FALSE
15	Málaga CF	Adnane Tig	22	Morocco	Left Winger	LW	NAC Breda	1200000	in	summer	Laliga	2015	FALSE
16	Torino FC	Afriyie Acqu	23	Ghana	Central Mid	CM	TSG Hoffe	3100000	in	summer	Serie A	2015	FALSE
17	Juventus F	Alberto Bri	23	Italy	Goalkeeper	GK	Ternana	2350000	in	summer	Serie A	2015	FALSE
18	SSC Napoli	Alberto Gra	20	Italy	Defensive	DM	Atalanta B	8000000	in	winter	Serie A	2015	FALSE
19	Swansea C	Alberto Pal	26	Italy	Centre-For	CF	Chievo Ver	8800000	in	winter	Premier League	2015	FALSE
20	FC Augsburg	Albian Ajet	18	Switzerland	Centre-For	CF	FC Basel	1000000	in	winter	1 Bundesliga	2015	FALSE
21	Hamburge	Albin Ekdal	25	Sweden	Defensive	DM	Cagliari Ca	4500000	in	summer	1 Bundesliga	2015	FALSE
22	FC Barcelo	Alex Vidal	25	Spain	Right Wing	RW	Sevilla FC	17000000	in	summer	Laliga	2015	FALSE
23	SL Benfica	Alejandro C	20	Spain	Left-Back	LB	Barcelona	2100000	in	winter	Liga Portug	2015	FALSE
24	Newcastle	Aleksandra	20	Serbia	Centre-For	CF	RSC Ander	18500000	in	summer	Premier League	2015	FALSE
25	Zenit St. P	Aleksandr	24	Russia	Second Str	SS	Dinamo M	2000000	in	winter	Premier Li	2015	FALSE
26	FC Schalke	Alessandro	21	Austria	Attacking M	AM	1.FC Nure	5000000	in	winter	1 Bundesliga	2015	FALSE
27	AC Milan	Alessio Ro	20	Italy	Centre-Back	CB	AS Roma	25000000	in	summer	Serie A	2015	FALSE
28	Central B	Alm. M. C.	25	England	Central B	CB	AS Roma	4000000	in	summer	Premier League	2015	FALSE

Hình 3-16: Kết quả sau khi lọc của bộ dữ liệu giá chuyển nhượng

Để hợp nhất 2 bộ dữ liệu với nhau một cách chính xác và đầy đủ thì phải thêm một số thuộc tính, cụ thể như sau (Hình 3-17):

- Với bộ dữ liệu giá chuyển nhượng tên của cầu thủ chứa nhiều ký tự đặc thù của mỗi quốc gia, và có tên không đầy đủ so với bộ dữ liệu cầu thủ. Vì vậy tôi đã thêm một trường là `normalized_name` để chuyển đổi tên cầu thủ về dạng bình thường.
- Với bộ dữ liệu cầu thủ tôi cũng làm tương tự với 2 cột là `player_name` và `short_name` để dễ dàng đối chiếu với bộ dữ liệu giá chuyển nhượng.

	A	B	C	D	E	F
1	club	name	normalized_name	age	nationality	
2	1. FC Köln	Luca Kilian	luca kilian	22	Germany	
3	1. FC Köln	Sargis Adamyan	sargis adamyan	29	Armenia	
4	1. FC Köln	Steffen Tigges	steffen tigges	23	Germany	
5	1. FC Köln	Eric Martel	eric martel	20	Germany	
6	1. FC Union Berlin	Jordan	jordan	26	United States	
7	1. FC Union Berlin	Jamie Leweling	jamie leweling	21	Germany	
8	1. FC Union Berlin	Morten Thorsby	morten thorsby	26	Norway	
9	1. FC Union Berlin	Josip Juranovic	josip juranovic	27	Croatia	
10	1. FC Union Berlin	Aïssa Laïdouni	aïssa laidouni	26	Tunisia	

	A	B	C	D	E	F
1	sofifa_id	player_url	short_name	player_name	player_short_name	long_name
2	158023	https://sofifa.com/player/158023/lionel-messi/210002	L. Messi	lionel messi	L. messi	Lionel Andrés Messi Cuccittini
3	20801	https://sofifa.com/player/20801/c-ronaldo-dos-santos-aveiro/210002	Cristiano Ronaldo	c ronaldo dos santos aveiro	cristiano ronaldo	Cristiano Ronaldo dos Santos Aveiro
4	200389	https://sofifa.com/player/200389/jan-oblak/210002	J. Oblak	jan oblak	j. oblak	Jan Oblak
5	188545	https://sofifa.com/player/188545/robert-lewandowski/210002	R. Lewandowski	robert lewandowski	r. lewandowski	Robert Lewandowski
6	190871	https://sofifa.com/player/190871/neymar-da-silva-santos-jr/210002	Neymar Jr	neymar da silva santos jr	neymar jr	Neymar da Silva Santos Júnior
7	192985	https://sofifa.com/player/192985/kevin-de-bruyne/210002	K. De Bruyne	kevin de bruyne	k. de bruyne	Kevin De Bruyne
8	231747	https://sofifa.com/player/231747/kylian-mbappe/210002	K. Mbappe	kylian mbappe	k. mbappe	Kylian Mbappé Lottin
9	192448	https://sofifa.com/player/192448/marc-andre-ter-stegen/210002	M. ter Stegen	marc andre ter stegen	m. ter stegen	Marc-André ter Stegen
10	203376	https://sofifa.com/player/203376/virgil-van-dijk/210002	V. van Dijk	virgil van dijk	v. van dijk	Virgil van Dijk

Hình 3-17: Kết quả sau khi thêm trường vào 2 bộ dữ liệu

Sau khi đã thêm các trường mới, tôi tiến hành kết hợp 2 bộ dữ liệu với nhau bằng cách đối chiếu giữa `normalized_name` và `nationality` của bộ giá trị chuyển nhượng với `player_name`, `player_short_name` và `nationality` của bộ dữ liệu cầu thủ. Bởi vì đây là 2 trường không thể nhầm lẫn, khi xác định được tên chính xác và quốc tịch của cầu thủ thì các thuộc tính khác sẽ khớp với nhau. Kết quả sau khi hợp nhất 2 bộ dữ liệu, tôi thu được 4158 quan sát, những dữ liệu không khớp sẽ bị bỏ qua.

3.3 Tiền xử lý dữ liệu:

3.3.1 Lựa chọn thuộc tính (Feature Selection):

Từ tập dữ liệu đã hợp nhất, có một số thuộc tính sẽ bị loại bỏ vì chúng không mang lại giá trị cho đề án này. Các thuộc tính `sofifa_id`, `player_name`, `short_name`, `long_name`, `name` vì những thuộc tính này chỉ có giá trị duy nhất, do đó không hữu ích trong việc đưa vào mô hình học máy.

Trong tập dữ liệu FIFA, có một thuộc tính được gọi là `league_rank`. Biến `league_rank` được sử dụng trong các trò chơi FIFA được phát hành từ năm 2015 đến năm 2020 để mô tả cấp độ của giải đấu và có thang điểm từ 0-4. Thuộc tính này không mang giá trị đối với các cầu thủ nên cũng sẽ bị loại bỏ.

Các thuộc tính `age`, `nationality`, `league_name`, `club_name`, `club` sẽ bị xóa bỏ vì trùng lặp với nhau. Thuộc tính `'value_eur'` thể hiện giá trị của một cầu thủ và `'wage_eur'` thể hiện mức lương của một cầu thủ theo FIFA. Vì không rõ các giá trị của các thuộc tính này được tính như thế nào và liệu chúng có thực tế hay không nên chúng bị loại trừ.

Một số biến bắt nguồn từ trò chơi FIFA không cung cấp thông tin về kỹ năng hoặc đặc điểm của cầu thủ. Các thuộc tính loại bỏ bao gồm: `player_positions`, `work_rate`, `body_type`, `real_face`, `release_clause_eur`, `player_tags`, `team_position`, `team_jersey_number`, `loaned_from`, `joined`, `nation_position`, `nation_jersey_number`, `pace`, `shooting`, `passing`, `dribbling`, `defending`, `physic`, `gk_diving`, `gk_handling`, `gk_kicking`, `gk_reflexes`, `gk_speed`, `gk_positioning`, `player_traits`, `short_pos`, `normalized_name`, `market_value`, `movement`, `is_loan`, `loan_status`, `dealing_club`, `dealing_country`, `international_reputation`, `weak_foot`, `skill_moves`.

3.3.2 Xử lý dữ liệu thiếu (Handle Missing Data):

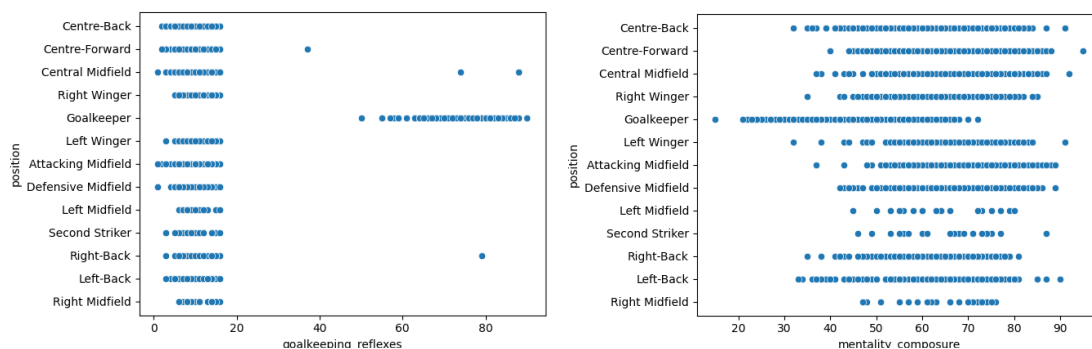
Tôi đã sử dụng `isna()` để tạo một DataFrame mới với các giá trị boolean (True nếu là NaN, False nếu không phải), sau đó gọi phương thức `any()` để xác định xem có ít nhất một giá trị True nào đó trong mỗi cột không.

```
missing_data = df1.isna().any()
missing_data[missing_data].index
```

Sử dụng `df1.isna().sum()` tôi xác định được số dữ liệu đang thiếu của mỗi thuộc tính, với ‘mentality_composure’ là 884 và ‘goalkeeping_reflexes’ là 20.

Dựa vào biểu đồ (Hình 3-18) tôi thấy rằng các giá trị của ‘goalkeeping_reflexes’ phân phối theo các vị trí của cầu thủ. Với đặc thù là khả năng của thủ môn nên các chỉ số của các vị trí khác sẽ thấp hơn so với goalkeeper, vì vậy nên không thể điền các dữ liệu còn thiếu bằng cách tính trung bình của cả cột thuộc tính đó được như vậy sẽ khiến các giá trị các vị trí khác sẽ cao bất thường và các giá trị của nhóm goalkeeper sẽ bị giảm đáng kể.

Điều này trái so với bộ dữ liệu nên tôi đã tính trung bình của ‘goalkeeping_reflexes’ theo từng vị trí sau đó điền các giá trị đó với các dữ liệu thiếu tương ứng với nhóm vị trí của chúng.



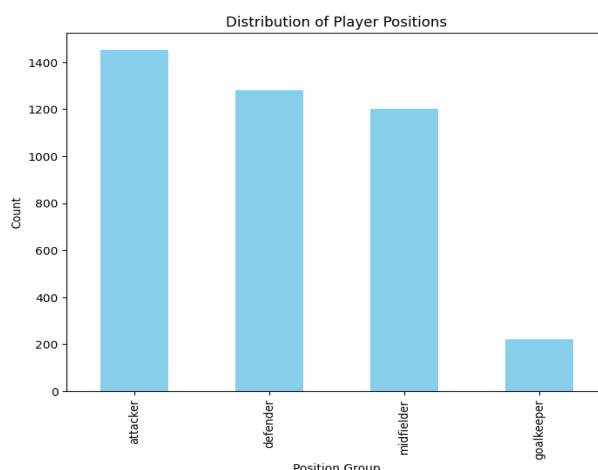
Hình 3-18: Phân bố dữ liệu của `goalkeeping_reflexes` và `mentality_composure` với các vị trí của cầu thủ

3.3.3 Kỹ thuật thuộc tính (Feature Engineering):

Vì tập dữ liệu khá nhỏ chỉ với 4158 mẫu quan sát nên các vị trí của thuộc tính ‘position’ sẽ không thể đánh giá tốt được với số mẫu ít như vậy. Do đó tôi đã nhóm các vị trí theo

khu vực họ chơi lại với nhau thành một thuộc tính là `position_group` với 4 giá trị như sau: `attacker`, `midfielder`, `defender`, `goalkeeper`.

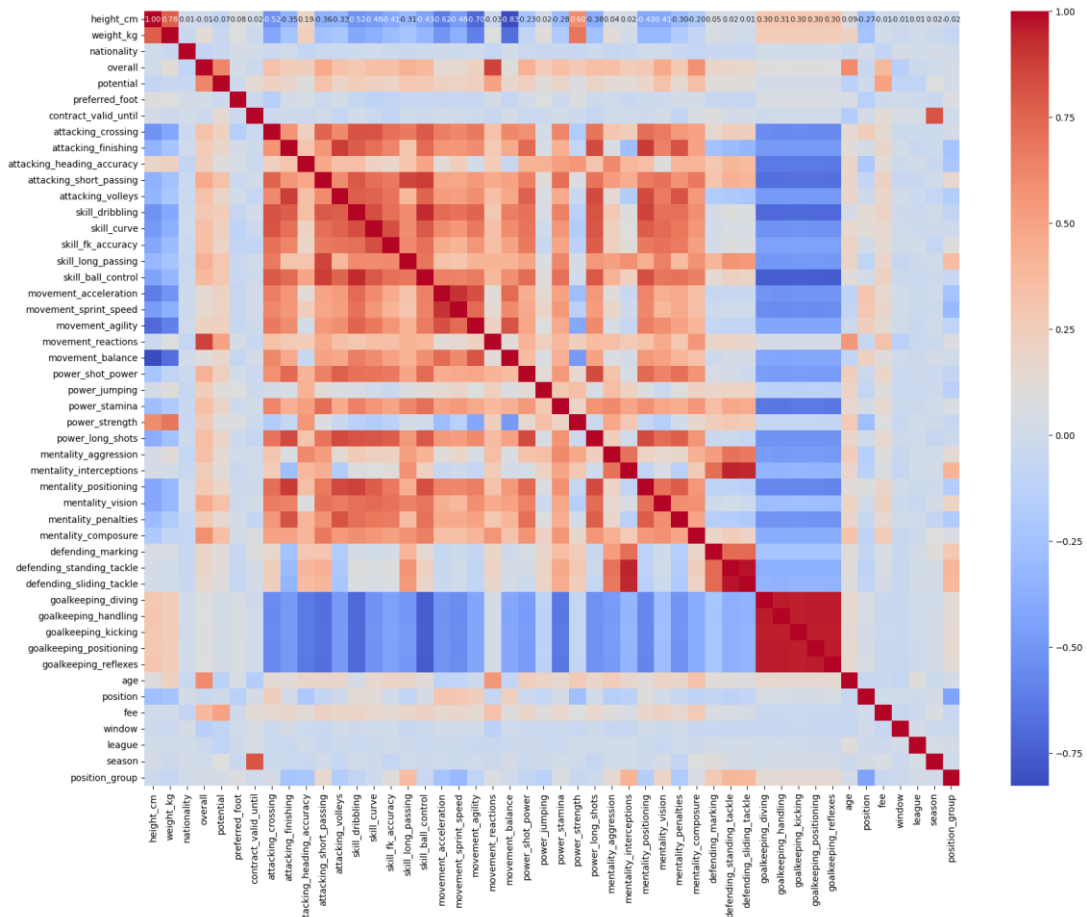
Trên thực tế thì các thủ môn sẽ có các bài đánh giá khác so với các vị trí khác trên sân. Cộng thêm số mẫu của vị trí thủ môn trong bộ dữ liệu cũng rất ít (Hình 3-19), điều này khiến cho việc so sánh thủ môn với các cầu thủ trên sân trở nên khó khăn và do đó chúng tôi sẽ loại thủ môn khỏi các mô hình chính của đề án này.



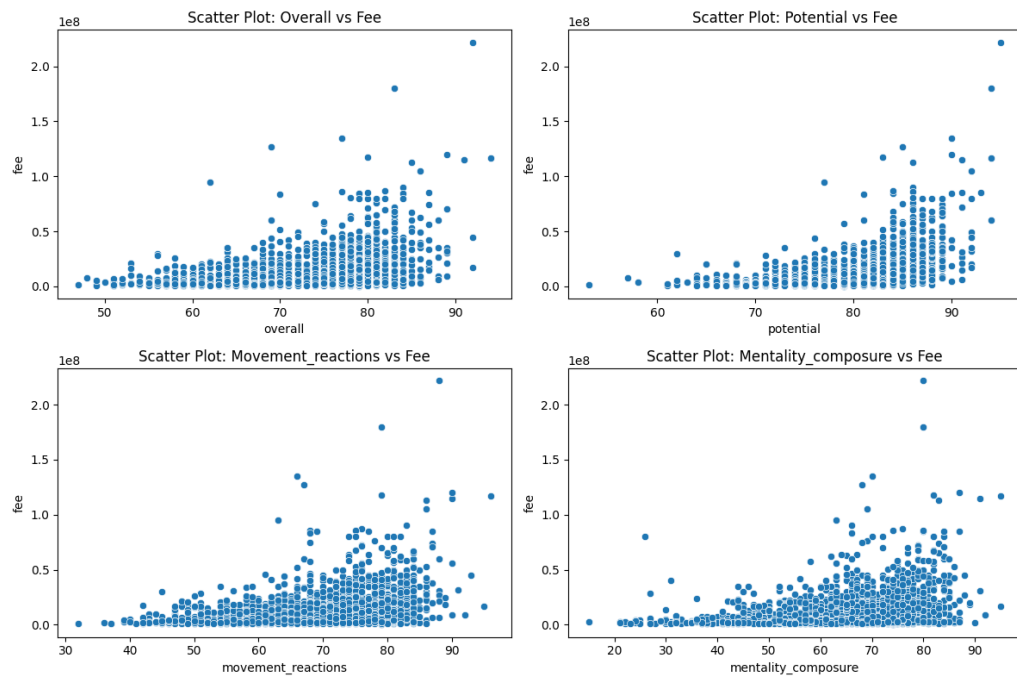
Hình 3-19: Phân bố dữ liệu theo nhóm các vị trí

3.3.4 Xử lý dữ liệu ngoại lai (*Remove Outlier*):

Dựa vào biểu đồ (Hình 3-20), thấy rằng có một vài biến có sự tương quan cao với biến mục tiêu (ở đây là thuộc tính ‘`fee`’) như ‘`overall`’, ‘`potentinal`’, ‘`movement_reactions`’, ‘`mentality_composure`’. Tôi sẽ trực quan hóa mối quan hệ giữa các biến tương quan mạnh nhất với biến mục tiêu ‘`fee`’ để xác định các giá trị ngoại lai (Hình 3-21).



Hình 3-20: Biểu đồ heatmap thể hiện sự tương quan giữa các thuộc tính



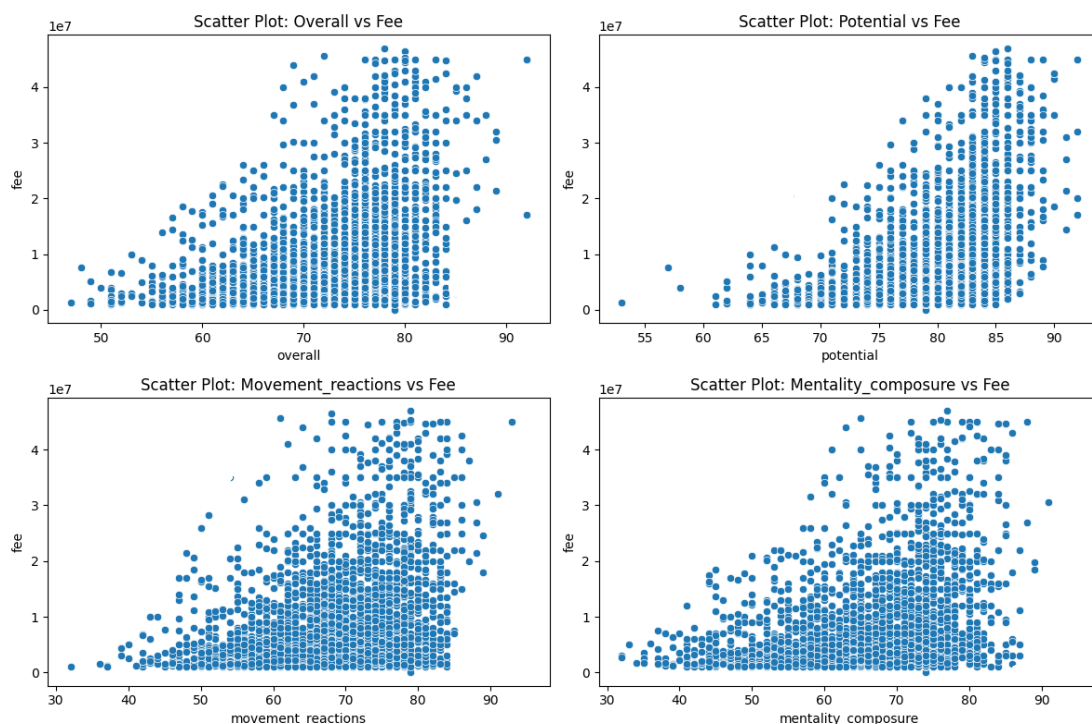
Hình 3-21: Biểu đồ trực quan mối quan hệ giữa các thuộc tính overall, potential, movement_reactions, mentality_composure với biến mục tiêu fee

Biểu đồ trong Hình 3-21 cho thấy có một vài điểm dữ liệu phi logic. Nhìn vào biểu đồ giữa ‘overall’ và ‘fee’ ta thấy có các điểm dữ liệu chỉ có ‘overall’ từ 60 -70 nhưng giá trị lại cao hơn mức 80 – 90, tương tự ‘potential’ cũng vậy. Các điểm dữ liệu này nằm không tuân theo mô hình chung của dữ liệu nên ta có thể xác định chính là dữ liệu ngoại lai và tiến hành xóa bỏ nó.

Trong bài này tôi sử dụng Z-score để loại bỏ các dữ liệu ngoại lai. Z-score là một phép đo thống kê đo khoảng cách của một giá trị từ trung bình của tập dữ liệu, tính theo đơn vị độ lệch chuẩn. Để loại bỏ dữ liệu ngoại lai sử dụng Z-score:

- Xác định Z-Score: Tính Z-score cho từng quan sát trong tập dữ liệu.
- Đặt ngưỡng Z-Score: Chọn một ngưỡng Z-score (thường là 2 hoặc 3) để xác định giá trị ngoại lai.
- Loại bỏ hoặc đánh dấu ngoại lai: Loại bỏ hoặc đánh dấu các giá trị có Z-score vượt quá ngưỡng.

Kết quả sau khi loại bỏ dữ liệu ngoại lai (Hình 3-22):



Hình 3-22: Biểu đồ trực quan mối quan hệ giữa các thuộc tính overall, potential, movement_reactions, mentality_composure với biến mục tiêu fee sau khi loại bỏ các dữ liệu ngoại lai

3.3.5 Chuẩn hóa và mã hóa dữ liệu:

Để thực hiện tiêu chuẩn hóa dữ liệu, chúng ta có thể sử dụng StandardScaler trong thư viện Scikit-learn. Công thức như sau:

Trong bài này các dữ liệu đều có một phạm vi nhất định không bị lệch quá nhiều ngoại trừ biến mục tiêu 'fee'. Vì vậy, tôi sẽ scale dữ liệu cho biến 'fee' để ổn định hiệu suất và cải thiện cho mô hình học máy.

Về mã hóa dữ liệu tôi sử dụng LabelEncoder, một thư viện của Python được dùng để chuyển đổi các giá trị của biến phân loại thành các biến số nguyên. Trong bài này các thuộc tính sau sẽ được sử dụng: 'nationality', 'preferred_foot', 'position_group', 'window', 'league'.

3.4 Xây dựng mô hình:

3.4.1 Mô hình LightGBM:

Mô hình được xây dựng ở đây là một Regressor sử dụng thuật toán LightGBM, một mô hình Gradient Boosting Framework hiệu quả. Mục tiêu của mô hình là dự đoán giá trị của biến mục tiêu 'fee' dựa trên các biến độc lập trong bộ dữ liệu.

Trước khi xây dựng mô hình, dữ liệu được chuẩn bị bằng cách mã hóa biến phân loại và chuẩn hóa và logarit biến 'fee'. Việc này có thể là để giảm biến động của dữ liệu và đảm bảo phân phối của nó gần với phân phối chuẩn.

Sau đó, tập dữ liệu được chia thành tập huấn luyện (X_{train} , y_{train}) và tập kiểm tra (X_{test} , y_{test}) sử dụng hàm 'train_test_split'.

Mô hình LightGBM được tinh chỉnh thông qua việc tìm kiếm siêu tham số sử dụng phương pháp Bayesian Optimization. Phạm vi của các siêu tham số được định nghĩa trong 'param_space', và các giá trị tốt nhất được chọn dựa trên độ đo 'neg_mean_squared_error'. Các tham số tốt nhất và giá trị đạt được cao nhất được in ra màn hình. Các tham số cụ thể như sau:

- `colsample_bytree` (0.7): Quyết định tỷ lệ số cột được lấy ngẫu nhiên khi xây dựng mỗi cây trong quá trình huấn luyện. Giá trị này giúp kiểm soát sự đa dạng của các cây trong ensemble.

- `learning_rate` (0.02): Tốc độ học của mô hình, tức là mức độ điều chỉnh trọng số sau mỗi bước huấn luyện. Learning rate càng nhỏ thì mô hình càng ổn định, nhưng có thể yêu cầu thêm thời gian để hội tụ.
- `max_depth` (8): Độ sâu tối đa của mỗi cây trong ensemble. Điều này kiểm soát độ phức tạp của mô hình và giúp tránh tình trạng quá mức fitting.
- `min_child_samples` (31): Số lượng mẫu tối thiểu được yêu cầu để tạo một nút lá. Điều này giúp kiểm soát sự phức tạp của cây và tránh tình trạng overfitting.
- `min_child_weight` (8): Trọng lượng tối thiểu của một nút lá. Giảm giá trị này có thể dẫn đến việc tăng độ phức tạp của mô hình.
- `min_split_gain` (0.08): Ngưỡng tối thiểu để thực hiện một phân chia. Điều này kiểm soát độ nhảy của mô hình khi phân loại dữ liệu.
- `n_estimators` (350): Số lượng cây (estimators) được xây dựng trong ensemble. Điều này kiểm soát số lượng bước huấn luyện.
- `num_leaves` (57): Số lượng lá tối đa trên mỗi cây. Số lượng lá càng lớn, mô hình càng có khả năng fit tốt dữ liệu huấn luyện, nhưng cũng có nguy cơ overfitting.
- `reg_alpha` (0.2): Hệ số alpha của L1 regularization. Hỗ trợ kiểm soát việc sử dụng các biến quan trọng.
- `reg_lambda` (0.9): Hệ số lambda của L2 regularization. Cũng là một hệ số kiểm soát regularization, giúp tránh overfitting.
- `subsample` (0.9): Tỷ lệ mẫu được sử dụng để huấn luyện mỗi cây. Điều này giúp kiểm soát độ phức tạp và đa dạng của mô hình.

Ta sử dụng một phạm vi cụ thể cho từng tham số sau đó sử dụng Bayesearchcv để kết hợp với cross-validation tìm ra bộ siêu tham số tốt nhất.

Sau khi xác định được các tham số tốt nhất, mô hình LightGBM được xây dựng lại (`best_model_bayes`) và được sử dụng để dự đoán trên tập kiểm tra. Các phương pháp đánh giá như Mean Squared Error (MSE), R-squared, Mean Absolute Error (MAE), và Root Mean Squared Error (RMSE) được sử dụng để đánh giá hiệu suất của mô hình trên tập kiểm tra.

3.4.2 Mô hình XGBoost:

Mô hình XGBoost được xây dựng là một mô hình học máy sử dụng thuật toán cực mạnh và hiệu quả, chuyên dụng cho bài toán dự đoán giá trị ‘fee’ trong bộ dữ liệu. Dưới đây là mô tả về cách mô hình hoạt động.

Đầu tiên, dữ liệu được chuẩn bị bằng cách tách các đặc trưng (biến độc lập) và biến mục tiêu (fee), sau đó thực hiện biến đổi mã hóa biến phân loại, chuẩn hóa và logarit thuộc tính ‘fee’ để làm giảm độ lớn của giá trị và cải thiện tính chất phân phối.

Tiếp theo, dữ liệu được chia thành tập huấn luyện (X_train, y_train) và tập kiểm tra (X_test, y_test) bằng cách sử dụng hàm ‘train_test_split’.

Mô hình XGBoost được chọn làm mô hình dự đoán. Một không gian siêu tham số để tìm kiếm được xác định (param_space_xgb). Trong trường hợp này, tìm kiếm siêu tham số được thực hiện bằng cách sử dụng Bayesian Optimization (BayesSearchCV). Các siêu tham số chi tiết:

- colsample_bylevel (1.0): Xác định tỷ lệ số cột được lấy ngẫu nhiên cho mỗi cấp độ cây (level). Giảm giá trị này có thể giúp kiểm soát việc overfitting.
- colsample_bynode (1.0): Tỷ lệ số cột được lấy ngẫu nhiên cho mỗi nút trong cây. Điều này kiểm soát độ đa dạng của cây.
- colsample_bytree (0.5): Xác định tỷ lệ số cột được lấy ngẫu nhiên cho mỗi cây trong ensemble. Giúp kiểm soát sự đa dạng của các cây.
- Gamma (0.0): Ngưỡng cắt cho cây để thực hiện một phân chia. Giảm giá trị này có thể kiểm soát sự phức tạp của mô hình và giảm overfitting.
- learning_rate (0.03): Tốc độ học của mô hình, tức là mức độ điều chỉnh trọng số sau mỗi bước huấn luyện. Giảm learning rate có thể làm tăng ổn định của mô hình.
- max_depth (3): Độ sâu tối đa của mỗi cây trong ensemble. Kiểm soát độ phức tạp của mô hình và giúp tránh overfitting.
- min_child_weight (3): Trọng lượng tối thiểu của mỗi nút lá. Giảm giá trị này có thể dẫn đến mô hình phức tạp hơn.
- n_estimators (700): Số lượng cây (estimators) trong ensemble. Kiểm soát số lượng bước huấn luyện.

- `reg_alpha` (0.04): Hệ số alpha của L1 regularization. Giúp kiểm soát việc sử dụng các biến quan trọng.
- `subsample` (0.6): Tỷ lệ mẫu được sử dụng để huấn luyện mỗi cây. Giảm giá trị này có thể kiểm soát sự đa dạng của mô hình.

Sau khi tìm kiếm siêu tham số, mô hình tốt nhất được chọn và lưu vào biến `'best_model_xgb'`. Mô hình này sau đó được sử dụng để dự đoán trên tập kiểm tra (`X_test`), và các kết quả được lưu vào biến `'predictions_xgb'`.

Cuối cùng, một số thông số đánh giá mô hình như Mean Squared Error (MSE), R-squared, Mean Absolute Error (MAE) và Root Mean Squared Error (RMSE) được in ra để đánh giá hiệu suất của mô hình trên tập kiểm tra.

3.4.3 Mô hình *Random Forest Regression*:

Mô hình `RandomForestRegressor` và thực hiện tinh chỉnh siêu tham số thông qua phương pháp Bayesian Optimization để cải thiện hiệu suất dự đoán giá trị `'fee'` trong bộ dữ liệu. Dưới đây là mô tả về cách xây dựng mô hình:

Chuẩn bị dữ liệu: Trước khi chia dữ liệu, dữ liệu đã được mã hóa biến phân loại, chuẩn hóa biến mục tiêu `'fee'`. Dữ liệu được chia thành hai phần chính: tập dữ liệu huấn luyện (`X_train, y_train`) và tập dữ liệu kiểm tra (`X_test, y_test`). Biến mục tiêu `'fee'` được chuyển đổi bằng hàm `'log1p'` để đảm bảo phân phối gần với phân phối chuẩn.

Xây dựng và tinh chỉnh mô hình `RandomForest`:

- `n_estimators` (300): Số lượng cây quyết định trong ensemble. Một số cây lớn có thể dẫn đến overfitting, trong khi một số cây nhỏ có thể làm giảm hiệu suất.
- `max_depth` (40): Độ sâu tối đa của mỗi cây quyết định. Giới hạn độ sâu giúp kiểm soát độ phức tạp của mô hình và ngăn chặn overfitting.
- `min_samples_split` (10): Số mẫu tối thiểu cần để một nút có thể được chia. Giảm giá trị này có thể dẫn đến các cây với các nhánh nhỏ hơn và có thể giảm overfitting.
- `min_samples_leaf` (10): Số mẫu tối thiểu cần để một lá của cây. Giảm giá trị này có thể làm tăng độ chính xác của lá nhưng cũng có thể dẫn đến overfitting.
- `Bootstrap` (true): Xác định liệu có nên thực hiện lấy mẫu có thay thế hay không. Lấy mẫu có thay thế giúp tăng tính ngẫu nhiên và đa dạng của các cây.

Tinh chỉnh tham số: Các tham số được lựa chọn thông qua Bayesian Optimization dựa trên độ đo 'neg_mean_squared_error', với mục tiêu là tối ưu hóa hiệu suất mô hình trên tập dữ liệu huấn luyện.

Kết quả tinh chỉnh tham số được in ra màn hình, bao gồm các giá trị tối ưu của số cây, độ sâu tối đa của cây, và các tham số khác. Mô hình tốt nhất được chọn dựa trên kết quả tinh chỉnh. Mô hình tinh chỉnh được đánh giá trên tập kiểm tra thông qua các số liệu đánh giá như Mean Squared Error (MSE), R-squared, Mean Absolute Error (MAE), và Root Mean Squared Error (RMSE) để đảm bảo khả năng dự đoán chính xác và hiệu quả của mô hình.

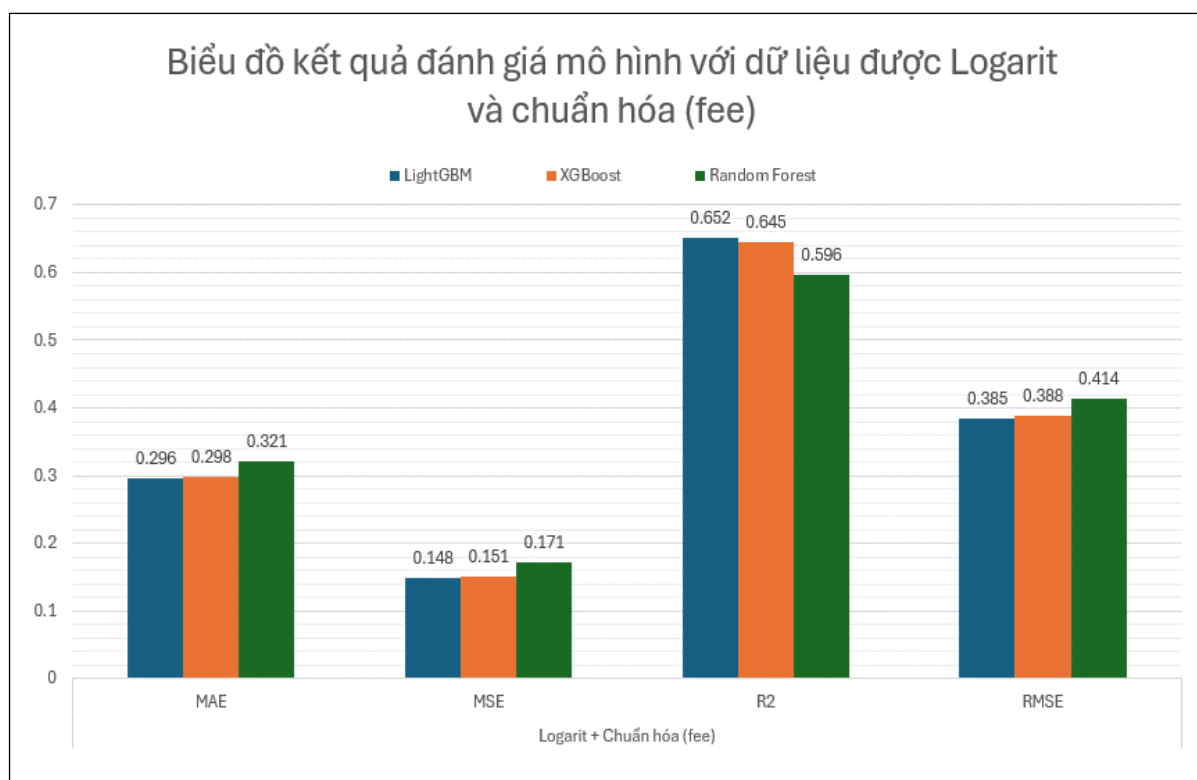
CHƯƠNG 4 : KẾT QUẢ VÀ ĐÁNH GIÁ MÔ HÌNH

4.1 Kết quả các độ đo của các mô hình:

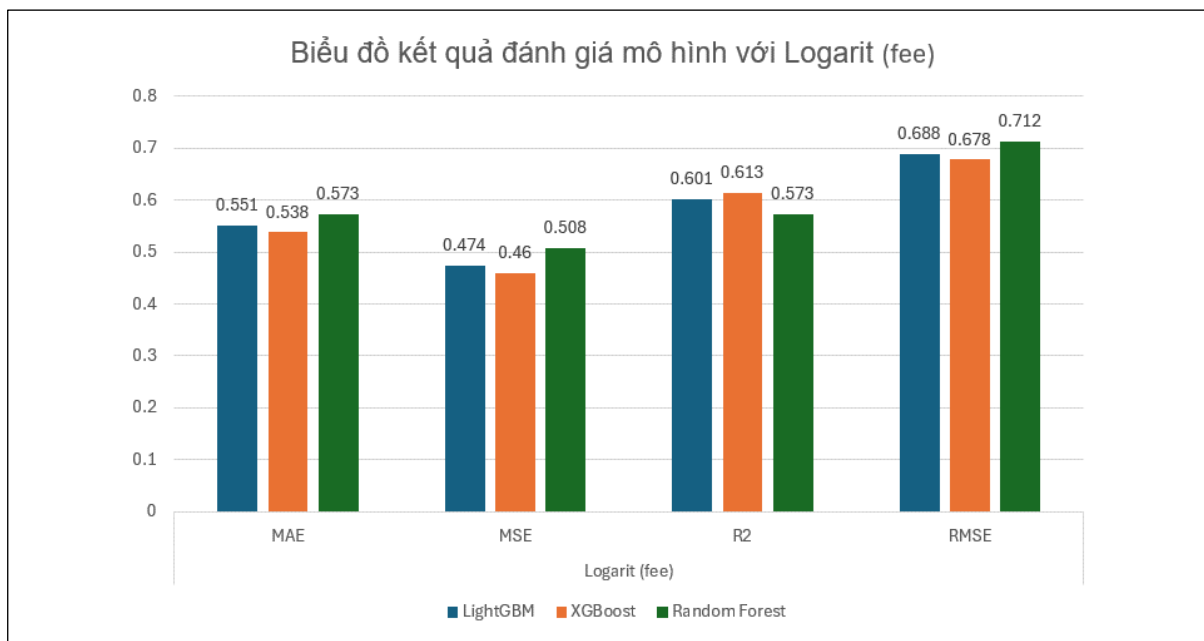
Trong bài này, tôi so sánh ba mô hình để xem mỗi mô hình dự đoán giá trị chuyển nhượng của các cầu thủ bóng đá chuyên nghiệp tốt đến mức nào. Hiệu suất của các mô hình được so sánh để chúng ta có thể tìm ra mô hình hoạt động tốt nhất là gì. Với các trường hợp phí chuyển nhượng sẽ được chuyển đổi theo logarit, chuẩn hóa, và kết hợp cả 2 phương thức. Kết quả ở bảng 4-1 sau:

	Logarit (fee)				Chuẩn hóa (fee)				Logarit + Chuẩn hóa (fee)			
	MAE	MSE	R2	RMSE	MAE	MSE	R2	RMSE	MAE	MSE	R2	RMSE
LightGBM	0.551	0.474	0.601	0.688	0.380	0.435	0.582	0.659	0.296	0.148	0.652	0.385
XGBoost	0.538	0.460	0.613	0.678	0.373	0.414	0.602	0.643	0.298	0.151	0.645	0.388
Random Forest	0.573	0.508	0.573	0.712	0.383	0.427	0.589	0.654	0.321	0.171	0.596	0.414

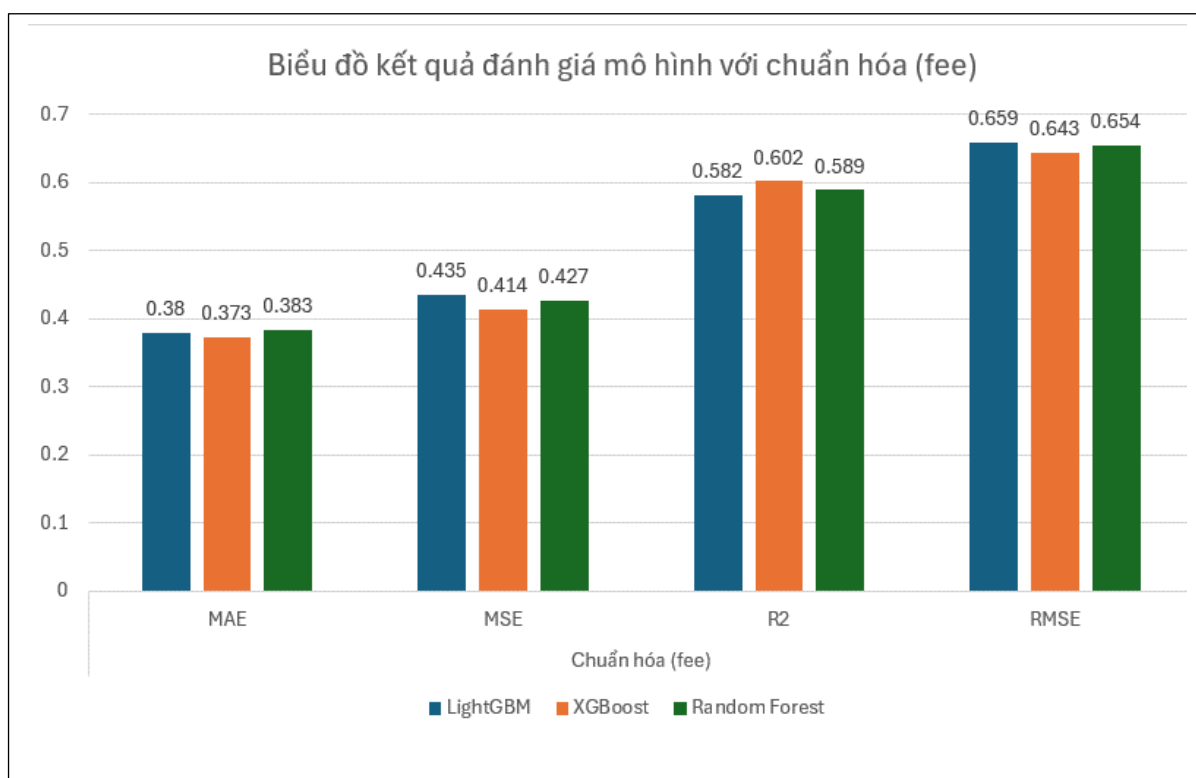
Bảng 4-1: Kết quả các độ đo của các mô hình học máy



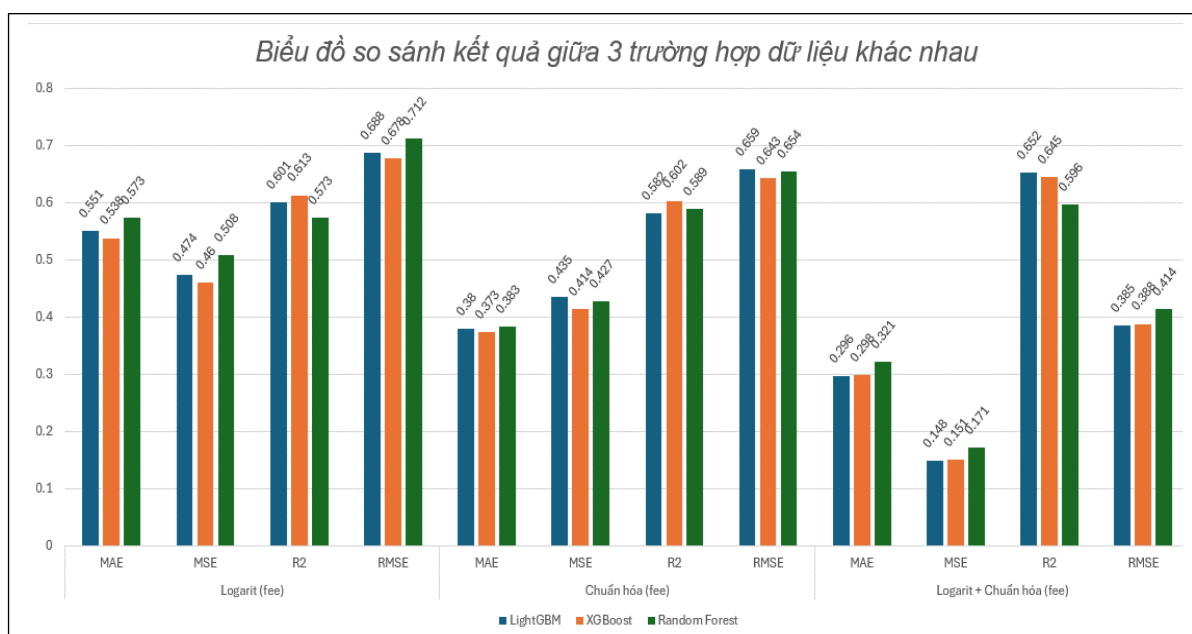
Hình 4-1: Biểu đồ kết quả đánh giá mô hình với Logarit và chuẩn hóa biến 'fee'



Hình 4-2: Biểu đồ kết quả đánh giá mô hình với Logarit (fee)



Hình 4-3: Biểu đồ kết quả đánh giá mô hình với chuẩn hóa (fee)



Hình 4-4: Biểu đồ so sánh kết quả giữa 3 trường hợp dữ liệu khác nhau

Nhận xét: Dựa vào kết quả hiệu suất của các mô hình (LightGBM, XGBoost, Random Forest) trong ba trường hợp tiền xử lý dữ liệu khác nhau (logarit, chuẩn hóa, logarit và chuẩn hóa), có thể rút ra một số nhận xét chi tiết.

Trong trường hợp sử dụng logarit cho biến mục tiêu ‘fee’, cả ba mô hình đều cho thấy hiệu suất tốt, nhất là với LightGBM có chỉ số R2 cao và các chỉ số MSE, RMSE thấp nhất. XGBoost cũng cho thấy hiệu suất tốt với R2 và các chỉ số MSE, RMSE khá cạnh tranh. Tuy nhiên, Random Forest có hiệu suất thấp hơn so với hai mô hình kia.

Khi áp dụng chuẩn hóa cho biến mục tiêu, LightGBM và XGBoost tiếp tục thể hiện hiệu suất tốt, với LightGBM có sự giảm mạnh về MSE và RMSE. Random Forest cũng có hiệu suất tốt, nhưng không sánh kịp với các mô hình còn lại.

Trong trường hợp logarit kết hợp chuẩn hóa, LightGBM xuất sắc với R2 cao và các chỉ số MSE, RMSE thấp nhất trong tất cả các trường hợp. XGBoost cũng duy trì hiệu suất tốt, đặc biệt là với MSE và RMSE. Random Forest cũng cho thấy sự cải thiện, nhưng vẫn thấp hơn so với các mô hình còn lại.

Điều đáng chú ý là kết hợp giữa logarit và chuẩn hóa biến mục tiêu ‘fee’ dường như mang lại kết quả tốt nhất, đặc biệt là đối với mô hình LightGBM. Sự linh hoạt của phương pháp này có thể giúp mô hình hiểu và dự đoán dữ liệu một cách hiệu quả hơn.

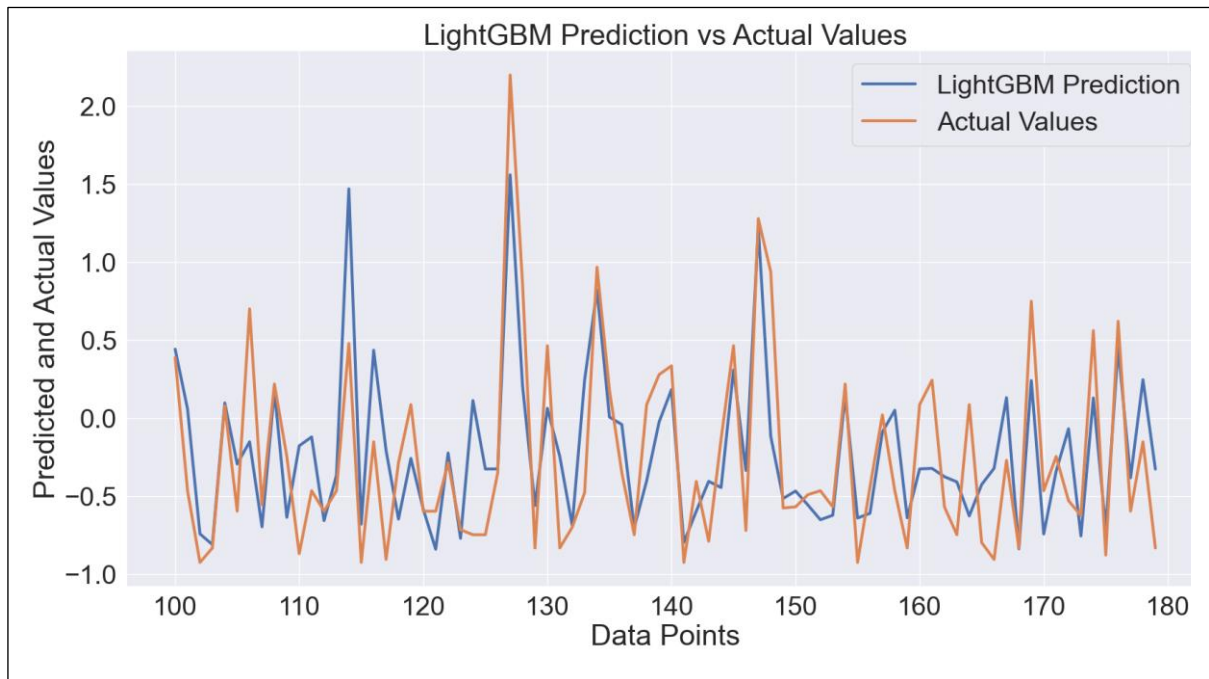
Trong tổng thể, việc lựa chọn mô hình cũng đóng vai trò quan trọng, với XGBoost và LightGBM thường cho hiệu suất tốt hơn so với Random Forest trong bối cảnh này.

Để có thể trực quan về việc dự đoán và thể hiện ra các giá trị của mô hình, tôi đã xây dựng một giao diện đơn giản để nhập dữ liệu đầu vào là các biến độc lập của mô hình. Giao diện sử dụng thư viện Tkinter của Python để thực hiện việc nhập, sau đó tôi sẽ chuyển đổi các dữ liệu đầu vào đó thành dữ liệu để có thể dự đoán bằng mô hình và trả về kết quả ra màn hình (Hình 4-5)

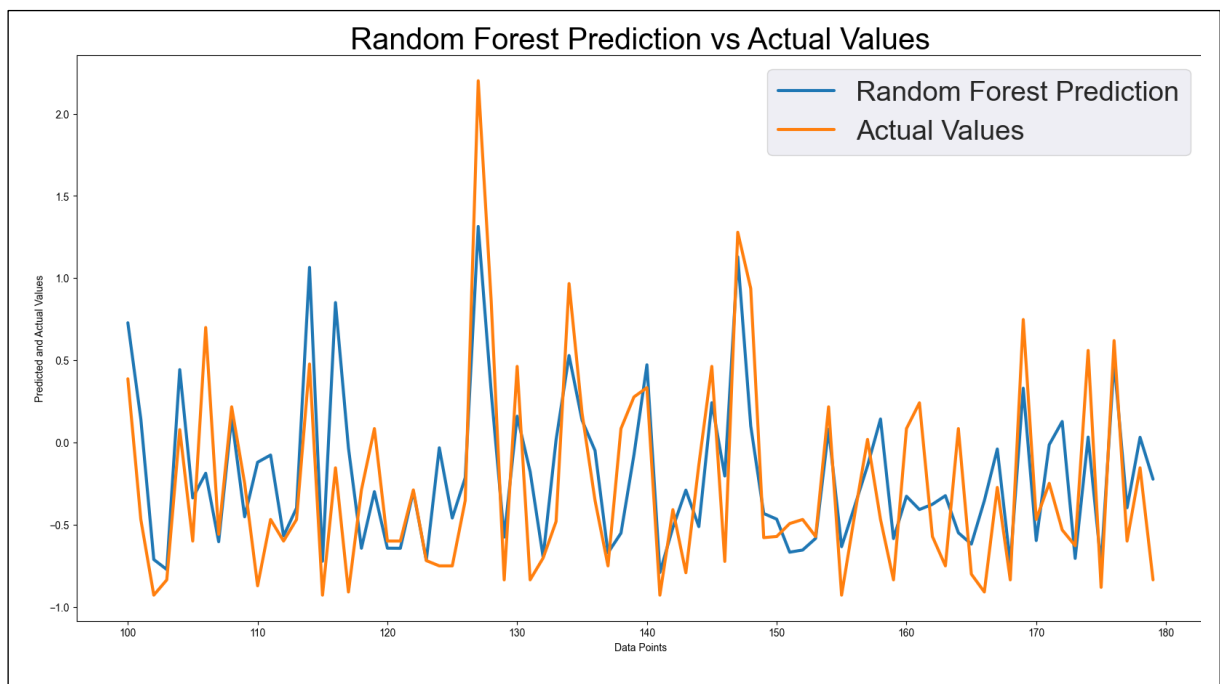
Hình 4-5: Giao diện dự đoán giá trị chuyển nhượng

4.2 Đánh giá mô hình:

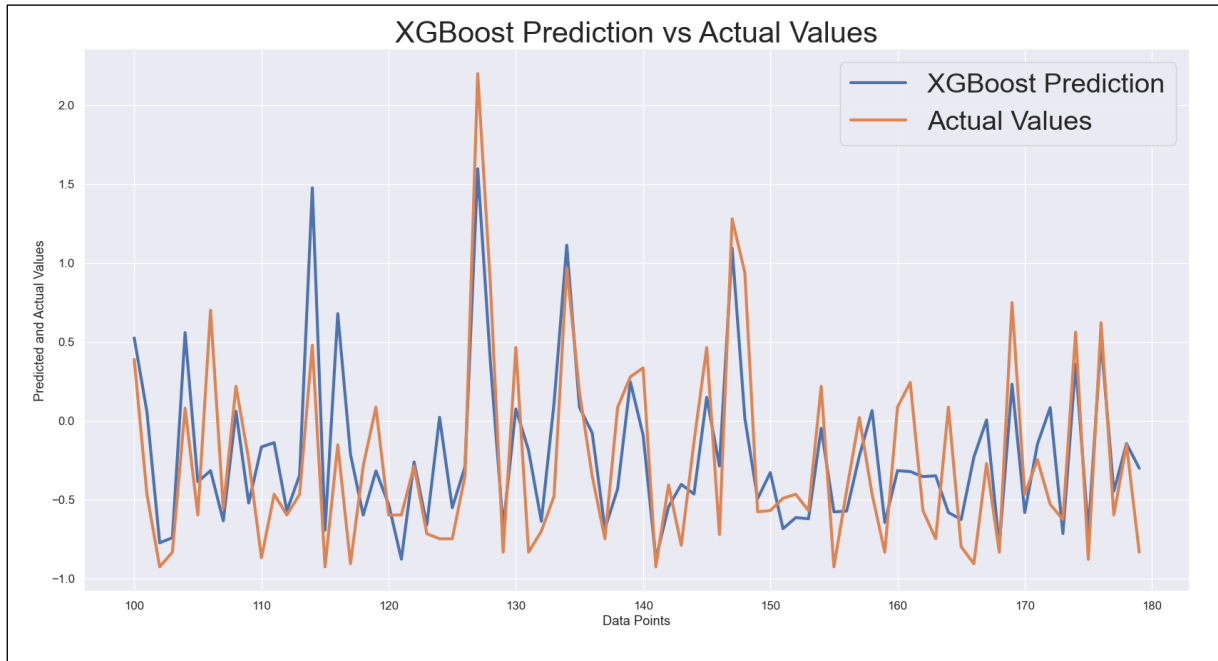
Các biểu đồ (Hình 4-6, 4-7, 4-8) thể hiện sự phân phối giữa giá trị thực tế và giá trị dự đoán. Dựa vào đây ta có thể đánh giá mô hình đang hoạt động tốt ở điểm nào và kém ở điểm nào.



Hình 4-6: Biểu đồ thể hiện giữa giá trị thực tế và giá trị dự đoán của mô hình LightGBM



Hình 4-7: Biểu đồ thể hiện giữa giá trị thực tế và giá trị dự đoán của mô hình Random Forest



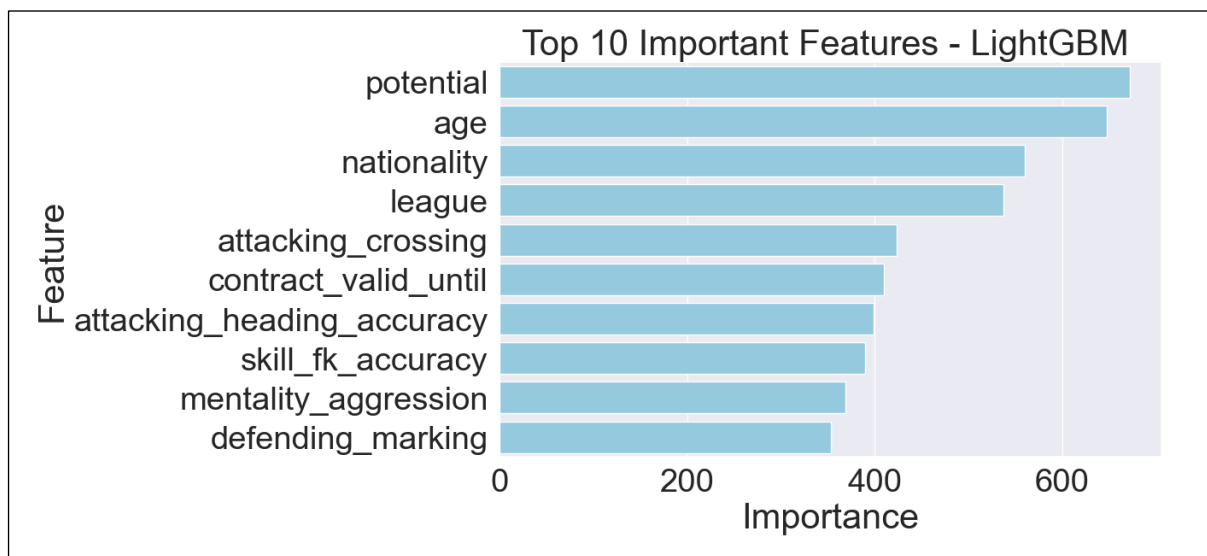
Hình 4-8: Biểu đồ thể hiện giữa giá trị thực tế và giá trị dự đoán của mô hình XGBoost

Với cùng một bộ dữ liệu nhất định, quan sát kết quả dự đoán từ ba mô hình Machine Learning khác nhau - LightGBM, XGBoost và Random Forest, cho thấy sự tương đồng chung trong hiệu suất của chúng. Tuy nhiên, sự khác biệt trở nên rõ ràng đặc biệt khi xem xét một số điểm dữ liệu cụ thể.

- Cụ thể, đối với các điểm dữ liệu như 128, 135 và 148, có sự nhất quán trong việc LightGBM và XGBoost đưa ra dự đoán gần với giá trị thực tế hơn so với mô hình Random Forest. Điều này có thể là do khả năng học tập linh hoạt và khả năng ổn định của các thuật toán cụ thể này.
- Tổng quan, phân phối giữa giá trị dự đoán và giá trị thực tế của cả ba mô hình có sự tương đồng. Tuy nhiên, khi quan sát các khoảng giá trị nhỏ, chẳng hạn từ 105 đến 115 và từ 150 đến 170, ta nhận thấy sự chênh lệch lớn hơn giữa giá trị dự đoán và giá trị thực tế. Điều này có thể là do thiếu số lượng dữ liệu đủ lớn trong những khoảng giá trị này, dẫn đến khả năng học tập và áp dụng kém.

Kết quả này cung cấp một cái nhìn tổng quát về hiệu suất ổn định của ba mô hình. Tuy nhiên, để đạt được hiệu suất tốt nhất, có thể cần xem xét cụ thể và điều chỉnh các yếu tố như tỷ lệ mẫu, xử lý dữ liệu ngoại lệ, hoặc thậm chí là tối ưu hóa các siêu tham số của mô hình. Điều này là quan trọng để đảm bảo rằng mô hình không chỉ hoạt động tốt với dữ liệu hiện tại mà còn tổng quát hóa tốt cho các tình huống dữ liệu mới.

Để trả lời câu hỏi nghiên cứu và có thể cho biết có thể dự đoán giá trị chuyển nhượng của các cầu thủ bóng đá chuyên nghiệp ở mức độ nào, chúng ta cần biết những đặc điểm nào là quan trọng nhất khi dự đoán giá trị này. Vì mô hình LightGBM cho kết quả tốt nhất nên chúng ta sẽ xem xét các tính năng quan trọng nhất trong mô hình này. Biến quan trọng nhất là ‘potential’, cho biết mức độ một cầu thủ có thể phát triển trong sự nghiệp của mình theo FIFA. Các thuộc tính quan trọng tiếp theo như tuổi, quốc tịch, giải đấu, chuyên bóng, cũng ảnh hưởng đến giá trị chuyển nhượng. Trong Hình 4-9 đặc điểm quan trọng nhất trong mô hình LightGBM được hiển thị.



Hình 4-9: Top 10 thuộc tính quan trọng nhất của mô hình LightGBM

Kết quả thu được từ mô hình dường như phản ánh khá chân thực với thực tế, đặc biệt là khi xem xét các yếu tố như tuổi của cầu thủ, quốc tịch không thuộc các quốc gia châu Âu nổi tiếng như Anh, Pháp, Đức, cũng như giải đấu ít cạnh tranh hơn hoặc thời hạn hợp đồng còn lại và một số chỉ số kỹ năng khác. Mô hình đã có khả năng ổn định và hiệu quả khi dự đoán giá trị chuyển nhượng, bắt chước cách các câu lạc bộ thực tế đánh giá giá trị của cầu thủ trong thị trường chuyển nhượng.

Kết quả thu được từ mô hình dự đoán giá trị chuyển nhượng của cầu thủ đưa ra những nhận định tích cực đối với câu hỏi nghiên cứu ban đầu: “Giá trị chuyển nhượng của các cầu thủ có thể được dự đoán ở mức độ nào dựa trên kỹ năng và đặc điểm cá nhân của họ?”. Mô hình đã chứng minh khả năng ưu việt trong việc dự đoán giá trị chuyển nhượng, với sự tích hợp hiệu quả của các đặc điểm và kỹ năng cá nhân của cầu thủ. Nhìn

chung, kết quả này cũng là một sự khẳng định cho việc sử dụng dữ liệu từ Sofifa là một nguồn thông tin quan trọng và có giá trị. Các thuộc tính quan trọng nhất, như được xác định thông qua mô hình, đều xuất phát từ dữ liệu này. Điều này cung cấp sự hiểu biết sâu sắc về những yếu tố nào ảnh hưởng đến giá trị chuyển nhượng và làm cơ sở cho việc đưa ra quyết định chiến lược khi tìm kiếm và đàm phán với cầu thủ.

KẾT LUẬN

Trong đồ án này, tôi muốn tìm hiểu xem giá trị chuyển nhượng của các cầu thủ bóng đá chuyên nghiệp có thể dự đoán được ở mức độ nào dựa trên kỹ năng và đặc điểm. Để kiểm tra điều này, tôi đã thu thập dữ liệu từ 2 nguồn chính là Sofifa và Transfermarkt để lấy dữ liệu về cầu thủ và giá trị chuyển nhượng. Sau đó xây dựng các mô hình học máy để dự đoán giá trị chuyển nhượng của các cầu thủ bóng đá chuyên nghiệp. Mô hình hoạt động tốt nhất là mô hình hồi quy LightGBM ($R^2 = 0,652$). Có thể có những yếu tố khác góp phần tạo ra phí chuyển nhượng thực tế mà tôi không xem xét trong đồ án này, điều này có thể gây ra hiệu suất không cao. Vì vậy trong tương lai, tôi sẽ cải tiến bằng cách đưa các biến mới vào mô hình để xem liệu điều này có ảnh hưởng đến hiệu suất của mô hình hay không, chẳng hạn như ngân sách chuyển nhượng của các câu lạc bộ, sức ảnh hưởng của họ trên mạng xã hội, cơ hội và các chỉ số chi tiết khác trong trận đấu.

Đồ án này cung cấp một cái nhìn sâu sắc vào những tính năng quan trọng nhất để dự đoán giá trị chuyển nhượng. Các thuộc tính từ trò chơi điện tử FIFA đã chứng minh là hữu ích trong việc này, với chín trong mười tính năng quan trọng nhất của mô hình hồi quy LightGBM xuất phát từ dữ liệu này. Các tính năng quan trọng nhất là tiềm năng, tuổi, quốc tịch, giải đấu, thời hạn hợp đồng và một số kỹ thuật như: chuyển bóng, đánh đầu, ... Dựa vào đó chúng ta có thể kết luận rằng giá trị chuyển nhượng có thể được dự đoán được với các kỹ năng và đặc điểm của cầu thủ.

Do thời gian có hạn và trình độ hiểu biết của còn hạn chế nên đồ án này không thể tránh khỏi những thiếu sót. Rất mong sẽ nhận được những lời khuyên, lời đánh giá từ quý thầy, cô để đồ án này được hoàn thiện hơn.

TÀI LIỆU THAM KHẢO

- [1] Al-Asadi, M. A. and S. Tasdemir, Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques. *IEEE Access*, 10, 22631–22645. <https://doi.org/10.1109/ACCESS.2022.3154767>, 2022.
- [2] Kirschstein and Liebscher, Assessing the market values of soccer players—a robust analysis of data from German 1. and 2. Bundesliga. *Journal of Applied Statistics*, 46(7), 1336–1349. <https://doi.org/10.1080/02664763.2018.1540689>, 2019.
- [3] Yiğit, A. T., Samak, B. and T. Kaya, "Football Player Value Assessment Using Machine Learning Techniques. In S. and C. O. S. and O. B. and T. A. C. and S. I. U. Kahraman Cengiz and Cebi (Ed.), *Intelligent and Fuzzy Techniques in Big Data Analytics and Decision Making* (pp. 289–297).," 2020. [Online]. Available: Springer International Publishing, <https://link-springer-com.tilburguniversity.idm.oclc.org/chapter/10.1007/978-3-030->.
- [4] Behravan, I., & Razavi and S. M., A novel machine learning method for estimating football players' value in the transfer market. *Soft Computing*, 25(3), 2499–2511. <https://doi.org/10.1007/s00500-020-05319-3>, 2021.
- [5] Barbuscak and L., What Makes a Soccer Player Expensive? Analyzing the Transfer Activity of the Richest Soccer. In *Augsburg Honors Review* (Vol. 11). https://idun.augsburg.edu/honors_review Available at: https://idun.augsburg.edu/honors_review/vol11/iss1/5, 2018.
- [6] Poli, R., R. Besson, Ravenel and L., Econometric Approach to Assessing the Transfer Fees and Values of Professional Football Players. *Economies*, 10(1). <https://doi.org/10.3390/economies10010004>, 2022.