Kính chào các thầy, cô giáo trong hội đồng cùng toàn thể các bạn sinh viên có mặt tại buổi bảo vệ ngày hôm nay. Em tên là Nguyễn Thành Trung, sinh viên lớp 61TH6, khoa CNTT trường ĐHTL. Dưới sự hướng dẫn của thầy giáo Nguyễn Đắc Hiếu, em đã hoàn thành đồ án tốt nghiệp của mình với đề tài “**ỨNG DỤNG CÁC MÔ HÌNH HỌC MÁY DỰ ĐOÁN GIÁ TRỊ CHUYỂN NHƯỢNG CỦA CẦU THỦ BÓNG ĐÁ CHUYÊN NGHIỆP”**

Đồ án của e bao gồm 4 phần như sau: Tổng quan về đề tài, tổng quan về lý thuyết, quy trình thực hiện, kết quả và đánh giá.

**Phần đầu tiên tổng quan về đề tài:**

Bóng đá không chỉ là một môn thể thao phổ biến mà còn là yếu tố kết nối văn hóa, mang đến nhiều khung bậc cảm xúc khác nhau. Với sự yêu thích và đón nhận lớn như vậy, việc lựa chọn cầu thủ phù hợp trở thành một yếu tố quan trọng đối với sự thành công của mỗi đội bóng.

**Lý do chọn đề tài**

Trong bối cảnh nền bóng đá ngày càng phát triển, việc lựa chọn cầu thủ trở nên phức tạp do sự cạnh tranh giữa các đội bóng lớn và sự tăng giá của cầu thủ. Quản lý chuyển nhượng đòi hỏi sự hiểu biết sâu sắc về bóng đá và khả năng phân tích dữ liệu chính xác. Tuy nhiên, phương pháp tuyển chọn truyền thống còn nhiều hạn chế và rủi ro, dẫn đến việc đưa ra quyết định không chính xác. Do vậy việc có một công cụ dự đoán giá chuyển nhượng của các cầu thủ là một giải pháp hữu ích hỗ trợ việc lựa chọn và ra quyết định đúng đắn

**Tổng quan về lý thuyết:**

Học máy là một lĩnh vực trong trí tuệ nhân tạo (AI), 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.

Học máy được chia thành 3 loại chính: học có giám sát, học không giám sát và học tăng cường

Bài toán của em đang thuộc loại học có giám sát cụ thể là bài toán dự đoán.

**Tổng quan về các mô hình sử dụng trong bài:**

Các mô hình em đã sử dụng trong bài toán bao gồm, RFR, Lightgbm, XGBoost, chúng đều là các thuật toán thuộc ensemble learning. Do vậy Trước tiên em sẽ trình bày về khái niệm này. Ensemble Learning là một phương pháp kết hợp nhiều mô hình để tạo ra một mô hình mạnh mẽ hơn so với việc sử dụng một mô hình đơn lẻ. Các biến thể:

Boosting: Các mô hình được xây dựng tuần tự, mỗi mô hình cố gắng sửa lỗi của mô hình trước đó. Đây cũng là cơ sở cho 2 thuật toán LightGBM và XGB

Bagging: Sử dụng nhiều mô hình với dữ liệu huấn luyện được tạo ra bằng cách lấy mẫu ngẫu nhiên có thay thế từ dữ liệu gốc,

**Random Forest**

Điển hình cho kỹ thuật này chính là random forest, đây là một thuật toán kết hợp nhiều cây quyết định sau đó kết hợp chúng để tạo ra kết quả cuối cùng, với bài toán dự đoán thì kết quả sẽ là trung bình của tất cả các dự đoán.

**LightGBM và XGB**

Cả 2 thuật toán LightGBM và XGB đều được phát triển dựa trên khung Gradient Boosting, do đó chúng ta sẽ tìm hiểu Gradient boosting trước.

Gradient Boosting là một phương pháp cụ thể trong boosting, Ý tưởng cơ bản của phương pháp này là học từ sai số của mô hình trước đó và cố gắng giảm sai số đó trong mô hình tiếp theo. *Đầu tiên thuật toán sẽ tạo ra một dự đoán ban đầu . Tính toán độ chệch giữa dự đoán thực tế và dự đoán của mô hình hiện tại. Với mỗi mô hình mới được tạo, chúng sẽ được "học" từ độ chệch để giảm thiểu hàm mất mát và cải thiện khả năng dự đoán.*

Và dựa vào khung GB này đã có nhiều biến thể cải tiến về hiệu suất và tốc độ chạy nổi tiếng và phổ biến đó chính là lightgbm và xgboost.

Một số cải tiến nổi bật có thể kể đến như:

khả năng tối ưu hóa bộ nhớ và tối ưu hóa quá trình huấn luyện. Khả năng học song song cũng là một điểm mạnh, giúp giảm thời gian huấn luyện đặc biệt trên các tập dữ liệu lớn. Cả hai thuật toán cũng cải tiến việc ngăn chặn overfitting trên dữ liệu đào tạo.

Đối với LightGBM còn sử dụng histogram-based algorithms kết hợp với các kỹ thuật như GOSS và EFB để tăng tốc độ tính toán. Thêm nữa nó sử dụng phương pháp phát triển theo lá thay vì phát triern theo cấp như các thuật toán ensemble khác.

**Tiếp theo đến phần Quy trình thực hiện, ở phần này em sẽ thực hiện 4 bước chính là:**

Thu thập dữ liệu

Tổng quan các bước thực hiện:  
Đối với phần thu thập dữ liệu em sẽ có một sơ đồ tổng quan các bước thực hiện như sau,

Quá trình thu thập sẽ bắt đầu bằng việc gửi các yêu cầu đến trang web cần lấy dữ liệu.

em sẽ lấy dữ liệu về cầu thủ từ trang web Sofifa và dữ liệu về giá chuyển nhượng từ Transfermarkt. Cả hai nguồn dữ liệu này đã được chứng minh là đáng tin cậy và được sử dụng phổ biến trong cộng đồng nghiên cứu bóng đá.

Sau khi gửi yêu cầu thì Trang web sẽ phản hồi bằng cách trả về cấu trúc HTML của trang web đó. Sau đó sẽ sử dụng các kỹ thuật scraping để trích xuất thông tin cần thiết từ HTML và lưu trữ dữ liệu thu được vào các tập tin CSV.

Để thu thập dữ liệu về cầu thủ, em sẽ lấy danh sách liên kết của từng cầu thủ. Mỗi liên kết này sẽ chứa các thông tin chi tiết về cầu thủ, như được minh họa trong hình bên phải màn hình. Các vùng được em đánh dấu khung là những dữ liệu mà em sẽ tập trung thu thập.

Tiếp theo, em sẽ sử dụng thư viện Beautiful Soup để trích xuất các dữ liệu từ cấu trúc HTML, như đã được mô tả trong hình.

Đến với phần thu thập dữ liệu về giá chuyển nhượng, em sẽ tập trung vào các dữ liệu chuyển nhượng của từng câu lạc bộ. Các thuộc tính cần thu thập đã được em đánh dấu và minh họa trong hình. Và giá trị của côt fee là giá chuyển nhượng của các cầu thủ đồng thời đây sẽ là biến mục tiêu trong bộ dữ liệu của em.

Tương tự với phần thu thập trước, em cũng trích xuất các dữ liệu từ cấu trúc HTML của trang web.

Trong quá trình thu thập thì em nhận thấy rằng có các dữ liệu đặc biệt trong 2 cột là fee và market value, ví dụ như giá trị = ‘-‘ hay bằng ‘?’ vì vậy em sẽ xử lý chúng trước khi lấy dữ liệu.chi tiết đc minh họa như hình sau.

Sau khi thu thập em sẽ tiến hành làm sạch trước khi kết hợp với bộ dữ liệu cầu thủ.

Sau khi thu thập bộ dữ liệu cầu thủ từ Sofifa và dữ liệu về giá chuyển nhượng từ Transfermarkt,em sẽ hợp chúng thành một bộ dữ liệu. Em sẽ thực hiện quy trình này bằng cách so sánh tên và quốc tịch của các cầu thủ từ hai nguồn dữ liệu. Nếu trùng khớp, em sẽ lấy thông tin từ cả hai bộ dữ liệu và hợp nhất chúng thành một bộ dữ liệu chung, các bước thực hiện được mô tả chi tiết trong hình bên phải.

Sau quá trình này, em sẽ có được một bộ dữ liệu tổng cộng gồm 4158 mẫu, và đây sẽ là bộ dữ liệu em sẽ sử dụng trong đồ án của mình.

**Đến với bước thứ 2 trong quy trình thực hiện đó chính là tiền xử lý dữ liệu.**

Đầu tiên là xử lý dữ liệu thiếu, em sẽ xác định các cột chứa dữ liệu thiếu. Trong quá trình phân tích dữ liệu, em nhận thấy rằng các thuộc tính có sự liên quan đến vị trí của cầu thủ, ví dụ như hình thuộc tính phản xạ thủ môn. Thuộc tính này đặc thủ giành cho vị trí thủ môn có giá trị cao hơn so với các vị trí khác như tiền đạo hay hậu vệ.

Do đó, để điền vào các giá trị thiếu, em tính giá trị trung bình của thuộc tính theo từng vị trí. Việc này giúp đảm bảo rằng giá trị được điền vào sẽ phản ánh đúng cho từng vị trí cầu thủ, thay vì sử dụng một giá trị trung bình chung

Tiếp đến là phần lựa chọn thuộc tính Các thuộc tính mang giá trị duy nhất sẽ được loại bỏ. Loại bỏ các cột trùng lặp của 2 bộ dữ liệu. Các cột không cung cấp thông tin về kỹ năng và đặc điểm của cầu thủ cũng sẽ được loại bỏ

Cũng trong quá trình phân tích dữ liệu, em nhận thấy rằng phân phối của dữ liệu với các vị trí không đồng đều và có những vị trí có số lượng dữ liệu quá ít, làm cho chúng không đủ để sử dụng trong mô hình học máy. Do đó, em sẽ nhóm các vị trí thành một thuộc tính mới dựa trên phần sân mà cầu thủ thường hoạt động.

Cụ thể, em tạo ra một thuộc tính mới có 4 giá trị: "attacker" (tiền đạo), "defender" (hậu vệ), "midfielder" (tiền vệ), và "goalkeeper" (thủ môn)

Với phần xử lý dữ liệu ngoại lai: em sẽ xác định các thuộc tính có sự tương quan cao với biến mục tiêu ở đây là thuộc tính fee để tìm kiếm các điểm ngoại lai sau đó em sẽ sử dụng phương pháp z-score để loại bỏ các giá trị ngoại lai. Đây là trực quan trước và sau khi loại bỏ các dữ liệu ngoại lai. Phân phối của dữ liệu đã trở nên tuyến tính hơn.

**Xây dựng mô hình**

Đến với phần thứ 3 trong quy trình đó chính là xây dựng mô hình.

Sau khi tiền xử lý dữ liệu, em áp dụng chuyển đổi log cho biến mục tiêu, nhằm giảm sự chênh lệch giữa các giá trị lớn và giá trị nhỏ, từ đó làm cho phân phối của dữ liệu gần với phân phối chuẩn hơn. Tiếp theo, em chia dữ liệu thành hai phần, 80% cho quá trình huấn luyện và 20% cho quá trình kiểm tra.

Trong giai đoạn huấn luyện mô hình, em áp dụng các mô hình học máy và thực hiện điều chỉnh tham số bằng phương pháp Bayesian Optimization kết hợp với kỹ thuật cross-validation để tìm ra mô hình với tham số tối ưu nhất. Sau khi có mô hình đã được điều chỉnh, em sử dụng tập kiểm tra để dự đoán và tiến hành đánh giá mô hình với các độ đo.

Dưới đây là kết quả sau quá trình huấn luyện, với các phương pháp xử lý khác nhau. Phương pháp chuyển đổi log kết hợp với chuẩn hóa đã đem lại kết quả tốt nhất với R2 = 0.652.

Qua việc quan sát biểu đồ, ta nhận thấy các điểm dữ liệu phân tán và không hội tụ chặt với đường tuyến tính, có thể đưa ra kết luận rằng mô hình đang mắc sai lệch so với giá trị thực tế, cụ thể trong trường hợp giá trị thực tế thấp, mô hình dự đoán giá trị cao và ngược lại. Điều này có thể giải thích rằng dữ liệu về cầu thủ hiện tại chưa đủ thuộc tính để giải thích một cách chính xác giá trị chuyển nhượng

Và trong quá trình thực hiện đồ án, em có tham khảo các nghiên cứu liên quan và thu được kết quả sau. Trong một nghiên cứu sử dụng cùng bộ dữ liệu, mô hình RFR đạt R2 cao nhất là 0.617. Tuy nhiên, khi em áp dụng các mô hình mới và mạnh mẽ như XGBoost và LightGBM trong đồ án, kết quả đã được cải thiện hơn cụ thXGBoost: 0.645 và LightGBM: 0.652. Điều này chứng minh sự tiến bộ và tính hiệu quả của các phương pháp mà em áp dụng cùng với các các mô hình tiên tiến

Để làm cho quá trình trực quan trở nên thuận tiện hơn, em đã sử dụng Python để tạo ra một giao diện đơn giản giúp dự đoán giá trị chuyển nhượng.

**Em biết là về tgian trình bày em sẽ không đủ để có thể nêu được các nội dung của đề tài nên các thầy cô cần thêm thắc mắc gì về phần trình bày của em thì có thể đưa ra câu hỏi ạ**