PROJECT SUMMARY

I. Key Findings

Dựa trên kết quả phân tích dữ liệu và huấn luyện mô hình, nhóm nghiên cứu rút ra 4 nhận định cốt lõi:

1. **Hiệu suất vượt trội của CatBoost:** CatBoost đã chứng minh khả năng dự đoán vượt trội so với các mô hình XGBoost và Random Forest, đặc biệt trong việc xử lý các biến phân loại (categorical features) như Club và Nationality mà không cần One-Hot Encoding phức tạp.
2. **"Độ tuổi vàng":** Biểu đồ phân tích cho thấy mối tương quan phi tuyến tính giữa Age và Market Value. Giá trị cầu thủ thường tăng trưởng nhanh và đạt đỉnh trong khoảng **24 - 27 tuổi**, sau đó giảm dần khi qua ngưỡng 30, phản ánh đúng thực tế về vòng đời sự nghiệp cầu thủ.
3. **Hiệu ứng giải đấu:** Có sự phân hóa rõ rệt về giá trị giữa các giải đấu. Cầu thủ thi đấu tại **Top 5 giải VĐQG hàng đầu châu Âu** (Premier League, La Liga,...) có mức định giá khởi điểm cao hơn hẳn so với các giải đấu còn lại, cho thấy yếu tố "thương hiệu giải đấu" đóng vai trò quan trọng.
4. **Khám phá thú vị nhất:** Phân phối của Giá trị thị trường bị **lệch phải cực đại**. Điều này chỉ ra rằng thị trường chuyển nhượng bị chi phối bởi "Hiệu ứng siêu sao" (Superstar Effect) – nơi một nhóm rất nhỏ các cầu thủ (outliers) nắm giữ giá trị khổng lồ, trong khi đại đa số cầu thủ chuyên nghiệp chỉ có mức định giá trung bình thấp. Các mô hình máy học thường gặp khó khăn với các outlier này nếu không được xử lý log-transform phù hợp.

II. Limitations

1. **Dữ liệu:**

Tính thời điểm: Dữ liệu chỉ phản ánh giá trị tại thời điểm thu thập thiếu tính lịch sử để phân tích sự biến động giá trị theo thời gian.

Biến ẩn: Dữ liệu không có các thông tin quan trọng ảnh hưởng lớn đến giá chuyển nhượng như: *Thời hạn hợp đồng còn lại, Lịch sử chấn thương, Giá trị thương mại/truyền thông*.

Thiếu sót: Chưa thấy được performance của các Hậu vệ, Thủ môn để định giá cầu thủ

2. **Phân tích:**

   Mô hình hiện tại tập trung vào các chỉ số thống kê cơ bản (bàn thắng/kiến tạo) mà chưa đi sâu vào các chỉ số nâng cao (Advanced Metrics như xG, xA, Key Passes) do giới hạn của nguồn dữ liệu.

III. Future Directions

Nếu có thêm thời gian và nguồn lực, nhóm sẽ mở rộng nghiên cứu theo các hướng sau:

1. **Mở rộng dữ liệu:** Thu thập thêm dữ liệu trong 5-10 năm để chuyển bài toán từ hồi quy tĩnh sang dự đoán theo chuỗi thời gian.
2. **Phân tích nâng cao:** Tích hợp thêm các chỉ số quan trọng khác và phân tích cảm xúc tin tức từ báo chí/mạng xã hội để xem dư luận ảnh hưởng thế nào đến giá cầu thủ.
3. **Triển khai:** Xây dựng một Web App cho phép người dùng nhập tên, thông số cầu thủ và nhận về dự đoán giá trị theo thời gian thực tế.

IV. Individual Reflections

1. **Lê Minh Nhật (Nhóm trưởng - Modeling & Exploration)**

- **Thách thức:**
  - **Kỹ thuật:** Việc xử lý các biến phân loại có độ nhiễu cao như **current_club_name** (tên câu lạc bộ) và

country_of_citizenship (quốc tịch) là một bài toán hóc búa. Nếu sử dụng *One-Hot Encoding* thông thường, số lượng cột sẽ bùng nổ, gây ra "lời nguyền chiều dữ liệu".
    - **Phân tích:** Biến mục tiêu **market_value** ban đầu bị lệch phải rất nặng (vài siêu sao gánh phần lớn giá trị). Điều này khiến các mô hình học máy dễ bị chệch hướng bởi các điểm ngoại lai (outliers).
    - **Khái niệm:** Làm sao để định nghĩa "giá trị" khi nó không chỉ nằm ở bàn thắng? Việc kết hợp giữa số liệu hiệu suất (performance) và bối cảnh (context) của câu lạc bộ yêu cầu sự cân bằng tinh tế trong việc tạo đặc trưng (feature engineering).
    - **Thách thức lớn nhất** chính là **"Nghịch lý dữ liệu"**. Dữ liệu cho thấy danh tiếng CLB và giải đấu quan trọng hơn hiệu suất cá nhân. Là một người yêu và xem bóng đá nhiều năm, việc chấp nhận rằng một tiền đạo ghi 5 bàn ở Premier League có giá cao hơn tiền đạo ghi 20 bàn ở giải hạng thấp là một sự thật phũ phàng từ kết quả train model mà tôi phải thừa nhận qua con số thực tế.
- **Giải pháp:**
    - Sử dụng phép biến đổi Logarithm (***Ln***) cho **market_value** để đưa phân phối về dạng gần chuẩn, giúp mô hình hội tụ tốt hơn.
    - Tận dụng sức mạnh của **CatBoost**, một thuật toán được thiết kế riêng để xử lý các biến phân loại mà không cần tiền xử lý phức tạp, giúp giữ lại ý nghĩa của tên CLB hay giải đấu.
- **Bài học:**
    - **Kỹ năng kỹ thuật:** Thành thạo quy trình tối ưu hóa mô hình Boosting (XGBoost, CatBoost) và hiểu sâu về các độ đo sai số như MAE hay RMSE trong bài toán hồi quy.
    - **Cách tiếp cận phân tích:** Luôn luôn bắt đầu từ EDA (Khám phá dữ liệu). Biểu đồ tương quan (Correlation) đã chỉ dẫn tôi biết nên tập trung vào **total_minutes_played** thay vì chỉ nhìn vào số bàn thắng.
    - **Kiến thức Domain:** Hiểu về cấu trúc thị trường chuyển nhượng: Tuổi tác là rào cản, nhưng giải đấu là bệ phóng giá trị.

    - Điều bất ngờ là **last_season** (biến thời gian) có trọng số quan trọng nhất trong mô hình XGBoost (0.35). Điều này chứng tỏ lạm phát trong bóng đá là có thật và nó tác động mạnh mẽ đến mức làm lu mờ cả các chỉ số chuyên môn trong một số mô hình.
    - Dự án này cho thấy một điều: Mô hình CatBoost với $R^2 \approx 0.8$ cho thấy máy tính có thể nắm bắt "logic" của con người lên tới 80%. Nhưng 20% còn lại chính là "vẻ đẹp không thể dự đoán" của bóng đá: những cảm xúc, scandal, hay những bản hợp đồng điên rồ mà đa phần không có dữ liệu để tính toán, chỉ dựa trên cảm xúc của con người.
- **Kết luận rút ra:** Sau khi đi qua toàn bộ workflow từ xử lý dữ liệu thô đến khi ra được bảng xếp hạng hiệu suất giữa XGBoost và CatBoost, tôi rút ra kết luận: Mô hình **CatBoost** là sự lựa chọn ưu việt nhất cho bài toán này vì nó phản ánh đúng bản chất phân tầng của bóng đá (giải đấu > clb > cá nhân). Dự án này không chỉ là một bài tập về Machine Learning, mà là một lăng kính thực tế giúp ta thấy được thị trường chuyển nhượng vận hành theo những quy luật kinh tế chặt chẽ, đôi khi lạnh lùng, đằng sau những trận cầu rực lửa.

2. Huỳnh Đặng Ngọc Hân (Data Exploration & Preprocessing)

- **Thách thức:**
    - Dữ liệu thiếu quá nhiều, ngoài những giá trị trống thì có cả những giá trị được điền là “Missing” không để ý kỹ sẽ không thấy được. Cần xử lý missing values sao cho logic mà không làm mất bản chất của dữ liệu.
    - Khi tạo biến mới là ‘ga_per_90min’, các cầu thủ thi đấu dưới 90 phút một mùa sẽ có số ga cao vượt trội, gây nhiễu
    - Biến phân loại có quá nhiều giá trị. Nếu chỉ dùng one-hot encoding sẽ làm dữ liệu phình quá mức.
    - Các biến liên quan đến xuất thân của cầu thủ như “country_of_citizenship” hay “club” cần có thứ tự để model đánh giá được cầu thủ xuất thân như thế nào sẽ được đánh giá cao hơn.

    - Biến target là giá trị cầu thủ chênh lệch nhau quá nhiều.
    - Nguy cơ rò rỉ dữ liệu khi chia tập train/test.
- **Giải pháp:**
    - Với các biến định lượng, sử dụng median để điền cho giá trị trống. Với các biến phân loại, cân nhắc sử dụng mode với những biến có thể chấp nhận được và loại bỏ luôn nếu là biến quan trọng.
    - Với các cầu thủ thi đấu dưới 90 phút một mùa, số G/A mỗi 90 phút của cầu thủ được tính bằng đúng số bàn thắng (G) + số kiến tạo (A).
    - Với biến phân loại có nhiều giá trị và cần làm nổi bật thứ tự để so sánh, sử dụng target encoding để có model có thể so sánh được.
    - Sử dụng log transformation để nến khoảng cách với target để nén khoảng cách, giúp model dễ học hơn và tránh bị ảnh hưởng bởi outlier.
    - Để tránh rò rỉ dữ liệu, chia tập train test trước, sử dụng target encoding ở tập Train sau đó mới mapping sang tập Test.
- **Bài học:**
    - Học được cách encoding mới là target-encoding giúp xử lý được những biến có thứ tự và có nhiều giá trị khác nhau.
    - Phải xử lý dữ liệu nếu bị lệch nặng để tránh gây lỗi khi model học.
    - Không thể so sánh các mẫu có số lượng nhỏ với các mẫu có số lượng lớn, gây nên phán đoán sai.
    - Cần phải có kiến thức tốt về vấn đề mình đang làm để có thể đánh giá và xử lý một cách tốt nhất. Như ở đây là việc feature engineering những biến nào để tối ưu cho model.

3. Nguyễn Văn Khoa (Data Collection & Analysis)

- **Thách thức:**
    - Việc xác định tiêu chí cho các **CLB "hàng rẻ, chất lượng cao"** rất mơ hồ, khó cân bằng giữa giá trị chuyển nhượng thấp và hiệu suất thi đấu cao.

    - Dữ liệu quốc tịch khi phân tích **"Luồng di cư bóng đá"** có một số nơi không đồng nhất như Ukraine, Nga..
- **Giải pháp:**
    - Loại bỏ, không tính đến những nơi có dữ liệu quốc tịch không đồng nhất
    - Sử dụng thêm phương pháp thống kê để kiểm định sự chênh lệch có ý nghĩa hay không
    - Sử dụng biểu đồ phân tán (Scatter Plot) kết hợp đường xu hướng phi tuyến tính để nhìn ra được sự phân hóa về dữ liệu của các câu hỏi
- **Bài học:**
    - Học được cách chuyển đổi các câu hỏi nghiên cứu trừu tượng thành các biểu đồ trực quan và dễ hiểu.
    - Hiểu sâu hơn về thị trường chuyển nhượng: Giá trị cầu thủ không chỉ nằm ở tài năng sân cỏ mà còn bị chi phối mạnh bởi quốc tịch và thương hiệu giải đấu.