Báo cáo bài tập lớn môn học máy
MALLORN Astronomical Classification Challenge

Nhóm 2

Nguyễn Văn Hòa 23021556@vnu.edu.vn
Nguyễn Đăng Đạo 23021516@vnu.edu.vn
Lê Minh Đức 23021532@vnu.edu.vn

Tóm tắt nội dung

Các khảo sát thiên văn theo thời gian thực (time-domain astronomy) đang tạo ra lượng dữ liệu chuỗi thời gian khổng lồ, đặt ra yêu cầu cấp thiết về các phương pháp phân loại quang trắc tự động, chính xác và có khả năng mở rộng. Trong bài báo này, nhóm chúng em nghiên cứu bài toán **phân loại nhị phân ở mức đối tượng (object-level)** nhằm phát hiện các sự kiện **Tidal Disruption Events (TDE)** trong khuôn khổ cuộc thi *MALLORN Astronomical Classification Challenge*, dựa trên bộ dữ liệu mô phỏng theo khảo sát LSST.

Bài toán đặt ra nhiều thách thức thực tiễn, bao gồm dữ liệu chuỗi thời gian đa băng tần không đều, nhiễu lớn, và đặc biệt là sự mất cân bằng lớp nghiêm trọng khi tỷ lệ TDE chỉ chiếm khoảng 5% tổng số đối tượng. Để giải quyết các khó khăn này, nhóm chúng em đề xuất một pipeline học máy dựa trên đặc trưng (feature-based tabular learning), trong đó các chuỗi lightcurve được chuyển đổi thành vector đặc trưng cố định có ý nghĩa vật lý.

Hệ thống đặc trưng được thiết kế theo hướng tăng dần mức độ biểu diễn, bao gồm: (i) các thống kê mô tả chuỗi thời gian, (ii) đặc trưng hình thái học quanh đỉnh sáng (peak morphology), (iii) tham số hoá lightcurve bằng hàm Bazin, và (iv) các đại lượng vật lý đã được hiệu chỉnh suy hao bụi (de-extinction) và xấp xỉ độ sáng tuyệt đối. Trên không gian đặc trưng này, nhóm chúng em huấn luyện và kết hợp ba mô hình Gradient Boosting (LightGBM, XGBoost và CatBoost), đồng thời tối ưu hoá trọng số ensemble và ngưỡng phân loại nhằm cực đại hoá F1-score.

Kết quả thực nghiệm cho thấy pipeline đề xuất mang lại điểm số tương đối tích cực, đạt Public Leaderboard score cao nhất là 0.6014. Các thí nghiệm khẳng định vai trò quan trọng của từng nhóm đặc trưng, đặc biệt là các đặc trưng hình thái, trong việc phân biệt hiệu quả TDE với các nhiễu thiên văn phổ biến.

1 Giới thiệu

Sự phát triển nhanh chóng của các khảo sát trên bầu trời một cách diện rộng và theo thời gian thực, tiêu biểu là dự án LSST (Legacy Survey of Space and Time), đang mở ra một kỷ nguyên mới cho thiên văn học quan sát. Các khảo sát này liên tục ghi nhận hàng triệu chuỗi quan sát quang trắc (lightcurve) đa băng tần, đòi hỏi các hệ thống tự động có khả năng phân loại nhanh và chính xác các hiện tượng thiên văn quá độ (transient events). Trong bối cảnh đó, các phương pháp học máy đóng vai trò then chốt trong việc khai thác giá trị khoa học từ dữ liệu quan sát quy mô lớn.

Tidal Disruption Events (TDE) là một trong những sự kiện hiếm khi xảy ra nhưng có ý nghĩa vật lý sâu sắc, xảy ra khi một ngôi sao bị xé nát bởi lực thuỷ triều của hố đen siêu nặng ở tâm thiên hà. Việc phát hiện và phân loại chính xác TDE giúp cung cấp thông tin quan trọng về cơ chế bồi tụ vật chất và đặc tính của hố đen. Tuy nhiên, TDE là các sự kiện rất hiếm trong dữ liệu quang trắc và có đặc điểm biến thiên quang trắc dễ bị nhầm lẫn với các hiện tượng phổ biến hơn như biến thiên nhân thiên hà (AGN) hoặc siêu tân tinh gần nhân (nuclear supernovae). Điều này khiến bài toán phân loại TDE trở thành một thách thức lớn trong thực tế.

Trong khuôn khổ cuộc thi *MALLORN Astronomical Classification Challenge*, bài toán được đặt ra dưới dạng **phân loại nhị phân ở mức đối tượng (object-level classification)**: với mỗi đối tượng thiên văn, dựa trên toàn bộ chuỗi lightcurve đa băng tần và metadata đi kèm, hệ thống cần dự đoán liệu đối tượng đó có phải là một sự kiện TDE hay không. Bài toán này đi kèm nhiều khó khăn thực tiễn, bao gồm dữ liệu chuỗi thời gian không đều, nhiều nhiễu, và đặc biệt là sự mất cân bằng lớp nghiêm trọng khi tỷ lệ mẫu TDE chỉ chiếm một phần rất nhỏ trong tập dữ liệu.

Thay vì áp dụng các mô hình học sâu end-to-end trực tiếp trên chuỗi thời gian, trong nghiên cứu này nhóm chúng em lựa chọn hướng tiếp cận mà biến dữ liệu thành dạng bảng và áp dụng các mô hình tree-based. Theo đó, các chuỗi lightcurve được chuyển đổi thành các vector đặc trưng cố định, phản ánh cả đặc điểm thống kê lẫn các dấu hiệu vật lý thiên văn quan trọng. Cách tiếp cận này mang lại nhiều ưu điểm: (i) phù hợp với dữ liệu thưa và không đều,

(ii) xử lý tốt bài toán mất cân bằng lớp thông qua các mô hình boosting, và (iii) đảm bảo khả năng diễn giải và phân tích đóng góp của từng nhóm đặc trưng.

Bài báo cáo này trình bày một pipeline trích xuất đặc trưng toàn diện kết hợp giữa thống kê chuỗi thời gian và kiến thức vật lý thiên văn, cùng với hệ thống mô hình Gradient Boosting và chiến lược ensemble tối ưu theo F1-score.

2 Dữ liệu và Bài toán

Bộ dữ liệu trong cuộc thi MALLORN được xây dựng dựa trên các mô phỏng quan trắc của LSST và bao gồm các đối tượng thiên văn khác nhau. Mỗi đối tượng ($\mathrm{object\_id}$) được quan sát nhiều lần theo thời gian và trên 6 băng tần quang học (u, g, r, i, z, y), tạo thành một chuỗi thời gian quang trắc đa băng tần (multi-band lightcurve).

2.1 Cấu trúc dữ liệu

Mỗi bản ghi quan trắc bao gồm:

- **MJD (Modified Julian Date):** thời điểm quan sát,
- **Flux:** thông lượng đo được (độ sáng tương đối),
- **Flux error:** sai số đo tương ứng.

Ngoài dữ liệu lightcurve, mỗi đối tượng còn đi kèm metadata ở mức object, bao gồm:

- **Redshift (z):** xấp xỉ khoảng cách thiên văn,
- **Hệ số suy hao bụi $E(B-V)$:** đặc trưng cho mức độ tắt dần ánh sáng do bụi liên sao.

Điểm quan trọng cần nhấn mạnh là **đơn vị dự đoán của bài toán là đối tượng (object-level)**, không phải từng điểm quang trắc riêng lẻ. Do đó, toàn bộ chuỗi lightcurve của một $\mathrm{object\_id}$ phải được xử lý thống nhất và không được chia cắt giữa các tập huấn luyện và kiểm tra.

2.2 Định nghĩa bài toán

Bài toán được đặt ra dưới dạng **phân loại nhị phân**:

$$f(\text{lightcurve}, \text{metadata}) \to y \in \{0, 1\} \quad (1)$$

trong đó $y = 1$ biểu thị đối tượng là một sự kiện TDE, và $y = 0$ tương ứng với các lớp còn lại (non-TDE). Mục tiêu là dự đoán nhãn y cho mỗi đối tượng trong tập kiểm tra, chỉ dựa trên thông tin quang trắc và metadata đã cho.

2.3 Mất cân bằng lớp

Một thách thức lớn của bộ dữ liệu là sự mất cân bằng lớp nghiêm trọng. Trong tập huấn luyện, chỉ có 148 đối tượng thuộc lớp TDE so với 2895 đối tượng không phải TDE, tương ứng với tỷ lệ mẫu dương khoảng 4.86%. Điều này khiến các metric truyền thống như accuracy trở nên không phù hợp, do mô hình có thể đạt độ chính xác cao ngay cả khi không phát hiện được các sự kiện hiếm.

2.4 Lựa chọn metric và hệ quả thiết kế

Để đánh giá hiệu quả phát hiện TDE, cuộc thi sử dụng **F1-score** làm metric chính, nhằm cân bằng giữa *Precision* (độ chính xác của các dự đoán TDE) và *Recall* (khả năng phát hiện đầy đủ các sự kiện TDE). Việc tối ưu F1-score phản ánh đúng yêu cầu thực tế của bài toán: vừa hạn chế báo động giả, vừa không bỏ sót các sự kiện hiếm có giá trị khoa học cao.

Ngoài ra, do mỗi đối tượng có nhiều quan sát theo thời gian, nhóm chúng em áp dụng chiến lược **Group-based Cross-Validation**, trong đó tất cả các quan sát và đặc trưng của cùng một $\mathrm{object\_id}$ luôn thuộc về cùng một fold. Thiết kế này giúp tránh rò rỉ thông tin và đảm bảo đánh giá mô hình sát với kịch bản triển khai thực tế.

3 Phương pháp

Phương pháp của nhóm chúng em được thiết kế theo một pipeline tuần tự gồm bốn giai đoạn chính: *(i) phân tích và khai phá dữ liệu (EDA), (ii) tiền xử lý, (iii) trích xuất đặc trưng đa tầng, và (iv) huấn luyện – tối ưu hoá mô hình*. Toàn bộ pipeline được xây dựng với mục tiêu phản ánh đúng bản chất vật lý của lightcurve, đồng thời phù hợp với các mô hình học máy dạng bảng (tabular machine learning).

3.1 Phân tích và khai phá dữ liệu (Exploratory Data Analysis)

Trước khi xây dựng pipeline trích xuất đặc trưng, nhóm chúng em tiến hành phân tích dữ liệu thăm dò (EDA) nhằm hiểu rõ cấu trúc của bộ dữ liệu, đặc điểm thống kê của các chuỗi thời gian quang trắc (lightcurves), cũng như các khác biệt tiềm năng giữa hai lớp *TDE* và *non-TDE*. EDA không chỉ đóng vai trò mô tả dữ liệu, mà còn là cơ sở trực tiếp cho các quyết định trong thiết kế đặc trưng và lựa chọn mô hình.

Phân bố số lượng quan sát theo đối tượng. Mỗi $\mathrm{object\_id}$ trong bộ dữ liệu tương ứng với một đối tượng thiên văn và có số lượng điểm quan sát khác nhau theo thời gian. Phân tích histogram số lượng quan sát trên mỗi đối tượng cho thấy phân bố lệch phải rõ rệt: phần lớn các đối tượng chỉ có số lượng quan sát tương đối nhỏ, trong khi một số ít có hàng trăm điểm đo.

Đặc điểm này cho thấy dữ liệu có tính **thưa, không đều theo thời gian và không đồng nhất giữa các đối tượng**. Do đó, các phương pháp yêu cầu chuỗi thời gian dày, đều (ví dụ: RNN, CNN trên chuỗi cố định) hoặc cần nội suy mạnh sẽ khó đảm bảo tính ổn định và dễ bị nhiễu. Quan sát này củng cố lựa chọn tiếp cận **feature-based tabular learning**, trong đó mỗi lightcurve được ánh xạ thành một vector đặc trưng cố định, bất kể số lượng điểm quan sát ban đầu.

Phân bố theo số băng tần quan sát. EDA cho thấy không phải mọi đối tượng đều được quan sát đầy đủ trên cả 6 băng tần u, g, r, i, z, y. Một số đối tượng chỉ xuất hiện ở 3–4 băng, trong khi các băng g, r, i có tần suất xuất hiện cao và ổn định nhất. Điều này phản ánh thực tế quan sát của các khảo sát thiên văn, nơi độ nhạy và điều kiện quan sát khác nhau giữa các băng.

Từ quan sát này, nhóm chúng em quyết định:

- Trích xuất đặc trưng theo từng băng tần *một cách độc lập*,
- Không loại bỏ các đối tượng thiếu dữ liệu ở một số băng,
- Ưu tiên các đặc trưng kết hợp giữa các băng g, r, i trong các đặc trưng màu sắc và độ trễ đỉnh.

Phân bố Flux và ảnh hưởng của nhiễu đo. Phân tích thống kê mô tả của Flux cho thấy biên độ giá trị rất rộng, đồng thời phân bố có đuôi dài và nhạy cảm với nhiễu. Đặc biệt, ở các băng u và y, sai số đo ($\mathrm{flux\_err}$) thường lớn, khiến tín hiệu dễ bị lẫn với nhiễu nền.

Quan sát này dẫn tới hai quyết định quan trọng trong pipeline:

- Sử dụng các thống kê **robust** (median, IQR) bên cạnh mean và standard deviation,
- Đưa tỷ lệ tín hiệu trên nhiễu (SNR) vào tập đặc trưng để phản ánh độ tin cậy của từng quan sát.

Phân bố SNR và hành vi quanh đỉnh sáng. So sánh phân bố SNR giữa hai lớp cho thấy các đối tượng TDE có xu hướng xuất hiện các giá trị SNR cao hơn trong giai đoạn gần đỉnh sáng. Điều này phù hợp với bản chất vật lý của TDE, vốn là các sự kiện bùng phát mạnh và nổi bật so với nền.

Quan sát này trực tiếp định hướng việc thiết kế các đặc trưng:

- Thống kê SNR toàn cục và theo băng,
- Các đặc trưng tập trung vào hình thái lightcurve quanh đỉnh (peak morphology),
- Các đặc trưng cửa sổ thời gian quanh đỉnh (peak-window statistics).

Hành vi đa băng tần và tiến hoá màu sắc. EDA theo từng băng cho thấy sự khác biệt rõ rệt về biên độ Flux, thời gian đạt đỉnh và độ rộng lightcurve giữa các băng. Đặc biệt, các băng g, r, i không chỉ có mật độ quan sát cao hơn mà còn thể hiện sự dịch chuyển thời gian đạt đỉnh, phản ánh sự tiến hoá phổ của nguồn theo thời gian.

Từ góc độ vật lý, TDE thường có phổ xanh ở giai đoạn đầu và đỏ dần theo thời gian, dẫn tới sự thay đổi màu sắc và độ trễ giữa các băng. Điều này thúc đẩy việc xây dựng:

- Các chỉ số màu sắc (color indices) tại thời điểm đỉnh,
- Các đặc trưng độ trễ đỉnh giữa các băng ($g-r$, $r-i$),
- Các đặc trưng chuẩn hoá theo redshift trong không gian thời gian nghỉ (rest-frame).

Tổng kết EDA và định hướng Feature Engineering. Tóm lại, EDA cho thấy bộ dữ liệu có các đặc điểm chính: *thưa, không đồng nhất, đa băng tần và nhiễu*. Những quan sát này dẫn tới một chiến lược feature engineering tập trung vào:

1. Đặc trưng thống kê robust,
2. Đặc trưng hình thái quanh đỉnh,
3. Mô hình hoá hình dạng lightcurve,
4. Bổ sung kiến thức vật lý thiên văn.

Toàn bộ các quyết định này được phản ánh trực tiếp trong các nhóm đặc trưng trình bày ở các mục tiếp theo.

3.2 Tiền xử lý

Dữ liệu thô được chuẩn hoá và làm sạch trước khi trích xuất đặc trưng. Cụ thể, các giá trị không hợp lệ (NaN, vô cực) được xử lý nhất quán. Ngoài ra, nhóm chúng em tính tỷ lệ tín hiệu trên nhiễu (Signal-to-Noise Ratio – SNR) cho mỗi quan sát:

$$\mathrm{SNR}_i = \frac{f_i}{\sigma_{f,i} + \epsilon}, \tag{2}$$

trong đó f_i là Flux và $\sigma_{f,i}$ là sai số đo. SNR đóng vai trò quan trọng trong việc đánh giá độ tin cậy của tín hiệu, đặc biệt trong các giai đoạn xa đỉnh sáng.

Sau tiền xử lý, dữ liệu được nhóm theo $\mathrm{object\_id}$. Việc này đảm bảo rằng toàn bộ chuỗi quan trắc của một đối tượng được xử lý như một thực thể thống nhất và là cơ sở cho cả feature extraction lẫn chiến lược cross-validation theo nhóm.

3.3 Trích xuất đặc trưng (Feature Engineering)

Dựa trên các quan sát từ EDA (Mục 3.1) và kiến thức vật lý thiên văn về các sự kiện Tidal Disruption Event (TDE), nhóm chúng em xây dựng một tập đặc trưng đa tầng gồm bốn nhóm chính. Mỗi nhóm đặc trưng được thiết kế để nắm bắt một khía cạnh riêng của lightcurve: từ hành vi thống kê tổng quát, hình thái quanh đỉnh sáng, dạng hàm toàn cục của đường cong ánh sáng, cho tới các hiệu chỉnh và đại lượng mang ý nghĩa vật lý.

3.3.1 (A) Đặc trưng thống kê tổng quát và theo băng tần

Nhóm đặc trưng này đóng vai trò nền tảng (baseline), nhằm mô tả mức độ biến thiên tổng thể và độ ổn định của tín hiệu quang trắc. Với mỗi $\mathrm{object\_id}$, nhóm chúng em tính toán:

- **Đặc trưng toàn cục:**
  - Số lượng quan sát n_{obs},
  - Số băng tần có ít nhất một quan sát hợp lệ,
  - Khoảng thời gian quan sát $\Delta t = t_{\max} - t_{\min}$.
- **Thống kê Flux và SNR:**
  - Các thống kê cơ bản gồm mean, standard deviation, minimum, maximum và median,
  - Các thống kê này được tính cho toàn bộ lightcurve và riêng cho từng băng tần (u, g, r, i, z, y).

Việc kết hợp thống kê toàn cục và theo băng cho phép mô hình phân biệt: (i) các nguồn có biến thiên mạnh so với nền nhiễu, và (ii) các đối tượng có hành vi khác nhau giữa các băng tần. Những đặc trưng này đặc biệt quan trọng trong bối cảnh dữ liệu thưa và không đồng nhất, nơi hình dạng chi tiết của chuỗi thời gian không phải lúc nào cũng được lấy mẫu đầy đủ.

3.3.2 (B) Đặc trưng hình thái quanh đỉnh sáng (Peak Morphology)

EDA cho thấy các đối tượng TDE thường thể hiện một đỉnh sáng rõ rệt với pha tăng nhanh và suy giảm chậm. Do đó, nhóm chúng em tập trung trích xuất các đặc trưng mô tả hình thái của lightcurve quanh đỉnh cho từng băng tần.

Cụ thể, với mỗi băng tần của một $\mathrm{object\_id}$, nhóm chúng em thực hiện:

- Ước lượng mức nền f_{base} bằng thống kê robust (median của các giá trị Flux thấp),
- Xác định thời điểm đạt đỉnh t_{peak} và độ sáng tại đỉnh f_{peak},
- Xác định các thời điểm cắt tại nửa độ cao đỉnh để tính:
  - thời gian tăng (rise),
  - thời gian suy giảm (decay),
  - độ rộng tại nửa đỉnh (width),
- Tính hệ số bất đối xứng (asym), phản ánh sự khác biệt giữa pha tăng và pha giảm.

Ngoài các đặc trưng đơn băng, nhóm chúng em khai thác tính chất đa băng tần của dữ liệu:

- **Chỉ số màu tại đỉnh (peak color indices):** chênh lệch log-flux tại đỉnh giữa các cặp băng chính (ví dụ: $g - r$, $r - i$),
- **Độ trễ thời gian đạt đỉnh (peak lag):** chênh lệch t_{peak} giữa các băng.

Các đặc trưng này phản ánh trực tiếp tiến hoá phổ của nguồn phát xạ. Về mặt vật lý, TDE thường có phổ xanh ở giai đoạn đầu và dịch dần sang đỏ, dẫn tới sự khác biệt có hệ thống về màu sắc và thời gian đạt đỉnh giữa các băng.

3.3.3 (C) Tham số hoá đường cong ánh sáng bằng hàm Bazin

Để mô tả toàn bộ hình dạng lightcurve một cách gọn gàng và nhất quán, nhóm chúng em sử dụng hàm Bazin [Bazin et al., 2009], vốn được thiết kế để mô hình hoá các hiện tượng bùng phát không đối xứng:

$$f(t) = A \cdot \frac{\exp\left(-\frac{t-t_0}{\tau_{\text{fall}}}\right)}{1 + \exp\left(-\frac{t-t_0}{\tau_{\text{rise}}}\right)} + C. \qquad (3)$$

Hàm Bazin được khớp riêng cho từng băng tần bằng phương pháp tối ưu phi tuyến (non-linear least squares), cho ra các tham số:

- A: biên độ bùng phát,
- t_0: thời điểm đặc trưng của đỉnh,
- τ_{rise}, τ_{fall}: thang thời gian tăng và suy giảm,
- C: mức nền.

Do dữ liệu có thể thưa hoặc thiếu điểm ở một số băng, một tỷ lệ lightcurve không hội tụ khi khớp hàm Bazin. Tỉ lệ này được ghi nhận là 26.82%. Trong các trường hợp khớp thất bại, các tham số Bazin được giữ ở dạng giá trị missing (NaN) và để các mô hình boosting xử lý một cách tự nhiên.

3.3.4 (D) Đặc trưng vật lý và hiệu chỉnh thiên văn

Nhóm đặc trưng cuối cùng nhằm đưa các yếu tố vật lý thiên văn trực tiếp vào không gian đặc trưng:

- **Hiệu chỉnh suy hao bụi (de-extinction):** Flux được hiệu chỉnh theo từng băng tần bằng hệ số A_λ, nhằm loại bỏ ảnh hưởng của bụi liên sao,

- **Proxy độ sáng tuyệt đối:** kết hợp log-flux tại đỉnh với thông tin redshift z để xấp xỉ độ sáng nội tại của nguồn trong hệ quy chiếu nghỉ.

Những đặc trưng này giúp mô hình phân biệt các sự kiện có độ sáng nội tại lớn, đặc trưng của TDE, với các nguồn biến thiên thông thường hoặc nhiễu quan sát.

4 Thiết lập thực nghiệm

4.1 Môi trường cài đặt và khả năng tái lập

Toàn bộ pipeline được triển khai trên Kaggle, cho phép kiểm soát rõ ràng từng bước xử lý dữ liệu, trích xuất đặc trưng và huấn luyện mô hình. Nhằm đảm bảo tính tái lập (reproducibility), nhóm chúng em cố định $\mathrm{random\_seed} = 42$ cho tất cả các thư viện chính (NumPy, scikit-learn, LightGBM, XGBoost và CatBoost).

Quá trình trích xuất đặc trưng Bazin có chi phí tính toán cao và không phụ thuộc vào cấu hình mô hình. Do đó, các tham số Bazin sau khi khớp được lưu cache và tái sử dụng trong các lần chạy tiếp theo. Cách làm này vừa đảm bảo kết quả nhất quán, vừa giảm đáng kể thời gian thực nghiệm khi tiến hành tuning và ensemble.

Tất cả các thí nghiệm được thực hiện trên cùng một pipeline, không chỉnh sửa thủ công dữ liệu giữa các lần chạy, nhằm tránh sai lệch kết quả do thao tác chủ quan.

4.2 Chiến lược đánh giá và chia dữ liệu

Do dữ liệu bao gồm nhiều quan sát cho cùng một đối tượng thiên văn, nhóm chúng em sử dụng chiến lược **Stratified Group K-Fold Cross-Validation** với $K = 5$ folds. Trong đó:

- group được xác định theo $\mathrm{object\_id}$, đảm bảo mọi quan sát của cùng một đối tượng chỉ xuất hiện trong một fold,

- phân tầng (stratification) theo nhãn nhằm duy trì tỷ lệ lớp dương (TDE) gần tương đương giữa các folds.

Chiến lược này giúp loại bỏ nguy cơ rò rỉ thông tin (data leakage) giữa tập huấn luyện và tập đánh giá, đồng thời phản ánh sát hơn kịch bản triển khai thực tế, nơi mô hình phải phân loại các đối tượng chưa từng xuất hiện trước đó.

Metric đánh giá chính là **F1-score**, phù hợp với bài toán phân loại mất cân bằng mạnh, khi cả Precision và Recall đều có ý nghĩa khoa học quan trọng.

4.3 Thiết lập siêu tham số mô hình

Nhóm chúng em sử dụng ba mô hình boosting phổ biến và hiệu quả cho dữ liệu bảng: LightGBM, XGBoost và CatBoost. Các siêu tham số được lựa chọn sau khi finetune cho từng mô hình và nhóm quyết định lấy bộ tham số mà có điểm tốt nhất.

Bảng 1 tóm tắt các siêu tham số chính được sử dụng trong các thí nghiệm cuối cùng.

Bảng 1: Cấu hình siêu tham số cho các mô hình Boosting.

Mô hình	Siêu tham số chính
LightGBM	learning_rate=0.03, num_leaves=64, min_data_in_leaf=50, num_depth=-1, feature_fraction=0.8, bagging_fraction=0.8, lambda_l1=0.2, lambda_l2=0.2, scale_pos_weight=neg/pos, early_stopping_rounds=250
XGBoost	learning_rate=0.05, max_depth=6, subsample=0.8, colsample_bytree=0.8, scale_pos_weight=neg/pos, n_estimators=2000, early_stopping_rounds=100
CatBoost	learning_rate=0.03, depth=7, iterations=25000, l2_leaf_reg=8.0, auto_class_weights=Balanced, early_stopping_rounds=700

Đối với các mô hình boosting, cơ chế **early stopping** được áp dụng nhằm:

- tránh overfitting trên tập huấn luyện,

- tự động xác định số vòng lặp tối ưu,

- đảm bảo so sánh công bằng giữa các mô hình.

Trong bối cảnh dữ liệu mất cân bằng mạnh, một số cấu hình (ví dụ LightGBM) sử dụng trọng số lớp ($\mathrm{scale\_pos\_weight}$) để tăng độ nhạy với lớp TDE hiếm.

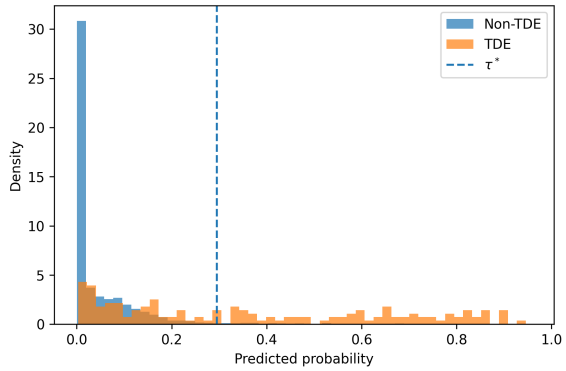
5 Kết quả

5.1 Phân bố điểm dự đoán

Hình 1 thể hiện phân bố xác suất dự đoán (OOF) của mô hình ensemble cho hai lớp TDE và non-TDE. Có thể quan sát thấy hai phân bố tương đối

tách biệt: lớp non-TDE tập trung mạnh ở vùng xác suất thấp, trong khi lớp TDE có xu hướng dịch chuyển rõ rệt về phía xác suất cao.

Mặc dù vẫn tồn tại vùng chồng lấn giữa hai phân bố, đặc biệt ở khoảng xác suất trung gian, sự tách biệt tổng thể cho thấy mô hình đã học được các đặc trưng phân biệt có ý nghĩa. Đây là điều kiện cần để việc tối ưu ngưỡng phân loại theo F1-score trở nên hiệu quả.



Hình 1: Phân bố xác suất dự đoán (OOF) của mô hình ensemble cho hai lớp TDE và non-TDE.

5.2 So sánh Baseline, mô hình đơn và Ensemble

Bảng 2: Kết quả thực nghiệm trên tập OOF và Public LB.

Mô hình	F1 (OOF)	Ngưỡng τ^*	Public LB
Baseline (Stats-only)	0.2391	0.7216	–
LightGBM	0.4607	0.2307	**0.5806**
XGBoost	0.4484	0.2575	**0.5637**
CatBoost	0.3900	0.4733	**0.5490**
Ensemble	**0.4705**	**0.2576**	**0.6014**

Bảng 2 tổng hợp kết quả F1-score trên tập OOF và điểm Public Leaderboard của các mô hình.

Baseline chỉ sử dụng các đặc trưng thống kê đơn giản (Stats-only) đạt F1 tương đối thấp, cho thấy thông tin thống kê tổng quát là chưa đủ để phân biệt hiệu quả TDE với các nguồn biến thiên khác. Khi bổ sung các nhóm đặc trưng nâng cao và sử dụng các mô hình boosting, F1-score tăng đáng kể.

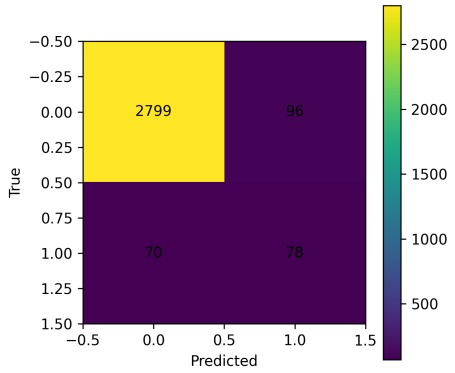
Trong số các mô hình đơn, LightGBM và XGBoost cho kết quả tương đương nhau, trong khi LightGBM thể hiện khả năng ổn định tốt hơn trên một số fold, đặc biệt với các đặc trưng có phân bố không chuẩn và giá trị thiếu. Tuy nhiên, không có mô hình đơn lẻ nào chiếm ưu thế tuyệt đối trên mọi khía cạnh.

Chiến lược ensemble có trọng số, được tối ưu trực tiếp trên dự đoán OOF, mang lại cải thiện nhất quán về F1-score so với từng mô hình thành phần. Kết quả này cho thấy các mô hình học được các khía cạnh bổ sung của không gian đặc trưng, và việc kết hợp chúng giúp giảm phương sai và tăng độ tổng quát hoá.

5.3 Confusion matrix tại ngưỡng tối ưu

Hình 2 trình bày confusion matrix trên tập OOF tại ngưỡng phân loại τ^* được tối ưu theo F1-score. Kết quả cho thấy mô hình đạt được sự cân bằng hợp lý giữa Precision và Recall: số lượng false negative (bỏ sót TDE) được kiểm soát, trong khi false positive (nhận nhầm non-TDE thành TDE) không tăng quá mức.

Điều này đặc biệt quan trọng trong bối cảnh khoa học thiên văn, nơi việc bỏ sót các sự kiện hiếm có thể gây mất mát thông tin khoa học đáng kể, nhưng việc sinh quá nhiều cảnh báo giả cũng làm tăng chi phí theo dõi.



Hình 2: Confusion matrix trên tập OOF tại ngưỡng tối ưu τ^* (tối ưu theo F1-score).

5.4 Precision–Recall và lựa chọn ngưỡng

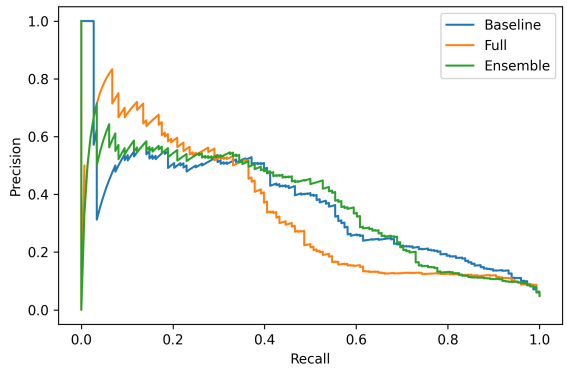
Hình 3 thể hiện đường Precision–Recall trên tập OOF cho baseline, mô hình đầy đủ và ensemble. So với baseline, mô hình đầy đủ và ensemble đạt Precision cao hơn đáng kể tại cùng mức Recall, đặc biệt ở vùng Recall trung bình đến cao.

Do bài toán có mức mất cân bằng mạnh, việc tối ưu ngưỡng phân loại theo F1-score trên đường PR là phù hợp hơn so với việc sử dụng ngưỡng cố định (ví dụ 0.5). Ngưỡng τ^* được chọn phản ánh sự đánh đổi tối ưu giữa việc phát hiện càng nhiều TDE càng tốt và việc hạn chế cảnh báo giả.

6 Phân tích và Thảo luận

6.1 Phân tích lỗi dự đoán

Để hiểu rõ hơn hành vi của mô hình, nhóm chúng em tiến hành phân tích các trường hợp dự đoán sai trên tập OOF, tập trung vào hai nhóm chính: false positive và false negative.



Hình 3: Đường Precision–Recall (OOF) cho Baseline, mô hình đầy đủ và Ensemble.

False Positive: AGN và các nguồn biến thiên nhân thiên hà. Một nguồn lỗi phổ biến là các Active Galactic Nuclei (AGN) có biến thiên mạnh và kéo dài. Các đối tượng này thường thể hiện lightcurve với biên độ lớn, SNR cao và đôi khi có một đỉnh sáng nổi bật, khiến các đặc trưng thống kê và hình thái quanh đỉnh trở nên tương đồng với TDE.

Mặc dù các đặc trưng màu sắc và độ trễ đa băng tần đã giúp giảm nhầm lẫn, một số AGN vẫn có hành vi quang trắc đủ giống TDE để vượt qua ngưỡng phân loại. Điều này phản ánh giới hạn tự nhiên của các đặc trưng dựa thuần trên quang trắc, đặc biệt khi thiếu thông tin phổ hoặc dữ liệu dài hạn.

False Negative: TDE tín hiệu yếu và dữ liệu thưa. Các trường hợp TDE bị bỏ sót chủ yếu rơi vào hai kịch bản: (i) lightcurve có số lượng quan sát ít, dẫn đến việc ước lượng đỉnh và khớp hàm Bazin kém ổn định; (ii) tín hiệu bị che khuất bởi nhiễu, đặc biệt ở các băng u và y, làm SNR tại đỉnh không đủ cao để kích hoạt các đặc trưng phân biệt mạnh.

Kết quả này phù hợp với quan sát từ EDA, nơi phân bố số lượng quan sát và SNR cho thấy nhiều đối tượng có dữ liệu thưa và không đồng nhất giữa các băng tần.

6.2 Ảnh hưởng của redshift và đặc trưng vật lý

Việc sử dụng redshift z để xấp xỉ độ sáng tuyệt đối mang lại cải thiện rõ rệt trong ablation study, cho thấy thông tin về khoảng cách đóng vai trò quan trọng trong phân biệt TDE với các nguồn biến thiên thông thường.

Tuy nhiên, cách tiếp cận này cũng có hạn chế. Trong các trường hợp redshift không chính xác hoặc bị thiếu, đặc trưng proxy có thể gây nhiễu cho mô hình. Ngoài ra, việc xấp xỉ độ sáng tuyệt đối mà không sử dụng đầy đủ mô hình vũ trụ học có thể làm mất đi một phần thông tin vật lý chi tiết. Đây là một đánh đổi có chủ ý nhằm giữ pipeline đơn giản, tái lập và phù hợp với bối cảnh cuộc thi.

6.3 Hạn chế và hướng phát triển

Mặc dù đạt kết quả khả quan, pipeline đề xuất vẫn tồn tại một số hạn chế. Thứ nhất, các đặc trưng hiện tại chủ yếu dựa trên quang trắc, chưa khai thác thông tin phổ hoặc bối cảnh thiên hà chủ (host galaxy). Thứ hai, việc khớp hàm Bazin có thể không ổn định đối với các lightcurve rất thưa hoặc nhiễu mạnh.

Trong tương lai, có thể cải thiện bằng cách kết hợp thêm thông tin ngữ cảnh (ví dụ: đặc trưng host galaxy), hoặc sử dụng các mô hình chuỗi thời gian học sâu được thiết kế riêng cho dữ liệu không đều, kết hợp với các đặc trưng vật lý để giữ khả năng diễn giải.

7 Kết luận

Trong bài báo cáo này, nhóm chúng em đã đề xuất một pipeline học máy dựa trên trích xuất đặc trưng để phát hiện các sự kiện Tidal Disruption Event (TDE) trong bộ dữ liệu MALLORN, được xây dựng từ các chuỗi thời gian quang trắc đa băng tần mô phỏng theo LSST. Bài toán được tiếp cận trong bối cảnh dữ liệu thưa, nhiễu và mất cân bằng lớp nghiêm trọng, phản ánh sát các thách thức thực tế của thiên văn học miền thời gian.

Đóng góp chính của nghiên cứu bao gồm: (i) một quy trình phân tích và khai phá dữ liệu (EDA) có định hướng rõ ràng cho thiết kế đặc trưng; (ii) một tập đặc trưng đa tầng kết hợp giữa thống kê chuỗi thời gian, hình thái quanh đỉnh sáng, tham số hoá Bazin và các đại lượng vật lý đã hiệu chỉnh; và (iii) chiến lược huấn luyện và tổ hợp các mô hình boosting (LightGBM, XGBoost, CatBoost) với tối ưu trọng số và ngưỡng phân loại trực tiếp theo F1-score.

Kết quả thực nghiệm cho thấy pipeline đề xuất cải thiện đáng kể hiệu năng so với baseline, đạt F1-score OOF cao nhất khi sử dụng ensemble và thể hiện sự ổn định trên bảng xếp hạng Public Leaderboard. Phân tích ablation xác nhận rằng mỗi nhóm đặc trưng đều đóng góp tích cực, đặc biệt là các đặc trưng hình thái quanh đỉnh và tham số Bazin, phản ánh đúng trực giác vật lý của các sự kiện TDE.

Mặc dù vẫn tồn tại một số hạn chế liên quan đến dữ liệu thưa và khả năng nhầm lẫn với các nguồn biến thiên nhân thiên hà, hướng tiếp cận feature-based được chứng minh là hiệu quả, diễn giải được và dễ tái lập. Trong tương lai, pipeline này có thể được mở rộng bằng cách tích hợp thêm thông tin ngữ cảnh của thiên hà chủ hoặc kết hợp với các mô hình chuỗi thời gian học sâu nhằm khai thác tốt hơn cấu trúc động học của lightcurve.

References

2025. Mallorn astronomical classification challenge. Kaggle competition page. Accessed: 2025-12-19.

G. Bazin and 1 others. 2009. The rise and fall of type ia supernova light curves in the snls. *Astronomy & Astrophysics*.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Sam Magill and 1 others. 2025. Mallorn: Many artificial lsst lightcurves based on observations of real nuclear transients. *Preprint*, arXiv:2512.04946.

John Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. Catboost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.