TRƯỜNG ĐẠI HỌC CÔNG NGHỆ
ĐẠI HỌC QUỐC GIA HÀ NỘI
—o0o—

MALLORN ASTRONOMICAL CLASSIFICATION CHALLENGE

Giảng viên: TS. Lê Đức Trọng

Môn học: Học máy

Lớp học phần: INT3405E_4

Thành viên nhóm:

Ngô Thị Ngọc Linh	23021608
Nguyễn Thị Hải Yến	23021756
Vũ Thùy Linh	23021612

Hà Nội – 2025

I. MÔ TẢ BÀI TOÁN:

1. Giới thiệu bài toán:

Bối cảnh và Thách thức: Đài thiên văn Vera C. Rubin sắp tiến hành Khảo sát LSST 10 năm, dự kiến phát hiện gấp 100 lần số lượng thiên thể thoáng qua (như siêu tân tinh) so với trước đây. Thách thức lớn là không đủ nguồn lực để phân loại quang phổ tất cả thiên thể, do đó cần xác định thiên thể quan trọng chỉ dựa vào đường cong ánh sáng.

Mục tiêu: Phát hiện TDEs

Sự kiện Xé rách Thủy triều (TDEs) xảy ra khi ngôi sao bị xé toạc bởi lực hấp dẫn của lỗ đen siêu khối lượng. Đây là hiện tượng hiếm (~100 quan sát) nhưng có giá trị khoa học cao trong nghiên cứu lỗ đen. LSST có thể mở rộng đáng kể mẫu nghiên cứu nếu phân loại hiệu quả.

Nhiệm vụ: Phát triển thuật toán học máy phân loại TDEs từ đường cong ánh sáng mô phỏng LSST (dựa trên dữ liệu thực từ ZTF). Dự đoán nhị phân: TDE (1) hoặc không phải TDE (0).

Tiêu chí đánh giá: Chỉ số F1

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

Chọn F1 vì dữ liệu mất cân bằng cao (TDEs rất hiếm), cân bằng giữa phát hiện đúng và tránh dương tính giả.

2. Mô tả dữ liệu:

Cấu trúc Files:

- Split [01-20]: Dữ liệu chia 20 phần bằng nhau
- train/test_full_lightcurves.csv: Chuỗi quan sát theo thời gian và bộ lọc
- train/test_log.csv: Thông tin bổ sung (redshift, tàn quang, phân loại, target)
- sample_submission.csv: Mẫu file nộp bài

Các cột chính trong Log File

Cột	Mô tả

object_id	Mã định danh (3 từ tiếng Sindarin)
Z	Độ dịch chuyển đỏ (train: chính xác; test: có sai số)
Z_err	Sai số redshift (chỉ test)
EBV	Hệ số tàn quang - đo mức độ che khuất bởi bụi
SpecType	Loại thiên thể thực (chỉ train)
target	1 = TDE, 0 = không phải TDE (chỉ train)

Các cột trong Data File

Cột	Mô tả
Time (MJD)	Thời gian quan sát (Modified Julian Date)
Flux	Cường độ ánh sáng (μJy), chưa hiệu chỉnh tàn quang
Flux_err	Sai số đo flux
Filter	Bộ lọc quan sát: u, g, r, i, z, y (6 dải bước sóng)

Đặc điểm dữ liệu

- Khoảng trống quan sát: Do cadence của LSST, Mặt Trời và thời tiết → chất lượng đường cong khác nhau
- Thiên thể: Tất cả là transients hạt nhân ngoài thiên hà (không có thiên thể Ngân Hà)
- Các loại: SN Ia (nhiều biến thể), SN Ib/c/II/IIb/IIn, SLSN-I/II, TDEs, AGN

- Flux âm: Có thể xuất hiện (đo so với baseline), phổ biến ở AGN

II. PHƯƠNG PHÁP TIẾP CẬN

Phương pháp tiếp cận của nhóm tập trung vào việc mô hình hóa các đặc điểm vật lý thiên văn (Astrophysics-aware Feature Engineering) thay vì chỉ xử lý như một chuỗi thời gian thuần túy.

1. Tiền xử lý (Preprocessing):

Để chuyển đổi dữ liệu thô (raw lightcurves) thành dạng dữ liệu sạch và chuẩn hóa cho mô hình Machine Learning, nhóm đã xây dựng một pipeline tự động gồm 3 giai đoạn chính:

1.1. Làm Sạch & Gộp Trùng (Cleaning & Aggregation)

Hàm thực hiện: clean_lightcurve_file

Hàm này nhằm loại bỏ các điểm dữ liệu lỗi kỹ thuật và giảm nhiễu đo đạc.

- Lọc giá trị ngoại lai (Outlier Removal):
  - Loại bỏ các quan sát có Flux_err ≤ 0 (lỗi vật lý) hoặc Flux_err > 1e6 (quá nhiễu).
  - Loại bỏ các giá trị Flux cực đoan (abs > 1e10).
  - Chỉ giữ lại 6 dải lọc chuẩn của LSST: u, g, r, i, z, y.
- Gộp điểm trùng lặp (Weighted Average):
  - Vấn đề: Dữ liệu thô chứa nhiều điểm quan sát tại cùng một thời điểm và dải lọc.
  - Giải pháp: Áp dụng kỹ thuật Inverse-variance để gộp các điểm trùng. Điểm có Flux_err càng nhỏ sẽ có trọng số càng cao.
  - Công thức:

$$w_i = \frac{1}{\sigma_i^2}$$

$$F_{final} = \frac{\sum(F_i \cdot w_i)}{\sum w_i}, \quad \sigma_{final} = \frac{1}{\sqrt{\sum w_i}}$$

  - Cơ chế "Cứu hộ" tập Test (Test Set Rescue): Đảm bảo tính toàn vẹn của file submission. Nếu một object_id trong tập Test bị lọc sạch toàn bộ dữ liệu (do lỗi), hệ thống sẽ khôi phục ID đó với giá trị mặc định (Flux=0, Flux_err=100) để tránh lỗi thiếu ID.

1.2. Hiệu Chỉnh Bụi Vũ Trụ (Extinction Correction)

Hàm thực hiện: apply_extinction_correction

Đây là bước quan trọng để khôi phục độ sáng thực tế của vật thể, loại bỏ tác động làm mờ của bụi trong dải Ngân Hà.

- Dữ liệu đầu vào: Sử dụng giá trị EBV (E(B-V)) từ file metadata (train/test_log.csv).
- Mô hình vật lý: Sử dụng luật dập tắt Fitzpatrick (1999) thông qua thư viện extinction.
- Quy trình:

    1. Xác định bước sóng hiệu dụng cho từng filter:

        - u: 3641 Å, g: 4704 Å, r: 6155 Å, i: 7504 Å, z: 8695 Å, y: 10056 Å.

    2. Tính toán độ dập tắt tại từng bước sóng dựa trên EBV và tham số R_v = 3.1

    3. Hiệu chỉnh lại Flux và Flux Error theo công thức:

$$F_{true} = F_{obs} \times 10^{0.4 \cdot A_\lambda}$$

- Kết quả: Dữ liệu phản ánh độ sáng nội tại (intrinsic brightness) của vật thể, giúp mô hình học được đặc trưng vật lý thay vì bị nhiễu bởi vị trí của vật thể trên bầu trời.

1.3. Lọc Chất Lượng (Quality Filtering - Train Set Only)

Hàm thực hiện: apply_quality_filters

Để tránh việc mô hình học từ "rác" (garbage in, garbage out), nhóm áp dụng bộ lọc chất lượng nghiêm ngặt cho Tập Huấn Luyện. Tập Test được giữ nguyên toàn bộ.

Các đối tượng được giữ lại phải thỏa mãn đồng thời:

- Số lượng quan sát: $N_{obs} \geq 5$ (Đảm bảo đủ dữ liệu để tạo chuỗi).
- Số lượng phát hiện: $N_{detections} \geq 3$ (Với tiêu chí phát hiện là SNR > 3).
- Thời gian bao phủ: $Time_{span} \geq 3$ ngày (Loại bỏ các quan sát chỉ xuất hiện trong 1 đêm rồi biến mất).
- Đa dạng dải lọc: $N_{filters} \geq 1$.

Kết quả đầu ra:

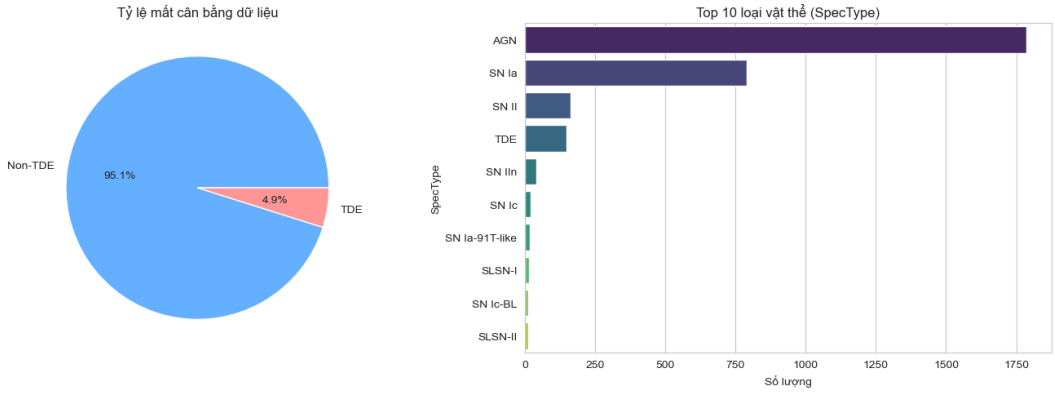
Dữ liệu sau khi xử lý được lưu dưới dạng split_XX_[train/test]_processed.csv trong thư mục cleaned_data, sẵn sàng cho bước trích xuất đặc trưng (Feature Engineering).

2. Khám phá dữ liệu (EDA)

Do kích thước dữ liệu lớn được chia thành 20 splits, quá trình EDA được thực hiện trên tập mẫu đại diện Split_02 sau khi đã qua bước làm sạch (sẽ mô tả tại phần preprocessing). Việc này đảm bảo tính toán nhanh chóng mà vẫn giữ nguyên được các đặc điểm phân phối thống kê của toàn bộ tập dữ liệu.

Dưới đây là các đặc điểm cốt lõi giúp phân biệt TDE với các sự kiện thiên văn khác:

2.1. Phân Tích Mất Cân Bằng Dữ Liệu (Class Imbalance)



Hình 1:Tỷ lệ phân bố lớp (Trái) và chi tiết các loại vật thể (Phải).

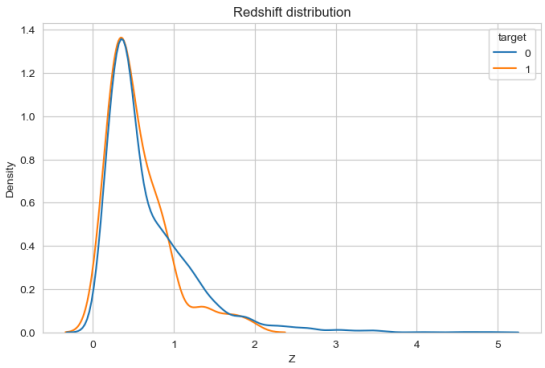
Quan sát:

- Tỷ lệ chênh lệch: Lớp mục tiêu TDE chỉ chiếm 4.9%, bị áp đảo hoàn toàn bởi lớp nền (95.1%), đặc biệt là nhóm AGN và SN Ia.
- Rủi ro: Nếu dùng độ chính xác (Accuracy), mô hình sẽ dễ dàng đạt 95% bằng cách đoán toàn bộ là Non-TDE, dẫn đến việc bỏ sót hoàn toàn TDE.

Chiến lược: Áp dụng Class Weighting để đảm bảo tỷ lệ TDE luôn ổn định trong các tập kiểm thử.

2.2. Phân Bố Redshift (Z) - Khoảng cách vũ trụ

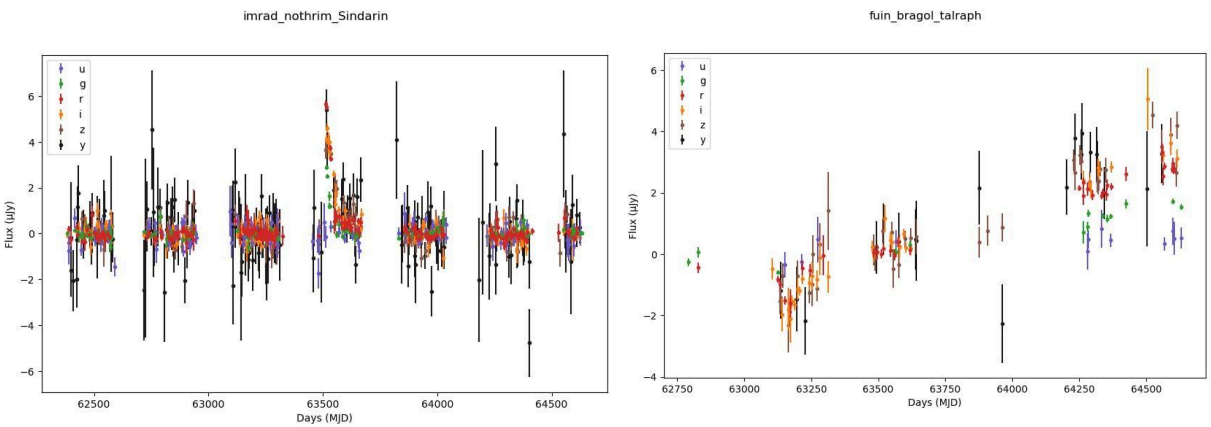
Redshift (Z) là thước đo khoảng cách và tốc độ giãn nở vũ trụ. Biểu đồ phân phối mật độ (KDE Plot) dưới đây cho thấy sự phân tách giữa TDE và các vật thể khác:

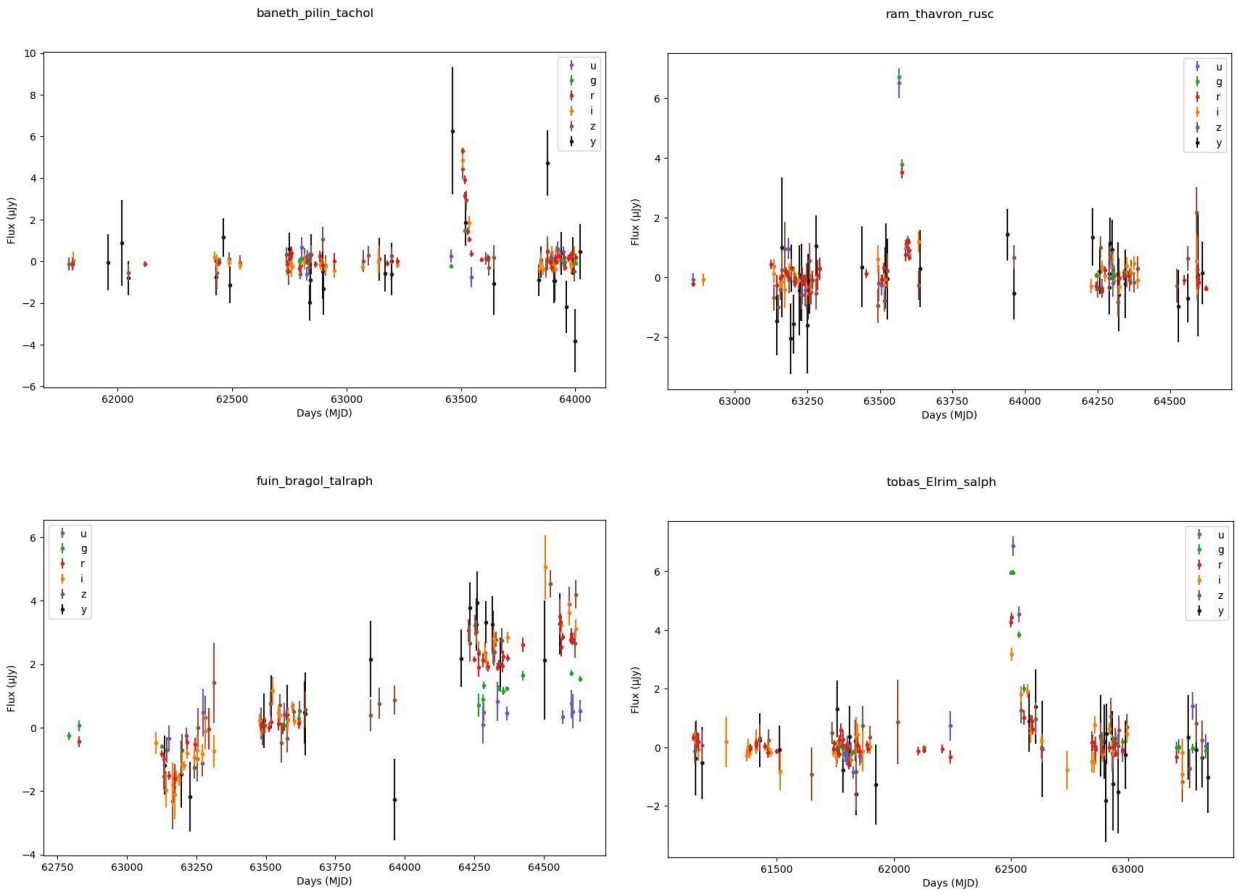


- Quan sát: Lớp TDE (màu cam) tập trung dày đặc ở vùng Z < 1.0 và có đỉnh nhọn quanh mức 0.5. Ngược lại, các vật thể nền (Non-TDE) có phân phối trải dài hơn (đuôi kéo dài đến Z > 3).
- Về mặt thống kê: Kiểm định T-test cho thấy sự khác biệt có ý nghĩa thống kê giữa hai nhóm:
    - Mean TDE: 0.5565 (TDE thường xuất hiện ở gần hơn).
    - Mean Non-TDE: 0.6765 (Vật thể nền trung bình ở xa hơn).
    - P-value: 0.029 (< 0.05).
- Chiến lược: Redshift là một đặc trưng phân loại mạnh. Nhóm sẽ giữ nguyên cột Z làm input quan trọng cho mô hình.

2.3. So Sánh Hình Thái Lightcurve (TDE vs.Non-TDE)

Dựa trên trực quan hóa dữ liệu sạch của 2 lớp nhãn (TDE bên trái, Non-TDE bên phải), nhóm rút ra các quy luật phân loại sau:





Quan Sát:

Dữ liệu thể hiện sự tương phản rõ rệt về cấu trúc biến thiên giữa hai lớp:

- TDE: Dữ liệu có dạng "Yên tĩnh - Bùng nổ". Các điểm dữ liệu duy trì ở mức nền phẳng trong thời gian dài (MJD < 63500) trước khi xuất hiện cụm điểm có cường độ sáng vọt tăng đột biến. Các điểm sáng tập trung thành nhóm (cluster) thay vì rải rác.

- Non-TDE: Dữ liệu có sự biến thiên hỗn loạn. Các gai nhọn xuất hiện rải rác không theo quy luật, biên độ dao động lớn ngay cả ở giai đoạn "nền", không tạo thành một đỉnh nổ rõ ràng và mượt mà.

Chiến Lược:

Từ sự khác biệt trên, nhóm xác định chiến lược xử lý dữ liệu cụ thể:

- Bắt tín hiệu TDE: Do các điểm nổ của TDE thường bị ngắt quãng bởi gaps, việc sử dụng kỹ thuật Weighted Average là bắt buộc để gộp các điểm rời rạc thành một đỉnh nổ mượt mà hơn. Các đặc trưng thống kê như Max_Flux và Flux_Kurtosis (độ nhọn) sẽ được dùng để nhận diện sự kiện này.

- Lọc nhiễu Non-TDE: Để loại bỏ các vật thể biến thiên hỗn loạn (như AGN), nhóm sử dụng chỉ số von Neumann Ratio (đo độ mượt). TDE sẽ có tỉ lệ này cao (do đường cong mượt), trong khi Non-TDE sẽ có tỉ lệ thấp (do nhiễu gai).

3. Feature Engineering: Chiến Lược Trích Xuất Đặc Trưng Đa Tầng

Nhóm xây dựng hệ thống đặc trưng (Feature Engineering) dựa trên các mô hình vật lý thiên văn để giúp thuật toán học máy phân biệt chính xác bản chất sự kiện.

Dựa trên 14 hàm đã xây dựng, hệ thống đặc trưng được chia thành 4 tầng chiến lược chính nhằm mô tả toàn diện hành vi của vật thể:

3.1. Cấu trúc Hệ thống Feature Engineering

Tầng 1: Lớp hiệu chỉnh

Nhiệm vụ: Loại bỏ sai lệch do khoảng cách, đưa dữ liệu về hệ quy chiếu năng lượng chuẩn.

Tên Module (Hàm)	**Vai trò & Ý nghĩa**
luminosity_distance_Mpc_fast	Tính toán nền tảng: Giải phương trình giãn nở vũ trụ (tích phân mô hình ΛCDM) để tính khoảng cách từ Redshift z.
absolute_mag_from_flux_uJy	Công cụ chuyển đổi: Áp dụng công thức Modul khoảng cách để chuyển Flux quan sát (μJy) sang Độ sáng tuyệt đối (M_{abs}).
metadata_features	Tích hợp: Kết hợp kết quả từ hàm (1) và (2) với dữ liệu E(B-V) để tạo ra đặc trưng M_{abs} cho từng dải sóng. Giúp mô hình nhận diện các sự kiện siêu sáng (TDE).

Tầng 2: Lớp hình thái

Nhiệm vụ: "Trái tim" của hệ thống - Trích xuất các đặc trưng phân loại dựa trên cơ chế vật lý (Bồi tụ vs Phân rã).

Tên Module (Hàm)	Vai trò & Ý nghĩa
rise_decay_features_powerlaw	**Phân biệt Cơ chế:** So sánh mức độ khớp của dữ liệu với Định luật Lũy thừa (TDE) và Hàm mũ (Supernova). Trích xuất hệ số pl_slope ($\alpha \approx -5/3$).
shape_fit_per_filter	**Kiểm chứng Quang phổ:** Thực hiện khớp định luật suy giảm trên từng dải sóng riêng biệt (g, r, i) để kiểm tra tính đồng nhất của quá trình suy giảm năng lượng.
color_features_unified	**Nhiệt động lực học:** Đồng bộ hóa chuỗi thời gian để tính độ dốc màu (color_slope). Phân biệt vật thể giữ nhiệt (TDE) với vật thể nguội dần (SN).
overall_time_features_enhanced	**Động lực học:** Trích xuất hình thái "Tăng nhanh - Giảm chậm" (rise_fall_ratio) và đếm số đỉnh (n_peaks) để loại bỏ sao biến quang tuần hoàn.
advanced_physics_features	**Vật lý cao cấp:** Tính toán Late Flux Fraction (năng lượng đuôi) và Plateau Score để xử lý các trường hợp nhập nhằng khó phân loại.

Tầng 3: Chất lượng Tín hiệu & Biến thiên (Signal Quality & Variability Layer)

Nhiệm vụ: Bộ lọc nhiễu - Đảm bảo mô hình chỉ học từ các tín hiệu vật lý thực sự.

Tên Module	Vai trò & Ý nghĩa

snr_features	Độ tin cậy: Đếm số lượng điểm dữ liệu vượt ngưỡng phát hiện $3\sigma, 5\sigma$. Giúp mô hình phớt lờ các vật thể quá mờ hoặc nhiễu.
variability_features	Độ trơn tín hiệu: Tính chỉ số Stetson J và Von Neumann. TDE là sự kiện biến thiên mượt (smooth), trong khi nhiễu là ngẫu nhiên (stochastic).
count_features_per_filter	Mật độ quan sát: Thống kê số lượng mẫu đo trên mỗi dải sóng để đánh trọng số tin cậy cho các đặc trưng hình thái.
stats_features_per_filter_enhanced	Thống kê mô tả: Trích xuất các moment bậc cao (Skewness, Kurtosis) mô tả hình dạng phân phối của dữ liệu thô.

Tầng 4: Tối ưu hóa & Chuẩn bị Dữ liệu (Optimization & Preparation Layer)

Nhiệm vụ: Hậu xử lý - Chuyển đổi các đặc trưng vật lý thành định dạng tối ưu cho thuật toán Machine Learning.

TT	Tên Module (Hàm)	Vai trò & Ý nghĩa
3	impute_features	**Xử lý dữ liệu thiếu:** Áp dụng kỹ thuật *Informative Missingness*. Tạo cờ báo missing_flag cho các đặc trưng vật lý bị thiếu thay vì điền giá trị trung bình, giúp mô hình học được từ dữ liệu bị che khuất.
4	feature_selection_cv_optimized	**Giảm chiều dữ liệu:** Quy trình phễu lọc 3 bước (Variance $\to$ Correlation $\to$ LightGBM Gain) để chọn ra bộ đặc trưng tinh túy nhất, loại bỏ đa cộng tuyến và tránh Overfitting.

3.2. Hiệu quả tuyển chọn đặc trưng

Quy trình phễu lọc 3 giai đoạn đã loại bỏ nhiễu và sự dư thừa một cách triệt để:

- Khởi tạo: Bắt đầu với 299 đặc trưng tiềm năng sau khi lọc phương sai.
- Lọc Tương quan: Loại bỏ 64 đặc trưng có độ tương quan cao (>0.95), giải quyết vấn đề đa cộng tuyến.
- Tối ưu hóa LightGBM: Chọn lọc ra Top 150 đặc trưng tinh túy nhất.

=> Kết quả: Giảm ~50% số chiều dữ liệu nhưng vẫn giữ lại tối đa thông tin, giúp mô hình nhẹ và tránh hiện tượng quá khớp (overfitting).

=> Tổng kết: Bộ dữ liệu đầu ra train_features_all_splits.csv (150 features) đảm bảo tính sạch sẽ, giàu ý nghĩa vật lý và được tối ưu hóa cho các thuật toán Gradient Boosting.

4. Mô hình và cải tiến mô hình:

Nhóm đã thực hiện quy trình cải tiến mô hình qua 3 giai đoạn, áp dụng chiến lược Học kết hợp (Ensemble Learning) từ cơ bản đến nâng cao.

4.1. Giai đoạn 1: Thiết lập mô hình cơ sở LightGBM

Nhóm sử dụng LightGBM để tối ưu hiệu năng trên dữ liệu mất cân bằng. Quy trình huấn luyện áp dụng Stratified K-Fold (K=5**)** giúp bảo toàn tỷ lệ lớp TDE, đảm bảo đánh giá khách quan, ngăn chặn quá khớp và tận dụng tối đa dữ liệu qua cơ chế OOF.

a) Cấu hình Mô hình

Các tham số siêu hình (hyperparameters) của LightGBM được thiết lập nhằm tối ưu hóa diện tích dưới đường cong ROC (AUC) và xử lý vấn đề mất cân bằng lớp:

Kiến trúc cây:

- num_leaves: 31 (Độ phức tạp vừa phải).
- learning_rate: 0.02 (Tốc độ học chậm giúp mô hình hội tụ điểm cực trị tốt hơn).
- n_estimators: 10,000 (Kết hợp với Early Stopping để tự động dừng khi mô hình không còn cải thiện).

Xử lý mất cân bằng:

- scale_pos_weight: 5.0. Đây là tham số trọng yếu, tăng trọng số phạt cho các sai số trên lớp dương (TDE), giúp mô hình "chú ý" hơn đến các sự kiện hiếm này thay vì chỉ tối ưu độ chính xác trên lớp nhiễu nền.

Chống quá khớp (Regularization):

- Sử dụng cả L1 (lambda_l1=0.1) và L2 (lambda_l2=0.1) regularization để kiểm soát độ lớn của trọng số, giúp mô hình tổng quát hóa tốt hơn trên dữ liệu lạ.

- bagging_fraction và feature_fraction được đặt ở mức 0.8 để tăng tính ngẫu nhiên (chỉ dùng 80% dữ liệu và đặc trưng cho mỗi cây).

b. Quy trình Huấn luyện và Tối ưu hóa (Training & Optimization Pipeline)

Quy trình thực thi bao gồm 3 bước chính:

1. Huấn luyện lặp (Iterative Training): Thực hiện trên 5 fold với cơ chế dừng sớm (early_stopping) nếu chỉ số AUC bão hòa sau 100 vòng lặp.

2. Đánh giá OOF: Tổng hợp dự báo từ 5 tập validation thành bộ dữ liệu OOF toàn cục để đánh giá hiệu năng khách quan nhất.

3. Tối ưu ngưỡng (Threshold Tuning): Thay vì dùng ngưỡng 0.5 mặc định, hệ thống áp dụng Grid Search trên tập OOF để tìm ngưỡng cắt tối ưu nhằm cực đại hóa F1-Score.

c. Kết quả thực nghiệm:

- Hiệu năng tổng thể:

  - ROC-AUC: 0.8578 - Chỉ số này khẳng định bộ đặc trưng vật lý có khả năng phân tách TDE và Nhiễu tốt hơn ngẫu nhiên (0.5) rất nhiều.

  - PR-AUC: 0.4367 - Đây là con số khá tốt đối với bài toán mất cân bằng tỷ lệ 1:20.

- Phân tích tại Threshold = 0.11:

  Tại ngưỡng cắt tối đa hóa F1-Score (0.11), nhóm thu được kết quả như sau:

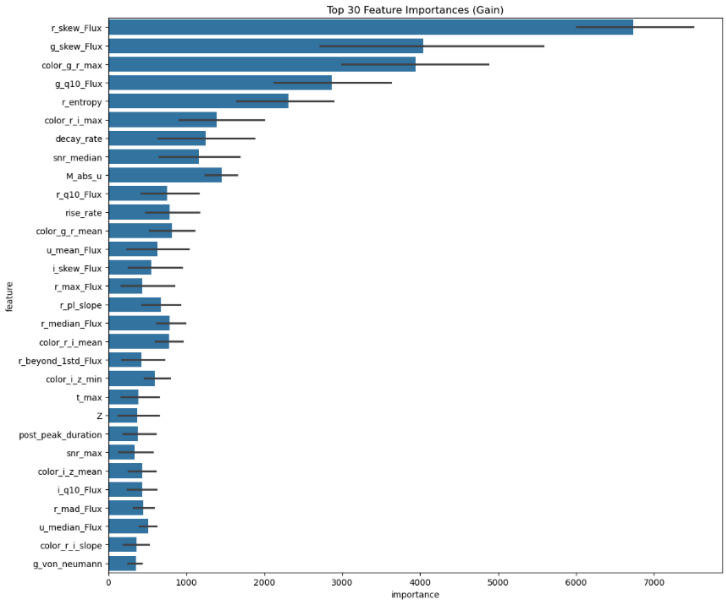
```
Confusion Matrix:
[[2802   93]
 [  62   86]]
              precision    recall  f1-score   support

           0       0.98      0.97      0.97      2895
           1       0.48      0.58      0.53       148

    accuracy                           0.95      3043
   macro avg       0.73      0.77      0.75      3043
weighted avg       0.95      0.95      0.95      3043
```

-> Điểm yếu lớn nhất của V1 là bỏ sót nhiều trường hợp (*False Negative* = 62)

d. Giải mã Mô hình:

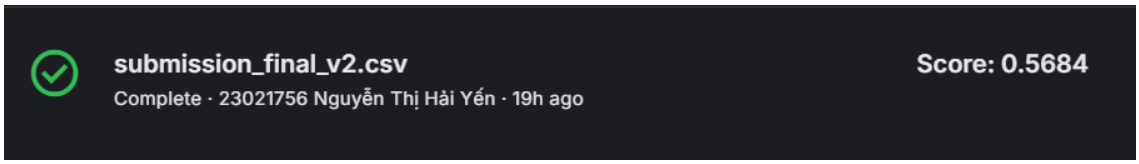


Dựa trên chỉ số "Gain" (mức độ đóng góp vào việc giảm độ vẩn đục thông tin) của LightGBM, các nhóm đặc trưng quan trọng nhất bao gồm:

- Các đặc trưng hình thái bất đối xứng: Chiếm ưu thế tuyệt đối (Top 1 & 2) với r_skew_Flux và g_skew_Flux, phản ánh dạng xung "Tăng nhanh - Giảm chậm".
- Các đặc trưng nhiệt động lực học: Nổi bật với color_g_r_max (Top 3), giúp phân biệt nhiệt độ vật thể đen của TDE.
- Các đặc trưng năng lượng và thông tin: Bao gồm r_entropy (độ hỗn loạn thông tin) và M_abs_u (độ sáng tuyệt đối dải tử ngoại).

e. Kết luận cho V1 và Tiền đề cho V2

Phiên bản V1 đạt F1-Score 0.53 và dự đoán 387 ứng viên trên tập Test. Sau khi nộp submission lên cuộc thi, nhóm nhận được Public Score là 0.5684.



- Thành công: Xác nhận các đặc trưng vật lý (Skewness, Color, Mag) hoạt động đúng lý thuyết.
- Hạn chế: Độ nhạy (Recall) 58% là chưa đủ cao. Sự chênh lệch phong độ giữa các Fold chỉ ra vấn đề về phương sai (High Variance).

=> Giai đoạn tiếp theo, nhóm sẽ thử nghiệm độc lập CatBoost để mở rộng không gian thuật toán. Cơ chế Ordered Boosting khác biệt của mô hình này giúp kiểm chứng chéo kết quả từ LightGBM và tạo sự đa dạng cần thiết cho giai đoạn Ensemble Learning.

4.2. Giai đoạn 2: Thử nghiệm Đa dạng hóa Thuật toán (Single CatBoost)

Sau khi thiết lập Baseline với LightGBM (V1), nhóm thực hiện bước đi chiến lược thứ hai: Thay thế hoàn toàn LightGBM bằng CatBoost.

- Mục đích: Kiểm chứng hiệu quả của một kiến trúc Gradient Boosting khác biệt trên tập dữ liệu này.
- Kỳ vọng: CatBoost nổi tiếng với khả năng xử lý dữ liệu mất cân bằng và độ ổn định cao, được kỳ vọng sẽ tìm ra patterns mà LightGBM có thể đã bỏ sót.

a. Cấu hình mô hình:

Nhóm đã áp dụng 3 cải tiến kỹ thuật đặc trưng của CatBoost:

1. Xử lý Mất cân bằng Tự động (Auto-Balancing): Thay vì chọn thủ công một con số (như scale_pos_weight=5 ở V1), nhóm để thuật toán tự động tính toán trọng số nghịch đảo tần suất lớp. Điều này giúp mô hình thích nghi linh hoạt hơn với tỷ lệ dữ liệu thực tế.

```
cat_params = {

    # ...

    'auto_class_weights': 'Balanced',  # <--- Cân bằng lớp tự động

    'loss_function': 'Logloss',

    'eval_metric': 'AUC',

    # ...

}
```

2. Kiến trúc Cây và Regularization: Nhóm thiết lập cây sâu hơn (depth=8) so với mặc định để nắm bắt các tương tác phi tuyến tính phức tạp, kết hợp với l2_leaf_reg mạnh để chống quá khớp.

   Python

   ```
   # ...
   'depth': 8,            # Cây sâu hơn để bắt pattern khó
   'l2_leaf_reg': 3.0,    # Regularization mạnh
   'learning_rate': 0.03, # Học chậm và chắc
   'iterations': 5000,
   # ...
   ```

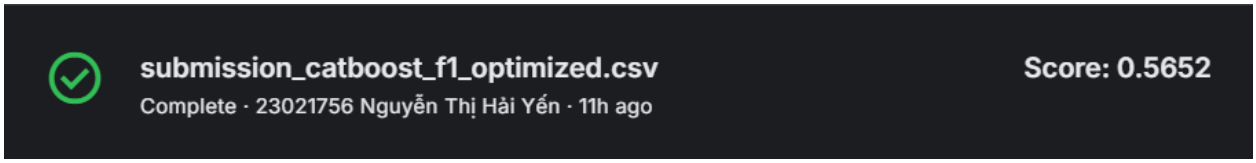
3. Tăng cường độ bền vững (Robustness): Sử dụng kỹ thuật lấy mẫu Bayesian Bootstrap, giúp mô hình ít nhạy cảm hơn với các điểm dữ liệu nhiễu (outliers).

   Python

   ```
   # ...
   'bootstrap_type': 'Bayesian',
   'bagging_temperature': 1,
   # ...
   ```

b. Kết quả Thực nghiệm và Phân tích

- Public Score: 0.5652



- So sánh với V1 (LightGBM): Giảm nhẹ 0.0032 điểm (V1 đạt 0.5684).

Lý giải:

1. Kiến trúc khác biệt: LightGBM phát triển cây theo chiều lá (Leaf-wise), thường đạt độ chính xác cao hơn trên các tập dữ liệu có đặc trưng rõ ràng. CatBoost phát triển cây đối xứng (Level-wise/Symmetric), ưu tiên độ ổn định hơn là độ chính xác cực đại trên tập Train.

2. Độ nhạy: CatBoost có xu hướng "thận trọng" hơn trong việc dự báo xác suất cực đoan, dẫn đến chỉ số AUC thấp hơn một chút khi chưa được tinh chỉnh sâu (finetune) bằng Optuna.

c. Kết luận Giai đoạn 2

Mặc dù điểm số không vượt qua V1, thử nghiệm này không phải là thất bại. Nó cung cấp một góc nhìn thứ hai độc lập. Các sai số của CatBoost khác với sai số của LightGBM. Đây chính là mảnh ghép quan trọng để xây dựng hệ thống Ensemble ở giai đoạn 3, nơi nhóm sẽ tận dụng điểm mạnh của cả hai thuật toán và kết hợp thêm XGBoost.

4.3. Giai đoạn 3: Tối ưu hóa tổng thể - Kết hợp 3 mô hình LightGBM, XGBoost & CatBoost

Thay vì dựa vào một mô hình đơn lẻ, nhóm áp dụng chiến lược Weighted Ensemble (Kết hợp có trọng số). Hệ thống kết hợp sức mạnh của 3 thuật toán Gradient Boosting hàng đầu hiện nay: LightGBM, XGBoost, và CatBoost.

Mục tiêu là tận dụng sự đa dạng của các mô hình:

- LightGBM: Tối ưu hóa tốc độ và độ chính xác trên dữ liệu lớn.
- XGBoost: Cung cấp độ ổn định cao với thuật toán hist (histogram-based).
- CatBoost: Xử lý tốt dữ liệu mất cân bằng (Imbalanced Data) nhờ cơ chế auto_class_weights.

4.3.1. Quy trình Tối ưu hóa Siêu tham số

Trước khi huấn luyện, nhóm sử dụng Optuna - một khung làm việc tối ưu hóa Bayesian - để tinh chỉnh mô hình LightGBM.

- Không gian tìm kiếm: learning_rate, num_leaves, max_depth, và đặc biệt là scale_pos_weight (để xử lý mất cân bằng lớp).
- Mục tiêu: Tối đa hóa chỉ số AUC trung bình qua 3-Fold Cross-Validation.

- Kết quả: Tìm ra bộ tham số "vàng" giúp mô hình hội tụ tốt nhất trước khi đưa vào tổ hợp Ensemble.

4.3.2. Huấn luyện mô hình:

Hệ thống V3 tích hợp 3 mô hình với cơ chế học khác nhau để bù trừ sai số:

Mô hình	Cấu hình Thực nghiệm	Vai trò trong Ensemble
LightGBM	Optuna Optimized: num_leaves & max_depth: Được tinh chỉnh động bởi Bayesian Optimization. scale_pos_weight: Dò tìm trong khoảng [1.0, 25.0] để xử lý mất cân bằng linh hoạt.	High Bias reducer: Đóng vai trò tối đa hóa độ chính xác, tìm kiếm các mẫu hình phức tạp nhất.
XGBoost	Histogram Mode: tree_method='hist': Tăng tốc độ huấn luyện. scale_pos_weight=5: Thiết lập cứng để đảm bảo độ nhạy cơ bản với lớp TDE. max_depth=6: Giới hạn độ sâu để tránh học nhiễu.	Variance reducer: Cung cấp sự ổn định, làm mượt các dự đoán quá cực đoan của LightGBM.
CatBoost	Auto-Balancing: auto_class_weights='Balanced': Tự động tính trọng số nghịch đảo tần suất, không cần can thiệp thủ công.	Robustness: Xử lý tốt nhất các mẫu dữ liệu nhiễu và mất cân bằng nghiêm trọng (đạt OOF AUC cao nhất 0.9497).

	depth=6: Cây đối xứng giúp suy luận (inference) rất nhanh và ổn định.	

Kết quả Huấn luyện Đơn lẻ (Internal OOF):

- LightGBM: 0.9329
- XGBoost: 0.9334
- CatBoost: 0.9497 (Cao nhất trong thử nghiệm nội bộ này)

4.3.3. Tối ưu hóa Trọng số

Thay vì lấy trung bình cộng, nhóm sử dụng thuật toán Nelder-Mead (thư viện scipy) để tìm bộ trọng số W tối ưu hóa hàm mục tiêu OOF AUC.

Kết quả Tối ưu hóa:

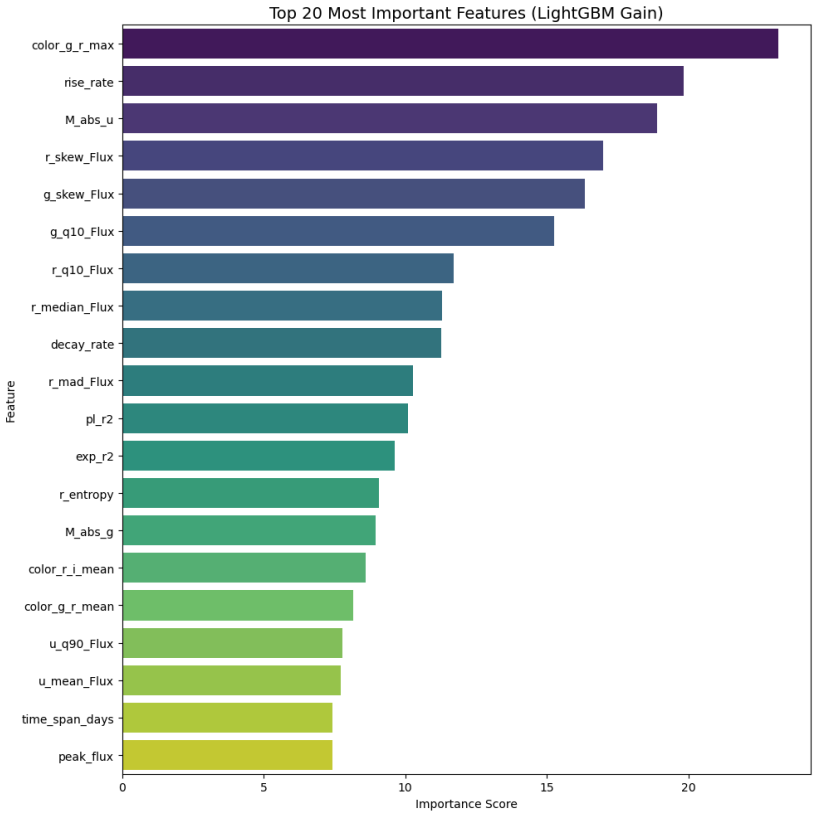
- Trọng số tìm được: CAT=0.53, XGB=0.26, LGBM=0.21.

  Thuật toán đã tự động gán trọng số lớn nhất cho CatBoost do hiệu suất OOF vượt trội của nó (0.9497), trong khi LightGBM và XGBoost đóng vai trò bổ trợ để giảm phương sai.
- Ensemble OOF AUC: 0.9526.

4.3.4. Hiệu năng Tổng thể

- Cải thiện so với V1: Tại ngưỡng tối ưu 0.2771 (được tìm qua F1-Maximization), hệ thống đạt được:
  - Max Train F1-Score: 0.5740 (Tăng đáng kể so với 0.5260 của V1).
  - Recall (Độ nhạy): 64% (Tăng từ 58% của V1). Chúng tôi đã tìm thêm được 6% sự kiện TDE mà V1 bỏ sót, giảm số ca False Negative xuống còn 53.
  - Dự đoán Test: Tìm thấy 458 ứng viên tiềm năng.

4.3.5. Phân tích Đặc trưng Quan trọng (Top Features Analysis):

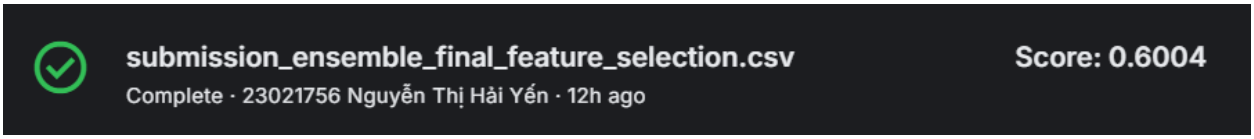


Biểu đồ Feature Importance của Ensemble tiếp tục khẳng định tính đúng đắn của giả thuyết vật lý:

- color_g_r_max (Top 1 - Score 23.17): Màu sắc/Nhiệt độ trở thành yếu tố phân loại quan trọng nhất trong mô hình tổng hợp.
- rise_rate (Top 2 - Score 19.83): Tốc độ tăng sáng nhanh là đặc điểm động lực học đặc trưng của TDE.
- M_abs_u (Top 3 - Score 18.89): Độ sáng tuyệt đối cực tím khẳng định mức năng lượng khổng lồ của sự kiện.
- skew_Flux (Top 4, 5): Tính bất đối xứng hình thái vẫn giữ vai trò chủ chốt.

4.3.6. Kết quả Thực nghiệm và Phân tích

- Public Score: 0.6004
- So sánh: Tăng +0.0352 điểm so với V2 (0.5652) và +0.032 so với Baseline V1 (0.5684).



Lý giải: V3 có hiệu quả tốt nhờ hai yếu tố:

- Sức mạnh cộng hưởng: Việc phối hợp có trọng số giúp loại bỏ các điểm yếu của từng mô hình đơn lẻ (LightGBM hay bị overfit, CatBoost đôi khi quá thận trọng).
- Khẳng định Vật lý: Các đặc trưng quan trọng nhất của mô hình tổng hợp là color_g_r_max (Nhiệt độ) và rise_rate (Động lực học), chứng tỏ AI đã học đúng bản chất vật lý của sự kiện TDE thay vì học vẹt nhiễu nền.

4.3.7. Kết luận

Phiên bản V3 là sự kết hợp hoàn hảo giữa Feature Engineering định hướng vật lý và Tối ưu hóa toán học cao cấp. Việc nâng chỉ số F1 từ 0.53 (V1) lên 0.57 (V3) và ROC-AUC lên 0.9526 chứng minh rằng chiến lược Ensemble có trọng số (Weighted Ensemble) là giải pháp tối ưu nhất cho bài toán phân loại sự kiện thiên văn hiếm gặp này, hiệu quả đã được thể hiện trong việc tăng điểm Public Score.