

MALLORN Astronomical Classification Challenge

Báo cáo phân tích &
xử lý bài toán

i



Our team

Nhóm 2:

Nguyễn Văn Hòa

Nguyễn Đăng Đạo

Lê Minh Đức



Nội dung trình bày

1. Giới thiệu bài toán
2. Exploratory Data Analysis
3. Feature Engineering
4. Modeling
5. Kết quả

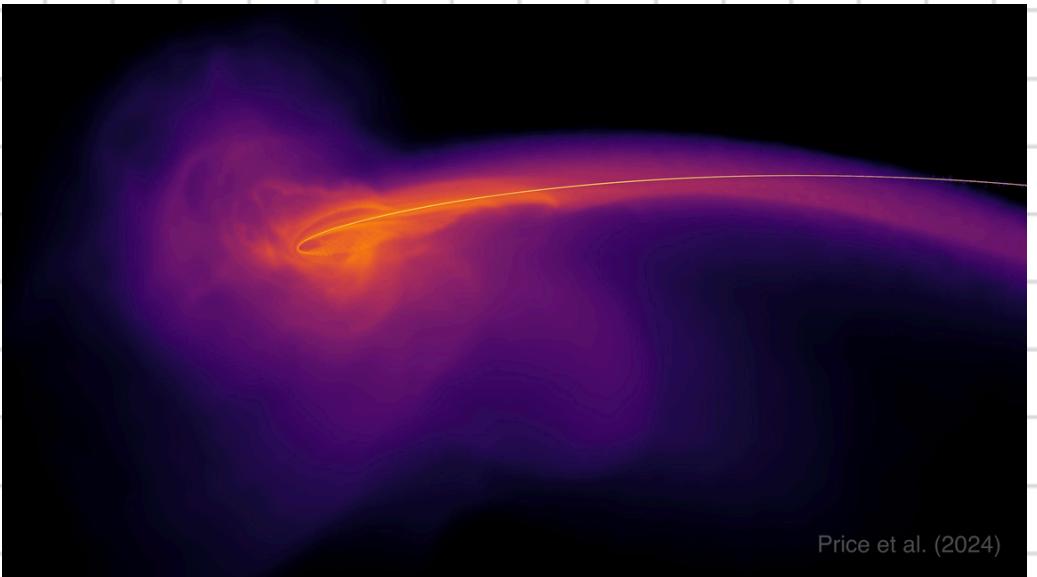


Giới thiệu bài toán

Bài toán: Phát hiện Tidal Disruption Events

(TDE) từ dữ liệu lightcurve:

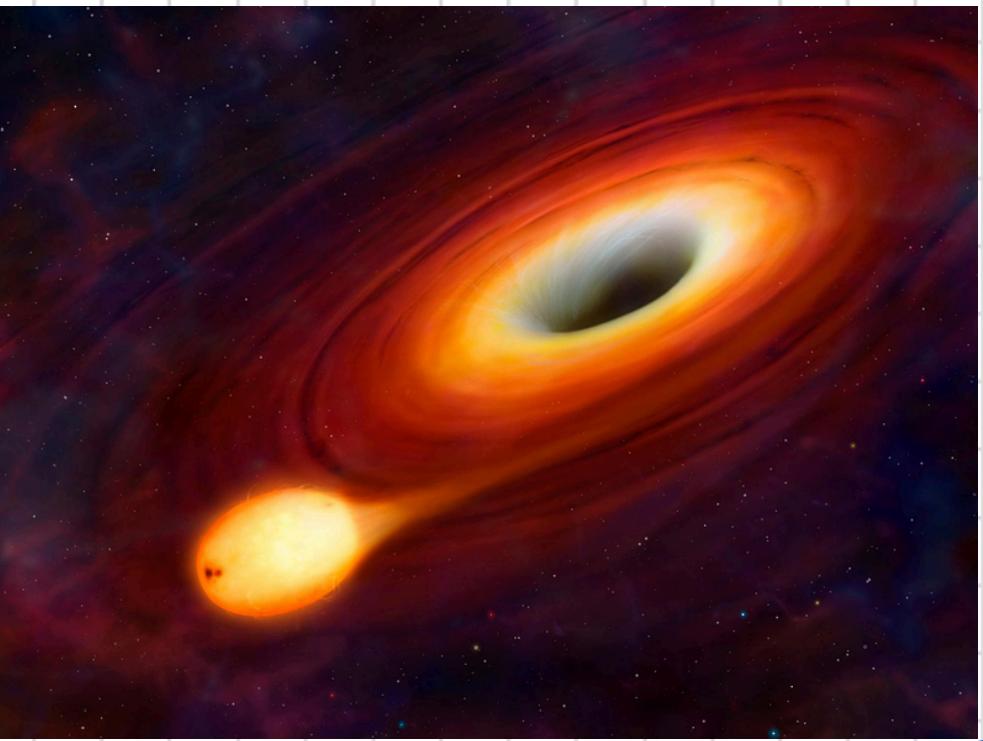
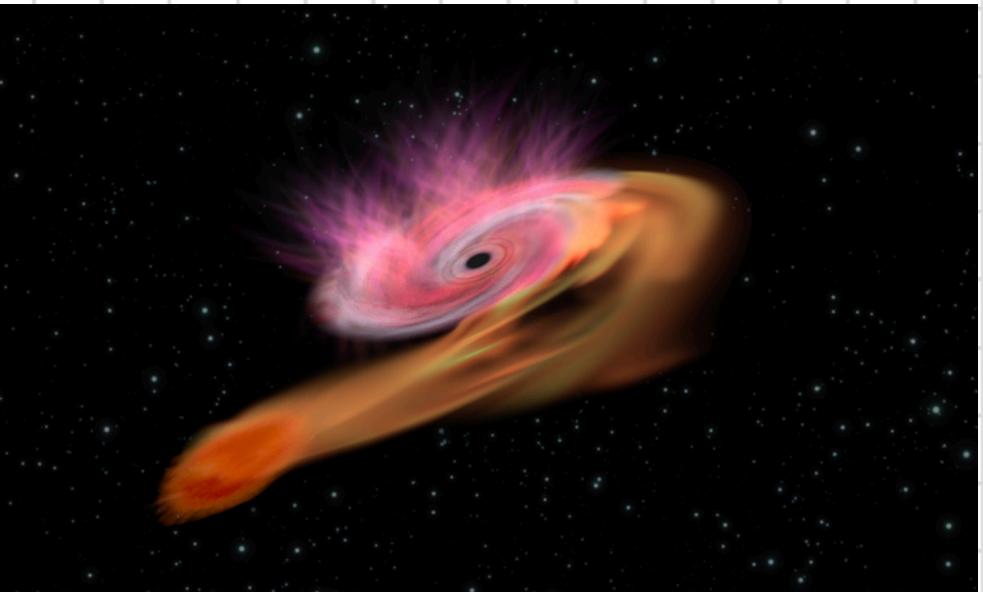
- **TDE là sự kiện hiếm khi một ngôi sao bị xé toạc bởi hố đen siêu nặng.**
- Trong các khảo sát lớn như LSST, việc phát hiện TDE cần được tự động hóa từ dữ liệu đa băng tần.



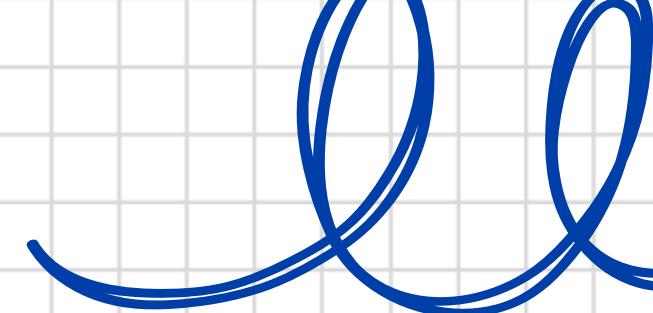
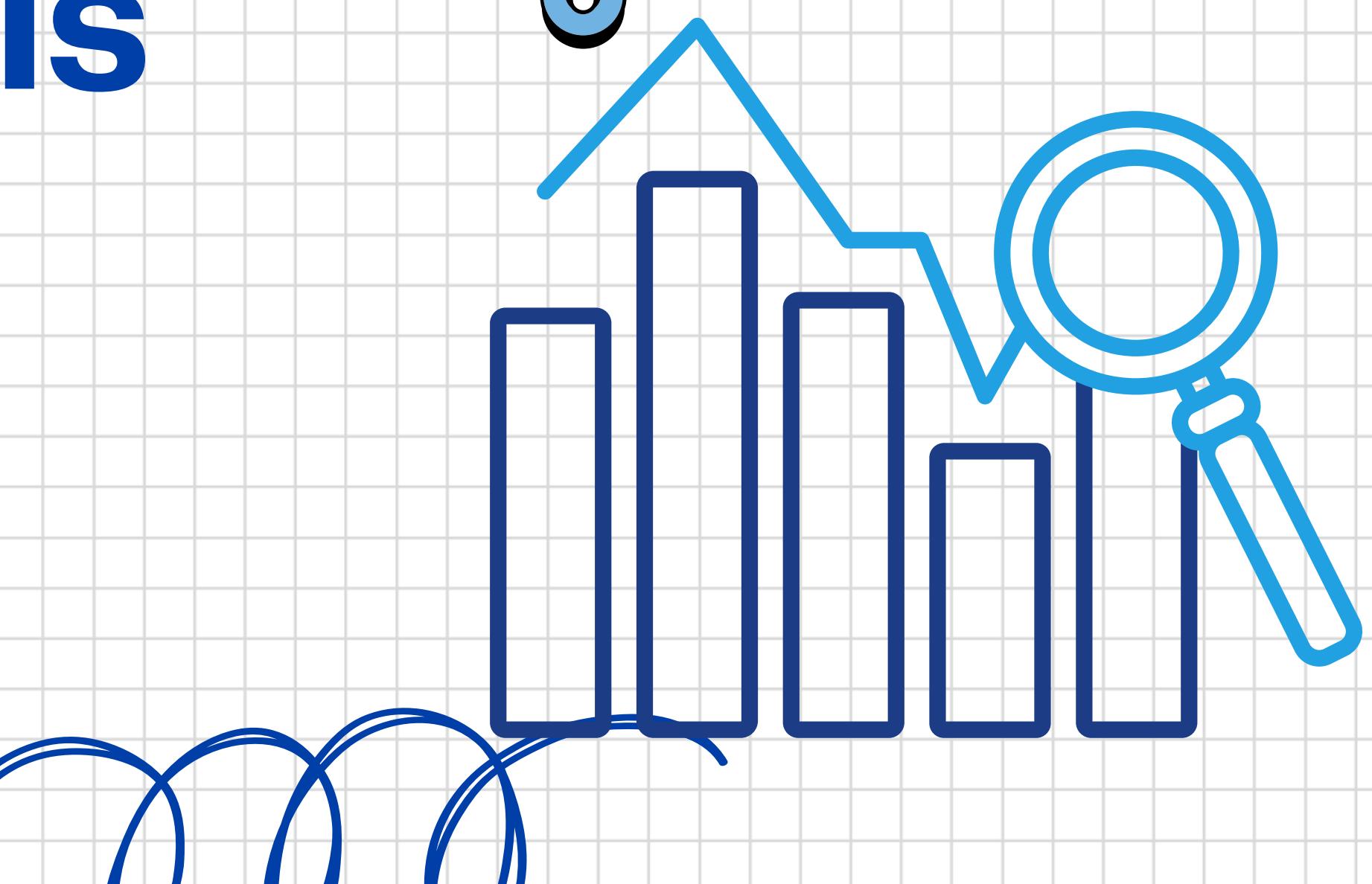
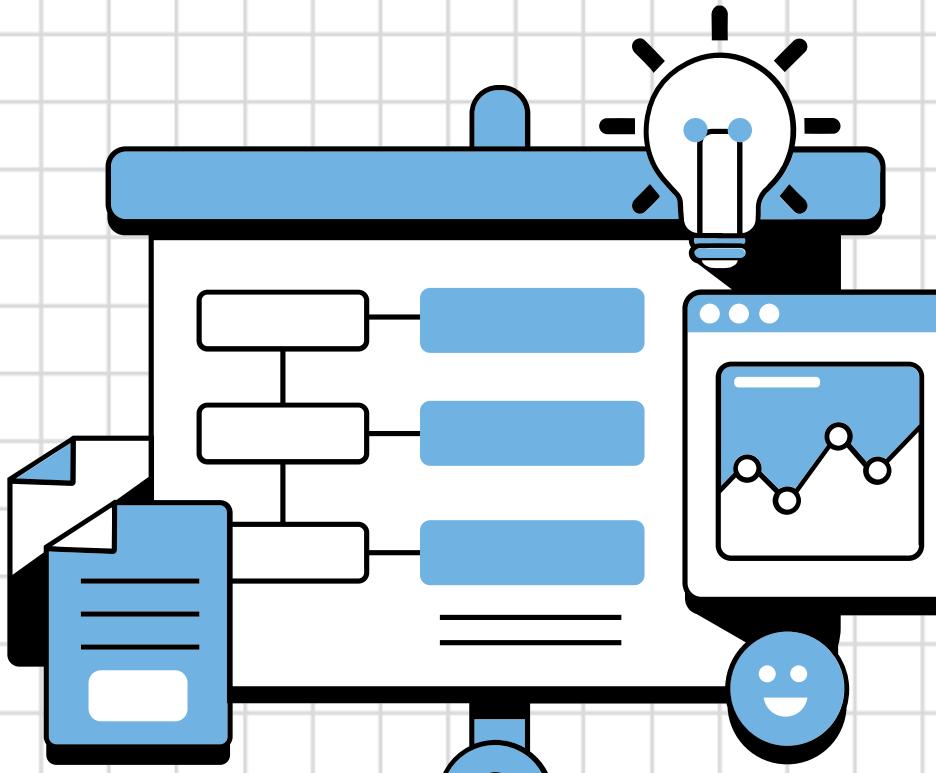
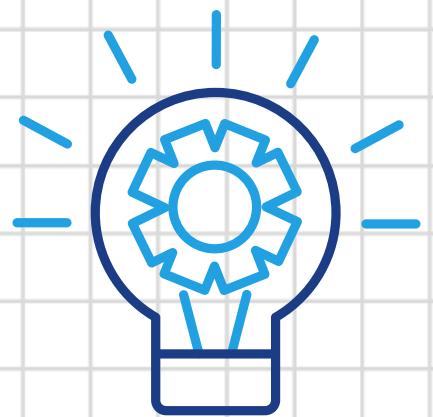
Giới thiệu bài toán

Đặc điểm, thách thức bài toán, metric chính:

- **Dữ liệu time-series**
- **Nhiều đo lớn**
- **Dữ liệu đa bằng tần, không đồng đều**
- **Mất cân bằng lớp nghiêm trọng**
- **Metric: F1-score**
- **Định hướng:**
 - **Phân tích EDA → Thiết kế features**
 - **Dùng mô hình tree-based**
 - **Ensemble để tăng độ ổn định**



Exploratory Data Analysis



EDA

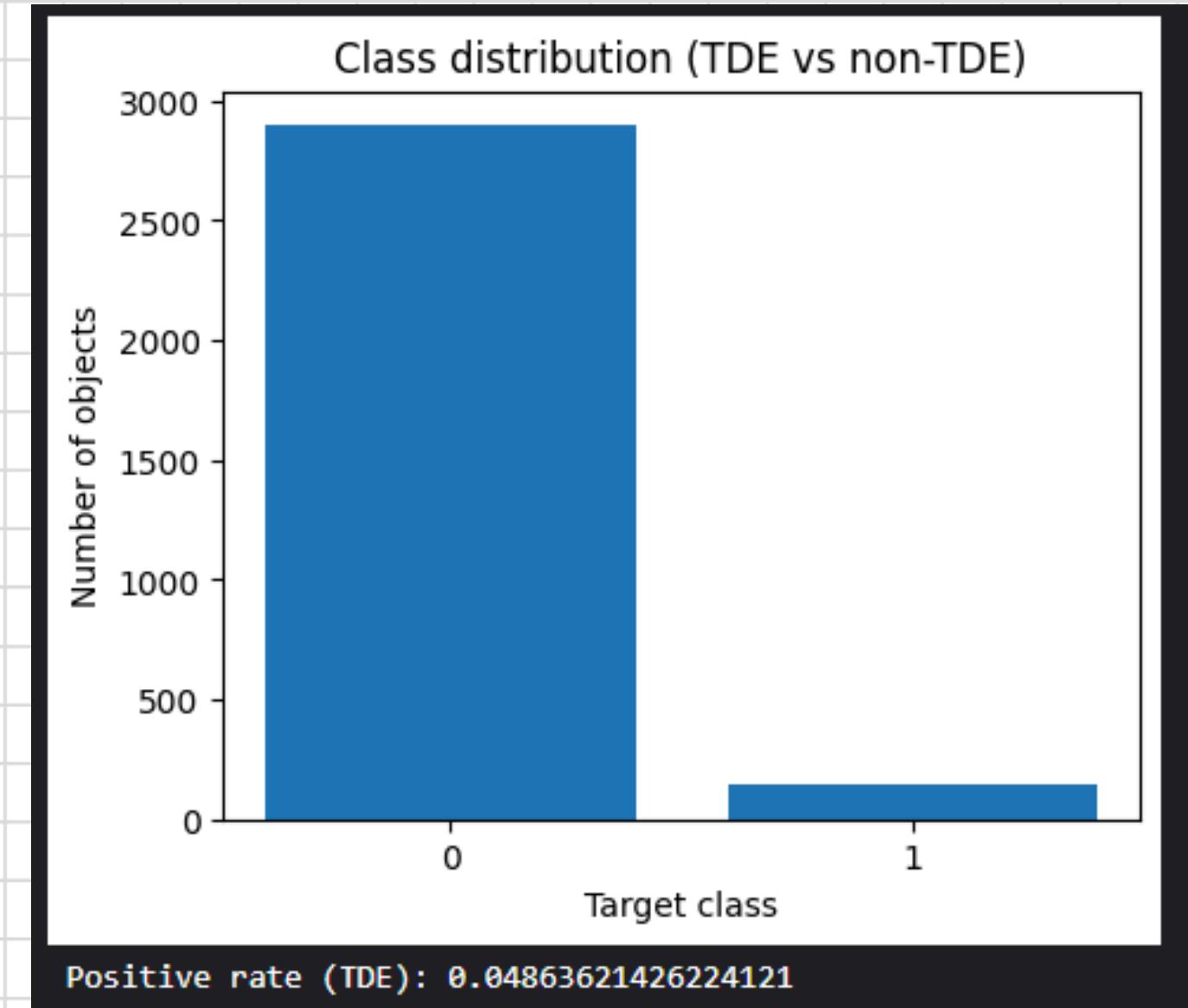
Phân bố nhãn

Nhận xét:

- Số lượng non-TDE >> TDE
- Tỷ lệ TDE ~ 4.86%
- Dataset rất mất cân bằng

→ Lựa chọn F1 score

→ Áp dụng scale_pos_weight và threshold tuning khi huấn luyện mô hình



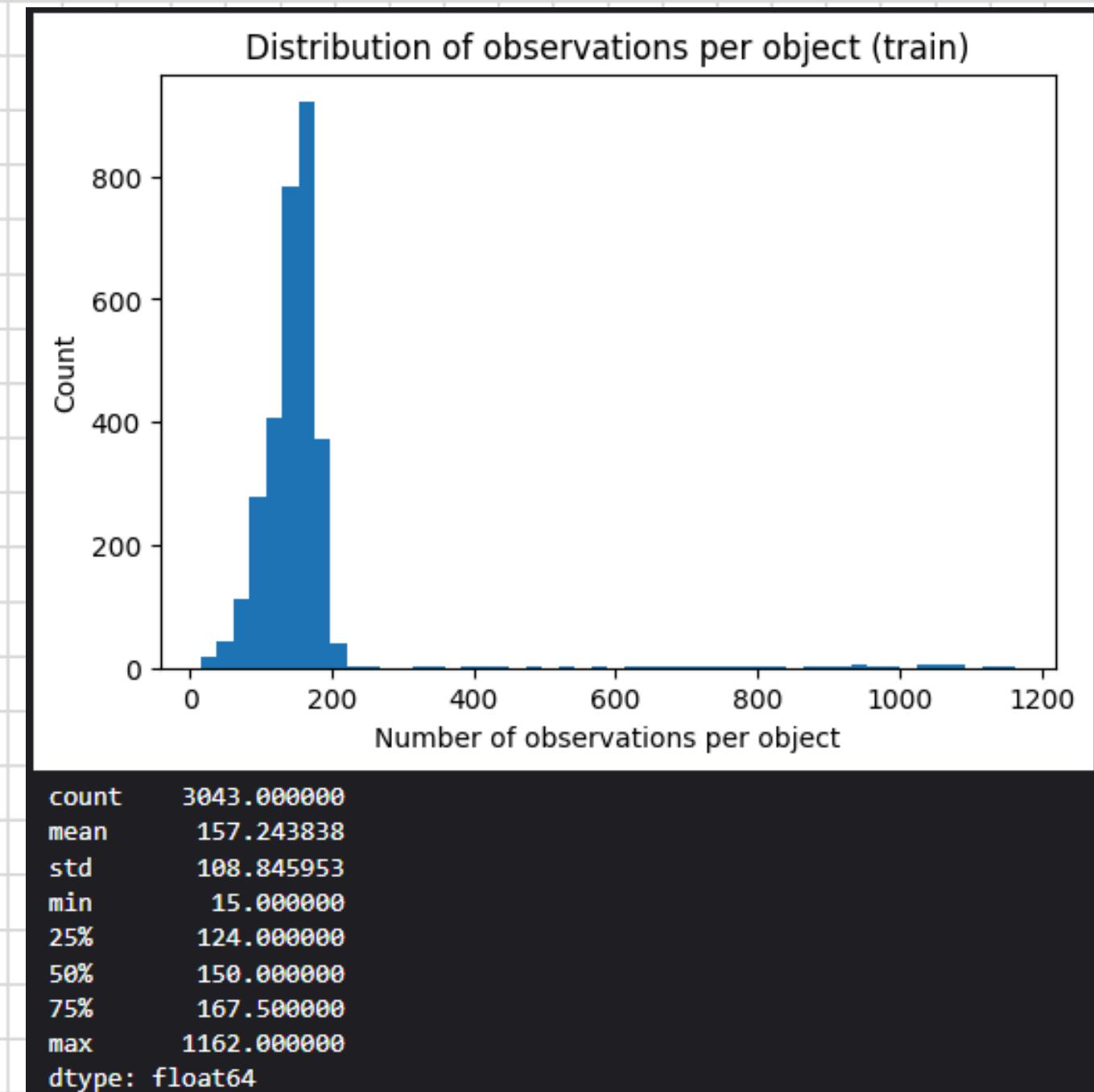
EDA

Số điểm quan sát trên mỗi object

Nhận xét:

- Số điểm quan sát trên mỗi object không đồng đều
- Có object rất ít điểm, có object rất nhiều điểm

→ Không thể dùng trực tiếp raw time-series



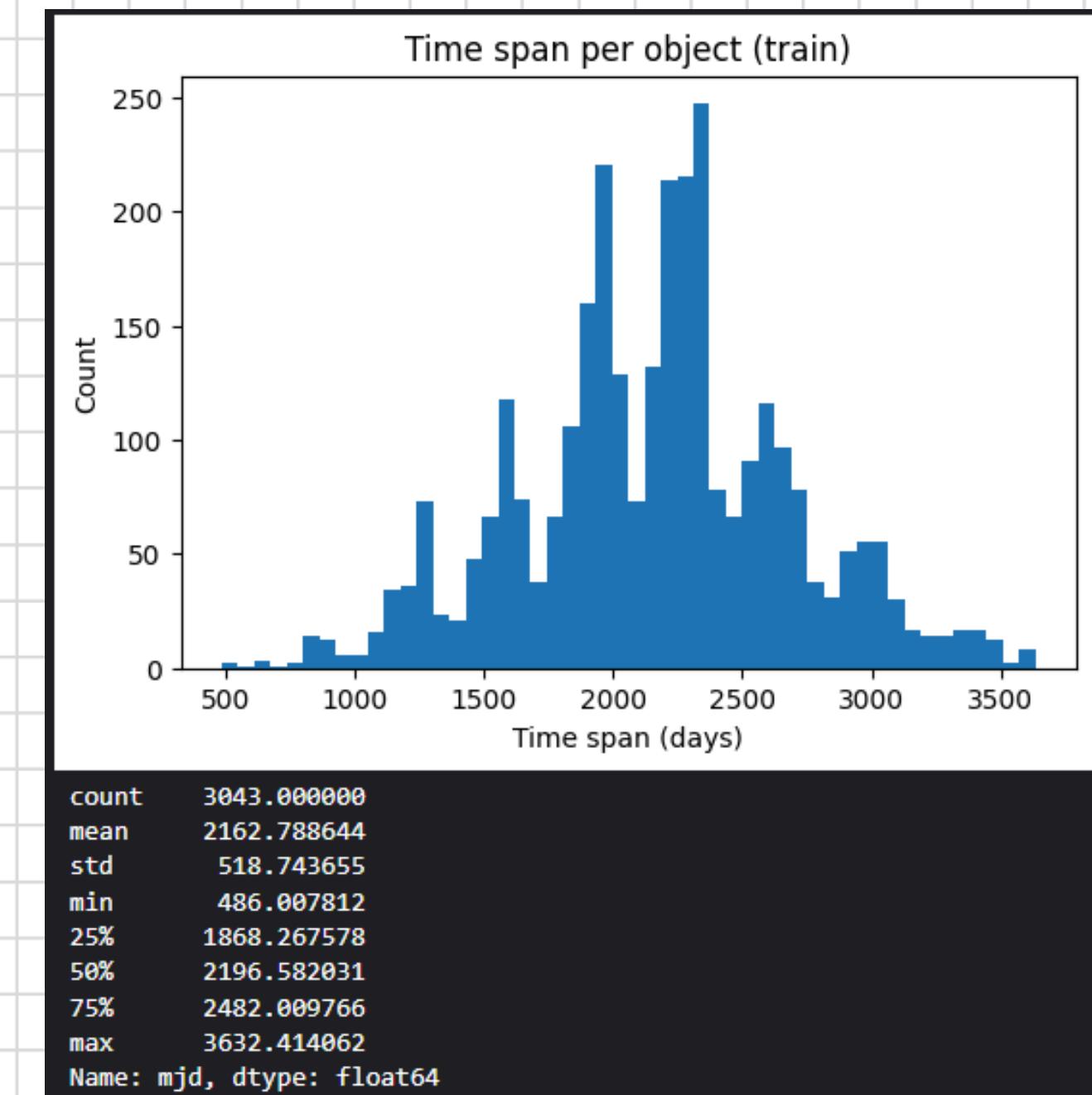
EDA

Time span của lightcurve

Nhận xét:

- Thời gian quan sát kéo dài
nhiều năm
- Phân bố time span khá rộng
- TDE là sự kiện bùng phát trong
một thời gian nhỏ

→ Phải thiết kế các features dựa
trên hình dạng đoblin sáng



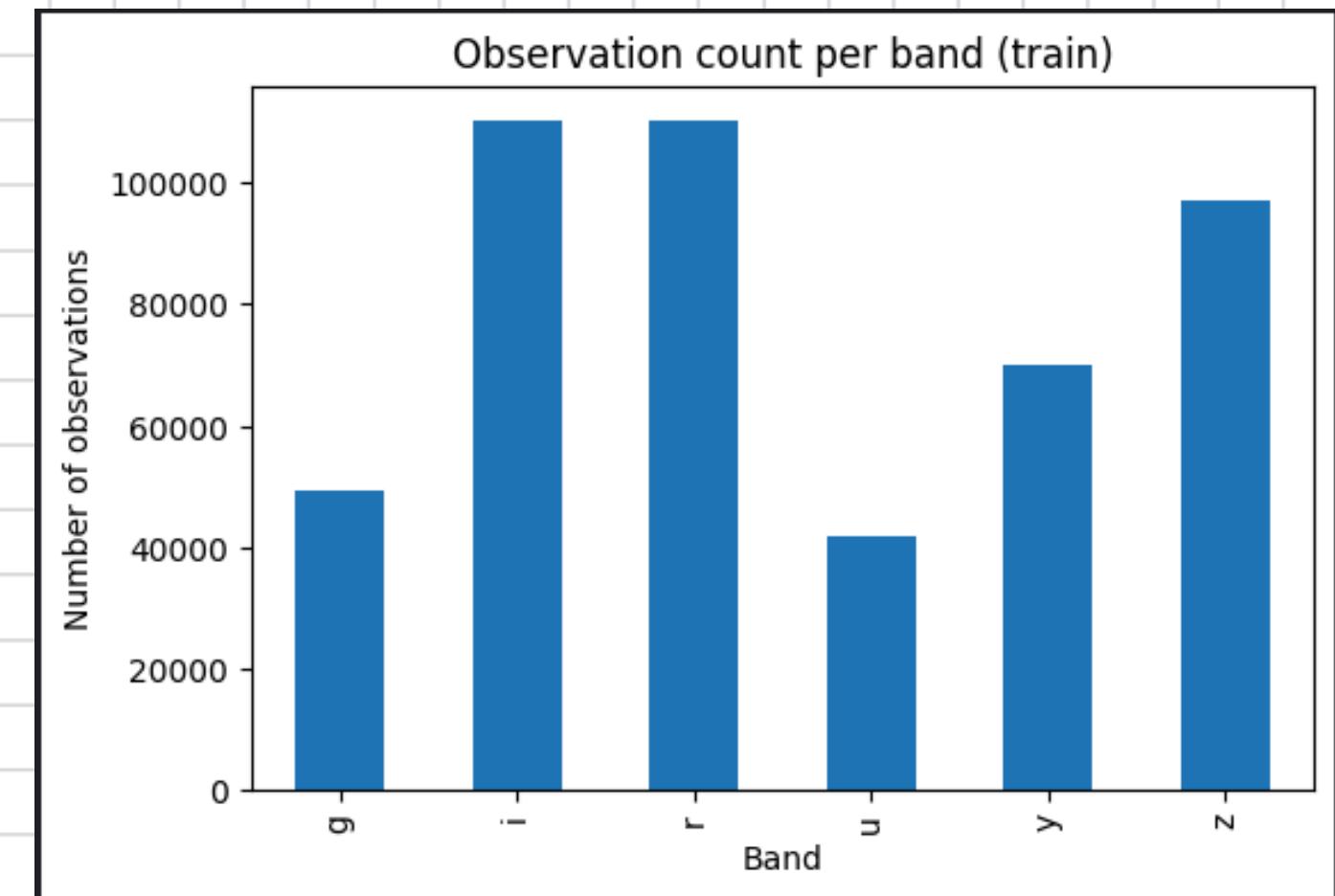
EDA

Phân bố quan sát theo băng tần

Nhận xét:

- r, i có nhiều quan sát nhất
- u và g ít hơn đáng kể
- Số lượng quan sát không đồng đều giữa các band

→ Phải thiết kế các features sao cho có thể xử lý được giá trị thiếu và các feature màu chỉ được tính khi đủ band liên quan



```
band
g    49145
i   110121
r   110464
u    41585
y    70020
z   97158
Name: count, dtype: int64
```

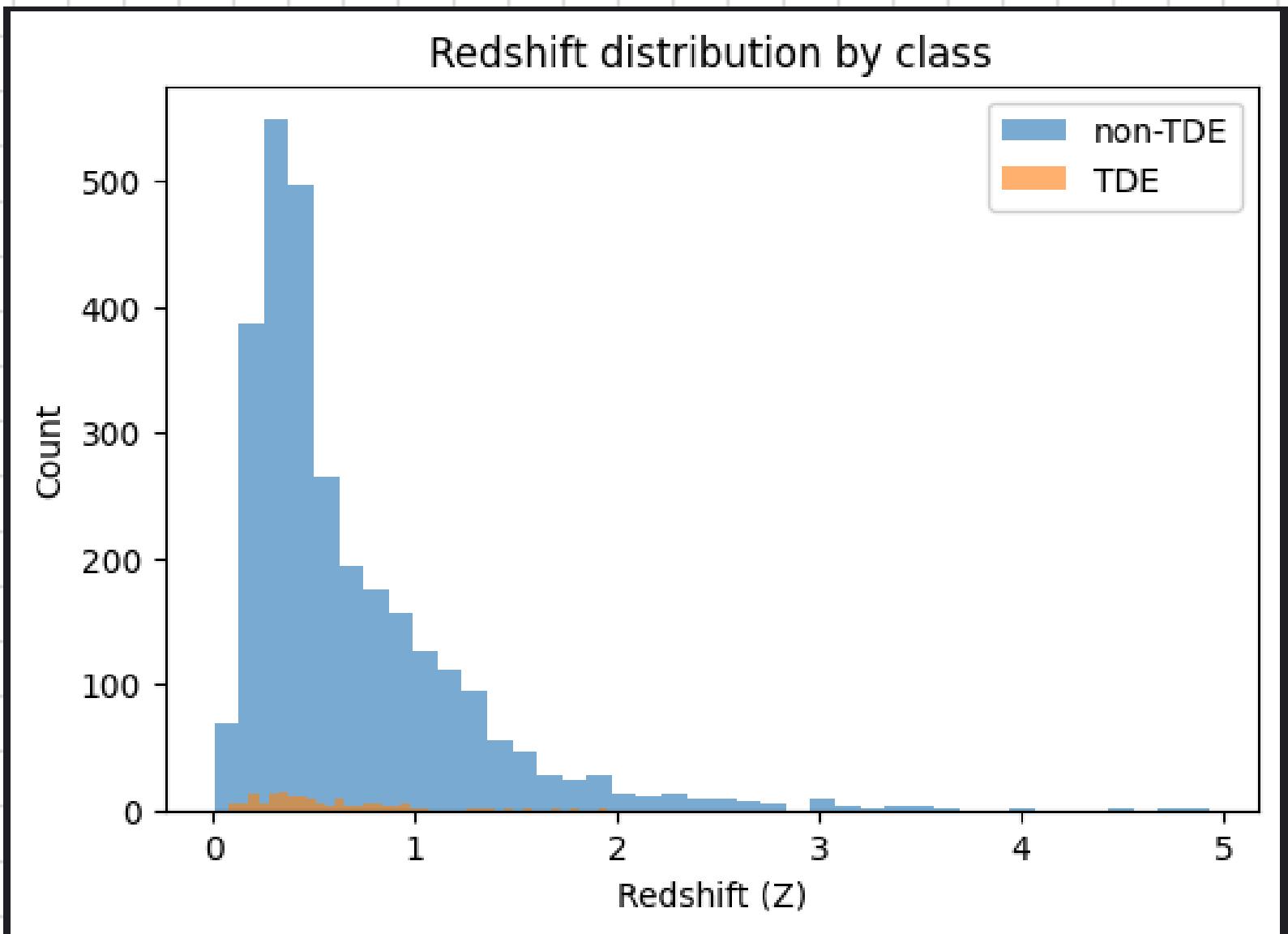
EDA

Redshift (Z) theo lớp

Nhận xét:

- Phân bố redshift giữa hai lớp khác nhau
- TDE chủ yếu xuất hiện ở Z thấp – trung bình
- Non-TDE có dải Z rộng hơn

→ Sự khác biệt cho thấy Z mang thông tin phân biệt nhất định



EDA

Kết luận

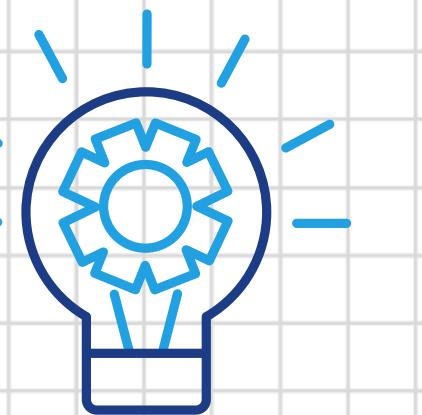
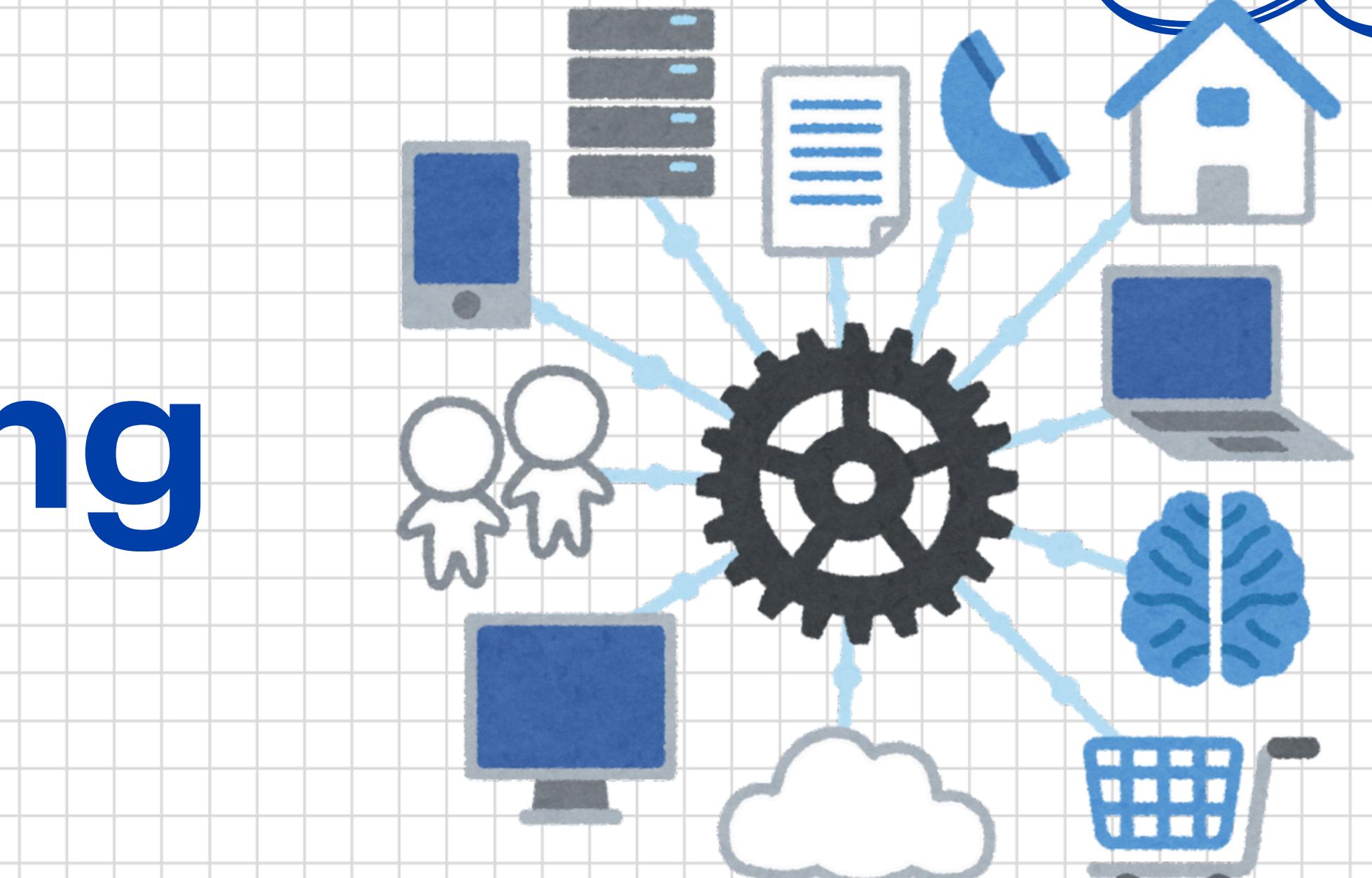
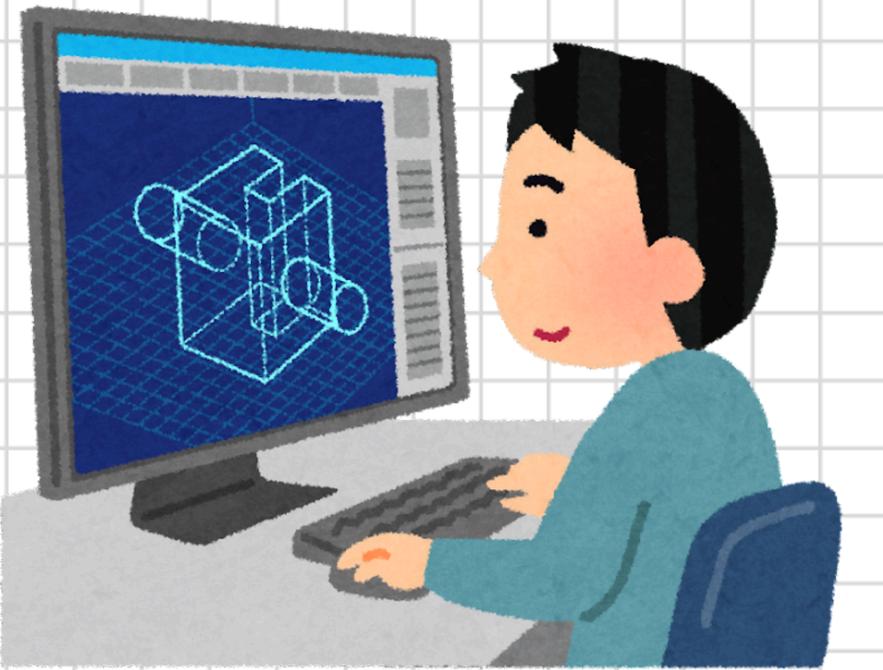
Nhận xét:

- **Mất cân bằng lớp**
- **Số điểm quan sát không đều**
- **Time span dài, TDE là peak**
- **Thiếu band**
- **Z khác nhau theo lớp**

→ Là nền tảng cho toàn bộ pipeline
mô hình phía sau, quyết định cách
thiết kế feature engineering

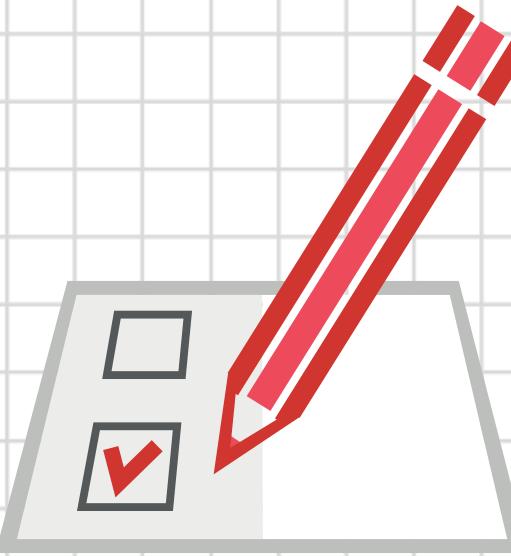
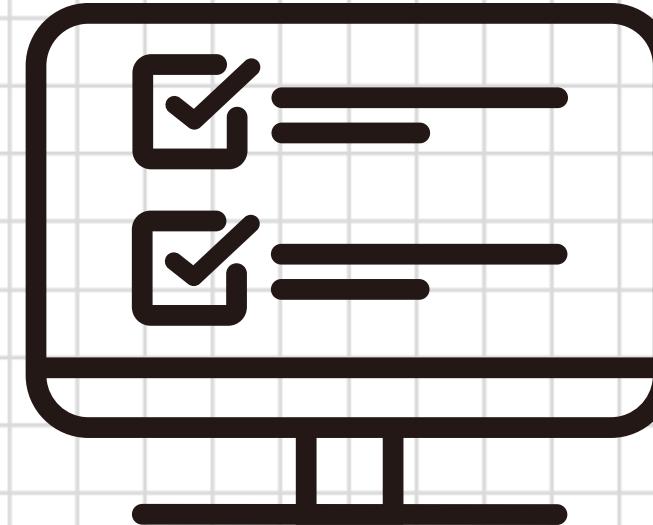


Feature Engineering



Feature Engineering

- Đơn vị dự đoán: object-level
- Không dùng raw time-series
- Trích xuất đặc trưng từ lightcurve
theo 3 nhóm:
 - Aggregate statistical features
 - Peak & shape-based features
 - Physics-informed features



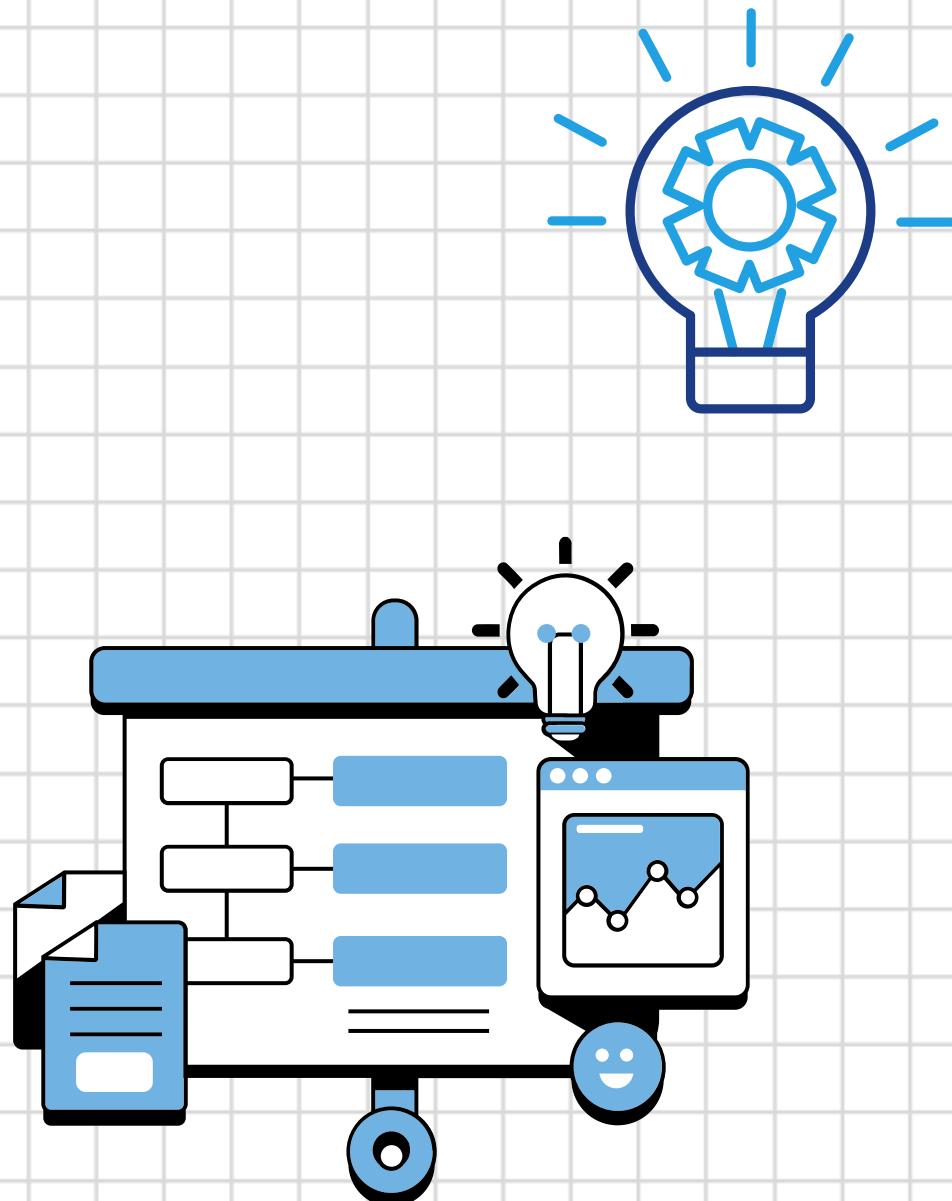
Feature Engineering

Aggregate Features

Trích xuất thống kê toàn cục:

- **n_obs:** tổng số điểm quan sát
- **n_band:** số băng tần được quan sát
- **flux_mean, flux_std, ...:** Thông kê độ sáng
- **snr_mean, snr_max, ...:** Thông kê tỷ số tín hiệu trên nhiễu
- **time_span:** thời gian quan sát

→ Mô tả mức độ biến thiên tổng thể của ánh sáng, mức độ nhiễu,...



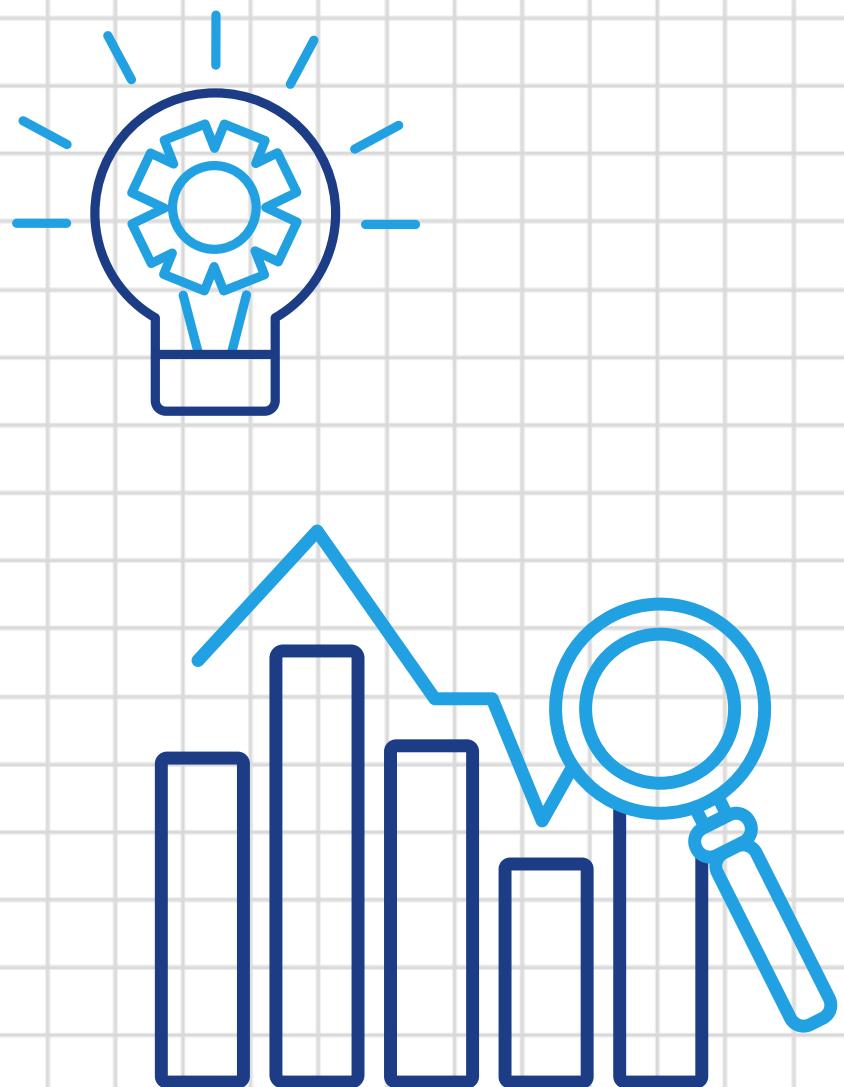
Feature Engineering

Per-band Aggregate Features

Thống kê riêng cho từng băng tần:

- `flux_mean_band_*`, `flux_std_band_*`, ...
- `snr_mean_band_*`, `snr_max_band_*`, ...
- `time_span_band_*`
- `n_obs_band_*`

→ **Do EDA cho thấy số lượng quan sát giữa các băng tần không đồng đều → trích xuất đặc trưng riêng cho từng band**

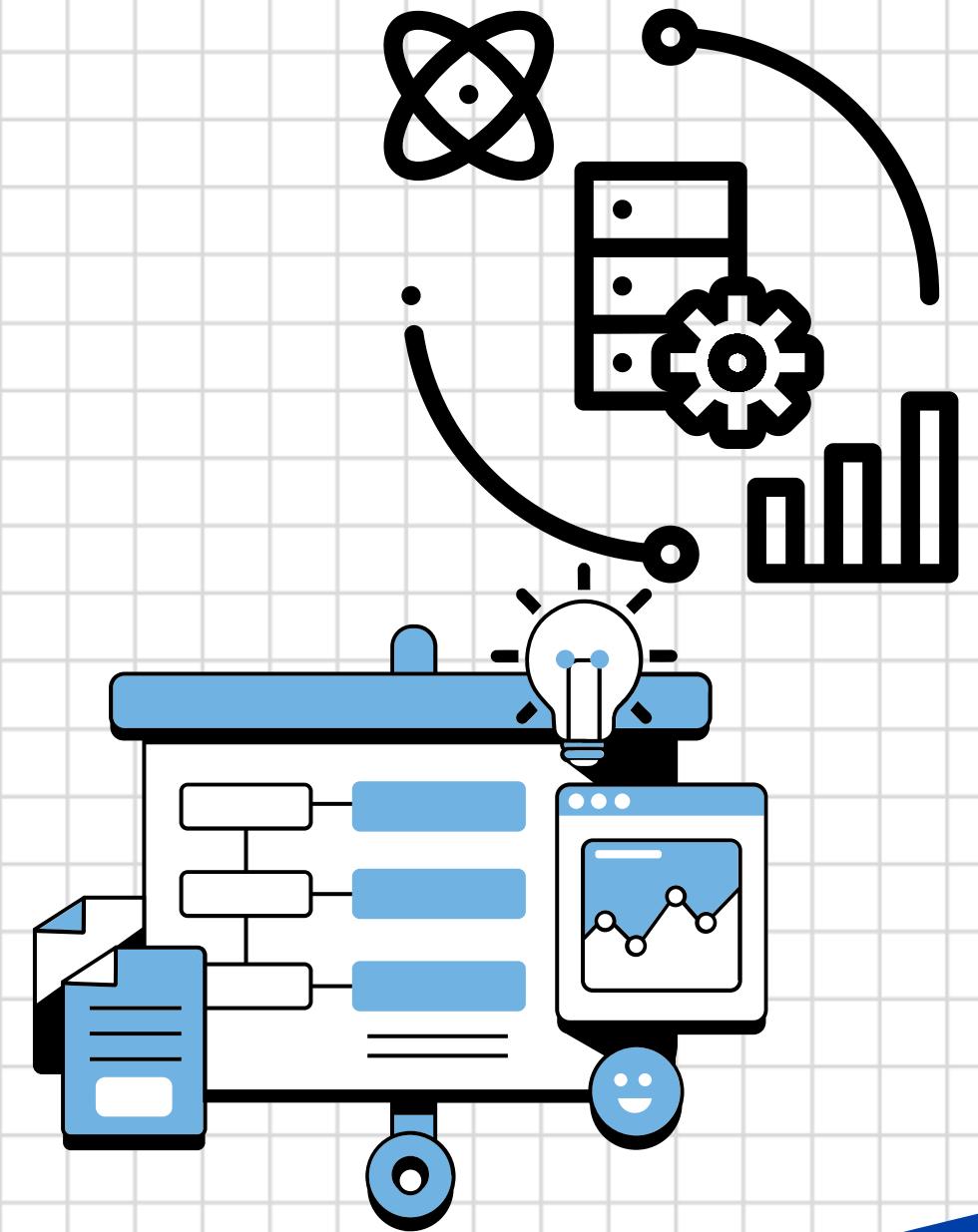


Feature Engineering

Peak-based Feature Engineering

Phân tích:

- TDE là sự kiện bùng phát ánh sáng
- Thông tin quan trọng sẽ nằm ở đỉnh sáng
- Thiết kế các features mô tả:
 - Vị trí đỉnh
 - Độ cao đỉnh
 - Hình dạng tăng-giảm quanh đỉnh



→ TDE chỉ xuất hiện như một đỉnh sáng ngắn →
tập trung trích xuất đặc trưng như trên

Feature Engineering

Peak Shape Features

Với mỗi band:

- **t_peak: thời điểm đạt đỉnh**
- **peak: độ cao đỉnh**
- **rise: thời gian tăng lên đỉnh**
- **decay: thời gian suy giảm sau đỉnh**
- **global_peak: đỉnh mạnh nhất trong các band**

→ Các đặc trưng này mô tả được hình dạng bùng phát ánh sáng



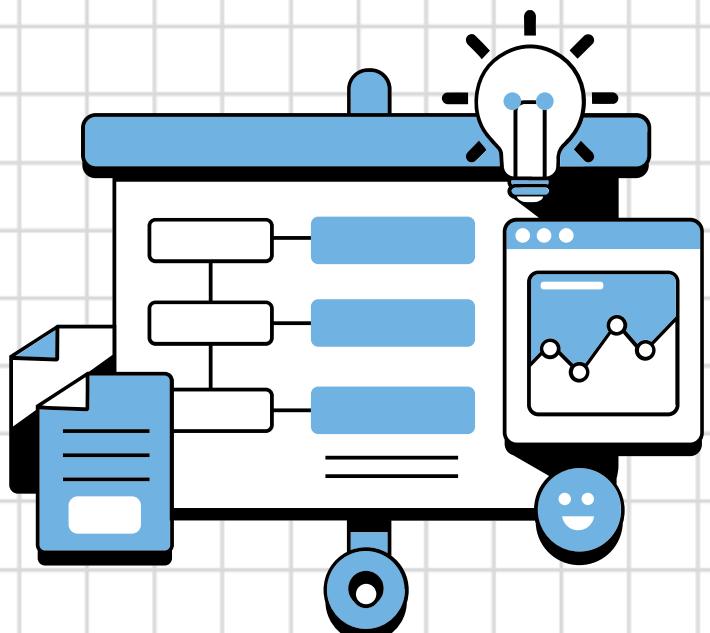
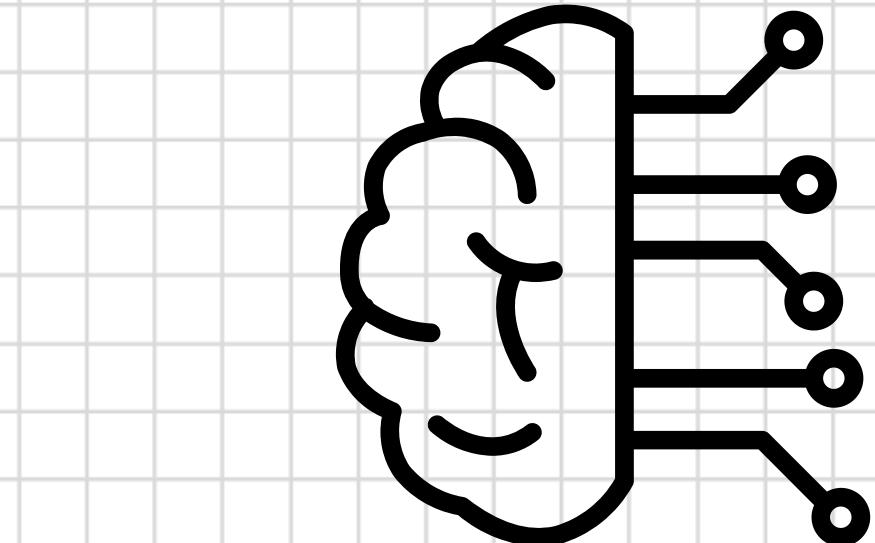
Feature Engineering

Physics-informed Features

Các features đặc biệt:

- color_g_r, color_u_g,...: màu sắc
- Peak flux được hiệu chỉnh theo band
- Log-flux: proxy cho độ sáng biểu kiến

→ Các đặc trưng vật lý giúp mô hình không chỉ học được từ thống kê mà còn từ ý nghĩa vật lý của ánh sáng



Feature Engineering

Xử lý missing data

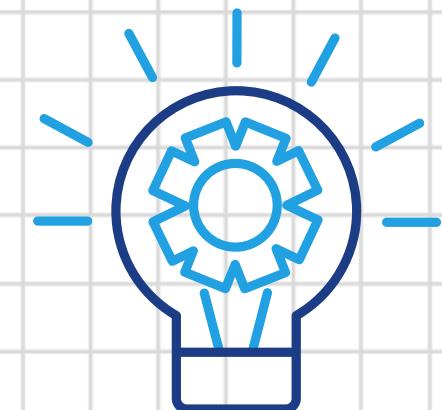
Với cách xử lý features ở trên thì missing sẽ xuất hiện khi:

- Object không có đủ band
- Không có đủ điểm để tính thống kê

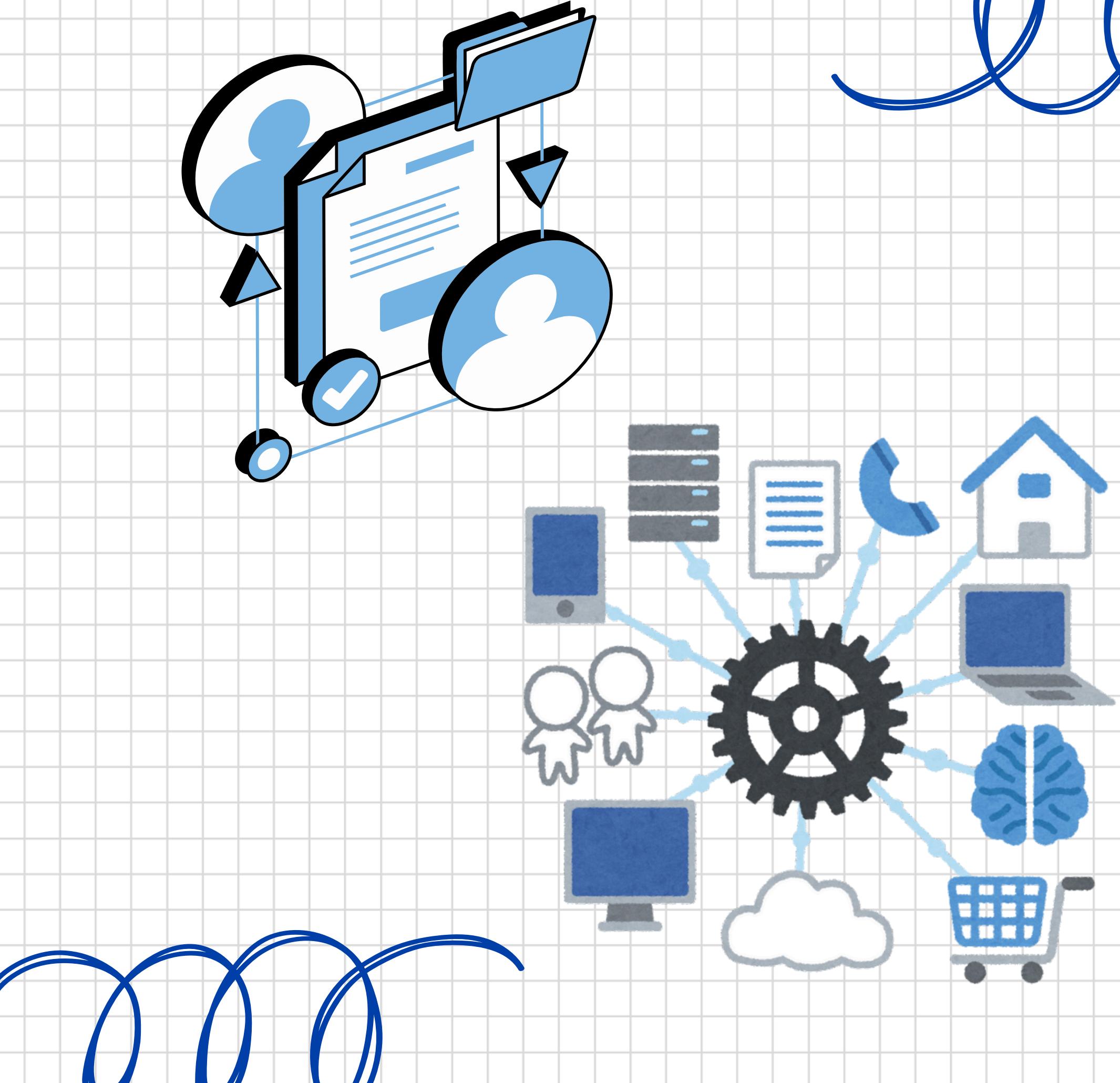
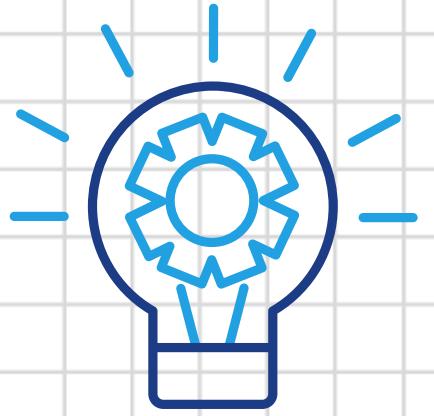
Chiến lược:

- Giữ NaN ở giai đoạn features
- Khi train model: chuyển sang numeric
- Fill bằng -999

→ Quyết định không bỏ object thiếu band vì sẽ làm mất đi
giữ liệu quý giá, fill bằng -999 để mô hình cây tự học cách
xử lý missing data



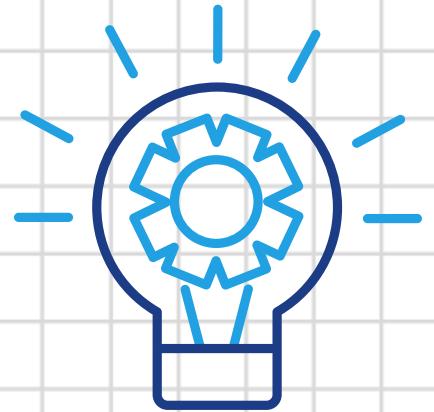
Modeling



Modeling

Định hướng:

- **Bài toán: Binary classification (TDE vs non-TDE)**
- **Đặc điểm dữ liệu:**
 - Feature tabular (sau feature engineering)
 - Mật cân bằng lớp nghiêm trọng
- **Chiến lược:**
 - Mô hình cây quyết định tăng cường (Boosting)
 - Cross-validation theo group
 - Tối ưu F1-score

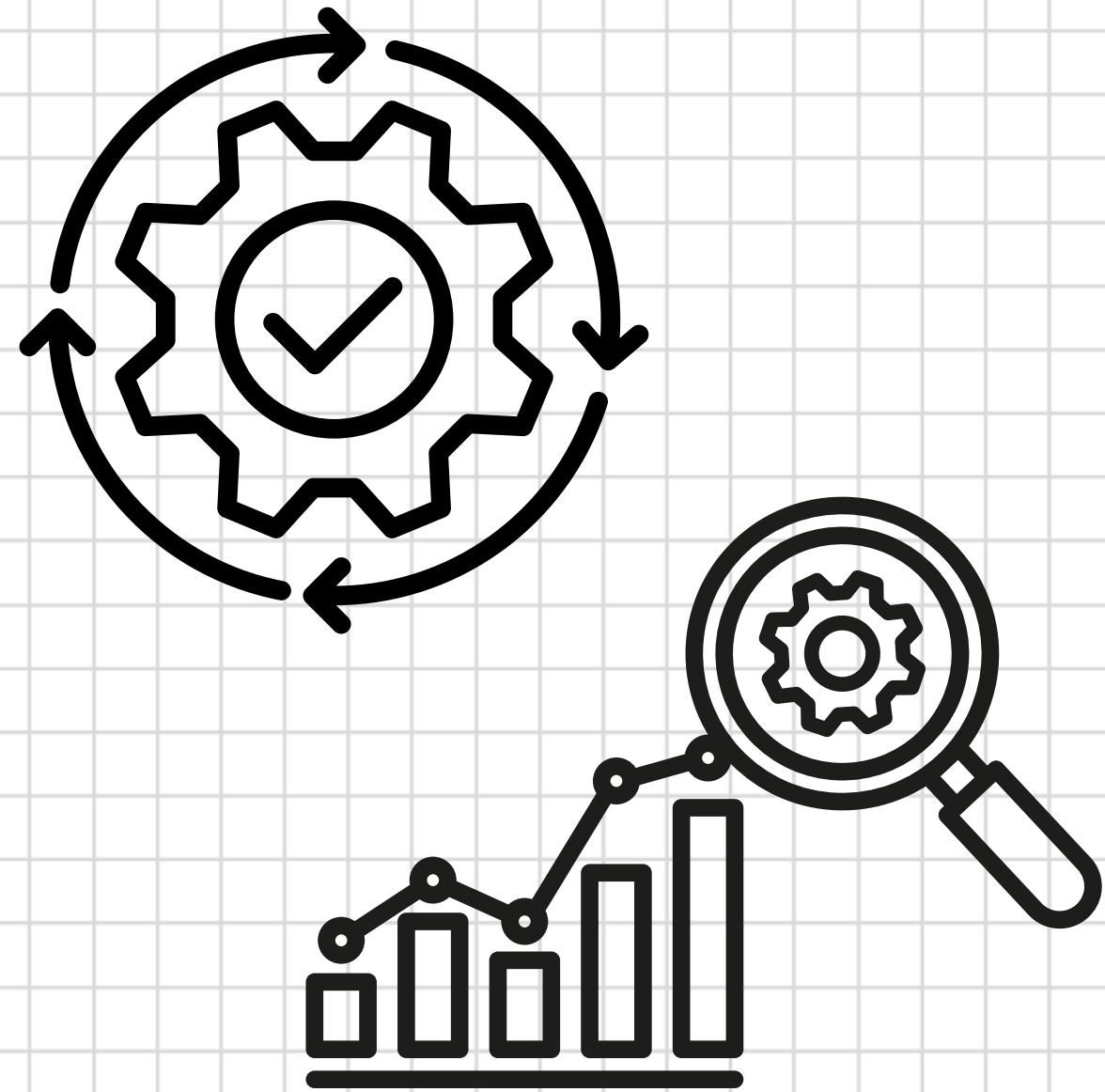


Modeling

Vì sao chọn mô hình cây boosting:

- Không yêu cầu chuẩn hoá feature
- Xử lý tốt:
 - Feature không tuyến tính
 - Missing value (sentinel)
- Hiệu quả với dữ liệu tabular
- Dễ ensemble

→ So với các mô hình thuần tuý, các mô hình boosting như LightGBM, CatBoost và XGBoost tỏ ra rất hiệu quả với dữ liệu tabular có nhiều đặc trưng phức tạp



Modeling

Các mô hình được sử dụng:

- LightGBM: Tốc độ nhanh, hiệu quả
- CatBoost: Bổ sung khả năng học các quan hệ phức tạp
- XGBoost: Bổ sung thêm góc nhìn khác để bước ensemble hiệu quả hơn

→ Sử dụng nhiều mô hình boosting khác nhau nhằm tận dụng sự khác biệt trong cách học của từng thuật toán

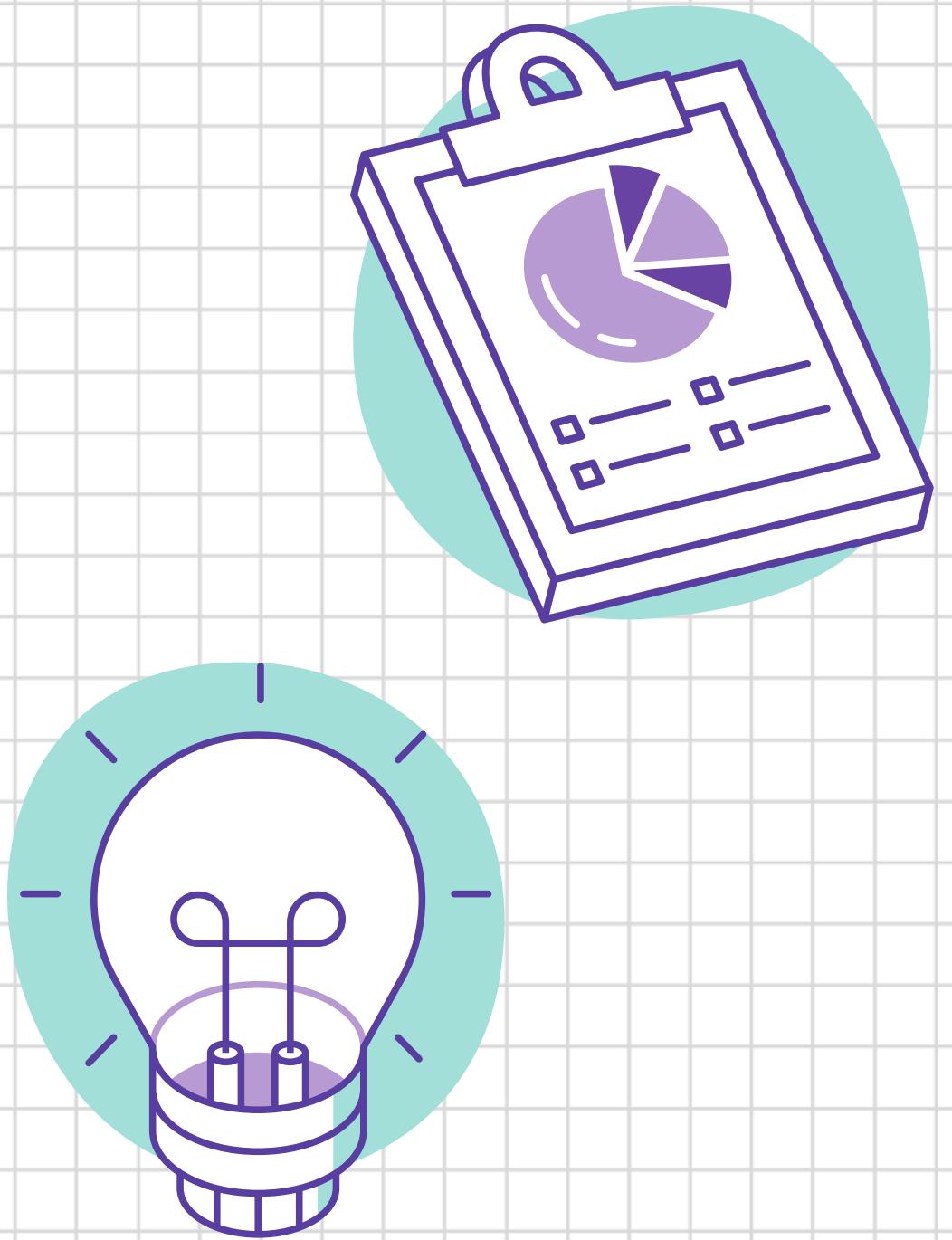


Modeling

Threshold Tuning cho F1 score

- Mô hình trả về xác suất
- Không cố định threshold = 0.5
- Quét threshold $\in [0.01, 0.99]$
- Chọn threshold tối ưu F1 trên OOF

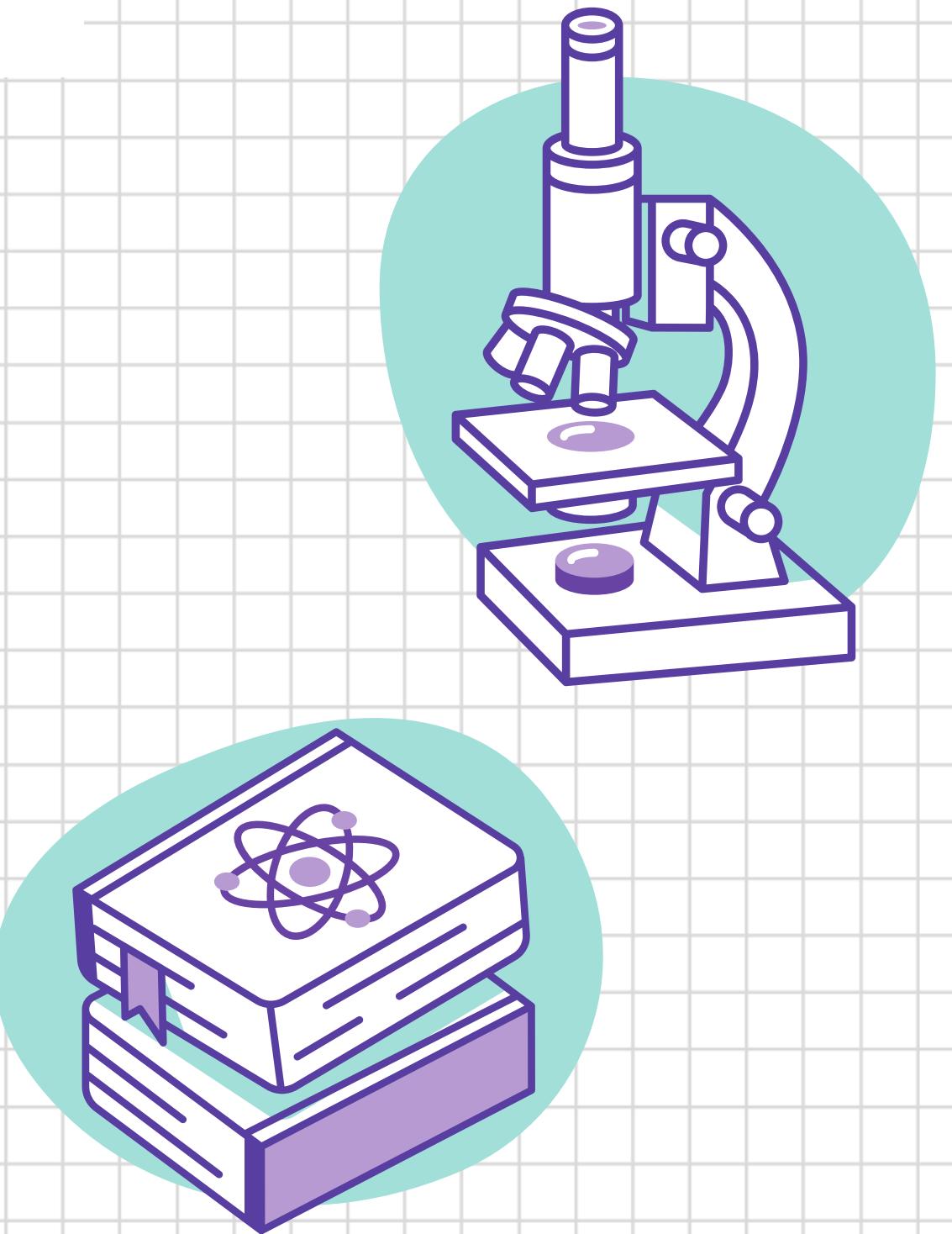
→ Do dữ liệu mêt cân bằng → ngưỡng phân loại 0.5 không tối ưu → tìm ngưỡng phân loại phù hợp nhất để cân bằng giữa precision và recall



Modeling

Ensemble: Kết hợp nhiều mô hình

- Ensemble xác suất từ:
 - LightGBM
 - CatBoost
 - XGBoost
 - Lý do:
 - Giảm phương sai
 - Tận dụng ưu điểm từng mô hình
- Mỗi mô hình có xu hướng mắc lỗi khác nhau → Ensemble giúp giảm rủi ro dự đoán sai và cải thiện kết quả



Modeling

Tối ưu trọng số Ensemble

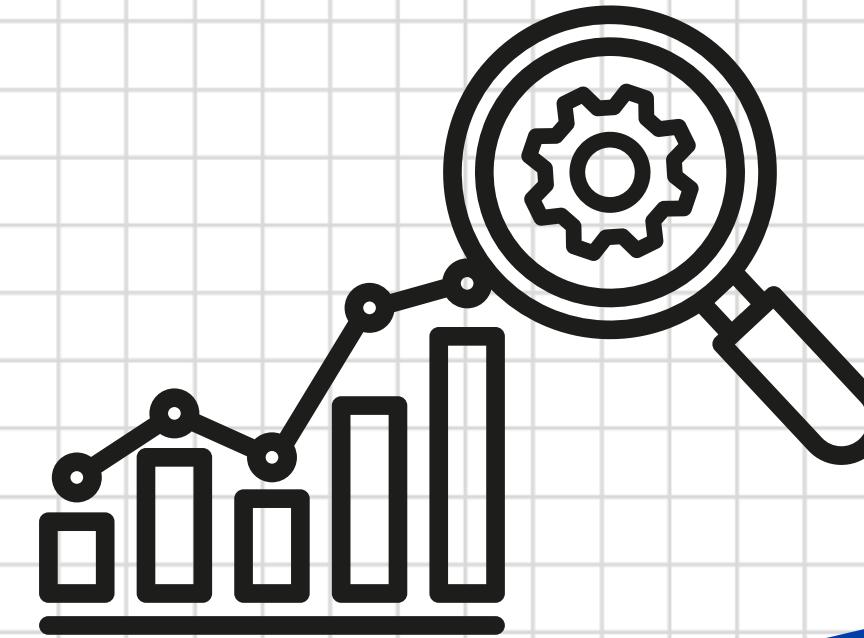
- Linear blend:

$$P = w_{xgb}P_{xgb} + w_{lgb}P_{lgb} + w_{cat}P_{cat}$$

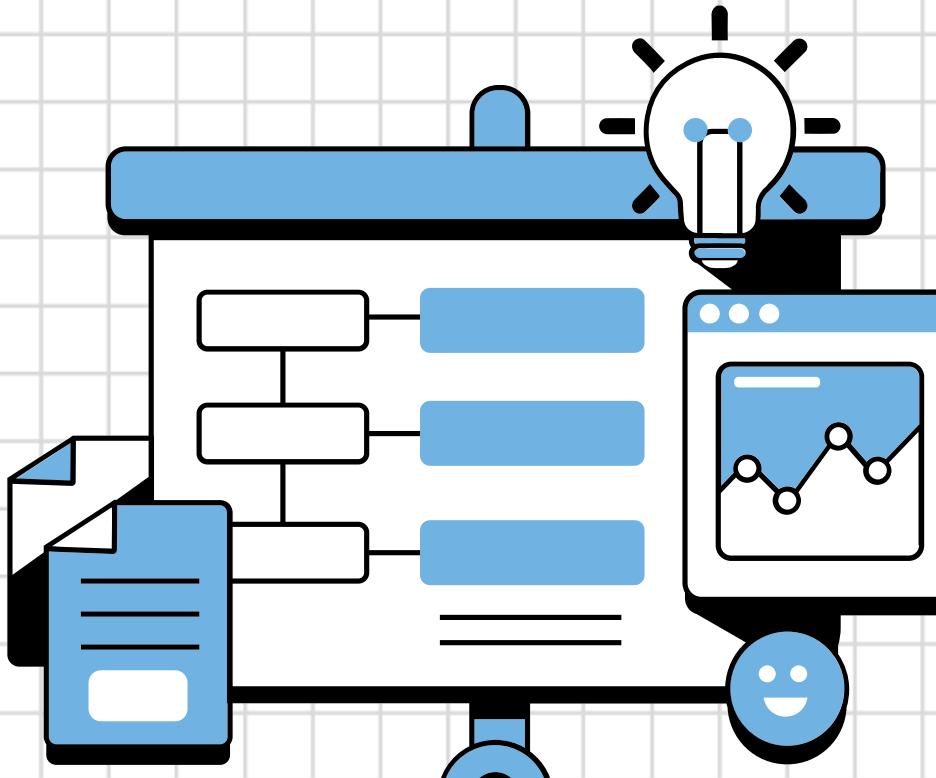
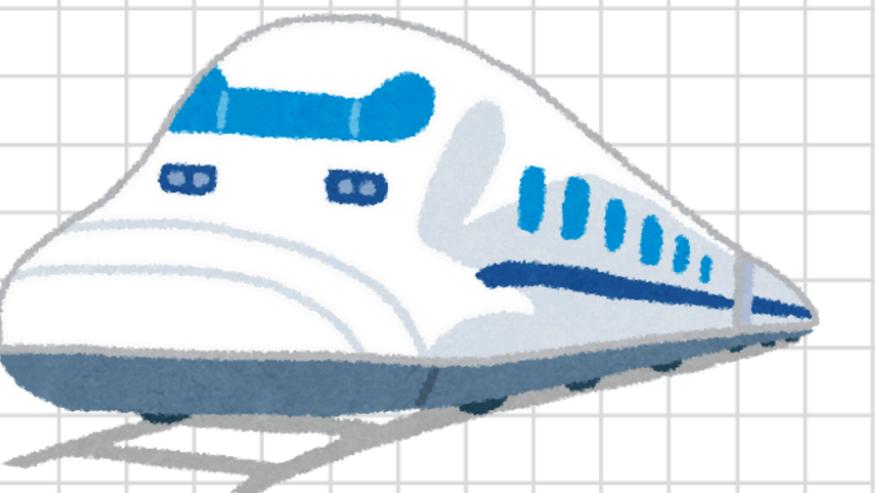
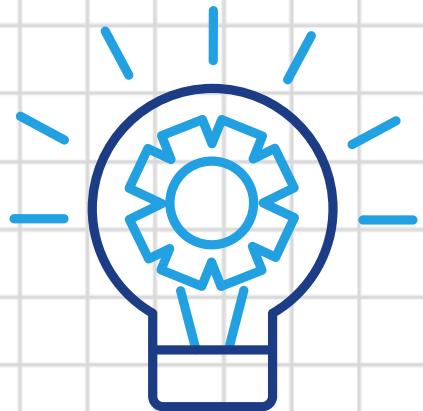
- Điều kiện:

- $w_i \geq 0$
- $\sum w_i = 1$

→ Không chọn trọng số bằng cảm tính mà
thực hiện grid search trên OOF prediction
để tìm bộ tối ưu



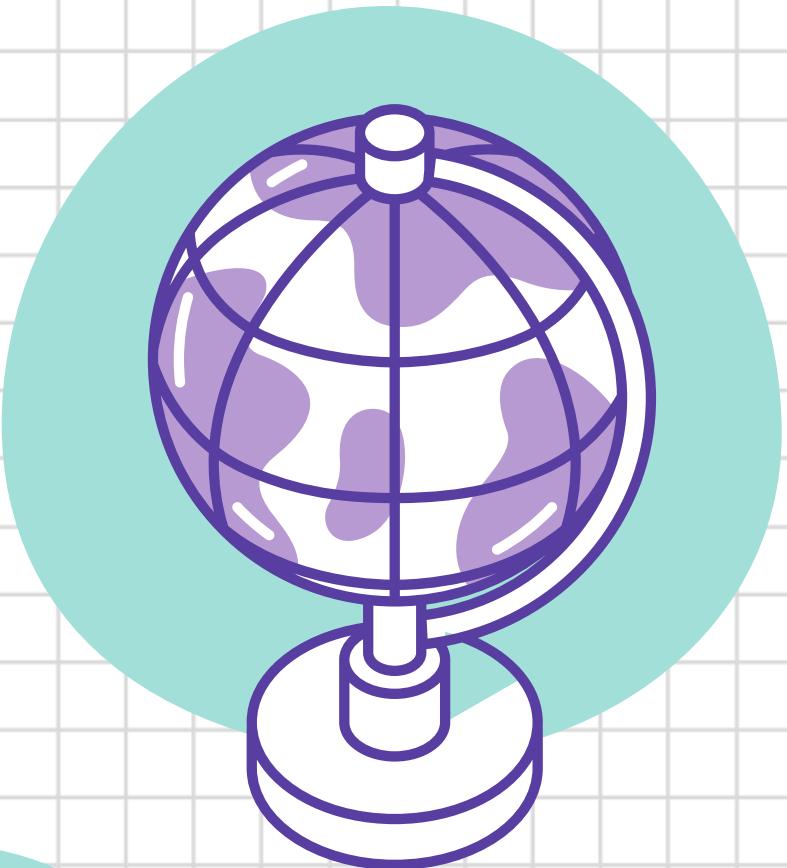
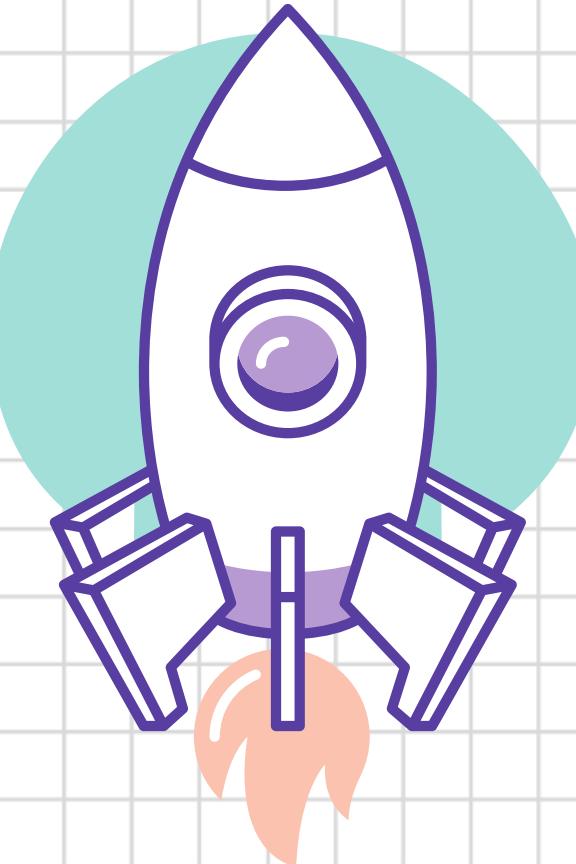
Kết Quả



Kết quả

Cách đánh giá:

- Đánh giá dựa trên:
 - Out-of-Fold (OOF) prediction
 - Cross-validation theo group
- Metric chính: F1-score
- Threshold được tối ưu trên OOF

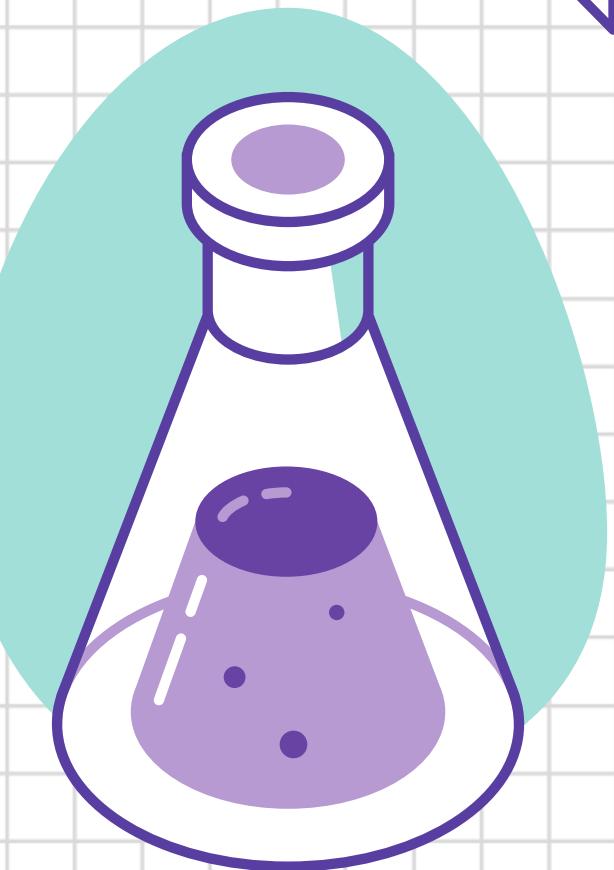
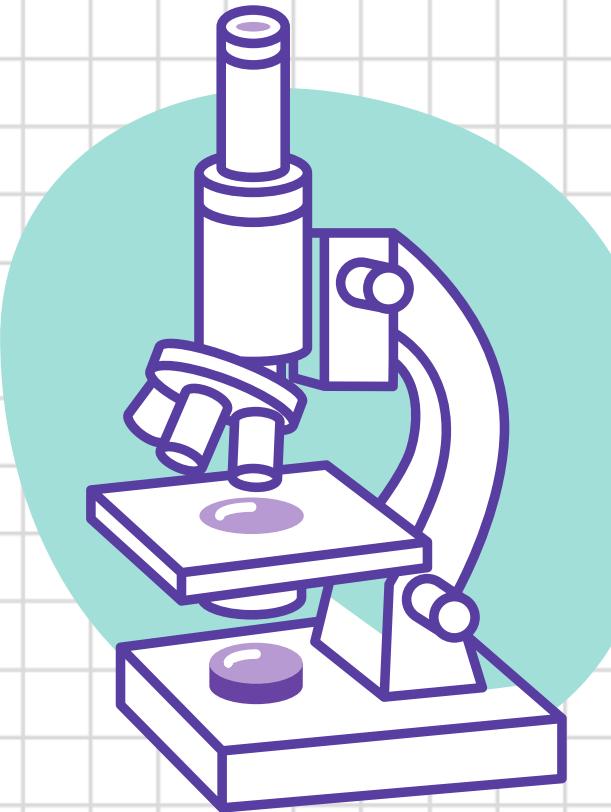


Kết quả

Kết quả sau finetuned của từng mô hình(F1):

- LightGBM: 0.5806
- CatBoost: 0.5490
- XGBoost: 0.5637

→ Mỗi mô hình đơn lẻ có kết quả khá tương đồng, tuy nhiên mỗi mô hình vẫn có những điểm mạnh và hạn chế riêng

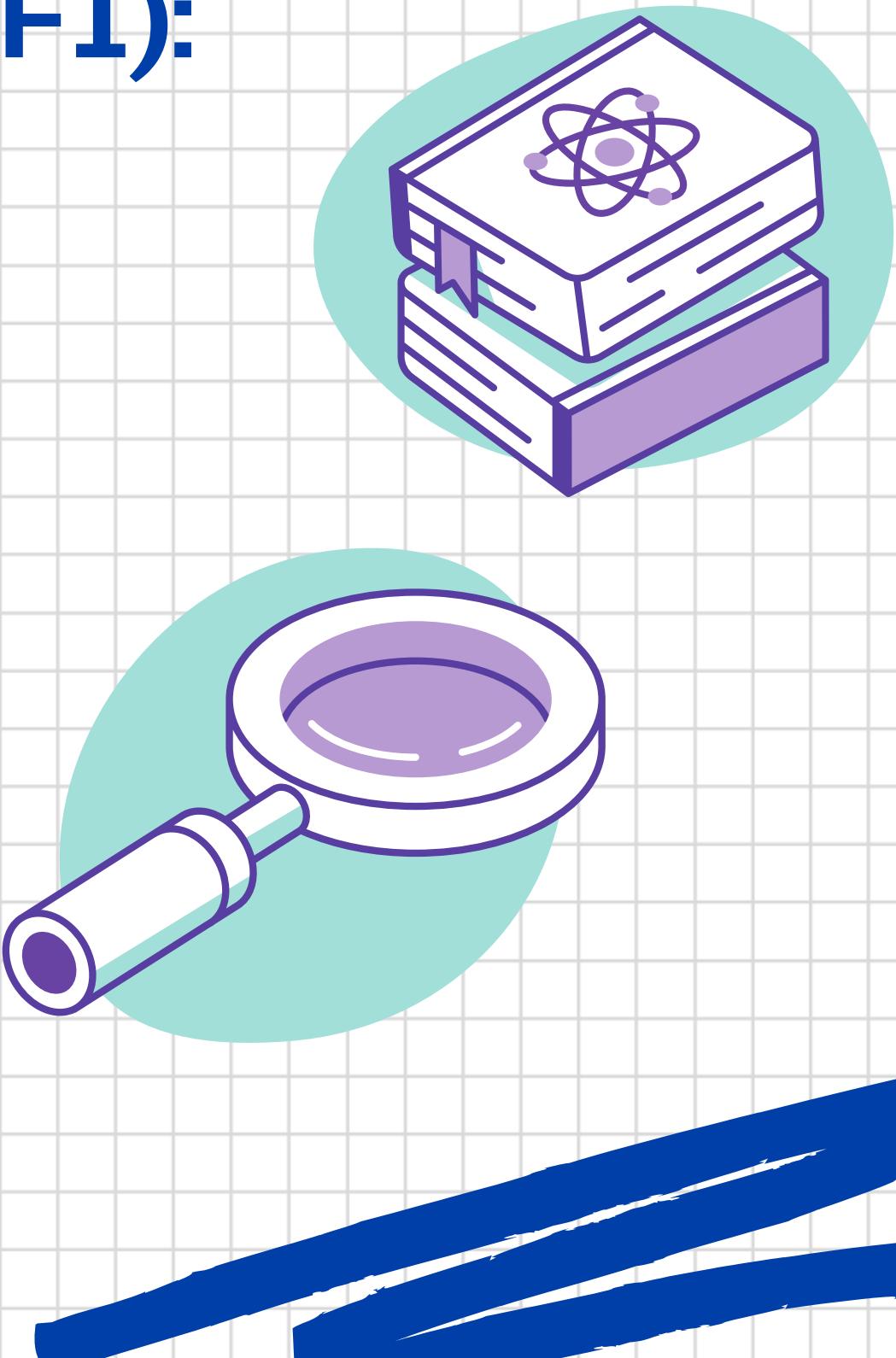


Kết quả

Kết quả sau khi ensemble các mô hình (F1):

- LightGBM + XGBoost: 0.5852
- LightGBM + CatBoost: 0.5758
- CatBoost + XGBoost: 0.5978
- CatBoost + XGBoost + LightGBM: 0.6014

→ Ensemble tận dụng sự khác biệt trong cách học của từng mô hình → mang lại kết quả tốt hơn hẳn so với các mô hình đơn lẻ



Kết luận

- **Bài toán phát hiện TDE từ lightcurve đa băng tần:**
 - Là bài toán khó
 - Dữ liệu cực kỳ mêt cân bằng
 - Dữ liệu giàu nhiễu



Kết luận

- Nhóm tiếp cận bài toán theo quy trình rất chặt chẽ:
 - Phân tích dữ liệu (EDA) để hiểu bản chất
 - Thiết kế feature engineering dựa trên EDA
 - Xây dựng và đánh giá mô hình với Group Cross-Validation, tối ưu F1-score
 - Sử dụng và finetune các mô hình boosting (LightGBM, CatBoost, XGBoost)
 - Kết hợp ensemble cho kết quả ổn định và vượt trội so với mô hình đơn lẻ



**Thank you for
your listening**

