

TRUSTEE

... Tree-based Regression
for Undergraduate Student Tracking and
Educational Explainability ...

NỘI DUNG CHÍNH

1. Giới thiệu bài toán
2. Xử lý dữ liệu
3. Phương pháp đề xuất
4. Kết quả thực nghiệm
5. Kết luận

GIỚI THIỆU BÀI TOÁN

- Thực trạng : Tình trạng sinh viên đăng ký nhiều nhưng tỷ lệ trượt môn/cảnh báo học vụ cao đang là thách thức lớn
- Mục tiêu : Xây dựng mô hình hồi quy dự báo số tín chỉ sinh viên hoàn thành ngay từ đầu kỳ
- Hướng tiếp cận:
 - Dữ liệu: Năng lực nền tảng + Lịch sử học tập + Kế hoạch đăng ký
 - Công nghệ: Tree-based Models + Explainable AI
 - Đích đến: Giải pháp cảnh báo sớm + Cá nhân hóa đào tạo



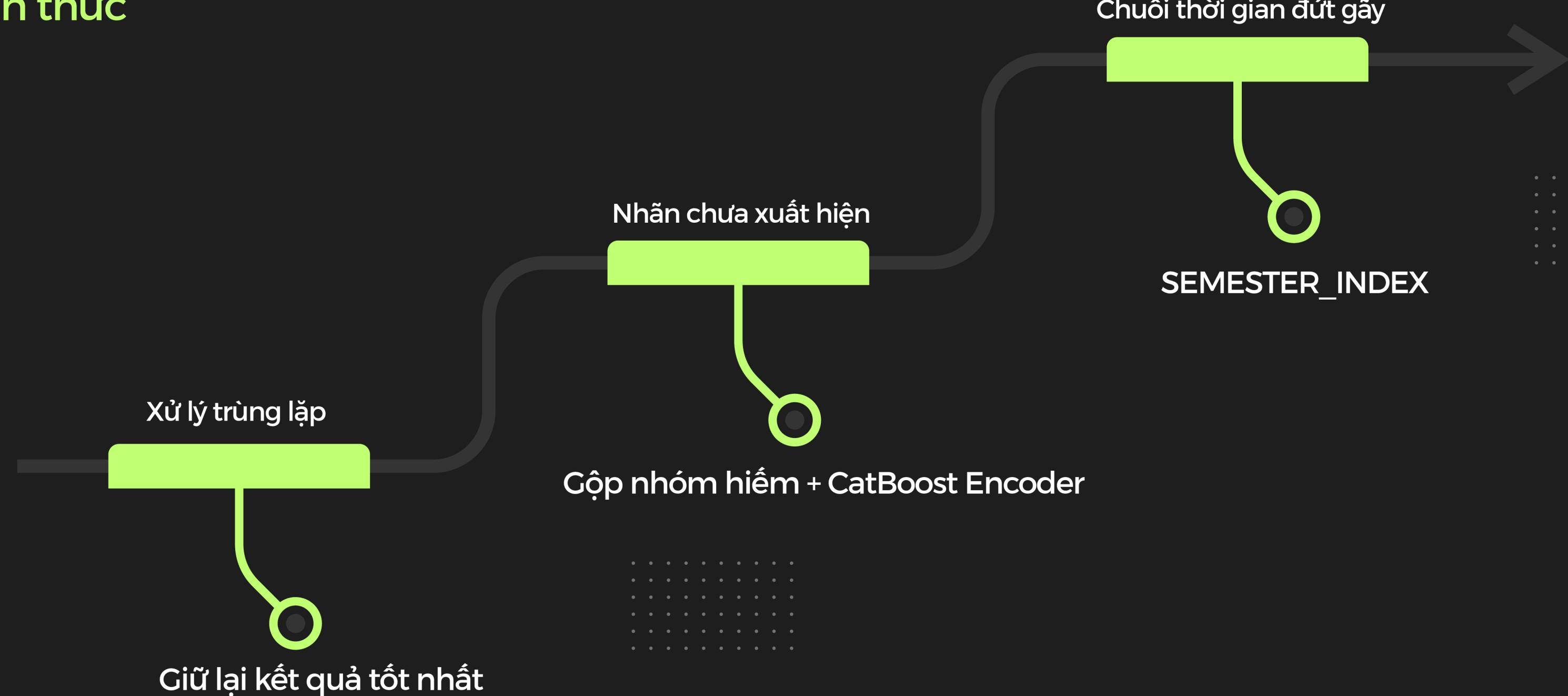
XỬ LÝ DỮ LIỆU

Dữ liệu tổng quan

Bảng dữ liệu	addmission.csv	academic_records.csv	test.csv
Nội dung	Thông tin tuyển sinh	Lịch sử học tập	Tập kiểm thử
Số lượng mẫu	30.217 sinh viên	20.381 sinh viên	16.502 sinh viên
Các đặc trưng	Năm tuyển sinh, phương thức xét tuyển, tổ hợp môn xét tuyển...	GPA, CPA, số tín chỉ đăng kí, số tín hoàn thành	Số tín chỉ đăng kí HK1 2024-2025

XỬ LÝ DỮ LIỆU

Thách thức



DỮ LIỆU BỔ SUNG

Phổ điểm THPT Quốc gia (2020 - 2024)

Z_SCORE

Độ lệch chuẩn (So với
mặt bằng chung năm
đó)

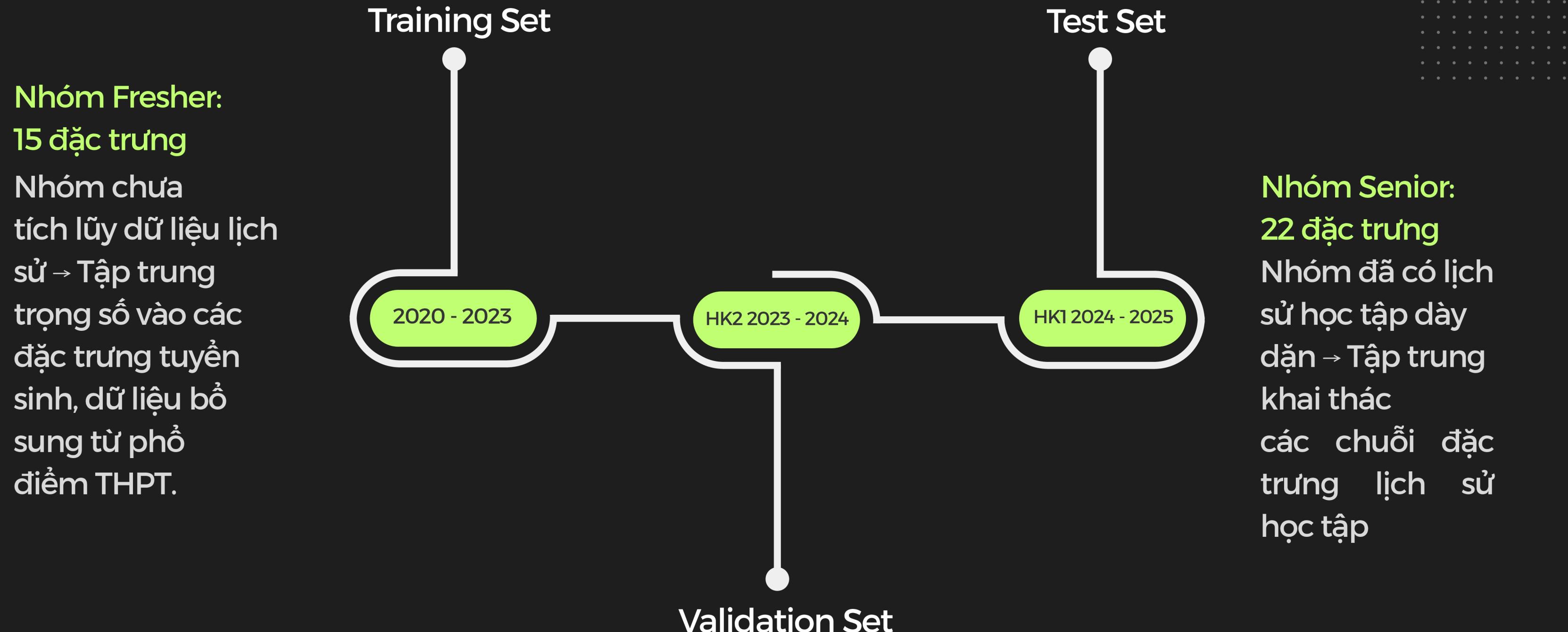
SCORE_GAP

Khoảng cách an toàn
(Điểm thi - Điểm
chuẩn)

ENTRY_RANK

Thứ hạng phần trăm
trong khóa

PHÂN CHIA DỮ LIỆU



PHƯƠNG PHÁP ĐỀ XUẤT

Dự báo trực tiếp

$$y_{credits} = f_{credits}(X)$$

- Ánh xạ trực tiếp từ đặc trưng đầu vào sang số tín hoàn thành

Dự báo phần dư

$$y_{gap} = TC_{DANGKY} - f_{gap}(X)$$

- Giảm phương sai: Số tín chỉ trượt thường nhỏ và hội tụ
- Biến mục tiêu ổn định hơn

Dự báo tỉ lệ

$$y_{ratio} = TC_{DANGKY} \times f_{ratio}(X)$$

- Chuẩn hóa quy mô: Loại bỏ bias do số lượng đăng ký quá ít hoặc quá nhiều

MÔ HÌNH ĐỀ XUẤT



EXPLAINABLE AI

...

WHY?

- Công cụ: SHAP
- Định lượng mức độ ảnh hưởng
- So sánh trực diện sinh viên Rủi ro vs. Sinh viên Tốt

IS IT RELIABLE?

- Công cụ: LIME
- Giải thích từng trường hợp cụ thể
- Kiểm tra độ tin cậy của các dự báo

HOW TO FIX?

- Công cụ: DiCE
- Sinh ra các kịch bản giả định
- Dự báo tỷ lệ hoàn thành của sinh viên về ngưỡng an toàn

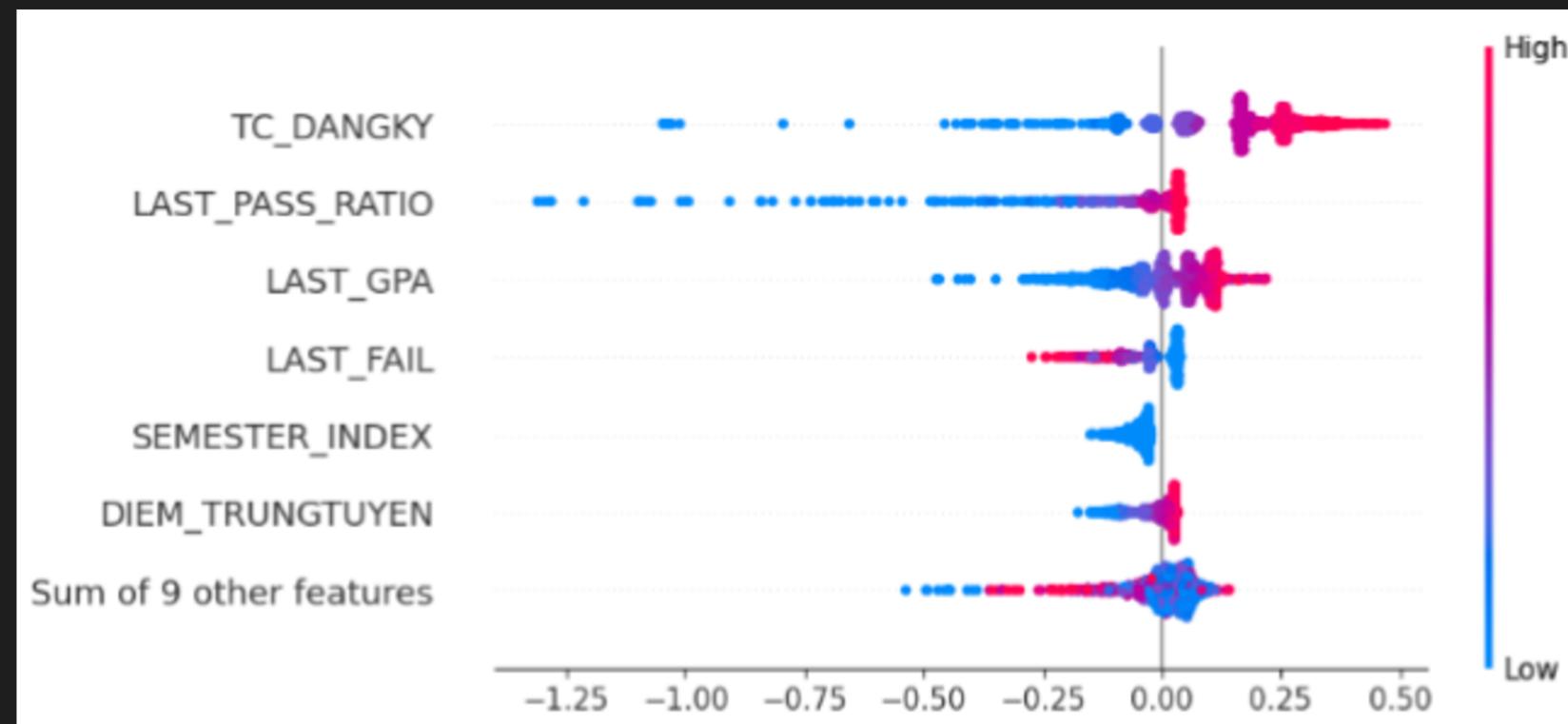


KẾT QUẢ THỰC NGHIỆM

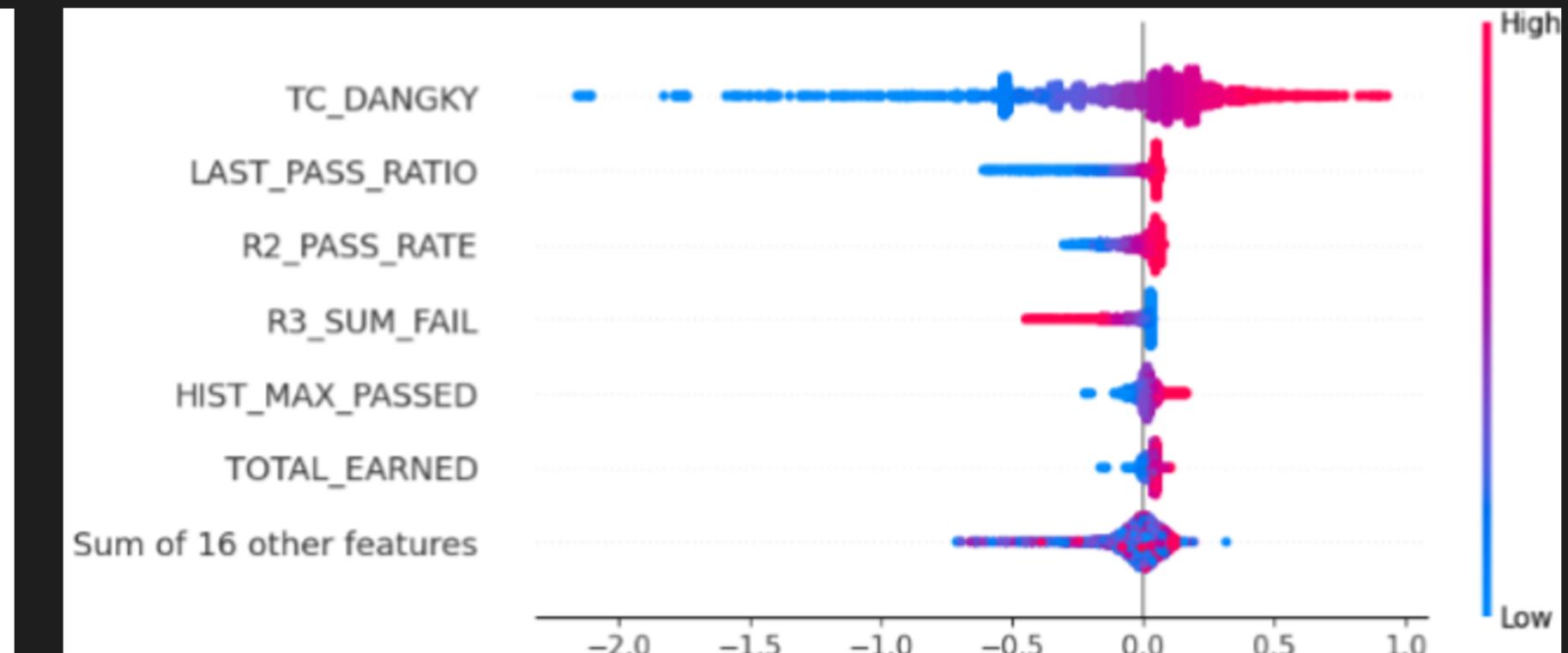
Hướng tiếp cận	Mô hình	Validation Set				Public Test Set
		R-squared	RMSE	MSE	wMAPE	RMSE
Direct Credits Prediction	Decision Tree	0.6794	3.9749	15.8002	0.1928	3.7589
	Random Forest	0.7069	3.7831	14.3115	0.1832	3.5332
	XGBoost	0.7149	3.7488	14.0534	0.1786	3.5431
	LightGBM	0.7147	3.7499	14.0620	0.1777	3.5040
Gap Prediction	Decision Tree	0.6948	3.8785	15.0427	0.1911	3.6859
	Random Forest	0.7148	3.7491	14.0556	0.1809	3.5161
	XGBoost	0.7152	3.7469	14.0395	0.1778	3.5203
	LightGBM	0.7178	3.7295	13.9091	0.1781	3.5171
Ratio Prediction	Decision Tree	0.6961	3.8703	14.9789	0.1870	3.6492
	Random Forest	0.6446	4.1857	17.5198	0.1985	3.5758
	XGBoost	0.7173	3.7328	13.9304	0.1792	3.5371
	LightGBM	0.7166	3.7375	13.9687	0.1800	3.5205

KẾT QUẢ THỰC NGHIỆM

PHÂN TÍCH TOÀN CỤC: SHAP



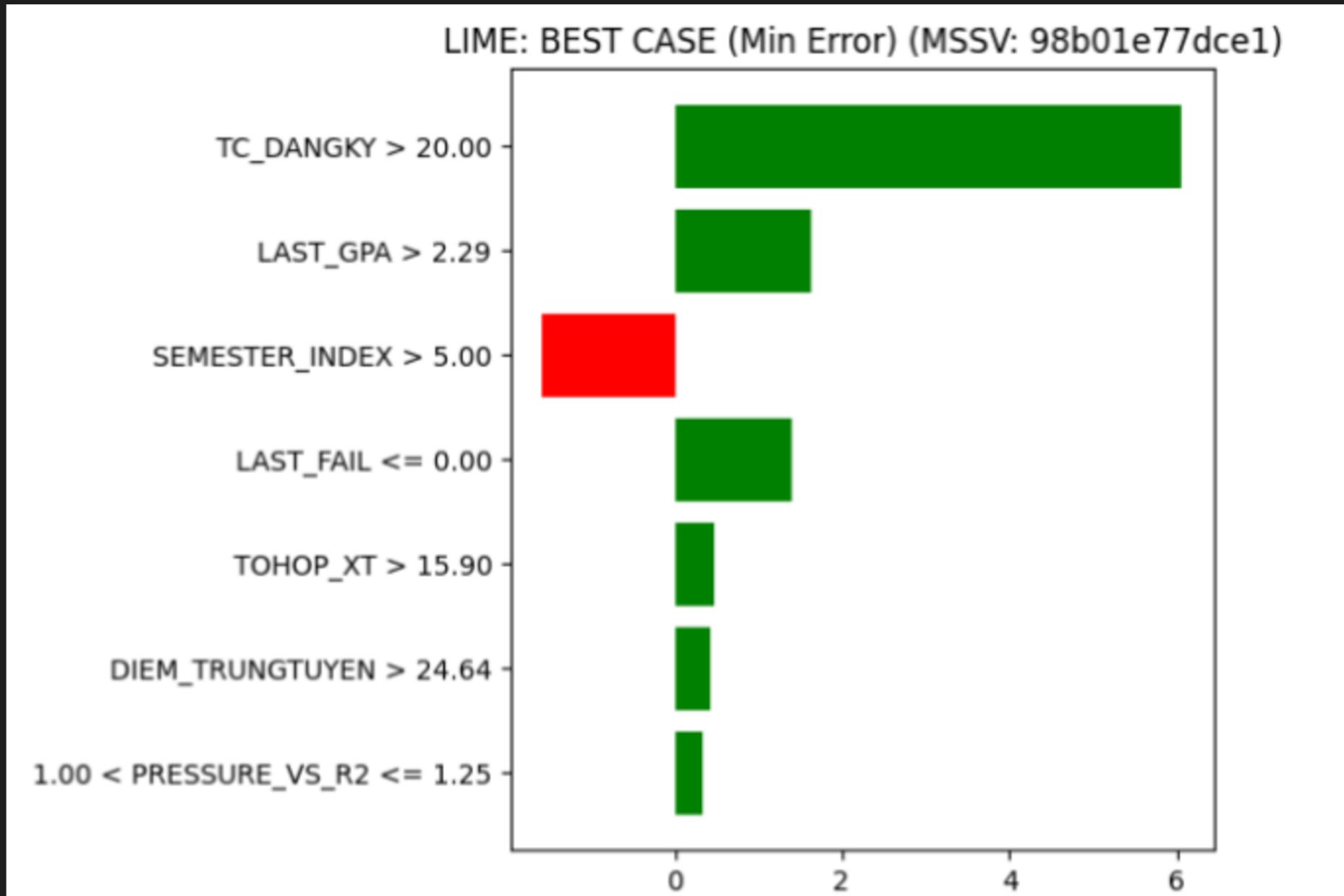
Mô hình Fresher



Mô hình Senior

- Nhóm Fresher dựa vào năng lực đầu vào và sự thích nghi ban đầu
- Nhóm Senior dựa vào sự bền bỉ và lịch sử tích lũy

KẾT QUẢ THỰC NGHIỆM



PHÂN TÍCH CỤC BỘ : LIME

Mô hình không chỉ đạt độ chính xác cao về mặt thống kê mà còn học được các quy luật nhân quả đúng đắn từ dữ liệu thực tế.

KẾT QUẢ THỰC NGHIỆM

PHÂN TÍCH PHẢN CHỨNG: DiCE

	Trạng thái hiện tại	Kịch bản 1	Kịch bản 2
Tỉ lệ hoàn thành	0.48	0.78	0.85
LAST_PASS_RATIO	0.46	0.65	0.80
R2_PASS_RATE	0.16	0.82	0.73
FAIL_TREND_R2	-3.00		
TOTAL_EARNED	100.0	38.8	58.2

Bảng gợi ý cải thiện cho sinh viên bc6bd14ea87e

Kết quả từ DiCE cung cấp cơ sở định lượng để cô vấn học tập đặt mục tiêu cụ thể cho sinh viên

KẾT LUẬN

Lợi thế

Khung giải thích đa tầng (SHAP, LIME, DiCE) giúp chuyển hóa các dự báo kỹ thuật thành các khuyến nghị hành động

Kết quả

LightGBM sử dụng hàm mắt mát Tweedie đạt hiệu năng vượt trội (**RMSE = 3.5040**)

Mở rộng

Tập trung mở rộng nguồn dữ liệu đa chiều (như mức độ chuyên cần, hoạt động tương tác) và thử nghiệm các kiến trúc Deep learning khác

Hạn chế

Sự phụ thuộc vào dữ liệu kết quả học tập mang tính thời điểm, chưa phản ánh được các biến động hành vi liên tục trong quá trình học

Thank You

Our github:

