DATAFLOW 2026

LEARNING PROGRESS PREDICTION

GROUP'S NAME: MULTOUR

THÀNH VIÊN

- Nguyễn Việt Anh
- Nguyễn Thị Hà
- Cao Chí Cường
- Hoàng Thị Thanh Nhàn

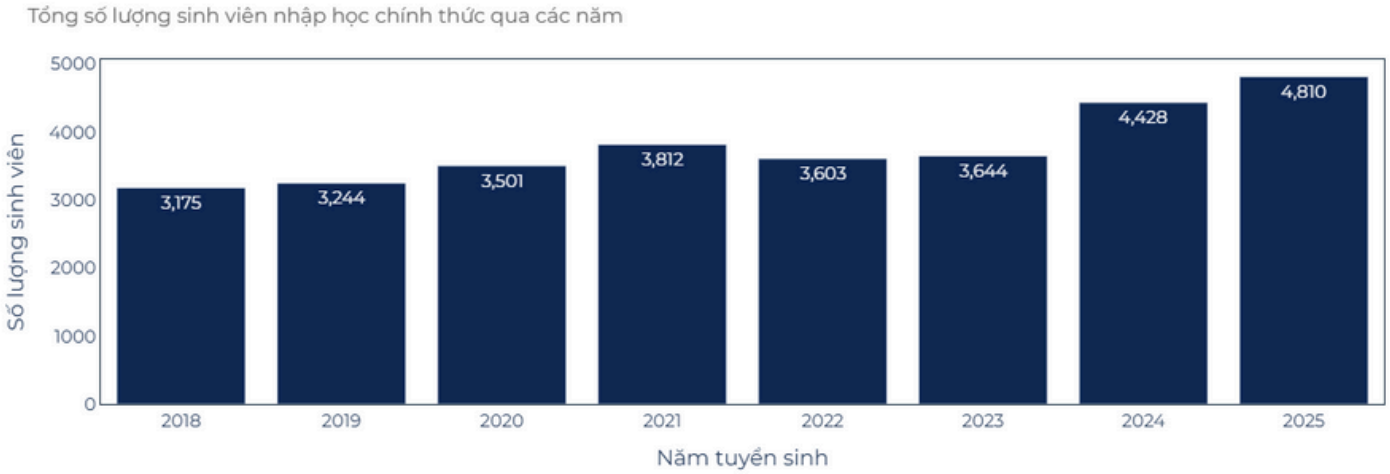
I. Giới thiệu bài toán

Xuất phát từ thực trạng sinh viên Đại học U gặp khó khăn trong việc đảm bảo tiến độ học tập, đề tài tập trung giải quyết bài toán **dự báo số tín chỉ thực tế mà sinh viên sẽ hoàn thành** dựa trên năng lực đầu vào và lịch sử học tập. Mục tiêu cốt lõi là phát hiện sớm những trường hợp có nguy cơ rủi ro ngay từ đầu kỳ, từ đó cung cấp cơ sở tin cậy để đội ngũ Cố vấn học tập và Nhà trường chủ động đưa ra các phương án hỗ trợ kịp thời. Giải pháp không chỉ yêu cầu độ chính xác cao về mặt kỹ thuật mà còn cần đảm bảo tính minh bạch (Explainable AI) để phục vụ hiệu quả cho công tác ra quyết định và xây dựng chính sách giáo dục.

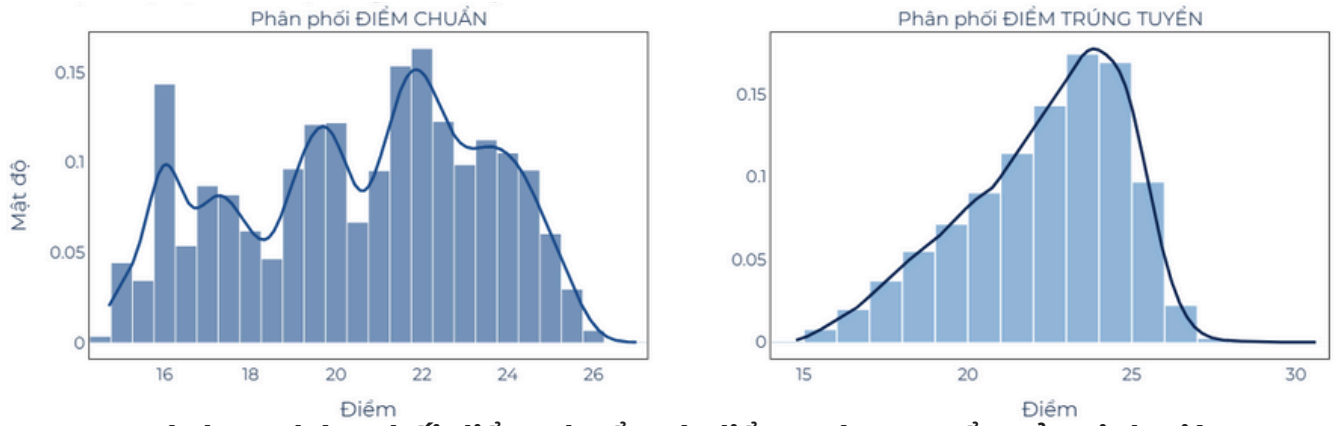
II. Phân tích dữ liệu

TỔNG QUAN

- Bảng **admission.csv**: 30217 bản ghi (mỗi bản ghi tương ứng với thông tin tuyển sinh của 1 sinh viên)
- Bảng **academic_records.csv**: 105726 bản ghi, ghi lại quá trình học tập của sinh viên từ HK1 (2020-2021) đến HK2 (2023-2024) (mỗi bản ghi tương ứng với kết quả của 1 sinh viên trong 1 học kỳ)

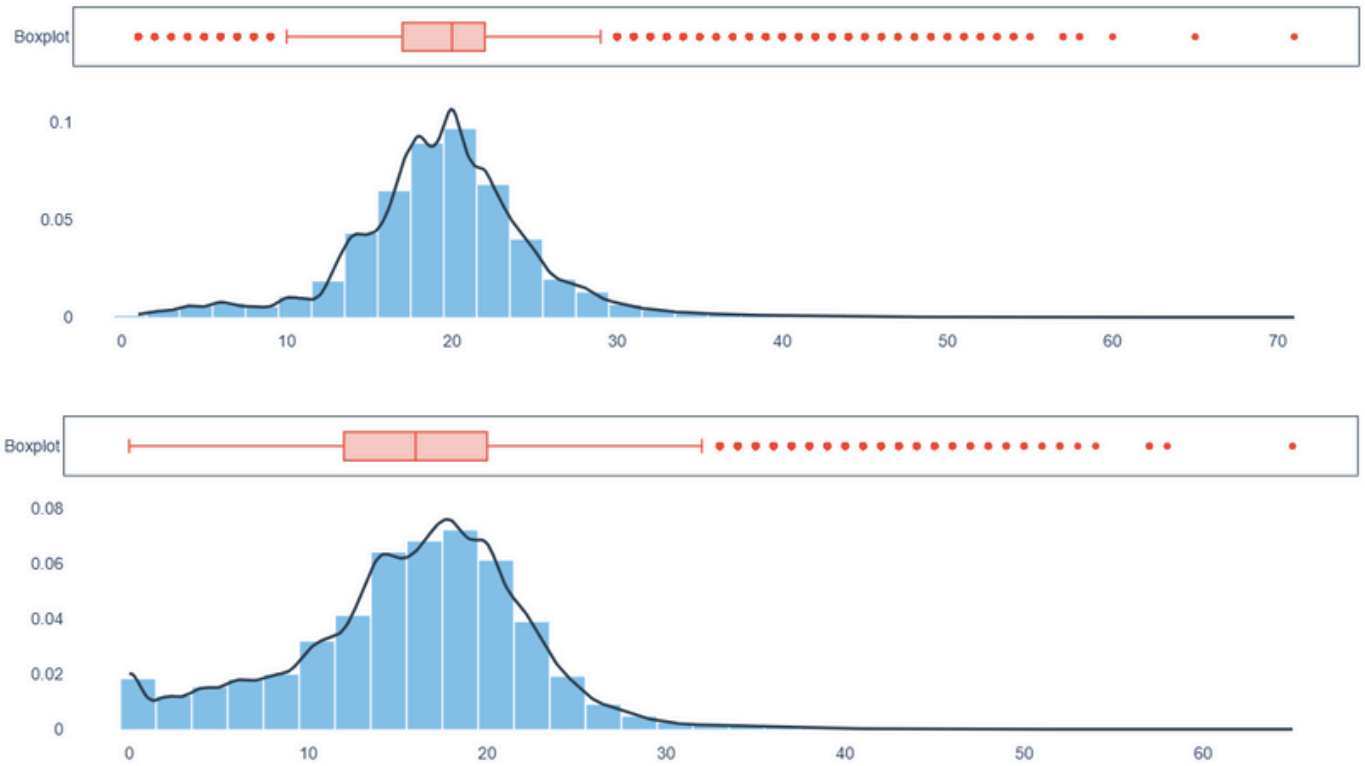


Hình 1. Phân phối sinh viên theo năm tuyển sinh



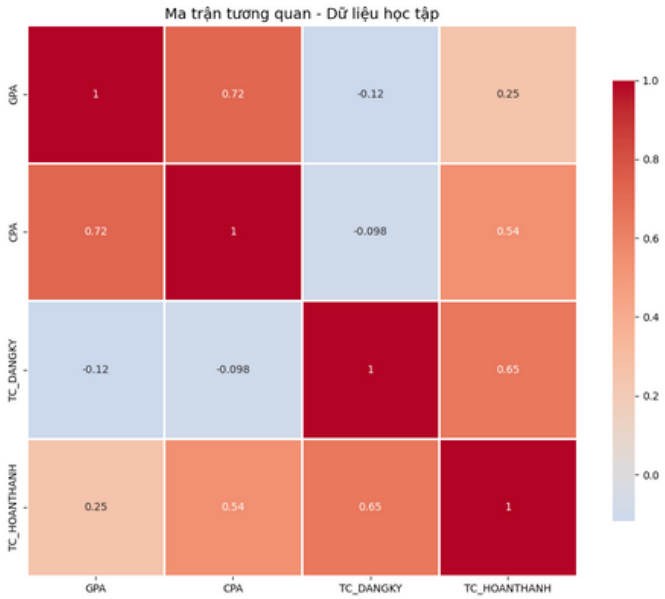
Hình 2. Phân phối điểm chuẩn và điểm trúng tuyển của sinh viên

Điểm chuẩn có độ phân tán lớn và xuất hiện nhiều đỉnh, phản ánh sự phân hóa về điểm đầu vào giữa các ngành, khoa trong trường. Điểm trúng tuyển tập trung hơn và đạt đỉnh ở mức 23-25 điểm, cho thấy chất lượng đầu vào của sinh viên trường U so với các trường đại học khác ở mức khá cao



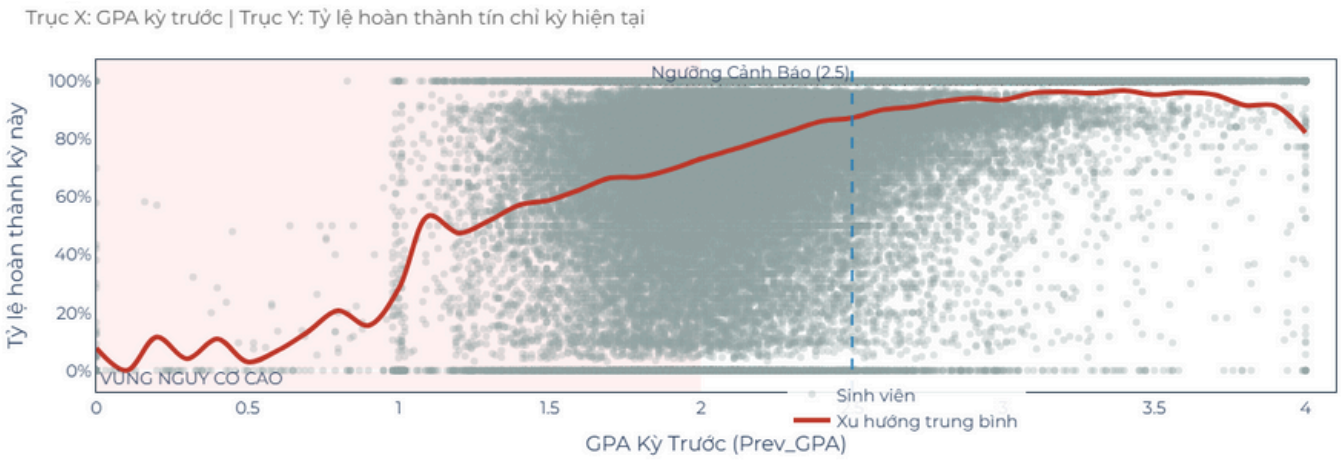
Hình 3. Phân phối tín chỉ đăng kí và tín chỉ hoàn thành

- Sinh viên tập trung đăng ký trong khoảng 15 – 22 tín chỉ/học kỳ
- Đa số sinh viên có khả năng hoàn thành 100% số tín chỉ đăng ký trong các học kỳ bình thường
- Xuất hiện các trường hợp đăng ký đột biến (> 25 - 30 tín chỉ). Ở nhóm này, tỷ lệ hoàn thành có xu hướng sụt giảm rõ rệt, cho thấy sự mất cân bằng giữa kỳ vọng và năng lực thực tế.
- Khoảng cách đăng ký - hoàn thành nới rộng hơn ở các học kỳ cuối hoặc các học kỳ có GPA thấp.



Hình 4. Ma trận tương quan giữa các biến

- GPA và CPA có mối tương quan mạnh (r = 0,72), phản ánh sự nhất quán trong kết quả học tập.
- TC hoàn thành tương quan khá cao với TC đăng ký (r = 0,65) và CPA (r = 0,54), cho thấy sinh viên đăng ký và tích lũy nhiều tín chỉ thường đạt kết quả học tập tốt hơn.
- Ngược lại, GPA/CPA gần như không tương quan với TC đăng ký (hệ số nhỏ, âm), cho thấy số tín chỉ đăng ký không phản ánh trực tiếp chất lượng điểm số.
- Những sinh viên có CPA (điểm tích lũy) ổn định thường có số tín chỉ hoàn thành bằng đúng số tín chỉ đăng ký, cho thấy khả năng tự đánh giá năng lực bản thân tốt.



Hình 5. Mối quan hệ giữa GPA kỳ trước và tỷ lệ hoàn thành tín chỉ

- Sinh viên có GPA thấp (đặc biệt < 1.0) có nguy cơ hoàn thành tín chỉ thấp
- Từ ngưỡng cảnh báo 2.5 trở lên, tỷ lệ hoàn thành tăng cao và ổn định
  → Kết quả học tập kém ở kỳ trước làm gia tăng rủi ro học tập ở kỳ sau.

III. Phương pháp xử lí

1. Tiền xử lí dữ liệu

- Loại bỏ dữ liệu thiếu và trùng lặp
- Chuẩn hóa các cột về đúng dạng dữ liệu
- Xử lí các dữ liệu bị lỗi về logic
  - điểm GPA, CPA > 4.0
  - TC_DANGKY = 0
  - DIEM_TRUNGTUYEN < DIEM_CHUAN
  - TC_HOANTHANH > TC_DANGKY

2. Feature Engineering

NHÓM 01. LAG FEATURE

Tạo các biến lag(1) để tránh bị data leakage cho các biến liên quan tới target (GPA, CPA, TC_DANGKY, TC_HOANTHANH)

$$prevX = X_{t-1}$$

NHÓM 02. ADMISSION FEATURE

Tên biến	Mô tả ý nghĩa	Công thức / Logic tính toán
`diem_vuot_chuan`	Chênh lệch giữa tổng điểm trúng tuyển thực tế của sinh viên và điểm chuẩn đầu vào của ngành. Biến này phản ánh mức độ vượt trội về năng lực đầu vào so với mặt bằng chung.	DIEM_TRUNGTUYEN − DIEM_CHUAN
`nam_tuoi`	Khoảng thời gian (số năm) tính từ năm sinh viên bắt đầu nhập học đến năm dự báo (2026).	2026 − NAM_TUYENSINH
`semester_number`	Số thứ tự của học kỳ hiện tại đối với từng sinh viên. Biến này xác định sinh viên đang ở giai đoạn nào trong lộ trình đào tạo.	$\text{Index}_{\text{kỳ học}} + 1$

NHÓM 03. HISTORY FEATURES

Tên biến	Mô tả ý nghĩa	Công thức / Logic tính toán
`prev_gpa_cpa_diff`	Đánh giá xu hướng phong độ học tập hiện tại so với trung bình tích lũy. Giá trị dương cho thấy sự tiến bộ gần đây, giá trị âm báo hiệu sự sa sút.	$\text{Prev\_GPA} - \text{Prev\_CPA}$
`prev_completion_rate`	Tỷ lệ hoàn thành tín chỉ của học kỳ gần nhất. Phản ánh mức độ hiệu quả và kỷ luật trong học tập của sinh viên ở kỳ trước đó.	$\frac{\text{Prev\_TC\_HOANTHANH}}{\text{Prev\_TC\_DANGKY}}$
`load_factor`	Hệ số tải (Áp lực học tập). So sánh khối lượng đăng ký kỳ hiện tại với năng lực hoàn thành trung bình (dựa trên trung bình trượt 5 kỳ gần nhất).	$\frac{\text{TC\_DANGKY}}{\text{RollingMean}(\text{TC\_HT}, k=5)}$
`failed_last_sem`	Biến cờ (Binary) xác định sinh viên có bị rớt môn (không hoàn thành đủ số tín chỉ đăng ký) ở học kỳ trước hay không.	$\begin{cases} 1, & \text{nếu Prev\_TC\_HT} < \text{Prev\_TC\_DK} \\ 0, & \text{ngược lại} \end{cases}$

NHÓM 04. TREND FEATURES

Tên biến	Mô tả ý nghĩa	Công thức / Logic tính toán
`gpa_trend_slope`	Độ dốc xu hướng điểm (Linear Slope). Đánh giá đà tăng trưởng hoặc suy giảm của GPA trong 3 kỳ gần nhất bằng hồi quy tuyến tính.	$\text{Slope}(\text{Prev\_GPA}, k = 3)$
`gpa_volatility`	Độ biến động kết quả học tập (Standard Deviation). Đo lường sự ổn định phong độ trong 4 kỳ gần nhất; giá trị cao báo hiệu kết quả thất thường.	$\text{Std}(\text{Prev\_GPA}, k = 4)$
`total_credits_failed`	Tổng số tín chỉ rớt tích lũy. Chênh lệch giữa tổng khối lượng đăng ký và tổng tín chỉ thực tế hoàn thành từ đầu khóa đến hiện tại.	$\sum \text{Prev\_TC\_DK} - \sum \text{Prev\_TC\_HT}$
`accumulated_fail_ratio`	Tỷ lệ trượt tích lũy. Chuẩn hóa số tín chỉ rớt trên tổng khối lượng đăng ký, phản ánh mức độ nghiêm trọng của việc nợ môn.	$\frac{\text{total\_credits\_failed}}{\sum \text{Prev\_TC\_DK}}$
`credit_velocity`	Vận tốc tích lũy tín chỉ. Trung bình số tín chỉ sinh viên hoàn thành được trong mỗi học kỳ đã tham gia.	$\frac{\sum \text{Prev\_TC\_HT}}{N_{\text{semesters}}}$

NHÓM 05. RISK FEATURES

Tên biến	Mô tả ý nghĩa	Công thức / Logic tính toán
aggressive_recovery	Chỉ báo hành vi "Gỡ gạc"(Rủi ro cao). Xác định trường hợp sinh viên tăng khối lượng đăng ký ngay sau khi gặp thất bại ở học kỳ trước, tiềm ẩn nguy cơ quá tải và tiếp tục rớt môn.	$\begin{cases} 1, & \text{nếu failed\_last\_sem} \wedge (\text{TC\_DK}_t > \text{TC\_DK}_{t-1}) \\ 0, & \text{ngược lại} \end{cases}$
expected_real_credits	Tín chỉ kỳ vọng thực tế. Hiệu chỉnh khối lượng đăng ký hiện tại dựa trên tỷ lệ trượt tích lũy trong quá khứ để ước lượng năng lực hoàn thành thực tế.	$\text{TC\_DANGKY} \times (1 - \text{accumulated\_fail\_ratio})$

3. Target Transformation

Trước khi đưa vào mô hình, nhóm thực hiển chuyển đổi biến Target thành "COMPLETION_RATE" với công thức

$$\text{df['COMPLETION\_RATE']} = \frac{\text{df['TC\_HOANTHANH'])}}{\text{df['TC\_DANGKY'] + 1e-9}}$$

(ϵ) được thêm vào mẫu số để đảm bảo tính ổn định tính toán (numerical stability), tránh lỗi chia cho 0 trong trường hợp sinh viên hủy học phần hoặc dữ liệu lỗi.

Chuẩn hoá (normalization / standardization) giúp đưa các biến về cùng một thang đo để so sánh và học tốt hơn. Đưa các đặc trưng về cùng quy mô. Tránh việc biến có giá trị lớn lấn át biến có giá trị nhỏ. Giúp thuật toán học nhanh và ổn định hơn. Các thuật toán như Gradient Descent, Linear Regression, Logistic Regression, SVM, Neural Network hội tụ nhanh hơn, ít bị dao động. Tránh hiện tượng learning bị "lệch hướng".

IV. MÔ HÌNH

1. Experiment setup

Các mô hình được sử dụng trong bài là LightGBM, XGBoost, Catboost và Ensemble với phương pháp hyperparameter tuning là OPTUNA

2. Results

2.1. Tối ưu siêu tham số

Siêu tham số	Giá trị
learning_rate	0.0116
max_depth	7
subsample	0.7976
colsample_bytree	0.6741
reg_alpha (L1)	1.4838
reg_lambda (L2)	3.1691
min_child_weight	1

XGBoost

Siêu tham số	Giá trị
learning_rate	0.0236
max_depth	9
subsample	0.8652
feature_fraction	0.6481
lambda_l1 (L1 Reg)	1.7630
lambda_l2 (L2 Reg)	8.8040
min_child_samples	4

LightGBM

Siêu tham số	Giá trị
learning_rate	0.0977
depth	6
l2_leaf_reg	1.7558
min_data_in_leaf	46

CatBoost

Bảng 1: Bộ siêu tham số tối ưu của mô hình XGBoost, LightGBM và Catboost

2.2. Model Evaluation

- **TẬP VALID** : HK2 (2023 - 2024)

Mô hình	RMSE	R2	MAPE (%)
LightGBM	**3.6638**	**0.6999**	**2.5221**
XGBoost	3.6642	0.6998	2.5155
CatBoost	3.6889	0.6958	2.5409
Ensemble	3.7463	0.6862	2.5993

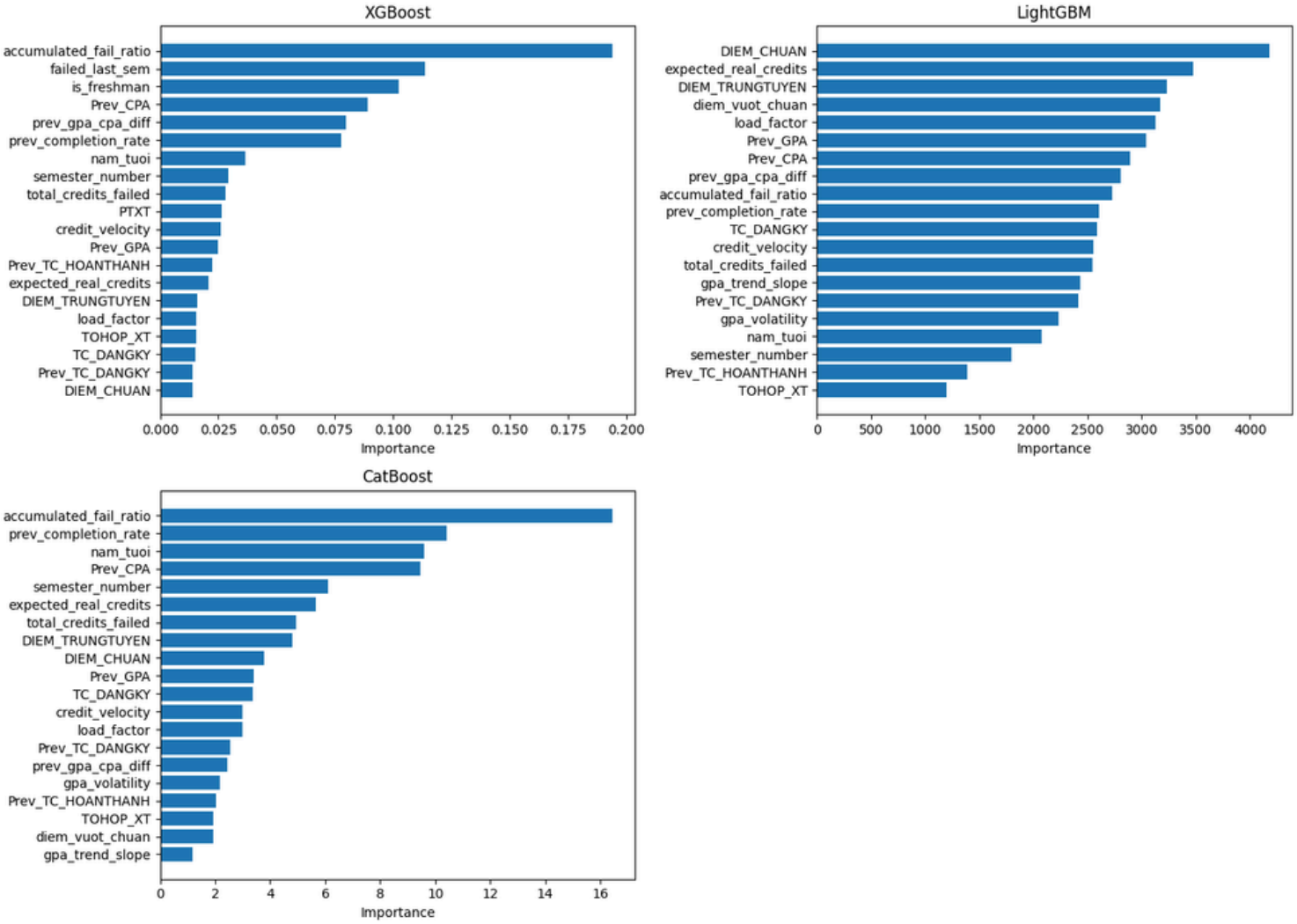
→ Mô hình tốt nhất trên tập VALID là **LightGBM**

- **TẬP TEST (40%)** : HK1 (2024 - 2025) - Kaggle

Mô hình	RMSE
LightGBM	**3.5514**
XGBoost	3.5634
CatBoost	3.5720
Ensemble	3.5556

→ Mô hình tốt nhất trên tập TEST là **LightGBM**

2.3. Feature Importance



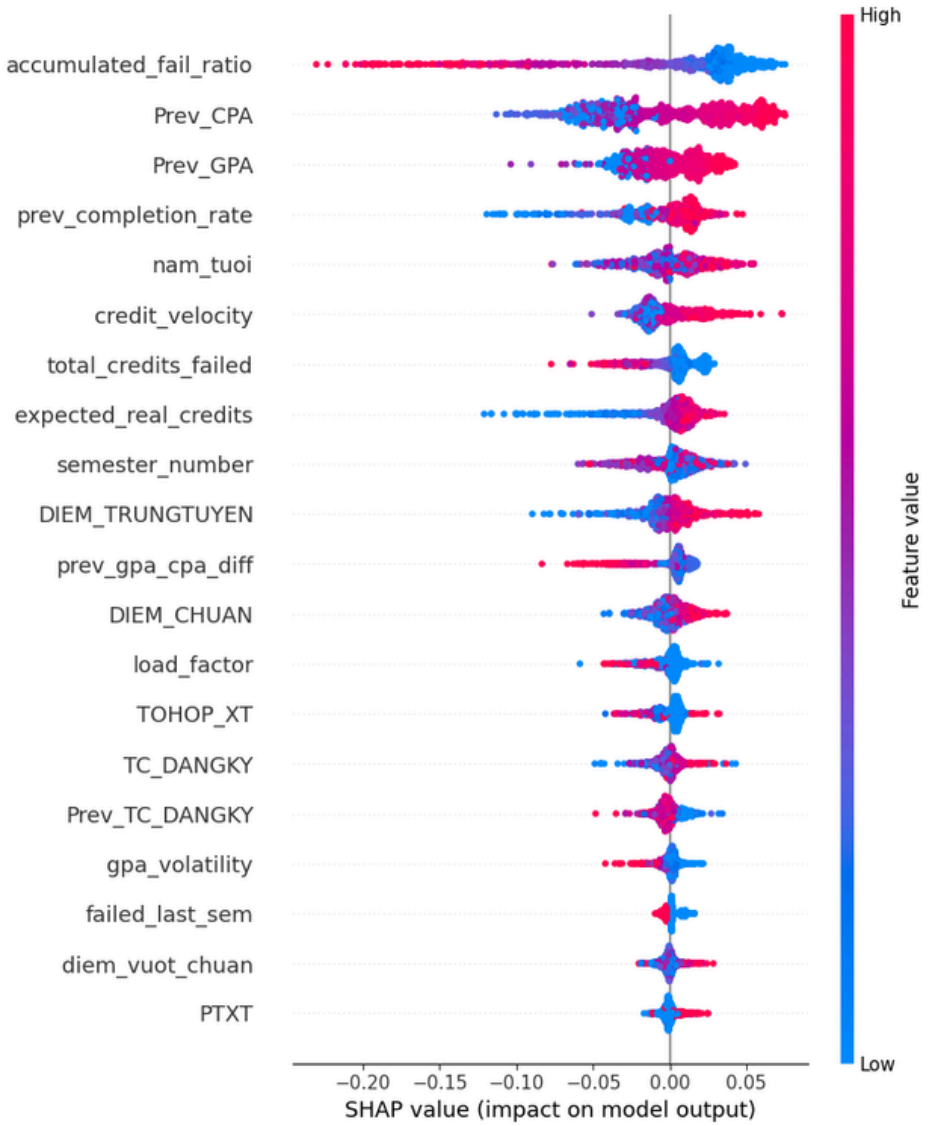
Hình 6. So sánh mức độ quan trọng của các biến trong các mô hình

- Ba mô hình (XGBoost, LightGBM, CatBoost) đều nhất quán rằng các chỉ số phản ánh **lịch sử học tập** là quan trọng nhất trong **dự đoán số tín chỉ sinh viên hoàn thành.**

- Tỷ lệ trượt tích lũy (accumulated_fail_ratio), CPA/GPA trước đó, tỷ lệ hoàn thành học phần, và các biến về kết quả học tập học kỳ trước có mức độ ảnh hưởng cao nhất..

→ Mô hình chủ yếu dựa vào hiệu quả học tập trong quá khứ để dự đoán khả năng hoàn thành tín chỉ

2.4. LIME/SHAP Analysis



Hình 7. Giải thích mô hình bằng SHAP: Ảnh hưởng của các biến đến kết quả dự báo rủi ro học tập

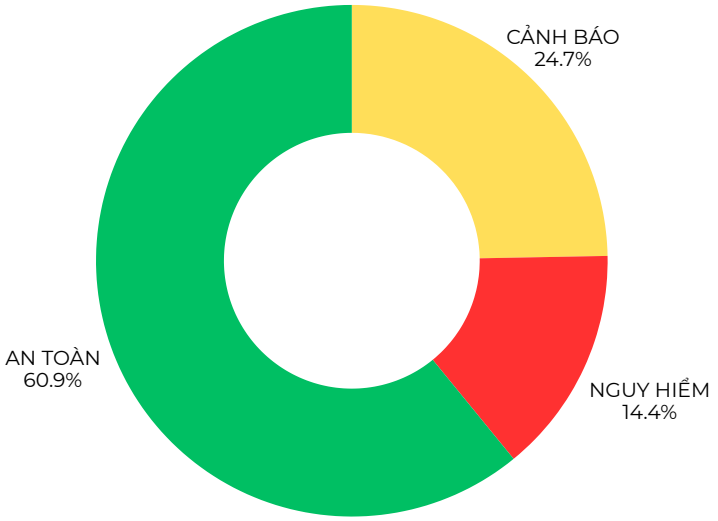
- Biến quan trọng nhất trong mô hình là **accumulated_fail_ratio:** accumulated_fail_ratio có ảnh hưởng mạnh nhất và tác động âm rõ rệt, tỷ lệ trượt tích lũy cao làm giảm đáng kể số tín chỉ dự đoán hoàn thành.
- **Prev_CPA, Prev_GPA** và **prev_completion_rate** có tác động dương: giá trị cao của các biến này làm tăng khả năng hoàn thành nhiều tín chỉ.
- **total_credits_failed** và **failed_last_sem** chủ yếu kéo dự đoán theo hướng giảm, phản ánh rủi ro học tập từ các học kỳ trước.
- **credit_velocity** và **expected_real_credits** cho thấy tác động hai chiều, hàm ý ảnh hưởng phụ thuộc vào mức độ đăng ký và khả năng đáp ứng tải học tập của sinh viên.
- Các biến như tuổi, số học kỳ, điểm thành phần có ảnh hưởng nhỏ hơn và phân bố SHAP gần 0.

V. Giải pháp

Nhóm xây dựng công thức tính tỷ lệ trượt dự báo dựa trên số tín chỉ đăng kí và số tín chỉ dự báo hoàn thành

$$\mathbf{TY\_LE\_TRUOT} = \frac{\mathbf{(TC\_DANGKY - PRED\_TCHOANTHANH)}}{\mathbf{TC\_DANGKY}}$$

NGUY HIỂM: TY_LE_TRUOT > 40 %
CẢNH BÁO: TY_LE_TRUOT in (20%, 40%)
AN TOÀN: TY_LE_TRUOT < 20 %



Hình 8. Tỷ lệ trượt dự báo

Nhà trường sẽ truy cập được vào một danh sách ghi rõ MA_SO_SV, TC_DANGKY, PHAN_LOAI như dưới đây, cột TC_GOI_Y sẽ hiện thông tin số tín chỉ đăng kí mà sinh viên nên đăng kí để đảm bảo TY_LE_TRUOT không vượt quá 20%.

MA_SO_SV	TC_DANGKY	PHAN_LOAI	TC_GOI_Y
00003e092652	20	CẢNH BÁO	18
00027b0dec4c	19	AN TOÀN	good
000e15519006	19	AN TOÀN	good
000ea6e12003	20	AN TOÀN	good
00109b845a3d	25	NGUY HIỂM	20

Bảng 2. Demo Hỗ trợ sinh viên đăng kí tín

1. Nhà trường mở "TUẦN LỄ ĐIỀU CHỈNH NGUYỆN VỌNG"

- vòng 1: sinh viên đăng kí bình thường
- vòng 2: sinh viên điều chỉnh lại số tín dựa theo mức cảnh báo của bản thân, nếu không, cần làm bản cam kết với CVHT về việc có thể hoàn thành đủ số tín đăng ký
- giáo viên nên đồng hành với sinh viên để lựa chọn mức tín chỉ phù hợp (sử dụng cột TC_GOI_Y)

2. Chính sách "Trần tín chỉ"

- Đối với sinh viên thuộc nhóm "NGUY HIỂM", phòng Quản lí Đào tạo sẽ giới hạn số tín chỉ được đăng kí của sinh viên, nếu sinh viên muốn đăng kí thêm phải nộp đơn giải trình và được CVHT phê duyệt

3. Chính sách đồng hành

- Nhà trường mở các lớp bồi dưỡng cho các bạn sinh viên thuộc nhóm "NGUY HIỂM" có mong muốn được học thêm để hoàn thành tín chỉ, người đứng lớp có thể là giáo viên hoặc các bạn sinh viên thuộc nhóm "AN TOÀN" (khuyến khích bằng việc cộng điểm rèn luyện, điểm đoàn nếu tham gia)
- Mở CLB Đồng hành để chia sẻ tài liệu, trao đổi học tập giữa các bạn trong trường

VI. REFERENCES

LINK GITHUB: Vanh41/Learning-Progress-Prediction: Dataflow 2026 Competition