DATAFLOW 2026

LEARNING PROGRESS PREDICTION

GROUP’S NAME: MULTOUR

MỤC LỤC

GIỚI THIỆU BÀI TOÁN

THÁCH THỨC:

- **Bối cảnh**: Đại học U là cơ sở giáo dục hàng đầu, khối lượng kiến thức lớn.
- **Vấn đề**: Tồn tại khoảng cách lớn giữa Số tín chỉ đăng ký và Số tín chỉ hoàn thành.
- **Hậu quả**: Sinh viên bị chậm tiến độ, dính cảnh báo học vụ, thậm chí buộc thôi học → Cần sự can thiệp sớm.

NHIỆM VỤ

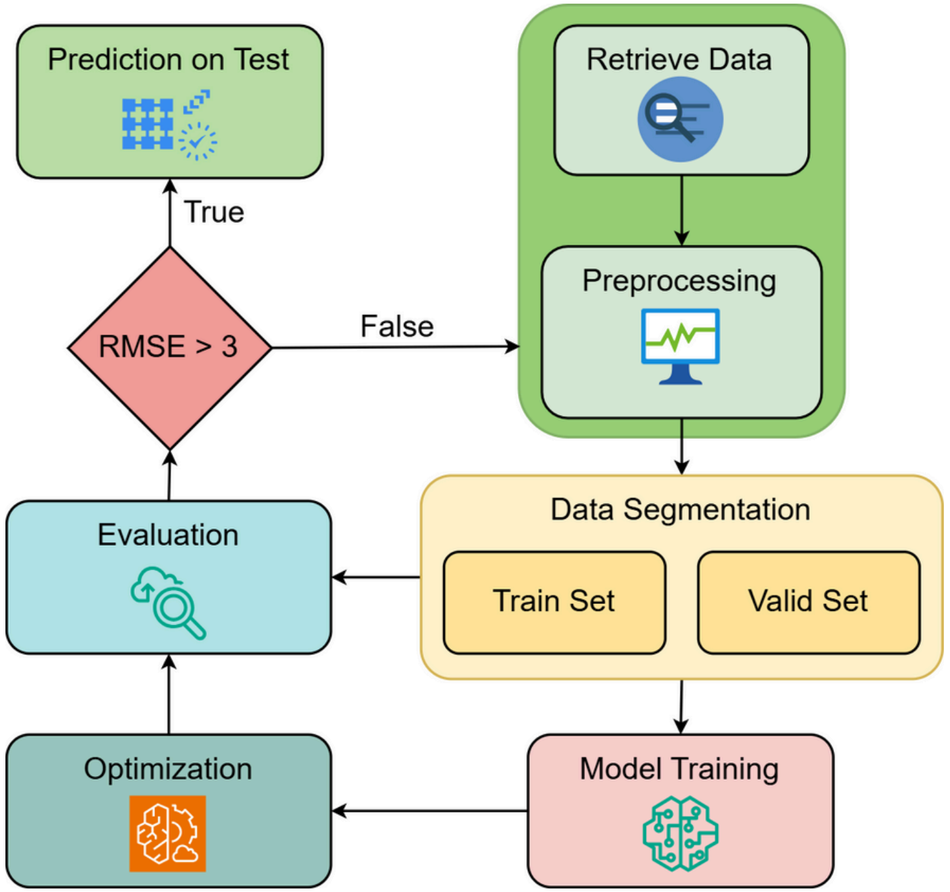
- Dự đoán số tín chỉ mỗi sinh viên đạt được sau khi kết thúc một kì học

YÊU CẦU

- Phát hiện nhóm sinh viên có nguy cơ không hoàn thành kế hoạch học tập ngay từ đầu kỳ.
- Hỗ trợ đội ngũ Cố vấn học tập và phòng đào tạo trong công tác điều phối, hỗ trợ sinh viên.
- Đề xuất các chính sách hỗ trợ học tập phù hợp.

QUY TRÌNH XỬ LÍ

- Bắt đầu bằng việc thu thập dữ liệu
- Thực hiện tiền xử lý nhằm làm sạch, chuẩn hóa và chuẩn bị dữ liệu đầu vào.
- Sau đó, dữ liệu được chia thành tập huấn luyện và tập kiểm tra.
- Mô hình được huấn luyện trên tập huấn luyện và tối ưu hóa các tham số.
- Kết quả mô hình được đánh giá bằng chỉ số RMSE. Nếu RMSE quá nhỏ hoặc lớn hơn ngưỡng cho phép, mô hình được điều chỉnh và huấn luyện lại. Khi RMSE đạt yêu cầu, mô hình được sử dụng để dự đoán trên tập dữ liệu kiểm tra.



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)

CHIA TẬP TRAIN/VALID/TEST

- **TRAIN** : HK1 (2020-2021) đến HK1 (2023-2024)
- **VALID** : HK2 (2023 - 2024)
- **TEST** : HK1 (2024-2025)

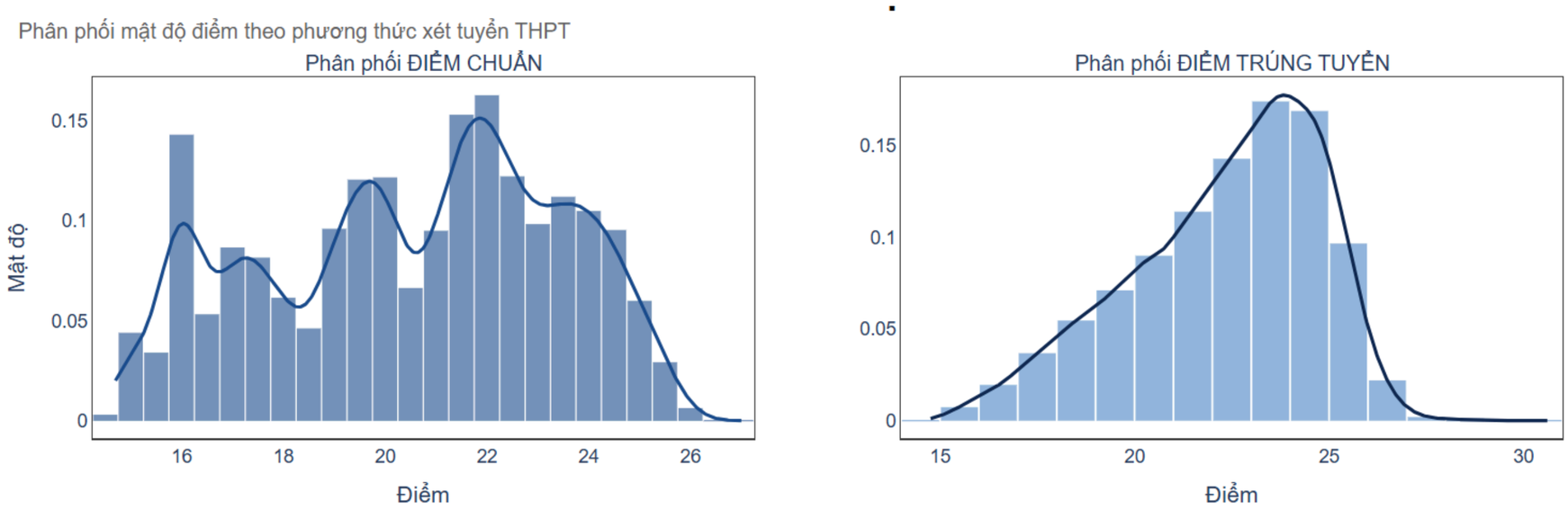
admission

MA_SO_SV	TOHOP_XT
NAM_TUYENSINH	DIEM_TRUNGTUYEN
PTXT	DIEM_CHUAN

academic_records

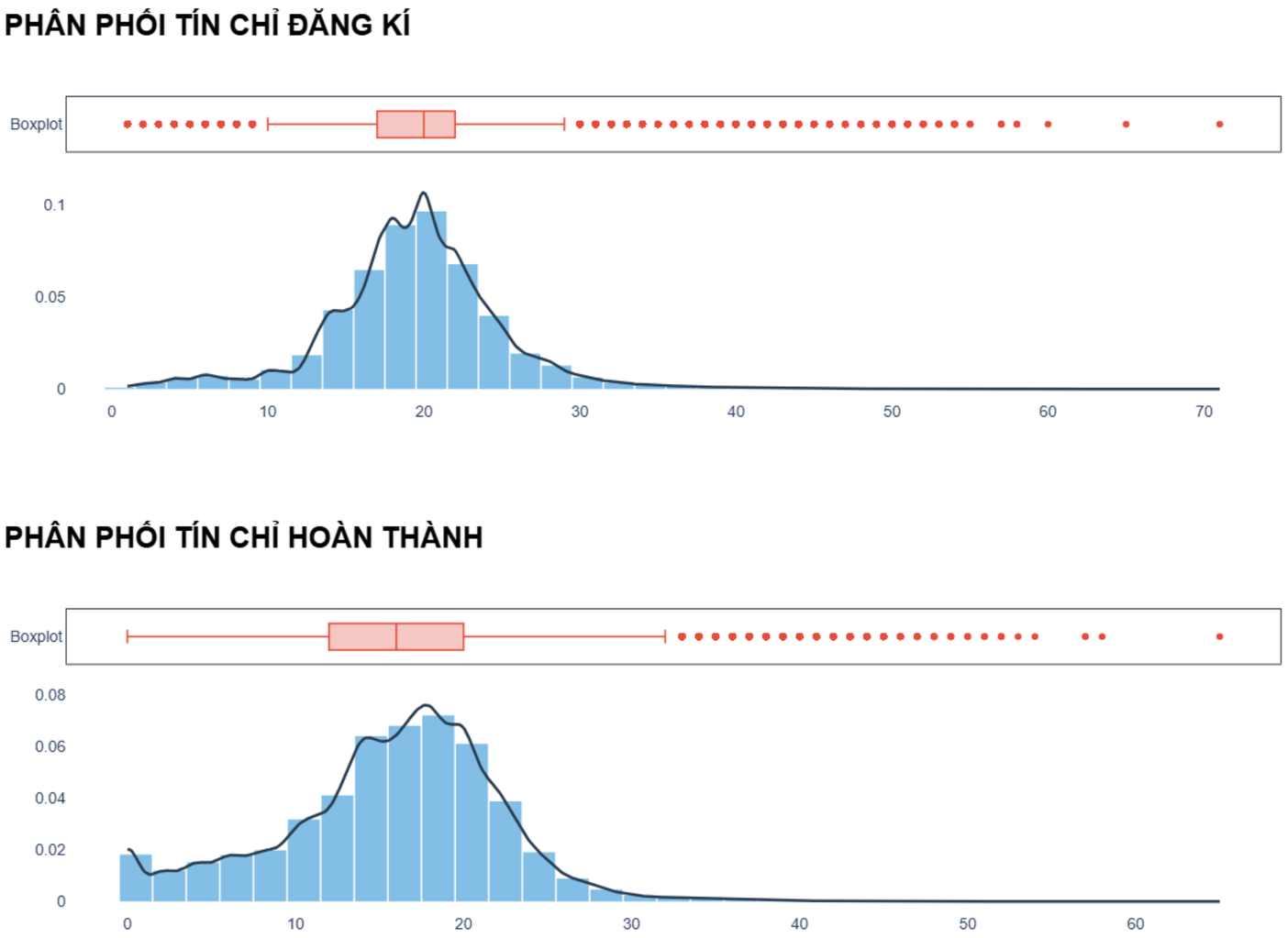
MA_SO_SV	GPA
HOC_KY	TC_DANGKY
CPA	TC_HOANTHANH

THỐNG KÊ MÔ TẢ



Hình 1. So sánh điểm chuẩn và điểm trúng tuyển thực tế

- Điểm trúng tuyển thực tế dịch chuyển sang phải và tập trung hơn so với điểm chuẩn, phản ánh mức điểm trúng tuyển nhìn chung cao hơn và ổn định hơn
- Điểm chuẩn có độ phân tán lớn và xuất hiện nhiều đỉnh, cho thấy sự khác biệt rõ rệt giữa các ngành/trường trong ngưỡng điểm xét tuyển.

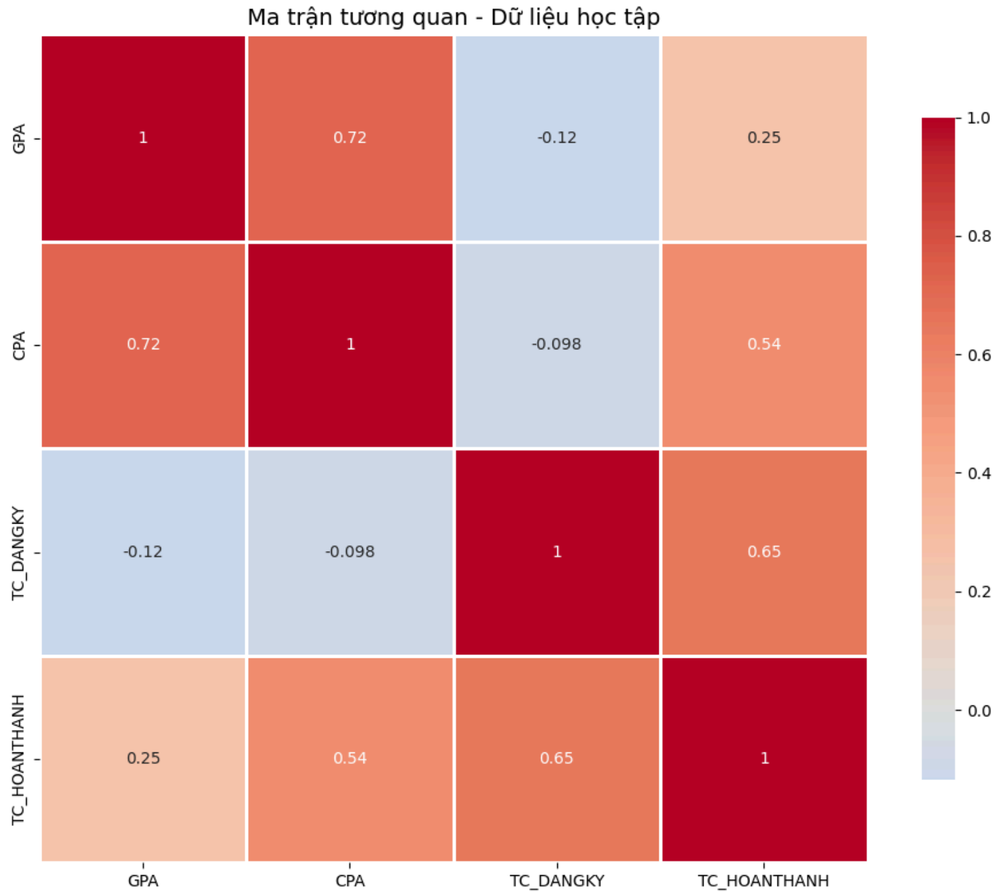


Hình 2. Phân phối tín chỉ đăng kí/hoàn thành.

- 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.

TƯƠNG QUAN

- Có sự tương quan **thuận** giữa GPA học kỳ trước và khả năng hoàn thành tín chỉ kỳ sau.
- 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 3. Ma trận tương quan giữa các biến

LÀM SẠCH DỮ LIỆU

B1: CHUẨN HÓA

- Đồng bộ dữ liệu: Chuyển MA_SO_SV về dạng chuỗi (String) để đảm bảo tính nhất quán khi tham chiếu.
- Xử lý kiểu số: Ép kiểu (Casting) an toàn cho các cột điểm số (Float) và tín chỉ (Int), xử lý NaN thành 0 ở các cột tín chỉ.
- Sắp xếp lại data bằng cách sort theo (MA_SO_SV, HOC_KY)

B2: KIỂM TRA LOGIC & LÀM SẠCH NHIỄU

- TC_HOANTHANH $\leqslant$ TC_DANGKY.
  - *Xử lý:* Dùng hàm min() để sửa lỗi nhập liệu (nếu có).
- Giới hạn GPA, CPA trong khoảng [0.0, 4.0].
  - *Xử lý:* Clipping dữ liệu (cắt bỏ giá trị ngoại lai vô lý).
- DIEM_TRUNGTUYEN $\not\geqslant$ DIEM_CHUAN.
  - *Xử lý:* Loại bỏ các dòng dữ liệu mâu thuẫn (Anomaly Detection).
- Lọc nhiễu: Loại bỏ các kỳ học "ảo" có TC_DANGKY = 0.

FEATURE ENGINEERING (1)

Tên biến	Mô tả
df['Prev_GPA'] df['Prev_CPA'] df['Prev_TC_HOANTHANH'] df['Prev_TC_DANGKY']	Dùng giá trị của quá khứ (t-1) làm đầu vào cho hiện tại (t) để đảm bảo không xảy ra Data Leakage $prev X = X_{t-1}$
df[prev_gpa_cpa_diff]	So sánh phong độ kỳ trước với năng lực tích lũy (CPA).
df[prev_completion_rate]	Tỷ lệ hoàn thành tín chỉ của kỳ trước
df[credit_velocity]	Tốc độ tính lũy tín chỉ trung bình $\frac{\sum_{i=1}^{t-1} TC\_HoanThanh_i}{\text{Semester Number}}$

FEATURE ENGINEERING (2)

Tên biến	Mô tả
df[accumulated_fail_ratio]	Tỷ lệ nợ môn tích lũy toàn lịch sử.
df[gpa_volatility]	Độ lệch chuẩn (đo sự thay đổi GPA).
df[aggressive_recovery]	Cờ rủi ro cao: Rớt môn kỳ trước nhưng kỳ này đăng ký nhiều hơn (Học gỡ). $(Fail_{t-1} > 0) \wedge (TC_t > TC_{t-1})$
df[load_factor]	Áp lực tải: So sánh số tín chỉ đăng ký kỳ này với sức học trung bình.

XÂY DỰNG MÔ HÌNH

MODEL XGBOOST

- max_depth = [3, 10]
- learning_rate = [0.005, 0.05]

MODEL LIGHTGBM

- max_depth = [3, 10]
- learning_rate = [0.005, 0.05]

MODEL ENSEMBLE

Layer 1: XBG, LGBM, CatBoost

Layer 2: Ridge

MODEL CATBOOST

- max_depth = [3, 10]
- learning_rate = [0.005, 0.05]

BEST MODEL

- 'learning_rate': 0.0236,
- 'max_depth': 9,
- 'subsample': 0.8652,
- 'feature_fraction': 0.6481,
- 'lambda_l1': 1.7629,
- 'lambda_l2': 8.8039,
- 'min_child_samples': 4

XÂY DỰNG MÔ HÌNH - KẾT QUẢ

- **TẬP VALID** : HK2 (2023 - 2024)

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

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

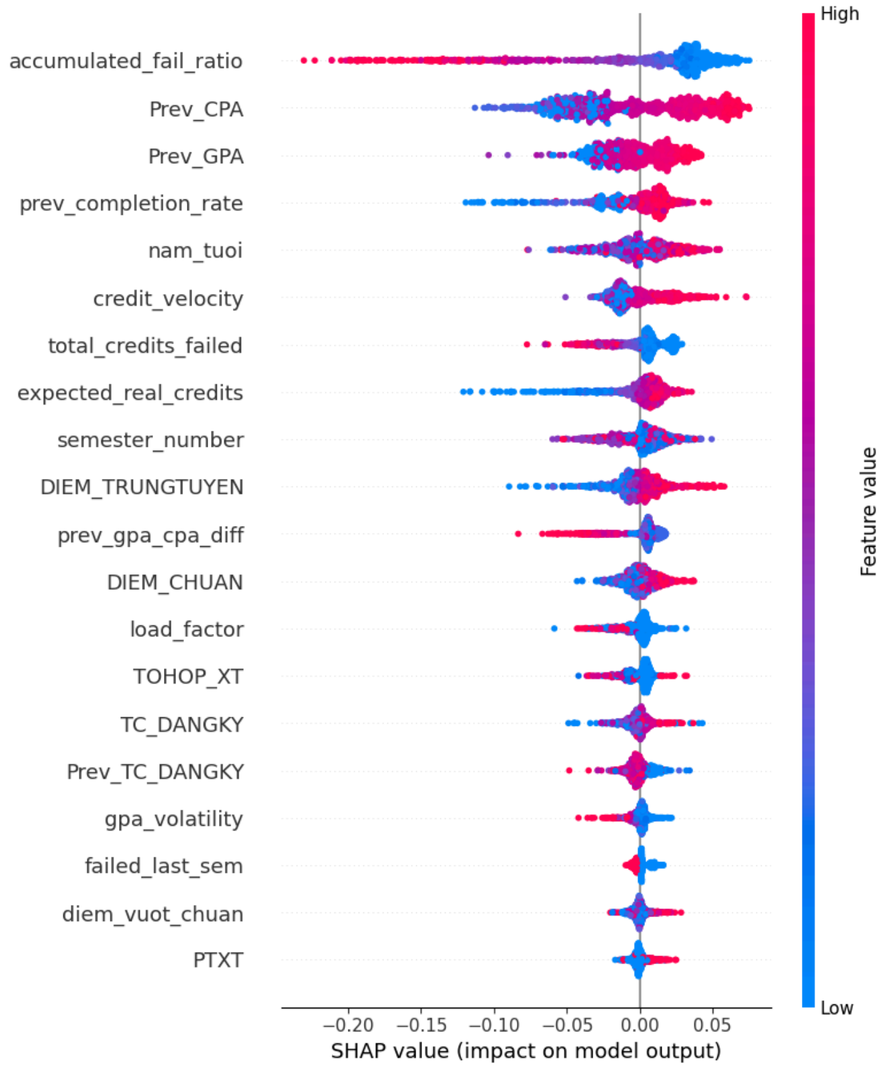
- **TẬP TEST** : HK1 (2024 - 2025) - Kaggle

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

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

KẾT QUẢ MÔ HÌNH

- 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.

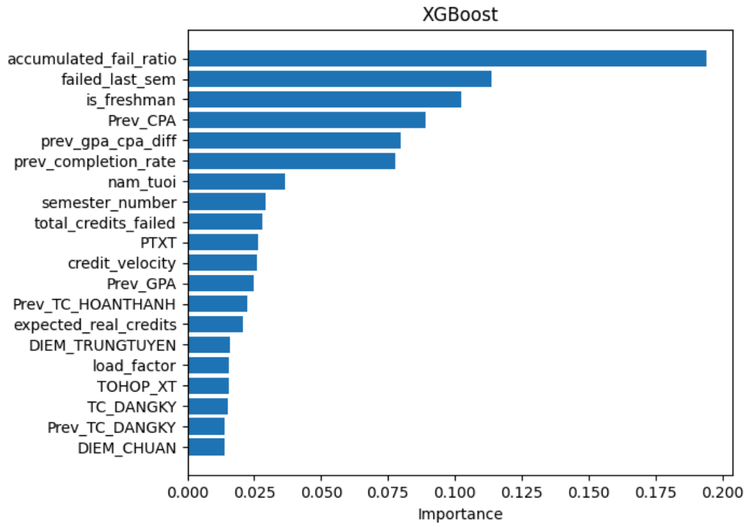


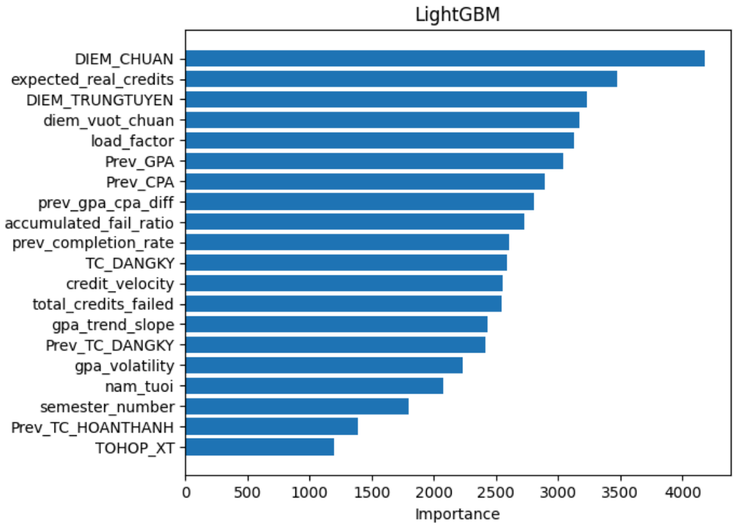
KẾT QUẢ MÔ HÌNH

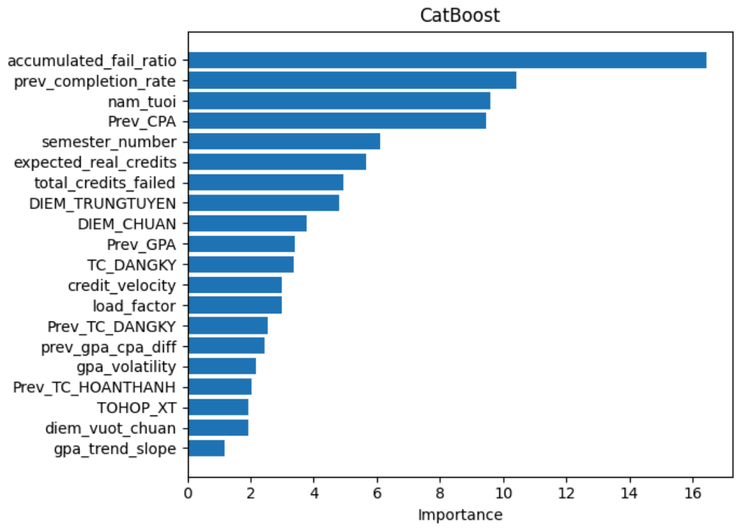
- Ba mô hình (XGBoost, LightGBM, CatBoost) đều nhất quán rằng các chỉ số phản ánh lịch sử học tập là quan trọng nhất trong dự đoán số tín chỉ sinh viên hoàn thành.

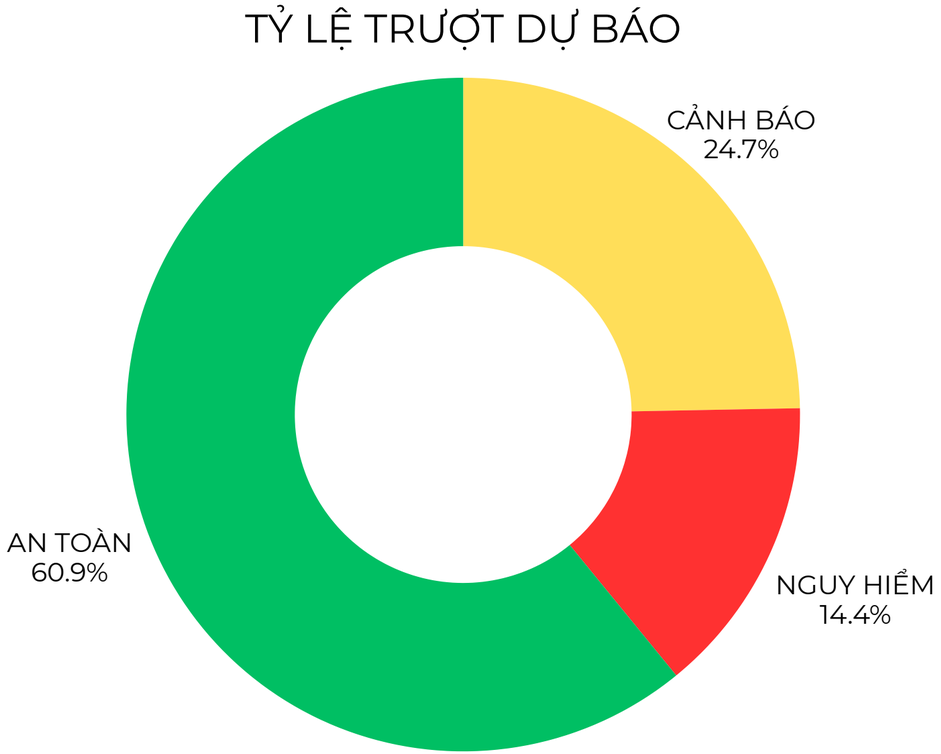
- Tỷ lệ trượt tích lũy (**accumulated_fail_ratio**), CPA/GPA trước đó, tỷ lệ hoàn thành học phần, và các biến về kết quả học tập học kỳ trước có mức độ ảnh hưởng cao nhất..

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









PHÂN NHÓM RỦI RO

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

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

DASHBOARD TƯƠNG TÁC

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

SỐ TÍN CHỈ GỢI Ý được tính để đảm bảo tỉ lệ trượt giảm xuống dưới 20% → mức ít rủi ro để sinh viên có thể hoàn thành được

GIẢI PHÁP

01. 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

02. 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

03. 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