Learning Progress Prediction - Dự báo Tiến độ Học tập

The GroundUp

Hà Nội 2026

Mục lục

1 Giới thiệu bài toán

Trong môi trường đại học, việc sinh viên không theo kịp tiến độ học tập (không hoàn thành đủ tín chỉ) là một bài toán quản trị nan giải. Đại học U đang tìm kiếm một giải pháp dựa trên dữ liệu để chuyển từ thế bị động (xử lý khi sinh viên đã bị cảnh báo) sang chủ động (hỗ trợ ngay khi có dấu hiệu rủi ro). Với vai trò là một chuyên gia dữ liệu, nhiệm vụ của bạn là xây dựng mô hình dự báo chính xác tiến độ học tập của sinh viên.

1.1 Tóm tắt bài toán

- Bối cảnh: Nhiều sinh viên gặp khó khăn trong việc hoàn thành tín chỉ, dẫn đến nguy cơ chậm tiến độ, bị cảnh báo học vụ hoặc thậm chí thôi học.
- Mục tiêu cốt lõi: Xây dựng mô hình dự báo khả năng hoàn thành lộ trình học tập của sinh viên dựa trên dữ liệu lịch sử.
- Kết quả kỳ vọng: Xác định sớm các đối tượng "nguy cơ cao"để nhà trường có biện pháp can thiệp, tư vấn kịp thời.

1.2 Phân tích bài toán (Góc nhìn Data Science)

- Loại bài toán: * Phân loại (Classification): Dự báo sinh viên có thuộc nhóm "Nguy cơ"hay "An toàn"hay không.
  - Hồi quy (Regression): Dự báo cụ thể số lượng tín chỉ sinh viên có khả năng hoàn thành trong kỳ tới.
- Các biến đầu vào tiềm năng (Features):
  - Dữ liệu học thuật: Điểm GPA các kỳ trước, số tín chỉ từng đăng ký so với số tín chỉ đã đạt, lịch sử nợ môn.
  - Dữ liệu hành vi: Tần suất tương tác trên hệ thống LMS, tỷ lệ chuyên cần.
  - Dữ liệu cá nhân: Hoàn cảnh gia đình, khu vực địa lý, phương thức xét tuyển đầu vào.

- Thách thức chính: * Xử lý dữ liệu mất cân bằng (thường số lượng sinh viên bị thôi học sẽ ít hơn nhiều so với sinh viên bình thường).
  - Đảm bảo tính giải thích được của mô hình (tại sao sinh viên đó lại bị dự báo là có nguy cơ?).

2 Exploratory Data Analysis(EDA) and Data Cleaning:

2.1 THU THẬP VÀ XỬ LÝ DỮ LIỆU ĐIỂM CHUẨN VNU (2018-2023)

1. **Tổng quan và Mục tiêu:** được thiết kế để thu thập (scrape), làm sạch (clean) và chuẩn hóa dữ liệu điểm chuẩn trúng tuyển đại học chính quy của khối ĐHQGHN trong giai đoạn 6 năm (2018-2023).
   Thách thức được giải quyết:

   - Dữ liệu nguồn phân tán trên nhiều URL khác nhau.
   - Định dạng bảng biểu (HTML Table) không đồng nhất giữa các năm (cấu trúc cột thay đổi).
   - Dữ liệu thô chứa nhiều nhiễu (kí tự lạ, lỗi nhập liệu, hợp nhất ô).
   - Thiếu thông tin định danh trường (School Identity) trong một số bảng dữ liệu tổng hợp.

2. **Kiến trúc Kỹ thuật và Thư viện:**Hệ thống sử dụng các thư viện Python chuyên dụng cho khoa học dữ liệu và khai thác web:

   (a) pandas (pd):
   - Đóng vai trò trung tâm trong việc phân tích cú pháp HTML (read_html) và thao tác trên Dataframe.
   - Sử dụng engine lxml và bs4 (BeautifulSoup4) làm backend để parse HTML lỏng lẻo.

   (b) requests: Gửi HTTP Request để lấy nội dung trang web.

   (c) re (Regular Expression): Xử lý chuỗi mạnh mẽ, dùng để bóc tách số liệu, chuẩn hóa tên và nhận diện mẫu (pattern recognition).

   (d) urllib3: Dùng để vô hiệu hóa cảnh báo bảo mật (InsecureRequestWarning) khi truy cập các trang web VNU có chứng chỉ SSL không xác thực hoặc hết hạn.

   (e) io.StringIO: Tạo luồng bộ nhớ đệm để Pandas đọc chuỗi HTML như một file vật lý.

3. **Các Module Chức năng Chi tiết**
   3.1. Module Chuẩn hóa (normalize_school_name, infer_school_from_code)
   Đây là "trí tuệ"của hệ thống, giúp định danh chính xác trường đại học thành viên dựa trên dữ liệu mập mờ.

   - Logic chuẩn hóa văn bản: Loại bỏ các tiền tố dư thừa như "TRƯỜNG", "KHOA", "VIỆN", "ĐẠI HỌC"và xử lý lỗi khoảng trắng kép (double spaces) bằng Regex.
   - Logic suy luận (Inference Engine):
     - Dựa trên Mã ngành (Prefix Mapping): Hệ thống sử dụng từ điển ánh xạ (Dictionary Hash Map) để chuyển đổi 3 ký tự đầu của mã ngành sang tên viết tắt chuẩn quốc tế của trường:
       * QHI → UET (ĐH Công nghệ)

    - * QHF → ULIS (ĐH Ngoại ngữ)
    - * QHT → HUS (ĐH Khoa học Tự nhiên)
    - * ... và các mã khác (QHE, QHL, QHS, QHY, v.v.).
  - Dựa trên Quy tắc nghiệp vụ (Heuristics):
    - * Mã bắt đầu bằng CN → UET.
    - * Mã bắt đầu bằng GD hoặc tên chứa "Sư phạm"→ UEd (ĐH Giáo dục).
    - * Tên chứa "Y khoa", "Dược học"→ UMP (Khoa Y Dược).
    - * Nhận diện các khoa trực thuộc đặc thù: Khoa Luật (UoL), Quốc tế (VNU-IS), Quản trị Kinh doanh (HSB), Liên ngành (SIS), Việt Nhật (VJU).

3.2 Module Làm sạch Điểm số (clean_score)Hàm này chịu trách nhiệm chuyển đổi dữ liệu chuỗi sang số thực (float) an toàn:

- Chuyển đổi dấu phẩy (,) sang dấu chấm (.) theo chuẩn số học quốc tế.
- Sử dụng Regex (+.?*) để trích xuất phần số trong các chuỗi nhiễu (ví dụ: "25.5 (tiêu chí phụ)").
- Trả về None nếu dữ liệu không hợp lệ.

3.3. Module Thu thập Chính (vnu_master_scraper)
Đây là hàm thực thi chính, chứa vòng lặp xử lý qua danh sách các URL nguồn (sources).

A. **Chiến lược Tải trang**

- Giả lập User-Agent của trình duyệt Chrome để tránh bị chặn bởi tường lửa: Mozilla/5.0....
- Thiết lập verify=False để bỏ qua lỗi SSL (thường gặp ở cổng thông tin giáo dục).
- Timeout thiết lập 30s để tránh treo tiến trình.

B. **Chiến lược Phân tích Bảng (Adaptive Parsing Logic)**
Hệ thống có khả năng tự thích nghi với 2 loại định dạng bảng chính xuất hiện trong dữ liệu lịch sử. Thuật toán quét từng dòng (row) và quyết định chiến lược xử lý:
Chiến lược 1: Định dạng Cổ điển (Giai đoạn 2018-2020)

- Đặc điểm: Cấu trúc bảng mở rộng theo chiều ngang. Cột 1 chứa mã trường (QHI, QHF...), và các cột tiếp theo chứa các cặp giá trị [Tổ hợp] - [Điểm chuẩn] lặp lại liên tiếp.
- Thuật toán xử lý:
  - Duyệt từ cột thứ 4 đến hết bảng với bước nhảy step=2. Kiểm tra tính hợp lệ của cặp dữ liệu: Cột chẵn phải là Mã tổ hợp (Regex: 3 ký tự, 1 chữ + 2 số, hoặc D01...), Cột lẻ phải là Điểm số hợp lệ ($13 <$ điểm $<= 40$). Nếu thỏa mãn, ghi nhận một bản ghi mới. Điều này cho phép một ngành có thể sinh ra nhiều dòng dữ liệu cho các tổ hợp khác nhau.

Chiến lược 2: Định dạng Tiêu chuẩn (Giai đoạn 2021-2023)

- Đặc điểm: Mỗi dòng là một ngành duy nhất, các cột Mã ngành, Tên ngành, Điểm phân bố cố định.
- Thuật toán xác định cột tự động: Hệ thống không gán cứng chỉ số cột mà quét tiêu đề (Header) để tìm vị trí các từ khóa: 'MÃ', 'TÊN', 'TỔ HỢP', 'ĐIỂM'.
- Cơ chế sửa lỗi (Error Correction):
  * Swap Detection: Năm 2023 có hiện tượng cột Mã ngành và Tên ngành bị đảo vị trí. Hệ thống tự động phát hiện bằng cách đo độ dài chuỗi (Mã < 8 ký tự, Tên > 8 ký tự) và kiểm tra ký tự số trong chuỗi để hoán đổi lại cho đúng.
  * Name Validation: Hàm is_valid_name kiểm tra xem tên ngành có phải là mã rác hoặc tên không có ý nghĩa hay không.

C. **Logic Hậu xử lý & Lọc nhiễu (Post-processing Filtering)**
Trước khi chấp nhận một bản ghi vào tập dữ liệu cuối cùng (all_data), dữ liệu phải qua "cửa kiểm soát":

(a) Logic Thang điểm:
- Mặc định thang 30.
- Riêng trường ULIS (ĐH Ngoại Ngữ) hoặc các ngành Ngôn ngữ được gán thang 40.

(b) Bộ lọc biên (Boundary Logic):
- Loại bỏ điểm < 13 (Điểm sàn hiếm khi thấp hơn mức này, thường là dữ liệu rác).
- Loại bỏ điểm > 31 (Trừ thang 40).
- Loại bỏ các trường hợp Mã ngành là NaN, None hoặc rỗng.

D. **Quy trình Xuất dữ liệu**

(a) Tạo DataFrame: Chuyển list các object thành pandas DataFrame.

(b) Khử trùng lặp (Deduplication):
- Sử dụng tập khóa phức hợp: Key = NAM_TUYENSINH, TRUONG, MA_NGANH, TEN_NGANH, TOHOP_XT, DIEM_CHUAN.
- Điều này đảm bảo giữ lại đầy đủ các biến thể xét tuyển của cùng một ngành nhưng loại bỏ các dòng bị scraper đọc trùng lặp do lỗi bảng HTML.

(c) Đánh số thứ tự (Indexing): Tái lập cột STT chạy liên tục từ 1.

(d) Sắp xếp & Định dạng: Đảm bảo thứ tự cột đầu ra đúng yêu cầu: STT | NAM | TRUONG | TOHOP | TEN | DIEM | THANG.

(e) Lưu trữ: Xuất ra file vnu_benchmark_final.csv với encoding utf-8-sig để hiển thị tốt tiếng Việt trên Excel.

2.2 Data Loading and Exploration

2.2.1 Tải tệp CSV rồi tạo dataframe.

Những phát hiện chính

TÊN BIẾN	TỆP NGUỒN	MÔ TẢ & LOGIC KINH DOANH	KIỂU DỮ LIỆU & VÍ DỤ
MA_SO_SV	Tất cả các tệp	**Student ID (Primary Key).** Mã định danh duy nhất cho mỗi sinh viên. Dùng để liên kết (JOIN) giữa bảng Lịch sử học tập, Tuyển sinh và Test.	`Categorical (String)` Bán tại: `SV001`, `20205123`
HOC_KY	Đào tạo / Kiểm tra	**Academic Term (Time Step).** Thời điểm diễn ra dữ liệu. Cấu trúc thường là "HK[Kỳ] [Năm học]". *Lưu ý xử lý: Cần tách thành 2 cột số: Year (2023) và Term (1/2/3).*	`String (Time)` Bán tại: `HK1 2023-2024`
Điểm trung bình	Xe lửa	**Grade Point Average (Semester).** Điểm trung bình của riêng học kỳ đó (thang 4.0 hoặc 10.0). *Ý nghĩa: Phản ánh phong độ ngắn hạn (Short-term Performance).*	`Numerical (Float)` Bán tại: `3.2`, `2.8`
CPA	Xe lửa	**Cumulative Point Average.** Điểm trung bình tích lũy từ khi nhập học đến thời điểm hiện tại. *Ý nghĩa: Phản ánh năng lực dài hạn (Long-term Ability).*	`Numerical (Float)` Bán tại: `3.05`
TC_DANGKY	Đào tạo / Kiểm tra	**Registered Credits (Input).** Số tín chỉ sinh viên đăng ký vào đầu kỳ. *Logic: Đây là "trần" (Upper Bound) của kết quả. TC_HOANTHANH không bao giờ > TC_DANGKY.*	`Numerical (Int)` Bán tại: `18`, `24`
TC_HOANTHANH	Chỉ có tàu hỏa	**Completed Credits (TARGET).** Số tín chỉ thực tế sinh viên tích lũy được (đậu) sau khi kết thúc kỳ. **Nhiệm vụ:** Dự báo giá trị này cho tập Test.	`Numerical (Int)` Bán tại: `15`, `18`
NAM_TUYENSINH	Nhập học	**Enrollment Year.** Năm sinh viên nhập học. Dùng để tính "Tuổi học đường" (School Age = Current Year - Enrollment Year).	`Numerical (Int)` Bán tại: `2020`
PTXT	Nhập học	**Admission Method.** Phương thức xét tuyển (VD: Điểm thi THPT, Học bạ, Tuyển thẳng...). *Lưu ý: Cần xử lý mã hóa (Encoding) vì đây là biến phân loại.*	`Categorical (Code)` Bán tại: `1`, `301`
TOHOP_XT	Nhập học	**Exam Subject Group.** Tổ hợp môn thi đầu vào (Khối thi). *Insight: Khối A00 (Toán-Lý-Hóa) thường có tư duy logic tốt hơn khối C (Văn-Sử-Địa) trong các môn kỹ thuật.*	`Categorical (String)` Bán tại: `A00`, `D01`
DIEM_TRUNGTUYEN	Nhập học	**Entrance Score.** Tổng điểm thi đầu vào của sinh viên. *Lưu ý: Cần chuẩn hóa theo năm vì đề thi mỗi năm có độ khó khác nhau.*	`Numerical (Float)` Bán tại: `26.5`
DIEM_CHUAN	Nhập học	**Benchmark Score.** Điểm sàn của ngành học. Dùng để tính `SCORE_SURPLUS` (Điểm thi - Điểm chuẩn) để xem sinh viên "dư sức" hay "vừa đủ đậu".	`Numerical (Float)` Bán tại: `25.0`

2.2.2 Tổng quan cơ bản

Những phát hiện chính

1. Cấu trúc dữ liệu: Time-Series trên dữ liệu Bảng (Panel Data)
   - Dữ liệu không phải là các dòng độc lập. Một sinh viên (MA_SO_SV) xuất hiện nhiều lần trong bảng Train (Academic Records) qua các kỳ (HOC_KY).
   - Chiến lược: Không được shuffle dữ liệu bừa bãi khi train. Phải dùng GroupKFold theo MA_SO_SV hoặc TimeSeriesSplit để tránh Data Leakage.
2. Ràng buộc Vật lý của Biến mục tiêu
   - TC_HOANTHANH $\leq$ TC_DANGKY
   - Chiến lược: Mô hình nên dự báo Tỷ lệ (Ratio) thay vì số tuyệt đối, sau đó nhân ngược lại với TC_DANGKY. Hoặc dùng mô hình 2 bước (Hurdle Model)
3. Vấn đề "Cold Start"tiềm ẩn
   - Tập Test là HK1 2024-2025.
   - Nếu trong tập Test có sinh viên nhập học năm 2024 (tức là Freshmen), họ sẽ KHÔNG có mặt trong bảng Train (Academic Records).
   - Chiến lược: Bắt buộc phải merge bảng Admission Info. Đối với sinh viên mới, thông tin Tuyển sinh (DIEM_TRUNGTUYEN, TOHOP_XT) là "phao cứu sinh"duy nhất.

2.2.3 Kiểm tra tính hợp lý

2.2.4 Xử lý Duplicate rows

Những phát hiện chính:
Không chỉ dọn rác, mà còn học từ rác.

```
***** Train(Academic Records) Quality Checks *****
No missing values
Duplicate rows: 24
Constant columns: None

***** Competition Test Quality Checks *****
No missing values
Duplicate rows: 0
Constant columns: ['HOC_KY']

***** Admission Info Quality Checks *****
No missing values
Duplicate rows: 0
Constant columns: None

***** VNU Benchmark Quality Checks *****
```

	missing_%
TOHOP_XT	6.982

```
Duplicate rows: 0
Constant columns: None
```

1. Dấu hiệu của "Sự bất ổn định hành chính"(Administrative Instability)

   - Phát hiện: Những sinh viên bị trùng lặp dữ liệu (đặc biệt là kiểu trùng MA_SO_SV + HOC_KY nhưng khác số liệu) thường rơi vào nhóm có sự thay đổi trạng thái học tập phức tạp: Đăng ký muộn, Hủy môn, Đăng ký học lại, hoặc Phúc khảo điểm.
   - Insight: Sinh viên có dữ liệu "bẩn"này thường là những sinh viên có vấn đề về lộ trình học. Họ không tuân theo lộ trình chuẩn.

2. Lỗi hệ thống mang tính "Cộng dồn"(Aggregation Glitch)

   - Phát hiện: Trường hợp 43 tín chỉ (trong khi thực tế là 13) cho thấy hệ thống Database của trường đang bị lỗi ở khâu GROUP BY hoặc SUM. Nó cộng gộp cả những môn đã hủy (Dropped courses) hoặc cộng gộp lịch sử học lại vào học kỳ hiện tại.
   - Insight: TC_DANGKY trong dữ liệu thô không hoàn toàn đáng tin cậy 100% nếu nó vượt quá ngưỡng 30. Ta cần đặt một "Trust Gate"(Cổng tin cậy) ở mức 30-35 tín.

2.3 Nhóm tính năng (Feature Groups)

2.3.1 Nhóm tính năng:academic_records.csv

Những phát hiện chính

1. HOC_KYThể loại Kẻ giả mạo
   - Hiện tượng: Đang xếp HOC_KY vào nhóm Categorical.
   - Rủi ro: Nếu One-Hot Encoding cột này (ví dụ: HK1_2023, HK2_2023...), mô hình sẽ mất hoàn toàn tính thứ tự thời gian. Nó sẽ có khả năng không hiểu là HK2 diễn ra sau HK1.
   - Insight: HOC_KY thực chất là Time-Series (Ordinal). Cần tách nó thành YEAR (Niên khóa) và TERM (Học kỳ) để mô hình thấy được xu hướng "Trưởng thành"hoặc "Trượt dốc"của sinh viên. ***
2. Sự vắng mặt của "Hồ sơ tĩnh"(The Missing Baseline) - Admission_Info.csv
   - Hiện tượng: Feature list này hoàn toàn thiếu thông tin Tuyển sinh (Admission_Info.csv).
   - Rủi ro: Chỉ có thông tin Kết quả học mà thiếu thông tin "Tĩnh"(Năng lực gốc). Một sinh viên có GPA 2.0 nhưng đầu vào 29 điểm (Thủ khoa lười học) rất khác với sinh viên GPA 2.0 nhưng đầu vào 15 điểm (Học yếu thực sự).
   - Insight: Cần Merge bảng Admission.
3. Bẫy "Đa cộng tuyến"(Multicollinearity Trap)
   - Hiện tượng: Có cả GPA (ngắn hạn) và CPA (dài hạn) trong nhóm Numeric.
   - Rủi ro: CPA thực chất là trung bình cộng của các GPA quá khứ. Hai biến này tương quan cực mạnh.
   - Insight: Sự chênh lệch giữa GPA và CPA (GAP = GPA - CPA) mới là tín hiệu vàng.
     - GAP < 0: Sinh viên đang sa sút (Phong độ $<$ Đẳng cấp).
     - GAP > 0: Sinh viên đang tiến bộ.

Hợp nhất - Sự kết hợp chiến lược Để thực hiện một pha Merge chuẩn xác, chúng ta không chỉ dùng lệnh pd.merge một cách vô tri. Chúng ta cần một quy trình Smart Merge bao gồm 4 lớp :

1. Pre-validation: Đảm bảo bảng bên phải Admission_Info.csv là duy nhất (Unique) theo MA_SO_SV.
2. Inference (Suy luận): Nếu thiếu năm nhập học, tự động suy luận từ kỳ học đầu tiên.

2.3.2 Nhóm tính năng:train

Những phát hiện chính:

1. PTXT (Phương thức xét tuyển) là "Kẻ trà trộn"(Imposter Numeric) ***

- Vấn đề: Hiện tại đang xếp PTXT vào nhóm Numeric. Tuy nhiên, trong tuyển sinh đại học, đây là biến Phân loại (Categorical). Ví dụ: 1 là Xét tuyển thẳng, 2 là Xét điểm thi THPT, 3 là Xét học bạ.
- Nguy cơ: Nếu để nguyên dạng số, mô hình sẽ hiểu sai rằng Phương thức 3 "lớn hơn"Phương thức 1, dẫn đến học sai quy luật.
- Hành động: Cần chuyển PTXT sang dạng Categorical (One-Hot Encoding hoặc Label Encoding).

2. Kho báu ẩn giấu SCORE_SURPLUS (Dư địa năng lực) ***
   - Vấn đề: Có DIEM_TRUNGTUYEN (Điểm thực của SV) và DIEM_CHUAN (Sàn của trường).
   - Giá trị: Sự chênh lệch giữa hai điểm này SURPLUS = TRUNGTUYEN - CHUAN mới là chỉ số vàng.
     - SURPLUS cao: Sinh viên giỏi vượt trội so với mặt bằng chung (Valedictorian mindset).
     - SURPLUS $\approx$ 0: Sinh viên đậu vớt (Survival mindset), nhóm này thường có nguy cơ rớt môn cao nhất ở năm đầu.
3. Thiếu vắng yếu tố "Thời gian thực"(School Age) ***
   - Vấn đề: Có NAM_TUYENSINH nhưng lại để nó là Numeric thuần túy. Nó không có ý nghĩa nếu đứng một mình.
   - Giá trị: Phải kết hợp NAM_TUYENSINH với Năm hiện tại (từ HOC_KY) để ra SCHOOL_AGE (Tuổi nghề sinh viên). Sinh viên năm 1 (Age=1) rớt môn vì bỡ ngỡ, Sinh viên năm 4 (Age=4) rớt môn vì chán nản hoặc đi làm thêm. Hai hành vi này khác hẳn nhau.

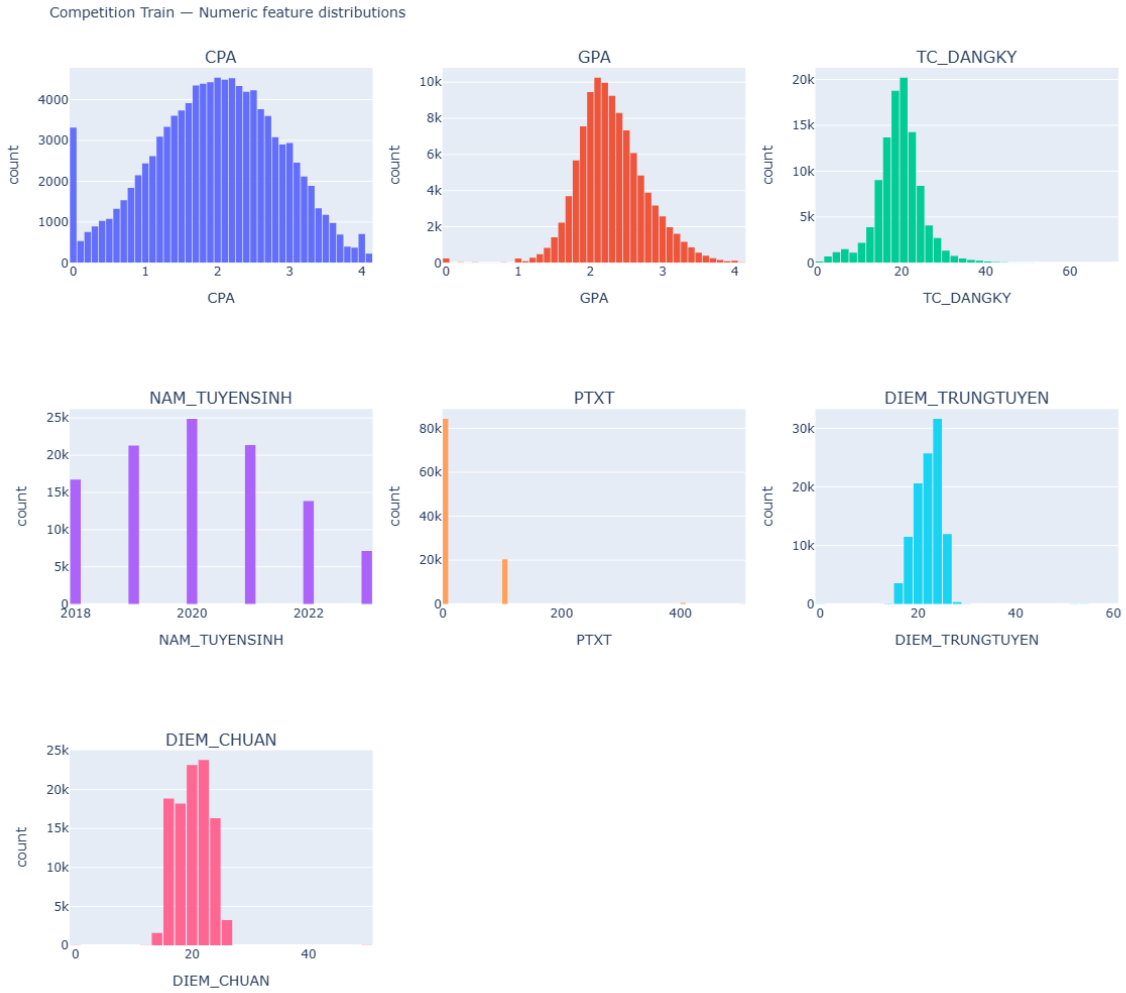
2.4 Phân tích các đặc điểm

2.4.1 Phân bố đặc trưng đơn biến (Dạng số/Dạng phân loại)

Phân bố đặc trưng đơn biến (dạng số)
Những phát hiện chính

1. "Siêu Tín Chỉ"và Lỗi Vật Lý (Physical Constraint Violation) ***
   - Dữ liệu: TC_DANGKY có Max = 71.00, trong khi 95% sinh viên chỉ đăng ký dưới 28 tín.
     - Đây là điều phi lý. Một học kỳ chuẩn chỉ có 15-20 tín. 71 tín tương đương với học... 3 năm trong 1 kỳ.
   - Có thể do lỗi cộng dồn (Group By sai) hoặc sinh viên đăng ký ảo rồi hủy hàng loạt nhưng hệ thống vẫn ghi nhận con số ban đầu.
   - Rủi ro: Mô hình Regression rất nhạy cảm với Outlier. Con số 71 này sẽ "bẻ cong"đường hồi quy, làm sai lệch dự báo cho nhóm sinh viên bình thường.

Dựa trên Quyết định 3626/QĐ-ĐHQGHN (Quy chế Đào tạo Đại học tại ĐHQGHN) và dữ liệu khác, xác định các ngưỡng xử lý như sau:



(a) Về Tín Chỉ Đăng Ký (Điều 20 - Quy chế Đào tạo):

- Quy định: Tối thiểu $\geq$2 / 3 khối lượng trung bình (khoảng 10-12 tín). Tối đa $\leq$ 3 / 2 khối lượng trung bình (khoảng 22-25 tín).
- Thực tế tại HUS: Các chương trình Tài năng/Tiên tiến có thể học nặng hơn, nhưng ngưỡng 35 tín chỉ/kỳ là giới hạn vật lý an toàn (gần gấp đôi chuẩn).
- 71 tín chỉ trong dữ liệu là LỖI (Error), không phải là biến động (Variance).
- Hành động: Áp dụng ngưỡng trên (cap_limit) ở mức 32 tín chỉ (con số an toàn cho HUS).

Giải pháp thông minh: Hoà hợp logic mỏ neo

- Mỏ neo (The Anchor): TC_HOANTHANH (Tín chỉ hoàn thành) là con số đáng tin cậy hơn (vì nó là kết quả cuối cùng sau khi thi). TC_DANGKY thường hay bị ảo (do đăng ký xong hủy, hoặc đăng ký nhầm).

  Nguyên tắc 1 (Ngưỡng trên Vật Lý): Không gì được vượt quá 35.

  Nguyên tắc 2 (Logic Bảo Toàn): TC_HOANTHANH không bao giờ được lớn hơn TC_DANGKY.

Khi xảy ra tình huống: chỉ TC_DANGKY > 35 hoặc ngược lại (TC_HOANTHANH > 35). Ví dụ: TC_DANGKY = 71, TC_HOANTHANH = 15. Tôi chia các ca lỗi (TC_DANGKY > 35) thành 3 nhóm hành vi dựa trên GPA, và xử lý khác nhau cho từng nhóm:

- **Tình huống A**: "Thiên Tài Bị Lỗi Hệ Thống"(GPA cao > 3.2): Phân tích: GPA cao nghĩa là học môn nào qua môn đó (điểm A/B+). Xác suất rớt môn cực thấp. Suy luận: Số tín chỉ đăng ký thực tế phải rất sát với số tín chỉ hoàn thành. Giải pháp: TC_DANGKY_Mới = TC_HOANTHANH. Hoặc ngược lại với trường hợp TC_HOANTHANH_Mới.
- **Tình huống B**: "Sinh Viên Trung Bình"(2.0 <= GPA <= 3.2) Phân tích: Học lực Khá. Có thể rớt 1-2 môn hoặc hủy môn nhưng không nhiều. Suy luận: Đăng ký thực tế có thể cao hơn Hoàn thành một chút (khoảng 10-20%). Giải pháp: TC_DANGKY_Mới = TC_HOANTHANH * 1.2 (nhưng không vượt trần 30).
- **Tình huống C**: "Sinh Viên Yếu Kém / Bỏ Cuộc"(GPA < 2.0) Phân tích: GPA thấp chứng tỏ rớt rất nhiều. Suy luận: Sinh viên này có thể đã đăng ký "Full tải"(25-30 tín) nhưng rớt sạch, chỉ qua được vài môn dễ. Giải pháp: TC_DANGKY_Mới = Ngưỡng Trần An Toàn (ví dụ 25 hoặc 30).

2. Sự Hỗn Loạn Thang Đo Điểm Đầu Vào (Scale Inconsistency) - Heterogeneous Scale Problem ***

   - Dữ liệu: DIEM_TRUNGTUYEN có Max = 59.06. Trong khi thang điểm thi THPT Quốc gia (3 môn) tối đa là 30 (hoặc 40 nếu nhân hệ số).
     - Con số 59.06 cho thấy dữ liệu đang trộn lẫn các thang điểm khác nhau. Có thể bao gồm cả điểm "Đánh giá năng lực"(thang 1200 quy đổi về 100?) hoặc điểm năng khiếu nhân hệ số lạ.
   - Máy tính không biết thang điểm. Nó sẽ hiểu 59 điểm là "thiên tài"so với 29 điểm (thủ khoa thang 30). Điều này gây nhiễu nghiêm trọng.

3. Nghịch lý CPA thấp hơn GPA (Performance Paradox) ***

   - Dữ liệu: Mean CPA (1.96) thấp hơn đáng kể so với Mean GPA (2.29). Đồng thời CPA có tới 3.14% điểm 0 (Zeros), trong khi GPA chỉ có 0.23
   - CPA là tích lũy của GPA. Về lý thuyết, CPA phải ổn định. Việc CPA trung bình thấp hơn GPA hiện tại cho thấy:
     - Hoặc sinh viên đang tiến bộ vượt bậc (Kỳ này GPA cao kéo CPA lên).
     - Hoặc (khả năng cao hơn): Nhóm 3.14% có CPA=0 kia là những "Zombie Account"(Tài khoản chết, bảo lưu dài hạn hoặc bỏ học từ đầu) đang kéo tụt trung bình chung xuống.

4. Phân tích Phương thức xét tuyển (PTXT): Trong tập dữ liệu, các phương thức được mã hóa bằng số. Dưới đây là sự phân bổ các phương thức chính:

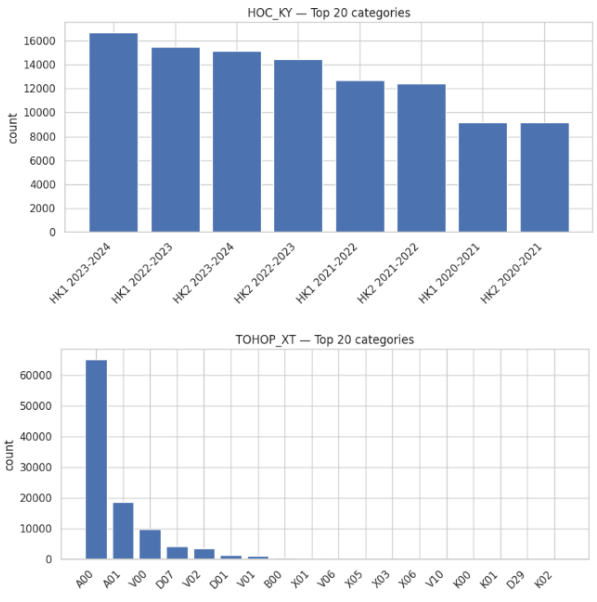
   - Mã PTXT '1': Đây là phương thức chiếm tỷ trọng lớn nhất với 13.712 sinh viên (chiếm khoảng 65%). Theo quy chuẩn của DHQGHN, mã này thường tương ứng với Xét tuyển dựa trên kết quả thi THPT Quốc gia (trước khi đổi mã) hoặc xét tuyển thẳng theo quy định.

- Mã PTXT ’100’: Đứng thứ hai với 7.052 sinh viên. Đây là mã phổ biến trong các năm gần đây (2020-2023) dành cho phương thức xét kết quả thi tốt nghiệp THPT.
- Các phương thức khác: * Mã ’409’ (124 sinh viên): Thường là xét tuyển kết hợp chứng chỉ quốc tế (IELTS, SAT...) với kết quả học tập/thi cử.
  - Mã ’402’ (30 sinh viên): Xét tuyển dựa trên bài thi Đánh giá năng lực (HSA) của ĐHQGHN.
  - Các mã ’200’, ’500’, ’3’: Chiếm tỷ lệ nhỏ, dành cho các đối tượng ưu tiên hoặc diện dự bị đại học.

2.4.2 Phân bố đặc trưng đơn biến (phân loại)

Phân tích Tổ hợp xét tuyển (TOHOP_XT): Dữ liệu cho thấy sự áp đảo của các khối tự nhiên và kỹ thuật:

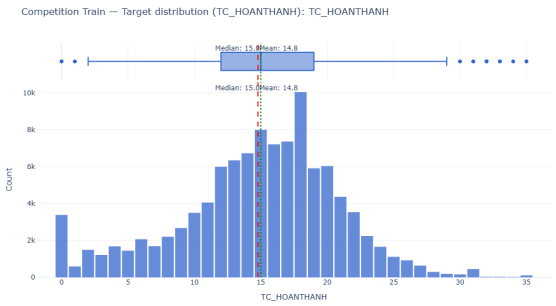
- A00 (Toán, Lý, Hóa): Là tổ hợp phổ biến nhất với 12.755 sinh viên, chiếm hơn 60% tổng lượng tuyển sinh. Điều này cho thấy khối ngành Kỹ thuật, Công nghệ và Tự nhiên chiếm ưu thế trong dữ liệu này.
- A01 (Toán, Lý, Anh): Đứng thứ hai với 3.712 sinh viên.
- V00 & V02: Các khối thi có môn năng khiếu (Vẽ) cũng xuất hiện đáng kể với tổng cộng hơn 2.600 sinh viên, cho thấy sự góp mặt của các ngành Kiến trúc/Quy hoạch.
- D07 (Toán, Hóa, Anh) & D01 (Toán, Văn, Anh): Lần lượt có 835 và 535 sinh viên trúng tuyển.



2.5 Quan hệ / Xu hướng

2.5.1 Phân phối mục tiêu huấn luyện

Những phát hiện chính



1. Cấu trúc "Hai Đỉnh Ngầm"(Latent Bimodality)

   - Dữ liệu: Mean (14.79) ≈ Trung vị (15,00) → Nhìn sơ qua tưởng là phân phối chuẩn (Normal Distribution).
   - Nhưng: Skewness = -0.41 (Lệch trái) và Risk Group = 20.81
   - Insight: Đây không phải là một đường cong hình chuông hoàn hảo. Nó là sự chồng lấn của 2 nhóm đối tượng:
     - Nhóm "Học ổn": Tập trung quanh mức 15-20 tín chỉ (đa số).
     - Nhóm "Gặp nạn"(The Strugglers): Một cái đuôi dài kéo về phía bên trái (0-10 tín).
   - Kết luận: Mô hình máy học (Model) sẽ rất dễ dự đoán tốt cho nhóm "Học ổn"nhưng dự đoán sai lệch (Over-predict) cho nhóm "Gặp nạn".

2. "Báo động Đỏ"ở ngưỡng 20% (The 20% Danger Zone)

   - Dữ liệu: Risk Group ($\leq$10 tín) chiếm tới 20.81%.
   - Insight: Cứ 5 sinh viên thì có 1 người hoàn thành $\leq$ 10 tín chỉ/kỳ. 10 tín chỉ thường chỉ tương đương 3 môn học. Đây là mức hiệu suất cực thấp, tiệm cận với cảnh báo học vụ.
   - Kết luận: Đây chính là nhóm Target quan trọng nhất. Nếu xây dựng mô hình dự báo rủi ro (Risk Prediction), độ chính xác trên nhóm 20% này mới là thước đo thành công, chứ không phải độ chính xác tổng thể (Accuracy).

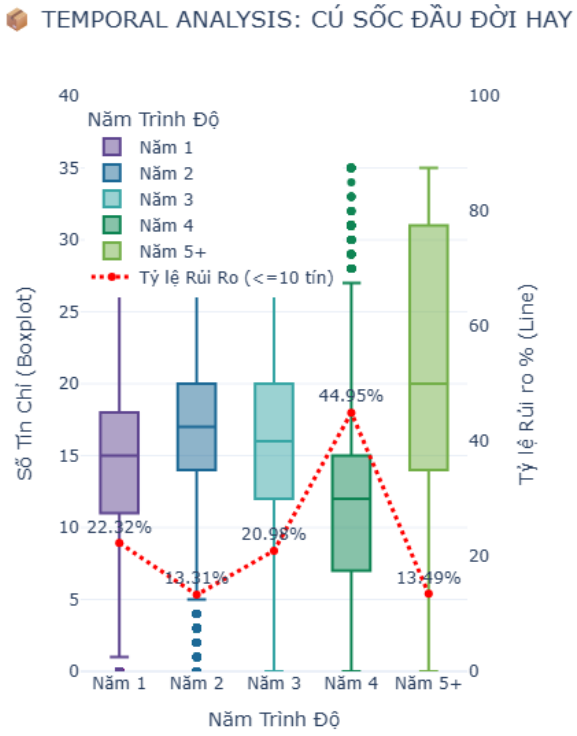
3. Hiệu quả của bước làm sạch trước đó

   - Dữ liệu: Max = 35, Perfect Score = 0.10%.
   - Insight: Việc chúng ta ép trần (Capping) ở mức 35 là hoàn toàn chính xác. Chỉ có 0.1% sinh viên chạm ngưỡng này, chứng tỏ không có hiện tượng "trần giả"(Artificial Ceiling) làm mất thông tin của sinh viên xuất sắc. Dữ liệu Min/Max đã nằm trong vùng an toàn vật lý.

   - Nếu Đường đỏ cao vút ở "Năm 1"(trên 25%):

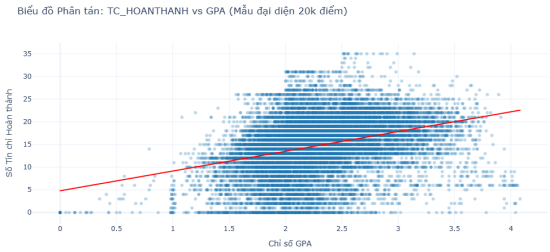
     Kết luận: Chính xác là "Cú sốc đầu đời". Sinh viên chưa quen cách học tín chỉ.

     Hành động: Cần thêm các feature về Vùng miền (Tỉnh lẻ vs Thành phố) hoặc Khối thi (Khối A00 học Toán tin sẽ đỡ sốc hơn Khối D học Toán tin?).



- Nếu Đường đỏ tăng dần về "Năm 4"hoặc "Năm 5+":
  Kết luận: Đây là "Sự kiệt sức"hoặc "Học dai dẳng". Sinh viên nợ môn dồn lại, hoặc đi làm sớm bỏ bê bài vở.
  Hành động: Cần thêm feature GPA Tích lũy (CPA) hoặc Số tín chỉ nợ để bắt tín hiệu chán nản này.
- Nếu Đường đỏ hình chữ U (Cao ở đầu và cuối, thấp ở giữa):
  Đây là mô hình kinh điển nhất: Sốc lúc đầu -> Quen dần (Năm 2, 3 ổn định) -> Đuối lúc cuối.

2.5.2 Mục tiêu so với các đặc điểm số



Những phát hiện chính

1. Biểu đồ phân tán GPA vs. Số Tín chỉ Hoàn thành (TC_HOANTHANH)

   - Tương quan Thuận: Đường OLS trendline dốc lên rõ rệt ($R^2 = 0,116$). Điều này xác nhận giả thuyết: Sinh viên có năng lực học thuật (GPA) cao thường có khả năng hoàn thành số lượng tín chỉ lớn hơn.

- Ngưỡng bão hòa: Nhìn vào biểu đồ Binned Trend, giá trị trung bình tăng mạnh từ GPA 1.6 đến 2.6, sau đó có xu hướng đi ngang hoặc tăng chậm lại ở nhóm GPA xuất sắc (>3.2).
- Ngưỡng bão hòa: Nhìn vào biểu đồ Binned Trend, giá trị trung bình tăng mạnh từ GPA 1.6 đến 2.6, sau đó có xu hướng đi ngang hoặc tăng chậm lại ở nhóm GPA xuất sắc (>3.2).
- Bất thường (Anomaly): Có một lượng lớn điểm dữ liệu nằm ở mức 0 tín chỉ hoàn thành trải dài trên mọi mức GPA.
- Câu hỏi chuyên sâu :
  - Tại sao sinh viên có GPA cao (trên 3.0) vẫn có những người hoàn thành 0 tín chỉ? Đây là do bảo lưu, bỏ học giữa chừng hay lỗi hệ thống dữ liệu?
  - Sự sụt giảm nhẹ ở nhóm GPA cao nhất (cuối biểu đồ đường) là do đâu? Có phải vì họ chọn các môn khó hơn, hay do họ ưu tiên chất lượng hơn số lượng?

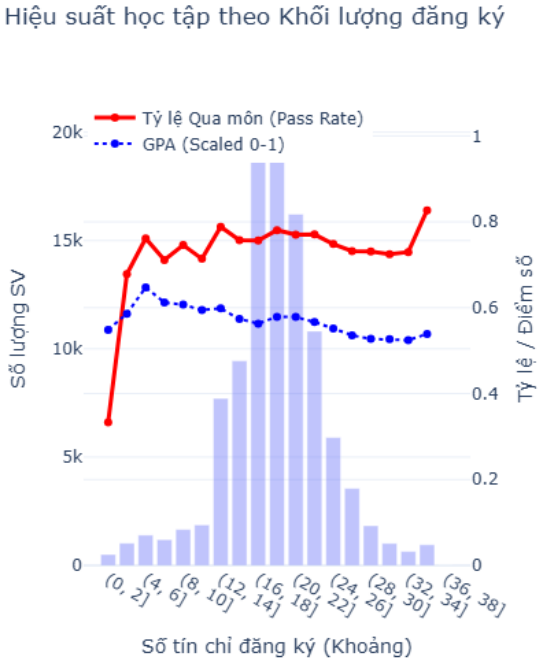
2. TC_DANGKY vs. Số Tín chỉ Hoàn thành

- Tương quan cực mạnh: $R^2 = 0,432$ là một chỉ số rất cao trong dữ liệu giáo dục. Biểu đồ Scatter tạo thành một hình phễu mở rộng.
- Hiệu suất hoàn thành: Đường xu hướng nằm dưới đường (nếu có). Điều này cho thấy khi đăng ký càng nhiều, rủi ro không hoàn thành hết (rớt môn/hủy môn) càng tăng.
- Điểm uốn: Biểu đồ đường cho thấy sự tăng trưởng gần như tuyến tính hoàn hảo.
- Câu hỏi chuyên sâu:
  - Tỷ lệ "rơi rụng"(TC_DANGKY - TC_HOANTHANH) có tăng đột biến ở ngưỡng đăng ký nào không? (Ví dụ: Đăng ký > 25 tín chỉ thì tỷ lệ trượt có cao hơn không?)
  - Những sinh viên đăng ký nhiều nhưng hoàn thành ít (phía dưới đường đỏ) có đặc điểm chung gì về ngành học hay năm tuyển sinh?

3. NAM_TUYENSINH vs. Số Tín chỉ Hoàn thành

- Tương quan Nghịch (nhẹ): Hệ số $-0,134$ cho thấy sinh viên các khóa gần đây (2022, 2023) có số tín chỉ hoàn thành trung bình thấp hơn các khóa cũ.
- Biến động theo năm: Biểu đồ đường cho thấy đỉnh cao vào năm 2020 và sụt giảm mạnh vào 2023.
- Câu hỏi chuyên sâu:
  - Sự sụt giảm năm 2023 có phải do sinh viên mới vào trường nên chưa có đủ thời gian tích lũy tín chỉ (Time-bias)?
  - Biến động năm 2020-2021 có liên quan gì đến việc học Online trong đại dịch COVID-19 không? (Sinh viên đăng ký và học được nhiều hơn hay ít hơn?)

Hiệu suất học tập theo khối lượng đăng kí: Ngưỡng quá tải là bao nhiêu? Đăng ký > 25 tín chỉ thì tỷ lệ trượt có cao hơn không? **Những phát hiện chính**



(a) Phân tích Tỷ lệ Qua môn (Pass Rate) và Ngưỡng 25 tín chỉ

- Đăng ký $>$ 25 tín chỉ có làm tỷ lệ trượt cao hơn không? * Thực tế ngược lại: Quan sát biểu đồ, từ ngưỡng 24-26 tín chỉ trở đi, đường màu đỏ không hề đi xuống mà có xu hướng đi ngang khá ổn định, thậm chí còn tăng vọt ở nhóm đăng ký rất cao (34-36 tín chỉ).
  - Lý giải: Những sinh viên "dám"đăng ký trên 25 tín chỉ thường là nhóm có năng lực học tập xuất sắc hoặc là nhosm chương trình tài năng. Do đó, tỷ lệ trượt của họ không cao hơn nhóm trung bình.
- Ngưỡng "Nguy hiểm"thực sự: Tỷ lệ trượt cao nhất (Pass rate thấp nhất) lại nằm ở nhóm đăng ký cực ít (0-2 tín chỉ). Điều này thường phản ánh những sinh viên có ý định bỏ học hoặc gặp sự cố cá nhân nghiêm trọng ngay từ đầu kỳ.

(b) Tìm kiếm "Ngưỡng Quá Tải"(Overload Point) qua GPA:
Dù Tỷ lệ qua môn (Pass Rate) ổn định, nhưng GPA (Đường nét đứt màu xanh) lại kể một câu chuyện khác về "chất lượng"học tập:

- Điểm bão hòa (Peak Performance): Hiệu suất học tập cao nhất nằm ở khoảng 4-6 tín chỉ (GPA đạt đỉnh).
- Xu hướng suy giảm dần: Từ sau ngưỡng 12 tín chỉ, đường GPA bắt đầu dốc xuống đều đặn.
- Xác định ngưỡng quá tải: * Khoảng từ 18 đến 24 tín chỉ là nơi có số lượng sinh viên tập trung đông nhất (các cột bar cao nhất).
  - Tuy nhiên, GPA ở khoảng này thấp hơn rõ rệt so với nhóm đăng ký ít. Đây chính là "Ngưỡng đánh đổi": Sinh viên chấp nhận GPA giảm nhẹ để lấy được số lượng tín chỉ nhiều hơn.

Tại sao GPA $>$ 3.0 mà hoàn thành 0 tín?
Những phát hiện chính

(a) Phân tích kết quả: Nhìn vào các mẫu cung cấp, thấy sự vô lý đến mức hiển nhiên:

- Bằng chứng 1 (Sự bất khả thi): Sinh viên 9453 (2023) có GPA 4.05. Đây là điểm tuyệt đối (A+). Không thể có chuyện một sinh viên đạt điểm tuyệt đối mà lại không hoàn thành tín chỉ nào (TC_HOANTHANH = 0).
- Bằng chứng 2 (Sự tồn tại): Sinh viên 25525 bị tình trạng này ở 2 học kỳ liên tiếp (HK1 & HK2 năm 2020). Nếu thực sự rớt hết 17-19 tín chỉ ở HK1, sinh viên này đã bị cảnh cáo học vụ hoặc buộc thôi học, không thể đăng ký tiếp HK2 với 19 tín chỉ được.
- Kết luận: Sinh viên này ĐÃ HỌC VÀ ĐÃ QUA MÔN. Con số GPA là đúng (vì nó cụ thể: 3.27, 3.44...), còn con số TC_HOANTHANH = 0 là SAI (do lỗi cập nhật dữ liệu, lỗi query, hoặc độ trễ hệ thống khi chốt sổ tín chỉ).

(b) Nguyên nhân sâu xa:

- Lỗi cập nhật: Điểm số (GPA) thường được giảng viên nhập trước. Trạng thái "Hoàn thành tín chỉ"thường được hệ thống chạy batch job sau. Có thể dữ liệu được trích xuất đúng lúc GPA đã có nhưng Tín chỉ chưa "nhảy"số.
- Học cải thiện/Học lại: Đôi khi hệ thống ghi nhận điểm mới vào GPA nhưng không cộng thêm tín chỉ tích lũy nếu đó là môn học lại (tuy nhiên TC_HOANTHANH theo kỳ vẫn phải có).

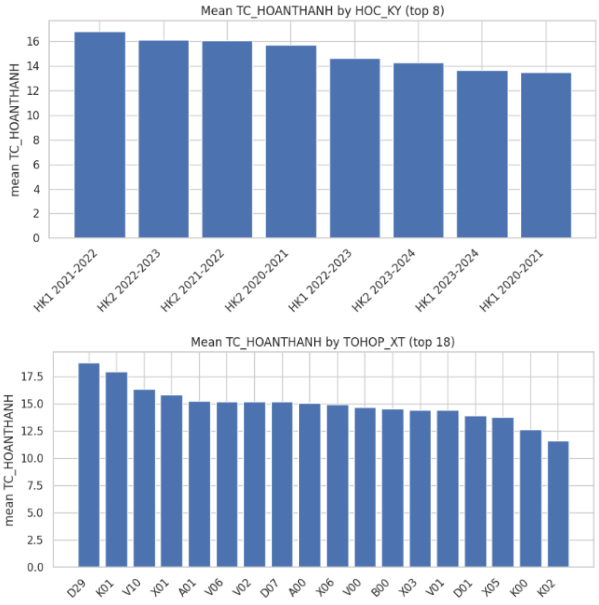
(c) Giải pháp: Tin vào điểm số:

- Dùng chiến thuật Imputation (Điền khuyết) dựa trên logic: "Nếu GPA cao, chắc chắn sinh viên đã hoàn thành số tín chỉ đã đăng ký". Thiết kế cho "Bảng Tập Luật Hồi Quy Tín Chỉ"(Credit Regression Ruleset) dựa trên phân phối xác suất thực tế.
- Nguyên lý: GPA càng cao $\rightarrow$ Tỷ lệ qua môn (Pass Rate) càng tiệm cận 100%.

2.5.3 Mục tiêu so với Phân loại

Những phát hiện chính

1. Hiện ứng "Bong bóng COVID"
   - Dữ liệu: Năm học 2021-2022 có số tín chỉ hoàn thành trung bình cao đột biến ($\approx$16.8).
   - Insight: Giai đoạn này học online, có thể việc đánh giá lỏng hơn hoặc sinh viên tận dụng thời gian ở nhà để đăng ký nhiều môn hơn.
   - Rủi ro: Mô hình học từ giai đoạn này sẽ bị "lạc quan tếu"(Over-optimistic) khi áp dụng cho năm 2023-2024 (vốn đã quay lại mức bình thường 13.6). Cần một feature IS_ONLINE_PERIOD để đánh dấu.
2. Cú sốc "Tân sinh viên"
   - Dữ liệu: Hầu hết các năm, Học kỳ 1 (HK1) đều có điểm trung bình thấp hơn Học kỳ 2 (HK2). Ví dụ 2022: HK1 (14.6) $<$ HK2 (16.1). Đặc biệt HK1 2023-2024 chỉ đạt 13.69.

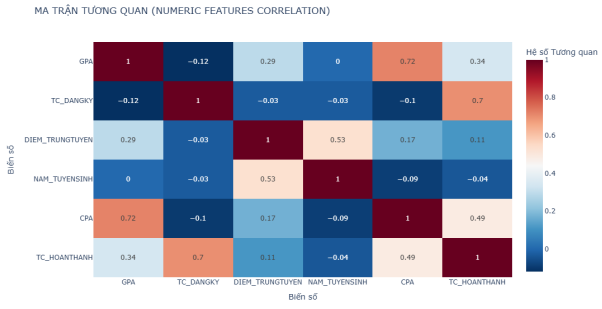


- Insight: Sinh viên năm nhất chưa quen với môi trường ĐH, dễ bị rớt môn hoặc đăng ký ít ở kỳ đầu. Đây là giai đoạn nhạy cảm nhất.

3. Sự phân hóa "Khối Tự Nhiên vs Xã Hội"

- Dữ liệu: Khối A00/A01 (Toán, Lý, Hóa/Anh) có phong độ ổn định và cao (≈15.1-15.2). Trong khi Khối D01 (Toán, Văn, Anh) thấp hơn hẳn (13.9).
- Insight: Nếu đây là trường thiên về Kỹ thuật/Công nghệ (như UET, HUS), sinh viên khối D01 đang gặp khó khăn thực sự với các môn đại cương tự nhiên (Giải tích, Vật lý).

2.5.4 Tương tác giữa các đặc điểm số



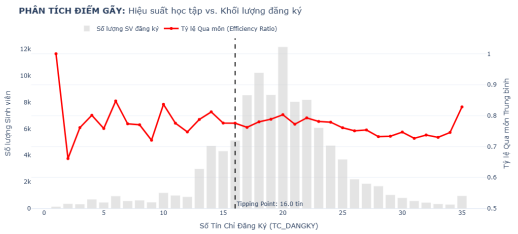
1. Quy luật "Muốn thắng phải dám chơi"

- Dữ liệu: TC_DANGKY có tương quan mạnh nhất (0.70), áp đảo hoàn toàn GPA (0.34).
- Insight: Để dự báo một sinh viên hoàn thành bao nhiêu tín chỉ, việc biết họ "tham vọng"thế nào (đăng ký bao nhiêu) quan trọng gấp đôi việc biết họ "giỏi"thế nào (GPA).

- Nghịch lý: Điều này ám chỉ rằng rào cản lớn nhất để hoàn thành nhiều tín chỉ không phải là năng lực học tập (độ khó môn học) mà là hành vi đăng ký (dám đăng ký nhiều hay không).

2. Sự chiến thắng của "Sức bền"
   - Dữ liệu: CPA (0.49) có tương quan cao hơn hẳn GPA (0.34).
   - Insight: Điểm trung bình tích lũy (CPA - Phong độ dài hạn) dự báo tốt hơn điểm kỳ này (GPA - Phong độ ngắn hạn). Những sinh viên có "gốc"tốt thường duy trì tiến độ hoàn thành tín chỉ ổn định hơn, bất chấp GPA kỳ này có thể biến động.
3. Điểm đầu vào
   - Dữ liệu: DIEM_TRUNGTUYEN chỉ có tương quan 0.11 (Rất yếu).
   - Insight: Đầu vào cao không đảm bảo đầu ra tốt. Môi trường đại học là một môi trường mới. Nếu mô hình quá tin vào điểm thi THPT, nó sẽ thất bại.
   - Giá trị của Feature Engineering:
     Tuy nhiên, khi bạn kết hợp nó thành ENTRY_MOMENTUM, tương quan tăng vọt lên 0.24 (Gấp đôi!). Điều này chứng minh rằng "Sự tiến bộ so với chính mình"quan trọng hơn điểm số tuyệt đối. ***
4. Feature "Áp lực tải"(Load Pressure)
   - Dữ liệu: LOAD_PRESSURE có tương quan 0.38.
   - Insight: Feature này (TC_DANGKY / GPA) thấp hơn TC_DANGKY gốc (0.70). Điều này cho thấy việc chia cho GPA đã làm "loãng"tín hiệu tích cực của việc đăng ký nhiều.
   - Kết luận: Feature này có thể không tốt để dự báo TC_HOANTHANH (số lượng), nhưng có thể cực tốt để dự báo Rủi ro rớt môn (Binary Classification).

Điểm gãy Hiệu suất theo Năng lực Học tập?



(a) Sự thật về Nhóm Yếu (GPA < 2.0) - "Càng đăng ký nhiều càng thể hiện sự quyết tâm"
   - Quan sát: Đường màu đỏ (Nhóm Yếu) có xu hướng đi lên khi số tín chỉ đăng ký tăng từ 10 đến 30.
   - Insight: Đây là một nghịch lý thú vị. Những sinh viên học yếu nhưng vẫn "dám"đăng ký 25-30 tín chỉ thường là những người đang có quyết tâm vực dậy cao độ hoặc đang phải học trả nợ môn tích cực để kịp ra trường. Sự gia tăng tỷ lệ qua môn cho thấy nỗ lực của họ đang có hiệu quả nhất định về mặt "số lượng".

(b) Sự thật về Nhóm Khá/Giỏi (GPA >= 3.2) - "Không thất bại trước đăng ký số lượng tín chỉ"

- Quan sát: Đường màu xanh lá (Nhóm Giỏi) gần như nằm ngang và duy trì ở mức sát 1.0 (100
- Insight: Đối với nhóm này, "Điểm gãy của sự tham lam"hầu như không tồn tại ở khía cạnh Tỷ lệ qua môn. Năng lực của họ đủ để bao quát khối lượng công việc khổng lồ.
- Phân tích sự đánh đổi giữa Số lượng tín chỉ (TC_DANGKY) và Chất lượng điểm số (GPA) đặc biệt cho nhóm sinh viên Giỏi.

(c) Sự thật về Nhóm Trung Bình (2.0 - 3.2) - "Điểm gãy thực sự"

- Quan sát: Đây là nhóm duy nhất đường biểu đồ (màu cam) có xu hướng dốc xuống rõ rệt sau ngưỡng 20 tín chỉ và chạm sát "Ngưỡng an toàn 80%".
- Insight: Nhóm Trung bình là nhóm nhạy cảm nhất với khối lượng. Họ không có nền tảng cực tốt như nhóm Giỏi, cũng không ở thế "đường cùng"như nhóm Yếu. Khi họ đăng ký quá 20 tín chỉ, rủi ro trượt môn bắt đầu xuất hiện rõ rệt do sự phân tán nguồn lực.

2.6 Kiểm tra tình trạng dữ liệu:

2.6.1 Giá trị ngoại lệ

Phân tích chuyên sâu Outliers của Target.

1. Hiệu quả tuyệt vời của bước "Data Cleaning Chỉ có 0.37% (394 SV) rơi vào "Bức tường số 0".

   - 394 sinh viên này khả năng cao là "Hard-core Failures"(Rớt thật sự) hoặc bỏ học giữa chừng sau khi chốt danh sách, chứ không phải lỗi hệ thống nữa. Đây là nhãn (Label) sạch cho bài toán dự báo Drop-out.

2. Vùng nguy hiểm thực sự: "Nhóm 1-2 Tín chỉ": Có tới 1,616 SV (2 tín) và 746 SV (1 tín). Tổng cộng nhóm này đông gấp 6 lần nhóm 0 tín.

   - Nhóm này nguy hiểm hơn nhóm 0 tín nhiều, vì họ vẫn đi học, vẫn tồn tại trên hệ thống, nhưng thực chất đã "chết lâm sàng"về mặt học thuật (Academic Zombie).
   - Thực tế: Trong số 2,362 sinh viên hoàn thành 1-2 tín chỉ, có tới 1,375 người (chiếm khoảng 58%) thuộc nhóm "Rớt thảm hại"(Đăng ký $\geq$ 10 tín nhưng chỉ qua 1-2 môn).
   - Kết luận: Giả thuyết "Học nhẹ/Thực tập"chỉ đúng với khoảng 42% (987 SV). Phần lớn còn lại là nhóm Rủi ro cực cao (Academic Zombies) cần cảnh báo ngay lập tức.
   - Hành động: Tạo feature IS_FAIL_SHOCK = 1 cho nhóm 1,375 sinh viên này.

2.6.2 Sự thay đổi phân bổ (Đào tạo so với Kiểm tra)

2.6.3 So sánh kết quả thi với bộ dữ liệu gốc

- Sự trôi lệch mục tiêu

- Kiểm tra rò rỉ đơn giản
- Biểu đồ đơn biến - Tập dữ liệu gốc.

3 Kỹ thuật trích chọn đặc trưng

3.1 Chuyển đổi các đặc điểm thành dạng phù hợp cho việc lập mô hình.

3.1.1 Mã hóa phân loại

- Mã hóa One-Hot
- Mã hóa nhãn

3.2 TARGET ENCODING (CON DAO HAI LƯỠI SẮC BÉN)

Về Phân Loại Năm Đào Tạo (Điều 23 - Quy chế Đào tạo):Quy định: Sinh viên được xếp năm trình độ dựa trên tín chỉ tích lũy, KHÔNG dựa trên năm nhập học.
Năm 1: < 35 tín chỉ.
Năm 2: $35 - 70$ tín chỉ.
Năm 3: $71 - 105$ tín chỉ.
Năm 4: > 105 tín chỉ.Phán quyết: Việc dùng NAM_TUYENSINH để tính năm học là sai bản chất quy chế tín chỉ.Hành động: Tạo feature mới STUDENT_YEAR_LEVEL được tính động dựa trên tổng TC_HOANTHANH tích lũy.

3.3 CÁC TÍNH NĂNG CỬA SỔ VÀ CỬA TRƯỢT

3.4 INTERACTION "LOAD vs. ABILITY"(TỶ LỆ TẢI)

4 Phát triển mô hình

Thực thi các thuật toán cơ bản.

4.1 CHIẾN LƯỢC VALIDATION: GROUP K-FOLD

- Kiểm định chéo.
- Tập hợp.
- Xử lý hậu kỳ & Hiệu chuẩn
- Trích xuất các đặc điểm quan trọng.

5 Kết quả và thảo luận

5.1 Tóm tắt EDA