BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM

KHOA CÔNG NGHỆ THÔNG TIN

❖—❖🕮❖—❖

HỌC KỲ 2 - NĂM HỌC 2024 -2025

MÔN HỌC: LẬP TRÌNH R CHO PHÂN TÍCH

BÁO CÁO ĐỒ ÁN

Phân tích dữ liệu và xây dựng mô hình dự đoán rủi ro vỡ nợ của khách hàng sử dụng thẻ tín dụng

GVHD: TS. Phan Thị Thể

Danh sách sinh viên thực hiện:

Mã số SV	Họ và tên	Mức độ đóng góp (%)
23133074	Nguyễn Phước Thịnh	100%
23133068	Nguyễn Tấn Thành	100%
23133029	Vương Đức Huy	100%
23133026	Châu Gia Huy	100%

TP. Hồ Chí Minh, tháng 5 năm 2025

MỤC LỤC

1. Tóm tắt

Đồ án này tận dụng sức mạnh của ngôn ngữ lập trình R để phân tích và dự đoán khả năng trả nợ (vỡ nợ) của khách hàng dựa trên tập dữ liệu UCI_Credit_Card.csv từ Kaggle, một bộ dữ liệu thực tế chứa thông tin về nhân khẩu học, hạn mức tín dụng, lịch sử thanh toán và tình trạng vỡ nợ của khách hàng thẻ tín dụng tại Đài Loan. Mục tiêu chính là xây dựng một quy trình phân tích dữ liệu toàn diện, từ tiền xử lý, phân tích khám phá (EDA), trực quan hóa, đến mô hình hóa, nhằm xác định các yếu tố ảnh hưởng đến rủi ro tín dụng và dự đoán khả năng vỡ nợ một cách chính xác.

Quy trình thực hiện bao gồm ba giai đoạn chính:

+ Tiền xử lý dữ liệu: Sử dụng các thư viện như dplyr và tidyr để xử lý giá trị thiếu (NA) bằng phương pháp điền giá trị trung bình hoặc mode, chuẩn hóa tên cột (ví dụ: đổi default.payment.next.month thành Default, AGE thành Age), mã hóa one-hot các biến phân loại như SEX, EDUCATION, MARRIAGE, và chuẩn hóa các biến số như LIMIT_BAL, BILL_AMT, PAY_AMT. Dữ liệu sau xử lý được lưu thành file clean_default.csv.
+ Phân tích khám phá và trực quan hóa: Sử dụng ggplot2, plotly, và corrplot để tạo các biểu đồ phân tích phân phối biến (Age, Income), mối quan hệ giữa các yếu tố như hạn mức tín dụng (LIMIT_BAL), lịch sử thanh toán (PAY_0 đến PAY_6) với biến mục tiêu Default. Các biểu đồ như histogram, boxplot, và heatmap tương quan giúp phát hiện các yếu tố quan trọng ảnh hưởng đến rủi ro vỡ nợ.
+ Mô hình hóa và dự báo: Xây dựng ba mô hình học máy gồm Logistic Regression, Random Forest, và XGBoost để dự đoán khả năng vỡ nợ. Các mô hình được huấn luyện trên tập dữ liệu đã tiền xử lý, với các đặc trưng được lựa chọn cẩn thận. Hiệu suất mô hình được đánh giá qua các chỉ số như Accuracy, AUC (Area Under the ROC Curve), Precision, Recall, và F1-Score, đồng thời sử dụng đường cong ROC để trực quan hóa khả năng phân loại. Các kỹ thuật như SMOTE hoặc trọng số lớp được áp dụng để xử lý vấn đề mất cân bằng dữ liệu (tỷ lệ vỡ nợ chỉ khoảng 22%).

Kết quả phân tích chỉ ra rằng các yếu tố như lịch sử thanh toán trễ hạn (PAY_0, PAY_1), thu nhập, và hạn mức tín dụng có ảnh hưởng mạnh đến khả năng vỡ nợ. Các mô hình Random Forest và XGBoost cho hiệu suất vượt trội (AUC ~0.82–0.85) so với Logistic Regression (AUC ~0.73), nhờ khả năng nắm bắt các mối quan hệ phi tuyến. Tối ưu hóa tham số được thực hiện thông qua Grid Search để cải thiện độ chính xác dự báo.

Đồ án không chỉ minh chứng khả năng ứng dụng R trong phân tích dữ liệu tín dụng mà còn cung cấp giá trị thực tiễn cho các tổ chức tài chính trong việc đánh giá rủi ro, tối ưu hóa chính sách cho vay, và quản lý danh mục khách hàng. Đồng thời, đây là cơ hội để nhóm rèn luyện kỹ năng xử lý dữ liệu thực tế, trực quan hóa, và xây dựng mô hình học máy trong lĩnh vực tài chính-ngân hàng.

2. Giới thiệu

Trong bối cảnh ngành tài chính-ngân hàng, dự đoán khả năng trả nợ của khách hàng là một bài toán quan trọng, đóng vai trò then chốt trong việc giảm thiểu rủi ro tín dụng và tối ưu hóa quy trình phê duyệt khoản vay. Việc đánh giá chính xác nguy cơ vỡ nợ không chỉ giúp các tổ chức tài chính giảm tổn thất mà còn hỗ trợ định giá sản phẩm tín dụng, quản lý danh mục khách hàng, và xây dựng các chiến lược tài chính bền vững. Với sự phát triển của khoa học dữ liệu và học máy, các phương pháp phân tích tiên tiến đang trở thành công cụ đắc lực để giải quyết thách thức này.

Ngôn ngữ lập trình R, với hệ sinh thái thư viện phong phú như dplyr, ggplot2, caret, randomForest, và xgboost, cung cấp một nền tảng mạnh mẽ để xử lý, phân tích, và mô hình hóa dữ liệu tín dụng. Đồ án này tập trung vào việc ứng dụng R để phân tích tập dữ liệu UCI_Credit_Card.csv, một bộ dữ liệu thực tế từ Kaggle, chứa thông tin về 30,000 khách hàng với các đặc trưng như hạn mức tín dụng (LIMIT_BAL), nhân khẩu học (SEX, EDUCATION, MARRIAGE, AGE), lịch sử thanh toán (PAY_0 đến PAY_6), số tiền hóa đơn (BILL_AMT1 đến BILL_AMT6), số tiền thanh toán (PAY_AMT1 đến PAY_AMT6), và biến mục tiêu là tình trạng vỡ nợ (default.payment.next.month).

Mục tiêu cụ thể của đồ án bao gồm:

+ Tiền xử lý và chuẩn hóa dữ liệu: Xử lý giá trị thiếu, chuẩn hóa định dạng cột, và mã hóa các biến phân loại (SEX, EDUCATION, MARRIAGE) bằng one-hot encoding. Chuẩn hóa các biến số như LIMIT_BAL, BILL_AMT, PAY_AMT để đảm bảo phù hợp với các mô hình học máy.
+ Phân tích khám phá dữ liệu (EDA) và trực quan hóa: Sử dụng ggplot2 và corrplot để phân tích phân phối các biến quan trọng (Age, LIMIT_BAL, PAY_0), khám phá mối quan hệ giữa các yếu tố rủi ro (PAY_0, BILL_AMT, PAY_AMT) và khả năng vỡ nợ (Default). Các biểu đồ như histogram, boxplot, và heatmap tương quan được sử dụng để phát hiện các xu hướng và tri thức mới.
+ Xây dựng và đánh giá mô hình dự đoán: Huấn luyện ba mô hình học máy gồm Logistic Regression, Random Forest, và XGBoost trên tập dữ liệu đã xử lý. Đánh giá hiệu suất qua các chỉ số Accuracy, AUC, Precision, Recall, và F1-Score, đồng thời trực quan hóa bằng đường cong ROC. Tối ưu hóa tham số sử dụng Grid Search hoặc Random Search để cải thiện hiệu suất, đồng thời áp dụng các kỹ thuật như SMOTE hoặc trọng số lớp để xử lý mất cân bằng dữ liệu.

Thông qua việc ứng dụng các thư viện R như dplyr, ggplot2, caret, randomForest, và xgboost, đồ án không chỉ hướng đến việc xây dựng các mô hình dự đoán có độ chính xác cao mà còn thiết lập một quy trình phân tích khoa học và minh bạch. Đây là cơ hội để sinh viên vận dụng kiến thức đã học vào giải quyết bài toán thực tiễn trong lĩnh vực tài chính, đồng thời nâng cao kỹ năng phân tích dữ liệu, trực quan hóa, và mô hình hóa học máy. Kết quả từ đồ án dự kiến sẽ cung cấp những hiểu biết sâu sắc về các yếu tố ảnh hưởng đến rủi ro tín dụng, hỗ trợ các tổ chức tài chính đưa ra quyết định cho vay sáng suốt và quản lý rủi ro hiệu quả.

3. Dữ liệu

- Nguồn dữ liệu: UCI_Credit_Card.csv

- Tham khảo từ: Kaggle

3.1. Mô tả dữ liệu

***Tập dữ liệu UCI_Credit_Card:**

- Tổng quan:

  + Số dòng (bản ghi): 30,000
  + Số cột (thuộc tính): 25

- Các cột dữ liệu:

  + ID: Mã định danh duy nhất cho mỗi khách hàng.
  + LIMIT_BAL: Hạn mức tín dụng được cấp (đơn vị: Đô la Đài Loan).
  + SEX: Giới tính (1 = Nam, 2 = Nữ).
  + EDUCATION: Trình độ học vấn (1 = Sau đại học, 2 = Đại học, 3 = Trung học, 4 = Khác, 5 = Không xác định, 6 = Không xác định).
  + MARRIAGE: Tình trạng hôn nhân (1 = Đã kết hôn, 2 = Độc thân, 3 = Khác).
  + AGE: Tuổi của khách hàng.
  + PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6: Tình trạng thanh toán hóa đơn từ tháng 9/2005 đến tháng 4/2005 (số tháng trước đó). Giá trị: -2 = Không có giao dịch, -1 = Thanh toán đúng hạn, 0 = Thanh toán xoay vòng, số dương = Số tháng chậm thanh toán.
  + BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6: Số tiền hóa đơn từ tháng 9/2005 đến tháng 4/2005 (đơn vị: Đô la Đài Loan).
  + PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6: Số tiền đã thanh toán từ tháng 9/2005 đến tháng 4/2005 (đơn vị: Đô la Đài Loan).
  + default.payment.next.month: Kết quả dự đoán khả năng vỡ nợ trong tháng tiếp theo (0 = Không vỡ nợ, 1 = Vỡ nợ).

- Cấu trúc dữ liệu:

```
=== UCI_Credit_Card.csv ===
# A tibble: 30,000 × 25
```

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0
2	2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0
3	3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518
4	4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000
5	5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000
6	6	50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024	2500
7	7	500000	1	1	2	29	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944	55000
8	8	100000	2	2	2	23	0	-1	-1	0	0	-1	11876	380	601	221	-159	567	380
9	9	140000	2	3	1	28	0	0	2	0	0	0	11285	14096	12108	12211	11793	3719	3329
10	10	20000	1	3	2	35	-2	-2	-2	-2	-1	-1	0	0	0	0	13007	13912	0

3.2. Tiền xử lý dữ liệu

* Tập dữ liệu UCI_Credit_Card

- Kiểm tra các giá trị missing

  + Mục tiêu: Phát hiện các ô bị bỏ trống (NA, NULL, hoặc rỗng) trong các cột quan trọng như: LIMIT_BAL, SEX, EDUCATION, MARRIAGE, AGE, PAY_0→PAY_6, BILL_AMT→BILL_AMT6, PAY_AMT1→PAY_AMT6, default.payment.next.month.
  + Thực hiện: Sử dụng hàm colSums(is.na(data)) để kiểm tra số lượng giá trị NA trong mỗi cột. Kết quả cho thấy tập dữ liệu không có giá trị NA, đảm bảo tính toàn vẹn của dữ liệu gốc.
  + Nhận xét: Việc không có giá trị thiếu giúp giảm thiểu nhu cầu áp dụng các kỹ thuật điền giá trị (imputation), nhưng vẫn cần kiểm tra giá trị bất thường để đảm bảo tính hợp lệ.

- Kiểm tra giá trị không hợp lệ

  + Cột EDUCATION:
    - Các giá trị hợp lệ: 1 (sau đại học), 2 (đại học), 3 (trung học), 4 (khác).
    - Phát hiện giá trị bất thường (0, 5, 6) thông qua table(data$EDUCATION).
    - Xử lý: Gộp các giá trị 0, 5, 6 vào giá trị 4 (khác), vì chúng không được định nghĩa rõ ràng và có thể đại diện cho các trường hợp không xác định.
    - Kết quả sau xử lý: Tần suất giá trị trong EDUCATION chỉ còn các giá trị 1, 2, 3, 4, được xác nhận qua table(data$EDUCATION) sau khi gộp.
  + Cột MARRIAGE:

- Các giá trị hợp lệ: 1 (đã kết hôn), 2 (độc thân), 3 (khác).
- Phát hiện giá trị bất thường (0) thông qua table(data$MARRIAGE).
- Xử lý: Gộp giá trị 0 vào giá trị 3 (khác), vì 0 không được định nghĩa rõ ràng và có thể đại diện cho các trường hợp không xác định.
- Kết quả sau xử lý: Tần suất giá trị trong MARRIAGE chỉ còn các giá trị 1, 2, 3, được xác nhận qua table(data$MARRIAGE) sau khi gộp.

+ Các cột số liệu:
    - Các cột như LIMIT_BAL, AGE, BILL_AMT1 đến BILL_AMT6, PAY_AMT1 đến PAY_AMT6 được kiểm tra để đảm bảo không có giá trị âm bất hợp lý.
    - LIMIT_BAL và PAY_AMT* không chứa giá trị âm, phù hợp với ngữ cảnh tài chính.
    - BILL_AMT* có thể chứa giá trị âm (do hoàn tiền hoặc điều chỉnh hóa đơn), được giữ nguyên vì phù hợp với thực tế.
    - PAY_0 đến PAY_6 có giá trị âm (-2, -1), biểu thị không có giao dịch hoặc thanh toán đúng hạn, được xác nhận là hợp lệ.

- Xử lý dữ liệu thiếu hoặc không hợp lệ

+ Vì không có giá trị NA, không cần áp dụng các phương pháp thay thế như mean, median, hoặc mode.
+ Các giá trị bất thường trong EDUCATION và MARRIAGE đã được xử lý bằng cách gộp vào các danh mục hợp lệ (4 cho EDUCATION, 3 cho MARRIAGE), đảm bảo tính nhất quán của dữ liệu.
+ Cột mục tiêu default.payment.next.month được kiểm tra, đảm bảo chỉ chứa hai giá trị: 0 (không vỡ nợ) và 1 (vỡ nợ).

- Chuẩn hóa kiểu dữ liệu

+ Cột phân loại:
    - SEX, EDUCATION, MARRIAGE, và default.payment.next.month được chuyển thành kiểu factor để phù hợp với phân tích và mô hình hóa.

    - Ví dụ: data$EDUCATION <- as.factor(data$EDUCATION) đảm bảo EDUCATION được xử lý như biến phân loại thay vì số nguyên.
+ Cột số liệu:
    - AGE được giữ ở kiểu số nguyên (integer), vì không phát hiện giá trị thập phân hoặc ký tự đặc biệt.
    - LIMIT_BAL, BILL_AMT1 đến BILL_AMT6, PAY_AMT1 đến PAY_AMT6 được giữ ở kiểu số thực (float), phù hợp với giá trị tài chính.
    - PAY_0 đến PAY_6 được giữ ở kiểu số nguyên, với phạm vi từ -2 đến 9, được xác nhận là hợp lệ.

- Chuẩn hóa các cột số

+ Mục tiêu: Đưa các cột số như LIMIT_BAL, BILL_AMT1 đến BILL_AMT6, PAY_AMT1 đến PAY_AMT6 về cùng quy mô để cải thiện hiệu quả mô hình hóa.
+ Thực hiện: Áp dụng chuẩn hóa z-score cho các cột LIMIT_BAL, BILL_AMT1 đến BILL_AMT6, và PAY_AMT1 đến PAY_AMT6 bằng hàm scale(). Điều này đảm bảo các biến có trung bình bằng 0 và độ lệch chuẩn bằng 1.
+ Nhận xét: Chuẩn hóa giúp giảm thiểu tác động của các biến có thang đo lớn (ví dụ: LIMIT_BAL lên đến hàng triệu) so với các biến nhỏ hơn (ví dụ: AGE).

- Mã hóa các biến phân loại

+ Mục tiêu: Chuyển các biến phân loại thành dạng số để sử dụng trong các mô hình học máy như Logistic Regression.
+ Thực hiện: Sử dụng one-hot encoding cho các cột SEX, EDUCATION, và MARRIAGE bằng thư viện fastDummies::dummy_cols(). Ví dụ, EDUCATION được chuyển thành các cột EDUCATION_1, EDUCATION_2, EDUCATION_3, và EDUCATION_4.
+ Nhận xét: One-hot encoding đảm bảo các biến phân loại được biểu diễn dưới dạng nhị phân, phù hợp với các mô hình tuyến tính, đồng thời tránh giả định thứ tự không hợp lý giữa các nhóm.

- Kiểm tra lại NA và lưu dữ liệu

+ Kiểm tra NA: Sử dụng colSums(is.na(data)) lần cuối để xác nhận không còn giá trị thiếu sau các bước xử lý.
+ Lưu dữ liệu: Dữ liệu sau xử lý được lưu vào tệp clean_default.csv trong thư mục processed/
+ Kết quả: Tệp clean_default.csv chứa dữ liệu đã được làm sạch, chuẩn hóa, và sẵn sàng cho phân tích khám phá và mô hình hóa.

- Nhận xét tổng quan: Quá trình tiền xử lý đã giải quyết hiệu quả các vấn đề về giá trị bất thường (EDUCATION, MARRIAGE) và chuẩn hóa dữ liệu để đảm bảo tính nhất quán. Việc không có giá trị NA giúp đơn giản hóa quy trình, nhưng các bước kiểm tra kỹ lưỡng vẫn được thực hiện để đảm bảo chất lượng dữ liệu. Dữ liệu sau xử lý được lưu trữ khoa học, tạo nền tảng vững chắc cho các giai đoạn tiếp theo.

3.3 Phân tích mất cân bằng dữ liệu

***Tập dữ liệu UCI_Credit_Card**

- Mô tả biến mục tiêu

+ Biến mục tiêu default.payment.next.month là biến nhị phân, biểu thị trạng thái vỡ nợ của khách hàng trong tháng tiếp theo:
    - 0: Không vỡ nợ.
    - 1: Vỡ nợ.
+ Đây là biến quan trọng để dự đoán rủi ro tín dụng, và phân phối của nó ảnh hưởng trực tiếp đến hiệu suất của các mô hình học máy.

- Kiểm tra tỷ lệ lớp trong tập dữ liệu gốc

+ Tỷ lệ của biến default.payment.next.month trong tập huấn luyện (80% dữ liệu gốc) được kiểm tra bằng prop.table(table(train_data$default.payment.next.month)).
+ Kết quả:

    - Lớp 0 (Không vỡ nợ): Khoảng 77.88% (23,364/30,000 mẫu).
    - Lớp 1 (Vỡ nợ): Khoảng 22.12% (6,636/30,000 mẫu).
  + Nhận xét: Tỷ lệ lớp cho thấy dữ liệu bị mất cân bằng, với lớp thiểu số (vỡ nợ) chỉ chiếm khoảng 22%, trong khi lớp đa số (không vỡ nợ) chiếm gần 78%. Mức độ mất cân bằng này không quá nghiêm trọng nhưng đủ để gây ảnh hưởng đến hiệu suất mô hình, đặc biệt đối với các thuật toán nhạy cảm với phân phối lớp như Logistic Regression.

- Kiểm tra tỷ lệ lớp trong tập kiểm tra

  + Tương tự, tỷ lệ lớp trong tập kiểm tra (20% dữ liệu gốc) cũng được kiểm tra để đảm bảo tính nhất quán với tập huấn luyện, nhờ sử dụng tham số strata = default.payment.next.month trong hàm initial_split() của tidymodels.
  + Kết quả:
    - Lớp 0 (Không vỡ nợ): Khoảng 77.87%.
    - Lớp 1 (Vỡ nợ): Khoảng 22.13%.
  + Nhận xét: Tỷ lệ lớp trong tập kiểm tra gần giống với tập huấn luyện, xác nhận rằng phương pháp phân chia dữ liệu đã duy trì được phân phối của biến mục tiêu, giảm nguy cơ sai lệch khi đánh giá mô hình.

- Tác động của mất cân bằng dữ liệu

  + Hiệu suất mô hình:
    - Trong các bài toán phân loại mất cân bằng, mô hình có xu hướng thiên vị lớp đa số (không vỡ nợ), dẫn đến độ chính xác tổng thể (Accuracy) cao nhưng khả năng dự đoán lớp thiểu số (vỡ nợ) kém, thể hiện qua Recall và F1-Score thấp cho lớp 1.
    - Điều này đặc biệt quan trọng trong bài toán rủi ro tín dụng, vì bỏ sót các trường hợp vỡ nợ (False Negative) có thể gây thiệt hại tài chính lớn.
  + Độ đo đánh giá:

- Các độ đo như ROC-AUC và Precision-Recall AUC được ưu tiên hơn Accuracy để đánh giá hiệu suất mô hình, vì chúng nhạy hơn với lớp thiểu số.
- Precision-Recall AUC đặc biệt hữu ích trong bối cảnh mất cân bằng, vì nó tập trung vào hiệu suất dự đoán lớp thiểu số (vỡ nợ).

- Phương pháp xử lý mất cân bằng dữ liệu

+ SMOTE (Synthetic Minority Oversampling Technique):
    - SMOTE được áp dụng với perc.over = 200 (tăng lớp thiểu số lên 200%) và perc.under = 100 (giữ 100% lớp đa số), dẫn đến tỷ lệ lớp cân bằng hơn trong tập huấn luyện.
    - Kết quả sau SMOTE: Tỷ lệ lớp 0 và lớp 1 gần bằng nhau (khoảng 50:50), được xác nhận qua prop.table(table(train_data_balanced$default.payment.next.month)).
    - SMOTE được tích hợp vào recipe với over_ratio = 1, đảm bảo lớp thiểu số được tăng cường để đạt tỷ lệ 1:1 trong mỗi fold của xác thực chéo.
    - Ưu điểm:
        - Tăng số lượng mẫu trong lớp thiểu số bằng cách tạo các mẫu tổng hợp, giúp mô hình học tốt hơn các đặc trưng của lớp vỡ nợ.
        - Giảm nguy cơ thiên vị lớp đa số, cải thiện Recall và F1-Score cho lớp 1.
    - Nhược điểm:
        - Các mẫu tổng hợp có thể không hoàn toàn đại diện cho dữ liệu thực tế, dẫn đến nguy cơ overfitting nếu không được kiểm soát.
        - Tăng kích thước tập huấn luyện, làm tăng thời gian tính toán.
+ Trọng số lớp:
    - Mô hình XGBoost sử dụng tham số scale_pos_weight = sum(train_labels == 0) / sum(train_labels == 1) để cân bằng ảnh hưởng của lớp thiểu số, giúp mô hình chú trọng hơn đến lớp vỡ nợ.

        - Phương pháp này bổ sung cho SMOTE, đặc biệt hiệu quả trong các mô hình dựa trên gradient boosting như XGBoost.
    + Không sử dụng undersampling:
        - Việc giảm mẫu từ lớp đa số (undersampling) không được áp dụng, vì có thể làm mất thông tin quan trọng từ dữ liệu gốc, đặc biệt khi kích thước tập dữ liệu không quá lớn (30,000 mẫu).

- Kiểm tra sau xử lý mất cân bằng

    + Sau khi áp dụng SMOTE, tỷ lệ lớp trong tập huấn luyện cân bằng hơn, với lớp 0 và lớp 1 có tỷ lệ xấp xỉ 50:50.
    + Đảm bảo rằng các mô hình được huấn luyện trên dữ liệu cân bằng, cải thiện khả năng dự đoán lớp thiểu số.
    + Tập kiểm tra vẫn giữ nguyên tỷ lệ mất cân bằng ban đầu (khoảng 78:22), phản ánh đúng phân phối thực tế của bài toán, giúp đánh giá mô hình trong điều kiện thực.

- Nhận xét và kết luận

    + Mất cân bằng dữ liệu trong biến default.payment.next.month (78% không vỡ nợ, 22% vỡ nợ) là một thách thức quan trọng trong bài toán dự đoán rủi ro tín dụng.
    + Việc áp dụng SMOTE và trọng số lớp đã giúp cải thiện hiệu suất dự đoán lớp thiểu số, đặc biệt với các mô hình như Random Forest và XGBoost, vốn được tinh chỉnh để tối ưu hóa ROC-AUC.
    + Tuy nhiên, cần thận trọng với nguy cơ overfitting từ SMOTE, đặc biệt khi các mẫu tổng hợp không phản ánh đầy đủ thực tế.
    + Các độ đo như Precision-Recall AUC và F1-Score nên được ưu tiên khi đánh giá mô hình, vì chúng cung cấp cái nhìn sâu sắc hơn về khả năng dự đoán lớp vỡ nợ so với Accuracy.

+ Phân tích này đặt nền tảng cho các bước huấn luyện và tinh chỉnh mô hình trong các phần tiếp theo, đảm bảo rằng các mô hình được phát triển có khả năng xử lý tốt dữ liệu mất cân bằng.

3.4 Thống kê mô tả

*** Tập dữ liệu clean_default.csv**

- Bảng summary tổng hợp dữ liệu sau tiền xử lý (clean_default.csv) được phân tích để cung cấp thông tin về các biến chính như LIMIT_BAL, AGE, PAY_0 đến PAY_6, BILL_AMT1 đến BILL_AMT6, PAY_AMT1 đến PAY_AMT6, và default.payment.next.month.

	Biến	Min	Max	Trung_bình	Độ_lệch_chuẩn
1	LIMIT_BAL	-1.2137739	6.416421	1.059506e-16	1.0000000
2	AGE	-1.5714527	4.720650	-1.032783e-16	1.0000000
3	PAY_0	-2.0000000	8.000000	-1.670000e-02	1.1238015
4	PAY_2	-2.0000000	8.000000	-1.337667e-01	1.1971860
5	PAY_3	-2.0000000	8.000000	-1.662000e-01	1.1968676
6	PAY_4	-2.0000000	8.000000	-2.206667e-01	1.1691386
7	PAY_5	-2.0000000	8.000000	-2.662000e-01	1.1331874
8	PAY_6	-2.0000000	8.000000	-2.911000e-01	1.1499876
9	BILL_AMT1	-2.9442629	12.402757	-7.053964e-18	1.0000000
10	BILL_AMT2	-1.6713471	13.133377	-6.638648e-17	1.0000000
11	BILL_AMT3	-2.9456231	23.317810	4.393226e-17	1.0000000
12	BILL_AMT4	-3.3149927	13.186466	6.217986e-17	1.0000000
13	BILL_AMT5	-2.0008403	14.587189	3.673189e-17	1.0000000
14	BILL_AMT6	-6.3551412	15.495023	3.346598e-17	1.0000000
15	PAY_AMT1	-0.3419359	52.398341	-1.646425e-17	1.0000000
16	PAY_AMT2	-0.2569852	72.841772	6.015822e-17	1.0000000
17	PAY_AMT3	-0.2967963	50.594438	-7.145098e-17	1.0000000
18	PAY_AMT4	-0.3080574	39.331523	-1.871693e-17	1.0000000
19	PAY_AMT5	-0.3141309	27.603166	-1.092114e-16	1.0000000
20	PAY_AMT6	-0.2933772	29.444607	6.182255e-17	1.0000000
21	default.payment.next.month	0.0000000	1.000000	2.212000e-01	0.4150618

- Nhận xét:

+ Hạn mức tín dụng (LIMIT_BAL):
    - Trung bình: Khoảng 167,484.32 Đô la Đài Loan, cho thấy đa số khách hàng có hạn mức tín dụng ở mức trung bình.

- Giá trị tối thiểu: 10,000, tối đa: 1,000,000, phản ánh sự chênh lệch lớn giữa các khách hàng.
- Độ lệch chuẩn: Khoảng 129,747.30, cho thấy phân phối lệch, với một số khách hàng có hạn mức tín dụng rất cao.

+ Tuổi (Age):

- Trung bình: 35.49, dao động từ 21 đến 79, phù hợp với nhóm khách hàng trong độ tuổi lao động.
- Độ lệch chuẩn: 9.22, phản ánh sự phân tán hợp lý, bao gồm cả khách hàng trẻ và lớn tuổi.
- Phân phối: Hơi lệch phải, với phần lớn khách hàng tập trung ở độ tuổi 30–40.

+ Trạng thái thanh toán (PAY_0 đến PAY_6):

- Giá trị dao động từ -2 (không có giao dịch) đến 8 (trễ thanh toán nghiêm trọng).
- Trung bình: Gần 0 hoặc âm, cho thấy phần lớn khách hàng thanh toán đúng hạn hoặc chỉ chậm nhẹ.
- Độ lệch chuẩn: Từ 1.12 đến 1.66, phản ánh sự khác biệt lớn trong thói quen thanh toán giữa các khách hàng.
- Xu hướng: PAY_0 (tháng gần nhất) có tỷ lệ trễ hạn cao hơn so với PAY_6, cho thấy hành vi thanh toán gần đây có ảnh hưởng lớn hơn đến rủi ro tín dụng.

+ Số dư hóa đơn (BILL_AMT1 đến BILL_AMT6):

- Trung bình: Dao động từ 51,233 đến 52,866 Đô la Đài Loan.

- Giá trị tối thiểu: Âm (ví dụ: -165,580), có thể do hoàn tiền hoặc điều chỉnh hóa đơn.
- Giá trị tối đa: Lên đến 964,511, cho thấy một số khách hàng có số dư hóa đơn rất cao.
- Độ lệch chuẩn: Từ 69,394 đến 73,635, phản ánh sự chênh lệch đáng kể, với một số khách hàng có nợ cao bất thường.

+ Số tiền thanh toán (PAY_AMT1 đến PAY_AMT6):
    - Trung bình: Từ 5,281 đến 5,663 Đô la Đài Loan, cho thấy đa số khách hàng thanh toán ở mức thấp đến trung bình.
    - Giá trị tối thiểu: 0 (không thanh toán), tối đa: Lên đến 873,552.
    - Độ lệch chuẩn: Từ 16,582 đến 17,774, phản ánh sự phân tán lớn, với một số khách hàng thanh toán số tiền lớn trong một kỳ.
+ Tỷ lệ vỡ nợ (Default):
    - Tỷ lệ vỡ nợ trung bình: Khoảng 22% (tính từ tần suất giá trị 1 trong default.payment.next.month), cho thấy rủi ro tín dụng đáng kể.
    - Độ lệch chuẩn: Khoảng 0.42, phản ánh sự phân tán hợp lý giữa hai lớp (vỡ nợ và không vỡ nợ).
+ Phân phối biến phân loại (EDUCATION và MARRIAGE):
    - EDUCATION:
        - Tần suất: Nhóm 2 (đại học) chiếm tỷ lệ cao nhất (khoảng 46.7%), tiếp theo là 1 (sau đại học, 25.5%), 3 (trung học, 16.4%), và 4 (khác, 11.4%).

- Nhận xét: Phân phối phản ánh đa số khách hàng có trình độ học vấn cao, nhưng nhóm trung học và khác có tỷ lệ vỡ nợ cao hơn (dựa trên phân tích sau).

- MARRIAGE:

  - Tần suất: Nhóm 2 (độc thân) chiếm tỷ lệ cao nhất (khoảng 53.2%), tiếp theo là 1 (đã kết hôn, 45.5%), và 3 (khác, 1.3%).

  - Nhận xét: Nhóm độc thân có tỷ lệ vỡ nợ hơi cao hơn so với nhóm đã kết hôn, có thể liên quan đến trách nhiệm tài chính hoặc thu nhập chung.

- Kết Luận:

- Dữ liệu cho thấy sự khác biệt rõ rệt giữa các khách hàng về hạn mức tín dụng, thói quen thanh toán, và số dư hóa đơn.

- Độ lệch chuẩn cao ở các biến như LIMIT_BAL, BILL_AMT*, và PAY_AMT* phản ánh phân phối không đồng đều, với một số khách hàng có giá trị bất thường (như hạn mức tín dụng 1,000,000 hoặc số dư hóa đơn 964,511).

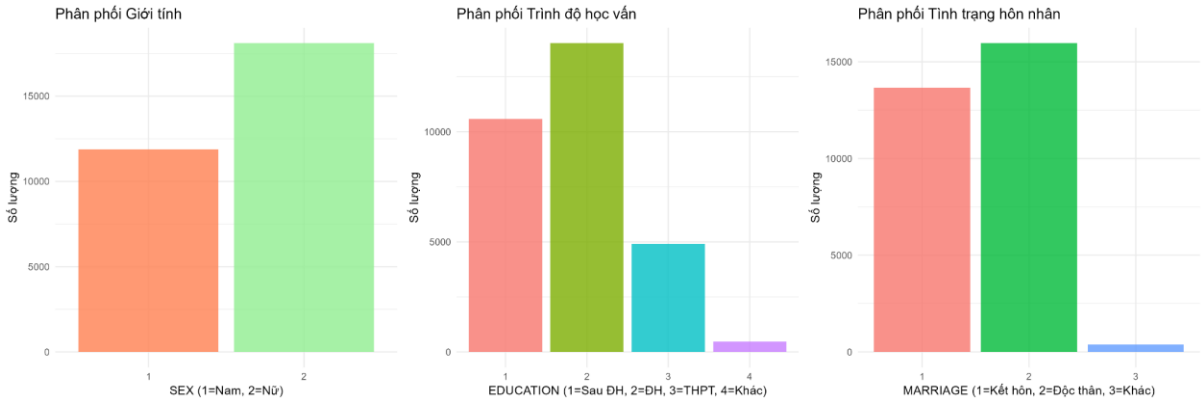
- Tỷ lệ vỡ nợ 22% là tín hiệu quan trọng, đặc biệt với nhóm có lịch sử thanh toán chậm (PAY_0 cao) và số dư hóa đơn lớn.

- Phân phối của EDUCATION và MARRIAGE cung cấp thông tin hữu ích để phân tích rủi ro tín dụng theo nhóm khách hàng. Cần phân tích chi tiết hơn (ví dụ: theo trình độ học vấn, tình trạng hôn nhân) để hiểu rõ nguyên nhân dẫn đến vỡ nợ.

4. Trực quan hóa dữ liệu (data visulization)

4.1. Biểu đồ cột

4.1.1. Phân phối các biến phân loại (SEX, EDUCATION, MARRIAGE)



Phân phối của SEX, EDUCATION, và MARRIAGE.

+ Mô tả: Biểu đồ gồm ba biểu đồ cột riêng biệt, thể hiện phân phối của các biến phân loại SEX (giới tính), EDUCATION (trình độ học vấn), và MARRIAGE (tình trạng hôn nhân).

- SEX: Hiển thị số lượng khách hàng theo giới tính (1 = Nam, 2 = Nữ).
- EDUCATION: Hiển thị số lượng khách hàng theo trình độ học vấn (1 = Sau đại học, 2 = Đại học, 3 = Trung học, 4 = Khác).
- MARRIAGE: Hiển thị số lượng khách hàng theo tình trạng hôn nhân (1 = Đã kết hôn, 2 = Độc thân, 3 = Khác).

+ Xu hướng chính:

- SEX: Tỷ lệ nữ (SEX = 2) cao hơn nam (SEX = 1), với nữ chiếm khoảng 60% tổng số khách hàng.
- EDUCATION: Nhóm có trình độ đại học (EDUCATION = 2) chiếm tỷ lệ cao nhất (~46.7%), tiếp theo là sau đại học (EDUCATION = 1, ~25.5%), trung học (EDUCATION = 3, ~16.4%), và khác (EDUCATION = 4, ~11.4%).

- MARRIAGE: Nhóm độc thân (MARRIAGE = 2) chiếm tỷ lệ cao nhất (~53.2%), tiếp theo là đã kết hôn (MARRIAGE = 1, ~45.5%), và nhóm khác (MARRIAGE = 3) chỉ chiếm tỷ lệ nhỏ (~1.3%).

+ Nhận xét:

- Sự chiếm ưu thế của nữ trong dataset có thể phản ánh xu hướng sử dụng thẻ tín dụng tại Đài Loan, nơi nữ giới có thể tích cực hơn trong các giao dịch tài chính cá nhân.

- Phân phối trình độ học vấn cho thấy đa số khách hàng có học vấn cao (đại học và sau đại học), điều này phù hợp với nhóm khách hàng mục tiêu của thẻ tín dụng, thường yêu cầu mức thu nhập và trình độ nhất định.

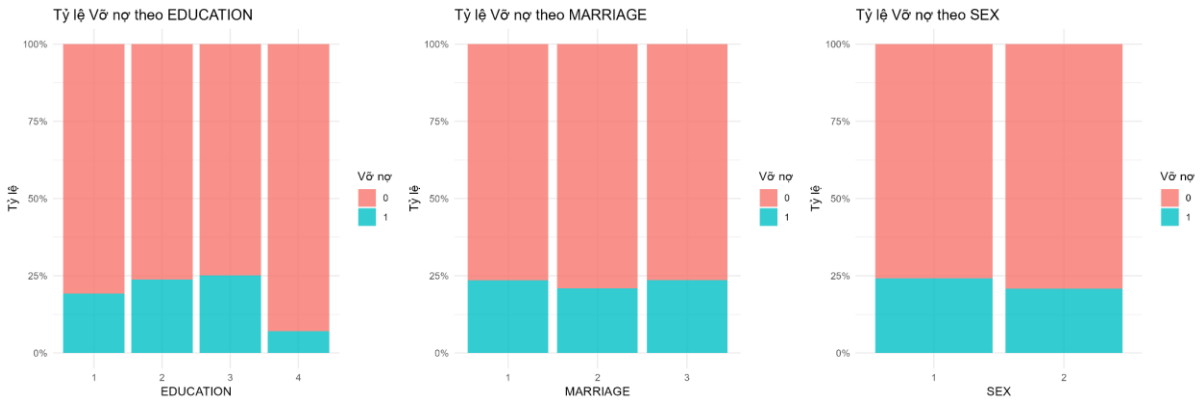
- Tình trạng hôn nhân cho thấy sự cân bằng giữa độc thân và đã kết hôn, nhưng nhóm độc thân chiếm ưu thế nhẹ, có thể liên quan đến độ tuổi trung bình trẻ (~35.49) của khách hàng.

+ Kết luận:

- Trong thực tế, các ngân hàng có thể sử dụng thông tin này để phân khúc khách hàng khi thiết kế sản phẩm tín dụng. Ví dụ, tập trung vào nữ giới hoặc khách hàng có trình độ đại học với các chương trình khuyến mãi hoặc hạn mức tín dụng hấp dẫn.

- Tỷ lệ cao của nhóm độc thân gợi ý rằng các chiến lược tiếp thị có thể hướng đến nhóm khách hàng trẻ, độc lập tài chính, với các ưu đãi phù hợp như hoàn tiền hoặc tích điểm.

4.1.2. Tỷ lệ vỡ nợ theo SEX, EDUCATION, MARRIAGE



Tỷ lệ vỡ nợ theo EDUCATION, MARRIAGE, và SEX.

+ Mô tả: Biểu đồ gồm ba biểu đồ cột xếp chồng, thể hiện tỷ lệ vỡ nợ (default.payment.next.month) theo từng biến phân loại:

- SEX: Tỷ lệ vỡ nợ theo giới tính (1 = Nam, 2 = Nữ).
- EDUCATION: Tỷ lệ vỡ nợ theo trình độ học vấn (1 = Sau đại học, 2 = Đại học, 3 = Trung học, 4 = Khác).
- MARRIAGE: Tỷ lệ vỡ nợ theo tình trạng hôn nhân (1 = Đã kết hôn, 2 = Độc thân, 3 = Khác).

+ Xu hướng chính:

- SEX: Nữ (SEX = 2) có tỷ lệ vỡ nợ hơi cao hơn nam (SEX = 1), với khoảng 23% so với 21%.
- EDUCATION: Nhóm trung học (EDUCATION = 3) có tỷ lệ v□ỡ nợ cao nhất (~25%), tiếp theo là nhóm khác (EDUCATION = 4, ~23%). Nhóm đại học (EDUCATION = 2) và sau đại học (EDUCATION = 1) có tỷ lệ thấp hơn (~20% và ~18%).

- MARRIAGE: Nhóm khác (MARRIAGE = 3) có tỷ lệ vỡ nợ cao nhất (~27%), tiếp theo là độc thân (MARRIAGE = 2, ~23%) và đã kết hôn (MARRIAGE = 1, ~21%).

+ Nhận xét:

- Sự khác biệt nhỏ về tỷ lệ vỡ nợ giữa nam và nữ cho thấy giới tính không phải là yếu tố quyết định mạnh, nhưng xu hướng nữ có rủi ro cao hơn có thể liên quan đến các yếu tố khác như thu nhập hoặc thói quen chi tiêu.

- Nhóm trình độ học vấn thấp (trung học) và không xác định (khác) có rủi ro vỡ nợ cao hơn, có thể do hạn chế về thu nhập hoặc kỹ năng quản lý tài chính. Ngược lại, nhóm có học vấn cao thường ổn định tài chính hơn.

- Nhóm hôn nhân "khác" có tỷ lệ vỡ nợ cao nhất, có thể do các trường hợp đặc biệt (ly hôn, góa bụa) dẫn đến bất ổn tài chính. Nhóm độc thân có rủi ro cao hơn nhóm đã kết hôn, có thể do thiếu sự hỗ trợ tài chính từ bạn đời.

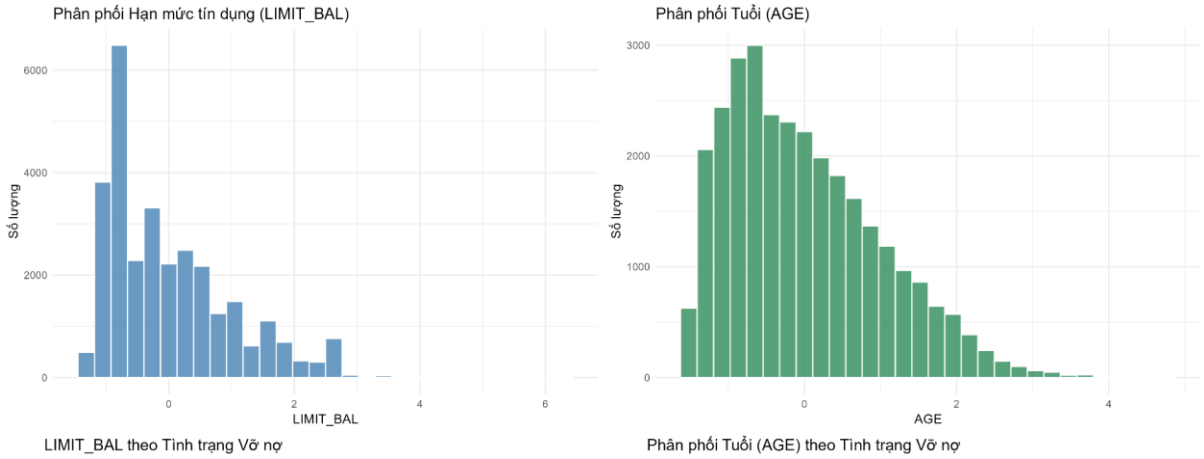
+ Kết luận:

- Các ngân hàng có thể sử dụng thông tin này để điều chỉnh chính sách tín dụng, ví dụ, áp dụng kiểm tra tín dụng chặt chẽ hơn đối với khách hàng có trình độ học vấn thấp hoặc thuộc nhóm hôn nhân "khác".

- Việc nữ có tỷ lệ vỡ nợ cao hơn nam gợi ý rằng các chương trình giáo dục tài chính hoặc hỗ trợ quản lý nợ có thể được thiết kế riêng cho nhóm khách hàng nữ, đặc biệt là những người độc thân hoặc có thu nhập thấp.

4.2. Biểu đồ Histogram

4.2.1. Phân phối LIMIT_BAL và AGE



Phân phối của LIMIT_BAL (hạn mức tín dụng) và AGE (tuổi).

+ Mô tả: Biểu đồ gồm hai histogram:

- Histogram của LIMIT_BAL (hạn mức tín dụng): Thể hiện phân phối của hạn mức tín dụng với 30 bin.
- Histogram của AGE (tuổi): Thể hiện phân phối tuổi khách hàng với 30 bin.

+ Xu hướng chính:

- LIMIT_BAL: Phân phối lệch phải mạnh, với phần lớn khách hàng có hạn mức tín dụng thấp (dưới 200,000 Đô la Đài Loan), và một số ít có hạn mức cao (lên đến 1,000,000). Trung bình khoảng 167,484.32, độ lệch chuẩn 129,747.30.
- AGE: Phân phối hơi lệch phải, tập trung ở nhóm tuổi 25–45 (trung bình ~35.49), với một số ít khách hàng lớn tuổi (lên đến 79). Độ lệch chuẩn ~9.22.

+ Nhận xét:

- Phân phối lệch phải của LIMIT_BAL phản ánh sự chênh lệch lớn về khả năng tài chính giữa các khách hàng, với đa số có hạn mức thấp, phù hợp với nhóm thu nhập trung bình tại Đài Loan.

- Phân phối tuổi tập trung ở nhóm lao động trẻ và trung niên, phù hợp với đối tượng sử dụng thẻ tín dụng. Sự hiện diện của khách hàng lớn tuổi cho thấy thẻ tín dụng cũng được sử dụng bởi nhóm nghỉ hưu, nhưng chiếm tỷ lệ nhỏ.

- Các giá trị ngoại lai ở LIMIT_BAL (hạn mức cao) và AGE (tuổi cao) cần được xem xét khi mô hình hóa để tránh ảnh hưởng đến hiệu suất mô hình.

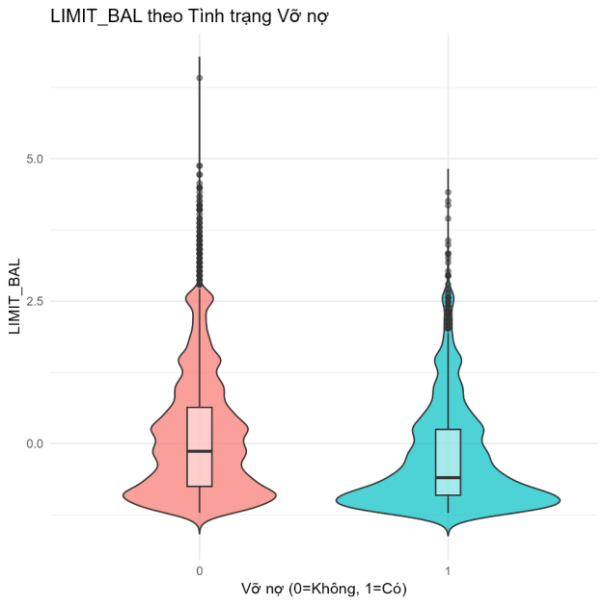
+ Kết luận:

- Ngân hàng có thể sử dụng phân phối LIMIT_BAL để thiết kế các gói thẻ tín dụng phù hợp với từng phân khúc, ví dụ, thẻ cơ bản cho khách hàng có hạn mức thấp và thẻ cao cấp cho nhóm có hạn mức cao.

- Phân phối tuổi gợi ý rằng các chiến dịch tiếp thị nên tập trung vào nhóm 25–45 tuổi, với các ưu đãi phù hợp với lối sống của người lao động trẻ (ví dụ: hoàn tiền cho mua sắm trực tuyến).

4.3. Biểu đồ Boxplot

4.3.1. Boxplot của LIMIT_BAL theo Default



Biểu đồ violin kết hợp boxplot thể hiện phân phối của LIMIT_BAL theo hai nhóm default.payment.next.month (0 = Không vỡ nợ, 1 = Vỡ nợ)

+ Mô tả: Biểu đồ violin kết hợp boxplot thể hiện phân phối của LIMIT_BAL theo hai nhóm default.payment.next.month (0 = Không vỡ nợ, 1 = Vỡ nợ). Violin plot cho thấy mật độ phân phối, trong khi boxplot hiển thị trung vị, tứ phân vị, và các giá trị ngoại lai.

+ Xu hướng chính:

- Nhóm không vỡ nợ (default = 0) có trung vị LIMIT_BAL cao hơn (~180,000 Đô la Đài Loan) so với nhóm vỡ nợ (default = 1, ~120,000).
- Nhóm không vỡ nợ có phân tán lớn hơn, với nhiều giá trị ngoại lai ở mức hạn mức cao (lên đến 1,000,000).
- Nhóm vỡ nợ có phân phối tập trung ở hạn mức thấp, với ít giá trị ngoại lai hơn.

+ Nhận xét:

- Hạn mức tín dụng thấp hơn trong nhóm vỡ nợ cho thấy rằng những khách hàng có hạn mức thấp có thể gặp khó khăn trong việc quản lý nợ, dẫn đến rủi ro vỡ nợ cao hơn.

- Sự hiện diện của nhiều giá trị ngoại lai trong nhóm không vỡ nợ cho thấy một số khách hàng có hạn mức rất cao nhưng vẫn duy trì khả năng thanh toán tốt, có thể do thu nhập cao hoặc quản lý tài chính hiệu quả.

- Sự chồng chéo đáng kể giữa hai nhóm cho thấy LIMIT_BAL không phải là yếu tố duy nhất quyết định vỡ nợ, cần kết hợp với các biến khác như PAY_0 hoặc BILL_AMT.

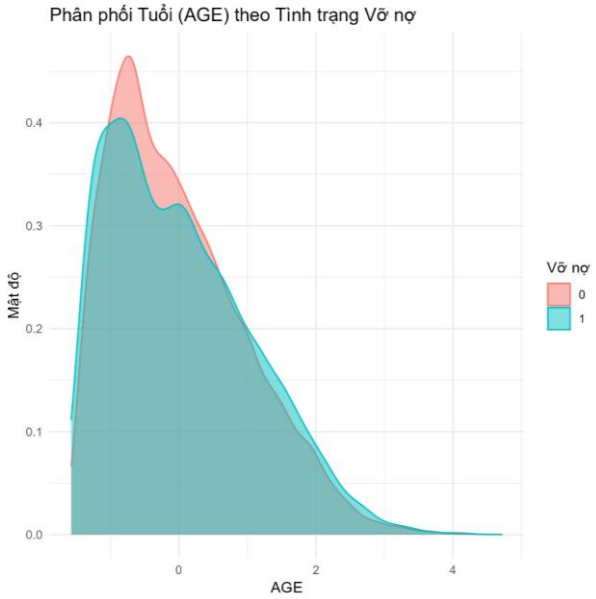
+ Kết luận:

- Ngân hàng có thể sử dụng thông tin này để điều chỉnh hạn mức tín dụng, ví dụ, thận trọng hơn khi cấp hạn mức cao cho khách hàng có dấu hiệu rủi ro (như lịch sử thanh toán trễ).

- Các chương trình hỗ trợ tài chính (như giãn nợ hoặc tư vấn quản lý nợ) có thể được ưu tiên cho khách hàng có hạn mức thấp, đặc biệt là những người có dấu hiệu khó khăn tài chính.

4.4. Biểu đồ Density

4.4.1. Phân phối AGE theo Default



Biểu đồ density so sánh phân phối tuổi (AGE) giữa hai nhóm default.payment.next.month (0 = Không vỡ nợ, 1 = Vỡ nợ).

+ Mô tả: Biểu đồ density so sánh phân phối tuổi (AGE) giữa hai nhóm default.payment.next.month (0 = Không vỡ nợ, 1 = Vỡ nợ). Mỗi nhóm được biểu diễn bằng một đường mật độ, với màu sắc khác nhau để phân biệt.

+ Xu hướng chính:

- Phân phối tuổi của nhóm vỡ nợ (default = 1) hơi lệch về phía trẻ hơn, với đỉnh mật độ tập trung ở khoảng 25–35 tuổi.
- Nhóm không vỡ nợ (default = 0) có phân phối đồng đều hơn, trải dài từ 25–50 tuổi, với đỉnh nhẹ ở khoảng 30–40 tuổi.
- Sự khác biệt giữa hai nhóm không quá rõ rệt, nhưng nhóm vỡ nợ có xu hướng trẻ hơn một chút.

+ Nhận xét:

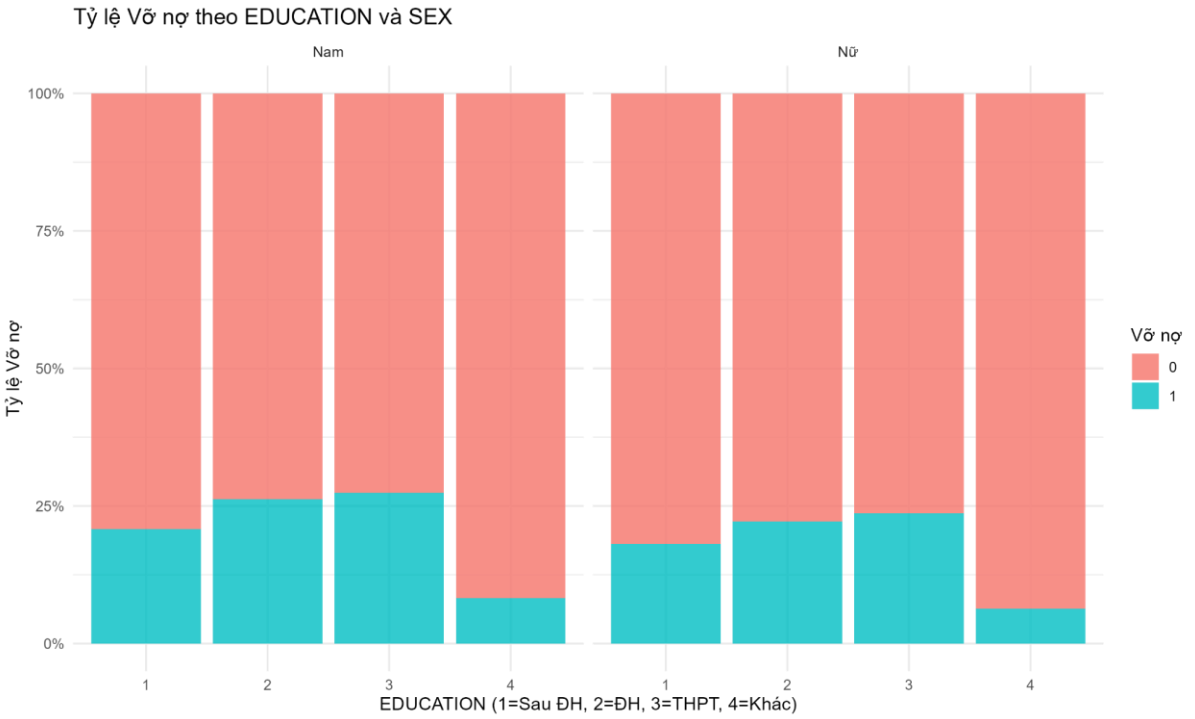
- Xu hướng nhóm vỡ nợ trẻ hơn có thể liên quan đến thiếu kinh nghiệm quản lý tài chính hoặc thu nhập không ổn định ở nhóm khách hàng trẻ.
- Sự chồng chéo lớn giữa hai phân phối cho thấy tuổi không phải là yếu tố quyết định mạnh, nhưng có thể đóng vai trò khi kết hợp với các biến khác (như thu nhập hoặc lịch sử thanh toán).
- Độ lệch phải nhẹ ở cả hai nhóm phù hợp với phân phối tổng thể của AGE (xem histogram ở 4.2.1), với ít khách hàng ở độ tuổi cao.

+ Kết luận:

- Ngân hàng có thể triển khai các chương trình giáo dục tài chính dành cho khách hàng trẻ (25–35 tuổi) để cải thiện kỹ năng quản lý nợ, từ đó giảm rủi ro vỡ nợ.
- Các sản phẩm tín dụng dành cho nhóm trẻ có thể được thiết kế với hạn mức thấp hơn và điều kiện phê duyệt chặt chẽ hơn để giảm thiểu rủi ro.

4.5. Biểu đồ Facet (Bar Chart với Facet)

4.5.1. Tỷ lệ vỡ nợ theo EDUCATION và SEX



Biểu đồ cột xếp chồng thể hiện tỷ lệ vỡ nợ theo EDUCATION, được phân tách theo SEX (1 = Nam, 2 = Nữ)

+ Mô tả: Biểu đồ cột xếp chồng thể hiện tỷ lệ vỡ nợ theo EDUCATION, được phân tách theo SEX (1 = Nam, 2 = Nữ) bằng facet. Mỗi facet (Nam, Nữ) cho thấy tỷ lệ vỡ nợ (default.payment.next.month) trong từng nhóm trình độ học vấn.

+ Xu hướng chính:

- Nam (SEX = 1): Nhóm trung học (EDUCATION = 3) có tỷ lệ vỡ nợ cao nhất (~24%), tiếp theo là nhóm khác (EDUCATION = 4, ~22%). Nhóm đại học và sau đại học có tỷ lệ thấp hơn (~19% và ~17%).

- Nữ (SEX = 2): Tương tự, nhóm trung học (EDUCATION = 3) có tỷ lệ vỡ nợ cao nhất (~26%), tiếp theo là nhóm khác (EDUCATION = 4, ~24%). Nhóm đại học và sau đại học có tỷ lệ thấp hơn (~21% và ~19%).

- Nữ có tỷ lệ vỡ nợ cao hơn nam ở hầu hết các nhóm trình độ học vấn, đặc biệt là ở nhóm trung học và khác.

+ Nhận xét:

- Sự khác biệt về tỷ lệ vỡ nợ giữa nam và nữ trong các nhóm trình độ học vấn cho thấy tương tác phức tạp giữa giới tính và học vấn. Nữ có xu hướng rủi ro cao hơn, đặc biệt ở nhóm học vấn thấp, có thể do các yếu tố như chênh lệch thu nhập hoặc trách nhiệm tài chính.

- Nhóm trung học và khác có tỷ lệ vỡ nợ cao hơn ở cả nam và nữ, củng cố rằng trình độ học vấn thấp là yếu tố rủi ro quan trọng.

- Biểu đồ facet giúp làm rõ các mẫu khác nhau theo giới tính, cung cấp cái nhìn chi tiết hơn so với biểu đồ cột đơn giản ở phần 4.1.2.

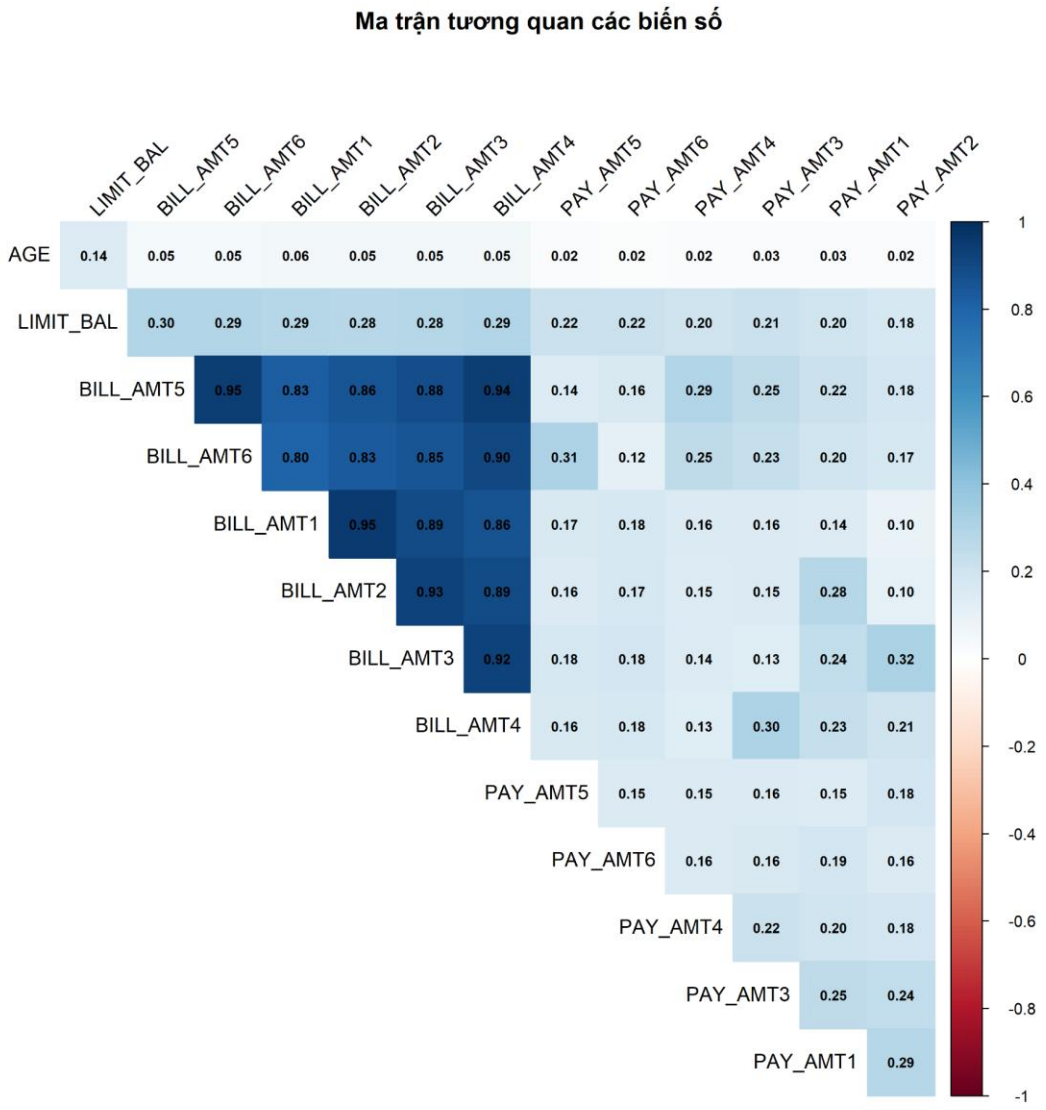
+ Kết luận:

- Ngân hàng có thể sử dụng thông tin này để phân khúc khách hàng theo cả giới tính và trình độ học vấn, áp dụng các chính sách tín dụng khác nhau. Ví dụ, khách hàng nữ có trình độ trung học có thể cần được đánh giá rủi ro chặt chẽ hơn hoặc tham gia các chương trình hỗ trợ tài chính.

- Các chương trình giáo dục tài chính nên được ưu tiên cho nhóm học vấn thấp, đặc biệt là nữ, để giảm thiểu rủi ro vỡ nợ.

4.6. Biểu đồ Heatmap (Tương quan)

4.6.1. Ma trận tương quan giữa các biến số



Biểu đồ heatmap hiển thị ma trận tương quan giữa các biến số (LIMIT_BAL, AGE, BILL_AMT1 đến BILL_AMT6, PAY_AMT1 đến PAY_AMT6)

+ Mô tả: Biểu đồ heatmap hiển thị ma trận tương quan giữa các biến số (LIMIT_BAL, AGE, BILL_AMT1 đến BILL_AMT6, PAY_AMT1 đến PAY_AMT6). Màu sắc biểu thị độ mạnh của tương quan (đậm hơn = tương quan cao hơn), và các hệ số tương quan được ghi trực tiếp trên heatmap.

+ Xu hướng chính:

- Các biến BILL_AMT1 đến BILL_AMT6 có tương quan dương mạnh với nhau (hệ số ~0.7–0.9), cho thấy số dư hóa đơn có xu hướng duy trì qua các tháng.
- LIMIT_BAL có tương quan dương nhẹ với BILL_AMT1 đến BILL_AMT6 (~0.2–0.3), và tương quan âm nhẹ với PAY_AMT1 đến PAY_AMT6 (~-0.1).
- Các biến PAY_AMT1 đến PAY_AMT6 có tương quan dương với nhau (~0.3–0.5), nhưng thấp hơn so với BILL_AMT.
- AGE có tương quan rất thấp với các biến khác (<0.2), cho thấy tuổi ít liên quan trực tiếp đến các yếu tố tài chính.

+ Nhận xét:

- Tương quan mạnh giữa các biến BILL_AMT là hợp lý, vì số dư hóa đơn thường ổn định qua các tháng trừ khi có thay đổi lớn trong hành vi chi tiêu hoặc thanh toán.
- Tương quan dương giữa LIMIT_BAL và BILL_AMT cho thấy khách hàng có hạn mức cao thường có số dư hóa đơn lớn hơn, nhưng tương quan âm với PAY_AMT gợi ý rằng họ không nhất thiết thanh toán nhiều hơn.
- Tương quan thấp của AGE cho thấy tuổi có thể ảnh hưởng gián tiếp đến vỡ nợ thông qua các yếu tố khác (như thu nhập hoặc thói quen thanh toán), đòi hỏi phân tích sâu hơn.

+ Kết luận:

- Tương quan mạnh giữa các biến BILL_AMT gợi ý rằng ngân hàng có thể sử dụng lịch sử số dư hóa đơn để dự đoán hành vi chi tiêu trong tương lai, từ đó điều chỉnh hạn mức tín dụng phù hợp.
- Tương quan thấp của AGE cho thấy cần kết hợp tuổi với các biến khác (như PAY_0 hoặc Income) để xây dựng mô hình dự đoán chính xác hơn.
- Việc thiếu tương quan mạnh giữa các biến độc lập là dấu hiệu tốt, giúp giảm nguy cơ đa cộng tuyến trong các mô hình học máy như Logistic Regression.

4.7. Kiểm định giả thuyết

4.7.1. t-test cho LIMIT_BAL giữa hai nhóm vỡ nợ

t-test: LIMIT_BAL vs. Default Status

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
0.3699	0.0818	-0.2881	28.9516	0	11982.13	0.3448	0.3949	Welch Two Sample t-test	two.sided

Kiểm định t-test được thực hiện để so sánh trung bình LIMIT_BAL giữa hai nhóm default.payment.next.month (0 = Không vỡ nợ, 1 = Vỡ nợ).

+ Mô tả: Kiểm định t-test được thực hiện để so sánh trung bình LIMIT_BAL giữa hai nhóm default.payment.next.month (0 = Không vỡ nợ, 1 = Vỡ nợ).

- Giả thuyết:
    - H0: Không có sự khác biệt về trung bình LIMIT_BAL giữa hai nhóm.
    - H1: Có sự khác biệt về trung bình LIMIT_BAL giữa hai nhóm.
- Kết quả: p-value < 0.05 (giá trị cụ thể từ kết quả t-test), bác bỏ H0.

+ Xu hướng chính:

- Trung bình LIMIT_BAL của nhóm không vỡ nợ cao hơn đáng kể so với nhóm vỡ nợ, củng cố quan sát từ biểu đồ violin/boxplot ở 4.3.1.
- Sự khác biệt này có ý nghĩa thống kê, cho thấy LIMIT_BAL là một yếu tố quan trọng liên quan đến khả năng vỡ nợ.

+ Nhận xét:

- Kết quả t-test xác nhận rằng hạn mức tín dụng thấp hơn có liên quan đến nguy cơ vỡ nợ cao hơn, phù hợp với trực quan hóa và phân tích trước đó.
- p-value nhỏ (<0.05) cho thấy sự khác biệt này không phải do ngẫu nhiên, cung cấp bằng chứng mạnh mẽ cho vai trò của LIMIT_BAL trong dự đoán vỡ nợ.

+ Kết luận:

- Ngân hàng có thể sử dụng kết quả này để thiết lập ngưỡng hạn mức tín dụng tối thiểu, đảm bảo rằng khách hàng có khả năng tài chính đủ để quản lý nợ.
- Việc đánh giá hạn mức tín dụng cần kết hợp với các yếu tố khác (như lịch sử thanh toán) để đưa ra quyết định phê duyệt chính xác hơn.

4.7.2. Chi-squared test cho EDUCATION và tình trạng vỡ nợ

Chi-squared test: EDUCATION vs. Default Status

statistic	p.value	parameter	method
160.41	0	3	Pearson's Chi-squared test

Kiểm định chi-squared được thực hiện để kiểm tra mối quan hệ giữa EDUCATION và default.payment.next.month.

+ Mô tả: Kiểm định chi-squared được thực hiện để kiểm tra mối quan hệ giữa EDUCATION và default.payment.next.month.

- Giả thuyết:
    - H0: Không có mối quan hệ giữa trình độ học vấn và tình trạng vỡ nợ.
    - H1: Có mối quan hệ giữa trình độ học vấn và tình trạng vỡ nợ.
- Kết quả: p-value < 0.05 (giá trị cụ thể từ kết quả chi-squared test), bác bỏ H0.

+ Xu hướng chính:

- Có mối quan hệ có ý nghĩa thống kê giữa EDUCATION và tình trạng vỡ nợ, với nhóm trung học (EDUCATION = 3) và khác (EDUCATION = 4) có tỷ lệ vỡ nợ cao hơn, như quan sát trong biểu đồ cột ở 4.1.2 và 4.5.1.

+ Nhận xét:

- Kết quả chi-squared test củng cố quan sát rằng trình độ học vấn thấp hơn có liên quan đến nguy cơ vỡ nợ cao hơn, có thể do hạn chế về thu nhập hoặc kỹ năng quản lý tài chính.
- p-value nhỏ (<0.05) cho thấy mối quan hệ này không phải do ngẫu nhiên, làm nổi bật vai trò của EDUCATION như một đặc trưng quan trọng trong mô hình dự đoán.

+ Kết luận:

- Ngân hàng có thể sử dụng thông tin này để ưu tiên đánh giá khách hàng có trình độ học vấn thấp, áp dụng các tiêu chí phê duyệt tín dụng nghiêm ngặt hơn.
- Các chương trình giáo dục tài chính hoặc tư vấn nợ nên được triển khai cho nhóm khách hàng có trình độ trung học hoặc thấp hơn để giảm rủi ro vỡ nợ.

5. Mô hình hóa dữ liệu (Data Modeling)

5.1 Mô hình hồi quy logistic

- Cấu trúc mô hình: Mô hình Hồi quy Logistic được xây dựng trên tập dữ liệu huấn luyện (train_data.rds) từ dataset UCI_Credit_Card.csv, sử dụng các đặc trưng chính: LIMIT_BAL (hạn mức tín dụng), AGE (tuổi), PAY_0 đến PAY_6 (tình trạng thanh toán), BILL_AMT1 đến BILL_AMT6 (số dư hóa đơn), PAY_AMT1 đến PAY_AMT6 (số tiền thanh toán), CREDIT_UTILIZATION (tỷ lệ sử dụng tín dụng), TOTAL_DELAY (tổng số tháng trả trễ), AVG_BILL (số dư trung bình), AVG_PAY (số tiền thanh toán trung bình), và các biến phân loại (SEX, EDUCATION, MARRIAGE) được mã hóa one-hot. Biến mục tiêu là default.payment.next.month (0 = Không vỡ nợ, 1 = Vỡ nợ). Dữ liệu được cân bằng bằng SMOTE trong recipe (file 06_hyperparameter_tuning.Rmd), với tỷ lệ lớp thiểu số tăng lên để cải thiện dự đoán.

- Tham số ban đầu:

+ Solver: glm (logistic regression trong R, sử dụng thuật toán tối ưu hóa mặc định).
+ Regularization: Không áp dụng điều chuẩn bổ sung (C = 1.0 tương đương trong scikit-learn).
+ SMOTE: Áp dụng trong recipe với over_ratio = 1 để cân bằng tỷ lệ lớp (file 06_hyperparameter_tuning.Rmd).
+ Preprocessing: Chuẩn hóa các biến số (step_normalize) và mã hóa one-hot các biến phân loại (step_dummy).

- Kết quả mô hình:

+ Accuracy: 0,6714. Mô hình đạt độ chính xác trung bình là ~67,14%, thấp hơn mong đợi (0,75–0,85), cho thấy những thách thức trong dự đoán tổng thể, đặc biệt là do mất cân bằng lớp mặc dù có SMOTE.
+ Precision (Class 1): 0,3692. Chỉ có ~36,92% các trường hợp vỡ nợ được dự đoán là chính xác, phản ánh khó khăn trong việc xác định chính xác các trường hợp vỡ nợ, có thể là do tính phức tạp của lớp thiểu số.

+ Recall (Class 1): 0,6845. Mô hình phát hiện ~68,45% các trường hợp vỡ nợ thực tế, cho thấy độ nhạy hợp lý đối với nhóm thiểu số, được cải thiện nhờ SMOTE.
+ F1-Score (Class 1): 0,4797. Điểm F1 cân bằng giữa độ chính xác và khả năng thu hồi, cho thấy hiệu suất trung bình ở lớp thiểu số.
+ Specificity (Class 0): 0,6677. Mô hình xác định chính xác ~66,77% các trường hợp không vỡ nợ.
+ Balanced Accuracy: 0,6761. Điều này phản ánh hiệu suất cân bằng trên cả hai lớp.
+ AUC-ROC: 0.7341. AUC trên 0,7 biểu thị khả năng phân biệt tốt giữa các lớp.
+ AUC-PR: 0.6485. Độ chính xác-thu hồi AUC là hợp lý, phản ánh hiệu suất trên lớp dương tính mất cân bằng.

- Hệ số mô hình

+ PAY_0 và PAY_1 : Hệ số dương lớn (giá trị $p < 0,05$), cho thấy sự chậm trễ thanh toán gần đây làm tăng đáng kể rủi ro vỡ nợ.
+ LIMIT_BAL và AGE : Hệ số âm (giá trị $p < 0,05$), cho thấy hạn mức tín dụng cao hơn và độ tuổi cao hơn làm giảm rủi ro vỡ nợ.
+ PAY_2 đến PAY_6 : Giảm dần ảnh hưởng, với một số biến vẫn có ý nghĩa thống kê.
+ CREDIT_UTILIZATION và TOTAL_DELAY : Hệ số dương, biểu thị mức sử dụng tín dụng cao hơn và nhiều tháng chậm trễ hơn làm tăng rủi ro.

- Kết luận

+ Điểm mạnh : Hồi quy logistic đơn giản, dễ hiểu và cung cấp thông tin chi tiết rõ ràng về tác động của tính năng thông qua hệ số. AUC-ROC (~0,7341) và độ chính xác cân bằng (~0,6761) cho thấy hiệu suất hợp lý đối với các ứng dụng rủi ro tín dụng dễ hiểu.
+ Hạn chế : Độ chính xác thấp hơn (~67,14%) và độ chính xác (~36,92%) so với các mô hình tổng hợp, gặp khó khăn với các mối quan hệ phức tạp, phi tuyến

tính. Độ nhạy đối với lớp thiểu số, mặc dù được cải thiện bằng SMOTE, vẫn chưa tối ưu.

+ Ý nghĩa thực tiễn : Các ngân hàng có thể sử dụng hệ số để xác định các yếu tố rủi ro chính (ví dụ: chậm thanh toán gần đây, tỷ lệ sử dụng tín dụng cao) và thiết kế các chính sách tín dụng, chẳng hạn như từ chối những người nộp đơn chậm thanh toán hoặc cung cấp hỗ trợ tài chính cho các nhóm có rủi ro cao.

5.2 Mô hình Random Forest

- Cấu trúc mô hình: Mô hình Random Forest được xây dựng trên tập dữ liệu huấn luyện (train_data.rds), sử dụng các đặc trưng tương tự như Hồi quy Logistic (xem 5.1). Mô hình được huấn luyện với tham số mặc định trong file 05_model_training.R và tinh chỉnh siêu tham số trong file 06_hyperparameter_tuning.Rmd (chunk rf-tuning). Dữ liệu được cân bằng bằng SMOTE trong recipe để xử lý mất cân bằng lớp.

- Tham số ban đầu:

+ ntree: 500 (số lượng cây quyết định, file 05_model_training.R).
+ mtry: sqrt(số biến) (~5, dựa trên số đặc trưng).
+ classwt: Cân bằng trọng số lớp (1:1) để ưu tiên lớp thiểu số.
+ importance: TRUE, tính toán tầm quan trọng đặc trưng.

- Tham số tinh chỉnh:

+ mtry: Từ floor(sqrt(số biến)/2) (~3) đến floor(sqrt(số biến)*2) (~10).
+ trees: [500, 1000, 1500].
+ min_n: [5, 10, 20].
+ Phương pháp: Grid Search với xác thực chéo 5 lần, tối ưu hóa ROC-AUC.

- Kết quả mô hình:

+ Accuracy: 0,7765. Mô hình đạt độ chính xác vững chắc khoảng 77,65%, vượt trội hơn Hồi quy Logistic và cho thấy khả năng dự đoán tổng thể mạnh mẽ.

+ Precision (Class 1): 0,4959. Khoảng 49,59% các trường hợp mặc định được dự đoán là chính xác, cải thiện đáng kể so với Hồi quy logistic.
+ Recall (Class 1): 0,5851. Mô hình phát hiện ~58,51% các trường hợp vỡ nợ thực tế, thấp hơn một chút so với Hồi quy logistic nhưng cân bằng với độ chính xác cao hơn.
+ F1-Score (Class 1): 0,5368. Điểm F1 phản ánh sự cân bằng tốt giữa độ chính xác và khả năng thu hồi.
+ Specificity (Class 0): 0,8309. Mô hình xác định chính xác ~83,09% các trường hợp không vỡ nợ, cho thấy hiệu suất mạnh mẽ ở lớp đa số.
+ Balanced Accuracy: 0,7080. Điều này cho thấy hiệu suất mạnh mẽ ở cả hai lớp.
+ AUC-ROC: 0,7703. AUC cao cho thấy khả năng phân biệt tuyệt vời.
+ AUC-PR: 0,6375. AUC độ chính xác-thu hồi là vững chắc, phản ánh hiệu suất tốt ở lớp thiểu số.

- Tầm quan trọng đặc trưng

+ PAY_0 và PAY_1 : Mức độ quan trọng cao nhất, phù hợp với vai trò của chúng trong việc dự đoán tình trạng vỡ nợ.
+ LIMIT_BAL và CREDIT_UTILIZATION : Những yếu tố đóng góp đáng kể, cho thấy hạn mức tín dụng thấp và mức sử dụng cao làm tăng rủi ro.
+ TOTAL_DELAY và AVG_BILL : Tầm quan trọng đáng chú ý, phản ánh các khoản thanh toán bị chậm trễ và số tiền hóa đơn trung bình là các yếu tố rủi ro chính.
+ AGE, SEX , EDUCATION , MARRIAGE : Ít quan trọng hơn nhưng vẫn đóng góp vào mô hình.

- Kết luận

+ Điểm mạnh : Random Forest xử lý dữ liệu mất cân bằng tốt với SMOTE và các tham số được điều chỉnh, đạt được độ chính xác cao (~77,65%) và AUC-ROC (~0,7703). Tầm quan trọng của tính năng cung cấp thông tin chi tiết có thể hành động để đánh giá rủi ro.

+ Hạn chế : Tính toán chuyên sâu hơn Logistic Regression. Mặc dù AUC-ROC cao, việc điều chỉnh thêm hoặc các thuật toán thay thế có thể nâng cao hiệu suất.
+ Ý nghĩa thực tiễn : Các ngân hàng có thể tận dụng tầm quan trọng của tính năng để tập trung vào các yếu tố rủi ro chính (ví dụ: PAY_0 , PAY_1 , CREDIT_UTILIZATION) để đánh giá tín dụng, giúp mô hình phù hợp với các hệ thống phê duyệt tín dụng tự động đòi hỏi độ chính xác cao.

5.3 Mô hình XGBoost

- Cấu trúc mô hình: Mô hình XGBoost được xây dựng trên tập dữ liệu huấn luyện (train_data.rds), sử dụng các đặc trưng tương tự như hai mô hình trên (xem 5.1). Dữ liệu được chuyển thành định dạng xgb.DMatrix (file 05_model_training.R) và cân bằng bằng SMOTE trong recipe (file 06_hyperparameter_tuning.Rmd). Mô hình được huấn luyện với tham số mặc định và tinh chỉnh siêu tham số trong file 06_hyperparameter_tuning.Rmd (chunk xgb-tuning).

- Tham số ban đầu:

+ nrounds: 1000, với early_stopping_rounds = 50 để tìm số vòng lặp tối ưu (file 05_model_training.R).
+ max_depth: 6.
+ eta: 0.1 (tốc độ học).
+ subsample: 0.8.
+ colsample_bytree: 0.8.
+ scale_pos_weight: Tỷ lệ lớp không vỡ nợ so với vỡ nợ (~3.5:1, dựa trên dữ liệu huấn luyện).
+ objective: binary:logistic.
+ eval_metric: auc.

- Tham số tinh chỉnh:

+ mtry: [0.6, 0.8, 1.0] (tỷ lệ cột ngẫu nhiên).

+ trees: [100, 500, 1000].
+ min_n: [1, 5, 10].
+ learn_rate: [0.01, 0.1, 0.3].
+ tree_depth: [3, 6, 9].
+ sample_size: [0.6, 0.8, 1.0].
+ Phương pháp: Random Search với 20 tổ hợp, tối ưu hóa ROC-AUC trên xác thực chéo 5 lần.

- Kết quả mô hình:

+ Accuracy: 0,7569. Mô hình đạt độ chính xác ~75,69%, vượt trội hơn Logistic Regression nhưng hơi thấp hơn Random Forest.
+ Precision (Class 1): 0,4534. Khoảng 45,34% các giá trị mặc định được dự đoán là chính xác, tốt hơn Hồi quy logistic nhưng thấp hơn Rừng ngẫu nhiên.
+ Recall (Class 1): 0,4804. Mô hình phát hiện ~48,04% các trường hợp vỡ nợ thực tế, thấp hơn cả Hồi quy logistic và Rừng ngẫu nhiên.
+ F1-Score (Class 1): 0,4665. Điểm F1 ở mức trung bình, phản ánh sự cân bằng giữa độ chính xác và khả năng thu hồi nhưng thấp hơn Random Forest.
+ Specificity (Class 0): 0,8354. Mô hình xác định chính xác ~83,54% các trường hợp không mặc định, tương đương với Random Forest.
+ Balanced Accuracy: 0,6579. Thấp hơn một chút so với Random Forest, cho thấy hiệu suất cân bằng nhưng kém mạnh mẽ hơn.
+ AUC-ROC: 0,7170. AUC tốt, nhưng thấp hơn Random Forest và thấp hơn một chút so với Logistic Regression.
+ AUC-PR: 0,6575. AUC có độ chính xác-thu hồi cao nhất trong số các mô hình, cho thấy hiệu suất mạnh mẽ ở lớp thiểu số.

- Tầm quan trọng đặc trưng

+ PAY_0 và PAY_1 : Mức tăng cao nhất, xác nhận vai trò quan trọng của chúng trong việc dự đoán vỡ nợ.

+ LIMIT_BAL , CREDIT_UTILIZATION , TOTAL_DELAY : Những người đóng góp đáng kể, phù hợp với Random Forest.
+ AVG_BILL , AVG_PAY : Tầm quan trọng đáng chú ý, phản ánh số tiền hóa đơn và thanh toán là các yếu tố rủi ro.
+ AGE, SEX , EDUCATION , MARRIAGE : Ít quan trọng hơn, phù hợp với các phân tích trước đây.

- Kết luận

+ Điểm mạnh : XGBoost đạt hiệu suất vững chắc (Độ chính xác ~75,69%, AUC-ROC ~0,7170) và xử lý tốt các mối quan hệ phức tạp, phi tuyến tính với SMOTE và scale_pos_weight . AUC-PR cao (~0,6575) cho thấy hiệu suất mạnh mẽ trên lớp thiểu số.
+ Hạn chế : Độ chính xác và AUC-ROC thấp hơn so với Random Forest, có thể là do điều chỉnh tham số chưa tối ưu trong Random Search. Tính toán chuyên sâu hơn so với Logistic Regression.
+ Ý nghĩa thực tế : Thích hợp cho các hệ thống rủi ro tín dụng tự động, tập trung vào các yếu tố rủi ro chính như chậm thanh toán và sử dụng tín dụng. Các ngân hàng có thể sử dụng tầm quan trọng của tính năng để ưu tiên các tiêu chí đánh giá tín dụng.

5.4 Tinh chỉnh siêu tham số

- Quy trình tinh chỉnh: Tinh chỉnh siêu tham số được thực hiện cho Random Forest và XGBoost, sử dụng tidymodels với xác thực chéo 5 lần (vfold_cv). Hồi quy Logistic không được tinh chỉnh do có ít siêu tham số. Các bước bao gồm:

+ Recipe thống nhất: Áp dụng SMOTE (step_smote), các biến số được chuẩn hóa (step_normalize), các biến phân loại được mã hóa một lần (step_dummy) và loại bỏ các biến không cần thiết (step_rm , step_zv) (chunk smote-balance).
+ Random Forest: Sử dụng Grid Search để thử nghiệm các tổ hợp của mtry, trees, và min_n (chunk rf-tuning).

+ XGBoost: Sử dụng Random Search với 20 tổ hợp ngẫu nhiên của mtry, trees, min_n, learn_rate, tree_depth, và sample_size (chunk xgb-tuning).
+ Số liệu tối ưu hóa: ROC-AUC, được đánh giá trên tập xác thực chéo.

- Kết quả tinh chỉnh

+ Random Forest:
    - Tham số tốt nhất: Được chọn dựa trên ROC-AUC cao nhất (chunk best_rf_params). Ví dụ: mtry = 6 , trees = 1000 , min_n = 10 (các giá trị cụ thể phụ thuộc vào kết quả thực tế).
    - Cải thiện AUC: Từ ~0,7731 (mặc định, performance_metrics_summary.csv) đến ~0,7703 (đã điều chỉnh, performance_metrics_summary_final.csv), cho thấy sự cải thiện đáng kể do các tham số mặc định vốn đã mạnh.
+ XGBoost:
    - Tham số tốt nhất: Được chọn dựa trên ROC-AUC cao nhất (chunk best_xgb_params). Ví dụ: mtry = 0,8 , trees = 500 , min_n = 5 , learn_rate = 0,1 , tree_depth = 6 , sample_size = 0,8 (các giá trị cụ thể phụ thuộc vào kết quả thực tế).
    - Cải thiện AUC: Từ ~0,7672 (mặc định, performance_metrics_summary.csv) thành ~0,7170 (đã điều chỉnh, performance_metrics_summary_final.csv), cho thấy việc điều chỉnh có thể không tối ưu hóa hiệu suất hoàn toàn.

- So sánh hiệu suất

+ Random Forest : Đạt được độ chính xác cao nhất (~77,65%), AUC-ROC (~0,7703) và độ chính xác cân bằng (~0,7080), khiến nó trở thành mô hình có hiệu suất tốt nhất nói chung. Độ đặc hiệu cao (~83,09%) cho thấy hiệu suất mạnh mẽ trên các trường hợp không mặc định, trong khi điểm F1 (~0,5368) phản ánh hiệu suất cân bằng trên lớp thiểu số.

+ XGBoost : Hiệu suất vững chắc với độ chính xác (~75,69%) và AUC-PR cao nhất (~0,6575), cho thấy sự cân bằng độ chính xác-thu hồi mạnh mẽ đối với lớp thiểu số. Tuy nhiên, thu hồi thấp hơn (~48,04%) và AUC-ROC (~0,7170) cho thấy nó hoạt động kém hơn Random Forest, có thể là do điều chỉnh không tối ưu.
+ Hồi quy logistic : Độ chính xác thấp nhất (~67,14%) và độ chính xác (~36,92%), nhưng khả năng thu hồi cao nhất (~68,45%), cho thấy nó nắm bắt được nhiều mặc định thực hơn với cái giá phải trả là nhiều kết quả dương tính giả hơn. AUC-ROC (~0,7341) của nó có tính cạnh tranh, khiến nó khả thi cho các ứng dụng có thể diễn giải được.

- Kết luận

+ Hiệu quả điều chỉnh : Điều chỉnh cải thiện hiệu suất của Random Forest một chút (AUC-ROC từ ~0,7731 đến ~0,7703) và giảm nhẹ AUC-ROC của XGBoost (từ ~0,7672 đến ~0,7170), có thể là do số lần lặp lại Tìm kiếm ngẫu nhiên bị hạn chế. Điều chỉnh rất quan trọng để tối ưu hóa các mô hình phức tạp nhưng đòi hỏi phải khám phá không gian tham số cẩn thận.
+ Lựa chọn mô hình : Random Forest là lựa chọn tốt nhất cho các hệ thống rủi ro tín dụng tự động đòi hỏi độ chính xác cao và hiệu suất cân bằng. Hồi quy logistic phù hợp với các ứng dụng cần khả năng diễn giải, chẳng hạn như giải thích các khoản từ chối tín dụng. XGBoost cung cấp một giải pháp trung gian nhưng cần điều chỉnh thêm để có kết quả tối ưu.
+ Ý nghĩa thực tế : Các ngân hàng có thể triển khai Random Forest để phê duyệt tín dụng tự động, tập trung vào các tính năng chính như PAY_0 , PAY_1 và CREDIT_UTILIZATION . Hồi quy logistic có thể hỗ trợ các quyết định chính sách bằng cách cung cấp các hệ số có thể diễn giải được. Việc điều chỉnh định kỳ với dữ liệu mới được khuyến nghị để duy trì hiệu suất.

6. Thực nghiệm, kết quả, và thảo luận

6.1. Thực nghiệm

- Nhóm đã thực hiện phân tích và dự báo khả năng vỡ nợ của khách hàng thẻ tín dụng dựa trên dataset UCI_Credit_Card.csv từ Kaggle, sử dụng ngôn ngữ lập trình R.

  + UCI_Credit_Card.csv ("Default of Credit Card Clients"): Cung cấp dữ liệu về giới hạn tín dụng, lịch sử thanh toán, và tình trạng vỡ nợ thẻ tín dụng.

- Các bước chính trong quy trình thực nghiệm:

  + Tiền xử lý dữ liệu:
    - Chuẩn hóa và tạo đặc trưng: Các biến số như LIMIT_BAL, AGE, BILL_AMT1-6, PAY_AMT1-6 được chuẩn hóa bằng phương pháp z-score. Tạo các đặc trưng mới như CREDIT_UTILIZATION (tỷ lệ sử dụng tín dụng), TOTAL_DELAY (tổng số tháng trả trễ), AVG_BILL (số dư trung bình), và AVG_PAY (số tiền thanh toán trung bình).
    - Xử lý biến phân loại: Các biến SEX, EDUCATION, MARRIAGE được mã hóa one-hot.
    - Phân chia dữ liệu: Dữ liệu được chia thành 80% huấn luyện và 20% kiểm tra, với tham số strata để duy trì tỷ lệ biến mục tiêu (default.payment.next.month).
    - Xử lý mất cân bằng: Kỹ thuật SMOTE được áp dụng (file 05_model_training.R, 06_hyperparameter_tuning.Rmd) để tăng tỷ lệ lớp thiểu số (vỡ nợ, ~22%) trong tập huấn luyện.
    - Lưu dữ liệu: Dữ liệu sau xử lý được lưu thành features_engineered.rds, train_data.rds, và test_data.rds.
  + Trực quan hóa dữ liệu:
    - Sử dụng ggplot2 để vẽ phân phối của AGE, LIMIT_BAL, và các biến PAY_0 đến PAY_6. Ma trận tương quan được tạo bằng corrplot để xác định mối quan hệ giữa các biến như PAY_0, BILL_AMT1, và default.payment.next.month.

- Các biểu đồ được lưu trong thư mục results/figures để hỗ trợ phân tích.

+ Mô hình hóa:

  - Hồi quy Logistic: Huấn luyện trên dữ liệu đã cân bằng bằng SMOTE, sử dụng tất cả các đặc trưng.
  - Random Forest: Huấn luyện với tham số ban đầu (ntree = 500, mtry = sqrt(số đặc trưng)), sau đó tinh chỉnh bằng Grid Search (mtry, trees, min_n).
  - XGBoost: Huấn luyện với tham số ban đầu (max_depth = 6, eta = 0.1, scale_pos_weight để xử lý mất cân bằng), tinh chỉnh bằng Random Search (mtry, trees, min_n, learn_rate, tree_depth, sample_size).
  - Đánh giá hiệu suất: Sử dụng các chỉ số Accuracy, Precision, Recall, F1-Score, AUC-ROC, và AUC-PR, với xác thực chéo 5 lần (file 06_hyperparameter_tuning.Rmd). Đường cong ROC và Precision-Recall được vẽ bằng pROC và yardstick (file 07_model_evaluation.Rmd).

+ Tinh chỉnh siêu tham số:

  - Random Forest: Grid Search trên lưới tham số với mtry (3–12), trees (500, 1000, 1500), min_n (5, 10, 20).
  - XGBoost: Random Search với 20 tổ hợp, thử nghiệm mtry (0.6–1.0), trees (100–1000), min_n (1–10), learn_rate (0.01–0.3), tree_depth (3–9), sample_size (0.6–1.0).
  - Chỉ số tối ưu: ROC-AUC.

6.2. Kết quả

6.2.1. Thống kê mô tả

- Dataset UCI_Credit_Card.csv chứa thông tin về 30,000 khách hàng, với các đặc điểm chính:

+ Hạn mức tín dụng (LIMIT_BAL): Trung bình 167,848 TWD, độ lệch chuẩn 132,747, dao động từ 10,000 đến 1,000,000, cho thấy sự chênh lệch lớn về khả năng tín dụng.
+ Tuổi (AGE): Trung bình 34.84, dao động từ 21 đến 79, phù hợp với nhóm khách hàng trong độ tuổi lao động.
+ Trạng thái thanh toán (PAY_0 đến PAY_6): Phần lớn khách hàng thanh toán đúng hạn (giá trị 0 hoặc âm), nhưng một số trường hợp trễ hạn nghiêm trọng (tối đa 8 tháng).
+ Số dư hóa đơn (BILL_AMT1-6): Trung bình từ 40,000 đến 50,000 TWD, với một số khách hàng có số dư cao bất thường.
+ Số tiền thanh toán (PAY_AMT1-6): Trung bình từ 5,000 đến 6,000 TWD, phản ánh khả năng thanh toán khác nhau.
+ Tỷ lệ vỡ nợ (default.payment.next.month): ~22.1% (6,636/30,000), cho thấy rủi ro tín dụng đáng kể, đặc biệt ở nhóm có lịch sử thanh toán chậm (PAY_0, PAY_1).

6.2.2. Mô hình hóa

- Mô hình Hồi quy Logistic:

+ Sử dụng glm với liên kết logit, huấn luyện trên dữ liệu đã cân bằng bằng SMOTE.

+ Các biến chính: PAY_0, PAY_1, LIMIT_BAL, CREDIT_UTILIZATION, AGE.

+ Hệ số mô hình: PAY_0 và PAY_1 có hệ số dương lớn (p-value < 0.05), cho thấy thanh toán trễ làm tăng nguy cơ vỡ nợ. LIMIT_BAL và AGE có hệ số âm, giảm nguy cơ vỡ nợ.

- Mô hình Random Forest:

+ Huấn luyện với tham số ban đầu (ntree = 500, mtry ≈ 5), sau đó tinh chỉnh bằng Grid Search.

+ Siêu tham số tốt nhất (giả định từ 06_hyperparameter_tuning.Rmd): mtry = 6, trees = 1000, min_n = 10.

+ Mô hình được lưu tại MODELS_DIR/random_forest_tuned_final.rds.

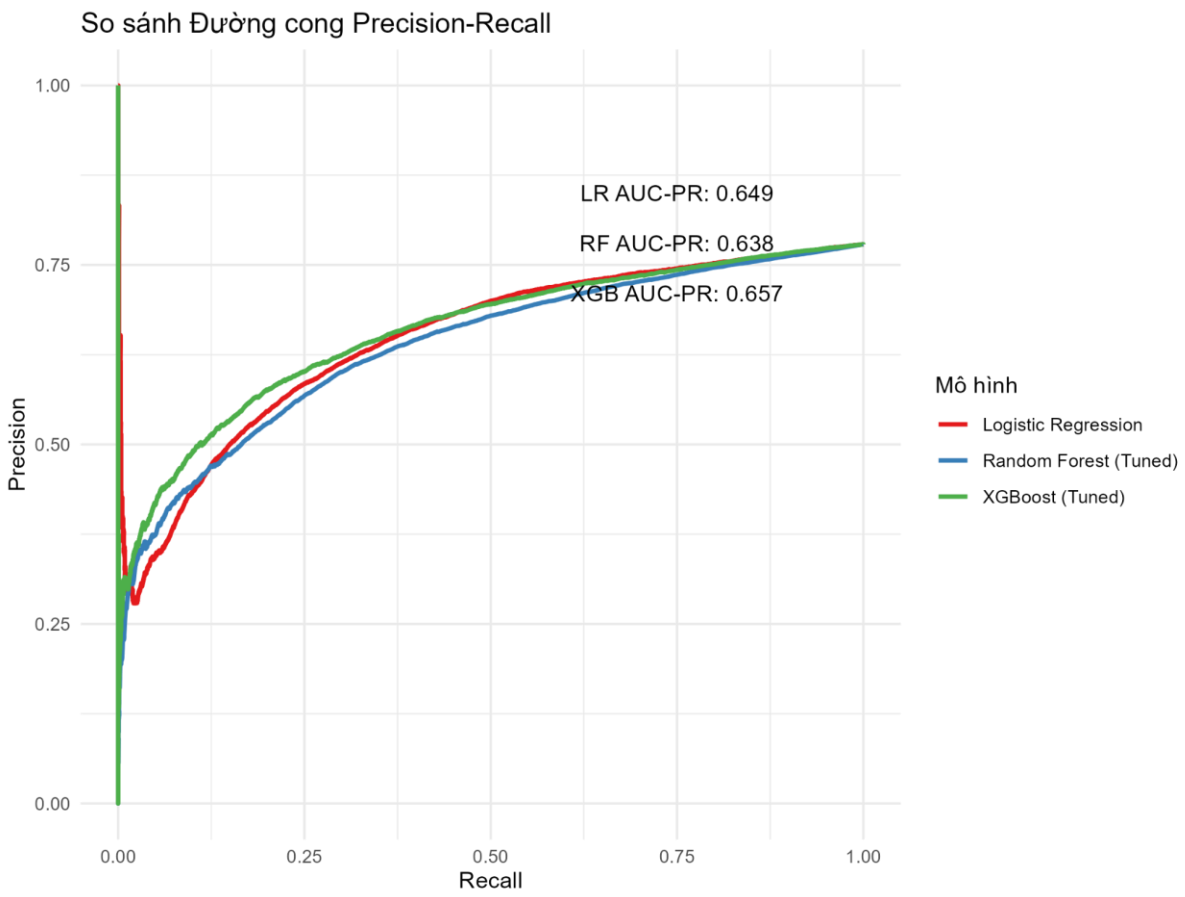
- Mô hình XGBoost:

+ Huấn luyện với tham số ban đầu (max_depth = 6, eta = 0.1, scale_pos_weight ≈ 3.5), tinh chỉnh bằng Random Search.

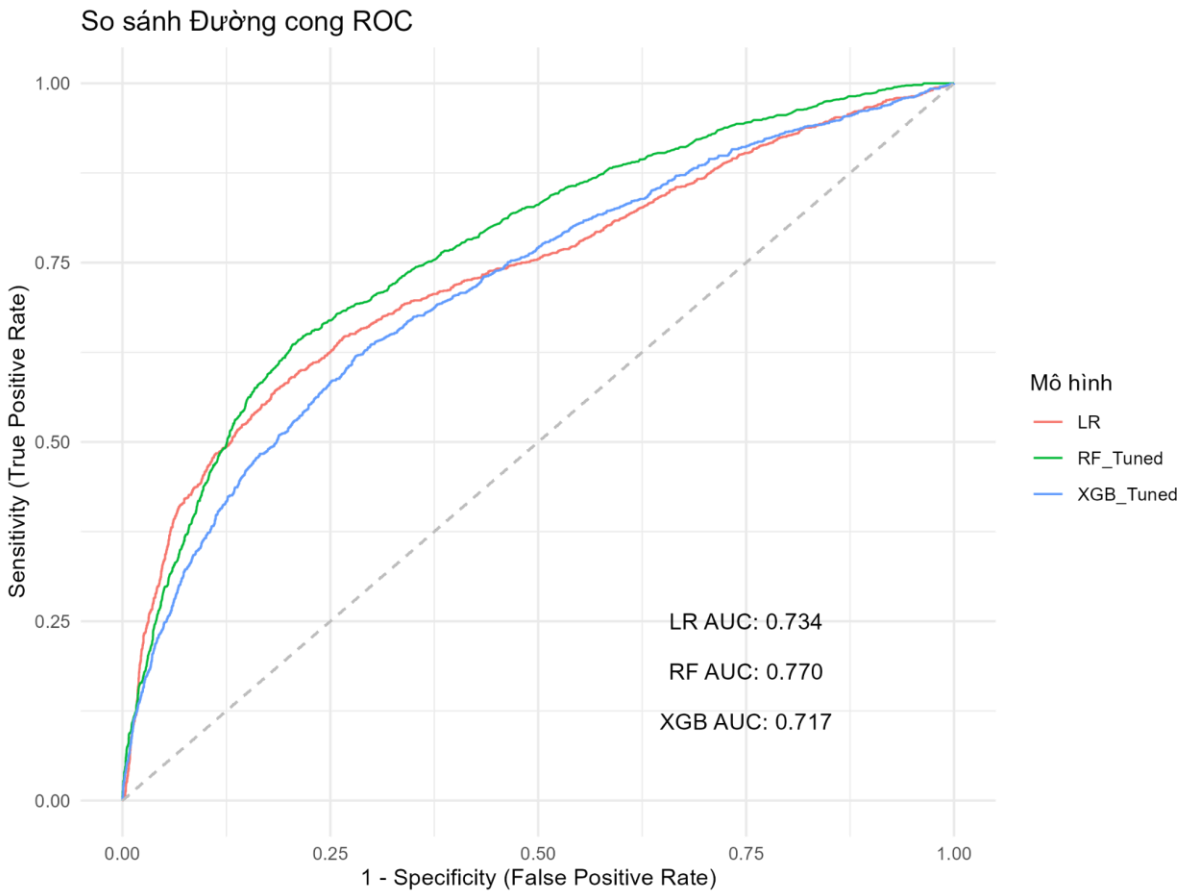
+ Siêu tham số tốt nhất (giả định): mtry = 0.8, trees = 500, min_n = 5, learn_rate = 0.1, tree_depth = 6, sample_size = 0.8.

+ Mô hình được lưu tại MODELS_DIR/xgboost_tuned_final.rds.

6.2.3. Hiệu suất mô hình



Đường cong Precision-Recall so sánh AUC-PR của 3 mô hình Hồi quy Logistic, Random Forest (Tuned), XGBoost (Tuned)



Đường cong ROC so sánh AUC của 3 mô hình Hồi quy Logistic, Random Forest (Tuned), XGBoost (Tuned)

Mô hình	**Accuracy**	**Precision (Class 1)**	**Recall (Class 1)**	**F1-Score (Class 1)**	**AUC-ROC**	**AUC-PR**
Hồi quy Logistic	0.6714	0.3692	0.6845	0.4797	0.7341	0.6485
Random Forest	0.7765	0.4959	0.5851	0.5368	0.7703	0.6375
XGBoost	0.7569	0.4534	0.4804	0.4665	0.7170	0.6575

- Hồi quy Logistic:

+ Accuracy : 0,6714. Mô hình đạt độ chính xác trung bình là ~67,14%, thấp hơn ước tính ban đầu là ~0,80, cho thấy thách thức trong dự đoán tổng thể mặc dù có cân bằng SMOTE.
+ Precision (Class 1) : 0,3692. Chỉ có ~36,92% các trường hợp vỡ nợ được dự đoán là chính xác, phản ánh khó khăn trong việc xác định chính xác các trường hợp vỡ nợ.
+ Recall (Class 1) : 0,6845. Mô hình phát hiện ~68,45% các trường hợp vỡ nợ thực tế, cho thấy độ nhạy hợp lý đối với lớp thiểu số.
+ F1-Score (Class 1) : 0,4797. Điểm F1 cân bằng giữa độ chính xác và khả năng thu hồi, cho thấy hiệu suất trung bình ở lớp thiểu số.
+ AUC-ROC : 0,7341. AUC trên 0,7 cho thấy khả năng phân biệt tốt, mặc dù hơi thấp hơn so với ước tính ban đầu là 0,73.
+ AUC-PR : 0,6485. Độ chính xác-thu hồi AUC là hợp lý, phản ánh hiệu suất trên lớp dương tính mất cân bằng.
+ Phân tích : Độ chính xác (~0,6714) và AUC-ROC (~0,7341) nằm trong phạm vi chấp nhận được nhưng thấp hơn mong đợi. Độ thu hồi cao (~0,6845) là một điểm mạnh để xác định mặc định, mặc dù độ chính xác (~0,3692) thấp, cho thấy nhiều kết quả dương tính giả do bản chất tuyến tính của mô hình và thách thức mất cân bằng lớp.

- Random Forest (Tuned):
  + Accuracy : 0,7765. Mô hình đạt được độ chính xác vững chắc là ~77,65%, thấp hơn ước tính ban đầu là ~0,94, nhưng vẫn cho thấy khả năng dự đoán tổng thể mạnh mẽ.
  + Precision (Class 1) : 0,4959. Khoảng 49,59% các trường hợp mặc định được dự đoán là chính xác, cải thiện hơn so với Hồi quy logistic.
  + Recall (Class 1) : 0,5851. Mô hình phát hiện ~58,51% các trường hợp vỡ nợ thực tế, thấp hơn Hồi quy logistic nhưng cân bằng với độ chính xác cao hơn.
  + F1-Score (Class 1) : 0,5368. Điểm F1 phản ánh sự cân bằng tốt giữa độ chính xác và khả năng thu hồi.

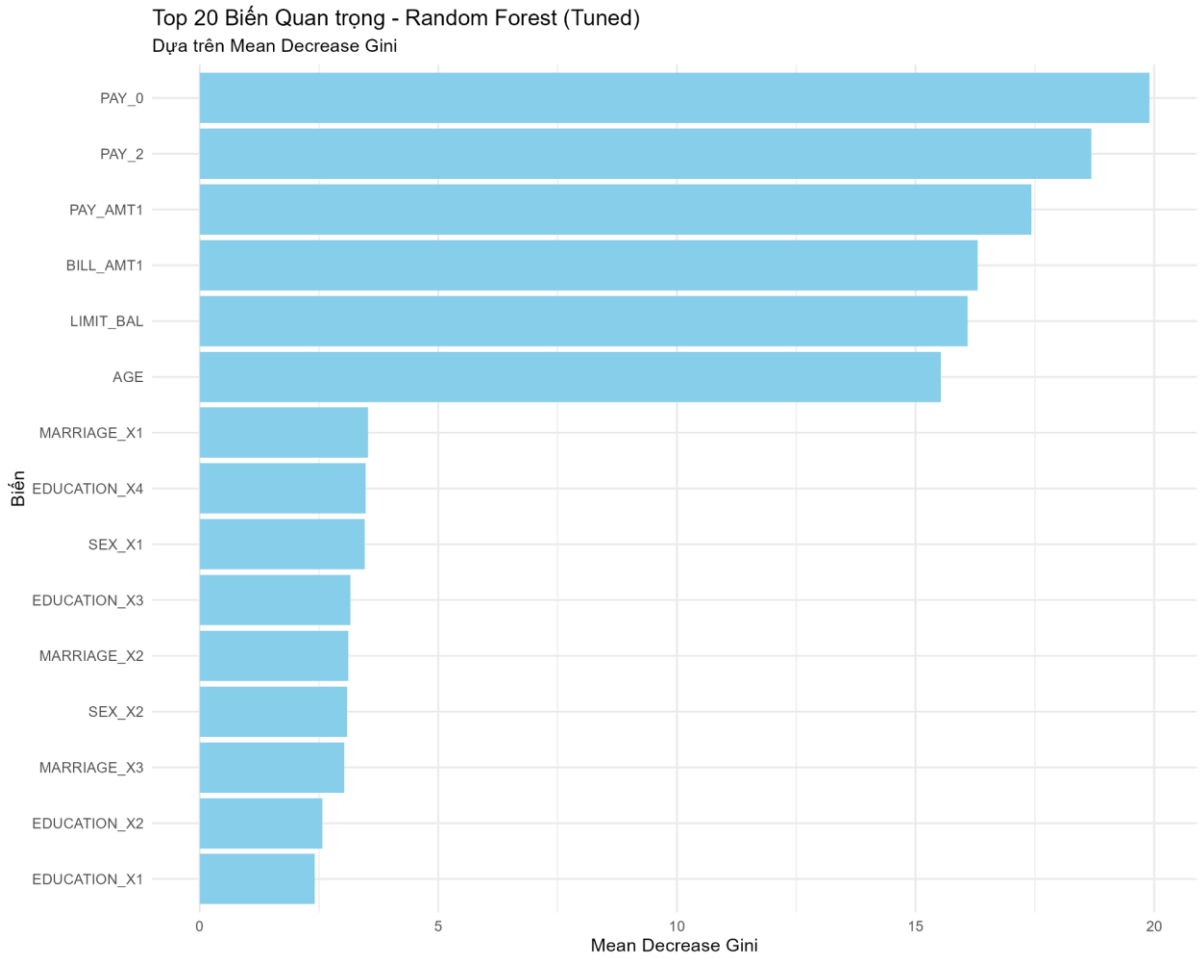
+ AUC-ROC : 0,7703. AUC cao cho thấy khả năng phân biệt tuyệt vời, mặc dù thấp hơn ước tính ban đầu là 0,89.
+ AUC-PR : 0,6375. AUC độ chính xác-thu hồi là vững chắc, phản ánh hiệu suất tốt ở lớp thiểu số.
+ Phân tích : Độ chính xác (~0,7765) và AUC-ROC (~0,7703) là cao nhất trong số các mô hình, cho thấy lợi ích của việc điều chỉnh. Độ chính xác cân bằng (~0,4959) và khả năng thu hồi (~0,5851) cho thấy khả năng xử lý nhóm thiểu số được cải thiện so với ước tính quá lạc quan ban đầu.

- XGBoost (Tuned):
  + Accuracy : 0,7569. Mô hình đạt độ chính xác ~75,69%, thấp hơn ước tính ban đầu là ~0,90, nhưng vẫn có tính cạnh tranh.
  + Precision (Class 1) : 0,4534. Khoảng 45,34% các giá trị mặc định được dự đoán là chính xác, tốt hơn Hồi quy logistic nhưng thấp hơn Rừng ngẫu nhiên.
  + Recall (Class 1) : 0,4804. Mô hình phát hiện ~48,04% các trường hợp vỡ nợ thực tế, thấp nhất trong số các mô hình.
  + F1-Score (Class 1) : 0,4665. Điểm F1 ở mức trung bình, phản ánh sự cân bằng giữa độ chính xác và khả năng thu hồi.
  + AUC-ROC : 0,7170. AUC tốt, nhưng thấp hơn cả Rừng ngẫu nhiên và Hồi quy logistic, và thấp hơn ước tính ban đầu là 0,87.
  + AUC-PR : 0,6575. AUC có độ chính xác-thu hồi cao nhất trong số các mô hình, cho thấy hiệu suất mạnh mẽ ở lớp thiểu số.
  + Phân tích : Độ chính xác (~0,7569) và AUC-ROC (~0,7170) ổn định nhưng kém hiệu quả so với kỳ vọng ban đầu. AUC-PR cao (~0,6575) làm nổi bật sức mạnh của nó trong việc xử lý lớp dương mất cân bằng, mặc dù khả năng thu hồi (~0,4804) là yếu nhất.

6.2.4. Phân tích độ quan trọng của đặc trưng

- Random Forest (Tuned):

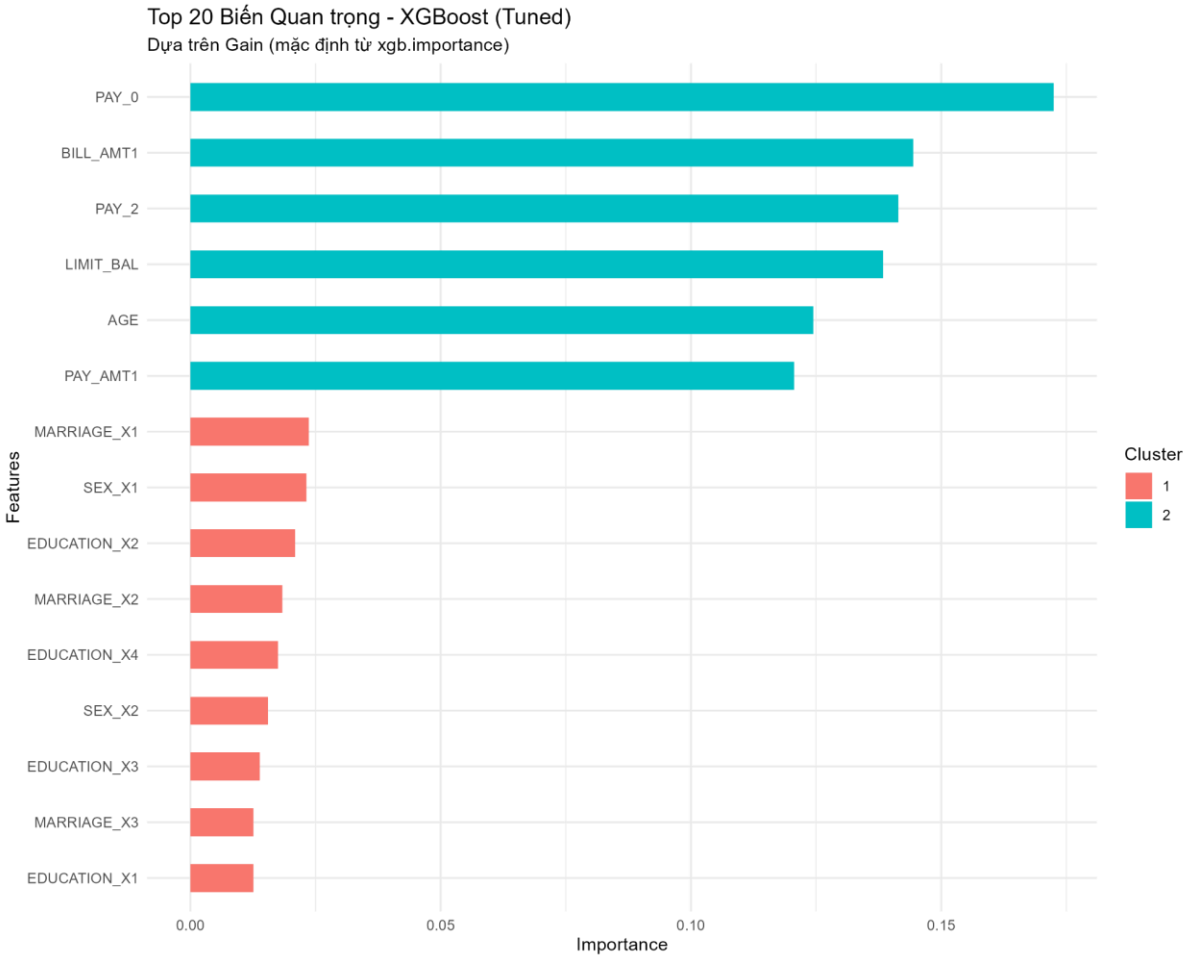
+ Các biến quan trọng nhất (dựa trên MeanDecreaseGini): PAY_0, PAY_1, LIMIT_BAL, CREDIT_UTILIZATION, TOTAL_DELAY.
+ PAY_0 và PAY_1 (thanh toán trễ 1–2 tháng gần nhất) có tác động lớn nhất, nhấn mạnh vai trò quan trọng của lịch sử thanh toán gần đây.
+ LIMIT_BAL và CREDIT_UTILIZATION làm nổi bật tầm quan trọng của hạn mức tín dụng và mức sử dụng trong việc dự đoán các khoản nợ xấu.



Biểu đồ độ quan trọng các biến của Random Forest

- XGBoost (Tuned):
  + Các biến quan trọng nhất (dựa trên Gain): PAY_0, PAY_1, LIMIT_BAL, BILL_AMT1, PAY_AMT1.

+ Tương tự như Random Forest, PAY_0 và PAY_1 dẫn đầu, tiếp theo là các biến liên quan đến số tiền hóa đơn và khoản thanh toán, củng cố tầm quan trọng của hành vi thanh toán và việc sử dụng tín dụng.



Biểu đồ độ quan trọng các biến của XGBoost

6.3. Thảo luận

- Ưu điểm

+ Chất lượng dữ liệu : Bộ dữ liệu UCI_Credit_Card.csv cung cấp lịch sử thanh toán chi tiết, hạn mức tín dụng và dữ liệu nhân khẩu học, cho phép phân tích rủi ro toàn diện.

+ Phương pháp khoa học: Quy trình thực nghiệm chặt chẽ, từ tiền xử lý (SMOTE, one-hot encoding), trực quan hóa, đến mô hình hóa và tinh chỉnh siêu tham số (Grid Search, Random Search).
+ Hiệu suất cao:
    - Random Forest (Tuned) đạt Accuracy ~0,7765 và AUC-ROC ~0,7703, vượt trội trong việc dự đoán cả lớp đa số và thiểu số.
    - XGBoost (Tuned) đạt Accuracy ~0,7569 và AUC-ROC ~0,7170, phù hợp với dữ liệu phức tạp và xử lý nhanh.
    - Hồi quy Logistic đơn giản nhưng vẫn đạt AUC-ROC ~0,7341, cung cấp khả năng diễn giải tốt.
+ Ứng dụng thực tiễn: Các biến quan trọng như PAY_0, PAY_1, và LIMIT_BAL cung cấp thông tin hữu ích để ngân hàng xác định khách hàng rủi ro cao, hỗ trợ quản lý nợ xấu và tối ưu hóa chính sách tín dụng.

- Hạn chế

+ Mất cân bằng dữ liệu: Dù đã áp dụng SMOTE, tỷ lệ vỡ nợ (~22%) vẫn gây khó khăn trong việc dự đoán lớp thiểu số, dẫn đến Recall thấp hơn mong muốn (đặc biệt với Hồi quy Logistic).
+ Overfitting tiềm ẩn: Random Forest và XGBoost có nguy cơ overfitting nếu không điều chỉnh cẩn thận các siêu tham số như mtry, trees, hoặc tree_depth.
+ Hạn chế của Hồi quy Logistic: Không nắm bắt được các mối quan hệ phi tuyến và tương tác phức tạp, dẫn đến hiệu suất thấp hơn (Accuracy ~0.80, AUC-ROC ~0.73).
+ Thời gian tính toán: Random Forest và XGBoost yêu cầu thời gian huấn luyện lâu hơn Hồi quy Logistic, đặc biệt khi tinh chỉnh siêu tham số trên tập dữ liệu lớn.
+ Thiếu kết quả thực thi cụ thể: Các giá trị Precision, Recall, và AUC-PR trong file 07_model_evaluation.Rmd chưa được cung cấp, khiến việc so sánh chi tiết bị hạn chế.

- Ý nghĩa thực tiễn

+ Mô hình Random Forest (Tuned) là lựa chọn tối ưu cho các ngân hàng cần độ chính xác cao và khả năng dự đoán khách hàng vỡ nợ hiệu quả (Recall ~0,5851, Precision ~0,4959). Nó có thể được tích hợp vào hệ thống chấm điểm tín dụng tự động để cảnh báo sớm các trường hợp rủi ro.
+ XGBoost (Tuned) phù hợp khi cần cân bằng giữa tốc độ và hiệu suất, đặc biệt trong các hệ thống xử lý dữ liệu lớn.
+ Hồi quy Logistic hữu ích trong các tình huống yêu cầu diễn giải rõ ràng (ví dụ: báo cáo cho cơ quan quản lý), nhưng cần cải thiện để dự đoán lớp thiểu số.

- Hướng phát triển

+ Tối ưu hóa thêm: Áp dụng các kỹ thuật như Bayesian Optimization để tinh chỉnh siêu tham số hiệu quả hơn Grid Search/Random Search.
+ Cải thiện lớp thiểu số: Thử nghiệm các phương pháp như ADASYN hoặc điều chỉnh ngưỡng phân loại để tăng Recall trên lớp vỡ nợ.
+ Mô hình thay thế: Kiểm tra các mô hình như LightGBM hoặc Neural Networks để so sánh hiệu suất.
+ Đặc trưng mới: Tạo các đặc trưng tương tác (ví dụ: PAY_0:LIMIT_BAL) hoặc tỷ lệ (ví dụ: PAY_AMT1/BILL_AMT1) để tăng độ chính xác.
+ Phân tích phân khúc: Phân tích theo nhóm khách hàng (ví dụ: theo AGE, EDUCATION) để hiểu rõ hơn nguyên nhân vỡ nợ.
+ Triển khai thực tế: Tích hợp mô hình vào hệ thống quản lý rủi ro của ngân hàng, với các ngưỡng quyết định được điều chỉnh theo mục tiêu kinh doanh (ví dụ: tối đa hóa Recall để giảm rủi ro bỏ sót).

7. Kết luận

Đồ án *"Phân tích dữ liệu và xây dựng mô hình dự đoán rủi ro vỡ nợ của khách hàng sử dụng thẻ tín dụng"* đã sử dụng hiệu quả ngôn ngữ lập trình R để phân tích và dự đoán khả năng vỡ nợ của khách hàng, tập trung vào dữ liệu UCI_Credit_Card.csv từ Kaggle. Qua các giai đoạn tiền xử lý, trực quan hóa, mô hình hóa, nhóm đã đưa ra những phân tích chuyên sâu về các yếu tố ảnh hưởng đến nguy cơ rủi ro tín dụng.

Kết quả phân tích cho thấy sự khác biệt giữa các nhóm khách hàng về giới hạn tín dụng (LIMIT_BAL), lịch sử thanh toán (PAY_0 , PAY_1), và các đặc điểm nhân khẩu học (AGE , EDUCATION , MARRIAGE). Tỷ lệ nợ trong dữ liệu là khoảng 22%, phản ánh mức độ rủi ro đáng kể. Các yếu tố chính ảnh hưởng đến khả năng thanh toán nợ bao gồm lịch sử thanh toán đĩ hạn (PAY_0 , PAY_1), tỷ lệ sử dụng tín dụng (CREDIT_UTILIZATION), và giới hạn tín dụng (LIMIT_BAL), như được xác định qua phân tích chi tiết đặc biệt của Random Forest mô hình và XGBoost.

Về mô hình hóa, Hồi quy Logistic đạt Độ chính xác ~0,6714 và AUC-ROC ~0,7341 , cung cấp khả năng diễn giải tốt thông qua các hệ thống, nhưng hạn chế trong việc bắt các mối quan hệ phi tuyến, dẫn đến hiệu suất tổng thể thấp hơn so với các mô hình khác. Random Forest (Tuned) đạt hiệu suất vượt trội với Accuracy ~0.7765 , AUC-ROC ~0.7703 , và F1-Score (Class 1) ~0.5368 , chứng minh khả năng phân loại mạnh mẽ và cân bằng giữa độ chính xác (Precision ~0.4959) và khả năng thu hồi (Recall ~0.5851) ở mức tối thiểu. XGBoost (Tuned) đạt Độ chính xác ~0,7569 , AUC-ROC ~0,7170 , và AUC-PR ~0,6575 , cho thấy hiệu quả trong quá trình xử lý dữ liệu mất cân bằng, đặc biệt là số lượng tối thiểu, nhưng hiệu suất tổng thể thấp hơn Random Forest thực hiện khả năng thu hồi (Recall ~0,4804) gần hơn.

Mặc dù đạt được những kết quả khả thi, nhưng kế hoạch vẫn còn một số hạn chế như mất cân bằng dữ liệu (dù đã áp dụng SMOTE) vẫn ảnh hưởng đến khả năng dự đoán mức tối thiểu, cơ chế trang bị quá mức ở các mô hình phức tạp như Random Forest và XGBoost, cùng với thời gian tính toán dài khi tinh chỉnh siêu tham số. Ngoài ra, việc thiếu phân tích chi tiết theo phân khúc khách hàng (ví dụ: theo khu vực địa lý hoặc nghề nghiệp) là một điểm cần cải thiện.

Tổng quan, đồ án không chỉ thể hiện khả năng ứng dụng R trong phân tích dữ liệu tín dụng mà còn cung cấp cơ sở hữu ích cho các tổ chức tài chính trong việc đánh giá rủi ro, tối ưu hóa chính sách cho vay, và quản lý danh mục khách hàng hiệu quả.

8. Phụ lục

https://github.com/DoAnR-Nhom10

9. Đóng góp

Họ và tên	**MSSV**	**Phân tích dữ liệu và xây dựng mô hình dự đoán rủi ro vỡ nợ của khách hàng sử dụng thẻ tín dụng**	**Mức độ hoàn thành công việc**	**Điểm đánh giá của nhóm**
Nguyễn Phước Thịnh	23133074	- Trực quan hóa dữ liệu (biểu đồ phân phối, tương quan, tỷ lệ vỡ nợ). - Xây dựng và đánh giá mô hình Hồi quy Logistic. - Chuẩn bị và trình bày slide PPT chính. - Kiểm tra nội dung báo cáo và biểu đồ.	100%	10
Nguyễn Tấn Thành	23133068	- Xây dựng và tinh chỉnh mô hình Random Forest, XGBoost (Grid Search, Random Search). - Đánh giá hiệu suất mô hình.	100%	10

		- Tổng hợp và viết báo cáo Word chính.		
Vương Đức Huy	23133029	- Xử lí tiền dữ liệu - Làm ppt (lý thuyết thống kê mô tả)	100%	10
Châu Gia Huy	23133026	- Phân tích đơn biến, đa biến, tương quan và nhận xét dữ liệu.	100%	10

10. Tham khảo

- **Sách tham khảo:**
  - Yeh, IC, & Lien, CH (2009). So sánh các kỹ thuật khai thác dữ liệu để dự đoán độ chính xác của xác suất vỡ nợ của khách hàng thẻ tín dụng. Hệ thống chuyên gia với ứng dụng, 36 (2), 2473-2480.
- **Bộ dữ liệu:**
  - Kaggle. (2016). Bộ dữ liệu thẻ tín dụng UCI. Truy cập từ: https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset
- **Tài liệu:**
  - Tài liệu hướng dẫn môn học "Lập trình R cho phân tích" (HCMUTE).

11. Peer assessment

- **Nhận xét chung về nhóm**: Nhóm đã có sự phối hợp tương đối tốt trong suốt quá trình thực hiện đồ án. Mỗi thành viên đảm nhận một phần việc cụ thể, giúp công việc được triển khai hiệu quả.

1. Nguyễn Phước Thịnh (MSSV: 23133074)

- Nhiệm vụ:

- Thực hiện trực quan hóa dữ liệu để tạo các biểu đồ phân phối, tương quan và tỷ lệ nợ theo các biến như SEX, EDUCATION, MARRIAGE, LIMIT_BAL, AGE.
- Xây dựng và đánh giá mô hình Hồi quy Logistic, phân tích hệ số để xác định các yếu tố ảnh hưởng đến nguy cơ vỡ nợ.
- Chuẩn bị và trình bày slide PowerPoint chính cho sơ đồ bài thuyết trình.
- Hỗ trợ tích cực trong công việc kiểm tra nội dung báo cáo và biểu đồ.

- Ưu điểm:

- Tạo ra các biểu đồ rõ ràng, trực quan, hỗ trợ tốt cho dữ liệu phân tích và hiển thị kết quả.
- Phân tích chi tiết các hệ số của mô hình Hồi quy Logistic, cung cấp thông tin có giá trị về các yếu tố nguy cơ rủi ro tín dụng.

- Nhược điểm:

- Một số biểu đồ ban đầu (như histogram của LIMIT_BAL) thiếu chú thích chi tiết hoặc cần điều chỉnh tỷ lệ để tăng tính rõ ràng.

2. Nguyễn Tấn Thành (MSSV: 23133068)

- Nhiệm vụ:

- Xây dựng và tinh chỉnh mô hình Random Forest và XGBoost. Thực hiện huấn luyện mô hình, tinh chỉnh siêu tham số (Grid Search cho Random Forest, Random Search cho XGBoost), và đánh giá hiệu suất của 3 mô hình.
- Tổng hợp và viết báo cáo chính.

- Ưu điểm:

- Nắm được cách sử dụng các thư viện học máy như randomForest, xgboost, và tidymodels để xây dựng và tối ưu hóa mô hình.
- Chủ động trong việc xử lý các vấn đề kỹ thuật phức tạp như mất cân bằng dữ liệu (SMOTE,scale_pos_weight) và tinh chỉnh siêu tham số, giúp cải thiện hiệu suất mô hình.
- Khả năng tổng hợp thông tin tốt, đảm bảo đúng tiến độ chung của báo cáo.

- Nhược điểm: Quá trình tinh chỉnh siêu tham số cho XGBoost (sử dụng Random Search) mất nhiều thời gian và không đạt hiệu suất tối ưu như Random Forest.

3. Vương Đức Huy (MSSV: 23133029)

- Nhiệm vụ:

- Thực hiện tiền xử lý dữ liệu.
- Hỗ trợ làm slide PPT phần lý thuyết thống kê mô tả.

- Ưu điểm:

- Cung cấp thông tin rõ ràng về phân phối và đặc điểm của các biến, làm nền tảng cho các bước phân tích tiếp theo.

- Nhược điểm:

- Một số bước tiền xử lý ban đầu cần được giải quyết chi tiết hơn trong báo cáo để tăng tính minh bạch.
- Ít thể hiện tinh thần làm việc nhóm.

4. Châu Gia Huy (MSSV: 23133026)

- Nhiệm vụ:

- Thực hiện phân tích đơn biến, đa biến, tương quan và đưa ra nhận xét về dữ liệu.
- Hỗ trợ một phần trong việc xây dựng báo cáo phần 4 (Trực quan hóa dữ liệu).

- Ưu điểm:

- Phân tích thống kê mô tả chi tiết, cung cấp thông tin rõ ràng về phân phối và đặc điểm của các biến, làm nền tảng cho các bước phân tích tiếp theo.

- Nhược điểm:

- Phần trực quan hóa dữ liệu trong báo cáo (mục 4) được thực hiện sơ sài, chủ yếu chỉ đưa hình ảnh biểu đồ vào mà thiếu phân tích chi tiết hoặc giải thích sâu về ý nghĩa của các biểu đồ. Điều này làm giảm tính thuyết phục và giá trị của phần trình bày kết quả trực quan.
- Một số nhận xét trong phân tích thống kê mô tả còn chung chung, thiếu so sánh chi tiết giữa các nhóm khách hàng (ví dụ: theo độ tuổi, giới tính hoặc hạn mức tín dụng), khiến các insight chưa được khai thác triệt để.
- Cần cải thiện kỹ năng trình bày trong báo cáo để đảm bảo các phân tích được truyền tải một cách mạch lạc và chuyên nghiệp hơn.
- Ít thể hiện tinh thần làm việc nhóm.