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

Mã môn học: RPAN233577_01
GVHD: TS.Phan Thị Thể

Thành viên nhóm 10:
23133074 - Nguyễn Phước Thịnh
23133068 - Nguyễn Tấn Thành
23133029 - Vương Đức Huy
23133026 - Châu Gia Huy

Mục Lục

Vì sao phân tích vỡ nợ quan trọng?

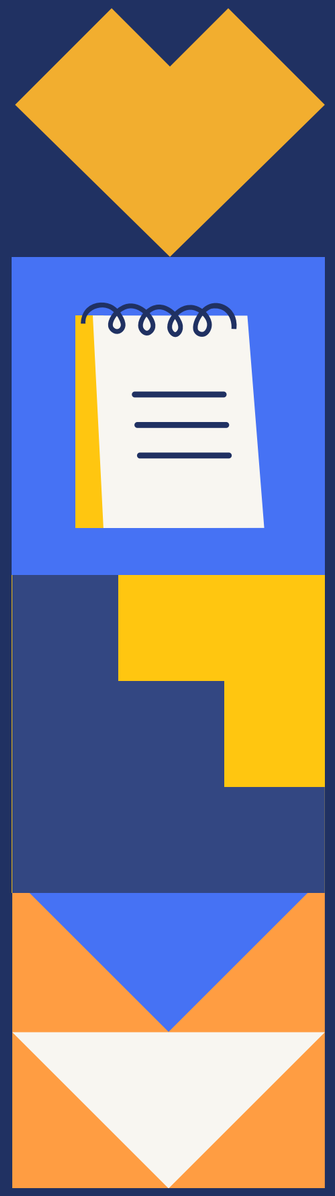
Vấn đề then chốt trong quản lý rủi ro tín dụng của các tổ chức tài chính

- Mất vốn, ảnh hưởng dòng tiền
- Tăng chi phí xử lý nợ xấu
- Ảnh hưởng đến uy tín và an toàn tài chính toàn hệ thống

PHẦN 1

Tổng quan dữ liệu

Mục tiêu chính



Mục tiêu	Phạm vi
Phân tích dữ liệu UCI Credit Card	Kaggle, UCI
Xây dựng mô hình dự đoán khả năng vỡ nợ	Mô hình: Random Forest Logistic Regression, Decision Tree, v.v.

Default of Credit card

Nguồn : kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset/data

```
CI_Credit_Card.csv ===
ibble: 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	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0
2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0
3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518
4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000
5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000
6	50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024	2500
7	500000	1	1	2	29	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944	55000
8	100000	2	2	2	23	0	-1	-1	0	0	-1	11876	380	601	221	-159	567	380
9	140000	2	3	1	28	0	0	2	0	0	0	11285	14096	12108	12211	11793	3719	3329
10	20000	1	3	2	35	-2	-2	-2	-2	-1	-1	0	0	0	0	13007	13912	0

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

Biến chính default.payment.next.month

- Là biến mục tiêu cần dự đoán
  (0 = Không vỡ nợ, 1 = Vỡ nợ).
- Chức năng: xác định xem người dùng sẽ thanh toán đúng hạn vào tháng kế tiếp hay không.

- 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).

Xử lí dữ liệu

Missing Value

Kiểm tra giá trị NA

```
Kiểm tra giá trị thiếu:
> colSums(is.na(data))
                        ID                  LIMIT_BAL                        SEX                  EDUCATION
                         0                          0                          0                          0
                  MARRIAGE                        AGE                      PAY_0                      PAY_2
                         0                          0                          0                          0
                     PAY_3                      PAY_4                      PAY_5                      PAY_6
                         0                          0                          0                          0
                  BILL_AMT1                  BILL_AMT2                  BILL_AMT3                  BILL_AMT4
                         0                          0                          0                          0
                  BILL_AMT5                  BILL_AMT6                   PAY_AMT1                   PAY_AMT2
                         0                          0                          0                          0
                   PAY_AMT3                   PAY_AMT4                   PAY_AMT5                   PAY_AMT6
                         0                          0                          0                          0
default.payment.next.month
                         0
```

Giúp giảm thiểu nhu cầu áp dụng các kỹ thuật điền giá trị

Missing Value

Xử lí các giá trị không xác định

```r
```{r }
# EDUCATION: Gộp 0, 5, 6 vào 4 (others)
cat("Tần suất EDUCATION trước xử lý:\n")
table(data$EDUCATION)
data$EDUCATION[data$EDUCATION %in% c(0, 5, 6)] <- 4
cat("Tần suất EDUCATION sau xử lý:\n")
table(data$EDUCATION)
```
```

```
Tần suất EDUCATION trước xử lý:

    0     1     2     3     4     5     6
   14 10585 14030  4917   123   280    51
Tần suất EDUCATION sau xử lý:

    1     2     3     4
10585 14030  4917   468
```

```r
```{r }
# MARRIAGE: Gộp 0 vào 3 (others)
cat("Tần suất MARRIAGE trước xử lý:\n")
table(data$MARRIAGE)
data$MARRIAGE[data$MARRIAGE == 0] <- 3
cat("Tần suất MARRIAGE sau xử lý:\n")
table(data$MARRIAGE)
```
```

```
Tần suất MARRIAGE trước xử lý:

    0     1     2     3
   54 13659 15964   323
Tần suất MARRIAGE sau xử lý:

    1     2     3
13659 15964   377
```

Chuẩn hóa

Chuẩn hóa kiểu dữ liệu

Là quá trình chuyển đổi các cột dữ liệu về đúng kiểu phù hợp với ý nghĩa và mục đích sử dụng của chúng, như số nguyên, số thực, hoặc biến phân loại.

Chuyển về factor để mô hình nhận diện đúng là biến rời rạc, không phải số lượng liên tục.
→ Giúp mô hình phân biệt nhóm, tránh hiểu nhầm về thứ tự.

```
# Chuyển lại các biến phân loại thành factor
data$EDUCATION <- as.factor(data$EDUCATION)
data$MARRIAGE <- as.factor(data$MARRIAGE)
data$SEX <- as.factor(data$SEX)
data$default.payment.next.month <- as.factor(data$default.payment.next.month)
```

SEX, EDUCATION, MARRIAGE → factor (biến phân loại).

AGE, PAY_0 → integer.

LIMIT_BAL, BILL_AMT, PAY_AMT → numeric (số thực).

Chuẩn hóa

Chuẩn hóa các cột số

Là kỹ thuật đưa các đặc trưng số về cùng một thang đo (thường là trung bình 0, độ lệch chuẩn 1) để mô hình hoạt động hiệu quả hơn.

Các mô hình như Logistic Regression nhạy cảm với thang đo. Nếu không chuẩn hóa, các biến có giá trị lớn (ví dụ LIMIT_BAL) sẽ chi phối mô hình.

```r
# Chuẩn hóa các biến số
num_vars <- c("LIMIT_BAL", "AGE", paste0("BILL_AMT", 1:6), paste0("PAY_AMT", 1:6),
              "CREDIT_UTILIZATION", "TOTAL_DELAY", "AVG_BILL", "AVG_PAY")
data[num_vars] <- scale(data[num_vars])
```

$$z = \frac{x - \mu}{\sigma}$$

Công thức z-score

Lợi ích:

- Cải thiện hiệu quả học máy
- Tránh thiên lệch do thang đo lớn nhỏ khác nhau

Trước khi chuẩn hóa

ID	LIMIT_BAL	SEX	EDUCATIC	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	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default.pay
1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0	689	0	0	0	0	1
2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000	1
3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000	0
4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000	0
5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689	679	0
6	50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024	2500	1815	657	1000	1000	800	0
7	########	1	1	2	29	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944	55000	40000	38000	20239	13750	13770	0
8	########	2	2	2	23	0	-1	-1	0	0	-1	11876	380	601	221	-159	567	380	601	0	581	1687	1542	0
9	140000	2	3	1	28	0	0	2	0	0	0	11285	14096	12108	12211	11793	3719	3329	0	432	1000	1000	1000	0
10	20000	1	3	2	35	-2	-2	-2	-2	-1	-1	0	0	0	0	13007	13912	0	0	0	13007	1122	0	0
11	########	2	3	2	34	0	0	2	0	0	-1	11073	9787	5535	2513	1828	3731	2306	12	50	300	3738	66	0
12	260000	2	1	2	51	-1	-1	-1	-1	-1	2	12261	21670	9966	8517	22287	13668	21818	9966	8583	22301	0	3640	0
13	630000	2	2	2	41	-1	0	-1	-1	-1	-1	12137	6500	6500	6500	6500	2870	1000	6500	6500	6500	2870	0	0
14	70000	1	2	2	30	1	2	2	0	0	2	65802	67369	65701	66782	36137	36894	3200	0	3000	3000	1500	0	1
15	250000	1	1	2	29	0	0	0	0	0	0	70887	67060	63561	59696	56875	55512	3000	3000	3000	3000	3000	3000	0

Sau khi chuẩn hóa

ID	LIMIT_BAL	SEX	EDUCATIC	MARRIAGI	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT	BILL_AMT	BILL_AMT	BILL_AMT	BILL_AMT	BILL_AMT	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default.pa
1	-1.1367	2	2	1	-1.246	2	2	-1	-1	-2	-2	-0.64249	-0.64739	-0.66798	-0.67249	-0.66305	-0.65271	-0.34194	-0.22708	-0.2968	-0.30806	-0.31413	-0.29338	1
2	-0.36597	2	2	2	-1.02903	-1	2	0	0	0	2	-0.65921	-0.66674	-0.63924	-0.62163	-0.60622	-0.59796	-0.34194	-0.21358	-0.24	-0.24423	-0.31413	-0.18088	1
3	-0.59719	2	2	2	-0.16115	0	0	0	0	0	0	-0.29855	-0.49389	-0.4824	-0.44972	-0.41718	-0.39162	-0.25029	-0.19188	-0.24	-0.24423	-0.24868	-0.01212	0
4	-0.90548	2	2	1	0.1643	0	0	0	0	0	0	-0.05749	-0.01329	0.03285	-0.23237	-0.18673	-0.15658	-0.22119	-0.16936	-0.22864	-0.23784	-0.24416	-0.23713	0
5	-0.90548	1	2	1	2.33399	-1	0	-1	0	0	0	-0.57861	-0.61131	-0.16119	-0.34699	-0.34813	-0.33148	-0.22119	1.33501	0.27116	0.26643	-0.26903	-0.25518	0
6	-0.90548	1	1	2	0.1643	0	0	0	0	0	0	0.17894	0.11085	0.15277	-0.37102	-0.34035	-0.31648	-0.191	-0.17821	-0.25948	-0.24423	-0.24868	-0.24838	0
7	2.56279	1	1	2	-0.70358	0	0	0	0	0	0	4.30146	5.098	5.73897	7.7626	7.28145	7.3055	2.97866	1.47906	1.86144	0.98384	0.58584	0.4812	0
8	-0.52012	2	2	2	-1.35448	0	-1	-1	0	0	-1	-0.53435	-0.68563	-0.66925	-0.66905	-0.66566	-0.64319	-0.31899	-0.2309	-0.2968	-0.27097	-0.20371	-0.20664	0
9	-0.21183	2	3	1	-0.81206	0	0	2	0	0	0	-0.54238	-0.49292	-0.50332	-0.48268	-0.46907	-0.59027	-0.14095	-0.25699	-0.27226	-0.24423	-0.24868	-0.23713	0
10	-1.1367	1	3	2	-0.05267	-2	-2	-2	-2	-1	-1	-0.69563	-0.69097	-0.67792	-0.67249	-0.44911	-0.41911	-0.34194	-0.25699	-0.2968	0.5222	-0.24069	-0.29338	0
11	0.25061	2	3	2	-0.16115	0	0	2	0	0	-1	-0.54526	-0.55346	-0.5981	-0.63342	-0.63298	-0.59006	-0.20271	-0.25646	-0.29396	-0.28891	-0.06947	-0.28966	0
12	0.71304	2	1	2	1.68308	-1	-1	-1	-1	-1	2	-0.52912	-0.38651	-0.53421	-0.5401	-0.29647	-0.42321	0.97532	0.17555	0.19068	1.11546	-0.31413	-0.08862	0
13	3.56473	2	2	2	0.59824	-1	0	-1	-1	-1	-1	-0.53081	-0.59965	-0.58419	-0.57145	-0.55613	-0.60452	-0.28156	0.02512	0.07238	0.10685	-0.12628	-0.29338	0
14	-0.75134	1	2	2	-0.59509	1	2	2	0	0	2	0.19798	0.25557	0.26947	0.36558	-0.06866	-0.03321	-0.14874	-0.25699	-0.12641	-0.11656	-0.21595	-0.29338	1
15	0.63597	1	1	2	-0.70358	0	0	0	0	0	0	0.26704	0.25123	0.23862	0.25544	0.27244	0.27941	-0.16081	-0.12678	-0.12641	-0.11656	-0.11777	-0.12462	0

PHẦN 2

Thống kê và trực quan dữ liệu

1.Thống kê mô tả

MIN (GIÁ TRỊ NHỎ NHẤT) VÀ MAX (GIÁ TRỊ LỚN NHẤT):

- MIN LÀ GIÁ TRỊ NHỎ NHẤT TRONG TẬP DỮ LIỆU.
- MAX LÀ GIÁ TRỊ LỚN NHẤT TRONG TẬP DỮ LIỆU.
- CẢ HAI CHỈ SỐ NÀY GIÚP XÁC ĐỊNH PHẠM VI (RANGE) CỦA DỮ LIỆU
- KHOẢNG CÁCH GIỮA GIÁ TRỊ NHỎ NHẤT VÀ LỚN NHẤT.

MEAN (TRUNG BÌNH CỘNG):

- LÀ GIÁ TRỊ TRUNG BÌNH CỦA TẤT CẢ CÁC ĐIỂM DỮ LIỆU TRONG TẬP.
- CHỈ SỐ PHỔ BIẾN NHẤT ĐỂ MÔ TẢ TRUNG TÂM CỦA DỮ LIỆU.
- CÔNG THỨC TÍNH:

$$\text{Mean} = \frac{\sum X_i}{n}$$

1.Thống kê mô tả

MEDIAN (TRUNG VỊ):

- LÀ GIÁ TRỊ GIỮA TRONG MỘT TẬP DỮ LIỆU ĐÃ ĐƯỢC SẮP XẾP THEO THỨ TỰ
- TỪ NHỎ ĐẾN LỚN. NẾU CÓ SỐ LƯỢNG DỮ LIỆU CHẴN, TRUNG VỊ LÀ TRUNG
- BÌNH CỦA HAI GIÁ TRỊ GIỮA.
- TRUNG VỊ GIÚP LOẠI BỎ ẢNH HƯỞNG CỦA CÁC GIÁ TRỊ NGOẠI LAI.

SD (ĐỘ LỆCH CHUẨN):

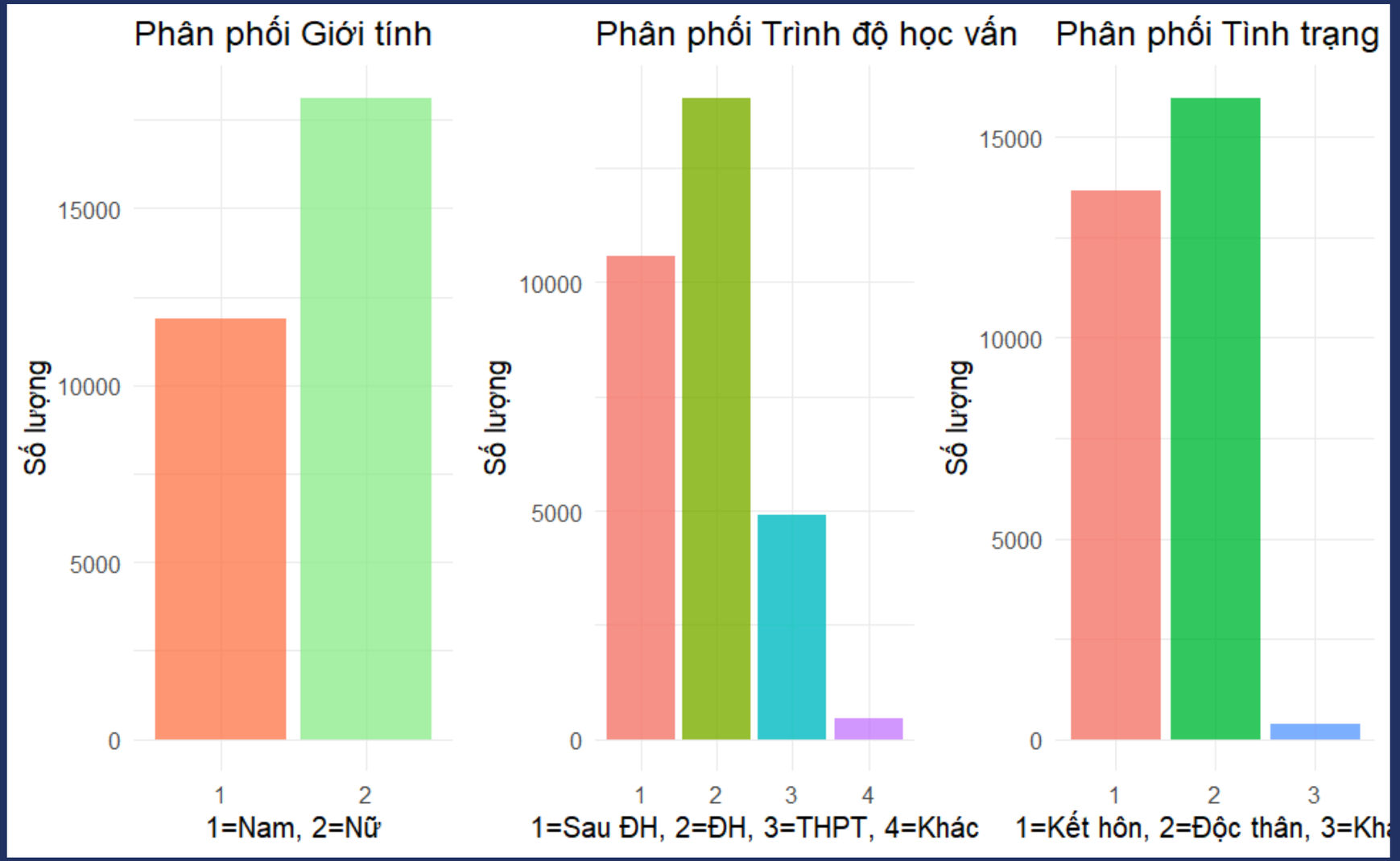
- ĐO ĐỘ PHÂN TÁN CỦA DỮ LIỆU, TỨC LÀ MỨC ĐỘ MÀ CÁC GIÁ TRỊ DỮ LIỆU
- PHÂN TÁN QUANH GIÁ TRỊ TRUNG BÌNH.
- CÔNG THỨC TÍNH:

$$\mathrm{SD} = \sqrt{\frac{\sum (X_i - \mu)^2}{n}}$$

	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

2. Trực quan hóa dữ liệu

Phân phối của các biến phân loại SEX, EDUCATION, và MARRIAGE.



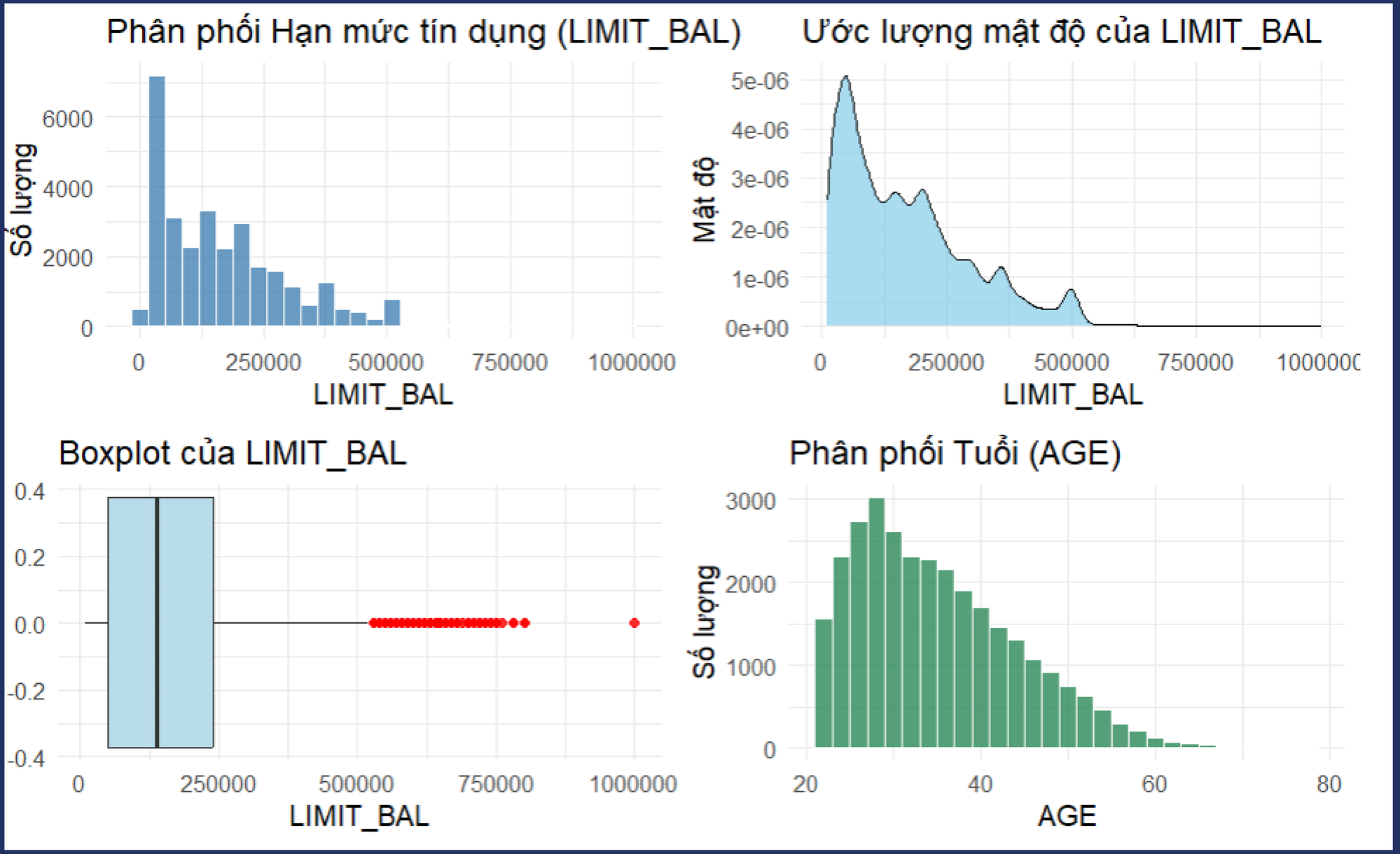
Nhận Xét:

- Phân phối Giới tính: Tỷ lệ nữ cao hơn nam, cho thấy tập khách hàng có sự chênh lệch giới tính, với nữ chiếm ưu thế. Điều này có thể phản ánh thẻ tín dụng được sử dụng nhiều hơn bởi phụ nữ trong tập dữ liệu.

- Phân phối EDUCATION: Nhóm đại học có số lượng cao nhất, tiếp theo là cao học , trung học , và khác . Cho thấy sự đa dạng trong trình độ học vấn.

- Phân phối MARRIAGE: Độc thân và đã kết hôn chiếm đa số, nhóm khác ít hơn. Điều này phản ánh khách hàng chủ yếu ở trạng thái hôn nhân rõ ràng (kết hôn hoặc độc thân).

Phân phối của LIMIT_BAL và AGE



Nhận Xét:

Histogram (LIMIT_BAL và AGE):

LIMIT_BAL có phân phối lệch phải giá trị tập trung ở vùng thấp cho thấy ngân hàng có xu hướng cấp hạn mức tín dụng thấp cho đa số khách hàng(khách hàng thuộc nhóm rủi ro cao)

AGE có phân phối gần đối xứng, với đỉnh ở khoảng 20-40. Tập khách hàng chủ yếu là người trẻ, có thể là đối tượng mới tham gia thị trường tín dụng, dẫn đến nhu cầu sử dụng thẻ tín dụng cao.

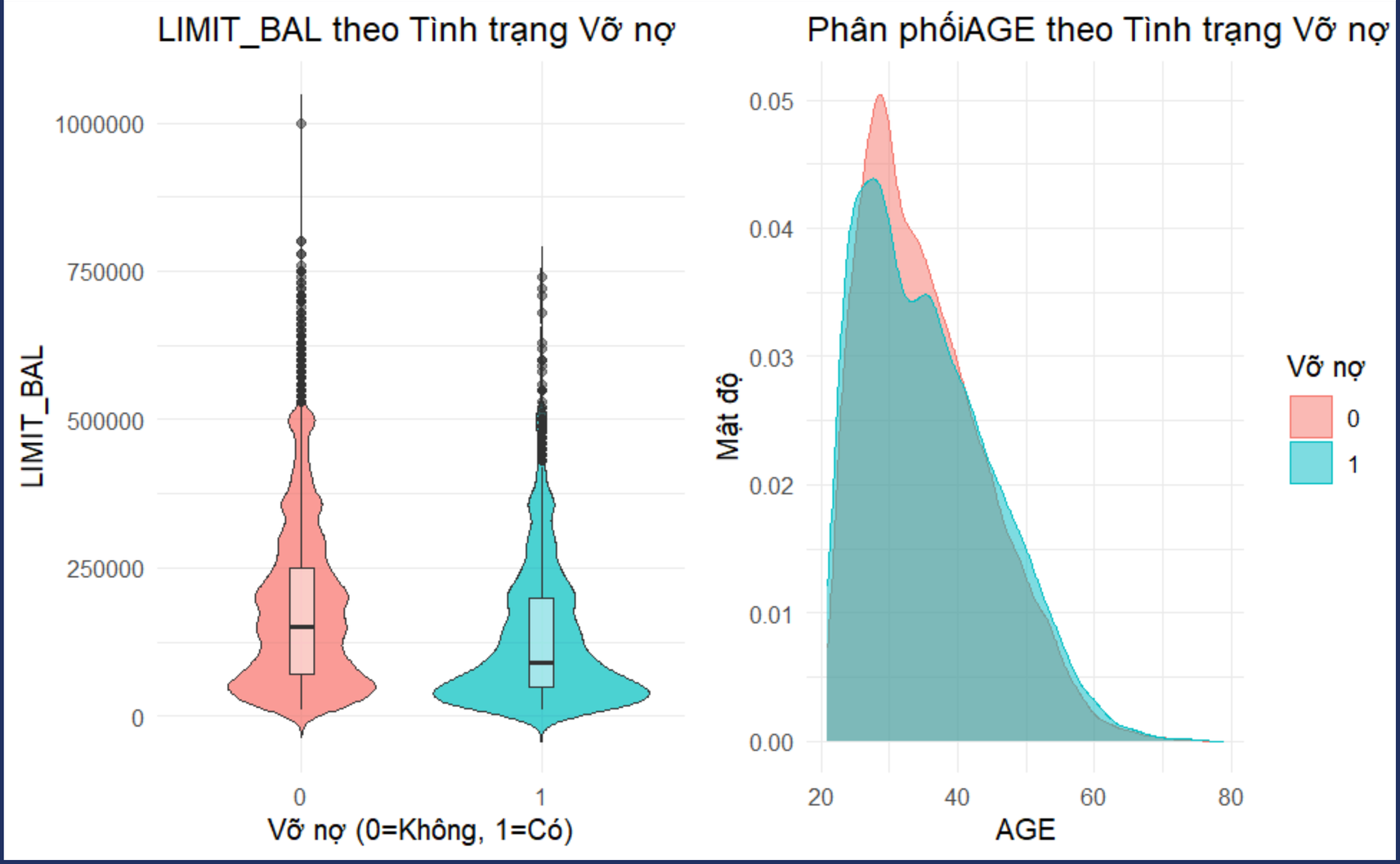
Density Plot (LIMIT_BAL):

Xác nhận phân phối lệch phải của LIMIT_BAL, với mật độ cao ở vùng giá trị âm. Tương tự histogram, điều này nhấn mạnh rằng ngân hàng có xu hướng phân bổ hạn mức tín dụng thấp cho đa số khách hàng, có thể để giảm rủi ro tín dụng.

Boxplot (LIMIT_BAL):

Hiển thị trung vị gần 0, với một số giá trị ngoại lai ở vùng dương (hạn mức cao), Các giá trị ngoại lai này có thể là những khách hàng đặc biệt có thu nhập cao, được ngân hàng cấp hạn mức tín dụng cao hơn nhiều so với trung bình.

Mối quan hệ giữa LIMIT_BAL, AGE theo tình trạng vỡ nợ



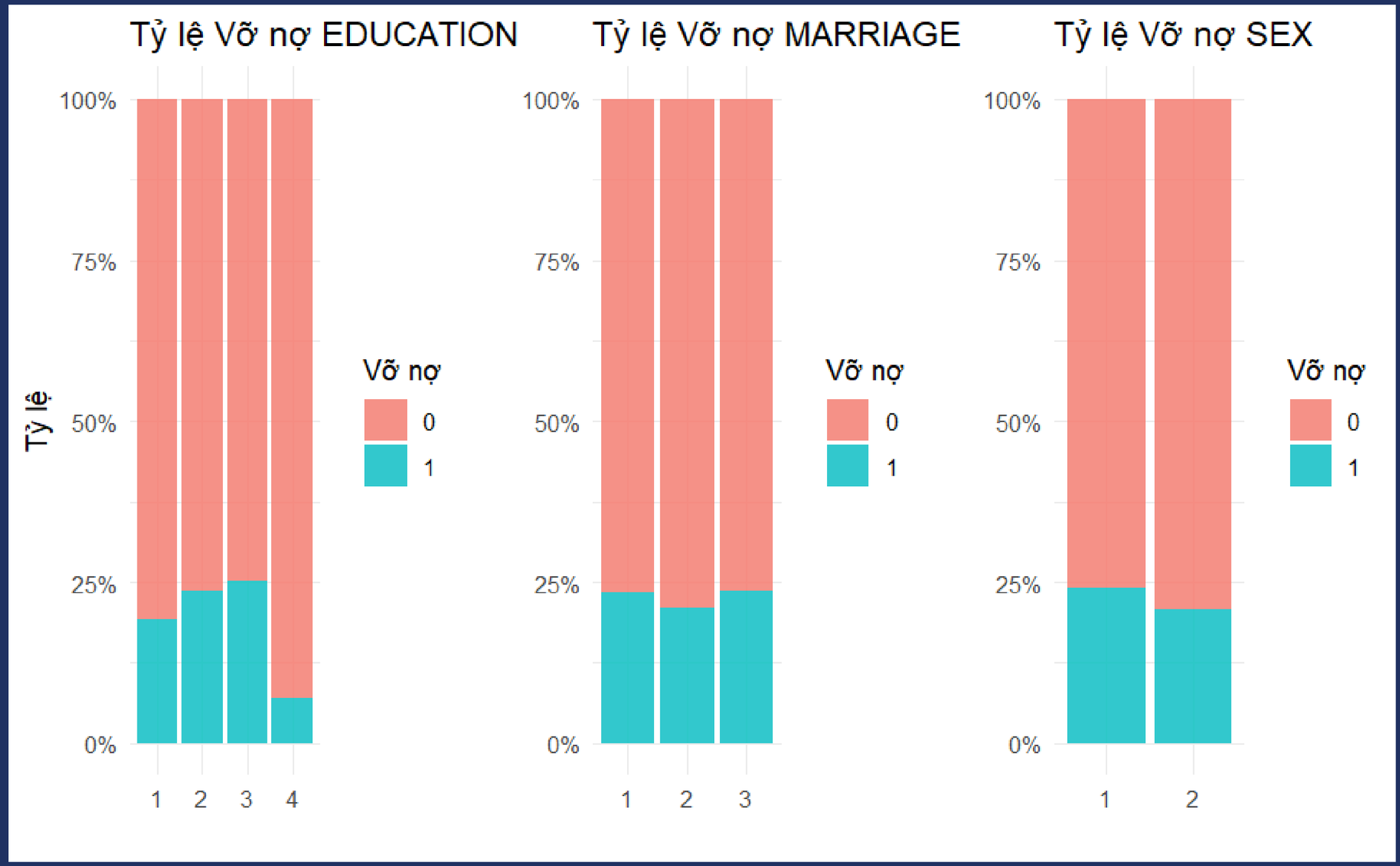
Violin plot(LIMIT_BAL):

Nhóm không vỡ nợ (0) có hạn mức cao hơn (phân phối rộng, giá trị dương), nhóm vỡ nợ (1) tập trung ở hạn mức thấp (giá trị âm), với trung vị thấp hơn, phản ánh rủi ro tài chính cao hơn.

Density plot(AGE):

Nhóm vỡ nợ tập trung ở độ tuổi trẻ (20-30), nhóm không vỡ nợ phân bố rộng hơn (20-60), cho thấy khách hàng trẻ có nguy cơ vỡ nợ cao hơn.

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

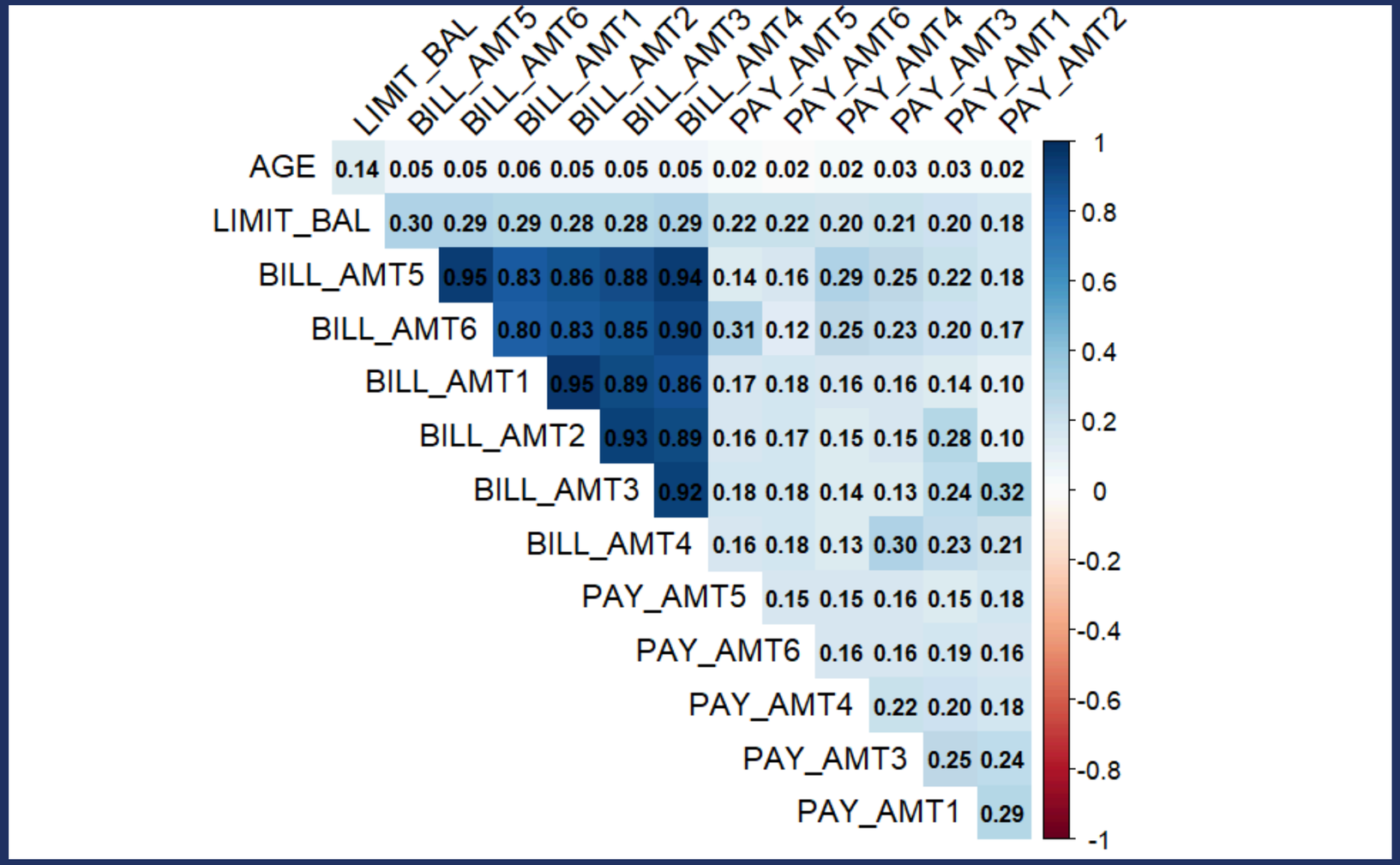


Nhận Xét:

Biểu đồ Stacked barplot

- Tỷ lệ Vỡ nợ EDUCATION: Nhóm 3 (THPT) và 2 (ĐH) có tỷ lệ vỡ nợ (màu xanh) cao hơn nhóm 1 (Sau ĐH) và 4(Khác), phản ánh lứa tuổi còn đi học dùng tiền quá mức hoặc dính vào lừa đảo, tệ nạn xã hội.
- Tỷ lệ Vỡ nợ MARRIAGE: Nhóm 3 (Khác) và nhóm 1 (Kết hôn) có tỷ lệ vỡ nợ cao nhất, so với nhóm 2 (Độc thân) cho thấy việc độc thân ít bị vỡ nợ nhất
- Tỷ lệ Vỡ nợ SEX: Tỷ lệ vỡ nợ giữa nam (1) và nữ (2) tương đối cân bằng, nhưng nữ có phần thấp hơn nhẹ, phản ánh giới tính không phải yếu tố chính ảnh hưởng đến vỡ nợ, nhưng nữ có thể quản lý tài chính tốt hơn một chút.

Ma trận tương quan giữa các biến số



Biểu đồ heatmap hiển thị ma trận tương quan giữa các biến số . 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.

PHẦN 3

LÝ THUYẾT HỌC MÁY

CÁC MÔ HÌNH HUẤN LUYỆN DỰ ĐOÁN

- Hồi Quy Logistic
- Random Forest
- XGBoost

Hồi Quy Logistic

Hồi quy Logistic là một mô hình tuyến tính được sử dụng để phân loại nhị phân

$$P(y = 1|X) = \frac{1}{1 + e^{-z}}$$

Z = β0 + β1*X1 + ... + βn*Xn
với βi là các hệ số của mô hình và Xi là các biến đặc trưng. Mô hình sử dụng hàm mất mát log-loss để tối ưu hóa các tham số, đảm bảo khả năng phân biệt giữa hai lớp.

Mục đích:

Trong bài toán này, mô hình dự đoán xác suất vỡ nợ của khách hàng (0 = Không vỡ nợ, 1 = Vỡ nợ)

- Dữ liệu được tiền xử lý bằng cách chuẩn hóa các biến số (sử dụng scale() để đưa về phân phối chuẩn với trung bình 0 và độ lệch chuẩn 1) nhằm đảm bảo các biến có cùng thang đo.

- Để xử lý mất cân bằng dữ liệu (lớp vỡ nợ chiếm tỷ lệ thấp), kỹ thuật SMOTE (Synthetic Minority Oversampling Technique) được áp dụng với tham số over_ratio = 1, giúp cân bằng tỷ lệ giữa hai lớp

Hồi Quy Logistic

Tác động của các đặc trưng qua hệ số mô hình

Cột Estimate: Giá trị hệ số của từng biến.

- **PAY_0, PAY_1 có hệ số dương (Estimate > 0), cho thấy thanhtoán trễ trong 1-2 tháng gần nhất làm tăng nguy cơ vỡ nợ.**
- **LIMIT_BAL có hệ số âm (Estimate < 0), ngụ ý hạn mức tín dụng cao làm giảm nguy cơ vỡ nợ.**

Cột Pr(>|z|): p-value để kiểm tra ý nghĩa thống kê (p-value < 0.05 cho thấy biến có ý nghĩa thống kê).

Biến	Estimate	p-value	Ý nghĩa thống kê	Ghi chú/Ảnhhưởng chính
Intercept	-1.655	< 2e-16	***	Hệ số chặn, log-odds ban đầu
ID	-0.00000262	0.065	.	Không ý nghĩa rõ ràng, có thể loại bỏ
LIMIT_BAL	-0.0689	4.89e-05	***	Hạn mức tăng → xác suất vỡ nợ giảm
AGE	+0.0743	3.66e-07	***	Tuổi cao hơn → xác suất vỡ nợ tăng

PAY_0	**+0.616**	**< 2e-16**	***	**Trễ hạn gần nhất → ảnh hưởng mạnh nhất**
PAY_2	+0.143	1.55e-15	***	Trễ hạn tháng thứ 2 → nguy cơ tăng
PAY_3	+0.0876	7.54e-06	***	Trễ hạn tháng thứ 3 → nguy cơ tăng
PAY_4	+0.0485	0.0255	*	Trễ hạn tháng thứ 4 → nguy cơ tăng nhẹ
PAY_5	+0.0854	0.000245	***	Trễ hạn tháng thứ 5 → nguy cơ tăng
PAY_6	-0.002	0.918		Không ý nghĩa
BILL_AMT1	-0.514	2.55e-14	***	Hóa đơn cao tháng 1 → nguy cơ giảm
BILL_AMT2	+0.1858	0.0312	*	Hóa đơn tháng 2 → nguy cơ tăng nhẹ

BILL_AMT3	**+0.1636**	**0.0229**	*	**Hóa đơn tháng 3 → nguy cơ tăng nhẹ**
BILL_AMT4	-0.0401	0.566		Không ý nghĩa
BILL_AMT5	-0.0569	0.479		Không ý nghĩa
PAY_AMT6	-0.03599	0.054	.	Biên độ yếu
PAY_AMT1	-0.245	4.25e-16	***	Trả nhiều tháng 1 → giảm nguy cơ
PAY_AMT2	-0.263	3.01e-12	***	Trả nhiều tháng 2 → giảm nguy cơ
PAY_AMT3	-0.0072	0.744		Không ý nghĩa
PAY_AMT4	-0.0553	0.0099	**	Trả nhiều tháng 4 → giảm nhẹ nguy cơ

PAY_AMT5	**-0.077**	**0.0006**	***	**Trả nhiều tháng 5 → giảm nguy cơ**
CREDIT_UTILIZATION	+0.0953	0.303		Không ý nghĩa
TOTAL_DELAY	-0.144	6.80e-12	***	Tổng ngày chậm trả giảm → nguy cơ giảm
AVG_BILL	NA	NA	NA	Bị loại vì trùng lặp (đa cộng tuyến)
AVG_PAY	NA	NA	NA	Bị loại vì trùng lặp
EDUCATION_1	+1.243	< 2e-16	***	So với baseline (EDUCATION_4), nhóm 1 có nguy cơ cao hơn
EDUCATION_2	+1.184	2.73e-16	***	Nhóm 2 cũng nguy cơ cao hơn
EDUCATION_3	+1.120	2.46e-14	***	Tương tự

EDUCATION_4	**NA**	**NA**	**NA**	**Baseline → không cần hiển thị hệ số**
MARRIAGE_1	+0.1973	0.0807	.	Có thể liên quan nhẹ
MARRIAGE_2	-0.0119	0.917		Không ý nghĩa
MARRIAGE_3	NA	NA	NA	Baseline
SEX_1 (Nam)	+0.101	7.94e-05	***	Nam có nguy cơ vỡ nợ cao hơn nữ
SEX_2 (Nữ)	NA	NA	NA	Baseline

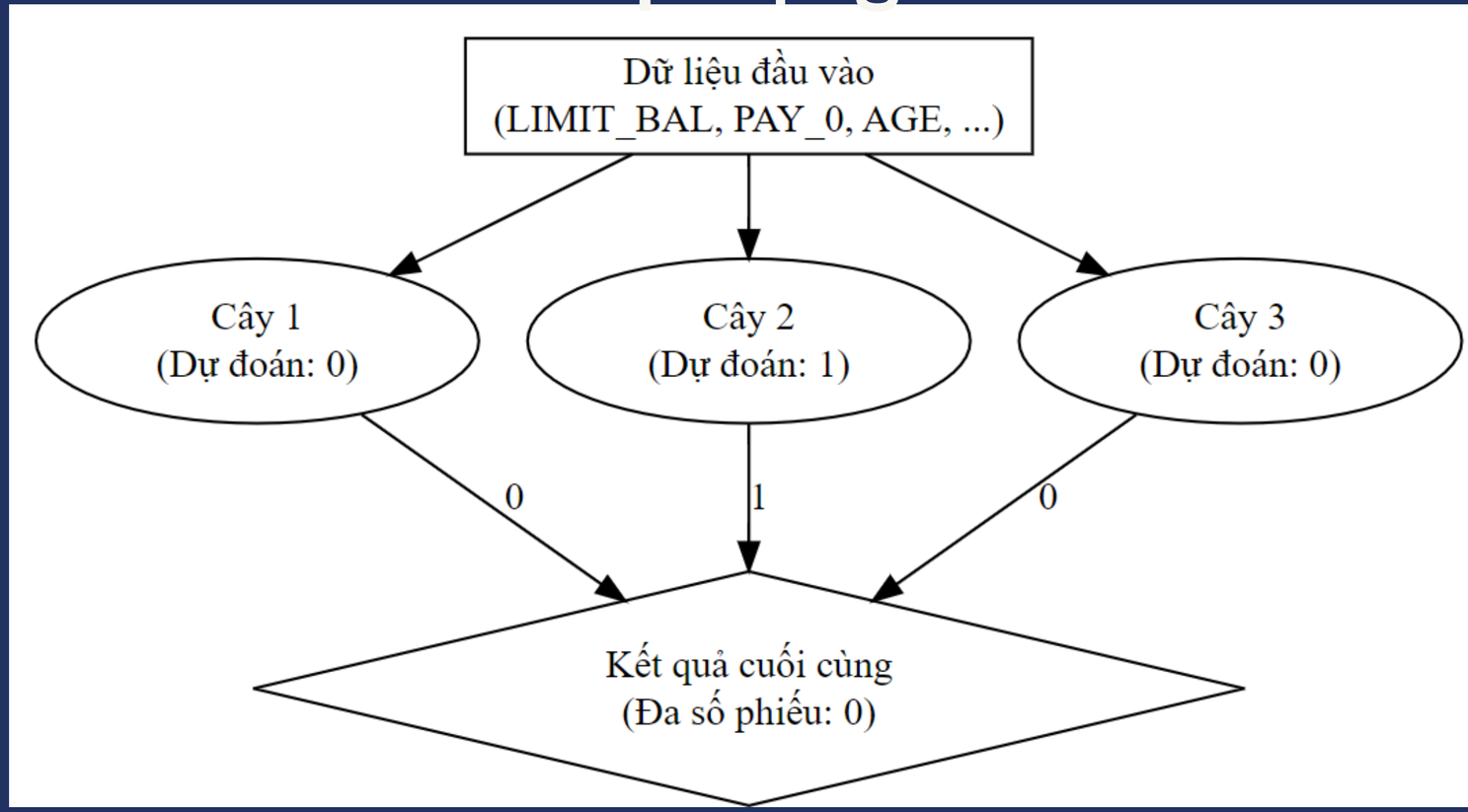
*****: Rất có ý nghĩa (p < 0.001)**
****: Có ý nghĩa (p < 0.01)**
*** : Có ý nghĩa (p < 0.05)**
. : Gần có ý nghĩa (p < 0.1)

Random Forest

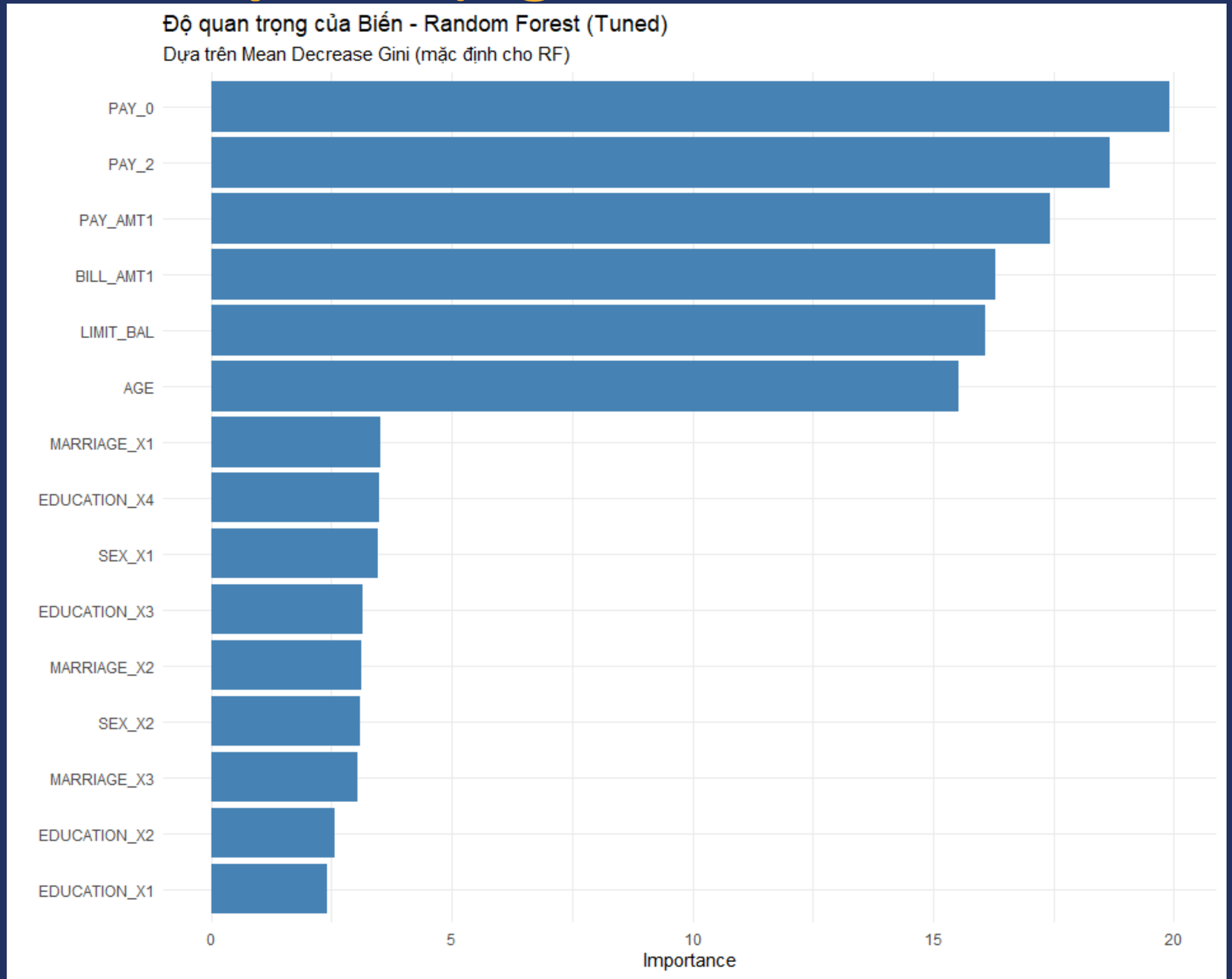
Là một tập hợp mô hình học máy dựa trên cây quyết định, cải thiện độ chính xác và giảm overfitting. Thích hợp để phân tích các mối quan hệ phi tuyến tính và xác định tầm quan trọng của các đặc trưng.

Mô hình được tinh chỉnh bằng phương pháp Grid Search với xác thực chéo 5 lần để tối ưu hóa số liệu ROC-AUC. Grid Search thử nghiệm các tổ hợp tham số trong không gian: mtry từ 3 đến 10, trees từ 500 đến 1500, và min_n từ 5 đến 20. Quá trình này đảm bảo mô hình đạt hiệu suất tốt nhất trên dữ liệu huấn luyện.

Áp dụng



Tầm quan trọng của biến



- PAY_0 và PAY_2 (lịch sử thanh toán trễ gần nhất) có tác động mạnh nhất, tiếp theo là PAY_AMT1, BILL_AMT1.
- Các biến như LIMIT_BAL, AGE cũng đóng góp đáng kể, trong khi SEX, EDUCATION, MARRIAGE ít ảnh hưởng hơn

=> giúp ngân hàng cải thiện quy trình đánh giá tín dụng

XGBoost

XGBoost (Extreme Gradient Boosting) là một thuật toán học máy tiên tiến dựa trên nguyên lý Gradient Boosting.

Mô hình tối ưu hóa một hàm mất mát (log-loss cho phân loại nhị phân) bằng cách sử dụng gradient descent, đồng thời áp dụng các kỹ thuật như regularization (điều chuẩn) và xử lý song song để tăng hiệu suất

$$L(y, \hat{y}) = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})]$$

y: Nhãn thực tế (`0` hoặc `1`).

$\hat{y}$: Xác suất dự đoán.

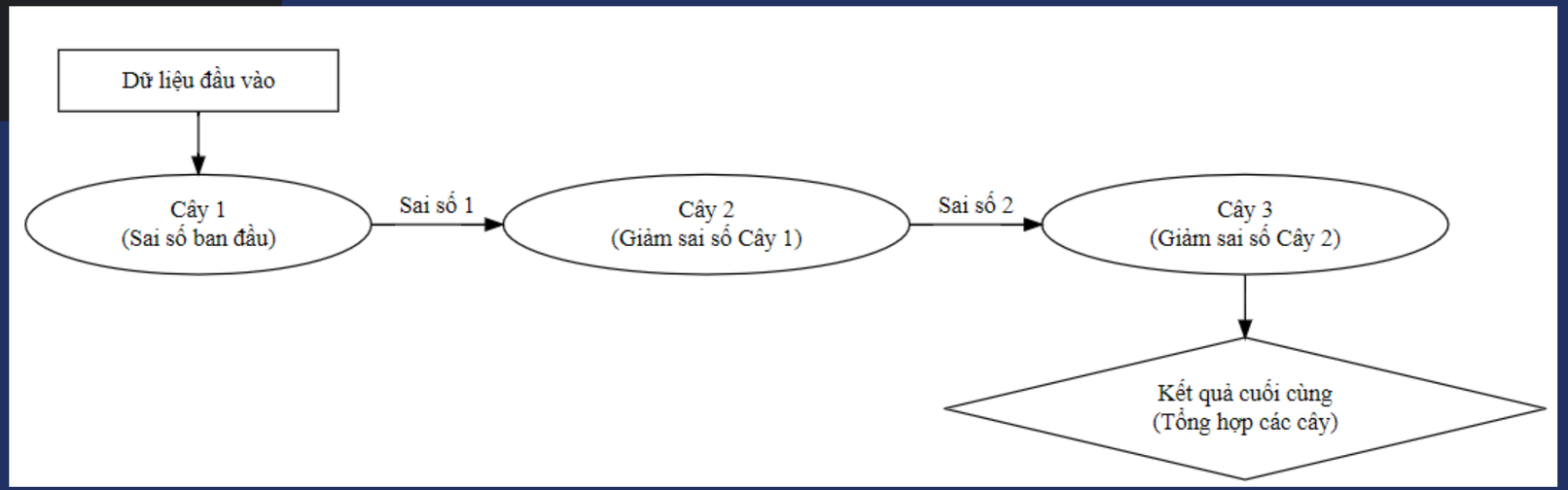
Mô hình được tinh chỉnh bằng phương pháp Random Search với :

- 20 tổ hợp ngẫu nhiên
- kết hợp với xác thực chéo 5 lần để tối ưu hóa ROC-AUC.

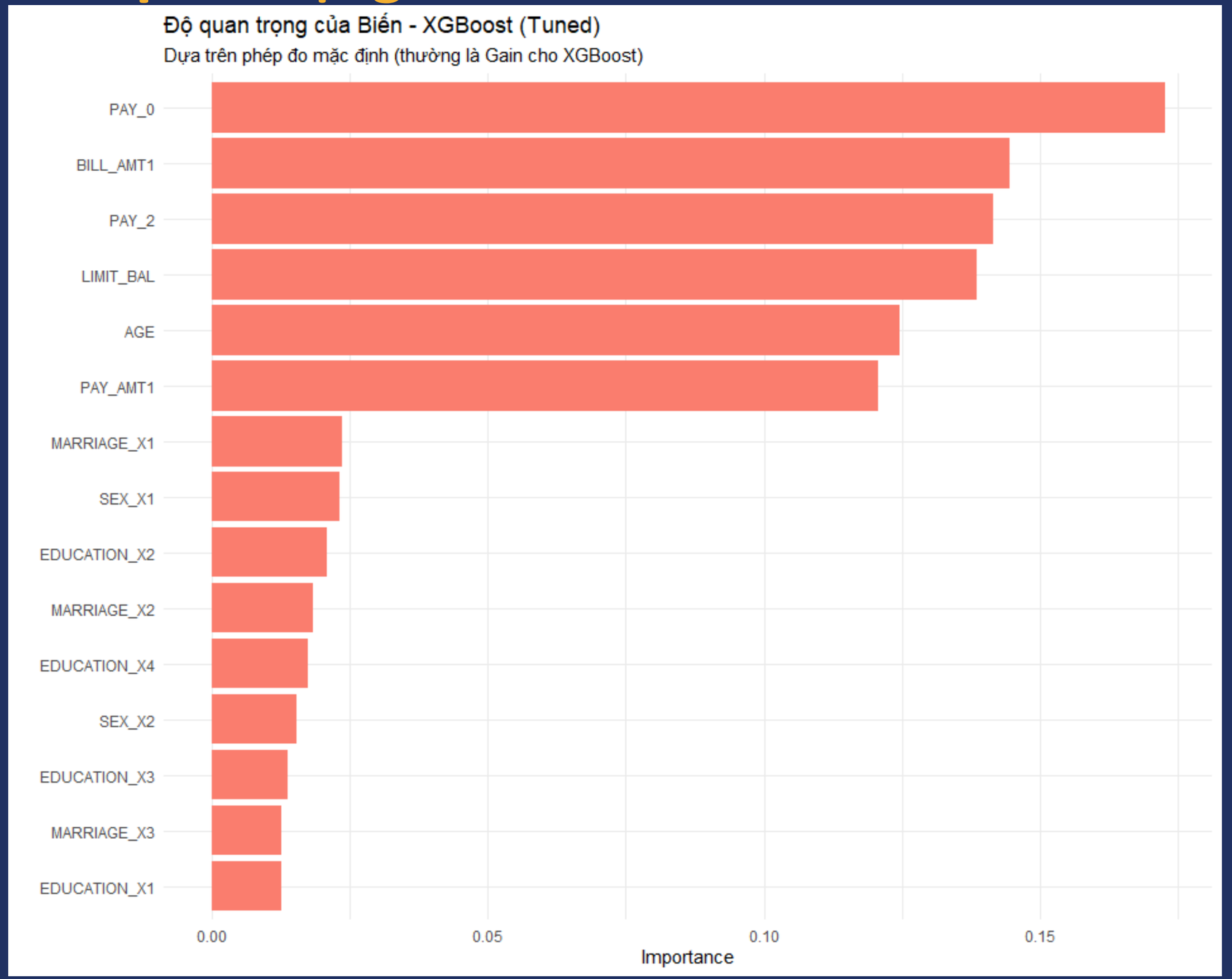
Không gian tham số bao gồm:

- mtry từ 0.6 đến 1.0
- trees từ 100 đến 1000
- min_n từ 1 đến 10
- learn_rate từ 0.01 đến 0.3, tree_depth từ 3 đến 9
- sample_size từ 0.6 đến 1.0.

=> Phương pháp này giúp tiết kiệm thời gian tính toán mà vẫn tìm được tổ hợp tham số tối ưu.



Tầm quan trọng của biến



XGBoost

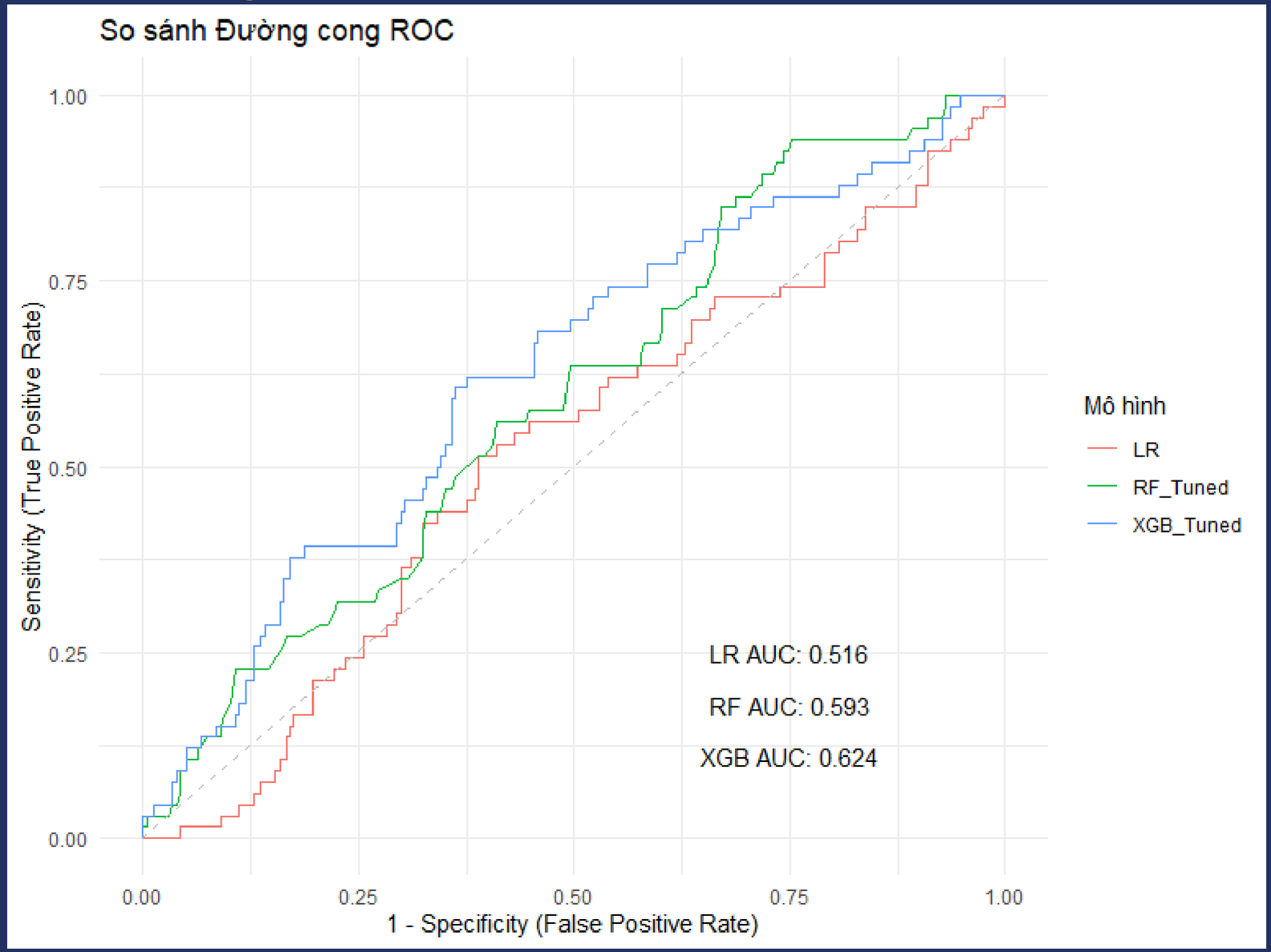
Nhận xét

Biểu đồ độ quan trọng biến từ mô hình XGBoost với thước đo Gain cho thấy:

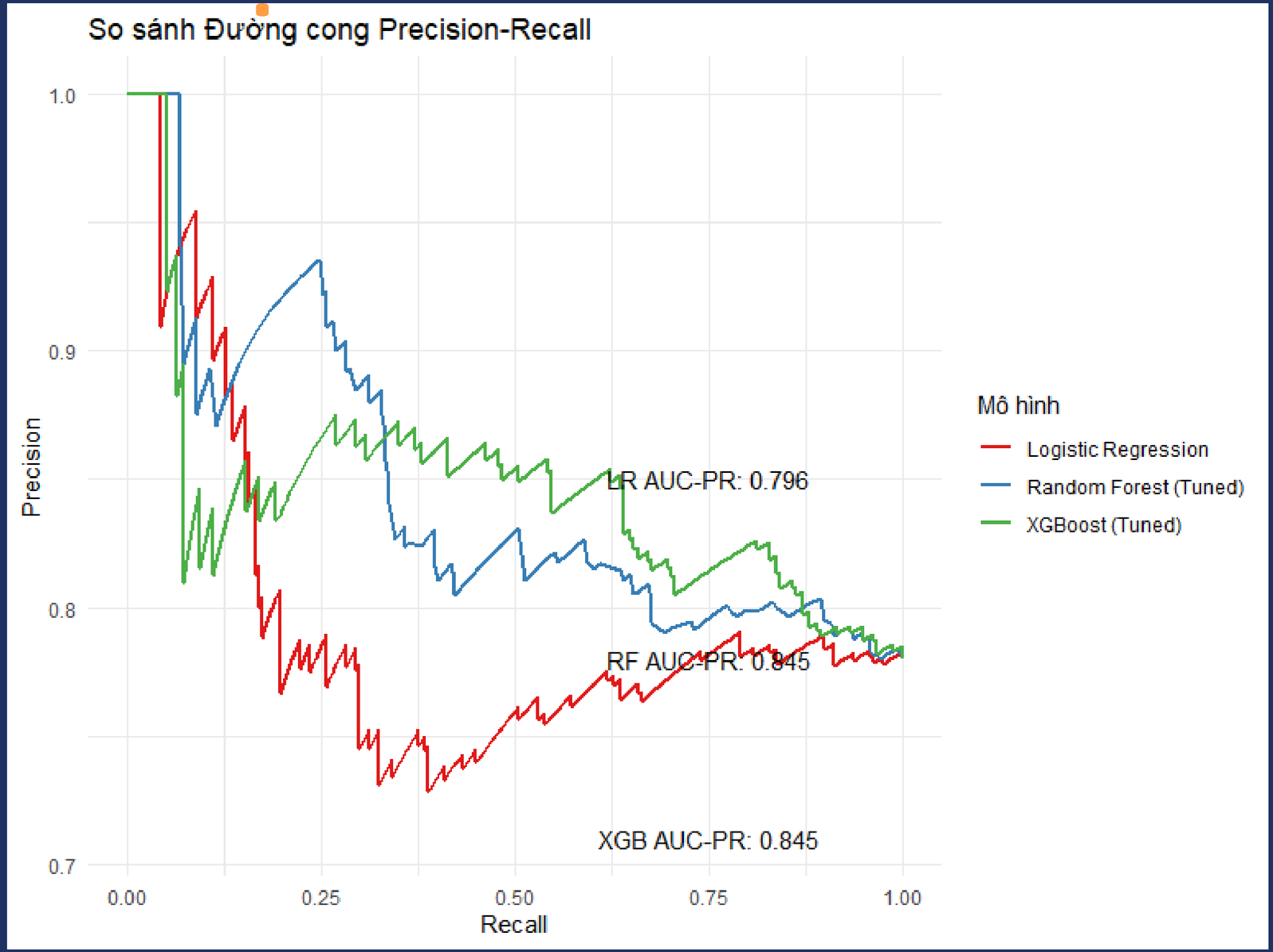
- **PAY_0 và BILL_AMT1 là yếu tố quan trọng nhất, nhấn mạnh vai trò của thanh toán trễ tháng gần nhất và số dư hóa đơn**
- **PAY_2 và LIMIT_BAL cũng ảnh hưởng đáng kể, trong khi AGE có tác động trung bình**
- **Các biến nhân khẩu học như MARRIAGE_X1, SEX_X1, và EDUCATION_X2 có mức độ quan trọng thấp.**

=> Kết quả khẳng định hành vi tài chính (thanh toán trễ, số dư) là yếu tố chính dự đoán vỡ nợ, hữu ích cho việc giám sát tín dụng, dù các biến nhân khẩu học cần phân tích thêm về tương tác.

So Sánh độ chính xác



So Sánh độ chính xác



So Sánh độ chính xác

Mô hình	Accuracy	Precision	Recall	F1-Score	AUC-ROC	AUC-PR
Random Forest	0.6714	0.3692	0.6845	0.4797	0.7341	0.6485
Hồi quy logistic	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

KẾT LUẬN

Ưu điểm

- Ứng dụng thực tế cao, sử dụng dữ liệu tín dụng thật.
- Quy trình xử lý bài bản: từ tiền xử lý, EDA, trực quan, đến mô hình hóa.
- Áp dụng đa mô hình: Logistic Regression, Random Forest, XGBoost.
- Trực quan hóa rõ ràng, dễ hiểu qua biểu đồ đa dạng.
- Phân tích đặc trưng hợp lý, hỗ trợ ra quyết định tín dụng.

Hạn chế

- Độ chính xác chưa cao ở một số mô hình (như Logistic).
- Precision và Recall lớp vỡ nợ còn thấp, dễ bỏ sót khách hàng rủi ro.
- Chưa thử nghiệm thêm mô hình khác để so sánh toàn diện.
- Chưa triển khai thực tế, mới dừng ở mức mô phỏng.

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.

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

Thank You!