Bài tập lớn: Quản trị rủi ro định lượng 2

Nguyễn Thị Trà - 11226303

Tóm tắt nội dung

Thẻ tín dụng ngày càng được sử dụng phổ biến nhờ vào chức năng chi tiêu vượt hạn mức, qua đó hình thành mối quan hệ tín dụng giữa khách hàng và các tổ chức tài chính. Tuy nhiên, việc không thanh toán đúng hạn khoản nợ thẻ tín dụng có thể dẫn đến những hậu quả tiêu cực, như ảnh hưởng xấu đến hồ sơ tín dụng của khách hàng và gây thiệt hại kinh tế cho các tổ chức tài chính. Bài nghiên cứu này phân tích dữ liệu về vỡ nợ thẻ tín dụng tại Đài Loan trong giai đoạn từ tháng 4 đến tháng 9 năm 2005, với mục tiêu hỗ trợ các tổ chức tài chính nâng cao hiệu quả trong việc giám sát và quản lý rủi ro tín dụng thẻ.

1 Giới thiệu

Trong xã hội hiện đại, thẻ tín dụng đóng vai trò thiết yếu trong các giao dịch hàng ngày của mỗi cá nhân, từ mua sắm trực tiếp cho đến các giao dịch trực tuyến. Tuy nhiên, việc cấp tín dụng có thể tiềm ẩn rủi ro nếu khách hàng không thể thanh toán khoản nợ đúng hạn. Điều này tạo ra rủi ro tín dụng, có thể gây thiệt hại tài chính đáng kể cho các ngân hàng và tổ chức tài chính. Để hạn chế rủi ro từ các khách hàng có khả năng không trả nợ, các ngân hàng và tổ chức tài chính thường áp dụng quy trình xét duyệt tín dụng nghiêm ngặt. Rủi ro tín dụng xảy ra khi khách hàng không thực hiện nghĩa vụ thanh toán đúng hạn trong một khoảng thời gian nhất định. Việc không dự báo được khả năng vỡ nợ trong tương lai có thể gây ra thiệt hại tài chính nghiêm trọng cho doanh nghiệp. Do đó, việc xây dựng mô hình dự đoán khả năng vỡ nợ trở thành nhiệm vụ quan trọng, giúp các công ty quyết định có nên cấp thẻ tín dụng cho khách hàng hay không.

Mục tiêu của nghiên cứu này là xây dựng mô hình dự đoán tối ưu nhất, sử dụng các thông tin khách hàng thu thập được từ quá trình đăng ký thẻ tín dụng. Từ đó, giúp các ngân hàng và tổ chức tài chính cải thiện khả năng quản lý rủi ro tín dụng, giảm thiểu thiệt hại tài chính từ các khoản nợ xấu. Việc xây dựng mô hình dự đoán khả năng vỡ nợ không chỉ giúp các tổ chức tài chính ra quyết định chính xác hơn trong việc cấp tín dụng mà còn hỗ trợ quá trình phát triển các chiến lược tín dụng hiệu quả hơn, tối ưu hóa việc sử dụng nguồn lực tài chính.

Về bản chất, đây là một bài toán phân loại nhị phân, nhằm dự đoán khả năng khách hàng không thanh toán khoản nợ trong kỳ thanh toán kế tiếp. Dữ liệu nghiên cứu sẽ bao gồm các đặc điểm cá nhân của khách hàng và lịch sử tín dụng, cùng với các yếu tố tài chính liên quan. Các mô hình phân tích sẽ được xây dựng và đánh giá bao gồm hồi quy Logistic trên biến gốc và WOE, KNN, Random Forest. Nghiên cứu tiến hành huấn luyện và so sánh hiệu quả nhiều mô hình dự đoán dựa trên một số chỉ tiêu như ma trận nhầm lẫn, đường cong ROC, các chỉ số AUC, accuracy, sensitivity và gini.

2 Cơ sở lý thuyết

2.1 Khái niệm về vỡ nợ

Khi một khách hàng chấp nhận sử dụng thẻ tín dụng do ngân hàng hoặc tổ chức phát hành cung cấp, họ đồng ý với một số điều khoản nhất định, trong đó yêu cầu thực hiện khoản thanh toán tối thiểu trước ngày đến hạn được ghi trên bảng sao kê thẻ tín dụng. Nếu khách hàng không thanh toán khoản nợ đúng hạn, tổ chức phát hành sẽ đánh dấu tài khoản thẻ tín dụng đó là vỡ nợ, có thể áp dụng mức lãi suất phạt, giảm hạn mức tín dụng, và trong trường hợp vi phạm nghiêm trọng, thậm chí đóng tài khoản thẻ.

Trong nỗ lực mở rộng thị phần, các ngân hàng phát hành thẻ đôi khi cấp thẻ tín dụng cho những khách hàng không đủ điều kiện mà không thẩm định kỹ lưỡng khả năng thanh toán nợ của họ. Khi những chủ thẻ này chi tiêu quá mức để mua hàng hóa, dịch vụ mà không cân nhắc khả năng hoàn trả, họ sẽ tích lũy các khoản nợ lớn ngoài khả năng chi trả. Do đó, việc các ngân hàng phát hành thẻ tín dụng dự báo chính xác khả năng vỡ nợ của khách hàng đóng vai trò then chốt đối với công tác phân tích rủi ro và ra quyết định phê duyệt hồ sơ cấp thẻ.

2.2 Tổng quan các mô hình

Mục tiêu chính của nghiên cứu này là so sánh kết quả của mô hình, bao gồm Hồi quy Logistic trên biến gốc và WOE, KNN (K-nearest Neighbor), Random Forest, dựa trên độ chính xác trong phân loại và hiệu quả trong việc dự đoán xác suất thực tế mà một cá nhân thuộc vào nhóm thực sự.

Để đánh giá độ chính xác trong phân loại, các chỉ số như tỷ lệ dự đoán đúng phân loại dựa trên ma trận nhầm lẫn, đường cong ROC, diện tích dưới đường cong (AUC), accuracy, sensitivity và gini được sử dụng.

2.2.1 Weight of Evidence - WOE

WOE (Weight of Evidence) là một kỹ thuật phổ biến trong việc tạo đặc trưng và lựa chọn đặc trưng, đặc biệt hiệu quả trong việc xây dựng mô hình scorecard. Phương pháp này xếp hạng các biến theo các mức độ khác nhau như mạnh, trung bình, yếu, không có tác động, dựa trên khả năng dự báo nợ xấu. Tiêu chí để xếp hạng các biến là chỉ số giá trị thông tin IV (information value), được tính toán từ phương pháp WOE. Đồng thời, phương pháp này cũng tạo ra các đặc trưng cho từng biến, giúp đo lường sự khác biệt trong phân phối giữa nhóm **default** và **not default**. Cụ thể, phương pháp WOE sẽ áp dụng các kỹ thuật xử lý khác nhau đối với biến liên tục và biến phân loại:

Với biến liên tục: WOE sẽ gán nhãn cho mỗi quan sát theo giá trị của bin mà nó thuộc về. Các bin này là các khoảng liên tiếp, được xác định sao cho số lượng quan sát trong mỗi bin là đều nhau. Để xác định các bin, ta cần quyết định số lượng bin, và đầu mút của các khoảng này có thể được hình dung là các quantile.

Với biến phân loại: WOE có thể xem mỗi nhóm là một bin, hoặc nhóm các nhóm có số lượng quan sát ít vào cùng một bin. Hơn nữa, mức độ chênh lệch giữa phân phối **default/ not default**, được đo bằng chỉ số WOE, có thể giúp nhận diện các nhóm có tính chất phân loại tương tự. Nếu giá trị WOE của các nhóm gần giống nhau, chúng có thể được nhóm lại. Trường hợp Null cũng có thể được xem như một nhóm riêng biệt nếu số lượng của nó đủ lớn, hoặc có thể nhóm nó vào các nhóm khác nếu nó là thiểu số.

IV (Information Value): Chỉ số giá trị thông tin, có tác dụng đánh giá một biến có sức mạnh trong việc phân loại nợ xấu hay không. Công thức tính IV:

$$\text{IV} = \sum_{i=1}^{n} (\%\text{Good}_i - \%\text{Bad}_i) \times \text{WOE}_i$$

Ta nhận thấy IV luôn nhận giá trị dương vì WOE_i và $(\%\text{Good}_i - \%\text{Bad}_i)$ đồng biến. Giá trị IV sẽ cho ta biết mức độ chênh lệch của %Good và %Bad ở mỗi bin là nhiều hay ít. Nếu IV cao thì sự khác biệt trong phân phối giữa %Good và %Bad sẽ lớn và biến hữu ích hơn trong việc phân loại hồ sơ, và trái lại IV nhỏ thì biến ít hữu ích trong việc phân loại hồ sơ. Một số tài liệu cũng đưa ra tiêu chuẩn phân loại sức mạnh của biến theo giá trị IV như bên dưới:

- $IV \leq 0.02$: Biến không có tác dụng trong việc phân loại hồ sơ (Good/Bad)
- $IV \in (0.02, 0.1]$: yếu
- $IV \in (0.1, 0.3]$: trung bình
- $IV \in (0.3, 0.5]$: mạnh
- $IV > 0.5$: Biến rất mạnh, tuy nhiên trường hợp này cần được điều tra lại để tránh trường hợp biến có mối quan hệ trực tiếp để định nghĩa hồ sơ (good/bad).

2.2.2 Mô hình hồi quy Logistic

Mô hình Binary Logistic là mô hình hồi quy sử dụng biến phụ thuộc dạng nhị phân để ước lượng xác suất một sự kiện sẽ xảy ra với những thông tin của biến độc lập $(X_1, X_2, ..., X_n)$ có được. Từ biến phụ thuộc nhị phân, một thủ tục sẽ được dùng để dự đoán xác suất sự kiện xảy ra theo quy tắc nếu xác suất được dự đoán lớn hơn 0.5 thì kết quả dự đoán sẽ là "có"xảy ra sự kiện, và ngược lại thì kết quả dự đoán sẽ là "không".

Xác suất để sự kiện xảy ra $P(Y = 1)$ như sau:

$$P(Y = 1) = \frac{e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)}}{1 + e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)}}$$

Xác suất không xảy ra sự kiện $P(Y = 0)$ như sau:

$$P(Y = 0) = 1 - \frac{e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)}}{1 + e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)}}$$

Khi sử dụng để dự báo khả năng vỡ nợ của khách hàng, $P(Y)$ càng lớn hơn 0.5 thì nguy cơ khách hàng vỡ nợ càng cao và ngược lại.

2.2.3 Mô hình KNN

K-Nearest Neighbors (KNN) là một trong những thuật toán học có giám sát đơn giản nhất, được ứng dụng rộng rãi trong khai phá dữ liệu và học máy. Đây là một thuật toán **không tham số** (non-parametric) thường được sử dụng trong các bài toán phân loại.

Đối với một mẫu dữ liệu mới, KNN tiến hành tìm kiếm các điểm dữ liệu trong khu vực lân cận gần nhất và gán nhãn của mẫu đó dựa trên nhãn của các "láng giềng" gần nhất. Tham số k là số nguyên, quy định số lượng điểm láng giềng cần xét.

- Với $k = 1$, mẫu mới sẽ được gán nhãn của láng giềng duy nhất gần nhất.
- Với $k > 1$, nhãn được xác định bằng cách **bỏ phiếu đa số** (majority voting) từ k láng giềng, hoặc thông qua **gán trọng số** theo khoảng cách.

Trong nghiên cứu này, vòng lặp `for` sẽ được sử dụng để kiểm tra hiệu suất mô hình với nhiều giá trị k khác nhau, từ đó lựa chọn giá trị k tối ưu dựa trên độ chính xác.

Thuật toán KNN dựa trên giả định rằng: *các điểm dữ liệu có đặc điểm tương tự sẽ nằm gần nhau trong không gian đặc trưng*. Do đó, việc tìm ra các điểm lân cận gần nhất đòi hỏi xác định khoảng cách giữa các điểm dữ liệu. Một số công thức phổ biến để tính khoảng cách giữa hai điểm dữ liệu x và y (với k thuộc tính) bao gồm:

- **Khoảng cách Euclid (Euclidean distance):**

$$d(x, y) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

- **Khoảng cách Manhattan (Manhattan distance):**

$$d(x, y) = \sum_{i=1}^{k} |x_i - y_i|$$

- **Khoảng cách Minkowski (Minkowski distance):**

$$d(x, y) = \left(\sum_{i=1}^{k} |x_i - y_i|^q\right)^{1/q}$$

với q là một tham số xác định mức độ tổng quát hóa giữa các công thức khoảng cách.

2.2.4 Mô hình Random Forest

Mô hình Random Forest được huấn luyện dựa trên sự phối hợp giữa luật kết hợp (ensembling) và quá trình lấy mẫu tái lặp (bootstrapping). Cụ thể thuật toán này tạo ra nhiều cây quyết định mà mỗi cây quyết định được huấn luyện dựa trên nhiều mẫu con khác nhau và kết quả dự báo là bầu cử (voting) từ toàn bộ những cây quyết định. Như vậy một kết quả dự báo được tổng hợp từ nhiều mô hình nên kết quả của chúng sẽ không bị chệch. Đồng thời kết hợp kết quả dự báo từ nhiều mô hình sẽ có phương sai nhỏ hơn so với chỉ một mô hình. Điều này giúp cho mô hình khắc phục được hiện tượng quá khớp.

2.3 Phương pháp chấm điểm

Điểm tín nhiệm (credit score) cho từng khách hàng được tính toán dựa trên từng đặc trưng (feature), cụ thể là các bin trong biến liên tục hoặc nhãn của biến phân loại. Việc tính điểm được chuẩn hóa theo công thức:

$$\text{Score} = \mathit{offset} + \text{Factor} \cdot \ln \mathit{odds} \tag{1}$$

Với $\mathit{odds} = \frac{p}{1-p}$ là tỷ lệ dự báo default/ not default.

2.4 Các chỉ tiêu đánh giá

Các chỉ số đánh giá hiệu suất phân loại so sánh nhãn thực tế của từng quan sát với dự đoán của mô hình phân loại. Để minh họa sự tương quan giữa dự đoán nhị phân và phân phối thực tế, ta có thể xây dựng một ma trận nhầm lẫn (confusion matrix).

	Thực tế 0 (Actual Negative)	Thực tế 1 (Actual Positive)
Dự đoán 0 (Predicted Negative)	TN	FN
Dự đoán 1 (Predicted Positive)	FP	TP

Một số thuật ngữ thường dùng trong ma trận nhầm lẫn:

- **True Positive (TP)**: số lượng mẫu thực sự 1 được dự đoán chính xác.
- **False Negative (FN)**: số lượng mẫu thực sự 1 bị dự đoán sai thành 0.
- **False Positive (FP)**: số lượng mẫu thực sự 0 bị dự đoán sai thành 1.
- **True Negative (TN)**: số lượng mẫu thực sự 0 được dự đoán chính xác.

Từ ma trận nhầm lẫn, ta sử dụng một số chỉ số để so sánh kết quả các mô hình.
Accuracy thể hiện khả năng dự báo đúng của mô hình, được tính bằng:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP}$$

Sensitivity (Recall, True Positive Rate) được tính bằng công thức:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity (True Negative Rate) được tính bằng công thức:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Chỉ số AUC (area under curve) đo lường phần diện tích nằm dưới đường cong ROC cho biết khả năng phân loại của các tín dụng default/ not default của mô hình là mạnh hay yếu. AUC $\in [0, 1]$, giá trị của nó càng lớn thì mô hình càng tốt.

Hệ số Gini là phép chuyển đổi trực tiếp từ AUC. Nó dựa trên khả năng phân biệt giữa các trường hợp tích cực và tiêu cực của mô hình, được biểu diễn bằng đường cong ROC. Phép tính bao gồm tính diện tích dưới đường cong ROC và áp dụng công thức để chuyển đổi thành hệ số Gini: $Gini = (AUC \times 2) - 1$.

3 Dữ liệu

3.1 Tổng quan dữ liệu

Bài nghiên cứu sử dụng dữ liệu *Default of Credit Card Clients* của UCI Machine Learning Repository, có sẵn tại liên kết sau. Tập dữ liệu gốc bao gồm 30,000 quan sát, nhưng trong khuôn

khổ bài tập, tôi đã lựa chọn ngẫu nhiên 15,000 quan sát để phân tích. Mỗi quan sát có 24 thuộc tính chứa thông tin về tình trạng vỡ nợ, các yếu tố nhân khẩu học, dữ liệu tín dụng, lịch sử thanh toán và các sao kê thẻ tín dụng của các khách hàng tại Đài Loan từ tháng 4 năm 2005 đến tháng 9 năm 2005.

Nhóm các biến đầu tiên chứa thông tin về thông tin cá nhân của khách hàng:

- **ID**: ID của mỗi khách hàng, biến phân loại.
- **LIMIT_BAL**: Số tiền tín dụng cấp cho khách hàng (đơn vị: Đài tệ NT), bao gồm tín dụng cá nhân và tín dụng gia đình/bổ sung.
- **SEX**: Giới tính, biến phân loại (1 = nam, 2 = nữ).
- **EDUCATION**: Trình độ học vấn, biến phân loại (1 = sau đại học, 2 = đại học, 3 = trung học, 4 = khác, 5 = không xác định).
- **MARRIAGE**: Tình trạng hôn nhân, biến phân loại (1 = đã kết hôn, 2 = độc thân, 3 = khác).
- **AGE**: Tuổi (năm), biến số học.

Nhóm các thuộc tính tiếp theo liên quan đến tình trạng thanh toán chậm trong các tháng cụ thể:

- **PAY_0**: Tình trạng thanh toán trong tháng 9 năm 2005 (-1 = thanh toán đúng hạn, 1 = chậm một tháng, 2 = chậm hai tháng, ..., 8 = chậm tám tháng, 9 = chậm chín tháng trở lên).
- **PAY_2**: Tình trạng thanh toán trong tháng 8 năm 2005 (theo thang đo như trên).
- **PAY_3**: Tình trạng thanh toán trong tháng 7 năm 2005 (theo thang đo như trên).
- **PAY_4**: Tình trạng thanh toán trong tháng 6 năm 2005 (theo thang đo như trên).
- **PAY_5**: Tình trạng thanh toán trong tháng 5 năm 2005 (theo thang đo như trên).
- **PAY_6**: Tình trạng thanh toán trong tháng 4 năm 2005 (theo thang đo như trên).

Các biến khác liên quan đến số tiền trong các sao kê hàng tháng (tức là báo cáo hàng tháng mà các công ty thẻ tín dụng gửi cho chủ thẻ):

- **BILL_AMT1**: Số tiền sao kê trong tháng 9 năm 2005 (Đài tệ NT).
- **BILL_AMT2**: Số tiền sao kê trong tháng 8 năm 2005 (Đài tệ NT).
- **BILL_AMT3**: Số tiền sao kê trong tháng 7 năm 2005 (Đài tệ NT).
- **BILL_AMT4**: Số tiền sao kê trong tháng 6 năm 2005 (Đài tệ NT).
- **BILL_AMT5**: Số tiền sao kê trong tháng 5 năm 2005 (Đài tệ NT).
- **BILL_AMT6**: Số tiền sao kê trong tháng 4 năm 2005 (Đài tệ NT).

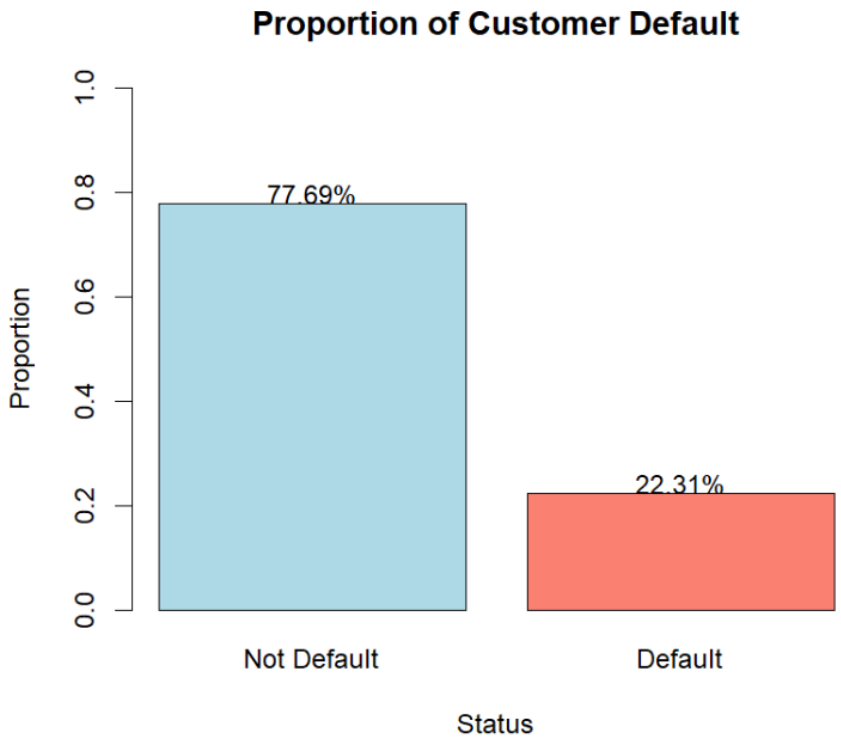
Các biến tiếp theo liên quan đến số tiền thanh toán trong các tháng cụ thể:

- **PAY_AMT1**: Số tiền thanh toán trong tháng 9 năm 2005 (Đài tệ NT).
- **PAY_AMT2**: Số tiền thanh toán trong tháng 8 năm 2005 (Đài tệ NT).
- **PAY_AMT3**: Số tiền thanh toán trong tháng 7 năm 2005 (Đài tệ NT).
- **PAY_AMT4**: Số tiền thanh toán trong tháng 6 năm 2005 (Đài tệ NT).
- **PAY_AMT5**: Số tiền thanh toán trong tháng 5 năm 2005 (Đài tệ NT).
- **PAY_AMT6**: Số tiền thanh toán trong tháng 4 năm 2005 (Đài tệ NT).

Biến cuối cùng là biến cần dự đoán:

- **default.payment.next.month**: Chỉ thị liệu chủ thẻ có vỡ nợ hay không trong tháng tiếp theo (1 = vỡ nợ, 0 = không vỡ nợ).

Mục tiêu chính của tập dữ liệu là phân loại các khách hàng dự đoán sẽ vỡ nợ trong tháng tiếp theo, dựa trên cột **default.payment.next.month**, được thiết lập là "0"đối với khách hàng không vỡ nợ và "1"đối với khách hàng vỡ nợ. Do đó, đây là một bài toán phân loại nhị phân trên một tập dữ liệu không cân bằng, tỷ lệ cụ thể như hình 1.



Hình 1: Tỷ lệ khách hàng vỡ nợ

Để có cái nhìn tổng quan về dữ liệu, ta có bảng thống kê mô tả dưới đây:

Bảng 1: Thống kê mô tả các biến trong tập dữ liệu

Biến	Min	1st Qu.	Median	Mean	3rd Qu.	Max
ID	2	7466	15064	15054	22662	29999
LIMIT_BAL	10000	50000	140000	167121	240000	780000
SEX	1.0	1.0	2.0	1.608	2.0	2.0
EDUCATION	0.0	1.0	2.0	1.852	2.0	6.0
MARRIAGE	0.0	1.0	2.0	1.552	2.0	3.0
AGE	21	28	34	35.38	41	79
PAY_1	-2.0	-1.0	0.0	-0.0098	0.0	8.0
PAY_2	-2.0	-1.0	0.0	-0.1179	0.0	7.0
PAY_3	-2.0	-1.0	0.0	-0.1532	0.0	8.0
PAY_4	-2.0	-1.0	0.0	-0.2098	0.0	8.0
PAY_5	-2.0	-1.0	0.0	-0.2645	0.0	7.0
PAY_6	-2.0	-1.0	0.0	-0.2861	0.0	8.0
BILL_AMT1	-15308	3757	22468	51901	68028	746814
BILL_AMT2	-69777	3078	21197	49831	65123	743970
BILL_AMT3	-157264	2844	20205	47717	61111	1664089
BILL_AMT4	-170000	2464	19189	43872	55774	706864
BILL_AMT5	-81334	1814	18336	40871	50680	587067
BILL_AMT6	-339603	1272	17175	39422	49699	527566
PAY_AMT1	0	1000	2148	5808	5085	873552
PAY_AMT2	0	898.8	2031.5	6038.4	5000.0	1684259
PAY_AMT3	0	392	1850	5214	4646	889043
PAY_AMT4	0	313	1500	4854	4168	621000
PAY_AMT5	0	275	1500	4868	4145	388071
PAY_AMT6	0	198	1500	5262	4080	443001
default	0.0	0.0	0.0	0.2231	0.0	1.0

3.2 Xử lý dữ liệu

Dữ liệu này không chứa giá trị bị thiếu. Tuy nhiên, vẫn tồn tại một số vấn đề cần xử lý trước khi tiến hành phân tích.

Về thuộc tính, biến **EDUCATION** xuất hiện thêm ba loại giá trị không được mô tả trong tập dữ liệu gốc (0, 5 và 6). Tương tự, biến **MARRIAGE** cũng có giá trị 0 không phù hợp với các nhóm đã định nghĩa. Để khắc phục, các giá trị 0, 5 và 6 của **EDUCATION** được gộp chung vào nhóm "khác"(loại 4), và giá trị 0 của **MARRIAGE** được gộp vào nhóm "khác"(loại 3).

Từ bảng 1, ta thấy số dư hóa đơn (**BILL_AMTn**) âm phản ánh việc khách hàng đã thanh toán trước các khoản nợ, do đó, cần thực hiện biến đổi các giá trị đó về 0. Tương tự, các biến **PAY_n** thể hiện số tháng chậm thanh toán, trong đó giá trị -1 được sử dụng để chỉ những khách hàng thanh toán đúng hạn. Tuy nhiên, vẫn còn tồn tại các giá trị -2 và 0, dẫn đến sự không nhất quán trong cách diễn giải dữ liệu. Do đó, cần điều chỉnh lại nhãn, quy định giá trị 0 đại diện cho việc thanh toán đúng hạn nhằm đảm bảo tính nhất quán và chính xác cho quá trình phân tích.

Đối với các giá trị ngoại lai, ta cần làm sạch dữ liệu bằng cách loại bỏ chúng. Cụ thể, bất kỳ giá trị nào nhỏ hơn $Q1 - 1.5 \times IQR$ hoặc lớn hơn $Q3 + 1.5 \times IQR$ sẽ được coi là ngoại lai và sẽ bị

xoá bỏ.

Ngoài ra, trong quá trình kiểm tra dữ liệu, ta cũng phát hiện ra một số trường hợp gán nhãn **default.payment.next.month** không chính xác: có khách hàng luôn thanh toán đúng hạn nhưng vẫn bị gán nhãn vỡ nợ, và ngược lại. Do đó, cần tiến hành hiệu chỉnh lại nhãn mục tiêu dựa trên thông tin tình trạng thanh toán (từ **PAY_0** đến **PAY_6**) nhằm đảm bảo độ chính xác cho mô hình phân tích sau này.

Để thuận tiện hơn cho việc xử lý và trình bày trong quá trình nghiên cứu, biến **PAY_0** được đổi tên thành **PAY_1** nhằm đồng bộ hóa với các biến thanh toán khác, phù hợp hơn với thông tin mô tả, và biến **default.payment.next.month** được gọi tắt là **default**. Cột **ID** được loại bỏ bởi không cần thiết trong mô hình nghiên cứu.

Đồng thời, ta cần đưa các biến về đúng dạng. Cụ thể, các biến **SEX, EDUCATION, MARRIAGE, PAY_1, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6** được đưa về dạng `factor`.

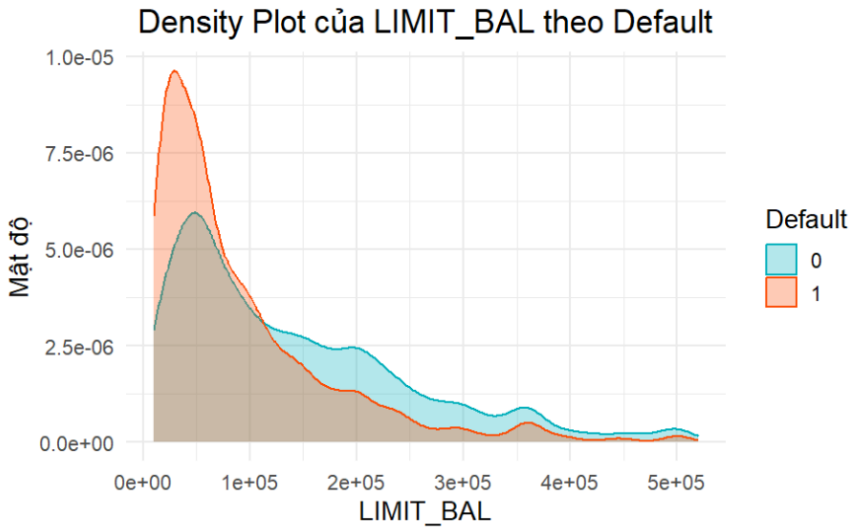
Sau khi hoàn thành các bước xử lý dữ liệu, kích thước dữ liệu còn lại là 9762 dòng $\times$ 24 cột.

3.3 Trực quan hoá dữ liệu

Trước khi xây dựng mô hình dự đoán, việc trực quan hóa dữ liệu là bước cần thiết nhằm hiểu rõ đặc điểm phân phối của các biến cũng như mối liên hệ giữa các yếu tố đầu vào và biến mục tiêu.

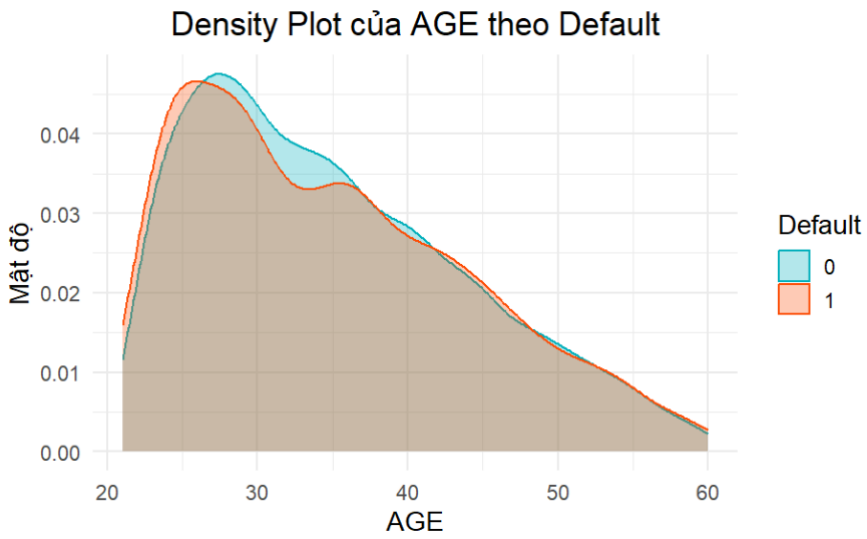
Ta xem xét mối quan hệ của một số biến và khả năng trả nợ.

Một trong những biến quan trọng trong bộ dữ liệu là **LIMIT_BAL** – hạn mức tín dụng được cấp cho khách hàng. Biến này thể hiện số tiền tối đa mà khách hàng có thể chi tiêu thông qua thẻ tín dụng.



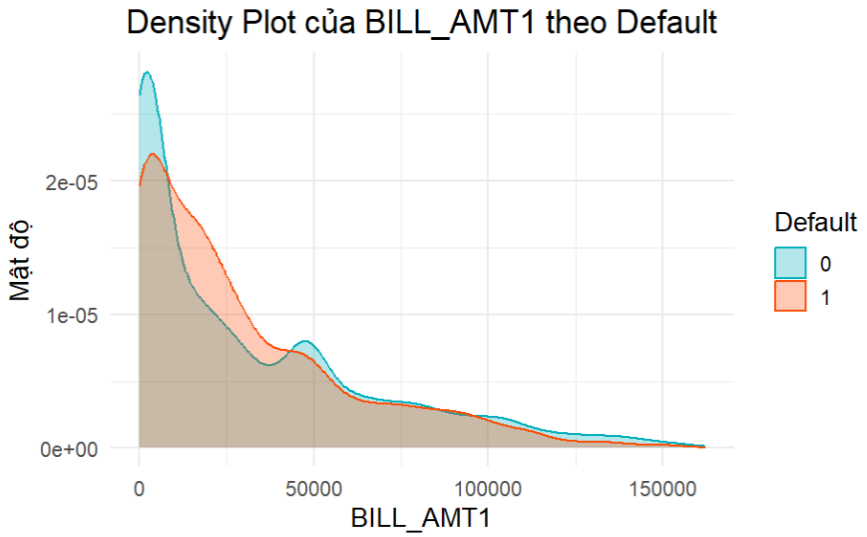
Hình 2: Biểu đồ mật độ của hạn mức tín dụng theo trạng thái vỡ nợ

Ta có thể thấy rằng nhóm khách hàng vỡ nợ (default = 1) có xu hướng tập trung ở mức hạn mức thấp hơn so với nhóm không vỡ nợ (default = 0). Điều này gợi ý rằng những khách hàng có hạn mức tín dụng thấp có khả năng vỡ nợ cao hơn, từ đó cho thấy **LIMIT_BAL** là một biến có tiềm năng phân biệt giữa hai nhóm khách hàng.



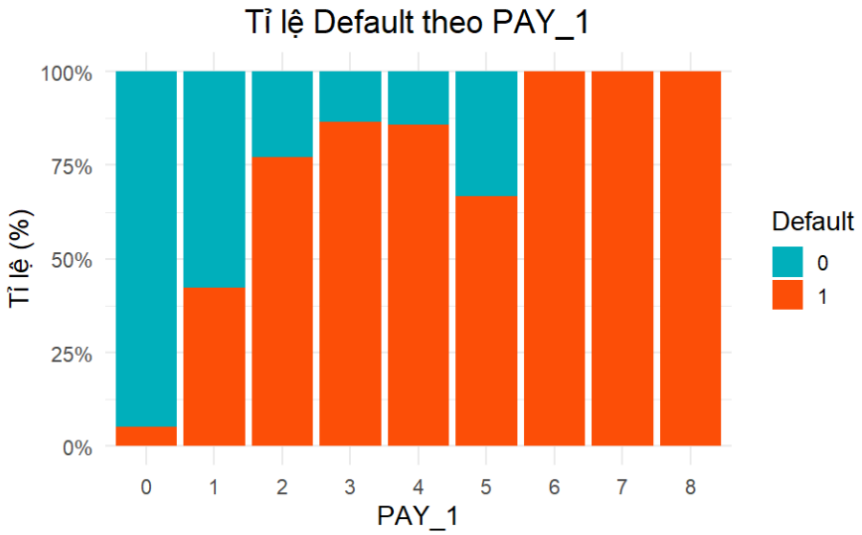
Hình 3: Biểu đồ mật độ của độ tuổi theo trạng thái vỡ nợ

Tuy nhiên, biểu đồ mật độ cho thấy phân bố tuổi của hai nhóm khách hàng (vỡ nợ và không vỡ nợ) có hình dạng tương đối giống nhau, tập trung chủ yếu ở độ tuổi từ 25 đến 40. Như vậy, **AGE** có khả năng không mạnh trong việc phân loại, đánh giá nguy cơ vỡ nợ.



Hình 4: Biểu đồ mật độ của số tiền sao kê trong tháng 9 năm 2005 theo trạng thái vỡ nợ

Biểu đồ cho thấy rằng đa số khách hàng, bất kể có vỡ nợ hay không, đều có số tiền hóa đơn tháng 9 năm 2005 ở mức thấp.



Hình 5: Tỷ lệ Vỡ nợ theo Tình trạng Thanh toán tháng 9 năm 2005

Có thể thấy, khách hàng thanh toán đúng hạn có nguy cơ vỡ nợ rất thấp. Khi thời gian chậm thanh toán tăng lên, nguy cơ vỡ nợ cũng tăng lên đáng kể. Đặc biệt, những khách hàng đã chậm thanh toán từ một tháng trở lên có nguy cơ vỡ nợ cao hơn nhiều so với những khách hàng thanh toán đúng hạn.

4 Kết quả

Bài nghiên cứu chia bộ dữ liệu thành hai phần. Một phần là tập huấn luyện, chiếm 70% dữ liệu, sẽ được sử dụng để huấn luyện mô hình, và phần còn lại là tập kiểm tra, chiếm 30% dữ liệu còn lại, sẽ được dùng để đánh giá mô hình.

Ta có kết quả Information Value (IV) của từng biến. IV dùng để đo độ tương quan giữa biến độc lập và biến dự đoán (ở dạng binary).

Bảng 2: Information Value (IV) of Variables

No.	Variable	IV	No.	Variable	IV	No.	Variable	IV
6	PAY_1	2.4763	16	BILL_AMT5	0.1210	22	PAY_AMT5	0.0624
7	PAY_2	1.6276	15	BILL_AMT4	0.1082	23	PAY_AMT6	0.0430
8	PAY_3	1.2773	13	BILL_AMT2	0.1029	5	AGE	0.0178
9	PAY_4	1.1657	14	BILL_AMT3	0.0969	2	SEX	0.0092
10	PAY_5	1.0643	19	PAY_AMT2	0.0931	4	MARRIAGE	0.0014
11	PAY_6	0.9854	21	PAY_AMT4	0.0909			
1	LIMIT_BAL	0.2962	20	PAY_AMT3	0.0899			
18	PAY_AMT1	0.1503	12	BILL_AMT1	0.0857			
17	BILL_AMT6	0.1256	3	EDUCATION	0.0745			

Các biến có IV nhỏ hơn 0.02 không có tác dụng trong việc phân loại hồ sơ vỡ nợ. Do đó, ba biến **AGE**, **SEX**, và **MARRIAGE** sẽ bị loại khỏi mô hình. Các biến còn lại đều có tác dụng hỗ

trợ phân loại hồ sơ, với mức độ khác nhau.

Trong đó, nhóm biến có sức mạnh phân loại cao nhất là các biến **PAY_n**, đặc biệt **PAY_1** có IV cao nhất (2.4763). Điều này hợp lý vì nó phản ánh tình trạng thanh toán ở thời điểm gần với thời điểm dự đoán nhất.

Các biến có sức mạnh phân loại trung bình bao gồm **LIMIT_BAL** và **PAY_AMT1**. Một số biến khác như **BILL_AMT6**, **BILL_AMT5**, **BILL_AMT4**, **BILL_AMT2** cũng có vai trò hỗ trợ phân loại ở mức trung bình-thấp.

Cuối cùng, nhóm biến còn lại như **BILL_AMT3**, **PAY_AMT2**, **PAY_AMT4**, **PAY_AMT3**, **BILL_AMT1**, **EDUCATION**, **PAY_AMT5**, và **PAY_AMT6** có sức mạnh phân loại yếu nhưng vẫn được giữ lại do giá trị hỗ trợ phân biệt hồ sơ vỡ nợ.

4.1 Mô hình hồi quy Logistic

4.1.1 Mô hình hồi quy Logistic với WOE

Phương trình hồi quy logistic trong credit scorecard sẽ không hồi quy trực tiếp trên các biến gốc mà thay vào đó giá trị WOE ở từng biến sẽ được sử dụng thay thế để làm đầu vào. Ta sẽ tính toán các biến WOE bằng cách map mỗi khoảng bin tương ứng với giá trị WOE của nó.

Xây dựng phương trình hồi quy logistic các biến đầu vào WOE, sau đó sử dụng stepwise để chọn ra mô hình tốt nhất với các biến được lựa chọn.

Kết quả ước lượng cho mô hình logistic với WOE được thể hiện ở bảng 3

Biến	**Estimate**	**p-value**
Intercept	-1.35071	$< 2 \times 10^{-16}$ ***
LIMIT_BAL_woe	0.42588	1.37×10^{-6} ***
EDUCATION_woe	-1.37709	0.000255 ***
PAY_1_woe	0.81611	$< 2 \times 10^{-16}$ ***
PAY_2_woe	0.07628	0.103003
PAY_3_woe	0.32383	4.22×10^{-10} ***
PAY_4_woe	0.24696	9.58×10^{-5} ***
PAY_5_woe	0.11120	0.080973 .
PAY_6_woe	0.60433	$< 2 \times 10^{-16}$ ***
PAY_AMT1_woe	0.23394	0.028031 *
PAY_AMT3_woe	0.45951	0.005562 **
PAY_AMT6_woe	0.30268	0.106360

Bảng 3: Kết quả hồi quy logistic với các biến WOE.
Ghi chú: ******* p<0.001; ****** p<0.01; ***** p<0.05; **.** p<0.1

Mô hình dự đoán chính xác 88.56% kết quả dự báo trong mẫu, cụ thể được thể hiện qua ma trận nhầm lẫn:

		Actual	
		0	**1**
Predicted	**0**	5208	515
	1	264	825

Bảng 4: Ma trận nhầm lẫn trong mẫu của mô hình logistic với WOE

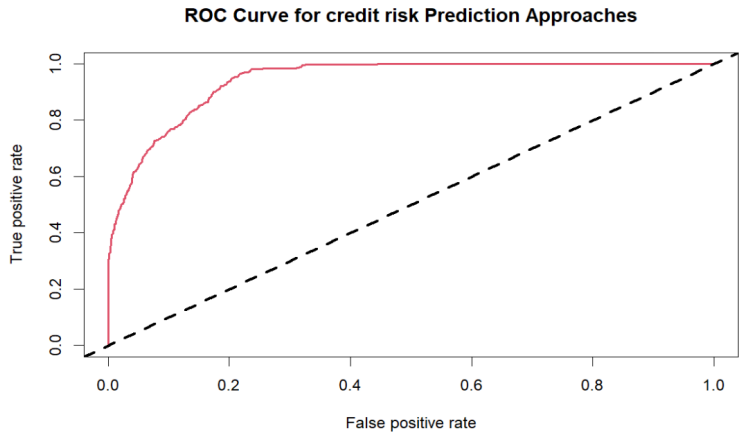
Sử dụng mô hình đã ước lượng để dự báo ngoài mẫu, thu được độ chính xác là 81.92%.

		Actual	
		0	**1**
Predicted	**0**	2277	231
	1	96	346

Bảng 5: Ma trận nhầm lẫn ngoài mẫu của mô hình logistic với WOE

Từ ma trận nhầm lẫn, ta cũng có được các giá trị sensitivity, specificity lần lượt là: 0.5996534; 0.9595449.

Đồng thời, thể hiện khả năng phân loại của mô hình qua đường cong ROC. Thu được các chỉ số AUC= 0.9420992, gini= 0.8841984



Hình 6: Đường cong ROC cho mô hình Logistic với WOE

4.1.2 Mô hình hồi quy Logistic với biến gốc

Đối với biến gốc, trong bước đầu tiên, em tiến hành xây dựng hai mô hình hồi quy logistic cơ bản: Mô hình đầy đủ bao gồm tất cả các biến giải thích và mô hình rỗng chỉ bao gồm hằng số mà không có biến giải thích nào.

Áp dụng phương pháp lựa chọn biến Stepwise để tìm kiếm mô hình tối ưu. Quá trình lựa chọn được thực hiện theo hai hướng: từ mô hình rỗng đến mô hình đầy đủ và kết hợp cả thêm và bớt biến. Trong quá trình lựa chọn mô hình, hai tiêu chí AIC, BIC được sử dụng để đánh giá. Kết quả thu được hai mô hình logistic tốt nhất: một mô hình tối ưu theo AIC **(1)** và một mô hình tối ưu theo BIC **(2)**.

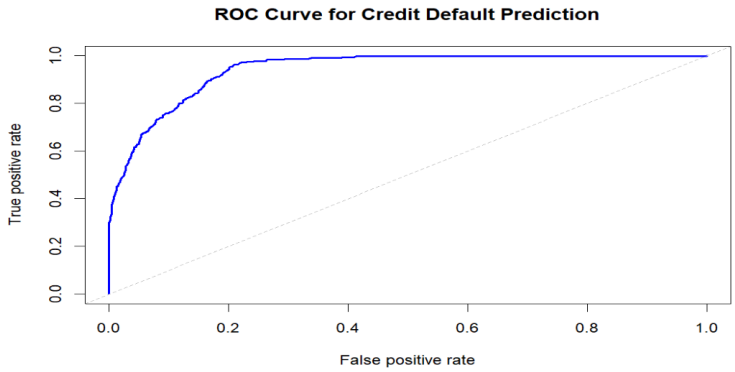
Mô hình **(1)** có AIC= 3475.36 và gồm 11 biến: PAY_1, PAY_6, PAY_3, PAY_4, EDUCATION, LIMIT_BAL, PAY_AMT4, PAY_AMT3, PAY_2, BILL_AMT1, BILL_AMT6. Mô hình có khả năng dự báo 88.62% trong mẫu.

Kết quả dự báo ngoài mẫu của mô hình được thể hiện qua ma trận nhầm lẫn với chỉ số sensitivity= 0.6169844, specificity= 0.9561736.

		Actual	
		0	**1**
Predicted	**0**	2269	221
	1	104	356

Bảng 6: Ma trận nhầm lẫn ngoài mẫu của mô hình **(1)**

Hệ số AUC của mô hình **(1)** là 0.9406984. Từ đó, tính được gini= 0.8813968.



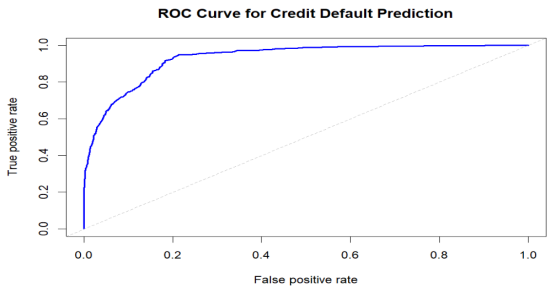
Hình 7: Đường cong ROC của mô hình **(1)**

Mô hình **(2)** có BIC= 3706.297 và gồm 6 biến: PAY_1, PAY_6, PAY_3, PAY_AMT3, PAY_AMT4, LIMIT_BAL. Khả năng dự báo trong mẫu của mô hình là 88.86%.

Các chỉ tiêu đánh giá ngoài mẫu của mô hình lần lượt là: sensitivity= 0.6412478, specificity= 0.9502739, AUC= 0.930868, gini= 0.861736. Cụ thể, kết quả bảng ma trận nhầm lẫn, đường cong ROC được thể hiện qua bảng 7, và hình 8 dưới đây:

		Actual	
		0	**1**
Predicted	**0**	2255	207
	1	118	370

Bảng 7: Ma trận nhầm lẫn ngoài mẫu của mô hình **(2)**



Hình 8: Đường cong ROC của mô hình **(2)**

4.2 Mô hình KNN

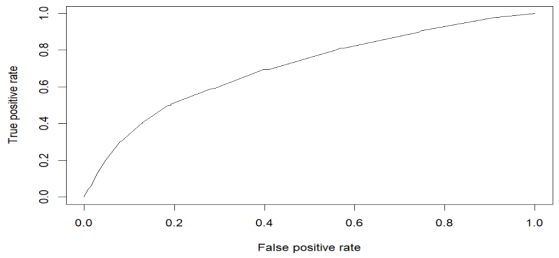
Trong nghiên cứu này, em đã tiến hành biến đổi biến mục tiêu **default** thành kiểu dữ liệu nhân tố `factor` đối với cả tập huấn luyện và tập kiểm tra. Mô hình KNN sau đó được xây dựng bằng cách sử dụng tập huấn luyện với biến mục tiêu là **default** và toàn bộ các biến còn lại làm biến giải thích. Quy trình đánh giá mô hình áp dụng phương pháp kiểm định chéo 10 phần (10-fold cross-validation) nhằm tự động tìm kiếm giá trị tối ưu của tham số k trong phạm vi 9 giá trị khác nhau.

Kết quả huấn luyện cho thấy giá trị $k = 19$ mang lại sai số phân loại thấp nhất. Với $k = 19$, mô hình dự đoán đúng khoảng 80.2% các trường hợp trong mẫu.

Kết quả dự báo ngoài mẫu của mô hình KNN cụ thể như sau:

		Actual	
		0	**1**
Predicted	**0**	2309	508
	1	64	69

Bảng 8: Ma trận nhầm lẫn ngoài mẫu của mô hình KNN



Hình 9: Đường cong ROC của mô hình KNN

Từ bảng 8 và hình 9, ta có thể thấy mô hình KNN kém hiệu quả trong việc phân loại **default/ non default**. Cụ thể, các chỉ số về khả năng dự báo của mô hình trong tập kiểm tra là: sensitivity = 0.1195841, specificity = 0.9730299, AUC = 0.7047394, gini = 0.4094788.

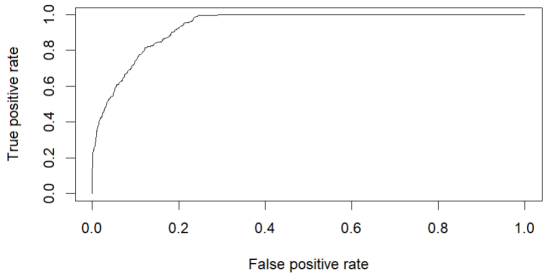
4.3 Mô hình Random Forest

Trong quá trình thực nghiệm, mô hình Random Forest được huấn luyện trên 6812 mẫu với 20 biến dự báo và biến mục tiêu dạng nhị phân (0, 1). Dữ liệu được đánh giá bằng phương pháp 10-fold cross-validation nhằm đảm bảo tính ổn định của mô hình. Quá trình tinh chỉnh tham số với các giá trị `mtry` khác nhau cho thấy rằng khi `mtry` = 2, độ chính xác (Accuracy) đạt 86,52% với chỉ số Kappa là 0,4438. Khi tăng `mtry` lên 32, mô hình đạt độ chính xác cao nhất 88,33% và Kappa 0,6211, thể hiện khả năng phân loại được cải thiện rõ rệt. Tuy nhiên, khi tiếp tục tăng `mtry` lên 62, độ chính xác giảm nhẹ còn 87,98% và Kappa cũng giảm xuống 0,6102. Dựa trên tiêu chí chọn mô hình có độ chính xác cao nhất, Random Forest với `mtry` = 32 được lựa chọn làm mô hình cuối cùng. Kết quả này cho thấy Random Forest không chỉ đạt độ chính xác cao mà còn có khả năng phân biệt hai lớp mục tiêu khá tốt, phù hợp để áp dụng vào bài toán phân loại đang nghiên cứu.

Đối với kết quả dự báo ngoài mẫu, mô hình đạt được độ chính xác là 87.49%. Các chỉ số đánh giá bao gồm: sensitivity = 0.6256499, specificity = 0.9355247, AUC = 0.9369999, gini = 0.8739999.

		Actual	
		0	**1**
Predicted	**0**	2220	216
	1	153	361

Bảng 9: Ma trận nhầm lẫn ngoài mẫu của mô hình Random Forest



Hình 10: Đường cong ROC của mô hình Random Forest

4.4 Lựa chọn mô hình

Từ kết quả các mô hình trên, ta có bảng tổng kết hiệu suất của các mô hình khi dự đoán ngoài mẫu:

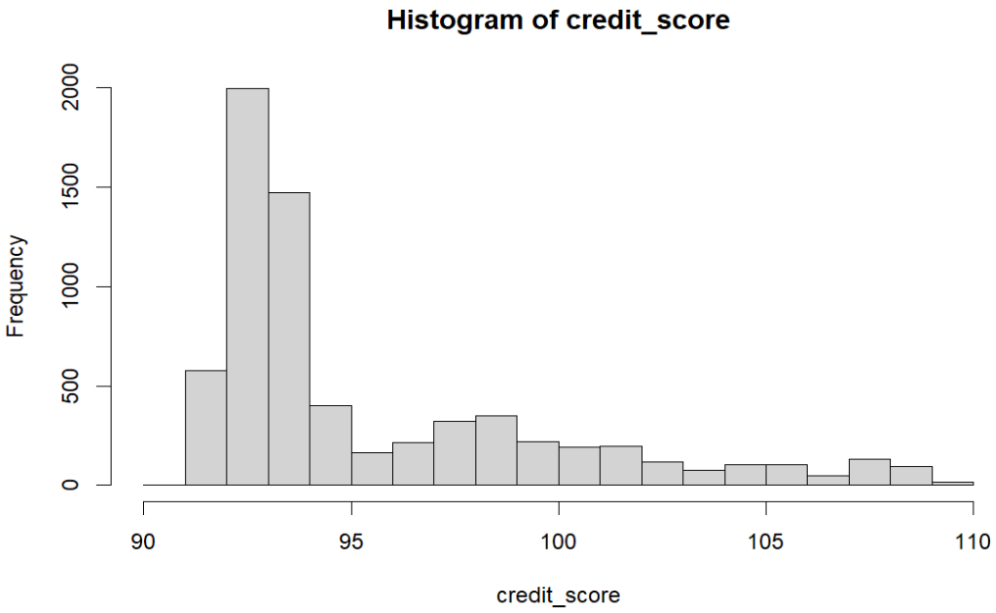
Mô hình	**Ma trận nhầm lẫn**	**AUC**	**Gini**
Logistic với WOE	Accuracy: 81.92% Sensitivity: 59.97% Specificity: 95.95%	94.21%	88.42%
Logistic với biến gốc **(1)**	Accuracy: 88.95% Sensitivity: 61.70% Specificity: 95.62%	94.07%	88.14%
Logistic với biến gốc **(2)**	Accuracy: 88.98% Sensitivity: 64.12% Specificity: 95.03%	93.09%	86.17%
KNN	Accuracy: 80.61% Sensitivity: 11.96% Specificity: 97.30%	70.47%	40.95%
Random Forest	Accuracy: 87.49% Sensitivity: 62.56% Specificity: 93.55%	93.70%	87.40%

Bảng 10: So sánh hiệu suất các mô hình

Với bài toán phân loại và dự báo khả năng vỡ nợ, ta chú trọng vào khả năng dự báo đúng lượng khách hàng vỡ nợ, hay là chỉ số sensitivity của mô hình. Do đó, ngoài việc lựa chọn mô hình tốt, chúng ta cũng cần lựa chọn những mô hình có sensitivity tốt.

Dựa vào bảng 10 ở trên thì mô hình Logistic với WOE cho AUC, gini tốt nhất (94.21%, 88.42%). Đồng thời, ta cũng thấy mô hình Logistic với WOE có sensitivity gần như không kém mô hình Logistic với biến gốc lựa chọn theo tiêu chí BIC quá nhiều, mà có rất nhiều ưu điểm như mô hình rất dễ giải thích và đơn giản để có thể huấn luyện, kiểm thử, tinh chỉnh trong thời gian ngắn, do đó chọn mô hình Logistic với WOE làm mô hình cuối cùng để áp dụng vào bài toán chấm điểm.

Áp dụng công thức tính Scorecard với *offset*= 100, Factor= 2, ta có được kết quả chấm điểm tín dụng cho từng khách hàng.



Hình 11: Đồ thị tần suất của Scorecard

Đồ thị hình 11 cho thấy phân phối lệch phải. Phần lớn điểm tín dụng của khách hàng ở mức 92-95 điểm. Sau mốc 95, tần suất giảm dần, cho thấy số lượng cá nhân có điểm tín dụng cao hơn trở nên ít dần. Tuỳ vào yêu cầu của từng ngân hàng, doanh nghiệp mà xác định ngưỡng điểm tín dụng cho khách hàng vay.

5 Kết luận

Trong bối cảnh nhu cầu kiểm soát rủi ro tín dụng ngày càng cao, nghiên cứu này đã xây dựng và đánh giá hiệu quả các mô hình phân loại nhằm dự đoán khả năng vỡ nợ của khách hàng dựa trên dữ liệu đăng ký thẻ tín dụng. Quá trình thực hiện bao gồm các bước tiền xử lý dữ liệu, biến đổi các biến đầu vào bằng phương pháp WOE (Weight of Evidence), và so sánh hiệu quả giữa các mô hình như hồi quy Logistic, KNN, và Random Forest.

Kết quả cho thấy so với các mô hình học máy, mô hình hồi quy Logistic cho hiệu suất tốt nhất, với độ chính xác cao và khả năng phân tách khách hàng tiềm năng vỡ nợ rõ ràng hơn. Mô hình này không chỉ đơn giản trong việc triển khai mà còn dễ dàng diễn giải, giúp các ngân hàng hiểu rõ mức độ ảnh hưởng của từng biến đến khả năng vỡ nợ.

Trên cơ sở mô hình tốt nhất được lựa chọn, nghiên cứu đã xây dựng hệ thống tính điểm tín dụng (credit scoring) cho khách hàng. Thang điểm này giúp lượng hóa mức độ rủi ro của từng cá nhân, từ đó hỗ trợ các tổ chức tài chính đưa ra quyết định cấp tín dụng một cách chính xác và hiệu quả hơn.

Từ kết quả đạt được, nghiên cứu gợi mở một số hướng phát triển tiếp theo như: thử nghiệm các mô hình nâng cao khác như Gradient Boosting để cải thiện hiệu quả phân loại. Việc kết hợp mô hình Logistic với các mô hình học máy cũng là một hướng đi hứa hẹn, góp phần nâng cao độ chính xác và giảm thiểu rủi ro trong hoạt động cấp tín dụng của ngân hàng.

Tài liệu

[1] Quách Hải Yến, Vũ Minh Tuấn, và Lê Phương Hà. *Mô hình chấm điểm tín dụng trong hoạt động quản trị rủi ro của ngân hàng thương mại tại Việt Nam: Ứng dụng mô hình Logistic*. Ngân hàng TMCP Quân đội; Tập đoàn Công nghiệp - Viễn thông Quân đội; Viện Chiến lược Chuyển đổi số, 2023. Truy cập từ: https://doi.org/10.38203/jiem.vi.102023.1089

[2] Phạm Đình Khánh. *Xây dựng model scorecard*, trong sách trực tuyến *Machine Learning cơ bản với ứng dụng trong Tài chính - Ngân hàng*. Truy cập tại: https://phamdinhkhanh.github.io/deepai-book/ch_ml/creditScorecard.html#xay-dung-model-scorecard