ĐẠI HỌC KINH TẾ QUỐC DÂN

KHOA TOÁN KINH TẾ

BÀI TẬP CÁ NHÂN

QUẢN TRỊ RỦI RO ĐỊNH LƯỢNG 2

Xây dựng và so sánh các mô hình dự báo vỡ nợ để phát triển hệ thống chấm điểm tín dụng

Lớp học phần: TOTC1121(224)_01

Giảng viên: Đinh Thị Hồng Thêu

Sinh viên: Nghiêm Gia Phương

MSV: 11225223

Hà Nội, 2025

Mục lục

TÓM TẮT 1

1 Giới thiệu **2**
1.1 Lý do nghiên cứu 2
1.2 Mục tiêu nghiên cứu 2
1.3 Tầm quan trọng của nghiên cứu 2
1.4 Phạm vi nghiên cứu 2

2 Cơ sở lý thuyết **4**
2.1 Tổng quan về lý thuyết 4
2.2 Kĩ thuật biến đổi và lựa chọn biến đầu vào 4
2.3 Các mô hình xếp hạng tín dụng 5
2.3.1 Mô hình Logit 5
2.3.2 Cây quyết định (Decision Tree- DT) và rừng ngẫu nhiên (Random Forest . 5
2.3.3 K Láng giềng gần nhất (K-Nearest Neighbor- KNN) 6

3 Dữ liệu **8**
3.1 Nguồn dữ liệu 8
3.2 Xử lí dữ liệu 9
3.2.1 Chuẩn hóa tên cột 9
3.2.2 Xử lý giá trị không hợp lệ trong các biến 9
3.2.3 Điều chỉnh nhãn "vỡ nợ" 9
3.3 Thống kê mô tả 9
3.4 Kiểm tra dữ liệu và xử lý giá trị bất thường bằng phương pháp IQR 11
3.5 Chọn biến bằng cách tính chỉ số IV 11
3.6 Chia tập dữ liệu 12

4 Kết quả các mô hình **13**
4.1 Mô hình Logit 13
4.1.1 Mô hình Logit với biến gốc 13
4.1.2 Mô hình Logit với WOE 14
4.1.3 So sánh mô hình Logit với biến gốc và mô hình Logit với WOE 15
4.2 Rừng ngẫu nhiên (Random Forest) 16
4.3 K Láng giềng gần nhất (K-Nearest Neighbors - KNN) 17
4.4 So sánh các mô hình Logit, Random Forest và KNN 17
4.5 Xây dựng hệ thống tính điểm tín dụng (Credit Scoring) từ mô hình Logit với WOE 18

5 Kết luận và thảo luận **19**

Tài liệu tham khảo **20**

Tóm tắt

Bài nghiên cứu này tập trung vào xây dựng và so sánh các mô hình dự đoán khả năng vỡ nợ của khách hàng vay tiêu dùng dựa trên bộ dữ liệu Default of Credit Card Clients. Các mô hình được triển khai bao gồm mô hình Logit với biến gốc, mô hình Logit với biến đã mã hóa theo Weight of Evidence (WOE), Random Forest và K-Nearest Neighbors (KNN). Dữ liệu được chia thành tập huấn luyện và kiểm tra để đánh giá hiệu suất mô hình thông qua các chỉ số như Sensitivity, Specificity, AUC (diện tích dưới đường cong ROC) và Gini. Kết quả cho thấy mô hình Logit sử dụng biến WOE đạt Sensitivity cao nhất trong khi vẫn duy trì Specificity, AUC và Gini ở mức tốt, cho thấy sự phù hợp trong bài toán dự đoán rủi ro tín dụng. Từ mô hình này, hệ thống tính điểm tín dụng (scorecard) được xây dựng nhằm chuyển đổi xác suất vỡ nợ thành điểm tín dụng đơn giản, hỗ trợ ra quyết định trong thực tiễn.

Chương 1

Giới thiệu

1.1 Lý do nghiên cứu

Trong bối cảnh thị trường tín dụng tiêu dùng tại Việt Nam và trên thế giới ngày càng mở rộng, rủi ro tín dụng đã trở thành một trong những thách thức lớn đối với các tổ chức tài chính. Việc khách hàng không hoàn trả khoản vay đúng hạn không chỉ ảnh hưởng đến hiệu quả hoạt động kinh doanh mà còn có thể tạo ra hệ lụy lan tỏa đến toàn hệ thống tài chính. Trong khi đó, các tổ chức tín dụng thường phải ra quyết định dựa trên thông tin không hoàn hảo, khiến cho việc đánh giá mức độ rủi ro của từng cá nhân trở nên phức tạp. Do đó, nhu cầu xây dựng các mô hình định lượng có khả năng dự báo khả năng vỡ nợ ngày càng cấp thiết.

1.2 Mục tiêu nghiên cứu

Mục tiêu chính của nghiên cứu là xây dựng, đánh giá và so sánh các mô hình dự đoán khả năng vỡ nợ của khách hàng vay tiêu dùng, từ đó lựa chọn mô hình tối ưu để phát triển hệ thống chấm điểm tín dụng (credit scorecard). Cụ thể, nghiên cứu tiến hành huấn luyện các mô hình Logit với biến gốc, Logit với biến có mã hóa WOE, Random Forest và K-Nearest Neighbors trên tập dữ liệu khách hàng thẻ tín dụng, từ đó so sánh hiệu suất dự báo của các mô hình dựa trên tập kiểm tra với các chỉ số đo lường phổ biến như Sensitivity, Specificity, AUC và Gini. Mô hình tốt nhất sẽ được sử dụng để xây dựng hệ thống điểm tín dụng nhằm hỗ trợ quyết định cho vay của các tổ chức tài chính.

1.3 Tầm quan trọng của nghiên cứu

Việc áp dụng các mô hình định lượng trong đánh giá rủi ro tín dụng không chỉ giúp tăng tính khách quan và hiệu quả của các quyết định cho vay, mà còn góp phần hạn chế nợ xấu và cải thiện chất lượng danh mục tín dụng của các tổ chức tài chính. Trong thực tiễn, hệ thống điểm tín dụng đóng vai trò như một công cụ ra quyết định nhanh chóng, giúp xác định mức độ rủi ro của khách hàng tiềm năng một cách trực quan và nhất quán. Ngoài ra, việc nghiên cứu so sánh hiệu suất các mô hình còn cung cấp cơ sở khoa học để lựa chọn kỹ thuật phù hợp với dữ liệu thực tế tại Việt Nam.

1.4 Phạm vi nghiên cứu

Nghiên cứu sử dụng bộ dữ liệu "Default of Credit Card Clients", bao gồm 30.000 quan sát về khách hàng sử dụng thẻ tín dụng của một tổ chức tài chính tại Đài Loan trong giai đoạn từ tháng 4

đến tháng 9 năm 2005, được thu thập trên nền tảng Kaggle. Để thuận tiện cho xử lý và phân tích, mẫu được thu hẹp còn 10.000 quan sát được chọn ngẫu nhiên. Dữ liệu bao gồm các thông tin về mã định danh, nhân khẩu học, hạn mức tín dụng, lịch sử thanh toán, số dư hóa đơn hàng tháng, số tiền thanh toán hàng tháng và kết quả về tình trạng vỡ nợ trong tháng kế tiếp. Mặc dù dữ liệu có chứa thông tin theo tháng trong vòng sáu tháng, mỗi quan sát là độc lập và không hình thành chuỗi thời gian liên tục cho từng khách hàng; do đó, phạm vi nghiên cứu không bao gồm phân tích chuỗi thời gian theo nghĩa truyền thống hoặc cập nhật theo dòng dữ liệu thực tế. Các mô hình được huấn luyện trên tập huấn luyện và đánh giá trên tập kiểm tra nhằm đảm bảo tính khách quan trong đánh giá hiệu suất. Kết quả mô hình chưa được triển khai kiểm định trên dữ liệu của tổ chức tín dụng tại Việt Nam. Tuy nhiên, nghiên cứu này có thể làm cơ sở cho việc phát triển và điều chỉnh mô hình trong tương lai, khi dữ liệu thực tế tại Việt Nam được thu thập đầy đủ.

Chương 2

Cơ sở lý thuyết

2.1 Tổng quan về lý thuyết

Theo Điều 20 Luật các tổ chức tín dụng 2010 quy định: Cấp tín dụng là việc ngân hàng "thỏa thuận để tổ chức, cá nhân sử dụng một khoản tiền hoặc cam kết cho phép sử dụng một khoản tiền theo nguyên tắc có hoàn trả bằng nghiệp vụ cho vay, chiết khấu, cho thuê tài chính, bao thanh toán, bảo lãnh ngân hàng và các nghiệp vụ cấp tín dụng khác" (Quốc hội nước Cộng hòa Xã hội Chủ nghĩa Việt Nam, 2010, tr. 12)[1]. Dựa trên khái niệm này, tín dụng đối với khách hàng cá nhân có thể được hiểu là việc ngân hàng chuyển giao quyền sử dụng vốn cho cá nhân trong một khoảng thời gian nhất định, với yêu cầu hoàn trả cả gốc và lãi theo thỏa thuận.

Rủi ro tín dụng sẽ phát sinh khi các cá nhân được phê duyệt các khoản vay. Theo Ủy ban Basel về Giám sát Ngân hàng (BCBS, 2000) [2] rủi ro tín dụng (credit risk) được định nghĩa là người vay có khả năng không đáp ứng các nghĩa vụ của mình theo các điều khoản đã thỏa thuận. Theo khoản 24 Điều 2 Thông tư 41/2016/TTNHNN giải thích chi tiết hơn: "Rủi ro tín dụng là rủi ro do khách hàng không thực hiện hoặc không có khả năng thực hiện một phần hoặc toàn bộ nghĩa vụ trả nợ theo hợp đồng hoặc thỏa thuận với ngân hàng, chi nhánh ngân hàng nước ngoài" (Ngân hàng Nhà nước Việt Nam, 2016, tr. 02)[3].

Khi rủi ro tín dụng được lượng hóa, nó được thể hiện dưới dạng xác suất vỡ nợ (Probability of Default - PD). Xác suất vỡ nợ đóng vai trò then chốt trong các mô hình phân tích rủi ro tín dụng và trong công tác quản lý rủi ro của ngân hàng.

2.2 Kĩ thuật biến đổi và lựa chọn biến đầu vào

Trong phân tích dữ liệu tín dụng, việc biến đổi và lựa chọn biến số đầu vào là một bước quan trọng nhằm tăng hiệu quả của mô hình dự báo. Trong đó, WOE (Weight of Evidence) và IV (Information Value) là hai phương pháp phổ biến được sử dụng để đánh giá mức độ liên quan giữa một biến độc lập với khả năng xảy ra vỡ nợ của khách hàng.

WOE là kỹ thuật mã hóa biến đầu vào dựa trên phân phối của biến mục tiêu trong từng nhóm. WOE được tính theo công thức:

$$WOE_i = \ln\left(\frac{\text{Tỷ lệ khách hàng không vỡ nợ trong nhóm } i}{\text{Tỷ lệ khách hàng vỡ nợ trong nhóm } i}\right)$$

Giá trị WOE giúp biến đổi các biến thành dạng liên tục có tính đơn điệu và dễ dàng giải thích hơn trong mô hình hồi quy logistic.

IV là chỉ số dùng để đánh giá mức độ hữu ích của một biến trong việc phân biệt giữa hai nhóm (vỡ nợ và không vỡ nợ). Công thức tính IV là:

$$IV = \sum_{i=1}^{n} (P_{\text{good},i} - P_{\text{bad},i}) \cdot WOE_i$$

Trong đó $P_{\text{good},i}$ và $P_{\text{bad},i}$ lần lượt là tỷ lệ khách hàng không vỡ nợ và vỡ nợ trong nhóm thứ i. Ý nghĩa của IV thường được diễn giải như sau:

Giá trị IV	Mức độ phân biệt
< 0.02	Không có giá trị dự báo
$0.02 - 0.1$	Yếu
$0.1 - 0.3$	Trung bình
$0.3 - 0.5$	Mạnh
> 0.5	Rất mạnh

Bảng 2.1: Mức độ phân biệt dựa trên giá trị IV

Việc áp dụng WOE và IV giúp lựa chọn các biến đầu vào phù hợp và cải thiện độ ổn định của mô hình dự đoán xác suất vỡ nợ, đặc biệt trong mô hình hồi quy logistic.

2.3 Các mô hình xếp hạng tín dụng

2.3.1 Mô hình Logit

Mô hình Logit là một dạng hồi quy trong đó biến phụ thuộc Y là biến nhị phân, chỉ nhận giá trị 0 hoặc 1. Các biến độc lập trong mô hình có thể ở dạng nhị phân, rời rạc hoặc liên tục. Trong xếp hạng tín dụng, biến Y biểu thị trạng thái vỡ nợ của khách hàng: $Y = 1$ nếu khách hàng không thực hiện được nghĩa vụ thanh toán (vỡ nợ) và $Y = 0$ nếu khách hàng thực hiện đúng nghĩa vụ trả nợ (không vỡ nợ).

Các biến độc lập đại diện cho thông tin định tính và định lượng về khách hàng, như thu nhập, độ tuổi, giới tính, trình độ học vấn,... Sau khi thực hiện hồi quy, mô hình thu được:

$$\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Từ đó, xác suất khách hàng vỡ nợ được xác định bởi công thức:

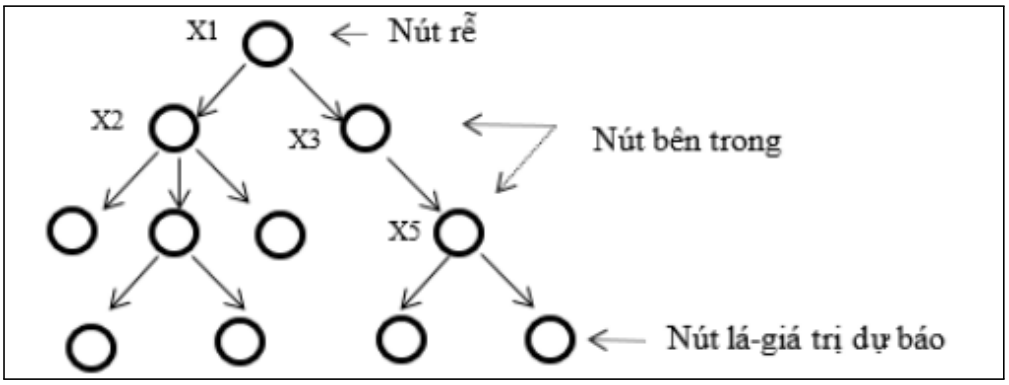
$$P = \frac{1}{1 + e^{-\hat{Y}}}$$

Giá trị P dao động trong khoảng từ 0 đến 1. Để phân loại khách hàng, P được so sánh với một ngưỡng xác định trước. Trong bài nghiên cứu này, ngưỡng phân loại chọn là 0.5: nếu $P \geq 0.5$, khách hàng được dự báo sẽ vỡ nợ (default); ngược lại nếu $P < 0.5$, khách hàng được dự báo sẽ không vỡ nợ.

2.3.2 Cây quyết định (Decision Tree- DT) và rừng ngẫu nhiên (Random Forest

Cây quyết định (Decision Tree - DT) là một mô hình phân lớp sử dụng các quy tắc quyết định dựa trên các thuộc tính đầu vào. Cấu trúc của mô hình gồm các nút: nút rễ (root node), các nút bên trong (internal nodes) và các nút lá (leaf nodes). Mỗi nút trong cây quyết định đại diện cho một thuộc tính, và các nhánh kết nối giữa các nút thể hiện các giá trị cụ thể của thuộc tính đó. Tại mỗi nút lá, mô hình đưa ra một giá trị dự đoán cho biến mục tiêu. Quá trình phân loại sẽ diễn ra theo

các bước phân chia các thuộc tính cho đến khi đạt đến các nút lá, từ đó xác định được kết quả dự báo. (Hình 2.1)



Hình 2.1: Cây quyết định (Decision Tree)

Nguồn: Abdou (2011) [4]

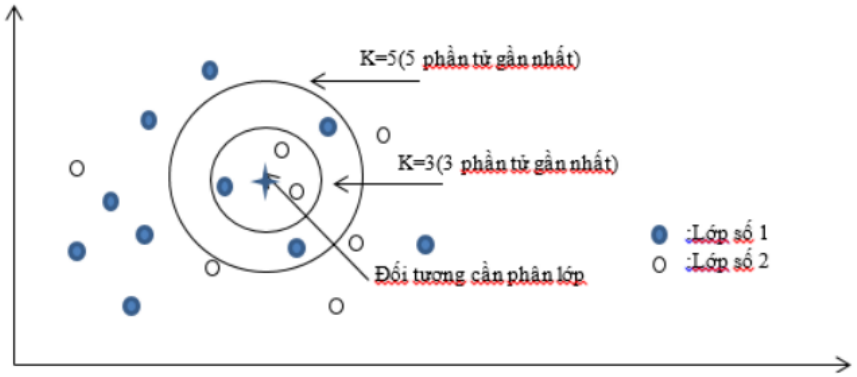
Mô hình cây quyết định xây dựng các quy tắc phân lớp từ các thuộc tính đầu vào của dữ liệu và tiếp tục cho đến khi đạt được giá trị mục tiêu. Những luật quyết định này từ các đường đi giữa các nút giúp xác định mức độ rủi ro của khách hàng trong bài toán xếp hạng tín dụng.

Tuy nhiên, cây quyết định thường gặp vấn đề về quá khớp (overfitting), đặc biệt khi dữ liệu huấn luyện có sự biến động lớn. Để khắc phục hạn chế này, mô hình Rừng ngẫu nhiên (Random Forest) được phát triển. Rừng ngẫu nhiên bao gồm một tập hợp các cây quyết định, mỗi cây được xây dựng từ các mẫu dữ liệu ngẫu nhiên và các tập con của các thuộc tính. Dự đoán cuối cùng được đưa ra dựa trên kết quả đa số của các cây trong rừng, giúp giảm thiểu vấn đề quá khớp và cải thiện độ chính xác dự đoán, đặc biệt khi làm việc với các tập dữ liệu phức tạp như dữ liệu tín dụng.

2.3.3 K Láng giềng gần nhất (K-Nearest Neighbor- KNN)

Phương pháp KNN (K-Nearest Neighbors) là một kỹ thuật học máy dùng để phân lớp các đối tượng dựa trên việc xác định K đối tượng gần nhất trong dữ liệu huấn luyện. Khi phân loại một đối tượng mới, lớp của đối tượng này được xác định từ K điểm gần nhất trong tập dữ liệu huấn luyện. Lớp dự đoán cho đối tượng sẽ là lớp mà có số lượng điểm trong K điểm gần nhất nhiều nhất.

Từ Hình 2.2 có thể thấy, nếu chọn $K = 3$, nghĩa là ta sẽ tìm ba điểm gần nhất với điểm cần phân loại. Nếu trong ba điểm này, một điểm thuộc lớp 1 và hai điểm thuộc lớp 2, thì đối tượng cần phân lớp sẽ được xếp vào lớp 2, vì lớp này có số lượng điểm nhiều hơn. Tương tự, nếu $K = 5$, khi tìm năm điểm gần nhất, trong đó ba điểm thuộc lớp 1 và hai điểm thuộc lớp 2, đối tượng sẽ được phân vào lớp 1.



Hình 2.2: K Láng giềng gần nhất (K-Nearest neighbor - KNN)

Nguồn: Marinakis và cộng sự (2008) [5]

Với dữ liệu thực tế, mỗi đối tượng có thể có nhiều thuộc tính, tương ứng với không gian đa chiều. Do đó, việc tính khoảng cách giữa các điểm không chỉ đơn giản là trong không gian 2 chiều, mà cần tính khoảng cách giữa các điểm trong không gian nhiều chiều. Khoảng cách này thường được tính bằng công thức Euclidean, như sau:

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

Trong đó x và y là hai điểm cần tính khoảng cách, và n là số chiều của không gian (số thuộc tính của biến mục tiêu).

Chương 3

Dữ liệu

3.1 Nguồn dữ liệu

Bộ dữ liệu được sử dụng là "Default of Credit Card Clients Dataset"được công bố trên nền tảng Kaggle, bao gồm thông tin về nhân khẩu học, hành vi tín dụng, lịch sử thanh toán và trạng thái vỡ nợ của 30.000 khách hàng sử dụng thẻ tín dụng tại Đài Loan trong khoảng thời gian từ tháng 4/2005 đến tháng 9/2005. Tuy nhiên, trong nghiên cứu này, chỉ một phần dữ liệu ngẫu nhiên được chọn, bao gồm 10.000 quan sát từ tổng số 30.000 quan sát.

Mặc dù dữ liệu không được thu thập tại Việt Nam, các biến trong bộ dữ liệu phản ánh đầy đủ các yếu tố thường được sử dụng trong mô hình đánh giá rủi ro tín dụng cá nhân ở Việt Nam như giới tính, độ tuổi, tình trạng hôn nhân, trình độ học vấn, hạn mức tín dụng, lịch sử thanh toán, và tình trạng trả nợ đúng hạn. Do đó, bộ dữ liệu này được xem là phù hợp để nghiên cứu và thử nghiệm các mô hình phân loại rủi ro tín dụng trong bối cảnh Việt Nam.

Bộ dữ liệu bao gồm 25 biến, được phân loại thành các nhóm sau:

- Mã định danh:
  - ID: Mã số định danh của khách hàng
- Đặc điểm nhân khẩu học:
  - SEX: Giới tính của khách hàng (1 = Nam, 2 = Nữ)
  - EDUCATION: Trình độ học vấn của khách hàng (1 = Sau đại học, 2 = Đại học, 3 = Trung học phổ thông, 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.
- Thông tin tín dụng:
  - LIMIT_BAL: Hạn mức tín dụng của khách hàng (đơn vị: NT$).
- Lịch sử thanh toán:
  - PAY_0 đến PAY_6 (không có PAY_1): Tình trạng thanh toán trong 6 tháng gần nhất, với các mức độ từ -1 (trả đúng hạn), 1 (chậm thanh toán 1 tháng) đến 9 (chậm thanh toán 9 tháng hoặc hơn).
- Số dư hóa đơn hàng tháng:
  - BILL_AMT1 đến BILL_AMT6: Số dư hóa đơn của khách hàng trong 6 tháng gần nhất.
- Số tiền thanh toán hàng tháng:
  - PAY_AMT1 đến PAY_AMT6: Số tiền khách hàng thanh toán trong 6 tháng gần nhất.
- Biến mục tiêu:

– default.payment.next.month: Trạng thái vỡ nợ của khách hàng trong tháng tiếp theo (1 = Vỡ nợ, 0 = Không vỡ nợ).

3.2 Xử lí dữ liệu

3.2.1 Chuẩn hóa tên cột

Trong bộ dữ liệu ban đầu, các cột phản ánh tình trạng thanh toán của khách hàng được đặt tên từ PAY_0 đến PAY_6 (không có PAY_1), tương ứng với tình trạng thanh toán của khách hàng trong các tháng từ tháng 9/2005 trở về trước. Tuy nhiên, do cấu trúc dữ liệu này có thể gây nhầm lẫn khi xử lý, cột PAY_0 được đổi tên thành PAY_1 để thống nhất với thứ tự từ tháng 4/2005 đến tháng 9/2005. Cụ thể, PAY_1 phản ánh tình trạng thanh toán trong tháng 9/2005, PAY_2 cho tháng 8/2005, và tiếp tục như vậy cho đến PAY_6 cho tháng 4/2005. Việc đổi tên này giúp dễ dàng hơn khi tham chiếu đến các tháng cụ thể trong quá trình phân tích và mô hình hóa.

3.2.2 Xử lý giá trị không hợp lệ trong các biến

Trong bộ dữ liệu, một số biến chứa các giá trị không hợp lệ hoặc không xác định, cần được xử lý để đảm bảo tính đồng nhất và chuẩn hóa trong quá trình phân tích.

Cụ thể, đối với biến EDUCATION, các giá trị 0 (không hợp lệ), 5 và 6 (không xác định) được thay thế bằng 4 (Khác). Việc gộp này nhằm giảm thiểu sự phân tán không cần thiết và tăng khả năng diễn giải hợp lý về trình độ học vấn của khách hàng.

Tương tự, trong biến MARRIAGE, giá trị 0 (không hợp lệ) được thay bằng 3 (Khác) để thống nhất cách phân loại tình trạng hôn nhân trong dữ liệu.

Bên cạnh đó, các cột số dư hóa đơn từ BILL_AMT1 đến BILL_AMT6 có thể chứa các giá trị âm, điều này không phù hợp trong bối cảnh dữ liệu tài chính. Do đó, toàn bộ các giá trị âm trong các cột này được thay thế bằng 0 nhằm đảm bảo tính hợp lý và tránh sai lệch trong các phân tích tiếp theo.

Cuối cùng, các biến từ PAY_1 đến PAY_6 phản ánh tình trạng thanh toán của khách hàng trong 6 tháng gần nhất. Các giá trị -2 và -1 (ngụ ý thanh toán đúng hạn) được quy đổi thành 0 để chuẩn hóa dữ liệu. Việc này giúp đơn giản hóa biểu diễn và tạo điều kiện thuận lợi cho quá trình xây dựng mô hình dự đoán.

3.2.3 Điều chỉnh nhãn "vỡ nợ"

Nhãn "vỡ nợ"trong cột default.payment.next.month cần được điều chỉnh để đảm bảo tính chính xác. Nếu một khách hàng không có khoản thanh toán chậm nào trong 6 tháng nhưng lại bị gán nhãn "vỡ nợ", nhãn này được sửa thành "không vỡ nợ". Ngược lại, nếu khách hàng có nợ trong cả 6 tháng nhưng lại bị gán nhãn "không vỡ nợ", nhãn này sẽ được sửa thành "vỡ nợ". Việc này giúp loại bỏ các nhãn sai và tạo ra một bộ dữ liệu chính xác hơn cho việc xây dựng mô hình.

3.3 Thống kê mô tả

Để phân tích dữ liệu, thống kê mô tả được sử dụng để hiểu rõ hơn về các đặc điểm của bộ dữ liệu.

Biến LIMIT_BAL, đại diện cho hạn mức tín dụng của khách hàng, có giá trị tối thiểu là 10000 và tối đa là 500000, với giá trị trung bình là 119748. Điều này phản ánh sự đa dạng về khả năng tài chính của khách hàng, từ những người có hạn mức tín dụng thấp đến những người có hạn mức cao.

Bảng 3.1: Các chỉ số thống kê mô tả của các biến trong bộ dữ liệu

Biến	**Min**	**Mean**	**Max**
LIMIT_BAL	10000	119748	500000
PAY_1	0.0000	0.3959	8.0000
PAY_2	0.0000	0.4158	7.0000
PAY_3	0.0000	0.3948	8.0000
PAY_4	0.0000	0.3398	8.0000
PAY_5	0.0000	0.2922	8.0000
PAY_6	0.0000	0.2976	8.0000
BILL_AMT1	0	42330	189003
BILL_AMT2	0	40819	182435
BILL_AMT3	0	38797	173325
BILL_AMT4	0	35729	161573
BILL_AMT5	0	33349	151194
BILL_AMT6	0	31969	146194
PAY_AMT1	0	2646	11662
PAY_AMT2	0	2543	11236
PAY_AMT3	0	2201	11000.0
PAY_AMT4	0	1991	10902
PAY_AMT5	0	1987	10800
PAY_AMT6	0	1917	10400

Biến	**Giá trị**	**Tần suất**
SEX	1 (Nam)	3974
	2 (Nữ)	6026
EDUCATION	1 (Sau đại học)	2919
	2 (Đại học)	5108
	3 (Trung học phổ thông)	1862
	4 (Khác)	111
MARRIAGE	1 (Đã kết hôn)	4342
	2 (Độc thân)	5517
	3 (Khác)	141
default.payment.next.month	0 (Không vỡ nợ)	8237
	1 (Vỡ nợ)	1763

Các biến PAY_1 đến PAY_6, đại diện cho tình trạng thanh toán của khách hàng trong sáu tháng trước, đều có giá trị tối thiểu là 0 (thanh toán đúng hạn) và tối đa lên tới 8 (tương ứng với việc chậm thanh toán 8 tháng). Tuy nhiên, giá trị trung bình của các biến này dao động từ 0.29 đến 0.42, cho thấy phần lớn khách hàng thanh toán đúng hạn hoặc chỉ có sự chậm trễ nhẹ. Điều này chỉ ra rằng đa số khách hàng có khả năng thanh toán ổn định.

Với các biến về số dư hóa đơn như BILL_AMT1 đến BILL_AMT6, ta cũng thấy sự phân tán rõ rệt. Mặc dù giá trị tối thiểu của các biến này là 0, nhưng giá trị tối đa có thể lên tới hơn 190000, với giá trị trung bình dao động từ 31969 đến 42330. Điều này phản ánh sự khác biệt lớn giữa các khách hàng về mức độ sử dụng tín dụng và số dư hóa đơn cần thanh toán.

Các biến SEX, EDUCATION, MARRIAGE cho thấy cơ cấu phân bố khá rõ ràng giữa các nhóm. Cụ thể, biến giới tính (SEX) cho thấy số lượng khách hàng nữ chiếm ưu thế với 6026 người, so với 3974 khách hàng nam. Về trình độ học vấn (EDUCATION), phần lớn khách hàng có trình độ đại học (5108 người), tiếp đến là sau đại học (2919 người), trung học phổ thông (1862 người), và nhóm còn lại có trình độ khác chỉ chiếm 111 người. Biến tình trạng hôn nhân (MARRIAGE)

cho thấy số lượng người độc thân (5517 người) cao hơn một chút so với nhóm đã kết hôn (4342 người), trong khi nhóm "khác" chỉ có 141 người.

Đối với biến mục tiêu default.payment.next.month (đại diện cho việc khách hàng có khả năng vỡ nợ trong tháng tiếp theo hay không), dữ liệu cho thấy có 1763 khách hàng vỡ nợ, tương đương 17.63% tổng số, trong khi số còn lại (8237 khách hàng) không rơi vào tình trạng vỡ nợ. Điều này phản ánh rằng hiện tượng vỡ nợ là hiện tượng ít xảy ra hơn trong bộ dữ liệu, tuy nhiên vẫn là một tỷ lệ đáng kể cần quan tâm khi xây dựng mô hình dự báo. Tỷ lệ này cho thấy sự mất cân bằng nhãn không quá nghiêm trọng, và trong thực tế, mức phân bố như vậy vẫn thường được chấp nhận trong nhiều nghiên cứu mô hình hóa hành vi tín dụng.

3.4 Kiểm tra dữ liệu và xử lý giá trị bất thường bằng phương pháp IQR

Trước khi tiến hành các bước phân tích và xây dựng mô hình, bộ dữ liệu đã được kiểm tra chất lượng nhằm đảm bảo tính toàn vẹn. Kết quả cho thấy tất cả các biến không có giá trị thiếu, do đó không cần áp dụng các kỹ thuật xử lý giá trị thiếu như nội suy hay loại bỏ quan sát.

Tiếp theo, để giảm thiểu ảnh hưởng của các giá trị ngoại lai (outliers) đối với kết quả phân tích, phương pháp khoảng tứ phân vị (Interquartile Range – IQR) đã được áp dụng để phát hiện và loại bỏ các điểm dữ liệu bất thường ở các biến định lượng. Các biến dạng thứ hạng như PAY_1 đến PAY_6, cũng như biến phân loại và biến mục tiêu, đã được loại trừ khỏi quá trình phát hiện ngoại lai để đảm bảo tính hợp lý.

Bảng 3.2: Số lượng giá trị ngoại lai theo từng biến (IQR)

Biến	**Ngoại lai**
LIMIT_BAL	419
AGE	0
BILL_AMT1	255
BILL_AMT2	257
BILL_AMT3	379
BILL_AMT4	443
BILL_AMT5	433
BILL_AMT6	395
PAY_AMT1	419
PAY_AMT2	464
PAY_AMT3	447
PAY_AMT4	334
PAY_AMT5	318
PAY_AMT6	290

Kết quả phân tích cho thấy một số biến như LIMIT_BAL, các biến từ BILL_AMT1 đến BILL_AMT6, và từ PAY_AMT1 đến PAY_AMT6 xuất hiện số lượng giá trị ngoại lai tương đối lớn, trong đó, riêng biến PAY_AMT2 có tới 464 điểm ngoại lai. Sau khi loại bỏ các quan sát chứa ít nhất một ngoại lai trong các biến này, số lượng dòng dữ liệu giảm từ 10000 xuống còn 7726, tức đã loại bỏ tổng cộng 2274 quan sát khỏi tập dữ liệu ban đầu.

Việc loại bỏ các quan sát này cũng ảnh hưởng đến tỷ lệ phân bố của biến mục tiêu default.payment.next.month. Cụ thể, số lượng khách hàng không vỡ nợ giảm xuống còn 6229 người (chiếm khoảng 80.62%), trong khi số người vỡ nợ còn lại là 1497 (tương đương 19.38%). Tỷ lệ này thể hiện sự mất cân bằng không quá nghiêm trọng và vẫn được xem là chấp nhận được trong các bài toán phân loại thực tiễn. Việc xử lý outliers theo phương pháp IQR không chỉ giúp làm sạch dữ liệu mà còn giảm thiểu khả năng mô hình bị ảnh hưởng bởi các giá trị cực đoan trong quá trình huấn luyện.

3.5 Chọn biến bằng cách tính chỉ số IV

Trong bước này, chỉ số IV (Information Value) được tính toán để đánh giá mức độ liên quan giữa các biến độc lập và biến mục tiêu (default.payment.next.month). Chỉ số IV cho phép xác định những biến có ảnh hưởng lớn nhất đến việc dự đoán khả năng vỡ nợ của khách hàng.

Bảng 3.3: Giá trị Information Value (IV) của các biến độc lập

Biến	IV
PAY_1	2.491
PAY_2	1.548
PAY_3	1.201
PAY_4	1.122
PAY_5	1.017
PAY_6	0.906
LIMIT_BAL	0.279
PAY_AMT1	0.239
PAY_AMT2	0.195
PAY_AMT3	0.156
PAY_AMT4	0.152

Biến	IV
BILL_AMT1	0.138
PAY_AMT6	0.120
BILL_AMT2	0.112
PAY_AMT5	0.105
BILL_AMT3	0.098
BILL_AMT6	0.087
BILL_AMT5	0.081
BILL_AMT4	0.072
EDUCATION	0.048
AGE	0.011
MARRIAGE	0.006
SEX	0.002

Kết quả tính toán IV cho các biến trong bộ dữ liệu cho thấy một số biến có chỉ số IV cao, đặc biệt là các biến liên quan đến lịch sử thanh toán như PAY_1, PAY_2, PAY_3, với các chỉ số IV lần lượt là 2.491, 1.548 và 1.201. Điều này cho thấy các biến này có mối quan hệ mạnh mẽ với khả năng vỡ nợ của khách hàng và sẽ được giữ lại trong mô hình.

Trong khi đó, các biến như AGE, MARRIAGE và SEX có chỉ số IV rất thấp, lần lượt là 0.011, 0.006 và 0.002. Những biến này không có khả năng phân biệt rõ ràng giữa các nhóm khách hàng vỡ nợ và không vỡ nợ, do đó chúng sẽ được loại bỏ khỏi mô hình. Việc loại bỏ các biến có IV thấp giúp giảm sự phức tạp của mô hình, đồng thời đảm bảo rằng mô hình chỉ sử dụng các biến có ảnh hưởng đáng kể đến biến mục tiêu.

Sau khi loại bỏ các biến có IV nhỏ hơn 0.02, các biến còn lại sẽ được sử dụng để xây dựng mô hình phân loại. Những biến này có khả năng phân biệt rõ ràng giữa các nhóm khách hàng, giúp cải thiện hiệu quả dự đoán của mô hình.

3.6 Chia tập dữ liệu

Sau khi loại bỏ các biến có chỉ số Information Value (IV) thấp cũng như biến không mang tính giải thích như ID, bộ dữ liệu đã được tinh gọn để chỉ giữ lại những biến có đóng góp đáng kể trong việc dự báo khả năng vỡ nợ. Cụ thể, tập dữ liệu sau xử lý bao gồm các biến có giá trị IV từ 0.02 trở lên, đảm bảo tính hiệu quả và chính xác trong việc xây dựng mô hình phân loại.

Sau khi hoàn thiện bước chọn biến, dữ liệu được chia thành hai phần: tập huấn luyện (training set) và tập kiểm tra (testing set), theo tỷ lệ 70:30. Việc chia tách được thực hiện ngẫu nhiên nhưng có kiểm soát bằng cách cố định hạt giống khởi tạo (seed = 10), nhằm đảm bảo tính tái lập cho quá trình huấn luyện mô hình. Tập huấn luyện sẽ được sử dụng để xây dựng mô hình dự báo, trong khi tập kiểm tra sẽ giúp đánh giá khả năng dự đoán của mô hình trên dữ liệu mới, nhằm kiểm tra mức độ tổng quát và độ chính xác của mô hình khi áp dụng vào thực tế. Ở đây, phương pháp chia sử dụng kỹ thuật lấy mẫu phân tầng (stratified sampling), giúp duy trì tỷ lệ phân phối của biến mục tiêu - cụ thể là tỷ lệ giữa các khách hàng vỡ nợ và không vỡ nợ - trong cả hai tập train và test. Điều này đảm bảo rằng mô hình được huấn luyện và kiểm tra trên các tập dữ liệu có phân phối tương tự nhau, từ đó tránh thiên lệch và nâng cao độ tin cậy trong quá trình đánh giá hiệu suất dự báo.

Chương 4

Kết quả các mô hình

4.1 Mô hình Logit

4.1.1 Mô hình Logit với biến gốc

Kết quả mô hình hồi quy Logit ban đầu cho thấy các biến liên quan đến lịch sử thanh toán như PAY_1, PAY_4, PAY_5, và PAY_6 có ý nghĩa thống kê cao, phù hợp với kết quả phân tích IV trước đó. Bên cạnh đó, một số yếu tố khác như LIMIT_BAL và một vài mức độ của biến EDUCATION cũng có ảnh hưởng nhất định đến xác suất vỡ nợ.

Để đơn giản hóa mô hình và loại bỏ các biến không có ý nghĩa thống kê, phương pháp lựa chọn biến theo hồi quy lùi từng bước (backward stepwise selection) đã được áp dụng. Mô hình cuối cùng giữ lại 10 biến bao gồm: LIMIT_BAL, EDUCATION, PAY_1, PAY_3, PAY_4, PAY_5, PAY_6 và PAY_AMT1. Hầu hết các biến trong mô hình sau khi lựa chọn đều có ý nghĩa thống kê với mức ý nghĩa 5%, đặc biệt là PAY_1, biến có hệ số lớn nhất, cho thấy đây là yếu tố ảnh hưởng mạnh đến khả năng vỡ nợ của khách hàng.

```
Call:
glm(formula = default.payment.next.month ~ LIMIT_BAL + EDUCATION +
    PAY_1 + PAY_3 + PAY_4 + PAY_5 + PAY_6 + PAY_AMT1, family = binomial(link = "logit"),
    data = train.data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.800e+00  1.527e-01 -18.334  < 2e-16 ***
LIMIT_BAL   -2.127e-06  7.857e-07  -2.707 0.006789 **
EDUCATION2  -2.618e-01  1.228e-01  -2.133 0.032931 *
EDUCATION3  -8.921e-02  1.468e-01  -0.608 0.543429
EDUCATION4  -1.422e+01  2.694e+02  -0.053 0.957913
PAY_1        1.618e+00  6.026e-02  26.846  < 2e-16 ***
PAY_3        2.774e-01  6.223e-02   4.457 8.30e-06 ***
PAY_4        4.589e-01  7.239e-02   6.339 2.31e-10 ***
PAY_5        3.344e-01  7.735e-02   4.324 1.53e-05 ***
PAY_6        5.307e-01  6.729e-02   7.887 3.10e-15 ***
PAY_AMT1    -1.140e-04  3.451e-05  -3.304 0.000953 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hình 4.1: Kết quả mô hình Logit với biến gốc

Để đánh giá hiệu quả mô hình, ma trận nhầm lẫn (confusion matrix) đã được sử dụng trên cả tập huấn luyện và tập kiểm tra. Trên tập huấn luyện, tỷ lệ phân loại đúng tổng thể đạt khoảng 88.7%, trong đó tỷ lệ dự đoán đúng nhóm khách hàng không vỡ nợ là 77.1% và nhóm vỡ nợ là 11.6%. Trên tập kiểm tra, mô hình vẫn giữ được độ chính xác tương đối ổn định với tổng tỷ lệ đúng là khoảng 87.0%, bao gồm 74.9% khách hàng không vỡ nợ và 12.0% khách hàng vỡ nợ được phân loại đúng.

Điều này cho thấy mô hình vẫn duy trì khả năng phân biệt tốt giữa các nhóm khách hàng vỡ nợ và không vỡ nợ khi được áp dụng vào tập kiểm tra.

```
train.pred1          0          1
          0 0.77112220 0.07321131
          1 0.03937881 0.11628767
```

Hình 4.2: Confusion matrix trên tập train

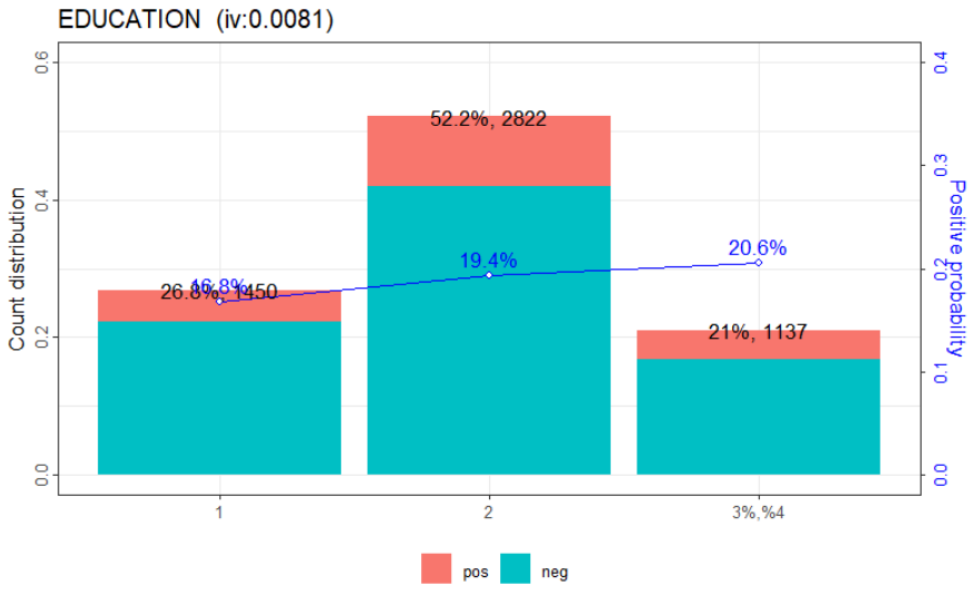
```
test.pred1          0          1
          0 0.74924471 0.08329737
          1 0.04704359 0.12041433
```

Hình 4.3: Confusion matrix trên tập test

4.1.2 Mô hình Logit với WOE

Một mô hình hồi quy Logit khác cũng được xây dựng bằng cách sử dụng các biến đã được chuyển đổi theo phương pháp WOE (Weight of Evidence). Phương pháp này giúp cải thiện khả năng phân tích dữ liệu, đặc biệt là khi làm việc với các biến phân loại, nhờ vào việc biến đổi các giá trị của chúng thành các chỉ số dễ so sánh và có thể diễn giải được. Các biến WOE có thể giúp giảm thiểu sự ảnh hưởng của các giá trị ngoại lệ và phân phối lệch, từ đó tăng cường độ chính xác của mô hình.

Bước đầu tiên trong việc xây dựng mô hình là phân loại các biến độc lập theo WOE trên tập huấn luyện. Sau đó, các biểu đồ WOE đã được tạo ra để trực quan hóa phân phối và sự thay đổi của các biến. Dưới đây là một ví dụ về phân phối của biến EDUCATION sau khi phân loại theo WOE. Có thể thấy, tỷ lệ khách hàng vỡ nợ trong nhóm đầu tiên (sau đại học) là 16.8%, trong nhóm thứ hai (đại học) là 19.4% và trong nhóm thứ ba (Trung học phổ thông và Khác) là 20.6%. Mặc dù sự khác biệt không quá lớn, xu hướng chung cho thấy khách hàng có trình độ học vấn cao hơn có xác suất vỡ nợ thấp hơn.



Hình 4.4: Biểu đồ phân phối biến EDUCATION sau phân loại theo WOE

Với mô hình hồi quy Logit ban đầu, tất cả các biến đã được đưa vào. Kết quả cho thấy một số biến như LIMIT_BAL_woe, PAY_1_woe, PAY_4_woe, PAY_5_woe, PAY_6_woe, PAY_AMT1_woe, PAY_AMT2_woe và PAY_AMT3_woe có mức độ ảnh hưởng đáng kể đối với khả năng dự báo vỡ nợ (với p-value < 0.1), trong khi các biến còn lại không có sự ảnh hưởng đáng kể đến kết quả dự báo (p-value > 0.1).

Để tối ưu hóa mô hình, phương pháp lựa chọn biến theo hồi quy lùi từng bước (backward stepwise selection) đã được áp dụng nhằm loại bỏ các biến không quan trọng, đồng thời cải thiện độ chính xác của mô hình. Sau khi lọc, mô hình hồi quy Logit cuối cùng giữ lại 10 biến bao gồm: LIMIT_BAL_woe, EDUCATION_woe, PAY_1_woe, PAY_4_woe, PAY_5_woe, PAY_6_woe, PAY_AMT1_woe, PAY_AMT2_woe, PAY_AMT3_woe và PAY_AMT6_woe. Hầu hết các biến trong mô hình đều có ý nghĩa thống kê với mức ý nghĩa 10%, đặc biệt là PAY_1_woe, biến có hệ số lớn nhất, cho thấy đây là yếu tổ ảnh hưởng mạnh đến khả năng vỡ nợ của khách hàng.

```
Call:
glm(formula = default.payment.next.month ~ LIMIT_BAL_woe + EDUCATION_woe +
    PAY_1_woe + PAY_4_woe + PAY_5_woe + PAY_6_woe + PAY_AMT1_woe +
    PAY_AMT2_woe + PAY_AMT3_woe + PAY_AMT6_woe, family = binomial(link = "logit"),
    data = train.data_woe)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.45826    0.05317 -27.427  < 2e-16 ***
LIMIT_BAL_woe  0.34764    0.09917   3.506 0.000456 ***
EDUCATION_woe -0.83193    0.57354  -1.451 0.146912
PAY_1_woe      0.86042    0.03061  28.113  < 2e-16 ***
PAY_4_woe      0.34235    0.06694   5.114 3.15e-07 ***
PAY_5_woe      0.29480    0.06579   4.481 7.43e-06 ***
PAY_6_woe      0.48866    0.06016   8.122 4.59e-16 ***
PAY_AMT1_woe   0.13267    0.07406   1.791 0.073232 .
PAY_AMT2_woe   0.34764    0.08452   4.113 3.90e-05 ***
PAY_AMT3_woe   0.33640    0.10891   3.089 0.002010 **
PAY_AMT6_woe   0.24028    0.13823   1.738 0.082169 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hình 4.5: Kết quả mô hình Logit với WOE

Để đánh giá hiệu quả mô hình, ma trận nhầm lẫn (confusion matrix) đã được sử dụng trên cả tập huấn luyện và tập kiểm tra. Trên tập huấn luyện, tỷ lệ phân loại đúng tổng thể đạt khoảng 88.4%, trong đó tỷ lệ dự đoán đúng nhóm khách hàng không vỡ nợ là 76.6% và nhóm vỡ nợ là 11.8%. Trên tập kiểm tra, mô hình vẫn giữ được độ chính xác tương đối ổn định với tổng tỷ lệ đúng là khoảng 87.2%, bao gồm 75.1% khách hàng không vỡ nợ và 12.1% khách hàng vỡ nợ được phân loại đúng. Điều này cho thấy mô hình vẫn duy trì khả năng phân biệt tốt giữa các nhóm khách hàng vỡ nợ và không vỡ nợ khi được áp dụng vào tập kiểm tra.

```
train.pred2          0          1
          0 0.76594565 0.07117767
          1 0.04455537 0.11832132
```

Hình 4.6: Confusion matrix trên tập train

```
test.pred2          0          1
          0 0.75097108 0.08243418
          1 0.04531722 0.12127751
```

Hình 4.7: Confusion matrix trên tập test

4.1.3 So sánh mô hình Logit với biến gốc và mô hình Logit với WOE

Nhằm đánh giá hiệu quả của mô hình phân loại, đặc biệt trong bối cảnh dự đoán khả năng vỡ nợ, một số chỉ số quan trọng đã được sử dụng, bao gồm:

Sensitivity: phản ánh tỷ lệ các trường hợp thực sự vỡ nợ được mô hình dự đoán đúng. Chỉ số này quan trọng khi việc bỏ sót khách hàng có khả năng vỡ nợ là điều cần tránh. Sensitivity được tính bằng công thức:

$$\text{Sensitivity} = \frac{\text{Số dự đoán đúng nhóm vỡ nợ}}{\text{Tổng số khách hàng thực sự vỡ nợ}}$$

Specificity: Là tỷ lệ các khách hàng không vỡ nợ được mô hình nhận diện chính xác. Chỉ số này phản ánh khả năng mô hình tránh đưa ra các cảnh báo sai đối với khách hàng tốt.

$$\text{Specificity} = \frac{\text{Số dự đoán đúng nhóm không vỡ nợ}}{\text{Tổng số khách hàng thực sự không vỡ nợ}}$$

AUC (Area Under the Curve): Là diện tích dưới đường cong ROC (Receiver Operating Characteristic), thể hiện khả năng phân biệt giữa hai nhóm (vỡ nợ và không vỡ nợ). AUC nằm trong khoảng từ 0.5 đến 1, với giá trị gần 1 cho thấy mô hình phân biệt hai nhóm rất tốt.

Gini index: Là một chỉ số tương quan chặt chẽ với AUC, được tính bằng công thức:

$$\text{Gini} = 2 \times \text{AUC} - 1$$

Chỉ số Gini cao cho thấy mô hình có khả năng phân loại mạnh giữa hai nhóm khách hàng.

Bảng 4.1: So sánh hiệu suất hai mô hình Logit trên tập kiểm tra

Mô hình	Sensitivity	Specificity	AUC	Gini
Logit với biến gốc	0.5911	0.9409	0.9300	0.8600
Logit với WOE	0.5953	0.9431	0.9281	0.8562

Kết quả so sánh giữa hai mô hình Logit cho thấy hiệu suất dự đoán của chúng tương đối tương đồng. Cụ thể, mô hình sử dụng biến đã chuyển đổi theo WOE có chỉ số sensitivity và specificity nhỉnh hơn, cho thấy khả năng phát hiện đúng khách hàng vỡ nợ và không vỡ nợ được cải thiện đôi chút. Tuy nhiên, mô hình sử dụng biến gốc lại đạt giá trị AUC và Gini cao hơn, phản ánh khả năng phân biệt giữa hai nhóm khách hàng tốt hơn. Nhìn chung, sự chênh lệch giữa hai mô hình là không đáng kể, tuy nhiên tùy vào mục tiêu phân tích (ưu tiên phát hiện khách hàng rủi ro hay phân loại tổng thể chính xác), người phân tích có thể lựa chọn mô hình phù hợp. Trong bối cảnh này, nếu ưu tiên được đặt vào việc phát hiện các trường hợp có nguy cơ vỡ nợ, mô hình sử dụng biến WOE là lựa chọn hợp lý.

4.2 Rừng ngẫu nhiên (Random Forest)

Mô hình Random Forest được huấn luyện trên tập train, với biến mục tiêu là default.payment.next.month. Quá trình huấn luyện sử dụng phương pháp Cross-Validation với 10 lần lặp lại để đảm bảo tính ổn định và khách quan trong đánh giá mô hình. Tham số mtry (số biến được chọn ngẫu nhiên tại mỗi cây) được tinh chỉnh trong quá trình huấn luyện, và giá trị tối ưu là mtry = 12, tương ứng với tỷ lệ dự đoán đúng (Accuracy) cao nhất đạt 88.50%.

Sau khi huấn luyện, mô hình được áp dụng lên tập kiểm tra để dự đoán xác suất khách hàng vỡ nợ. Với ngưỡng phân loại 0.5, mô hình đạt được các chỉ số hiệu suất như sau:

Sensitivity đạt 58.90%, phản ánh khả năng phát hiện đúng các trường hợp vỡ nợ. Đây là một chỉ số đặc biệt quan trọng trong bài toán quản trị rủi ro tín dụng, vì ngân hàng quan tâm đến việc nhận diện đúng những khách hàng có khả năng không trả nợ. Giá trị này cho thấy mô hình mới chỉ dự đoán đúng khoảng 59% số khách hàng thực sự vỡ nợ.

Ngược lại, Specificity đạt 94.47%, nghĩa là mô hình nhận diện rất tốt các khách hàng không vỡ nợ. Điều này dẫn đến xu hướng nghiêng về việc "an toàn", tức là mô hình tránh gán nhãn vỡ nợ một cách quá mức, nhưng lại bỏ sót một phần đáng kể các trường hợp rủi ro thực sự.

Tóm lại, mô hình có xu hướng dự đoán tốt nhóm khách hàng không vỡ nợ hơn là nhóm có nguy cơ vỡ nợ.

Bên cạnh đó, đánh giá độ quan trọng của các biến trong mô hình cho thấy biến quan trọng nhất là PAY_1 – phản ánh tình trạng thanh toán gần nhất của khách hàng. Các biến tiếp theo có ảnh

hưởng đáng kể bao gồm PAY_2, PAY_4, PAY_6, PAY_5, PAY_AMT1, và BILL_AMT1. Những biến liên quan đến lịch sử thanh toán và dư nợ đóng vai trò quan trọng trong việc dự đoán khả năng vỡ nợ, cho thấy tính hợp lý của mô hình trong bối cảnh thực tế tài chính.

	Overall <dbl>
PAY_1	100.000000
PAY_2	29.372338
PAY_4	17.671530
PAY_6	12.921700
PAY_5	12.800251
PAY_AMT1	12.671198
BILL_AMT1	12.343070
BILL_AMT6	11.997319
PAY_AMT4	11.355293
PAY_AMT2	11.353121

1-10 of 20 rows

Hình 4.8: Biến và độ quan trọng trong mô hình Random Forest

4.3 K Láng giềng gần nhất (K-Nearest Neighbors - KNN)

Tiếp theo, mô hình K-Nearest Neighbors (KNN) được áp dụng để dự đoán khả năng vỡ nợ của khách hàng. Quá trình huấn luyện được thực hiện trên tập huấn luyện với 10-fold cross-validation và giá trị k được chọn tối ưu dựa trên tỷ lệ dự báo đúng (Accuracy). Kết quả cho thấy mô hình đạt Accuracy cao nhất là 80.90% tại k = 11.

Khi đánh giá trên tập kiểm tra, chỉ số Sensitivity (khả năng nhận diện đúng các trường hợp vỡ nợ) chỉ đạt 11.65%, trong khi Specificity (khả năng nhận diện đúng các trường hợp không vỡ nợ) lại rất cao (96.59%). Điều này cho thấy mô hình có xu hướng dự đoán đa số khách hàng là không vỡ nợ, dẫn đến khả năng nhận diện sai các khách hàng có rủi ro.

Nhìn chung, mô hình KNN hoạt động kém trong việc phát hiện các khách hàng vỡ nợ – nhóm mục tiêu quan trọng trong bối cảnh quản trị rủi ro tín dụng. Điều này thể hiện rõ qua các chỉ số Sensitivity thấp. Do đó, mô hình này không phù hợp để áp dụng trong thực tiễn nếu mục tiêu là dự báo rủi ro vỡ nợ.

4.4 So sánh các mô hình Logit, Random Forest và KNN

Bảng 4.2: So sánh hiệu suất các mô hình Logit, Random Forest và KNN trên tập kiểm tra

Mô hình	**Sensitivity**	**Specificity**	**AUC**	**Gini**
Logit với WOE	0.5953	0.9431	0.9281	0.8562
Random Forest	0.5890	0.9447	0.9397	0.8795
KNN	0.1165	0.9659	0.6720	0.3439

Có thể thấy, mô hình Random Forest có chỉ số AUC và Gini cao nhất (AUC = 0.9397, Gini = 0.8795), phản ánh khả năng phân biệt hai lớp tốt hơn so với hai mô hình còn lại. Mô hình Logit với WOE cũng thể hiện hiệu suất cao và ổn định, với AUC = 0.9281 và Gini = 0.8562, chỉ thấp hơn Random Forest một chút, trong khi lại có Sensitivity cao nhất (đạt 0.5953) – ngụ ý phát hiện được nhiều trường hợp vỡ nợ hơn.

Ngược lại, mô hình KNN cho kết quả kém hơn rõ rệt về mọi phương diện, đặc biệt là Sensitivity rất thấp (chỉ đạt 0.1165), cho thấy mô hình này bỏ sót phần lớn các trường hợp vỡ nợ. Dù độ Specificity khá cao (đạt 0.9659), nó không đủ bù đắp cho hiệu suất tổng thể thấp, thể hiện rõ qua AUC và Gini rất thấp so với hai mô hình còn lại.

Tóm lại, giữa ba mô hình, Random Forest có hiệu suất tổng thể tốt nhất, trong khi Logit với WOE là lựa chọn hiệu quả, dễ triển khai hơn do tính đơn giản và mô hình KNN tỏ ra không phù hợp trong bối cảnh này. Trong trường hợp ưu tiên được đặt vào việc phát hiện các trường hợp có nguy cơ vỡ nợ, mô hình Logit với WOE là mô hình được chọn.

4.5 Xây dựng hệ thống tính điểm tín dụng (Credit Scoring) từ mô hình Logit với WOE

Kết quả từ mô hình Logit với WOE được sử dụng làm cơ sở để xây dựng bảng điểm (scorecard). Mục đích của việc này là chuyển xác suất vỡ nợ thành điểm tín dụng, giúp dễ dàng đánh giá và phân loại khách hàng dựa trên mức độ rủi ro tín dụng của họ. Để tính điểm tín dụng, trước tiên ta tính z-score, một đại lượng biểu thị tỷ lệ giữa xác suất vỡ nợ và xác suất không vỡ nợ, theo công thức:

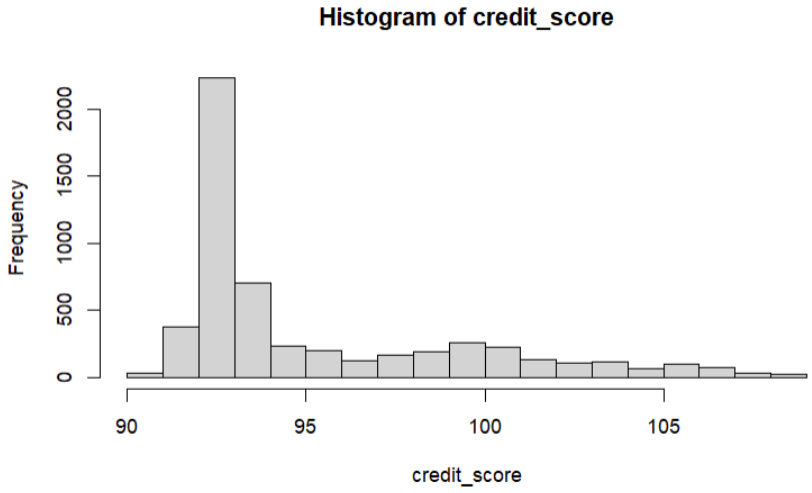
$$z\_score = \log\left(\frac{p}{1-p}\right)$$

Trong đó, p là xác suất vỡ nợ của khách hàng. Sau khi tính được z-score, điểm tín dụng được tính theo công thức:

$$credit\_score = 100 + 2 \times z\_score$$

Ở đây, 100 là offset, và 2 là scaling factor. Scaling factor điều chỉnh số điểm tăng lên hoặc giảm xuống khi thay đổi xác suất vỡ nợ, trong khi offset là giá trị điểm ban đầu.

Cuối cùng, điểm tín dụng được tính toán cho từng khách hàng sẽ được sử dụng để phân loại khách hàng theo mức độ rủi ro tín dụng của họ. Biểu đồ phân phối điểm tín dụng cũng được xây dựng để trực quan hóa kết quả và hỗ trợ việc phân tích mức độ rủi ro của các khách hàng trong tập huấn luyện.



Hình 4.9: Biểu đồ phân phối điểm tín dụng

Từ biểu đồ có thể thấy, số lượng khách hàng có điểm tín dụng nằm trong khoảng từ 90 đến 95 cao hơn hẳn so với các khoảng điểm còn lại.

Chương 5

Kết luận và thảo luận

Nghiên cứu này đã tiến hành so sánh hiệu suất của các mô hình phân loại bao gồm: mô hình Logit với biến gốc, mô hình Logit với biến đã chuyển đổi theo Weight of Evidence (WOE), Random Forest và K-Nearest Neighbors (KNN) trong việc dự đoán khả năng vỡ nợ của khách hàng. Kết quả cho thấy mô hình Logit với WOE thể hiện Sensitivity cao nhất, đồng thời có chỉ số AUC và Gini ổn định, cho thấy khả năng phân biệt tốt giữa hai nhóm khách hàng. Mô hình Random Forest tuy có độ chính xác tổng thể cao nhưng lại không cải thiện đáng kể về Sensitivity, trong khi mô hình KNN thể hiện hiệu suất kém rõ rệt do Sensitivity rất thấp.

Từ đó, mô hình Logit với WOE được lựa chọn để xây dựng thang điểm tín dụng (scorecard). Điểm tín dụng được tính toán từ xác suất vỡ nợ thông qua điểm z-score và được quy đổi về thang điểm tuyến tính để dễ áp dụng trong thực tiễn. Phân phối điểm tín dụng cho thấy phần lớn khách hàng nằm trong khoảng điểm từ 90 đến 95, cho thấy đây là nhóm chiếm tỷ trọng lớn nhất trong tập dữ liệu huấn luyện.

Kết quả này cho thấy việc áp dụng WOE không chỉ giúp cải thiện hiệu suất mô hình mà còn hỗ trợ việc xây dựng thang điểm tín dụng một cách hiệu quả và dễ diễn giải. Tuy nhiên, nghiên cứu cũng còn một số hạn chế, như việc chưa tối ưu ngưỡng phân loại hay chưa đánh giá mô hình trên dữ liệu ngoài mẫu (out-of-time). Các yếu tố này có thể được xem xét trong các nghiên cứu tiếp theo để nâng cao độ chính xác và tính ứng dụng thực tiễn của mô hình.

Tài liệu tham khảo

[1] Việt Nam. "Luật các tổ chức tín dụng". **in**(2010).

[2] Basle Committee on Banking Supervision **and** Bank for International Settlements. *Principles for the management of credit risk*. Bank for International Settlements, 2000.

[3] Ngân hàng Nhà nước Việt Nam. *Thông tư số 41/2016/TT-NHNN: Quy định tỷ lệ an toàn vốn đối với ngân hàng, chi nhánh ngân hàng nước ngoài*. 2016. URL: `https://chinhphu.vn/default.aspx?pageid=27160&docid=188256`.

[4] Hussein A Abdou **and** John Pointon. "Credit scoring, statistical techniques and evaluation criteria: a review of the literature". **in***Intelligent systems in accounting, finance and management*: 18.2-3 (2011), **pages** 59–88.

[5] Yannis Marinakis **andothers**. "Optimization of nearest neighbor classifiers via metaheuristic algorithms for credit risk assessment". **in***Journal of Global Optimization*: 42 (2008), **pages** 279–293.