TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN
KHOA TOÁN KINH TẾ

CHUYÊN ĐỀ THỰC TẬP

Chuyên ngành: Toán Kinh tế

Đề tài: Ứng dụng, đánh giá, và so sánh một số mô hình phân loại vào việc phân loại khách hàng thẻ tín dụng

Sinh viên thực hiện: Nguyễn Đức Hiếu

Mã sinh viên: 11131371

Lớp: Toán Kinh tế 55

Giảng viên hướng dẫn: PGS. Nguyễn Thị Minh

Hà Nội, Ngày 17 tháng 4 năm 2017

LỜI MỞ ĐẦU

Đối với các ngân hàng việc chấm điểm tín dụng và phân loại các khách hàng là yếu tố thiết yếu cho lợi nhuận của ngân hàng. Phương pháp truyền thống của việc ra quyết định có cho một cá nhân cụ thể vay hay không là dựa trên đánh giá cảm tính dựa trên kinh nghiệm cá nhân. Tuy nhiên, sự phát triển về quy mô của nền kinh tế đã tạo ra sức ép về nhu cầu vay, đi kèm với đó là sự cạnh tranh giữa các ngân hàng và công nghệ máy tính ngày càng phát triển đã khiến cho việc sử dụng các mô hình thống kê trong việc phân loại các khách hàng tín dụng là bắt buộc đối với các ngân hàng trên thế giới mà ở Việt Nam cũng không phải là ngoại lệ.

Vậy, phương pháp ước lượng nào có thể giúp chúng ta xây dựng được hệ thống chấm điểm tín dụng chính xác nhất? Đã có một số nghiên cứu mang tính chất so sánh hiệu năng giữa các mô hình (Baesens et al. 2003; Xiao, Zhao, and Fei 2006; Lessmann et al. 2015). Sự khác biệt về hiệu năng của các phương pháp khác nhau là có, tuy nhiên hầu như là không đáng kể, và không phải các mô hình hiệu quả hơn đều là các mô hình mới và tân tiến. Theo Thomas (2010), cách hiệu quả để xây dựng một hệ thống lượng định hiệu quả là phối hợp nhiều mô hình khác nhau thay vì tìm kiếm một mô hình toàn diện có thể áp dụng với tất cả các ngân hàng.

Trong bài này, chúng ta sẽ tiếp cận đến một số phương pháp phân loại các khách hàng tín dụng phổ biến hiện nay và rút ra một số kết luận về việc sử dụng các phương pháp khác nhau sao cho hợp lý. Bài viết này được bố cục như sau:

- **Chương 1** đưa ra một cái nhìn tổng quan về lĩnh vực quản trị rủi ro tín dụng trong ngân hàng và đưa ra một số vấn đề của việc chấm điểm tín dụng tại các ngân hàng Việt Nam.
- Các mô hình được thực hiện trong bài này sẽ được giới thiệu ở **Chương 2**, đi kèm với đó là một số chỉ tiêu sẽ được dùng để đánh giá mô hình trong bài này.

- Trong **Chương 3**, chúng ta sẽ ứng dụng các phương pháp được giới thiệu ở **Chương 2** trong một bộ số liệu mẫu về các khách hàng thẻ tín dụng trong một ngân hàng ở Đài Loan.
- Kết quả của các mô hình sẽ được thảo luận ở **Chương 4**, cùng với một số kết luận rút ra được sau khi áp dụng mô hình.

Đề tài này được soạn thảo bằng LaTeX kết hợp với `Sweave` và `knitr` (Xie 2015). Tất cả phân tích được thực hiện trên phần mềm thống kê R version 3.3.3 (2017-03-06) (R Core Team 2017), các phân tích cụ thể được thực hiện sử dụng các gói mở rộng `caret`(Jed Wing et al. 2016), `glmnet`(Friedman, Hastie, and Tibshirani 2010), `tidyverse` (Wickham 2017)...

Em xin cảm ơn giáo viên hướng dẫn, cô Nguyễn Thị Minh, trưởng khoa Toán Ứng dụng trong Kinh tế, cùng với các thầy cô giáo khác trong khoa đã tạo điều kiện cho em thực hiện đề tài này.

MỤC LỤC

Lời mở đầu **1**

Mục lục **4**

Danh sách bảng **5**

Danh sách hình **6**

1 Tổng quan về quản trị rủi ro tín dụng đối với khách hàng cá nhân **7**
1.1 Một số khái niệm 7
1.2 Thực trạng của việc chấm điểm tín dụng tại Việt Nam 7
1.3 Kết luận 7

2 Các phương pháp phân loại khách hàng vay thẻ tín dụng **8**
2.1 Các mô hình phân loại 8
2.1.1 Mô hình logit 8
2.1.1.1 Khái niệm 8
2.1.1.2 Ước lượng mô hình logit 9
2.1.1.3 Diễn giải kết quả ước lượng mô hình 10
2.1.2 Mô hình phân loại tuyến tính 10
2.1.3 Mô hình SVM (Support Vector Machine) 10
2.2 Đánh giá mô hình 10
2.2.1 Đường ROC và phần diện tích dưới đường cong (AUC) . . 10
2.2.2 Thang đo H 10

3 Tình huống nghiên cứu **11**
3.1 Số liệu và các biến số 11

3.2 Ứng dụng mô hình logit 15
3.2.1 Tiền xử lý bộ số liệu 15
3.2.2 Ước lượng mô hình 15
3.2.3 Kiểm tra hiệu quả của mô hình 16
3.3 Ứng dụng mô hình phân loại tuyến tính 20
3.4 Ứng dụng mô hình SVM 20

4 Kết luận **21**

A Thông tin về phiên làm việc trên R **22**

Tài liệu tham khảo **25**

DANH SÁCH BẢNG

3.1 Hệ số ước lượng 19

DANH SÁCH HÌNH VẼ

3.1 Ma trận hệ số tương quan Pearson 13
3.2 Phép chiếu bộ số liệu trên hai thành phần chính. 14
3.3 Giá trị ước lượng của các hệ số theo chiều tăng của $\log(\lambda)$ 16
3.4 Giá trị trung bình của Deviance tương ứng với mỗi giá trị tương ứng của λ 17
3.5 Confusion matrix cho mô hình logit (phương pháp Lasso) 17

CHƯƠNG 1

TỔNG QUAN VỀ QUẢN TRỊ RỦI RO TÍN DỤNG ĐỐI VỚI KHÁCH HÀNG CÁ NHÂN

1.1 MỘT SỐ KHÁI NIỆM

1.2 THỰC TRẠNG CỦA VIỆC CHẤM ĐIỂM TÍN DỤNG TẠI VIỆT NAM

1.3 KẾT LUẬN

CHƯƠNG 2

CÁC PHƯƠNG PHÁP PHÂN LOẠI KHÁCH HÀNG VAY THẺ TÍN DỤNG

2.1 CÁC MÔ HÌNH PHÂN LOẠI

2.1.1 Mô hình logit

2.1.1.1 Khái niệm

Mô hình hồi quy Logistic (hay logit) được dùng để nghiên cứu mối quan hệ giữa xác suất của các biến nhị phân hoặc phân loại và các biến giải thích khác. Hướng tiếp cận của mô hình Logistic cho bài toán phân loại là bằng cách ước lượng giá trị xác suất $P(y=1|X)$ như sau:

$$P(y=1|X) = \frac{e^{\beta_0+\sum\beta_1X_1+\beta_2X_2+...+\beta_nX_n}}{1+e^{\beta_0+\sum\beta_1X_1+\beta_2X_2+...+\beta_nX_n}}$$

Với y là biến dùng để phân loại, chỉ nhận hai giá trị 0 hoặc 1, X là các vector của biến độc lập, β_0, β_1, β_2, ..., β_n là các hệ số cần ước lượng.

Hay còn được viết dưới dạng:

$$\log\frac{P(y=1|X)}{P(y=0|X)} = \beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n$$

Trong trường hợp các biến độc lập là biến phân loại không so sánh được (ví dụ: Giới tính, dân tộc, v..v..) chúng ta đưa các biến này vào mô hình bằng cách sử dụng một nhóm các biến giả tương ứng với từng giá trị khác nhau của biến phân loại.

2.1.1.2 Ước lượng mô hình logit

a, Ước lượng hợp lý tối đa

Các hệ số β thường được ước lượng bằng phương pháp ước lượng hợp lý tối đa (Hosmer Jr, Lemeshow, and Sturdivant 2013), sử dụng hàm hợp lý có điều kiện G đối với mỗi giá trị của X. Hàm hợp lý logarit cho N quan sát được viết như sau:

$$L(\theta) = \sum_{i=1}^{N} P_{g_i}(x_i;\theta)$$

, với $p_k(x_i;\theta) = P(G = k|X = x_i;\theta)$.

Trong trường hợp biến phụ thuộc Y chỉ có 2 giá trị: $(0,1)$, ta có thể mã hóa 2 nhóm của g_i thành$y_i = 1$ khi $g_i = 1$ và $y_i = 0$ khi $g_i = 2$. Khi đó logarit của hàm hợp lý có thể viết lại như sau:

$$L(\beta) = \sum_{i=1}^{N} \{y_i log(x_i;\beta) + (1 - y_i) log(x_i;\beta)\}$$

Tối ưu hóa hàm L sẽ cho chúng ta ước lượng hợp lý tối đa cho các hệ số β trong mô hình.

b, Ràng buộc L1 hay mô hình Lasso

Một vấn đề mô hình logit hay gặp phải đó là hiện tượng đa cộng tuyến giữa các biến khi số lượng biến p tăng lên. Hậu quả của hiện tượng này là các ước lượng cho hệ số β thường là có sai số lớn, mặc dù ước lượng vẫn là không chệch. Nói cách khác, các giá trị β ước lượng được thường có hiệu quả kém khi áp dụng trên mẫu mới, mặc dù mô hình vẫn có độ chính xác cao khi áp dụng trên bộ số liệu mẫu dùng để ước lượng ra mô hình.

Để xử lý vấn đề này, chúng ta có thể áp dụng nhiều phương pháp để loại biến ra khỏi mô hình, hoặc sử dụng các phương pháp ước lượng khác mà các biến có ý nghĩa thống kê thấp bị loại ra khỏi mô hình trong quá trình ước lượng.

Phương pháp Lasso là một cải tiến của các mô hình tuyến tính, trong mô hình này, chúng ta áp dụng thêm ràng bộc L1 đối với hàm hợp lý tối đa. Áp dụng với mô hình Logit, thay vì tối ưu hàm hợp lý tối đa, chúng ta tối ưu:

$$\max_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} [y_i log(x_i;\beta) + (1-y_i) log(x_i;\beta)] - \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Sử dụng các giá trị khác nhau của tham số λ, phương pháp lasso thu nhỏ giá trị ước lượng của các β so với phương pháp tối đa hóa hàm hợp lý truyền thống. Vì các giá trị thu nhỏ của β có thể giảm về 0 và loại ra khỏi mô hình nếu λ đủ lớn, phương pháp lasso có thể được dùng để thay thế cho việc chọn biến trong các mô hình đa biến.

Đồng thời đối với lasso, ta thường không ràng buộc các hệ số chặn, và các biến giải thích phải được chuẩn hóa để ràng buộc chung cho các β là có ý nghĩa. Vì hàm tối ưu của chúng ta là hàm lõm, lời giải có thể tính sử dụng các phương pháp phi tuyến.

2.1.1.3 Diễn giải kết quả ước lượng mô hình

2.1.2 Mô hình phân loại tuyến tính

2.1.3 Mô hình SVM (Support Vector Machine)

2.2 ĐÁNH GIÁ MÔ HÌNH

2.2.1 Đường ROC và phần diện tích dưới đường cong (AUC)

2.2.2 Thang đo H

CHƯƠNG 3

TÌNH HUỐNG NGHIÊN CỨU

3.1 SỐ LIỆU VÀ CÁC BIẾN SỐ

Chúng ta thực hành trên bộ số liệu mẫu bao gồm 30000 quan sát và 25 biến bao gồm tình trạng trả nợ, các thông tin nhân khẩu học cơ bản cùng với số liệu về tín dụng và tình trạng hồ sơ của các khách hàng thẻ tín dụng ở Đài Loan từ tháng 4 năm 2005 đến tháng 9 năm 2005.

Các tên biến đã được thay đổi để tiện lợi cho việc đọc hiểu và phân tích, cụ thể như sau:

`ID` Số ID của mỗi khách hàng tín dụng

`LIMIT_BAL` Lượng tín dụng cho vay tính bằng Đô la Đài Loan (bao gồm cả các khoản vay cá nhân và các khoản vay với thẻ tín dụng phụ)

`SEX` Giới tính (0=Nữ, 1=Nam)

`EDUCATION` (1=sau đại học, 2=đại học, 3=phổ thông, 4=khác)

`MARRIAGE` Trạng thái hôn nhân (1=đã cưới, 2=độc thân, 3=khác)

`AGE` Số tuổi tính bằng năm

`PAY_0` Tình trạng hồ sơ vào thời điểm tháng 9/2005 (-1=trả đúng hạn, 1=chậm 1 tháng, 2=chậm 2 tháng, ... 8=chậm 8 tháng, 9=chậm 9 tháng hoặc nhiều hơn)

`PAY_2` Tình trạng hồ sơ vào thời điểm tháng 8/2005 (thang điểm như trên)

`PAY_3` Tình trạng hồ sơ vào thời điểm tháng 7/2005 (thang điểm như trên)

`PAY_4` Tình trạng hồ sơ vào thời điểm tháng 6/2005 (thang điểm như trên)

`PAY_5` Tình trạng hồ sơ vào thời điểm tháng 5/2005 (thang điểm như trên)

`PAY_6` Tình trạng hồ sơ vào thời điểm tháng 4/2005 (thang điểm như trên)

`BILL_AMT1` Hóa đơn thanh toán vào thời điểm 9/2005 (Đô la Đài Loan)

`BILL_AMT2` Hóa đơn thanh toán vào thời điểm 8/2005 (Đô la Đài Loan)

`BILL_AMT3` Hóa đơn thanh toán vào thời điểm 7/2005 (Đô la Đài Loan)

`BILL_AMT4` Hóa đơn thanh toán vào thời điểm 6/2005 (Đô la Đài Loan)

`BILL_AMT5` Hóa đơn thanh toán vào thời điểm 5/2005 (Đô la Đài Loan)

`BILL_AMT6` Hóa đơn thanh toán vào thời điểm 4/2005 (Đô la Đài Loan)

`PAY_AMT1` Lượng tiền đã thanh toán vào thời điểm tháng 9/2015 (Đô la Đài Loan)

`PAY_AMT2` Lượng tiền đã thanh toán vào thời điểm tháng 8/2015 (Đô la Đài Loan)

`PAY_AMT3` Lượng tiền đã thanh toán vào thời điểm tháng 7/2015 (Đô la Đài Loan)

`PAY_AMT4` Lượng tiền đã thanh toán vào thời điểm tháng 6/2015 (Đô la Đài Loan)

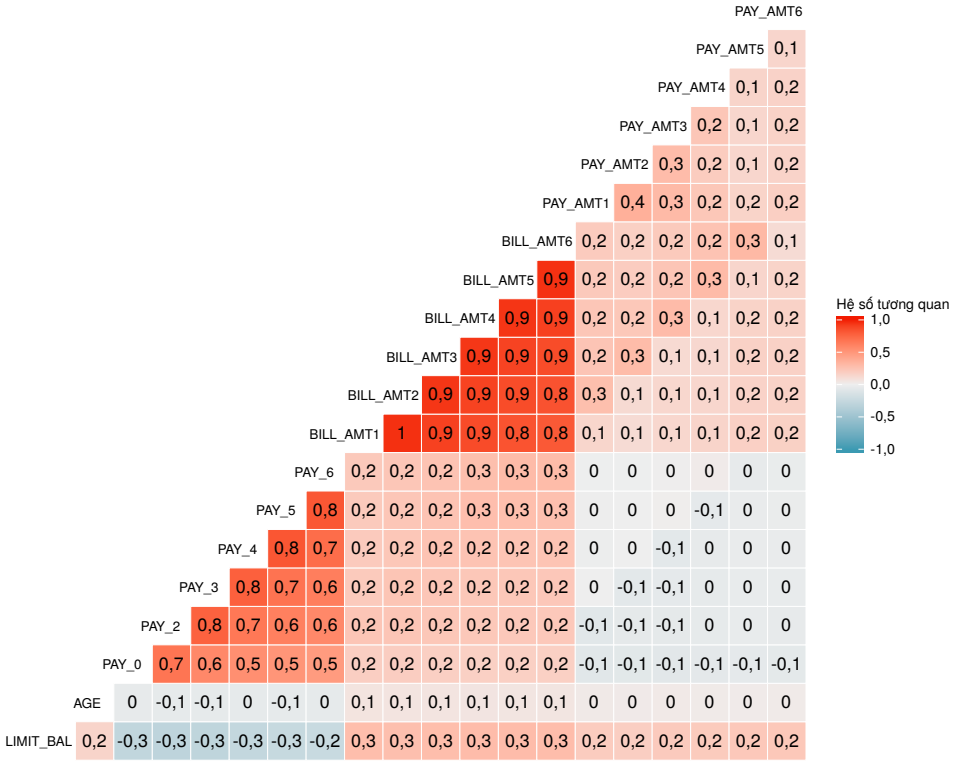
`PAY_AMT5` Lượng tiền đã thanh toán vào thời điểm tháng 5/2015 (Đô la Đài Loan)

`PAY_AMT6` Lượng tiền đã thanh toán vào thời điểm tháng 4/2015 (Đô la Đài Loan)

`DEFAULT` Có trả nợ hay không (1=có, 0=không)

Hình 3.1 (trang 13) mô tả ma trận hệ số tương quan Pearson giữa các biến số trong bộ số liệu. Lưu ý tương quan giữa các biến trong nhóm biến `PAY` (tình trạng hồ sơ) và giữa các biến trong nhóm biến `BILL_AMT` (hoá đơn thanh toán) là khá cao, thể hiện sự tương đồng cao về mặt thông tin thể hiện của các biến này. Trong số các biến trong bộ số liệu, các biến `PAY` là có thể hiện tương quan dương với biến `DEFAULT`, gợi ý rằng chúng ta có thể sử dụng biến này là biến chính để dự đoán tỉ lệ vỡ nợ của khách hàng.

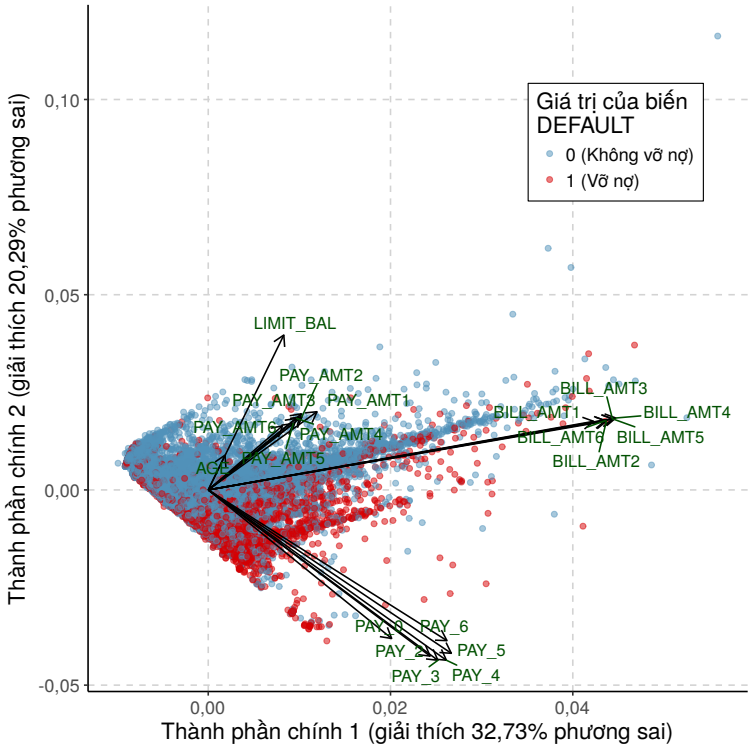
Để có cái nhìn cụ thể hơn vào bộ số liệu này, chúng ta sử dụng phương pháp phân tích thành phần chính (PCA - Principal Component Analysis) để phân tích bộ số liệu. Với phương pháp này, chúng ta tìm một hệ tọa độ trực giao mới để thể hiện bộ số liệu, sao cho với thành phần chính thứ nhất (chiều thứ nhất của hệ tọa độ mới) thể hiện được nhiều nhất có thể thông tin của bộ số liệu, thành phần chính



Hình 3.1: Ma trận hệ số tương quan Pearson giữa các biến trong bộ số liệu.

thứ hai (chiều thứ hai của hệ tọa độ mới) thể hiện nhiều nhất có thể lượng thông tin còn lại của bộ số liệu, v...v... Lưu ý rằng vì các biến trong bộ số liệu có thang đo khác nhau, để đảm bảo hiệu quả cho phương pháp phân tích đa biến này, chúng ta chuẩn hóa các biến trước khi thực hiện PCA. Đồng thời, các biến phân loại như `EDUCATION` và `MARRIAGE` cũng được lược bỏ.

Phép chiếu của các biến và các quan sát trong bộ số liệu trên hai thành phần chính đầu tiên được thể hiện trong hình 3.2 (trang 14), với mỗi véc tơ thể hiện một biến và mỗi điểm thể hiện một quan sát trong bộ số liệu. Các quan sát thuộc vào nhóm vỡ nợ (biến `DEFAULT` bằng 1) có màu đỏ và các quan sát thuộc nhóm không vỡ nợ (biến `DEFAULT` bằng 0) có màu xanh. Quan sát đồ thị này, chúng ta nhận thấy các quan sát thuộc nhóm vỡ nợ (màu đỏ) tập trung nhiều ở phía dưới đồ thị, hay là giá trị của các biến này chiếu trên thành phần chính thứ 2 (trục tung) là thấp hơn. Như chúng ta nhận xét ở ma trận hệ số tương quan phía trên, véc tơ chiếu các biến thuộc cùng nhóm `PAY`, `PAY_ATM` và `BILL_ATM` nằm khá gần nhau, thể hiện mức độ tương quan cao giữa các biến số thuộc cùng một trong ba nhóm này. Các biến thuộc nhóm `PAY` có hướng trùng với hướng phân bố của các quan sát thuộc nhóm vỡ nợ, trong khi các biến thuộc nhóm `PAY_ATM` có hướng trùng với hướng phân bố



Hình 3.2: Phép chiếu bộ số liệu trên hai thành phần chính.

của các quan sát thuộc nhóm không vỡ nợ, gợi ý tiềm năng dùng để dự báo của các nhóm biến này. Ngoài ra các quan sát nhóm vỡ nợ cũng có xu hướng thể hiện cao trên biến `SEX`. Nhóm các quan sát không vỡ nợ cũng phân bố nhiều theo chiều tăng của các biến `AGE` và `LIMIT_BAL`.

Lưu ý rằng đồ thị 3.2 chỉ thể hiện 53,02 phần trăm lượng thông tin của bộ số liệu, chưa kể các biến phân loại như `EDUCATION` hay `MARRIAGE`. Bằng cách sử dụng các phương pháp phân tích cụ thể hơn, chúng ta có thể đưa ra một mô hình phân loại chính xác hơn đối với khả năng vỡ nợ của các khách hàng dùng thẻ tín dụng trong bộ số liệu này.

Trong 30000 quan sát của bộ số liệu này, sau khi làm sạch, chúng ta chọn ngẫu nhiên 75% (Khoảng hơn 22000 quan sát) để xây dựng các mô hình. Số lượng quan sát còn lại sẽ được dùng để kiểm tra các mô hình xây dựng được và cho chúng ta cái nhìn về hiệu quả của chúng.

3.2 ỨNG DỤNG MÔ HÌNH LOGIT

3.2.1 Tiền xử lý bộ số liệu

Mô hình logit giả định xác suất dự báo $P(y = 1|X)$ phân phối chuẩn với trung bình là 0,5 và thực hiện ước lượng có hiệu quả hơn trên bộ số liệu có tỉ lệ biến phụ thiộc là 0,5. Tuy nhiên trong bộ số liệu chúng ta nghiên cứu này, tỷ lệ $\texttt{DEFAULT} = 1$ là 22.34%. Để đảm bảo cho hiệu quả của mô hình logit, trong bộ số liệu dùng để ước lượng mô hình, chúng ta lấy tập con của số mẫu thuộc nhóm $\texttt{DEFAULT} = 1$ một cách ngẫu nhiên, sao cho tỷ lệ $P(\texttt{DEFAULT} = 1)$ trong bộ số liệu ước lượng bây giờ là 0,5.

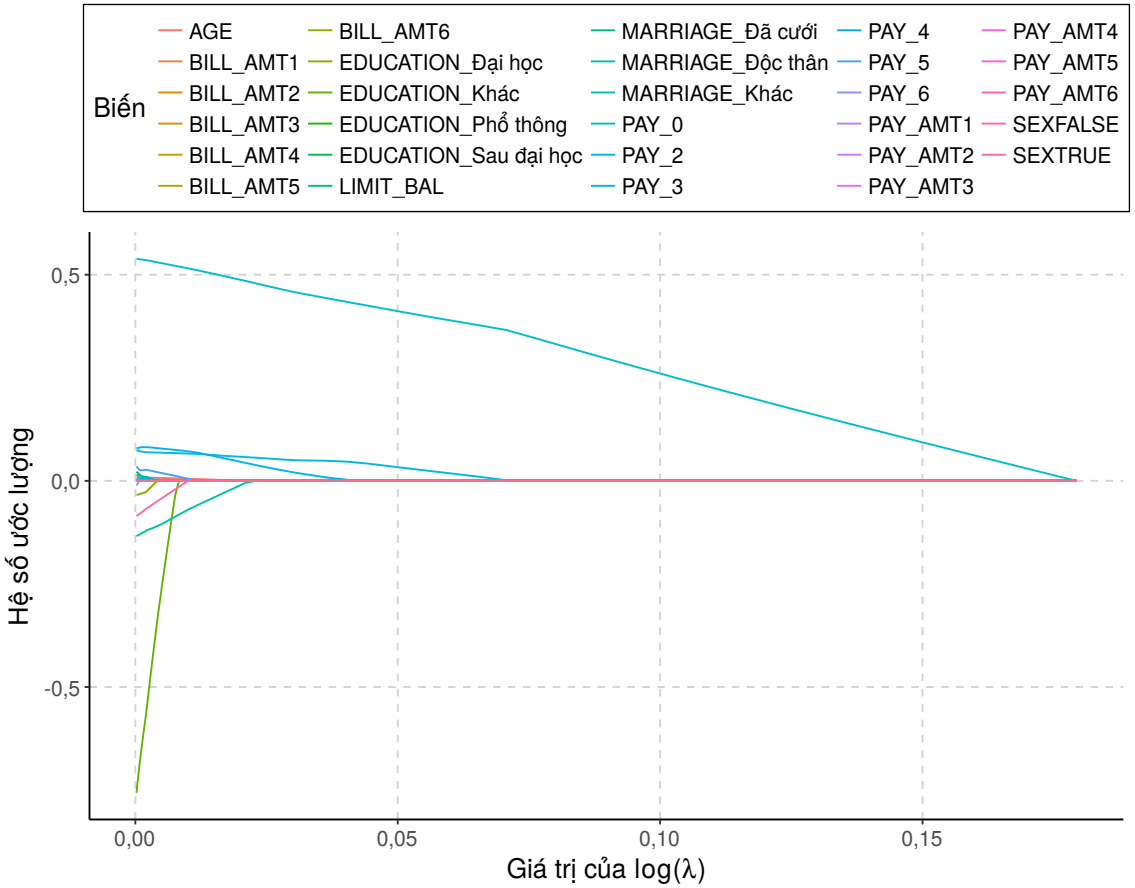
3.2.2 Ước lượng mô hình

Chúng ta thực hiện mô hình logit, sử dụng phương pháp Lasso để giới hạn giá trị của các hệ số ước lượng. Với mỗi gía trị của tham số λ, giá trị của các hệ số ước lượng β càng bị ràng buộc chặt, các hệ số ước lượng có giá trị bằng 0 có thể coi là bị loại khỏi mô hình.

Hình 3.3 mô tả xu hướng của các hệ số ước lượng β khi giá trị của $\log\lambda$ thay đổi. Với hướng tăng của $\log\lambda$ các biến thuộc nhóm `BILL` nhanh chóng hội tụ về 0, thể hiện mức ý nghĩa thống kê thấp của các hệ số này trong mô hình. Trong khi đó, các biến thuộc nhóm `PAY` chậm hội tụ về 0 hơn, với biến `PAY_0` là biến cuối cùng có hệ số ước lượng β tương ứng hội tụ về 0. Với các giá trị λ lớn hơn từ sau thời điểm này, có thể nói mô hình chỉ còn hệ số chặn β_0.

Để xác định giá trị hợp lý cho tham số λ trong mô hình này. Chúng ta thực hiện bằng cách chia bộ số liệu thử nghiệm thành 10 phần nhỏ, chạy mô hình logit sử dụng phương pháp Lasso này trên từng bộ số liệu con này, rồi kiểm định kết quả mô hình dựa trên các bộ số liệu còn lại. Từ kết quả của 10 phép ước lượng kiểm định chéo này, chúng ta có thể tính được giá trị trung bình của Deviance tại các giá trị của λ, qua đó chúng ta có thể xác định được giá trị của λ mà Deviance là thấp nhất.

Hình 3.4 mô tả mối quan hệ của Deviance (trên trục tung) khi giá trị của λ thay đổi (giá trị của λ trên trục hoành đã được logarit hóa). Các chấm màu đỏ biểu diễn giá trị Deviance trung bình của kiểm định chéo tại các giá trị khác nhau của $\log(\lambda)$, thanh màu đen thể hiện sai số chuẩn tương ứng của giá trị trung bình này. Đường kẻ dọc màu xanh đậm đánh dấu giá trị của $\log(\lambda)$ mà tại đó gía trị của Deviance là



Hình 3.3: Giá trị ước lượng của các hệ số theo chiều tăng của $\log(\lambda)$

thấp nhất($\lambda \approx 0{,}001421$), tương đương với giá trị của λ mà chúng ta lựa chọn cho mô hình này.

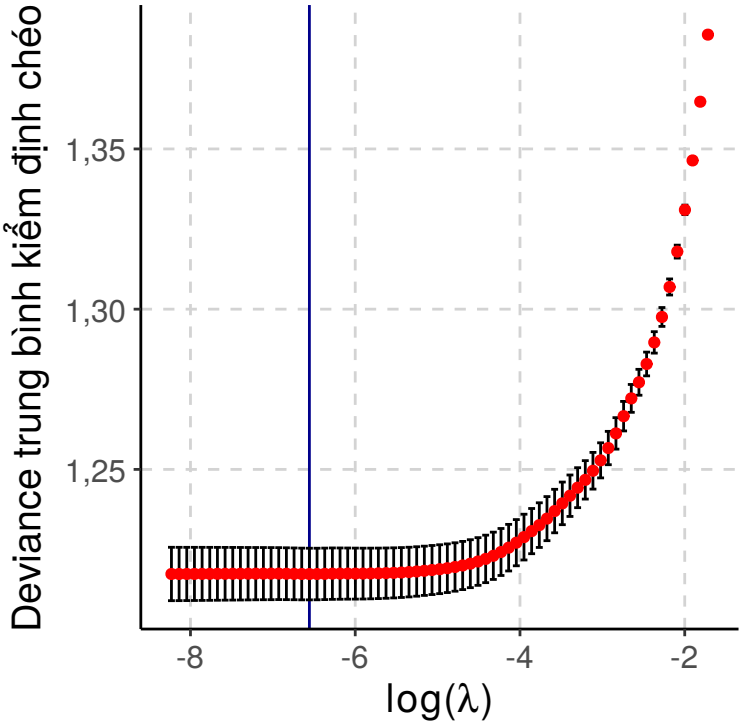
Hệ số ước lượng β của mô hình được thể hiện trong bảng 3.1).

3.2.3 Kiểm tra hiệu quả của mô hình

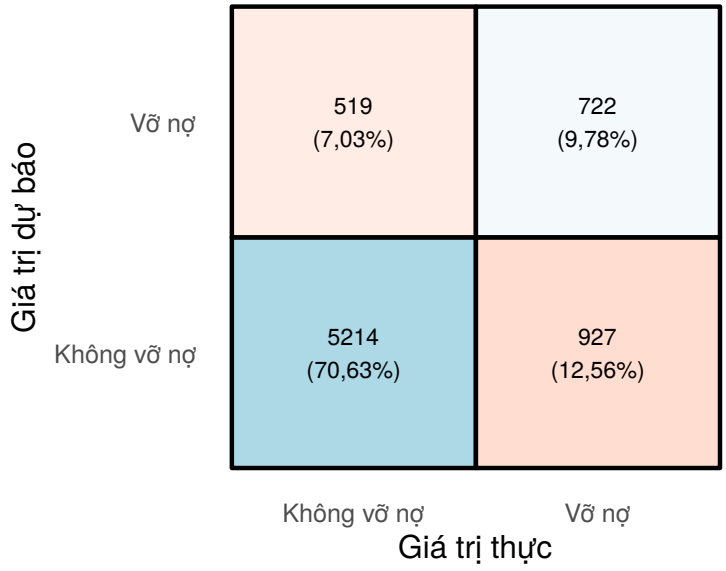
Chúng ta sử dụng mô hình ước lượng được để dự báo trên tập số liệu kiểm tra. Hình 3.5 minh họa một ma trận thể hiện độ chính xác của mô hình khi áp dụng lên tập số liệu kiểm tra, với màu xanh thể hiện các quan sát được dự đoán đúng và màu đỏ thể hiện các quan sát được dự đoán sai.

Trong tổng cộng 7382 dòng của bộ số liệu dùng để kiểm tra, có 1649 trường hợp vỡ nợ, trong đó mô hình dự đoán đúng 722 trường hợp. Đồng thời, trong các quan sát thuộc nhóm không vỡ nợ còn lại, có 519 quan sát bị mô hình đánh giá nhầm là có vỡ nợ.

Nhìn chung tỉ lệ dự đoán chính xác của mô hình trên tập số liệu kiểm tra là 80,41%, tỷ lệ của sai lầm loại I (không vỡ nợ nhưng mô hình dự đoán là có) là



Hình 3.4: Giá trị trung bình của Deviance tương ứng với mỗi giá trị tương ứng của λ



Hình 3.5: Confusion matrix cho mô hình logit (phương pháp Lasso)

7,03%, tỷ lệ của sai lầm loại II (có vỡ nợ nhưng dự đoán là không vỡ nợ) là 12,56%.

	Tên biến	Hệ số
1	Hệ số chặn	0,0488178
2	LIMIT_BAL	-7e-07
3	SEXFALSE	-0,0741289
4	EDUCATION_Sau đại học	0,0072849
5	EDUCATION_Đại học	-0,0296278
6	EDUCATION_Khác	-0,6249096
7	MARRIAGE_Đã cưới	0,0109769
8	MARRIAGE_Độc thân	-0,1256848
9	AGE	0,008362
10	PAY_0	0,5365426
11	PAY_2	0,0705397
12	PAY_3	0,08174
13	PAY_4	0,0045188
14	PAY_5	0,0259231
15	BILL_AMT1	-2e-06
16	BILL_AMT4	7e-07
17	PAY_AMT1	-1,5e-05
18	PAY_AMT2	-5,7e-06
19	PAY_AMT3	-4,1e-06
20	PAY_AMT4	-4,4e-06
21	PAY_AMT5	-1,2e-06
22	PAY_AMT6	-3,6e-06

Bảng 3.1: Hệ số ước lượng

3.3 ỨNG DỤNG MÔ HÌNH PHÂN LOẠI TUYẾN TÍNH

3.4 ỨNG DỤNG MÔ HÌNH SVM

CHƯƠNG 4

KẾT LUẬN

PHỤ LỤC A

THÔNG TIN VỀ PHIÊN LÀM VIỆC TRÊN R

```
## R version 3.3.3 (2017-03-06)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 16.04.2 LTS
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8    LC_NUMERIC=C
##  [3] LC_TIME=vi_VN           LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=vi_VN       LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=vi_VN          LC_NAME=C
##  [9] LC_ADDRESS=C            LC_TELEPHONE=C
## [11] LC_MEASUREMENT=vi_VN    LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets
## [6] methods   base
##
## other attached packages:
##  [1] glmnet_2.0-5    foreach_1.4.3   Matrix_1.2-8
##  [4] ggfortify_0.4.1 GGally_1.3.0    caret_6.0-73
##  [7] ggplot2_2.2.1   lattice_0.20-35 broom_0.4.2
## [10] dplyr_0.5.0     tidyr_0.6.1     readr_1.1.0
## [13] knitr_1.15.1
##
## loaded via a namespace (and not attached):
##  [1] latex2exp_0.4.0    Rcpp_0.12.10
##  [3] RColorBrewer_1.1-2 nloptr_1.0.4
##  [5] plyr_1.8.4         iterators_1.0.8
```

```
##  [7] tools_3.3.3         digest_0.6.12
##  [9] lme4_1.1-12         evaluate_0.10
## [11] tibble_1.3.0        gtable_0.2.0
## [13] nlme_3.1-131        mgcv_1.8-16
## [15] psych_1.7.3.21      DBI_0.6-1
## [17] ggrepel_0.6.5       parallel_3.3.3
## [19] SparseM_1.76        gridExtra_2.2.1
## [21] stringr_1.2.0       MatrixModels_0.4-1
## [23] hms_0.3             stats4_3.3.3
## [25] grid_3.3.3          nnet_7.3-12
## [27] reshape_0.8.6       R6_2.2.0
## [29] foreign_0.8-67      minqa_1.2.4
## [31] reshape2_1.4.2      car_2.1-4
## [33] magrittr_1.5        scales_0.4.1
## [35] codetools_0.2-15    ModelMetrics_1.1.0
## [37] MASS_7.3-45         splines_3.3.3
## [39] assertthat_0.1      pbkrtest_0.4-7
## [41] mnormt_1.5-5        xtable_1.8-2
## [43] colorspace_1.3-2    labeling_0.3
## [45] quantreg_5.29       stringi_1.1.5
## [47] lazyeval_0.2.0      munsell_0.4.3
```

TÀI LIỆU THAM KHẢO

1. R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

2. Wickham, Hadley (2017). *tidyverse: Easily Install and Load 'Tidyverse' Packages*. R package version 1.1.1. URL: https://CRAN.R-project.org/package=tidyverse.

3. Jed Wing, Max Kuhn. Contributions from et al. (2016). *caret: Classification and Regression Training*. R package version 6.0-73. URL: https://CRAN.R-project.org/package=caret.

4. Lessmann, Stefan et al. (2015). "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research". In: *European Journal of Operational Research* 247.1, pp. 124–136.

5. Xie, Yihui (2015). *Dynamic Documents with R and knitr*. 2nd. ISBN 978-1498716963. Boca Raton, Florida: Chapman and Hall/CRC. URL: http://yihui.name/knitr/.

6. Hosmer Jr, David W, Stanley Lemeshow, and Rodney X Sturdivant (2013). *Applied logistic regression*. Vol. 398. John Wiley & Sons.

7. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1, pp. 1–22. URL: http://www.jstatsoft.org/v33/i01/.

8. Thomas, Lyn C (2010). "Consumer finance: Challenges for operational research". In: *Journal of the Operational Research Society* 61.1, pp. 41–52.

9. Xiao, Wenbing, Qian Zhao, and Qi Fei (2006). “A comparative study of data mining methods in consumer loans credit scoring management”. In: *Journal of Systems Science and Systems Engineering* 15.4, pp. 419–435.

10. Baesens, Bart et al. (2003). “Benchmarking state-of-the-art classification algorithms for credit scoring”. In: *Journal of the operational research society* 54.6, pp. 627–635.