

TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN  
KHOA TOÁN KINH TẾ



## CHUYÊN ĐỀ THỰC TẬP

<b>Chuyên ngành:</b>	Toán Kinh tế
<b>Đề tài:</b>	Ứng dụng, đánh giá, và so sánh một số mô hình phân loại trong phân loại khách hàng thẻ tín dụng
<b>Sinh viên thực hiện:</b>	Nguyễn Đức Hiếu
<b>Mã sinh viên:</b>	11131371
<b>Lớp:</b>	Toán Kinh tế 55
<b>Giảng viên hướng dẫn:</b>	PGS. Nguyễn Thị Minh

---

Hà Nội, Ngày 3 tháng 5 năm 2017

## MỤC LỤC

<b>Mục lục</b>	<b>2</b>
<b>Danh sách bảng</b>	<b>3</b>
<b>Danh sách hình</b>	<b>4</b>
<b>Lời mở đầu</b>	<b>5</b>
<b>CHƯƠNG I: Tổng quan về quản trị rủi ro tín dụng đối với khách hàng cá nhân.</b>	<b>7</b>
1.1 Tổng quan về rủi ro tín dụng và vấn đề quản trị rủi ro trong ngân hàng	7
1.1.1 Rủi ro và các khái niệm liên quan trong ngân hàng	7
1.1.1.1 Tín dụng	7
1.1.1.2 Rủi ro tín dụng trong ngân hàng	7
1.1.2 Quản trị rủi ro trong ngân hàng	7
1.1.2.1 Hoạt động quản trị rủi ro trong ngân hàng	7
1.1.2.2 Quản trị rủi ro tín dụng trong lịch sử	7
1.1.2.3	7
1.2 Thực trạng của việc chấm điểm tín dụng tại Việt Nam	7
1.3 Kết luận	7
<b>CHƯƠNG II: Các Mô hình phân loại khách hàng vay thẻ tín dụng</b>	<b>8</b>
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 SVM (Support Vector Machine)	10

<b>CHƯƠNG III: Tình huống nghiên cứu.</b>	<b>11</b>
3.1 Số liệu và các biến số	11
3.2 Ứng dụng mô hình logit	15
3.3 Ứng dụng mô hình SVM	15
<b>CHƯƠNG IV: Kết luận.</b>	<b>16</b>
<b>Tài liệu tham khảo.</b>	<b>18</b>

## **DANH SÁCH BẢNG**

## 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

## LỜI MỞ ĐẦU

Đối với các ngân hàng việc chấm điểm tín dụng và phân loại các khách hàng là một trong những khâu thiết yếu cho quy trình quản trị rủi ro 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  $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$  kết hợp với Sweave và knitr (Xie 2015). Tất cả phân tích được thực hiện trên phần mềm thống kê R version 3.4.0 (2017-04-21) (R Core Team 2017), các phân tích cụ thể được thực hiện sử dụng các gói mở rộng caret(Jed Wing et al. 2016), tidyverse (Wickham 2017)... Mô hình logit được thực hiện với gói glmnet(Friedman, Hastie, and Tibshirani 2010). Mô hình SVM được thực hiện với gói kernlab(Karatzoglou et al. 2004), là một giao diện của phần mềm LIBSVM (Chang and Lin 2011) trong môi trường R.

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

## **CHƯƠNG I**

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

### **1.1 TỔNG QUAN VỀ RỦI RO TÍN DỤNG VÀ VẤN ĐỀ QUẢN TRỊ RỦI RO TRONG NGÂN HÀNG**

#### **1.1.1 Rủi ro và các khái niệm liên quan trong ngân hàng**

##### **1.1.1.1 Tín dụng**

##### **1.1.1.2 Rủi ro tín dụng trong ngân hàng**

#### **1.1.2 Quản trị rủi ro trong ngân hàng**

##### **1.1.2.1 Hoạt động quản trị rủi ro trong ngân hàng**

##### **1.1.2.2 Quản trị rủi ro tín dụng trong lịch sử**

##### **1.1.2.3**

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

### **1.3 KẾT LUẬN**



## CHƯƠNG II

# CÁC MÔ HÌNH 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:

$$P(y = 1|X) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n}}$$

Với  $y$  là biến dùng để phân loại, chỉ nhận hai giá trị 0 hoặc 1,  $X$  là các vector của biến độc lập,  $\beta_0, \beta_1, \beta_2, \dots, \beta_n$  là các hệ số cần ước lượng.

Hay còn được viết dưới dạng:

$$\log \frac{P(y = 1|X)}{P(y = 0|X)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

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  $\beta$  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  $\beta$  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  $\beta$  thường là có sai số lớn, mặc dù ước lượng vẫn là không chệch. Nói cách khác, các giá trị  $\beta$  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 buộc 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  $\lambda$ , phương pháp lasso thu nhỏ giá trị ước lượng của các  $\beta$  so với phương pháp tối đa hóa hàm hợp lý truyền thống. Vì các giá trị thu nhỏ của  $\beta$  có thể giảm về 0 và loại ra khỏi mô hình nếu  $\lambda$  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  $\beta$  là có ý nghĩa. Vì hàm tối ưu của chúng ta là hàm lồi, 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 SVM (Support Vector Machine)

## CHƯƠNG III

### 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)

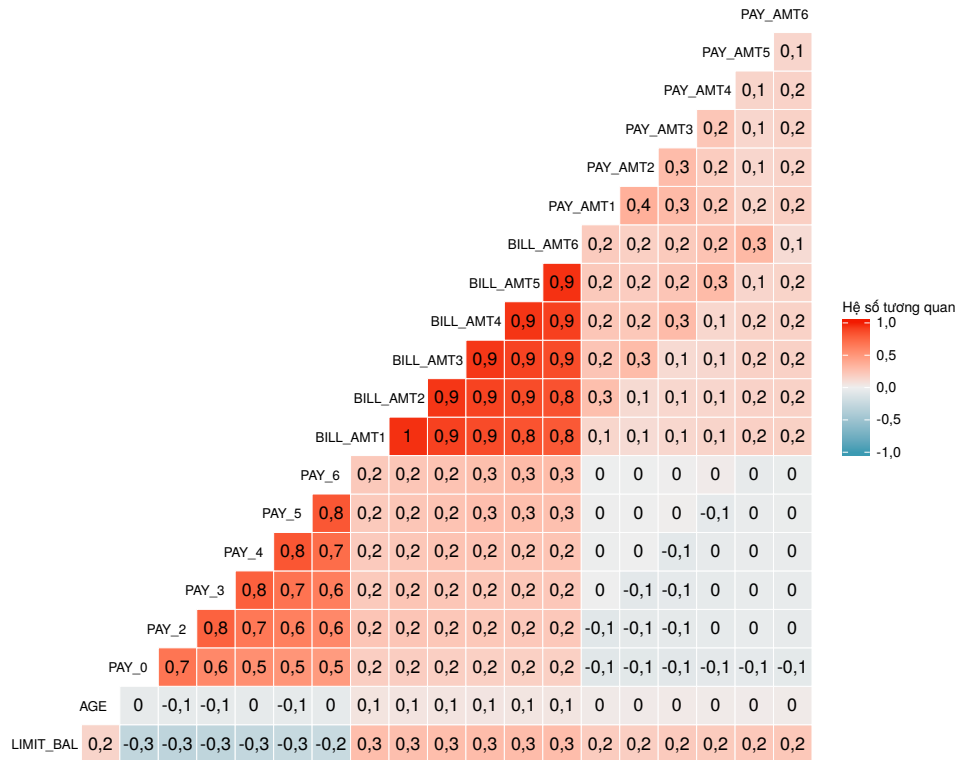
PAY\_AMT5 Lượng tiền đã thanh toán vào thời điểm tháng 5/2015 (Đô la Đài Loan)

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.

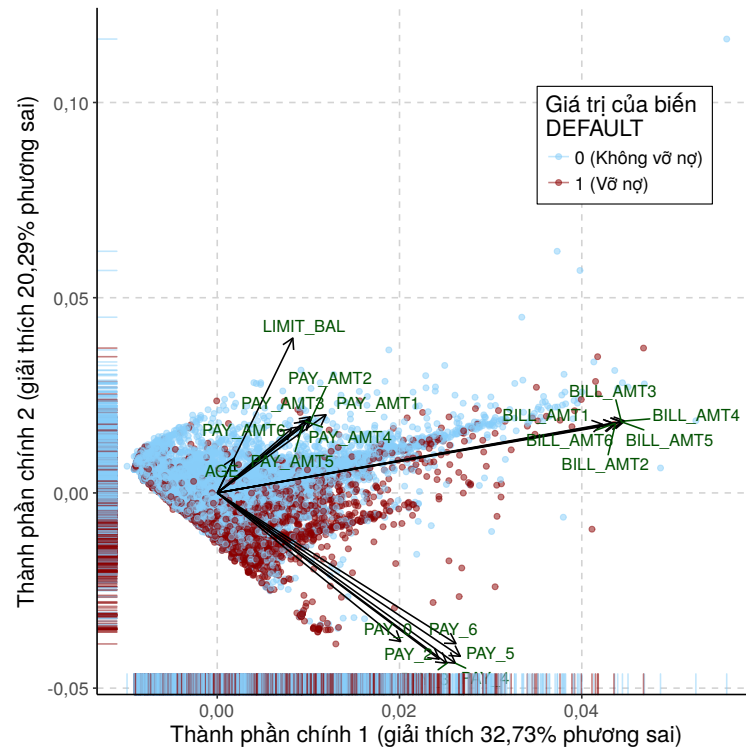
Để có cái nhìn cụ thể hơn vào bộ số liệu này, chúng ta sử dụng phương pháp phân tích thành phần chính (PCA - Principal Component Analysis) để phân tích bộ số liệu. Với phương pháp này, chúng ta tìm một hệ tọa độ trực giao mới để thể hiện bộ số liệu, sao cho với thành phần chính thứ nhất (chiều thứ nhất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

### 3.3 ỨNG DỤNG MÔ HÌNH SVM

Trong xây dựng một mô hình SVM, việc lựa chọn kernel thích hợp và việc điều chỉnh để có được tham số thích hợp cho kernel đó là yếu tố vô cùng quan trọng. “[A practical guide to support vector classification](#)” đưa ra một quy trình để có thể đạt được kết quả chấp nhận được đối với hầu hết tất cả mọi trường hợp. Quy trình đó như sau:

- Biến đổi số liệu phù hợp với gói ứng dụng SVM
- Thực hiện chuẩn hoá bộ số liệu
- Sử dụng kernel RBF  $K(x, y) = e^{-\gamma \|x - y\|^2}$



## **CHƯƠNG IV**

### **KẾT LUẬN**

## 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. Chang, Chih-Chung and Chih-Jen Lin (2011). “LIBSVM: A library for support vector machines”. In: *ACM Transactions on Intelligent Systems and Technology* 2 (3). Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, 27:1–27:27.

8. 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/>.
9. Thomas, Lyn C (2010). “Consumer finance: Challenges for operational research”. In: *Journal of the Operational Research Society* 61.1, pp. 41–52.
10. 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.
11. Karatzoglou, Alexandros et al. (2004). “kernlab – An S4 Package for Kernel Methods in R”. In: *Journal of Statistical Software* 11.9, pp. 1–20. URL: <http://www.jstatsoft.org/v11/i09/>.
12. 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.
13. Hsu, Chih-Wei, Chih-Chung Chang, Chih-Jen Lin, et al. “A practical guide to support vector classification”. In: