PHƯƠNG PHÁP HỌC MÁY TRONG PHÁT HIỆN GIAN LẬN THẺ TÍN DỤNG - MỘT NGHIÊN CỨU THỰC NGHIỆM

Nguyễn Thị Liên
Đại học Kinh tế Quốc dân
Email: Lientkt@neu.edu.vn
Nguyễn Thị Thu Trang
Đại học Kinh tế Quốc dân
Email: Thutrang21@gmail.com
Nguyễn Chiến Thắng
Ngân hàng TMCP Việt Nam Thịnh Vượng
Email: Thangnc9@outlook.com

Ngày nhận: 27/8/2018
Ngày nhận bản sửa: 02/10/2018
Ngày duyệt đăng: 15/10/2018

Tóm tắt:
Nghiên cứu giới thiệu các phương pháp thống kê và học máy để phát hiện gian lận thẻ tín dụng tại ngân hàng thương mại. Bằng việc sử dụng 284807 giao dịch thẻ tín dụng của châu Âu trong tháng 09/2013, nghiên cứu ứng dụng các mô hình được sử dụng trong thực tế hiện nay như mô hình Logistic, mạng Bayesian (Bayesian Network), cây quyết định (Decision trees), phương pháp Stacking (Stacked generalization). Ngoài ra, nghiên cứu cũng đưa ra một số cách xử lý trong trường hợp dữ liệu mất cân bằng. Thông qua kết quả so sánh các mô hình và xử lý dữ liệu mất cân bằng, các ngân hàng thương mại ở Việt Nam có thể lựa chọn ứng dụng để kiểm soát phát hiện gian lận thẻ tín dụng.
Từ khóa: Gian lận thẻ tín dụng, phát hiện gian lận, học máy.

Machine Learning Method in Credit Card Fraud Detection – An Experimental Research

Abstract:
This study aims to introduce the application of machine learning techniques to credit card fraud detection. Using the credit card data which includes 284807 observations of Europe in September 2013, the research conducted and compared methods from currently used in practice as Logistic regression, Bayes network, Decision trees. The research also deals with the case of using unbalanced data. Through the comparison of models, commercial banks in Vietnam can find appropriate methods and control the process of credit card fraud detection.
Keywords: Credit card fraud detection, fraud detection, machine learning.

1. Đặt vấn đề

Hành vi gian lận là những hành vi cố ý làm sai lệch thông tin do một tổ chức, cá nhân hoặc bên thứ ba thực hiện. Đó là hành vi không hợp pháp nhằm chủ ý lừa gạt, đưa các thông tin không chính xác để thu được lợi ích bất hợp pháp. Gian lận phát sinh khi hội tụ các yếu tố: Cố ý trình bày sai một yếu tố hay sự kiện quan trọng, kết quả trình bày sai làm cho người bị hại tin vào kết quả đó, người bị hại dựa vào kết quả trình bày sai để ra các quyết định, gây ra khoản

lỗ, thiệt hại về tiền, tài sản.

Hiện nay, có nhiều cách phân loại hành vi gian lận khác nhau. Xét theo chủ thể thực hiện, hành vi gian lận có thể đến từ đối tượng là cán bộ, nhân viên ngân hàng lợi dụng chức vụ, quyền hạn để tạo dựng hồ sơ, giấy tờ giả, sổ tiết kiệm khống, giả mạo chữ ký của khách hàng gửi tiền nhằm tham ô, sử dụng bút toán giả, thu tiền nợ vay không nhập quỹ, lập hồ sơ vay khống hoặc hồ sơ ghi tăng số tiền vay để rút tiền, không thẩm định hoặc cố tình thẩm định sai tài sản thế chấp. Hành vi gian lận có thể đến từ đối tượng là khách hàng của ngân hàng, thực hiện hành vi lừa đảo, tự tạo dựng hồ sơ dự án, giả mạo hợp đồng, lập hồ sơ vay, thế chấp tài sản, phương án kinh doanh, phương án trả nợ giả; hoặc nâng giá trị tài sản thế chấp tăng lên nhiều lần, tài sản thế chấp không đủ giấy tờ pháp lý hoặc đang có tranh chấp, lập dự án khống có thật, đột nhập mạng ngân hàng, trộm cắp mật khẩu, tạo lệnh chuyển tiền giả nhằm chiếm đoạt tài sản. Mặt khác, hành vi cấu kết giữa cán bộ ngân hàng và đối tượng bên ngoài như hối lộ nhân viên ngân hàng để tạo điều kiện cho việc chiếm đoạt tài sản, rửa tiền, sử dụng công nghệ cao để thực hiện hành vi phạm pháp trộm cắp thông tin thẻ ngân hàng. Hành vi cấu kết giữa cán bộ ngân hàng và đối tượng bên ngoài.

Xét theo lĩnh vực thực hiện, gian lận có thể chia thành một số loại như gian lận thẻ tín dụng (Credit Card Fraud), gian lận công nghệ cao (Telecommunication Fraud), xâm nhập máy tính (Computer Intrusion), gian lận phá sản (Bankruptcy Fraud), trộm cắp thẻ, thẻ giả hoặc rửa tiền.

Những vụ gian lận có thể diễn ra ở thời điểm bất kỳ với quy mô và số lượng không thể biết trước, hậu quả tiềm ẩn những nguy cơ rủi ro cho ngân hàng. Dự báo trước hiện tượng gian lận không chỉ áp dụng với những khách hàng của ngân hàng, mà còn áp dụng với cả các nhân viên trong ngân hàng, giảm thiểu rủi ro từ việc nhân viên cố tình giả mạo hồ sơ hay cho vay quá hạn mức để đạt được doanh số. Hoạt động các ngân hàng liên tục cải tiến quy trình quản lý rủi ro và phát hiện gian lận hiệu quả hơn bằng cách kiểm duyệt xen lẫn giữa lấy ý kiến của chuyên gia và chấm điểm dựa trên mô hình, đem lại kết quả tốt hơn, giảm thiểu nhiều chi phí.

Trong lĩnh vực quản lý thẻ tín dụng, tình trạng gian lận càng ngày càng gia tăng. Theo những công bố báo cáo năm 2015 của tổ chức phát hành thẻ tín dụng Visa và Mastercard, tỷ lệ gian lận ở Việt Nam vẫn còn thấp hơn so với các nước khác trên thế giới. Năm 2015, tỷ lệ số tiền bị mất cắp do gian lận trên toàn thế giới là 0.07%, tương đương 21 tỷ USD, trong đó tỷ lệ gian lận ở Việt Nam vẫn khá thấp – 0,023%. Tuy nhiên, với sự phát triển nhanh chóng của Việt Nam, các mối nguy cơ tiềm ẩn về gian lận trong giao dịch tín dụng vẫn còn rất lớn.

2. Tổng quan nghiên cứu

Phát hiện gian lận thẻ tín dụng là một nhiệm vụ khó khăn khi sử dụng thủ tục thông thường. Với sự phát triển của hệ thống các dịch vụ cung cấp của ngân hàng, các mô hình phát hiện gian lận thẻ tín dụng đã trở nên có ý nghĩa trong cả lý thuyết và thực tiễn. Các loại dữ liệu thống kê về gian lận là dữ liệu dạng số. Cũng như các dữ liệu dạng phân loại (classification) khác, các biến sử dụng trong phát hiện gian lận cũng có các biến định tính hoặc định lượng, biến phụ thuộc là biến phân loại nhị phân với mục tiêu phân biệt trạng thái gian lận hoặc không gian lận (mã hóa thành Fraud - Legal hoặc 1-0).

Không giống với bình thường, các dữ liệu về gian lận bị mất cân bằng, thường có trên 99% quan sát là không gian lận và dưới 1% quan sát gian lận cần tìm ra. Các bộ dữ liệu được các ngân hàng công bố nghiên cứu cho thấy tỷ lệ quan sát gian lận còn nhỏ hơn – thường dưới 0,3%. Do tỷ lệ chênh lệch quá cao, mật độ gian lận quá thấp, các quan sát gian lận phân bố theo tính ngẫu nhiên nên không thể tuân theo phân phối chuẩn. Nếu số biến trong dữ liệu quá lớn, các phương pháp giảm chiều dữ liệu như phân tích thành phần chính (PCA) cũng không khả thi, do PCA chỉ phù hợp với dữ liệu có phân phối chuẩn, gần chuẩn hoặc có quan hệ tuyến tính với nhau.

Những mô hình nghiên cứu tập trung theo định hướng thống kê hoặc dựa trên trí tuệ nhân tạo (artificial intelligent - AI). Phương pháp sử dụng tùy thuộc vào các giả định, các yếu tố đầu vào. Ghosh & Reilly (1994) đã sử dụng một mạng nơron (Neural network) để phát hiện gian lận dựa trên một mẫu bao gồm các tài khoản thẻ tín dụng của tổ chức phát hành, cho thấy mạng nơron phát hiện thấy nhiều gian lận hơn. Hanagandi & cộng sự (1996) sử dụng lịch sử thông tin về giao dịch thẻ tín dụng để tạo mô hình xây dựng điểm số gian lận, đã được kiểm định thỏa mãn trong thực tế. Hansen & cộng sự (1996) đã sử dụng mô hình phản ứng định lượng dự đoán gian lận (bao gồm các hồi quy Probit và Logit) quản lý dựa trên một tập hợp dữ liệu được phát triển bởi một công ty kế toán quốc tế. Các kết quả cho thấy

khả năng tiên đoán tốt cho cả hai giả định chi phí đối xứng và không đối xứng. Haimowitz & Schwarz (1997) sử dụng kỹ thuật phân nhóm (Clustering techniques) để dự báo hành vi gian lận thẻ tín dụng. Dorronsoro & cộng sự (1997) xây dựng một hệ thống trực tuyến để phát hiện gian lận thẻ tín dụng hoạt động dựa trên nhóm phân loại nơron. Để đảm bảo cấu trúc của mô hình, một mô hình dạng phi tuyến Fisher sử dụng phân tích khác biệt, kết quả rất khả quan. Ogwueleka (2011) đã phát hiện gian lận thẻ tín dụng bằng mạng nơron nhân tạo.

3. Cơ sở lý tuyết và phương pháp nghiên cứu

Những quan sát sử dụng phân tích gian lận sẽ chia ra làm hai nhóm: Gian lận (fraud) và không gian lận (legal), được đặt là biến nhị phân (gồm hai giá trị 0 – nếu quan sát là legal và 1 nếu quan sát là fraud). Thông thường, phân tích dữ liệu cho các biến nhóm nhị phân được thực hiện bằng một số phương pháp như: mô hình Logistics (Logistics regression), cây quyết định (Decision Tree), mạng Bayes (Baysian network). Tuy nhiên, dữ liệu về gian lận khác với dữ liệu thông thường: Một tỷ lệ gian lận thấp làm cho dữ liệu mất cân bằng, gây ra cho kết quả chia nhóm bằng sử dụng các phương pháp phân tích dữ liệu nhị phân rất khó khăn, hoặc không thể thực hiện được. Vì vậy, để thực hiện được các phương pháp nêu trên, nghiên cứu sử dụng một số kỹ thuật chọn mẫu được sử dụng để giải quyết vấn đề dữ liệu mất cân bằng.

3.1. Phương pháp xử lý dữ liệu

Một số phương pháp sử dụng trong kỹ thuật xử lý mẫu mất cân bằng bao gồm:

Phương pháp lấy lại mẫu được thực hiện bằng cách hai cách như tăng tỷ lệ các quan sát có giá trị bằng 1 từ rất thấp lên gấp nhiều lần (Oversampling) hoặc giảm tỷ lệ các quan sát có giá trị bằng 0 (Resampling) được nghiên cứu bởi Solberg (1996). Nghiên cứu của Japkowicz (2000) và nghiên cứu của Phua & cộng sự (2004) giải quyết vấn đề dữ liệu phân nhóm mất cân bằng. Dữ liệu được xử lý bắt đầu từ việc tách dữ liệu thành hai phần gồm phần giá trị bằng 0 (legal) và phần có giá trị bằng 1 (fraud). Do dữ liệu mất cân bằng, phần giá trị 1, chiếm tỷ lệ rất nhỏ, sẽ được nhân các quan sát lên gấp nhiều lần. Đồng thời, phần giá trị bằng 0, chiếm tỷ lệ rất lớn, sẽ được giảm số lượng đi nhiều lần bằng thuật toán loại ngẫu nhiên (random), để đảm bảo tính chất ngẫu nhiên của dữ liệu. Dữ liệu cuối cùng được ghép từ hai phần đã xử lý có tỷ lệ gian lận cao hơn ban đầu rất nhiều.

Phương pháp SMOTE (Synthetic Minority Oversampling Technique) thực hiện bằng cách tăng tỷ lệ gian lận trong dữ liệu mất cân bằng từ thuật toán KNN (K Nearest Neighbor) được giới thiệu bởi Batista & cộng sự (2004). Các quan sát có giá trị 1 được tạo thêm có các đặc tính số liệu gần với các quan sát gian lận ban đầu. Thuật toán KNN sử dụng những quan sát gần với nhau để tạo nên một nhóm và tìm ra quan sát tâm của nhóm. Dựa trên khoảng cách giữa các quan sát, phương pháp tìm ra K quan sát lân cận đạt yêu cầu sao cho khoảng cách từ quan sát đó đến quan sát tâm không lớn hơn khoảng cách tối đa đặt ra. Kết quả cuối cùng tạo thêm các quan sát mới ở giữa các quan sát gian lận ban đầu.

3.2. Phương pháp nghiên cứu

3.2.1. Phát hiện quan sát ngoại lai

Kỹ thuật phát hiện gian lận (Outlier Detection) theo Bolton & Hand (2002) thực hiện thông qua phân tích dữ liệu để phát hiện được các dữ liệu cá biệt, bất thường. Phương pháp phát hiện dựa trên phân tích các dữ liệu bình thường trong quá khứ để phát hiện gian lận. Phần lớn các outliers thường tách ra xa khỏi xu hướng so với các quan sát còn lại.

Phát hiện dữ liệu gian lận thông qua các quan sát outliers sẽ có thể là dạng giá trị cực tiểu (Left outlier), giá trị ngoại lai cực đại (Right outlier) hay giá trị ngoại lai đại điện cho một phân lớp (Representative outlier). Các quan sát ngoại lai có thể được phát hiện bằng nhiều phương pháp như thông qua kỹ thuật đồ thị, thống kê đồ thị lịch sử và một số phương pháp khác. Các tiêu chuẩn để đánh giá phát hiện sai phạm trong trường hợp dữ liệu đặc biệt sẽ được các chuyên gia xây dựng, phân tích và tìm ra dựa trên dữ liệu để phát hiện các khách hàng có hành vi lừa đảo.

3.2.2. Kỹ thuật phân cụm

Kỹ thuật phân cụm (Clustering techniques) phân nhóm thành những nhóm khác nhau, được giới thiệu bởi Bolton & Hand (2002). Phát hiện sai phạm dựa trên các luồng dữ liệu đặc biệt là những quan sát khi đạt hoặc rơi vào một vùng giá trị xác định sẽ được nhận diện là gian lận hoặc không gian lận (100% gian lận hoặc 100% không gian lận). Thông thường, ngân hàng sẽ tách các quan sát đặc biệt để có chính sách riêng với mỗi khách hàng. Khi khách hàng bị từ chối có nghĩa là tỷ lệ cao các quan sát này luôn xấu, ngân hàng sẽ đánh giá mức độ ưu tiên thấp hoặc đưa vào danh sách cần chú ý (Blacklist). Ngược lại, những khách hàng luôn luôn tốt sẽ luôn được chấp nhận sớm và đánh dấu độ ưu tiên cao hơn.

3.2.3. Hồi quy Logit

Mô hình hồi quy Logit được giới thiệu bởi Berkson (1944) là một công cụ sử dụng phổ biến trong phân tích dữ liệu với biến nhị phân. Một số phát triển của Altman & cộng sự (1994); Flitman (1997) sử dụng trong phân tích mô hình hồi quy đa biến, phân tích khác biệt.

Mô hình hồi quy Logit xác định xác suất xảy ra sự kiện Y =1 như sau:

$$P(\mathrm{Y} = 1/X_1, \dots, X_n) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}{1+ e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}$$

Trong đó, X_1,..., X_n là giá trị của biến độc lập.

Mô hình hồi quy logit có thể được sử dụng để ước lượng tỷ lệ log(odds) cho mỗi biến độc lập của mô hình (Ohlson, 1980):

$$\ln\frac{(\mathrm{Y} = 1/X_1, \dots, X_n)}{(\mathrm{Y} = 0/X_1, \dots, X_n)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Các tham số β_n được ước lượng bằng phương pháp hợp lý tối đa (Maximum Likelihood – ML). Mô hình Logit được được sử dụng với nhiều dạng dữ liệu, ít những điều kiện ràng buộc, hiệu quả khi áp dụng vào thực tế, dễ giải thích kết quả, có khả năng theo dõi, chẩn đoán và hiệu chỉnh để kết quả phù hợp với thực tế.

3.2.4. Cây quyết định

Cây quyết định (Decision Tree) là một mô hình phân loại được giới thiệu bởi Belson (1959), được sử dụng rộng rãi trong nhiều lĩnh vực khác nhau. Sau khi giới thiệu về hệ thống phương pháp học máy (Machine learning), cây quyết định đã đã được phát triển hơn với các thuật toán C4.5 bởi Quinlan (1996) và thuật toán ID3 bởi Quinlan (1986).

Decision Tree là một cây phân loại có cấu trúc được phân lớp các đối tượng dựa vào dãy các luật. Các biến độc lập và thuộc tính có thể thuộc các kiểu dữ liệu khác nhau như nhị phân (binary), định danh (nominal), thứ bậc (ordinal), dữ liệu định lượng (quantitative). Để xác định biến nào sử dụng phân loại trước, biến nào sử dụng sau, trọng số thông tin (Entropy) ứng với mỗi biến được tính toán, giá trị thông tin càng cao, biến đó càng mang nhiều thông tin phân loại.

Giả sử biến X, có n giá trị x_1,x_2,..,x_n. Giả sử với mỗi x_i sẽ có xác suất để sự kiện xảy ra tương ứng là P_i, thỏa mãn điều kiện: $0 \le p_i \le 1$ và $\sum_{i=1}^{n} p_i = 1$

Trọng số thông tin của biến X: $\mathrm{E(X)} = -\sum_{i=1}^{n} p_i \log_2(p_i)$

3.2.5. Phương pháp Bayesian

Phương pháp Bayesian (Bayesian Network) được ứng dụng để phân lớp dựa trên xác suất có điều kiện. Cũng như hàm Logistic, kết quả của Bayesian là một xác suất có giá trị từ 0 đến 1 (thể hiện xác suất xảy ra của sự kiện từ 0% đến 100%), các biến liên kết với nhau bằng mối liên kết xác suất.

Phương pháp Bayesian được phát triển từ định lý Bayes trong xác suất thống kê, theo Carlin & Louis (2010), phương pháp Bayesian thiên về thống kê hơn là hồi quy. Phương pháp Bayesian khá hiệu quả và dễ sử dụng, không yêu cầu các điều kiện của dữ liệu, có thể làm trên cả dữ liệu số hay chữ. Với bộ số liệu nhỏ hoặc bị mất cân bằng, phương pháp càng hiệu quả, khi mà các phương pháp khác không thực hiện được, hoặc phải xử lý dữ liệu rất nhiều thao tác.

Với mục đích phát hiện gian lận, mạng Bayesian sẽ được xây dựng với quy tắc Bayes cùng với điều kiện P(Y=1) + P(Y=0) = 1 được viết như sau:

$P(Y=1 \mid X) = [P(X \mid Y=1)P(Y=1)]/P(X)$

$P(Y=0 \mid X)=[P(X \mid Y=0)P(Y=0)]/P(X)$

$P(Y=0 \mid X)=[P(X \mid Y=0)P(Y=0)]/P(X)$

Trong đó: $P(X)=P(Y=1)P(X \mid Y=1)+P(Y=0) P(X \mid Y=0)$

Các thành phần được tính như sau: P(Y=1) chính là tỷ lệ sai phạm của mẫu sử dụng để chạy mô hình. Với giả thiết các biến độc lập nhau:

$$P(X|\mathrm{Y} = 1) = \prod_{k=1}^{n} P(x_k|\mathrm{Y} = 1)$$

$$P(x_k|\mathrm{Y} = 1) = s_{ik}/s$$

Trong đó: *S* là số lượng gian lận trong mẫu, S_{ik} là số lượng gian lận thuộc phân lớp *i* của biến X_k

3.2.6. Phương pháp Stacking

Phương pháp Stacking (Stacked generalization) là một phương pháp để tăng tính chính xác kết quả đầu ra được giới thiệu bởi Wolpert (1992), Smyth & Wolpert (1998). Thuật toán Stacking sử dụng nhiều mô hình học đơn lẻ để ra kết quả của riêng với mỗi mô hình, sau đó sử dụng chính kết quả đó để kết hợp với dữ liệu ban đầu, dự đoán lại với một mô hình khác dựa trên bộ dữ liệu mới tạo ra để tìm kết quả cuối cùng. Phương pháp Stacking là một dạng cụ thể của lớp các phương pháp tập hợp (Ensemble method), được giới thiệu bởi Dietterich (2000), được kiểm chứng khá hiệu quả và được phát triển rộng rãi trong lĩnh vực học máy và chạy dữ liệu nhiều lần để tự cải thiện kết quả dùng trí tuệ nhân tạo AI.

3.3. Phương pháp đánh giá

Bảng 1: Ma trận nhầm lẫn

		Dự báo	
		0	1
Thực tế	0	TN	FP
	1	FN	TP

Nguồn: Townsend (1971, 44).

Để đánh giá độ chính xác của mô hình, nghiên cứu sử dụng ma trận nhầm lẫn (confusion matrix) để đánh giá mức độ dự báo chính xác của mô hình, được Townsend (1971) giới thiệu (Bảng 1).

Độ chính xác (Accuracy) = $\frac{TN+TP}{TN+FN+TP+FP}$

Trong đó:

TN (True Negative): Số quan sát dự đoán đúng không gian lận Y= 0;

TP (True Positive): Số quan sát dự đoán đúng gian lận Y= 1;

FP (False Positive): Số quan sát dự đoán sai Y= 0 thành Y=1;

FN (False Negative): Số quan sát dự đoán sai Y = 1 thành Y= 0.

4. Minh họa kết quả

4.1. Dữ liệu sử dụng

Bộ số liệu sử dụng 492 quan sát gian lận trên tổng số 284807 quan sát đã thực hiện giao dịch thẻ tín dụng, tỷ lệ gian lận 0,172%, của Châu Âu trong tháng 09/2013 theo nguồn www.kaggle.com. Các dữ liệu liên quan trực tiếp đến giao dịch và tài khoản đã được giao dịch nên cần được mã hóa.

4.2. Xử lý dữ liệu đầu vào

Để có sự tương ứng trong kết quả so sánh, dữ liệu ở 2 phương pháp được tăng tỷ lệ gian lận lên xấp xỉ 18-20 lần so với dữ liệu gốc.

Dữ liệu xử lý bằng phương pháp tăng quan sát gian lận (Oversampling): Lặp lại các quan sát gian lận 20 lần, không làm thay đổi cấu trúc dữ liệu ban đầu, các quan sát gian lận được tạo ra giống hoàn toàn các quan sát gian lận ban đầu của bộ dữ liệu gốc.

Dữ liệu xử lý bằng phương pháp tạo ra các quan sát gian lận mới (SMOTE): Quanh mỗi quan sát ban đầu, tìm tối đa 15 quan sát lân cận với khoảng cách không quá 0,1 (theo thang đo khoảng cách DGOWER trong phần mềm SAS). Với mỗi một quan sát gốc sẽ có thêm tối đa 15 quan sát xung quanh tạo thành một nhóm. Quan sát gốc sẽ lần lượt cùng các quan sát xung quanh tạo ra các quan sát mới theo công thức:

$X_{new} = X_a + (X_a - X_b)*\alpha$

Trong đó:

X_{new} là giá trị của quan sát mới được tạo ra;

X_a là giá trị của quan sát ban đầu;

X_b là giá trị của một trong 15 quan sát lân cận với quan sát ban đầu;

α là một giá trị ngẫu nhiên từ 0 đến 1.

Các quan sát mới tạo ra sẽ làm thay đổi cấu trúc cũ của dữ liệu, nhưng sẽ tạo ra các cụm "đám mây gian lận" để dễ phân biệt hơn, nhưng cũng dễ nhầm lẫn với các quan sát không ain lận gần quan sát gian lận hơn. Kết quả dữ liệu thu được chi tiết theo Bảng 3.

4.3. Lựa chọn biến cho mô hình Logit

Bảng 2: Thông tin biến sử dụng

Biến	Thông tin
Class	Biến phân loại với 1- Gian lận và 0 - Không gian lận
V1 đến V28	Thông tin cá nhân của chủ thẻ
Time	Thời gian tính bằng giây kể từ giao dịch đang ghi nhận đến giao dịch đầu tiên được ghi nhận trong ngày
Amount	Số tiền giao dịch của chủ thẻ (đơn vị - USD)

Nguồn: Kaggle (2018).

Phương pháp lựa chọn biến và phân nhóm giá trị trong biến sử dụng WOE (Weights Of Evidence) được tính cho từng nhóm thể hiện khả năng dự báo của biến:

$$WOE_i = \ln\left(\frac{\text{Dist legal}_i}{\text{Dist fraud}_i}\right) \times 100$$

Trong đó: Dist legal$_i$ là tần suất tích lũy của các quan sát không gian lận trong nhóm I;

Dist fraud$_i$ là tần suất tích lũy của các quan sát gian lận trong nhóm i.

Các biến được chọn dựa trên hệ số giá trị thông tin IV (Information Value) được xác định theo công thức:

$$IV = \sum_{i=1}^{k}(\text{Dist legal}_i - \text{Dist fraud}_i) * WOE_i$$ với k là số nhóm của biến

Biến được lựa chọn phù hợp nếu có hệ số giá trị thông tin IV thuộc khoảng 0.1 đến 0.5. Theo kết quả tính toán được, một số biến sau được lựa chọn vào mô hình với hệ số IV tương ứng.

Kết quả và phân tích thực hiện với các biến theo Bảng 4 cho các mô hình Logit, Decision Tree, Bayes Network và Stacking được trình bày tại Bảng 5.

Ma trận nhầm lẫn sử dụng với điểm cắt (Cutoff) là 0,5%. Mức Cutoff này sẽ khác với thực tế tùy vào từng ngân hàng. Hai thông số được tập trung chú ý trong bảng ma trận nhầm lẫn là số quan sát dự đoán đúng TP (True Positive) và TN (True Negative). Với số quan sát dự đoán sai FP (False Positive), khách hàng bị đánh giá nhầm không gian lận thành gian lận càng ít càng tốt. Khi mô hình loại nhầm quá nhiều khách hàng không gian lận (bị nhầm thành gian lận) không chỉ thiệt hại cho ngân hàng về lợi nhuận có thể có từ những khách hàng đó, mà còn ảnh hưởng tới danh tiếng hay hình ảnh của ngân hàng.

Kết quả thể hiện độ chính xác khá tốt với bộ dữ liệu gốc với các mô hình (ngoại trừ mô hình Logistic không sử dụng cho dữ liệu mất cân bằng, cụ thể:

Mô hình (1) thể hiện kết quả tính toán được của mô hình Logistic với bộ dữ liệu Oversampling và bộ dữ liệu SMOTE với các biến đã được phân nhóm theo WOE. Kết quả dự báo với dữ liệu SMOTE cho kết quả cao. Số quan sát dự báo đạt độ chính xác 67,58%. Mô hình Logistic không thể sử dụng được với dữ liệu chưa qua xử lý do bộ dữ liệu gốc sử dụng mất cân bằng, tỷ lệ fraud chỉ 0,172%. Mặt khác, điều kiện để thực hiện phân nhóm khi phân tích WOE không thể có 100% không gian lận hoặc 100% gian lận. Cho dù đã sử dụng phương pháp biến đổi dữ liệu, với 8 biến được chọn, ước lượng kết quả cho Logit không cho ra kết quả tốt nhất nhưng độ ổn định tương đối cao, thể hiện mức chênh lệch độ chính xác giữa các kết quả của mô hình nhỏ. Kết quả chỉ ra tính ưu việt của mô hình Logit trong sử dụng với các bộ dữ liệu bình thường, dữ liệu không quá mất cân bằng (0,5% - 1% tỷ lệ Fraud). Bằng cách nhân đôi quan sát gian lận, mô hình Logit kết hợp với phân nhóm WOE vẫn là mô hình được yêu thích và sử dụng rộng rãi do tính ổn định, dễ kiểm soát, dễ triển khai trong thực tế.

Mô hình (2) sử dụng cây quyết định kết hợp với các bộ số liệu cho kết quả chính xác và ổn định cao. Trong đó, phương pháp xử lý dữ liệu SMOTE cho kết quả tốt nhất, loại bỏ được 429 quan sát fraud và đánh giá nhầm 118 quan sát không gian lận thành gian lận – một con số chấp nhận được khi so sánh với số quan sát phân loại nhầm của các mô hình. Nguyên nhân do phương pháp SMOTE đã biến đổi cấu trúc ban đầu của bộ dữ liệu thành bộ dữ liệu có cấu trúc thích hợp nhất với mô hình cây quyết định. Bộ số liệu để cây quyết định hoạt động hiệu quả nhất là số liệu phân tán thành từng cụm. Như vậy, các luật của cây quyết định sẽ dễ dàng phân biệt từng cụm mà không cần số quan sát mỗi cụm phải ít nhất xấp xỉ 5% tổng số quan sát. Phương pháp SMOTE lại xây dựng bộ dữ liệu ban đầu thành bộ dữ liệu mới dựa trên phương pháp tìm những quan sát gần nhau, thuận lợi cho cây quyết định phân loại. Tuy cho kết quả tốt nhưng phương pháp SMOTE kết hợp với cây quyết định có phức tạp hơn sử dụng trực tiếp mô hình mạng Bayesian hay hồi quy Logit bởi việc nhân lượng quan sát gian lận lên. Đồng thời,

Bảng 3: Kết quả xử lý dữ liệu đầu vào

Phương pháp xử lý dữ liệu	Số quan sát	Tỷ lệ gian lận (%)
Dữ liệu gốc	284.807	0,172
Dữ liệu Oversampling	294.155	3,35
Dữ liệu SMOTE	293.466	3,12

Nguồn: Tác giả tính toán dựa trên bộ dữ liệu gốc.

Bảng 4: Kết quả IV chọn biến chạy mô hình Logit với bộ dữ liệu Oversampling và SMOTE

Biến	Oversampling	SMOTE
V3	0,3425	0,3254
V4	0,3655	0,37
V10	0,4264	0,2702
V11	0,3873	0,4246
V12	0,4253	0,3897
V14	0,4908	0,4113
V16	0,31	0,3115
V17	0,4207	0,3998

Nguồn: Tác giả tính toán dựa trên bộ dữ liệu xử lý.

với khối lượng tính toán lớn, tốn nhiều thời gian tùy thuộc vào số lượng quan sát gian lận và lượng biến có của dữ liệu, nên phương pháp này được cân nhắc khi sử dụng.

Mô hình (3) sử dụng mạng Bayesian cho ra kết quả không nổi trội nhất, nhưng cân bằng. Bayesian đã giải quyết được vấn đề mà Logit không làm được, mặc dù 2 phương pháp này là ánh xạ của nhau. Mạng Bayesian tuy loại được số gian lận ra nhiều nhất (413 quan sát) nhưng xác định nhầm 1714 quan sát không gian lận thành gian lận. Kết quả cho thấy có rất nhiều phân loại sai, mô hình mạng Bayesian đưa ra kết quả không ổn định, không phù hợp với các bộ dữ liệu trong nghiên cứu.

Mô hình (4) sử dụng phương pháp Stacking sử dụng cho kết quả ổn định tương đương với mô hình Logit. Phương pháp Stacking được sử dụng kết hợp kết quả dự báo của cả 2 mô hình mạnh nhất để tạo ra thang xác suất đánh giá quan sát cuối cùng, cho thấy kết quả cải thiện đáng kể so với đơn mô hình. Mô hình cho kết quả chính xác đều trên 60%, tỷ lệ loại các quan sát gian lận cao nhất trong các mô hình sử dụng. Trong đó, 427 quan sát gian lận bị loại với bộ dữ liệu SMOTE trong khi dữ liệu không gian lận bị loại cũng chấp nhận được. Ưu điểm của phương pháp là có thể kết hợp nhiều mô hình nhỏ, có sức đánh giá yếu lại với nhau tạo ra một mô hình đánh giá tốt hơn, dễ sử dụng nhưng lại không phức tạp, hoàn toàn có thể kiểm soát.

5. Kết luận

Các phương pháp xử lý dữ liệu và mô hình phát hiện gian lận đều không chiếm ưu thế hoàn toàn. Người sử dụng phải cân nhắc giữa tính hợp lý, độ ổn định và sức mạnh cũng như tính phức tạp khi thực hiện với mỗi mô hình hay phương pháp. Công cụ lấy mẫu theo phương pháp Oversampling có ưu điểm dễ

Bảng 5: Kết quả chạy mô hình

Mô hình	Dữ liệu sử dụng	FN	TN	FP	TP	Độ chính xác
Logistic (1)	SMOTE	88	284	260	404	66,41%
	Oversampling	86	284	245	406	**67,58%**
Cây quyết định (2)	SMOTE	63	284	118	429	79,76%
	Oversampling	108	284	106	384	75,74%
	Gốc	99	284	31	393	**83,90%**
Mạng Bayesian (3)	SMOTE	74	283	1359	418	32,85%
	Oversampling	79	283	1714	413	27,95%
	Gốc	90	284	294	402	**64,11%**
Stacking (4)	SMOTE	65	284	348	427	63,26%
	Oversampling	79	284	287	413	**65,57%**
	Gốc	77	284	301	415	64,90%

Nguồn: Tác giả tính toán dựa trên bộ dữ liệu xử lý.

thực hiện, tuy tăng về kích thước dữ liệu nhưng các quan sát được lặp lại nên trong một số trường hợp phương pháp thể hiện tính không hiệu quả. Phương pháp SMOTE giải quyết được vấn đề dữ liệu mất cân bằng. Tuy nhiên, phương pháp chỉ hiệu quả với một số lượng gian lận rất thấp, khối lượng tính toán phức tạp và phải thực hiện lặp đi lặp lại nhiều lần. Trong quá trình thực hiện, phương pháp tạo ra các quan sát nhiễu và phải xử lý dữ liệu bị thiếu trước khi thực hiện tăng số quan sát.

Với phân tích phân nhóm sử dụng WOE, mô hình Logistic là phương pháp truyền thống được cải tiến, đơn giản, hiệu quả, dễ áp dụng trên nhiều hệ thống, tính ổn định tốt, kiểm soát hoàn toàn được mô hình. Tuy nhiên, mô hình Logistic gặp phải vấn đề dữ liệu mất cân bằng và phải phân tích WOE hợp lý trước khi chạy mô hình. Mạng Bayesian là phương pháp áp dụng được trên tất cả các loại dữ liệu. Phương pháp đem lại kết quả tốt nhưng dễ bị phân loại nhầm bởi các quan sát nhiễu hay lặp lại. Kết quả cho thấy mạng Bayesian ứng dụng hiệu quả nhất với các mô hình có lượng quan sát ít. Kết quả trên cho thấy phương pháp phân loại sử dụng cây quyết định là mô hình đơn giản nhất và không gặp phải vấn đề với tất cả các loại biến, mất dữ liệu, dữ liệu ngoại lai. Tuy nhiên, phương pháp cây quyết định chỉ hiệu quả với dữ liệu có tính phân cụm cao, dữ liệu được chia thành nhiều nhánh nhỏ dễ bị làm cho kết quả quá đúng với mẫu (overfitting).

Ngân hàng thương mại ở Việt Nam sử dụng mô hình Logistic kết hợp với cây quyết định với biến phân nhóm WOE trên phương pháp xử lý Oversampling phù hợp hơn cả. Phương pháp này có thể xử lý được các nhóm biến bị thiếu, ngoại lai, các biến định tính có thứ bậc. Ngân hàng có thể dễ dàng kiểm tra tính ổn định của các biến, nhóm biến, đồng bộ trên hệ thống chung của ngân hàng để thuận hiện xây dựng hệ thống tự động đánh giá từng khách hàng.

Ngoài ra, ngân hàng có thể cân nhắc sử dụng phương pháp SMOTE để xử lý số liệu. Mặc dù khó thực hiện hơn các cách khác nhưng ở những bộ dữ liệu có tỷ lệ gian lận rất thấp (nhỏ hơn 0,05%), phương pháp này sẽ mang lại một kết quả rất tốt. Ngân hàng cũng cần kiểm soát các quan sát mới tạo ra để có thể chắc chắn các quan sát này sẽ tồn tại trong thực tế là quan sát gian lận.

Tài liệu tham khảo:

Altman, E.I., Marco, G. & Varetto, F. (1994), 'Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience)', *Journal of banking & finance*, 18(3), 505-529.

Batista, G.E., Prati, R.C. & Monard, M.C. (2004), 'A study of the behavior of several methods for balancing machine learning training data', *ACM SIGKDD Explorations Newsletter*, 6(1), 20-29.

Belson, W.A. (1959), 'Matching and prediction on the principle of biological classification', *Applied Statistics*, 8(2), 65-75.

Berkson, J. (1944), 'Application of the logistic function to bio-assay', *Journal of the American Statistical Association*, 39(227), 357-365.

Bolton, R.J. & Hand, D.J. (2002), 'Statistical fraud detection: A review', *Statistical Science*, 17(3), 235-249.

Carlin, B.P. & Louis, T.A. (2010), *Bayes and empirical Bayes methods for data analysis*, Chapman and Hall/CRC, Florida, US.

Dietterich, T.G. (2000), 'Ensemble methods in machine learning', In *International workshop on multiple classifier systems* (1-15), Springer, Berlin, Heidelberg.

Dorronsoro, J.R., Ginel, F., Sánchez, C.R. & Santa Cruz, C. (1997), 'Neural fraud detection in credit card operations', *IEEE Transactions on Neural Networks*, retrieved on September 18th 2018, from <https://repositorio.uam.es/handle/10486/663701/>.

Flitman, A.M. (1997), 'Towards analysing student failures: neural networks compared with regression analysis and multiple discriminant analysis', *Computers & Operations Research*, 24(4), 367-377.

Ghosh, S. & Reilly, D.L. (1994), 'Credit card fraud detection with a neural-network', In *Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences* (3, 621-630), retrieved on September 18th 2018, from <https://ieeexplore.ieee.org/abstract/document/323314/>.

Haimowitz, I.J. & Schwarz, H. (1997), 'Clustering and prediction for credit line optimization', In *Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection & Risk Management* (29-33), retrieved on September 18th 2018, from <http://www.aaai.org/Library/Workshops/1997/ws97-07-006.php>.

Hanagandi, V., Dhar, A. & Buescher, K. (1996), 'Density-based clustering and radial basis function modeling to generate credit card fraud scores', In *Proceedings of the IEEE/IAFE 1996 Conference on Computational Intelligence for Financial Engineering* (247-251), IEEE, New York City, USA.

Hansen, J.V., McDonald, J.B., Messier Jr, W.F. & Bell, T.B. (1996), 'A generalized qualitative-response model and the analysis of management fraud', *Management Science*, 42(7), 1022-1032.

Japkowicz, N. (2000), 'The class imbalance problem: Significance and strategies', In *Proceedings of the Internationall Conference on Artificial Intelligence*, retrieved on September 18th 2018, from <https://pdfs.semanticscholar.org/907b/02c6322d0e7dff6b0201b03e3d2c6bc1d38f.pdf>.

Kaggle (2018), *Credit Card Fraud Detection*, retrieved on September 18th 2018, from <https://www.kaggle.com/mlg-ulb/creditcardfraud>.

Ogwueleka, F.N. (2011), 'Data mining application in credit card fraud detection system', *Journal of Engineering Science and Technology*, 6(3), 311-322.

Ohlson, J.A. (1980), 'Financial ratios and the probabilistic prediction of bankruptcy', *Journal of Accounting Research*, 109-131.

Phua, C., Alahakoon, D. & Lee, V. (2004), 'Minority report in fraud detection: classification of skewed data', *ACM SIGKDD Explorations Newsletter*, 6(1), 50-59, New York, USA.

Quinlan, J.R. (1996), 'Bagging, boosting, and C4. 5', in *AAAI-96 Proceedings*, 725-730.

Quinlan, J.R. (1986), 'Induction of decision trees', *Machine Learning*, 1(1), 81-106.

Smyth, P. & Wolpert, D. (1998), 'Stacked density estimation', In *Advances in neural information processing systems* (668-674), retrieved on September 18th 2018, from <http://papers.nips.cc/paper/1353-stacked-density-estimation.pdf/>.

Solberg, A.H.S., Taxt, T. & Jain, A.K. (1996), 'A Markov random field model for classification of multisource satellite imagery', *IEEE Transactions on Geoscience and Remote Sensing*, 34(1), 100-113, retrieved on September 18th 2018, from <https://ieeexplore.ieee.org/abstract/document/481897/>.

Townsend, J.T. (1971), 'Theoretical analysis of an alphabetic confusion matrix', *Perception & Psychophysics*, 9(1), 40-50.

Wolpert, D.H. (1992), 'Stacked generalization', *Neural Networks*, 5(2), 241-259.