TRƯỜNG ĐẠI HỌC KINH TẾ TP. HỒ CHÍ MINH

UEH
UNIVERSITY

KỶ YẾU HỘI THẢO KHOA HỌC QUỐC GIA

HỆ THỐNG THÔNG TIN TRONG KINH DOANH VÀ QUẢN LÝ

ISBM21

UEH PUBLISHING HOUSE NXB KINH TẾ TP. HỒ CHÍ MINH

ỨNG DỤNG MÔ HÌNH CATBOOST TRONG NHẬN DIỆN GIAN LẬN THẺ TÍN DỤNG

Dương Minh Trúc, Nguyễn Duy Nghĩa, Châu Thị Bích Trâm, Nguyễn Văn Hồ, và Hồ Trung Thành

Trường Đại học Kinh tế - Luật, Đại học Quốc gia Tp. Hồ Chí Minh

Tóm tắt: *Sự phát triển vượt bậc của công nghệ làm cho người tiêu dùng thay đổi hình thức thanh toán hướng tới một xã hội không tiền mặt, vô tình tạo cơ hội để các hành vi chiếm đoạt tài sản, lừa đảo qua thẻ tín dụng diễn ra thường xuyên hơn. Do đó, nhu cầu về một mô hình có thể ngăn chặn hay giảm được nguy cơ này là điều rất cấp thiết. Nhiều nghiên cứu trước đây tập trung phát hiện gian lận thẻ tín dụng với các phương pháp các Học máy khác nhau. Tuy nhiên, chưa nhiều nghiên cứu tiến hành trên dữ liệu lớn cũng như chưa xây dựng chi tiết cho mô hình với các đặc điểm nhận dạng người dùng để phát hiện ra các hình thức gian lận. Do đó, các mô hình này chưa đáp ứng được việc giảm tỷ lệ cảnh báo sai khi được cung cấp với luồng dữ liệu quy mô lớn. Trong bài báo này, tác giả sử dụng tập dữ liệu "Phát hiện gian lận IEEE-CIS" của Vesta Corporation để xây dựng mô hình CatBoost cải thiện độ chính xác trong việc tìm thấy và giảm tình trạng gian lận khi thanh toán giao dịch bằng thẻ tín dụng. Thử nghiệm cho thấy mô hình mang lại kết quả tốt, đạt giá trị là 82,30%.*

Từ khóa: CatBoost, nhận diện gian lận, định danh người dùng, xây dựng đặc trưng, biến đổi đặc trưng.

1. MỞ ĐẦU

Thương mại điện tử đã thay đổi nhanh chóng trong các thập kỷ qua. Giao dịch trực tuyến được sử dụng ngày càng nhiều tạo điều kiện cho các phương thức thanh toán bằng thẻ trở nên phổ biến. Tuy nhiên, sự thay đổi trong phương thức giao dịch này cũng góp phần gia tăng các hoạt động gian lận. 'Gian lận' là những hành vi cố ý làm sai lệch thông tin kinh tế, tài chính, hành vi gian lận là loại hành vi có chủ ý thường gắn liền với tính vụ lợi [1]. Việc phát hiện gian lận là một quá trình xác định các hành vi bất thường của chủ thẻ khi đối chiếu với lịch sử giao dịch của thẻ. Dựa trên những khác biệt, cảnh báo sẽ được gửi nếu các giao dịch có nguy cơ vượt quá ngưỡng được phân loại là gian lận. Sự gian lận trong giao dịch thường được thực hiện thông qua việc truy cập trái phép vào thông tin thẻ như số tài khoản, số thẻ tín dụng [2], địa chỉ email, số điện thoại [3], và các thông tin khác để lấy cắp tiền.

Để ngăn chặn gian lận thẻ, ngành tài chính - ngân hàng đã bỏ ra rất nhiều tiền của, nguồn lực để xây dựng các hệ thống phát hiện gian lận, nhiều thuật toán Học máy đã được sử dụng, từ loại cổ điển như Logistic Regression, Support Vector Machines, Decision Tree, Hidden Markov Models cho đến các phương pháp hiện đại như CatBoost - một giải pháp hứa hẹn nhất nhờ kết quả và hiệu suất phát hiện giao dịch gian lận. Thực tế cho thấy mô hình Mạng nơ-ron nhân tạo cũng được sử dụng trong bài toán xác định gian lận thẻ tín dụng, tuy nhiên, hạn chế

của thuật toán này là không thể kết hợp lịch sử giao dịch của người dùng, điều này là lợi thế của các mô hình dựa trên cây quyết định như CatBoost. Nhóm tác giả nhận thấy CatBoost là phương pháp tối ưu trong việc xử lý dữ liệu cùng lúc của cả người dùng mới và người dùng lịch sử.

Cấu trúc phần còn lại của bài báo như sau: phần 2 đánh giá, so sánh ngắn gọn về các nghiên cứu trước đây về các kỹ thuật Học máy dùng trong phát hiện gian lận thẻ. Phần 3 trình bày cơ sở toán học của thuật toán CatBoost. Phần 4 mô tả chi tiết các thông tin về mô hình được đề xuất. Phần 5 đánh giá kết quả thu và điểm qua các ý tưởng để phát triển, nghiên cứu thêm trong tương lai.

2. CÁC NGHIÊN CỨU LIÊN QUAN

Trong phần này, bài báo sẽ nói về các nghiên cứu liên quan đến phát hiện gian lận thẻ tín dụng đã được thực hiện. Shiyang Xuan và các cộng sự trong nghiên cứu [4] đã sử dụng mô hình Random Forest dựa trên CART. Phương pháp của họ là sử dụng dữ liệu lịch sử giao dịch dựa trên các đặc điểm để phân biệt giữa hành vi giao dịch thông thường và gian lận. Bộ dữ liệu được cung cấp bởi công ty thương mại điện tử ở Trung Quốc với 62 đặc trưng và hơn 3 triệu giao dịch với 82.000 giao dịch trong số đó là các gian lận. Tỷ lệ giữa giao dịch bình thường và bất thường được sử dụng là 5:1. Nghiên cứu mang lại độ chính xác (accuracy) là 98,67%, Precision là 32,68% và Recall là 59,62%. Mô hình đem đến kết quả với mức độ chính xác tương đối cao, nhưng tỷ lệ false positive cũng cao, nên có khả năng hệ thống phát hiện này gây phiền nhiễu cho những khách hàng có giao dịch hợp pháp bị xác định nhầm là gian lận. Tuy mô hình Random Forest cung cấp kết quả chính xác cao, nhưng chỉ được áp dụng trên tập dữ liệu nhỏ, không phù hợp với các tổ chức tài chính và doanh nghiệp lớn. Mặc dù có thể áp dụng Logistic Regression và SAE cho dữ liệu lớn, nhưng mang lại độ chính xác thấp và không phù hợp để sử dụng trong thực tế. Điều đó cũng đã được xác nhận trong nghiên cứu [5] của Aya Abdel Naby và các cộng sự.

Hasan Najadat và các cộng sự trong nghiên cứu [6] đã áp dụng BiLSTM-MaxPooling-BiGRU-MaxPooling để dự đoán gian lận. Các tác giả cũng áp dụng Naive-based, Voting, AdaBoosting, Random Forest, Decision Tree và Logistic Regression để so sánh hiệu quả của từng mô hình. Bài báo sử dụng tập dữ liệu có sự mất cân bằng cao. Các tác giả đã sử dụng Kỹ thuật lấy mẫu ngẫu nhiên (SMOTE) để xử lý sự mất cân bằng này. Kết quả là các mô hình Học sâu kết hợp ba kỹ thuật lấy mẫu đạt độ chính xác cao hơn đáng kể so với các mô hình Học máy. Kết quả cao nhất của các phương thức dựa trên Học máy là 81,00%. Tuy nhiên, khi các tác giả kết hợp BiLMST-MaxPooling-BiGRU-MaxPooling với kỹ thuật lấy mẫu ngẫu nhiên, họ đã đạt được độ chính xác 91,37%. Từ những nghiên cứu trên, có thể thấy rằng nếu chỉ riêng các thuật toán Học máy chưa thật sự hiệu quả trên tập dữ liệu lớn phức tạp như trình bày trên.

Trong một nghiên cứu khác [5], các tác giả đã thử nghiệm với Deep Learning để phát hiện gian lận với mục đích độ chính xác cao. Nhóm tác giả đã sử dụng mô hình Decision Tree, SAE, RBM với độ chính xác lần lượt là 90,49%, 80,52% và 91,53%. SAE thấp so với Decision Tree và RBM.

3. CƠ SỞ LÝ THUYẾT

3.1. Tổng quan về bản chất thuật toán CatBoost

Giả sử ta có một tập dữ liệu các quan sát $D = \{(x_k, y_k)\}_{k=1..n}$, *với* $x_k = (x_k^1, \ldots, x_k^m)$ là một vectơ tùy ý có m phần tử và $y_k \in R$ là biến mục tiêu dạng nhị phân hoặc số bất kỳ.

Mỗi cặp quan sát (x_k, y_k) là các mẫu độc lập và có phân phối đồng nhất tuân theo phân phối $P(\cdot,\cdot)$. Mục tiêu là cần huấn luyện hàm số $F: R^m \rightarrow R$ sao cho cực tiểu hoá kỳ vọng của hàm mất mát: $L(F) := EL(y, F(x))$, với $L(\cdot,\cdot)$ là hàm mất mát trơn và (x, y) là một cặp dữ liệu trong tập thử nghiệm lấy mẫu từ P độc lập với tập huấn luyện D [10].

Thuật toán tăng cường độ dốc xây dựng một phép toán lặp chuỗi các hàm xấp xỉ:

$F^t: R^m \rightarrow R$, $t = 0, 1,...$ đến khi hiệu quả mô hình không tăng thêm thì dừng lại. Cụ thể, với mỗi hàm F^t nhận được từ hàm F^{t-1} trước đó theo dạng cộng dồn: $F^t = F^{t-1} + \alpha h^t$, với α là bước nhảy và h^t là hàm số được chọn từ họ hàm H sao cho $h^t: R^m \rightarrow R$ (hàm dự đoán cơ sở) để cực tiểu hóa kỳ vọng của hàm mất mát ở công thức (1):

$$h^t = \arg\min_{h \in H} L(F^{t-1} + h) = \arg\min_{h \in H} EL(y, F^{t-1}(x) + h(x)) \quad (1)$$

Bài toán tìm điểm cực tiểu thường được tiếp cận bằng phương pháp Newton sử dụng ước lượng gần đúng bậc hai của hàm $L(F^{t-1} + h^t)$ tại F^{t-1} hoặc bằng cách lấy bước đạo hàm (âm). Cả hai cách đều thuộc loại giảm dần độ dốc đạo hàm. Cụ thể, bước đạo hàm h được chọn sao cho $h^t(x) \approx -g^t(x, y)$, với $-g^t(x, y) := \frac{\partial L(y,s)}{\partial s} \mid_{s = F^{t-1}(x)}$.

Thông thường, ước lượng bình phương cực tiểu sẽ được sử dụng như công thức (2):

$$h^t = arg\ min_{h \in H} E(-g^t(x, y) - h(x))^2 \quad (2)$$

CatBoost [10] là một thuật toán sử dụng tăng cường độ dốc dùng cây quyết định nhị phân làm hàm dự đoán cơ sở. Cây quyết định là mô hình xây dựng bằng phân vùng đệ quy không gian đặc trưng R^m thành nhiều vùng riêng biệt với nhau gọi là các nút cây, dựa trên giá trị của một số đặc trưng chia nhỏ. Đặc trưng thường là các biến nhị phân xác định một số biến x^k vượt ngưỡng t mà $a = \mathbb{1}\{x^k > t\}$, với x^k là biến số hoặc biến nhị phân. Mỗi vùng cuối (lá của cây) được gắn một giá trị, là ước lượng của biến đầu ra y trong vùng được gắn nhãn dự đoán. Khi đó, cây quyết định h có thể viết thành $h(x) = \sum_{j=1}^{J} b_j \mathbb{1}_{\{x \in R_j\}}$, với R_j là các vùng riêng biệt tương ứng với lá của cây.

3.2. Thuật toán tăng cường có thứ tự với biến phân loại

Giả sử ta có mô hình cần học với I cây. Để phần lỗi (residual) $r^{I-1}(x_k, y_k)$ không chuyển dịch, cần có F^{I-1} huấn luyện mà không có mẫu x^k. Vì ta đang cần phần lỗi không bị thiên lệch cho tất cả dữ liệu tập huấn luyện, không một dữ liệu nào sẽ được sử dụng để huấn luyện F^{I-1} [10], điều khiển cho quá trình huấn luyện dường như bất khả thi ngay từ đầu. Tuy nhiên, việc duy trì một hệ các mô hình khác nhau theo dữ liệu dùng huấn luyện là hoàn toàn có thể. Khi đó, để tính phần lỗi cho một dữ liệu, có thể tiến hành huấn luyện mô hình loại bỏ lỗi bằng nguyên tắc thứ tự. Để minh họa cho ý tưởng này, giả sử ta có một hoán vị ngẫu nhiên σ của tập dữ liệu huấn

luyện và duy trì n mô hình hỗ trợ M_1, ..., M_n khác nhau sao cho mô hình M_i được học chỉ sử dụng duy nhất i dữ liệu đầu tiên trong hoán vị. Tại mỗi bước, để có được lỗi của mẫu thứ j, mô hình M_{j-1} sẽ được sử dụng. Giải thuật 1 (pseudo-code) [10] minh họa dưới đây được gọi là tăng cường có thứ tự.

ORDERED BOOSTING [10]

input : $\{(x_k,\ y_k)\}_{k=1}^{n}$, I;

$\sigma \leftarrow$ *random permutation of [1, n];*

$M_i \leftarrow 0$ *for* $i = 1..n$;

for $t \leftarrow 1$ ***to*** I ***do***

  for $i \leftarrow 1$ ***to*** n ***do***

    $r_i \leftarrow y_i - M_{\sigma(i)-1}(x_i)$;

  for $i \leftarrow 1$ *to* n *do*

    $\Delta M \leftarrow LearnModel((x_j , r_j): \sigma(j) \leq i)$;

    $M_i \leftarrow M_i + \Delta M$;

  return M_n

Tuy nhiên, thuật toán này không khả thi trong hầu hết các nhiệm vụ thực tế do cần phải huấn luyện n mô hình khác nhau, làm tăng độ phức tạp và bộ nhớ không cần thiết lên n lần. Do đó, thuật toán CatBoost đã được chỉnh sửa để triển khai trên cơ sở thuật toán tăng độ dốc với cây quyết định (GBDT) là các yếu tố dự báo cơ sở.

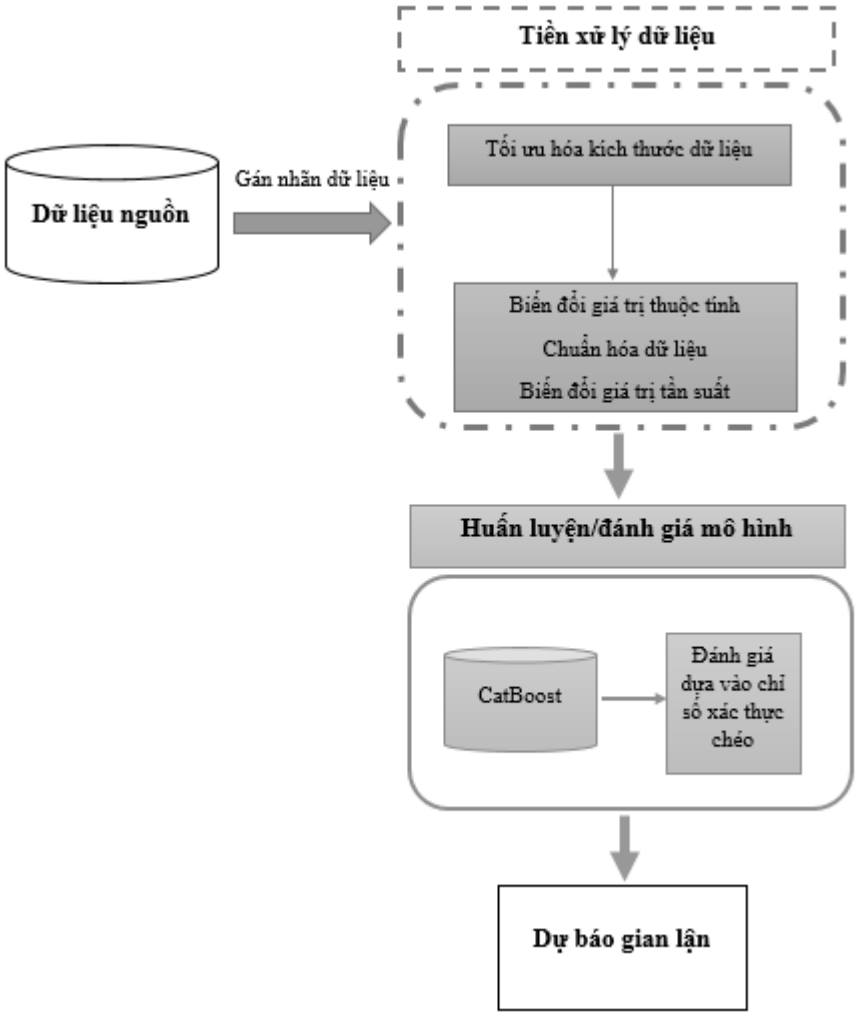
Tăng cường có thứ tự với các biến phân loại là việc sử dụng các hoán vị ngẫu nhiên σ_{cat} và σ_{boost} của tập huấn luyện sao cho $\sigma_{cat} = \sigma_{boost}$ để tránh việc chuyển dịch dự đoán xảy ra. Điều này đảm bảo biến mục tiêu y_i sẽ không được sử dụng để huấn luyện M_i.

4. MÔ HÌNH ĐỀ XUẤT

4.1. Tổng quan về bản chất thuật toán CatBoost

Hình 1. mô tả cách tiếp cận của nghiên cứu trong phát hiện hành vi gian lận bằng thuật toán CatBoost. Quá trình bắt đầu với tối ưu hóa kích thước dữ liệu nhằm giảm mức sử dụng bộ nhớ. Sau đó, phân tích khám phá được thực hiện để kiểm tra dữ liệu về các mẫu, xu hướng hoặc mối liên hệ giữa các biến và giữa biến mục tiêu với các biến khác. Trong bước này, có thể thấy rằng dữ liệu trong tập huấn luyện và tập thử nghiệm có các tập khách hàng khác nhau trong mỗi dữ liệu. Nhóm tác giả thực hiện đánh giá chéo (k-fold cross-validation) phân chia theo thời gian hàng tháng để tạo đặc trưng về định danh người dùng. Thực tế quan trọng ở đây là chúng ta có sự chồng chéo nhận dạng người dùng giữa các fold. Nhận dạng người dùng chồng chéo nhiều hơn trong khoảng thời gian gần nhau hơn. Ví dụ: Fold0 (Giáng sinh) và fold cuối cùng hoạt động kém hơn những fold ở giữa, vì có nhiều nhận dạng người dùng hơn chồng lên các fold huấn luyện. Ngoài ra, một số kỹ thuật khai thác dữ liệu như biến đổi đặc trưng, lựa chọn đặc

trưng, xây dựng mới đặc trưng cũng được áp dụng riêng cho từng phần chồng chéo và không trùng lặp. Chi tiết cho từng bước được làm rõ trong các phần tiếp theo.



Hình 7. Mô hình nghiên cứu đề xuất (Nguồn: Tác giả)

4.2. Nguồn dữ liệu

Để xây dựng được một mô hình tốt cho bất kỳ mô hình Học máy nào nhằm đưa ra kết quả phù hợp với thực tế, tập dữ liệu phức tạp là điều rất cần thiết. Do đó, nhóm tác giả đã chọn tập dữ liệu IEEE-CIS [7] vì bao gồm nhiều đặc điểm điển hình cho các vấn đề thực tế ảnh hưởng đến phát hiện gian lận, bao gồm khối lượng dữ liệu lớn, tổng giao dịch hợp pháp nhiều hơn gấp nhiều lần các giao dịch gian lận và đặc trưng có liên quan đến thẻ như các biến thời gian, số tiền giao dịch, địa chỉ, thông tin kết nối mạng (IP, ISP, Proxy,…) và chữ ký số (UA / browser / os / version,…) được liên kết với các giao dịch. Tập dữ liệu bao gồm bao gồm tổng cộng 433 cột và 590.540 dòng là các giao dịch thực tế được cung cấp bởi Vesta Corporation, công ty chuyên về các giải pháp thanh toán trong lĩnh vực thương mại điện tử.

Phương thức được sử dụng là phân loại nhị phân, theo chủ sở hữu tập dữ liệu, giao dịch được ký hiệu là “isFraud = 1” khi có báo cáo hoàn trả trên thẻ và tất cả các giao dịch sau đó có liên

kết với tài khoản người dùng, địa chỉ email và các thông tin liên quan khác, cũng được gán nhãn là gian lận. Nếu chủ thẻ không báo cáo trong vòng 120 ngày thì những giao dịch bị nghi ngờ là gian lận sẽ tự động được coi là hợp pháp (isFraud = 0). Nói cách khác, khi một thẻ bị báo cáo là gian lận, tài khoản đó sẽ được chuyển đổi thành isFraud = 1.

Tập dữ liệu bao gồm tập giao dịch và tập định danh người dùng, mỗi loại được chia thành tập huấn luyện và tập thử nghiệm. Điểm đáng chú ý của tập dữ liệu là biến mục tiêu, isFraud, được biểu diễn dưới dạng nhị phân, chỉ tồn tại trong tập huấn luyện, có nghĩa là mô hình nên được chia nhỏ để mang lại kết quả đánh giá chính xác hơn. Cả tập giao dịch và tập định danh sẽ được liên kết với nhau qua TransactionID. Để có thể hiểu hơn về tập dữ liệu, **Bảng 1.** mô tả dữ liệu dựa trên giải thích của Vesta được đưa ra như sau:

Bảng 1. Mô tả dữ liệu

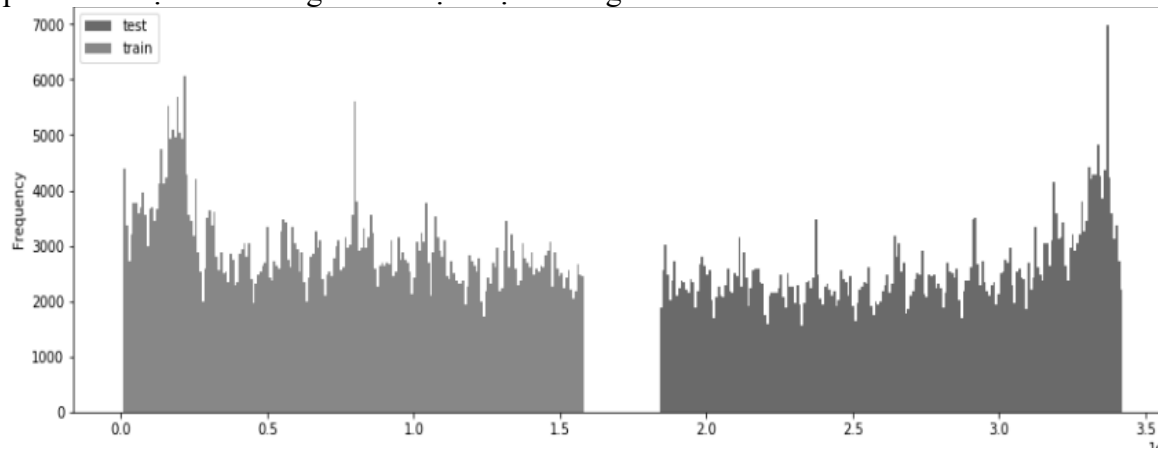
TRANSACTION	
ĐẶC TRƯNG	MÔ TẢ
TransactionDT	Khoảng thời gian giữa hai khung thời gian tham chiếu
TransactionAMT	Số tiền thanh toán giao dịch bằng USD
ProductCD	Mã sản phẩm, sản phẩm cho từng giao dịch
Card1 - Card6	Thông tin thẻ được dung thanh toán. Chẳng hạn như loại thẻ thanh toán, ngân hàng phát hành, quốc gia và các thông tin khác.
Addr	Địa chỉ
Dist	Khoảng cách giữa bên bán và bên mua
C1-C14	Biến đếm. Chẳng hạn như có bao nhiêu địa chỉ được tìm thấy có liên quan đến thẻ thanh toán và các biến khác.
D1-D15	Khoảng thời gian giữa hai thời điểm. Chẳng hạn như số ngày giữa hai giao dịch gần nhất và các khoảng thời gian khác
M1-M9	Mã đối chiếu, chẳng hạn như tên trên thẻ và địa chỉ, và các thông tin khác
Vxxx	Vesta đã tạo thêm các đặc trưng và mã hóa nó
IDENTITY	
DeviceType	Loại thiết bị được khách hàng sử dụng khi giao dịch
DeviceInfo	Thông tin thiết bị đươc sử dụng

Id_01 - Id_11	Các đặc điểm nhận dạng bằng số, chẳng hạn như xếp hạng thiết bị, xếp hạng tên miền ip, xếp hạng proxy, số lần đăng nhập tài khoản giống như dấu vân tay hành vi / lần đăng nhập không thành công, thời gian tài khoản ở lại trên trang và các đặc điểm khác

4.3. Tiền xử lý dữ liệu

4.3.1. Phân tích dữ liệu khám phá

Khoảng 3,5% các giao dịch trên tập huấn luyện là gian lận và phần lớn các cột có giá trị bị thiếu. Ngoài ra, có các cột chỉ có một giá trị duy nhất hoặc rỗng, và rất nhiều biến liên tục và một số biến phân loại. Một đặc điểm rất quan trọng của tập dữ liệu là tập huấn luyện và tập thử nghiệm được phân tách dựa trên chuỗi thời gian. Khoảng thời gian trong tập thử nghiệm và tập huấn luyện cách nhau một tháng, theo đó tập huấn luyện gồm những dữ liệu trong giai đoạn đầu của chuỗi thời gian và tập thử nghiệm gồm những dữ liệu trong giai đoạn cuối của chuỗi thời gian, điều này đã được thể hiện trong **Hình 2**. Với đặc điểm trên, nhóm tác giả đã tiến hành phân tách dựa trên thời gian để thực hiện đánh giá chéo.



Hình 8. Đồ thị phân phối tần suất của TransactionDT trong tập huấn luyện và thử nghiệm

Nguồn: Tác giả

4.3.2. Trích chọn đặc trưng

Bước này liên quan đến việc tạo các đặc trưng mới, bao gồm nhận dạng người dùng (user identification - UID) và các đặc trưng dạng số dựa trên UID. Trong phát hiện gian lận, trạng thái người dùng rất quan trong vì mô hình dự đoán của người dùng cũ không thể áp dụng cho người dùng mới. Với đặc điểm rằng tập dữ liệu này được kết hợp giữa người dùng mới và cũ, nhóm tác giả tạo thêm một số loại UID dựa trên 6 thẻ và cột địa chỉ, cụ thể là UID1 (dựa trên Card1 và Card2), UID2 (dựa trên UID1, Card3 và Card5), UID3 (dựa trên UID2, Addr1 và Addr2), UID4 (dựa trên UID3 và P_emaildomain), UID5 (dựa trên UID4 và R_emaildomain). Đối với tổng hợp UID, nó thực sự là danh sách 5 loại UID được nhóm theo giá trị trung bình và độ lệch chuẩn.

4.3.3. Lựa chọn đặc trưng

Chọn đúng bộ đặc trưng làm đầu vào cho mô hình là một trong những đóng góp cực kỳ quan trọng giúp đạt hiệu quả cao nghiên cứu. Đối với CatBoost, nghiên cứu bắt đầu với kỹ thuật mã hóa tần số (frequency encoding) hầu hết tất cả các đặc trưng, sau đó chọn ra đặc trưng bằng cách sử dụng LightGBM và đánh giá chéo (cross-validation). Quá trình được lặp lại để loại bỏ các đặc trưng cho đến khi danh sách đặc trưng không giúp hiệu quả mô hình tăng thêm.

4.4. Mô hình

Mô hình được đề xuất xây dựng dựa trên thuật toán CatBoost, dự đoán về các phần chồng chéo và không chồng chéo tương ứng sẽ được kết hợp thành một đầu ra duy nhất. Nhóm tác giả sử dụng CatBoost để xem có thể cải thiện tỷ lệ dự đoán cho những người dùng trùng lặp hay không. Hai lợi ích của CatBoost là xử lý tự động các biến phân loại với hiệu suất mạnh mẽ hơn so với cách triển khai GBDT khác. Do đó, đối với một tập hợp dữ liệu có nhiều biến phân loại và có dữ liệu như bộ dữ liệu IEEE-CIS, kết quả huấn luyện được cải thiện mà không cần chuyển biến phân loại thành số. CatBoost mạnh mẽ vì không cần tinh chỉnh tham số [9]. Với những ưu điểm trên cho thấy CatBoost cho kết quả tốt về cả về tốc độ và độ chính xác. **Bảng 2.** Trình bày tham số trong mô hình CatBoost.

Bảng 2. Điều chỉnh tham số trong mô hình CatBoost

THAM SỐ	*MÔ TẢ THAM SỐ*	*GIÁ TRỊ*
learning_rate	Được sử dụng để giảm bước gradient	0,07
loss_function	Số liệu được sử dụng để huấn luyện	Log-loss
depth	Depth of the tree	8

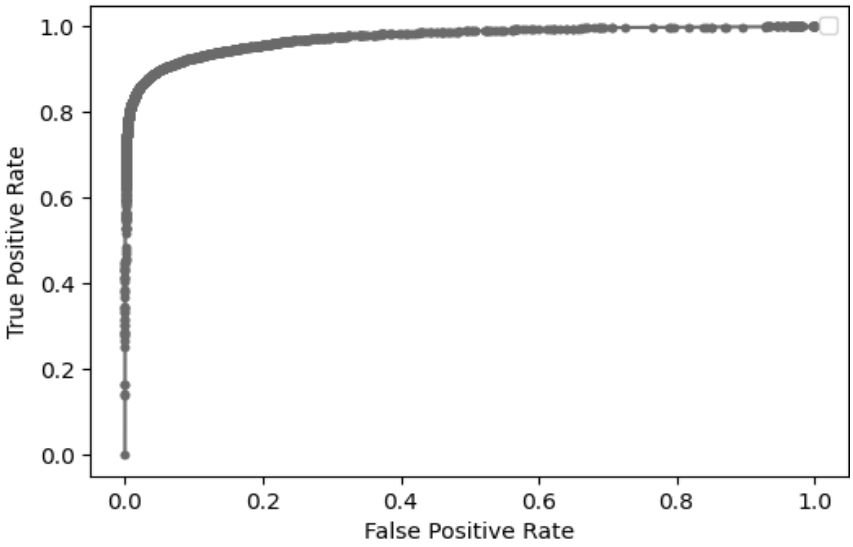
5. KẾT QUẢ THỬ NGHIỆM VÀ THẢO LUẬN

Trong phần này, nghiên cứu sẽ đề xuất xây dựng mô hình CatBoost giúp xử lý hai loại dữ liệu: mới và cũ của người dùng. Giá trị dự đoán của mô hình được biểu diễn dưới dạng xác suất có giá trị trong [0, 1] với xác suất bằng 0 ứng với người dùng không gian lận, và ngược lại bằng 1 ứng với người dùng gian lận. Kết quả được biểu diễn dưới dạng đồ thị phân phối tần suất như **Hình 3.** dưới đây

Hình 3. Đồ thị phân phối tần suất khả năng gian lận thẻ tín dụng (Nguồn: Tác giả)

Đồ thị phân phối như **Hình 3**. cho thấy nếu người dùng có xác suất ở miền giá trị trong đoạn [0,2, 1] càng cao thì khả năng họ có gian lận là càng lớn. Tuy nhiên vì kết quả dự đoán đang ở dạng xác suất với miền giá trị trong đoạn [0, 1], để có thể gửi cảnh báo về gian lận thẻ, kết quả dự đoán cần phải được chuyển thành dạng isFraud = 1 hoặc isFraud = 0, bằng cách xác định một ngưỡng mà tại đó giao dịch được gán nhãn là gian lận. Cụ thể đối với kết quả thu được từ **Hình 3**., ngưỡng xác định phải là một số lớn hơn hoặc bằng 0,2 và gần với 0,2 nhất có thể đối với kết quả ở mỗi tiêu chí. Để tìm ra ngưỡng này, nhóm tác giả thực hiện thông qua đánh giá hiệu quả mô hình với 2 tiêu chí như sau:

Tiêu chí 1: Dựa trên ROC – AUCBan đầu, đánh giá mô hình thông qua việc so sánh kết quả thực (ở dạng phân loại 0, 1) và kết quả dự đoán (ở dạng xác suất) được biểu diễn như **Hình 4.**



Hình 4. Diện tích dưới đường cong ROC (Nguồn: Tác giả)

Bảng 3. Độ chính xác của mô hình tương ứng với các ngưỡng chọn

NGƯỠNG	*ĐỘ CHÍNH XÁC*	*AUC*
0,20	0,986	0,898

0,21	0,987	0,897
0,22	0,987	0,895
0,23	0,987	0,894
0,24	0,987	0,892
0,25	0,988	0,891
0,26	0,988	0,889
0,27	0,988	0,888
0,28	0,988	0,886
0,29	0,988	0,884
0,30	0,988	0,882

Dựa vào **Bảng 3.**, ngưỡng xác định là 0,25, nghĩa là bất kỳ giao dịch nào có xác suất dự đoán lớn hơn hoặc bằng 0,25 sẽ được cho là gian lận và các biện pháp cảnh báo, ngăn chặn sẽ được kích hoạt. Sở dĩ, nhóm tác giả chọn ngưỡng 0,25 vì tại đây khả năng dự đoán của mô hình cho kết quả với độ chính xác cao nhất và đồng thời AUC vẫn tương đối cao. Tiêu chí độ chính xác thường được sử dụng trong các bài toán nhằm mục đích phân loại vì giúp phản ánh sự hiệu quả của mô hình một cách trực tiếp bằng việc chỉ ra số trường hợp dự đoán đúng trên tổng số trường hợp xảy ra.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

Tuy nhiên chỉ số này tỏ ra kém hiệu quả đối với trường hợp mẫu mất cân bằng nghiêm trọng như tập dữ liệu đang được sử dụng xây dựng mô hình. Bởi lẽ, kết quả dự báo của mô hình sẽ bị thiên lệch về phía lớp đa số do số trường hợp giao dịch thẻ tín dụng hợp pháp là quá nhiều ảnh hưởng đến khả năng dự đoán của mô hình, theo đó mô hình sẽ khả năng cao nhận định toàn bộ giao dịch không có gian lận. Vì thế, để xác định mô hình dự đoán tốt nhằm đưa vào thực tế không nên chỉ dựa vào tiêu chí độ chính xác.

Do dữ liệu có sự mất cân bằng đáng kể giữa giao dịch gian lận và không gian lận như vậy, nhóm tác giả sẽ sử dụng thêm tiêu chí Precision và Recall để đánh giá do nó có những tính chất phù hợp sau:

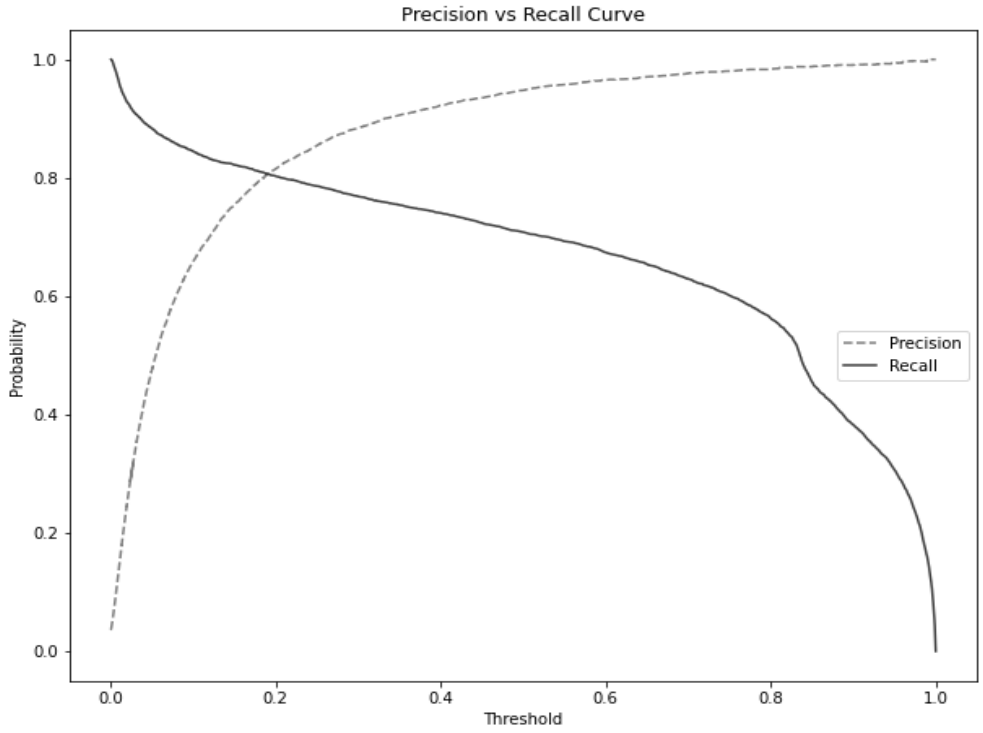
Cả hai đều cho thấy được mức độ dự báo chính xác của gian lận thẻ tín dụng thực sự (True Positive). Đây là nhóm được ưu tiên phân loại chính xác hơn vì thiệt hại gây ra bởi nó lớn hơn.

Precision đánh giá tỷ lệ dự báo chính xác gian lận trong tổng số trường hợp được dự báo là gian lận.

Recall đánh giá tỷ lệ dự báo chính xác gian lận khi giao dịch thẻ tín dụng đó theo bản chất là gian lận thật sự.

Tiêu chí 2: Dựa trên Precision - Recall

Các đường cong Precision và Recall giao nhau tại ngưỡng 0,2 trong **Hình 5.** dưới đây, chứng tỏ mức độ chính xác trong dự đoán gian lận đối với các giao dịch có xác suất lớn hơn hoặc bằng ngưỡng này là rất lớn. Điều này nhất quán với kết quả đồ thị phân phối tần suất tại **Hình 3.**



Hình 5. Đường cong Precision-Recall tương ứng với các ngưỡng (Nguồn: Tác giả)

Tuy nhiên, để có thể phân loại giao dịch hợp pháp hay gian lận, cần dựa vào cả giá trị F1-score, là tham số điều hòa Precision và Recall, giúp đưa ra kết quả tối ưu nhất cho cả hai tham số đánh giá này.

$$\text{F1-score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Bảng 4. Giá trị F1-score tương ứng với các ngưỡng chọn

NGƯỠNG	*F1-SCORE*	*AUC*
0,20	0,809	0,865
0,21	0,812	0,865
0,22	0,814	0,865
0,23	0,816	0,865
0,24	0,817	0,865

0,25	0,819	0,865
0,26	0,821	0,865
0,27	0,822	0,865
0,28	0,822	0,865
0,29	0,823	0,865
0,30	0,822	0,865

Từ **Bảng 4.** cho thấy, ngưỡng tại 0,29 sẽ trả về giá trị F1-score cao nhất (0,823), còn AUC = 0,865 không đổi với mọi ngưỡng.

Kết hợp hai tiêu chí, nhóm tác giả đề xuất chọn ngưỡng tại 0,29 để gán nhãn gian lận cho mọi giao dịch có xác suất cao hơn hoặc bằng giá trị này nhằm đáp ứng mục tiêu dự đoán chính xác, giảm tỷ lệ cảnh bảo sai làm ảnh hưởng đến khách hàng có các giao dịch hợp pháp vì ngưỡng xác định với F1-score lớn hơn ngưỡng xác định với độ chính xác và F1-score không bị tác động mạnh bởi sự mất cân bằng phân lớp như tham số độ chính xác.

6. KẾT LUẬN

Bài báo trình bày những thiệt hại đáng kể gần đây của các hành vi giao dịch gian lận trong ngành tài chính dựa trên CatBoost nhằm nâng cao hiệu quả phát hiện và giải quyết vấn đề liên quan một cách đáng kể. Thông qua phương pháp này, mô hình đã loại bỏ nhiều đặc trưng dư thừa và có khả năng cao có thể làm sai lệch kết quả mô hình trong nhận diện gian lận. Kết quả thực nghiệm có thể thấy rằng, thuật toán CatBoost cho kết quả tốt về cả thời gian thực nghiệm và độ chính xác và khả năng ứng dụng cao vào thực tế nhận diện gian lận. Trong tương lai, nhóm tác giả có kế hoạch tiếp tục cải tiến mô hình để dự báo với tập dữ liệu có thời gian thực kết hợp với các mô hình học sâu (Deep Learning) để mô hình ngày càng phù hợp trong các tình huống thực tế.

Tài liệu tham khảo

Zabihollah Rezaee. (2011). *Financial Statement Fraud – Prevention and Detection* (2nd edition). John Wiley & Sons, Inc., 2

Zareapoor, M., Yang, J. (2017). *A Novel Strategy for Mining Highly Imbalanced Data in Credit Card Transactions*. Intelligent Automation & Soft Computing, 1-7

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., Baesens, B. (2019). *The value of big data for credit scoring: enhancing financial inclusion using mobile phone data and social network analytics*. Applied Soft Computing, Vol. 74, 26-39

S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang and C. Jiang. (2018). *Random Forest for Credit Card Fraud Detection*. 15th IEEE International Conference on Networking, Sensing and Control (ICNSC), 1-6

A. M. Mubalaike and E. Adali. (2018). *Deep Learning Approach for Intelligent Financial Fraud Detection System*. 3rd IEEE International Conference on Computer Science and Engineering (UBMK), 598-603

H. Najadat, O. Altiti, A. A. Aqouleh and M. Younes. (2020). *Credit Card Fraud Detection Based on Machine and Deep Learning*. 11th IEEE International Conference on Information and Communication Systems (ICICS), 204-208

7. Dataset. *IEEE – Fraud Detection*. Kaggle. https://www.kaggle.com/c/ieee-fraud-detection/data

John T. Hancock and Taghi M. Khoshgoftaar. (2020). *CatBoost for big data: an interdisciplinary review*. Journal of Big Data, 7(1)

Prathima G., Sai Tejasri Y., Gayathri Devi D., Vamsi Kaladhar P., and Vegesna Prudhvi R. (2021). *Detection of Credit Card Fraudulent Transactions using Boosting Algorithms*. Journal of Emerging Technologies and Innovative Research (JETIR), 8(2), 2031-2036

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, Andrey Gulin (2019). *CatBoost: unbiased boosting with categorical features*. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)