TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 20, Số 10 (2023): 1746-1760
Website: https://journal.hcmue.edu.vn
ISSN: 2734-9918

HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 20, No. 10 (2023): 1746-1760
https://doi.org/10.54607/hcmue.js.20.10.3906(2023)

Bài báo nghiên cứu

GIẢI THÍCH ĐẶC TRƯNG THẺ TÍN DỤNG THEO PHƯƠNG PHÁP LIME VÀ SHAP SAU GIAI ĐOẠN HỌC SÂU

Nguyễn Quốc Huy**, *Từ Lãng Phiêu
Trường Đại học Sài Gòn, Việt Nam
**Tác giả liên hệ: Nguyễn Quốc Huy – Email: nqhuy@sgu.edu.vn*
Ngày nhận bài: 09-8-2023; ngày nhận bài sửa: 19-10-2023; ngày duyệt đăng: 20-10-2023

TÓM TẮT

Giải thích đặc trưng sau giai đoạn huấn luyện (XAI) là hướng nghiên cứu cần thiết trong thực tiễn ứng dụng AI, và được cộng đồng nghiên cứu AI quan tâm đông đảo. XAI có hai phương pháp tiếp cận phổ biến chính là LIME và SHAP rất hiệu quả trong việc giải thích các mô hình sau khi huấn luyện. Có nhiều tài liệu hướng dẫn sử dụng thư viện LIME và SHAP để giải thích các mô hình sau khi huấn luyện, nhưng rất ít tài liệu đề cập đến mô hình toán học bên trong LIME và SHAP, và điều này gây khó khăn trong việc nghiên cứu. Bài báo mô tả chi tiết các bước thực hiện của LIME và SHAP trên dữ liệu nhỏ, và tiến hành giải thích các đặc trưng sau khi thực nghiệm phân loại trên dữ liệu thẻ tín dụng bằng phương pháp học sâu. Kết quả thực nghiệm và việc giải thích mang lại thông tin khá thú vị khi giải thích đặc trưng theo phương pháp LIME cũng như theo phương pháp SHAP.

Từ khóa: phân lớp thẻ tín dụng; LIME; Học máy; SHAP; XAI

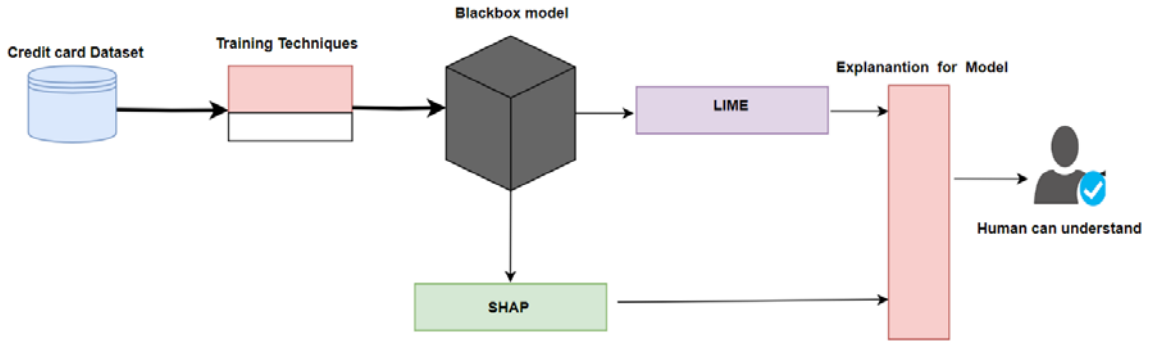
1. Giới thiệu

XAI (giải thích các quyết định từ AI) cung cấp các kĩ thuật, quy trình và hệ thống cho phép con người tin tưởng và tín nhiệm vào việc ra quyết định của máy. Các doanh nghiệp dựa vào công cụ do AI cung cấp, công cụ này giải thích các bằng chứng rõ ràng để hỗ trợ kết quả và giúp cho người dùng thấy rõ tính thuyết phục vào việc ra quyết định của máy. Chỉ với AI, ta chỉ nhận được quyết định sau quá trình học máy. Nhưng với XAI, ta có thể giải thích kết quả quyết định một cách thuyết phục hơn khi biết được lí do ra quyết định. Trong thực tế, khách hàng sẽ vui vẻ đón nhận quyết định và lí do giải thích hợp lí hơn là chỉ nhận kết quả quyết định mà không biết lí do.

Đi sâu vào tìm hiểu XAI (Molnar, 2020; Guidotti et al., 2020) và cách xây một hệ thống AI có độ tín nhiệm cao có khả năng diễn giải lí do ra quyết định của một mô hình, có hai phương pháp phổ biến dùng để giải thích mô hình là LIME và SHAP. Bài báo tập trung vào tìm hiểu XAI (Chen et al., 2018; Zhang et al., 2022) và cách xây dựng các hệ thống AI có độ tin cậy cao, có khả năng giải thích lí do đằng sau quyết định của các mô hình AI. Nó

Cite this article as: Nguyen Quoc Huy, & Tu Lang Phieu (2023). Interpretation of credit card features by LIME and SHAP after deep learning. *Ho Chi Minh City University of Education Journal of Science, 20*(10), 1746-1760.

tập trung vào hai phương pháp phổ biến để giải thích mô hình, đó là LIME (Local Interpretable Model-agnostic Explanations) và SHAP (SHapley Additive exPlanations) (Lundberg et al., 2017; Gramegna & Giudici, 2021). Ngoài ra, bài báo này cung cấp một góc nhìn dễ hiểu về XAI, sử dụng ví dụ về dự đoán khả năng vỡ nợ trong dữ liệu thẻ tín dụng để minh họa cách LIME và SHAP có thể được áp dụng để giải thích kết quả của mô hình AI sau khi dữ liệu này được mô hình học máy dự đoán kết quả bằng phương pháp học sâu (Dighe et al., 2018; Zhang, 2000). Để dễ hình dung, bài báo thực hiện theo các thứ tự công việc như Hình 1.



Hình 1. *Quy trình giải thích mô hình AI*

Phương pháp SHAP (SHApley Additive exPlanations) giải thích kết quả phân lớp cho một trường hợp theo góc độ toàn cục bằng cách tính toán sự đóng góp của từng giá trị đặc trưng mà trường hợp đó thể hiện. Thay vì cung cấp góc nhìn chung về mô hình trên toàn bộ tập dữ liệu như SHAP, LIME (Local Interpretable Model – agnostic Explanations) giải thích cục bộ về mô hình bất khả tri) tập trung vào việc giải thích dự đoán của mô hình cho các trường hợp riêng lẻ.

2. Nội dung nghiên cứu

Phần này phân biệt hai hướng tiếp cận chính trong XAI (Ribeiro et al., 2017) để giải thích những kết quả đã được phân lớp. Giả sử ta có tập quan sát như sau:

Bảng 1. *Bảng dữ liệu quan sát*

STT	X1	X2	X3	Kết quả dự đoán
1	40	50	Xuân	1
2	45	51	Hạ	0
3	47	52	Thu	1
4	49	54	Đông	0
5	50	56	Xuân	1

Trên bảng dữ liệu quan sát này, giá trị của thuộc tính X1 và X2 là giá trị liên tục. Giá trị của thuộc tính X3 là giá trị rời rạc. Phương pháp LIME xử lí trên dữ liệu liên tục và rời rạc là khác nhau, với dữ liệu quan sát này ta có thể thao tác đồng thời trên dữ liệu liên tục và rời rạc.

2.1. Phương pháp LIME

Phương pháp LIME (Ribeiro et al., 2016; Ribeiro & Singh, 2020) là phương pháp giải

thích cục bộ, phương pháp này xấp xỉ thành một mô hình tuyến tính dựa trên vùng dữ liệu lân cận xung quanh cá thể cần được giải thích để xác định vai trò của từng đặc trưng quan trọng của cá thể đó. Công thức xấp xỉ cần thiết của LIME:

$LIME(x_i) = argmin_{g \in G} L(f,g,\pi_x)$ (1)

Trong đó, $\boldsymbol{LIME}(\boldsymbol{x_i})$: là giá trị giải thích của mô hình tại điểm dữ liệu cụ thể $(\boldsymbol{x_i})$, xác định mức độ đóng góp của các đặc trưng vào dự đoán của mô hình tại điểm dữ liệu $(\boldsymbol{x_i})$, $\boldsymbol{argmin}\_\{\boldsymbol{g} \in \boldsymbol{G}\}$: là công thức tối ưu, để tìm kiếm mô hình diễn giải $(\boldsymbol{g})$ trong tập $(\boldsymbol{G})$ sao cho khoảng cách giữa dự đoán của mô hình diễn giải $(\boldsymbol{g})$ và dự đoán của mô hình gốc $(\boldsymbol{f})$ tại điểm dữ liệu $(\boldsymbol{x_i})$ là nhỏ nhất.

(G): là tập hợp các mô hình diễn giải có thể được sử dụng để giải thích mô hình huấn luyện. Thường thì $(\boldsymbol{G})$ chứa các mô hình đơn giản như hồi quy tuyến tính

$(\boldsymbol{f})$: Đây là mô hình gốc mà ta muốn giải thích kết quả của mô hình $(\boldsymbol{f})$

$(\boldsymbol{g})$: Đây là mô hình diễn giải, nó là một thành phần của tập $(\boldsymbol{G})$ và được sử dụng để xấp xỉ dự đoán của mô hình gốc (f) tại điểm dữ liệu $(\boldsymbol{x_i})$

$(\boldsymbol{\pi_x})$: hàm phân phối xác suất trên tập hợp dữ liệu nhiễu xung quanh điểm dữ liệu $(\boldsymbol{x_i})$, tạo ra các mẫu dữ liệu xung quanh $(\boldsymbol{x_i})$để xây dựng mô hình diễn giải $(\boldsymbol{g})$.

Phương pháp làm nhiễu từ dữ liệu x_i theo phương pháp sau:

- Tính giá trị trung bình $(\boldsymbol{\mu})$ và độ lệch chuẩn $(\boldsymbol{\sigma})$ của từng thuộc tính của cá thể x_i trên tập dữ liệu quan sát.
- Tính giá trị nhiễu của xi bằng hàm ngược erfinv F(xi) = $\frac{\mathbf{1}}{\mathbf{2}}$(1+erf($\frac{\boldsymbol{x_i - \mu}}{\boldsymbol{\sigma\sqrt{2}}}$)) Với F(x') =U. Khi đó x' = $\boldsymbol{\mu + \sigma\sqrt{2}x\ erfinv(2U - 1)}$. Với U là giá trị ngẫu nhiên từ phân phối chuẩn[0,1]
- Tính xác suất dự đoán mô hình giải thích theo hướng xấp xỉ. $P(y=0|\ x) = \frac{\mathbf{1}}{\mathbf{1+}\ \boldsymbol{e^{-f(x)}}}$, với $f(x) = w_0 + w_1 x'_1 + w_2 x'_2 + \dots + w_n . x'_n$ là hàm quyết định xấp xỉ. $x'_1, x'_2, \dots, x'_n$ là giá trị nhiễu từ tập quan sát x_i .
- Tính toán độ tương đồng từ tập quan sát và tập nhiễu, để đánh giá mức độ ảnh hưởng tập cá thể trong tập nhiễu:

$w_i = exp(\frac{\boldsymbol{-d_i^2}}{\boldsymbol{2\sigma^2}})$

Trong đó d_i được tính theo cách tính khoảng cách Euclid (Gu et al., 2018).

Dữ liệu quan sát của Bảng 1 (gọi là *X*). Giả sử dữ liệu này đã phân lớp theo kĩ thuật học sâu theo kết quả phân lớp (0,1). Quy trình giải thích cá thể x_2 của tập X theo phương pháp LIME bắt đầu từ việc tạo những cá thể nhiễu quanh cá thể x_2. Thuộc tính X1, X2 là thuộc tính có giá trị liên tục, tạo tập giá trị nhiễu X1', X2' trên đặc trưng X1, X2, ta sử dụng phân phối chuẩn với các thông số như giá trị trung bình $(\boldsymbol{\mu})$ và độ lệch chuẩn $(\boldsymbol{\sigma})$ của đặc trưng X1, X2. Trên cá thể x2 của tập X (dòng số 2 trên Bảng 1) Giả sử muốn tạo 5 dòng nhiễu cho cá thể quan sát X1=45 và X2=51.

- *Bước 1.* Tính trung bình (μ) và độ lệch chuẩn (σ)

$\mu_{X1} = \frac{40+45+47+49+50}{5} = 46.2, \mu_{X2} = \frac{50+51+52+54+56}{5} = 52.6$

$$\sigma_{X1} = \sqrt{\frac{(40-46.2)^2 + (45-46.2)^2 + (47-46.2)^2 + (49-46.2)^2 + (50-46.2)^2}{5}}$$

$$\approx 4.25$$

- *Bước 2.* Tạo các giá trị nhiễu $(X1')$ bằng cách sử dụng hàm ngược của CDF của phân phối chuẩn (inverse CDF hay erfinv). Hàm phân phối tích lũy (CDF) của phân phối chuẩn được cho bởi công thức:

$$F(x) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right). Vậy, F(X1') = \frac{1}{2}\left(1 + \text{erf}\left(\frac{X1'-\mu}{\sigma\sqrt{2}}\right)\right)$$

$$X1' = \mu + \sigma\sqrt{2} \times \text{erfinv}(2U - 1)$$

Giả sử $U = 0.2$, ta tính như sau:

$$X1' = 46.2 + 4.25\sqrt{2} \times erfinv(2 \times 0.2 - 1)$$

Ta được:

$$X1' \approx 46.2 + 4.25\sqrt{2} \times erfinv(-0.6)$$

$$X1' \approx 45.3766$$

Tương tự, sử dụng phân phối chuẩn (normal distribution) để tạo các giá trị nhiễu cho X1 và X2. Với các giá trị $(U_1 = 0.2)$, $(U_2 = 0.5)$, $(U_3 = 0.8)$, $(U_4 = 0.1)$, $(U_5 = 0.9)$ lấy từ phân phối đều. Áp dụng công thức trên với các giá trị (μ_{X1}) và (σ_{X1}) để tính toán giá trị X1' tương ứng với mỗi giá trị (U_i):

$$X1'_1 = 46.2 + \sigma_{X1} \times \sqrt{2} \times \text{erfinv}(2 \times U_1 - 1)$$

$$X1'_2 = 46.2 + \sigma_{X1} \times \sqrt{2} \times \text{erfinv}(2 \times U_2 - 1)$$

$$X1'_3 = 46.2 + \sigma_{X1} \times \sqrt{2} \times \text{erfinv}(2 \times U_3 - 1)$$

$$X\,1'_4 = 46.2 + \sigma_{X1} \times \sqrt{2} \times \text{erfinv}(2 \times U_4 - 1)$$

$$X1'_5 = 46.2 + \sigma_{X1} \times \sqrt{2} \times \text{erfinv}(2 \times U_5 - 1)$$

Riêng đối với X3, vì đây là thuộc tính có giá trị rời rạc với các giá trị {"Xuân", "Hạ", "Thu", "Đông"} ứng với các giá trị {1, 2, 3, 4}, cần sử dụng phân phối đều để tạo các giá trị nhiễu cho X3. Bước thực hiện như sau:

- Tạo 5 giá trị ngẫu nhiên từ phân phối chuẩn cho $X1$ và $X2$ (giả sử ta có $(X1'_1, X1'_2, X1'_3, X1'_4, X1'_5)$ và $(X2'_1, X2'_2, X2'_3, X2'_4, X2'_5)$
- Tạo 5 giá trị ngẫu nhiên từ phân phối đều trong khoảng [0, 1] cho X3 (giả sử ta có $(U_1, U_2, U_3, U_4, U_5)$).

Tạo các số ngẫu nhiên (U_i) từ phân phối đều trong khoảng [0, 1]:

- Khoảng xác định phân chia:

$(0 \leq U < \frac{1}{4}) \rightarrow Xuân$

$(\frac{1}{4} \leq U < \frac{1}{2}) \rightarrow$ Hạ

$(\frac{1}{2} \leq U < \frac{3}{4}) \rightarrow$ Thu

$(\frac{3}{4} \leq U \leq 1) \rightarrow$ Đông

Với $(U_1 = 0.26), (U_2 = 0.85), (U_3 = 0.28), (U_4 = 0.37), (U_5 = 0.45)$, ta có X3' = {Hạ, Đông, Hạ, Hạ, Hạ}. Kết quả tạo được 5 dòng nhiễu dữ liệu cho X1, X2 và X3, với giá trị quan sát X1=45, X2=51 và X3="Hạ", bằng cách sử dụng phân phối chuẩn cho X1 và X2 và phân phối đều cho X3 đã mã hóa nhãn:

Bảng 2. *Dữ liệu sau khi tạo nhiễu từ cá thể x2*

STT	X1'	X2'	X3'
1	45.8	51.15	2
2	45.27	50.82	4
3	44.96	51.35	2
4	44.82	51.12	2
5	44.94	51.54	2

- *Bước 3.*

Dựa trên bảng dữ liệu đã nhiễu, xác suất dự đoán được xác định bởi hàm quyết định của mô hình là (x) , trong đó x là véc tơ của các đặc trưng ($x = [X1', X2', X3']$)).

Đối với phân loại nhị phân, đầu ra của hàm quyết định được chuyển đổi thành xác suất bằng cách sử dụng hàm sigmoid

$$P(y = 0|x) = \frac{1}{1 + e^{-f(x)}}$$

trong đó $\big(P(y = 0|x)\big)$ là xác suất thuộc $class$ 0

hàm quyết định $f(x)$ như sau:

$$f(x) = w_0 + w_1 X1' + w_2 X2' + w_3 X3'$$

Đối với điểm dữ liệu đầu tiên xác suất cho $class$ 0:

$$f(x_1) = w_0 + w_1 \times 45{,}38 + w_2 \times 51{,}15 + w_3 \times 2$$

$$P(y = 0|x_1) = \frac{1}{1 + e^{-f(x_1)}}$$

Đối với điểm dữ liệu đầu tiên xác suất cho $class$ 1:

$$f(x_1) = w_0 + w_1 \times 45.38 + w_2 \times 51.15 + w_3 \times 2$$

$$P(y = 1|x_1) = \frac{1}{1 + e^{-f(x_1)}}$$

Tương tự, ta sẽ tính được cho các giá trị còn lại:

$(P(y = 0|x_3)), (P(y = 0|x_4)), (P(y = 0|x_5))$

Ta có bảng dự đoán (Bảng 3) như sau:

Bảng 3. *Bảng dữ liệu dự đoán tập làm nhiễu*

Dự đoán cho class 0	Dự đoán cho class 1
0.75	0.25
0.65	0.35
0.80	0.20
0.85	0.15
0.70	0.30

- *Bước 4*. Chúng ta sẽ tính toán độ tương đồng của mỗi trường hợp nhiễu với trường hợp mẫu quan sát bằng một hàm tương đồng, dựa theo tính khoảng cách Euclid. Sau đó, áp dụng hàm phân phối Gaussian lên các giá trị đó để chuyển đổi thành các trọng số. Điều này giúp đánh giá mức độ ảnh hưởng của từng trường hợp nhiễu đến dự đoán của mô hình phức tạp. Tính toán trọng số điểm tương đồng (similarity) dựa trên khoảng cách Euclid giữa trường hợp nhiễu và trường hợp tập quan sát:
- Tính khoảng cách Euclid giữa trường hợp tập quan sát và từng trường hợp nhiễu:

  + Khoảng cách giữa trường hợp tập quan sát và trường hợp nhiễu 1:

  $d1 = \sqrt{((45 - 45.38)^2 + (51 - 50.97)^2 + (0 - 0)^2)} = 0.492$

  Tính toán tương tự, ta có:

  + Khoảng cách giữa trường hợp tập quan sát và trường hợp hợp nhiễu 2, 3, 4, 5:

  $d2 = 0.367$, $d3 = 0.216$, $d4 = 0.281$, $d5 = 0.595$

  Kết quả của khoảng cách trường hợp tập quan sát với tập nhiễu như Bảng 4, cột 1:

Bảng 4. *Độ lệch giữa dữ liệu quan sát và nhiễu*

STT	*Lệch giữa quan sát và nhiễu*	Điểm tương đồng
1	0.492	0.792
2	0.367	0.864
3	0.216	0.933
4	0.281	0.899
5	0.595	0.721

- Áp dụng hàm phân phối Gaussian để chuyển đổi các khoảng cách thành điểm tương đồng (similarity score) trong Bảng 4, cột 2:

  + Trọng số (similarity score) cho trường hợp nhiễu của dòng 1:

  $$w1 = exp\left(-d1^2/(2 * \sigma^2)\right) = \exp\left(-\frac{0.492^2}{2 \cdot 0.5^2}\right) \approx 0.792$$

  Tính toán tương tự, ta có:

  + Trọng số (similarity score) cho trường hợp nhiễu của dòng 2, 3, 4, 5:

  $w2 \approx 0.864$, $w3 \approx 0.933$, $w4 \approx 0.899$, $w5 \approx 0.721$

  Khoảng cách Euclide được chuẩn hóa thành giá trị tương đồng trong khoảng từ 0 đến 1. Điểm mẫu dữ liệu có khoảng cách càng gần thì điểm tương đồng càng lớn. Các điểm có khoảng cách xa điểm mẫu sẽ có điểm tương đồng ngày càng nhỏ.

2.2. Phương pháp SHAP

Phương pháp SHAP (Lundberg & Su-In, 2017) là phương pháp giải thích toàn cục,

phương pháp này ứng dụng lí thuyết trò chơi để đánh giá mức độ đóng góp của từng thuộc tính trên toàn dữ liệu quan sát. Cho một hàm giá trị (value function) của trò chơi $(\boldsymbol{v}(\boldsymbol{S}))$, trong đó $(\boldsymbol{S})$ là một tập hợp con của tập hợp các đặc trưng. Giả sử $(\boldsymbol{n})$ là số lượng đặc trưng.

Hàm giá trị Shapley được định nghĩa như sau:

$$[\boldsymbol{\phi_i}(\boldsymbol{v}) = \sum_{\boldsymbol{S}\subseteq \boldsymbol{N}\setminus\{\boldsymbol{i}\}} \frac{|\boldsymbol{S}|!\,(\boldsymbol{n}-|\boldsymbol{S}|-\mathbf{1})!}{\boldsymbol{n}!}[\boldsymbol{v}(\boldsymbol{S}\cup\{\boldsymbol{i}\})-\boldsymbol{v}(\boldsymbol{S})]]$$

trong đó:

- $(\boldsymbol{\phi_i}(\boldsymbol{v}))$ là đóng góp của đặc trưng thứ $(\boldsymbol{i})$ vào hàm giá trị).
- $(\boldsymbol{N})$ là tập hợp gồm $(\boldsymbol{n})$ đặc trưng (từ $\mathbf{1}$, ..., (n)).
- $(\boldsymbol{S})$ là một tập hợp con của $(\boldsymbol{N})$ không chứa đặc trưng $(\boldsymbol{i})$.
- $(|\boldsymbol{S}|$ là số lượng phần tử trong tập hợp (S).

Giả sử ta có một mô hình dự đoán mức độ đóng góp dựa trên ba đặc trưng: X1, X2 và X3. Ta muốn tính đóng góp của từng đặc trưng vào giá trị dự đoán của mô hình. Giá trị giống như Bảng 1 nhưng thay đặc trưng X3 của dòng đầu tiên là 60. Như vậy, ta có một dữ liệu gồm ba đặc trưng (X1, X2, X3) có tổng ba đặc trưng đóng góp bằng 160 (theo phép cộng thông thường là 150). Ta có, bảng dữ liệu như sau:

Bảng 5. *Đóng góp giá trị của các đặc trưng*

Đặc trưng	Năng lực cá nhân	Đóng góp vào nhóm
X1	40	**42.5**
X2	50	**52**
X3	60	**65**
X1,X2	95	95
X1,X3	110	110
X2,X3	120	120
X1, X2, X3	160	160

Quan sát Bảng 5, ta thấy một cá nhân X1 chỉ có năng lực là 40 khi làm việc độc lập, nhưng khi làm nhóm (cụ thể làm với đồng đội X2, X3) thì năng lực đóng góp lại cao hơn so với làm độc lập(42.5 thay vì 40). Để tính được năng lực đóng góp khi làm việc nhóm. Ta gọi giá trị năng lực nhóm với hàm sau: $[\boldsymbol{v}(\boldsymbol{S}) = \text{X1} + \text{X2} + \text{X3}]$. Trong đó,

- Đóng góp của X1 $((\boldsymbol{\phi_1}(\boldsymbol{v}))$: $(\text{X1} + \frac{\mathbf{1}}{\mathbf{2}} \times [(\text{X1} + \text{X2} + \text{X3}) - (\text{X2} + \text{X3})])$
- Đóng góp của X2 $((\boldsymbol{\phi_2}(\boldsymbol{v}))$: $(\text{X3} + \frac{\mathbf{1}}{\mathbf{2}} \times [(\text{X2} + \text{X3}) - (\text{X1} + \text{X2} + \text{X3})])$
- Đóng góp của X3 $((\boldsymbol{\phi_3}(\boldsymbol{v}))$: $(\text{X3} + \frac{\mathbf{1}}{\mathbf{2}} \times [(\text{X3}) - (\text{X1} + \text{X2} + \text{X3})])$

Đây là mô hình tính giá trị Shapley. Giá trị Shapley cho một nhân viên là trung bình có trọng số của tất cả đóng góp cận biên của người chơi trong tất cả các kết hợp có thể có và trọng số là xác suất xảy ra của kết hợp các đặc trưng. Dưới đây, các quy trình tính toán giá trị Shapley.

Bước 1. Giả sử chúng ta biết mức độ đóng góp của từng đặc trưng và mức độ đóng góp của các cặp đặc trưng trong ba đặc trưng như bảng 6 (cột giá trị Năng lực cá nhân).

Bước 2. Đóng góp cận biên và cách tính đóng góp cận biên được mô tả qua ví dụ sau:

X1X2 đóng góp 95, vậy đóng góp cận biên của X2 là: X1X2 – X1 = 95 - 40 = 55
X2X3 đóng góp 120, vậy đóng góp cận biên của X2 là: X2X3 – X3 = 120 - 60 = 60
X2 có năng lực cá nhân = 50

Bước 3. Tính giá trị Shapley (Bảng 6). Giá trị Shapley cho một đặc trưng là trung bình có trọng số của tất cả đóng góp cận biên của đặc trưng trong tất cả các kết hợp có thể có và trọng số là xác suất xảy ra kết hợp.

Bảng 6. *Tính giá trị các đặc trưng đóng góp giá trị cận biên*

Tổ hợp liên minh	**Xác suất xảy ra kết hợp**	**Đóng góp cận biên X1**	**Đóng góp cận biên X2**	**Đóng góp cận biên X3**
X1,X2,X3	*1/6*	*X1= 40*	*X1,X2 – X1 = 95 – 40 =55*	*X1,X2,X3 – X1,X2 = 160 – 95 = 65*
X1,X3,X2	*1/6*	*X1= 40*	*X1,X3,X2 – X1,X3 = 160 - 110 =50*	*X1,X3 – X1 = 110 -40= 70*
X2,X1,X3	*1/6*	*X2,X1 – X2 = 95 -50 = 45*	*X2 = 50*	*X2,X1,X3 – X2,X1 = 160 – 95 = 65*
X2,X3,X1	*1/6*	*X2,X3,X1 – X2,X3 = 160 – 120 = 40*	*X2 = 50*	*X2,X3 – X2 = 120 -50 = 70*
X3,X1,X2	*1/6*	*X3,X1 – X3 = 110 - 60 = 50*	*X3,X1,X2 – X3,X1 = 160 – 110 = 50*	*X3 = 60*
X3,X2,X1	*1/6*	*X3,X2,X1 – X3,X2 = 160 – 120 = 40*	*X3,X2 – X3 = 120 – 60*	*X3= 60*
Shap Value =Trung bình trọng số đóng góp cận biên của (Xác suất xảy ra kết hợp)(Đóng góp cận biên)*		*40*1/6+40*1/6 +45*1/6+40*1/6 +50*1/6+40*1/6 = 42.5*	*=52.5*	*= 65*

2.3. Dữ liệu thực nghiệm

Dữ liệu (https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients), bao gồm 30.000 trường hợp (quan sát) với 24 đặc trưng, một biến phụ thuộc và 23 biến độc lập. Biến phụ thuộc có tên là default.payment.next.month (Có = 1, Không = 0). Dữ liệu này được sử dụng trong khá nhiều bài báo uy tín về phân tích dữ liệu ngân hàng.

Một mô hình mạng neural được sử dụng để phân loại dữ liệu và tất cả công việc được thực hiện trên môi trường Colab của google, và ngôn ngữ nguồn mở và các thư viện như Python, Keras, Tensorflow, thư viện phân tích và trực quan hóa dữ liệu như: matplot, seaborn, pandas, pandas profiling và numpy. Chi tiết về tập dữ liệu được đưa ra trong Bảng 7.

Bảng 7. *Thông tin dữ liệu dùng trong thực nghiệm*

STT	Đặc trưng	Mô tả
1	LIMIT_BAL	Số tiền tín dụng được phép dùng
2	SEX	Giới tính(1=Nam, 2= Nữ)
3	EDUCATION	Học vấn(1= đại học, 2= đang học đại học, 3= đang học cấp ba, 4 = khác)
4	MARRIAGE	Tình trạng hôn nhân(1=kết hôn, 2 = độc thân, 3= tình trạng khác)
5	AGE	Tuổi(năm)
6	PAY_0	Lịch sử thanh toán thẻ cách 6 tháng (-1 = đúng hạn, 1= chậm 1 tháng, 2 = chậm 2 tháng, 8 = chậm 8 tháng, 9 = chậm 9 tháng)
7	PAY_2	Lịch sử thanh toán thẻ cách 5 tháng (phạm vi như PAY_0)
8	PAY_3	Lịch sử thanh toán thẻ cách 4 tháng (phạm vi như PAY_0)
9	PAY_4	Lịch sử thanh toán thẻ cách 3 tháng (phạm vi như PAY_0)
10	PAY_5	Lịch sử thanh toán thẻ cách 2 tháng (phạm vi như PAY_0)
11	PAY_6	Lịch sử thanh toán thẻ cách 1 tháng (phạm vi như PAY_0)
12	BILL_ATM1	Số tiền sao kê hóa đơn cách 6 tháng
13	BILL_ATM2	Số tiền sao kê hóa đơn cách 5 tháng
14	BILL_ATM3	Số tiền sao kê hóa đơn cách 4 tháng
15	BILL_ATM4	Số tiền sao kê hóa đơn cách 3 tháng
16	BILL_ATM5	Số tiền sao kê hóa đơn cách 2 tháng
17	BILL_ATM6	Số tiền sao kê hóa đơn cách 1 tháng
18	PAY_ATM1	Số tiền thanh toán trước đó 6 tháng gần nhất
19	PAY_ATM2	Số tiền thanh toán trước đó 5 tháng gần nhất
20	PAY_ATM3	Số tiền thanh toán trước đó 4 tháng gần nhất
21	PAY_ATM4	Số tiền thanh toán trước đó 3 tháng gần nhất
22	PAY_ATM5	Số tiền thanh toán trước đó 2 tháng gần nhất
23	PAY_ATM6	Số tiền thanh toán trước đó 1 tháng gần nhất

2.4. Kĩ thuật thực nghiệm

- Dữ liệu đầu vào: một file.CSV.
- Lớp: Lớp đầu vào, lớp fully-connected, lớp đầu ra.
- Hàm mất mát: Đo độ khác biệt giữa giá trị dự đoán và giá trị thực tế.
- Tối ưu hóa: quá trình huấn luyện là quá trình tìm tập hợp trọng số và độ lệch tốt nhất nhằm cực tiểu hóa hàm mất mát.

Kiểm tra dữ liệu của mô hình trước khi đưa vào huấn luyện: Nếu kiểu dữ liệu không phù hợp với các đối tượng thì cần được chuyển đổi. Trong tập dữ liệu, biến MARRIAGE và EDUCATION được chuyển đổi thành biến giả trước khi biến giả có 23 biến độc lập hiện được mở rộng lên tới 28 biến, phương thức pandas get dummies được sử dụng để tạo các biến giả. Để kiểm thử dữ liệu chia 75% cho huấn luyện và 25 %. Thư viện python ScikitLearn được sử dụng để phân tách dữ liệu thành mục đích huấn luyện và thử nghiệm bằng chức năng "train_test_split".

- Tham số đầu tiên là dữ liệu huấn luyện được cung cấp 'X_train' 'y_train'.
- Tham số thứ ba là '*batch_size'* được sử dụng để cập nhật trọng số sau một số quan sát nhất định mà chúng ta có thể sử dụng bằng cách quan sát tổng số.
- Tham số thứ tư là *epochs* là số lượng bạn muốn huấn luyện mô hình.

- Để tránh hiện tượng quá khớp hoặc không khớp khi huấn luyện trên tập dữ liệu nhóm chúng tôi phân thành 2 tập khác nhau:
- Mỗi tập huấn luyện chúng tôi sử dụng trong 80% được trộn lẫn và chọn ngẫu nhiên để xây dựng mô hình cấu trúc và xác định thuật toán tối ưu.
- Tập dữ liệu được chọn để kiểm tra 20% trong tập dữ liệu được chọn.

Trong Bảng 8 có 3 tham số, Layer(type), Output Shape, Param:

Layer(typer): Là kiểu kết nối lớp fully connected

Output Shape (số chiều đầu ra) với dense_3 có output shape là (None, 12), trong đó có 12 neural và "None" là đại diện cho số lượng mẫu tham gia huấn luyện:

- Lớp đầu tiên là một lớp Dense với 12 neuron, có tổng cộng 288 tham số.
- Lớp thứ hai là một lớp Dense với 12 neuron, có tổng cộng 156 tham số.
- Lớp thứ ba là một lớp Dense với 1 neuron, có tổng cộng 13 tham số.

Tổng số tham số cần học trong mô hình là 457

Bảng 8. *Kiến trúc mô hình huấn luyện mạng neural network*

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_3 (Dense)             (None, 12)                288

 dense_4 (Dense)             (None, 12)                156

 dense_5 (Dense)             (None, 1)                 13

=================================================================
Total params: 457
Trainable params: 457
Non-trainable params: 0
_________________________________________________________________
```

Sau khi huấn luyện, độ chính xác của mô hình được mô tả trong Bảng 9.

- *Phương pháp đánh giá*

Bài báo sử dụng độ đo Precision, Recall, F1-Score để đánh giá mô hình: Việc phân lớp được xác định {0 – khách hàng không thanh toán thẻ tín dụng trong tháng tới, 1 – khách hàng có thanh toán thẻ tín dụng trong tháng tới}. Theo Bảng 9 thì có 84% trường hợp dự đoán đúng, Precision cao cho lớp 0 (0.84) cho biết có khá nhiều các mẫu được dự đoán đúng là thuộc lớp 0. Recall thấp cho lớp 1 (0.34) có thể cho thấy rằng mô hình có thể bỏ sót một số mẫu thực sự thuộc lớp 1. Tuy nhiên giá trị F1 vẫn cao cho kết quả dự đoán ở lớp 0. Tổng số lượng khách hàng không thanh toán thẻ tín dụng trong tháng tới của tập dữ liệu là 4703, trong khi lớp còn lại thì số lượng là 1297.

Bảng 9. *Các giá trị của ma trận lỗi sau khi học sâu*

	Precision	**Recall**	**F1-score**	**Support**
0	**0.84**	**0.96**	**0.90**	**4703**
1	**0.70**	**0.34**	**0.46**	**1297**

2.5. *Phương pháp LIME với tập dữ liệu mẫu*

Chọn một mẫu dữ liệu (hoặc điểm dữ liệu) là một thực thể từ tập dữ liệu huấn luyện hoặc làm điểm dữ liệu cần giải thích, như Hình 2 dưới đây:

ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	...	BILL_AMT6	PAY_AMT1	PAY_AMT2	...	PAY_AMT6	default.payment.next.month
19286	160000	2	2	2	32	0	0	0	...	7072	1222	1147	...	766	1

Hình 2. *Thông tin một cá thể trong tập huấn luyện*

Thông tin chi tiết dữ liệu của khách hàng có hồ sơ 19286 như trong Hình 2 được giải thích như sau:

Với LIMIT_BAL = 160000, số tiền tín dụng cho phép khá cao, nếu người dùng không thanh toán đầy đủ số tiền này thì có khả năng họ sẽ thanh toán trễ hẹn

AGE = 32 tuổi. Đối với độ tuổi này do còn trẻ nên khả năng sẽ ảnh hưởng đến khả năng quản lí tài chính và việc thanh toán.

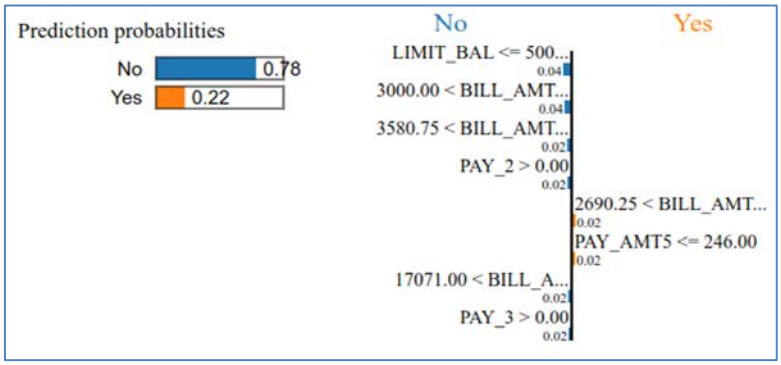
(PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6) đều có giá trị là 0 nên lịch sử thanh toán trong các tháng trước được thực hiện đúng hạn. Tuy nhiên, việc thanh toán đúng hạn cũng không dự đoán được tương lai khách hàng sẽ không thanh toán trễ hẹn.

Từ BILL_ATM1 đến BILL_ATM6 cho thấy người này đã dùng thẻ để thanh toán hóa đơn liên tục và số tiền trả cho hóa đơn tăng dần theo các tháng cho thấy có thể dẫn đến áp lực thanh toán thẻ tín dụng trong tương lai.

Số tiền thanh toán PAY_ATM1 = 1222, PAY_ATM2 = 1147, PAY_ATM3=2000, PAY_ATM4=2000 và PAY_ATM6 = 766 là rất thấp cho thấy khả năng thanh toán trễ hẹn của tháng tiếp theo.

Tóm lại, từ dữ liệu của khách hàng của hồ sơ khách hàng 19286, cho thấy có một số đặc trưng tác động tới tình hình tài chính, tuổi tác và hành vi thanh toán trong quá khứ cho thấy có khả năng trễ hạn trong thanh toán thẻ tín dụng tháng tiếp theo.

2.6. *Kết quả thực nghiệm với phương pháp SHAP*

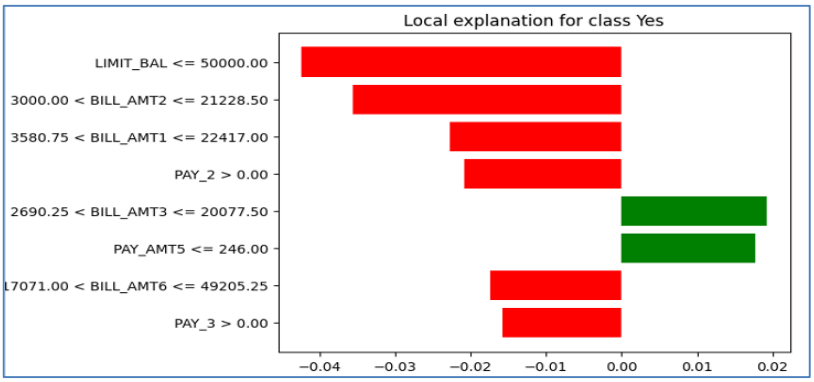


Hình 3. *Kết quả thực nghiệm với LIME*

Từ Hình 3 cho thấy kết quả dự đoán là 78% cho thấy khách hàng sẽ không thanh toán thẻ tín dụng đúng hạn. Chỉ có 22% còn lại khách hàng sẽ thanh toán đúng hạn.

Phần màu xanh bên trái cho thấy các đặc trưng tác động đến kết quả dự đoán 78%.

Phần màu cam bên phải cho thấy chỉ có 2 đặc trưng BILL_AMT3 và PAY_ATM5 tác động đến kết quả 22%.

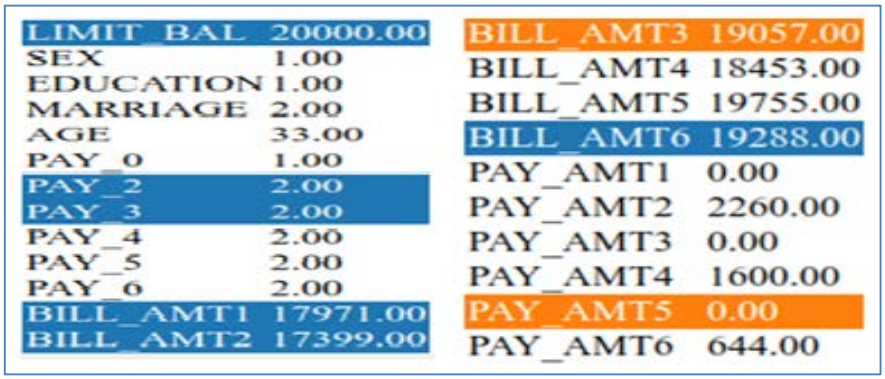


Hình 4. *Mức độ ảnh hưởng đặc trưng*

Trong Hình 4 mô tả thị chi tiết mức độ ảnh hưởng đến kết quả dự đoán của Hình 3. Kết quả trên cho thấy rằng các yếu tố đóng góp đến kết quả như sau:

Mức độ ảnh hưởng đến kết quả dự đoán 78%, khách hàng không thanh toán thẻ tín dụng trong tháng tiếp theo có đặc trưng LIMIT_BAL tác động cao nhất, PAY_3 tác động thấp nhất.

Mức độ ảnh hưởng đến kết quả dự đoán 22%, khách hàng không thanh toán thẻ tín dụng trong tháng tiếp theo có đặc trưng BILL_AMT3 và PAY_ATM5 tác động gần như nhau.



Hình 5. *Đặc trưng ảnh hưởng phân lớp*

Trong Hình 5 cho thấy các đặc trưng được tô màu xanh tác động đến kết quả dự đoán 78% khách hàng sẽ không có khả năng thanh toán thẻ tín dụng đúng hạng, trạng thái hoàn trả trong các tháng PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6 đều bằng 2 (trả chậm 2 tháng). Điều này cho thấy khách hàng thanh toán thẻ tín dụng bị trì hoãn.

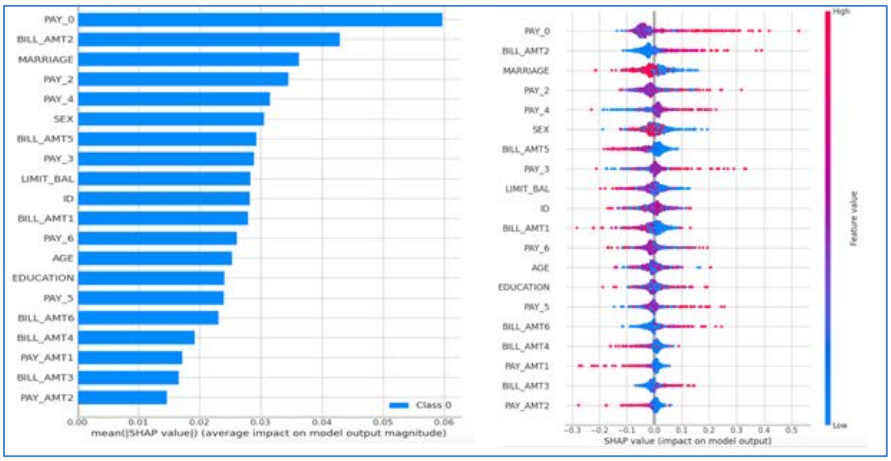
Ngoài ra, đặc trưng PAY_AMT3 (Số tiền thanh toán trước 4 tháng gần nhất) cũng bằng 0 cho thấy khách hàng không thanh toán trả nợ cho tháng trước đó trong các trạng thái hoàn trả trước đó từ PAY_0, PAY_2 đến PAY_6 đã trả chậm 2 tháng, cho thấy khách hàng có áp lực tài chính.

Số tiền sao kê tháng BILL_AMT1, BILL_AMT3, BILL_AMT5 đều tăng, cho thấy khách hàng đang sử dụng thẻ tín dụng để thanh toán liên tục, cho thấy khả năng khách hàng

đang ở tình trạng sử dụng tài chính khá cao dẫn đến khả năng chậm thanh toán cũng cao.

Đặc trưng LIMIT_BAL cao có nghĩa khả năng thanh toán trễ hẹn cũng cao. Vì vậy, khả năng thanh toán thẻ tín dụng tháng tới sẽ ở mức thấp.

Đặc trưng được tô màu xanh tác động đến kết quả dự đoán 22% khách hàng sẽ có khả năng thanh toán thẻ tín dụng đúng hạng, chỉ có 2 đặc trưng BILL_ATM3 = 19057(số tiền thanh toán hóa đơn có giá trị rất cao) và PAY_ATM5 =0 (khách hàng không thanh toán số tiền thanh toán trước đó 2 tháng). Tuy nhiên, chỉ có 2 đặc trưng nên mức độ đóng góp vào kết quả dự đoán cho kết quả 20% khách hàng có khả năng thanh toán thẻ tín dụng rất thấp.



Hình 6. *Mức độ quan trọng của đặc trưng đóng góp kết quả đầu ra*

Từ Hình 6 hiển thị giải thích chung và đưa ra tổng quan về tổng độ lớn SHAP của các đặc trưng được sắp xếp theo tổng giá trị SHAP trên tất cả tập mẫu. Trong đó màu xanh cho biết giá trị đặc trưng ở mức thấp và màu đỏ giá trị đặc trưng ở mức cao.

Trong hình cho thấy các đặc trưng MARRIAGE**,** PAY_0, BILL_ATM1, PAY_0, PAY_2, PAY_5, PAY_6, BILL_ATM6 đóng góp kết quả dự đoán thấp hơn các đặc trưng còn lại trong bảng dữ liệu.

3. Kết luận

Ưu điểm của phương pháp LIME, đó là tính độc lập với mô hình. LIME không phụ thuộc vào bản chất của mô hình. Hoạt động được trên mọi mô hình máy học. Tuy nhiên, phương pháp LIME cũng có một số hạn chế như: Đầu tiên, giải thích LIME không ổn định vì nó phụ thuộc vào rất nhiều thông số như độ rộng nhân, số tính năng, số mẫu, độ rời rạc. Tiếp theo, LIME là một kĩ thuật post hoc dựa trên giả định rằng mô hình phải tuyến tính cục bộ. Nếu mô hình quá phức tạp và không tuyến tính cục bộ, thì giải thích cục bộ có thể không chính xác. Vì vậy, bạn cần tìm một kích thước của một điểm đang xét đến phải rộng. Cuối cùng, nếu lặp lại quy trình lấy mẫu, thì những lời giải thích đưa ra có thể khác. Vì mỗi lần LIME có thể lấy mẫu các quan sát khác nhau. Loại không ổn định này khiến người được giải thích khó tin vào những lời giải thích.

Đối với phương pháp SHAP thì có những ưu điểm như: SHAP có nền tảng dựa trên lí thuyết trò chơi. Dự đoán được phân phối công bằng giữa các giá trị đặc trưng. Thu được các

giải thích trái ngược nhau, so sánh dự đoán với trung bình dự đoán SHAP triển khai theo mô hình dựa trên cây. Diễn giải toàn cục gồm mức quan trọng của đặc trưng, tính phụ thuộc đặc trưng. Tuy nhiên, phương pháp này tốn khá nhiều chi phí tính toán, vì vậy không khả thi việc tính nhiều các giá trị Shapley cần thiết cho các diễn giải mô hình toàn cục.

Tích hợp với các Hệ thống để đưa ra quyết định: mục tiêu của XAI là cho phép con người ra quyết định trong vòng lặp. Nghiên cứu trong tương lai có thể tập trung vào việc tích hợp các giải thích LIME và SHAP vào các hệ thống ra quyết định để trao quyền cho người dùng đưa ra các quyết định sáng suốt và có trách nhiệm dựa trên các dự đoán của mô hình. Sẽ tiếp tục thực nghiệm với các dữ liệu về hình ảnh, để đánh giá giữa dữ liệu văn bản và dữ liệu hình ảnh.

❖ ***Tuyên bố về quyền lợi:*** *Các tác giả xác nhận hoàn toàn không có xung đột về quyền lợi.*

TÀI LIỆU THAM KHẢO

Asha, R. B., & Suresh Kumar, K. R. (2021). Credit card fraud detection using artificial neural network. *Global Transitions Proceedings, 2*(1), 35-41. https://doi.org/10.1016/j.gltp.2021.01.006

Chen, J., Song, L., Wainwright, M. J., & Jordan, M. I. (2018). Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning (ICML)* (pp. 883-892).

Dataset (2016). *Default of Credit Card Clients*. https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients

Dighe, D., Patil, S., & Kokate, S. (2018). Detection of credit card fraud transactions using machine learning algorithms and neural networks: a comparative study. In *Proceedings of IEEE* (pp.1-6).

Gu, L., Zhou, N., & Zhao, Y. (2018). An Euclidean Distance Based on Tensor Product Graph Diffusion Related Attribute Value Embedding for Nominal Data Clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), 3101-3108. https://doi.org/10.1609/aaai.v32i1.11681

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2020). A Survey of Methods for Interpreting Machine Learning Models. *ACM Computing Surveys*, *52*(5), 1-52.

Gramegna, A., & Giudici, P. (2021). SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk. *Frontiers in Artificial Intelligence, 4,* Article 752558. https://doi.org/10.3389/frai.2021.752558

Kanmani, W. S., & Jayapradha, B. (2017). Prediction of default customer in banking sector using artificial neural network. *International Journal of Research in Information Technology and Computing*, *5*, 293-296.

Lundberg, S. M., & Lee, S. I. (2017). Explaining black box models with Shapley values. In *Advances in Neural Information Processing Systems* (pp. 4765-4774).

Molnar, C. (2020). *Interpretable Machine Learning*. Leanpub. https://christophm.github.io/interpretable-ml-book/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2017). A unified approach to interpretable machine learning. In *Proceedings of the 34th International Conference on Machine Learning* (ICML) (pp. 1165-1174).

Ribeiro, M. T., & Singh, S. (2020*).* "Why should I trust you?" Explaining the predictions of any machine learning classifier with LIME. *Data Mining and Knowledge Discovery*, *34*(3), 818-835.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). LIME: Local interpretable model-agnostic explanations. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 2145-2154).

Zhang, G. P. (2000). *Neural networks for classification: A survey*. IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, *30*(4), 451-462.

Zhang, B., Wu, B., & Xu, L. (2022). Interpretable Machine Learning for Natural Language Processing. *ACM Computing Surveys*, *55*(3), 1-40.

INTERPRETATION OF CREDIT CARD FEATURES BY LIME AND SHAP AFTER DEEP LEARNING

Nguyen Quoc Huy**, *Tu Lang Phieu

Saigon University, Vietnam

**Corresponding author: Nguyen Quoc Huy – Email: nqhuy@sgu.edu.vn*

Received: August 09, 2023; Revised: October 19, 2023; Accepted: October 20, 2023

ABSTRACT

Feature Interpretation after training is a necessary research trend in AI application and is of great interest to the AI research community. Explainable AI (XAI) has two main popular approaches, LIME and SHAP, which are very effective in explaining the models after training. There are many documents for using LIME and SHAP libraries to explain the models after training, but very few documents discuss the mathematical model inside LIME and SHAP. This makes it difficult to study. The paper describes in detail the steps of LIME and SHAP on small data and then explains its features after classifying credit card data by deep learning. The experimental results and interpretation show interesting information in interpreting the features by the LIME and SHAP methods.

Keywords: credit card clasification; LIME; machine learning; SHAP; XAI