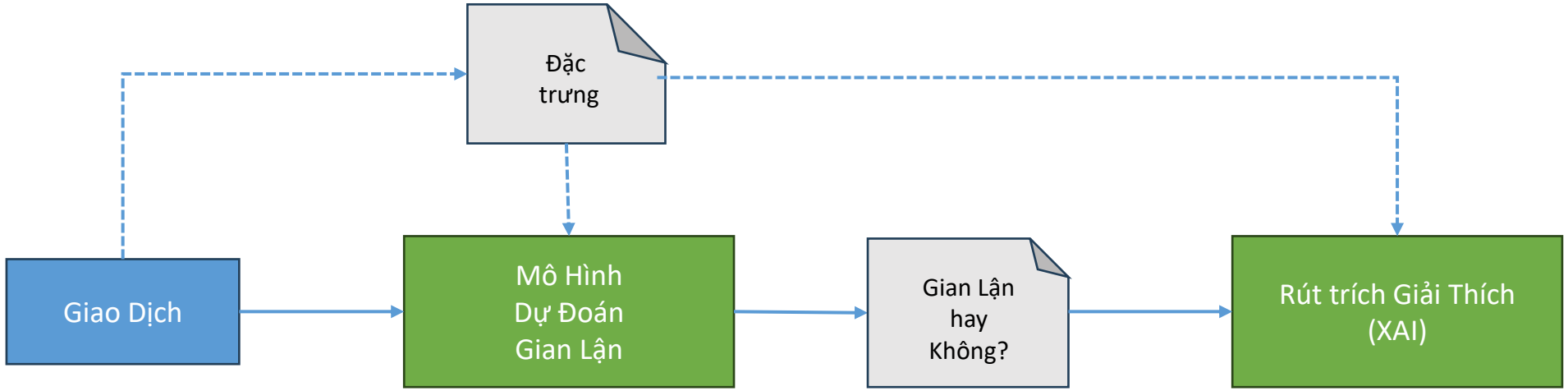
NGHIÊN CỨU XAI ĐỂ GIẢI THÍCH CÁC YẾU TỐ GIAN LẬN

6/2025

MỤC LỤC

1. TỔNG QUAN VỀ ĐỀ TÀI

- Qui trình dự đoán gian lận



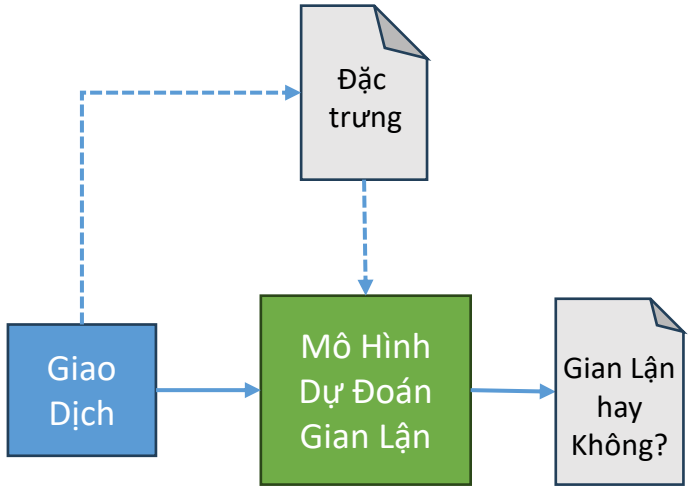
Hình 1. Qui trình dự đoán gian lận

- ❑ Hành vi gian lận trong giao dịch tài chính thường gây thiệt hại lớn
- ❑ Gian lận là hành vi tương đối hiếm → Cần có 1 hệ thống AI để lọc ra các nghi ngờ → Giảm chi phí hoạt động
- ❑ Các giao dịch được đánh dấu nghi ngờ, sẽ được các chuyên gia xem xét tính hợp pháp

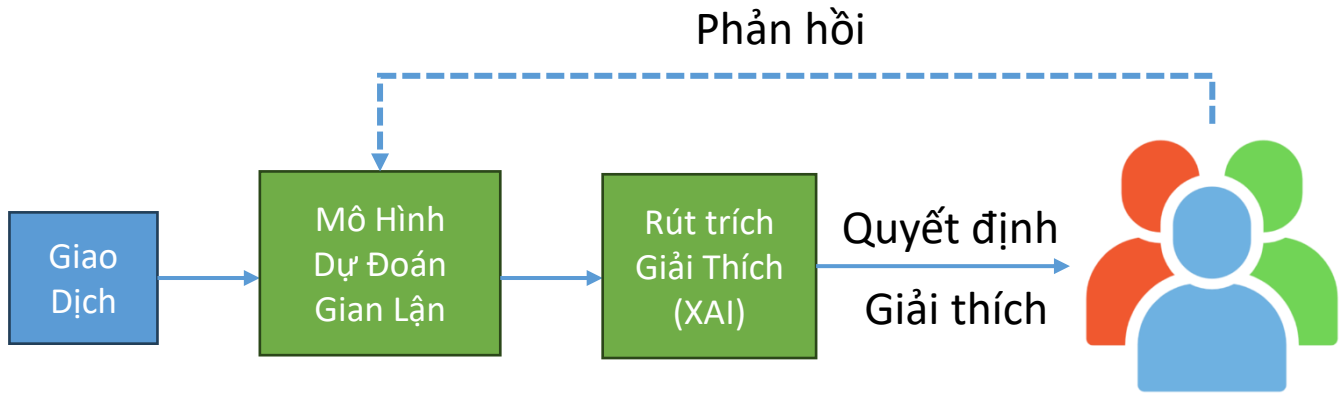
[1] I. Zitouni et al., "***Explainable AI in fraud detection***," *Milliman Insights*, vol. 12, no. 22, pp. 1-4, Dec. 2022. [Online]. Available: https://www.milliman.com/en/insight/explainable-ai-in-fraud-detection#1.

1.1. ĐỘNG LỰC NGHIÊN CỨU

- AI đã cách mạng hóa nhiều lĩnh vực nhưng gặp trở ngại trong việc chấp nhận rộng rãi do thiếu minh bạch và khả năng giải thích trong các mô hình "hộp đen".



Hình 2. Mô hình AI không có XAI



Hình 3. Mô hình AI có XAI

- Trong phát hiện gian lận, XAI được tích hợp để giải thích lý do tại sao một giao dịch không gian lận hoặc bị đánh dấu là gian lận — từ đó giúp con người hiểu, tin cậy và hành động hiệu quả hơn.

1.2. MỤC TIÊU NGHIÊN CỨU

- Các mô hình AI hiện nay thường gặp tình trạng mất cân bằng dữ liệu, dẫn đến nhiều kết quả dương tính giả (false positives), tức là giao dịch hợp lệ bị đánh dấu gian lận. ***Lưu ý***: mô hình có xu hướng đẩy về dương tính giả thay cho âm tính giả (không hợp lệ mà đánh dấu hợp lệ) để tránh rủi ro.
- XAI hỗ trợ các điều tra viên bằng cách cung cấp lý do tại sao AI đánh dấu một giao dịch là gian lận, giúp tăng hiệu quả và giảm chi phí vận hành.
- XAI giúp hỗ trợ giải thích trong các mô hình hộp đen (black-box models).
- Kết hợp các kỹ thuật như SHAP[1] (SHapley Additive exPlanations) để giải thích rõ hơn về các quyết định của mô hình.

[1] S. M. Lundberg et al., "***A unified approach to interpreting model predictions***," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

2. CÁC VẤN ĐỀ LIÊN QUAN

Bảng 1. So sánh mô hình hộp đen vs. hộp trắng

Đặc điểm	Black Box Models (Hộp Đen)	White Box Models (Hộp Trắng)
Định nghĩa	Các mô hình AI có cấu trúc và quyết định không thể giải thích rõ ràng hoặc trực quan bởi con người	Các mô hình có thể hiểu và giải thích rõ ràng cách thức hoạt động và ra quyết định
Độ phức tạp	Rất phức tạp, thường liên quan đến các thuật toán nâng cao như mạng nơ-ron sâu (DNN) hoặc cây quyết định phức hợp.	Đơn giản, sử dụng các phương pháp toán học hoặc thống kê dễ hiểu như hồi quy tuyến tính hoặc cây quyết định cơ bản.
Khả năng giải thích	Khó hoặc không thể giải thích rõ ràng cách một quyết định được đưa ra.	Dễ dàng giải thích từng bước trong quá trình xử lý và ra quyết định.
Ưu điểm	- Hiệu suất cao trong các bài toán phức tạp. - Có khả năng xử lý dữ liệu phi tuyến và khối lượng lớn.	- Minh bạch và dễ hiểu. - Thích hợp cho các lĩnh vực đòi hỏi tính trách nhiệm và sự tin cậy cao.
Nhược điểm	- Thiếu minh bạch và khó xây dựng niềm tin. - Yêu cầu công cụ như XAI (Explainable AI) để làm rõ quyết định.	- Hiệu suất thấp hơn trong các bài toán phức tạp. - Khó xử lý dữ liệu phi tuyến và không phù hợp với dữ liệu lớn.

2.1. LIME & SHAP TRONG GIẢI THÍCH MÔ HÌNH HỘP ĐEN

Bảng 2. LIME và SHAP

Yếu tố	LIME	SHAP
Nguyên lý	Mô hình giả lập cục bộ, gần đúng	Giá trị Shapley, theo lý thuyết trò chơi
Độ chính xác	Gần đúng, có thể sai lệch	Chính xác và vững chắc về lý thuyết
Hiệu suất tính toán	Nhanh, nhưng yêu cầu huấn luyện mô hình giả lập xấp xĩ	Chậm nhưng tối ưu với TreeSHAP
Ứng dụng	Giải thích cục bộ cho từng mô hình học máy	Giải thích cục bộ và toàn cục, mô hình học máy phức tạp
Phạm vi giải thích	Cục bộ	Giải thích cục bộ và toàn cục

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "'***Why should I trust you?': Explaining the predictions of any classifier***," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1135–1144, 2016.
[2] S. M. Lundberg et al., "***A unified approach to interpreting model predictions***," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

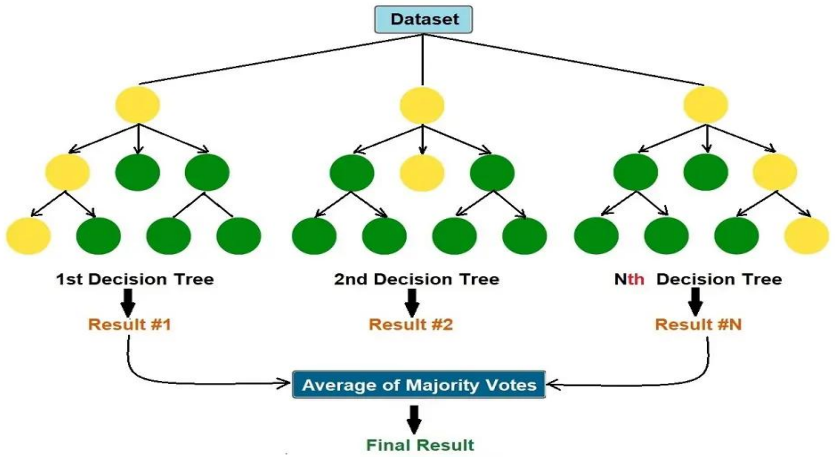
2.2. CÁC MÔ HÌNH MÁY HỌC DỰ ĐOÁN GIAN LẬN

Random Forest

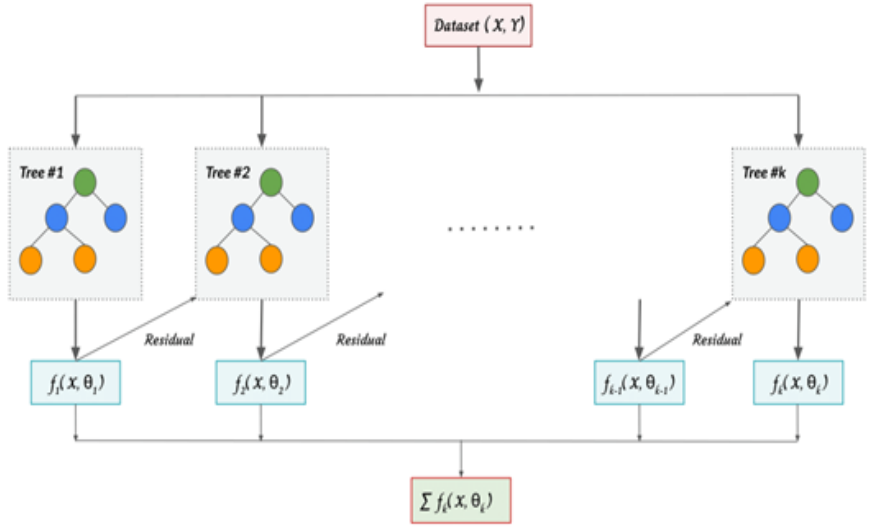
- Dùng nhiều cây quyết định độc lập để dự đoán.
- Giúp giảm sai sót do cây đơn lẻ gây ra, tăng độ ổn định và chính xác.
- Thích hợp với dữ liệu có nhiều biến và phức tạp.

XGBoost

- Tạo các cây liên tiếp nhau, mỗi cây sửa lỗi của cây trước.
- Giúp mô hình học nhanh và chính xác hơn, đặc biệt với dữ liệu mất cân bằng hoặc nhiễu.
- Thường cho hiệu suất cao trong các bài toán phức tạp.



Hình 5. Minh họa mô hình Random Forest



Hình 6. Minh họa mô hình XGBoost

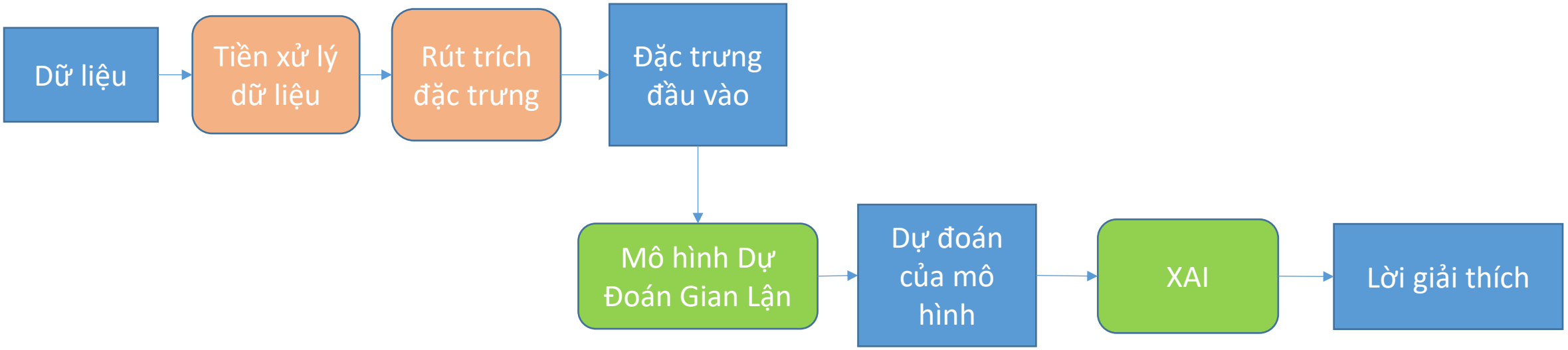
[1] C. Liu, Y. Chan, S. H. A. Kazmi, and H. Fu, "***Financial fraud detection model: Based on Random Forest***," *Int. J. Econ. Finance*, vol. 7, no. 7, pp. 178–188, 2015.

[2] T. Chen and C. Guestrin, "***XGBoost: A scalable tree boosting system***," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, San Francisco, CA, USA, 2016, pp. 785–794.

3. PHƯƠNG PHÁP ĐỀ XUẤT

3.1. Quy trình tổng thể

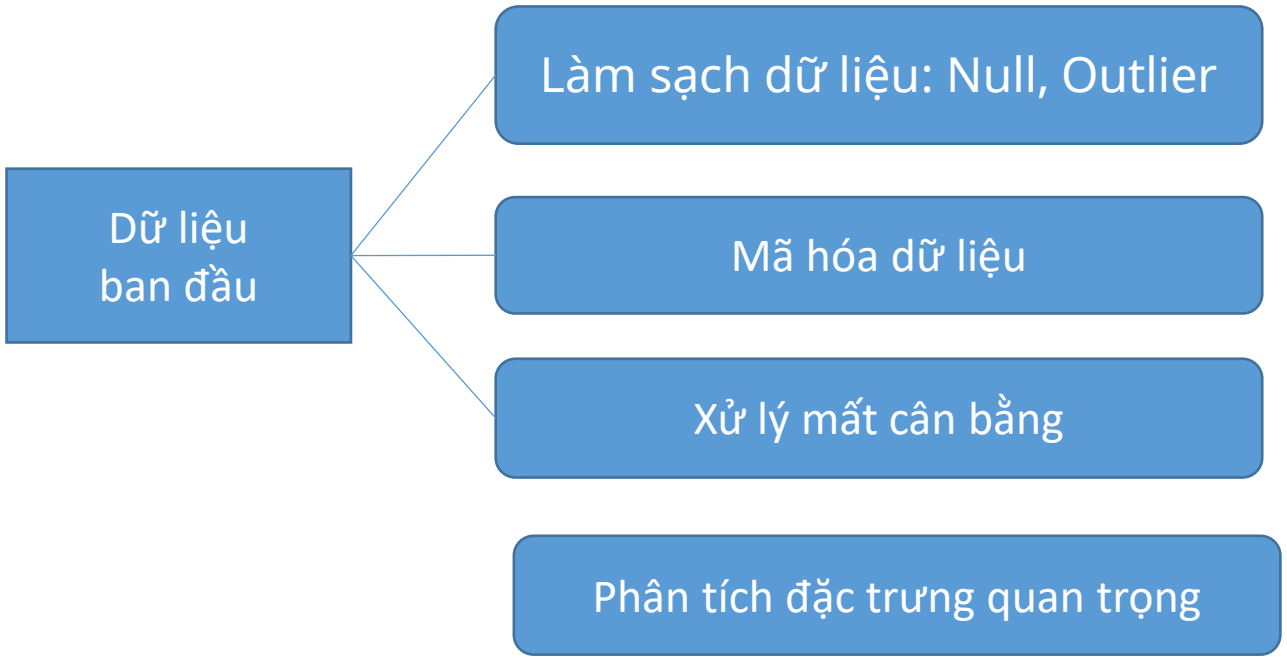
- Tiến hành huấn luyện mô hình hộp đen trên dữ liệu bảng.
- Áp dụng phương pháp XAI để tạo ra lời giải thích



Hình 7. Tổng quan phương pháp đề xuất

3.2. TIỀN XỬ LÝ DỮ LIỆU VÀ RÚT TRÍCH ĐẶC TRƯNG

- Mã hóa dữ liệu sử dụng Label Encoder
- Xử lý mất cân bằng với SMOTE
- Đánh giá tầm quan trọng đặc trưng dựa trên mô hình

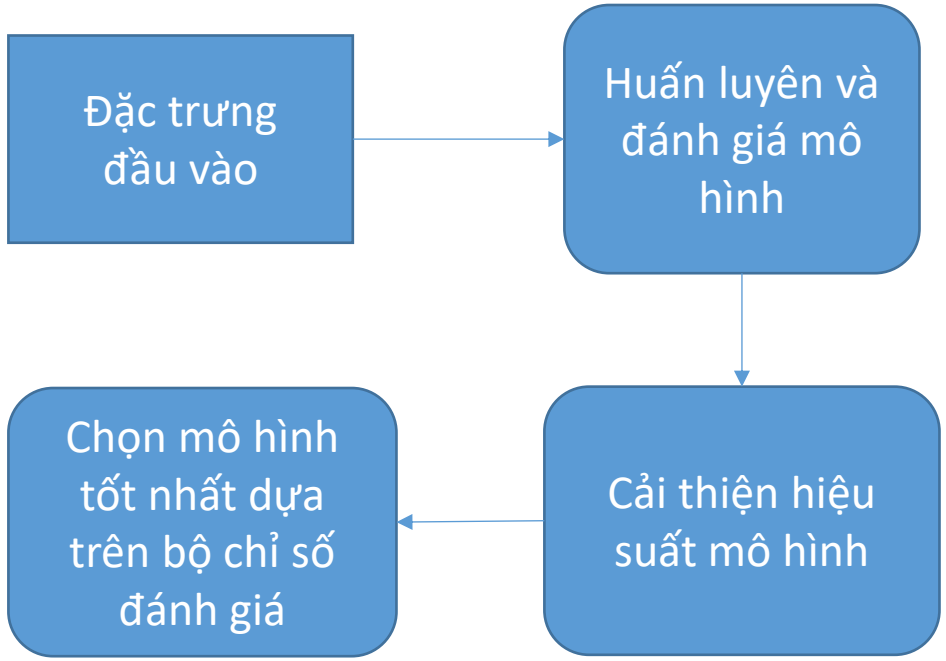


Hình 8. Tiền xử lý dữ liệu

- Với **Random Forest**, chỉ số được sử dụng phổ biến là **Mean Decrease in Impurity (MDI)** – đo lường mức độ giảm độ hỗn tạp (impurity) khi chia dữ liệu tại một đặc trưng nào đó.
- Với **XGBoost**, chỉ số **Gain** được sử dụng để thể hiện mức độ cải thiện hàm mục tiêu khi một đặc trưng được chọn để tách nhánh.
- Mục tiêu: xây dựng một **tập đặc trưng tinh gọn, hiệu quả và dễ giải thích.**

3.3. HUẤN LUYỆN VÀ ĐÁNH GIÁ MÔ HÌNH

- Cải thiện hiệu suất mô hình bằng tối ưu hóa tham số với GridSeacrhCV
- Chọn mô hình tốt nhất để áp dụng XAI

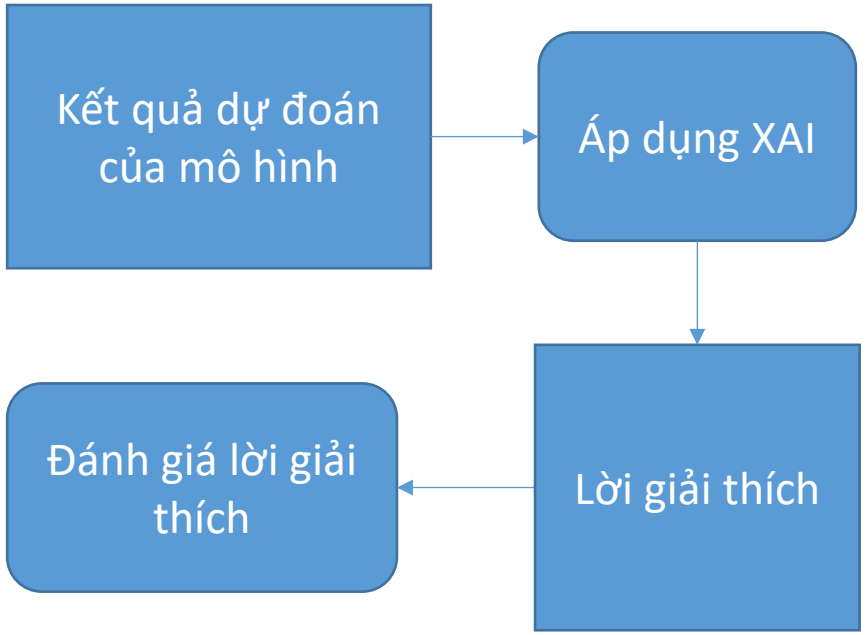


Hình 9. Huấn luyện và đánh giá mô hình

- **Xác định không gian tham số**: Bao gồm các siêu tham số đặc trưng như số lượng cây (n_estimators), độ sâu tối đa (max_depth), hệ số điều chỉnh (C, gamma với SVM), hoặc số lượng lân cận (n_neighbors trong KNN).
- **Đánh giá mô hình với cross-validation**: Mỗi cấu hình siêu tham số được đánh giá qua nhiều lần chia ngẫu nhiên để đảm bảo kết quả không phụ thuộc vào ngẫu nhiên.
- **Huấn luyện mô hình cuối cùng**: Mô hình với tập siêu tham số tối ưu sẽ được huấn luyện lại trên toàn bộ tập huấn luyện để xây dựng mô hình cuối cùng.

3.4. ÁP DỤNG XAI TẠO LỜI GIẢI THÍCH VÀ ĐÁNH GIÁ

- XAI được áp dụng để giải thích các dự đoán của mô hình học máy.
- Lời giải thích giúp đánh giá độ tin cậy và hiểu rõ cơ chế ra quyết định.



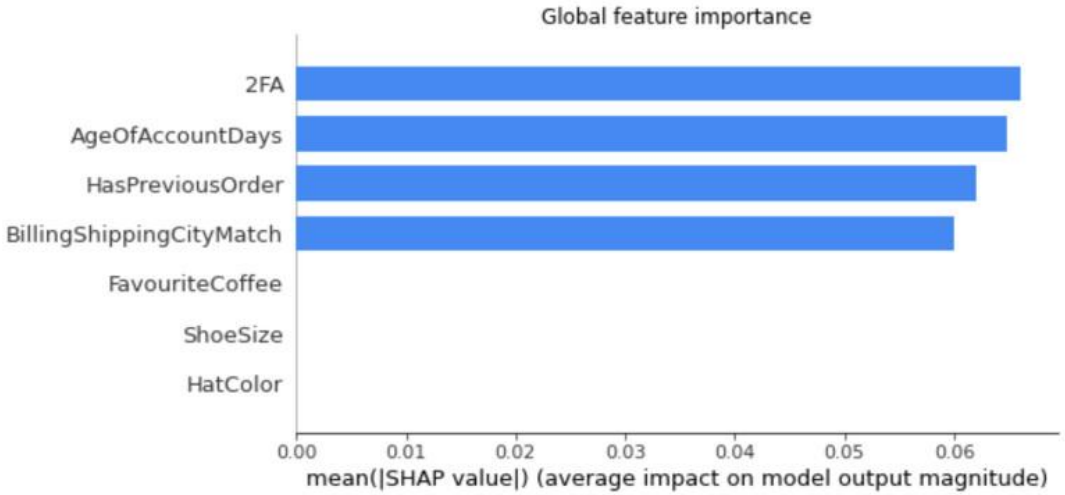
Hình 10. XAI để giải thích mô hình

- Phương pháp **SHAP** được sử dụng nhằm cung cấp giải thích **ở cả cấp độ toàn cục (global) và cục bộ (local)**. SHAP dựa trên lý thuyết trò chơi và giá trị Shapley để đo lường mức độ đóng góp của từng đặc trưng vào đầu ra của mô hình.

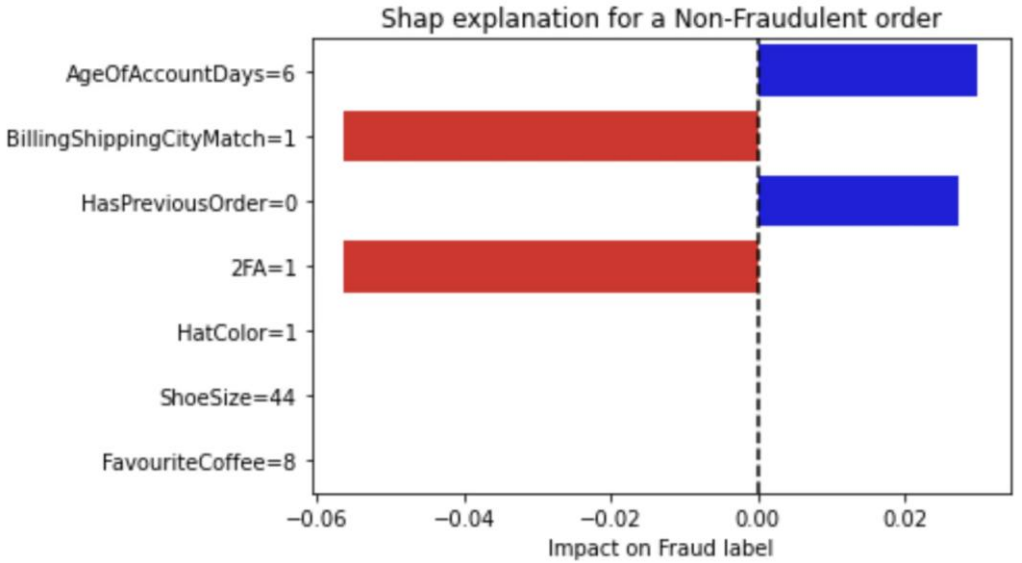
- **LIME** tập trung vào giải thích **cục bộ** – tức là giải thích cụ thể một mẫu dữ liệu tại một thời điểm. Phương pháp này phù hợp để kiểm tra **các trường hợp cá biệt** hoặc "truy vết" nguyên nhân vì sao một giao dịch cụ thể bị gắn nhãn là gian lận.

3.4. ÁP DỤNG XAI TẠO LỜI GIẢI THÍCH VÀ ĐÁNH GIÁ

- Giải thích dùng SHAP



Hình 11. Đặc trưng quan trọng toàn cục



Hình 12. Giải thích cục bộ cho giao dịch không gian lận

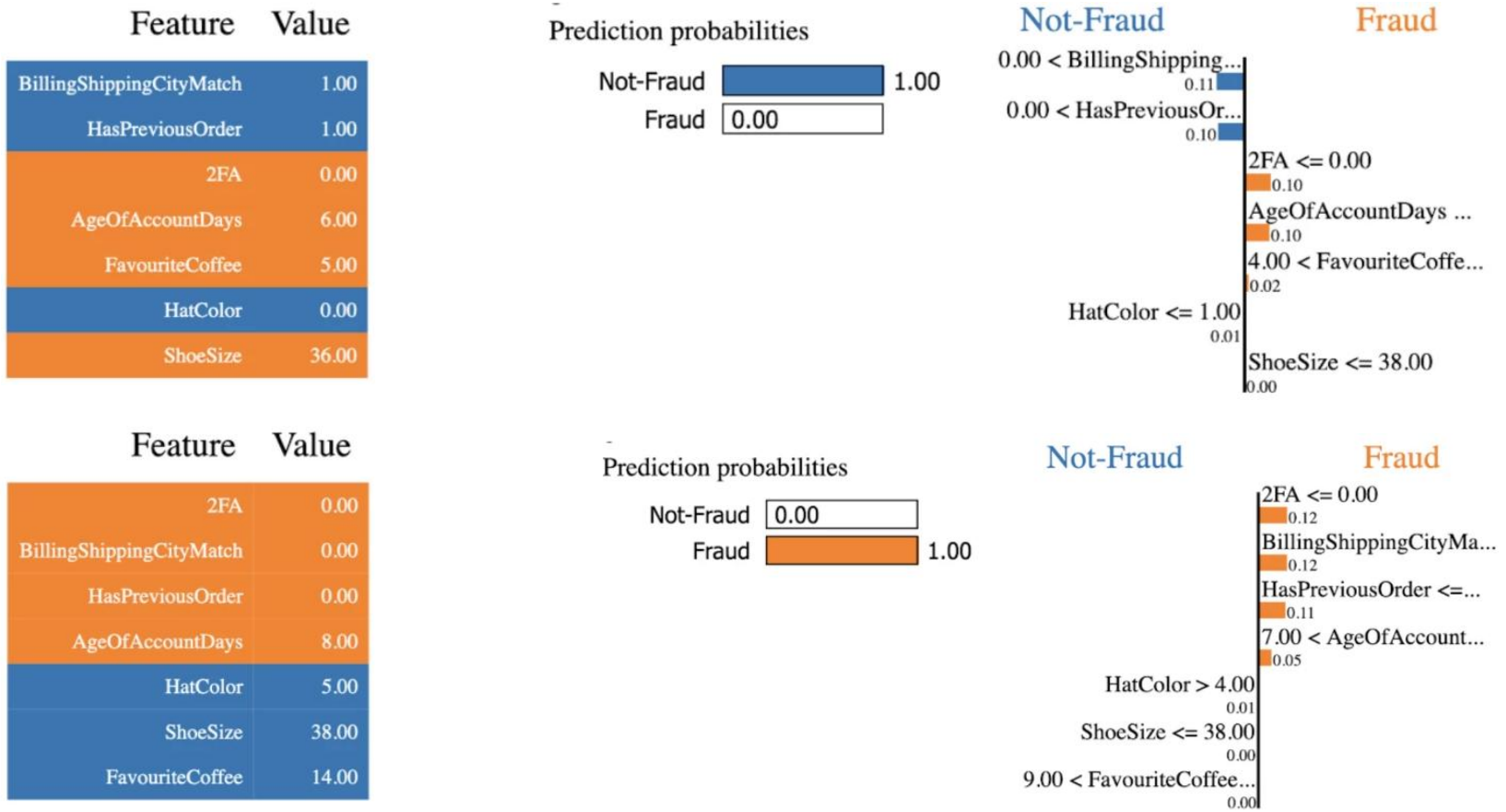
+ **2FA** (Xác thực hai yếu tố): Các tài khoản có 2FA thường ít bị gắn cờ gian lận hơn, vì 2FA tăng cường bảo mật.
+ **AgeOfAccountDays** (Số ngày tồn tại của tài khoản): Tài khoản lâu năm thường được đánh giá là đáng tin cậy hơn.
+ **HasPreviousOrder** (Lịch sử đơn hàng trước đó): Tài khoản có lịch sử giao dịch thường giảm nguy cơ bị gắn cờ gian lận.
+ **BillingShippingCityMatch** (Địa chỉ thanh toán và giao hàng khớp nhau): Sự không khớp giữa địa chỉ thanh toán và giao hàng là dấu hiệu phổ biến của gian lận.

+ Giá trị dương (màu xanh): Đặc trưng làm tăng khả năng giao dịch **không gian lận**.
+ Giá trị âm (màu đỏ): Đặc trưng làm tăng khả năng giao dịch **bị nghi ngờ gian lận**.

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "'***Why should I trust you?': Explaining the predictions of any classifier***," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1135–1144, 2016.

3.4. ÁP DỤNG XAI TẠO LỜI GIẢI THÍCH VÀ ĐÁNH GIÁ

- Giải thích dùng LIME



+ Các đặc trưng **tích cực** (màu xanh lá) đóng góp vào việc phân loại giao dịch là **an toàn**, Ví dụ: địa chỉ thanh toán và giao hàng khớp nhau.
+ Tài khoản có lịch sử đặt hàng trước đó.
+ Có kích hoạt xác thực hai yếu tố (2FA).

+ Địa chỉ thanh toán và giao hàng không khớp.
+ Tài khoản mới hoặc không có lịch sử giao dịch.
+ Không kích hoạt 2FA.
+ Các đặc điểm bất thường khác, như kích thước giao dịch lớn bất thường hoặc vị trí giao dịch không hợp lệ.

Hình 13. Giải thích LIME cho một đơn hàng an toàn (trên) và một đơn hàng gian lận (dưới)

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "'***Why should I trust you?': Explaining the predictions of any classifier***," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1135–1144, 2016.

4. THỰC NGHIỆM VÀ KẾT QUẢ

4.1. Dữ liệu thực nghiệm

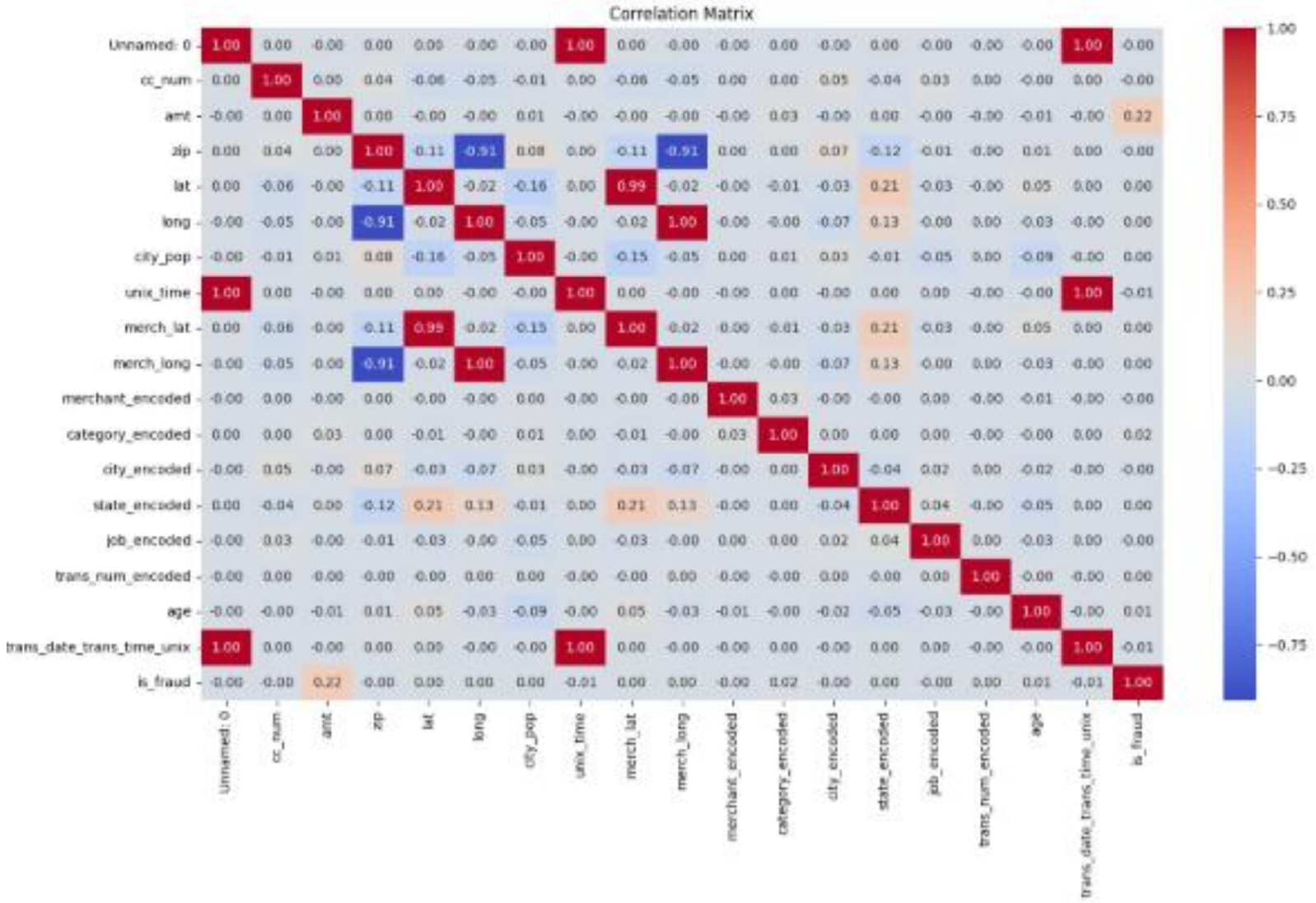
- Bộ dữ liệu phát hiện gian lận thẻ tín dụng (Kaggle)
- Tập dữ liệu mô phỏng các giao dịch thẻ tín dụng diễn ra trong giai đoạn từ ngày 01/01/2019 đến 31/12/2020.
- Tổng cộng, bộ dữ liệu bao gồm **1.296.675 giao dịch**, trong đó có **7.506 giao dịch gian lận**, chiếm khoảng **0.58%**, thể hiện bài toán mất cân bằng dữ liệu đặc trưng trong lĩnh vực tài chính.
- Mỗi bản ghi trong tập dữ liệu gồm **23 trường thông tin**, bao phủ cả thông tin nhân khẩu học và hành vi giao dịch

[1] **Fraud detection** [Data set]. (n.d.). *Kaggle*. https://www.kaggle.com/datasets/kartik2112/fraud-detection
[2] A. Correa Bahnsen, A. Stojanovic, D. Aouada, and B. Ottersten, "Feature engineering strategies for credit card fraud detection," *Expert Systems with Applications*, vol. 51, pp. 134–142, 2016.

4.1. DỮ LIỆU THỰC NGHIỆM

- **Unnamed: 0 (*int64*):** Chỉ số dòng (index tự động).
- **trans_date_trans_time (*object*):** Thời gian giao dịch ở định dạng chuỗi (ngày-giờ).
- **cc_num (*int64*):** Mã số thẻ tín dụng (đã ẩn danh).
- **merchant (*object*):** Tên/ID của cửa hàng.
- **category (*object*):** Danh mục giao dịch (ví dụ: shopping, travel...).
- **amt (*float64*):** Số tiền giao dịch.
- **first, last (*object*):** Họ và tên người dùng (đã ẩn danh).
- **gender (*object*):** Giới tính người dùng.
- **street, city, state, zip (*object*/ *int64*):** Thông tin địa chỉ người dùng.
- **lat, long (*float64*):** Tọa độ địa lý (vĩ độ và kinh độ) của người dùng.
- **city_pop (*int64*):** Dân số thành phố nơi người dùng cư trú.
- **job (*object*):** Nghề nghiệp của người dùng.
- **dob (*object*):** Ngày sinh của người dùng.
- **trans_num (*object*):** Mã giao dịch.
- **unix_time (*int64*):** Thời gian giao dịch ở dạng Unix timestamp.
- **merch_lat, merch_long (*float64*):** Tọa độ địa lý của cửa hàng.
- **is_fraud (*int64*):** Nhãn mục tiêu — 1 nếu là giao dịch gian lận, 0 nếu bình thường.

PHÂN TÍCH DỮ LIỆU HUẤN LUYỆN



Hình 14. Phân tích tương quan

- **Tỉ lệ giữa các lớp:** số lượng giao dịch gian lận chiếm 0.58%, lớp giao dịch hợp lệ chiếm 99.42%.
- **Phân tích tương quan với nhãn mục tiêu is_fraud**
- Một tập đặc trưng được lựa chọn dựa trên phân tích để **giảm nhiễu, hạn chế đa cộng tuyến, và tăng tốc quá trình huấn luyện**

PHÂN CHIA DỮ LIỆU THỰC NGHIỆM

amt
category
age
city
state
lat
long
merch_lat
merch_long
trans_date_trans_time

Hình 15. Các đặc trưng chọn lựa

- **Chọn 10 đặc trưng quan trọng:**
  - **amt**: Gian lận thường đi kèm với các giao dịch giá trị cao bất thường.
  - **category**: Một số loại hình giao dịch dễ bị gian lận hơn (ví dụ: online shopping, entertainment).
  - **age**: Nhóm tuổi có thể liên quan đến hành vi tiêu dùng và mức độ bị nhắm tới.
  - **city, state**: Vị trí địa lý có thể ảnh hưởng đến tần suất và loại gian lận.
  - **lat, long, merch_lat, merch_long**: So sánh vị trí người dùng và cửa hàng giúp phát hiện những giao dịch bất thường về mặt địa lý. (ví dụ người ở New York nhưng giao dịch tại California)
  - **trans_date_trans_time_unix**: Gian lận có thể xảy ra vào những khung giờ bất thường (ví dụ: nửa đêm).

- **Toàn bộ 7.506 mẫu giao dịch gian lận** được giữ nguyên, sau đó chia đều vào hai tập huấn luyện và kiểm tra (mỗi tập 3.753 mẫu).
- Với mỗi tập, bổ sung thêm các mẫu không gian lận được chọn ngẫu nhiên sao cho tỷ lệ gian lận đạt **khoảng 3%** trong mỗi tập. Tỷ lệ này cao hơn thực tế, nhưng **giúp cân bằng hiệu quả huấn luyện và tăng khả năng nhận diện lớp thiểu số**.

4.2. ĐỘ ĐO ĐÁNH GIÁ HIỆU SUẤT

Các chỉ số đánh giá mô hình:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Accuracy là một chỉ số đo lường mô hình dự đoán đúng bao nhiêu phần trăm trên toàn bộ tập dữ liệu.

Precision (Độ chính xác dương) đo lường tỷ lệ các giao dịch mà mô hình dự đoán là gian lận thực sự là gian lận

Recall (Độ nhạy) thể hiện tỷ lệ các giao dịch gian lận thực sự được mô hình phát hiện.

F1-Score là chỉ số tổng hợp giữa Precision và Recall, hữu ích khi cần cân bằng giữa khả năng phát hiện gian lận và hạn chế báo động sai

ĐÁNH GIÁ LỜI GIẢI THÍCH CỦA XAI

Các chỉ số đánh giá lời giải thích:

$$Fidelity = \frac{1}{|\aleph(x)|} \sum_{x' \epsilon \aleph(x)} f(x') - g(x')$$

- f: mô hình gốc
- g: mô hình giải thích (ví dụ mô hình hồi quy tuyến tính cục bộ
- N(x): Các điểm lân cận của điểm cần giải thích x

Fildelity[1]l à độ trung thực của lời giải thích
đo lường mức độ mà lời giải thích phản ánh đúng hành vi của mô hình gốc.
Fidelity càng gần 1 thì càng tốt (lời giải thích càng gần với hành vi mô hình gốc trong vùng lân cận).

$$Stability(x_1, x_2) = \frac{||E(x_1) - E(x_2)}{||x_1 - x_2||}$$

- E(x): lời giải thích cho điểm x(thường là vector độ quan trọng của các đặc trưng)
- x1,x2: hai điểm đầu vào gần nhau

Stability là độ ổn định của lời giải thích
Chỉ số stability càng nhỏ → lời giải thích càng ổn định.

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "'***Why should I trust you?': Explaining the predictions of any classifier***," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

4.3. THỰC NGHIỆM 1: HUẤN LUYỆN MÔ HÌNH CƠ SỞ

Bảng 2. So sánh hiệu suất mô hình học máy trong phát hiện gian lận

Mô hình	Accuracy	Precision	Recall	F1-score	AUC-ROC
Random Forest	0.9900	0.9259	0.7253	0.8131	0.9879
XGBoost	0.9897	0.7853	0.9065	0.8414	0.9956

- **Random Forest** đạt Precision cao hơn (0.9259), cho thấy khả năng hạn chế dự đoán sai dương tính (false positives) tốt hơn. Điều này phù hợp trong các tình huống cần giảm thiểu cảnh báo sai.
- **XGBoost** lại vượt trội về Recall (0.9065), F1-score (0.8414) và AUC-ROC (0.9956), cho thấy mô hình có khả năng phát hiện đầy đủ hơn các trường hợp gian lận, đồng thời duy trì sự cân bằng giữa độ chính xác và độ bao phủ.

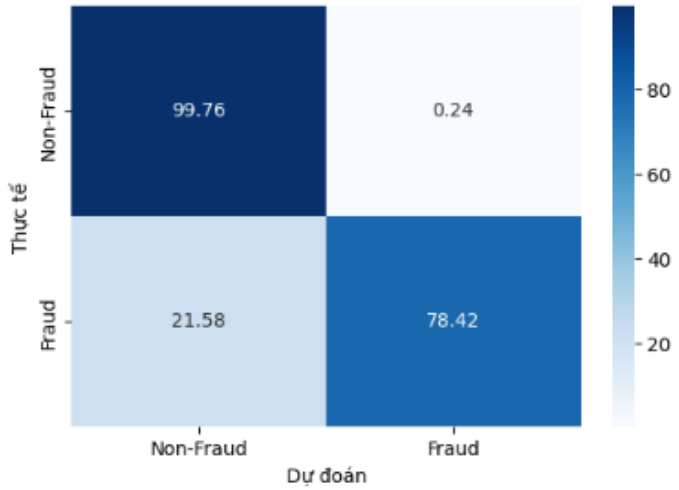
4.4. THỰC NGHIỆM 2: TỐI ƯU HÓA MÔ HÌNH

Kết quả đánh giá mô hình XGBoost trên tập kiểm tra

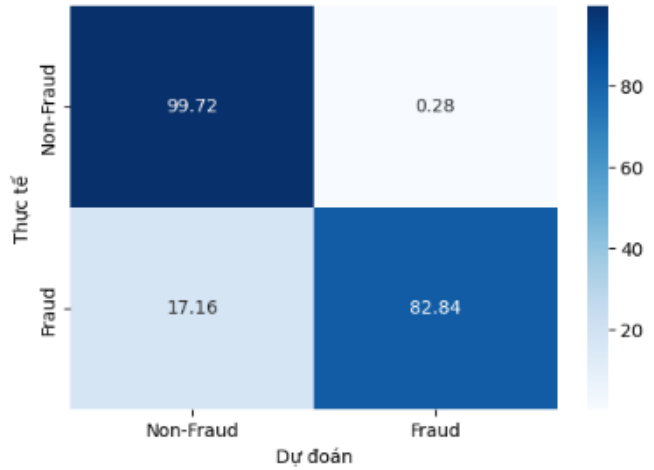
Mô hình	Accuracy	Precision	Recall	F1-score	AUC-ROC
Mô hình cơ sở	0.9887	0.7558	0.9227	0.8310	0.9958
Mô hình tối ưu	0.9922	0.9030	0.8284	0.8641	0.9955

Kết quả đánh giá mô hình Random Forest trên tập kiểm tra

Mô hình	Accuracy	Precision	Recall	F1-score	AUC-ROC
Mô hình cơ sở	0.9908	0.9384	0.7423	0.8289	0.9910
Mô hình tối ưu	0.9913	0.9128	0.7842	0.8436	0.9939



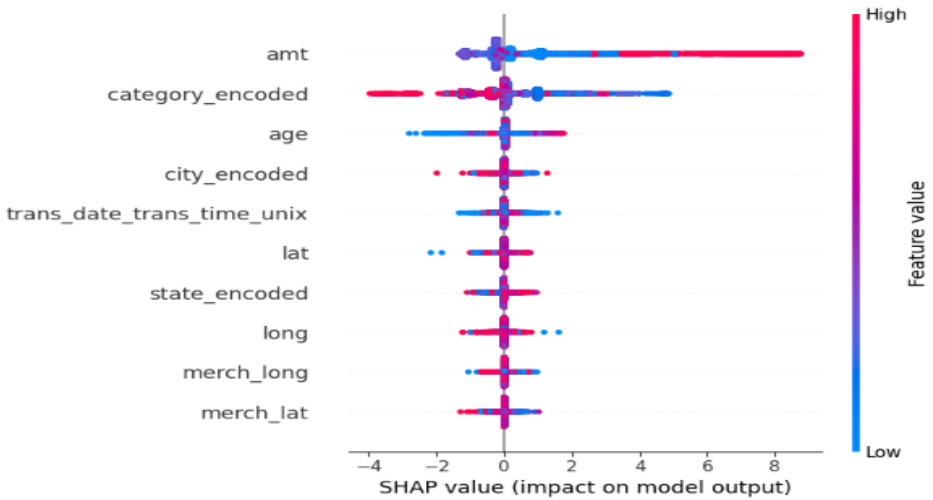
Ma trận nhầm lẫn của mô hình Random Forest (sau tối ưu)



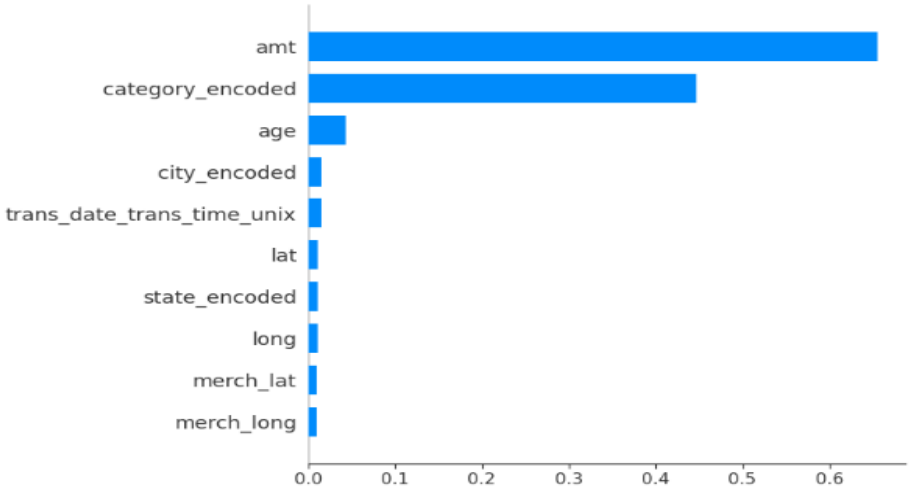
Ma trận nhầm lẫn của XGBoost sau tối ưu

4.4. THỰC NGHIỆM 3: ỨNG DỤNG XAI GIẢI THÍCH MÔ HÌNH

Giải thích dùng SHAP



Mức độ ảnh hưởng của từng đặc trưng (Beeswarm Plot)



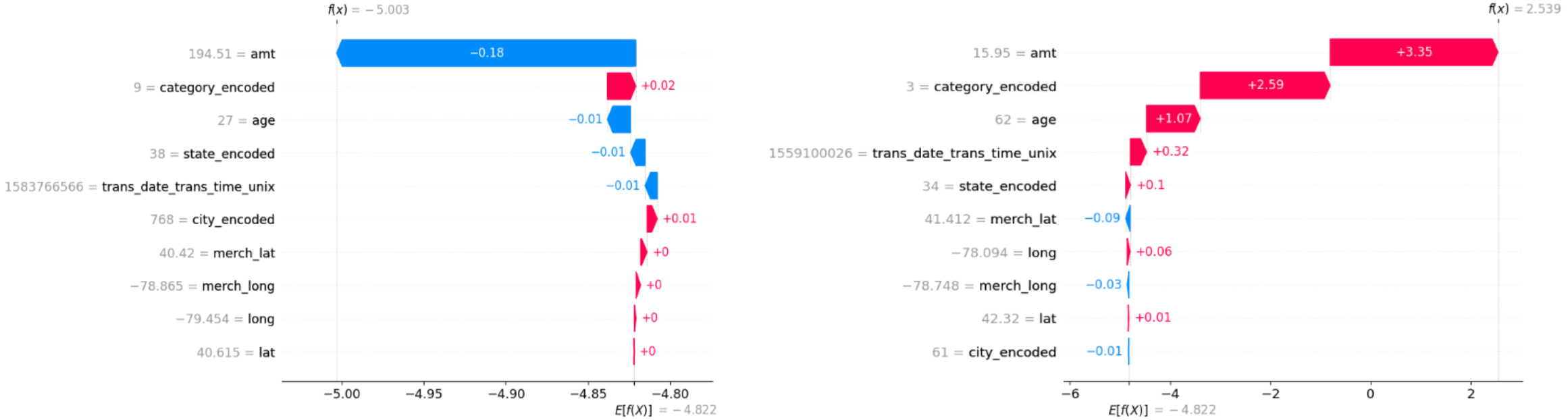
Mức độ ảnh hưởng của từng đặc trưng (Bar Plot)

+ Trục Y: Các đặc trưng được liệt kê từ trên xuống theo **mức độ ảnh hưởng trung bình** đến mô hình.
+ Trục X: Giá trị SHAP thể hiện mức độ mỗi đặc trưng **đóng góp vào việc tăng hay giảm xác suất dự đoán**
  - Giá trị dương: Đặc trưng làm tăng khả năng giao dịch **không gian lận**.
  - Giá trị âm : Đặc trưng làm tăng khả năng giao dịch **bị nghi ngờ gian lận**.
+ Giá trị đặc trưng **cao** (màu đỏ), giá trị đặc trưng **thấp** (màu xanh)

Trục Y: đại diện cho **mức độ ảnh hưởng** của từng biến đến dự đoán (dù là Fraud hay Non-Fraud).
+Trục X: độ quan trọng trung bình của mỗi đặc trưng (feature) đối với mô hình — được tính bằng giá trị tuyệt đối trung bình của SHAP trên toàn bộ tập dữ liệu.
+Không nói đến chiều ảnh hưởng (âm/dương), chỉ nói đến **tầm quan trọng**.

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "'***Why should I trust you?': Explaining the predictions of any classifier***," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1135–1144, 2016.

4.4. THỰC NGHIỆM 3: ỨNG DỤNG XAI GIẢI THÍCH MÔ HÌNH



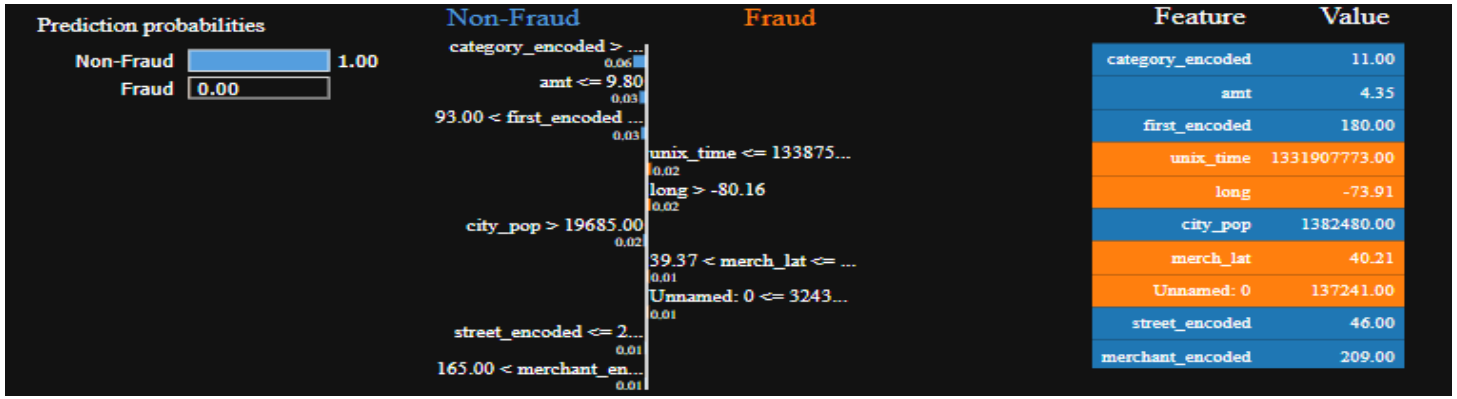
Hình 10. Giải thích cục bộ của SHAP cho giao dịch không gian lận (trái) và gian lận (phải)

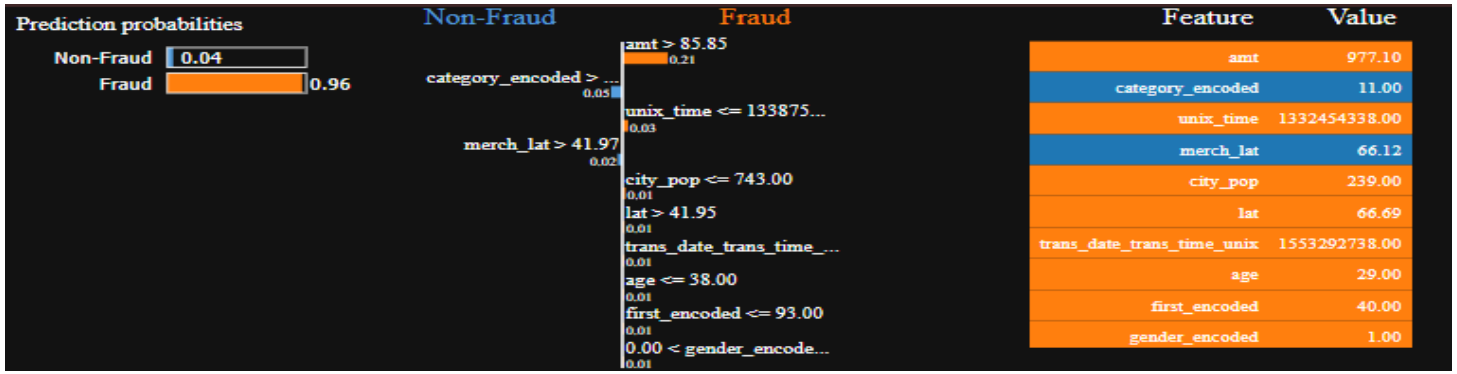
f(x) = -5.003 → mô hình tin rằng giao dịch này không gian lận
Đặc trưng ủng hộ không gian lận (giá trị âm – màu xanh):
+ amt = 194.51 *(tác động mạnh nhất)* → Giá trị giao dịch ở mức trung bình, không bất thường.
+ age = 27 → Độ tuổi trẻ, phù hợp với hành vi giao dịch bình thường.
+ state = 38 → Bang này không có dấu hiệu bất thường trong lịch sử dữ liệu.
+ time_unix & city → Thời gian & mã thành phố không nằm trong nhóm nghi ngờ.
Đặc trưng ủng hộ gian lận (giá trị dương – màu đỏ)
+category = 9 → Danh mục có một chút liên quan đến giao dịch gian lận trong dữ liệu.

f(x) = 2.539 → mô hình tin rằng giao dịch này gian lận
Nhiều đặc trưng ủng hộ gian lận
+ category_encoded = 3
→ Danh mục này có xu hướng xuất hiện trong các giao dịch gian lận trước đó.
+ age = 62
→ Người lớn tuổi, ít khả năng thực hiện giao dịch kiểu này.
+ trans_date_trans_time_unix
→ Thời điểm không phổ biến với các giao dịch bình thường.

4.4. THỰC NGHIỆM 3: ỨNG DỤNG XAI GIẢI THÍCH MÔ HÌNH

Giải thích dùng LIME





Hình 11. Giải thích LIME cho một giao dịch an toàn (trên) và một đơn hàng gian lận (dưới)

+ Non-Fraud: 1.00 → Mô hình rất **chắc chắn** rằng đây là giao dịch hợp lệ (không gian lận).

Đặc trưng ủng hộ không gian lận:

+ amt ≤ 9.80

→ Giao dịch giá trị thấp → thường không phải mục tiêu của gian lận.

+ category

→ Danh mục giao dịch phổ biến, ít liên quan đến gian lận.

Đặc trưng ủng hộ gian lận:

+unix_time ≤ 133875...

→ Giao dịch xảy ra vào thời điểm bất thường (có thể ngoài giờ hành chính).

+ merch_lat ≤ 39.37

→ Vị trí cửa hàng cũng có thể thuộc vùng đáng ngờ.

+ amt > 85.85

→ Giao dịch có giá trị cao → dễ trở thành mục tiêu để gian lận.

+ unix_time ≤ 133875...

→ Thời điểm giao dịch nằm trong khoảng giờ/đêm nghi ngờ.

+ age = 29.0

→ Có thể nằm trong nhóm tuổi bị nghi ngờ nhiều.

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "***'Why should I trust you?': Explaining the predictions of any classifier***," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1135–1144, 2016.

5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Kết luận

- Các mô hình học máy tích hợp XAI cho thấy tiềm năng rõ rệt trong việc nâng cao hiệu quả và minh bạch của hệ thống phát hiện gian lận:
- Kết hợp giữa Random Forest, XGBoost với SHAP & LIME cho kết quả ổn định và chính xác.
- Các kỹ thuật XAI như SHAP, LIME giúp mô hình trở nên dễ hiểu và dễ giải thích hơn ở cả cấp độ tổng thể lẫn cá nhân.
- Kết quả chứng minh tính khả thi của việc áp dụng XAI vào bài toán phát hiện gian lận tài chính.

Hướng phát triển

- Sử dụng thêm bộ dữ liệu đa dạng hơn để tăng độ khách quan của mô hình.
- Thử nghiệm các mô hình mạnh hơn như LightGBM, DNN cùng các kỹ thuật XAI tiên tiến (Counterfactuals, Anchors...).
- Đánh giá định lượng lời giải thích bằng các chỉ số như fidelity, stability để đảm bảo độ tin cậy.

CÁM ƠN QUÝ THẦY CÔ ĐÃ LẮNG NGHE!