

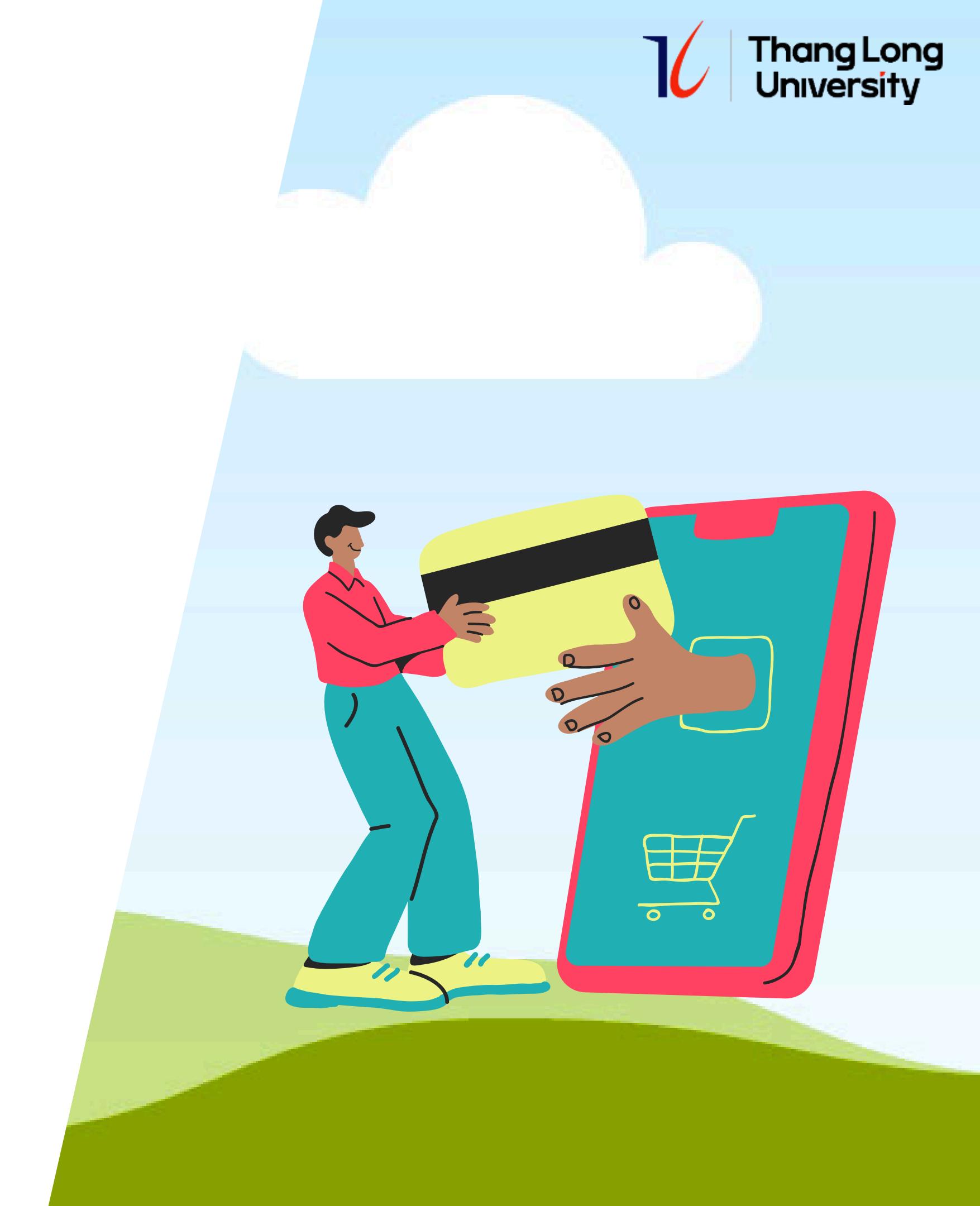
# CREDIT CARD FRAUD DATA

SỐNG NHƯ NHỮNG ĐÓA HOA

A44948  
Phan Thu Hiền

A46277  
Trần Thị Ngọc Huế

A44639  
Phạm Huyền Trang



# TỔNG QUAN



I. Business Case

II. Phương pháp sử dụng

III. Kết quả mô hình

IV. Kết luận

# I. BUSSINESS CASE

## Bussiness problem

Gian lận thẻ tín dụng gây tổn hại tài chính và làm giảm uy tín của tổ chức tài chính.

## Scope

Phát hiện gian lận thẻ tín dụng dựa trên mô hình AutoEncoder.  
Áp dụng kỹ thuật SMOTE để cân bằng dữ liệu và Logistic Regression để phân loại gian lận.

## Goals

Phát triển một hệ thống gian lận thẻ tín dụng có thể nhận diện các giao dịch gian lận.  
Tăng cường khả năng phát hiện các giao dịch gian lận với tỷ lệ Recall cao nhằm giảm thiểu các tổn thất tài chính.

## MVP

Webside phát hiện gian lận tự động sử dụng Autoencoder kết hợp Logistic Regression.

## Stakeholders

Ngân hàng và tổ chức tài chính  
khách hàng,...

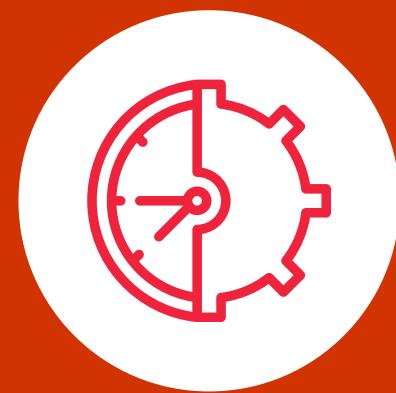


# GIỚI THIỆU VỀ DATASET



## Nguồn gốc

Tập dữ liệu được cung cấp bởi nghiên cứu liên quan đến các giao dịch thẻ tín dụng của khách hàng Châu Âu trong tháng 9 năm 2013.



## Quy mô

Tập dữ liệu bao gồm 284,807 giao dịch.  
Trong đó có 492 giao dịch là gian lận, chiếm khoảng 0,172% tổng số giao dịch.



## Các biến số

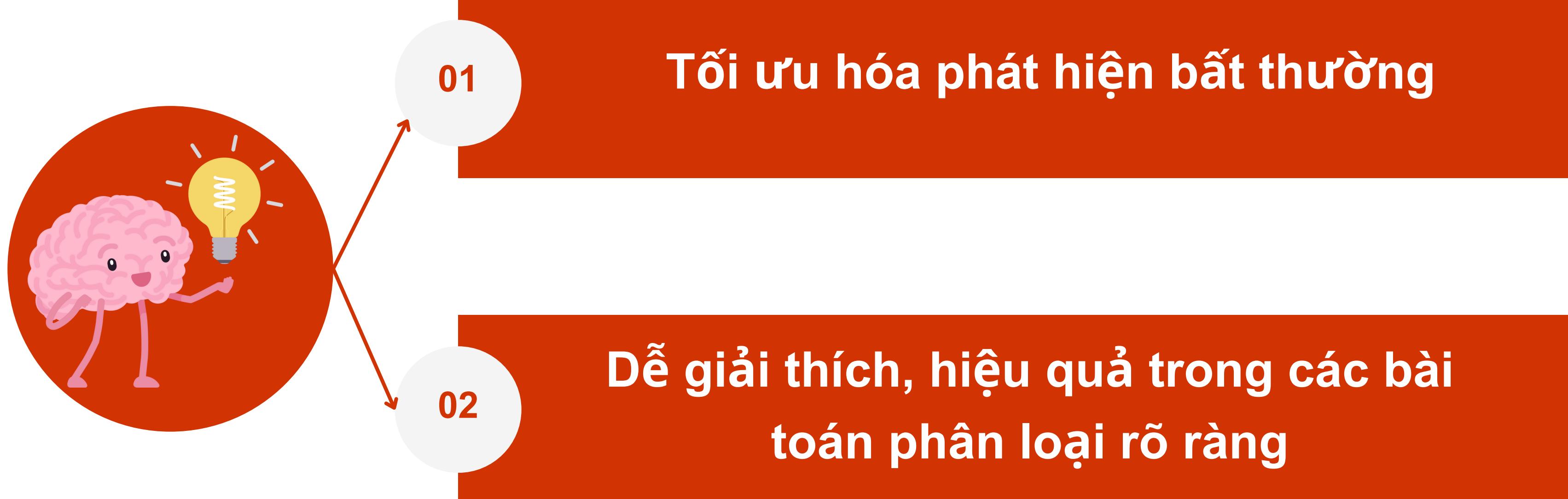
Tập dữ liệu gồm 30 biến, với các đặc trưng V1 đến V28 đã được mã hóa ẩn danh bằng PCA để bảo mật.  
Amount: Số tiền giao dịch.  
Class: Nhãn của giao dịch.  
0 - giao dịch bình thường.  
1 - giao dịch gian lận.



## Thách thức chính

Dữ liệu không cân bằng: Tỷ lệ giao dịch bình thường và gian lận rất chênh lệch.  
Mã hóa dữ liệu: Các biến số đã được mã hóa nên không thể dựa vào thông tin thực tế của giao dịch.

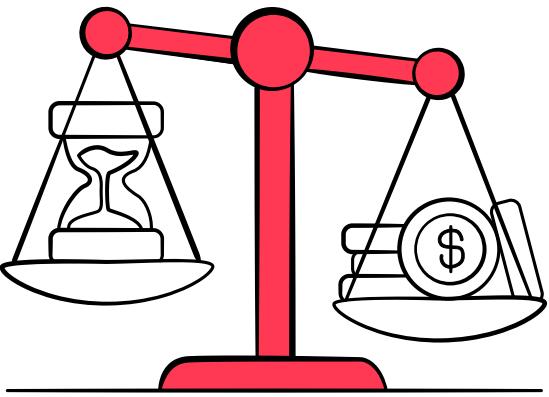
# Lý do lựa chọn kết hợp hai mô hình //



## II. PHƯƠNG PHÁP SỬ DỤNG



# TIỀN XỬ LÝ DỮ LIỆU



## LOẠI BỎ CỘT DỮ LIỆU TIME

20%

Không có giá trị trong việc phát hiện bài toán.

## TÁCH DỮ LIỆU

80%

Chia tệp dữ liệu theo tỷ lệ 80-20 giúp đánh giá hiệu quả mô hình trên cả tập huấn luyện và kiểm tra.

## CHUẨN HOÁ CỘT DỮ LIỆU AMOUT

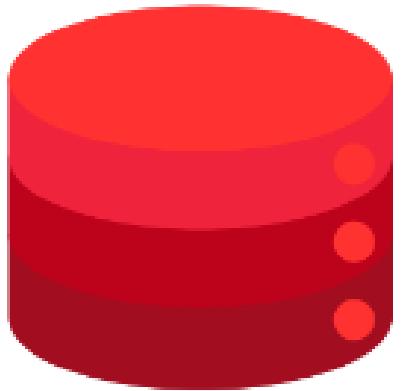
Để đảm bảo các đặc trưng có cùng quy mô và ảnh hưởng đồng đều trong quá trình huấn luyện mô hình.

## CÂN BẰNG DỮ LIỆU SMOTE

Để giúp mô hình không bị thiên vị.  
Undersampling: Không đảm bảo tính trọn vẹn của dữ liệu.  
Oversampling: Khó tránh khỏi overfitting.

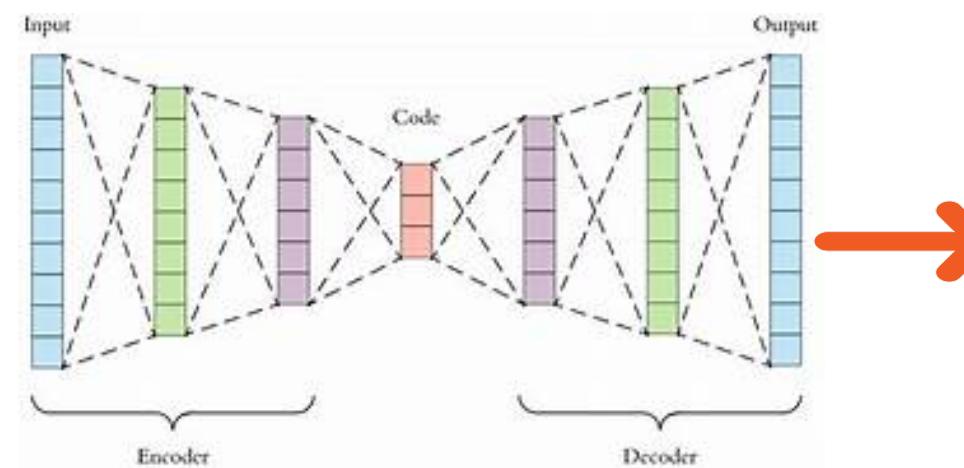
# ÁP DỤNG MÔ HÌNH

## Dữ liệu đầu vào



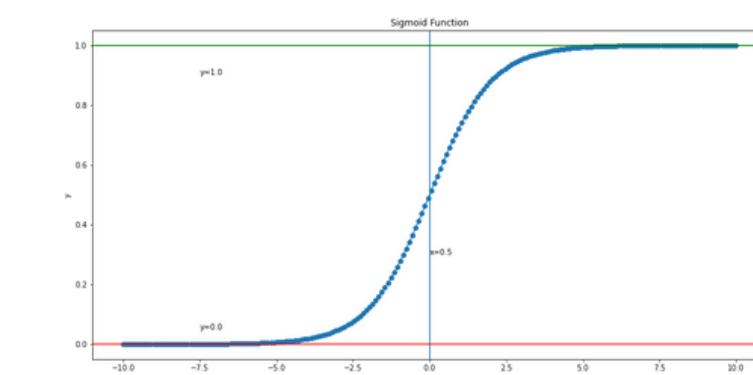
V1, V2, V3, V4 , V5 ,..., V28  
Amount

## Xử lý với Autoencoder



Dữ liệu đầu vào được chuẩn hóa và đưa vào Autoencoder để tái tạo.  
Lỗi tái tạo được tính toán cho mỗi giao dịch.

## Xử lý với Logistic Regression

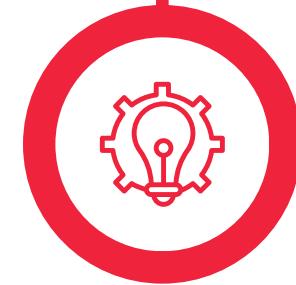
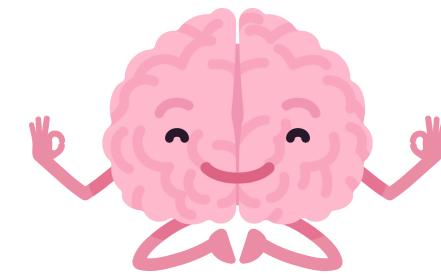


Lỗi tái tạo từ Autoencoder được dùng làm đầu vào cho Logistic Regression để phân loại giao dịch bình thường hoặc gian lận.

## Kết quả



Phân loại được giao dịch.



## Huấn luyện Autoencoder

- Mục đích của Autoencoder là học cấu trúc của các giao dịch ‘Normal’ để có thể phát hiện ra các điểm khác biệt.
- Sử dụng sai số tái tạo để gán nhãn: Những giao dịch lớn hơn ngưỡng được coi là ‘Fraud’, những giao dịch có sai số nhỏ hơn ngưỡng được coi là ‘Normal’.

## Cân bằng dữ liệu Smote

Áp dụng SMOTE để cân bằng số lượng mẫu ‘Fraud’ và ‘Normal’ giúp Logistic Regression học từ cả lớp thiểu số và đa số.



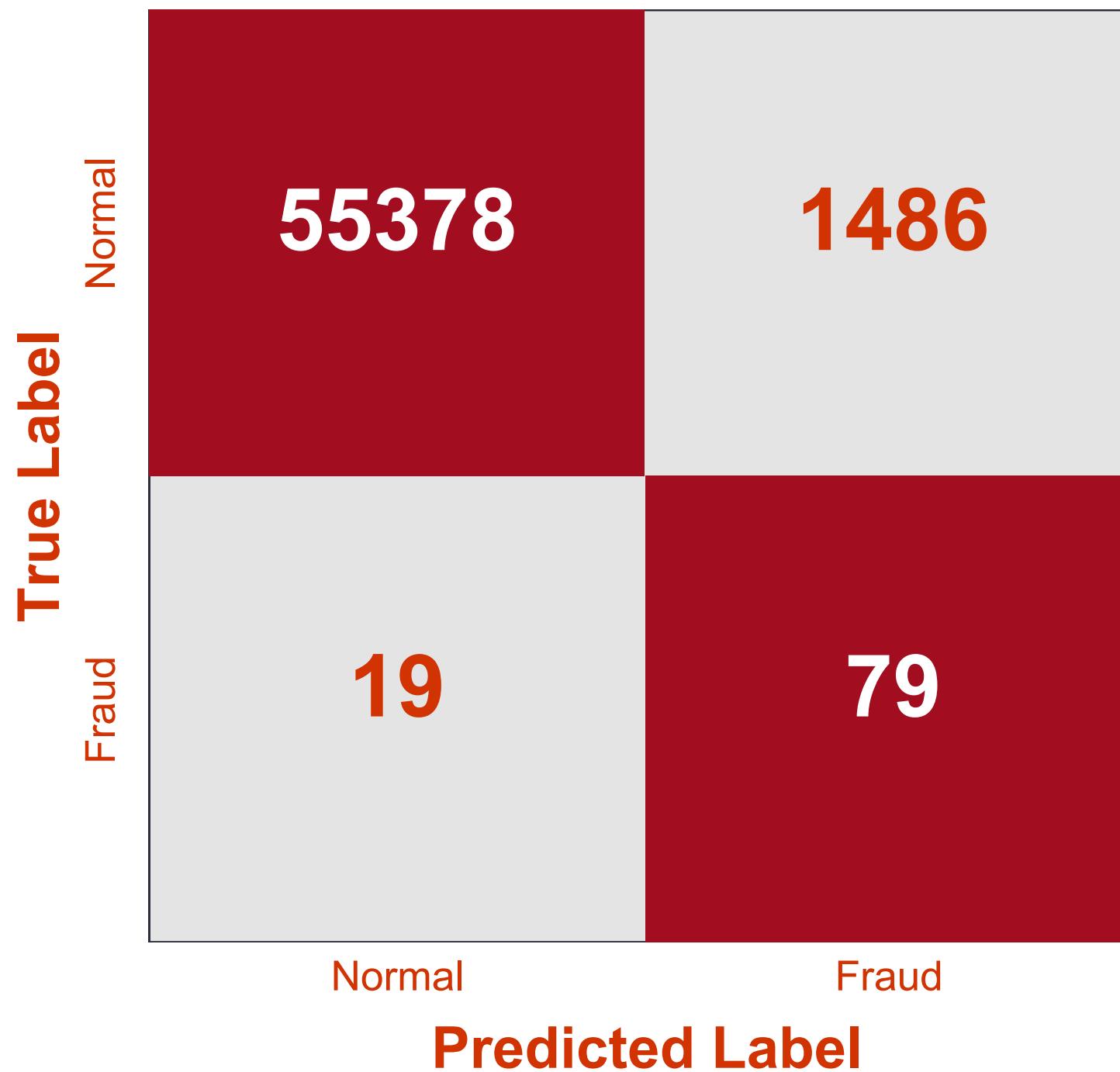
## Huấn luyện Logistic Regression

Logistic Regression sẽ sử dụng những đặc trưng của các giao dịch do Autoencoder phát hiện và dựa vào đó để dự đoán một cách chính xác hơn.

### III. KẾT QUẢ MÔ HÌNH



# Confusion Matrix



Accuracy (Độ chính xác)

**97.15%**

Recall (Khả năng phát hiện gian lận)

**80.61%**

Precision (Độ chính xác dự đoán gian lận)

**5.04%**

F1-score

**9.6%**

# Tối ưu phát hiện gian lận: Ưu tiên Recall và F1-score

## Tầm quan trọng của Recall

- **Thiệt hại từ False Negatives (FN):** gây thiệt hại trực tiếp cho tổ chức.
- **Bảo đảm an ninh tài chính:** Recall càng cao, hệ thống càng phát hiện ra nhiều gian lận. Điều này đồng nghĩa với việc tổ chức sẽ tránh được nhiều tổn thất tài chính.
- **Giảm thiểu thiệt hại dài hạn:** Phát hiện một giao lận gian lận có thể ngăn chặn những hành vi gian lận tiếp theo của cùng một đối tượng.

## Tối ưu hóa phát hiện gian lận với F1-score

- **F1-score phản ánh tốt về hiệu suất tổng thể của mô hình:** ưu tiên không bỏ sót gian lận, nhưng cũng không thể bỏ qua việc giảm thiểu cảnh báo nhầm.
- **F1-score hợp lý hơn Accuracy:** dù Accuracy cao nhưng không phản ánh đúng hiệu quả mô hình do mất công bằng dữ liệu.

# ĐÁNH GIÁ MÔ HÌNH: //

## Ưu điểm

**Phát hiện gian lận hiệu quả:** Ưu tiên recall giúp mô hình phát hiện được các giao dịch gian lận, giảm thiểu thiệt hại tài chính và bảo vệ tổ chức.

**Kết hợp hai kỹ thuật mạnh mẽ:**

+ **Autoencoder** phát hiện bất thường dựa trên lối tái tạo, hiệu quả với dữ liệu mất công bằng.

+ **Logistic Regression**: Cung cấp khả năng phân loại chính xác dựa trên lối tái tạo từ Autoencoder.



## Nhược điểm

**Precision thấp** dẫn đến nhiều cảnh báo nhầm gây rắc rối cho khách hàng và đội ngũ xử lý.

**Hiệu suất tổng thể F1-score thấp** cho thấy sự cân bằng giữa Recall và Precision chưa tối ưu, cần cải thiện.

**Chi phí thực hiện xác minh:** Cảnh báo nhầm cần kiểm tra thủ công, tốn kém và tiêu tốn thời gian.

**Độ phức tạp cao:** Tích hợp Autoencoder và Logistic có thể phức tạp trong triển khai và bảo trì.



# MÔ TẢ BÀI TOÁN

Giả sử áp dụng với trường hợp:	100,000 giao dịch
Với số lượng giao dịch gian lận:	150 giao dịch.
Chi phí bồi thường cho giao dịch gian lận không được phát hiện:	5,000,000 VND.
Chi phí bồi thường cho giao dịch gian lận bị phát hiện sai:	2,000,000 VND.
Thời gian xử lý 1 giao dịch nghi ngờ gian lận:	10 phút.



# Kết quả

SỐ GIAO DỊCH GIAN LẬN : 150

	<b>KHÔNG SỬ DỤNG MÔ HÌNH</b>	<b>SỬ DỤNG MÔ HÌNH</b>
SỐ GIAN LẬN BỊ PHÁT HIỆN SAI	75	120
SỐ GIAN LẬN KHÔNG BỊ PHÁT HIỆN SAI	75	30
TỔNG CHI PHÍ BỒI THƯỜNG	525.000.000	210.000.000
TỔNG THỜI GIAN XỬ LÍ	12.5	5

Tiết kiệm được: 525,000,000 VND - 210,000,000 VND = 315,000,000 VND

Tiết kiệm được: 12.5 giờ - 5 giờ = 7.5 giờ

## IV. KẾT LUẬN

### KHÔNG SỬ DỤNG MÔ HÌNH

**Chi phí bồi thường cho giao dịch gian lận không được phát hiện và phát hiện sai cao, gây thiệt hại tài chính đáng kể.**

**Thời gian xử lý giao dịch nghi ngờ gian lận** dài, ảnh hưởng đến trải nghiệm khách hàng và hiệu quả công việc.



### SỬ DỤNG MÔ HÌNH

**Giảm Chi Phí Bồi Thường:** Mô hình Autoencoder và Logistic Regression kết hợp giúp giảm thiểu số lượng giao dịch gian lận không được phát hiện và giảm số lượng cảnh báo gian lận sai, tiết kiệm chi phí bồi thường.

**Tiết Kiệm Thời Gian:** Cải thiện thời gian xử lý các giao dịch nghi ngờ gian lận, từ đó giảm tải công việc cho nhân viên và nâng cao hiệu quả công việc.



# GIÁ TRỊ KINH DOANH



## Tiết kiệm chi phí, giảm thiệt hại

Giảm thiểu thiệt hại từ các giao dịch gian lận không được phát hiện.

## Giảm chi phí điều tra và xử lý

Tiết kiệm chi phí liên quan đến điều tra và xử lý giao dịch gian lận.

## Cải thiện trải nghiệm khách hàng

Giảm thiểu số lượng giao dịch hợp lệ bị đánh giá sai, từ đó nâng cao sự hài lòng của Khách hàng.

## Tăng cường uy tín và lợi thế cạnh tranh

Tạo sự tin tưởng và gia tăng số lượng Khách hàng mới.

MVP



## Credit Card Fraud



Thank  
you!

## SỐNG NHƯ NHỮNG ĐOÁ HOA

A44948  
Phan Thu Hiền

A46277  
Trần Thị Ngọc Huế

A44639  
Phạm Huyền Trang

