

[MENU](#)[ANALYSIS](#)[CONTACT](#)[DATA ANALYSIS](#)

Sử dụng mô hình **machine learning** nhằm phát hiện **gian lận** trong **hoạt động giao dịch** của ngân hàng

Đội thi: ATTACKER ATTACKER





NỘI DUNG

01

GIỚI THIỆU CHUNG

Tổng quan về bộ dữ liệu và quy trình dự đoán

02

XỬ LÝ DỮ LIỆU

Tiến hành xử lý dữ liệu trước khi đưa vào mô hình

03

XÂY DỰNG MÔ HÌNH

Sử dụng các mô hình khác nhau và lựa chọn mô hình tốt nhất

04

DỰ ĐOÁN KẾT QUẢ

Tiến hành dự đoán và báo cáo kết quả





MENU

ANALYSIS

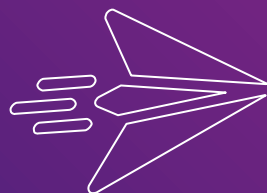
CONTACT

DATA ANALYSIS



01

GIỚI THIỆU CHUNG



[MENU](#)[ANALYSIS](#)[CONTACT](#)[DATA ANALYSIS](#)

GIỚI THIỆU CHUNG



Bộ dữ liệu

Từ ngân hàng

Thông tin chi tiết
của các giao dịch

48030 quan sát

66 biến

Thời gian khách hàng
thực hiện giao dịch, địa
chỉ, giới tính, nghề
nghiệp khách hàng, v.v

36 biến bị ẩn
thông tin

54 biến thiếu từ
40% dữ liệu trở lên





QUY TRÌNH XÂY DỰNG MÔ HÌNH





MENU

ANALYSIS

CONTACT

DATA ANALYSIS



02

XỬ LÝ DỮ LIỆU





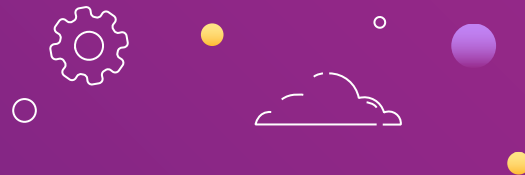
MENU

ANALYSIS

CONTACT

DATA ANALYSIS

XỬ LÝ DỮ LIỆU



Lỗi định dạng



Bỏ các biến không cần thiết

Lỗi thiếu dữ liệu



Kiểm tra mức độ tương quan





LỖ ĐỊNH DẠNG

time_1	time_2
2018-08-12T01:31:43Z	2018-08-12T01:31:43Z
2017-03-03T15:43:40.733Z	2017-03-03T15:43:40.733Z
2018-07-31T14:02:40Z	2018-07-31T14:02:40Z
2018-01-17T02:56:19.277Z	2018-01-17T02:56:19.277Z
2019-01-02T02:41:12.015Z	2019-01-02T07:12:45Z

trans_location
153 XÃ' Viáºłt Nghá»† TÃ©nh Quáo-n BÃ-nh Tháoình
649/40 Äía»†n BiÃãn Pháo§, P.25, Q.BÃ-nh Tháoình
ÄÄ Láºit, LÃcm Äá»“ng
38D1 NGUYá»...N THáo« THáo-P, KP2, P. TÃ,N PHONG, Q7
8 Tráoình HÆ°ng Äáoìo, P3, ÄÄ Láºit, LÃ

Định dạng datetime có
timezone

Lỗi định dạng mã hóa





LOẠI BỎ CÁC BIẾN KHÔNG CẦN THIẾT

Các biến về thời gian

time_1

time_2

date_1

date_2

.....

Các biến kiểu văn bản

address

mer_name

job

.....



[MENU](#)[ANALYSIS](#)[CONTACT](#)[DATA ANALYSIS](#)

LỖI THIẾU DỮ LIỆU



54 biến thiếu từ 40% dữ liệu trở lên



Điền dữ liệu bị thiếu theo những giá trị đã có



[MENU](#)[ANALYSIS](#)[CONTACT](#)[DATA ANALYSIS](#)

LỖI THIẾU DỮ LIỆU



time_1	time_2	Field_11	cat_1	cat_2	cat_3	cat_4	cat_5	date_1	mer_des	mul_rate	value	cat_6
2017-01-23T18:00:32.69Z	2017-01-23T18:00:32.69Z		C2	P1	1		1	#####		2.26	3,367,400	1

Các biến có xu hướng thiếu cùng nhau





CHIA BỘ DỮ LIỆU THÀNH HAI PHẦN

Bảng 10: Số lượng các giá trị bị thiếu của các biến có đủ giá trị time_1

Tên biến	Số lượng dữ liệu bị thiếu	Phần trăm dữ liệu bị thiếu (%)
id	0	0.00
label	0	0.00
time_1	0	0.00
time_2	0	0.00
Field_11	16312	70.95
cat_1	0	0.00
cat_2	0	0.00
cat_3	0	0.00
cat_4	16341	71.08
cat_5	0	0.00
date_1	8852	38.50
mer_des	15421	67.07
mul_rate	0	0.00
value	0	0.00
cat_6	0	0.00
num_date_review	0	0.00

Bảng 11: Số lượng các giá trị bị thiếu của các biến bị thiếu giá trị time_1

Tên biến	Số lượng dữ liệu bị thiếu	Phần trăm dữ liệu bị thiếu (%)
id	0	0.00
label	0	0.00
time_1	25039	100.00
time_2	25039	100.00
Field_11	25039	100.00
cat_1	25039	100.00
cat_2	25039	100.00
cat_3	25039	100.00
cat_4	25039	100.00
cat_5	25039	100.00
date_1	25039	100.00
mer_des	25039	100.00
mul_rate	25039	100.00
value	25039	100.00
cat_6	25039	100.00
num_date_review	25039	100.00





LỖI THIẾU DỮ LIỆU

Các biến thiếu trên 40%

Tiến hành loại bỏ

Các biến thiếu dưới 40%

Điền dữ liệu

Nhiều outlier

Trung vị (median)

Ít outlier

Trung bình (mean)





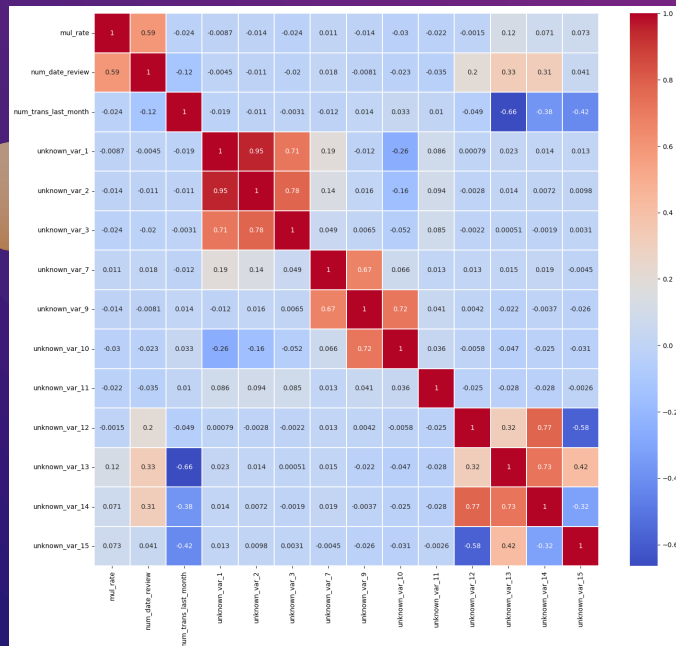
MENU

ANALYSIS

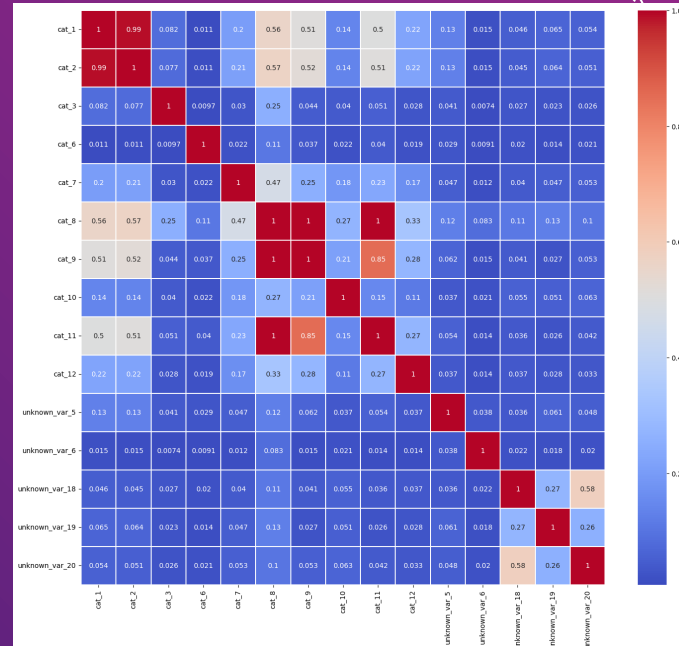
CONTACT

DATA ANALYSIS

KIỂM TRA MỨC ĐỘ TƯƠNG QUAN



Đồ thị hệ số tương quan của các biến số



Đồ thị hệ số tương quan của các biến phân loại



MENU

ANALYSIS

CONTACT

DATA ANALYSIS



03

XÂY DỰNG MÔ HÌNH





MENU

ANALYSIS

CONTACT

DATA ANALYSIS

MÔ HÌNH SỬ DỤNG

01

LOGISTIC REGRESSION

02

DEEP LEARNING (FFN)

03

RANDOM FOREST

04

SUPPORT VECTOR MACHINE (SVM)





ĐÁNH GIÁ MÔ HÌNH

01

ACCURACY

Độ chính xác của mô hình (lấy số dự đoán đúng chia cho toàn bộ các dự đoán)

03

RECALL

$TP/(TP+FN)$ cho chúng ta biết tỉ lệ positive được dự đoán đúng trên tổng số tất cả các positive

02

PRECISION

$TP/(TP+FP)$ cho chúng ta biết thực sự có bao nhiêu dự đoán Positive được dự đoán đúng

04

F1

Chỉ số F1 trung hòa giữa precision và recall





SO SÁNH GIỮA CÁC MÔ HÌNH

01

LOGISTIC REGRESSION

F1 (MODEL 1) = 0.093

F1 (MODEL 2) = 0.005

02

DEEP LEARNING (FFN)

F1 (MODEL 1) = 0.3445

F1 (MODEL 2) = 0.3032

03

RANDOM FOREST

F1 (MODEL 1) = 0.0393

F1 (MODEL 2) = 0.193

04

SUPPORT VECTOR MACHINE (SVM)

F1 (MODEL 1) = 0

F1 (MODEL 2) = 0





MENU

ANALYSIS

CONTACT

DATA ANALYSIS



04

DỰ ĐOÁN KẾT QUẢ





KẾT QUẢ MÔ HÌNH ĐƯỢC CHỌN



CHỈ SỐ	KẾT QUẢ
Accuracy	0.596
Precision	0.26
Recall	0.345
F1	0.209

- 59.6% kết quả được dự đoán là chính xác
- Trong các dự đoán positive có 34.5% dự đoán đúng và Recall bằng 0.209 đồng nghĩa với việc bỏ lỡ 79.1% giá trị positive
- Từ hai giá trị trên chúng ta có trung bình F1 bằng 0.26



MA TRẬN LỖI

- 674 biến là 0 được dự đoán là 1 (FN)
- 1346 biến là 1 được dự đoán là 0 (FP)
- Mô hình còn chưa thực sự hiệu quả
- Vấn đề phát sinh có thể nằm ở bước xử lý dữ liệu, do số lượng dữ liệu bị thiếu lớn
- Để khắc phục vấn đề này và cải thiện kết quả nghiên cứu thì cần phải nghiên cứu thêm và có thêm nhiều dữ liệu hơn về các biến số và thông tin của ngân hàng

		Dự đoán	
		0	1
Thực tế	0	2625	674
	1	1346	355





MENU

ANALYSIS

CONTACT

DATA ANALYSIS



GIAO DIỆN ỨNG DỤNG



ATTACKER ATTACKER





MENU

ANALYSIS

CONTACT

DATA ANALYSIS

Các chức năng

Chức năng xác thực

Cấp quyền truy cập

Dashboard báo cáo
quản trị bộ thông tin
về các giao dịch

Quản lý và phát hiện
các giao dịch gian
lận





Chức năng chính ứng dụng từ mô hình

Quản lý phát hiện gian lận



- Toàn bộ giao dịch được dự đoán
- Giao dịch gian lận
- Giao dịch không gian lận

Thống kê dữ liệu giao dịch



- Các bảng biểu, sơ đồ thống kê số giao dịch gian lận vs số giao dịch thông thường theo tháng/quý/năm
- Số tiền thất thoát do gian lận



[MENU](#)[ANALYSIS](#)[CONTACT](#)

TOÀN BỘ GIAO DỊCH

[DATA ANALYSIS](#)

MAIN MENU

[Dashboard](#)[Normal Transactions Datable](#)[Charts](#)

PAGES

[Authentication](#)[Errors](#)

Hi, Quang

Datable

[Dashboard](#) > Datable

10 entries per page

Search...

label	id	time_1	time_2	cat_1	cat_2	cat_3	cat_4	cat_5
Normal	31502							
Normal	31752	2/17/2017 6:51:53 PM	3/19/2019 9:42:37 AM	C1	P2	1	UI	1
Normal	63913	10/11/2019 2:49:16 PM	11/25/2019 2:10:04 AM	C1	P2	1	UI	1
Normal	87530							
Normal	24828	11/12/2019 3:31:21 PM	11/13/2019 9:55:15 AM	C1	P2	1		1
Normal	86442	10/15/2019 3:55:14 PM	10/16/2019 2:22:48 PM	C1	P2	1		1
Normal	87268							



[MENU](#)[ANALYSIS](#)[CONTACT](#)

GIAO DỊCH GIẢN LẶN

[DATA ANALYSIS](#)

MAIN MENU

[Dashboard](#)[Datatable](#)[Charts](#)

PAGES

[Authentication](#)[Errors](#)[Hi, Quang](#)

Fraud Transactions Datatable

[Dashboard](#) > [Datatable](#) > [Fraud Transactions Datatable](#)

10

entries per page

Search...

label	id	time_1	time_2	cat_1	cat_2	cat_3	cat_4	cat_5
Fraud	52725	3/15/2019 2:27:31 PM	3/19/2019 9:42:37 AM	C1	P2	1		1
Fraud	78992	2/28/2019 8:27:46 AM	2/28/2019 10:06:39 AM	C1	P2	1	GX	1
Fraud	52965	8/12/2019 5:16:03 PM	8/15/2019 8:48:17 AM	C1	P2	1		1
Fraud	80918	11/13/2018 10:20:30 AM	12/5/2018 3:10:05 AM	C1	P2	1	UI	1
Fraud	20927							
Fraud	23203	12/14/2019 8:16:18 AM	12/16/2019 9:16:12 AM	C2	P1	1		1



MENU

ANALYSIS

CONTACT

GIAO DỊCH THÔNG THƯỜNG

DATA ANALYSIS



MAIN MENU

[Dashboard](#)[Datatable](#)[Charts](#)

PAGES

[Authentication](#)[Errors](#)

Hi, Quang

Normal Transactions Datatable

[Dashboard](#) > [Datatable](#) > Normal Transactions Datatable

10

entries per page

Search...

label	id	time_1	time_2	cat_1	cat_2	cat_3	cat_4	cat_5
Normal	31502							
Normal	31752	2/17/2017 6:51:53 PM	3/19/2019 9:42:37 AM	C1	P2	1	UI	1
35%	63913	10/11/2019 2:49:16 PM	11/25/2019 2:10:04 AM	C1	P2	1	UI	1
Normal	87530							
Normal	24828	11/12/2019 3:31:21 PM	11/13/2019 9:55:15 AM	C1	P2	1		1
Normal	86442	10/15/2019 3:55:14 PM	10/16/2019 2:22:48 PM	C1	P2	1		1
Normal	87268							

[MENU](#)[ANALYSIS](#)[CONTACT](#)

THỐNG KÊ SỐ GIAO DỊCH GIẢN LẬN

[DATA ANALYSIS](#)

MAIN MENU

[Dashboard](#)[Fraud Detection Management](#)[Fraud Detection Statistics](#)

PAGES

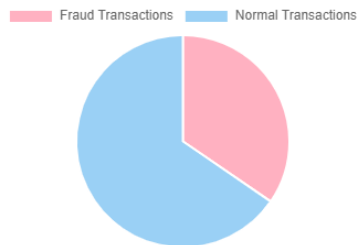
[Authentication](#)[Errors](#)

Hi, Quang

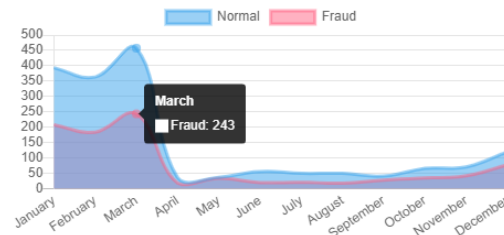
Fraud Detection Statistics

[Dashboard](#) > [ChartJS](#)

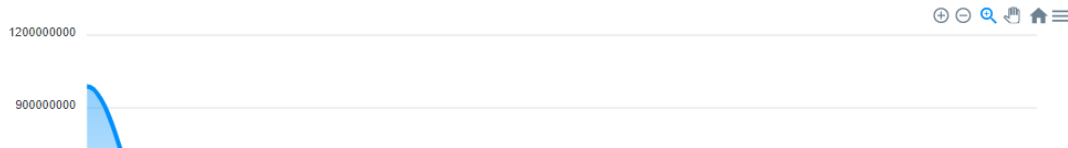
Pie Chart Fraud Proportion



Number of Fraud and Normal Trans Monthly in 2017



Total Value of Normal and Fraud Transactions monthly 2017



THỐNG KÊ TỔNG GIÁ TRỊ GIAO DỊCH GIẢN LẬN

DATA ANALYSIS



MAIN MENU

Dashboard

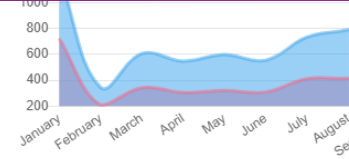
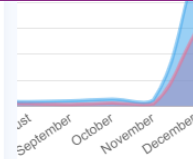
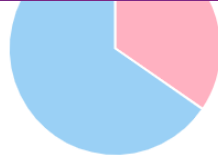
Fraud Detection
Management

Fraud Detection Statistics

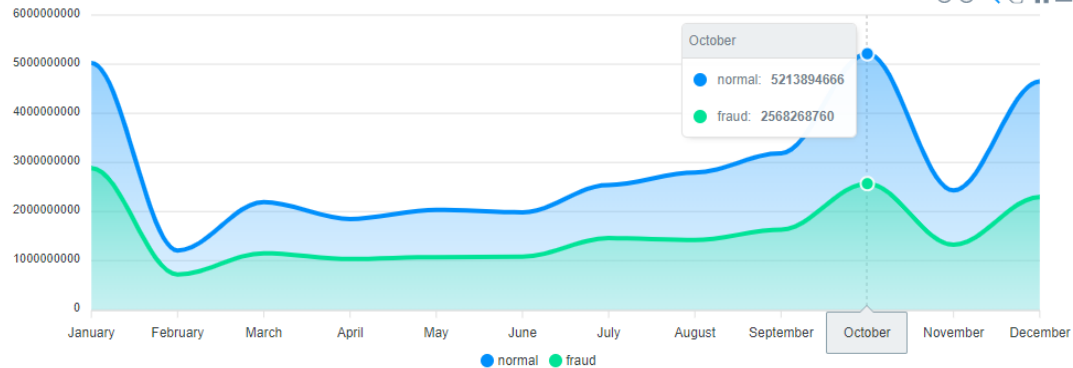
PAGES

Authentication

Errors



Total Value of Normal and Fraud Transactions monthly 2019





MENU

ANALYSIS

CONTACT

DATA ANALYSIS

THANK YOU

FOR LISTENING

