VEF Final Project. Credit Cards Fraud Detection.

NHÓM 5.

PHAN TÙNG DƯƠNG.

NGUYỄN MẠNH TUẤN.

Table of contents

1. Tổng quan mục tiêu dự án.

- Các giao dịch thông qua thẻ tín dụng có nguy cơ bị lừa đảo.
- Mục tiêu dự án: Phát hiện các giao dịch lừa đảo.

2. Cách tiếp cận vấn đề.

- Dựa trên dữ liệu (Data driven).
- Dựa trên dữ liệu có sẵn về 2 loại giao dịch lừa đảo (1) và không lừa đảo (0), học quy tắc để nhận biết một giao dịch có lừa đảo hay không.

3. Tổng quan về dữ liệu.

- Dữ liệu Credit Card Fraud Detection từ Kaggle. (https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud).
- Dữ liệu có 284,807 bản ghi, với 492 bản ghi về giao dịch lừa đảo.
- Dữ liệu có 28 thuộc tính (đã được trích xuất thông qua chuyển đổi PCA. Ngoài ra, còn 2 thuộc tính nguyên bản là Time và Amount.
- Tỷ lệ giao dịch lừa đảo/Tổng số giao dịch: 0.172%

4. Giới thiệu về dữ liệu.

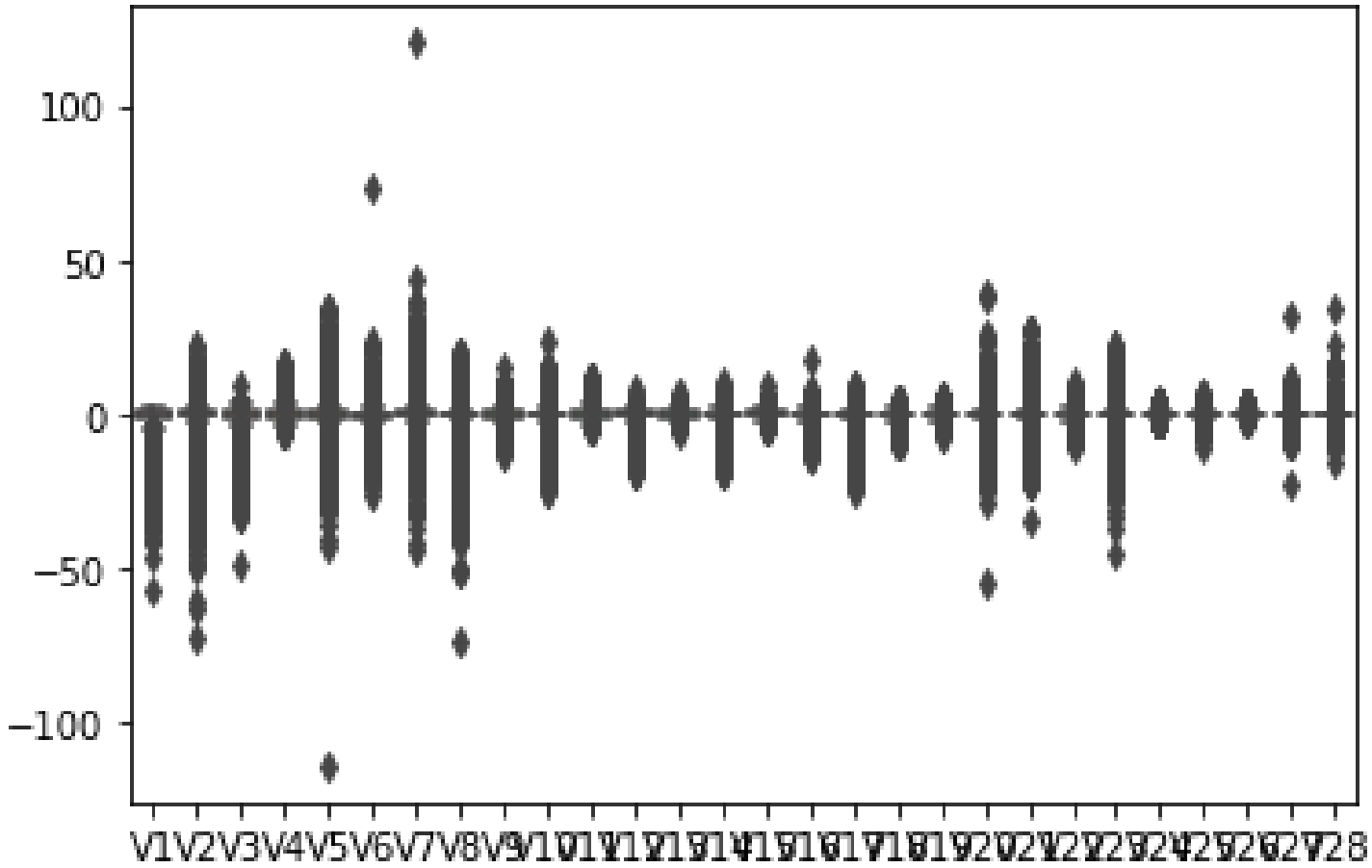
Giới thiệu chi tiết cột, data type, định nghĩa mỗi bảng dữ liệu.

Column	**Data type**	**Description**
V1, V2,…,V28	Float	PCA transformed features.
Time	Float	Seconds elapsed between each transactions and the first transaction in the dataset.
Amount	Float	Transaction amount.

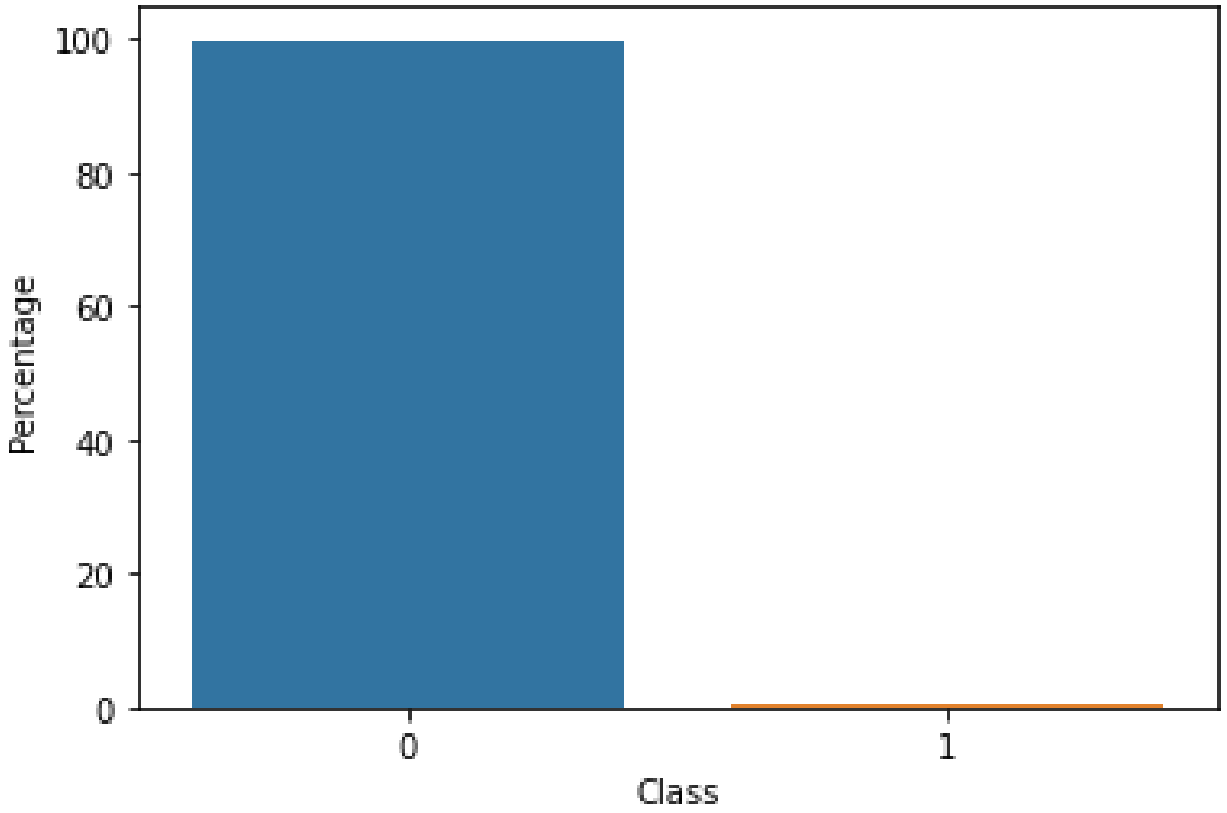
5. Phân tích khám phá.

	Time	Amount
Count	284807.000000	284807.000000
Mean	94813.859575	88.349619
Std	47488.145955	250.120109
Min	0.000000	0.000000
25%	54201.500000	5.600000
50%	84692.000000	22.000000
75%	139320.500000	77.165000
Max	172792.000000	25691.160000

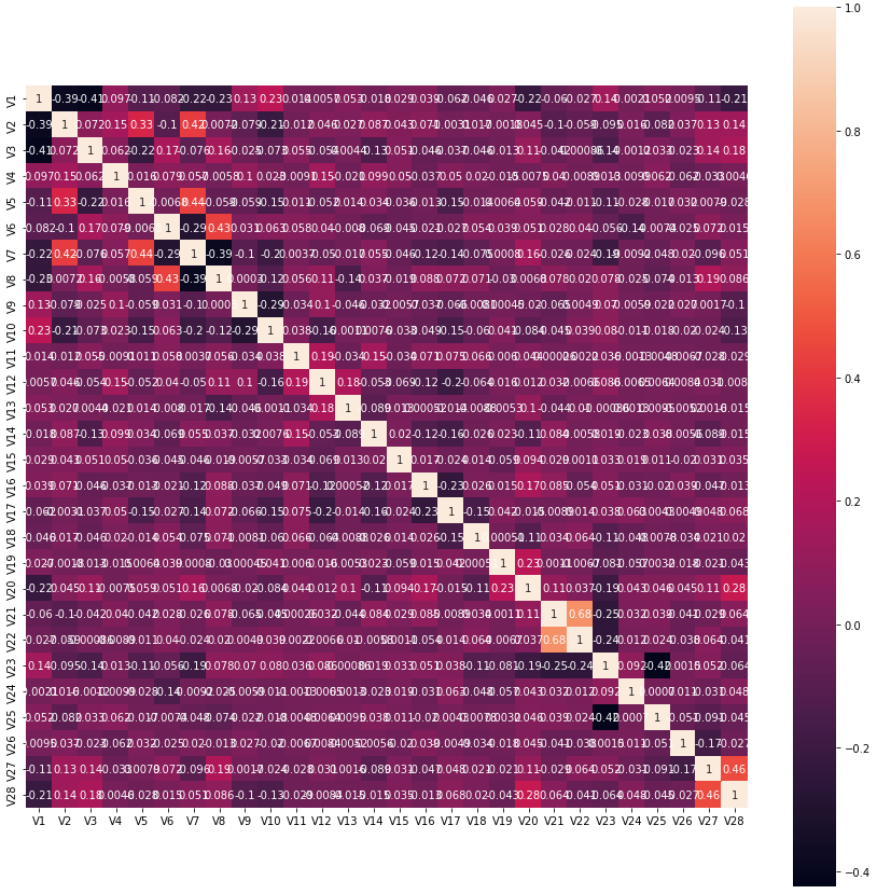
5. Phân tích khám phá.



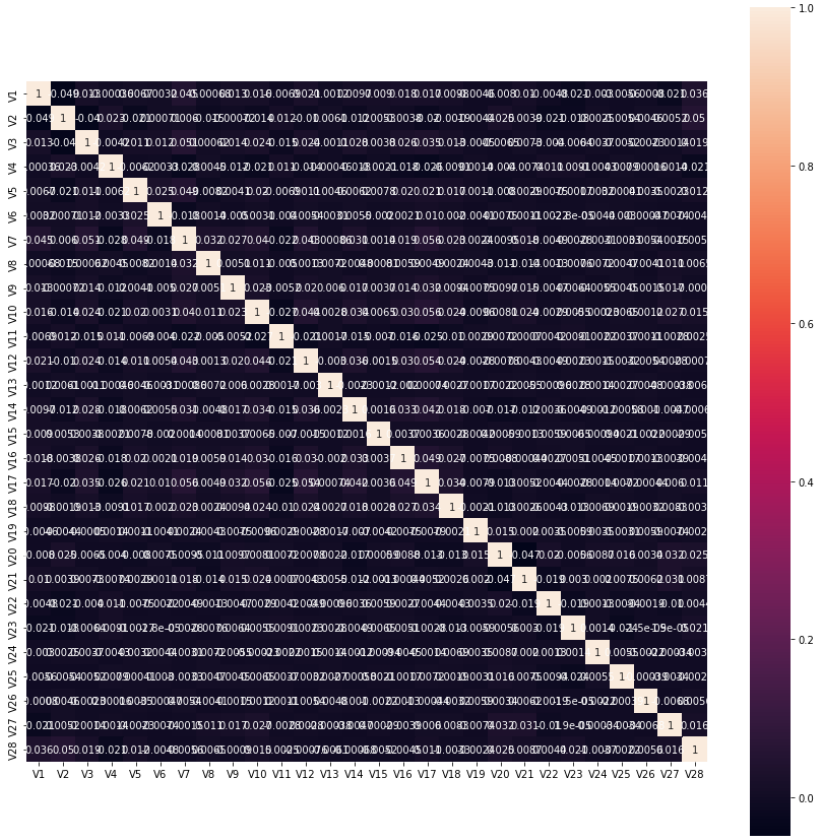
5. Phân tích khám phá.



5. Phân tích khám phá – Spearman matrix.



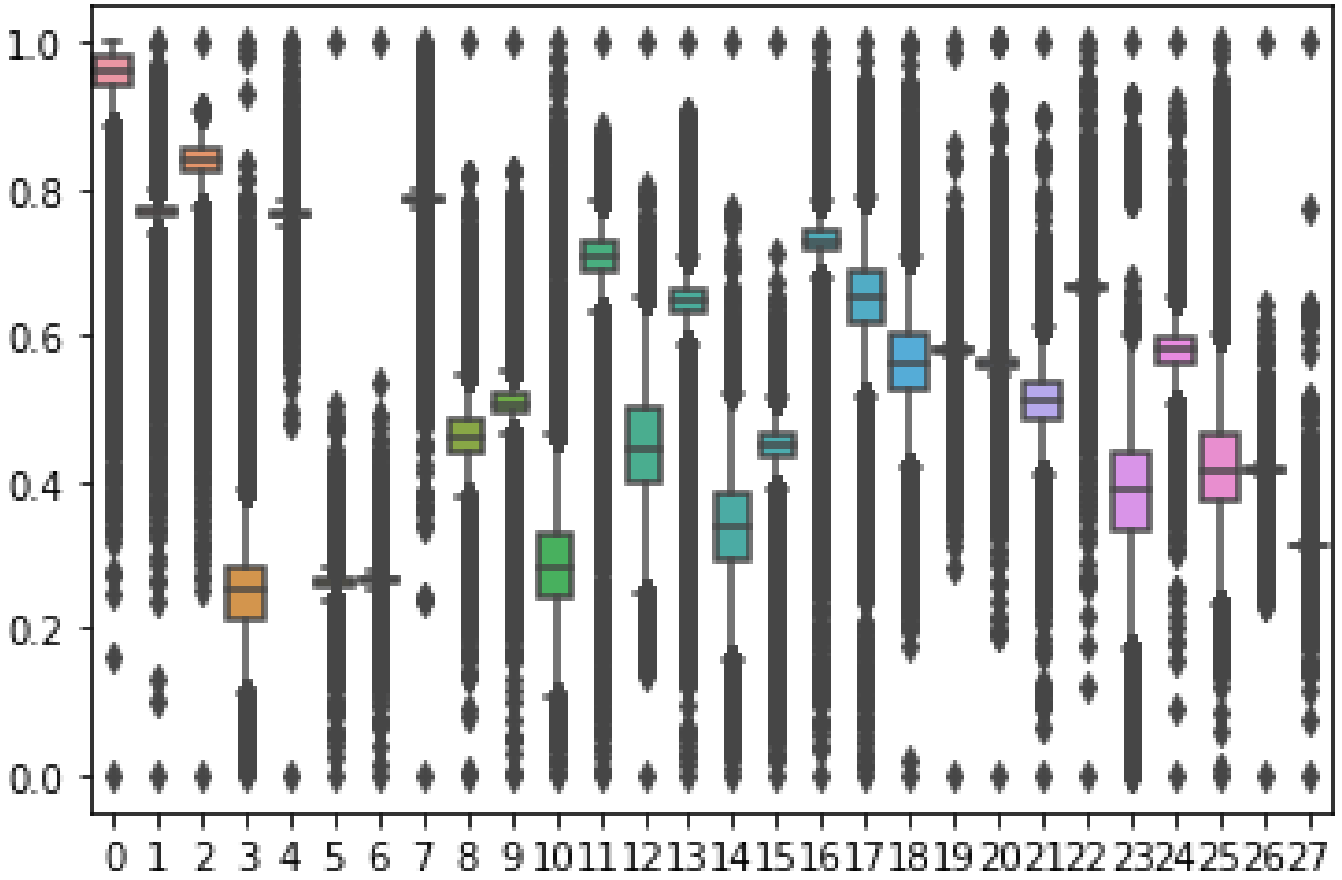
5. Phân tích khám phá – Pearson matrix.



6. Cách giải quyết và kết quả

- Giải quyết vấn đề Scale dữ liệu.
- Giải quyết vấn đề dữ liệu Imbalanced.
- Train Test Split.
- Giải quyết bài toán phân loại.
- Kết quả.

6.1. Scale dữ liệu – MinMax Scaler.



6.2. Giải quyết vấn đề dữ liệu Imbalanced.

- Sử dụng SMOTE (Synthetic Minority Oversampling Technique).

```
[ ]   1 df_label.value_counts()
```

```
0    284315
1    284315
Name: Class, dtype: int64
```

6.3. Train Test Split.

- Random State: 42.
- Test Size: 0.2.

6.4. Giải bài toán phân loại.

- Logistic Regression.
- Decision Tree.
- Random Forest.

6.5. Kết quả chạy.

Metrics sử dụng: Area Under the Precision Recall Curve (AUPRC).

Phương pháp sử dụng.	AUPRC
Logistic Regression.	0.4656359412344510 7
Decision Tree.	0.4975694599048091 6
Random Forest.	9.651323986170635e-05

7. Kết luận.

- Các mô hình Logistic Regression, Decision Tree, Random Forest đều cho ra kết quả phân loại tốt.
- Mô hình cho kết quả phân loại tốt nhất là Decision Tree (AUPRC = 0.4975694599048091 6).