FPT University

Môn học: MAS291

Lớp: SE1821

Khảo sát các phương pháp xác suất thống kê trong phát hiện bất thường

Nhóm 1:

Khưu Trọng Quân - SE192441
Huỳnh Khả Tú - SE182134
Lê Đức Trung Thi - DE180553
Phan Tấn Hải - SE180327
Nguyễn Phúc Đạt - SE182125
Nguyễn Xuân Trường - SE182688
Nguyễn Thị Lộc Nhi - SE182735
Nguyễn Vĩnh Khang - SE182879

Giảng viên:

Cô Lý Ánh Dương

FPT UNIVERSITY

Bảng phân công nhiệm vụ

STT	Họ Tên	Phân công nhiệm vụ
1	Khưu Trọng Quân	Code, thực nghiệm, slide + thuyết trình
2	Huỳnh Khả Tú	IQR method, thực hiện so sánh và cải tiến
3	Lê Đức Trung Thi	Tìm kiếm dataset, tiền xử lý dữ liệu
4	Phan Tấn Hải	Z-score Gaussian Distribution
5	Nguyễn Phúc Đạt	Lý do chọn đề tài, vẽ đồ thị
6	Nguyễn Xuân Trường	CDF Poisson Distribution
7	Nguyễn Thị Lộc Nhi	Phát biểu bài toán, khái niệm cơ bản
8	Nguyễn Vĩnh Khang	PMF Poisson Distribution

1 Phát biểu bài toán

Phát hiện bất thường là một lĩnh vực quan trọng trong phân tích dữ liệu, với ứng dụng rộng rãi trong bảo mật, tài chính, y tế và sản xuất. Mục tiêu là xác định các điểm dữ liệu bất thường, giúp phát hiện sự cố, lỗi hệ thống hoặc gian lận. Trong bối cảnh dữ liệu ngày càng lớn, việc phát hiện bất thường hiệu quả giúp đảm bảo an toàn và tối ưu hóa hệ thống. Nghiên cứu này khảo sát một số phương pháp xác suất thống kê như một cách tiếp cận cơ bản để nhận diện mẫu dữ liệu ngẫu nhiên, đồng thời so sánh với dữ liệu nhãn gốc nhằm đánh giá hiệu quả và đề xuất các hướng cải tiến nhằm nâng cao khả năng phát hiện bất thường.

2 Phân tích bài toán

2.1 Giới thiệu (Introduction)

2.1.1 Dữ liệu đầu vào (Input)

Một file creditcard.csv [1] (đã được lưu vào *github*) đầu vào bao gồm các cột:

- Cột 1: Time - biểu diễn thời gian theo giây, mỗi giây có nhiều lần rút tiền
- Cột 2: Amount - số tiền rút ra sau mỗi lần rút tiền
- Cột 3: Class - dữ liệu được đánh dấu là bình thường (giá trị cột class bằng 0) hay bất thường (giá trị cột class bằng 1)

2.1.2 Dữ liệu đầu ra (Output)

- Mô hình xuất ra các giá trị so sánh độ chính xác của các phương pháp xác suất thống kê với dữ liệu gốc (cột Class)

2.1.3 Giải thích, định nghĩa từ khóa [2]

Để hiểu rõ hơn về dự án (project), việc định nghĩa, khái niệm hóa các keywords là vô cùng quan trọng.

- **Dữ liệu bất thường** (hay còn gọi là dữ liệu ngoại lai, outlier,...) là các điểm dữ liệu có sự khác biệt đáng kể so với phần lớn dữ liệu còn lại. Trong thực tế, dữ liệu bất thường xuất hiện bởi một số nguyên nhân như gian lận tài chính (giao dịch bất thường), lỗi hệ thống (dữ liệu sai lệch), thay đổi đột ngột trong xu hướng (mùa lễ hội, sự kiện đặc biệt), sự cố IT (tấn công DDoS), hoặc bất thường trong dữ liệu y tế (kết quả xét nghiệm bất thường),...

- **Phát hiện bất thường** (Anomaly detection) là quá trình xác định các điểm dữ liệu, sự kiện hoặc quan sát khác biệt đáng kể so với phần lớn dữ liệu còn lại. Trong thực tế, các điểm bất thường này có thể biểu thị lỗi hệ thống, gian lận, thay đổi đột ngột hoặc các sự kiện quan trọng cần được chú ý.

- **Ngưỡng** (threshold) là một giá trị giới hạn dùng để xác định dữ liệu có phải là bất thường hay không. Nếu một giá trị vượt qua ngưỡng hoặc nằm ngoài vùng phạm vi của các ngưỡng, nó có thể được coi là bất thường.

- **True positive** (TP - dương tính thật): dữ liệu trên mô hình và thực tế đều là bất thường

- **True negative** (TN - âm tính thật): dữ liệu trên mô hình và thực tế đều là bình thường

- **False positive** (FP - dương tính giả): dữ liệu trên mô hình là bất thường trong khi thực tế đều là bình thường

- **False negative** (FN - âm tính giả): dữ liệu trên mô hình là bình thường trong khi thực tế đều là bất thường

- **Accuracy** là chỉ số đo lường độ chính xác của mô hình bằng cách tính tỉ lệ dự đoán đúng trên tổng số mẫu, được tính bằng công thức:

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

- **Precision** là tỷ lệ mẫu dự đoán đúng trong số tất cả các mẫu được dự đoán là bất thường. Precision được xác định bằng công thức:

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

- **Recall** là tỷ lệ mẫu thực sự bất thường được dự đoán đúng so với tổng số mẫu bất thường. Recall được xác định bằng công thức:

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

- **F1-score** cân bằng Precision và Recall, được xác định bằng công thức:

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

2.1.4 Các phương pháp xác suất thống kê

- Định lý phân phối Poisson [3]: một phân phối xác suất rời rạc mô tả số lần xảy ra của một sự kiện trong một khoảng thời gian hoặc không gian cố định. Có 2 hàm quan trọng đối với phân phối Poisson (hàm khối xác suất PMF và hàm phân phối tích lũy):
  - Hàm khối xác suất (PMF - Probability Mass Function): mô tả xác suất của một giá trị cụ thể k xảy ra, được tính theo công thức như sau:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{1}$$

    Trong đó:
    - λ là trung bình số lần xảy ra sự kiện trong một khoảng thời gian nhất định
    - k là số lần sự kiện xảy ra
    - e là hằng số Euler (2.718)
  - Hàm phân phối tích lũy (CDF - Cumulative Distribution Function): mô tả xác suất tích lũy, tức là xác suất để biến ngẫu nhiên nhận giá trị nhỏ hơn hoặc bằng một số cụ thể k, được tính theo công thức:

$$P(X \leq k) = \sum_{i=0}^{k} \frac{\lambda^i e^{-\lambda}}{i!} \tag{2}$$

    Trong đó:
    - λ là giá trị trung bình (mean) của phân phối Poisson
    - k là số lần sự kiện xảy ra

* e là hằng số Euler (2.718)
* i là biến số đếm cho hàm tổng

- Định lý phân phối Chuẩn (Gaussian distribution) [4]: một phân phối xác suất liên tục có dạng hình chuông và đối xứng quanh giá trị trung bình, đặc trưng bởi hai tham số μ và σ. Ở đây, ta sử dụng phương pháp phát hiện giá trị bất thường dựa trên Z-score trong phân phối Gaussian, được biểu diễn dưới dạng công thức:

$$Z = \frac{x - \mu}{\sigma} \tag{3}$$

  Trong đó:

  – x là giá trị quan sát
  – μ là giá trị trung bình của dữ liệu
  – σ là độ lệch chuẩn

- Phương pháp khoảng tứ phân vị (IQR - Interquartile Range) [5]: phương pháp không giả định phân phối dữ liệu, nên phù hợp cho dữ liệu có phân phối lệch hoặc có outliers mạnh, được dùng để đo độ phân tán của dữ liệu bằng cách tính khoảng cách giữa Q1 (phân vị thứ nhất - 25%) và Q3 (phân vị thứ ba - 75%). Để áp dụng phương pháp này, ta tiến hành tính toán dựa trên công thức:

$$IQR = Q3 - Q1 \tag{4}$$

  Trong đó:

  – $Q1$ là phân vị thứ nhất (25% của dữ liệu).
  – $Q3$ là phân vị thứ ba (75% của dữ liệu).
  – IQR đo khoảng giữa 50% dữ liệu trung tâm, giúp loại bỏ ảnh hưởng của giá trị cực đoan (outliers).

Liên quan đến lý do lựa chọn các hàm phân phối xác suất thống kê cho tác vụ phát hiện bất thường, chúng có độ tương thích nhất định với yêu cầu của bài toán. Trước tiên, phân phối Poisson thường được sử dụng để phát hiện bất thường vì nó phù hợp trong việc tìm kiếm những dữ liệu có độ hiếm

cao. Tiếp theo, Z-score trong phân phối chuẩn (Gaussian) giúp phát hiện bất thường bằng cách đo lường mức độ chênh lệch của một điểm dữ liệu so với giá trị trung bình theo đơn vị độ lệch chuẩn. Đối với dữ liệu tuân theo phân phối Gaussian, khoảng 99.7% dữ liệu nằm trong khoảng [-3σ, +3σ]. Do đó, những dữ liệu nằm ngoài khoảng này có thể được xem là bất thường. Cuối cùng, phương pháp sử dụng IQR xác định giá trị bất thường bằng cách kiểm tra xem một giá trị có nằm ngoài phạm vi cho phép hay không, từ đó xác định tính bất thường của nó.

2.1.5 Các bước thực hiện

1. Tiền xử lý dữ liệu: chuyển đổi đơn vị của dữ liệu Time và tính log của dữ liệu Amount có sẵn để thuận tiện trong tính toán phân phối Poisson và Gaussian.

2. Áp dụng công thức tính toán của từng loại phương pháp đã đề cập ở phần 2.1.4 với giá trị Time hoặc Amount (hoặc log_amount với phân phối Gaussian).

3. Thiết lập các ngưỡng tiêu chuẩn (thresholds) để phân biệt dữ liệu bình thường và dữ liệu bất thường.

4. Xuất ra dataset mới với các cột dữ liệu khác tương đương với sự xác định giá trị bình thường (giá trị của dòng đó bằng 0) và giá trị bất thường (giá trị của dòng đó bằng 1).

5. Đánh giá bằng các chỉ số tiêu chuẩn (Accuracy, Precision, Recall, F1-score) để xác định độ chính xác của từng loại phân phối với dữ liệu gốc (dựa trên cột Class), đồng thời đưa ra bảng so sánh các loại phân phối với dữ liệu gốc thông qua các chỉ số tiêu chuẩn.

6. Đề xuất phương pháp cải tiến và so sánh hiệu suất.

7. Tổng kết về ưu nhược điểm của phương pháp xác suất thống kê, sau đó đưa ra Future works.

2.2 Đóng góp chính (Main Contribution)

2.2.1 Dữ liệu đầu vào

Dữ liệu đầu vào có dạng như Figure 1. Như đã đề cập ở phần 2.1.1, cột đầu tiên của dữ liệu là thời gian (tính theo giây; nếu có trùng lặp, nghĩa là tại giây đó có nhiều giao dịch). Cột tiếp theo là số tiền giao dịch tại thời điểm tương ứng, và cột cuối cùng là nhãn: 0 nếu giao dịch là bình thường (không gian lận) và 1 nếu giao dịch là bất thường (gian lận).

Time	Amount	Class
0	149.62	0
0	2.69	0
1	378.66	0
1	123.5	0
2	69.99	0
2	3.67	0
4	4.99	0
7	40.8	0
7	93.2	0
9	3.68	0
10	7.8	0
10	9.99	0
10	121.5	0

406	0	1
472	529	1

172777	0.89	0
172778	9.87	0
172780	60	0
172782	5.49	0
172782	24.05	0
172783	79.99	0
172784	2.68	0
172785	2.69	0
172786	0.77	0
172787	24.79	0
172788	67.88	0
172788	10	0
172792	217	0

Figure 1: Dataset for project input

2.2.2 Tiền xử lý dữ liệu

Bắt đầu với bước 1, ta thực hiện tiền xử lý dữ liệu Time và Amount để làm cơ sở tính toán cho bước 2

- Đối với dữ liệu Time, ta cần tiến hành chuyển đổi sang đơn vị phút để tương tác với dữ liệu liên quan đến số sự kiện xảy ra trong một khoảng thời gian cố định (Formula 5). Sau đó tính trung bình số giao dịch mỗi phút để xác định lambda_poisson λ (Formula 6).

$$\text{Minute} = \frac{\text{Time}}{60} \tag{5}$$

$$\lambda = \frac{\sum \text{số giao dịch trong mỗi phút}}{\text{số phút quan sát}} \tag{6}$$

- Tính log cho dữ liệu Amount để thuận lợi cho việc tính toán phân phối Gaussian (kết quả tính toán được biểu diễn ở Figure 2). Công thức được sử dụng bằng cách tính logarit cơ số e với Amount đã cộng thêm 1 đơn vị (tránh trường hợp Amount = 0):

$$\text{Log\_Amount} = \ln(1 + \text{Amount}) \tag{7}$$

Time	Amount	Class	Minute	Log_Amount
0	149.62	0	0	5.014760109
0	2.69	0	0	1.305626458
1	378.66	0	0	5.939276115
1	123.5	0	0	4.824305716
2	69.99	0	0	4.262539022
2	3.67	0	0	1.541159072
4	4.99	0	0	1.790091412
7	40.8	0	0	3.73289634
7	93.2	0	0	4.545420182
9	3.68	0	0	1.54329811
10	7.8	0	0	2.174751721
10	9.99	0	0	2.396985768
10	121.5	0	0	4.80811103

Figure 2: LogAmount calculation

2.2.3 Tính toán xác suất thống kê

Như đã đề cập ở bước 2, ta áp dụng các công thức tính toán của từng phương pháp với giá trị Time hoặc Amount (đã qua hoặc chưa qua tiền xử lý). Thực nghiệm được thực hiện bằng ngôn ngữ lập trình Python trên IDE Visual Studio Code (VS Code), tính toán dữ liệu trong tập dataset theo từng phương pháp phân phối xác suất thống kê. Các công thức và giải thích chi tiết về các phương pháp này đã được trình bày ở phần 2.1.4.

- Phân phối Poisson: áp dụng các hàm PMF và CDF để tính toán phân phối đối với dữ liệu Time (dữ liệu rời rạc thỏa mãn điều kiện áp dụng phân phối). Kết quả thu được thể hiện trong Figure 3 và 4

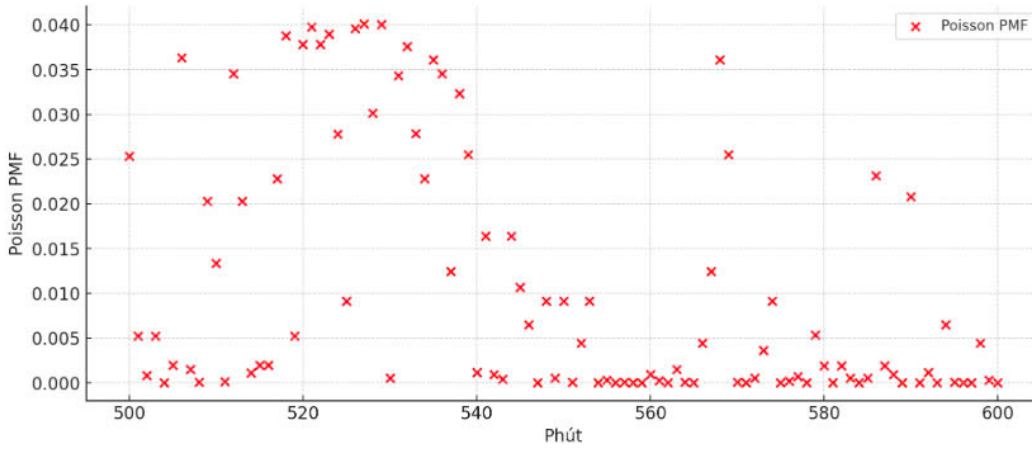

Figure 3: Poisson PMF từ phút 500 đến 600

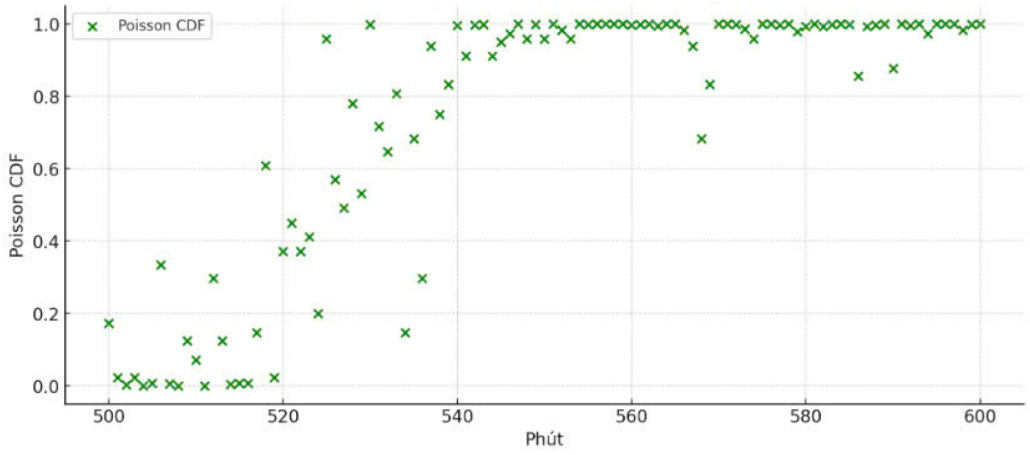

Figure 4: Poisson CDF từ phút 500 đến 600

- Phân phối Gaussian: áp dụng Z-score để tính toán phân phối cho dữ liệu Amount. Kết quả thu được được biểu diễn trong khoảng 30.000–32.000 giây để làm rõ sự phân biệt và ứng dụng của Z-score, như thể hiện trong Figure 5.

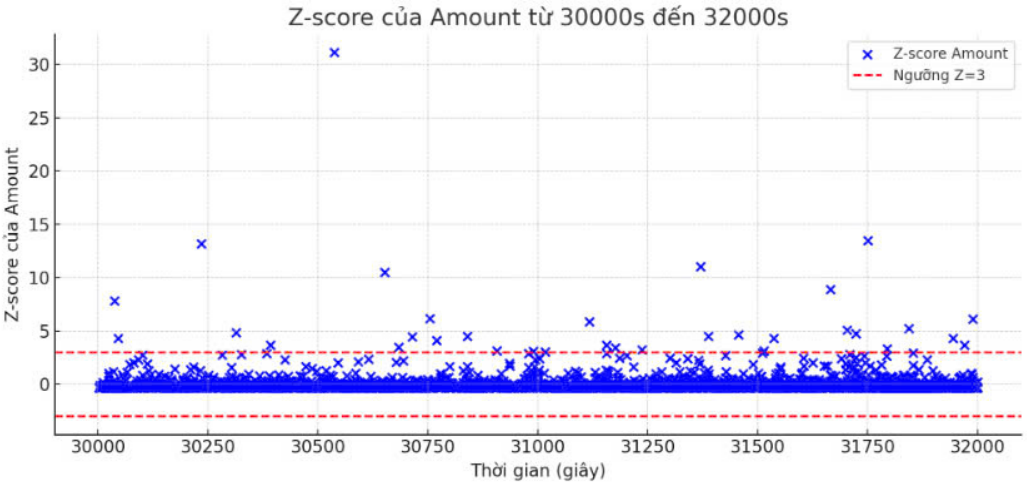

Figure 5: Z-score của Amount từ 30000s đến 32000s

- Phương pháp khoảng tứ phân vị (IQR): Từ việc lấy 25% dữ liệu dưới ($Q1$) và 75% dữ liệu dưới ($Q3$), ta có thể tính được IQR để đo độ phân tán của 50% dữ liệu trung tâm. Từ đó, ta có thể xác định ngưỡng (threshold) để tìm ra dữ liệu bất thường. Kết quả IQR từ dữ liệu được mô tả thông qua Figure 6.

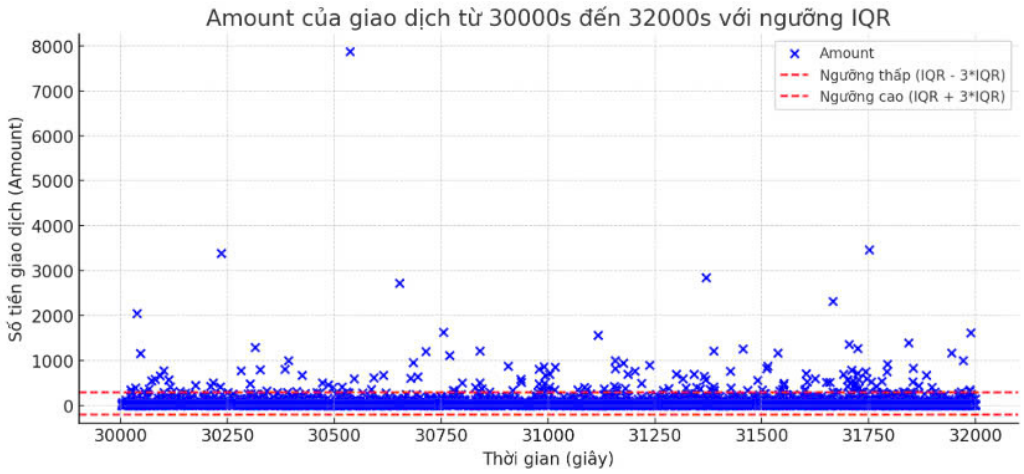

Figure 6: Amount của giao dịch từ 30000s đến 32000s với ngưỡng IQR

2.2.4 Thiết lập ngưỡng (threshold) tiêu chuẩn

Để phân biệt dữ liệu thuộc loại bình thường hay bất thường, việc thiết lập ngưỡng là vô cùng quan trọng. Nếu không có ngưỡng, rất khó xác định ranh giới giữa hai loại dữ liệu này.

Trong bất kỳ phương pháp phân phối xác suất thống kê nào, ta đều cần thiết lập một hoặc nhiều ngưỡng. Có hai loại ngưỡng chính trong phát hiện dữ liệu bất thường: ngưỡng tĩnh (static threshold) và ngưỡng động (dynamic threshold). Ngưỡng tĩnh được xác định trước bằng một giá trị cố định, trong khi ngưỡng động được tính toán dựa trên phân phối của dữ liệu.

Trong dự án này, ta thiết lập ngưỡng tĩnh (có thực nghiệm) đối với phân phối Poisson, ngưỡng tĩnh (mặc định) với phân phối Gaussian và ngưỡng động với phương pháp IQR. Ưu điểm của ngưỡng tĩnh là có thể tối ưu hóa việc phát hiện phát hiện nhiều gian lận hơn, nhưng lại có thể dính nhiều false positives (FP). Ngược lại, ngưỡng động giúp giảm FP bằng cách điều chỉnh theo phân bố dữ liệu, nhưng có thể bỏ sót gian lận (điểm bất thường)

nếu dữ liệu biến động lớn hoặc gian lận xảy ra thường xuyên, làm ngưỡng thích nghi theo.

Với phân phối Poisson, ngưỡng tĩnh được tinh chỉnh và xác định thông qua thực nghiệm (code được đính kèm trên *github*). Kết quả cho thấy ngưỡng tối ưu là PMF = 0.001 và CDF = 0.05 nếu ưu tiên Recall. Ngược lại, nếu muốn cải thiện Precision (dù vẫn thấp), các bộ giá trị PMF = 0.005, CDF = 0.02 hoặc PMF = 0.005, CDF = 0.01 có hiệu suất tương đương nhưng giúp giảm số lượng false positives (FP) đôi chút.

Với phân phối Gaussian, ngưỡng thường được xác định theo công thức $\tau = \mu \pm k\sigma$, trong đó μ là giá trị trung bình, σ là độ lệch chuẩn, và k là hệ số điều chỉnh. Theo phân phối chuẩn $Z \sim \mathcal{N}(0,1)$ với $\mu = 0$, $\sigma = 1$, giá trị phổ biến là $k = 3$, dẫn đến ngưỡng z-threshold được xác định là ± 3.

Với phương pháp khoảng tứ phân vị (IQR), các giá trị $Q1$ và $Q3$ được sử dụng để xác định ngưỡng. Giá trị ngưỡng sẽ thay đổi theo từng tập dữ liệu khác nhau do sự thay đổi của $Q1$ và $Q3$. Ngưỡng dưới và ngưỡng trên được tính theo hai công thức sau:

$$\text{low\_threshold} = Q1 - 3 \times IQR \tag{8}$$

$$\text{high\_threshold} = Q3 + 3 \times IQR \tag{9}$$

2.2.5 Thực nghiệm và kết quả (Experimental Results)

Sau khi tính toán các phương pháp phân phối xác suất thống kê và tinh chỉnh ngưỡng để phát hiện bất thường, ta tiến hành thực nghiệm bằng cách triển khai code và thu được kết quả (một tập dataset mới, đã được lưu trữ trên *github*). Dataset mới được xây dựng từ dataset gốc, nhưng bổ sung các cột để biểu diễn quá trình triển khai thực tế. Code chi tiết được mô tả trong Listing 1 và cũng được lưu trữ trên *github*.

```
import pandas as pd
import numpy as np
from scipy.stats import poisson

df = pd.read_csv("creditcard.csv")
df["Minute"] = df["Time"] // 60
df["Log_Amount"] = np.log1p(df["Amount"])
df["Z_Score"] = (df["Log_Amount"] - df["Log_Amount"].mean())
    / df["Log_Amount"].std()

```

```
transaction_counts = df["Minute"].value_counts().sort_index()
lambda_poisson = transaction_counts.mean()

pmf_threshold = 0.001
cdf_threshold = 0.05

def detect_anomaly_pmf(k, lambda_poisson, pmf_threshold):
    return poisson.pmf(k, lambda_poisson) < pmf_threshold

def detect_anomaly_cdf(k, lambda_poisson, cdf_threshold):
    low_prob = poisson.cdf(k, lambda_poisson)
    high_prob = 1 - poisson.cdf(k, lambda_poisson)
    return (low_prob < cdf_threshold) or (high_prob <
    cdf_threshold)

pmf_anomalous_minutes = [minute for minute, count in
    transaction_counts.items() if detect_anomaly_pmf(count,
    lambda_poisson, pmf_threshold)]
cdf_anomalous_minutes = [minute for minute, count in
    transaction_counts.items() if detect_anomaly_cdf(count,
    lambda_poisson, cdf_threshold)]

df["Poisson_PMF_Anomaly"] = df["Minute"].isin(
    pmf_anomalous_minutes).astype(int)
df["Poisson_CDF_Anomaly"] = df["Minute"].isin(
    cdf_anomalous_minutes).astype(int)

q1 = df["Amount"].quantile(0.25)
q3 = df["Amount"].quantile(0.75)
iqr = q3 - q1
low_amount = q1 - 3 * iqr
high_amount = q3 + 3 * iqr
df["IQR_Amount_Anomaly"] = ((df["Amount"] < low_amount) | (df
    ["Amount"] > high_amount)).astype(int)

z_threshold = 3
df["Gaussian_Anomaly"] = ((df["Z_Score"] > z_threshold) | (df
    ["Z_Score"] < -z_threshold)).astype(int)

df.to_csv("creditcard_final_anomalies.csv", index=False)

print("File creditcard_final_anomalies.csv was created.")
```

Listing 1: Implementation code for anomaly detection

Dataset mới hình thành từ quá trình thực nghiệm sẽ được mô tả tại Figure 7. Tuy nhiên, việc đọc kết quả một cách trực tiếp với dữ liệu lớn là vô cùng khó khăn. Vậy nên, ở phần 2.2.6, ta sẽ thực hiện đánh giá với Dataset mới và đưa ra kết luận chi tiết.

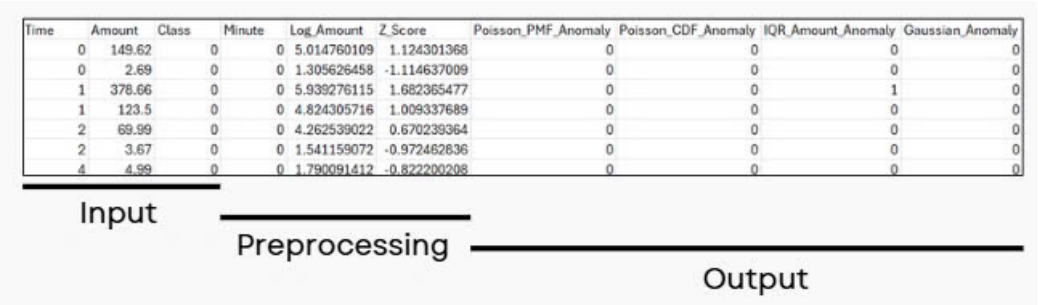

Figure 7: Dataset mới được tạo ra từ quá trình Implementation

2.2.6 Đánh giá

Đánh giá là bước quan trọng để xác định hiệu suất của mô hình xác suất thống kê trong dự án. Trong bài toán phát hiện bất thường, các chỉ số đánh giá phổ biến bao gồm Accuracy, Precision, Recall và F1-score, như đã đề cập ở phần 2.1.3. Dưới đây là kết quả tính toán các chỉ số này cho từng phương pháp phân phối xác suất thống kê (mã nguồn đánh giá được mô tả ở Listing 2 và đã được đính kèm trên *github*).

```
from sklearn.metrics import accuracy_score, precision_score,
    recall_score, f1_score
import pandas as pd

df = pd.read_csv("creditcard_final_anomalies.csv")

accuracy_pmf = accuracy_score(df['Class'], df['
    Poisson_PMF_Anomaly'])
precision_pmf = precision_score(df['Class'], df['
    Poisson_PMF_Anomaly'])
recall_pmf = recall_score(df['Class'], df['
    Poisson_PMF_Anomaly'])
f1_pmf = f1_score(df['Class'], df['Poisson_PMF_Anomaly'])

accuracy_cdf = accuracy_score(df['Class'], df['
    Poisson_CDF_Anomaly'])
```

```
precision_cdf = precision_score(df['Class'], df['
    Poisson_CDF_Anomaly'])
recall_cdf = recall_score(df['Class'], df['
    Poisson_CDF_Anomaly'])
f1_cdf = f1_score(df['Class'], df['Poisson_CDF_Anomaly'])

accuracy_iqr = accuracy_score(df['Class'], df['
    IQR_Amount_Anomaly'])
precision_iqr = precision_score(df['Class'], df['
    IQR_Amount_Anomaly'])
recall_iqr = recall_score(df['Class'], df['IQR_Amount_Anomaly
    '])
f1_iqr = f1_score(df['Class'], df['IQR_Amount_Anomaly'])

accuracy_gaussian = accuracy_score(df['Class'], df['
    Gaussian_Anomaly'])
precision_gaussian = precision_score(df['Class'], df['
    Gaussian_Anomaly'])
recall_gaussian = recall_score(df['Class'], df['
    Gaussian_Anomaly'])
f1_gaussian = f1_score(df['Class'], df['Gaussian_Anomaly'])

print("Poisson PMF:")
print(f"Accuracy: {accuracy_pmf:.4f}")
print(f"Precision: {precision_pmf:.4f}")
print(f"Recall: {recall_pmf:.4f}")
print(f"F1-score: {f1_pmf:.4f}")
print()

print("Poisson CDF:")
print(f"Accuracy: {accuracy_cdf:.4f}")
print(f"Precision: {precision_cdf:.4f}")
print(f"Recall: {recall_cdf:.4f}")
print(f"F1-score: {f1_cdf:.4f}")
print()

print("IQR Amount:")
print(f"Accuracy: {accuracy_iqr:.4f}")
print(f"Precision: {precision_iqr:.4f}")
print(f"Recall: {recall_iqr:.4f}")
print(f"F1-score: {f1_iqr:.4f}")
print()

print("Gaussian (Z-score):")
print(f"Accuracy: {accuracy_gaussian:.4f}")
```

```
print(f"Precision: {precision_gaussian:.4f}")
print(f"Recall: {recall_gaussian:.4f}")
print(f"F1-score: {f1_gaussian:.4f}")
```

Listing 2: Evaluation code for anomaly detection

Table 1: **Bảng so sánh các phương pháp xác suất thống kê**

Phương pháp	Accuracy	Precision	Recall	F1-score
Poisson PMF	0.2830	0.0020	0.8130	0.0039
Poisson CDF	0.1765	0.0018	0.8557	0.0036
IQR Amount	0.9322	0.0034	0.1321	0.0067
Gaussian (Z-score)	0.9975	0.0000	0.0000	0.0000

Table 1 cho thấy các phương pháp xác suất thống kê có nhiều hạn chế trong phát hiện bất thường. Poisson CDF đạt Recall cao nhất (0.8557), cho thấy khả năng phát hiện bất thường tốt, nhưng Precision rất thấp (0.0018), dẫn đến nhiều cảnh báo sai (False Positives). Poisson PMF có hiệu suất tương tự nhưng kém hơn CDF một chút. Phương pháp IQR Amount có Accuracy cao hơn Poisson (0.3922) nhưng Recall thấp (0.1321), tức là bỏ sót nhiều trường hợp bất thường. Precision của IQR nhỉnh hơn Poisson nhưng vẫn rất thấp (0.0034). Trong khi đó, Gaussian (Z-score) có Accuracy cực kỳ cao (0.9975) nhưng Precision, Recall và F1-score đều bằng 0, chứng tỏ phương pháp này không phát hiện được bất thường nào, có thể do dữ liệu không tuân theo phân phối chuẩn. Nhìn chung, không có phương pháp nào cân bằng tốt giữa Precision và Recall (những chỉ số đánh giá quan trọng về tác vụ phát hiện bất thường).

2.2.7 So sánh và cải tiến

Qua việc đánh giá ở phần 2.2.6, ta nhận thấy phương pháp xác suất thống kê trong phát hiện bất thường có một số hạn chế. Thứ nhất, phương pháp này thường giả định rằng dữ liệu tuân theo một phân phối xác suất cụ thể, chẳng hạn như phân phối chuẩn (Gaussian), điều này có thể không chính xác trong nhiều trường hợp, dẫn đến kết quả sai lệch. Thứ hai, khi dữ liệu có tính phụ thuộc cao, phương pháp này có thể không hoạt động hiệu quả, đặc biệt trong trường hợp dữ liệu có tính biến động mạnh. Thứ ba, các phương pháp này cũng gặp khó khăn khi đối diện với các dữ liệu bất thường chưa

biết hoặc không tuân theo các phân phối xác suất đã định sẵn. Cuối cùng, tính linh hoạt của các mô hình này khá hạn chế khi áp dụng vào các bài toán phát hiện bất thường trong những lĩnh vực có đặc điểm dữ liệu thay đổi hoặc không đồng đều theo thời gian.

Để nâng cao hiệu quả phát hiện bất thường so với các phương pháp xác suất thống kê (như Poisson, Z-score, IQR), các phương pháp học máy (Machine Learning) và học sâu (Deep Learning) là những lựa chọn mạnh mẽ, giúp cải thiện độ chính xác và khả năng phát hiện bất thường. Dưới đây là một số phương pháp tổng quan có thể mang lại kết quả tốt hơn so với các phương pháp truyền thống:

1. **Học máy** là lĩnh vực cho phép các hệ thống học hỏi từ dữ liệu mà không cần phải lập trình cụ thể cho từng nhiệm vụ. Các mô hình học máy có thể được sử dụng để phát hiện bất thường, và những phương pháp này có thể tự động phát hiện các mẫu và mối quan hệ phức tạp trong dữ liệu.

2. **Học sâu** là một nhánh của học máy sử dụng mạng nơ-ron nhân tạo với nhiều lớp (layers), giúp hệ thống học hỏi từ dữ liệu ở cấp độ cao hơn. Các mô hình học sâu có thể rất hiệu quả trong việc phát hiện bất thường trong dữ liệu phức tạp hoặc không có cấu trúc rõ ràng.

Để hiểu rõ hơn về sự tối ưu hiệu quả của các phương pháp học máy và học sâu trong tác vụ phát hiện bất thường so với phương pháp xác suất thống kê, ta cần thực hiện so sánh ưu nhược điểm giữa chúng với phương pháp xác suất thống kê. Sự so sánh sẽ được thể hiện ở bảng 2.

Table 2: **Ưu nhược điểm của hai phương pháp**

Phương pháp	**Ưu điểm**	**Nhược điểm**
Xác suất thống kê	- Đơn giản, dễ hiểu và dễ triển khai.	- Không linh hoạt với dữ liệu không tuân theo phân phối giả định.
	- Tính toán nhanh, không yêu cầu nhiều tài nguyên.	- Khó xử lý dữ liệu có cấu trúc phức tạp.
	- Hiệu quả với dữ liệu có phân phối có quy tắc (phân phối chuẩn, phân phối Poisson,...)	- Hiệu suất giảm khi số lượng dữ liệu lớn hoặc có nhiều nhiễu.
Học máy và Học sâu	- Không quan tâm đến phân phối dữ liệu.	- Cần nhiều dữ liệu huấn luyện để đạt hiệu suất cao.
	- Có thể phát hiện bất thường phức tạp, kể cả trong dữ liệu phi tuyến.	- Tính toán phức tạp
	- Tích hợp tốt với hệ thống dữ liệu lớn, có thể áp dụng GPU để tăng tốc xử lý.	- Khó diễn giải, đặc biệt với các mô hình học sâu.

2.2.8 Tổng kết và dự định tương lai (Conclusion and Future Works)

Báo cáo khảo sát các phương pháp xác suất thống kê trong phát hiện bất thường với mục tiêu xác định các điểm dữ liệu bất thường trong tập dữ liệu tài chính. Các phương pháp được sử dụng gồm phân phối Poisson, phân phối Gaussian (Z-score) và khoảng tứ phân vị (IQR). Kết quả cho thấy các phương pháp thống kê có độ chính xác thấp, đặc biệt Gaussian không phát hiện được bất thường (F1-score = 0), và Poisson có Recall cao nhưng Precision rất thấp. Nhược điểm chính là sự phụ thuộc vào giả định về phân phối dữ liệu, làm giảm hiệu suất nếu dữ liệu không tuân theo mô hình dự đoán. Báo cáo đề xuất sử dụng mô hình học máy và học sâu để cải thiện khả năng phát hiện bất thường nhờ tính linh hoạt cao hơn. Và trong tương lai, nghiên cứu có thể mở rộng theo các hướng sau:

1. Tích hợp học máy và học sâu bằng cách áp dụng các mô hình để cải thiện khả năng phát hiện bất thường so với các phương pháp thống kê truyền thống.

2. Kết hợp xác suất thống kê với học máy nhằm tận dụng ưu điểm của cả hai, giúp giảm tỷ lệ sai sót và tăng độ chính xác.

3. Thử nghiệm trên nhiều loại dữ liệu khác nhau để đánh giá tính tổng quát của các phương pháp.

4. Thực nghiệm với tập dữ liệu lớn hơn kết hợp với GPU để đánh giá cụ thể hơn về hiệu suất xử lý của mô hình.

References

[1] Credit Card Fraud Detection

[2] Powers, David MW. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." arXiv preprint arXiv:2010.16061 (2020).

[3] Consul, Prem C., and Gaurav C. Jain. "A generalization of the Poisson distribution." Technometrics 15.4 (1973): 791-799.

[4] Do, Chuong B. "The multivariate Gaussian distribution." Section Notes, Lecture on Machine Learning, CS 229 (2008).

[5] Vinutha, H. P., B. Poornima, and B. M. Sagar. "Detection of outliers using interquartile range technique from intrusion dataset." Information and decision sciences: Proceedings of the 6th international conference on ficta. Springer Singapore, 2018.