**ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỆ THỐNG THÔNG TIN**



**BÁO CÁO ĐỒ ÁN**

**HỆ HỖ TRỢ QUYẾT ĐỊNH**

**ĐỀ TÀI**

**BANKSEALER, A DECISION SUPPORT SYSTEM FOR ONLINE BANKING FRAUD ANALYSIS AND INVESTIGATION**

**BANKSEALER, HỆ HỖ TRỢ QUYẾT ĐỊNH ĐỂ PHÂN TÍCH VÀ ĐIỀU TRA GIAN LẬN NGÂN HÀNG TRỰC TUYẾN**

Giảng viên hướng dẫn:

TS. Dương Minh Đức

KS. Quan Chí Khánh An

Sinh viên thực hiện:

Nguyễn Thị Quỳnh Anh 18520004

Hoàng Ngọc Phương Dung 18520023

Lê Văn Thắng 18520355

Triệu Thị Thùy Dung 18520640

**TP. Hồ Chí Minh, tháng … năm ….**

**LỜI CẢM ƠN**

Để hoàn thành đề tài “Banksealer, hệ hỗ trợ để phân tích và điều tra gian lận ngân hàng trực tuyến”. Ngoài sự nỗ lực của các thành viên, nhóm còn nhận được sự giúp đỡ của nhiều tập thể và cá nhân. Đặc biệt xin gửi lời cảm ơn đến TS. Dương Minh Đức và KS Quan Chí Khánh An – giảng viên môn “Hệ hỗ trợ quyết định” đã tận tình giảng dạy, trang bị cho nhóm những kiến thức quý báu trong kì học vừa qua để nhóm có thể hoàn thành tốt đồ án. Trong thời gian học tập, nhóm chúng em không chỉ nhận được những kiến thức bổ ích về chuyên môn mà còn ở những lĩnh vực khác. Những trải nghiệm quý báu đó không chỉ giúp em hoàn thành tốt bài báo cáo mà còn là hành trang theo em trong suốt thời gian học tập và làm việc sau này.

Tuy nhiên trong quá trình nghiên cứu đề tài, do kiến thức chuyên ngành còn hạn

chế nên nhóm chúng em vẫn còn nhiều thiếu sót khi tìm hiểu, đánh giá và trình bày về

đề tài. Rất mong nhận được sự quan tâm, đóng góp của thầy để đồ án của chúng em

được đầy đủ và hoàn chỉnh hơn.

Cuối cùng xin kính chúc quý thầy dồi dào sức khỏe và thành công trong sự nghiệp

cao quý.

Chúng em xin chân thành cảm ơn.

Hồ Chí Minh, ngày.....tháng 6 năm 2021

Nhóm sinh viên

**NHẬN XÉT CỦA GIẢNG VIÊN**

**................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................**

## MỤC LỤC

[MỤC LỤC 4](#_Toc75446864)

[PHẦN 1: GIỚI THIỆU 6](#_Toc75446865)

[1.1. Mục tiêu: 6](#_Toc75446866)

[1.2 Phạm vi: 6](#_Toc75446867)

[1.2. Đối tượng nghiên cứu: 6](#_Toc75446868)

[1.3. Kết quả nghiên cứu: 6](#_Toc75446869)

[1.4. Mô tả tổng quát bài báo: 6](#_Toc75446870)

[1.4.1. Input: 6](#_Toc75446871)

[1.4.2. Process: 7](#_Toc75446872)

[1.4.3. Output: 7](#_Toc75446873)

[PHẦN 2: CÁC NGHIÊN CỨU VÀ HƯỚNG TIẾP CẬN LIÊN QUAN 8](#_Toc75446874)

[PHẦN 3: MÔ HÌNH/ THUẬT TOÁN ĐỀ XUẤT CẢI TIẾN 13](#_Toc75446875)

[3.1. Vẽ mô hình: 13](#_Toc75446876)

[3.2. Giải thích các thành phần: 13](#_Toc75446877)

[3.3. Đóng góp mới của tác giả: 14](#_Toc75446878)

[PHẦN 4: BỘ DỮ LIỆU THỰC NGHIỆM 15](#_Toc75446879)

[4.1. Mô tả bộ dữ liệu: 15](#_Toc75446880)

[4.1.1. Nguồn bộ dữ liệu: 15](#_Toc75446881)

[4.1.2. Mô tả cấu trúc các thành phần dữ liệu và các thuộc tính: 15](#_Toc75446882)

[4.2. Đánh giá bộ dữ liệu (số lượng mẫu, tính ngẫu nhiên, độ tin cậy) 20](#_Toc75446883)

[PHẦN 5: KẾT QUẢ THỰC NGHIỆM VÀ ĐÁNH GIÁ 21](#_Toc75446884)

[5.1. Cài đặt thực nghiệm: 21](#_Toc75446885)

[5.1.1. Giới thiệu thuật toán: 21](#_Toc75446886)

[5.1.2. Cài đặt mô hình: 23](#_Toc75446887)

[5.2. Đánh giá: 51](#_Toc75446888)

[5.3. Ưu điểm và hạn chế: 52](#_Toc75446889)

[5.3.1. Ưu điểm: 52](#_Toc75446890)

[5.3.2. Hạn chế: 52](#_Toc75446891)

[5.4. Độ tin cậy, chính xác, hiệu suất: 52](#_Toc75446892)

[PHẦN 6: ĐỀ XUẤT HƯỚNG NGHIÊN CỨU 53](#_Toc75446893)

[PHẦN 7: KẾ HOẠCH THỰC HIỆN 54](#_Toc75446894)

[PHẦN 8: TÀI LIỆU THAM KHẢO 55](#_Toc75446895)

## PHẦN 1: GIỚI THIỆU

## Mục tiêu:

Tạo ra một mô hình chung cho kỹ thuật phát hiện bất thường nửa giám sát online dựa trên sự kết hợp giữa các mô hình khác nhau để tìm ra các loại gian lận khác nhau

## 1.2 Phạm vi:

Dữ liệu được tạo từ mẫu dữ liệu giao dịch tổng hợp do một ngân hàng ở Tây Ban Nha cung cấp. Dữ liệu này chứa hàng nghìn bản ghi dữ liệu giao dịch trong sáu tháng, từ tháng 11 năm 2012 đến tháng 4 năm 2013 bị hạn chế bởi mã zip ở Madrid và Barcelona.

## Đối tượng nghiên cứu:

Dữ liệu chứa các giao dịch của khách hàng liên quan đến việc chuyển khoản (Bank transfer) (ví dụ, chuyển tiền từ bất kỳ tài khoản ngân hàng nào đến tài khoản ngân hàng khác).

## Kết quả nghiên cứu:

* Cho biết số điểm bất thường của các giao dịch.
* Hiệu suất và tài nguyên:
  + Thời gian yêu cầu tối đa sẽ là 4 phút mỗi ngày cho ngữ cảnh chuyển khoản ngân hàng BANKSEALER phù hợp với giám sát gian lận trực tuyến.
  + Mức tiêu thụ RAM cao nhất tăng gần như tuyến tính với số ngày và bậc hai với số người dùng.

## Mô tả tổng quát bài báo:

Diagram

Description automatically generated

### Input:

Dữ liệu giao dịch của khách hàng liên quan đến các giao dịch tại ngân hàng.

### Process:

* Làm sạch dữ liệu.
* Sử dụng một profile cục bộ, toàn bộ và theo thời gian, được xây dựng trong suốt training phase.
* Local profiling nhằm mô tả đặc điểm thói quen chi tiêu của mỗi người dùng. Trong suốt training, tổng hợp các giao dịch của người dùng và ước tính từng phân phối đặc điểm bằng biểu đồ. Trong thời gian chạy (runtime), nhóm tác giả sử dụng phương pháp HBOS ( Goldstein and Dengel, 2012 ) để tính toán điểm bất thường của mỗi giao dịch mới.
* Global profiling nhằm mô tả "các lớp" của thói quen chi tiêu. Trong quá trình training, đầu tiên nhóm tác giả tạo ra một profile toàn bộ cho mỗi người dùng và sau đó phân cụm các profile kết quả bằng cách sử dụng phiên bản lặp của DBSCAN (iterative version of the DBSCAN). Cuối cùng, đối với mỗi profile toàn bộ, nhóm tác giả tính toán điểm CBLOF – điểm này cho nhà phân tích biết profile ở mức độ bất thường nào đối với cụm gần nhất của nó. Profile toàn bộ cũng được tận dụng để tìm profile cục bộ cho người dùng mới hoặc under-trained. Cơ sở lý luận là những người dùng thuộc cùng một cụm thể hiện thói quen chi tiêu có profile cục bộ tương tự.
* Temporal profiling để đối phó với những hành vi gian lận - khai thác sự lặp lại của các giao dịch trông có vẻ hợp lý theo thời gian. Nhóm tác giả sử dụng khoảng thời gian (time window) – có thể dễ dàng chọn kích thước tài nguyên phần cứng có sẵn. Trong phạm vi khoảng thời gian, trong training, nhóm tác giả tổng hợp các giao dịch của mỗi người dùng theo thời gian với tần suất lấy mẫu hàng ngày và tính toán giá trị trung bình các mẫu và phương sai của các đặc điểm số. Chúng được sử dụng làm ngưỡng trong runtime để tính điểm bất thường.
* Cập nhật profile và điểm số bằng cách sử dụng hàm số mũ hệ số đếm, được biểu thị bằng khoảng thời gian W và tần suất lấy mẫu tương ứng.

### Output:

Điểm bất thường trong mỗi giao dịch.

# PHẦN 2: CÁC NGHIÊN CỨU VÀ HƯỚNG TIẾP CẬN LIÊN QUAN

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nhóm tác giả** | **Mục tiêu** | **Phương pháp** | **Kết quả** | **Khả năng kế thừa và áp dụng** |
| **BANKSEALER:** Michele Carminati, Roberto Caron, Federico Maggi, Ilenia Epifani, Stefano Zanero | Hỗ trợ việc phân tích của các gian lận (mới) và những bất thường | Phương pháp định hình 3 lần (threefold profiling), không giám sát và bán giám sát | Độ chính xác: 74% các vụ gian lận đối với tập dữ liệu chuyển khoản ngân hàng, 100% cho bộ dữ liệu nạp điện thoại và 93% cho tập dữ liệu thẻ trả trước đối với người dùng well-trained và thậm chí cao hơn đối với người dùng under-trained. | Bài báo |
| Hiệu suất và tài nguyên:  -Thời gian yêu cầu tối đa sẽ là 4 phút mỗi ngày cho ngữ cảnh chuyển khoản ngân hàng BANKSEALER phù hợp với giám sát gian lận trực tuyến.  -Mức tiêu thụ RAM cao nhất tăng gần như tuyến tính với số ngày và bậc hai với số người dùng. |
| **Chandola et al (2009) Anomaly Detection : A Survey:** VARUN CHANDOLA(University of Minnesota)  ARINDAM BANERJEE  (University of Minnesota)  VIPIN KUMAR  (University of Minnesota)  **Phua et al. A Comprehensive Survey of Data Mining-based Fraud Detection Research:**  CLIFTON PHUA  VINCENT LEE KATE SMITH  ROSS GAYLER  **Bolton & David:**  R. J. Bolton, David, **Statistical fraud detection: a review**, Stat Sci 17. | Phát hiện lừa đảo, tập trung chủ yếu vào thẻ tín dụng | Khảo sát và các phương pháp khai thác dữ liệu | Phát hiện lừa đảo nhưng không đưa ra được lý do kết luận cho nhà phân tích | Không đề cập |
| **Bay and Pazzani, 2001.**  Bay SD, Pazzani MJ. **Detecting group differences: mining contrast**  **sets**. Data Min Knowl Discov 2001;5(3):213e46 | Phát hiện lừa đảo ngân hàng | Các phương pháp có giám sát dựa trên những dạng tương phản và bộ tương phản | Yêu cầu những mẫu thử được dán nhãn | Nhược điểm chính của phương pháp có giám sát: Yêu cầu những mẫu thử được dán nhãn |
| **Aggelis (2006) Aggelis V. Offline internet banking fraud detection**. In: ARES, IEEE  Computer Society; 2006. p. 904e5. | Phát hiện lừa đảo ngân hàng | Hệ thống phát hiện lừa đảo Internet banking dựa trên quy tắc (rule-based), có giám sát | Không hoạt động trong thời gian thực (real-time), yêu cầu những mẫu thử được dán nhãn | Nhược điểm chính của phương pháp có giám sát: Yêu cầu những mẫu thử được dán nhãn |
| **Wei et al. (2013) Effective detection of sophisticated online banking fraud on extremely imbalanced data:**  Wei Wei, Jinjiu Li, Longbing Cao, Yuming Ou, Jiahang Chen. | Phát hiện lừa đảo ngân hàng | Phương pháp không giám sát, xem xét sự phụ thuộc giữa các sự kiện tại các điểm khác nhau trong cùng lúc nhằm giảm thiểu những thiếu sót của các mẫu khai thác tương phản | Giải quyết những nhật ký của ứng dụng web online banking, và vì thế không phát hiện các vụ lừa đảo nhiều như việc tương tác bất quy tắc với ứng dụng, không đưa ra được lý do kết luận cho nhà phân tích | Nhược điểm chính của phương pháp không giám sát và bán giám sát hiện nay: Không đưa ra lý do kết luận cho nhà phân tích, điều này khiến việc điều tra thủ công và xác nhận trở nên khó khăn hơn. |
| **Mhamane and Lobo (2012) Internet banking fraud detection using HMM:**  Sunil S Mhamane  (Walchand of Institute of Technology, Solapur, India)  L.M.R.J Lobo  (DepartmentOf Computer Science and Engineering, Walchand of Institute of Technology, Solapur, India) | Cơ chế phát hiện hiệu quả để nhận dạng người dùng hợp pháp và truy vết các hoạt động bất hợp pháp của họ | Không giám sát, dùng Hidden Markov Model HMMs | Không đưa ra được lý do kết luận cho nhà phân tích | Nhược điểm chính của phương pháp không giám sát và bán giám sát hiện nay: Không đưa ra lý do kết luận cho nhà phân tích, điều này khiến việc điều tra thủ công và xác nhận trở nên khó khăn hơn. |
| **Kovach and Ruggierio (2011) Kovach S, Ruggiero W. Online banking fraud detection based on**  **local and global behavior.** In: ICDS 2011: the Fifth Intl. Conf. on  digital society; 2011. p. 166e71. | Phát hiện lừa đảo ngân hàng | Mô hình không giám sát về các quan sát cục bộ và toàn bộ hành vi của người dùng, và tin vào phân tích khác biệt để phát hiện các vụ lừa đảo như những sai lệch so với hành vi bình thường. Bằng chứng này được củng cố hay bị suy giảm bởi hành vi của người dùng toàn cầu. | Hạn chế chủ yếu của phương pháp này là sự thu thập dữ liệu phải xảy ra ở phía người dùng (client side), điều này làm cho nó trở nên bế tắc khi triển khai trong bối cảnh lớn và thực tế, không đưa ra được lý do kết luận cho nhà phân tích | Nhược điểm chính của phương pháp không giám sát và bán giám sát hiện nay: Không đưa ra lý do kết luận cho nhà phân tích, điều này khiến việc điều tra thủ công và xác nhận trở nên khó khăn hơn. |

# PHẦN 3: MÔ HÌNH/ THUẬT TOÁN ĐỀ XUẤT CẢI TIẾN

## 3.1. Vẽ mô hình:

Diagram

Description automatically generated

## 3.2. Giải thích các thành phần:

* Sử dụng một profile cục bộ, toàn bộ và theo thời gian, được xây dựng trong suốt training phase.
* **Local profiling** nhằm mô tả đặc điểm thói quen chi tiêu của mỗi người dùng.
* Trong suốt training, nhóm tác giả tổng hợp các giao dịch của người dùng và ước tính từng phân phối đặc điểm bằng biểu đồ. Chính xác hơn, nhóm tác giả tính toán phân phối biên thực nghiệm của đặc điểm giao dịch của người dùng.
* Trong thời gian chạy (runtime), nhóm tác giả sử dụng phương pháp HBOS ( Goldstein and Dengel, 2012 ) để tính toán điểm bất thường của mỗi giao dịch mới. HBOS tính toán xác suất của một giao dịch theo phân phối biên đã học. Nhóm tác giả cải thiện HBOS để giải thích sự khác biệt của từng đặc điểm và cho phép nhà phân tích cân nhắc các đặc điểm khác nhau theo ưu tiên của tổ chức.
* **Global profiling** nhằm mô tả "các lớp" của thói quen chi tiêu.
* Trong quá trình training, đầu tiên nhóm tác giả tạo ra một profile toàn bộ cho mỗi người dùng và sau đó phân cụm các profile kết quả bằng cách sử dụng phiên bản lặp của DBSCAN (iterative version of the DBSCAN).
* Cuối cùng, đối với mỗi profile toàn bộ, nhóm tác giả tính toán điểm CBLOF – điểm này cho nhà phân tích biết profile ở mức độ bất thường nào đối với cụm gần nhất của nó. Profile toàn bộ cũng được tận dụng để tìm profile cục bộ cho người dùng mới hoặc under-trained. Cơ sở lý luận là những người dùng thuộc cùng một cụm thể hiện thói quen chi tiêu có profile cục bộ tương tự.
* **Temporal profiling** để đối phó với những hành vi gian lận - khai thác sự lặp lại của các giao dịch trông có vẻ hợp lý theo thời gian. Nhóm tác giả xây dựng một profile theo thời gian cho mỗi người dùng có đủ số lượng giao dịch trong quá khứ, bởi vì các giao dịch không thường xuyên thì có phương sai cao, không thích hợp cho loại phân tích này. Nhóm tác giả sử dụng khoảng thời gian (time window) – có thể dễ dàng chọn kích thước tài nguyên phần cứng có sẵn.
* Trong quá trình training, đối với mỗi người dùng, nhóm tác giả trích xuất các đặc điểm tổng hợp sau: tổng số tiền, tổng số giao dịch và số lượng giao dịch hàng ngày tối đa. Trong quá trình training, nhóm tác giả tính giá trị trung bình và độ lệch chuẩn cho từng đặc điểm,và đặt ngưỡng ở giá trị trung bình cộng với độ lệch chuẩn.
* Trong runtime, cho mỗi người dùng và theo tần suất lấy mẫu, nhóm tác giả tính toán giá trị tích lũy cho mỗi đặc điểm đã nói ở trên. Sau đó, nhóm tác giả tính tổng delta dương giữa mỗi giá trị tích lũy và ngưỡng tương ứng để hình thành điểm bất thường.

## 3.3. Đóng góp mới của tác giả:

* Phân tích chuyên sâu về tệp dữ liệu online banking đời thực, trong đó tác giả nêu bật lên những thách thức và ý nghĩa của việc giải quyết tình trạng thiếu thốn tệp dữ liệu trong phạm vi nghiên cứu này.
* Một framework chung cho kỹ thuật phát hiện bất thường nửa giám sát online dựa trên sự kết hợp giữa các mô hình khác nhau để tìm ra các loại gian lận khác nhau. Phương pháp này ghi điểm với ý nghĩa thống kê rõ ràng, là có khả năng thích nghi với các nguồn tĩnh và có thể giải quyết trượt khái niệm (concept drift) và sự thiếu thốn dữ liệu.
* Một đánh giá gần như toàn diện xuyên suốt một chuỗi các cuộc tấn công thực tế và trong bối cảnh thực tế, nhờ vào sự triển khai cho một ngân hàng thương mại lớn.

# PHẦN 4: BỘ DỮ LIỆU THỰC NGHIỆM

## 4.1. Mô tả bộ dữ liệu:

### 4.1.1. Nguồn bộ dữ liệu:

* Dữ liệu được lấy từ Kaggle.com, là một cộng đồng trực tuyến về khoa học dữ liệu và Merching learning nổi tiếng. Bộ dữ liệu được tạo ra từ Banksim.
* Banksim: Là một trình mô phỏng các khoản thanh toán qua ngân hàng dựa trên một mẫu dữ liệu có sẵn. Mục đích chính của BankSim là tạo ra dữ liệu tổng hợp có thể được sử dụng để nghiên cứu phát hiện gian lận. Mục tiêu lớn nhất là BankSim có thể sử dụng được để lập mô hình các tình huống liên quan đến kết hợp thanh toán thông thường và các chữ ký gian lận đã biết trong nội bộ. Các tập dữ liệu do BankSim tạo ra không chứa thông tin cá nhân hoặc không đảm bảo về các giao dịch khách hàng cá nhân và hợp pháp
* Dữ liệu được tạo từ mẫu dữ liệu giao dịch tổng hợp do một ngân hàng ở Tây Ban Nha cung cấp. Dữ liệu này chứa hàng nghìn bản ghi dữ liệu giao dịch trong sáu tháng, từ tháng 11 năm 2012 đến tháng 4 năm 2013 bị hạn chế bởi mã zip ở Madrid và Barcelona. Tức là, dữ liệu này đủ để phản ánh các điều kiện thanh toán hiện tại, nhưng được tổng hợp để không gây rủi ro về bảo mật của khách hàng.

### 4.1.2. Mô tả cấu trúc các thành phần dữ liệu và các thuộc tính:

* Tệp dữ liệu có 9 cột thuộc tính và 1 cột thuộc tính quyết định.
* Các thuộc tính gồm có:
* Step: ngày kể từ khi bắt đầu mô phỏng dữ liệu. Có 180 steps nên dữ liệu mô phỏng chạy trong 6 tháng.
  + Customer: id của khách hàng.
  + zipCodeOrigin: nguồn của mã bưu chính.
  + Merchant: id của người bán.
  + zipMerchant: mã bưu chính của người bán.
  + Age: Phân loại tuổi. Có thể có sự liên quan giữa độ tuổi và các hành vi gian lận. Ví dụ: kẻ lừa đảo có xu hướng thực hiện hành vi này với những khách hàng dưới 18.
* Số lượng các giá trị của thuộc tính Age trong tệp dữ liệu:

Ảnh có chứa bàn

Mô tả được tạo tự động

Trong đó: 0: <= 18

1: 19 – 25

2: 26 – 35. Có số lượng nhiều nhất trong tệp dữ liệu.

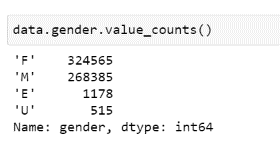
3: 36 – 45

4: 46 – 55

5: 56 – 65

6: > 65

* U: Không xác định được độ tuổi. Có số lượng ít nhất trong tệp dữ liệu.
* Gender: Giới tính của khách hàng
* Số lượng các giá trị của thuộc tính Gender trong tệp dữ liệu:



Trong đó: E: Doanh nghiệp

F: Nữ. Có số lượng nhiều nhất trong tệp dữ liệu.

M: Nam

U: Không xác định được giới tính.

* Category: Phân loại việc mua hàng, giao dịch vào việc gì. Tương tự với độ tuổi, phân loại việc mua hàng có thể xem xét mối liên hệ trong hành vi lừa đảo với hành động mua sắm của khách hàng. Ví dụ: kẻ lừa đảo thường nhắm tới những tài khoản chi trả cho việc giải trí,…
* Số lượng các giá trị của thuộc tính Category trong tệp dữ liệu:

Ảnh có chứa bàn

Mô tả được tạo tự động

Trong đó: es\_barsandrestaurants: bar và nhà hàng

es\_contents: nội dung

es\_fashion: thời trang

es\_food: đồ ăn

es\_health: sức khỏe

es\_home: nhà ở

es\_hotelservices: dịch vụ khách sạn

es\_hyper: các giao dịch khác

es\_leisure: giải trí. Có số lượng ít nhất trong tệp dữ liệu.

es\_otherservices: các dịch vụ khác

es\_sportsandtoys: thể thao và đồ chơi

es\_tech: công nghệ

es\_transportation: phương tiện. Có số lượng nhiều nhất trong tệp dữ liệu.

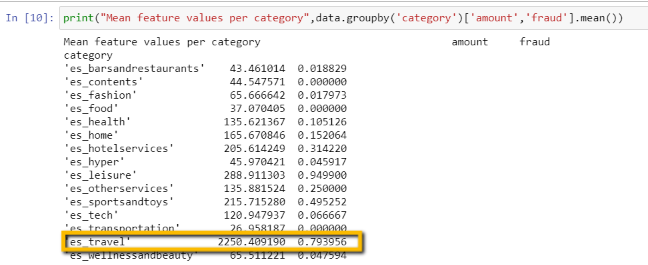
es\_travel: du lịch

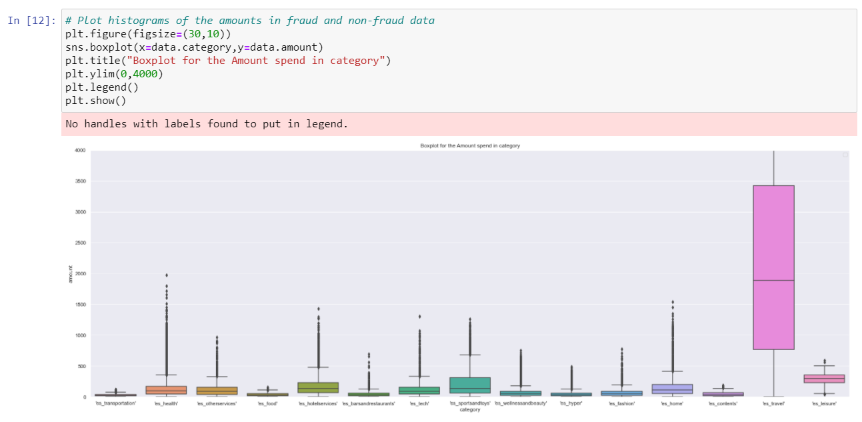
es\_wellnessandbeauty: làm đẹp

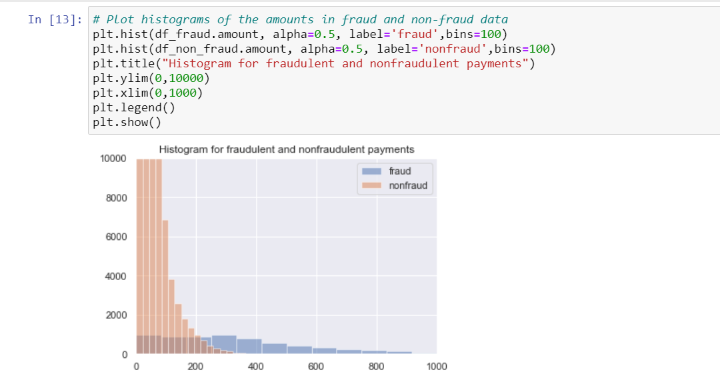
* Amount: Số tiền giao dịch.
* Fraud: Thuộc tính quyết định có gian lận hay không. Có là 1, Không là 0.
* Tệp dữ liệu này rất mất cân bằng:

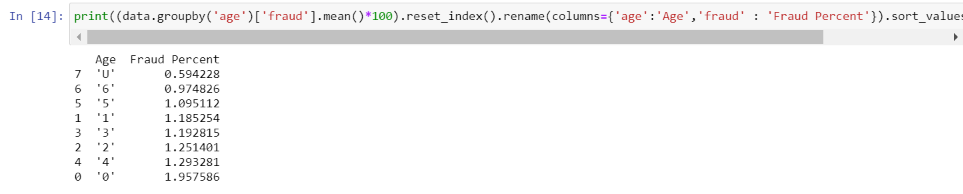


* Các đối tượng lừa đảo thường nhắm vào danh mục giải trí và du lịch (vì có mức chi tiêu cao), độ tuổi <=18, với mức lừa đảo chiếm đa số khoảng 500$:









Trong đó: Nhóm 0 là 0-18 tuổi, nhóm 6 là >= 65 tuổi.

## 4.2. Đánh giá bộ dữ liệu (số lượng mẫu, tính ngẫu nhiên, độ tin cậy)

* Bộ dữ liệu gồm 594643 dòng, đảm bảo tính đa dạng và khách quan khi tạo mô hình.
* Dữ liệu ở các cột đều không bị thiếu.

Ảnh có chứa văn bản

Mô tả được tạo tự động

* Là dữ liệu mô phỏng ngẫu nhiên bằng BankSim dựa trên mẫu dữ liệu thật của một ngân hàng Tây Ban Nha cung cấp, nên bộ dữ liệu không phải là dữ liệu thật, các giá trị không giống dữ liệu gốc (BankSim tạo ra dữ liệu gần đúng một cách hữu ích với các khía cạnh có liên quan tới dữ liệu thực).
* Bộ dữ liệu mất cân bằng. Trong 594643 dòng, có 587443 dòng là các khoản thanh toán thông thường và 7200 dòng là giao dịch gian lận. Đây là vấn đề thường gặp trong các bộ dữ liệu dự đoán gian lận, lừa đảo.
* Bộ dữ liệu không chứa các thuộc tính tương tự với bài báo đề tài. Vì lý do bảo mật nên tác giả không thể đưa dữ liệu cho nhóm. Tuy nhiên, theo bài báo thì bộ dữ liệu được ẩn danh toàn bộ trừ các thuộc tính như Amount, IP, IBAN. Bộ dữ liệu nhóm tìm được chỉ có Amount và Fraud là thuộc tính tương đương.

# PHẦN 5: KẾT QUẢ THỰC NGHIỆM VÀ ĐÁNH GIÁ

## 5.1. Cài đặt thực nghiệm:

### 5.1.1. Giới thiệu thuật toán:

#### 5.1.1.1. PCA (Principal component analysis):

* PCA là một kỹ thuật giảm kích thước tuyến tính có thể được sử dụng để trích xuất thông tin từ không gian chiều cao bằng cách chiếu nó vào không gian con có chiều cao thấp hơn. Nó cố gắng duy trì các thành phần thiết yếu có nhiều biến thể hơn của dữ liệu và loại bỏ các phần không thiết yếu với ít biến thể hơn.
* Kích thước không là gì ngoài các tính năng đại diện cho dữ liệu. Ví dụ: một hình ảnh 28\*28 có 784 yếu tố hình ảnh (pixel) là các kích thước hoặc đặc điểm cùng đại diện cho hình ảnh đó.
* Lưu ý đây là một kỹ thuật giảm kích thước không giám sát, chúng ta có thể phân cụm các điểm dữ liệu tương tự dựa trên mối tương quan về tính năng giữa chúng mà không cần bất kỳ sự giám sát (hoặc nhãn) nào.
* Áp dụng PCA để:
  + Trực quan hóa dữ liệu.
  + Tăng tốc thuật toán Machine Learning (ML).

#### 5.1.1.2. Thuật toán HBOS:

* HBOS là một thuật toán dựa trên khoảng cách không được giám sát để nắm bắt các điểm ngoại lệ. Một lựa chọn phù hợp để điều trị ngoại lai toàn cầu nhưng không phải là một lựa chọn ấn tượng cho những dị thường cục bộ. Khi một điểm dữ liệu được đặt xa hơn phần còn lại của dữ liệu, thì nó được gọi là ngoại lệ toàn cầu.
* Trong khi, điểm kỳ lạ cục bộ là một điểm dữ liệu nằm trong giới hạn của toàn bộ dữ liệu, nhưng được đặt ở một khoảng cách hợp lý so với vùng lân cận cục bộ của nó. Thuật toán giả định các tính năng hoàn toàn độc lập (không đa cộng tuyến) với nhau.
* Ưu điểm của thuật toán:
  + Tương đối nhanh hơn khi so sánh với cách tiếp cận dựa trên cụm và dựa trên hàng gần nhất.
  + HBOS hoạt động trong thời gian tuyến tính O(n) trong trường hợp chiều rộng thùng cố định hoặc trong O(n\*log(n)) sử dụng chiều rộng thùng.

#### 5.1.1.3. Khoảng cách Mahanobis:

* Khoảng cách Mahalanobis cung cấp một cách để đo mức độ tương tự của một số điều kiện đối với một tập hợp các điều kiện đã biết. Nó tính hiệp phương sai giữa các biến. Nói cách khác, khoảng cách Mahalanobis được sử dụng để tìm các ngoại lệ trong một tập dữ liệu.
* Nó được tính như sau: D2 = ( x – m )T C-1 ( x – m )

Trong đó: D2 = khoảng cách Mahalanobis

x = Vector dữ liệu

m = Vector giá trị trung bình của các biến độc lập

C-1 = Ma trận phương sai nghịch đảo của các biến độc lập

T = Vector hoán vị

#### 5.1.1.4. Thuật toán DBSCAN (Density\_Based Spatial Clustering of Application with Noise):

* DBSCAN là một thuật toán phân cụm dựa trên mật độ. Các cụm được hình thành dựa trên các thông số về mật độ.
* Thuật toán DBSCAN nhận 2 tham số; ε —epilon, là bán kính của các điểm cốt lõi và số lượng điểm dữ liệu tối thiểu trong cụm.
* Các thuật toán DBSCAN có thể được trừu tượng hóa thành các bước chủ yếu sau:
* Tìm số điểm trong phạm vi bao vây của một điểm và xác định các điểm cốt lõi.
* Tìm các thành phần được kết nối của các điểm cốt lõi trong vùng lân cận.
* Gán mỗi điểm ranh giới cho một cụm điểm cốt lõi gần đó, nếu điểm đó nằm trong vùng lân cận còn không thì gán cho điểm nhiễu.

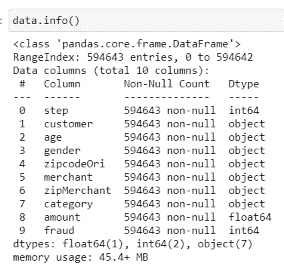
#### 5.1.1.5. Thuật toán CBOLF (Cluster-based Local Outlier Factor)

* CBLOF tính toán điểm ngoại lệ dựa trên hệ số ngoại lệ cục bộ dựa trên cụm. Điểm bất thường được tính bằng khoảng cách của mỗi cá thể đến trung tâm cụm của nó nhân với các cá thể thuộc cụm của nó (thư viện pyOD bao gồm việc triển khai CBLOF).

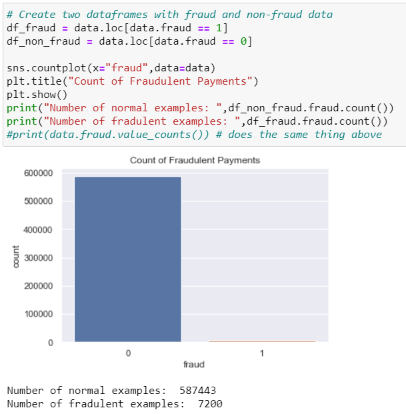
### 5.1.2. Cài đặt mô hình:

#### 5.1.2.1. Phân tích và trực quan hóa dữ liệu

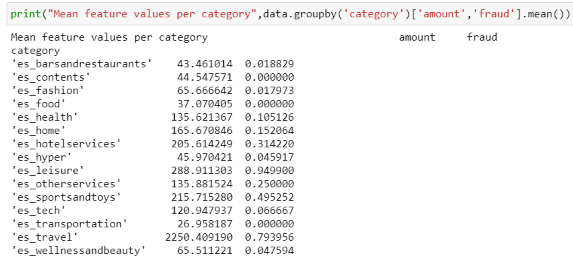
* Bộ dữ liệu gồm 594643 dòng và 10 cột, không chứa missing value:

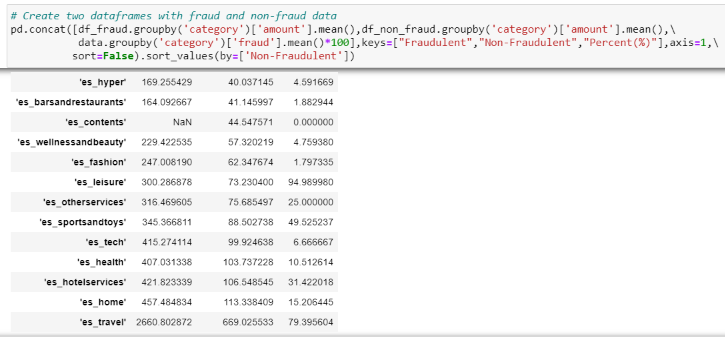


* Có thể thấy dữ liệu rất mất cân bằng với 587443 giao dịch là không lừa đảo và chỉ 7200 giao dịch là lừa đảo. Đây cũng là đặc trưng thường thấy của các bộ dữ liệu dự đoán gian lận.

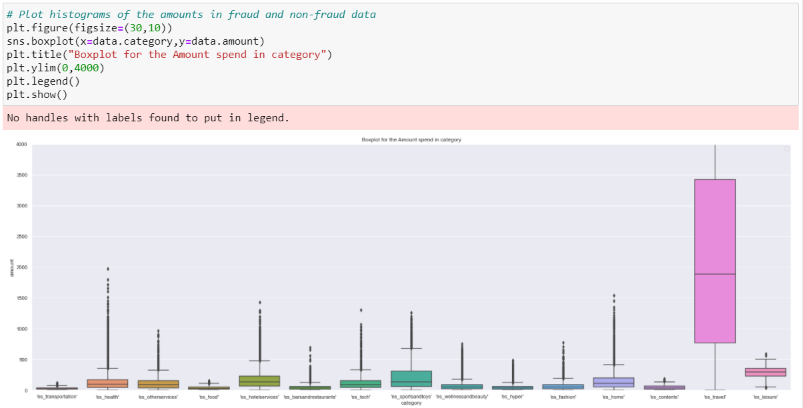


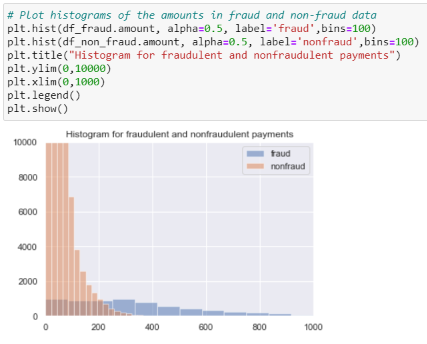
* Xem số tiền trung bình và phần tram gian lận è Có thể thấy giải trí và du lịch được kẻ lừa đảo chọn lựa nhiều hơn vì có chi tiêu cao hơn:



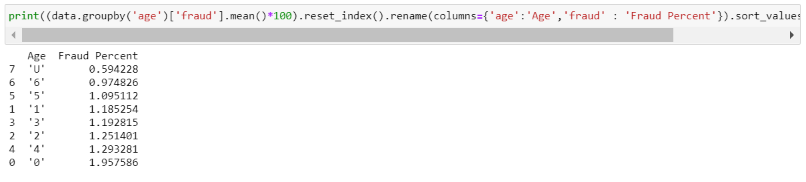


* Số tiền gian lận xảy ra ở mức tương đương nhau (khoảng 0-500) cho các loại danh mục, trừ danh mục du lịch tăng rất cao:



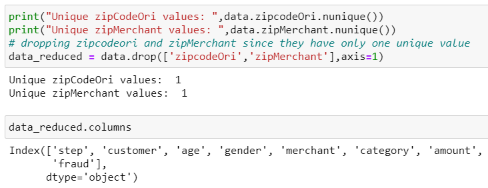


* Các giao dịch gian lận cũng thường xảy ra ở nhóm tuổi 0 (<=18):

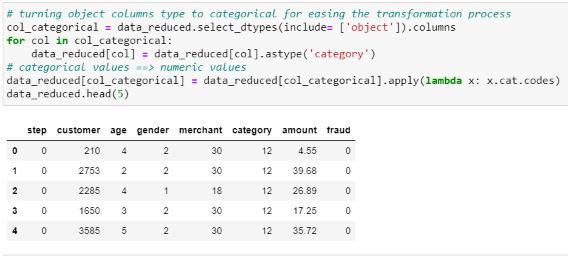


#### 5.1.2.2. Tiền xử lý dữ liệu

* Chỉ có 1 giá trị zipCodeOri và zipMerchant duy nhất nên nhóm sẽ tiến hành xóa bỏ. Lúc này data\_reduced có 8 cột:



* Tiến hành số hóa dữ liệu để dễ dàng chạy các thuật toán:

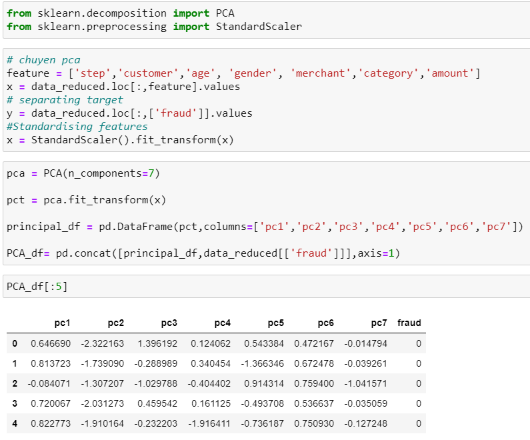


* Chia train và test với tỷ lệ 80:20:



#### 5.1.2.3. Local Profile:

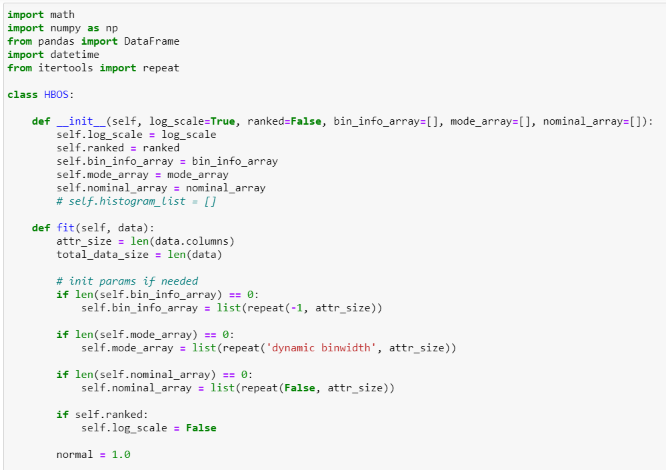
* Sử dụng PCA để giảm chiều dữ liệu và tăng tốc chạy thuật toán:



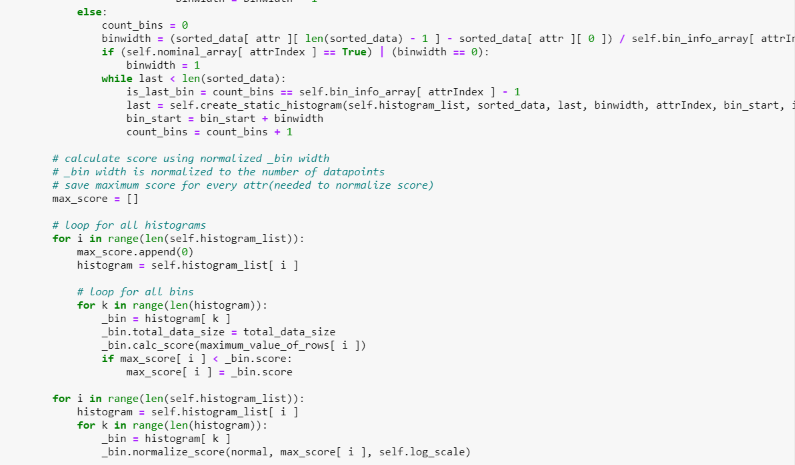
* Xóa thuộc tính “fraud” để chạy thuật toán HBOS:



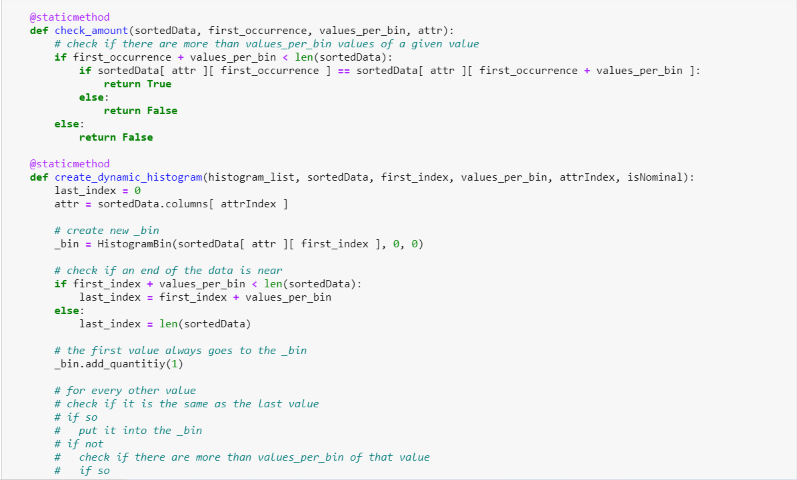
* Cài đặt thuật toán HBOS:

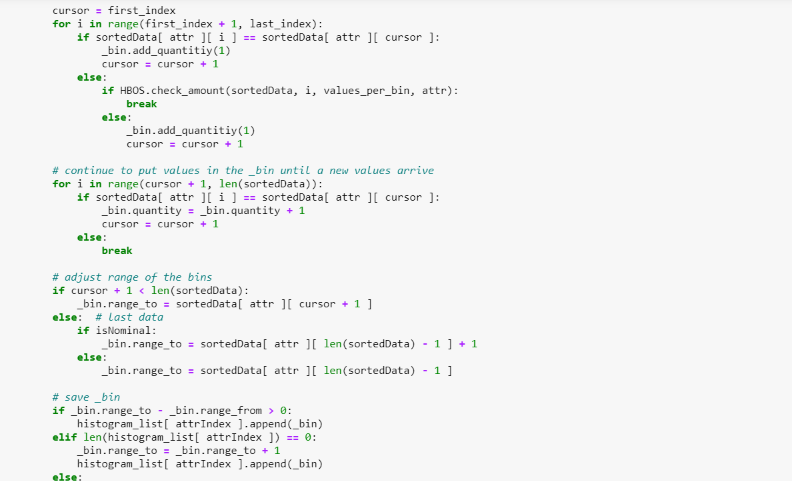


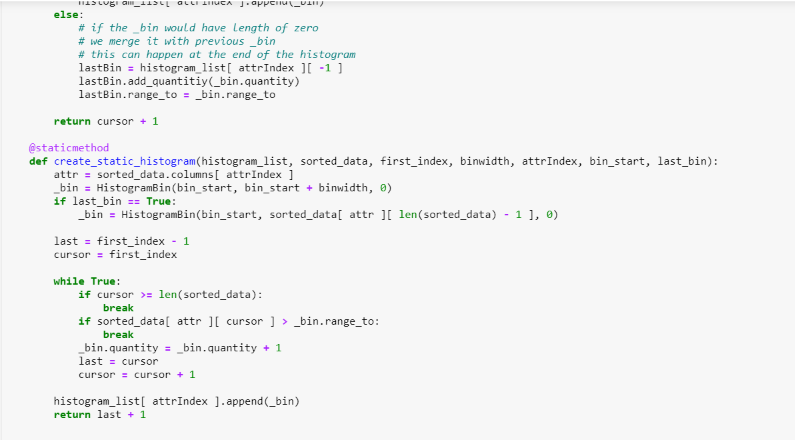


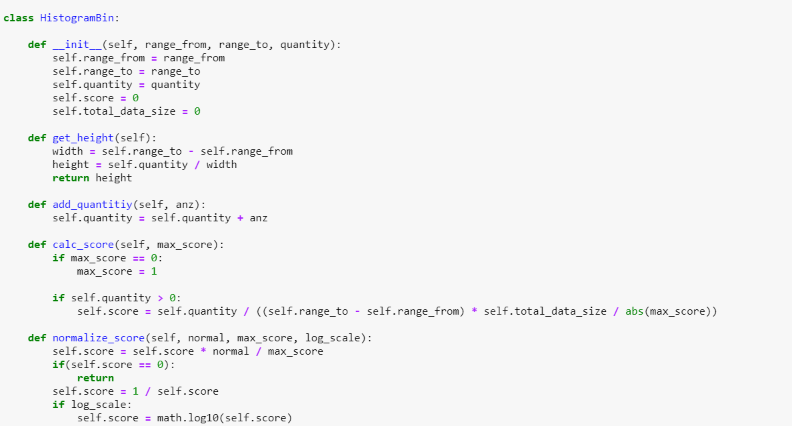




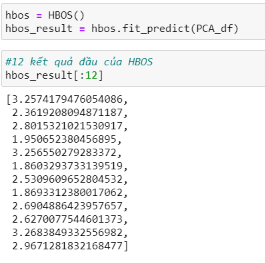








* Chạy thuật toán HBOS trên data đã PCA và xem kết quả. Output của thuật toán HBOS chính là điểm bất thường mà bài báo đề cập. Điểm bất thường càng cao thì khả năng giao dịch là gian lận càng cao.



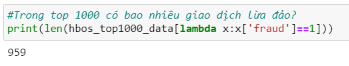
* Thêm điểm hbos vào data\_reduced:



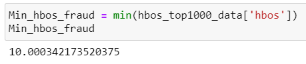
* Lấy top 1000 giao dịch có điểm hbos cao nhất và xem 15 kết quả đầu. Có thể thấy 15/15 giao dịch này là giao dịch lừa đảo:



* Trong 1000 giao dịch này có 959 giao dịch là giao dịch lừa đảo. Tương ứng khoảng 96%.



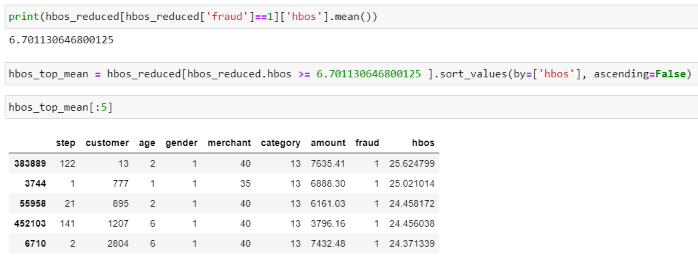
* Điểm hbos nhỏ nhất trong top 1000:



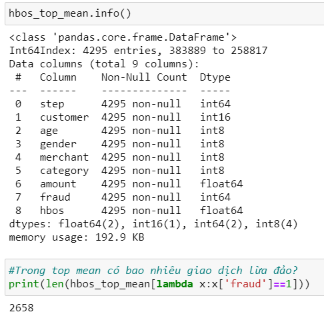
* Điểm hbos trung bình mà giao dịch là lừa đảo:



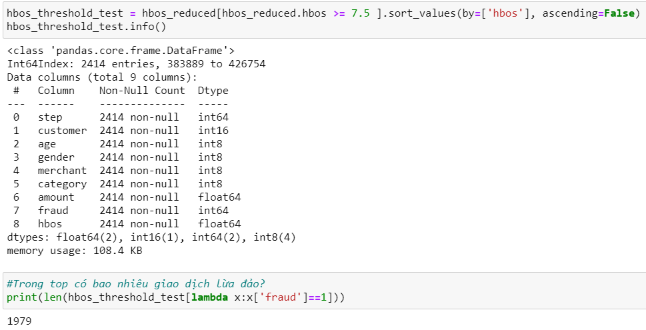
* Lấy top các giao dịch là lừa đảo với điểm hbos từ giá trị trung bình trở lên:



* Xem kết quả cho thấy trong 4295 giao dịch chỉ có 2658 giao dịch là lừa đảo. Tương ứng khoảng 62%.



* Test thử tương tự với giá trị hbos là 7.5. Có thể thấy 1979/2414 giao dịch là lừa đảo, tương ứng với ~82%.



* Ngưỡng hbos tham khảo cho nhà phân tích ngân hàng là khoảng 7.5 trở lên.

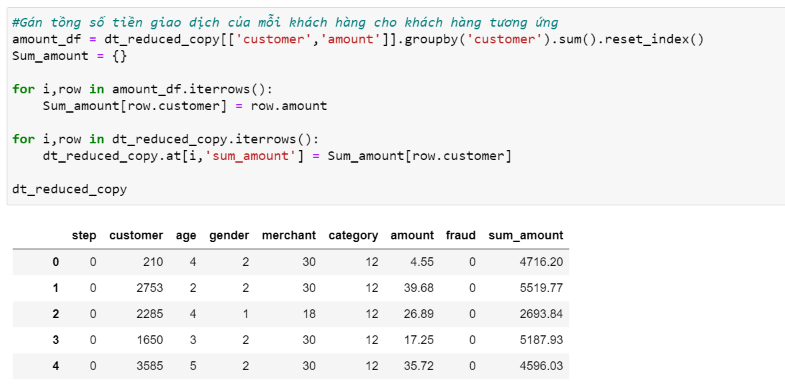
#### 5.1.2.4. Global Profile:

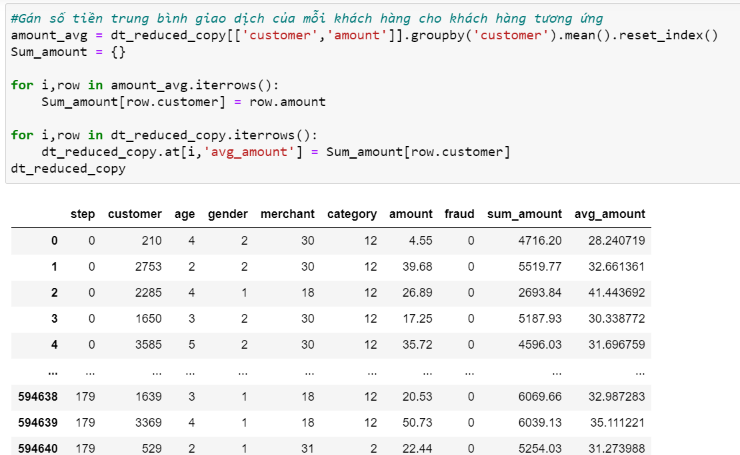
* Tạo dataframe mới là bản copy của dataset “data\_reduced” vì đã được chuyển về dạng số, dễ tính toán hơn.

Ảnh có chứa bàn

Mô tả được tạo tự động

* Bài báo coi mỗi khách hàng đại diện như một vector đặc điểm của sáu thành phần: trung bình số tiền giao dịch, tổng số tiền giao dịch, khoảng thời gian trung bình giữa các giao dịch liên tiếp, số lượng giao dịch thực hiện từ nước ngoài, số lượng giao dịch có tài khoản thụ hưởng ở nước ngoài (chỉ dành cho chuyển khoản ngân hàng), số lượng giao dịch được thực hiện.
* Do sự khác biệt về tệp dữ liệu, chỉ tính được 4 thành phần là: tổng số tiền giao dịch, số tiền trung bình trong các giao dịch, số lần giao dịch và thời gian giao dịch trung bình của mỗi khách hàng.





Ảnh có chứa bàn

Mô tả được tạo tự động

Ảnh có chứa bàn

Mô tả được tạo tự động

* Loại bỏ các khách hàng trùng nhau. Lúc này data còn 4112 hàng, tức là có 4112 khách hàng.

Ảnh có chứa bàn

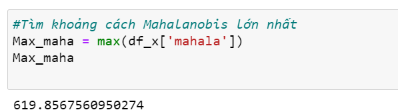
Mô tả được tạo tự động

* Tính khoảng cách Mahalanobis trước khi sử dụng thuật toán DBSCAN.

Ảnh có chứa văn bản

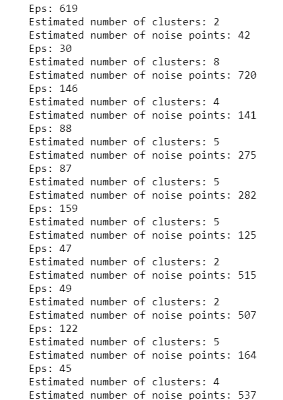
Mô tả được tạo tự động

* Tìm khoảng cách Mahalanobis lớn nhất để làm ε cho DBSCAN. Bài báo cũng sử dụng thuật toán DBSCAN với ε là khoảng cách Mahalanobis lớn nhất.



* Theo bài báo, lặp thuật toán DBSCAN 10 lần để giảm số ε

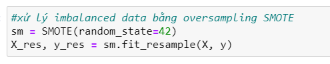




* Import thư viện để sử dụng CBLOF



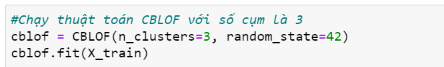
* Vì dataset rất mất cân bằng nên xử lý bằng phương pháp oversampling SMOTE để tăng số lượng gian lận lên, giảm độ mất cân bằng của bộ dữ liệu.



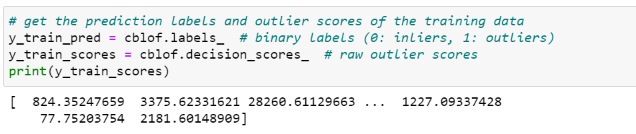
* Chia dữ liệu train và test theo tỉ lệ 80:20



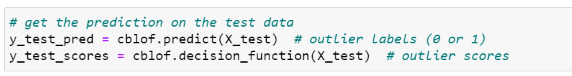
* Chạy thuật toán CBLOF với số cụm là 3.



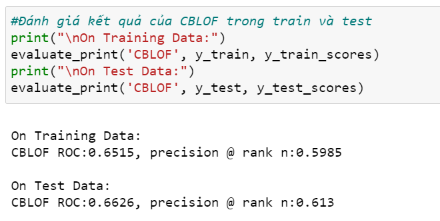
* Lấy các nhãn dự đoán và các điểm bất thường từ data train.



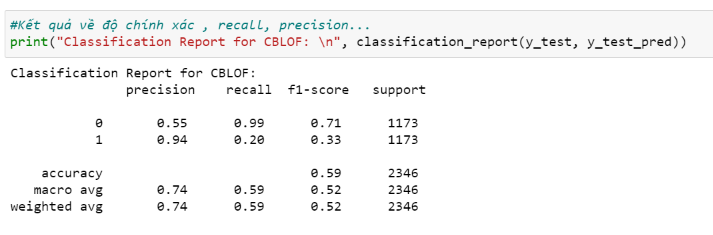
* Dự đoán trên data test.



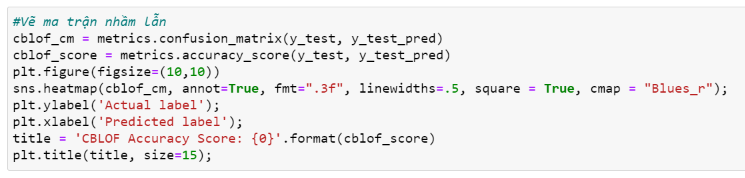
* Đánh giá kết quả của thuật toán CBLOF trong 2 data train và test.



* Đánh giá kết quả trên các chỉ số precision, recall, f1-score.



* Vẽ ma trận nhầm lẫn để hiểu rõ hơn về kết quả. Từ ma trận nhầm lẫn, có thể thấy kết quả dự đoán sai khá lớn. Có tới 942 kết quả dự đoán không gian lận nhưng thực tế là gian lận.

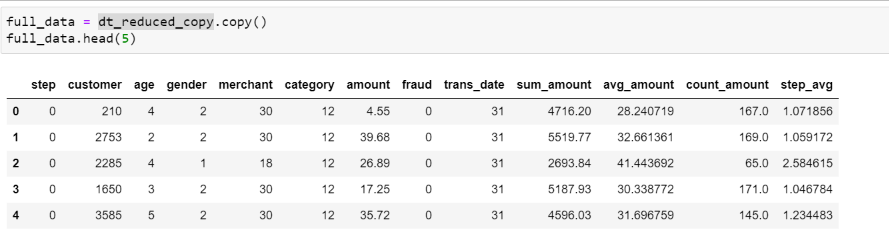




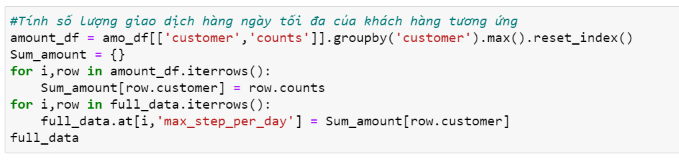
* Kết quả của Global Profiling: Độ chinh xác của thuật toán tương đối thấp, đạt 0.6. Theo kết quả của bài báo thì Global Profiling tính điểm bất thường trong thói quen chi tiêu của 1 khách hàng so với toàn bộ khách hàng. Sự khác biệt giữa kết quả của nhóm với bài báo phần nhiều là do bộ dữ liệu không có đủ thuộc tính như bài báo để tính toán.

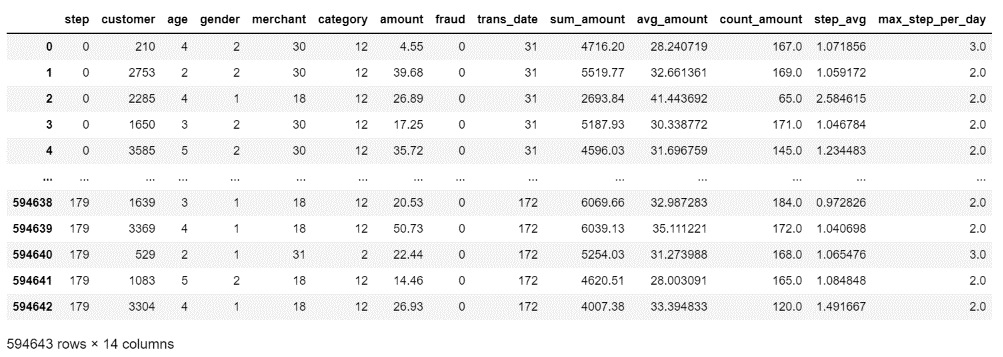
#### 5.1.2.5. Temporal Profile:

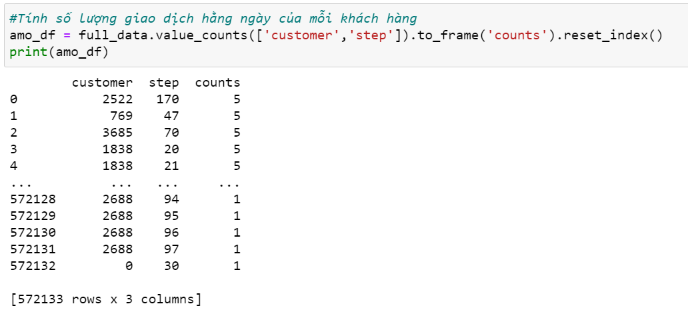
* Tạo 1 dataframe mới cho Temporal Profile để tính toán và chỉnh sửa thuận tiện hơn.



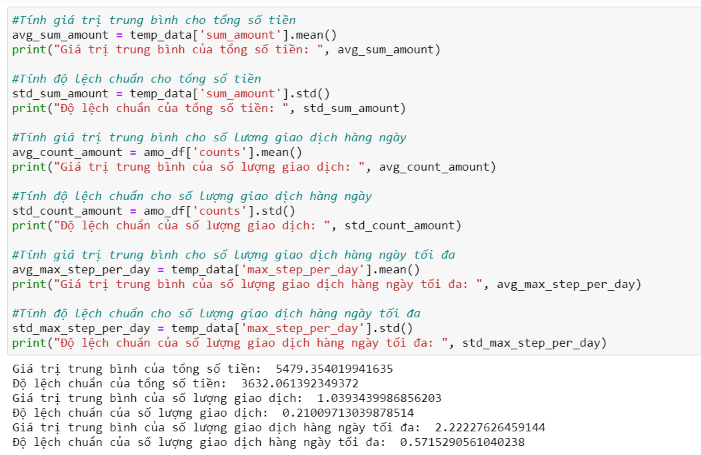
* Tính số lượng giao dịch hằng ngày và số lượng giao dịch hằng ngày tối đa của mỗi khách hàng.



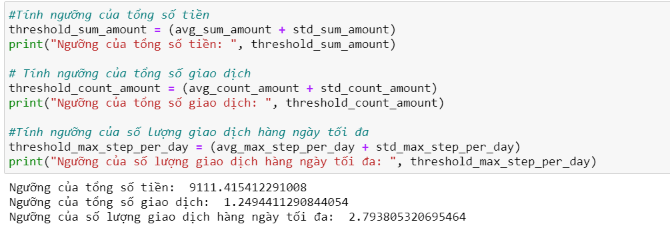




- Sau đó, tính giá trị trung bình và độ lệch chuẩn của từng đặc điểm là: tổng số giao dịch, số lượng giao dịch hàng ngày và số giao dịch hằng ngày tối đa của mỗi khách hàng.

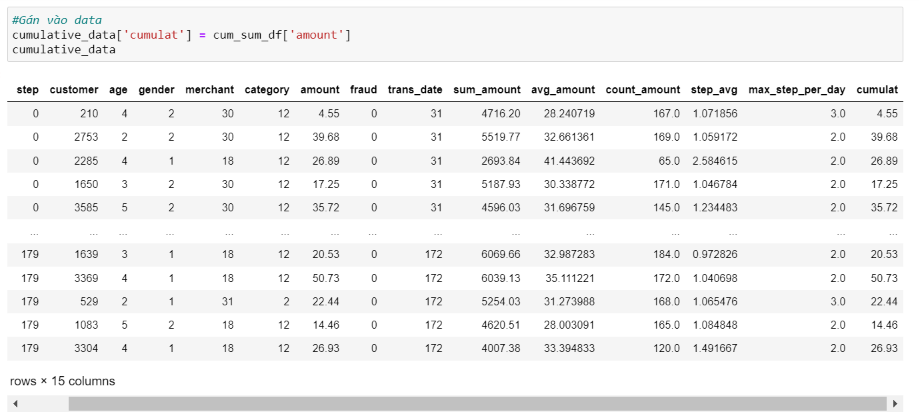


- Đặt ngưỡng giá trị bằng tổng của giá trị trung bình và độ lệch chuẩn cho mỗi đặc điểm đã nêu trên.

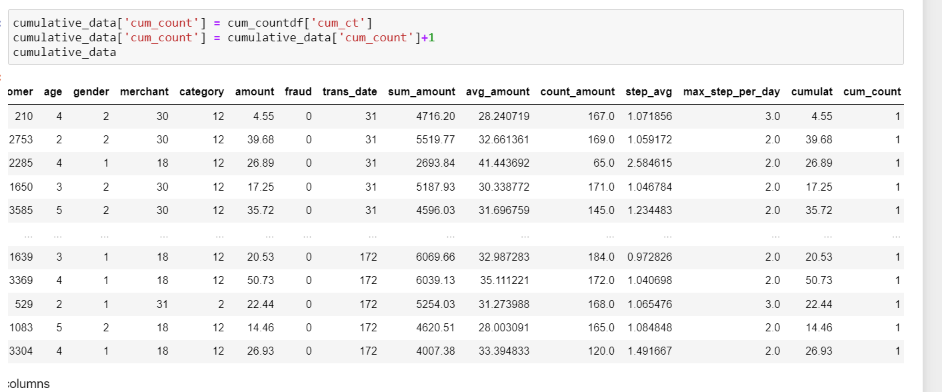


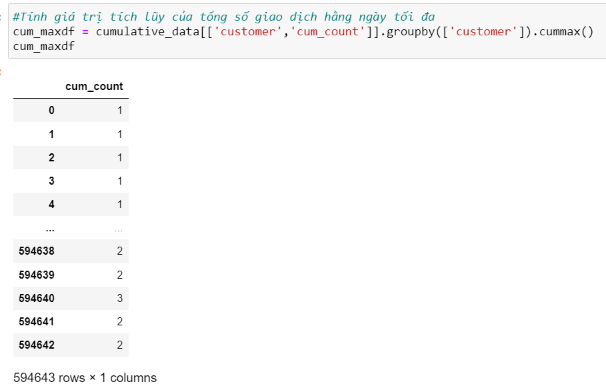
- Theo tần suất lấy mẫu, tính giá trị tích lũy của 3 đặc điểm đã đề cập đến.

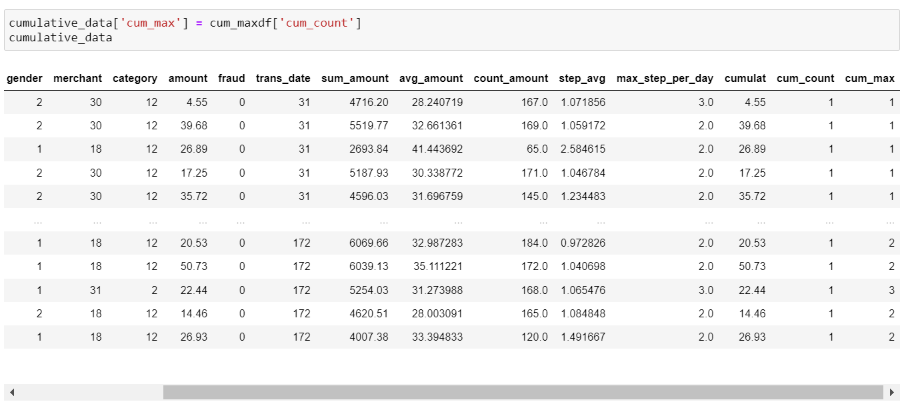




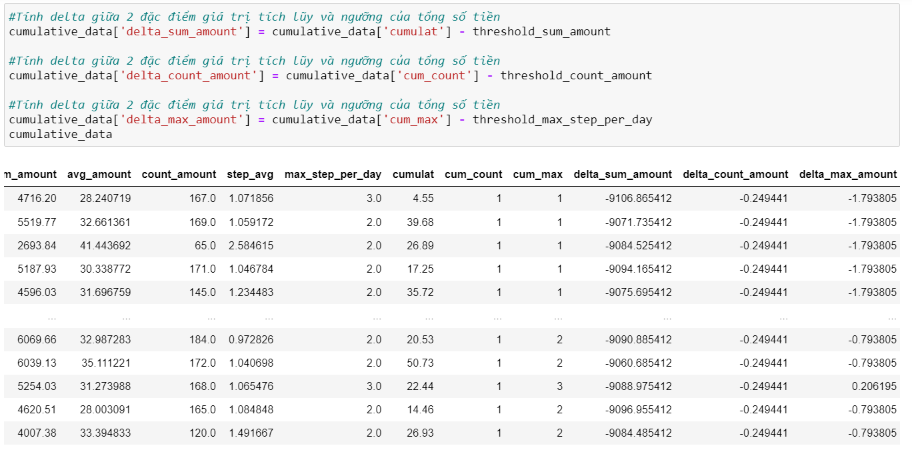




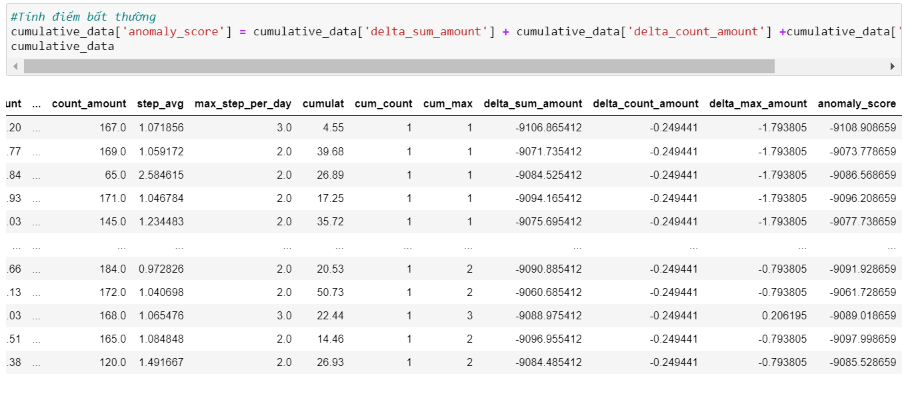




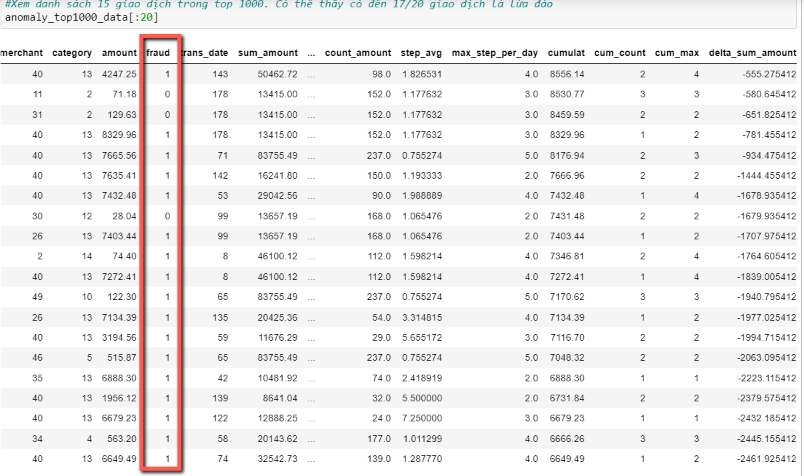
* Tính delta giữa mỗi giá trị tích lũy và ngưỡng tương ứng của các đặc điểm.



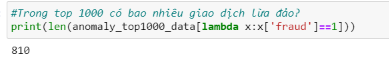
* Tính điểm bất thường bằng tổng delta đã tính.



* Lấy 1000 giao dịch có điểm bất thường cao nhất và xem danh sách 20 giao dịch có điểm bất thường cao nhất. Có 17/20 là giao dịch lừa đảo.



* Trong top 1000 giao dịch có điểm bất thường cao nhất có tới 810 giao dịch lừa đảo.



## 5.2. Đánh giá:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Bài báo | Thực nghiệm | Lý do giống/khác nhau |
| Bộ dữ liệu | Không mô tả chi tiết, nhưng có chứa IP/IBAN, CC… | Sử dụng dataset khác cũng là các giao dịch ngân hàng nhưng không chứa IP, IBAN… | Vì lý do bảo mật nên tác giả không cung cấp bộ dữ liệu. |
| Output | Điểm bất thường 🡺 Xếp hạng rủi ro gian lận | Điểm bất thường | Xếp hạng rủi ro gian lận cần có chuyên gia phân tích ngân hàng |
| Local profile | Thuật toán HBOS | Thuật toán HBOS | Đã tìm hiểu và chạy thành công |
| Global profile | Gom cụm bằng thuật toán DBSCAN  Tính điểm bất thường bằng thuật toán CBLOF | Gom cụm bằng thuật toán DBSCAN  Tính điểm bất thường bằng thuật toán CBLOF | Kết quả có độ chính xác khá thấp do sự khác biệt về dữ liệu. |
| Temporal profile | Ngưỡng và điểm bất thường tính bởi delta dương của giá trị tích lũy và ngưỡng tương ứng | Ngưỡng và điểm bất thường tính bởi hiệu giữa giá trị tích lũy và ngưỡng tương ứng | Tính điểm bất thường bằng hiệu cho kết quả chính xác cao hơn đối với dataset nhóm đang dung |
| Kết quả | Nhà phân tích ngân hàng nhận định đúng 98% giao dịch sau khi xem điểm bất thường ở cả 3 profile | Chỉ tính ra điểm bất thường, có thể đề xuất thêm ngưỡng tham khảo | Không có nhà phân tích ngân hàng để đánh giá |

## 5.3. Ưu điểm và hạn chế:

### 5.3.1. Ưu điểm:

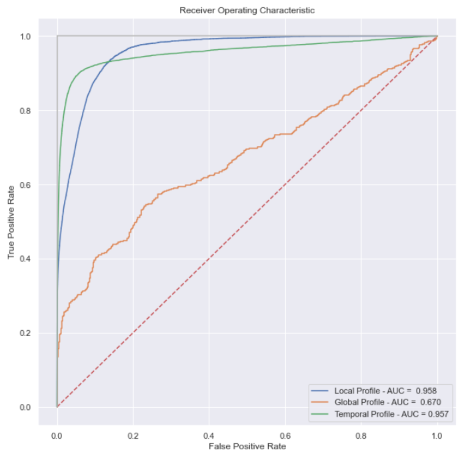
* Cài đặt được các thuật toán mà tác giả sử dụng.
* Đạt được kết quả khá tốt ở Local và Temporal profile.
* Có cải tiến để phù hợp với dataset của nhóm.

### 5.3.2. Hạn chế:

* Global profile có kết quả chưa tốt vì véc-tơ sử dụng không đủ chiều (dataset thiếu).
* Không thể đánh giá tính chính xác vì không có nhà phân tích ngân hàng.

## 5.4. Độ tin cậy, chính xác, hiệu suất:

* **Độ chính xác:** Đánh giá dựa trên AUC (Area Under the Curve) - ROC curve (Receiver Operating Characteristic). AUC – ROC curve là một phép đo hiệu suất cho các vấn đề phân lớp mà có nhiều ngưỡng khác nhau. ROC là đường cong xác suất và AUC là mức độ hoặc thước đo khả năng phân tách. Chúng cho biết khả năng phân biệt giữa các lớp như thế nào. AUC càng cao thì mô hình càng tốt trong việc dự đoán lớp 0 là 0 và lớp 1 là 1. Đường cong ROC được vẽ với TPR (True Positive Rate) trên FPR (False Positive Rate) mà TPR nằm trên trục y và FPR nằm trên trục x.



* Qua sơ đồ nhóm đã vẽ, có thể thấy ở ba lớp thì lớp Local và Temporal có kết quả về độ đo AUC rất cao. Do đó, 2 lớp này cho kết quả dự đoán các giao dịch gian lận và không gian lận chính xác.
* Lớp Global Profile có kết quả độ đo AUC tương đối thấp bằng 0.67. Điều này có nghĩa là có 67% lớp này dự đoán được giao dịch là gian lận hay không.
* Nên dựa vào Local và Temporal Profile để quyết định xem giao dịch có gian lận hay không.

# PHẦN 6: ĐỀ XUẤT HƯỚNG NGHIÊN CỨU

* Hạn chế chủ yếu của phương pháp này là sự thu thập dữ liệu phải xảy ra ở phía người dùng (client side), điều này làm cho nó trở nên bế tắc khi triển khai trong bối cảnh lớn và thực tế. Do đó cần có sự tham gia và cung cấp các dữ liệu về lừa đảo trong ngân hàng.
* Phát triển thêm Global profile để cho kết quả chính xác cao hơn.
* Hạn chế bởi mức tiêu phụ RAM trong khi chạy Local và Global profile. Hạn chế này có thể được giảm nhẹ bằng việc áp dụng giải thuật khác.

# PHẦN 7: KẾ HOẠCH THỰC HIỆN

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Thời gian | Công việc | Ghi chú | Đánh giá tiến độ | |
| 2/3/2021 | Thành lập nhóm |  | Hoàn thành | |
| 9/3/2021-25/3/2021 | Thực hiện Task 2, trả lời 4 câu hỏi để hiểu về đề tài thực hiện |  | Hoàn thành đúng hạn | |
| 23/3/2021-29/3/2021 | Tìm hiểu chi tiết về bài báo, tìm dataset tương đương với bài báo để thực hiện |  | Hoàn thành đúng hạn. | |
| 30/3/2021-12/4/2021 | Phân tích 4 giai đoạn DSS cho đề tài thực hiện. Chạy cài đặt với bộ dữ liệu tìm được. | Trong quá trình tiến hành, tìm thấy 1 bộ dữ liệu khác phù hợp để chạy thực nghiệm hơn nên đã thay đổi. Thực hiện phân tích dữ liệu của dataset mới | Hoàn thành đúng hạn. | |
| 20/4/2021-27/4/2021 | Vẽ sơ đồ và mô tả các thành phần của DSS cho đề tài. | Trong quá trình thực nghiệm, gặp khó khăn vì dataset tìm được đã PCA, không biết nhãn. Do đó, không thể tính toán giống hướng dẫn bài báo. | Hoàn thành đúng hạn. |
| 27/4/2021-11/5/2021 | Tìm dataset khác và phân tích chi tiết, cũng như đánh giá dataset để tiến hành cài đặt thực nghiệm với bộ dataset này |  | Hoàn thành đúng hạn. |
| 11/5/2021-18/5/2021 | Tiếp tục chạy cài đặt thực nghiệm với bộ dữ liệu đã chọn. Bảng đối sánh kết quả của nhóm vả tác giả |  | Hoàn thành đúng hạn. |
| 19/5/2021-25/5/2021 | Tiếp tục chạy cài đặt thực nghiệm. |  | Hoàn thành đúng hạn. |
| 25/5/2021-1/6/2021 | Tiếp tục chạy cài đặt thực nghiệm. Viết file báo cáo, file tóm tắt nội dung đồ án, file đề xuất hướng nghiên cứu. |  | Hoàn thành đúng hạn. |

# PHẦN 8: TÀI LIỆU THAM KHẢO

[1] Edgar Alonso Lopez-Rojas, Stefan Axelsson (2014). BankSim: A Bank Payment Simulation for Fraud Detection Research. Từ https://www.researchgate.net/publication/265736405\_BankSim\_A\_Bank\_Payment\_Simulation\_for\_Fraud\_Detection\_Research

[2] [Financial Fraud Detection](https://drive.google.com/file/d/1OLj7C3X9BErjnmL3YSn9wNrK_tkIjgS5/view?usp=sharing)

[3] [Understanding AUC and ROC Curves](https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5)