ATTACKER 2022

BÁO CÁO

SỬ DỤNG HỌC MÁY TRONG DỰ ĐOÁN GIAN LẬN TRONG HOẠT ĐỘNG TÍN DỤNG CỦA NGÂN HÀNG

Đội thi: ATTACKER ATTACKER

MỤC LỤC

I. Phân tích khám phá dữ liệu (EDA)

A. Mô tả dữ liệu:

Bộ dữ liệu "train_attacker_2022" bao gồm 48030 quan sát (dòng) và 66 biến (cột). Đây là thông tin chi tiết của các giao dịch của khách hàng được ghi chép bởi ngân hàng. Bảng ở dưới là mô tả chung của dữ liệu:

Bảng 1. Mô tả chung [1]

Vào những phần sau của khám phá dữ liệu, chúng tôi sẽ đưa ra các giả thuyết về những dữ liệu đã bị ẩn.

B. Thống kê mô tả

a. Dữ liệu định lượng:

Xử lý một số biến bị định dạng sai trong bộ dữ liệu (các biến cat, dob).
Đồng thời do các biến unknown chưa được xác định, phần thống kê mô tả dữ liệu định lượng sẽ không bao gồm những biến trên.
Biến value có tên biến bị định dạng lỗi " value ", chúng tôi thay đổi thành tên đúng "value"; value có dạng object, chúng tôi đổi thành dạng số (float). Biến review_value có tên biến bị định dạng lỗi " review_value", chúng tôi thay đổi thành tên đúng "review_value"; review_value có dạng object, đổi thành dạng số (float)
Sử dụng phần mềm Python và nhận được kết quả thống kê như sau:

Bảng 2. Thống kê dữ liệu định lượng [2]

Nhìn chung, chúng tôi thấy có sự tương đồng giữa số lượng quan sát các biến không bị trống (khoảng 22991/ 48030 quan sát giữa các biến với dữ liệu từ ngân hàng, 21644/48030 quan sát từ dữ liệu tài khoản xã hội của khách hàng). Về giá trị của từng biến, hệ số nhân (mul_rate) có khoảng cách khá lớn giữa giá trị bình quân và giá trị lớn nhất. Với các vị trí 25%, trung vị, 75% đều là 0, ta có thể suy ra rằng hệ số này phân phối phần lớn về 0. Chúng tôi sẽ kiểm tra kỹ hơn trong phần đồ thị phân phối.
Các biến còn lại: giá trị giao dịch, số ngày kiểm duyệt, giá trị kiểm duyệt, số lượng giao dịch, lượng bạn bè, lượng người đăng ký đều có tính chất tương tự với hệ số nhân.

b. Dữ liệu định tính:

Như phần trên, các biến định tính đã được chuyển thành loại dữ liệu đúng
Chúng tôi có được bảng tổng kết các biến định tính như sau:

Bảng 3. Thống kê dữ liệu định tính [3]

Nhận xét về số lượng các biến:
ID giao dịch và đánh giá giao dịch có đầy đủ dữ liệu. Các biến phân loại (cat_1,..) trừ cat_4, cat_8, cat_12 đều chỉ có 22991/48030 (~48%). Tương tự vậy, biến giới tính, ID địa chỉ khách hàng đăng ký, ID của người bán cũng có khoảng 48% dữ liệu. Biến vị trí hiện tại (thành phố, quốc gia) có 25% dữ liệu; biến quê quán (thành phố, quốc gia) có khoảng 23% số quan sát.
Điều này có thể cho thấy rằng có một xu hướng chung trong dữ liệu thiếu. Chúng tôi sẽ kiểm tra kĩ hơn vấn đề này ở phần sau.
Nhận xét chung về thống kê mô tả từng biến định tính:
id: ID giao dịch có 36027 giá trị đơn nhất, nghĩa khoảng 25% giao dịch bị lặp lại. Với giao dịch xuất hiện thường xuyên nhất là 39888 với tần suất là 6 lần.
label: Số lượng giao dịch trong tập dữ liệu có 65% là giao dịch chính thống (không lừa đảo)
cat_1, cat_2: Dữ liệu phân loại 1, 2 đều có 2 loại với loại xuất hiện thường xuyên nhất lần lượt là C1 và P2 chiếm ~ 54%, không có quá nhiều chênh lệch về số lượng giữa 2 loại trong từng biến cat_1, cat_2.

cat_3, cat_4, cat_5, cat_6: Dữ liệu đều có đa dạng giá trị đơn nhất, nhưng đặc điểm chung giữa các biến này là, giá trị thông dụng nhất chiếm phần lớn trong dữ liệu sẵn có hầu hết lớn hơn 90% đặc biệt là biến cat_5: những quan sát có dữ liệu đều là thuộc 1 nhóm. Về cat_4, mặc dù có 12 giá trị đơn nhất, chỉ có UI là xuất hiện thường xuyên với tần suất cao.
Tương tự các biến cat_7 đến cat_12, mặc dù có nhiều nhóm nhưng tần suất giao dịch hầu hết tập trung chỉ vào một số nhóm cụ thể. Ta sẽ kiểm tra kỹ hơn ở phần phân phối trong đồ thị.
sex: Về giới tính của khách hàng, tần suất khá cân bằng giữa nam và nữ (57%, 43%)
location_id: ID địa chỉ khách hàng đăng ký có 34 nhóm với nhóm DN là phổ biến nhất chiếm hơn 59%
mer_id: ID của người bán khá đa dạng với 10160 nhóm trong 22991 lượng dữ liệu sẵn có, ID người bán có nhiều giao dịch nhất trong dữ liệu là BO0001Z, xuất hiện trong 715 quan sát.
trans_currency: Loại tiền tệ của giao dịch phần lớn là VN (đồng) chiếm hơn 93% số dữ liệu sẵn có.
com_type: Về phân loại khách hàng doanh nghiệp, ta có 9 nhóm, nhưng nhóm Vùng 1 chiếm hơn 50% số quan sát.
social_sex_info, social_location_id: Về thông tin trên mạng xã hội của khách hàng, 59% tài khoản có dữ liệu là nam giới, có thể do số lượng thông tin thu thập được chưa được đầy đủ nên có sự khác biệt giữa thông tin trên mạng xã hội và thông tin từ ngân hàng. ID địa chỉ trên tài khoản xã hội của khách hàng trên 96% là ở Việt Nam.
current_location_city, current_location_country: Địa chỉ hiện tại của khách hàng về thành phố: phần lớn từ Hồ Chí Minh khoảng 20%, về quốc gia: là Việt Nam trên 97%.
hometown_location_city, hometown_location_country: Thông tin quê quán khách hàng: thành phố khác với địa chỉ hiện tại 11% là ở Hà Nội, còn quốc gia vẫn là Việt Nam (98%)

c. Các biến chưa xác định và khác

Sau quan sát dữ liệu, chúng tôi đưa ra:

Bảng 4. Giả định loại dữ liệu của các biến chưa xác định [4]

Các biến unknown_var_5, unknown_var_6, unknown_var_11, unknown_var_16, unknown_var_18, unknown_var_19, unknown_var_20 được xác định là biến định tính trong thang đo Likirt.
Các biến còn lại sẽ là biến định lượng.

Bảng 5. Thống kê biến unknown định tính [5]

Về số lượng quan sát sẵn có, unknown_var_18, 19, 20 có lượng giống nhau, thể hiện mối quan hệ trong nguồn dữ liệu. Tương tự với các biến định tính chúng tôi đã khám phá ở trên, phần lớn dữ liệu tập trung vào một vài nhóm nhất định. Đặc biệt ở unknown_var_6 và unknown_var_16 trên 90% dữ liệu.

Bảng 6. Thống kê biến unknown định lượng [6]

Có thể thấy, lượng dữ liệu của biến unknown_var_13, 14, 15 giống nhau (22991). unknown_var_1, 2,3,4 đều có giá trị nhỏ nhất là 1 với giá trị lớn nhất khá gần nhau. Từ các bách phân vị, giá trị trung bình, chúng tôi nhận thấy hầu hết các giá trị nằm về phía bên tay trái (nhỏ hơn), trong khi giá trị lớn nhất khá khác biệt (có thể là dữ liệu ngoại lai). Về các biến unknown_var_7, unknown_var_9, unknown_var_10, unknown_var_17, giá trị chạy từ âm đến dương, hầu hết các giá trị nằm về phía âm (nhỏ hơn). Các biến unknown_var_12, 13 chạy từ 0 đến 1 với các giá trị trung bình là 0.31, 0.42.
Biến unknown_var_14 có giá trị từ -1 đến 1. Biến unknown_var_15 nhận giá trị từ -1 đến 2 với bách phân vị thứ 75 là 0 nghĩa là 75% quan sát có sẵn nhận giá trị nhỏ hơn hoặc bằng 0.

Các dữ liệu khác

- Dữ liệu thời gian

Bảng 7. Thống kê dữ liệu thời gian [7]

Chúng tôi nhận thấy biến thời gian giao dịch có số lượng bằng nhiều biến đã khám phá ở trên 22991. Thời gian xuất hiện thường xuyên nhất ở cả time_1 và time_2 là giống nhau, tương tự với ngày tháng của Field_11 và date_4

- Dữ liệu văn bản

Bảng 8. Thống kê dữ liệu văn bản [8]

Lượng thông tin của người bán hàng có rất ít 7570/48030, không chỉ vậy dữ liệu xuất hiện thường xuyên nhất trong biến này là dấu ".", chưa rõ ràng là dấu này phản ánh điều gì trong biến.
Địa chỉ khách hàng, tên người bán, công việc khách hàng đều có số lượng là 22991. Về địa chỉ khách hàng và tên người bán, có khá nhiều giá trị đơn nhất. Từ thông tin biến công việc khách hàng, ta nhận thấy "Doanh nghiệp có vốn đầu tư nước ngoài" là phổ biến nhất chiếm 24% quan sát sẵn có của biến. Cụ thể về chi tiết công việc khách hàng, công nhân là khách hàng thực hiện nhiều giao dịch được ghi lại trong dữ liệu nhất với 8.87% giao dịch.
Vị trí giao dịch cũng có khá nhiều giá trị đơn nhất, với vị trí khách hàng thường xuyên thực hiện giao dịch nhất là "153 Xô Viết Nghệ Tĩnh Quận Bình Thạnh" với 337 giao dịch

C. Dữ liệu thiếu

Bảng 9. Thông tin dữ liệu thiếu [9]

Từ bảng, chúng tôi quan sát được là nhiều dữ liệu thiếu cùng số lượng, có nhiều biến thiếu nhiều hơn 40% dữ liệu. Vấn đề này sẽ cần được xử lý.

D. Phân phối của dữ liệu:

1. Dữ liệu định tính:

Chúng tôi vẽ phân phối của các biến định tính, trong mục này sẽ thể hiện những đồ thị quan trọng, bất thường, và nhận xét.

Hình 1. Đồ thị loại giao dịch [I]

Số lượng giao dịch chính thống gần gấp đôi lượng giao dịch lừa đảo.
Về biến cat_5

Hình 2. Đồ thị biến cat_5 [II]

Tất cả dữ liệu sẵn có của cat_5 đều chỉ thuộc vào một nhóm.

2. Dữ liệu định lượng

Hình 3. Đồ thị biến hệ số nhân [III]

Phân phối của biến hệ số nhân có xu hướng lệch phải. Chúng tôi có thể biến đổi dữ liệu thành dạng Log để nhận được phân phối chuẩn.
Tương tự vậy, các biến giá trị giao dịch, số ngày kiểm định giao dịch, giá trị giao dịch kiểm định, lượng bạn bè, người đăng ký trên tài khoản mạng xã hội của khách hàng cũng lệch phải.

Hình 4. Đồ thị biến số lượng giao dịch của các khách hàng trong tháng trước [IV]

Trong khi đó, phân phối của số lượng giao dịch của các khách hàng trong tháng trước có sự không quy luật.
Về các biến chưa xác định, phần lớn cũng có phân phối lệch phải. Với ngoại lệ như

Hình 5. Đồ thị biến unknown_var_7 và unknown_var_9 [V]

Biến unknown_var_7, 9 gần với phân phối chuẩn

Hình 6. Đồ thị biến unknown_var_10, unknown_var_13, unknown_var_14, unknown_var_15 [VI]

Hình 7. Đồ thị biến unknown_var_17 [VII]

Unknown_var_10, 13, 14, 15, 17 có giao động trong phân phối.

II. Làm sạch dữ liệu

A. Lỗi định dạng

Chúng tôi tiến hành xử lý vấn đề đầu tiên của bộ dữ liệu là về lỗi định dạng các biến. Cụ thể, dữ liệu gốc chứa các biến sau ở định dạng lỗi:

- time_1, time_2,... đang ở định dạng datetime có timezone
- trans_location, job, com_type, job_detail, current_location_city, hometown_location_city lỗi định dạng mã hóa

Cách xử lý lỗi đã trình bày ở trên gồm 2 hướng giải quyết:

- Đưa định dạng time_1, time_2 về datetime
- Mã hóa UTF8 cho trans_location, job, com_type, job_detail, current_location_city, hometown_location_city

B. Loại bỏ các biến không dùng đến

Trong bài báo cáo này, chúng tôi sẽ loại bỏ các biến không dùng đến trong quá trình chạy mô hình machine learning. Các biến không cần thiết bao gồm:
id, time_1, time_2, date_1, date_2, date_3, date_4, address, mer_id, mer_name, trans_location, job, com_type, job_detail, current_location_city, current_location_country, hometown_location_city, hometown_location_country, cat_5

C. Xử lý dữ liệu bị thiếu

Tiếp theo, chúng tôi tiến hành xử lý vấn đề dữ liệu bị thiếu. Dựa trên quan sát, bộ dữ liệu cho thấy có nhiều biến có đặc điểm thiếu cùng nhau (missing-together variables). Cụ thể, với các bản ghi bị thiếu giá trị ở biến time_1, thì các biến khác như time_2, Field_11, date_3, date_4, v.v cũng bị thiếu theo. Điều này có thể được thể hiện qua bảng tính số lượng các giá trị bị thiếu của các biến có đủ giá trị time_1 và các biến bị thiếu giá trị time_1 dưới đây:

Bảng 10. Số lượng các giá trị bị thiếu của các biến có đủ giá trị time_1 [10]

Bảng 11. Số lượng các giá trị bị thiếu của các biến bị thiếu giá trị time_1 [11]

Từ bảng trên có thể chỉ ra được, trong số 25039 các quan sát bị thiếu giá trị time_1, thì số lượng các giá trị bị thiếu của biến time_2, Field_11, date_3, date_4 đều bằng nhau và bằng 25039 giá trị (tức là thiếu 100% số lượng các giá trị). Thêm vào đó, ở các biến khác cũng cho thấy số lượng dữ liệu bị thiếu tương đối nhiều và cũng ở mức 100%. Như vậy, có thể kết luận là các biến time_1, time_2, Field_11, v.v thể hiện đặc điểm thiếu cùng với nhau.
Bởi vì số lượng các quan sát bị thiếu giá trị time_1 là rất lớn (25039 quan sát trên tổng số 48030 quan sát, tương ứng 52.13%), cách xử lý thông thường là tiến hành bỏ các quan sát là không khả thi. Để xử lý vấn đề này, chúng tôi tiến hành chia bộ dữ liệu gốc thành 2 phần dữ liệu và sau đó sẽ chạy các mô hình với từng phần dữ liệu:

- Phần dữ liệu thứ nhất bao gồm các quan sát có đủ giá trị time_1
- Phần dữ liệu thứ hai gồm các quan sát thiếu giá trị time_1

Giải pháp xử lý dữ liệu bị thiếu đối với hai phần dữ liệu này đều thống nhất như nhau và gồm các bước sau:

- Với các biến thiếu đáng kể các quan sát (thiếu từ 40% trở lên) thì không có giá trị trong phân tích dữ liệu nên sẽ tiến hành loại bỏ các biến này.
- Với các biến thiếu từ 40% dữ liệu trở xuống, tiến hành điền lại các giá trị bị thiếu cho các biến đó bằng cách thay thế các giá trị bị thiếu bằng giá trị trung bình (mean), trung vị (median), và yếu vị (mode) của tập dữ liệu đã có. Đối với các biến số cần điền dữ liệu, nếu xuất hiện nhiều giá trị outlier thì sẽ áp dụng cách điền dữ liệu thiếu bằng yếu vị (median), nếu số lượng các outlier tương đối ít thì sẽ áp dụng cách điền dữ liệu bằng trung bình (mean)

Đồ thị boxplot của phần dữ liệu bao gồm các quan sát có đủ giá trị time_1 được thể hiện ở hình 8 dưới đây

Hình 8. Đồ thị boxplot của phần dữ liệu bao gồm các quan sát có đủ giá trị time_1 [VIII]

Đồ thị boxplot của phần dữ liệu bao gồm các quan sát thiếu giá trị time_1 được thể hiện trong hình 9 dưới đây

Hình 9. Đồ thị boxplot của phần dữ liệu bao gồm các quan sát thiếu giá trị time_1 [IX]

D. Kiểm tra mức độ tương quan giữa các biến

Đa cộng tuyến cao (multicollinearity) là một hiện tượng thường gặp trong phân tích dữ liệu. Sự xuất hiện của hiện tượng đa cộng tuyến làm các mô hình bị ước lượng sai và chệch khỏi giá trị chính xác. Chúng tôi xử lý vấn đề này bằng cách kiểm tra hệ số tương quan giữa các biến trong dữ liệu, đối với hai biến có hệ số tương quan từ 0.9 trở lên thì sẽ tiến hành loại bỏ một trong hai biến.
Đối với phần dữ liệu bao gồm các quan sát có đủ giá trị time_1, hệ số tương quan của các biến số và các biến phân loại được thể hiện ở hình 10 và hình 11 lần lượt dưới đây:

Hình 10. Đồ thị hệ số tương quan của các biến số của dữ liệu có time_1 [X]

Hình 11. Đồ thị hệ số tương quan Crammer's V của các biến phân loại của dữ liệu có time_1 [XI]

Đối với phần dữ liệu bao gồm các quan sát thiếu giá trị time_1, hệ số tương quan của các biến số và các biến phân loại được thể hiện ở hình 12 và hình 13 lần lượt dưới đây:

Hình 12. Đồ thị hệ số tương quan Crammer's V của các biến số của dữ liệu thiếu time_1 [XII]

Đối với phần dữ liệu thiếu giá trị của time_1 thì chỉ có duy nhất một biến phân loại (biến unknown_var_5) nên không có đồ thị tương quan giữa các biến phân loại (chi tiết ở Bảng 11)

III. Xây dựng mô hình

A. Mô hình hồi quy Logistic

Vì bộ dữ liệu được định dạng và gắn nhãn (labeled data) nên nhóm nghiên cứu trước tiên sử dụng các mô hình học có giám sát (supervised learning). Trước hết, chúng tôi sử dụng mô hình hồi quy logistics vì tính đơn giản, phù hợp của mô hình và khả năng giải thích cao của nó.

Mô hình hồi quy logistic là một mô hình thống kê dùng cho các bài toán phân loại và dự báo. Mô hình này ước tính khả năng xảy ra của một biến phân loại dựa trên các biến độc lập. Với bài toán phát hiện gian lận, mục tiêu của chúng tôi là xác định giao dịch có gian lận hay không (biến nhị thức) – trùng khớp với kĩ thuật và mục tiêu của mô hình logistics. Về ý tưởng hình thành mô hình, logistics gần giống với mô hình hồi quy tuyến tính. Điểm khác biệt là mô hình logistics dùng hàm sigmoid để giới hạn giá trị của biến phụ thuộc trong khoảng từ 0 đến 1 – đưa ra dự đoán về khả năng xảy ra của sự kiện.

Sau khi làm sạch dữ liệu, điền các dữ liệu trống, và loại bỏ các biến tương quan cũng như các biến định danh không có giá trị dự đoán, chúng tôi chia thành 2 mô hình bao gồm: mô hình dữ liệu có đủ số liệu time_1 (mô hình 1) và mô hình thiếu dữ liệu time_1 (mô hình 2)

Kết quả thu được khi chạy mô hình hồi quy tuyến tính là:

Bảng 12. Chỉ số của mô hình hồi quy Logistic

	Model_1	Model_2
Accuracy	0.639	0.654
Precision	0.344	0.5
Recall	0.054	0.002
F1	0.093	0.005

Độ chính xác (Accuracy) của 2 mô hình khá gần nhau xấp xỉ 63.9% và 65.4% lần lượt. Mô hình có khả năng dự đoán chính xác khoảng hai phần ba số lượng quan sát.

Chỉ số Precision giữa 2 mô hình cũng xấp xỉ nằm ở mức 0.344 và 0.5. Chỉ số Recall và F1 của 2 mô hình tương đối thấp cho thấy hiệu quả dự báo không tốt.

B. Mô hình Deep Learning (Mạng nơ-ron)

Do để phát hiện lừa đảo trong tài chính, mô hình mạng nơ-ron nhân tạo có hiệu quả cao và vô cùng phổ biến, chúng tôi quyết định áp dụng mô hình này trong bài nghiên cứu.
Mạng nơ-ron có sự tương đồng chuẩn mạnh với những phương pháp thống kê như đồ thị đường cong và phân tích hồi quy.[1]

Mạng nơ-ron là một phương thức trong lĩnh vực trí tuệ nhân tạo, được sử dụng để dạy máy tính xử lý dữ liệu theo cách được lấy cảm hứng từ bộ não con người. Đây là một loại quy trình máy học, được gọi là deep learning, sử dụng các nút hoặc nơ-ron liên kết với nhau trong một cấu trúc phân lớp tương tự như bộ não con người. Phương thức này tạo ra một hệ thống thích ứng được máy tính sử dụng để học hỏi từ sai lầm của chúng và liên tục cải thiện.[2]

Tương tự như trên, chúng tôi chạy mô hình deep learning trên 2 phần dữ liệu: có đủ số liệu time_1 (mô hình 1) và mô hình thiếu dữ liệu time_1 (mô hình 2).

Kết quả thu được khi chạy mô hình deep learning là:

Bảng 13. Chỉ số của mô hình Deep learning

	Model_1	Model_2
Accuracy	0.5392	0.5704
Precision	0.3369	0.2688
Recall	0.3525	0.3476

F1	0.3445	0.3032

Nhìn chung các chỉ số của mô hình Deep learning không quá khả quan: dự đoán chính xác chỉ được hơn nửa. Đồng thời, các chỉ số Precision, Recall và F1 cũng không quá cao chỉ tầm hơn 30%.

C. Mô hình Random Forest

Mô hình decision tree được tạo nên từ những chuỗi câu lệnh if-else (dựa trên các biến độc lập) từ đó tạo nên một lưu đồ giúp chúng tôi dự đoán kết quả dựa trên tập dữ liệu. Ý tưởng của cây quyết định là chia tập dữ liệu thành các tập dữ liệu nhỏ hơn dựa trên những đặc điểm được mô tả. Mỗi câu hỏi giúp một cá nhân đi đến quyết định cuối cùng, quyết định này sẽ được biểu thị bằng nút lá. Các quan sát phù hợp với tiêu chí sẽ đi theo nhánh "Có" và những quan sát không phù hợp sẽ đi theo con đường thay thế.Dựa trên đó, mô hình random forest được tạo thành từ nhiều mô hình decision tree một cách ngẫu nhiên. Random forest được coi là một trong những mô hình học có giám sát được ứng dụng trong bài toán phân loại. So với decision tree, random forest hiệu quả hơn vì nó loại bỏ được tính thiên vị trong mô hình, giảm bớt vấn đề sự quá vừa trong dữ liệu (overfitting data).

Mô hình rừng cây được huấn luyện dựa trên sự phối hợp giữa luật kết hợp (ensembling) và quá trình lấy mẫu tái lặp (boostrapping). Cụ thể thuật toán này tạo ra nhiều cây quyết định mà mỗi cây quyết định được huấn luyện dựa trên nhiều mẫu con khác nhau và kết quả dự báo là bầu cử (voting) từ toàn bộ những cây quyết định.[3] Như vậy một kết quả dự báo được tổng hợp từ nhiều mô hình nên kết quả của chúng sẽ không bị chệch. Đồng thời kết hợp kết quả dự báo từ nhiều mô hình sẽ có phương sai nhỏ hơn so với chỉ một mô hình.
Chúng tôi chạy mô hình này trên 2 dữ liệu: có đủ số liệu time_1 (mô hình 1) và mô hình thiếu dữ liệu time_1 (mô hình 2):

Kết quả thu được khi chạy mô hình là:

Bảng 14. Chỉ số của mô hình Random Forest

	Model_1	Model_2
Accuracy	0.638	0.616
Precision	0.309	0.344
Recall	0.0209	0.134
F1	0.0393	0.193

Về tỷ lệ dự đoán chuẩn xác, mô hình Random Forest có kết quả cao nhất lên tới 63.8% và 61.6%, lần lượt. Về chỉ số precision, của mô hình này rơi vào khoảng 31-34%. Chỉ số recall và f1 tương đối thấp. Ở mô hình 1, hai chỉ số này vào khoảng 3%; còn ở mô hình 2, hai chỉ số này nằm tại khoảng 15%.

D. Mô hình SVM

Mô hình Máy Vector hỗ trợ, hay còn gọi là Support Vector Machine (viết tắt SVM) là một thuật toán giúp

tìm ra một siêu phẳng phân cách tối ưu để có thể phân chia dữ liệu thành các lớp khác nhau. Máy Vector hỗ trợ là một thuật toán phổ biến nhất trong học máy, được sử dụng để phân loại, hồi quy và phát hiện điểm dữ liệu bất thường. [4]

Mục tiêu của thuật toán SVM là tạo ra đường hoặc ranh giới quyết định tốt nhất có thể phân tách không gian n chiều thành các lớp để chúng tôi có thể dễ dàng đặt điểm dữ liệu mới vào đúng phân lớp trong tương lai. Ranh giới quyết định tốt nhất này được gọi là siêu phẳng. [5]

Đối với các dữ liệu có nhiều biến thì không thể phân chia các lớp theo tuyến tính (không linear separable). Với trường hợp này thì có thể dùng một kỹ thuật là sử dụng kernel để chuyển đổi dữ liệu ban đầu từ không phân biệt tuyến tính sang không gian mới, ở không gian mới này, dữ liệu trở nên phân biệt tuyến tính.[6]

Bảng 15. Chỉ số của mô hình SVM

	Model_1	Model_2
Accuracy	0.64057	0.65794
Precision	0	0
Recall	0	0
F1	0	0

Mặc dù chỉ số Accuracy của mô hình SVM khá cao, song mô hình trả về cho chỉ số Precision bằng 0, nghĩa là toàn bộ các quan sát trong dữ liệu test đều được dự đoán là 0. Điều đó cũng dẫn tới các chỉ số còn lại đều kém (riêng chỉ số F1, về mặt công thức toán học là không tồn tại vì Precision và Recall đều bằng 0). Do đó, mô hình SVM không cho thấy sự khả quan về khả năng dự đoán gian lận.

IV. So sánh và chọn mô hình

Do bản chất của vấn đề nghiên cứu là phát hiện gian lận trong hoạt động giao dịch của ngân hàng, vì vậy chỉ số F1 cần được ưu tiên cao nhất. Cụ thể vì, Recall cao có nghĩa tỉ lệ bỏ sót các sample positive thực thấp - tối thiểu việc nhận nhầm các nhãn Positive thực thành False Negative.
Dựa trên các phương pháp đã thực hiện ở trên, và với ưu tiên chỉ số F1, chúng tôi quyết định lựa chọn mô hình học sâu (Deep Learning).

V. Tiến hành dự đoán

Sau khi đã chọn được mô hình, chúng tôi tiến hành dự đoán.
Dữ liệu từ file test được xử lý bằng cách fill các biến theo các giá trị trung vị (median) và yếu vị (mode) tương tự như phần xử lý dữ liệu đã trình bày ở trên. Đối với các quan sát có đủ giá trị của time_1, chúng tôi áp dụng Model_1 đã thu được ở phần trước, và đối với các quan sát thiếu giá trị của time_1, chúng tôi áp dụng Model_2 cũng đã được lập từ phần trước.
Kết quả thu được được thể hiện dưới Bảng 16 sau và Hình 10 Ma trận lỗi (Confusion Matrix):

Bảng 16. Chỉ số kết quả dự đoán

Chỉ số	Kết quả
Accuracy	0.596

F1-score	0.565
Precision	0.345
Recall	0.209

Hình 10. Ma trận lỗi (Confusion Matrix)

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

Dựa vào kết quả thu được, có thể dễ dàng nhận ra số lượng các dự đoán là 0, thực tế là 1 (false negative) tương đối lớn. Bởi vì vấn đề nghiên cứu ở đây là phát hiện gian lận trong hoạt động khách hàng của một ngân hàng, với số lượng false negative lớn như vậy cho thấy mô hình học máy vẫn chưa thực sự hiệu quả.

Vấn đề phát sinh có thể nằm ở bước xử lý dữ liệu, do còn điền dữ liệu theo các phỏng đoán và 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.

PHỤ LỤC

Bảng 1. Mô tả chung [1]

Tên biến	Mô tả	Loại dữ liệu
id	ID giao dịch	Định tính
label	Giao dịch được đánh giá là lừa đảo (1) hoặc không lừa đảo (0)	Định tính
time_1	Thời gian giao dịch được thực hiện bởi khách hàng	Thời gian
time_2	Thời gian giao dịch được hoàn thành	Thời gian
Field_11, date_1 - date_4	Dữ liệu thời gian	Thời gian
cat_1 - cat_12	Dữ liệu phân loại	Định tính
mer_des	Thông tin của người bán	Văn bản
mul_rate	Hệ số nhân	Định lượng
value	Giá trị giao dịch	Định lượng
num_date_review	Số ngày giao dịch nằm trong hệ thống kiểm duyệt của ngân hàng	Định lượng
review_value	Giá trị giao dịch có thể bị kiểm duyệt	Định lượng
dob	Ngày sinh của khách hàng	Thời gian
sex	Giới tính của khách hàng	Định tính
address	Địa chỉ khách hàng đăng ký	Văn bản
location_id	ID địa chỉ khách hàng đăng ký	Định tính
mer_id	ID của người bán	Định tính
mer_name	Tên người bán	Văn bản
trans_location	Địa điểm thực hiện giao dịch	Văn bản
trans_currency	Loại tiền tệ của giao dịch	Định tính
job	Công việc của khách hàng	Văn bản
num_trans_last_month	Số lượng giao dịch khách hàng thực hiện trong tháng trước	Định lượng
com_type	Phân loại khách hàng doanh nghiệp	Định tính
job_detail	Chức vụ công việc khách hàng	Văn bản
unknown_var_1 - unknown_var_20	Chưa xác định	Chưa xác định
social_friend_count	Lượng bạn bè trên tài khoản xã hội của khách hàng	Định lượng
social_sex_info	Thông tin giới tính trên tài khoản xã hội của khách hàng	Định tính
social_subcriber_count	Lượng người đăng ký trên tài khoản xã hội của khách hàng	Định lượng
social_location_id	ID địa chỉ trên tài khoản xã hội của khách hàng	Định tính

current_location_city	Vị trí hiện tại của khách hàng (Thành phố)	Định tính
current_location_country	Vị trí hiện tại của khách hàng (Quốc gia)	Định tính
hometown_location_city	Quê quán của khách hàng (Thành phố)	Định tính
hometown_location_country	Quê quán của khách hàng (Quốc gia)	Định tính

Bảng 2. Thống kê dữ liệu định lượng [2]

Tên biến	Count*	Mean*	Std*	Min*	25%*	50%*	75%*	Max*
mul_rate	22991	0.27	0.84	0	0	0	0	6.86
value	22991	3621327	2297407	0	1490000	3899480	4500000	29800000
num_date_review	22991	31.51	46.54	0	1	14	40	464
review_value	22991	222518.84	840799.44	0	0	0	0	20000000
num_trans_last_month	22991	15.02	17.1	3	4	5	33	160
social_friend_count	21644	491.5	999	0	0	29	477	5000
social_subcriber_count	21644	190.61	2290.01	0	0	0	0	174916

*Count: Số lượng quan sát không bị trống của biến (thiếu dữ liệu)
*Mean: Giá trị bình quân của biến
*Std: Độ lệch chuẩn của biến
*Min: Giá trị nhỏ nhất của biến
*25%: Tứ phân vị thứ nhất/bách phân vị thứ 25 - vị trí có 25% trên tổng số quan sát nhận giá trị nhỏ hơn hoặc bằng giá trị tại điểm đó
*50%: Tứ phân vị thứ hai/bách phân vị thứ 50/trung vị - giá trị giữa của quan sát, 50% tổng số quan sát nhận giá trị nhỏ hơn hoặc bằng giá trị tại điểm bách phân vị thứ 50
*75%: Tứ phân vị thứ ba/bách phân vị thứ 75 - vị trí có 75% trên tổng số quan sát nhận giá trị nhỏ hơn hoặc bằng giá trị tại điểm đó
*Max: Giá trị lớn nhất của biến

Bảng 3. Thống kê dữ liệu định tính [3]

Tên biến	Count**	Unique**	Top**	Freq**	Per**
id	48030	36027	39888	6	24.99%
label	48030	2	0	31426	65.43%
cat_1	22991	2	C1	12406	53.96%
cat_2	22991	2	P2	12512	54.42%
cat_3	22991	6	1	20979	91.25%
cat_4	6650	12	UI	4219	63.44%
cat_5	22991	1	1	22991	100.00%
cat_6	22991	3	1	22923	99.70%

sex	22991	2	MALE	13149	57.19%
location_id	22991	34	DN	13690	59.55%
mer_id	22991	10160	BO0001Z	715	3.11%
cat_7	22991	5	0	9440	41.06%
trans_currency	12895	41	VN	12089	93.75%
cat_8	22958	216	YN	5591	24.35%
cat_9	22991	5	I	15921	69.25%
cat_10	22991	10	2	11055	48.08%
cat_11	22991	5	0	16747	72.84%
com_type	21743	9	Vùng 1	11055	50.84%
cat_12	22300	8	D	4656	20.88%
social_sex_info	21058	2	male	12465	59.19%
social_location_id	21644	24	vi_VN	20829	96.23%
current_location_city	12232	827	Ho Chi Minh City	2421	19.79%
current_location_country	12232	47	Vietnam	11962	97.79%
hometown_location_city	11163	835	Hanoi	1198	10.73%
hometown_location_country	11163	35	Vietnam	10980	98.36%

**Count: Số lượng quan sát không bị trống của biến (thiếu dữ liệu)
**Unique: Số lượng giá trị đơn nhất (không lặp lại)
**Top: Giá trị thông dụng nhất (xuất hiện thường xuyên nhất)
**Freq: Tần suất xuất hiện của giá trị thông dụng nhất
**Per: Phần trăm của tần suất xuất hiện của giá trị thông dụng nhất trên số lượng quan sát không bị trống của biến

Bảng 4. Giả định loại dữ liệu của các biến chưa xác định [4]

Tên biến	Loại dữ liệu		Tên biến	Loại dữ liệu
unknown_var_1	Định lượng		unknown_var_11	Định tính
unknown_var_2	Định lượng		unknown_var_12	Định lượng
unknown_var_3	Định lượng		unknown_var_13	Định lượng
unknown_var_4	Định lượng		unknown_var_14	Định lượng
unknown_var_5	Định tính		unknown_var_15	Định lượng
unknown_var_6	Định tính		unknown_var_16	Định tính
unknown_var_7	Định lượng		unknown_var_17	Định lượng

unknown_var_8	Định lượng		unknown_var_18	Định tính
unknown_var_9	Định lượng		unknown_var_19	Định tính
unknown_var_10	Định lượng		unknown_var_20	Định tính

Bảng 5. Thống kê biến unknown định tính [5]

Tên biến	Count	Unique	Top	Freq
unknown_var_5	47805	23	1	7971
unknown_var_6	45728	5	1	41945
unknown_var_11	48030	4	1	32760
unknown_var_16	26433	2	1	26086
unknown_var_18	21405	14	4	4660
unknown_var_19	21405	11	0	10200
unknown_var_20	21405	15	4	4087

Bảng 6. Thống kê biến unknown định lượng [6]

Tên biến	Count	Mean	Std	Min	25%	50%	75%	Max
unknown_var_1	37902	8.23	9.75	1	3	7	11	810
unknown_var_2	36680	7.53	9.35	1	3	6	10	800
unknown_var_3	32189	5.65	8.68	1	2	5	7	791
unknown_var_4	20704	3.79	6.81	1	1	3	5	781
unknown_var_7	32907	20.75	40.18	-267	4	16	35	242
unknown_var_8	25091	30.91	36.08	0	5	16	46	326
unknown_var_9	32836	-1.80	32.22	-267	-6	2	10	233
unknown_var_10	32535	-32.18	63.52	-290	-69	0	1	233
unknown_var_12	22300	0.31	0.32	0	0	0	1	1
unknown_var_13	22991	0.42	0.28	0	0	0	1	1
unknown_var_14	22991	0.34	0.27	-1	0	0	1	1
unknown_var_15	22991	0.15	0.41	-1	0	0	0	2
unknown_var_17	18529	1877.64	1292.92	-31	717	1801	2906	8034

Bảng 7. Thống kê dữ liệu thời gian [7]

Tên biến	Count	Unique	Top	Freq	Per
time_1	22991	19415	2017-03-24T20:10:37.62Z	226	0.98%

time_2	22991	16891	2017-03-24T20:10:37.62Z	223	0.97%
Field_11	6679	800	10/31/2019	474	7.10%
date_1	14139	1490	8/31/2013	498	3.52%
date_2	22446	1475	12/28/2018	983	4.38%
date_3	20894	1656	1/1/2015	1589	7.61%
date_4	6171	608	10/31/2019	472	7.65%

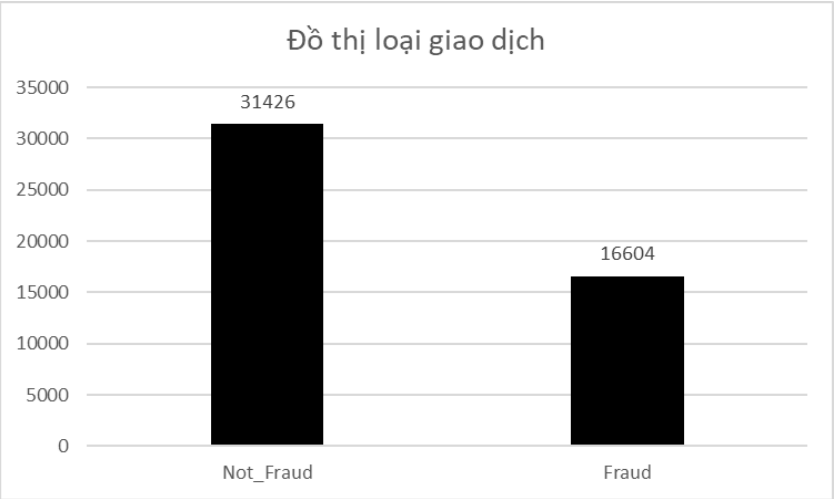
Bảng 8. Thống kê dữ liệu văn bản [8]

Tên biến	Count	Unique	Top	Freq	Per
mer_des	7570	5967	.	284	3.75%
address	22991	18915	TT Dịch vụ Việc Làm Tp.HCM	258	1.12%
mer_name	22991	12964	Trung Tâm Dịch Vụ Việc Làm Thành Phố Hồ Chí Minh	337	1.47%
trans_location	21217	11462	153 Xô Viết Nghệ Tĩnh Quận Bình Thạnh	337	1.59%
job	22991	1229	Doanh nghiệp có vốn đầu tư nước ngoài	5591	24.32%
job_detail	8271	3240	Công nhân	734	8.87%

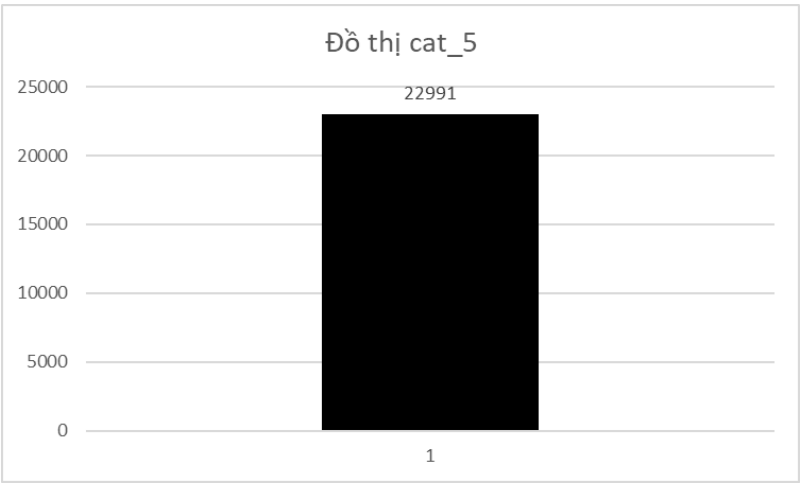
Bảng 9. Thông tin dữ liệu thiếu [9]

Tên biến	Số lượng dữ liệu bị thiếu	
id, label, unknown_var_11	0	0
unknown_var_5	225	0.47
unknown_var_6	2302	4.79
unknown_var_1	10128	21.09
unknown_var_2	11350	23.63
unknown_var_7	15123	31.49
unknown_var_9	15194	31.63
unknown_var_10	15495	32.26
unknown_var_3	15841	32.98

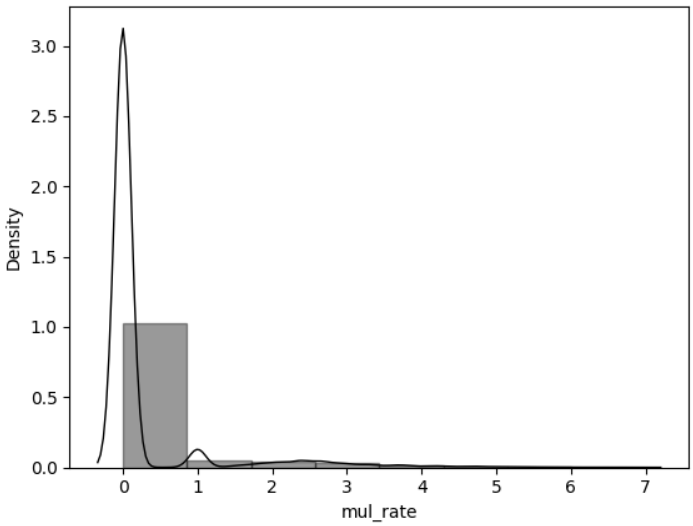
unknown_var_16	21597	44.97
unknown_var_8	22939	47.76
time_1, time_2, cat_1 - 3, cat_5 - 7, cat_9 ,cat_11, job, num_trans_last_month,cat_10, unknown_var_13 -15, mer_name, mer_id, address, location_id, , mul_rate, value, num_date_review, sex, review_value, dob	25039	52.13
cat_8	25072	52.2
date_2	25584	53.27
unknown_var_12, cat_12	25730	53.57
com_type	26287	54.73
social_subcriber_count, social_location_id, social_friend_count	26386	54.94
unknown_var_18 - 20	26625	55.43
trans_location	26813	55.83
social_sex_info	26972	56.16
date_3	27136	56.5
unknown_var_4	27326	56.89
unknown_var_17	29501	61.42
date_1	33891	70.56
trans_currency	35135	73.15
current_location_city, current_location_country	35798	74.53
hometown_location_city, hometown_location_country	36867	76.76
job_detail	39759	82.78
mer_des	40460	84.24
Field_11	41351	86.09
cat_4	41380	86.15
date_4	41859	87.15



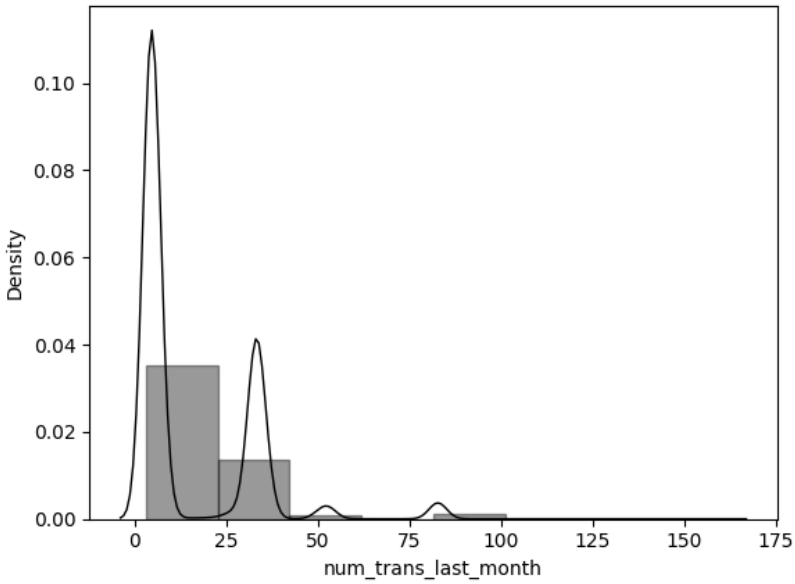
Hình 1. Đồ thị loại giao dịch [I]



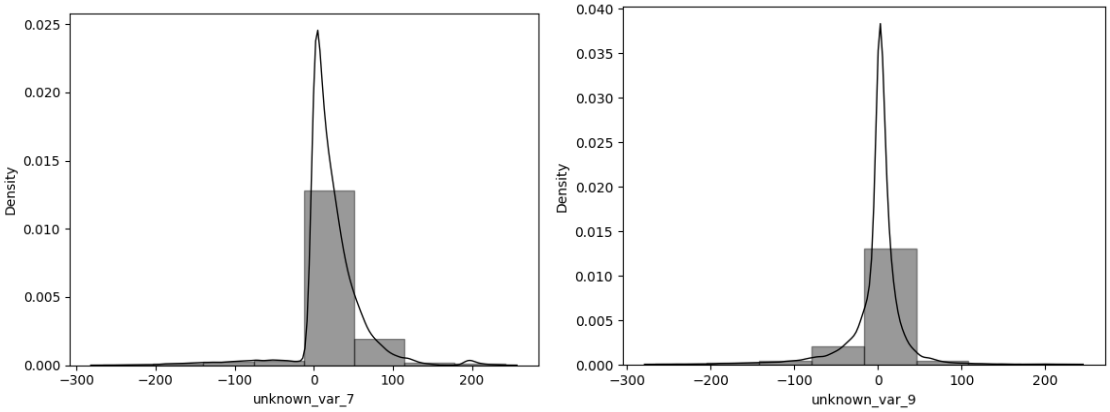
Hình 2. Đồ thị biến cat_5 [II]



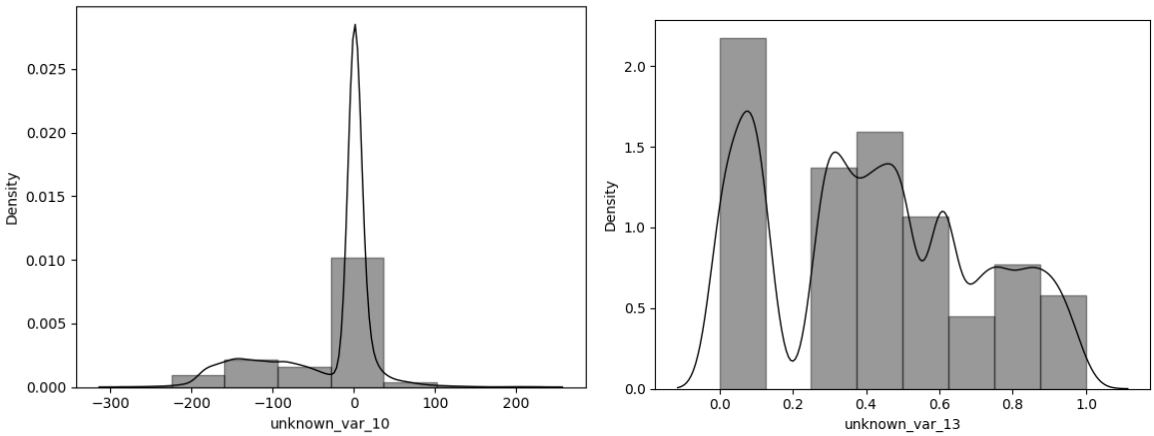
Hình 3. Đồ thị biến hệ số nhân [III]

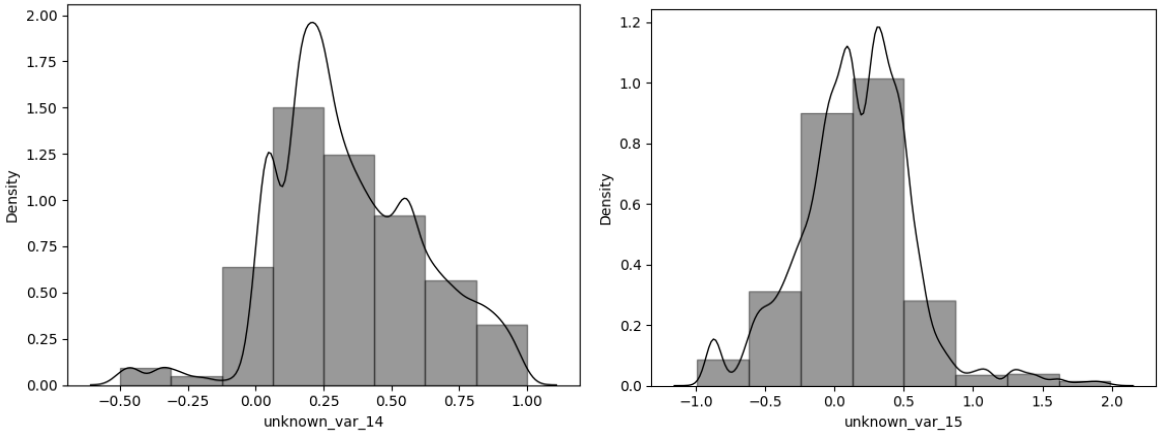


Hình 4. Đồ thị biến số lượng giao dịch của các khách hàng trong tháng trước [IV]

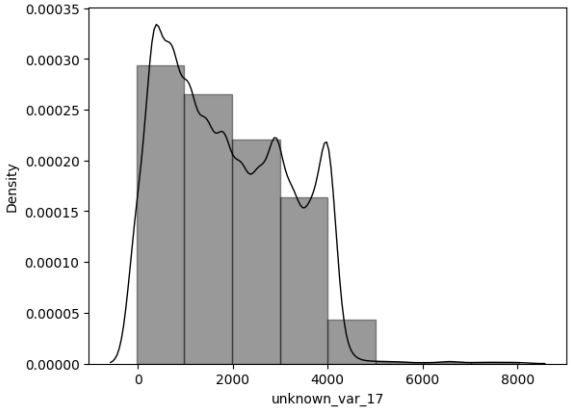


Hình 5. Đồ thị biến unknown_var_7 và unknown_var_9 [V]

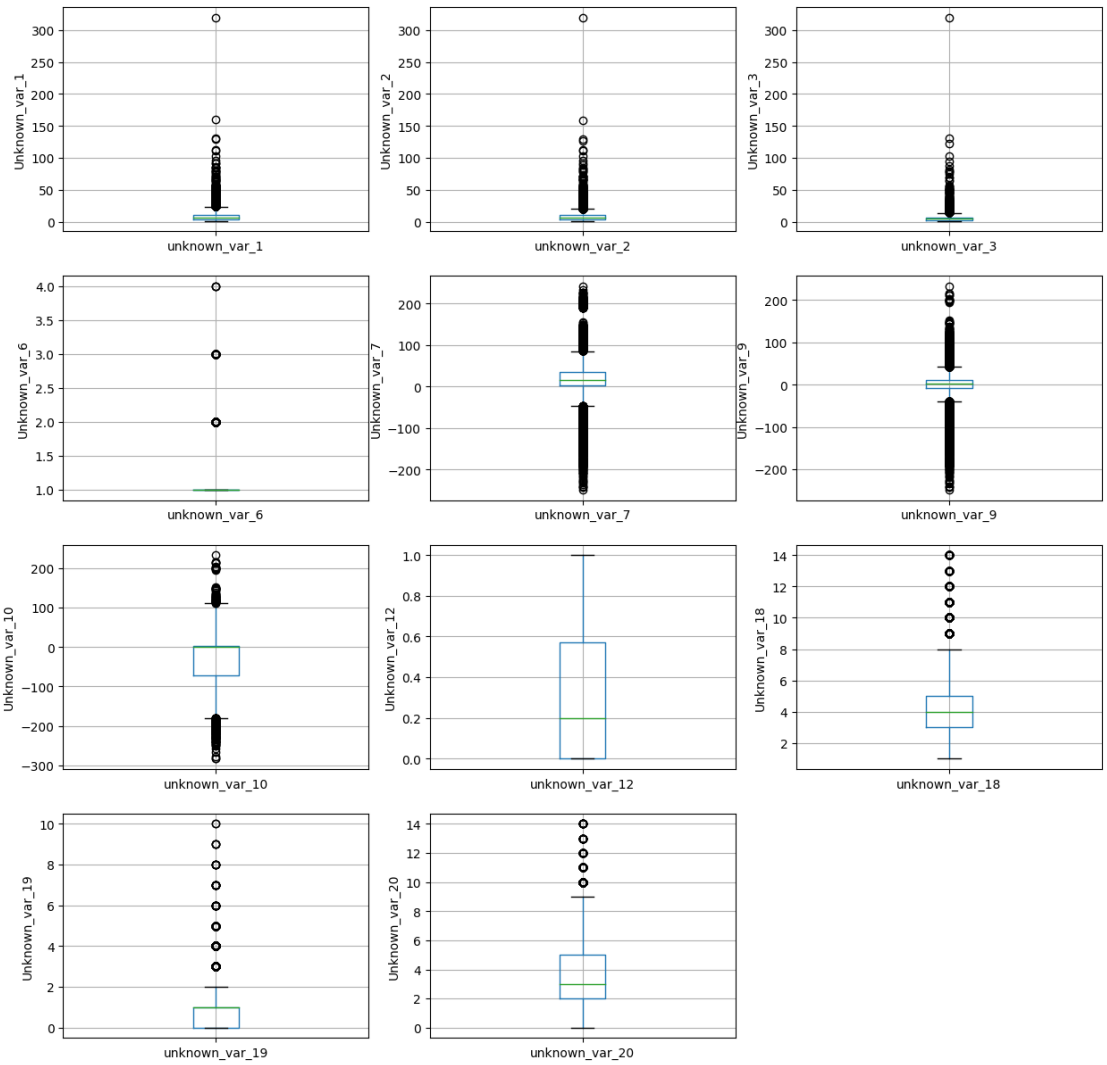




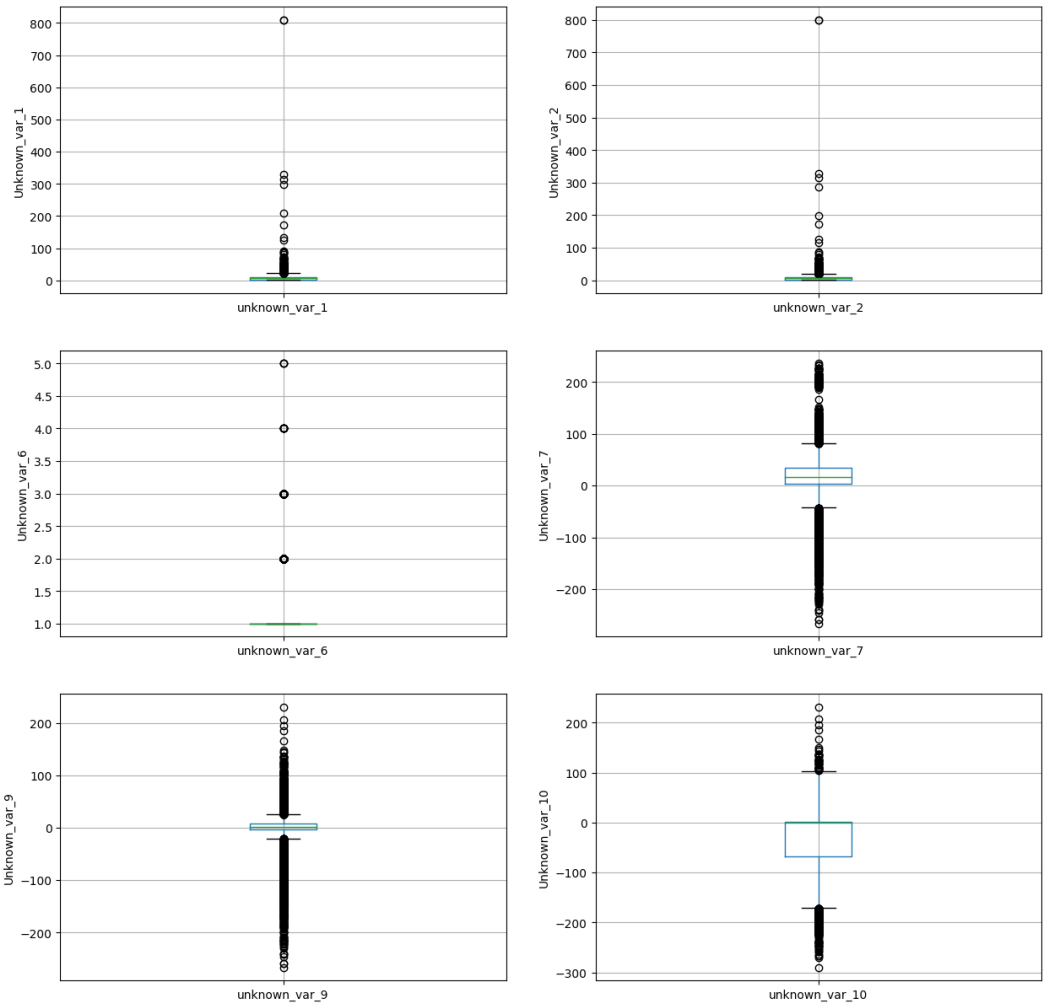
Hình 6. Đồ thị biến unknown_var_10, unknown_var_13, unknown_var_14, unknown_var_15 [VI]



Hình 7. Đồ thị biến unknown_var_17 [VII]



Hình 8. Đồ thị boxplot của phần dữ liệu bao gồm các quan sát có đủ giá trị time_1 [VIII]



Hình 9. Đồ thị boxplot của phần dữ liệu bao gồm các quan sát thiếu giá trị time_1 [IX]

Bảng 10. Số lượng các giá trị bị thiếu của các biến có đủ giá trị time_1 [10]

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
review_value	0	0.00
date_2	545	2.37
date_3	2097	9.12
date_4	16820	73.16
dob	0	0.00
sex	0	0.00
address	0	0.00
location_id	0	0.00
mer_id	0	0.00
mer_name	0	0.00
cat_7	0	0.00
trans_location	1774	7.72
trans_currency	10096	43.91
cat_8	33	0.14

job	0	0.00
num_trans_last_month	0	0.00
cat_9	0	0.00
cat_10	0	0.00
cat_11	0	0.00
com_type	1248	5.43
cat_12	691	3.01
job_detail	14720	64.03
unknown_var_1	2819	12.26
unknown_var_2	3493	15.19
unknown_var_3	5962	25.93
unknown_var_4	12215	53.13
unknown_var_5	59	0.26
unknown_var_6	610	2.65
unknown_var_7	7534	32.77
unknown_var_8	11634	50.60
unknown_var_9	7561	32.89
unknown_var_10	7656	33.30
unknown_var_11	0	0.00
unknown_var_12	691	3.01
unknown_var_13	0	0.00
unknown_var_14	0	0.00
unknown_var_15	0	0.00
unknown_var_16	9469	41.19
unknown_var_17	11184	48.65
unknown_var_18	3537	15.38
unknown_var_19	3537	15.38
unknown_var_20	3537	15.38
social_friend_count	11263	48.99
social_sex_info	11610	50.50
social_subcriber_count	11263	48.99

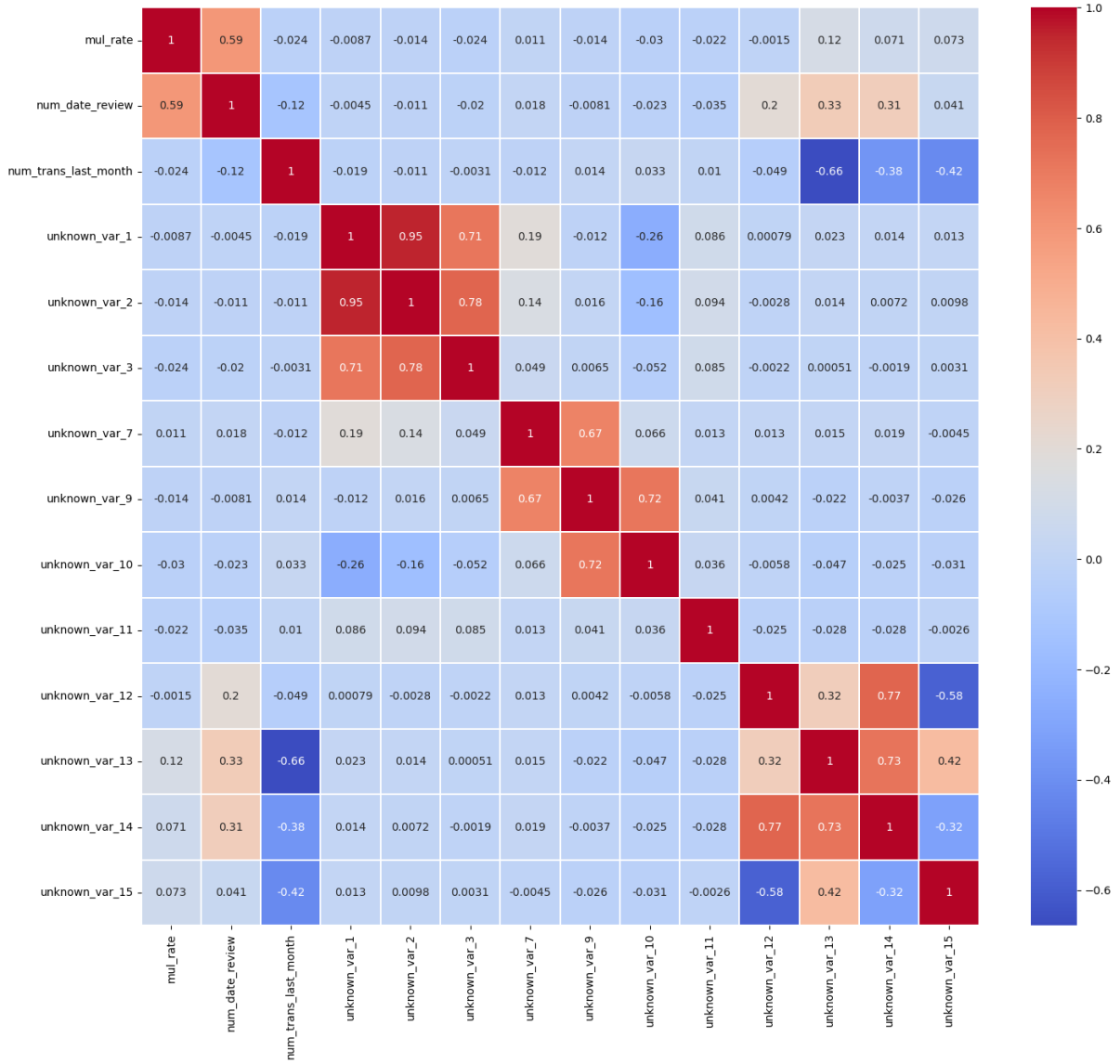
social_location_id	11263	48.99
current_location_city	16278	70.80
current_location_country	16278	70.80
hometown_location_city	16850	73.29
hometown_location_country	16850	73.29

Bảng 11. Số lượng các giá trị bị thiếu của các biến bị thiếu giá trị time_1 [11]

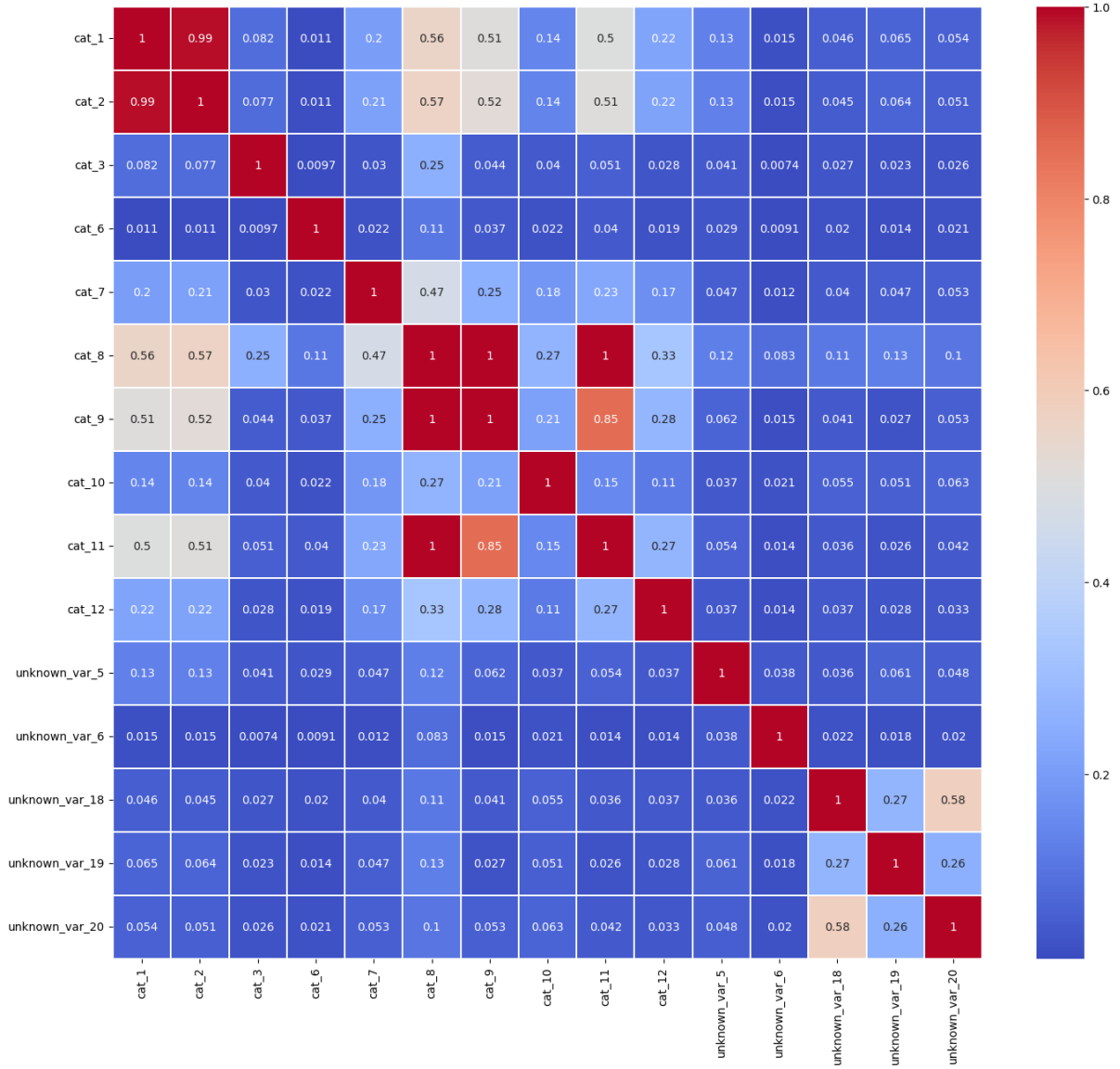
Tên biến	Số lượng dữ liệu bị thiếu	Phần trăm dữ liệu bị thiếu (%)
id	0	0.00
label	0	0.00
time_1	25039	100.00
time_2	25039	100.00
Field_11	25039	100.00
cat_1	25039	100.00
cat_2	25039	100.00
cat_3	25039	100.00
cat_4	25039	100.00
cat_5	25039	100.00
date_1	25039	100.00
mer_des	25039	100.00
mul_rate	25039	100.00
value	25039	100.00
cat_6	25039	100.00
num_date_review	25039	100.00
review_value	25039	100.00
date_2	25039	100.00
date_3	25039	100.00
date_4	25039	100.00
dob	25039	100.00
sex	25039	100.00
address	25039	100.00
location_id	25039	100.00
mer_id	25039	100.00
mer_name	25039	100.00
cat_7	25039	100.00
trans_location	25039	100.00
trans_currency	25039	100.00

cat_8	25039	100.00
job	25039	100.00
num_trans_last_month	25039	100.00
cat_9	25039	100.00
cat_10	25039	100.00
cat_11	25039	100.00
com_type	25039	100.00
cat_12	25039	100.00
job_detail	25039	100.00
unknown_var_1	7309	29.19
unknown_var_2	7857	31.38
unknown_var_3	9879	39.45
unknown_var_4	15111	60.35
unknown_var_5	166	0.66
unknown_var_6	1692	6.76
unknown_var_7	7589	30.31
unknown_var_8	11305	45.15
unknown_var_9	7633	30.48
unknown_var_10	7839	31.31
unknown_var_11	0	0.00
unknown_var_12	25039	100.00
unknown_var_13	25039	100.00
unknown_var_14	25039	100.00
unknown_var_15	25039	100.00
unknown_var_16	12128	48.44
unknown_var_17	18317	73.15
unknown_var_18	23088	92.21
unknown_var_19	23088	92.21
unknown_var_20	23088	92.21
social_friend_count	15123	60.40
social_sex_info	15362	61.35

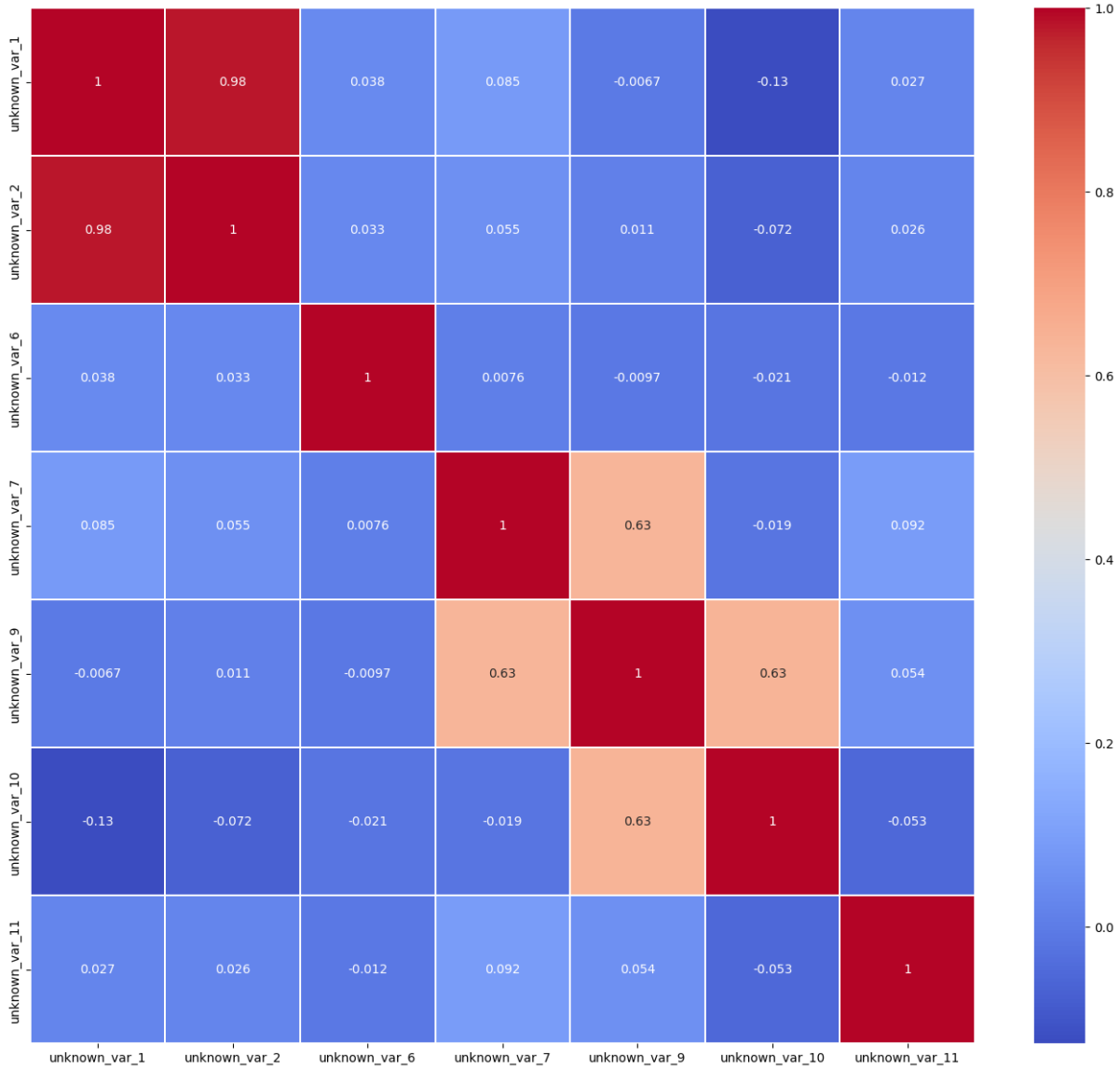
social_subcriber_count	15123	60.40
social_location_id	15123	60.40
current_location_city	19520	77.96
current_location_country	19520	77.96
hometown_location_city	20017	79.94
hometown_location_country	20017	79.94



Hình 10. Đồ thị hệ số tương quan của các biến số [X]



Hình 11. Đồ thị hệ số tương quan của các biến phân loại [XI]



Hình 12. Đồ thị hệ số tương quan của các biến số của dữ liệu thiếu time_1 [XII]

Nguồn tham khảo

[1] ITNavi, “Tổng quan về Neural Network(mạng Nơ Ron nhân tạo) là gì?,” *itnavi.com*, 2021. https://itnavi.com.vn/blog/neural-network-la-gi.

[2] Amazon, “Mạng nơ-ron là gì?,” *aws.amazon.com*. https://aws.amazon.com/vi/what-is/neural-network/.

[3] N. G. Ngh and N. H. Vi, “Nghiên c ứ u m ộ t s ố mô hình d ự báo d ị ch t ả d ự a trên khai phá d ữ li ệ u và phân tích không gian ứ ng d ụ ng công ngh ệ gis,” 2018.

[4] M. Pham, “Máy vector hỗ trợ (Support Vector Machine - SVM) là gì?,” *vietnambiz*, 2020. https://vietnambiz.vn/may-vector-ho-tro-support-vector-machine-svm-la-gi-20200226223210903.htm.

[5] Javapoint, “Support Vector Machine Algorithm,” *javapoint.com*. https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm.

[6] MLCB, “Kernel Support Vector Machine,” *machinelearningcoban.com*. https://machinelearningcoban.com/2017/04/22/kernelsmv/.