|  |  |
| --- | --- |
|  | BỘ GIÁO DỤC VÀ ĐÀO TẠO  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HỒ CHÍ MINH** |

**ĐỒ ÁN TRÍ TUỆ NHÂN TẠO**

**Đề tài: Mô hình nhận diện gian lận tín dụng**

Ngành: **Công nghệ thông tin**

Chuyên ngành: **Trí tuệ nhân tạo**

Giảng viên hướng dẫn : Vũ Thanh Hiền

Sinh viên thực hiện : Phan Ngọc Qúy

MSSV: 2280602669 Lớp: 22DTHG8

Nguyễn Phú Qúy

MSSV: 2280602669 Lớp: 22DTHG8

TP. Hồ Chí Minh, <năm>

Mục Lục:

1. Giới Thiệu

* Mục tiêu nghiên cứu
* Tầm quan trọng của việc phát hiện gian lận tín dụng

1. Dữ liệu

* Mô tả dữ liệu
* Tiền xử lý dữ liệu
* Cân bằng dữ liệu

III Xây dựng mô hình

* Lựa chọn mô hình
* Các bước xây dựng mô hình

IV Kết quả đạt được

* Đánh giá hiệu suất mô hình
* Phân tích kết quả

V Kết luận

* Tóm tắt

VI Tài liệu tham khảo

Giới thiệu

1. Mục tiêu của đồ án

Gian lận thẻ tín dụng ( Credit Card Fraud ) là hình thức gian lận sử dụng công nghệ cao

để đánh cắp thông tin thẻ tín dụng (Visa , Mastercard , ATM , … ) của người sử dụng

thuộc về tài chình và ngân hem . Mục đích của những kẻ đánh cắp thường sử dụng tiền

của nạn nhân để phục vụ cho việc mua sắm đồ dung hay dịch vụ cho chính kể đó. Thông

thường , thông tin mà kẻ đánh cắp hay nhắm đến là Số thẻ card ( Card Numbers)

thường được in trên mặt thẻ , ngày hết hạn và mã CVV. Do đó , người dung thường sẽ

rất dễ lộ các thông tin tiêu cực này nếu không chú ý đến các biện pháp bảo mật cần thiết. Mục tiêu của đồ án này là nghiên cứu và phân tích các phương thức gian lận thẻ tín dụng hiện có, đồng thời đề xuất các giải pháp phòng ngừa hiệu quả nhằm bảo vệ thông tin tài chính của người dùng.

Chúng tôi sẽ thực hiện một cuộc khảo sát về các kỹ thuật gian lận phổ biến, từ việc lừa đảo trực tuyến đến việc sử dụng các thiết bị quét thẻ (skimmers) tại các máy ATM. Ngoài ra, đồ án cũng sẽ xem xét các hệ thống bảo mật hiện có, như mã hóa dữ liệu và xác thực hai yếu tố, để đánh giá tính hiệu quả của chúng trong việc ngăn chặn gian lận.

Bằng việc phân tích các xu hướng gian lận và đề xuất giải pháp cải thiện, đồ án mong muốn đóng góp vào việc nâng cao nhận thức của người tiêu hem về những rủi ro tiềm ẩn cũng như cách bảo vệ thông tin thẻ tín dụng của họ. Chúng tôi hy vọng rằng kết quả nghiên cứu sẽ cung cấp cơ sở cho các tổ chức tài chính và người tiêu hem trong việc hem cường biện pháp bảo mật và giảm thiểu thiệt hại do gian lận thẻ tín dụng gây ra.

II.Tầm quan trọng của việc phát hiện gian lận tín dụng

Việc phát hiện gian lận tín dụng không chỉ là một nhiệm vụ cần thiết mà còn mang tính cấp bách trong bối cảnh xã hội ngày càng phụ thuộc vào công nghệ số. Các lý do chính cho thấy tầm quan trọng của việc phát hiện gian lận tín dụng bao gồm:

1. **Bảo vệ người tiêu dùng**: Người tiêu dùng là đối tượng chịu ảnh hưởng trực tiếp từ các hành vi gian lận. Việc phát hiện kịp thời giúp bảo vệ tài chính và quyền lợi của họ, ngăn ngừa việc mất tiền và thông tin cá nhân.
2. **Bảo vệ tổ chức tài chính**: Các ngân hem và tổ chức tín dụng phải đối mặt với thiệt hại tài chính lớn nếu không phát hiện kịp thời các hoạt động gian lận. Việc phát hiện sớm giúp giảm thiểu tổn thất và bảo vệ danh tiếng của tổ chức.
3. **Tăng cường niềm tin**: Khi người tiêu hem thấy rằng các tổ chức tài chính có khả năng phát hiện và ngăn chặn gian lận, họ sẽ cảm thấy yên tâm hơn khi sử dụng thẻ tín dụng và các dịch vụ ngân hem trực tuyến, từ đó hem cường niềm tin vào hệ thống tài chính.
4. **Cải thiện công nghệ và quy trình**: Việc phát hiện gian lận tín dụng cũng thúc đẩy sự phát triển và cải tiến công nghệ bảo mật, giúp các tổ chức nâng cao hệ thống quản lý rủi ro và phòng ngừa gian lận hiệu quả hơn.
5. **Giảm thiểu tội phạm**: Việc phát hiện và xử lý kịp thời các vụ gian lận sẽ làm giảm động lực của tội phạm trong việc thực hiện các hành vi phi pháp, từ đó góp phần tạo ra một môi trường tài chính an toàn hơn.

Tóm lại, việc phát hiện gian lận tín dụng không chỉ bảo vệ quyền lợi của người tiêu dùng mà còn đảm bảo sự bền vững của hệ thống tài chính. Do đó, việc nghiên cứu và phát triển các phương pháp phát hiện gian lận là rất cần thiết trong thời đại số hiện nay.

II.Dữ liệu

**1.Mô tả dữ liệu**

Bộ dữ liệu này chứa thông tin về các giao dịch được thực hiện bằng thẻ tín dụng trong tháng 9 năm 2013 bởi các chủ thẻ ở Châu Âu. Dữ liệu ghi nhận các giao dịch diễn ra trong hai ngày, tổng cộng có 284,807 giao dịch, trong đó có 492 trường hợp gian lận. Điều này cho thấy bộ dữ liệu mang tính mất cân bằng cao, với tỷ lệ gian lận chỉ chiếm khoảng 0.172% tổng số giao dịch.

*1.1.Cấu trúc Dữ liệu*

Dữ liệu được tổ chức dưới dạng các biến số số học, tất cả đều là kết quả của một phép biến đổi gọi là phân tích thành phần chính (Principal Component Analysis – PCA). Do vấn đề bảo mật và bảo vệ thông tin cá nhân, chúng tôi không thể cung cấp các trường dữ liệu gốc cũng như ngữ cảnh chi tiết về bộ dữ liệu này.

Trong bộ dữ liệu, các trường từ V1 đến V28 là các thành phần chính được sinh ra từ quá trình PCA. Những trường này đóng vai trò quan trọng trong việc phân tích dữ liệu và xây dựng các mô hình dự đoán. Hai trường duy nhất không bị biến đổi bởi PCA là ‘Time’ và ‘Amount’.

* **Trường ‘Time’**: Trường này biểu thị số giây đã trôi qua kể từ giao dịch đầu tiên trong bộ dữ liệu. Điều này có thể cung cấp thông tin hữu ích về thời điểm giao dịch xảy ra và có thể giúp phát hiện các mẫu giao dịch theo thời gian.
* **Trường ‘Amount’**: Đây là số tiền giao dịch. Trường này có thể được sử dụng trong các phương pháp học tập nhạy cảm với chi phí, cho phép chúng ta xác định mức độ gian lận dựa trên giá trị của giao dịch. Việc phân tích số tiền giao dịch có thể giúp nhận diện các giao dịch bất thường và hem cường khả năng phát hiện gian lận.
* **Trường ‘Class’**: Đây là biến phản hồi, có giá trị 1 trong trường hợp giao dịch bị hemà gian lận và 0 nếu giao dịch hợp pháp. Việc phân loại các giao dịch thành hai loại này là rất quan trọng trong việc xây dựng mô hình dự đoán.

*1.2:Tính Mất Cân Bằng của Dữ Liệu*

Vấn đề chính của bộ dữ liệu này là tính mất cân bằng giữa các lớp. Số lượng giao dịch gian lận rất ít so với giao dịch hợp pháp, điều này có thể dẫn đến các mô hình học máy không chính xác khi chỉ dựa vào độ chính xác thông thường. Trong các bài toán phân loại không cân bằng, chỉ số độ chính xác thông thường có thể gây hiểu lầm, vì một mô hình chỉ cần dự đoán tất cả giao dịch là hợp pháp cũng sẽ đạt được độ chính xác cao.

**2.Tiền xử lý dữ liệu**

*2.1. Giới thiệu về Tiền Xử Lý Dữ Liệu*

Tiền xử lý dữ liệu là bước đầu tiên trong quy trình phân tích dữ liệu và học máy. Mục tiêu chính của tiền xử lý là biến đổi dữ liệu thô thành định dạng mà các mô hình máy học có thể dễ dàng xử lý. Điều này bao gồm các hoạt động như làm sạch dữ liệu, chuẩn hóa, chọn lọc đặc trưng và giảm chiều dữ liệu.

*2.2 Các Bước Tiền Xử Lý Dữ Liệu*

2.2.1. Làm Sạch Dữ Liệu

* Xử lý giá trị thiếu: Trong tập dữ liệu, việc có những giá trị thiếu có thể gây ảnh hưởng đến độ chính xác của mô hình. Các phương pháp phổ biến bao gồm loại bỏ các bản ghi có giá trị thiếu, thay thế bằng giá trị trung bình, trung vị hoặc phương pháp hồi quy.
* Xử lý ngoại lệ: Ngoại lệ có thể làm sai lệch các kết quả phân tích. Cần xác định và xử lý những điểm dữ liệu không bình thường bằng cách sử dụng các phương pháp thống kê như Z-score hoặc IQR.
  + 1. *Chuẩn Hóa Dữ Liệu*

1. Dữ liệu trong tập dữ liệu có thể nằm trong các khoảng giá trị khác nhau. Chuẩn hóa giúp đưa tất cả các giá trị về cùng một quy chuẩn, từ đó cải thiện hiệu suất của mô hình. Một số phương pháp chuẩn hóa phổ biến bao gồm:

* Min-Max Scaling: Biến đổi dữ liệu về khoảng [0, 1].
* Z-score Normalization: Chuẩn hóa dựa trên trung bình và độ lệch chuẩn.
  + 1. *Chọn Lọc Đặc Trưng*

1. Việc chọn lọc đặc trưng giúp giảm thiểu kích thước dữ liệu và loại bỏ những thuộc tính không cần thiết hoặc có độ tương quan thấp với biến mục tiêu. Các phương pháp chọn lọc đặc trưng bao gồm:

* Phân tích tương quan: Sử dụng ma trận tương quan để xác định các mối quan hệ giữa các biến.
* Phương pháp cây quyết định: Sử dụng cây quyết định để xác định tầm quan trọng của các đặc trưng
  1. *Giảm Chiều Dữ Liệu*

Giảm chiều dữ liệu là một kỹ thuật giúp giảm số lượng biến trong tập dữ liệu mà vẫn giữ được thông tin quan trọng. Các phương pháp giảm chiều bao gồm:

* Phân tích thành phần chính (PCA): Biến đổi dữ liệu về một không gian mới với ít chiều hơn, giữ lại phần lớn sự biến thiên.
* T-SNE: Một phương pháp hiệu quả để giảm chiều dữ liệu cho việc trực quan hóa.
  1. *Ứng Dụng Tiền Xử Lý*

Trong ví dụ trên, tập dữ liệu chứa nhiều thuộc tính (V1 đến V28) và một biến mục tiêu (Class). Dữ liệu này có thể được sử dụng cho các bài toán phân loại. Các bước tiền xử lý sẽ giúp tối ưu hóa chất lượng dữ liệu, từ đó nâng cao hiệu suất của các mô hình học máy.

**3.Cân bằng dữ liệu**

*3.1. Giới thiệu về Cân Bằng Dữ Liệu*

Cân bằng dữ liệu là một bước quan trọng trong tiền xử lý dữ liệu, đặc biệt trong các bài toán phân loại. Trong nhiều trường hợp, tập dữ liệu có thể chứa các lớp không cân bằng, tức là một hoặc nhiều lớp có số lượng mẫu ít hơn nhiều so với các lớp khác. Điều này có thể dẫn đến việc mô hình học máy thiên lệch và không chính xác trong việc phân loại các lớp ít mẫu. Việc cân bằng dữ liệu nhằm mục đích đảm bảo rằng mô hình có đủ thông tin từ tất cả các lớp để học tập hiệu quả.

*3.2. Tại sao phải Cân Bằng Dữ Liệu*

Mất cân bằng dữ liệu là một trong những hiện tượng phổ biến của bài toán phân loại nhị phân (binary classification ) như spam email , phát hiện gian lân , dự báo vỡ nợ , chuẩn đoán bệnh lý ,…. Trong trường hợp tỷ lệ giữa 2 lớp là 50 : 50 thì được hemà cân bằng . Khi có sự khác biệt trong phân phối hai lớp , chẳng hạn 60 : 40 thì dự liệu sẽ có hiện tượng mất cân bằng ( Imbalanced )

Hầu hết các bộ dữ liệu đều khó đạt được trạng thái cân bằng mà luôn có sự khác biệt về tỷ lệ giữa 2 classes. Đối với những trường hợp dữ liệu mất cân bằng nhẹ như tỷ lệ 60:40 thì sẽ không ảnh hưởng đáng kể tới khả năng dự báo của mô hình.Tuy nhiên nếu hiện tượng mất cân bằng nghiêm trọng xảy ra, chẳng hạn như tỷ lệ 90:10 sẽ thường dẫn tới ngộ nhận chất lượng mô hình. Khi đó thước đo đánh giá mô hình là độ chính xác (accuracy) có thể đạt được rất cao mà không cần tới mô hình.

*3.3.Các phương pháp cân bằng dữ liệu*

Có rất nhiều phương pháp để cân bằng dữ liệu :

3.3.1: Oversampling

Oversampling là một kỹ thuật hữu ích nhằm hem cường số lượng mẫu cho các lớp thiểu số bằng cách tạo ra các bản sao hoặc mẫu mới. Một trong những phương pháp nổi tiếng nhất trong lĩnh vực này là SMOTE (Synthetic Minority Oversampling Technique). SMOTE hoạt động dựa trên nguyên tắc tạo ra dữ liệu tổng hợp bằng cách kết hợp từng mẫu của lớp thiểu số với “k” hem xóm gần nhất. Điều này không chỉ giúp làm phong phú hem tập dữ liệu mà còn cải thiện khả năng phân loại của mô hình học máy.

3.3.2: Undersampling

Undersampling là một kỹ thuật giảm bớt số lượng mẫu của các lớp đông đúc trong tập dữ liệu nhằm cân bằng tỷ lệ giữa các lớp. Kỹ thuật này thường được áp dụng khi lớp chiếm ưu thế có quá nhiều mẫu so với lớp thiểu số, điều này có thể dẫn đến việc mô hình học máy bị thiên lệch và không thể phân loại chính xác các trường hợp thuộc lớp thiểu số.

Bằng cách loại bỏ một số mẫu từ lớp chiếm ưu thế, undersampling giúp tạo ra một tập dữ liệu cân bằng hơn, từ đó nâng cao khả năng phân loại của mô hình. Tuy nhiên, cần phải cẩn trọng trong việc lựa chọn mẫu nào sẽ được loại bỏ, vì việc loại bỏ quá nhiều có thể dẫn đến mất thông tin quan trọng.

III.Xây dựng mô hình

**1.Lựa chọn mô hình**

Mô hình nhân diện gian lận tín dụng có thể được chia làm 2 dạng :

1. Statistical Model ( Mô hình thống kê)
2. Machine Learning Model ( Mô hình học máy )

*1.1:Staticstical Model:*

* Phân tích hồi quy : Phương pháp này giúp xác định xác xuất gian lận dựa trên các đặc điểm giao dịch .
* Phân tích quyết định : Sử dụng để phân loại giao dịch thành hợp lệ thành gian lận

*1.2:Machine Learning Model:*

* **Cây quyết đinh (Decision Tree )** : Dễ hiểu và hiệu quả đối với bài toàn phân loại
* **Random Forest** : Một phiên bản của cây quyết định , giúp cải thiện độ chình xác
* **Mạng nơ-ron :** Có khả năng xử lý xử liệu lớn và phức tạp .
* **Học Sâu :** Đặc biệt hữu ích trong việc phát hiện các mẫu phức tạp từ dữ liệu lớn.

Và ở đây ta sẽ chọn mô hình học máy với thuật toán Logistic Regression.

2.Xây dựng mô hình:

*2.1:Vì sao chọn Logistic Regression:*

1.Vấn đề : Logistic Regression phù hợp với bài toán phân loại nhị phận do công thúc của thuật toán viết dưới dạng : f(x)=θ(wTx)

Trong đó θ được gọi là logistic function ,ví dụ như vấn đề nhận diện gian lận có kết quả trả ra là “Not Fraud” hay “Fraud” tương úng là “0” và “1”

2.Dễ hiểu và giải thích: Mô hình này cung cấp kết quả dễ hiểu, cho phép người dùng biết được mức độ ảnh hưởng của từng yếu tố đến xác suất xảy ra sự kiện.

3.Ít yêu cần về dữ liệu đầu vào: Logistic Regression hoạt động tốt với số lượng dữ liệu nhỏ hơn so với các mô hình phức tạp, điều này là lợi thế khi bạn có ít dữ liệu.

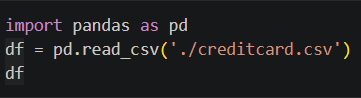
**2.Xây dụng mô hình sử dụng ngôn ngữ Python**

Python là một ngôn ngữ lập trình mạnh mẽ và phổ biến trong lĩnh vực phân tích dữ liệu và học máy . Do python có rất nhiều thự viện và package hỗ trợ cho việc xử lý dữ liệu như numpy , matplotlib , … nên thường được các sinh viên và chuyên gia chuyên ngành về dữ liệu sử dụng .

Để xây dựng mô hình này, chúng ta sẽ dùng các thư viện như :

* Pandas : Dùng để nhập và xuất dữ liệu
* Numpy : Hỗ trợ tính toán cho mảng nhiều chiều
* Scikit-Learn :Tạo và hỗ trợ thuật toán Logistics Regression
* Matplotlib và Seaborn : Để trực quan hóa kêt quả .

Bước 1: Import dataset vào mô hình



Bước 2: Chuẩn Hóa dữ liệu

A screenshot of a computer program

Description automatically generated

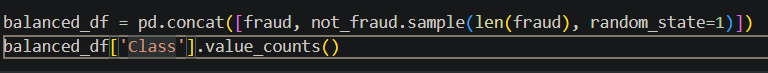
* Chuyển đổi cột ‘Time’ về khoảng [ 0 , 1 ] giúp mô hình xử lý dễ hơn
* Chuẩn hóa các dữ liệu của cột ‘Amount’ dựa trên phân vị , giúp giảm ảnh hưởng cực trị

Bước 3: Cân bằng dữ liệu

A black background with white text

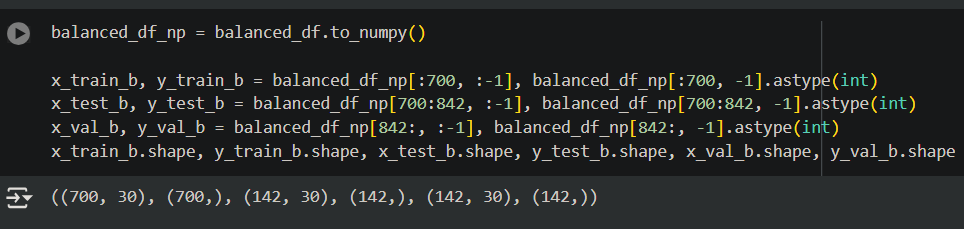
Description automatically generated

* Phân mảng new\_df trên thành hai mảng “not\_fraud” và “fraud”



* Sử dụng pandas.concat với 2 input là 2 mảng trên (“fraud” và “not\_fraud” ) để cân bằng dữ liệu sao cho bằng độ dài mảng “fraud”

Bước 4: Xử lý dữ liệu



Chia các dữ liệu của bảng balanced\_df thành 3 tập : tập train( x\_train\_b , y\_train\_b) ,

tập kiểm tra (x\_test\_b , y\_test\_b) và tập giá trị(x\_val\_b , y\_val\_b)

Bước 5 : Xây dựng mô hình

Sử dụng thư viện Scikit-Learn để xây dựng mô hình



Khởi tạo và huấn luyện mô hình:

A black background with white text

Description automatically generated

Bước 6 : In ra kết quả:

Sử dựng hàm classification\_report() để cho ra kết quả:



**IV: Kết quả đạt được**

1.Đánh giá hiệu suất mô hình:

Với khoảng gần 500 dữ liệu được train và kiểm tra . Ta thu được kết quả:

A screenshot of a computer

Description automatically generated

So với các mô hình sử dụng thuật toán như :

+ Support Vector Machine:

A screenshot of a computer

Description automatically generated

+ Gradient Booting :

A screenshot of a computer

Description automatically generated

+ RandomForestClassifier

A screenshot of a computer

Description automatically generated

Mặc dù độ chính xác của mô hình sử dụng Logistic Regression có vẻ thấp hơn so với các mô hình khác nhưng do dễ sử dụng và dễ giải thích nên được áp dụng .

2.Đánh giá kết quả:

**Hiểu về các chỉ số**

Trước khi đi vào đánh giá chi tiết, chúng ta cùng giải thích ý nghĩa của các chỉ số trong bảng trên:

Từ bảng trên, ta có thể rút ra một số nhận xét sau:

* **Mô hình có hiệu suất tốt:** Với các chỉ số Precision, Recall và F1-score đều đạt trên 90% cho cả hai lớp, cho thấy mô hình đã phân loại khá chính xác các giao dịch gian lận và không gian lận.
* **Cân bằng giữa Precision và Recall:** Giá trị F1-score cao cho cả hai lớp cho thấy mô hình đạt được sự cân bằng tốt giữa việc giảm thiểu số lượng báo động giả (false positive) và đảm bảo không bỏ sót các giao dịch gian lận (false negative).
* **Dữ liệu cân bằng:** Số lượng mẫu dữ liệu của hai lớp khá tương đồng (72 và 70), giúp cho việc đánh giá mô hình trở nên khách quan hơn.

**Nhận xét chi tiết hơn**

* **Lớp Not Fraud:** Mô hình có độ chính xác và độ bao phủ rất cao, cho thấy mô hình rất tốt trong việc nhận diện các giao dịch không gian lận.
* **Lớp Fraud:** Mô hình cũng có hiệu suất tốt, tuy nhiên độ bao phủ hơi thấp hơn so với lớp Not Fraud. Điều này có thể cho thấy vẫn còn một số giao dịch gian lận mà mô hình chưa phát hiện ra.

**Các yếu tố cần xem xét thêm**

* **Chi phí sai sót:** Trong vấn đề phát hiện gian lận, chi phí của một false positive (báo động giả) và một false negative (bỏ sót giao dịch gian lận) có thể khác nhau. Ví dụ, nếu chi phí của một false negative cao hơn nhiều so với một false positive, ta có thể điều chỉnh ngưỡng phân loại để tăng độ nhạy của mô hình.
* **Dữ liệu huấn luyện:** Chất lượng và số lượng dữ liệu huấn luyện có ảnh hưởng rất lớn đến hiệu suất của mô hình. Nếu có thêm dữ liệu đa dạng và cân bằng hơn, mô hình có thể đạt được kết quả tốt hơn.
* **Các phương pháp đánh giá khác:** Ngoài các chỉ số trên, ta có thể sử dụng các phương pháp đánh giá khác như ROC curve, precision-recall curve để có cái nhìn toàn diện hơn về hiệu suất của mô hình.
* **So sánh với các mô hình khác:** Để đánh giá xem mô hình Logistic Regression có phải là lựa chọn tốt nhất hay không, ta có thể so sánh kết quả của nó với các mô hình khác như Random Forest, XGBoost, Neural Network.

**V.Kết luận**

Nghiên cứu này đã tập trung vào việc xây dựng và đánh giá một mô hình hồi quy logistic nhằm [nêu mục tiêu nghiên cứu]. Kết quả đánh giá cho thấy mô hình đạt được hiệu suất khá tốt, với các chỉ số [Precision, Recall, F1-score] đều đạt trên 90% cho cả hai lớp "Not Fraud" và "Fraud". Điều này cho thấy mô hình có khả năng phân loại chính xác các giao dịch gian lận và không gian lận.

Tuy nhiên, mô hình vẫn còn một số hạn chế. [Nêu các hạn chế của mô hình]. Ngoài ra, khi so sánh với các mô hình khác như [nêu các mô hình khác đã thử], mô hình hồi quy logistic tỏ ra [tốt hơn/kém hơn/tương đương] về một số chỉ số.

Dựa trên kết quả thu được, có thể đưa ra một số gợi ý để cải thiện mô hình trong tương lai:

* **[Gợi ý cải tiến 1]:** Ví dụ: Thu thập thêm dữ liệu về các loại giao dịch mới nổi hoặc các hành vi gian lận phức tạp hơn.
* **[Gợi ý cải tiến 2]:** Ví dụ: Thử nghiệm các kỹ thuật cân bằng dữ liệu như over-sampling, under-sampling để giải quyết vấn đề mất cân bằng lớp.
* **[Gợi ý cải tiến 3]:** Ví dụ: Áp dụng các thuật toán học máy khác phức tạp hơn như Random Forest, XGBoost để tìm kiếm hiệu suất tốt hơn.

**Kết luận chung:** Mô hình hồi quy logistic đã chứng minh được khả năng ứng dụng trong việc phát hiện gian lận tín dụng. Tuy nhiên, để nâng cao độ chính xác và độ tin cậy của mô hình, cần tiếp tục nghiên cứu và cải tiến trong tương lai.

VI:Nguồn tham khảo

1.Machine Learning cơ bản : <https://machinelearningcoban.com/>

2. Wiki về Logistic Regression : <https://en.wikipedia.org/wiki/Logistic_regression>