PHÁT HIỆN GIAN LẬN TÍN DỤNG VỚI KỸ THUẬT HỌC MÁY, THUẬT TOÁN RANDOM FOREST

ThS. NGUYỄN DƯƠNG HÙNG
Khoa Hệ thống thông tin quản lý, Học viện Ngân hàng
TS. NGUYỄN HỮU XUÂN TRƯỜNG
Bộ môn Toán Kinh tế, Học viện Chính sách và Phát triển

Học máy là một lĩnh vực được các doanh nghiệp và các tổ chức nghiên cứu và ứng dụng, đặc biệt là trong lĩnh vực tài chính, ngân hàng. Từ các trợ lý ảo như Siri và Cortana, đến các Chatbots được tạo ra bởi Facebook và Google, trí tuệ nhân tạo đang ngày càng tác động mạnh mẽ đến các lĩnh vực kinh tế - xã hội, trong đó có lĩnh vực ngân hàng. Ngành Ngân hàng với việc phát hiện gian lận tín dụng là một ví dụ cụ thể. Hệ thống phát hiện gian lận tín dụng được áp dụng vào hệ thống ngân hàng từ những năm 2000. Tuy nhiên, những hệ thống này cho tới nay vẫn còn những hạn chế cần được bổ sung, chỉnh sửa để đáp ứng được yêu cầu quản trị rủi ro tín dụng trong điều kiện nền kinh tế hội nhập hiện nay và cần tiến xa hơn nữa cho tương lai. Trong bài viết này, chúng tôi sử dụng thuật toán Rừng Ngẫu nhiên (Random forest) là một thuật toán trong lớp các thuật toán của trí tuệ nhân tạo để phát hiện gian lận tín dụng trong các ngân hàng thương mại.

Từ khóa: Trí tuệ nhân tạo, học máy, khai phá dữ liệu, gian lận, tín dụng

1. Giới thiệu

Gian lận thẻ tín dụng là hình thức tội phạm sử dụng công nghệ cao để có được thông tin hoặc thẻ tín dụng của người sở hữu nhằm thực hiện các hành vi bất hợp pháp. Các hình thức gian lận thẻ tín dụng bao gồm:

- Tội phạm sử dụng thẻ lấy cắp của nạn nhân để thanh toán nhằm đánh cắp tiền trong tài khoản của nạn nhân hoặc rút tiền mặt tại các máy ATM.

- Sử dụng công nghệ cao như phần mềm độc hại, lừa đảo để đánh cắp thông tin thẻ của nạn nhân, từ đó có thể thực hiện các hành vi bất hợp pháp như làm giả thẻ, thanh toán mạo danh…

Tình trạng gian lận thẻ tín dụng diễn ra rất phức tạp trên thế giới nói chung và ở Việt Nam nói riêng. Hằng năm, tội phạm thẻ tín dụng thực hiện hàng trăm vụ gian lận và gây thiệt hại hàng triệu USD cho Việt Nam. Vì vậy, các đơn vị quản lý nhà nước và các ngân hàng thương mại, tổ chức tín dụng đã thực hiện nhiều biện pháp phòng, tránh. Các nhóm biện pháp này chủ yếu tập trung ở một số khía cạnh như sau:

- *Đối với các tổ chức phát hành thẻ và đơn vị chấp nhận thẻ:* Ban hành các quy tắc phòng chống gian lận thẻ tín dụng; phổ biến cụ thể các hình thức gian lận thẻ tín dụng cho cán bộ, nhân viên; kiểm tra, giám sát các thiết bị thanh toán chấp nhận thẻ; áp dụng các loại thẻ thông minh an toàn...

- *Đối với khách hàng:* Hướng dẫn cách thức sử dụng và bảo quản thẻ an toàn; yêu cầu sử dụng các biện pháp phòng tránh như OTP, tin báo biến động số dư...

Tuy nhiên, với thực tế về tội phạm công nghệ trong lĩnh vực tài chính, ngân hàng ngày càng phức tạp hiện nay, hoạt động phòng chống gian lận thẻ tín dụng cần phải áp dụng những giải pháp thông minh (trí tuệ nhân tạo và học máy), có khả năng giám sát và phát hiện gian lận 24/7.

Tính đến nay, đã có một số công trình nghiên cứu đề xuất phát hiện gian lận thẻ tín dụng theo hướng tiếp cận trí tuệ nhân tạo và học máy, thậm chí còn có những giải pháp đã được triển khai trong thực tế như Predator - Vũ khí siêu hình, giải pháp tự trị của GBG (công ty công nghệ hàng đầu thế giới về phòng, chống gian lận).

Trong bài viết này, chúng tôi phát triển một mô hình phát hiện gian lận thẻ tín dụng dựa trên thuật toán học máy Random forest - một trong những kỹ thuật được đánh giá là phù hợp cho lớp bài toán phát hiện gian lận. So sánh với một số giải pháp đã đề xuất, mô hình của chúng tôi có độ chính xác cao hơn khi thực nghiệm trên bộ dữ liệu thật ccFraud.csv gồm 1 triệu đối tượng khách hàng được thu thập từ các ngân hàng được các nhà khoa

học cung cấp tại: https://packages.revolutionanalytics.com/datasets/. Phần tiếp theo trình bày khái quát về kỹ thuật học máy (một lĩnh vực của trí tuệ nhân tạo) Random forest; sau đó mô hình phát hiện gian lận thẻ tín dụng dựa trên thuật toán học máy Random forest sẽ được xây dựng và đánh giá trong phần ba; cuối cùng những vấn đề đã đạt được trong nghiên cứu này cũng như công việc dự định trong tương lai sẽ được trình bày ở phần kết luận.

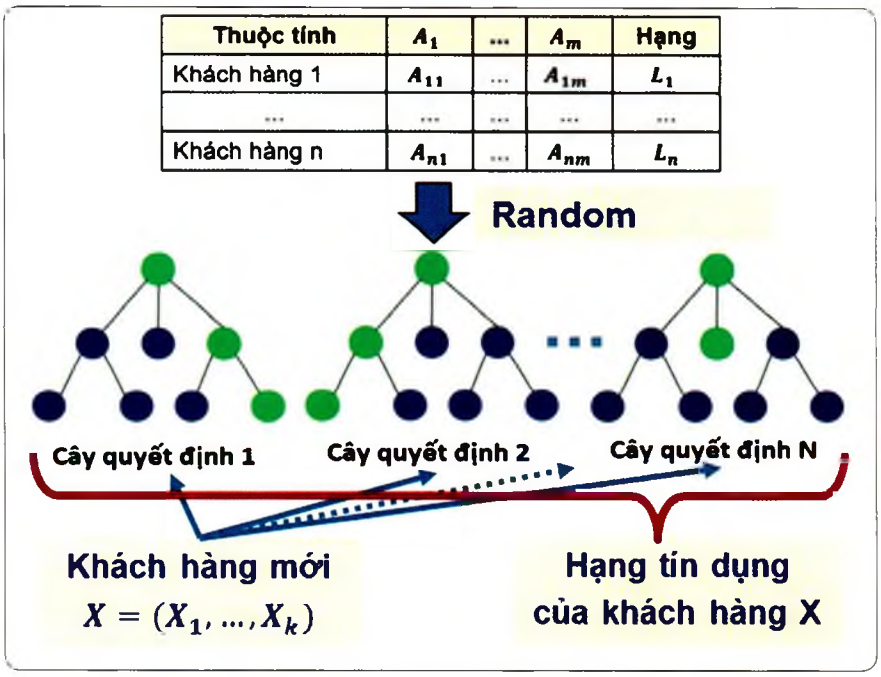
2. Phương pháp phân lớp Random forest

2.1. Sơ lược về thuật toán Random forest

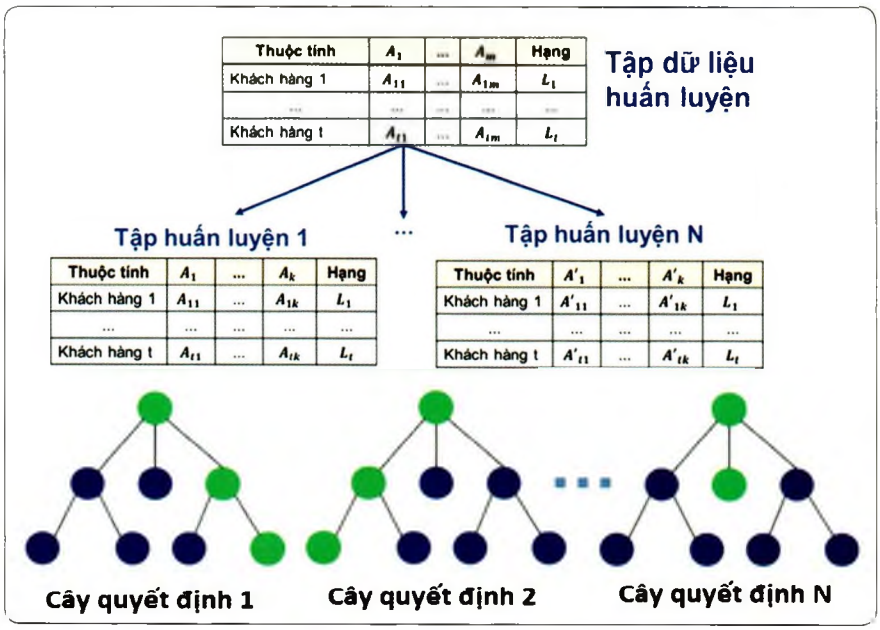
Trong học máy, Random forest là một kỹ thuật khá tiêu biểu có thể được sử dụng trong cả phân lớp và hồi quy. Về cơ bản, Random forest xây dựng nhiều cây quyết định (Decision tree) và tổng hợp kết quả của những cây này để đạt được đầu ra cuối cùng. (Hình 1)

Ý tưởng đầu tiên của thuật toán Random forest được giới thiệu bởi Tin Kam Ho (nhà khoa học máy tính tại IBM Watson Health). Trong nghiên cứu này, tác giả đã sử dụng phương pháp lựa chọn ngẫu nhiên tập con thuộc tính của bộ thuộc tính để đưa vào xây dựng các cây. Sau đó, một phiên bản mở rộng của thuật toán này đã được phát triển độc lập bởi hai nhà khoa học Leo Brleiman và Adele Cutler (Mỹ) bằng cách tạo ra các tập con từ tập dữ liệu ban đầu trước khi đưa vào xây dựng các cây. Gần đây, khi nhắc tới thuật toán Random forest, người ta thường hiểu rằng thuật toán này xây dựng các cây bằng cả hai phương pháp tiền xử lý dữ liệu ở trên: Trước tiên tạo ra các tập dữ liệu con, sau đó với mỗi tập dữ liệu con này chỉ giữ lại một số thuộc tính ngẫu nhiên được chọn. Nói một cách dễ hiểu, thuật toán Random forest tạo ra nhiều Decision tree, mỗi cây được xây dựng dùng thuật toán trên tập dữ liệu khác nhau và dùng tập thuộc tính khác nhau. Sau đó kết quả dự đoán đầu ra sẽ được tổng hợp từ các cây quyết định này. Trong các lĩnh vực ứng dụng, Random forest được dùng như một "hộp đen" bởi không dễ giải thích cơ chế làm việc của thuật toán này. (Hình 2)

Hình 1. Ví dụ về thuật toán Random forest



Hình 2. Ví dụ về cơ chế thực hiện của Random forest



2.2. Ý tưởng thuật toán Decision tree ID3

Ở nội dung này, chúng tôi trình bày lại ý tưởng của thuật toán Decision tree ID3 (Iterative Dichotomiser 3) - một trong những thuật toán điển hình được sử dụng trong thuật toán Random forest. ID3 là một thuật toán được John Ross Quinlan (trường Đại học Sydney, Australia) phát minh, để tạo Decision tree từ một tập dữ liệu cho trước.

Ý tưởng của ID3 như sau: Thuật toán ID3 xác định thứ tự của thuộc tính cần được xem xét tại mỗi bước. Với các bài toán có nhiều thuộc tính và mỗi thuộc tính có nhiều giá trị khác nhau, việc tìm được phương án tối ưu thường là không khả thi. Thay vào đó, một cách đơn giản thường được sử dụng là tại mỗi bước, một thuộc tính tốt nhất sẽ được chọn ra dựa trên một tiêu chuẩn. Với mỗi thuộc tính được chọn, ta chia dữ liệu vào các nhánh tương ứng với các giá trị của thuộc tính đó rồi tiếp tục áp dụng phương pháp này cho mỗi nhánh. Việc chọn ra thuộc tính tốt nhất ở mỗi bước như thế này được gọi là cách chọn tham lam (greedy). Cách chọn này có thể không phải là tối ưu mà sẽ gần với cách làm tối ưu. Ngoài ra, cách làm này khiến cho bài toán cần giải quyết trở nên đơn giản hơn.

Sau mỗi câu hỏi, dữ liệu được phân chia vào từng nhánh tương ứng với các câu trả lời cho câu hỏi đó. Câu hỏi ở đây chính là một thuộc tính, câu trả lời chính là giá trị của thuộc tính đó. Để đánh giá chất lượng của một cách phân chia, chúng ta cần đi tìm một phép đo.

Ta thấy rằng, một phép phân chia là tốt nhất nếu dữ liệu trong mỗi nhánh hoàn toàn thuộc vào một lớp (class) - khi đó, mỗi nhánh có thể được coi là một nút lá, tức là ta không cần phân chia thêm nữa. Nếu dữ liệu trong các nhánh vẫn lẫn vào nhau theo tỷ lệ lớn, phép phân chia đó được cho rằng chưa thực sự tốt. Từ nhận xét này, cần có một hàm số đo độ đồng nhất (purity) hoặc độ không đồng nhất (impurity) của một phép phân chia. Hàm số này sẽ cho giá trị thấp nhất nếu dữ liệu trong mỗi nhánh nằm trong cùng một lớp (tinh khiết nhất), và cho giá trị cao nếu mỗi nhánh có chứa dữ liệu thuộc nhiều lớp khác nhau.

Hàm số Entropy được dùng nhiều trong lý thuyết thông tin là hàm có các đặc điểm này. Cho một phân phối xác suất của một biến rời rạc x có thể nhận n giá trị khác nhau $x_1, x_2 \ldots, x_n$. Giả sử rằng xác suất để x nhận các giá trị này là $p_i = p(x = x_i)$ với $0 \le p_i \le 1$:

$$\sum_{i=1}^{n} p_i = 1$$

Ký hiệu phân phối này là $p=(p_1,p_2,\ldots,p_n)$. Entropy của phân phối được định nghĩa là:

$$H(p) = \sum_{i=1}^{n} p_i \log(p_i)$$

Những tính chất này, khiến hàm Entropy sử dụng trong việc đo độ hỗn độn của một phép phân chia của ID3. Vì vậy, thuật toán ID3 còn được gọi là thuật toán Decision tree dựa trên độ đo của hàm Entropy.

Tiếp theo, độ lợi thông tin IG(A) được định nghĩa là thước đo sự khác biệt trong hàm Entropy từ trước đến sau khi tập hợp S được phân chia trên một thuộc tính A. Nói cách khác, mức độ không thuần nhất trong S đã được giảm sau khi tách tập hợp S trên thuộc tính A.

$$IG(S, A) = H(S) - \sum_{t \in T} p(t)H(t) = H(S) - H(S|A)$$

Trong đó:

H(S) - Entropy của bộ S.

T - Các tập hợp con được tạo từ việc tách tập hợp tách theo thuộc tính A.

$S = U_{t \in T} t$

P(t) - Tỷ lệ số phần tử trong t với số phần tử trong tập hợp S.

H(t) - Entropy của tập hợp con *t*.

Trong ID3, độ lợi thông tin có thể được tính toán (thay vì tính Entropy) cho mỗi thuộc tính còn lại. Thuộc tính có mức tăng thông tin lớn nhất được sử dụng để tách tập hợp S ở lần lặp đang xét.

Chi tiết về thuật toán ID3 được trình bày như sau:

Đầu vào: Tập mẫu huấn luyện S, tập thuộc tính phân lớp C, tập thuộc tính A.

Đầu ra: Decision tree.

Thuật toán:

- Bước 1: Tạo Nút gốc cho Decision tree.
- Bước 2: Nếu tất cả các mẫu huấn luyện đều có giá trị của nhãn là P, trở về cây có một nút duy nhất là Nút gốc với nhãn P.
- Bước 3: Nếu A rỗng, trở về cây có một nút duy nhất là Nút gốc với nhãn là giá trị phổ biến nhất trong C.
- Bước 4:
  + Gọi X là một thuộc tính trong A phân lớp S tốt nhất.
  + Gán nhãn cho nút gốc với tên thuộc tính X.
  + A = A – {X}.
  + Cho từng giá trị v của X.
  + Thêm một nhánh mới dưới Nút gốc với X = v.
  + Xác định tập con S_v ứng với X = v.
  + Nếu S_v rỗng thì thêm dưới nhánh mới này một nút lá có nhãn là giá trị phổ biến nhất của thuộc tính quyết định trong S.
  + Ngược lại thêm cây con vào dưới nhánh này bằng cách gọi đệ quy ID3 (S_v, C, A-{X}).
- Bước 5: Trở về Nút gốc.

2.3. Ưu điểm và nhược điểm của Random forest

Ưu điểm:

- Random forest có thể giải quyết cả bài toán phân lớp và hồi quy.

- Chất lượng mô hình dự báo thường tốt hơn các thuật toán cây quyết định khác.

- Không gặp phải vấn đề quá khớp dữ liệu (overfitting).

Nhược điểm:

Do khâu chia tập dữ liệu ban đầu thành các tập con của thuật toán Random forest mang nhiều tính chất ngẫu nhiên nên khả năng diễn giải của thuật toán bị hạn chế. Chính vì thế, người dùng thường coi nó như "hộp đen" khi sử dụng thuật toán này.

3. Giải pháp dự báo giao dịch thẻ tín dụng lừa đảo dựa trên kỹ thuật Random forest

3.1. Phát biểu bài toán

Như vậy, chúng ta thấy rằng, một lĩnh vực khác trong ứng dụng trí tuệ nhân tạo có thể được sử dụng trong ngành Ngân hàng với mục đích phát hiện gian lận. Với sự giúp đỡ của các thuật toán trí tuệ nhân tạo, các hành động gian lận ngày càng được phát hiện nhiều hơn. Có hai phương pháp tiếp cận phổ biến đã được phát triển bởi tổ chức tài chính để phát hiện các mô hình gian lận.

- Phương pháp tiếp cận thứ nhất, các ngân hàng thương mại cần phải sử dụng đến kho dữ liệu của bên thứ ba và sử dụng các kỹ thuật trí tuệ nhân tạo để xác định mô hình gian lận, sau đó, các ngân hàng có thể tham chiếu chéo các mẫu với cơ sở dữ liệu riêng của mình.

- Phương pháp thứ hai, gian lận được nhận dạng dựa trên các mẫu thông tin nội bộ riêng của mình mà không phải nhờ vào bên thứ ba. Tuy nhiên, trên thực tế hầu hết các ngân hàng đang sử dụng kết hợp cả hai phương pháp tiếp cận trên.

Trong phần tiếp theo của bài viết, chúng tôi trình bày một phương pháp phát hiện gian lận sử dụng thuật toán học máy và dữ liệu lịch sử của các ngân hàng. Ý tưởng của phương pháp là sử dụng bộ dữ liệu mà các ngân hàng đang lưu trữ và các lớp thuật toán học máy để tạo ra các mô hình nhằm phát hiện đâu là khách hàng có khả năng gian lan lận trong số hàng triệu các khách hàng đang giao dịch với ngân hàng.

Bài toán có thể phát biểu dưới dạng mô hình toán học ngắn gọn như sau: Gọi X là tập dữ liệu gồm k thuộc tính về n khách hàng cần đánh giá khả xem họ có phải là đối tượng gian lận hay không. Gọi C là tập các giá trị (gồm hai giá trị 0 và 1) để đánh dấu khách hàng có gian lận hay không ($C \in \{0, 1\}$). Ta gọi $f : X \rightarrow C$ là hàm xác định khách hàng có gian lận hay không. Mục tiêu của bài toán là cần tính toán $f(xi) \in \{0, 1\}$, $\forall i = 1,...n$.

3.2. Mô tả dữ liệu

Dữ liệu để thực nghiệm cho thuật toán Random forest trong bài báo cáo này là bộ dữ liệu ccFraud.csv đã nói ở phần Giới thiệu. Các đối tượng khách hàng này gồm 8 thuộc tính cơ bản có ảnh hưởng nhiều nhất tới việc dự báo. Các thuộc tính, sau khi tiền xử lý với các thư viện mã nguồn mở và ngôn ngữ lập trình Python và lưu dưới dạng file excel với tên: ccFraud.csv. (Bảng 1, Bảng 2)

3.3. Triển khai thực nghiệm và đánh giá kết quả

Trong quá trình thử nghiệm, chúng tôi sử dụng quy trình thực hiện theo quy trình học máy. Quy trình này có thể tóm tắt bằng các bước thực hiện trên Python và các thư viện học máy trên nền tảng của Jupyter Lab. (Bước 1 - 5)

Bảng 1: Cấu trúc cụ thể của bộ dữ liệu

STT	Tên biến	Giải thích
1	FraudRisk	Biến phụ thuộc (mục tiêu dự đoán), dự đoán khách hàng có gian lận tín dụng hay không (0 - không, 1 - có gian lận)
2	Gender	Giới tính (Nam là 1; nữ là 2)
3	State	Khu vực khách hàng sử dụng dịch vụ
4	Cardholder	Khách hàng có sử dụng card (Có sử dụng là 1; không sử dụng là 0)
5	Balance	Số dư trong tài khoản của khách hàng
6	numTrans	Số giao dịch trong hệ thống ngân hàng
7	numIntlTrans	Số giao dịch trong hệ thống liên ngân hàng
8	CreditLine	Hạn mức tín dụng, là mức dư nợ vay tối đa được duy trì trong một thời hạn nhất định mà ngân hàng và khách hàng thỏa thuận trong hợp đồng tín dụng.

Bảng 2: Một số bản ghi ví dụ trong bộ dữ liệu

CustID	Gender	State	Cardholder	Balance	NumTrans	NumIntlTrans	CreditLine	FraudRisk
1	1	35	1	3000	4	14	2	0
4	1	15	1	0	12	0	5	0
22940	2	44	1	17000	45	0	16	1
271356	2	5	1	10397	100	0	12	1
271407	1	23	1	16724	13	6	35	1

Bước 1: Khai báo các thư viện sử dụng.

```
import pandas as pd
import numpy as np
import re
import sklearn
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
from collections import Counter
from sklearn.ensemble import RandomForestClassifier, VotingClassifier
from sklearn.model_selection import GridSearchCV, cross_val_score, StratifiedKFold, learning_curve
from sklearn.feature_selection import SelectFromModel, SelectKBest
from sklearn.pipeline import make_pipeline
from sklearn.model_selection import StratifiedKFold
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import GridSearchCV
sns.set(style='white', context='notebook', palette='deep')
pd.options.display.max_columns = 100
```

Bước 2: Đọc dữ liệu từ bộ nhớ ngoài.

```
data = pd.read_csv("D:\Datasets\Dataset for ML\ccFraud.csv")
print(data.shape)
```

Bước 3: Mô tả các giá trị thống kê cơ bản.

	custID	gender	state	cardholder	balance	numTrans	numIntlTrans	creditLine	fraudRisk
count	1.000000e+07	1.000000e+07	1.000000e+07	1.000000e+07	1.000000e+07	1.000000e+07	1.000000e+07	1.000000e+07	1.000000e+07
mean	5.000000e+06	1.382177e+00	2.466127e+01	1.030004e+00	4.109920e+03	2.893519e+01	4.047190e+00	9.134469e+00	5.960140e-02
std	2.886751e+06	4.859195e-01	1.497012e+01	1.705991e-01	3.996847e+03	2.655378e+01	8.602970e+00	9.641974e+00	2.367469e-01
min	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00
25%	2.500001e+06	1.000000e+00	1.000000e+01	1.000000e+00	0.000000e+00	1.000000e+01	0.000000e+00	4.000000e+00	0.000000e+00
50%	5.000000e+06	1.000000e+00	2.400000e+01	1.000000e+00	3.706000e+03	1.900000e+01	0.000000e+00	6.000000e+00	0.000000e+00
75%	7.500000e+06	2.000000e+00	3.800000e+01	1.000000e+00	6.000000e+03	3.900000e+01	4.000000e+00	1.100000e+01	0.000000e+00
max	1.000000e+07	2.000000e+00	5.100000e+01	2.000000e+00	4.148500e+04	1.000000e+02	6.000000e+01	7.500000e+01	1.000000e+00

Bước 4: Xây dựng mô hình bằng huấn luyện mô hình trên bộ dữ liệu huấn luyện sau khi thực hiện chia bộ dữ liệu thành hai phần theo tỷ lệ 70% là tập dữ liệu huấn luyện, 30% là tập dữ liệu kiểm tra.

```
# Splitting the data into Train and Test
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 0)
```

Bước 5: Sử dụng mô hình để đánh giá kết quả:

```
from sklearn.ensemble import RandomForestClassifier
Model=RandomForestClassifier(max_depth=3)
Model.fit(X_train,y_train)
y_pred=Model.predict(X_test)
```

3.4. Kết quả và lưu ý

Sau khi thực hiện mô hình, chúng ta thu được ma trận kết quả sau: Khi sử dụng 3.000.000 đối tượng khách hàng của tập dữ liệu kiểm tra chạy qua mô hình, kết quả đạt được với độ chính xác là 95% và được diễn giải trong Bảng 3.

Trong đó:

- 2.818.811 đối tượng khách hàng thực tế không có nghi ngờ gian lận và khi cho chạy qua mô hình cho kết quả là không nghi ngờ gian lận.
- 23.825 đối tượng khách hàng thực tế có nghi ngờ gian lận và khi cho chạy qua mô hình cho kết quả là có nghi ngờ gian lận.
- 2.354 đối tượng khách hàng thực tế không nghi ngờ gian lận và khi cho chạy qua mô hình cho kết quả là có nghi ngờ gian lận.
- 155.010 đối tượng khách hàng thực tế có nghi ngờ gian lận và khi cho chạy qua mô hình cho kết quả là không có nghi ngờ gian lận

Phần trên của bài viết đã trình bày quy trình sử dụng thuật toán Random forest khi tìm kiếm thông tin từ dữ liệu ngân hàng nhằm phân lớp khách hàng có nghi ngờ gian lận trong tín dụng hay không. Để có kết quả mang tính ứng dụng phù hợp với thực tế hơn, chúng ta cần phải thực hiện thuật toán này trên bộ dữ liệu thu thập được từ các ngân hàng thương mại tại Việt Nam. Đồng thời, cần tìm hiểu thêm tình hình thực tế để từ đó cải tiến chương trình, thay đổi các tham số cài để bài toán phù hợp với thực tế tại Việt Nam.

Bảng 3

	Không gian lận (thực tế)	Gian lận (thực tế)
Không gian lận (dự đoán)	2.818.811	155.010
Gian lận (dự đoán)	2.354	23.825

4. Kết luận và đánh giá

Chúng ta đã hệ thống hóa cơ sở lý thuyết về dữ liệu cũng như phân tích và nghiên cứu các vấn đề liên quan nhằm đưa ra giải pháp và áp dụng vào quy trình phát hiện gian lận tín dụng. Việc nghiên cứu áp dụng các mô hình mới là cần thiết để nâng cao tính chính xác, độ tin cậy, tính khách quan khi ra quyết định cho vay. Qua quá trình này, chúng ta có thể đưa ra được những đánh giá tổng quát sau:

Thứ nhất, ứng dụng công nghệ học máy vào phát hiện gian lận tín dụng của ngân hàng là một phương pháp hiện đại đang dần chiếm ưu thế khi tối thiểu hóa được quy trình thẩm định tín dụng tại các ngân hàng. Với công nghệ học máy, các ngân hàng hoàn toàn có thể dùng các thuật toán dựa trên các kho dữ liệu đã có sẵn về khách hàng để đánh giá một cách khách quan và hiệu quả về tín dụng khách hàng.

Thứ hai, có thể nói rằng, việc ứng dụng học máy vào lĩnh vực tín dụng làm giảm đáng kể rủi ro ngân hàng. Ngành Ngân hàng tính đến thời điểm hiện tại vẫn chưa thật sự tiếp cận toàn diện đến ứng dụng kỹ thuật số, vì thế, các rủi ro khi làm việc với giấy tờ như thất lạc, sai số là điều khó có thể tránh khỏi.

Thứ ba, ứng dụng công nghệ học máy cũng giúp cho thời gian thực hiện mỗi lần đánh giá tín dụng nói riêng và các công việc của ngân hàng nói chung trở nên nhanh hơn và đáng tin cậy hơn. Sở dĩ như vậy là bởi khả năng tính toán và đưa ra quyết định của con người là có hạn, trong khi đó học máy cũng có thể làm được điều tương tự với tốc độ nhanh hơn gấp nhiều lần. Không chỉ tốc độ, các chỉ tiêu đánh giá khách hàng đã được mở rộng hơn, từ đó khiến cho các đánh giá mang tính khách quan hơn và có chiều sâu hơn. Ngoài ra, các phương thức trên cũng là một chỉ tiêu mới được đưa ra nhằm đa dạng hóa khả năng thanh toán cho khách hàng, cho thấy sự linh hoạt ứng biến tốt của ngân hàng.

Cuối cùng, như đã nói ở trên, khi chúng ta ứng dụng công nghệ của học máy vào trong việc phát hiện gian lận của khách hàng nói riêng và hoạt động tài chính của ngân hàng nói chung, thời gian xử lý của học máy là rất nhanh và tiện lợi, điều đó đồng nghĩa với ngân hàng sẽ tiếp cận được nhiều khách hàng hơn. Lượng khách hàng lớn hơn sẽ đem lại doanh thu cao hơn cho ngân hàng, đi đôi với đó là chi phí nhân sự và chi phí quản lý giảm xuống đáng kể. Khả năng thu thập thông tin của học máy rất nhanh và từ nhiều nguồn đáng tin cậy là một bước tiến lớn, khi mà các tổ chức tín dụng hiện giờ chưa áp dụng hoặc mới áp dụng được một phần vào trong việc đánh giá. Bước tiến này sẽ là tiền đề cho các tổ chức tín dụng khác học hỏi theo và phát triển, đem lại cho khách hàng những trải nghiệm hoàn toàn mới lạ và tốt nhất.■

TÀI LIỆU THAM KHẢO:

1. Analysis of Financial Credit Risk Using Machine Learning, 2017.
2. Data mining techniques: study, analysis, prevention & detection for financial cyber crime and frauds, 2010.
3. https://baotintuc.vn/thong-cao-bao-chi/gbg-su-dung-hoc-may-va-ai-de-phat-hien-gian-lan-trong-giao-dich-the-tin-dung-thanh-toan-so-20200528151051901.htm
4. https://ichi.pro/vi/phat-hien-gian-lan-the-tin-dung-48935557595017
5. Ho, Tin Kam. "Random decision forests." Proceedings of 3rd international conference on document analysis and recognition. Vol. 1. IEEE, 1995.
6. Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.
7. Leo Breiman and Adele Cutler, Package 'randomForest' Breiman and Cutler's Random Forests for Classification and, 2018.