Anomaly Detection in Credit Card Transaction

Tăng Minh Hiển
Dept. of Computer Science
Vietnam National University – Ho Chi Minh City
Ho Chi Minh City, Vietnam
21520229@gm.uit.edu.vn

Nguyễn Thái Thành Long
Deft. of Computer Science
Vietnam National University – Ho Chi Minh City
Ho Chi Minh City, Vietnam
21520334@gm.uit.edu.vn

Châu Thiên Long
Dept. of Computer Science
Vietnam National University – Ho Chi Minh City
Ho Chi Minh City, Vietnam
21520331@gm.uit.edu.vn

***Tóm tắt* — Thẻ tín dụng là một trong các hình thức giao dịch phổ biến hiện nay trong thời đại phát triển của các nền tảng thanh toán trực tuyến phục vụ cho Thương mại điện tử. Hình thức giao dịch này đem lại rất nhiều lợi ích như có thể nhanh chóng t. Trong phạm vi đồ án môn học, nhóm trình bày về đồ án Nhận dạng bất thường trong các giao dịch thẻ tín dụng, áp dụng công nghệ dữ liệu lớn Apache Pyspark và xử lý data streaming Kafka. Hệ thống được xây dựng với mục đích học tập và nghiên cứu. Cụ thể, nhóm thực hiện xây dựng và huấn luyện các mô hình Máy học có thể huấn luyện được các mô hình**

Từ khóa—Anomaly detection, Credit card, Big data, Machine Learning, Kafka, Apache Spark

I. GIỚI THIỆU

Thẻ tín dụng là loại thẻ cho phép chủ thẻ thực hiện giao dịch trong phạm vi hạn mức tín dụng đã được cấp theo thỏa thuận giữa người dùng với tổ chức phát hành thẻ. Nói một cách đơn giản, thẻ tín dụng là loại thẻ giúp bạn mua hàng trước và thanh toán lại cho ngân hàng sau [1]. Giao dịch qua thẻ tín dụng đang là loại hình giao dịch phổ biến trong thời đại kinh tế điện tử phát triển mạnh mẽ gần đây. Trong năm 2023 đã ghi nhận 724 tỷ giao dịch thẻ tín dụng với trung bình 1.98 tỷ giao dịch mỗi ngày, hoặc 22 950 giao dịch mỗi giây trên toàn cầu [2]. Còn tại Việt Nam, hình thức giao dịch thẻ tín dụng nằm trong top 2 hình thức giao dịch phổ biến chỉ xếp sau hình thức giao dịch tiền mặt. Cụ thể, tính đến tháng 3 năm 2024 ghi nhận 150.6 triệu thẻ tín dụng đang lưu hành. Trong khi đó, giá trị giao dịch thẻ tín dụng ghi nhận được là 1.3 triệu giao dịch với khoảng 10 nghìn tỉ đồng, tăng cả về số lượng và giá trị so với cùng kỳ 2023 [3].

Hình thức giao dịch này cung cấp nhiều tiện ích như cung cấp nhiều tiện ích cho người dùng có thể thực hiện các giao dịch offline và online. Tuy nhiên, hình thức này lại tiềm ẩn nhiều lỗ hổng, trong đó đáng chú ý là lỗ hổng về các giao dịch bất thường trong hình thức giao dịch này. Giao dịch bất thường có thể định nghĩa là các giao dịch thực hiện bởi một bên không được ủy quyền thực hiện giao dịch và không có sự chấp thuận chủ sở hữu hợp pháp hoặc tổ chức của thẻ tín dụng. Trong những trường hợp này, những kẻ lừa đảo có thể sử dụng thẻ vì lợi ích cá nhân của mình, làm cạn kiệt tài nguyên của thẻ hoặc cho đến khi bị bắt hoặc thẻ bị khóa. Mục đích của các giao dịch này nhằm thu lợi các cá nhân, tổ chức và dẫn đến hậu quả tài chính cho chủ thẻ cũng như công ty chủ thẻ. Việc tạo giao dịch gian lận có thể được tạo ra từ việc đánh cắp thẻ tín dụng để thực hiện giao dịch offline hoặc đánh cắp thông tin thẻ để thực hiện giao dịch online. Mục đích của các giao dịch nhằm đến trục lợi cho cá nhân, tổ chức và gây ra hậu quả thất thoát tài chính cho chủ thẻ cũng như công ty của thẻ.

Vì thế đã có các nghiên cứu nhằm tìm kiếm các phương pháp phát hiện các giao dịch bất thường. Một số phương pháp đã đề xuất như kiểm tra quá trình giao dịch từng thẻ và tìm ra các đặc trưng giữa các giao dịch bất thường và giao dịch thông thường dựa trên các thông tin của một giao dịch gồm thông tin và dữ liệu liên quan đến việc mua hàng, thời gian đã trôi qua kể từ lần mua cuối cùng, số tiền được sử dụng để mua hàng, v.v. Tuy nhiên cần hiểu rằng, không thể có đủ tài nguyên về nhân lực và thời gian để có thể điều tra từng giao dịch một trong số rất nhiều giao dịch hằng ngày [4].

Từ đó, chủ đề "Anomaly Detection in Credit Card Transaction" là một chủ đề học thuật được sự quan tâm lớn từ các nhà nghiên cứu cũng như các công ty thẻ tín dụng, nơi các cá nhân và tổ chức đang từng ngày cố gắng xác định giải pháp tốt nhất cho vấn đề này và cũng như cần cập nhật để theo kịp các cách tiếp cận luôn thay đổi được của những kẻ tạo ra các giao dịch bất thường này. Nhóm nhận thấy nhiều thách thức từ bài toán này bao gồm việc cần xử lý dữ liệu theo thời gian thực (real-time data) và đưa ra kết quả nhanh nhất ngay sau khi giao dịch đã được thực hiện. Bên cạnh đó là xử lý trên số lượng lớn giao dịch hằng ngày, nên yêu cầu hệ thống có thể làm việc với các nền tảng xử lý Big Data nhằm thực hiện vấn đề này. Bên cạnh đó, vấn đề mất cân bằng dữ liệu là vấn đề cần phải giải quyết khi số lượng các giao dịch thường sẽ rất ít các giao dịch bất thường so với số lượng các giao dịch hằng ngày. Và thách thức về độ chính xác trong việc dự đoán xem giao dịch đó gian lận hay không bởi nếu bỏ lỡ giao dịch bất thường, điều này sẽ gây nhiều thiệt hại như đã đề cập trên. Tuy nhiên, nhóm cũng nhận thấy tiềm năng của bài toán này như đây là đã có nhiều nền tảng giúp xử lý dữ liệu real-time cũng như Big Data hỗ trợ xây dựng ứng dụng này. Với sự phát triển của Máy học, các nhà nghiên cứu đã xây dựng các mô hình dự đoán dựa trên các thuật toán của Máy học để giải quyết bài toán này, như SVM, Logistic Regression, mạng học sâu Neural Network,... với nhiều cách xây dựng và tiền xử lý dữ liệu khác nhau, đã có các kết quả tích cực từ các mô hình này.

Bài toán của nhóm thực hiện trong đồ án này được mô tả như sau:

- Input:
  - Tập dữ liệu các giao dịch thẻ tín dụng gồm các giao dịch đã được phân loại đáng ngờ hay không
  - Các giao dịch mới chưa được phân loại gian lận hay không
- Output: Dự báo các giao dịch mới xem có thuộc giao dịch gian lận hay không

Các đóng góp mà nhóm thực hiện bài toán này bao gồm:

- Áp dụng các phương pháp tiền xử lý dữ liệu gồm Robust Scaler và phương pháp xử lý mất cân bằng dữ liệu cho bài toán.
- Thực hiện huấn luyện các mô hình Máy học từ bộ dữ liệu mà nhóm lựa chọn bằng framework Apache Pyspark và so sánh giữa các phương pháp nhóm đề xuất.
- Áp dụng nền tảng dữ liệu lớn Apache Pyspark và data streaming bằng Kafka cho dữ liệu thực tế nhằm hiện thực hệ thống mà nhóm xây dựng.

Trong báo cáo này, nhóm trình bày các mục bao gồm Các nghiên cứu liên quan (phần II), Hệ thống đề xuất (Phần III), Thực nghiệm và kết quả (Phần IV), Kết luận (Phần V).

II. CÁC NGHIÊN CỨU LIÊN QUAN

Phát hiện bất thường trong giao dịch thẻ tín dụng là một lĩnh vực nghiên cứu quan trọng nhằm ngăn chặn gian lận tài chính. Nhiều phường pháp đã được để xuất qua các năm, sử dụng các kỹ thuật học máy và thống kê khác nhau. Sau đây là các nghiên cứu liên quan mà nhóm đã tìm kiếm liên quan đến bài toán này và bộ dữ liệu mà nhóm sử dụng.

A. *Các phương pháp áp dụng*

Đã có nhiều nghiên cứu trước của bài toán này, sau đây là tóm tắt các phương pháp mà nhóm tìm được.

1) Các phương pháp thống kê

Các phương pháp phát hiện bất thường ban đầu chủ yếu dựa vào các kỹ thuật thống kê. Bolton và Hand (2002) đã nghiên cứu các kỹ thuật thống kê như hệ thống dựa trên luật, mạng Bayesian, và phân tích nhóm ngang hàng để phát hiện các giao dịch bất thường [5]. Các phương pháp này dựa vào việc xây dựng các mô hình từ dữ liệu giao dịch để xác định các giao dịch có thể là gian lận. Tuy nhiên, các phương pháp này thường gặp phải tỷ lệ dương tính giả cao và khó khăn trong việc thích nghi với những thay đổi trong hành vi gian lận.

2) Các phương pháp Máy học có giám sát

Sự phát triển của học máy đã đem lại nhiều cải tiến cho lĩnh vực này. Các phương pháp học có giám sát như cây quyết định, máy vector hỗ trợ (SVM), và mạng neural đã được áp dụng để cải thiện độ chính xác của các hệ thống phát hiện gian lận. Bhattacharyya et al. (2011) đã chứng minh rằng việc kết hợp nhiều bộ phân loại có thể nâng cao hiệu suất phát hiện [6]. Các phương pháp như Random Forests và Gradient Boosting Machines đã cho thấy hiệu quả trong việc nắm bắt các mẫu giao dịch phức tạp. Tuy nhiên, những phương pháp này yêu cầu dữ liệu có nhãn, điều này không phải lúc nào cũng có sẵn.

3) Các phương pháp Máy học không giám sát và bán giám sát

Do khó khăn trong việc thu thập dữ liệu có nhãn, các phương pháp học không giám sát và bán giám sát đã được nghiên cứu sâu rộng. Các thuật toán như K-means, DBSCAN, Isolation Forests và One-Class SVM được sử dụng để phát hiện các điểm bất thường mà không cần dữ liệu có nhãn. Chandola và cộng sự (2009) đã cung cấp một khảo sát toàn diện về các kỹ thuật phát hiện bất thường, nhấn mạnh hiệu quả của các phương pháp này trong việc xác định các mẫu bất thường có thể chỉ ra gian lận [7].

4) Các phương pháp Neural Network

Những tiến bộ gần đây trong học sâu đã mở ra nhiều cơ hội mới cho việc phát hiện bất thường. Các mô hình như Autoencoders và mạng neural hồi quy (RNNs) đặc biệt hiệu quả trong việc xử lý dữ liệu tuần tự và nắm bắt các phụ thuộc theo thời gian trong giao dịch. LSTM (Long Short-Term Memory) đã được áp dụng thành công trong việc phát hiện các bất thường dựa trên chuỗi thời gian của giao dịch. Hơn nữa, các mạng đối kháng sinh (GANs) đã được sử dụng để tạo ra dữ liệu tổng hợp, nâng cao khả năng học tập của các mô hình (Fiore và cộng sự, 2019) [8].

5) Các mô hình Học kết hợp và phương pháp Ensemble Learning

Để tận dụng ưu điểm của các phương pháp khác nhau, các mô hình kết hợp và phương pháp ensemble đã được đề xuất. Các phương pháp này kết hợp các kỹ thuật thống kê, học máy, và học sâu để đạt được độ chính xác và độ bền cao hơn. Ví dụ, hệ thống kết hợp các kỹ thuật xử lý đặc trưng với các mô hình học sâu đã cho thấy kết quả hứa hẹn (Nguyễn và cộng sự, 2018) .

Mặc dù đã có nhiều tiến bộ, việc triển khai các hệ thống phát hiện bất thường trong thực tế vẫn đối mặt với nhiều thách thức. Những thách thức này bao gồm sự mất cân bằng giữa giao dịch gian lận và không gian lận, sự thay đổi liên tục của các chiến thuật gian lận, và nhu cầu xử lý dữ liệu thời gian thực. Các giải pháp như học trực tuyến và các thuật toán thích ứng đang được nghiên cứu để giải quyết những vấn đề này.

Tóm lại, lĩnh vực phát hiện bất thường trong giao dịch thẻ tín dụng rất phong phú với các phương pháp đa dạng, mỗi phương pháp đóng góp vào việc chống lại gian lận tài chính. Nghiên cứu này nhằm xây dựng trên những phương pháp hiện có, đề xuất một cách tiếp cận mới tích hợp để nâng cao khả năng phát hiện.

B. *Kết quả thực nghiệm*

Nhóm đã thực hiện thống kê các kết quả mà các nghiên cứu trước đã thực hiện, các kết quả được thống kê bao gồm thống kê các kết quả tốt bài toán này và thống kê kết quả của các mô hình đã thực hiện trên bộ dữ liệu Credit Card Transaction mà nhóm thực hiện.

Bảng II-1: Bảng các phương pháp và kết quả thực hiện của các bài báo

Reference	Method	Best result
Giulia Moschini, Regis Houssou, Jerome Bovay (2021) [9]	ARIMA, K Means	Precision: 34.29% Recall: 69.57% F-Measure: 36.19%
Soumaya Ounacer, Hicham Ait El Bour, Younes Oubrahim, Mohamed Yassine Ghoumari and Mohamed Azzouazi (2018) [10]	LOF, One class SVM, K-Means, Isolation Forest	FI Score: 5.44% Accuracy: 95.12% AUC Score: 91.68%
Suraya Nurain Kalid, Keng-Hoong NG, Gee-Kok Tong And Kok-Chin Khor (2020) [11]	Naive Bayes, C4.5, Random Forest, Random Tree, Logistic, Multilayer Perceptron, KNN	AUC: 97.5% True Positive Rate: 95.5%
Shanshan Jiang, Ruiting Dong, Jie Wang, Min Xia (2023) [12]	SVM, Decision Tree, XG Boost, KNN, Random Forest, LSTM,	Precision: 97.95% Recall: 75.53% F1 Score: 86.56% AUC: 95.15%

Reference	Method	Best result
	CNN, MLP, AE, UAAD-FDNet	

Bên dưới là bảng kết quả của các nghiên cứu trước trên bộ dữ liệu Credit Card Transaction (đính kèm đường dẫn hyperlink tại đây).

Bảng II-2: Bảng kết quả tốt nhất của bài toán Nhận dạng giao dịch bất thường trong giao dịch Thẻ tín dụng trên bộ dữ liệu Credit Card Transactinon

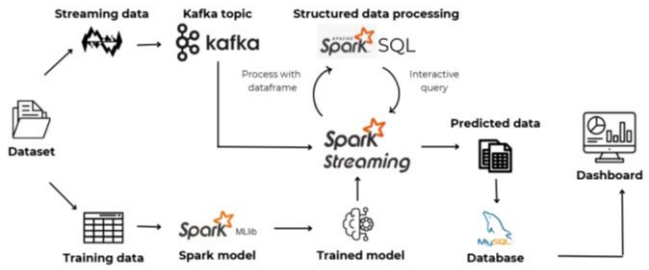
Reference	Method	Best result (Model)
Guansong Pang, Chunhua Shen , Anton van den Hengel (2019) [13]	DevNet, REPEN, DSVDD, FSNet, iForest	AUC-ROC: 0.980 ± 0.001 (DevNet model) AUPRC: 0.690 ± 0.002 (DevNet model)
Hongzuo Xu, Guansong Pang , Yijie Wang, Yongjun Wang (2023) [14]	DIF, EIF, PID, LeSiNN, IF, eRDP, eREPEN, eDSVDD, eRECON	AUC-ROC: 0.953 ± 0.002 (DIF model) AUPRC: 0.480 ± 0.026 (eDSVDD model)
Grover, Prince and Li, Zheng and Liu, Jianbo and Zablocki, Jakub and Zhou, Hao and Xu, Julia and Cheng, Anqi (2022) [15]	AFD OFI, AFD TFI, AutoGluon, H2O, Auto-sklearn	AUC-ROC: 0.992

III. HỆ THỐNG ĐỀ XUẤT

Hệ thống đề xuất của nhóm được chia thành 2 phần gồm: Xử lý Online và Xử lý Offline. Với hệ thống Xử lý Online, nhóm thực hiện huấn luyện các mô hình Máy học dựa trên bộ dữ liệu Credit Card Transaction, từ đó thu được các mô hình Máy học có thể nhận diện các giao dịch gian lận. Với Xử lý Offline, nhóm thực hiện xây dựng hệ thống phát hiện giao dịch gian lận theo thời gian thực, trong đó thực hiện đưa dữ liệu vào dưới dạng data streaming và lần lượt dự đoán các dữ liệu dựa trên các mô hình đã được huấn luyện ở phần Xử lý Offline.

A. *Tổng quan kiến trúc hệ thống*

Hình bên dưới thể hiện kiến trúc hệ thống mà nhóm xây dựng cho đồ án này.



Hình III-1: Kiến trúc hệ thống của ứng dụng phát hiện bất thường trong giao dịch thẻ tín dụng

Cụ thể về kiến trúc này, nhóm thực hiện chia dữ liệu thành 2 phần gồm: Streaming data và Training data. Trong đó, với Training data, thực hiện , nhóm thiết kế pipeline gồm bước tiền xử lý dữ liệu và huấn luyện các mô hình khác nhau, từ đó đưa ra đánh giá, so sánh các phương pháp và lựa chọn mô hình có kết quả tốt nhất. Với mô hình tốt nhất này, sẽ được đóng gói vào pipeline. Đối với Streaming data, dữ liệu được truyền bằng Kafka và được đưa vào các components được tích hợp bên trong thư viện Spark Streaming. Spark SQL được sử dụng để tái cấu trúc lại khung dữ liệu phù hợp với yêu cầu đầu vào của mô hình được đào tạo. Các mẫu dữ liệu đã được gắn nhãn được lưu trữ trong cơ sở dữ liệu, sau đó được quản lý, trực quan hóa và truy vấn thông qua dashboard. Dưới đây là phần trình bày chi tiết các hệ thống offline và hệ thống online mà nhóm thực hiện.

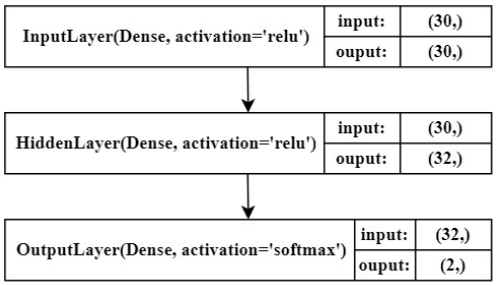
B. *Hệ thống offline*

Trong kiến trúc này, hệ thống xử lý offline là hệ thống nhóm thực hiện huấn luyện các mô hình Máy học dựa trên bộ dữ liệu Credit Card Transaction (đính kèm đường dẫn hyperlink tại đây). Trong đó, để thử nghiệm các mô hình cũng như tìm ra mô hình tối ưu cũng như có kết quả tốt nhất trên bộ dữ liệu mà nhóm xây dựng, nhóm đã xây dựng nhiều pipeline khác nhau kết hợp các bước xử lý khác nhau và các mô hình Máy học khác nhau. Trong các pipeline, nhóm tạo ra 1 baseline model gồm đưa dữ liệu không qua bước tiền xử lý, sau đó, cho mô hình Logistic Regression học. Đây là phương pháp baseline để thử nghiệm và so sánh với các phương pháp tiền xử lý và mô hình Máy học khác.

Với bước tiền xử lý dữ liệu, nhóm lựa chọn 2 bước gồm: thực hiện chuẩn hóa dữ liệu bằng Robust Scaler và thực hiện cân bằng dữ liệu bằng các phương pháp xử lý mất cân bằng dữ liệu. Trong đó, việc thực hiện chuẩn hóa dữ liệu được thực hiện bằng biến đổi Robust Scaler. Là phương pháp chuẩn hóa, sử dụng giá trị trung vị và median để chuẩn hóa dữ liệu sao cho chuẩn hóa dữ liệu gần lại với giá trị trung vị và chia cho giá trị IQR.; đây là phép biến đổi phù hợp với dữ liệu có các giá trị ngoại lệ và phân phối dữ liệu không đối xứng; phù hợp với thể hiện của dữ liệu đã được biểu diễn [16]. Sau đó, với dữ liệu huấn luyện, thực hiện các biện pháp giải quyết mất cân bằng dữ liệu gồm các phương pháp Undersampling - Ngẫu nhiên loại bỏ một số mẫu của lớp chiếm đa số, Oversampling các dữ liệu ở nhãn thiểu số (1 – Fraud) - Thêm các dữ liệu ngẫu nhiên ở lớp thiểu số, Oversampling sử dụng thư viện SMOTE - Tạo thêm các dữ liệu giả bằng cách nội suy giữa các mẫu hiện có của lớp thiểu số. Các phương pháp này đều hướng đến giải quyết mất cân bằng dữ liệu mà bản thân dữ liệu đang có do hiện thực cho thấy số lượng giao dịch bất thường là rất nhỏ so với giao dịch không bất thường.

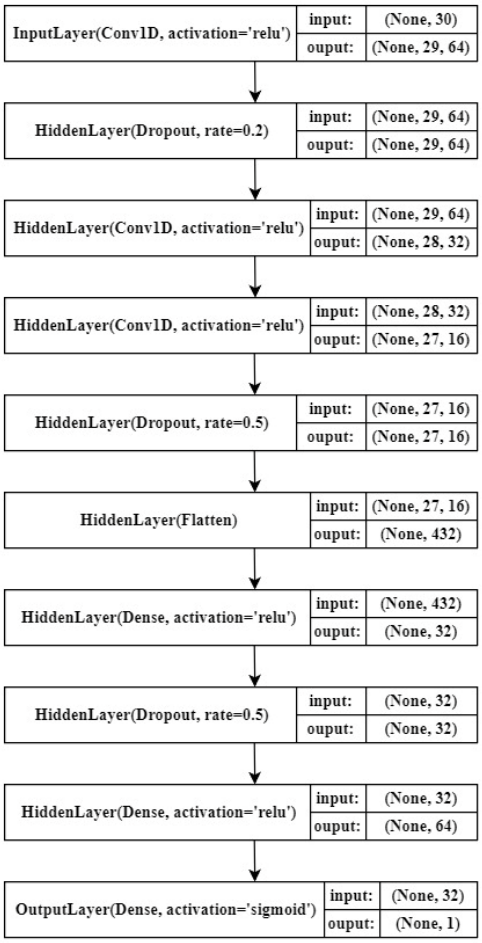
Với các mô hình Máy học, nhóm lựa chọn các mô hình Máy học có khả năng phân loại nhị phân tốt, cũng như các mô hình đạt được hiệu quả tốt ở một số domain liên quan đến bài toán nhóm xây dựng là Anomaly Detection cũng như tham khảo mô hình trên các notebook trên Kaggle nhằm thử nghiệm với bài toán này, các mô hình được lựa chọn bao gồm Logistic Regression, Support Vector Machine, XGBoost, Random Forest và Neural Network. Trong đó, với mô hình mạng Neural Network, nhóm đã thực hiện tham khảo dựa trên các notebook đi trước của bài toán này. Nhóm đã sử dụng 2 mô hình mạng Neural Network cho bài toán này với kiến trúc được thể hiện ở các hình bên dưới.

Mạng Neural Network đầu tiên nhóm sử dụng, được tham khảo từ đường dẫn tại đây, có kiến trúc như bên dưới:

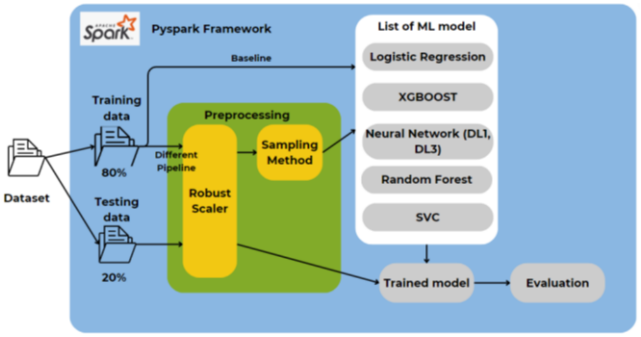


Hình III-2: Kiến trúc mạng Neural Network 1 (ký hiệu DL1)

Mạng Neural Network thứ hai được nhóm sử dụng, với mã được đánh dấu là DL3, được nhóm lấy tham khảo từ đường dẫn tại đây. Đây là kiến trúc sử dụng ý tưởng của cấu trúc mạng CNN và thiết kế lại dùng cho dữ liệu dạng bảng như bài toán mà nhóm thực hiện.



Hình III-3: Kiến trúc mạng Neural Network 2 (ký hiệu DL3)



Hình III-4: Sơ đồ các bước trong pipeline thực hiện huấn luyện và đánh giá hiệu suất mô hình

Để đưa ra nhận xét tổng quan về phương pháp và đánh giá từng phương pháp, nhóm thực hiện tổ chức các pipeline khác nhau, với mỗi pipeline là mỗi cách kết hợp các phương pháp mà nhóm đã lựa chọn ở từng bước. Tổng quan các bước được hiện thực ở hình III-4. Cụ thể các pipeline nhóm đã tiến hành thử nghiệm được liệt kê như bên dưới:

- Baseline: Logistic Regression (Không qua xử lý tiền xử lý dữ liệu)
- Pipeline 1: Robust Scaler + Undersampling + Logistic Regression
- Pipeline 2: Robust Scaler + Undersampling + Random Forest
- Pipeline 3: Robust Scaler + Oversampling duplicating minority class + Logistic Regression.
- Pipeline 4: Robust Scaler + Oversampling duplicating minority class + Random Forest
- Pipeline 5: Robust Scaler + DL1
- Pipeline 6: Robust Scaler + DL3
- Pipeline 7: Robust Scaler + SVC
- Pipeline 8: Robust Scaler + XGBoost
- Pipeline 9: RobustScaler + OverSampling(SMOTE library) + LogisticRegression
- Pipeline 10: RobustScaler + OverSampling(SMOTE library) + RandomForest

C. Hệ thống online

Hệ thống online bao gồm các bước xử lý:

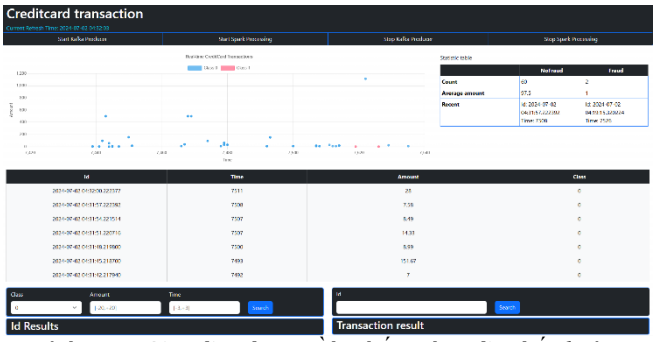
- Bước 1: Load dữ liệu dạng data streaming: Với dữ liệu streaming được phân chia từ dataset, đưa dữ liệu vào theo dạng data streaming và dữ liệu được lưu vào Kafka.

- Bước 2: Pipeline mô hình có kết quả tốt nhất (ở đây là XGBOOST) được lưu lại các thông số và load vào Spark Streaming. Các dữ liệu streaming được đưa vào xử lý theo dataframe và xử lý theo pipeline đã lưu lại để đưa ra các dự đoán bất thường hay không. Quá trình này lặp lại liên tục theo stream dữ liệu.

- Bước 3: Kết quả dự đoán sau khi được đưa ra thì được lưu lại trong database được lưu dưới dạng MySQL. Cuối cùng, với demo được xây dựng, nhóm thực hiện truy vấn cơ sở dữ liệu này và hiển thị kết quả này dưới dạng dashboard.

D. *Ứng dụng web demo*

Nhóm đã xây dựng ứng dụng web để hiện thực cho hệ thống mà nhóm đã xây dựng. Ứng dụng là tổng hợp từ 2 hệ thống xử lý offline và online mà nhóm đã đề cập ở trên.



Hình III-5: Giao diện demo về hệ thống nhận diện bất thường trong giao dịch thẻ tín dụng theo thời gian thực

IV. THỰC NGHIỆM VÀ KẾT QUẢ

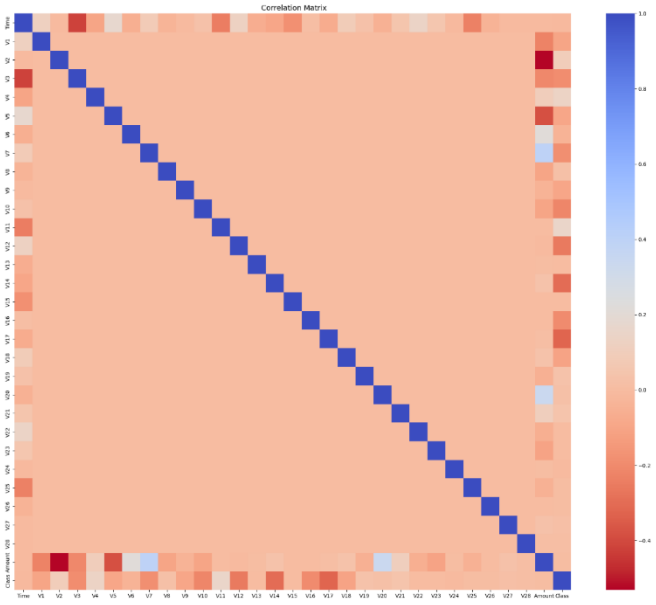
A. *Bộ dữ liệu*

Bộ dữ liệu được lựa chọn là Credit Card Fraud Detection từ thư viện dữ liệu Kaggle (đính kèm đường dẫn hyperlink tại đây) của tác giả Machine Learning Group.

Bộ dữ liệu gồm các giao dịch sử dụng thẻ tín dụng được lấy trong 2 ngày vào tháng 9 năm 2013 bởi các chủ thẻ tín dụng ở Châu Âu. Bộ dữ liệu gồm có 31 cột và 284 807 dòng dữ liệu. Cụ thể, có 30 cột là feature của dữ liệu gồm cột dữ liệu "Time" chứa số giây trôi qua giữa giao dịch tiếp theo của từng dòng dữ liệu so với giao dịch đầu tiên và cột "Amount" cho biết giá trị giao dịch được thực hiện. Còn lại là 28 cột dữ liệu đã được mã hóa bởi tác giả với tên gọi lần lượt là V1, V2, V3,..., V27, V28; đây là các dữ liệu đã được biến đổi qua phép biến đổi PCA. Theo thông tin từ bộ dữ liệu, vì lý do bản quyền và quyền riêng tư của bên cung cấp dữ liệu mà tác giả cung cấp 28 cột này thay vì dữ liệu ban đầu. Đây cũng là trở ngại lớn cho nhóm nếu muốn thực hiện các hướng phát triển mới cho bài toán khi sử dụng bộ dữ liệu này. Bên cạnh đó, gồm 1 cột target thể hiện phân loại giao dịch là gian lận (Fraud) hay không với mã hóa là 1 – Fraud / Gian lận và 0 – No Fraud / Không gian lận.

Tập dữ liệu có sự mất cân bằng lớn về số nhãn dữ liệu. Đây là điều dễ thấy khi trong thực tế số lượng giao dịch mang dấu hiệu gian lận thường rất nhỏ so với các giao dịch bình thường. Thống kê về số nhãn dữ liệu của bộ dữ liệu này cho thấy, có 492 vụ gian lận trong số 284.807 giao dịch, chiếm 0,172% tổng số giao dịch. Còn lại là 284 315 dữ liệu mang nhãn 0 - No Fraud, chiếm 99.83% tổng số dữ liệu.

Xét trên ma trận tương quan, cho thấy không có sự tương quan cao giữa các thuộc tính trong bộ dữ liệu. Vì vậy nhóm sẽ ưu tiên việc không lựa chọn thuộc tính mà sử dụng đầy đủ các thuộc tính của bộ dữ liệu cho việc huấn luyện mô hình.



Hình IV-1: Ma trận tương quan của bộ dữ liệu Credit Card Transaction

B. *Độ đo đánh giá*

Các độ đo được nhóm sử dụng để đánh giá hiệu năng của mô hình bao gồm: Macro Precision, Macro Recall, Macro F1, AUPRC (Area under the precision-recall curve) và AUC-ROC score (Area under the ROC curve).

1) Độ đo Precision, Recall, F1

Công thức của các độ đo Macro Precision, Macro Recall và Macro F1 score:

- Macro Precision:

$$\text{Macro Precision} = \frac{\sum_{i=0}^{n} \text{Precision}_i}{n} \text{ với n là số class} \quad (1)$$

- Macro Recall:

$$\text{Macro Recall} = \frac{\sum_{i=0}^{n} \text{Recall}_i}{n} \text{ với n là số class} \quad (2)$$

- Macro F1 score:

$$\text{Macro F1 score} = \frac{\sum_{i=0}^{n} \text{F1}_i}{n} \text{ với n là số class} \quad (3)$$

Các độ đo Precision, Recall, F1 là các độ đo được dùng cho bài toán phân loại (Classification). Với bài toán này, mặc dù có sự mất cân bằng lớn về dữ liệu, nhưng nhóm đã quyết định lựa chọn cách tính Macro trên 3 chỉ số này, thay vì Weighted. Lý do cho lựa chọn này nằm ở vấn đề, nhóm muốn đánh giá công bằng giữa 2 nhãn 0 và 1 mà không quan tâm đến số lượng dữ liệu tương ứng của 2 nhãn này; đồng thời cũng không che giấu hiệu suất chưa tốt của nhãn 1 – Fraud, có số lượng nhãn ít hơn. Để đánh giá hiệu suất của mô hình trên dự đoán các dữ liệu của nhãn dữ liệu 1 – Fraud, nhóm đã ghi nhận và so sánh kết quả của ba độ đo này trên class 1.

2) Độ đo AUPRC

AUPRC (Area under the precision-recall curve) một độ đo hiệu suất của các mô hình phân loại, mô hình phát hiện bất thường, đặc biệt hữu ích trong các bài toán mất cân bằng dữ liệu. Độ đo này xuất phát từ Precision-Recall Curve là đồ thị biểu diễn biểu diễn mối quan hệ giữa Precision (Độ chính xác) và Recall (Độ nhạy) ở các ngưỡng phân loại khác nhau.

Sau đó, thực hiện tính AUC (Diện tích phía dưới đường cong) của PR Curve thì ta thu được AUPRC.

Bởi vì độ đo này không sử dụng giá trị True Negative, do đó AUPRC sẽ không bị ảnh hưởng bởi số lượng lớn giá trị AUPRC trong dự đoán. Đây là đặc điểm cực kỳ phù hợp với các bộ dữ liệu có sự chênh lệch lớn về dữ liệu, ví dụ như 98% Negative và 2% Positive, vì độ đo này chỉ tập trung vào các giá trị mà mô hình dự đoán với các dữ liệu thuộc 2% Positive. Nếu mô hình xử lý tốt các trường hợp Positive, AUPRC cao và ngược lại. Do đó, đây là một độ đo phù hợp với các bộ dữ liệu có sự mất cân bằng về số lượng dữ liệu Positive. [17]

3) Độ đo AUC-ROC score

AUC-ROC score (diện tích bên dưới đường cong ROC) là thước đo quan trọng trong bài toán phân loại nhị phân thể hiện khả năng phân loại của mô hình. AUC-ROC score có khoảng giá trị từ 0 đến 1. Độ đo này càng cao thì mô hình càng tốt [18].

Độ đo này được xác định từ 2 thành phần là AUC và ROC. Trong đó, đường cong ROC (Receiver Operating Characteristic) là biểu đồ thể hiện hiệu suất của mô hình phân loại tại các giá trị ngưỡng khác nhau. Đường cong này thể hiện mối quan hệ giữa hai tham số: TPR (True Positive Rate) và FPR (False Positive Rate) tại các ngưỡng khác nhau [19].

AUC: Diện tích dưới đường cong ROC, một giá trị từ 0 đến 1, đại diện cho khả năng phân biệt giữa các lớp của mô hình.

- AUC = 1: Mô hình hoàn hảo.
- AUC = 0.5: Mô hình không có khả năng phân biệt, tương đương với đoán ngẫu nhiên.
- AUC < 0.5: Mô hình có hiệu suất kém hơn so với đoán ngẫu nhiên.

C. Kết quả

Nhóm đã thực hiện các thử nghiệm với các pipeline đã nêu trên và đưa ra các kết quả dựa trên các kết quả mà nhóm đã nêu. Trong đó, phần đầu là trình bày kết quả Macro Precision, Macro Recall, Macro F1, AUPRC và AUC-ROC trên các pipeline nhóm triển khai. Tiếp theo là kết quả của các độ đo Precision, Recall và F1 trên class Fraud của bài toán

a) Kết quả trên độ đo Macro Precision, Macro Recall, Macro F1, AUPRC, AUC-ROC

Với bảng kết quả IV-1, đây là phần so sánh giữa hiệu năng tổng quát của mô hình, tức thể hiện kết quả phân lớp tổng quát của cả class 0 và class 1 của từng pipeline với dữ liệu kiểm thử của nhóm. Các kết quả so sánh ở đây gồm Macro Precision, Macro Recall và Macro F1 score. Với 2 độ đo AUPRC và AUC-ROC, nhóm dùng chỉ số này bên cạnh so sánh giữa các pipeline mà nhóm tiến hành mà còn so sánh với các kết quả nghiên cứu trước đó mà nhóm đã tìm được.

Bảng IV-1: Bảng kết quả các độ đo Macro Precision, Macro Recall, Macro F1, AUPRC, AUC-ROC pipeline nhóm xây dựng

Pipeline	Macro Precision	Macro Recall	Macro F1	AUPRC	AUC-ROC
Baseline	0.95	0.83	0.88	0.72	0.96
Pipeline 1	0.76	0.90	0.81	0.72	0.98
Pipeline 2	0.84	0.91	0.87	0.74	0.98
Pipeline 3	0.63	0.91	0.70	0.73	0.99
Pipeline 4	0.83	0.91	0.86	0.72	0.98
Pipeline 5	0.93	0.92	0.93	0.91	0.99
Pipeline 6	0.93	0.92	0.93	0.78	0.98
Pipeline 7	0.89	0.87	0.88	0.71	0.91
Pipeline 8	0.97	0.89	0.93	0.73	0.98
Pipeline 9	0.53	0.94	0.56	0.73	0.99
Pipeline 10	0.61	0.92	0.68	0.73	0.98

b) Độ đo Precision, Recall và F1 trên class 1 - Fraud

Bảng IV-2 thực hiện xem xét các chỉ số Precision, Recall và F1 score trên lớp 1, là lớp thể hiện các giao dịch có dấu hiệu gian lận. Đây là cũng là một kết quả quan trọng mà nhóm muốn xem xét và so sánh hiệu quả của từng pipeline lên class 1 là như thế nào.

Bảng IV-2: Bảng kết quả độ đo Precision, Recall và F1 score trên class 1 của pipeline thử nghiệm

Pipeline	Precision 1	Recall 1	F1 1
Baseline	0.90	0.65	0.76
Pipeline 1	0.52	0.80	0.63
Pipeline 2	0.68	0.81	0.74
Pipeline 3	0.26	0.82	0.39
Pipeline 4	0.65	0.81	0.72
Pipeline 5	0.86	0.85	0.85
Pipeline 6	0.85	0.78	0.81
Pipeline 7	0.79	0.74	0.76
Pipeline 8	0.96	0.77	0.85
Pipeline 9	0.07	0.91	0.12
Pipeline 10	0.23	0.84	0.36

c) Nhận xét

Nhìn chung, các mô hình của nhóm thực nghiệm được lựa chọn ở mức đơn giản, dễ cài đặt nhưng kết quả đánh giá tương đối tốt. Với thông số AUC-ROC đều đạt sấp sĩ mức 0.9, cho thấy các mô hình có thể phân loại các dữ liệu tốt. Tuy nhiên AUPRC lại cho kết quả khoảng 0.7 đến 0.8 ở hầu hết các pipeline thực hiện, cho thấy được khả năng phân loại các giao dịch bất thường ở các model vẫn còn chưa tốt. So sánh với các nghiên cứu trước, có thể nhận thấy kết quả AUPRC của các model nhóm xây dựng cho kết quả tốt hơn so với kết quả của các bài báo trước; trong đó có pipeline 6, cho kết quả AUPRC cao nhất là 0.91. Với AUC-ROC, kết quả mà nhóm xây dựng được cũng gần với kết quả mà bài báo đã công bố.

Đầu tiên, đề cập đến baseline model thực hiện với dữ liệu không qua xử lý, có thể thấy mô hình baseline đã có kết quả khá gần so với kết quả tốt nhất qua nghiên cứu trước đó. Bên cạnh đó, với baseline model cho kết quả phân lớp trên class 1 tương đối tốt, với precision cao và F1 có chỉ số tốt, tuy nhiên recall chỉ ở mức khá tốt.

Thứ hai, với các pipeline chỉ dùng tiền xử lý RobustScaler (gồm pipeline từ 5 đến 9), có thể thấy, hầu hết các kết quả đều cho thấy tốt hơn so với phương pháp baseline. Trong đó, nổi bật là 3 mô hình DL và XGBoost là cho các kết quả tốt nhất khi so sánh kết quả độ đo so với tất cả các pipeline còn lại. Trên class 1, XGBoost cho kết quả tốt nhất ở Precision, F1 (lần lượt là 0.96 – 0.85), tiếp đến là model DL1 (0.86 – 0.85), DL2 (0.85 – 0.85) và DL3 (0.85 – 0.81). Cho thấy hiệu quả tốt của bước chuẩn hóa đến dự đoán của mô hình.

Thứ ba, ta đề cập đến pipeline có thực hiện tiền xử lý RobustScaler + OverSampling (SMOTE library), có kết quả thấp nhất, thấp hơn khá nhiều so với baseline. 2 pipeline đều phản ánh kết quả recall tốt, tuy nhiên kết quả precision và F1 đều thấp hơn khá nhiều với baseline. Khi xem kết quả của precision và F1 trên class 1, các pipeline này đều thể hiện kết quả nghèo nàn, với pipeline 10: precision trên class 1 là 0.07, F1 là 0.12; với pipeline 11: class 1 có precision là 0.23 và F1 là 0.36.

Thứ tư, với pipeline có sử dụng phương pháp tiền xử lý Robust Scaler và Oversampling duplicating minority class thì

cách tiền xử lý này không cho kết quả quá tốt với mô hình sử dụng Logistic Regression, với Macro Precision là 0.63 và Macro F1 là 0.70, Precision trên class 1 chỉ đạt 0.26 và F1 score trên class 1 là 0.39. Với mô hình Random Forest cho kết quả tốt hơn Logistic Regression (lần lượt là 0.83 – 0.86 – 0.65 – 0.72) nhưng vẫn thấp hơn kết quả baseline.

Thứ năm, với pipeline sử dụng phương pháp Robust Scaler và Undersampling, so với baseline thì mô hình Logistic Regression cho kết quả thấp hơn. Ngược lại, với Random Forest đã có sự cải thiện ở chỉ số Recall nhưng Precision và F1 lại thấp hơn. Chiều hướng tương tự với kết quả trên class 1 của model Random Forest so với Baseline.

Như vậy, có thể thấy rằng kết quả của các pipeline sử dụng phương pháp Resampling đang chưa đem lại hiệu quả tốt nếu so sánh với baseline ban đầu. Trong khi đó, Robust Scaler cho kết quả tác động tích cực hơn, khi cải thiện được kết quả so với baseline đã triển khai. Nhóm cũng nhận thấy rằng với các mô hình Neural Network và Ensemble Learning (gồm XGBoost và Random Forest) với ưu điểm của từng loại mô hình đã có kết quả tốt hơn so với baseline ban đầu.

Ngoài ra, các mô hình trên chỉ dùng với các cài đặt tham số mặc định. Nhóm chưa thực hiện tìm các bộ tham số tối ưu cho các mô hình cũng như các kiến trúc tối ưu cho bài toán.

V. KẾT LUẬN

A. *Đánh giá*

Từ kết quả mà nhóm ghi nhận được, nhóm rút ra các đánh giá các ưu điểm và nhược điểm của hệ thống nhóm xây dựng được như sau

1) Ưu điểm

- Kết quả của nhóm tốt và không chênh lệch so với các bài báo.
- Nhóm có thử qua nhiều cách xử lý dữ liệu mất cân bằng và sử dụng đa dạng độ đo cho việc đánh giá để đưa ra mô hình tốt nhất.
- Nhóm có thử qua nhiều mô hình và các mô hình học sâu thu được kết quả thực nghiệm khá tốt. Các mô hình có thời gian huấn luyện không dài và phù hợp với tài nguyên mà nhóm có.
- Áp dụng framework BigData Pyspark cho bài toán của nhóm và thực hiện data streaming khi thực hiện demo để xử lý data theo time series, làm cho hệ thống có thể đi gần với thực tế của bài toán là phát hiện giao lận trong các giao dịch theo thời gian thực.
- Áp dụng kiến thức và đóng gói Docker để ứng dụng có thể đóng gói và chạy trên nhiều máy khác nhau.

2) Nhược điểm

- Dữ liệu đã bị biến đổi, gây hạn chế trong việc tìm hiểu mối quan hệ giữa thuộc tính và xây dựng mô hình
- Do vấn đề về dữ liệu nên chưa áp dụng các mô hình tốt nhất của bài toán này như mô hình GNN, tận dụng được đồ thị để biểu diễn được mối quan hệ giữa các chủ thể trong giao dịch. Hay mô hình RNN để tận dụng xử lý dữ liệu chuỗi thời gian (Time Series),....
- Vẫn còn hạn chế và chưa đa dạng ở các bước tiền xử lý dữ liệu. Hiện tại nhóm chỉ thực hiện scaler cho các thuộc tính Amount và Time cho bước tiền xử lý dữ liệu.
- Mô hình chưa được thử nghiệm tối ưu tham số.

B. *Hướng phát triển*

Từ những đánh giá về ưu điểm và nhược điểm mà nhóm rút ra ở trên, nhóm đưa ra một số hướng cải thiện, phát triển cho hệ thống mà nhóm đã xây dựng, cụ thể như sau:

- Sử dụng các bộ dữ liệu khác, không thực hiện biến đổi để thực hiện xây dựng và train mô hình.
- Áp dụng mô hình SOTA của lĩnh vực này như mô hình GNN hoặc các mạng Học sâu RNN, GANs,... để cải thiện thêm hiệu suất bài toán này. Cũng như có thể áp dụng các kỹ thuật Transfer Learning để dùng các mô hình SOTA áp dụng cụ thể cho bộ dữ liệu được sử dụng
- Thực hiện thêm các thực nghiệm về tối ưu các tham số trong các mô hình trên mà nhóm đã xây dựng nhằm cải thiện hiệu suất mô hình.
- Tìm hiểu và cải tiến các biện pháp tiền xử lý dữ liệu, đặc biệt là giải quyết vấn đề mất cân bằng dữ liệu.
- Cải thiện về tính ứng dụng về hệ thống mà nhóm xây dựng, đưa hệ thống của nhóm sát với thực tế ứng dụng hiện có đang hoạt động, như sử dụng trong các giao dịch thẻ tín dụng ở các ngân hàng hiện nay.

LỜI CẢM ƠN

Để hoàn thành đồ án này, nhóm chúng em xin gửi lời cảm ơn chân thành nhất đến với trường Đại học Công nghệ Thông tin cùng khoa Khoa học và Kỹ thuật thông tin đã tạo mọi điều kiện tốt nhất để tất cả sinh viên chúng em được trải nghiệm và học tập với môn học Phân tích dữ liệu lớn. Trong xuyên suốt khoảng thời gian học tập, chúng em đã được tiếp thu không chỉ kiến thức về chuyên môn về Công nghệ thông tin, về chuyên môn lĩnh vực Dữ liệu lớn mà còn là kiến thức về kỹ năng trình bày thuyết trình và làm việc nhóm. Đây là những kiến thức vô cùng quan trọng và cần thiết để hoàn thành đồ án và làm hành trang của chúng em cho công việc trong tương lai.

Tiếp đến, là lời cảm ơn sâu sắc đến với TS. Đỗ Trọng Hợp, giảng viên giảng dạy lý thuyết môn này và hướng dẫn thực hành và đồ án lần này của nhóm. Dưới sự hướng dẫn, định hướng về chuyên môn cùng với những lời góp ý sâu sắc và giúp đỡ tận tình của thầy đã giúp sức rất nhiều cho nhóm chúng em để hoàn thành đồ án môn học lần này. Cùng với đó, qua quá trình làm việc và học tập với thầy, chúng em cũng được thu nạp thêm những tri thức mới về lĩnh vực này cũng như các kiến thức về nghiên cứu các bài báo khoa học.

Nhóm chúng em đã dành nhiều thời gian để cố gắng tìm hiểu và hoàn thành đồ án môn học tốt nhất trong khả năng trong khả năng nhóm chúng em. Dẫu vậy, khó có thể những thiếu sót vì những thiếu sót nhất định về kiến thức, kỹ năng. Vì thế, qua đồ án này, nhóm em cũng xin lắng nghe những góp ý bổ sung từ thầy để nhóm có thể hoàn thiện hơn ở đề tài này.

PHÂN CÔNG CÔNG VIỆC

Họ và tên	Công việc	Mức độ hoàn thành
Tăng Minh Hiển (21520229)	Tham gia xây dựng bài toán và lựa chọn bộ dữ liệu Thực hiện thử nghiệm Pipeline 1, 2, 3, 4, 5, 7, 9, 10 Ghi nhận các thử nghiệm của bài toán, sửa lỗi báo cáo.	100
Châu Thiên Long (21520331)	Tham gia xây dựng bài toán và lựa chọn bộ dữ liệu	100

	Thực hiện thực nghiệm Pipeline 8 Demo streaming với kafka với đóng gói docker, trang web demo.	
Nguyễn Thái Thành Long (21520334)	Tham gia xây dựng bài toán và lựa chọn bộ dữ liệu Xây dựng mô hình Baseline, Pipeline 6 Ghi nhận thử nghiệm và viết báo cáo, slide	100

VI. TÀI LIỆU THAM KHẢO

[1] "Ngân hàng TNHH Một thành viên HSBC (Việt Nam)," Cách sử dụng thẻ tín dụng, [Online]. Available: https://www.hsbc.com.vn/credit-cards/how-do-credit-cards-work/. [Accessed 26 06 2024].

[2] C. O. S. Research, "Number of Credit Card Transactions per Second, Day & Year," 2024.

[3] M. Trúc, "VietStock," Pháp luật TPHCM, 22 05 2024. [Online]. Available: https://vietstock.vn/2024/05/giao-dich-the-tin-dung-tang-cao-757-1192001.htm. [Accessed 05 07 2024].

[4] M. Rezapour, "Anomaly Detection using Unsupervised Methods: Credit Card Fraud Case Study," International Journal of Advanced Computer Science and Applications (IJACSA) , 2019.

[5] Bolton, Richard J and Hand, David J, "Statistical fraud detection: A review," *Statistical science*, vol. 17, no. 3, pp. 235-255, 2002.

[6] Bhattacharyya, Siddhartha and Jha, Sanjeev and Tharakunnel, Kurian and Westland, J Christopher, "Data mining for credit card fraud: A comparative study," *Decision support systems*, vol. 50, no. 3, pp. 602-613, 2011.

[7] Chandola, Varun and Banerjee, Arindam and Kumar, Vipin, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1-58, 2009.

[8] Fiore, Ugo and De Santis, Alfredo and Perla, Francesca and Zanetti, Paolo and Palmieri, Francesco, "Using generative adversarial networks for improving classification effectiveness in credit card fraud detection," *Information Sciences*, vol. 479, pp. 448-455, 2019.

[9] Giulia Moschini, Régis Houssou, Jérôme Bovay, Stephan Robert-Nicoud, "Anomaly and Fraud Detection in Credit Card Transactions Using the ARIMA Mode," in *The 7th International Conference on Time Series and Forecasting*, 2021.

[10] Ounacer, Soumaya and El Bour, Hicham Ait and Oubrahim, Younes and Ghoumari, Mohamed Yassine and Azzouazi, Mohamed, "Using Isolation Forest in anomaly detection: the case of credit card transactions," *Periodicals of Engineering and Natural Sciences*, vol. 6, no. 2, pp. 394-400, 2018.

[11] S. N. a. N. K.-H. a. T. G.-K. a. K. K.-C. Kalid, "A Multiple Classifiers System for Anomaly Detection in Credit Card Data With Unbalanced and Overlapped Classes," *IEEE Access*, vol. 8, pp. 28210-28221, 2020.

[12] Shanshan Jiang, Ruiting Dong, Jie Wang and Min Xia, "Credit Card Fraud Detection Based on Unsupervised Attentional Anomaly Detection Network," *Systems*, vol. 305, p. 11, 2023.

[13] Guansong Pang, Chunhua Shen, Anton van den Hengel, "Deep Anomaly Detection with Deviation Networks," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019.

[14] Xu, Hongzuo and Pang, Guansong and Wang, Yijie and Wang, Yongjun, "Deep isolation forest for anomaly detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12591-12604, 2023.

[15] Grover, Prince and Li, Zheng and Liu, Jianbo and Zablocki, Jakub and Zhou, Hao and Xu, Julia and Cheng, Anqi, "Fdb: Fraud dataset benchmark," *arXiv e-prints*, pp. arXiv--2208, 2022.

[16] A. Singh, "Analytics Vidhya," KNN algorithm: Introduction to K-Nearest Neighbors Algorithm for Regression, 13 2 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/.

[17] P. R. D. MD, "Glass Box," Machine Learning and Medicine - Measuring Performance: AUPRC and Average Precision, 2 March 2019. [Online]. Available: https://glassboxmedicine.com/2019/03/02/measuring-performance-auprc/#:~:text=The%20area%20under%20the%20precision,about%20finding%20the%20positive%20examples.. [Accessed 01 July 2024].

[18] E. A. Team, "EVIDENTLY AI," How to explain the ROC curve and ROC AUC score?, [Online]. Available: https://www.evidentlyai.com/classification-metrics/explain-roc-curve#:~:text=The%20ROC%20AUC%20score%20is%20the%20area%20under%20the%20ROC,and%201%20indicates%20perfect%20performance. [Accessed 30 06 2024].

[19] G. Developers, "Google Developer - Machine Learning," Google, 18 07 2022. [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=vi#:~:text=An%20ROC%20curve%20(receiver%20operating,False%20Positive%20Rate. [Accessed 30 6 2024].