Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.32, NO.1, Feb. 2022 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2022.32.1.25

X-FDS : 게임 결제 로그 기반 XAI적용 이상 거래탐지 모델 연구

이 영 헌,[1†] 김 휘 강[2‡]
[1,2]고려대학교 정보보호대학원 (대학원생, 교수)

Why Should I Ban You! : X-FDS (Explainable FDS) Model Based on Online Game Payment Log

Young Hun Lee,[1†] Huy Kang Kim[2‡]
[1,2]School of Cybersecurity, Korea University (Graduate student, Professor)

요 약

게임에 대한 결제 수단과 방식이 다양해지는 가운데, 관련된 금융사고가 이용자와 게임사에 심각한 문제를 야기하고 있다. 최근 게임 결제 시스템에 대해 게임사는 이상거래탐지시스템(FDS)을 도입하여 금융 사고를 방지하고 있다. 하지만, FDS는 지속적으로 탐지 패턴을 변경해야 하므로 효과적이지 않고 판단 결과에 따른 근거를 제시할 수 없다. 본 논문에서는 실제 게임회사의 결제 로그 데이터 중 이상거래를 분석하여 관련된 피처를 생성하였다. 비지도 학습 모델중 하나인 오토인코더를 사용하여 이상거래를 탐지하는 모델을 구축하였으며, 그 결과 85% 이상의 정확도를 얻을 수 있었다. 이를 XAI-SHAP을 적용한 X-FDS를 사용하여 이상 거래탐지에 대한 영향력이 가장 높은 피처는 나라, 거래 금액과 거래 매체, 이용자의 나이임을 알 수 있었다. 이를 바탕으로 제시한 모델의 판단 결과에 편향성을 주는 피처에 가중치를 세부 조정하여 최종적으로 정확도 94%의 개선된 탐지 모델을 도출하였다.

ABSTRACT

With the diversification of payment methods and games, related financial accidents are causing serious problems for users and game companies. Recently, game companies have introduced an Fraud Detection System (FDS) for game payment systems to prevent financial incident. However, FDS is ineffective and cannot provide major evidence based on judgment results, as it requires constant change of detection patterns. In this paper, we analyze abnormal transactions among payment log data of real game companies to generate related features. One of the unsupervised learning models, Autoencoder, was used to build a model to detect abnormal transactions, which resulted in over 85% accuracy. Using X-FDS (Explainable FDS) with XAI-SHAP, we could understand that the variables with the highest explanation for anomaly detection were the amount of transaction, transaction medium, and the age of users. Based on X-FDS, we derive an improved detection model with an accuracy of 94% was finally derived by fine-tuning the importance of features that adversely affect the proposed model.

Keywords: FDS, Machine Learning Algorithm, XAI

Received(11. 17. 2021), Modified(1st: 12. 03. 2021, 2nd: 12. 28. 2021), Accepted(01. 04. 2022)

† 주저자, dlddudfkr@korea.ac.kr
‡ 교신저자, cenda@korea.ac.kr(Corresponding author)

I. 서 론

글로벌 게임 시장은 2021년 약 28억 명의 플레이어가 1,890억 달러를 벌어들이는 등 계속 성장할 것으로 예상된다[1]. 그리고 다양한 결제 수단이 등장하면서 국가에 상관없이 온라인게임 콘텐츠에 대한 결제가 가능해졌다. 온라인게임 결제 시스템의 경우 현실과 달리 결제와 동시에 이용자에게 직접 서비스를 제공하거나 가상의 상품이나 아이템으로 빠르게 교환할 수 있다. 공격자는 이러한 부분을 악용하여 게임 아이템을 구매하거나 비정상적인 결제를 통해 금전적 이익을 얻는다. 온라인게임에서 이상 거래가 발생하더라도 사용자가 접속하기 전까지 알아차리기 어렵다. 또한, 게임 제작사, 결제사, 통신사 등 다양한 기업의 이해관계로 인해 이상 거래에 대한 분석 및 추적이 쉽지 않다. 게임 제작사는 지속해서 발생하는 온라인게임 서비스 금융사기를 방지하기 위해 결제시스템에 FDS (Fraud Detection System)를 도입하고 있다. FDS는 결제 또는 거래 중 발생하는 로그 분석을 통해 이상 거래를 탐지하는 시스템을 말한다. 이상 거래로 판단되는 이용자는 온라인게임 서비스 중단, 법적 조치 등의 제재를 받는다.

MMORPG의 특성상 많은 사람이 참여하면서 현금거래에 대한 사기가 발생하고 있다. 이에 대해 적절히 대응하지 못하면 사용자는 불만을 느끼고 떠나게 되면서 점차 게임 산업 발전에 악영향을 주게 된다. 특히 게임 내 금융사고의 경우 금액이 소액인 경우가 많고, 즉시 인지하기 힘들어 피해가 발생하더라도 탐지에 어려움을 겪고 있다[2].

또한, 핀테크가 발전하면서 다양한 거래 수단을 통해 이상 거래 행위를 시도하고 있는 사례가 늘고 있다[3]. 이에 대한 피해를 최소화하기 위한 신속하고 정확한 대응책의 필요성이 대두되고 있다. 이상 거래 행위에 대한 제재나 법적 조치가 진행될 경우, 정확한 판단을 위해 행위자에게 정확한 근거를 제시할 수 있어야 한다.

현재까지의 게임 산업의 FDS 모델의 경우 관리자가 지속적인 수정 및 추가적인 rule 설정을 통해 이상 거래를 분류하고 있다. 하지만, 많은 양의 데이터를 기반으로 정확성, 신속성 등이 중시되고 있으므로 점차 탐지 과정의 자동화가 중요해지고 있다. 따라서 본 논문에서는 딥러닝을 적용한 이상 거래탐지 모델을 구현하고 설명 가능한 인공지능 모델을 적용했다. 이후, 탐지 모델 결과에 대한 근거와 다양한 이상 결제 패턴에 영향을 미치는 피처를 시각적 설명력을 바탕으로 모델을 개선했다.

본 연구는 2장에서 다양한 산업의 금융거래와 관련된 이상 거래탐지시스템과 관련한 연구에 기술하였다. 3장에서는 데이터를 이상 거래탐지 모델에 적용하기 위해 딥러닝과 설명 가능한 인공지능 모델을 분석하여 활용하는 방법론을 제시하였다. 4장에서는 실제 온라인 게임 결제 데이터에 관해 설명하고, 비지도학습인 오토인코더를 통한 이상 거래탐지 모델을 구현했다. 이후, XAI 알고리즘이 적용된 X-FDS를 사용하여 생성된 예측 모델에 대한 특징을 분석했다. 해당 결과를 바탕으로 탐지 모델을 개선하고 기존 FDS에 사용되는 XGBoost, LightGBM 모델과 비교분석을 통해 성능을 평가했다.

II. 관련 연구

최근 FDS는 인공지능 모델을 사용하며 이상 거래를 탐지하는 연구가 진행되고 있다. 여러 산업에서 사용되는 FDS 중 대표적으로 금융권에서 사용되고 있는 FDS를 탐지 방법, 데이터 등 정보를 Table 1.에 정리하였다.

본 절에서는 연구와 관련된 주요 FDS의 기존 모델 분석 및 탐지 방법과 설명가능 인공지능 모델에 관해 설명한다.

Table 1. Previous research of FDS

Category	Detection Method	Technique	Data
Finance [4]-[7]	Rule based	Score & Rule	Credit card Fraud
	Supervised learning	XGBoost	
		LightGBM	
		Random Forest	Stock Market Fraud
	Unsupervised learning	Auto-encoder	Credit card Fraud

2.1 지도학습 기반 이상 거래 탐지시스템

기존 FDS에 관한 연구에서는 레이블이 포함된 데이터를 사용하여 이상 거래를 탐지하는 방법이 주로 사용되었다. 초기의 FDS 모델에서는 통계 알고리즘을 적용하여 의사결정트리, 랜덤 포레스트와 같

은 지도학습 모델을 통해 이상 패턴을 기반으로 탐지했다[8]. 지도학습은 분류 또는 회귀 알고리즘을 통해 결과를 예측한다. 학습 결과로 도출된 값들은 각 데이터를 구분하는 '분류'로 활용할 수 있고 입력값과 출력값 간의 일반적인 관계적 특성을 도출한다. 특정 조건을 충족하거나 적절한 범주 규칙을 사용하기 위해 회귀 모델을 사용할 수 있다.

하지만 기존 모델의 경우 데이터 마이닝 기술을 통해 탐지 방법을 개선하는 모델을 도출하는 공통점이 있다. 일반적으로 머신러닝 알고리즘의 정확한 탐지 성능을 위해서는 많은 양의 레이블 처리된 학습 데이터를 준비해야 한다[9]. 레이블을 처리하기 위해 담당자가 많은 양의 데이터를 수기로 처리해야 하므로 시간과 노력이 필요하며 비대칭적인 데이터를 활용하게 된다. 이러한 한계를 극복하기 위해 정상 데이터를 활용하여 학습할 수 있는 비지도학습 모델을 이용한 연구가 진행되고 있다.

2.2 비지도학습 기반 이상 거래 탐지시스템

지도학습 기반 이상 거래 탐지시스템의 모델을 생성하기 위한 데이터 라벨링과 이상 거래 데이터 수집의 한계에 따라 사람의 개입 없이 결과를 분류 할 수 있는 연구가 진행되었다. 비지도학습을 적용한 FDS 모델은 데이터에 군집화를 진행하여 K-Means, 오토인코더 등 알고리즘을 활용해 탐지를 진행했다. 비지도학습 기반 FDS는 지도학습과 달리 이상 거래로 판정된 데이터를 학습하지 않는다. 입력된 데이터에 대해 군집화를 진행하여 그룹을 나누고 이상, 정상 거래를 판단하게 된다. 금융거래에서 정상 거래가 많이 발생하여 정상거래 그룹이 생성되었을 때, 새로 입력된 이상 거래 데이터와 수치상으로 일정한 차이가 발생하게 된다. 각 그룹의 차이에 대한 값을 비교하여 이상 거래로 최종 예측할 수 있다.

비지도 학습 모델이 적용된 FDS 연구로는 클러스터링 알고리즘인 DPC(Density Peak Clustering)를 적용하여 노드의 중심으로부터의 밀도를 측정하는 방법을 적용했다. 군집화된 영역을 식별하고 분석하여 의료 보험에서 발생하는 이상거래를 탐지하는 모델을 제안하였다[10].

다만, 현재까지 제안된 FDS에 대한 비지도학습 모델은 정상 거래가 상대적으로 많다는 가정을 통해 학습을 진행하므로 해당 가정이 틀리면 오탐이 급격하게 늘어나는 문제가 발생한다.

2.3 XAI

많은 분야에서 인공지능 모델을 활용하여 문제를 해결하는 방법이 연구되고 있으며, 이를 실생활에 적용하여 효율적인 서비스가 제공되고 있다. 하지만 기술의 발달로 인공지능 모델의 네트워크 복잡성과 처리해야 할 데이터의 양이 늘어나면서 모델에 관한 결과 해석과 신뢰성이 문제가 되기 시작했다[11].

이러한 문제를 해결하기 위해 인공지능 모델의 판단 결과를 해석할 수 있는 XAI (eXplainable AI)가 연구되고 있다. XAI는 블랙박스와 같이 학습 과정이 복잡하고 추론이 어려운 모델을 해석하기 위한 기술이다. 데이터를 기반으로 결론에 도달하는 과정을 새로운 인공지능 프로세스를 사용하여 모델을 해석할 수 있다[12]. 모델 결과 설명을 위해 민감도 분석, 개별 조건 예측, 부분 종속 구성과 같은 기존 모델을 추적하는 알고리즘 기술이 주로 사용된다. 수학적 계산을 통해 모델에 관한 판단 근거를 시각적으로 도출하여 해석할 수 있다. 이는 사용자가 모델을 신뢰하고 복잡성을 완화하기 위해 설명할 수 있는 인공지능 모델을 참고하여 최종 의사결정을 내릴 수 있다는 장점이 있다. 참고로, DARPA (Defense Advanced Research Projects Agency)는 2021년까지 사용자가 인공지능 내부를 이해할 수 있도록 설명 모델 및 인터페이스 프로그램을 개발하고 결과를 신뢰하고 효율적으로 작업을 수행하는 것을 목표로 하고 있다[13].

III. 비지도학습 모델 활용 FDS 방법론

이상 거래에 대한 기준을 바탕으로 정상과 이상 거래 분류를 위한 피처를 설정하고 XAI에 관해 연구하여 다음과 같은 방법론을 적용하였다.

3.1 이상 거래 데이터 피처 설정

FDS는 사용자의 정보와 정상거래 패턴을 기반으로 생성되며, 이상 거래가 발생하면 이를 탐지하는 방식이다. 정확도가 높은 FDS를 만들기 위해서는 모든 거래 데이터 중에서 합리적인 이상거래 데이터 기준과 피처의 기준이 설정이 필수적이다. 특히 온라인 게임 결제의 경우 일반 금융거래 데이터와 다른 피처를 포함하고 있으므로, 일반 금융 FDS를 그대로 도입할 경우 실제 정상거래에 대해 이상 거래로

판단하게 되는 오류를 범할 수 있다. 데이터의 피처 설정을 위한 구체적인 데이터 전처리 과정에 대해서는 4.3에서 소개한다.

3.2 비지도학습 기반 모델링

본 논문에서 이상 거래를 탐지하기 위해서 비지도 학습 모델링 방법의 하나인 오토인코더를 사용하는 것을 제안한다. 기존 FDS의 경우 지도학습으로 비정상 거래 데이터를 활용하여 이상 거래 탐지 모델을 생성하는 경우가 많다[14]. 하지만 현실에서는 비정상 거래 데이터가 매우 적으며 혼합된 거래 데이터를 사용할 경우 FDS 생성 시 왜곡될 가능성이 존재한다.

비지도학습을 사용할 경우 수집된 거래정보에 대해서 레이블을 하지 않아도 자동으로 분석하여 그룹화했다. 오토인코더는 인코더와 디코더로 구성되어 입력을 출력으로 복사하는 신경망 기반 머신러닝 모델이다[15]. Fig. 1.의 구성은 네트워크의 단일 계층을 의미하고, 숫자는 뉴런의 수를 의미한다.

레이블이 없는 데이터가 입력되더라도 인코딩 및 디코딩 과정을 통해 해당 데이터 특성을 기반으로 학습이 진행된다. 인코더는 입력을 최적화된 매개변수로 변환하여 데이터 특성을 효과적으로 추출할 수 있는 계층이다. 디코더는 매개변수를 출력으로 변환하여 원래의 입력값을 최대한 복원할 수 있는 계층으로 구성된다. 이러한 레이어의 구조는 오토인코더가 단순히 입력을 출력에 직접 복사하는 것을 방지하고 데이터를 효율적으로 표현하는 방법을 배우도록 조절한다. 오토인코더는 입력과 출력이 같은 네트워크 구조로 되어있어서 최종적으로 입력을 재구성하여 결과를 출력한다[16]. 이러한 학습을 통해 손실 값을 최소화하여 최적화된 오토인코더 모델을 생성할 수 있다. 인코더와 디코더에 대한 정확한 이해를 위해 아래 수식으로 설명할 수 있다[17].

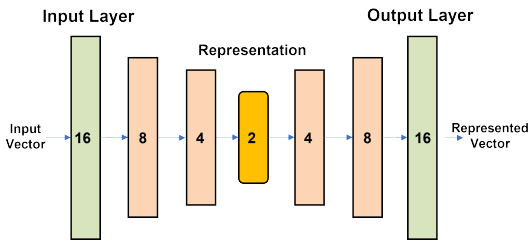

Fig. 1. Architecture of the Autoencoder

$$x \in [0,1]^D \tag{1}$$

$$e(x) = s(W_{enc}x + b_{enc}) \tag{2}$$

$$d(r) = s(W_{dec}r + b_{dec}) \tag{3}$$

$$Loss : L(x.y) = \| \| f_\theta(x) - yvert \|^2 \tag{4}$$

$$\theta^* = \arg min\ _\theta(\sum_{x \in D} L(x,y)) \tag{5}$$

- x: An input feature vector
- y: The target vector
- $e(x)$: An Encoder
- $d(x)$: A Decoder
- s: An activation function
- D: Dimension of features
- W_{enc}, W_{dec}: Weight Matrix of the encoder and the decode
- b_{enc}, b_{dec}: Biases of the encoder and the decoder
- θ: The parameters of encoder and decoder

오토인코더 모델은 정상 데이터를 기반으로 학습을 진행하여 모델 최적화를 진행한다. 비지도 학습 모델의 효율적인 학습을 위해 이상 탐지 데이터는 제외하고 입력한다. 오토인코더 모델이 입력받은 매개변수를 최적화하여 추가로 입력되는 정상 데이터에 대해서는 재구성 손실 값이 최소화된다. 하지만, 다른 데이터에 대해서는 최적화되어 있지 않기 때문에 재구성 수식 (4)를 통해 입력과 출력에 대한 재구성 손실값이 정상 데이터보다 높게 나오게 된다.

생성된 오토인코더 모델의 입력 데이터에서 비정상 데이터를 식별하는 분류 기준을 만들었다. 학습 모델에서 도출된 손실 값을 기준으로 입력되는 데이터가 비정상 거래인지 구분한다. 여러 실험의 결과를 통해 이상 거래 탐지 성능을 극대화하기 위해 임계값을 설정한다. 해당 임계값 설정을 통해 유효성을 검사하여 데이터를 정상과 이상 거래를 효과적으로 판별할 수 있다[18].

3.3 XAI 모델링

앞서 생성된 이상 거래 탐지시스템에 대하여 SHAP 프레임워크를 사용한다. SHAP은 게임이론[19]에서 도출된 Shapley Value와 피처 간 독립성을 기반하여 만들어졌다[20]. Shapley Value는 하나의 변수의 중요성을 알기 위해 여러 변수의 조합을 구성한 다음 변수의 유무에 따른 평균 변화를 구함으로써 얻을 수 있다. 이러한 방식으로 모델에 대한 각 기능의 기여도를 수치로 표현할 수 있다. 생성된 모델에 대해 피처의 Shapley Value를 구하는 수식을 표현하면 다음과 같다.

$$\chi_i(v) = \sum_{S \in N_i} \frac{|S|!(n-|S|-1)!}{n!}(v(S \cup \{i\}) - v(S)) \quad (6)$$

- $\chi_i(v)$: Shapley Value of the i^{th} data
- n: Total number of data
- S: All data except i^{th} data
- $v(S)$: The contribution of the remaining subsets to the value except for the i^{th} data
- $(v(S \cup \{i\}) - v(S))$: Total contribution including i^{th} data

SHAP은 모델에 대한 출력을 각 피처의 기여도로 분류하며, Shapley Value는 음수일 수 있다. 이는 해당 피처가 모델 예측에 부정적인 영향을 미치고 있다고 해석할 수 있다. 단일 특성의 첨가 전후를 비교하여 값을 계산할 수 있고 모든 한계 기여도의 평균을 통해 설명 근거로 사용한다[20].

IV. 실험 설계

4.1 실험 목표

본 논문에서 제안한 이상거래 탐지 모델에 대해 해석 가능성, 탐지 성능, 불균형 데이터에 대한 효과적 학습 방안, 총 3가지 측면으로 실험 목표를 설정했으며, 검증을 위한 실험을 설계했다. 자세한 설명은 아래와 같다.

4.1.1 FDS 모델의 설명 가능성

FDS의 탐지 모델에 인간의 설명 가능성을 추가한다. 여러 XAI 방안 중 Shapley Value를 이용하여 인공지능 모델을 해석한다. 단순 탐지 결과만 도출하는 기존 FDS 모델과 달리, 본 연구의 모델은 의사결정을 위한 수치적 데이터와 시각적 데이터를 모두 도출한다. XAI를 통해 모델의 판단 결과를 이해하고 피처들이 판단에 어떤 영향을 미치는지 확인할 수 있다. 모델 및 결과 개선이 필요한 경우 해당 부분을 수정하여 신뢰성을 높일 수 있다. 모델 설명도 평가를 측정하기 위해 탐지 결과에서 특징 중요도를 보여주는 그래프를 만들고 특징이 모델에 미치는 긍정적인 영향과 부정적인 영향을 찾고 개선을 통해 성능을 최적화를 진행한다.

4.1.2 탐지 성능

생성한 비지도 학습 모델에 대한 탐지 성능을 검증한다. 주어진 데이터를 기반으로 생성한 모델이 이상 거래를 얼마나 탐지하는지 평가하기 위해 실험을 설계했다. 레이블이 지정된 데이터를 두 개의 다른 데이터 세트로 분리하여 모델이 특정 데이터에 대해 학습되지 않도록 한다.

본 연구에서 활용한 비지도학습 모델인 오토인코더의 탐지 성능 확인을 위해 동일한 비율의 정상, 이상 데이터를 가지고 지도학습 모델인 XGBoost, LightGBM과 성능 비교를 진행한다. 해당 모델들은 FDS에 많이 사용되는 지도학습 모델이며, 이진 트리로 구성된 구조로 노드마다 최적의 임계값을 결정하여 패턴을 일반화하여 이상 거래를이 탐지할 수 있다. 지도학습 모델 중에서 불균형한 데이터에 대해 좋은 성능을 보이고, 학습속도가 빠른 장점이 있다. 이후 비지도학습 모델과 XAI를 적용하여 모델 튜닝을 진행한 X-FDS의 성능 비교를 통해 개선된 성능 결과를 최종 도출한다.

4.1.3 불균형 데이터의 효과적 학습

마지막으로, 불균형 데이터에 대한 활용 및 학습 방안을 제시한다. 기존 탐지 모델의 경우 많은 양의 레이블 처리된 데이터를 기반으로 하고 있지만, 실제 데이터는 불균형 문제를 가지고 있다. 본 연구에서는 실제 게임 내 정상 및 이상 거래에 대한 불균형 데이

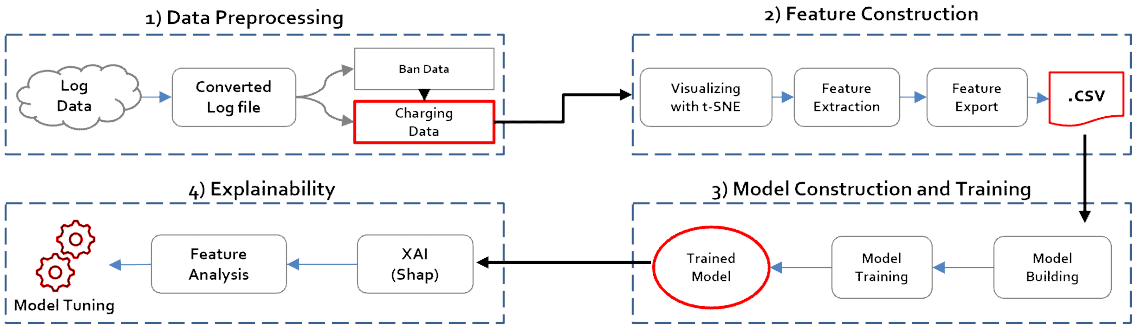

Fig. 2. Proposed Method (X-FDS)

터 세트를 사용하여 효과적으로 탐지할 수 있는 모델을 제시한다. 대량의 정상 거래 데이터를 훈련시킨 다음 이상 거래에 대해 탐지 성능을 유지하는지 검증하는 실험을 설계했다. 이전 연구에서 주로 활용된 XGBoost, LightGBM 모델을 통해 제안한 모델과 성능을 비교했다.

실험을 진행하기 위한 과정은 Fig. 2.와 같다. 크게 데이터 전처리, 피처 선정, 탐지 모델 생성, XAI 적용 및 모델 튜닝으로 나누어 진행된다.

4.2 데이터 셋

본 연구는 국내 온라인 게임 회사인 N사로부터 받은 다양한 MMORPG 게임 데이터를 활용하여 실험이 진행되었다.

Table 2. Description of Payment Log

Dataset	
Object	Payment Log of the MMORPG 'N'
Date	2014.11.01.~2015.05.01.
Feature	Account ID (GUID)
	Customer ID
	Login ID (Email)
	Payment Date
	Payment Item
	Count
	Total Amount
	Payment Methods
	IP Address
Size	899 MB
Number of Benign Users	6,923,342
Number of Banned Users	655,171

MMORPG는 대규모의 사용자가 온라인을 통해 캐릭터를 생성하고 경쟁을 통해 성장하는 구조로 되어있다. 이용자는 게임 내 여러 활동을 통해 얻은 게임 재화를 사용이 필수적이다. 게임 재화를 구매하거나 편의 서비스를 이용하기 위하여 사이버머니를 구매하는 행위를 '충전'이라고 하며, 이를 통해 빠르게 캐릭터를 성장시켜 경쟁의 우위를 차지할 수 있다. 본 연구에서는 이상 거래 탐지를 위하여 충전 기록을 이용하며, 정상과 이상 거래 충전 기록으로 나누어져 있다. Table 2.는 실험에 사용한 충전 기록 데이터에 대한 정보를 나타낸다.

4.3 데이터탐색 (EDA) 및 전처리

게임 내 이상 거래에 대한 적절한 기준을 명확하게 하려면 기존 피처들을 바탕으로 추가로 생성하거나 제거하는 전처리 과정이 필요하다. 선행 연구에서 사용한 데이터를 분석하여 문자열 데이터를 숫자 벡터 공간에 매핑하여 처리하는 과정을 수행했다. 문자열 데이터의 장점은 해석 가능성을 가지고 있지만, 개인정보 보호를 위해 익명화 처리가 필요하다[21]. 또한, 문자열 데이터의 경우 학습하는 속도가 상대적으로 느리고 컴퓨팅 파워가 많이 필요하여, 빠른 탐지 결과를 도출해야 하는 본 연구에는 적합하지 않다. 따라서 문자열이 포함된 데이터 집합을 빈도에 따라 숫자로 변환하는 벡터화 프로세스를 진행했다.

시각화된 초기 데이터의 특징 분포가 정상 및 이상을 개별 특징으로 식별하기에 불충분하다는 것을 확인했다. 정상과 이상 거래 데이터의 차이점은 많은 피처에서 확인할 수 있지만, 고유한 패턴을 드러내기 위해 더 많은 전처리가 필요하다. 이러한 필요성을 충족시키기 위해 피처에 대한 데이터를 정상과 이상 거래로 나누어 라벨링을 진행했다. 모든 데이터에 대

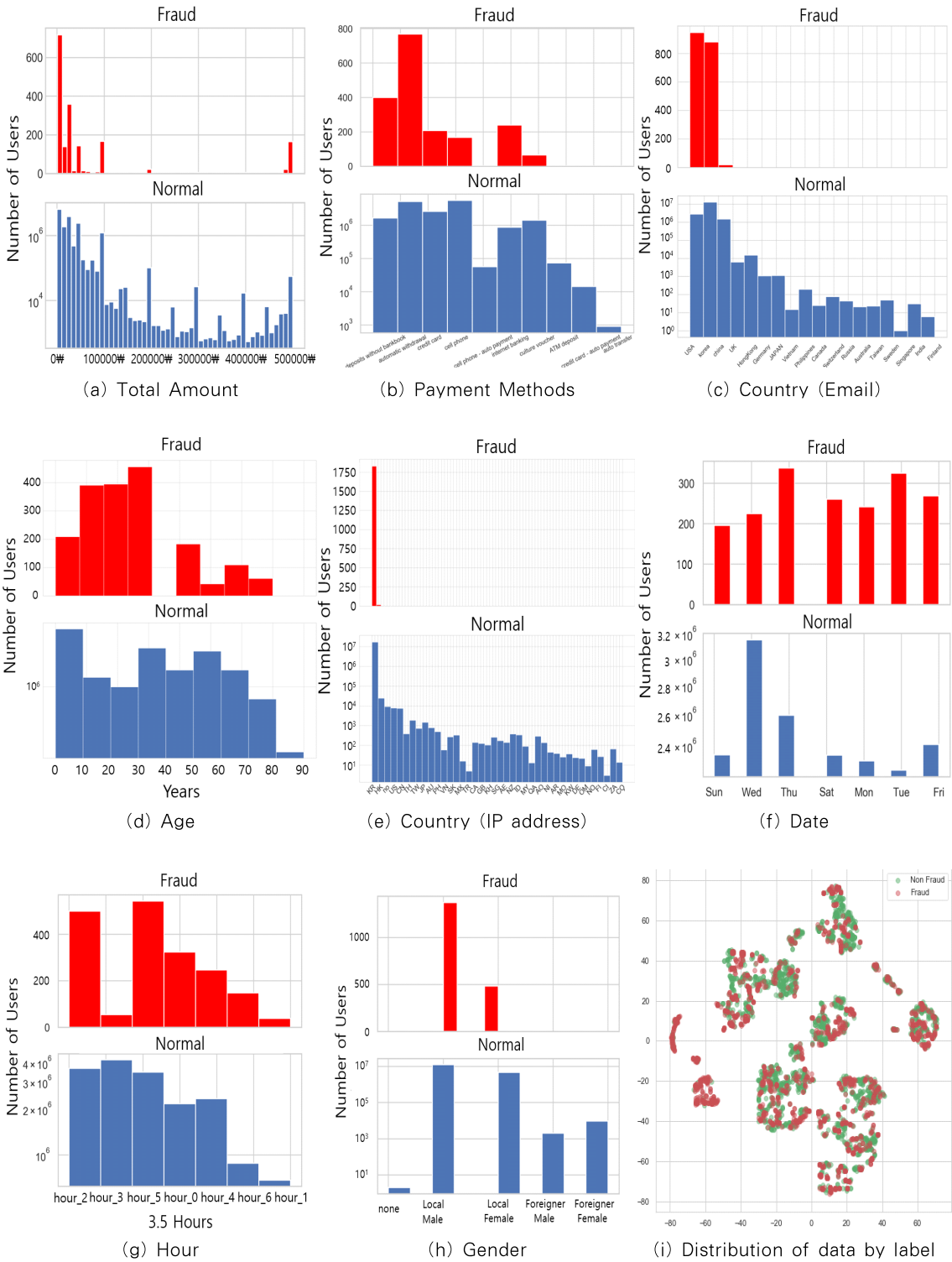

Fig. 3. Visualized distribution of multiple features for fraud and benign data

한 시각화된 결과는 Fig. 3.과 같으며, 몇몇 피처에서 명확하지 않은 패턴의 차이를 확인할 수 있었다. 빈도 기반 벡터화를 정상 및 이상 거래 데이터에 적용하여 구체적으로 활용될 수 있도록 전처리를 진행했다.

데이터의 특징 분포를 이해하기 위해 무작위로 데이터를 선택하여 분석하였다. 데이터의 각 특징의 특성을 자세히 조사하고 분포를 시각화하여 뚜렷한 차이점이 있는지 확인했다. 행렬 형태의 특징을 시각화하기 위해 t-Distributed Stochastic Neighbor Embedding (t-SNE)을 적용했다[22]. 원본 데이터를 2차원 공간으로 축소하고 벡터 공간으로 시각화 하였고, Fig. 3.(i)처럼 데이터가 여러 특성에 따라 분포하고 있음을 확인 할 수 있었다.

각 피처에 대한 모델 입력 벡터 값의 차이가 매우 큰 상태에서 모델을 학습하면 심층 신경망 학습에 악영향을 가져올 수 있다. 탐지 결과가 심하게 왜곡될 수 있으므로 정확한 탐지 결과를 얻기 어렵다. 따라서 수치 데이터를 최적화하여 탐지 모델 학습에 입력 벡터로 활용될 수 있도록 정규화를 사용했다. 정규화 과정으로 피처 간의 절대 수치 값을 상대적 차이로 변환한다. 정규화 프로세스는 각 데이터를 특정 범위로 변환하고 최솟값과 최댓값을 사용하여 전체 데이터에서 특정 데이터 간의 차이를 비교하고 정량화한다.

또한, 모델의 정확한 학습을 위해 이상 거래 이외의 사유로 제재를 받는 경우, 게임 작업장, 불건전 언어 사용, 불법 프로그램, 계정 도용, 게임 규율 위반, 결측치가 존재하는 경우에 데이터는 제외하였으며, X-FDS 생성에 적합한 피처 분석은 Fig. 3.와 같이 진행하였다. 성별 및 날짜와 시간과 같은 문자열 데이터의 경우 실수형 데이터 범위로 변환하였고 이를 통해 선택한 피처는 Table 3.과 같다.

4.4 이상 거래 탐지 모델 생성

앞서 주어진 데이터로부터 이상 거래탐지 모델을 생성하기 위해 전처리 과정을 거쳐 6,676,600명의 정상 데이터와 2,282명의 이상 거래 데이터를 csv 파일 형태로 추출하였으며, 실제 로그와 비슷한 비율로 구성하기 위해 데이터 비율을 학습에 90%를 사용하고, 10%는 검증에 사용하였다. 또한, 이상 거래 데이터에 대한 일정 수준의 탐지 성능을 가진 모델을 만들기 위해서 파라미터를 조절하였다.

이상 거래탐지 모델은 비지도학습인 오토인코더를 사용하였으며, 데이터 입력, 모델 생성, 학습 및 검증 총 3개의 부분으로 나누어 구현하였다.

데이터 입력 부분은 추출한 csv 파일을 읽고, 랜덤하게 섞인 뒤, 지정된 비율로 학습과 검증 데이터가 입력되게 된다. 학습에 적절하도록 데이터 X를 0과 1 사이의 X'값으로 변경해주는 Feature Scaling을 진행하며 그 정의는 아래와 같다[23].

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

모델 생성 부분은 오토인코더가 구현된 부분으로, 인코더와 디코더로 나누어져 있다. 데이터를 입력 레이어를 통해 각각 지정된 노드로 되어있는 5개의 네트워크를 통과하고, 최종적인 아웃풋은 Sigmoid 함

Table 3. Preprocessed Features of Payment Log

Classification	Feature	Description
Payment Information	Total Amount	Total amount of charging
	Payment Methods	Methods used for charging (e.g., credit card, bank transfer, ..)
Processed Payment Data	Gender	Gender of the users
	Date	Date of charging by users
	Hour	Time of charging by users
	Age	Users age in years
	Country (IP address)	Country from user's IP address
	Country (Email)	Country of the email domain
	Label	Labels of the data (e.g., benign, fraud)

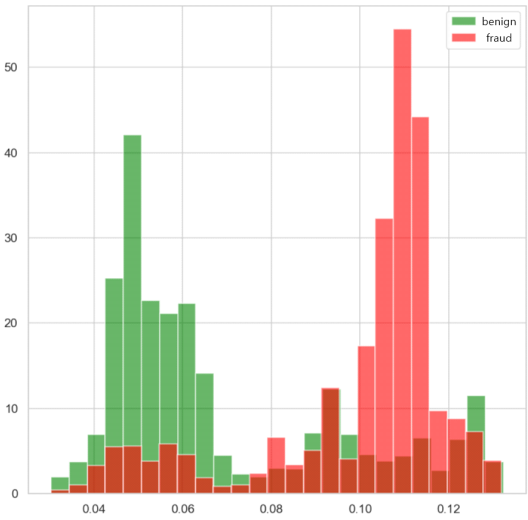

Fig. 4. Distribution of the reconstruction loss

수가 활성화된 출력 레이어를 거쳐 생성된다. Adam 옵티마이저를 통해 각 레이어에 대한 학습과 평가 단계를 거쳐 발생하는 손실 값을 최적화하고 결과를 반환하도록 구성하였다.

본 연구에 사용되는 오토인코더는 정상 데이터를 충분히 제공하여 재구성 손실 값을 계산할 수 있다. 이를 통해 오토인코더는 추가적인 이상 거래 데이터를 입력하게 되면 정상 데이터 입력 대비 낮은 재구성 손실 값이 출력될 것으로 예측할 수 있다. 이러한 재구성 손실 값의 차이를 비정상 거래를 탐지하기 위한 모델의 구분자로 사용하였다.

학습 및 검증 부분에서 임계치를 설정할 수 있으며, 입력값에 따른 출력 값이 임계값을 넘어갈 경우 이상 거래로 판단하도록 설정하였다. 그에 따른 결과는 Fig. 4.과 같으며 임계값의 변화에 따라 탐지 성능이 달라지는 결과를 확인할 수 있었다. 임계치가 작을수록 검출 확률이 일정 이상 증가하였고 0.08 미만이 되었을 때 검출하지 못하였다.

4.5 평가 방법

데이터 입력을 기반으로 임계값을 기준으로 이상 거래를 탐지하는 모델을 제안했고 임계값을 지속적으로 개선하여 탐지 성능을 최적화할 수 있다. 생성한 모델의 성능은 False Positive Error, False Negative Error의 비율로 나타낼 수 있다.

- False Positive Error: Identify benign as fraud
- False Negative Error: Identify fraud as benign

실험의 정확도, 정밀도 및 재현율을 통해 탐지 성능을 평가했다. TP (True Positive), TN (True Negative), FP (False Positive), FN (False Negative), F1-Score를 설정하고 아래 수식을 사용하여 각각의 값을 계산했다.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

$$\mathrm{Precision} = \frac{TP}{TP+FP} \tag{8}$$

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

$$F1-Score = 2 \times \frac{\mathrm{Precision} \times Recall}{\mathrm{Precision} + Recall} \tag{10}$$

V. 실험 결과

5.1 XAI 적용 및 피처 재조정

최종 이상 거래 탐지 모델인 X-FDS를 도출하기 위해 비지도 학습인 오토인코더를 기반으로 실험을 진행했다. 초기 오토인코더 기반 이상 거래 탐지 모델에 학습 데이터를 입력하고 다른 조작 없이 모델을 학습시켰다. 첫 번째 훈련된 모델에 대해 실험 평가를 기록하고 XAI를 적용하여 피드백을 통해 성능을 개선한 결과와 비교를 진행하였다. SHAP 프레임워크를 활용하여 입력 값에 대한 모델의 판단 결과에 대한 분석 그래프는 Fig. 5.와 같다.

X-FDS 모델로부터 각 피처에 대해 도출된 Shapley Value를 이용하여 그래프 일부를 얻을 수 있었다. 그래프의 X축과 Y축은 각각 데이터의 번호와 피처별 Shapley Value를 모두 더한 값을 의미한다. 파란색은 입력 데이터에 대한 모델의 판단에 부정적인 영향을, 빨간색은 긍정적인 영향을 나타낸다. 이를 통해 각 입력 데이터에 대해 모델에 영향을 미치는 피처를 확인하여 모델을 해석하는 것이 가능하다. 그래프에서 각 데이터를 선택할 때 가장 큰 영향을 미치는 피처의 Shapley Value를 세부적으로 확인할 수 있었다.

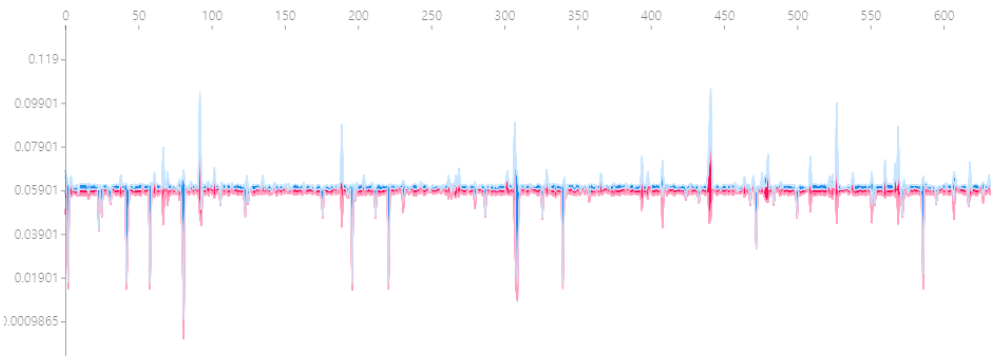

Fig. 5. A sample of XAI result based on Shapley Values

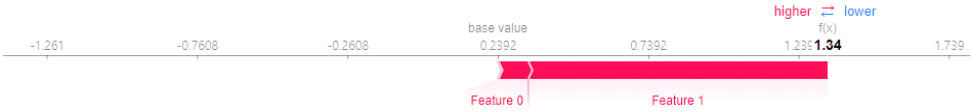

Fig. 6. A detailed XAI result of model misjudgment result by feature prejudice data

추가로 SHAP 프레임워크를 사용하여 모델의 의사 결정 성능에 부정적인 영향을 미치는 요소를 분석했다. 탐지 모델이 잘못 판단한 데이터를 분석한 결과 훈련에 사용된 특정 피처가 이상 거래 판단에 과도한 영향을 미치는 것으로 확인되었다. 예를 들어, Fig. 6.과 같이 입력된 특정 데이터에서 Feature 1 〔Country (IP address)〕의 영향이 매우 높아 레이블이 이상거래 데이터임에도 불구하고 정상 거래로 탐지되었음을 확인할 수 있었다. 해당 피처와 데이터를 세부적으로 분석해본 결과 게임 접속 및 이벤트 참여를 위한 VPN 사용이 많아지며 IP address가 특정 국가에 집중되고 있었다. 이러한 분석을 바탕으로 해당 피처 영향력의 가중치를 낮추고 모델을 수정을 진행했다.

5.2 X-FDS 모델 탐지 결과

FDS 모델은 빠른 학습 결과와 비대칭 데이터의 단점을 줄이는 것이 중요하다. Fig. 7.은 소수의 학습 데이터에 대한 한계를 해결할 수 있음을 검증했다. X-FDS와 Autoencoder 모델은 50,000개의 데이터 학습이 도달하기 전에 일정 성능에 수렴했다. 그러나 LightGBM 및 XGBoost 모델은 적은 량의 학습 데이터가 성능에 악영향을 미치는 것으로 나타났다.

기존 연구했던 XAI 모델 중 SHAP 프레임워크를 통해 생성된 비지도 학습에 대하여 피드백을 분석하여 피처들의 중요도를 확인하고 개선하는 방안을 적용했다. SHAP에서 제공하는 Explainer를 통해 각각의 피처 중요도와 전체 데이터에 대한 영향 그래프를 분석할 수 있었다. 이상 거래탐지 모델에 영향을 많이 주는 순서로는 나라, 거래 금액이 가장 영향력이 높았고, 이어서 거래 매체, 나이, 시간 순서로 영향을 주고 있었음을 확인할 수 있었다, 특히, 성별은 이상 거래탐지 결정 여부에 큰 영향을 미치지 않는 것을 확인할 수 있었다.

XAI를 통해 수정된 Country (IP address)와 삭제된 Gender 데이터로 학습을 진행하여 성능비교를 할 수 있도록 X-LightGBM, X-XGBoost모델을 각각 생성했다. 일부 모델의 경우 성능 하락이 발생했지만 제안한 X-FDS의 경우 불필요한 요소가 제거 또는 수정되어 가장 높은 성능 향상이 도출된 것을 확인할 수 있었다.

위의 결과로 생성된 X-FDS 분류에 따른 데이터의 특성을 확인하기 위해 t-SNE를 사용하여 2차원 벡터 공간으로 시각화를 진행한 결과 Fig. 8.과 같은 그래프를 도출할 수 있었다. 원본 데이터를 벡터 공간으로 진행한 Fig. 3.(i)과 비교했을 때, 정상과 이상 거래 데이터 분류가 개선된 것을 확인할 수 있었다. 이에 따라 X-FDS는 예측에 중요한 피처를

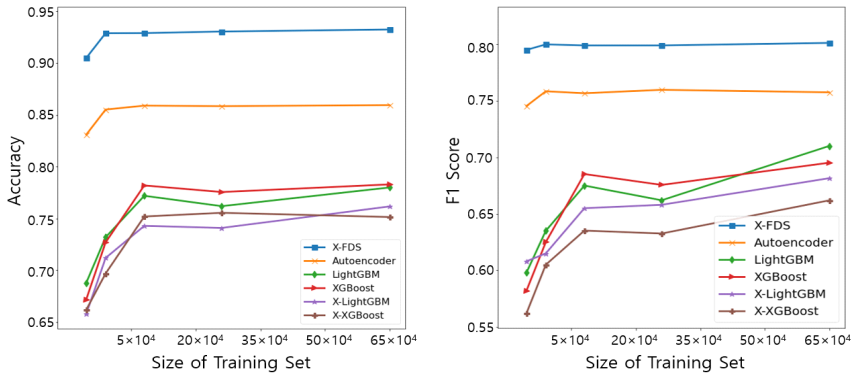

Fig. 7. Model performance along with various training

Table 4. The detection performance compared with previously suggested methods

Model	Training Set		Test Set		Detection Performance			
	Benign	Fraud	Benign	Fraud	Accuracy	Precision	Recall	F1 Score
Autoencoder	6,676,600	-	1,800	200	0.8531	0.7819	0.7771	0.7795
LightGBM	6,676,600	2,082	1,800	200	0.7831	0.7251	0.7013	0.7130
XGBoost	6,676,600	2,082	1,800	200	0.7829	0.7018	0.6813	0.6913
X-FDS	**6,676,600**	**-**	**1,800**	**200**	**0.9412**	**0.8715**	**0.7998**	**0.8341**
X-LightGBM	6,676,600	2,082	1,800	200	0.7617	0.6922	0.6713	0.6815
X-XGBoost	6,676,600	2,082	1,800	200	0.7515	0.6711	0.6531	0.6619

추출할 수 있는 장점이 있었으며, 이상 거래 탐지 모델 정확도에 대한 타당성을 강화하고 개선할 수 있었다. XAI 결과 피드백을 통하여 최종적으로 개선된 모델인 X-FDS를 생성할 수 있다.

시각화를 통해 본 연구에서 진행한 접근 방식이 이전 탐지 모델 방법과 비교하여 탐지 모델의 효율성을 강조할 수 있었다. 기존 연구와 오토인코더 모델보다 X-FDS는 Accuracy, Precision, Recall, F1-Score 성능이 전체적으로 증가한 것을 확인할 수 있었다.

VI. Discussion and Limitation

이상 거래를 탐지하는 기존 방법은 이상 거래 데이터 학습량에 의존적이다. 실제 온라인게임 결제 데이터에서 정상 데이터만을 가지고 비지도학습 모델을 기반으로 탐지 방법을 연구하는 것은 데이터 학습 관점에서 큰 의미가 있다. 기존 금융권에서 사용되던 FDS 모델을 활용하여 온라인게임 결제 데이터에 적용하기 위해서는 다양한 전처리 과정이 필요하다. 본 연구는 온라인게임에 관련된 IP address, 시간, 결제 방법 등을 활용하여 특성화된 탐지 모델을 선제적으로 연구한 부분에 의의가 있다.

탐지 모델을 공격하기 위해 공격자가 학습 데이터를 임의로 조작할 경우 실제 탐지 성능이 크게 떨어지는 문제가 발생할 수 있다. 하지만, X-FDS가 초기 정상 거래 데이터로 충분한 학습이 완료되었을 경우 Fig. 6.처럼 설명력을 가지고 분석하여 탐지할 수 있다는 장점이 있다.

또한, 기존 온라인게임의 이상 거래탐지 모델은 블랙박스와 같아 사용자가 해석할 수 없는 문제를 가지고 있어 신뢰성의 문제가 발생할 수 있는 단점이 있었다. 본 연구에서 XAI를 통해 각 피처의 영향도를 시각적으로 분석 가능하다는 것과 악영향을 미치는 피처들의 영향도를 감소시키고 모델에 대한 해석력을 강화하여 신뢰성이 개선되었다. 주요 기여 내용을 정리하면 다음과 같다.

- 실제 게임 사용자의 거래기록 데이터를 학습에 활용할 수 있도록 가공하고 분석을 진행함
- 오토인코더를 활용한 탐지 모델을 통해 불균형한 데이터임에도 이상 거래를 탐지하였음

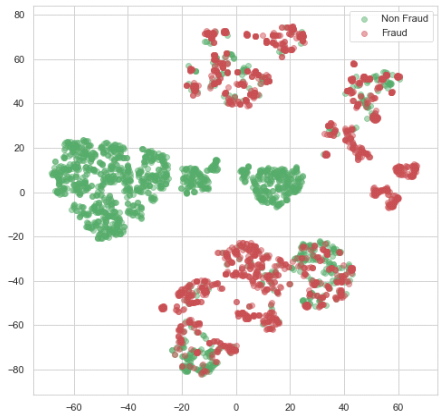

Fig. 8. Result of Benign and Fraud distribution for X-FDS

- XAI를 통해 탐지 모델의 편향된 판단 원인을 분석 및 수정하여 성능을 개선할 수 있음을 실험을 통해 확인함

최근 간편 결제 등 새로운 수단을 활용하는 사례가 점차 늘어나고 있다. 해당 데이터에 대한 추가적인 조사를 통해 모델의 범용성을 높일 필요가 있다. 본 연구에서 참고한 데이터는 결측치가 존재하고 구성된 일부 피처 값이 다르게 입력되어있어 다량의 데이터를 제거하여 진행하였다. 데이터가 추정되거나 변조되어 제공될 경우 모델 신뢰성이 떨어질 수 있으므로 초기 원본 데이터를 정확히 수집하고 검토한다면 이러한 단점을 보완할 수 있을 것으로 판단된다.

VII. 결 론

본 연구에서는 국내외에서 실제 서비스 되고 있는 MMORPG의 충전 로그 데이터를 이용하여 이상 거래가 발생할 수 있는 피처를 분석하고 비지도학습 모델인 오토인코더의 학습 데이터로 활용하여 이상 거래탐지 모델을 생성하였다. 실험 결과에 의하면 임계치 설정과 피처에 따라 이상 거래탐지에 관한 결과가 달라지는 것을 확인할 수 있었으며, 이에 대한 XAI 결과를 통해 개선된 모델을 통해 기존 지도학습 모델 방식과 비교하는 방안을 제안하였다. 정확도에 따른 신뢰성을 향상, 편향성을 개선하고 모델의 판단 근거를 제시할 수 있었다. 추후 LIME과 같은 다른 해석 모델을 통한 해석 피드백으로 확인하지 못했던 모델 개선과 관련된 추가 연구가 필요할 것이다.

References

[1] Newzoo. "Newzoo's games trends to watch in 2021", https://newzoo.com/insights/articles/newzoos-games-trends-to-watch-in-2021/, Accessed: Nov. 2021.

[2] Kim, Hana, Kwak, Byung Il, and Kim, Huy Kang, "A study on the identity theft detection model in MMORPGs," Journal of The Korea Institute of Information Security & Cryptology, vol. 25, no. 3, pp. 627-637, Jun. 2015.

[3] W. Y. Moon and S. D. Kim, "Adaptive fraud detection framework for fintech based on machine learning," Advanced Science Letters, vol. 23, no. 10, pp. 10167-10171, Oct. 2017.

[4] Ge, Dingling, et al. "Credit card fraud detection using lightgbm model." 2020 International Conference on E-Commerce and Internet Technology (ECIT), pp. 232-236, IEEE, Apr. 2020.

[5] C. V. Priscilla and D. P. Prabha, "Influence of optimizing xgboost to handle class imbalance in credit card fraud detection," in 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT). pp. 1309-1315, IEEE, Aug. 2020.

[6] C. Liu, Y. Chan, S. H. Alam Kazmi, and H. Fu, "Financial fraud detection model: Based on random forest," International journal of economics and finance, vol. 7, no. 7, Jun. 2015.

[7] S. Misra, S. Thakur, M. Ghosh, and S. K. Saha, "An autoencoder based model for detecting fraudulent credit card transaction," Procedia Computer Science, vol. 167, pp. 254-262, Apr. 2020.

[8] A. Agrawal, S. Kumar, and A. K.

Mishra, “Credit card fraud detection: A case study,” in 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 5 - 7, IEEE, Mar. 2015.

[9] Y.-G. Cheong, K. Park, H. Kim, J. Kim, and S. Hyun, “Machine Learning Based Intrusion Detection Systems for Class Imbalanced Datasets,” Journal of The Korea Institute of Information Security & Cryptology, vol. 27, no. 6, pp. 1385 - 1395, Dec. 2017.

[10] C. Sun, Z. Yan, Q. Li, Y. Zheng, X. Lu, and L. Cui, “Abnormal group-based joint medical fraud detection,” IEEE Access, vol. 7, pp. 13589 - 13596, Dec. 2018.

[11] A. Adadi and M. Berrada, “Peeking inside the black-box: a survey on explainable artificial intelligence (xai),” IEEE access, vol. 6, pp. 52138 - 52160, Sep. 2018.

[12] A. B. Arrieta, N. D´ıaz-Rodr´ıguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garc´ıa, S. Gil-L´opez, D. Molina, R. Benjamins et al., “Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward re sponsible ai,” Information Fusion, vol. 58, pp. 82 - 115, Jan. 2020.

[13] D. Gunning and D. Aha, “Darpa’s explainable artificial intelligence (xai) program,” AI Magazine, vol. 40, no. 2, pp. 44 - 58, Jun. 2019.

[14] G. E. Melo-Acosta, F. Duitama-Mu˜noz, and J. D. Arias-Londo˜no, “Fraud detec tion in big data using supervised and semi-supervised learning techniques,” in 2017 IEEE Colombian conference on communications and computing (COLCOM), pp. 1 - 6, IEEE, Aug. 2017.

[15] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, “Greedy layer-wise training of deep networks,” in Advances in neural information processing systems, pp. 153 - 160, Dec. 2007.

[16] S. Lange and M. Riedmiller, “Deep auto-encoder neural networks in reinforce ment learning,” in The 2010 International Joint Conference on Neural Networks (IJCNN), pp. 1 - 8, IEEE, Jul. 2010.

[17] F. Farahnakian and J. Heikkonen, “A deep auto-encoder based approach for intrusion detection system,” in 2018 20th International Conference on Advanced Communication Technology (ICACT), pp. 178 - 183, IEEE, Feb. 2018.

[18] M. Al-Shabi, “Credit card fraud detection using autoencoder model in unbalanced datasets,” Journal of Advances in Mathematics and Computer Science, pp. 1 - 16, Aug. 2019.

[19] L. S. Shapley, “A value for n-person games,” Published by Princeton University Press, 1953.

[20] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Proceedings of the 31st international conference on neural information processing systems, pp. 4768 - 4777, Dec. 2017.

[21] J. Y. Chun and G. Noh, “Suggestions for Applications of Anonymous Data under the Revised Data Privacy Acts,” Journal of The Korea Institute of Information Security & Cryptology, vol. 30, no. 3, pp. 503 - 512, Jun. 2020.

[22] S. Arora, W. Hu, and P. K. Kothari, “An analysis of the t-sne algorithm

for data visualization," in Conference On Learning Theory. PMLR, pp. 1455 - 1462, Jul. 2018.

[23] N. Kasa, A. Dahbura, C. Ravoori, and S. Adams, "Improving credit card fraud detection by profiling and clustering accounts," in 2019 Systems and Information Engineering Design Symposium (SIEDS), pp. 1 - 6, IEEE, Apr. 2019.

〈저 자 소 개〉

이 영 헌 (Young Hun Lee) 학생회원
2019년 2월: 서울과학기술대학교 컴퓨터공학과 졸업
2020년 2월~현재: 고려대학교 정보보호학과 석사과정
〈관심분야〉 금융보안, 차량보안, 데이터마이닝

김 휘 강 (Huy Kang Kim) 종신회원
1998년 2월: KAIST 산업경영학과 학사
2000년 2월: KAIST 산업공학과 석사
2009년 2월: KAIST 산업 및 시스템공학과 박사
2004년 5월~2010년 2월: 엔씨소프트 정보보안실장, Technical Director
2010년 3월~2014년 12월: 고려대학교 정보보호대학원 조교수
2015년 1월~2020년 2월: 고려대학교 정보보호대학원 부교수
2020년 3월~현재: 고려대학교 정보보호대학원 교수
〈관심분야〉 온라인게임 보안, 자동차 보안, 침입탐지시스템, 네트워크 보안