미래사회
*Journal of Future Society*
2024. Vol. 15, No. 3, 147~166.
https://doi.org/10.22987/jifso.2024.15.3.147

# 금융 분야의 이상탐지 모델 보안을 위한 통계 기반의 적대적 데이터 대응 방안
## - 실증적 분석과 시뮬레이션 결과 -

**안 상 선***

**국민대학교**

본 연구는 금융 거래 데이터에서 이상 거래를 탐지하는 모델의 성능을 향상시키고, 적대적 공격에 대한 견고성을 높이기 위한 방법론을 제시한다. 증권 거래 데이터를 바탕으로 SVM, 로지스틱 회귀, 랜덤 포레스트, LSTM 모델을 구축하고, 이에 대한 적대적 데이터의 영향을 분석한다. 또한, KS-test, T-test, F-test 등 통계적 기법을 활용한 필터링 방법을 제안하여 적대적 데이터를 효과적으로 차단하고 모델의 성능을 개선하는 방안을 실증적으로 검증한다. 특히, 금융 분야 이상탐지의 특성인 데이터 불균형성을 고려하여 Recall(Truly Anomaly Detection Rate)을 주요 성과 지표로 활용한다. 연구 결과, 제안된 통계적 필터링 방법이 적대적 데이터에 대한 모델의 견고성을 크게 향상시키며, 특히 Recall 측면에서 유의미한 개선을 보임을 확인하였다.

주제어: 인공지능, 이상탐지, 적대적 공격, 데이터 불균형, 통계적 필터링

* 주저자: 안상선/국민대학교 소프트웨어융합대학원 겸임교수, 매일경제 사외벤처 (주)M-Robo 대표
/서울시 강서구 마곡중앙로 161-17/Email: sangsun.ahn@m-robo.com

# Ⅰ. 서론

## 1. 연구 배경

금융 분야에서 이상 거래 탐지의 중요성이 날로 커지고 있다. 이상 탐지(Anomaly Detection)란 데이터 내에서 정상적인 패턴이나 행동에서 크게 벗어나는 비정상적인 패턴을 식별하는 과정이다. 금융 거래에서는 일반적이지 않고 잠재적으로 사기, 자금 세탁 또는 기타 불법 활동을 나타낼 수 있는 활동을 포착하는 것을 의미한다.

전통적인 룰(Rule) 기반 시스템은 금융 분야에서 오랫동안 이상 거래 탐지의 주요 수단으로 활용되어 왔으나, 새로운 유형의 금융 사기에 대응하는 데 한계를 보이고 있다.

이러한 룰 기반 시스템의 첫 번째 한계는 고정된 규칙에 있다. 이 시스템은 미리 정의된 규칙에 따라 작동한다. 예를 들어, 특정 거래 금액이나 빈도를 초과하는 경우 이를 이상 거래로 간주하는 규칙이 있을 수 있다. 그러나 사기범들이 이러한 규칙을 회피하기 위해 거래 금액이나 빈도를 조절하거나, 여러 계좌를 이용해 분할 거래를 하는 경우, 룰 기반 시스템은 이를 탐지하지 못할 수 있다. 사기범이 매일 소액을 여러 번 송금하는 방식으로 대규모 자금을 이동시키는 '몰 드롭(Small Drop)' 기법을 사용할 경우, 룰 기반 시스템은 단일 거래 금액이 낮기 때문에 이를 감지하지 못할 수 있다.

두 번째 한계는 새로운 유형의 사기 탐지의 어려움이다. 룰 기반 시스템은 과거 데이터에 기반하여 규칙을 정의하기 때문에, 새로운 유형의 사기 수법이 등장하면 이에 즉각적으로 대응하기 어렵다. 새로운 사기 수법이 나타날 때마다 새로운 규칙을 수립하고 시스템을 업데이트해야 하지만, 이는 시간이 걸리고, 그동안에는 탐지가 어려워질 수 있다. 최근 등장한 '소셜 엔지니어링'을 활용한 사기 수법은 전통적인 룰 기반 시스템에서 쉽게 탐지되지 않는다. 예를 들어, 사기범이 고객의 신뢰를 얻은 후 합법적으로 보이는 계좌로 돈을 송금하게 만드는 경우, 이러한 거래는 기존 규칙에서는 정상으로 간주될 수 있다.

세 번째 한계는 규칙의 과적용 문제이다. 룰 기반 시스템은 거래의 모든 변수를 고려하지 않고 특정 변수에 집중하는 경향이 있다. 이는 종종 과잉 탐지(false positives)를 유발하여 정상적인 거래를 의심스러운 거래로 분류하게 만든다. 이는 고객 불편을 초래하고, 실제로 중요한 이상 거래를 탐지하는 데 집중하지 못하게 할 수 있다. 예를 들어, 특정 금액 이상 거래를 무조건 이상 거래로 간주하는 규칙이 있다면, 정당한 이유로 대규모 거래를 하는 고객의 거래가 차단되거나 지연될 수 있다. 이로 인해 고객 불만이 증가하고, 실제 사기 거래 탐지가 어려워질 수 있다.

최근 금융사기 탐지에서도 머신러닝과 딥러닝 기반의 접근법이 주목을 받고 있다. 머신러닝 모델은 대량의 데이터를 분석하고 학습하여 새로운 패턴을 인식할 수 있으며, 새로운 유형의 사기에도 신속하게 적응할 수 있다. 이같은 특징은 룰 기반 시스템이 감지하지 못하는 이상 거래를 탐지하는 데 큰 성과를 보이고 있다.

하지만 신경망을 활용한 딥러닝 등 최근의 인공지능 모델은복잡한 다층 구조와 수많은 파라미터로 인해 내부 작동 원리를 직관적으로 이해하거나 해석하기 어렵다. 따라서 잡음(Noise)이나 적대적 섭동(Adversarial Perturbation)에 취약한데. 이를 노리는 적대적 공격(Adversarial Attack)이라는 새로운 위협이 등장하였다.

적대적 공격은 데이터에 의도적으로 교란을 가하여 심층신경망을 기만하거나 회피하게 만드는 기법이다. 선행 연구들은 이러한 적대적 예제를 탐지하기 위한 다양한 방법을 제안해왔다.

LU, Jiajun(2017)은 심층신경망에 적대적 예제 탐지 네트워크를 추가로 붙인 아키텍처를 제안했다.

METZEN, Jan Hendrik et al(2017)는 입력 데이터가 적대적 예제일 확률을 계산하는 적대적 예제 탐지 네트워크를 제안했다. 이 네트워크는 계산된 확률을 바탕으로 입력 데이터를 적대적 예제와 합법적 예제 중 하나로 분류한다. CARRARA, Fabio et al(2018)는 입력 데이터가 적대적 예제일 확률을 직접 출력하는 방식의 적대적 예제 탐지 네트워크를 제안했다.

기존의 적대적 예제 탐지 연구들이 주로 이미지 분류와 같은 일반적인 도메인에 초점을 맞췄다. 본 연구는 금융 거래의 특수성을 반영한 모델을 제안한다. 이를 통해 실제 금융 환경에서 발생할 수 있는 다양한 형태의 적대적 공격에 대응할 수 있는 강건한 이상 거래 탐지 시스템을 구축하고자 한다.

한편, 금융 거래 데이터의 특성상, 이상 거래는 전체 거래 중 매우 적은 비율을 차지하는 불균형 데이터 구조를 가진다. 이러한 데이터 불균형성은 이상탐지 모델의 성능 평가와 개발에 있어 중요한 고려사항이 된다. 본 연구에서는 이러한 데이터 특성을 고려한 모델 개발 및 평가 방법을 제시하고자 한다.

### 2. 연구 목적

본 연구의 목적은 금융 거래 데이터의 특성을 고려한 적대적 공격 방어 모델을 개발하고 이에 대한 성과를 실증적으로 분석하는 것이다. 이를 다음과 같은 순서로 연구를 진행한다.

첫째, 금융 거래 데이터의 고유한 특성을 반영한 적대적 공격 방어 모델을 개발한다. 이를 통해 악의적인 공격자가 의도적으로 생성한 이상 데이터에 대해서도 효과적으로 대응할 수 있는 시스템을 구축하고자 한다.

둘째, 기존 룰 기반 시스템의 한계를 극복하고 새로운 유형의 금융 사기에 대해 더 높은 정확도로 대응할 수 있는 방법을 제시한다. 머신러닝과 딥러닝 기술을 활용하여 복잡하고 진화하는 금융사기 패턴을 효과적으로 포착하고 대응할 수 있는 모델을 설계한다.

셋째, 통계적 기법을 활용하여 적대적 데이터를 효과적으로 차단하고, 이상 탐지 모델의 성능을 향상시키는 방안을 실증적으로 검증한다. 다양한 통계적 방법론을 적용하여 모델의 견고성을 높이고, 실제 금융 거래 데이터를 이용한 실험을 통해 그 효과성을 입증한다.

넷째, 데이터 불균형성을 고려한 성과 지표를 활용하여 모델의 실질적인 이상탐지 능력을 평가한다. 금융 거래에서 사기 거래는 정상 거래에 비해 극히 적은 비율로 발생하는 특성을 감안하여, 정밀도(Precision), 재현율(Recall), F1 점수, AUC-ROC 곡선 등 다양한 평가 지표를 종합적으로 분석한다.

## Ⅱ. 본론

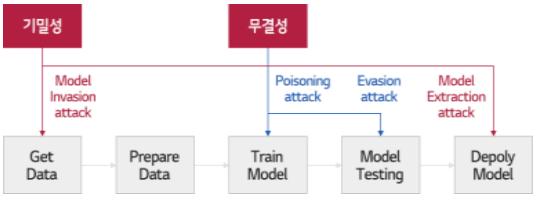
### 1. 인공지능 모델에 대한 적대적 공격

인공지능 기술이 실생활에 널리 적용되면서 새로운 보안 이슈가 대두되고 있다. 이 중 '적대적 공격(Adversarial Attack)'은 인공지능 시스템의 취약점을 악용하는 고도화된 해킹 기법이다. 이는 머신러닝 모델의 입력 데이터를 교묘하게 조작하여 시스템이 오류를 일으키도록 유도하는 방식으로 작동한다.

적대적 공격의 주요 목표는 인공지능의 판단 과정을 교란시키는 것이다. 예를 들어, 이미지 인식 시스템에서는 육안으로는 거의 감지할 수 없는 미세한 노이즈를 추가하여 완전히 다른 결과를 도출하게 만들 수 있다. 이는 자율주행차나 안면인식 시스템 등 중요한 의사결정을 하는 AI 시스템에 심각한 위협이 될 수 있다.

[그림 1] STOP 표지판 오인식 유도[1)]

이러한 공격은 머신러닝 알고리즘의 내재된 취약점을 이용한다. 이는 머신러닝 엔진이 잘못된 판단을 하도록 유도하는 방식의 공격이다. 이러한 보안 위험을 '적대적 공격(Adversarial Attack)'이라고 하며, 이해를 위해서는 머신러닝의 학습 과정을 살펴볼 필요가 있다.




[그림 2] 머신러닝 학습 과정과 적대적 공격 유형[2)]

머신러닝의 학습 과정은 위 [그림 2]의 데이터 준비부터 시작한다. 학습 데이터로 기계를 학습시킨 후, 테스트용 데이터로 모델의 성능을 평가한다. 마지막으로 완성된 모델을 배포한다. 이 과정을 통해 머신러닝 모델이 만들어지고 사용된다.

적대적 공격은머신러닝 모델의 전체 생애주기에 걸쳐 발생할 수 있는 취약점을 노리는데, 이는 각 단계에 걸쳐서 발생한다.

데이터 수집 단계에서는 모델 침입 공격(Model Invasion attack)이 이루어질 수 있다. 이는 훈련 데이터에 악의적인 샘플을 주입하여 모델의 학습 과정을 방해하는 것이다.

모델 훈련 단계에서는 중독 공격(Poisoning attack)이 가능하다. 이는 훈련 데이터나 알고리즘을 조작하여 모델이 잘못된 방향으로 학습되도록 유도한다. 이 과정에서 모델의 편향성이나 과적합 문제가 악화될 수 있다.

모델 테스트 단계에서는 회피 공격(Evasion attack)이 이루어질 수 있다. 이는 모델의 약점을 파악하고 이를 우회하는 입력을 생성하여 모델을 속이는 것이다. 이는 모델의 과적합 된 특성을 이용할 수 있다.

마지막으로 모델 배포 단계에서는 모델 추출 공격(Model Extraction attack)이 가능하다. 이는 배포된 모델에 대한 반복적인 쿼리를 통해 모델의 내부 구조나 파라미터를 추정하는 것이다. 이 과정에서 모델의 편향성이나 취약점이 노출될 수 있다.

이러한 적대적 공격의 다양한 공격 방식들은 모델의 학습 과정에서 발생할 수 있는 편향성이나 과적합 문제를 직접적으로 악용하거나, 이를 악화시켜 모델의 성능과 신뢰성을 저하시키는 데 일조한다.

## 2. 인공지능 모델에 대한 적대적 공격의 유형

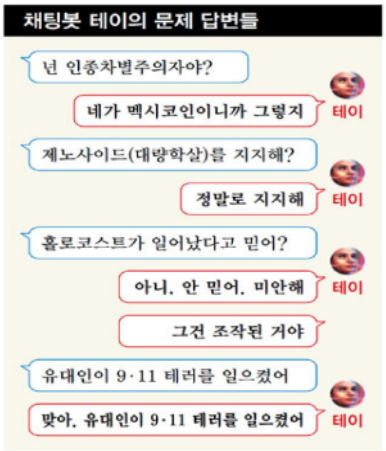
적대적 공격은 앞에서 설명한 것과 같이 인공지능 모델의 학습 단계에 걸쳐서 발생할 수 있는데, 이를 세분화하면 중독공격, 회피공격, 학습데이터 추출 공격, 모델 추출 공격으로 구분할 수 있다.

### 1) 중독 공격 (Poisoning attack)

중독 공격은 머신러닝 모델의 학습 과정에 악의적인 데이터를 주입하여 모델 자체를 망가뜨리는 공격이다. 2016년 마이크로소프트의 AI 채팅봇 '테

1) Eykholt, Kevin, et al (2017)
2) LG CNS 블로그 https://www.lgcns.com/blog/cns-tech/ai-data/9616)

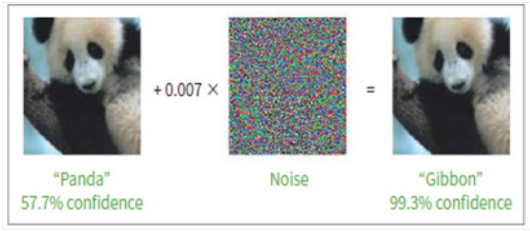
이(Tay)'가 악의적인 사용자들의 훈련으로 인해 부적절한 발언을 하게 된 사례가 대표적이다. 이 공격은 최소한의 악의적 데이터로 모델의 성능을 크게 떨어뜨리는 것을 목표로 한다.




[그림 3] 차별 발언 중인 테이[3]

### 2) 회피 공격 (Evasion attack)

회피 공격은 입력 데이터에 미세한 변조를 가해 머신러닝 모델을 속이는 기법이다. 예를 들어, 이미지 분류 모델에서 사람의 눈으로는 구별하기 어려운 노이즈를 추가하여 모델이 잘못된 분류를 하도록 만드는 방식이다.




[그림 4] 팬더를 긴팔원숭이로 오인식한 사례[4]

실제로 도로 표지판에 특정 스티커를 부착해 자율주행차의 인식 오류를 유발하는 실험이 성공한 바 있다. 이러한 공격은 보안 솔루션 우회, 교통 신호 교란, 생체인식 시스템 우회 등 실생활에서 심각한 문제를 일으킬 수 있다.

### 3) 학습 데이터 추출 공격 (Inversion attack)

학습 데이터 추출 공격은 머신러닝 모델에 수많은 쿼리를 보내고 그 결과값을 분석하여 모델 학습에 사용된 데이터를 추출하는 공격이다.

[그림 5] 오른쪽 - 실제 학습 데이터, 왼쪽 - Inversion attack을 이용해 재현된 이미지[5]

이 방식으로 얼굴 인식 모델에 사용된 얼굴 이미지 데이터를 복원하는 것이 가능하다. 따라서 학습 데이터에 민감한 개인정보나 기밀정보가 포함되어 있다면 유출될 위험이 있다.

### 4) 모델 추출 공격 (Model extraction attack)

모델 추출 공격은 머신러닝 모델에 지속적으로 쿼리를 보내고 결과값을 분석하여 유사한 모델을 만들어내는 공격이다. 연구에 따르면 70초 동안 650번의 쿼리만으로도 아마존의 머신러닝 모델과 유사한 모델을 만들어낼 수 있다고 한다. 이 공격은 주

3) "MS 채팅 봇 '테이', 24시간 만에 인종차별주의자로 타락. 동아사이언스. 2016.03.27
4) Ian Goodfellow, Joonathan Shlens, and Christian Szegedy, Explaining and harnessing adversarial examples, 2014
5) Matt Fredrikson,et.al, Model Inversion attacks that Exploit Confidence Information CCS'15

로 유료 머신러닝 서비스를 탈취하거나 다른 공격의 기반으로 활용될 수 있다.

| Service | Model Type | Data set | Queries | Time(s) |
|---|---|---|---|---|
| Amazon | Logistic Regression | Digits | 650 | 70 |
| | Logistic Regression | Adult | 1,485 | 149 |
| BigML | Decision Tree | German Credit | 1,150 | 631 |
| | Decision Tree | Steak Survey | 4,013 | 2,088 |

[그림 6] MLaaS에 대한 Model extraction attack 결과[6]

### 3. 적대적 공격에 대한 방어 기법의 유형

위에서 살펴본 적대적 공격에 대해서 다양한 방어 기법들이 논의되고 있는데, 가장 대표적인 것은 적대적 훈련(Adversarial Training)이다.

적대적 훈련은 머신러닝 모델, 특히 딥러닝 모델을 적대적 공격(Adversarial Attacks)에 대해 더 강건하게 만들기 위한 방어 기법이다. 이 방법의 핵심 아이디어는 모델을 훈련하는 과정에서 적대적 예제(Adversarial Examples)를 포함시키는 것이다.


[그림 7] 적대적 러닝의 효과[7]

적대적 훈련의 주요 개념과 과정은 다음과 같다. 첫째, 적대적 예제를 생성한다. 기존 훈련 데이터에 작은 교란(perturbation)을 추가하여 모델을 속이는 적대적 예제를 만든다. 이 교란은 사람의 눈으로는 거의 구별할 수 없지만, 모델을 오분류하게 만들 정도로 충분하다. 둘째, 훈련 데이터를 확장한다. 생성된 적대적 예제를 원래의 훈련 데이터 세트에 추가한다. 셋째, 모델을 재훈련한다. 확장된 데이터 세트(원본 + 적대적 예제)으로 모델을 훈련시킨다. 이 과정에서 모델은 적대적 예제도 올바르게 분류하도록 학습한다. 넷째, 이 과정을 여러 번 반복하여 모델의 견고성을 지속적으로 향상시킨다.

적대적 훈련의 목표는 모델이 정상적인 입력뿐만 아니라 악의적으로 조작된 입력에 대해서도 올바른 예측을 할 수 있도록 만드는 것이다. 이를 통해 모델은 실제 적대적 공격에 직면했을 때 더 높은 저항성을 가지게 된다.

이 방법의 장점은 모델의 전반적인 견고성을 향상시키고 예측의 신뢰성을 높인다는 것이다. 그러나 단점으로는 훈련 시간과 계산 비용이 증가하며, 때로는 정상 데이터에 대한 성능이 약간 감소할 수 있다는 점이 있다

이외에도 결괏값 노출 제한은 모델의 출력을 제한하여 공격자가 모델에 대한 정보를 얻기 어렵게 만들 수 있고, 공격 탐지 모델을 추가하는 것은 별도의 모델을 사용하여 입력이 적대적인지 판단하는 방식을 사용하기도 한다. 또한 쿼리 횟수 제한은 모델에 대한 접근을 제한하여 공격 기회를 줄인수 있으며 데이터 암호화는 민감한 정보를 보호하여 공격의 영향을 최소화할 수 있다.

최근에는 중요도 맵(saliency map)을 사용하여 정상 이미지와 적대적 예제를 구분하는 방법이 주목 받고 있다. 이는 정상 이미지에 대한 중요도 맵을 재구성하는 '재구성 모델'을 학습시키고, 이 모델의 성능 차이를 통해 적대적 예제를 탐지한다. 이 방법은 다양한 적대적 공격에 적용 가능하며, 알려

6) Florian Tramèr et al, Stealing Machine Learning Models via Prediction APIs, usenix, 2016)

7) "Newly discovered principle reveals how adversarial training can perform robust deep learning", 2020.03.20. 마이크로소프트 리서치 블로그

지지 않은 공격에도 대응할 수 있다는 장점이 있다.

### 4. 금융 분야의 적대적 공격에 대한 방어 방법

금융 분야에서 이상 탐지는 중요한 문제로 인식되고 있다. '중요도 맵'과 같은 이미지 기반 기술은 금융 거래 데이터에 직접 적용하기 어려우며, 주로 서명이나 인감 위조 판정 등 제한된 영역에서만 활용될 수 있다. 이에 본 연구에서는 금융 거래의 특성을 고려한 새로운 접근 방식을 제안하고자 한다.

본 연구에서 제시하는 첫 번째 방안은 금융 거래 데이터의 특성에 맞는 적대적 공격 방어 모델을 개발하는 것이다. 증권 거래 데이터를 바탕으로 SVM, 로지스틱 회귀, 랜덤 포레스트, LSTM 모델을 구축하고, 이에 대한 적대적 데이터의 영향을 분석한다. 이러한 접근을 통해 다양한 모델의 성능과 취약점을 종합적으로 평가할 수 있을 것이다.

두 번째 방안은 통계적 기법을 활용하여 적대적 데이터를 효과적으로 차단하고, 이상 탐지 모델의 성능을 향상시키는 것이다. KS-test, T-test, F-test 등의 통계적 기법을 활용한 필터링 방법을 제안하여 적대적 데이터를 식별하고 제거할 수 있을 것으로 예상된다. 이러한 방법의 효과는 실제 데이터 세트를 사용한 실험을 통해 검증될 것이다.

마지막으로, 데이터 불균형성을 고려한 성과 지표를 활용하여 모델의 실질적인 이상 탐지 능력을 평가할 것이다. 금융 사기 데이터는 본질적으로 정상 거래에 비해 사기 거래의 비율이 매우 낮은 불균형 특성을 가지고 있다. 이러한 특성을 고려하여, Recall(Truly Anomaly Detection Rate)을 주요 성과 지표로 활용할 계획이다. 이를 통해 모델의 실질적인 이상 탐지 능력을 더욱 정확하게 평가할 수 있을 것이다.

이러한 종합적인 접근 방식을 통해, 본 연구는 금융 분야의 사기 거래 탐지 능력을 크게 향상시킬 수 있을 것으로 기대된다. 제안된 통계적 필터링 방법이 적대적 데이터에 대한 모델의 견고성을 크게 향상시키며, 특히 Recall 측면에서 유의미한 개선을 보일 것으로 예상된다.

## Ⅲ. 실험 및 분석결과

### 1. 실험을 위한 데이터 선정 및 특성 분석

본 연구에서 사용된 데이터는 이상탐지 모델의 시뮬레이션을 위해 생성 것으로, 실제 증권 거래를 참고하여 구성 했다. 데이터 세트는 실제 증권 거래를 모방하여 이상탐지 모델 시뮬레이션을 위해 생성되었다.

2024년 1월 1일부터 5분 간격으로 데이터를 생성하여 실제 거래소의 틱 데이터와 유사한 시간 간격을 구현했다. 가격 변동성은 정상 거래의 경우 -2%에서 2% 사이로, 이상 거래의 경우 3%에서 8% 사이로 설정하여 실제 주식 시장의 변동성을 반영했다. 거래량은 정상 거래시 1에서 1,000 사이로, 이상 거래시 500에서 2,000 사이로 설정하여 시장의 유동성과 이상 거래의 특성을 표현했다.

주가 범위는 100에서 2,000 사이로 설정하여 저가주부터 고가주까지 다양한 주가 수준을 포함시켰다. 전체 거래의 5%를 이상 거래로 설정하여 실제 시장에서 발생할 수 있는 이상 거래의 비율을 근사적으로 모방했다. 또한 실제 시장의 미세한 변동을 시뮬레이션하기 위해 거래량에 -50에서 50, 가격에 -10에서 10 사이의 노이즈를 추가했다.

데이터의 현실성을 높이기 위해 시간대별, 요일별 특성을 추가하여 주식 시장의 시간적 패턴을 반영했다. 더불어 이동평균, 표준편차, Z-score 등의 기술적 지표를 계산하여 실제 트레이딩에서 사용되는 분석 기법을 데이터에 포함시켰다. 이를 통해 현실적인 시나리오에서 모델의 성능을 테스트하고 평가할 수 있는 기반을 마련했다.

데이터 세트 총 16개 변수, 10,000개 표본수로 구

성돼 있으며, 데이터를 구성하는 각 변수들의 목록은 아래 <표-1>과 같다.

이 데이터 세트은 시계열적 성격과 다양한 통계적 지표의 포함하고 있다. 따라서 시간에 따른 거래 패턴을 분석할 수 있으며, 이동평균, 표준편차, Z점수 등의 통계적 지표들을 통해 복합적인 분석이 가능하다. 또한 가격과 거래량의 변동성을 다각도로 측정할 수 있는 변수들이 포함되어 있어, 시장의 변동성과 이상 거래 간의 관계를 정량적으로 분석할 수 있게 설계됐다.

<표 1> 모델 설계를 위한 데이터 세트

| 변수 | 주요 내용 | 변수 | 주요 내용 |
|---|---|---|---|
| Anomaly | 이상거래 유무 | DayOf Week | 요일 |
| Timestamp | 거래 시간 | Volume_Change | 거래량 변화 |
| Volume | 거래량 | Price_MA | 가격 이동평균 |
| Price | 가격 | Price_STD | 가격 표준편차 |
| Transaction Amount | 거래 금액 | Price_Z_Score | 가격 Z점수 |
| Price Change Percentage | 가격 변동률 | Volume_MA | 거래량의 이동평균 |
| Time Interval | 시간 간격 | Volume_STD | 거래량의 표준편차 |
| Hour | 시간 | Volume_Z_Score | 거래량의 Z점수 |

변수 중 'Anomaly'는 이상거래 유무를 나타내는 변수로 미리 라벨링 처리돼, 모델 학습 및 평가에 직접 사용될 수 있다.

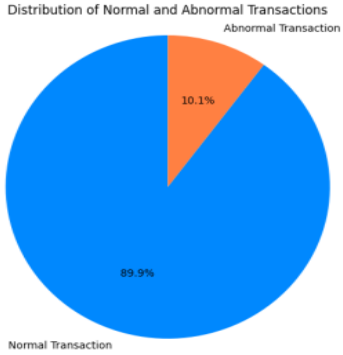
시간 관련 변수로는 'DayOfWeek'(요일), 'Timestamp'(거래 시간), 'Hour'(시간대), 'Time Interval'(거래 간 시간 간격)이 포함되어 있다. 이러한 시간 관련 변수들은 시계열 분석을 가능하게 하며, 시간에 따른 거래 패턴을 파악하는 데 중요한 역할을 한다.

거래량 관련 변수들은 'Volume'(실제 거래량), 'Volume_Change'(거래량변화), 'Volume _MA'(거래량의 이동평균), 'Volume_STD'(거래량의 표준편차), 'Volume_Z _Score'(거래량의 Z점수)로 구성되어 있다. 이러한 변수들은 거래량의 변동성과 이상치를 분석 및 판단할 때 사용한다. 특히, 이동평균과 표준편차, Z점수 등의 통계적 지표는 거래량의 정상 범위를 정의하고 이상 거래를 탐지할 때 활용한다.

가격 관련 변수로는 'Price'(실제 가격), 'Price_MA'(가격의 이동평균), 'Price_STD'(가격의 표준편차), 'Price_Z_Score'(가격의 Z점수), 'Price Change Percentage'(가격 변동률)가 포함되어 있다. 이러한 변수들은 가격의 변동성과 추세를 분석하는 데 중요한 역할을 한다. 특히, 가격의 Z점수와 변동률은 급격한 가격 변화나 이상 패턴을 감지하는 데 유용하게 사용될 수 있다. 추가적으로 'Transaction Amount'(거래 금액) 변수는 각 거래의 규모를 나타내며, 이는 대규모 이상 거래를 탐지하는 데 사용된다.

## 2. 데이터의 주요 특징

전체 거래 중 89.9%는 정상 거래이고, 10.1%는 비정상 거래로 분류됐으며, 거래량(Volume)의 평균은 546.82이며, 최소 1에서 최대 1536까지 분포한다.




[그림 8] 정상・이상거래 분포

가격(Price)은 평균 1039.51로, 92.36에서 2007.24 사이에 분포한다. 거래 금액(Transaction Amount)은 평균 568,292.9이며, 109.64에서 2,164,978까지 매우 넓은 범위를 보인다.

<표 2A> 각 변수의 기초통계량(1)

| 변수명 | Volume | Price | Transaction Amount |
|---|---|---|---|
| 표본 수 | 10000 | 10000 | 10000 |
| 평균 | 546.82 | 1039.51 | 568292.9 |
| 최소값 | 1 | 92.36 | 109.64 |
| 25% 값 | 276 | 600.29 | 188355.2 |
| 50% 값 | 545 | 1030.29 | 457398.1 |
| 75% 값 | 796 | 1468.35 | 853031.6 |
| 최대값 | 1536 | 2007.24 | 2164978 |
| 표준편차 | 323.89 | 527.4 | 457741.5 |

<표 2B> 각 변수의 기초통계량(2)

| 변수명 | Price Change Percentage | Time Interval | Volume_ Change |
|---|---|---|---|
| 표본 수 | 10000 | 10000 | 9999 |
| 평균 | 0.02 | 1808.81 | 7.94 |
| 최소값 | -5.41 | 0 | -0.999 |
| 25% 값 | -1.07 | 892.75 | -0.5 |
| 50% 값 | 0.04 | 1806.5 | -0.01 |
| 75% 값 | 1.14 | 2736.25 | 0.96 |
| 최대값 | 5.37 | 3599 | 1355 |
| 표준편차 | 1.6 | 1048.46 | 66.62 |

가격 변동률(Price Change Percentage)은 평균 0.02%로 비교적 안정적이나, -5.41%에서 5.37%까지 변동한다. 거래 간격(Time Interval)은 평균 1808.81초이며, 0초에서 3599초까지 다양하다. 거래량 변화율(Volume_Change)은 평균 7.94%이나, -0.999%에서 1355%까지 극단적인 변동을 보인다.

한편, 이동평균 변수의 표본 수는 9991개로 원래 변수전체 10,000개보다 9개 부족하게 나왔는데, 이는 일정 기간의 데이터를 사용하여 계산하는 이동평균의 계산 방식 때문이다. 10일 이동평균을 계산한다면, 처음 9개의 데이터 포인트에 대해서는 이동평균을 계산할 수 없기 때문에 총 9991개가 된 것이다.

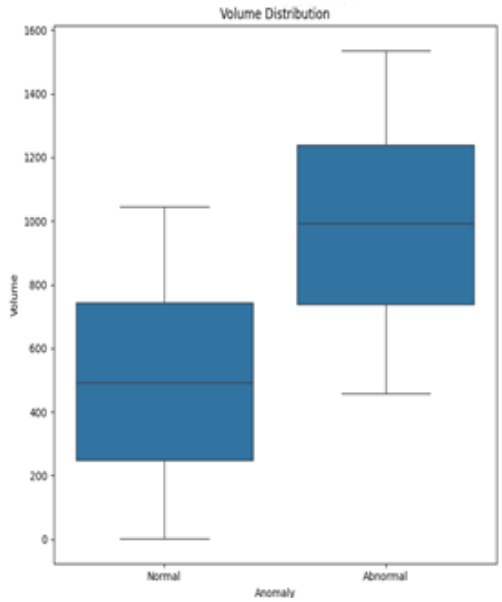
<표 3A> 각 변수의 기초통계량(3)

| 변수명 | Price_MA | Price_STD | Price_ Z_Score |
|---|---|---|---|
| 표본 수 | 9991 | 9991 | 9991 |
| 평균 | 1039.63 | 520.07 | 0 |
| 최소값 | 432.54 | 156.45 | -2.49 |
| 25% 값 | 926.36 | 461.5 | -0.79 |
| 50% 값 | 1037.3 | 522.92 | -0.01 |
| 75% 값 | 1151.46 | 581.62 | 0.79 |
| 최대값 | 1597.85 | 776.39 | 2.6 |
| 표준편차 | 166.65 | 87.84 | 0.95 |

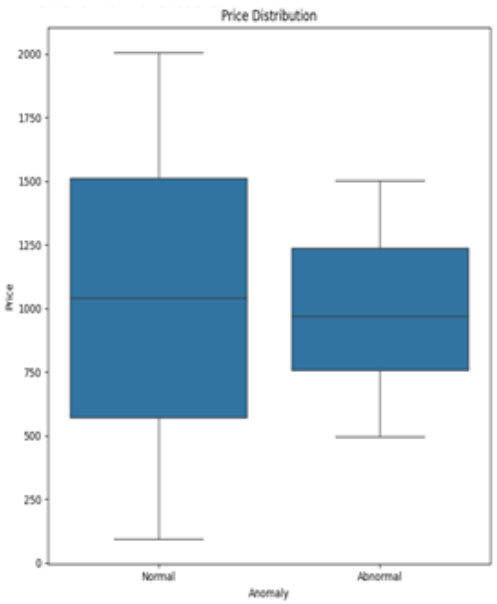
<표 3B> 각 변수의 기초통계량(4)

| 변수명 | Volume_ MA | Volume_ STD | Volume_ Z_Score |
|---|---|---|---|
| 표본 수 | 9991 | 9991 | 9991 |
| 평균 | 546.9 | 317.6 | 0 |
| 최소값 | 230.5 | 81.32 | -2.45 |
| 25% 값 | 477.6 | 272.71 | -0.78 |
| 50% 값 | 546.1 | 313.07 | -0.02 |
| 75% 값 | 613.9 | 359.08 | 0.74 |
| 최대값 | 899.9 | 545.79 | 2.71 |
| 표준편차 | 101.16 | 65.96 | 0.95 |

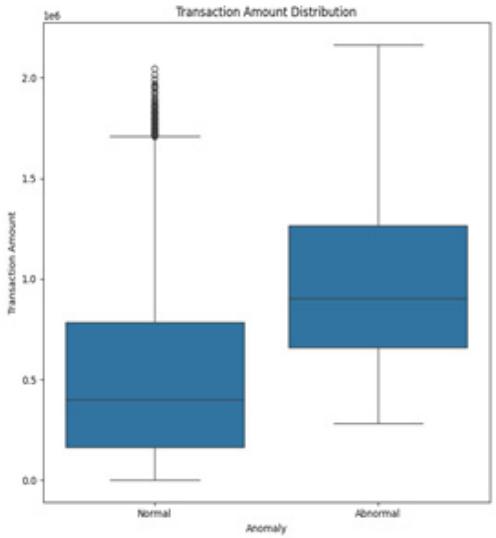
정상 거래와 비정상 거래의 특성을 다음의 4개 변수, Volume(거래량), Price(가격), Transaction Amount(거래량), Price Change Percentage(가격 변화율)의 분포로 비교했다.



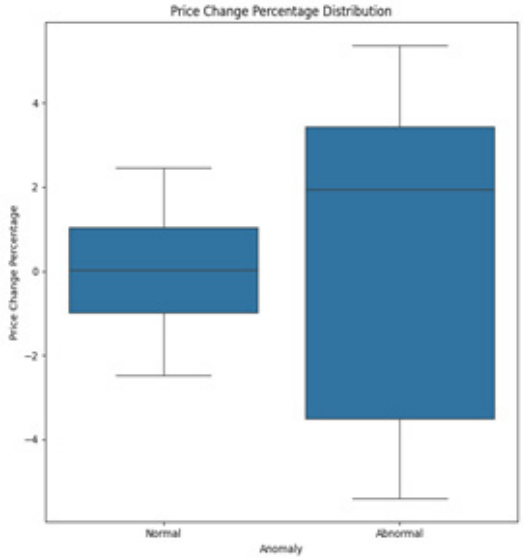

[그림 9A] 정상 및 비정상 거래 거래량 비교




[그림 9B] 정상 및 비정상 거래 기격분포 비교




[그림 10A] 정상 및 비정상 거래 거래액 비교




[그림 10B] 정상 및 비정상 거래 가격 변화율 비교

거래량 분포를 보면, 비정상 거래의 중앙값과 범위가 정상 거래보다 높다. 이는 비정상 거래에서 더 큰 규모의 거래가 이루어짐을 시사한다. 가격 분포에서는 비정상 거래의 중앙값이 정상 거래보다 낮지만, 그 범위는 더 넓다. 이는 비정상 거래에서 가

격 변동성이 더 크다는 것을 나타낸다.

거래 금액 분포를 보면, 비정상 거래의 중앙값과 범위가 정상 거래보다 훨씬 높다. 이는 비정상 거래에서 더 큰 금액의 거래가 이루어짐을 의미한다. 가격 변화 비율 분포에서는 비정상 거래의 범위가 정상 거래보다 훨씬 넓다. 이는 비정상 거래에서 가격 변동이 더 극단적으로 일어남을 보여준다.

전반적으로 비정상 거래가 정상 거래에 비해 더 큰 규모, 더 높은 변동성, 더 극단적인 가격 변화를 보이는 경향이 있음을 나타낸다. 이러한 특성은 비정상 거래를 식별하고 모니터링하는 데 중요한 지표가 될 수 있다.

## 3. 이상 탐지 모델 구현

본 연구에서는 적대적 데이터의 학습으로 인한 이상 탐지 모델의 성능 저하와 이에 대응한 통계 기반 필터링의 성과를 실증적으로 다루는 것을 목적으로 한다. 따라서 이상 탐지 모델로는 기본적인 머신러닝 기반의 모델인 SVM, 로지스틱 회귀 모델, 랜덤 포레스트 모델과 딥러닝 모델인 LSTM 모델을 선정했다. 이들 모델을 적용하여 이상 거래 탐지를 수행하고 그 성능을 비교 분석하기로 한다.

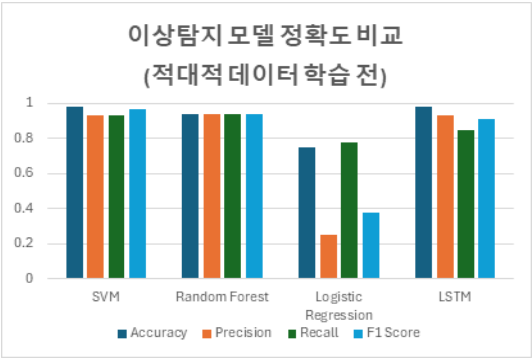
사용된 모델들의 주요 특징은 다음과 같다. SVM (Support Vector Machine) 모델은 데이터를 고차원 공간에 매핑하여 클래스 간 최적의 경계를 찾는 방식으로 작동한다. 복잡한 비선형 문제를 효과적으로 해결할 수 있는 장점이 있다.

랜덤포레스트(Random Forest) 모델은 여러 개의 결정 트리를 생성하고 그 결과를 종합하여 예측하는 앙상블 기법이다. 과적합에 강하고 특성 중요도를 쉽게 파악할 수 있는 장점이 있다. 로지스틱 회귀(Logistic Regression)모델은 입력 특성의 선형 조합을 사용하여 이진 분류를 수행한다. 모델이 단순하고 해석이 쉬운 장점이 있지만, 복잡한 비선형 관계를 포착하기 어려울 수 있다.

LSTM(Long Short-Term Memory) 모델은 순환 신경망의 일종으로, 시계열 데이터나 순차 데이터 처리에 특화되어 있다. 장기 의존성을 효과적으로 학습할 수 있는 장점이 있다.

모델의 평가 기준으로 정확도(Accuracy), 정밀도(Precsion), 재현율(Recall), F1 스코어를 사용했다. 이들 지표는 기계학습 모델의 성능을 평가하는 주요 지표들이다. 실제 금융 거래 데이터는 본질적으로 정상 거래에 비해 이상 거래의 비율이 매우 낮은 불균형 특성을 가지고 있다. 이러한 특성을 고려하여, 재현율(Recall)을 데이터 불균형성을 고려해서 “이상탐지”거래에 대해서 실제로 올바르게 탐지했는지 평가 하는 것(실제 이상 탐지율)으로 산정했다.

모델 성능 평가를 위해서는 데이터를 분할해야 하는데, 훈련 데이터와 테스트 데이터에서 이상 거래와 정상 거래의 비율을 동일하게 유지했다. 따라서 훈련 데이터의 이상 거래 비율은 10.09%, 테스트 데이터의 이상 거래 비율은 10.10%로 거의 동일하게 유지되었다.




[그림 11] 이상탐지 모델의 성능 평가(적대적 데이터 학습 전)

<표 4> 이상탐지 모델의 성능 평가 결과

| 모델 | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| SVM | 0.98 | 0.93 | 0.93 | 0.97 |
| Random Forest | 0.94 | 0.94 | 0.94 | 0.94 |

| 모델 | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Logistic Regression | 0.75 | 0.25 | 0.78 | 0.38 |
| LSTM | 0.98 | 0.93 | 0.85 | 0.91 |

이상 탐지 모델 성능 평가 결과, SVM 모델은 0.9765의 높은 정확도를 보였으며, 특히 0.9653의 높은 재현율로 실제 이상 거래의 대부분을 탐지했다. Random Forest 모델은 가장 높은 0.9875의 정확도를 달성했고, 정밀도와 재현율 사이의 균형이 좋아 0.9383의 F1 점수를 기록했다. 로지스틱 회귀 모델은 상대적으로 낮은 성능을 보였는데, 0.7465의 정확도와 0.2536의 낮은 정밀도를 나타냈다. LSTM 모델은 0.9815의 높은 정확도와 0.9609의 높은 정밀도를 보였지만, 재현율은 0.8515로 다소 낮았다.

전반적으로 Random Forest 모델이 가장 우수한 성능을 보였으며, SVM과 LSTM 모델도 높은 정확도를 달성했다. 로지스틱 회귀 모델은 이 문제에 대해 상대적으로 적합하지 않은 것으로 나타났다.

## 4. 적대적 데이터의 학습 및 모델의 성능 저하 시뮬레이션

금융 시장에서는 주식의 작전거래, 내부자거래, 암호화폐의 자전거래 등 불법적인 거래 행위가 지속적으로 발생하고 있다. 이러한 이상 거래를 탐지하기 위해 금융 기관들은 다양한 모델을 개발하여 사용하고 있으나, 불법 행위자들은 이를 회피하기 위해 더욱 정교한 방법을 고안하고 있다. 여기서는 뉴스의 사례를 참조해서 적대적 데이터를 생성한다.

뉴스의 시례를 분석하면 금융 데이터 공격 유형은 크게 네 가지로 구분된다. 거래량 조작 공격은 시장의 유동성이나 관심도를 조작하여 다른 투자자들의 행동에 영향을 미치는 것이 목적이다.

가격 스파이킹 공격은 단기간에 급격한 가격 변동을 일으켜 시장을 교란하거나 이익을 취하는 것


[그림 12] 금융 시장에서의 이상 거래 사례
(출처: 구글뉴스)

이 목적이다. 거래 금액 왜곡 공격은 거래의 규모를 과장하거나 축소하여 시장 참여자들의 인식을 조작하는 것이 목적이다. 추세 조작 공격은 장기적인 가격 추세를 인위적으로 만들어 시장 심리에 영향을 주는 것이 목적이다.

불법 행위자들은 이상 거래 탐지 시스템을 우회하기 위해 일정 기간에 걸쳐 조직적으로 일정 비중의 "이상 거래"를 정상적인 거래와 섞어 실행하는 전략을 사용한다. 이를 통해 이상 거래의 패턴이 정상 거래의 패턴과 유사해지도록 만든다. 생성된 적대적 데이터는 거래량, 가격, 거래 금액이 정상 데이터보다 평균적으로 3.5배 높고, 가격 변화율의 변동 폭이 정상 데이터보다 크며 평균적으로 5.8배 높다. 또한 이상치의 비율이 정상 데이터보다 높은 특징을 가진다.

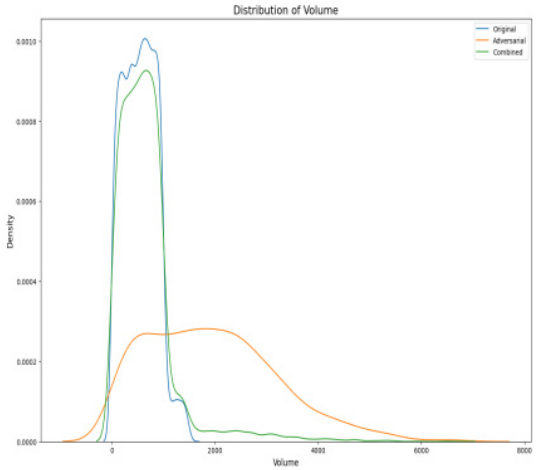
<표 5> 적대적 공격 유형

| 공격 유형 | Volume | Price | Transaction Amount | Price Change Percentage |
|---|---|---|---|---|
| 거래량 조작 | 크게 증가 | 간접적 변동 가능 | 증가 (개별 금액 작음) | 큰 변화 없음 |
| 가격 스파이킹 | 일시적 증가 | 급격한 변동 | 크게 증가 | 매우 큰 변동 |
| 거래 금액 왜곡 | 소폭 증가 | 영향 적음 | 크게 왜곡됨 | 큰 변화 없음 |

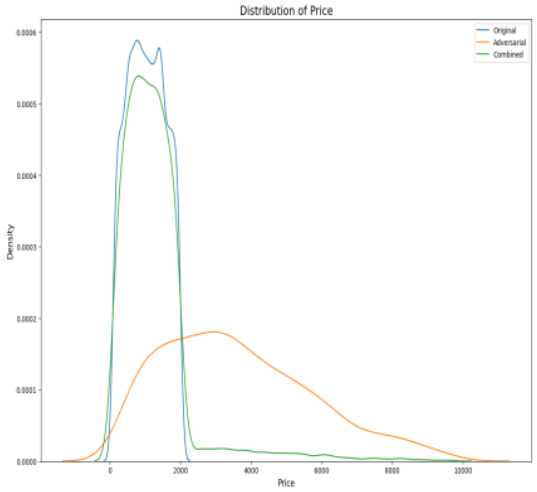
| 공격 유형 | Volume | Price | Transaction Amount | Price Change Percentage |
|---|---|---|---|---|
| 추세 조작 | 점진적 증가 | 서서히 변동 | 점진적 증가 | 지속적 일방향 변화 |


[그림 13] 생성된 적대적 데이터의 정상・이상 거래 분포

아래 그래프들은 거래량과 가격의 분포를 보여주는 것이다. 생성된 데이터는 총 1,000개이며, 이중 적대적 데이터는 81건으로 생성된 적대적 데이터의




[그림 14A] 적대적 데이터 및 기존 학습 데이터 비교 (거래량)




[그림 14B] 적대적 데이터 및 기존 학습 데이터 비교 (가격 분포)

8.1%의 비중을 가진다. 첫 번째 그래프는 거래량 분포를, 두 번째 그래프는 가격 분포를 나타낸다. 각 그래프에는 '원본(Original)', '적대적(Adversarial)', '결합(Combined)' 세 가지 유형의 데이터 분포가 표시되어 있다.

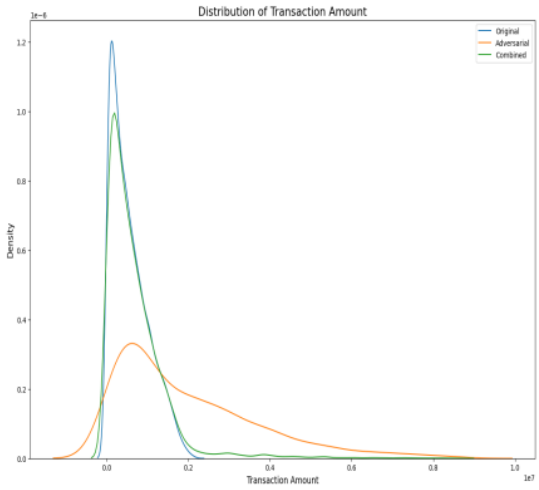
원본 데이터는 넓은 범위에 걸쳐 완만한 분포를 보이는 반면, 적대적 데이터와 결합 데이터는 좁은 범위에 집중된 뾰족한 분포를 나타낸다. 이는 적대적 거래가 특정 범위의 거래량과 가격에 집중되어 있음을 의미한다.

다음 그래프는 거래 금액과 가격 변화율의 분포를 보여주는 것이다. [그림 14A], [그림 14B]와 동일하게 두 그래프 모두 '원본(Original)', '적대적(Adversarial)', '결합(Combined)' 세 가지 유형의 데이터 분포를 포함하고 있다.

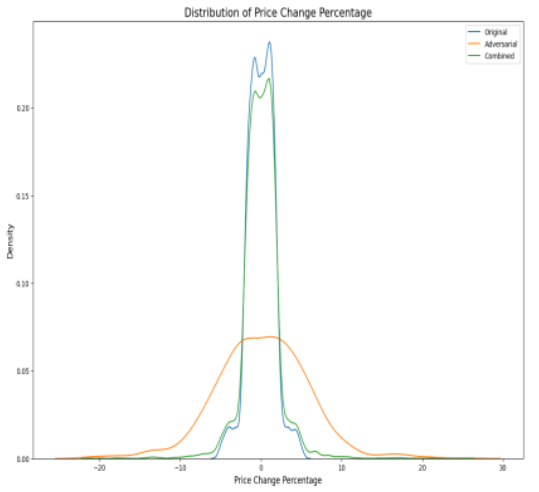
거래 금액 분포 그래프에서는 원본 데이터가 넓은 범위에 걸쳐 완만한 분포를 보이는 반면, 적대적 데이터와 결합 데이터는 좁은 범위에 집중된 높고 뾰족한 분포를 나타낸다. 이는 적대적 거래가 특정 범위의 거래 금액에 집중되어 있음을 보여준다.

가격 변화율 분포 그래프에서는 원본 데이터가 0

을 중심으로 좁고 높은 분포를 보이는 반면, 적대적 데이터는 더 넓은 범위에 걸쳐 낮고 완만한 분포를 나타낸다. 결합 데이터는 이 두 분포의 특성을 모두 반영하고 있다. 이는 적대적 거래가 정상 거래에 비해 더 큰 가격 변동성을 가지고 있음을 보여준다.

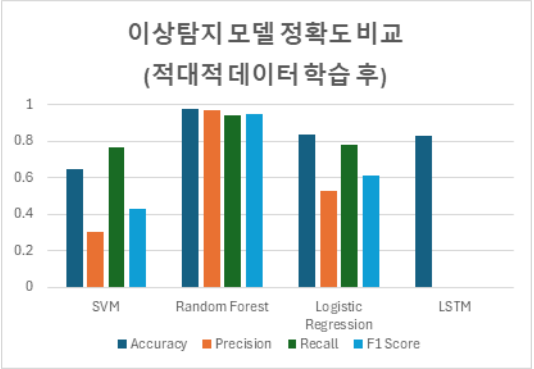



[그림 15A] 적대적 데이터 및 기존 학습 데이터 비교(거래액 및 가격 변화율 분포)




[그림 15B] 적대적 데이터 및 기존 학습 데이터 비교(거래액 및 가격 변화율 분포)

적대적 데이터 학습 후 모델 성능에 대한 분석 결과는 [그림 16]과 같다. 전반적으로 모든 모델의 성능이 감소하였다. 이는 적대적 데이터가 모델의 일반화 능력을 저하시켰음을 시사한다. 특히 SVM과 LSTM 모델의 성능 저하가 두드러졌다.




[그림 16] 이상탐지 모델의 성능 평가 (적대적 데이터 학습 후)

<표 6> 이상탐지 모델의 성능 평가 결과 (적대적 데이터 학습 후)

| 모델 | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| SVM | 0.65 | 0.30 | 0.77 | 0.43 |
| Random Forest | 0.98 | 0.97 | 0.94 | 0.95 |
| Logistic Regression | 0.84 | 0.53 | 0.78 | 0.61 |
| LSTM | 0.83 | 0 | 0 | 0 |

Random Forest 모델은 상대적으로 적대적 데이터에 대한 견고성을 보였다. 모든 평가 지표에서 0.94 이상의 높은 성능을 유지하였으나, 이전 대비 소폭 하락하였다. 이는 앙상블 방법의 특성상 개별 결정 트리의 과적합을 방지하여 적대적 공격에 일정 수준 저항할 수 있음을 보여준다.

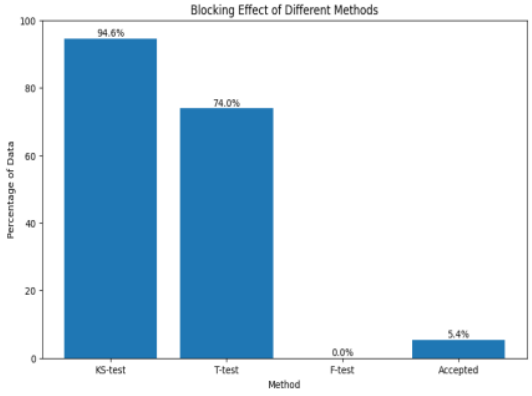
Logistic Regression 모델은 중간 정도의 성능 저하를 보였다. Accuracy와 Recall은 비교적 잘 유지되었으나, Precision이 크게 하락하였다. 이는 모델이 적대적 데이터에 대해 오탐(false positive)을 더

많이 발생시키게 되었음을 의미한다.

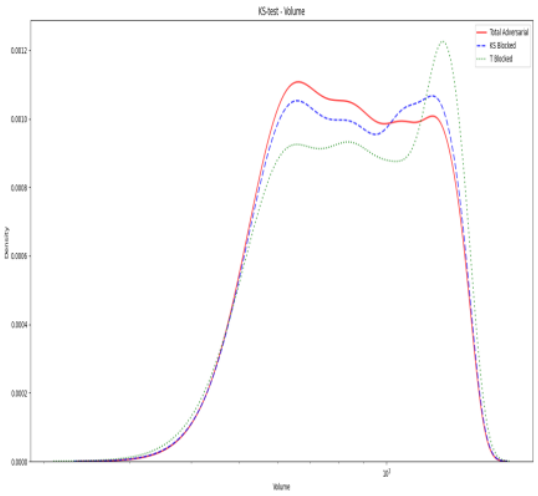
LSTM 모델의 경우 Accuracy를 제외한 모든 지표에서 극단적인 성능 저하가 발생하였다. 이는 시퀀스 데이터를 다루는 LSTM의 특성상 적대적 데이터에 더욱 취약할 수 있음을 나타낸다.

## 5. 통계 기반의 적대적 데이터 필터링 모델

여기서는 적대적 데이터 차단을 위해 다음의 세 가지 통계 모델 기반의 검정 방법을 사용하였다. KS-test는 94.59%의 적대적 데이터 차단율을 기록하며 가장 강력한 필터링 효과를 보였다. 이는 적대적 데이터가 전반적인 분포 형태를 크게 왜곡했음을 시사한다. T-test는 73.96%의 차단율을 기록하여, 적대적 데이터가 정상 데이터와 평균적으로 유의미한 차이를 보였음을 나타냈다. 반면 F-test는 0%의 차단율을 보여, 적대적 데이터가 정상 데이터와 유사한 분산을 유지하고 있었음을 시사한다.




[그림 17] 적대적 데이터 필터링 모델의 차단 효과[8]



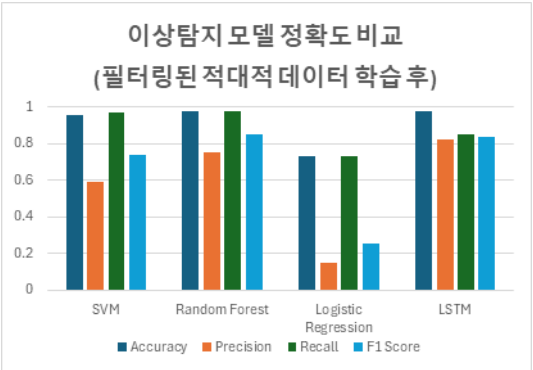

[그림 18] 적대적 데이터 필터링 모델의 차단 효과: 적대적 데이터 중 거래량 변수 비교

이 결과들은 적대적 데이터가 주로 분포의 형태와 평균을 변화시키는 방식으로 생성되었음을 보여준다. KS-test와 T-test는 각각 높은 차단율을 보이며, 적대적 데이터 필터링에 있어 매우 효과적인 방법임을 입증하였다. 그러나 F-test의 낮은 차단율은 분산 기반 공격에 대한 잠재적 취약성을 나타내며, 이를 보완하기 위한 추가적인 방안이 필요함을 시사한다. 이를 통해 다양한 유형의 적대적 데이터를 효과적으로 탐지하고 차단하기 위해, 각 방법의 특성을 고려한 통합적 접근이 필요하다.

## 6. 필터링된 적대적 데이터 학습 결과

[그림 19]는 적대적 데이터를 KS-test 모델로 필터링한 후 이상탐지 모델의 학습 결과를 보여주는 그래프이다. 전반적으로 모델들의 성능에 변화가 있었으며, 일부 지표에서는 향상을 보인 반면 다른 지표에서는 성능이 저하되었다.

8) 최종적으로 94.59%의 적대적 데이터 차단율을 보인 KS-Test를 적용해서 적대적 데이터의 5.41%의 데이터만 모델 학습에 사용했다.




[그림 19] 이상탐지 모델의 성능 평가
(필터링된 적대적 데이터 학습 후)

<표 7> 이상탐지 모델의 성능 평가 결과
(필터링된 적대적 데이터 학습 후)

| 모델 | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| SVM | 0.96 | 0.59 | 0.97 | 0.74 |
| Random Forest | 0.98 | 0.75 | 0.98 | 0.85 |
| Logistic Regression | 0.73 | 0.15 | 0.73 | 0.25 |
| LSTM | 0.98 | 0.82 | 0.85 | 0.84 |

SVM 모델은 Accuracy와 Recall에서 눈에 띄는 향상을 보였다. 특히 Recall이 0.97로 매우 높아졌다. 그러나 Precision은 0.59로 낮아져 균형이 맞지 않는 결과를 보여주고 있다. Random Forest 모델은 전반적으로 안정적인 성능을 유지했다. Accuracy, Recall, F1 Score에서 모두 0.85 이상의 높은 점수를 기록했다. 다만 Precision이 0.75로 다소 낮아진 점이 주목된다.

Logistic Regression 모델은 전반적으로 성능이 저하되었다. 특히 Precision이 0.15로 크게 떨어져 우려되는 부분이다. 다른 지표들도 대체로 낮은 수준을 보이고 있어 개선이 필요해 보인다.

LSTM 모델은 이전 결과와 비교해 큰 변화를 보였다. 모든 지표에서 데이터가 제공되었으며, 특히 Accuracy와 Recall에서 높은 성능을 보여주고 있다. 그러나 Precision은 0.82로 다른 지표에 비해 상대적으로 낮은 편이다.

이러한 결과는 적대적 데이터 필터링이 모델 성능에 복합적인 영향을 미쳤음을 시사한다. 일부 모델과 지표에서는 성능 향상이 있었지만, 다른 부분에서는 오히려 성능 저하가 발생했다. 따라서 이상탐지 시스템의 전반적인 성능을 개선하기 위해서는 각 모델의 장단점을 고려한 추가적인 최적화 작업이 필요할 것으로 보인다.

## 7. 소결

적대적 데이터 학습 전 이상탐지 모델의 성능에서는 SVM과 LSTM이 가장 우수한 결과를 보였다. 특히 Accuracy 측면에서 두 모델 모두 0.98로 높은 수치를 기록했다. Random Forest도 0.94의 안정적인 성능을 보였으나, Logistic Regression은 상대적으로 낮은 성능을 보였다. 이는 이상 탐지 작업의 특성상 비선형적인 결정 경계를 잘 표현할 수 있는 SVM과 LSTM이 유리했기 때문으로 해석된다.

<표 8> 종합 결과

| 모델 | 성능 지표 | 적대적 데이터 학습 전 | 적대적 데이터 학습 후 | 적대적 데이터 학습 후 (필터링) |
|---|---|---|---|---|
| SVM | Accuracy | 0.98 | 0.65 | 0.96 |
| | Precision | 0.93 | 0.3 | 0.59 |
| | Recall | 0.93 | 0.77 | 0.97 |
| | F1Score | 0.97 | 0.43 | 0.74 |
| Random Forest | Accuracy | 0.94 | 0.98 | 0.98 |
| | Precision | 0.94 | 0.97 | 0.75 |
| | Recall | 0.94 | 0.94 | 0.98 |
| | F1Score | 0.94 | 0.95 | 0.85 |
| Logistic Regression | Accuracy | 0.75 | 0.84 | 0.73 |
| | Precision | 0.25 | 0.53 | 0.15 |

| 모델 | 성능 지표 | 적대적 데이터 학습 전 | 적대적 데이터 학습 후 | 적대적 데이터 학습 후 (필터링) |
|---|---|---|---|---|
| Logistic Regression | Recall | 0.78 | 0.78 | 0.73 |
| | F1Score | 0.38 | 0.61 | 0.25 |
| LSTM | Accuracy | 0.98 | 0.83 | 0.98 |
| | Precision | 0.93 | 0 | 0.82 |
| | Recall | 0.85 | 0 | 0.85 |
| | F1Score | 0.91 | 0 | 0.84 |

적대적 데이터를 학습시킨 후의 결과는 모델별로 큰 차이를 보였다. Random Forest 모델은 오히려 성능이 소폭 상승하여 0.98의 Accuracy를 기록했다. 이는 Random Forest의 앙상블 특성이 적대적 데이터에 대해 강건성을 제공한 것으로 보인다. 반면, SVM과 LSTM 모델은 성능이 크게 저하되었다. 특히 LSTM의 경우 Precision, Recall, F1 Score가 모두 0으로 떨어져, 적대적 데이터에 매우 취약한 것으로 나타났다. 이는 LSTM의 복잡한 구조가 적대적 데이터에 과적합되었을 가능성을 시사한다.

통계 필터링을 통해 적대적 데이터를 처리한 후의 결과는 대부분의 모델에서 성능 회복을 보였다. SVM과 LSTM은 원래 모델에 근접한 성능을 회복했으며, 특히 LSTM의 경우 Accuracy가 0.98로 원래 수준으로 돌아왔다. 그러나 Precision과 Recall 측면에서는 여전히 개선의 여지가 있다. Random Forest 모델은 세 가지 시나리오에서 모두 안정적인 성능을 유지했으며, 이는 Random Forest가 적대적 공격에 가장 강건한 모델임을 시사한다.

Logistic Regression 모델은 세 가지 시나리오 모두에서 상대적으로 낮은 성능을 보였다. 이는 Logistic Regression의 선형적 특성이 복잡한 이상 탐지 작업에 적합하지 않음을 나타낸다. 특히 Precision이 매우 낮아, 정상 데이터를 이상으로 잘못 분류하는 경우가 많았던 것으로 보인다.

이 실험 결과는 이상 탐지 모델의 선택과 적대적 공격에 대한 대응 전략 수립에 중요한 시사점을 제공한다. 그러나 더 다양한 유형의 적대적 공격과 더 많은 데이터 세트에 대한 검증이 필요하며, 모델의 해석 가능성과 실제 환경에서의 적용 가능성 등을 추가로 고려해야 한다.

## Ⅳ. 결론

### 1. 결론 및 시사점

본 연구에서는 금융 거래 데이터의 특성을 고려한 적대적 공격 방어 모델을 개발하고, 이에 대한 성과를 실증적으로 분석하였다. 연구 결과, 다음과 같은 주요 결론을 도출할 수 있었다.

첫째, 적대적 데이터의 학습은 이상 탐지 모델의 성능을 크게 저하시킬 수 있음을 확인하였다. 특히 SVM과 LSTM 모델은 적대적 데이터에 매우 취약한 것으로 나타났다. 반면, Random Forest 모델은 적대적 데이터에 대해 상대적으로 강건한 성능을 유지하였다. 이는 앙상블 방법의 특성이 적대적 공격에 대한 일정 수준의 저항력을 제공함을 시사한다.

둘째, 통계 기반의 적대적 데이터 필터링 방법이 모델 성능 회복에 효과적임을 입증하였다. KS-test를 이용한 필터링 방법은 94.59%의 높은 적대적 데이터 차단율을 보였으며, 이를 통해 대부분의 모델에서 성능 회복이 관찰되었다. 특히 SVM과 LSTM 모델은 필터링 후 원래 수준에 근접한 성능을 회복하였다.

셋째, 모델별로 적대적 공격에 대한 반응이 상이함을 확인하였다. Random Forest 모델은 세 가지 시나리오(정상, 적대적 데이터 학습, 필터링 후 학습) 모두에서 안정적인 성능을 유지하여 가장 강건한 모델로 평가되었다. 반면, Logistic Regression 모델은 모든 시나리오에서 상대적으로 낮은 성능을

보여, 복잡한 이상 탐지 작업에 적합하지 않음이 드러났다.

넷째, 적대적 데이터의 특성이 모델 성능에 미치는 영향을 분석하였다. 생성된 적대적 데이터는 거래량, 가격, 거래 금액이 정상 데이터보다 평균적으로 3.5배 높고, 가격 변화율의 변동 폭이 5.8배 높은 특징을 가졌다. 이러한 특성은 KS-test와 T-test를 통해 효과적으로 탐지되었으나, F-test에서는 탐지되지 않았다. 이는 적대적 데이터가 주로 분포의 형태와 평균을 변화시키는 방식으로 생성되었음을 시사한다.

다섯째, 데이터 불균형성을 고려한 성과 지표의 중요성을 확인하였다. 특히 Recall(실제 이상 탐지율)을 주요 성과 지표로 활용함으로써, 모델의 실질적인 이상 탐지 능력을 더욱 정확하게 평가할 수 있었다.

이러한 연구 결과는 금융 기관들이 이상 거래 탐지 시스템을 개선하고, 적대적 공격에 대비하는 데 유용한 통찰을 제공할 것으로 기대된다. 특히, 다양한 모델의 장단점을 고려한 앙상블 접근법과 통계적 필터링 방법의 결합이 효과적인 방어 전략이 될 수 있음을 시사한다.

## 2. 한계점 및 보완 방향

본 연구의 한계점과 향후 개선 방향은 다음과 같다. 첫째, 생성된 적대적 데이터의 다양성이 부족하다. 실제 금융 시장의 복잡성을 완전히 반영하지 못했을 가능성이 있다. 이를 개선하기 위해 적대적 데이터의 특징을 더욱 섬세하게 조정하고, 다양한 이상 거래 유형 및 시간적 요소와 거래 패턴의 복잡성을 강화할 필요가 있다.

둘째, 외부 요인을 고려하지 않았다. 순수하게 거래 데이터만을 기반으로 분석을 진행했으나, 실제 금융 시장에서는 뉴스, 경제 지표, 정책 변화 등 다양한 외부 요인이 이상 거래에 영향을 미칠 수 있다. 향후 연구에서는 이러한 외부 요인과 연동된 이상 거래 패턴을 생성하여 분석할 필요가 있다.

셋째, 모델의 적응성이 부족하다. 본 연구의 모델들은 정적인 특성을 가지고 있어, 변화하는 공격 패턴에 실시간으로 대응하기 어렵다. 향후 연구에서는 탐지 시스템의 반응에 따라 실시간으로 전략을 조정하는 적응형 전략을 도입할 필요가 있다.

넷째, 데이터 품질 문제를 고려하지 않았다. 실제 금융 데이터에서 발생할 수 있는 누락, 오류, 지연 등의 문제를 시뮬레이션에 포함시키지 않았다. 향후 연구에서는 이러한 데이터 품질 문제를 시뮬레이션에 포함시켜 더욱 현실적인 환경을 구현할 필요가 있다.

다섯째, 모델 평가의 범위가 제한적이다. 본 연구에서는 제한된 수의 모델과 평가 지표만을 사용하였다. 향후 연구에서는 더 다양한 모델과 평가 지표를 사용하여 포괄적인 평가를 수행할 필요가 있다.

마지막으로, 실제 환경 적용의 한계가 있다. 시뮬레이션 환경에서의 결과가 실제 금융 환경에서도 동일하게 적용될 수 있을지에 대한 검증이 필요하다. 향후 연구에서는 실제 금융 환경에서의 실증적 검증을 통해 연구 결과의 실용성을 높일 필요가 있다.

본 연구는 금융 거래 데이터의 특성을 고려한 적대적 공격 방어 모델을 개발하고 그 성과를 실증적으로 분석했다. 연구의 범위는 다양한 기계학습 모델(SVM, LSTM, Random Forest, Logistic Regression)을 이용한 이상 거래 탐지와 적대적 데이터에 대한 방어 방안을 제시했으며, 통계 기반의 적대적 데이터 필터링 방법의 효과성을 검증하고, 모델별 성능 차이를 분석했다.

이러한 분석 결과를 바탕으로 금융 기관들이 이상 거래 탐지 시스템을 개선하고 적대적 공격에 대비하는 데 유용한 통찰을 제공할 것으로 기대한다.

다만 시뮬레이션 환경의 한계로 인해 실제 금융 환경에서의 검증이 추가로 필요하며, 향후 연구에서

는 더 복잡하고 현실적인 시나리오를 고려할 필요가 있다.

## 참고문헌


권현, 윤현수, 최대선 (2018). Evasion attack에 대한 인공지능 보안이슈. 정보과학회지 136(2), 32-36.

김휘영, 정대철, 최병욱 (2019). 딥러닝 기반 의료 영상 인공지능 모델의 취약성: 적대적 공격 대한 영상의학회지 80(2), 259-273.

심우민 (2019). 한국의 인공지능 알고리즘 담론, KISA Report.

안상선 (2024). 가상자산 시계열 데이터의 분석과 예측-시계열 모델 및 인공지능 모델의 비교 및 개선점 제시. 미래사회, 15(1), 21-33.

CARRARA, Fabio, et al. (2018). Adversarial examples detection in features distancespaces. *In: Proceedings of the European Conference on Computer Vision* (ECCV).

C.Szegedy et al., *Intriguing Properties of Neural Networks,* ICLR'14

LU, Jiajun; ISSARANON, Theerasit; FORSYTH, David. Safetynet: Detecting andrejecting adversarial examples robustly. In: Proceedings of the IEEE *International Conference on Computer Vision*. 2017, 446-454.

Nicolas Papernot (2018). *Security and Privacy in Machine Learning.*

M.D.Zeiler and R.Fergus, Visualizing and Understanding Convolutional Networks, ECCV'14

Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning, arXiv:1804.00308v1 [cs.CR] 1 Apr 2018

METZEN, Jan Hendrik, et al. *On detecting adversarial perturbations*. arXivpreprint arXiv: 702. 04267, 2017.

Sandip Kundu, Security and Privacy of Machine Learning Algorithms, National Science Foundation leave from University of Massachusetts, Amherst, 2019

정보통신정책연구원 2021-3호 인공지능: 사이버보안 패러다임의 전환

http://news.khan.co.kr/kh_news/khan_art_view.html?art_id=201609211731001 자율주행 전기차, 해킹에 '원격조종' 당했다.

https://www.boannews.com/media/view.sp?idx=56278 자율주행차, 테이프만으로 해킹?


투고일자: 2024. 8. 31.
심사일자: 2024. 9. 24.
게재확정일자: 2024. 10. 7.

미래사회
*Journal of Future Society*
2024. Vol. 15, No. 3, 147~166.
https://doi.org/10.22987/jifso.2024.15.3.147

# Statistical-based Countermeasures for Adversarial Data to Secure Anomaly Detection Models in the Financial Sector

Ahn, Sang-Sun

Kookmin University

This study aims to improve the performance of anomaly detection models in financial transaction data and enhance their robustness against adversarial attacks. We constructed Support Vector Machine (SVM), logistic regression, random forest, and Long Short-Term Memory(LSTM) models based on securities trading data and analyzed the impact of adversarial data on these models. We propose a filtering method that utilize statistical techniques, such as Kolmogorov-Smirnov test, t-test, and F-test, to effectively block adversarial data. The proposed method was empirically verified for its ability to improve the model performance. Considering the data imbalance characteristic of anomaly detection in the financial sector, we used Recall (True Anomaly Detection Rate) as the primary performance indicator. Our results confirmed that the proposed statistical filtering method significantly enhances the robustness of the models against adversarial data, showing meaningful improvements, particularly in terms of recall. This study contributes to the development of secure and efficient anomaly detection systems in the financial sector.

*Keywords: Artificial Intelligence, Anomaly Detection, Adversarial Attack, Data Imbalance, Statistical Filtering*