JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 27 No. 3, pp. 173-182, March 2022
https://doi.org/10.9708/jksci.2022.27.03.173

A Study on Fraud Detection in the C2C Used Trade Market Using Doc2vec

Do Hyun Lim*, Hyunchul Ahn*

*Master's Candidate, Graduate School of Business IT, Kookmin University, Seoul, Korea
*Professor, Graduate School of Business IT, Kookmin University, Seoul, Korea

[Abstract]

In this paper, we propose a machine learning model that can prevent fraudulent transactions in advance and interpret them using the XAI approach. For the experiment, we collected a real data set of 12,258 mobile phone sales posts from Joonggonara, a major domestic online C2C resale trading platform. Characteristics of the text corresponding to the post body were extracted using Doc2vec, dimensionality was reduced through PCA, and various derived variables were created based on previous research. To mitigate the data imbalance problem in the preprocessing stage, a complex sampling method that combines oversampling and undersampling was applied. Then, various machine learning models were built to detect fraudulent postings. As a result of the analysis, LightGBM showed the best performance compared to other machine learning models. And as a result of SHAP, if the price is unreasonably low compared to the market price and if there is no indication of the transaction area, there was a high probability that it was a fraudulent post. Also, high price, no safe transaction, the more the courier transaction, and the higher the ratio of 0 in the price also led to fraud.

▸ **Key words**: Fraud Detection, Online C2C Resale Market, Doc2vec, LightGBM, SHAP

[요 약]

본 논문에서는 사기 거래를 사전에 예방하고 XAI 접근 방식을 사용하여 해석할 수 있는 기계학습 모델을 제안한다. 실험을 위해 국내 주요 온라인 C2C 재판매 거래 플랫폼인 중고나라에서 휴대폰 판매 게시물 1만2,258개에 대한 실제 데이터셋을 수집했다. 게시물 본문에 해당하는 텍스트를 Doc2vec을 이용해 특성을 추출했고 PCA를 통해 차원축소를 했으며, 이전 연구를 바탕으로 다양한 파생변수가 만들어졌다. 전처리 단계에서 데이터 불균형 문제를 해결하기 위해 오버샘플링과 언더샘플링을 결합한 복합샘플링 방법이 적용되었다. 이러한 특성을 기반으로 사기성 게시물을 탐지하는 기계학습 모델들이 학습되었다. 분석 결과 LightGBM이 다른 기계학습 모델에 비해 가장 우수한 성능을 보였다. 그리고, SHAP을 이용한 분석 결과, 시세에 비해 터무니없게 가격이 쌀수록, 거래지역 표기가 없을수록, 가격이 높을수록, 안전거래를 하지 않을수록, 택배거래를 할수록, 가격 중 0의 비율이 많을수록 사기 게시글일 확률이 높았다.

▸ **주제어**: 사기 탐지, 온라인 C2C 중고거래 시장, Doc2vec, LightGBM, SHAP

• First Author: Do Hyun Lim, Corresponding Author: Hyunchul Ahn
*Do Hyun Lim (ehgus7011@naver.com), Graduate School of Business IT, Kookmin University
*Hyunchul Ahn (hcahn@kookmin.ac.kr), Graduate School of Business IT, Kookmin University
• Received: 2022. 02. 17, Revised: 2022. 02. 17, Accepted: 2022. 03. 04.

Copyright © 2022 The Korea Society of Computer and Information
http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945
www.kci.go.kr

I. Introduction

스마트폰의 대중적 보급화와 더불어 MZ세대를 중심으로 중고거래가 인기를 끌면서 C2C(Consumer to Consumer) 시장은 빠르게 성장하고 있다[1]. 중고 거래 시장 규모는 2008년 4조원에서 2020년 20조원을 달성하며 꾸준히 성장세에 있다. 특히, 코로나(COVID-19) 시대에 접어들면서 중고시장은 더욱 폭발적으로 성장했는데, 플랫폼 기업들이 안정성과 편의성을 보강한 모바일 플랫폼을 내놓기 시작하면서 거래 품목과 이용자층이 확대된 것으로 보인다[2]. 닐슨코리아의 조사에 따르면, 코로나 이후 국내 중고 거래 앱 사용자가 1,000만명을 상회하는 것으로 알려져 있다.

그러나 중고 거래 시장이 확대됨에 따라 자원의 선순환이라는 긍정적 측면 뿐만 아니라 사기 거래 건수 증가라는 부정적 측면도 두각을 나타냈다. 경찰청 조사에 따르면, 중고 거래 사기 피해는 2014년 총 4만 5,877건에서 2017년 6만 7,589건, 2020년 12만 3,168건으로 폭증했다[3].

피해자들의 피해사례를 신고받는 The Cheat라는 사이트에 따르면, 사기 피매 물품은 2006년에서 2022년까지 1위가 휴대폰 및 주변기기, 2위가 티켓 및 상품권, 3위가 패션 및 의류, 4위가 컴퓨터 및 주변기기, 5위가 가전 및 전자제품인 것으로 집계되었다[4].

중고 거래는 소비자가 구매, 판매 모두를 하는 거래 형태이다[5]. 소비자가 판매까지 하다 보니 판매자의 역할을 수행하는데 있어 정보의 비대칭성 문제가 생길 수밖에 없다. 판매자는 거래를 성사시키기 위하여 구매자에게 제품 품질에 대한 신뢰를 보이려고 노력하고 구매자는 정보의 비대칭성을 해결하기 위해 제품 상태, 판매자의 평판, 판매자의 주장의 질에 의존하게 된다[6]. 정보의 비대칭성 문제와 함께 거래가 일회성이라는 점을 사기꾼이 악용하여 사기 거래가 발생한다[7].

따라서 본 연구에서는 중고시장의 성장세와 함께 증가하는 사기 게시글의 특징을 파악하고 이를 토대로 한 기계학습(Machine Learning) 기반 탐지 모델을 제안한다. 특히 본 연구에서는 기존 연구에서 시도되지 않았던 Doc2vec을 활용해 더욱 효과적으로 게시글의 텍스트 정보를 기반으로 사기 거래 탐지의 정확도를 높이는 방안을 제시한다. 아울러 XAI 기법 중 하나인 SHAP을 이용해, 중고 사기 거래가 갖는 주요 특징에 대해 살펴본다.

II. Preliminaries

1. Related works

사기탐지를 위한 학계의 관심과 연구가 지속되었는데, 전통적인 사기탐지 연구는 금융사기 탐지 쪽이 많았다. 표 1은 사기탐지 분야의 선행연구들을 보여준다. 이에 관련된 연구로는 오토인코더를 전처리기로 활용하고 Random Forest로 분류를 시도함으로써 실시간으로 유입되는 데이터에 대해 지도 학습의 속도를 개선한 연구[8], 불법대출과 불법방문판매에 해당하는 글을 실시간으로 분류하는 알고리즘을 개발한 연구[9], 순차패턴 마이닝을 선급 전자지급수단의 금융서비스 거래내역에 적용함으로써 이상 금융거래를 탐지하는 모델을 만든 연구[10] 등이 있었다.

그러나 C2C 거래 플랫폼에서 사기 탐지 연구는 많이 이뤄지지 않았다. 온라인 C2C 거래에서 일어나는 사기를 탐지한 국외연구들은 대부분 쉴드 비딩(Shield Bidding)을 탐지한 연구들이다. 쉴드 비딩은 사기의 종류 중 하나로, 판매자가 한 명 혹은 다수 입찰자와 짜고 경매의 가격을 올리는 행위이다[11]. 대표적인 국외연구로는 경매사이트인 eBay의 데이터를 토대로 쉴드 비딩의 행위를 클러스터링하고 샘플링 방법을 적용 후, 다양한 기계학습 모델을 제안한 연구[11], SVM과 ANN으로 쉴드 비딩을 예측하고 Fuzzy logic으로 사기를 판별한 연구[12]가 있었다.

대표적인 국내연구로는 LDA로 게시글의 특성을 추출하고 XGBoost로 사기탐지 모델을 만든 연구[13], 더치트 API를 사용해 실제 사기 글의 사기패턴과 사기판별 요소를 밝힌

Table 1. Summary of the prior studies on Fraud detection area

Purpose	Technique	Ref.
To speed up supervised learning	Auto Encoder, Random Forest	[8]
To classify fraudulent posts	MLE, Term Frequency	[9]
To detect abnormal financial transactions	Sequential Pattern Analysis	[10]
To classify fraudulent posts	Clustering, Sampling	[11]
To classify fraudulent posts	SVM, ANN	[12]
To classify fraudulent posts	LDA, XGBoost	[13]

www.kci.go.kr

연구[14], 사기예방을 위해 C2C 중고거래 모바일 앱의 디자인을 제시한 연구[5]가 있었다. 이동우 외[13]는 본 연구와 주제 및 방법론에서 비슷한 점이 많으나 다음과 같은 한계점을 가졌다. 첫째, 텍스트 데이터를 임베딩하는 최신 방법론을 적용하지 못했다. 둘째, 게시글에는 다양한 정보가 있는데 이를 십분 활용하여 다양한 파생변수를 만들지 못했다.

이에 본 연구에서는 텍스트 데이터를 임베딩하는 비교적 최신 방법론인 Doc2vec을 적용하고 다양한 파생변수를 만들어 사기 게시글의 특징을 살펴볼 것이다.

2. Theoretical Background

2.1 Doc2vec

Doc2vec은 단어에서 단어 시퀀스로 임베딩 학습을 확장하기 위해 Word2vec에서 확장된 기법이다[30]. Doc2vec은 Word2vec에 비해 문서 전체에 대한 고려가 잘 이루어진다[15]. Doc2vec은 PV-DM모델과 PV-DBOW 모델로 나뉜다. Doc2vec에서 문장들은 고유한 값을 가진 단어 벡터로 맵핑된다. 맵핑된 단어 매트릭스는 합쳐져서 문장의 다음 단어를 예측하는 데 입력변수로 사용된다[16]. 이런 알고리즘을 사용하는 모델이 바로 PV-DM이다. PV-DM은 문장 벡터(Paragraph Vector)를 가지고 있는데, 문장 벡터는 문맥에서 빠진 정보를 기억하는 역할을 한다. 이때, 연구자가 벡터 크기(Vector Size)를 설정할 수 있는 데, 모델은 벡터 크기만큼의 문장을 학습하고 창(Window)만큼 옆으로 이동하면서 다음 단어를 예측하는 방식을 취한다[15]. 반면, PV-DBOW모델은 Word2vec의 Skig-gram모델과 비슷한데, 문장 벡터만으로 창 크기만큼의 단어를 무작위로 예측하는 방식이다[16].

2.2 LightGBM(Light Gradient Boosting Machine)

LightGBM은 Microsoft에서 배포한 오픈소스 기반의 GBDT(Gradient Boosting Decision Tree) 모델이다[17]. LightGBM은 기존의 기계학습 모델과 비교해 빠른 연산속도와 높은 정확도로 많은 인기를 얻고 있다. LightGBM은 기존의 GBDT 모델이 정보 이득(Information gain)의 평가를 위하여 모든 가능한 분기점을 나누는 방식을 택해야 했는데, 이 방식을 통해서라면 데이터를 모두 스캔해야 했기에 생기는 연산 복잡성이 높아지는 문제를 해결하고자 하는 시도에서 고안되었다. 정보이득이란 분류를 통한 정보의 취득을 수치화시킨 것인데, 데이터 구분에 대한 지표가 된다. LightGBM은 GBDT 모델의 연산 복잡성 문제를 GOSS(Gradient-based One-side Sampling)와 EFB(Exclusive Feature Bundling)으로 극복하였다.

GOSS는 데이터 중 기울기(Gradient)가 큰 부분은 유지하고 기울기(Gradient)가 작은 부분은 랜덤하게 제거하는 기술인데, 이를 통해 정보이득의 손실을 최소화하면서 데이터 수를 줄일 수 있었다. EFB는 효과적으로 변수의 수를 줄이는 기법으로, 변수 사이의 공간이 빔으로써 생기는 상호배타적인 속성을 가지는 변수들을 묶는 방식이다. LightGBM은 GOSS와 EFB를 통해 정확도를 거의 유지한 채 빠른 훈련 속도를 자랑한다[18].

2.3 SHAP(SHapley Additive exPlanations)

SHAP은 게임이론의 섀플리 값을 이용한 알고리즘이다. 섀플리 값은 모델의 예측값에 대한 변수의 기여도이다. 섀플리 값은 변수의 유무에 따라 다른 모델의 예측값 간의 차이에 가중치를 곱해 계산한다. 가중치는 가용 가능한 경우의 수 중에서 특정 변수를 뺏을 때의 경우의 수로 나누어 구한다[19]. 이렇게 SHAP을 사용하면 모델의 예측값에 대한 변수의 영향력과 입력변수에 따른 예측값의 변화 정도를 알 수 있다.

기존 기계학습 기법에서 변수 중요도는 변수를 돌아가며 값을 바꾸고, 변화가 예측에 주는 영향력을 계산한다. 하지만 이 기법은 변수 간의 의존성을 추정할 수가 없어서 상관관계가 있는 데이터는 변수 의존성 문제가 생긴다. 반면, SHAP은 변수 간 의존성을 고려하여 변수의 영향력을 계산하기 때문에 변수가 모델에 부(-)의 영향을 미친 것까지 보여준다[20].

III. Proposed Fraud Detection Model

1. Proposed classification model and structure

본 연구의 전체적인 모형은 그림 1과 같다. 우선 크롤링한 중고나라 데이터를 결측치가 있는 경우를 제외하고 파생변수를 만들었다. 범주형 변수는 원-핫 인코딩하였다. 텍스트 데이터에 해당하는 본문은 Doc2vec으로 벡터화한 후, 심재승 외[21]를 참조하여 주성분 분석(PCA)으로 차원축소를 하였다. 파생변수와 벡터를 합친 뒤, 특징 선택을 하였다. 모델링 과정에서는 훈련 데이터와 테스트 데이터를 만들기 위해 7:3의 비율로 나누어 주었다. 복합 샘플링이 적용된 훈련 데이터를 각각 LightGBM, XGBoost, Random Forest, SVM, Logistic Regression에 학습시켜 모델을 구축하고 테스트 데이터로 모델평가를 수행하였다. 마지막으로 XAI 기법인 SHAP을 적용해 변수의 영향력을 확인하였다. 실험은 Windows 10, RAM 16GB의 PC에서 Python 3.8 환경에서 수행되었다.

www.kci.go.kr

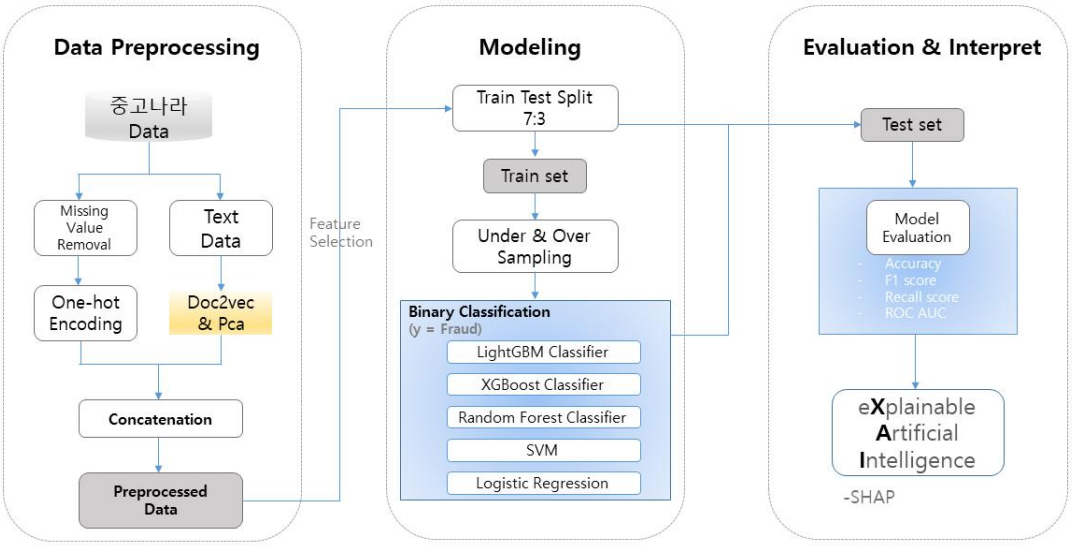

Fig. 1. Research Model

2. Data Organization and Preprocessing

2.1 Data Organization

'중고나라'는 'Naver'가 제공하고 있는 café형태의 온라인 C2C 거래 게시판이다. '중고나라'는 2020년 기준 거래 규모 5조원 수준으로[22] 2022년 2월 12일 기준 회원수 1,880만 명 이상이 가입되어 있다. '중고나라'에서는 판매자가 판매제품에 대한 정보를 업로드하고 구매자 게시판에서 게시글을 읽고 연락을 취하는 형태의 거래가 이루어지고 있다. '중고나라'에는 상품권부터 자동차까지 일상생활에 필요한 거의 모든 중고물품이 거래되고 있으며 물품 별로 범주화된 게시판이 존재하고 지역별 게시판도 있다. 게시글은 제목, 이미지, 가격, 상품상태, 결제 방법, 배송방법, 거래지역, 판매자 정보, 본문 등으로 구성되어 있으며 본문은 판매자가 자유롭게 작성할 수 있다. 또한, '중고나라'에서는 구매자가 상품구매 시, 안전결제를 할 수 있도록 네이버 페이를 통한 에스크로도 제공하고 있다.

2021년 8월부터 2021년 12월까지 약 4개월간 데이터를 수집하였으며 2021년 8월 18일부터 2021년 11월 22일까지의 게시글이 수집되었다. Python에서 Selenium 라이브러리를 활용하여 데이터를 수집하였다. '휴대폰 판매/매입' 게시판에서 휴대폰 기종 및 주변기기를 판매하는 게시글만을 대상으로 하였다. 휴대폰 및 주변기기 게시글만을 대상으로 한 이유는 사기 피해사례 검색 서비스를 제공하는 'The Cheat' 사이트의 통계에 따라 2006년에서 2022년까지 휴대폰 및 주변기기의 사기건수가 166,870건으로 피해물품 중 가장 많았기 때문이다. 회원등급이 '셀러회원'인 게시글의 경우는 데이터 수집 대상에서 제외하였는데, 중고나라 셀러회원은 중고나라에서 지정한 업자이기 때문에 사기 게시글을 올릴 가능성이 낮았기 때문이다. 셀러회원을 제외할 때는 '네이버 카페 필터링' 프로그램을 사용하였다. 총 12,258건의 게시글을 수집하였으며 URL, 글번호, 작성날짜, 닉네임, 가격, 제목, 본문, 연락처, 조회수, 댓글 수, 결제방법, 이미지 등을 수집하였다.

2.2 Data Labeling

사기 게시글을 탐지하기 위해서는 모델이 사기와 비사기를 학습할 수 있도록 타겟변수에 사기 여부를 라벨링 해주어야 한다. 라벨링 작업에는 사기 정보공유 사이트인 'The Cheat'의 API를 이용하였다. 'The Cheat'는 온라인 사기의 피해자가 사기꾼의 계좌번호와 휴대폰 번호 등 사기꾼 정보를 바탕으로 한 신고를 받는다. 'The Cheat'는 이런 정보를 바탕으로 다른 구매자가 물품 구매 전에 판매자의 휴대폰 번호 혹은 계좌번호 등을 조회할 수 있게 하여 현재 진행되는 거래가 사기인지 아닌지를 확인할 수 있는 서비스를 제공한다. 조회 서비스는 3개월 이내의 거래에서 사기꾼의 휴대폰 번호 혹은 계좌번호가 있어야만 검색이 되기 때문에 그 이전의 거래나 게시글에 휴대폰 정보가 없으면 사기 게시글 여부를 판단할 수 없다. 따라서 휴대폰 정보가 없거나 조회 시점에서 3개월이 지난 데이터를 제외한 후, 총 4,670건의 데이터를 분석에 사용하였다. 이 중, 사기 게시글로 분류된 데이터는 90건이다.

2.3 Data Preprocessing

판매 글만을 분석대상으로 했으므로 전처리 과정에서 예약 글, 교환 글, 제목이 중복되는 글 등을 삭제하였고 휴대폰이 아닌 휴대폰 케이스를 판매하는 글도 삭제하였다. 결측치가 하나라도 있으면 전체 행을 삭제하였다.

시세와 가격의 차이 변수는 총 2개로, 시세와 가격의 차이 변수는 중고폰 시세가에서 판매가격을 뺀 값에서 음수를 0으로 치환한 양수만(시세-가격) 변수와 시세에서 판매

www.kci.go.kr

가격을 뺀 값을 절대값으로 치환한 절대값(시세-가격) 변수로 구성되어 있다. 시세와 가격의 차이를 만들 때는 2022년 1월 11일 기준 세티즌의 휴대폰 중고가격 시세표를 참조하였다. 세티즌은 거의 모든 휴대폰 기종의 중고시세가 정보를 제공한다. 상품상태가 미개봉은 상급, 거의 새것은 평균, 사용감 있음은 하급으로 매칭하여 시세를 산출하였다. 시세표에 없는 시세는 최근 3개의 게시글 판매가격의 평균으로 대치하였다.

카카오톡 유도는 게시글 본문에서 카카오톡으로 채팅을 유도하는지를 뜻한다. 최근 신종 사기 수법 중, 구매자에게 메신저 채팅을 유도하고 사기꾼이 안전결제로 위장한 피싱 사이트에서 결제를 유도함으로써 구매자를 속이는 방식이 있었다[14]. 따라서 카카오톡 유도가 신종 사기 수법의 일환이라고 판단하였고 카카오톡 유도를 변수로 선정하였다. 카카오톡 유도는 게시글 본문에 카카오톡 아이디를 작성하였는지 여부를 바탕으로 만들었다.

게시글 올린 횟수는 2021년 8월 18일부터 2021년 11월 22일까지의 기간에서 게시글을 중복해서 올린 횟수로 산출하였다. 사기 게시글은 반복적으로 글을 올리기 때문에 사기 게시글 판별에 게시글 올린 횟수가 도움될 것이란 가정을 하여 게시글 올린 횟수를 변수로 선정하였다. 제조사는 휴대폰 기종의 제조사를 말한다. 어떤 제조사에서 만든 기종이 사기에 많이 쓰이는지를 확인하기 위해 제조사를 변수로 선정하였다.

거래지역 표시 여부는 판매자가 게시글에 직거래 지역을 표시한 여부를 말한다. 사기의 경우 거래지역을 표시하지 않을 것이라는 가정하에 거래지역 표시 여부를 변수로 선정하였다.

가격은 판매자가 게시글의 가격란에 제시한 상품 가격을 뜻한다. 터무니없이 싼 가격의 물건은 사기일 가능성이 높다[14]. 따라서 가격을 변수로 선정하였다. 가격은 판매자가 올린 가격을 그대로 사용하였으며 가격란에 제시된 가격이 터무니없이 작거나 높을 경우는 본문에 나와 있는 가격을 가격 변수로 변환해주었다.

글자수는 판매자가 본문에 작성한 글자의 수를 뜻한다. 본문의 길이를 띄어쓰기를 포함하여 글자수로 변환하였다. 상품 상태는 판매자가 상품 상태란에 제시한 상품의 상태로 미개봉, 거의 새것, 사용감 있음으로 분류된다.

결제 방법은 판매자가 제시한 결제 방법을 말하며 안전결제, 판매자와 협의로 분류된다. 네이버 페이 안전 결제 서비스를 이용하면 사기를 당할 확률이 낮기에 결제 방법에 따라 사기 게시글 여부가 달라지리라 판단하여 결제 방법을 변수로 선정하였다. 네이버 페이를 통해 에스크로 서비스를 이용할 수 있으면 안전결제로 그 외에는 판매자와 협의로 치환해주었다.

배송방법은 판매자가 상품을 구매자에게 양도하는 방법으로 내용설명란에 적어놓은 것을 말하며 직거래, 택배거래, 온라인전송, 기타로 분류된다. 배송방법은 판매자가 1개 이상의 방법(예: 택배거래, 직거래)을 제시하는 경우가 많기에 원-핫 인코딩을 해주어도 배송방법이 중복되었다. 배송방법이 공란이거나 판매자와 직접 연락하라는 문장이 쓰여 있는 경우는 기타로 처리해주었다.

중고나라에 게시글을 올릴 때 판매자가 자신의 휴대폰 번호를 정규 표현 식으로 올리지 않는 경우가 있다.(예: 010-오오삼삼-717사) 폰 마스킹 여부는 판매자가 휴대폰 번호 올릴 때, 정규 표현 식을 사용하지 않았는지 아닌지를 말한다. 0의 비율은 가격 중 0의 개수를 비율로 나타낸 것으로 사기 게시글은 0의 비율이 높을 것이라는 가정하에 변수로 선정하였다. 판매자의 이메일은 게시글에 공개한 메일주소를 뜻한다. 선행연구에 따르면 메일 자체로는 사

Table 2. Data Conversion

Data before conversion	Data after conversion
Market price(ex. 140,000) - Price(ex. 160,000)	Max(Market price-Price, 0) (ex. 0))
Kakaotalk ID (ex. *****5213)	Kakaotalk ID(Revealed 1, Otherwise 0)
Duplicate number of posts(ex. 2)	Duplicate number of posts(ex. 2)
Mobile phone model(ex. Galaxy Fold 2)	Manufacturer(ex. Samsung)
Transaction area(ex. Pil-dong)	Whether to reveal the transaction area(Revealed 1, Otherwise 0)
Price(ex. 230,000 won)	Price(ex. 230,000)
Number of letters(ex. Galaxy S9 Urgent Sale)	The number of letters(ex. 18)
Product status(ex. Unopened)	Product status(ex. Unopened)
Payment method(ex. Naver pay remittance)	Payment method(ex. Safe payment)
Delivery method(ex. Direct transaction)	Delivery method (ex. 1: Direct transaction, 0: Delivery transaction)
Phone number(ex. 010-삼3오5-7II2)	Whether it's masked or not (ex. Masked 1, Otherwise 0)
Price(ex. 20,000 won)	The ratio of 0(ex. 80%)
Mail(ex. *****62@naver.com)	Whether e-mail address is revealed(ex. Revealed 1, Otherwise 0)
Writing date(ex. 2021-8-19)	Writing day(ex. Monday)
Writing time(ex. 15:29)	Writing hour(ex. 15)
Comments(ex. I buy it for 200,000 won)	Whether there's a comment or not(ex. Yes 1, No 0)

www.kci.go.kr

기 글을 판단하는 데 도움이 되지 않는다고 판단하였으나 메일을 공개한 사람과 공개하지 않은 사람 간 차이가 있을 것으로 판단된다[13]. 따라서 메일 공개 여부를 변수로 치환하였다. 작성된 일자 및 작성 시간은 판매자가 게시글을 작성한 일자와 시간이다.

댓글은 게시글에 달린 짤막한 글을 말한다. 댓글 자체로는 사기 글을 판단하기 어렵고 사기 게시글은 정상 게시글과는 달리 주목할 만한 특징이 있어 댓글이 달릴 확률이 높다고 판단하여 댓글 여부를 변수로 선정하였다.

표 2는 이러한 데이터 변환의 예시를 나타내고 있다.

3. Doc2vec: Sentence Vector Extraction

본 연구에서는 Doc2vec을 사용하여 게시글 본문에 있는 텍스트 데이터의 특징을 추출하였다. 전처리 과정에서는 이모지, 이모티콘, 특수문자 등을 제거해주었으며 한국어에 특화된 형태소 분석기인 KoNLpy 라이브러리의 Komoran을 사용하여 '형태소'만을 추출하였다. 파라미터 설정 시, 최상혁 외[23]의 연구결과를 참조하여 벡터 크기 = 200, 창 크기(windows) = 5, 모델은 PV-DM 모델을 활용하였다. 벡터 크기를 200차원으로 설정하였으므로 총 200차원의 희소행렬이 만들어졌는데, 이를 그대로 모델에 학습시키면 '차원의 저주' 문제를 초래할 수 있다. 이를 해결하기 위해, 주성분분석을 사용하였고 200차원을 7차원으로 축소하였다. 7차원으로 줄인 이유는 200차원의 10%에 해당하는 20차원 이내에서 차원의 수를 바꿔가면서 ROC-AUC score를 측정한 결과, 7차원일 때 ROC-AUC score가 가장 높았기 때문이다.

IV. Model Construction and Evaluation

1. Data Preparation

총 4,670개의 데이터를 7:3의 비율로 훈련 데이터와 테스트 데이터로 나누었다. 훈련 데이터는 사기가 63건, 비사기가 3,206건이다. 언더샘플링 후, 비사기의 건수를 1,488건으로 줄였고 오버샘플링 후, 사기 1,489건-비사기 1,488건으로 5:5의 비율로 맞추었다. 총 2,977건의 훈련 데이터를 학습에 사용하였다.

2. Data Sampling

사기의 특성상 사기의 건수가 사기가 아닌 것의 건수보다 적을 수밖에 없다. 이런 불균형 데이터셋 문제는 모델의 성능을 저하하고 모델의 성능에 반영된다[24]. 따라서 불균형 데이터셋에 대한 처리가 모델학습 전에 선행되어야 한다. 본 연구에서도 전체 데이터 셋 중, 30%의 해당하는 테스트 데이터를 제외한 3,269건의 훈련 데이터 중에서 63건만이 사기 게시글이다. 사기 게시글의 수를 전체 게시글의 수로 나눈 비율이 약 2%로 계층 간 데이터 불균형이 있다. 이를 해결하고자 언더샘플링(Undersampling)과 오버샘플링(Oversampling)을 모두 사용한 복합샘플링(Edited Nearest Neighbours + Adasyn)을 적용하였다. Edited Nearest Neighbours(ENN)와 Adasyn은 이규남 외[25]의 연구결과를 참조했을 때, 10개의 데이터 셋에서 평균 F1_score를 산출한 결과, 각각 0.8이상으로 준수한 성능을 보여준 바 있다.

3. Feature Selection

특징선택은 노이즈와 중복되는 특성을 제거하는 데 가장 인기 있는 차원 축소기법 중 하나이다[26]. 특징선택은 성능을 향상시키고 계산 복잡성을 낮추고 더 나은 일반화된 모델을 만들고 요구되는 저장공간을 줄일 수 있다[26]. 본 연구에서는 특징선택 방법 중 일변량 모델인 Filter model을 사용하였다. 연속형 변수에 대해서는 F-test를, 범주형 변수에 대해서는 Chi-square test를 수행하였다. 변수들의 p-value를 오름차순으로 정렬 후, LightGBM 모델의 ROC_AUC score를 기준으로 차례대로 변수의 수를 늘려가면서 점수를 매기고 최종적으로 62개의 변수 중 26개의 변수를 골라내었다. 최종 선택된 변수는 다음의 표 3과 같다.

4. Analysis Results

LightGBM 모델의 성능을 평가하기 위해 다양한 기계학습 방법론과 비교하였다. 기계학습 방법론은 SVM(Support Vector Machine), LR(Logistic Regression), XGBoost(Extreme Gradient Boosting), RF(Random Forest) 등 총 4가지를 사용하였으며, 평가지표로는 정확도(Accuracy)와 ROC-AUC-score, Recall-score, F1-score를 사용하였다. 불균형 데이터셋에서는 모형을 구축했을 때, 높은 정확도를 얻을 수 있지만, Recall, F1 score와 같은 다른 평가지표에서 좋은 점수를 받지 못할 수 있다[24]. 따라서 본 연구에서는 모델 성능에 대한 더 정확한 판단을 위해 정확도 외에 3가지 지표를 더 살펴보았다. 예측성능 비교결과는 표 4와 같다.

LightGBM의 경우, Recall score와 ROC-AUC score, F1 score에서 다른 기계학습 모델보다 높은 평가점수를 기록하였다. 특히, 같은 Gradient Boosting 알고리즘을 사용하는 XGBoost와 비교해서 F1 score에서 약 6%가량 점수가 높게 나왔음을 알 수 있다.

www.kci.go.kr

Table 3. Variables after Feature Selection

Feature	Description	Feature	Description
The ratio of 0	Percentage of zero in price	Writing time / 4am	Whether the posting time is 4am(yes/no)
Whether to indicate the transaction area	1:indicated, 0:hidden	Writing time / 8am	Whether the posting time is 8am(yes/no)
Duplicate number of posts	Number of duplicate posts during data collection period	Writing time / 12pm	Whether the posting time is 12pm(yes/no)
Payment method / Safe payment	Whether secure payment is available(1:yes, 0:no)	Writing time / 21pm	Whether the posting time is 21pm(yes/no)
Payment method / Direct payment	Whether direct payment is available(1:yes, 0:no)	Writing day / Friday	Whether the date of posting is Friday(yes/no)
The number of letters	The number of letters in the post body	Writing day / Thursday	Whether the date of posting is Thursday(yes/no)
Whether there's a comment or not	1:comment made, 0:not made	Writing day / Monday	Whether the date of posting is Monday(yes/no)
Product status / Almost new	Whether the product is in like-new condition(yes/no)	Writing day / Sunday	Whether the date of posting is Sunday(yes/no)
Product status / Unopened	Whether the condition of the product is new(yes/no)	Writing day / Tuesday	Whether the date of posting is Tuesday(yes/no)
Product status-feeling of use	Whether the condition of the product is used(yes/no)	Manufacturer OO	Whether the cell phone manufacturer is OO(yes/no)
Positive number only(Market price-Price)	Max(0, Market price-Price)	Courier delivery	Whether the item is delivered via courier delivery(yes/no)
Online delivery	Whether the item is delivered online(yes/no)	pca3	The 3rd element of the PCA vector
Writing time 0am	Whether the posting time is 0am(yes/no)	pca4	The 4th element of the PCA vector

Table 4. Performance Comparison among Predictive Models

Predictive Models	Evaluation indicators			
	Accuracy	ROC-AUC	Recall	F1 score
LightGBM	**0.912**	**0.755**	**0.592**	**0.206**
XGBoost	0.881	0.703	0.518	0.144
LR	0.737	0.630	0.518	0.070
SVM	0.737	0.630	0.518	0.070
RF	0.954	0.649	0.333	0.219

Table 5. The result of Doc2vec

LightGBM	Before Doc2vec	After Doc2vec
Accuracy	0.877	**0.912**
ROC-AUC	0.701	**0.755**
Recall	0.518	**0.592**
F1 score	0.140	**0.206**

또한 Doc2vec으로 텍스트 정보를 모델학습에 추가한 것이 모델성능을 높였는지를 확인하기 위해 Doc2vec 적용 전과 후 모델성능의 차이를 비교하였다. 다음의 표 5는 Doc2vec적용 전과 후 모델성능의 차이를 보여준다. 이 결과를 통해 Doc2vec적용 후, 모델의 모든 성능이 향상되었음을 확인할 수 있었다. 따라서 Doc2vec으로 추출한 벡터화된 텍스트 데이터가 변수로 쓰였을 때, 모델의 성능 향상에 도움이 되는 것을 알 수 있다.

본 연구에서는 전체적인 특징의 중요도를 살펴보기 위해, 특징 선택 단계에서 제거한 특징들도 포함하여 모델을 돌린 후 특징 중요도를 살펴보았다. 그림 2는 SHAP의 요약 그래프(summary_plot)로, 섀플리 값이 높은 변수가 순서대로 나타나 있다. 붉은색으로 갈수록 특성이 사기 게시글일 확률에 양(+)의 영향력을 미쳤다는 것을 의미하고 파란색으로 갈수록 특성이 사기 게시글일 확률에 음(-)의 영향력을 미쳤다는 의미이다.

www.kci.go.kr

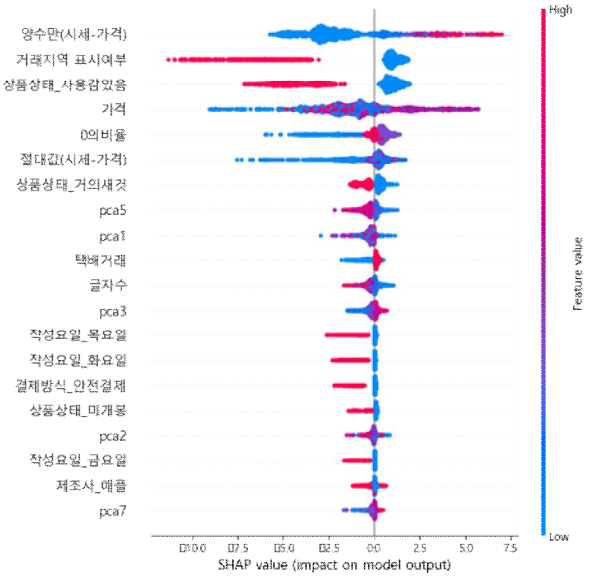

Fig. 2. SHAP Summary Plot

예를 들어, 거래지역 표시 여부가 없으면(0) 사기 게시글일 확률이 커진다는 의미이고 거래지역 표시 여부가 있으면(1) 사기 게시글일 확률이 낮아진다는 의미이다. 이렇게 그림을 해석해보면 시세에서 가격을 뺀 차이의 양수 값이 높을수록 사기 게시글일 확률이 높다는 의미인데 즉, 시세보다 가격이 많이 쌀수록 사기 게시글일 확률이 높다는 의미이다. 거래지역 표시를 안 할수록 사기 게시글일 확률이 높다는 의미이다. 또한, 가격이 높을수록, 안전거래를 안할수록, 택배거래를 할수록, 0의 비율이 높을수록 사기 게시글일 확률이 높았다. 또한, 상품상태가 사용감이 있을수록 사기 게시글이 아닐 확률이 높음을 알 수 있다.

SHAP으로는 특정 특징의 영향력도 간단하게 진단해 볼 수 있다[20]. 그림 3은 시세에서 가격을 뺀 차이의 양수 값과 사기 게시글인지 아닌지를 그래프로 나타낸 결과이다. 시세에서 가격을 뺀 차이가 양수일수록 사기 게시글일 확률이 높았다. 반대로 시세와 가격이 차이가 나지 않을수록 사기 게시글일 확률이 낮았다.

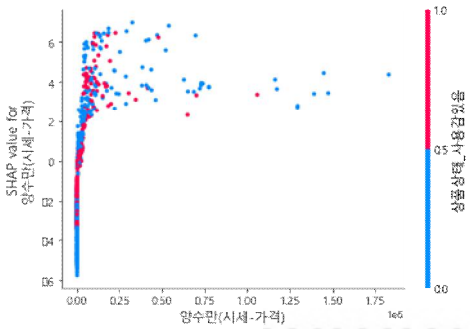

Fig. 3. SHAP Dependence Plot

분석 결과에 따르면, 사기 게시글의 상품 가격은 시세보다 많이 싼 것으로 나타났다. 이는 대부분의 사기 게시글들이 구매자들의 이목을 끌기 위해 상품의 가격을 터무니없이 저렴하게 책정한다는 현장호 외[14]의 연구결과와 일맥상통한다. 또한 거래지역 표시를 안 하는 것이 사기 게시글의 특징이었다. 거래지역은 직거래하는 장소를 말하므로 거래지역을 표시하지 않았다는 것은 직거래를 하지 않겠다는 의미로 해석된다. 사기꾼이 직거래하지 않는 이유를 추측하면, 직거래하게 된다면 구매자가 상품을 직접 눈으로 볼 수 있게끔 해야 하기 때문인데 사기꾼이 원래 없던 상품을 가지고 갈 수 없기 때문이다.

상품상태가 사용감이 있는 상품을 판매하지 않는 것이 사기 게시글의 특징이었다. 이는 달리 말하면, '미개봉 및 새 상품'을 판매하는 것이 사기 게시글의 특징이라고 볼 수 있다. 이런 연구결과는 현장호 외[14]의 연구결과와도 부합한다. 현장호 외[14]에서도 사기 판별 요소 중 하나로 미개봉 및 새 상품 여부를 제시하고 있는데, 중고 물품의 특성상 단순히 가격이 저렴하다는 이유만으로는 구매자의 이목을 끌기에는 부족하기 때문에, 다소 사용감이 있는 물품보다는 새 상품 혹은 미개봉 제품으로 사기 매물을 작성하는 것으로 보인다.

사기 게시글의 경우, 상품 가격이 높은 특징을 보였다. 높은 가격의 제품으로 사기를 치는 이유를 추측하면, 상품의 가격이 높은데도 불구하고 앞서 살펴본 바와 같이 시세보다 가격을 낮게 책정함으로써 구매자의 관심을 끌기 위함이다. 또한, 사기는 한 명의 피해자를 대상으로는 일회성에 그치기 때문에 사기꾼이 한 번에 많은 돈을 획득하기 위하여 가격을 높게 책정하는 것으로 보인다.

가격 중 '0의 비율'이 높은 것도 사기 게시글의 특징이었다. 이런 연구결과는 이동우 외[13]와도 부합된다. 사기꾼이 세밀한 가격을 제시하지 않는 이유를 추측해보면, 사기 게시글은 반복적으로 올려야 하는데 게시글을 올릴 때마다 세밀한 가격을 적기에 어려움이 있기 때문이다.

배송방식을 택배거래로 하는 것도 사기 게시글의 특징 중 하나이다. 사기는 앞서 설명한 바와 같이, 직거래로 이루어지기 어려워서 사기꾼이 배송방식을 택배거래로 해놓은 것으로 보인다. 거래방식을 직거래로 해놓으면 구매자 측에서 직접 만나자고 했을 때, 변명을 생각하기 어려워서 애초에 배송방식을 택배거래로 해놓은 것으로 보인다. 배송방식은 사기꾼이 바꿀 수 없어서 사기를 판별하는 데 있어 중요한 지표로 보인다.

본문의 글자 수가 적은 것도 사기 게시글의 특징이었다. 이는 상품에 대한 정보를 적게 제공하는 것인데, 사기꾼들

www.kci.go.kr

이 매물을 가지고 있지 않기 때문에 상세한 정보를 제공하기 불가능하거나 어렵기 때문으로 보인다. 반면, 작성 요일은 다른 특성들과 비교했을 때, 상대적으로 덜 중요한 특징을 보였다. 나아가 작성 시간대는 특성 중요도에서 모두 20위권 밖으로 중요하지 않은 특징을 보였다.

Ⅴ. Conclusions

1. Academic, practical implications

본 연구의 학술적 시사점은 다음과 같다. 지금까지 C2C 중고거래 환경에서 텍스트 데이터를 활용한 연구가 일부 있었으나 구체적으로 Doc2vec을 사용하여 텍스트 데이터의 특징을 뽑아낸 연구는 없었다. 또한 시세와 가격의 차이, 결제방법, 배송방법, 카카오톡 유도 여부 등 다양한 파생변수들을 만들어서 사기 게시글의 특징을 살펴본 연구도 드물었다. 이렇게 다양한 파생변수와 Doc2vec으로 텍스트 데이터까지 모델에 학습시킨 연구는 기존에 시도되지 않았기에, 이러한 점이 기존의 연구들과 가장 차별화되는 점이라고 볼 수 있다.

실무적 관점으로 볼 때, 본 연구는 향후 C2C 중고거래 플랫폼 운영자들에게 사기 게시글 판별의 초석이 될 수 있다. 즉, C2C 중고거래 플랫폼 운영자들은 본 연구의 결과를 바탕으로 사기 게시글의 유무를 판별해낼 수 있는 서비스를 만들 수 있을 것으로 기대된다. 사기 게시글의 유무를 판별하는 서비스가 확대되면 성장하고 있는 C2C 중고거래 시장에서 사기로 인한 사회적 비용을 줄일 수 있을 것이다.

2. Limitations and future research

본 연구의 한계점은 다음과 같다. 첫째, 분석을 위한 절대적인 데이터의 양이 적었다. 분석을 위해 사기판별이 불가능한 데이터를 어쩔 수 없이 제외해야 했지만 수집한 데이터의 2/3 데이터를 제외해야 했기에 데이터의 손실이 컸다. 데이터양을 늘릴수록 모델은 더욱 좋은 성능을 내게 되므로 향후 데이터를 더욱 많이 모아 사기판별 모델을 만든다면 더욱 우수한 성능이 기대된다.

둘째, 본 연구에서는 게시물의 이미지 데이터도 수집하였으나 수집된 이미지 데이터를 분석에 사용하지 못했다. 현장호 외[14]의 연구에 따르면 사기 매물 판별 요소에는 이미지 도용이 있었다. 따라서 도용된 이미지의 특징을 뽑아 모델에 학습시키면 좋은 성능을 낼 것으로 기대된다. 향후 CNN 등 딥러닝 모델을 사용하여 이미지의 특징을 추출한 뒤, 모델에 학습시켜볼 예정이다.

셋째, 본 연구에서는 휴대폰 및 주변기기 게시글만을 대상으로 했으므로, 다른 물품 카테고리에 해당하는 게시글에 대한 모델 구축을 수행하지 못했다. 즉, 다른 물품 카테고리에 해당하는 사기의 특징을 판별할 수 없었다. 향후에는 다양한 물품 카테고리에서의 사기를 모델에 학습시킨다면 실용성이 한층 강화될 것으로 기대된다.

ACKNOWLEDGMENT

This research was supported by the BK21 FOUR (Fostering Outstanding Universities for Research) funded by the Ministry of Education and National Research Foundation of Korea. This work was also supported by TheCheat(thecheat.co.kr).

REFERENCES

[1] C. S. Wu, F. F. Cheng, and D. C. Yen, “The influence of seller, auctioneer, and bidder factors on trust in online auctions,” Journal of Organizational Computing and Electronic Commerce, Vol. 24, No. 1, Jan. 2014, pp. 36-57. DOI: 10.1080/10919392.2014.866502

[2] M. J. Jin, and J. G. Kim, Motivation Factors Affecting the Use of C2C Secondhand Trading Platforms, Proceedings of the Spring Conference of Korean Institute of Industrial Engineers, pp. 2645-2673, Korea, Jun. 2021.

[3] S. M. Kim, Platform-based second-hand market fraud is active... 120,000 people lost 89.7 billion won last year, Nov. 2021, https://www.donga.com/news/Economy/article/all/20211124/110445133/1

[4] The Cheat, Fraudulent case statistics, https://thecheat.co.kr/rb/?mod=_statistics

[5] Y. J. Kim and Y. R. Koo, A Study on the Service Design for safe C2C Used Trading - Based on the mobile APP, Proceedings of the Winter Conference of Korean Society of Design Science, pp. 173-174, Korea, Nov. 2019.

[6] H. H. Park, “Analysis of Sales Information of Secondhand Clothing Goods on the C2C Secondhand Trading Platform,” Fashion & Textile Research Journal, Vol. 23, No. 3, pp. 358-369, Jun. 2021. DOI: 10.5805/SFTI.2021.23.3.358

[7] A. Dimoka,, Y. Hong, and P. A. Pavlou, “On product uncertainty in online markets: Theory and evidence,” MIS Quarterly, Vol. 36, No. 2, pp. 395-426, Jun. 2012. DOI: 10.2307/41703461

[8] Y. H. Lee, H. M. Kou and H. J. Kim, “Efficient Supervised Credit Card Fraud Detection Technique using Autoencoder,” The Journal

www.kci.go.kr

of Korean Institute of Information Scientists and Engineers, Vol. 25, No. 1, pp. 1-8, Jan. 2019. DOI: 10.5626/KTCP.2019.25.1.1

[9] S. J. Choi, J. W. Lee and O. B. Kwon, "Financial Fraud Detection using Text Mining Analysis against Municipal Cybercriminality," Journal of Intelligence and Information Systems, Vol. 23, No. 3, pp. 119-138, Sep. 2017. DOI: 10.13088/jiis.2017.23.3.119

[10] B. H. Choi and N. W. Cho, "A Study on the Fraud Detection through Sequential Pattern Analysis: Focused on Transactions of Electronic Prepayment," The Journal of Society for e-Business, Vol. 26, No. 3, pp. 21-32, Aug. 2021. DOI: 10.7838/jsebs.2021.26.3.021

[11] F. Anowar and S. Sadaoui, "Detection of Auction Fraud in Commercial Sites," The Journal of Theoretical and Applied Electronic Commerce Research, Vol. 15, No. 1, pp. 81-98, Jan. 2020. DOI: 10.4067/S0718-18762020000100107

[12] W. U. H. Abidi, M. S Daoud and B. Ihnaini, "Real-Time Shill Bidding Fraud Detection Empowered With Fussed Machine Learning," IEEE Access, Vol. 9, pp. 113612-113621, Jun. 2021. DOI:10.1109/ACCESS.2021.3098628

[13] D. W. Lee and J. Y. Min, "A Study on the Fraud Detection in an Online Second-hand Market by Using Topic Modeling and Machine Learning," Information Systems Review, Vol. 23, No. 4, Nov. 2021. DOI:10.14329/isr.2021.23.4.045

[14] J. H. Hyun, D. Y. Lim and C. Y. Lee, "A proposal on necessity of preventing fraud damage in C2C used trading markets: Focusing on fraud red flags," The Journal of Police Science, Vol. 21, No. 1, pp. 249-272, Korea, Mar. 2021.

[15] Y. S. Kim, H. S. Moon, and J. K. Kim, "Self Introduction Essay Classification Using Doc2Vec for Efficient Job Matching," The Journal of Information Technology Services, Vol. 19, No. 1, pp. 103-112, Feb. 2020. DOI:10.9716/KITS.2020.19.1.103

[16] Q. Le, and T. Mikolov, Distributed Representations of Sentences and Documents, Proceedings of the 31st International Conference on Machine Learning, Vol. 32, No. 2, pp. 1188-1196.

[17] D. Wang, Y. Zhang, and Y. Zhao, LightGBM: an effective miRNA classification method in breast cancer patients, Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics, pp. 7-11, Oct. 2017. DOI:10.1145/3155077.3155079

[18] G. Ke et al., LightGBM: A Highly Efficient Gradient Boosting Decision Tree, Advances in Neural Information Processing Systems, Vol. 30, pp. 3149-3157, 2017.

[19] S. M.. Lundberg and S. I. Lee, A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems, Vol. 30, pp. 4768-4777, 2017.

[20] J. H. Ahn. "XAI: Explainable Artificial Intelligence Dissects Artificial Intelligence," WikiBooks, pp.253-258, 2020

[21] J. S. Shim, J. J. Lee, I. T. Jeong, and H. C. Ahn, A Study on Korean Fake news Detection Model Using Word Embedding, Proceedings of the Korean Society of Computer Information Conference, Vol. 28, No. 2, pp. 199-202, Korea, July 2020.

[22] G. W. Kim, Junggonara: Last year's transaction amount exceeded 5 trillion won...43% increase from the previous year, Mar. 2021, https://www.news1.kr/articles/?4245907

[23] S. Choi, J. Seol, and S. G. Lee, On Word Embedding Models and Parameters Optimized for Korean, Proceedings of Annual Conference on Human and Language Technology, pp. 252-256, Oct. 2016.

[24] R. Mohammed, J. Rawashdeh and M. Abdullah, Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results, Proceedings of 2020 11th International Conference on Information and Communication Systems (ICICS), pp. 243-248, 2020. DOI: 10.1109/ICICS49469.2020.239556.

[25] K. N. Lee, J. T. Lim, K. Soo Bok, and J. S Yoo, "Handling Method of Imbalance Data for Machine Learning : Focused on Sampling," The Journal of the Korea Contents Association, Vol. 19, No. 11, pp. 567-577, Nov. 2019. DOI:10.5392/JKCA.2019.19.11.567

[26] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," Data classification: Algorithms and applications, Vol. 37, 2014

Authors

Do Hyun Lim obtained a BS in Industrial Psychology from the Department of Industrial Psychology at Kwangwoon University. He is currently enrolled in the master's program of the Graduate School of Business IT at Kookmin University. His main areas of interest are machine learning, fraud detection, and deep learning.

Hyunchul Ahn received a BS in Industrial Management from KAIST, and a ME and PhD from KAIST Graduate School of Management. He is currently working as a professor of the Graduate School of Business IT at Kookmin University. His main research areas include AI applications in finance and customer relationship management as well as behavioral models related to information system acceptance.

www.kci.go.kr