The Transactions of the Korea Information Processing Society
TKIPS Vol.13, No.5, pp.217~220
eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2024.13.5.217

Evaluating the Efficiency of Models for Predicting Seismic Building Damage

Chae Song Hwa† · Yujin Lim††

ABSTRACT

Predicting earthquake occurrences accurately is challenging, and preparing all buildings with seismic design for such random events is a difficult task. Analyzing building features to predict potential damage and reinforcing vulnerabilities based on this analysis can minimize damages even in buildings without seismic design. Therefore, research analyzing the efficiency of building damage prediction models is essential. In this paper, we compare the accuracy of earthquake damage prediction models using machine learning classification algorithms, including Random Forest, Extreme Gradient Boosting, LightGBM, and CatBoost, utilizing data from buildings damaged during the 2015 Nepal earthquake.

Keywords : Earthquake, Earthquake Damage Prediction, Machine Learning(ml)

지진으로 인한 건물 손상 예측 모델의 효율성 분석

채 송 화† · 임 유 진††

요 약

지진 발생은 정확히 예측하기 어렵고, 이러한 무작위성을 갖는 사건에 대비하여 모든 건물에 내진 설계를 도입하는 것은 현실적으로 어려운 과제이다. 건물의 특징 분석을 통한 건물 손상 예측을 기반으로 건물의 취약점을 보완한다면, 내진 설계를 도입하지 않은 건물에서도 피해를 최소화할 수 있으므로 건물 손상 예측 모델의 효율성을 분석하는 연구가 필요하다. 본 논문에서는 2015년 네팔 대지진으로 인해 손상된 건물 데이터를 활용하여 Random Forest, Extreme Gradient Boosting, LightGBM, CatBoost 기계학습 분류 알고리즘을 사용하여 지진 피해 예측 모델의 정확도를 비교하였다.

키워드 : 지진, 지진 피해 예측, 기계학습

1. 서 론

2015년 4월 25일, 네팔 카트만두 북서쪽에서 규모 7.8의 강력한 지진이 발생했다. 17일 후, 규모 6.8의 여진이 발생하였고 이 두 차례의 지진으로 인해 수십만 명의 사람들이 집을 잃었을 뿐만 아니라 많은 문화재 또한 상당수 파괴되어 상당한 재산 피해가 발생했다. 또한, 최소 8천여 명의 사망자와 22,000여 명의 부상자가 발생하는 등, 인명 피해 또한 극심했다. 대규모의 지진이었기 때문에 피해가 극심한 것은 불가피한 결과였지만, 피해 규모의 근본적인 이유로는 네팔 건축물의 대부분이 내진 설계를 고려하지 않은 흙벽돌, 짚 지붕과 같은 취약한 재료로 지어졌다는 점을 지적할 수 있다[1].

내진 설계를 도입하여 건물을 재건축하거나 기존 구조물을 적절히 보강하는 것으로 지진 피해를 줄일 수 있지만, 이를 모든 건물에 적용하는 것은 현실적으로 어려운 일이다. 따라서 지진으로 인한 건물 피해를 예측하는 모델은 지진 발생 전에 건물의 취약점을 파악하여 보강 작업을 수행함으로써 잠재적인 인명 피해를 사전에 방지하고 건물의 내구성을 높이는 데 도움이 될 것이다. 또한, 2015년 네팔 지진의 경우 피해 평가 조사가 9개월이나 걸렸다는 사실을 고려할 때, 지진 발생 시 인적 자원과 물적 자원을 투입하지 않아도 현장의 피해 정도를 신속하게 파악할 수 있는 피해 예측 모델은 긴급 상황에서 빠른 대응을 가능하게 하고 복구 작업의 우선순위를 정하는 데 도움이 된다. 따라서 지진으로 인한 건물 피해를 예측하는 모델의 성능을 평가하는 것은 제한된 자원을 효과적으로 활용하여 지진으로 인한 피해를 최소화하고 빠른 복구에 도움을 줄 수 있을 것이라 예상된다.

※ 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 ICT혁신인재 4.0 사업의 연구결과로 수행되었음(IITP-2024-RS-2022-00156299).
※ 이 논문은 2023년 ACK 2023의 우수논문으로 “지진 데이터를 이용한 건물 피해 예측 모델의 성능 분석”의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 숙명여자대학교 인공지능공학부
†† 종신회원 : 숙명여자대학교 인공지능공학부 교수
Manuscript Received : December 28, 2023
First Revision : April 1, 2024
Accepted : April 19, 2024
* Corresponding Author : Yujin Lim(yujin91@sookmyung.ac.kr)

2. 데이터 수집 및 전처리

2.1 데이터 수집

본 논문에서는 2015년 네팔 대지진으로 인해 피해를 입은 11개 지역의 건물 특성 정보 및 해당 건물들의 지진 피해 정도를 나타내는 데이터 세트를 활용하여 지진으로 인한 건물 피해 정도를 예측하기에 적절한 모델을 찾고자 하였다. 해당 데이터 세트는 2015년 네팔 지진 이후, 네팔의 국가계획위원회와 중앙 통계국이 주택 재건을 위해 실시한 피해 평가 조사를 통해 수집된 데이터로, 건물의 바닥, 토대, 지붕, 나이 등 총 762,106개 건물의 특성 정보와 지진으로 인한 피해 정도를 나타내는 Damage Grade로 구성되어 있다[2]. 피해 정도는 Grade 1부터 5까지 다섯 단계로 구분되며, 숫자가 증가함에 따라 피해 정도가 증가함을 의미한다.

2.2 데이터 전처리

데이터를 학습에 활용하기 전, 데이터를 무작위로 섞은 후 전체 데이터 셋 중 20%를 테스트 세트로 지정하였다. 학습에 사용할 80%의 데이터는 모델 학습에 적합한 형태를 갖추도록 전처리를 실행하였다.

실제 데이터에는 도메인을 더 정확하게 표현하기 위해 다양하고 많은 특징이 도입되는데, 이는 목표와 무관하거나 중복된 특징을 발생시킬 수 있다. 이러한 특징들은 직접적인 연관성이 없더라도 학습 과정에 영향을 미칠 수 있다[3]. 따라서, 학습 알고리즘의 실행 시간을 줄이고 성능을 향상시키기 위해 학습 전에 상관계수가 0.7 이상인 높은 상관성을 가진 특징들을 시각화하고 제거하였다(Fig. 1).

모델에 입력된 최종 특징은 지진 발생 이전 건물 층수, 건물 연식, 건물 주추 영역, 지표면 상태, 건물 토대 유형, 건물 지붕 유형, 건물 바닥 유형, 건물 소재지, 평면 구조, 건물 상부 구조 유형, 지진 발생 이후 건물 상태로 구성되었다.

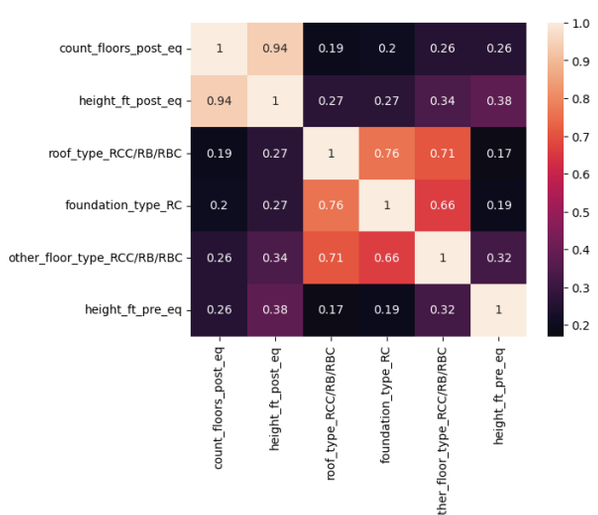

Fig. 1. Correlation Coefficient Between Features

Table 1. Removed Features

Features	Definition
count_floors_post_eq	Count of floors in the building after the earthquake
height_ft_post_eq	Height of the building after the earthquake
roof_type_RCC/RB/RBC	Type of building roof with reinforced materials
foundation_type_RC	Type of building foundation with reinforced concrete
other_floor_type RCC/RB/RBC	Type of building floor with reinforced materials
heigth_ft_pre_eq	Height of the building before the earthquake

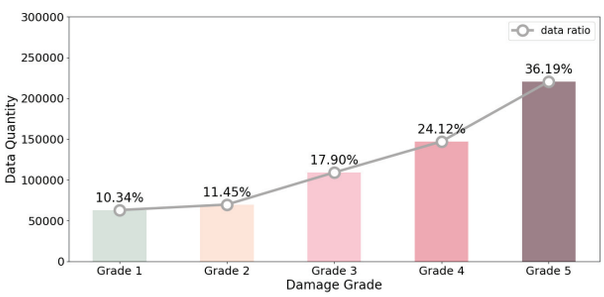

Fig. 2. Class Ratios Before SMOTE Processing

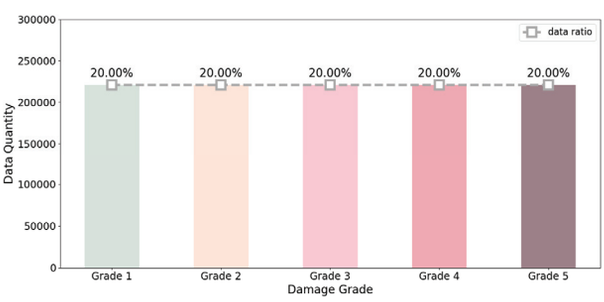

Fig. 3. Class Ratios After SMOTE Processing

사용 데이터의 종속 변수인 Damage Grade는 Fig. 2와 같이 총 다섯 단계로 구분되어 있다. 하지만 Grade 4와 Grade 5의 비율이 약 60%로 불균형한 데이터의 양상을 띠고 있다. 불균형한 데이터는 분류기 성능을 저하하고, 다수 클래스보다 소수 클래스의 오 분류를 발생시킬 수 있다.

따라서 데이터 불균형 문제를 해결하기 위해 합성 데이터를 오버 샘플링하는 기법인 SMOTE를 데이터에 적용하여 그림 3과 같이 각 Grade 간의 비율을 균형 있게 조정하였다[4]. 성능 평가의 신뢰도를 위해 SMOTE를 통해 데이터를 오버 샘플링 하는 기법은 학습 데이터에만 적용되었다.

건물의 특징을 나타내는 데이터인 만큼, 주로 범주형 변수로 이루어져 있다. 이 범주형 변수들은 값 간에 관련이 상관관계가 없고 명확하게 구분되어 있다는 특성이 있으므로 학습에 활용하기 위해 원 핫 인코딩을 통해 수치형 또는 이진 변수로

변환하였다. 숫자형 변수에 대해서는 데이터의 균형을 위해 표준화와 정규화를 진행하였으며, 낮은 분산 값을 가진 변수를 제거하였다.

3. 모델 학습 및 성능 평가

3.1 모델의 특성 및 설계

본 논문에서는 Random Forest(RF), Extreme Gradient Boosting(XGB), LightGBM(LGBM), CatBoost, 총 네 가지의 기계학습 분류 알고리즘을 사용하여 건물 지진 피해 예측을 수행하였다.

RF는 이진, 숫자, 범주형 특징을 쉽게 처리하며 단일 트리에 비해 높은 성능을 제공하고 낮은 오류율, 뛰어난 정밀도 및 효율성과 함께 탁월한 노이즈 저항성을 제공하기 때문에 예측 모델에 적합하다는 장점이 있다[5].

XGB는 기존의 Gradient Tree Boosting 알고리즘의 약점인 과적합을 방지하기 위해 추가적인 파라미터를 도입한 지도 학습 알고리즘으로, 수렴 속도와 일반화 능력을 만족시키는 것으로 입증되었다[6].

LightGBM은 효율적인 병렬 훈련을 통해 빠른 훈련 속도를 제공하는 알고리즘으로, 많은 양의 데이터를 효율적으로 처리하면서도 낮은 메모리를 소비하는 특징을 가지고 있어 다양한 환경에서 신속하게 모델을 학습할 수 있다는 장점이 있다[7].

CatBoost는 효율적이고 합리적인 처리를 위해 클래스 변수를 제공함과 동시에 높은 정확도를 유지한다. 또, 데이터의 특정 부분에 대한 지나친 학습을 방지하여 과적합 발생을 줄일 수 있는 알고리즘이다[8].

이와같은 네 가지 모델은 모두 의사결정 트리를 기반으로 한 앙상블 모델이고, 앙상블 모델은 다수의 약한 분류기를 결합하여 과적합을 효과적으로 관리하는 특성이 있어, 단일 의사결정 트리보다 일반화 능력이 뛰어나다. RF는 배깅을, XGB와 LGBM은 부스팅을, CatBoost는 부스팅의 변형을 사용하고 있으므로 세 가지의 성능 평가지표를 통해 앙상블 모델 내에서도 각 기법의 차이에 따른 성능을 비교하고자 하였다. 각 기계학습 모델에서 사용된 하이퍼파라미터 값은 다음과 같다.

Table 2. Hyperparameters used in Models

	Estimators	Max depth	Learning rate
RF	100		
XGB	100	3	0.1
LGMB	100		0.1
CatBoost	100	6	0.03

3.2 실험 결과

1) 정확도

본 논문에서 네 기계학습 기법을 SMOTE를 적용하지 않은 데이터와 SMOTE를 적용하여 Grade 간의 불균형을 해결한 데이터를 사용한 결과, 성능상의 큰 차이가 없음을 확인하였다.

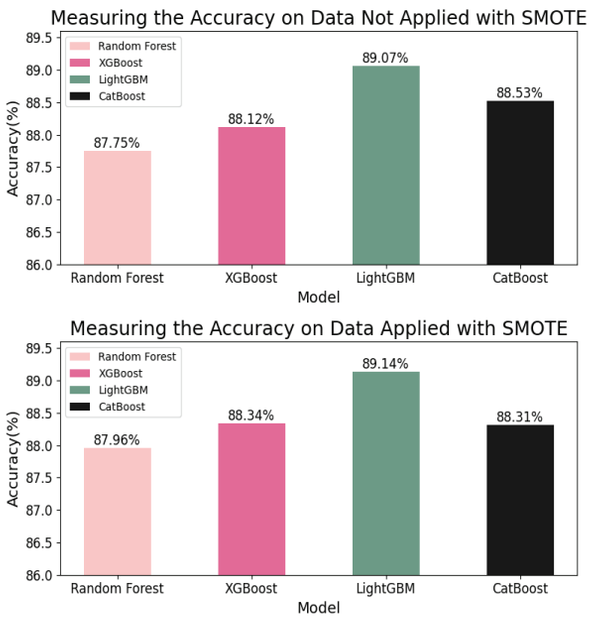

Fig. 4. Comparison of Accuracy

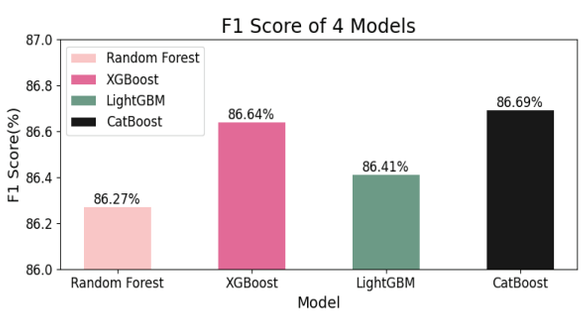

Fig. 5. F1 Score of 4 Models

2) F1 Score

모델의 전반적인 분류 성능을 평가하기 위해 혼동 행렬을 기반으로 각 클래스의 F1 Score의 평균값인 macro average를 도출했을 때, 네 개의 모델 간의 성능은 유사함을 확인할 수 있다.

3) AUC

AUC는 ROC 곡선 아래의 면적으로, 이진 및 다중 클래스 분류 모델의 성능을 평가하는 데 널리 사용된다[9]. 해당 연구는 피해 정도를 다섯 단계로 나누는 다중 클래스 분류 모델이기 때문에 AUC를 성능 평가지표로 활용하였다.

Fig. 6에서 보는 바와 같이 다중 클래스의 ROC 커브를 이용한 AUC 또한 정확도, F1 Score와 비슷하게 유사한 추세의 차이를 보이나 네 개의 모델 성능이 유사함을 확인할 수 있다. 네 개의 모델 하이퍼파라미터에서 생성되는 결정 트리의 개수가 같고, 그 외의 하이퍼파라미터의 설정이 유사하여 네 개의 모델의 성능이 비슷한 결과를 보이는 것으로 추측된다.

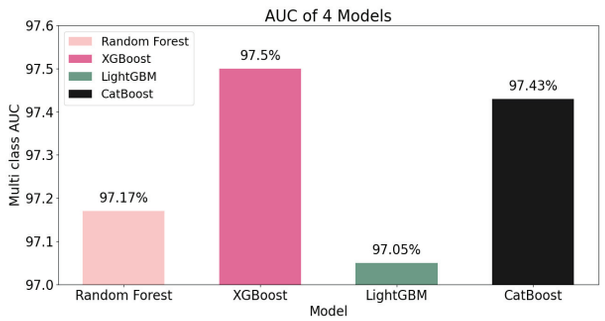

Fig. 6. Comparison of AUC Among Models

4. 결 론

본 논문에서는 지진 데이터를 통해 RF, XGB, LGBM, CatBoost와 같은 기계학습 분류 알고리즘을 활용하여 건물 지진 피해를 예측하는 모델의 성능을 평가하였다. 향후 지진으로 인한 인명 및 자산 피해를 최소화하기 위해 더욱 다양한 모델을 활용한 추가적인 연구가 필요하다. 본 연구에서는 기계학습만을 사용하였고, 데이터는 건물의 구성 요소 특성만을 고려했으나, 지진은 시간에 따라 변화하고 지역에 따른 피해 정도가 다르다는 특징을 가지고 있다. 이에 향후 시계열 데이터와 공간 데이터를 함께 고려하여 기계학습뿐만 아니라 딥러닝 모델에도 적용한다면, 지진의 동적이고 복잡한 패턴을 더 잘 이해할 수 있어 더욱 실용적인 연구가 될 것이라 예상된다. 따라서, 향후 다양한 데이터를 활용한 추가적인 연구를 진행할 예정이다.

References

[1] D. Gautam, and H. Chaulagain, "Structural performance and associated lessons to be learned from world earthquakes in Nepal after 25 April 2015 (MW 7.8) Gorkha earthquake," *Engineering Failure Analysis*, Vol.68, 2016.

[2] 2015 Nepal Earthquake: Open Data Portal [Internet], http://eq2015.klldev.org/#/

[3] C. C. Aggarwal, "Data classification: Algorithms and applications," Yorktown Heights, NY, United States, IBM T. J. Watson Research Center, 2018.

[4] Gede Angga Pradipta, "SMOTE for Handling Imbalanced Data Problem: A Review," *2021 Sixth International Conference on Informatics and Computing (ICIC)*, 03 November 2021.

[5] P. M. Chanal and M. S. Kakkasageri, "Random forest algorithm based device authentication in IoT," *IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 14-16 Jul. 2023.

[6] W. Feng, X. Gao, G. Dauphin, and Y.Quan, "Rotation XGBoost based method for hyperspectral image classification with limited training samples," *IEEE International Conference on Image Processing (ICIP)*, 08 Oct. 2023.

[7] J. Huang and W. Chen, "A study on category classification based on LightGBM for signal feature extraction and k-means clustering," *IEEE 5th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, 14 Jul. 2023.

[8] N. Acharya, A. K. Singh, A. K. Dwivedi, and P. Kannadaguli, "Indian food segmentation and calorie estimation using catboost and masked convolutional neural networks," *International Conference on Network, Multimedia and Information Technology (NMITCON)*, 01 Sep. 2023.

[9] B. V. Calster, V. V. Belle, G. Condous, T. Bourne, D. Timmerman, and S. V. Huffel, "Multi-class AUC metrics and weighted alternatives," *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Jun. 2008.

채 송 화

https://orcid.org/0009-0001-6921-8994
e-mail : watermelon97@smwu.ac.kr
2019년 ~ 현 재 숙명여자대학교 인공지능공학부
관심분야 : Machine Learning

임 유 진

https://orcid.org/0000-0002-3076-8040
e-mail : yujin91@sookmyung.ac.kr
2000년 숙명여자대학교 전산학과(박사)
2013년 일본 Tohoku University, Department of Information Sciences(박사)
2022년 ~ 2015년 수원대학교 정보미디어학과 부교수
2016년 ~ 현 재 숙명여자대학교 인공지능공학부 교수
관심분야 : 강화학습, IoT, Edge Computing