

Catboost를 활용한 건설사고 데이터 기반 위험 공정 사전관리를 위한 기초연구

한재영*, 김기남*, 이민재*

*충남대학교 토목공학과

jyh10240@naver.com

Risk Factor Management Based on Construction Accident Data Using Catboost

Jae-Yeong Han*, Ki-Nam Kim*, Min-Jae Lee*

*Dept. of Civil Engineering, Chungnam National University

요 약

최근 건설현장에서는 중대재해 예방과 고위험 작업에 대한 사전관리를 위해 데이터 기반의 예측 기술 도입이 활발히 이루어지고 있다. 본 연구는 건설공사 안전관리 종합정보망(CSI)에서 제공하는 위험요소 프로파일 데이터를 기반으로 위험성 허용가능 여부에 따라서 '중점관리 대상' 여부를 사전에 분류할 수 있는 모델을 개발하는 데 목적이 있다. 분석에는 범주형 변수 처리에 강점을 가진 CatBoost 알고리즘을 적용하였으며, 가능성과 심각성을 기반으로 위험성 기반의 위험 허용가능 여부와 총 8개의 독립변수를 활용하여 모델을 구축하였다.

1. 서론

최근 건설업에서는 건설사고 예방을 위한 대응책으로 다양한 스마트 안전기술과 데이터 기반의 위험관리 체계가 개발되고 있다. 특히 2022년 중대재해처벌법 시행 이후 대형 건설사를 중심으로 스마트 안전장비, AI CCTV등의 현장 도입 및 연구가 지속적으로 이루어지고 있으며, 국토안전관리원 및 고용노동부에서도 스마트 건설안전 기술을 적극 지원하고 있다. 이에 따라 AI 기반 예측 시스템이나 사고 방지 및 예방 시스템에 대한 데이터 분석의 관심도 높아지고 있다.

국토안전관리원에서 제공하는 위험요소 프로파일은 건설현장의 사고사례를 작업공정별, 구조물별로 정리하여 사고 예방 및 위험관리 강화를 위해 구축한 데이터베이스이다. 데이터베이스는 시설물 종류, 공종, 위험발생 객체, 위험발생 위치, 작업 프로세스, 물적 피해, 인적 피해, 사고 원인, 위험성, 저감대책 으로 구성되어 있다. 이러한 위험요소에 대한 사전예방 중심의 관리로의 전환이 요구되고 있다.

이에 본 연구에서는 CatBoost 알고리즘을 활용하여, 건설현장의 위험요소 프로파일 데이터를 기반으로 '중점관리 대상'과 '일상관리 대상' 여부를 사전에 예측할 수 있는 분류 모델을 제안한다. 또한, 기계학습을 통한 작업 단위의 위험도를 사전에 분류하고, 이를 기반으로 효율적인 안전관리 계획 지원 체계를 마련하는 것을 목표로 한다.

2. 문헌고찰

건설현장에서의 사고를 예방하고, 사고 발생 가능성을 사전에 예측하기 위한 다양한 연구가 활발히 진행되고 있다. 최근에는 다양한 AI 기술을 활용하여 스마트 안전기술과 빅데이터에 관한 연구가 활발히 진행되고 있다.

전정호 외(2024)는 Auto ML 기반의 이진 및 다중 분류 모델을 구축하여 건설사고의 발생 여부와 유형을 예측하고자 하였으며, 머신러닝 자동화 기법을 활용하여 건설 안전 데이터 분석의 효율성과 예측력을 향상하고자 하였다.

김태훈 외(2023)는 건설재해 빅데이터를 분석하여 건축공사 공정별 위험도를 평가하는 시스템을 개발하였다. 특히 공정별 사고 발생률과 위험도를 정량화하고 시각화함으로써, 현장 작업계획에서 활용 가능한 정량적 기준을 제시하였다. 이러한 연구는 작업 공정 단위에서의 위험 요소를 체계적으로 정리하고 사전 예방 체계를 설계하는 데 기초자료로 활용될 수 있다.

전민중(2021)은 교통 흐름 예측에 CatBoost 알고리즘을 적용하여 다른 머신러닝 모델들 보다 성능이 높음을 확인하였다. 해당 연구는 범주형 변수가 많은 복잡한 데이터에서 학습이 가능하다는 CatBoost의 장점을 확인하였다.

김요한 외(2021)는 건설산업 내 안전 및 재해관리에 인공지능 기술을 적용한 다양한 최신 사례를 제시하였다.

AI 인공지능 기술을 활용한 건설안전 데이터 분석 사례는 다수

존재하지만, CatBoost 알고리즘을 적용한 연구는 다른 Auto ML 분석에 비해 사례가 충분하지 않는 상황이다. 본 연구는 CatBoost 알고리즘을 활용하여 위험요소 프로파일을 활용한 이진 분류 모델을 개발하고, 중점관리 대상 여부를 사전에 예측함으로써 안전관리계획 수립 체계에 기여하고자 한다.

3. 연구 방법

본 연구에서는 건설 현장에서 작업 단위의 위험도를 사전에 분류하여 효율적인 안전관리 계획 지원하는 것을 목표로 한다. 본 연구에서는 건설공사 안전관리 종합정보망(CSI: Construction Safety Information)에서 제공하는 ‘위험요소 프로파일’ 데이터를 연구에 활용하였다. 해당 데이터는 각 공사 현장에서 발생 가능한 위험요소들을 작업유형, 발생 위치, 사고 결과, 심각도 등의 기준으로 정리한 표준화된 프로파일로, 건설사고 사례 총 33,127건의 사례와 18개 항목으로 구성되어 있다.

범주형 데이터 분석의 특화된 알고리즘인 Catboost 알고리즘을 사용했으며, 입력 변수로는 범주형 변수 시설물분류(중), 세부공종(중), 위험 발생 객체(대), 위험 발생 객체, 위험 발생 위치(대), 위험 발생 위치(중), 위험 발생 위치(소), 작업 프로세스 총 8개 범주형 변수로 구성하였다. 또한, 국토안전관리원의 위험성 결정 기준과 위험성 허용기준을 참고하여 사고 빈도와 사고 심각성을 기반으로 위험성 수준을 High/Medium/Low로 분류하였고, 이를 통해 High/Medium은 중점관리 대상, 일상관리 대상으로 재분류하고, 위험도 허용기준 이진 분류 모델을 구축하고 데이터를 분석하였다.

3.1 데이터 전처리 및 변수 정의

위험요소 프로파일 데이터 총 34,052건의 데이터 중에서 사고 빈도, 사고 심각성 등 주요 변수의 결측값을 포함하는 경우를 제거하여, 최종적으로 33,127건의 데이터를 사용하였다. 또한 사고 빈도와 사고 심각성을 곱하여 'High', 'Medium', 'Low'로 구분하였다. 이를 국토안전관리원의 위험성 허용 기준에 근거하였다. [표 1]의 행은 사고 심각성, 열은 사고 가능성으로 구성되어있으며, 'High' 및 'Medium'은 ‘중점관리 대상’, 'Low'은 ‘일상관리 대상’으로 분류하였다. 분석에 사용된 독립 변수로는 시설물분류_중, 세부공종_중, 위험발생객체_대, 위험발생객체, 위험발생위치(대), 위험발생위치(중), 위험발생위치(소), 작업프로세스 총 8개이다.

3.2 이진 분류 모델 구축 및 평가

CatBoost 모델의 하이퍼파라미터는 Optuna 프레임워크를 통해 최적화하였다. Optuna는 Bayesian Optimization 기반의 탐색 기법을 적용하여, 지정된 성능 지표를 최대화하는 방향으로 반복적으로 하이퍼파라미터를 조정한다. 본 연구에서는 중점관리 대

상을 일상관리 대상으로 잘못 예측할 경우를 중시하여, 중점관리 대상의 재현율과 F2score를 모델 성능 평가 지표로 설정하였다.

이후 모델을 학습한 결과, F2-score는 0.9069로 나타났으며, 특히 중점관리 대상에 대한 재현율이 0.90로 측정되었다. 이는 본 모델이 위험한 작업을 효과적으로 사전 식별할 수 있음을 의미한다.

모델 학습은 전체 데이터를 학습, 검증, 테스트 60:20:20으로 분할하여 과적합을 방지하였다. 주요 탐색 대상 파라미터는 [표 2]와 같다. 학습 과정에서는 Class Imbalance 문제를 고려하여 Class_Weight를 적용하고, 마지막으로 60:20:20 비율의 학습, 검증, 테스트 데이터 분할을 통해 검증 성능을 평가하였다.

3.3 모델 성능 평가

테스트 데이터에 대한 성능 평가는 혼동행렬을 기준으로 정확도(Accuracy), 정밀도(Precision), 재현율(Recall) 그리고 F2-score를 주요 지표로 사용하였다. 정확도는 전체 예측 중에서 올바르게 예측된 비율을 말하며, 중점관리 대상과 일상관리 대상의 클래스 불균형으로 인하여 정확도는 실질적인 성능 평가로 사용하기에는 어렵다고 판단하였다. 정밀도는 중점관리 대상으로 예측한 것 중에서 실제 중점관리 대상의 비율을 의미하며, 재현율은 실제 중점관리 대상 중에서 모델이 중점관리 대상으로 올바르게 예측한 비율로 사고예방과 관련된 상황에서 False Negative를 줄이는 것이 매우 중요하다.

본 연구에서는 FN(False Negative)의 위험의 가중치를 두어 F-beta score, $\beta=2$, F2-score를 평가 지표로 활용하였다.

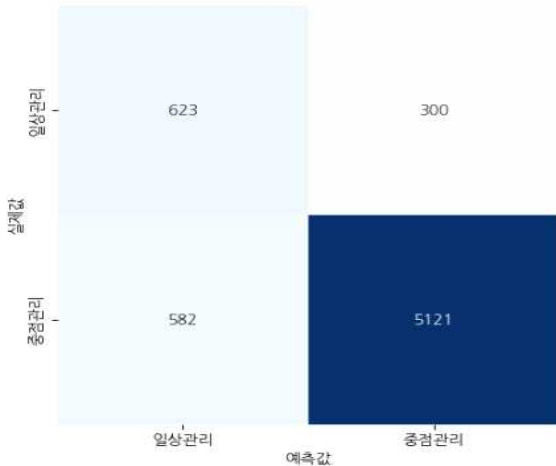
[표 1] 분류 모델 성능 평가

	정밀도	재현율	정확도	F2-score
일상관리	0.52	0.67	0.87	0.9069
중점관리	0.94	0.90		

CatBoost 모델은 Optuna를 통해 하이퍼파라미터를 최적화한 후 테스트 데이터를 기반으로 혼동행렬과 특성 중요도를 통하여 이진 분류의 성능과 주요 변수의 기여도에 대하여 분석했다.

테스트 데이터에 대한 Catboost 이진 분류 모델의 혼동 행렬은 다음과 같다. True Positive 5,121건, True Negative 623건, False Positive 300건, False Negative 582건으로 [그림 1]의 결과를 기록하였다.

[그림 1] Catboost 이진 분류 혼동행렬



CatBoost의 내장된 기능을 활용하여 도출된 변수 중요도 분석 위험발생위치_소, 작업프로세스, 세부공종_중, 위험발생위치_대, 시설물분류_중의 변수가 상대적으로 높은 기여도를 보였다. 이는 모델이 중점관리 대상 여부를 판단할 때, 시설물분류보다도 작업의 공간적 위치나 작업 프로세스에 민감하게 반응하고 있음을 보여준다.

특히 가장 중요한 변수로 나타난 '위험발생위치_소'는 작업의 위치를 상부, 하부, 내부, 고소 등 세부적인 공간에서의 위험성 차이를 정밀하게 반영할 수 있는 요소로서, 향후 사고예방 시스템 설계에 중요한 기준으로 활용될 수 있다.

이러한 결과는 안전관리계획에 있어서 고위험 조건에 대한 선별적 집중관리, 공정별 안전계획 수립, 사고예방 단계에서 실질적 기여 할 수 있는 결과를 나타낸다.

부를 사전에 분류할 수 있는 체계를 제안하였다. 문자형으로 구성된 데이터에 대해 데이터 전처리 없이도 높은 성능을 확보할 수 있는 CatBoost의 특성을 활용하였다.

모델은 테스트 데이터에서 F2-score 기준 0.9069의 성능을 나타냈으며, 특히 중점관리 대상 재현율 0.90으로 실제 중점관리 대상을 상대적으로 높은 성능을 보였다.

또한, 변수 중요도 분석을 통해 '위험발생위치_소', '작업프로세스', '세부공종_중' 정보가 중점관리 대상과 일상관리를 분류하는 주요한 영향을 미친다는 점을 확인할 수 있었다. 사고 발생 대책에 관하여 세부 작업 구역별로 차별화하여 위험평가와 관리 대책 등을 세워야 할 필요가 있다. 또한 단순한 공정 구분만으로 사고 위험성을 판별하기보다 공정별 세부 작업 절차에 관하여 위험평가를 실시할 필요가 있음을 보여준다.

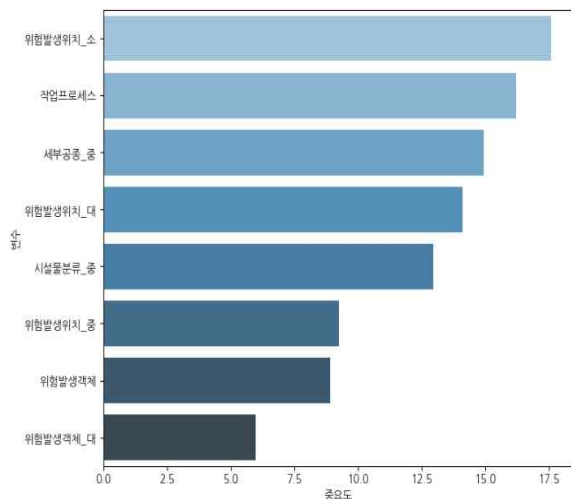
향후 연구에서는 다양한 분류 모델들과 성능들을 비교하고, 이진 분류 및 위험성 결정에 관한 요인 분석 등의 연구 분석을 수행할 예정이다.

또한, 공정별 위험에 따른 저감 대책 및 사고원인을 대응시켜 시각화 모듈을 구성하여 현장에서의 위험관리 체계에 실질적인 기여를 할 수 있을 것으로 기대된다.

감사의 글

본 연구는 2025년 국토교통부 스마트건설기술개발사업의 연구비 지원(RS-2020-KA156208)에 의해 수행되었으며 본 연구의 지원에 감사드립니다.

[그림 2] Catboost 분류 모델 변수 중요도



4. 결론

본 연구에서는 건설공사 위험요소 데이터를 기반으로 CatBoost 이진 분류 모델을 구축하여, 작업 단위의 중점관리 여

참고문헌

- [1] 전정호, 최수연, 윤성배, 용선진, 허영기 “AutoML 기반의 이진, 다중 분류 모델 구축을 통한 건설 사고 발생 및 유형 예측”, 대한건축학회논문집, 제40권 9호, pp. 247-254, 9월, 2024년
- [2] 김태훈, 김기남, 이민재, “건설재해 빅데이터 분석을 통한 건축공사 공정별 위험도 평가 시스템 개발”, 대한건축학회논문집, 제39권 1호, pp. 331-336, 1월, 2023년
- [3] 전민중 “Catboost 알고리즘을 통한 교통흐름 예측에 관한 연구”, 한국산학기술학회논문지, 제22권 3호, pp. 58-64, 2021년.
- [4] 김형관, “건설산업의 안전 및 재해 관련 인공지능,” 건설관리 : 한국건설관리학회 학회지, 제22권 4호, 2021년, pp. 4-8, 2021년