pISSN 1598-2033
eISSN 2233-5706

J. Korea Inst. Build. Constr. Vol. 21, No. 2 : 165-174 / Apr, 2021
https://doi.org/10.5345/JKIBC.2021.21.2.165

www.jkibc.org

건설 현장에서 발생한 업무상 재해가 근로손실일수 심각도에 미치는 특징 중요도 분석

Analysis of the Feature Importance of Occupational Accidents Occurring at Construction Sites on the Severity of Lost Workdays

강 경 수[1] 최 재 현[2] 류 한 국[3*]

Kang, Kyung-Su[1] Choi, Jae-Hyun[2] Ryu, Han-Guk[3*]

Researcher, Construction Engineering and Management Institute, Sahmyook University, Nowon-Gu, Seoul, 01795, Korea [1]

Professor, School of Architecturral Engineering, Korea University of Technology and Education (Koreatech), Dongnam-gu, Cheonan-si, Chungcheongnam-do, 31253, Korea [2]

Professor, Department of Architectural, Sahmyook University, Nowon-Gu, Seoul, 01795, Korea [3]

Abstract

The construction industry causes the most accidents and fatalities among all industries. Although many efforts have been made to reduce safety accidents in construction, the study on the lost workdays that return to work place is insufficient. Therefore, this study proposes a model that classifies the lost workdays lost into moderate and severity, and derives the importance of variable and analyzes important factors through the trained random forest model. We analyze the learning process of the random forest which is a black box model, and extracted important variables that impact on the severity of the lost workdays through the extracted feature importance. The factors existing inside were analyzed through the extracted variables. The purpose of this study is to analyze the accident case data at the construction site through a random forest model and to review variables that have a high impact on the lost workdays. In the future, this sutdy can apply to improve construction safety management and reduce the accident of industrial accidents.

Keywords : construction safety, machine learning, random forest, feature importance

1. 서 론

고용노동부의 2019년 산업재해현황분석 보고서에 따르면, 대한민국 전체 산업 분야에서 발생한 산업 재해 중 건설업이 차지하는 비중이 24.91%이며, 사망자는 517명으로 가장 많다[1]. 건설 현장에서 산업 재해 발생 빈도와 사망자 수가 높은 원인으로 건설 현장이 지닌 특징을 세 가지로 정리할 수 있다. 첫째, 건설 현장은 외부와 차단된 통제 가능한 실내 영역에서 작업하는 제조업과 달리 대부분 실외에서 작업하므로 온도, 습도, 바람 등의 날씨 조건과 같은 외부 위험요인에 추가로 노출되어 있다[2]. 또한, 건설공사는 착공부터 타워크레인과 같은 가설시설물로 공사를 수행하므로 영구적이지 않은 가변 상황에 노출되어 있고, 중장비, 자재, 작업자들이 현장에 혼재한다[3,4]. 이러한 혼잡하고 유동적인 건설 현장의 특징으로 소수의 안전 관리자가 작업장과 작업자들을 체계적으로 관리하기 매우 어렵게 만든다. 둘째, 흐름공정(flow shop) 또는 개별공정(job shop) 일정계획으로 고정된 장소에서 반복 작

Received : March 16, 2021
Revision received : April 7, 2021
Accepted : April 9, 2021
* Corresponding author : Ryu, Han-Guk
[Tel: 82-2-3399-1853, E-mail: ryuhanguk333@gmail.com]

업을 수행하는 작업장은 산업 재해를 통제하고 개선할 수 있는 반면에 건설 현장은 프로젝트 일정계획으로 단일성과 유일성으로 인하여 새로운 프로젝트를 수행할 때마다 새로운 안전 위험요인을 제거해나가면서 작업해야 하므로 안전관리가 매우 어려운 실정이다. 셋째, 건설 프로젝트는 대형화, 고층화, 비정형화되어감에 따라 고소 작업이 많아지면서 떨어짐과 같은 더욱 어렵고 위험한 작업으로 구성되어 있다. 따라서 타 산업과 비교해 건설업의 안전 재해는 중대 재해의 형태로 발생할 가능성이 매우 크다[5-7].

건설 현장에서 발생한 산업 재해의 결과로 근로자가 치명적인 육체적·정신적 손상을 입게 될 경우, 회복 시간으로 인해 발생한 근로손실일수에 따라 가족의 생계 및 고용유지가 어려워질 수 있다. 근로손실일수는 재해로 인해 피해를 본 당사자뿐만 아니라 생산성 저하와 건설 프로젝트 지연 등과 같은 영향을 미치고 그들이 속한 가족, 구성원, 고용주에게 정신적·경제적 영향을 미치며 나아가 사회적 비용뿐만 아니라 국가적으로 인적 자원의 손실을 야기 시킨다[8-11].

이에 많은 연구가 건설 현장의 업무상 재해 유형을 분석하였고, 예방을 위한 큰 노력이 있었다. 최근 데이터 분석 기법의 발전으로 데이터 마이닝, 머신러닝 등과 같은 데이터 기반 지도학습 방법과 해석 가능한 인공지능 방법론 등의 개발로 전통적인 통계 기법을 벗어나 재해 유형에 미치는 중요 특징을 추출하는데 활발히 적용되고 있다[12-14]. 하지만 산업재해 발생 후 따르는 근로손실일수에 영향을 미치는 중요 특징과 특징 안의 세부 요인에 관한 연구는 매우 적은 편이다. 작업자의 근로손실일수에 결정을 미치는 중요 특징을 추출하여 원인을 체계적으로 파악하고 기준을 설정하는 것이 근로자의 직장 복귀와 건설 생산성 향상에 필요하다. 즉, 업무상 재해 유형별 영향을 미치는 원인 분석과 달리 또 다른 관점에서 산업 재해 후 발생하는 근로손실일수, 산재 장애, 사회적 비용과 같은 부분들도 고려한다면 또 다른 통찰력을 발견할 수 있는 연구라 판단한다.

본 연구는 건설 현장에서 업무상 재해로 인한 부상자를 대상으로 근로손실일수를 결정짓는 중요 특징을 발견하고자 한다. 수집한 데이터를 기반으로 근로손실일수에서 심각도를 정의하고 이에 영향을 미치는 특징을 데이터마이닝으로 접근하여 모형을 학습하고, 특징 중요도 추출 기법으로 결과를 분석한다.

2. 근로손실일수

산업 재해의 결과는 부상과 사망으로 분류할 수 있다. 사망은 사회적인 비용이 가장 큰 부분이며 특히, 축적된 학습으로 숙련된 인적 자원이 무용지물이 되며, 부상의 경우는 장기간 요양으로 인한 근로손실일수가 증가하면서 생계 곤란, 고용유지 불안정 그리고 장애 발생 시 노동시장 퇴출 등과 같은 다양한 문제에 직면한다[15]. 산업 재해에서 말하는 근로손실일수는 근로자가 업무상 재해로 육체적·정신적 피해를 치료하기 위한 요양으로 정해진 업무시간을 정상적으로 수행하지 못하고 발생하는 손실된 시간을 뜻한다[16,17].

근로손실일수는 생산성의 손실과 근무하는 근로자들의 작업부담증가 등 조직의 비효율성을 초래한다. 결근으로 인한 근로손실일수는 질환에 기인한 결근보다는 업무상 재해로 인한 결근이 더 많으며 근로손실일수가 매우 크다[18]. 또한, 근로자의 나이가 증가할수록 재해자 수와 평균 근로손실일수가 급격한 증가세를 보인다. 일반적으로 나이가 들수록 근로자의 작업 능력이 신체적 능력이나 인지능력에서 떨어지기 때문으로 나타났다[19].

산업 재해로 인한 산재보험 환자의 평균 재원일 수와 타 보험제도의 평균 재원일 수에서도 차이가 크다. 일반병원과 특수병원 전체에서 산재보험 환자의 평균은 30.2일, 건강보험 8.1일, 의료급여 19.2일, 자동차보험 12.5일, 일반 및 기타 16.7일로 가장 높다[20]. 산재 환자는 장기 요양 환자가 많아 요양비의 부담이 커지며 전반적인 사회적 비용의 증가를 의미한다[21]. 근로자가 장기간 직장에 복귀하지 못하는 경우 회사에서 해고되거나 업무의 제한이 있다. 직장으로 돌아오더라도 근로자는 질병을 완전히 회복하지 못하고 업무를 수행한다[22,23].

특히, 건설 현장에서 발생한 업무상 재해 유형은 떨어짐, 넘어짐, 맞음, 절단, 끼임, 부딪힘 등이다. 떨어짐 사고는 떨어진 높이에 따라 심각한 신체적 손상이나 사망으로 이어지며 부상이라고 해도 근로손실일수가 매우 길다. 떨어짐으로 인해 발생하는 근로손실일수는 평균 44일로 나타났으며, 이로 인한 근로자 보상 및 입원비용이 가장 높은 재해로 연구되었다. 영국에서는 일과 관련된 건강 악화, 업무상 재해로 인하여 근로자 1인당 근로손실일수에 영향을 미치는 중요 질병들은 1,540만 건 중 660만 건이 스트레스, 우울증, 근골격계 질환에 영향을 주는 것으로 나타났다. 평균적으로, 산업 재해로

고통받는 사람은 16.5일의 근로손실일수로 나타났다. 질병으로는 19.8일, 스트레스 우울증 또는 불안과 같은 정신적 문제는 25.8일 그리고 근골격계 질환은 14.0일이다[24]. 건설업에서 산업 재해로 인해 가장 높은 근로손실일수를 나타내는 직종은 중장비 운전자, 철근 근로자, 콘크리트 근로자, 방수 근로자, 석공업 근로자, 일반 근로자 또는 목공 노동자 등이 있다. 특히 중장비 운전자의 경우 194일로 가장 높았으며, 비용은 약 1,600만 달러로 추정되었다[25].

국내와 해외의 산업 재해를 재해 발생 빈도를 통해 단순하게 추정하는 것은 매우 위험하지만, 근로손실일수를 빈도와 강도를 환산하여 체계적으로 비교할 수 있도록 하여 OECD 회원국 중 미국, 일본, 프랑스 등을 포함한 대부분 국가에서 산업재해통계자료로 활용되고 있는 국제적으로 통용되는 지표이다. 산업재해통계 업무처리 규정에 따르면 강도율은 근로시간 합계 1,000시간당 요양으로 인한 근로손실일수로 8단계를 정의한다[26]. 이를 통해 근로손실일수를 특정 기준으로 심각도로 정의하고 이에 영향을 미치는 특징을 분석한다면 건설업에서 발생하는 업무상 재해를 좀 더 객관적으로 평가할 수 있을 것이다.

3. 연구 방법론

3.1 연구설계

본 연구는 한국산업안전보건공단에서 수집한 데이터로 모든 산업 분야에서 산재 보상보험 처리된 사망자와 부상자에 대한 사고 당시와 이후의 상세한 정보를 기록한 산업 재해 데이터이다. 본 연구는 설명한 데이터를 기반으로 근로손실일수를 분석하였다. 연구설계는 다음 Figure 1과 같다.

첫째, 산업 재해 데이터를 가공하여 분류 문제로 접근하기 위해 근로손실일수 심각도를 탐색적 자료 분석하고 심각도 기준을 정의한다.

둘째, 모형이 데이터를 학습할 수 있도록 원본 데이터를 정제하고 데이터를 모형이 학습하는 과정과 새로 정의된 근로손실일수를 분류하는 모형의 성능을 평가한다.

셋째, 두 가지 특징 중요도 기법으로 특징의 중요도를 산출한다. 특징은 통계학에서 독립변수를 뜻한다. 추출된 특징을 분석하여 근로손실일수의 심각도에 따른 관계를 파악하고 해석한다.

인과관계를 분석할 수 있는 전통적인 통계 기법들이 있지만, 최근 다양한 분야에서 좋은 성과를 나타내는 데이터 마이닝과 기계학습 모형들이 좋은 대안이 될 수 있다. 특히, 정형 데이터 학습에 활발히 사용되고 있는 랜덤 포레스트 모형을 선택해서 데이터의 잠재된 특징을 발견해보고자 하였다. 기존 의사결정나무와 다르게 확률모형인 랜덤 포레스트는 블랙박스 모형이지만, 모형을 해석할 수 있는 다양한 기법들이 연구되고 있다. 랜덤 포레스트 모형 해석에 사용되는 기본적인 MDI(mean-decrease impurity)와 확장된 개념인 MDA(mean-decrease accuracy) 두 가지를 통해 근로손실일수 심각도에 영향을 미치는 잠재된 특징을 발견하고자 한다.

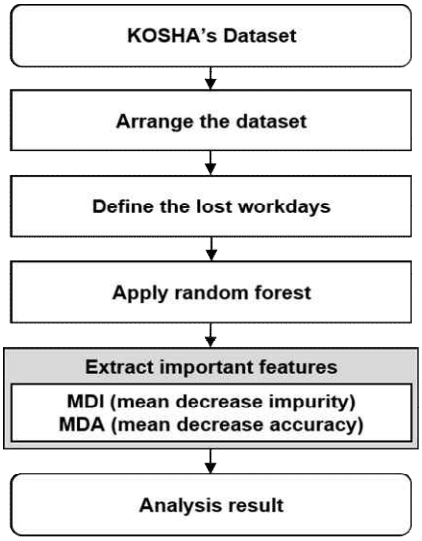

Figure 1. Process of the study

본 연구는 2008년부터 2014년까지 전체 산업 분야 중 건설 현장에서 발생한 8,921명의 재해자를 대상으로 하였다. 본 데이터는 총 55개 변수가 있으나, 연구목적에 맞게 20개를 선택하였고 변수들의 특징으로는 순위형(ordinal) 또는 명목형(nominal)으로 전부 범주형 변수들이다. 변수 중 결측값을 가진 것은 공사금액(construction fund) 18개, 고용 형태(employment type) 49개, 경력(career period) 463개, 안전 보호구 종류(personal protective equipment) 7,751개, 작업내용(work content) 2개, 떨어진 높이(heights) 6,911개이다. 결측치가 많은 안전 보호구 종류와 떨어진 높이는 다음과 같이 전처리하였다. 사고가 발생한 후 조사하였을 때 안전 보호구를 착용하지 않은 상태로 결측치는 착용안함으로, 떨어진 높이는 업무상 재해 유형과 비교하여 떨어짐 재하가 아닌 대상자는 0으로 처리하였다. 산업 재해 데이터의 변수명과 세부 요인은 다음 Table 1과 같다.

Table 1. Summary of dataset on occupational accident/injures in construction sites

Variable names	Factors (No. of factors)
Age	50-54, 55-59, 45-49,... (12)
Company size	<5, 5-9, 15-29,... (11)
Construction funds	300k-500k, 2,000k-5,000k,... (18)
Occupation	architectural woodworker, mining worker,... (128)
Employment types	daily, Full-time, Temporary,... (5)
Career periods	under 1 month, 10-20 years, over 20 years,... (14)
Occupational injury	nonfatal, Fatal (2)
Diagnose	injury of bone, berve, spinal cord, Intracranial injury traumatic,... (21)
Injured body part	upper limb, Soma(body), Head,... (9)
Accident types	fall from height, collision, rollover,... (9)
Original-cause materials	scaffolding or workstep, Stairway or ladder,... (50)
Assailing material	outdoor floor or ground, Indoor floor, form,... (372)
Protective equipment	safety harness, safety helment, safety shoes,... (11)
Work process	finishing work areas, Steel structure or concrete construction work areas,... (41)
Work content	mechanical equipment installation and maintenance work, form assembly and disassembly work,... (48)
Unsafe state	not installed and defective edge protection, Toe board was not installed or defect,... (80)
Unsafe behavior	not wearing or poor wearing protective equipment, omitted or defective work platform or ladder,... (97)
Fall height	3m-5m, under 2m, 2m-3m,... (8)
Fall places	ladder, Scaffolding structure,... (39)
Lost workdays	>= 181, 29-90, 91-180,... (4)

3.2 분석 방법

3.2.1 심각도 정의

산업 재해 데이터에 기록된 근로손실일수 값은 연속형(continuous)이 아닌 범주형(categorical)이다. 근로손실일수 변수는 본 연구에서 목표변수로 사용될 값으로, 총 네 가지 값으로 구성되어 있다. 28일 이하 808개, 29일에서 90일 3,237개 91일에서 180일 2,008개 그리고 181일 이상 1,873개로 불균형 클래스를 가진 데이터이다. 불균형 클래스를 가진 데이터를 모형에 학습용으로 사용하게 되면 한쪽 클래스로 편향될 수 있어 새로운 데이터를 예측하지 못하는 부정확한 모형이 만들어진다. 본 연구는 데이터가 가진 근로손실일수를 불균형 클래스 문제를 해소하기 위해 네 개의 클래스를 균형을 맞추기 위해 이진 클래스(binary class)로 정의하였다. 이는 국내에서 중대 재해를 기준으로 하는 일수와 같아 근로손실일수 심각도(injury severity)는 180일 이하는 보통(moderate), 181일 이상은 심각(severe)으로 구분하였다. 학습에 사용될 각 클래스의 개수는 보통 6,064개이며 심각은 5,161개이다. 심각도에 따른 영향 요인들을 분석하기 위한 기계학습 기법과 중요도 추출을 위한 방법은 다음과 같다.

3.2.2 랜덤 포레스트

하나의 모형을 학습하는 것보다 여러 개의 모형을 생성한 다음 결합하여 예측 정확성을 높이는 것을 앙상블(ensemble)이라고 한다. 랜덤 포레스트는 기계학습에서 대표적인 앙상블 모형으로 다수의 의사결정나무(decision tree)를 생성하여 하나의 모형으로 사용한다. 의사결정나무는 계층 구조로 이루어진 노드(nodes)와 에지(edges)의 집합으로 구성되어 있으며 이름 그대로 결정을 내리기 위한 트리 구조이다. 데이터에서 발견되는 패턴을 예측 가능한 규칙들의 조합으로 나타내며 대상을 좁혀 나가는 방법으로 의사결정나무에는 ID3, C4.5, CART 등과 같은 다양한 종류의 알고리즘이 있다[27].

랜덤 포레스트는 CART(classification and regression tree)를 사용한다[28]. CART는 다른 알고리즘과 다르게 이진 트리(binary tree)만 생성하며 리프 노드(leaf node)를 제외한 나머지 노드는 자식 노드를 두 개씩 가진다. 각 노드는 데이터를 특징과 임계 값(threshold)으로 구성된 각 분할에 대해 데이터를 불순도를 계산하여 하위 집합을 생성한다. 각 노드의 불순도를 계산하는 방법으로 지니(Gini)와 엔트로피(entropy)가 있다. 엔트로피는 분자의 무질서함을 측정하는 것으로 분자가 안정화되면 엔트로피는 0에 가깝다. 지니와 엔트로피는 크게 차이가 없으며 비슷한 트리를 만들어낸다. 하지만 지니 불순도는 연산시간이 빠르지만, 불균형 클래스에서 한쪽으로 치우치는 경향이 존재하고 엔트로피는 좀 더 균형 잡힌 트리를 생성한다.

의사결정나무의 작동방식은 매우 직관적이며 해석이 쉽지만, 결과 또는 성능의 변동 폭이 매우 크다는 단점을 가지고 있다. 특히 학습 데이터에 따라 생성되는 의사결정나무는 무작위성에 따라 매우 달라서 과적합(overfitting)되기 쉽고 일반화하기 어렵다. 그리고 계층적 접근방식으로 학습 도중에 에러가 발생하면 다음 단계로 전파되는 특징을 가지고 있다.

랜덤 포레스트는 의사결정나무의 단점을 극복하고 월등히

높은 정확성, 간편하고 빠른 학습, 변수 소거(elimination) 없이 수천 개의 특징을 다룰 수 있으며 무작위성으로 생성되는 의사결정나무의 형태가 조금씩 달라 비상관화(decorrelation)되어 일반화(generalization) 성능을 향상하게 시킨다. 무작위성은 부트스트랩(bootstrap)을 통해 조금씩 다른 훈련 데이터를 학습한 의사결정나무를 결합하는 방법이다. 이를 배깅(bagging)이라고 한다. 또한, 훈련 단계에서 임의 노드 최적화를 통해 목적 함수를 최대로 만드는 노드 분할 함수의 최적화된 매개변숫값을 구한다. 배깅[29]을 통해 모형을 훈련하는 과정은: (1) 부트스트랩으로 n개의 훈련 데이터를 생성한다. (2) T개의 결정 트리를 훈련한다. (3) 훈련된 결정 트리들을 하나의 모형으로 결합하여 평균(average) 또는 투표(vote) 방식을 사용하여 결괏값을 예측한다.

3.2.3 특징 중요도

데이터 마이닝을 통해 데이터에서 통계적 규칙이나 패턴을 분석하여 가치 있는 정보를 추출하는 과정이며, 이를 데이터베이스 속의 지식 발견이라 통찰력을 발견할 수 있다. 특징 중요도는 학습된 모형에서 종속 변수에 영향을 미치는 독립변수들의 영향도이다. 피처 엔지니어링(feature engineering)에서 머신러닝 모형을 작동하기 위해 데이터에 대한 도메인 지식을 활용하여 특징 또는 정형 데이터셋에서 속성, 독립변수라 불리는 것들을 새로 만들거나 정리하는 과정이다. 다른 말로, 머신러닝 모델을 위한 데이터 테이블의 칼럼을 조합하여 파생 변수를 생성하거나 수많은 변수 중에서 종속 변수에 영향을 끼치는 변수를 선택하거나 삭제하는 작업을 의미한다. 피처 엔지니어링은 모델 성능에 미치는 영향이 크기 때문에 데이터마이닝/머신러닝 응용에 있어서 중요한 단계이며, 전문성과 시간과 비용이 많이 드는 작업이다.

다양한 모형들이 특징에서 중요도를 산출할 수 있다. 랜덤 포레스트는 각 특징의 중요도를 계산해서 순위를 매길 수 있다. 이를 특징 중요도라고 하며 상대적 특징의 중요도를 계산하여 학습된 랜덤 포레스트 모형을 해석할 수 있다. 의사결정나무의 특징 중요도는 일부 특징을 완전히 배제하지만, 무작위성을 가진 랜덤 포레스트 모형은 거의 모든 특징에 대해 평가할 수 있다. 이러한 장점으로 랜덤 포레스트 모형은 의사결정나무 모형보다 더 넓은 시각으로 데이터를 살펴볼 수 있다.

특징 중요도는 어느 특징을 사용한 노드가 평균적으로 불순도를 얼마만큼 감소시키는지 확인하여 중요도를 계산한다. 특징 중요도를 도출하기 위한 대표적인 방법은 평균 불순물 또는 지니 중요도 감소(mean decrease impurity or Gini importance)라고 한다. DT의 특징 중요도는 노드에 사용된 특징별 (현재 노드의 표본 비율 불순도)-(왼쪽 자식 노드의 표본 비율 불순도)-(오른쪽 자식 노드의 표본 비율 불순도)이다. 표본 비율은 트리 전체 샘플 수에 대한 비율이며 RF의 특징 중요도는 각 DT의 특징 중요도를 모두 계산하여 더한 후 중요도의 합이 1이 되도록 전체 합으로 나누어 정규화(normalize)한다[30]. 그러나 평균 감소 불순도를 사용한 특징 중요도는 연속된 값이나 범주형 데이터에서 편향이 발생하여 매 결과가 달라질 수 있어서 신뢰성이 낮다[31,32]. 이를 극복하기 위해 순열(permutation) 방법을 적용한 평균 감소 정확도(mean decrease accuracy)를 사용하였다.

평균 감소 정확도는 순열 중요도(permutation importance)라고도 불리며 기본 개념은 각 특징의 값을 치환하고 해당 순열이 모형의 정확도 점수 측정에 얼마나 부정적인 영향을 미치는지 측정하는 것이다. 이는 학습된 모형이 특정한 특징 없이 기본 모형에 대비하여 성능의 차이가 발생하는지를 확인한다. 그러나 순열 중요도의 단점은 각 특징에 대한 모형을 재학습하는데 들어가는 계산 비용이 많이 든다는 것이다.

4. 분석 결과 및 고찰

이 장에서는 학습된 데이터에서 근로손실일수에 관한 탐색적 자료 분석, 랜덤 포레스트 모형을 학습 방법 및 성능 평가, 그리고 두 가지 특징 중요도를 도출하여 각 특징을 분석하고 논의한다. 근로손실일수 심각도를 분석하기 위해 Ubuntu 20.04 환경에서 Python 3.8 그리고 pandas 1.2, scikit-learn 0.24 버전의 라이브러리를 사용하였다.

4.1 근로손실일수의 탐색적 자료 분석

건설 현장에서 업무상 재해로 인해 다친 작업자들의 근로손실일수 빈도를 분석하면 Figure 2과 같다. KOSHA의 산업재해 데이터에는 총 4개의 범주형 값으로 구성되어 있다. 28일 이하의 근로손실일수가 가장 적으며 91일에서 180일, 29일에서 90일 그리고 181일 이상 근로손실일수 순으로 빈도가 높게 나타났다. 직접적인 비교는 어렵지만, 미국 Department of Labor의 보고서에 따르면, 2019년 건설에서 발생한 근로손실일수는 평균 79일이다[33].

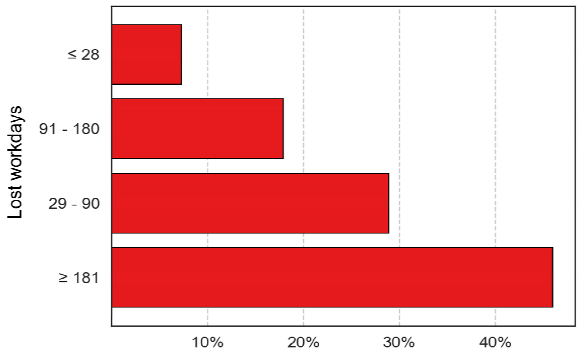

Figure 2. Frequent lost workdays based on the occupational accidents

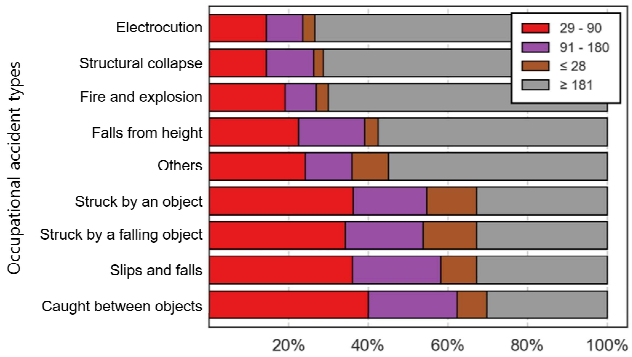

Figure 3. Occupational accident types based on the lost workdays

우리나라 건설 현장에서 사고가 발생하면 근로손실일수가 181일 이상으로 작업자의 심각한 상태로 이어진다. 해당 데이터는 산재보험 처리된 기준으로 기록된 데이터이기 때문에 산업재해현황분석 재해자 2만 명 이상과 수가 일치하지 않는 것은 가벼운 부상은 산재 처리를 하지 않았다고 짐작해볼 수 있다. 근로손실일수 181일 이상을 제외한 나머지 3개의 근로손실일수를 합하여 심각과 보통 클래스로 구분하였다. 그리고 생존한 근로자를 기준으로 중대 재해를 분석하기 위해 사망으로 기록된 근로자의 데이터는 제거하였다.

9가지 업무상 재해 유형에 따른 근로손실일수의 빈도 분석 결과는 Figure 3과 같다. 감전, 붕괴, 화재의 경우, 181일 이상이 가장 높게 나타나는 재해 형태로, 사고가 발생하면 심각한 신체적 손상을 받는 사고로 나타났다. 건설 현장에서 전체 27,211건에서 화재폭발파열 111건, 무너짐 360건, 기타에 포함된 감전 416건이다. 재해가 발생하면 근로손실일수가 매우 길어지는 중대 재해로 해석된다. 두 번째 특징이 나뉘는 재해 유형은 떨어짐(fall from height)과 기타(others)이다.

나머지 끼임, 물체에 맞음, 전복, 등은 중대 재해가 상대적으로 낮으며 28일 이하의 비율도 다른 사고들에 비해 높은 것으로 나타났다. 3대 다발 재해는 감김 끼임과 같은 안전사고이기 때문에 생명에 지장을 주는 정도의 재해는 아닌 것으로 판단된다.

181일 이상 근로손실일수가 나타나는 업무상 재해 유형은 감전, 붕괴, 화재, 추락으로 심각한 사고로 나타났으며, 이와 반대는 부딪힘, 물체에 맞음, 넘어짐, 깔림-뒤집힘, 끼임과 같은 사고는 근로손실일수가 대부분 29 – 90 사이로 나타났다.

4.2 랜덤 포레스트 모형 구축 결과

데이터 클리닝과 전처리 단계를 거쳐 모형을 구축하는 데 사용하는 학습 데이터셋과 테스트 데이터셋(test set)을 8:2로 나눴다. 그리고 구축할 모형에 사용될 하이퍼 파라미터를 설정하고 *k*-교차 검증(*k*-cross validation)을 5겹으로 사용하여 학습한 뒤 결과를 평가하고 일반화(generalization) 성능이 높을 때까지 반복한다.

랜덤 포레스트에는 학습 과정에서 과적합을 방지하고 일반화 성능을 높이기 위한 다양한 하이퍼 파라미터가 있다. 하이퍼 파라미터는 모형이 훈련 데이터에 과적합되지 않고 새로운 데이터에서도 예측 성능이 뛰어나도록 일반화시키는 것이다. 랜덤 포레스트는 의사결정나무를 기반으로 모형이 만들어지기 때문에 학습 데이터셋에 대한 제약사항이 없다. 따라서 제한을 두지 않으면 훈련 데이터에 아주 가깝게 맞추려고 해서 과적합이 되기 쉽다. 의사결정나무는 훈련되기 전에 파라미터 수가 결정되지 않기 때문이다. 이런 모형을 비파라미터 모형(non-parametric model)이라고 한다. 반대로 선형모형 같은 파라미터 모형은 미리 정의된 모형 파라미터 수를 가지므로 자유도가 제한되고 과적합될 위험이 줄어들지만, 과소적합될 경향이 높다. 이를 방지하는 방법으로는 하이퍼 파라미터를 조정하는 것이다. 하이퍼 파라미터에는 대표적으로 의사결정나무의 최대 깊이 제한(max depth), 의사결정나무의 자식 노드로 분할되기 위해 부모 노드가 가져야 할 최소 데이터 수(min sample split), 의사결정나무의 최종 노드의 수(max leaf nodes), 최종 노드에서 최소 가중치를 고려한 데이터의 수(min weight fraction leaf) 등이 있다. Figure 4는 파라미터 변화에 대한 예시이며 하이퍼 파라미터 중 최대 깊이에 따른 성능 변화를 나타낸다.

모형의 성능을 높이기 위해 하이퍼 파라미터를 조정하는

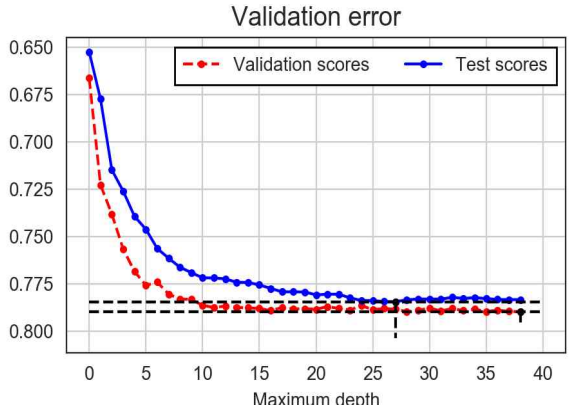

Figure 4. Example of variation based on hyper-parmeter of max depth

다양한 방법론이 존재한다. 가장 좋은 접근 방법은 모든 경우의 수를 고려하여 최적의 파라미터를 찾는 것이 가장 좋겠지만 현실적으로 시간적·물리적 제약으로 불가능하다. 따라서 랜덤 포레스트 모형의 대표적인 하이퍼 파라미터 다섯 가지를 선택하고 해당 값의 구간을 지정한 뒤 무작위 탐색을 통해 랜덤 포레스트를 최적화하였다. Figure 5는 최적화 가정에서 모형의 성능이 수렴해가는 과정을 시각화한 것이다. 최적화 과정은 50회 반복하였으며 반복 횟수 20회 중반 구간부터 성능의 변화가 없으며 이 구간에 나타난 하이퍼 파라미터가 최적해는 아니지만, 근사해로 받아들이고 모형의 학습을 종료하였다.

모형의 전체적인 성능은 검증 데이터 정확도 80.01%, ROC-AUC 84.2%로 도출되었다. 근로손실일수의 심각은 잘 구분되지만, 보통은 상대적으로 낮은 성능을 나타낸다. 그러나 심각한 건설 재해의 관리가 보통의 건설 재해보다 더 중점적으로 관리되어야 한다는 측면에서 분석할 가치가 높다고 판단하였다.

4.3 근로손실일수 심각도의 특징 중요도

모형의 분류 성능을 평가해 선택된 모형을 사용하여 근로손실일수 분류에 영향을 미치는 특징 중요도를 두 가지 기법을 통해 Figure 5와 같이 시각화하였다.

상단 MDI에서 특징 중요도 순위는 질병명(diagnose)과 부상 부위(injured body parts) 등으로 각 특징의 상자 그림(box plot)을 살펴보면, 중앙값(median)이 비슷하고 중요도의 편차가 매우 심한 것을 알 수 있다. 하지만 하단의 MDA를 사용한 중요도를 보면 편차가 상대적으로 적고 질병명과 부상 부위의 중앙값 차이가 MDI와 다르게 큰 것을 볼 수 있다.

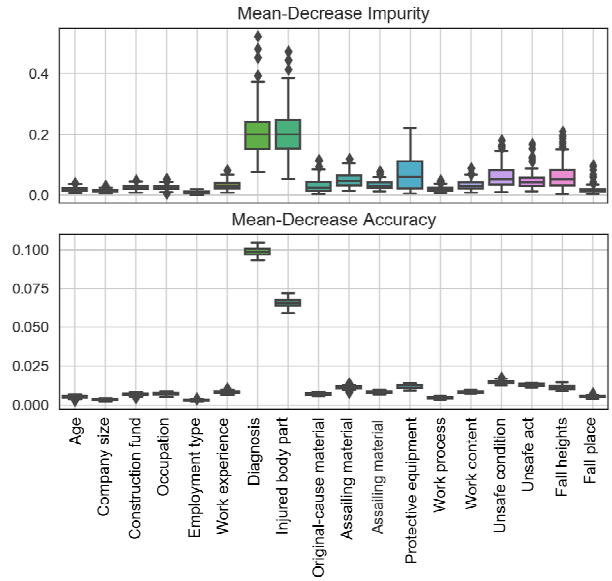

Figure 5. Comparison of feature importances between MDI and MDA

특히 MDI는 랜덤 포레스트의 무작위성으로 인해 실험마다 질병명과 부상 부위의 중요도가 달라진다. MDI는 각 의사결정나무에서 계산된 중요도 값의 불순도의 분산이 큰 것을 확인할 수 있다. 따라서 배깅 과정에서 데이터를 추출할 때마다 특징 중요도가 달라질 수 있고 무작위성으로 인해 중요도 값이 달라질 수 있다. 이와 반대로 MDA 기법으로 추출한 중요도의 편차가 MDI보다 상대적으로 적다. 따라서 본 연구는 특징 중요도를 추출하는 연구에서 기본적인 MDI보다 MDA 방법을 제안한다. MDA는 매번 새로운 모형을 학습할 때 변수를 하나씩 제거하면서 해당 변수가 학습에 영향이 없다고 판단하는 과정을 통해 중요도를 추출하기 때문에 좀 더 높은 신뢰성을 가진다.

분석 결과, MDI의 특징 중요도의 순위는 질병명, 부상 부위, 안전 보호구 종류, 불안전 상태, 떨어진 높이, 불안전 행동, 기인물, 업무상 재해 유형 순으로 나타났으며, MDA의 중요도 순위는 질병명, 부상 부위, 불안전 상태, 불안전 행동, 떨어진 높이 등의 순서로 도출되었다. 업무상 재해 유형을 예측하는 기존 연구와 비교하면 업무상 재해 유형별 판별에 영향이 높은 변수는 발생 계절, 프로젝트 형태로 나타났으며[12], 인공신경망을 적용한 연구에서는 발생 계절, 프로젝트 유형, 고용형태 등으로 중요도가 높다고 밝혔다. 하지만 기존 연구에서 사용한 데이터는 관련 변수가 적어 재해 유형에 크게 상관없다고 볼 수 있는 결론만 도출할 수 있었다[13]. 계절의 중요도가 높은 이유는 특정 계절에 사고의 빈도가 집중되어있기 때

문이다[2]. 본 연구와 같이 랜덤 포레스트로 특징 중요도를 뽑아낸 연구는 나이, 층수, 작업유형, 직종, 근무 연수, 재해 횟수, 계절 공사, 종류 등으로 나타났다[14]. 이러한 이유는 학습에 사용된 데이터의 설계와 건설안전에서 이론적인 불안전 행동이나 불안전 상태와 행동 등과 같이 상세한 정보를 수집하지 못했기에 나타난 경향으로 파악된다. 본 연구는 이런 것들과 불안전 행동, 기존 이론에서 나타난 불안전 행동 상태 등과 같은 수치를 정량적으로 증명되었고, 특징의 중요도로 유의미하다고 판단된다.

5. 결 론

건설 현장에서 업무상 재해로 인해 발생한 근로손실일수는 개인적, 기업적 그리고 사회적 측면에서 큰 비용과 인적 손실을 일으킨다. 이러한 근로손실일수를 예측하고 예방하는 것은 개인의 회복 기간과 관련 지출 비용, 직장 복귀 시기, 건설 프로젝트의 지연 및 중단을 최소화하고, 사회적 비용을 절감할 수 있다. 고 건설 프로젝트의 지연 및 중단을 최소화할 수 있다.

본 연구는 데이터 마이닝을 통해 건설 현장에서 발생한 업무상 재해가 근로손실일수 심각도에 미치는 영향을 분석했다. 심각도에 미치는 영향 변수를 추출하기 위해 특징 중요도 기법으로 수치화하였고, 기존 연구들과 차이점을 발견했다. 연구의 결과는 다음과 같다. 첫째, 사고 데이터를 학습한 랜덤 포레스트 모형의 분류 성능은 84.2%로 심각(181일 이상) 클래스의 분류 성능이 매우 높다. 둘째, 두 가지 특징 중요도 적용하여 신뢰성 높은 MDA 방법을 제안하고 근로손실일수 심각도에 영향을 미치는 특징을 추출하였다. 추출된 중요한 특징은 질병명, 부상 부위, 불안전 상태, 안전 보호구 착용 여부이다. 셋째, 데이터 학습 기반 모형에서 가장 중요한 부분 데이터로, 기존 연구의 데이터와 비교하여 특징 중요도가 바뀌는 패턴을 분석하였다.

특징 중요도를 통해 근로손실일수는 산업 재해 발생 후 파생되는 특징이기 때문에 질병명과 부상 부위에 영향을 크게 받는다. 하지만 불안전 상태와 불안전 행동을 예방할 수 있다면 업무상 재해와 근로손실일수가 절감될 수 있다. 산업 재해를 예방하기 위해서는 안전 교육, 안전 불감증, 안전 보호구 착용을 잘하며 감시·관리가 중요하다. 본 데이터에서 재해의 절반 이상이 안전 보호구를 착용하지 않은 상태였기 때문에 안전 보호구의 착용 상태를 확인하는 것이 매우 중요하다고 분석되었다.

본 연구에서는 기존의 건설안전 연구에서 제시한 특징 중요도 추출 기법과 제안한 방법에 따라 추출한 특징의 차이와 모형이 학습한 데이터에 따라 유의미한 분석 결과에 영향을 미칠 수 있다는 것을 확인하였다. 또한 본 연구는 기존의 업무상 재해 유형이 아닌 근로손실일수를 다시 심각도로 정의하여 이에 영향을 미치는 중요한 특징을 도출하고 분석하였다. 건설 현장의 안전 관리자는 건설 현장에서 기록된 데이터를 모형에 적용하여 건설안전사고로 인해 발생한 근로손실일수를 최소화하기 위한 영향 요인들을 파악함으로써 건설안전사고를 예방하고 개인, 기업, 사회적 비용을 최소화할 수 있다. 차후 특징 중요도로 추출한 특징을 기반으로 특징 내 요인들이 근로손실일수에 영향을 미치는 연관관계와 같은 구체적인 연구가 필요하다.

요 약

건설업은 전체 산업 분야 중에서 가장 많은 재해와 사망자를 발생시키는 산업 분야이다. 건설안전 재해를 줄이기 위한 큰 노력이 진행되어왔지만, 사망사고를 제외한 근로자의 업무 복귀시간까지 회복되는 근로손실일수에 관한 연구는 매우 적은 편이다. 따라서 본 연구는 근로손실일수를 심각도로 정의하여 이를 분류하는 모형을 제안하고 학습된 모형을 통해 특징 중요도를 도출하고 중요한 특징을 분석하고자 하였다. 블랙박스 모형인 랜덤 포레스트의 학습 과정을 해석하고 추출된 특징 중요도를 통해 근로손실일수 심각도에 영향력을 행사하는 중요 변수를 추출하였다. 추출된 특징을 통해 내부에 존재하는 요인들을 분석하였다. 본 연구의 목적은 건설 현장에서 발생한 사고 사례 데이터를 랜덤 포레스트 모형을 통해 분석하고자 하였다. 근로손실일수의 심각도에 미치는 중요한 특징을 도출해 체계적으로 관리한다면 건설 재해를 예방할 수 있다.

키워드 : 건설안전, 근로손실일수, 데이터마이닝, 랜덤 포레스트, 특징 중요도

Funding

This paper was supported by the Sahmyook University Research Fund in 2020.

ORCID

Kyung-Su Kang, https://orcid.org/0000-0002-6955-578X
Jae-Hyun Choi, https://orcid.org/0000-0002-8199-0311
Han-Guk Ryu, https://orcid.org/0000-0001-6101-560X

References

1. Occupational Safety & Health Research Institute. Analysis of Industrial Accidents in 2019. Ulsan (Korea): Korea Occupational Safety and Health Agency; 2020. p. 7-640.
2. Kang KS, Ryu HG. Predicting types of occupational accidents at construction sites in Korea using random forest model. Safety Science. 2019 Dec;120:226-36. https://doi.org/10.1016/j.ssci.2019.06.034
3. Ale BJM, Bellamy LJ, Baksteen H, Damen M, Goossens LHJ, Hale AR, Mud M, Papazoglou IA, Whiston JY. Accidents in the construction industry in the netherlands: An analysis of accident reports using storybuilder. Reliability Engineering & System Safety. 2008 Oct;93(10):1523-33. https://doi.org/10.1016/j.ress.2007.09.004
4. Swuste P, Frijters A, Guldenmund F. Is it possible to influence safety in the building sector?: A literature review extending from 1980 until the present. Safety science. 2012 Jun;50(5):1333-43. https://doi.org/10.1016/j.ssci.2011.12.036
5. Liao CW, Perng YH. Data mining for occupational injuries in the Taiwan construction industry. Safety science. 2008 Aug;46(7):1091-102. https://doi.org/10.1016/j.ssci.2007.04.007
6. Winge S, Albrechtsen E. Accident types and barrier failures in the construction industry. Safety science. 2018 Jun;105:158-66. https://doi.org/10.1016/j.ssci.2018.02.006
7. Hosseinian SS, Torghabeh ZJ. Major theories of construction accident causation models: A literature review. International Journal of Advances in Engineering & Technology. 2012 Sep;4(2):53-66.
8. Lopez-Alonso M, Ibarrondo-Davila MP, Rubio-Gamez MC, Munoz TG. The impact of health and safety investment on construction company costs. Safety science. 2013 Dec;60:151-9. https://doi.org/10.1016/j.ssci.2013.06.013
9. Santana VS, Villaveces A, Bangdiwala SI, Runyan CW, Albuquerque-Oliveira PR. Workdays lost due to occupational injuries among young workers in Brazil. American journal of industrial medicine. 2012 Oct;55(10):917-25. https://doi.org/10.1002/ajim.22099. Epub 2012 Jul 27
10. Manu P, Ankrah N, Proverbs D, Suresh S. An approach for determining the extent of contribution of construction project features to accident causation. Safety Science. 2010 Jul;48(6):687-92. https://doi.org/10.1016/j.ssci.2010.03.001
11. Boden LI, Biddle EA, Spieler EA. Social and economic impacts of workplace illness and injury: current and future directions for research. American Journal of Industrial Medicine. 2001 Oct;40(4):398-402. https://doi.org/10.1002/ajim.10013
12. Cho YR, Kim YC, Shin YS. Prediction model of construction safety accidents using decision tree technique. Journal of the Korea Institute of Building Construction. 2017 Jun;17(3):295-303. https://doi.org/10.5345/JKIBC.2017.17.3.295
13. Kim YC, Yoo WS, Shin YS. Application of artificial neural networks to prediction of construction safety accidents. The Journal of the Korean Society of Hazard Mitigation. 2017 Fed;17(1):7-14. https://doi.org/10.9798/KOSHAM.2017.17.1.7
14. Kim EJ. Prediction model for construction safety accidents using random forest. Journal of The Regional Association of Architectural Institute of Korea. 2020 Oct;22(5):81-7.
15. Weil D. Valuing the economic consequences of work injury and illness: a comparison of methods and findings. American Journal of Industrial Medicine. 2001 Oct;40(4):418-37. https://doi.org/10.1002/ajim.1114
16. Jeong WI, Lee KS, Jeon YI. Occupational accidents and foregone working days. Journal of Korean Economics Studies. 2011 Jun;29(2):139-74.
17. Choi JW, Kim TW, Lee CS. Effects of weather factors on the work loss days of the elderly workers. Korean Journal of Construction Engineering and Management. 2019 Jan;20(1):41-51. https://doi.org/10.6106/KJCEM.2019.20.1.041
18. Lee DB, Lee TY, Jo YC, Lee YS, Sim UT. Analysis of absenteeism factors of workers in manufacturing industries. Industrial Health. 1995 May;85:2-9.
19. Yoon JH, Jang SR, Im HK. Occupational accident prevention policy for middle-aged and old workers according to the trend of occupational accident. Ergonomics Society of Korea 2008 Conference; 2008 May 23; Gumi-si, Korea. Seoul (Korea): Ergonomics Society of Korea; 2008. p. 8-11.
20. Korea Health Industry Development Institute. Hospital management status survey - The average number of hospital stays for inpatients by type of salary [Internet]. Cheongju (Korea): Korea Health Industry Development Institute; 2015 Set 30. Available from: https://kosis.kr/statHtml/statHtml.do?orgId=358&tblId=DT_358N_H409101
21. Jeong WM, Park CY, Koo JW, Roh YM. Predictors of return to work in occupational injured workers. Annals of Occupational and Environmental Medicine 2003;15(2):119-31.
22. Jo JH. A study on the Causes Analysis and Preventive Measures by Disaster types in Construction Fields. Conference of Korea Safety Management and Science; 2011 Nov; Cheonan-si, South Korea. Incheon (Korea): Korea Safety Management and Scienc

e; 2011. p. 23-34.
23. Ruser JW. The changing composition of lost-workday injuries. Monthly Labor Review. 1999 Jun;122(6):11-9.
24. Health and Safety Executive. Working days lost in Great Britain [Internet]. Merseyside (England): Health and Safety Executive; 2017. Available from: http://www.hse.gov.uk/statistics/dayslost.htm
25. Yang YS, Park JH, Lee CS. Accident risk analysis of construction workers by occupation. Journal of the Architectural Institue of Korea Structure & Construction. 2009;25(10):149-56.
26. Kim SG, An HY, Lee EH. A comparative study on the changes in indicators of occupational accidents and social and economic activities in OECD countries. Ulsan-si (Korea): Occupational Safety and Health Research Institute. 2009 Dec. Report No.: 2009-72-1264.
27. Kotsiantis SB. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering. 2007 Jun;160:3-24.
28. Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees. CRC press. 1984. p. 237-51.
29. Geron A. Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. 1st ed. Sebastopol (CA): O'Reilly Media. 2017. p. 189-211.
30. David D. Summing feature importance in Scikit-learn for a set of features [Internet]. StackExchange; 2017 [updated 2020 Sep 11; cited 2021 Feb 12]. Available from: http://bit.ly/2TYUjnu
31. Breiman L. Random forests. Machine learning. 2001 Oct;45(1):5-32.
32. Parr T, Turgutlu K, Csiszar C, Howard J. Beware Default Random Forest Importances. explained.ai [Preprint]. 2018 [cited 2021 Feb 12]. Available from: https://explained.ai/rf-importance/
33. U.S. Bureau of Labor Statistics. Employer-Reported Workplace Injuries and Illnesses (Annual) News Release [Internet]. U.S. Department of Labor, Washington (DC): U.S. Bureau of Labor Statistics; 2020 Apr 11 [updated 2020 Nov 4; cited 2021 Feb 28]. Available from: https://www.bls.gov/news.release/osh.htm