# 시계열 기반 지역별 기상 정보를 활용한 119 신고 건수 예측을 위한 CatBoost 모델 제안

참가번호: 250664
팀명: 핀토레스트
공모주제: 소방데이터와 날씨 빅데이터를 융합한 119신고 건수 예측

## 1. 분석 배경 및 목표

최근 기후 변화는 전례 없는 자연재해의 발생 빈도와 규모를 증대시키고 있으며, 이는 인명 및 재산 피해의 심각성을 가중시키고 있다. 행정안전부의 '2023 재해연보(자연재난)'에 의하면, 2023년 자연재해로 인한 재산 피해액이 약 3,700억 원에 달하여, 이는 국민의 안전과 재산 보호에 대한 국가적 부담을 심화시키고 있음을 시사한다. 현재의 사건 사고 대응 방식은 신고 접수 후 이루어지는 사후적, 수동적 형태에 머물러 있으며, 긴급 상황 발생 시 골든타임 확보의 어려움과 자원 배분의 비효율성을 야기함으로써 피해 예방 및 최소화에 한계를 내포한다.

본 연구는 이러한 문제의식에서 출발하여, 기상 상황과 119 신고 건수 간의 유의미한 상관관계를 규명하고, 이를 기반으로 지역별 기상 정보를 활용한 119 신고 건수 예측 모델을 개발하는 것을 목표로 한다. 실제로 한국방재학회 논문 '기상요인이 119 구급서비스에 미치는 영향(2020)' 등 다수의 선행 연구들은 기온, 강수량, 습도와 같은 기상 요인이 119 구급 출동 건수에 유의미한 영향을 미침을 실증하고 있다. 119 신고 건수가 1~2건인 일반적인 경우를 안정적으로 예측하면서도, 신고 건수가 많아 추가적인 소방 자원이 필요한 경우를 예측할 수 있다면,, 관련 인력 및 자원의 효율적인 분배 및 투입을 통해 제한된 시간 내에 최대의 재산 및 인명 피해 절감을 달성할 수 있을 것이다. 이는 소방 서비스의 질적 향상과 더불어 국민의 안전을 선제적으로 확보하는 데 크게 기여할 것으로 기대된다. 본 연구에서는 신고 건수가 5건 이상에 해당하는 날은 고(高)신고일, 5건 미만은 저(低)신고로 간주한다.

## 2. 분석 데이터 및 전처리

| 변수명 | 설명 | 변수명 | 설명 |
|---|---|---|---|
| tm | 신고접수 날짜 | ta_max_min | 일교차 |
| address_city | 시/도명 | hm_min | 일 강수량 |
| address_gu | 군/구명 | hm_max | 일 최대 순간 풍속 |
| sub_address | 읍/면/동 명 | ws_max | 일 최대 10분 평균 풍속 |
| stn | AWS 지점 코드 | ws_ins_max | 일 최소 상대 습도 |
| ta_max | 일 최저 기온 | rn_day | 일 최대 상대 습도 |
| ta_min | 일 최고 기온 | call_count | 일 신고 건수 |

[표 1] 기상청에서 제공한 원천 데이터

기상청에서 제공한 지역별 기상상황 및 접수된 119 신고 건수에 대한 시계열 데이터에 대해 진행된 데이터 가공 및 생성은 다음과 같다.

공통 사안

- 각 컬럼 명칭의 접두사 &call119_train.& 및 공백 문자 제거.
- 데이터 입력 값 중 99.0은 결측치로 간주해 제거.
- test 데이터에서 아직 작성되지 않은 call_count 같은 컬럼에 대해 기본값 0으로 설정.

날짜 데이터 전처리 및 파생 변수 생성

- tm: 데이터 형식을 단순문자열(YYYYMMDD)에서 datetime 형식으로 변환.
- weekday: 0=월요일, …, 6=일요일로 정의.
- month: 월 정보는 1, 2, …, 12로 정의함.
- weekday_sin, weekday_cos: 요일의 주기성을 삼각함수를 통해 정규화.
- month_sin, month_cos: 월 단위 계절성 반영.
- is_holiday: 학습 및 검증 기간에 대한 공휴일 여부를 이진 분류해 인코딩.

기상 관련 파생 변수 생성

- hm_diff: 일 습도차, 최대 상대 습도에서 최소 상대 습도를 뺀 값.
- is_heatwave: 일 최고기온 33℃ 이상인 경우 폭염으로 판단.
- is_heavy_rain: 일 강수량 50mm 이상인 경우 폭우로 판단.

지역 관련 변수 인코딩

- stn, address_gu, sub_address 등: One-Hot Encoding을 적용.

목표값 변환

- 로그 스케일링을 통해 call_count의 정규성을 보정하고 고(高)신고 건수에 대한 수치 완화. 예측 이후 역변환 적용. 이를 통해 고(高)신고 건수 관련 outlier의 영향을 최소화하면서 전반적인 신고 건수 예측 안정성 확보를 위한 조치. (y = log(1+ x))

# 3. 분석방법

본 예측 모델은 Gradient Boosting 계열의 머신러닝 라이브러리 CatBoost Regressor를 기반으로 구축하였다. CatBoost는 과적합에 강인하며, 빠른 학습 속도를 제공하는 특징이 있다. 데이터의 특성과 분포를 고려하여 로그 변환된 타겟(log1p), 시간 및 기상 파생 변수, 그리고 지역 변수의 One-hot encoding을 적극적으로 활용하였다. 본 절에서는 모델 학습 방식과 구조를 구체적으로 설명한다.

119 신고 건수 데이터는 전반적으로 1~5건 수준의 저(低)신고가 대부분을 차지하는 비대칭적인 분포를 가지며, 일부 지역·일자에 5건 이상의 고(高)신고가 드물게 나타나는 특성을 가진다. 이러한 데이터에서는 일반적인 선형 회귀나 단순 트리 모델로는 고(高)신고 대응이 어렵고, 반대로 고(高)신고에 민감한 모델은 전체적인 RMSE 성능을 저하시킬 수 있다.
CatBoost는 잔차 기반 학습을 통해, 고(高)신고와 같은 특이값을 반복적으로 학습하면서도 저(低)신고의 일반적 분포를 무시하지 않으며, log1p 변환과 결합 시, 소수 고(高)신고 사례의 영향을 부드럽게 반영하면서 전체적인 예측 안정성을 유지할 수 있다. 또한, One-hot 인코딩된 지역 변수가 수백 개에 이르는 상황에서도 빠르게 학습되며, 과적합 없이 지역 특성을 반영할 수 있다. 이러한 점에서 CatBoost는 과적합 제어와 계산 효율성, 그리고 특이 분포 대응력 측면에서 본 예측 과제에 가장 적합한 선택지라고 판단하였다.

## 3.1. 타겟 변수 및 스케일 변환

119 신고 건수는 1~5건이 대부분인 불균형 데이터로, 고(高)신고가 소수 존재한다. 이를 보정하고 회귀 모델이 적절히 학습되도록 하기 위해 타겟 변수 call_count에 대해 로그 변환(np.log1p)을 적용하였다. 학습 시에는 변환된 값을 사용하고, 예측 후에는 np.expm1()로 다시 원래 스케일로 복원한다. 로그 변환을 통해 이상치(고(高)신고)의 영향을 완화하고, 데이터 분포를 정규화하여 모델 전체 구간에서 균형 있는 학습이 가능하도록 유도하였다.

## 3.2. 모델 구조 및 작동 원리

본 예측 모델은 CatBoost Regressor를 다음과 같은 설정으로 학습하였다.

| 파라미터 | 설정값 |
|---|---|
| iterations | 500 |
| learning_rate | 0.05 |
| depth | 6 |
| loss_function | RMSE |
| verbose | 100 |

[표 2] CatBoost Regressor 파라미터

Boosting 반복 횟수, tree의 개수를 500으로, 학습률은 0.05, 결정 트리의 최대 깊이는 6으로 설정하였다. 평균 제곱근 오차 기준으로 모델을 최적화하였고, 학습 로그 또한 출력하도록 설정하여 모델을 설계하였다.

본 예측 모델은 Gradient Boosting Decision Tree(GBDT) 방식으로 반복 학습을 수행한다. CatBoost는 수백 개의 약한 결정 트리(weak learner)를 순차적으로 학습시켜 예측 성능을 점진적으로 향상시키는 방식으로 동작한다. 본 모델의 학습 절차는 다음과 같다.

(1) 초기 예측값 설정
학습은 전체 데이터의 타겟 변수(call_count)의 로그 변환 값에 대해, 평균값 또는 기본 설정값으로 초기 예측값을 설정하는 것에서 시작된다. 본 모델에서는 np.log1p 변환을 통해 고(高)신고와 저(低)신고 간의 스케일 차이를 줄이고, 학습 안정성을 확보하였다.

(2) 잔차 기반 반복 학습 (Boosting 단계)
모델은 각 반복마다 현재까지의 예측값과 실제값의 차이인 잔차(residual)를 계산하고, 이를 보정하기 위한 새로운 결정 트리를 학습한다. 이 트리는 데이터의 특정 구간(예: 공휴일, 기온 급상승, 강우량 변화 등)에 대한 오차를 집중적으로 줄이도록 설계된다. 이 과정은 본 모델에서는 500회 반복(iterations=500)으로 설정하였다.

(3) 예측값 누적 갱신
각 반복에서 학습된 트리는 전체 예측값에 소폭 가중되어 더해지며 누적된다. 즉, 예측값은 &초기 예측값 + 각 트리의 예측값 가중합& 형태로 점진적으로 보정된다. 이처럼 누적된 예측은 최종적으로 로그 스케일에서 생성되며, 모델 출력은 np.expm1()을 통해 다시 원래의 신고 건수 단위로 역변환된다.

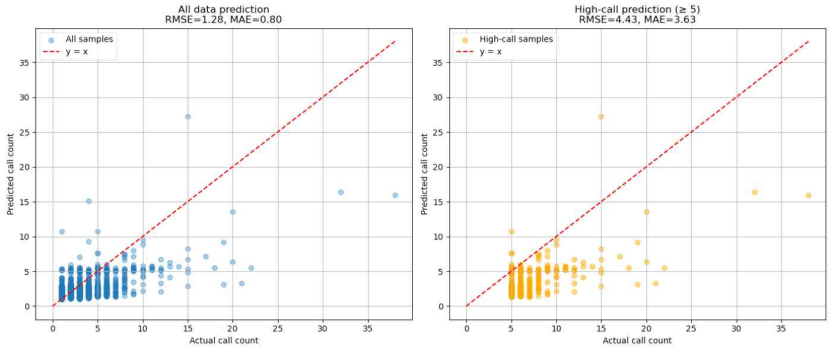
# 4. 결과 및 검증

## 4.1 실제 신고 건수와 예측 신고 건수 산점도 분석

아래 산점도는 2020년부터 2023년 데이터를 기반으로 학습한 모델의 예측 성능을 시각화한 것이다. 좌측 그래프는 예측의 전체 검증 데이터에 대한 예측 결과이며, 우측 그래프는 실제로 신고 건수가 5건 이상인 데이터에 대해 필터링하여 시각화한 결과이다. 각 그래프에서 x축은 실제 신고 건수, y축은 모델이 예측한 신고 건수를 의미하며 y=x 선에 해당하는 빨간 점선에 가까울수록 예측이 실제에 근접했음을 의미한다.

모델은 전체적으로 신고 건수가 1~5건인 구간에서 예측값이 실제 값과 유사하게 분포하고 있으며, 이 구간에서는 비교적 안정적이고 정확한 예측을 수행하는 것으로 나타났다. 반면 실제 신고 건수가 5건 이상인 데이터의 경우, 예측값이 다소 낮게 산출되는 경향이 있으며, 평균적으로 5건 수준에서 예측이 이루어졌다.
특히 실제 신고건수가 10건 이상인 일부 사례에 대해서도 예측값이 5건 이상으로 산출된

경우가 관측되었으며, 이는 모델이 제한된 학습 정보 속에서도 일부 고(高)신고일 패턴을 학습했음을 시사한다. 고(高)신고일 예측 정확도는 비교적 낮으나, 일부 조건에서 의미 있는 반응을 보였다는 점은 모델의 실용 가능성을 뒷받침한다.

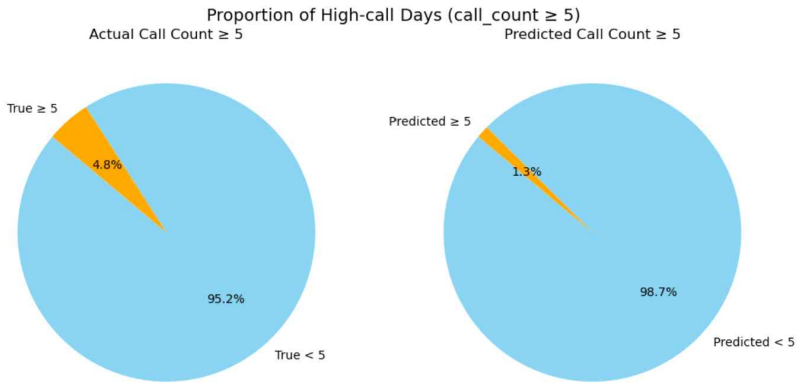



[그림 1] 실제 신고 건수와 예측 신고 건수 산점도

## 4.2 결론

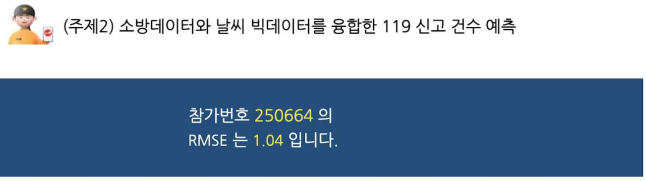
본 모델은 저(低)신고일 구간에 대해 안정적인 예측 성능을 보였으며, 일부 고(高)신고일 조건에 대해서도 평균에 근접한 보수적인 예측 반응을 나타냈다. 실제 신고 건수가 10건을 초과하는 극단적 사례에 대해서는 전체 학습 데이터에서의 비중이 매우 낮고, 기상 조건도 불규칙하게 분포되어 있어 예측이 평균값에 수렴하는 경향을 보였다.

아래 그림 2는 2020~2023년도 중 무작위 추출한 검증 데이터셋에서 신고건수가 5건 이상에 해당하는 실제 비율과 예측 비율을 각각 시각화한 것이다. 실제 고신고일은 약 4.8%를 차지하였으나, 모델은 이 중 약 1.3%만을 5건 이상으로 예측하였다. 이는 고신고일 발생 조건에 대해 모델이 전반적으로 보수적인 예측을 수행하고 있음을 보여준다. 그러나 앞서 그림 1에서 확인한 바와 같이, 예측값이 5건 이상으로 출력되는 상당 수가 실제 고(高)신고일에 해당하는 날과 일치하였으며, 이는 모델이 단순 회귀를 넘어서 일정 수준의 위험 탐지 기능을 갖추고 있음을 의미한다. 이러한 결과는 자원 배치나 사전 경고 체계와 같은 실무적 응용에서 유용하게 활용될 수 있으며, 향후 외부 요인을 추가로 반영할 경우 고(高)신고일 예측의 정확도 향상도 기대할 수 있다.




[그림 2] 고신고일 실제 발생 비율과 모델 예측 비율 비교

## 4.3 최종 검증 결과




[그림 3] 테스트 결과

2024년도 call_count 예측 최종 검증 결과, RMSE = 1.04 이다. 2020년부터 2023년까지의 데이터를 통한 검증 성능이 RMSE = 1.2767이었던 것으로 미루어보아 이는 내부 검증 성능보다 더 우수한 결과로, 모델이 실 데이터에 대해 일반화 성능을 안정적으로 유지하고 있음을 보여준다.

# 5. 활용 방안 및 기대효과

본 모델을 통해 예측된 119 신고 건수는 소방 업무 전반에 걸쳐 다양하게 활용되어, 궁극적으로 국민안전 강화와 소방 자원의 효율적 운용에 크게 기여할 것이다. 구체적인 활용 방안과 기대 효과는 다음과 같다.

| | |
|---|---|
| 인력 및 장비의 선제적 배치 | 예측한 신고 건수를 바탕으로 소방 인력을 미리 증원하거나 교대 근무를 조정하여 비상 상황에 즉각적으로 대응할 수 있으며, 이는 신고 폭주로 인한 현장 도착 지연을 최소화하고 골든타임 확보율을 높이는 데 결정적인 역할을 할 것이다. 또한, 폭우(침수 장비)나 폭염(구급차, 냉방 장비)과 같이 예상되는 재난 유형에 맞춰 필요한 소방 장비와 물자를 사전에 배치하고 점검함으로써, 재난 발생 시 신속하고 효과적인 초동 대응이 가능해진다. |
| 재난 대응 전략 수립 및 교육 훈련 | 예측된 패턴은 다양한 재난 시나리오 개발의 기반이 되며, 이를 통해 소방관들의 현장 대응 훈련을 강화하여 실제 상황 발생 시 대응 능력을 향상시킬 수 있다. 장기적인 관점에서는 예측 모델의 결과를 활용하여 소방 예산을 더욱 효율적으로 배분하고, 소방서 신설, 장비 보강, 인력 충원 등 중장기 계획을 수립하는 데 과학적인 근거를 제공할 수 있다. |
| 유관기관과의 협력 강화 | 예측 정보를 경찰, 지자체 등 유관기관과 공유함으로써, 재난 발생 시 통합적인 대응 체계를 구축하고 각 기관의 역할 분담을 명확히 하여 혼란을 줄이고 시너지를 창출할 수 있다. |
| 신고 종류 별 예측으로의 확장성 | 이 모델을 기반으로 단순히 신고 건수 뿐만아니라 신고 종류별 신고 건수 예측으로 확장하여 사고 및 피해 발생의 범주를 예측해 병의원, 경찰, 지자체 등과 공조하여 사고 예방 및 신속한 대응에 기여할 수 있는 잠재적 확장성을 내재하고 있다. |

[표3] 활용 방안 및 기대 효과

궁극적으로 본 모델은 기존의 수동적인 사후 대응 방식에서 벗어나, 선제적이고 능동적인 재난 관리를 가능하게 하며, 제한된 소방 자원으로 최대의 인명 및 재산 피해 절감 효과를 이끌어낼 것으로 기대된다.