대한민국 지진 규모 예측 모델

이름: 진용은

학번: 2318044

Github: https://github.com/jinyongeun/Korea_earthquake_magnitude_prediction_model

1. 안전 관련 머신러닝 모델 개발 관련 요약

a. 프로젝트에 관한 전체 내용을 요약

2007 년부터 2023 년까지 대한민국에서 발생한 지진을 분석하여, 지진의 규모를 사전에 예측하는 머신러닝 모델이다. 이를 통해 자연재해로 인한 위험을 사전에 평가하고, 효과적인 재난 대비 계획을 수립하는 데 기여하고자 한다.

2. 개발 목적

a. 머신러닝 모델 활용 대상: 지진 관측 기관, 재난 대비 관련 정부/민간 기관

b. 개발의 의의: 지진 발생 규모를 사전에 예측하여 재난 대응 능력을 강화한고, 지진이 자주 발생하는 위험 지역에 대한 우선적인 자원 배분 및 효과적인 재난 복구 지원을 가능하게 할 것이다.

c. 데이터의 어떠한 독립 변수를 사용하여 어떠한 종속 변수를 예측하는지

위도(Latitude), 경도(Longitude), 깊이(Depth), 연도(Year)를 독립변수로 사용하며, 규모(Magnitude)를 종속변수로 설정하여 규모를 예측한다.

3. 배경지식

a. 데이터 관련 사회 문제 설명

지진은 예측이 어려운 대표적인 자연재해 중 하나이다. 규모와 빈도는 인구 밀집 지역과 주요 인프라에 치명적인 영향을 미칠 수 있으므로, 이를 효과적으로 예측하는 것이 필수적이다.

b. 머신러닝 모델 관련 설명 등

머신러닝 기술은 대규모 데이터를 학습하고 복잡한 패턴을 식별하는 데 매우 효과적이다. 본 프로젝트에서는 랜덤 포레스트(Random Forest) 모델을 활용하여 지진 규모를 예측한다. 랜덤 포레스트는 여러 의사결정 나무를 결합한 앙상블 기법으로, 높은 예측 정확도와 안정성을 제공하는 모델이다.

4. 개발 내용

a. 데이터에 대한 구체적 설명 및 시각화

데이터는 총 44 개의 관측값으로 구성되어 있으며, 주요 속성으로는 Date, Latitude, Longitude, Magnitude 이다.

데이터를 시각화한 결과는 다음과 같다. 규모 분포 그래프는 지진의 규모가 주로 4 에서 5 사이에 집중되어 있음을 보여준다. 이는 대부분의 지진이 경미한 수준임을 나타내며, 드물게 더 큰 규모의 지진도 발생한다. 상관관계 히트맵은 Latitude, Longitude 와 Magnitude 간의 상관관계가 약하다는 것을 보여준다. 이는 지리적 위치와 지진 규모 간에 명확한 상관관계가 없음을 시사한다. 지도에서는 지진 발생 위치와 규모를 시각적으로 확인할 수 있다. 지도 상의 원 크기는 지진 규모에 비례하며, 위험도가 높은 지역을 한눈에 확인할 수 있다.

b. 데이터에 대한 설명 이후, 어떤 것을 예측하고자 하는지 구체적으로 설명

본 프로젝트는 지진의 규모를 사전에 예측하는 것을 목표로 한다. 본 프로젝트는 위도(Latitude), 경도(Longitude), 깊이(Depth), 연도(Year)를 독립변수로 사용하며, 규모(Magnitude)를 종속변수로 설정하였다.

c. 머신러닝 모델 선정 이유

랜덤 포레스트 회귀(RandomForestRegressor)는 비선형 데이터와 상호작용 효과를 잘 학습할 수 있기 때문에 주요 모델로 선정했다. 이 모델은 다양한 데이터 특징을 고려하며 과적합을 방지할 수 있는 장점이 있다. 비교 모델로 선형 회귀(Linear Regression)를 사용하였으며, 이는 간단한 선형 관계를 가정하여 랜덤 포레스트와의 성능 차이를 비교하기 위해 선정했다.

d. 사용할 성능 지표

MSE(Mean Squared Error)는 예측값과 실제값 간의 평균 제곱 오차를 측정하며, 값이 낮을수록 모델의 예측이 정확함을 의미한다. R2 Score 는 모델이 데이터를 얼마나 잘 설명하는지를 나타내며, 1 에 가까울수록 좋은 성능을 의미한다.

5. 개발 결과

a. 성능 지표에 따른 머신러닝 모델 성능 평가

랜덤 포레스트 회귀 모델의 MSE 는 0.49 이고, R2 Score 는 0.03 이다. 이는 랜덤 포레스트 모델이 지진 데이터를 처리하는 데 있어 제한적인 성능을

보였음을 나타낸다. 모델이 데이터 내 비선형 패턴을 충분히 학습하지 못했거나, 독립변수가 종속변수를 충분히 설명하지 못했을 가능성이 있다.

랜덤 포레스트 모델의 예측 결과를 시각화한 실제값 대 예측값 산점도에서는 일부 데이터가 실제값에서 크게 벗어나는 경향을 보인다.

b. 머신러닝 모델의 성능 결과에 대한 해석

랜덤 포레스트 회귀 모델은 지진 규모 예측에서 제한적인 성능을 보였다. 모델 성능이 낮았던 원인으로는 독립변수의 부족, 데이터의 크기가 작음, 그리고 지진 규모와 위치 간의 약한 상관관계 등이 있다. 이를 해결하기 위해 추가적인 데이터 수집과 변수를 고려한 확장된 모델링이 필요하다.

6. 결론

a. 머신러닝 모델 개발에 관한 간략한 요약 및 결과 설명

본 프로젝트에서는 지진 데이터를 활용하여 규모를 예측하는 머신러닝 모델을 개발하였다. 랜덤 포레스트 회귀 모델이 주요 모델로 사용되었으나, 제한적인 예측 성능을 보였다.

b. 개발 의의 등

지진 규모 예측 모델은 자연재해에 대한 사전 대비 및 대책 마련에 활용될 수 있다. 특히, 지리적 정보와 결합된 분석은 특정 지역의 위험도를 평가하는 데 기여할 수 있다.

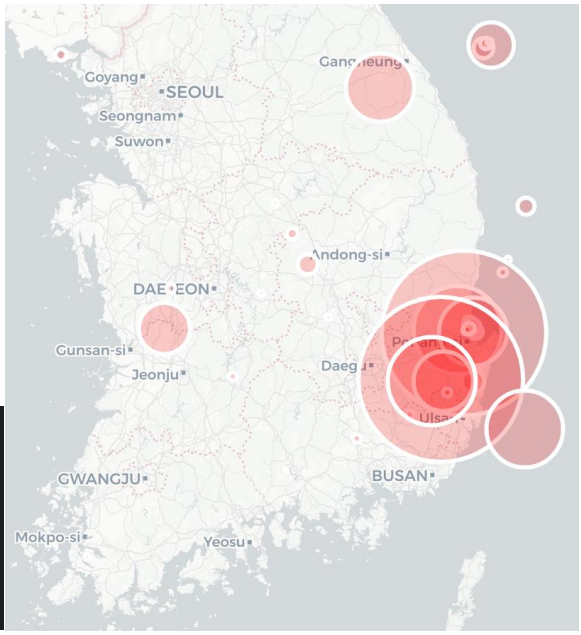
c. 머신러닝 모델의 한계

위치 정보 외에 다른 독립변수가 포함되지 않아 모델의 예측 정확도에 제한이 있었다.

d. 개발 과정 및 후기

지진에 대한 시각화 자료로 지도가 알맞을 것 같아서 지도를 만드는 코드를 작성했고, 지진이 난 장소를 규모에 따라 다르게 표시하기 위해 코드를 추가로 작성하였다.

```python
# 지도 생성
map = folium.Map(location=[37, 127], zoom_start=7, tiles="cartodb positron")
for i in range(names['Date'].count()):
    folium.Circle(
        location=[names['Latitude'][i], names['Longitude'][i]],
        radius=names['Magnitude'][i] ** 6.5,
        color="white",
        fill_color="Red"
    ).add_to(map)
```



처음에 위도와 경도, 연도를 독립 변수를 사용했지만, 정확도가 낮을 것이라고 생각해서 깊이를 추가 했다.

```python
features = names[["Latitude", "Longitude"]].copy()
features['Year'] = names['Date'].dt.year
target = names['Magnitude']
features = names[["Latitude", "Longitude", "Depth"]].copy()
features['Year'] = names['Date'].dt.year
target = names['Magnitude']
```

아직 머신러닝에 대한 이해도가 높지 않아서 개발하는 과정에서 여러 어려운 상황이 발생했다. 하지만 오류가 생기거나 혼자 해결하기 어려운 일이 발생했을 때, AI 를 이용하여 해결할 수 있었다. AI 도 머신러닝의 한 분야인 딥러닝으로 개발된 것이라 머신러닝에 대한 관심이 커졌다. 또한 머신러닝 모델을 개발하면서 데이터셋들을 분석하고 이를 통해 예측하는 것이 신기했다.

머신러닝 모델 성능 평가 점수가 낮아서 아쉬웠다. 더 많은 양 데이터가 있었더라면 머신러닝 모델의 정확도가 올랐을 것이라고 생각한다.