머신러닝 결과 보고서

1.0 주요 기능

1.1 머신러닝 주요 기능

- 사용자가 지도 클릭을 하면 위경도를 받아와서 기후/지형/위성 데이터를 분석하여 산불 피해 면적과 확산 방향, 확산 거리/속도를 예측한다. (혹은 DB에 있는 데이터를 계산하여 예측)

- 산불 위험도를 수치화해 실시간 평가 가능하다.

- 복합 데이터 기반 중요한 변수 자동 분석한다.

- 예측 불확실성(신뢰 구간) 제공으로 대응력 강화한다.

1.2 기대효과 및 활용 방안

- 피해 가능성이 높은 지역과 시간대를 예측해 인력, 장비를 효율적으로 배치하여 대응 효율을 극대화 시킬 수 있다.

- 신뢰 구간까지 반영한 예측 결과로, 위험 수준에 맞춰 대응 우선순위를 합리적으로 결정하기 때문에 긴급 대응 속도를 향상 시킬 수 있다.

- 장기 데이터와 결합해 산불 발생 경향 변화를 분석, 미래 재난 대비 전략 수립에 활용 가능하다. 즉, 기후 변화와 산불 패턴 분석에 기여할 수 있다.

2.0 데이터 설명

2.1 공공 DB 출처 및 데이터 수집 방식

- NASA POWER API : 기후 데이터 수집
    - 산불 발생날 기준 하루 전, 후로 72시간 기후 데이터 추출
- GEE(Google Earth Engine) : 지형 산림 데이터 수집
    - 고도, 경사, 방위각 (NASA/NASADEM_HGT/001) : 입력받은 위도/경도 좌표를 중심으로 5x5 픽셀 크기의 정사각형 영역의 고도, 경사, 방위각 데이터를 추출
    - 식생 NDVI (MODIS/006/MCD12Q1): 입력한 날짜를 기준으로 전후 기간동안의 평균 NDVI 값을 계산 후, 데이터를 바탕으로 이미지를 생성한 후, 해당 이미지에서 입력된 좌표의 NDVI 값을 추출

- 산림률 (UMD/hansen/global_forest_change_2023_v1_11): 기준 연도인 2000년의 수목 피복률 데이터를 확보하고, 이후 매년 산림이 감소한 위치를 기록한 데이터를 추출하여 입력된 좌표 주변 영역의 평균 산림률을 계산

- 산림청: 과거 강원도 산불 데이터 수집

  - 2011년부터 2024년까지의 강원도 모든 개별 산불 데이터 수집

2.2 주요 데이터 항목

1. 기상 데이터 (NASA Power API)

온도/습도 관련

- T2M: 2미터 높이 기온 (°C)

- RH2M: 2미터 상대습도 (%)

- PS: 지표면 기압 (kPa)

- ALLSKY_SFC_SW_DWN: 지표면 태양복사량 (MJ/m^2/day)

바람 관련

- WS2M/WS10M: 2m/10m 높이 풍속 (m/s)

- WD2M/WD10M: 2m/10m 높이 풍향 (도)

강수량 관련

- PRECTOTCORR: 시간당 강수량 (mm)

- total_precip_Nd_start: N일간 총 강수량

- dry_days_Nd_start: N일간 건조일 수 (강수량 < 1mm)

- consecutive_dry_days_start: 연속 건조일 수

2. FWI 지수 (Forest Fire Weather Index)

수분 코드

- FFMC (Fine Fuel Moisture Code): 세부 연료 수분 코드

- DMC (Duff Moisture Code): 부식토 수분 코드

- DC (Drought Code): 가뭄 코드

위험 지수

- ISI (Initial Spread Index): 초기 확산 지수

- BUI (Buildup Index): 축적 지수

- FWI (Fire Weather Index): 종합 화재 위험 지수

3. 지형 데이터 (Google Earth Engine)

고도 관련

- elevation_mean/std/min/max: 평균/표준편차/최소/최대 고도 (m)

경사 관련

- slope_mean/std/min/max: 평균/표준편차/최소/최대 경사도 (도)

사면 방향 관련

- aspect_mode/std: 주사면방향/표준편차 (도)

- aspect_north_ratio: 북향 사면 비율 (315-45도)

- aspect_south_ratio: 남향 사면 비율 (135-225도)

식생 관련

- ndvi_before: 화재 발생 전 정규화식생지수

- treecover_pre_fire_5x5: 화재 전 산림피복률

- ndvi_stress: NDVI 스트레스 지수

4. 시계열 피처 (0h~171h)

시간별 변화량

- t2m_change_X_Yh: X시간~Y시간 간 온도 변화량

- ws10m_change_X_Yh: X시간~Y시간 간 풍속 변화량

- rh2m_change_X_Yh: X시간~Y시간 간 습도 변화량

통계 피처

- T2M_mean/std/max/min: 온도 통계량

- WS10M_mean/std/max/min: 풍속 통계량

- RH2M_mean/std/max/min: 습도 통계량

5. 파생 피처 (Feature Engineering)

조합 지수

- dryness_index: 건조도 지수 = T2M × (100-RH2M)/100
- wind_humidity_ratio: 풍속/습도 비율
- wind_temp_product: 풍속×온도 곱
- dry_to_rain_ratio_30d: 30일 건조/강수일 비율

위험 플래그

- hot_dry_combo: 고온건조 조합 (T2M>30°C & RH2M<30%)
- high_wind_flag: 강풍 플래그 (WS10M>10m/s)
- low_humidity_flag: 저습도 플래그 (RH2M<30%)
- extreme_hot_flag: 극고온 플래그 (T2M>35°C)

지형 효과

- slope_south_combo: 남향 급경사 조합
- south_steep_effect: 남향 경사 효과
- terrain_var_effect: 지형 변동성 효과

6. 시공간 피처

날짜/시간

- startyear/startmonth/startday: 시작 연/월/일
- fire_month: 화재 발생 월
- is_spring/summer/autumn/winter: 계절 더미 변수

위치

- lat/lng: 위도/경도
- region_name: 지역명

7. 예측 목표 변수

- 면적 모델: 산불 확산 면적 (ha)
- 속도 모델: 확산 속도 클래스 (저속/중속/고속)

- 방향 모델: 확산 방향 클래스 (8방위)

이 피처들은 563개(속도), 447개(방향), 51개(면적) 규모로 각 모델별로 최적화되어 있다.

3. 데이터 전처리

3.1 결측치 처리, 이상치 제거 또는 수정

- 결측치 : 데이터 수집 과정(API, GEE)에서 일부 데이터가 누락되어 결측치(NaN)가 발생 또한 산불의 데이터의 심한 클래스 불균형이 심함 (대형 산불 데이터가 극심하게 적음)

- 해결:
    - 대체값을 사용: NASA API에서 데이터를 가져올 때, 특정 매개변수(예: 10M 풍속/풍향)가 누락되면 2M 매개변수(예: 2M 풍속/풍향)로 대체
    - 이전 값 채우기 : 시간별 데이터를 처리할 때, 이전 시간대의 유효한 값으로 현재 시간대의 결측치를 채움.
    - NaN을 0으로 간주: np.nansum이나 np.nanmean과 같은 함수를 사용하여 강수량이나 FWI 지수 계산 시, NaN 값을 0으로 간주하거나 계산에서 제외하여 결측치로 인한 오류를 방지
    - 핵심 컬럼의 결측치 제거: 모델 학습에 필수적인 타겟 변수나 핵심 피처에 결측치가 있는 행을 제거
    - 남은 결측치 0으로 채우기: 피처 스케일링 전에 모든 피처 DataFrame의 남은 결측치를 0으로 채움. 이는 모델이 NaN 값을 처리하지 못하는 것을 방지하고, 결측치가 특정 의미(예: 0)를 가질 수 있다고 가정하는 방식을 사용

- 이상치: 클래스 불균형으로 인해 중요 변수 이상치 발생

- 해결:
    - 필터링: 피해 면적을 예측하는 모델을 학습 시킬 때 극단적인 이상치를 제거함. 피해 면적이 0이거나 상위 1%에 해당하는 매우 큰 값(이상치)을 가진 데이터를 학습에서 제외시킴.
    - 스케일링: 평균/표준편차 대신 중앙값/사분위수 범위를 사용으로 극단적인 이상치가 스케일링 과정에 미치는 영향을 최소화 시킴.

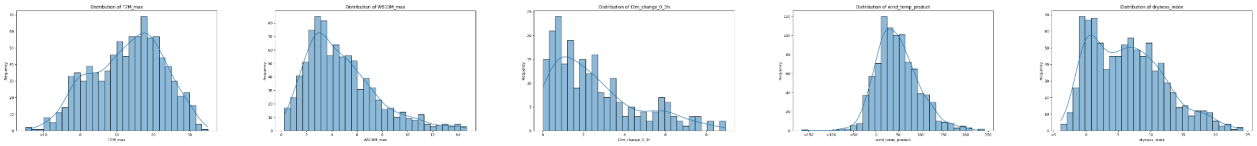
3.2 범주형 변수 분포

1. 정수 인코딩:
    - 타겟 변수인 확산 속도 클래스(spread_speed_class)와 확산 방향 클래스(spread_direction_class)는 각각 0, 1, 2 또는 0-7과 같은 정수 값으로 인코딩

2. 이진 플래그 생성:
    - is_spring, is_summer, is_autumn, is_winter와 같이 특정 범주(예: 계절)에 해당하는지 여부를 나타내는 0 또는 1의 이진 피처를 생성

3. 트리 기반 모델 특성 활용

3.3 변수 분포 및 시각화

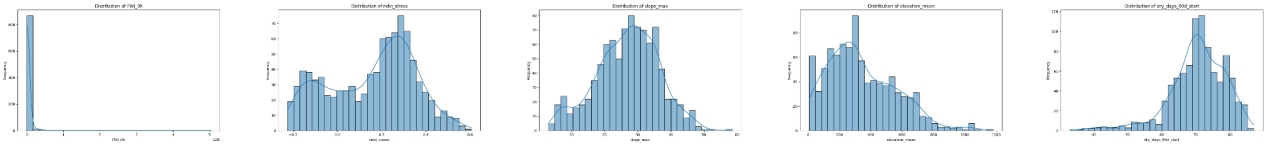


먼저 산불 피해면적 예측 모델의 변수 분포도이다. 사용 변수가 많아 가장 중요한 5개를 사진으로 나타냈다.
왼쪽부터 차례로 T2M_max(최고 기온), WS10M_max(최대 풍속), t2m_change_0_3h(3시간동안의 온도 변화), wind_temp_product(바람과 온도의 상호작용), dryness_index(건조도 지수) 이다.

T2M_max와 wind_temp_product는 정규 분포를 보인다. 즉 안정적인 예측이 가능하다고 해석된다.

WS10M_max와 t2m_change_0_3h, dryness_index는 우편향을 보이며 극값이 중요하다고 할 수 있다.

이를 통해 피해 면적 예측 모델은 기후 변수와 급격한 변화에 더 민감하다고 할 수 있다.



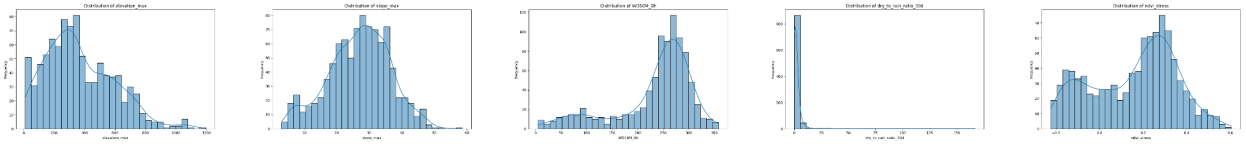
두번째는 방향 예측 모델의 변수 분포도이다. 마찬가지로 사용 변수가 많아 가장 중요한 5개를 사진으로 나타냈다.
왼쪽부터 차례로 FWI_0h(산불 기상 지수), ndvi_stress(식생 스트레스), slope_max(최대 경사도), elevation mean(평균 고도), dry_days_90d _start(90일 건조일수) 이다.

Slope_max와 dry_days_90d_start는 정규 분포를 보이며 안정적인 예측이 가능해보인다.

FWI_0h와 elevation_mean은 각각 극우편향과 우편향을 보이면서 임계점이 존재하며 극값이 중요하다고 해석된다.

ndvi_stress는 이봉분포로 보이며, 두가지 상태로 구분되는 모습을 보인다.

방향 모델은 식생 스트레스와 지형 특성이 핵심이라고 해석된다.



마지막은 속도 예측 모델 변수 분포도이다. 왼쪽부터 차례로 elevation_max(최대 고도), slope_max(최대 경사도), WD10M_0h(바람 방향), dry_to_rain_ratio_30d(30일 건조/ 강수비), ndvi_stress(식생 스트레스) 이다.

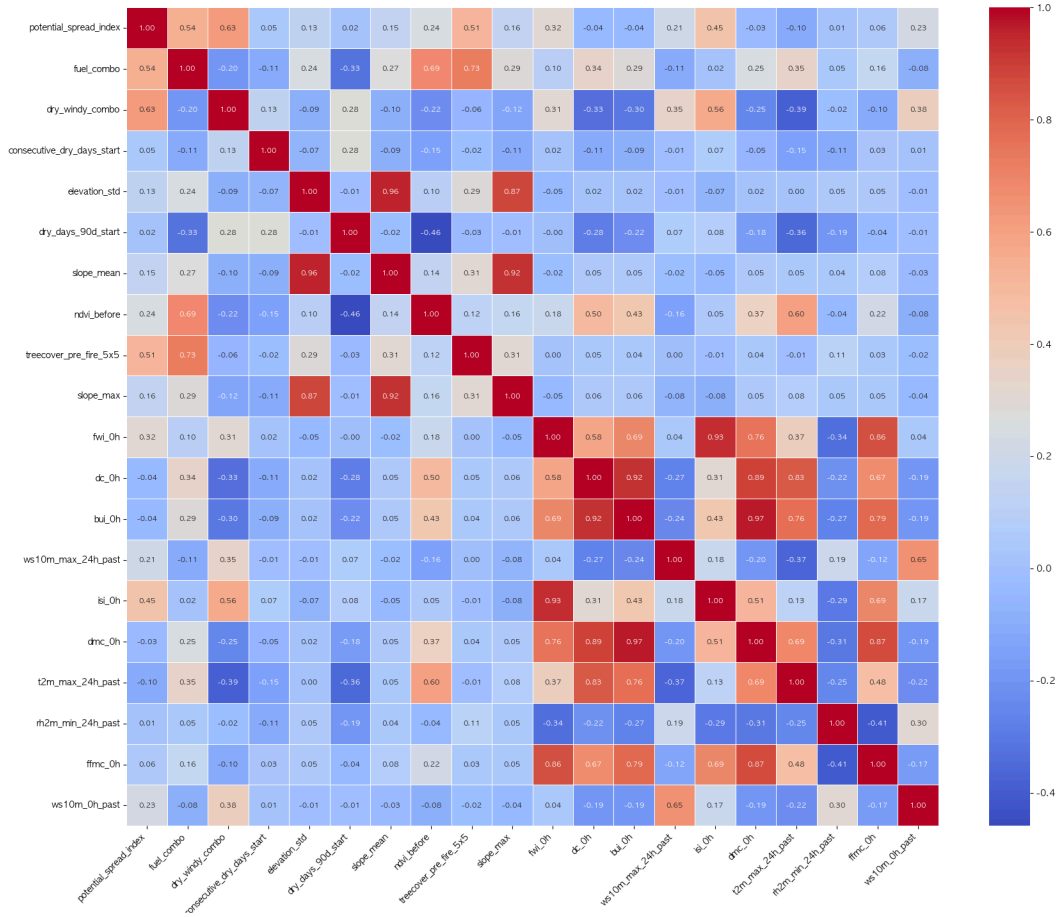
Slope_max와 WD10M_0h는 정규 분포로 안정적인 예측이 가능해보인다.

Elevation_max와 dry_to_rain_ratio_30d는 각각 우편향과 극우편향을 보여 극값이 중요하고 임계점이 존재하는 것으로 보여진다.

Ndvi_stress는 마찬가지로 이봉분포이며, 두가지 상태로 구분되는 모습을 보였다.

속도 모델은 지형 변수가 주도하되 기후 조건이 예측값을 조절해주는 것으로 해석된다.

3.4 변수간 상관관계 분석



FWI 관련 지수들 간의 강한 관계

- fwi_0h(산불기상지수)는 isi_0h (초기확산지수)와 0.93으로 매우 강한 양의 관계를 보인다. 이는 산불의 초기 확산 위험이 높을수록 전체적인 산불 위험도가 높아진다는 논리적인 관계를 보여준다.

- bui_0h (축적건조지수)는 dmc_0h (부식질수분지수) 및 dc_0h (가뭄지수)와 매우 강한 양의 관계(각각 0.97, 0.92)를 보인다. 이는 토양과 낙엽층이 건조할수록(가뭄이 심할수록) 축적된 건조 위험이 높아진다는 것을 의미한다.

건조도와 산불 위험

- ffmc_0h (실효습도지수)는 dmc_0h와 0.87의 강한 관계를 보인다. 이는 표면의 낙엽 등이 건조할수록, 그 아래 토양층도 건조할 가능성이 높다는 것을 의미한다.

주목할 만한 음의 상관관계

- dry_windy_combo (건조하고 바람부는 조건)와 dc_0h (가뭄지수)는 -0.33으로 약한 음의 관계를 보인다. 이는 단기적인 바람이나 건조함보다는, 장기적인 가뭄이 산불 위험에 다른 방식으로 영향을 줄 수 있음을 의미한다.

지형/식생과의 관계

- potential_spread_index (잠재 확산 지수)는 fuel_combo (연료 결합 지수), dry_windy_combo와 강한 양의 관계(각각 0.54, 0.63)를 보인다. 이는 연료가 많고 건조하며 바람이 불 때 잠재적인 확산 위험이 커진다는 것을 명확히 보여준다.

4. 모델링 기법 및 이유

4.1 적용한 머신러닝 알고리즘 목록 및 선택 이유

모든 머신러닝 예측 모델(피해면적, 확산 방향, 확산 속도)은 랜덤 포레스트(Random Forest)기반으로 만들어졌다.

피해 면적 예측 모델은 랜덤 포레스트 회귀(RandomForestRegressor) 를 이용하여 만들어졌고, 확산 방향과 확산 속도 모델은 랜덤 포레스트 분류(RandomForestClassifier)를 이용하여 만들어졌다.

랜덤 포레스트를 사용한 이유

1. 높은 정확도와 안정성:

- 랜덤 포레스트는 수많은 Decision Tree를 만들어 그 결과를 종합하는 '앙상블' 기법을 사용하여 단일 모델보다 훨씬 더 정확하고 안정적인 예측 성능을 보여줌.

2. 다재다능:

- 하나의 알고리즘으로 숫자 값을 예측하는 회귀(Regressor)와 카테고리를 예측하는 분류(Classifier) 작업을 모두 수행할 수 있어, 현재 프로젝트의 모든 예측 모델을 일관성 있게 구축할 수 있다.

3. Overfitting 방지:

- 여러 개의 나무가 서로 다른 데이터를 학습하고 그 결과를 평균 내므로, 학습 데이터에만 너무 치우쳐 새로운 데이터에 약해지는 오버피팅 현상을 효과적으로 방지한다.

4. 비선형 관계 학습 능력:

- 기온이 특정 지점을 넘으면 산불 위험이 급격히 커지는 것처럼, 변수와 결과 간의 복잡하고 비선형적인 관계를 잘 학습할 수 있다.

5. 피처 중요도 제공:

- 어떤 피처(변수)가 예측에 더 큰 영향을 미쳤는지 알려주는 '피처 중요도'를 계산해 준다. 이는 모델의 예측 결과를 해석하고, 어떤 변수가 중요한지 이해하는 데 도움이 된다.

4.2 교차 검증**(Cross-Validation)** 방식

K-Fold Cross-Validation

1. 분할 방식: 5-겹 교차 검증 (5-Fold Cross-Validation)

- 전체 데이터를 무작위로 5개의 그룹으로 나눈다.

2. 학습 및 검증 과정:

- 첫 번째 그룹을 검증(validation)용으로 사용하고, 나머지 4개 그룹을 학습(training)용으로 사용하여 모델을 학습하고 성능을 평가한다.

- 두 번째 그룹을 검증용으로, 나머지 4개 그룹을 학습용으로 사용하여 두 번째 평가를 진행한다.

- 이 과정을 총 5번 반복하여, 모든 그룹이 한 번씩 검증용으로 사용되도록 한다.

3. 최종 성능:

- 5번의 평가에서 나온 성능 점수들의 평균을 계산하여 모델의 최종 성능을 측정한다.

4. 기타 설정:

- shuffle=True: 데이터를 5개로 나누기 전에 무작위로 섞어, 데이터가 특정 순서로 정렬되어 있을 때 발생할 수 있는 편향을 방지한다.

- random_state=42: 무작위로 섞을 때 항상 동일한 방식으로 섞이도록 하여, 코드를 다시 실행해도 항상 같은 결과가 나오도록 보장한다.

5. 모델학습 결과 및 성능 평가

5.1 평가 지표 선정

1. 피해 면적 예측 (회귀 모델)

- RMSE (Root Mean Squared Error, 평균 제곱근 오차)

- 선택 이유: 모델의 예측값과 실제값의 차이를 측정하는 가장 대표적인 지표. 실제값과 차이가 큰 예측(큰 오차)에 대해 더 강한 패널티를 부여하는 특징이 있음. 따라서 모델이 매우 동떨어진 예측을 하는 것을 방지하고, 전반적으로 안정적인 예측을 하도록 유도하는 데 효과적임. GridSearchCV에서 최적의 모델을 찾는 핵심 기준으로 사용되었다.

- 해석: 값이 낮을수록 모델의 예측이 정확하다는 의미이다.

- R^2 (R-squared, 결정 계수)

- 선택 이유: 모델이 데이터의 변동성을 얼마나 잘 설명하는지를 직관적인 비율(0~1 사이)로 보여주는 지표. 모델의 설명력을 종합적으로 판단하기에 좋은 지표.

- 해석: 값이 1에 가까울수록 모델이 데이터를 잘 설명한다는 의미이다.

2. 확산 속도/방향 예측 (분류 모델)

- F1-Score (가중 평균)

- 선택 이유: 정확도(Accuracy)가 가질 수 있는 함정(예: 데이터가 불균형할 때 한쪽으로만 예측해도 높게 나옴)을 피하기 위해 사용. F1-Score는 정밀도(Precision)와 재현율(Recall)의 조화 평균으로, 두 지표를 모두 균형 있게 고려한다. 특히 weighted F1-score는 각 클래스의 샘플 수에 따라 가중치를 부여하여 평균을 내므로, 데이터 불균형이 있더라도 모델의 성능을 공정하게 평가할 수 있다. GridSearchCV에서 최적의 모델을 찾는 핵심 기준으로 사용되었다.

- 해석: 값이 1에 가까울수록 분류 성능이 좋다는 의미이다.

- Classification Report (분류 보고서)

- 모델의 최종 성능을 종합적으로 확인하기 위해 사용.

- Accuracy (정확도): 전체 예측 중 올바르게 예측한 비율. 가장 직관적이다.

- Precision (정밀도): 모델 예측의 신뢰도를 나타낸다.

- Recall (재현율): 모델이 얼마나 빠짐없이 잘 찾아내는지를 나타낸다.

5.2 모델 별 성능 비교 및 시각화

1. 피해 면적 예측 모델(회귀)

- R^2 Score**:** 0.8877

- RMSE: 0.3395

해석: 모델이 피해 면적의 변동성을 약 89% 설명할 수 있다. 하지만 여전히 약 11% 변동성은 설명하지 못 한다고 해석된다.

2. 확산 속도 예측 모델(분류)

- 정밀도(Precision): 0.99

- 재현율(Recall): 0.99

- F1-Score: 0.99

해석: 확산 속도 예측 모델은 거의 완벽한 분류 성능을 보인다. 오탐/누락이 거의 없고 실제 산불 확산 속도 카테고리를 잘 분류하고 있다고 해석할 수 있다.

3. 확산 방향 예측 모델(분류)

- 정밀도(Precision): 0.97

- 재현율(Recall):0.96

- F1-Score: 0.96

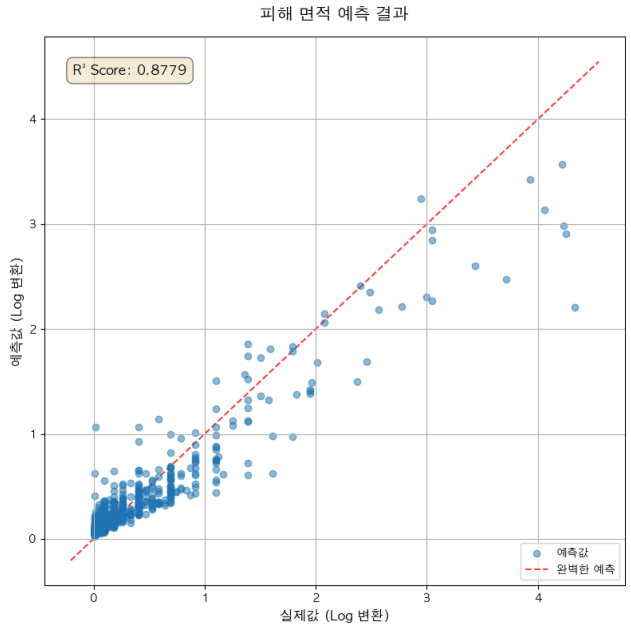
해석: 방향 예측 모델도 매우 우수하다. 다만 속도 모델 보다는 점수가 약간 낮다. 이는 풍향은 기상과 지형, 복합적인 요인에 따라 변동성이 크기 때문에 예측하기 어려운 부분이 존재한다는 것을 알 수 있다.

추가 설명: 확산 속도와 확산 방향 예측의 높은 점수로 인해 데이터 누수를 의심해 보았지만 이미 풍향과 관련된 피처(wd10_xh)관련 피처는 다 제거하고 학습을 시켰기 때문에 데이터 누수는 없었다. 풍속과 관련된 피처(ws10_xh)는 모델이 풍속, fwi, 기후, 지형 등의 복합적인 관계를 학습하여 최종 확산 방향/속도는 어떻게 되는지를 판단할때 필요한 중요한 피처 중 하나이기 때문에 제거하지 않았다.

6. 예측 결과 및 향후 개선 방향

6.1 실제 값과 예측 값 비교 및 해석

1. 피해면적 예측 결과

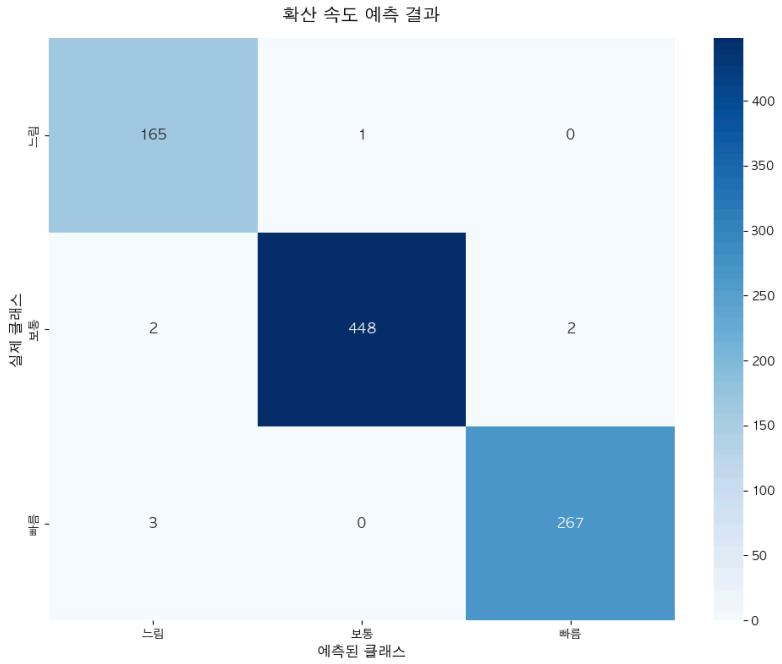


- R^2 Score: 0.8779

해석: 모델이 피해 면적의 변동성을 약 88% 설명할 수 있다고 해석된다. 산점도를 보면 대부분의 점들이 붉은 선과 가깝게 있어 대체로 실제 값과 유사하게 예측하고 있다는 것을 알 수 있다. 그러나

오른쪽으로 갈 수록 규모가 큰 산불의 경우 예측값이 실제 값보다 낮게 나온다. 이는 모델이 매우 큰 규모의 산불은 과소 예측하는 경향이 있다고 볼 수 있다.

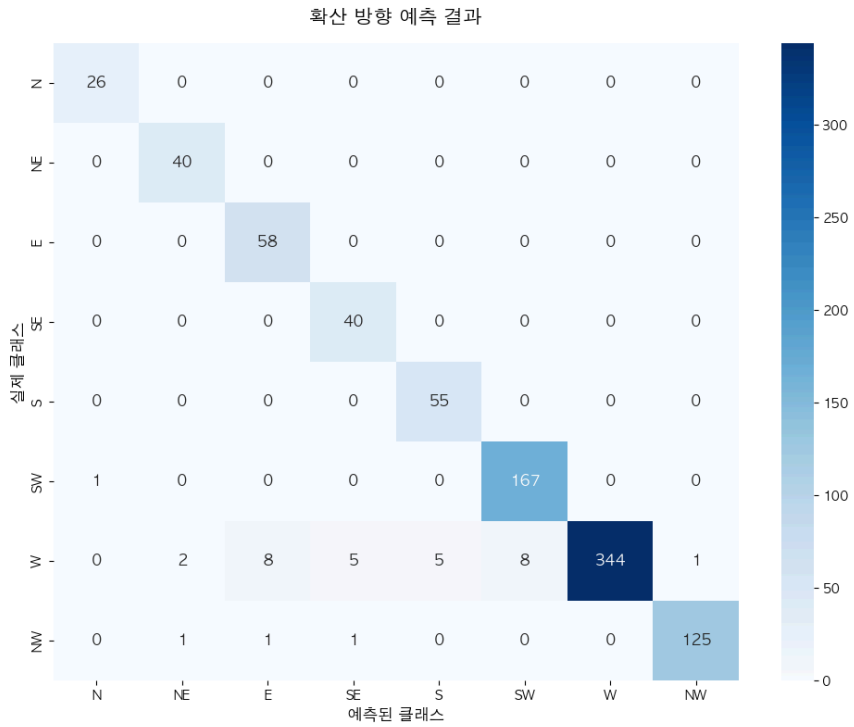
2. 확산 속도 예측 결과



- 전반적인 성능: 대각선(파란색이 진한칸)에 대부분의 값이 집중 되어 있다.

해석: 매우 높은 예측 성능을 보인다. 크게 눈에 띄는 오류는 없다.

3. 확산 방향 예측 모델



- 전반적인 성능: 속도 모델과 마찬가지로 대각선에 값이 압도적으로 많아 전반적으로 뛰어난 예측 성능을 보여준다.

해석: 사진에서 보이는 가장 큰 오류는 실제 방향이 W(서쪽)인 산불 8개를 E(동쪽)으로, 8개를 SW(남서)로, 5개는 SE(남동)과 S(남쪽)으로 예측한 것이다. 이는 모델이 서풍 계열의 바람이 불때 지형이나 다른 요인으로 인해 확산 방향이 미세하게 남쪽 혹은 동쪽으로 치우치는 경우를 실제보다 더 많이 예측하는 경우가 있다고 해석된다. 하지만 틀린 예측의 수가 맞은 예측의 수에 비해 매우 적어, 확산 방향 예측 모델의 신뢰도는 매우 높다고 할 수 있다.

6.2 향후 개선 방향

1. 초대형 산불 데이터 추가:

- 현재 산불 피해 면적 예측 모델은 대형 산불의 경우 피해면적을 과소 예측하는 경향을 보인다. 이는 앞서 말했다시피 산불 데이터 클래스의 불균형 문제로 실제 발생했던 초대형 산불 데이터를 더 많이 수집하여 학습 시키면 극단적인 상황에 대한 예측 정확도가 높아질 것이라고 예상된다.

2. 딥러닝 모델 도입:

- 시계열 모델을 도입하여 시간에 따라 변화하는 산불의 동적인 특성을 더 잘 모델링할 수 있도록 한다.

- 이미지 기반 모델(CNN)을 도입하여 위성 사진과 지도를 이미지 데이터로 간주하여 이미지 자체를 학습 시키는 딥러닝 모델을 통해서 공간적인 확산 패턴을 예측하는 방식으로 발전시킬 수 있다.