빼앗긴 산에도 봄은 오는가 :

과거 산불 데이터를 기반으로 한 사유림 보험 가입비 책정 제안

삼정KPMG Future Academy 4기

마.파.두.부 (마운틴 **파**이어 **두**목과 **부**하들)

김다은 | 문수현 | 박서형 | 이혁재

목 차

1. 프로젝트 개요

1-1. 주제 선정 배경

봄, 가을에 집중 발생하던 산불이 지구 온난화와 도시화 등의 이유로 계절에 관계 없이 다수 발생하는 산불 연중화 현상이 심화되고 있다. 문제는, 산불로 사유지 피해를 입더라도 산불은 국가가 운영하는 풍수해보험 보장 대상에 포함되지 않아 제대로 된 보상을 받을 수 없다는 것이다. 국내 전체 산림 중 약 70%가 개인이 소유한 사유림이며 산의 소유주인 산주 인구는 매년 증가세에 있다. 산림 피해 보장 보험 상품이 부재한 상황 속에서, 국내 산주 220만 명을 타겟으로 한 새로운 민간 보험 상품을 제시한다.

1-2. 프로젝트 목적

과거 산불 데이터를 기반으로 시기별 발생 빈도, 기후 조건, 산불 다발 지역, 피해 규모 등을 분석하여 사유림 피해 보장 보험 가입비 책정을 위한 인사이트를 얻는다. 고객의 가입 희망 시기, 사유림 소재지, 사유림 면적 등을 파악하여 보험 가입비를 가늠해본다.

2. 프로젝트 진행 절차

2-1. 데이터 선정

산림청에서 배포한 산불 발생 데이터 분석. 2015년 1월 부터 2025년 2월까지 산불 발생 지역, 발생 시점, 진화 완료 시점, 발생 원인, 피해 면적 등의 데이터를 선택하였다. 이후, 산불과 기상 조건의 상관관계 파악을 위해 기상청 일일 데이터와 병합하였다.

2-2. 데이터 전처리

- 필요성: 산림청 데이터는 화재 발생 시점과 진화 종료 시점이 일치하지 않는 등 수기로 입력됨이 추정되는 자료이다. 이에 따라 관측소 별로 상이했던 시간 데이터 형식을 통일하였다. 더불어, 산불이 발생하기 쉬운 기상 여건을 파악하기 위하여 산불이 나지 않은 날과 산불 발생일의 기후 조건을 수집하였다. 이 과정에서, 기상청에서 배포한 관측소별 일일 기온, 풍속, 습도 데이터를 지역 별로 정렬 후 병합하는 전처리 작업을 수행하였다.

- 과정:
    1) NaN값 제거
    2) 개명된 지역명은 혼돈 방지를 위해 제거(특정 기간 자료 부재)
    3) 진화 종료 시간에서 화재 발생 시간을 차감하여 이상치 수정
    4) 날짜와 지역을 기준으로 데이터 병합
    5) 피해 규모 2헥타르 미만을 '초기 진압 성공'으로 규정하고 작업 수행

2-3. 데이터 강화

1) 데이터를 이용한 화재 위험지역 분류 → 위험지역에 따른 보험 전략 강화
2) 화재 발생 기간 전체 데이터 확보, 전처리 → 화재 자체에 대한 예측력 강화

2-4. EDA

산불 발생 빈도 순위, 피해 면적 순위, 주요 발생 원인, 산불 확산에 악영향을 끼치는 기후 조건 등을 파악하기 위한 탐색적 데이터 분석 수행

2-5. Machine Learning

1) 데이터 인코딩
2) 타겟 및 머신러닝 목적 설정
    a) 분류
        - 산불 발생 가능성
          Random Forest Accuracy Score : 0.94
        - 위험 산불 발생 가능성
          피해면적, 진압시간 median 값 (0.1)
          이상일시 0 : 위험산불,
          이하일시 1 : 낮은위험산불
          Logistic Regression Accuracy Score : 0.94
          Random Forest Accuracy Score : 0.99
        - 진압시간: (0 : 110분 이하, 1 : 110분 이상)
          Logistic Regression Accuracy Score : 0.69

b) 회귀

- 피해면적

  Linear Regression R^2 : 0.03

  Polynomial Regression R^2 : -3.32

  XGBoost Regressor R^2 : 0.14

- 진압시간

  Linear Regression R^2 : 0.99

  XGBoost Regressor R^2 : 0.43

c) 비지도 학습: 클러스터링

- 낮은 위험 산불 군집화

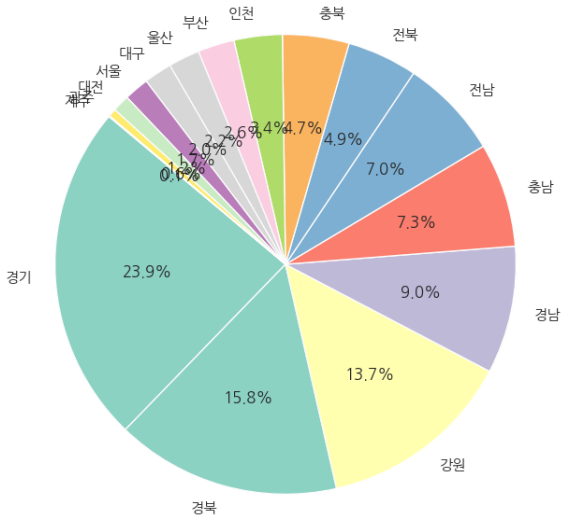
  기존 전처리 및 피처 엔지니어링 (월 별 계절, 진압시간)

  이후 다음과 같은 절차로 군집화를 실행

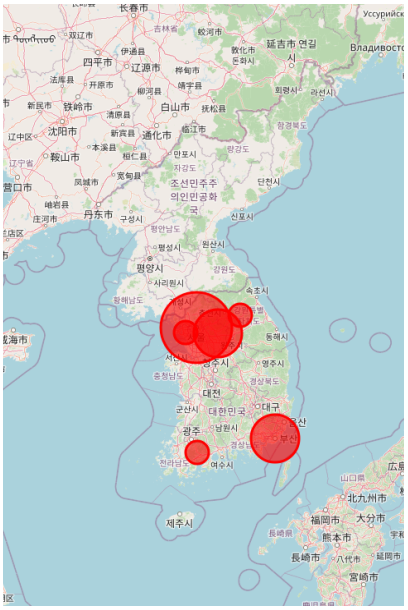
  1) 차원 축소
  2) N-Components: 3-선정
  3) K-Means 알고리즘 선정
  4) Elbow Method로 k = 3 사용
  5) 군집 분리도 측정: Davies-Bouldin Index로 1.9839 확인
  6) 군집 피처 중요도 시각화
  7) 계절에 의한 군집 분포 확인
  8) 계절과 군집 상관관계 측정: Chi Square Test로 P-value 0.0으로 강한 상관관계 확인

3. 데이터 인사이트

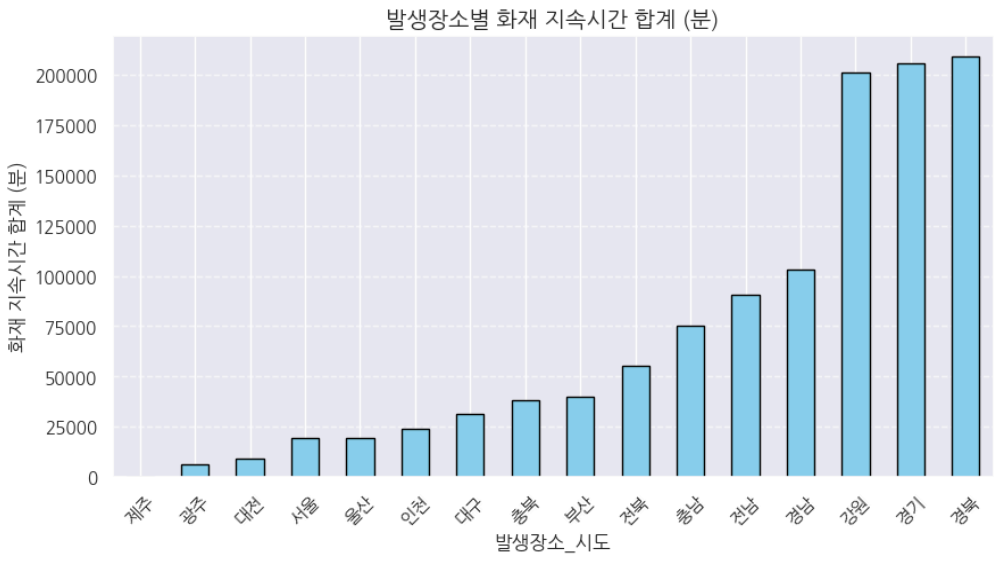
3-1. 산불 발생 장소 빈도, 강도



시도 별 발생 비율 파이 차트

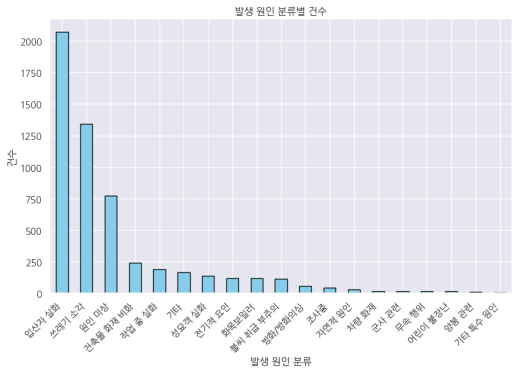


산불 발생 빈도 지도 시각화



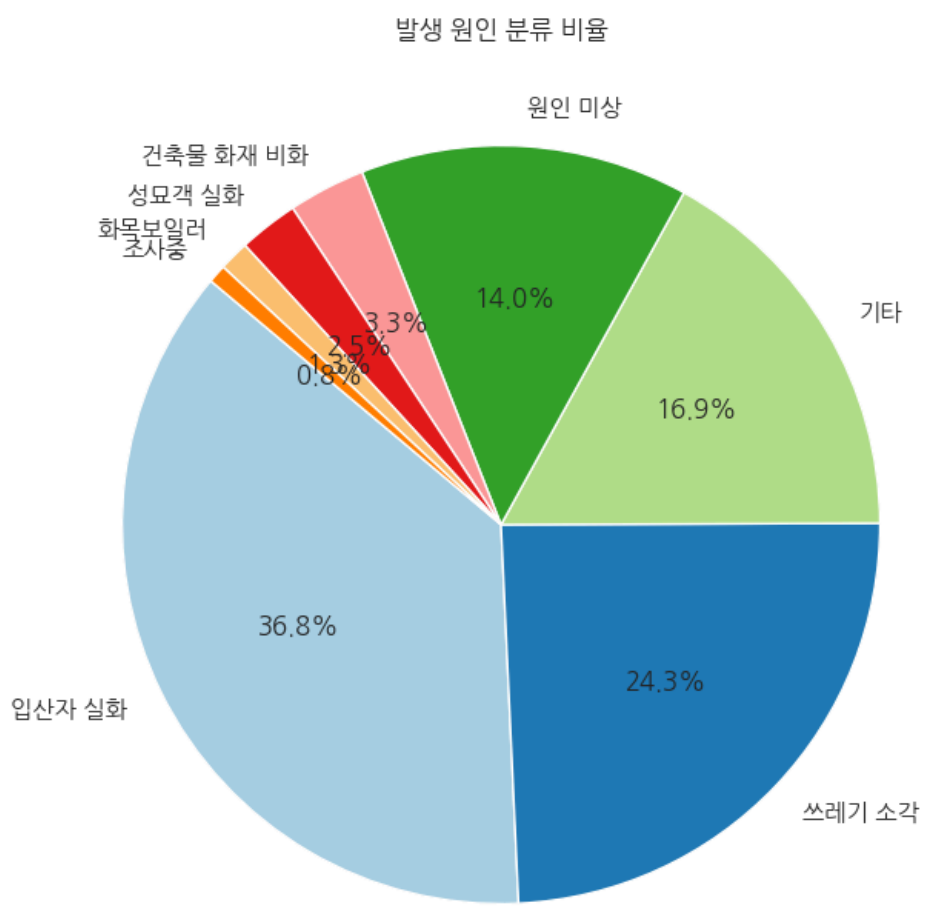
- 발생 빈도 순위: ① 경기 ② 경북 ③ 강원
- 화재 지속시간 순위: ① 경북 ② 경기 ③ 강원

3-2. 산불 발생 원인



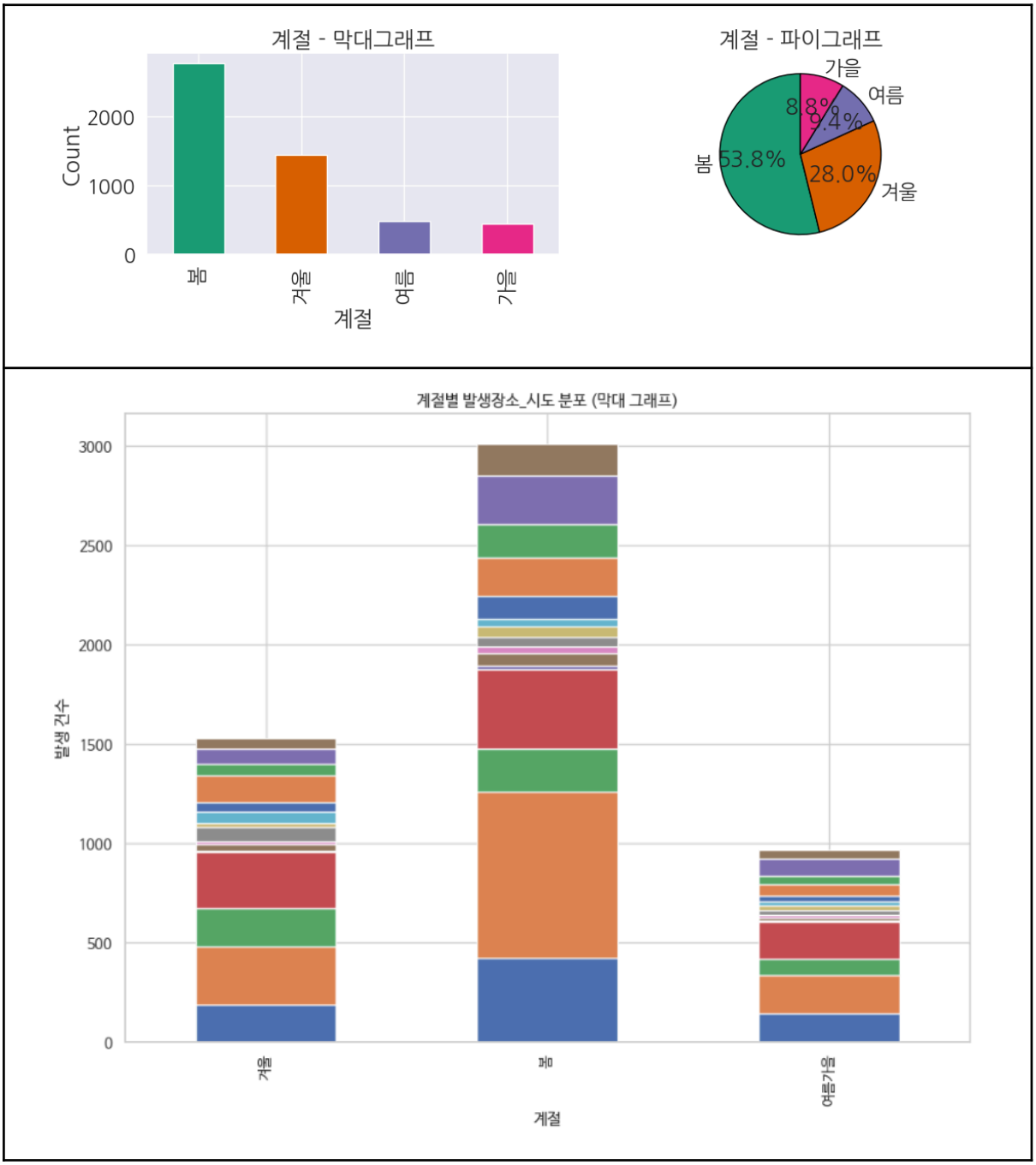
발생 원인별 히스토그램

발생 원인 빈도 분석에 따른 워드클라우드



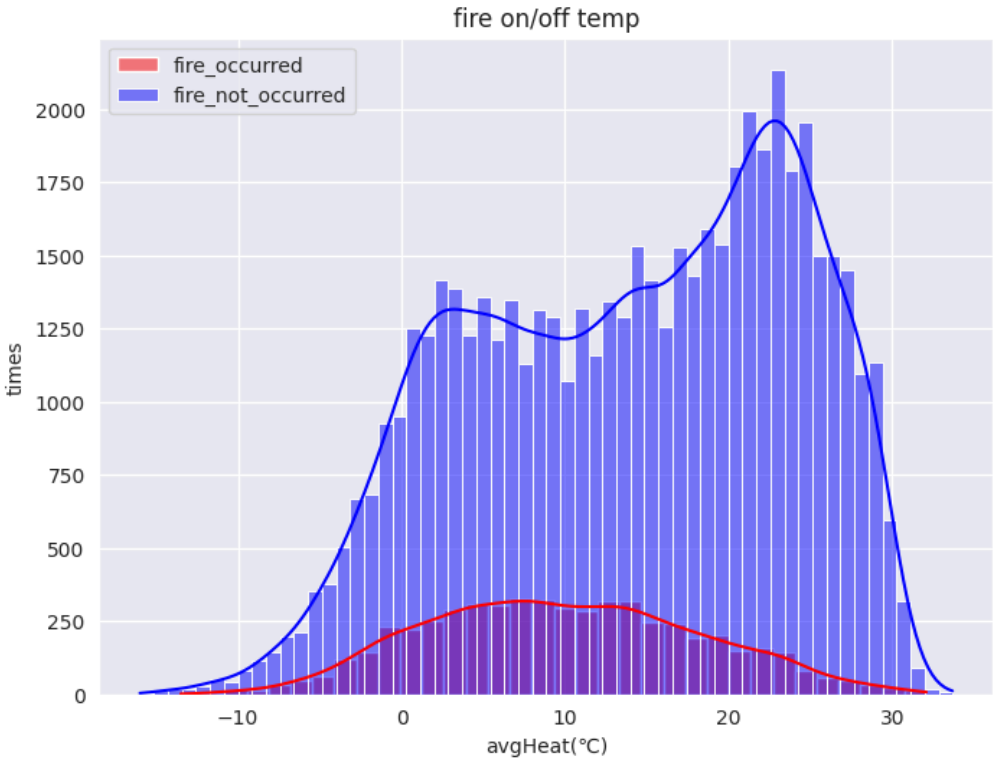
- 발생 원인 순위: ① 입산자 실화 ② 쓰레기 소각 ③ 원인 미상

3-3. 산불 발생 시기

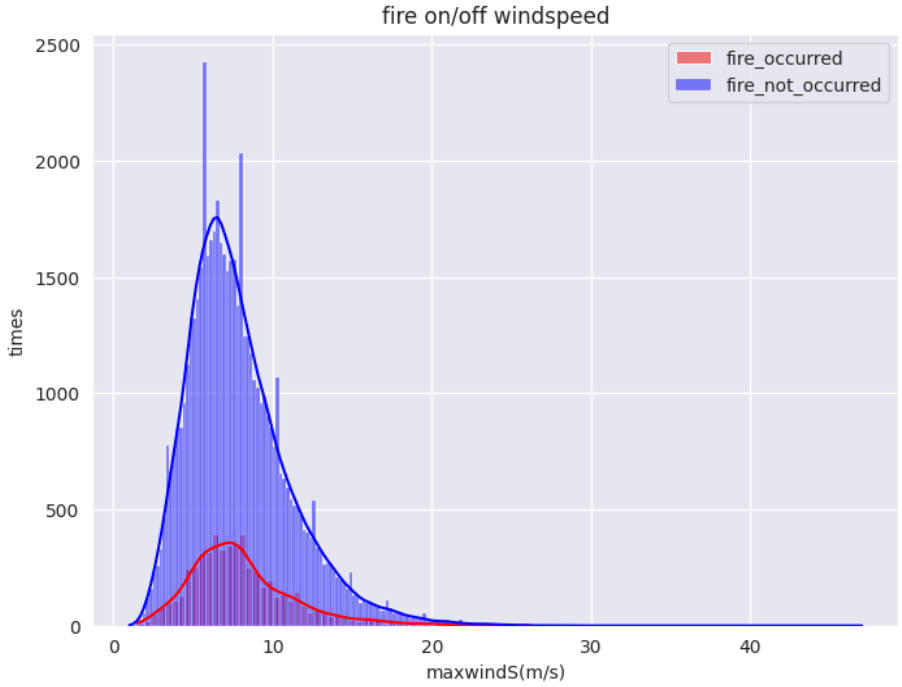


- 발생 시기 순위: ① 봄 ② 겨울 ③ 여름 ④ 가을
- 발생 장소 누적 그래프를 통해 계절 빈도 분포와 순위 재확인

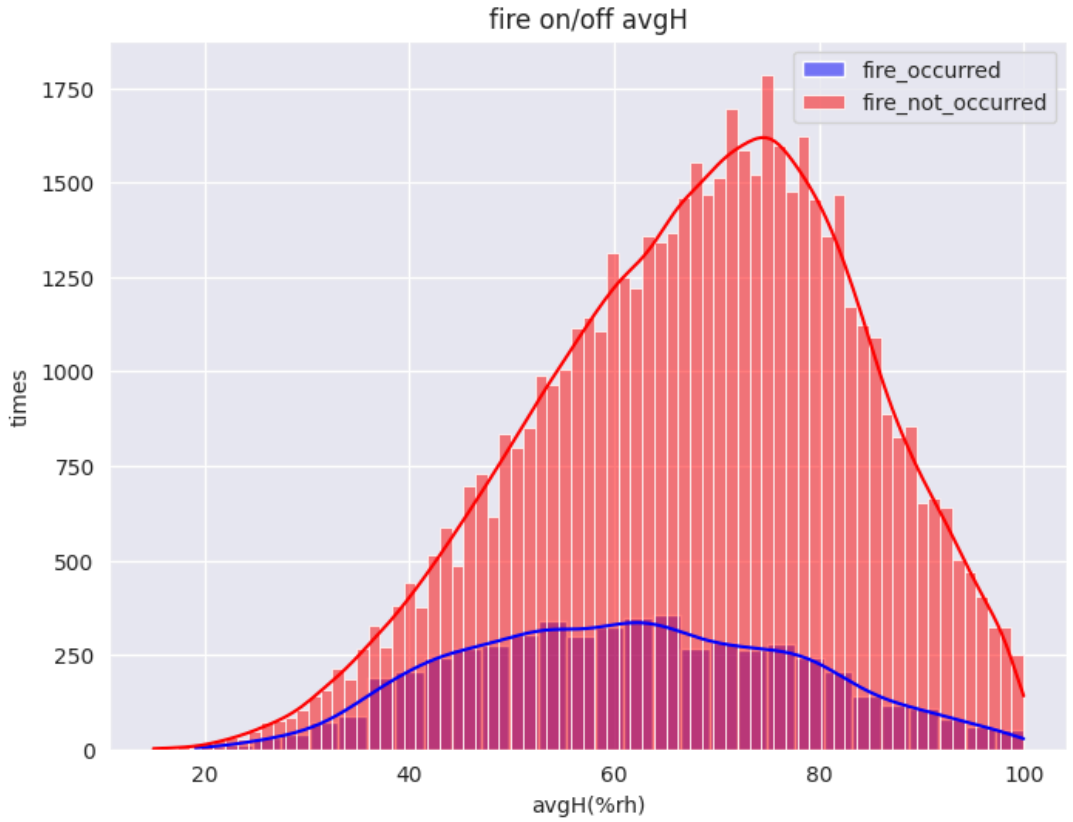
3-4. 산불 발생 기후



화재발생시 기온 분포: 기온이 높은 계절(여름)을 제외한 봄, 겨울에 발생함을 확인



화재발생시 풍속 분포: 산불 확산을 가속화하는 풍속인 6m/s 주위에 많이 분포된 것을 확인



화재발생시 습도 분포: 습도가 아주 낮은 시기(겨울), 습도가 높은 시기(여름)에는 입산객이 비교적 적어 산불 발생 빈도 역시 낮아짐

3-5. 산불 발생 예측

- 기온과 습도를 고려한 계절적 특성 생성1
  - 클러스터 0 (따뜻한 계절): 4월~10월
  - 클러스터 1 (추운 계절): 11~3월
- 기존 모델정확도 0.72 → 0.94, 데이터 및 코드 보완 후 0.94 → 0.95로 향상
- 기온과 습도를 고려한 계절적 특성 생성2
  - 클러스터0,1,2,로 분류
- 장소에 따른 산불 발생 가능성 예측가능

4. 프로젝트 제안

4-1. 실적용 제안

- 민간 보험 상품: 국가 보험에 포함되지 않은 산불 피해를 보장해주는 민간 보험 상품 제작을 통해 사유림 산주들의 산불 피해 보장을 도모한다. 산불 머신러닝 프로젝트 도출 결과를 6,000평 이하 사유림 소지자 대상 산불 보험 상품에 활용한다.
- 가입비 진단 웹서비스: 사유림 보험 가입 희망자의 사유림 소재지와 가입 희망 시기 등을 참고하여 산불 위험군 여부와 예상 가입비를 간단하게 진단해주는 웹 서비스를 생성, 보험 가입을 유도하는 하나의 광고 컨텐츠로도 활용할 수 있다. 유사한 서비스로는 수령 가능한 세금 환급금을 예측하는 삼쩜삼이 있음.

4-2. 가입비 책정 원리

- 가입 희망자의 위험도 계산:
    - 고려 요소:
        - 사유림 소재지(지역): 산불 다발 발생 지역일수록 가중치 부여
        - 사유림 면적 (평)
        - 가입 기간 (봄, 여름/가을, 겨울): 산불 발생 빈도 높을수록 가중치 부여
        - 기상 데이터 (기온, 풍속, 습도): 기후변화로 인해 계절의 기상 및 기간이 상의할 수 있는 가능성을 알고리즘에 고려하여 화재 발생 위험을 계산
- 가입비: 평 당 1,000원 책정
  (e.g) 3,025평(1 헥타르) 소유주일 경우, 가입비 3,025,000원 책정.
  이후 보험사 별로 희망하는 부가 보험료 액수 추가 책정.
- 보험사는 사유림 면적(최대 6000평), 희망 가입 기간(봄, 여름, 가을, 겨울), 위치(시도), 기온, 습도, 풍속 데이터를 활용해 1~100 위험 산불 발생 점수를 산출함. 과거 산불 발생 유사한 데이터 군집을 기반으로 클러스터를 분류하고, 이상 기온 및 지역별 가중치를 반영하여 보험 가입금을 계산함.

4-3. 위험 산불 지수 알고리즘

- 다음과 같은 알고리즘으로 위험산불 지수를 계산:

(1) 클러스터 2 산불위험지수

−[1+exp(−(−277.8672+(0.1147×Tmean)+(0.09195×Area)−(0.02595×Rh)−(0.2069×Wmean)+Σβ시도+Σβ계절))−1]−1- \left[1 + \exp{\left(-\left(-277.8672 + (0.1147 \times Tmean) + (0.09195 \times Area) - (0.02595 \times Rh) - (0.2069 \times Wmean) + \sum \beta_{\text{시도}} + \sum \beta_{\text{계절}} \right) \right)} - 1 \right] - 1−[1+exp(−(−277.8672+(0.1147×Tmean)+(0.09195×Area)−(0.02595×Rh)−(0.2069×Wmean)+Σβ시도+Σβ계절))−1]−1

(2) 클러스터 0 산불위험지수

−[1+exp(−(−369.5906+(0.03904×Tmean)+(0.12045×Area)+(0.00214×Rh)−(0.08024×Wmean)+Σβ시도+Σβ계절))−1]−1- \left[1 + \exp{\left(-\left(-369.5906 + (0.03904 \times Tmean) + (0.12045 \times Area) + (0.00214 \times Rh) - (0.08024 \times Wmean) + \sum \beta_{\text{시도}} + \sum \beta_{\text{계절}} \right) \right)} - 1 \right] - 1−[1+exp(−(−369.5906+(0.03904×Tmean)+(0.12045×Area)+(0.00214×Rh)−(0.08024×Wmean)+Σβ시도+Σβ계절))−1]−1

(3) 클러스터 1 산불위험지수

−[1+exp(−(−237.9762+(0.0929×Tmean)+(0.07847×Area)+(0.02235×Rh)−(0.3851×Wmean)+Σβ시도+Σβ계절))−1]−1- \left[1 + \exp{\left(-\left(-237.9762 + (0.0929 \times Tmean) + (0.07847 \times Area) + (0.02235 \times Rh) - (0.3851 \times Wmean) + \sum \beta_{\text{시도}} + \sum \beta_{\text{계절}} \right) \right)} - 1 \right] - 1−[1+exp(−(−237.9762+(0.0929×Tmean)+(0.07847×Area)+(0.02235×Rh)−(0.3851×Wmean)+Σβ시도+Σβ계절))−1]−1

- 알고리즘으로 산출한 위험지수로 다음과 같은 함수로 보험 가입비를 계산함:
  보험 가입비 =사유림 평수x (1000원) ×(1+ 100 (산불위험지수)×Risk Multiplier)
  - 평당 가입비 1000원으로 고정하여 계산
  - Risk Multiplier 2로 고정하여 계산

5. 개선점

5-1. 문제점

1) 비교 데이터 부재

- 산불이 발생하지 않은 날의 기후 데이터 부재로 기후 조건에 따른 산불 발생 확률 비교 분석이 불가능하며 특성간 상관관계 분석에 어려움을 겪음

→ 피드백 후 해결

2) 데이터간 수치

- 실제 기상학을 연계하여 풍속, 습도, 기온을 연계하여 가중치와 실제 영향의 정도를 정확하게고려해 반영하지 못했음

- 가장 큰 영향을 미치는 요인은 **최대순간풍속(m/s)**이며, 그 다음으로 **평균습도(%rh)**, **평균기온(℃)** 순으로 영향

3) 오픈API

- 오픈API에서 받아온 데이터가 제한적이었음. 보유한 데이터 기간보다 짧아 활용이 불가능한 상황이었음.

3) 용어 통일

- 팀원 간 용어 통일 부재로 커뮤니케이션 애로사항이 있었음

4) 텍스트 전처리 문제

- 텍스트 처리에 SentenceTransformer('jhgan/ko-sbert-sts')사용했으나 불완전

5-2. 해결방안

1) 추가 데이터 수집: 산불이 발생하지 않은 날의 기후 데이터, 산불 진화 자원 데이터, 산림 품종 데이터 등 연관 데이터를 추가 수집하여 머신러닝에 활용 가능한 특성 수 확대
2) 데이터 추가 학습 및 상세화: 각 데이터(습도, 온도, 풍속)에 대해 추가적인 학습과 데이터 상세화 필요. 지역 정보를 확대하여 산불 발생 장소의 고도 데이터를 추가하면 지역적 특색도 반영 가능해짐.
3) 오픈API: 오픈API에 대한 리퀘스트 작업이 성공적으로 수행되었다면 더욱 양질의 머신러닝이 가능해짐.
4) 발달된 모델의 한글 텍스트 전처리 필요

6. 데이터 출처

1. 산림청 알고리즘
   https://www.forest.go.kr/newkfsweb/html/HtmlPage.do?pg=/fgis/UI_KFS_5002_030202.html&orgId=fgis&mn=KFS_02_04_03_05_02
2. 기상청 데이터
   https://data.kma.go.kr/climate/RankState/selectRankStatisticsDivisionList.do
3. 산림청 공공데이터 개방
   https://www.forest.go.kr/kfsweb/kfi/kfs/frfr/selectFrfrStatsNow.do
4. 산림청 웹 서비스 참고 링크
   http://forestfire.nifos.go.kr/menu.action?menuNum=1
5. 낮은 위험의 산불 발생 횟수 기사
   https://www.yna.co.kr/view/AKR20191002057900063
6. 국립산림과학원 (2025.02.24) 산불위험 지수 vs 산불 발생 횟수 비교 그래프
   https://www.ohmynews.com/NWS_Web/View/img_pg.aspx?CNTN_CD=IE003419588
7. API 데이터 사이트 출처 정리 (실패한 API 데이터) 기상청, API 허브
   https://apihub.kma.go.kr/
8. 생활안전정보
   https://safemap.go.kr/opna/data/dataView.do