네이버 본사에서 UFO 관측하기

1. 양태양 (19930120)
2. [tyami@naver.com](mailto:tyami@naver.com)
3. 양태양 (19930120) / 김재훈 ( ) / 김종수 ( )



UFO를 보려면 어떻게 해야할까 블라블라 글을 쓰자

***분석 데이터 설명***

분석에는 Kaggle에 공개된 오픈데이터 중, “미국의 UFO 관측날의 대기 질 데이터 (UFO Sightings + Air Quality)”와 “서울시 대기 질 데이터 (Air pollutants measured in Seoul)”를 사용하였다. 각 데이터셋의 변수 이름과 설명은 아래와 같다.

1. **미국의 UFO 관측날의 대기 질 데이터 (UFO Sightings + Air Quality)**

* <https://www.kaggle.com/infof422henni/ufo-air-quality>
* UFO sightings 데이터 (<https://www.kaggle.com/NUFORC/ufo-sightings>) 중 미국 지역 데이터와 U.S. Pollution Data (<https://www.kaggle.com/sogun3/uspollution>) 를 결합한 데이터셋
* 63,173개의 데이터 샘플, 24개의 변수
  1. State.Code
  2. City
  3. State
  4. Day
  5. Month
  6. Year
  7. Hour
  8. NO2.Mean: The arithmetic mean of concentration of NO2 within a given day
  9. NO2.1st.Max.Value: The maximum value obtained for NO2 concentration in a given day
  10. NO2.1st.Max.Hour: The hour when the maximum NO2 concentration was recorded in a given day
  11. NO2.AQI: The calculated air quality index of NO2 within a given day
  12. O3.Mean
  13. O3.1st.Max.Value
  14. O3.1st.Max.Hour
  15. O3.AQI
  16. SO2.Mean
  17. SO2.1st.Max.Value
  18. SO2.1st.Max.Hour
  19. SO2.AQI
  20. CO.Mean
  21. CO.1st.Max.Value
  22. CO.1st.Max.Hour
  23. CO.AQI
  24. ET: equal to 1 if a sighting has occured and 0 otherwise

1. **서울시 대기 질 데이터 (Air pollutants measured in Seoul)**

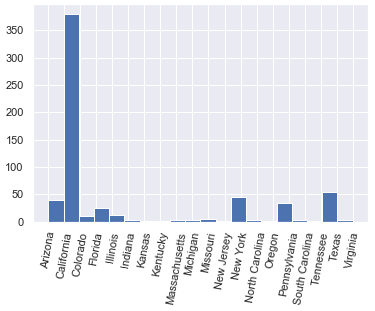
* <https://www.kaggle.com/jihyeseo/seoulairreport>
* 2017년 11월 17일부터 11월 24일까지 1시간 간격으로 측정된 서울시 지역구별 공기 질 데이터셋
* 4,225개의 데이터 샘플, 8개의 변수
  1. 측정일시
  2. 측정소명
  3. 이산화질소 (NO2) 농도 (ppm)
  4. 오존 (O3) 농도 (ppm)
  5. 일산화탄소 (CO) 농도 (ppm)
  6. 아황산가스 (SO2) 농도 (ppm)
  7. 미세먼지 (㎍/㎥)
  8. 초미세먼지 (㎍/㎥)

두 데이터셋간 동일한 데이터를 사용하기 위해 “UFO sightings + Air Quality” 데이터 중 \*.1st.Max, \*.1st.Max, \*.AQI 변수와 “Air pollutants measured in Seoul” 데이터셋의 미세먼지, 초미세먼지 변수를 제거하였으며, 각 데이터셋의 변수 단위를 맞추는 작업을 진행했다.

*UFO가 잘 관측되는 조건은 무엇이 있을까?*

위 질문에 답하기 위해 “UFO sightings + Air Quality” 데이터셋 중 UFO가 관측된 샘플 (ET=1)로부터 크게 다음과 같은 3가지 조건을 추출하여 분석하였다: 지역, 시간, 대기조건

1. **지역: 해안과 내륙 지방 중 어디에서 UFO가 많이 보일까?**
   1. 어느 주 (state)에서 UFO가 많이 관측되었을까?



**Insight:** California 주가 압도적으로 많은 편이며, Texas, New York, Pennsylvania, Arizona 주가 그 뒤를 이었다.

* 1. UFO가 관측된 city 위치를 해안(coast) /내륙(inland) 지역으로 나누어 보자

앞의 결과와 같이 state 별 분석이 가능하다. 하지만 UFO 운전수에게 인간의 행정구역은 크게 중요치 않을 것이다. 따라서 **각 UFO가 관측된 도시(city)의 위치가 해안으로부터 20km 내에 있을 경우를 “coast”, 그 외를 “inland”로 나누었다**. 또한 아래와 같이 간단하게 해안 지역의 면적을 계산하였다.

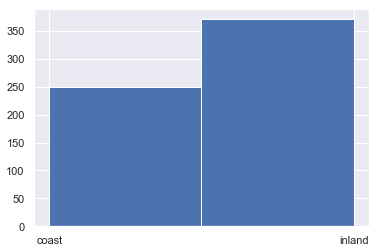
- 알래스카 주를 뺀 미국 본토 국토 총 면적: 8,080,510 km^2

- 해안(coast) 지역 총 면적 = 해안 지역의 총 길이 (9,027km) \* 20km = 180,540 km^2

- 내륙(inland) 지역 총 면적 = 8,080,510 km^2 - 180,540 km^2

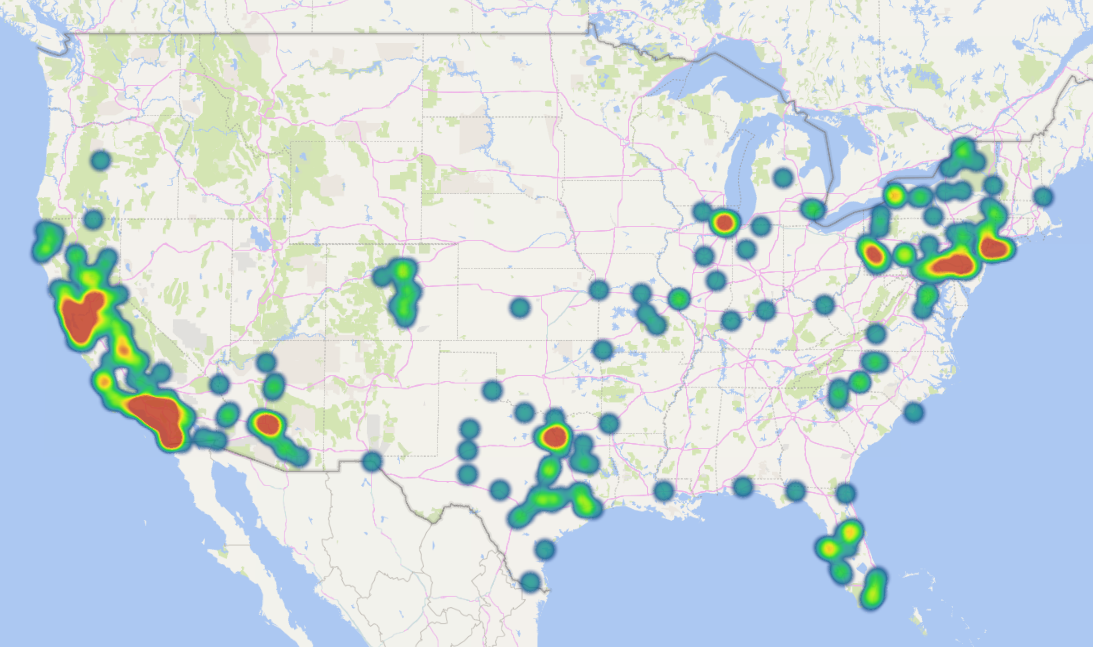
**- 해안(coast) : 내륙 (inland) 면적 비 = 2.23 : 97.77**

* 1. 해안과 내륙 지방 중 어디에서 UFO가 많이 보일까?



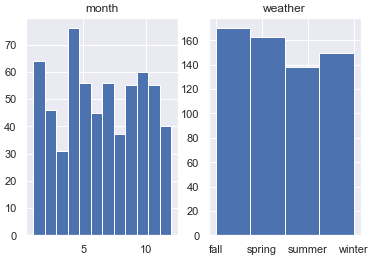
**Insight:** 내륙 지역에서 관측된 경우가 좀 더 많은 편이다. 하지만 미국 전체 국토 중 우리가 정의한 Coast 지역의 면적이 압도적으로 적다는 점을 감안하면 해안 지역에서 UFO가 더 많이 관측되었다고 볼 수 있다.

* 1. 미국 본토 지도 위에 관측위치 분포를 그려보자



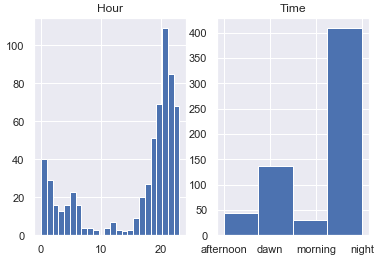
**Insight:** 앞의 결과와 동일하게 대부분의 UFO 관측 보고가 해안과 가까운 지역 (예: 캘리포니아 주)에서 이루어짐을 알 수 있다.

1. **시간: UFO가 많이 보이는 계절, 시간대가 있을까?**
   1. 월 🡪 계절 별 count hist plot



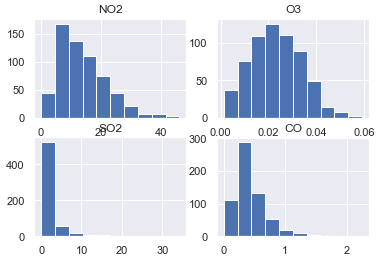
**Insight:** 월별로는 차이가 나기는 하지만, 계절별로는 큰 차이가 없다. 각 계절간 n수가 3개씩뿐 안되기 때문에 별도의 통계분석은 하지 않았다.

* 1. 시간 🡪 시간대 별 count hist plot



**Insight:** 밤 > 새벽 > 오후 > 아침

1. **대기조건**
   1. 각 대기조건별 count hist plot

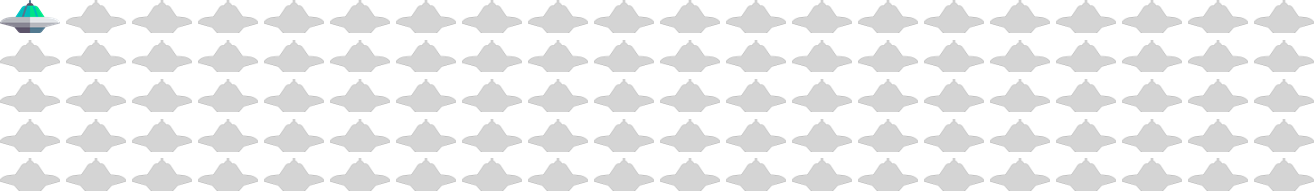


**Insight:**

*지역, 시간, 대기정보 조건을 바탕으로 UFO 관측가능여부를 알 수 있을까?*

위 질문에 답하기 위해 “UFO sightings + Air Quality” 데이터셋으로 예측 모델을 훈련시킨 후, **“**Air pollutants measured in Seoul” 데이터셋에서 UFO 관측 가능여부를 예측해보기로 했다.

1. **Imbalanced dataset 문제 해결**
   1. “UFO sightings + Air Quality” 데이터셋의 문제점



“UFO sightings + Air Quality” 데이터셋은 **UFO 관측 시의 샘플(ET=1)이 622개로 전체 데이터 63,174개의 1%가 채 되지 않는 imbalanced dataset (Imbalance Ratio, IR = 101.57)이다**. Decision tree, Logistic regression 등 대부분의 분류 알고리즘은 imbalanced dataset을 훈련시킬 경우 샘플 수가 많은 쪽에 편향된다.

* 1. Imbalanced dataset 해결 방안 {1명}
  + Resampling

Imbalanced dataset을 처리하는 가장 간단한 방법은, 다수 그룹(ET=0)의 데이터샘플을 undersampling하거나 소수 그룹(ET=1)의 데이터샘플을 oversampling하 하는 것이다. 이 외에도 SMOTE (Synthetic Minority Over-sampling Technique)를 비롯한 복잡한 알고리즘들이 있다. Undersampling과 oversampling은 각각 데이터에 있는 중요한 정보가 소실되거나, 훈련 데이터 중복으로 과적합(Overfitting)될 가능성이 존재한다. 따라서 **이번 분석에서는 random undersampling, random oversampling 및 SMOTE resampling을 모두 시행하여 결과를 비교해보았다**.

* + Class weight balancing

Imbalanced dataset을 처리하는 두번째 방법은 모델 훈련 시 각 class 분류에 weight를 주는 방법이다. 즉, 소수 그룹에 해당하는 샘플을 분류한 결과를 더 중요하게 여긴다. 우리가 사용한**데이터셋의 IR이 너무 높아 resampling 방법만으로는 imbalanced dataset 문제가 해결되기 어렵다고 판단하여 class weight balancing 또한 함께 사용하였다.**

1. **[Training] 모델을 훈련시켜보자 {2명}**
   1. UFO 관측조건 예측 모델 선택

**UFO 관측 예측을 위한 모델로** **Support Vector Machine, Decision Tree Classifier, Random Forest Classifier**를 선정하였다. 위 모델들은 Python 언어의 Scikit-learn package로 간단히 구현할 수 있으며, class weight balancing을 파라미터로 간단하게 설정할 수 있어 선정되었다.

* 1. UFO 관측조건 예측 모델 훈련

일단 Hold-out 으로 훈련, 테스트 데이터 split.

10-CV로 바꿀 예정.

* 1. UFO 관측조건 예측 모델 평가방법

정확도 (Accuracy)는 accuracy paradox 문제로 Imbalanced dataset에 적합한 지표가 아니다. 따라서 **모델을 평가하는 지표로 AUC (Area Under ROC Curve)와 F1 score를 함께 사용했다**. 두 지표는 아래와 같은 Confusion matrix에서 계산될 수 있다. 우리 분석에서는 Scikit-learn package의 roc\_auc\_score, f1\_score 함수를 이용하여 계산하였다.

|  |  |  |
| --- | --- | --- |
|  | UFO 관측조건으로 예측 | UFO 관측조건 X로 예측 |
| 실제 UFO 관측조건 | *True Positive (TP)* | *False Negative (FN)* |
| 실제 UFO 관측조건 X | *False Positive (FP)* | *True Negative (TN)* |

* + AUC

AUC는 특이도(Specificity)와 민감도(Sensitivity)의 산술평균으로 정의되며, 0.5에 가까울수록 모델의 성능이 없고 1에 가까울수록 최고의 성능을 내는 것을 의미한다.

* + F1 score

F1 score는 정밀도(Precision)와 재현율(Recall)의 조화평균으로 정의되며, 큰 값일수록 모델의 성능이 좋은 것으로 평가한다.

* 1. UFO 관측조건 예측 모델 성능 비교

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Resampler | Classifier | Accuracy | AUC | F1 score |
| *Random under sampling* | *Support Vector Machine* | 0.731 | 0.801 | 0.552 |
| *Random under sampling* | *Decision Tree Classifier* | 0.768 | 0.699 | 0.541 |
| *Random under sampling* | *Random Forest Classifier* | 0.806 | 0.815 | 0.498 |
| *Random over sampling* | *Support Vector Machine* | 0.745 | 0.805 | **0.142** |
| *Random over sampling* | *Decision Tree Classifier* | **0.987** | **0.937** | **0.812** |
| *Random over sampling* | *Random Forest Classifier* | **0.994** | **0.959** | **0.897** |
| *SMOTE* | *Support Vector Machine* | 0.741 | 0.834 | **0.162** |
| *SMOTE* | *Decision Tree Classifier* | **0.977** | 0.787 | 0.623 |
| *SMOTE* | *Random Forest Classifier* | **0.983** | **0.920** | 0.647 |

* 1. 앙상블 모델 (가능하면)

Sampling 방법 앙상블?

SVM + DT + RF 앙상블?

1. **[Test 1]**

“UFO sightings + Air Quality” 데이터셋으로 예측 모델을 훈련시킨 후, **“**Air pollutants measured in Seoul” 데이터셋에서 UFO 관측 가능여부를 예측해보기로 했다. 하지만 서울시 대기 데이터의 경우 UFO가 관측되었는지는 알 수 없기에 정확도를 계산하는 것은 불가능하다.

**###내용###**

1. **[Test] 네이버 위치로 모델을 테스트해보자**

우리가 본선에 진출한다면 합숙하게 될 장소에서 UFO를 보자

###내용###

본선 날짜에 해당하는 이전 연도들의 대기 데이터 획득해서 돌려보기? 간단한 확률 정도 내볼 수 있을 듯.