네이버 커넥트원에서 UFO를 볼 수 있을까?

1. 양태양 (19930120)
2. tyyang@unist.ac.kr
3. 양태양 (19930120) / 김재훈 (19941213) / 김종수 (19941219)

"사람들의 발길이 잘 닿지 않는 숲 속, 밤 안개와 더불어 무척 스산한 분위기……"

사람들이 기억하는 영화 속 대부분의 UFO 목격 장면은 위와 같다고 할 수 있다. '어두운 밤', '숲 속', '안개' 등과 같은 조건들은 우리 눈에 잘 목격되지 않는 미확인 비행 물체를 보기위한 필수 조건일 것일까? 나아가서 어떤 환경조건을 만족시켜야 그들을 만나기 쉬워질까? 실제 사람들이 마주쳤었던 그 당시의 시간, 위치, 환경 정보들을 바탕으로 UFO 를 마주칠 확률을 도출해 낼 수 있을지 궁금하여 분석을 시작하게 되었다.

과연 SF 영화 속 항상 등장하는 배경은 합리적인 것일까?

분석 데이터 설명

분석에는 'Kaggle'에 공개된 오픈데이터 중, "미국의 UFO 관측날의 대기 성분 데이터 (UFO Sightings + Air Quality)"와 "서울시 대기 성분 데이터 (Air pollutants measured in Seoul)"를 사용하였다. 각 데이터셋의 변수 이름과 설명은 아래와 같다.

1. 미국의 UFO 관측날의 대기 성분 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

1.1 State. Code: US EPA(미국 환경보호청)에 등록된 각 주의 인식 코드

1.2 City: 도시 명

1.3 State: 미국 총 50개의 주

1.4 Day: 당시 날짜

1.5 Month: 당시 월

1.6 Year: 당시 년도(2000~2008년)

1.7 Hour: 당시 시간

1.8 NO2. Mean(ppb): 주어진 날의 NO2 농도의 산술 평균. NO2는 공기중의 질소와 산소가 연료의 연소시에 반응하여 생성됨, 과거 LA스모그 사건 및 산성비의 한 원인물질로서 중요하게 취급됨.

1.9 NO2. 1st.Max. Value(ppb): 주어진 날의 NO2 농도 최대 값

1.10 NO2. 1st.Max. Hour: 주어진 날에 최대 NO2 농도가 기록 된 시간

1.11 NO2. AQI: 주어진 하루 동안의 NO2의 대기 성분 지수, AQI가 증가할수록 더 많은 수의 인구가 건강상의 악영향을 받을 가능성이 높음

1.12 O3. Mean(ppm): 주어진 날의 O3 농도의 산술 평균. O3는 대기 중에 배출된 질소산화물과 휘발성 유기화합물(VOCs)등이 자외선과 광화학 반응을 일으켜 생성된 2차 오염물질임.

1.13 O3. 1st.Max. Value(ppm): 주어진 날의 O3 농도 최대 값

1.14 O3. 1st.Max. Hour: 주어진 날에 최대 O3 농도가 기록 된 시간

1.15 O3. AQI: 주어진 하루 동안의 O3의 대기 성분 지수

1.16 SO2. Mean(ppb): 주어진 날의 SO2 농도의 산술 평균. SO2(황산화물)은 1차적 오염 물이며 화석 원료의 연소에 의하여 방출됨. 대기오염물질 중 가스상태의 것으로서 가장 문제시됨

1.17 SO2. 1st. Max. Value(ppb): 주어진 날의 SO2 농도 최대 값

1.18 SO2. 1st. Max. Hour: 주어진 날에 최대 SO2 농도가 기록 된 시간

1.19 SO2. AQI: 주어진 하루 동안의 SO2 대기 성분 지수

1.20 CO. Mean(ppm): 주어진 날의 CO 농도의 산술 평균. 무색, 무취의 유독성 가스로서 탄소성분이 불완전 연소되었을 때 발생

1.21 CO. 1st. Max. Value(ppm): 주어진 날의 CO 농도 최대 값

1.22 CO. 1st. Max. Hour: 주어진 날에 최대 CO 농도가 기록 된 시간

1.23 CO. AQI: 주어진 하루 동안의 CO의 대기 성분 지수

1.24 ET: UFO(미확인 비행 물체)관찰이 이루어진 경우 1, 이루어지지 않은 경우를 0으로 본다

2. 서울시 대기 성분 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

2.1 측정일시: 측정을 시작한 날짜와 시간

2.2 측정소명: 측정을 한 서울시 내의 위치

2.3 이산화질소 (NO2) 농도 (ppm): 1시간 동안 측정한 이산화질소의 평균 농도, 14년 경기의 연평균 농도가 0.034ppm으로 최고, 광주가 0.019ppm으로 최저

2.4 오존 (O3) 농도 (ppm): 1시간 동안 측정한 오존의 평균 농도, 14년 서울의 경우 0.023ppm 인천과 경기는 0.026, 0.025 ppm

2.5 일산화탄소 (CO) 농도 (ppm): 1시간 동안 측정한 일산화질소의 평균 농도, 수도권 평균 오염도는 0.5~0.6 ppm

2.6 아황산가스 (SO2) 농도 (ppm): 1시간 동안 측정한 아황산가스의 평균 농도

2.7 미세먼지 (㎍/㎥): 1시간 동안 측정한 미세먼지의 평균 농도, 입자의 지름이 10㎛ 이하인 경우 미세먼지로 불리며 호흡시 폐포에 침투함

2.8 초미세먼지 (㎍/㎥): 1시간동안 측정한 초미세먼지의 평균 농도, 입자의 지름이 2.5㎛ 이하인 경우 초미세먼지로 불리며 호흡시 폐포에 침투한다.

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.

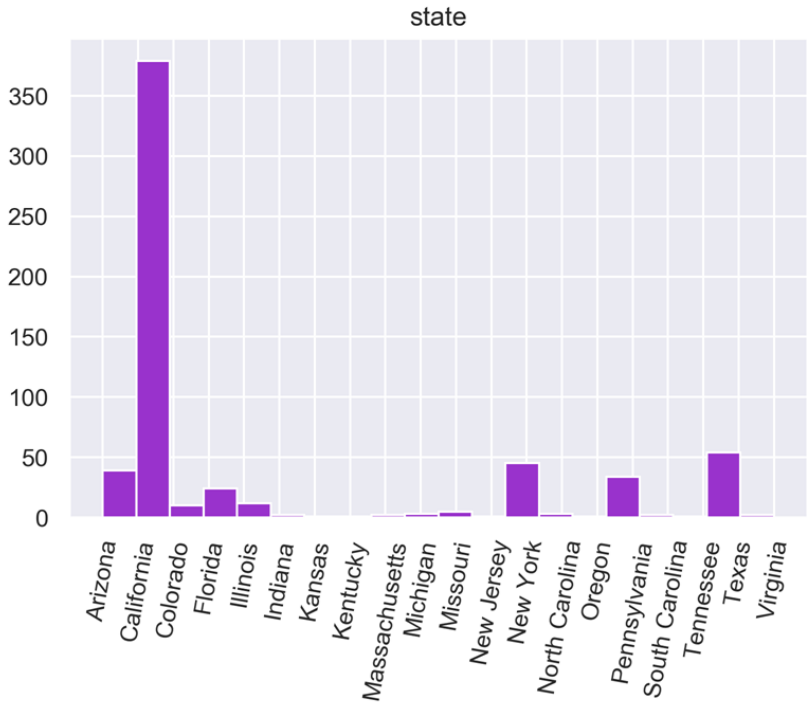
“UFO sightings + Air Quality” 데이터 중 구체적인 위치가 제시되지 않은 데이터의 경우 (“Not in a city”) 분석에 사용하지 않았다.

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, 대기 성분

1 지역: 해안과 내륙 지방 중 어디에서 UFO가 많이 보일까?

1.1 어느 주 (state)에서 UFO가 많이 관측되었을까?



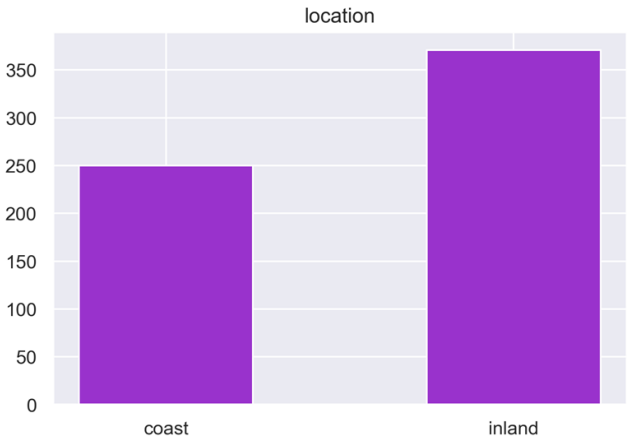
Insight: California 주가 압도적으로 많은 편이며, Texas, New York, Pennsylvania, Arizona 주가 그 뒤를 이었다.

1.2 UFO가 관측된 city 위치를 해안(coast)과 내륙(inland) 지역으로 나누어 보자

앞의 결과와 같이 state 별 분석이 가능하지만 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.

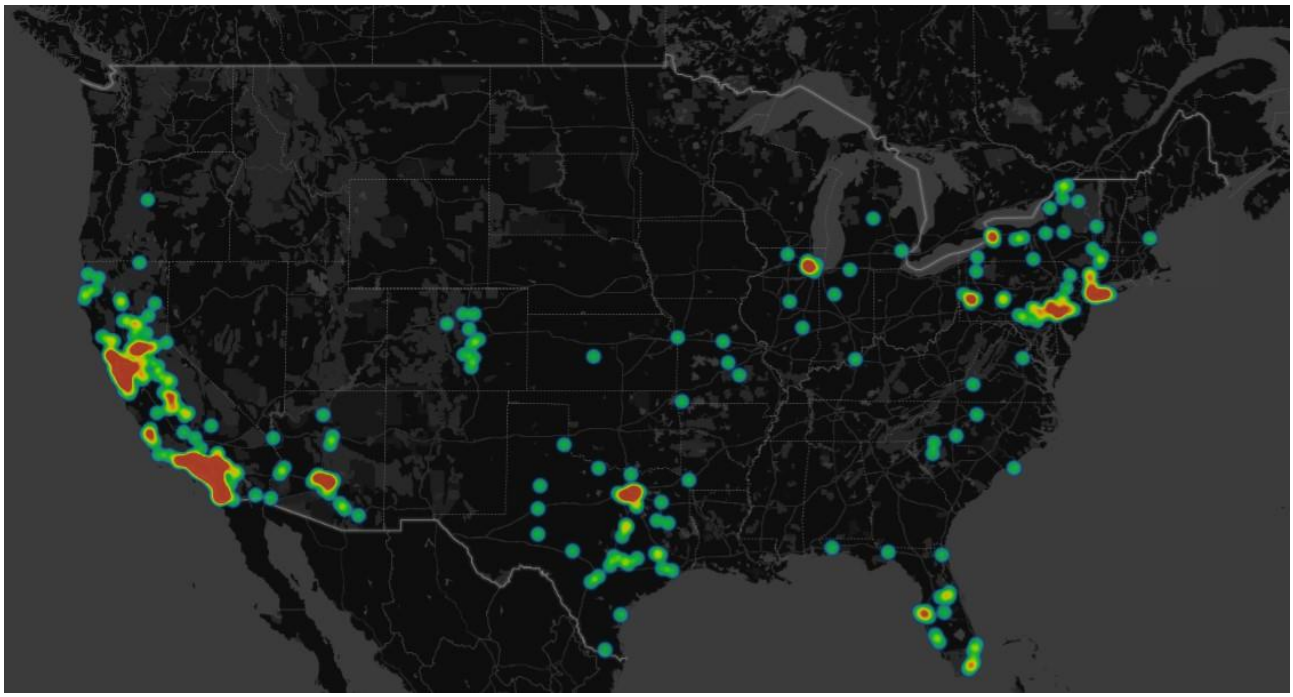
- 알래스카 주를 뺀 미국 본토 국토 총 면적: 8,080,510 km^2
- 해안(coast) 지역 총 면적 = 해안 지역의 총 길이 (9,027km) * 20km = 180,540 km^2
- 내륙(inland) 지역 총 면적 = 8,080,510 km^2 - 180,540 km^2
- **해안(coast) : 내륙(inland) 면적 비 = 2.23 : 97.77**

1.3 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. 따라서 "지역"은 UFO 관측 적합조건에 영향을 준다고 볼 수 있다.

1.4 미국 본토 지도 위에 관측위치 분포를 그려보자



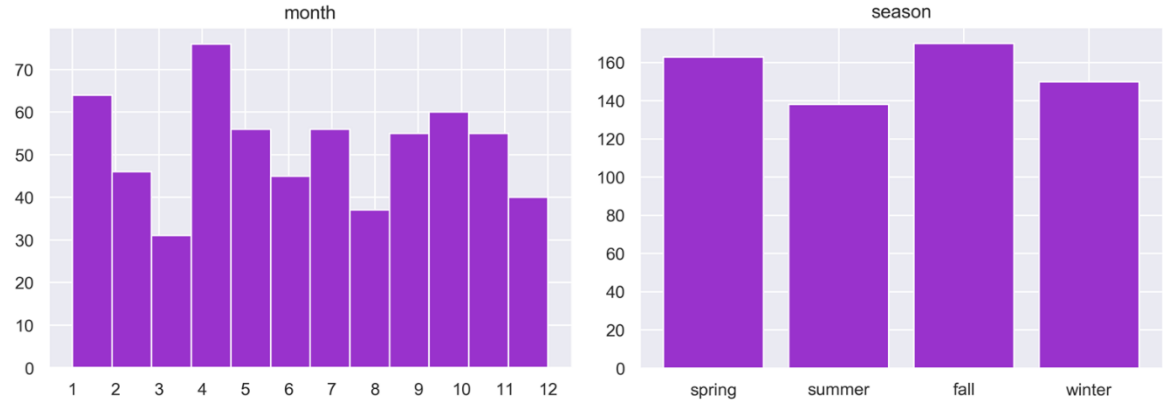
Insight: 앞의 결과와 동일하게 대부분의 UFO 관측 보고가 해안과 가까운 지역 (예: 캘리포니아 주)에서 이루어짐을 알 수 있다.

2 시간: UFO가 많이 보이는 계절, 시간대가 있을까?

2.1 UFO가 많이 보이는 "계절"은 언제일까?

아래와 같이 month 변수를 season 변수로 나누어 UFO 관측 시 데이터 분포를 살펴보았다.

Spring: 3-5월	Summer: 6-8월
Fall: 9-11월	Winter: 12-2월

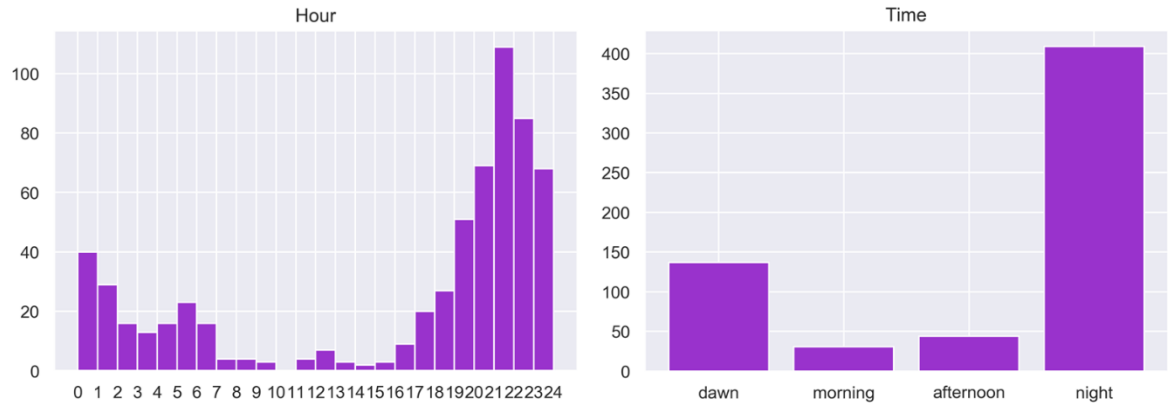


Insight: 월별로 조금씩 차이가 나기는 하지만, 계절단위로 보게 되면 큰 차이는 없다. 따라서 "계절"은 UFO 관측에 크게 중요한 요인이 아닌 것으로 보인다. 각 계절간 n수가 3개뿐 안되기 때문에 별도의 통계분석은 하지 않았다.

2.2 UFO가 많이 보이는 "시간"은 언제일까?

아래와 같이 hour 변수를 time 변수로로 나누어 UFO 관측 시 데이터 분포를 살펴보았다.

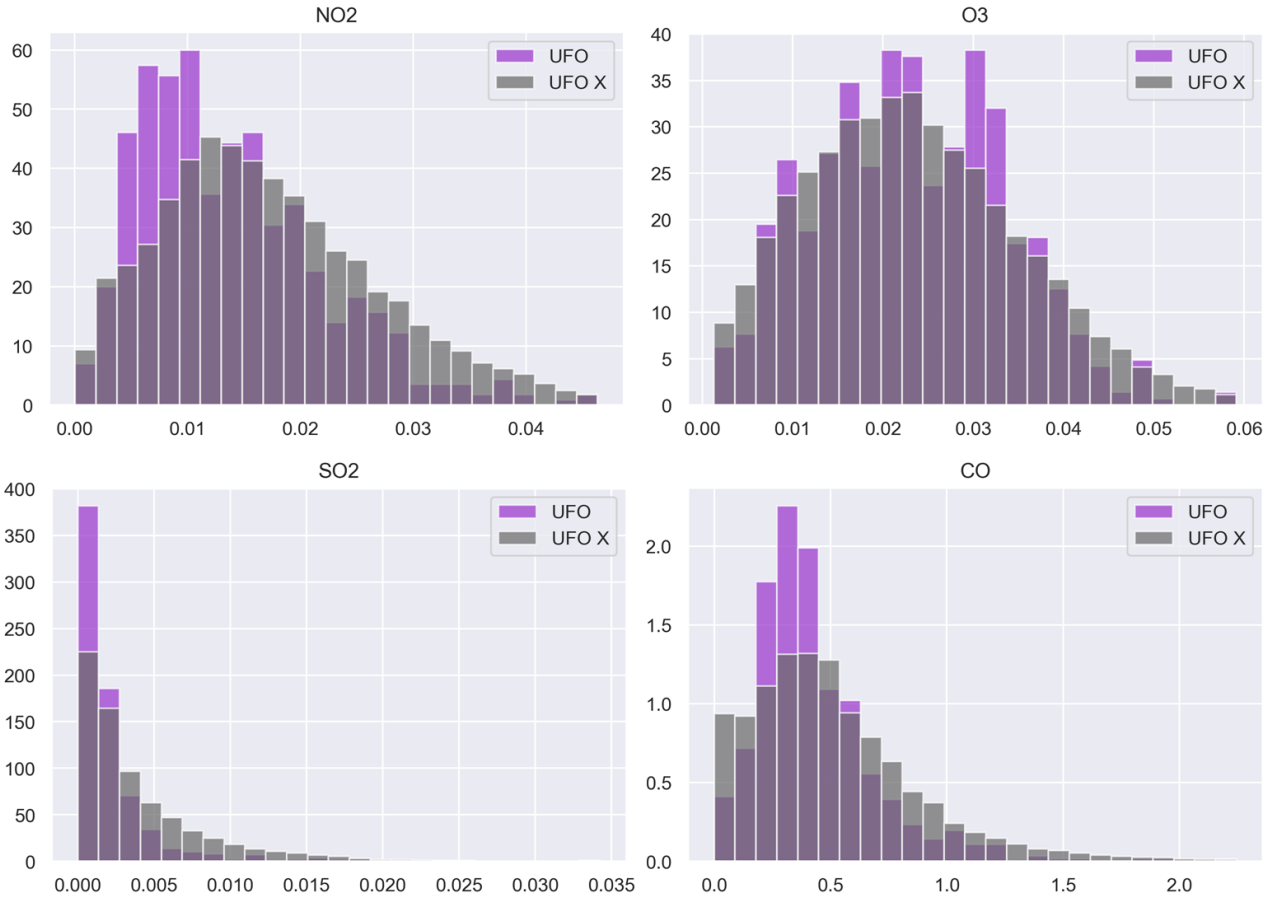
Dawn: 0-6시	Morning: 6-12시
Afternoon: 12-18시	Night: 18-24시



Insight: 오후, 아침처럼 날이 밝을 때 보다는 밤, 새벽에 UFO가 많이 관측되었다. UFO를 관측하기 위해서는 "시간" 조건이 중요하다는 것을 알 수 있다.

3 UFO가 많이 관측되는 대기 성분 조건이 있을까?

UFO가 관측되었을 때와 평상시로 나누어 각 성분의 농도를 확률밀도함수 (probability density function, pdf)로 계산하여 그려보았다



Insight: 대기 성분을 나타내는 4가지 지표 중 UFO가 관측될 때(UFO)와 관측되지 않았을 때(UFO X) 가장 차이가 두드러지게 나타나는 지표는 NO2 농도다. CO, SO2농도가 그 다음으로 눈에 띄며, O3 농도는 큰 분포차가 나타나지 않았다. 분포 사이의 유사도는 Mutual information 등을 계산하여 비교할 수 있겠으나 본 분석에서는 시행하지 않았다.

NO2, SO2, CO는 모두 자동차 배기가스의 주된 성분이자 광화학 스모그의 원인들이고, 그중 NO2가 가장 주된 스모그 원인임을 고려할 때 **대기 오염으로 인한 가시거리 감소가 적은 날 밤하늘에 UFO를 볼 수 있다**는 결론을 도출해볼 수 있다.

그러나 오존(O3)의 농도가 평소보다 높을 때 UFO가 자주 관측되는 현상도 발견되었다. 또한 일산화탄소(CO)의 농도 또한 어느 정도 높을 때 UFO가 더 많이 발견되는 것을 볼 수 있었다.

과연 머신러닝의 결과도 같은 결과를 보여줄까?

우리 나라에서 UFO를 볼 수 있었을까?

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 "UFO sightings + Air Quality" 데이터셋의 문제점: Imbalanced dataset

"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.

- **Re-sampling**

  Imbalanced dataset을 처리하는 가장 간단한 방법은, 다수 그룹(ET=0)의 데이터샘플을 under sampling하거나 소수 그룹(ET=1)의 데이터샘플을 over sampling하는 것이다. 이 외에도 두 방법을 모두 사용하는 hybrid sampling 방식과 k-Nearest Neighborhood 알고리즘을 이용한 SMOTE (Synthetic Minority Over-sampling Technique) 등의 방법들이 있다. Under sampling과 over sampling은 각각 데이터에 있는 중요한 정보가 소실되거나, 훈련 데이터 중복으로 과적합(Overfitting)될 가능성이 존재한다. 각 Resampler마다 장단점이 존재하기에 **이번 분석에서는 random under sampling, random over sampling 및 SMOTE resampling을 모두 시행하여 결과를 비교해보았다**.

- **Class weight balancing**

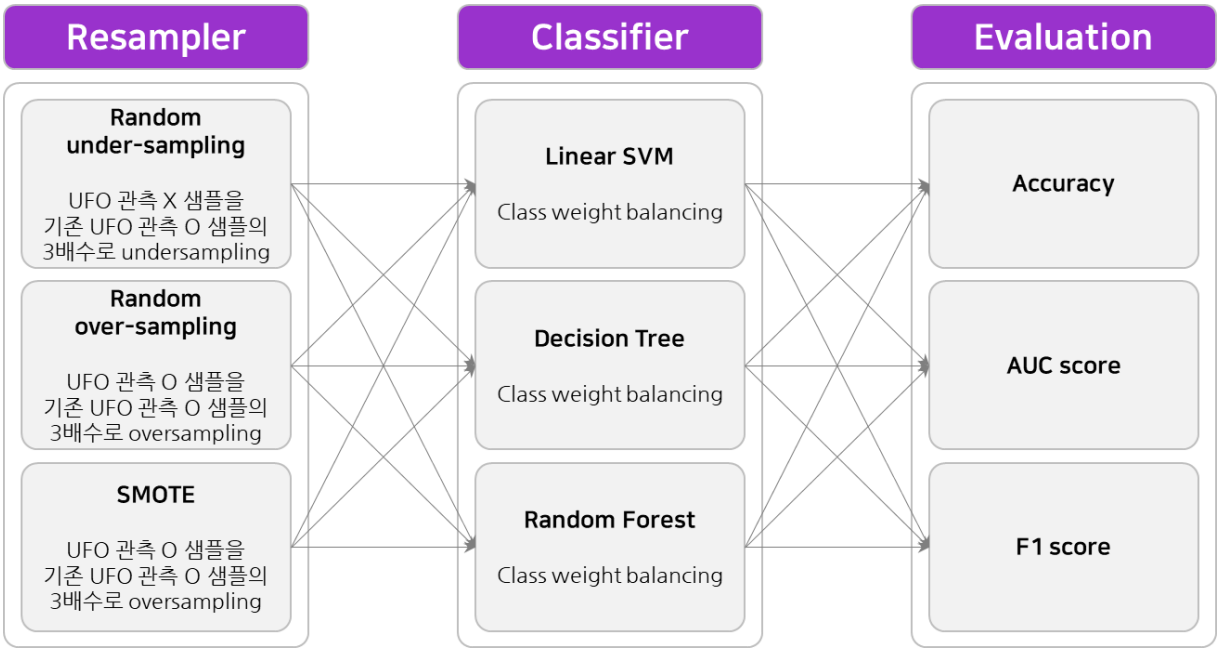
  Imbalanced dataset을 처리하는 두번째 방법은 모델 훈련 시 각 class 분류에 weight를 주는 방법이다. 즉, 소수 그룹에 해당하는 샘플을 분류한 결과를 더 중요하게 여긴다. 우리가 사용한 **데이터셋의 IR이 너무 높아 resampling 방법만으로는 imbalanced dataset 문제가 해결되기 어렵다고 판단하여 class weight balancing 또한 함께 사용하였다.**

2 UFO 관측 적합조건 예측 모델

2.1 UFO 관측 적합조건 예측 모델 선택

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

2.2 UFO 관측 적합조건 예측 모델 훈련



[변수] 모델에 입력으로 들어가는 **14개의 변수**는 아래와 같다. 이 중 10개의 **Categorical variables는 one-hot encoding** 된 형태로 모델에 사용되었다.

- [*Location*] Categorical variables (2) - Location, Coast
- [*Season*] Categorical variables (4) - Spring, Summer, Fall, Winter
- [*Time*] Categorical variables (4) - Dawn, Morning, Afternoon, Night
- [*Air quality*] Numerical variables (4) - NO2, O3, SO2, CO

[Resampler 와 Classifier] 3개의 샘플링 방법과 3개의 분류 모델로 이루어진 파이프라인으로 모델을 훈련시켰다. 모델간 비교를 위해 트레이닝 데이터와 테스트 데이터를 7:3 비율로 나누어 평가지표를 계산하였다 (hold-out). Cross-validation 방법을 이용하면 더 정교화된 모델 비교가 가능하겠지만, 모델 훈련에 소요되는 시간이 너무 오래 걸려 hold-out 방법을 사용하였다.

[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 관측조건 O 예측	UFO 관측조건 X 예측
실제 UFO 관측조건 O	*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.

$$AUC = \frac{Specificity + Sensitivity}{2}$$

$$Specificity = \frac{TN}{FP + TN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

➢ F1 score

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

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

2.3 UFO 관측 적합조건 예측 모델 성능 비교 및 최종 모델 선정

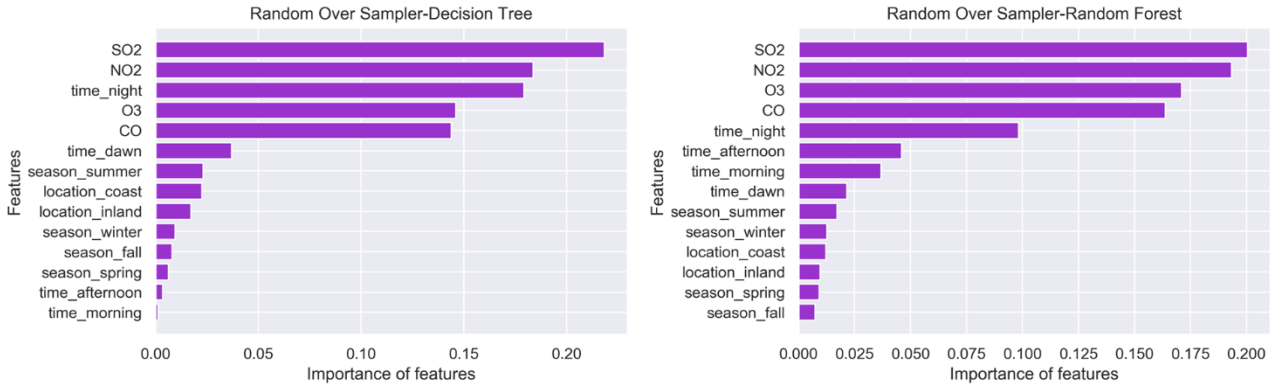
Decision Tree, Random Forest 분류기에서 전체적인 성능이 좋게 나타났다.

Resampler	Classifier	Accuracy	AUC	F1 score
Random under sampling	*Support Vector Machine*	0.731	0.782	0.538
Random under sampling	*Decision Tree Classifier*	0.768	0.700	0.544
Random under sampling	*Random Forest Classifier*	0.822	0.814	0.543
Random over sampling	*Support Vector Machine*	0.745	0.803	0.142
Random over sampling	*Decision Tree Classifier*	0.987	0.937	0.813
Random over sampling	*Random Forest Classifier*	0.994	0.956	0.899
SMOTE	*Support Vector Machine*	0.746	0.818	0.148
SMOTE	*Decision Tree Classifier*	0.967	0.704	0.453
SMOTE	*Random Forest Classifier*	0.976	0.883	0.429

Insight: Random over sampling-Decision Tree classifier 조합과 Random over sampling-Random Forest classifier 조합에서 AUC, F1 score 성능이 모두 좋게 나타났다. 편의를 위해 각각을 **DT, RF 모델**로 칭한다.

2.4 변수별 중요도 확인

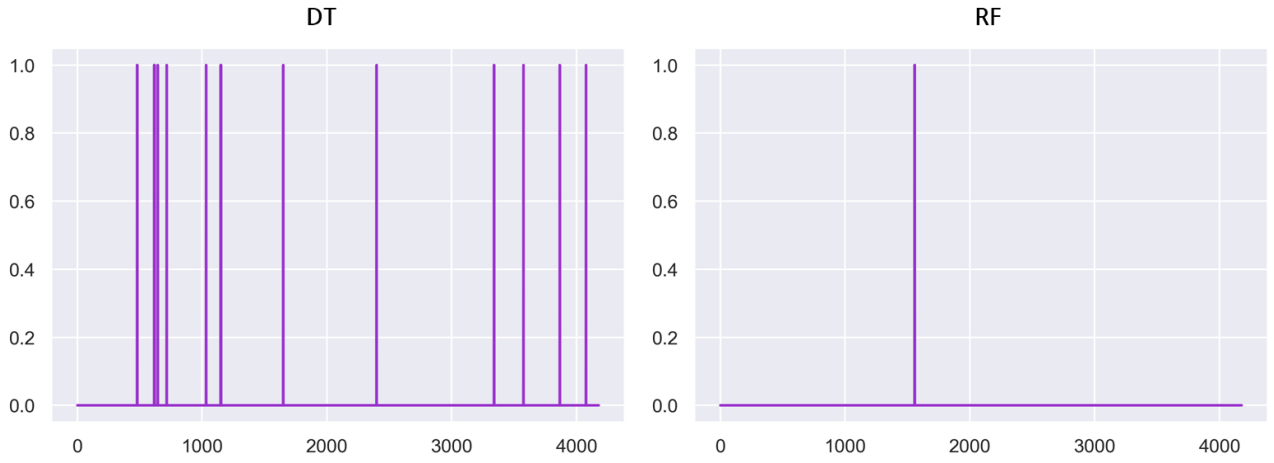
앞의 분석을 통해 UFO 관측 적합조건 예측 모델 성능이 좋게 나타난 두 조합(DT, RF)을 찾을 수 있었다. Decision Tree와 Random Forest는 모델에서 사용된 각각의 변수들의 information gain 평균으로 어떤 변수가 중요한지 비교할 수 있다. 이 과정은 Scikit-learn package의 feature_importances_함수를 이용해 구현할 수 있다.



Insight: 두 모델에서 조금씩 차이가 있기는 했으나, **대기 성분조건이 UFO 관측 적합조건을 예측하는데 가장 중요**한 역할을 하고 있음을 알 수 있다. "새벽/오전/오후/밤" 변수가 중위권에 계속해서 위치한 것을 볼 때, **"시간" 조건 또한 중요한 변수**로 여겨진다.

2.5 서울시의 UFO 관측 적합조건 예측

DT, RF 모델로 서울시의 UFO 관측 적합조건 여부를 예측해보았다.



Insight: DT 모델에서는 총 10건의 적합조건이 발견된 반면, RF 모델은 2건의 적합조건이 발견되었다. 따라서 우리는 **각 모델이 각각 Liberal, Conservative한 특징을 갖는다**고 할 수 있을 것이다.

네이버 커넥트원에서는 UFO를 볼 수 있을까?

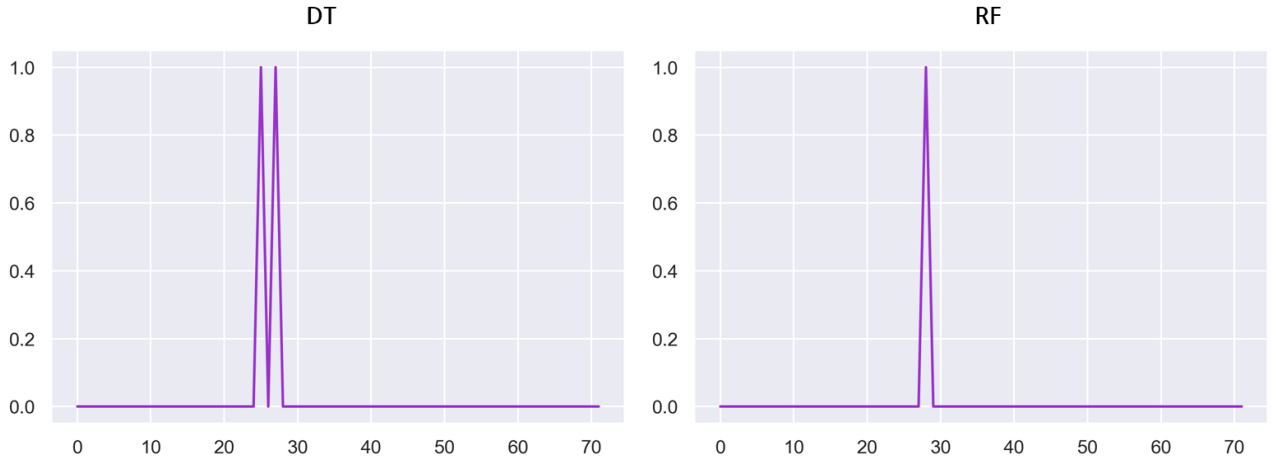
우리가 오프라인 교육 대상자로 선발된다면 집중교육을 받게 될 커넥트원에서는 UFO를 볼 수 있을까?

1 데이터 수집

대기 성분 데이터를 한달 후까지 예측할 수 있다면 위 질문에 답할 수 있겠지만, 여기에는 너무 많은 Confounding factor가 존재하여 예측이 어렵다. 따라서 본 분석에서는 **과거 동일한 날짜 (8/27-29)에 UFO를 관측할 수 있었을 지 알아보았다**. 국내 각 지역의 시간별 대기 성분 데이터는 에어코리아 (http://www.airkorea.or.kr/realSearch)에서 찾을 수 있었지만 아쉽게도 이 사이트는 2017년에 개설되어, 이전 년도의 데이터는 구할 수 없었다.

2 예측

과연 2017년도 8월 27일-29일은 UFO를 볼 수 있는 시기였을까?



Insight: DT 모델에서는 총 2건의 적합조건이 발견된 반면, RF 모델은 1건의 적합조건이 발견되었다. 앞서 각 모델이 Liberal/Conservative한 특성을 갖는다고 했던 것과 일치하는 결과를 보인다. 아래 테이블은 두 모델이 예측한 두 시점의 변수 데이터들이다. **두 모델 모두 비슷한 시기가 UFO 관측에 적합한 조건이라고 예측하고 있다.**

index	실제시간	location	season	time	NO2	O3	SO2	CO
25	8/29 01:00	inland	summer	dawn	0.004	0.029	0.003	0.4
27	8/28 23:00	Inland	summer	night	0.005	0.033	0.002	0.4
28	8/28 22:00	inland	summer	night	0.008	0.032	0.002	0.4

만약 올해 같은 시기에 비슷한 대기조건이 유지된다면, 8월 29일 자정 전후 시간대가 UFO를 보기 적합한 조건이라고 말할 수 있다. 다 함께 나가서 UFO를 찾아보는 건 어떨까?