내 주변에는 어떤 생물이?

팀 통마늘
신보현

INDEX

ch 1 문제 인식 및 분석 주제

ch 2 분석 방법 & EDA

ch 3 Feature Engineering

ch 4 Modeling & Interpretation

ch 5 한계 및 의의

Chapter 1

문제 인식 및 분석 주제

envbigdata

문제 인식

인천 수돗물 사태

2020.07.26일자 한경 신문

인천 수돗물 '깔따구' 유충 3건 발견…총 257건

입력 2020.07.26 14:38 | 수정 2020.07.26 14:38

- 올해 발생한 인천 수돗물 사태는 깔따구 유충이 원인으로 지목됨
- 수돗물 사태로 인해서 인천 시민들이 큰 피해를 받음

멸종 위기 동식물을 보호하기 위한 노력

2020.12.24일자 물산업신문

한강청, '한강의 멸종위기 야생동물' 도감 발간

- 멸종위기 동식물은 생태계의 질서를 유지하기 위해 필요한 행동임
- 하지만 멸종위기 동물이 희소한만큼 보호하기가 어려운게 현실임

문제 인식

데이터를 이용해서 동식물의 서식 여부를 미리 유추할 수 있다면?

- 깔따구류와 같은 유해 동물의 서식 여부를 미리 알 수 있다면 인천 수돗물 사태와 같은 상황을 막을 수 있다.
- 멸종위기 동식물의 서식 여부를 미리 알 수 있다면 그 지역에 생태학자를 파견해 해당 종을 면밀하게 관찰할 수 있다.
- 멸종위기 동식물뿐만 아니라 생태학자가 연구하고 싶은 동식물 종의 서식 여부를 미리 알 수 있다면 연구에 큰 도움이 될 것이다.

분석 주제

동식물이 관측된 데이터를 이용해서 분포 지역을 분석해보자

- 종별로 관측된 데이터는 그 종의 서식 패턴을 담고 있을 것이다.
- 관측 데이터를 분석해서 종의 서식 패턴을 분석해보자
- 더 나아가서 동식물 종이 관측되지 않은 지역의 관측 여부를 예측해보자

기대 효과

유해 동식물 차단, 생태계 질서에 기여, 효율적인 생태계 연구

- 유해 생물 종의 서식 여부를 미리 유추함으로써 인천 수돗물 사태 등을 방지함
- 멸종위기 동식물을 파악하여 보호함으로서 생태계 질서에 기여할 수 있음
- 생태학자가 관심있는 동식물 종을 탐구하고자 할 때, 연구를 수행할 지역을 효율적으로 선택할 수 있음

분석의 큰 틀

맞추고자 하는 것(관심있는 것) [반응변수]

- 특정 지역의 동식물 거주 여부
- 0, 1의 이진 형태 값 (모델 결과는 확률)

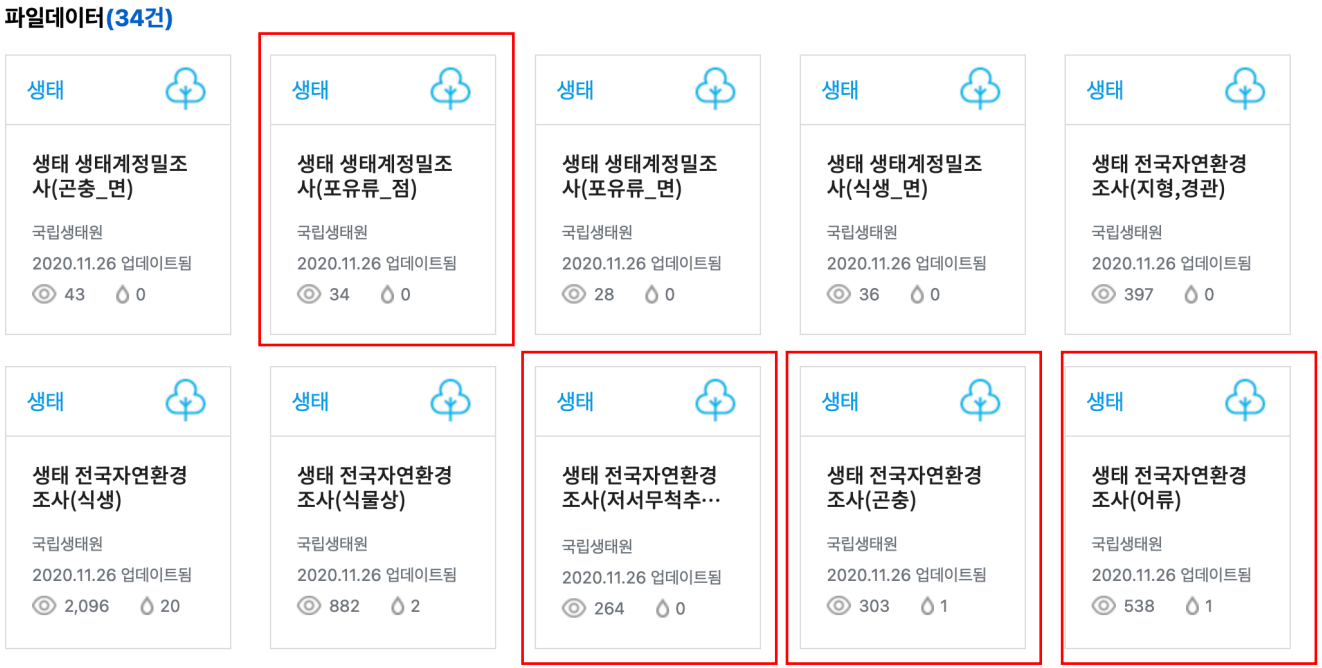
필요한 것 [독립변수]

- 환경 변수, 위도 경도 등..

사용 데이터

환경 빅데이터 플랫폼

- 특정 지역의 동식물 거주 여부 [반응 변수]
- 생태학자가 연구한 지역에서 위도 경도 형태로 저장됨
- 포유류, 양서파충류, 곤충, 조류, 자서무척추동물, 식물의 생태 데이터를 사용함



사용 데이터

Chelsa 기후 데이터

CHELSA

Climatologies at high resolution for the earth's land surface areas

- 환경 변수는 특정 지역별로 제공되어야 함
- 하지만 국내 환경 데이터 (기온, 풍속 등)는 관측치 지점 단위로만 제공됨
- 예를 들어서, 특정 위도, 경도의 기온을 알고 싶으면 그곳과 가장 가까운 관측치의 기온을 사용할 수밖에 없음
- 본 연구에서는 동식물 종이 발견된 위치의 환경 데이터가 중요하다고 생각되어, 위도 및 경도별로 환경 데이터를 제공하는 Chelsa 기후 데이터를 사용하기로 함
- 이를 동식물 종의 서식 여부를 예측하기 위한 독립변수로 사용하였음

Chapter 2

분석 방법 및 EDA

envbigdata

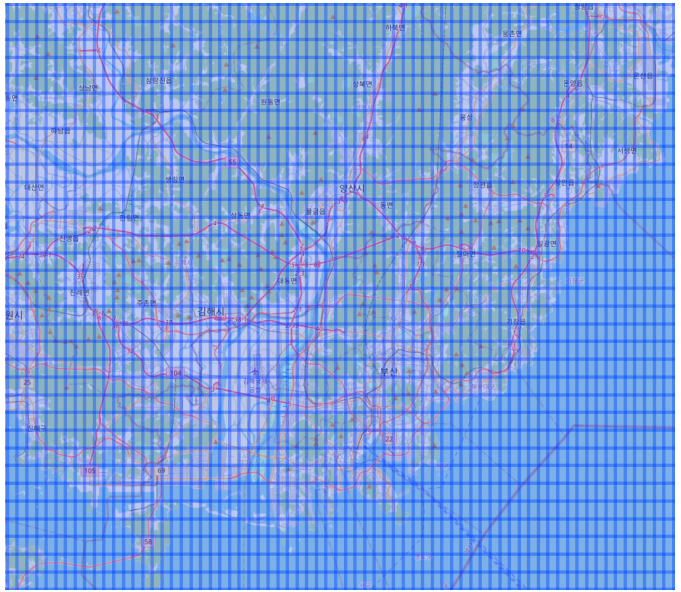
분석의 기본 단위

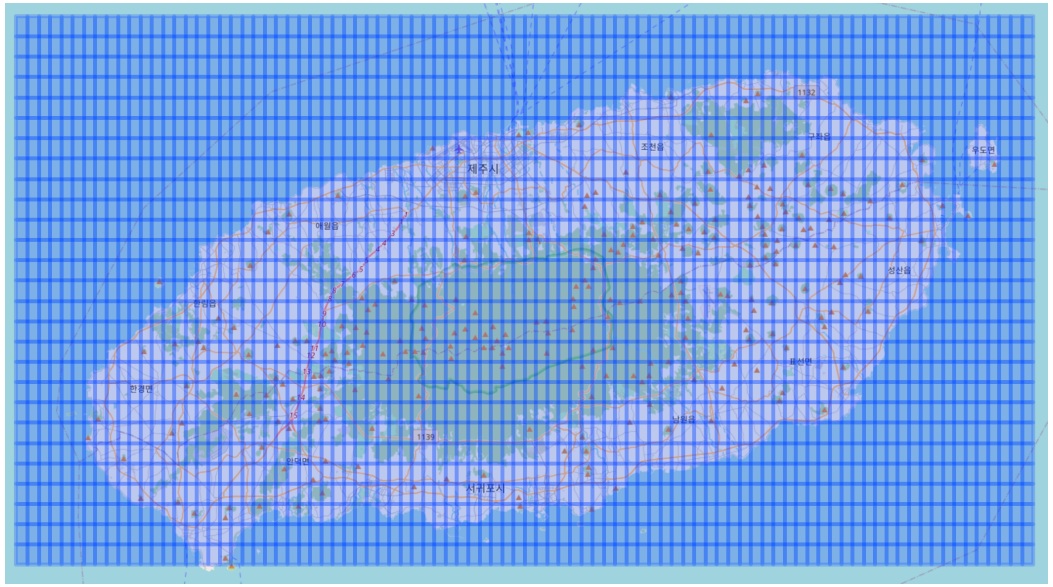
- 동식물 종의 관측 데이터는 위도 및 경도로 저장되었지만 데이터를 살펴본 결과 특정 조사에서 관측된 종들은 모두 동일한 위도와 경도로 저장되었음
- 따라서 기본 단위를 위도, 경도가 아닌 특정지역으로 정하는 것이 낫다고 판단함

분석방법

분석의 기본 단위

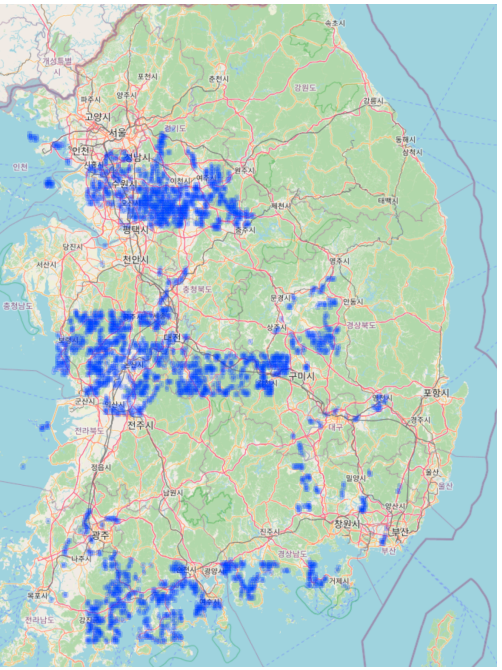
- 특정 지역의 크기는 1km by 1km(이하 그리드라고 칭함)로 정하였음
- 이보다 더 작다면 데이터가 희소(Sparse)해지고 더 크다면 생태학자들이 지역을 탐색하는데 어려움이 있을 것이라 판단했기 때문
- 예를 들어, 한반도 및 제주도를 아래와 같이 그리드로 나눔
- 왼쪽 그림은 한반도를 그리드로 나눈 결과 일부를 부산 지역으로 확대하여 살펴본 것임



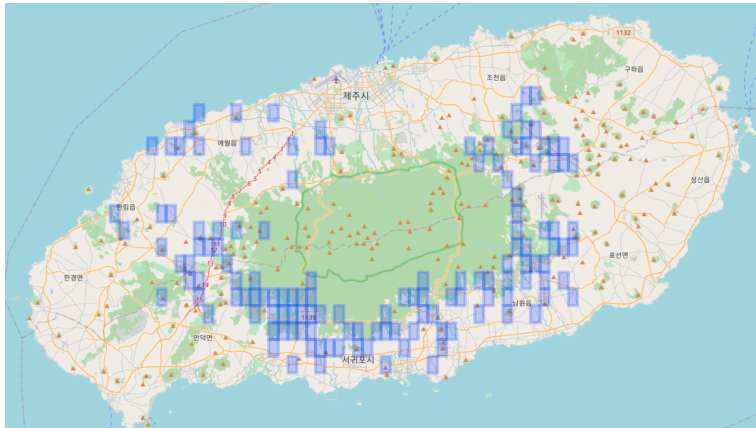


분석방법

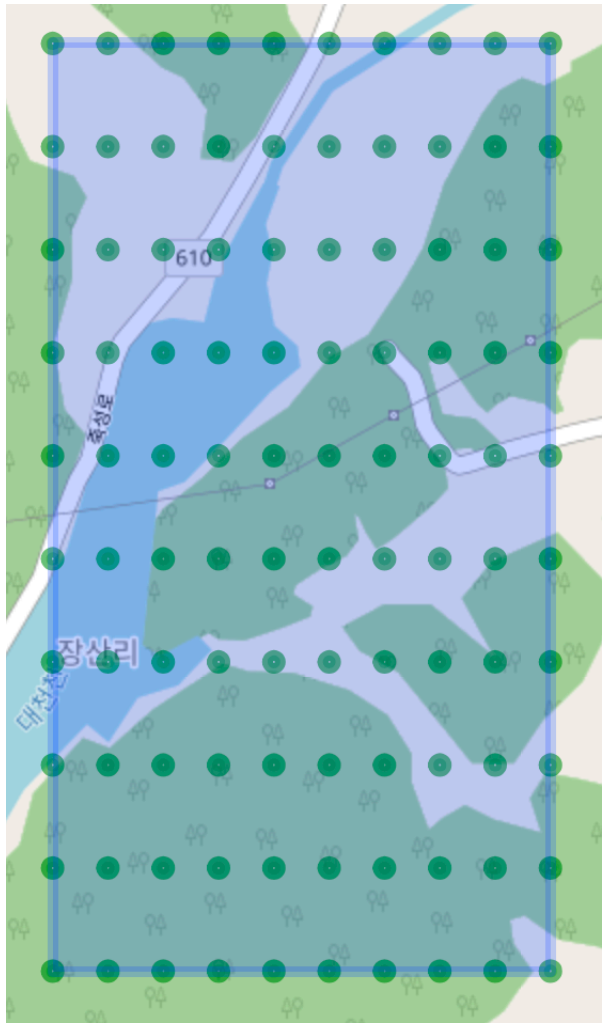
분석의 기본 단위



- 그리드로 나눈 모든 지역을 사용하는 것이 아니라, 그 중 연구가 진행된 적이 있는 곳만 데이터로 사용함
- 예를 들어, 한반도에서 조류가 관측된 적이 있는 그리드는 왼쪽 그림과 같고 제주도에서는 오른쪽 그림과 같음
- 조류가 관측된 적이 있는 그리드만 훈련 및 테스트 데이터로 사용함
- 해당 그리드에 생물종이 관측되었다면 1, 그렇지 않다면 0으로 기록함



분석방법



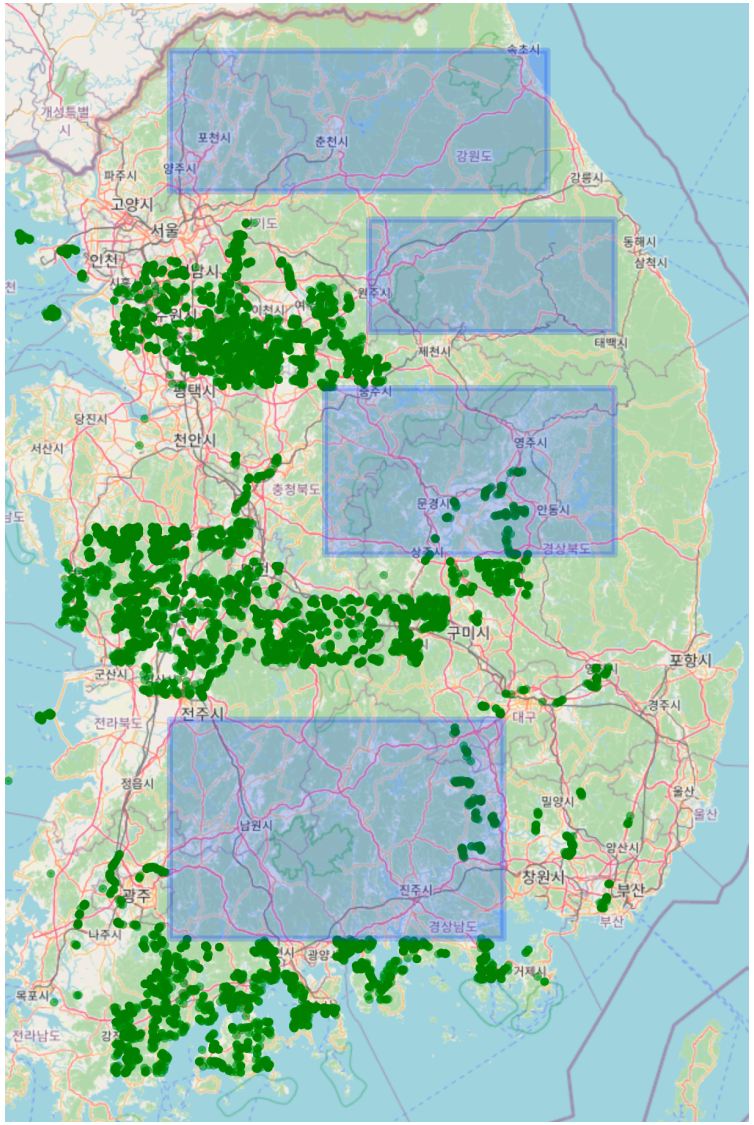
분석의 기본 단위

- 왼쪽 그림은 선정된 그리드 중 하나의 예시임
- 이 그리드에는 각 생물종의 서식 여부가 0 및 1로 기록되어 있음
- 독립변수를 생성하기 위해, 해당 지역의 환경 변수를 도출해야 함
- 이는 왼쪽 그림과 같이 일정한 간격의 좌표를 뽑아서, 해당 좌표의 환경 변수 데이터를 합하였음
- 자세한 사항은 Feature Engineering에서 살펴봄

최종 데이터 형태

그리드 구분	변수1	변수2	변수3	…	변수4	y (서식 여부)
1	100	200	300	…	400	0
2	100	200	300	…	400	1

분석방법



훈련 및 검증 방향

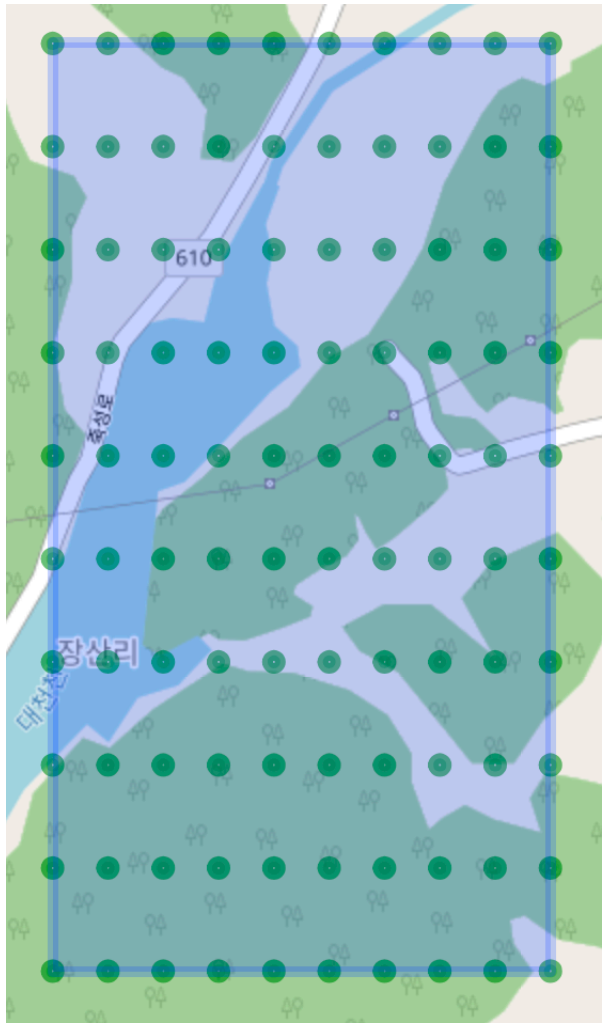
- 왼쪽 그림에서 초록색으로 표시된 그리드가 훈련 및 테스트 데이터로 쓰일 지역임
- 파란색 네모로 표시된 지역이 탐사되지 않은 지역으로, 모형이 예측할 지역임
- 파란색 네모와 같은 미지의 지역에 대한 생물 종의 서식 여부를 확률로 반환하는 것이 최종 목표

Chapter 3

Feature Engineering

envbigdata

Feature Engineering



독립 변수 도출

- 왼쪽 그림은 기본 단위 그리드 중 한 예시임
- 이 그리드의 독립 변수를 생성해야 함
- 이를 위해 그리드에서 일정한 간격의 좌표를 정하고 이 좌표들의 환경 변수를 평균내어서 하나의 값으로 만듦
- 예를 들어, 왼쪽 그림의 그리드에는 총 100개의 좌표가 있고 100개 좌표의 기온을 평균내어서 해당 그리드의 평균 기온을 도출함
- 이런 방식으로 풍속, 습도 등의 독립변수를 생성
- data source는 Chelsa로 사용함

Feature Engineering

CHELSA

Climatologies at high resolution for the earth's land surface areas

WorldClim

외부 데이터

- 기상청의 데이터 관측 장소는 흩어져 있으므로 그리드별로 상이한 환경 변수를 도출하기가 어려움
- Chelsa, WorldClim 데이터는 위도, 경도별로 환경 데이터를 제공함
- 물론, 위도 경도별로 환경 데이터를 모두 관측한 것이 아니라, spatially interpolated 된 데이터임
- 각 data source로부터 생물기후 데이터를 추출함

WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas, SE Fick et al., International journal of climatology 2017.

Climatologies at high resolution for the earth's land surface areas, DN Karger et al., Nature 2017.

Feature Engineering

추출 데이터

변수	설명	변수	설명
bio1	mean annual air temperature	bio12	annual precipitation amount
bio2	mean diurnal air temperature range	bio13	precipitation aomunt of the wettest month
bio3	isothermality (ratio of diurnal variation to annual variation in temperatures)	bio14	precipitation amount of the driest month
bio4	temperature seasonality	bio15	precipitation seasonality
bio5	mean daily maximum air temperature of the warmest month	bio16	mean monthly precipitation amount of the wettest quarter
bio6	mean daily minimum air temperature of the coldest month	bio17	mean monthly precipitation amount of the driest quarter
bio7	annual range of air temperature	bio18	mean monthly precipitation amount of the warmest quarter
bio8	mean daily air temperature of the wettest quarter	bio19	mean monthly precipitation amount of the coldest quarter
bio9	mean daily air temperature of the driest quarter	pet	potential evapotranspiration
bio10	mean daily air temperature of the warmest quarter	rh	relative humidity
bio11	mean daily air temperature of the coldest quarter	srad	surface solar radiation downwards

최종 데이터 형태

그리드 구분	bio1	bio2	bio3	…	srad	y (서식 여부)
1	100	200	300	…	400	0
2	100	200	300	…	400	1

동식물별 데이터 개수

	조류	곤충	자서무척추동물	식물
데이터 총 개수	3116	789	559	913

Chapter 4

Modeling & Interpretation

envbigdata

Modeling

모델링 흐름

Modeling

1. 종별로 다른 데이터 구성

- 조류, 곤충, 자서무척추동물, 식물별로 관측된 지역이 상이하므로 데이터 개수가 상이함
- 조류 내에서도 조류 종별로 그리드에 관측 여부가 상이함
- 따라서 종별로 데이터를 다르게 구성
- 예를 들어 조류에서 섬휘파람새, 직박구리의 데이터를 다르게 구성함
- 또한 자서무척추 동물의 깔따구류 데이터도 다르게 구성함

Modeling

2. 모형 선정

통계 모형

- 이진값을 분류하기 위한 대표적인 통계 모형은 로지스틱 회귀가 있음
- 로지스틱 회귀는 MLE에 기반한 추론 방법임
- MLE에 기반한 추론은 불균형 데이터 추론에 취약함

Tree 계열 모형(Ensemble, Boosting)

- 비선형 모형임
- 통계 모형에 비해 해석력은 떨어지지만 정확도는 확보할 수 있음
- tree interpreter 등을 통해서 모델링 결과를 해석할 수 있음
- 불균형 데이터 문제도 있고 데이터 개수가 많지 않으므로 Tree 계열 모형을 선택
- 최종 선택 모형

Modeling

2. 파라미터 튜닝

GridSearch Optimization

- 파라미터 공간을 Discrete하게 탐색
- 파라미터 공간을 늘리면 탐색 시간이 매우 오래 걸릴 수 있음
- 파라미터 공간을 Discrete하게 제한해야 하기 때문에 Optimal Point을 포함하지 못할 수도 있음

Bayesian Optimization (hyperopt)

- 파라미터 공간을 Continuous하게 탐색
- Continous하게 탐색하기 때문에 파라미터 공간을 늘리더라도 탐색 시간이 선형적으로 증가하지 않음
- GridSearch에 비해서 파라미터 공간의 제약이 상대적으로 적음
- 최종 선택 방법

Modeling

2. 파라미터 튜닝

RandomForest

- 과적합 방지 파라미터
  - max_depth
  - min_samples_split
  - min_samples_leaf
- 기타 파라미터
  - n_estimators
  - max_features

XGB

- 과적합 방지 파라미터
  - max_depth
  - min_child_weight
  - reg_alpha
  - reg_lambda
- 기타 파라미터
  - n_estimators
  - learning_rate
  - colsample_bytree

- 특히, 과적합이 우려되어서 관련 파라미터를 세심하게 튜닝하였다.
- Boosting 모형은 훈련시에 early_stopping을 통해서 과적합을 피하고자 하였다.

Modeling

결과

	조류	곤충	자서무척추동물	식물
고려한 종 개수 (수행한 모형 개수)	28	865	200	698
ROC AUC가 0.7 이상인 종 개수	10	80	71	61
ROC AUC 평균 값	0.86	0.77	0.77	0.78
ROC AUC 최대값	0.97 (섬휘파람새)	0.97 (제주왕딱정벌레)	0.99 (늪깔따구아과류)	0.97 (참식나무)

결과 분석

- 불균형 데이터인 종이 다수 있었기 때문에 결과를 비교하는 지표로는 ROC AUC 을 사용하였음
- 조류의 28개 종에 대한 모형의 ROC AUC 평균 값이 다른 종에 비해서 0.86으로 가장 높았음
- 조류, 곤충, 자서무척추동물, 식물별 가장 높은 ROC AUC는 0.97이 높을 정도로 테스트 데이터에 대해서 매우 정확한 성능을 보였음

Interpretation & Usage

Tree Interpreter을 이용한 해석 예시 1: 깔따구류 분석

※ 깔따구류 모형의 ROC AUC: 0.89

관심있는 종 및 지역 선택

- 깔따구류가 인천에 서식하는지 알고 싶음

해당 지역의 변수 생성

- 인천 그리드에 대한 생물 기후 변수를 추출함

서식 여부를 확률로 계산함

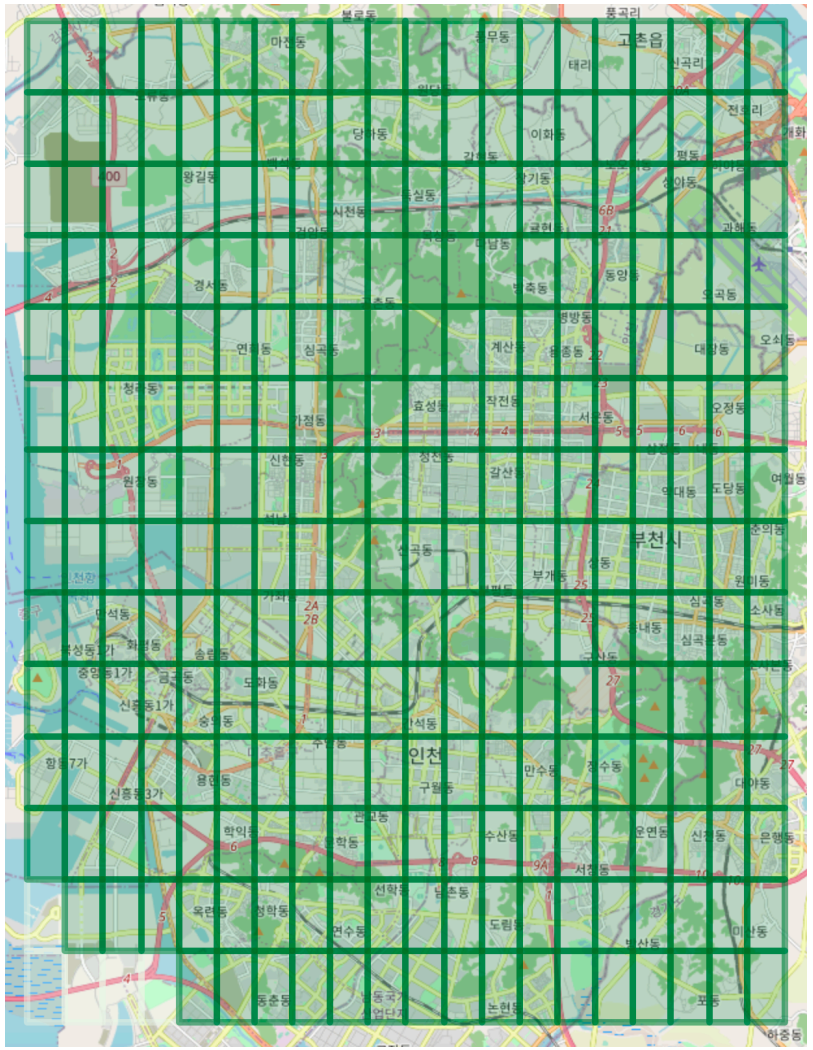
- 인천 그리드에 대해서 서식 여부를 계산한 결과, 확률이 높게 나옴

요인 분석

- 강수량, 습도 변수가 높은 영향을 미침을 밝힘

Interpretation & Usage

Tree Interpreter을 이용한 해석 예시 1: 깔따구류 분석



- 왼쪽 그림은 인천 지역 그리드에 대해서 깔따구류 서식 여부를 모형을 통해서 예측한 뒤, 확률을 초록색으로 시각화한 결과임
- 색이 모두 진한데, 높은 확률로 깔따구류가 서식한다고 해석할 수 있음
- 이와 같이 유해한 생물의 서식 여부를 미리 추론할 수 있다면 인천 수돗물 사태 등을 방지할 수 있을 것임

Interpretation & Usage

Tree Interpreter을 이용한 해석 예시 2: 섬휘파람새 분석

※ 섬휘파람새 모형의 ROC AUC: 0.97

관심있는 종 및 지역 선택

- 섬휘파람새가 전국적으로 어떻게 서식하는지 알고 싶음

해당 지역의 변수 생성

- 탐사되지 않은 지점에 대한 생물기후 변수를 추출함

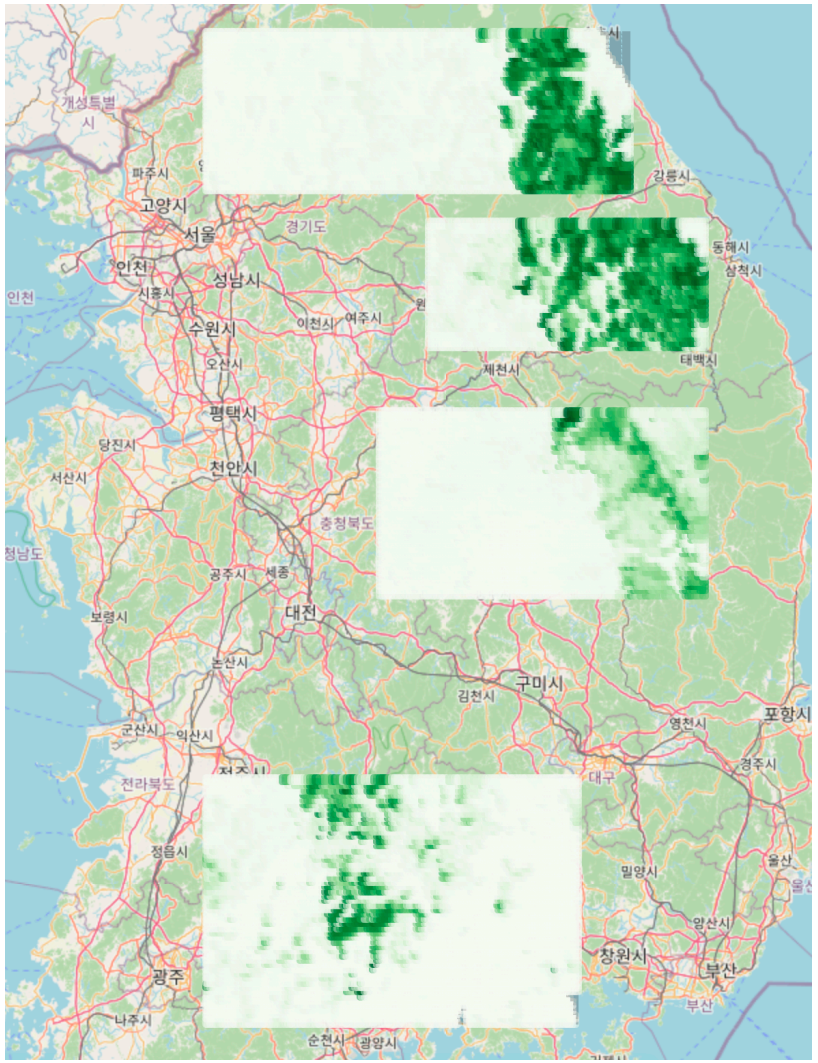
서식 여부를 확률로 계산함

- 각 그리드에 섬휘파람새가 서식할 확률을 계산함

요인 분석

- 시즌별 최저 기온, 태양광 등의 변수가 높은 영향을 미침을 밝힘

Tree Interpreter을 이용한 해석 예시 2: 섬휘파람새 분석



- 왼쪽 그림은 탐사되지 않은 지역에 섬휘파람새가 서식할 확률을 시각화한 것임
- 색이 진할수록 섬휘파람새가 서식할 확률이 높음을 의미함
- 이와 같이 생물 종이 특정 지역에 서식할 확률을 구할 수 있다면, 생태학자가 확률이 높은 지역을 중심으로 탐사하여 효율적인 연구를 진행할 수 있음

Chapter 5

한계 및 의의

envbigdata

한계

멸종 위기 동물에 대한 아쉬움

- 본 연구의 주요 목표 중 하나는 멸종 위기 동물의 서식지 분포를 추론하는 모형도 생성하여, 생태계 질서를 올바르게 유지하는데 기여를 하는 것이었음
- 하지만 멸종 위기 동물이 서식한다고 관측된 장소는 매우 적었으며 그에 따라서 엄청난 불균형의 데이터가 생성됨 (99.99% vs 0.01% 정도)
- 또한 데이터의 개수가 많은 것도 아니라서 RF, XGB로 적절한 정확도를 산출하는 모형을 만들어낼 수 없었음

한계

예측 단위에 대한 아쉬움

- 본 연구의 기본 예측 단위는 그리드(1km by 1km)이기 때문에 이 그리드에 관심 있는 생물 종이 서식하는지 여부를 중심으로 논의가 전개됨
- 이는 앞서 언급했듯이, 한번의 탐사가 진행될 때, 모두 동일한 위도 및 경도로 생물 종의 관측 여부가 기록되기 때문에 좌표(위도 및 경도) 단위로 분석을 진행하기에는 무리가 있다고 판단됨
- 만약 생태 데이터가 위도, 경도로 관측이 되었다면 본 연구의 기본 단위도 위도, 경도일 것임
- 만약 기본 단위가 위도, 경도라면 시공간 통계학의 Point Process을 사용해야하며 이를 통해 더 정확한 분석을 할 수 있을 것으로 보임

의의

국민의 삶의 질 증진

- 앞서, 깔따구류와 같은 유해 생물이 인천에 서식할 확률을 계산해 보았음
- 깔따구류뿐만 아니라 곱등이와 같이 유해한 생물의 서식 확률을 계산하여 인천 수돗물 사태와 같이 국민의 삶을 저해하는 사건을 미연에 방지할 수 있을 것임

의의

생태학자의 효율적인 연구

- 본 연구를 통해 관심있는 종이 특정 지역에 서식할 확률을 계산할 수 있음
- 이를 통해, 생태학자가 연구하고자 하는 생물이 특정 지역에 서식할 확률을 계산함으로써 생태학자가 좀 더 효율적인 연구를 진행할 수 있을 것으로 기대됨

감사합니다