머신러닝을 활용한 동물 멸종 위기 등급 예측

한상현

I. 서론

1.1 연구의 필요성

수많은 동물의 멸종 위기 단계를 일일이 추적하고, 단계 변화 시기를 종별로 연구하여 예측하는 것은 시간이 오래 걸릴뿐더러 사람의 손길이 닿지 않는 공간에는 미처 파악하지 못하는 종들이 생겨날 수도 있다. 하나의 종을 연구하고 추적하는 데 걸리는 시간은 짧게는 한 달, 길게는 몇 개월이 걸린다고 한다.

조사를 진행하는 도중에도 개체 수는 변화할 것이고, 멸종 위기 동물들에 대해서는 지속적이고 빠른 조사를 기반으로 등급이 매겨져야 한다. 본 연구는 현장 조사 혹은 관찰 이외에도 활용할 수 있는 데이터를 이용하여 멸종 위기종을 발견해 냄으로 동물 조사의 시간적/공간적 비용 감소를 기대한다.

1.2 멸종 위기 등급 현황

세계자연보전연맹(IUCN)은 지구상의 모든 개체를 아래 6개 등급으로 분류하려 한다.

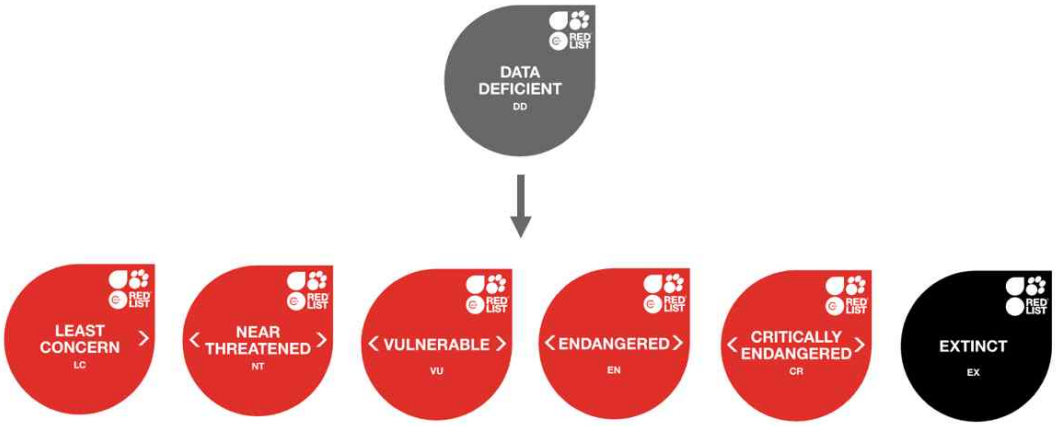
절멸 (Extinct)
위급 (Critically Endangered)
위기 (Endangered)
취약 (Vulnerable)
준 위험 (Near Threatened)
관심 대상 (Least Concern)

위에서부터 아래 순으로 멸종 위기에서 멀어지는 것이다. 하지만 앞서 말한 것과 같이 하나의 종을 연구하고 추적하는 데 걸리는 시간과 비용, 그리고 환경적 불가능성 때문에 “정보 부족 (Data Deficient)”이라는 등급이 있다. 조사를 미처 하지 못했거나, 불가피한 이유로 개체 수 파악이 불가능할 때, 위 6개의 등급에 속하지 않는 개체들에 부여되는 등급이다. 쉽게 말해, “아직은 파악되지 않았음”이라는 의미로 해석될 수 있다.

1.3 연구 목표

국제 자연 보전 연맹(IUCN)이 분류하는 멸종 위기 단계 중 정보 부족 (Data Deficiency)에 속한 종들은 단순히 정보가 적다는 이유로 분석이 불가하다는 판단이

내려진다. 예를 들어, 범고래 또한 분명 지구상에 살아 숨 쉬고 존재하고 있지만, 개체 수가 파악되지 않아 정보 부족군으로 분류되고 있다. 멸종 위기 범주가 아니므로, 이 범주에 속한 분류군은 보전의 필요성이 높다 하더라도 보전조치의 명백한 대상이 아닐 수 있다. 단순히 개체의 관찰된 수와 서식지에 의존하여 정보가 부족하다고 판단하는 기존의 방법과는 달리 본인은 개체 수에 대한 정보가 부족하여도 사용할 수 있는 동물의 생물학적, 사회학적 정보를 변수로 사용하였다. 단순 개체 수 추적이 아닌 종의 상위분류, 하위분류, 현재 서식지 지형, 생리학적 정보, 인간의 개입 등 다양한 변수들을 토대로 머신러닝을 사용하여 정보 부족 종들의 멸종 단계를 예측하였다.
다시 말해, 본 연구를 통해, 아직 정보 부족으로 남아있는 개체들에게 6개 등급 중 하나를 부여하는 작업을 수행하였다.



<Picture 1. 연구 목표>

1.4 선행연구

지난 2018년 12월, 오하이오 주립대학에서 공동 연구된 자료에서 본 연구와 유사한 접근법을 보여준 바 있다. 그들은 150,000개 이상의 식물 종의 보존상태를 예측하는데 머신러닝 기법을 활용하였다. 전 세계적으로 식물의 종이 매우 많음에도 불구하고 현재 보존상태가 정확히 알려진 종은 매우 적다. 따라서 대부분의 식물 종의 보존상태가 미확인된 상태였고, 관리가 필요한 '위험' 군의 종이 제대로 파악되지 않고 있다는 문제가 제시되고 있었다. 오하이오 주립대학 연구팀은 서식지 특징, 지역 날씨 패턴, 물리적 변수 등을 활용하여 종을 위기에 빠뜨릴 패턴을 추가하였다. 결과적으로 그들은 머신러닝 기법을 활용하여 지역별 멸종위험지수를 적합하여 전 세계 식물 종 위험 지도를 제시하면서 고위험 종을 효과적으로 파악할 수 있다는 결론을 도출하였다.

본 연구는 전 세계 포유류 종을 대상으로 데이터를 활용하여 종의 보존상태를 예측하는 모델을 구축한다는 측면에서 위 선행연구와 비슷한 맥락을 가진다.

Ⅱ. 데이터

2.1 데이터 수집

데이터는 IUCN Redlist 공식 사이트에 게시되어 있는 정보를 웹 크롤링하여 수집한다. (https://www.iucnredlist.org/search) 사이트에는 지구상 모든 개체의 멸종 위기 등급과, 생물학적 종 구분, 서식지, 위협요소, 종 보존을 위한 노력, 주 서식 국가 등이 포함되어 있다. 본인은 연구를 위해 직접 크롤링을 통해 데이터를 수집하였다.

수집된 변수들에 대한 설명은 다음과 같다.

Phylum: 생물 분류 — 문 (e.g. 연체동물, 척삭동물, 완족동물 등)

Class: 생물 분류 — 강 (e.g. 포유류, 양서류, 파충류 등)

Range: 서식 나라 (e.g. 대한민국, 미국, 페루 등)

Population Trend: 인구 변화 추세 (Increasing, Decreasing, Stable)

System: 육지/민물/바다 생물 구분

Habitat: 서식 환경 (e.g. 동굴, 숲, 심해, 초원 등)

Threats: 위험 종류 (e.g. 지구온난화, 산림 파괴, 질병, 환경 오염 등)

Use Trade: 인간의 활용 분야 (e.g. 음식, 사용 없음, 제약 등)

Conservation Action: 보호 정책 종류 (e.g. 교육, 실질적 보호, 연구 등)

2.2 변수 전처리

수집한 데이터는 모두 범주형 변수들로 이루어져 있다. 그러나 이를 머신러닝의 학습데이터로 입력할 때에는 반드시 숫자의 형태로 바꾸어 주어야 한다. 이를 위해 가변수를 만들어 사용하였다. 만약 해당 범주에 속하면 1, 아니면 0인 형태로 데이터를 변경해 주었다. 또, 한 범주변수 내에 level이 너무 많다면 희귀 level을 Others로 통합하여 주었다. 예를 들어 3개의 level을 가지는 System 변수의 경우 다음과 같이 변형된다.

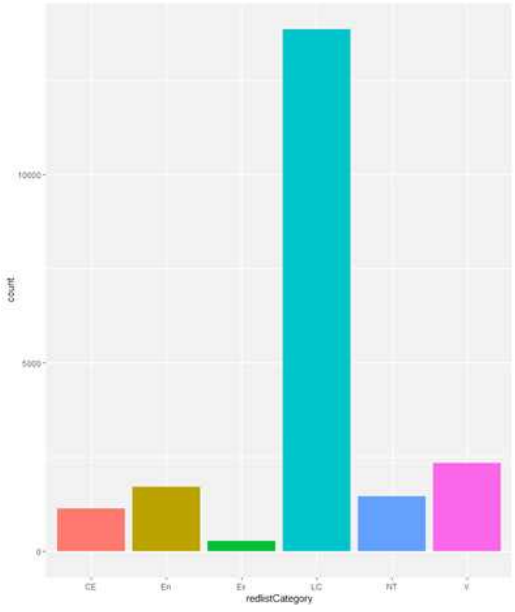
전)

동물	Systems
A	Freshwater
B	Marine
C	Marine, Terrestrial

후)

동물	Freshwater	Marine	Terrestrial
A	1	0	0
B	0	1	0
C	0	1	1

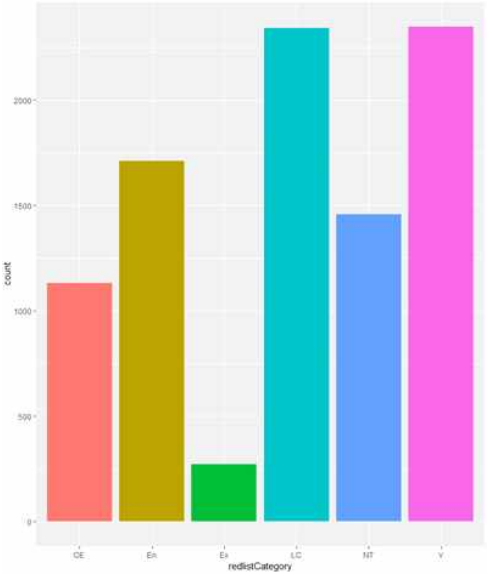
2.3 비대칭 데이터 전처리



<Picture 2. 기존 데이터 클래스 분포>

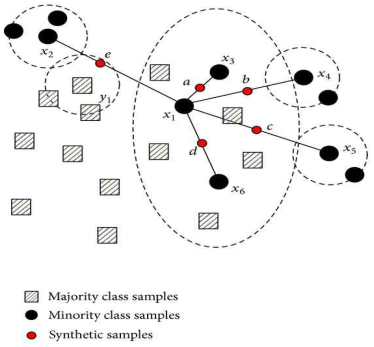
동물의 멸종 단계를 의미하는 반응변수 Redlist Category는 그 특성상 불균형할 수밖에 없다. 수집한 데이터에서 6개 등급 간의 개수 차이가 현저함을 확인했다. 일반적으로 이항 분류의 경우 threshold 조절로 비대칭 데이터 문제를 어느 정도 해결할 수 있지만, 다항 분류 문제에서는 이 방법을 사용할 수 없다. 또한, 단순히 소수 클래스를 복사하는 over-sampling이나 다수 클래스 중 일부만 사용하는 under-sampling을 사용할 경우 과적합 문제 혹은 데이터의 과도한 손실 등의 문제가 발생할 수 있다. 따라서 좀 더 체계적이고 심층적인 접근을 통해 비대칭 데이터 문제를 해결하였다.

일차적으로 가장 데이터 수가 많은 관심 대상 (Least Concern)의 경우 어느 정도 under-sampling 해주었다. 이를 통해 6개 Class 사이에 큰 개수 차이가 발생하지 않도록 한다.



<Picture 3. Under-sampling 이후 데이터 클래스 분포>

다음으로는 소수 클래스의 관측값을 보간하여 합성해내는 Synthetic Minority Over-sampling Technique (SMOTE) 기법을 사용하였다. 일반적인 over-sampling과 다르게 SMOTE는 k개의 이웃 간 거리 차를 구하여 그 차이에 0과 1 사이의 임의의 값을 곱하여 새로운 샘플을 만들어내는 형식이다. 즉 수가 적은 샘플들 (소수 클래스) 사이에서 임의로 새로운 샘플을 새롭게 뽑는 것인데, 이 과정에서 부트스트래핑과 KNN 방식이 사용되는 원리다. 단순히 적은 샘플들을 복제해내는 일반적인 over-sampling보다 편향으로 인한 과적합 문제가 덜 발생하고, 합리적인 인공데이터가 형성된다는 점에서 SMOTE를 사용하였다.

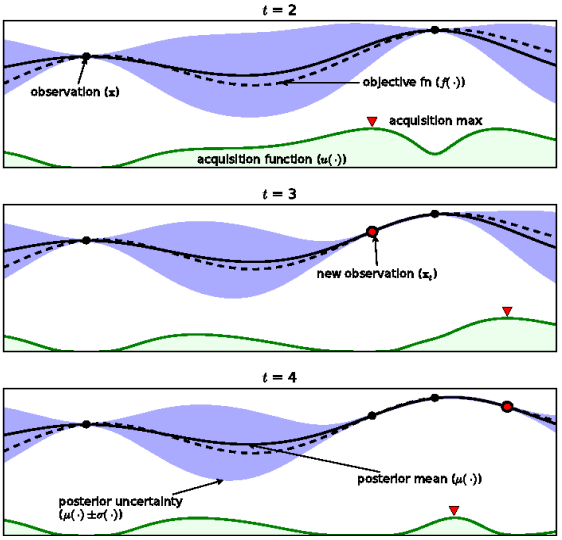


<Picture 4. SMOTE 알고리즘>

이때 SMOTE에는 over-sampling의 정도를 조절하는 하이퍼파라미터 (perc.over)가 있고 5개의 클래스에 적용하니 총 5개의 파라미터를 지정해 주어야 한다. 하지만 SMOTE의 알고리즘이 1 vs Many가 아닌 1 vs 1로 샘플을 생성해내기 때문에 가능

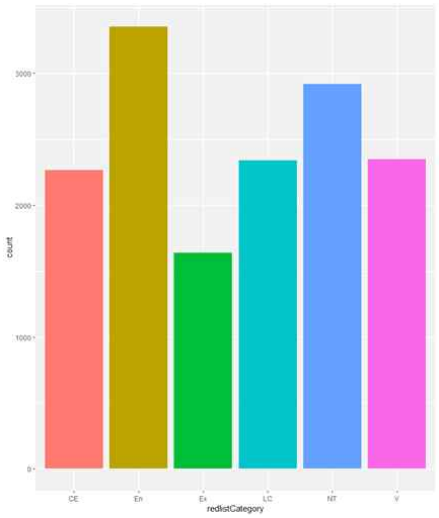
한 하이퍼 파라미터의 조합이 너무 많다는 문제가 있었다. 본인은 최적의 파라미터를 설정하기 위해 Bayesian Optimization을 적용하였다.

Bayesian Optimization은 머신러닝 모델들의 하이퍼파라미터를 튜닝하는 방법 중 하나로 Gaussian Process에 기반을 둔다. 불확실성이 큰 지점으로 이동하며 하이퍼파라미터에 따른 손실함수를 보다 효율적이고 확실하게 추정하는 알고리즘이다. 자세한 과정은 아래 그림을 참조한다.



<Picture 5. Bayesian Optimization 설명>

Bayesian Optimization을 거쳐 최적의 SMOTE 비율로 데이터를 재구성하여 본격적인 모델링을 위한 전처리 작업을 모두 마무리하였다.



<Picture 6. SMOTE 이후 데이터 클래스 분포>

Ⅲ. 예측모형

예측모형은 정보 부족군을 제외한, 6개 등급에 속한 개체들을 Train 데이터로 받아 학습을 시킨다. 총 6개의 머신러닝 기법을 비교하여 Cross-Validation Accuracy가 가장 높은 모형을 최종 모형으로 선정하였다. 비교에 사용된 모형은: Naive Bayes Classifier, Extreme Gradient Boosting, Neural Network, Random Forest와 K-nearest neighbour, CatBoost이다. 아래는 각 모델의 Cross-Validation Accuracy를 나타낸 표이다.
평가지표로는 accuracy를 사용한 이유는 어느 정도 비대칭 데이터 문제를 해결한 시점에서 굳이 AUC, F1 score를 사용하기보다 직관적인 accuracy를 사용하는 것이 적합했기 때문이다.

Model	NBC	XGBoost	NN	RF	KNN	CatBoost
Accuracy	0.3326	0.6673	0.6583	0.6610	0.6636	0.6011

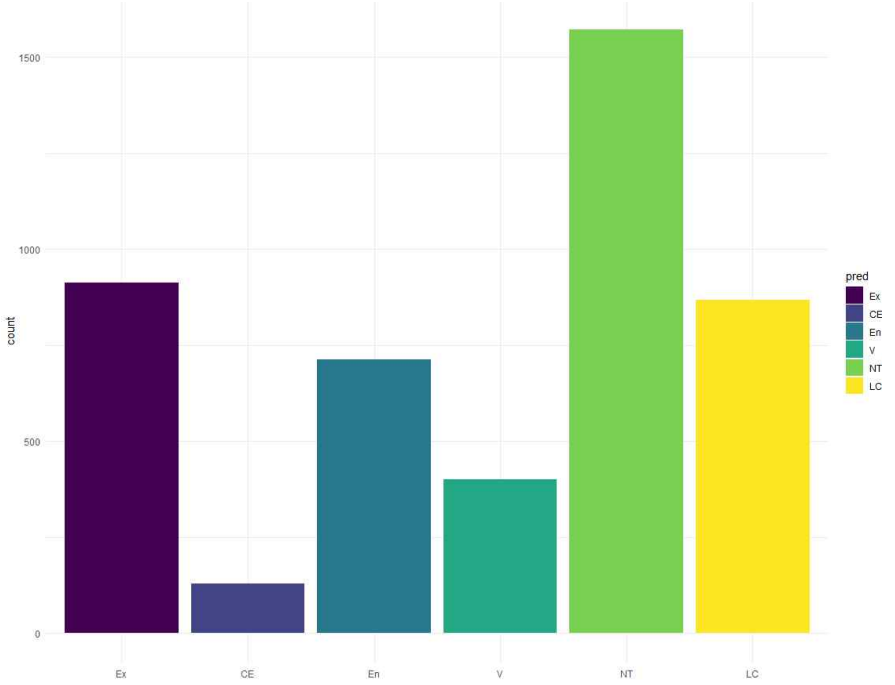
<Picture 7. 모델 Accuracy>

최종적으로 선정된 모델은 66.73%의 가장 높은 정확도를 보여준 Extreme Gradient Boosting 모형이었다. 해당 모형의 튜닝 파라미터로는 max_depth=5, min_child_weight=1, subsample=0.8, colsample_by_tree: 0.9를 사용하였다. 이 또한 Grid Search를 통해 찾은 최적의 파라미터 값이다.

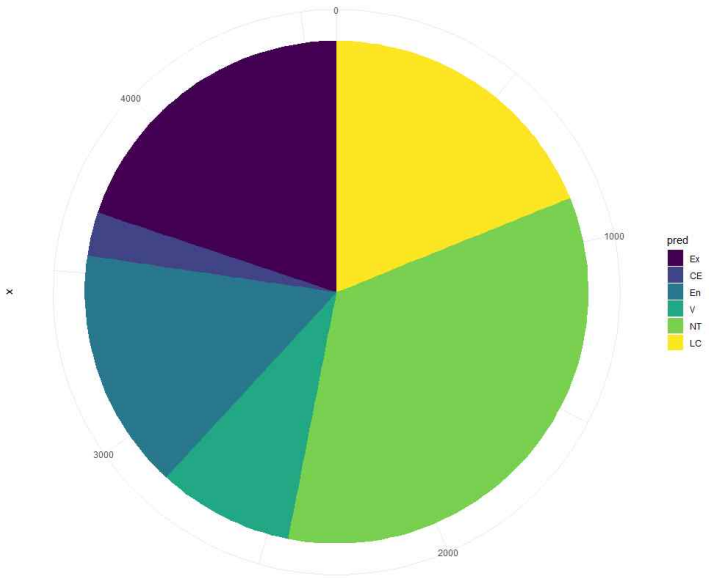
자원적 한계로 부족했던 데이터와 데이터 그 자체의 특성을 생각했을 때, 정확도 50%를 넘기는 것을 초기 목표로 설정했지만, 정밀한 변수 전처리와 비대칭 데이터를 해결하기 위한 노력, 그리고 세밀한 신경망 학습으로 좋은 성능의 모형을 적합 시킬 수 있었다.

Ⅳ. 결과

학습된 모형을 사용하여, 최종적으로 목표했던 정보 부족군을 Test 데이터로 두고 예측을 진행하였다.



<Picture 9. 정보 부족군 예측값 — bar plot>



<Picture 10. 정보 부족군 예측값 — pie chart>

위의 표에서 확인할 수 있듯이 대부분의 정보 부족군은 준 위험군으로 분류되었다. 그 뒤로 멸종 (Extinct), 관심 대상 (Least Concern), 위기 (Endangered), 취약 (Vulnerable), 위급 (Critically Endangered) 순으로 분류되었다. 이는 각각, 34.2%, 19.9%, 18.9%, 15.5%, 8.7%, 2.8%를 차지하고 있다.

이를 통해 알 수 있는 사실은, 정보 부족군으로 분류된 많은 개체가 실제로 멸종의 위기에 처해있고, 이를 해결하기 위해 해당 개체들에 대한 인간의 적극적인 보호 정책 수립이 필수적이다.

Ⅳ. 기대효과 및 활용

본인은 인공신경망을 사용하여 '정보 부족'으로 인해 분류되지 않았던 종들에 대해 종별로 멸종 위기 단계를 예측하였다. 이를 통해, 정보 부족군에 속한 개체들이 주로 어떤 멸종 위기 등급에 놓여있는지 파악할 수 있었다.

이에 그치지 않고, 본 예측값을 전문가들에게 전달하여, 새로운 멸종 위기 등급을 부여받은 '정보 부족' 관측치들 각각을 직접 조사하고, 개체 하나하나에 대한 보호 정책을 새롭게 구체적으로 수립할 수 있을 것이다. 이것이 이 분석의 가장 최종적이고 이상적인 목표라고 생각한다.

추가로, 시민과 연구자들의 현장연구를 토대로 더 심화한 데이터들을 모아 각 개체의 서식지 이동, 먹이 변화 등을 추적하는 연구를 진행할 수 있다면, 본 분석에서 세웠던 모형의 정확도와 성능이 향상될 것이며, 이에 따라 더 효과적인 종 보전 정책을 수립할 수 있을 것으로 기대된다. 이는 단순히 멸종 위기 동물을 생포하여 동물원에 가두고 보호하는 방법보다는 훨씬 인도적이며 생태를 보전할 방법이며, 향후 멸종 위기 등급 선정에서도 많은 도움을 줄 것으로 기대된다.