Linear & genralized linear model

Modeling the presence and absence of island scrub jay

김수연

CHAPTER

Dataset

2008년 가을 동안 캘리포니아 산타 크루즈 섬에서 307개 조사 지점에서 발견된 island scrub jay에 대한 데이터셋

- 데이터셋 크기: 5625 개

- 변수: island scrub jay의 존재 여부 (isj=1: 존재, isj=0: 부재), x위치, y 위치, 고도, forest 비율, chaparral 비율

- NA가 존재하는 데이터 개수(행): 5322 개

분석 목표

Goal1. 데이터셋에 주어진 변수를 바탕으로 Scrub jay가 선호하는 지형 유형 이해하기

Goal2. 다음과 같은 2개의 모델을 사용하여 섬 전체에서 scrub jay의 존재 여부 예측하기
기존 변수는 그대로 사용하고 그 변수를 활용하여 새로운 변수를 추가로 정의하였을 때,
1) 새로운 변수 중 오분류율을 가장 낮게 하는 변수의 조합과 기존 변수를 사용한 모델
2) 기존 변수와 새로운 변수 중 오분류율을 가장 낮게 하는 변수의 조합을 사용한 모델

변수 설정

Target variable: island scrub jay의 존재 여부 (isj=1: 존재, isj=0: 부재)

Predictor variable: x, y 위치, 고도, forest 비율, chaparral 비율

반응변수 NA 처리

	isj	x 위치	y 위치	고도	forest 비율	chaparral 비율
NA 행	5318	0	0	2838	2838	2838

-반응변수 isj에서의 NA가 존재하는 데이터는 총 5318개이므로 데이터셋의 95%가 NA값이다. 이때, 반응변수 isj에서의 NA값은 해당 위치에서 scrub jay에 대한 데이터가 수집되지 않았음을 의미한다.

- 해당 데이터셋에 대한 추가 정보를 찾을 수 없으므로 해당 변수에 값을 대체하는 방법은 사용하지 않는다. 즉, train dataset에 반응변수가 NA인 데이터(행)은 사용하지 않는다.

- 최종 모델을 train dataset에 사용되지 않은 2484개의 dataset에 적용하여 scrub jay의 존재여부를 예측하고자 한다. 이때, 2484개의 dataset는 반응변수 isj만 NA값이고 나머지 변수의 값은 NA가 아닌 dataset이다.

설명변수 NA 처리

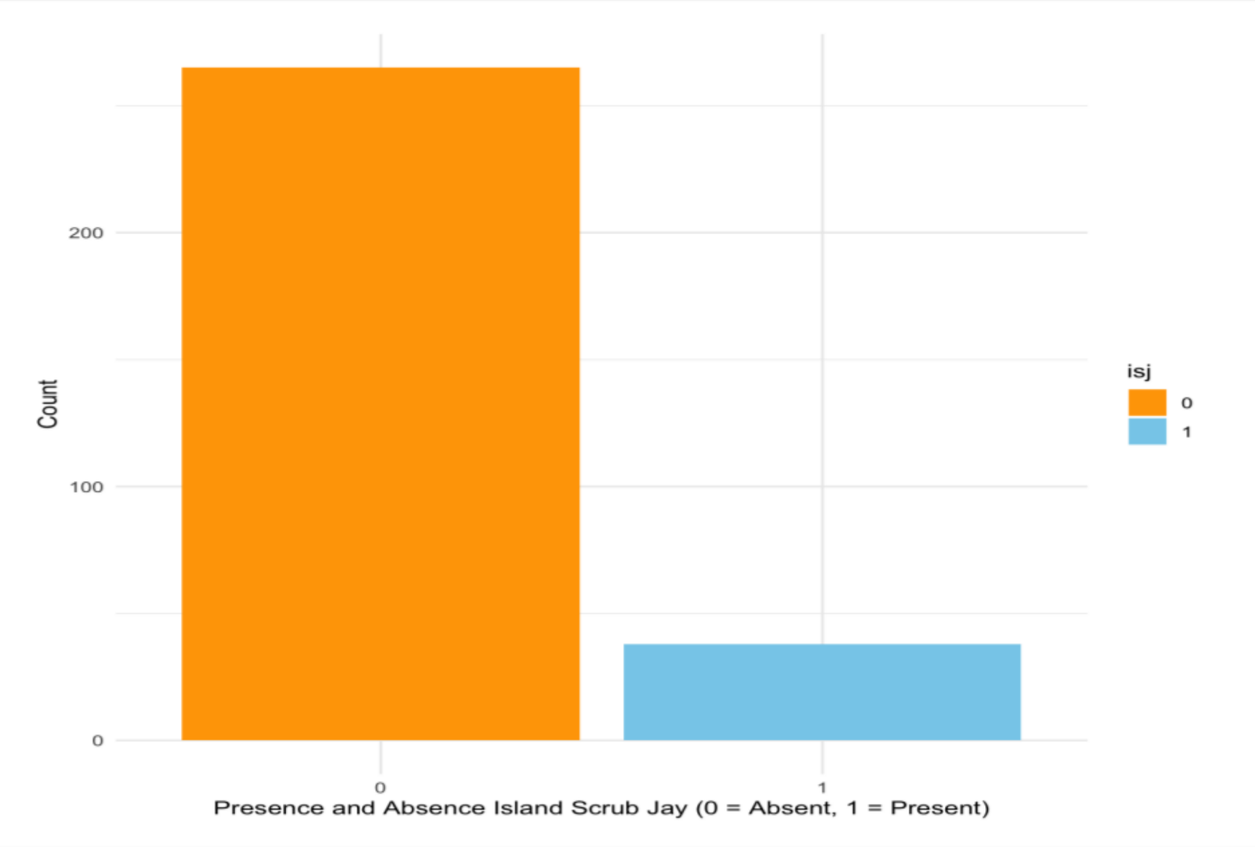
	isj	x 위치	y 위치	고도	forest 비율	chaparral 비율
NA 행	0	0	0	4	4	4

반응변수 isj에서의 NA가 존재하는 데이터를 삭제한 후, 다시 NA가 존재하는 데이터(행) 수를 확인하면 위의 표와 같다.

- 변수 고도가 존재하지 않으면 변수 forest 비율, chaparral 비율에 대한 데이터도 존재하지 않는다.

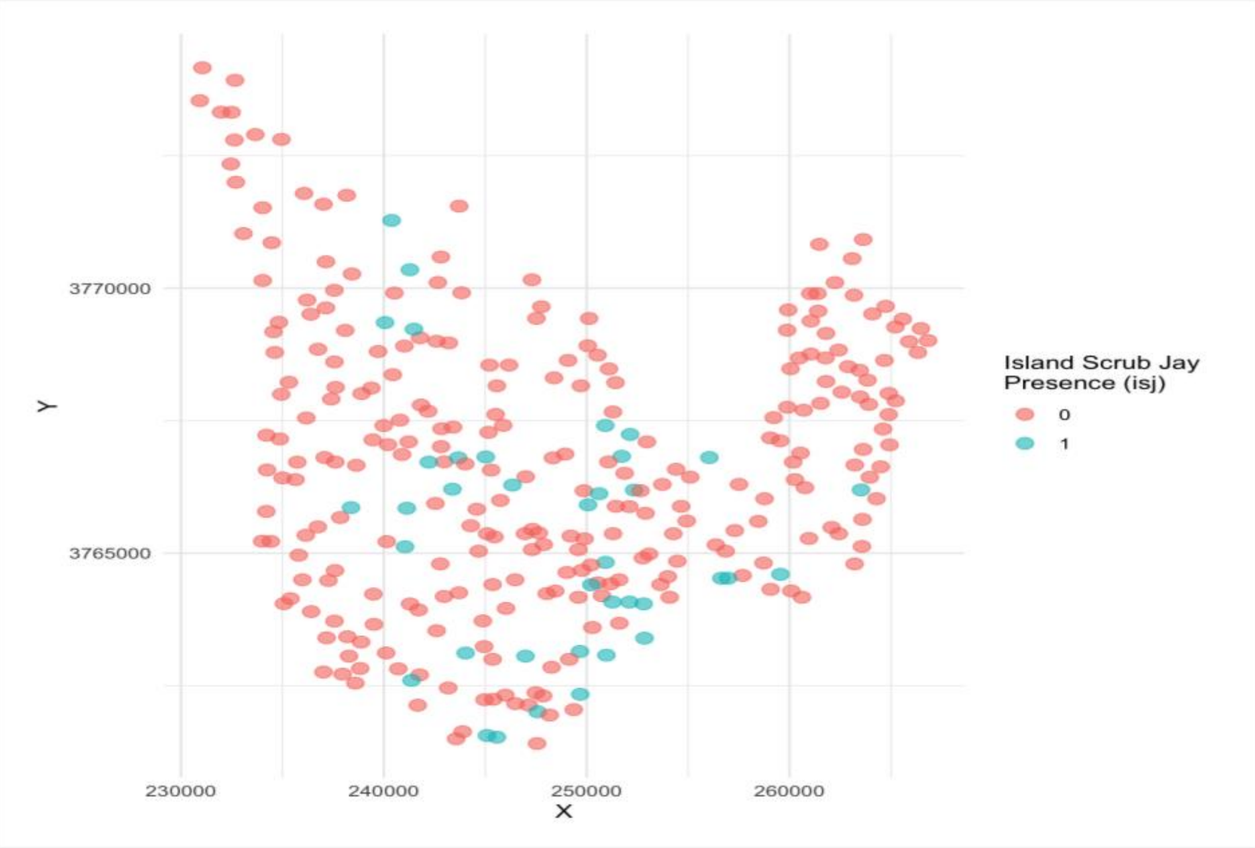
- 총 303개의 데이터 중 약 1%가 NA가 존재하는 데이터이므로 train dataset에 해당 데이터는 사용하지 않는다.

반응변수 시각화



island scrub jay의 존재하지 않는 데이터가 존재하는 데이터보다 더 많다.

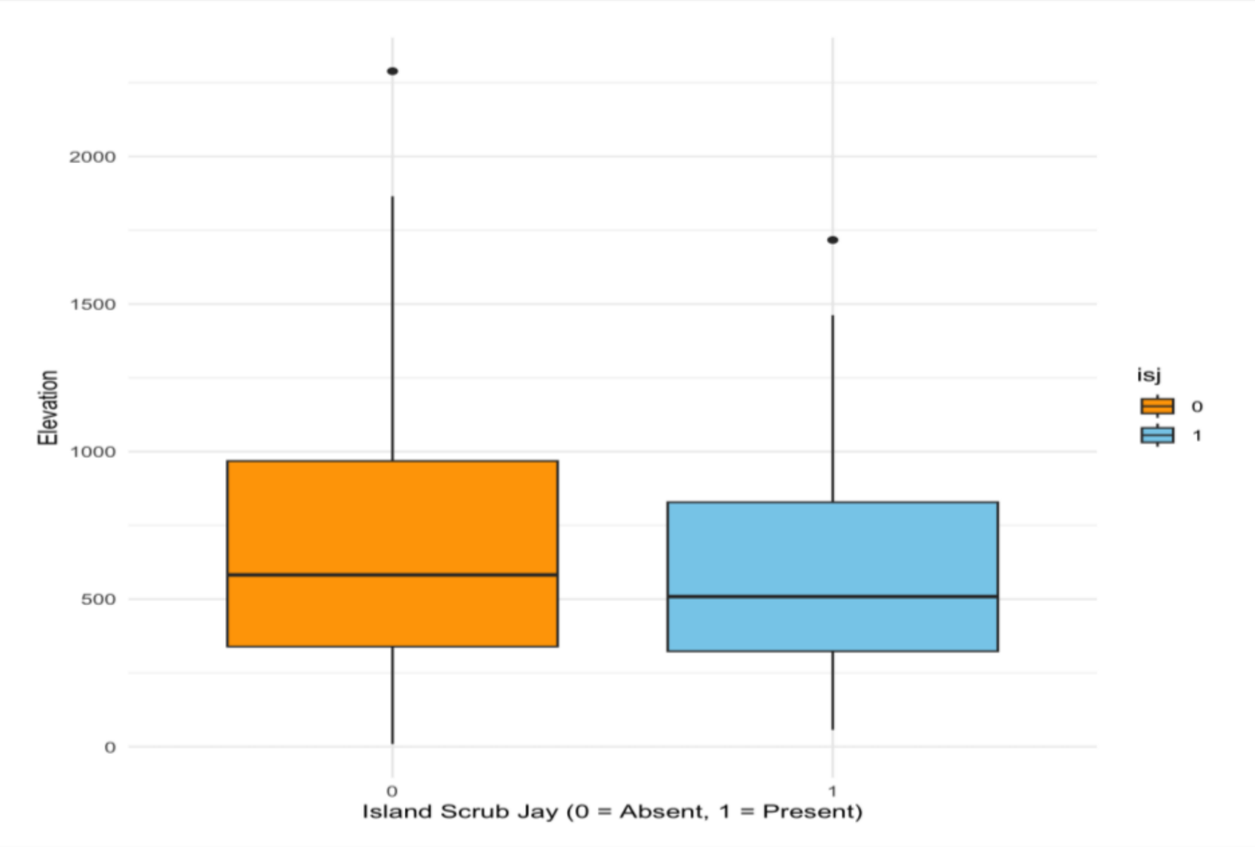
설명변수 x,y위치 시각화



[scrub jay 없는 지역 (isj=0)]
- X, Y 좌표 전역에 고르게 분포한다. 즉, 특정 위치에 집중되어있지 않다.

[scrub jay 존재하는 지역 (isj=1)]
- 특정 위치에서 집중적으로 나타난다. 특히, X 좌표가 약 240,000~250,000, Y 좌표가3765000~3770000 사이에서 집중되어 있다.

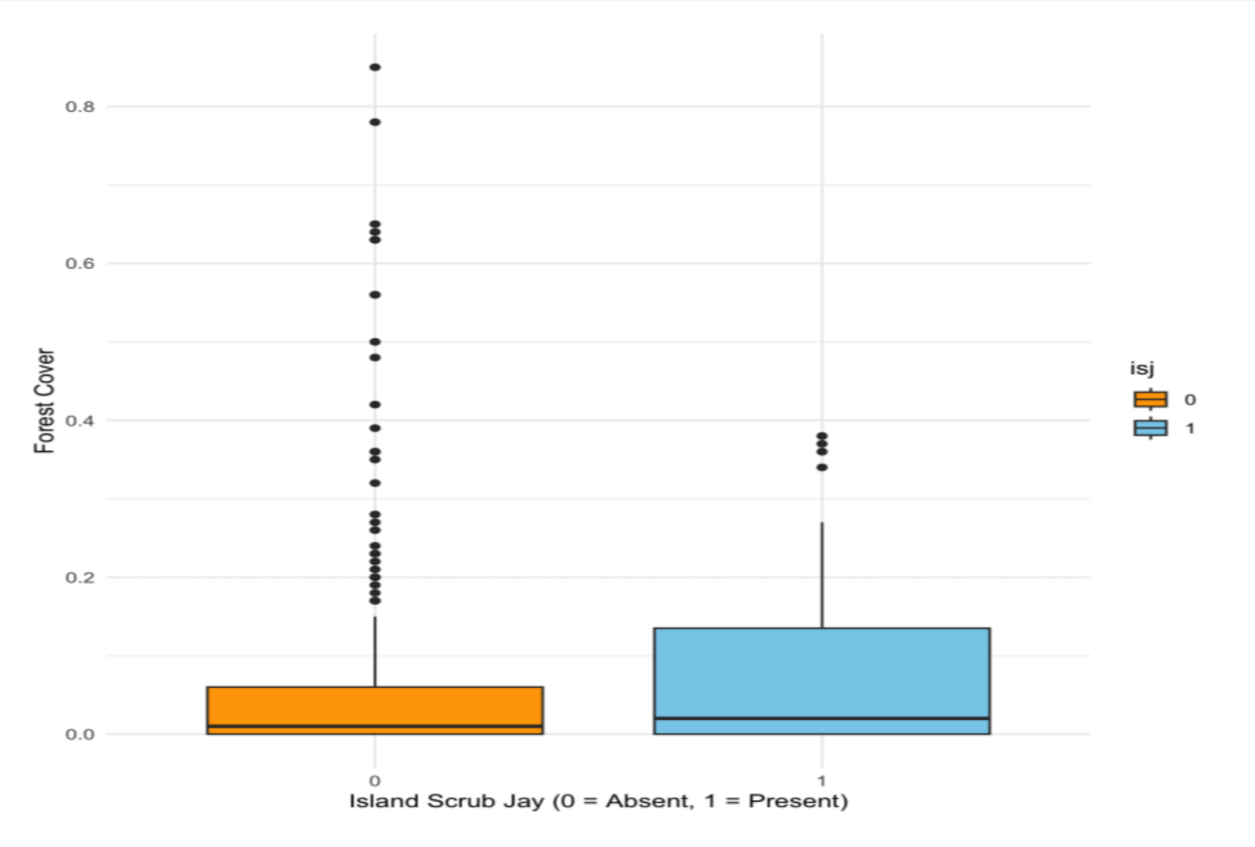
설명변수 elev 시각화



[scrub jay 없는 지역 (isj=0)]
- 고도 분포가 넓고 일부 이상치(outlier)가 존재한다.

scrub jay 존재하는 지역 (isj=1)]
- scrub jay가 존재하지 않는 지역보다 분포 폭이 더 좁다.

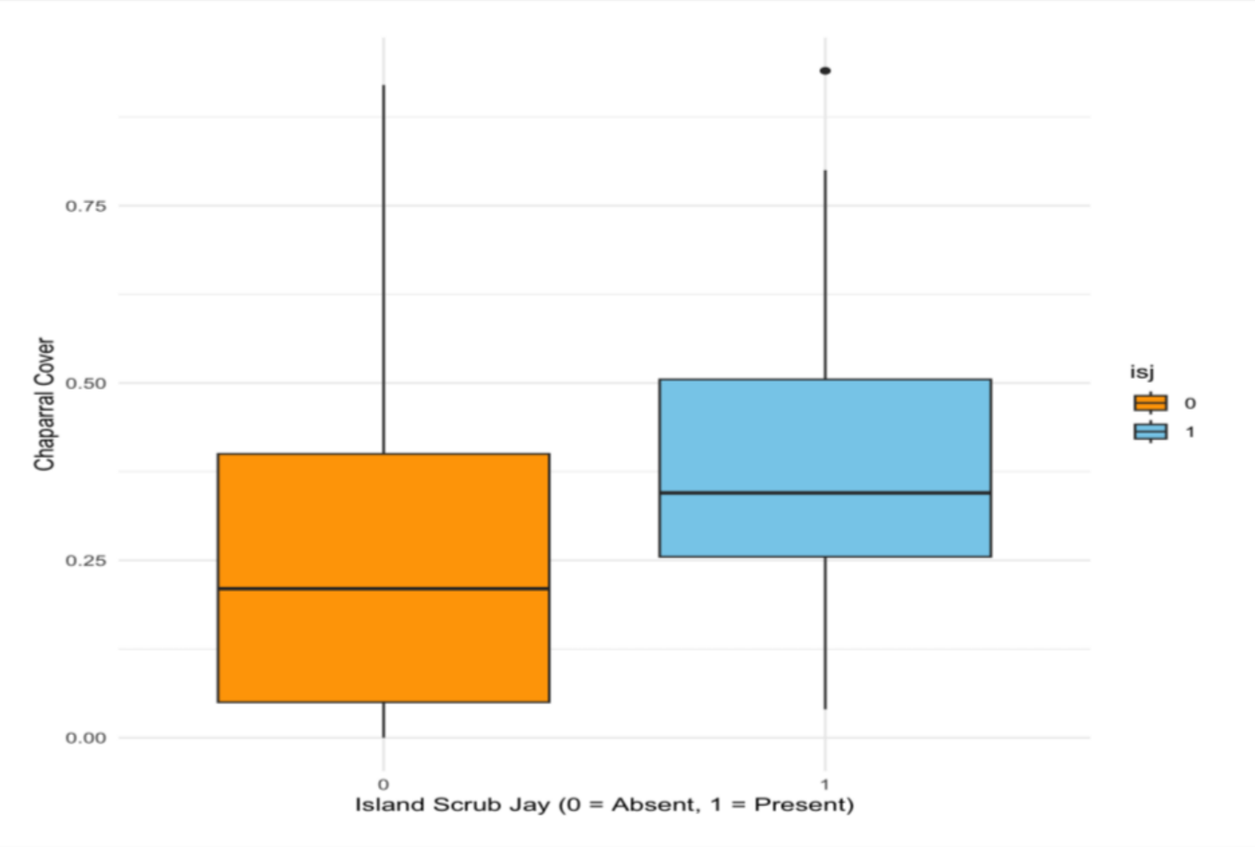
설명변수 forest 시각화



[scrub jay 없는 지역 (isj=0)]
- forest 비율은 대부분 0~0.1 사이에 분포하며, 중앙값은 약 0.05로 매우 낮다.
- forest 비율이 0.6~0.8까지 높은 곳도 존재한다.

[scrub jay 존재하는 지역 (isj=1)]
- forest 비율이 더 넓게 분포한다.
- Outlier가 거의 없다.

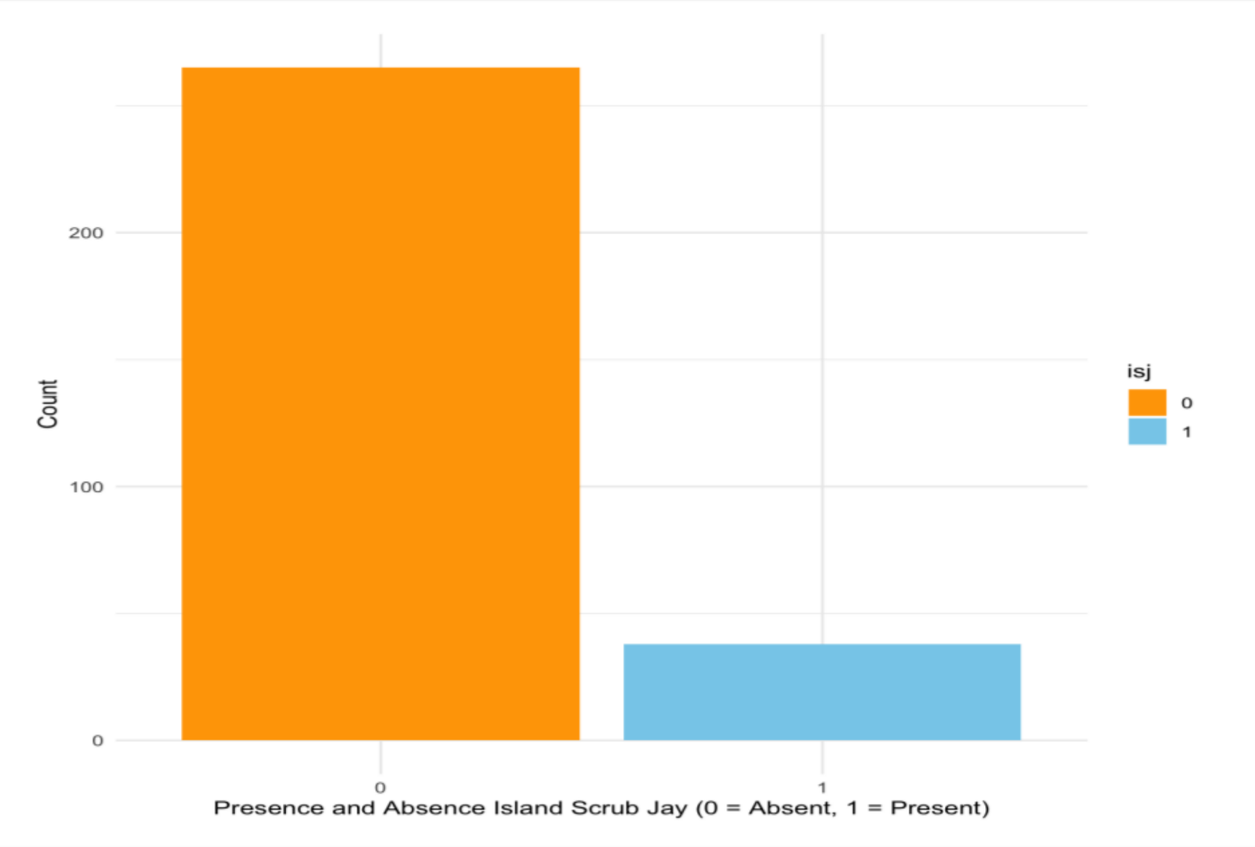
설명변수 chap 시각화



[scrub jay 없는 지역 (isj=0)]
- 대부분의 데이터가 약 0.1~0.45 사이에 분포한다.
- Chaparral 비율이 낮은 지역에서 scrub jay가 없다.

[scrub jay 존재하는 지역 (isj=1)]
- 중앙값이 scrub jay가 존재하지 않는 지역에 비해 더 높은 값을 갖는다.
- Chaparral 비율이 약 0.6~0.7와 같은 높은 값에서도 관찰된다.
- Outlier가 하나 존재하는데 Chaparral 비율이 매우 높은 지역에서도 scrub jay가 존재할 수도 있다고 할 수 있다.

반응변수 시각화



island scrub jay의 존재하지 않는 데이터가 존재하는 데이터보다 더 많다.

-> 사용할 데이터 303개 중, isj=1인 데이터의 개수는 38로 데이터가 편중되어 있음을 알 수 있다. 이때 반응변수가 1 또는 0이므로 binary regression으로 probit regression과 logistic regression을 사용할 수 있다. 설명변수와 반응변수 간의 관계 해석을 쉽게 하기 위해 logistic regression을 사용하겠다.

사용할 모델

Logistic regression을 사용하고자 한다. 그 이유는 다음과 같다.

1) 반응변수가 1 또는 0이므로 binary regression을 사용한다. Probit regression도 binary regression에 속하지만, logistic regression을 사용하는 경우가 변수 해석을 더 용이하게 하므로 해당 모델을 사용한다,

2) Scrub jay가 선호하는 지역의 유형을 파악하기 위해 가설 검정을 활용하고자 한다. 즉, 가설 검정을 통해 설명변수가 반응변수에 유의미한 영향을 미치는지를 확인하기 위해 해당 모델을 사용한다.

분석 흐름

1. Raw dataset에 polynomial과 intersection을 추가함으로써 기존의 변수를 활용하여 새로운 변수를 정의한다. 이때 변수 간의 상호작용과 비선형성을 고려하기 위해 새로운 변수를 사용하고자 한다.

2. Train data와 Test data을 7:3 비율로 나눈다.

3-1. 기존 설명변수와 새롭게 정의된 설명변수들 중에서 오분류율을 가장 적게 만들어주는 변수의 조합(기존 설명변수는 그대로 사용)을 찾기 위해 for문을 사용한다. 즉, train data로 모델을 적합한 후, test data에 대한 오분류율을 구한다.

4-1. 가설 검정을 진행하여 반응변수에 유의미한 영향을 미치는 설명변수 찾는다.

3-2. 기존 설명변수와 새롭게 정의된 설명변수들 중에서 AIC가 가장 작은 변수의 조합(기존 설명변수는 그대로 사용X)을 찾기 위해 for문을 사용한다. 즉, train data로 모델을 적합한 후, test data에 대한 오분류율을 구한다.

4-2. 가설 검정을 진행하여 반응변수에 유의미한 영향을 미치는 설명변수 찾는다.

5. 3-1과 3-2에서 정의한 모델 중 오분류율이 가장 낮은 모델의 결과를 시각화하고 해당 모델을 앞서 설명한 2484개의 dataset에 적용한다.

Step1. 새로운 변수 정의

Raw dataset에 polynomial과 intersection을 추가함으로써 기존의 변수를 활용하여 다음과 같이 15개의 새로운 변수를 정의한다.

변수명	x2	y2	elev2	forest2	chap2	x.y	x.elev
변수 정의	x^2	y^2	$elev^2$	$forest^2$	$chap^2$	x*y	x*elev

변수명	x.forest	x.chap	y.elev	y.forest	y.chap	elev.forest	elev.chap	forest.chap
변수 정의	x*forest	x*chap	y*elev	y*forest	y*chap	elev*forest	elev*chap	forest*chap

Step2. Train, test dataset

Train data와 Test data을 7:3 비율로 나눈다.

Step3. 최적의 변수 조합 찾기

303개 중, isj=1인 데이터의 개수는 38로 데이터가 데이터가 편중되어 있음을 알 수 있다.
scrub jay가 있는 곳을 더 많이 찾아내고자 하므로 threshold를 0.5가 아닌 0.3으로 조정한다. 이에 따라 scrub jay가 없는 곳을 있는 곳이라고 판단할 오류는 더 커지겠지만, 놓치고 있던 장소를 발견함으로써 얻는 가치는 이를 감수할 수 있을 것이라고 생각하므로 threshold를 0.3으로 한다.

Step3-1. 최적의 변수 조합 찾기

설명변수들 중에서 오분류율을 가장 적게 만들어주는 변수의 조합을 찾으면 다음과 같다. 이때, 오분류율은 0.1318681이다.

1) x 위치

2) y 위치

3) 고도 (변수 이름: elev)

4) forest 비율 (변수 이름: forest)

5) chaparral 비율 (변수 이름: chap)

6) 변수 x와 변수 forest의 교호 작용 (변수 이름: x.forest)

7) 변수 forest와 변수 chap의 교호 작용(변수 이름: forest.chap)

Step4. 가설 검정

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.557400   0.286765  -8.918   <2e-16 ***
x           -0.002711   0.281697  -0.010   0.9923
y           -0.873567   0.357028  -2.447   0.0144 *
elev        -0.171497   0.232909  -0.736   0.4615
forest      -3.657522   6.020581  -0.608   0.5435
chap         0.353577   0.183035   1.932   0.0534 .
x.forest     3.775835   5.935980   0.636   0.5247
forest.chap  0.159482   0.244024   0.654   0.5134
```

유의한 변수에 대한 가설검정만을 설명하겠다.
H0: 변수 y는 반응변수에 유의미한 영향을 미치지 않는다.
H1: 변수 y는 반응변수에 유의미한 영향을 미친다.

-> 유의수준 0.05에서 p-value는 0.0144이므로 귀무가설을 기각한다. 즉, 변수 y는 반응변수에 유의미한 영향을 미친다.

=> 설명변수 y를 제외한 모든 변수는 반응변수에 유의미한 영향을 미치지 않는다.

Step3-2. 최적의 변수 조합 찾기

설명변수들 중에서 AIC가 가장 작은 변수의 조합을 찾으면 다음과 같다. 이때, 오분류율은 0.1428571이다.

1) y 위치

Step4. 가설 검정

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3171     0.2397  -9.665  < 2e-16 ***
y            -0.8681     0.2875  -3.020  0.00253 **
---
```

유의한 변수에 대한 가설검정만을 설명하겠다.
H0: 변수 y는 반응변수에 유의미한 영향을 미치지 않는다.
H1: 변수 y는 반응변수에 유의미한 영향을 미친다.

-> 유의수준 0.05에서 p-value는 0.00253이므로 귀무가설을 기각한다. 즉, 변수 y는 반응변수에 유의미한 영향을 미친다.

=> 설명변수 y를 제외한 모든 변수는 반응변수에 유의미한 영향을 미치지 않는다.

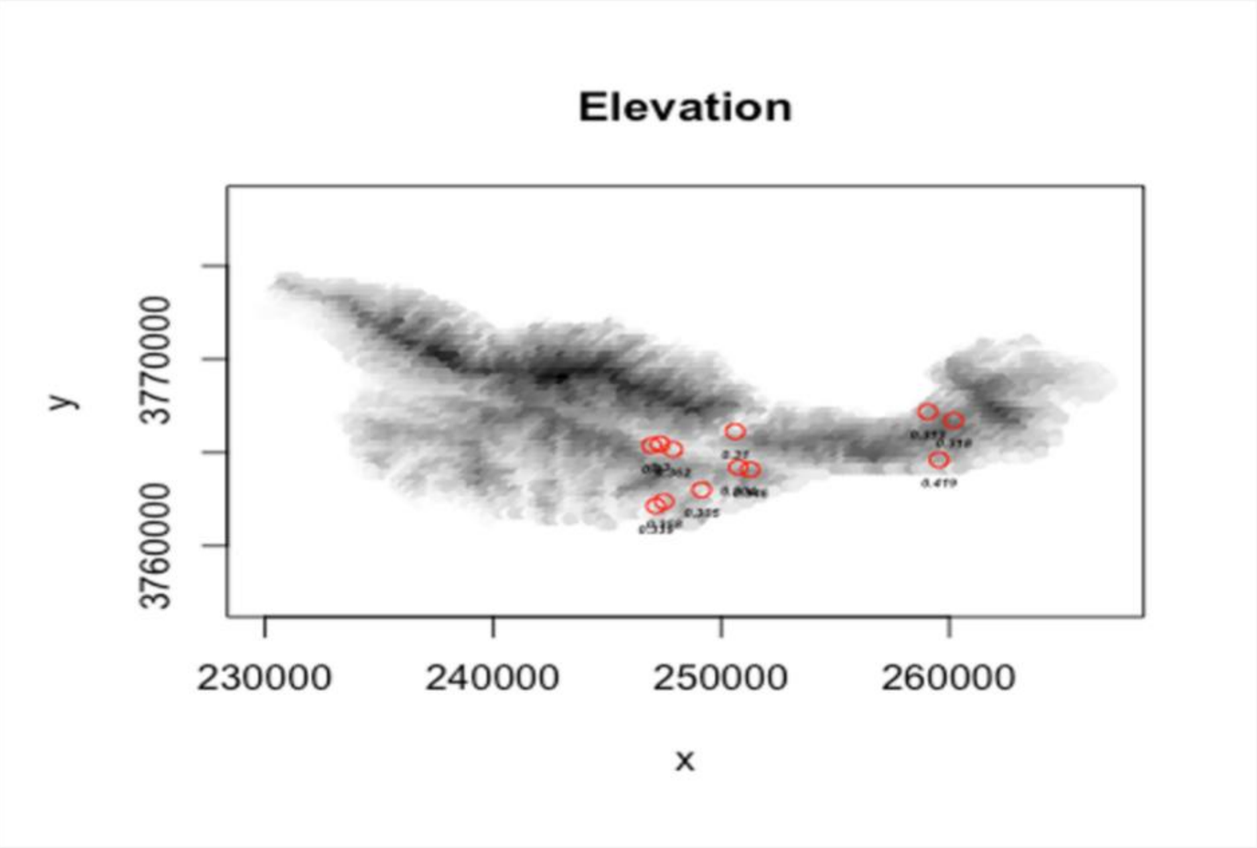
Step5.최종 모델

Step3-1에서 정의한 모델의 오분류율이 step3-2에서 정의한 모델보다 더 적으므로 Step3-1에서 정의한 모델의 오분류율을 사용하겠다.

logit(P)=−2.557400−0.002711·x−0.873567·y−0.171497·elev−3.657522·forest+0.353577·chap+3.775835·(x·forest)+0.159482·(forest·chap)

1) 모든 설명변수가 0일 때, 약 7.2%의 확률로 scrub jay가 존재한다.
2) 다른 변수들이 고정되어 있을 때, x 좌표가 1 단위 증가할 때 odds는 약 0.997291배로 감소한다
3) 다른 변수들이 고정되어 있을 때, y 좌표가 1 단위 증가할 때 odds는 약 0.41766배로 감소한다.
4) 다른 변수들이 고정되어 있을 때, 고도가 1 단위 증가할 때 odds는 약 0.84242배로 감소한다.
5) 다른 변수들이 고정되어 있을 때, forest 비율이 1 단위 증가할 때 odds는 약 0.02567배로 감소한다.
6) 다른 변수들이 고정되어 있을 때, chaparral 비율이 1 단위 증가할 때 odds는 약 1.42408배로 증가한다.
7) 다른 변수들이 고정되어 있을 때, x 좌표와 forest 비율의 상호작용이 1 단위 증가할 때 odds는 약 43.58182배로 증가한다.
8) 다른 변수들이 고정되어 있을 때, forest 비율과 chaparral 비율의 상호작용이 1 단위 증가할 때 odds는 exp(0.159482)이므로 약 1.17302배로 증가한다.

Step5.최종 모델

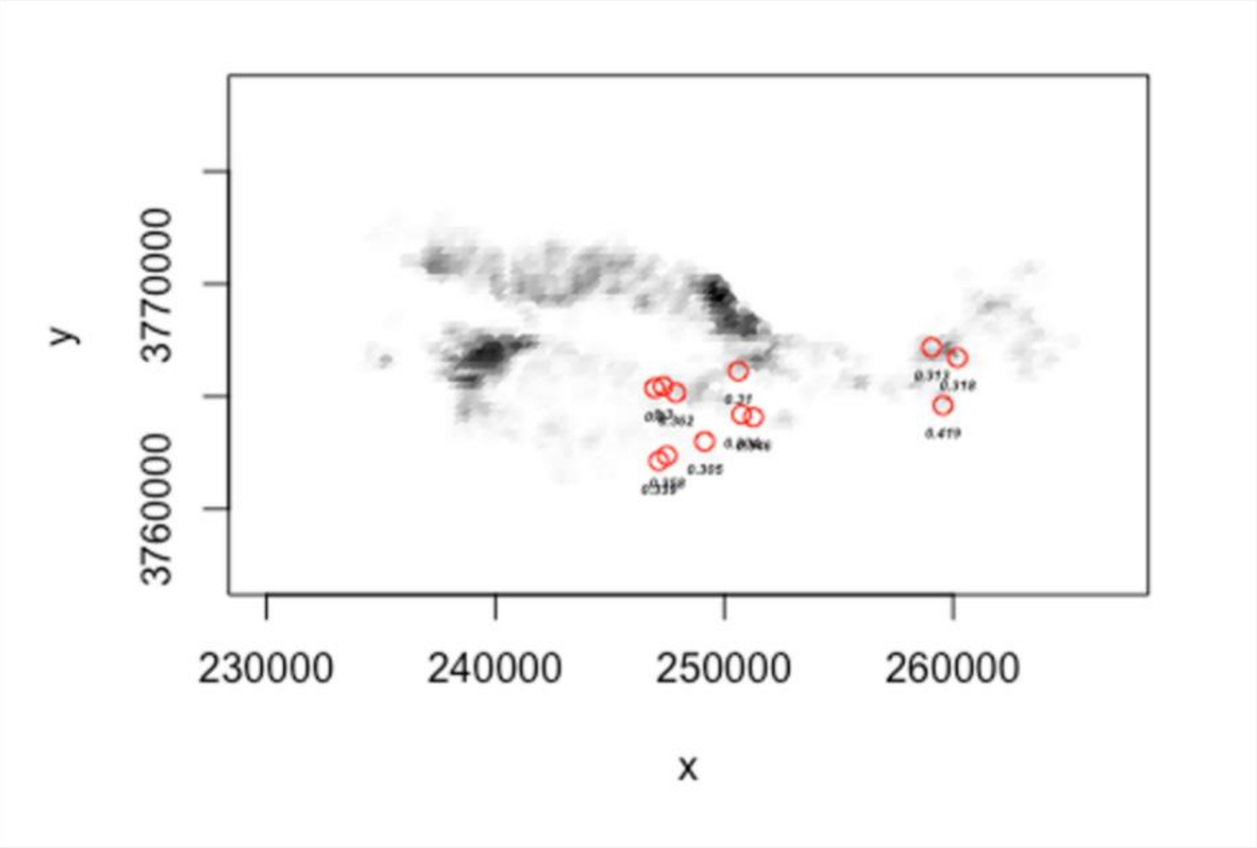


1) 고도

*반응변수의 값이 NA가 아닌 데이터들에서의 예측확률을 나타낸 그래프로 적합한 모델이 scrub jay가 존재한다고 판단한 곳에만 o 표시를 하겠다.

고도가 높은 곳보다는 적은 곳에 주로 scrub jay가 존재한다.

Step5.최종 모델

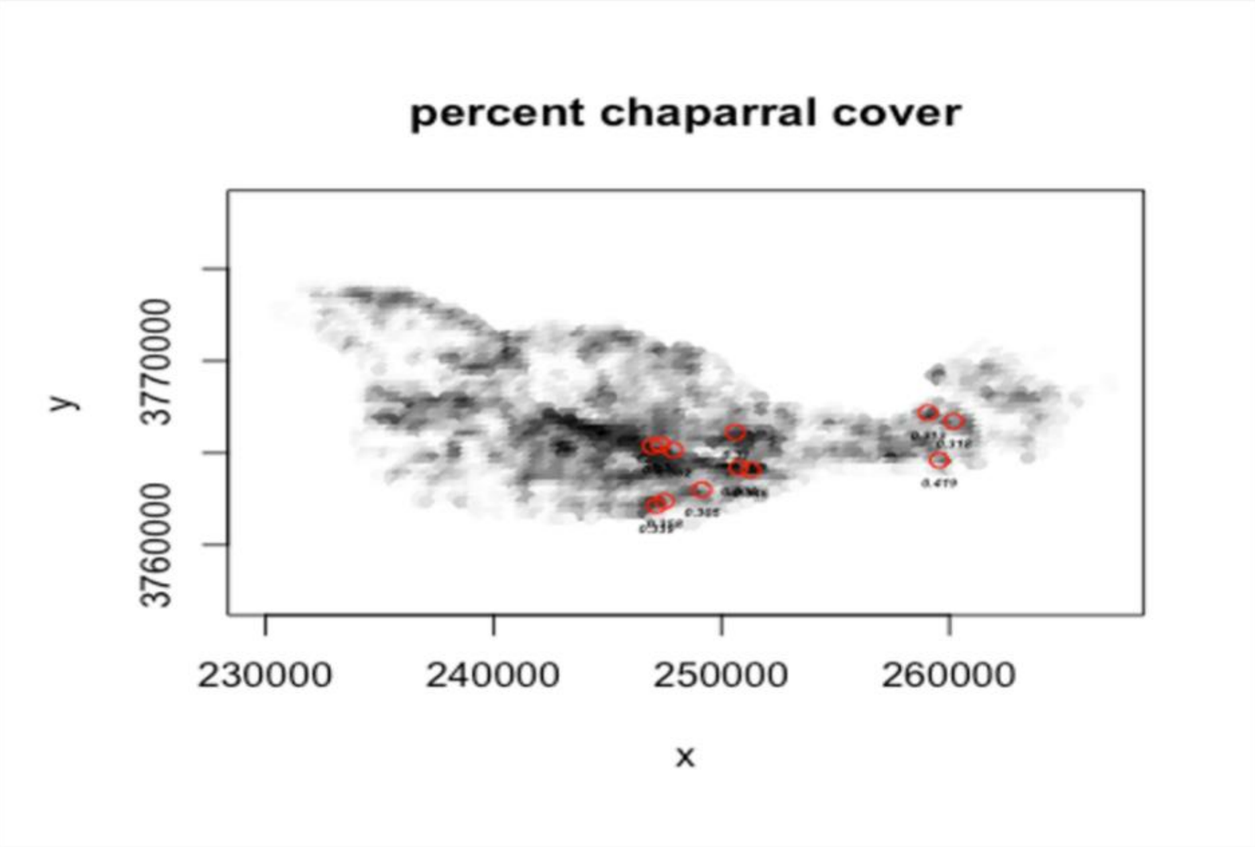


2) Forest 비율

*반응변수의 값이 NA가 아닌 데이터들에서의 예측확률을 나타낸 그래프로 적합한 모델이 scrub jay가 존재한다고 판단한 곳에만 o 표시를 하겠다.

Forest 비율이 적은 곳에 주로 scrub jay가 존재한다.

Step5.최종 모델

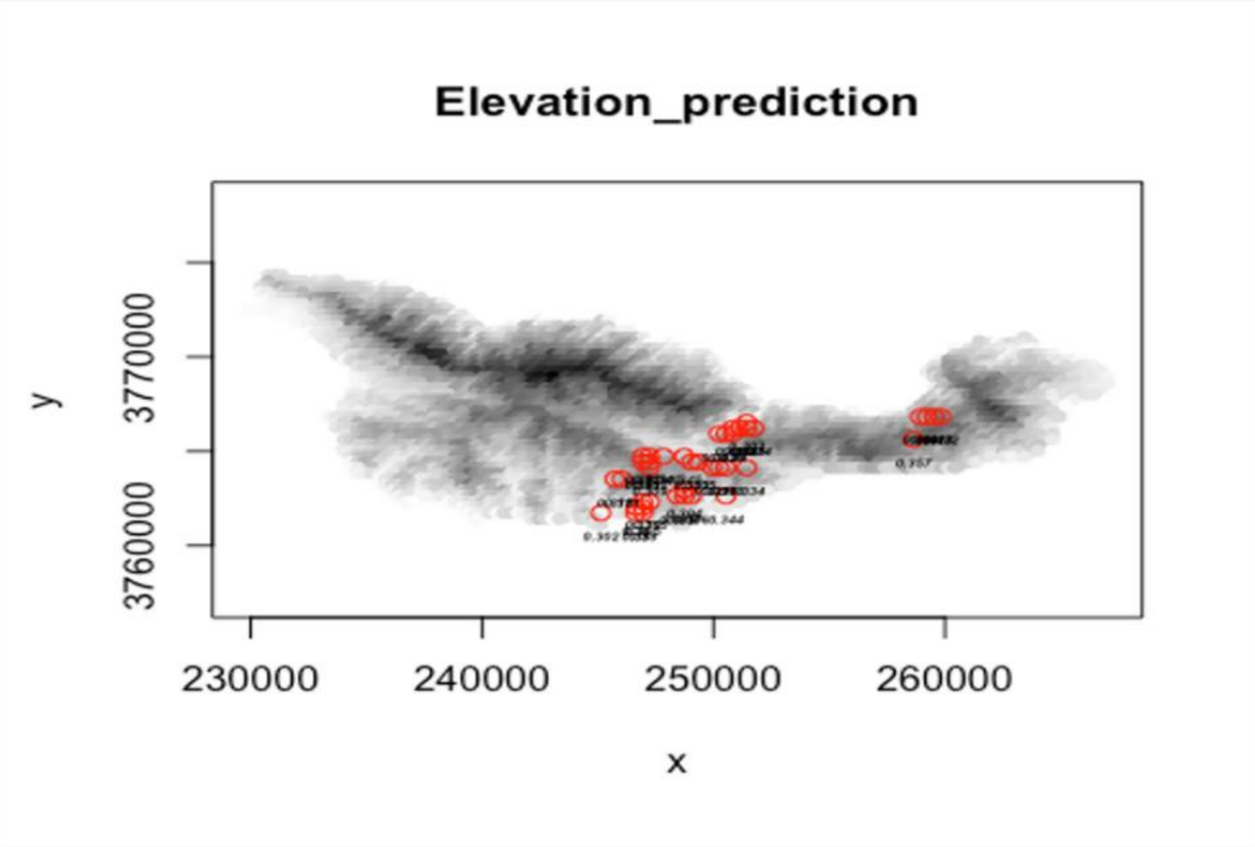


3) Chaparral 비율

*반응변수의 값이 NA가 아닌 데이터들에서의 예측확률을 나타낸 그래프로 적합한 모델이 scrub jay가 존재한다고 판단한 곳에만 o 표시를 하겠다.

Chaparral 비율이 높은 곳에 주로 scrub jay가 존재한다.

2484개 데이터셋 활용

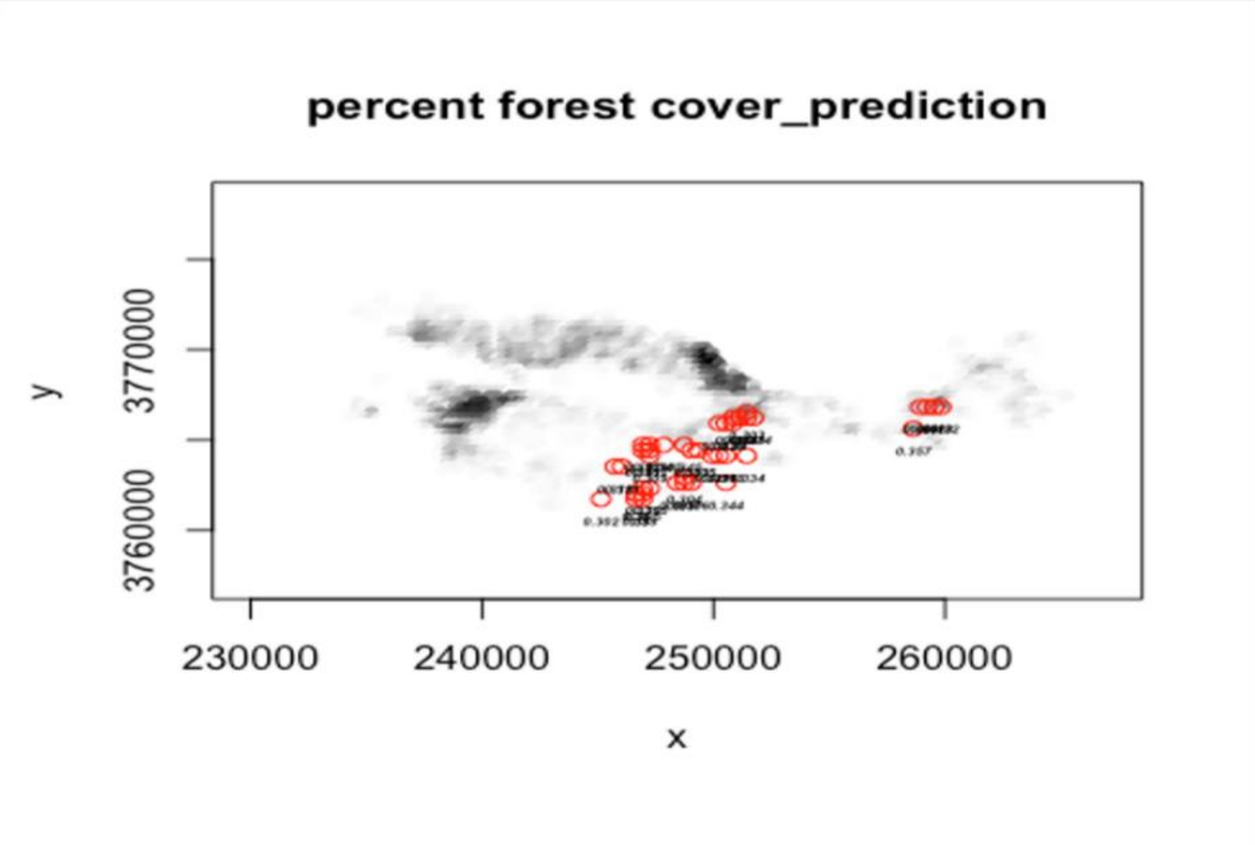


1) 고도

*반응변수의 값이 NA가 아닌 데이터들에서의 예측확률을 나타낸 그래프로 적합한 모델이 scrub jay가 존재한다고 판단한 곳에만 o 표시를 하겠다.

고도가 높은 곳보다는 적은 곳에 주로 scrub jay가 존재한다.

2484개 데이터셋 활용

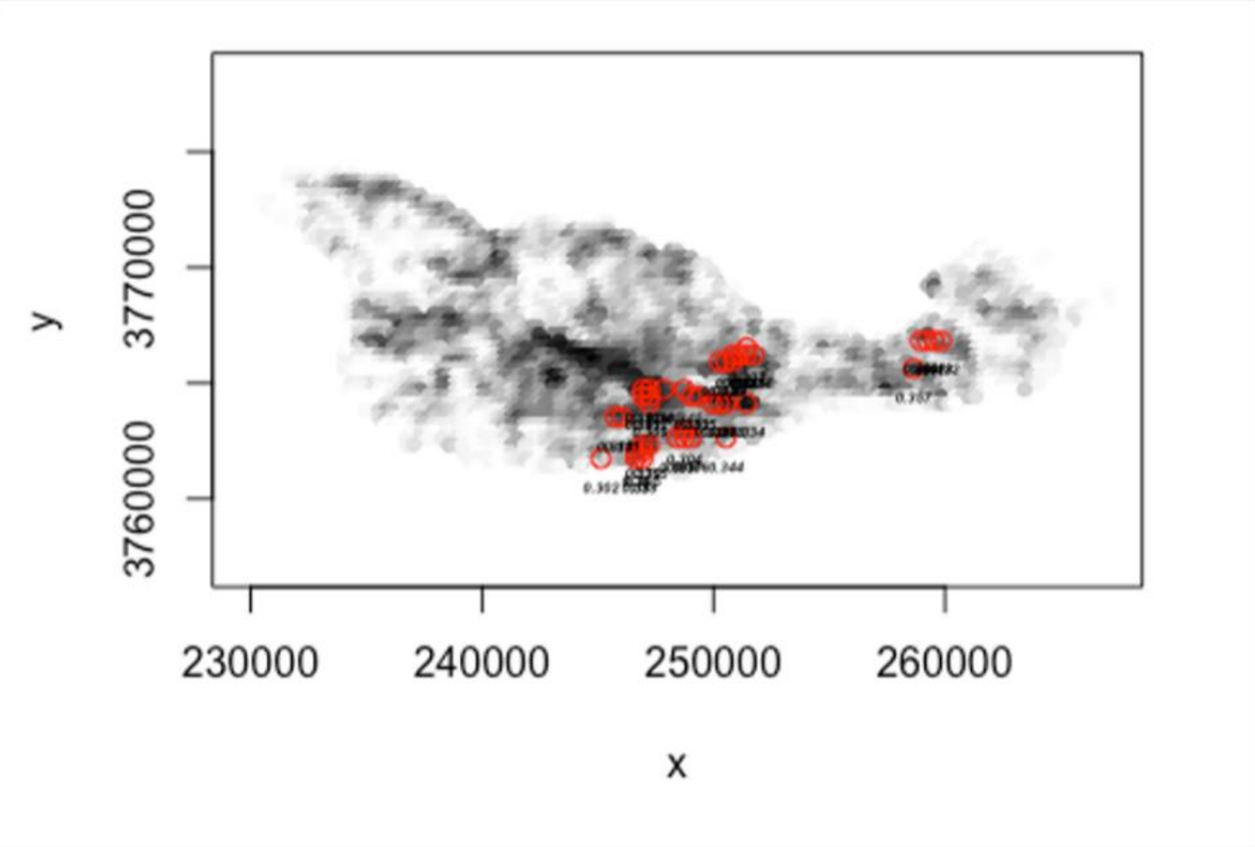


2) Forest 비율

*반응변수의 값이 NA가 아닌 데이터들에서의 예측확률을 나타낸 그래프로 적합한 모델이 scrub jay가 존재한다고 판단한 곳에만 o 표시를 하겠다.

Forest 비율이 적은 곳에 주로 scrub jay가 존재한다.

2484개 데이터셋 활용



3) Chaparral 비율

*반응변수의 값이 NA가 아닌 데이터들에서의 예측확률을 나타낸 그래프로 적합한 모델이 scrub jay가 존재한다고 판단한 곳에만 o 표시를 하겠다.

Chaparral 비율이 높은 곳에 주로 scrub jay가 존재한다.

마무리

해당 분석은 설명변수와 반응변수 간의 관계를 해석하고자 진행한 것으로, 해석에 집중한 프로젝트이다.
다음 기회가 된다면 머신러닝 혹은 딥러닝 기법을 활용하여 예측 성능에 집중한 프로젝트를 진행하고자 한다.