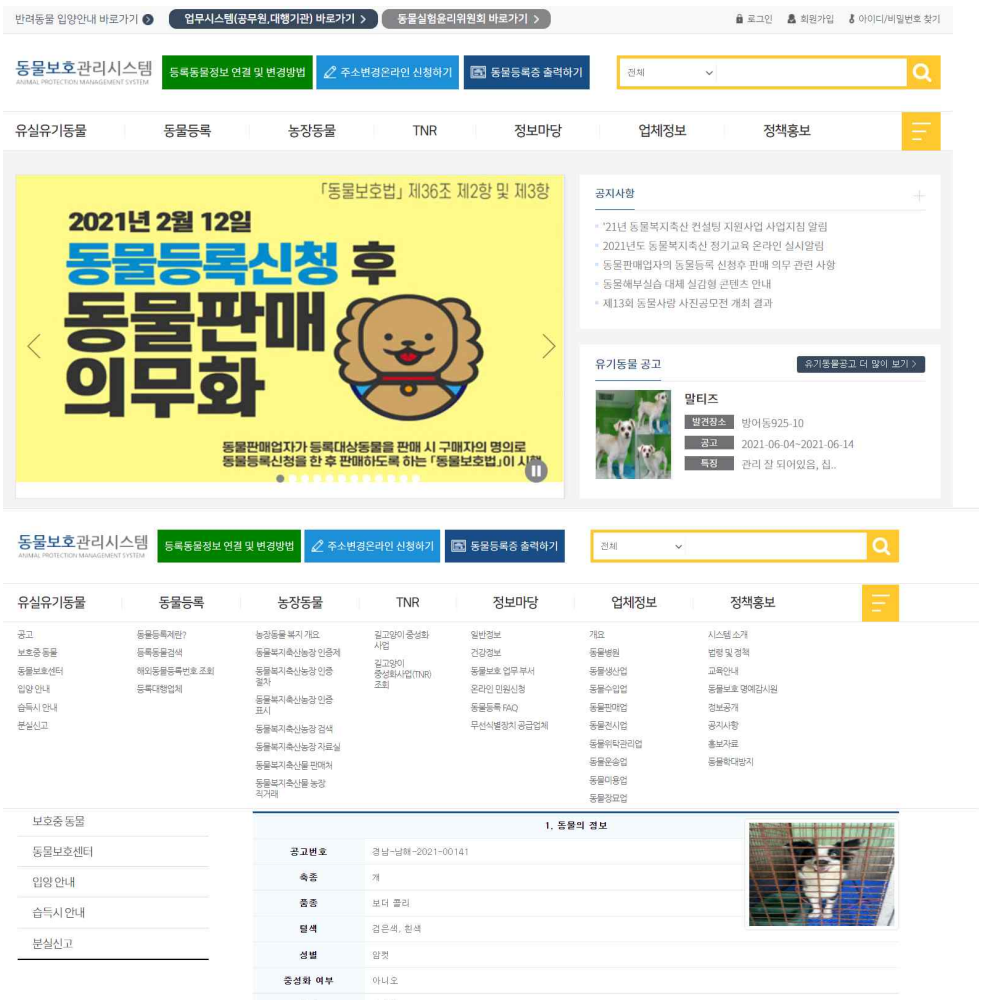
유기동물 입양율 예측

동물보호관리시스템(animal.go.kr)내 유기견, 유기묘의 실시간성 데이터 수집-저장-모델링(categorical)-시각화 파이프라인 구축

현대사회에 들어, 수많은 사람들이 반려동물을 키우기 시작했고, 이후 경제적 부담 등 여러 문제들로 인해 기르다가 버려지는 '유기동물'들도 적지않게 발생하게 되었다.
이러한 상황에서 유기동물 입양율을 예측하고, 이를 통해 각 지역의 보호소 관리자들에게 더 나은 운영방향성을 찾는데 조금이라도 도움이 되어보고자 해당 주제로 분석을 하기로 결정하였다.

데이터 수집 및 1차적 전처리 (Python)



유기동물의 정보를 얻기 위해서는 해당 오픈api의 각 개체 별 url을 클릭해야 했기에

파이썬의 sellenium 모듈을 사용하여 데이터를 수집하였다. 최소 10일간의 공고기간 후에 입양 및 안락사 등의 결정을 할 수 있기에 (보호소별 15일의 공고기간을 두는 곳도 있었다)
21년 4월 한 달간의 데이터를 사용.

사진 이미지를 제외한 html형식의 각 유기동물 개체별 자세히보기 내 모든 정보를 beautifulsoup모듈을 통해 필요한 형식으로 정리하여 csv파일로 저장.

유기동물의 상태값은 아래와 같이 작성되어 있었는데, 원 주인에게의 반환과 자연사, 방사는 제외하여 [보호중, 안락사를 0] [입양을 1]로 모형설계 전 전처리 수행.
'보호중','종료(안락사)','종료(입양)','종료(반환)','종료(자연사)',‘종료(방사)’

설명변수로 사용할 나머지 정보들은 개와 고양이의 축종, 성별, 품종, 나이, 중성화여부, 보호소 지역, 특징이다. 각 보호소의 관계자가 자연어 텍스트로 작성하는 특징란을 제외하고 나머지 변수들은 숫자 인덱스로 구분만 하였고 factor로 범주화는 Rstudio에서 작업.
(나이age 변수는 수치형으로 그대로 사용.)

state	sep	age	kind	sex	neuter	do	si	date	feature
1	0	2	0	1	0	인천광역시	중구	2021-04-30	5,6개월추정되며, 비교적 양호한편임
1	0	4	1	0	0	대구광역시	달성군	2021-04-30	온순함,관리가 안되어있음
1	1	1	0	1	0	부산광역시	해운대구	2021-04-30	영양불량사태,기력저하,검정색
0	0	1	0	1	0	부산광역시	강서구	2021-04-30	중구1-26호,온순한성격 얼굴,꼬리부분만 조금 진한 털
1	1	1	1	0	0	서울특별시	동작구	2021-04-30	꼬리가 꺽임
1	0	6	0	0	0	경상북도	경주시	2021-04-30	많이 사나운편

(*sep 열은 개와 고양이 구분)
(*state 1 : 입양)

모델링 (R)

온순함,심장사상충감염
4-63, 칩없음, 산에서 발견됨, 많이 마른 상태, 온순하고 예쁜 아이
4-62, 칩없음, 분홍색 끊어진 하네스 착용, 작고 예쁘고 온순한 성격
어미 안보인지 며칠된 새끼고양이 구조
얼굴에 털이 복슬복슬하게 많이났고 겁이조금 있다
순한 어린강아지로 하네스를 착용하고 있음.
없음
순하고 사람을 잘따르는 아이, 전반적 복부 발적, 피부 상태 좋지 않음,
경계심 있으나 순함

위와 같은 특징란을 변수로 사용하기 위해 자연어 처리를 먼저 해주었다. 긍정감정과 부정감정으로 나눈 군산대 감성사전에 반려동물에게 많이 사용되는 아래와 같은 묘사나 질병 단어들을 추가했고, 특징란의 단어들과 매칭될 때 1점을 부여하였다. 긍정사전에서 얻은 점수에서 부정사전에서 얻은 점수를 빼서 최종 점수를 수치형 변수로 사용했다.

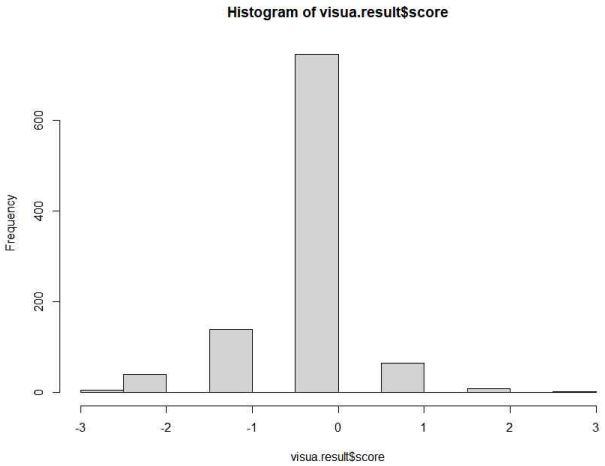
눈병	약물중독으	기아상태	마비
안검탈출증	재채기	영양불량상	식욕부진
눈이 좋지않음	호흡곤란	설사	심장사상충
곰팡이성	쇠약	콧물	장염
백내장	허약	눈꼽	기력
시력	야윔	종양	호흡
시력상실	탈수	저체온증	몸 가누지못함
눈안보임	탈진	기력없음	피오줌
안질환	저체온증	탈장	결막염
안구		혈변	안염
각막궤양			눈염증
치석심함			
치석증			
구내염			

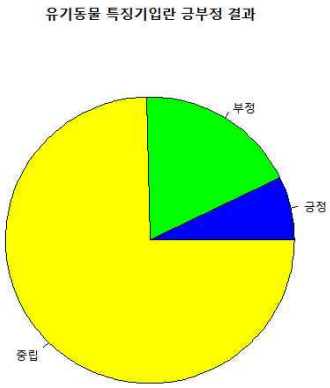
```r
attach(life0)

positi = readLines("pos_pol_word.txt", encoding = 'UTF-8')
head(positi)
negati = readLines("neg_pol_word.txt", encoding = 'UTF-8')
is.vector(negati)

emotion = function(sentences, positive, negative) {

  scores = laply(sentences, function(sentence, positive, negative) {

    sentence = gsub('[[:punct:]]', '', sentence) # 문장부호 제거
    sentence = gsub('[[:cntrl:]]', '', sentence) # 특수문자 제거
    sentence = gsub('\\d+', '', sentence) # 숫자 제거

    word.list = str_split(sentence, '\\s+') # 공백기준으로 단어생성 \\s+ :공백
    words = unlist(word.list)

    pos.matches = match(words, positive) # words의 단어를 positive에서 matching
    neg.matches = match(words, negative)

    pos.matches = !is.na(pos.matches)
    neg.matches = !is.na(neg.matches)

    score = sum(pos.matches) - sum(neg.matches)
    return(score)

  }, positive, negative)

  scores.df = data.frame(score = scores, text = sentences)
  return(scores.df)

}
```

처리 결과, 점수의 분포와 긍정/부정/중립의 비율을 아래와 같았다.

```
feascore = as.data.frame(life1$featurescore)
feascore$remark[life1$featurescore > 0] = '긍정'
feascore$remark[life1$featurescore == 0] = '중립'
feascore$remark[life1$featurescore < 0] = '부정'

score_table = table(feascore$remark)
score_table

pie(score_table, main = '유기동물 특징기입란 긍부정 결과',
    col = c('blue', 'green', 'yellow'), radius=1.0)


hist(life1$featurescore)
```

```
> score_table

긍정 부정 중립
 441  535 3719
```





안락사와 공고기간 이후에도 보호중인 상태 대비 입양될 확률을 구하기 위해 Logistic regression을 사용했다. (행정구역의 두 column중 시군구별은 사용하지 않았고 도별만 사용)

```
catdog = as.factor(catdog) ; kind = as.factor(kind) ; sex = as.factor(sex)
neuter = as.factor(neuter) ; location = as.factor(location)
summary(life1)

obj = glm(state ~ . ,data = life1, family = binomial)
summary(obj)
plot(obj)

obj1 = glm(state ~ .-location, data = life1, family = binomial)
summary(obj1)
plot(obj1)

life2 = cbind(life1, feascore$remark)
colnames(life2)
colnames(life2)[9] = 'featuredirec'

attach(life2)
obj2 = glm(state ~. -location -featurescore, data = life2, family = binomial)
summary(obj2)

# plot(state)
model.zero = glm(state ~. -featurescore ,data = life2, family = binomial)
install.packages("MASS")
library(MASS)
step.backward = stepAIC(model.zero, direction = "backward")


step.backward.obj = stepAIC(obj, direction = "backward")
step.backward.obj1 = stepAIC(obj1, direction = "backward")

obj.loca = glm(state ~ location, life2, family = binomial)
summary(obj.loca)

dim(life0)

set.seed(511)
life0[1, ]
train = sort(sample(1:4695, 3050))
test = setdiff(1:4695, train)
```

안락사와 공고기간 이후에도 보호중인 상태 대비 입양된 상태의 로그오즈비인 logit을 설명변수들의 선형적 결합으로 파악할 수 있었고, 각 개체의 입양될 확률을 추정할 수 있다.
특징란에 기입된 텍스트를 점수화 한 featurescore 변수 대신 긍정/부정/중립으로만 범주화한 featuredirec을 사용해 모형을 구축했을 시, 설명력이 조금 더 떨어지는 것을 확인할 수 있었다.

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.31642    0.16745  -1.890  0.05881 .
catdog               -0.40229    0.13267  -3.032  0.00243 **
age                  -0.26191    0.02083 -12.575  < 2e-16 ***
kind                  1.88226    0.11084  16.982  < 2e-16 ***
sex                   0.15601    0.06701   2.328  0.01991 *
neuter                0.58754    0.20046   2.931  0.00338 **
location경기도        0.55644    0.17080   3.258  0.00112 **
location경상남도     -0.08164    0.18163  -0.449  0.65310
location경상북도      0.27541    0.18148   1.518  0.12911
location광주광역시   -0.09335    0.29651  -0.315  0.75289
location대구광역시    1.41575    0.31746   4.460 8.21e-06 ***
location대전광역시    1.86623    0.90639   2.059  0.03950 *
location부산광역시    1.10450    0.24657   4.480 7.48e-06 ***
location서울특별시    2.02396    0.41560   4.870 1.12e-06 ***
location세종특별자치시 -0.53321    0.55048  -0.969  0.33273
location울산광역시    0.48276    0.26243   1.840  0.06583 .
location인천광역시    1.28836    0.23805   5.412 6.23e-08 ***
location전라남도      0.09633    0.19739   0.488  0.62555
location전라북도      0.46177    0.18436   2.505  0.01225 *
location제주특별자치도 -1.05577    0.21136  -4.995 5.88e-07 ***
location충청남도      0.50110    0.19710   2.542  0.01101 *
location충청북도      0.45151    0.22103   2.043  0.04108 *
featurescore          0.29626    0.06090   4.865 1.15e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6508.3  on 4694  degrees of freedom
Residual deviance: 5479.1  on 4672  degrees of freedom
AIC: 5525.1

Number of Fisher Scoring iterations: 4
```

몇 개의 행정지역factor를 제외한 설명변수들이 log[입양되지 않을 확률 대비 입양될 확률]에 유의한 영향을 끼치고 있음을 볼 수 있다. 특징점수(featurescore)가 1점 증가 시 입양될 오즈비가 exp[0.2962]만큼 승법적으로 증가한다고 해석 가능하다.

```
Call:
glm(formula = state ~ . - location - featurescore, family = binomial,
    data = life2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0441  -1.0454   0.5141   1.1921   2.3836

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.33237    0.11547   2.878 0.003997 **
catdog            -0.15572    0.12682  -1.228 0.219485
age               -0.24144    0.02010 -12.013  < 2e-16 ***
kind               1.86601    0.10665  17.496  < 2e-16 ***
sex                0.10832    0.06451   1.679 0.093117 .
neuter             0.76148    0.19589   3.887 0.000101 ***
featuredirec부정 -0.61176    0.14335  -4.268 1.98e-05 ***
featuredirec중립 -0.40966    0.10863  -3.771 0.000163 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6508.3  on 4694  degrees of freedom
Residual deviance: 5736.5  on 4687  degrees of freedom
AIC: 5752.5
```

```
Call:
glm(formula = state ~ . - location, family = binomial, data = life1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0090  -1.0516   0.4828   1.1888   2.3603

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.06483    0.06186  -1.048   0.2946
catdog       -0.14085    0.12659  -1.113   0.2658
age          -0.23858    0.02002 -11.917  < 2e-16 ***
kind          1.86416    0.10669  17.473  < 2e-16 ***
sex           0.10821    0.06449   1.678   0.0934 .
neuter        0.76844    0.19614   3.918 8.94e-05 ***
featurescore  0.27661    0.05782   4.784 1.72e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6508.3  on 4694  degrees of freedom
Residual deviance: 5732.9  on 4688  degrees of freedom
AIC: 5746.9

Number of Fisher Scoring iterations: 4
```

지역을 구분하지 않았을 때 입양될 확률에 있어 유기묘와 유기견의 유의한 차이는 없지만 지역별 보호소의 상황이 모두 다르므로 catdog변수를 loacation변수와 함께 사용하였다.

전체 설명변수를 포함한 모형에서 변수를 줄여나가는 Backward 방식으로 변수선택을 하였다. 처음의 전체모형의 아카이케 정보 Criterion이 가장 낮았으며 GVIF^(1/2*자유도)가 모두 2.0 이하로 설명변수들 간의 다중공선성이 없다고 판단해 모형을 수립하였다.

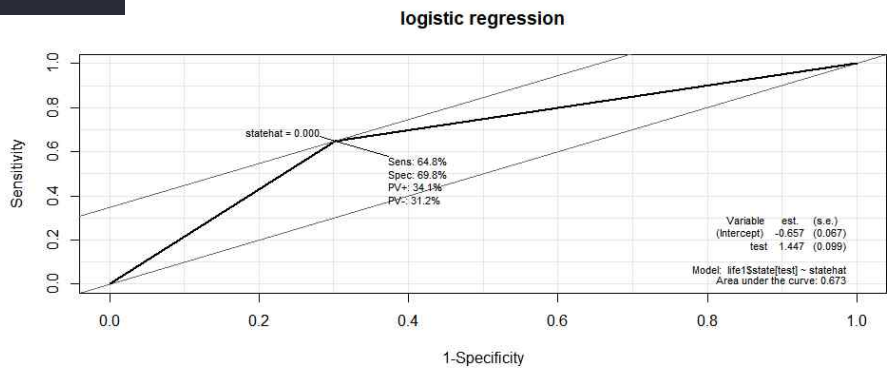
```
> library(MASS)
> step.backward = stepAIC(model.zero, direction = "backward")
Start:  AIC=5533.25
state ~ (catdog + age + kind + sex + neuter + location + featurescore +
    featuredirec) - featurescore

               Df Deviance    AIC
<none>              5485.2 5533.2
- sex            1   5490.6 5536.6
- neuter         1   5493.9 5539.9
- catdog         1   5494.9 5540.9
- featuredirec   2   5503.4 5547.4
- age            1   5666.0 5712.0
- location      16   5736.5 5752.5
- kind           1   5820.4 5866.4
>
>
> step.backward.obj = stepAIC(obj, direction = "backward")
```

```
> vif(obj.train)
                GVIF Df GVIF^(1/(2*Df))
catdog       2.200525  1        1.483417
age          1.572249  1        1.253894
kind         2.337333  1        1.528834
sex          1.057624  1        1.028408
neuter       1.190244  1        1.090983
location     1.305302 16        1.008361
featurescore 1.111068  1        1.054072
```

4695개의 row들중 60% 만을 train set으로 사용하였으며, 나머지 40%는 test set으로 남겨두어 검증하였다.

```
train = sort(sample(1:nrow(life1), 0.6*nrow(life1)))
test = setdiff(1:nrow(life1), train)

obj.train = glm(state ~ . , data = life1, subset = train, family = binomial)
result = summary(obj.train)
phat = predict.glm(obj.train, newdata = life1[test, ], type = "response")
statehat = ifelse(phat>=0.5, 1, 0)

tt = table(statereal = life1$state[test], statehat = statehat)
install.packages("Epi")
library(Epi)
gof = ROC(test = statehat, stat = life1$state[test], plot = "ROC", AUC = TRUE,
          main = "logistic regression")

print(result)
print(gof)

write.csv(tt, 'Accura table.csv')
```

```
> tt
         statehat
statereal   0   1
        0 646 280
        1 335 617
>
```



특이도와 민감도가 seed를 달리했을 때 65~70% 사이를 유지하였다.

1. 유기동물의 이미지 데이터 사용 및 분석
2. 유기묘와 유기견 각각 모델링
3. kind 변수는 순종과 믹스(견/묘) 로만 나눈 것인데 유기견의 경우 소형-대형견으로 변수 추가

시 예측력이 올라갈 수 있다고 예상하였다.