Report on Black Bullhead Presence Predication

Xinrui Wang 1004741078; Bisong Zhou 1004738741;Congyuan Lian 1004700228

MGEC11H3: Introduction to Regression Analysis

2021/8/10

a)

Plot black bullhead versus other variables. See figure 1. Since the value of black bullhead only has 0 and 1, we use boxplot and logistic regression model.

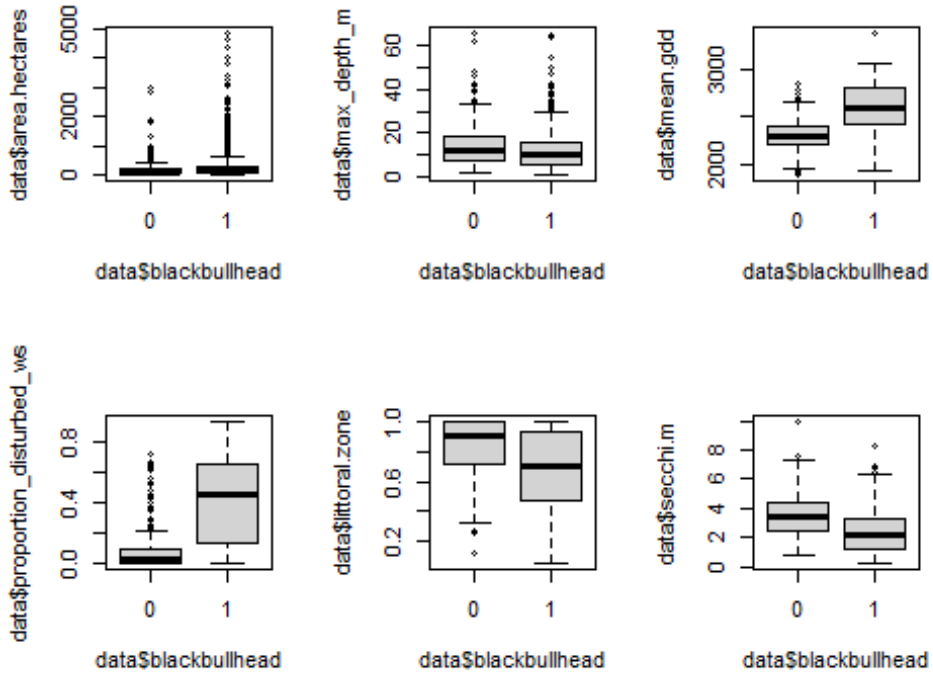

Figure 1 boxplots

Boxplot suggest that max_depth has no effect.

Generalized Linear Model 1:

Black Bullhead~ Area + Depth + gdd + proportion_disturbed + Littoral + Secchi

we got the following output. p-value of max_depth_m > 0.05 is also confirming that Depth is not significant, so we will drop Depth.

```
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)          -1.529e+01  1.905e+00  -8.024 1.02e-15 ***
## area                  1.247e-03  2.974e-04   4.194 2.74e-05 ***
## depth                 1.988e-02  1.653e-02   1.202  0.22922
## gdd                   6.274e-03  7.506e-04   8.359  < 2e-16 ***
## proportion_disturbed  2.772e+00  6.394e-01   4.335 1.46e-05 ***
## littoral              1.400e+00  6.598e-01   2.122  0.03382 *
## secchi               -3.341e-01  1.049e-01  -3.186  0.00144 **
…

## AIC: 853.94
```

Model 2:

Black Bullhead~ Area + gdd + proportion_disturbed + Littoral + Secchi

```
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.473e+01  1.840e+00  -8.002 1.23e-15 ***
## area                  1.332e-03  2.939e-04   4.534 5.79e-06 ***
## gdd                   6.205e-03  7.477e-04   8.299  < 2e-16 ***
## proportion_disturbed  2.763e+00  6.387e-01   4.325 1.52e-05 ***
## littoral              8.461e-01  4.738e-01   1.786  0.07414 .
## secchi               -2.370e-01  6.630e-02  -3.575  0.00035 ***
…

## AIC: 853.4
```

As we can see from the above output, at 5% significance level, littoral.zone is not significant, we will remove the variable Littoral.

Model 3:

Black Bullhead~ Area + gdd + proportion_disturbed + Secchi

```
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.358e+01  1.707e+00  -7.957 1.76e-15 ***
## area                  1.187e-03  2.736e-04   4.337 1.44e-05 ***
## gdd                   5.972e-03  7.310e-04   8.170 3.08e-16 ***
## proportion_disturbed  2.773e+00  6.409e-01   4.327 1.51e-05 ***
## secchi               -1.981e-01  6.211e-02  -3.190  0.00142 **
…

## AIC: 854.6
```

From the above output, all p-value are smaller than 5%. i.e. all variables in model 3 are significant.

Plot model 3. The diagnostic plots suggest that this model is not good. The QQ-plot in figure 2 indicates that the normality assumption may be violated. Thus, some transformations are needed.

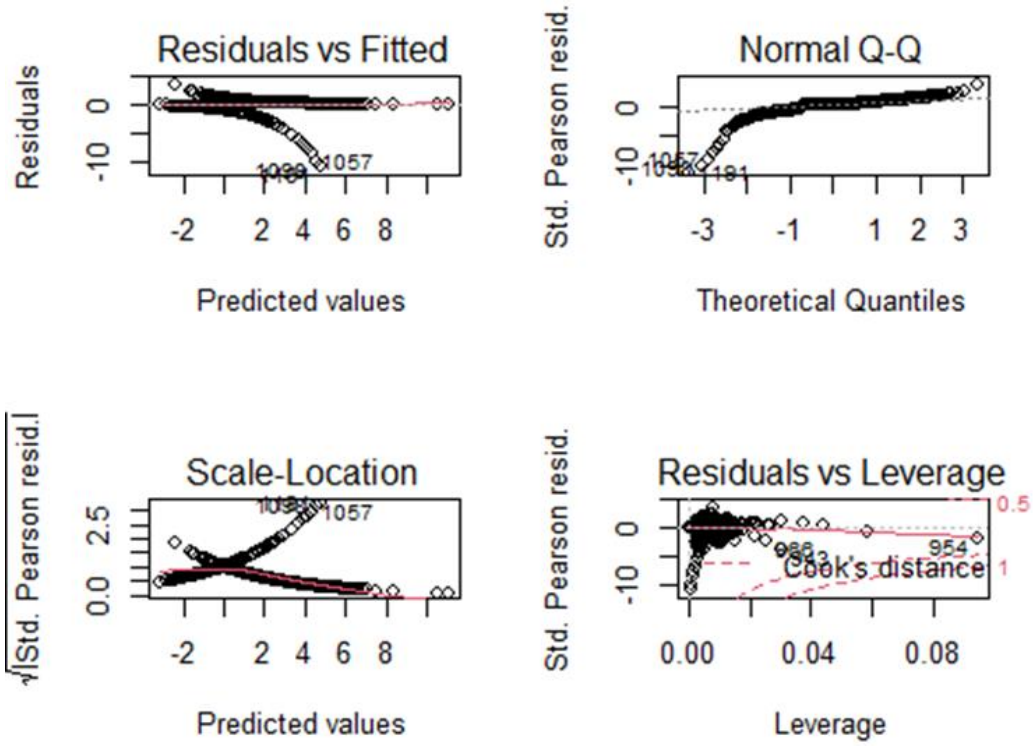

Figure 2 diagnostic plots of model 3

Try x-log transformation in model 4:

Black Bullhead~ log(Area + gdd + proportion_disturbed + Secchi)

Plot model 4. The diagnostic plots in Figure 3 show that model 4 provides a much better fit than 3.

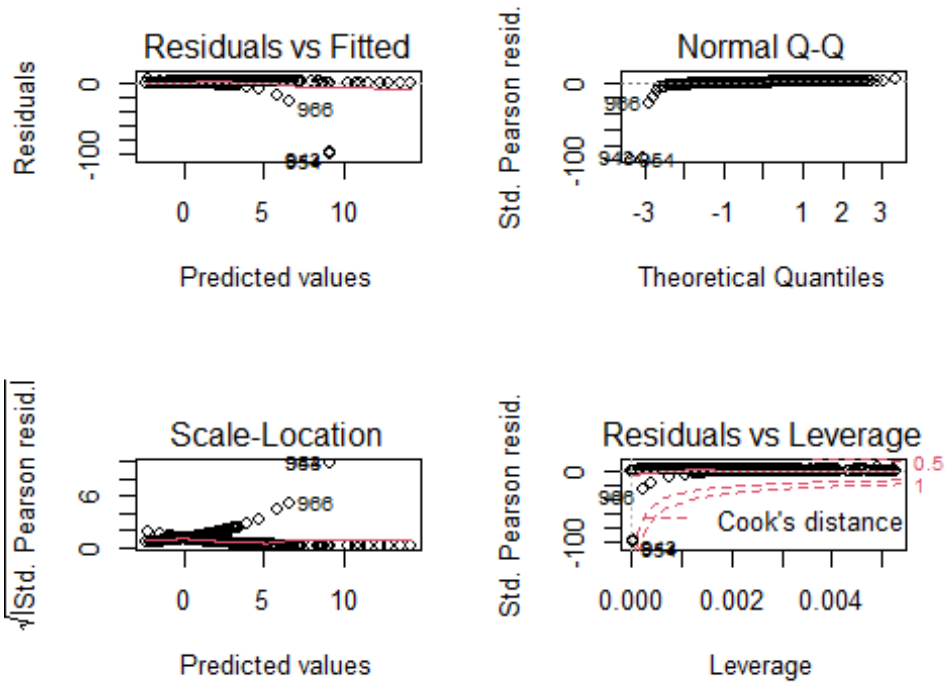

Figure 3 diagnostic plots of model 4

```
Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)

(Intercept)                             -95.4331     7.0029  -13.63   <2e-16 ***
log(area+gdd+proportion_disturbed+secchi) 12.2584     0.8915   13.75   <2e-16 ***
…
## AIC: 1041.1
```

Thus, model 4 is

Black Bullhead = -95.4331 + 12.2584 * log (Area + gdd + proportion_disturbed + Scchi)

To check the existence of multicollinearity, consider Variance Inflation Factor of model 1.

```
##                 area                depth                  gdd
##             1.181359             3.123487             1.623795
## proportion_disturbed             littoral              secchi
##             1.530166             2.503512             2.862349
```

All of them are smaller than 5, so there is no multicollinearity among them.

The best fit model is model 2 because it has the smallest AIC among all.

The complete equation is:

$$\eta = -14.73 + 0.001332 \times area + 0.006205 \times gdd + 2.763 \times proportion\_distributed + 0.8461 \times littoral - 0.237 \times secchi$$

$$\pi = \frac{e^{\eta}}{e^{\eta} + 1}$$

$$\pi = \frac{e^{-14.73+0.001332\times area+0.006205\times gdd+2.763\times proportion\_distributed+0.8461\times littoral-0.237\times secchi}}{e^{-14.73+0.001332\times area+0.006205\times gdd+2.763\times proportion\_distributed+0.8461\times littoral-0.237\times secchi} + 1}$$

b)

If the proportion of disturbed watershed increases by 10%, η increase by 0.2763, which means $\log\frac{\pi}{1-\pi}$ increase 0.2763, and then after calculation, the odds $\frac{\pi}{1-\pi}$ should be $e^{0.2763}$ = 1.3182.

In other words, there is 31.82% increase in the presence of black bullhead per unit increase in the proportion of disturbed watershed. Therefore, black bullhead is 3.182% more likely to present when the proportion of disturbed watershed increases 10%.

c)

If strategy 1 is applied, the expected black bullhead presence will be 925.

If strategy 2 is adopted, the expected black bullhead presence will be 1024.

As we are aimed at decreasing black bullhead presence in lakes, strategy 2 should be recommended.