Amphibian Presence Prediction

Jerry Ngo

Intro

- Based on satellite information on amphibian appearance, give prediction on the occurrence of amphibian based on the site’s attribute.
- Applying model like decision tree, support vector machine, artificial neural network on the data set.

Data Set

Data Set Information

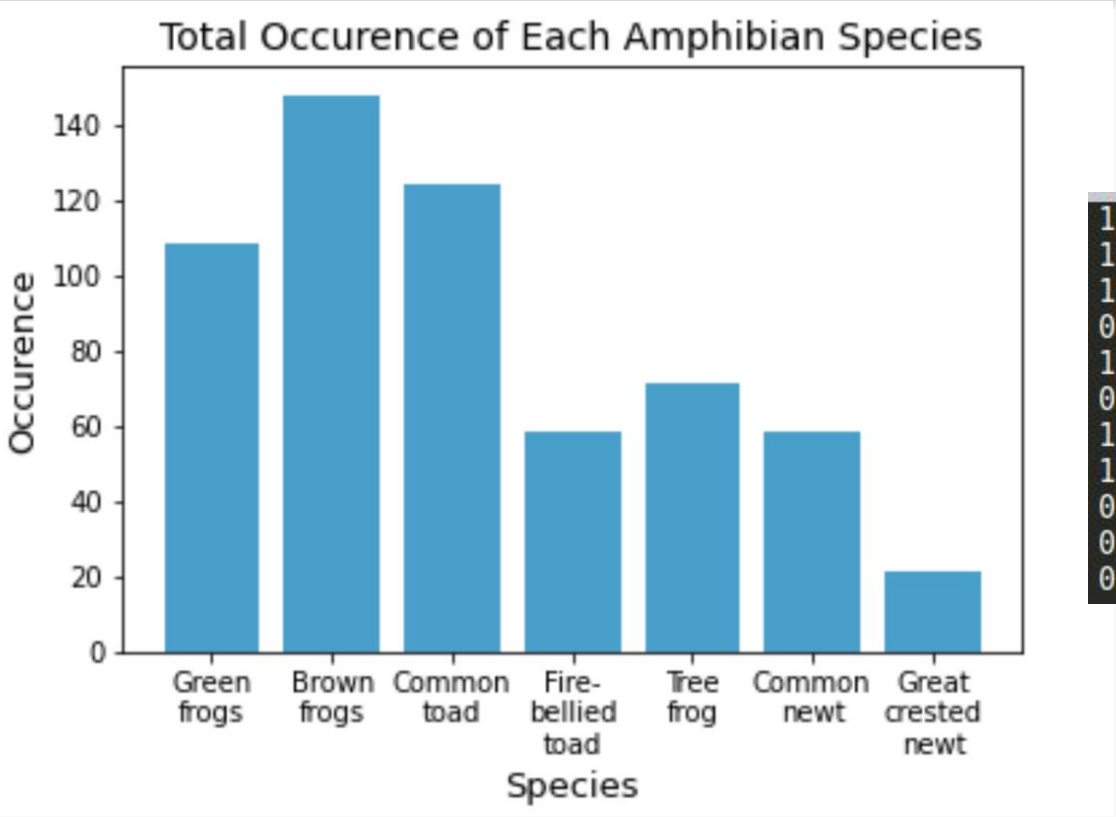
- Was prepared for the environmental impact assessment reports for two planned road in Poland.
- Gives the amphibian population with 189 occurence sites along with 16 attributes of each site.
- 7 amphibian species

Data Set Preprocess

- Data is well collected and presented
- Apply min-max normalization
- Select attribute for training
- Concatenate label
- Split training and validation set

	SR	NR	TR	VR	SUR1	...	RR	BR	MR	CR	label
0	0.004352	0.000000	0.000000	1.00	0.384615	...	0.0	0.0	0.0	0.0	0
1	0.005222	0.000000	0.285714	0.25	0.692308	...	0.1	0.1	0.0	0.0	110010
2	0.000870	0.000000	0.285714	0.25	0.692308	...	0.1	0.1	0.0	0.0	110010
3	0.001741	0.000000	0.285714	0.00	0.384615	...	0.0	0.0	0.0	0.0	10000
4	0.004352	0.166667	0.000000	1.00	0.692308	...	0.0	0.5	0.0	0.0	111011
..	...	...	...	...	...	...	...	...	...	...	...
119	0.008703	0.000000	0.285714	0.00	0.076923	...	0.5	0.1	0.0	1.0	0
120	0.020888	0.000000	0.000000	1.00	0.076923	...	0.5	0.1	0.0	0.0	1110110
121	0.002176	0.000000	0.000000	0.00	0.076923	...	0.1	0.1	0.0	0.0	1110000
122	0.003481	0.000000	0.000000	1.00	0.000000	...	0.5	0.5	0.0	0.0	1111010
123	0.001741	0.000000	0.785714	0.75	0.076923	...	0.1	0.0	0.0	0.0	110000

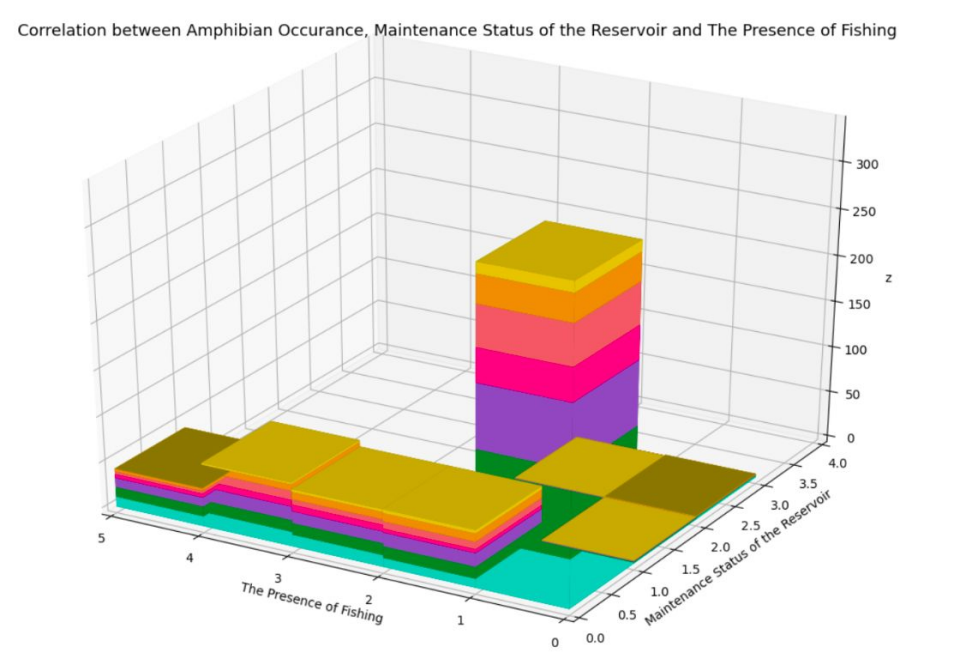
Data Set Analyzation

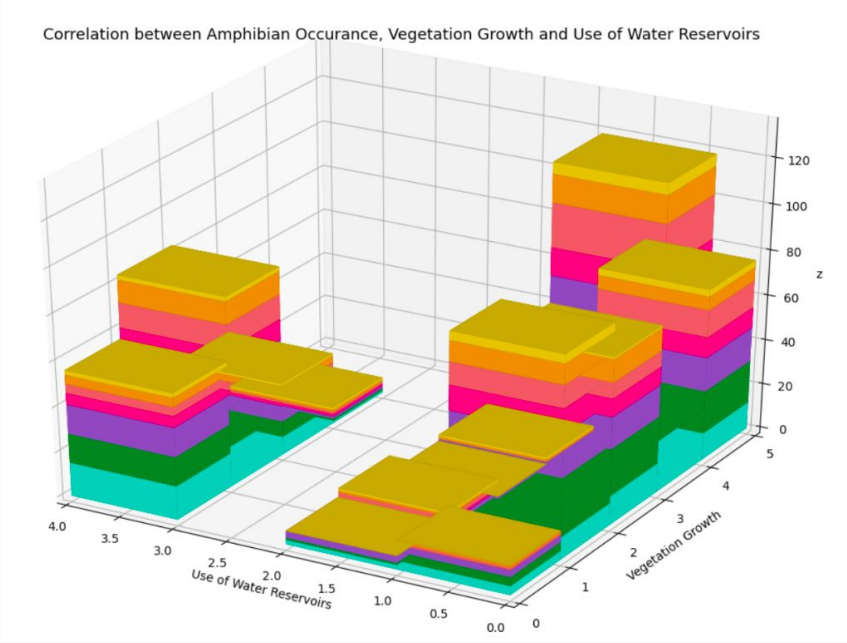



```
1101010 3  0111011 2  1010010 2  1000010 2
1010100 1  0111111 2  0101000 1  0010100 1
1101000 2  0001101 1  0000000 6  0100000 23
0011100 1  0001000 1  1100110 1  1000100 1
1100000 7  1010000 5  0110100 7  0010000 4
0100100 1  1110110 6  1111010 7  1111110 12
1110010 3  0110010 3  1111100 7  1001000 2
1000000 7  0001100 1  1111000 3  1110111 3
0000100 4  1110011 1  1110000 12  1111111 10
0111000 2  1110100 9  0100010 1  1110101 1
0110000 19  0010101 1  1001100 1  [Finished
```

Distribution is not evenly distributed

Data Set Analyzation





No attributes that can be used to easily predict the species occurrence

Machine Learning Model

Flawed Models

Support Vector Machine

- Radial Basis Function kernel with the degree of 3
- Accuracy of binary label data set: 65% with validation set | 67.7% with LOOCV
- Accuracy of constructing 7 SVMs corresponding with each species: 65.27%

```
Green frog
[[23  1]
 [21 20]]

Brown frog
[[ 1 10]
 [ 5 49]]

Common toad
[[13  6]
 [23 23]]

Fire-bellied toad
[[36  4]
 [20  5]]

Tree frog
[[32  7]
 [22  4]]

Common newt
[[35  4]
 [22  4]]

Great crested newt
[[51  1]
 [13  0]]

Average accuracy: 0.6505494505494506
```

True Negative	False Positive
False Negative	True Positive

Artificial Neural Network

- 20 hidden layers with the size of 9
- RELU for activation
- Limited-memory BFGS solver for weight optimization
- Accuracy: 67% with validation set | 64% with LOOCV

```
Accuracy: Green frog
[[16  8]
 [13 28]]

Brown frog
[[ 3  8]
 [16 38]]

Common toad
[[10  9]
 [12 34]]

Fire-bellied toad
[[32  8]
 [13 12]]

Tree frog
[[25 14]
 [14 12]]

Common newt
[[32  7]
 [14 12]]

Great crested newt
[[51  1]
 [13  0]]

0.6703296703296703
```

Decision Tree

- Use gini for split criteria
- Max depth of 4
- Accuracy: 68.1% with validation set | 65.1% with LOOCV

```
Accuracy: Green frog
[[21  3]
 [25 16]]
Brown frog
[[ 1 10]
 [ 8 46]]
Common toad
[[11  8]
 [10 36]]
Fire-bellied toad
[[40  0]
 [19  6]]
Tree frog
[[35  4]
 [19  7]]
Common newt
[[36  3]
 [25  1]]
Great crested newt
[[52  0]
 [13  0]]
0.676923076923077
```

Updated Models

Support Vector Machine

- Tune the punishment of soft margin, kernel (poly, RBF), the degree of the poly kernel and balanced the class weight for imbalance species of each SVM
- Accuracy of binary label data set: 70.33%

```
Green frog: 67.6923076923077
[[17  7]
 [14 27]]

Brown frog: 73.84615384615385
[[ 1 10]
 [ 7 47]]

Common toad: 72.3076923076923
[[ 7 12]
 [ 6 40]]

Fire-bellied toad: 60.0
[[28 12]
 [14 11]]

Tree frog: 67.6923076923077
[[26 13]
 [ 8 18]]

Common newt: 67.6923076923077
[[26 13]
 [ 8 18]]

Great crested newt: 83.07692307692308
[[49  3]
 [ 8  5]]

Average accuracy: 70.32967032967034
```

True Negative	False Positive
False Negative	True Positive

Artificial Neural Network

- Tune the number and the size of hidden layer for each ANN
- lbfgs for activation
- Limited-memory BFGS solver for weight optimization
- Accuracy: 72.75%

```
Green frog: 78.46153846153847
[[16  8]
 [ 6 35]]

Brown frog: 81.53846153846153
[[ 1 10]
 [ 2 52]]

Common toad: 70.76923076923077
[[ 9 10]
 [ 9 37]]

Fire-bellied toad: 69.23076923076923
[[38  2]
 [18  7]]

Tree frog: 73.84615384615385
[[35  4]
 [13 13]]

Common newt: 64.61538461538461
[[33  6]
 [17  9]]

Great crested newt: 70.76923076923077
[[43  9]
 [10  3]]

Average accuracy: 72.74725274725274
```

Decision Tree

- Tune splitting criteria, maximum tree depth, number of features to consider when looking for the best split and whether it uses should balance the class weight for each Decision Tree
- Accuracy: 69.45%

```
Green frog: 70.76923076923077
[[14 10]
 [ 9 32]]

Brown frog: 73.84615384615385
[[ 2  9]
 [ 8 46]]

Common toad: 66.15384615384615
[[12  7]
 [15 31]]

Fire-bellied toad: 66.15384615384615
[[33  7]
 [15 10]]

Tree frog: 64.61538461538461
[[31  8]
 [15 11]]

Common newt: 66.15384615384615
[[25 14]
 [ 8 18]]

Great crested newt: 78.46153846153847
[[45  7]
 [ 7  6]]

Average accuracy: 69.45054945054945
```

Ensemble Models

Random Forest

- Set the max depth of 4
- Learning rate: 0.72
- Random state of 6
- Accuracy: 69.9%

```
Accuracy: Green frog
[[21  3]
 [17 24]]

Brown frog
[[ 2  9]
 [ 6 48]]

Common toad
[[11  8]
 [19 27]]

Fire-bellied toad
[[36  4]
 [14 11]]

Tree frog
[[30  9]
 [16 10]]

Common newt
[[34  5]
 [15 11]]

Great crested newt
[[51  1]
 [11  2]]

0.6989010989010989
```

AdaBoost

- Use the previous decision tree classifier for the base classifier
- Learning rate: 0.72
- Random state of 5
- Accuracy: 70.1% with validation set

```
Accuracy: Green frog
[[22  2]
 [25 16]]

Brown frog
[[ 1 10]
 [ 1 53]]

Common toad
[[12  7]
 [20 26]]

Fire-bellied toad
[[38  2]
 [15 10]]

Tree frog
[[36  3]
 [18  8]]

Common newt
[[37  2]
 [19  7]]

Great crested newt
[[52  0]
 [12  1]]

0.701098901098901
```

Stacking using Bayes Optimal Classifier

- Use 6-Fold to choose the best training set to train the Bayes classifer using the base learners (SVM, Decision Tree, ANN)
- Use Gaussian Naive Bayes classifier as meta classifier
- Accuracy: 69.9%

```
Green frog
[[20  4]
 [13 28]]

Brown frog
[[ 0 11]
 [ 2 52]]

Common toad
[[13  6]
 [17 29]]

Fire-bellied toad
[[35  5]
 [20  5]]

Tree frog
[[35  4]
 [17  9]]

Common newt
[[39  0]
 [26  0]]

Great crested newt
[[52  0]
 [12  1]]

0.6989010989010989
```

Classifier	Accuracy
SVM	70.33%
KNN	64.6%
ANN	72.75%
Decision Tree	69.45%
Random Forest	69.9%
Bayes Stacking	69.9%
AdaBoost	70.1%

- Overall ANN based on the decision tree has the best performance of 72.75%
- Because of the uneven distribution and a really small instance, it is hard to get a better accuracy

Thanks!

Phuc Ngo, Beloit College
ngoph@beloit.com