Predicting Oil Exposure in Shore Birds

By Tom Cairns, Gabriel Deacon, Ben Kosiborod, Amy Tang

Introduction

Oil has major impacts on wildlife

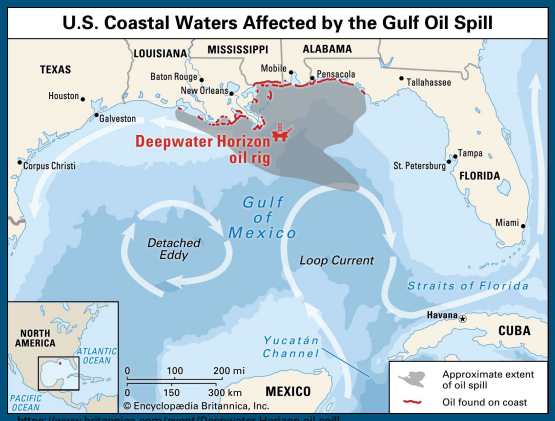
- Reduces water repellent properties of feathers
- Anemia
- Reduced immune function

https://www.motherjones.com/environment/2019/04/deepwater-horizon-bp-oil-spill/

Introduction

Identify markers in birds most impacted by oil spills

Create a model to predict if a bird has been impacted



https://www.britannica.com/event/Deepwater-Horizon-oil-spill

Data

- Report generated to study Deepwater Horizon oil spill in 2010
  - Deepwater Horizon (MC 252) Oil Spill Natural Resource Damage Assessment and Restoration
- Report includes:
  - PDF - written report including explanation of the variables
  - CSV - containing the data

Band Number	Classification	Species	Heinz Body Count	Reticulocytes Count	Plasma Haptoglobin (g/dl)	Plasma Ferritin (ng/ml)	Packed Cell Volume (%)	Hemoglobin (g/dl)	MCHC (g/dl)	Total Solids	PAH (ng/ml)	Heterophil (%)	Heterophil Count (10^9/L)	Lymphocyte (%)	Lymphocyte Count (10^9/L)
1055-04860	Reference	AMOY	0	0	43 NR	NR		48.5 NR	NR		3.8 ND		65	4550	34	2380
1106-06098	Reference	AMOY	0	0	45 NR	NR		47 NR	NR		4.1 ND		58	4640	23	1840
1106-14620	Reference	AMOY				0.34	42.72	45 NR	NR		2.9 ND		55	4840	22	1760
1106-14621	Reference	AMOY				0.23	26.19	47 NR	NR		3.1 ND		64	2560	22	880
1106-14622	Reference	AMOY				0.26	67.99	46 NR	NR		2.6 ND		62	2480	26	1040
1106-14623	Reference	AMOY				0.22	57.41	37 NR	NR		3.1	7.6	59	4130	33	2310
1106-14624	Reference	AMOY				0.24	46.94	48 NR	NR		3.2 ND		60	3600	39	2340
1106-14625	Reference	AMOY				0.25	46.28	48 NR	NR		3.1	6.1	42	1680	52	2080
1106-14626	Reference	AMOY				0.2	21.55	46 NR	NR		3 ND		53	3710	41	2870
1106-14627	Reference	AMOY	0	42	0.17	12.57		55 NR	NR		4.2 ND		68	2720	26	1040
1106-14628	Reference	AMOY	0	44	0.22	16.7		46.5 NR	NR		3.8 ND		59	2950	32	1600
1106-14629	Reference	AMOY	0	50	0.23	53.21		43.5 NR	NR		3.9 ND		58	2320	37	1480
1106-14630	Reference	AMOY	0	35	0.27	49.58		50 NR	NR		3.8 ND		69	4140	17	1020
1106-14631	Reference	AMOY	0	42	0.26	60.42		48.5 NR	NR		4 ND		76	3800	12	600
1106-14632	Reference	AMOY	0	61 NR	NR			51 NR	NR		3.8 ND		68	4760	14	980
1106-14642	Reference	AMOY		NR	NR	NR		NR	NR	NR	NR	NR	NR	NR	NR	NR
1106-14643	Reference	AMOY		NR	NR	NR		NR	NR	NR	NR	NR	NR	NR	NR	NR
1106-14645	Reference	AMOY		NR	NR	NR		NR	NR	NR	NR	NR	NR	NR	NR	NR
1106-14646	Reference	AMOY		NR	NR	NR		NR	NR	NR	NR	NR	NR	NR	NR	NR
1106-14647	Reference	AMOY		NR	NR	NR		NR	NR	NR	NR	NR	NR	NR	NR	NR
1106-14648	Reference	AMOY		NR	NR	NR		NR	NR	NR	NR	NR	NR	NR	NR	NR
1106-14649	Reference	AMOY		NR	NR	NR		NR	NR	NR	NR	NR	NR	NR	NR	NR
1106-14650	Reference	AMOY		NR	NR	NR		NR	NR	NR	NR	NR	NR	NR	NR	NR
1106-14651	Reference	AMOY		NR	NR	NR		NR	NR	NR	NR	NR	NR	NR	NR	NR
1106-14652	Reference	AMOY		NR	NR	NR		NR	NR	NR	NR	NR	NR	NR	NR	NR
1106-14653	Reference	AMOY		NR	NR	NR		NR	NR	NR	NR	NR	NR	NR	NR	NR
1106-14654	Reference	AMOY		NR	NR	NR		NR	NR	NR	NR	NR	NR	NR	NR	NR
1106-14658	Reference	AMOY		NR	NR	NR		NR	NR	NR	NR	NR	NR	NR	NR	NR
1106-14698	Reference	AMOY			0.44	67.35		47 NR	NR		2.7 ND		62	2480	29	1160
1106-14699	Reference	AMOY			0.22	44.22		48 NR	NR		3.9 ND		51	3570	40	2800
1106-16301	Impacted	AMOY			0.23	30.95		27	13	46.14814815	4 ND		56	6400	32	4000
1106-16302	Impacted	AMOY			0.24	19.42		46	7.2	15.65217391	4.8	11.3	60	3600	36	2160
1106-16303	Impacted	AMOY	0	55	0.4	50.81		38	15.4	40.52631579	4.8 ND		42	3360	38	3040
1106-16304	Impacted	AMOY	0	51 NR	NR			43	19.9	49.27906977	4.2 ND		36	2160	53	3180
1106-16305	Impacted	AMOY		NR	NR	NR		NR	NR	NR	ND		71	3550	15	750
1106-16306	Impacted	AMOY		NR	NR			41 NR	NR		4.8 ND		57	3990	26	1820
1106-16307	Impacted	AMOY	10	82 NR	NR			34	12.8	37.64705882	3.8	9.4	49	2940	32	1920
1106-16308	Impacted	AMOY	0	66	0.22	14.17		40	15.7	39.25	4.6 ND		49	4410	26	2340
1106-16309	Impacted	AMOY	36	75	0.17	22.65		38	15.4	40.52631579	4.8 ND		42	2940	37	2590
1106-16310	Impacted	AMOY	0	48	0.42	50.93	NR	NR			4.8 ND		47	3760	41	3280
1106-16311	Impacted	AMOY	24	74	0.25	40.76		41	9.5	22.99302326	4.2 ND		33	2420	48	3360
1106-16312	Impacted	AMOY	0	69	0.22	38.01		41	14.9	36.34146341	4.4 NR		42	2940	48	3360
1106-16313	Impacted	AMOY	42	84 NR	NR			38	14.7	38.68421053	4 NR		45	3150	25	1750
1106-16314	Impacted	AMOY	0	54	0.21	17.45 NR		NR	NR	NR	NR		60	4200	27	1890
1106-16315	Impacted	AMOY	0	56	0.25	76.59		44	16.3	37.04654545	4.8 NR		32	2560	57	4560
1106-16317	Impacted	AMOY	0	61 NR	NR			44	22	50	4.4 NR		32	1600	47	2350
1106-16318	Impacted	AMOY	0	61 NR	NR	NR		NR	NR	NR	NR		51	3570	18	1260
1106-16319	Impacted	AMOY	31	73 NR	NR			47	15.9	33.82978723	4.9 NR		39	3900	47	4700
1106-16320	Impacted	AMOY	0	70 NR	NR			44	16.7	37.95454545	4.8 NR		58	2320	26	1040

Variables include: Band-number, species, classification (whether the bird was impacted or a reference), various physiological markers (PAH, Heinz Body Counts, Reticulocytes)

Shape: 1257 rows (samples) X 67 columns (variables)*

** prior to cleaning the data*

Central Hypothesis:

Oil-spill impacted birds have different levels of physiological markers compared to non-impacted (reference) birds.

Questions and Hypotheses

Questions:

1. Does an increase of PAHs (ng/mL) in the blood indicate that a bird was impacted by an oil spill?
2. Are the reticulocyte counts between impacted and non-impacted birds different?
3. Are the metabolisms of impacted and non-impacted birds different, and can this be seen a physiological marker?

Hypotheses:

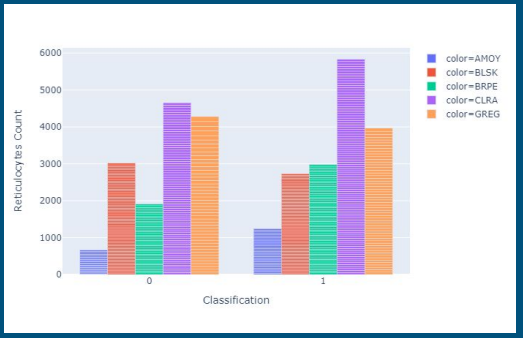
1. An increase in PAHs will indicate that a bird was impacted by oil pollution.
2. There is a decrease in reticulocyte counts for impacted birds compared to non-impacted birds.
3. There is a decrease in the metabolic marker, uric acid, in impacted birds compared to non-impacted birds.

Supervised Machine Learning Algorithms

- Tested 5 different algorithms:
  - K-Nearest Neighbors-based classification (KNeighborsClassifier)
  - Linear Support Vector Classification (LinearSVC)
  - Gaussian Naive Bayes (GaussianNB)
  - Decision Tree classifier (DecisionTreeClassifier)
  - Logistic Regression classifier (LogisticRegression)
- **Used Decision Tree classifier**

Results overview

- Data Cleaning
  - Dropped rows that had no useful data
  - Replaced empty/no record values with NaN
  - Cleaned up continuous data
  - Cleaned up categorical data
- Data Exploration
  - Explored relationship of data related to hypotheses



Results overview cont'd

- Feature Selection & Model Construction
  - Selected features using Decision Tree Classifier-based feature selection
  - Tested proposed hypotheses using aggregation by species and independent samples t-test
- Model Evaluation & Optimization
  - Scaled features
  - Cross validation approach *and* percentage-split approach
  - Evaluated accuracy on selected & scaled features on 5 different classifiers
  - Hyperparameter tuned all 5 models
- Model Testing
  - Tested our best scoring model (Decision Tree Classifier) on our test set
  - Used confusion matrix and classification report to report metrics of our selected model

Hypothesis Discussion

- Null hypothesis: means are not significantly different between reference and impacted birds (by species) for each feature we tested
  - PAH (ng/mL)
  - Reticulocytes Count
  - Uric Acid (mg/dL)
- Alternative hypothesis: means are significantly different between reference and impacted birds (by species) for each feature we tested
- We failed to reject all 3 of our null hypotheses, as for each hypothesis, not all the species showed a significant difference between means of the reference and impacted groups

ML Discussion

- Of the 5 classifiers that were evaluated, we ended up using the Decision Tree Classifier due to its high accuracy across both test/train split and cross validation metrics, as well as the fact that we believed it had the least likelihood of underfitting.
- There was no benefit from performing hyperparameter tuning on the Decision Tree Classifier, and even though other models did benefit from hyperparameter tuning, they still performed worse than the Decision Tree model.

ML Discussion

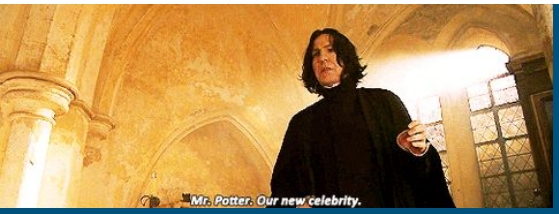
- The results from the classification report constructed from running the model against the testing data indicated that our model correctly labels reference birds 90% of the time and impacted birds 94% of the time.
- These scores could be improved with more data--this could either be in the form of more records (rows) in the dataset, or it could be more complete data in columns that we elected not to use as features for our model due to issues with the existing data in those columns.

Confusion Matrix

	Predicted reference	Predicted impacted
Actual reference	60	5
Actual impacted	7	74

Classification Report

	precision	recall	f1-score	support
0	0.90	0.92	0.91	65
1	0.94	0.91	0.92	81
accuracy			0.92	146
macro avg	0.92	0.92	0.92	146
weighted avg	0.92	0.92	0.92	146


Mr. Potter. Our new celebrity.


It's leviOsa.

Thank you


I'll be in my bedroom, making no noise and pretending
that I don't exist.