**Predicting turtle bycatch in a pelagic longline fishery with machine learning models**

**Problem statement:** determine the factors/variables that contribute most to the capture of sea turtles in a pelagic longline fishery, and save the model to predict the probability of future captures with a new set of independent variables.

A change in bycatch can mea any of these 3 things: 1) change in population size, 2) change in effort, 3) change in fishing gear (i.e. mitigation measures/devices, area).

The turtles catches appear in 1% of the sets. The turtles are caught alive. Since finding a pattern in the reason for capture is difficult, making the avoidance difficult, the priority should be on ensuring the turtles are released alive and the gear+bait used minimises the chance of capture.

**Data pre-processing:**

* SQL turtle data check.sql (SA1 SAold2 combined before cleaning.csv)
* SA1 SAold2 combined cleaning.csv (Phase 1,2,3,3\_1),
* Turtle bycatch prediction EDA.ipynb

Dependent variable: the catches per set must be comparable so measured in number of turtles caught/1000 hooks.

* Catch records for Green, Hawksbill, Olive Ridley, Leatherback, Loggerhead, Turtle (Marine (unid) turtles). (i) Total number and (ii) CPUE calculated as no./1000 hooks

Independent variables:

The data was split between two databases, observer records from 2002-2011 and observer records from 2012-2018. The common fields between the two datasets are:

* Fleet
* Year
* Month
* Day
* Lat
* Long
* No. hooks
* SST
* No. buoys
* No. light sticks

No data is available on what has been shown in previous studies to impact on turtle catches, bait type and hook type.

Data cleaning:

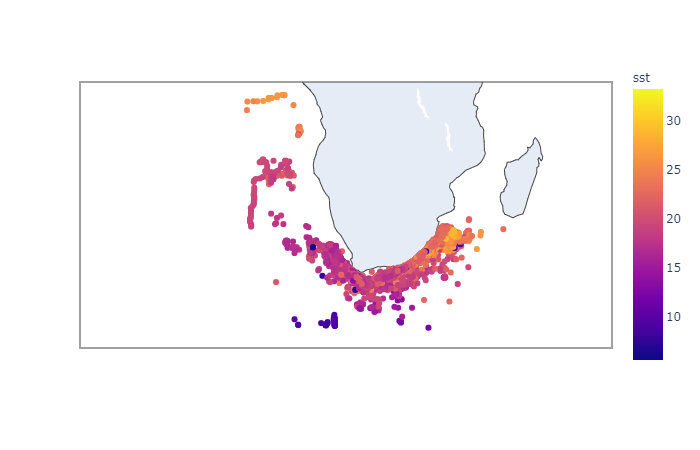
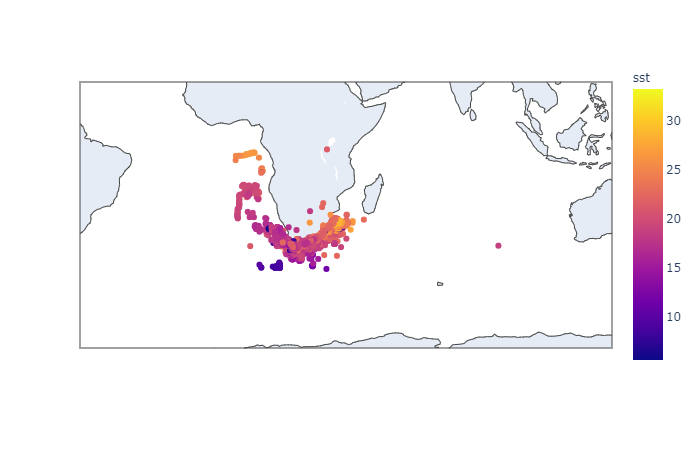
* Number of sets before cleaning: 11019
* Number of turtles caught before cleaning: 185
* Number of sets with turtles before cleaning:176 (1.59% of sets)

MYSQL:

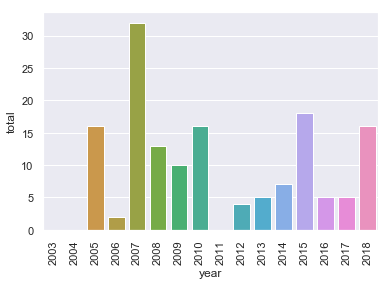
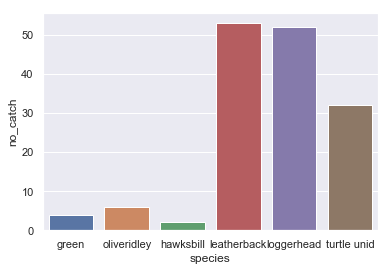
* Year: 208, 209, 330, 710, 1958 corrected. 3 0 records removed
* Lat: 1 zero record removed
* Long: 1 zero record removed
* SST: 2 large outliers (197, 198). SST > 50 (293 records) converted from Fahrenheit to Celsius. 383 zero records removed. SST between 1 and 5 removed due to uncertainty (909 records)
* No. hooks: 14 zero records removed. Histogram plot (python) shows outliers > 5500 hooks, removed 5 records
* Lightsticks: 8711 zero records. Removed this variable
* No. buoys: 10 zero records. Check collinearity with number of hooks to consider removing
* Turtle CPUE: number of turtles / 1000 hooks
* Turtle\_YN: Turtle catch presence/absence
* Fleet: JV=1, SA=0.

Python and MYSQL:

* Plotted Lat and Long with plotly package: removed 29 records on land in MySQL (Set\_id:8742\_3,11904-53, 6861-36, 6902-8, 11904-13, 11701-1, 9012-3, 11701-9, 4493-2, 12840-3, 8871-2, 10480-68, 11046-35, 11046-25, 12772-36, 12127-41, 8536-16, 10476-23, 12130-14, 12773-43, 6907-26, 6695-10, 8339-65, 9193-2, 8702-19, 8702-25, 10478-21, 7299-9, 11554-19, 7244-58



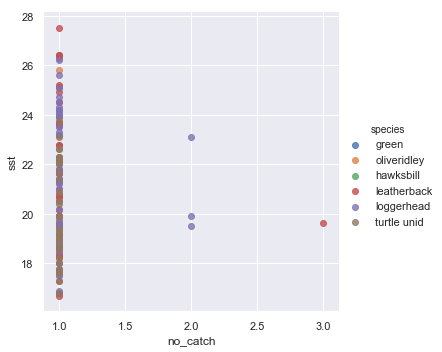
* Number of sets after cleaning: 9662
* Number of turtles caught after cleaning: 149
* Number of sets with turtles after cleaning: 141 (1.46% of sets)

No pattern in catches since for the time period.

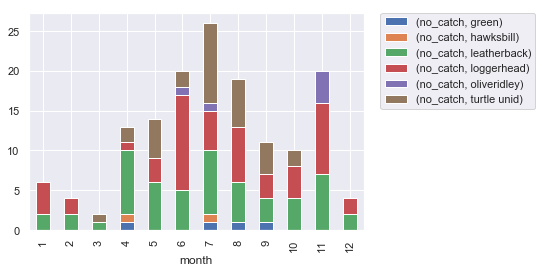
Leatherback and loggerhead turtles dominated the catches, followed by unidentified turtles.

Relationship between dependent and independent variables:

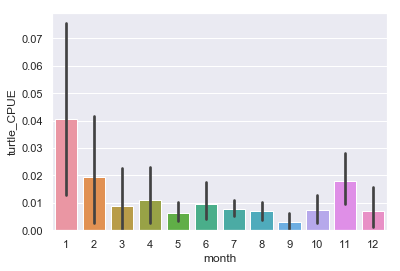
Relationship between SST and turtle catch – no visible relationship



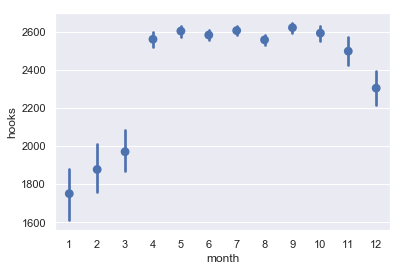
Relationship between month and turtle catch: most turtles caught between April and November.



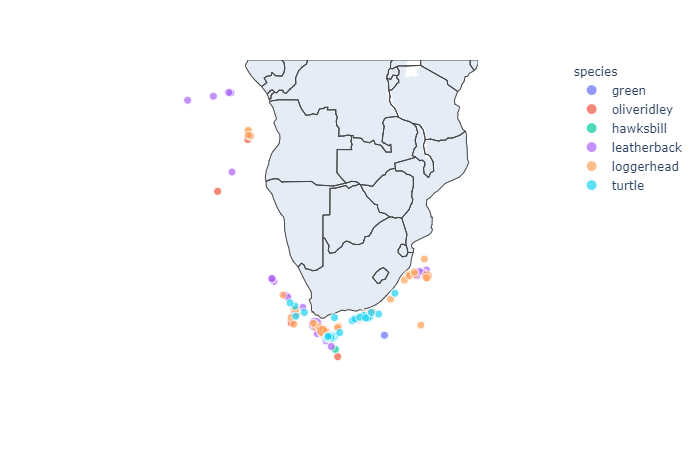
Relationship between month and CPUE: highest CPUE in Jan and Feb but skewed since these months had the least number of hooks set.



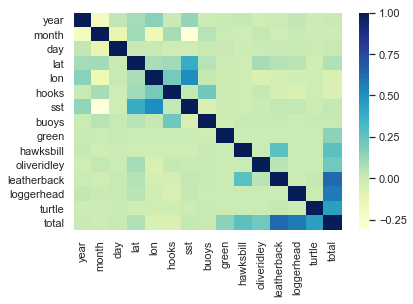
Average hooks per month:



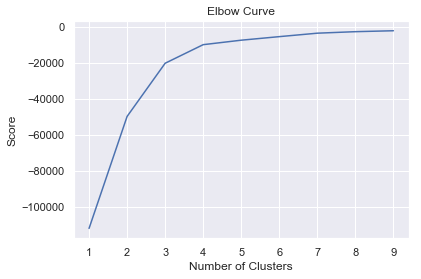
Location of turtle catches

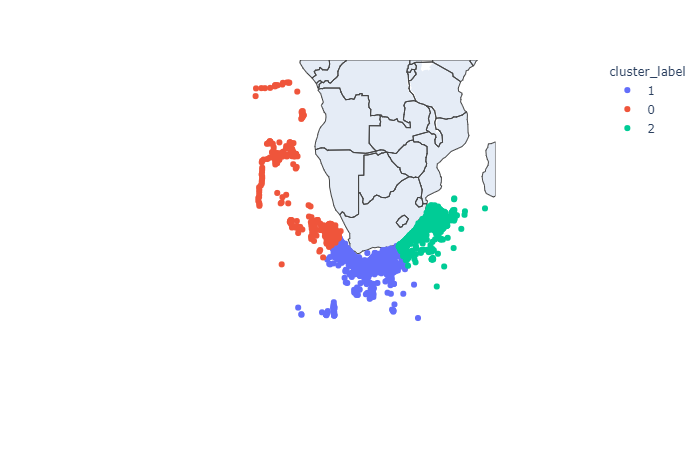


Multicollinearity: SST+Lat, SST+Long, Hooks+Buoys, Leatherback+Hawksbill, Hooks+Long



Cluster coordinates into areas (category type):

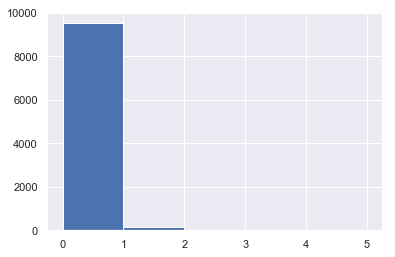




* Placed sst, buoys, hooks into a category, instead of Label Encode.

**Check normality and distribution of y-variable and determine the best model:**

Machine learning models don’t require the data to be normally distributed.



**Model creation: Decision Tree Classifier (CART)**

<https://stackabuse.com/decision-trees-in-python-with-scikit-learn/>

* turtle\_data\_ready.csv
* Turtle bycatch prediction EDA.ipynb

[Independent variables can be grouped together into one variable through cluster analysis. (K-means clustering). Could be included with 2012-2018 data that includes the setting strategy (Setting speed, Vessel speed, Mainline length, Branchline length, Nr branchlines betwn buoys = Hooks per basket)]

Final independent variables:

* Fleet – 1 JV, 0 SA
* Month - category
* Year - category
* Lat+Long = cluster 0, 1, 2
* No. buoys = 1 (<200), 2 (200 <= sst < 250), 3 (250 <= sst < 275), 4 (>=25)
* SST = 1 (<10), 2 (10 <= sst < 15), 3 (15 <= sst < 20), 4 (20 <= sst < 25), 5 (>=25)
* No. hooks = 1 (<1000), 2 (1000 <= sst < 1500), 3 (1500 <= sst < 2000), 4 (2000 <= sst < 2500), 5 (2500 <= sst < 3000), 6 (>= 3000)

Final dependent variable:

The presence (“1”) or absence (“0”) of a turtle in the set. The total catch and CPUE is too low to be used.

Overfitting versus Underfitting a model:

When a model is overfitted, it can predict the result of the current data well, but has poor predictability for new data. Solution: hide a small part of the database from the algorithm. Train the model based on most of the data, and test the data on the small part.

**Model testing**

The goal is to not adjust the model any further.

The result indicates how often (% accuracy) the model will predict the level of absenteeism correctly. If the test result was 10-20% less than the trained model, then the model is overfitted and will fail in real life.

Accuracy score:

The score indicates the percentage of the observations that the model had learnt to classify correctly.

**Importance of each variable**

It is not necessary that the more important a feature is then the higher its node is at the decision tree.

fleetnr - 0

month - 0.07774299

hooks - 0.55928324

sst - 0.21696896

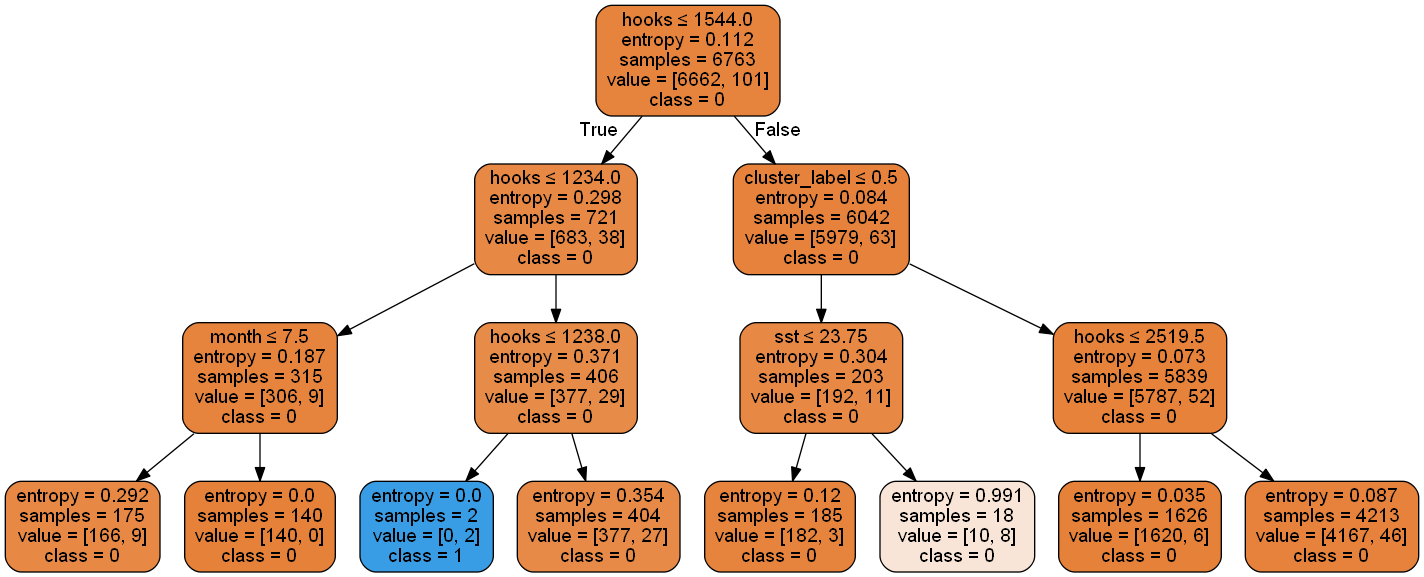
buoys - 0

cluster\_label - 0.14600481

**Probability estimates:** indicate, per record, the probability of resulting in “0 (no turtle caught)” or “1 (turte caught)”. If probability is <0.5, then “0” result, if >0.5 then a “1”.

**Save the model** – the model can be loaded to run on new data

**Visualise the tree**

****

Weak prediction of turtle catches.

Hooks <1544, values [0 = 6662, 1 = 101]

Hooks < 1234, values [0 = 683, 1 = 38]

Hooks < 1238, values [0 = 377, 1 = 27] -> turtle catch

Other variables used in the tree: month <7.5, cluster\_label <0.5, sst <23.75