Global Shark Attacks
Predicting Fatality

Introduction

Shark attacks are a rare occurrence but are becoming more frequent as coastal populations increase. The global shark attack data has been accumulated over the last few centuries and we can train a model to predict whether an attack is fatal or non-fatal.

Three models will be tuned then compared:

- Random Forest
- Support Vector Machine
- Logistic Regression
- Naive Bayes

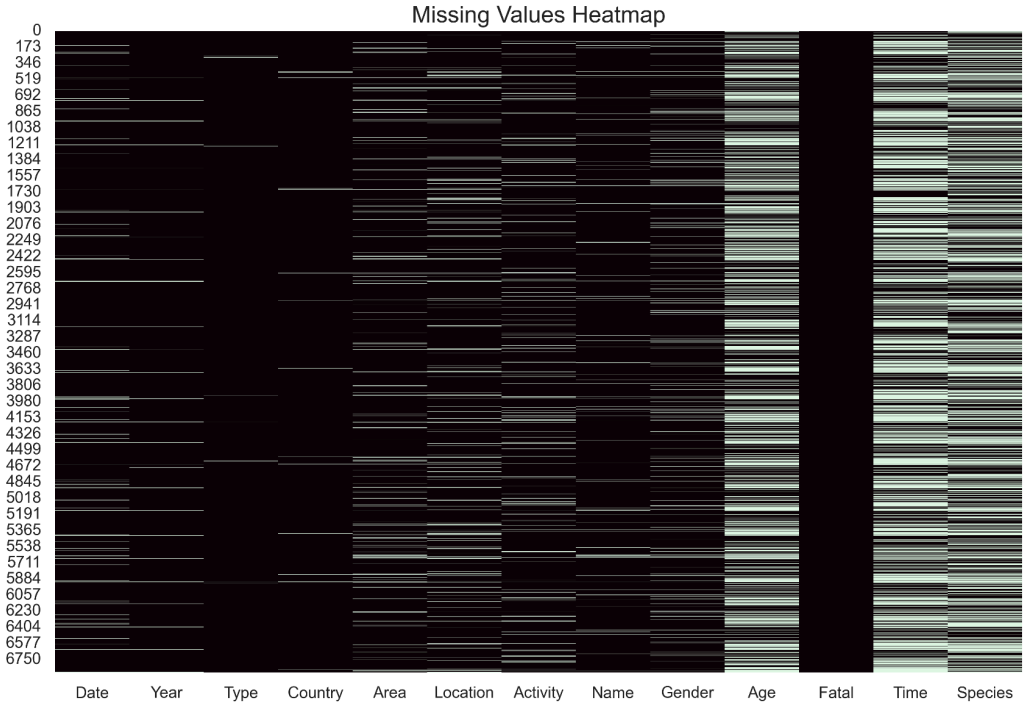
Requirements

The following libraries are necessary to run the Jupyter notebook:

- numpy
- pandas
- warnings
- matplotlib
- seaborn

Procedure

The activity and species values were consolidated to remove extra whitespace, spelling errors with descriptive text converted into categories. Missing data was quite frequent in the Age, Time and Species columns.



Age was cleaned by setting the missing values to the median rather than the mean since the distribution was not gaussian. Age was the only numeric feature.

Time was not used in prediction since it can be difficult to determine values to fill missing data with.

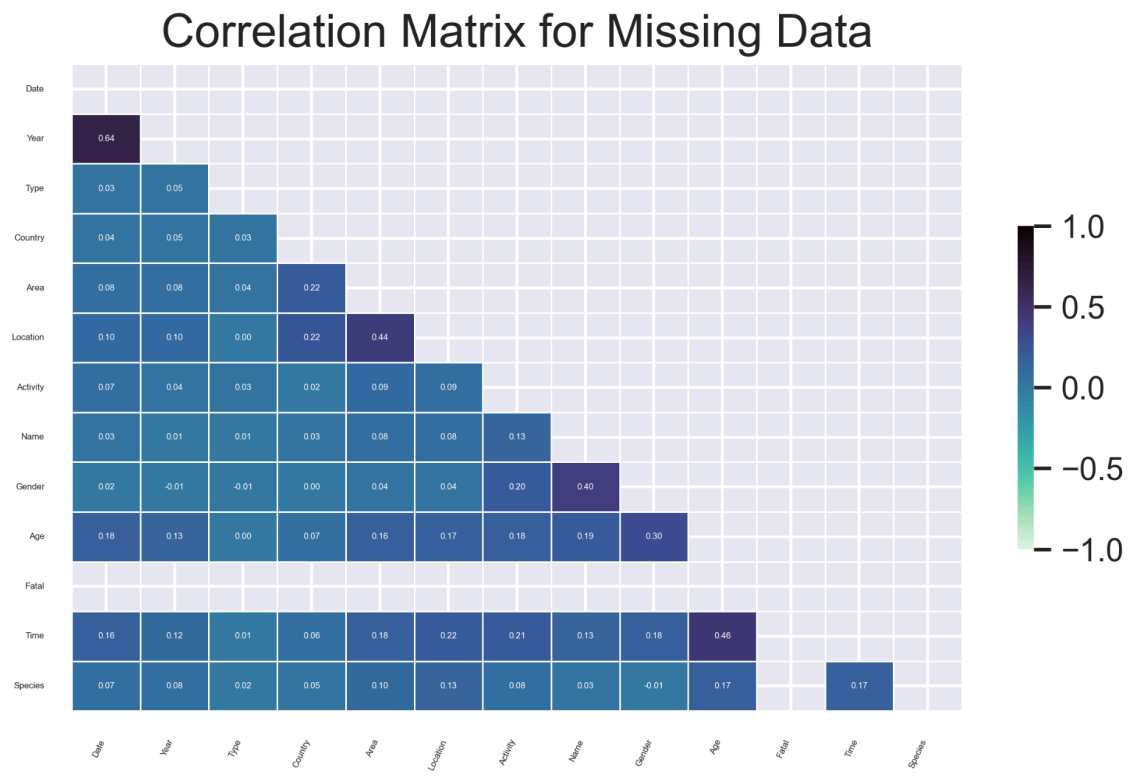
Species missing data was significant. There were 3 categories of missing data.

- Other
- NA
- Description of estimated shark length

The 3rd category was created as a value and it played a role in the SVM model scores. If we have a description of length, the victim did not die or there were witnesses (nearshore).

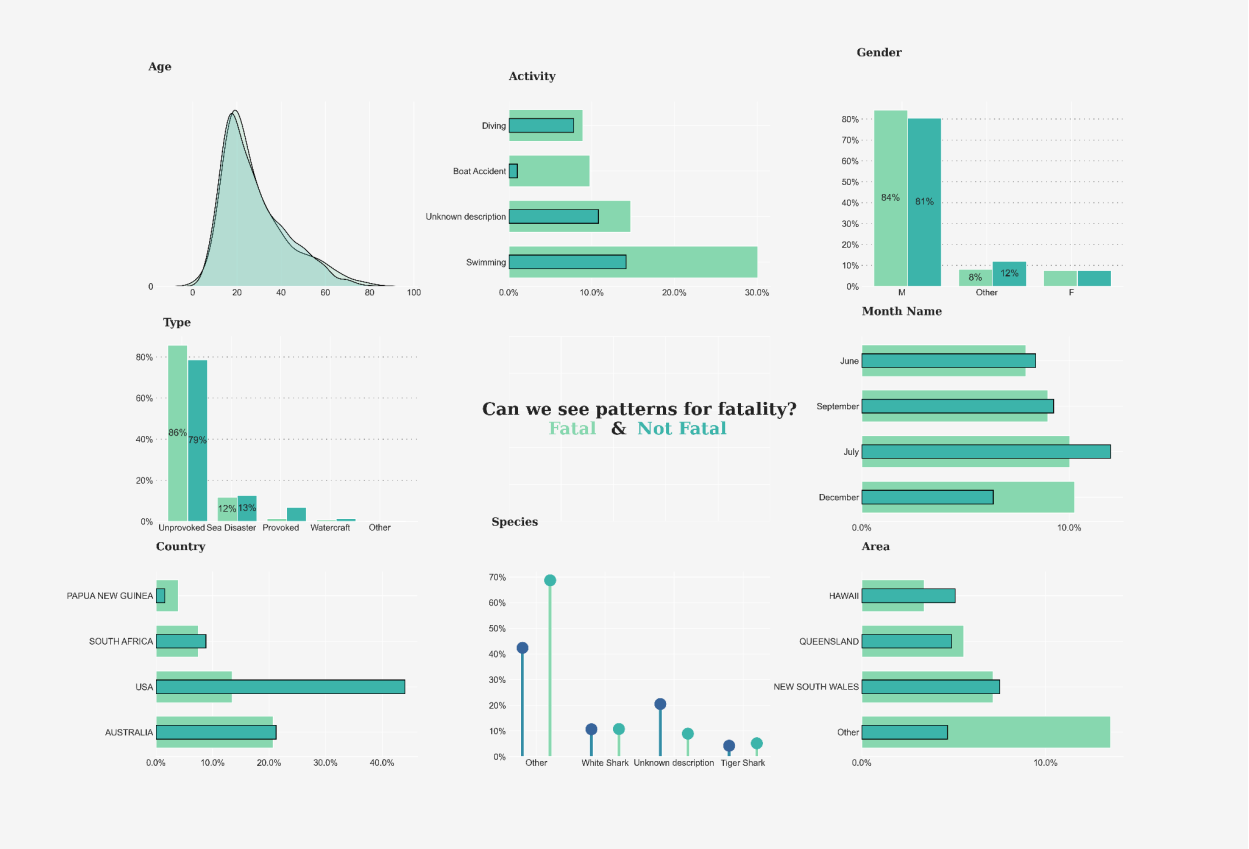
I chose not to set the missing data to the mode because it would taint the data with too much synthetic info. The vast majority of the cases (over 80%) would become Great Whites which is invalid. For example, most attacks in Florida are NOT Great Whites. I tested this by running a prediction with the models and achieved poor F1 scores.

Activity was consolidated into simple categories such as Surfing, Swimming, Floating.



Most of the data is categorical which requires a Chi-Square Test. The strongest correlation is between Area and Country at -0.45. Fatal with Year is -0.31 which means shark attacks

are becoming less fatal perhaps due to improvements in rescue, an increase in witnesses to help the victim to shore and medical care.

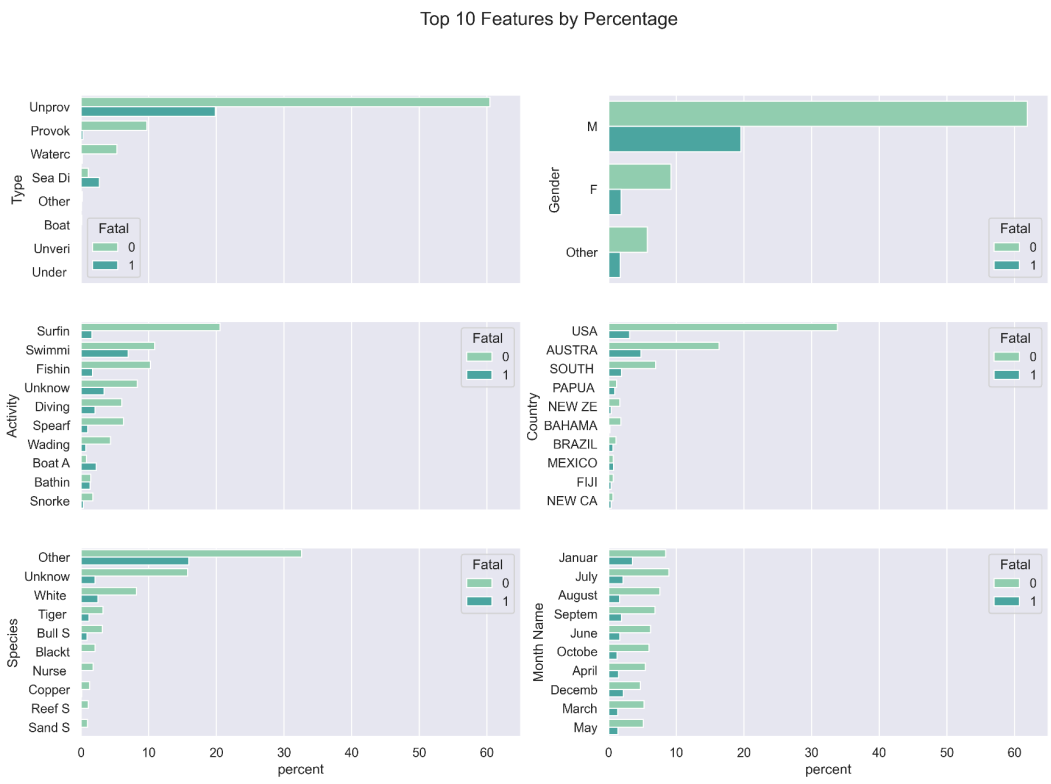


The above plots each feature ordered by fatal percentage. Swimming has the most fatalities of activity at about 30%. Boat accidents are next at about 10%. In fact, boat accident shark attacks have an overwhelmingly high number of fatalities relative to non-fatal outcomes.

Some factors that may come into play include access to rescue, lack of barrier between shark and victim(s) and more than one person in the water for multiple days until being rescued.

It's also evident that December has a high number of fatalities relative to non-fatalities. This may be related to shark migration behavior.

The US has a very high number of non-fatal attacks. I think Florida skews this since most US attacks are in Florida.



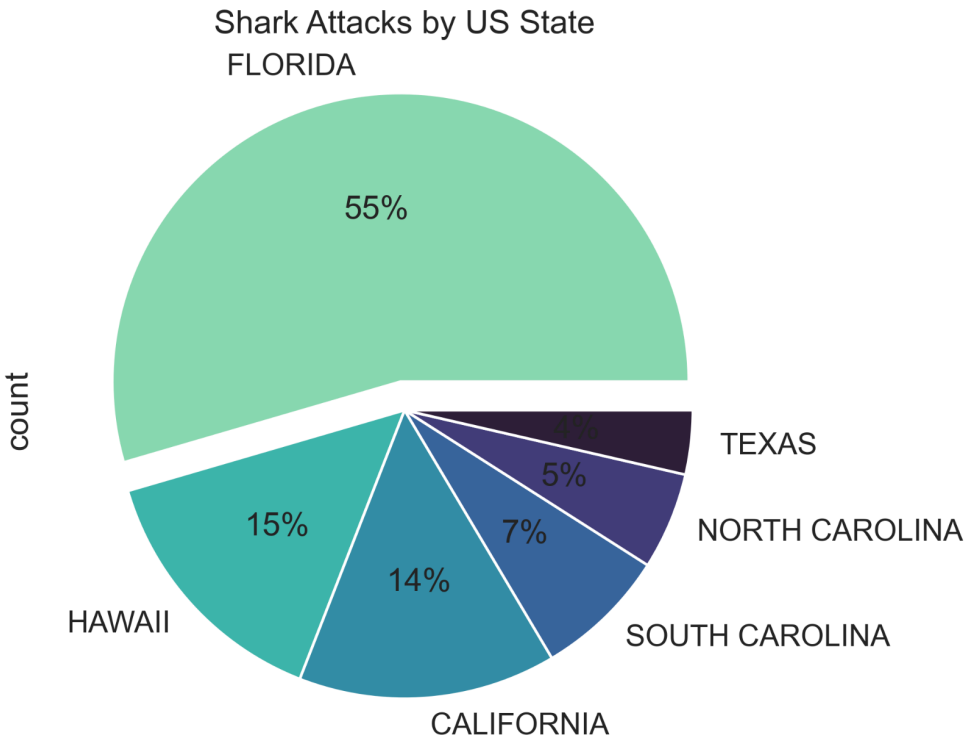
The above chart is plotting each feature using hue and percentage as the stat.

An interesting followup would be to geo-encode the location, area, country features to get the latitude longitude. Then create bounding boxes which can be encoded. In this run, I would include a KNN model comparison.

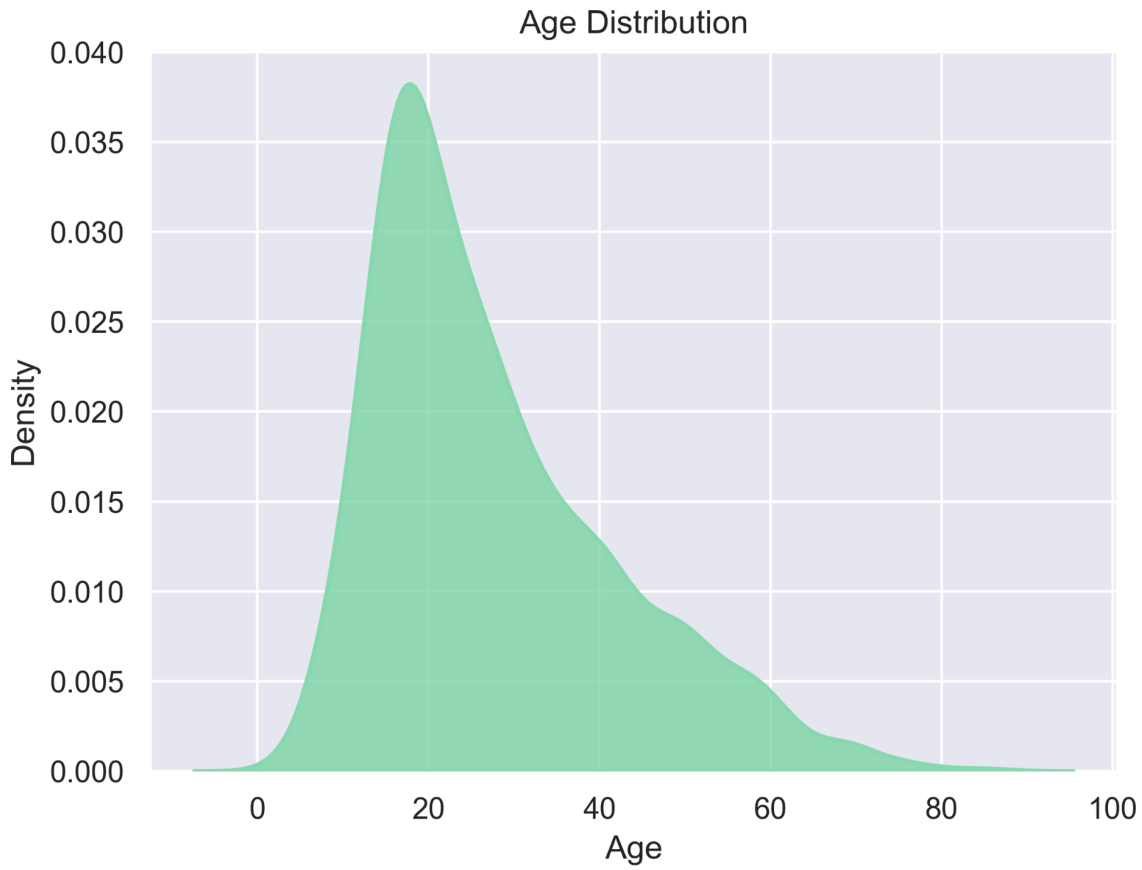
The issue I had with geo-encoding was that the Oceanic locations were invalid. I will need to manually get the actual coordinates as multiple sources were invalid. Another options is getting a-hold of the Florida dataset which is guarded for research purposes only.

I would also consider creating a new category of prey behavior (splashing, jumping, floating, swimming) and nearshore vs offshore vs open ocean activities. A couple of insights: access

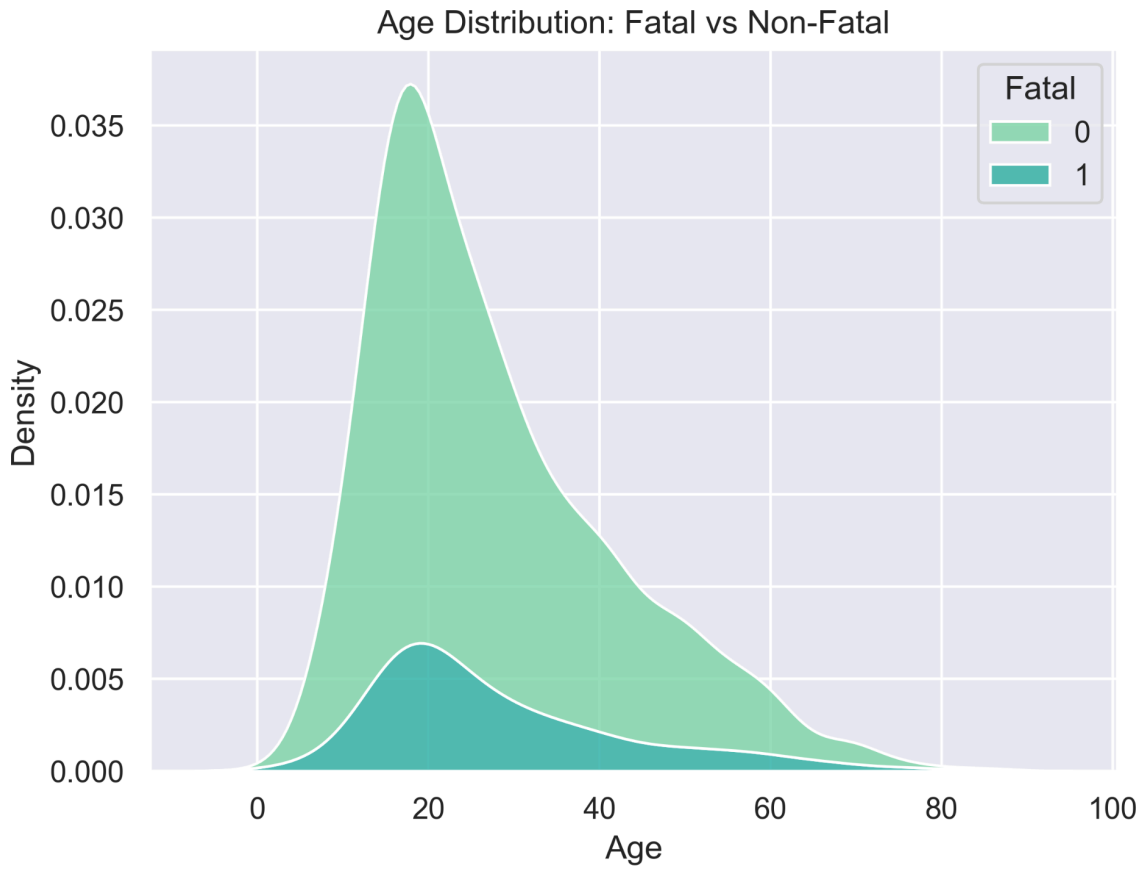
to rescue is probably important. Life guards are able to monitor nearshore activities such as surfing but not offshore or open ocean. Also, defensibility may be important. A surfboard acts as a barrier whereas swimming has no barrier to the shark.



Florida has the largest number of shark attacks, however most are not fatal. Also most attacks are not the top 3 most dangerous species (Great White, Tiger, Bull).

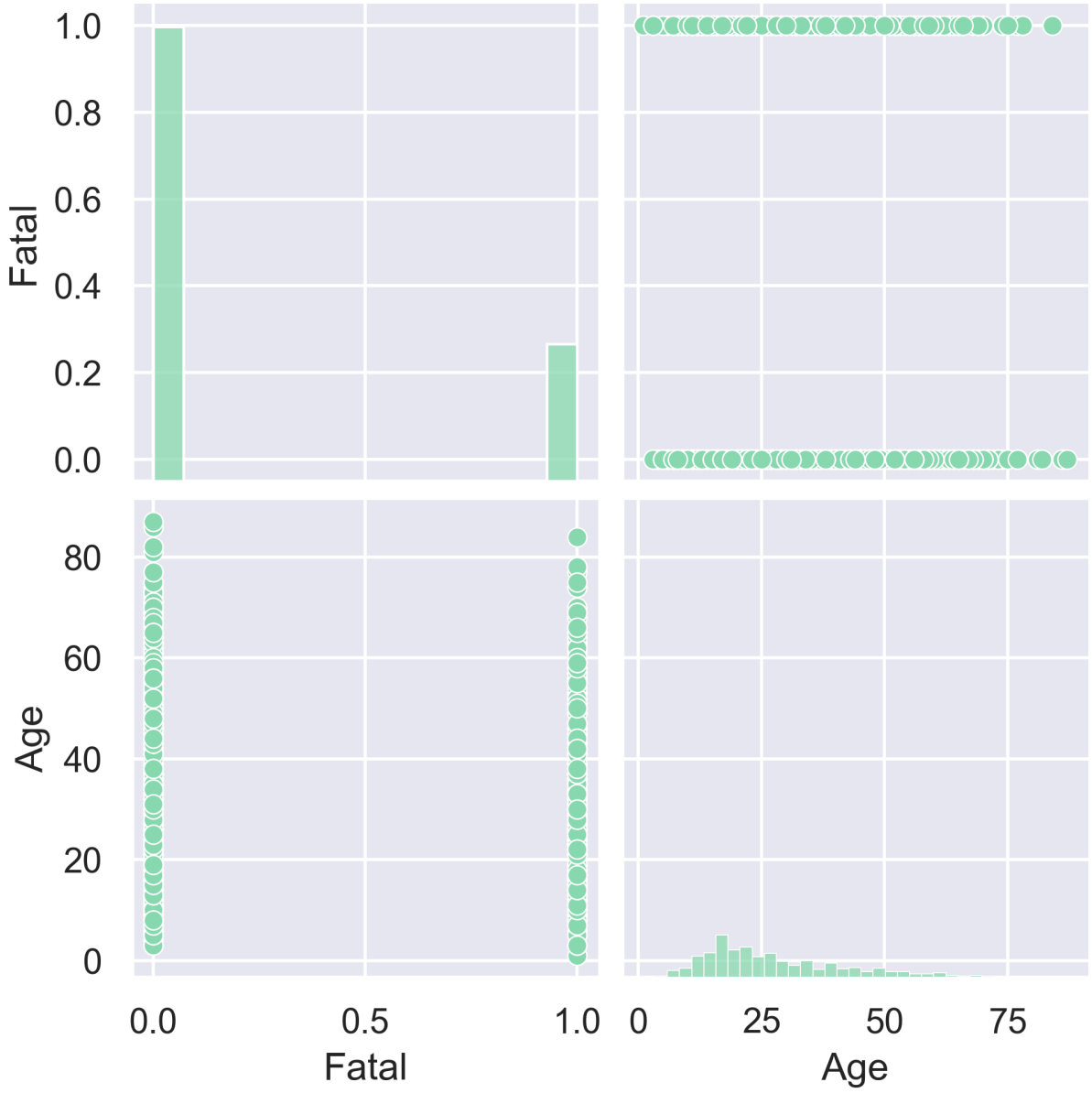


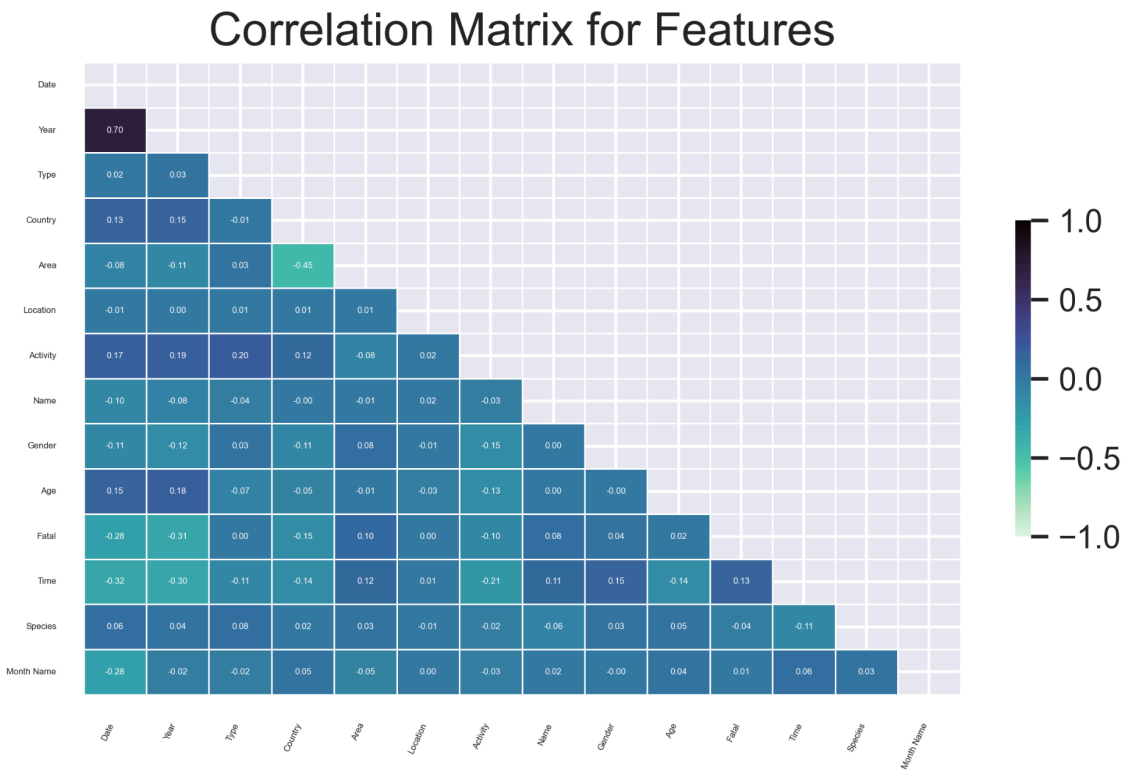
The age distribution (before the median is applied to the missing values) is skewed left. This may be because more younger people in the 18-30 range are in the water.



The age distribution for fatal is skewed only slightly right of non-fatal.

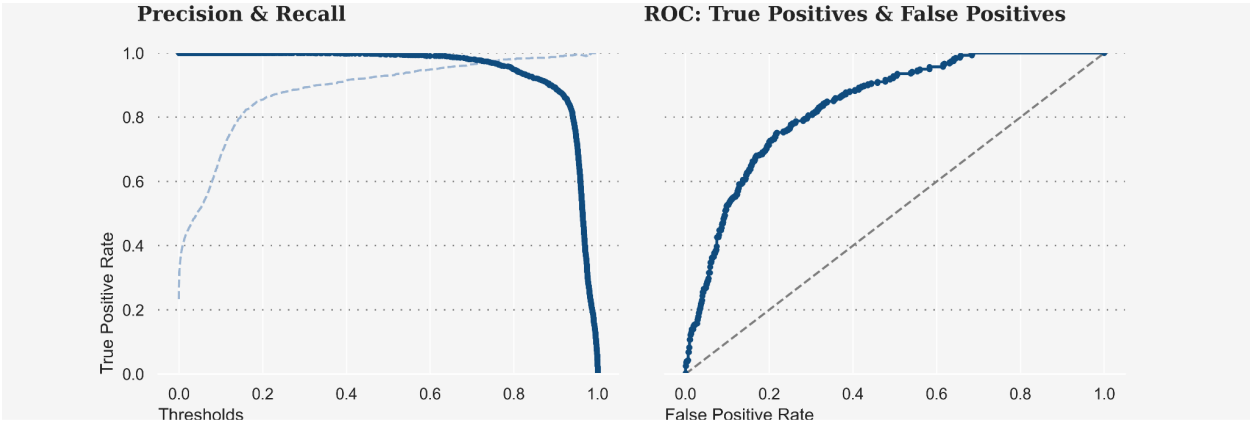
Next, I plotted the correlation between the only numerical feature and the target fatal with pairplots.


1.0
0.8
0.6
0.4
0.2
0.0
Fatal
80
60
40
20
0
Age
0.0
0.5
1.0
Fatal
0
25
50
75
Age

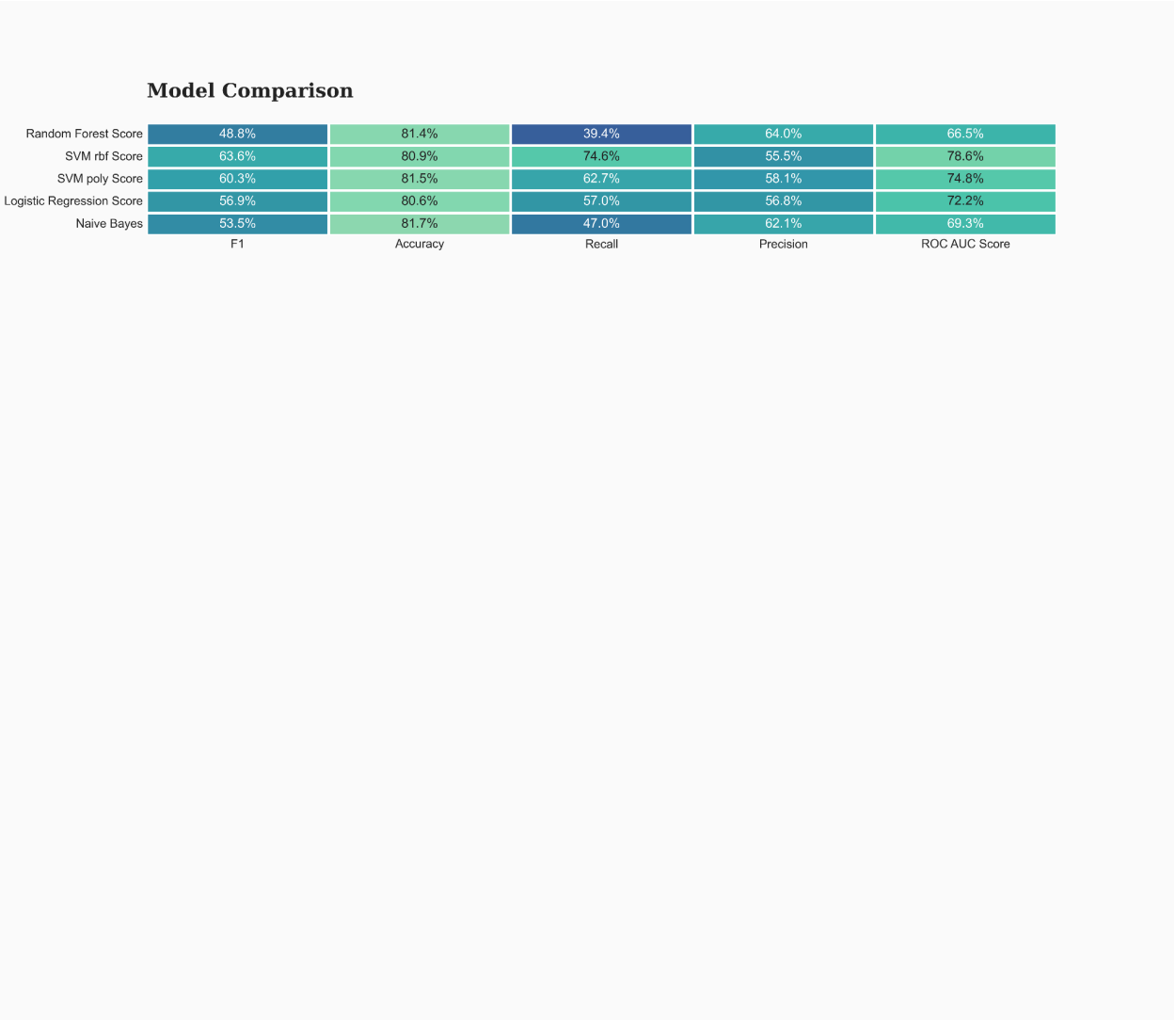


The plot below shows Precision and Recall as well as ROC curves for Logistic Regression.

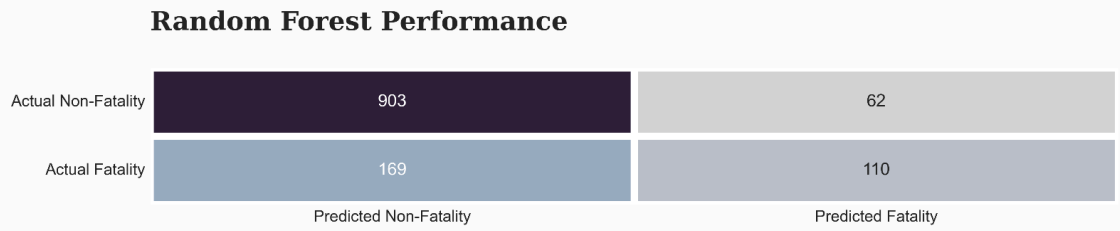
We can see that Precision falls off dramatically at about 88%.

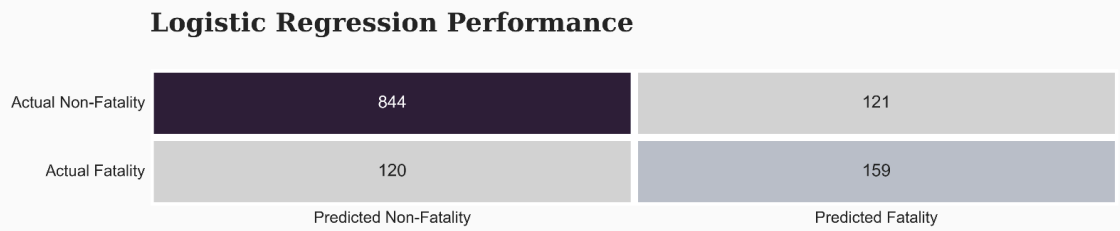


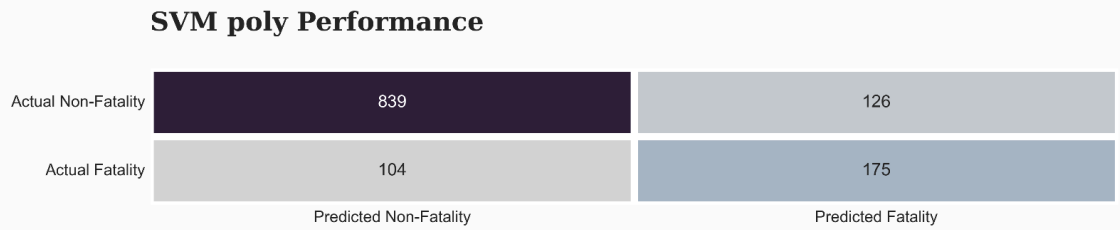
The model comparison shows SVM outperforming both Random Forest and Logistic Regression. The complexity of the data is enough to give SVM the advantage. Also, I wonder if this can mean oversampling is occuring.

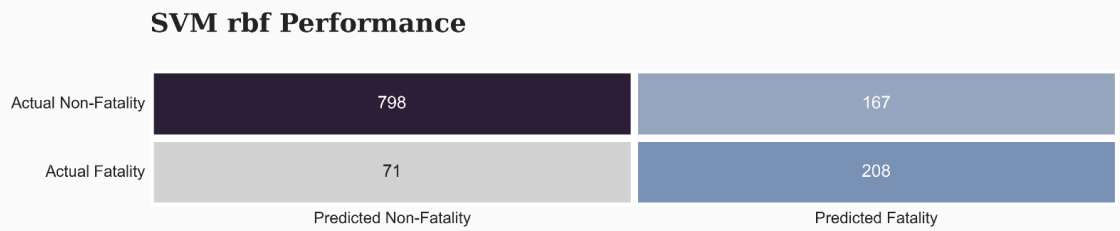


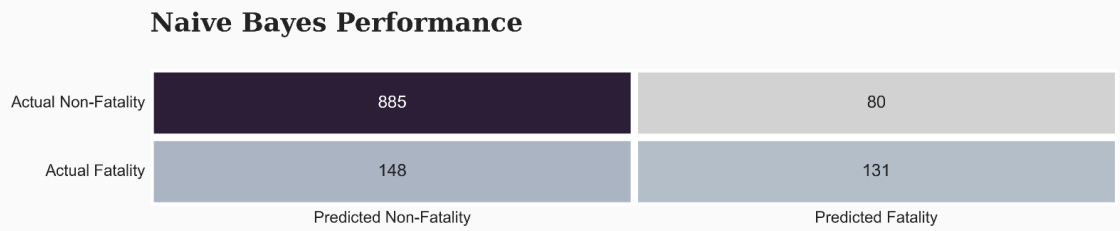
The confusion matrix for each model is shown below:

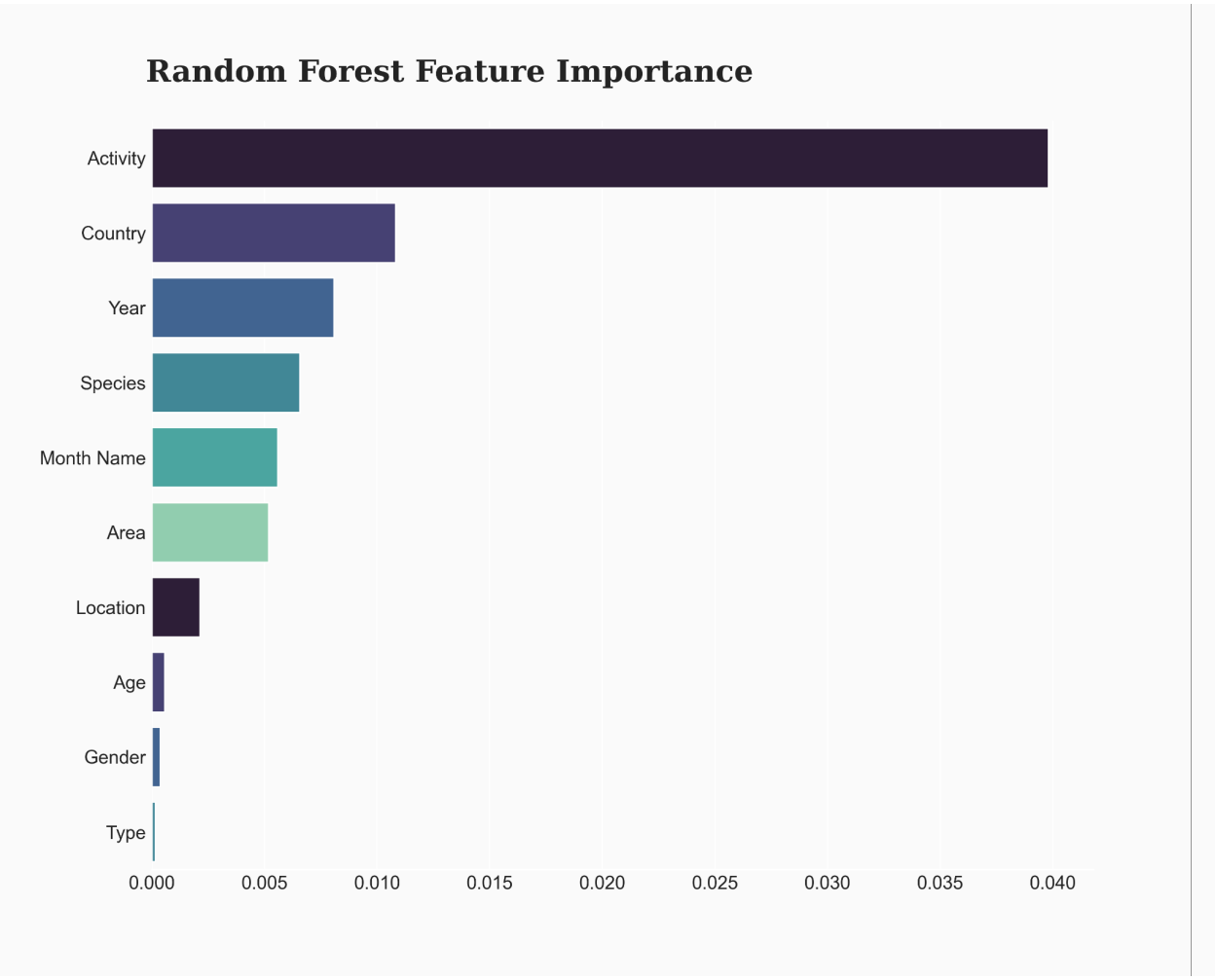

Random Forest Performance
Actual Non-Fatality
Actual Fatality
903
62
169
110
Predicted Non-Fatality
Predicted Fatality


Logistic Regression Performance
Actual Non-Fatality
Actual Fatality
844
121
120
159
Predicted Non-Fatality
Predicted Fatality


SVM poly Performance
Actual Non-Fatality
Actual Fatality
839
126
104
175
Predicted Non-Fatality
Predicted Fatality


SVM rbf Performance
Actual Non-Fatality
Actual Fatality
798
167
71
208
Predicted Non-Fatality
Predicted Fatality


Naive Bayes Performance
Actual Non-Fatality
Actual Fatality
885
80
148
131
Predicted Non-Fatality
Predicted Fatality



For Random Forest, the feature chosen as the most important is Activity. Next is Country, then Year.

Conclusion

The tuned SVM models were able to predict whether an attack is fatal with pretty decent performance for the full dataset. The accuracy for all models was in the low 80s percent. This was surprising due to the lack of shark species information. The models heavily favored human activity as important over all other features.

I think the results can be improved with accurate geo-encoding of locations which will reflect human population density, and shark migration patterns. This will add more real data which perhaps will offset the large amount of missing species data.