Modeling Shark Attack Occurrence: A Supervised Learning Approach

Introduction/Background

Shark attacks are extremely rare events with great consequences for the victims involved. These events have become widely popularized in modern culture and media because of their violent and novel nature. That said, public exposure to these events widely inflated relative to their actual occurrence. Individual risk, as reported by the ISAF is 1 in 11.5 million, but this can vary based on the profile of the individual (e.g. individual's number of visits to the beach, activity while in the water, etc.).

With this in mind, shark attacks are still events that occur with risk that can be modeled. And where risk can be modeled, individuals can assess and make judgements for themselves on how to navigate through the world. This paper will explore various factors associated with shark attack occurrences, and will attempt to quantify the risk involved so that individuals can make judgements for themselves on how to interpret that risk.

The main sources of data for these events is The International Shark Attack File (ISAF). The ISAF is a catalog of all these events and is a widely trusted database on shark attack occurrences globally, curated by researchers at the University of Florida (special acknowledgement and thanks to Dr. Gavin Naylor and Joseph Miguez for this data). To help decrease model complexity and to provide a manageable project scope, only a subset of this data was pulled: specifically, unprovoked attack occurrences on the Atlantic Coast of Florida. The possibility of heterogeneity in pulling a specific dataset such as this, rather than global data is a limitation that will be discussed further in the conclusion of this paper.

Problem Statement

The main problem that is being solved by this project is: Given temporal and weather data, can a shark attack occurrence be reliably predicted? This is not to say that the aim of this project is to model ***individual*** risk of attack, rather the ***general*** occurrence of an attack given the features at hand. This is an important distinction that is necessary to understand when viewing the data and results.

Hypothesis

The two general questions tested in this project are:

1. Can temporal features (i.e. time of day, month of year, and lunar phases) help provide indication of a shark attack occurrence?

2. Can weather features (i.e. barometric pressure, temperature, and the change of these features over time) provide indication of shark attack occurrence?

Regarding temporal features, it is widely known shark attacks occur more often on a seasonal basis. These events happen more often in warmer months due to more beachgoers and shark migratory patterns. Additionally, it is known that more attacks occur during day hours due to the fact that more people visit the beach during the day rather than night. Furthermore, it is also thought that the lunar phases have an effect on shark attack occurrence, due to general lunar illumination. These features and their effect on shark attack occurrence will be tested via the models later mentioned.

Regarding weather features, the interaction that is of interest for this project is barometric pressure, temperature, and the delta and rate of change between these variables over time. As all aquatic life are sensitive to dramatic shifts in their environment, these models will test whether or not these shifts influence shark behavior in a way that has an effect on shark attack occurrence. The causal nature of this affect is not being measured (i.e. whether sharks become more aggressive or whether they flee an area when there are dramatic shifts in these features), rather the overall correlation between these whether events and shark attack occurrences themselves.

Data Cleansing

One of the main challenges in this project was data sanitization. The ISAF is a near-century old project, with data going back from the 1500s- present. Data drift is an inevitable threat with a project as long-standing as this, and sanitization is a normal part of the modeling process. Why and how values are inputted have changed over time, and the data reflects this. Multiple cleansing and parsing functions were used to create a sanitized output.

The first of these were functions created to parse through the 'time-of-attack' feature, which is one of the most critical to this analysis. As the data varied between written phrases such as "Forenoon" and "Afternoon", the usage of AM/PM, military time, and general ranges such as "2:00- 4:00", a variety of functions were created to parse this information. Additionally, functions were created to "bucket-ize" this time of attack feature and create new features based off of this. This includes the creation of a general time-of-day feature that shows whether the attack occurred in morning, afternoon, evening, or night, which was useful for visualizations.

Additional features were also created from the 'date_at_location' feature to help create more modeling value. The first of these was the creation of general months, and seasons, based on the date. The second was the creation of features related to the lunar phases. Three functions were created to calculate the general lunar phase (new, first quarter, full, last quarter), the general lunar illumination fraction (expressed as a fraction for modeling purposes), and general lunar illumination percentage (expressed as a percentage for visualization purposes).

Another thing to note is that unnecessary columns were dropped (such as body parts involved in the attack) to reduce noise while modeling and to curtail the dataframe. Additionally, all provoked attacks were dropped from the data, as the scope of this project pertains to solely unprovoked instances. As provoked instances vary widely, the risk is that they would have skewed the results of the analysis.

One of the byproducts of this project is a cleaned output of data, as well as a data cleansing pipeline that can both be used for further projects such as this.

A Note on Weather Data

Finally, weather data was selected only from three counties for ease of analysis. Temporal data, still remained the full dataset, but due to complications with metostat libraries, weather information had to be limited. This can be expanded in future model iterations, but for this specific project, these three counties were analyzed: Brevard, Duval, and Volusia.

EDA and Basic Visualizations

Exploratory data analysis and the creation of visuals is a key step for a project such as this, because it allows one the ability to see what key features they are going to use in modelling and how these features might affect the model. Many visualizations were created to make sense of the data. The visualizations shared below are the most relevant to the modelling that was subsequently done.

Temporal Visualizations

The below visualizations related to count of shark attack occurrences across different measures of time are quite simple in and of themselves, but they do show us a few different things to look out for in the subsequent modeling. The first is that, unsurprisingly, shark attacks for this dataset tend to skew towards daytime hours, specifically in the mid-day to afternoon. The second is that shark attacks for this dataset skew toward late summer, early fall months. The third is that shark attacks for this dataset may have a slight uptick around new moons.

Finally, based on the correlation heat-map, there seems to be a correlation between attack binary (shark attack or not) and month, as well as hour. It should be noted that this is a point-biserial correlation (correlation between linear and binary values), so the correlation may not look as high as it actually is. The models below will prove out whether these are strong indicators or not

All this being said, these visualizations aren’t enough to determine whether these imbalances are simply indicative of the dataset itself, or can be used to infer future occurrences. In other words, are these results statistically significant?

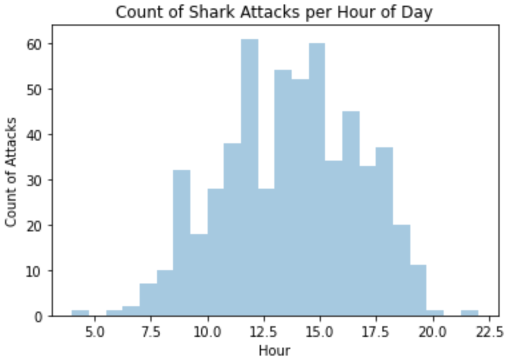


Figure 1: Count of Shark attacks per hour of day

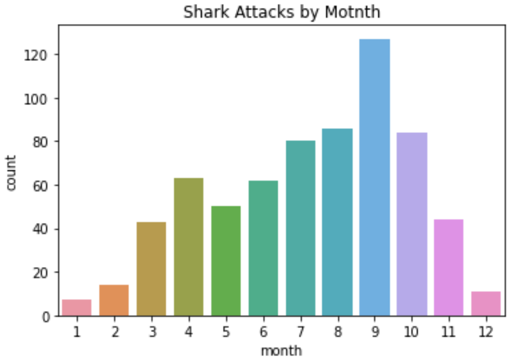


Figure 2: Count of Shark attacks per Month

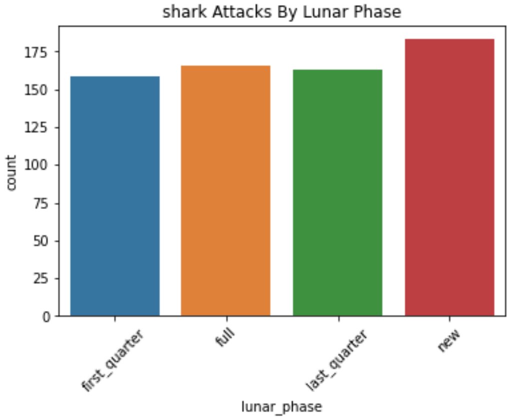


Figure 3: Count of Shark attacks per Lunar Phase

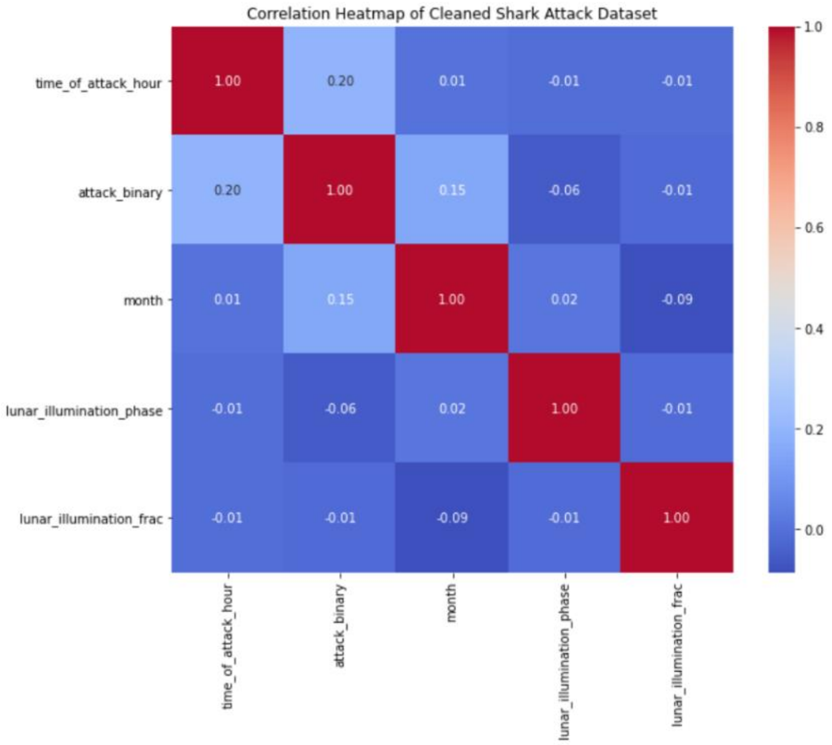


Figure 3.5: Correlation Heatmap of Temporal Features

Weather Visualizations

The below visualizations are some significant features that were found to be of note when inspecting the weather data. Data was pulled in from meteostat, which is a python library that contains weather data pulled from NOAA weather stations. With this, a new data byproduct was created, enriching the original shark attack data with meteorological features. A further note regarding this byproduct is that non-attack occurrences were created for comparison to the attack occurrences, but this will be discussed later in this paper.

These are basic box-and-whisker plots that show the distribution of attacks across three different features: temperature, change in pressure, and rate of change (or slope) of pressure. Other features were visualized, but these were the most pertinent to modeling. As one can see by the below visualizations, there is a slight skew toward higher temperature on the attack values vs non-attack values. Additionally, the change in pressure and slope of pressure change give an even slighter skew higher. Moving to the modeling phase, these observations will be tested for their statistical significance.

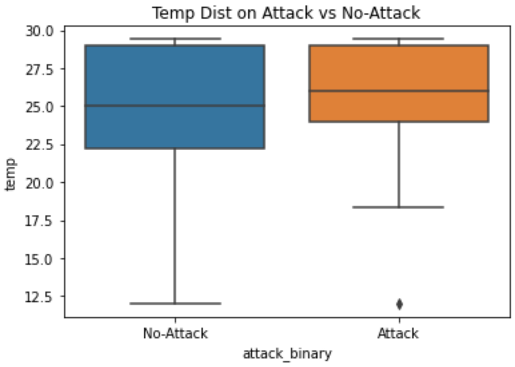


Figure 4: Temperature of Attack vs Non-Attack Occurrences

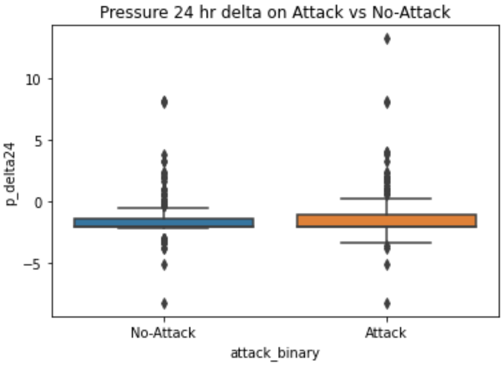


Figure 5: 24 Hour Pressure Change of Attack vs Non-Attack Occurrences

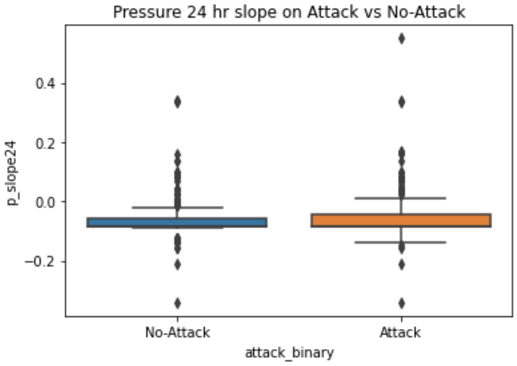


Figure 6: 24 Hour Pressure Slope of Attack vs Non-Attack Occurrences

Modeling Approach

Logistic Regression

When approaching how to model this problem, the nature of the what is being modeled was assessed. In this case, the outcome that was to be predicted was binary, whether a shark attack occurred or not based on the predictor at hand. In this case, logistic regression is a clear choice, as the predicted class in this method of supervised learning is a binary case.

Accounting for Non-Shark Attack Occurrences (Negative Events)

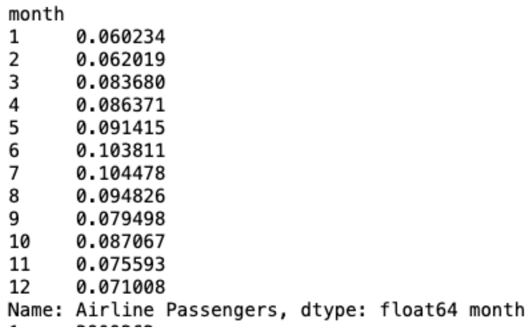
After selecting logistic regression as the method of supervised learning to solve this problem, one main issue arose, which is the absence of negative cases. The data involved in this project is confirmed shark attack occurrences, and we do not have confirmed non-occurrences. This seems a silly distinction to make, as this is an obvious thought. There would be no point in recording non-shark attack occurrences, because they happen every day all around the world. The reason it is important is that the model will have to somehow account for these negative events.

To do so, the solution implemented in this particular project was the creation of synthetic negatives. Synthetic negatives allow one to account for all the non-shark attack occurrences. A reasonable question to ask would be: Wouldn't the synthetic negatives just represent the entire date-length of the dataset where attacks didn't occur (e.g. in this case, that would be every non-shark attack month, day, hour since 1931)? Unfortunately, one cannot feed a model this number of negative events, as shark attacks are such rare occurrences, the amount of non-attack occurrences are so large it greatly outweighs the attack occurrences. In this case, an equal length of data is sufficient to remove the model bias that would be had if one were to overwhelm it with negative events.

Creating Synthetic Negatives- Weighting Beachgoer Seasonality

To create synthetic negatives the same length as the existing dataset, the simplest approach is to use a random choice function pick non-attack occurrences that are not within the existing dataset. While this is easy to implement, there are some drawbacks to this. For this particular problem, the drawback is that shark attacks only occur when people are present. Furthermore, there are more or less people present at beaches in a) certain times of the year (seasonality) and b) in certain times of the day. If one was to randomly grab days using a random choice function, the randomization would not account for this seasonality and time of day where beachgoers are present.

So, how does one go about normalizing the random choice function to account for beachgoer seasonality? In this instance, data was taken from a publicly accessible dataset created by the University of West Florida in the form of monthly flight inflow data into Florida Coastal airports. Using this data, monthly weights were applied to the randomization to account for the seasonality of beachgoers.



```
month
1     0.060234
2     0.062019
3     0.083680
4     0.086371
5     0.091415
6     0.103811
7     0.104478
8     0.094826
9     0.079498
10    0.087067
11    0.075593
12    0.071008
Name: Airline Passengers, dtype: float64 month
```

Figure 7: Weighted Monthly Seasonality for Beachgoer Data

It should be noted, that this weighting is not a perfect representation of when individuals go to the beach, but it does provide a directionally correct indication of beachgoer seasonality. Additionally, although this data is captured on a monthly basis, this particular project did not account for time-of-day (hourly) beach going patterns. This is a note that was created to be implemented in future iterations of this modeling to increase efficacy. Finally, in the case of temporal models, seasonality was not able to be implemented as the synthetic negatives were randomized on an hourly basis to match the dataset. The weather data was specifically where seasonally-adjusted synthetic negatives were implemented. The limitation of the temporal models' synthetic negatives is noted and will fixed for future iterations.

Modeling Results

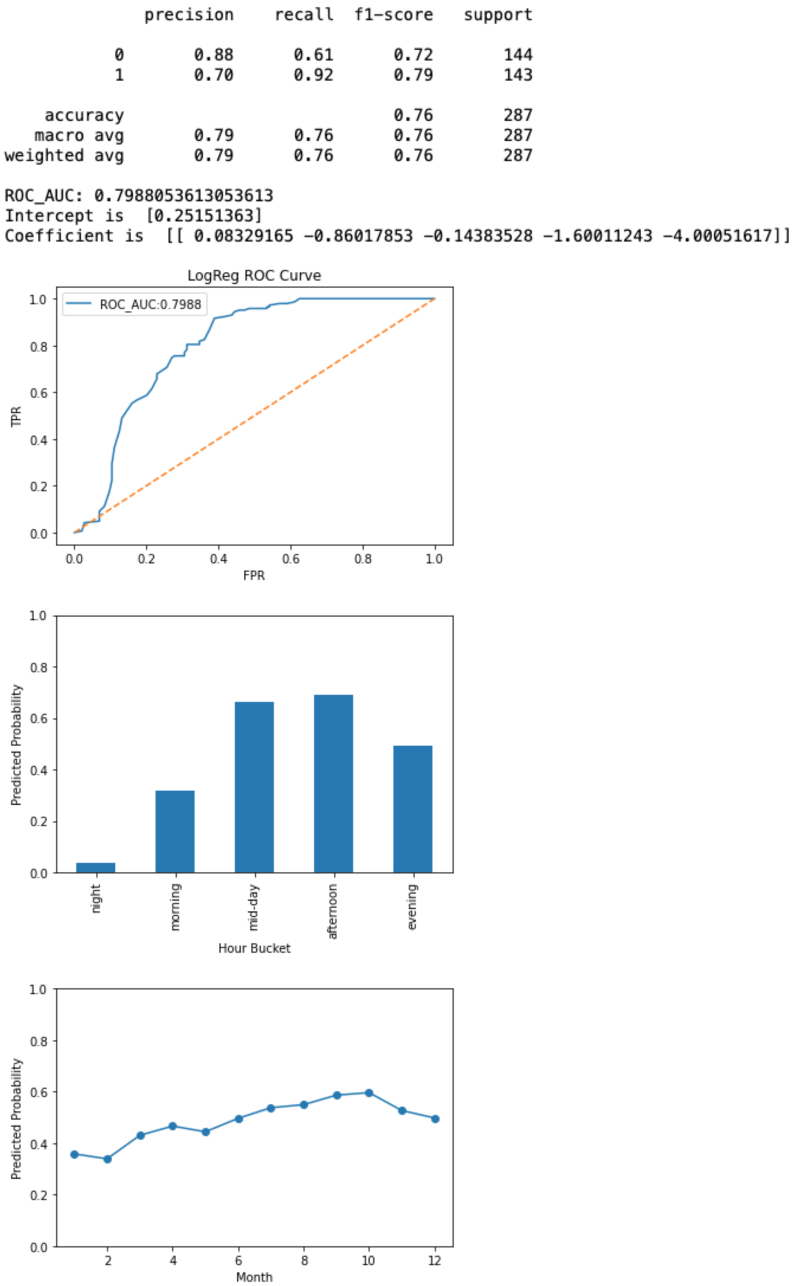
In this project, 10 separate Logistic regression models were created. They are as follows:

1. Shark Attack Occurrence ~ Lunar Illumination,
2. Shark Attack Occurrence ~ Month of Attack,
3. Shark Attack Occurrence ~ Hour of Attack,
4. Shark Attack Occurrence ~ Hourly Bucket (Time of Day),
5. Shark Attack Occurrence ~ Lunar Phase,
6. Shark Attack Occurrence ~ Month of Attack + Hourly Bucket (Time of Day),
7. Shark Attack Occurrence ~ Temperature,
8. Shark Attack Occurrence ~ 24 Hour Pressure Change,
9. Shark Attack Occurrence ~ 24 Hour Pressure Rate of Change,
10. Shark Attack Occurrence ~ Temp + 24 Hour Pressure Change + 24 Hour Pressure Rate of Change

For the sake of brevity, only two of the most significant results will be listed in this section from both the Temporal and Weather category, but all results will be in the companion Jupyter Notebook.

For the temporal data "**Shark Attack Occurrence ~ Month of Attack + Hourly Bucket (Time of Day**" yielded the most significant results. As you can see by the results below, combining these two features yielded a combined ROC AUC score of ~80%. This is a significant outcome, and means this combo model has a strong chance of predicting shark attack occurrence vs non-occurrence.

In the predicted probability charts below for the monthly and hourly data in this model, one can see that the distribution of attack probability is similar to the real-world distribution of shark attacks that was seen in the EDA visualizations earlier. This is a robust result, because it shows the model is predicting in a similar fashion to real world outcomes.



```
              precision    recall  f1-score   support

           0       0.88      0.61      0.72       144
           1       0.70      0.92      0.79       143

    accuracy                           0.76       287
   macro avg       0.79      0.76      0.76       287
weighted avg       0.79      0.76      0.76       287

ROC_AUC: 0.7988053613053613
Intercept is  [0.25151363]
Coefficient is  [[ 0.08329165 -0.86017853 -0.14383528 -1.60011243 -4.00051617]]
```

Figure 8: Logistic Regression Results for “Shark Attack Occurrence ~ Month of Attack + Hour of Attack”

Additionally, providing the hourly and monthly distribution of attack vs non-attack below to visually see how the model parsed through positive and negative events in the test and train data:

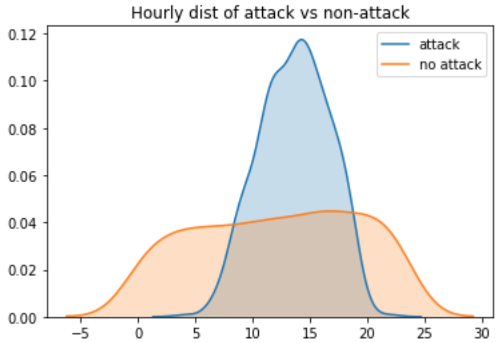


Figure 9: Modeled Hourly Distribution of attack vs non-attack occurrences

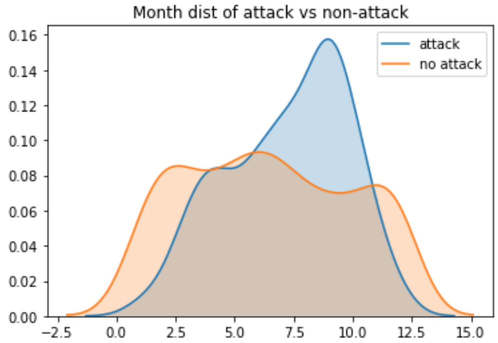
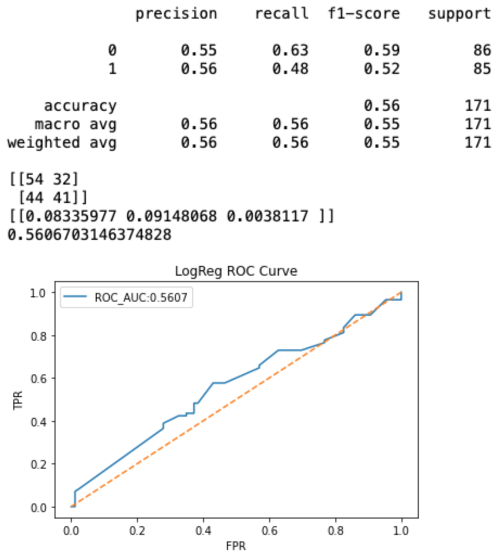


Figure 10: Modeled Monthly Distribution of attack vs non-attack occurrences

For the Weather data, "**Shark Attack Occurrence ~ Temp + 24 Hour Pressure Change + 24 Hour Pressure Rate of Change"** yielded the most significant result. Unfortunately, this model was little better than a coin flip in predicting shark attack occurrences from these three variables stacked, with an ROC AUC score barely over 50%



```
              precision    recall  f1-score   support

           0       0.55      0.63      0.59        86
           1       0.56      0.48      0.52        85

    accuracy                           0.56       171
   macro avg       0.56      0.56      0.55       171
weighted avg       0.56      0.56      0.55       171

[[54 32]
 [44 41]]
[[0.08335977 0.09148068 0.0038117 ]]
0.5606703146374828
```

Figure 11: Logistic regression results for "Shark Attack Occurrence ~ Temp + 24 Hour Pressure Change + 24 Hour Pressure Rate of Change"

Model	Encoding	ROC-AUC	Notes
Shark Attack Occurrence ~ Lunar Illumination	Continuous	~0.50	Weak; nearly random performance.
Shark Attack Occurrence ~ Month of Attack	Cyclical (sin/cos)	~0.69	Not a strong predictor; small seasonal effect visible.
Shark Attack Occurrence ~ Hour of Attack	Cyclical (sin/cos)	~0.76	Moderate strength; improves over month alone.
Shark Attack Occurrence ~ Hourly Bucket (Time of Day)	Categorical	~0.78	Strongest single predictor; captures diurnal patterns well.
Shark Attack Occurrence ~ Lunar Phase	Categorical (New, Full, etc.)	~0.45	Weak; nearly random performance.
Shark Attack Occurrence ~ Month of Attack + Hourly Bucket (Time of Day)	One-hot month + Categorical	~0.80	Best performing combined model; strong diurnal + seasonal signal.
Shark Attack Occurrence ~ Temperature	Continuous (°C)	~0.53	Weak; nearly random performance.
Shark Attack Occurrence ~ 24 Hour Pressure Change	Continuous (Δ pressure, hPa)	~0.54	Weak; nearly random performance.
Shark Attack Occurrence ~ 24 Hour Pressure Rate of Change	Continuous (slope, hPa/hr)	~0.54	Weak; nearly random performance.
Shark Attack Occurrence ~ Temp + 24 Hour Pressure Change + 24 Hour Pressure Rate of Change	Continuous (3 features)	~0.56	Adding multiple meteorological features doesn't improve model; near random.

Discussion and Conclusion

Discussion of Results

Looking at the results of the modeling done on this dataset, it is easy for one to tell that the temporal indicators provided the results of the highest value. Looking at the overall ROC AUC score of ~80% on the Hourly Bucket and Month day combination, this model is provides high efficacy and will accurately predict the occurrence of an attack 76% of the time when given date-time info. In terms of which model should be iterated upon "Shark Attack Occurrence ~ Month of Attack + Hourly Bucket (Time of Day)" is the most promising.

Other temporal data features looked promising as well, but may need further refinement. Lunar phases and lunar illumination, which were previously hypothesized to be a potential indicator, but yielded poor results when modelled. This information in itself is useful though in that it tells us that Lunar Phases may have no correlation whatsoever with shark attack occurrences. In this respect, Lunar temporal data requires further investigation.

In regards to the meteorological features- Those that were chosen for modeling (Temperature, 24 Hour Pressure Delta, and 24 Hour Pressure Rate of Change) were extremely poor indicators of shark attack occurrence. There may be further credence to the use of these features as indicators, if modeling is enhanced, but for the time being they lack a significance in the current iteration.

Limitations of Current Model

One notable limitation for the current model is that synthetic negative events for temporal data were completely randomly generated and in future iterations, implementing seasonality and time-of-day into the random choice function would increase model efficacy Additionally time-of-day generation in both random choice functions for temporal and weather data, if beachgoer hourly data is captured would increase model efficacy

Furthermore, lack of expertise in ichthyology and shark research could have caused disregard for potential in other features that could increase the model efficacy. In future iterations, working with the ISAF researchers and gaining first-hand knowledge in shark behavior could help direct additional features to include.

Additionally, as previously noted, the weather analysis was done on only three counties within the dataset, and future iterations of this model should include the entire dataset if possible.

Finally, bias in the geographical nature of data, being that this subset only included Florida Atlantic coast attacks could have very well skewed the results of certain models. Running a global or continent based model could drown out these regional biases. Additionally, splitting the data along these regional lines could yield interesting and different results.

Future Applications and Model Iterations

One potential application of a model such as this, when further refined to include additional features to increase overall efficacy, would be the use of a risk indicator of sorts for shark attack occurrence given a specific set of circumstances. Think of a "Smokey the Bear: Wildfire Risk"-type indicator that could provide the generalized risk of a shark attack occurrence. This model could be further refined to add features such as geographical location, tidal and current affects, storm systems in nearby area, species indicators, and so on. With further expert input on the model features to be included, one could more greatly refine to provide individuals with valuable risk information when making decisions about their beach going activities.

References

Florida Museum. (n.d.). Beach injuries and fatalities: Comparing risks. Florida Museum of Natural History. Retrieved September 25, 2025, from https://www.floridamuseum.ufl.edu/shark-attacks/odds/compare-risk/beach-injuries-fatalities/

Florida Museum. (n.d.). International Shark Attack File. Florida Museum of Natural History. Retrieved September 25, 2025, from https://www.floridamuseum.ufl.edu/shark-attacks/

University of West Florida, Haas Center. (n.d.). Tourism indicators. University of West Florida. Retrieved September 25, 2025, from https://uwf.edu/centers/haas-center/explore-the-economy/tourism-indicators/

French, L. A., Midway, S. R., Evans, D. H., & Burgess, G. H. (2021). Shark side of the moon: Are shark attacks related to lunar phase? Frontiers in Marine Science, 8, 745221.https://doi.org/10.3389/fmars.2021.745221

OpenAI. (2025). ChatGPT (version GPT-5) [Large language model]. OpenAI. (Used for error trace debugging and function usage examples.)

Google DeepMind. (2025). Gemini [Large language model]. Google DeepMind. (Used for error trace debugging and function usage examples.)

Special acknowledgement: Dr. Gavin Naylor and Joseph Miguez, Florida Program for Shark Research, for providing ISAF data.