Initial Analysis of Multivariate Factors for Prediction of Shark Presence and Attacks on the Coast of North Carolina

Sonal Kaulkar[1] and Lavanya Vinodh[1] and Pamela Thompson[1,2][0000-0002-1955-4121]

[1] University of North Carolina at Charlotte, Department of Computer Science, Charlotte, NC, USA
[2] Catawba College, Salisbury, NC, USA
plthomps@uncc.edu

Abstract. Classification, association rules and clustering are used in the study to improve understanding of the presence of sharks in near shore waters during tourist seasons in middle Atlantic coastal waters, specifically North Carolina. The Global Shark Attack File combined with data on environmental, biotic and meteorological factors is prepared for analysis using the CRISP-DM process. In future work, combined inputs including a standardized hashtag for twitter mining, real time weather and water information, and data on crab and turtle presence will provide real-time input to an app or a dashboard providing early warning of shark presence.

Keywords: Shark Attack, Clustering, Early Warning Shark Attack App, Knowledge Discovery Process, Association Rules, Multivariate Analysis, Classification, Balancing Data

1 Introduction

Shark attack incidents are one of the most well-known animal related phenomenon with wide spread documentation on the news and social media. Exploratory data analysis of the Global Shark Attack File [1] shows the East Coast of the US experienced an unusual number of attacks during the Summer of 2015 with sixteen total attacks during the season in North and South Carolina. This study was initiated as a research project for the Knowledge Discovery in Databases graduate class at the University of North Carolina at Charlotte during July of 2015 and has continued. The increase in shark attack incidents throughout the summer and the lack of effective early warning systems for East Coast tourists create interest in applying knowledge discovery techniques to the prediction of sharks in near coast waters along East Coast beaches. In the US, North Carolina and South Carolina together rank second after Florida with respect to the incidence of shark attacks [1] [17] [20].

1.1 Objective

In this study, analysis is performed to find the potential impact of certain meteorological, environmental and anthropogenic factors using data mining following the Cross-industry standard process for data mining (CRISP-DM). These factors could be a potential cause for the rise of shark attacks on the coast of North Carolina [4]. The objective of this research is to improve our understanding of the presence of sharks during tourist seasons in middle Atlantic and south eastern coastal waters, specifically North Carolina for this study. Our study will focus on the analysis of existing data from the Global Shark Attack File, weather, wind and water data from NOAA, calculated moon phase dates, fish, crab and turtle populations. The quantitative analysis on this data will lead to new and interesting knowledge that will provide the basis for an app providing advanced information on the likelihood of sharks in coastal waters where tourists swim, surf and wade. A future focus of this research is to analyze social media activity relative to shark presence. A recommendation for a standardized way to tweet will be considered with interested and strategic partners in order to ultimately provide an additional feed to a shark sighting app such as Dorsal, co-founded by professional surfer Sarah Beardmore [21].

1.2 Methods

The principal data source for the incident of shark attack in North Carolina was taken from the incident log of the Shark Research Institute [1] defined as the "Global Shark Attack File" (GSAF). The project constrained the study of incidences of shark attacks to be limited to the coast of North Carolina from 2009 to 2016 for the months of May to September.

The methodology adopted for the project was the usage of CRISP-DM model [16]. The process started with business understanding and data understanding phase of understanding the shark attacks which was enhanced further by the interaction with Dr. Chuck Bangle Shark Researcher of East Carolina University. The Data Preparation necessitated the collection of data sources for all the potential factors that led to higher rate of shark attacks. The next phase employed was modelling of the collected data to gain insights in the data and find the relevant factors that showed an impact on the incidences of shark attacks. The input file used for modeling was aggregated on a daily basis for all the data sources collected for data mining.

2 Data Understanding

Data Understanding constituted of finding the important factors and the study of their contribution to the rise of shark attacks. The GSAF incident log was used as a starting point to understand the incidences of shark attack. Exploratory Data Analysis was performed using a variety of tools including Excel, Weka and R.

For EDA the GSAF file was filtered for the years of 2009 to 2016 for the months of May to September for both the beaches of North and South Carolina states. Location was a preliminary attribute chosen based on the fact that place of incident can be a starting point for judging the deciding factors [4]. It was found that of all the beaches Myrtle Beach had the highest number of shark attacks. The input file used for data mining for all the consequent phases also included the area which is the state, species, beach and county. Bull sharks are one of the shark species most involved in fatal attacks against humans [5][3]. Plotting a bar chart of Species of sharks involved in the incident revealed that bull sharks are the larger in number in these incidences.

Anthropologic factors like very large involvement of human settlement can be a reason to attract sharks [4]. A less populated shoreline with limited access to humans has less number of shark attacks. A histogram of activities of humans against the count of shark attacks was plotted. The histogram reflected that activities such as surfing and swimming on the shore proved more fatal to humans with respect to shark attacks.

Lunar cycles have always been controversial to be considered as a factor to the rise of shark attacks [6]. It is said that lunar cycles are unlikely to have an effect on incidence of shark attack [15] however it is common knowledge to fisherman that lunar cycles affect the catch [7]. To explore the relevance of lunar phases to shark presence in near shore waters an algorithm in Java was used to calculate the moon phases for all the days of May to September from 2009 to 2016 for inclusion in the study.

The set of moon phases belonged to one of the categories new moon, full moon, first quarter, waxing gibbous, waning gibbous, third quarter, waning crescent, or waxing crescent. The calculated moon phase by date was then integrated with the GSAF file using R code for the purpose of performing exploratory data analysis on the effect of lunar cycles on shark attacks. Histogram analysis shows attacks are almost equal in both the cases new moon and full moon. Domain knowledge on moon phases and the effect on tides shows that when the moon and sun align, known as full or new moon, the pull is at its strongest, causing the tides to be at their highest and lowest. This is known as spring tides. The change from high tides to low tides and back again happens very quickly which may cause sharks to move in areas closer to where people swim when the tide is low. The effect on tides occurs over a span of days leading up to the change in the lunar cycle. The effect duration, therefore, is not on the day that is typically depicted as the new or full moon but can span several days before and after [6][7].

Meteorological factors impact the occurrences of shark attacks along the East coast; warm temperatures increase the presence of food (crabs, fish that migrate to warmer waters and this increases the likelihood of attack as sharks follow the food source. The encounters with humans also become more prevalent with summer months from May to September when there is a rise in water temperature that bring in humans as well as sharks near the shoreline [2][9]. The data source for the all the

weather parameters is taken from National Oceanic and Atmospheric Administration [8] [12]. Exploratory data analysis revealed that high temperature favors shark attacks. As the temperature of the water increases, the sharks tend to move into favorable locations [21].

Increases in temperature have resulted into scarce rainfall, thereby increasing the salinity of water which may cause more movement of sharks to the shoreline away from the estuaries. Hence precipitation in atmosphere along the shoreline is taken as an attribute in the prediction of shark attacks. Wind speed is another crucial factor which may attracting sharks to the shore along with their sources of food [13]. The higher speed of the winds moves clearer, warmer water to the shore. When that happens the food sources such as mullet, bait fish and menhaden move closer to the shore and the sharks follow their food sources [10] [4]. Wind speed data from NOAA [12] was included in the study for analysis.

Salinity of water affects the habitat of sharks. Sharks usually are found near estuaries except for the bull shark which may thrive in open sea water and is responsible for many of the incidents of attacks in the study data. Salinity data was taken from National Estuarine Research Reserve System [11] which is monitoring program SWMP or system-wide monitoring program, and provides long-term data on water quality, weather, biological communities, habitat, and land-use and land-cover characteristics [11]. Exploratory data analysis was performed on the salinity and the number of attacks that occur in the study data are highest when the salinity reached the optimal/desired levels.

An increase in rainfall causes rivers to swell and thereby causes turbidity to increase, reducing the visibility and making it harder for people and sharks to see each other. This can result in attacks [13]. Thus, the more turbid the water, the greater may be the probability of occurrence of attack incidents. Low levels of dissolved oxygen can cause marine life to become very lethargic. Aquatic animals involving sharks move towards the shore to get more oxygen. This effect is called as “Jubilee” effect by the local communities who are involved in fishing [12]. Dissolved Oxygen along with Salinity and Turbidity is collected from NERRS [11] every 30 minutes from each of the 26 NERRS [11] sites out of which the data collected from the East Cribbing site is taken since it is closest with the Wrightsville Beach, NC location which represents a centralized location for the North Carolina shoreline and beaches. Ecology is another factor that drives sharks towards the shore. Sharks follow close behind whales, turtles, menhaden who surged northwards towards the heat wave in 2015 [13] [14]. Turtle data was collected in terms of false crawling and nesting from Dr. Matthew Godfrey (State Coordinator) from North Carolina Wildlife Resources Commission. False crawls are when turtles come to shore to lay eggs but do not succeed and go back to the ocean to try again another day. Nesting is when the turtles have succeeded and a nest is documented.

Blue crabs are prevalent along the east coast of the United States and are particularly prevalent in North Carolina. These crustaceans can survive in different environments and migrate towards the Carolinas in summer for higher spawning rates. This movement in-

creases the blue crabs in number along the east coast [15]. Blue crab movement is also affected by the tides with crabs moving from inlets to estuaries in increasing distance during full and new moon periods. Sharks along the east coast prey on the blue crabs that are prevalent in the region. Daily Blue crab landings from 1994 to 2004 was collected from Alan Bain, the chair of the NC Division of Marine Fisheries. This data is integrated with the prediction of shark attacks and shows high crab landings favor attack.

South West winds blow in a north easterly direction but do not move water in that direction. Due to the earth's rotation water takes a 90-degree toll and the water temperature drops down by 10 degrees making the lower cold water rise to the top making water murkier during higher winds. This may be a reason for sharks to get confused with its intended prey resulting in shark attacks on humans, not fish in murky waters. ERRS was used as a data source for wind direction which contained readings in degrees which were then converted in one of the directions using Java code. On performing initial exploratory data analysis, it was found that attacks are almost equal in both the cases south-southwest and southwest winds.

3. Data Preparation

Data merging and preparation followed exploratory data analysis and domain knowledge discovery. The data preparation presented a significant task with data sources from several different sources requiring merging with the global shark attack file primarily by date. The global shark attack file shows only dates where attacks occur. For learning purposes, the study including dates where attacks did not occur in order to gain knowledge on the presence and absence of sharks in near shore waters based on documented attacks. For supervised learning, a binary attack flag representing Attack Y or N and other variables as predictor variables were merged with the global shark attack file by date. Another file was maintained for exploratory data analysis which consisted only of instances with the attack flag equal to Yes.

The Global Shark Attack File containing the incident log of attacks and enhanced with dates not having attacks provided the basis for the principal file into which other predictor variables were then added as columns. Turtle data constituted turtle activities like nesting and false crawls. The turtle data that was obtained was observatory data. For unification of format, the nesting and the false crawls were merged together to form a single discrete numerical attribute named Turtle Activity. This modified feature was then aggregated date wise to get a daily count of turtle activity and merged into the GSAF file using R code. Moon phase and lunar cycles have their effect on a moving average of three days before and after a particular moon phase [19]. Thus, the daily data of moon phases including values for full moon and new moon was extended to three days before and after the calculated date. This feature was used as one of the predictor variables in modeling. A moving average method using R was applied to the precipitation data from NOAA [12] to account for the

nature of effects rainfall has on water and other factors. Wind direction data from NERRS [11] was given in the form of degrees of rotation every 30 minutes. This data was then aggregated with the mean calculation to calculate the average rotation of the wind on a daily basis. On careful understanding of the impact of changing temperature on the surge of sharks towards the shore, change in water temperature between consecutive days was used as a predictor rather than raw temperature values. Mean calculation of water temperature on a daily basis is aggregated and then a difference between the current and previous day's temperature is used as a predictor variable for temperature.

Binning of numerical variables of the data collected was employed for most of the predictor variables [16]. Before using turtle activity data in analysis, the numerical predictor was discretized based on location and daily count. On plotting histograms, it was assured that the turtle activity discretized on a daily basis correlated more with shark attacks and was thus used in the prediction in the modeling phase. Similarly, time of attack from GSAF, salinity, turbidity, station pressure, and wind speed numerical variables were discretized based on three bin equal width binning for the initial analysis.

Missing values were evident in the merged data. Weather data collected from NOAA [12] had 17% of Not Applicable or Not Available (NA). To solve this missing data issue, NAs were removed by replacing it with the mean of the value aggregated over the day from the raw data. This method was applied for Dissolved Oxygen, Salinity, Turbidity, Water Temperature, Precipitation, Station pressure and Wind Speed. Another hurdle with respect with missing data was with respect to Crab landings data from 1994 to 2015. Out of the available data from 1994 to 2015, data was filtered from 2009 to 2015. There were no readings for crab landings for all Sundays. To replace these missing values R code was written to take average of the crab landings of Saturday and Monday as the value for Sunday. Data for 2016 year was created using estimation and regression analysis using Excel. Linear Regression was used with 95% confidence to predict the Crab landings.

Min-max Normalization was chosen standardize all numeric variables. The final input file was improved by removing duplicate records using R duplicated function.

3.1 Data Imbalance

The GSAF file was filtered for the state of North Carolina for the duration of May to September between the years 2009 to 2016. Thus, the number of records with Attack equal to Yes were 65 in total. When the GSAF file was combined with the file containing all dates it led to a data imbalance problem. In order to solve the problem, a stratified sampling of Attack=No subset was prepared so that 1/3 of records remain with adequate representation for each class. To evaluate the efficiency of

the solution applied to the data imbalance problem, a two sample Z test for difference in proportions was used to verify the partitions.

4. Modeling

Classification, clustering and association rule analysis were performed for the prediction of shark attacks along the coast of North Carolina.

4.1 Classification

The method used for classification was Naïve Bayes Classifier using J48 with Weka Tool. The major task for classification was to find the potential factors of the included predictors that lead to shark attack. The algorithm used for Attribute selection was Weka's Info Gain Attribute Evaluator. Of the 12 variables given as input to the Attribute evaluator, the rankings provided by the attribute evaluator ranked Wind Speed as the highest and the Station Pressure as the lowest. The classification model generated an unpruned J48 tree giving 149 correctly classified instances out of total of 186 instances. The mean absolute error was .21 with a weighted average true positive rate of .80, a false positive rate of .24, precision and recall were .80 each. The total confidence of the model was 80.1075% weighted average.

4.2 Clustering

The K-means algorithm was used for clustering in Weka Tool [16]. The clustering resulted into two distinct clusters with Cluster 0 with target variable as Attack = Yes and a second Cluster 1 with target variable as Attack = No. The nine highly ranked predictor variables used in the classification algorithm were used for clustering. The Mean Absolute error was 0.2172, Root mean squared error: 0.4178, Relative Absolute Error: 45.7399%, Root relative squared error: 85.7572% on Total Number of Instances: 186. It was found that Cluster 0 had a lower turtle activity proportion than Cluster 1. Also, the Wind Speed for Cluster 0 was lower in proportion than Cluster 1. The direction of winds for Cluster 1 is South West Winds (SWW). The salinity is High for Cluster 1 while for Cluster 0 it is medium the lunar cycle also showed a significant change. In Cluster 0 the moon phase is First Quarter while for records of the shark attack category in Cluster 1 it is a full moon. The number of instances belonging to Cluster 0 were 53% of the data and that belonging to Cluster 1 was 47% of the data (Figure 1).

Fig. 1. Final Cluster Centroids

Final Cluster Centroids

	Cluster#		
Attribute	Full Data	0	1
	(186)	(99)	(87)
Attack	No	No	Yes
DO2	Med	Med	Med
Salinity	High	Med	High
Temperature	High	High	High
Precipitation	Low	Low	Low
Direction	SSW	SSW	SW
Moon Phase	Full	New	Full
Wind Speed	0.34	0.27	0.43
Turtle	0.15	0.10	0.21
Time taken to build model: 0.01 seconds			
Model and evaluation on training set			
Clustered Instances			
0	99(53%)		
1	87(47%)		

4.3 Association Rules

Association rule modeling was employed using the apriori algorithm in order to find the best rules for consequent Attack = Yes. The variables used in Apriori were Attack flag, Dissolved Oxygen, Wind Speed, Crab Landings, Wind Direction, Extended Moon Phase, and Temperature Change as the variables in the consequent part.

The result of the association model gave minimum support of 0.1 and a minimum confidence of 0.75. The best rule with the highest confidence is: {Crab Landings = High, Wind Direction = SW, Moon Phase = New Moon} => {Attack = Yes} with confidence of 0.78.

5. Discussion

The data preparation and preliminary analysis performed in this study have provided factors that influence the presence of sharks based on shark attack data from the Global Shark Attack File. These factors such as weather conditions and environmental factors based on location can be input to an application which would then provide people with an assurance level of how safe it is to surf, swim or wade on beach [19]. This data can be captured real time and provided to a supervised neural network which would present its output node as the level of safety for a particular beach. A warning system like this can be combined with real time surveillance with newer methods involving drone and blimp technologies with image detection capabilities. Another way

this data mining project can be deployed is to give the relevant attributes as input to existing apps in the market that monitor activities related to shark attack such as the Dorsal app co-founded by professional surfer Sarah Beardmore. Apps like Dorsal can be used for monitoring purposes and to guide people to be safe before going into the ocean [19] [21].

In addition to the predictors of interest from this study, a standardized hash tag can be used to gain new understanding on the presence of sharks and the interaction of features from the weather and environment. In developing a uniform hash tag to report attacks or events that predict sharks in near shore waters, procedures such as those outlined in the United Nations Office for the Coordination of Humanitarian Affairs [18] are recommended.

6. Conclusion

The data analysis for prediction of shark attacks on the coast of North Carolina provided useful insights about the factors that lead to shark attacks and shark presence. Different modeling techniques gave the relevance of each of the different potential factors and how they could contribute to the prediction of shark attacks and presence. To better understand the potential factors, attribute relevance using information gain was run on the training data. Wind speed and direction though rarely mentioned in articles and social media did prove to be more relevant in our results. Also, ecological variants like changes in crab landings and lunar cycles also played an important role in contributing to the factors for shark attacks to occur. On a whole this study gave a list of all the meteorological, ecological and environmental factors that could be the potential reason for the surge in shark attacks for the coast of North Carolina over the years from 2009 to 2016 and particularly in 2015

.

Limitations of the study include the fact that prediction of shark presence is based only on actual attacks; there is a lack of data on actual sightings of sharks in near shore waters not necessarily leading to attacks. With the addition of shark sighting data from a test of a standardized twitter hash tag and extension of the study to South Carolina and other Eastern states, the researchers hope to improve on the results which can lead to a test of an actual recommender system.

References

1. Shark Research Institute (SRI) homepage, Global Shark Attack File Incident Log, http://www.sharkattackfile.net, last accessed 2018/5/28.
2. Fordie, J.: The Shark Attacks In North Carolina, Explained. http://www.outsideonline.com/1999256/shark-attacks-north-carolina-explained (2015).
3. Burgess, George.: http://pilotonline.com/news/local/environment/expect-more-sharks-off-north-carolina-this-summer-experts-say/article_dd48563a-6bc7-5c7c-b811-

2b92b1e0db3d.html (2015)
4. Amin, Ritter and Wetzel.: An Estimation of Shark Attack Risk for the Coast of North Carolina and South Carolina. Journal of Coastal Research 31(5):1253-1259 (2015).
5. Palermo, Elizabeth.: Shark Bites Two: Possible Explanations for Attacks. http://www.livescience.com/51218-shark-attacks-north-carolina.html (2015).
6. Ritter, E., Zambesi, A., and Amin, R.: Do Lunar Cycles Influence Shark Attacks? The Open Fish Science Journal: (October, 2015).
7. Mockler, Butch.: Pick the Right Time to Fish. http://www.farmersalmanac.com, last accessed 2018/4/20.
8. NOAA Estuaries homepage, http://oceanservice.noaa.gov/education/tutorial_estuaries/est10_monitor.html, last accessed 2018/5/21.
9. Gardenette, N.: Warm Summer Waters Induce Shark Migration to East Coast Beaches. http://www.accuweather.com/en/weather-news/summer-shark-migration-ocean-east/29413314 (June, 2014).
10. Griffin, D.: Weather May Have Contributed to Attacks. http://www.cnn.com/2001/US/09/05/shark.attack/ (2001).
11. NERRS.: http://nerrs.noaa.gov/research (2016).
12. NOAA Ocean Service Education.: Dissolved Oxygen, http://oceanservice.noaa.gov/education/kits/estuaries/media/supp_estuar10d_disolvedox.html, last accessed 2018/5/21.
13. Hampton, J.: Expect more sharks off North Carolina this summer, experts say, https://pilotonline.com/news/local/environment/expect-more-sharks-off-north-carolina-this-summer-experts-say/article_dd48563a-6bc7-5c7c-b811-2b92b1e0db3d.html (March 2016).
14. Hammerschlag, N., Broderick, A., Coker, J., Coyne, M., Dodd, M., Frick, M., Godfrey, M., Godley, B.: Evaluating the landscape of fear between apex predatory sharks and mobile sea turtles across a large dynamic seascape. Ecological Society of America (August 2015).
15. Heikkinen, N.: Blue crabs migrate north as ocean warms, Scientific American (March 2015).
16. Larose, D., Larose, C.: Discovery Knowledge in Data, Wiley p. 138 (2014).
17. Haelle, T.: Shark bites are up, but attack risk is down? Scientific American (July 2015).
18. United Nations Office for the Coordination of Humanitarian Affairs: Hashtag Standards for Emergencies, https://www.unocha.org/publication/policy-briefs-studies/hashtag-standards-emergencies, last accessed 2018/5/21.
19. Dr. P. Thompson homepage: http://www.profthompson.net, last accessed 2018/5/24.
20. Ferretti, F., Chapple, T., Jorgensen, S., Micheli, F.: Reconciling predator conservation with public safety. Frontiers in Ecology and the Environment (August 2015).
21. Dorsal homepage, http://www.dorsalwatch.com, last accessed 2018/5/24.
22. Payne, N., Meyer, C., Smith, J., Houghton, J., Barnett, A., Holmes, B., Nakamura, I., Papastamatiou, Y., Royer, M., Coffey, D., Anderson, J., Hutchinson, M., Sato, K., Halsey, L.: Combining abundance and performance data reveals how temperature regulates coastal occurrences and activity of a roaming apex predator. Global Change Biology (March 2018).