Chith Sabesh
chithsabesh@gmail.com

GLOBAL SHARK ATTACK REPORT

OVERVIEW

The problem here deals with analysing and examining shark attacks based on various informations like country,gender etc. Project aims at coming up with a model for classifying fatal and non-fatal shark attacks based on both conventional techniques and text analysis.

THE DATA

The data is available to us from Kaggle (https://www.kaggle.com/teajay/global-shark-attacks) and hence no scraping is required. However there is a bit of cleaning and wrangling required.

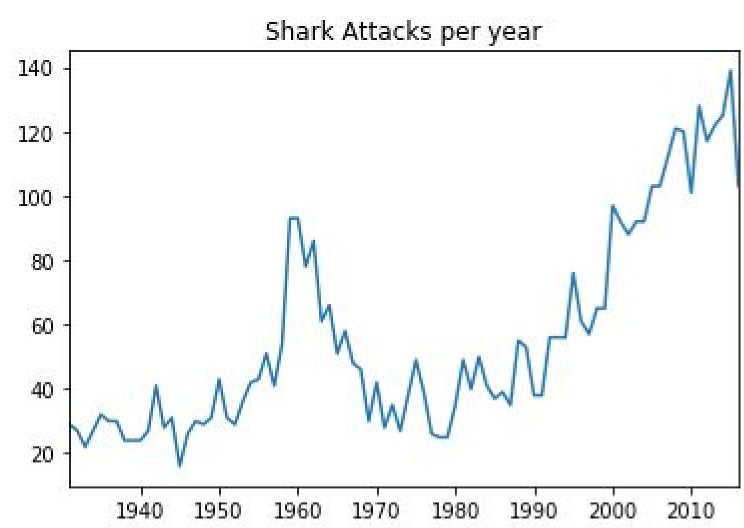
DATA WRANGLING

1) Missing values were treated appropriately for most of the features. Time was one such feature which had to be dropped since imputing time was found to be not possible.
2) Age was imputed with the mean of all ages.
3) The activities were cleaned to include only relevant activities.

EXPLORATORY DATA ANALYSIS

The following section can be divided to many other sections for ease of work.

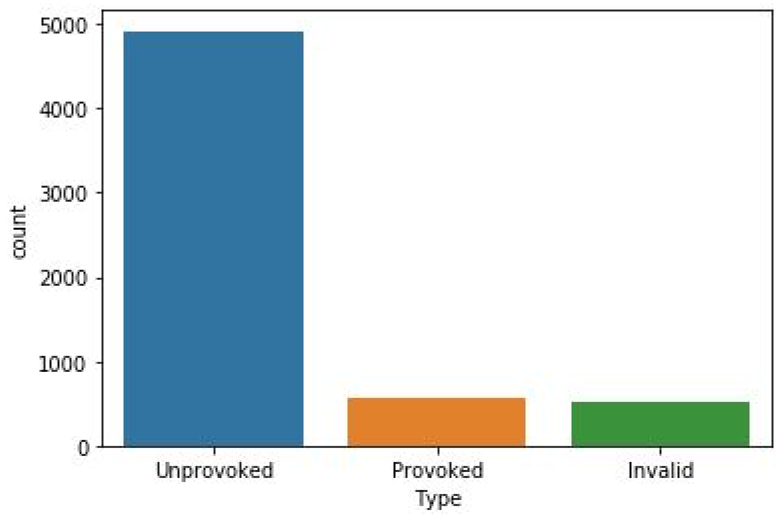
Plotting Year of attacks:



Shark attacks took a sharp increase after around 1980 can be attributed to the fact that humans started entering into the Oceans more and more and reporting of attacks became better

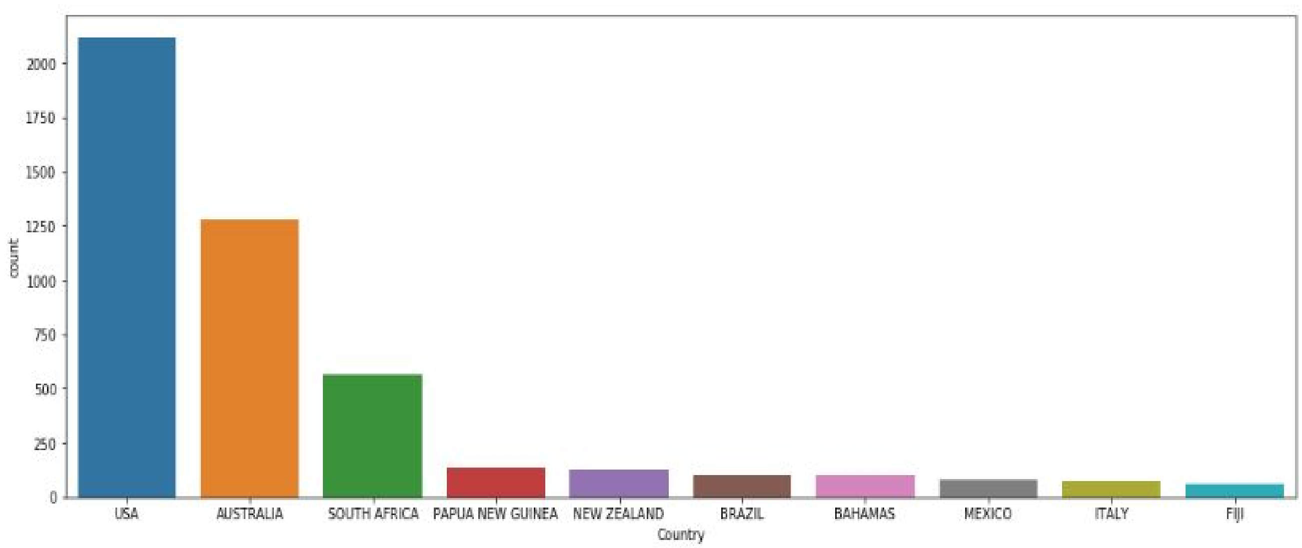
Chith Sabesh
chithsabesh@gmail.com

ATTACK TYPES



Most of the attacks are unprovoked which is reasonable since no one ever provokes a shark except maybe fishermen.

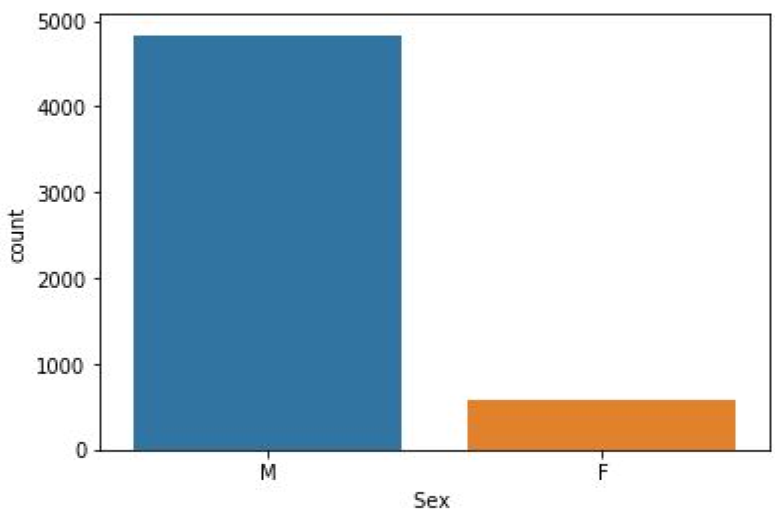
COUNTRIES



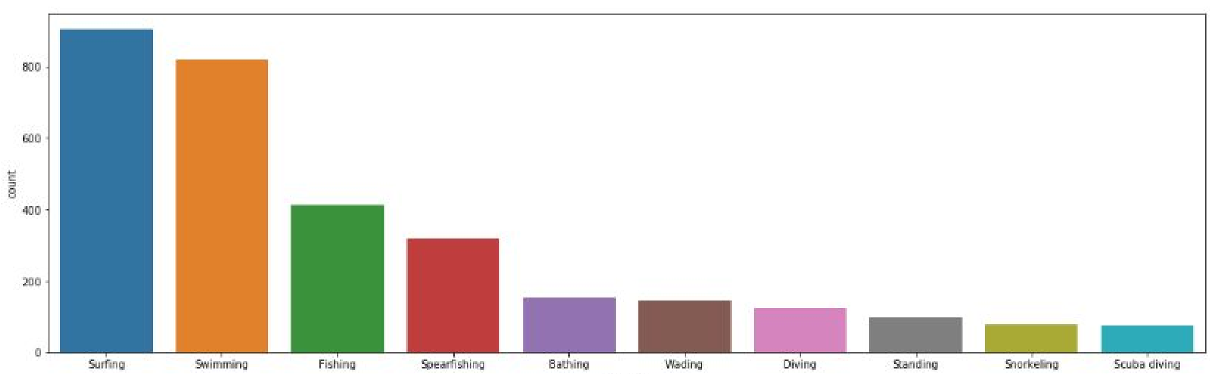
Most of these countries have a large coastline and hence large fishing industries.

Chith Sabesh
chithsabesh@gmail.com

GENDER



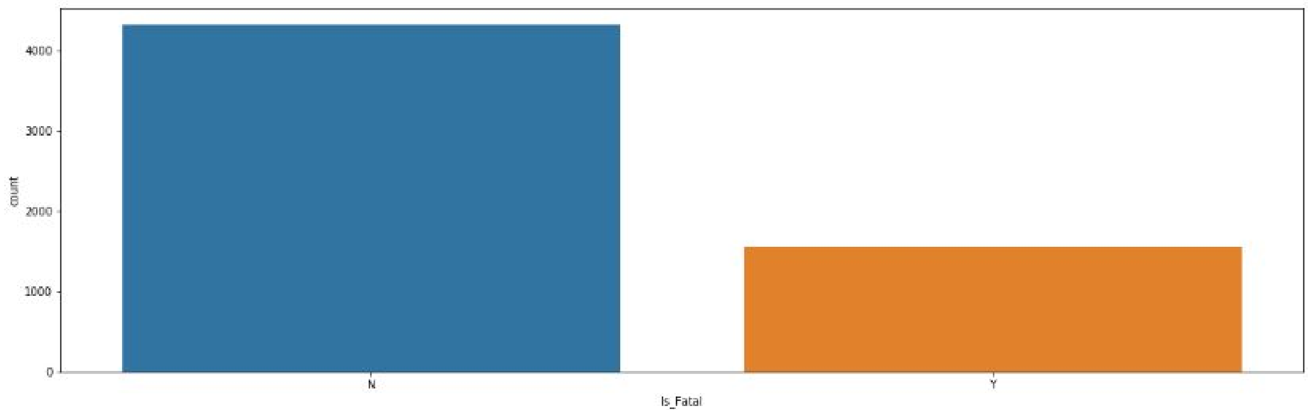
ACTIVITY



Surfing is the most attacked activity followed by swimming

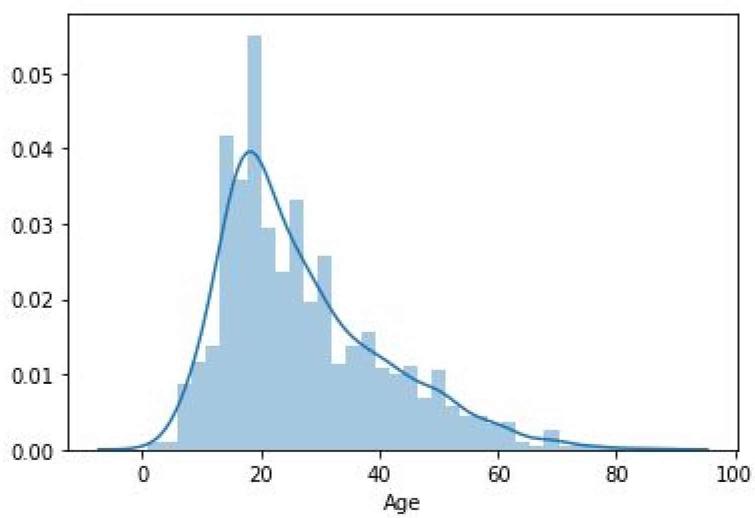
Chith Sabesh
chithsabesh@gmail.com

Fatalities to Non-Fatalities



Most of the attacks are non-fatal

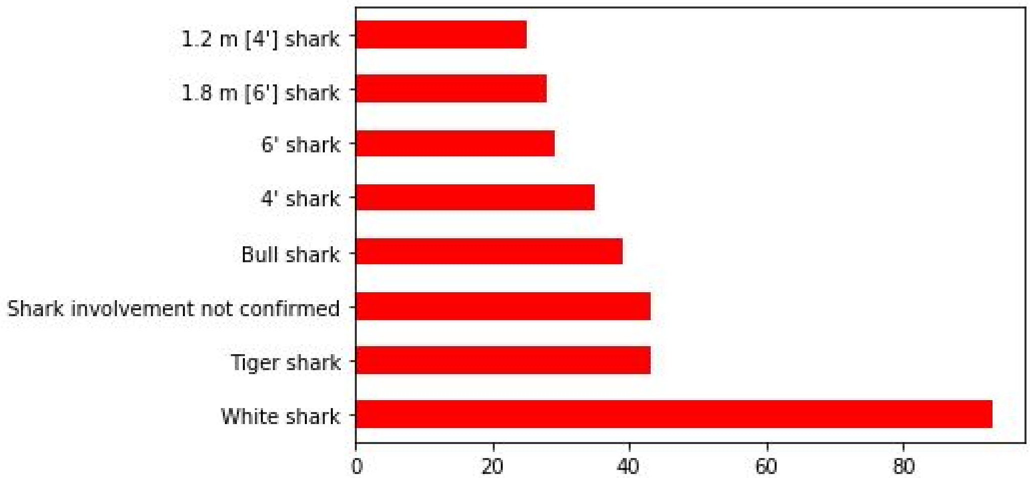
AGE DISTRIBUTION



The mean age of attacks is 27.08.There is a slight positive skew with maximum peak between 15 and 35. This is due to the fact that most of the teen and middle age adults are the ones who enter the water the most.

Chith Sabesh
chithsabesh@gmail.com

SPECIES



The species column is very unclean and hence only some analysis could be done and it was found that White Shark the biggest of all carnivorous sharks has attacked the most.

WORD CLOUDS

NON-FATAL

Chith Sabesh
chithsabesh@gmail.com

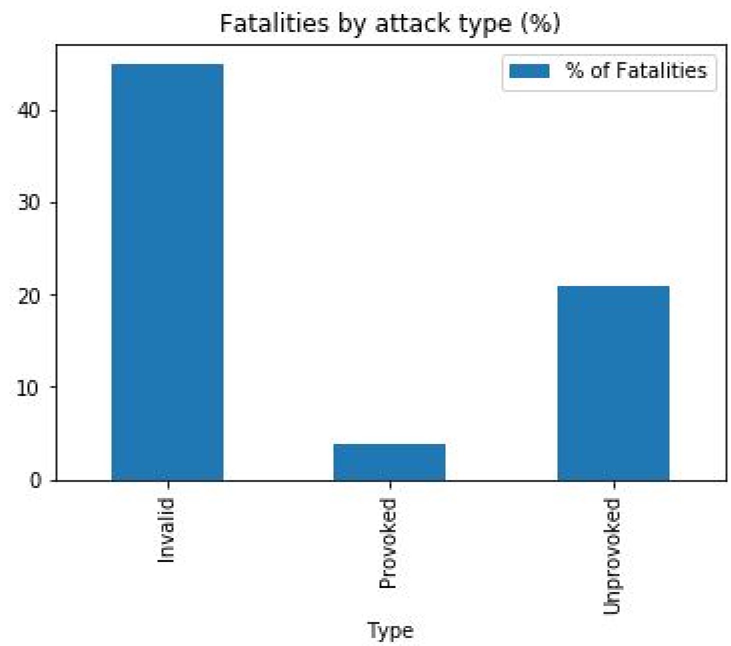
FATAL

Chith Sabesh
chithsabesh@gmail.com

Clear difference between both the word clouds

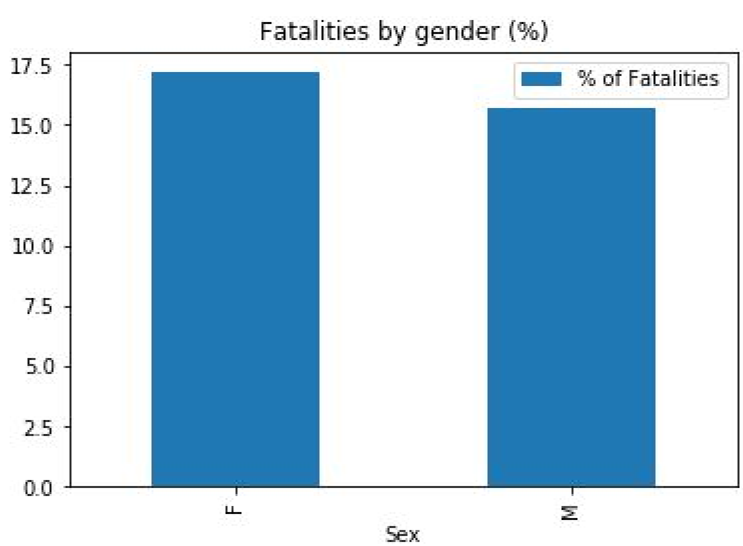
DOES PROVOCATION LEAD TO FATALITY

Chith Sabesh
chithsabesh@gmail.com



No unprovoked attacks always tend to have more fatalities here both Unprovoked and Invalid have high fatalities (Since most shark attacks happen due to mistaken identity and not due to any provocation)

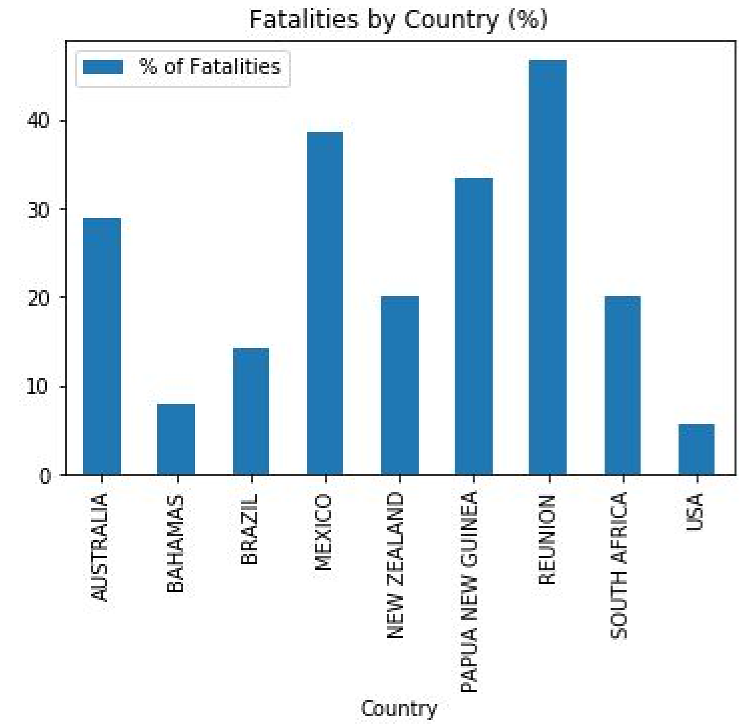
DO FEMALES DIE MORE



Strangely females have higher fatalities than males

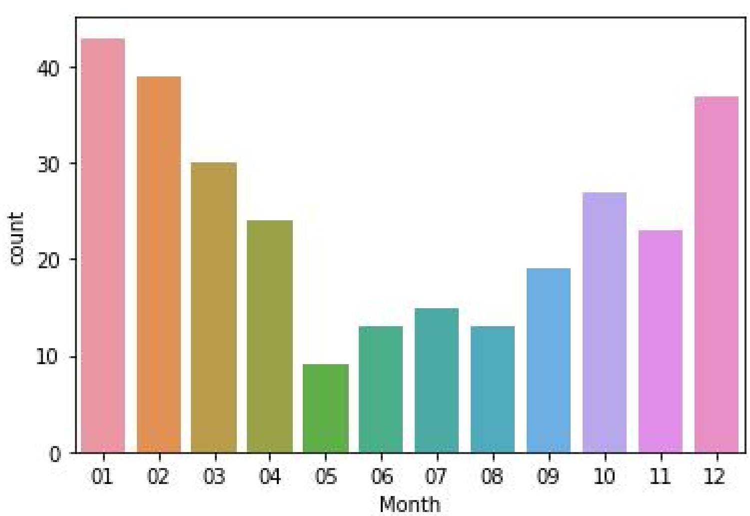
FATALITIES BY COUNTRY

Chith Sabesh
chithsabesh@gmail.com



Here it should be known that Reunion has only about 15 attacks out of which nearly 50 percent of attacks are fatal. The case is same for both Mexico and PAPUA NEW Guinea

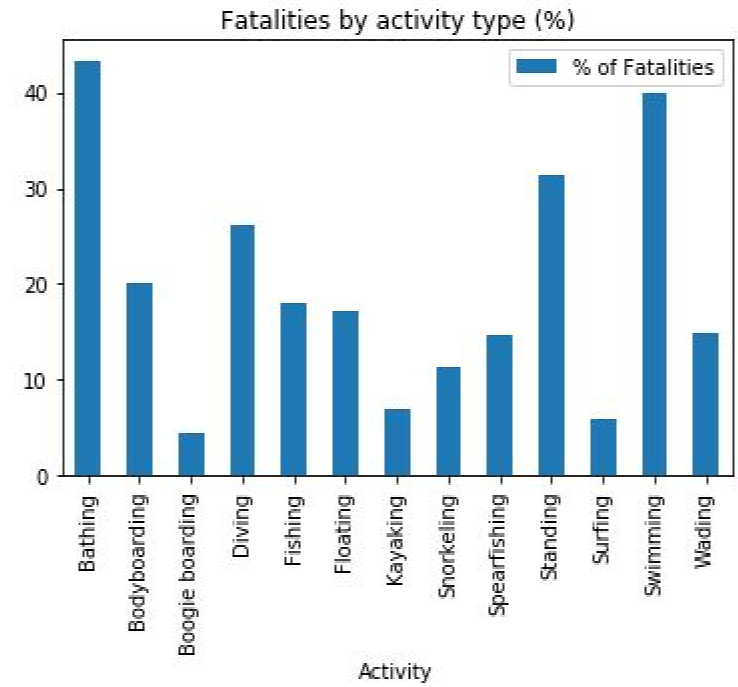
MONTH WISE ATTACKS IN AUSTRALIA



Attacks happen in the height of summer not only in Australia but all over the world and hence no point in checking for fatalities month wise

Chith Sabesh
chithsabesh@gmail.com

Fatalities By Activity



Although surfing has the most attacks it is not the most dangerous that distinction goes to bathing and swimming.Reason for this could be that sharks are attracted to splashing activities in the water

INFERENTIAL STATISTICS

This section presents the results of inferential statistics methods applied on two hypothesis tests namely:

1.) Relationship between fatality and activity
2.) Relationship between fatality and gender

Relationship between fatality and activity

This test was performed to test whether there is a relationship between fatality and activity in other words to see if the activity influences the fatality. To do this we only took the top 10 activities since there are many other activities that are big strings and they also don't make sense.Since both the variables are categorical variables we do the chi-square test to check for dependencies.We made a contingency table using the two variables

Chith Sabesh
chithsabesh@gmail.com

Is_Fatal	N	Y	All
Activity_new			
Bathing	75	61	136
Bodyboarding	95	24	119
Boogie boarding	46	2	48
Diving	260	86	346
Fishing	309	63	372
Floating	32	6	38
Kayaking	41	3	44
Playing	20	1	21
Snorkeling	64	8	72
Spearfishing	270	44	314
Standing	122	49	171
Surfing	894	53	947
Swimming	541	346	887
Wading	197	33	230
All	2966	779	3745

1.There is significant relationship between fatality and activity
2.P-value obtained was 4.14 x 10^-69

Relationship between fatality and gender
Here again we perform chi-square test for finding dependencies between the two variables

Is_Fatal	N	Y	All
Sex			
F	362	77	439
M	2600	702	3302
All	2962	779	3741

1.There is no significant relationship between gender and fatality
2.P-Value obtained here was 0.516

Chith Sabesh
chithsabesh@gmail.com

FEATURE ENGINEERING

The feature engineering done here was to create a month feature from the case number feature. Features with an enormous amount of missing values like species and features like time which cannot be imputed were dropped. Then the activities were binned and only the activities which had a count of greater than 15 were taken in the final train dataset.Finally one-hot encoding was done to all the categorical variables.

MACHINE LEARNING

The next step was to build a classifier which classifies the fatalities to non-fatallities.

FEATURES USED

1) Attack Type
2) Gender
3) Month
4) Age
5) Country
6) Activity

MODELS USED

1) Logistic Regression
2) Decision Tree
3) Random Forest
4) Ada-Boost
5) XG-Boost
6) Gaussian Naive bayes
7) Multinomial Naive bayes

Since this dataset was imbalanced(i.e it had a 79-21 class percentage) SMOTE had to be applied on it to make it balanced.

Chith Sabesh
chithsabesh@gmail.com

MODELS PERFORMANCE

	Logistic Regression	Decision Tree	Random Forest	Ada Boost	XGBOOST
TP	164.000000	198.000000	155.000000	132.000000	134.000000
TN	648.000000	466.000000	665.000000	731.000000	749.000000
FP	81.000000	47.000000	90.000000	113.000000	121.000000
FN	235.000000	417.000000	218.000000	152.000000	134.000000
SENSITIVITY	0.411028	0.321951	0.415550	0.464789	0.500000
SPECIFICITY	0.888889	0.908382	0.880795	0.866114	0.860920
Precision	0.669388	0.808163	0.632653	0.538776	0.525490
f1_score	0.509317	0.460465	0.501618	0.499055	0.512428

Since this was an imbalanced class problem the accuracy metric becomes the f1-score.

From the above table we can say that XG-Boost performs best but it will be very obvious that it performs best due to its mathematical complexity. But for a small dataset like this it would be better to go with Logistic Regression or Random Forest.

HYPERPARAMETER TUNING

Tuning both logistic regression and random forest did not yield better results The best c parameter for Logistic regression was 1.
The best parameters for Random Forest was

```
{'bootstrap': True,
 'max_depth': 30,
 'max_features': 3,
 'min_samples_leaf': 3,
 'min_samples_split': 8,
 'n_estimators': 300}
```

CLASSIFYING USING TEXT FEATURE

There is a feature named injury which describes the type of injuries that person has got. So we will try to classify fatalities with that text column.

CLEANING THE INJURY COLUMN

1) Converting all text to lower case
2) Removing stopwords
3) Applying Stemming

Chith Sabesh
chithsabesh@gmail.com

The next step was to create a Document Term Matrix to that text column.This is a highly sparse matrix. With a sparsity of 99.5%

Both Gaussian Naive bayes and Multinomial Naive bayes were used in this classification.

The multinomial naive bayes performs better than the Gaussian Naive Bayes with a very good sensitivity of 98.300

CONCLUSION

This report highlights the process of Data Wrangling,Exploratory Data Analysis,Inferential Statistics,Feature Engineering and Machine Learning done on the global shark attacks dataset and a Logistic Regression classifier was built to classify fatalities.It should also be noted that unsupervised clustering was also applied to this dataset but there was not much inference we could gather.