**LM StarGazer - Design of a Decision Support Tool and Data Science to Predict Illegal Fishing**

A.Mityushina, A. Misra, E.Taylor, R. Skaddan, S.Jagtap

Contents

[Abstract 3](#_30j0zll)

[Context Analysis 4](#_3znysh7)

[Need Statement 4](#_2et92p0)

[Stakeholder Analysis 5](#_tyjcwt)

[Concept of Operations 5](#_sqyw64)

[Data analysis](#_4d34og8) 6

[Data Sources](#_17dp8vu) 6

[Preparation](#_2s8eyo1) 6

[Data Inconsistencies](#_3cqmetx) 7

[Data Conditioning](#_35nkun2) 8

[Code Revision and Modification](#_1ksv4uv) 8

Attribute transformation9

Adding Attribute 9

[Satellite Imagery](#_1rvwp1q) 9

[Analysis](#_4bvk7pj) 9

[Predictive Modeling](#_3j2qqm3) 11

Position Model 12

Future [Predictive Modeling](#_3j2qqm3) 12

Anomaly Detection 13

Bayesian Net14

Image Extraction 15

[Proof of Concept 1](#_4i7ojhp)5

[Visualization 1](#_1664s55)6

[Results 1](#_2bn6wsx)8

[Conclusion 1](#_3q5sasy)9

[Future Work](#_1pxezwc) 19

[References](#_25b2l0r) 21

[Biographies](#_nmf14n) 22

[Appendix](#_111kx3o) 22

**Abstract*:***

Illegal, unreported and unregulated (IUU) fishing is a worldwide issue contributing to ecological devastation. It is estimated that 30% of the total fishing yield is due to IUU, making it a top priority for global law enforcement services. Fishing in restricted areas can have detrimental effects on the ocean floor and fish populations pushing species towards extinction and causing damage to ecosystems on a global scale. Enforcement agencies vary depending on the country, therefore the primary stakeholder in the United States is the Coast Guard. One of the primary inefficiencies of detecting a vessel’s location/activity is the lack of data. The Automated Identification System (AIS) which broadcasts ship information is mandatory on all large vessels but is inconsistent, easy to modify, and not continuous. This project develops a decision support system which assists law enforcement in the apprehension of vessels who are partaking in illegal fishing by observing vessel behavior, predicting the likelihood of a ship crossing into a restricted area, and providing a satellite image of the area.  The system is gauged in supplementing law enforcement with additional information and evidence of vessel activity. The behavior and data analysis can be extrapolated to other vessel identification scenarios, making this project valuable to other situations with similar actors.

Currently there are no tools which can make a prediction about the ‘status’ (fishing or not fishing) of a vessel, determine if the vessel is in a restricted region, and recommend a course of action. There is a need for a vessel identification system that uses the available data and quickly provides recommendations to the personnel monitoring an identified region. By improving the identification process and allocating missions to the apprehension vessels, it is expected to have an increase in response time by 10%.

AIS data is used as a baseline for the identification of illegal fishing, but there are a few other ways to supplement those datasets. The current method is for a station to detect a ship’s location through radar towers. The radar detection provides more precise and continuous information than the AIS feed. A drawback is that the tower has a limited range and is primarily restricted by land. Another method of catching a vessel engaged in illegal fishing is by having patrol boats conduct random walks around protected areas. Our system will not replace the existing methods, but can aid in the identification process. By creating a tool that adapts and ‘learns’ vessel behavior, the prediction component for vessel status becomes more accurate as data is fed into the system. With the ability to coordinate behavior and location, it is possible to predict a likelihood of illegal activity. If a likelihood exceeds a value of .7 (on a scale of 0 to 1) then the system checks for available satellite coverage. If a satellite image is available then a portion of the image with the vessel’s position is captured and added to the output file. With a likelihood of above .7, a condensed report will be displayed to the user containing all available information about the vessel in a comprehensive manner.

The tool can be partitioned into two key functions. The first model is a regression model based on AIS location data, speed of the vessel, type of the vessel. This model can predict if the vessel is fishing or not fishing. The second model uses light analysis to identify low light emission from a satellite image. The identification of low light targets assists with vessel identification and tracking. This is an important component as the current knowledge dictates that much of the illegal fishing occurs after the sun has set. The system recognizes and flags a vessel operating with an identified behavior, then the data is enriched by coordinating satellite imagery. The model will be trained on historical AIS data and satellite imagery, tests are conducted by feeding individual vessel data sets with known behavior and determine if the output report matches the expected information. This process of vessel identification through AIS tracking, and a confirmation through satellite imagery decreases the response time therefore augments the analysis for vessel apprehension. Vessels identified with a high risk of illegal activity are communicated to the user through a display. The identification of a vessel engaged in illegal behavior is identified through the support tool, but the ultimate decision to engage in apprehending the vessel will be left to the individual.

## **Context Analysis:**

Illegal fishing is a worldwide concern and through global efforts, it is reducing in quantity through stricter regulations and improved detection methodologies. The impact of overfishing a species can change global ecosystems and have unexpected consequences by damaging other dependent species and the environments they live in. Roughly 11 to 20 million tons of fish is caught illegally, equaling about 30% of the fishing market [1]. By sectioning off and protecting regions, the ability for vessels to dredge, deplete and damage the environment is reduced. Unfortunately, protection of the identified areas is complex, costly, and potentially dangerous. With minimal data and few resources to allocate patrolling large bodies of water, illegal fishing continues to thrive.

The primary apprehension unit for the United States is the Coast Guard, who intersects vessels operating in restricted areas around the coast. Fishing detection and vessel apprehension is heavily dependent on data. The decision to engage with a suspect vessel is costly, potentially dangerous, and should not be done without absolute certainty. While it is possible to track a vessel and know it’s relative position through the Automatic Identification System (AIS), the Coast Guard supplements the data source with radar detection. The radar detection provides a precise and continuous track of vessels, but only covers major ports and is unable to detect every vessel on the coastline. Therefore, adding an analysis to the available data improves the detection process.

Previous work was conducted by another George Mason University Capstone group which built a regression model to predict if a vessel is fishing or not-fishing [2]. Our system incorporates the previous work and furthers the analysis by relating the location points to the fishing status to create a prediction of if the vessel is conducting fishing in a protected area or not.

Lastly, illegal unreported and unregulated fishing is not the only application of this methodology. It is possible to relate a regression model and a position prediction model to other scenarios such as drug and human trafficking with boat transportation. While this model is outfitted to detect illegal fishing, the process of having a model which learns vessel behavior and can make an analysis about its position, known path, and where it may be going is valuable to different enforcement agencies.

## **Need Statement:**

As stated above, IUU can have detrimental effects on the ocean floor and fish populations, pushing species towards extinction and causing damage to ecosystems on a global scale. Vessels conducting IUU fishing do not adhere to global rules regarding bycatch limits or quotas, further harming fish populations.

Not only does IUU affect the environment in general, it also interferes with other fisheries trying to fulfill their sustainability promises. IUU fishing bring extra fish into the market, reducing prices of fish and resulting in less profit for legal fisheries.

Therefore, this project fulfills the need to improve the detection process and determine which boats are conducting IUU fishing. The faster and more thorough the analysis in the detection of illegal fishing results in more boats being stopped, therefore the less of an impact IUU fishing will have on our environment and economy.

## **Stakeholder Analysis:**

The primary stakeholder is the fishing vessels. They want as little government interference as possible and the fact that they have to register, request to leave port, and have a tracker on their ship is viewed as a hassle and a waste of time. Then, whenever the fishermen encounter the coastguard it has the same viewpoint as other forms of law enforcement agencies; general feelings of disgust until the ship is in trouble. Also, not all of illegal fishing is intentional. It is possible to set nets in a legal location and then the current changes causing the nets to drift into a protected region. Then upon retrieval, the ship must travel into the restricted area and run the risk of being caught or losing their equipment.

The user of this system is the United States Coast Guard or another countries vessel apprehension organization. Therefore, not only does the analysis have to be thorough but the output and user interface must be comprehensible and usable so that anyone can understand the situation. The Coast Guard is a branch of the US military that is tasked with enforcing US laws in open waters. The commitment to arrest, apprehend, or punish a captain/vessel can be a cumbersome task. The proper team must be assembled with the correct vehicle and adequate supplies, therefore the decision to engage with a large vessel is taken under thorough consideration. To achieve the decision to board and apprehend a vessel, there must be an overwhelming circumstance with data to formulate the foundation to risk engagement. This system would aid the Coast Guard by running the system which prioritizes active vessels with the highest likelihood of conducting illegal fishing and producing an automated report with all available data.

## **Concept of Operations.**

Throughout the day, a Coast Guard user would conduct their standard surveillance operations but is able to switch the computer monitor to the systems user interface. The interface provides a map of the monitored region with vessels recorded tracks, a list of unique vessel identification numbers with the likelihood of illegal activity, and condensed reports for selected vessels; the system defaults to the vessel with the highest rank. A user could select the unique vessel id and view the condensed report containing the logic for the illegal activity detection and a satellite image if available. The report can be saved and used as evidence to initiate an investigation.

LM Stargazer will be incorporated into the surveillance component of the existing system. The user would identify a region to monitor and run the system throughout the day. Detection of illegal fishing vessels is identified by surveillance of designated areas and collecting data. The system continually pulls data from public data sources, such as Global Fishing Watch, and identifies a list of recommended vessels to investigate within the identified region.

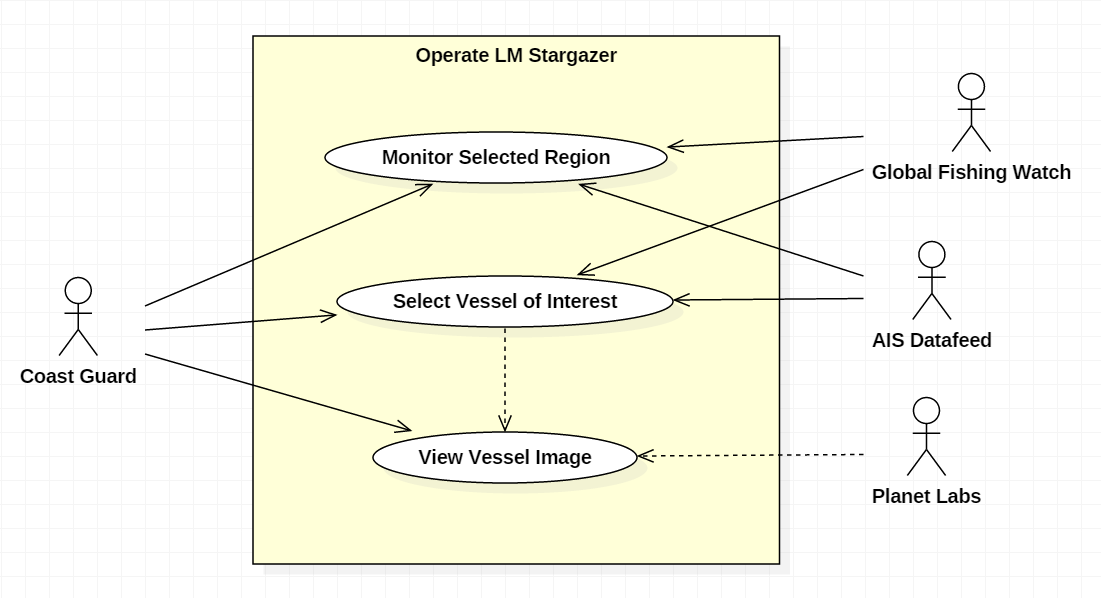


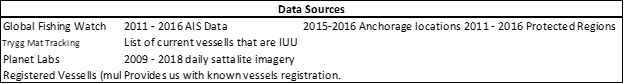
Figure 1. Use Case Diagram

In the image above, it describes the actors involved with the basic functionality of the system. The Coast Guard only interacts with the system and does not communicate with the data sources directly. The Coast Guard user has the ability to monitor the selected region or select a vessel and find out supplementary information about its path. To incorporate the system into the detection process, the current methodologies used by the Coast Guard can be improved by prioritizing the list of predicted vessels fishing in an illegal location.

## **Data analysis**

Data analysis included data collection, data exploration, data validation, and data cleaning. The major sources of the data used in this project are Global Fishing Watch (GFW) AIS, Point of Interest and Planet Labs Satellite imagery data.

### Data Sources:

**  
 Table 1: Data Sources Utilized**

Data Preparation:Global Fishing Watch (GFW) is a new technology platform developed by Google, Sky Truth, and Oceana that uses satellite Automated Information Systems (AIS) data to monitor fishing activity around the world in near real-time. GFW provides the user with vessel identity, fishing activity, transshipment information, and anonymized AIS data.

The fishing hours of vessels on each day have been combined to get the fishing hours for gear types. Fishing vessels were categorized into four groups of fishing gear type: longline vessels, trawler vessels, Fixed Gear vessels and seiner vessels. Every vessel with different gear type had a distinct behavior: A longline fishing vessel is described to use long line, with baited hooks attached at intervals by means of branch lines called snoods to catch fish and once the lines are laid the vessel moves at very slow speed. The time required to fish in this manner varies but may take up to a full day. A trawler fishing vessel is described to capture fish by dragging a net behind the vessel while moving at a very slow speed. These vessels will typically fish from 3-5 hours. A purse seiner fishing vessel searches for large schools of fish, and once identified this method uses net called a seine, that hangs vertically in the water with its bottom edge held down by weights and its top edge buoyed by floats. And the fixed gear: a category that includes set longlines, set gillnets, and pots and traps. This behavior and inputs are used to create training data which can be summarized as vessel track information generated by AIS data and classified at each track point with the classification of fishing, not fishing, or unknown.

Global Fishing Watch sourced AIS data about vessels present in the ocean, contains latitude, longitudes, time stamps along with the MMSI (Maritime Mobile Service Identity) which acts as the primary key which uniquely identifies each ship in our study. MMSI itself is a composite primary key containing a lot of information about the vessel, hence is an authentic entity to act as a primary key. We studied AIS data of three types of vessels differing primarily in the fishing gear they use namely Longliner, Purse Seine and Trawler. Another set of data used was acquired from Kristina’s research group, which had each MMSI hand labelled as ‘is-fishing’ and ‘not-fishing’. We used MMSI and time-stamp as anchors to merge the two data sets to give as a merged output giving us each MMSI matched to its time stamp, precise location and the final label (is-fishing and not fishing).

## 

## Data Inconsistencies: There are few inconsistencies observed in the data.

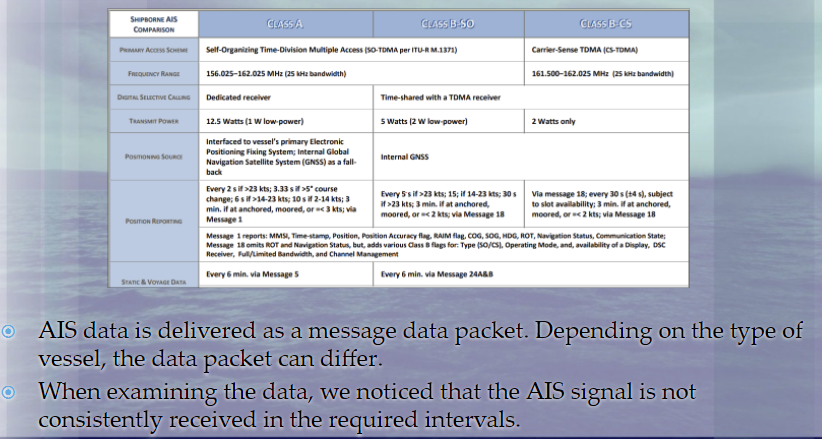


Figure 2: Data Inconsistencies

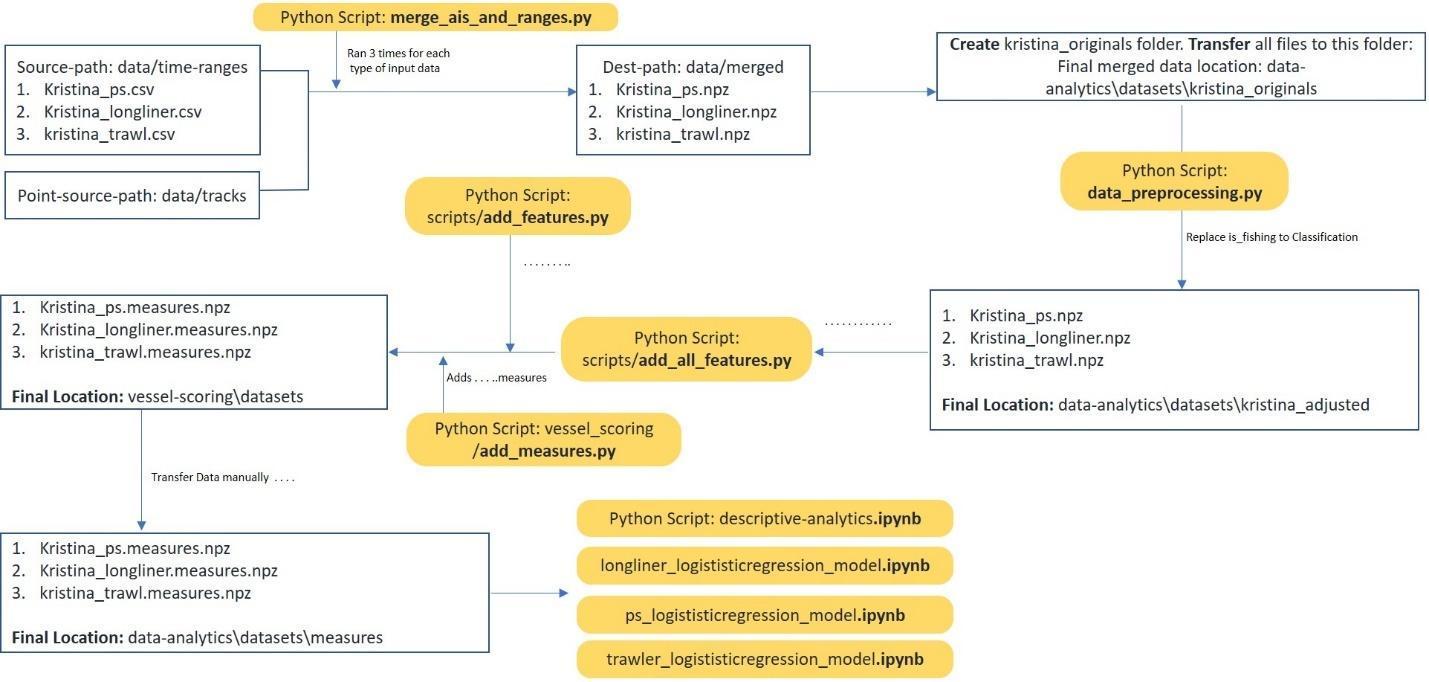


Figure 3. Process Flow

Data Conditioning. The training Data was accessed from the Global Fishing Watch GitHub repository, which included several merged datasets. To access the merged training datasets the "prepare.sh" script inside the repository had to be run. The "prepare.sh" script is used to run the python codes that merge different datasets to combine separate vessel track files into a combined gear-specific file; i.e., one file for each longliner, trawler, purse seine, fixed\_gear, squid\_jigger and other\_fishing equipment.  After obtaining all the training datasets, the data was cleaned by removing all the rows with missing or duplicate data.

Code Revision and Modification.The set of scripts acquired from the previous research group were modified to Python 3.6 to build a futuristic repo of scripts capable of getting support from latest Python 3.6 framework as well as allowing developers and data analysts to seek best help from Python developer community as most of the developers have already switched to Python 3.6. Also, the code revision has ensured more capability of data analytics and has widened the freedom to try new data manipulation strategies with newer libraries being introduced to Python 3 framework. Code revision also alleviates the concern of the prior code becoming totally useless as older Python libraries used will be sooner or later deprecated, with almost no community support. At some places to reproduce the same kind of result a modified algorithm was used in the best interest to get the latest scripts functional in Python 3 framework.

Data Processing: In real life scenarios the data collected is not continuous hence we observe many nulls. There are many ways to handle nulls

1. drop null(NaN) or infinity(inf) values
2. assign class average as the value of the missing attribute
3. assign model value in place of nulls

We have used the drop method as there is very less number of missing tuples in our dataset.

Attribute transformation: Attribute transformation generally adjusts the data by replacing an attribute by one or more new attributes, functionally dependent on the original one, to enable further investigation. This is typically performed before creating predictive models to bypass some limitations of the modeling algorithms used or improve model performance.  The transformed ‘is\_fishing’ to ‘classification’, as it directly matched our requirement to address vessels as classified as fishing or not.

Adding Attributes.We had very little information given in the form of AIS data but through more attributes we can find other aspects of the scenario which can help us understand data better. Like from latitude and longitude values we could find the *course* and combining this with time stamp we can find at the average speed of the boat. Initially we had 8 attributes in the data and could derive 70 more attributes by computing averages rolled over specific windows of time intervals.

Satellite Imagery.The team has partnered with Planet Labs to acquire Satellite imagery from 2009 to 2018. The data is accessible to the team via an API call which allows the final system to request an image with determined end points to be clipped and shipped to the user interface. At a later time this image will be analyzed to determine the position of a protected are in comparison to the boat.

## **Analysis:**

We have run the logistic regression models developed by prior teams based on Dalhousie University Machine Learning algorithm developed to learn tracks and predict fishing behavior. The model predicted if a vessel is fishing or not fishing. Below are results of the logistic regression models developed by prior team for various Gears.

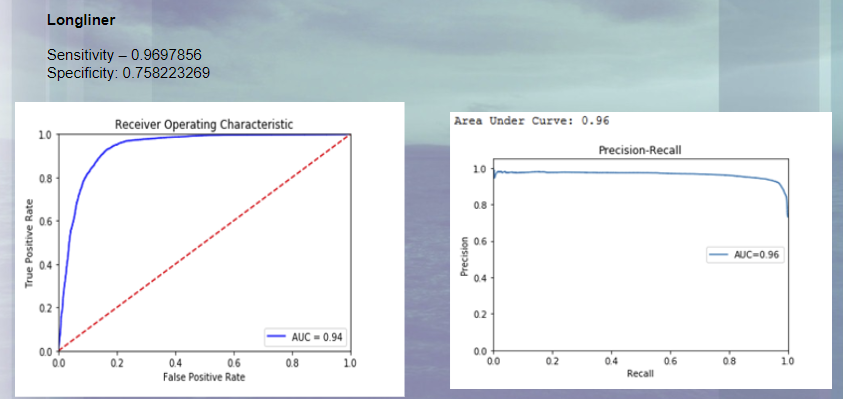


Figure 4: Logistic Regression Results for Longliner

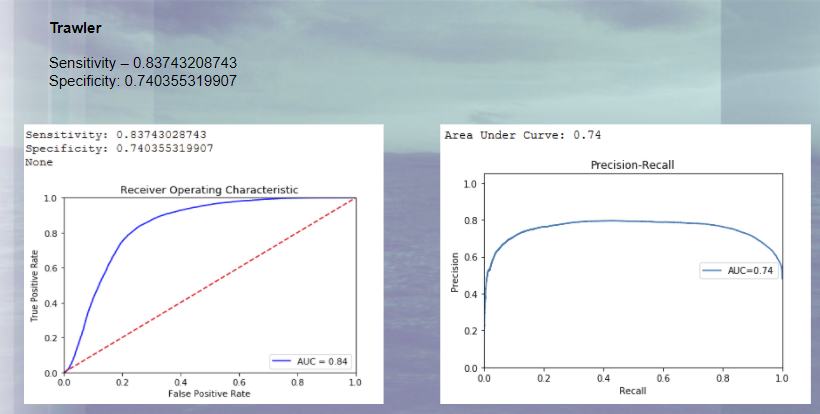


Figure 5: Logistic Regression Results for Trawler

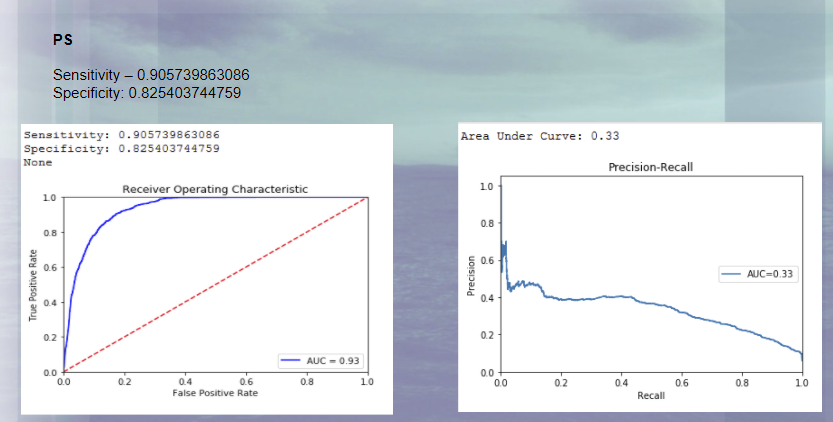


Figure 6: Logistic Regression Results for PS

From the above results we have observed that as the models are using all the attributes without any dimensionality reduction methods, it is probably overfitting to the data.

To improve the model developed by the prior team we have developed a system that include Six components. A **Regression model** – to predict the behavior of a vessel and predict if it is fishing or not fishing, a **position model** – to predict overlap between a ship and identified regions, a **Monte Carlo prediction** – to predict the next location the vessel will be, an **Anomaly detection** model used to detect outliers in dataset, a **Bayesian net** – to use the generated information to form a prediction about the likelihood of a vessel engaged in illegal fishing. And finally, when the image is available, an **image extraction** – to accesses Plant Labs API and collect an image from a selected region and specified timeframe.

Predictive Modeling. **Gradient Boosted Regression Model**. From prior work, the team was given a baseline for improved model performance with a logistic regression model over Boerder’s models. Logistic regression is a heavily trusted and used tool in risk modeling, prediction models (Quantitative Structure-Activity Relationship), and discovering gene interactions. It has shortcomings as it can overfit and it can predict poorly on multi-level factor datasets. It has advantages over classifiers such as random forest models when needing ranking and probability estimation. For our purposes, ranking and probability estimation is needed, but our findings on other types of modeling will be discussed in further sections. To overcome some of the shortcomings of the logistic regression, the team tried using a gradient boosted mechanism with variable transformation. The new logistic regression model utilizes an R package called Gradient Boosted Machine (GBM) with a logic function. Further, the new model attempts at increasing usability of the model by combining three models into one without losing the importance of variables and gear type. Meaning, instead of building three R models for the three gear types, we have interactions of variables with variable called gear type. This handles the change in the model’s coefficients for various scenarios.

First, we examined the data for each gear type separately (i.e. longliner dataset). We then remove all unknowns (i.e. ‘-1’) from the training dataset. The unknowns can be a data condition where the behavior of the vessel was not known or mislabeled. Removing this from the training helps us reduce noise in the model. Further, we want to be careful in our modeling not to overfit the data, as the known labeling can have missed labels due to less than 100% accuracy when the labels were created.

This left us with a known dataset of fishing and non-fishing observations for a specific gear type. We had 138, 666 observations with 96 variables for longliner, 166910 for purse saline, and 323958 for trawler. the three gear type datasets. We also included two variables geartype and geartype\_code. These variables will be used for the final interactions in the model. We also create a dataset for all observations with out unknowns and with the two new variables added (i.e. geartype and geartype\_code).

During the modeling part, we utilize R’s package registerDoParallel to help speed up the processing time. With out parallelization, the run time of the model is 45 minutes as compared to 16 minutes. (We assess the processing speeds with R’s function system.time() to help record and later overcome long training times with 3GB dataset).

The GBM package is initially ran with 0.05 step size (i.e. shrinkage), 800 trees, bernouilli distribution to get the probabilities, interaction depth of 2, 5 cross validation folds, and 80% train function.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Iter | TrainDeviance | ValidDeviance | StepSize | Improve |
| 1 | 1.1144 | 1.1152 | 0.05 | 0.0314 |
| 2 | 1.0598 | 1.0603 | 0.05 | 0.0274 |
| 3 | 1.0121 | 1.0124 | 0.05 | 0.0239 |
| 4 | 0.9701 | 0.9703 | 0.05 | 0.0211 |
| 5 | 0.9321 | 0.9324 | 0.05 | 0.019 |
| 6 | 0.8913 | 0.8919 | 0.05 | 0.0205 |
| 7 | 0.854 | 0.8549 | 0.05 | 0.0186 |
| 8 | 0.8247 | 0.8255 | 0.05 | 0.0146 |
|  |  |  |  |  |
| 480 | 0.1798 | 0.1814 | 0.05 | 0.0004 |
| 500 | 0.1769 | 0.1787 | 0.05 | 0.0002 |
|  |  |  |  |  |
| 660 | 0.1582 | 0.1608 | 0.05 | 0.0001 |
|  |  |  |  |  |
| 780 | 0.1478 | 0.1508 | 0.05 | 0.00000 |
| 800 | 0.1462 | 0.1494 | 0.05 | 0.00000 |

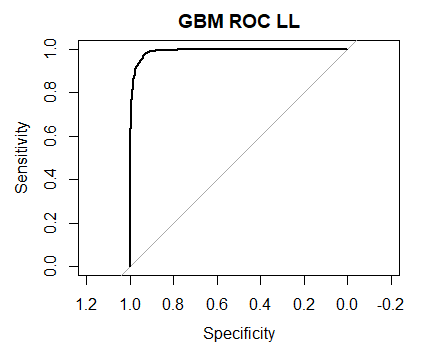
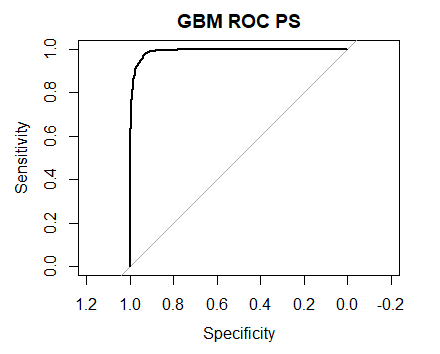
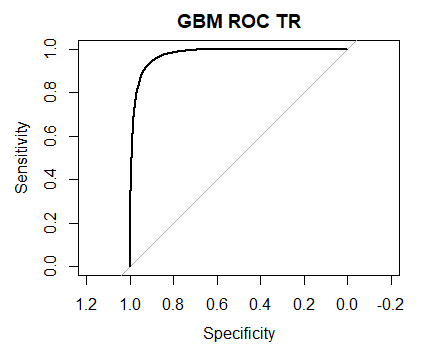
The gbm model starts to overfit around 700 trees in that we no longer see an improvement form additional tress, but we do see a reduction in TrainingDeviance (i.e. 20% hold out) and ValidDeviance (5 fold cross validation). The algorithm will pick the best iteration as 800 trees, but we set it to 700 to reduce overfitting to noise in the data. Again, we want to be careful as our data had original hand labeling which is prone to human error. We repeat this step for PS and TS datasets. Best iteration for

Then we run the gbm function for summarization which allows us to view the top contributors in explaining our variance in the dataset. For longliner, we find that the following 9 attributes contribute to 92.96% of variance in the data. We could select more attributes, but we want to make a model with the least number of attributes that will reduce its chances of being susceptible to noise.

|  |  |
| --- | --- |
| measure\_coursestddev\_43200 | 26.46993 |
| distance\_from\_port | 21.73128 |
| measure\_pos\_86400 | 20.8393 |
| measure\_coursestddev\_86400 | 6.974363 |
| measure\_coursestddev\_21600 | 5.669901 |
| distance\_from\_shore | 4.320157 |
| measure\_pos\_43200 | 4.258876 |
| measure\_speedavg\_21600 | 1.726125 |
| measure\_pos\_21600 | 0.969156 |
| KEY: | |
| 86400 | 24 hours |
| 43200 | 12 hours |
| 21600 | 6 hours |
| 10800 | 3 hours |
| 3600 | 1 hour |
| 1800 | 30 minutes |
| 900 | 15 minutes |

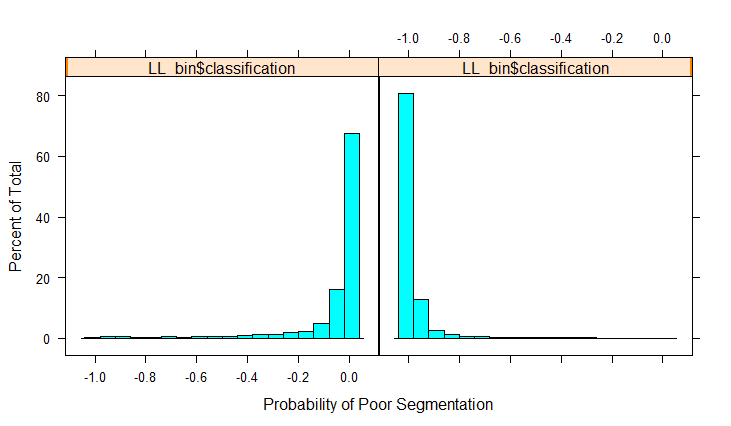
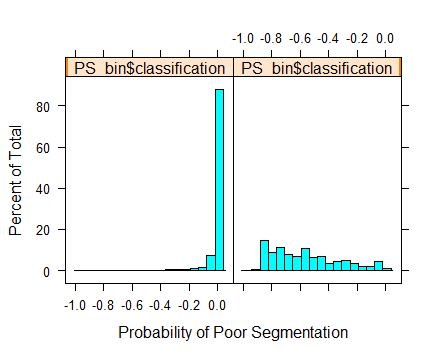
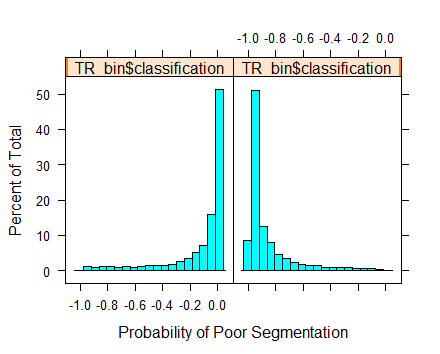
From these we observe that deviation in course at 24, 12, and 6 hours play a significant role in determining if the vessel is fishing. Distance from port and distance from shore are other attributes that come important for the longliner. Positioning of the vessel, which is its change in lat and long, is another set of variables that comes important at 24, 12, and 6 hours. Lastly, average speed at 6 hours comes predictive.

Validation of each individual model yields the following results:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **AUC** | **BestIteration** | **Learning Rate** | **Training Error** | **Validation Error** |
| LL | 0.998 | 700 | 0.05 | 0.1525 | 0.1638 |
| PS | 0.9904 | 500 | 0.05 | 0.0803 | 0.0814 |
| TR | 0.998 | 700 | 0.05 | 0.3989 | 0.4022 |

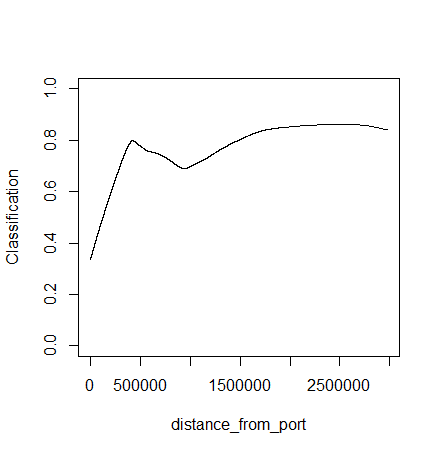
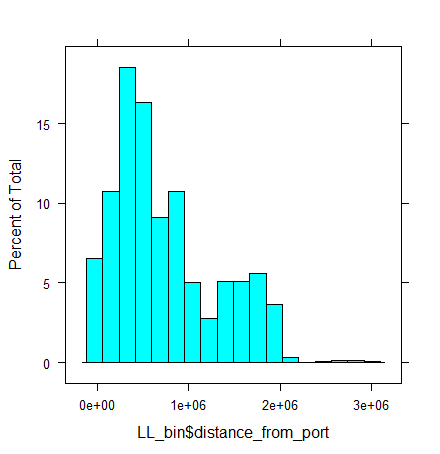
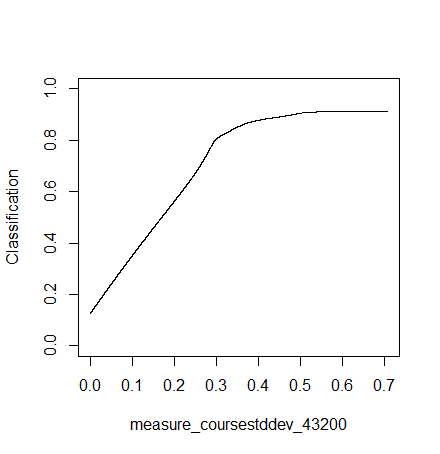
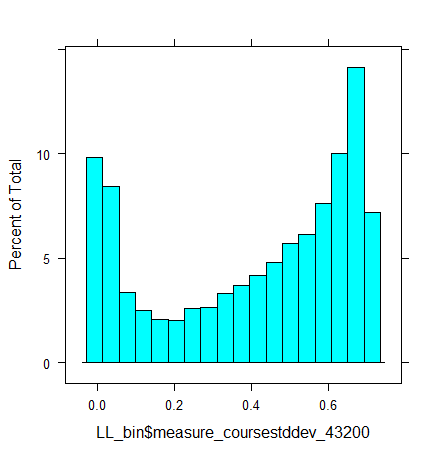
To make a usable model, we now must understand if these variables make intuitive sense for the business. The high AUC curves can also be a concern for overfitting.

The segmentation in longliner and trawler are favorable, but the purse seine seems to distribute the probability evenly, while the true data is highly favoring non-fishing.

From prior knowledge gathered in the analytics section of this paper, we know that vessels that are out on the water greater than four hours are significantly more likely to fish vs those under four hours. This validates that our most important attributes for segmenting fishing and not fishing comes for variables with rolling measure greater that 4 hours (i.e. 6 hours or greater).

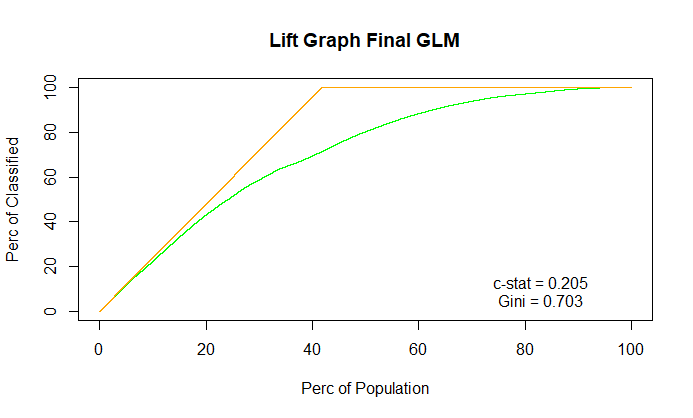
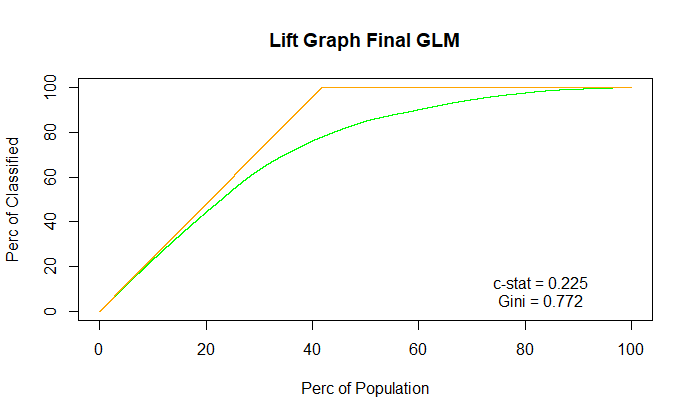
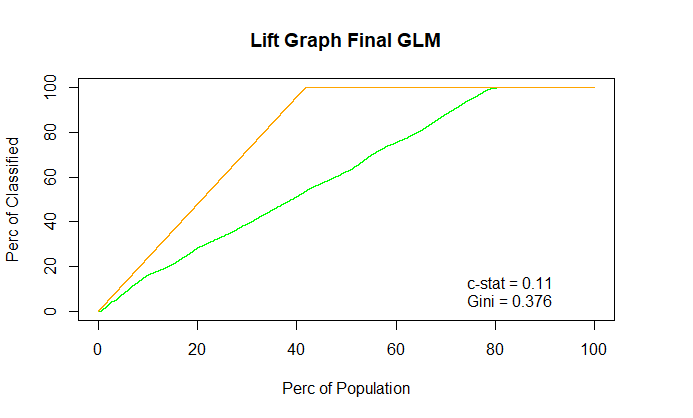
In the next section of code, we look at histograms of each variable and its correlation to the response variable. Again, we are looking for outliers that do not fit the business knowledge of the domain.



|  |
| --- |
| > sum(is.na(LL))/prod(dim(LL)) |
| [1] 4.091385e-06 |
| > sum(is.na(PS))/prod(dim(PS)) |
| [1] 1.040625e-06 |
| > sum(is.na(TR))/prod(dim(TR)) |
| [1] 4.173786e-05 |

In R, we look at the missing rate of the entire dataset we would use for training. Here we see that the missing rate is insignificant. This is great, because we do not need to create a solution for treating missing variables. However, we are not assuming that this dataset is a holistic representation of all fishing and non – fishing scenarios since we can still have incomplete data in the sense of human error.

The final model is build using the glm ‘logit’ model (as opposed to ‘Bernoulli’ in variable assessment). All three gear type models have the same top 9 variables that account for over 93% of variance in the dataset. The coefficients are different for each. This confirms our hypothesis that hat gear type is an important variable to determine the true fishing behavior of the vessel. We will add variables one by one to see how the lift chart improves. After we assess single variables in the model, we then add interactions of vessel flag plus the predictive variables to see if the performance improves.



The three images depict model performance with 1 variable, 7, and 30. This is based on the logic regression model that can be used for any vessel type.

Position Model.We have developed a position model that has been coded in python and uses the AIS data to determine how close the recorded points are to a point of interest. Points of interest include restricted fishing regions, ports and anchorage zones. Because the AIS data doesn’t provide a continuous data feed and there is a significant amount of time between data points, uniform distributions are used to predict the positions a vessel could be in between data points. This was done by using the midpoint between the intervals and associating an area of a circle as a uniform distribution. Uniform distributions are assumed due to the lack of information and uncertainty of a vessels path between points.

To calculate the area overlap, the points of interest are estimated by finding the center of the identified region and estimating the area of the circle to match size of the identified region. To calculate the area overlap between the vessel the point of interest, the equation to calculate different sized circles overlap, equation (1), was used. This prediction provides a percentage of how likely a vessel was to cross into a protected region. To achieve an overlap greater than 50%, the calculated midpoint must be within the radius of the protected area. Therefore, midpoint values which are outside the circle, yet yield an overlap greater than 30%, are flagged for consideration.

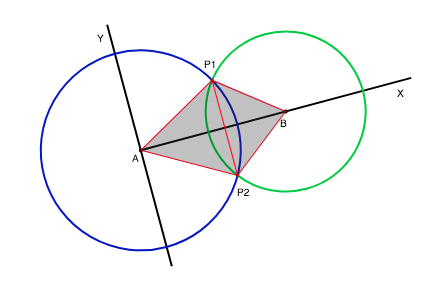


Figure 9. Area Overlap

Equation 1

Future Predictive Model**.** A Monte Carlo analysis varies the direction and speed to extrapolate possible future positions and calculates how many of the next positions are located within an identified region. This model is intended to provide a proactive component to the identification system, instead of being purely reactive. Because it is impossible to predict the decision making of the captain of various ships, uniform distributions are used to replicate possible course of actions. One thousand various paths are generated for each vessel point and the number of positions that are within a protected region returns the prediction of a vessel to move into a restricted region. One thousand replications were chosen to maintain a usable processing time of under one minute.

Anomaly Detection Model.

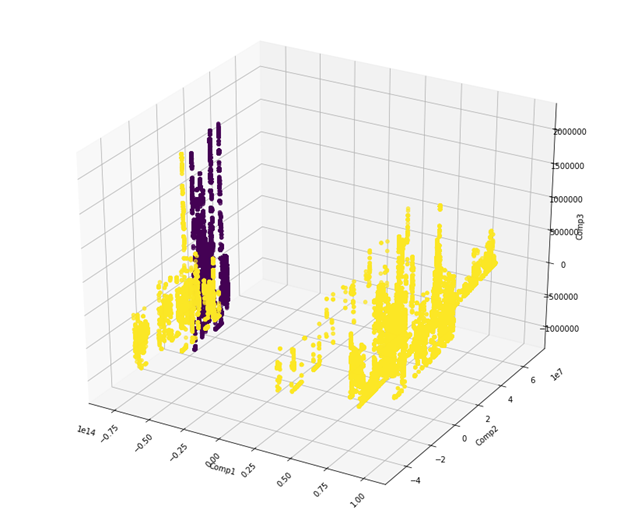


Figure 10 . Outlier Detection Scatter Plot

We have built the Outlier detection model to detect distinct data points. Outliers are observations that are at abnormal distances from the other values in the sample. It is important to characterize normal observations before abnormal observations can be noticed. There are some formal tests that test for single outliers and some that test for multiple outliers. Our model uses Tietjen-Moore test to detect various outliers in the dataset.

As our datasets contain numerous attributes we have performed PCA for dimensionality reduction. We reduced the number variables to 3 using PCA and used the 3 attributes for our outlier detection model. The Outlier detection model returns a Boolean array with True if points are outliers and False otherwise. Figure 10 shows the results of outlier detection model on a scatter plot. The data points plotted in Purple are the outliers detected by the model.

When we saved the outliers in a dataset we have noticed that it contained path of one single MMSI. After plotting (Figure 11) the points on the map we observed there are few points where vessel is classified as not fishing, but the activity of the vessel shows that it is fishing.. Also we have observed that the vessel has visited 2 different ports, which increases the likelihood of the vessel fishing illegally. The color scheme in Figure 15 is to show position over time, ranging from red to yellow to light blue to dark blue.

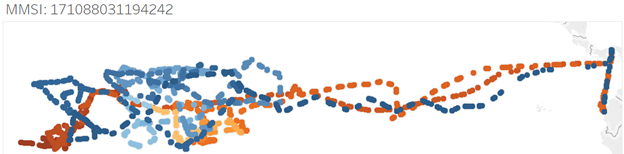


Figure 11. Path of the outlier vessel

Bayesian Net.

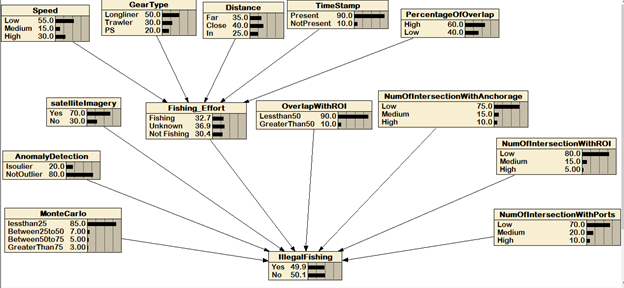


Figure 12. Netica Diagram of Illegal Fishing Rank

We have built a Netica diagram in Figure 11 to get a high-level overview of various attributes that feed into our model and to get the probability of Illegal fishing. As the diagram above shows that different attribute contributes towards the proportion of IUU. All these attributes are assigned certain weights depending on their effect on IUU. As fishing effort is a major contributing factor from all the factors that affect the IUU therefore, the probability of Fishing effort calculation depends on 5 major attributes (gear type, distance from port, speed, timestamp, and percentage of overlap).

The report generation is a summary of the ship that was just selected, and includes information such as: regression model results, prediction model, IUU prediction, and prediction. Below this report is an image dock, which shows real satellite images, of which are in the latitude/longitude bin where the selected ship is and/or was.

Image ExtractionUsing Planet Labs’ library of images, an API is utilized to feed coordinates and save an image to a storage device. The image retrieval is initiated when the likelihood of illegal fishing exceeds a rating of 0.6. A region containing the last known location of the vessel is provided and the API searches for an available image. If an image is found, the image is saved for the user to view. Future work would be to add a layer of the location of protected areas in red. This would provide the user context to the image, allowing them to internalize the image faster. This modified image is linked to the condensed report that is available to the user to inspect.

The report generation is a summary of the ship that is selected, and includes information such as: regression model results, position overlap prediction model, the Monte Carlo analysis, and the IUU prediction. Below the report is an image dock, which shows real satellite images, of which are in the latitude/longitude bin where the selected ship is and/or was

**Proof of Concept:**

To assess and verify the abilities of the system a sample dataset has been identified and used to test the performance of the system. A vessel was identified as fishing on Jan 4th, 2013 off of the coast of California. In a Planet Labs database a satellite image was identified to contain the longlines vessel. The vessel has recorded longitude and latitude points from the AIS transponder, and some of the potential vessel paths intersect with protected regions making this an optimal trial dataset. The system used the AIS information and matches the points to the linear regression to provides a likelihood of if the vessel is fishing or not fishing. Then the system creates additional data points by calculating the midpoints and determining a prediction of how likely the vessel went into the protected area. This is done by assigning the vessel a uniform distribution of the area of a circle, diameter of 2.5 miles, and assigning the protected area a circle with the respective size of the area. By calculating the overlap, it provides an estimate of how likely the vessel crossed into the restricted area. Because there were sections of the path that exceeded an overlap of 50%, then the model searched Planet Labs for an image within the respective timestamp, and location.

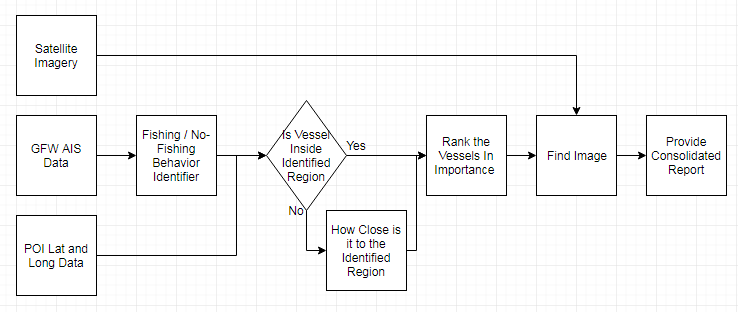
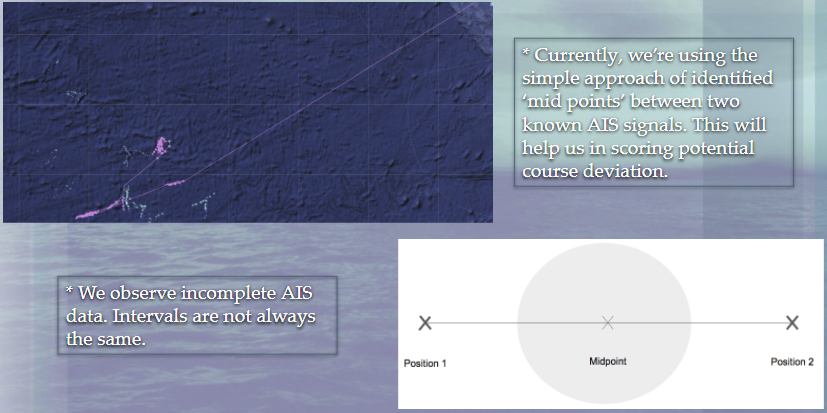
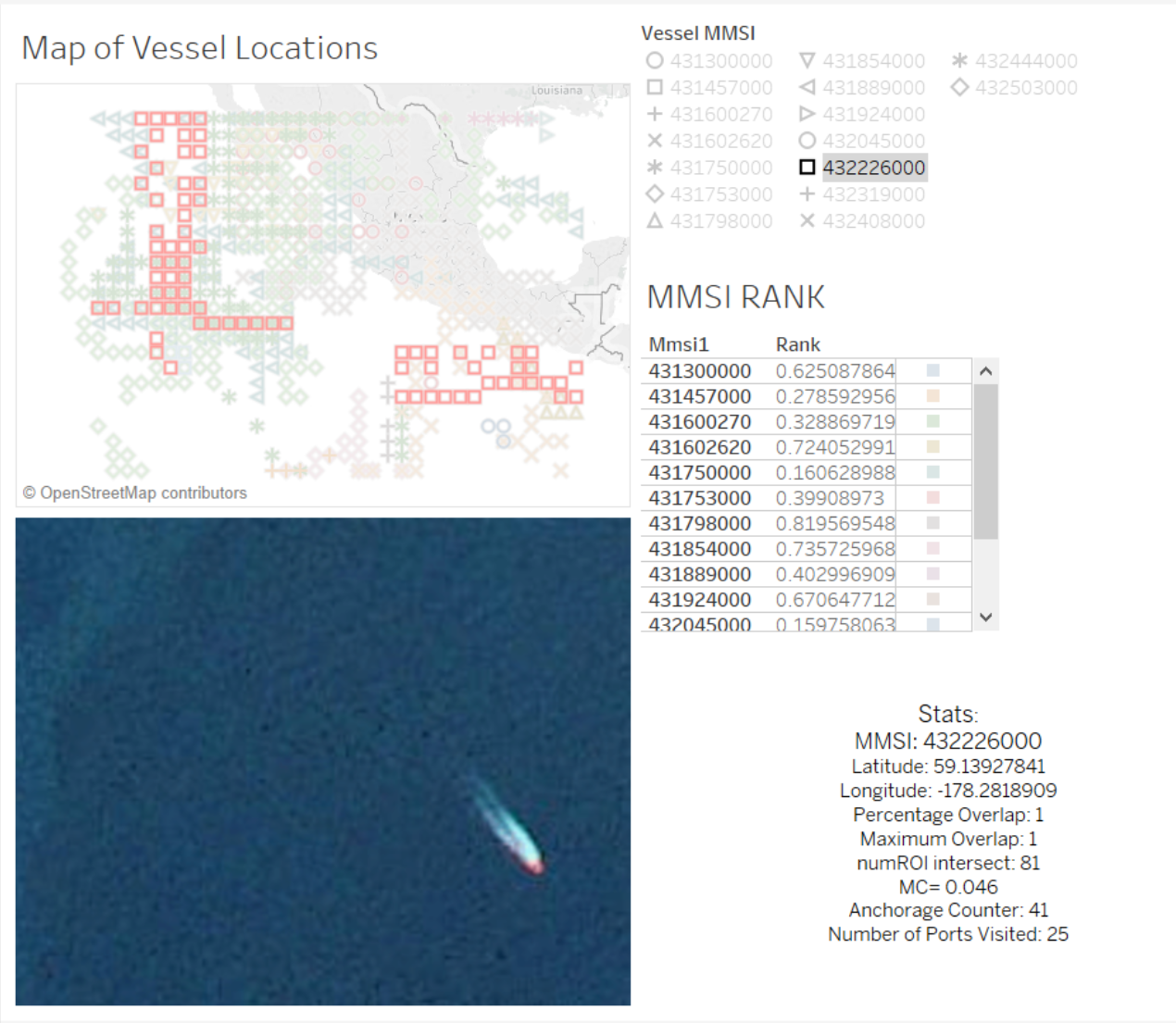


Figure 13: Diagram of Determination for a Singular Vessel

  
Figure 14: Easy-to-Understand Explanation of the Midpoint Logic Utilized with the AIS Data

***Visualization:***

Our finalized UI is in the form of a Tableau project. The best way to describe the setup would be to describe it left side and then the right side, with Figure 15 to show a convenient Screenshot.

  
Figure 15: Screen-Shot of the Tableau Visualization Dashboard

The left side of the user interface contains two parts: A map with color-coded density squares and an image. This example is near Mexican and Central American waters. The map is populated with data, and each vessel is a different color/shape representation that can be selected from the right-hand side of the map. Within the map, all of the latitude/longitude bins are visible, and all display at the North-Westernmost point of the respective bins. The map is a prime example of the possibility for not having perfect accuracy due to only having precision to 0.1°. The image is a displayed image that could possibly be the vessel in question (future teams could make this particular feature more robust).

The right side of the interface (top to bottom) has a ranking table, and a mini-report. For the region in question, it displays the top ten MMSIs and displays their Rank. The Rank is an indicator of likelihood of fishing illegally. If a person clicks on the MMSI, it will activate the results. These results are basic statistics about the vessel in question. For this instance, the vessel with MMSI 432226000 has an instance of overlap with a restricted area and a large number of ports visited, which both greatly increase the probability of having fished illegally.

## **Results:**

## Below mentioned are few of the important results observed from various analysis

## The figure below shows the data distribution for various gears. Most of the data obtained from GFW is supporting the longliner and trawler gears. Therefore, most of our models have been developed on these 2 gears, as we have high observations supporting training and test models.

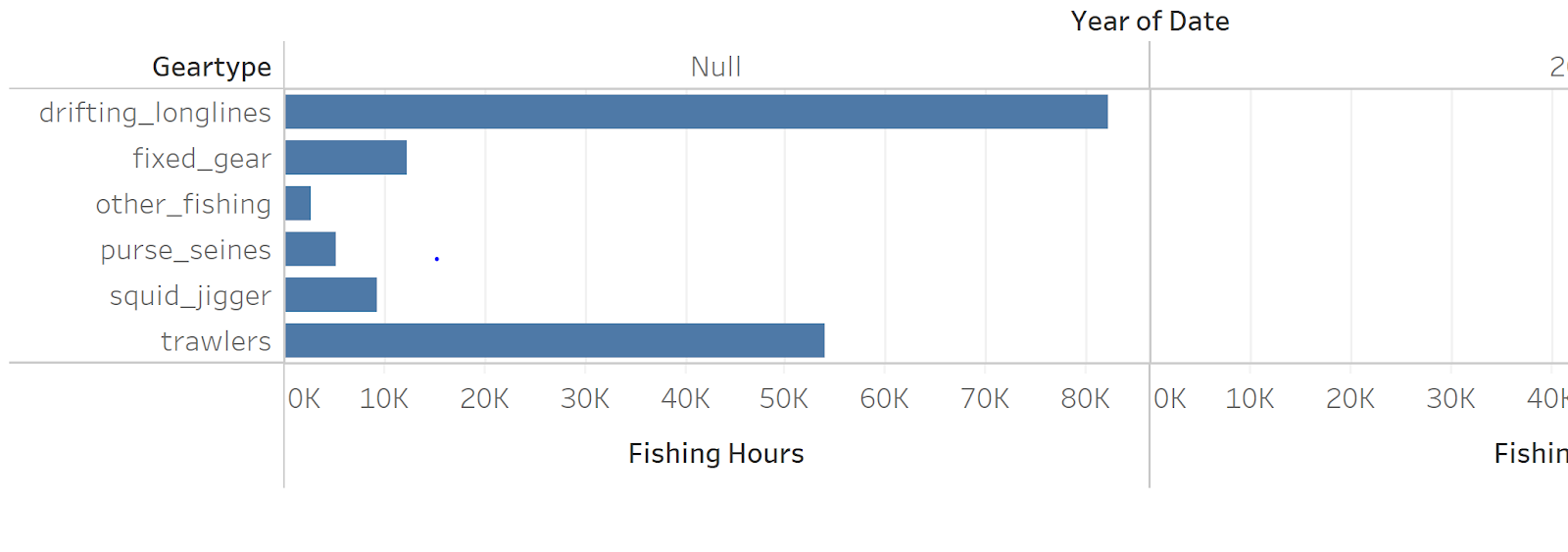


Figure 16: Dataset distribution for various gears

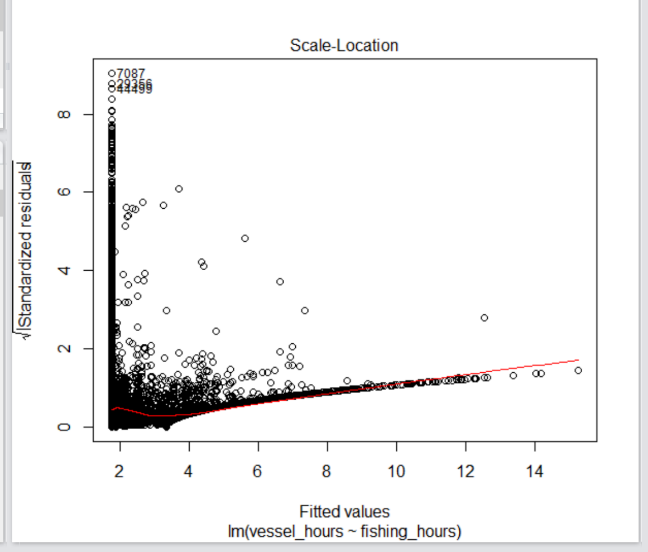
We have performed regression on the fishing hours dataset and the results show, if a vessel has been in waters for more than 4 hours then it is highly likely that the vessel is fishing. 

Figure 17: Regression graph vessel hours and fishing hours

### 

### After plotting the vessel activity on the map we have observed that majority of fishing Occurs in specific areas highlighted in the figure below.

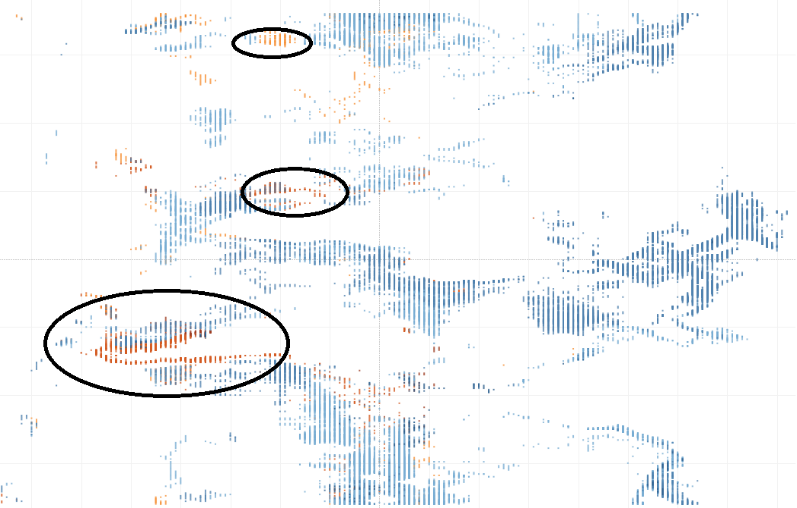


Figure 18: Majority Fishing Areas

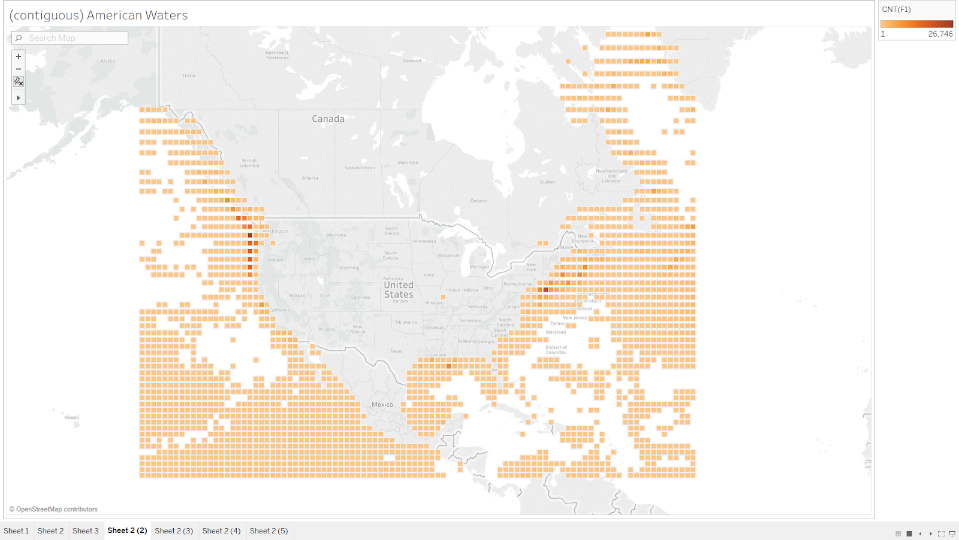


Figure 19: Majority Fishing Areas in USA

**Conclusion:**

The models used in the identification of illegal fishing are robust enough to handle various types of equipment. The logistic regression and Monte Carlo analysis adjust their functions to the vessel’s equipment type, allowing for a more accurate analysis versus a broad generalization of boats. Sacrificing some accuracy, the logistic regression model has a reduced likelihood of overfitting while maintaining a high level of accuracy of determining whether a vessel is fishing.

After, by extrapolating predicted positions from recorded locations, the position model is able to increase its capability and transform the model into a proactive analysis. Through the combination of a logistic regression, position model, bayesian net, and image extraction, an analysis on a vessel’s status is able to be predicted unlike any system currently available. This system does not replace the current methods of identification, but strengthens the analysis of the collected data to support the decision process of monitoring or apprehending a vessel.

## **Future Work:**

There are many issues with illegal, unregulated, and unreported fishing. What we have found is that there is not one solution to the problem, but there is rather a different approach or a ensemble of approaches best suited for this problem. Below we will describe a few ways the solution could be improved upon.

Incorporation of Light Source and RADAR Last year, the Indonesian government started to release their VMS signals allowing for GFW to add vessels by using light emitted from the vessel. This helps fill in the gap of vessels that are not registered for AIS or are trying to stay off the radar. A future team could use the VMS data available from GFW to develop tracks and assess if the vessels are fishing or not fishing. They should also add this information to the in/out models. Further, if radar information is available, the teams can incorporate this source to the DSS as well. This will enrich the models and make the IUU Fishing score a better representation of both bad and good cases.

Image AnalysisWhile this team attempted to breach the gap of image analysis, further study is required. The next team can pick up available image vessel data from Kaggle competition (over 5,000 images to train on). Many students have already showed impressive results in identifying vessels in images. The next team can use this to work from and further add not only if the vessel is in the image, but also the trajectory and speed. (https://www.kaggle.com/rhammell/ships-in-satellite-imagery)

Data Gathering An idea for a future team is to gather all IUU samples for known MMSI, then do Fuzzy Augmentation to create a match on training set of known IUU. Then, they could develop a model to train on IUU cases. The reason this is a good idea would be that a real model on IUU behavior would be developed. However, the problem with that plan would be issues and inability to ensure consistent data wrangling, collection, or accuracy.

In-House Changes While our group has succeeded in many areas, there is always room for improvement. For example: future teams could potentially obtain more data to finally be able to produce models for equipment types that are currently not represented. Also, more functionality could be added to the Tableau visualization or a future team could create a more streamlined process to get from analyzing the data to feeding it into the database.

## 

## **References:**

1. Agnew, David J., et al. “Estimating the Worldwide Extent of Illegal Fishing.” *PLOS ONE*, Public Library of Science, journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0004570.

2. Byrnes, Jarred, et al. “IUU Fishing Detection Final Report.” *IUU\_Fishing*, 2017, seor.vse.gmu.edu/~klaskey/Capstone/IUU\_Fishing/index.html.

3. Global Fishing Watch. [2018]. ​www.globalfishingwatch.org​

4. Jsoma. “Greek PM Sentiment Analysis before and after Signing the MoU with EU[Project] · Issue #100 · Jsoma/Data-Studio-Projects.” *GitHub*, github.com/jsoma/data-studio-projects/issues/100.

5. Planetlabs.com, planetlabs.com/.

6. Revkin, Andrew C. “How Digital Tracking of Rogue Fishing Can Safeguard Vast Ocean Reserves.” *The New York Times*, The New York Times, 15 Sept. 2016, dotearth.blogs.nytimes.com/2016/09/15/how-digital-tracking-of-rogue-fishing-can-safeguard-vast-ocean-reserves/.

7. Souza, Erico N. de, et al. “Correction: Improving Fishing Pattern Detection from Satellite AIS Using Data Mining and Machine Learning.” *PLOS ONE*, Public Library of Science, doi.org/10.1371/journal.pone.0163760.

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## **Biographies:**

### Anya

**Education:** Finishing last semester of the Master of Science in Data Analytics Engineering with a Predictive Analytics Concentration. Hold a Bachelor of Science in Criminal Justice from Texas State University.

**Work Experience:** Data and Decision Scientist at Fundation (FinTech). Work cooperatively on the design, development and testing of statistically grounded insights for various business needs including risk modeling. Prior work experience as a Data Analyst at a Non-Profit working on reporting, data processing, database migrations, and segmentation analysis.

**Tools:** Microsoft SQL Server, SQL language, R Programming, Excel (Advanced). Learning and eager to use on a project: Python, AWS.

**Interests:** Behavior and financial modeling, NLP, Image analysis.

***LinkedIn****:* https://www.linkedin.com/in/mityushina/

### Spandana

Is a graduate student, pursuing her Master’s in Data Analytics Engineering at George Mason University. Also, a Graduate Teaching Assistant with the AIT department under Dr. Ioulia Rytikova and with the SEOR department under Dr. Paulo costa and Kathy laskey. Holds a bachelor's degree in Information Technology from Osmania university in India.

**Tools**: MySQL, R, Weka and Tableau.

**Interests**: Predictive analytics, NPL, Sentiment analysis.

***LinkedIn****:* <https://www.linkedin.com/in/spandana-jagtap-4915a980/>

### Emma:

BS Science in Physics from Guilford College, with departmental honors, and a Mathematics minor. Undergraduate thesis: The Evolution of the Relationship Between the Temporal Decay Index and the Spectral Index of Gamma-Ray Bursts.

**Tools**: MatLab, Unix shell, LabView, various other physics-oriented tools, Python, R, SQL, Tableau, Spark and now ElasticSearch.

**Interests**: Numbers, learning new things, challenges, tea

### Abhishek:

Is pursuing his degree in Master’s Data Analytics Engineering Program. He is a TA to IT 213, IT 315, IT 431 courses of the Information Sciences and Technology Department of Volgenau School of Engineering. He hold a bachelor’s degree in Biomedical Engineering. Prior experience in molecular mand eLearning, fabricating microneedles for a Bill and Melinda Gates Foundation project, and developing e-learning products for Microsoft and Symantec and Symantec.

***LinkedIn****:* <https://www.linkedin.com/in/abhishekmishra11/>

Raymond Skaddan:

He graduated in 2017 from George Mason University with a Bachelor of Science in Systems Engineering and is also a recipient of the Carl M. Harris Memorial Scholarship. While completing his Masters of Science in Systems engineering, he has been the Teacher Assistant for the undergraduate capstone senior design course (Syst 490 & 495). Outside of school, he has worked at Zodiac Aerospace for the past four years where he has been able to supplement his degree with mechanical, electrical and software engineering. In July of 2018, he will begin working at Northrop Grumman in Melborne Florida to apply his Systems Engineering background to the aerospace industry.

**Appendix:**

### Definition of Terms

* AIS - Automatic  Identification System
* EEZs  - Exclusive Economic Zones
* GIS - Geographic Information Systems
* IUU - Illegal,  Unreported, and Unregulated
* MBSE  - Model Based Systems Engineering
* MPAs  - Marine Protected Area
* RFMOs  - Regional Fishery Management Organization
  + SPRFMO - South Pacific
  + CCAMLR - Antarctic
  + WCPFC - Western and Central Pacific
  + IATTC - Eastern Pacific
  + ICCAT - Atlantic
  + IOTC - Indian Ocean
  + CCSBT - South Atlantic, Indian Ocean and part of Pacific
  + CLAV  - combined tuna RFMO lists
  + FFA - Pacific Islands EEZ areas for member countries
  + SICA - Central America
  + CTMFA - Argentine-Uraguayan border area
  + NPFC - North Pacific
* SYSML  - Systems Modeling Language

### 

### 

### 

### 

### 

### Field DescriptionsThe key attributes in the datasets are

MMSI (Type- Integer) - Maritime Mobile Service Identities (MMSIs) are nine-digit numbers used by automatic identification systems (AIS) and certain other equipment to uniquely identify a ship or a coast radio station.

Flag (Type- String) – The attribute contains an iso3 value for the flag state of the vessel. Only for vessels that have been matched to registries or have known values.

Gear Type (Type- String) – This attribute holds the great type used by a particular vessel. There are 6 different types of gears that can be carried by different vessels, options include trawlers, purse\_seines, squid\_jigger, fixed\_gear, other\_fishing and drifting\_longlines

Fishing\_Hours (Type- Float) -The attribute contains the hours that vessels of a particular r gear type and flag were fishing in this grid cell on a particular day

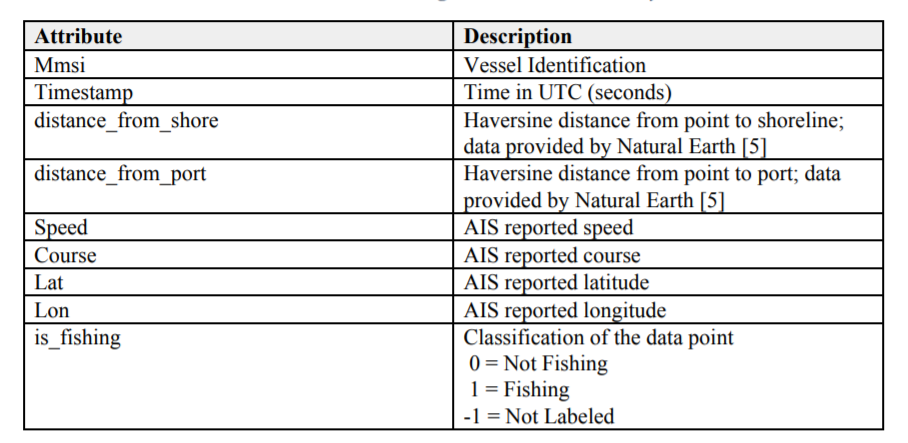
Vessel\_Hours (Type- Float)- The attribute contains the hours that vessels of a gear type and flag were present in this grid cell on a particular day.

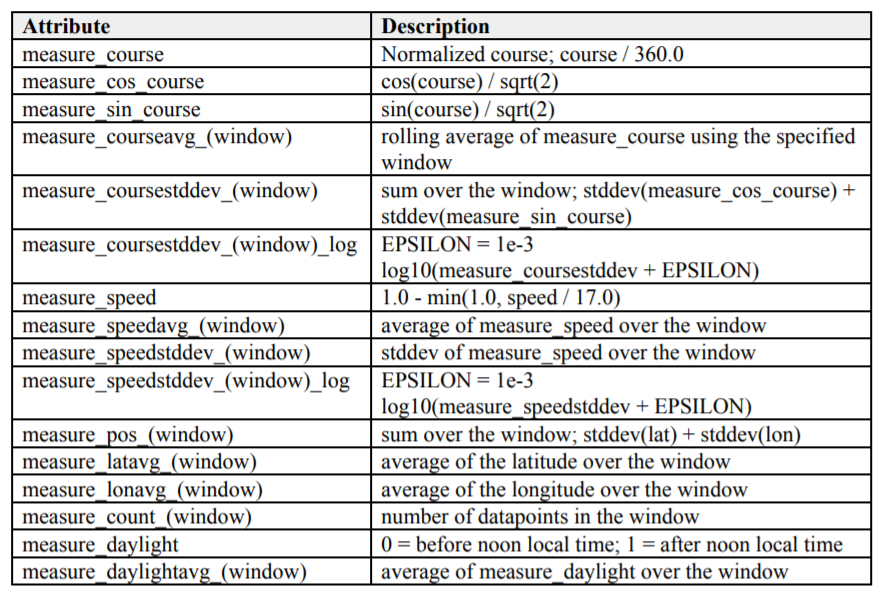
Lat\_Bin (Type- Float) – The attribute contains latitude information a vessel with particular MMSI. This helps in mapping the dataset to satellite imagery.

Lon\_Bin(Type- Float) - The attribute contains longitude information a vessel with particular MMSI. This helps in mapping the dataset to satellite imagery.

### 

### Data Dictionary:





### 

### Assumptions:

* If end user will have access to ‘on demand’ imagery.
* Focus on California Coast.
* Not all vessels will turn off their AIS transmitter while fishing illegally.
* Previously used libraries are valid and ran the statistical analysis accurately.
* Prior models were well tested and properly identifies vessels of interest
* Data is still relevant and there is no population change from 2015 to now (data used was for 2011-2015 cases)

## Data Repository:

* Our Work: <https://github.com/AnyaMit/IUU-Fishing-Detection-DAEN-690-Capstone>
* Prior Team Work: <http://seor.vse.gmu.edu/~klaskey/Capstone/IUU_Fishing/index.html>

Variables:

X

+ measure\_coursestddev\_1800\_log

+ measure\_courseavg\_43200

+ course

+ measure\_sin\_course

+ measure\_daylightavg\_900

+ measure\_speedstddev\_10800

+ speed

+ measure\_pos\_86400

+ measure\_daylightavg\_43200

+ measure\_latavg\_86400

+ distance\_from\_port

+ measure\_lonavg\_3600

+ measure\_lonavg\_1800

+ measure\_courseavg\_10800

+ distance\_from\_shore

+ measure\_courseavg\_1800

+ measure\_speedstddev\_86400

+ timestamp

+ measure\_coursestddev\_86400

+ measure\_speedstddev\_21600

+ measure\_daylightavg\_3600

+ measure\_courseavg\_86400

+ measure\_speedstddev\_3600\_log

+ measure\_coursestddev\_900

+ measure\_coursestddev\_10800

+ measure\_coursestddev\_43200

+ measure\_daylightavg\_1800

+ measure\_coursestddev\_3600\_log

+ measure\_lonavg\_900

+ measure\_distance\_from\_port

+ measure\_coursestddev\_10800\_log

+ measure\_coursestddev\_900\_log

+ measure\_latavg\_21600

+ mmsi

+ measure\_course

+ measure\_lonavg\_21600

+ measure\_daylight

+ measure\_pos\_10800

+ measure\_count\_10800

+ measure\_coursestddev\_1800

+ measure\_speedstddev\_43200

+ measure\_daylightavg\_21600

+ measure\_courseavg\_21600

+ measure\_speedavg\_86400

+ measure\_pos\_900

+ classification

+ measure\_coursestddev\_21600

+ measure\_count\_1800

+ measure\_speed

+ measure\_latavg\_43200

+ measure\_latavg\_10800

+ measure\_speedavg\_1800

+ measure\_count\_3600

+ lon

+ measure\_speedavg\_10800

+ measure\_speedavg\_900

+ measure\_daylightavg\_10800

+ measure\_count\_21600

+ measure\_courseavg\_900

+ measure\_speedstddev\_3600

+ measure\_pos\_43200

+ measure\_latavg\_3600

+ measure\_coursestddev\_21600\_log

+ measure\_speedavg\_43200

+ measure\_courseavg\_3600

+ measure\_latavg\_900

+ measure\_daylightavg\_86400

+ measure\_count\_43200

+ measure\_coursestddev\_3600

+ measure\_lonavg\_86400

+ measure\_lonavg\_43200

+ measure\_cos\_course

+ measure\_pos\_3600

+ measure\_speedstddev\_900

+ measure\_speedstddev\_21600\_log

+ measure\_count\_86400

+ measure\_pos\_1800

+ measure\_coursestddev\_86400\_log

+ measure\_speedavg\_3600

+ measure\_speedstddev\_1800

+ measure\_coursestddev\_43200\_log

+ measure\_latavg\_1800

+ measure\_speedavg\_21600

+ measure\_count\_900

+ measure\_speedstddev\_86400\_log

+ measure\_lonavg\_10800

+ measure\_speedstddev\_10800\_log

+ lat

+ measure\_speedstddev\_43200\_log

+ measure\_pos\_21600

+ measure\_speedstddev\_900\_log

+ measure\_speedstddev\_1800\_log

+ geartype

+ geartype\_code

+ pred01

+ pred02

+ geartype\_code\_flag\_LL

+ geartype\_code\_flag\_PS

+ geartype\_code\_flag\_TR

+ pred\_final\_2

+ pred\_final

Final Model Formula:

