Data-Driven Framework for Prediction of Fish Catches: A Norwegian Case Study

Mrinal Yadav1, Ekaterina Kim2, Bjørnar Brende Smestad2,

Bjørn Egil Asbjørnslett2, Anirban Bhattacharyya1

Department of Ocean Engineering and Naval Architecture, Indian Institute of Technology Kharagpur Department of Marine Technology*,* Norwegian University of Science and Technology

## *Abstract*— The fishing industry is identified as an important sector accounting for 4.6% of the total Norwegian Export value. Global changes in climatic variables have impacted and continue to impact marine fish and aquaculture production, where machine learning (ML) methods are yet to be extensively used to study aquatic systems in Norway. The method proposed in this document aims to find, combine and explore relevant fishing activities data with focus on activities in Norway and develop data-specific tools for visualization, observing accessing, forecasting and managing fisheries. In this paper we explore a spatio-temporal dataset that is a combination of AIS data (i.e., information on trajectories of fishing vessels) and the corresponding fish catch (i.e., the quantity and type of fish caught). The overall data describes the fishing activities over The Norwegian Sea and The North Sea for the past two decades. The problem we try to solve is the prediction of catch given some underlying conditions. Since it’s a prediction task hence we use tree-based regression models like Random Forest, XGBoost and LightGBM. Here we have made our own custom objective and custom evaluation function for LightGBM algorithm. The approach is to model for specific vessel groups, geographical locations and species. The study explores the relationship between physical parameters of the vessels for each year and uses that relation for further analysis. The study reflects the dependence of catch on physical parameters of vessels like length, gross tonnage and power as well as the impact of geographic locations (latitude and longitude), species of fish targeted, tools (gears) with which fishing is done and also the time (month) in which the fishing is to be done. We have also included a feature of product condition (code) in our analysis since the total catch is the total weight of fish being delivered hence it is important to have the information beforehand about how the delivery of the fish caught is planned. For example, you can catch the same number of fish but deliver it in frozen format (ice) so in this case the total weight will also include the weight of ice, so one will get a larger weight but the actual fish weight could be the same. In this paper for the purpose of specific analysis we have considered vessels with length less than 10m (which constitute 30.2% of all vessels) and 4 different fish species codes for cod species (‘1022 - Cod’, ‘102201 – Norwegian Cod’, ‘102202- Northeast Arctic Cod’, ‘102204 – Other Cod’). Also, we have divided the model into southern and northern part based on latitudes of fishing locations in a sense that we have divided the range of latitudes into two halves, the upper half is the northern model whereas the lower half is the southern model. However, one can easily modify the codes for any different length group and different species of fishes to predict the catch they can obtain with certain error bar. The preliminary results demonstrate that results are good for southern locations and northern region still needs improvement. With the proposed approach we were able to achieve mean absolute error 47.6 Kg in average on a test set. Our predictive results are preliminary in both temporal data horizon that we are able to explore and in the limited set of learning techniques that are employed in this task. However, it does not explore fluctuations in catch caused by environmental variation or any political interference. In the rapidly warming region, it is of vital importance to understand how stocks may be further affected by climate change in addition to fishing pressure.

## (Can we add this line)-

## It is likely that other centers of intense fishing activities are in possession of similar data and could use the methods similar to the ones proposed here in their local context.

# **Introduction**

Fish catch has two main importance that are food security needs and generation of export revenue. Fish catch plays a vital part in meeting the food security needs of Norway, where per capita fish consumption was 19.5kg in 2020 [[1](mailto:https://www.statista.com/statistics/643484/per-capita-consumption-of-fish-and-fish-products-in-norway/)]. Then there is the revenue that is generated from the export of fish and seafood. In 2019 the export value of seafood from Norwegian aquaculture amounted to roughly 76.5 billion Norwegian kroner (8.4 billion USD). Seafood from Norwegian fisheries was exported with a value of approximately 30.8 billion Norwegian kroner (3.4 billion USD) [[2](mailto:https://www.statista.com/statistics/665969/export-value-of-seafood-from-fisheries-and-aquaculture-in-norway/)]. Populations of fish are greatly influenced by several physical, geographical and temporal factors. The prediction of fish catch will be important to the fisheries management as well as the fishers. With the help of this study, we can predict the quantity of fish catch given the underlying conditions that will be presented by the fishers in form of location they want to fish, the vessel they will be using for fishing, the time of year (month) they will be fishing and the gears and tools they would be using for that purpose. Fish catch is not just dependent upon the physical and geographical characteristics but also on climatic and environmental variables such as sea-level rise, sea surface temperature variation, humidity variance, etc. However, in this paper we have not incorporated those climatic factors. Earlier approaches have been made to quantify the relationship between climatic variables and the fish catch [[3](mailto:https://ieeexplore.ieee.org/document/1519023)]. Several methods, i.e., the autoregressive integrated moving average (ARIMA), seasonal ARIMA, vector autoregression (VAR), neural network, nonlinear autoregressive (NARX), wavelet, etc., have become popular among researchers for predicting short-term fish landings. Anuja et al. showed the forecasting of marine fish production in Tamil Nadu using the ARIMA model [[4](mailto:https://journals.ansfoundation.org/index.php/jans/article/view/1252)]. However, these models are usually fitted with single time-series data, therefore often produce unsatisfactory predictions when multi-dimensional data is fetched as inputs. To overcome this problem, researchers came up with the idea of using machine learning (ML) methods [[5](mailto:https://content.iospress.com/articles/journal-of-ambient-intelligence-and-smart-environments/ais210604),[6](mailto:https://ndpublisher.in/admin/issues/EAv63n4e.pdf)]. Although several types of research have been done so far on estimating fish landings considering ecological variables [[7](mailto:https://cdnsciencepub.com/doi/abs/10.1139/f82-036?journalCode=cjfas),[8](mailto:https://www.sciencedirect.com/science/article/abs/pii/S0034425710000295?via%3Dihub)].

**(Can we add this line)- However, in Norway, as far as we are aware, no one has implemented the ML-based method to predict and analyze fish landings of the coastal area.**

A single ML model does not perform best in different time-series predictions. In this case, a good choice would be to use ensemble-based ML models for enhancing the prediction accuracy. In this research we have used Tree based method Random Forest and gradient boosting approach in which we have used XGBoost and LightGBM algorithm for the prediction of catch. The main research questions that guide this work is: *How accurately can we predict the catch for a particular trip given the underlying variables?* As a case study we have selected the region of the North Sea and a specific vessel type (length less than 10m) and specific species (cod), The reason for taking cod species is that in history of Norway it has constituted a substantial part of fish stock being the most valued species at NOK 3.7 billion, while herring was worth NOK 2.2 billion back in 2007 [[9](mailto:https://www.ssb.no/historisk-statistikk/emner/jord-skog-jakt-og-fiske/_/attachment/inline/1c30f10a-d015-4db3-8894-9543a3ea3d5f:229476c7ae06e7f9ae32b1a9a510bc5875404914/nos_d428_en.pdf)] . However, the presented approach can be applied to other species and vessels. The hypothesis is that we can predict the catch at a given location and time of the year, provided the vessel length, gears and tools to use and the information on targeted species. To address the research question, we analyzed and enhanced AIS data for the years 2001 - 2021 and have explored the predictive capabilities of two popular supervised machine learning models (Random Forest, XGBoost and LightGBM). Being able to predict fish catch for a trip, is important for the planning of trip. The contributions of this work are the proposal of a framework that: (i) cleans and preprocess the data by analyzing the year wise trends to fill in the missing data and removing the irrelevant and wrong (illogical) entries; and (ii) applies Machine Learning (Random Forest, XGBoost and LightGBM) to predict catch with certain error.

This paper is structured as follows: We first briefly present the AIS data along with the methods used to transform data into a state suitable to train and validate the different machine learning models. Next, we analyze and compare predictions from three supervised machine learning models, i.e., the Random Forest model, the XGBoost model, and LightGBM (custom) for test dataset followed by a brief discussion about the capabilities and limitations of the models and modelling techniques. One of the best model implementations (model) is provided on GitHub (link), and the presented approach could be used to support planning and active management of fishing vessels in the North Sea region.

# Proposed Framework

I will take care of this part

# Case Study

To demonstrate the feasibility of the proposed framework a case study is used.

# **Material and Methods**

## Data

The catch-and-vessel data (AIS) used for the analysis are retrieved from Directorate of Fisheries, Norway. The data available here contains the catch data from the landing and final banknote register linked to vessel data from the marketing register. Catch data and vessel data are combined using the last capture date and the vessel's validity ; the last catch date must be in a period when the vessel is valid in order to be cooped against the vessel. The Directorate of Fisheries’ practice has been to consider this data to be public. The time-series data used here is for the past two decades (2001-2021). The dataset contained information about all fishing and trapping in the sea taken by Norwegian-registered fishing vessels as well as the capture of foreign vessels landed in Noreg. Each year has a different file and each file contains data for a catch year, data for a catch year equal to the current year are updated every night while data for previous two years are updated every weekend. The data provided about 1.2 million entries in each file which corresponded to different fishing trips, however the number of unique fishing vessels was about 6000 for each year. A total of 133(non-unique) features were provided for each entry which described a particular fishing trip and the corresponding fishing vessel. To describe the fishing trip information such as reception station, landing municipality, production municipality, etc were provided. For describing the vessel information such as vessel name, vessel id, vessel country, vessel type along with the information on physical characteristics of the vessel like vessel length, gross tonnage and engine power were provided. Data on gears and tools used for fishing was provided in terms of gear group, gear sub-group and gear codes. Information on geographical locations of fishing were provided in terms of latitudes, longitudes and location codes. Data on species of fishes was provided in terms of species group, species-subgroup and finally species code. Finally, the information on product condition and catch value was also provided.

*Variables:*

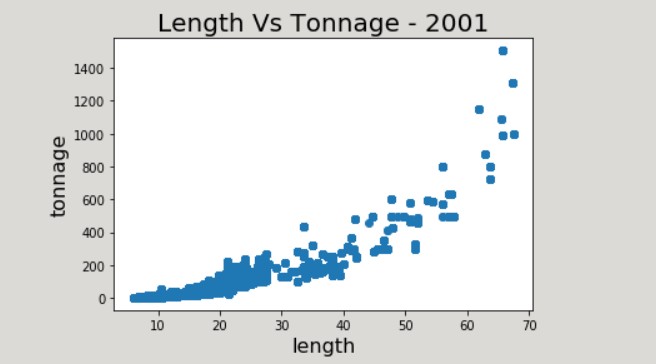
The data file for each year had a number of features describing the trip and the fishing vessel, however we have only included physical, geographical and temporal variables for the purpose of the research. For the physical features we have considered the vessel length, gross tonnage, and engine power of the vessel. The type of gear/tool (in terms of codes) used for fishing. For considering the geographical aspect we have used latitude and longitudes of fishing areas. We have also considered the target species of fish and the product condition. The target variable for our model will be the catch value. The AIS data was cleaned and enhanced before modelling as described in the following section.

**Data Pre-processing:**

The objective of the preprocessing is to produce a clean dataset that preserve the original information but also includes additional information from the input variables like latitudes and longitudes.

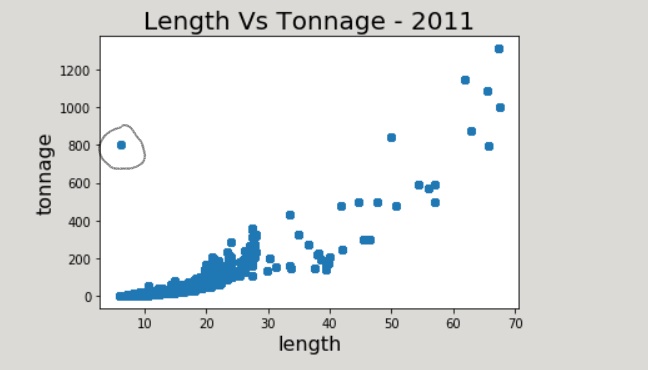
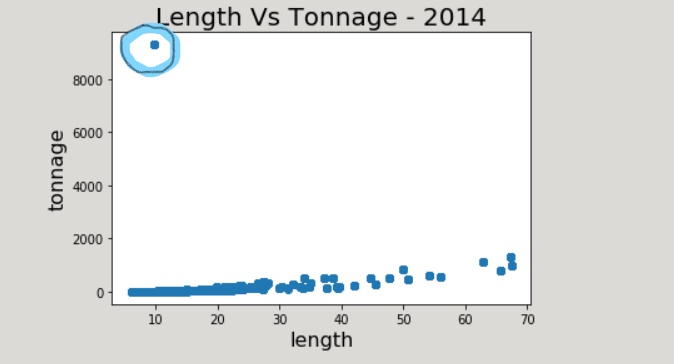
The first step is that we have encoded the species information in terms of species group code, species sub-group code and species code in string format to preserve their identity. We have checked for illogical entries where engine power and gross tonnage had zero values that did not make any sense, so we removed them from our data. For gross tonnage two columns were provided, gross tonnage – other and gross tonnage – 1969, for each file either of the one column had greater than 90% Nan values, so we considered the alternative gross tonnage in that case. Also, we checked if there is a zero value in one of the gross tonnage columns and a non-zero value in other, we have used that value for further analysis. Then we corrected the format of length column by encoding it to integer value which was initially a string. The catch date column was also a string type so converted it date-time format for further analysis. Similarly, we converted latitude, longitude and catch into integer formats that were initially strings.

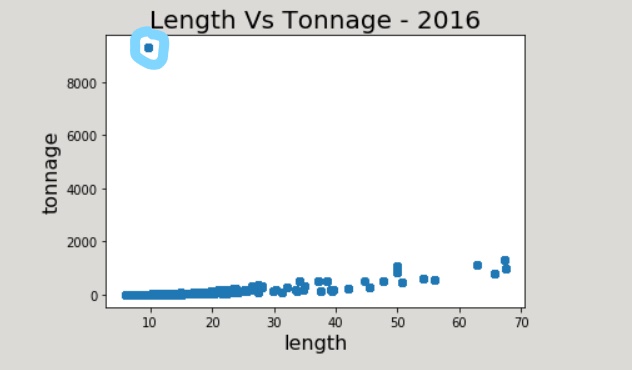
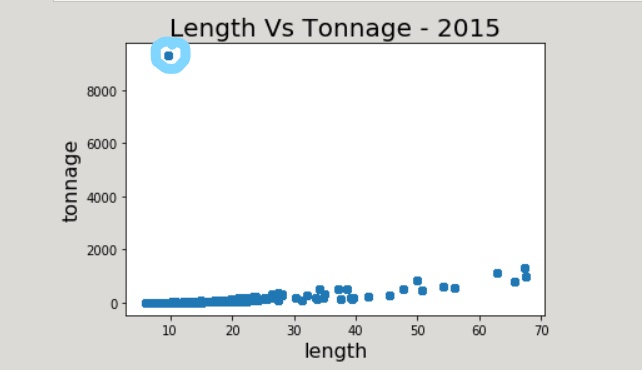
Now we handle the Nan values in important columns we observe that for each year the number of nan values in length and engine power column are exactly the same which constitutes about 7% of total entries for each year. Since 7% is not a high value hence we remove the entries where length has nan values, the engine power is also taken care of with this step. In some years the number of missing values for length and power are not exactly the same but overlapping, so after handling length column if some missing entries were still left in power column they were explicitly removed. We also check for nan values in latitude and longitude columns and observe that for each year the number of missing values is same in both the columns and is approximately 4% of the total data, since we cannot impute latitude and longitude and also 4% is not very high so we remove those entries. However, the number of Nan values in tonnage column is quite high (approximately 17%) and we cannot afford to lose on this much amount of data so we impute the tonnage column to fill in the missing entries. For filling in the missing values of tonnage we have used length variable , we have used the log-linear relationship between length and tonnage [[10](mailto:file:///C:/Users/Mrinal/Downloads/ISSF-2013-08-Relationship-between-Gross-Tonnage-and-Overall-Length-for-Vessels-on-the-ICCAT-Record%20(2).pdf)] to impute the tonnage values for each year. First, we plot length vs tonnage values for each year and observe a linear logarithmic relation between them.



**Figure 1**: A typical relationship between the length and the gross tonnage for vessels delivering catch in 2001

This is the general trend followed by the vessels in the remaining years except for 2011, 2014, 2015 and 2016 where we encountered anomalies that were due to wrong entries of tonnage values.



**Figure 2**: Abnormal entries of tonnage for the year 2011, 2014, 2015 and 2016.

These anomalies were due to wrong entry of tonnage value. For this we checked the vessel name to see if there was any other entry with correct tonnage value but were unable to find any, also the number of abnormal entries was very less so we decided to remove them. After removal of those entries, the general trend was followed for all the years. Then we have filled the missing values in tonnage column using the log-linear relation: **log(tonnage) = log(length)\*a + b**

For each of the 21 years the coefficients (a, b) will have different values that will fit into the data of respective years and to check the goodness of fit we will have the R-square score as the measure.

We used MATLAB for curve fitting with the available length-tonnage values to get an optimal log-linear relation with a high R-square score as evident in Figure 4 , which we can use to impute values of tonnage. For each year we obtain an optimal pair of coefficients (a,b) and the corresponding R-square score.

**Figure 3**: (a, b) coefficient) values vs year

**Figure 4**: R-square values vs year

We also check for abnormal entries of data in our target variable (product weight) column and found out that about 1 million (in total) of them were zero in spite of having all other information in dependent variables which does not makes sense and is probably error in data. Since much can’t be done hence these entries were removed from data. We also add a year and a month column to our data with the help of the date column, since month and year are broader terms and more generalizable (or representable). After getting the data complete, we combine the data for all the years and then perform elementary analysis and clustering.

**Elementary Analysis and Clustering:**

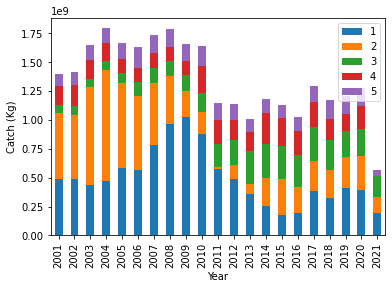
We have performed elementary analysis on data (complete with cleaning) and then did clustering of complete data this is just to get an idea of what the data clusters might look like with respect to parameters like vessel length and the species of fish.

Since the data is cleaned to a certain level, we perform cluster analysis where we cluster the vessels based on the characteristic that might affect the catch. For purpose of analysis, we will be considering top 5 species (180 total species) with respect to catch.

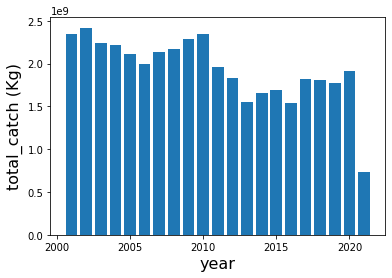


**Figure 5**: Species by catch over 2001-2021

**(Should I write something about this fig or its self-explanatory?)**



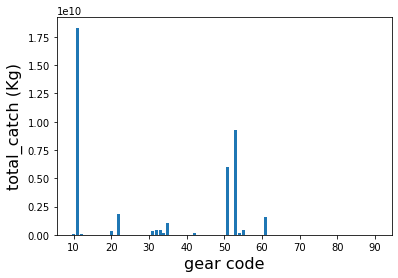
**Figure 6**: Total catch over the years for species 1: Norwegian spring-spawning herring; 2: Blue whiting; 3: Northeast Arctic cod; 4: Mackerel and 5: Pollock



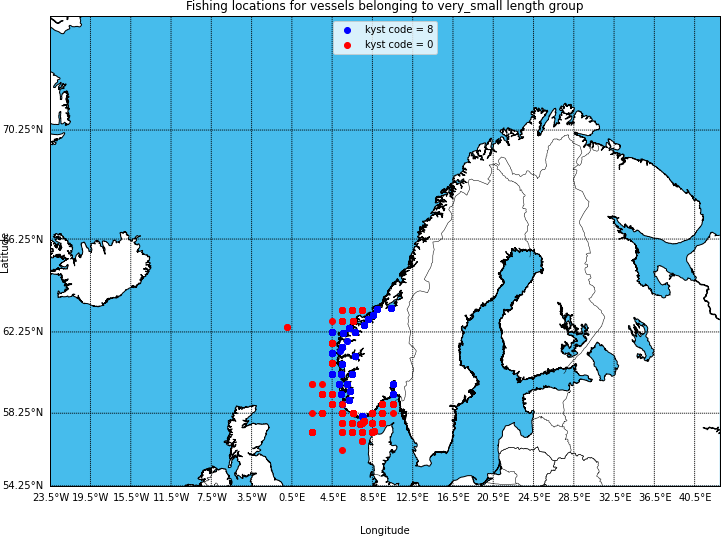
**Figure 7**: Variation of total catch over the years (all species)

The general trend is decrease in catch over the years – **(Reasons)**

Moreover, we can infer that certain gears/tools have been excessively used for fishing and hence have caught more fishes than other tools.



**Figure 8**: Gear distribution with respect to total catch



**Figure 9**: Representing fishing locations within a radius of 20 nautical mile with blue and that outside the radius with red.(south-data)

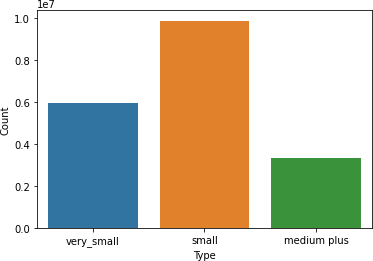
Figure 8 shows that the gears 11, 51 and 53 have caught the most fish over the past years. They correspond to the following gears – 11: Purse seine; 53: Floating trawl and 51: Bottom trawl.

In our analysis there are 239 unique fishing locations (cleaned data) out of which 151 locations are in the north model and remaining 88 locations are there in the south model and here in this paper we focus on the south model. Figure 9 shows the fishing location for south model. As we move away from the coast the sea depth will start to increase making the water waves to be high and more dangerous for fishing.

**Clustering:**

We are analyzing the top 5 species by catch that is shown in Figure 5.

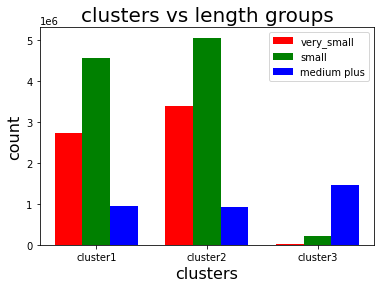
For segregation of vessels into different groups we consider the vessel length to be the deciding factor as the other physical parameters like gross tonnage and power are directly related to it. We divide the vessels into three groups namely very small (vessels length less than 10m), small (vessel length between 10-15m) and medium plus (vessel length greater than 15m).



**Figure 10**: Vessel groups with respect to vessel length

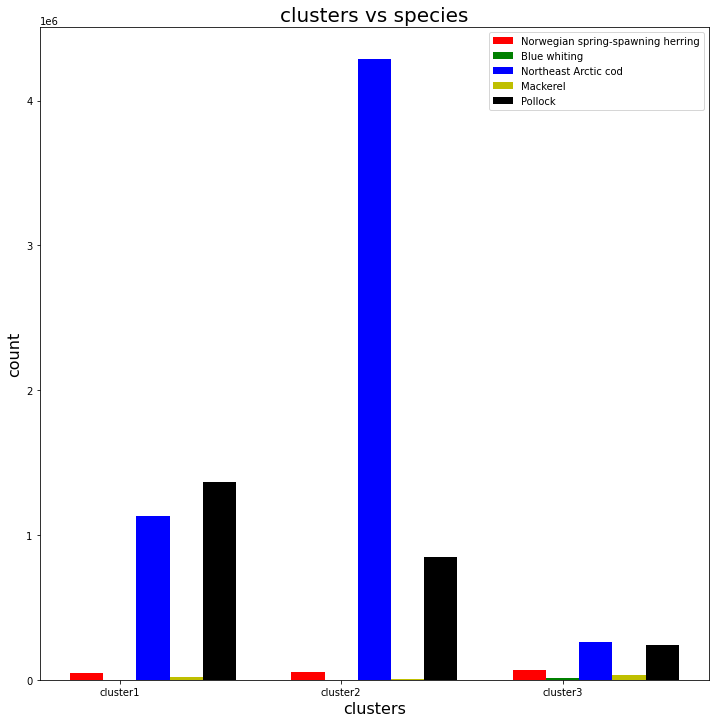
We clustered our data using K-means with pytorch and obtain three as the optimal number of clusters using elbow method. The clusters represent following composition for vessel groups.

In terms of Length range:



**Figure 11:** Results of length groups within each cluster

In terms of species composition:



**Figure 12**: Results of species distribution within each cluster

Figure 10 represents the cluster composition with respect to various length group of vessels. We can infer that cluster 1 and cluster 2 have similar kind of composition of different vessel groups. These groups of vessels have majority of small and very small vessels. Whereas the cluster 3 represents a group of

vessels that are bigger in size.

Figure 11 represents the cluster composition with respect to various length group of vessels. We can observe that cluster 3 fleets caught a similar number of fishes of each kind whereas cluster 2 fleets specifically targeted the Norwegian spring-spawning herring species and cluster 3 fleets caught Norwegian spring-spawning herring and pollock more than they caught other species. **The number of fish catch for each cluster(fleet) is directly related to the number of vessels in that fleet**. Since cluster 3 had the least number of vessels, hence it has the lowest total catch, cluster 2 being the most populated one resulted in highest total catch and cluster 1 had intermediate amount of total catch corresponding to its number of vessels.

Figure 12 – information/inference

**Some more data preprocessing: Preparing data for input in model**

We form our AIS data frame by considering the dependent variables like length, gear code, species code, coast/ocean code, latitude and longitude, product condition code, month and our target variable that is product weight. In this paper we have analyzed specific group of vessels (length lesser than 10m) and species (specific codes for cod fish). We created two more column namely gear code frequency and product condition code frequency which indicated the number of times a particular gear/tool was used and the number of times the product had a particular condition respectively. We have removed those entries where the frequency of gear code and product condition code was less than 1000 to handle outliers. We observe that the catch is a very dispersed value. We calculated the number of times the catch was in a particular range and got the following results.

|  |  |
| --- | --- |
| Catch range (Kg) | Frequency of occurrence |
| x <= 10 | 525839 |
| x <= 2000 | 2359026 |
| 2000< x <= 4000 | 5300 |
| 4000< x <= 6000 | 296 |
| 6000< x <= 8000 | 20 |
| 8000< x <= 10000 | 3 |
| 10000< x <= 12000 | 2 |

**(Should I put a bar graph for this information?)**  
Table 1: Distribution of Product weight

as we see that the distribution is quite dispersed, making certain ranges act like a outlier(noise). Hence, we just consider the catch from 10kg to 4000kg.

Now we divide our data into 2 parts based on longitudes and latitudes, we call one the south part and the other one the north part. In this paper we discuss the south part of data. We take the range of latitudes and arithmetically divide it into south and north. This south part is essentially the fishing locations in the North Sea. Figure 9 shows the fishing locations in the south data part.

Analyzing target variable: product weight

The target variable has high variance as evident from the Figure 13.

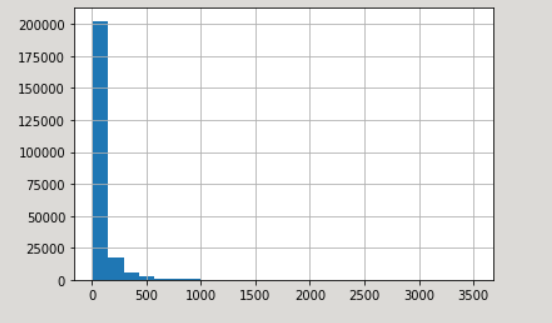


Figure 13: Distribution of product weight (before transform)

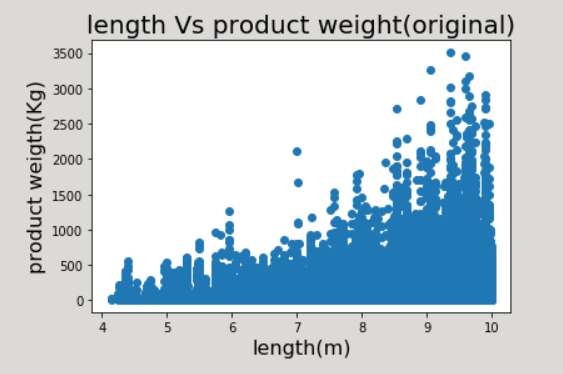


Figure 14: Length of vessel vs product weight (before transform)

While decision trees have a natural resistance to outliers, boosted trees are susceptible, since new trees are built off the residual. Hence, we apply a log transform to our target to have good predictions from the model.

Transform of target variable:

**Product weight 🡪 log (product weight)/length**

After transform:

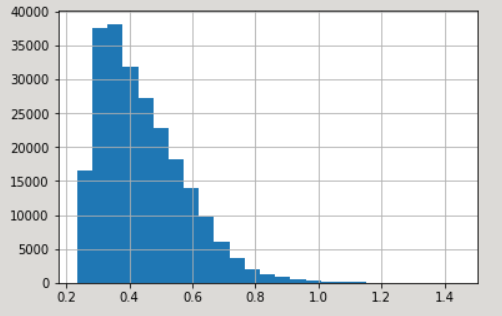


Figure 15: Distribution of product weight (after transform)

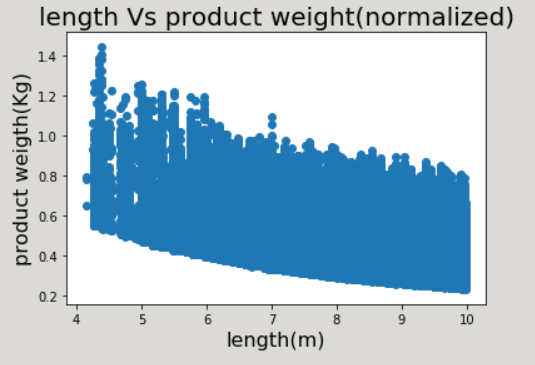
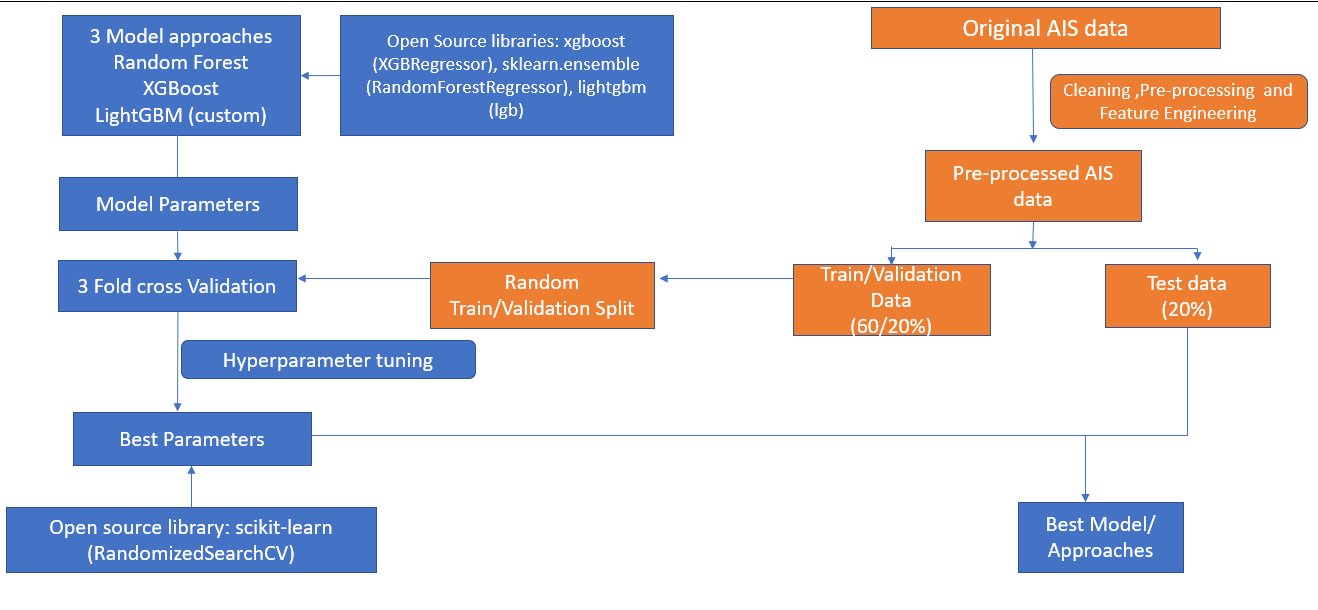


Figure 16: Length of vessel vs product weight (after transform)

**Modelling Approach:**

In this study we have considered the following factors that can contribute to change in the product weight: vessel length, gear/tools for fishing, fishing location, target species, product condition, month, season and distance from North Sea. The task is to investigate how accurate we can predict product weight given these conditions.



We have systematically explored predictive capabilities of three popular supervised machine learning approaches such as Random Forests (Breiman, 2001) XGBoost (EXtreme Gradient Boosting) (Chen and Guestrin, 2016), and LightGBM (Ke et al., 2017) which are recognized for predicting patterns in data and admirably on tabular datasets. These methods are based on a decision threes algorithm (Quinlan, 1986) but are different in how the trees are created and built and in how the predictions of the threes are summed up. The mathematics behind the decision trees can be found in e.g., Breiman et al. (1984) whereas the implementation of the above approaches can be found in Pedregosa et al. (2011), thus only a short summary is provided below. For catch prediction, we have provided the implementation of one of the best-performing models (algo) on GitHub (link). In case Random Forest and XGBoost algorithm we have used the default custom objective and custom evaluation function while in case of LightGBM algorithm we have made and used our own clutom objective and custom evaluation function.

*Random Forest Approach*

The random forest predictor H(𝚯) (Fig. 6) is formed by taking the average over K tree predictors {h(X, Θk)}. 𝐻(𝚯) = { ℎ ( 𝑿, 𝚯𝑘 )}, 𝑘 = 1, …, 𝐾 (1) where X is the input data matrix with n rows and p features such as vessel’s position, route, etc.; 𝚯 = [Θ1, …, ΘK], represents the parameters in H and includes splitting variables and their splitting values; K is the total number of trees in the model. These parameters are obtained by training data X and Y, where Y is the outcome vector containing vessels speeds. Through the fitted forest predictor, for any set of features Xi, i=1, …, n, we obtain the speed prediction from each tree in H.

*Gradient boosting approaches*

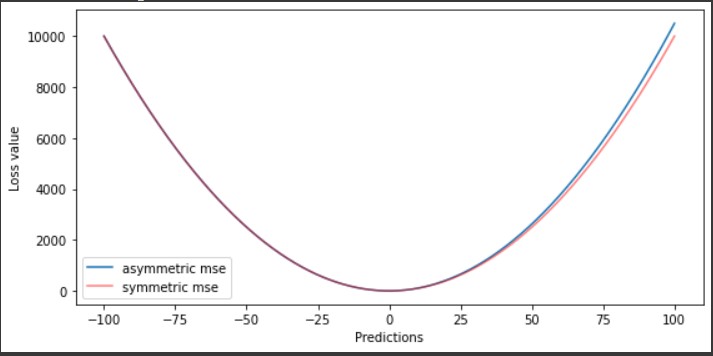
Like the random forest approach, gradient boosting approach is a set of decision trees. There are two main differences: (1) how trees are built and (2) how the trees’ predictions are combined. The random forest builds each tree independently while gradient boosting builds one tree at a time. The random forest combines the predictions at the end of the process (by averaging) while gradient boosting combines the tree’s predictions along the way. XGboost method applies horizontal three growth, whereas LightGBM applies vertical growth (see Fig. 7).

The model (or the exact mathematical structure by which the prediction of speed is made from the input variables), including the model parameters were determined by learning from the two years of enhanced AIS data as described in the following paragraphs.

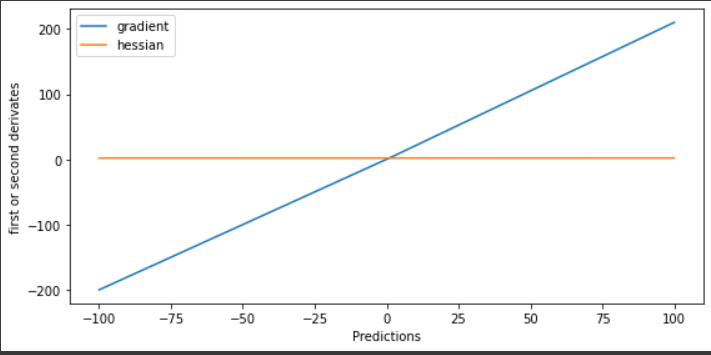
*Creating LightGBM custom function*

Our task is the prediction of catch given certain parameters like physical properties of the vessel, the (gear, mesh) tools with which they fish and geographical location (latitude, longitude) at which they fish, the targeted species of fish and also the time of the year(month) they want to fish. So, we can have two types of incorrect predictions, overestimated predictions and underestimated predictions. In case of default loss function both of these incorrect predictions would have been penalized equally however we don't want that because in our case if we have the overestimated prediction and the fisherman is not able to fetch the predicted amount of catch in accordance to which he had earlier done investment , then it will lead to the loss of fisherman and we don't want that , so instead of giving equal penalty to both kinds of incorrect predictions we give more penalty to overestimated predictions and lesser to underestimated predictions because this cannot lead to the loss in any case unlike overestimated predictions . so, we design our custom loss function in a similar manner. Any loss function for a regression problem is based to optimize on mean squared error, so we multiply the error by the penalty factor for each error. The penalty factor we use is p = 1.05, because it seems to be the convergence point, that is no further significant change in MAE after this point. The penalty will be included in the gradient, hessian as well as the evaluation function

Following is the plot of the gradient of our custom objective function for regression (penalty included) – asymmetric loss vs the default objective function for regression – symmetric loss



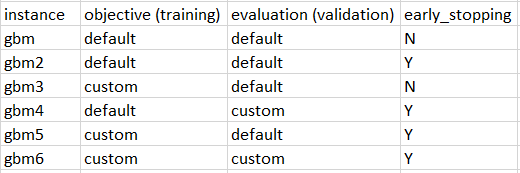
**Fig 15**: Gradient–1 order derivative of squared error



**Fig 16:** Hessian – 1 order derivative of squared error.

Moreover, we create six instances of LightGBM model to get different insights and so that we can have results to compare with each other.

The following represents what each instance signifies:



**Fig 17:** Different Lightgbm models for training

The model gbm6 is the model with custom objective and custom evaluation function where we have also applied early stopping. We observe that this model provides the least MAE on test set hence we use this model for further analysis.

**Feature engineering:**

We have introduced new features into the preprocessed AIS data These features are described in the following paragraphs and include North Sea distance and season feature.

***North Sea distance***

The number of fish in a particular region depends upon the depth of water column at that location. Depth is known to influence many factors on reef ecosystems for both coral and reef fish communities mainly due to light attenuation, changes in water temperature and pressure. The location of North Sea from which we calculate haversine distances to fishing locations is (56.88, 3.51). The distance captures the sense of relative number of fish that can be found at a particular fishing location.

***Season feature***

Since the case study focuses on Norway, hence we need to inculcate the fact that there are different fishing seasons. Cod fishing in Norway varies moderately throughout the year. High season is March to June and rest is considered low season. (For analysis purpose). We have considered 1 for high season and 0 for low season.

**Model training and validation:**

The data set was split into a training set (60%), a validation (20%) set and a test (20%) set. To divide data into these three categories, we randomly split each data point into the tree sets with the constraint that each set contains the same proportion of species codes. We use the stratify feature for this purpose. Also, the species code is a string format information to preserve its essence. For inputting to the model, we one hot encode the species code.



Table 2

After training each of the three models: Random Forest, XGBoost, and LightGBM, we exposed it to the validation dataset for the estimation and optimization of its performance. The optimization is done by tuning the hyperparameters of the model using 3-fold cross-validation. The hyperparameters for each model are listed in Table 2. The hyperparameters of all the models in different input data settings were tuned using RandomSearchCV method from the scikit-learn library (Pedregosa et al., 2011). After optimizing the hyperparameters of the models, we tested its predictive capability on the test set.

**Performance metrics:**

Error measures (performance metrics) are vital component of the model evaluation. In this study we employed mean absolute error (MAE) (see Eq. (2)) and standard deviation of the absolute error (STD) (see Eq. (3)) to evaluate the modelling approaches.

(2)

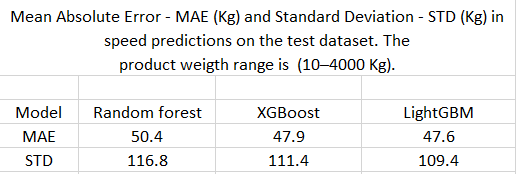
(3)

where n is the number of data samples, yi – single model prediction of product weight (Kg), xi – corresponding AIS record of product weight (kg), and 𝜇 – average of absolute error.

**Model limitations**

The models’ parameters are based on the historical AIS data (2001, 2021) from Norwegian fisheries. Any limitations of the underlying training data have been transferred to the models. The models will not be able to produce predictions for data points beyond the scope of AIS training data (i.e., to capture any trends in data outside the observed training set).

Also, the study does not incorporate the environmental variables and any political tensions that might also affect catch.



**Results and Discussion**

This paper has addressed the following research question: how accurately can one predict product weight (catch) by knowing underlying conditions as mentioned in the paper? As a case study we have selected the region of the North Sea (South Norway) and a specific vessel type and species. It was hypothesized that it is possible to predict the catch at a given location and time of the year, provided a representative historical dataset of catch. We have analyzed and enhanced AIS data for the period of 2001–2021 (two full decades) and have explored the ability of three distinct supervised machine learning models (Random Forest, XGBoost, and LightGBM) to capture complex relationship in the data and then tested predictive capabilities of several models.

The following paragraphs present and discuss results of this evaluation. After 3-fold cross-validation of the models (with parameters in Table 2; see values highlighted in black. Mean absolute error (Kg) and standard deviation (Kg) in catch predictions on the test dataset can be found in Table 3. Based on the MAE scores presented in Table 3, it can be concluded that LightGBM model with our custom defined objective and evaluation function with additional features was better than other approaches with MAE up to 47.7 Kg and the standard deviation of 109.2 Kg on a test dataset (product weight ranging between 10 and 4000Kg). The feature importance of the best performing model is presented in Figure 18. The vertical axis is the normalized mean decrease in impurity score or Gini importance (Louppe, 2014). Each feature importance was calculated as the sum over the number of splits (across all trees) that include the feature, proportionally to the number of samples it splits (ref. scikit-learn library implementation, Pedregosa et al., 2011). The higher the value of the importance score, the more important the feature is. Length was the most important feature to explain the catch variations in our dataset, followed by month, longitude, gear code, latitude, product condition code and distance from North Sea. Other features (coast code, species code and season) were found to be less important but still contributed to the predictive capabilities of the models.

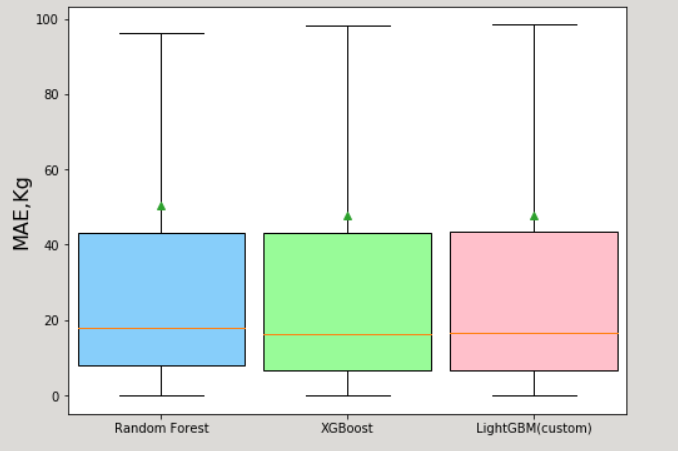


Figure 17: Box-plot of MAE for different models

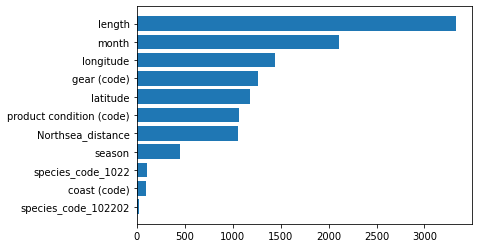


Figure 18: Feature importance of LightGBM (custom) model

Furthermore, these predictions do not exceed the maximum values of the AIS speed or the minimum of the AIS speed.

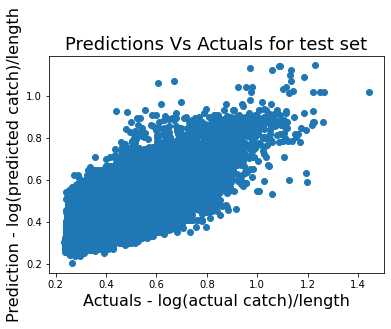
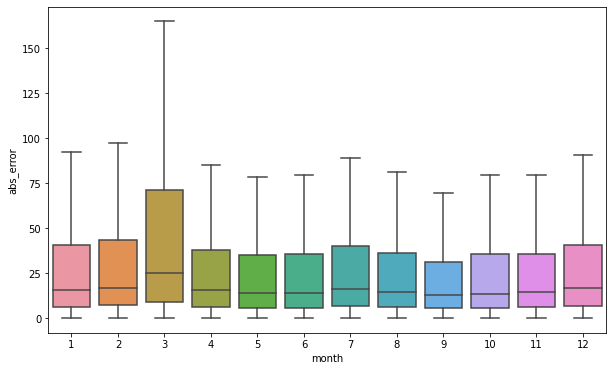


Figure 19: Actual vs Predicted values ligthgbm mode

Figure 20: Box-plot for month wise error

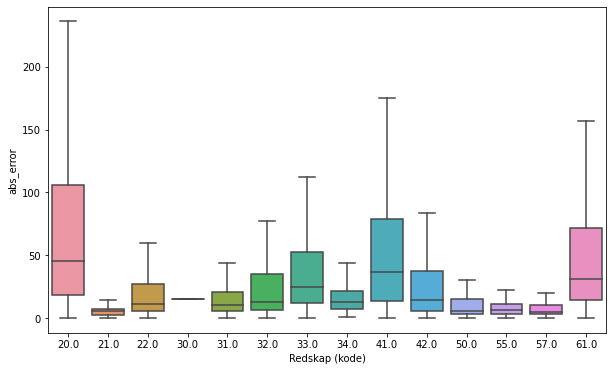


Figure 21: Box-plot for gear wise error

From the month analysis in Figure 20 we can see that the average error for all months except March is lower than 50kg, this indicates we need to pay more attention to transactions or activities happening in the month of march. For March the average absolute error is above 70kg.

From the gear analysis in Figure 21 we can say that we need to be more alert while fishing with gear/tools having code 20-undefined yearn, 41-ruser and 61-spinning rod as these are the gears for which the average error is higher than 50 kg. For the rest of the gears, it's less than 50 kg so it's fine.

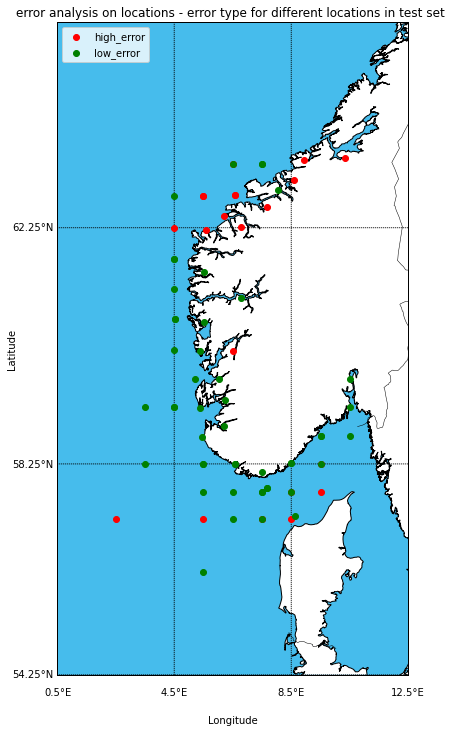


Figure 22: Location wise error (Green - low error < 50kg, Red - high error > 50kg)

There are a total 72 unique fishing locations for south Norway (cleaned data - test set) and we set the threshold to be 50 kg, that is if the error in that region is higher than threshold, it is said to be a region with high error whereas the region with error lesser than threshold is a region of low error. This can signify where they have trustworthy people and where they do not. From the location analysis we can say that we need to pay more attention to the activities of two locations one is north region of North Sea bounded by longitude 4.5E and 8.5E above the latitude 62.25N on the coastline of Norway and the other is just above the coastline of Denmark as these fishing locations have a cluster of red dots indicating that these are the regions of high error. Majority of the high errors are in the region below Trondheim to Haugsbygada

**Conclusion:**

The Fish catch has two main importance that are food security needs and generation of export revenue. In this study we have investigated a possibility of predicting catch using historical AIS data. First, we have analyzed AIS data retrieved from Directorate of Fisheries, Norway, and then we trained three machine learning models on the enhanced AIS data (extra information about North Sea distance and the season). Then we evaluated the models’ performance on the validation data to tune the hyperparameters for the best model. Next, we have exposed the models to test data to evaluate their predictive capability. The main results of this study show that (1) Supervised machine learning methods can be used for predicting catch (10 − 4000 kg) with mean absolute error up to approximately 47.6Kg in average. (2) Majority of the high errors are in the region below Trondheim to haugsbygada. The month of March experienced the highest error, along with these certain gears like 20-undefined yarn, 41-ruser and 61-Spinning rod have also resulted in high error with respect to others. (3) The LightGBM, model built on custom objective and evaluation functions with additional features, is a better modeling approach in a view of the considered dataset and modelling techniques. (3) geographical information such as a distance from the predefined places (North Sea) can be engineered to enhance the predictive capabilities of the models. We provide implementation of one the best models (LightGBM model for catch prediction) on GitHub(to-do). The presented machine learning approach to catch prediction could be considered as a source of additional and complimentary information for tactical planning (months ahead) of vessel trips in the North Sea as well as to support development of new strategies that can include the effect of climatic variables as well. Future work should focus on supplementing existing AIS data with the data from the environmental data, development and training of the models based on a continuously growing dataset.

//// **Declaration of Competing Interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. /////

**Acknowledgments**

The computations were performed on resources provided by UNINETT Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway. We thank the Norwegian Coastal Administration for providing access to the AIS data.

**References**

**Rough work**

**References:**

1. [**https://www.statista.com/statistics/643484/per-capita-consumption-of-fish-and-fish-products-in-norway/**](mailto:https://www.statista.com/statistics/643484/per-capita-consumption-of-fish-and-fish-products-in-norway/)
2. [**https://www.statista.com/statistics/665969/export-value-of-seafood-from-fisheries-and-aquaculture-in-norway/**](mailto:https://www.statista.com/statistics/665969/export-value-of-seafood-from-fisheries-and-aquaculture-in-norway/)
3. [**https://ieeexplore.ieee.org/document/1519023**](https://ieeexplore.ieee.org/document/1519023)
4. [**https://journals.ansfoundation.org/index.php/jans/article/view/1252**](mailto:https://journals.ansfoundation.org/index.php/jans/article/view/1252)
5. [**https://content.iospress.com/articles/journal-of-ambient-intelligence-and-smart-environments/ais210604**](mailto:https://content.iospress.com/articles/journal-of-ambient-intelligence-and-smart-environments/ais210604)
6. [**https://ndpublisher.in/admin/issues/EAv63n4e.pdf**](mailto:https://ndpublisher.in/admin/issues/EAv63n4e.pdf)
7. [**https://cdnsciencepub.com/doi/abs/10.1139/f82-036?journalCode=cjfas**](mailto:https://cdnsciencepub.com/doi/abs/10.1139/f82-036?journalCode=cjfas)
8. [**https://www.sciencedirect.com/science/article/abs/pii/S0034425710000295?via%3Dihub**](mailto:https://www.sciencedirect.com/science/article/abs/pii/S0034425710000295?via%3Dihub)