Interpretable AI for Fisheries Data

Aida Ashrafi[1 [0009−0000−9117−592X]], Katja Enberg[2 [0000−0002−0045−7604]], and Bjørnar Tessem[1 [0000−0003−2623−2689]]

[1] Dept. of Information Science and Media Studies, University of Bergen, Norway
[2] Dept. of Biological Sciences, University of Bergen, Norway
{aida.ashrafi,katja.enberg,bjornar.tessem}@uib.no

Abstract. Sustainable use of fish resources is essential, and decision makers such as the Norwegian Directorate of Fisheries (NDF) must take proactive measures to prevent Illegal, Unreported, and Unregulated (IUU) fishing activities. With access to large volumes of open datasets, machine learning (ML) models can play a key role in automating the detection of hidden patterns indicative of such activities. One valuable dataset is the collection of catch reports, where fishermen record the details of their fishing operations. Previous research has explored the use of ML models to predict expected catch quantities. By comparing these predictions with the actual reported values, potential violations of regulations can be identified. However, to ensure trust in the model's outputs and to gain deeper insight into the data, this paper applies interpretable Artificial Intelligence (AI) methods and visualization techniques to analyze prediction errors. We investigate feature importance and examine how the most influential features affect the model's output patterns. The results are promising, demonstrating that it is possible to provide transparency in the use of ML models for fisheries data. This approach enables domain experts to better understand, trust, and make informed use of the model's findings in the future.

Keywords: Sustainability · Fisheries · Catch reports · Prediction error · Interpretable AI.

1 Introduction

Fisheries is a vital industry, especially in Norway, as they provide both food and economic benefits to society. To ensure that resources are preserved for future generations and to meet the sustainable development goals related to life below water, it is crucial to combat Illegal, Unreported, and Unregulated (IUU) fishing. The primary challenge lies in monitoring fishing vessels, as they often operate far from the reach of fisheries experts and coast guards. In Norway, to mitigate this, fishermen are required to submit reports detailing the potential resources harvested during these trips, which are essential for analysis and ensuring sustainability.

The amount of fisheries data available today is greater than ever, yet there are few really efficient automated methods to process it to deliver almost real-time

analyses of fisheries activities. One optimal approach to assist decision-makers in uncovering hidden patterns related to IUU fishing is through the use of ML models. While fishermen's reports may contain some inaccuracies, the majority of them remain reliable and useful for the purpose. ML algorithms can identify recurring patterns over time, enabling the detection of deviations or outliers that may indicate potential IUU activities.

NDF provides several publicly available datasets, including catch reports from fishermen. These reports contain information such as fishing intervals, time and location, vessel length, and the species caught. However, the datasets do not include a direct definition of what constitutes a "regular" or "irregular" fishing operation. That said, Tessem et al. [12] have proposed a method for identifying overall deviations in a vessel's behavior, which could help highlight potential irregularities. To better understand the data and the decisions made by the model, the next step is to employ interpretable AI techniques and visualization methods.

In this paper, similarly to [12], we predict the catch weight using features from the catch report dataset. Since deviations in the error distribution compared to the rest of the population can indicate abnormal behavior, our goal is to first predict the error using the same set of features. The error is the difference between the predicted and the true values. Then, we analyze the error by exploring the influence of input features, with a particular focus on time. Through data visualization and interpretable AI techniques, we aim to address questions such as:

- What are the most influential features, and how do they impact the model's predictions?
- Is there a noticeable pattern in the error distribution over time?
- Are there specific time periods when the error is higher, and if so, what might be the potential reasons behind this?
- Does the error distribution pattern vary over the years?

In the following section, we provide an overview of the problem background and related work concerning the selected methods. In Section 3, we detail our methodological approach, from dataset preparation and feature selection to the prediction task and the models employed, complemented by visualizations and plots that aid in interpreting the outcomes. Section 4 offers a discussion of key insights and the broader implications of our findings. Finally, in Section 5, we summarize the work and outline potential directions for future research.

2 Background and related work

2.1 Interpretable AI

A well-known definition of interpretability comes from [9], which states: "Interpretability is the degree to which a human can understand the cause of a decision." Interpretable AI focuses on answering the question, "Why did the

model produce a specific output?" It emphasizes the transparency of the model rather than just the results it generates [1]. In applied ML projects like ours, providing explanations and interpretations is crucial for debugging the model, understanding potential biases, and, most importantly, building trust with the domain experts we collaborate with [5].

One type of interpretable method is post-hoc, which includes a subcategory called model-agnostic methods. Model-agnostic methods are further divided into local and global approaches. Post-hoc interpretability means applying interpretability techniques after the model has been trained. Model-agnostic methods do not consider the internal workings of the model; instead, they analyze how the model's output varies in response to changes in the input features. Local interpretation methods focus on explaining individual predictions. Shapley values, a type of local attribution method, break down a single prediction into the sum of its feature effects, providing a detailed "zoom-in" view that highlights unusual cases.

In contrast, Partial Dependence Plots (PDPs) are global methods that illustrate the overall marginal effect of one or two features on the output of the model. Accumulated Local Effects (ALE) plots are also global methods that measure the influence of a feature on the model output, but they do so in a locally aware way. Unlike PDPs, ALE does not assume feature independence and avoids creating implausible data combinations.[10].

Surrogate models, as a model-agnostic method, involve replacing a black box model with a simpler model for interpretation goals; basically, using more ML to interpret the original ML model we have [10].

We proposed a method inspired by the surrogate model approach to better interpret the behavior of our ML model. To further understand the model's output, we first computed feature importance scores. Then, we leveraged a combination of interpretability techniques, including Shapley values, PDPs, and ALE plots, to analyze how different features influence the model's predictions. Here, Shapley values serve as a global explanation technique.

2.2 Data visualization

Data visualization is an essential tool for identifying patterns and trends during the exploratory phase of analysis [8]. By using graphs and charts, large datasets can be represented in a way that makes the information easier for our brains to process and understand intuitively. For example, in [4], several visualization techniques such as bubble charts, parallel coordinates, line graphs, and box plots were used to uncover patterns in whole-person health data.

In this paper, we present various visualizations of both the dataset and the model's output. All the plots generated through interpretable ML methods bring us one step closer to understanding how different features influence the trends observed in the model's output, which can, in turn, assist decision-makers in making more informed future decisions. To the best of our knowledge, this is one of the few studies focused on visualizing Norwegian catch reports.

2.3 AI and data science for sustainable fisheries

According to Rubbens et al. [11], ML, as a subfield of AI, provides a powerful alternative to manual processing of large-scale fisheries data. Its speed, flexibility, and ability to reduce noise and bias make it particularly valuable. Various ML techniques have been applied to identify IUU fishing activities, with these methods detecting characteristic patterns associated with such operations by analyzing vessel trajectories [6], [2]. Additionally, anomaly detection, whether based on individual catch reports [3] or a collection of reports, can help identify potential deviations from normal fishing behavior [12].

In [12], potential collective anomalies were identified by predicting the logarithm of the sum of the catch round weight. The error is defined as the difference between the logarithm of the predicted value and the logarithm of the reported value. Subsequently, a T-test is applied to detect vessels with error distributions that deviate significantly from the rest of the population. These vessels are then flagged as potential anomalies with unusual behavior.

Since there are not many works around the interpretable AI for fisheries, in this paper, we aim to have a closer look at the output of the model from [12] and interpret the effect of features on that. To achieve this, we adopted the same approach as in [12], utilizing XGBoost to calculate error values. But since the error is even more important than the catch itself, we employed a Decision Tree and a Random Forest to predict these error values based on the same set of features. This simple method was inspired by the surrogate models. We obtain the ranking of features using impurity-based feature importance. Additionally, the error values are plotted over time as one of the input features, allowing us to identify trends across different months or years. Moreover, Shapley values are used to illustrate how time influences the error. PDP and ALE are included to provide further insight into the effect of different important features on the error.

3 Methodology

3.1 Data preparation

One of the open datasets provided by the NDF [7] is the Logbook, which includes Daily Catch Activity (DCA) reports for fishing operations in Norwegian waters. We used the same model and feature set as in [12]. Our experiments began with data from bottom trawlers in 2018, and we then applied the same methodology to data from subsequent years to identify trends and compare similarities and differences over time.

The dataset contains several redundant features, including various coded fields and the vessel-related information that were excluded from our analysis. Each catch can span multiple entries in the dataset, with different species recorded. Therefore, we grouped all reports corresponding to a single catch and summed the round weight of all species involved. This total round weight served as the target feature for our prediction task. The final set of input features includes:

- The start and stop positions of the catch interval (latitude and longitude),
- The start and stop times of the catch interval,
- The duration of the catch,
- The length of the vessel, and
- The main species caught, defined as the species with the highest weight in each catch.

To reduce noise and identify reliable patterns, we filtered the dataset to include only records where the main species had over 1,000 occurrences. For each year considered, the dataset contains over 30,000 reports. Date is a cyclic feature, so we used the trigonometric transformations to convert it into sin and cos features.

3.2 choice of the ML model and the prediction task

In [12], the data is grouped into reports by each vessel, and then the behavior of each vessel in terms of reporting the total catch is compared with the rest of the vessels. Among the models explored in their work, we selected XGBoost to predict the logarithm of the total catch, following the same approach. However, as noted by Tessem et al. [12], deviations in the distribution of prediction errors can indicate potential anomalies or inconsistencies in catch reporting. Based on this insight, we define a new prediction task.

Our approach is inspired by the concept of surrogate models, where an interpretable ML model, such as a decision tree in our case, is used to explain the behavior of another model, like XGBoost. The key difference is that, unlike traditional surrogate modeling, we did not train the surrogate on the same prediction task. Instead of predicting the total catch, we use a decision tree to predict the model's error directly from the same set of input features.

This approach allows us to investigate how different features influence the error and may help uncover systematic patterns or biases. While the decision tree tends to overfit the data, this is acceptable, and maybe even wanted for our purposes, as our goal is not to build a highly accurate predictive model, but rather to analyze and interpret the sources of error. XGBoost, while effective for catch prediction, is less suitable for directly modeling the error from the same features.

3.3 Feature importance

The performance of the Random Forest model in terms of the r^2 score is significantly better than XGBoost, but slightly worse than the Decision Tree. Despite this, Random Forest remains a strong candidate for this task due to its balance between predictive performance and interpretability. Additionally, it provides useful insights through impurity-based feature importance.

Random Forest in Scikit-Learn utilizes a method known as impurity-based feature importance to evaluate the relevance of input features. This method is often referred to as Mean Decrease Impurity (MDI) or Gini importance. MDI measures how much each feature contributes to reducing impurity across the

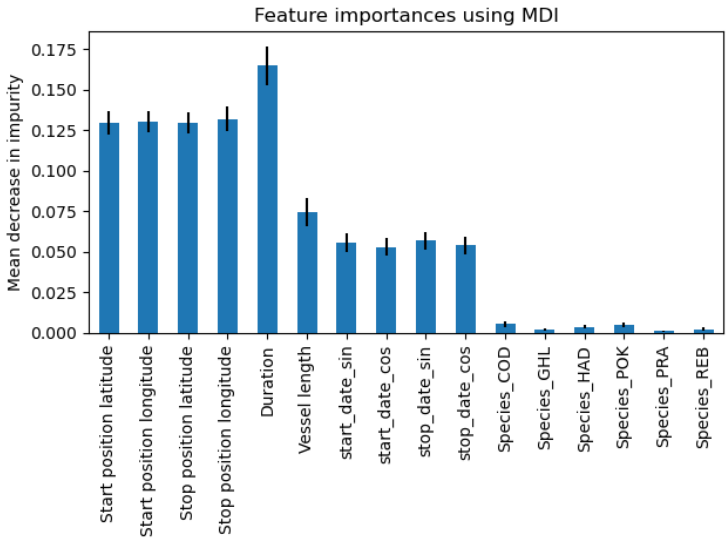

Fig. 1: feature importance scores using MDI for the 2018 dataset when using a Random Forest model to predict the error.

decision trees in the forest. Specifically, it calculates the reduction in Gini impurity achieved by each feature when it is used to split a node. In the context of a Random Forest, the MDI for a given feature is computed as the average impurity reduction that the feature contributes, aggregated over all the trees in the ensemble. Features with higher average impurity reduction are considered more important. Figure 1 shows the feature importance scores for the 2018 dataset when using a Random Forest model to predict the error. While the feature importance scores may vary slightly across different years, the ranking of features in terms of importance remains consistent with the results from 2018. Duration is the most important feature for predicting the error, followed by position, vessel length, and the time of the catch, in that order of importance.

3.4 PDPs

Furthermore, we employed PDPs to investigate the relationship between the most important features, specifically, the duration of the catch and the length of the vessel, and the target variable, which in this case is the prediction error.

To generate a PDP, we first train a model, then systematically vary the values of a selected feature across its range for all instances in the dataset, while keeping the other features fixed. For each value of the chosen feature, we compute the model's average predicted outcome. By plotting these feature values against their corresponding average predictions, we can visualize the marginal effect of that feature on the model's output.

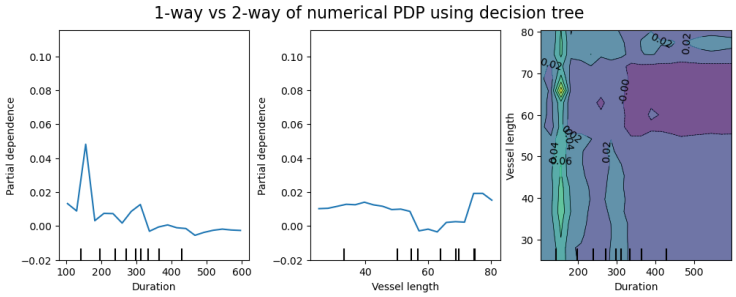

Fig. 2: PDPs to show the influence of duration of the catch and length of the vessel on the prediction error value for 2018 data.

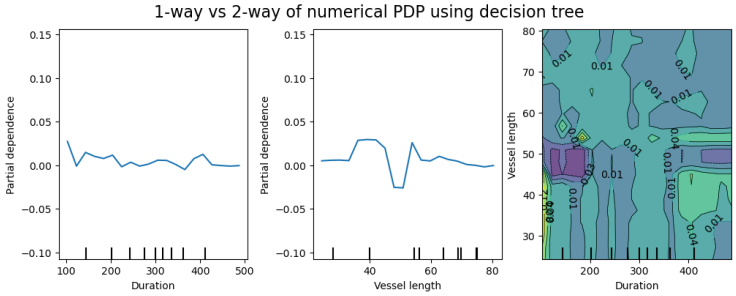

Fig. 3: PDPs to show the influence of duration of the catch and length of the vessel on the prediction error value for 2019 data.

Figure 2 presents two one-way PDPs and one two-way PDP for data from 2018. The one-way PDPs illustrate the effect of a single feature on the prediction error. The left plot shows how the duration of the catch influences the error, while the middle plot demonstrates the effect of vessel length. The right plot shows a two-way PDP, highlighting the combined influence of duration and vessel length on the prediction error.

In the middle plot, we observe that larger vessels tend to have smaller prediction errors, indicating that their reported catches align more closely with the model's predictions. This changes for very large vessels. The PDP for duration reveals a noticeable peak in error around a 160-minute duration. We hypothesize that this may be due to fishing operations that were started but aborted early, possibly because little or no fish were found, resulting in irregular or unexpected reporting patterns.

Despite some fluctuations, longer durations generally lead to smaller prediction errors, suggesting more consistent or predictable catch behavior. However, this trend reverses slightly beyond approximately 450 minutes, where the error begins to increase again. According to domain experts, bottom trawling operations lasting more than 400 minutes are considered unusually long and may be possible reporting errors. It is possible that fishermen who are less precise in reporting unusually long durations may also be less accurate in reporting total catch.

From the right plot, which shows the two-way PDP for duration and vessel length, we can identify a distinct region, around 160 minutes duration and 67 meters vessel length, with the highest prediction error (depicted as a very small deeper area in yellow color). This observation aligns with patterns seen in the one-way PDPs. Additionally, for durations exceeding 350 minutes, prediction errors tend to remain low, particularly for vessels between 55 and 72 meters in length.

We also provide PDPs for the same features using the data from 2019 in Figure 3. While there are some similarities with the 2018 plots, for example, peaks in prediction error appear around similar durations (approximately 100, 300, and 400 minutes), the magnitude of these peaks is visibly different.

Another consistent pattern is that larger vessels tend to have lower prediction errors. However, in the 2019 data, this trend remains stable even for the largest vessels, whereas in 2018, a slight increase in error was observed for vessels at the extreme upper end of the length range.

These observations suggest that, although duration and vessel length show some consistent influence on prediction error across years, they may not be the most reliable indicators when assessing model error over time. It's also important to note that changes in regulations between years can significantly affect fishing practices, which may contribute to changes in reporting behavior.

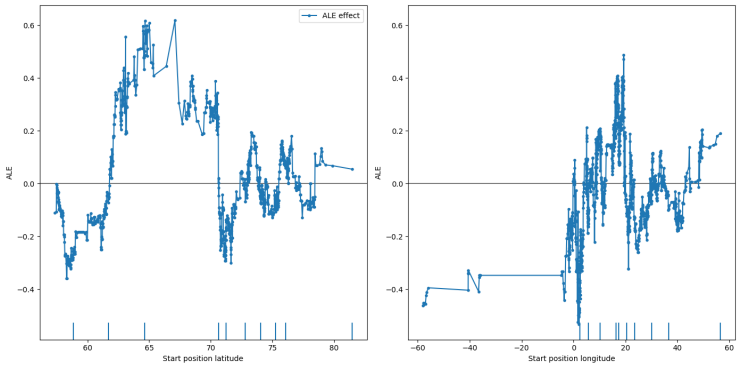

Fig. 4: ALE plots to show the effects of latitude and longitude on the prediction error for 2018.

3.5 ALE plots

Another important feature to examine is position, represented by latitude and longitude. Since these two features are naturally correlated, using ALE plots are more appropriate and reliable.

Figure 4 shows the ALE plots for latitude and longitude for 2018. From the left plot, we observe that the error increases when moving from south to north (i.e., from latitudes below 62° to those above). This observation aligns with insights from domain experts, who suggest that fishing operations in the southern regions tend to have greater transparency in reporting. A similar trend is also evident in the 2019 data, as shown in the left plot of Figure 5.

The effect of longitude on prediction error shows a more noticeable difference between 2018 and 2019 compared to the effect of latitude; this can be seen in the right plots from Figures 4 and 5. In 2018, the peak occurs around longitude 17°, while in 2019 it shifts westward, peaking around 4°. However, in both years, the largest positive effect is observed at longitudes below 20°. Taking into account the effects of both latitude and longitude, we can conclude that the prediction error tends to increase in the region between approximately the positions of Trondheim and Tromsø.

So far, we have investigated the most important features contributing to prediction error. Certain patterns are visible, offering insights and transparency into how different features affect prediction error over the course of a year. However, due to the complex and dynamic nature of the data, which may be influenced by external factors such as environmental variability, fishing behavior, and regulation changes, we conclude that no single feature can be considered a consistently reliable basis for interpreting prediction error patterns across different years.

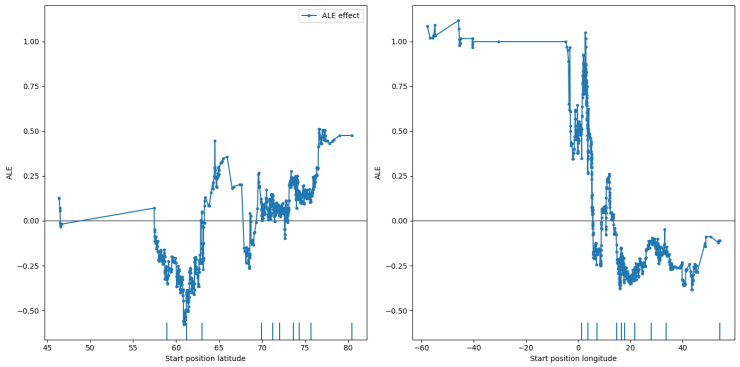

Fig. 5: ALE plots to show the effects of latitude and longitude on the prediction error for 2019.

3.6 Error over time and Shapley values

Figures 6 and 7 show the smoothed average error (in blue) and Shapley values for time (in green) across different months of the years 2018 and 2019. The Shapley values reflect the influence of time on the prediction error. In both years, the model tends to underpredict the reported catch weight, which explains why the error values are predominantly negative in these plots.

Notably, the first few months of each year exhibit more negative errors compared to the rest of the year. This indicates that reported catch values were generally higher than predicted during this period. We hypothesize that this could be due to underreporting at the end of the previous year, possibly by vessels that had exceeded their quotas. These unreported catches may have been reported in the early months of the following year, leading to the observed discrepancy.

In both years, this trend appears to shift around the spring months, with errors becoming less negative or more balanced. The Shapley values show a generally negative effect of time on the prediction error. Although the effect fluctuates, it tends to decrease gradually over the course of the year.

We presented plots from two consecutive years, as space is limited. However, the method is easily extendable to additional years. The trends, similarities, and differences observed in other years are comparable to those found between these two, making them representative for illustration purposes.

4 Discussion

The models used in this work are not perfect; The XGBoost model consistently predicts lower catch values compared to the reported values for almost every vessel, which indicates a systematic bias and suggests that the model's predictions may lack balance and realism. On the other hand, the Decision Tree model overfits when used to predict the prediction error. However, given the inherent

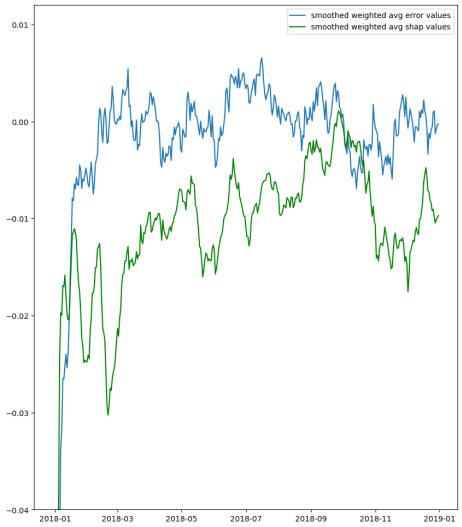

Fig. 6: Error and Shaply values over time for 2018.

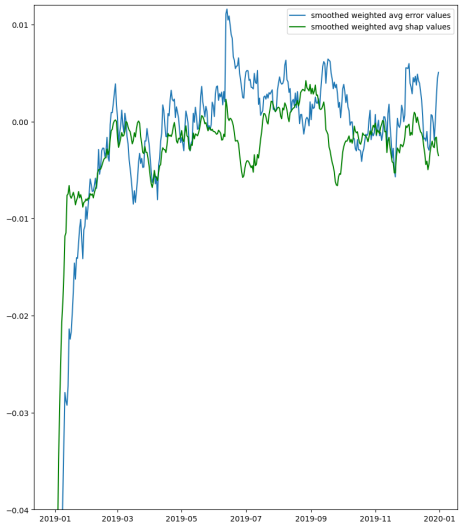

Fig. 7: Error and Shaply values over time for 2019.

randomness in the data and its strong dependence on environmental factors and yearly changes in regulations, this type of in-depth analysis remains valuable.

Additionally, interpreting the model's output and examining which feature values are associated with larger prediction errors, we found it difficult to identify consistent patterns of high error values that could be reliably linked to specific values of important features across different years. While this inconsistency may be partly explained by year-to-year variations in regulations, we believe a significant contributing factor is, once again, the inherent complexity and variability of the data itself.

Despite this challenge, we can still see some similar patterns over the years, and through them we can gain insights into potential misreporting of catch values. Such interpretations can be particularly useful for domain experts, as they may offer a deeper understanding of patterns in the data and provide early indicators of irregular activity. Identifying feature values linked to unusually high errors could help pinpoint areas where suspicious or non-compliant fishing operations are more likely to occur. This, in turn, can guide future monitoring efforts and help focus regulatory attention on the most relevant cases.

Moreover, plots and visualizations are generally the preferred method of communication for domain experts, making this work a valuable step toward more effective interaction and collaboration with them.

More advanced models, such as neural networks for tabular data, were also tested for predicting the total catch. However, despite their longer training times and increased complexity, the results were comparable to those of simpler models. Therefore, we chose to continue with simpler approaches, concluding that a significant part of the challenge lies in the inherent nature of the data, as previously discussed.

5 Conclusions and further work

Fishing catch reports from trawlers represent one of the largest datasets available over the years for applying machine learning models. Predicting the total catch is the most natural starting point. However, we aimed to explore how different features influence the model's output by predicting the prediction error directly from catch-related features. Here, prediction error is defined as the difference between the predicted catch and the reported one.

To investigate this, we applied interpretability and visualization techniques to analyze the influence of various features on this new target variable, using data from 2018 and 2019. Due to the data's dependency on environmental conditions and annual changes in regulations, it is challenging to identify consistent patterns across years. Nevertheless, we were able to uncover meaningful insights into the impact of key features within individual years, as well as some recurring trends across multiple years.

These findings can support domain experts in better understanding and trusting the model's behavior, ultimately helping them make more informed decisions based on the observed patterns.

In [12], a list of vessels with deviated catch reports is provided as potential collective anomalies. To offer experts more context beyond just the vessel names, a possible direction for future work is to visualize statistics of these candidate vessels. This would help better assess how significantly they deviate compared to others, using detailed information such as catch position and species.

Another potential direction for future work is to identify a more informative prediction task beyond total catch, for example, predicting the species composition and their distribution within a catch. This approach could reveal more insightful patterns and enhance the interpretability of the results.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE access **6**, 52138–52160 (2018)
2. Ashrafi, A., Tessem, B., Enberg, K.: Detection of Fishing Activities from Vessel Trajectories. In: Nurcan, S., Opdahl, A.L., Mouratidis, H., Tsohou, A. (eds.) Research Challenges in Information Science: Information Science and the Connected World. pp. 105–120. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-33080-3_7
3. Ashrafi, A., Tessem, B., Enberg, K.: Analysing Unlabeled Data with Randomness and Noise: The Case of Fishery Catch Reports. Swedish Artificial Intelligence Society pp. 204–212 (Jun 2024). https://doi.org/10.3384/ecp208023, https://ecp.ep.liu.se/index.php/sais/article/view/1015
4. Austin, R.R., Mathiason, M.A., Monsen, K.A.: Using data visualization to detect patterns in whole-person health data. Research in nursing & health **45**(4), 466–476 (2022)
5. Brennen, A.: What do people really want when they say they want" explainable ai?" we asked 60 stakeholders. In: Extended abstracts of the 2020 CHI conference on human factors in computing systems. pp. 1–7 (2020)
6. Chuaysi, B., Kiattisin, S.: Fishing vessels behavior identification for combating iuu fishing: Enable traceability at sea. Wireless Personal Communications **115**(4), 2971–2993 (2020)
7. Directorate of Fisheries: Electronic Reporting Systems, https://www.fiskeridir.no/English/Fisheries/Electronic-Reporting-Systems
8. Leban, G., Zupan, B., Vidmar, G., Bratko, I.: Vizrank: Data visualization guided by machine learning. Data Mining and Knowledge Discovery **13**(2), 119–136 (2006)
9. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019)
10. Molnar, C.: Interpretable machine learning. Lulu. com (2020)
11. Rubbens, P., Brodie, S., Cordier, T., Destro Barcellos, D., Devos, P., Fernandes-Salvador, J.A., Fincham, J.I., Gomes, A., Handegard, N.O., Howell, K., et al.: Machine learning in marine ecology: an overview of techniques and applications. ICES Journal of Marine Science **80**(7), 1829–1853 (2023)
12. Tessem, B., Ashrafi, A.: Collective anomaly detection in fisheries. In: Norsk IKT-konferanse for forskning og utdanning. No. 2 (2024)