Exploration of machine learning algorithms for maritime risk applications

Sabine Knapp[1] and Michel Van de Velden[2]

Econometric Institute Report 2021-03

Erasmus University

Abstract

To manage and pre-empt incident risks effectively by maritime stakeholders, predicted incident probabilities at ship level have different application aspects such as enhanced targeting for ship inspections, improved domain awareness and improving risk exposure assessments for strategic planning and asset allocations to manage risk exposure. Using a unique and comprehensive global dataset from 2014 to 2020 of 1.2 million observations, this study explores 144 model variants from the field of machine learning (18 random forest variants for 8 incident endpoints of interest) with the aim to enhance prediction capabilities to be used in maritime applications. An additional point of interest is to determine and highlight the relative importance of over 500 evaluated covariates. The results differ for each endpoint of interest and confirm that random forest methods improve prediction capabilities, based on a full year of out of sample evaluation. Targeting the top 10% most risky vessels would lead to an improvement of predictions by 2.7 to 4.9 compared to random selection. Balanced random forests and random forests with balanced training variants outperform regular random forests where the end selection of the variants also depends on the aggregation type and use of probabilities in the application areas of interest. The most important covariate groups to predict incident risk are related to beneficial ownership, the safety management company, size and age of the vessel and the importance of these factors is similar across the endpoint of interest considered here.

Keywords: ship specific risk, safety quality, reducing false negative events, risk exposure estimation; machine learning, case weighting, subsampling, random forest, sampling, evaluation metrics, top decile lift, variable importance, machine learning

[1] Corresponding author: Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, Netherlands, email knapp@ese.eur.nl
[2] Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, Netherlands, email vandevelden@ese.eur.nl

1. Introduction

Coastal states, ship owners and regulators face increased challenges due to volatilities in the economy, different uses of coastal states and complex operating environments as for instance the effect due to Covid19 has shown (Knapp, 2021). Potential consequences are loss of assets, lives and damages to the marine environment. The overall aim of this paper is to develop and test new methods from the field of machine learning (random forests) that can be used in various application areas where ship specific probabilities are needed. The associated data challenges for the creation of the risk formulas and prediction routines are also discussed in the relevant sections.

Ship specific risk is expressed as a probability of incident which can also be translated into percentile ranks or manipulated in other ways to serve the application area of interest. Estimating individual safety qualities of vessels is crucial for all application areas and the current approach is based on global fleet, incident and inspection data from 2014 to 2019 and 2020 for an out of sample validation exercise. Over 580 covariates are evaluated and eight incident end points of interest are considered as follows:1= very serious and serious incidents (VSS), 2= pollution (VSS), 3= loss of life and injuries (VSS), 4= fire and explosions (VSS), 5= collisions and powered groundings (VSS), 6= mobility related incidents (VSS) that include all failures that can lead to drifting such as engine failures, black outs, steering gear failures, propulsion failures, 7= hull related failures (VSS) and 8= equipment failures (VSS). All endpoints of interest are restricted to very serious and serious incidents (VSS) as data qualities are not good enough to consider less serious incidents as separate category.

Four main application areas of interest are identified and described along with some relevant literature. They are as follows: 1) improved targeting of vessels, 2) enhance domain awareness, 3) strategic planning and asset allocation and 4) quantifying potential incident consequences in monetary terms.

Vessel Targeting: For maritime administrations (flags and Port state control) or industry vetting organizations, it is important to be able to target high risk vessels for inspection thereby reducing false negative events (missing risky vessels) and to focus inspection priority risk areas. Targeting of vessels for inspections is an important risk control option to pre-empt incidents from happening and Heij et al (2011), Heij and Knapp (2012, 2018, 2019) and Knapp and Heij (2020) demonstrate the need to improve targeting ships for inspections including the focus on vessel inspection priority areas. Currently, the shipping industry is primarily using detentions as a proxy for safety quality (Knapp, 2006, Knapp and Heij, 2020) and to target vessels for inspections but Knapp and Heij (2020) demonstrate that false negative events (missing risky vessels) can be reduced by combining the two main risk dimensions (incident risk and detention). Ship specific probabilities transferred into percentile ranks are used to combine the two dimensions and to benchmark vessels. This manuscript further investigates the incident type endpoints of interest that can be used for targeting.

Domain awareness and strategic planning: For coastal states and search and rescue organizations, improved domain awareness can assist in knowing as early as possible when potentially dangerous situations are developing so that mitigation strategies can be deployed For medium to long term strategic planning aspects, risk assessments and sensitivity analysis can assist regulators to decide on deployment and financing of mitigation assets. Van der Hoorn and Knapp (2015, 2019) define total risk exposure as the integration of risk at the individual ship level with risk due to vessel traffic densities, physical environmental criteria (wind, currents, tides, bathymetry) and coastal sensitivities. They propose a multi-layered approach to account for all risk factors including uncertainties. In most cases the risk at the individual ship level is not considered in risk exposure assessments (DNV, 2013, BRISK, 2012, Hansen, 2007) and all vessels are assumed to have the same safety quality, which is unrealistic given that safety quality at the ship level can vary considerably (Heij and Knapp 2012, Heij et al. 2011, Knapp 2006).

Based on Van der Hoorn and Knapp (2015, 2019), two application outcomes exist. The first one is improved domain awareness where ship specific probabilities or percentile ranks are attached to ship positioning data (AIS data). Improved domain awareness can help reducing incidents as stakeholders know as early as possible if a potential risky situation is occurring or about to develop that could be mitigated with the appropriate intervention. The second application area relates to risk exposure assessments and prediction scenarios. Prediction scenarios help simulating the change in risk exposure given change in underlying factors (eg. ship traffic, risk profiles of vessels). For risk exposure estimations, probabilities rather than percentile ranks are needed as they form part of further calculations (eg. expected number of incidents) including spatial distributions across large or small areas of interest.

Quantifying potential incident costs: For insurers, safety management companies, ship owners and coastal states, estimating potential incident costs can be useful as they form the basis for cost benefit analysis, for determining insurance premiums (Knapp and Heij (2017), Knapp et al (2011)) or for quantifying changes in risk exposure due to unforeseen events such as Covid19 as well as quantifying the effect of inspections in reduction risk exposure (Knapp, 2021). In these application areas, probabilities are estimated and are used in different ways to estimate the effect of interest.

Different methods can be used to estimate the safety quality of a vessel such as logit models, Bayesian networks or survival analysis (Knapp, 2006, 2011, Hansen 2007, Knapp and Franses 2007, Bijwaard and Knapp 2009, Vander Horn and Knapp, 2015, 2019, Knapp and Heij 2020) or more recently random forests (Yan et al, 2021). Many classic statistical methods, rely on rather specific model assumptions. For example, one popular approach that has been applied in the maritime industry is the logistic regression model. In this method, probabilities are modelled by linking them to a set of independent predictor variables in such a way that the log-odds are linearly related to the predictor variables. For further details on logit models, refer to Verbeek (2008) or Greene (2000). An important advantage of a logistic regression model is that the signs of the coefficients of the model are easy to interpret. It is however difficult to assess and interpret the importance of the individual covariates when they are measured on different scales and/or when they are of different measurement types (e.g., numerical or categorical). Interpretation becomes increasingly cumbersome if many (e.g., more than 500) are evaluated using groups of (categorical) variables as some categories are omitted as they serve as benchmark variables, whereas others may have coefficients that do not significantly differ from zero. Finally, in logistic regression, interactions need to be explicitly formulated, before running the analysis, and incorporated into the model.

In this paper, alternative prediction methods from the field of Machine Learning are considered. Several definitions of machine learning exist, the term here indicates methods that are largely algorithmic rather than model based. That is, instead of explicitly formulating a statistical model, and corresponding distributional assumptions, the covariates are mapped to the response variable in such a way that out-of-sample predictions are optimized. See Breiman (2001) for a more elaborate treatment of the differences between statistical and machine learning.

To deal with the severe class imbalance problems, asymmetric measures are proposed, and methods can be adapted to better deal with the data's asymmetric properties. Such adaptations include the use of balanced sampling as was done by Lemmens and Croux (2006). They propose using an equal amount of `zero' and `one' observations in the sample used for modelling a binary dependent variable. The idea being that a larger relative proportion of one observation would improve the model's ability to predict ones. Lemmens and Croux (2006) point out that balanced sampling can lead to biased prediction results due to an overly positive model and therefore propose two bias corrections. Another approach concerns the use of asymmetric misclassification costs. If it is costlier to wrongfully predict a `one' than it is to wrongfully predict a `zero', the model should be better at predicting ones.

It appears that, despite its growing popularity in several research areas (e.g., marketing, finance, image and pattern recognition), machine learning methods so far have not been

developed for maritime applications related to incident probabilities. Recently, Yan et al (2021) considered the application of random forests for port state control detention and reported promising results and a similar direction is used by applying several random forest models to a unique and comprehensive global data set containing incident data.

Data needed for analysis in the maritime industry is often scarce as detention and incidents are rare events. The only exception would be vessel positioning data which is high frequency data in the range of millions of observations. Regardless of the frequency of data, the results of any method depend on the quality of its input data. That is, the data needed to estimate the models, as well as the "new" data to which the models should be applied to assess their performance. Experience with previous logistic regression models indicate that the risk formulas do not change significantly over time – for instance the effect of age, size, ship type, flag, class or even ownership on the probability of incident does not change overnight and the coefficients appear to remain valid for three to five years (Heij and Knapp, 2019, Knapp and Heij 2020).

The remainder of this paper is organized as follows. Section 2 provides a description of the datasets and covariates used. Section 3 then describes model variants as well as the selection of evaluation metrics that are used. Section 4 presents and discusses the prediction results of all models and the importance of covariates for selected models are visualized. Section 5 concludes the paper and highlights future areas of research. The analysis is supported by Annexes that provide the relevant supportive documentation of the logic of the approach used here.

2. Data sources, preparation and the selection of covariates

Data aspects are complex in the maritime industry. The events of interest are typically rare, and data is often of poor quality and difficult to access. Data sources for incidents must always be combined from multiple sources to reduce bias due to underreporting of incidents. The various challenges exist for both crucial steps that are needed to develop, test and use formulas that can estimate probabilities. – First, the creation of the training data (combination of data and years of data) to estimate formulas or create algorithms which requires extensive cleaning and processing and second, the processing and combining of several sources of new out of sample input data to create input parameters to predict probabilities at a given point in time using the developed risk formulas or algorithms. For data preparations, PHP was used while for forest estimations and evaluation metrics, the free statistical programming language R (R Core Team, 2021) was used.

This publication dealt with both aspects and the following data sources were used:

- Global ship particulars, inspection and incident data from 2014 to 2019 to develop the formulas. The data sources used are IHS-Maritime, LLIS, IMO and the USCG. The inspection data covers data from the main PSC MoU's (Paris MoU, Tokyo MoU, USCG and Vina del Mar and is based on Knapp (2006) to account for the inspection history of vessels.
- New out of sample input data from 2020 of global ship particulars and global incident data. Four time periods are chosen for the out of sample evaluation (January to March, April to June, July to September and October to December 2020) and it is assumed that estimated probabilities are valid for a maximum of 3 months (Knapp and Heij, 2020). This means estimated probabilities for incidents as of January (or April) are matched with empirical data from January to March (April to June) and so forth.

Incident data has different type of challenges for either the training data creation or the input data feeds. The creation of the training data to create risk formulas is complex and a lengthy process. Due to the poor data quality of incident data and to overcome under-reporting, global incident information was combined from four different sources (IHS Markit, LLI, IMO and AMSA) and duplicates are eliminated. Since different data providers use different definitions

to identify a very serious or serious incident, the data is manually reclassified. The remaining incidents according to the IMO definitions for seriousness, which are very serious (including total loss), serious and less serious (IMO 2000) For this paper, the main end point of interest is very serious and serious incidents (VSS). Less serious incidents are excluded due to the underreporting.

Since incidents typically have multiple consequences, the first event of a chain of events is identified so that incident types can be considered (eg. loss of life, collisions, pollution, fire and explosion etc.). Without these data preparations, incident data are not suitable for the estimation of risk prediction models. These preparations result in the following datasets:

- Training data: Global incident and world fleet data (721,767 observations; 2014 to 2019). This matrix is used for the estimation of the incident type models.
- Test data: Out of sample data from January to December 2020 consisting of 509,118 observations (129,365 individual vessels, which is the world fleet of all vessels above 100gt consisting of all major ship types (general cargo, dry bulk, container, tankers, passenger vessels, other ship types), fishing vessels, tugs and large commercial yachts). Ships under construction are excluded from the estimation routine.

For exact counts for each dependent variable for the training and out of sample data, please refer to the result section. The initial selection of variables is based on the literature such as Knapp (2006, 2011) and Knapp and Heij (2020) and an overview of the main variable groups used in the analysis is provided in Table 1.

Table 1: High level overview of number of variables used

Variable groups	**Data Type**	**Nr of variables or categories**
Size, age	continuous	2
Ship Types	categorical	9
Flag	categorical	198
Class	categorical	101
Main engine designer	categorical	207
Main engine builder location	categorical	37
Safety management company location	categorical	5
Beneficial owner company location	categorical	5
Maritime expertise		
Company presence and years of existence	categorical	6
Previous histories:		
Previous inspections, detentions, incidents (VS, S, and LS)	continuous	6
Previous changes in ship particulars		4
Interaction variables		
Shipyard country groups with age groups	categorical	8
Total variables evaluated		**588**

The variable groups that are used are standard ship particulars such as ship type, age, size (GRT), classification society, flags, main engine designer and builder besides the country of country of built (grouped into four groups) and interaction effects with age groups (0-2 and above 14 years of age representing high age risk while 3-14 years of ship age represent low age risk). Changes of ship particulars overtime such as flag changes, beneficial ownership changes, DoC company changes and class changes within 3 years prior to event date of interest (detention or incident) are considered.

Safety Management Companies (DoC company) and group beneficial owner country of location are classified according to income and their years of existence as proxy to experience. Due to the number of variables, the safety management company location and the beneficial owner locations had to be regrouped based on the World Bank classification of countries

indicating their income/development as follows: 1) high income, 2) upper middle income, 3) lower middle income, 4) low income and 5) unknown. Maritime expertise is expressed as concentration of ownership companies, DoC companies, main engine builders, main engine designers in a particular country. Lagged inspection, deficiency and incident history of the vessel (within 1 year prior to event date) is added to account for the vessel history with respect to those covariates.

The out of sample data for 2020 consists of the same variables as those present in the training data used for estimating the random forests. New input data is processed using a custom-made software written in PHP and is transformed in such a way that they matched the information used to estimate the models The empirical data to evaluate the predictions were manually reclassified using the same methodology as with the training data to be able to assign the correct degree of seriousness (VSS) ad to identify the first event of a chain of events with respect to incident types.

3. Methods, model variants and evaluation metrics used

3.1. Classification Trees and Random Forests

Classification trees (Breiman et al, 1984) arrive at predictions by sequentially splitting the data based on "tests" for individual variables. That is, splits, usually binary, are created by evaluating statements on individual variables, for example: “age” > 10. Predictions are made by considering the sequence of tests and choosing the majority class in the final data partition. Constructing a classification tree boils down to finding a sequence of tests that leads to accurate predictions. A common way to do this is to adopt a so-called greedy algorithm that determines the variable and test in such a way that a split (partitioning) leads to the largest immediate reduction of variance in the response variable. For a categorical response variable, a common measure for such variance (typically referred to in classification tree context as impurity), is the Gini index. The Gini index for a classification problem involving *C* classes, is defined as

$$Gini = \Sigma_c^C \ (\hat{p}_c(1 - \hat{p}_c))$$

where $\hat{p}_c$ denotes the observed proportion in class *c*. In building the classification tree, a weighted average of the Gini indices is calculated for each possible partitioning. From all these options, the partitioning is chosen that results in the greatest improvement (decrease) of this weighted average. Note that, for a binary classification problem (*C=2*), the maximum value of the Gini index is *0.5* and corresponds to the situation in which the observations are equally distributed over the two classes. The smallest value for the Gini index is zero, which is attained when the partitioning contains only values of the same class.

Classification trees are fast, unbiased, and allow for an easy interpretation. However, classification trees are known to be instable. In particular, small changes in the data can heavily impact the classifications. In other words, although unbiased predictions are obtained, the prediction variance is large, making these predictions less reliable. To overcome this problem, Breiman (1996; 2001) proposed to create multiple tree classifiers and combine the results of such a so-called ensemble to generate predictions. A popular ensemble method that combines the results of classification trees into one superior classifier is a random forest (Breiman, 2001). Random forests (RF) combine results of several classification trees, obtained by building trees on samples of the (training) data and generally provide the best performing prediction methods.

Random forests are typically constructed by using bootstrap samples. That is, from the original *n* observations in the training data, *B* samples of size *n* are drawn with replacement. Then, on each of these bootstrap samples, a tree is fitted resulting in *B* such classification trees. This "forest" of *B* classification trees can be used to make predictions by parsing data through the *B* trees, resulting in *B* predictions which are then combined into one final prediction by

selecting the most often predicted class. Alternatively, one can calculate for each class the proportions of predicted classes by the *B* trees and take the average of these proportions. It should be noted, however, that the goal of each single classification tree is to arrive at “pure” distributions in its leaves. The aim is to arrive at final distributions where all observations are of the same class. Hence, the proportions in the final leaves are biased towards zero and one. Consequently, for a binary classification problem, the proportions obtained in a random forest are biased towards zero and one. Nonetheless, the proportions do provide a measure for the certainty with which a certain class can be predicted and, although they may not correctly reflect probabilities, the ranking of proportions is meaningful and can be used to assess confidence of predictions.

It can be shown (Breiman, 2001) that combining multiple weak classifiers is effective if individual classifiers are independent. By construction, classifiers constructed on bootstrap samples (i.e., based on sampling with replacement) are not independent. Moreover, when certain variables are strongly correlated to the outcome (response) variable, such variables tend to be selected as splitting variables in the early stages of tree construction. Hence, differences in trees and predictions based on the different bootstrap samples may be limited as will be the effect on final predictions.

To increase independence among trees, and hence obtain better and more stable predictions, Breiman (2001) proposed to "decorrelate" trees by adding a random feature set selection step. That is, instead of considering all variables in deciding which variable to split on (and hence selecting the "best" from all available variables), only a random subset of variables is considered for each split of the tree. From this random subset, the "best" variable is selected. Hence, in addition to creating *B* bootstrap samples, the trees constructed for each of these samples, are created by considering different random samples of predictor variables in each split. Constructing an ensemble in this fashion creates the random forest.

A random forest consists of *B* trees, and each tree is constructed by recursive partitioning using subsets of *m* variables at each split. The *B* and *m* can be seen as parameters of the random forest. Values for these parameters must be decided upon before construction of the final forest. Often such parameters are referred to as tuning parameters as there is no obvious way to select optimal values for them. Typically, a "tuning" procedure is used in which different plausible values are assessed after which the "best" value of such a set is selected. For random forest, Breiman (2001) showed that predictions cannot become worse when increasing the number of trees *B.* Therefore, tuning of this number *B* amounts to finding a minimum size of the forest.

For the size of the random subset, the default value in classification scenarios, equals the squared root of the number of features. However, this value can be "tuned" by considering a range of values and observing (prediction) performances for the different values. A convenient property of random forests is that by using bootstrap samples, out-of-sample evaluation can easily be assessed by considering observations not included in the separate tree constructions. These observations are typically referred to as out-of-bag. Keeping track of these out-of-bag observations throughout the ensemble makes it possible to assess out-of-sample (i.e., prediction) performance of a forest without the necessity of creating a separate train-validation split of the data. Hence, one can use the out-of-bag observations to tune parameters for a random forest.

As a random forest consists of many tree classifiers, a straightforward interpretation of the model is not possible. There are, however, measures to assess which variables play an important role in the forest and this is shown in Section 4.

3.2. Threshold adjustment, weighting and sampling

Class-imbalance, that is, one class occurs (much) more often than the other, leads to underestimation of the probability for the minority class. In the case of classification trees (and ensemble methods using such trees) this problem can be readily understood as the Gini index

used to create the splits takes on its minimum or best value, when all observations belong to the same class. Hence, the Gini index for an imbalanced data set is close to its minimum at the start and it may be difficult to improve on this. There are several ways to correct for this problem and three different approaches aimed at resolving this issue are considered. – threshold adjustment, weighting and sampling.

If the desired outcome is simply the class (zero or one), one could change the threshold used to assign predictions to classes. For example, instead of predicting a one if the estimated probability is larger than 0.5, one could predict a one if the predicted probability is larger than the "base-rate". That is, the proportion of ones in the sample (in our case the average incident rate for each endpoint of interest). This type of correction does not alter the predicted probabilities and as such does not correct for the downwards biased probabilities.

Another approach to account for the imbalance problems is by adding weights either to the observations, that is, use case weights and hence assigning more/less importance to certain observations, or by incorporating such weights in the objective function. That is, in the calculation of the Gini index.

A third way to deal with the class imbalance, is to create a (more) balanced sample. That is, rather than using the actual sample, a sample is created for which the proportions of zeros and ones are 50/50. The easiest way to create such a sample, is by selecting all observations of the minority class and selecting a sample of the same size of the majority class. This is called "undersampling" as not all observations are used. Alternatively, one can draw (with replacement) from the minority class until the sample sizes are equal or more balanced. This approach is called upsampling. Note that, upsampling results in a data set with duplicated observations. There also exist variants in which, rather than simply resampling the original data, "noise" is added to allow for some variation in the upsampled values. Upsampling methods are not explicitly considered in the present approach. It is not difficult to see that case weighting and sampling are closely related but they are not identical as in the weighted variant all observations are still present, whereas undersampling removes several observations from the training data.

For random forest, there is an additional option that can be used to mitigate the loss of observations incurred by undersampling. Recall that to build a random forest, sampling is used. The samples are typically drawn with replacement from the training data. That is, the trees are constructed based on so-called bootstrap samples. However, the sampling scheme in random forests can be adapted in such a way that the trees are built on different samples. Furthermore, different samples to be balanced can be chosen. That is, for each tree a balances sample from the complete training data is drawn. In that way, the loss of data resulting from undersampling is reduced. This approach, which can be referred to as balanced random forests (BRF), is described in Chen et al (2004) and is implemented in the popular *R* packages *RandomForest* and *Ranger*.

Although the balanced random forest idea is not new and it is implemented in popular R packages, it does not yet appear to be used often in practice. In Yan et al (2021), balanced forests were used and outperformed the “regular” random forest in predictions of detentions.

3.3. Empirical analysis combinations and vote aggregation

Eight end points of interest are considered which are of interest for the various application areas explained in the introduction. One general incident type model (VSS) and seven incident types all with degree VSS. Table 2 provides a list of the model combinations that were used – a total of 18 variants for the 8 endpoints of interest, hence a total of 144 combinations. In addition to applying “ordinary” random forest (RF), various other options dealing with the class-imbalances explained in the previous section are used. The analysis considers random forests trained on balanced samples (RF_BS) using under sampling of the training data as well as the balanced random forest (BRF) as by Chen et al. (2004). Moreover, for all random forests,

aggregation of results is considered using both - majority voting as well as averaging of probabilities.

Recall that a random forest consists of *B* classification trees. A prediction is obtained by aggregating the *B* individual predictions. The two aggregation types considered are: 1) majority voting 2) probability aggregation. For majority voting, the class predictions of each tree (i.e., the predicted class in an individual tree) is considered and the proportions of predicted classes over all trees is calculated. For probability aggregation, the average predicted leaf proportions over all trees in the forest is calculated.

Table 2: Summary of model variants used

Group	Variant	Explanation
1	RF_m_16	Regular RF, m =16, majority votes aggregation
1	RF_p_16	Regular RF, m =16, probability votes aggregation
1	RF_m_32	Regular RF, m =32, majority votes aggregation
1	RF_p_32	Regular RF, m =32, probability votes aggregation
1	RF_m_8	Regular RF, m =8, majority votes aggregation
1	RF_p_8	Regular RF, m =8, probability votes aggregation
2	BRF_m_16	Balanced RF, m =16, majority votes aggregation
2	BRF_p_16	Balanced RF, m =16, probability votes aggregation
2	BRF_m_32	Balanced RF, m =32, majority votes aggregation
2	BRF_p_32	Balanced RF, m =32, probability votes aggregation
2	BRF_m_8	Balanced RF, m =8, majority votes aggregation
2	BRF_p_8	Balanced RF, m =8, probability votes aggregation
3	RF_BS_m_16	RF balanced training data, m =16, majority votes aggregation
3	RF_BS_p_16	RF balanced training data, m =16, probability votes aggregation
3	RF_BS_m_32	RF balanced training data, m =32, majority votes aggregation
3	RF_BS_p_32	RF balanced training data, m =32, probability votes aggregation
3	RF_BS_m_8	RF balanced training data, m =8, majority votes aggregation
3	RF_BS_p_8	RF balanced training data, m =8, probability votes aggregation

Notes: m= majority voting aggregation, p=probability votes aggregation, the numbers correspond to m, the number of variables considered for splitting. The default value for the data sets is 16. The number of trees for all models is 500

The model variants in Table 2 can be divided in three main groups: 1) Regular random forest (RF), 2) Balanced random forests (BRF) and 3) Random Forest on balanced samples (RF_BS). For each of these groups, both the results using the default number of variables to split on (*m* is equal to the square root of the number of variables in the training data) as well as this default divided by two and the default times two is considered. Although tuning of *m* for a regular random forest is relatively straightforward, given the size of our data set, tuning can be quite time consuming. Hence, the gains of tuning need to be considered with respect to the increased model estimation times. Initial experiments on tuning in trying different values for *m* and assessing its impact on predictions, turned out to be limited and it was decided to only consider the three options for *m* given in Table 2.

3.4. Evaluation metrics and out of sample data

To assess prediction performance, the data is split into train and test data. The training set or training data (from 2014 to 2019) is used for estimation of all models. The out of sample test data set from January to December 2020 is only used for evaluation. As first step using the outcome of the random forest routines, probabilities are estimated at a certain time period and are assumed to be valid for a maximum of 3 months (see Knapp and Heij, 2020). As second

step, empirical data for 2020 is matched with the estimated probabilities and evaluation metrics are calculated using the following setup:

- *Period 1: Probabilities as of January 2020 – empirical data from January to March 2020*
- *Period 2: Probabilities as of April 2020 – empirical data from April to June 2020*
- *Period 3: Probabilities as of July 2020 – empirical data from July to September 2020*
- *Period 4: Probabilities as of October 2020 – empirical data from November to December 2020*

As explained in the data section, input data comes from various sources and is cleaned and processed to calculate input parameters to estimate the probabilities for each end point of interest. It should be noted that the out of sample data for 2020 is of lesser data quality than the training data (the five-year period) used for estimating the model as missing data cannot be corrected. Furthermore, it also includes new vessels and the data for 2020 is influenced by the Covid19 pandemic in the shipping industry (Knapp 2021) and might not be the most representative year in shipping.

To assess performance, the observed or empirical data for 2020 is linked with the estimated probabilities for the four time periods described above where the probabilities are transferred into classifications by assigning observations to the class with the highest predicted probability. This is equivalent to assigning class 1 if the estimated probability is larger than 0.5 or some other value such as average incident rates. The prediction results can then be summarized in a so-called confusion matrix (Table 3) where predictions are related to the actual (observed) classes.

Table 3: Classification Table

	Predicted class = 0 (no-incident)	Predicted class = 1 (incident)
Actual class = 0 (no-incident)	True negative (TN)	False positive (FP)
Actual class = 1 (incident)	False negative (FN)	True positive (TP)

Based on the classifications in Table 3, various measures that are often used in classification problems (Yan et al (2021) are as follows: accuracy (TN+TP)/(TN+TP+FP+FN), recall (TP/(TP+FN)), precision (TP/(TP+FP)), and the F-measure (the harmonic mean of recall and precision). Although easy to calculate and interpret, these measures are in severely unbalanced setting (with many more non-incidents like is the case here) rather limited. Predicting no incident for all observation would in this unbalanced setting yield a very high accuracy, even though it fails to predict any incident.

Another measure for overall fit is the H-measure (Hand, 2009). The H-measure was designed to overcome an important flaw in the area under the receiver operating characteristics curve (AUC) measure, that is often used for model comparison. AUC uses different misclassification cost distributions for different classifiers. In a model comparison setting, this invalidates this measure as it implies that misclassifying observations is evaluated differently among the compared classifiers. How to assess the different types of misclassifications is a property of the problem at hand, rather than the chosen classification methods. For a fair comparison, the use of misclassification costs should not differ among methods. Hand (2009) proposed the H-measure as simple alternative to the AUC which lies between zero and one, with one indicating perfect classification, as an overall measure of fit. The default implementation of the H-measure in the R-package *hmeasure* as described by Anagnostopoulos et al (2019) is used in the current approach.

The main interest for the application areas is to either classify vessels (reduce false negative events as they can be very costly) or to estimate probabilities. Based on the various limitations of the evaluation metrics, this study considers the so-called top-decile lift. The top decile lift

considers the highest 10% estimated “probabilities” and compares the accuracy (hit rate) in that group to the base-rate, that is, the proportion of incidents in the complete sample. It assesses how much better predictions are in the top 10% compared to random guessing.

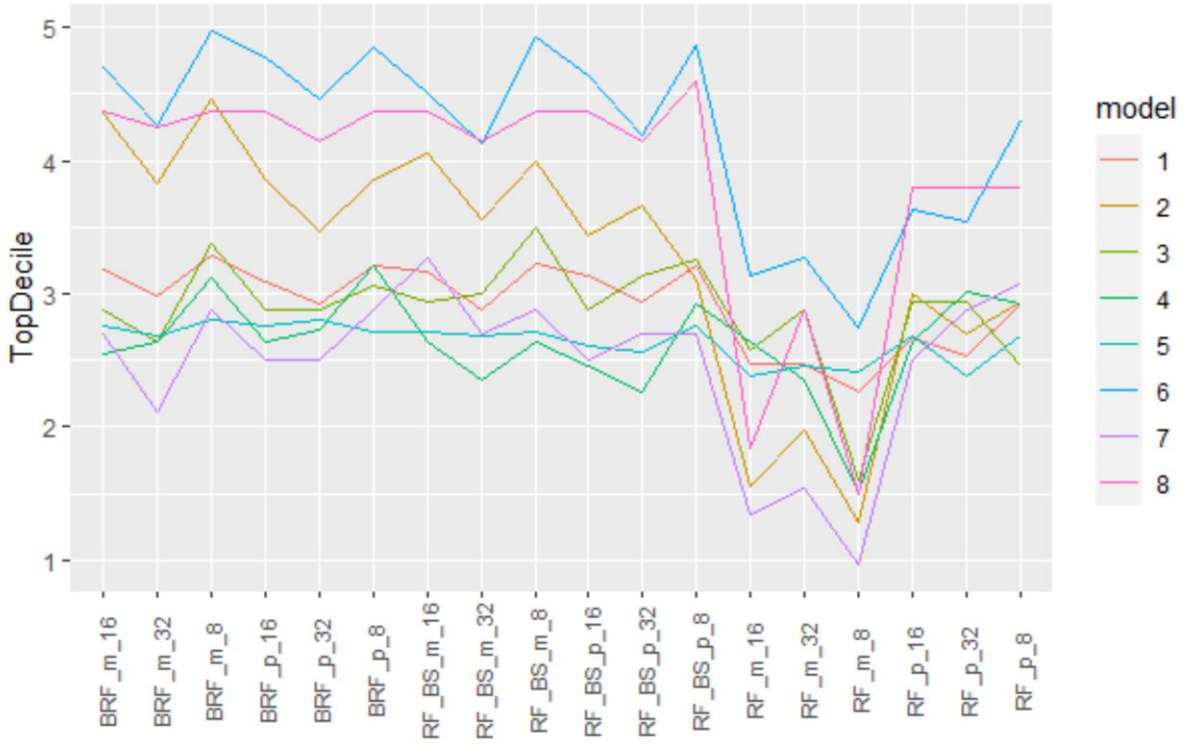
If the predicted probabilities are good, the top decile lift is large. This measure is useful for practitioners as it is not uncommon to target only those cases for which the predicted probabilities are largest. Moreover, the top-decile lift is somewhat insensitive to the bias in the probability estimates as it only considers the ranking of probabilities.

4. Results and Discussions

4.1. Selection of model variant to predict incidents

Figure 1 and Table 4 summarize the top decile lift for each of the model variants and Appendix A (Table A.1 to A.8) provides the other evaluation metrics explained in the previous section which are not further interpreted here due to the various limitations highlighted previously. The higher the top decile lift, the better the model variant performs on out of sample data for the year 2020. As can be seen from Figure 1 and Table 4, the regular random forest variants (RF), variants 1 to 6 in Table 2 are outperformed by the other variants.

Figure 1: Top decile lift for each model variant (out of sample data 2020)



Notes: 1= very serious and serious incidents, 2= pollution, 3= loss of life and injuries, 4= fire and explosions, 5= collisions and powered groundings, 6= mobility related incidents that include all failures that can lead to drifting (eg. engine failures, black outs, steering gear failures, propulsion failures), 7= hull related failures, 8= equipment failures. Abbreviations for the RF variants, please refer to Table 2

The best three model variants are highlighted in bold for each incident type in Table 4. To give an example on how to interpret the numbers in the Table, for method BRF_m_8 with a top decile lift of 3.293 for VSS incidents, the interpretation of this figure means that, if vessels with the top 10% highest predicted probabilities are targeted, the method successfully predicts 3.293 times as many VSS than random selection would do.

Table 4: Summary of results for incident (VSS) and incident types (VSS) – top decile lift for OOS 2020 data

	Model variants/endpoint of interest	1=VSS	2=POL	3=LIFE	4=FIRE	5=COLPOW	6=MOB	7=HULL	8=EQUIP
	Nr of 1's. (matrix): 2014 to 2019	**10,443**	**572**	**904**	**920**	**3,844**	**3,274**	**601**	**589**
	Nr of 1's. (OOS): 2020	**1,573**	**303**	**163**	**106**	**261**	**513**	**52**	**87**
	Incident rate (matrix): 2014 to 2019	0.014469	0.000792	0.001252	0.001275	0.005326	0.004536	0.000833	0.000816
	Incident rate (OOS): 2020	0.012159	0.002342	0.001260	0.000819	0.002018	0.003966	0.000402	0.000673
1	RF_m	2.467	1.551	2.577	2.641	2.375	3.138	1.346	1.839
2	RF_p	2.670	3.003	2.945	2.641	2.682	3.626	2.500	3.793
3	RF_m_32	2.473	1.980	2.883	2.358	2.452	3.275	1.538	2.874
4	RF_p_32	2.537	2.706	2.945	**3.019**	2.375	3.548	2.885	3.793
5	RF_m_8	2.257	1.287	1.595	1.509	2.414	2.749	0.962	1.494
6	RF_p_8	2.924	2.937	2.454	2.925	2.682	4.288	**3.077**	3.793
7	BRF_m	3.179	**4.356**	2.883	2.547	**2.759**	4.698	2.692	**4.368**
8	BRF_p	3.090	3.861	2.883	2.641	**2.759**	4.776	2.500	**4.368**
9	BRF_m_32	2.982	3.828	2.638	2.641	2.682	4.269	2.115	4.253
10	BRF_p_32	2.931	3.465	2.883	2.736	**2.797**	4.464	2.500	4.138
11	BRF_m_8	**3.293**	**4.455**	**3.374**	**3.113**	**2.797**	**4.971**	**2.885**	**4.368**
12	BRF_p_8	3.204	3.861	3.067	**3.208**	2.720	4.854	**2.885**	**4.368**
13	RF_BS_m	3.160	**4.059**	2.945	2.641	2.720	4.503	**3.269**	**4.368**
14	RF_BS_p	3.140	3.432	2.883	2.453	2.605	4.639	2.500	**4.368**
15	RF_BS_m_32	2.886	3.564	3.006	2.358	2.682	4.133	2.692	4.138
16	RF_BS_p_32	2.943	3.663	3.129	2.264	2.567	4.191	2.692	4.138
17	RF_BS_m_8	**3.229**	3.993	**3.497**	2.641	2.720	**4.932**	**2.885**	**4.368**
18	RF_BS_p_8	**3.217**	3.102	**3.252**	2.925	**2.759**	**4.873**	2.692	**4.598**

Note: VSS = very serious and serious incidents, POL = pollution, LIFE = loss of life and injuries, FIRE= fire and explosions, COLPOW= collisions and powered groundings, MOB = mobility related incidents that include all failures that can lead to drifting (eg. engine failures, black outs, steering gear failures, propulsion failures), HULL= hull related failures, EQUIP = equipment failures

In Table 4, one can see that the “regular” random forest models are mostly outperformed by the two balanced variants. It appears, that balancing, either within the forest (balanced random forest) or before fitting the model (random forest on balanced data) is beneficial. Comparing the two balancing variants, one can observe that for most endpoints of interest, a relatively small value for *m* performs well. A small value for *m* means that a smaller random subset is used in each split. Smaller values of this parameter make models more robust against overfit.

Although there is not one method/model that is best in all scenarios, the balanced random forest with *m* equal to 8, and majority voting to aggregate results, performs in the top three for all endpoints of interest. In addition, differences between the three best methods for each endpoint of interest are typically very small.

As this study is the first in its kind, it is recommended to implement the three best methods going forward and to re-evaluate them using new out of sample data in the future. In practice, this could be a yearly evaluation routine besides the re-estimation of the forests themselves which could be done every 3 to 5 years or at shorter time intervals. The risk factors do not change dramatically and fast over time in the shipping industry but the composition and quality of the new input data will change every month.

Based on the results, it is not recommended to use regular random forest variants (RF) for any of the application areas. The selection between balanced random forests (BRF variants) and random forests based on balanced training data (RF_BS) depends on the endpoint of interest. Note from Table 4, that the empirical data for 2020 some incident types is very low such as 52 for hull related failures or 87 for equipment related failures and should therefore be interpreted with caution. Given the application areas of interest that are considered, the selection of the end models between BRF and RF_BS variants will depend on their end use application.

For the area related to targeting of vessels and domain awareness where percentile ranks can be used as vessels are benchmarked, variant BRF_m_8 (majority votes) can be recommended. For this variant, the main purpose is to reduce false negative events and to indicate high risk vessels. The spread of the estimated probabilities across the world fleet is less important since the probabilities are not used for any further calculation. In addition, for loss of life and injuries, RF_BS_m_8 performs well, while for hull related failures, RF_BS_m outperforms the other variants. For equipment failures, RF_BS_p_8 can be recommended but others are not far behind.

For the application areas related to improved risk assessment and the quantification of incident costs probability aggregation may be more appropriate if probability estimates are required, hence variant BRF_m_p (probability aggregation) can be recommended with limitations. Differences in performances between the different aggregation methods for all end points appear to be small. However, assessing the quality of the exact probability estimates is difficult and none of the performance measures directly capture this aspect. Calibration plots, where the observed frequencies are compared to the expected frequencies corresponding to the estimated probabilities, are a typically a useful tool to appraise probability estimates. However, for applications, such calibration plots, when applied to the out-of-sample data, merely illustrate how the estimated probabilities, for all models, are biased towards the majority class (no incident) due to the severe imbalance for all end points. A reliable assessment of the actual probability estimates, rather than the ranking of such probabilities, for the different models is therefore not possible.

4.2. Importance of covariates

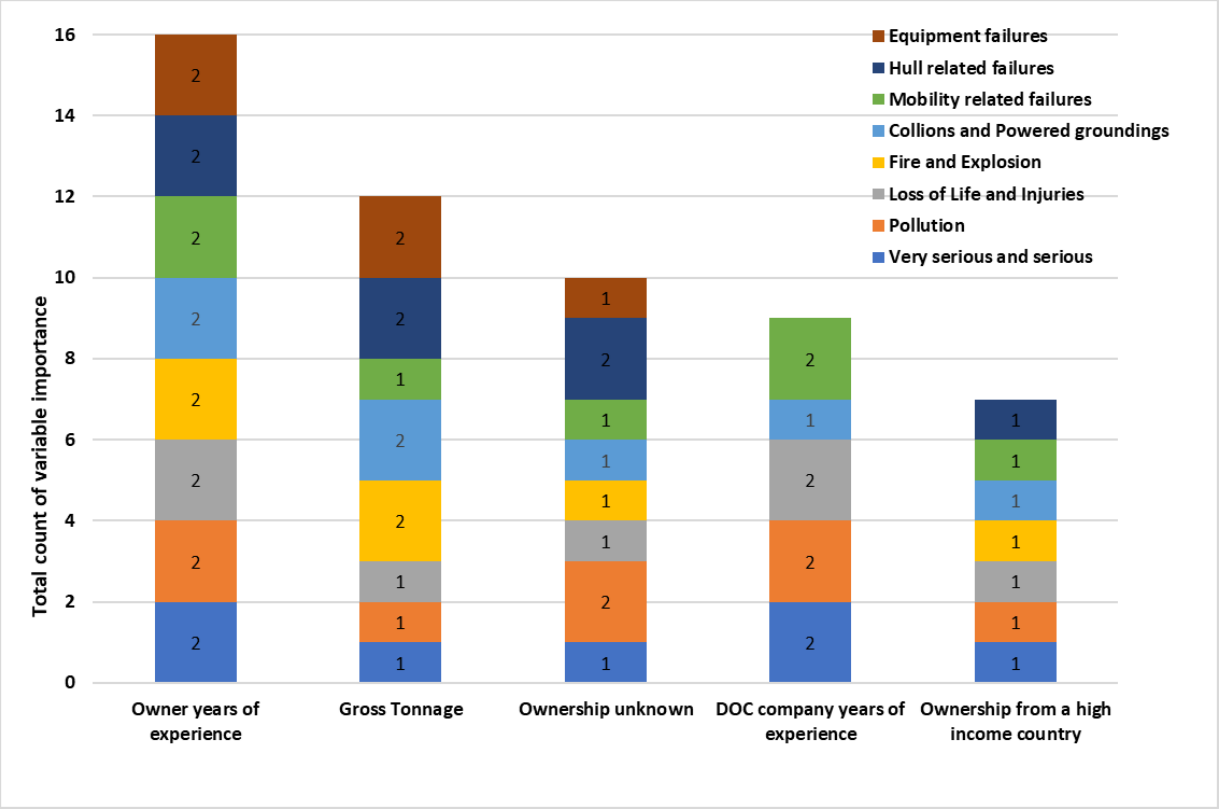
This section looks at the importance of the covariates for method BRF_m_8 (balanced random forest majority votes). One way to do so is by calculating how much a variable contributed to the construction of the forest. This is done by calculating for each variable how much it contributed to the total decrease of variance, as measured by the decrease in the Gini index. The resulting measure is referred to as Gini importance.

Alternatively, one can consider the impact of a variable on the final prediction accuracy. The influence of each variable on the predictions is considered by comparing the random forest predictions with predictions obtained when the values of a variable are randomly permuted. If predictions are not affected by such permutations, this indicates that the variable has no impact on the final predictions. On the other hand, if predictions get worse, the variable is important in making correct predictions. This type of importance is denoted here as permutation importance. To assess the impact on predictions, the OOS observations are used.

Together these importance measures give an indication about which variables played a role in the construction (Gini importance) and predictions (permutation importance) of the forest. Note that the importance measures do not provide information about the direction of the effect. That is, without further study, one does not know whether higher/lower values for a variable lead to an increase or decrease in the probabilities for a certain class.

For each model type, two plots are created and are provided in Appendix B along with a Table which explains the abbreviations of the covariates shown in the Figures. To provide an example of how to interpret the plots, refer to for instance to Figure B1 in Appendix B (model VSS). On can see that the variable OWN_YEX (Ownership years of experience), is both important when considering building the forest (Gini importance) as in accurately predicting incidents. On the other hand, the variable GRT, is important when constructing the forest, indicating that the variable is often selected to split on and leads to a large reduction in the Gini index. However, the permutation importance indicates that this variable is less important for making accurate predictions. The reason for this could be that when the variable is ignored, other variables pick up on similar effects. Figures 2 summarizes the importance of covariates based on the top 5 most influential covariates – either based on the Gini importance or the permutation importance.

Figure 2: Summary of variable importance by incident type: top 5 covariates



The counts are based on the plots for each incident type shown in Appendix B where the overall total of 80 points (5 times 2 times eight models) can be assigned. The most influential variables are mostly related to beneficial ownership and DOC company variables, GRT and Age. Their importance is also similar across all incident types where GRT is most influential for pollution, loss of life and fire and explosion based on the Gini importance.

The results underline the importance of the quality of beneficial ownership and the DOC company – both based on the Gini importance and the permutation importance. Standard variables that are believed to be important such as ship type, flag and classification society do not necessarily come out as the top influential covariates for prediction when using random forests variants but are found within the top 20 influential covariates of all incident types along with shipyard countries, main engine designer and main engine builder. Previous inspection histories (previous number of inspections and deficiencies) and incidents (less serious incidents) are influential for all incident type models within the top 20.

5. Conclusions

This study explores a total of 144 model variants (18 random forest variants of 8 endpoints of interest) with the aim to predict incident type probabilities that can be used for various application areas of interest in the maritime industry. Secondary area of importance is to highlight the importance of covariates.

The study also highlights the importance of the data related aspects as those are always complex and challenging in the maritime industry since the endpoints of interest are rare events and data is scares and often of poor quality and not easy to access. For this reason, data sources for incidents must always be combined from multiple sources to reduce bias due to underreporting of incidents.

The results are mixed and to deal with the severe imbalance in our variable of interest, balancing is recommended. Selection between balanced random forests (BRF) and random forests using balanced training data (RF_BS) may depend on the endpoint of interest and application area to be used. Overall, balanced random forest variants with a relatively low number for the number of variables to consider for splitting (i.e., m=8) perform best. For the area related to targeting of vessels and enhanced domain awareness, variant BRF_m_8 (majority votes) can be recommended. For this application area, probabilities can be converted into percentile ranks for benchmarking of vessels and the main goal is to minimize false negative events. In using the probabilities from this variant, it is less important whether probabilities are biased toward zero or one. For the application areas related to improved risk assessment and the quantification of incident costs, it is important to calculate probabilities and to get the spread the world fleet. For such applications, variant BRF_p_8 (probability aggregation) may be more appropriate from a theoretical point of view.

Selection of the best performing model is often not clear and it is recommended to implement the best 3 or 5 performing model variants and re-evaluate them after some time using new out of sample data. This could be a yearly exercise as the fleet composition and market conditions change over time.

With respect to the importance of the covariates, the most important covariate groups are related to beneficial ownership, the DOC company, GRT and Age and the importance of these factors is similar across the endpoint of interest. The standard variables such as ship type, flag, classification society are not that important when using random forests. Previous inspection and incident histories are relevant for all incident types within the top 20 most influential variables.

Future research areas to be included are related to detention in addition to incident types based on global inspection data.

References

Anagnostopoulos, C. Hand, D.J., and Adams, N.M. (2019), Measuring classification performance: the hmeasure package. https://cran.r-project.org/web/packages/hmeasure/vignettes/hmeasure.pdf

Bijwaard G and Knapp S (2009), Analysis of Ship Life Cycles – The Impact of Economic Cycles and Ship Inspections, Marine Policy, 33(2): 350-369.

BRISK (2012), Project on sub-regional risk of spill of oil and hazardous substances in the Baltic Sea, Risk Method note, Danish Admiralty, http://www.brisk.helcom.fi/.

Breiman, L. Bagging predictors. Mach Learn **24,** 123–140 (1996). https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random Forests. Machine Learning **45,** 5–32 (2001). https://doi.org/10.1023

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984), Classification And Regression Trees (1st ed.). Routledge.

Chen, C., Liaw, A., and Breiman L. (2004), Using random forest to learn imbalanced data, 110, Tech-report of University of California, Berkeley, pp. 1–12.

Det Norske Veritas (2013), North East Shipping Risk Assessment, Report prepared for the Australian Maritime Safety Authority, Report Nr 14OFICX-4.

Hand, D.J. (2009), Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning 77: 103–123.

Hansen PH (2007), IWRAP II, Working Document, Basic Modelling Principles for Prediction of Collision and Grounding Frequencies, Technical University of Denmark.

Greene H W, 2000, Econometric Analysis, Fourth Edition, Prentice Hall, New Jersey.

Heij C, Bijwaard G and Knapp S (2011), Ship Inspection Strategies: Effects on Maritime Safety and Environmental Protection, Transportation Research Part D, 16, 42-48.

Heij C, Knapp S (2012), Evaluation of safety and environmental risk at individual ship and company level, Transportation Research Part D 17 228–236.

Heij, C, Knapp S (2018), S, Predictive power of inspection outcomes for future shipping accidents – an empirical appraisal with special attention for human factor aspects, Maritime Policy and Management, Volume 45 Issue 5, 1–18.

Heij C, Knapp S (2019), Shipping Inspections, Detentions, and Incidents: An Empirical Analysis of Risk Dimensions, Maritime Policy and Management - Volume 47, Issue 7.

IMO (2000), MSC/Circ. 953, MEPC/Circ. 372, Reports on Marine Casualties and Incidents, Revised harmonized reporting procedures, adopted 14th December 2000, IMO, London.

Knapp S, (2006), The Econometrics of Maritime Safety – Recommendations to improve safety at sea, Doctoral Thesis, Erasmus University Rotterdam.

Knapp, S, Franses, P.H (2007), A global view on port state control - econometric analysis of the differences across port state control regimes. Maritime Policy and Management 34, 453-483.

Knapp S, Bijwaard G and Heij C (2011), Estimated Incident Cost Savings in Shipping due to Inspections, Accident Analysis and Prevention, 43: 1532–1539.

Knapp S and Heij C (2017), Evaluation of total risk exposure and insurance premiums in the maritime industry, Transportation Research Part D: Transport and Environment (2017), Volume 54, 321–334.

Knapp S, Heij C (2020), Improved strategies for the maritime industry to target vessels for inspections and to select inspection priority areas, Safety 2020, 6, 18.

Knapp S (2021), Quantification and analysis of risk exposure in the maritime industry: Averted incident costs due to inspections and the effect of SARS-CoV-2 (Covid19), Safety 2021, 7(2), 43.

Lemmens, A., & Croux, C. (2006), Bagging and boosting classification trees to predict churn. Journal of. Marketing Research, 43(2), 276-286.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

Vander Hoorn S, Knapp S (2015), A multi-layered risk exposure assessment approach for the shipping industry, Transportation Research Part A: Policy and Practice (2015) Volume 78, 21–33.

Vander Hoorn S, Knapp S (2019), Predicting traffic and risk exposure in the maritime industry, Safety 2019, 5(3), 42.

Verbeek, M, 2008, A Guide to Modern Econometrics, Third edition, John Wiley and Sons Ltd, Chichester.

Yan R, Wang S, Peng C (2021), An Artificial Intelligence Model Considering Data Imbalance for Ship Selection in Port State Control Based on Detention Probabilities, Journal of Computational Science 48, 101257.

Appendix A: Evaluation metrics

Table A.1: Evaluation metrics: incident (VSS)

		OOS sample 2020: 1,573 observations							
	Model variants	**H**	**AUC**	**ER**	**Precision**	**TPR**	**FPR**	**F**	**TopDecile**
1	RF_m	0.08272	0.65723	0.00309	na	0.000	0.000	na	2.467
2	RF_p	0.14274	0.73653	0.00309	na	0.000	0.000	na	2.670
3	RF_m_32	0.08662	0.66795	0.00309	0.00000	0.000	0.000	0.00000	2.473
4	RF_p_32	0.12475	0.72123	0.00309	na	0.000	0.000	na	2.537
5	RF_m_8	0.05933	0.61538	0.00309	na	0.000	0.000	na	2.257
6	RF_p_8	0.16243	0.74879	0.00309	na	0.000	0.000	na	2.924
7	BRF_m	0.15973	0.74476	0.16795	0.00838	0.455	0.167	0.01645	3.179
8	BRF_p	0.16911	0.75396	0.17144	0.00828	0.459	0.170	0.01627	3.090
9	BRF_m_32	0.15034	0.74014	0.16315	0.00793	0.418	0.162	0.01557	2.982
10	BRF_p_32	0.15840	0.74657	0.16722	0.00791	0.427	0.166	0.01554	2.931
11	BRF_m_8	0.17040	0.75244	0.18821	0.00806	0.491	0.187	0.01586	3.293
12	BRF_p_8	0.18020	0.75976	0.18836	0.00807	0.492	0.187	0.01589	3.204
13	RF_BS_m	0.15795	0.74386	0.17480	0.00796	0.449	0.174	0.01564	3.160
14	RF_BS_p	0.16625	0.75161	0.17517	0.00806	0.456	0.174	0.01585	3.140
15	RF_BS_m_32	0.14939	0.73792	0.16969	0.00799	0.438	0.168	0.01570	2.886
16	RF_BS_p_32	0.15552	0.74356	0.16899	0.00804	0.439	0.168	0.01579	2.943
17	RF_BS_m_8	0.16664	0.75201	0.19424	0.00797	0.501	0.193	0.01569	3.229
18	RF_BS_p_8	0.18086	0.75819	0.19562	0.00797	0.505	0.195	0.01569	3.217

Abbreviations: H = H measure, AUC = area under the curve, ER = error rate, TPR= true positive rate, FPR= false positive rate, F= harmonic mean of recall (TPR) and precision, na= not available

Table A.2: Evaluation metrics: pollution (VSS)

		OOS sample 2020: 303 observations							
	Model variants	**H**	**AUC**	**ER**	**Precision**	**TPR**	**FPR**	**F**	**TopDecile**
1	RF_m	0.00971	0.52907	0.00060	na	0.000	0.000	na	1.551
2	RF_p	0.18988	0.76484	0.00060	na	0.000	0.000	na	3.003
3	RF_m_32	0.02298	0.54893	0.00060	na	0.000	0.000	na	1.980
4	RF_p_32	0.13594	0.70706	0.00060	na	0.000	0.000	na	2.706
5	RF_m_8	0.00441	0.50883	0.00060	na	0.000	0.000	na	1.287
6	RF_p_8	0.20609	0.77739	0.00060	na	0.000	0.000	na	2.937
7	BRF_m	0.28653	0.82061	0.27564	0.00162	0.749	0.276	0.00322	4.356
8	BRF_p	0.26504	0.81046	0.28544	0.00159	0.766	0.285	0.00318	3.861
9	BRF_m_32	0.24978	0.80305	0.27538	0.00157	0.726	0.275	0.00313	3.828
10	BRF_p_32	0.25264	0.80504	0.28114	0.00159	0.752	0.281	0.00318	3.465
11	BRF_m_8	0.30099	0.82524	0.30471	0.00165	0.845	0.305	0.00329	4.455
12	BRF_p_8	0.27545	0.81379	0.30592	0.00158	0.812	0.306	0.00315	3.861
13	RF_BS_m	0.27882	0.81725	0.29181	0.00159	0.779	0.292	0.00317	4.059
14	RF_BS_p	0.24833	0.80340	0.29917	0.00154	0.776	0.299	0.00308	3.432
15	RF_BS_m_32	0.23962	0.79917	0.29262	0.00150	0.736	0.293	0.00298	3.564
16	RF_BS_p_32	0.25056	0.80240	0.29448	0.00153	0.759	0.295	0.00306	3.663
17	RF_BS_m_8	0.28565	0.81956	0.30177	0.00160	0.812	0.302	0.00319	3.993
18	RF_BS_p_8	0.26930	0.80904	0.30889	0.00158	0.822	0.309	0.00316	3.102

Abbreviations: H = H measure, AUC = area under the curve, ER = error rate, TPR= true positive rate, FPR= false positive rate, F= harmonic mean of recall (TPR) and precision, na= not available

Table A.3: Evaluation metrics: loss of life and injuries (VSS)

		OOS sample 2020: 163 observations							
	Model variants	**H**	**AUC**	**ER**	**Precision**	**TPR**	**FPR**	**F**	**TopDecile**
1	RF_m	0.07039	0.58430	0.00032	na	0.000	0.000	na	2.577
2	RF_p	0.16374	0.72677	0.00032	na	0.000	0.000	na	2.945
3	RF_m_32	0.08780	0.60248	0.00032	na	0.000	0.000	na	2.883
4	RF_p_32	0.14325	0.69761	0.00032	na	0.000	0.000	na	2.945
5	RF_m_8	0.04392	0.54808	0.00032	na	0.000	0.000	na	1.595
6	RF_p_8	0.19318	0.74168	0.00032	na	0.000	0.000	na	2.454
7	BRF_m	0.18241	0.74049	0.23022	0.00078	0.558	0.230	0.00155	2.883
8	BRF_p	0.19276	0.74064	0.21957	0.00078	0.534	0.219	0.00155	2.883
9	BRF_m_32	0.16781	0.72444	0.20615	0.00073	0.472	0.206	0.00147	2.638
10	BRF_p_32	0.17329	0.72791	0.20381	0.00075	0.479	0.204	0.00150	2.883
11	BRF_m_8	0.20800	0.75755	0.27537	0.00079	0.681	0.275	0.00158	3.374
12	BRF_p_8	0.20400	0.75623	0.25417	0.00081	0.644	0.254	0.00162	3.067
13	RF_BS_m	0.18800	0.74308	0.22828	0.00081	0.577	0.228	0.00161	2.945
14	RF_BS_p	0.18936	0.74294	0.22332	0.00077	0.540	0.223	0.00155	2.883
15	RF_BS_m_32	0.17134	0.73306	0.20941	0.00079	0.515	0.209	0.00157	3.006
16	RF_BS_p_32	0.17977	0.73432	0.21060	0.00082	0.540	0.211	0.00164	3.129
17	RF_BS_m_8	0.20300	0.75633	0.26552	0.00079	0.656	0.265	0.00158	3.497
18	RF_BS_p_8	0.20386	0.75526	0.25091	0.00079	0.620	0.251	0.00158	3.252

Abbreviations: H = H measure, AUC = area under the curve, ER = error rate, TPR= true positive rate, FPR= false positive rate, F= harmonic mean of recall (TPR) and precision, na= not available

Table A.4: Evaluation metrics: fire and explosions (VSS)

		OOS sample 2020: 106 observations							
	Model variants	**H**	**AUC**	**ER**	**Precision**	**TPR**	**FPR**	**F**	**TopDecile**
1	RF_m	0.05538	0.58201	0.00021	na	0.000	0.000	na	2.641
2	RF_p	0.12154	0.68804	0.00021	na	0.000	0.000	na	2.641
3	RF_m_32	0.06966	0.60169	0.00021	na	0.000	0.000	na	2.358
4	RF_p_32	0.12184	0.67260	0.00021	na	0.000	0.000	na	3.019
5	RF_m_8	0.01420	0.53031	0.00021	na	0.000	0.000	na	1.509
6	RF_p_8	0.14277	0.71280	0.00021	na	0.000	0.000	na	2.925
7	BRF_m	0.15574	0.71324	0.17776	0.00053	0.453	0.178	0.00106	2.547
8	BRF_p	0.15931	0.71607	0.18510	0.00056	0.500	0.185	0.00112	2.641
9	BRF_m_32	0.13826	0.70737	0.17657	0.00050	0.425	0.176	0.00100	2.641
10	BRF_p_32	0.13777	0.70549	0.18098	0.00052	0.453	0.181	0.00104	2.736
11	BRF_m_8	0.17253	0.72099	0.19484	0.00051	0.481	0.195	0.00103	3.113
12	BRF_p_8	0.15817	0.71685	0.19913	0.00049	0.472	0.199	0.00099	3.208
13	RF_BS_m	0.14257	0.70438	0.20898	0.00048	0.481	0.209	0.00096	2.641
14	RF_BS_p	0.13741	0.70707	0.21789	0.00050	0.519	0.218	0.00099	2.453
15	RF_BS_m_32	0.11255	0.68892	0.20901	0.00046	0.462	0.209	0.00092	2.358
16	RF_BS_p_32	0.12724	0.69383	0.21074	0.00048	0.481	0.211	0.00095	2.264
17	RF_BS_m_8	0.13793	0.70631	0.21801	0.00048	0.500	0.218	0.00095	2.641
18	RF_BS_p_8	0.14408	0.70820	0.23305	0.00044	0.490	0.233	0.00088	2.925

Abbreviations: H = H measure, AUC = area under the curve, ER = error rate, TPR= true positive rate, FPR= false positive rate, F= harmonic mean of recall (TPR) and precision, na= not available

Table A.5: Evaluation metrics: collisions and powered groundings (VSS)

		OOS sample 2020: 261 observations							
	Model variants	**H**	**AUC**	**ER**	**Precision**	**TPR**	**FPR**	**F**	**TopDecile**
1	RF_m	0.05861	0.61359	0.00051	na	0.000	0.000	na	2.375
2	RF_p	0.11338	0.69900	0.00051	na	0.000	0.000	na	2.682
3	RF_m_32	0.07617	0.63798	0.00051	na	0.000	0.000	na	2.452
4	RF_p_32	0.10985	0.69915	0.00051	na	0.000	0.000	na	2.375
5	RF_m_8	0.06432	0.60052	0.00051	na	0.000	0.000	na	2.414
6	RF_p_8	0.13855	0.71771	0.00051	na	0.000	0.000	na	2.682
7	BRF_m	0.12733	0.70841	0.18102	0.00120	0.425	0.181	0.00240	2.759
8	BRF_p	0.13161	0.71322	0.18270	0.00126	0.448	0.183	0.00251	2.759
9	BRF_m_32	0.11451	0.70125	0.17064	0.00121	0.402	0.170	0.00241	2.682
10	BRF_p_32	0.12122	0.70751	0.16935	0.00121	0.398	0.169	0.00241	2.797
11	BRF_m_8	0.13217	0.71428	0.20732	0.00119	0.483	0.207	0.00238	2.797
12	BRF_p_8	0.13239	0.71527	0.20545	0.00119	0.475	0.205	0.00237	2.720
13	RF_BS_m	0.12716	0.70578	0.18769	0.00125	0.456	0.188	0.00248	2.720
14	RF_BS_p	0.12651	0.70952	0.18654	0.00125	0.456	0.186	0.00250	2.605
15	RF_BS_m_32	0.11372	0.69669	0.17634	0.00119	0.410	0.176	0.00238	2.682
16	RF_BS_p_32	0.11787	0.70286	0.17687	0.00126	0.433	0.177	0.00250	2.567
17	RF_BS_m_8	0.13580	0.71063	0.20695	0.00117	0.471	0.207	0.00233	2.720
18	RF_BS_p_8	0.13703	0.71255	0.20717	0.00119	0.483	0.207	0.00238	2.759

Abbreviations: H = H measure, AUC = area under the curve, ER = error rate, TPR= true positive rate, FPR= false positive rate, F= harmonic mean of recall (TPR) and precision, na= not available

Table A.6: Evaluation metrics: mobility related failures (VSS)

		OOS sample 2020: 513 observations							
	Model variants	**H**	**AUC**	**ER**	**Precision**	**TPR**	**FPR**	**F**	**TopDecile**
1	RF_m	0.12500	0.66152	0.00101	na	0.000	0.000	na	3.138
2	RF_p	0.21693	0.77065	0.00101	na	0.000	0.000	na	3.626
3	RF_m_32	0.13712	0.68042	0.00101	na	0.000	0.000	na	3.275
4	RF_p_32	0.19017	0.75173	0.00101	na	0.000	0.000	na	3.548
5	RF_m_8	0.08622	0.61114	0.00101	na	0.000	0.000	na	2.749
6	RF_p_8	0.26138	0.79974	0.00101	na	0.000	0.000	na	4.288
7	BRF_m	0.26898	0.78974	0.11893	0.00431	0.509	0.119	0.00855	4.698
8	BRF_p	0.27789	0.79527	0.12191	0.00422	0.511	0.122	0.00837	4.776
9	BRF_m_32	0.24294	0.78132	0.11540	0.00395	0.452	0.115	0.00784	4.269
10	BRF_p_32	0.25251	0.78473	0.11694	0.00412	0.478	0.117	0.00816	4.464
11	BRF_m_8	0.28928	0.79892	0.13005	0.00421	0.544	0.130	0.00836	4.971
12	BRF_p_8	0.29233	0.80456	0.13323	0.00414	0.548	0.133	0.00822	4.854
13	RF_BS_m	0.25804	0.78859	0.12640	0.00409	0.513	0.126	0.00811	4.503
14	RF_BS_p	0.26606	0.79352	0.12606	0.00410	0.513	0.126	0.00813	4.639
15	RF_BS_m_32	0.23240	0.77898	0.12532	0.00387	0.481	0.125	0.00768	4.133
16	RF_BS_p_32	0.23673	0.78183	0.12556	0.00380	0.474	0.125	0.00755	4.191
17	RF_BS_m_8	0.28838	0.80057	0.13541	0.00407	0.548	0.135	0.00809	4.932
18	RF_BS_p_8	0.28710	0.80219	0.13869	0.00413	0.569	0.138	0.00820	4.873

Abbreviations: H = H measure, AUC = area under the curve, ER = error rate, TPR= true positive rate, FPR= false positive rate, F= harmonic mean of recall (TPR) and precision, na= not available

Table A.7: Evaluation metrics: hull related failures (VSS)

		OOS sample 2020: 52 observations							
	Model variants	**H**	**AUC**	**ER**	**Precision**	**TPR**	**FPR**	**F**	**TopDecile**
1	RF_m	0.00238	0.51283	0.00010	na	0.000	0.000	na	1.346
2	RF_p	0.10908	0.64831	0.00010	na	0.000	0.000	na	2.500
3	RF_m_32	0.01455	0.52749	0.00010	na	0.000	0.000	na	1.538
4	RF_p_32	0.09730	0.63460	0.00010	na	0.000	0.000	na	2.885
5	RF_m_8	0.00002	0.50084	0.00010	na	0.000	0.000	na	0.962
6	RF_p_8	0.12474	0.68511	0.00010	na	0.000	0.000	na	3.077
7	BRF_m	0.14616	0.70399	0.16092	0.00023	0.365	0.161	0.00046	2.692
8	BRF_p	0.15797	0.70570	0.16236	0.00024	0.385	0.162	0.00048	2.500
9	BRF_m_32	0.14135	0.70136	0.15825	0.00021	0.327	0.158	0.00042	2.115
10	BRF_p_32	0.13596	0.69658	0.15590	0.00023	0.346	0.156	0.00045	2.500
11	BRF_m_8	0.15582	0.69914	0.17379	0.00026	0.442	0.174	0.00052	2.885
12	BRF_p_8	0.16229	0.70085	0.16687	0.00026	0.423	0.167	0.00052	2.885
13	RF_BS_m	0.15878	0.68901	0.16269	0.00024	0.385	0.163	0.00048	3.269
14	RF_BS_p	0.14964	0.68797	0.15832	0.00026	0.404	0.158	0.00052	2.500
15	RF_BS_m_32	0.13680	0.68639	0.15713	0.00025	0.385	0.157	0.00050	2.692
16	RF_BS_p_32	0.13580	0.68398	0.15093	0.00022	0.327	0.151	0.00044	2.692
17	RF_BS_m_8	0.15698	0.68220	0.17333	0.00025	0.423	0.173	0.00050	2.885
18	RF_BS_p_8	0.16371	0.68645	0.16405	0.00026	0.423	0.164	0.00053	2.692

Abbreviations: H = H measure, AUC = area under the curve, ER = error rate, TPR= true positive rate, FPR= false positive rate, F= harmonic mean of recall (TPR) and precision, na= not available

Table A.8: Evaluation metrics: equipment related failures (VSS)

		OOS sample 2020: 87 observations							
	Model variants	**H**	**AUC**	**ER**	**Precision**	**TPR**	**FPR**	**F**	**TopDecile**
1	RF_m	0.06486	0.54622	0.00017	na	0.000	0.000	na	1.839
2	RF_p	0.22561	0.74761	0.00017	na	0.000	0.000	na	3.793
3	RF_m_32	0.10867	0.60416	0.00017	na	0.000	0.000	na	2.874
4	RF_p_32	0.18165	0.69087	0.00017	na	0.000	0.000	na	3.793
5	RF_m_8	0.02823	0.53083	0.00017	na	0.000	0.000	na	1.494
6	RF_p_8	0.24457	0.76565	0.00017	na	0.000	0.000	na	3.793
7	BRF_m	0.28467	0.79354	0.16425	0.00056	0.540	0.164	0.00112	4.368
8	BRF_p	0.28294	0.79272	0.16618	0.00054	0.529	0.166	0.00109	4.368
9	BRF_m_32	0.28937	0.79245	0.15873	0.00059	0.552	0.159	0.00119	4.253
10	BRF_p_32	0.28875	0.79092	0.16407	0.00056	0.540	0.164	0.00112	4.138
11	BRF_m_8	0.27816	0.79077	0.18828	0.00053	0.586	0.188	0.00106	4.368
12	BRF_p_8	0.29484	0.79813	0.19771	0.00053	0.609	0.198	0.00105	4.368
13	RF_BS_m	0.29634	0.80099	0.20114	0.00051	0.598	0.201	0.00101	4.368
14	RF_BS_p	0.28741	0.79925	0.20058	0.00050	0.586	0.201	0.00100	4.368
15	RF_BS_m_32	0.27598	0.79324	0.19075	0.00053	0.586	0.191	0.00105	4.138
16	RF_BS_p_32	0.27394	0.79345	0.19463	0.00050	0.575	0.195	0.00101	4.138
17	RF_BS_m_8	0.29606	0.80206	0.22874	0.00049	0.655	0.229	0.00098	4.368
18	RF_BS_p_8	0.30021	0.80273	0.22973	0.00047	0.632	0.230	0.00094	4.598

Abbreviations: H = H measure, AUC = area under the curve, ER = error rate, TPR= true positive rate, FPR= false positive rate, F= harmonic mean of recall (TPR) and precision, na= not available

Appendix B: Importance plots and list of abbreviations used

Table B.1: List of Abbreviations used in importance plots

Abbreviation	Explanation
AGE	Age of vessel
AGE_high	Age risk group high (0 to 2 and above 14 years)
AGE_low	Age risk group low (3 to 14 years)
CL_AB	Class - American Bureau of Shipping
CL_NK	Class - Nippon Kaiji Kyokai
CL_RS	Class - Russian Maritime Register
CL_UNK	Class Unknown
CL_VL	Class - Det Norske Veritas
CLCH3Y	Class Changes within 3 years
DOC_pres	DOC presence
DOC_UM	DOC company from upper middle income
DOC_UNK	DOC company unknown
DOC_YEX	DOC company years of experience
DOCH3Y	DOC changes within 3 years
FL_CAN	Flag Canada
FL_UNK	Flag Unknown
FL_USA	Flag USA
GRT	Gross Tonnage
LCAS_LS	Nr of less serious incidents within 365 days
LINSPECT	Previous inspection records (365 days)
LTOTALDEF	Previous number of deficiencies (365 days)
MEB_JPN	Main engine builder located in Japan
MEB_KRS	Main engine builder located in South Korea
MED_EMD	Main engine designer - EMD (Electro-Motive)
MED_MBW	Main engine designer - MAN-B&W
OWC_HIGH	DOC company from a high-income country
OWN_HIGH	Ownership from a high-income country
OWN_pres	Ownership presence
OWN_UM	Owner from upper middle income
OWN_UNK	Ownership unknown
OWN_YEX	Owner years of experience
OWNCH3Y	Ownership changes within 3 years
ST_DRY	Dry bulk carrier
ST_FISH	Fishing vessel
ST_GEN	General cargo Ship
ST_OTH	Other ship type
ST_PAX	Passenger vessel
ST_TANK	Tanker
ST_TUG	Tug
SY3	Shipyard Country Group 3 (medium risk)
SY4	Shipyard Country Group 4 (low risk)

Figure B.1: Visualization of variable importance: incident (VSS)

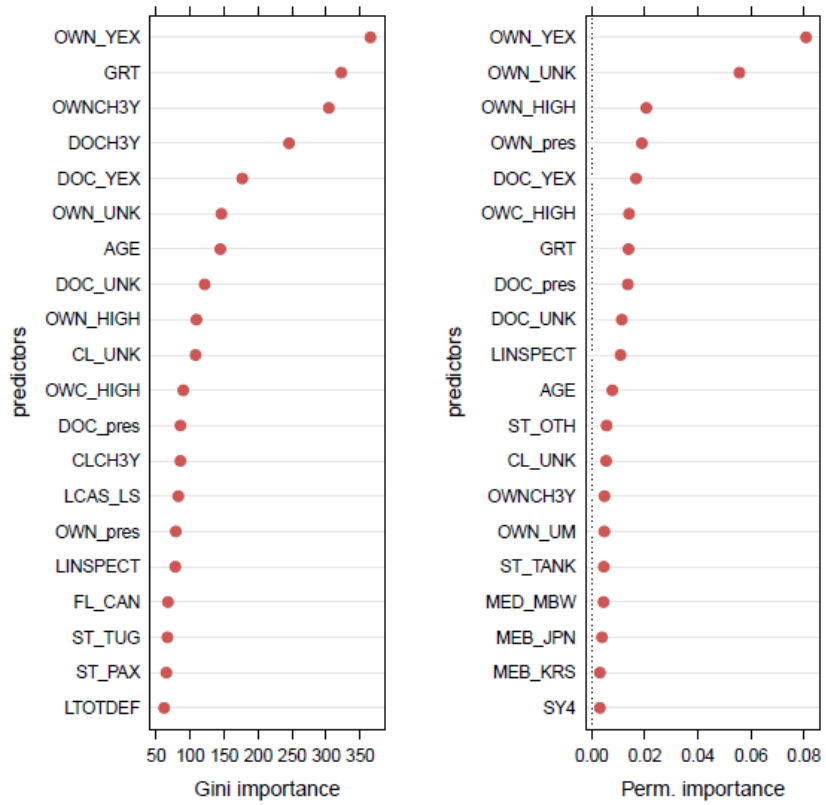


Figure B.2: Visualization of variable importance: pollution (VSS)

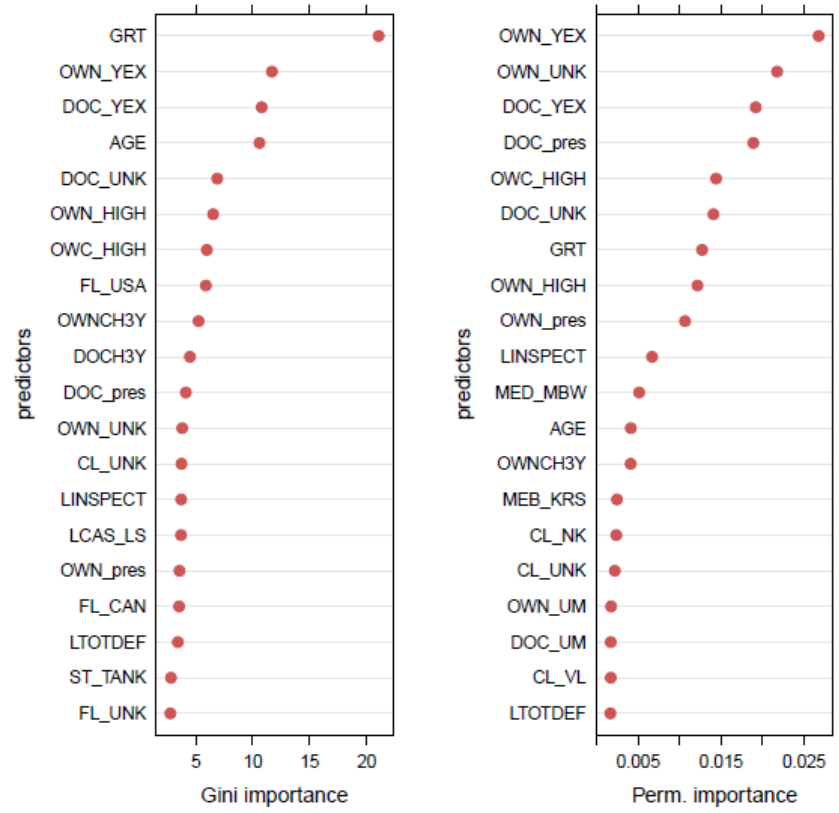


Figure B.3: Visualization of variable importance: loss of life and injuries (VSS)

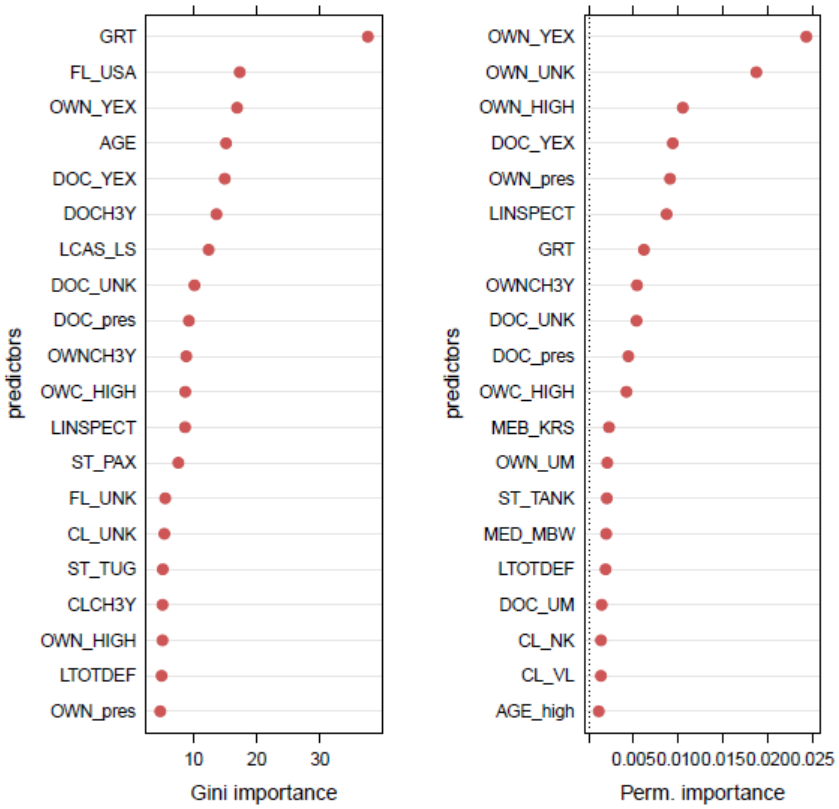


Figure B.4: Visualization of variable importance: fire and explosions (VSS)

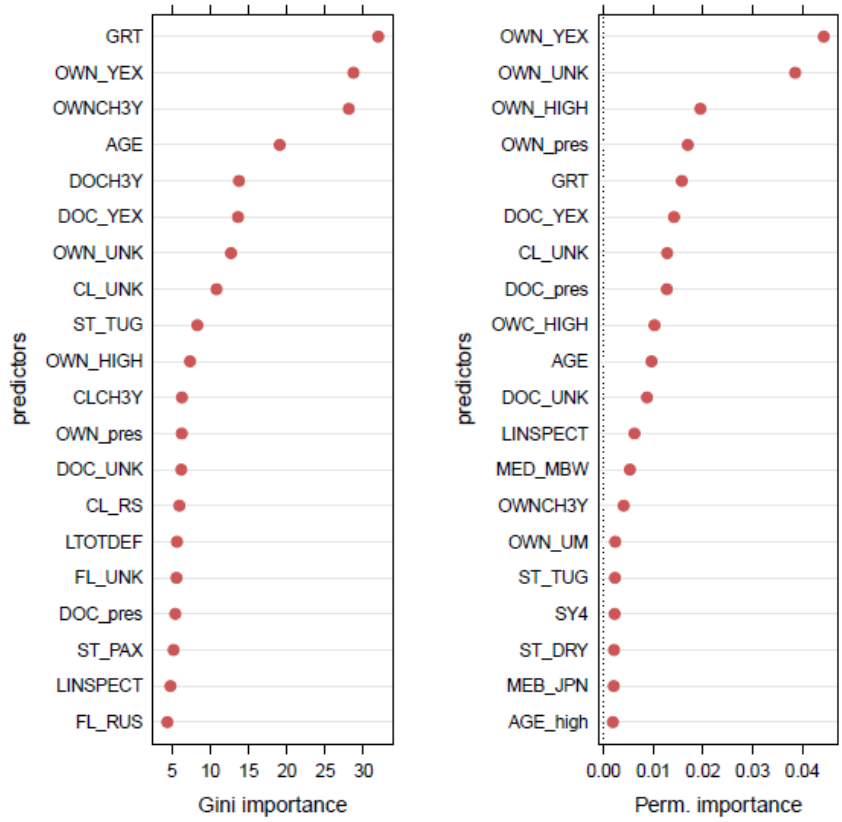


Figure B.5: Visualization of variable importance: collisions and powered groundings (VSS)

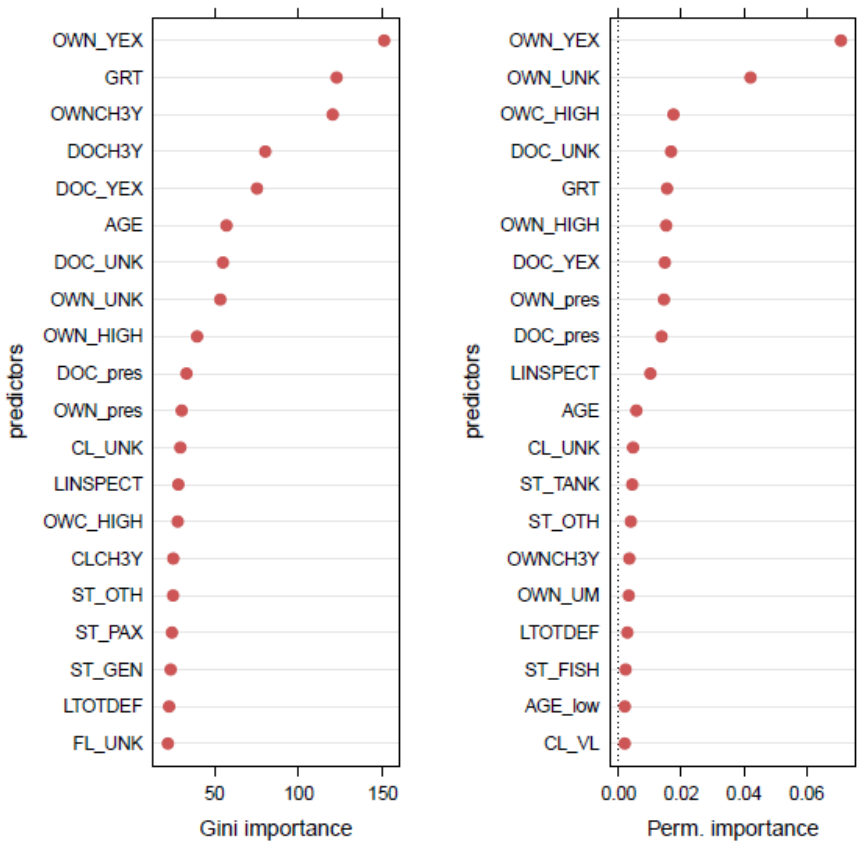


Figure B.6: Visualization of variable importance: mobility related failures (VSS)

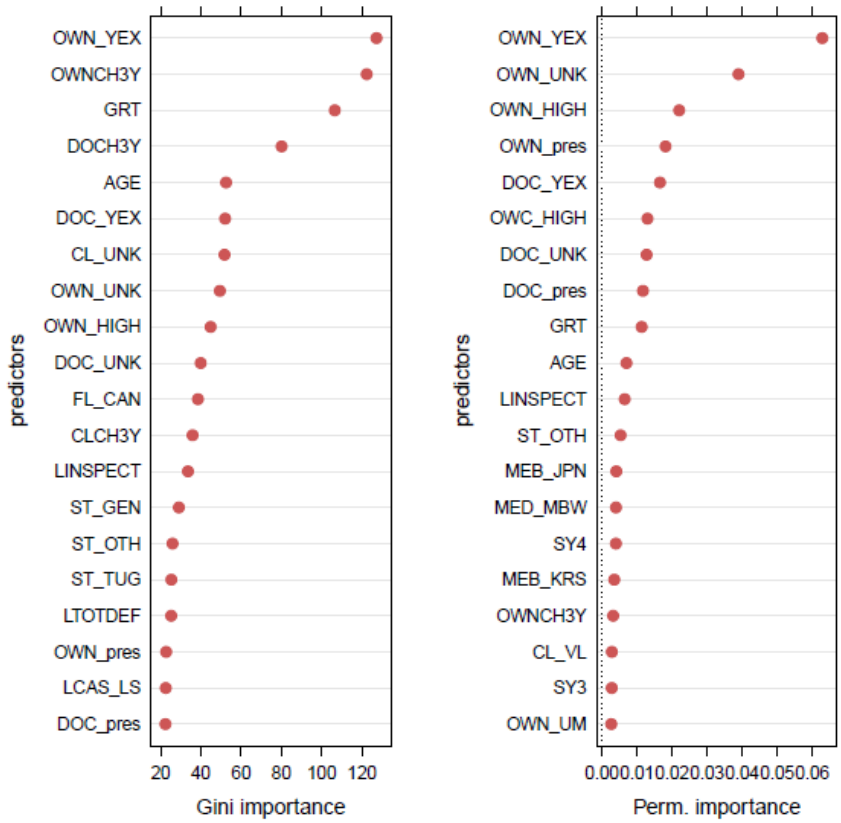


Figure B.7: Visualization of variable importance: hull related failures (VSS)

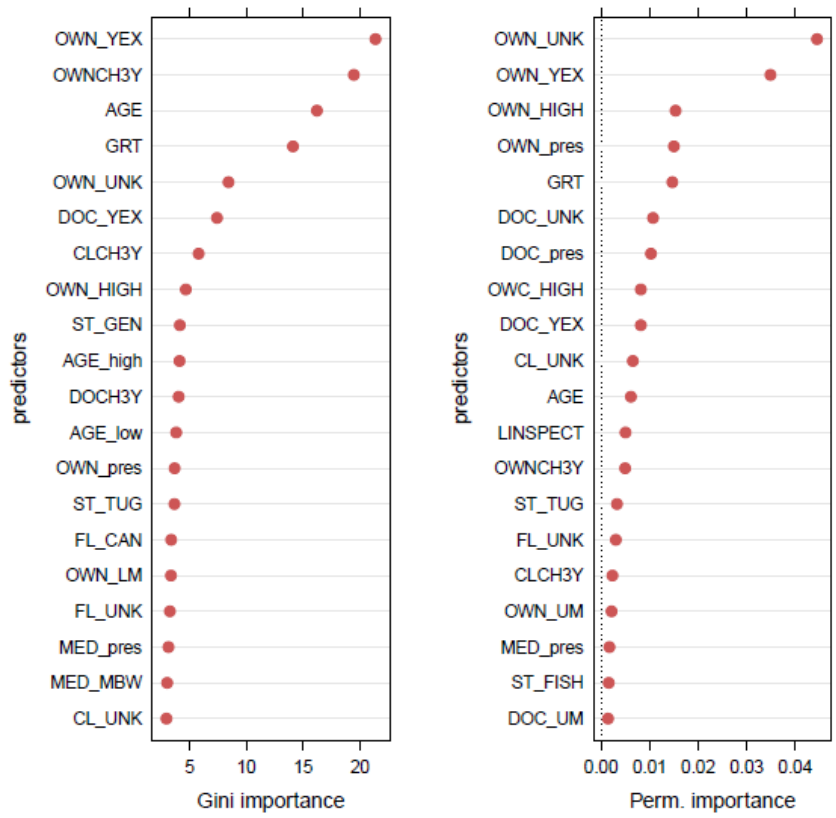


Figure B.8: Visualization of variable importance: equipment related failures (VSS)