Predicting inspection outcomes and evaluating port state control targeting using random forests

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

Econometric Institute Report 2024-01

Erasmus University

Abstract

This study uses global inspection data of 790k inspections and 1.5 million deficiencies (2013 to 2021) which is complemented by 500k incidents and ship particulars of 132k unique vessels. The results show that over 70% of ships that had very serious and serious incidents (2020 to 2021) were not inspected and only 2.5% were detained. The global averages of percentage of inspections without deficiencies is around 50% with high variability across the port state control (PSC) regimes (2013 to 2021). Since there is ample room for improvement to target risky vessels for inspection, it is not recommended to continue with the status quo of the industry by using detention alone as proxy to target future risk. Instead, the study develops 13 prediction models for detention and deficiency types using ML methods by evaluating over 400 risk factors. The results vary across the endpoint of interest but overall, the normal random forests variants outperform the other variants. The top 5 most influential covariates towards prediction are found to be the size of the vessel (GRT), age, previous number of deficiencies within 365 days prior to the inspection, the year of existence of the beneficial owner and safety manager company. These prediction models can be combined with incident type models to enhance targeting of risky vessels and reduce future incidents compared to the current status quo of 70% false negative events.

Keywords: detention, deficiencies, incidents, machine learning, case weighting, subsampling, effectiveness of PSC targeting, importance of covariates towards prediction

[1] Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, Netherlands, phone +61-466827029, 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

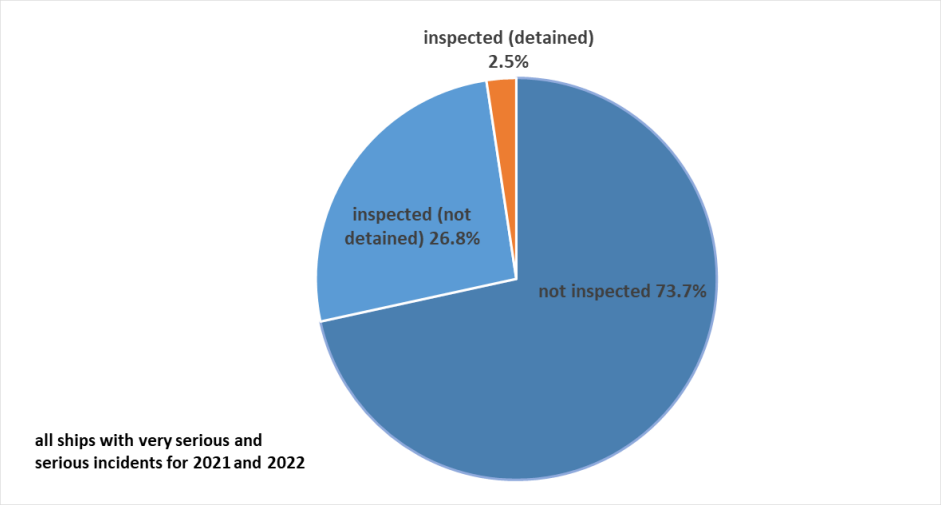
Currently the status quo of the industry is to use detention only to target ships for inspections while deficiency types and incident types are not considered by all Port State Control Regimes (PSC regimes) (Knapp 2006, Knapp and Heij, 2020). At the global level, there are ten regimes organized in Memoranda of Understanding (MoU) that cover the world regions (see Figure 2) with each regime using its own way to target ships and only one country using prediction models. Knapp and Heij (2020) demonstrate that targeting based on combined risk dimensions (using detention, incident, incident types and two deficiency types) can improve hit rates and reduce false negative predictions (i.e, missing a risky vessel). The approach presented in this study adds to this philosophy by adding more deficiency type models beside detention, which can then be combined with incident type models to improve targeting.

This study develops and tests thirteen prediction models based on machine learning (ML) and extends previous work on improved targeting for Port State Control (PSC) and domain awareness which can be used in conjunction with incident type models developed initially by Knapp and Heij (2020) or by Knapp and Van de Velden (2023). Knapp and Heij (2020) only use two deficiency models besides detention to combine with incident types and Knapp and Van de Velden (2023) only develop incident type models. This exploration study develops a detention and 12 deficiency type models.

The study uses a global inspection dataset going back to 2013 of 790k inspections and 1.5 million deficiencies (2013 to 2021). For the development of the detention and deficiency models, most recent data (2014 to 2019) is used for the train data and various ML methods are explored evaluating over 400 risk factors. Out of sample data (test data) of a period of 2 years (2020 to 2021) is used to assess predictive performance. Moreover, the influence of the various factors concerning the predictions is identified and visualized using variable importance plots. The study also uses 500k incident data (very serious and serious incidents) for the same period to provide a high-level evaluation of the effectiveness of the current targeting using detention only.

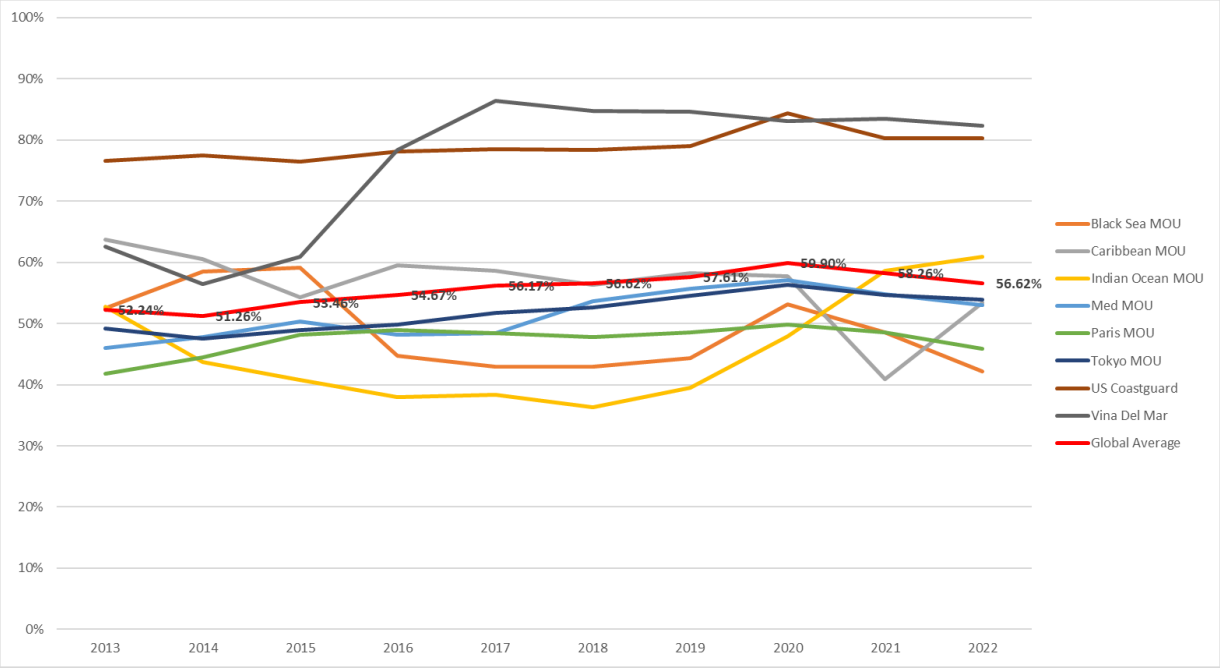
Figure 1 provides a high-level overview of the percentage of vessels that were not selected for inspection but had incidents within a three-month time frame from the time they could have been inspected – about 74% of the world fleet. Only 2.5 % were detained. It should be stated that one cannot observe the percentage of vessels that were inspected and did not have an incident, hence only a high-level overview can be provided which indicates room for improvement.

Figure 1: Overview of ships with incidents not selected for inspection (2020 to 2021)



Going back further in time, Figure 2 shows the percentage of inspections without deficiencies for the main PSC regimes with a global average of around 50% and with high variability across the regimes. This means that at least 50% of the resources allocated to inspections globally is not used effectively.

Figure 2: Percentage of inspections without deficiencies (2013 to 2021)



The high-level overviews provided in Figures 1 and 2 clearly support that targeting of PSC can be improved in the future taking the philosophy developed by Knapp and Heij (2020) into account and by combining detention, deficiency, and incident type models in the future.

The current study provides an important component by developing detention and twelve deficiency models using ML methods which can then be combined with the incident type models from Knapp and Van de Velden (2023). The development of improved targeting metrics that combine incident, detention and deficiency type models is beyond the scope of this study. The scope of this analysis is to highlight the need to improve targeting by using predictive models that account for the main factors that influence safety qualities of vessels. The current status quo for targeting vessels for inspections does not use predictive models, only uses historical inspection outcomes (eg. detention) to indicate future risk and puts heavy emphasis on flag and class performance rather than other important factors that can influence safety qualities of vessels.

The setup of this paper is as follows. In Section 2, the development of the dataset and variables used in this study is explained while Section 3 introduces the model variants and explains the methodology to evaluate models. Section 4 discusses the results and visualizes the importance of the variables towards prediction using importance plots. Section 5 ends with providing the main conclusions and recommendations.

2. Data and variables

The study uses a global inspection dataset going back to 2013 of 790k inspections and 1.5 million deficiencies (2013 to 2021). It also uses a global incident dataset of 500k incidents (very serious and serious) to provide a high-level overview of the effectiveness of the current state of the art in targeting ships for inspections. Note that most data needed for analysis in the maritime industry is rather scarce as detention and in particular incidents are rare events. The only

exception would be vessel positioning data which is high frequency data in the range of millions of observations.

For the detention and deficiency type models, the dataset is spit into train data, data used to construct the models, and test data, out of sample data used to assess predictive performance of the models. For the train data, 212,228 global inspections (6,070 detentions) from January 2014 to June 2019 form the basis to develop the formulas. Please refer to Annex A.1 for basic statics of the inspection data used as train data. For the test data to evaluate the models and to provide the high-level overview of the effectiveness of targeting, the out of sample data comprises of 1,029,726 observation (January 2020 to December 2021) of the world fleet with 133,252 inspections and 2,704 detentions comprising of 132k unique vessels.

Incident data from four different sources (S&P Global, USCG, AMSA ad LLIS) was added to the out of sample data. Incident data was reclassified to meet the definitions of seriousness used by IMO (IMO 2000) and only very serious and serious incidents (VSS) are considered as less serious incidents are highly underreported.

The inspection data covers data from the main PSC MoU's (Paris MoU, Tokyo MoU, Vina del Mar, USCG). Missing data, such as ship particulars at the time of the inspection, is complemented by using data from IHS Maritime. Only initial inspections are considered and follow up inspections were excluded to reduce a possible source of bias. Furthermore, inspection data is biased since the selection of ships that are inspected is guided by various target factors of the various Port State Control Memoranda of Understanding (MoU)s. For this reason, it is better to combine data from various MoU's rather than just one country or one region. Table 1 provides an overview of the number of variables in each main group and the data type for each main group.

Table 1: High level overview of number of variables

Variable groups	Type	Nr of factors
Size, age	continuous	2
Ship Types	categorical	9
Flag	categorical	151
Class	categorical	81
Main engine designer	categorical	115
Main engine builder location	categorical	30
Safety management company location	categorical	5
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		**422**

The initial selection of variables is based on the literature such as Knapp (2006), Knapp and Heij (2020) and Knapp and Van de Velden (2023). An overview of the main variable groups is provided in Table 1. We can distinguish:

1. Ship particulars such size, the age, the ship type, flags, main engine builder and designed and classification societies.
2. The country where the ship was built which is grouped into four groups and interaction effects with 2 age groups (0-2 and above 14 years of age representing high age risk and 3-14 years of ship age represent low age risk).

3. The country of location of the Safety Management Companies (DoC company) and group beneficial owner which are classified according to income based on the World Bank classification such as: 1) high income, 2) upper middle income, 3) lower middle income, 4) low income and 5) unknown.
4. The year of existence of safety management and beneficial owner which serves as proxy to their experience and quality. This is further complemented by an indicator that expressed the concentration of maritime industries such as ownership companies, safety management companies, engine designers and builders. The concentration acts as proxy to knowledge spill over and safety quality.
5. Lagged inspection, deficiency and incident history of the vessel (within 1 year prior to event date) and changes of ship particulars overtime such as flag changes, ownership changes, DoC company changes and class changes within 3 years prior to event date of interest

Note that there are over 600 individual deficiency codes and 29 main deficiency groups. The groups were regrouped into 12 groups reflecting inspection areas that are found to be useful for inspections and for domain awareness, and which could also be combined with the incident types used by Knapp and Van de Velden (2023) in the future.

Since vessels can have more than one deficiency for each deficiency group during inspections, the variables are reclassified into 0 and 1 indicating none or at least one deficiency) for each of the deficiency groups. Table 2 provides the counts for the dependent variables for the train and test data.

Table 2: High level overview of data used for model development and testing

	Train data (2014 to 2019)		**Test data (2020 to 2021)**	
	indicator	**sum**	**indicator**	**sum**
Inspected	212,228	-	99,944	133,252
Detained	6,070	-	2,602	2,704
Certificates and Qualifications	32,394	50,947	9,783	16,729
Maritime Labor Convention	28,583	45,965	12,063	22,413
Structural and Watertight Integrity	29,001	43,048	8,934	14,433
Propulsion and Machinery	14,466	19,223	5,117	7,938
Life Saving and Fire Appliances	56,622	108,966	16,521	33,349
Emergency systems and alarms	19,190	24,104	7,265	10,140
Safety of Nav. and Radio Com.	36,102	59,261	10,929	19,527
Safety Management (ISM)	24,415	31,555	6,272	9,065
Marpol A1 to A3	6,351	7,119	1,992	2,397
Marpol A4 and 5	8,578	9,402	2,897	3,548
Marpol A6	4,429	4,863	997	1,122
Ballast Water and Antifouling	1,547	1,724	1,677	2,002
Total deficiencies	**261,678**	**406,177**	**84,447**	**142,663**

Note: indicator means for the train data at least 1 deficiency per inspection, and for the out of test data at least one inspection, detention or deficiency per period

The endpoints of interest are as follows:

- detained
- Group 1: Certificates and Qualifications (Code groups 01100, 01200, 01300)
- Group 2: Maritime Labor Convention (Code groups 18100, 18200, 18300, 18400)

- Group 3: Structural Conditions and Watertight Integrity (Code groups 02100, 03100)
- Group 4: Propulsion and Auxiliary Machinery (Code group 13100)
- Group 5: Life Saving Appliances and Fire Safety (Code groups 1100, 07100)
- Group 6: Emergency Systems and Alarms (Code groups 04100, 08100)
- Group 7: Safety of Navigation and Radio Communications (Code groups 10100, 05100)
- Group 8: Safety Management (ISM-15100, Cargo Operations-06100 and Dangerous Goods-12100, Other – 99101, 99102)
- Group 9: MARPOL Annex 1 to 3 (Oil-14100, Chemicals-14200, 14300)
- Group 10: MARPOL Annex 4 and 5 (Sewage-14400, Garbage-14500)
- Group 11: MARPOL Annex 6 (Air Pollution-14600)
- Group 12: Ballast Water and Anti Fouling (Code groups 14700, 14800)

The only deficiency group excluded from the analysis is ISPS (security) since the dataset does not have enough observations for this type of deficiency group. Furthermore, empirical data of incident data and ship particular data of the world fleet is used for general evaluation of the models but to also filter out ships that could have been inspected but did not get selected.

3. Combination of model variants and model evaluation

The present study considers 13 end points of interest in total (detention plus 12 deficiency types). Table 3 provides a list of the model combinations that were used – a total of 18 variants for the 13 endpoints of interest, hence a total of 234 combinations. The machine learning models that we consider here are all random forest variants. For a general overview of random forests, class-imbalance and tuning please refer to Knapp and Van de Velden (2023) and Breiman (1996, 2001) and Breiman et al (1984).

Table 3: Summary of model variants

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 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

Table 3 shows three model groups. 1) Regular random forests (RF), 2) Balanced random forests (BRF) by Chen et al. (2004) and 3) Random Forest on balanced samples (RF_BS) (regular random forests based on (under sampled) balanced samples of the training data. Based on initial experiments on tuning we considered, for each group, three options for *m*, that is, the number of randomly selected variables to be considered at splits. In particular, either 8, 16 or 32 variables were considered. Table 3 provides a summary of all considered model variants.

Moreover, for all random forests, aggregation of results is considered using both majority voting as well as averaging of probabilities. For majority voting, the class predictions of each tree are 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. To estimate and evaluate the models, R is used.

As shown in Table 2, the test data are from January 2020 to December 2021. To evaluate the models, probabilities are estimated at a certain time with the assumption that they are valid for 3 months (see Knapp and Heij, 2020 and Knapp and Van de Velden, 2023). Second, observed data is matched with the estimated probabilities and evaluation metrics are calculated using the following setup and eight periods:

- *P1: Probabilities as of January 2020 – empirical data from January 2020 to March 2020*
- *P2: Probabilities as of April 2020 – empirical data from April 2020 to June 2020*
- *P3: Probabilities as of July 2020 – empirical data from July 2020 to September 2020*
- *P4: Probabilities as of October 2020 – empirical data from October to December 2020*
- *P5: Probabilities as of January 2021 – empirical data from January to March 2021*
- *P6: Probabilities as of April 2021 – empirical data from April to June 2021*
- *P7: Probabilities as of July 2021 – empirical data from July to September 2021*
- *P8: Probabilities as of October 2021 – empirical data from November to December 2021*

The main interest for targeting of vessels is to reduce false negative events. A false negative event is when a risky vessel is missed since it was not selected for inspection and then has an incident with very serious or serious consequences which can be very costly.

Based on Knapp and Van de Velden (2023) who explain the various evaluation metric limitations, the top-decile lift is considered here. It compares the 10% highest estimated probabilities with random selection of vessels. If the predicted probabilities are good, the top decile lift is large.

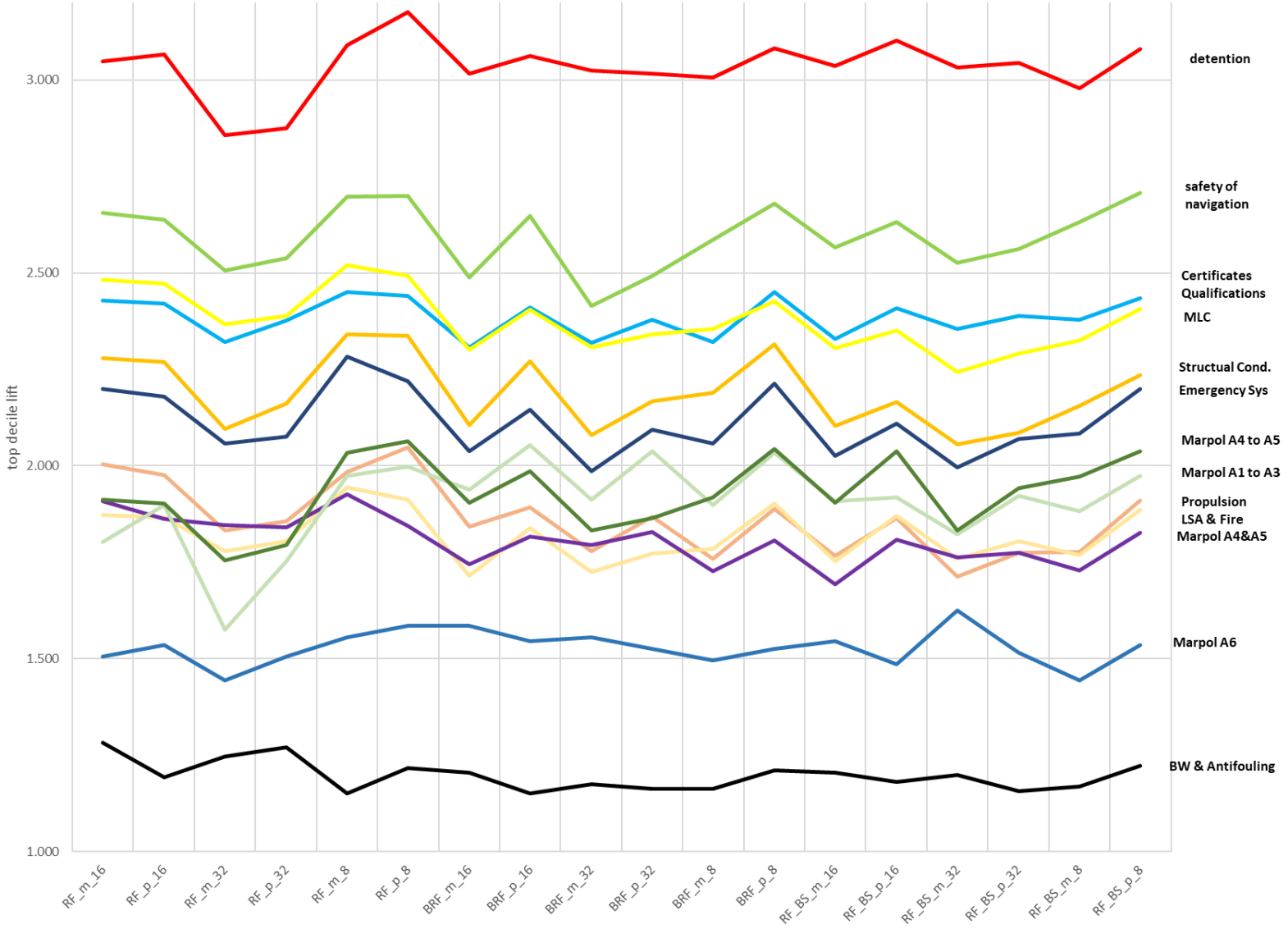
4. Results and Discussions

4.1. Model selection for detention and deficiency models

Figure 3 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 Knapp and Van de Velden (2023) 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 the test data (2020 and 2021).

Detention is easier to predict than individual deficiency groups of which detection depends upon the training and background of the inspector (Knapp, 2006). It is therefore no surprise to see this difference. Deficiency groups related to safety of navigation and certificates and qualifications as well as the Maritime Labor Convention follow in second and third place while MARPOL Annex 6 (air emissions) and Ballast Water and Antifouling have the worst top decile lift as they are harder to predict, and inspectors are less experiences in these areas as they are relative new areas.

Figure 3: Top decile lift for each model variant (test data 2021 and 2020)



Abbreviations: 1: detained, 2: Certificates and Qualifications, 3: Maritime Labor Convention, 4: Structural Conditions and Watertight Integrity, 5: Propulsion and Auxiliary Machinery, 6: Life Saving Appliances and Fire Safety, 7: Emergency Systems and Alarms, 8: Safety of Navigation and Radio Communications, 9: Safety Management (ISM), 10: MARPOL Annex 1 to 3 (Oil and chemicals), 11: MARPOL Annex 4 and 5 (Sewage and Garbage-14500), 12: MARPOL Annex 6 (Air Pollution), 13: Ballast Water and Anti Fouling

Table 4: Summary of results for detention and deficiency models – top decile lift (train data for 2020 to 2021)

	Dependent Variable		Deficiency Groups											
	Model variants	**Detention**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
1	RF_m_18	3.048	2.428	2.482	2.279	2.003	1.873	2.200	2.655	**1.908**	1.802	1.912	1.505	**1.282**
2	RF_p_18	3.067	2.420	2.472	2.269	1.976	1.867	2.179	2.637	1.863	1.898	1.902	1.535	1.193
3	RF_m_32	2.856	2.320	2.366	2.095	1.833	1.778	2.058	2.506	1.847	1.576	1.754	1.444	1.246
4	RF_p_32	2.875	2.376	2.388	2.161	1.857	1.804	2.076	2.538	1.841	1.752	1.795	1.505	1.270
5	RF_m_8	3.090	**2.451**	**2.520**	**2.341**	1.984	**1.944**	**2.282**	2.698	1.927	1.973	2.033	1.555	1.151
6	RF_p_8	**3.175**	2.441	2.492	2.337	**2.048**	1.912	2.220	2.700	1.844	1.998	**2.064**	1.585	1.217
7	BRF_m_16	3.017	2.307	2.301	2.106	1.843	1.714	2.037	2.489	1.745	1.938	1.905	1.585	1.205
8	BRF_p_16	3.063	2.410	2.404	2.270	1.892	1.838	2.146	2.647	1.817	**2.053**	1.985	1.545	1.151
9	BRF_m_32	3.025	2.318	2.306	2.079	1.778	1.725	1.986	2.415	1.794	1.913	1.833	1.555	1.175
10	BRF_p_32	3.017	2.378	2.341	2.167	1.868	1.772	2.094	2.493	1.828	2.038	1.864	1.525	1.163
11	BRF_m_8	3.006	2.321	2.354	2.189	1.759	1.785	2.057	2.586	1.726	1.898	1.919	1.495	1.163
12	BRF_p_8	3.082	**2.451**	2.427	2.315	1.888	1.902	2.213	2.679	1.806	2.033	2.044	1.525	1.211
13	RF_BS_m_16	3.036	2.329	2.304	2.104	1.767	1.753	2.025	2.566	1.692	1.908	1.905	1.545	1.205
14	RF_BS_p_16	3.102	2.408	2.351	2.166	1.864	1.870	2.110	2.631	1.809	1.918	2.037	1.485	1.181
15	RF_BS_m_32	3.032	2.355	2.242	2.056	1.712	1.758	1.995	2.525	1.763	1.822	1.833	**1.625**	1.199
16	RF_BS_p_32	3.044	2.388	2.290	2.086	1.775	1.805	2.069	2.562	1.774	1.923	1.943	1.515	1.157
17	RF_BS_m_8	2.979	2.379	2.324	2.155	1.777	1.768	2.084	2.632	1.729	1.883	1.971	1.444	1.169
18	RF_BS_p_8	3.079	2.434	2.406	2.234	1.911	1.887	2.200	**2.708**	1.826	1.973	2.037	1.535	1.222

Abbreviations:1=detained, 2: Certificates and Qualifications, 3: Maritime Labor Convention, 4: Structural Conditions and Watertight Integrity, 5: Propulsion and Auxiliary Machinery, 6: Life Saving Appliances and Fire Safety, 7: Emergency Systems and Alarms, 8: Safety of Navigation and Radio Communications, 9: Safety Management (ISM), 10: MARPOL Annex 1 to 3 (Oil and chemicals), 11: MARPOL Annex 4 and 5 (Sewage and Garbage-14500), 12: MARPOL Annex 6 (Air Pollution), 13: Ballast Water and Anti Fouling

In Table 4, the best performing model based on the two-year test data is highlighted in bold. Not surprisingly, the result varies across the dependent variable but overall, the normal random forests variants RF (m8 and p8) outperform the other variants for most dependent variables. Variants BRF (m8 and p8) are possible alternatives. For Safety of Navigation, variant RF_BS (p8) is the best and for MARPOL Annex VI, variant RF_BS (m32) performs best.

It is recommended to choose the best three models and re-evaluate their performance every year with new out of sample data. Especially for the areas that are relatively new for inspectors and where inspections are not as straight forward, detection and prediction is more difficult compared to classic deficiencies such as certificates, qualifications or areas related to the safety of navigation.

Overall, all models perform well and better than random selection. Deficiency type models along with detention should not be used as the sole metric to target ships for inspections but can enhance targeting of risky vessels where the endpoint of interest is to reduce false negative events related to future incidents.

4.2. Importance of covariates for detention and deficiencies

The last section visualizes the importance of covariates towards prediction for each of the models considered here. To visualize the importance of variables towards the endpoint of interest, one can calculate the contribution of each variable towards the construction of the prediction. This is accomplished by calculating the contribution for each variable to the total decrease of variance, which is measured by the decrease in the Gini index used to determine split in the trees. The resulting measure is referred to as *Gini importance*.

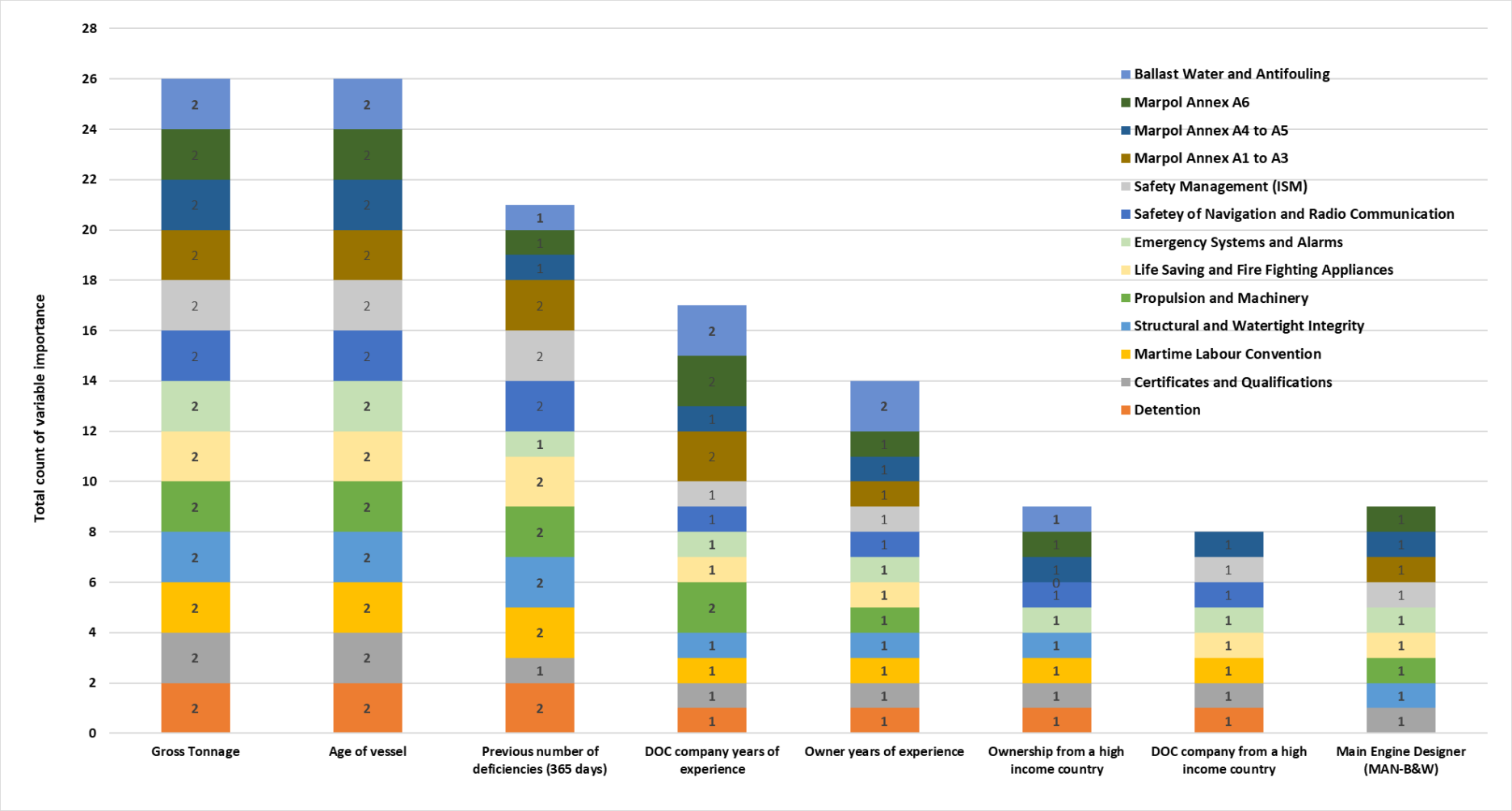
Another way to assess the importance is to look at the impact of a variable towards the final prediction accuracy. The influence of each variable on the predictions is considered by comparing the random forest predictions with predictions obtained when the values of a variable are randomly permuted. 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*.

Appendix C shows the importance plots for each model type using model RF_m_8 (Regular RF, m =8, majority votes aggregation). 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 all endpoints of interest, two plots are provided in Appendix C along with a Legend Table explaining the abbreviations used in the plots. For detention (Table C1.), one can see that variables GRT and Age are important followed by the total number of previous deficiencies. Figure 4 summarizes the importance of covariates based on aggregating the top 5 most influential covariates for all 26 plots (either based on the Gini importance or the permutation importance).

One can observe from Figure 4 that the most influential variables are GRT, age, the previous number of deficiencies within 365 days prior to the inspection, the number of years of existence of the beneficial owner and safety manager company, the location of ownership and safety management company and one main engine designer. The results are similar across all endpoints of interest. Variables that are given prime importance in targeting vessels for inspections currently is the flag. The importance plots do not confirm this as flag is not within the top 5 and only 2 flags (Panama and Belize) appear within the top 20 most influential variables. This further supports that a targeting regime based primarily on flag and only historical inspection history to prevent future incidents is not the most efficient way of targeting ships for inspections.

Figure 4: Summary of importance of top 5 covariates by model type



A comparison with the importance plots from Knapp and Van de Velden (2003) for incident type models shows similar results with variables associated with beneficial ownership and the safety management company besides age and size of the vessel been influential for prediction. It is important to notice here again that it is unknown whether these are positive or negative effects. They are important with respect to prediction and to decreasing variance as explained previously.

5. Conclusions and Recommendations

This exploration study uses a unique global inspection dataset going back to 2013 of 790k inspections and 1.5 million deficiencies (2013 to 2021) and is complemented by global incident data (500k very serious and serious incidents) and world fleet data of 132k unique vessels for the time period 2000 to 2001.

Our results clearly indicate that there is room for improvement in targeting vessels for inspections. Over 70% of ships that had very serious and serious incidents (2020 to 2021) were not inspected and only 2.5% were detained. Going back in time to 2013, the percentage of inspections without deficiencies for the main PSC regimes show a global average of around 50% with high variability across the regimes.

The study develops prediction models for a total of 13 endpoints of interest (detention plus 12 deficiency type models) using machine learning, thereby evaluating 18 variants and 234 combinations in total. The results show that all models perform better than random selection. Although results vary across the dependent variable, it appears that regular random forests variants with only 8 variables randomly selected at each split (m8 and p8) outperform other variants for most dependent variables. Variants BRF (m8 and p8) are possible alternatives. It is recommended to implement the best 3 model types and re-evaluate them with new out of sample data after some time.

The 5 most influential covariates towards prediction are GRT, age, previous number of deficiencies, the year of existence of the beneficial owner and safety manager company. Targeting factors currently used in the industry such as flag are not among the most influential variables and only two flags appear when looking at the top 20 factors.

Given that there is ample room for improvement in targeting vessels for inspections to reduce future incidents, the recommendation based on this study is to change the way ships are targeted for inspection. To improve the reduction of false negative events, it is recommended to use better methods as currently used and to combine incident type models similar to Knapp and Heij (2020) with incident type models. The models developed here form one component towards this goal and the authors are currently working on a combined and revised targeting metric based on 21 models.

References

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.

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, Michel van de Velden (2023), Exploration of machine learning methods for maritime risk predictions, Maritime Policy and Management, https://doi.org/10.1080/03088839.2023.2209788

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/

Appendix A.1: Detention and mean deficiencies by year (train data)

Year	Not Detained Count	Detained Count	Inspections Sum	Det rate %	Deficiencies Indicator*)	Deficiencies Sum	Deficiencies Mean
2014	38,322	1,320	39,642	3.3%	54,459	88,464	2.23
2015	39,279	1,337	40,616	3.3%	51,228	79,909	1.97
2016	38,777	1,145	39,922	2.9%	48,737	75,990	1.90
2017	39,412	1,052	40,464	2.6%	47,638	73,543	1.82
2018	38,899	963	39,862	2.4%	47,612	71,161	1.79
2019	11,469	253	11,722	2.2%	12,004	18,193	1.55
Total	**206,158**	**6,070**	**212,228**	**2.9%**	**261,678**	**407,260**	**1.92**

**) Note: at least 1 deficiency per deficiency group*

Appendix B: Evaluation metrics

Table B.1: Evaluation metrics: Detention

Model	H	AUC	ER	Precision	TPR	FPR	F	TopDecile
RF_m	0.123	0.691	0.026	0.500	0.001	0.000	0.002	3.048
RF_p	0.131	0.705	0.026	0.500	0.001	0.000	0.002	3.067
RF_m_32	0.112	0.689	0.026	0.172	0.002	0.000	0.004	2.856
RF_p_32	0.118	0.697	0.026	0.286	0.001	0.000	0.002	2.875
RF_m_8	0.120	0.679	0.026		0.000	0.000		3.090
RF_p_8	0.139	0.712	0.026		0.000	0.000		3.175
BRF_m	0.137	0.715	0.285	0.053	0.584	0.281	0.097	3.017
BRF_p	0.140	0.717	0.297	0.051	0.596	0.294	0.095	3.063
BRF_m_32	0.136	0.714	0.330	0.049	0.630	0.329	0.090	3.025
BRF_p_32	0.141	0.718	0.325	0.050	0.637	0.324	0.092	3.017
BRF_m_8	0.134	0.711	0.251	0.055	0.533	0.245	0.100	3.006
BRF_p_8	0.142	0.718	0.269	0.054	0.564	0.265	0.098	3.082
RF_BS_m	0.136	0.714	0.285	0.052	0.577	0.282	0.095	3.036
RF_BS_p	0.140	0.717	0.293	0.051	0.587	0.290	0.094	3.102
RF_BS_m_32	0.133	0.711	0.308	0.050	0.605	0.306	0.093	3.032
RF_BS_p_32	0.138	0.715	0.313	0.050	0.611	0.311	0.092	3.044
RF_BS_m_8	0.133	0.710	0.256	0.055	0.542	0.251	0.099	2.979
RF_BS_p_8	0.140	0.715	0.270	0.053	0.552	0.265	0.096	3.079

Notes: 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, n.a. = not available.

Table B.2: Evaluation metrics: Certificates and Qualifications

Model	H	AUC	ER	Precision	TPR	FPR	F	TopDecile
RF_m	0.095	0.662	0.116	0.261	0.101	0.031	0.145	2.428
RF_p	0.098	0.666	0.115	0.264	0.100	0.030	0.145	2.420
RF_m_32	0.086	0.656	0.123	0.248	0.125	0.041	0.166	2.320
RF_p_32	0.089	0.659	0.121	0.251	0.117	0.038	0.160	2.376
RF_m_8	0.096	0.661	0.111	0.264	0.075	0.023	0.117	2.451
RF_p_8	0.101	0.669	0.112	0.263	0.077	0.024	0.119	2.441
BRF_m	0.092	0.664	0.317	0.166	0.559	0.304	0.256	2.307
BRF_p	0.098	0.668	0.331	0.163	0.577	0.321	0.254	2.410
BRF_m_32	0.087	0.659	0.356	0.154	0.588	0.350	0.244	2.318
BRF_p_32	0.093	0.663	0.354	0.155	0.589	0.348	0.246	2.378
BRF_m_8	0.097	0.666	0.299	0.174	0.546	0.282	0.263	2.321
BRF_p_8	0.101	0.670	0.312	0.170	0.563	0.299	0.261	2.451
RF_BS_m	0.093	0.664	0.322	0.164	0.562	0.310	0.254	2.329
RF_BS_p	0.097	0.666	0.333	0.162	0.575	0.323	0.253	2.408
RF_BS_m_32	0.089	0.659	0.360	0.152	0.587	0.355	0.242	2.355
RF_BS_p_32	0.092	0.662	0.356	0.155	0.592	0.351	0.245	2.388
RF_BS_m_8	0.097	0.666	0.300	0.172	0.543	0.283	0.262	2.379
RF_BS_p_8	0.100	0.669	0.315	0.168	0.561	0.302	0.258	2.434

Notes: 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, n.a. = not available.

Table B.3: Evaluation metrics Maritime Labour Convention

Model	H	AUC	ER	Precision	TPR	FPR	F	TopDecile
RF_m	0.108	0.675	0.124	0.389	0.052	0.011	0.092	2.482
RF_p	0.111	0.682	0.124	0.397	0.054	0.011	0.096	2.472
RF_m_32	0.100	0.673	0.128	0.346	0.071	0.019	0.118	2.366
RF_p_32	0.103	0.675	0.126	0.368	0.064	0.015	0.108	2.388
RF_m_8	0.108	0.670	0.122	0.433	0.034	0.006	0.063	2.520
RF_p_8	0.116	0.685	0.123	0.415	0.037	0.007	0.069	2.492
BRF_m	0.104	0.679	0.370	0.190	0.634	0.370	0.293	2.301
BRF_p	0.110	0.683	0.380	0.188	0.647	0.384	0.291	2.404
BRF_m_32	0.097	0.673	0.410	0.179	0.667	0.421	0.282	2.306
BRF_p_32	0.104	0.678	0.413	0.179	0.674	0.425	0.283	2.341
BRF_m_8	0.106	0.679	0.339	0.199	0.595	0.329	0.298	2.354
BRF_p_8	0.114	0.685	0.352	0.195	0.613	0.347	0.296	2.427
RF_BS_m	0.100	0.676	0.373	0.187	0.627	0.373	0.289	2.304
RF_BS_p	0.105	0.679	0.382	0.186	0.640	0.385	0.288	2.351
RF_BS_m_32	0.093	0.670	0.415	0.177	0.670	0.427	0.280	2.242
RF_BS_p_32	0.097	0.673	0.419	0.176	0.672	0.431	0.279	2.290
RF_BS_m_8	0.103	0.676	0.339	0.197	0.591	0.330	0.296	2.324
RF_BS_p_8	0.110	0.682	0.355	0.193	0.608	0.350	0.293	2.406

Notes: 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, n.a. = not available.

Table B.4: Evaluation metrics Structural and Watertight Integrity

Model	H	AUC	ER	Precision	TPR	FPR	F	TopDecile
RF_m	0.085	0.667	0.098	0.209	0.035	0.013	0.060	2.279
RF_p	0.088	0.672	0.097	0.218	0.035	0.012	0.060	2.269
RF_m_32	0.077	0.662	0.106	0.196	0.059	0.024	0.091	2.095
RF_p_32	0.081	0.665	0.103	0.211	0.054	0.020	0.086	2.161
RF_m_8	0.084	0.661	0.094	0.209	0.017	0.006	0.032	2.341
RF_p_8	0.094	0.676	0.094	0.224	0.019	0.007	0.035	2.337
BRF_m	0.085	0.670	0.356	0.143	0.597	0.352	0.231	2.106
BRF_p	0.089	0.673	0.371	0.141	0.616	0.370	0.229	2.270
BRF_m_32	0.078	0.664	0.408	0.133	0.648	0.414	0.221	2.079
BRF_p_32	0.083	0.669	0.411	0.134	0.658	0.418	0.222	2.167
BRF_m_8	0.087	0.671	0.318	0.150	0.548	0.305	0.235	2.189
BRF_p_8	0.092	0.675	0.339	0.146	0.576	0.331	0.233	2.315
RF_BS_m	0.083	0.669	0.358	0.141	0.592	0.353	0.228	2.104
RF_BS_p	0.086	0.671	0.369	0.140	0.610	0.367	0.228	2.166
RF_BS_m_32	0.077	0.663	0.404	0.133	0.640	0.409	0.221	2.056
RF_BS_p_32	0.080	0.666	0.403	0.134	0.645	0.408	0.222	2.086
RF_BS_m_8	0.086	0.671	0.316	0.149	0.536	0.301	0.233	2.155
RF_BS_p_8	0.089	0.673	0.334	0.146	0.565	0.324	0.232	2.234

Notes: 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, n.a. = not available.

Table B.5: Evaluation metrics Propulsion and Machinery

Model	H	AUC	ER	Precision	TPR	FPR	F	TopDecile
RF_m	0.062	0.649	0.051	0.000	0.000	0.000	0.000	2.003
RF_p	0.067	0.658	0.051	0.000	0.000	0.000	0.000	1.976
RF_m_32	0.057	0.646	0.052	0.095	0.001	0.000	0.002	1.833
RF_p_32	0.060	0.648	0.051	0.000	0.000	0.000	0.000	1.857
RF_m_8	0.059	0.643	0.051		0.000	0.000		1.984
RF_p_8	0.071	0.661	0.051		0.000	0.000		2.048
BRF_m	0.066	0.657	0.398	0.078	0.630	0.400	0.139	1.843
BRF_p	0.068	0.660	0.412	0.078	0.649	0.415	0.139	1.892
BRF_m_32	0.062	0.652	0.431	0.076	0.666	0.436	0.137	1.778
BRF_p_32	0.065	0.656	0.441	0.075	0.675	0.447	0.136	1.868
BRF_m_8	0.065	0.654	0.375	0.080	0.599	0.374	0.141	1.759
BRF_p_8	0.068	0.661	0.387	0.080	0.620	0.387	0.141	1.888
RF_BS_m	0.061	0.652	0.408	0.077	0.636	0.410	0.138	1.767
RF_BS_p	0.065	0.655	0.422	0.077	0.659	0.427	0.138	1.864
RF_BS_m_32	0.056	0.645	0.446	0.074	0.673	0.452	0.134	1.712
RF_BS_p_32	0.060	0.650	0.450	0.075	0.688	0.458	0.135	1.775
RF_BS_m_8	0.064	0.653	0.387	0.079	0.618	0.387	0.141	1.777
RF_BS_p_8	0.067	0.659	0.396	0.079	0.630	0.398	0.140	1.911

Notes: 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, n.a. = not available.

Table B.6: Evaluation metrics Life Saving and Fire Appliances

Model	H	AUC	ER	Precision	TPR	FPR	F	TopDecile
RF_m	0.076	0.653	0.204	0.310	0.189	0.083	0.235	1.873
RF_p	0.077	0.656	0.206	0.307	0.194	0.087	0.238	1.867
RF_m_32	0.066	0.645	0.221	0.289	0.230	0.112	0.256	1.778
RF_p_32	0.068	0.646	0.218	0.292	0.224	0.108	0.253	1.804
RF_m_8	0.077	0.652	0.195	0.325	0.165	0.068	0.219	1.944
RF_p_8	0.081	0.659	0.196	0.319	0.164	0.069	0.217	1.912
BRF_m	0.073	0.652	0.375	0.243	0.599	0.370	0.345	1.714
BRF_p	0.077	0.656	0.396	0.237	0.630	0.402	0.345	1.838
BRF_m_32	0.064	0.643	0.428	0.225	0.651	0.444	0.334	1.725
BRF_p_32	0.068	0.647	0.432	0.225	0.660	0.450	0.336	1.772
BRF_m_8	0.077	0.656	0.345	0.253	0.557	0.325	0.348	1.785
BRF_p_8	0.081	0.659	0.365	0.247	0.591	0.356	0.349	1.902
RF_BS_m	0.073	0.653	0.375	0.243	0.599	0.370	0.346	1.753
RF_BS_p	0.076	0.655	0.394	0.237	0.625	0.398	0.344	1.870
RF_BS_m_32	0.065	0.643	0.425	0.225	0.646	0.439	0.334	1.758
RF_BS_p_32	0.068	0.646	0.426	0.226	0.650	0.441	0.335	1.805
RF_BS_m_8	0.078	0.657	0.341	0.254	0.547	0.319	0.347	1.768
RF_BS_p_8	0.079	0.658	0.365	0.246	0.586	0.355	0.347	1.887

Notes: 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, n.a. = not available.

Table B.7: Evaluation metrics Emergency Systems and Alarms

Model	H	AUC	ER	Precision	TPR	FPR	F	TopDecile
RF_m	0.071	0.649	0.073	0.181	0.002	0.001	0.004	2.200
RF_p	0.077	0.661	0.073	0.195	0.002	0.001	0.005	2.179
RF_m_32	0.063	0.646	0.075	0.181	0.010	0.003	0.018	2.058
RF_p_32	0.067	0.650	0.074	0.191	0.005	0.002	0.010	2.076
RF_m_8	0.073	0.644	0.073	0.133	0.000	0.000	0.001	2.282
RF_p_8	0.082	0.665	0.073	0.172	0.001	0.000	0.001	2.220
BRF_m	0.075	0.662	0.369	0.114	0.603	0.367	0.192	2.037
BRF_p	0.078	0.664	0.385	0.112	0.619	0.386	0.189	2.146
BRF_m_32	0.066	0.652	0.421	0.106	0.643	0.426	0.182	1.986
BRF_p_32	0.072	0.657	0.419	0.107	0.650	0.424	0.184	2.094
BRF_m_8	0.075	0.661	0.328	0.119	0.552	0.319	0.196	2.057
BRF_p_8	0.081	0.666	0.355	0.116	0.588	0.350	0.194	2.213
RF_BS_m	0.071	0.658	0.375	0.112	0.597	0.373	0.188	2.025
RF_BS_p	0.073	0.659	0.389	0.111	0.621	0.389	0.188	2.110
RF_BS_m_32	0.063	0.648	0.420	0.105	0.638	0.425	0.181	1.995
RF_BS_p_32	0.067	0.652	0.417	0.106	0.639	0.422	0.182	2.069
RF_BS_m_8	0.074	0.661	0.328	0.118	0.542	0.318	0.194	2.084
RF_BS_p_8	0.076	0.661	0.351	0.115	0.575	0.345	0.192	2.200

Notes: 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, n.a. = not available.

Table B.8: Evaluation metrics Safety of Navigation and Radio Communication

Model	H	AUC	ER	Precision	TPR	FPR	F	TopDecile
RF_m	0.126	0.691	0.131	0.308	0.157	0.044	0.208	2.655
RF_p	0.130	0.699	0.130	0.308	0.154	0.043	0.206	2.637
RF_m_32	0.115	0.688	0.138	0.285	0.170	0.053	0.213	2.506
RF_p_32	0.119	0.691	0.135	0.291	0.164	0.049	0.209	2.538
RF_m_8	0.127	0.686	0.127	0.315	0.138	0.037	0.192	2.698
RF_p_8	0.136	0.702	0.127	0.321	0.142	0.037	0.197	2.700
BRF_m	0.123	0.696	0.333	0.188	0.617	0.327	0.288	2.489
BRF_p	0.131	0.700	0.344	0.185	0.632	0.341	0.287	2.647
BRF_m_32	0.114	0.690	0.379	0.174	0.658	0.383	0.275	2.415
BRF_p_32	0.121	0.694	0.379	0.174	0.659	0.383	0.276	2.493
BRF_m_8	0.129	0.698	0.299	0.200	0.578	0.284	0.297	2.586
BRF_p_8	0.135	0.703	0.314	0.196	0.601	0.304	0.295	2.679
RF_BS_m	0.126	0.698	0.339	0.187	0.628	0.335	0.289	2.566
RF_BS_p	0.132	0.701	0.349	0.185	0.642	0.348	0.287	2.631
RF_BS_m_32	0.119	0.692	0.382	0.173	0.662	0.388	0.275	2.525
RF_BS_p_32	0.124	0.696	0.382	0.174	0.669	0.389	0.277	2.562
RF_BS_m_8	0.132	0.700	0.298	0.202	0.586	0.284	0.301	2.632
RF_BS_p_8	0.137	0.704	0.316	0.196	0.607	0.307	0.296	2.708

Notes: 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, n.a. = not available.

Table B.9: Evaluation metrics Safety Management (ISM)

Model	H	AUC	ER	Precision	TPR	FPR	F	TopDecile
RF_m	0.045	0.614	0.064	0.213	0.005	0.001	0.010	1.908
RF_p	0.045	0.618	0.064	0.179	0.004	0.001	0.008	1.863
RF_m_32	0.039	0.612	0.067	0.150	0.015	0.006	0.028	1.847
RF_p_32	0.042	0.615	0.065	0.192	0.012	0.003	0.022	1.841
RF_m_8	0.045	0.608	0.063	0.156	0.001	0.000	0.002	1.927
RF_p_8	0.049	0.623	0.063	0.175	0.001	0.000	0.002	1.844
BRF_m	0.041	0.621	0.362	0.090	0.522	0.354	0.153	1.745
BRF_p	0.046	0.622	0.381	0.089	0.546	0.376	0.152	1.817
BRF_m_32	0.039	0.615	0.416	0.086	0.580	0.415	0.149	1.794
BRF_p_32	0.043	0.618	0.416	0.085	0.580	0.416	0.149	1.828
BRF_m_8	0.042	0.622	0.324	0.093	0.475	0.311	0.155	1.726
BRF_p_8	0.045	0.624	0.350	0.092	0.513	0.340	0.155	1.806
RF_BS_m	0.040	0.619	0.360	0.090	0.518	0.352	0.153	1.692
RF_BS_p	0.043	0.620	0.377	0.088	0.536	0.371	0.152	1.809
RF_BS_m_32	0.037	0.612	0.410	0.085	0.565	0.408	0.148	1.763
RF_BS_p_32	0.039	0.614	0.410	0.085	0.567	0.409	0.148	1.774
RF_BS_m_8	0.041	0.620	0.320	0.093	0.465	0.305	0.154	1.729
RF_BS_p_8	0.045	0.623	0.345	0.092	0.505	0.335	0.155	1.826

Notes: 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, n.a. = not available.

Table B.10: Evaluation metrics MARPOL A1 to A3- Oil and HNS

Model	H	AUC	ER	Precision	TPR	FPR	F	TopDecile
RF_m	0.038	0.615	0.020	0.000	0.000	0.000	0.000	1.802
RF_p	0.043	0.622	0.020		0.000	0.000		1.898
RF_m_32	0.030	0.606	0.020	0.026	0.001	0.000	0.001	1.576
RF_p_32	0.036	0.613	0.020	0.000	0.000	0.000	0.000	1.752
RF_m_8	0.041	0.615	0.020		0.000	0.000		1.973
RF_p_8	0.048	0.630	0.020		0.000	0.000		1.998
BRF_m	0.049	0.633	0.398	0.029	0.590	0.398	0.056	1.938
BRF_p	0.053	0.638	0.419	0.029	0.608	0.420	0.055	2.053
BRF_m_32	0.048	0.632	0.452	0.028	0.651	0.454	0.054	1.913
BRF_p_32	0.051	0.635	0.459	0.028	0.648	0.461	0.053	2.038
BRF_m_8	0.049	0.633	0.359	0.030	0.545	0.357	0.057	1.898
BRF_p_8	0.054	0.639	0.385	0.030	0.579	0.384	0.057	2.033
RF_BS_m	0.048	0.633	0.420	0.029	0.609	0.420	0.055	1.908
RF_BS_p	0.050	0.634	0.437	0.028	0.623	0.438	0.054	1.918
RF_BS_m_32	0.044	0.626	0.463	0.027	0.641	0.465	0.052	1.822
RF_BS_p_32	0.047	0.630	0.468	0.027	0.653	0.471	0.053	1.923
RF_BS_m_8	0.048	0.631	0.371	0.030	0.561	0.370	0.057	1.883
RF_BS_p_8	0.052	0.638	0.404	0.030	0.605	0.405	0.056	1.973

Notes: 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, n.a. = not available.

Table B.11: Evaluation metrics MARPOL A4 and 5: Sewage and Garbage

Model	H	AUC	ER	Precision	TPR	FPR	F	TopDecile
RF_m	0.040	0.615	0.029	0.000	0.000	0.000	0.000	1.912
RF_p	0.047	0.629	0.029	0.000	0.000	0.000	0.000	1.902
RF_m_32	0.035	0.613	0.029	0.068	0.001	0.001	0.003	1.754
RF_p_32	0.038	0.618	0.029	0.071	0.000	0.000	0.001	1.795
RF_m_8	0.040	0.610	0.029	0.000	0.000	0.000	0.000	2.033
RF_p_8	0.056	0.641	0.029		0.000	0.000		2.064
BRF_m	0.054	0.641	0.386	0.044	0.591	0.386	0.081	1.905
BRF_p	0.057	0.645	0.401	0.043	0.609	0.401	0.081	1.985
BRF_m_32	0.046	0.631	0.441	0.041	0.637	0.443	0.077	1.833
BRF_p_32	0.051	0.638	0.445	0.041	0.649	0.448	0.078	1.864
BRF_m_8	0.056	0.643	0.335	0.046	0.532	0.331	0.084	1.919
BRF_p_8	0.061	0.649	0.363	0.045	0.563	0.361	0.082	2.044
RF_BS_m	0.054	0.642	0.391	0.043	0.583	0.390	0.080	1.905
RF_BS_p	0.057	0.644	0.401	0.043	0.606	0.401	0.080	2.037
RF_BS_m_32	0.047	0.634	0.436	0.041	0.625	0.438	0.077	1.833
RF_BS_p_32	0.052	0.638	0.438	0.041	0.638	0.440	0.078	1.943
RF_BS_m_8	0.056	0.643	0.333	0.045	0.520	0.329	0.083	1.971
RF_BS_p_8	0.060	0.648	0.362	0.044	0.559	0.359	0.082	2.037

Notes: 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, n.a. = not available.

Table B.12: Evaluation metrics MARPOL A6: Air Pollution

Model	H	AUC	ER	Precision	TPR	FPR	F	TopDecile
RF_m	0.018	0.574	0.010		0.000	0.000		1.505
RF_p	0.021	0.580	0.010		0.000	0.000		1.535
RF_m_32	0.014	0.565	0.010	0.063	0.001	0.000	0.002	1.444
RF_p_32	0.019	0.574	0.010		0.000	0.000		1.505
RF_m_8	0.018	0.576	0.010		0.000	0.000		1.555
RF_p_8	0.024	0.586	0.010		0.000	0.000		1.585
BRF_m	0.023	0.589	0.349	0.014	0.479	0.347	0.027	1.585
BRF_p	0.023	0.589	0.368	0.014	0.498	0.367	0.026	1.545
BRF_m_32	0.022	0.588	0.405	0.013	0.537	0.405	0.026	1.555
BRF_p_32	0.023	0.587	0.398	0.013	0.521	0.398	0.025	1.525
BRF_m_8	0.021	0.585	0.304	0.014	0.416	0.301	0.027	1.495
BRF_p_8	0.023	0.589	0.341	0.014	0.461	0.339	0.026	1.525
RF_BS_m	0.023	0.589	0.370	0.014	0.502	0.369	0.026	1.545
RF_BS_p	0.024	0.589	0.387	0.013	0.520	0.386	0.026	1.485
RF_BS_m_32	0.023	0.587	0.418	0.013	0.552	0.418	0.026	1.625
RF_BS_p_32	0.025	0.590	0.414	0.013	0.551	0.413	0.026	1.515
RF_BS_m_8	0.021	0.584	0.318	0.014	0.431	0.315	0.026	1.444
RF_BS_p_8	0.021	0.585	0.352	0.013	0.470	0.350	0.026	1.535

Notes: 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, n.a. = not available.

Table B.13: Evaluation metrics Ballast Water and Antifouling

Model	H	AUC	ER	Precision	TPR	FPR	F	TopDecile
RF_m	0.005	0.532	0.017		0.000	0.000		1.282
RF_p	0.006	0.536	0.017		0.000	0.000		1.193
RF_m_32	0.004	0.526	0.017	0.071	0.001	0.000	0.001	1.246
RF_p_32	0.005	0.526	0.017		0.000	0.000		1.270
RF_m_8	0.006	0.537	0.017		0.000	0.000		1.151
RF_p_8	0.007	0.541	0.017		0.000	0.000		1.217
BRF_m	0.011	0.558	0.293	0.020	0.351	0.287	0.039	1.205
BRF_p	0.012	0.559	0.291	0.020	0.346	0.285	0.038	1.151
BRF_m_32	0.009	0.547	0.371	0.019	0.421	0.367	0.037	1.175
BRF_p_32	0.010	0.551	0.356	0.019	0.410	0.352	0.037	1.163
BRF_m_8	0.016	0.570	0.222	0.022	0.280	0.213	0.041	1.163
BRF_p_8	0.015	0.568	0.236	0.022	0.302	0.228	0.041	1.211
RF_BS_m	0.014	0.565	0.320	0.021	0.399	0.316	0.040	1.205
RF_BS_p	0.013	0.562	0.312	0.021	0.388	0.307	0.040	1.181
RF_BS_m_32	0.011	0.558	0.369	0.021	0.450	0.366	0.039	1.199
RF_BS_p_32	0.012	0.557	0.363	0.021	0.443	0.360	0.039	1.157
RF_BS_m_8	0.016	0.570	0.262	0.023	0.352	0.255	0.043	1.169
RF_BS_p_8	0.017	0.573	0.270	0.023	0.365	0.264	0.043	1.222

Notes: 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, n.a. = not available.

Appendix C: Importance Plots

Legend for 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_BV	Class - Bureau Veritas
CL_NK	Class - Nippon Kaiji Kyokai
CL_UNKN	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
OC_HIGH	DOC company from a high income country
DOC_YEX	DOC company years of experience
DOCH3Y	DOC changes within 3 years
FLCH3Y	Flag changes within 3 years
FL_BZE	Flag Belize
FL_PAN	Flag Panama
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)
LDET	Previous number of detentions (365 days)
MEB_CHR	Main engine builder located in China
MEB_GEU	Main engine builder located in Germany
MEB_JPN	Main engine builder located in Japan
MEB_KRS	Main engine builder located in South Korea
MED_CST	Main engine designer (Chinese Std. Type)
MED_HAN	Main engine designer (Hanshin)
MED_MBW	Main engine designer (MAN-B&W)
WN_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_GEN	General cargo Ship
ST_TANK	Tanker
SY2	Ship Yard Country Group 2 (high risk)
SY3	Ship Yard Country Group 3 (medium risk)
SY4	Ship Yard Country Group 4 (low risk)

Figure C.1: Importance plots: Detention

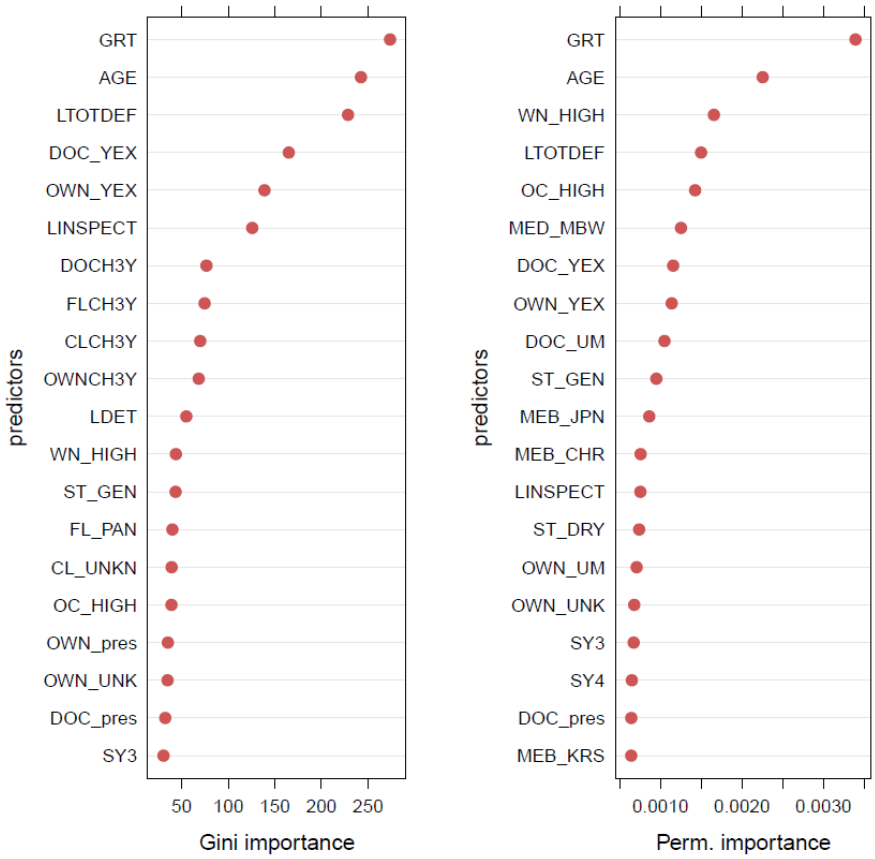


Figure C.2: Importance plots: Certificates and Qualifications

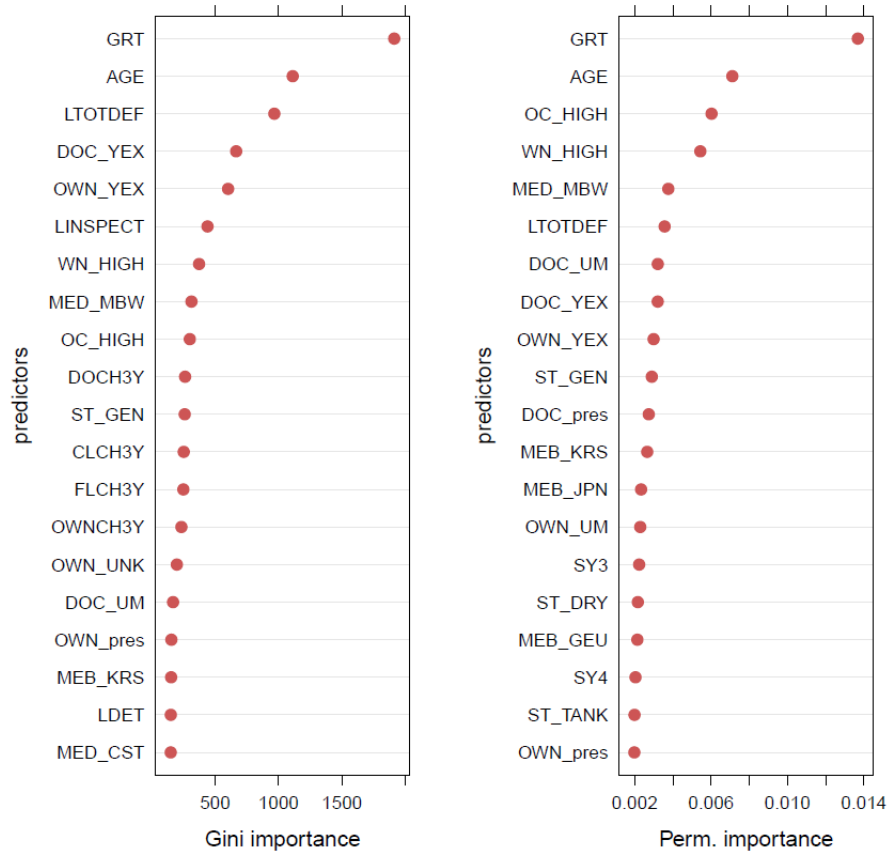


Figure C.3: Importance plots: Maritime Labour Convention

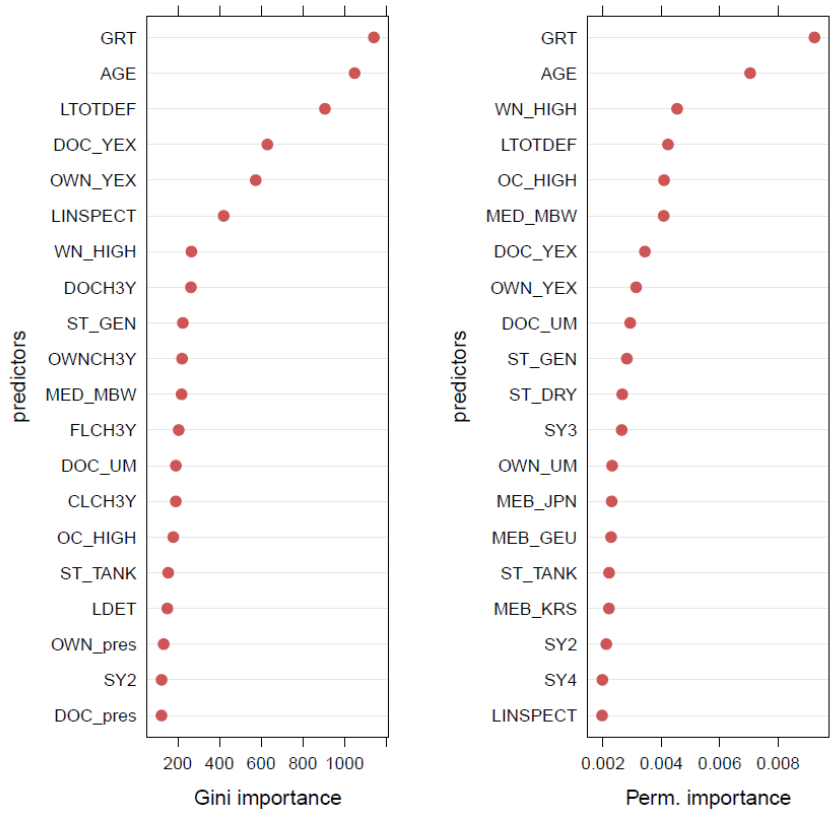


Figure C.4: Importance plots: Structural and Watertight Integrity

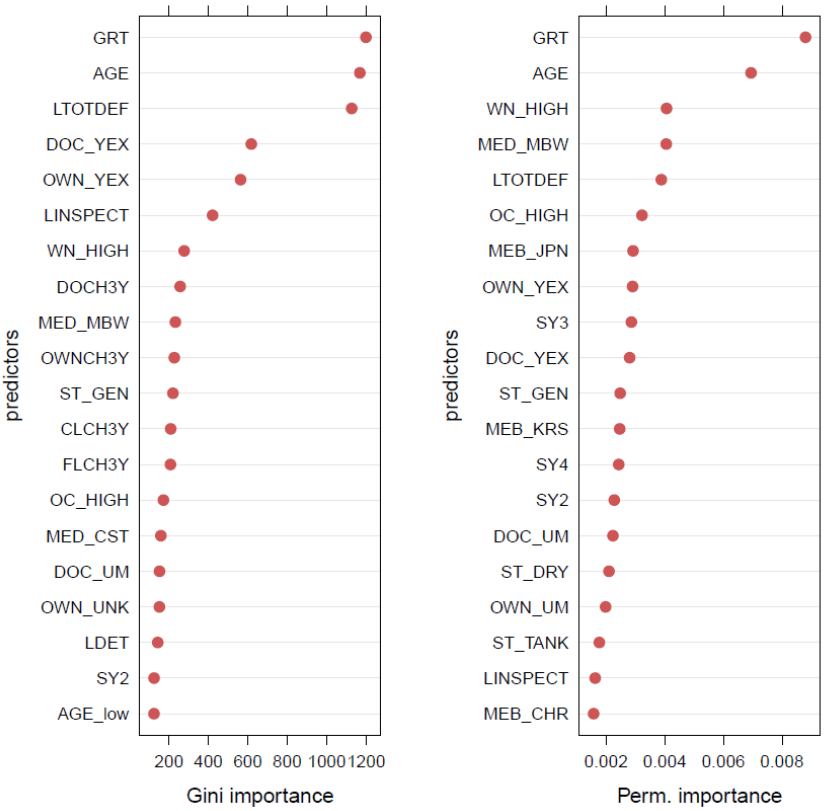


Figure C.5: Importance plots: Propulsion and Machinery

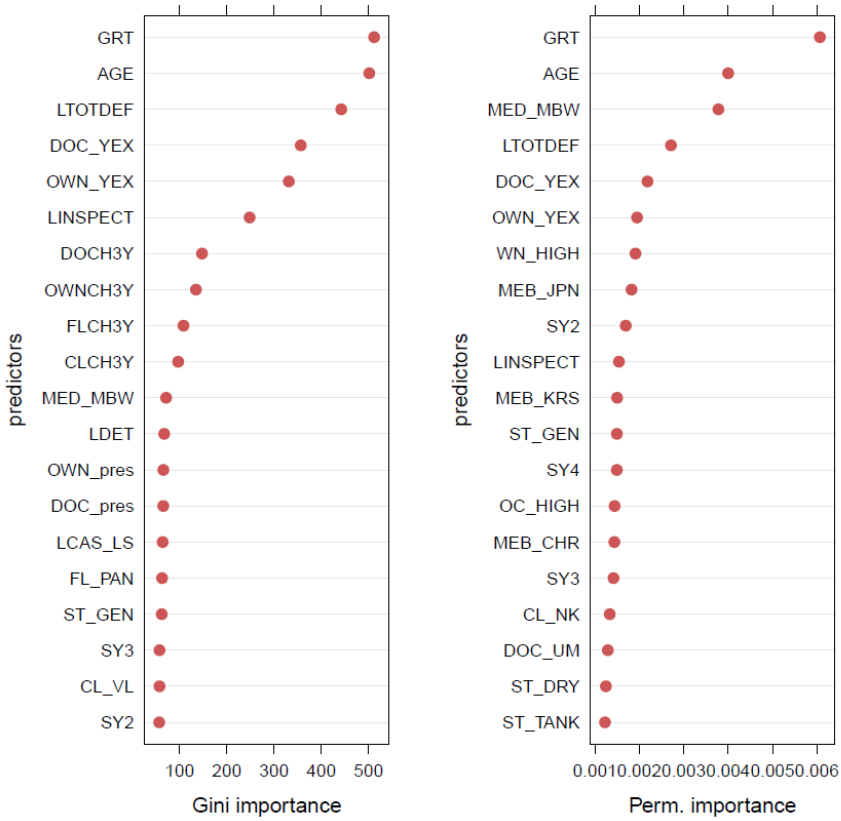


Figure C.6: Importance plots: Life Saving and Fire Appliances

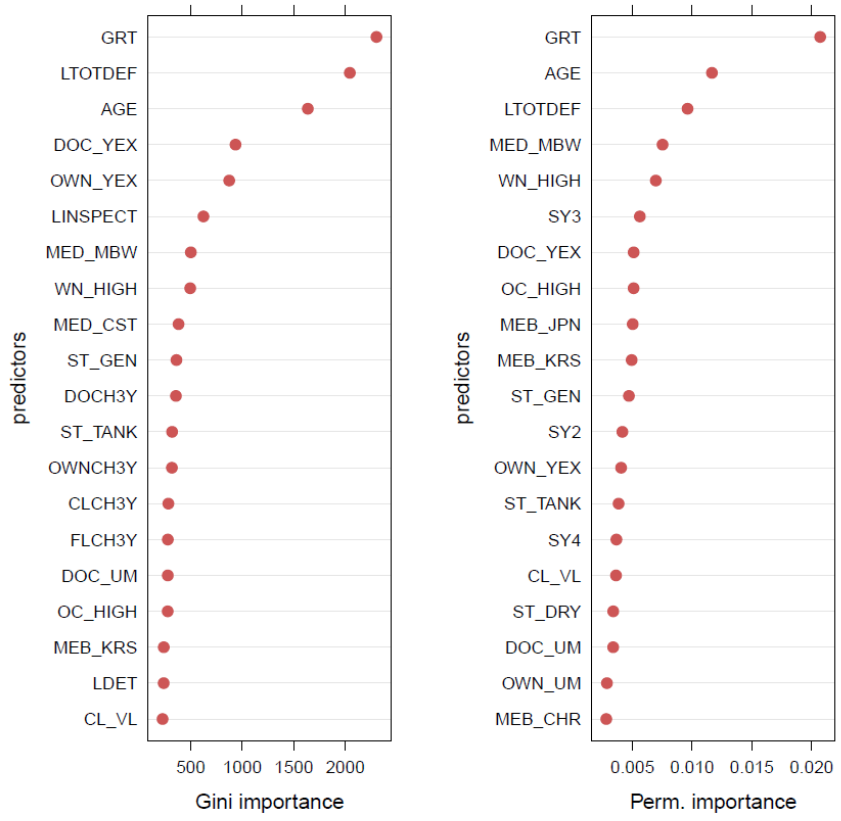


Figure C.7: Importance plots: Emergency Systems and Alarms

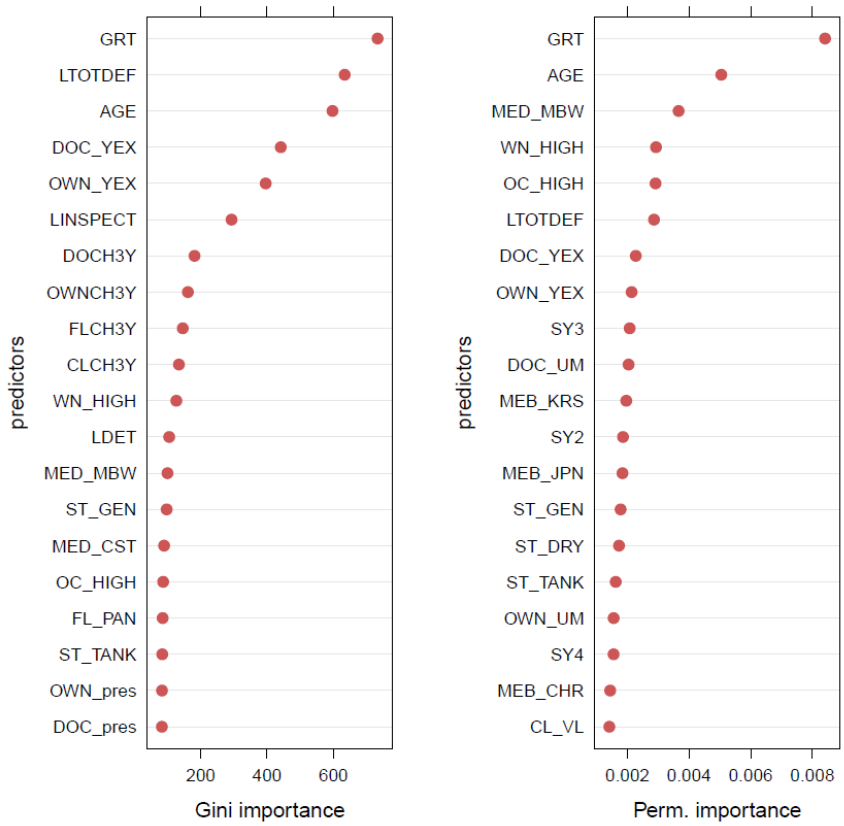


Figure C.8: Importance plots: Safety of Navigation and Radio Communication

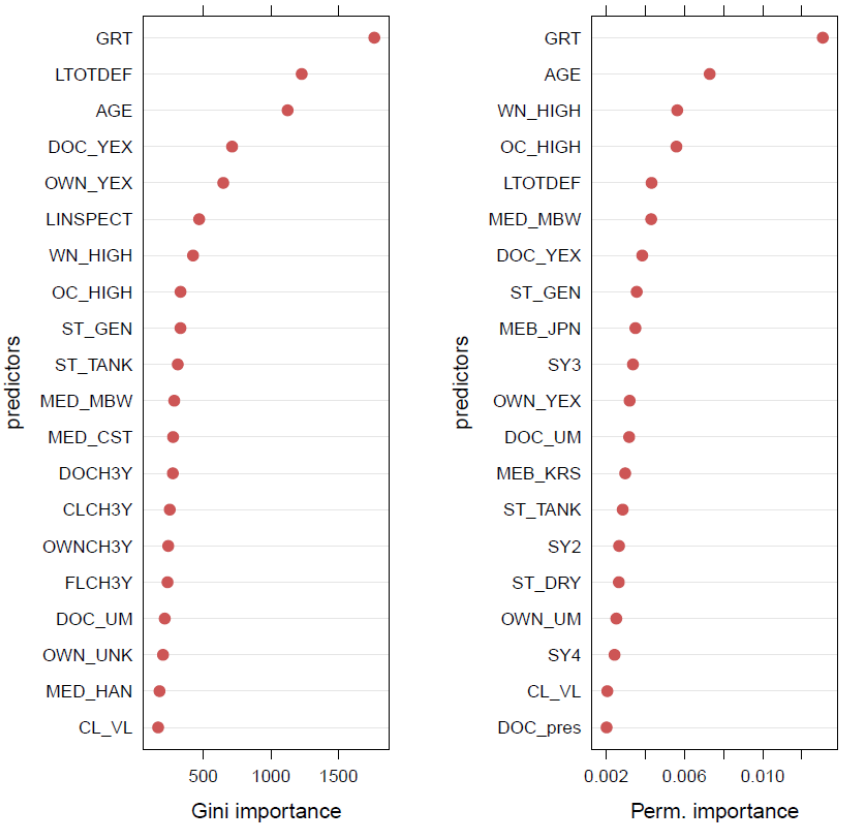


Figure C.9: Importance plots: Safety Management (ISM)

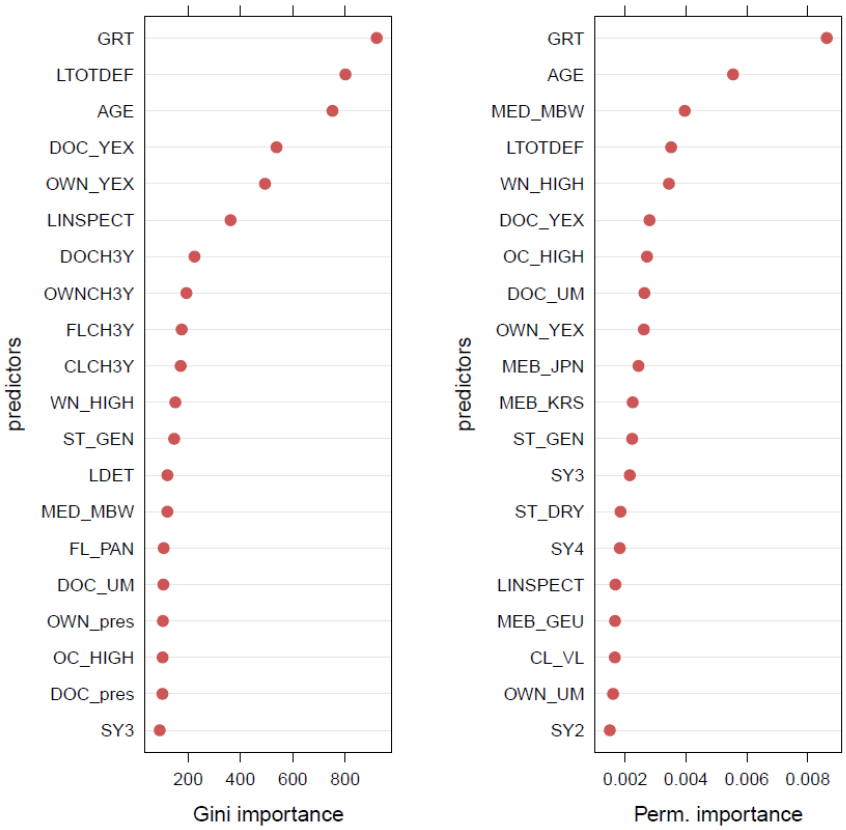


Figure C.10: Importance plots: MARPOL A1 to A3- Oil and HNS

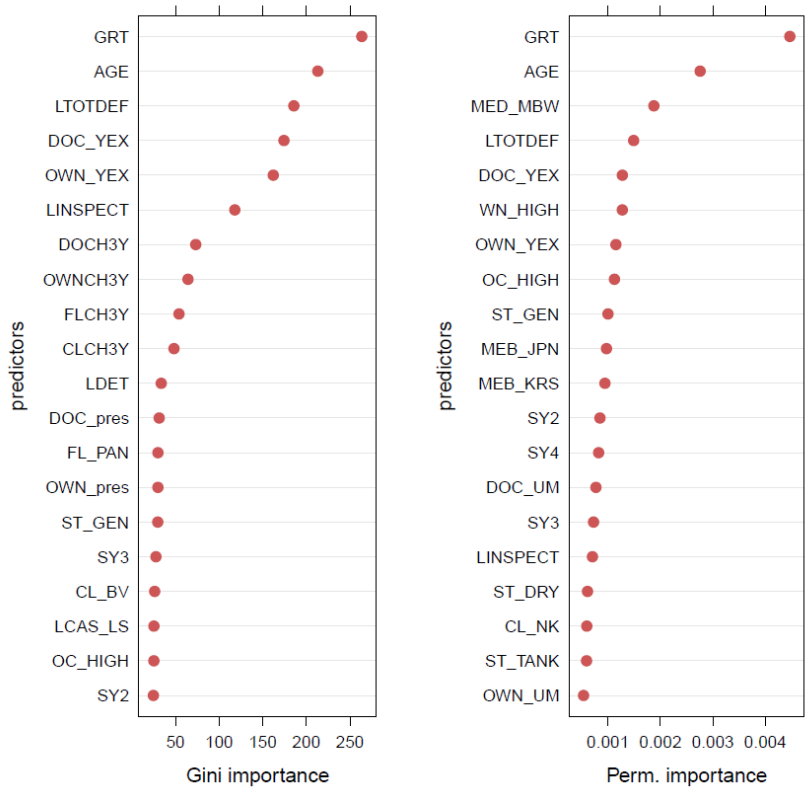


Figure C.11: Importance plots: MARPOL A4 and 5: Sewage and Garbage

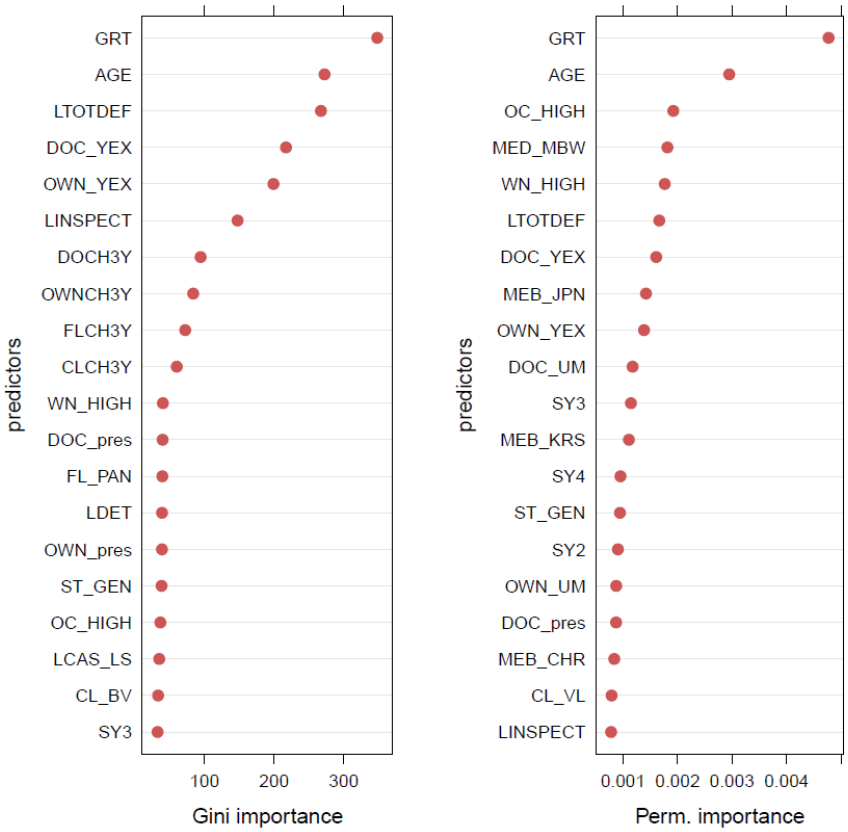


Figure C.12: Importance plots: MARPOL A6: Air Pollution

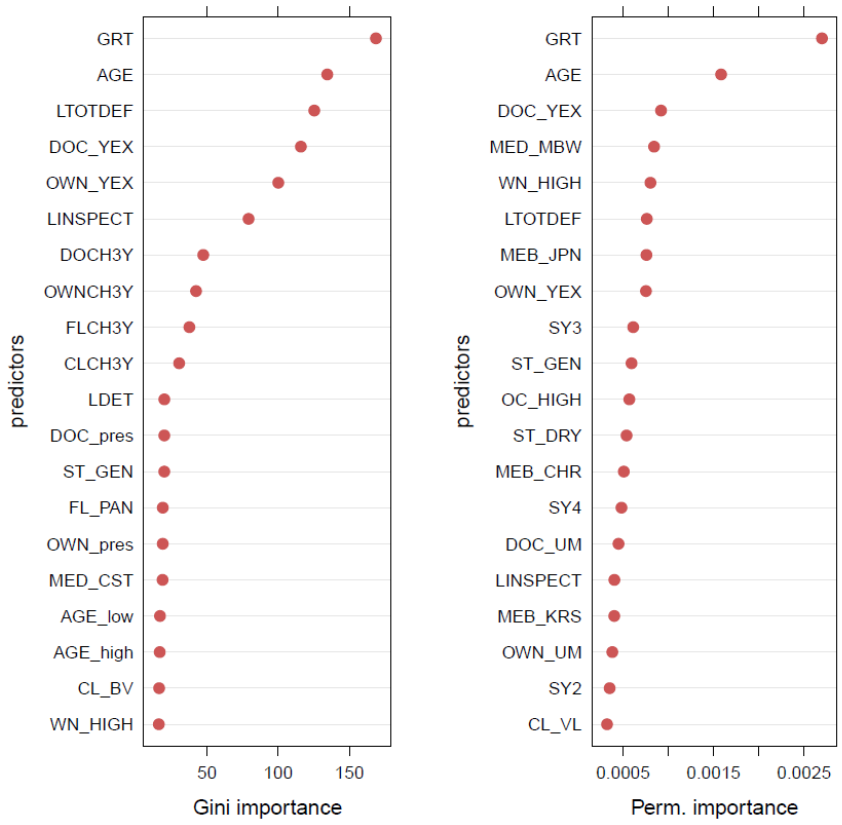


Figure C.13: Importance plots: Ballast Water and Antifouling