Universiteit Leiden

Master Computer Science

Explainable AI for ship design analysis with AIS and static ship data

Name:	Lucas van Rooij
Student ID:	s1695185
Date:	21-06-2022
Specialisation:	Artificial Intelligence
1st supervisor:	Bas van Stein
2nd supervisor:	Anna Kononova

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Decisions made in the early phases of ship design have a large influence on the building and operational costs of a vessel. In order to support decision making in this phase, big data and machine learning techniques can be of great use. This thesis shows how Explainable Artificial Intelligence(XAI) and Global Sensitivity Analysis (GSA) combined with Autonomous Identification System (AIS) and static ship data can be used to find important design characteristics of ships. A data collection framework is setup that collects AIS data over a five month time period. Static ship design data is used to predict performance related target features that are calculated from AIS data. By applying XAI and GSA methods to the regression models that predict these target features, we gain insight in how design features influence the performance of ships. We find that for most ship types the overall length is the most important design feature for speed related target features. For rotation related target features we also find that the draught is an important design feature.

Contents

1 Introduction

Over the recent years the application of Artificial Intelligence (AI) in the maritime industry has seen an increase [1]. Examples of recent applications range from fleet allocation to ETA prediction [2]. These applications have caused a large increase in efficiency and performance of vessels throughout the whole maritime industry [2]. A sub sector of the maritime industry that has only recently seen an increase in applications of AI, is the naval architect industry [3]. Applications of AI and data driven decision making can be of huge benefit for naval architects, as many early phase design decisions previously relied on the experience of engineers or a handful of reference vessels [4]. Any wrong assumptions made in this phase can lead to higher production, or operational costs, which highlights the importance of this phase [5]. In order to come up with more cost efficient and better performing ships, engineers need a clear picture of a ship's desired operational profile. Here, data driven technologies can provide engineers a helping hand. Previous applications of AI systems in the naval architect industry mainly focused on data collected from a relatively small group of vessels. Examples of these are Neural Network aided design of ship hull structures [6], and propulsion power optimum calculation via genetic algorithms [7].

Together with an increase in applications of AI methods, the predictive capabilities of AI models has also increased [8]. As the performance of these models increase, they also tend become more complex and their workings harder to understand. In order to understand the workings of complex black box predictive models, Sensitivity Analysis (SA) and Explainable AI (XAI) can help. These methods define measures or visualisations of important input features of predictive black box models. In this research we will combine these tools with data collected from a large group of vessels in order to obtain a high level view on what ship design parameters are distinctive for a vessel's performance capabilities. The data used in this work consists of Autonomous Identification System (AIS) data, for which we create a data collection framework. Next to AIS data we use static ship data provided by C-Job, a Dutch Naval Architect bureau. The contribution of this work is two-fold:

1. We apply XAI and SA methods to a Multi-Output regression model built by an Autosklearn 2.0 AutoML pipeline.
2. We use AIS data to analyze the world fleet in order to find important design characteristics for commercial ships.

The remaining of this thesis is structured as follows: Section 2 describes the background, related work and methods used in this work. Section 3 defines the problem statement. Hereafter, Section 4 explains how the data is collected, processed, and stored. Section 5 then shows the experiments and results, which are discussed in Section 6. Finally, Section 7 concludes.

2 Background and Related Work

This Section describes the background and related work of the methods and data that are used in this research. The methods include: regression models, SA techniques, and XAI techniques. These methods are selected as the combination of these methods provides an overall view of the important features in the data.

2.1 Related AIS data research

The Autonomous Identification System (AIS) has mainly been developed as a system that can be used for collision avoidance and navigational safety for almost all ships in the commercial fleet. The system was initially developed in 2002, but had limited coverage over only the coastal area waters. In 2008 satellite AIS transponders were introduced which increased the coverage to almost all areas in the world. With the increase in coverage and data quality new applications for the AIS appeared. Examples of this are, ship trajectory extraction and prediction [9], ship activity tracking, tracking for environmental monitoring [10], or tracking in restricted waters such as locks and canals [11]. Next to vessel tracking, AIS data can also be used to measure various port statistics such as the number of daily visiting ships, or waiting times [12]. To the best of our knowledge AIS data has never been used to compare ship design characteristics of the commercial fleet on a worldwide scale. More on how AIS data is used in this research is explained in Section 4.1.

2.2 Regression models

In previous works regression models have been used in order to predict numerical values given a set of input features [13]. These regression models can then be analysed in order to find which input features have the highest contribution to the output as explained later on in Section 2.5. The following subsections discuss how Linear regression, Support Vector Regression, XGBoost, and Random Forest regression have previously been used and how they can be used for SA and XAI methods.

2.2.1 Linear regression

Linear regression is a regression model often used for it's simplicity and interpretability [14]. Linear regression uses a linear combination of independent variables to predict the dependent variable. Linear regression models the relation

$$y = X\beta + \epsilon$$

Where y is the dependent variable, X is the vector of independent variables or input features, β are the regression coefficients or weights, and ϵ is a noise term. Linear regression uses Ordinary Least Squares (OLS) to find the weights that minimize the error. Furthermore, linear regression assumes that the relation between the dependent and independent variables is linear. In the case where the input data is non-linear, alternative regression models are needed.

2.2.2 Support Vector Regression

Support vector regression (SVR) [15] is an alternation to the widely used Support Vector Machine (SVM) [16] classification algorithm. SVR tries to find a line or hyperplane that fits the data, while giving the user the flexibility to define how much error in the model is acceptable. If the hyperplane can not be found in the current dimension, a kernel is used to transform the data into a higher dimension before searching again for a hyperplane. The application of kernels allows SVR to also model non-linear data. Training an SVR model works by solving the following objective function

$$min \;\; \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}|\xi_i|$$

With constraint function

$$|y_i - \mathbf{w}_i\mathbf{x}_i| \leq \epsilon + |\xi_i|$$

The term C in the objective function defines how far datapoints are allowed to lie outside of the decision boundaries. ξ_i is the distance of sample i to the decision boundary. The decision boundaries are defined by ϵ, which is a parameter set by the user. Figure 1 shows a one dimensional example of an SVR where the red line is the fitted line that is used to predict new samples. The two grey lines are the decision boundaries that lie a distance ϵ from the fitted line.

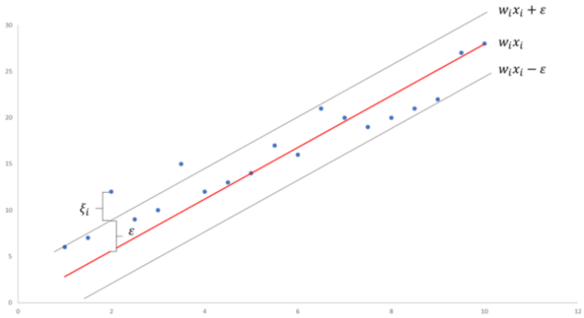


Fig. 1: Overview of a one-dimensional SVR model [17]. The red line is used to make predictions.

2.2.3 Gradient boosting with XGboost

Gradient boosting is a tree based classification or regression algorithm [18] and relies on three main ideas. First, Gradient boosting uses a loss function that needs to be optimized. Second, it uses weak learners to make predictions, weak learners are cheap and simple models that often perform slightly better than a random model [19]. Finally gradient boosting uses an additive model to combine the predictions of weak learners to minimize the loss function. The loss function must be differentiable, for most regression cases the squared error function is used. Most often, regression trees with a limited number of splits are used as weak learners. The additive model adds the trees one at a time using only the mispredictions to construct new weak learners. Extreme gradient boosting [20], or XGboost is an open source implementation of the gradient boosting algorithm that is widely used in practice. In addition to gradient boosting, XGboost uses L1 and L2 regularization which improve model generalization.

2.2.4 Random Forest Regression

Random Forest (RF) regression is a regression algorithm that similar to gradient boosting uses an ensemble of multiple models to come up with a more accurate prediction than a single model [21]. The algorithm selects a subset of k random features. For these k features it then builds a regression tree. These two steps are repeated N times, resulting in N decision trees. In order to come up with a prediction for an unseen sample, the model predicts a value for each regression tree and then averages the prediction of all N trees. The algorithm is often used for its strong performance on regression tasks [22, 23]. The disadvantage of RF regression models is that they lack interpretability and tend to overfit faster than other methods.

2.3 Cross validation and hyperparameter optimization

In many cases, the default hyperparameter settings of a machine learning algorithm yield a good performance. However, some improvements can often be made by optimizing the combination of hyperparameters. In order to obtain a trustworthy report of the performance of a model under a certain combination of hyperparameters a model needs to be evaluated on as much unseen data as possible, without sacrificing training data. In other works this is done by performing k-fold cross validation [24]. A schematic overview of 5-fold cross validation is shown in Figure 2.This method partitions the training data into k folds, each time training a model on different $k-1$ folds, and evaluating on the remaining fold. After training the model k times, the performance over the k evaluations is averaged and reported as the score for the model under the current hyperparameter combination. A specific case of k-fold cross validation is leave-one-out cross validation. This method works the same as cross validation, but takes the number of samples n as value for k. The best combination of hyper parameters is the combination that yields the best performance over the k, for k-fold, or n, for leave-one-out, validations. In practice,

k-fold is often used for it's computational benefits over leave-one-out. After optimizing the hyperparameters in combination with cross validation, a final evaluation can be made on the test set in order to see how the optimized model performs on unseen data.

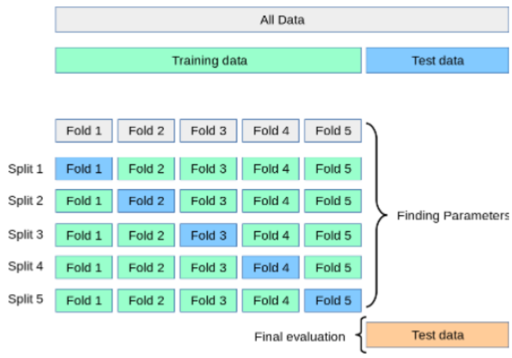


Fig. 2: 5-fold cross validation. Cross validation is performed after each time that a new combination of hyperparameter values is introduced. Only the final optimized model is evaluated on the test data. Image from [25].

2.4 AutoML pipeline

Automated machine learning helps machine learning engineers with the task of designing machine learning pipelines and has recently achieved promising results [26]. The *Auto-sklearn 2.0* framework, an improvement of *Auto-sklearn 1.0*, offers an implementation of an AutoML system that can perform well on large datasets under strict time constraints by using a meta-feature-free meta-learning technique and applying a bandit strategy for budget allocation. As *Auto-sklearn 2.0* is an improvement on it's precedor, Section 2.4.1 explains the general workings of *Auto-sklearn 1.0* before Section *Auto-sklearn 2.0* is discussed in 2.4.2.

2.4.1 Auto-sklearn 1.0

Figure 3 gives a schematic overview of the pipeline that *Auto-sklearn 1.0* executes for training input $\{\mathbf{X}_{train}, \mathbf{Y}_{train}\}$, test input $\mathbf{X}_{test}$, time budget b and loss function $\mathcal{L}$. The overview shows that the first step is the meta-learning phase; an offline phase used to create a set of meta features for 140 datasets from the OpenML repository [27]. For each dataset the meta-features are evaluated and Bayesian Optimization (BO) is performed on ML frameworks constructed from a selection of 15 regression algorithms, 14 preprocessing methods, and 4 data preprocessing methods. If an iteration of BO yields good performance on the dataset in question, the instantiation of the ML framework that is currently being evaluated is stored. Then for the training input $\{\mathbf{X}_{train}, \mathbf{Y}_{train}\}$ the $k = 25$ closest datasets in meta-feature space are used for evaluation before BO is started on their results. After the BO phase, an ensemble is built consisting of the best found ML frameworks. The ensemble is then used to make predictions $\hat{\mathbf{Y}}_{test}$ for $\mathbf{X}_{test}$.

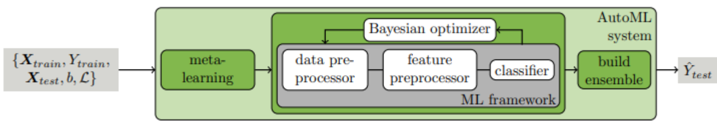


Fig. 3: Schematic overview of the Auto-sklearn 1.0 pipeline [28].

2.4.2 Auto-sklearn 2.0

Figure 4 gives a schematic overview of the AutoML pipeline implemented by *Auto-sklearn 2.0*. Just as for *Auto-sklearn 1.0*, the pipeline has $\mathbf{X}_{train}$, $\mathbf{Y}_{train}$ as training input together with $\mathbf{X}_{test}$ which is used to generate the output $\hat{\mathbf{Y}}_{test}$, b is the time budget and $\mathcal{L}$ is the loss function that needs to be minimized. The pipeline differs from *Auto-sklearn 1.0* in the two phases before the AutoML system is executed. In the first phase a budget allocation together with a model selection strategy is selected, and in the second phase a portfolio of candidate pipelines is constructed.

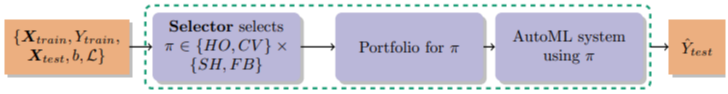


Fig. 4: Schematic overview of the Autosklearn 2.0 pipeline [29].

In the first phase *Auto-sklearn 2.0* selects an optimization policy π, which is a combination of AutoML system hyperparameters and components to be used during a run of the AutoML system. π consists of a model selection strategy, either Hold Out (HO), or Cross-Validation (CV) and a budget allocation strategy, either Successive Halving (SH), or Full Budget (FB). HO splits the set into one training and one test part, CV splits the data as explained in section 2.3. SH works by setting a minimal and maximal budget for evaluating a pipeline. Then, it iteratively selects $\frac{1}{\eta}$ pipelines with the lowest generalization error, multiplies their budget by η, removes the pipelines that are not selected and repeats the process. The process continues until there is one pipeline left or the maximal budget has been spent, in this case the pipeline with the current lowest error is selected.

Making a decision about which policy to use is done by the policy selector. The policy selector works by fitting an RF classifier that predicts whether policy π_A outperforms π_B for each policy given the meta-features of the current dataset. The meta-features used in this phase are very simple and can be calculated in linear time: the number of data points, and the number of features. These meta-features are different from the more complex meta-features used by *Auto-sklearn 1.0*.

After selecting the optimization policy π, a portfolio with ML pipelines is created from a finite set of candidate pipelines $\mathcal{C} = \{\lambda_1, ..., \lambda_l\}$. This step replaces the meta-feature calculation phase in *Auto-sklearn 1.0*. *Autosklearn 2.0* then uses a set of meta-datasets $\mathbf{D}_{meta} = \{\mathcal{D}_1, ..., \mathcal{D}_{|D_{meta}|}\}$ to build a portfolio $\mathcal{P}$ that performs well on $\mathbf{D}_{meta}$. The portfolio is built by starting with an empty set and adding candidates $\lambda^* \in \mathcal{C}$ in a greedy way to $\mathcal{P}$ such that the generalization error over all meta-datasets is reduced the most. This procedure continues until $|\mathcal{P}|$ reaches a predefined limit value. The generalization error on a single dataset $\mathcal{D}$ is the error of the best performing pipeline $\lambda \in \mathcal{P}$ on $\mathcal{D}$. The pipelines in $\mathcal{P}$ are then evaluated on $\mathbf{X}_{train}$ with policy π in order to come up with a final model, or ensemble of models.

2.5 Global sensitivity analysis and Explainable AI methods

Global Sensitivity Analysis (GSA) and Explainable AI (XAI) methods help in understanding the workings of a complex model by determining the contributions of individual features to the output of a model [30]. It gives an overall view on the influence of inputs on model outputs. GSA methods aim to determine how much the output of a model generally changes with a change in input. Consider a model f, the goal is to understand the relationship

$$y = f(\mathbf{x})$$

Here f can be any function or model ranging from a simple linear regression to a complex neural network [31]. $\mathbf{x} = |x_1, x_2, ..., x_m| \in \mathbf{X} \subseteq \mathbb{R}^m$, is a vector of model inputs that determines the model output. The goal of GSA and XAI is to understand how much change in the model output $y \in \mathbb{R}$ can be attributed to a specific x_i. GSA and XAI use metrics and visualisations that show how much a change in output can be attributed to a single, or combination of input features. In previous works GSA has been proven to be useful for many applications such as: determining the optimal lot size for production systems [32], understanding the workings of chemical models [33], or measuring the influence of input parameters on biological models [34].

GSA can be divided into variance based, derivative based, and density based methods. Examples of variance based methods are: Fourier Amplitude Sensitivity Test (FAST), and Sobol indices. Derivative based methods used in earlier works are the Morris method and Derivative Based Global Sensitivity Measures (DGSM). Previously used density based GSA methods are DELTA, and PAWN. A short overview of the workings of these methods is given in the following subsections.

2.5.1 FAST

FAST decomposes the variance of the model into partial variances that can be attributed to individual input features by using a periodic sampling method and a Fourier transform [35].

2.5.2 Sobol indices

Sobol indices are a form of variance based GSA [36], and can be used to determine the importance of input features. This method decomposes the variance of the output of a machine learning model into fractions that can be attributed to features, or combinations of features. A function $Y = f(\mathbf{X})$ can be decomposed as follows

$$Y = f_0 + \sum_{i=1}^{d} f_i(X_i) + \sum_{i<j}^{d} f_{ij}(X_i, X_j) + \cdots + f_{1,2,\ldots,d}(X_1, X_2, ..., X_d)$$

With constant f_0, and f_i as function of X_i. The variance of Y can then be expressed as:

$$V(Y) = \sum_{i=1}^{d} V_i + \sum_{i<j}^{d} V_{ij} + \cdots + V_{1,\ldots,d}$$

Where

$$V_i = V(E(Y|X_i))$$

And

$$V_{ij} = V(E(Y \mid X_i, X_j)) - V_i - V_j$$

Here $E(Y|X_i)$ stands for the expectation over Y given X_i, which can be calculated by taking the mean over Y for all X_i. The first order Sobol index can then be calculated as a measure of sensitivity S of feature i on model output Y as

$$S_i = \frac{V_i}{V(Y)}$$

The first order Sobol index measures the effect of alternating X_i alone averaged over variations of other input features. By dividing it over the total variance it is measured as a fractional contribution. Higher order indices can be calculated by dividing V_{ij}, V_{ijk} and so on over $V(Y)$. The first- and higher-order Sobol indices quantify the importance of each feature, or combination of features with respect to the output variance. Evaluating all indices for a large number of features can be problematic as the number of evaluations is quadratic with the number of input features. In practice the total-order Sobol index is often calculated to overcome this problem. The total-order Sobol index is a measure of the contribution to the output variance of the i-th feature including all variance caused by the interactions with other features. It can be calculated as follows

$$S_{Ti} = \frac{E_{\mathbf{X}\sim i}(V_{X_i}(Y \mid \mathbf{X}_{\sim i}))}{V(Y)}$$

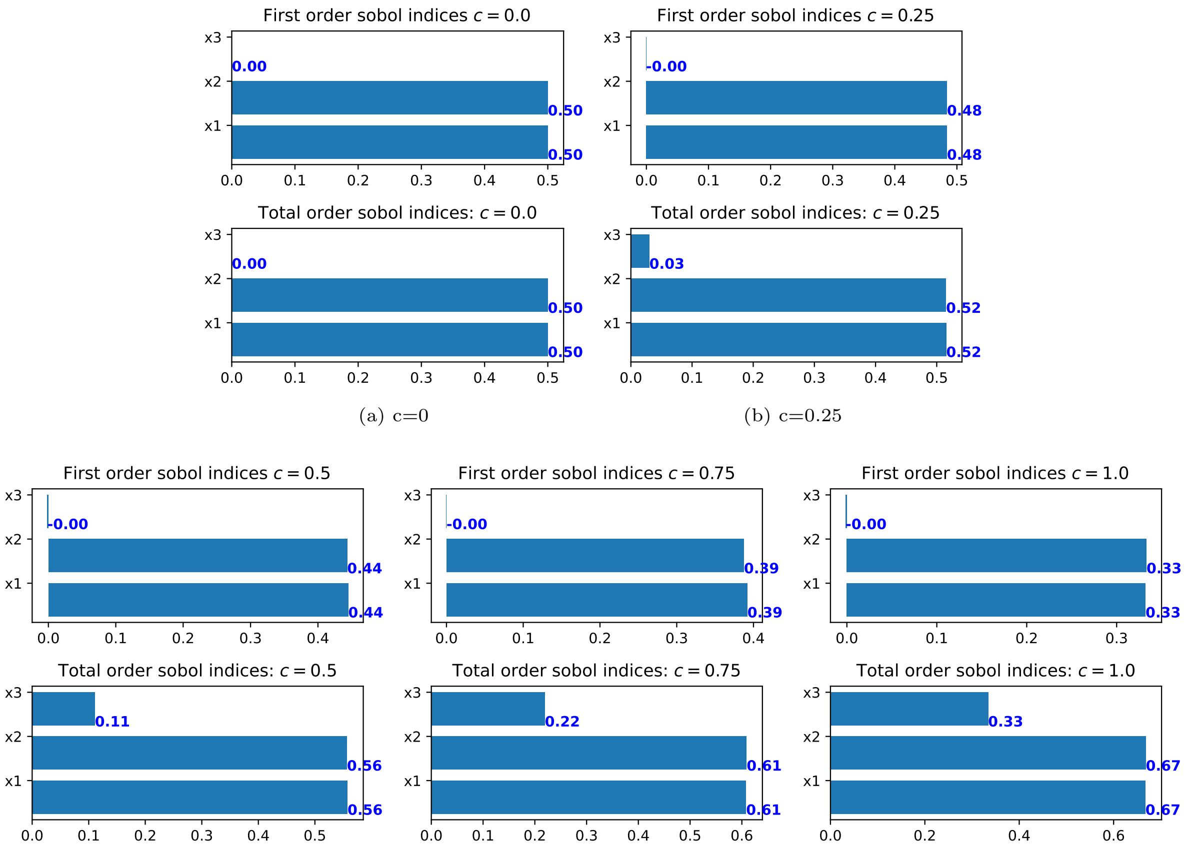
Here $E_{\mathbf{X}\sim i}(V_{X_i}(Y \mid \mathbf{X}_{\sim i}))$ stands for the expected variance in model output Y when all but the i-th feature are fixed.

Test case

Figure 5 shows the first- and total order Sobol indices for a simple test function:

$$f(x_1, x_2, x_3) = x_1 + x_2 + c \cdot rand(0,1) \cdot x_3 \tag{1}$$

The function is evaluated for values between 0 and 1 for x_1, x_2, x_3. c is a parameter that determines how much noise is added to the function and is valued $\{0, 0.25, 0.5, 0.75, 1\}$. x_1 and x_2 intuitively have the same influence on the output in the case that there is no noise present. The visualisation shows that the first order index for x_3 is 0.0 even when the amount of noise for this variable is increased. Furthermore, inspection of the visualisation shows that a higher amount of noise results in a larger value for the total order index of x_3. In earlier work by Kenneth et al. [37] it is shown that the Sobol indices can be well approximated and the correct order of variable importance is maintained even when the analyzed models contain significant error.



(a) c=0

(b) c=0.25

(c) c=0.5

(d) c=0.75

(e) c=1

Fig. 5: Sobol indices for test equation 1 for various values of c.

2.5.3 Morris method

The Morris method [38], also known as the One-Step-At-A-Time (OAT) method aims to identify non-influential features in the input set. The Morris methods classifies each input feature into one of three categories. Input with negligible effect, input with large linear effect without interaction, and input having large non-linear and/or interaction effects. The method performs r OAT experiments over an input space that is discretized into a d-dimensional grid with n input levels. The elementary effect $E_j^{(i)}$ can then be calculated at the i-th repetition for the j-th variable as:

$$E_j^{(i)} = \frac{f(\mathbf{X}^{(i)} + \Delta e_j) - f(\mathbf{X}^{(i)})}{\Delta}$$

Here Δ is a multiple of $\frac{1}{n-1}$ and e_j a vector of the canonical base. From this, the mean of the absolute value of elementary effects μ_j^* and the standard deviation of the elementary effects σ_j can be calculated as

$$\mu_j^* = \frac{1}{r} \sum_{i=1}^{r} \mid E_j^{(i)} \mid$$

$$\sigma_j = \sqrt{\frac{1}{r}\sum_{i=1}^{r}(E_j^{(i)} - \sum_{i=1}^{r} E_j^{(i)})^2}$$

The larger the value of μ_j^* for a feature j, the more the input value of this feature contributes to the divergence in output. σ_j gives information about the type of interaction effects that a feature has. A low value of σ_j indicates that the elementary effect has a low variation due to the input, suggesting that the effect of a perturbation is the same for all inputs and thus a linear relation between the input and output. If there is a high value of σ_j, elementary effects have large variations due to the input, suggesting that a variable with high σ_j has a non-linear effect on the output.

2.5.4 Derivative Based Global Sensitivity Measures

The previously mentioned variance based methods often require a large number of model evaluations in order to sample the model outputs. As models become more complex, the evaluation time or cost of a model can also increase. A faster method that relies less on model evaluations is called Derivative Based Global Sensitivity Measures (DGSM) [39]. The method focuses on averaging local derivatives by using Monte Carlo sampling methods. The DGSM method calculates derivative based importance measures $v_1, ..., v_n$

$$v_i = \int_{H^n} \left(\frac{\partial f}{\partial x_i}\right)^2 dx$$

A smaller value for v_i indicates a less influential features x_i. The importance measure is similar to the Morris importance measure μ^*.

2.5.5 Density based Sensitivity measures

The DELTA(δ)-index [40] is a global model free sensitivity measure, and is based on the Probability Density Function (PDF) of the output of the model. The method assesses the influence of the complete input distribution on the complete output distribution without referencing to a specific moment of the output. PAWN [41] is a sensitivity index that relies on the Cumulative Distribution Function (CDF). The PAWN index for a given parameter is calculated by taking the mean, median an maximum over all Kolmogorov–Smirnov (KS) test values. The KS test values can be calculated by taking the distance between the unconditional output distribution and the conditional output distribution. The unconditional output distribution can be calculated by moving all parameters simultaneous while the conditional distribution can be calculated by fixing the input parameter of interest to a certain value.

2.5.6 Partial Dependence Plots

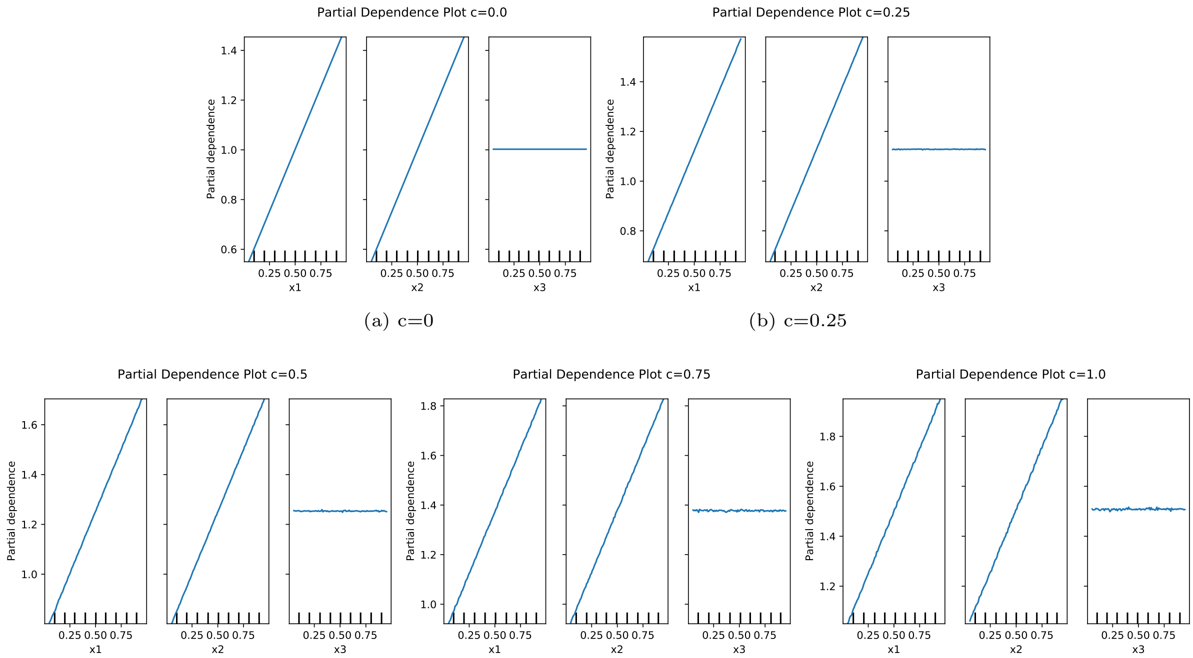
Once the importance of each input features has been determined it is possible to visualise the type of relation between the input feature and the output target. Partial Dependency Plots (PDP) [42] is an XAI method that can show whether the target variable and the input features have linear, monotonic, or more complex relationships. PDP's show the outcome of a model at a value of a feature when this value is substituted for all samples in the training set. PDP's divide the set of input features into to subsets: C and S. S is the set of features that are investigated, and C is the set of the remaining features. PDP's marginalize the model output over the distribution of the features in set C, so that the plot shows the relation between the model output and the features in S. In practice the PD function is estimated by calculating the average output in the training data

$$\hat{f}_s(x_s) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_s, x_C^{(i)})$$

The output of this function expresses the average marginal effect of the features in S on the prediction of the model. $x_C^{(i)}$ are the feature values from the training set from the features that are not investigated.

Test case

An example PDP for test equation 1 can be seen in Figure 6. The visualisation shows the PDP's for various values of c. Even for relatively high values of c the noise has very little effect on the plots. Experiments in Section 5.6 show how the r^2 value of a model influences the interpretability of PDP's.



(a) c=0 (b) c=0.25

(c) c=0.5 (d) c=0.75 (e) c=1

Fig. 6: Partial dependence plots for test equation 1 with $c = \{0, 0.25, 0.5, 0.75, 1\}$.

When making the PDP's the assumption is made that the features in S are independent of the features in C. The downside of this method is that when this assumption is violated it will cause the PDP to include data points with very unlikely or even impossible values. Examples of this are ships with a very large length and very narrow width, or very little engine power. These examples do not occur in real life, but are evaluated for making the PDP. A disadvantage of PDP's is thus that the assumption is made that the features are independent. The advantage of this methods is that it provides the user with a visual interpretation of the importance and effect of a feature.

2.5.7 Accumulated Local Effect plots

Accumulated Local Effect (ALE) plots [43] are a faster alternative to PDP's and in addition overcome the problem that PDP's assume feature independence. Just as with PDP's, ALE plots reduce the complex model function to a function that is dependent on only one or two features. The ALE plot, in contradiction to PDP's, show the change in prediction of a model over only a small window z of a value of a feature. The difference in prediction can be seen as the effect of a feature on a single instance in a certain interval. The uncentered ALE function is defined as:

$$\hat{\tilde{f}}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i:x_j^{(i)} \in N_j(k)} [\hat{f}(z_{k,j}, x_{\setminus j}^{(i)}) - \hat{f}(z_{k-1,j}, x_{\setminus j}^{(i)})]$$

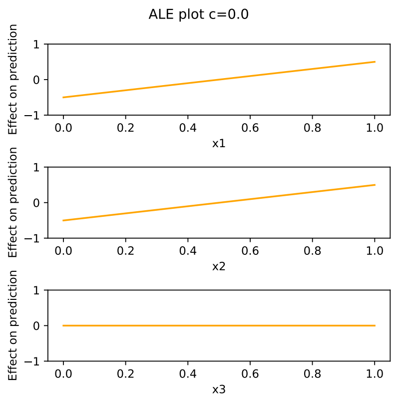
Here, the sum on the right side of the equation adds the effect of all samples within interval or neighborhood $N_j(k)$. This sum is then divided by the number of samples in the interval in order to obtain the average difference of predictions in this interval. The left summation then accumulates all the averages over all intervals. The following equation then centers the effect such that the mean effect is 0.

$$\hat{f}_{j,ALE,centered}(x) = \hat{\tilde{f}}_{j,ALE}(x) - \frac{1}{n}\sum_{i=1}^{n}\hat{\tilde{f}}_{j,ALE}(x_j^{(i)})$$

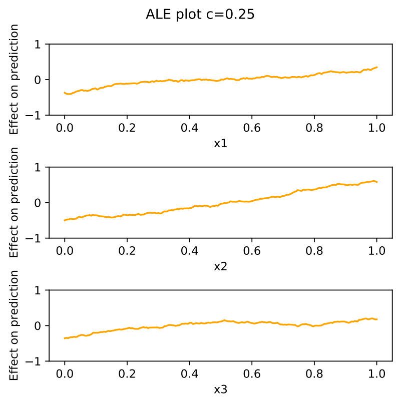
The value of this function can be interpreted as the main effect of a feature at a certain value compared to the average prediction of the data. When the ALE function has, for instance, an estimate of -4 at $x_j = 2$ it indicates that when the j-th feature has as value 2 the prediction is lower by 4 compared to the average prediction. The advantage of ALE plots is that they are not biased in a situation where features are correlated.

Test case

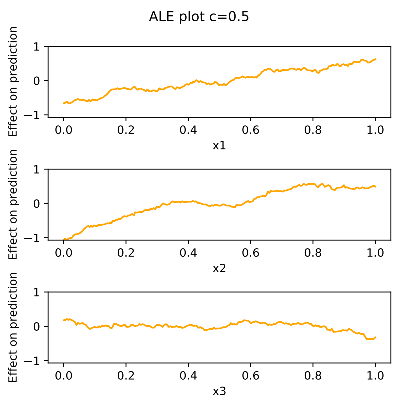
An example ALE plot for test equation 1 can be seen in Figure 7. The visualisation shows that as the noise level increases, the ALE plots become less stable. Experiments in Section 5.7 show how the r^2 value of a model influences the interpretability of the ALE plot.



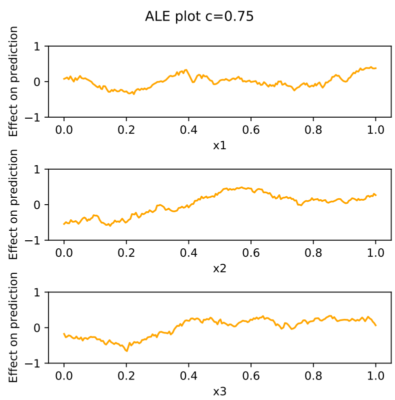
(a) c=0



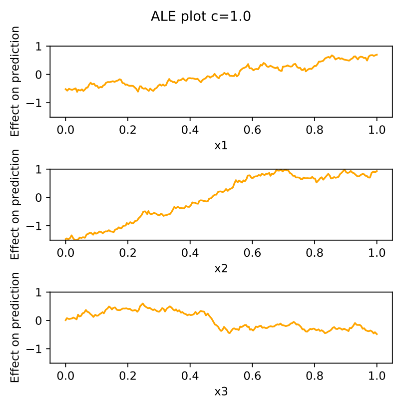
(b) c=0.25



(c) c=0.5



(d) c=0.75



(e) c=1

Fig. 7: Ale plots for test equation 1 with $c = \{0, 0.25, 0.5, 0.75, 1\}$.

3 Problem Statement

The ship design profession has existed for hundreds of years. In the past, most early phase design decisions were based on a handful of reference vessels, or made by experienced engineers that relied on the knowledge they had built up over their career. Examples of early phase design decisions are decisions about a ships measurements or engine power. The problem with this approach is that a wrong estimation by an engineer, or wrong sample of reference vessels can lead to flaws in the design, such as too many or little engine power for the length or width of the ship. These flaws can lead to higher building, or operating costs.

As big data has proven it's impact in many different industries over the recent years, it is also starting to gain ground in the shipping industry [44]. Previous big data projects in the shipping industry have shown promising results in vessel route planning and ETA prediction [45], ship chartering analysis [2], but also for static ship data based ship design optimization[46]. Most ship design optimization big data studies focus on static data, or data collected from sensors of either one or a handful of vessels. The results and conclusions from these studies thus only apply to a limited number of ships. To the best of the author's knowledge there is currently no study that uses the performance of all vessels in the commercial fleet to search for distinctive ship design characteristics.

By using the AIS it has become possible to gather data about the performance of all vessels in the commercial fleet. In order to learn from this vast amount of data, an efficient way of collecting, storing, and analyzing all incoming data from the AIS is needed. More on the data collection and processing will be explained in Section 4. Once the data is collected it can be analyzed by building a regression model that takes as input a vector $\mathbf{X}$ that contains features $\{x_1, x_2, ..., x_m\} \in \mathbf{X} \subseteq \mathbb{R}^m$. The output of this regression model is also a vector of a size than can range from 1 up to n; the number of target features that need to be predicted. As the regression algorithms produce a black box model, further analysis is needed in order understand how the input features affect the output of the model. The Sobol indices, explained in Section 2.5.2, can be used to determine which input features have the largest influence on the output. How these features influence the output can then be visualised by creating PDP's and ALE plots, as explained in Section 2.5.6, and 2.5.7. By understanding how the model produces the output values it is possible to gain insight in which input features, and thus which ships design parameters, are important for determining the model output, and thus the performance of a ship.

4 Data

The data used for this research can be divided into two main branches; AIS data, and static ship data. This section gives an in depth overview of how both data sources are collected, stored and processed.

4.1 AIS data

The AIS is a transceiver based automatic tracking system that aims to improve the safety of maritime traffic. The goal of the system is to provide insights in the whereabouts of vessels so that collisions can be avoided. The system does this by constantly transmitting a vessel's identity, location, sailing speed, and course along with additional information about the destination and identity of the vessel. Most vessels are equipped with a Very High Frequency (VHF) AIS transceiver that allows local AIS data to be received and plotted on a chart plotter. Simultaneously, the transceiver sends out AIS data to other nearby receivers. The range of VHF receivers is approximately between 10-20 nautical miles. Due to the relatively short range of VHF receivers it is not possible to capture all worldwide AIS data with just one receiver. In order to overcome this issue, many third party applications have been developed that collect the AIS data from multiple local transponders and combine them in order to come up with a wider coverage. The third party application used in this research is AISHub [47] as they provide an easy to use API that allows the raw AIS strings to be collected. Normally, an API key can only be obtained by contributing to the AISHub network of AIS transceivers, meaning that you have to install an own AIS transponder and connect it to the AISHub network. As we managed to obtain an API key via the Transferring Operational Data into Design Information for Ships (TODDIS) project [48] we did not install our own AIS transceiver.

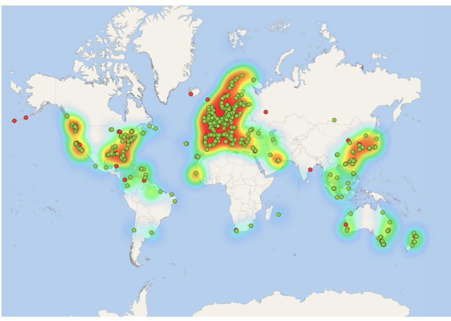


Fig. 8: AISHub coverage

AIShub aims to connect as many local AIS transceivers in order to obtain the best possible coverage. All connected stations are on land, or within coastal areas, meaning that there is no coverage on open ocean. The total coverage of AIShub can be seen in Figure 8. The map shows a strong coverage in Europe, East coast North-America, and South-East Asia.

4.1.1 Data collection

AIShub continuously logs the incoming AIS signals, and allows for one request to collect the most recent data per API key per minute. For this research a Google Cloud environment is setup that performs the API request, receives the data, and subsequently stores the data in a database. A cloud function performs a call to the AIShub API once every minute. The API then responds with a .csv file containing approximately 34,000 rows, each row corresponding to a vessel with an AIS transmitter, and 22 columns containing the features corresponding to each vessel. Table 1 describes the features that are logged. Once a .csv file is received from the API the cloud function appends it's contents to a Google Bigquery database where all collected AIS data is stored.

Table 1: Definition of features in raw AIS data

Feature name	Explanation	Range
MMSI	Maritime Mobile Service Identity	N/A
TSTAMP	Timestamp in UTC date/time format	N/A
Longitude	Geographical longitude in degrees	[-180,180]
Latitude	Geographical latitude in degrees	[-90,90]
COG	Course over ground	[0,360]
SOG	Speed over ground	>0
Heading	Current heading of vessel at time of last message	[0,360]
Navstat	Navigational status, indicates what operation the ship is performing	[0,15]
IMO	IMO ship identification number	N/A
Name	Ship name	N/A
Type	Vessel Type	[1,99]
Callsign	Vessel callsign	N/A
A	Distance from transceiver to bow in meters	N/A
B	Distance from transceiver to stern in meters	N/A
C	Distance from transceiver to port in meters	N/A
D	Distance from transceiver to starboard in meters	N/A
DRAUGHT	Draught of vessel in meters	>0
DEST	Vessel destination	N/A
ETA	Estimated time of arrival in UTC date/time format	N/A

The data collection process has started in september 2021 and is still going on at the moment of writing. At the moment of writing the Google Bigquery database contains approximately 7.5 billion rows with a total size of 1.2 TB. A schematic overview of the AIS data collection pipeline is given in Figure 9.

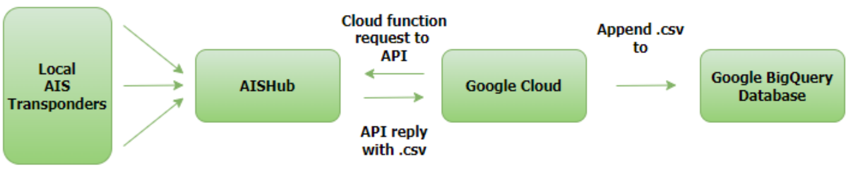


Fig. 9: Schematic overview of the AIS data collection pipeline

4.1.2 Data preprocessing

From this large set of raw AIS data, information can be extracted about a vessels operational performance capabilities. The columns of interest are: TSTAMP, Latitude, Longitude, COG, SOG, Heading, Navstat, IMO, A,B,C,D and Draught. The definition of these columns can be seen in Table 1. Furthermore, a subset of the data is selected so that only ocean going vessels are considered. This is done by only selecting vessels with a length of more than 80 meters, as recommended by a domain expert.

Next, the samples with feature values that lie outside the allowed ranges shown in Table 1 are filtered out. Examples of this are samples that have a value of 511 for heading, which means that the heading is unknown. Any further samples that have unrealistic feature values, such as an SOG >40 knots, are also removed. Finally, the navigational status of a vessel is checked for each record, and removed if the navigational status is 1 (at anchor) or 5 (moored). The resulting subset, containing data from October up to February, is then downloaded from the Google Bigquery database and saved as a Hierarchical Data Format (HDF) file, a compressed file format designed for handling large datasets. The data from September is not considered as the data collection process was not continuously active in this period.

Even though the dataset spans a five month time window, many vessels have a relatively limited number of records in the dataset as shown in Figure 10. In order to get a good grasp of the maximum operational performance capabilities of a vessel, a minimum number of data points per vessel is needed. Furthermore, these data points need to be from a variety of days. Simply looking at the data from one day might give a biased view of a vessels performance as the conditions on that day might be especially dis- or advantageous. By trial and error it is determined that vessels in the dataset that have transmitted less than 300 data points, or on less than 4 unique days will be removed from the dataset. As only data points that have an AIS status different from 1, or 5 are selected from the Bigquery database, it is guaranteed that these data points are not from ships laying still in harbor.

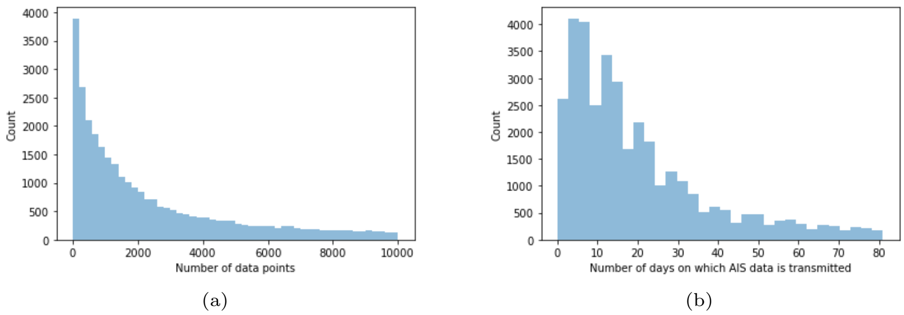


Fig. 10: For many vessels the number of logged data points is relatively low. Similarly, for many vessels the logged data points have been transmitted over a small number of days.

4.1.3 Feature extraction

The resulting dataset is 9 GB, has 197 644 394 rows and 12 columns describing the behaviour of 35 645 unique vessels over a period of 5 months. The next step is to derive features from the AIS data that give information about the operational performance capabilities of a ship. This is done by calculating the following features for each vessel.

Max speed
The maximum SOG that a vessel has sailed. This feature is of interest to a naval architect because it shows how fast a ship can potentially sail. As SOG is part of the AIS data string, it can be simply extracted directly for each ship in the dataset.

Max rotation
The maximum rotation in degrees that a vessel has made in a time window of one minute. The feature is of interest for a naval architect because it is a quantification of the maneuverability of a ship. Rotation is not part of the raw AIS data, and thus needs to be derived. The rotation per minute can be calculated from AIS data by taking the difference in heading between two consecutive AIS records, dividing the difference in heading by the difference in time in seconds and then multiplying this by 60. An example edge case exists when the heading of record 1 is 5, and the heading of record 2 is 355. The difference between these two records is then 350, while the ship has probably only rotated by 10 degrees. If the difference between the two records is larger than 180, the difference in heading is subtracted from 360 to account for the edge cases where the heading of two records are

both close to 360 and 0. The formula used for calculating the rotation between two records is:

$$\Delta_t = TSTAMP_2 - TSTAMP_1$$
$$\Delta_h = \begin{cases} 360 - abs(h_1 - h_2), & \text{if } abs(h_1 - h_2) > 180 \\ abs(h_1 - h_2), & \text{otherwise} \end{cases}$$
$$rotation = \frac{\Delta_h}{\Delta_t} \times 60$$

Here Δ_t stands for the difference in time, Δ_h for the difference in heading, h_1 for the heading of record 1 and h_2 for the heading of record 2. We have empirically found that using consecutive AIS records that are no further than two minutes apart provide us with the best results.

Max acceleration

The maximum acceleration that a vessel exhibits between two consecutive AIS records. The feature is of interest to a naval architect because acceleration plays a role in the speed and maneuverability of a ship. Acceleration is not part of the raw AIS data but can be derived by taking the difference in SOG between two consecutive AIS records, dividing this difference by the difference in time between the records, and multiplying this by 60. Here we also take a maximum of two minutes for two records to be consecutive. In formula:

$$\Delta_{SOG} = SOG_1 - SOG_2$$
$$\Delta_t = TSTAMP_2 - TSTAMP_1$$
$$Acceleration = \frac{\Delta_{SOG}}{\Delta_t}$$

For acceleration, we only look at postive results of this formula. The negative values are used for deceleration.

Max deceleration

The maximum deceleration that a vessel exhibits between two consecutive AIS records. As deceleration influences the maneuverability of a ship, it is an important measure for a naval architect. Deceleration can be derived by the same formula as for acceleration, but this time only looking at negative values. The same time window as for rotation and acceleration is used here.

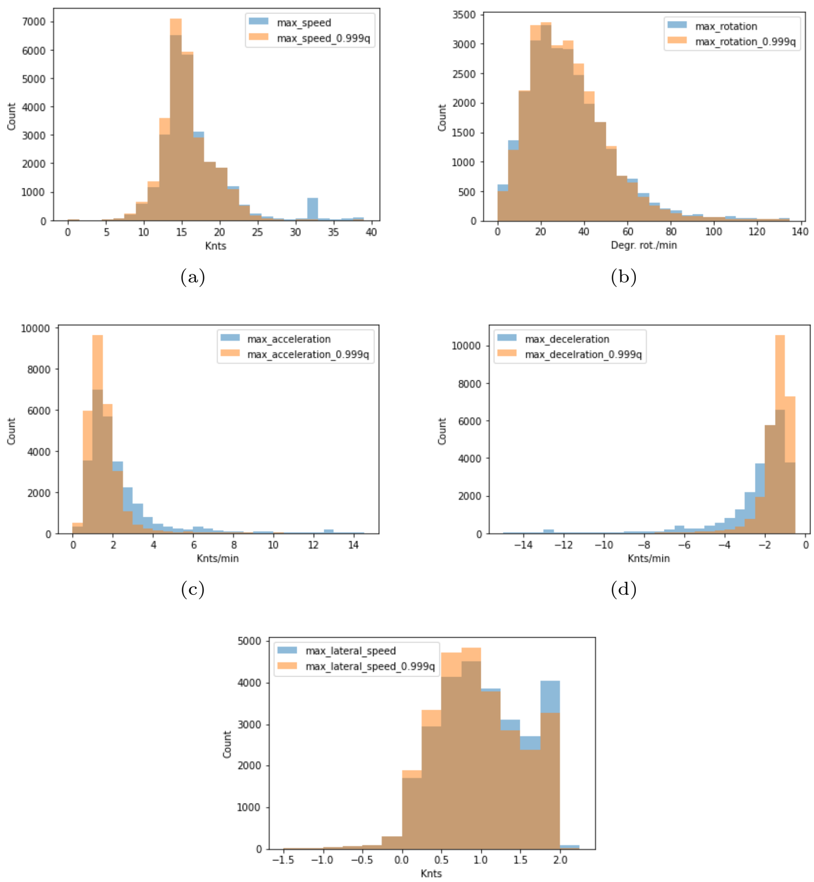
Max lateral speed

The maximum lateral speed that a vessel exhibits. Lateral speed is also a quantification of the maneuverability of a ship, and can thus be of interest to a naval architect. The lateral speed is calculated as follows:

$$SOG_{lat} = \sin(rotation) \cdot SOG$$

Next to these target features, some additional input features are calculated from the AIS data. For each target feature, the draught of the vessel at the moment that the vessel produced the target value is logged. Furthermore, for each vessel the number of tugboats in the proximity is calculated at the moment the maximum rotation value is logged. This value is calculated by using the latitude and longitude coordinates in the AIS data to check within a radius of 600 meters, in a time interval of 1 minute before the max rotation is logged, and 1 minute after the max rotation is logged if there are vessels present that transmit AIS data with ship type 52, as this is the vessel code for tugboats. The time interval and radius of the proximity are set after consulting a human expert, combined with some trial and error.

The calculation of aforementioned target features can be influenced by missing values, or noise which is naturally present in AIS data [49]. AIS is not a strongly typed language, so for the columns DEST, DRAUGHT, ETA, and NAVSTAT, users are free to fill in strings where real valued numbers are expected, or leave certain fields unfilled. In order to account for outliers that occur as a result of this while calculating the maximum value for each ship, the 0.999 quantile value for each feature is also calculated. Figure 11 shows the distributions of the target features for both the maximum value, and the 0.999 quantile value. As expected, most of the 0.999 quantile values seem to have a slightly lower value than the absolute maximum values. For acceleration and deceleration, the 0.999 quantile values have many more occurrences in the lower valued bins than the absolute maximum value. This suggests that using the absolute maximum value is more susceptible to outliers. In Section 5 we determine which feature yields the best results.



(e)

Fig. 11: Distributions of features calculated from AIS data. Taking the absolute maximum value seems more susceptible to outliers than taking the 0.999 quantile value.

4.1.4 Outlier removal

In order to take account for the remaining outliers in the target the Inter Quartile Range (IQR) method [50] is used. This method calculates an upper and lower bound for a feature. All feature values below the lower bound, or above the upper bound are selected as outliers. The bounds are calculated as follows

$$\begin{aligned} Q1 &= 1st\ \ quartile \\ Q3 &= 3rd\ \ quartile \\ IQR &= Q3 - Q1 \\ Upperbound &= Q3 + 1.5 \cdot IQR \\ Lowerbound &= Q1 - 1.5 \cdot IQR \end{aligned}$$

4.2 Static ship data

The second branch of data used in this research is static ship data. The dataset with static ship data comes from earlier work [46] where the data is provided by a reference database from C-Job Naval Architects. C-Job Naval Architects is a worldwide independent ship design and engineering company with its headquarters in the Netherlands. The vision of C-job is "A sustainable maritime industry within one generation." The company now employs over 180 in-house maritime engineers and naval architects in seven offices across the globe. C-Job has a passion for everything afloat. Whether it is new build or conversions, C-Job is renowned for ground-breaking new vessels in a broad range of maritime sectors.

The dataset contains 230 851 unique vessels, each described by 129 features. A description of all features in the dataset can be found in the table A1. Some features are specific for certain ship types, which means that not all features contain a value for each ship. After calculating the performance features from the AIS data as explained in Section 4.1, the instance is merged with additional information about the ship design from the reference database. Each instance in the resulting dataset now contains the features calculated from the AIS data, as well as the static ship information describing the ship's design. As each row now describes the performance characteristics and design of a unique vessel, we have a much smaller dataset containing 27 343 rows. This is less than the initial 34 000 vessels that we collect AIS data of because non-ocean going vessels, and vessels that have not transmitted sufficient data points on a variety of days are filtered out as explained in 4.1.

Figure 12 shows a histogram of the top-10 most frequently occurring ship types. The distribution is

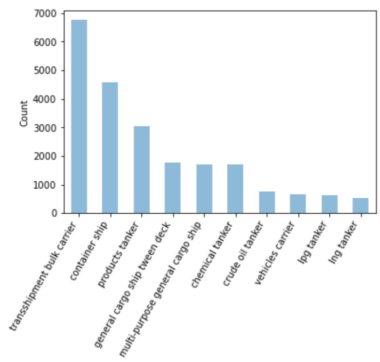
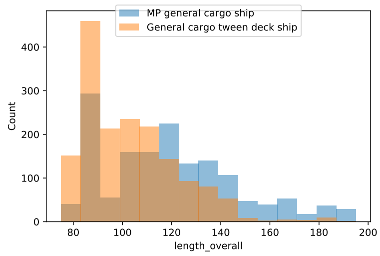
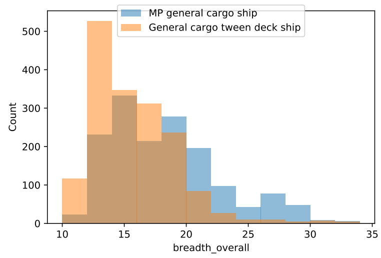


Fig. 12: Histogram of most frequent ship types in dataset.

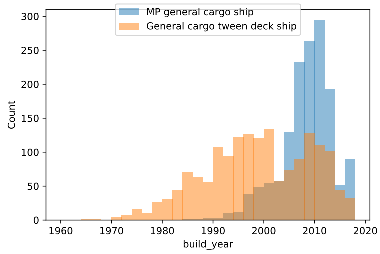
skewed to the left with Transshipment Bulk Carrier, Container ship, and Products Tanker clearly being the most frequently occurring ship types. Furthermore, two ship types seem very similar due to their name; General Cargo ship tween deck, and Multi-Purpose General Cargo ship. The distributions of the most distinctive ship design parameters of these ship types can be seen in Figure 13. The distributions show that Multi-Purpose General cargo ships tend to be larger and more recently built than General Cargo ship tween decks. It can also be noticed that there is a spike in ship occurrences of approximately 90 meters. Ships of this length correspond to ships with a gross tonnage of 3000 tonnes. Ships below this gross tonnage are required to only have one captain and one helmsman, while ships with a higher gross tonnage are required to have one captain with higher qualifications and two helmsmen, resulting in higher operational costs. In order to check if these two distributions are statistically significantly different we apply the Mann-Whitney U test [51], as this test does not assume any specific distribution of the data. The null hypothesis of this test states that two distributions are the same. A p-value < 0.05 rejects the null-hypothesis.



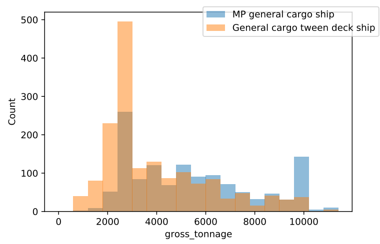
(a) Mann-Whitney U-test: $p = 0.0$. Null hypothesis rejected.



(b) Mann-Whitney U-test: $p = 0.0$. Null hypothesis rejected.



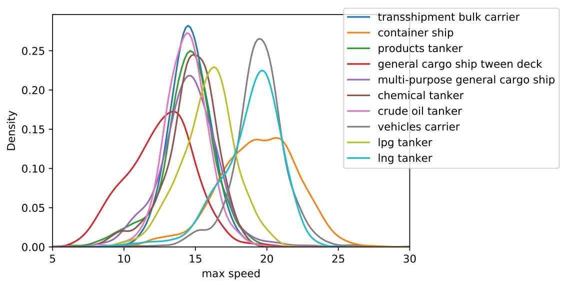
(c) Mann-Whitney U-test: $p = 0.0$. Null hypothesis rejected.



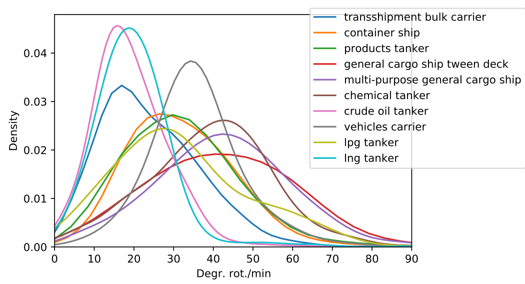
(d) Mann-Whitney U-test: $p = 0.0$. Null hypothesis rejected.

Fig. 13: Distributions of design parameters of General Cargo ship tween deck and Multi-Purpose General cargo ship showing that they are different ship types, even though their names are similar.

As each ship type is designed for its own unique purpose we expect there to be differences between the performance features of the different ship types. Figure 14 shows the distributions of the performance features calculated from AIS data for the 10 most frequently occurring ship types in our dataset. For max speed we see that container ships, Liquified Natural Gas (LNG) tankers, and vehicles carriers overall seem to have a higher max speed compared to the other ship types. For max rotation we see a less clear distinction, but crude oil tankers and LNG tankers seem to have an overall lower max rotation value.



(a)



(b)

Fig. 14: Kernel Density plots of performance features calculated from AIS data.

4.3 Imputing Missing values

The reference database with static ship data, discussed in Section 4.2, contains many columns with missing values. A missing value can occur in two cases; if the feature is not applicable to a ship, the feature "number_of_lorries" for instance, is only applicable for vehicle carrier ships and not for other ship types. In order to take account for this, for all ship types only the features that have a missing value percentage smaller than a predetermined threshold are selected. By manually inspecting the number of missing values for each ship type we have set this percentage to 30%. In almost all cases this percentage caused all features that are applicable to a ship to be selected, without selecting inapplicable features.

The second case where missing values can occur is when the value is simply unknown. If a feature value is unknown it can be derived from similar samples in the dataset. In order to impute the missing value the k-Nearest Neighbour (KNN) imputing method [52] can be applied. This method assigns each sample to k neighbours based on the Euclidian distance between the feature values and then fills in the missing value as the mean feature value of the k neighbours. For categorical features, all feature values should first be transformed to numerical values.

5 Experiments & Results

This section contains the experiments and results of this research. Linear regression and SVR are trained as baseline algorithms, and used as comparison for the effect of various feature engineering steps. Additional regression algorithms and a multi-output AutoML pipeline are introduced in order to further improve the model performance. The AutoML pipeline is then further analyzed with SA and XAI techniques.

5.1 Baselines

The first experiment contains baseline models for each unique ship type and target feature. The algorithms used for the baseline models are Linear regression and SVR as described in Section 2.2.1 and 2.2.2. Before the models are trained, the target feature outliers are removed by using the IQR method as described in Section 4.1. Furthermore, for each ship type only the input features from the reference database that contain less than 30% missing values are used. This value is selected after measuring the decrease in r^2 scores when removing an x percentage of feature values and filling them with the KNN imputing method. The results of this preliminary experiment can be seen in Table A2. For each feature that contains any remaining missing values the value is filled by using the KNN imputing method described in Section 4.3. The models are trained with the default hyperparameters and their average r^2 scores over 5-fold cross validation are reported. The results can be found in Table 2.

Table 2: Average r^2 scores over 5 cross validation folds for the 10 most frequently occurring ship types in the dataset. The two baseline algorithms used are Linear Regression and Support Vector Regression.

		Max lat. speed	Max acc.	Max decel.	Max rot.	Max speed
Transshipment bulk carrier	Lin reg	0.02	0.13	0.09	0.14	0.11
	SVR	0.02	0.18	0.07	0.13	0.13
Container ship	Lin reg	0.06	0.33	0.25	0.48	0.38
	SVR	0.04	0.38	0.25	0.51	0.42
Products tanker	Lin reg	0.02	0.22	0.20	0.25	0.15
	SVR	0.00	0.23	0.24	0.24	0.21
General cargo ship tween deck	Lin reg	0.03	0.18	0.14	0.19	0.46
	SVR	0.08	0.17	0.09	0.22	0.48
Multi Purpose general cargo ship	Lin reg	0.00	0.25	0.15	0.26	0.30
	SVR	0.09	0.25	0.15	0.27	0.37
Chemical tanker	Lin reg	0.02	0.25	0.20	0.14	0.16
	SVR	0.02	0.26	0.20	0.12	0.16
Crude oil tanker	Lin reg	0.01	0.00	0.03	0.12	0.00
	SVR	0.03	0.00	0.03	0.14	0.00
Ro-ro cargo	Lin reg	-0.03	0.21	0.19	0.15	0.57
	SVR	-0.05	0.22	0.19	0.04	0.58
LPG tanker	Lin reg	0.02	0.26	0.31	0.29	0.22
	SVR	0.00	0.25	0.30	0.26	0.23
LNG tanker	Lin reg	0.00	0.02	0.02	0.01	0.11
	SVR	-0.1	0.08	0.08	0.01	0.08

The results show that for many ship types and many features the models fail to learn any relationship in the data. However, for Container ship, General Cargo (GC) ship tween deck, Multi Purpose (MP) general cargo ship, and Ro-ro cargo ship with target features maximum rotation and maximum speed the r^2 scores show that the models have learned some relation between the input data and target feature. As only for these four ship types and two target features the results are promising, the experiments in the following sections are focused on this subset of ship types and target features.

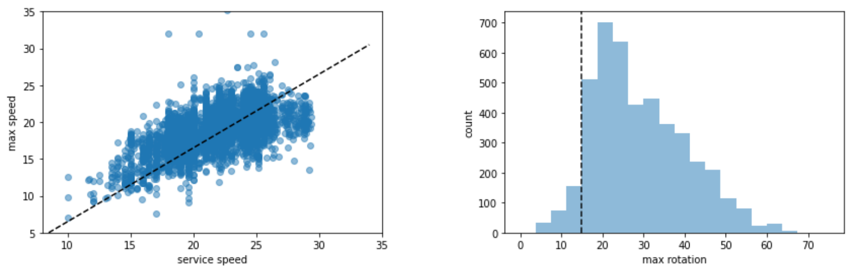
5.2 Feature engineering

The performance of regression models can be increased by performing various feature engineering steps. The experiments in this subsection show how further outlier removal, and adjusting the definition of target features can improve model performance.

5.2.1 Improved outlier removal

The IQR method explained in Section 4.1 is a statistical method for detecting outliers in the target features. Next to this statistical way of removing outliers, domain specific knowledge can be applied to remove any remaining outliers or noise in the data that might decrease the performance of the regression models. The maximum speed that is measured, is calculated by taking the absolute maximum value of all SOG values of a ship. For some ships, however, all collected data is collected at moments where the ship is not sailing at full power. This is often the case for Container ships, as they are known for sailing on a scheduled service. For these ships, there is no point in sailing at full power unless they are behind on their schedule. When they are sailing on schedule they do not sail at full power in order to save fuel. For ships that have only sailed on schedule, the maximum recorded speed does not resemble the true maximum speed of the ship. In order to detect and remove samples of ships that have not sailed at full power the maximum recorded SOG is compared to the service speed. For each ship, the reference database provides a service speed which is the estimated average speed that a ship sails under regular load and weather circumstances. Together with a human expert we determine that ships with a maximum speed recorded in the AIS data lower than 3.5 knots below their service speed certainly have not sailed at maximum power. As a result, all ships that have a maximum recorded speed lower than 3.5 knots below the service speed are removed from the dataset. The distribution of the recorded maximum speed versus the corresponding service speed of container ships and the cut off that is made can be seen in Figure 15a.

For the feature max rotation, outlier samples can be removed in a similar way. For some vessels only data points that are transmitted during a straight trajectory are recorded. The recorded max rotation over these trajectories is not representative of a ships rotational capabilities simply because the ship barely had to rotate. For this reason, samples that have a recorded max rotation of 15 degrees per minute or lower are removed. This threshold has been determined, just as for max speed, in consultation with a human expert. The distribution of max rotation and the cut off point can be seen in Figure 15b.



(a) (b)

Fig. 15: Plots showing how outliers are removed for max speed and max rotation for ship type Container ship. For a, samples below the dotted line are removed. For b, samples left of the dotted line are removed. Similar plots for GC ship tween deck, MP general cargo ship, and Ro-ro cargo can be found in Figure A1, A2, and A3.

Table 3 shows the results of retraining the baseline models on data after applying aforementioned outlier removal steps. The r^2 scores increase for almost all ship types for both features, and especially for Container ships.

Table 3: Baseline model average r^2 score over 5-fold cross validation pre and post feature engineering. The highest results for each ship type, target feature, and algorithm are marked in bold.

		Max rotation		Max speed	
		Pre FE	Post FE	Pre FE	Post FE
Container ship	**Lin reg.**	0.48	**0.49**	0.38	**0.48**
	SVR	0.51	**0.52**	0.42	**0.52**
General cargo ship tween deck	**Lin reg.**	0.19	**0.20**	0.46	**0.47**
	SVR	**0.22**	0.21	0.48	0.48
MP general cargo ship	**Lin reg.**	0.26	**0.27**	0.30	**0.31**
	SVR	0.27	**0.29**	0.36	0.36
Ro-ro cargo ship	**Lin reg.**	0.12	**0.15**	**0.57**	0.54
	SVR	**0.15**	0.08	**0.58**	0.56

5.2.2 Additional target features

The outlier removal step has as disadvantage that almost 50% of the samples of Container ships are discarded due to the max speed values that have been gathered from ships that are not sailing at full power. Removing such a large subset from the dataset means that the model misses out on information from 50% of the samples, which might prevent the model from generalizing. In order to remove less ships from the dataset but still use a target feature that gives information about a ships speed, we define a new target feature called median cruising speed. This feature gives information about the median speed of a ship, and is thus less dependant on whether the collected data points come from ships that are sailing at full power. The median cruising speed is calculated by taking the median of all speed values that lie within the range

$$x \cdot SOG_{max} < SOG < SOG_{max}$$
$$x \in \{0.6, 0.7, 0.8, 0.9\}$$

This range is used in order to disregard data points that are recorded while ships are sailing in ports, or other waters where their movements are restricted.

Table 4 shows the results of retraining the baseline models for the new target features. For each ship type, a different value for x results in the highest r^2 score. Furthermore, the r^2 scores for MP general cargo ship and GC ship tween deck are higher than when max speed is used as target feature. For Container ships the results show that the r^2 score has dropped in comparison to max speed, but only slightly. Moreover, the table shows that using the 0.999 quantile value of SOG, results in a higher r^2 score than when using the absolute maximum SOG value. Because of this, the 0.999 quantile SOG value for max speed is used in all remaining experiments.

In the case of max speed and max speed 0.999q only 2467 container ships are considered due to outlier removal, whereas for median cruise speed 4569 container ships are considered. The median cruise speed thus gives a broader view of the speed capacities of container ships.

A similar experiment has been performed where the mode over all data points in an interval is taken instead of the median. As the resulting r^2 scores of the models built for these features are too low to provide useful insights, they are presented in the appendix in table A3. The low r^2 scores for target features where the mode is taken instead of the median might be explained due to the fact that some vessels might have multiple very often occurring, but very different values for SOG. When taking the mode only the most frequent value is considered while all other frequently occurring values are discarded, thus providing a limited view.

Table 4: Results of different target features that quantify the speed capabilities of multiple ship types.

		median cruise speed x 0.6	0.7	0.8	0.9	max speed 0.999q	max speed
Container ship	**Lin reg.**	0.33	0.46	0.51	0.45	0.56	0.38
	SVR	0.36	0.49	0.54	0.50	0.59	0.42
General cargo ship tween deck	**Lin reg.**	0.59	0.59	0.54	0.52	0.48	0.47
	SVR	0.60	0.59	0.57	0.57	0.49	0.48
MP general cargo ship	**Lin reg.**	0.51	0.48	0.40	0.32	0.35	0.31
	SVR	0.59	0.57	0.48	0.44	0.37	0.36
Ro-Ro cargo	**Lin reg.**	0.58	0.62	0.67	0.62	0.57	0.57
	SVR	0.58	0.59	0.64	0.65	0.59	0.58

By slightly changing the definition of the speed related target features the r^2 scores for regression models that predict the speed capacity for GC ship tween deck, MP general cargo ship, and Ro-Ro cargo ships have been improved. Analogous, the r^2 score can be improved for regression models that predict the rotational capabilities of ships. Currently the absolute maximum value is taken as maximum rotational speed. This value might be influenced by noise or outliers. The 0.999 quantile value might give a better insight in the rotational capabilities of the ship, as it is less influenced by outliers which is visualised in Figure 11. The baseline regression algorithms are trained again for both max rotation and max rotation 0.999q. The average r^2 scores over 5-fold cross validation are compared in Table 5. The table shows that for Container ships the models clearly perform better when we use the 0.999 quantile value for max rotation, but for General cargo ship tween deck, and Multi-purpose general cargo ship, taking the absolute maximum rotation value works better. For Ro-Ro cargo we see that for both target features the models perform very poor.

Table 5: Comparison of average r^2 scores over 5-fold cross validation for absolute max rotation and 0.999 quantile max rotation.

		max rotation	max rotation 0.999q
Container ship	**Lin reg**	0.49	0.58
	SVR	0.52	0.58
General cargo ship tween deck	**Lin reg**	0.31	0.19
	SVR	0.33	0.18
Multi-purpose general cargo ship	**Lin reg**	0.35	0.27
	SVR	0.33	0.27
Ro-Ro cargo	**Lin reg**	0.04	0.08
	SVR	0.07	0.00

5.3 Model improvements

The model performance can be further improved by using more advanced regression algorithms, and by optimizing the hyperparameters of these algorithms. Before tuning the hyperparameters, XGBoost regressor and a Random forest regressor are trained with default hyperparameters for each ship type and target feature. The same methods for filling missing values, and removing outliers as in Section 5.1 are used. The average r^2 score over 5-fold cross validation is reported in Table 6 which shows that RF regressor outperforms XGboost in all cases. When comparing the results to the baseline results in Table 4, and 5 it also shows that RF regressor outperforms all baseline models.

Table 6: Average r^2 score over 5-fold cross-validation of XGBoost regressor and RF regressor on all target features.

		median cruise speed						
		x						
		0.6	**0.7**	**0.8**	**0.9**	**max speed 0.999q**	**max rot.**	**max rot. 0.999q**
Container ship	**XGBoost reg.**	0.31	0.46	0.53	0.49	0.59	0.50	0.53
	RF reg.	0.39	0.52	0.57	0.54	0.64	0.55	0.59
GC ship tween deck	**XGBoost reg.**	0.58	0.59	0.55	0.54	0.49	0.28	0.04
	RF reg.	0.62	0.62	0.60	0.59	0.52	0.40	0.21
MP GC ship	**XGBoost reg.**	0.57	0.52	0.44	0.39	0.36	0.28	0.13
	RF reg.	0.59	0.55	0.48	0.42	0.41	0.39	0.29
Ro-Ro cargo	**XGBoost reg.**	0.56	0.64	0.66	0.61	0.66	0.14	0.14
	RF. reg	0.63	0.67	0.66	0.62	0.65	0.19	0.08

5.3.1 Hyperparameter optimization

As RF has outperformed all other tested algorithms, hyper parameter tuning is performed only to the RF models. In order to optimize the hyperparameters a grid search is performed over the grid shown in Table 7. The average r^2 scores over 5-fold cross-validation of the models with optimized hyperparameter combinations can be found in Table 8.

Table 7: Hyper parameter grid used for optimizing Random Forest model.

Param name	**Values**
max_depth	{8,10,12}
max_features	{3,4,5}
min_samples_leaf	{3,4,5}
min_samples_split	{3,5,7}
n_estimators	{100,200,300}

Table 8: Average r^2 scores over 5-fold cross-validation of Random forest regressor models with tuned hyper parameters.

	median cruise speed						
	x						
	0.6	**0.7**	**0.8**	**0.9**	**max speed 0.999q**	**max rot.**	**max rot. 0.999q**
Container ship	0.41	0.53	0.58	0.55	0.66	0.56	0.60
GC ship tween deck	0.64	0.63	0.62	0.61	0.56	0.41	0.23
MP GC ship	0.61	0.57	0.50	0.45	0.43	0.40	0.31
Ro-ro cargo	0.64	0.69	0.68	0.68	0.66	0.19	0.14

The results from optimizing the hyperparameters show that there is not a single target feature that obtains the highest r^2 score for all ship types. Instead, the results show that for each ship type, slightly different definitions of the target features are needed in order to create a regression model that can make accurate predictions about a ship's speed or rotational capabilities. The combination of these slightly different target features might provide the regression algorithms useful information that it can use in order to make better predictions.

5.4 Multi-output regression

A multi-output regressor is trained that predicts for one ship type all target features. Predicting all target features at once might exploit the extra information that is enclosed in the relationship between the target features. In order to prevent having to repeat all experiments performed in the previous subsections, an AutoML pipeline is used that takes care of selecting algorithms, fine tuning models, and finally building an ensemble of the individual models. The AutoML pipeline is implemented in *Auto-sklearn 2.0* which requires the setting of only two hyperparameters; *time_left_for_this_task* which is the time limit in seconds for finding appropriate models, and *per_run_time_limit* which is the time limit for a call to a single machine learning pipeline. The time limits are set as recommended in the documentation to respectively 7 hours, and 35 seconds. An overview of the ensemble of the resulting model for each ship type can be found in tables A4, A5, A6, and A7. The performance of the resulting models is evaluated by reporting the r^2 score on the test set in Table 9.

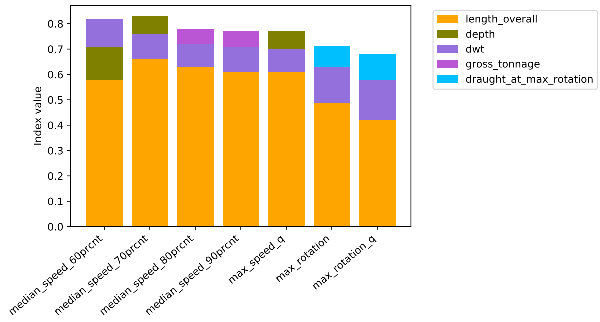
Table 9: r^2 score of AutoML models on test set.

	median cruising speed x						
	0.6	**0.7**	**0.8**	**0.9**	**max speed 0.999q**	**max rot.**	**max rot 0.999q**
Container ship	0.52	0.67	0.72	0.68	0.67	0.52	0.56
GC ship tween deck	0.59	0.60	0.56	0.54	0.56	0.32	0.31
MP GC ship	0.52	0.47	0.39	0.32	0.34	0.27	0.26
Ro-Ro cargo	0.82	0.81	0.83	0.81	0.82	0.46	0.51

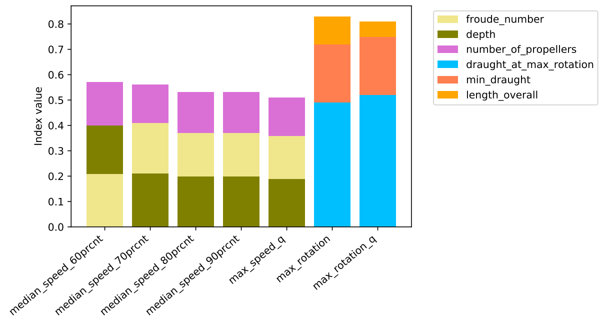
The table shows that for some target features the r^2 score has improved compared to the r^2 scores of the optimized random forest models from Table 8, especially for Ro-Ro cargo ships. For other features we see that the performance has dropped, which can be explained by the fact that the AutoML pipeline aims to produce a model that results in the highest overall r^2 score instead of focusing on individual r^2 scores.

5.5 Sobol indices

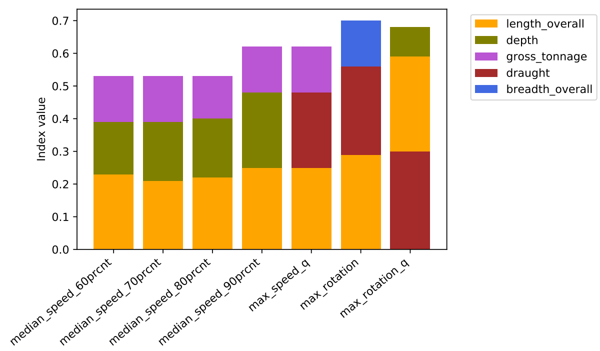
The models described in Section 5.4 work in a black-box fashion, it accepts input and produces output, but it is not clear to the user how the model makes decisions in order to produce the output. In order to gain insight in which ship design parameters have influence on the performance of a ship it is important to understand the working of the regression models. This subsection shows which input features explain the largest part of the variance of the model output, and thus which input features are most important for determining the value of the target features. The model that we choose to analyze is the multi-output regression model described in Section 5.4 as this allows us to see how multiple target features are influenced by the input features. For each ship type and target feature the first order, and total order Sobol index are calculated in order to gain insight in the effect of a feature with and without interaction with other features. As there exists very little difference between the resulting first order, and total order index, we report the total order index in Figure 16, and the first order index in Figure A5. The plots show that for all target features for ship type Container ship the input feature *length_overall* has the highest total order Sobol index. Although with lower index values than Container ships, GC ship tween decks has *depth* as the highest ranked Sobol index for most speed related target features. Furthermore, the draught related features obtain the highest Sobol index values for the target features related to rotation for this ship type. MP general cargo ships show a less clear picture of which features have the highest indices, but in most cases *length_overall* is the most important input feature for both speed and rotation related target features. For Ro-Ro cargo ships, the Sobol indices show that *length_overall* is the most important feature for all speed related target features, while also being one of the most important for both rotation related target features.



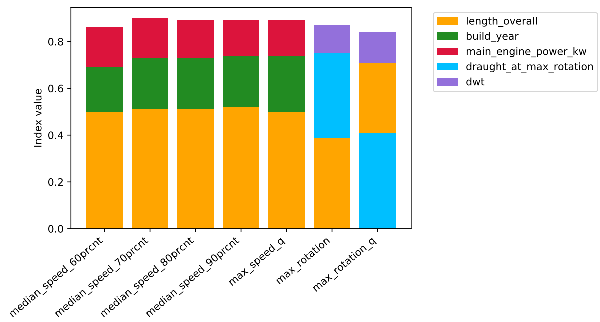
(a) Container ship



(b) GC ship tween deck



(c) MP general cargo ship



(d) Ro-Ro cargo ship

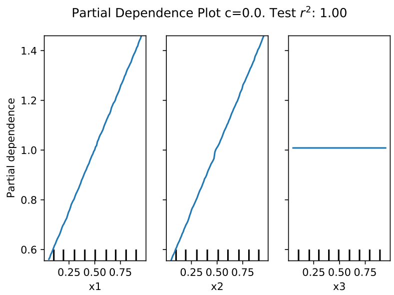
Fig. 16: Total order sobol indices for all ship types and target features.

5.6 Partial dependence plots

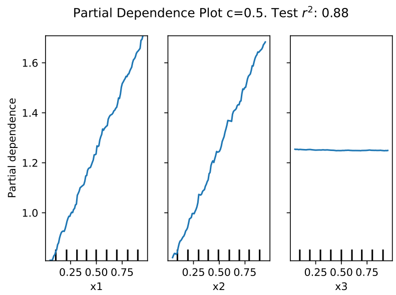
The Sobol indices give insights in which input features have the largest influence on the model output variance. In order to investigate in which way these features influence the model output partial dependency plots are created for the feature with the highest ranked Sobol indices. The plots show what kind of relation there exists between changing the value of an input feature and the effect on the target feature.

5.6.1 Test function

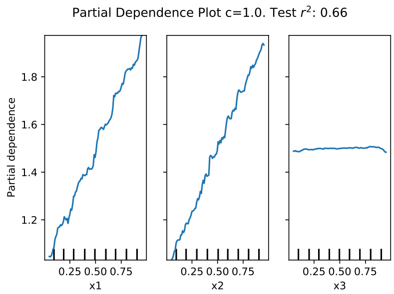
In order to determine if the PDP's show the correct relation between the input features and the target feature even when the r^2 score of the analyzed model is relatively low, Figure 17 shows the PDP's for test equation 1 with $c = \{0, 0.5, 1, 1.5, 2\}$. We fitted an RF regressor to 7k training samples and evaluated the performance of the model on 3k test samples. The title of the plots show that as the amount of noise added to the function increases, the r^2 score on the test set decreases. From the function definition it is clear that x_1 and x_2 both contribute equally and linearly to the output of f, and that x_3 only contributes to the output when $c > 0$. The plots show that even when the r^2 score is relatively low (< 0.5) the PDP's still approximate a correct visualisation of the influence of the different input parameters.



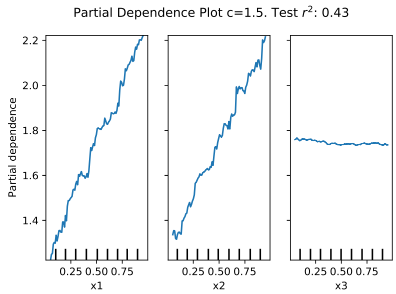
(a) c=0



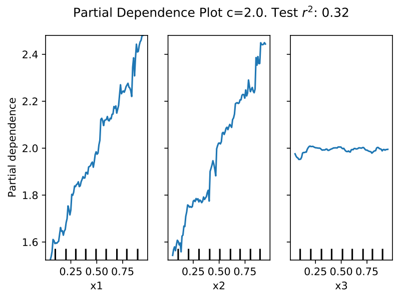
(b) c=0.5



(c) c=1



(d) c=1.5

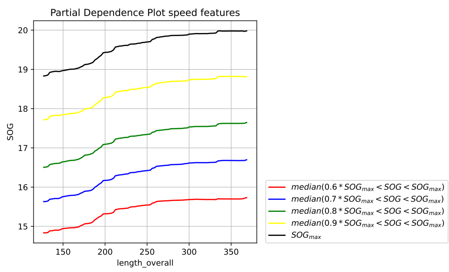


(e) c=2

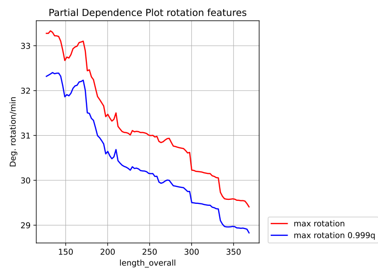
Fig. 17: Partial dependence plots for RF regressor fitted to test equation 1 with $c = \{0, 0.5, 1, 1.5, 2\}$.

5.6.2 Application to AutoML model

For each ship type the PDP's are created for features with the highest ranked sobol indices. As experiments have shown that there is little to no difference between the rankings of the first and total order indices, the ranking from the total order index is used for determining which features are plotted.



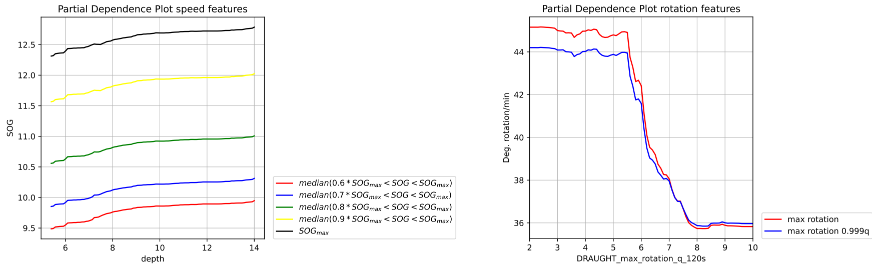
(a)



(b)

Fig. 18: Partial dependence plots for speed related (a) and (b) rotation related target features of container ships for input feature *length_overall*.

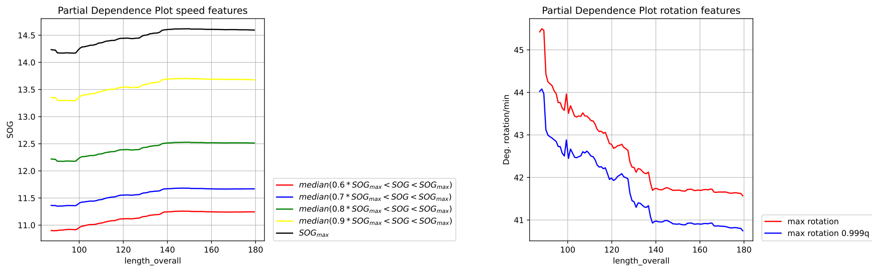
Figure 18a shows that there exists a positive linear trend between the speed related target features and the overall length of Container ships. The lines have similar gradients for each target feature although the intercept is higher for target features that resemble more closely to the maximum speed of a vessel. Furthermore, it can be noticed from Figure 18b that for the rotation related target features there exists a negative trend between the target features and the overall length of the vessel.



(a) (b)

Fig. 19: Both plots show partial dependence for GC ship tween deck.(a) shows the Partial dependence plot for input feature *depth* for all speed related target features. (b) shows the draught at max rotation and draught at max rotation0.999q

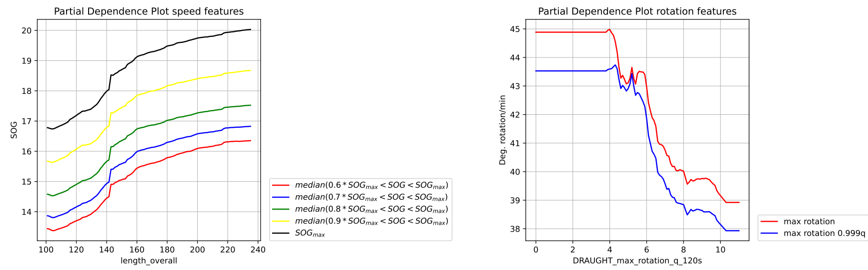
Figure 19a shows that there exists a positive trend between the speed related target features of GC ship tween decks and their depth, although the gradient of the trend is relatively small. As the Froude number had a similar Sobol index as depth for the speed related target features, we have also created a PDP with Froude number as input feature. As the resulting line is almost flat we have appended the resulting plot in Figure A4. From 19b it can be seen that for rotation related target features there is a relatively large drop in degrees of rotation per minute for ships with a draught at max rotation between 6 and 8 meters. For ships that have a draught at max rotation that lie above this interval there is a lower degree of rotation per minute, while there is a higher degree of rotation per minute for ships that have a draught at max rotation that lies below this interval.



(a) (b)

Fig. 20: Both plots show partial dependence for MP general cargo ships for input feature *length_overall*.

For MP general gargo ships Figure 20a shows a slightly positive trend between the speed related features and the overall length of a ship. Figure 20b shows a negative trend between the overall length of the ship and the maximum degrees of rotation per minute.



(a) (b)

Fig. 21: Both plots show partial dependence for Ro-Ro cargo ships for input feature *length_overall* and draught at max speed/draught at max speed0.999q.

The plots in Figure 21a show that for Ro-Ro cargo ships there exists a positive trend between the overall length of a ship and the speed related features, just as for Container ships. As for the rotation related target features, the plot in Figure 21b shows a negative trend between the draught at max rotation and the rotational capabilities of a ship.

As for most ship types the overall length of the ships is the most important feature for the speed related target features, Figure 22 shows a PDP for all ship types with on the x-axis the overall length, and on the y-axis the SOG_{max} of the vessels. Two conclusions can be drawn from this figure: the SOG_{max} of Ro-Ro cargo ships is influenced the most by the overall length of the ship. Furthermore we see that Container ships and Ro-Ro cargo ships are larger, and have a larger range of length values than GC ship tween deck and MP cargo ships.

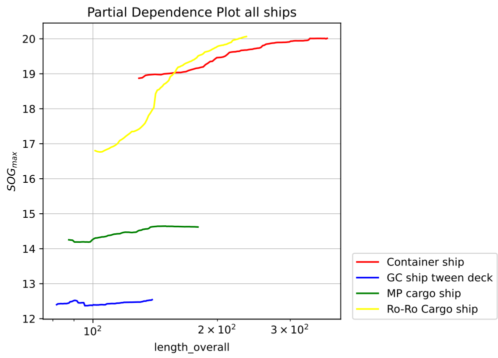


Fig. 22: PDP showing the effect of the overall length of a ship for all ships types on the maximum SOG of the ship. X-axis is in logarithmic scale.

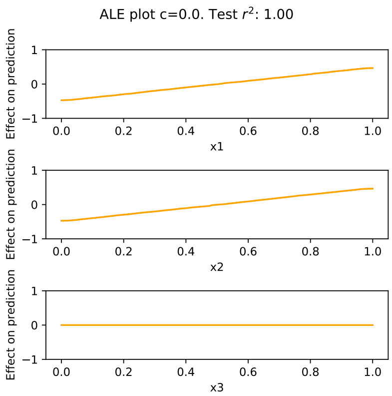
As the partial dependence plots are calculated by averaging over all samples we risk using samples that are unlikely to occur in real life such as very large ships that are very narrow, or have very low engine power. In order to check the validity of the PDP results, the next section shows the experiments performed for ALE plots to see if they show similar trends as the PDP's.

5.7 ALE plots

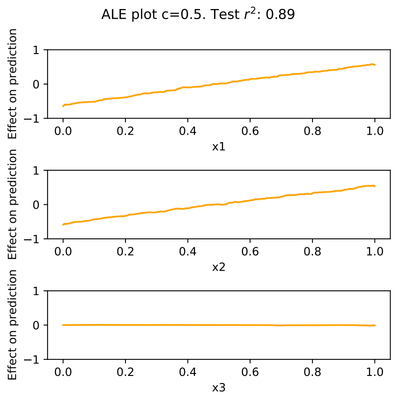
The PDP's show what kind of relation there is between the most influential input features and the absolute value of the target feature. ALE plots can complement these results by visualising how much the prediction of the target features change if we slightly adjust an input feature.

5.7.1 Test function

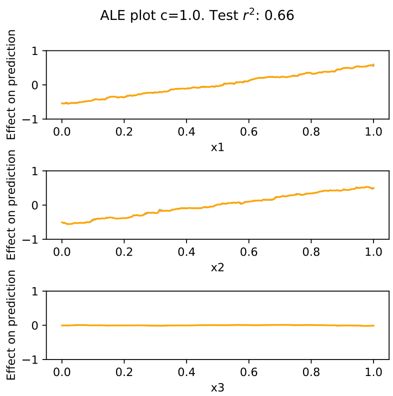
Figure 23 shows the ALE plots for test equation 1 with noise term $c = \{0, 0.5, 1, 1.5, 2\}$ that is modeled by an RF regressor. The goal is to see if the ALE plots model the correct relation between the input features and target feature even when the r^2 score of the analyzed model is relatively low. The regressor is trained on 7k training samples, and evaluated on 3k test samples. The titles of the plots show that when the noise level increases, the r^2 score on the test set decreases. The plots show, similar to the PDP's, that even when the r^2 score of the model that is being analyzed is relatively low, the ALE plots still approximate a correct visualisation of the influence of the input parameters.



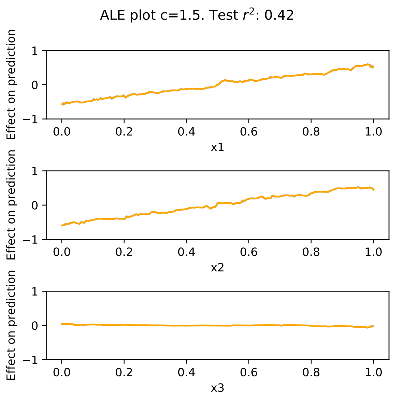
(a) c=0



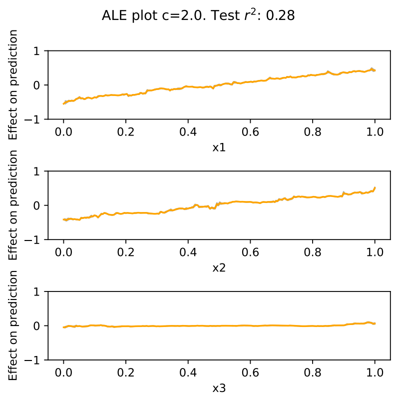
(b) c=0.5



(c) c=1



(d) c=1.5

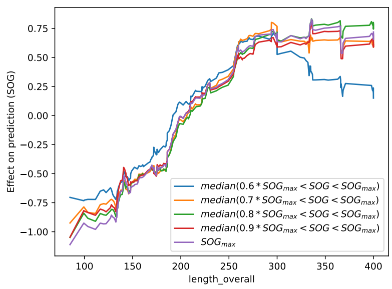


(e) c=2

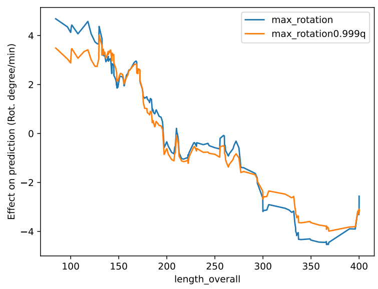
Fig. 23: Ale plots for RF regressor fitted to test equation 1 with $c = \{0, 0.5, 1, 1.5, 2\}$.

5.7.2 Application to AutoML model

For all ship types and target features the influence of input parameters on the prediction of the AutoML model created in Section 5.4 is visualised by creating ALE plots.



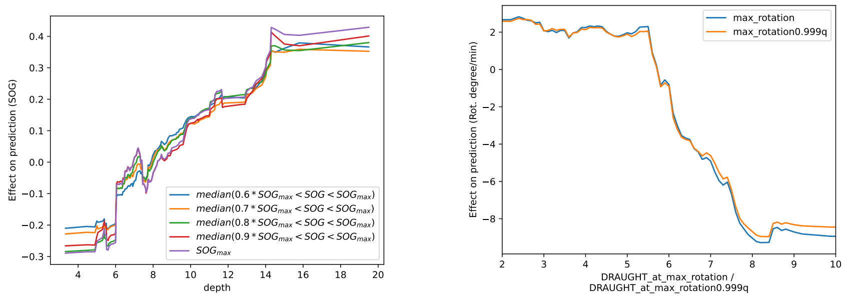
(a)



(b)

Fig. 24: Both figures show ALE plots for Container ships for input feature *length_overall*.

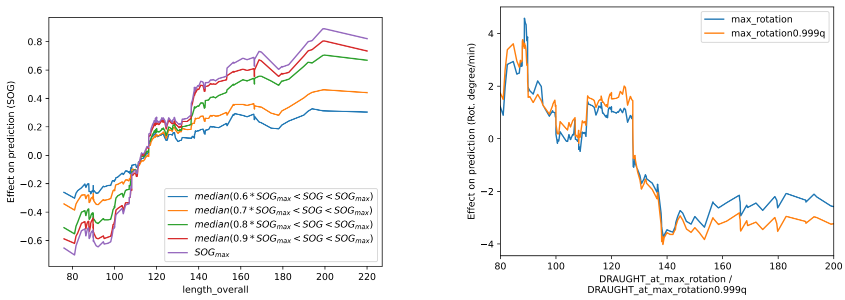
Figure 24 shows that there exists a positive trend between the overall length of container ships and the speed related target features. The plots show a more oscillating line than for the PDP's, this can be explained by the fact that ALE plots calculate an average of the target feature over samples that lie in a small interval, while PDP's use all samples in the dataset which results in a smoother line. What can also be noticed is that there is less difference between the intercept of the ALE lines than for the PDP lines. This can be attributed to the fact that the y-axes for ALE's show only the effect on prediction, while the y-axes for PDP's show the absolute value of the target features. When we look at the ALE plot for rotation related features we see that the predicted value of degrees of rotation per minute decreases when the overall length of a ship increases, just as we have seen in the PDP's.



(a) (b)

Fig. 25: Both figures show ALE plots for GC ship tween decks. (a) for *depth* and (b) for *draught_at_max_rotation/draught_at_max_rotation_0.999q*.

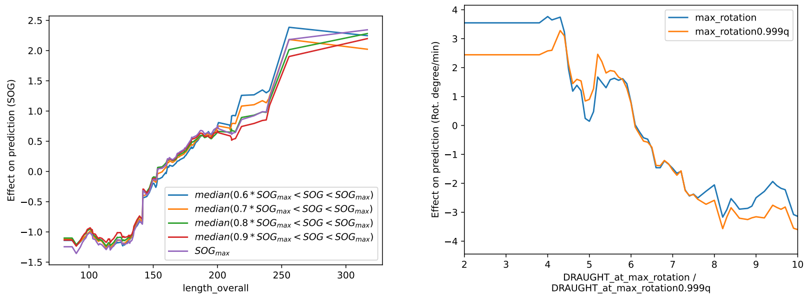
In Figure 25 the plots show that for GC ship tween decks there exists a positive trend between the predicted speed related target features and the depth of the ship, although the effect is very small as it ranges from −0.3 to 0.4 over the complete range of depth values. As the Sobol index for input feature froude number had a similar value, an ALE plot for froude number and the speed related target features is appended in Figure A4. The ALE plot for rotation related target features shows little to no effect on the prediction for ships with a draught at max rotation between 2 and 6 meters. Between 6 and 8 meters we see a sharp drop in effect on predicted degrees of rotation per minute. From 8 to 10 meters we again see little to no effect on the predicted target value. This is again, the same trend as we have seen in the PDP's.



(a) (b)

Fig. 26: Both figures show ALE plots for MP general cargo ships. (a) for input feature *length_overall* and (b) for *draught_at_max_rotation/ draught_at_max_rotation0.999q*.

The ALE plots for Multi Purpose general cargo ships can be found in Figure 26. Here the plots again show a positive trend between the overall length of a ship and the predicted target value for speed related target features. Furthermore, it can be noticed that for rotation related features the overall length of a ship has a negative effect on predicted degrees of rotation per minute. Figure 26b shows a relatively large difference between ships with an overall length between 80 and 140 meters. For ships longer than 140 meters the plot shows that there is little to no difference in prediction, just as for the PDP's in Figure 20.



(a) (b)

Fig. 27: Both figures show ALE plots for Ro-ro cargo ships. (a) for input feature *length_overall*, and (b) for *draught_at_max_rotation/ draught_at_max_rotation0.999q*.

Figure 27 shows the ALE plots for Ro-ro Cargo ships. 27a shows that there exists a positive trend between the overall length of the ships and the speed related target features, while 27b shows a negative trend between the predicted degrees of rotation per minute and the draught at max rotation of a ship. This trend is similar to the trends shown in the PDP's in Figure 21. Only a limited number of Ro-Ro cargo ships have a length of more than 250m, so the large increase in effect on prediction that the plot shows after 250m is based on a limited number of samples.

6 Discussion of the results

The experiments in Section 5 showed that using a combination of AIS data and static ship data does not always result in good regression models that can accurately predict ship performance related target features. However, for Container ships, MP cargo ships, GC ship tween deck, and Ro-ro cargo ships good regression models have been created for speed and rotation related target features. In some cases taking the absolute maximum value for a target feature was susceptible to outliers. In order to overcome this, the 0.999th quantile value was used. Further outliers were taken account of by using the Inter Quartile Range method. Other feature engineering steps included slightly altering the definition of target features which resulted in more samples that were no longer considered outliers. The use of more advanced regression algorithms, and tuning the hyperparameters of these algorithms resulted in even higher r^2 scores. The performance of the models was further increased by training a multi-output AutoML pipeline that could exploit the relation between the multiple target features. A multi-output AutoML model resulted in the highest r^2 scores for most target features and ship types. Analyzing the multi-output model shows how the multiple targets change to a change in input. The Sobol indices for these models showed that for Container ships, the overall length of the ship is the most important input feature for both speed related and rotation related target features. For GC ship tween deck and Ro-ro cargo the results also show that the overall length is the most important input feature for the speed related target features. For MP general cargo ships the depth is the most important input feature for the speed related target feature. When looking at the rotation related target features we see for MP general cargo, GC ship tween deck, and Ro-ro cargo ships that the draught at the moment the maximum rotation is logged is the most important input feature.

The fact that it was not possible to create good regression models for all ship types and target features means that for most ship types and target features, the static ship parameters do not contain enough information to explain the variance in the target features. Especially for the maximum lateral speed the model completely fails to learn any patterns in the data. This can be attributed to the fact that the calculation of the lateral speed assumes that the AIS transponder is placed in the middle of the vessel. A domain expert pointed that in some cases the transponder can be placed on either the bow, the stern, or somewhere in between on the ship. The position of the AIS transponder greatly influences the calculation of the lateral speed, as it uses the distance to bow and distance to stern. To the best of our knowledge, there is unfortunately no automated way to determine the position of the AIS transponder so that this issue can be avoided. Including more information about factors that were not possible to account for in this work, such as accurate weather data or data about engine settings might increase the r^2 scores of these models.

The GSA and XAI methods used in Section 5.5, 5.6, and 5.7 show how the input features influence the output of the model, and thus how the input features influence the speed and rotational capabilities of a ship. Experiments on a test function showed that even when the r^2 is lower than 0.5 the PDP's and ALE plots still give a good approximation of the influence of the input parameters. Most ship types had the overall length of the ship as the feature with the highest ranked Sobol index. Furthermore, the PDP and ALE plots showed a positive trend between the predicted speed capabilities and overall length of ships. An increase of the overall length of a ship often means that the breadth of the ship increases as well, but only up to a certain level, as ships often have to be narrow enough to fit through locks used in busy sailing routes. Increasing the width of a ship is associated with the largest increase in resistance called wave resistance. On the other hand, increasing the length of a ship is associated with a much smaller increase in resistance called boundary layer induced friction [53]. To account for the increase in resistance for longer and wider ships, these vessels are often installed with extra engine power. Even though the engine power was one of the input features, it is never selected as most important inpute feature. As the breadth of Container ships tend to max out around 60 meters, all extra engine power installed after this can be used to compensate for extra length of a ship. As the extra resistance as a result of a longer ship is much smaller than that of a wider ship, the results show that longer container ships exploit the extra engine power in a more efficient way than shorter or broader ships.

The influence of the overall length of a vessel on the rotational capabilities of Container ships can be explained by the fact that a shorter ship has to displace less water when turning than a longer ship. These explanations of the importance of the overall length of a ship are confirmed by the PDP's and ALE plots in Section 5.6 and 5.7 which show that there exists a negative trend between the rotational capabilities of Container ships and the overall length.

When looking at the Sobol indices for the speed related features for GC ship tween decks there is no single feature that explains a large part of the model output variance, but rather three separate features that all explain a relatively small part of it. The depth is the most important input feature for speed related target features. The PDP's and ALE plots show that there is a slightly positive trend between the depth and the speed related target features. For the rotation related target features the regression models have a relatively low r^2 score. A domain expert's explanation for this is that the rotational capabilities of these ships are heavily determined by how many bow thrusters they have, and how powerful they are. Unfortunately, the reference database did not contain sufficient information about the bow thrusters for this ship type, causing it to be excluded from our analysis. When looking at the Sobol indices for the rotation related features, the draught at the moment the maximum rotation value was logged is the most important feature. This can be explained by the fact that when a vessel has a lower draught it has to displace less water when turning, which allows it to turn faster. This explanation is confirmed by the PDP's and ALE plots which show a sharp drop in the predicted degrees of rotation per minute when ships have a higher draught.

The results of the experiments for MP General Cargo ships again show that the overall length of the ship is the most important input feature for both the speed and rotation related target features. The same explanations hold here as for the Container ships.

For Ro-ro cargo ships the r^2 scores of the regression models for speed related target features were higher than for any other ship type. The Sobol indices showed that for speed related target features the overall length of a ship is the most important input feature. The PDP's and ALE plots show similar trends as for Container ships and Multi Purpose general cargo ships; a positive trend between the overall length of the vessel and the speed and rotation related target features. What is different for Ro-ro cargo ships is that there is a much steeper line for speed related target features in both the PDP's and ALE plots compared to the two previously mentioned ship types.

A limiting factor in this research was that some important information, such as accurate weather data or information about a ships engine settings was not available. Furthermore, the coverage of AIS hub is mainly focused on coastal waters, and does not cover the open oceans between the continents. In future work, using a source of AIS data that has more coverage might increase the quality of the data and thus the performance of the regression models.

7 Conclusion

To the best of our knowledge this is the first work where AIS data is used to analyse the performance of a large volume of ships with relation to their design characteristics. During this research, AIS data from over 30.000 vessels during a period of multiple months has been collected. Exploratory analysis has shown that there exists a large difference between the number of occurrences of different ship types, and that AIShub is not able to provide a continuous input stream of data for all ships over the complete time period of interest. Because of this a subset was used containing data from frequently occurring vessel types, that transmit data over a multiple day time span and with a sufficient number of total transmitted data points.

In order to gain insights into how ship design and performance are related we have used static ship data to predict the performance related target values calculated from AIS data. The results showed that that only for a subset of the data good performing regression models could be created. For other ship types and target features better regression models could have been made if extra information was available such as accurate weather data, or the engine settings of a ship.

For the ship types and target features for which good regression models were created there is no single combination of feature engineering steps and algorithms that performs best for each ship type and target feature. However, for most ship types the best performing model resulted from a multi-output AutoML pipeline. Experiments on a test function showed that even for r^2 values below 0.5 the PDP´s and ALE plots provide a good approximation of the influence of the input features on the target feature. When analyzing the AutoML models with GSA and XAI techniques it can be concluded that for Container ships the overall length is the most important feature for both the speed related, and rotation related target features. When increasing the length the PDP's and ALE plots show that ships tend to sail faster, and rotate slower. The results for GC ship tween deck show that the depth, and Froude number are equally important for the speed related features. For both input features the PDP's and ALE plots show that there exists a small positive trend between the input features and speed related target features. When looking at rotation related target features, the Sobol indices show that the draught at max rotation is the most important feature. For MP General cargo ships the same trends hold for both the speed related and rotation related target features as for Container ships. Finally, for Ro-Ro cargo ships the results also show that the overall length of the ship is the most important feature for the speed related target features while the draught at max rotation is the most important feature for the rotation related target features. The PDP's and ALE plots show a positive trend, much steeper than for other ship types, between the overall length of Ro-Ro cargo ships and the speed related target features, and a negative trend between the draught at max rotation and rotation related target features.

This work can be extended in the future by applying the same methods to AIS data from a provider that has a larger coverage. A larger coverage might increase the quality of the data, or expose different operational characteristics of ships. Including accurate weather data, or more onboard sensor data of ships migh also lead to interesting insights. Another interesting approach would be to skip the feature calculation step and apply a time series approach where a Recurrent Neural Network automatically derives the important features.

Appendix A

Table A1: Feature descriptions of static ship data

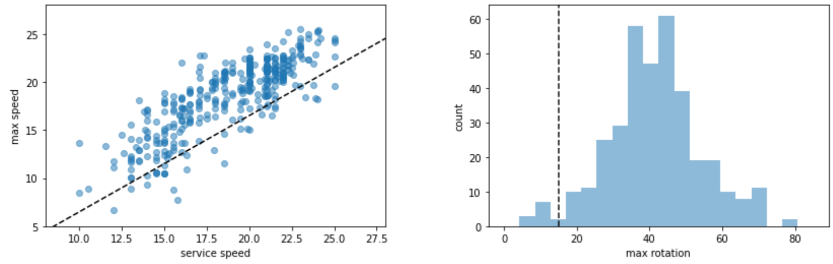
Explanation	**Type**	**Range**	**Feature name**
additional service info	Extra information about a vessel	categorical	N/A
vessel_age	Years passed since vessel launch	numerical	>0
volume_of_asphalt	Volume of asphalt cargo	numerical	>0
auxiliary_engine_power_hp	Power of the engine that powers everything except a ships movement. Measured in horsepower	numerical	>0
auxiliary_engine_power_kw	Power of the engine that powers everything except a ships movement. Measured in kilowatts	numerical	>0
volume_of_bale	Volume of bale in m^3	numerical	>0
number_of_berths	Number of beds	numerical	>0
bollard_pull	Measure of pulling force in ton	numerical	>0
breadth_moulded	Maximum breadth measured amidship to the moulded line of the frame in a ship	numerical	>0
breadth_overall at the widest point	Overall width of the ship measured numerical	>0	
build_year	Year in which ship has been built	numerical	>1960
bulb	Wether the ship has a bulb or not	bool	yes/no
number_of_cabins	Number of cabins on board	numerical	>0
cad	cad	numerical	>0
callsign	Radio call name	string	N/A
capacity_of_cargo_pumps	Capacity of cargo pumps in m^3/h	numerical	>0
number_of_cargo_pumps	Number of cargo pumps	numerical	>0
total_volume_of_cargo	Total volume of cargo in m^3	numerical	>0
number_of_cars	Car capacity	numerical	>0
block_coefficient	Ratio of the underwater volume of a ship to the volume of a rectangular block having the same overall length, breadth and depth	numerical	[0.1]
classification_society technical inspections	Organization that executes categorical	N/A	
comp_gross_tonnage	Amount of work necessary to build a ship	numerical	>0
construction_month	Month in which construction of ship finished	numerical	[0,12]
construction_year	Year in which construction of ship finished	numerical	>1960
total_number_of_cranes	Number of installed cranes	numerical	>0
total_crane_swl	Safe working load of cranes in tonnes	numerical	>0
number_of_crew	Number of crew on board	numerical	>0
continous_surface_rating		numerical	>0
number_of_decks	The number of decks on the ship	numerical	>0
delivery_month	Month in which the ship was delivered	numerical	>0
depth	Distance from top of keel to top of the deck beam op uppermost continuous deck	numerical	>0
displacement	Volume of the water that a ship pushes aside when floating in m^3	numerical	>0
draught	Distance from keel to waterline	numerical	>0
length_of_dry_holds	Length over dry hold areas	numerical	>0
number_of_dry_holds	Number of dry holds	numerical	>0
dwt	Deadweight tonnage; amount of mass a ship can transport	numerical	>0
dwt_by_formula	Deadweight measured by formula	numerical	>0
dwt_div_lsw	DWT divided by light ship weight	numerical	>0
number_of_electric_engines	Number of electric engines	numerical	>0
number_of_engine_cylinders	Number of engine cylinders	numerical	>0
designer_of_engine	Company that designed engine	numerical	>0

engine_maximum_continous_rating	Maximum power output of engine while running continuously at safe limits and conditions	numerical	>0
number_of_engines	Number of engines	numerical	>0
engine_revolutions_per_minute	Number of revolutions per minute of engine	numerical	>0
stroke_of_engine	Type of power cycle used in engine	numerical	[2,4]
type_of_engine	Type of engine	categorical	N/A
flag	Flag under which the ship sails	categorical	N/A
froude_number	Measure of the ratio of the inertia force on an element of fluid to the weight of the fluid element	numerical	[0,1]
gas_fuelled	Whether the ship is gas fuelled	bool	T/F
volume_of_grain	Max capacity of grain in mˆ3	numerical	>0
gross_tonnage	Internal volume of a ship in tonnes	numerical	>0
height	Height in m	numerical	>0
hopper_volume		numerical	>0
hull	Hull type	categorical	N/A
hull_construction	Technique used for constructing the hull	categorical	N/A
hull_material	Material used for constructing hull	categorical	N/A
hull_type	Type of hull	categorical	N/A
ice_class	If vessel has an ice class	categorical	N/A
insulated_volume			
months_from_keel_layed_to_launch	Months between start of keel and launch of ship	numerical	>0
heigh_of_lanes	Height of lanes	numerical	>0
length_of_lanes	Length of lanes	numerical	>0
width_of_lanes	Width of lanes	numerical	>0
months_from_launch_to_commission	Months from launch to commission	numerical	>0
lb+d			
lb+t			
lbd			
ldt			
length_between_perpendiculars	Length of the ship along the summer load line from the forward surface of the stem	numerical	>0
length_overall	Length of the ship measured from two outermost points of the ship	numerical	>0
ligthship_weight	Weight of the ship without fuel, passengers, cargo, and water in tonnes	numerical	>0
total_volume_of_liquid_cargo	Volume of liquid cargo that the ship can hold in mˆ3	numerical	>0
total_volume_of_liquified_gas	Volume of liquid gas that the ship can hold in mˆ3	numerical	>0
number_of_lorries	Number of lorries that the ship can hold in mˆ3	numerical	>0
main_engine_power_hp	Main engine power in horse power	numerical	>0
main_engine_power_kw	Main engine power in kw	numerical	>0
total_maximum_continuous_rating	Maximum continuous rated power output in kw	numerical	>0
mmsi	Maritime Mobile Service Identity number	numerical	>0
net_tonnage	Dimensionless index calculated from the total moulded volume of a ships cargo space	numerical	>0
newbuilt_price	Price of the newbuilt ship	numerical	>0
old_name	Previous name of the ship	str	N/A
operator	Operator of the ship	str	N/A
owner_country	Country of the owner	categorical	N/A
number_of_passengers	Maximum number of passengers	numerical	>0
number_of_propellers	Number of propellers	numerical	>0
type_of_propellers	Type of propellers	categorical	N/A
number_of_railway_wagons	Number of railway wagons	numerical	>0
length_of_ramp	Length of ramp	numerical	>0
location_of_ramps	Location of ramps	numerical	>0
number_of_ramps	Number of ramps	numerical	>0
swl_of_ramps	Safe working load of ramps	numerical	>0
width_of_ramps	Width of ramps	numerical	>0

propeller_revolutions_per_minute	Number of revolutions that a single propeller can make	numerical	>0
volume_of_segregated_ballast	Volume of segregated ballast in m^3	numerical	>0
service_speed	Service speed of ship	numerical	>0
ship_name	Ship name	str	N/A
ship_type_1	Ship division	categorical	N/A
ship_type_2	Ship category	categorical	N/A
ship_type_3	Ship type	categorical	N/A
ship_type_4	Ship type	categorical	N/A
ship_type_5	Ship type	categorical	N/A
ship type additional	Additional ship type information	str	N/A
ship_type_total	All ship types combined	str	N/A
ship_owner	Owner of the ship	str	N/A
ship_builder	Builder of ship	str	N/A
total_volume_of_slop	Total volume of slop tanks in m^3	numerical	>0
specials	Special features of ship	str	N/A
status	Status code	categorical	N/A
status additional	Additional information about status	str	N/A
standard_design_name	Name of standard design	categorical	N/A
tons_per_cm_immersion	Mass that has to be loaded to change draught in salt water by 1 cm	numerical	>0
number_of_teu	Maximum container capacity (twenty foot equivalent units)	numerical	>0
number_of_teu_on_deck	Maximum number of teu on deck	numerical	>0
number_of_teu_in_holds	Maximum number of teu in holds	numerical	>0
number_of_thrusters_aft	Number of thrusters at rear of ship	numerical	>0
number_of_thrusters_fwd	Number of thrusters at bow of ship	numerical	>0
power_of_thrusters_aft	Power of thrusters at rear of ship in kw	numerical	>0
power_of_thrusters_fwd	Power of thrusters forward	numerical	>0
number_of_trailers	Number of trailers that fit on ship	numerical	>0
length_of_tween_holds			
number_of_tween_holds			
type_size	Type and size combinde	categorical	>0
yard_country	Country of ship yard	categorical	>0

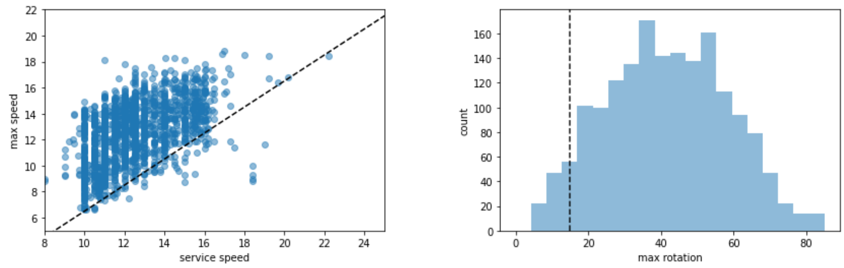
Table A2: Table showing the average r^2 scores over 5-fold cross validation of RF regression model when deleting a percentage of the input values and filling them with the KNN imputing method.

	median cruise speed x						
percentage of samples removed	**0.6**	**0.7**	**0.8**	**0.9**	**max speed**	**max rotation**	**max rotation0.999q**
0	0.54	0.69	0.71	0.69	0.66	0.49	0.48
5	0.53	0.70	0.70	0.68	0.65	0.48	0.48
10	0.53	0.69	0.70	0.68	0.65	0.48	0.48
15	0.54	0.70	0.70	0.68	0.65	0.48	0.48
20	0.52	0.66	0.68	0.65	0.62	0.48	0.47
25	0.51	0.66	0.66	0.64	0.62	0.48	0.45
30	0.51	0.65	0.66	0.61	0.61	0.45	0.43



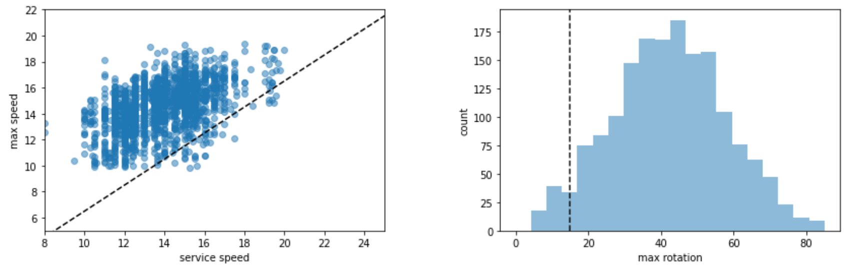
(a) (b)

Fig. A3: Plots showing how outliers are removed for max speed and max rotation for ship type ro-ro cargo ship. For a, samples below the dotted line are removed. For b, samples left of the dotted line are removed. A much smaller number of samples is deleted when compared to container ships. This can be attributed to the fact that ro-ro cargo ships do not sail on scheduled services, and thus sail at full power more often.



(a) (b)

Fig. A1: Plots showing how outliers are removed for max speed and max rotation for ship type General cargo ship tween deck. For a, samples below the dotted line are removed. For b, samples left of the dotted line are removed. A much smaller number of samples is deleted when compared to container ships. This can be attributed to the fact that General Cargo ship tween deck ships do not sail on scheduled services, and thus sail at full power more often



(a) (b)

Fig. A2: Plots showing how outliers are removed for max speed and max rotation for ship type multi purpose general cargo ship. For a, samples below the dotted line are removed. For b, samples left of the dotted line are removed. A much smaller number of samples is deleted when compared to container ships. This can be attributed to the fact that Multi-purpose general cargo ships do not sail on scheduled services, and thus sail at full power more often

Table A3: Average r^2 scores over 5-fold CV for modus cruise speed Container Ships, General Cargo Ship Tween Decks, and Multi-Purpose General Cargo Ships

	modus cruise speed			
	0.6	**0.7**	**0.8**	**0.9**
Container ship	0.22	0.38	0.50	0.51
GC ship tween deck	0.51	0.52	0.53	0.52
MP GC ship	0.47	0.52	0.48	0.37

Table A4: Ensemble details of AutoML model for container ship.

Ensemble weight	**Model type**	**Loss on validation**
0.04	Extra trees	0.42
0.02	Extra trees	0.42
0.02	Extra trees	0.42
0.04	Extra trees	0.42
0.36	Extra trees	0.42
0.20	Extra trees	0.42
0.32	Extra trees	0.54

Table A5: Ensemble details of AutoML model for General Cargo ship tween deck.

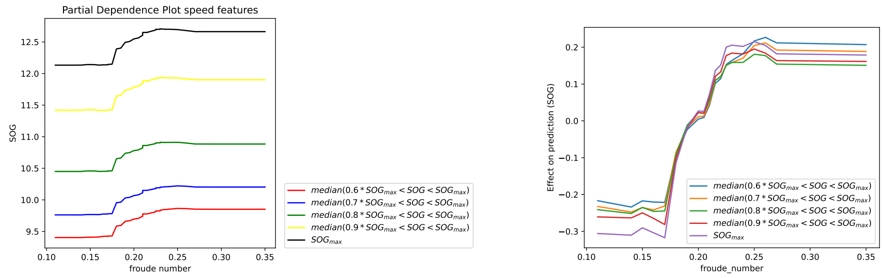
Ensemble weight	**Model type**	**Loss on validation**
0.34	Random Forest	0.54
0.26	Random Forest	0.54
0.02	Random Forest	0.54
0.12	Random Forest	0.54
0.12	Random Forest	0.54
0.02	Extra trees	0.54
0.02	Extra trees	0.54
0.02	Extra trees	0.54
0.02	Extra trees	0.54
0.04	Extra trees	0.54
0.02	Extra trees	0.54

Table A6: Ensemble details of AutoML model for Multi Purpose General Cargo ship.

Ensemble weight	**Model type**	**Loss on validation**
0.02	Extra trees	0.55
0.04	Extra trees	0.55
0.06	Extra trees	0.55
0.02	Extra trees	0.55
0.02	Extra trees	0.55
0.02	Extra trees	0.55
0.02	Extra trees	0.55
0.04	Extra trees	0.55
0.04	Extra trees	0.55
0.06	Extra trees	0.55
0.06	Extra trees	0.55
0.06	Extra trees	0.55
0.02	Extra trees	0.55
0.02	Extra trees	0.55
0.02	Extra trees	0.55
0.06	Extra trees	0.55
0.02	Extra trees	0.55
0.02	Extra trees	0.55
0.04	Extra trees	0.55
0.02	Extra trees	0.55
0.08	Extra trees	0.56
0.24	Extra trees	0.57

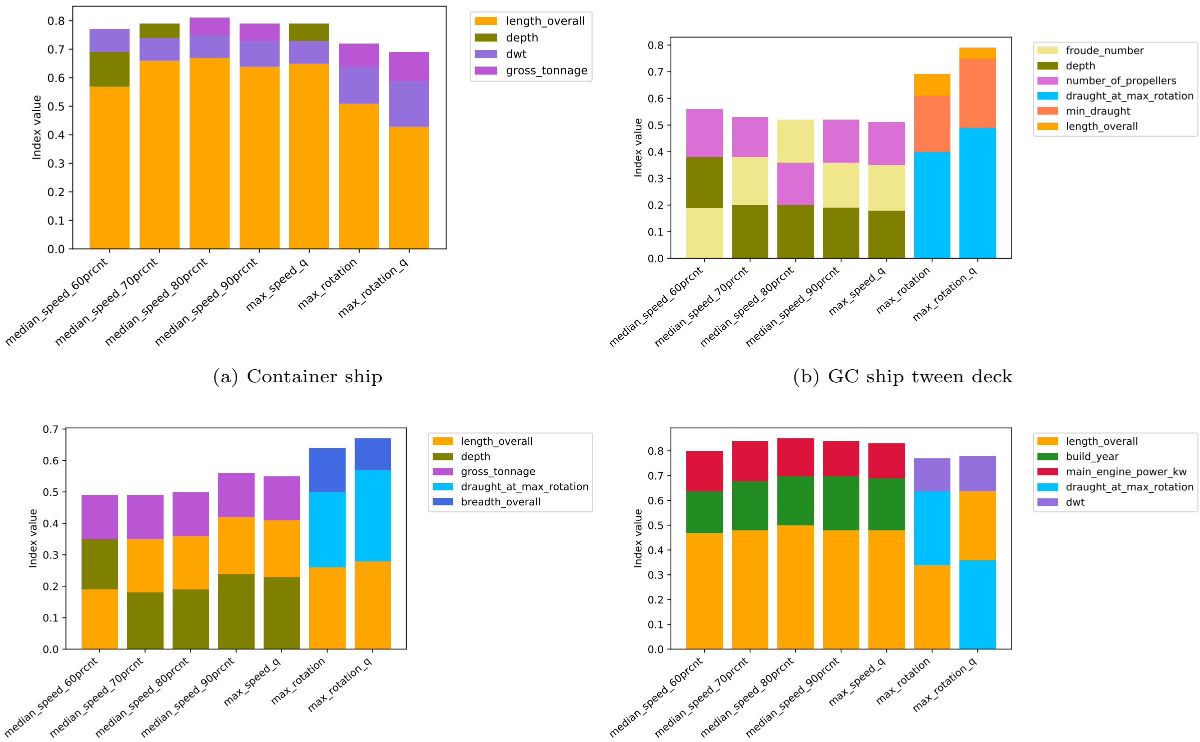
Table A7: Ensemble details of AutoML model for Ro-ro cargo ship.

Ensemble weight	Model type	Loss on validation
0.46	Extra trees	0.44
0.22	Extra trees	0.46
0.18	Random forest	0.46
0.02	Extra trees	0.47
0.12	Random forest	0.47



(a) (b)

Fig. A4: (a)PDP for GC ship tween deck for speed related target features and froude number as input feature. (b) ALE for GC ship tween deck for speed related target features and froude number as input feature.



(a) Container ship

(b) GC ship tween deck

(c) MP general cargo ship

(d) Ro-Ro cargo ship

Fig. A5: First order sobol indices for all ship types and target features.

References

[1] Munim ZH, Dushenko M, Jimenez VJ, Shakil MH, Imset M. Big data and artificial intelligence in the maritime industry: a bibliometric review and future research directions. Maritime Policy & Management. 2020;47(5):577–597. https://doi.org/10.1080/03088839.2020.1788731. https://arxiv.org/abs/https://doi.org/10.1080/03088839.2020.1788731.

[2] Ming Z, De quan M, qiu LJ. A dynamic vessel fleet planning model in uncertain circumstances. In: 2009 IEEE International Conference on Automation and Logistics; 2009. p. 379–384.

[3] Panda J.: Machine Learning for Naval Architecture, Ocean and Marine Engineering.

[4] Gaspar HM, Rhodes DH, Ross AM, Ove Erikstad S. Addressing complexity aspects in conceptual ship design: A systems engineering approach. Journal of Ship Production and Design. 2012;28(04):145–159.

[5] Keane RG, Deschamps L, Maguire S. Reducing Detail Design and Construction Work Content by Cost-Effective Decisions in Early Stage Naval Ship Design. In: SNAME Maritime Convention. OnePetro; 2014. .

[6] Ao Y, Li Y, Gong J, Li S. An artificial intelligence-aided design (AIAD) of ship hull structures. Journal of Ocean Engineering and Science. 2021;https://doi.org/https://doi.org/10.1016/j.joes.2021.11.003.

[7] Abramowski T. Application of Artificial Intelligence Methods to Preliminary Design of Ships and Ship Performance Optimization. Naval Engineers Journal. 2013 09;125:101–112.

[8] Kumar V, Garg M. Predictive analytics: a review of trends and techniques. International Journal of Computer Applications. 2018;182(1):31–37.

[9] V FA, G P, M V. MARITIME TRAFFIC NETWORKS. From Historical Positioning Data to Unsupervised Maritime Traffic Monitoring [Full paper]. IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS. 2018;19(3):722–732. https://doi.org/10.1109/TITS.2017.2699635.

[10] Winther M, Christensen JH, Plejdrup MS, Ravn ES, Eriksson ÓF, Kristensen HO. Emission inventories for ships in the arctic based on satellite sampled AIS data. Atmospheric Environment. 2014 Jul;91:1–14. https://doi.org/10.1016/j.atmosenv.2014.03.006.

[11] Mitchell K, Scully B. Waterway Performance Monitoring with Automatic Identification System Data. Transportation Research Record Journal of the Transportation Research Board. 2014 12;2426:20. https://doi.org/10.3141/2426-03.

[12] Chen L, Zhang D, Ma X, Wang L, Li S, Wu Z, et al. Container Port Performance Measurement and Comparison Leveraging Ship GPS Traces and Maritime Open Data. IEEE Transactions on Intelligent Transportation Systems. 2015 12;p. 1–16. https://doi.org/10.1109/TITS.2015.2498409.

[13] Ananth CV, Kleinbaum DG. Regression models for ordinal responses: a review of methods and applications. International Journal of Epidemiology. 1997 12;26(6):1323–1333. https://doi.org/10.1093/ije/26.6.1323. https://arxiv.org/abs/https://academic.oup.com/ije/article-pdf/26/6/1323/18477637/261323.pdf.

[14] Nathans LL, Oswald FL, Nimon K. Interpreting multiple linear regression: a guidebook of variable importance. Practical assessment, research & evaluation. 2012;17(9):n9.

[15] Awad M, Khanna R. Support vector regression. In: Efficient learning machines. Springer; 2015. p. 67–80.

[16] Noble WS. What is a support vector machine? Nature biotechnology. 2006;24(12):1565–1567.

[17] Kleynhans T, Montanaro M, Gerace A, Kanan C. Predicting Top-of-Atmosphere Thermal Radiance Using MERRA-2 Atmospheric Data with Deep Learning. Remote Sensing. 2017 11;9:1133. https:

//doi.org/10.3390/rs9111133.

[18] Friedman J. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics. 2000 11;29. https://doi.org/10.1214/aos/1013203451.

[19] Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences. 1997;55(1):119–139. https://doi.org/https://doi.org/10.1006/jcss.1997.1504.

[20] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. CoRR. 2016;abs/1603.02754. https://arxiv.org/abs/1603.02754.

[21] Breiman L. Random Forests. Machine Learning. 2001 Oct;45(1):5–32. https://doi.org/10.1023/A:1010933404324.

[22] Ouedraogo I, Defourny P, Vanclooster M. Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale. Hydrogeology Journal. 2019;27(3):1081–1098.

[23] Shataee S, Kalbi S, Fallah A, Pelz D. Forest attribute imputation using machine-learning methods and ASTER data: comparison of k-NN, SVR and random forest regression algorithms. International journal of remote sensing. 2012;33(19):6254–6280.

[24] Refaeilzadeh P, Tang L, Liu H. Cross-validation. Encyclopedia of database systems. 2009;5:532–538.

[25] SKlearn.: Cross validation example. Accessed: 11-04-2022. https://scikit-learn.org/stable/modules/cross_validation.html.

[26] Hutter F, Kotthoff L, Vanschoren J. Automatic Machine Learning: Methods, Systems, Challenges. Springer; 2019.

[27] Vanschoren J, van Rijn JN, Bischl B, Torgo L. OpenML. ACM SIGKDD Explorations Newsletter. 2014 jun;15(2):49–60. https://doi.org/10.1145/2641190.2641198.

[28] Feurer M, Klein A, Eggensperger K, Springenberg JT, Blum M, Hutter F. Efficient and Robust Automated Machine Learning. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. NIPS'15. Cambridge, MA, USA: MIT Press; 2015. p. 2755–2763.

[29] Feurer M, Eggensperger K, Falkner S, Lindauer M, Hutter F. Auto-Sklearn 2.0: The Next Generation. CoRR. 2020;abs/2007.04074. https://arxiv.org/abs/2007.04074.

[30] Helton JC, Johnson JD, Sallaberry CJ, Storlie CB. Survey of sampling-based methods for uncertainty and sensitivity analysis. Reliability Engineering System Safety. 2006;91(10):1175–1209. The Fourth International Conference on Sensitivity Analysis of Model Output (SAMO 2004). https://doi.org/https://doi.org/10.1016/j.ress.2005.11.017.

[31] Yeung DS, Cloete I, Shi D, wY Ng W. Sensitivity analysis for neural networks. Springer; 2010.

[32] Harris FW. How Many Parts to Make at Once. Oper Res. 1990;38:947–950.

[33] Saltelli A, Ratto M, Tarantola S, Campolongo F. Sensitivity analysis for chemical models. Chemical reviews. 2005;105(7):2811–2828.

[34] Zi Z. Sensitivity analysis approaches applied to systems biology models. IET systems biology. 2011;5(6):336–346.

[35] Cukier RI, Fortuin CM, Shuler KE, Petschek AG, Schaibly JH. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory. The Journal of Chemical Physics. 1973;59(8):3873–3878. https://doi.org/10.1063/1.1680571. https://arxiv.org/abs/https://doi.org/10.1063/1.1680571.

[36] Nossent J, Elsen P, Bauwens W. Sobol'sensitivity analysis of a complex environmental model. Environmental Modelling & Software. 2011;26(12):1515–1525.

[37] Jutz K. Accuracy of data-based sensitivity indices. SIAM Undergraduate Research Online. 2016 01;9. https://doi.org/10.1137/15S014757.

[38] Morris MD. Factorial sampling plans for preliminary computational experiments. Quality Engineering. 1991;37:307–310.

[39] Sobol I, Kucherenko S. Derivative based global sensitivity measures. Procedia - Social and Behavioral Sciences. 2010 12;2:7745–7746. https://doi.org/10.1016/j.sbspro.2010.05.208.

[40] Borgonovo E. A new uncertainty importance measure. Reliability Engineering System Safety. 2007;92(6):771–784. https://doi.org/https://doi.org/10.1016/j.ress.2006.04.015.

[41] Pianosi F, Wagener T. A simple and efficient method for global sensitivity analysis based on cumulative distribution functions. Environmental Modelling Software. 2015;67:1–11. https://doi.org/https://doi.org/10.1016/j.envsoft.2015.01.004.

[42] Molnar C, Freiesleben T, König G, Casalicchio G, Wright MN, Bischl B.: Relating the Partial Dependence Plot and Permutation Feature Importance to the Data Generating Process.

[43] Apley DW, Zhu J.: Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models.

[44] Rødseth Perera L, Mo B. Big data in shipping - Challenges and opportunities; 2016. .

[45] Park K, Sim S, Bae H. Vessel estimated time of arrival prediction system based on a path-finding algorithm. Maritime Transport Research. 2021;2:100012. https://doi.org/https://doi.org/10.1016/j.martra.2021.100012.

[46] Winter Rd, Stein Bv, Bäck T, Bertram V. Ship design performance and cost optimization with machine learning. COMPIT'21. 2021;p. 185–196.

[47] AISHub.: AISHub webpage. Accessed: 25-03-2022. https://www.aishub.net.

[48] Stenden N.: TODDIS Project. Accessed: 17-05-2022. https://www.nhlstenden.com/nieuws-en-artikelen/toddis-project-gebruik-van-big-data-voor-innovatieve-scheepsontwerpen.

[49] Kim KI, Lee KM. Mining of missing ship trajectory pattern in automatic identification system. International Journal of Engineering Technology. 2018;7(2.12):167–170.

[50] H P V, Poornima B, Sagar B. In: Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset; 2018. p. 511–518.

[51] Nachar N. The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution. Tutorials in Quantitative Methods for Psychology. 2008 03;4. https://doi.org/10.20982/tqmp.04.1.p013.

[52] Cunningham P, Delany SJ. k-Nearest Neighbour Classifiers: 2nd Edition (with Python examples). CoRR. 2020;abs/2004.04523. https://arxiv.org/abs/2004.04523.

[53] Zeng Q, Hekkenberg R, Thill C, Hopman H. Scale effects on the wave-making resistance of ships sailing in shallow water. Ocean Engineering. 2020;212:107654. https://doi.org/https://doi.org/10.1016/j.oceaneng.2020.107654.