# Pricing the second-hand dry bulk vessel through stacking ensemble with add-on plain feedforward neural networks

Jingzhou Zhao
*Cass Business School, City University of London, London, UK*

Received 23 June 2023
Revised 16 January 2024
15 March 2024
Accepted 17 April 2024

## Abstract

**Purpose** – The accurate valuation of second-hand vessels has become a prominent subject of interest among investors, necessitating regular impairment tests. Previous literature has predominantly concentrated on inferring a vessel's price through parameter estimation but has overlooked the prediction accuracy. With the increasing adoption of machine learning for pricing physical assets, this paper aims to quantify potential factors in a non-parametric manner. Furthermore, it seeks to evaluate whether the devised method can serve as an efficient means of valuation.

**Design/methodology/approach** – This paper proposes a stacking ensemble approach with add-on feedforward neural networks, taking four tree-driven models as base learners. The proposed method is applied to a training dataset collected from public sources. Then, the performance is assessed on the test dataset and compared with a benchmark model, commonly used in previous studies.

**Findings** – The results on the test dataset indicate that the designed method not only outperforms base learners under statistical metrics but also surpasses the benchmark GAM in terms of accuracy. Notably, 73% of the testing points fall within the less-than-10% error range. The designed method can leverage the predictive power of base learners by incrementally adding a small amount of target value through residuals and harnessing feature engineering capability from neural networks.

**Originality/value** – This paper marks the pioneering use of the stacking ensemble in vessel pricing within the literature. The impressive performance positions it as an efficient desktop valuation tool for market users.

**Keywords** Second-hand dry bulk vessel valuation, Maritime economics, Machine learning, Neural network, Stacking

**Paper type** Research paper

## 1. Introduction

Research into second-hand vessel pricing has been well-developed in the past 2 decades. Many econometric models have been applied to explain the relationship between the endogenous or exogenous variables and the price itself. However, these papers tend to use benchmark or adjusted time-series data and rarely assess the accuracy and robustness of the methods in predicting actual prices. This paper focuses on real transactions and develops an efficient desktop valuation workflow that can accurately price the fair market value. Finally, it investigates the accuracy of predicted values on the test dataset from 2014 to 2022.

The second-hand vessel market thrives on price fluctuations (Stopford, 2009). The banks who lend against the mortgage on the purchased vessels often require a certified value of their collaterals over the life of their loan to keep the borrower maintaining the value level of their assets. Leasing houses typically need to estimate the future remaining value of the vessel when the period ends. Companies that are prepared for an Initial public offering (IPO) or draw up vessel-backed securities generally need a fair market value of the fleet. These contemporary demands underscore the necessity for a precise and efficient method to cater to the frequent needs of valuation.

Maritime Business Review
Vol. 9 No. 2, 2024
pp. 145-159
Emerald Publishing Limited
2397-3757
DOI 10.1108/MABR-06-2023-0043

Earlier studies attempted to model historical data and determine the influence of relevant variables on second-hand prices. Charemza and Gronicki (1981) propose equations for both supply and demand to adjust ship prices based on the freight rate and other shipping activities. Haralambides *et al.* (2004) fit a theoretical error correction model to construct an equilibrium function, suggesting that the demand for second-hand vessels is determined by the time charter rate, newbuilding price, and financing cost. Simultaneously, the supply side is expressed as the ratio of the order book to the fleet. The empirical result shows that newbuilding price and time charter rate have the most substantial effect. This finding is also discussed by Fan *et al.* (2021) through the cox proportional hazards model, which states that second-hand prices are more sensitive to expected revenue than cost. There are also opposing views on the equilibrium theory. Beenstock (1985) emphasises that the traditional supply and demand theory does not apply to vessel pricing because a ship is an asset with a very long lifespan, contrasting the equations derived from freight rate and market activity posed by Charemza and Gronicki (1981).

To address the non-linear relationship between the price and related predictors, Ådland and Koekebakker (2007) applied a non-parametric multivariate density estimation (MDE) model to the second-hand Handysize prices. This approach is designed with three factors: age, deadweight and 1-year time charter rate, as the properties of the MDE's approach tend to deteriorate when considering more variables. The empirical results indicate that, although the non-parametric model attempts to capture non-linearity, it cannot perfectly fit real vessel prices due to other factors, such as engine type and country of the shipyard. Ådland *et al.* (2018) found that sale price and energy efficiency exhibit a negative relationship under the same market conditions and with identical vessel specifications. This negative relationship is less intense in the booming dry bulk market from 2003 to 2008 than in other periods. In the exploration of the best timing of investment and divestment to create trading strategies in the second-hand market, Kalouptsidi (2014) illustrates that the most significant investment volatility is observed when the time to build declines and is more volatile than a constant time to build. These findings provide evidence that, for certain specific independent variables, the inconsistent effect necessitates the use of a non-linear model to explain the varying levels of influence.

The employment of the stacking method with an add-on neural network has received increased interest in asset valuation. According to Mohammed and Kora (2023), the effectiveness of ensemble methods is contingent upon various factors, such as the training of baseline models and how they are integrated. Additionally, in forecasting U.S. market excess return, a simple ensembling structure can outperform models like mallows model averaging, complete subset regression and elastic net regression (Zhao and Cheng, 2022). From the wide application, the stacking method with add-on neural networks alleviates the constraints of multi-collinearity or the manual judgement in the addition of interaction terms, which classical statistical methods may encounter.

The rest of this paper is structured as follows: Section 2 clarifies the data source for this paper and outlines the initial data handling before model fitting, while Section 3 describes the machine learning and benchmark models involved. Section 4 presents the results, focussing on accuracy and robustness. The conclusion and the future improvement are discussed in Section 5.

## 2. Data

Regarding feature selection, while factors like charter rate and age are known to impact prices, other variables may be contentious due to data limitations or methodology applied. Previous studies suggest that time charter rates have a significant impact on ship prices (Stopford, 2009). They suggest that when these rates are high, a five-year-old ship typically values at four to six times its annual earnings. Other factors such as age and scrap prices are also considered. Corrosive cargoes or inadequate maintenance can shorten a ship's economic lifespan. If a ship's market value falls below its scrap value, it is typically sold for scrapping.

Given the recent trend towards green technology, the value of extra green onboard installations is also considered in this paper.

The dry bulk carrier sales data used in this article are obtained and then collated from various shipbrokers' public weekly reports published on the Hellenic Shipping News (2022) Website including Allied Shipbroking, Intermodal, Xclusiv and Advanced Shipping and Trading. In practice, a single transaction reported by several shipbrokers has a higher confidence level, while it is also possible that a record is only mentioned by one broker. For the sake of achieving low data bias, the transactions reported by more than two brokers are included in this paper. The period of the raw data ranged from 01/01/2014 to 31/12/2022, containing 3667 raw records.

The deadweight threshold selected for this paper is greater than or equal to 25,000 dwt, contrasting with some of the previous journals focussing only on specific subsectors. The rationale behind this all-size range selection is that in the seaborn market, these neighbour sub-sectors normally are substitute choices that could be compared by the chartering parties. In practice, Panamax bulk carrier is commonly chosen to carry iron ore when its freight rate becomes more attractive than Capesize and sometimes even the Handysize carriers can be used to ship iron ore.

A vessel's technical specifications include continuous variables such as deadweight, age at sale and categorical variables like country of build, scrubber installation, eco-design and ballast water treatment installation, all of which are extracted from the brokers' comments for this paper. A vessel's market-related indicators include the 1-year time charter rate, newbuilding price, scrap price and the 3-month London interbank offered rate (libor) at the time of sale. In this paper, for the sake of value consistency, the benchmark 1-year time charter rate and newbuilding price are collated based on the popular time-series from brokers, including: Very Large Ore Carrier (VLOC) (400 k ± 2.5%), Newcastlemax (208 k ± 2.5%), Capesize (170 k/180 k ± 2.5%), Kamsarmax (82 k ± 2.5%), Panamax (75 k ± 2.5%), Ultramax (58 k ± 2.5%) and Handy (32 k/38 k ± 2.5%).

Specifically, the scrap price is calculated by the product of light displacement tonnage and the steel scrap price ($/LDT). After collation and cleaning, the clean dataset was reduced from 3,667 to 3,643 records after removing the auction or bloc sales. Table 1 shows the descriptive statistics of the variables. Since the status of scrubber-fitted or eco-design tends to impact 1-year time charter rate under various market circumstances, together with the newbuilding price, they will be considered only as benchmark variables in the model fitting phase.

Data partitioning is an art of work that needs to consider the bias, variance and computing burden (James *et al.*, 2013). Large training datasets always come with a low bias, high variance and more computing time (Hastie *et al.*, 2009). Besides this fact, a more complex model could also lead to high variance and potential overfitting issues, which may generate big testing errors for a new dataset. In this paper, with 10 features and 3643 transaction records, the dataset can be classified as having a small input size. Therefore, following the convention in practice, it is randomly divided into 70% training data and 30% test data, as suggested by Afshin *et al.* (2018).

## 3. Method

This sector is structured into three parts. The first part covers the applied supervised machine learning methods, which include four tree-based learners and an add-on neural network, along with the final stacking process. The second part focuses on generalised additive models (GAMs) as the benchmark method. Lastly, the third part delves into model evaluation measurements.

### *3.1 Applied supervised machine learning methods*

*3.1.1 Random forest.* To overcome the overfitting issue faced by a single decision tree, a random forest is employed as a collection of multiple random decision trees, significantly

**Table 1.**
Descriptive statistics of the numerical and dummy variables

| | Variable name | Symbol | Minimum | Mean | Maximum |
|---|---|---|---|---|---|
| Numerical Variables | Price ($ Mn) | price | 1.50 | 13.62 | 61.00 |
| | 1 Year Time Charter Rate ($k/Day) Benchmark* | tc | 3.56 | 13.59 | 38.00 |
| | Newbuild Price ($ Mn) Benchmark | nb | 19.50 | 28.83 | 69.03 |
| | Scrap Price ($ Mn) Benchmark | scrap | 1.47 | 4.78 | 18.62 |
| | Deadweight (tonne) | dwt | 25,000 | 68,860 | 234,000 |
| | Age at sale | age | 0.00 | 11.31 | 30.00 |
| | 3 month US Dollar LIBOR | libor | 0.11 | 1.15 | 4.77 |

| | Variable name | Category | Symbol | Value | Percentage (%) |
|---|---|---|---|---|---|
| Dummy Variables | Country of build | Japan | country_jp | 1 | 58.9 |
| | | China | country_cn | 1 | 26.8 |
| | | South Korea | country_kr | 1 | 8.6 |
| | | Other | country_o | 1 | 5.7 |
| | Scrubber Installation | No | scrubber | 0 | 93.6 |
| | | Yes | | 1 | 6.3 |
| | Ballast water treatment installation | No | bwts | 0 | 80.4 |
| | | Yes | | 1 | 19.6 |
| | Eco | No | eco | 0 | 99.1 |
| | | Yes | | 1 | 0.9 |

**Note(s):** * Linear interpolation is applied based on the deadweight difference with neighbouring benchmarks for those not falling into the popular benchmarks
**Source(s):** Table by the author; data collected and collated from Hellenic Shipping News (2022)

reducing sensitivity to the training data. This is achieved by utilising bootstrapping (sampling with replacement) during sampling, ensuring that training datasets for each tree are less correlated and less sensitive to the original training data (James *et al.*, 2013). Random feature selection further reduces the correlation between the trees, enhancing the diversity. The finding of Breiman (2001) demonstrates that aggregating several trees produces more precise results when using a bootstrapped sample.

*3.1.2 Stochastic gradient boosting.* In Gradient Boosting (GBoost), trees are created sequentially, and the method is more robust to overfitting due to its learning rate $\eta$, i.e., scaling rate. The data is denoted as $\{(x_i, y_i)\}_{i=1}^{n}$ where $x_i$ refers to all the features of a single sample while $y_i$ refers to the value of the response variable and $n$ is the number of samples. In this paper, the loss function is defined as

$$L(y_i, F(x)) = \frac{1}{2}(y_i - F(x))^2 \tag{1}$$

where $F(x)$ is the predicted value. The first step is to initialise the model with a constant value $F_0(x)$ which minimises the value of the loss function through Equation (2).

$$F_0(x) = argmin_{\gamma} \sum_{i=1}^{n} L(y_i, F(x)) \tag{2}$$

where $F_0(x)$ is the initial value and it's typically equal to the average of the values of the response variable. The second step is to build the sequential decision trees using a loop that handles pseudo residuals $r_{im}$ shown in Equation (3), rather than the original values. The number of decision trees is denoted as $m$ and the model starts from the first tree, i.e., $m = 1$.

$$r_{im} = -\left[\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}\right] \text{for i} = 1 \ldots \text{n} \tag{3}$$

Like a normal regression decision tree, those pseudo residuals will fall into the terminal leave nodes $R_{j,m}$ where $j$ represents the index of leaf in a tree and $m$ is the index of the tree. For each terminal region, the optimal $\gamma$ needs to be found that could minimise the loss function through Equation (4). $F_{m-1}(x_i)$ is the previous predicted value.

$$\gamma_{jm} = argmin_{\gamma} \sum L(y_i, F_{m-1}(x_i) + \gamma) \; x_i \in R_{ij} \tag{4}$$

After optimal $\gamma_{jm}$ is obtained which is also a pseudo residual, the new predicted value could be updated by adding the previous predicted value $F_{m-1}(x_i)$ with the product of learning rate $\eta$ and $\gamma_{jm}$ as shown in Equation (5).

$$F_m(x) = F_{m-1}(x) + \eta \sum_{i=1}^{n} \gamma_{jm} \; x_i \in R_{ij} \tag{5}$$

When the first loop (from Equations (3)–(5)) is finished, the second loop will continue to progress as the above steps stated to create another tree using new pseudo residuals until the last tree is fitted and the $F_M(x)$ is the final predicted value for that sample.

*3.1.3 Extreme Gradient Boosting.* Extreme gradient boosting (XGBoost) was first introduced by Chen and Guestrin (2016) and like GB, the trees are also fitted in sequence based on the pseudo residual shown in Equation (3). The loss function for XGBoost combines the loss function for GBoost with the regularisation term $\frac{1}{2}\lambda O_{Value}^2$ as shown in Equation (6).

$$L(y_i, p_i) = \sum_{i=1}^{n} L(y_i, p_{i-1} + O_{Value}) + \gamma T + \frac{1}{2}\lambda O_{Value}^2 \tag{6}$$

The goal is to find the output value $O_{Value}$ that can minimise this loss function. The notations $p_i$ and $p_{i-1}$ represent the current and previous predicted values, respectively. The threshold $\gamma$ and number of leaves $T$ are used to encourage tree pruning and prevent overfitting. Moreover, $\lambda$ is included also as a regularisation factor, where a larger value results in a smaller output value to prevent overfitting.

In practice, the second order Taylor approximation shown in Equation (7) is used to substitute $L(y_i, p_{i-1} + O_{Value})$ to ease the calculation.

$$L(y_i, p_{i-1} + O_{Value}) \approx L(y_i, p_{i-1}) + \left[\frac{d}{dp_i} L(y_i, p_{i-1})\right] O_{Value} + \frac{1}{2}\left[\frac{d^2}{dp_i^2}(y_i, p_{i-1})\right] O_{Value}^2 \tag{7}$$

By adding and solving the loss function for each observation, the output value and the similarity score are calculated as Equations (8) and (9).

$$O_{Value} = -\frac{\left(\frac{d}{dp_1} L(y_1, p_1) + \frac{d}{dp_2} L(y_2, p_2) + \ldots + \frac{d}{dp_n} L(y_n, n)\right)}{\left(\frac{d^2}{dp_1^2}(y_1, p_1) + \frac{d^2}{dp_2^2}(y_2, p_2) + \ldots + \frac{d^2}{dp_n^2}(y_n, p_n) + \lambda\right)} \tag{8}$$

$$Similarity\ Score = \frac{\left(\frac{d}{dp_1} L(y_1, p_1) + \frac{d}{dp_2} L(y_2, p_2) + \ldots + \frac{d}{dp_n} L(y_n, n)\right)^2}{\left(\frac{d^2}{dp_1^2}(y_1, p_1) + \frac{d^2}{dp_2^2}(y_2, p_2) + \ldots + \frac{d^2}{dp_n^2}(y_n, p_n) + \lambda\right)} \tag{9}$$

The gain is calculated by subtracting the sum of the similarity score of two leaves and the previous node, which is used to find the best split threshold. Like the process in GB, once the first loop (from Equations (7)–(9)) concludes, the second loop iterates through the same steps to build another tree using new pseudo residuals, until the last tree is fitted. The final predicted value is represented by $F_M(x)$ shown in Equation (5).

*3.1.4 CatBoost.* Developed by Yandex in 2017, CatBoost is believed to be another tree-based member of the gradient boosting family, which is especially highly capable of handling categorical values. To reduce the chance of overfitting, a common issue in GB, Liudmila *et al.* (2019) suggests a weighted sampling method named minimal variance sampling (MVS) in Catboost for efficient feature splitting. In other words, for each iteration, the features are selected that can maximise each tree's accuracy, but the speed is slower than the XGBoost.

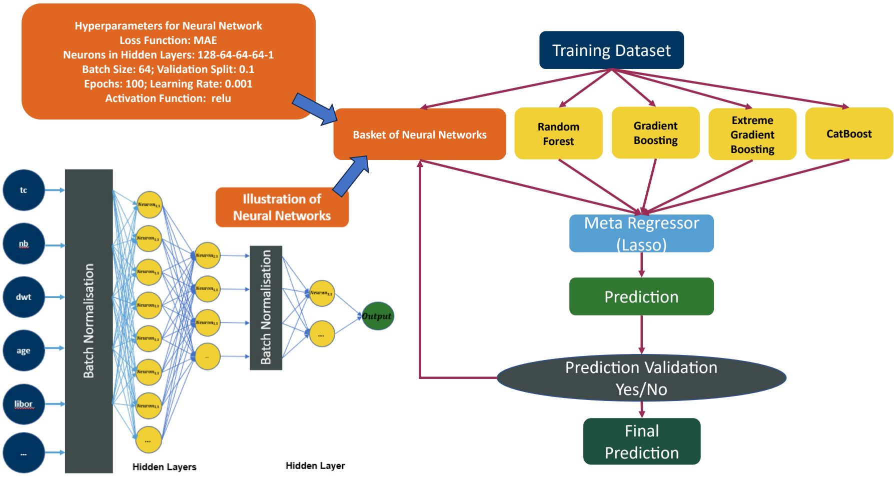
*3.1.5 Stacking with feedforward neural network.* The stacking algorithms typically leverage the strengths of the trained models, and select a meta-learner (Boehmke and Greenwell, 2019). The meta-learner is designed to combine the strength of each model and minimise the relative weakness (Reid and Grudić, 2009). To mitigate the issue of overfitting, the paper employs a regularised linear model – Lasso as the meta-learner (Li *et al.*, 2021).

As a versatile structure, feedforward neural network (FNN) can effectively fit data with even two hidden layers to approximate a continuous function (Paluzo-Hidalgo *et al.*, 2020), according to the universal approximation theorem (Hornik, 1991; Lewicki and Marino, 2003).

An advantage of FNN is the feature engineering through sequential layers rather than manually trying numerous combinations (Heaton, 2016). The add-on FNN in this paper is constructed by reducing half of the number of neurons for the next hidden layer and applying batch normalisation every two layers. Figure 1 represents the construction of the stacking workflow and the simple FNN.

### *3.2 Benchmark generalised additive models (GAMs)*

The GAM is chosen as the benchmark method in this paper because it can recognise non-linear relationships and consider interactions within variables through smooth functions (Müller, 2011), akin to feature selection on each node in tree methods and neuron handling various activation functions in neural networks. Moreover, it relaxes the prerequisite assumptions of the general linear model, such as the normality of residuals, constant variance of residuals and even the independence of observations.


Hyperparameters for Neural Network
Loss Function: MAE
Neurons in Hidden Layers: 128-64-64-64-1
Batch Size: 64; Validation Split: 0.1
Epochs: 100; Learning Rate: 0.001
Activation Function: relu

**Source(s):** Figure by the author

**Figure 1.**
Stacking workflow involving neural network models

In the previous literature, traditional statistical GAM has yielded good results given its flexible characteristics. Köhn (2009) finds that GAM effectively identify the non-linear relationship between ship price with predictors. A chemical tanker pricing model proposed by (Ådland and Köhn, 2018) suggests that GAM is a suitable function for capturing asset-specific factors. The GAM is structured as in Equation (10) where $\beta_0$ is the intercept and $n$ is the number of predictor variables. $f_n(x_n)$ represents the smooth function of $x_n$ and $g(E(y_i))$ is the link function that connects the expected value of the target variable $y_i$ to a linear combination of the non-linear functions of $x_n$.

$$g(E(y_i)) = \beta_0 + f_1(x_1) + f_2(x_2) \ldots + f_n(x_n) \tag{10}$$

### *3.3 Model evaluation*

In the first half of this section, mean squared error (MSE), mean absolute error (MAE) and $R^2$ are employed to measure the fitness of the models. The second half utilises Shapley additive explanations (SHAP) to evaluate the features importance.

$R^2$ represents the percentage of the variance in target variable that is explained by the features in the fitted model. Its value ranges from 0 to 1, with a higher value indicating a better fitness. MSE calculates the average of the squared differences between actual and predicted values, while MAE calculates the average of the absolute differences. Both MSE and MAE measure how far the models predicted values are from the actual ones. Those three measurements are shown in the following equations, where $y_i$ is the actual target value, $\overline{y_i}$ is the mean of actual target value and $\widehat{y_i}$ is the predicted target value.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{\sum_{i=1}^{n} \left(y_i - \overline{y_i}\right)^2} \tag{11}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \tag{12}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \widehat{y_i}| \tag{13}$$

SHAP provides convenience in breaking down the contribution of each feature to the prediction by considering all possible coalitions of features.

$$\varphi_m(v) = \frac{1}{p} \sum_{1}^{S} \frac{v(S \cup \{m\}) - v(S))}{\binom{p-1}{k(S)}}, m = 1, 2, , \ldots p \tag{14}$$

The SHAP value is calculated as in Equation (14), $v(S \cup \{m\})$ is the value after feature $m$ joins sub-set $S$, $k(S)$ is the size of $S$ and $p$ is the number of features.

## 4. Empirical analysis

### *4.1 Application of the models*

After fitting the models on the training dataset, the predictions are made using the given features and the performance is illustrated by the breakdown of the prediction errors shown in Figure 2. From the results, 47% of the stacking method test errors fall within the

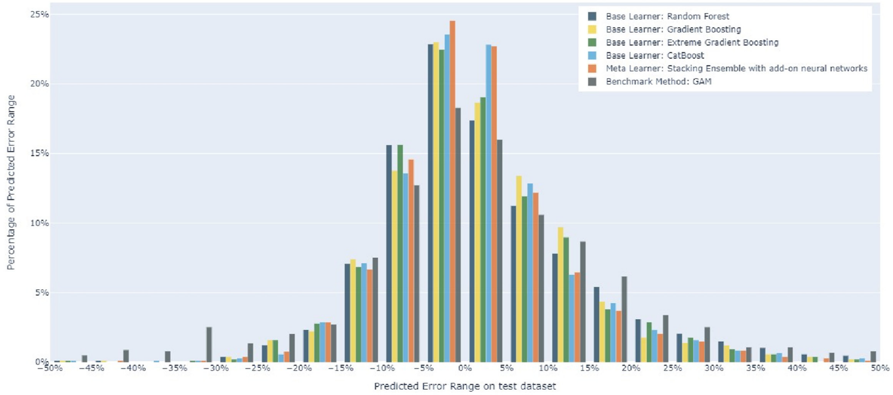


**Figure 2.**
Distribution of prediction errors on the test dataset

**Source(s):** Figure by the author

less-than-5% error band and 73% are within less-than-10%, demonstrating superior performance compared to other base learners and the benchmark GAM.

Figure 3 illustrates the cross-validation (k = 10) chart based on MAE, indicating that the model fitting improves with the inclusion of more data. Specifically, the dataset is split into 10 times, with 9 folds used as the training dataset and the remaining 1-fold used for testing. The MAE for all the models improves as more data is added. Therefore, the length of the training dataset can be extended in future applications under current hyperparameters.

### *4.2 Evaluation of the model's fitness*

In Table 2, performance metrics (MAE, MSE and $R^2$) are compared for different models using a 70–30% data split. The stacking method outperforms four base learners and a

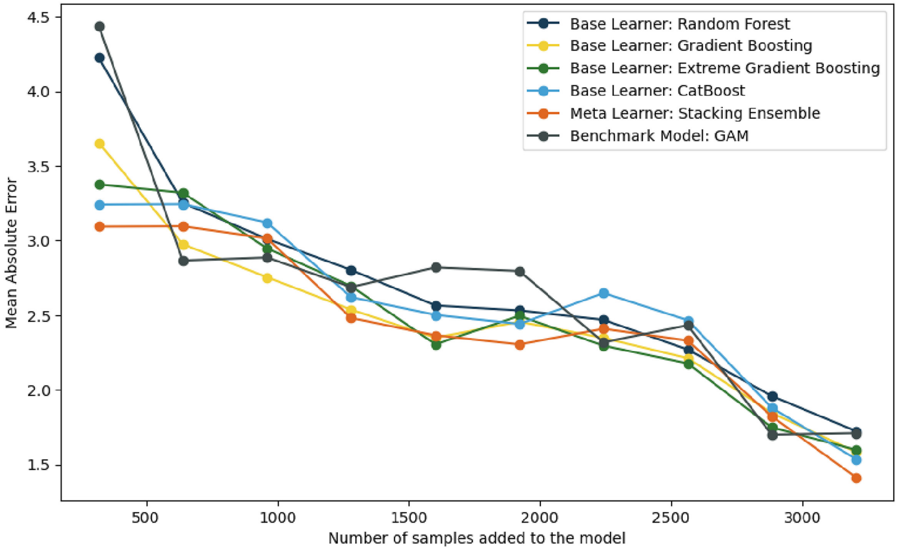


**Figure 3.**
10-k cross-validation charts from machine learning models

**Source(s):** Figure by the author

| | Training/test dataset: 70–30% | | | | | | Training/test dataset: 60–40% | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Base learners | | | Meta learner | Bench-mark | | Base learners | | | Meta learner | Bench-mark |
| | RF | GB | XGB | CatBoost | Stacking | GAM | RF | GB | XGB | CatBoost | Stacking | GAM |
| Training MAE | 0.70154 | 0.37991 | 0.41722 | 0.46168 | 0.43706 | 1.36310 | 0.70563 | 0.43931 | 0.45577 | 0.47990 | 0.45620 | 1.39216 |
| Test MAE | 1.24579 | 1.03017 | 1.07755 | 0.97881 | 0.96270 | 1.50149 | 1.25653 | 1.05379 | 1.08361 | 0.98311 | 0.97197 | 1.50892 |
| Training MSE | 0.92544 | 0.36585 | 0.30289 | 0.36945 | 0.34814 | 3.79013 | 0.93198 | 0.49329 | 0.35565 | 0.40144 | 0.37481 | 3.94446 |
| Test MSE | 3.63813 | 2.71538 | 3.03371 | 2.63263 | 2.54754 | 5.29204 | 3.72530 | 2.85965 | 3.20808 | 2.77902 | 2.64387 | 5.44161 |
| Training R Squared | 0.98597 | 0.99445 | 0.99541 | 0.99440 | 0.99472 | 0.94162 | 0.98584 | 0.99251 | 0.99460 | 0.99390 | 0.99431 | 0.94009 |
| Test R Squared | 0.94818 | 0.96165 | 0.95639 | 0.96215 | 0.96338 | 0.92392 | 0.94770 | 0.95889 | 0.95581 | 0.96076 | 0.96266 | 0.92315 |

| | Training/test dataset: 50–50% | | | | | |
|---|---|---|---|---|---|---|
| | | Base learners | | | Meta learner | Bench-mark |
| | RF | GB | XGB | CatBoost | Stacking | GAM |
| Training MAE | 0.70576 | 0.46162 | 0.46868 | 0.51 344 | 0.46760 | 1.40881 |
| Test MAE | 1.26267 | 1.06490 | 1.08805 | 0.99500 | 0.97412 | 1.53278 |
| Training MSE | 0.93900 | 0.51414 | 0.37927 | 0.45142 | 0.39451 | 4.06723 |
| Test MSE | 3.86655 | 3.15064 | 3.29724 | 3.11162 | 2.96726 | 5.88229 |
| Training R Squared | 0.98554 | 0.99208 | 0.99416 | 0.99305 | 0.99392 | 0.93833 |
| Test R Squared | 0.94739 | 0.95778 | 0.95470 | 0.95830 | 0.96023 | 0.92117 |

**Source(s):** Table by the author

**Table 2.**
Predicted errors of RF, GB, XGB, CatBoost, Stacking and GAM

benchmark model, showing the smallest prediction errors (MAE: 0.9627, MSE: 2.54754) and the highest $R^2$ (0.96338) on the test dataset. The 70–30% split method proves optimal among three splitting plans, indicating that the proposed model effectively learns from more data, potentially improving generalisation. Subsequent analysis will be based on this plan.

Furthermore, Figure 4 displays a comparison of the predicted values with the actual values on both the training and test datasets. Consistent with the MSE and MAE values in Table 2, GB and Extreme Gradient Boosting (XGB) all show two relatively thinner quasi-blue lines representing the predicted values on training dataset, contrasting with the dispersed red points on test dataset nearby. The Quantile-Quantile (QQ)-plot of training residuals from the GB even indicates a departure from a normal distribution. In practice, it is preferable to observe a close dispersion both in the training and test datasets (Boehmke and Greenwell, 2019). Fortunately, the ensemble construction provides the opportunity to incorporate base learners such as Random Forest (RF) and CatBoost that do not suffer from overfitting issues. The benchmark GAM displayed the highest variety under all three metrics.

### *4.3 Evaluation of the feature importance*

Figure 5 shows the feature importance using the stacking ensemble on the training dataset, where each dot represents a single sample. The horizontal axis displays the SHAP value, and the colour density illustrates whether the value of this feature is large or small.

From Figure 5, the long right tail of the newbuilding price suggests that a higher newbuilding price has more influence than the 1-year time charter rate. This relationship also holds for the age at sale; Both old and young ages can have the greatest effect on the vessel's pricing. The variables, such as newbuilding price benchmark, 1-year time charter rate benchmark, scrap price and deadweight, exhibit a higher positive contribution to the price, while the age at sale has a negative effect. This aligns with previous studies on practical valuation methods (Ådland and Köhn, 2018; Stopford, 2009), as well as Pruyn *et al.* (2011), who point out that the newbuilding price is also a significant predictor in pricing a vessel. For variables such as Japan-built, scrubber-installed and eco-design, which may have an

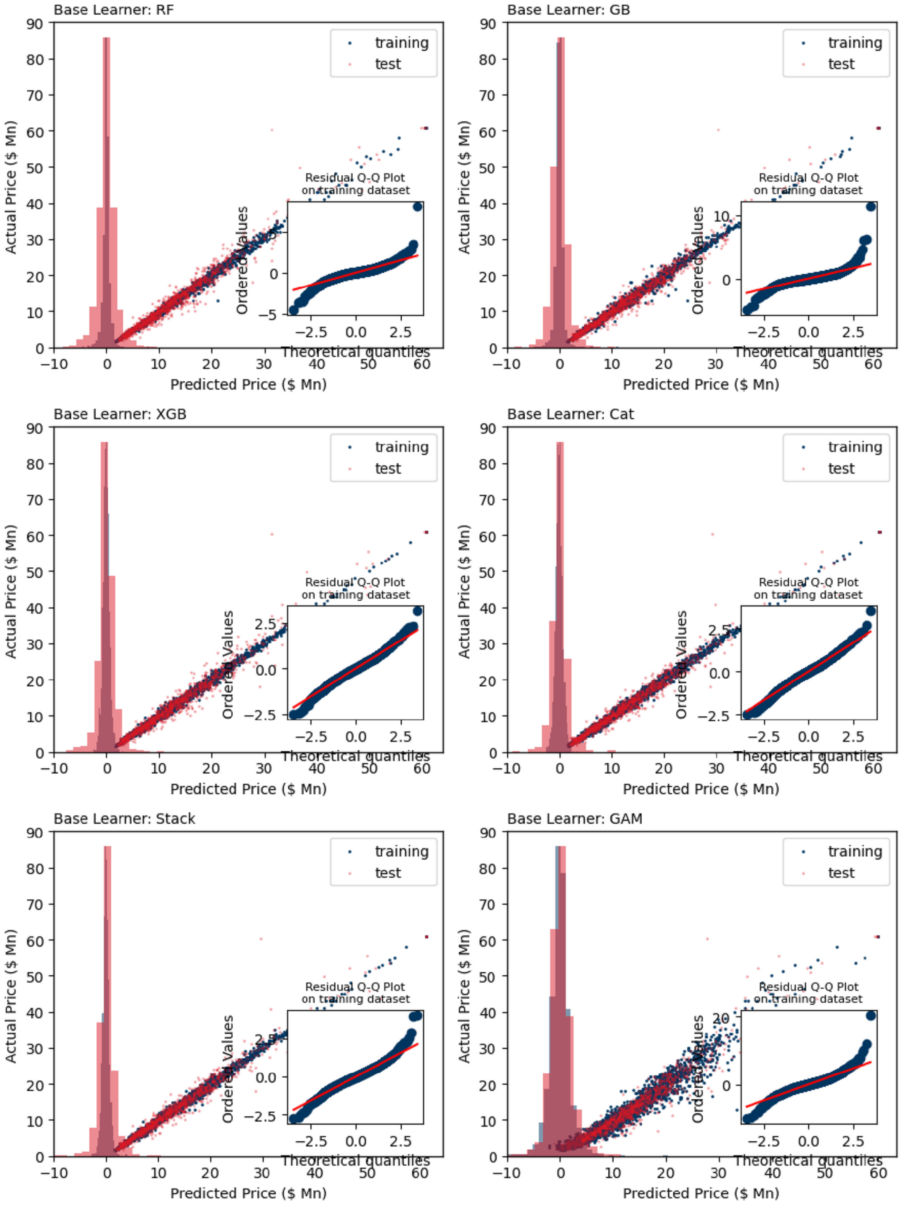
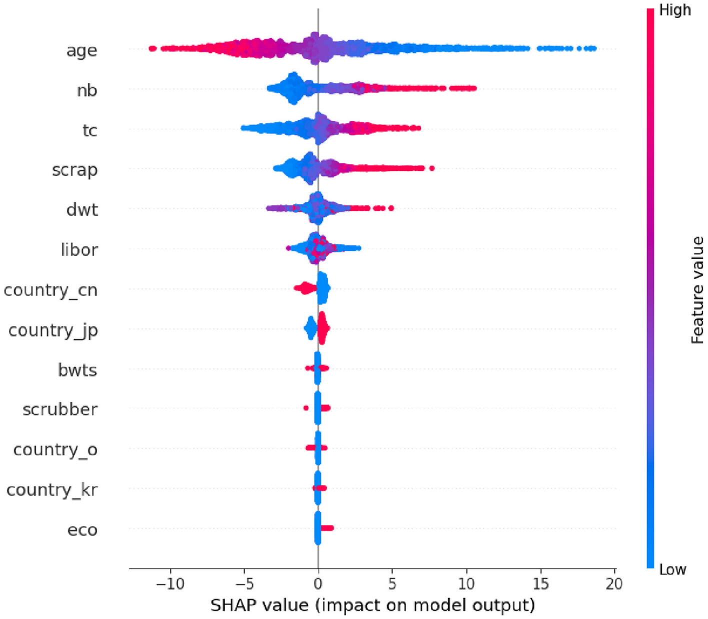


**Figure 4.**
Predicted versus actual values

**Source(s):** Figure by the author; Actual Price from Hellenic Shipping News (2022); Predicted Price from the applied model

incomplete data collection issue, they all demonstrate an intuitive result, typically commanding a positive premium compared to vessels built in China or other places. It is worth noting that deadweight provides a mixed contribution except the big values. Although it may initially seem difficult to interpret, this aspect actually highlights another advantage of



**Source(s):** Figure by the author

**Figure 5.** Feature importance based on SHAP of stacking ensemble

Machine Learning (ML) methods compared to traditional coefficient inference methods, which often rely solely on the final significance result table for judgement. However, the SHAP chart, displaying all values, can mitigate this drawback by providing a more comprehensive understanding and facilitating informed judgements.

As proposed by previous research (Ådland and Köhn, 2018; Köhn, 2009), the result of GAM is shown in Table 3. Likewise, the hyperparameter for the number of *splines* in GAM is set to 30 in this paper after various trials. The effective degree of freedom (EDoF) of 117.1114 indicates that there exists considerable flexibility. Similarly, the individual EDoFs reflect the

| | EDoF | *p*-value | Significance code | | |
|---|---|---|---|---|---|
| dwt | 17.4 | 1.11E−16 | *** | Effective DoF | 117.111 |
| tc | 20.5 | 1.11E−16 | *** | Log Likelihood | −6837.018 |
| nb | 18.7 | 1.11E−16 | *** | AIC | 13910.259 |
| scrap | 16.6 | 1.11E−16 | *** | AICc | 13921.834 |
| age | 19.8 | 1.11E−16 | *** | GCV | 5.804 |
| libor | 18.4 | 1.11E−16 | *** | Scale | 5.327 |
| scrubber | 1.0 | 5.20E−03 | ** | Pseudo R-Squared | 0.924 |
| bwts | 1.0 | 9.19E−01 | | | |
| eco | 1.0 | 3.71E−08 | *** | | |
| country_jp | 0.9 | 9.81E−01 | | | |
| country_kr | 0.9 | 1.60E−02 | * | | |
| country_cn | 0.9 | 1.11E−16 | *** | | |
| country_o | 0.00 | 5.76E−13 | *** | | |
| intercept | 0.00 | 1.11E−16 | *** | | |

**Note(s):** ‘***’: 0.001; ‘**’: 0.01; ‘*’: 0.05; ‘.’: 0.1; ‘ ’: 1
**Source(s):** Table by the author

**Table 3.** Result of benchmark GAM with smoothed terms

degree of non-linearity. Applying a 5% threshold for *p* value, all variables, except bwts and country_jp, are statistically significant, rejecting the null hypothesis of a linear relationship.

One of the big differences between the stacking ensemble and GAM lies in the model construction. The stacking ensemble can leverage the predictive power of the tree-driven base learners by adding a small amount through a series of trees, and the feature engineering power of the add-on neural network. To some extent, with the use of *splines*, GAM fits several polynomial functions to approximate the non-linear relationship of each independent variable with the target, rather than fitting a single polynomial function, necessitating more manual judgement.

Furthermore, unlike in a neural network, the addition of the optimal interaction terms in GAM normally requires trials and judgement. To ensure the robustness of the GAM and avoid high-degree multicollinearity, low-degree polynomials are typically employed. However, in tree-driven methods, through bootstrapping and aggregation, the same effect can be achieved more conveniently given the efforts spent in this paper.

However, higher variance is observed for vessels priced at more than $30 million in all methods. One possible reason is the limited data, making it challenging to train the models. It is also not easy to collect all ranges of data in terms of vessel specification and transaction details. Another limitation concerns the optimal selection of the supervised models used in this study. Since the primary objective of this study is not to identify the best way to choose base learners, it's satisfying to observe the meta-learner combining the strengths of the chosen models. This paper follows guidance on selecting diverse base learners (Wolpert, 1992), taking into account RF's robustness, GB's boosting capabilities, XGB's efficiency, and CatBoost's effective handling of categorical variables.

## 5. Conclusion

The approach of this paper examines whether the stacking ensemble method with add-on FNN models can improve the valuation of second-hand vessels in the dry bulk carrier market by contrasting the parameters inference methods used in previous studies. Additionally, the commonly used GAM method is evaluated as a benchmark.

The data ranging from 01/01/2014 to 31/12/2022 are split into 70% as training dataset for model fitting and 30% as test dataset for performance comparison. The designed stacking ensemble approach results in the best performance with a prediction accuracy of 74% on test dataset falling into the less-than-10% error range, with the lowest MAE and MSE of 0.9627 and 2.54754, contrasting 1.50149 and 5.29204 from benchmark GAM.

For traditional statistical models, multicollinearity and feature engineering typically requires trials and judgement in the addition of interactions and polynomials. In this paper, RF, GB, XGB and CatBoost are employed. These four diverse base models are trained independently from the perspective of the meta-model, considering their robustness, boosting capabilities, efficiency and good handling of categorical variables, respectively. They are expected to capture different characteristics from the same training dataset. Additionally, along with the feature engineering capability of the add-on neural network, they collectively serve as a convenient desktop valuation tool.

For the feature importance, the SHAP chart illustrates that the age at sale, newbuilding price and 1-year time charter rate are the three most important features from the stacking ensemble. One of the base learners, Catboost, can effectively handle dummy variables such as 'country_jp,' which is deemed insignificant by GAM. The adaptable stacking approach, as intended, offers the possibility of leveraging the strengths of individual methods, providing users with significant freedom to customise their base learners, particularly for those operating in rapidly evolving methodological areas of asset valuation.

There is also room for improvement in this analysis in future development. The first one comes from the data's completeness and reliability side, such as the vessel's actual condition at the time of sale; certain terms and conditions of the sale contract; whether the ship has an incoming special survey; whether there is a subsidy or long-term charter contract involved, etc. Data reliability is another area for improvement. For example, some records may misreport important installation information or other relevant terms in the contract, or they may contain typographical errors, necessitating careful handling of outlier removal. It is essential to ensure that the removal of outliers is assessed carefully to determine whether it enhances the validity of your analysis or introduce bias. Lastly the number of transactions is a limitation; the trading of second-hand vessels is not as frequent as the freight market, which will bring pressure on supervised machine learning methods in correctly capturing the pattern and may lead to unstable performance. Furthermore, this method can be applied to other shipping sectors like oil tankers and containerships since they also have a given size of transaction volume, but specialised vessels such as chemical tankers and gas carriers need much more qualitative determination of their values.

## References

Ådland, R. and Koekebakker, S. (2007), "Ship valuation using cross-sectional sales data: a multivariate non-parametric approach", *Maritime Economics and Logistics*, Vol. 9 No. 2, pp. 105-118, doi: 10.1057/palgrave.mel.9100174.

Ådland, R. and Köhn, S. (2018), "Semiparametric valuation of heterogeneous assets", in Mathew, J., Lim, C., Ma, L., Sands, D., Cholette, M. and Borghesani, P. (Eds), *Asset Intelligence through Integration and Interoperability and Contemporary Vibration Engineering Technologies. Lecture Notes in Mechanical Engineering*, Springer, Cham, pp. 23-30.

Ådland, R., Cariou, P. and Wolff, F.-C. (2018), "Does energy efficiency affect ship values in the second-hand market?", *Transportation Research Part A: Policy and Practice*, Vol. 111, pp. 347-359, doi: 10.1016/j.tra.2018.03.031.

Afshin, G., Vladik, K. and Olga, K. (2018), "Why 70/30 or 80/20 relation between training and testing sets: a pedagogical explanation", *Departmental Technical Reports (CS)*, Vol. 11 No. 2, pp. 105-111.

Beenstock, M. (1985), "A theory of ship prices", *Maritime Policy and Management*, Vol. 12 No. 3, pp. 215-225, doi: 10.1080/03088838500000028.

Boehmke, B. and Greenwell, B.M. (2019), *Hands-On Machine Learning with R*, Chapman and Hall/CRC, New York.

Breiman, L. (2001), "Random forests", *Machine Learning*, Vol. 45 No. 1, pp. 5-32, doi: 10.1023/a:1010933404324.

Charemza, W. and Gronicki, M. (1981), "An econometric model of world shipping and shipbuilding", *Maritime Policy and Management*, Vol. 8 No. 1, pp. 21-30, doi: 10.1080/03088838100000019.

Chen, T. and Guestrin, C. (2016), "XGBoost: a scalable tree boosting system", *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining - KDD '16*, pp. 785-794, doi: 10.1145/2939672.2939785.

Fan, L., Gu, B. and Yin, J. (2021), "Investment incentive analysis for second-hand vessels", *Transport Policy*, Vol. 106, pp. 215-225, doi: 10.1016/j.tranpol.2021.04.001.

Haralambides, H.E., Tsolakis, S.D. and Cridland, C. (2004), "Econometric modelling of newbuilding and secondhand ship prices", *Research in Transportation Economics*, Vol. 12 No. 1, pp. 65-105, doi: 10.1016/s0739-8859(04)12003-9.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

Heaton, J. (2016), “An empirical analysis of feature engineering for predictive modelling”, in *SoutheastCon 2016*, pp. 1-6.

Hellenic Shipping News Worldwide (2022), *Online Daily Newspaper on Hellenic and International Shipping 2022,* available at: Hellenicshippingnews.com

Hornik, K. (1991), “Approximation capabilities of multilayer feedforward networks”, *Neural Networks*, Vol. 4 No. 2, pp. 251-257, doi: 10.1016/0893-6080(91)90009-t.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), *An Introduction to Statistical Learning*, Springer, New York, NY.

Kalouptsidi, M. (2014), “Time to build and fluctuations in bulk shipping”, *American Economic Review*, Vol. 104 No. 2, pp. 564-608, doi: 10.1257/aer.104.2.564.

Köhn, S. (2009), *Generalized Additive Models in the Context of Shipping Economics*, Doctoral Thesis, University of Leicester.

Lewicki, G. and Marino, G. (2003), “Approximation by superpositions of a sigmoidal function”, *Zeitschrift für Analysis und ihre Anwendungen*, Vol. 22 No. 2, pp. 463-470, doi: 10.4171/zaa/1156.

Li, M., Yan, C. and Liu, W. (2021), “The network loan risk prediction model based on convolutional neural network and Stacking fusion model”, *Applied Soft Computing*, Vol. 113, 107961, doi: 10.1016/j.asoc.2021.107961.

Liudmila, P., Gusev, G., Vorobev, A., Dorogush, A.V. and Gulin, A. (2019), “CatBoost: unbiased boosting with categorical features”, *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montréal, Canada.

Mohammed, A. and Kora, R. (2023), “A comprehensive review on ensemble deep learning: opportunities and challenges”, *Journal of King Saud University - Computer and Information Sciences*, Vol. 35 No. 2, pp. 757-774, doi: 10.1016/j.jksuci.2023.01.014.

Müller, M. (2011), “Generalized linear models”, in Gentle, J.E., Härdle, W.K. and Mori, Y. (Eds), *Handbook of Computational Statistics: Concepts and Methods*, Springer Berlin, Heidelberg, pp. 681-709.

Paluzo-Hidalgo, E., Gonzalez-Diaz, R. and Gutiérrez-Naranjo, M.A. (2020), “Two-hidden-layer feed-forward networks are universal approximators: a constructive approach”, *Neural Networks*, Vol. 131, pp. 29-36, doi: 10.1016/j.neunet.2020.07.021.

Pruyn, J.F.J., van de Voorde, E. and Meersman, H. (2011), “Second hand vessel value estimation in maritime economics: a review of the past 20 years and the proposal of an elementary method”, *Maritime Economics and Logistics*, Vol. 13 No. 2, pp. 213-236, doi: 10.1057/mel.2011.6.

Reid, S. and Grudić, G.Z. (2009), “Regularized linear models in stacked generalization”, *International Workshop on Multiple Classifier Systems*, Berlin, Heidelberg, pp. 112-121, doi: 10.1007/978-3-642-02326-2_12.

Stopford, M. (2009), *Maritime Economics*, Routledge, London.

Wolpert, D.H. (1992), “Stacked generalization”, *Neural Networks*, Vol. 5 No. 2, pp. 241-259, doi: 10.1016/s0893-6080(05)80023-1.

Zhao, A.B. and Cheng, T. (2022), “Stock return prediction: stacking a variety of models”, *Journal of Empirical Finance*, Vol. 67, pp. 288-317, doi: 10.1016/j.jempfin.2022.04.001.

## Further reading

Ådland, R., Jia, H., Harvei, H.C.O. and Jørgensen, J. (2021), “Second-hand vessel valuation: an extreme gradient boosting approach”, *Maritime Policy and Management*, Vol. 50, pp. 1-18, doi: 10.1080/03088839.2021.1969601.

Clintworth, M., Lyridis, D. and Boulougouris, E. (2021), “Financial risk assessment in shipping: a holistic machine learning based methodology”, *Maritime Economics and Logistics*, Vol. 25 No. 1, pp. 90-121, doi: 10.1057/s41278-020-00183-2.

Dai, L., Hu, H. and Zhang, D. (2015), "An empirical analysis of freight rate and vessel price volatility transmission in global dry bulk shipping market", *Journal of Traffic and Transportation Engineering (English Edition)*, Vol. 2 No. 5, pp. 353-361, doi: 10.1016/j.jtte.2015.08.007.

George, K., Anna, M. and Xakousti-Afroditi, M. (2021), "Environmental regulation on the energy-intensive container ship sector: a restraint or opportunity?", *Marine Policy*, Vol. 125, 104278, doi: 10.1016/j.marpol.2020.104278.

Geurts, P., Ernst, D. and Wehenkel, L. (2006), "Extremely randomized trees", *Machine Learning*, Vol. 63 No. 1, pp. 3-42, doi: 10.1007/s10994-006-6226-1.

Hale, C. and Vanags, A. (1992), "The market for second-hand ships: some results on efficiency using cointegration", *Maritime Policy and Management*, Vol. 19 No. 1, pp. 31-39, doi: 10.1080/03088839200000003.

Kavussanos, M.G. and Alizadeh, A.H. (2002), "Efficient pricing of ships in the dry bulk sector of the shipping industry", *Maritime Policy and Management*, Vol. 29 No. 3, pp. 303-330, doi: 10.1080/03088830210132588.

Ke, L., Liu, Q., Ng, A.K.Y. and Shi, W. (2022), "Quantitative modelling of shipping freight rates: developments in the past 20 years", *Maritime Policy and Management*, pp. 1-19, doi: 10.1080/03088839.2022.2138595.

Kilian, L. (2009), "Not all oil price shocks are alike: disentangling demand and supply shocks in the crude oil market", *American Economic Review*, Vol. 99 No. 3, pp. 1053-1069, doi: 10.1257/aer.99.3.1053.

Lee, S. and Chung, J.Y. (2019), "The machine learning-based dropout early warning system for improving the performance of dropout prediction", *Applied Sciences*, Vol. 9 No. 15, p. 3093, doi: 10.3390/app9153093.

Li, Y. (2018), "Feature extraction and learning effect analysis for MOOCs users based on data nining", *International Journal of Emerging Technologies in Learning (iJET)*, Vol. 13 No. 10, p. 108, doi: 10.3991/ijet.v13i10.9456.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A.Y.M., Mousa Fardoun, H. and Ventura, S. (2015), "Early dropout prediction using data mining: a case study with high school students", *Expert Systems*, Vol. 33 No. 1, pp. 107-124, doi: 10.1111/exsy.12135.

Nordhausen, K. (2014), "An introduction to statistical learning-with applications in R by gareth james, daniela witten, trevor hastie and robert tibshirani", *International Statistical Review*, Vol. 82 No. 1, pp. 156-157, doi: 10.1111/insr.12051_19.

**Corresponding author**
Jingzhou Zhao can be contacted at: jingzhou.zhao@bayes.city.ac.uk