Norwegian School of Economics
Bergen, Spring 2021

# Vessel Valuation in Maritime Industry Using Machine Learning

Babak Ebrahimi and Arrold Jahjolli

Supervisor: Haiying Jia

Master thesis, Economics and Business Administration

Major: Financial Economics and Economics

NORWEGIAN SCHOOL OF ECONOMICS

# Acknowledgements

This thesis is written as the final part of our Master of Science in Economics and Business Administration degree, within our Majors in Financial Economics and Economics, at the Norwegian School of Economics - NHH.

We would like to thank our supervisor, Haiying Jia, for sharing her insights and giving us constructive feedback throughout the whole process. Her insights and interest in the field of maritime economics has been very valuable for our thesis.

Norwegian School of Economics

Bergen, December 2021

Babak Ebrahimi Arrold Jahjolli

# Abstract

This thesis focuses on the application of machine learning for vessel valuation. In the following paper, we present four different models and conclude that supervised machine learning models such as Catboost exhibit predictive prowess in estimating vessel prices. The CatBoost model is compared against a PLS/PCA model, Lasso and a traditional linear regression model. We find conclusive evidence that linear regression is not effective in predicting vessel prices.

Furthermore, CatBoost proves to be an ideal solution to vessel valuation due to its natural ability to encode categorical variables efficiently. The model found that the most important variables that affect price are age at sale, freight rates and one-year yield bond prices. The findings support previous research in this topic. Another reason why CatBoost can be very useful for vessel valuation is that the algorithm uses an extreme gradient boosting approach that makes it immune to multicollinearity between predictors. The results from CatBoost exhibit the lowest measurement errors and do not indicate any signs of overfitting.

The data used in this thesis is provided by the Clarkson World Fleet register. The data set contains more than 17,700 observations focused on five different vessel types: bulk, container, gas carrier, tanker, tanker chem. There are roughly 22 numerical variables and 20 categorical ones. Other macro variables such as interest rates, freight rates and exchange rates were added into the models to gauge the overall effect that macroeconomics has on vessel prices.

Machine learning algorithms facilitate accurate predictions by analyzing numerous independent variables and presenting the top factors that mostly affect the dependent variable (i.e. vessel price). Considering the efficiency and accuracy that machine learning algorithms like CatBoost offer, we suggest that CatBoost is extremely useful for asset valuation in the maritime industry.

**Keywords** – Vessel Valuation, Machine Learning, CatBoost, PCA, Lasso Regression

# Contents

# List of Figures

# List of Tables

# 1. Introduction

The shipping industry is the backbone of intercontinental trade and therefore one of the main contributing factors in today’s globalized markets. Seaborne transportation accounts for nearly of 90% of international trade (Iqbal, 2020). Vessel industry market is a multi-billion-dollar industry that is projected to steadily grow in the future as the demand for vessels is closely aligned with general economic growth.

As local economies experience economic growth, their import/export needs affect shipping demands which in turn affect vessel prices. Global marine vessel market is estimated to grow to 220 billion dollars by 2026 from 201 billion dollars in 2018 (Fortune Business Insight, 2018). As the vessel market grows, an all-encompassing accurate price model for ships would be extremely beneficial for the shipping industry. Investors, brokers, and shipping companies would be provided with better estimation figures for vessel which will lead to an increase in information symmetry for the vessel market.

The most prominent approach for vessel valuation has been using non- and semi-parametric models for the last decade. However, machine learning algorithms have been adopted considerably in the last years in various financial valuation models. Aubry et al. (2019) found that machine learning is particularly helpful for assets that are associated with high price uncertainty. It is not a mystery that vessel prices are highly volatile as Stopford (1988, p.383) argues that second-hand prices will respond sharply to changes in market conditions, and it is not uncommon for the prices paid to double, or halve, within a period of a few months. Considering the volatility of the vessel price markets and the many factors that affect such prices, we propose to introduce a CatBoost approach in forecasting ship prices in the shipping industry.

Our paper aims to present an accurate machine learning pricing model that is applicable to different vessel types in the shipping industry. The proposed model attempts to accurately predict vessel prices based on three important components such as: 1) vessel specific features (DWT, ship environmental impact, age of ship, engine propulsion, beam, SO scrubbers, GT, etc.), 2) macroeconomic factors (interest rates, forex exchange rates), and 3) other miscellaneous industry categorical variables that might affect vessel prices (country of buyer, country of seller, country of ship builder).

Extreme Gradient Boosting algorithms have experienced a rapid adoption rate in asset valuation models. Based on the available data, applying a CatBoost algorithm is believed to be suitable. The proposed algorithm is extremely efficient in tackling problems with categorical variables which is extremely useful for the purposes of this thesis.

Furthermore, we intend to compare the proposed CatBoost model against three other models. The models that CatBoost will be compared against are: 1) a traditional linear regression, 2) a Lasso regression model, and 3) a PCA model, respectively. Each model analyzes more than 40 different variables that according to economic theory affect vessel prices. Afterwards, we recommend whether machine learning is a viable approach in vessel valuation and draw conclusions on the relationships that the independent variables have on the dependent variable i.e. vessel price.

The rest of this paper is constructed as follows. First, we conduct a literature review on the research regarding vessel valuation throughout the last two decades. Second, we review the data used in this master thesis and explain any shortcoming in the data and/or any data estimations that we conducted due to missing data points. Further, we talk about our methodology and we analyze and compare our findings. Finally, we present the conclusion followed with the respective limitations and future recommendations about our study.

# 2. Literature Review

In this section, we will conduct a literature review where we will highlight the valuation methods used in previous research for vessels in the shipping industry. Throughout the last three decades of research in vessel valuation, the attention of researchers has switched from testing if the efficient market hypothesis holds to more dynamic nonlinear approaches. Adland and Strandenes (2006) reject the idea that EMH can explain the bulk freight market. By using kernel smoothing of the spot freight rates, they show that is possible to achieve significant profits from trading information which is the opposite of what EMH stands for. Pruyn et al. (2011) argue that the research on the efficient market hypothesis on vessel pricing is nonconclusive in explaining the pricing models in the shipping industry. They argue that using non- and semi-parametric models would provide a better estimation model for vessel prices.

Since the application of econometric models for vessel valuation has gained much popularity in the research community, our efforts will be focused on this particular stream of research. Kohn (2008) tackles the issue of vessel valuation in the shipping industry by introducing a semi-parametric estimation using the GAM framework. The results from his paper state that vessel pricing model is a nonlinear function of the main factors such as: age of the vessel, vessel size, vessel type and market conditions (i.e. charter rates and newbuilding prices). His research indicates that using a semi-parametric model for ship valuation tends to be more precise in estimating price compared to the previous valuation based on the EMH framework.

Adland and Kohn (2019) further prove that a multivariate semi-parametric valuation model for chemical tankers using a GAM framework outperforms linear estimation models. Hence, they conclude that the GAM is an adequate framework to model how prices are affected from vessel-specific factors. The GAM framework is an extension of a generalized linear model with a combination of linear predictors and the sum of smooth functions of explanatory variables (Wood, 2006). The GAM framework allows for flexible specifications of the dependence of the response variable on the covariates by specifying the model in terms of “smooth functions: rather than parametric relationships.

Gu et al. (2018) argues the importance of machine learning in the field of asset pricing since ML algorithms simplify the investigation of economic mechanisms necessary for valuation of assets. Machine learning algorithms have been rapidly adopted in asset pricing. However, ML application in vessel valuation is still somehow at early stages. To the best of our knowledge, there is only one paper that has attempted the application of a ML algorithm regarding second-hand vessel valuation.

Harvei and Jorgensen (2019) attempted in their paper to compare the precision of the established GAM framework versus a machine learning algorithm such as: Extreme Gradient Boosting approach (XGBoost). They concluded that the XGBoost algorithm improved desktop valuation accuracy by missing only 16.6% on out of sample predictions. In the paper, they argue that the application of machine learning was promising but it was highly encouraged that more complex algorithms were constructed in the future.

Hence, in this paper we will be conducting a CatBoost approach since this framework allows the inclusion of categorical variables in contrast of the Harvei and Jorgensen paper whose study included mostly quantitative variables. This thesis is a clear extension of the Harvei and Jorgensen paper where we expand the number of variables from 27 to 42 and our data covers different vessel types in contrast to only bulker vessel. A big emphasis is put onto including categorical variables that based on economic theory are believed to affect prices such as: builder country, seller country, manufacture country etc.

The proposed machine learning model in this paper is expected to analyze and show how different categorical and numerical variables each affect the overall vessel price. This study aims to further extend the pre-existing research that analyze the main factors that affect the value of ships in the maritime industry.

# 3. DATA

The database used in the following paper is extracted from the Clarkson World Fleet Registry. The data is extremely comprehensive, and it includes more than 17,700 observations about five different vessel types such as: bulk, container, gas carrier, tanker and tanker chem. The descriptive statistics of important vessel specific features grouped by their type are presented below in table 3.1.

The sale prices of the vessels range from January 1996 to December 2020 with a total of 16,058 transactions for all five of the aforementioned vessel types. In many instances, the quoted sale price for many vessels is omitted due to the fact that buyers acquire ships en bloc. There is a total of 1767 observations that have omitted sale prices. To circumvent the problem of the omitted quote prices, an average value is assigned to each individual vessel with the same en bloc number divided over the total en bloc acquisition price. Further data cleaning is conducted to make sure that the sale prices are all expressed in millions of US dollars.

We create an additional variable named “Age at Sale”. This variable is calculated by subtracting the sale date from the build date. According to economic theory, the age of the ship directly affects vessel price because fixed assets depreciate over time. The building year for the vessels ranges from 1962 until 2022 (i.e. ships that are sold but are yet to be operational). The deadweight “DWT” of the ships in our database ranges from 600 to 555,051 with an average DWT of 61,860.

Additionally, our dataset includes vessel specific numerical variables such as; GT, LOA, Beam, Draught, Speed, etc. The variables are specific for each ship and differences in these features is believed to affect the sale price as well. For an extensive overview of the descriptive statistic for each numerical variable, please refer to Appendix A1.

**Table 3.1**: Key Numerical Variables Breakdown By Vessel Type

| Variable | Minimum | Mean | Maximum | Missing Data |
|---|---|---|---|---|
| **PRICE (mln USD)** | | | | |
| *Bulk* | 0.5 | 15.99 | 1670 | 5.92% |
| *Container* | 0.5 | 16.51 | 171 | 16.79% |
| *Gas Carrier* | 0.3 | 29.57 | 390 | 27.51% |
| *Tanker* | 0.2 | 26.05 | 330 | 9.69% |
| *Tanker Chem* | 0.6 | 19.18 | 150 | 12.26% |
| **Age at Sale** | | | | |
| *Bulk* | -6.6 | 13.27 | 40.8 | 0.38% |
| *Container* | -4.4 | 11.06 | 35.8 | 7.81% |
| *Gas Carrier* | -3.1 | 13.16 | 44 | 13.03% |
| *Tanker* | -4.2 | 12.54 | 47.6 | 0.14% |
| *Tanker Chem* | -2.2 | 10.96 | 38.7 | 0% |
| **DWT** | | | | |
| *Bulk* | 10,055 | 60,019.82 | 403,508 | 0% |
| *Container* | 2,210 | 32,347.19 | 199,272 | 0% |
| *Gas Carrier* | 759 | 22,493.67 | 96,889 | 0.1% |
| *Tanker* | 600 | 107,852.76 | 555,051 | 0.06% |
| *Tanker Chem* | 976 | 35,598.6 | 105,715 | 0% |

To capture all variables that might affect sale prices, we have extended the list of our variables outside vessel specific features. We have added freight rates and two additional macroeconomic variables that might affect sale prices such as: 10-year bond yield and forex exchange rates presented in table 3.2.

Interest rate fluctuations are thought to affect the cost of financing in the shipping industry. As a result, fluctuations in interest rates might affect vessel prices and it would be interesting to measure the true effect of the rates to our proposed ship valuation model. Hence, we use as proxy the 1-year T-bill rate. The 1-year note is considered as the most appropriate benchmark to gauge the overall behavior of interest rates.

The same logic is applied for the inclusion of the forex exchange rates in the model. Considering that buyers and sellers operate in different countries, it means that fluctuations of exchange rates influence vessel prices. To account for such fluctuations, we have included a forex exchange proxy variable in our valuation model. We have extracted daily conversion rates from January 1996 to December 2020 where we match the forex rates with the date that the transaction took place. This would make possible that all the quoted ship prices are in millions USD which would make for a better comparison between prices.

The last macroeconomic variable that is added into the data is the freight rates. The freight rate is the quoted price that a shipping company receives for transporting goods from point A to B. The freight rates vary based on destination, cargo weight and the type of vessel used for transportation. Hence, different vessel types have slightly different rates. Generally, a higher a freight rate translates into a higher perceived value for that specific ship. In the data extracted from Clarkson, the average freight rate is $16,384 with a maximum value of over $98,000. These numbers depend heavily on the type of vessel being considered.

**Table 3.2**: Descriptive Statistics of Macroeconomic Variable

| **Variable** | **Minimum** | **Mean** | **Maximum** | **Missing Data** |
|---|---|---|---|---|
| **Freight Rate** (in Usd) | | | | |
| *Bulk* | $ 3,504 | $ 15,624.47 | $ 68,871 | 0% |
| *Container* | $ 5,520 | $ 10,826.64 | $ 25,749 | 38.38% |
| *Gas Carrier* | $ 10,498 | $ 21,009.32 | $ 59,533 | 0% |
| *Tanker* | $ 5,752 | $ 25,172.44 | $ 98,094 | 0% |
| *Tanker Chem* | $ 8000 | $ 17,762.70 | $ 48,000 | 37.18% |
| **Interest Rates** | 0.09% | 2.16% | 6.39% | 0% |
| **Forex Exchange Rate** | | | | |
| *USD/YEN* | 75.72 | 156.92 | 358.44 | 0% |
| *EUR/USD* | 0.83 | 1.2 | 1.6 | 0% |
| *USD/Yuan* | 1.53 | 6.22 | 8.74 | 0% |

The database used in the analysis suffered from many missing values for the TEU variable. Approximately 78.3% or 13900 data points are missing in our database. Since, the analysis in this thesis is partly focused on gas carriers, tankers and tanker chem vessels. We substitute a zero value to every vessel that qualifies in the aforementioned vessel types. The rest of the missing data points that fall under the bulk and container vessel types, we substitute these missing values with the median TEU for the respective vessel types. The rationale of using median values for  the missing TEU data points for the container vessel type is that there are only two observations missing. Hence, replacing these two data points with median values will have a miniscule effect on a database with more than 17,700 observations. The same process is done with the missing TEU values of the bulk vessel types.

Furthermore, the missing values for the *PumpCapacityCum* and *PumpsTotal* follow the same logic. We insert a zero value for all vessel type that do not have pumps. The rest of missing values are filled with the median values for each respective vessel type. Last, we insert average values for the missing values for the *MainBunkerCapacity, GrainCapacity, CapacityCUM, HoldTotalNo, HatchesTotalNo.*

The filling average values are filtered based on vessel types, GT, LOA, and several other vessel specific features. The rationale behind this approach is that when vessels share a lot of similar features when it comes to vessel type, size and volume; then it would be fair to assume that the missing values for a specific vessel feature would be close to an average number from other similar vessels in the database. Hence, it makes sense to take such averages and generate a proxy data point that is the mean of the capacity of all the other similar vessels.

Another important part of the dataset is the categorical portion of the data. There are more than 20 categorical variables that cover different vessel features. Below, we present the key categorical variables and their respective breakdown. The presence of more than 20 categorical variables somehow makes the model more complex than a model containing only numerical variable. However, in the next sections we will be arguing that adding categorical variables will indeed improve our accuracy and give a better picture on the variables that affect vessel price.

As presented in table 3.3, the ships in the database have been built from 44 different countries around the world. The most frequent builder country in database is Japan with a 44.7% frequency or 7938 ships built. The closest builder competitors are South Korea and China with a frequency of 18.3% and 16.0% respectively.

The buyer’s country for the observations in the database consists of 98 distinct nations. The most frequent buyer country is Greece with 4619 vessels bought or 26% of the transactions. Greece is followed by China and Germany with 11.4% and 4.9% respectively. The seller countries follow a similar distribution. Greece is still in the first place with 3421 vessels sold or a frequency of 19.3%. Followed closely by Japan and Germany with 14.6% and 10.2%

respectively. Of the 17,700 vessels recorded in the database, 87.2% or 15481 are currently in service and fully operational. Another 1651 vessels are laden.

**Table 3.3**: Frequency for Key Categorical Variables

| ***Type at Sale*** | ***Builder Country*** | ***Buyer Country*** | ***Seller Country*** | ***Speed Category*** |
|---|---|---|---|---|
| Bulk *53.1%* | Japan *44.7%* | Greece *26%* | Greece 19.3% | Service *87.2%* |
| Tanker *19.7%* | South Korea *18.3%* | Unknown *16.3%* | Japan *14.6%* | Laden *9.3%* |
| Container *16.2%* | China *16%* | China *11.4%* | Germany *10.2%* | Other *3.5%* |
| TankChem *5.5%* | Germany *4.5%* | Germany *4.9%* | China *5.4%* | |
| Gas Carrier *5.4%* | Other *16.5%* | Other *41.4%* | Other *50.5%* | |

Other categorical variables that will be part of this analysis include environmental features (i.e. EcoEI_engine, SOxScrubber, Energy saving Technology, Environmental summary), engine specific features (i.e. Propulsion Type, Gear Summary, Main Bunker Fuel type, Power type), and other additional variables like en bloc. Please refer to appendix A2 for a detailed overview of the categorical variables is presented above.

## 3.1 Data Cleaning

The data received from Clarkson underwent a lengthy data processing/cleaning. The dataset was in daily frequency and the first step was to convert the weekly freight rates to daily rates. We transformed the freight rates to daily rates by using the *reindex* function from Python. This function extends weekly data to daily by assigning the same weekly rate to each date that falls in that specific week. Data was further manipulated by creating a new column named *build date* by merging the month and year data from the provided dataset. This generated column would later be used to calculate *age_at_sale* variable that was mentioned in the data section above. Further formatting of the sale date was conducted to match the format of the *build date* column generated above by using the *panda* data frame.

Each quoted price was filtered to check for the currency used in the transaction. Data points that were in currencies different than USD were assigned the appropriate daily exchange rate to convert the sale price into American Dollars. The data was cleaned from punctuations like semi-colon/comma values and replaced with *N/A* values. We also created a dummy variable for en bloc vessels where 1 entails the vessel is a part of an en bloc sale and 0 entails the opposite. Other dummy variables were created during the data cleaning. For example, variables like *Energy Saving Technologies ES Summary*, *SOx Scrubber 1 Retrofit Date*, and *Eco Electronic engine* are dummy variables that take a value of 1 if they possess these features and a value of 0 if they do not. Panda data frame *pd.get_dummies* in Python was used for creating these variables.

Furthermore, many other variables were eliminated from the analysis due to irrelevancy. Variables like "Seller Company", "CVN", "Vessel Name" do not seem important in having an effect in vessel prices. These categorical variables were discarded to preserve the accuracy of our models.

Variables with missing data points were imputed by the most frequent value in the column based on the vessel type. The imputed variables were later renamed as their *variable name_Imputed* (e.g. *GearSummary_Imputed*). Categorical data that had missing values were replaced with *Unknown* so after encoding using the *JamesSteinEncoder* these missing data points would have the same numerical value. The *JamesSteinEncoder* proved to be the most accurate encoder among five different encoding approaches by generating a score of 0.63.

# 4. Methodology

In this section, an analysis is conducted on the four prediction models used throughout this paper (i.e. Linear regression, Lasso regression, PCA/PLS and Catboost). The proposed prediction models both have their own advantages and disadvantages when applied in the dataset from section 3. The validation of each model is done by applying a k-cross validation approach. As will be discussed below, the aforementioned technique provides the best alternative when it comes to managing the bias-variance tradeoff of the end results. The end goal of this analysis is to provide a model that exhibits a satisfying degree of accuracy while limiting the overfitting effects of the model.

## 4.1 Linear Regression

Linear regression attempts to model a relationship between two or more variables by fitting a linear equation to the observed data. Generally, the dependent variable is regressed against one or more independent variables and a linear equation is constructed as follows for n predictors:

$$y_i = \beta_1 \, x_{i1} + \beta_2 \, x_{i2} + \cdots + \beta_p \, x_{ip} + \varepsilon_i \,,$$

Where the value of β minimizes the sum of square residuals (Kutner et al., 2005).

$$S(b) = \sum_{i=1}^{n} (y_i - x_i^{\mathrm{T}} b)^2 = (y - Xb)^{\mathrm{T}} (y - Xb),$$

In the next section, several linear regressions will be regressed stepwise against different independent variables (i.e. age, age^2, DWT, DWT*age, and other macro variables) in order to analyze the effect that each variable has on the dependent variable vessel price.

## 4.2 LASSO Regression

Lasso is a regression method for linear models that uses shrinkage. This regression is sort of “the middle ground” of the subset selection and ridge regression. Similar to subset selection, Lasso performs variable selection by eliminating irrelevant variables. Moreover, like to ridge regression, Lasso uses a shrinkage penalty $\lambda$ that controls the relative impact on the regression estimates (James et al 2013).

This regression is suitable for models showing high levels of multicollinearity. Considering the variety of the independent variables in the database, the lasso regression is an appropriate technique to utilize to eliminate parameters that are irrelevant for the response variable. The 'lasso' minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant, because of the nature of this constraint, it tends to produce some coefficients that are exactly 0 and hence gives interpretable models (Tibshirani 1996). The regression is formulated like the following:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (4.1)$$

Where:

$X_i$ *are the explanatory variables.*

$Y_i$ *is the dependent variable.*

$\beta$ *is the coefficient* estimate.

$\lambda$ *is the amount of* shrinkage.

$\lambda=0$ *would mean that all features are important and considered in the model and it's the same as the with a classic OLS regression where the residual sum of squares is considered to build a predictive model.*

$\lambda=\infty$ *implies no feature is considered, as* $\lambda$ *moves closer to infinity the more features are eliminated or given a zero score*

*The bias increases with increase in* $\lambda$

*Variance increases with decrease in* $\lambda$*.*

Lasso estimator incorporates into the least-squares optimization a penalty function $L_1$ which leads to producing coefficients estimates with a value of zero for independent variables that do not have or have little effect on the dependent variable.

Despite the fact that Lasso estimator is appealing for prediction purposes, it is well-documented that such estimates are bias especially for the non-zero coefficients due to regularization (Feng et al. 2020). Even though bias is present in the Lasso regression, the regression method is stable in generating accurate prediction models. Lasso enjoys favorable properties of both subset selection and ridge regression; it produces interpretable models like the subset selection and exhibits stability of ridge regression (Tibshirani 1996).

## 4.3 CatBoost Model

Categorical Boosting (CatBoost) is a supervised machine learning algorithm that was developed in 2017 by a Russian technology company called Yandex. CatBoost uses a gradient boosting decision tree (GBDT) implementation. The main advantage of CatBoost compared to its counterparts is how effectively and effortlessly the algorithm processes categorical data. For many years, gradient boosting has remained the primary method for learning problems with heterogeneous features, noisy data, and complex dependencies (Prokhorenkova et. al 2019).

The gradient boosted is a minimization non-parametric function where a set of [$x_i$,$y_i$] and expected output values $y_i$ construct a collection of functions $F^0$, $F^1$,…,$F^t$,…., $F^m$, given a loss function L($y_i$,$F^t$). After $F^t$ functions are constructed, the algorithm can improve estimates of $y_i$, by finding another function $F^{t+1}= F^t + h^{t+1}$ that minimizes the expected value of the loss function (Friedman 1999). The CatBoost algorithm makes some minor approximations to the Friedman’s GBDT technique by adding a (1/n) term to the loss function to obtain an estimate for $h^{t+1}$ like the following:

$$h^{t+1} \approx \underset{h \in H}{\operatorname{argmin}} \frac{1}{n} \left( \frac{\partial \mathcal{L} y}{\partial F^t} - h \right)^2 . \qquad (4.2)$$

Compared to other GBDT algorithms (i.e. XGBoost,LightBoost), CatBoost brings two innovations: Ordered Target Statistics and Ordered Boosting (Hanckock et. al 2020). CatBoost deals in two distinct ways for the categorical data depending on the type. For low cardinality categorical data, the algorithm uses one-hot encoding. Ordered Target statistics is the technique that the algorithm uses for encoding high cardinality categorical variables. A target statistic is a value we calculate from the ground truth output values associated with particular values of a categorical attribute in a dataset (Mici-Barreca, 2001). There is an extensive research area behind the reasoning why encoding categorical variables this way is superior to one-hot encoding. But this goes beyond the scope of this thesis.

However, it is important to emphasize that the algorithm is designed to minimize the expected value of the loss function L, and thus avoiding "target leakage"[1] which makes CatBoost outperform other supervised ML algorithm when tested in highly cardinal categorical datasets (Prokhorenkova et. al 2019).

The second important feature of the algorithm is that CatBoost uses Oblivious Decision Trees (ODT's) in making Decision Trees. ODTs are balanced, less prone to overfitting, and allow speeding up execution at testing time significantly (Prokhorenkova et. al 2019). Since ODTs are full binary trees, the number of comparisons to reach a leaf node is the minimum number of comparisons to reach the maximum number of leaf nodes, which may yield more efficient executions than deeper Decision Trees that are not completely filled (Hanckock et. al 2020).

Considering the nature of data in section 3, CatBoost appears to be a solid methodological option to fit a valuation model for the aforementioned dataset. The algorithm is an implementation of Gradient Boosted Decision Trees that avoids conditional shift (e.g. Ordered Target Statistics) and prediction shift (e.g. Ordered Boosting). CatBoost's use of Ordered Target Statistics and Ordered Boosting make it a good choice for datasets with categorical variables that are sparse, or infrequently occur with specific target values, since these techniques guarantee that, given some unusual training examples, CatBoost will involve other examples to update its estimate for the unusual example systematically (Hancock, 2020).

The Catboost model as an inherently supervised learning approach is designed to train algorithms into classifying data or predicting outcomes accurately by using labeled inputs and outputs, the model can measure its accuracy and learn over time. This learning algorithm is designed for classifying and predicting new data accurately. The accuracy obtained from these models tends to increase as the training data becomes bigger since the algorithm learns itself

[1] Target leakage is defined in terms of conditional shift. Target leakage occurs when the expected encoded value $x^i_k$ is shifted under the condition $y_k = v$. This is an overftting condition in the sense that in the fitting process the model can exploit the correlation between $x^k$ and $y_k$ during training, but the correlation will not exist during testing due to the difference in expected values (Prokhorenkova 2019).

as more data is fed into it. Supervised learning is ideal for tackling problems like classification, plant control, and prediction models (Sathya 2020).

The fact that the data being analyzed is extensive and filled with categorical variables, would make Catboost an ideal approach to follow in this thesis to generate an accurate prediction model. The overarching goal of supervised learning algorithms is to make accurate predictions rather than inference, a quality that makes such algorithms effective for generating a valuation model for the purposes of this thesis.

## 4.4 PLS Regression Using the PCA Dimension Reduction

The data matrix from the database being analyzed for the purposes of this thesis consists of 17,700 rows of observations and 42 columns of variables. For convenience, the rows with observations are denoted as "n" and the columns as "p". This data matrix **X** has an (n x p) dimension.

Principal Component Analysis (PCA) is a technique for deriving low-dimensional features from a large set of variables. This approach is heavily used in unsupervised machine learning. However, the dimension reduction feature of PCA allows this technique to be regression friendly and thus allow the opportunity for model prediction. Principal component analysis is different from other regression technique that either focus on subset selection or coefficient shrinkage. Instead of using the original predictors $X_1$, $X_2$…..$X_p$ like the Lasso, PCA transforms the predictors into principal components $Z_1$,$Z_2$……$Z_m$ where M<p (James et. al 2013).

The principal component is as followed:

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j \qquad (4.3)$$

For constants where $\phi_{1m}$,$\phi_{2m}$,……$\phi_{pm}$ and m=1,…,M then the linear regression model is fitted (James et. al 2013):

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \quad i = 1, \ldots, n, \tag{4.4}$$

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}. \tag{4.5}$$

When regression coefficients from (4.4) are fewer than the original predictors $X_p$, then there is dimension reduction which mostly outperform least squares regression (James et. al 2013). Dimension reduction constrains the estimated $\beta_j$ coefficients and possibly bias the coefficient estimates. In cases where p is large relative to n then selecting $M<<p$ can significantly reduce variance of fitted coefficients (James et. al 2013). From (4.3), if $M=p$, this means that there is no dimension reduction, and the results would be the same as taking the least squares regression to the original independent variables.

The first part of this sub-section covered the predictor transformation from $\mathbf{X}_p$ to $\mathbf{Z}_\mathbf{m}$. After transforming the predictors, there two possible approaches in selecting the $Z_{1,}Z_2 \ldots Z_m$ that fit the model accurately: 1) principal components regression (PCR) and 2) partial least squares (PLS).

However, this thesis will focus on the partial least squares approach because the PLS approach attempts to find direction that help explain both response variable and predictors (James et al 2013). When computing $Z_1$ from (4.3), PLS places the highest weight on variables that are most strongly related to the response.

Principal Component Analysis (PCA) falls under the unsupervised learning algorithm category because such an algorithm analyses unlabeled input data. PCA like many other unsupervised algorithms searches for patterns between data without human intervention.

The lack of human intervention refers to the ability to learn and organize information without providing an error signal to evaluate the potential solution; the lack of direction for the learning algorithm in unsupervised learning can sometime be advantageous, since it lets the algorithm to look back for patterns that have not been previously considered (Sathya 2020).

The strengths of these kinds of algorithms are in tackling problems like clustering, association, and dimension reduction. Even though unsupervised learning algorithms are more competent in inference rather than prediction, including the algorithm would be beneficial for the purposes of this analysis because the results may show potential relationships between variables that have not been considered before in maritime economic theory. Another reason why including PCA is a viable option is that the amount of data being used is extremely extensive which will help in increasing the overall accuracy of the model.

## 4.5 K-Fold Cross-Validation

To validate the accuracy of the proposed valuation models, the data is divided into 80% of training set and 20% of testing data. Gholamy et al. (2018) show that empirical studies obtain the best results free of overfitting when 80% of the sample is used as a training set and 20% is allocated for testing the model. Proposing a different allocation ratio between training and test data would open the possibility of generating a model extremely specific for the database from section 3 and thus have no real accuracy when applied into a different set of data (i.e. overfitting).

The training of the different models is done on the training set using k-fold cross-validation when k=5. This method divides the training set into k number of subsets, or folds, of almost equal non-overlapping size (James et al., 2013). In this analysis, 80% of the k folds (i.e. 5-fold cross validation) consist of the training set and 20% of the k folds to the testing set. After division, one-fold is held out as a test set and the models are fitted on the training set folds. This process is repeated five times where each time a different subset is used as a test set. As a result, five different error metrics are generated for each iteration and an average error is calculated from the five error indicators calculated for each iteration.

Another viable alternative of k-fold Cross-Validation is Leave-one-out Cross-Validation (LOOCV). The LOOCV method takes one observation from the dataset say ($x_1$,$y_1$) as a test set and uses the rest of the data as a training set in order to fit the model. This process is done n-times where each observation is used one time as a testing set. This technique provides unbiased estimates since each training set contains n-1 observations (James et al. 2013). However, there is a bias-variance trade off that makes LOOCV not so viable for validation techniques. LOOCV always has higher variance than the k-fold CV because LOOCV averages the output of n-fitted models where each model is trained on an identical training set, and therefore are highly correlated with each other.

Empirical studies show that k-fold CV where k=5 or 10 do not exhibit any excessive level of bias or high variance (James et al. 2013). Another reason why LOOCV is inferior to k-fold CV is that inherently LOOCV requires excessive amounts of computational power since this technique has n-iterations that need to be performed. As discussed in section 3, the database used in this thesis is extremely big and thus it would be challenging in applying a LOOCV technique. As a result, a five-fold cross-validation is deemed suitable for the purposes of this thesis.

## 4.6 Model Evaluation

There are many error metrics in the literature that help give a better understanding about the accuracy of a fitted model in the out of sample data. Even though each error metric has its own strengths and drawbacks, their power can be fully harnessed if they are compared all at once together rather than one standalone metric.

For the purposes of this thesis, the attention will be focused on both scale dependent accuracy measures and scale-independent accuracy measures. The intended metrics used in this thesis are $R^2$, MSE/RMSE, and Mape.

The first metric will be $R^2$. This measure indicates the percentage of the variance in the dependent variable that is explained from the independent variables collectively. $R^2$ can take values from 0 to 1. Generally, a higher r-squared indicates a better fit for the model and a low r-squared can indicate low predictive power. However, this is not always the case which is why it is extremely important that other metrics are considered in the accuracy measures. However, the adjusted $R^2$ is only a good measure for how well the model ft the training data, not out-of-sample predictive power (James et al., 2013).

Scale dependent metrics are calculated using various transformations of the forecast errors $e_t$ (i.e. $e_t$=y^hat$_t$ - $y_t$). Scale dependent measures have measurement units depending on the original data, hence it these measures should not be compared with error measures from other data since it wont be accurate ( Bodea et al. 2014). The most frequently cited scale-dependent metrics typically require the transformation of the forecast error using absolute value and squaring and are formally expressed as (Bodea et. 2014):

$$\text{Mean Squared Error}\left(MSE\right)=\frac{1}{n}\cdot\sum_{t=1}^{n}e_t^2$$

$$\text{Root Mean Squared Error}\left(RMSE\right)=\sqrt{MSE},$$

(4.6)

where $e_t$= $y_t$ - y^hat$_t$ , $A_t$= is the abosolute error $e_t$ and n is the number of observations. The RMSE will be calculated both for the training and test set and is extremely convenient of testing accuracy.

The last metric used for model accuracy will be MAPE. The advantage of MAPE is that it is scale-independent hence this metric can be compared to other metrics belonging to other data sets. One of the criticisms that MAPE has often received lies in the percentage error being undefined and/or infinite when Yt is zero or numerically unstable when observations Yt approach zero (Bodea et al. 2014). The formula for the aforementioned metric is as followed:

$$\text{Mean Absolute Percentage Error}\left(MAPE\right)=\frac{1}{n}\cdot\sum_{t=1}^{n}\left(100\cdot\left|\frac{e_t}{Y_t}\right|\right).\qquad(4.7)$$

# 5. Empirical Analysis

In this section, the methodology described in section 4 will be conducted on the dataset from section 3 and thus generating different pricing models by following four different approaches (i.e. linear regression, LASSO, Catboost, PCA/PLS). For each model, measurement error metrics for the training and test set will be provided in order to compare for accuracy. For the predictive models to be robust, multicollinearity issues should not persist after applying the techniques from section 4. Refer to Appendix A8 for a correlation heatmap of the variables with each other. Initially, multicollinearity does not pose a problem for predictive power, but it makes interpretation of variable effects more complicated (Paul, 2006). However, it should be pointed out that the proposed Lasso, CatBoost, and PCA model respectively handle multicollinearity effectively and the results generated from each model is not affected by such issue.

The proposed LASSO regression is expected to generate estimates that are not affected by multicollinearity. Schreiber-Gregory et al. (2018) argues that statistical theory and machine learning have made great strides in creating regularization techniques that are designed to help generalize models with highly complex relationships (such as multicollinearity). In its most simplistic form, regularization adds a penalty to model parameters, so the model generalizes the data instead of overfitting (a side effect of multicollinearity).

CatBoost as a gradient boosting decision tree technique, provides multiple regularization parameters to help reduce model complexity and guard against overfitting (Boehmke, 2020). For example, if there are 2 features that are 99% correlated when algorithm is splitting the tree it will chose only one of them. Furthermore, this algorithm applies regularization penalty to each coefficient making the results generated from such model robust and free from any multicollinearity issue. Refer to appendix A3 for a correlation heatmap of the CatBoost model.

Similar to the other models, the PCA model also proves to be very effective in eliminating the issue of multicollinearity in large data with many predictors. Principal component analysis is one of these measures (for eliminating multicollinearity) and uses the manipulation and analyzation of data matrices to reduce covariate dimensions, while maximizing the amount of variation (Perez, 2017). The PCA generates principal components where the first component contains the most variability in the data. The same process is repeated in generating the second component that contains the second most amount of variability in the data but with the condition that the first and second component matrices are orthogonal (i.e. the product of the matrices is 0). Refer to appendix A4 for a correlation heatmap for all the PCA variables showing that all the selected variables are completely not correlated with each other.

## 5.1 Fitting the Models

Models are generated by using the programming language Python and R (for linear regression). The respective python packages are loaded into the software to run the regression for each regression technique. Price of Vessel is used as the dependent or response variable. For the predictors, a total of 44 variables are loaded into Python where 20 of these variables are categorical and 24 are numerical respectively. Regarding the CatBoost algorithm, the data are fed into the code without being encoded since this particular technique does an outstanding job handling categorical variables. The categorical variables in the LASSO and PCA algorithm has been encoded before being used in the regression. The nature of these two techniques, make it necessary to encode the categorical variables in order to generate any regression output.

All the models proposed in section 4 will be tested for their accuracy by splitting the dataset into 80% training set and 20% test set. The intended method to be used is k-fold cross validation in order to gauge each model's accuracy. In modern research, the go-to levels for k are 5 or 10 times. However, James et al. (2013) argues that a k=5 is a sensible trade-off between prediction error bias and variance.

The first regression model is presented in table 5.1. As discussed in section 4.1, we run a step-by-step linear regression where one additional independent variable is introduced in each regression step. There are a total of six linear step-by-step models where independent variables are regressed as follows: age, age^2, DWT, age*Dwt, loa*Beam*Draught, freight rate, interest rate.

**Table 5.1** Step-by-step Linear Regression and Lasso coefficients

| Dependent variable: Vessel price | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 | Lasso |
| AgeatSale | -0.581*** | -0.587*** | -0.517*** | -0.520*** | -0.508*** | -0.509*** | -0.552*** | -0.435 |
| | t = -84.958 | t = -84.012 | t = -82.844 | t = -84.082 | t = -80.682 | t = -80.907 | t = -90.529 | |
| AgeatSale^2 | | -0.029*** | -0.020*** | -0.031*** | -0.028*** | -0.029*** | -0.048*** | |
| | | t = -4.899 | t = -3.689 | t = -5.551 | t = -5.213 | t = -5.452 | t = -8.900 | |
| DWT | | | 0.424*** | 0.419*** | 0.166*** | 0.163*** | -0.05 | 0.161 |
| | | | t = 72.890 | t = 73.027 | t = 3.860 | t = 3.817 | t = -0.843 | |
| AgeatSale*DWT | | | | -0.057*** | -0.063*** | -0.067*** | -0.061*** | |
| | | | | t = -7.965 | t = -8.854 | t = -9.338 | t = -8.633 | |
| LOA*Beam*Draught | | | | | -0.016*** | -0.016*** | -0.011*** | |
| | | | | | t = -4.981 | t = -5.162 | t = -3.554 | |
| Oneyearyeild | | | | | | 0.061*** | 0.052*** | 0.05 |
| | | | | | | t = 11.526 | t = 10.248 | |
| Freightrate | | | | | | | 0.238*** | 0.241 |
| | | | | | | | t = 35.460 | |

The coefficient estimates for models 1 through 7 are all statistically significant at the 99% confidence interval. However, upon running a multicollinearity test (VIF), we find that there exists multicollinearity between the independent variables and therefore making these estimates bias. James et al. (2013) argues that in a model sometimes the predictive power is more important than inference thus making this procedure of handling multicollinearity repetitive. However, James et al. (2013) argues that including variables with degree of multicollinearity can cause problems when the initial predictive power of the variables is unknown and lead to a model with weak assumptions. The age and price is negatively correlated in every model. It should be mentioned that the number of variables used in the linear model is very limited thus potentially making the model vulnerable to omitted variable bias.

Next, we run a Lasso regression presented in table 5.1 in Python using the GridSearchCV function. The lasso regression is a regression technique that eliminates unimportant variables. We put all 44 variables into the Lasso regression. The categorical variables are already encoded by using the JamesSteinEncoder. The Lasso estimates that 18 variables are considered important and thus have an effect in price. The coefficient estimates from Lasso estimates are grouped with the estimates from the linear regression for comparison.

An important factor for the Lasso model adjustment is the choosing the right λ tuning parameter. As discussed in the previous section, the λ parameter is the amount of shrinkage for the regression. We will be fine tuning the lasso λ parameter for the model by constructing a graph of different values of the lambda against the negative mean absolute errors. Figure 5.1 shows the different error values for different λ values.

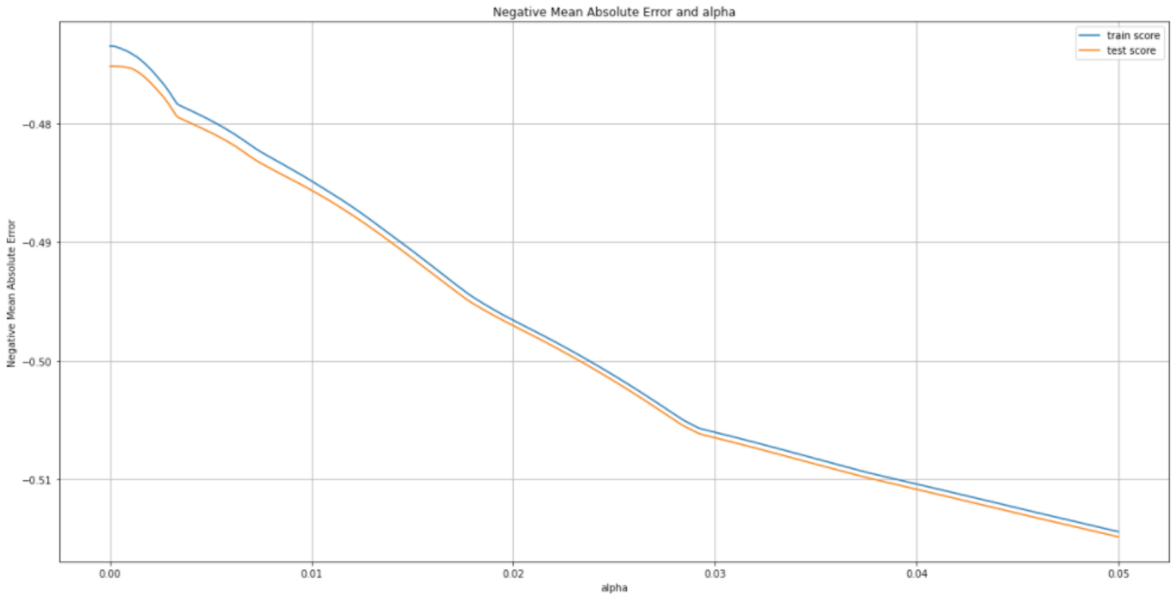


**Figure 5.1** Negative Mean Abs Errors and Alpha (or λ in our notation)

As seen from the graph, the test and train sets exhibit different values of MAE for different levels of λ. Hence, we assign the value of the 0.0002 for the lasso parameter. As can be seen in figure 5.1, this parameter minimizes the mean absolute error. Moreover, a penalty term of 0.05 is assigned for the purposes of variable selection. Estimates with coefficients lower than 0.05 will be removed since these variables are considered not important to the regression. The final variable selection from Lasso shows an 18 variable subset model from the original 44 variable that were initially used.

Next, we fit the CatBoost model presented in figure 5.2 by using the Catboost library from Python. The CatBoost algorithm uses an advanced extreme boosting approach where each variable is given an individual importance score. Each importance feature represents the difference between the loss value of the model with that specific feature and without it. The output received from CatBoost gives a lengthy table where for each variable there is a specific importance score. It is important to emphasize that the CatBoost model is not meant to be a variable selection or dimension reduction approach. The algorithm gives an importance score to every variable even though for many of the variables is extremely small or close to zero.

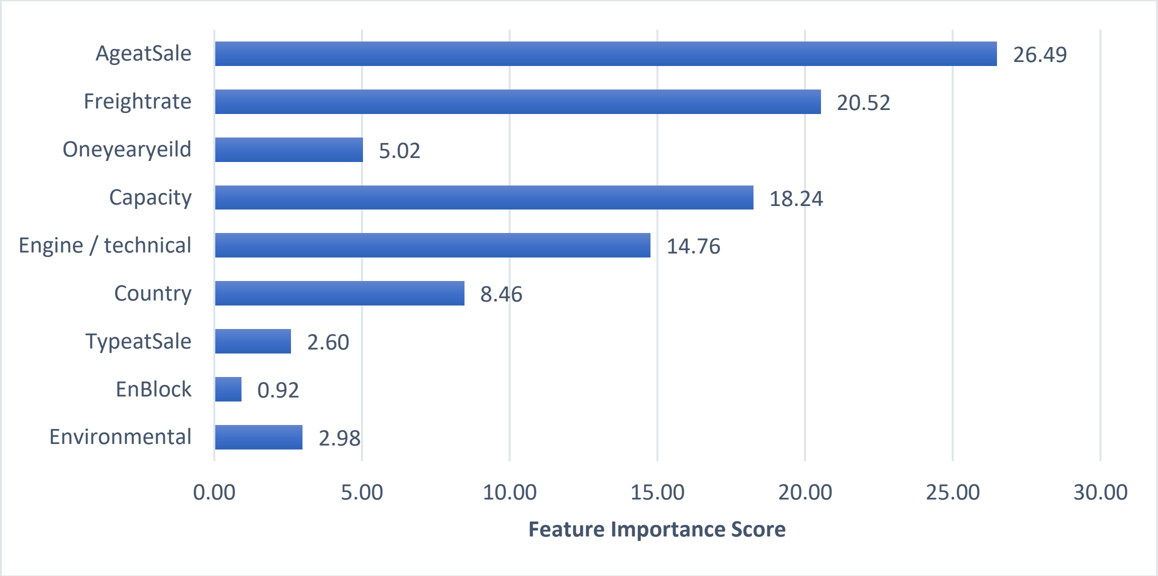


**Figure 5.2** CatBoost Regression Results Grouped by Category

The figure above is an edited graph where variables are grouped in similar categories. Capacity group includes vessel specific features like DWT, Loa, GT, etc. Engine group represents all the variables that are related to speed and other engine related aspects. Appendix A7 shows the CatBoost importance feature scores for the most important variables that are not grouped like in figure 5.2.

Lastly, we fit the data into the PCA algorithm. The nature of this algorithm is focused on dimension reduction where the selected variables maximize the variance in the data. PCA choses principal components for every stage of the regression. The first principal component is the one variable that has the highest amount of variation. Then in the second stage the algorithm choses the second component that maximizes the variation with the condition that the principal components selected are linearly independent.

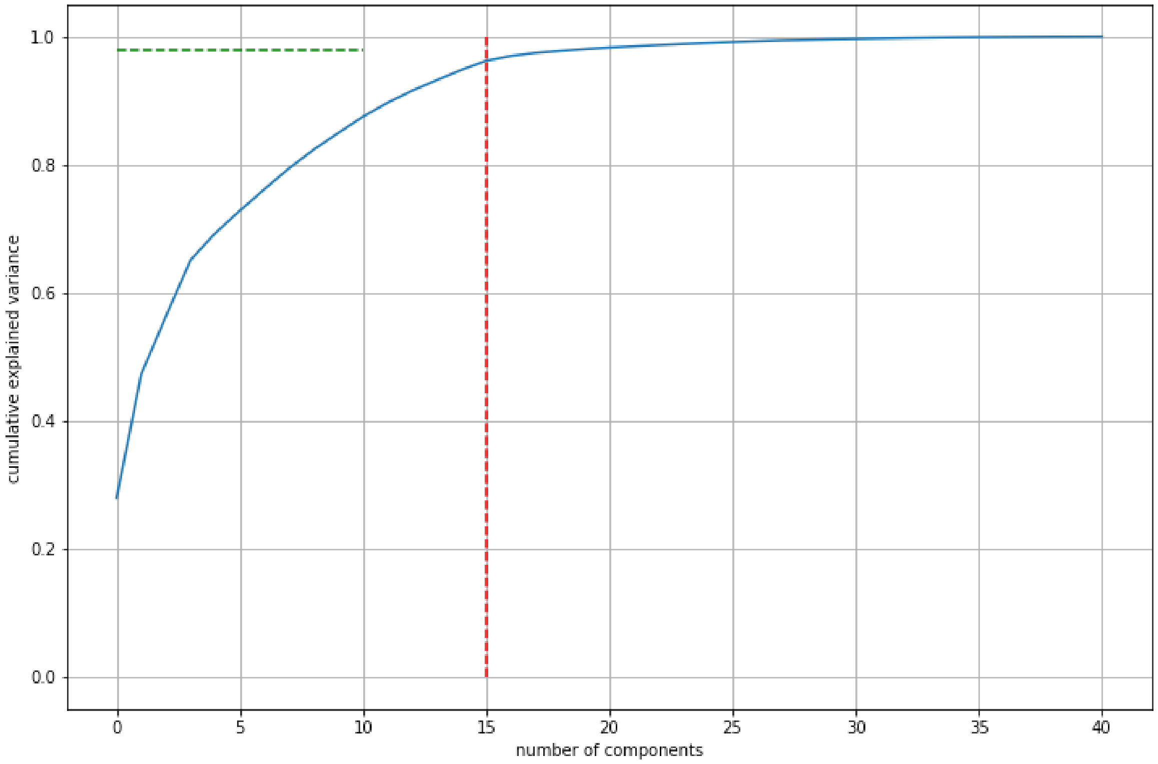


**Figure 5.3** The PCA Component Selection

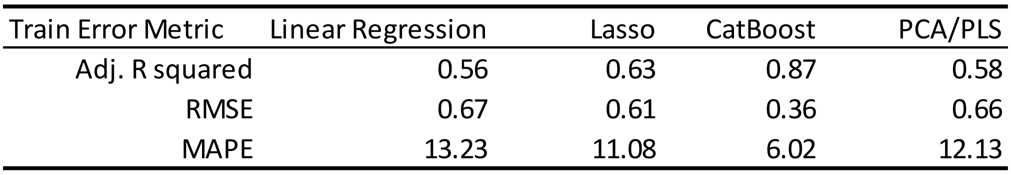
As seen from the graph the PCA has selected only 15 principal components that maximize the variance. The selected components from PCA have a cumulative variance of roughly 0.97 of the entire dataset. We then fit the partial least square regression into the dimension reduced data to test the out of sample performance and convey the accuracy error metrics.

In the upcoming section, we will evaluate the fitted models that were explained above. The analysis will focus on measurement error metrics, and we will interpret any relationships that variables show with each other. The main purpose of this thesis is to recommend a model that is fairly accurate in predicting vessel prices. Hence, the ideal model should have strong predictive powers but also not show signs of extreme overfitting.

## 5.2 Evaluation of Models

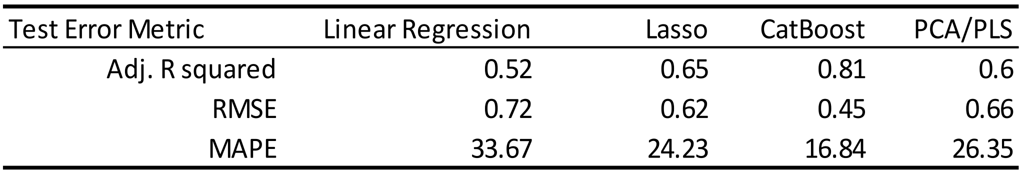
Table 5.2 and 5.3 shows the error metrics of each model being analyzed for the purposes of this thesis. The Linear regression error metrics for both training and test set are compared against the Lasso, PCA and CatBoost metrics, respectively. James et al. (2013) argues that if test errors are slightly larger than training errors then it indicates a sensible tradeoff between overfitting and underfitting the training data. However, in the case that the test errors are excessively large, this is an indication of model overfitting and that the model is not suitable for out-of-sample predictions. Table 5.2 and 5.3 show that the error metrics of training are lower than the test set. Thus, indicating that models are not underfitting. However, the test error metrics for the linear regression and Lasso indicate large error metrics when compared to the training set metrics.

**Table 5.2** Evaluation Metrics for train set



| Train Error Metric | Linear Regression | Lasso | CatBoost | PCA/PLS |
|---:|---:|---:|---:|---:|
| Adj. R squared | 0.56 | 0.63 | 0.87 | 0.58 |
| RMSE | 0.67 | 0.61 | 0.36 | 0.66 |
| MAPE | 13.23 | 11.08 | 6.02 | 12.13 |

**Table 5.3** Evaluation Metrics for test set



| Test Error Metric | Linear Regression | Lasso | CatBoost | PCA/PLS |
|---:|---:|---:|---:|---:|
| Adj. R squared | 0.52 | 0.65 | 0.81 | 0.6 |
| RMSE | 0.72 | 0.62 | 0.45 | 0.66 |
| MAPE | 33.67 | 24.23 | 16.84 | 26.35 |

The error metrics tables show that the linear regression and Lasso have relatively high scores compared to the other machine learning algorithms. The linear regression exhibits a low R squared coupled with a high RMSE and an even higher MAPE test error. These numbers suggests that the linear model exhibits the lowest prediction power among the other models. Comparing the linear model with LASSO, we can see that Lasso has slightly better error metrics. This is due to the fact that LASSO introduces a penalty error term that takes out irrelevant variables and thus putting a higher emphasis only on the variables that truly affect price. However, even though LASSO performs better, it still is not accurate enough for out of sample predictions since the test MAPE and RMSE are considerably higher than the machine learning algorithms. Hence, we rule out the first two models and we will further compare CatBoost and PCA in order to suggest the model that is optimal for vessel price prediction.

Appendix A5 and A6, shows the test prediction performance of CatBoost and PCA categorized by vessel type. As seen from the graph Catboost exhibits predictions that are considerably closer to the 45-degree line compared to PCA. Gas Carrier and Tanker predictions are the two vessel types that show the biggest difference in terms of performance between the two algorithms. Even though, the error metrics for PCA and CatBoost are not very different, the prediction performance in appendix A5 and A6 points to choosing the CatBoost as the algorithm that we should proceed analyzing further.

From Table 5.3, it can be observed that the Catboost model exhibits the highest R$^2$ score out of all the models. An R squared score of 0.81 means that the model explains 81% of the variance in the testing set observations' prices. Even though a high R$^2$ is an auspicious sign of a strong predictive model, it can be misleading for analysis to only focus on such a measure. The rest of the error metrics (RMSE and Mape) both show that Catboost has the lowest metrics and seems to predict the data more efficiently than the other models.

The CatBoost offers many advantages that would make this model suitable for vessel valuation in the industry. As a result, this model will be taken for further analysis for the following reasons. Firstly, the algorithm's sole purpose is to handle categorical data and generate a model that predicts with reliability and efficiency (i.e. in terms of computational power). Secondly, the algorithm exhibits the lowest error metrics from the mix of the proposed models which make CatBoost an overall reliable model to predict vessel prices.

Based on figure 5.2, the most important features of the model are age, freight and interest rates. This finding supports economic theory suggesting that the older a ship gets the more the price is affected by such a feature. Furthermore, the higher the freight rate of a vessel is the more valuable that ship is perceived in the industry. Hence, it is sensible for such a variable to affect ship prices as shown in figure 5.2. The most interesting finding is the relative importance that the one-year yield bonds have on vessel price. Furthermore, CatBoost results confirm the findings from Koray and Cetin (2020) that argue that vessel prices are influenced by vessel age, tonnage, shipyard built, and other specific vessel features.

To further analyze the marginal contribution that each important feature brings into the model, the Shap values for the CatBoost model are presented. The Shap values are the most important tool for explaining results from machine learning models. The benefits of using this tool in our model evaluation are several: 1) global interpretability, 2) local interpretability, and 3) Shap values can be used for any regression tree-based model (Lundberg et al. 2017).

Shap values are calculated in a way that the numerical value presents the difference that the model prediction would have with and without that feature included in the model.

As presented in figure 5.4, the variables that are labeled in blue show how those features lower the model prediction by the respective value presented. The red variables express the numerical value that the variable helps increase the prediction of the model.

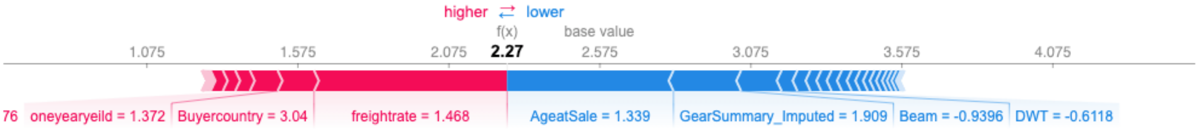


**Figure 5.4** Shap Value of the CatBoost model

The difference between the blue bars and red bars explains the overall difference between our CatBoost model and the base model.[2] It is worth noting that the color of the variable indicates whether that particular variable increases or decreases predictability. Figure 5.4 suggests that age does hurt the prediction power of the model. This can be explained in practice by the fact that older ships might be harder for brokers to value since these vessels have older technology and history of repairs that might affect value.

To further examine the contribution of the variables we present another shap value graph. Figure 5.5 shows two important qualities 1) the contribution of the variable in the overall model and 2) the relationship of the variables with the target variable price.

[2] Base value E(y_hat) is "the value that would be predicted if we did not know any features for the current output." In other words, it is the mean of all branch predictions, or mean(yhat).

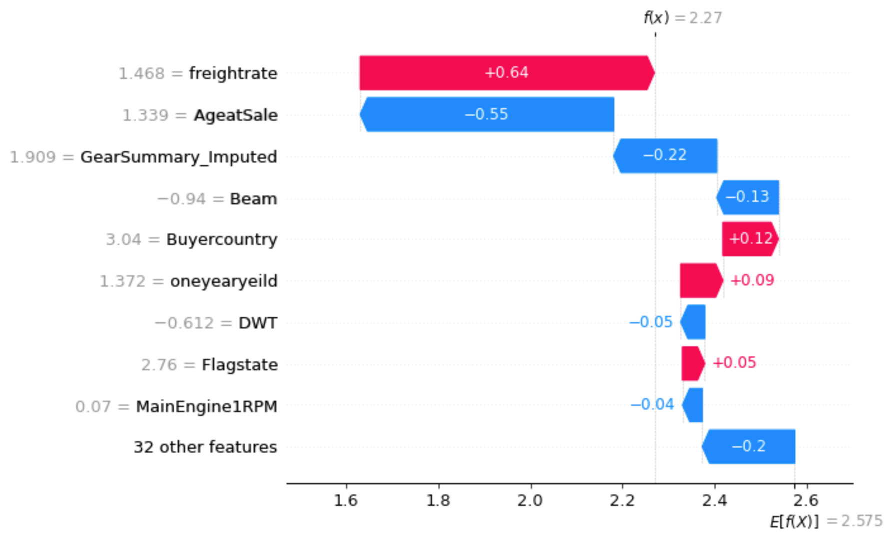


**Figure 5.5** Shap Values ranked by contribution importance

As seen in the figure 5.5, freight rate has a high and positive impact on the vessel price. Whereas a blue negative shap value for AgeatSale shows a high negative contribution that age has with prices. The Shap value of 0.64 means that if freight rates were not included in the data then the model would be hurt by changing the predicted price by 33.2% and thus deteriorating the model’s predictive powers the most.[3]

Another aspect that should be analyzed is the relationship that the most important independent variables exhibit with each other and also with the target variable. As a result, we construct two Shap dependence plots. This particular approach makes possible to analyze the marginal effect that one feature has in the overall model and how it interacts with another independent variable. The purpose of including this analysis is to inspect the relationship that the independent variables have with the target variable.

[3] Shap values are scalable in percentages by taking the effect of one variable and comparing against the effect of the rest.

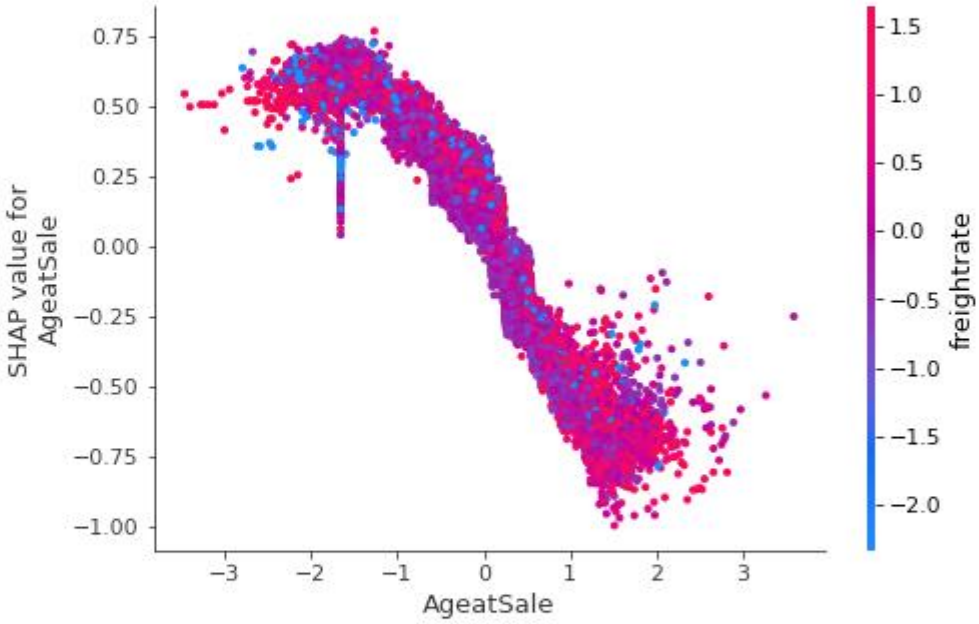


**Figure 5.6** Age and Freight Rate Dependence Plot

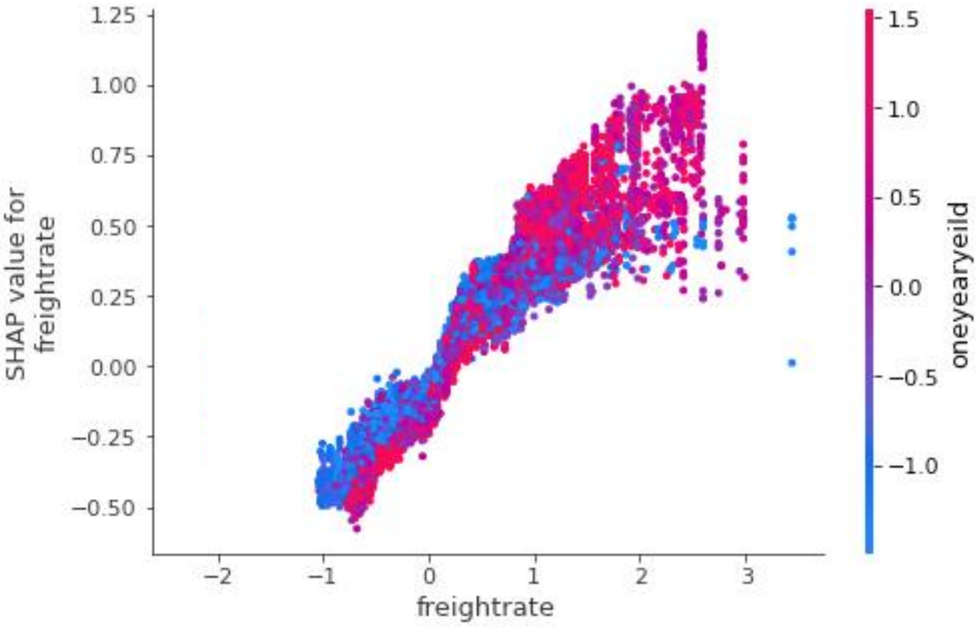


**Figure 5.7** Freight Rate and OneYearYield Dependence Plot

As seen in figure 5.6 there is a negative relationship between the age variable and the dependent variable. Furthermore, the relationship can be considered as non-linear, thus confirming previous research discussed in the literature review section. An interesting aspect is analyzing the age and freight rates interaction with each other. There are two important points that need to be emphasized: 1) for vessel with negative age[4] freight rates on average decreases the prediction power for vessel price and 2) for positive vessel ages, freight rates increase the prediction for vessel prices. This implication seems to be logical since freight rates for ships that have been operational for some years tend to help more in predicting a vessel's price rather than a ship that is still being built in shipyard and is yet to be operational.

Another interesting relationship is shown in the figure 5.7 between freight rates and one year yield and how they interact with the target variable price. The graph suggests a positive linear relationship between freight rates and interest rates. Furthermore, the graph shows that for vessels with high freight rates the interest rate improves the prediction power. In practice, this makes economic sense because vessels with high freight rates are considered more valuable and are more expensive. As a result, fluctuations in interest rates affect these vessel's prices more, thus improving the prediction effect that interest has on vessel prices.

[4] Negative age is when a ship is sold before being build completely and be operational.

# 6. Conclusion

The purpose of this thesis is to compare the accuracy of extreme gradient boosting (EGB) algorithms such as CatBoost against other regression models and conclude whether EGB algorithms perform better. We found evidence that the Catboost model on average outperforms on terms of accuracy the linear regression, Lasso regression, and the Principal Component Analysis model.

The results indicate that the Catboost model does a better job predicting vessel prices that are close to real prices. PLS/PCA model on average overestimates the predicted prices by showing an upward trend for vessel up to $10 million dollars. Above the cutoff value of $10 million dollars, PCA shows an underestimation of prices thus exhibiting opposite predictive errors sign values for vessels in different price brackets.

Another tradeoff between the Catboost model and the PCA model is that the former exhibits considerably better predictive power and the latter performs dimension reduction of the variables that hold most of the variation of the dependent variable. Considering that the goal of this thesis is to provide a reliable model that accurately predicts vessel prices, it is reasonable to conclude that the CatBoost algorithm is the appropriate model to adopt for vessel price predictions.

Another interesting point worth noting is that CatBoost identified the age at sale, freight rates, and 1-year yield as the three most important important variables in vessel valuation. These findings confirm previous research in vessel valuation (Adland and Koekebakker, 2007) that argue that age of ships and freight markets influence the valuation of vessels. Furthermore, CatBoost results further confirm the findings that argue that vessel prices are influenced by vessel age, tonnage, shipyard built, and other specific vessel features.

Although Catboost seems to be delivering promising results in terms of accuracy, this thesis has certain limitations. The data collected from Clarkson World Fleet Registry had many missing data points that were imputed in order to create a comprehensive dataset. Furthermore, the categorical variables were encoded by JamesStein encoder which even though was the most accurate encoder, it still has some error built in. Lastly, the omitted block sale prices accounted for more than 18% of all the quoted prices in the data. These numbers were replaced by an average sale price calculated over the entire block price acquisition which naturally affects negatively CatBoost’s accuracy.

In future, we recommend that machine learning algorithms should be further applied since these approaches show a promising potential. It would be interesting to see future studies that use Catboost, or similar gradient boosting algorithm, to analyze the impact that environmental features have on vessel price. As technology is advancing and as new regulations are forcing new vessels on  becoming more eco-friendly, we suspect that environmental features that were included in this paper will become more important and affect the final price considerably more.

# References

Adland, R. and Köhn, S. (2019). Semiparametric valuation of heterogeneous assets. In Mathew, J., Lim, C., Ma, L., Sands, D., Cholette, M. E., and Borghesani, P., editors, Asset Intelligence through Integration and Interoperability and Contemporary Vibration Engineering Technologies, chapter 3, pages 23–30. Springer, Cham.

Adland, Roar, and Siri Strandenes. "Market Efficiency in the Bulk Freight Market Revisited." *Maritime Policy & Management*, vol. 33, no. 2, 2006, pp. 107–117., doi:10.1080/03088830600612773.

Aubry, Mathieu, et al. "Machine Learning, Human Experts, and the Valuation of Real Assets." *EconStor*, Frankfurt a. M.: Goethe University Frankfurt, Center for Financial Studies (CFS), 1 Jan. 1970, www.econstor.eu/handle/10419/206414.

Bodea, Tudor, and Mark Ferguson. *Segmentation, Revenue Management, and Pricing Analytics*. Routledge, 2014.

FENG, GUANHAO, et al. "Taming the Factor Zoo: A Test of New Factors." *The Journal of Finance*, vol. 75, no. 3, 2020, pp. 1327–1370., doi:10.1111/jofi.12883.

Fortune Business Insight. "Marine Vessel Market Size, Share and Industry Analysis, by Type (Commercial Vessel, Passenger Ship, LNG/LPG Carrier, and Special Purpose Vessel), By System (Marine Engine System, Sensor System, Control System, Electrical System, Auxiliary System, and Communication System), and Regional Forecast, 2019- 2026." *Infographics - Marine Vessel Market*, 2 Feb. 2018, www.fortunebusinessinsights.com/infographics/marine-vessel-market-102699.

Friedman, Jerome H. "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics*, vol. 29, no. 5, 2001, doi:10.1214/aos/1013203451.

Gholamy, Afshin, et al. "Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation." Feb. 2018.

Greenwell, Bradley Boehmke & Brandon. "Hands-On Machine Learning with R." · *Bradley Boehmke*, 1 Feb. 2020, bradleyboehmke.github.io/HOML/.

Gu, S., Kelly, B., and Xiu, D. (2018). Empirical asset pricing via machine learning. Technical report, National Bureau of Economic Research.

Hancock, John, and Taghi M Khoshgoftaar. "CatBoost for Big Data: an Interdisciplinary Review." 2020, doi:10.21203/rs.3.rs-54646/v2.

Harvei, Hans Christian, and Julius Jorgensen. "Second-Hand Vessel Valuation A Generalized Additive Model and Extreme Gradient Boosting Approach." Dec. 2019.

Iqbal, Zain. "Overview - How The Shipping Industry Works & Key Catalysts." *Alpha Invesco*, 3 Mar. 2020, www.alphainvesco.com/blog/how-does-the-shipping-industry-work/.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning. Springer, New York.

Köhn, S. (2008). Generalized additive models in the context of shipping economics. PhD thesis, University of Leicester.

Koray, Murat, and Oktay Cetin. "A Combined Qualitative Ship Valuation Estimation Model." *WMU Journal of Maritime Affairs*, vol. 19, no. 2, 2020, pp. 205–217., doi:10.1007/s13437-020-00202-2.

Micci-Barreca, Daniele. "A Preprocessing Scheme for High-Cardinality Categorical Attributes in Classification and Prediction Problems." Mar. 2001.

Paul, R. K. (2006). Multicollinearity: Causes, effects and remedies. IASRI, New Delhi.

Perez, Lexi V. "Principal Component Analysis to Address Multicollinearity." 13 May 2017.

Prokhorenkova, Liudmila, et al. "CatBoost: Unbiased Boosting with Categorical Features." 2019.

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 & Logistics, 13(2):213–236.

Sathya, R., and Annamma Abraham. "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification." *International Journal of Advanced Research in Artificial Intelligence*, vol. 2, no. 2, 2013, doi:10.14569/ijarai.2013.020206.

Schreiber-Gregory, Deanna N. "Regulation Techniques for Multicollinearity: Lasso, Ridge, and Elastic Nets." 2018.

Stopford, M. (1988). Maritime Economics. Unwin Hyman, London.

Tibshirani, Robert. "Regression Shrinkage and Selection Via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, 1996, pp. 267–288., doi:10.1111/j.2517-6161.1996.tb02080.x.

Wood, Simon N. *Generalized Additive Models: an Introduction with R.* CRC Press/Taylor & Francis Group, 2006.

Wood SN, "Generalized additive models—an introduction with R. Chapman & Hall/ CRC, London", 2006

# Appendix

## A1 Descriptive Statistics of Numerical Variables

| Variable | Minimum | Mean | Maximum | Missing Data % |
|---|---|---|---|---|
| Dependent Variable | | | | |
| *PRICE (mln USD)* | *0.2* | *18.85* | *1670* | *10.0%* |
| Independent Variable | | | | |
| *TEU* | 0 | 2,003.0 | 19,224 | 78.3% |
| *DWT* | *600* | *61,546.78* | *555,051* | *0.1%* |
| *GT* | *488* | *35,994.93* | *261,453* | *0.1%* |
| *LOA* | *48.5* | *200.15* | *414.22* | *0.1%* |
| *BEAM* | *8.6* | *31.2* | *79* | *0.1%* |
| *DRAUGHT* | *2.5* | *11.95* | *28.6* | *0.1%* |
| *SPEED* | *6.5* | *15.15* | *26.5* | *2.1%* |
| *CONSUMPTION TPD* | *4* | *42.59* | *320* | *21.4%* |
| *HP TOTAL PROPULSION* | *275* | *15,457.80* | *149,555* | *0.1%* |
| *MAIN BUNKER CAPACITY* | *29.2* | *2698.7* | *23,330* | *35.8%* |
| *ENGINE RPM* | *55* | *164.48* | *6,700* | *0.5%* |
| *GRAIN CAPACITY* | *352* | *69,446.98* | *293,683* | *44.5%* |
| *CAPACITY CUM* | *61* | *87,121.69* | *66,3832* | *71.3%* |
| *PUMP CAP CUM* | *0* | *7,127.41* | *75,200* | *78.0%* |
| *PUMPS TOTAL NO* | *0* | *5.79* | *24* | *77.5%* |
| *HOLD TOTAL NO* | *1* | *5.75* | *18* | *34.1%* |
| *HATCHES TOTAL NO* | *1* | *6.85* | *51* | *34.5%* |
| *AGE at SALE* | *-6.6* | *12.6* | *47.6* | *2.2%* |
| *FREIGHT RATE* | *$ 0* | *$16,383* | *$ 98,094* | *0%* |
| *TEN YEAR YEILD* | *0.2%* | *1.85%* | *6.39%* | *0%* |

## A2 Descriptive Statistics of Categorical Variables

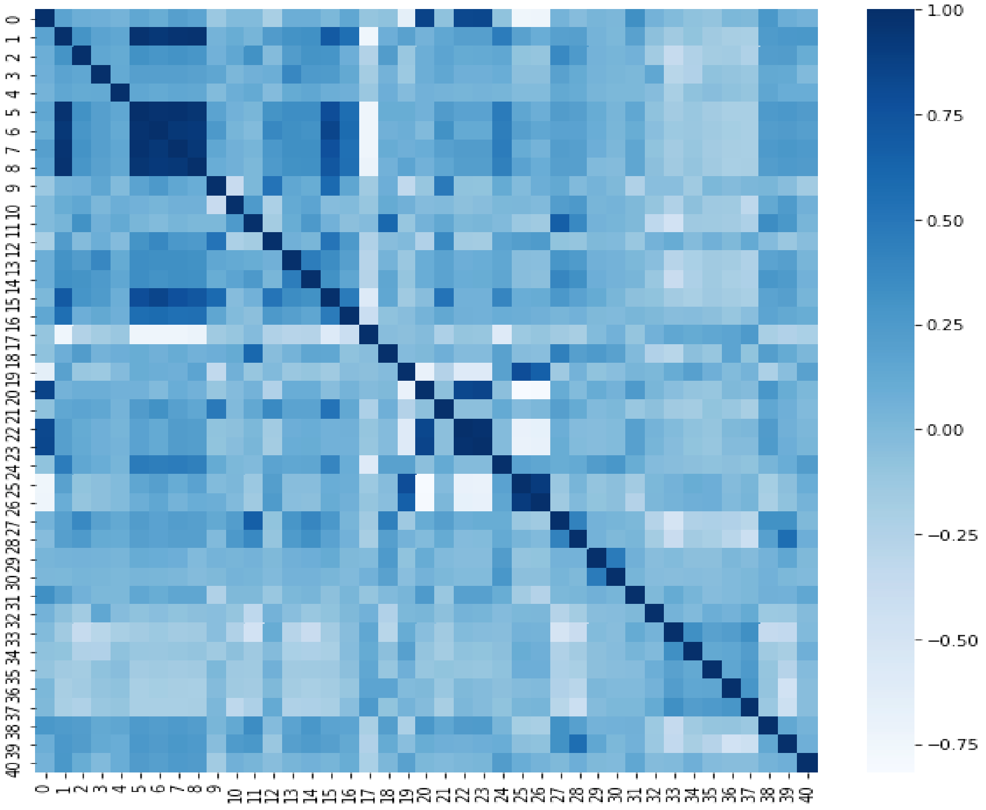
| ***Type at Sale*** | ***Builder Country*** | ***Buyer Country*** | ***Seller Country*** | ***Speed Category*** |
|---|---|---|---|---|
| Bulk *53.1%* | Japan *44.7%* | Greece *26%* | Greece 19.3% | Service *87.2%* |
| Tanker *19.7%* | South Korea *18.3%* | Unknown *16.3%* | Japan *14.6%* | Laden *9.3%* |
| Container *16.2%* | China *16%* | China *11.4%* | Germany *10.2%* | Other *3.5%* |
| TankChem *5.5%* | Germany *4.5%* | Germany *4.9%* | China *5.4%* | |
| Gas Carrier *5.4%* | Other *16.5%* | Other *41.4%* | Other *50.5%* | |

| ***Status*** | ***Owner Country*** | ***Flagstate*** | ***B. Fuel Type*** | ***Power Type*** |
|---|---|---|---|---|
| F *59.7%* | Greece *21.8%* | Panama *24%* | VLF IFO *56.3%* | Diesel 2 *89.8%* |
| D *39.9%* | China 14.*2%* | Liberia *13%* | IFO 380 *36.1%* | Diesel 4 *9.4%* |
| Other *0.3%* | Singapore *6.4%* | Marshall Is. *10.4%* | IFO 180 *4.8%* | Steam T *0.6%* |
| | Other *57.6%* | China *5.6%* | Other *1.1%* | Other *0.2%* |
| | | Other *47%* | Missing *1.7%* | |

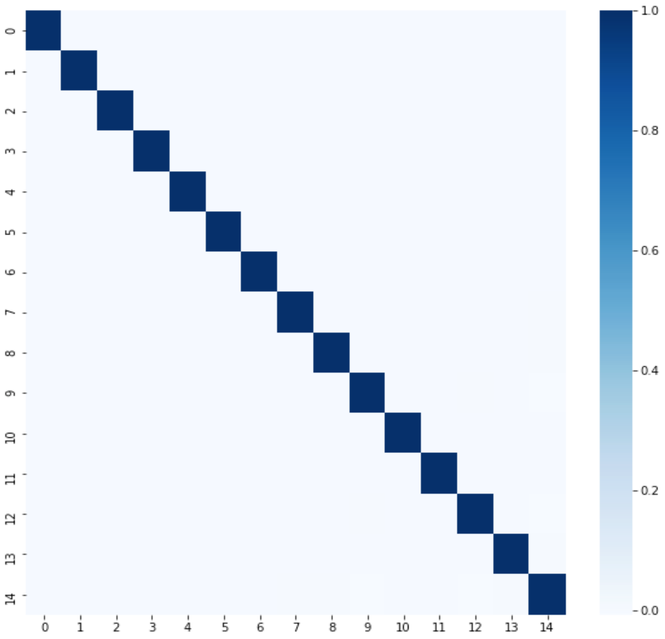
| ***Classification Soc.*** | ***BWMS Ind*** | ***LNG Contain. Type*** | ***Propulsion Type*** | ***En Bloc*** |
|---|---|---|---|---|
| N/A *42%* | **0** *82.2%* | Membrane *0.3%* | Mechanical *99.8%* | **1** *71.2%* |
| Nippon *12.3%* | **1** *17.8%* | Aluminium *0.2%* | Electric *0.2%* | **0** *28.8%* |
| DNV Group *9.2%* | | Missing *99.5%* | Hybrid *0.1%* | |
| Bureau *7.1%* | | | | |
| Other 29.*4%* | | | | |

| ***Energy Saving Tech*** | ***SOxScrubber*** | **EcoEI_engine** | **Gear Summary** | **Env. Summary** |
|---|---|---|---|---|
| **1** *96.6%* | **1** *95.2%* | **1** *93.7%* | 4x30t cr. *70.4%* | 1xBWTS *83.7%* |
| **0** *3.4%* | **0** *4.8%* | **0** *6.3%* | 4x25t cr. *7%* | Other *16.3%* |
| | | | 4x30.5t cr. 2% | |
| | | | 2x40t cr. 1.7% | |
| | | | Other *16.5%* | |

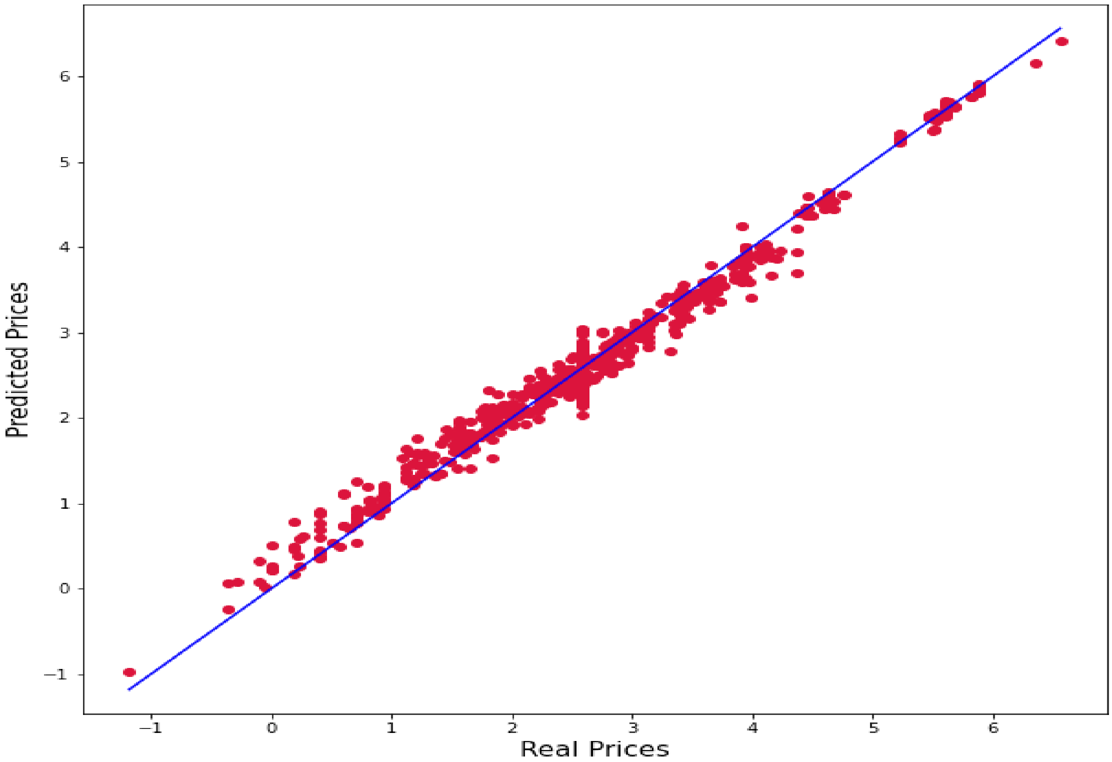
## A3 CatBoost Correlation Heatmap

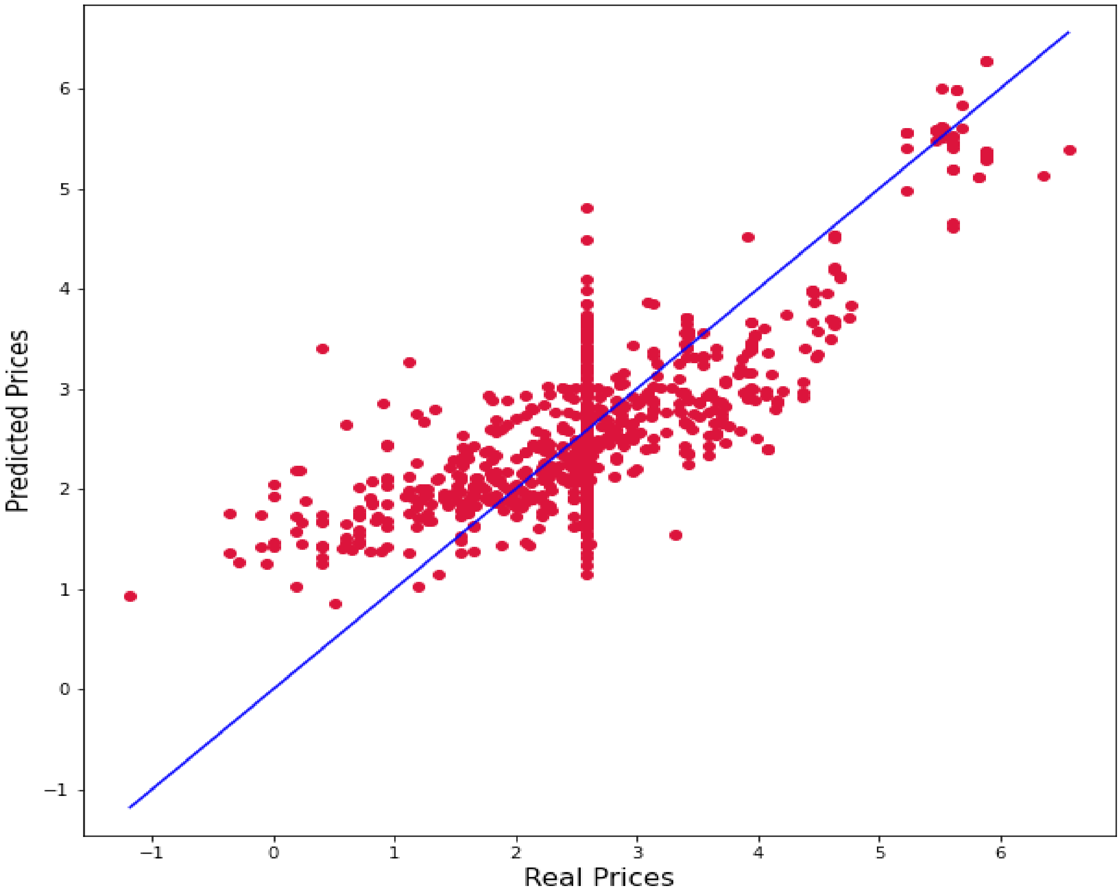


## A4 Correlation Heatmap of the Selected PCA Variables

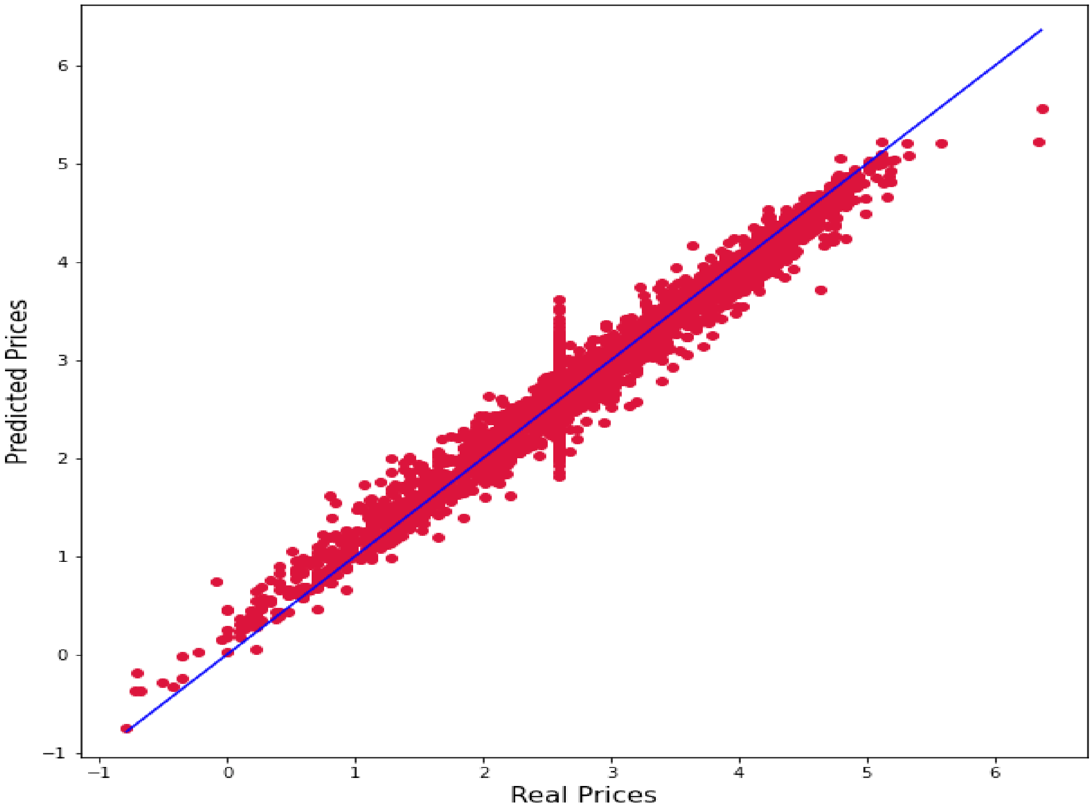


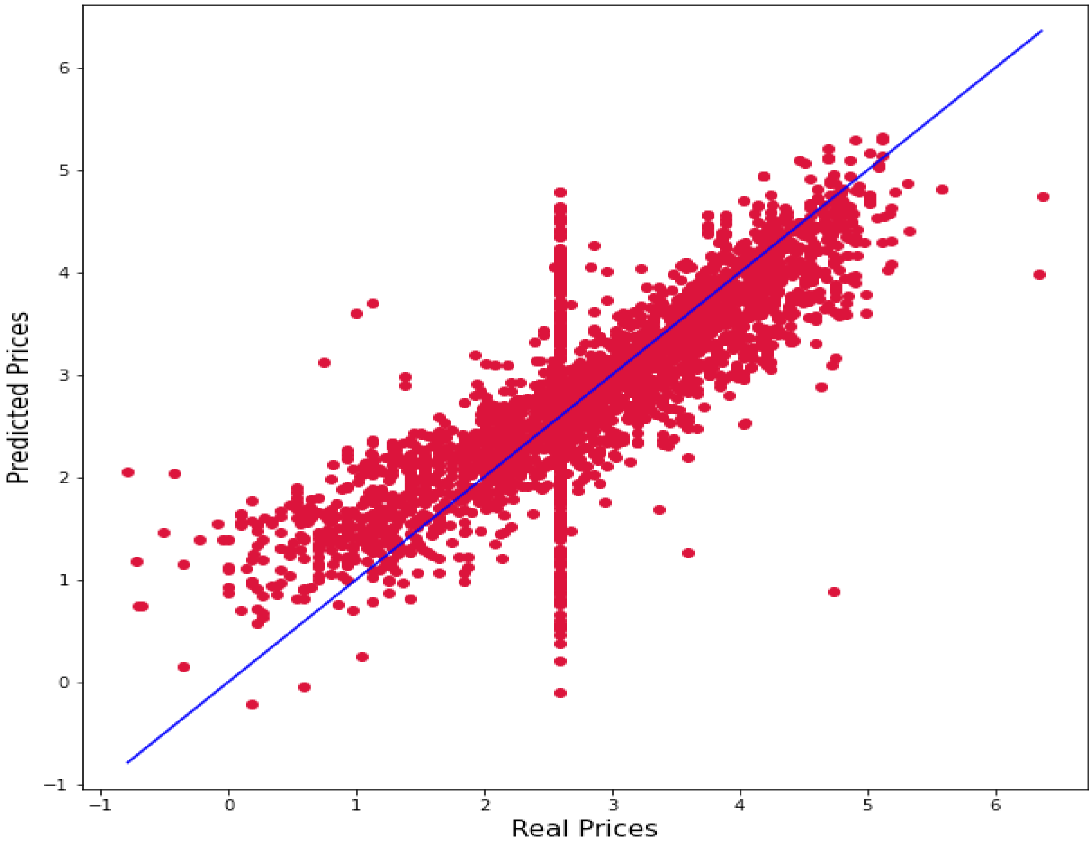
## A5 CatBoost (top graph) and PCA (bottom graph) Gas Carrier



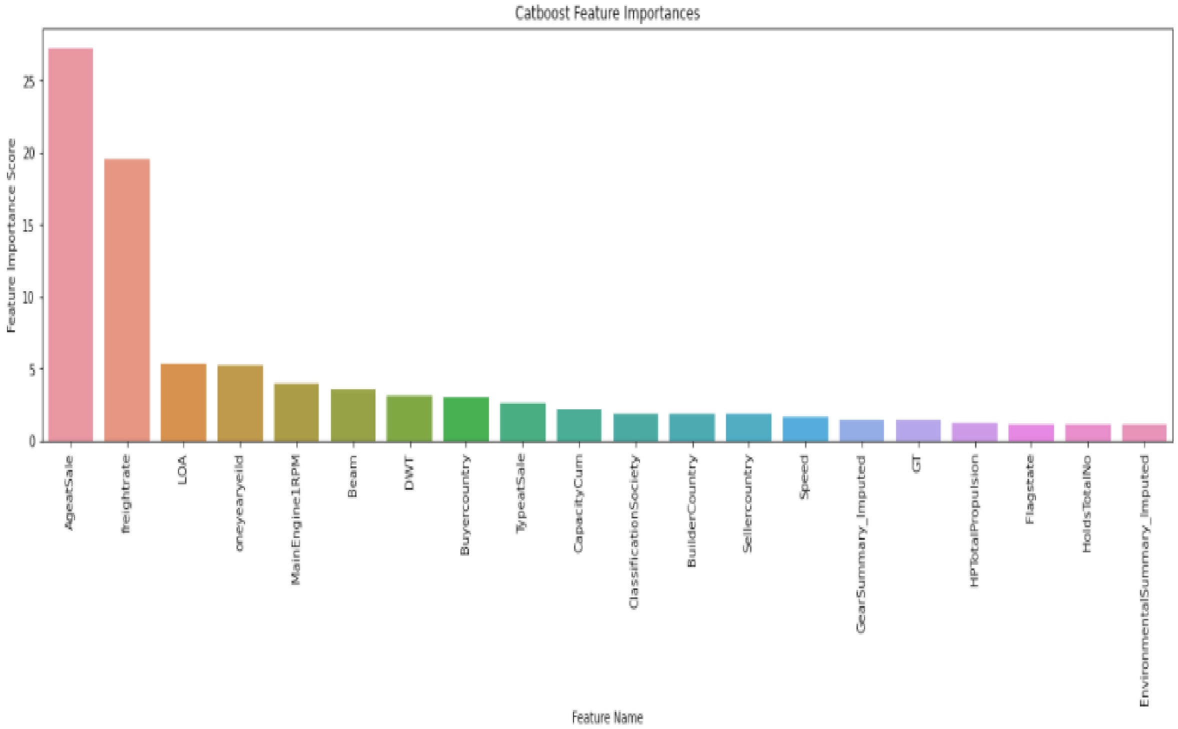


## A6 CatBoost (top graph) and PCA (bottom graph) Tanker

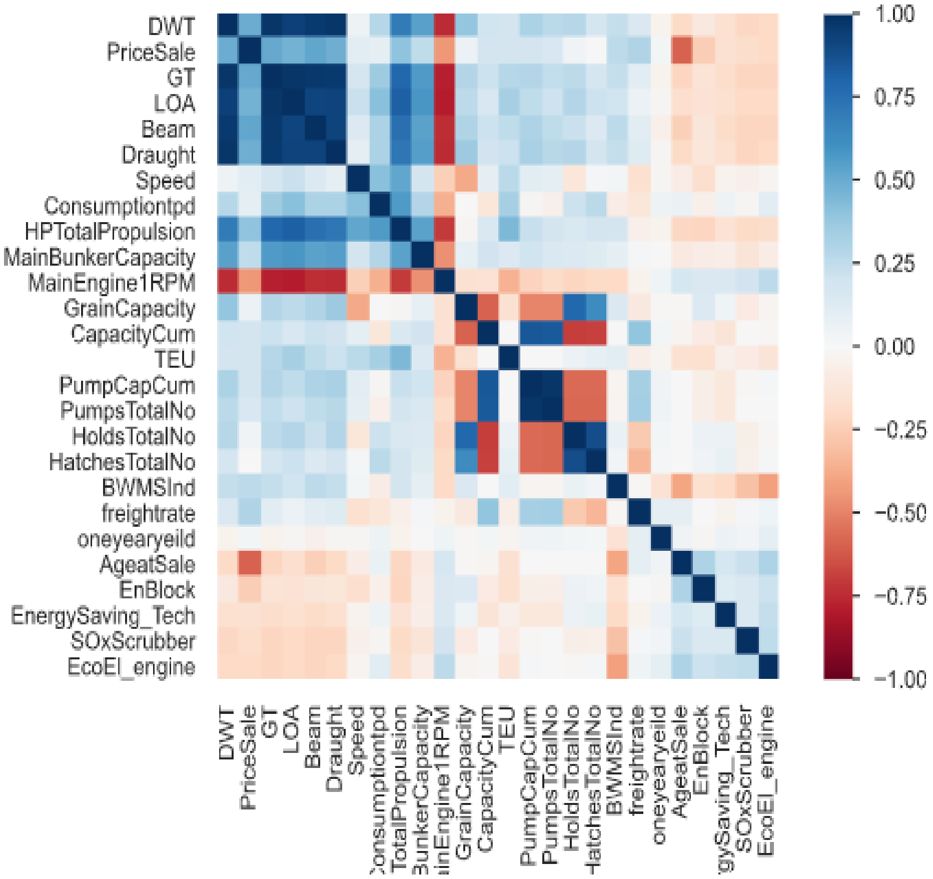




## A7 CatBoost Importance Scores



## A8 Pearson Correlation Heatmap of the Data


DWT
PriceSale
GT
LOA
Beam
Draught
Speed
Consumptiontpd
HPTotalPropulsion
MainBunkerCapacity
MainEngine1RPM
GrainCapacity
CapacityCum
TEU
PumpCapCum
PumpsTotalNo
HoldsTotalNo
HatchesTotalNo
BWMSInd
freightrate
oneyearyeild
AgeatSale
EnBlock
EnergySaving_Tech
SOxScrubber
EcoEl_engine