Norwegian School of Economics
Bergen, Fall 2019

Second-Hand Vessel Valuation

A Generalized Additive Model and Extreme Gradient Boosting Approach

Hans Christian Olsen Harvei and Julius Jørgensen

Supervisor: Roar Os Ådland

Master thesis, Economics and Business Administration

Major: Business Analytics and Financial Economics

NORWEGIAN SCHOOL OF ECONOMICS

Acknowledgements

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

First and foremost, we would like to thank our supervisor, Roar Os Ådland, for sharing his extensive knowledge and giving us constructive feedback throughout the whole process. His insights and interest in the field of maritime economics has been very valuable for our thesis. We would also like to thank the Norwegian Shipowners' Association for the grants provided to us.

Norwegian School of Economics

Bergen, December 2019

Hans Christian Olsen Harvei

Julius Jørgensen

Abstract

This thesis investigates the applicability of extreme gradient boosting (XGBoost) compared to the generalized additive model (GAM) approach to create a desktop valuation model of second-hand Handysize bulkers. The data basis is 1880 unique sales transactions in the period from January 1996 to September 2019 derived from Clarkson Research. This thesis contributes to existing literature by applying an XGBoost algorithm and a data-driven GAM approach to vessel valuation.

Using vessel-specific and market variables, we find evidence that the XGBoost algorithm is more suited for desktop valuation of Handysize bulk carriers than the GAM approach. The predictive power of XGBoost in this instance could be caused by its ability to model complex relationships between multiple variables. Supporting existing research in maritime economics, we find linear models to be inadequate at vessel valuation. When fitting the XGBoost model, vessel age at sale, timecharter rates and fuel efficiency index are identified as the most important variables. We also find vessels priced over $20 million to be significantly harder to predict. This could be a consequence of a scarce data basis, vessel characteristics not present in the data or the majority of these transactions occurring during the financial super cycle years of 2003 to 2009.

The XGBoost algorithm facilitates accurate predictions of vessel prices based on vessel characteristics and market conditions, and provides a useful machine learning framework for desktop valuation. The flexibility of the XGBoost algorithm can make it highly usable for investors, ship owners and other market players in the maritime industry.

Keywords – Second-Hand Vessel Valuation, Generalized Additive Models (GAM), Extreme Gradient Boosting (XGBoost)

Contents

List of Figures

List of Tables

1 Introduction

Shipping is an industry were the main asset, the ships, are frequently traded. Therefore, the second-hand market, also called the sale and purchase (S&P) market, plays an important economic role in the industry. *"The remarkable feature of this market is that ships worth tens of millions of dollars are traded like sacks of potatoes at a country market"*, Stopford (2009, p.198). Vessels have the characteristics of fixed assets, which means they are liquid, easy to resale, and valuable as scrap because of the hull construction (Thalassinos and Politis, 2014). Second-hand prices of vessels fluctuate rapidly, as Stopford (1988, p.383) notes: *"Typically, 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"*. The volatility in the market makes it important to purchase/sell a vessel at the right time, as these vessels are worth millions of dollars.

Volatility and frequent transactions in the S&P market emphasize the importance of a fast and reliable "desktop valuation" of vessels. Desktop valuation models provide an assessment of the value of an asset without the need for a physical inspection (Warren et al., 2005). Vessel-specific desktop valuation is particularly valuable to brokers, financiers and owners when brokers valuations are expensive or not available (Adland and Köhn, 2019). For investors trying to diversify their portfolio with real assets, in this case ships, an accurate valuation is also important. Real assets tend to have low correlation with financial assets, making them valuable for a diversified portfolio (Bodie et al., 2014).

The focus of this thesis is the S&P market for Handysize bulk carriers, where vessels are specified by a deadweight carrying capacity between 10000 and 40000 tonnes (Clarkson Research, 2019b). A bulk carrier is a single-deck ship designed to carry dry cargoes, such as ore, coal, sugar or cereals (Stopford, 2009). The trading flexibility, and moderate investment size of Handysize bulkers, ensures that the second-hand market for those ships remains sufficiently liquid throughout the shipping market cycle (Adland et al., 2018). As the Handysize vessels are relatively small, and less costly than bigger ships, they are a lot more frequently traded than other ship sizes.

In Maritime Economics by Stopford (2009), some main influences on the ship price at a specific point in time are proposed, two of them being freight rates (timecharter rates) and vessel age. For the freight market, highs and lows are transferred into the S&P market. The second influence is age, and vessel values are typically depreciated by 4-6% each year. Furthermore, there are many other ship-specific factors influencing the value of the vessel, like ship size, ship type, yard of build and technical specifications such as speed, fuel consumption, type of engine and cargo capacity.

Previous publications have suggested that a linear model is insufficient in desktop valuation of second-hand vessels (Adland and Koekebakker, 2007; Adland and Köhn, 2019). Köhn (2008) found that the generalized additive model (GAM) framework in the context of shipping economics yielded good results, suggesting GAMs to be a good way to model economic aspects of the maritime industry. Different machine learning approaches have proved useful in valuations of financial assets, and could therefore also be interesting in vessel valuation (Gu et al., 2018). Extreme gradient boosting (XGBoost) is a very powerful and effective scalable tree boosting machine learning algorithm, widely used by data scientists to achieve state-of-the-art predictive accuracy (Chen and Guestrin, 2016). Since GAM and XGBoost respectively are renowned for their flexibility and predictive power, we fit these models to a set of historical transaction data, and test which model provides the most accurate valuations of second-hand Handysize vessels.

This master's thesis, to the best of our knowledge, has several contributions to second-hand vessel desktop valuation. This is the first academic article on second-hand price modelling testing a data-driven iterated GAM. This approach tests a series of GAMs with different interaction terms, choosing the most accurate. Secondly, we contribute to the field of shipping economics by comparing an XGBoost model's performance in desktop valuation with the currently renowned GAM approach. This leads to our research question:

> *Is an extreme gradient boosting approach more suitable than a data-driven generalized additive model in valuation of second-hand Handysize bulk carriers?*

The remainder of this thesis is structured as follows. First, we review relevant existing literature regarding second-hand vessel valuation. Next, we describe the data and methodical framework for the thesis. Further, the results are compared and analyzed. Finally, a conclusion is presented, with limitations and recommendations for future research.

2 Literature Review

In this section we review relevant literature to our thesis from an application point of view, reviewing papers addressing second-hand vessel pricing, mainly through econometric models.

The literature on second-hand vessel valuation is mainly twofold. One stream is dedicated to testing whether the efficient market hypothesis (EMH) holds in the shipping market. The EMH is tested in many previous papers, such as Hale and Vanags (1992), Glen (1997), Kavussanos and Alizadeh (2002) and Adland and Koekebakker (2004). Kavussanos (1996a,b, 1997) creates models based on the autoregressive conditional heteroscedasticity, utilized for the purpose of second-hand vessel pricing. His main focus is the dynamics of price volatility for different sized second-hand ships, and concludes that volatility is lower in Handysize compared to the bigger vessel sizes. Pruyn et al. (2011) summarize the past 20 years of research on the EMH, and find it to still be inconclusive in shipping economics. The second stream, based on econometric models, is the main focus of the literature review.

In the maritime industry, many of the empirical researches about second-hand vessel pricing are based on time-series. Charemza and Gronicki (1981) suggest equations for supply and demand, where ship prices adjust over time by freight- and activity rates. Beenstock (1985) argues that vessels should be considered as a capital asset, and supply and demand analysis is inappropriate. He proposes a framework based on portfolio theory, with expected ship returns inversely related to alternative investments. In subsequent papers, Beenstock and Vergottis (1989a,b, 1993a,b) further develop this idea.

Tsolakis et al. (2003) continue the time-series work, but focus more on a structural approach in modelling second-hand prices. They create functions where the demand is expressed as timecharter rates, new-building price, second-hand price and the cost of capital (3-month LIBOR). Tsolakis et al. (2003) propose supply as a function of orderbook to fleet ratio and second-hand prices, and find an equilibrium point for the second-hand vessel price based on vector autoregression and non-structured models.

Adland and Koekebakker (2007) present an analysis of ship valuation using cross-sectional data based on actual sale and purchase transactions in the second-hand market for

Handysize bulk carriers. This approach allows investigation of price formation in the second-market free of broker bias and measurement error. Adland and Koekebakker (2007) applies a two- and a three-factor model for the second-hand prices. A flexible non-parametric vessel valuation function allows for the presence of non-linear relationships. Their findings conclude that a three-factor model is not capable of fully explaining the second-hand prices. According to them, this is because of the exclusion of several vessel-specific variables an experienced ship broker would have included, such as engine make and other technical specifications.

Köhn (2008) addresses current issues in maritime economics by the application of semi-parametric estimations, using the GAM framework in his PhD. His empirical results confirm the findings in recent literature; that ship valuation is a non-linear function of the main drivers such as ship size, age and market conditions. The major implication is that, from an investors point of view, the application of non-linear models for asset valuation allow for a more efficient capital allocation through a much more precise price estimation. He also addresses GAM as a promising way of quantitative modelling in shipping economics.

Adland et al. (2018) expand the number of variables from the Adland and Koekebakker (2007) study. Adland et al. (2018) investigate whether the energy efficiency of vessels is reflected in sales prices in the second-hand market for Handysize bulk carriers. Mainly using linear regressions, Adland et al. (2018) add a wide variety of technical specifications, such as country of build, buyer country and the number of previous sales. Energy efficiency is identified as a significant factor in determining the second-hand prices of vessels, with higher energy efficiency increasing sales prices.

Further, Adland and Köhn (2019) propose a multivariate semi-parametric valuation model for oceangoing chemical tankers using the GAM framework. The GAM outperform linear methods of estimation in terms of explanatory power. In conclusion they found that GAM provide an appropriate framework to model highly heterogeneous assets, with new vessel-specific factors having a significant impact on prices.

Gu et al. (2018) find that machine learning simplifies the investigation into economic mechanisms of asset pricing, and highlight the value of machine learning in the field of financial innovation. As machine learning has become a more prominent tool in valuation of financial assets, it is reasonable to further examine its possible applications in the perspective of real assets, and shipping economics. In previous research, different approaches have been used in vessel valuation, with the common denominator being a predetermined model structure. So far, the application of machine learning techniques deriving relationships between independent variables from the data sample, rather than economic intuition, is not largely represented in research regarding vessel valuation. Machine learning techniques determining model structure based on the data sample has obtained good results in other fields of real asset pricing. For instance, Park and Bae (2015) utilized an adaptive boosting approach (ADABoost), to predict real estate prices. Based on the findings of Park and Bae (2015) and Gu et al. (2018), it is interesting to test a more data-driven GAM approach and a gradient boosting approach in vessel valuation.

3 Data

The Handysize bulk carrier fleet consists of 3431 ships, as of the 31st of September 2019, with 1319 different owners from 68 different countries (Clarkson Research, 2019b). There are 686 owners owning only one Handysize bulker, and 1185 owners who own five or less. In other words, the ownership is quite diverse, implying a market power dispersed between many small market players. The vessels in the fleet have an average deadweight tonnes (DWT) capacity of 27986, and an average length of 168 meters. DWT is the weight a ship can carry when loaded to its marks, including cargo, fuel, fresh water, and crew (Stopford, 2009).

The sales transactions are obtained from Clarkson Research (2019b), for the period of January 1, 1996 to September 31, 2019. In total, the dataset contains 2434 sales transactions, including the sales prices, sales dates, buyer- and seller countries and if a sale was a part of a block sale. The dataset also includes the year and month the vessel was built, building company and in which country it was built. We also have certain technical specifications and vessel specific variables, such as size (DWT), grain capacity, the design speed, fuel consumption, engine power, horsepower, fuel type, engine manufacturer, number of holds, number of hatches and gear summary (cranes and derricks).

To complement the dataset, we add two variables as proxy for market conditions in the shipping industry, and one variable for cost of capital. The first is one year timecharter (TC) rates for a standard Handysize bulker quoted at 32000 deadweight tonnes, collected from Clarkson Research (2019a). The TC rates are the average daily price, in dollars, for the hire of a vessel in a yearly perspective (Stopford, 2009). The TC rates were quoted at 28000 DWT from 1996 until April 2002, and quoted at 32000 DWT from December 2016. In between it was quoted at 30000 DWT. We choose to get the values quoted at the same DWT for the whole period, and transform the values so all of them are quoted at 30000 DWT. Even though it might not be a one-to-one relationship between the quoted size and TC, we find it more accurate to quote them at the same size.

Secondly, we add orderbook to fleet (OTF) ratio in the Handysize bulker market. This ratio accounts for newly ordered ships divided by the total fleet (Clarkson Research, 2019a). OTF is another variable to account for the "temperature" in the market. Since the TC control for market conditions for ships already existing, we add the OTF ratio to account for new entries in the market. The third variable we add to the dataset is 3-month London Interbank Offered Rate (LIBOR) as the cost of capital (Clarkson Research, 2019a). LIBOR might impact predictions, as it can affect both supply and demand of second-hand vessels. A lower LIBOR could explain an increased frequency of loans and willingness to buy vessels. On the other hand, a low LIBOR could be an indicator of low returns, and therefore decrease the willingness to buy vessels.

The rationale for including TC, LIBOR and OTF is that the information should be available at the time of a sale. TC is released first day of the week, LIBOR is released daily and OTF is released the first day of the month.

Further, the data is cleaned by removing sales with untrustworthy currency conversions, auction sales, judicial sales and block sales. A large portion of the observations were removed, as they were part of a block sale, making it difficult to assign an accurate transaction value for each vessel. Transactions with missing data for any of the variables are also omitted. A total number of 1880 transactions remain.

Following the data cleaning, we create a new variable for the age of the vessel at sale. To fully utilize the grain capacity and DWT variables, we create a new variable for cubic utilization, defined as grain capacity divided by DWT. We create a variable for total gear capacity, which is the number of gears times the capacity of each gear. We find that 91% (1711 of 1880) of the observations' engine summary is either MAN, Mitsubishi or Sulzer. Therefore, we extract those values from the column, and end up with a dummy for engine manufacturer and a category called other.

In accordance with previous research, we create a Fuel Efficiency Index (FEI) variable. FEI was first defined by Adland et al. (2017) and has the following formula:

$$FEI = \frac{Fuel\,consumption}{DWT * Speed * 24} * 10^6 \tag{3.1}$$

The FEI measures fuel consumption on a "grams per tonnemile" basis, and a low FEI indicates a fuel efficient vessel.

After the creation of new variables and data cleaning, we end up with 1880 transactions, and a total of 17 explanatory variables. See Table 3.1 for the numerical variables' and 3.2

for dummy variables' descriptive statistics. Appendix A1 shows descriptive statistics for three different time periods, as well as the whole period.

Table 3.1: Numerical variables

Variable	Minimum	Mean	Maximum
Response variable			
Price (mUSD)	0.60	8.44	49.5
Explanatory variables			
Age (at sale)	1	18	40
Deadweight tonnes	10703	28722	39814
Cubic Utilization	0.60	1.28	2.06
Speed (knots)	10	13.9	17.7
Fuel consumption (tpd)	11.5	25	58
Horsepower	3300	9022	18700
Engine RPM	92	151	1200
Number of holds	2	4.9	9
Number of hatches	2	5	12
Gear capacity (tonnes)	0	102.8	350
Fuel efficiency index	1.41	2.68	6.37
One year timecharter rates ($/Day)	4375	12207	40800
Orderbook to fleet (%)	2.28	17	49.58
LIBOR 3-month (%)	0.32	2.93	6.96

Source: Clarkson Research (2019a,b).

Table 3.2: Dummy variables' distribution

Builder country		Engine manufacturer		Fuel type	
Japan	74.2%	MAN	38.4%	HFO	84.3%
China	6.9%	Mitsubishi	25.2%	IFO	15.7%
South Korea	4.3%	Sulzer	27.5%		
Brazil	2%	Other	8.9%		
Other	12.6%				

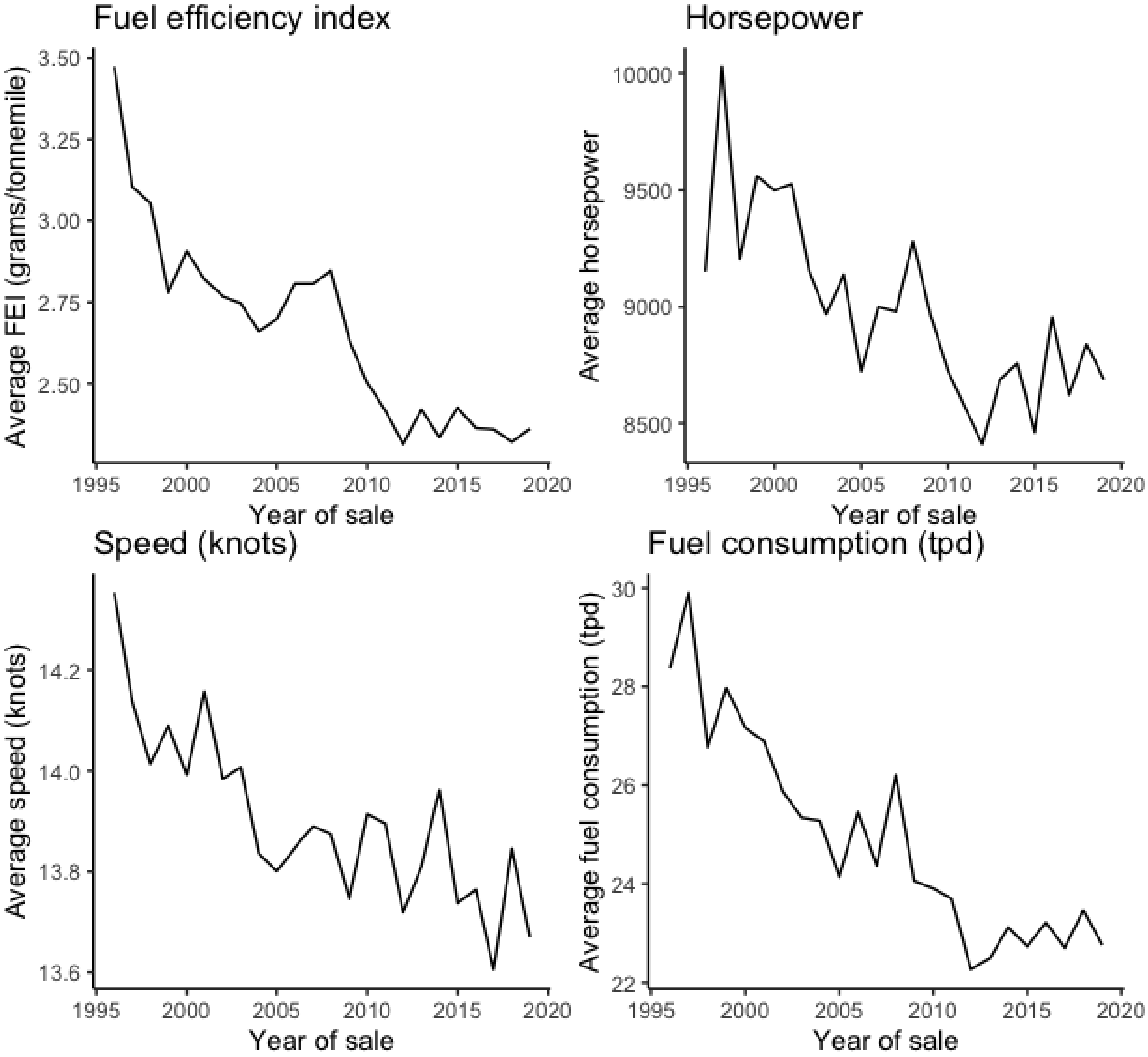
Source: Clarkson Research (2019b).

From Table 3.1, we can see that the average age at sale is 18 years, the average size of the ships is 28772 DWT, the average fuel consumption is 25 tonnes per day, and the average gear capacity is 102.8 tonnes. On average, the sales price of a Handysize bulker is $8.44m over the whole period. We can also see clear differences between the minimum and maximum values of the timecharter rates, respectively 4375 and 40800 dollars per day, averaging at 12207. An even bigger difference is observed in the orderbook to fleet ratio, with a minimum of 2.3%, average of 17% and a maximum of 49.6%. These values are good indicators of the volatility of respectively the freight market and the S&P market.

Table 3.2 presents the percentage distribution of the dummy variables in the dataset. Most of the vessels in the dataset was built in Japan (74.2%), 6.9% in China and 4.3% in South Korea. The engine manufacturers in the Handysize vessels are mainly MAN (38.4%), with Mitsubishi and Sulzer at respectively 25.2% and 27.5%. It is also eminent that most of the vessels use heavy fuel oil (HFO) with 84.3%, and a minority of 15.7% use intermediate fuel oil (IFO). To further investigate the data, it is split into groups divided by years.

Appendix A1 displays the development of the variables divided into three time periods, 1996-2003, 2004-2009 and 2010-2019, and for the full period. We also display standard deviation for sales price, $7m and median $6.25m for the whole period. From the table, we can see the price, TC and OTF were clearly highest during the 2004-2009 period. Another interesting observation from 2004-2009 is that the average DWT is lowest, at 27931, and the average vessel age at sale is highest, at 20.2 years. In addition, we can detect other interesting trends. Sales of vessels produced in China has increased from 1.4% to 13.4% from the first to the last period. Sales with intermediate fuel oil has increased from 8.2% to 25.5%, and vessels with a MAN engine are more frequently sold, increasing from 26.8% to 52.3%.

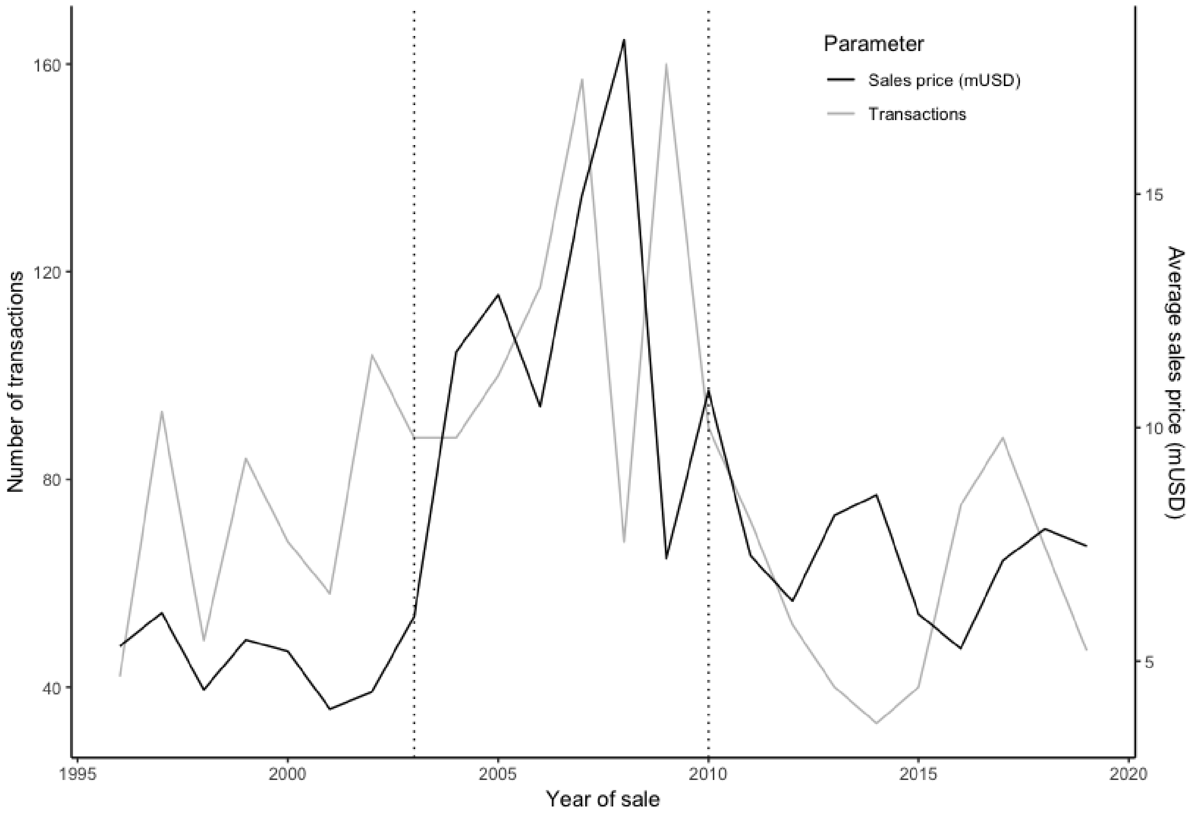
Figure 3.1: Trends in FEI, horsepower, speed and fuel consumption



Source: Clarkson Research (2019b).

Trends for a decreasing fuel efficiency index, fuel consumption, speed and horsepower are presented in Figure 3.1 on a yearly basis. Decreasing average fuel consumption, speed and horsepower indicates more fuel efficient vessels and environmental friendly ships. Reductions in speed and fuel consumption could reduce CO_2 emissions and operating costs, which is important for the environment and shipping companies (Lindstad et al., 2011). Financiers and vessel owners are increasingly conscious of the negative environmental consequences from maritime transport, and energy efficiency might be a competitive element when buying a second-hand vessel (Raucci et al., 2017). This, combined with technological enhancements, could pose an explanation of the improved fuel efficiency over time. Energy efficiency is also a significant determinant of the second-hand price, with improved energy efficiency increasing sales price (Adland et al., 2018).

Figure 3.2: Yearly average sales prices and transactions



Source: Clarkson Research (2019b).

In Figure 3.2, we can see the development of the average sales price and the number of transactions in the period of January 1996 - September 2019. The vertical dotted lines represent the years 2003 and 2009, dividing the groups as in Appendix A1. The period of 2003 to 2008 was the commodity market super cycle, and is clearly displayed in the graph (Erten and Ocampo, 2013). The average amount of yearly transactions is 78.3, ranging from 33 transactions in 2014 to 160 transactions in 2009. In the years of 2007 and 2009, the number of transactions, respectively 158 and 160, are more than doubled compared to the sample average. We can also see the impact of the financial crisis on sales prices from 2008 to 2009, changing from an average sales price of $18.3m to $7.2m.

4 Methodology

The focus of this section is explaining the different prediction methods to be applied in the analysis; generalized linear model, generalized additive model and extreme gradient boosting. These methods are used, as they are known to yield good prediction results and facilitate different degrees of model complexity (James et al., 2013).

4.1 Cross-validation and validation set

The dataset is divided into a training- and a test (validation) set, to have a basis for out-of-sample evaluation of the prediction models (James et al., 2013). Models are fitted to the training set and validated on the test set. The training set makes up 80% of the data, the test makes up the final 20%, and the observations that go in to each set is arbitrarily selected.

The training of the different models is done on the training set using k-fold cross-validation. This method divides the training set into k number of subsets, or folds, of almost equal non-overlapping size (James et al., 2013). The model is fitted k number of times, where one fold is held out as the test set, and the relevant error metrics are calculated for every held-out fold. The error metrics, explained in section 4.6, are used for optimizing the model and/or tune the model parameters. A large value for k would on average yield lower prediction error variance, but a higher bias in the different prediction errors. This is due to the trade-off between bias and variance. The higher the bias of a model, the lower its ability is to represent the relevant relationships in the data (James et al., 2013). The higher the variance of a model, the less suited a fitted model is for out-of-sample prediction. After the k-fold cross-validation is performed, a final model is returned.

The final model's fit to the training data forms the basis for calculating the training error metrics, and the predictions performed on the validation set are used to calculate the test error metrics (James et al., 2013). Another approach to cross-validation could have been leave-one-out cross-validation (LOOCV). The LOOCV approach uses a single observation (x1, y1) for the validation inside the training set, and the remaining observations (x2, y2), . . . , (xn, yn) would make up the training set. This approach demands high processing power, and would not be possible in this thesis, due to the use of complicated prediction methods. Also, the LOOCV would result in a high prediction variance, making the model less suitable for out-of-sample prediction (James et al., 2013).

4.2 Variable selection with the Boruta algorithm

To exclude unimportant variables from the prediction models, we apply the Boruta algorithm. The Boruta algorithm is applied using the R package *Boruta* (Kursa and Rudnicki, 2010), which utilizes an algorithm wrapped around the random forests algorithm from the R package *randomForest* (Liaw and Wiener, 2002). The Boruta algorithm creates a set of shadow variables in the dataset, with arbitrary values for each observation. A random forest model is fitted to predict the response variable, using the dataset with the shadow variables. After fitting the random forest model, the Boruta algorithm calculates the Z-score of each variable using the following formula:

$$Z = \frac{Average\, Loss}{Standard\, deviation\, of\, the\, variable} \tag{4.1}$$

The process of testing the Z-score for the variables is repeated until all of the variables are classified as important or unimportant. A variable is deemed important by the Boruta algorithm if the variable has a higher Z-score than any of the shadow variables (Kursa and Rudnicki, 2010).

4.3 Generalized linear model

Nelder and Wedderburn (1972) created the generalized linear model (GLM), which is a flexible generalization of a linear regression. The model allows for distributions other than Gaussian, and for a degree of non-linearity in the model structure (Wood, 2017). The structure of a GLM is:

$$g(\mu_i) = X_i \beta_i \tag{4.2}$$

In equation 4.2, $\mu_i \equiv \mathbf{E}(Y_i)$, g is a smooth monotonic "link function", X_i is the i^{th} row of a model matrix, X, β is a vector of unknown parameters, and $Y_i \sim$ $some\, exponential\, family\, distribution$ (Wood, 2017). If a Gaussian distributed GLM is assumed, the GLM is characterised by a dependent variable, whose distribution has mean $= \phi$, variance $= \sigma^2$ and $\phi = predicted\, g = \sum \hat{\beta}_i x_i$ (Nelder and Wedderburn, 1972).

In this case, equation 4.2 would look like a linear regression.

4.4 Generalized additive model

The generalized additive model (GAM), created by Hastie and Tibshirani (1986), is an extension of the GLM. The GAM fits a GLM with a linear predictor involving a smooth term which is the sum of smooth functions of explanatory variables (Wood, 2017). GAM replaces the linear form of $\sum \beta_i X_i$ by a sum of smooth functions $\sum s_i(X_i)$ (Hastie and Tibshirani, 1986). The general model is structured as follows:

$$g(\mu_i) = A_i\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) \tag{4.3}$$

where $g(\mu_i) \equiv \mathbf{E}(Y_i)$ and $Y_i \sim EF(\mu_i, \phi)$. Y_i is a response variable, $EF(\mu_i, \phi)$ denotes an exponential family distribution with mean μ_i, and scale parameter ϕ. A_i is a vector of explanatory variables for strictly parametric components, θ is the corresponding parameter vector, and f_i are smooth functions of the non-parametric model variables, x_k (Hastie and Tibshirani, 1986; Wood, 2017).

A GAM is additive because a separate f_i is calculated for each x_i and added together (James et al., 2013). The GAM is flexible, and can model non-linear relationships, making it possible to avoid specific assumptions about the functional form of these relations. Because of its additive structure, it is possible to examine effects of each x_i on g, holding the other variables constant.

Regression splines are created by specifying a set of knots, producing a sequence of basis functions and then using least squares to estimate the spline coefficients (James et al., 2013). Thin plate regression splines (TPRS) avoid the problem of subjectivity by automatizing the knot placement for the smoothing splines (Wood, 2017). TPRS are also relatively cheap to compute, and can be constructed for smooths of any number of predictor variables.

The TPRS approach is a general solution to the problem of estimating a smooth function at particular values of the predictors (Wood, 2017). Consider the problem of estimating $g(x)$ such that $y_i = g(x_i) + \epsilon_i$, where ϵ_i is a random error term and x is a d-vector. TPRS estimates g by finding the function $\hat{f}$ which minimizes

$$||y - f||^2 + \lambda J_{md}(f) \tag{4.4}$$

where y is the vector of y_i data and $f = [f(x_1), ..., f(x_n)]^T$ (Wood, 2017). $J_{md}(f)$ is a penalty function measuring the "wiggliness" of f, and λ is a smoothing parameter. λ controls the important bias-variance trade-off between smoothing the functions adequately and fitting f to the data specifically enough (Wood, 2017). In other words, λ is chosen to prevent underfitting and overfitting.

The problem of choosing the optimal smoothing parameter, λ, is solved using generalized cross-validation (GCV). GCV is the most frequently used way of determining the smoothing parameter, λ (Wood, 2017). If λ is too low, the data will be under-smoothed and if it is too high the data will be over-smoothed. In both cases, the estimate $\hat{f}$ will not be an accurate estimation of f. This means that the choice of lambda should be to make $\hat{f}$ as close to f as possible. Generalized cross-validation is used to minimize the following formula:

$$V_g = \frac{n \sum_{i=1}^{n} (y_i - \hat{f}_i)^2}{[n - tr(A)]^2} \tag{4.5}$$

Equation 4.5 returns the GCV score, which is an estimate of the mean square prediction error (Wood, 2017). The GCV approach refits the model to subsets of the data, before resampling it. Different values for λ are used, and the model with the best fit is chosen.

4.5 Extreme gradient boosting

Extreme gradient boosting (XGBoost) is an algorithm combining several weak learners, typically regression trees, to create a strong predictive model, created by Chen and Guestrin (2016). Boosting refers to the sequential ensemble of trees, where new trees are built based on residuals in previous trees (Friedman et al., 2000). Hence, the sequentially added trees are trained on the variance in the training set not adequately explained by the existing trees. New trees are added until the algorithm reaches the maximum number of trees, or the new trees do not add significant predictive power. At the beginning of the algorithm, each observation is predicted as the mean of all observed response variables, with an equal weight of $\frac{1}{N}$ (Friedman et al., 2008). The weights will be adjusted throughout every iteration. The more a prediction misses, the higher the weight of that observation in that tree, and the more the algorithm is forced to focus on this observation (Chen and Guestrin, 2016).

The process of fitting the XGBoost algorithm revolves around sequential tuning of a series of model parameters, and choosing the parameters that gives the most accurate model (Chen and Guestrin, 2016). Before explaining the effect of each parameter on the model, and in what order the parameters are tuned, it is noted that the *nrounds* (B) parameter is tested for multiple values in every tune iteration (James et al., 2013). B sets the maximum number of trees the algorithm ensembles. For each parameter tuning, the algorithm will return the optimal combination of input parameters, by minimizing the Loss function presented in equation 4.6.

$$\begin{gathered}\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \\ where\, \Omega(f) = \gamma T + \frac{1}{2}\lambda ||w||^2\end{gathered} \tag{4.6}$$

In equation 4.6, $\mathcal{L}(\phi)$ is the estimated predictive loss (l), which represents a second-degree approximation of the residuals (Chen and Guestrin, 2016). $\Omega(f)$ is a function penalizing complex models, and is meant to prevent overfitting. The number of leafs (T) in a tree is penalized through the Gamma (γ) parameter.

The first two parameters normally tuned are the learning rate (λ) and the max depth (d) of the regression trees (Friedman et al., 2008). The learning rate decides the weight placed on each sequentially added tree. A low λ will give a more robust model, as each step has less impact, but a low λ will also require a large number of trees, B, hence being computationally expensive. The max depth indicates how many levels the regression tress is allowed to have. Deep trees can lead to overfitting of the training set, as unnecessarily complex patterns can be modelled, whereas shallow trees can be excessively simplistic (James et al., 2013).

In the second tuning, *min_ child_ weight* is set, deciding the minimum required weight for all observations in a single tree (Kuhn and Johnson, 2013). A low *min_ child_ weight* can lead to overfitting, as the model can be prone for very specific trends. On the contrary, a high *min_ child_ weight* can lead to underfitting, as important trends are overlooked.

In the third tuning, *colsample_ bytree* and *subsample* is set. *colsample_ bytree* decides the fraction of variables to be sampled in a new tree (Chen and Guestrin, 2016). The *subsample* parameter decides the fraction of the training set used to train each tree. Small subsamples can lead to underfitting, and large subsamples can lead to overfitting. In the fourth tuning, Gamma (γ) is tuned. Gamma constraints the minimum effect each new tree can have on the loss function, before the algorithm is stopped (Kuhn and Johnson, 2013). In the fifth and sixth tunes, the model is sequentially tuned for learning rate and the exactly optimal maximum number of trees.

4.6 Model evaluation

To evaluate the different models' performances, the adjusted R^2, training and test root mean squared error (RMSE) and mean absolute percentage error (MAPE) are calculated. When comparing predictive models, it is advantageous to examine different error metrics, to get a nuanced basis for model evaluation.

A common measure of how well the model fits the dependent variable in the training set, is the correlation coefficient of determination, R^2 (Harel, 2009). This measure of the model fit lies between 0 and 1 (0% to 100%). Since R^2 always increases as more variables are added to the model, the adjusted R^2 is a better measure for how well the model fits the data. The adjusted R^2 will account for the number of independent variables in the model. However, the adjusted R^2 is only a good measure for how well the model fit the training data, not out-of-sample predictive power (James et al., 2013).

Mean squared error (MSE) is a measure of a models squared residuals (James et al., 2013). The root of the mean squared error (RMSE) is an error term on the same scale as the dependent variable, making it easy to interpret. The RMSE is calculated for both the training and the test set, and is an acknowledged way of testing model accuracy (James et al., 2013). The RMSE is calculated using the formula in equation 4.7.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{f}(x_i) - y_i)^2} \tag{4.7}$$

In equation 4.7, N is the number of predicted observations, y_i is the observed value and $\hat{f}(x_i)$ is the predicted value for observation i. As RMSE is calculated using squared residuals, large residuals will be heavily penalized by this error measurement. Hence, RMSE is suitable when it is appropriate to penalize a model for large residuals (James et al., 2013).

An error measurement that does not penalize large residuals as severely as RMSE is the mean absolute percentage error (MAPE), calculated using the following equation:

$$MAPE = \frac{1}{n}\sum_{i=1}^{N}\frac{|\hat{f}(x_i) - y_i|}{|y_i|} \tag{4.8}$$

where $\hat{f}$ is the model function, x_i is the list of explanatory variables (input) and y_i denotes the target variables (De Myttenaere et al., 2016). MAPE is a good error measurement when you have a varying specter of values in the response variable. As MAPE is calculated as a percentage, it takes the value of the observed response variable into account. Thus, penalization of the residuals is scaled to the observed value (James et al., 2013). Scaling can be a useful trait of the error metric when predicting a variable like price, as a $1 million residual is more significant for a price of $5 million than $50 million.

The combination of the adjusted R^2, RMSE and MAPE facilitates a nuanced and accurate basis for model evaluation.

5 Empirical Analysis

In this section, the methods described in section 4 will be applied to the dataset described in section 3, to predict second-hand Handysize bulk carrier prices. First, the Boruta algorithm will be applied to check variable importance in the independent variables. Further, we will present how the different models are fitted. Lastly, we will examine and compare the models and their results.

Variable selection is performed using the R package *Boruta* (Kursa and Rudnicki, 2010). The response variable is second-hand price, and all the predictors are used in the model with 11 iterations. The figure in Appendix A2 shows the relative importance of the predictors, where the horizontal axis represents the predictors and the vertical axis represents the variable importance. From a total of 17 independent variables, all of them are classified as important. According to Kursa and Rudnicki (2010), Boruta is a heuristic method, which seeks to find all relevant variables, including weakly relevant ones. Hence, many of the variables classified as important, might only be slightly relevant.

The Boruta algorithm's results indicate that all variables should be included in the predictions. In this thesis, the main focus is maximizing predictive power, rather than identifying inference. Therefore, we choose to initially include all the variables deemed important by the Boruta algorithm. As the Boruta algorithm is based on a random forest model, it will not be affected by problems with multicollinearity (Kursa and Rudnicki, 2010). Problems with multicollinearity could occur when two or more variables are highly correlated and account for the same variation in price. Initially, multicollinearity does not pose a problem for predictive power, but it makes interpretation of variable effects more complicated (Paul, 2006). However, including several variables that account for the same price variation can cause unstable predictions, when slight changes in the data is introduced (James et al., 2013; Paul, 2006). This will be further addressed when the GAMs are fitted.

5.1 Fitting the models

Models are created based on the logarithm of price as the response variable, mainly to avoid predicting negative prices. Another advantage of log-transforming the price is that we can assume a Gaussian distribution, as seen in the histogram of logarithmic price in Appendix A3. A response variable with a Gaussian distribution could reduce the computational cost of some models, by eliminating the need for a link function (Friedman et al., 2008). To properly scale the dollar values of the TC rates to the price variable, this variable is also transformed to the logarithmic scale.

All of the models presented will be estimated using a training set, arbitrarily including 80% of the observations. The last 20% will be used as out-of-sample validation of the models. K-fold cross-validation will be used when training the different models, as described in section 4. We use $k = 5$, as this facilitates a sensible trade-off between the prediction error bias and variance with relatively small datasets (James et al., 2013).

The first model we fit is a GLM, through the use of the *caret* R package (Kuhn et al., 2008). As the response variable is approximately distributed in a Gaussian manner, the GLM will be structured as a standard linear regression (Nelder and Wedderburn, 1972). The GLM is created as a benchmark to maximize predictive power, and is therefore fitted using all of the explanatory variables (James et al., 2013).

Next, we create a benchmark GAM using the *mgcv* package in R (Wood, 2017). When fitting the benchmark GAM, smoothing terms are created for all numeric variables in Figure 3.1, except the number of holds and hatchets, as their respective effective degrees of freedom (EDF) are too low to be smoothed. A variable's EDF is a measurement of the freedom a variable has to vary (Janson et al., 2015). The smoothing terms allow for non-linear relationships between the independent variables and price to be modelled, as explained in section 4.4. As with the GLM, the model is fit using all of the explanatory variables.

To obtain an interpretative GAM, a low degree of concurvity in the non-parametric terms and multicollinearity of the parametric terms is desirable. A high degree of concurvity for a variable indicates that its smoothing term could be approximated by one or more other smoothing terms. One could argue that predictive power is more important than variable interpretability, and that addressing multicollinearity and concurvity is a redundant procedure. However, including variables with a high degree of multicollinearity or concurvity can cause problems when the initial predictive power of the variables is unknown, and lead to the predictive model being based on faulty assumptions (James et al., 2013). Hence, addressing multicollinearity and concurvity makes the model more robust.

To ensure robustness, the model is set to penalize the smoothing terms and reduce the effect of variables which could be explained by other variables, effectively setting the effect of variables that cause a high degree of concurvity or multicollinearity to zero (Marra and Wood, 2011). A summary of the benchmark GAM is displayed in Appendix A4, indicating that the effect of the *speed* and *hp* variables are non-significant, and set to zero. The benchmark GAM is presented in the following equation:

$$\begin{aligned} g(E(price_i|.)) = \theta_0 + s(age_i) + s(dwt_i) + s(cubic_i) + s(speed_i) + s(fuelcon_i) \\ +s(hp_i) + s(rpm_i) + holds_i + hatches_i + s(gearcap_i) + s(fei_i) \\ +I_i^{builder} + I_i^{fueltype} + I_i^{engine} + s(tc1y_i) + s(otf_i) + s(libor_i) \end{aligned} \tag{5.1}$$

In the regression equation 5.1, s refers to the different smoothing terms of the numeric variables, and I indicate transformation from factorial variables to dummy variables.

To further assess the applicability of generalized additive modelling for vessel pricing, two sequentially tested interaction terms will be added to the GAM. In the same way as the benchmark GAM, an iterated GAM is fitted using the *mgcv* package in R (Wood, 2017). For the interaction terms, smoothing functions are created to model influence on price for all possible combinations of the two variables. Given the importance of *tc1y* and *age* in the Boruta plot in Appendix A2, these variables will make up the first factors of two created interaction terms. For the second factors in each of the interaction terms, all 14 of the numerical variables are tested, except for the variable making up the first factor of the term. For an overview of the numeric variables, see Table 3.1. The approach iterates through 169 models.

For each GAM, the cross-validated RMSE is returned and used to compare the model to the currently best performing model. As the cross-validated training error is a suitable proxy for out-of-sample prediction error, the cross-validated RMSE is used for model selection. If the training error calculated on the final model was used in model selection, instead of the cross-validation error, the method could be prone to choosing a model that overfit the data. The currently best performing model is saved, and replaced if a new combination of interaction terms yields a lower cross-validated RMSE. As the interaction terms are created by iterating through the variables, this is a highly data-driven approach to fit a GAM. The optimal interaction terms found are $tc1y * age$ and $age * dwt$.

One could argue that iterating though 169 models is excessive to find these interaction terms, as it could make intuitive sense to combine two important variables in $tc1y*age$ and a vessel-specific variable, dwt, with the age in $age * dwt$. However, this iterative approach is based on the assessment of predictive power rather than an educated judgement. This method could potentially uncover relationships between different variables not represented in conventional maritime economics.

Lastly, we fit an extreme gradient boosting model. As an XGBoost model is free to use any variable as much, or as little, as appropriate, this method is not affected by multicollinearity of variables, and produces a robust model (Ding et al., 2016). The sequential parameter tuning, with k-fold cross-validation, described in section 4.5 is applied to the training set, using the *caret* R package (Kuhn et al., 2008). The parameters selected are presented in Table 5.1.

Table 5.1: Parameter values - XGBoost

Parameter	B	d	λ	γ	colsample_bytree	min_child_weight	Subsample
Value	649	4	0.025	0.05	1	3	0.5

As seen in Table 5.1, the model uses 649 regression trees, each having a relatively low impact on the prediction, due to the low learning rate of 0.025. The values for *subsample* and *colsample_ bytree*, of respectively 0.5 and 1, indicate that 50% of the observations are used to train each tree, and all observations are eligible for use in each tree. The possibility to use any variables in all trees could be advantageous, as the Boruta algorithm uncovered a large difference in relative variable importance. A γ value larger than zero, and the relatively high *min_ child_ weight* value of 3, could balance out the maximum tree depth of 4, which makes for a pretty complex model (Chen and Guestrin, 2016). Overall, the combination of parameters could facilitate a balanced degree of model complexity (Kuhn and Johnson, 2013; Chen and Guestrin, 2016).

5.2 Evaluation of models

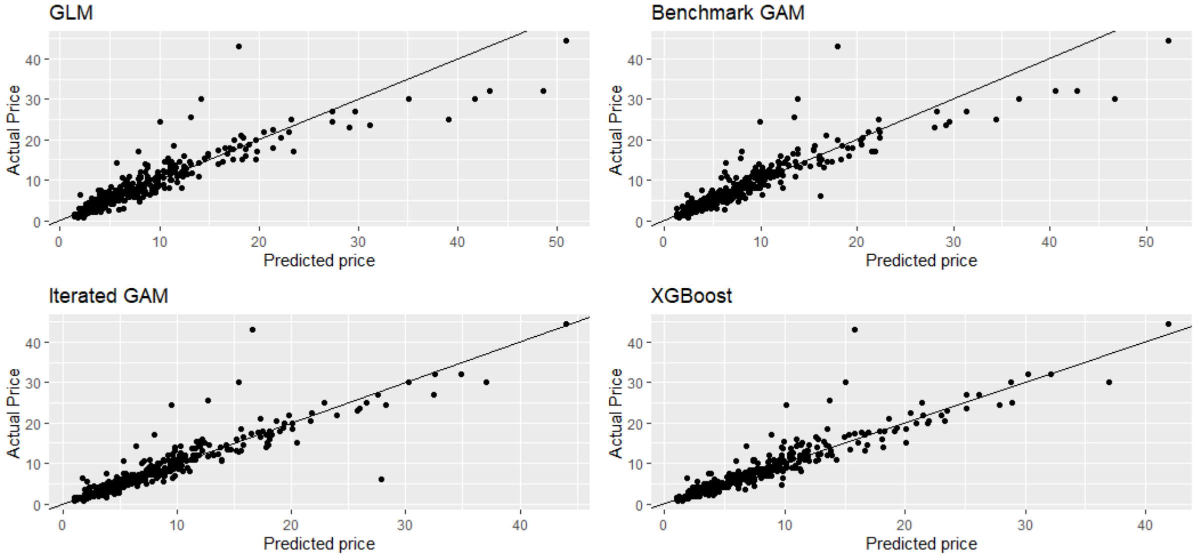
After fitting the models, their predictive performance are examined and compared. First, we compare the GLM to the benchmark GAM, then the benchmark GAM to the iterated GAM, and lastly we compare the iterated GAM to the XGBoost. The models' evaluation metrics are shown in Table 5.2, calculated as explained in section 4.6. As the *price* variable is log-transformed, the adjusted R^2 is manually calculated for how well the model explains deviance in the exponentially back-transformed *price* variable.

Table 5.2: Evaluation metrics

Evaluation metrics	GLM	Benchmark GAM	Iterated GAM	XGBoost
Adj. R^2	84.27%	88.06%	91.18%	89.9%
Train RMSE	2.8	2.44	2.1	2.24
Test RMSE	2.95	2.88	2.7	2.48
Train MAPE	19.95%	16.33%	14.73%	16.14%
Test MAPE	21.26%	18.74%	18.31%	16.6%

To visualize the out-of-sample predictions, the fitted prices for the test set are plotted against the actual prices in Figure 5.1. A 45° line is added to the plots, representing a perfect prediction. The closer an observation is to the 45° line, the better the prediction is.

Figure 5.1: Fitted- versus actual values in the validation set



Source: Clarkson Research (2019a,b).

As seen in Table 5.2, the adjusted R^2 for the GLM is 84.27%, indicating that the model explains about 84% of the variance in the training set observations' prices. The GLM's train- and test MAPE are respectively 19.95% and 21.26%. The RMSE is 2.8 for the training set and 2.95 for the test set. So far, the GLM seems to fit and predict the data in a reasonable manner, as the test errors are slightly higher than the training errors. Test errors larger than training errors, but not by a large margin, indicates a good trade-off between overfitting and underfitting the training data (James et al., 2013). If the test error measurements are very high compared to the training measurements, it might indicate that the model is overfitted to the training set, and that the model is not suitable for out-of-sample predictions.

From Figure 5.1 it appears that GLM does not capture certain relationships in the data, as the price of vessels costing over $20 million are generally being predicted too high or too low. Therefore, it is interesting to compare the model with GAMs, as they facilitate modeling of non-linear relationships.

The benchmark GAM seems to do an even better job than the GLM at both fitting the training data and predicting the test set. The train and test RMSE and MAPE are all lower than those of the GLM. The adjusted R^2 is higher, and indicates that the fitted model accounts for approximately 88% of the variation in price in the training set. Contradictory to the GLM, the benchmark GAM seems to detect some of the characteristics for the higher priced ships, as seen in Figure 5.1. However, prices over $20m still seem to be predicted somewhat low or high. The slight improvement of the GAM compared to the GLM could be explained by the modeling of non-linear relationships.

In Appendix A4, the EDF reflects the degree of non-linearity in the predictors in the GAMs, and the significance of these. The values are calculated to a null hypothesis of a linear relationship versus the alternative of a non-linear relationship for each smoothed variable (James et al., 2013). From Appendix A4, we can see that *speed* and *hp*, are the only variables with weak evidence for non-linearity and significance in the benchmark GAM. This implies that the model has deemed these variables as redundant, and set their effect on price to zero, indicating how the GAMs automatically addresses concurvity and multicollinearity. All the remaining smoothed variables have significantly non-linear relationships with price.

Further, we will examine the iterated GAM, and compare it to the benchmark GAM. For the iterated GAM, all the error metrics are lower than those of the benchmark GAM. As explained in section 4.6, the MAPE metric provides a good understanding of prediction error, as it is scaled to the observed values. The training and test MAPE, of respectively 14.73% and 18.31%, indicate that the training set is a lot more accurately predicted than the test set. The increased divergence between test and train error metrics, compared to benchmark GAM, could imply that the interaction terms leads the model to slightly overfit the training set. However, as the iterated GAM has lower train and test error metrics than the benchmark GAM, it seems to be better suited to fit the data and perform out-of-sample predictions.

As the iterated GAM has yielded the best results this far, it is interesting to compare it to the XGBoost approach. The test error metrics of the XGBoost model are slightly higher than the training error metrics, showing no clear indications of overfitting or underfitting. The test RMSE and test MAPE are both lower for the XGBoost model than for the other models. Although the training error measurements of the XGBoost is slightly higher than those of the iterated GAM, the XGBoost seems to be best suited for out-of-sample prediction.

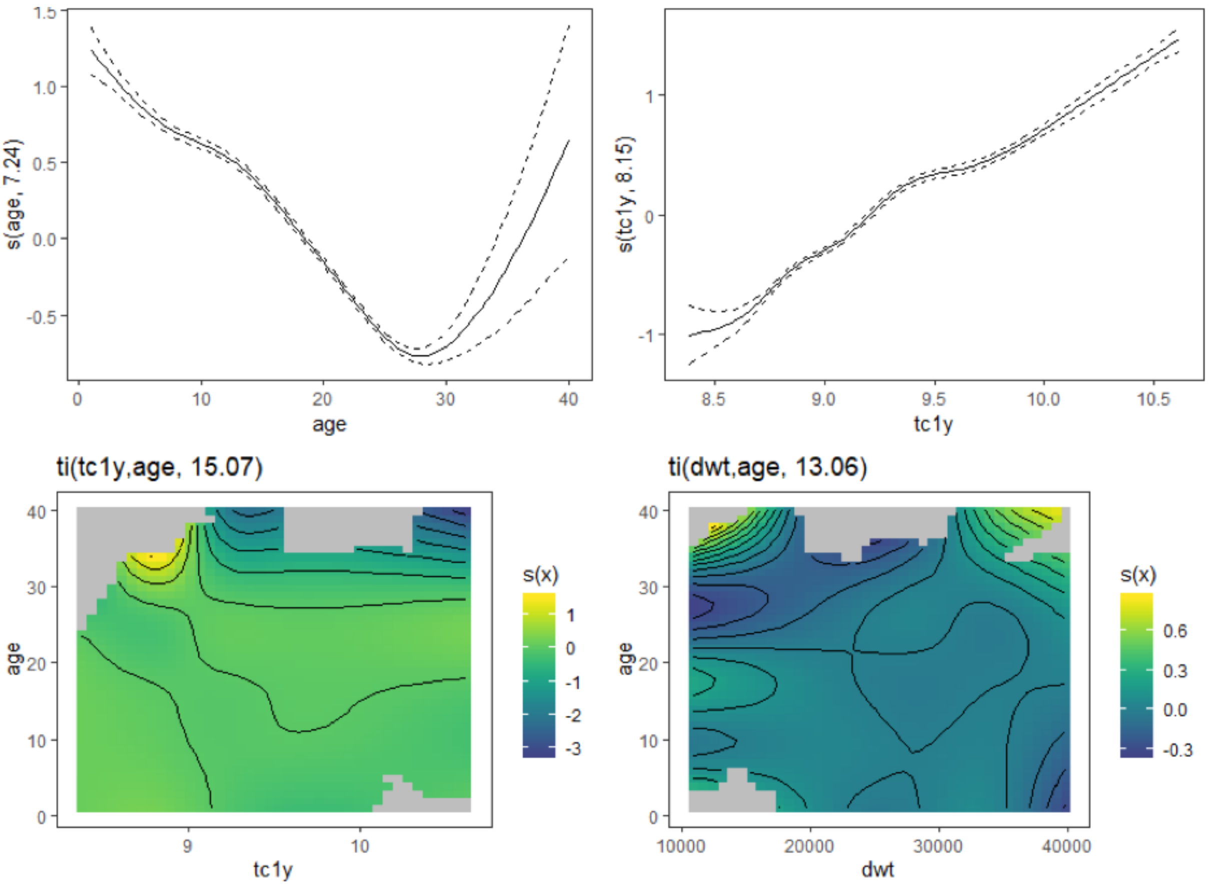
The application of precise desktop ship valuation can be highly useful for investors, ship owners or other market players in the maritime industry, with more accurate prediction leading to better informed decision making. As the average transaction value in the dataset is \$8.44m, increasing price prediction accuracy with a few percentage points could have a large economic impact. Due to the volatile S&P market for vessels, increased prediction accuracy could be important to facilitate well-timed sales or purchases.

In previous research, the GAM approach has yielded good results in desktop valuation of chemical tankers, and modelling shipping markets, mainly due to its flexibility (Köhn, 2008; Adland and Köhn, 2019). One of the major differences between the XGBoost approach and the GAM approach is that an XGBoost model makes predictions based on a series of regression trees, enabling complex relationships between all the variables in the dataset to be modeled. Where the GAM is limited to a set number of interaction terms, the XGBoost model can create decision rules for multiple variables in any branch of any tree, as well as model non-linearity in each separate variable. As the GAM is additive in nature, it cannot capture non-linear relationships between variables not included in interaction terms.

Application of XGBoost in vessel valuation could for instance allow the effect vessel-specific variables to be modeled differently, based on a set of market variables or other vessel-specific variables. This attribute of the XGBoost algorithm can be beneficial, as certain variables could affect the influence of other variables. Our analysis indicates that the XGBoost algorithm's ability to model complex relationships in the data makes the algorithm suitable for application in desktop valuation, as out-of-sample prediction accuracy is 1.7 percentage points better than the most accurate GAM.

Where the GAM arguably is advantageous to the XGBoost, is in interpretability of different variables' effect on the response variable (James et al., 2013). Because of the additive structure of the GAM, a variable's effect on price can be examined, holding all other variables constant. The smoothing plots from the iterated GAM facilitate interpretation of a single smoothed variable's effect on the price, whereas the XGBoost models a variable's effect in different trees, and in combination with other variables. Therefore, the iterated GAM will be the main focus when examining variable effects on price. An interpretation of the GAM's variables will not account for the complex relationships between variables modeled in the XGBoost model. Rather, it will give a general overview of impacts on price from specific variables in the training set. In Figure 5.2, smoothing plots for age, $tc1y$ and the interaction terms for $tc1y * age$ and $age * dwt$ in the iterated GAM are presented.

Figure 5.2: Smooth of tc1y, age, $tc1y * age$ and $age * dwt$



Source: Clarkson Research (2019a,b).

The black lines in Figure 5.2 for $tc1y$ and age in the two upper plots represent the smoothed variables, and the dotted lines are the confidence bands. The confidence bands indicate more uncertainty regarding the effect of smaller and higher values, mainly due to fewer observations. Relationships with second-hand price for age and $tc1y$ are strongly non-linear, as indicated in the plots and Appendix A4 from their high EDF and significance.

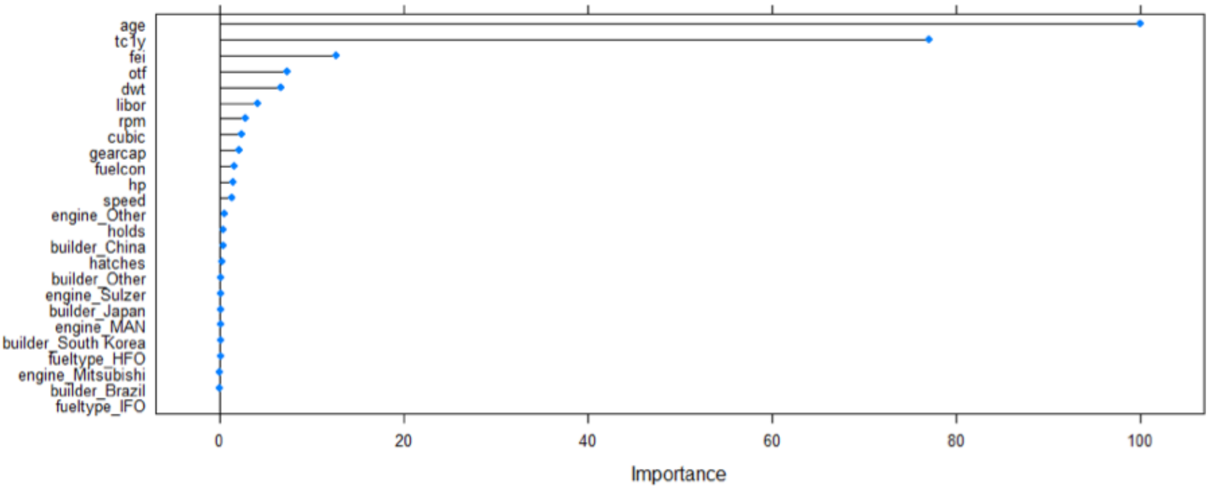
The positive influence of $tc1y$ on price could be explained by $tc1y$ containing a lot of information about possible future income. If the valuation was done through an income- or discounted cash flow approach, TC rates would probably be one of the best indicators for future income (Stopford, 2009). As *tc1y* accounts for the state of the freight market, mainly in the short term, a high *tc1y* indicates positive expectations for future earnings.

Vessel age at sale affects the second-hand price negatively until a certain point, around age 30, as seen in Figure 5.2. The *age* curve has larger confidence bands after the shift, indicating an increased degree of uncertainty. Depreciation reflects the loss of performance, higher maintenance, technical obsolescence and expectations about the economic life of an ageing vessel (Stopford, 2009). So far, the findings regarding the effect *age* has on price does not coincide with existing literature, as real assets are expected to depreciate at a constant or varying rate (Hulten and Wykoff, 1980). However, to fully explain the *tc1y* and *age* effects in the iterated GAM, they must be examined along with the interaction terms.

The two lower plots in Figure 5.2 show the interaction terms $tc1y \ast age$ and $age \ast dwt$. The scale of impact on price for each interaction term is shown on their right side. From the $tc1y \ast age$ plot, we can see that high values for vessel age mostly affect the price negatively, with only a few combinations of *tc1y* and *age* having a positive effect. The $tc1y \ast age$ seems to function as a counterweight for the *age* variable, as the interaction's effect on price is mainly negative for *age* over 30. From the interaction between $age \ast dwt$, it appears that small vessels and bigger vessels with an old vessel age affect the price positively, but at a low rate.

After examining different variables' impact on price in the iterated GAM, the relative importance of the variables in the XGBoost model are presented in Figure 5.3. Although the complexity of the XGBoost model makes it difficult to interpret the full effect of a variable on price, relative importance indicates which variables cause the model's high predictive power.

Figure 5.3: Variable Importance - XGBoost



Source: Clarkson Research (2019a,b).

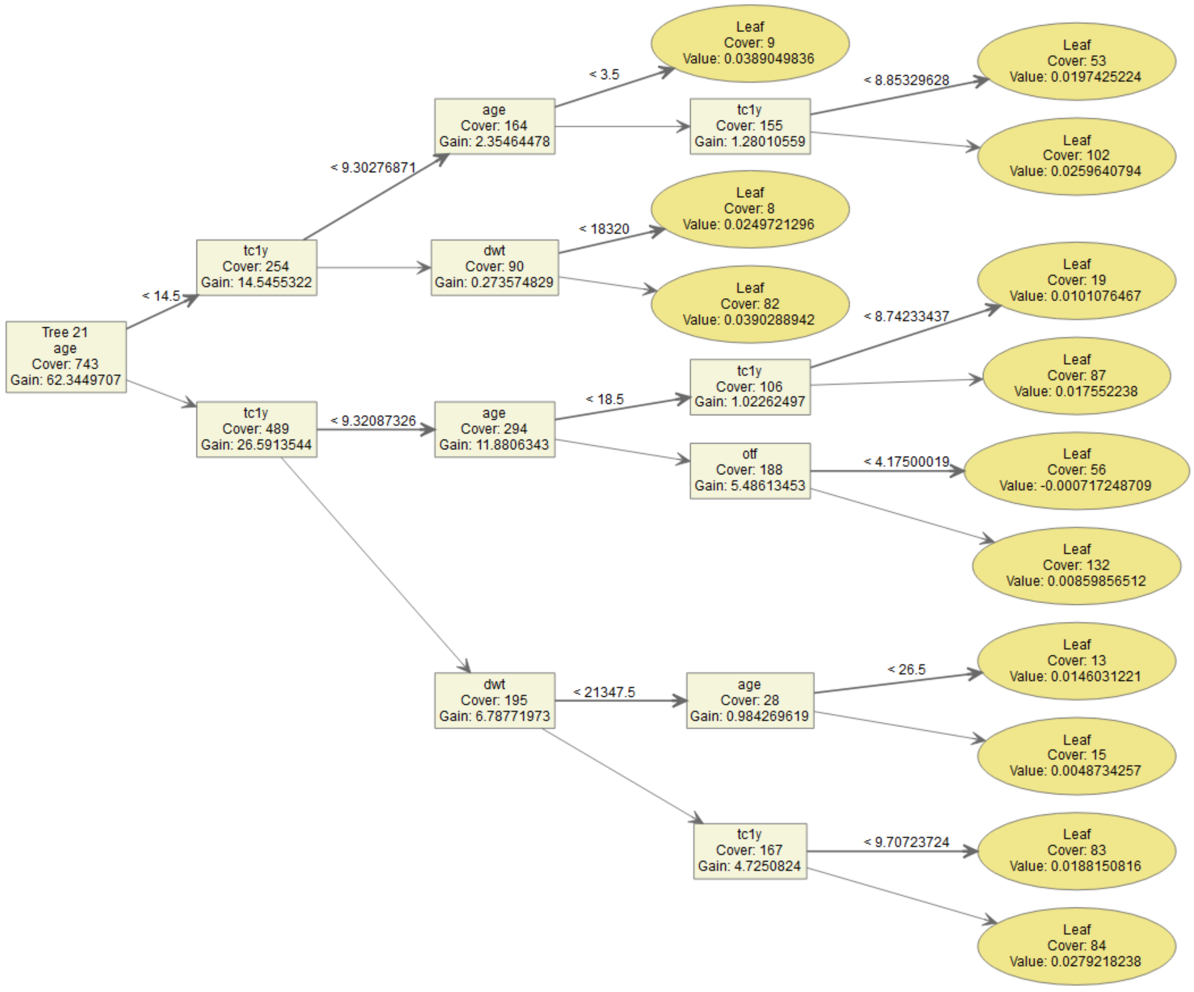
Figure 5.3 implies that *age* and *tc1y* are the most important variables, followed by *fei, otf* and *dwt*. In *Maritime Economics*, Stopford (2009) identified age and timecharter as two of the main influences of second-hand vessel prices, supporting the legitimacy of the XGBoost results. The importance of these variables is also coherent with the proposition of Beenstock (1985), who argued that vessel prices should be priced using a framework based on portfolio theory. If a ship is viewed as a capital asset in a portfolio, the future income creating higher return on investment and duration of the asset are key components in vessel pricing (Bodie et al., 2014). The importance of the depreciation from *age* and increasing price with *tc1y* are coherent with previous research regarding desktop valuation and shipping economics (Adland and Koekebakker, 2007; Adland and Köhn, 2019; Stopford, 2009).

The fuel efficiency index proposed by Adland et al. (2017) is the third most important variable in the XGBoost model. Adland et al. (2018) states that the fuel efficiency of vessels could be more beneficial when freight rates are low, due to potential cost savings. As the XGBoost model is based on a series of regression trees, such a relationship between variables can easily be modelled, emphasizing the applicability of the XGBoost algorithm for desktop vessel valuation.

Interestingly, *engine, builder* and *fueltype* are relatively unimportant in valuation of vessels. An initial expectation was that these variables would be more important when predicting the second-hand price. However, for *fueltype* this could be explained by 84.3% of the vessels sold using heavy fuel oil, as seen in Appendix A1. Hence, the relatively low importance of *fueltype* could be interpreted as a result of a homogeneous data basis, rather than lack of general predictive power. The same logic applies to *builder*, as most of the vessels (74.2%) are built in Japan. On the other hand, the *engine* variable has a more diverse data basis. Given the heterogeneous distribution of the *engine* variable, the relatively low importance could be interpreted as a lack of predictive power.

To improve the understanding of the tree based XGBoost model, we plot one of the regression trees in Figure 5.4. It is important to emphasize that this tree is only one of 649 trees, and that other variables will be included in other trees. However, the plotted tree gives a good indication of how the XGBoost model makes predictions. The boxes illustrate decision nodes, and the circles illustrate the tree's leafs. Starting from the left node, each variable proceeds to one of two branches, based on *age*. The two next nodes have decision rules regarding the *tc1y*. Depending on of what branch an observation follows, *age, tc1y, otf* and/or *dwt* are tested. The *Value* in each leaf is multiplied with the λ and added to the predictions of the variables in that leaf. This is repeated for every tree created by the algorithm, making up the total prediction. As *subsample* is 0.5, 50% of the variables will be sorted in the different leafs of a tree.

Figure 5.4: Regression tree - XGBoost



Source: Clarkson Research (2019a,b).

Although the overall predictive results seemingly implies good predictive accuracy, particularly for the XGBoost model, the plots of the fitted test set prices against the actual prices in Figure 5.1 indicate that all the models seem to have problems with a set of observations systematically being predicted too low. These observations start at an actual price of around $20 million, and the predictive error seems to increase as the actual value goes up, with a relatively similar deviation pattern for all the models. This might imply there are drivers of price not captured in the data.

To further examine the increase in predictive error for transactions over $20 million, the fitted training set prices are plotted against actual prices in Appendix A5. Similarly to the prediction of the test set values, the fitted values for certain observations in the training priced over $20 million are too low. Apparently, the systematic errors found in the test set predictions are also present in the training set. This trend could be present as a consequence of certain ship characteristics or market conditions not being represented in the dataset. For instance, Pruyn et al. (2011) argues for new-building price level as a considerable explanatory variable to use in second-hand price models. This is an example of a variable that could counteract the systematic prediction error occurring for vessels priced over $20 million.

Another explanation of the increase in predictive error for observations with prices over $20 million could be the lack of observations in this price range. Only 8% (149 of 1880) of the sales transactions are priced over $20 million. Looking further into those 149 observations, 89% of them were in the super cycle years from late 2003 to early 2009. As prices deviate more from the average sales price of $8.44m, and the amount of observations decreases with increased price, it is reasonable that those observations will be harder to predict. The scarce data basis for higher priced vessels can make it challenging to train the models and recognize certain characteristics of those ships.

A third explanation could be irrational market behavior. At times of extreme growth of prices in a commodity market, market players tend to invest on the basis of earlier price increase, and not on the fundamental information (Szyszka, 2010). These irrational investors will increase divergence from the true value of the asset. Since most of the transactions over $20m was during a super cycle, this could offer an explanation for why prices peaked in this period, hence explaining why our models predict the price too low.

One solution to this problem of lower predictive power for transactions over $20m could be to remove the whole super cycle period from our dataset. We did not want to do this because of the importance of being able to predict prices, also during a booming period. Also, to justify systematically omitting a set of variables based on time period, they should be considered as error inputs, not observations with unusual values (James et al., 2013). By removing these observations, we would indirectly assume an ability to predict a super cycle and its crash, which is unlikely (Van den Berg et al., 2008).

6 Conclusion

The main objective of this thesis has been to compare extreme gradient boosting to generalized additive models in desktop valuation of second-hand Handysize bulkers. We found evidence that the XGboost algorithm improves desktop valuation accuracy, compared to the GAM approach. The XGBoost yielded the best predictive accuracy, on average missing by 16.6% on out-of-sample predictions, a 1.7 percentage points decrease from the data-driven GAM's test mean absolute percentage error.

The XGBoost approach contributes to research on Handysize vessel valuation, mainly by facilitating complex variable relationships in a substantial degree compared to methods applied in previous research. However, the GAMs arguably facilitate variable inference in a better way than XGBoost, making them valuable for interpreting variable effects. An interesting aspect of the analysis was the systematic increase in predictive error for transactions over $20 million. As 89% of these transactions occurred between 2003 and 2009, these findings could emphasize the difficulty of predicting vessel prices during a super cycle.

We have found age at sale, timecharter rates and fuel efficiency index to be the three most important variables in valuation of second-hand prices, using the XGBoost algorithm. These findings coincides with previous research in maritime economics (Adland and Koekebakker, 2007; Stopford, 2009; Adland et al., 2018). In addition, the effective degrees of freedom and significance of variables found in the GAMs showed evidence of linear models to be insufficient in vessel valuation.

To sum up - the use of several vessel-specific variables and market variables yields promising prediction results. Notably, the XGBoost algorithm provide an appropriate framework in desktop valuation of second-hand Handysize bulk carriers. The application of the XGBoost algorithm in vessel valuation could be advantageous for investors, ship owners and other market players in the maritime industry.

Although our thesis presented a prediction model that yields promising results, it has certain limitations. The transaction data was collected from Clarkson Research. As the data basis is not cross checked with other data sources, there is a possibility of typing errors and incorrect values to be included in the data, potentially causing wrongful conclusions to be drawn. Also, as the block sales transactions were omitted, certain mechanisms in the S&P market might not have been addressed.

The iterated GAM presented was limited to testing two interaction terms, which might not have been sufficient to capture all relevant relationships between the independent variables. Also, the interaction terms in the iterated GAM were based on variable importance found in the Boruta algorithm. It is no guarantee that the important variables found by the Boruta algorithm are equally important in the GAMs, and testing more interaction combinations could be beneficial. The tested parameters in the tuning of the XGBoost model were also limited, due to computational cost. Although the process of parameter tuning the XGBoost model was done with caution, a more thorough tuning of the parameters could improve predictive results.

In future research, we recommend testing other, and maybe more complex, machine learning techniques. It would be interesting to see research papers using the XGBoost algorithm in other shipping sizes and classes. We also recommend experimenting with adding more variables, for instance delivery time for a newly built ship, a proxy for new building vessel price and the relationship between sellers and buyers. Adding new variables, and possibly excluding some of the variables used in this thesis, might lead to even more accurate desktop valuations.

References

Adland, R., Alger, H., Banyte, J., and Jia, H. (2017). Does fuel efficiency pay? empirical evidence from the drybulk timecharter market revisited. *Transportation Research Part A: Policy and Practice*, 95:1–12.

Adland, 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*, 111:347–359.

Adland, R. and Koekebakker, S. (2004). Market efficiency in the second-hand market for bulk ships. *Maritime Economics & Logistics*, 6(1):1–15.

Adland, R. and Koekebakker, S. (2007). Ship valuation using cross-sectional sales data: A multivariate non-parametric approach. *Maritime Economics and Logistics*, 9(2):105–118.

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.

Beenstock, M. (1985). A theory of ship prices. *Maritime Policy and Management*, 12(3):215–225.

Beenstock, M. and Vergottis, A. (1989a). An econometric model of the world market for dry cargo freight and shipping. *Applied Economics*, 21(3):339–356.

Beenstock, M. and Vergottis, A. (1989b). An econometric model of the world tanker market. *Journal of Transport Economics and Policy*, pages 263–280.

Beenstock, M. and Vergottis, A. (1993a). *Econometric modelling of world shipping.* Springer Science & Business Media.

Beenstock, M. and Vergottis, A. (1993b). The interdependence between the dry cargo and tanker markets. *Logistics and Transportation Review*, 29(1):3.

Bodie, Z., Kane, A., and Marcus, A. J. (2014). *Investments, 10th edition.* McGraw-Hill, New York.

Charemza, W. and Gronicki, M. (1981). An econometric model of world shipping and shipbuilding. *Maritime Policy & Management*, 8(1):21–30.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

Clarkson Research (2019a). Shipping intelligence network database. http://www.sin.clarksons.net.

Clarkson Research (2019b). World fleet register database. http://www.clarksons.net/wfr.

De Myttenaere, A., Golden, B., Le Grand, B., and Rossi, F. (2016). Mean absolute percentage error for regression models. *Neurocomputing*, 192:38–48.

Ding, C., Wang, D., Ma, X., and Li, H. (2016). Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees. *Sustainability*, 8(11):1100.

Erten, B. and Ocampo, J. A. (2013). Super cycles of commodity prices since the mid-nineteenth century. *World Development*, 44:14–30.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2000). Additive logistic regression: a statistical view of boosting. *The annals of statistics*, 28(2):337–407.

Glen, D. R. (1997). The market for second-hand ships: Further results on efficiency using cointegration analysis. *Maritime Policy and Management*, 24(3):245–260.

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

Hale, C. and Vanags, A. (1992). The market for second-hand ships: some results on efficiency using cointegration. *Maritime Policy and Management*, 19(1):31–39.

Harel, O. (2009). The estimation of r 2 and adjusted r 2 in incomplete data sets using multiple imputation. *Journal of Applied Statistics*, 36(10):1109–1118.

Hastie, T. J. and Tibshirani, R. (1986). Generalized additive models. *Statistical science*, 1(3):297–318.

Hulten, C. R. and Wykoff, F. C. (1980). *The measurement of economic depreciation*. Citeseer.

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

Janson, L., Fithian, W., and Hastie, T. J. (2015). Effective degrees of freedom: a flawed metaphor. *Biometrika*, 102(2):479–485.

Kavussanos, M. G. (1996a). Comparisons of volatility in the dry-cargo ship sector: Spot versus time charters, and smaller versus larger vessels. *Journal of Transport Economics and Policy*, 30:67–82.

Kavussanos, M. G. (1996b). Price risk modelling of different size vessels in the tanker industry using autoregressive conditional heteroskedastic (arch) models. *Logistics and Transportation Review*, 32(2):161–176.

Kavussanos, M. G. (1997). The dynamics of time-varying volatilities in different size second-hand ship prices of the dry-cargo sector. *Applied Economics*, 29(4):433–443.

Kavussanos, M. G. and Alizadeh, A. H. (2002). Efficient pricing of ships in the dry bulk sector of the shipping industry. *Maritime Policy & Management*, 29(3):303–330.

Kuhn, M. et al. (2008). Building predictive models in r using the caret package. *Journal of statistical software*, 28(5):1–26.

Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*, volume 26. Springer.

Kursa, M. B. and Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, 36(11):1–13.

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

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Lindstad, H., Asbjørnslett, B. E., and Strømman, A. H. (2011). Reductions in greenhouse gas emissions and cost by shipping at lower speeds. *Energy Policy*, 39(6):3456–3464.

Marra, G. and Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, 55(7):2372–2387.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Park, B. and Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data. *Expert Systems with Applications*, 42(6):2928–2934.

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

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.

Raucci, C., Prakash, V., Rojon, I., Smith, T., Rehmatulla, N., and Mitchell, J. (2017). Navigating decarbonisation: An approach to evaluate shipping's risks and opportunities associated with climate change mitigation policy. *UMAS: London, UK.*

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

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

Szyszka, A. (2010). Behavioral anatomy of the financial crisis. *Journal of Centrum Cathedra*, 3(2):121–135.

Thalassinos, E. I. and Politis, E. (2014). Valuation model for a second-hand vessel: Econometric analysis of the dry bulk sector. *Journal of Global Business and Technology*, 10(1).

Tsolakis, S. D., Cridland, C., and Haralambe, H. E. (2003). Econometric modelling of second-hand ship prices. *Maritime Economics and Logistics*, 5(4):347–377.

Van den Berg, J., Candelon, B., and Urbain, J.-P. (2008). A cautious note on the use of panel models to predict financial crises. *Economics Letters*, 101(1):80–83.

Warren, C., Elliott, P., et al. (2005). The valuation profession in australia: Profile, analysis and future directions. *Australian Property Journal*, 38(5):362.

Wood, S. (2017). *Generalized Additive Models : An Introduction with R, Second Edition.* Chapman & Hall / CRC Press, London.

Appendix

A1 Descriptive statistics by yearly periods

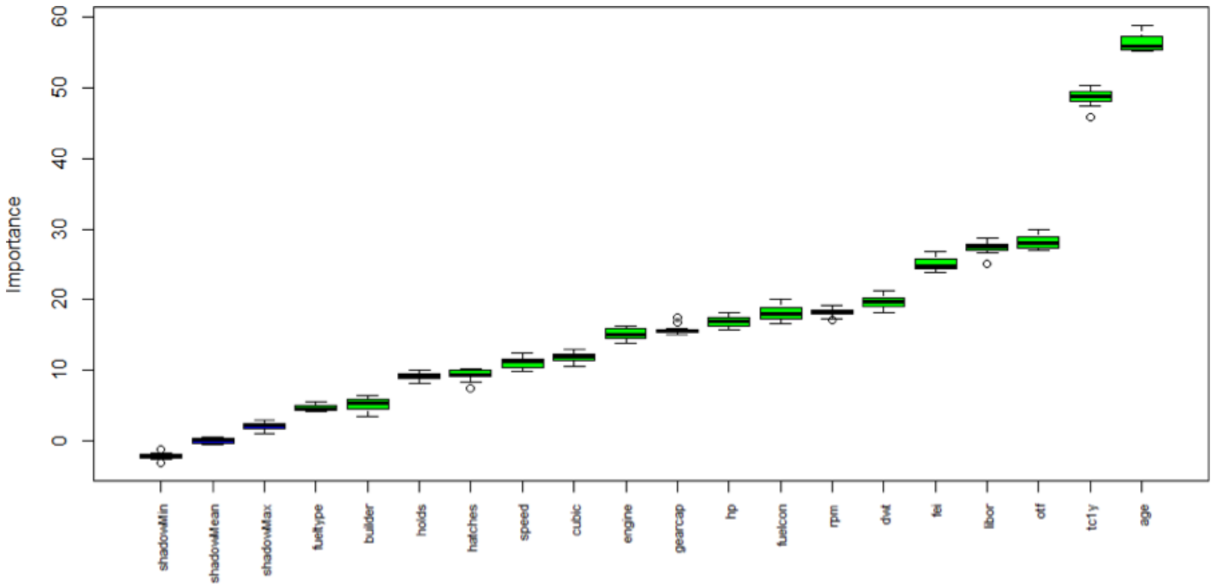
Table A1.1: Descriptive statistics by year

Vessel characteristics	Variable name	1996-2003	2004-2009	2010-2019	All
Sales price (million USD)	price				
Average		5.15	12	7.56	8.44
Standard deviation		3.7	8.9	4.9	7
Median		4.5	10	6.35	6.25
Explanatory variables					
Age at sale (years)	age	17.2	20.2	16.3	18
Size (DWT)	dwt	28592	27931	29751	28722
Cubic utilization	cubic	1.27	1.27	1.3	1.28
Gear capacity (tonnes)	gearcap	94.9	99.4	114.3	102.8
Speed (knots)	speed	14.1	13.8	13.8	13.9
Fuel consumption (tpd)	fuelcon	27.2	24.7	23.1	25
Engine (rpm)	rpm	160.2	152.6	139.2	151
Horsepower	hp	9404	8993	8685	9022
Number of holds	holds	4.9	4.8	4.9	4.9
Number of hatches	hatches	5.1	4.9	4.9	5
FEI (grams per tonnemile)	fei	2.9	2.7	2.4	2.7
Market variables					
Timecharter (1 year)	tc1y	7315	18767	9461	12207
Orderbook / fleet (%)	otf	3.93	26.48	18.84	17
LIBOR 3M (%)	libor	4.27	3.43	1.07	2.93
Builder country (%)	builder				
Japan		75.8	74.2	72.7	74.2
China		1.4	5.9	13.4	6.9
South Korea		3.6	3.5	5.8	4.3
Brazil		3.4	2.3	0.2	2
Other countries		15.8	14.1	7.9	12.6
Fuel type (%)	fueltype				
HFO		91.8	86.5	74.5	84.3
IFO		8.2	13.5	25.5	15.7
engine manufacturer (%)	engine				
MAN		26.8	35.9	52.3	38.4
Mitsubishi		18.4	24.4	32.8	25.2
Sulzer		39.3	30.9	12.1	27.5
Other manufacturers		15.5	8.8	2.8	8.9
Number of observations		586	690	604	1880

Source: Clarkson Research (2019a,b).

A2 Boruta model

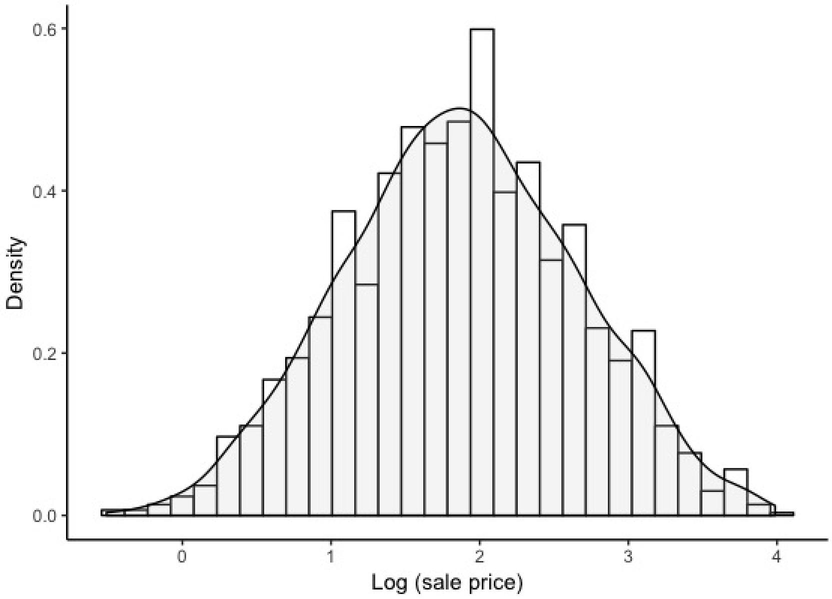
Figure A2.1: Boruta model



Source: Clarkson Research (2019a,b).

A3 Histogram of logarithmic sales price

Figure A3.1: Histogram of logarithmic sales price



Source: Clarkson Research (2019b).

A4 Smoothed terms for GAMs

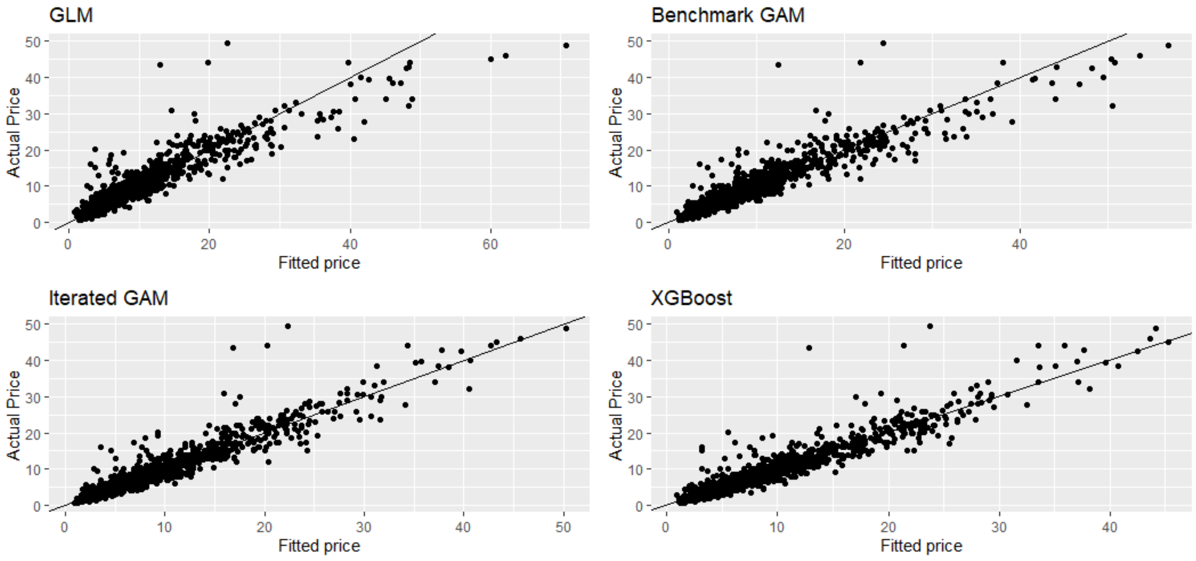
Table A4.1: Smoothed terms benchmark- and iterated GAM

	Benchmark GAM		Iterated GAM	
Variable	EDF	Significance	EDF	Significance
age	5.470	***	7.531	***
speed	7.859e-06		1.954e-06	
gearcap	7.437	**	7.544	**
tc1y	8.164	***	8.695	***
libor	6.812	***	6.259	***
rpm	6.823	***	3.332	***
otf	7.997	***	7.974	***
fuelcon	7.202	***	6.820	*
hp	7.234e-01		2.514e-06	
dwt	3.357	***	4.260	***
cubic	6.832	***	6.462	***
fei	7.377	***	7.121	***
tc1y * age			1.504e+01	***
age * dwt			1.369e+01	***

Source: Clarkson Research (2019a,b).

A5 Fitted- versus actual prices for the training set

Figure A5.1: Fitted- versus actual prices for the training set



Source: Clarkson Research (2019a,b).