New Insights on Loss Given Default for Shipping Finance: Parametric and Non-Parametric Estimations

Aida Salko
Department of Economics and Social Sciences
Sapienza University of Rome, Italy
E-mail address: aida.salko@uniroma1.it

Rita D'Ecclesia
Department of Statistics
Sapienza University of Rome, Italy
E-mail address: rita.decclesia@uniroma1.it

Abstract

This study analyzes different parametric and non-parametric modeling methods for estimating the Loss Given Default (LGD) of bank loans for shipping companies. The shipping industry is subject to several different risks which create the need to accurately measure the possible losses in order to estimate the LGDs for the banking industry. We use a unique database of defaulted loans in European banks involved in shipping finance. The aim of this study is twofold: to compare the performance of alternative LGD modeling methodologies in shipping finance and to provide some insights into what drives LGD in the shipping industry. We find that non-parametric methods, especially random forest, lead to a remarkable increase in the prediction accuracy and outperform the traditional statistical models in terms of both in-sample and out-of-sample results. To investigate the risk drivers in the shipping business, we use a variable importance measure built on the idea of the permutation importance. We find the energy index to be of paramount importance and the most important risk factor to estimate shipping finance LGD. We find that crude oil prices play a big role and may affect the financial health of shipping firms and then the LGDs of shipping loans.

Keywords*:* Loss Given Default, Shipping Finance, Forecasting, Machine Learning, Global Credit Data.

JEL Classification*: C53, G21, G32.*

1. Introduction

The shipping industry is the leading mode of transportation worldwide and is considered the backbone of global trade generating an annual income of almost $500 billion in freight rates representing approximately 5% of the total global economy.[1] It has always been a volatile and cyclical business and very complex affected by different factors involving high levels of capital investment, the characteristics of the company that needs capital, the legislation, financial markets, and different global economic indicators. Bank loans are considered historically the largest source and the most common way of financing vessels in the shipping industry (Harwood, 2006, Grammenos, 2010, Albertijn et al., 2011). However, during the financial crisis of 2007-2008, the maritime industry suffered a great recession associated with very low earnings of shipping freight markets, which affected banks to significantly reduce the amount of debt financing in this sector. Moreover, increased regulatory standards provided by Basel III Accord, have made loans to finance vessels less available and more expensive for ship-owners as well as less profitable for banks (Drobetz & Merikas, 2013).

The shipping industry is affected by two main risks: changes in operating cash flows and changes in the market value of assets. Accurate estimates of potential losses are essential for financial institutions in terms of lending and pricing strategies and the management of credit risk. Loss Given Default (LGD) is crucial to measure credit risk and recently has been at the center of many studies, both academic and commercial. The LGD concerning shipping finance is of great relevance for banks taking into account a wide set of risks associated with the sector, mainly driven by the very high volatility in freight rates. Kavussanos & Tsouknidis (2016), analyze several risks that make the cash-flow generating ability of ships uncertain for covering both the operating costs of the vessels and debt payments. In spite of its obvious importance, LGD in shipping finance still remains an unexplored topic in the academic literature mainly due to the lack of data.

To our knowledge, the only study focusing on this topic (Brumma & Winckle, 2017), reveals interesting findings indicating higher losses for banks in cases of selling the ships as well as a strong positive correlation between the loan-to-value (LTV) ratio and LGD. The authors also

[1] International Chamber of Shipping (2020). Annual Review 2020, Heroes at Sea. https://globalmaritimehub.com/wp-content/uploads/2020/10/Annual-Review-2020-Final.pdf

highlight the impact of macroeconomic downturns on the LGD curve over time in the context of the shipping industry.

In this study, we use a unique database of defaulted loans in European banks involved in shipping finance provided by Global Credit Data (GCD).[2] Using shipping finance loss data, we compare different parametric and non-parametric modeling methods to estimate and forecast LGD for shipping finance. The aims of this study are: i) to explore different approaches to estimating the LGD in shipping finance and ii) identify the risk drivers of LGD in the shipping industry. The unique dataset provides interesting features of the distribution and size of losses in the shipping industry as well as information on the risk factors for the industry. A key variable is the energy index and in particular crude oil prices, which affect LGD in the shipping finance of bank loans.

We consider four parametric models in our study: 1) a simple OLS regression, 2) Ridge regression, 3) Least Absolute Shrinkage Selector Operator (LASSO) regression, and 4) Net Elastic regression and we compare their performances with a wide set of machine learning algorithms. No consensus exist in the literature regarding the most accurate prediction methods. Many studies have addressed benchmarks of LGD prediction methods to provide a comprehensive assessment. Bellotti & Crook (2012) find that OLS regressions in combination with macroeconomic explanatory variables turn out to be the best forecasting approach. However, recent studies provide evidence that non-parametric methods outperform parametric methods in terms of prediction accuracy (see Bastos 2014, Loterman et al., 2012, Qi & Zhao, 2011, Yao et al., 2015, Bellotti et al., 2021).

To identify the main risk drivers, we use a variable importance measure built on the idea of the permutation importance. In this regard, we further explore what features drive the results of the algorithm's prediction. We find that non-parametric methods, especially the random forest one, provides the best accuracy and outperform the traditional statistical models both in-sample and out-of-sample performance.

We also find that the energy commodity index, the country of jurisdiction, followed by uncertainty index and collateral vessel-related characteristics turn out to be important drivers for an accurate prediction of LGD in shipping finance. To the best of our knowledge, this study is the first that

[2] Global Credit Data provides the largest LGD data base worldwide. The association consists of 55 banks from all over the world. See http://www.globalcreditdata.org/ for further information.

investigates modeling methods for LGD of bank loans in shipping finance and provides new insights to identify the risk drivers for LGD in the shipping industry.

The remainder of the paper is organized as follows. Section 2 summarizes the main studies related to banks in shipping finance. Section 3 briefly describes our dataset with a special focus on the underlying collateral structure in terms of vessel types. Section 4 introduces the parametric and non-parametric methods that are used for the estimation. Section 5 and section 6 report the empirical results and the concluding remarks.

2. Banks and shipping finance

Banks play a crucial role in the shipping industry as they provide the major source of financing vessels. Harwood (2006) states that the term loan is considered the most widely used financial instrument for debt capital in shipping. According to Stopford (2009), commercial banks make available most of the debt capital in shipping. Historically, European nations have been heavily involved in the maritime industry and nowadays, European banks still provide the main debt capital in shipping. However, the financial crisis of 2007-2008 had a huge impact on the shipping industry the banking sector was affected by it. According to Lozinskaia et al. (2017), after the crisis, slower growth of global demand for seaborne trade and a rising supply of vessels entering the market brought a sharp decrease in vessel values and charter rates leading to many loan defaults and bankruptcy of shipping companies and this was mostly due to the lack of reliable and accurate models for estimating the risk of lending to shipping companies.

Kavussanos & Tsouknidis (2016) show that the total value of the loans to shipping firms globally, reached a high of $115 billion in 2007 representing the 75% of external funding in the shipping industry and fell to $46 billion in 2012. In addition, Basel II/III capital requirements affect lending to the shipping industry.

Research on bank loans of shipping finance is limited and mainly focused on identifying the drivers which cause defaults in shipping loan agreements. Dimitras et al. (2002), use a sample of 17 shipping bank loans and propose a multi-criteria methodology for the evaluation of loan origination in shipping, identifying a utility function and the cut-off utility level for granting a loan. Grammenos (2010) analyzes the 6 “Cs” of credit in bank shipping finance as a sound credit

analysis method of assessing the default probability. In an earlier study, Grammenos et al. (2008) applies a binary logit model to predict the probability of default for high yield bonds issued by shipping companies. Gong et al. (2013) investigate the impact of the 2008 Global financial crisis on Hong Kong banks' ship financing practices and found a significant decrease in banks' shipping portfolios during the crisis. Mitroussi et al. (2016) use a binary logit model to analyze the criteria for assessing the security of shipping loans issued by banks. They examine thirty shipping loans during the period 2005–2009 and find as risk drivers for evaluating the performance of shipping loans: financial and non-financial factors, ship owners′ experience, employability and market risk indicators. Kavussanos & Tsouknidis (2016) propose a credit scoring model for the empirical assessment of default risk drivers of shipping bank loans. The authors find that the current and expected conditions in shipping freight markets, the risk appetite of the obligor, and the pricing of the loan are the main factors explaining the default probabilities of bank loans. Further, Lozinskaia et al. (2017) employ a logit model using a sample of 192 listed shipping companies to investigate the determinants of the probability of default. In line with previous studies, the authors find that financial and non-financial variables are important in assessing the creditworthiness of shipping companies. As all of the existing studies mainly focus on default risk in shipping loans, to our knowledge, there are no empirical studies concerning the LGD of bank loans in shipping finance.

3. Dataset

We use a unique loss database provided by Global Credit Data (GCD). The GCD association consists of 55 member banks including several global important banks, from all over the world and collect data over the last 20 years. We use data on defaulted shipping borrowers divided by: (1) the defaulted borrower, (2) the characteristics of the ships serving as collateral, and (3) loan-related factors. We analyze 363 defaulted loans with a shipping collateral whose country of jurisdiction is located in European countries.

LGD[3] is given as:

$$LGD = 1 - RR \tag{1}$$

where RR presents the Recovery Rate computed as:

$$RR = \frac{\sum i^{+} - \sum c^{-}}{EAD} \tag{2}$$

where (i^{+}) includes all the discounted incoming cash flows (i.e. principal, interest, received fees and commissions, post resolution payments etc.) and (c^{-}) consists of all the discounted direct and indirect costs (i.e. workout costs, legal expenses, liquidation expenses etc.) divided by the exposure at default (EAD). According to Basel II definition, a default occurs if an obligor is "unlikely to pay" or "past due more than 90 days on any material credit obligation". We refer to the default definition set by Basel II and therefore we restrict our data sample from the year 2000 in order to ensure a consistent default definition. In addition, we do not account for defaults after 2019 since the workouts of recent defaults are not necessarily completed and it might lead to an unrealistically long-term average LGD since lots of cases of short workout periods are present.[4] We also consider macroeconomic factors to identify the variables which affect the LGD dynamics following the European Banking Authority (2018).

Table 1 provides a summary of all the variables that are used in our study. The rich data set provided by GCD offers us a great variety of input variables to analyze different risk drivers for accurate LGD forecasts and estimations. We also include three main macroeconomic control variables: the Gross Domestic Product (GDP) growth rate, the news-based economic uncertainty index, and the commodity index. In a recent study, Gambetti et al., (2019) find that economic uncertainty turns out to be the most important systematic determinant for the recovery rate distributions. Following Gambetti et al. (2019), we use the original economic policy uncertainty index developed by Baker et al. (2015) which is based on the normalized volume of newspaper articles published in a given month containing expressions referring to economic policy

[3] In this study we model loan-level LGD. The economic LGD calculation is used where principal advance and financial claim are parts of the recovered amount.

[4] The resolution bias is addressed according to GCD methodology: https://www.globalcreditdata.org/system/files/documents/gcd_lgd_report_2020_appendix_01062020.pdf

uncertainty. This news-based indicator, commonly referred to as the economic policy uncertainty index, serves as a proxy for policy-related economic uncertainty.

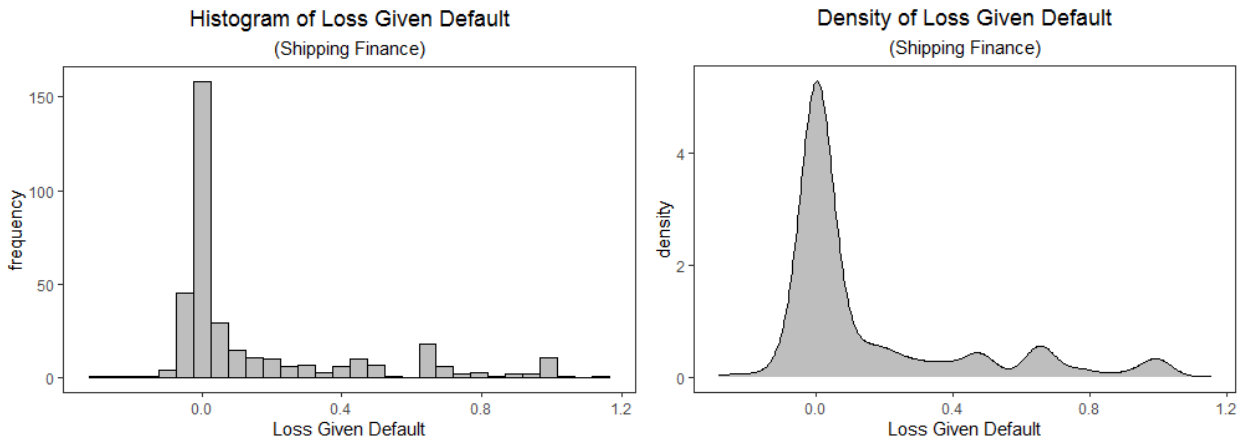
Table 1. Risk drivers used as explanatory variables of LGD shipping bank loans.

Type of information	Variable	Description	Variable Type
Loan-related characteristics	EAD	Exposure at Default (EUR).	Continuous
	Facility Type	Distinguished between medium term, short term, and other facilities.	Categorical
	Seniority	Divided into the categories as super senior, pari-passu, and non-senior.	Categorical
	Guarantee indicator	Indicator showing if loan has underlying protection in form of a guarantee or not.	Categorical
	Committed Indicator	The contractual obligation for the bank to "make the funds" when the facility is drawn by the client.	Categorical
	Country of Jurisdiction	Country of the loan contract.	Categorical
Collateral (Ships) related characteristics	LTV	Loan-to-Value Ratio	Continuous
	Vessel Type	i.e. container, tanker, dry cargo, cruise vessels, or offshore.	Categorical
	Vessel Size	i.e. oceangoing, seagoing or river/coastal.	Categorical
	Collateral Valuation Type	How or by whom the collateral has been valued.	Categorical
	Ranking security	Ranking priority of the collateral, i.e. first, second ore subsequent charge	Categorical
Entity-related characteristics	Debt to Assets Ratio	The entity total amount of debt relative to its assets in the 12-Month period before the default.	Continuous
	Asset Turnover Ratio	The ratio of the entity sales relative to the value of its assets in the 12-Month period before the default.	Continuous
	Industry Category	Industry that accounts for the largest percentage of the borrower's revenues.	Categorical
	Country of Residence	The legal country of the residence.	Categorical
Macroeconomic variables	GDP Growth	Quarterly European GDP growth rate when the loan has defaulted (%).	Continuous
	Uncertainty Index	Quarterly news-based Uncertainty Index for Europe retrieved from FRED.	Continuous
	Commodity Index	World Bank Monthly Energy Commodity Price Index.	Continuous

We refer to the World Bank Commodity Price Data and use the Energy Price Index which includes prices regarding coal (4.7%), crude oil (84.6%), and natural gas (10.7%). The choice of the commodity price index in our study is aimed to take into account the role of commodity prices which is known to be one of the main factors affecting the shipping industry. In the world economy, most of the commodities used as input for final products are transported by sea. Commodity prices drive the demand for commodities and many studies have found that commodity prices have an impact in the maritime industry.

In Figure 1, the distribution of LGDs is reported and they show a left-skewed distribution. The lowest LGD is -28.61% while the mean is given by 15.54% and the median by 0.9%. We can observe a few LGDs that fall below 0 and exceed the value of 1. Negative values of LGDs as well as values greater than one may appear in some cases due to high costs (administrative, legal, and liquidation expenses/financial penalties or high collateral recoveries).

Figure 1. Observed LGD values for Shipping Finance

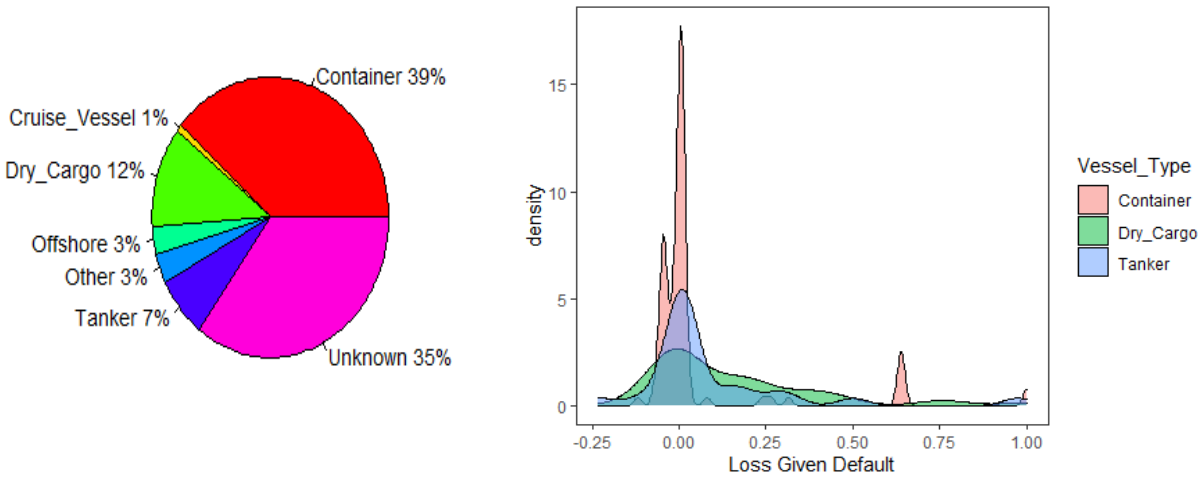


The crucial part of shipping is the vessels. Most of our database consists of container vessels (39%), followed by dry cargo vessels (12%) and tanker vessels (7%), Figure 2. Other types of the vessels such as offshore vessels or cruise vessels are also present in the database (3% and 1% respectively). In Figure 2a, the composition of vessels by vessel type is reported, in Figure 2b a comparison of the LGD distribution for the most relevant vessel types is reported.

We observe that the LGD distribution of these three vessel categories presents quite different features which have to be investigated further.

Figure 2. Vessel Types and their LGD distribution

(a) (b)



4. Methodology

We consider four parametric models: i) a simple OLS regression, ii) Ridge regression, iii) Least Absolute Shrinkage Selector Operator (LASSO) regression, and iv) Net Elastic regression. The main assumption when running linear regression is that the distribution of the error term is (approximately) normally distributed which is not the case for the LGDs, as shown in Figure 1. In this case linear regression may not be the adequate model to investigate LGD drivers as also found by Hartmann-Wendels et al., 2014, Zhang and Thomas, 2012, Kaposty et al, 2020. We also use five non-parametric methods for forecasting LGD to deal with non-linear and more complex relationships among variables.

4.1 Parametric Methods

Ridge, LASSO and Elastic Net Regression[5] are the so called "shrinkage methods" and are forms of regularized linear techniques found in General Linear Models. The main idea behind these regression models stands in shrinking the regression coefficients by imposing a penalty on their

[5] Hoerl (1962), Tibshirani (1996), Zou & Hastie (2005).

size. If we denote by $\lambda \geq 0$ the penalty factor that controls the amount of shrinkage, and the mixing factor $0 \leq \alpha \leq 1$, the regression methods are the solution to the minimization problem:

$$argmin_{\beta_0 \beta} \|Y - X\beta - \beta_0\|_2^2 + \lambda\big((1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1\big) \qquad (3)$$

where $Y = (Y_1, \ldots, Y_N)$ is the vector of training set observations, X denotes the $(N \times p)$ matrix of regressors, and $\beta = (\beta_1, \ldots, \beta_p)$ is the vector of unknown regression coefficients. The larger the value of λ, the greater the amount of shrinkage. When λ=0, we get the standard OLS regression. In this study we define penalized model by setting $\lambda > 0$ and different mixing factors: Ridge ($\alpha = 0$), LASSO ($\alpha = 1$), or Elastic net ($0 < \alpha < 1$). The main difference between a Ridge and a LASSO regression stands in how they alter the cost function. The Ridge regression is based on the idea of adding a penalty equivalent to the square of the magnitude of the coefficients. So the Ridge regression shrink the coefficients and reduces the model complexity and multicollinearity. In the case of LASSO regression, the penalty term includes the absolute weight and tends to make coefficients to absolute zero.

4.2 Non-Parametric Methods

In this study we use five different non-parametric methods. The model structure is not specified a priori but is greatly determined directly from the available training data. Mainly the number and the nature of the parameters are flexible and can depend on the training data.

4.2.1 Bagged Trees

First introduced by Breiman (1996), bagging is a powerful method that generates several versions of individual predictors and averages them to get a final aggregated predictor. Given a dataset, this approach generates new data sets by multiple bootstrapping (i.e. sampling with replacement) and therefore creating a decision tree for each of the new training sets. Consequently, the number of trees composing the ensemble equals the number of generated bootstrap samples. The main idea of bagging stands on the combination of the predictions of several base learners with the purpose to create more accurate output. The name “bagging” comes from the combination of both Bootstrapping and Aggregation that form the ensemble method.

4.2.2 Random Forest

Considered as an evolution of Breiman's original bagging algorithm, the random forest is another important ensemble strategy that incorporates randomized feature selection Breiman (2001). Random forest is a powerful rule-based algorithm formed as an ensemble of decision trees where each tree is trained on a different artificially created sample. Same as bagging, all the decision trees that form the random forest are different since each tree is built on a different random data. Random forest produces a final predictor under a different sampling mechanism. In the case of bagging trees, all features are considered at each split in the tree-growing phase. Contrary, the random forests differ in this aspect because only a random subset of available features is considered at each split. In other words, the random forest uses a random selection of features rather than using all features to grow the trees. This contributes to reducing the correlation and the variance of the ensemble prediction. The random forest uses an average of all single predictors to make a better final prediction. That is, if we have a full set of *n* features, then only a random sample of *m* features is chosen as split candidates when building random forest. It is important to emphasize that randomness is introduced only in the process of selecting features and not on splitting points of these features.

4.2.3 Boosted Trees

Boosting is another ensemble algorithm widely used in many statistical learning methods for regression and classification. Based on the gradient boosting machines algorithms presented by Friedman (2001), this model is a powerful ensemble strategy where the residuals of the model are fitted by many weak learners iteratively. Contrarily to bagged trees, boosted trees do not involve bootstrap samples but use the original data set to grow trees sequentially. This means that each tree is grown in sequence by using the information from the previously grown tree and depends on the results of the previous trees. In the case of bagging and random forest, trees are grown in parallel.

4.2.4 Neural Networks (NN)

Neural Networks (NN) is a non-linear and non-parametric modeling method that is inspired by the way that the human brain works, imitating the way that biological neurons signal to one another.

The main idea of the algorithm is to extract linear combinations of the explanatory variables and then model the target variable as non-linear functions of these explanatory variables. This machine learning algorithm typically consists of the following three layers: the input layer, presenting the raw information that can feed into the network; the hidden layers, transforming the input into something the output layer can use; and the output layer, which returns an output value that presents the prediction of the response variable. We use a two-layer feed-forward neural network with one hidden layer[6]. This is one of the simplest variants of neural networks which passes the information through various input nodes only in one direction, until it arrives to the output node. The model is described by:

$$LGD_i = \sum_{k=1}^{s} w_k g_k\left(b_k + \sum_{j=1}^{p} x_{ij}\beta_{kj}\right) + \mathrm{e}_i, \qquad i = 1, \dots, n \tag{4}$$

where *s* is the number of nodes , *p* is the number of input variables, w_k presents the weight of the *k*-th node (*k =1,…., s*), g_k is the activation function, b_k is a bias for the *k*-th neuron which can be interpreted as the intercept of the linear combinations of the inputs (*k =1,…., s*); β_{kj} represents the weight of the *j*th input to the network; x_{ij} denotes the information that is included on loan *i;* and $e_i \sim N(0, \sigma_e^2)$.

4.2.5 Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regression Splines (MARS) is another non-parametric and non-linear regression method introduced by Friedman (1991). The main idea of this modeling technique stands in building multiple linear regression models across the range of predictor values. The MARS algorithm is considered as an extension of linear models but it makes no assumptions about the relationship between the response variable and the predictor variables. The algorithm builds the models in two steps: first, it starts by partitioning the data, and second, it runs a linear regression model on each different partition.

In the first step, the algorithms create a range of predictor values which is partitioned into several groups, and for each of these, a separate linear regression is modeled. The connections between the separate regression lines are referred to as knots. Then, the idea of the MARS algorithm is to search for the best spots to place the knots. The model can be expresses as the following problem:

[6] We follow Foresee & Hagan (1997), MacKay (1992), Rodriguez & Gianola (2016), and Kaposty et al. (2020).

If *y* is the target output and $X = (X_1, \ldots, X_N)$ is a matrix of *N* input variables, let's assume that the data are generated from an unknown "true" model which would be expresses as:

$$y = f\left(X_{1,\ldots\ldots}X_N\right) + e = f(X) + e \tag{5}$$

where *e* is the distribution of the error. The function *f* is then approximated by applying some basis functions, which are splines (smooth polynomials), including piecewise linear and piecewise cubic functions. It can be formally written as:

$$max(0\,, x - t) = \begin{cases} x - t \;\; if \; x \geq t \\ 0, otherwise \end{cases} \tag{6}$$

indicating that only the positive part of the equation is used otherwise it is given a zero value. Finally, the MARS model *f(X)* is expressed as a linear combination of basis functions and their interactions:

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m \lambda_m(X) \tag{7}$$

where each $\lambda_m(X)$ is a basis function and the β coefficients are estimated using the least-squares method.

4.3 Error measures and predictive accuracy

Evaluating the predictive accuracy of our models is an essential part of the study. In order to assess the performance of our models, we need to quantify how well the predictions actually match the observed data. Therefore, we will use the root mean squared error (RMSE) and the mean absolute error (MAE) as the most commonly used measures of model performance. The root mean squared error (RMSE) is defined as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{8}$$

where y_i is the actual level of the variable; the $\hat{y}$ is the predicted variable.

The mean absolute error (MAE) which shows on average, how far is the model prediction from the true value is given as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | \, y_i - \hat{y}_i| \tag{9}$$

However, we are interested to assess the RMSE and MAE on a sample that is independent of that used in building the models. To achieve this, we will split our sample into two sets using a standard 70% - 30% random split. The first set is used to fit the model, i.e. the training set, and the second one is used to test its accuracy, i.e. the test set. Following this, the performance measures mentioned above are assessed in both sets. In addition, model hyper-parameters were tuned by using a ten-fold cross-validation on the training set. All the models were trained using the latest version of the Caret library in R (Kuhn, 2008, 2018; Kuhn & Johnson, 2013).

5. Results

The performance matrices of in-sample and out-of-sample results are presented in Table 3 below. We also report the R^2 values as the most intuitive measure of explanatory power. Both in-sample and out-of-sample performance measures of our algorithms are ranked in an improving order in terms of R^2, RMSE, and MAE.

Regarding errors between the realized and forecasted LGDs, we find that non-parametric methods produce the best forecasting results, and outperform parametric methods in terms of both in-sample and out-of-sample results. Non-parametric methods exhibit a proportion of explained variation in terms of R^2 measure, ranging from 44.20% to 81.15 % for in-sample results and from 14.31 % to 55.76 % for out-of-sample results.

We find that the random forest algorithm is the best method explaining 81.15% of volatility for the sample data ($R^2 = 0.81$) and for the out-of-sample data 55% ($R^2 = 0.55$). The bagged trees and boosted trees are the next best performing algorithms after random forest, among non-parametric methods. Other non-parametric models including NN and MARS are also well ahead of the parametric methods presenting the lowest errors both in-sample and out-of-sample as well as in terms of R^2 . Parametric models show a weaker performance in terms of predictive accuracy and explained variation: the OLS seems to produce the best results in terms of in-sample performance but at the same time it reports the weakest performance for the out-of-sample results. The LASSO regression and Elastic Net Regression turn out to produce the best out-of-sample results compared to all the parametric methods.

Table 3. Performance Matrix

	In-sample quality-of-fit measures			Out-of-sample quality-of-fit measures		
Methods	RMSE	MAE	R^2	RMSE	MAE	R^2
OLS	0.2149 (5)	0.1447 (5)	0.4319 (6)	0.2922 (9)	0.2002 (8)	0.1064 (9)
Ridge	0.2512 (9)	0.1887 (9)	0.3106 (9)	0.2752 (8)	0.2059 (9)	0.1104 (8)
LASSO	0.2364 (7)	0.1676 (7)	0.3453 (7)	0.2671 (6)	0.1907 (6)	0.1313 (6)
Elastic Net	0.2404 (8)	0.1761 (8)	0.3270 (8)	0.2696 (7)	0.1959 (7)	0.1148 (7)
Bagged Trees	0.1943 (3)	0.1273 (3)	0.5712 (3)	0.2598 (5)	0.1584 (3)	0.2461 (3)
Random Forest	**0.1265 (1)**	**0.0776 (1)**	**0.8115 (1)**	**0.1777 (1)**	**0.1213 (1)**	**0.5576 (1)**
Boosted Trees	0.1728 (2)	0.1168 (2)	0.6908 (2)	0.2457 (3)	0.1473 (2)	0.3321 (2)
NN	0.2174 (6)	0.1646 (6)	0.4552 (4)	0.2499 (4)	0.1809 (5)	0.1431 (5)
MARS	0.2106 (4)	0.1431 (4)	0.4420 (5)	0.2286 (2)	0.1658 (4)	0.1901 (4)

Notes: The numbers in brackets state the ranks of the models in terms of performance measures. The ranks range from 1 (best) to 9 (worst).

In summary, we find that the non-parametric models, show a clear advantage over the parametric models in terms of both in-sample and out-of-sample performances. The superiority of these algorithms in forecasting these data remains evident.

5.1 Variable Importance

In this section, we investigate the importance of LGD risk drivers in our models mainly to identify the factors which drive the LGD dynamics. We generate a visual comparison of all the input variables for every model by constructing a measure built on the idea of the permutation importance. The main idea of the permutation importance stands in two steps. First, we calculate an error measure for each prediction using the entire dataset. In this case, we choose MAE as the error measure. The calculation is based on the in-sample case. In the second step, we permute the values of each independent variable that is used for prediction and calculate the new error measure for each method. We then calculate the increase in the error measure relative to the non-permuted case. A variable is considered "important" if shuffling its values increases the model error, given that, the model relies on that variable for the prediction. The permutation feature importance measurement was first introduced by Breiman (2001) for the random forests algorithm. It can be expressed in the following steps:

Input: Trained model $\boldsymbol{F}$, feature matrix $\boldsymbol{X}$, target vector $\boldsymbol{y}$, error measure $\boldsymbol{L(y, F)}$.

1) Estimate the original model error $\boldsymbol{MAE^{orig} = L(y, f(X))}$ (in our case MAE)

2) For each feature $\boldsymbol{j = 1,...,n}$:

- Generate feature matrix $\boldsymbol{X^{perm}}$ by permuting feature $\boldsymbol{j}$ in the data **X**. This breaks the association between feature $\boldsymbol{j}$ and true outcome $\boldsymbol{y}$.
- Estimate error $\boldsymbol{MAE^{perm} = L(Y, f(X^{perm}))}$ based on the predictions of the permuted data.
- Calculate permutation feature importance $\boldsymbol{FI^j = MAE^{perm}/MAE^{orig}}$.

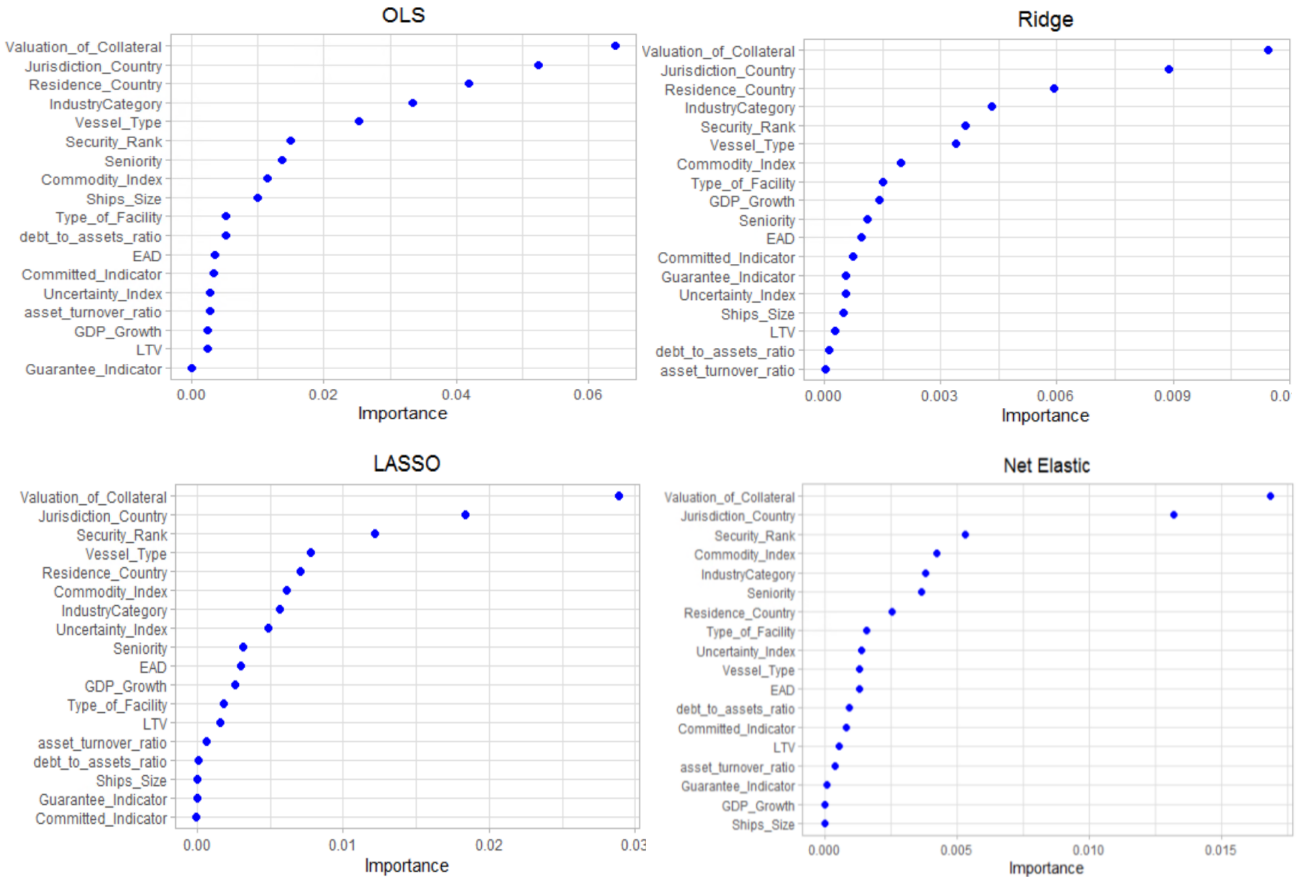
3) Sort features by descending **FI**.

Figure 3 presents variable importance rankings for all the parametric models, as measured by changes in the MAE when permuting a single variable. Figure 4 presents the results for all the other non-parametric methods. The variables are ranked by their importance in a decreasing order and the results reveal some important insights into what drives LGD of shipping finance. The most important variables considered by all the parametric and non-parametric methods, are a mixture of macroeconomic indicators, collateral related characteristics and loan file information.

The results of only parametric methods (Figure 3) show in all cases the country of jurisdiction and collateral valuation result to important features. In the case of non-parametric algorithms (Figure 4), all these algorithms generally agree in their decisions and rank the energy index and the jurisdiction country as the main input variables in forecasting LGDs of shipping loans. Country of jurisdiction refers to the country of the law applicable to the facility, i.e. the law of the loan contract and it is considered by all the prediction techniques as a significant risk driver of LGD. The reason can be the important role of country specific differences in terms of the regulatory and legal framework. Betz et al. (2016) find that considerable differences in the legal frameworks across countries lead to adjustments in the general lending behavior of creditors by affecting the time to resolution of loans which is itself positively correlated with the LGD of loan contracts. The permutation results reflect the importance of the macroeconomic environment, collateral-related characteristics as well as the significant role of the regulatory and legal framework in LGD modeling of shipping-related transactions. Looking at the random forest model results, we observe that the energy index is the first most important input variable, followed by the country of jurisdiction, and the uncertainty index. This result highlights the significant role of the macroeconomic environment in the shipping industry.

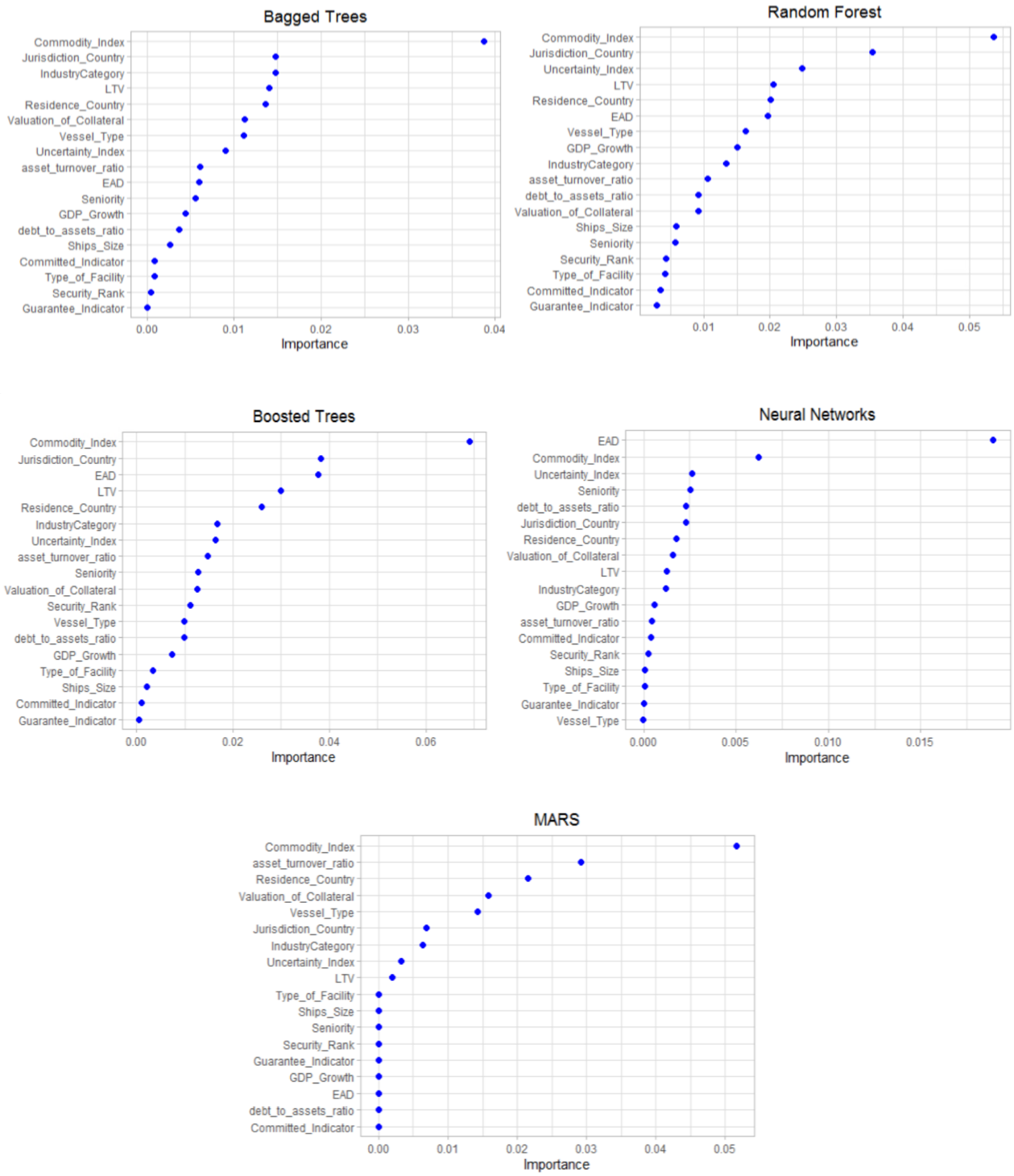
Particularly, the volatility of energy prices is a dominant factor in this industry, which of course is then related to the increase in economic uncertainty levels. The result is also in line with the findings of Gambetti et al. (2019), who show that economic uncertainty proves to be the most important systematic determinant of recovery rates, even if with a lower impact.

Figure 3. Variable Importance in Parametric Methods



LTV, residence country and EAD have all the same relevance in explaining LGD dynamics while the vessel type and GDP growth have lower impact, unlike the results found by Zannetos (1966) who considers the vessel as the "firm". Depending on the specific purpose of the vessels, different amounts of capital may be required implying higher risk of losses for banks. In addition, vessels are affected by great volatility in terms of freight rates and asset prices which make it even more problematic for lenders. In general, a careful loss modeling of a shipping company depending on the vessel type is required for banks, even if our models do not attribute high importance to this variable.

Figure 4. Variable Importance in Non-Parametric Methods



5.2 Effect of energy index in model prediction

The main contribution of this study relies in finding the energy index is considered as the most important risk driver of LGD in shipping finance. To further investigate and interpret the relationship between model forecasts and this risk driver, we present the marginal effect on the predicted outcome of the energy index by the best method - random forest. We do this by visualizing the Partial Dependence (PD) plots (Friedman, 2001), which is a popular tool used in the field of explainable machine learning. They are calculated after the model has been fit and attempt to visually explain what the model predicts on average when the value of the feature changes. In other words, their interpretation can be seen as the expected target response as a function of the input features.

The vertical axes of the plots represent the marginal impact of the independent variable to the dependent variable while the horizontal ones stand for the individual feature. In Figure 5 we present the dynamic of the energy commodity index which shows its peak in correspondence of the Global Financial Crisis (GFC) and then shows a very volatile dynamic from 2010 onwards. In Figure 6, we report the PD plot. As it can be seen, the highest peak of the index (Figure 5), around 170, is achieved during the financial crisis of 2008, this period is also the one with highest losses for the banks. The model suggests that as the index is increasing, up to the value of 120, there is a negative downtrend, meaning that it is associated with lower losses for the bank. However, the response changes immediately as the algorithm captures a strong positive signal in terms of higher losses for the banks as this index gets greater than 120. In other words, we infer that the bank should expect higher losses when the energy index jumps above a certain level.

Given that crude oil represents an 84.6% share of the energy index, the results are obviously driven by the price of crude oil. Oil plays an essential role in the global economy and particularly in the shipping industry. There are several studies finding evidence on the impact of changes in oil prices as one of the most important risk factors on the shipping industry[7]. Apparently, the role of the oil prices is considered as one of the main determinants on the profitability of shipping companies. In their study, Mayr & Tamvakis (1999) explain that if oil price increases due to more demand for imported crude oil, then the oil prices have beneficial effects on freight rates. Other authors like

[7] Grammenos & Arkoulis (2002) , Alizadeh & Nomikos (2006), Beverelli et al. (2010), Drobetz et al. (2010), Kutin et al., (2018), Sun et al. (2018).

El-Masry et al., (2010a), Drobetz et al., (2010) also find a positive relationship between an increase in oil prices and shipping stock returns, explaining that the demand for tanker freights is an initiated demand from oil. In a recent study, Maitra et al. (2020) show that there was an increase in the volatility co-movement between oil and liner shipping companies' stock returns during the 2007–09 global financial crisis, and the 2010–12 Eurozone debt crisis. In addition, as explained by Narayan & Sharma (2011), oil prices may also have a positive impact on transportation if the increase in the oil price is driven by an improvement in the overall economic growth due to more energy consumption. When oil prices increase as a result of increased energy demand, the shipping company stocks will have a better performance and generate lower losses with a result of lower LGD. This dynamic works for values of the energy index growing from 90 to 120, so in presence of higher energy index, LGD reports lower values. When it goes above 120, which is the value of the energy index during the GFC in 2008, a sharp increase of the LGD occurs. This may be explained by the fact that when oil prices, and therefore the energy index become extremely high, shipping companies experience losses due to higher operating costs and this may translate in larger number of defaults and higher LGDs. The higher operating costs generated by extremely high energy prices, caused shipping companies to save on fuel prices by using large vessels and reducing steaming, and looking for faster modes of transportation (Maitra et al. 2020). We find that crude oil prices large volatility affects LGDs in two directions: at first when prices increase due to an initial higher demand of crude oil, shipping companies benefit from a first increase of business but when oil prices keep on increasing, the positive effects on shipping companies revenues is offset by the increase in operating costs and uncertainty in the economic context.

Figure 5. Energy Index in years

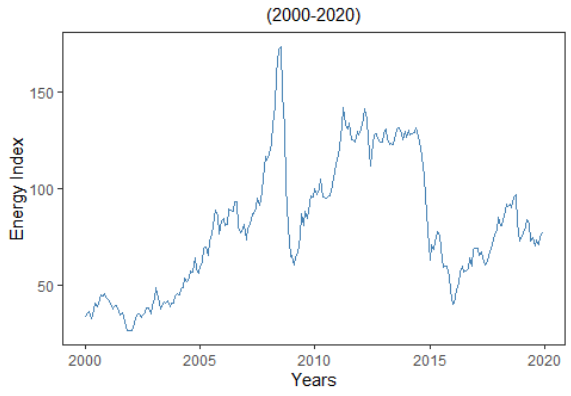
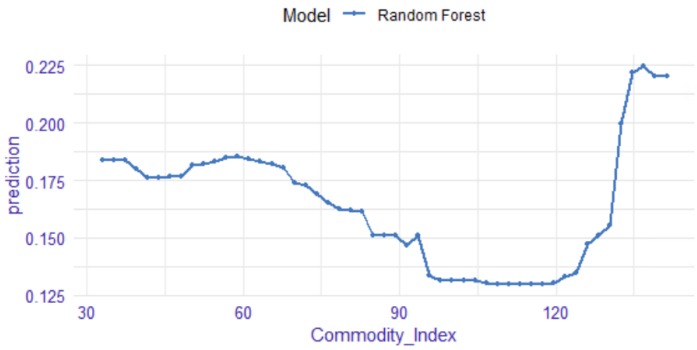
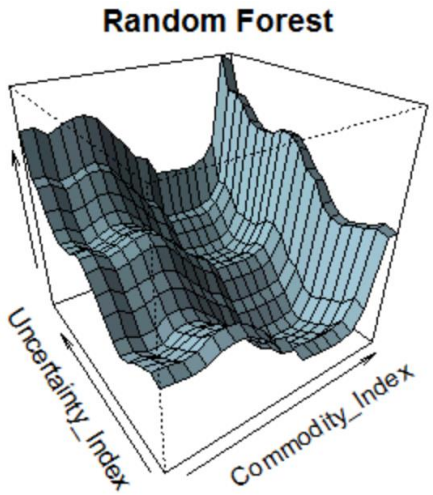


Figure 6. PD plot for Energy Index



Kang & Ratti (2013) show that increases in the real price of oil are associated with significant increases in economic policy uncertainty. The economic policy uncertainty is a transmission channel for the effect of oil price shocks on the economy. Gambetti et al. (2019) find economic uncertainty to be the most important systematic determinant of recovery rate distributions. However, an increase in economic uncertainty is followed by lower investment and reduction of oil demand, leading to less transportation by oil tankers and lower worldwide seaborne transportation of goods (Maitra et al. 2020). The random forest method identifies the uncertainty index as the second most important macroeconomic indicator after the energy index, in forecasting LGDs of shipping finance. Figure 7 presents a 3D PD plot of LGD and the two most important macroeconomic drivers.

Figure 7. 3D PD plot Energy Commodity Index and Uncertainty Index



The uncertainty index clearly conveys information related to the LGD in shipping finance: the higher the uncertainty index, the higher the LGD of shipping-related loans. In other words, the positive trend of the uncertainty index implies that higher uncertainty levels are associated with possible higher losses for the bank which is consistent with the findings of Gambetti et al. (2019).

5 Conclusion

Since the introduction of the Basel II Accord, the modeling of Loss Given Default (LGD) as a critical component in credit risk management, is increasing in importance. A special focus is also given to specialized lending exposures as one of the main parts of the regulatory framework. In this study we focus on the shipping finance loss data and try to identify the main risk drivers for LGDs.

We use different parametric and non-parametric approaches to predict LGD for shipping finance. The shipping industry is considered the backbone of global trade and the global economy but is affected by different risk factors which create the need for a more detailed loss modeling for the banks. Accurate estimates of potential losses are essential for financial institutions in terms of lending and pricing strategies and the management of credit risk.

W use parametric and non-parametric models to study LGDs dynamics, specifically a simple OLS regression, Ridge regression, LASSO regression, and Net Elastic regression, and a wide set of machine learning algorithms including bagged trees, random forest, boosted trees, NN, and MARS. To the best of our knowledge, this is the first study to conduct a comparative analysis of such a wide range of prediction methods in the context of shipping finance LGDs.

Even though simpler models are more interpretable and easier to implement, we find that non-parametric methods, especially random forest, lead to a remarkable increase in the prediction accuracy and outperform the traditional statistical models in both in-sample and out-of-sample results. The random forest model stood out as having the best forecasting performance among all the models.

Furthermore, we use a variable importance measure built on the idea of the permutation importance, to analyze the risk drivers with the greatest effects on the LGD for shipping finance prediction accuracy for each method. The importance of the explanatory variables is analyzed by computing the relative changes in the prediction errors when permuting a single variable. In this regard, we further explore what features drive the results of the algorithm's prediction. We find that both, parametric and non-parametric models, identify in general the same risk drivers for LGD prediction. The best accuracy is obtained using non-parametric methods and in particular using the random forest approach.

All the non-parametric models identify energy index as main driver in forecasting LGDs of shipping loans. For a further investigation of the effect of energy index on model prediction, we use the PD plots as a tool to explain the relationship between the energy index and the expected LGDs of shipping loans. We observe that the model captures a positive signal in terms of higher expectation of losses, as the volatility of the crude oil market increases sharply, as it happened during the last financial crisis.

The result highlights the dominant role played by crude oil prices which can deteriorate the financial health of shipping firms and therefore affect the LGDs of shipping loans. Other inputs such as the freight rates can be considered in LGD modeling of shipping finance for further research.

References

Albertijn, S., Bessler, W., Drobetz, W., (2011). Financing shipping companies and shipping operations: a risk-management perspective. Journal of Applied Corporate Finance 23, 70–82.

Alizadeh, A.H., & Nomikos, N., K. (2006) Trading strategies in the market for tankers, Maritime Policy & Management, 33:2, 119-140.

Baker, Scott R., Bloom, Nick and Davis, Stephen J., Economic Policy Uncertainty Index for Europe [EUEPUINDXM], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/EUEPUINDXM, December 1, 2021.

Basel I, C. on B.S., (1988). International convergence of capital measurement and capital standards. BIS report.

Basel II, C. on B.S., (2004). International convergence of capital measurement and capital standards: a revised framework. BIS report.

Basel III (2010). Group of governors and heads of Supervision announces higher global minimum capital standards. BIS report.

Bastos, J. A. (2014). Ensemble Predictions of Recovery Rates. *Journal of Financial Services Research*, *46*(2), 177–193.

Bellotti, A., Brigo, D., Gambetti, P., & Vrins, F. (2021). Forecasting recovery rates on non-performing loans with machine learning. *International Journal of Forecasting*, *37*(1), 428–444.

Bellotti, T., & Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, *January 2007*.

Betz, J., R. Kellner, D. Rosch (2016). What drives the time to resolution of defaulted ban loans? *Finance Research Letters* 18, 7–31.

Beverelli, C., Benamara, H., & Asariotis R. (2010). Oil prices and maritime freight rates: An empirical investigation United Nations Conference on Trade and Development (UNCTAD/DTL/TLB/2009/2)

Breiman, L. (1996). Bagging predictions. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Brumma, N., Winckle, P. (2017). Global Credit Data. (2017). *GCD Shipping Finance LGD Study 2017*. *May*, 1–5.

Dimitras, A.I., Petropoulos, T., Constantinidou, I., (2002). Multi-criteria evaluation of loan applications in shipping. J. Multi-Crit. Decis. Anal. 11, 237–246.

Drobetz, W., Schilling, D., & L. Tegtmeier (2010). Common risk factors in the returns of shipping stocks Maritime Policy Manage., 37 (2), pp. 93-120.

Drobetz, W., & Merikas, A. G. (2013). Maritime Financial Management. *Transportation Research Part E: Logistics and Transportation Review*, *52*, 1–2.

El-Masry, A.A., Olugbode, M., Pointon, J., (2010). The exposure of shipping firms' stock returns to financial risks and oil prices: a global perspective. Maritime Policy Manage., 37 (5) (2010), pp. 453-473

European Banking Authority (2018). Consultation paper: On guidelines for the estimation of LGD appropriate for an economic downturn ('downturn LGD estimation'). European Banking Authority (EBA/CP/2018/08)

Friedman, J. H. (1991). Multivariate adaptive regression splines. The Annals of Statistics, 19(1), 1–67

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232.

Foresee, F. D., & Hagan, M. T. (1997). Gauss-Newton approximation to Bayesian regularization. In Proceedings of the 1997 international joint conference on neural networks Vol. 3 (pp. 1930–1935).

Gambetti, P., Gauthier, G., & Vrins, F. (2019). Recovery rates: Uncertainty certainly matters. *Journal of Banking and Finance*, *106*, 371–383.

Gong, S. X., Y. Heng-Qing, and Y. Zeng. (2013). "Impacts of the Recent Financial Crisis on Ship Financing in Hong Kong: Aresearch Note." Maritime Policy and Management 40 (1): 1–9. doi:10.1080/03088839.2012.745202.

Grammenos, C. T., A. G. Arkoulis. (2002). "Macroeconomic Factors and International Shipping Stock Returns. " International Journal of Maritime Economics, 2002, 4, (81-99).

Grammenos, C. T., N. K. Nomikos, and Papapostolou N. C. (2008). "Estimating the Probability of Default for Shipping High Yield Bond Issues." *Transportation Research Part E: Logistics and Transportation Review* 44 (6): 1123–1138.

Grammenos, C. (2010). Revisiting credit risk, analysis and policy in bank shipping finance. T*he Handbook of Maritime Economics and Business, Informa Law*, 777–810.

Hartmann-Wendels T., Honal M. (2010). Do economic downturns have an impact on the loss given default of mobile lease contracts? – An empirical study for the German leasing market Credit and Capital Markets, 43 (1) (2010), pp. 65-96.

Harwood, S. (2006). Shipping Finance, 3rd Edition. London: Euromoney.

Hoerl, A.E. (1962) Application of Ridge Analysis to Regression Problems. Chemical Engineering Progress, 58, 54-59.

International Chamber of Shipping (2020). Annual Review 2020, Heroes at Sea. https://globalmaritimehub.com/wp-content/uploads/2020/10/Annual-Review-2020-Final.pdf

Kang W., Ratti R.A., (2013). Structural oil price shocks and policy uncertainty, *Economic Modelling*, Volume 35, Pages 314-319.

Kaposty F., Kriebel J., Löderbusch M. (2020). Predicting loss given default in leasing: A closer look at models and variable selection, *International Journal of Forecasting, Volume 36, Issue 2, 2020, Pages 248-266,*

Kavussanos, M. G., and Tsouknidis D. A. (2016). "Default Risk Drivers in Shipping Bank Loans." Transportation Research Part E: Logistics and Transportation Review 94: 71–94.

Kuhn, M. (2008). Building predictive models in R using the caret package. Journal of Statistical Software, Articles, 28(5), 1–26.

Kuhn, M. (2018). Caret: Classification and regression training. R package version 6.0-80, URL https://CRAN.R-project.org/package=caret.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. New York,Springer.

Kutin, N., Moussa, Z., Vallée, T., (2018). Factors behind the freight rates in the liner shipping industry. HAL archive Id: halshs-01828633. Available at: https://halshs.archives-ouvertes.fr/halshs-01828633.

Loterman, G., Brown, I., Martens, D., Mues, C., & Baesens, B. (2012). Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*, *28*(1), 161–170.

Lozinskaia, A., Merikas, A., and Penikas, H. (2017) Determinants of the probability of default: the case of the internationally listed shipping corporations, *Maritime Policy & Management*, 44:7, 837-858.

MacKay, D. J. C. (1992). Bayesian interpolation. Neural Computation, 4(3), 415–447.

Maitra, D., Chandra, S., Dash, S., (2020). Liner shipping industry and oil price volatility: Dynamic connectedness and portfolio diversification, Transportation Research Part E: Logistics and Transportation Review, Volume 138.

Mayr, T.P., Tamvakis, M.N. (1999). The dynamic relationships between paper petroleum refining and physical trade of crude oil into the United States. Maritime Policy Manage., 26 (2) (1999), pp. 127-136

Milborrow, S. (2018). earth: Multivariate adaptive regression splines. R package version 4.6.3, URL https://CRAN.R-project.org/package=earth.

Mitroussi, K., W. Abouarghoub, J. J. Haider, S. J. Pettit, and N. Tigka. (2016). "Performance Drivers of Shipping Loans: An Empirical Investigation." *International Journal of Production Economics* 171 (3): 438–452.

Narayan, P.K., Sharma, S.S. (2011). New evidence on oil price and firm returns. *Journal of Banking & Finance*, 35 (12), pp. 3253-3262

OECD (2021), Quarterly GDP (indicator). doi: 10.1787/b86d1fc8-en (Accessed on 25 September 2021)

R Core Team (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, https://www.R-project.org/.

Rodriguez, P. P., & Gianola, D. (2016). Bayesian regularization for feed-forward neural networks. Working Paper

Qi, M., & Zhao, X. (2011). Comparison of modeling methods for Loss Given Default. *Journal of Banking and Finance*, *35*(11), 2842–2855.

Stopford, M. (2009). *Maritme Economics* 3e. New York: Taylor & Francis e-Library.

Sun, X., Liu, H., Zheng, S., Chen, S., (2018). Combination hedging strategies for crude oil and dry bulk freight rates on the impacts of dynamic cross-market interaction. Maritime Policy Management, 45 (2) (2018), pp. 174-196.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the royal statistical society series b-methodological, 58, 267-288

Yao, X., Crook, J., & Andreeva, G. (2015). Support vector regression for loss given default modelling. *European Journal of Operational Research, 240*(2), 528–538.

Zannetos, Z. (1966). *The theory of oil tankship rates: an economic analysis of tankship operations.* Cambridge, USA: MIT Press.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B: Statistical Methodology, 67(2), 301-320

Zhang J., Thomas L.C. (2012). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. *International Journal of Forecasting, 28 (1) (2012), pp. 204-215.*