Mersin University Journal of Maritime Faculty

Mersin University Journal of Maritime Faculty (MEUJMAF)
Vol. 3, Issue 1, pp. 1-8, June 2021
ISSN 2687-6612, Turkey
DOI: 10.47512/meujmaf.923874
Research Article

MODELING OF GENERAL CARGO SHIP'S MAIN ENGINE POWERS WITH REGRESSION BASED MACHINE LEARNING ALGORITHMS: COMPARATIVE RESEARCH

Fatih Okumuş *[1] and Araks Ekmekçioğlu [2]

[1] Yıldız Technical University, Department of Naval Architecture and Marine Engineering, Istanbul, Turkey
ORCID ID 0000-0001-8414-5802
hfatihokumus@gmail.com

[2] Yıldız Technical University, Department of Marine Engineering, Istanbul, Turkey
ORCID ID 0000-0002-4821-0272
araks@yildiz.edu.tr

* Corresponding Author
Received: 21/04/2021 Accepted: 09/06/2021

ABSTRACT
This study, which allows estimating main engine power of new ships based on data from general cargo ships, consists of a series of mathematical relationships. Thanks to these mathematical relationships, it can be predicted main engine power according to length (L), gross tonnage (GT) and age of a general cargo ship. In this study, polynomial regression, K-Nearest Neighbors (KNN) regression and Gradient Boosting Machine (GBM) regression algorithms are used. By this means the relationships presented here, it is aimed to build ships that are environmentally friendly and can be sustained at a lower cost by using the main engine power of the new ships with high accuracy. In addition, the relationships presented here provide validation for computational fluid dynamics (CFDs) and other studies with empirical statements. As a result of the study, polynomial regression gives similar results with other studies in the literature. We also concluded that while KNN regression yields fast results, GBM regression algorithm provides more accurate solutions to estimate the ship's main engine power.

Keywords: *Machine learning, Regression algorithm, General cargo ship, Engine power, Prediction*

1. INTRODUCTION

Seaway is the most efficient transportation method in terms of energy efficiency (Li *et al*. 2020). Therefore, the demand for shipping has increased dramatically since the mid-1990s. Today, 90% of the transportation in the world is made by ships (Kaluza *et al*. 2010). However, this rapid increase in shipping supply has also opened up environmental problems. Although other methods of transportation have been subject to considerable environmental scrutiny, shipping has largely gone unnoticed. As a result of these problems, the International Maritime Organization (IMO) tried to prevent it with the implementation of Annex VI of the International Convention on the Prevention of Pollution from Ships (MARPOL) in 1997. The MARPOL convention sets the limits for the main air pollutants such as nitrogen oxides (NOx), carbon dioxide (CO2), sulfur oxides (SOx), particular matter (PM) in the exhaust gases of ships. The reduction of the amount of these pollutants is directly related to the selection of the main and auxiliary engine of the ships in accordance with the working conditions. In addition to emission, the appropriate selection of the ship main engine power is also beneficial in reducing operating costs. In Stopford's study (Stopford 2008), fuel consumption accounts for about two-thirds of ship's cruising costs and more than 25% of a ship's total operating costs.

Machine learning is a technique that examines the work and systems of algorithms that can predict by performing assumptions using mathematical and statistical methods from the possible inputs. Machine learning, which creates a model by making predictions from sample inputs, is a sub-discipline of artificial intelligence (Gheibi, Weyns, and Quin 2021).

Looking at the research on the application of machine learning on the maritime industry in the literature C. Trozzi (Trozzi 2010) proposed a model based on gross tonnage and ship type to predict the ship main engine power. It also provided a ratio dependent main engine power to estimate auxiliary engine power. Requia *et al*. (Requia, Coull, and Koutrakis 2019) examined and analyzed $PM_{2.5}$ factors with Ordinary King (OK) interpolation, hybrid interpolation and machine learning (forest-based regression) techniques. They determined that the forest based regression model offers the best performance because of the R^2 value is higher than 0.7. Peng *et al*. (Peng *et al*. 2020) examined the energy consumption of ships in Jingtang port of China and denoted their strategies to diminish energy consumption and suggested prediction models. They used Random Forest Regression, the Gradient Boosting Regression, Liner Regression, BP Network and K-Nearest Neighbor Regression machine learning models and analyzed 15 features that have an impact on ships' energy consumption as input. They determined that net tonnage, deadweight tonnage (DWT), actual weight and efficiency of facilities are the four most essential features to foresee the energy consumption of the ships. T. Cepowski (Cepowski 2019) proposed regression models for prediction of main engine power for tankers, bulk carriers and container ships. He concluded that main engine power affected nonlinearly from DWT and TEU (Twenty-foot Equivalent Unit) while the speed effects a linear. Gkerekos *et al*. (Gkerekos, Lazakis, and Theotokatos 2019) performed the ships' fuel oil consumption prediction using with machine learning algorithms Support Vector Machines (SVMs), Random Forest Regressors (RFRs), Extra Trees Regressors (ETRs), Artificial Neural Networks (ANNs), and ensemble methods. They stated that their results may be useful for accurate prediction of ships fuel oil consumption. Also, R^2 of approximately 90% was obtained through the best performing modeling approaches. Yan *et al*. (Yan, Wang, and Du 2020) suggested fuel consumption prediction and fuel reduction model for a dry bulk ship. They set up a fuel consumption prediction model that takes into account the ship sailing speed, cargo weight, sea and weather conditions by using the random forest regressor. They concluded that the requested model could reduce ship fuel consumption by 2-7% and the reduction in fuel consumption will also lead to lower CO_2 emissions. Uyanık *et al*. (Uyanık, Karatuğ, and Arslanoğlu 2020) studied that the fuel consumption optimization of a container ship with the help of multiple regression, ridge and lasso regression, support vector regression, tree-based algorithms and reinforcement algorithms. They compared the prediction models and stated that the predictions made by multiple regression and ridge regression yielded more accurate results. In addition, parameters such as main engine speed, cylinder values, cleaning air and shaft gauges were highly correlated with fuel consumption. Jeon *et al*. (Jeon *et al.* 2018) conducted a regression design using an artificial neural network (ANN) with big data analysis combining data acquisition, clustering, compression, and expansion to estimate host fuel consumption. In order to obtain a regression model with good predictions, they used various activation functions by changing the number of hidden layers and neurons in the ANN, and investigated the applications of regression analysis on efficiency and performance. Ekinci *et al*. (Ekinci *et al*. 2011) predicted the main design criteria in consequence of different machine learning methods in their studies. In the first part of the study, they determined the best / worst prediction criterion among all design parameter estimations.

There are many techniques in the literature that are used to calculate the total resistance or resistance components of ships. CFD (Computational Fluid Dynamics), panel methods, other numerical techniques, model experiments, experimental and statistical approaches are among the leading methods of these methods. In addition to these methods, the machine learning technique is also widely used in the literature to estimate ship main engine power. Some of these studies are summarized above. The most important feature that distinguishes this study from others is that the proposed algorithms offer acceptable results in more than one ship type. Thanks to the developing computer power, energy efficiency on ships is increasing day by day. The use of ship age within the entries will contribute to the preservation of the validity of the results in the future. This situation has been omitted in many studies in the literature. In addition, machine learning methods together with the inputs used make this study privileged

In this study, we use different machine learning-based regression methods in order to estimate the main engine power of the ships. The success of regression methods was determined with three different error methods: Root Mean Squared Error (RMSE), Mean

Absolute Error (MAE) and R-squared (R^2). In the next stage of work, they found which parameter was the most effective in estimating the main engine power and which machine learning method was the most successful. They stated that the best approximate parameter is length (LBP) and the worst is the velocity (V) and the most successful method is Model Trees (M5P).

2. MATERIALS AND METHOD

2.1. Data Set

In this study, data containing information from 2286 different general cargo ships were used. The data of the ships were collected by the authors. The dataset contains gross tonnage, year of manufacture, length, and the main engine power for each ship. While 80% of these data of these ships are used to train the model, 20% of them are used for testing. The gross tonnage of the ships varies between 74 and 162960. The oldest ship was produced in 1925, while the newest ship was built in 2018. The lengths of the ships were kept in a wide range from 18.25 m to 368 m. The main machine power and auxiliary machine power to be estimated vary between 147-72240 kW and 37-9600 kW, respectively. Table 1. provides statistical data on the ships.

Table 1. Statistical data of the data set

	Gross Tonnage	Length	Age	ME Power
Minimum	386	40.00	2	202
1St. Qu.	2811	95.63	11	1324
Median	5087	118.22	20	2880
Mean	11140	131.26	20.05	4226
3rd. Qu	16041	166.49	34	6480
Maximum	194817	333.00	56	36560

2.2. Accuracy control of predictions

The accuracy of the model's predictions is computed by comparing the actual power values of the main engine with the corresponding predicted values. Ten-fold cross-validation was used to as objectively and accurately evaluate the performance of the model. The dataset was randomly divided into ten parts. Nine of the detached parts were used to train the model, and one was used for testing. This process was repeated ten times, with each piece subject to testing. The predictive ability of the model is evaluated as the average performance of the model in all replicates. The Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) were used to determine the performance of the improved regression models.

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

As shown above, y_i and $\hat{y}_i$ respectively represent the actual power values and estimated power values. It is a quadratic metric that measures the magnitude of the error, often used to find the distance between the predictor's predicted values and the actual values of a machine learning model. RMSE is the standard deviation of prediction errors (residues). That is, residues are a measure of how far away the regression line is from data points. The RMSE value can vary from 0 to ∞, and the fact that its value is zero means that the model does not make any errors.

Average absolute error is an error measurement method used to control the difference between two continuous variables. The MAE controls the average vertical distance between the values predicted by the regression model and the best fit line between the actual values. Since the MAE value can be easily interpreted, it is frequently used in regression and time series problems. The MAE is a linear score reflecting the average magnitude of errors within a range of predictions, and all individual errors are equally weighted regardless of whether they are positive or negative. The MAE value can range from 0 to ∞. Negative focused scores i.e. lower valued estimators perform better. Analytical statement is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \quad (2)$$

R-squared measures the rate of variation in your dependent variable (Y) explained by your independent variables (X) for the regression model. Adjusts the adjusted R-squared statistic according to the number of independent variables in the model. The R^2 correlation coefficient is used to evaluate the performance of the models and is given as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(y_i - \overline{y}_i)} \quad (3)$$

$\overline{y}_i$ represents the mean value of y_i. It is a measure showing how close each data point is to the regression line with the R-Squared value. It is always positive and between 0 and 1.

2.3. Polynomial Regression

Regression is a method used to understand the relationship between one or more independent variables with a dependent variable. The dependent variable can be expressed with only one parameter, or it is possible to express it with more variables. If expressed in a model based on a single parameter, it is called a single regression, when expressed in two or more parameters, it is called multiple regression. Arguments do not always have to establish a linear relationship with the dependent variable. Some arguments can be expressed exponentially to increase the reliability of the model. Polynomial regression is used in such cases. For multiple exponents of the argument, the polynomial model is constructed as in Eq. (4).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2{}^2 + ... + \beta_p x_n{}^p + \varepsilon \quad (4)$$

In the equation, p refers to the polynomial degree of the independent variable, n refers to the number of independent variables.

In this part of the study, the polynomial degrees of the independent variables were investigated by using the data from the entire data set without any test-train distinction. Polynomial levels of the effects of three different independent (Length, Gross Tonnage, Age) variables on machine power were examined between one and five. Mean squares of error of polynomial levels were used to decide on the final model. Figure 1 also shows the mean square error of polynomial degrees. The expressions i, j and k are the polynomial levels of length, gross tonnage and age, respectively.

When figure 1 is examined, 2.nd degree polynomial is suitable for length, 5.th degree for gross tonnage and 2.nd degree for age. Table 2 contains RMSE, R^2, MAE errors about the polynomial model's train and test sets.

Table 2. Error values of the polynomial model.

	RMSE	R^2	MAE
Train	5174.02	0.808	3112.52
Test	5006.43	0.800	2955.65

2.4. K Nearest Neighbors – Regression

K-Nearest Neighbors (KNN) is one of the algorithms used for classification and regression in Supervised Learning. It is considered to be the simplest machine learning algorithm. With KNN, basically, the closest points to the new point are searched. K represents the amount of the closest neighbors of the unknown point. We choose k quantities of the algorithm (usually an odd number) to predict the results. KNN was used as a nonparametric technique in statistical prediction and pattern recognition in the early 1970s.

The KNN algorithm is predicted by the majority vote of its neighbors. The closest neighbors are found with a distance function. Eq. 5, 6 and 7 contains distance functions that are frequently used for regression (Chomboon *et al*. 2015)

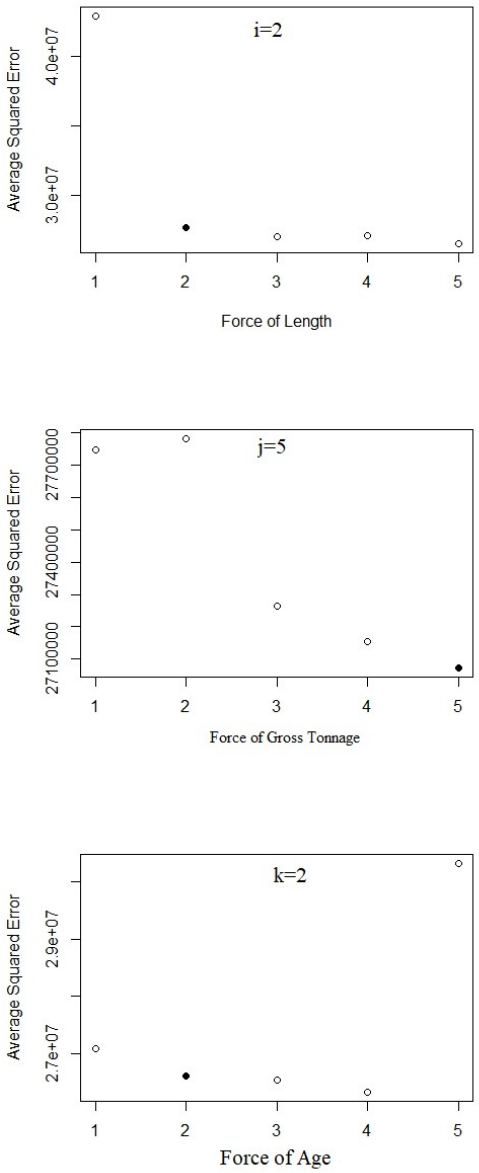

Fig. 1. The forces of independent variables

Euclidean $$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (5)$$

Manhattan $$\sum_{i=1}^{k}|x_i - y_i| \quad (6)$$

Minkowski $$\left(\sum_{i=1}^{k}\left(|x_i - y_i|\right)^q\right)^{1/q} \quad (7)$$

The three distance functions expressed in Equations 5, 6 and 7 are distance functions that can only be used in continuous variables. Generally, a large K value is more

sensitive as it reduces overall noise, there is no guarantee of time. Cross validation is another way to retrospectively determine a good K value using an independent data set to validate the K value.

In this part of the study, the number of neighbors was determined. Model 2 was designed to be used to estimate ship main engine power. The arguments used to estimate the outputs were not changed. To determine the number of neighbors, the number of neighbors between 1 and 10 were examined and determined according to RMSE values. Figure 2 shows the RMSE values of neighbor numbers.

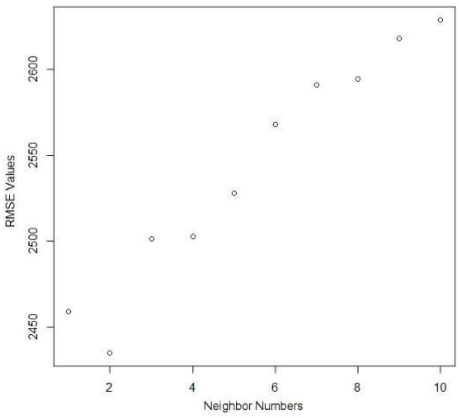

Fig. 2. RMSE values of neighbor numbers.

When Figure 2 is examined, the minimum error value for Model 2 is obtained when the neighbor number is 2. After the suitable neighbors were found, the model was trained with 80% of the data in the version set and tested on 20%. The results obtained were analyzed for both test and train sets with three different error calculation methods.

Table 3. Error values of the KNN model.

	RMSE	R^2	MAE
Train	989.76	0.925	396.36
Test	1536.01	0.839	676.415

2.5. Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) develops the conventional decision tree method by combining a statistical approach called augmentation. The main idea of this technique is to put together a series of "weak" models to create a single "strong" consensus model rather than creating an optimized model. In GBM, new decision trees are created sequentially, minimizing existing residual. Unlike standard regression models, in GBM, new decision trees are created by reducing the residuals at each step. In other words, optimization is made by adding trees in each step to reduce residues.

This method requires the most time as training time. Besides, there is a considerable amount of parameters that need to be determined from the outside. Initially, Model 3 was designed to estimate ship main engine power. Interaction dept, n.trees, shrinkage and n.minobsinnode variables were determined by tuning. Interaction depth 1 through 7 in 2 increments, n.trees between 1000 and 10,000 with 1000 increments, shrinkage value as 0.01 or 0.1 and n. minobsinnode value was searched between 10 and 20. The final values used for the model were n.trees = 3000, interaction.depth = 7, shrinkage = 0.01 and n.minobsinnode = 15. Figure 3 shows the effect of these variables on RMSE.

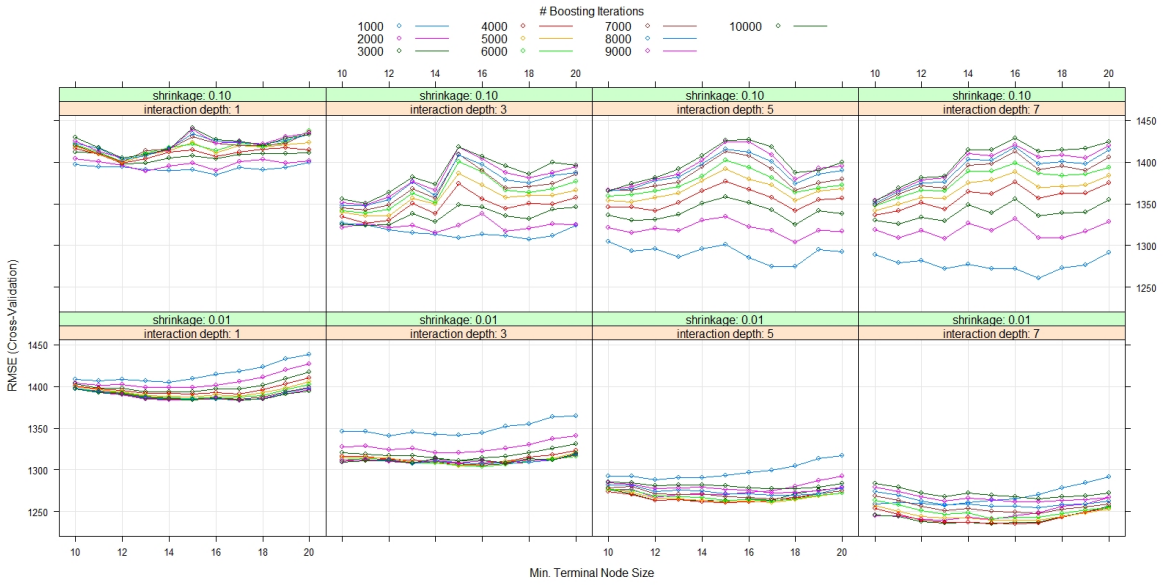

Fig. 3. Effects of variables on RMSE

The error rates for the final models created after the tuning process are listed in Table 4.

Table 4. Error values of the GBM model.

	RMSE	R^2	MAE
Train	408.49	0.987	273.41
Test	415.661	0.991	267.39

3. RESULTS AND DISCUSSION

Based on the length, gross tonnage and age data of 2286 different ships, the main engine power values were estimated in this study. While the gross tonnage values of the ships varied between 386 and 194817, their average was calculated as 11140. The length of the ship with the smallest length in the data set is 40 m, while the average length and maximum length values are 131.26 m and 333 m, respectively. In addition, the newest ship is 2 years old, while the oldest ship is 56 years old as of 2020. As a predictor, three different regression models (Polynomial, KNN and GBM) were studied. Models were trained in 80% of the test set and tested in 20%. The performance of the models was evaluated with ten-fold cross validation and RMSE, MAE and R^2 errors were calculated and interpreted.

In this study, a parametric study has also been done. For polynomial regression, the appropriate polynomial force was chosen for each independent variable. In addition, K value for KNN regression was examined at ten different levels and the optimum K value was determined as number 2. Finally, Interaction dept, n.trees, shrinkage and n.minobsinnode parameters were examined for GBM regression. The final values used for the model were n.trees = 3000, interaction.depth = 7, shrinkage = 0.01 and n.minobsinnode = 15.

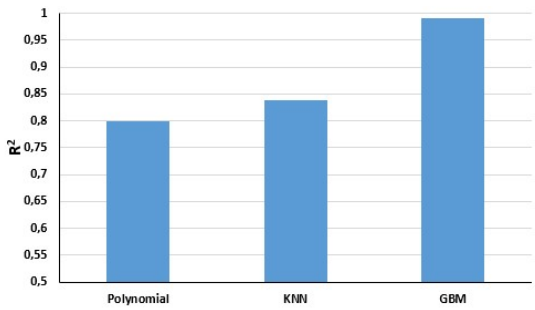

Fig. 4. Coefficients of determination of three models.

Basically, the closer R^2 error value is to one, the higher the success of the model. In Figure 4, the model contains R^2 error values calculated for three different models. As a result of the study, GBM algorithm has made the best approach to estimate main engine power of general cargo ships. The calculated R^2 value for the GBM algorithm is 0.991. However, the GBM algorithm is a method that takes quite a long time because it contains many variables. In addition, polynomial regression, which is a relatively easier method, has yielded results very close to the KNN algorithm and R^2 value is 0.800. Although KNN is quite simple in its application, it is not a suitable method for estimating main engine power of general cargo ships. Although KNN can show better results in small data sets, its success decreases in large data sets. The R^2 value obtained for KNN is 0.839.

Statistically, the mean absolute error (MAE) is a measure of errors between paired observations expressing for the same arguments. If the error value is close to zero, it means the success of the model. In Figure 5, the success of the models is evaluated on the basis of the MAE error value.

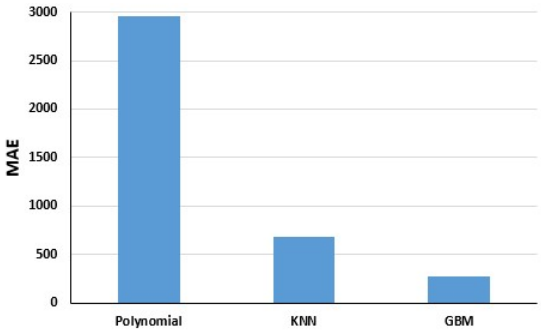

Fig. 5. Mean absolute error values of three models.

The MAE error values calculated for polynomial, KNN and GBM regressions are 2955.65, 676.41 and 267.39, respectively. The success criterion obtained in the R^2 error value did not change in the MAE. While GBM is the most successful algorithm in predicting the main engine power of general cargo ships, the weakest results are obtained by polynomial regression.

Another comparative criterion used in the study is RMSE, and similar to MAE, the success of the model increases as the error values approach zero. In Figure 6, the comparison of RMSE errors of the three models is visualized.

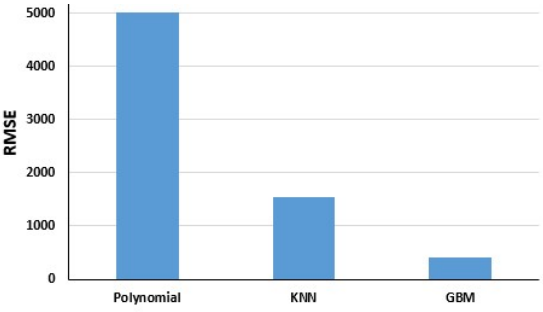

Fig. 6. Root mean squared error values of three models.

The relationship between the estimation data made with three different models and the actual data is given in Figure.7.

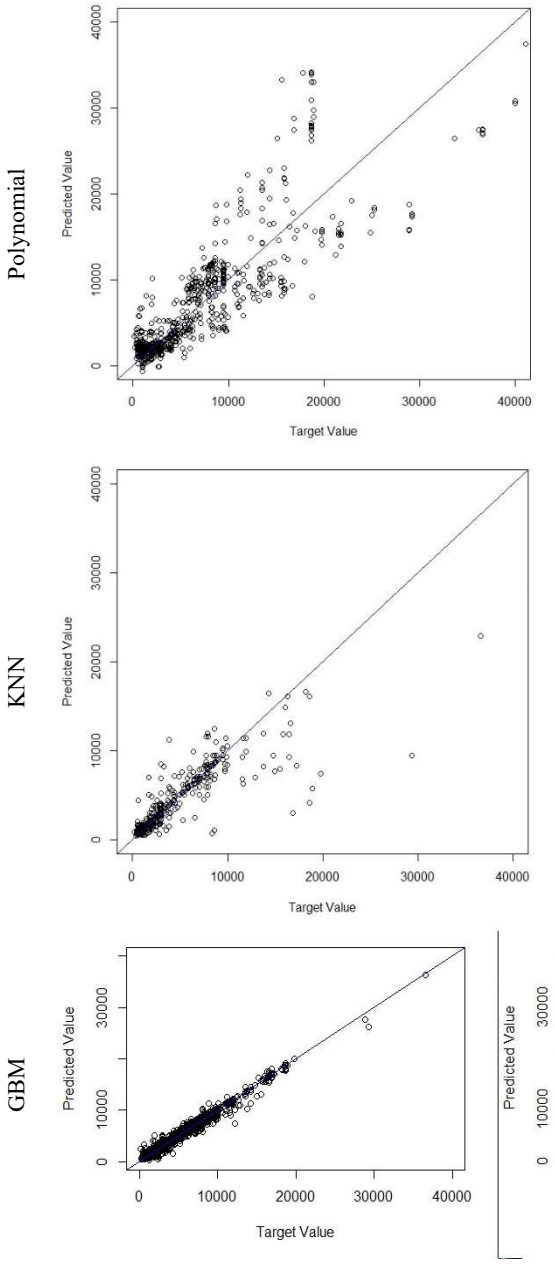

Fig. 7. Difference between target values and forecast values.

When Figure 7 is examined, it is seen that the points move away from the blue line when the predictive power of the models decreases. Also, Figure 7 provides information about the main engine power distribution of the ships in the data set.

Differences between actual values and estimated values are called residuals. The residual analysis method plays an important role in the validation of regression models, and enables the visualization of residuals. The difference between the values calculated with the help of models and the actual values is in Figure 8.

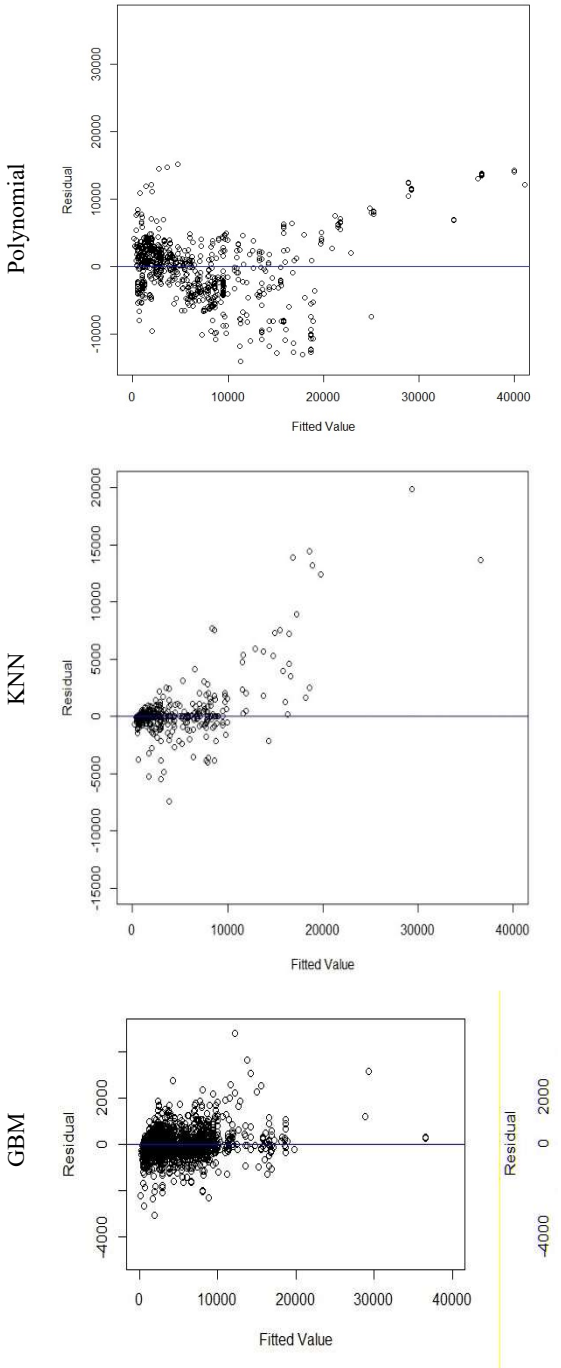

Fig. 8. Residuals.

In Figure 8, it is seen that the residuals increase away from the zero line in polynomial regression where the error rates are high. On the other hand, it is seen that GBM and KNN algorithms are located closer to the zero line thanks to their relatively high precision.

4. CONCLUSION

In this study, regression based algorithms (Polynomial, KNN, GBM) are used to estimate ship main and auxiliary engine powers. For each method, there are data preprocessing, data distribution determination, regression and performance evaluation steps, which are important stages of machine learning.

K-cross validation method was used to compare the performance of the models. Five different polynomial forces were investigated for each independent variable in the polynomial regression model. In addition, analyzes were performed to determine the optimum neighbor number for KNN regression and the optimum neighbor number was determined as number 2. In the study of 2286 general cargo ship samples, GBM was the algorithm that best predicted ship main engine power compared to R^2, RMSE and MAE. Although this method provided good results in the study, the excessive number of parameters to be determined externally and the time consuming nature appeared as the negative side of the method. Polynomial regression was revealed for three different error detection methods that it is not suitable for this data set. KNN regression could not exhibit the expected performance due to the large data set. The GBM regression is the optimum method for estimating the main engine power of general cargo ships, and it has proved highly sensitive.

REFERENCES

Cepowski, T. (2019). "Regression Formulas for The Estimation of Engine Total Power for Tankers, Container Ships and Bulk Carriers on The Basis of Cargo Capacity and Design Speed" *Polish Maritime Research,* 26 (1): 82–94. https://doi.org/10.2478/pomr-2019-0010.

Chomboon, K., Chujai, P., Teerarassammee, P., Kerdprasop, K. and Kerdprasop, N. (2015). "An Empirical Study of Distance Metrics for K-Nearest Neighbor Algorithm" *The 3rd International Conference on Industrial Application Engineering 2015 (ICIAE2015) Japan*, pp. 280–85. https://doi.org/10.12792/iciae2015.051.

Ekinci, S., Celebi, U.B., Bal, M., Amasyali, M.F. and Boyaci, U.K. (2011). "Predictions of Oil/Chemical Tanker Main Design Parameters Using Computational Intelligence Techniques" *Applied Soft Computing, The Impact of Soft Computing for the Progress of Artificial Intelligence*, 11 (2): 2356–66. https://doi.org/10.1016/j.asoc.2010.08.015.

Gheibi, O., Weyns,D. and Quin, F. (2021). "Applying Machine Learning in Self-Adaptive Systems: A Systematic Literature Review" *ArXiv*:2103.04112 [Cs], March. http://arxiv.org/abs/2103.04112.

Gkerekos, C., Lazakis, I. and Theotokatos, G. (2019). "Machine Learning Models for Predicting Ship Main Engine Fuel Oil Consumption: A Comparative Study" *Ocean Engineering*, 188 (September): 106282. https://doi.org/10.1016/j.oceaneng.2019.106282.

Jeon, M., Noh, Y., Shin, Y., Lim, O-K. Lee, I. and Cho, D. (2018). "Prediction of Ship Fuel Consumption by Using an Artificial Neural Network" *Journal of Mechanical Science and Technology* 32 (12): 5785–96. https://doi.org/10.1007/s12206-018-1126-4.

Kaluza, P., Kölzsch, A., Gastner, M.T. and Blasius, B. (2010). "The Complex Network of Global Cargo Ship Movements" *Journal of The Royal Society Interface,* 7 (48): 1093–1103. https://doi.org/10.1098/rsif.2009.0495.

Li, X., Sun, B., Guo, C., Du, W. and Li, Y. (2020). "Speed Optimization of a Container Ship on a given Route Considering Voluntary Speed Loss and Emissions" *Applied Ocean Research*, 94 (January): 101995. https://doi.org/10.1016/j.apor.2019.101995.

Peng Y., Liu, H., Li, X., Huang, J. and Wang, W. (2020). "Machine Learning Method for Energy Consumption Prediction of Ships in Port Considering Green Ports" *Journal of Cleaner Production*, 264: 121564. https://doi.org/10.1016/j.jclepro.2020.121564.

Requia, W.J., Coull, B.A., Koutrakis, P. (2019). "Evaluation of Predictive Capabilities of Ordinary Geostatistical Interpolation, Hybrid Interpolation, and Machine Learning Methods for Estimating PM2.5 Constituents over Space" *Environmental Research*, 175, pp. 421–33. https://doi.org/10.1016/j.envres.2019.05.025.

Stopford, M. (2008). *Maritime Economics*, 3e (3rd ed.) Routledge. https://doi.org/10.4324/9780203891742.

Trozzi, C. (2010). "Emission Estimate Methodology for Maritime Navigation" Co-Leader of the Combustion & Industry Expert Panel.

Uyanık, T., Karatuğ, Ç. and Arslanoğlu, Y. (2020). "Machine Learning Approach to Ship Fuel Consumption: A Case of Container Vessel" *Transportation Research Part D: Transport and Environment*, 84 (July): 102389. https://doi.org/10.1016/j.trd.2020.102389.

Yan R., Wang S., Du Y. (2020). "Development of a Two-Stage Ship Fuel Consumption Prediction and Reduction Model for a Dry Bulk Ship" *Transportation Research Part E: Logistics and Transportation Review*, 138 (July 2019): 101930. https://doi.org/10.1016/j.tre.2020.101930.