Explainable Machine Learning for Predicting Maritime Fuel Consumption and CO_2 Emissions with Minimized Computational Cost

Konstantinos Zervoudakis[1][0000-0002-2146-1493] and Stelios Tsafarakis[2][0000-0001-5535-4787]

School of Production Engineering and Management, Technical University of Crete, Chania, Greece
[1]kzervoudakis@tuc.gr
[2]tsafarakis@tuc.gr

Late Breaking Abstract

Fuel consumption and CO_2 emissions from maritime vessels are critical indicators both for operational cost efficiency and environmental sustainability. With tightening international regulations and the pressing need to reduce greenhouse gas emissions, predictive models that can anticipate fuel usage and emissions are very important. Machine Learning (ML), even though most of the times being computationally expensive, offers robust capabilities to analyze complex operational datasets and uncover patterns linked to fuel efficiency and emission levels (Bansal et al., 2022). The aim of this work is to develop an explainable, optimized ML approach to accurately predict ship fuel consumption and CO_2 emissions, thus supporting informed operational and regulatory decisions with the lowest computational cost as possible.

The main contribution of this work is to combine ML with optimization techniques that lie in the field of Computational Intelligence, for hyperparameter tuning and feature selection to maximize a performance metric while reducing the computational cost of the ML model. The optimized model is then analyzed using SHapley Additive exPlanations (SHAP) to explain how different features affect the predictions (Kim & Kim, 2022). This combined approach creates models that are not only accurate but also transparent, providing actionable insights for maritime vessels.

In this study, several ML regressors are used to predict ship fuel consumption and CO_2 emissions. These included Linear Regression (LR), Decision Trees (DT), Random Forest (RF), Gradient Boosting (GB) and Support Vector Machine (SVM). The models were trained on Ship Fuel Consumption & CO2 Emissions data retrieved from Kaggle containing detailed information of 1441 vessels, including records of ship voyages, like speed, draft, displacement, distance traveled, and the corresponding fuel consumption and CO_2 emissions. Cross-validation is also used to evaluate the performance of each model.

To improve the performance of the models a Non-dominated Sorting Genetic Algorithm II (NSGA II) was used to tune hyperparameters and select import features simultaneously. The goal was to maximize the R^2 regression metric, while reducing the computational complexity of the models. For that reason, several computational complexity

objective functions were coded specifically for each ML model. These functions allowed us to quantify the computational demands of training and using each algorithm, so that we could explicitly incorporate computational cost into the multi-objective optimization process together with the predictive performance. For example, regarding Random Forest, the computational complexity function took into account key factors such as the number of trees in the ensemble, the number of samples and features in the training data, the number of features considered at each split, and the maximum depth of the trees. This way, the function produced an estimate of how computationally intensive a Random Forest model is. In this way, it was ensured that the optimization process was guided not only by the goal of achieving high predictive accuracy, but also by an assessment of the required computational resources.

SHAP is applied to the best-performing optimized ML model to identify which factors were the most important ones in predicting ship fuel consumption and CO_2 emissions to make the model's decisions more transparent and easier to understand.

The results show that NSGA-II was able to improve the performance of all ML models. The optimized RF regressor emerged as the best regression model, achieving an R^2 of above 0.96 for both ship fuel consumption and CO_2 emissions with the lowest computational cost as possible.

The SHAP analysis provided clear insights into the factors that most influenced ship fuel consumption and CO_2 emissions. A beeswarm plot showed that the distance and the type of ship were the strongest predictors. Dependence plots revealed that tanker ships had greater fuel consumption and CO2 emissions. Waterfall plots were also generated to show how individual feature contributions add up to the final fuel consumption and CO_2 emissions prediction for single vessels. These plots break down the prediction step by step, starting from the average model output and adding or subtracting the impact of each feature.

This study shows that combining machine learning with multi-objective optimization and SHAP explainability can lead to more accurate predictions with less computational cost.

Keywords: machine learning, maritime fuel emissions, explainable AI

References

Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. *Decision Analytics Journal*, *3*, 100071. https://doi.org/10.1016/J.DAJOUR.2022.100071

Kim, Y., & Kim, Y. (2022). Explainable heat-related mortality with random forest and SHapley Additive exPlanations (SHAP) models. *Sustainable Cities and Society*, *79*, 103677. https://doi.org/10.1016/J.SCS.2022.103677