Data Analysis and Machine Learning Approaches for forecasting Vessel Carbon Emissions

Sai Sandeep Illuri

MS in Maritime Technology and Management
National University of Singapore (NUS)
Singapore
saisandeep.illuri@u.nus.edu

***Abstract*—This report presents a comprehensive methodology for optimizing carbon emission predictions in global logistics using advanced machine learning techniques. The study integrates predictive modeling to enhance sustainability in maritime operations. The proposed methodology involves data preprocessing, exploratory data analysis, feature engineering and selection, ML Modeling, hyperparameter tuning, and evaluation. The experimental results with over 85% R2 score demonstrate the effectiveness of methods including Ensemble, Boosting and Neural Networks in improving predictive accuracy for reducing environmental impact. This research contributes to data-driven decision-making frameworks for sustainable logistics.**

***Index Terms*—Maritime Data Analytics, Supply Chain Management, Risk Optimization, Sustainability, Data Analytics, Machine Learning, Carbon Emissions.**

I. Introduction

The global maritime industry, responsible for over 80% of world trade by volume, significantly contributes to environmental challenges, accounting for 2-3% of global CO2 emissions. This environmental impact, coupled with rising concerns over climate change, has spurred the maritime sector to adopt sustainable practices. Reducing carbon emissions in maritime logistics has become critical, aligning with global sustainability goals while offering potential economic benefits. Emissions arise from factors such as fuel consumption, operational inefficiencies, vessel design, route planning, and cargo handling. In response, stakeholders are exploring innovative solutions to mitigate emissions while maintaining operational efficiency.

Recent advancements in data-driven techniques, particularly machine learning (ML), hold significant promise for optimizing carbon emissions in the maritime industry. By leveraging datasets collected from vessels, ports, and supply chains, ML algorithms can identify patterns, make predictions, and suggest actionable solutions for emissions reduction. Techniques like supervised learning, unsupervised learning, and reinforcement learning enable predictive models for forecasting emissions. This work aims to explore the application of ML methods to predict and optimize emissions in maritime logistics, offering a proactive, data-driven approach to achieving sustainability and operational efficiency.

II. Methodology

In this project, the goal is to predict the average emissions of maritime vessels based on a set of given features. The methodology includes data preprocessing, exploratory data analysis, feature engineering and selection, ML Modeling, hyperparameter tuning, and evaluation. Below are the detailed steps followed to achieve the objective:

A. Data Exploration and Understanding

The initial step was to understand and analyze the raw dataset. This exploration helped identify key characteristics and potential issues in the data that might affect model performance.

1) Regression Analysis: A linear regression model was initially tested to examine whether a linear relationship exists between features and emissions. It was found that while some features showed moderate linearity with emissions, a linear regression model was not the most suitable for this task.

2) Outlier Detection: The dataset contained several extreme outliers across almost all features. After identifying these outliers, only extreme outliers were removed to maintain consistency between training and test datasets while ensuring the data distribution remains similar.

3) Correlation Analysis: The correlation matrix was computed, revealing that no features were highly correlated. However, three features showed notable correlations with the target variable (EMISSION), suggesting their potential importance for the predictive models.

4) Distribution Comparison: A comparison of distributions between the training and test data revealed that both datasets followed similar patterns, indicating that the model trained on the training data would generalize well to the test data.

5) Multi-Collinearity: A moderate level of multicollinearity was present among some features, which may be handled by employing techniques like appropriate ML Models and Principal Component Analysis (PCA) to reduce dimensionality and avoid overfitting.

B. Data Preprocessing

Data preprocessing was crucial for ensuring that the models could efficiently learn from the data. The following preprocessing steps were performed:

- Efficiency column: The 'Efficiency' column was one-hot encoded to convert categorical values into numerical format for machine learning models.
- Type column: Similarly, the 'Type' column was one-hot encoded to handle categorical data.
- Feature Removal: Unnecessary features such as the 'IMO', 'NAME', 'EFFICIENCY', 'REGISTERED' were removed, as they were not relevant for emission prediction.
- Outlier Removal: Extreme outliers identified in the previous step were removed from the training dataset to make the data more consistent and suitable for modeling.
- PCA & Feature Engineering: PCA was applied for dimensionality reduction, but it was found that the resulting features did not significantly improve the model's performance. Feature selection techniques were also explored but did not provide substantial improvements.

C. Model Selection and Training

Multiple ML models were trained and evaluated to find the best-performing model for predicting emissions. The models included linear and nonlinear regressors:

1) Linear Models:

- Linear Regression
- Lasso Regression
- Ridge Regression

2) Non-linear Models:

- Support Vector Machines (SVM) Regression
- K-Nearest Neighbors (KNN) Regression
- Decision Tree Regression
- Random Forest Regression
- AdaBoost Regression
- Gradient Boosting Machines (GBM)
- XGBoost Regression
- CatBoost Regression
- Optimized Deep Neural Networks (DNN)
- TabNet (a deep learning model)

D. Model Evaluation

Each model was evaluated based on its Root Mean Squared Error (RMSE) and R^2 score to assess performance. The results from different models were as follows:

1) Best Models:

- CatBoost Regression: RMSE = 64.83, R^2 = 0.8673
- Random Forest Regression: RMSE = 65.07, R^2 = 0.8663
- Gradient Boosting Regression: RMSE = 66.73, R^2 = 0.8594

2) Other Models: Several models like Linear Regression, Lasso, Ridge, SVM, and KNN performed well, but none of them outperformed CatBoost, Random Forest, or Gradient Boosting.

E. Hyperparameter Tuning and Optimization

To further improve the models, hyperparameter tuning was performed using grid search and cross-validation for models such as Random Forest, CatBoost, and Gradient Boosting. This helped fine-tune the models and yield the best results.

F. Final Model and Submission

After selecting CatBoost as the final model, predictions were generated on the test data. These predictions were saved in the required submission format (CSV with 'Id' and 'Predicted' columns) and uploaded to the Kaggle competition for evaluation.

The methodology followed in this project ensured the careful selection and optimization of models to predict emissions in maritime vessels, achieving high accuracy and minimal error, with the CatBoost model providing the best performance in both RMSE and R^2 scores, along with Random Forest.

III. Results and Discussion

The results demonstrate that tree-based ensemble methods (Random Forest and CatBoost) outperform traditional regression models in terms of predictive accuracy. Basic ML models delivered strong performance with lower computational demands. While deep learning may not be necessary, exploring simple architectures were used for identifying complex patterns in the dataset. A comparative analysis of machine learning techniques is summarized in Table I.

TABLE I
COMPARISON OF MODEL PERFORMANCE

Model	RMSE	R^2
Linear Regression	78.99	0.80
Lasso Regression	78.99	0.80
Ridge Regression	78.98	0.80
SVM Regression	122.96	0.52
KNN Regression	75.28	0.82
Decision Tree Regression	81.88	0.79
Random Forest Regression	65.73	0.86
AdaBoost Regression	111.25	0.61
CatBoost Regression	64.85	0.87
Best Gradient Boosting	66.67	0.86
Best XGBoost	66.41	0.86
Best Random Forest	65.05	0.87
Best CatBoost	64.83	0.87
CV CatBoost	70.83 ± 7.01	0.85 ± 0.01
DNN Model	71.37	0.84
TabNet	69.75	0.85

IV. Conclusion

This study acknowledges the potential of advanced machine learning models, particularly Random Forest and CatBoost, in optimizing carbon emission predictions for maritime logistics. The results show that these models provide decent improvements in predictive accuracy over traditional regression approaches. Although CatBoost outperformed, Random Forest showed almost same results. Basic ML models performed good with lower computational needs. Deep learning may not be required, but use of simple architectures could be a valuable research focus. These AI-driven strategies can optimize maritime operations, reducing carbon emissions and enhancing sustainability for green logistics.