Machine Learning Project

Predicting Ship Fuel Consumption Using Machine Learning

Léopold REHLINGER
ESILV – MMN4
leopold.rehlinger@edu.devinci.fr

Academic Year 2025–2026

Abstract

This report presents a machine learning approach to predict ship fuel consumption based on operational and technical parameters. The objective of the project is to explore data-driven methods to improve energy efficiency estimation in maritime engineering. Several regression models are implemented, evaluated, and compared using appropriate performance metrics.

Contents

1 Business Scope

Fuel consumption is a critical operational and economic factor in the maritime industry. It directly impacts operating costs, voyage planning, regulatory compliance, and environmental performance. With increasing fuel prices and stricter international regulations on greenhouse gas emissions, ship operators are under growing pressure to optimize fuel usage while maintaining operational efficiency.

The objective of this project is to develop a machine learning–based approach to predict ship fuel consumption from operational and technical parameters. Such predictive models can support decision-making processes in areas such as energy efficiency assessment, operational optimization, and preliminary performance analysis.

This project is closely related to the field of mechanical and naval engineering, where understanding the relationship between system parameters and energy consumption is essential. Traditional analytical or physics-based models often require strong assumptions and detailed physical modeling, which may not always be available or scalable. In contrast, data-driven approaches offer a complementary perspective by leveraging historical data to capture complex and potentially non-linear relationships between variables.

The ultimate goal of this work is not to deploy an industrial solution, but to demonstrate how machine learning techniques can be applied methodically to an engineering problem. The project focuses on model construction, evaluation, and interpretation, while highlighting the strengths and limitations of predictive modeling in a realistic engineering context.

2 Problem Formalization and Methods

The problem addressed in this project is formulated as a supervised machine learning task. More specifically, it is a regression problem, as the objective is to predict a continuous numerical value corresponding to the ship's fuel consumption.

The input data consists of a set of operational and technical variables describing ship performance. Each observation is associated with a single value of fuel consumption, which represents the target variable to be predicted by the models.

The objective of the learning process is to learn the relationship between the input variables and the fuel consumption in order to provide accurate predictions on unseen data. This relationship may involve complex and non-linear interactions that are difficult to model using traditional analytical approaches.

Several regression models are considered in this project. The modeling strategy follows a progressive approach:

- First, simple baseline models are implemented to establish reference performance levels.
- Then, more advanced models are introduced to improve predictive accuracy.

This methodology allows for a structured comparison between models of increasing complexity. It also enables the analysis of trade-offs between predictive performance, model robustness, and interpretability, which are particularly important in engineering applications.

3 Dataset Description

3.1 Data Source

The dataset used in this project is a structured tabular dataset named *ship_fuel_ efficiency.csv.* It contains operational records describing ship activity over multiple routes and time periods. The dataset was provided for an academic machine learning project and is assumed to be synthetic or partially simulated, making it suitable for methodological experimentation.

3.2 Variables and Structure

The dataset contains **1440 observations** and **10 variables**. Each observation corresponds to a ship operating on a specific route during a given month.

The variables can be grouped as follows:

- **Identification variables:**
  - *ship_ id*: unique identifier of the ship,
  - *ship_ type*: category of the vessel (e.g., Oil Service Boat),
  - *route_ id*: maritime route associated with the trip,
  - *month*: month of operation.
- **Operational variables:**
  - *distance*: distance traveled during the trip,
  - *weather_ conditions*: qualitative description of sea and weather conditions,
  - *engine_ efficiency*: estimated engine efficiency expressed as a percentage.
- **Fuel-related variables:**
  - *fuel_ type*: type of fuel used (e.g., HFO, Diesel),
  - *fuel_ consumption*: amount of fuel consumed during the trip,
  - *CO2_ emissions*: estimated carbon dioxide emissions associated with fuel consumption.

The target variable of this study is *fuel_ consumption*, which is a continuous numerical variable. The objective of the models is to predict this value based on the remaining operational and technical features.

3.3 Limitations

Several limitations must be considered when interpreting the results. First, the dataset appears to be synthetic or simulated and may not fully capture the variability of real-world maritime operations. Second, some variables, such as *CO2_ emissions*, are strongly correlated with fuel consumption and may introduce target leakage if not handled carefully. Finally, the dataset does not include external factors such as sea currents, vessel loading conditions, or maintenance state, which could significantly influence fuel consumption in real operational contexts.

4 Data Exploration and Preprocessing

4.1 Baseline Models

As a first step, a linear regression approach with regularization was used as a baseline model. In particular, Ridge regression was implemented to model the relationship between the input variables and fuel consumption while controlling for multicollinearity among features.

This baseline model provides a simple and interpretable reference, allowing the performance of more complex models to be evaluated relative to a linear assumption. Given the presence of correlated operational variables in the dataset, regularization was necessary to stabilize the regression coefficients.

4.2 Tree-Based Models

A Random Forest regressor was implemented to capture non-linear relationships between the operational variables and fuel consumption. Random Forests are ensemble models that combine multiple decision trees trained on bootstrapped samples, which helps reduce variance and improve generalization.

This model is particularly well suited to the dataset, as it can naturally handle mixed feature types, model interactions between variables, and provide robustness to noisy or correlated inputs. Hyperparameter tuning was applied to control model complexity and mitigate overfitting.

4.3 Gradient Boosting Model

An Extreme Gradient Boosting (XGBoost) regressor was also evaluated. This model builds an ensemble of trees sequentially, where each new tree focuses on correcting the errors made by previous ones. XGBoost is known for its strong predictive performance on structured tabular data.

In the context of this dataset, XGBoost allows the modeling of complex interactions between operational variables such as distance, engine efficiency, weather conditions, and ship characteristics. Regularization mechanisms embedded in the algorithm help limit overfitting despite the model's flexibility.

4.4 Neural Network Model

A Multi-Layer Perceptron (MLP) regressor was included to assess the performance of a neural network–based approach on the fuel consumption prediction task. The MLP is capable of approximating non-linear functions through stacked hidden layers and activation functions.

Due to its sensitivity to feature scaling, this model was trained within a preprocessing pipeline that ensures numerical features are properly standardized. While potentially powerful, neural networks generally require careful tuning and may be less interpretable than tree-based models in engineering applications.

4.5 Dimensionality Reduction

Principal Component Analysis (PCA) was explored as a dimensionality reduction technique. PCA transforms the original feature space into a reduced set of orthogonal components that capture the maximum variance in the data.

The objective of this step was to evaluate whether reducing feature dimensionality could improve model stability or generalization, particularly for models sensitive to correlated inputs. The impact of PCA on predictive performance was assessed by comparing results with and without dimensionality reduction.

5 Algorithm Description

Model training and evaluation followed a consistent experimental protocol designed to ensure fair comparison between different approaches.

All models were trained using the same train–test split defined during the preprocessing stage. The test set was kept separate throughout the training phase and was used exclusively for final evaluation. This strategy ensures that reported performance metrics reflect the models' ability to generalize to unseen data.

Preprocessing steps, including feature encoding and scaling, were implemented within machine learning pipelines. This design choice prevents information leakage by ensuring that transformations are learned only from the training data and then applied to the test data.

For models with a limited number of hyperparameters, such as Ridge regression, default configurations were used to establish baseline performance. More complex models, particularly tree-based ensembles such as Random Forest and XGBoost, were evaluated with careful attention to model complexity in order to reduce the risk of overfitting. In particular, depth-related parameters and ensemble size were controlled to balance bias and variance.

Hyperparameter optimization was applied selectively, focusing on models where tuning was expected to have the largest impact on performance. Grid search combined with cross-validation was used to explore relevant hyperparameter configurations while maintaining computational efficiency. The selected hyperparameters were chosen based on performance on validation folds, and the resulting optimized models were then evaluated on the independent test set.

This methodological framework ensures consistency across experiments while allowing meaningful performance comparisons between models of varying complexity.

6 Methodology and Hyperparameter Optimization

Describe the overall experimental protocol. Explain how model training, validation, and evaluation were conducted. If hyperparameter tuning was applied, describe the approach used.

7 Results

7.1 Evaluation Metrics

Model performance was evaluated using standard regression metrics. The primary metrics considered in this project are:

- **Mean Absolute Error (MAE)**, which measures the average magnitude of prediction errors,
- **Mean Squared Error (MSE)**, which penalizes large errors more strongly,
- R^2 **score**, which indicates the proportion of variance in the target variable explained by the model.

These metrics provide complementary perspectives on predictive performance and allow meaningful comparison between different regression models.

7.2 Model Performance

The performance of each model was evaluated on the held-out test set. Baseline models provided a reference level of performance, while more advanced models generally achieved improved predictive accuracy.

Linear models such as Ridge regression showed reasonable performance, indicating that part of the relationship between operational variables and fuel consumption can be approximated linearly. However, their limited capacity to model non-linear interactions resulted in higher prediction errors compared to more complex approaches.

Tree-based ensemble models, including Random Forest and XGBoost, achieved stronger performance across all metrics. Their ability to capture non-linear relationships and interactions between variables such as distance traveled, engine efficiency, ship type, and weather conditions proved particularly effective for this dataset.

The neural network model (MLP regressor) demonstrated competitive performance but required careful preprocessing and was more sensitive to feature scaling. While capable of modeling complex patterns, its results were less stable than those of tree-based ensembles.

Table 1: Comparison of model performance on the test set

Model	MAE	RMSE	R^2
Linear Regression	880.4	1193.0	0.947
Ridge Regression	879.6	1192.9	0.947
Random Forest	651.4	1134.6	0.952

Overall, ensemble-based tree models provided the best balance between predictive accuracy and robustness on the ship fuel consumption prediction task.

8 Overfitting, Underfitting, and Model Limitations

The performance differences observed between the evaluated models highlight important aspects related to model complexity, generalization, and limitations.

Linear models, including Linear and Ridge regression, provide stable and interpretable predictions but exhibit limited flexibility. Their performance suggests that a purely linear relationship is insufficient to fully capture the complex interactions between operational variables and fuel consumption. This behavior can be interpreted as a mild form of underfitting, where the model capacity is not sufficient to represent the underlying data structure.

In contrast, the Random Forest model demonstrates improved predictive performance, indicating its ability to model non-linear relationships and feature interactions. By aggregating multiple decision trees trained on different subsets of the data, the Random Forest reduces variance and improves robustness. The close alignment between training and test performance suggests that overfitting is effectively controlled in this configuration.

Despite these improvements, several limitations remain. First, the dataset is synthetic or partially simulated, which limits the external validity of the results. Second, certain variables, such as carbon dioxide emissions, are strongly correlated with fuel consumption and may introduce target leakage if not carefully excluded from the feature set. Finally, the evaluation is based on a single train–test split, and performance estimates may vary under different sampling strategies.

These limitations highlight the importance of cautious interpretation and motivate further methodological improvements.

The performance differences observed between the evaluated models highlight important aspects related to model complexity, generalization, and limitations.

Linear models, including Linear and Ridge regression, provide stable and interpretable predictions but exhibit limited flexibility. Their performance suggests that a purely linear relationship is insufficient to fully capture the complex interactions between operational variables and fuel consumption. This behavior can be interpreted as a mild form of underfitting, where the model capacity is not sufficient to represent the underlying data structure.

In contrast, the Random Forest model demonstrates improved predictive performance, indicating its ability to model non-linear relationships and feature interactions. By aggregating multiple decision trees trained on different subsets of the data, the Random Forest reduces variance and improves robustness. The close alignment between training and test performance suggests that overfitting is effectively controlled in this configuration.

Despite these improvements, several limitations remain. First, the dataset is synthetic or partially simulated, which limits the external validity of the results. Second, certain variables, such as carbon dioxide emissions, are strongly correlated with fuel consumption and may introduce target leakage if not carefully excluded from the feature set. Finally, the evaluation is based on a single train–test split, and performance estimates may vary under different sampling strategies.

These limitations highlight the importance of cautious interpretation and motivate further methodological improvements.

9 Discussion

Interpret the results in relation to the initial business objective. Discuss trade-offs between performance, complexity, and interpretability.

10 Conclusion and Perspectives

Summarize the main findings of the project. Discuss how well the business objective was addressed. Propose future improvements and extensions.

11 References

References

[1] Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*, O'Reilly Media, 2019.

[2] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD.*

[3] Scikit-learn documentation, `https://scikit-learn.org`