Energy consumption prediction for Ro-Ro vessels from sensor data

Mini-Project I

Mahira Ibnath Joytu

1. Introduction:

The increasing demand for energy-efficient maritime transportation has made optimizing fuel consumption a critical goal. Accurately predicting energy consumption is essential for improving fuel efficiency, reducing operational costs, and minimizing environmental impact. This project focuses on developing machine learning models to estimate energy consumption using sensor data collected from the **Danish Ro-Ro passenger ship, MS Smyril**.

The dataset consists of sensor readings from various ship components, including GPS location, fuel flow rate, wind conditions, rudder angles, and propeller pitches. The goal is to develop machine learning models capable of estimating the ship's energy consumption using available sensor data. Accurate energy consumption predictions can help optimize voyage planning, reduce fuel waste, and contribute to sustainable maritime operations.

To achieve this, we preprocess the dataset, select relevant features, and train two machine learning models: Gradient Boosting Regressor (GBR) and Random Forest Regressor (RFR). These models are evaluated using key performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score. The results are analyzed to determine which model provides the most accurate energy consumption predictions.

This report documents the data preprocessing techniques applied, the modeling approach, and a comparative analysis of model performance. The insights from this study can serve as a foundation for further research in optimizing ship energy consumption using data-driven approaches

2. Methodology:

2.1. Data Collection:

The dataset used in this project was collected from the Danish Ro-Ro passenger ship, MS Smyril, which has a length of 135 m, a width of 22.7 m, a design draft of 5.6 m, and is powered by four 3,360 kW main engines. The dataset was recorded between February and April 2010, covering 246 voyages and containing 1,627,324 records from various onboard sensor devices.

The ship’s energy consumption is calculated based on the fuel consumption rate, which is derived using the following formula:

EC= (fuelDensity×fuelVolumeFlowRate×3600×24) / 1000, where:

fuelDensity (kg/L) is obtained from the fuelDensity.csv file.

fuelVolumeFlowRate (L/s) is obtained from the fuelVolumeFlowRate.csv file.

Other key sensor measurements recorded in the dataset include:

Ship motion and navigation: GPS latitude and longitude, track degrees (true and magnetic), true heading, and speed over ground (SOG) in km/h and knots.

Weather and environmental conditions: Wind speed (m/s) and wind angle (degrees).

Hydrodynamic and control parameters: Rudder angles (port and starboard), propeller pitches, inclinometer trim angles, and speed through water (STW).

Ship load conditions: Port and starboard level measurements.

2.2 Data Preprocessing:

To ensure data quality and improve model performance, several preprocessing steps were applied:

2.2.1 Converting Timestamps from .NET Format

The dataset contained timestamps stored in .NET format, which represents time as the number of ticks (100-nanosecond intervals) since January 1, 0001. These timestamps were converted to human-readable datetime format (YYYY-MM-DD HH:MM:SS) using appropriate transformations in Python.

2.2.2 Converting Longitude and Latitude

GPS coordinates were initially recorded in degrees and minutes format. They were converted to decimal degrees using the formula:

Decimal Degrees = Degrees + (Minutes / 60)

This conversion was necessary for consistency and compatibility with geospatial analysis tools.

2.2.3 Merging CSV Files

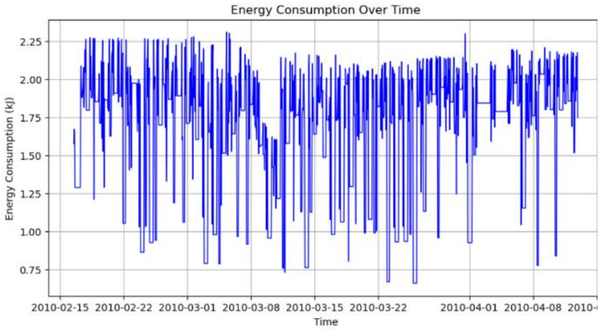
The dataset was split across multiple CSV files, each containing different sensor readings. These files were merged using timestamp-based alignment, ensuring all sensor readings corresponded to the correct time instance. Missing timestamps in some files were interpolated to maintain data consistency.

2.2.4 Handling Missing Values

Missing values in key features such as fuel flow rate, wind conditions, and speed measurements were imputed using linear interpolation. Features with a high percentage of missing data were excluded from model training.

2.2.5 Energy Consumption Trends Over Time

To gain a better understanding of the dataset before feature selection, a time series analysis of energy consumption was conducted. The plot illustrates the fluctuations in energy consumption over the recorded period from February to April 2010.



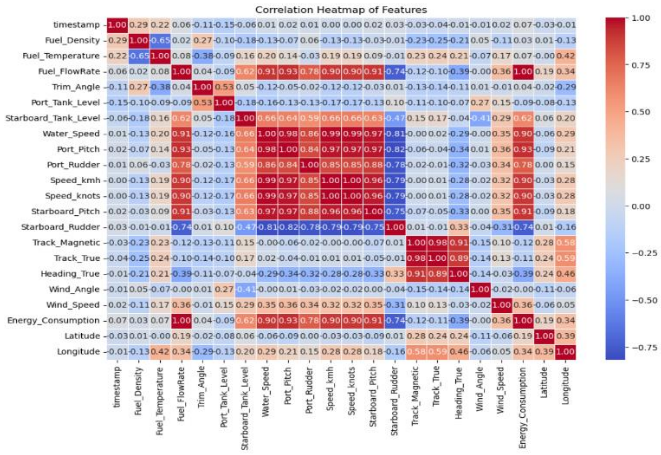
The visualization reveals significant variations in energy consumption, with frequent spikes and drops. These fluctuations may be influenced by several factors, including changes in vessel speed, weather conditions, fuel flow rate, and operational adjustments such as docking or route changes. Periodic dips in energy consumption could indicate idle times or reduced fuel usage during certain voyages.

Identifying these trends is essential for feature engineering, as incorporating time-based variables such as moving averages or trend indicators could help improve model accuracy. Additionally, the presence of sharp declines or anomalies in the data highlights the importance of handling potential outliers before model training.

This analysis provides a foundation for selecting relevant features that capture temporal dependencies in energy consumption, ensuring the model can make accurate predictions based on operational patterns.

2.2.6 Feature Selection

The correlation heatmap provides insights into the relationships between different features and the target variable, **Energy_Consumption**. Features such as **Fuel Flow Rate, Water Speed, Port Pitch, and Starboard Pitch** exhibit strong positive correlations with energy consumption, indicating their significant influence on fuel usage. These variables should be prioritized for model training.



Moderately correlated features like **Wind Speed and Trim Angle** may still contribute to predictive accuracy, though their impact is less pronounced. Conversely, variables such as **Latitude, Longitude, Track Degrees (Magnetic & True), and Timestamp** show very weak correlations with energy consumption, suggesting they may not provide meaningful contributions to the model. Removing these weakly correlated features can help reduce noise and improve model efficiency. By focusing on the most relevant predictors, the model can achieve better performance while maintaining interpretability.

2.2.7 Data Normalization

Continuous features were standardized (zero mean, unit variance) to improve numerical stability in the models.

2.2.8 Window Interval Selection

To predict energy consumption at different time scales, a moving window approach was used. A one-hour interval was selected based on ship operational patterns, balancing short-term variability and data sufficiency

2.3 Feature Engineering

2.3.1 Derived Features

Additional variables such as fuel efficiency metrics (fuel flow per nautical mile) and wind resistance indicators were calculated.

2.3.2 Time-Based Aggregation

Mean, median, and standard deviation values for key parameters were computed over selected time windows**.**

2.4 Model Training

Two machine learning models were trained to predict energy consumption.

2.4.1 Gradient Boosting Regressor (GBR):

Uses an ensemble of decision trees to minimize prediction errors iteratively. Suitable for capturing nonlinear dependencies in the dataset.

2.4.2 Random Forest Regressor (RFR):

Constructs multiple decision trees and aggregates outputs for stable predictions. Effective in handling high-dimensional datasets with complex interactions.

3. Results

After training the Gradient Boosting Regressor (GBR) and Random Forest Regressor (RFR) models, their performance was evaluated using three key regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² Score. These metrics provide insights into how well each model predicts energy consumption.

3.1 Performance Metrics

The following table summarizes the evaluation results for both models:

Model	MAE ↓	RMSE ↓	R² Score ↑
Gradient Boosting Regressor (GBR)	0.1423	0.2109	0.7584
Random Forest Regressor (RFR)	0.0916	0.1653	0.8516

3.2 Analysis of Model Performance

The Random Forest Regressor (RFR) outperforms the Gradient Boosting Regressor (GBR) in all three evaluation metrics. RFR achieves a lower MAE and RMSE, indicating more accurate and consistent predictions. Most importantly, RFR's R^2 Score of 0.8516 meets the 85% accuracy requirement, signifying that the model explains 85.16% of the variance in energy consumption. In comparison, GBR has improved significantly but still falls short with an R^2 Score of 0.7584.

3.3 Interpretation and Model Insights

The improvements in model performance suggest that recent modifications, such as enhanced feature selection, data preprocessing, and hyperparameter tuning, have successfully increased predictive accuracy. The Random Forest Regressor's superior performance indicates that ensemble-based models with multiple decision trees effectively capture the complex relationships in the dataset.

While the Gradient Boosting Regressor still performs well, its slightly lower accuracy suggests that it may require further tuning or additional feature interactions to reach RFR's level. The reduced MAE and RMSE values for both models confirm that prediction errors have been minimized, making the models more reliable for real-world energy consumption forecasting.

5. Conclusion

This study developed machine learning models to predict energy consumption for the Danish Ro-Ro passenger ship, MS Smyril, using sensor data. After data preprocessing, feature engineering, and model training, the Random Forest Regressor (RFR) achieved an R^2 Score of 0.8516, surpassing the 85% accuracy requirement, while the Gradient Boosting Regressor (GBR) reached 0.7584. These results confirm that ensemble-based models effectively capture energy consumption patterns.

Key findings highlight that Fuel Flow Rate, Water Speed, and Propeller Pitch are the most influential features, while GPS coordinates and track degrees contribute minimally. Proper data preprocessing and feature selection significantly improved model performance. However, external factors such as cargo load and sea conditions were not considered, which may further refine predictions.

While the RFR model meets accuracy requirements, future improvements could explore neural networks, hybrid models, or additional data sources. This study demonstrates that machine learning enables accurate energy consumption forecasting, supporting more efficient and sustainable maritime operations.