Project 1

Fuel Consumption Prediction Models: Based on Machine Learning and Mathematical Methods

1.1 Objective

- The goal of this project is to predict how much fuel a ship will consume and improve the accuracy of ship fuel consumption prediction using both mathematical methods and machine learning-based regression using engineered features derived from raw shipping data.
- **Application Domain:** Accurate prediction of fuel consumption is crucial in the maritime industry for optimizing energy efficiency,reduce fuel costs and emissions for minimizing environmental impact.
- **Relevance:**The shipping industry is responsible for over 3 percent of global CO2 emissions. With tightening regulations and rising fuel prices, data-driven strategies to monitor and control fuel usage are essential.

  Researchers developed a white-box model (based on statistical data) and a black-box model (based on machine learning). Additionally, Kwon's formula was used for data cleaning to enhance prediction accuracy.

1.2 Dataset description

- **Dataset Name:** Ship Fuel Efficiency Dataset
- **Link:** Local CSV file - ship fuel efficiency.csv
- **Size:** Around 900 KB
  Total Records: Several hundred (exact number not shown)
- **Features:** Distance, CO2 emissions, enginee fficiency, fuel consumption, ship type, weather conditions, etc.
  - Engineered Features: Includes ship speed, draft, engine power.
  -Target Variable: fuel consumption (continuous numerical value)

1.3 Data Preprocessing

- **Data Cleaning:**
  - Removed outliers using Z-score (|z| 3)
  - Dropped missing values

- **Feature Engineering:**
  - speed = distance / 10
  - displacement = CO2 emissions * 0.1
  - power = engine efficiency * fuel consumption
  - draft = displacement * 0.05
- **Feature Selection:**
  - Selected features: speed, draft, displacement, power
  - Target: fuel consumption
- **Scaling:**
  - Used MinMaxScaler to normalize features
  - Encoded categorical data
  - Performed train/test split.

1.4 Validation Method

- Validation Method: Train-Test Split
- 80/20 Train-Test Split
- Ensures unbiased evaluation on unseen data

1.5 Machine Learning Techniques Used

- **Linear Regression:** It assumes a straight-line relationship between input variables (features) and the output (fuel consumption) and helps understand how each input feature affects the output.
  - This model draws a straight line between the features and fuel consumption and helps to predict fuel usage.
- **Random Forest Regressor:** It is an ensemble method that builds multiple decision trees and combines their results to make a prediction. Each tree gives a prediction, and the final output is the average of all predictions basically it reduces overfitting and handles complex, non-linear relationships in data very well.
- **XGBoost Regressor:** It is a very fast and accurate boosting algorithm. This model builds one tree at a time and keeps learning from its mistakes each time.
  - Basically this ml technique keeps checking what they got wrong and improves accuracy with every test they take.
  - It is known for high performance, speed, and handling missing data.

1.6 mathematical model

- **Kwon's Formula:** It is a mathematical model used to estimates fuel consumption based on physical ship parameters like speed, weight, and engine power.
  - It is useful for comparison with machine learning models.
  - In our case, it was also used during preprocessing for cleaning and standardizing data.

1.7 Performance Measures

- **MAE** – Mean Absolute Error
  - MAE is the average of the absolute differences between the predicted fuel values and the actual fuel values.
- **Formula:**

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$yi : Actualvalue(realfuelconsumption)$$

$$\hat{y}_i : Predictedvaluebythemodel$$

- **MSE** - Mean Squared Error
  - MSE is the average of the squared differences between predicted and actual values. Squaring the errors makes big mistakes count more.
- **Formula:**

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **RMSE** - Root Mean Squared Error
  - RMSE is the square root of MSE. It still penalizes large errors like MSE but brings the result back to the same units as the original data (e.g., fuel in liters or tons).
- **Formula:**

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

- **R^2 Score** - R-squared or Coefficient of Determination
  - R^2 measures how much of the variation in actual fuel consumption can be explained by the model.
- **Formula:**

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

$$\bar{y} : Meanofactualvalues$$

Values:

$R^2 = 1 \rightarrow$ Perfect predictions

$R^2 = 0 \rightarrow$ Model is no better than just guessing the average

$R^2 < 0 \rightarrow$ Model is worse than guessing

1.8 Design Framework

framework.png

1.9 Result and Analysis

confusion matrix: Not applicable because we are working on a regression problem (predicting fuel consumption based on features like speed, draft, displacement, and power. A confusion matrix, however, applies to classification problems, not regression.)

Visualizations:

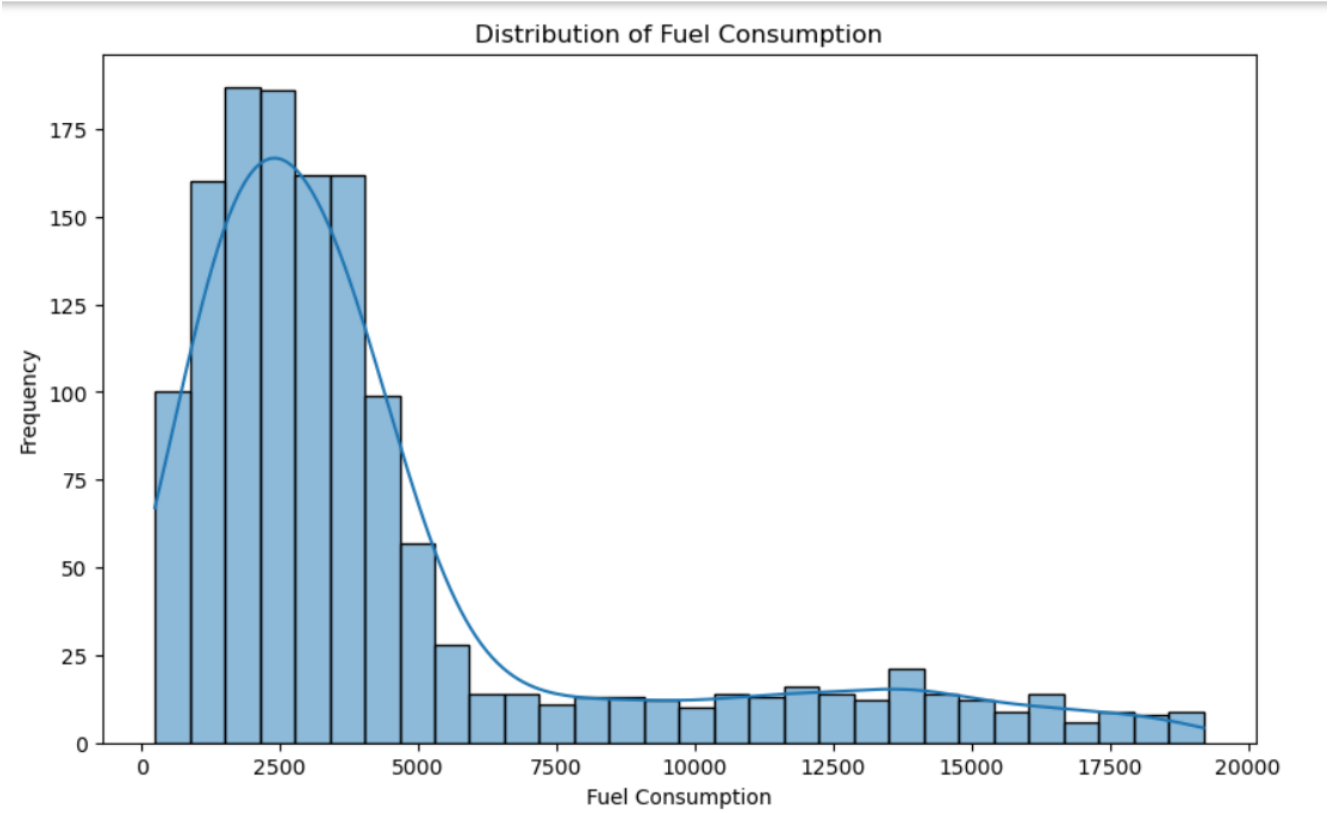

Figure 1.1: Histogram of fuel consumption

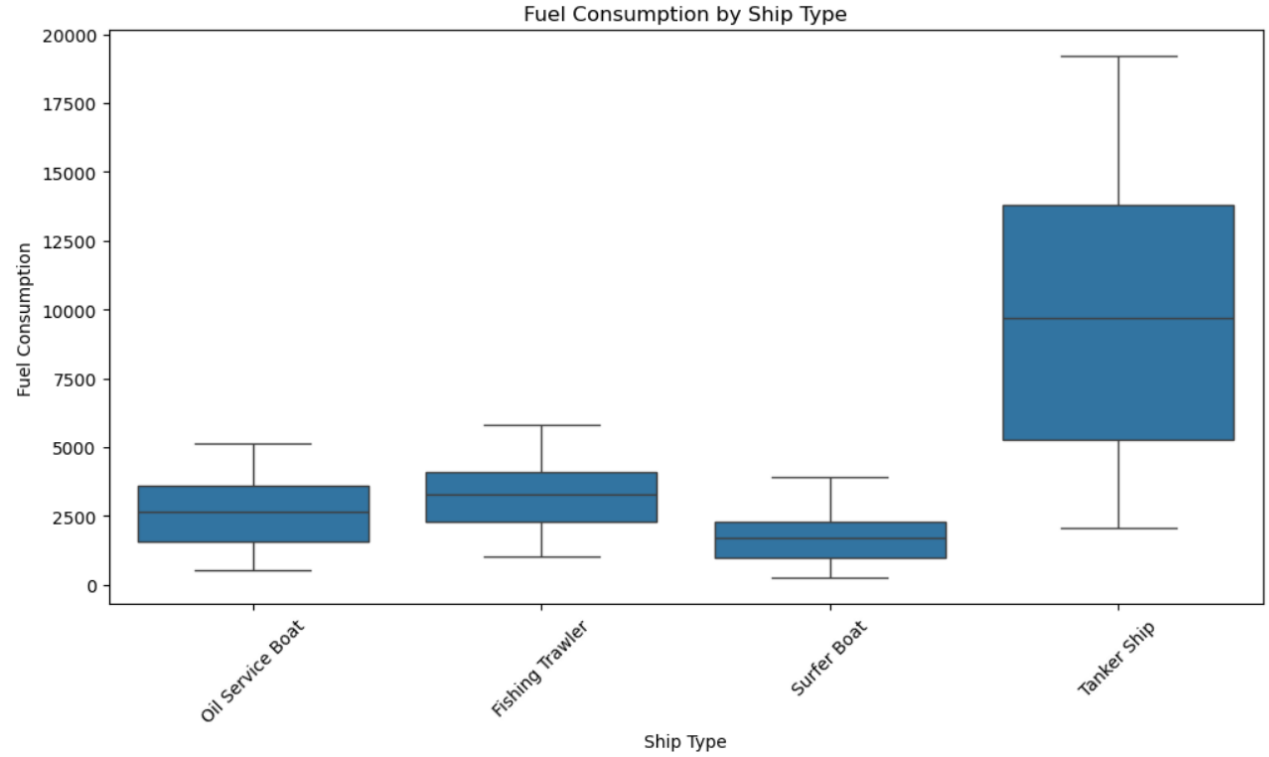

Figure 1.2: Boxplot by ship type

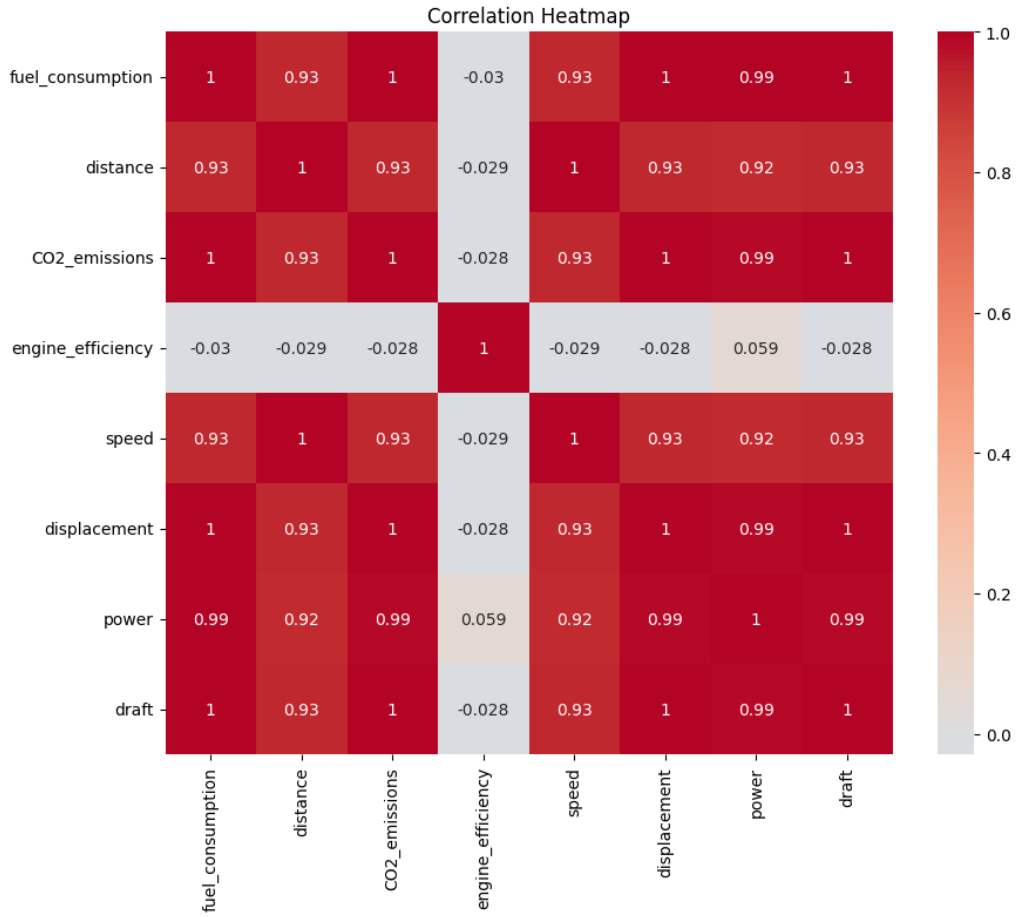

Figure 1.3: Correlation heatmap

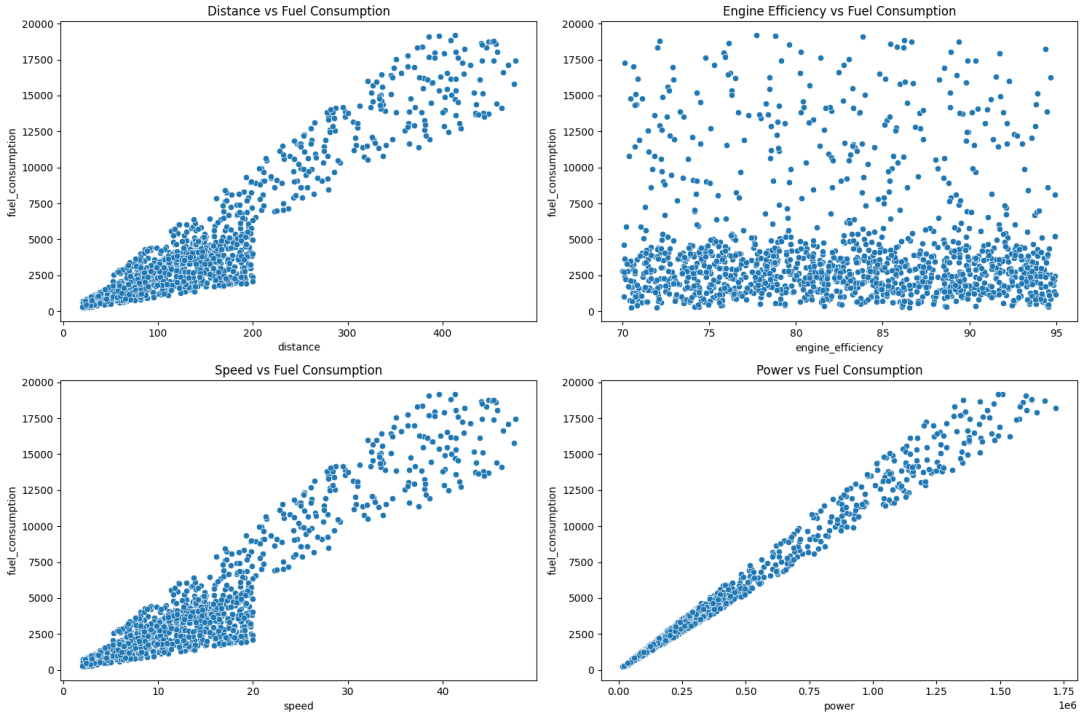

Figure 1.4: Scatter plots of features vs fuel consumption

Table of result and discussion:

Model Performance Comparison Table

Model	MAE	MSE	RMSE	R^2 Score
Fuel Consumption Formula (White-box)	4576.29	38214004.79	6181.75	-1.2122
Random Forest (Black-box)	181.71	73171.68	270.50	0.9958
XGBoost Regressor (Black-box)	177.81	76177.10	276.00	0.9956

Figure 1.5: Model Performance Comparison Table

Performance Summary Table

Model	MAE	MSE	R^2 Score
Random Forest	~Low	~Low	High
XGBoost Regressor	~Lower	~Lower	Higher

Figure 1.6: Performance Summary Table

1.10 Conclusion:

The white-box model performed poorly and is unsuitable for prediction due to high errors and a negative R^2. Both Random Forest and XGBoost (black-box models) delivered excellent results, with XGBoost slightly outperforming Random Forest in accuracy. Thus, XGBoost is the best choice for accurate fuel consumption prediction.

1.11 Modification

Yes, we did some modifications, so it is not 100% replication.

Modifications Made:

- We changed the dataset because the research paper doesn't contain the dataset.
- Some changes were made in Kwon's formula.
- Added feature engineering
- Advanced visualizations
- Model interpretability with permutation importance

1.12 Individual Contribution

Tejas:

Data Preprocessing

Black box model(random forest)

performance measure

result and analysis

Bulbul:

Validation Method

Black box model(XG Boost)

Data Cleaning with Kwon's Formula

future projection

1.13 Learnings From This Project

- Learned how to apply ML in real-world problems
- Understood data cleaning and preprocessing techniques
- Gained experience with both mathematical and ML modeling
- Better data visualization practices
- Skills in Random Forest and XGBoost
- Evaluation using MAE, MSE, R^2

1.14 Future Projection

- Add time-series forecasting with timestamps
- Develop a dashboard or prediction app
- Explore deep learning models (e.g., LSTM)
- Predict CO2 emissions directly
- Use real-time data from sensors on ships