# Final Project Report Round 2

Group partner:

Cheng Zhang(2054180)

Hongkun Liu(2068590)

### **Introduction**

The maritime shipping industry, crucial for global trade, relies on precise forecasting to optimize operations and ensure timely delivery of goods. In this context, our project embarked on leveraging advanced regression modeling techniques to predict future values in maritime shipping data. This endeavor aimed not only to enhance operational efficiency but also to navigate the complexities of maritime logistics with data-driven insights.

The project utilized a comprehensive dataset spanning from 1991 to the present day, marking a significant period for analysis due to the evolution of global trade patterns and shipping technologies. The dataset's time-series nature, incorporating monthly samples, presents a unique opportunity to apply predictive modeling for forecasting future outcomes based on historical trends.

### **Dataset Overview**

**Data Characteristics**:

* **Type**: Time series
* **Number of Features**: 463, including time as the first column, representing various aspects relevant to maritime shipping.
* **Number of Samples**: 397, each corresponding to a month's data from 1991 onwards.
* **Target IDs for Prediction**:
  + **Good Result**: IDs 542236, 67321
  + **Mid Result**: ID 549295
  + **Bad Result**: IDs 41108, 541982

Each row in the dataset encapsulates the monthly data of all features, serving as the independent variables for the model. The dependent variable or the target for each month is the intended value for the subsequent month, residing in the specified target column. This structure allows for a direct application of regression models to forecast future values based on the provided historical data.

**Data Preparation**: To prepare the dataset for modeling, it's pivotal to generate **X** (features) and **y** (target) matrices accurately. **X** encompasses all features for a given month, while **y** is a vector representing the target value for the next month. This preparation ensures that each sample in **X** aligns with its corresponding label in **y**, crucial for training predictive models.

**Training and Testing Sets**: Given the time series nature of the data, special attention was paid to avoid data leakage during model training and evaluation. The dataset was split into training and testing sets, with the last 36 months reserved for testing to assess the model's performance on recent data. This split respects the temporal sequence of the dataset, ensuring that the model learns from past data to predict future outcomes.

In summary, the dataset's comprehensive coverage of the maritime shipping industry, combined with careful preparation and consideration for time series analysis, sets a solid foundation for applying regression modeling techniques. The following sections of the report will delve into the methodology, model selection, results, and analysis, providing insights into the predictive modeling process and its implications for the maritime shipping industry.

The first Python script, **Linear Regression**, outlines the steps taken to apply a linear regression model to the maritime shipping dataset for forecasting. Here's a summary of the key components and actions within this script:

### **1.Linear Regression Summary**

* **Dataset Preparation**: The script begins by loading the dataset from a CSV file, **Sub\_Oil\_VLCC\_Monthly.csv**, and preparing the features (**X**) and the target variable (**Y**). It specifically targets one of the good result target IDs, "67321", for prediction. To align the features with the corresponding target values for next month's prediction, it drops the last row from **X** and the first row from **Y**.
* **Model Training and Testing**: It splits the data into training and testing sets based on the last 36 months, fitting a linear regression model on the training data. The script calculates predictions for the testing set.
* **Accuracy Calculation**: The script computes accuracy for each prediction using the formula **Accuracy = 100 \* (1 - abs((Y\_test - Y\_pred) / Y\_test))**, then calculates the average accuracy across all predictions.
* **Results Output**: Finally, it prepares an Excel file named **maritime\_shipping\_forecast\_results.xlsx** with the date, true target values, predicted values, and individual accuracies for the last 36 months. The script prints out the average accuracy of the model.

### **2. Polynomial Regression**

* **Model Description**: Applies a polynomial regression model of degree 2.
* **Data Preparation**: Similar to the linear regression script, it prepares **X** and **Y** with a focus on target ID "67321".
* **Evaluation**: Calculates the mean squared error (MSE), R2 score, and mean accuracy for the predictions.

### **3. Polynomial Regression with Normalization**

* **Model Description**: Extends the polynomial regression approach by incorporating MinMax normalization.
* **Target ID**: Changes to target a "bad result" ID "541982".
* **Normalization**: Applies MinMaxScaler to the features before performing polynomial regression.
* **Evaluation**: Computes the mean accuracy of the model predictions.

### **4. Lasso Regression**

* **Model Description**: Implements Lasso regression targeting the "bad result" ID "541982".
* **Parameter Tuning**: Utilizes a default alpha value of 1.0 for the Lasso model.
* **Evaluation**: Outputs the model's coefficients and calculates the mean accuracy of predictions.

### **5. Lasso Regression with Normalization**

* (The content for this script was not fully visible in the summary, but it likely follows a similar structure to the Lasso regression script with the addition of feature normalization.)

### **6. Random Forest**

* **Model Description**: Employs a Random Forest regressor targeting "bad result" ID "541982".
* **Configuration**: The model is configured with a default random state.
* **Evaluation**: The script calculates the mean accuracy for the Random Forest model's predictions.

Each script demonstrates a unique approach to handling the predictive modeling task, varying by regression model type, data preprocessing techniques (like normalization), and the specific target variable selected for prediction. These variations in methodology and target focus provide a broad overview of the modeling efforts undertaken to address the project's objectives.

The result table:

