**Final Project Report Round 3**

Group partner:

Cheng Zhang(2054180)

Hongkun Liu(2068590)

We have double check the output of dataset.

## 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.

## 2. Dataset Overview

### 2.1 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**: ID 542236, 67321

**Mid Result**: ID 549295

**Bad Result**: ID 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.

### 2.2 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.

#### 2.1.1 Lag

The lag parameter in regression, especially in time series analysis, plays a crucial role by incorporating the concept of time delay into the modeling process. It is used to account for the influence of past values on current values, a common phenomenon in sequential data. For this dataset, N rows can be concatenated to generate lag N and first N rows of X and last N rows of Y should be deleted to have the same row as the input. The snapshot of sample code for generating lag 1 is as follows:

文本

描述已自动生成

#### 2.1.2 Normalization

Normalization in regression is a critical preprocessing step involving the scaling of input features to ensure they have a standard scale or range. This process is essential because many machines learning algorithms, including regression models, perform better or converge faster when features are on a relatively similar scale and close to a normally distributed shape.

There are many kinds of techniques for Normalization, for this dataset, we select X Min-Max to scale all the features in X to a fixed range( 0 to 1). Following is the snapshot of implementing this kind of approaches:

文本

描述已自动生成

#### 2.1.3 PCA

Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction while preserving as much variance as possible. It's commonly used in data preprocessing for machine learning and data visualization tasks. It is worthy to mention that normalizing the data is a critical step before applying PCA, as PCA is sensitive to variances of the initial variables. Following is the snapshot of implementing this kind of approaches:

文本

描述已自动生成

### 2.3 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.

## 3.Regression Models

### 3.1 Linear Regression

The first model, 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:

* **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.

### 3.2 Polynomial Regression

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

### 3.3 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.

### 3.4 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.)

### 3.6 Random Forest

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

### 3.7 Decision Tree Regression

### 3.8 GB Regression

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.

### 3.7 Overall Results

The results of each model for 5 different IDs are summarized below:

## 4. Other Suggestions