**Project 1:**

Data type: time series

Number of features: 463 (the first column includes time)

Number of Samples: 397 (the first row includes feature IDs)

Target ID for prediction: 67321

Dear all,

I am excited to present an engaging project involving predictive modeling in the maritime shipping industry using a comprehensive dataset from 1991 to the present day. The ultimate objective is forecasting future values after the last recorded month.

This project offers a hands-on opportunity to apply regression modeling techniques to real-world data, emphasizing the challenges and nuances of forecasting in the dynamic maritime shipping industry.

The dataset is a tabular time series with a date in the first column. Each sample includes all features of a month as the input to the model (independent variables) and the intended value (in the specified column as a target) for the next month as the output of the model (dependent variable).

1. **Data preparation**

To generate X and y, we can consider the entire table as X and copy the target column in a vector as y. Remember that the label of sample t in X is in row t+1 in y.

A screenshot of a computer

Description automatically generated

So, neglecting the first row (IDs), you can delete the last sample of X (that doesn’t have any label) and also delete the first sample from y (that is not the label of any sample). This way, the sizes of X and y are the same, and the label of each sample is in the same row number. It is vital to double-check some random samples manually to be sure a sample and its label are in the right place in X and y. You can print some samples and check their labels in the initial Excel file.

Creating X\_train, y\_train, X\_test, and y\_test from X and y is very important. In time series, we must avoid data leakage, which means seeing a training sample between test samples or vice versa, considering the order of the samples.

A screenshot of a computer

Description automatically generated

Due to the temporal nature of time series data, we aim to assess the model's accuracy over the last three years available in the dataset. To achieve this, consider training the model on all samples from the beginning up to n-36, and then test it on the last 36 samples (n is the total number of samples excluding the last one that doesn’t include a label).

1. **Training and Testing the Model**

Train a regression model on the training set X\_train and evaluate its performance on the designated test set X\_test. You may start with a simple LinearRegression and improve your work using the other methods. I encourage you to explore different regression models, fine-tune parameters, and critically evaluate the model's performance.

1. **Evaluation**

Finally, create an Excel file comprising the data from the last 36 months for reporting and analysis. The first column should represent time, the second column the true target values, the third column the predicted values, and the fourth column the calculated accuracy using the formula:

Accuracy = 100\*(1-abs((actual-prediction)/actual))

Ultimately, the average accuracy across the 36 predictions will be calculated.

A screenshot of a computer

Description automatically generated

Please submit your code named your “group name” Version 1, an Excel file of the results, and a Word file of the report. I will add some tasks in the next rounds to the initial core I'm sharing now. You may modify or complete your code and update the results and the report every time.

Best of luck, and I look forward to your insightful analyses.

Regards,

Mehdi.