Data Science Intern at Data Glacier

Corrosion Detection & Severity Level Prediction Using Machine Learning

Week 8 : Deliverables

Name: Syed Sanaullah Shah

University: NUST, Islamabad

Email: iamkadhimi@gmail.com

Country: Pakistan

Specialization: Data Science

Batch Code: LISUM11

Date: 30 October 2022

Submitted to: Data Glacier

Table of Contents:

Problem Statement

Corrosion of pipelines has been recognized as a major predicament because it undermines the mechanical strength of pipelines, which can pose a disaster for the environment and the economy. Most of the time, the O&G industry uses mild steels. Ultrasonic testing during accelerated corrosion testing is a way to detect corrosion. AI is already using the scientific approach by popping up with hypotheses and doing simple tests to see if they are true. In the mid to long term, AI could have a big impact on how research is conducted, constructed, and acknowledged. However, using machine learning techniques to classify corrosion based on parameters like weight and thickness loss is still a new idea.

Our framework suggests using a new method that looks for corrosion through thickness and mass losses by NDT. With the help of machine learning methods and ultrasonic testing from accelerated corrosion testing, corrosion severity levels can be predicted based on data generated in the lab. Samples of mild steel could be evaluated for accelerated corrosion over four different time periods in a lab. Samples' thickness and mass loss were noted down in the dataset. Ultrasonic testing readings were done with an Epoch LT Ultrasonic testing 4

machine with a straight beam configuration, because the relationship between thickness and mass losses and the corrosion process is linear. For multi-class problem, four Corrosion severity levels have been created based on thickness & mass loss occurred during accelerated corrosion testing for which XGB, SVM and Random Forests showed cross validation accuracy score of 99.8%, 93.5%, and 91.5% respectively.

Data Understanding

Data acquisition was the first objective of our project where we simulated pipeline corrosive conditions to develop a framework where we could access the risk analysis by measuring the features that contribute to corrosion in oil and gas pipelines. Now that the first step has been achieved, it was time to go to the next phase of our project where programming and machine learning expertise were utilized extensively. The machine learning phase included data preparation, selection of machine learning methods, feature engineering and model deployment after final hyperparameter optimizations.

The results of the data generated via lab experiments are recorded. Dataset contains 11 columns and 48 data points from 12 samples. Every mild steel sample gave us 4 datapoints. The explanation of each column is as follows:

- Data Point: The unique unit of information extracted from lab experiments.
- Target: The labeling of output.
- Before[g]: The weight of sample recorded before corrosion.
- After[g]: The weight of sample recorded after corrosion.
- Weight Loss[g]: The sample wight loss in the recorded time period. It is obtained by subtracting after[g] from before[g].
- %Weight Loss: The percentage of weight loss. It is the obtained by dividing Weight Loss[g] and Before[g] then multiplying with 100.
- Before[mm]: The weight of sample recorded before corrosion
- After[mm]: The weight of sample recorded after corrosion.
- %Thickness Loss: The percentage of weight loss. It is the obtained by dividing Weight Loss[g] and Before[g] then multiplying with 100.
- Time: It denoted the corrosion time of the sample in the laboratory.

Data Preprocessing

The data cleaning process detects and removes the errors and inconsistencies present in the data and improves its quality. We solve this by eliminating missing values, noise reduction and outliers removal. Replace missing values by the mean or median or simply putting 0 in missing datapoints.

Outliers can be reduced by the use of binning methodology to smoothen the affects of outliers. Pandas is effectively used in cleaning the dataset and in the process of reckoning the outliers.

The major issue in our dataset is the lack of data as more data would be helpful. Since this data is generated in the lab with only handful of experiments thus the possibilities of data problems are negligible to non existant.