Data Glacier

Data Science Intern at Data Glacier

Project: Corrosion Detection & Severity Level Prediction using ML & Ensemble Modelling

Week 9: Deliverables

Name: Syed Sanaullah Shah+ Herbert+Jasmine

University: NUST

Github:

Country: Pakistan

Specialization: Data Science

Batch Code: LISUM11

Date: 30 September 2022

Submitted to: Data Glacier

Table of Contents:

1. Project Plan

Weeks	Date	plan
Weeks 07	August, 2022	Problem Statement, Data Collection, Data Report
Weeks 08	August, 2022	Data Preprocessing (Text Cleaning)
Weeks 09	September, 2022	Data Preprocessing (Preprocessing Operation + Feature Extraction)
Weeks 10	September, 2022	Building the Model
Weeks 11	September, 2022	Model Result Evaluation
Weeks 12	September, 2022	Streamlit Deployment
Weeks 13	September, 2022	Submission(Presentation)

2. Problem Statement

- Globally, the cost of corrosion is in the billions of dollars for every single economy on the planet. Corrosion failures have caused more than $2 trillion dollars in losses around the world. Machine learning is currently one of the most talked-about issues since it allows machines to learn from data and make predictions without having to be explicitly programmed for that job, and it can be done automatically without the assistance of a human. Our main objectives are as follows:
- Corrosion Severity Level detection and prediction from data collected via personal Lab Experiments.

- To check which Machine Learning model Performs well. Ensemble or Traditional Models

3. Data Collection

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 Point #	Target	Before [g]	After [g]	Weight Loss[g]	%Weight Loss	Thickness Before [mm]	Thickness After [mm]	Thickness Loss[mm]	%Thickness Loss	Time
3	A	66.7	64.56	2.14	3.208	6.12	5.97	0.15	2.450	1
4	A	67.68	64.83	2.85	4.210	6.08	5.94	0.14	2.302	1
14	B	74.37	69.65	4.72	6.346	6.25	5.89	0.36	5.760	2
15	B	66.7	61.94	4.76	7.136	6.12	5.82	0.3	4.901	2
27	C	66.7	59.79	6.91	10.359	6.12	5.74	0.38	6.209	3
28	C	67.68	60.34	7.34	10.845	6.08	5.61	0.47	7.730	3
47	D	66.39	58.35	8.04	12.110	6.02	5.49	0.53	8.803	4
48	D	65.04	57.93	7.11	10.931	6.01	5.48	0.53	8.818	4

4. Data Preprocessing

The dataset contains 48 instances, which is little by Machine Learning standards, but it's sufficient for getting started. All attributes are numerical with no repeating instances that save us the time consumed by categorical features encoding. The first step was to remove any duplicates and look for null-values. However, these were not observed in any of the features. Moving forward, we stat the results of exploratory data analysis. The initial investigation to understand the data and detect anomalies is to check for outlier detection using box plot graphical representation. Outliers, punctuations e.t.c were removed from columns and only cleaned data was used.

4.1 Feature Extraction

After looking at relationships between various features, it’s time to select only relevant features that had the most impact on the model performance. More precisely how accurate they will classify the classes in testing phase Tree derived feature importance is a very straightforward, fast and accurate method of selecting suitable features for machine learning. Thus. embedded

methods combine the qualities of filter and wrapper methods is used in feature engineering step. We will utilize random forests algorithm to extract key features.

4.2 Model Selection

The accuracy of your classifications depends on various factors such as effectiveness of the algorithm of choice, selection of parameters, and the amount of useful data at hand. After the completion of data pre-processing, the dataset will be ready for the use. It all starts with the data split. The train-test split step is used to estimate the performance of machine learning algorithms when they are used to classify data in classification tasks on the data kept separated from the training data of the model.

We will use the common split where 80% data is kept for training and the remaining 20% for testing. Train dataset is used to fit the machine learning model whereas test dataset to evaluate the fit of the machine learning model. It's good practice to use a (random)seed to ensure the reproducibility of the results.

4.3 Hyperparameters Tuning

In this part, we split the data into Train. And we split 80% for training and 20% for test. Data splitting is when data is divided into two or more subsets. Typically, with a two-part split, one part is used to evaluate or test the data and the other to train the model. Data splitting is an important aspect of data science, particularly for creating models based on data.

4.4 Evaluation Metrics

The performance of the proposed machine learning approaches using the ML and the ensemble classifiers in predicting corrosion levels will be evaluated by computing commonly used evaluation measures for binary classification referred to as classification accuracy (A), precision (P), recall (R) and F1-score, squared r and root mean squared error.
