Electrical Transformer Diagnostics by Using Machine learning Techniques.

Veera Avinash Chowdary Pusuluri
MSc in Data Analytics
National College of Ireland
x22162402@student.ncirl.ie

Abstract: The device that transfers the Electrical energy from one circuit to another circuit by magnetic coupling without a requirement of any relative motion in between its parts are called Electrical Transformers. At present in entire world uses sensors to check dissolved gases and moisture content in oil inside the electrical transformer to know these emerging faults in the transformer this work will surely help to forecast its trend. The goal of the entire project work is to create a highly diagnostic tool that can give you the Electrical Transformers current condition and the lifespan of any electrical transformer. To know the lifespan and current health condition of Electrical Transformer we can use Machine Learning techniques. This present report is to establish a core basis by using Machine Learning techniques for the overall insulation health condition for Electrical Transformers. There were two machine learning algorithms used to know the health condition of Electrical Transformer. To reach this certain goal we use precise Machine Learning Algorithms. One is Decision Tree and the other is Random Forest. The reason behind using these algorithms for Electrical transformers health condition is because Decision trees are used to solve classification problems. So, in this scenario here we are trying to know the transformers health and lifespan in today's world. In addition to this Random Forest as mentioned above.

1. Introduction

Electrical Transformer is a static device which means there is no running machinery present inside in it. The most expensive and crucial piece of hardware in an electrical power system is a power transformer Dissolved gas Analysis (DGA) is the most popular and widely used. Method for assessing the condition of transformers. The outcome of DGA may aid the utility in making decisions regarding and enhancing transformer maintenance. The defect component is also thought to be the most important among the other factors when establishing the transformer health index condition. This component has come from the dissolved gases. Failure of oil-filled transformers occurs due to various reasons they are, internal external short circuit faults, lightning and switching over voltages, bushing failures, and loose joints. The tool which helps for determining the faults in electrical transformer is DGA (Dissolved Gas Analysis). In general, DGA method is key gas method which is used to interpret fault gas concentrations or gas ratios depending upon the experience of practical experts rather than quantitative evidence.

Many studies have been carried out to develop machine Learning Algorithms for DGA assessment. Many studies have mentioned that by using machine learning algorithms for identifying faults based on the DGA method. All over no method uses the DPM (Duval Pentagon Method) for diagnostic reference. Therefore, this compares several machine learning techniques as a basis for fault identification. Due to the unavailability of DGA data we are using the machine learning technique for fault identification in Electrical Transformers. Reasons behind machine learning algorithm is due to avoid the problem of an unbalanced DGA dataset including decision tree. In addition to this, to achieve this particular goal we propose an exact Machine Learning based fault identification model by utilizing the random forest algorithm. The proposed random forest models adequate in diagnosing faults.

2.Problem Statement

The statement of this problem is, in today's world there were many Electrical Transformers being failed due to multiple reasons. They are, Firstly, Line surges. Line surges are very common in Electrical Transformers failures. Secondly, line faults. Line faults occurs when during the installation of transformer there will be some issues with wire lining by the installers. So, this might be one of the reasons for Electrical transformers. In addition to this thunderstorm. As we are aware of this due to heavy rains and thunderstorms occurs quite frequent. So, we can consider this as well one of the reasons for Electrical transformers failure.

Chronicle / Brant Ward

3. METHODOLOGY

The main aim of the job is to search models that provides great performance for the datasets. To get to know Electrical Transformer I have multiple classification and ensemble methods. Below are the different datasets I have taken for finding and solving problem in this Electrical Transformers.

Data Set-1

The dataset which I have taken for this is `Dataset_Year_2019.csv`

This dataset has 15,873 rows and 16 columns in it. Here burned transformers 2019 is my dependent variable. Also, we can call it as target variable. And rest all are my independent variable in this dataset.

```
LOCATION                                                    int64
POWER                                                     float64
SELF-PROTECTION                                             int64
Average earth discharge density DDT [Rays/km^2-a  o]      float64
Maximum ground discharge density DDT [Rays/km^2-a  o]     float64
Burning rate  [Failures/year]                             float64
Criticality according to previous study for ceramics level  int64
Removable connectors                                        int64
Type of clients                                            object
Number of users                                             int64
Electric power not supplied EENS [kWh]                    float64
Type of installation                                       object
Air network                                                 int64
Circuit Queue                                               int64
km of network LT:                                          object
Burned transformers 2019                                    int64
dtype: object
```

Data Set-2

The dataset which I have taken for this is `transformer-parameters.csv`

This dataset has 2,472 rows and 16 columns in it. Here Method_of_Earthing is my dependent variable. Also, we can call it as target variable. And rest all are my independent variable in this dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2472 entries, 0 to 2471
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Area                            2472 non-null   object
 1   GSP Group                       2472 non-null   object
 2   S/S or Busbar Name Node 1       2472 non-null   object
 3   S/S or Busbar Name Node 1 Voltage  2472 non-null   int64
 4   S/S or Busbar Name Node 2       2472 non-null   object
 5   S/S or Busbar Name Node 2 Voltage  2472 non-null   float64
 6   Transformer ID                  2373 non-null   float64
 7   Vector Group                    2472 non-null   object
 8   Positive Sequence Resistance R  2472 non-null   float64
 9   Positive Sequence Reactance X   2472 non-null   float64
 10  Zero Sequence Reactance X       2472 non-null   float64
 11  Tap Min                         2472 non-null   float64
 12  Tap Max                         2472 non-null   float64
 13  Tap Step                        931 non-null    float64
 14  Normal Rating                   2472 non-null   float64
 15  Method_of_Earthing              2472 non-null   object
dtypes: float64(9), int64(1), object(6)
memory usage: 309.1+ KB
```

Data Set-3

The dataset which I have taken for this is `two-winding-transformer-2021.csv`

This dataset has 2,581 rows and 17 columns in it. Here Method of Earthing is my dependent variable. Also, we can call it as target variable. And rest all are my independent variable in this dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2581 entries, 0 to 2580
Data columns (total 17 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Area                     2581 non-null   object
 1   GSP Group                2581 non-null   object
 2   Node 1                   2581 non-null   object
 3   Operating Voltage 1      2581 non-null   int64
 4   Node 2                   2581 non-null   object
 5   Operating Voltage 2      2581 non-null   object
 6   Status                   2581 non-null   object
 7   Vector Group             2571 non-null   object
 8   R1                       2581 non-null   float64
 9   X1                       2581 non-null   float64
 10  X0                       2581 non-null   float64
 11  Min Tap %                2581 non-null   float64
 12  Max Tap %                2581 non-null   float64
 13  Nominal Rating           2559 non-null   float64
 14  Emergency Rating         2555 non-null   float64
 15  Reverse Power Capability 2245 non-null   object
 16  Method of Earthing       2581 non-null   object
dtypes: float64(7), int64(1), object(9)
memory usage: 342.9+ KB
```

Description of Dataset

Descriptive statistics are used to summarize or describe the characteristics of datasets. By describing the mean value, variance, and data distribution of a dataset, descriptive statistics can be used to describe it. For each of the supplied variables, it computes the lowest, maximum, first, third, middle, and mean in descriptive statistics. According to descriptive statistics, which are frequently used to characterize datasets, the dataset's mean represents the average of all the values. It is calculated by summing up each value and dividing the result by the total number of values. The standard deviation shows how far the data deviate from the mean. The dataset's minimum value is known as minimum.

Dataset-1

ER	SELF-PROTECTION	Average earth discharge density DDT [Rays/km^2-a o]	Maximum ground discharge density DDT [Rays/km^2-a o]	Burning rate [Failures/year]	Criticality according to previous study for ceramics level	Removable connectors	Number of users	Electric power not supplied EENS [kWh]	Air network	Circuit Queue	Burned_transformers_2019
00	15873.000000	15873.000000	15873.000000	15873.000000	15873.000000	15873.000000	15873.000000	15873.000000	15873.000000	15873.000000	15873.000000
570	0.100044	2.230378	5.667857	0.083050	0.219240	0.486297	24.207396	546.705986	0.997921	0.442638	0.050841
785	0.300068	1.745835	3.542212	0.176791	0.413745	0.499828	31.792209	801.972170	0.045550	0.496714	0.219680
00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	81.000000	0.000000	0.000000	0.000000
00	0.000000	0.886588	3.100000	0.000000	0.000000	0.000000	5.000000	243.000000	1.000000	0.000000	0.000000
00	0.000000	1.351333	5.200000	0.000000	0.000000	0.000000	13.000000	399.600000	1.000000	0.000000	0.000000
00	0.000000	3.155602	6.600000	0.000000	0.000000	1.000000	29.000000	666.900000	1.000000	1.000000	0.000000
00	1.000000	6.818167	17.600000	2.000000	1.000000	1.000000	357.000000	54000.000000	1.000000	1.000000	1.000000

Descriptive Analysis of a Dataset -1

DATASET - 2

	S/S or Busbar Name Node 1 Voltage	S/S or Busbar Name Node 2 Voltage	Transformer ID	Positive Sequence Resistance R	Positive Sequence Reactance X	Zero Sequence Reactance X	Tap Min	Tap Max	Tap Step	Normal Rating
count	2472.000000	2472.000000	2.373000e+03	2472.000000	2472.00000	2472.000000	2472.000000	2472.000000	931.000000	2472.000000
mean	57.389563	14.311570	2.984813e+06	5.592315	90.39535	273.617910	-16.239239	5.733697	1.367991	22.702861
std	41.216693	8.779491	2.209604e+06	4.406451	35.24501	4015.936941	2.878400	2.977039	0.196765	21.314160
min	33.000000	6.600000	1.268900e+04	0.000000	10.50000	-103.768000	-24.000000	0.000000	1.000000	0.000000
25%	33.000000	11.000000	1.098619e+06	2.873500	75.85450	21.440250	-19.000000	5.000000	1.200000	10.000000
50%	33.000000	11.000000	3.092091e+06	5.006000	97.91450	84.420000	-15.000000	5.200000	1.400000	18.000000
75%	66.000000	11.000000	4.234469e+06	7.130000	111.11900	105.781250	-13.900000	5.700000	1.400000	24.000000
max	132.000000	66.000000	1.426570e+07	107.900000	283.00000	99900.000000	0.000000	20.000000	2.700000	120.000000

Descriptive Analysis of a Dataset - 2

DATASET – 3

	Operating Voltage 1	R1	X1	X0	Min Tap %	Max Tap %	Nominal Rating	Emergency Rating
count	2581.000000	2581.000000	2581.000000	2581.000000	2581.000000	2581.000000	2559.000000	2555.000000
mean	56.589694	5.407874	89.277832	270.191583	-15.785819	5.602983	25.741145	31.381969
std	40.781097	3.752275	36.034226	3930.871809	3.524142	3.418595	22.764294	27.233453
min	33.000000	0.000000	0.196000	-103.768000	-24.000000	-20.000000	1.000000	3.000000
25%	33.000000	2.823000	75.591000	21.219000	-19.000000	5.000000	11.500000	14.000000
50%	33.000000	4.995000	97.405000	80.294000	-14.000000	5.200000	18.000000	24.000000
75%	66.000000	6.779000	111.120000	104.178000	-13.900000	5.300000	30.000000	35.000000
max	132.000000	32.789000	283.000000	99900.000000	10.000000	20.000000	120.000000	156.000000

Descriptive Analysis of a Dataset - 3

In Dataset maximum is the most crucial value. Here we can discuss on Q1 & Q3. Q1 is quartile – 1 When bottom 25% is separated from the rest of the 75% in the dataset is known as first quartile also known as Q1.

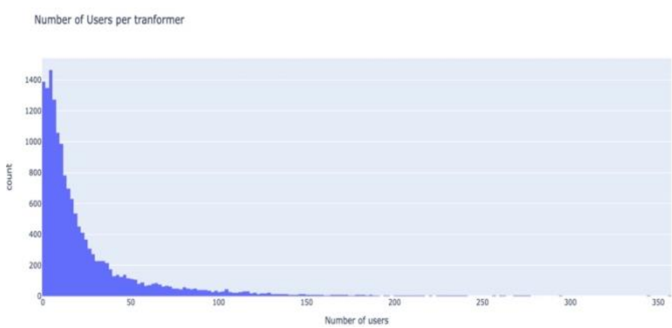
Data Preparation: Preparing and cleaning of data is required before running the analysis to check what factors impacting.

Data Normalization: This is a technique which is applied many times as a part of data preparation for Machine Learning. The major target of the normalization is to change the values of numerical columns in the dataset to use a common scale without losing information.

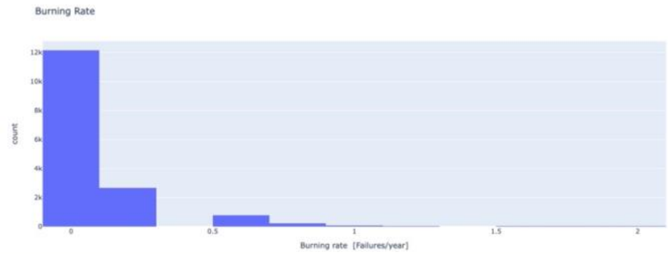
Exploratory Data Analysis (EDA)

Dataset - 1

In this exploratory data analysis first I have found the value counts of one element in dataset which is number of users in dataset – 1.



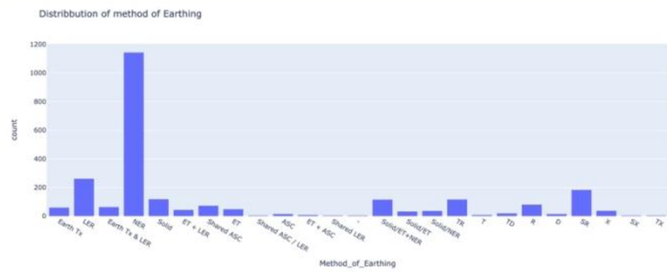
In addition to this I found the value counts of another element in dataset – 1 which is Burning rate [Failures/year]



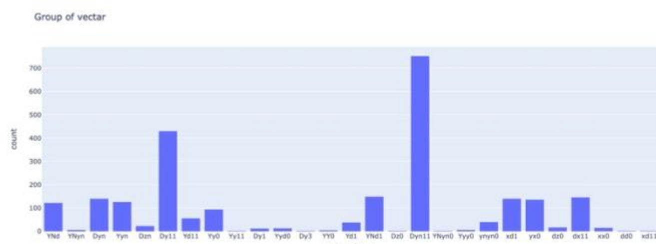
Just like above form Dataset – 1. I have done for dataset 2 and dataset 3 as shown below.

Dataset - 2

Here I have found the value counts for the element named method of earthing in dataset – 2

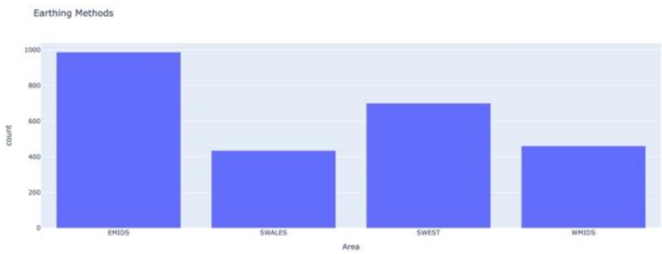


In addition to this below is the value counts for another element vector group in dataset – 2

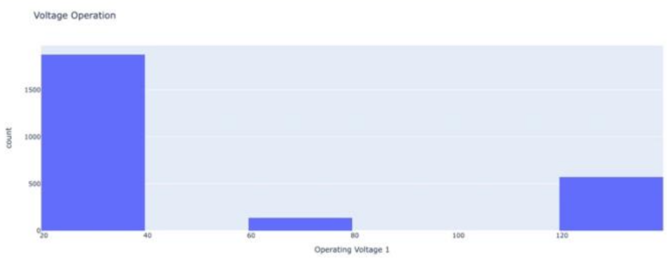


Dataset - 3

Value counts of the element which I found in dataset – 3 is Area.

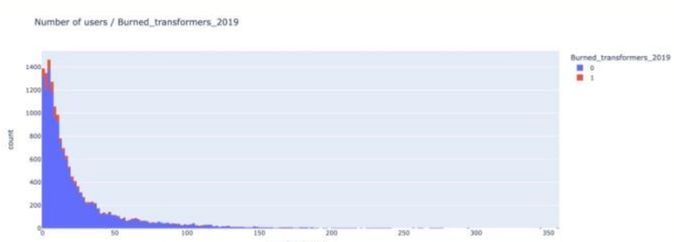


Below is the value counts plot for element named Operating voltage 1. Plotting as follows.

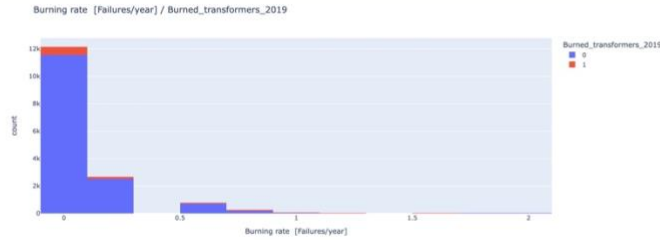


Comparison V/S in three datasets are below.

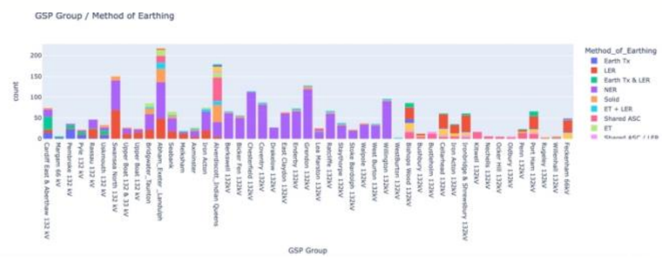
Histogram plot for Number of users v/s Burned transformers in 2019 in Dataset - 1



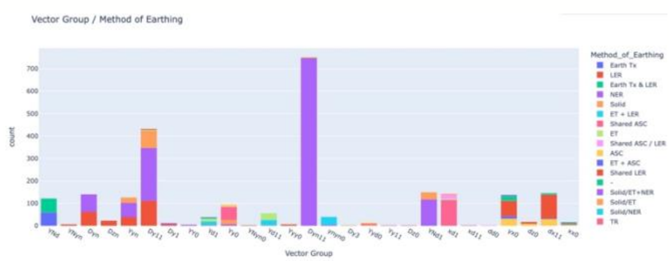
Histogram plot for Burning rate v/s Burned transformers in 2019 in Dataset - 1



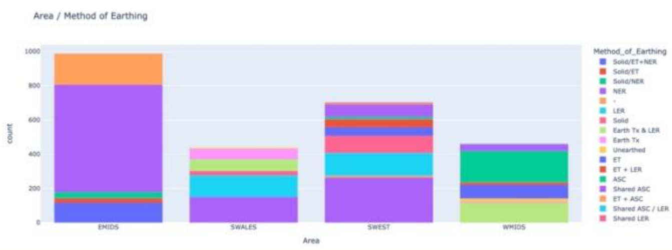
Plotting for GSP Group v/s method of earthing in Dataset – 2



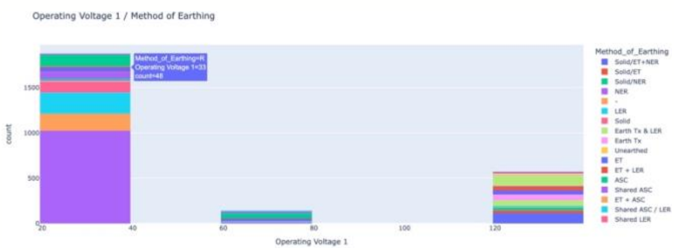
Plotting for vector group v/s method of earthing in dataset - 2



Plotting for Area v/s method of earthing in dataset – 3



Plotting for Operating voltage 1 v/s method of earthing in dataset – 3



4. Model Building

This is a process of developing a statistical representation of a process in today's world. The major target of this development is to predict how the system acts under multiple circumstances that affect the behaviour.

First-Model (Decision tree)

I utilized the sklearn and XGBoost libraries, which contain the decision tree method. Here we can pass a certain limit of arguments for each libraries we have taken. In addition to this after completion of building a model Accuracy and F1 score are the parameters I have taken because Accuracy gives number of data instances classifies correctly. In addition to this where F1 score gives the evaluation with its class wise performance.
Below are the results for Decision Tree Algorithm which I have got.

Decision Tree	SKlearn	XGBoost
Dataset-1 Accuracy	0.94	0.94
Dataset-1 F1 score	0.92	0.92
Dataset-2 Accuracy	0.70	0.84
Dataset-2 F1 score	0.68	0.83
Dataset-3 Accuracy	0.62	0.79
Dataset-3 F1 score	0.53	0.77

Second-Model (Random Forest)

I utilized the sklearn and XGBoost libraries, which contain the Random Forest method. First on discussing with Sklearn library in random forest algorithm we can pass 18 arguments. Secondly in XGBoost library in random forest algorithm we can pass 40 arguments. In addition to this where F1 score gives the evaluation with its class wise performance. Below are the results for Random Forest Algorithm which I have got.

Random Forest	SKlearn	XGBoost
Dataset-1 Accuracy	0.94	0.94
Dataset-1 F1 score	0.92	0.92
Dataset-2 Accuracy	0.76	0.78
Dataset-2 F1 score	0.69	0.74
Dataset-3 Accuracy	0.66	0.68

Dataset-3 F1 score	0.56	0.61

5.Conclusion

A transformer is an electrical energy transfer device that steps up or steps down the voltage from one alternating circuit to one or more other circuits. During this step-up and step-down process there will be power fluctuations at some other point. So, that failures happen in these Electrical transformers. At the end of feature selection process Decision tree and Random forest algorithms were used to train the model by using SKlearn and XGBoost libraries in it.

6.References

1. Rahman Azis Prasojo, “Precise transformer fault diagnosis via random Forest model enhanced by synthetic minority over-sampling technique”
2. Matti Lehnon, “ Precise transformer fault diagnosis via random Forest model enhanced by synthetic minority over-sampling technique ”
3. Vishnu Priya KJ, “Online Prediction of DGA results for Intelligent Condition Monitoring of Power Transformers”