FSM Online Internship Phase I Report
on

Remaining Usable Life Estimation (NASA Turbine Dataset)
In

Machine Learning

Submitted by

Sandesh Pabitwar
Modern Education Societies College of Engineering Pune

Under Mentorship of
Mr. Devesh Tarasia

IITD-AIA Foundation for Smart Manufacturing

Table of Content

1. Introduction

For the ML - 2 project data of turbofan engine is given. Each engine starts with different degrees of initial wear and manufacturing variation which is unknown to the user. This wear and variationare considered normal i.e., it is not considered a fault condition.
In the dataset the data of engine no, operational cycles, 3 operational settings, and 20 sensors measurement are given. The main objective of the Project is to predict the number of remaining operational cycles before failure in the test set, i.e., the number of operational cycles after the lastcycle that the engine will continue to operate.

1.1 Exploratory data analysis.

To understand and analyze the data Exploratory Data Analysis (EDA) is required. Main purpose of the EDA is to get deep insight into a data set and provide the specific outcomes that are useful for the training a Model.

1.2 What is Turbofan Engine?

Turbofan Engine is the most modern variation of basic gas turbine engine. In the turbofan engine, the core engine is surrounded by a fan in the front and an additional turbine at the rear. The fan and fan turbine are composed of many blades, like the core compressor and core turbine, and are connected to an additional shaft

2. Problem Definition

Predict the number of remaining operational cycles before failure in the test set, i.e., the number of operational cycles after the last cycle that the engine will continue to operate.
In simple words aim of this project is to build a machine learning model which can predict the Remaining useful life of engine.

3. Objectives

- To understand the given dataset.
- To find the pattern in given features.
- To detect outliers or anomalous events.
- To summaries dataset
- To get list of important features

4. Information About Dataset

Given dataset contains simulated data produced by a model-based simulation program, i.e., Commercial Modular Aero-Propulsion System Simulation (C-MAPSS), which was developed by NASA. The C-MAPSS dataset includes 4 sub-datasets that are composed of multi-viriate temporal data obtained from 21 sensors.

Each sub-dataset contains one training set and one test set. The Training datasets include run-to-failure sensor records of multiple aero-engines collected under different operational conditions and fault modes. Each engine unit starts with different degrees of initial wear and manufacturing variation that is unknown and considered to be healthy. As time progresses, the engine units begin to degrade until they reach the system failures, i.e., the last data entry corresponds to the time cycle that the engine unit is declared unhealthy.

On the other hand, the sensor records in the testing datasets terminate at some time before system failure, and the goal of this task is to estimate the remaining useful life of each engine in the test dataset. For verification, the actual RUL values for the testing engine units are also provided.

Dataset	Operating condition	Fault mode	Train trajectories	Test trajectories
FD001	1	1	100	100
FD002	6	1	260	259
FD003	1	2	100	100
FD004	6	2	248	249

5. Summary statistics

Training set 1 consists of *20631 rows × 26 columns*.

Table no.1: Statistical data for Operational Settings.

	OPsetting_1	OPsetting_2	OPsetting_3
count	20631.000000	20631.000000	20631.0
mean	-0.000009	0.000002	100.0
std	0.002187	0.000293	0.0
min	-0.008700	-0.000600	100.0
25%	-0.001500	-0.000200	100.0
50%	0.000000	0.000000	100.0
75%	0.001500	0.000300	100.0
max	0.008700	0.000600	100.0

Table no. 1 represents the statistical details of operational settings 1, 2 and 3.

Standard deviation of operational setting 3 is Zero means there is no deviation in values of operational setting 3 throughout the dataset. And the standard deviation of operational setting 1 and 2 are so less it shows that there is slight variation in values of operational setting 1 and 2.

Table 2: Statistical data for Sensors.

	count	mean	std	min	25%	50%	75%	max
sensor 1	20631.0	518.670000	0.000000e+00	518.6700	518.6700	518.6700	518.6700	518.6700
sensor 2	20631.0	642.680934	5.000533e-01	641.2100	642.3250	642.6400	643.0000	644.5300
sensor 3	20631.0	1590.523119	6.131150e+00	1571.0400	1586.2600	1590.1000	1594.3800	1616.9100
sensor 4	20631.0	1408.933782	9.000605e+00	1382.2500	1402.3600	1408.0400	1414.5550	1441.4900
sensor 5	20631.0	14.620000	1.776400e-15	14.6200	14.6200	14.6200	14.6200	14.6200
sensor 6	20631.0	21.609803	1.388985e-03	21.6000	21.6100	21.6100	21.6100	21.6100
sensor 7	20631.0	553.367711	8.850923e-01	549.8500	552.8100	553.4400	554.0100	556.0600
sensor 8	20631.0	2388.096652	7.098548e-02	2387.9000	2388.0500	2388.0900	2388.1400	2388.5600
sensor 9	20631.0	9065.242941	2.208288e+01	9021.7300	9053.1000	9060.6600	9069.4200	9244.5900
sensor 10	20631.0	1.300000	0.000000e+00	1.3000	1.3000	1.3000	1.3000	1.3000
sensor 11	20631.0	47.541168	2.670874e-01	46.8500	47.3500	47.5100	47.7000	48.5300
sensor 12	20631.0	521.413470	7.375534e-01	518.6900	520.9600	521.4800	521.9500	523.3800
sensor 13	20631.0	2388.096152	7.191892e-02	2387.8800	2388.0400	2388.0900	2388.1400	2388.5600
sensor 14	20631.0	8143.752722	1.907618e+01	8099.9400	8133.2450	8140.5400	8148.3100	8293.7200
sensor 15	20631.0	8.442146	3.750504e-02	8.3249	8.4149	8.4389	8.4656	8.5848
sensor 16	20631.0	0.030000	1.387812e-17	0.0300	0.0300	0.0300	0.0300	0.0300
sensor 17	20631.0	393.210654	1.548763e+00	388.0000	392.0000	393.0000	394.0000	400.0000
sensor 18	20631.0	2388.000000	0.000000e+00	2388.0000	2388.0000	2388.0000	2388.0000	2388.0000
sensor 19	20631.0	100.000000	0.000000e+00	100.0000	100.0000	100.0000	100.0000	100.0000
sensor 20	20631.0	38.816271	1.807464e-01	38.1400	38.7000	38.8300	38.9500	39.4300
sensor 21	20631.0	23.289705	1.082509e-01	22.8942	23.2218	23.2979	23.3668	23.6184

Table 2 represents the statistical details of sensor. From the table 2 it is clear that the standard deviation of sensor 1, 10, 18 and 19 is zero means there is no change in values of these sensors for all engines.

6. Visualization of data:

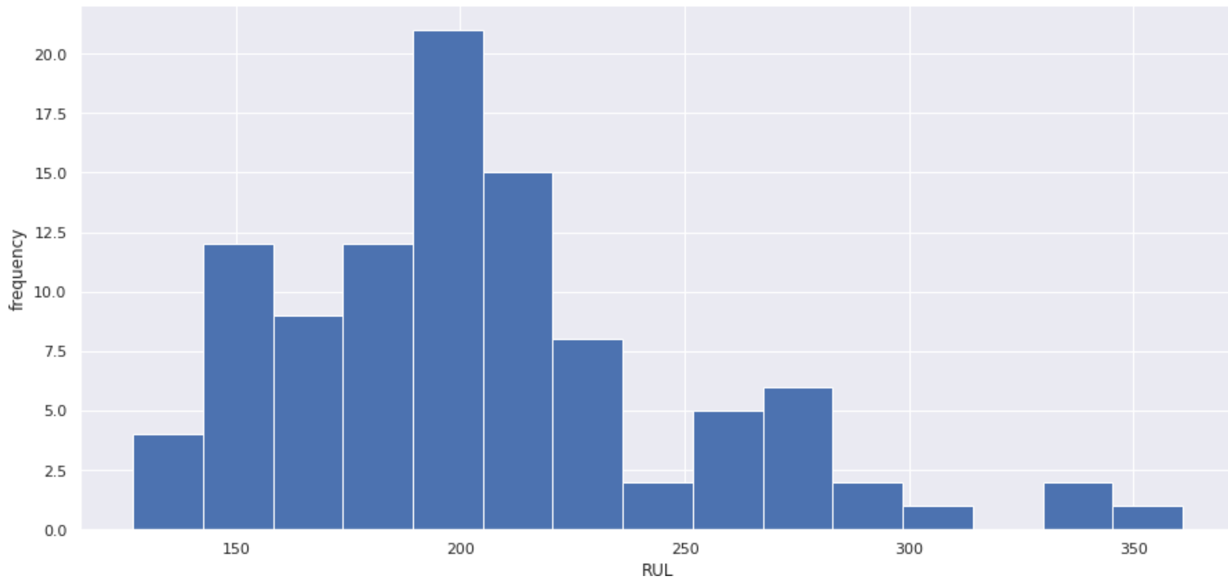

Figure 1 histogram RUL vs Frequency

Figure 1 shows the histogram of RUL vs Frequency. From figure 1 it is clear that most of the engines have RUL close to the 200.

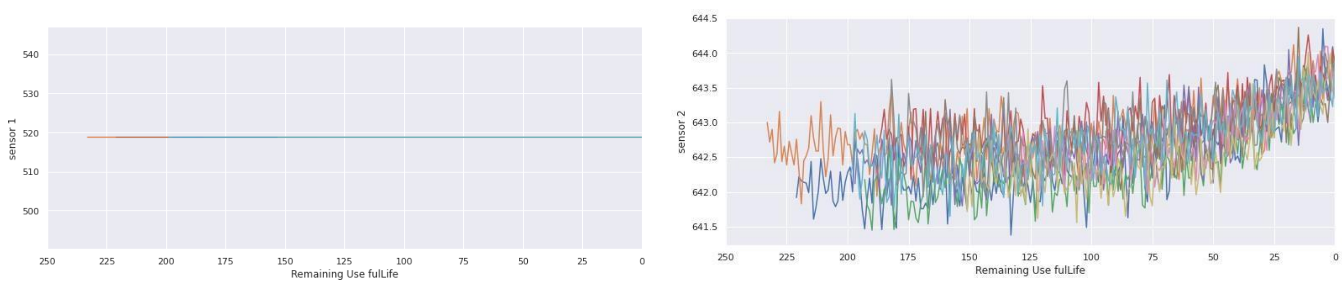

Figure 2 graph of sensor 1 and 2 against RUL

Figure 2 represents the graph of *graph of sensor 1 and 2 against RUL*

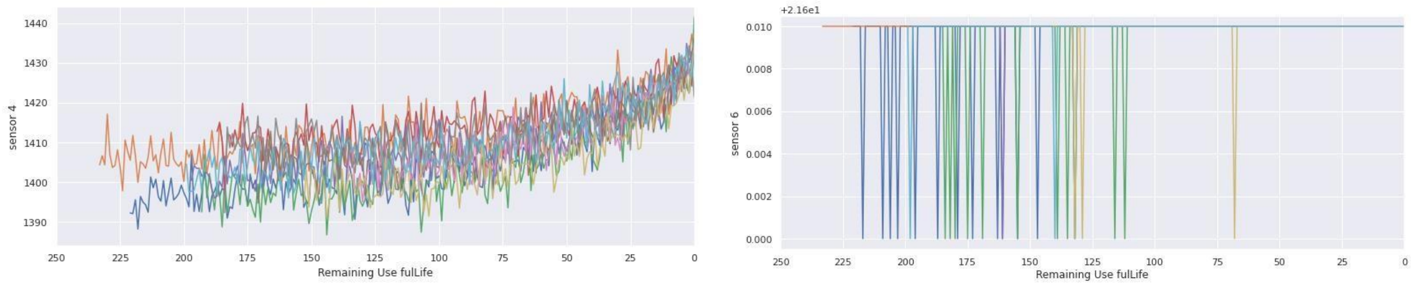

Figure 3 graph of sensor 4 and 6 against RUL

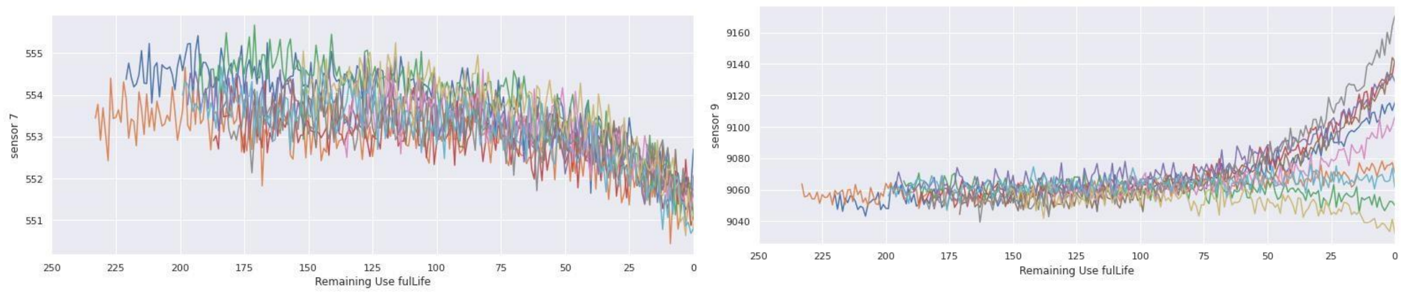

Figure 4 graph of sensor 7 and 9 against RUL

Out of 21 graphs Some of the graphs are shown by figure 2,3 and 4.

Graph of sensors 5,10,16,18,19 showing the same pattern as shown by sensor 1 in figure 1 i.e., flat line. It shows that they are not contributing to the Remaining Useful Life.

Sensor 2 is showing a rising trend, a similar pattern is observed for sensors 3, 4, 8, 11, 13, 15 and 17.

Sensor 7 is showing the deceasing trend and same pattern is observed for sensors 12, 20 and 21.

From plotting the sensors graph one thing is clear that values of sensors 1,5,10,16,18,19 are remaining constant so we can neglect those while training the model.

6.1 Heat Map:

To check the correlation heat map is plotted.

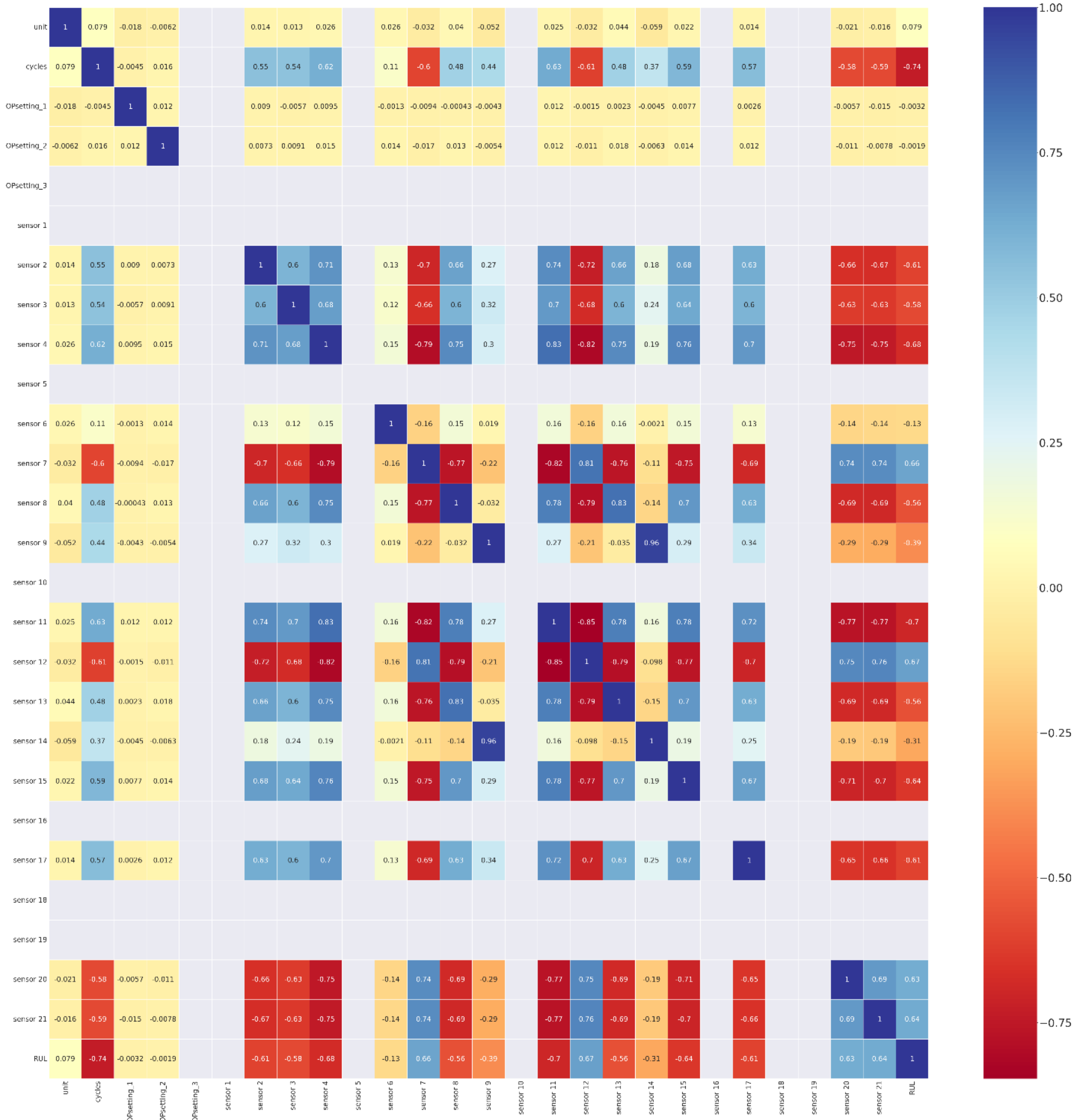

Figure 5 heat map

Figure 5 shows the heat map for all the features.

From figure 5 we can observe that 'OpSetting1', 'OpSetting2', 'OpSetting3', 'Sensor 1', 'Sensor 5', 'Sensor 6', 'Sensor 9', 'Sensor 10', 'Sensor 14', 'Sensor 16', 'Sensor 18', 'Sensor 19' these features are having correlation less than 0.5 with the RUL.

7. Key points

- No missing data in given dataset
- 100 time series in the training set 1, and 100 time series in the test set 1
- the statistics on the number of cycles aren't relevant, because we should only look at the statistics for the **maximum** number of cycles for each engine. However, the fact that the mean and median number of cycles for the training set are larger than for the test set, agrees with the fact that in the training set, the engines are followed until system failure. In the test set, the time series ends some time prior to system failure
- the following Features are constant, both in the training and in the test set, meaning that the operating condition was fixed or the sensor was broken/inactive: operational setting 3, sensor 1, sensor 5, sensor 10, sensor 16, sensor 18, sensor19. We can discard these variables from the analysis.
- Sensor 6 is oscillating between two values and its co-relation with RUL is less so we candiscard this feature.
- Sensors 7, 10, 20, 21 are highly corelated with the RUL. These are most important features.