Exploratory Data Analysis of the N-CMAPSS Dataset for Prognostics

Supratik Chatterjee
Engineering, Research & Development
Wipro Limited
India
supratik.chatterjee@outlook.in

Arvind Keprate
Department of Mechanical, Electronics and Chemical Engineering
Oslo Metropolitan University
Norway
arvind.keprate@oslomet.no

Abstract – **In the recent years, industries such as aeronautical, railway, and petroleum has transitioned from corrective/preventive maintenance to condition based maintenance (CBM). One of the enablers of CBM is Prognostics which primarily deals with prediction of remaining useful life of an engineering asset. Besides physics-based approaches, data driven methods are widely used for prognostics purposes, however the latter technique requires availability of run to failure datasets. In this manuscript authors have aimed at performing exploratory data analysis (EDA) on the New Commercial Modular Aero-Propulsion System Simulation (N-CMAPSS) dataset published by NASA. Although 8 datasets are publicly available, authors have chosen dataset 3 (DS03) for EDA in this paper which consists of 9.8 million instances and 47 features. The main aim of doing EDA is to gain better understanding of the dataset as it would facilitate in building a deep learning model that can be used for predicting RUL of the aircraft engines.**

Keywords - **Prognostics, Exploratory Data Analysis, N-CMAPSS dataset**

I. INTRODUCTION

Optimal inspection planning and maintenance of engineering assets in various industries is quintessential for maximizing safety and minimizing cost. Over the past 7 decades, the maintenance strategies have evolved from corrective maintenance to preventive maintenance and finally to condition based maintenance (CBM) as shown in Fig. 1 [1]. Advances in the field of sensor technology, data acquisitioning, data storage, and machine learning has made CBM more realistic. The greatest enabler of CBM is prognostics which as per ISO13381-1 [2], is defined as '*an estimation of time to failure and risk for one or more existing and future failure modes*'. In real life applications deployment of the prognostics models would provide early failure warnings of engineering systems, thus giving sufficient time to maintenance engineers to intervene the system before actual failure happens. Consequently, this would lead to reduction in machine downtime, enhanced system safety and considerable cost savings for the asset owners.

A number of approaches have been used by the researchers to perform prognostics in engineering domain as discussed by the authors in [3]. However, the two commonly used methods for building a prognostics model are physics-based approach and data driven approach. The former approach relies on employing closed form equations derived from the first principles (or fitted to experimental data) to estimate the Remaining Useful Life (RUL). Fatigue crack growth is one such degradation mechanism where physics-based model (such as Paris Law) is used for predicting the RUL. Authors have employed the aforementioned method to predict RUL of topside piping in the previous works [4,5,6]. However, physics-based models may not be available for all the physical phenomenon, hence under such circumstances researchers resort to data driven methods for performing prognostics.

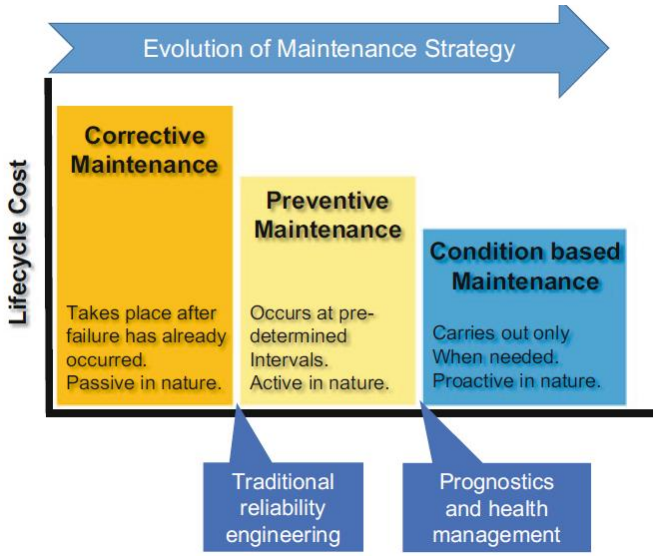

Fig. 1 - Evolving maintenance strategies [1].

Data-driven (DD) approaches utilize information from training dataset to recognize the characteristics of training data and finally make predictions about the future state. However, the success of DD approaches relies on collecting statistics of failures as a function of current state, which requires volumes of data [1]. Nevertheless, it is impractical to collect huge amount of failure data from safety critical equipment's, therefore researchers rely on simulations to generate synthetic data which can then be used to develop prognostic model. One such synthetic dataset of run-to-failure trajectories for a small fleet of aircraft engines under realistic flight conditions has been released by the NASA Ames Prognostics Center of Excellence (PCoE), in collaboration with ETH Zurich and Palo Alto Research Center (PARC). The dataset was generated with the Commercial Modular Aero-Propulsion System Simulation

(C-MAPSS) dynamical model and is referred as N-CMAPSS [7]. Although 8 datasets are publicly available, authors have chosen dataset 3 (DS03) for the exploratory data analysis (EDA) in this paper. EDA is performed in order to gain to the visual insights, within the DS03, by using various visualizations such as correlation matrix heatmap, box plot, KDE plot. The next obvious step is to build prediction models, especially non-linear deep learning models, which can then be used for predicting RUL of the aircraft engines.

The remaining of the manuscript is structured as follows. In Section II, a brief summary of N-CMAPSS dataset is presented, followed by the exploratory data analysis of the dataset in Section III, and a conclusion in Section IV.

II. N-CMAPSS DATASET

The N-CMAPSS dataset provides simulated run-to-failure trajectories of a small fleet of large turbofan engines the schematic of which is shown in Fig. 2. The data was synthetically generated using CMAPSS Engine simulator built in Simulink [8]. The schematic of the simulator is shown in Fig. 3 while the details of data generation methodology is discussed in [10]. On comparing Fig.3 and Fig.2, it can be seen that CMAPSS is a high-fidelity model of the turbofan engine.

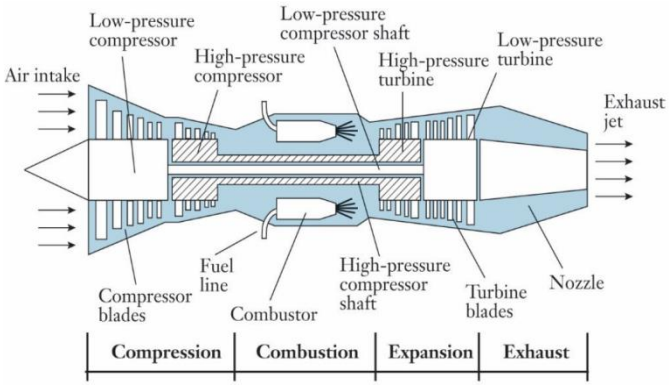

Fig. 2 - Schematic of Turbofan Engine [9].

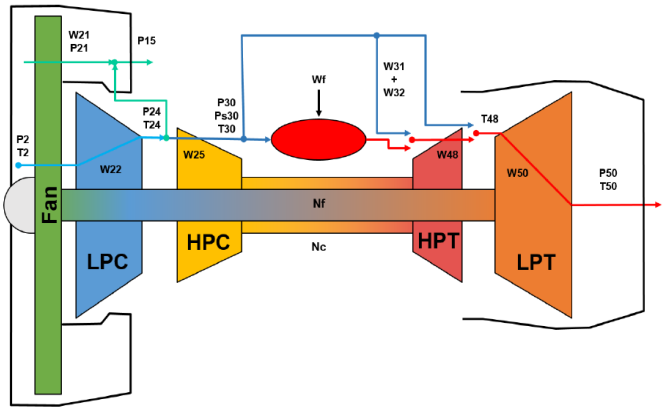

Fig. 3 - Schematic of CMAPSS Model [10].

In total 8 datasets are simulated using CMAPSS simulator, the overview of dataset is given in Table I. In this paper authors have considered DS03 for EDA. As depicted in Table I, DS03 has 9.8 million rows, 15 Units (i.e. data for 15 different turbofan engines), 3 flight classes and 1 failure modes. The data is in the form of .h5 files and the variables in the dataset are divided into 6 categories shown in TABLE II. For this paper authors have not considered X_v_* and T_* as these are not important for prognostics model building as discussed in [10]. The W_* consist of 4 parameters related to flight data (i.e., altitude, flight Mach number, throttle-resolve angle, and total temperature at fan inlet), while, X_s_* consists of 14 parameters as shown in TABLE III. The auxiliary variable (A_*) consists of 4 parameters namely unit number (categorical variable), flight class (categorical variable), health state and flight cycle number. Finally, Y_* consists of RUL cycles which is response in our prognostics problem, while remaining variables are the input variables. Furthermore, unobservable model health parameters (θ) as described in [10], have not been considered as part of this exploration, as prognostic problems can be solved without these parameters.

TABLE I
OVERVIEW OF DATASET

Name	# Units	Flight Classes	Failure Modes	Size
DS01	10	1, 2, 3	1	7.6 M
DS02	9	1, 2, 3	2	6.5 M
DS03	15	1, 2, 3	1	9.8 M
DS04	10	2, 3	1	10.0 M
DS05	10	1, 2, 3	1	6.9 M
DS06	10	1, 2, 3	1	6.8 M
DS07	10	1, 2, 3	1	7.2 M
DS08	54	1, 2, 3	1	35.6 M

TABLE II
DATASET VARIABLES CATEGORIES

Name [*= {"dev", "test"}]	Description
W_*	Scenario descriptors—w
X_s_*	Measurements—x_s
X_v_*	Virtual sensor—x_v
T_*	Health Parameters—θ
Y_*	RUL [in cycles]
A_*	Auxiliary data

TABLE III
PARAMETERS IN X_s

Symbol	Description	Units
Wf	Fuel flow	pps
Nf	Physical fan speed	rpm
Nc	Physical core speed	rpm
T24	Total temperature at LPC outlet	°R
T30	Total temperature at HPC outlet	°R
T48	Total temperature at HPT outlet	°R
T50	Total temperature at LPT outlet	°R
P15	Total pressure in bypass-duct	psia
P2	Total pressure at fan inlet	psia
P21	Total pressure at fan outlet	psia
P24	Total pressure at LPC outlet	psia
Ps30	Static pressure at HPC outlet	psia
P40	Total pressure at burner outlet	psia
P50	Total pressure at LPT outlet	psia

III. ILLUSTRATIVE CASE STUDY

A. Data Generation.

The data used in the case study is extracted and oriented using a custom build functions based on N-CMAPSS example data loading and exploration notebook. The details of which can

be found in [10]. The dataset itself is divided into two parts by default, namely development data ($\mathcal{D}$) and test data ($\mathcal{D}_{T^*}$), or train and test data respectively. Authors have decided to explore both combined and given split dataset to gain maximum insight. The two sets of data, one by keeping "dev" and "test" sets separate (D_{dev}, D_{test}) and another one by combining "dev" and "test" (D_{com}) were explored.

All sets of data contain extracted values from individual feature datasets: `W_*`, `X_s_*` and `A_*`, along with target dataset `Y_*`, where * = {"dev", "test"} accordingly. For ease of identification, before combining these datasets, each column of feature datasets, excluding target `Y_*` have been renamed by post-fixing words with the existing names according to it's intended original names as described in Table II. `Y_*` have been renamed to "RUL" (Remaining Useful Life) for all sets of data. A snippet of the concatenated data from one of the tables (D_{dev}) can be observed in Table. IV. The preliminary purpose of including `A_*` is to group data according to the unit number and flight classes when seems necessary.

TABLE IV
SNIPPET OF D_{dev} DATASET

alt_W	Mach_W	TRA_W	...	Ps30_X_s	unit_A	cycle_A	Fc_A	RUL
3010.0	0.342090	77.782646	...	417.04991	10.0	1.0	3.0	65
3020.0	0.342846	77.782646	...	416.97787	10.0	1.0	3.0	65
3030.0	0.343161	77.782646	...	416.87702	10.0	1.0	3.0	65
3033.0	0.343413	77.782646	...	416.83654	10.0	1.0	3.0	65
3042.0	0.344232	77.782646	...	416.83598	10.0	1.0	3.0	65

B. Exploratory Data Analysis.

As per CRISP-DM (Cross Industry Standard Process Data Mining) methodology, data understanding is one of the vital steps that must be performed before building a machine/deep learning model. EDA, utilizes various techniques in order to gain insights in the dataset and to find hidden patterns. The current datasets consist of 21 independent variables (excluding health status) and one dependent variable (RUL). Dimensions (row, column) of $D_{com,}$, D_{dev}, D_{test} are (9822837, 22), (5571277, 22) and (4251560, 22) respectively. No cell has missing or corrupt information across all datasets, $D_* = \{D_{com,}, D_{dev}, D_{test}\}$. There are three categorical variables (excluding health status) coming from `A_*`, named flight class; Fc_A = {1, 2, 3}, health status; hs_A = {1, 0} and unit_A = {x: x∈ $\mathbb{N}_1$ and 1≤x≤15}. Remaining independent variables are numerical in nature.

Due to enormous size of the dataset, it was necessary to subsample data when plotting individual points else the plots would be incomprehensible (for e.g., scatter plotting individual points of each pair of rows will make the plot not understandable and may take a prolonged time). In contrast, plotting distribution of features with whole dataset is possible with reasonable time and resource usage. Aim of this activity is to get a general overview understanding of the data, so that necessary and appropriate steps could be understood to solve prognostic problem(s) in future. The authors created two combined sets of figures, one for a general overview of the data and another one for detailed information.

Fig. 4 (in Appendix) is a combined representation of dependent & independent variable's scatter plots grouped by three class values and association matrix using seaborn [16] and sweetviz [15]. Scatter plots are being represented in the bottom triangle of the rectangle, separated from the upper triangle by diagonal approximate kernel density estimation. The upper triangle represents Pearson correlation among two numerical variables and correlation ratio among categorical variables & numerical variables [15]. In the case of scatter plots of Fig. 4, a random sub-sample of fraction 0.0001 or 982 instances of each variable from D_{com}, have been used. The reason for using sub-sampled data is explained earlier. In the case of correlation upper triangle, circles represent Pearson correlation coefficient and squares represent correlation ratio. The size and colour of the shapes inside the upper triangle are directly proportional to the value shown in the right-side legend of the figure. This reveals there are many features that are highly correlated with each other and not with the actual target. Table V shows association among features and target. The categorical association of unit_A and Fc_A with RUL are 0.21 and 0.14 respectively which are much higher than correlation values of scenario descriptors (alt_W and Mach_W). Furthermore, the uncertainty coefficient and correlation ratio of the two categorical variables have been shown in Table VI.

TABLE V
TARGET(RUL) ASSOCIATION VALUES

NUMERICAL ASSOCIATIONS (PEARSON, -1 to 1)		CATEGORICAL ASSOCIATIONS (CORRELATION RATIO, 0 to 1)	
cycle_A	-0.91	unit_A	0.21
Mach_W	-0.07	Fc_A	0.14
alt_W	-0.07		
P2_X_s	0.07		
T2_W	0.07		
T48_X_s	-0.07		
P50_X_s	0.07		
T50_X_s	-0.07		
P15_X_s	0.07		
P21_X_s	0.07		
P24_X_s	0.06		
TRA_W	-0.05		
P40_X_s	0.04		
Ps30_X_s	0.04		

TABLE VI
CATEGORICAL VARIABLE ASSOCIATION VALUES

unit_A	CATEGORICAL ASSOCIATIONS (UNCERTAINTY COEFFICIENT, 0 to 1)	Fc_A	CATEGORICAL ASSOCIATIONS (UNCERTAINTY COEFFICIENT, 0 to 1)
Fc_A	0.39	unit_A	0.39
	NUMERICAL ASSOCIATIONS (CORRELATION RATIO, 0 to 1)		NUMERICAL ASSOCIATIONS (CORRELATION RATIO, 0 to 1)
T2_W	0.48	T2_W	0.48
P2_X_s	0.47	P2_X_s	0.47
alt_W	0.47	alt_W	0.46
P21_X_s	0.45	P21_X_s	0.45
P15_X_s	0.45	P15_X_s	0.45
P50_X_s	0.45	P50_X_s	0.45
P24_X_s	0.43	P24_X_s	0.42
Mach_W	0.39	Mach_W	0.39
T24_X_s	0.33	T24_X_s	0.33
P40_X_s	0.30	P40_X_s	0.30
Ps30_X_s	0.30	Ps30_X_s	0.30
TRA_W	0.26	TRA_W	0.26
T50_X_s	0.24	T50_X_s	0.23
Wf_X_s	0.23	Wf_X_s	0.23

The pie chart of Fig. 5. shows that nearly half of the class type of D_{com} is of flight class 3 and classes count are not equally distributed (flight_class_1 = 1824707, flight_class_2 = 3094122, flight_class_3 = 4904008). Assigning a proportionate weight of target based on these count value ratios might help in creating a more accurate prognostic model in future. Furthermore, the bar chart of Fig.5 shows a good proportion of instances are present for each unit type. However, in Fig. 6 the bar plot unveils instances of each three flight class types are not present in every unit. Which makes unit-wise prognostic modelling of each class impossible.

Fig. 7 (in Appendix) shows top numerical features which are mostly not correlated with each other but having the most correlation score with the target (RUL in this case). Any feature having an absolute correlation value of more than or equal to 2nd quartile value of absolute correlations with the target (~0.02520) has been considered. Similarly, any feature having an absolute correlation value of less than or equal to 2nd quartile value of absolute correlations with each other (~0.65912) has

been considered. A consolidated description of D_{com} numerical variables is shown in TABLE VI.

TABLE VII
CONSOLIDATED TABLE DESCRIPTION OF NUMERICAL VARIABLES

	mean	std	min	25%	50%	75%	max
alt_W	15639.57	8083.52	3001.00	9086.00	14408.00	22622.00	35033.00
Mach_W	0.54	0.12	0.00	0.44	0.54	0.64	0.75
TRA_W	60.46	18.37	23.55	46.58	64.51	77.17	87.63
T2_W	490.20	19.88	421.38	473.77	494.29	506.58	534.38
T24_X_s	569.50	21.14	484.20	554.49	567.37	583.46	634.35
T30_X_s	1330.58	68.35	1068.82	1284.40	1326.38	1370.25	1534.37
T48_X_s	1641.08	124.04	944.50	1556.79	1650.40	1716.62	2006.07
T50_X_s	1130.93	62.73	690.19	1087.31	1121.36	1166.53	1372.75
P15_X_s	12.94	2.88	5.92	10.43	13.16	15.21	20.45
P2_X_s	10.10	2.42	4.37	7.91	10.38	12.02	15.68
P21_X_s	13.14	2.93	6.01	10.59	13.36	15.44	20.76
P24_X_s	15.95	3.45	6.91	13.10	15.99	18.38	26.45
Ps30_X_s	236.76	58.98	80.33	192.75	224.59	271.21	457.37
P40_X_s	240.96	59.80	82.09	196.33	228.85	275.98	463.83
P50_X_s	10.10	2.76	4.13	7.61	10.22	12.20	16.88
Nf_X_s	1957.05	187.06	1469.74	1838.71	2003.43	2114.77	2290.65
Nc_X_s	8237.51	226.85	7366.11	8084.66	8229.87	8372.57	8885.41
Wf_X_s	2.54	0.79	0.33	1.98	2.34	2.94	5.82
cycle_A	36.27	21.46	1.00	18.00	36.00	53.00	93.00
RUL	35.32	21.43	0.00	17.00	35.00	53.00	92.00

The Statistical descriptions of Table VII shows that the features are in different ranges. A scaling methodology such as min-max scaling could be used to bring variables in the same scale, which in turn could be useful to achieve a more accurate prognostics model. Furthermore, the unit-wise Box plot grouped by flight class in Fig. 8, depicts that features are highly spread and consist of a fairly large number of outliers. Therefore, applying a robust data scaling and transformation scheme could be helpful in creating an accurate prediction model as we did in our previous work [11].

Instead of plotting consolidated features of D_{com}, authors focused on a comparative analysis between selected D_{dev} and D_{test}. The data plot of Fig. 9 using sweetviz[15] shows that each colored and grouped feature bin pair highly matches with each other. However, line plot from each RUL against selected features shows RUL lines are mostly but not always in synchronization with each other. Still, this is a good split and can directly be used as training and testing data, without consolidating and recreating from the beginning.

While performing EDA authors spotted a non-intuitive relationship between cycle and RUL. Usually, the cycle should increase, and the target (RUL) should decrease till the target is 0 for a single unit. However, this is not the case. Upon closer inspection of D_{dev} and D_{test}, the authors found this is not how the data is, including where the target is 0. There is presence of multiple consecutive repeating RUL value instances from all units (unit_A) and class types (Fc_A), including when RUL is 0. Presumably, dataset creator(s) wanted to capture more instances of each RUL, including more data for the end of life reached engine's status. If unique RUL for each unit is required, then future researchers may drop duplicate instances of repeating RUL by appropriately grouped data.

Further, it is observed that despite RUL being 0, hs_A may or may not be 0 (False) in both D_{dev} and D_{test}. In total about 3.6 million and about 2.89 million instances from respective D_{dev} and D_{test} are discovered where RUL is not 0, i.e., some value more than 0 and health status (hs_A) is 0 (False), i.e., not 1 (True). In comparison, 84880 and 61628 values from respective D_{dev} and D_{test} have been observed where both values are 0 (could be perceived as False in case of hs_A). Authors relied on given RUL values for further analysis, rather than generating RUL from available features- cycle (cycle_A) and health status (hs_A), by finding maximum cycle per unit (unit_A) group till hs_A is 0 and consider that value as first (maximum) RUL for that unit group. Extracted feature, health status (hs_A) has been refrained from further exploration as this binary variable seemingly not correlated with target Fig. 10 shows behaviour of health status, hs_A.

C. Unsupervised clustering

Another interesting aspect of data could be seeing how numerical features might form clusters, as most of the features seem to have a mutual high correlation. This still can be part of data exploration as authors used sub-sampled 1% instances or 55713 instances of D_{dev}, to get an overview of labels (i.e. groups of instances) and outliers count. Appropriate stratification based on all categorical values has been applied while sub-sampling the data. As default train and test split is mostly related as observed and stated before, sub-sampling from any part will show the true nature of data clusters. Authors trained a density-based unsupervised model, OPTICS with cluster method as "xi" from famous scikit-learn machine learning package [12]. Maximum epsilon value, max(ε) or max_eps for considering the maximum distance between two samples as mutual neighbour have been set based on finding ascending sorted nearest neighbour distances of type "minkowski" with p-value 2, aka, Euclidean distance and thereafter finding "knee point " or "elbow point", based on [13]. max_eps was found as approximately 32.676. Fig. 11 shows the graph of found knee/elbow point.

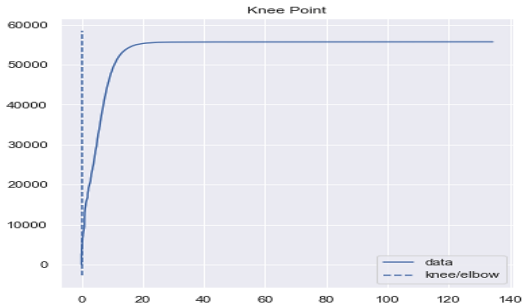

Fig. 11 - Knee point or max ε for OPTICS.

Minimum number of points required to form a dense region (minPts) has been set as select feature dimension of D_{dev} multiplied by 2, i.e., 6 in this case, based on [14]. Distance computation metric was "minkowski" with p-value 2. Other parameters were unaltered. Authors found 10 different clusters and a large number of outliers, i.e., 11 types of labels for these features. 55649 out of 55713 sub-samples or about 99.88% were not part of any clusters. Fig. 12 and Fig. 13 shows major cluster groups where group value count is more than 500 for outliers and all other groups respectively. It can be apprehended from these cluster labels that selection criteria for numerical features which are mostly not correlated with each other but having most correlation score with the target, can be relaxed further to get more such features, as commonality among features are being shown as extremely low.

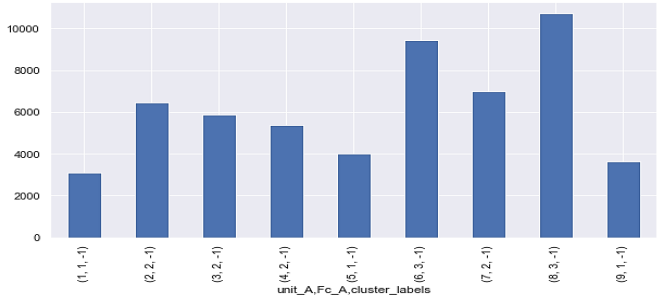

Fig.12 - Major outlier groups of (unit, flight class) pair.

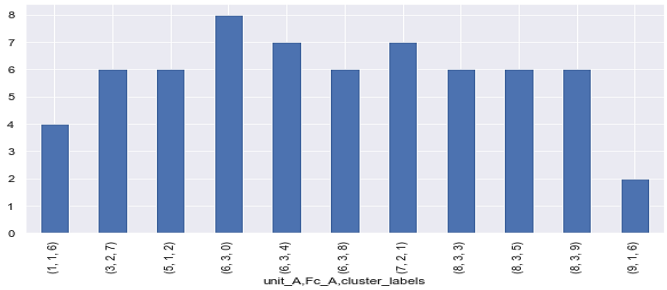

Fig. 13 - Major cluster groups of (unit, flight class) pair.

IV. CONCLUSION

Maintaining a good health of safety-critical equipment is vital for the success of the mission and for enhancing safety. Condition Based Maintenance (CBM) comes in handy under such situations as it prevents unwanted equipment downtime and is less costly than preventive maintenance. Prognostics is one of the enablers of CBM and is widely used in industries to predict the remaining useful life of machinery equipment. Data-driven prognostics rely on a spate amount of data which for safety-critical equipment is mainly generated synthetically. Authors performed exploratory data analysis (EDA) on one such run-to-failure data (termed as N-CMAPSS dataset) for a small fleet of aircraft engines under realistic conditions.

While performing EDA authors found out that most of the input features were not correlated to the target variable (RUL), nevertheless, some numerical variables showed a high degree of correlation among themselves. This information can be used for dimensionality reduction as parameters having high correlation can be represented by one parameter only while building a prognostics model. The input parameter having the highest correlation to RUL was cycle_A. However, authors spotted a non-intuitive relationship between the input parameter cycle and RUL as discussed in the paper. It was also discovered that for categorical variable (flight class), Fc_A_3 was most dominant among the three classes. Assigning a proportionate weight of target based on these count value ratios might help in creation of more accurate prognostic models in the future.

Likewise, the bar plot (Fig. 6) unveiled instances of each three flight class types are not present in every unit. Which makes unit wise prognostic modelling of each class impossible.

Moreover, the unit-wise box plot (Fig.8) grouped by flight class, depicts that features are highly spread and consist of a fairly large number of outliers. Therefore, applying a robust data scaling and transformation scheme could be helpful in creating an accurate prediction model.

DECELRATION

The views expressed in this article/presentation are author's own and Wipro does not subscribe to the substance, veracity, or truthfulness of the author's views. Conflict of interest declaration number 18022 has been acknowledged and accepted by Wipro on 06/07/2021.

REFERENCES

[1] N. H. Kim, D. An, and J. H. Choi, "Prognostics and health management of engineering systems: An Introduction," Springer International Publishing, Switzerland, 2016.

[2] ISO13381-1, Condition monitoring and diagnostics of machines -Prognostics – part1: General Guidelines, 2015.

[3] A. Keprate and R. M. C. Ratnayake, "Selecting a modelling approach for predicting remnant fatigue life of offshore topside piping," *IEEE International Conference on Industrial Engineering and Engineering Management,* Bali, Indonesia, 2016.

[4] A. Keprate, R. M. C. Ratnayake, and S. Sankararaman, "Minimizing hydrocarbon release from offshore piping by performing probabilistic fatigue life assessment," *Process Saf. Environ.*, vol. 106, pp. 34–51, 2017.

[5] A. Keprate and R. M. C. Ratnayake, "Remaining Fatigue life prediction of topside piping using response surface models," *IEEE International Conference on Industrial Engineering and Engineering Management,* Bangkok, Thailand, 2018.

[6] A. Keprate and R. M. C. Ratnayake, "Handling uncertainty in the remnant fatigue life assessment of offshore process pipework", *International Mechanical Engineering Congress and Exposition*, Phoenix, Arizona, USA, 2016.

[7] http://ti.arc.nasa.gov/project/prognostic-data-repository NASA Ames Research Center, Moffett Field, CA, USA.

[8] https://se.mathworks.com/products/simulink.html

[9] J. Wickert and K. Lewis, "An introduction to mechanical engineering", ISBN 978-1-305-63513-5, Cengage Learning, 2017.

[10] M. A. Chao, C. Kulkarni, K. Goebel, and O. Flink, "Aircraft engine run to failure dataset under real flight conditions for prognostics and diagnostics", Data 6, MDPI, Switzerland, 2021.

[11] S. Chatterjee and A. Keprate, "Predicting remaining fatigue life of topside piping using deep learning," International Conference on Applied Artificial Intelligence, Halden, Norway, 2021.

[12] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[13] V. Satopaa, J. Albrecht, D. Irwin and B. Raghavan, "Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior," 2011 31st International Conference on Distributed Computing Systems Workshops, 2011, pp. 166-171.

[14] Sander, J., Ester, M., Kriegel, HP. et al. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. Data Mining and Knowledge Discovery 2, 169–194 (1998).

[15] https://github.com/fbdesignpro/sweetviz

[16] Waskom, M. L., (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021, https://doi.org/10.21105/joss.03021.

APPENDIX

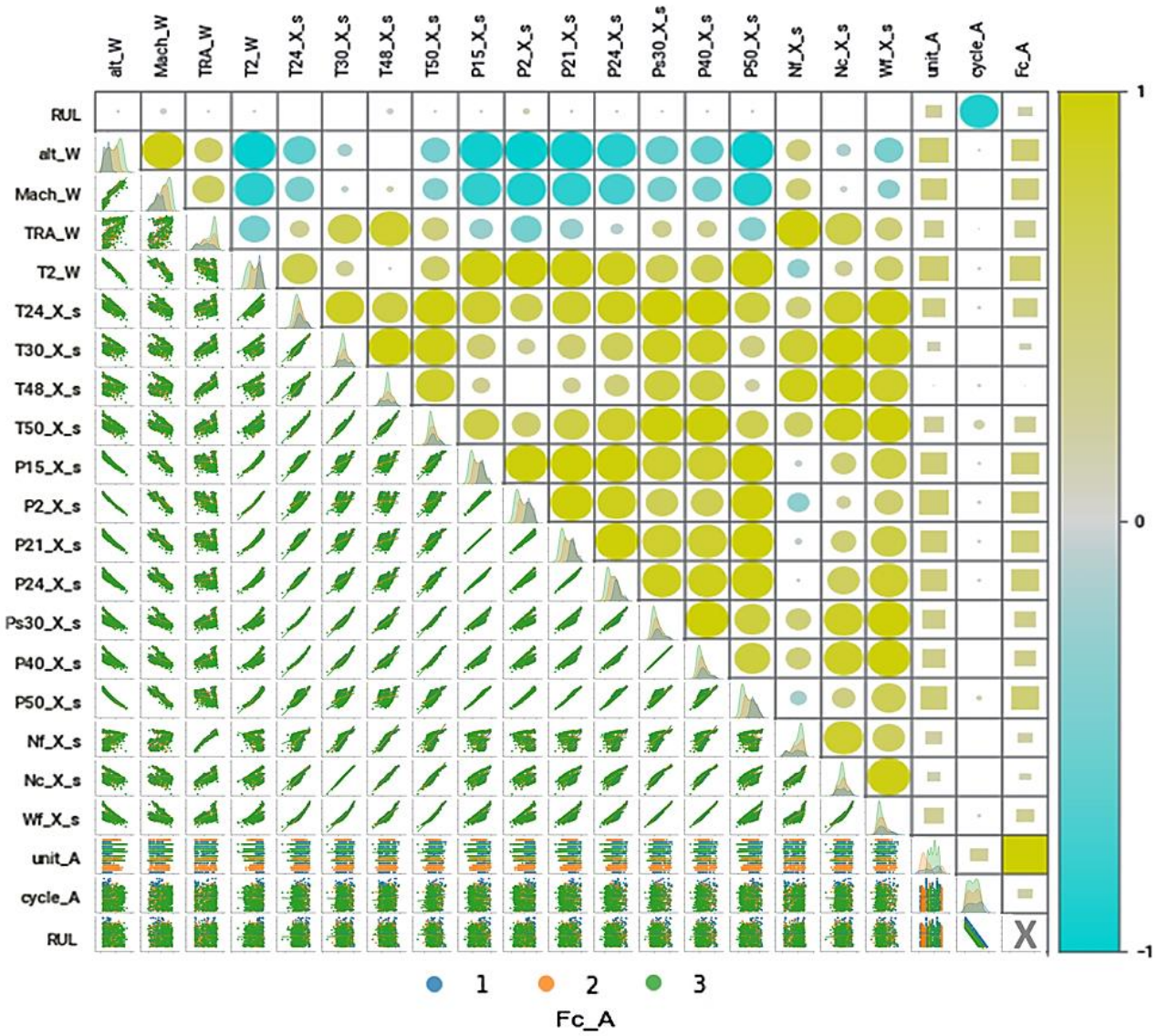

Fig. 4 - Combined association plot and coarse class-grouped scatter plot.

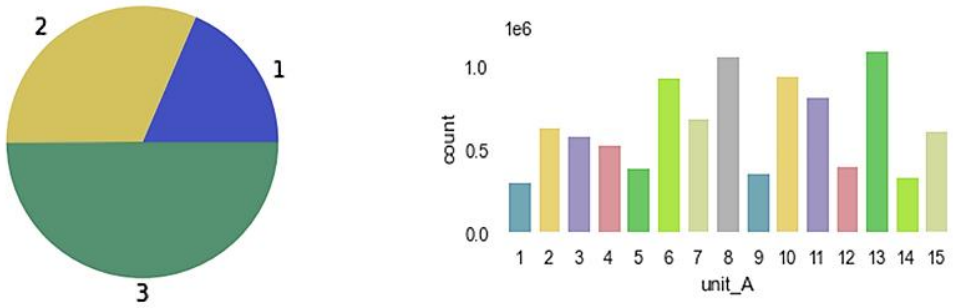

Fig. 5 - Pie chart on left depicting flight class distribution and bar chart on right depicting count of engine units.

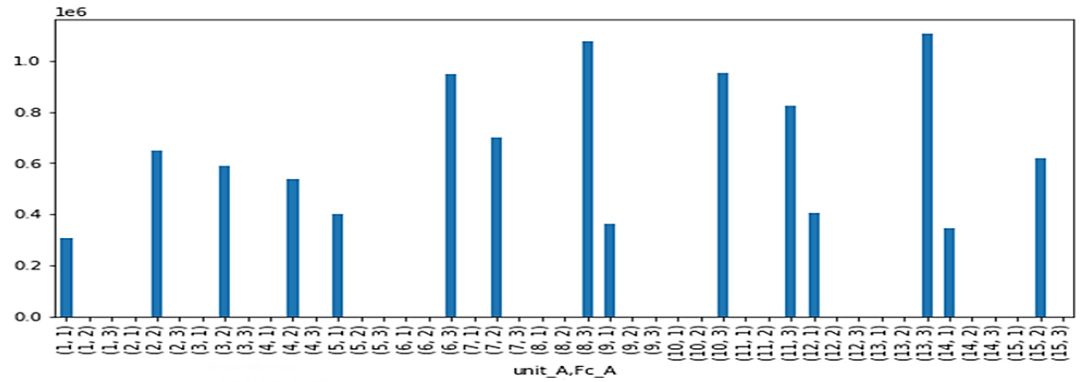

Fig. 6 - Bar plot of count distribution of each flight class among different engine units.

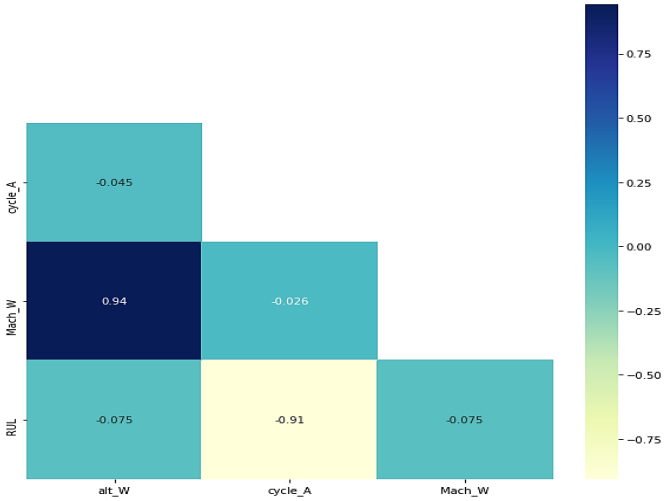

Fig 7 - Top features.

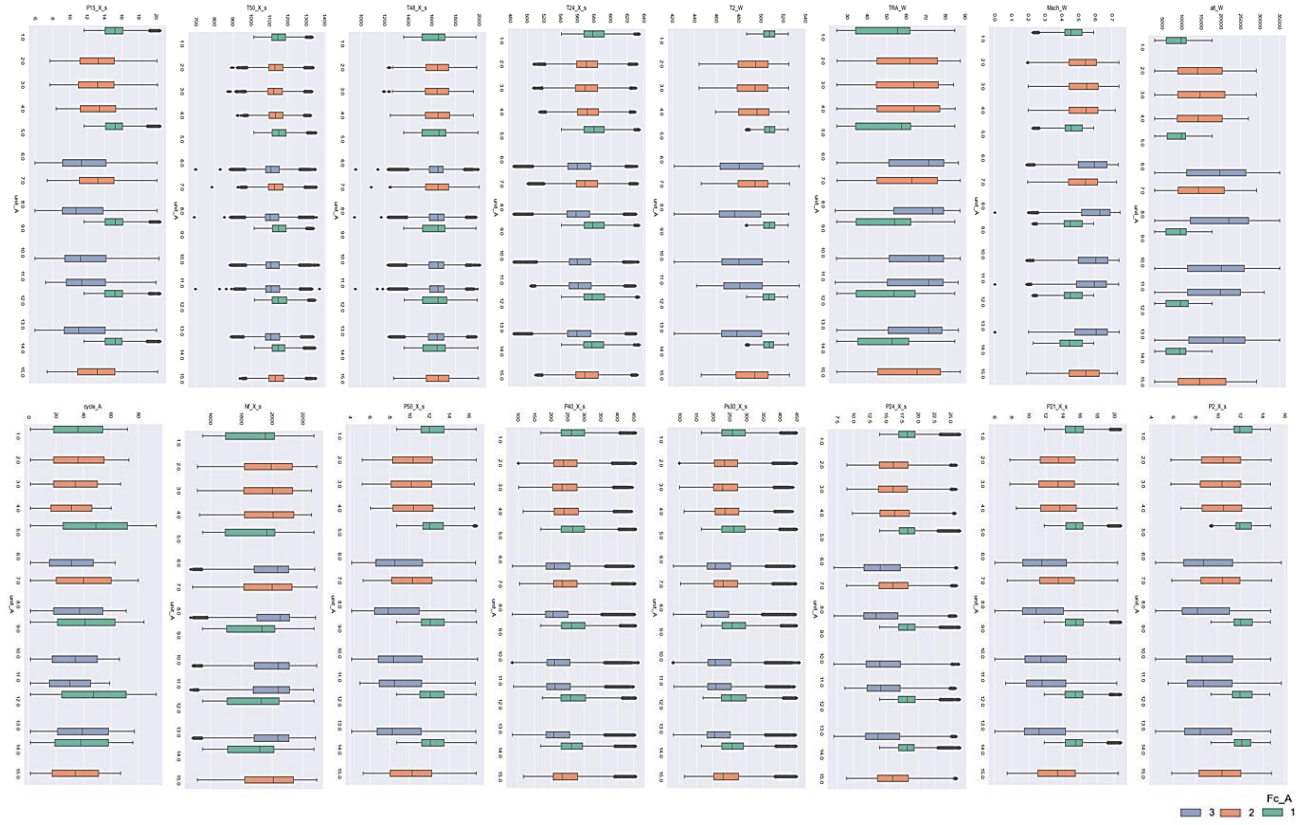

Fig. 8 - Box plot of most correlated features with RUL grouped by unit and flight class.

Fig. 9 - Detailed analysis of most correlated features of target.

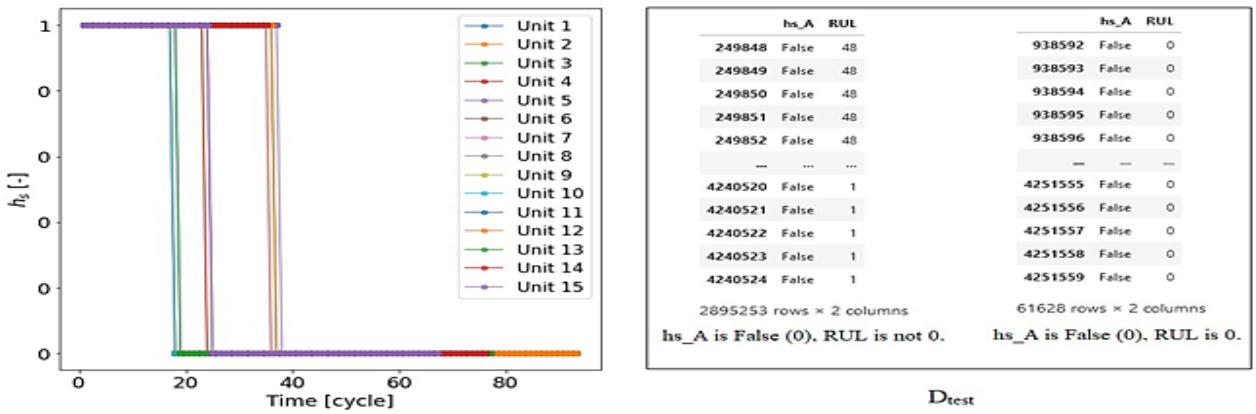

Fig. 10 - Behaviour of health status with respect to RUL and cycle.