Lappeenranta-Lahti University of Technology LUT
BM20A6100 Advanced Data Analysis and Machine Learning
Practical Assignment

Week 3

Data Pretreatment

Joseph Mugumya
Shubham Khaire
Maaz Tariq Bhatti

In continuation of last week's submission, where we primarily focused on exploratory data analysis, this week our main objective is data preparation for modelling.

Division of data into calibration, validation and test partitions

In preparation of our data for the model building, we divided the train data into two subsets: 80% will serve as the training set, allowing us to gauge model performance, while the remaining 20% will serve as a test set. This division helps us to test and fine-tune our models, ensuring they generalize well to unseen data.

By having this validation set, we can optimize our models' parameters and configurations, enhancing their accuracy and reliability.

Feature Selection

		Count	Mean	St.dev	Min	P25	P50	P75	Max
1	unit number	20631	51.5066	29.2276	1	26	52	77	100
2	time in cycles	20631	108.8079	68.8810	1	52	104	156	362
3	op setting 1	20631	-8.8701e-06	0.0022	-0.0087	-0.0015	0	0.0015	0.0087
4	op setting 2	20631	2.3508e-06	0.0003	-6.0000e-04	-2.0000e-04	0	3.0000e-04	6.0000e-04
5	op setting 3	20631	100	0	100	100	100	100	100
6	sensor measurement 1	20631	518.6700	0	518.6700	518.6700	518.6700	518.6700	518.6700
7	sensor measurement 2	20631	642.6809	0.5001	641.2100	642.3225	642.6400	643	644.5300
8	sensor measurement 3	20631	1.5905e+03	6.1311	1.5710e+03	1.5863e+03	1.5901e+03	1.5944e+03	1.6169e+03
9	sensor measurement 4	20631	1.4089e+03	9.0006	1.3822e+03	1.4024e+03	1.4080e+03	1.4146e+03	1.4415e+03
10	sensor measurement5	20631	14.6200	0	14.6200	14.6200	14.6200	14.6200	14.6200
11	sensor measurement 6	20631	21.6098	0.0014	21.6000	21.6100	21.6100	21.6100	21.6100
12	sensor measurement 7	20631	553.3677	0.8851	549.8500	552.8100	553.4400	554.0100	556.0600
13	sensor measurement 8	20631	2.3881e+03	0.0710	2.3879e+03	2.3881e+03	2.3881e+03	2.3881e+03	2.3886e+03
14	sensor measurement 9	20631	9.0652e+03	22.0829	9.0217e+03	9.0531e+03	9.0607e+03	9.0694e+03	9.2446e+03
15	sensor measurement 10	20631	1.3000	0	1.3000	1.3000	1.3000	1.3000	1.3000
16	sensor measurement 11	20631	47.5412	0.2671	46.8500	47.3500	47.5100	47.7000	48.5300
17	sensor measurement 12	20631	521.4135	0.7376	518.6900	520.9600	521.4800	521.9500	523.3800
18	sensor measurement 13	20631	2.3881e+03	0.0719	2.3879e+03	2.3880e+03	2.3881e+03	2.3881e+03	2.3886e+03
19	sensor measurement 14	20631	8.1438e+03	19.0762	8.0999e+03	8.1332e+03	8.1405e+03	8.1483e+03	8.2937e+03
20	sensor measurement 15	20631	8.4421	0.0375	8.3249	8.4149	8.4389	8.4656	8.5848
21	sensor measurement 16	20631	0.0300	0	0.0300	0.0300	0.0300	0.0300	0.0300
22	sensor measurement 17	20631	393.2107	1.5488	388	392	393	394	400
23	sensor_measurement 18	20631	2388	0	2388	2388	2388	2388	2388
24	sensor measurement 19	20631	100	0	100	100	100	100	100
25	sensor measurement 20	20631	38.8163	0.1807	38.1400	38.7000	38.8300	38.9500	39.4300
26	sensor measurement 21	20631	23.2897	0.1083	22.8942	23.2218	23.2979	23.3668	23.6184

Figure 1: Table showing the summary of our dataset. 7 Predictors have 0 st.deviation

Based on our data summary analysis, we have identified certain variables with a standard deviation of 0. This implies that these variables will not contribute to our modelling process and will not provide any additional information. Therefore, as part of our data preparation, we have removed predictors 6, 10, 11, 15, 21, 23, and 24.

Since our predictions are solely based on sensor data, we have retained only those predictors that represent the sensors. Consequently, we are left with 14 predictors- all Sensor Measurements, that will be utilized in constructing our model. Sensor Measurements (2,3,4,7,8,9,11,12,13,15,17,20,21)

Scaling and Normalization

In data preparation, scaling and normalization is a crucial step, as it ensures that all predictors are on the same scale. This prevents certain predictors from dominating the model process solely due to their larger values. In this way the model will capture the more meaningful patterns in the data. Thus, accurate predictions.

In our case we used the z-core technique achieving a mean of 0 and St. Deviation of 1.

$$Z = \frac{x - \mu}{\sigma}$$

Where:

- Z is the Z-score of the data point x.
- x is the individual data point.
- μ is the mean (average) of the dataset.
- σ is the standard deviation of the dataset.

From the box plots displayed below, a striking contrast emerges between the variations in scaled and normalized data as opposed to unscaled data. In the box plot illustrating the scaled and normalized data, the variations are notably reduced, resulting in a clearer and more interpretable representation compared to the unscaled data.

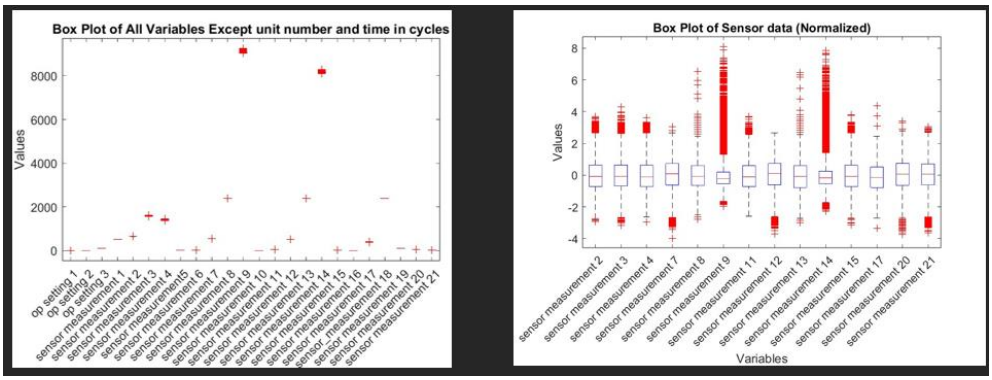

Figure 2: Comparison between the scaled and unscaled data using Box Plots

PCA

Principal component Analysis, a technique used in data analysis for dimensionality reduction. It works by retaining only the essential information by way of maximizing the Variance in the data.

We therefore applied Principal Component Analysis (PCA) to our dataset consisting of 14 predictors. MATLAB's built-in PCA function was utilized for this purpose. PCA also helped us to identify the principal components within a dataset, each of which captures different aspects of data varianceas explained below.

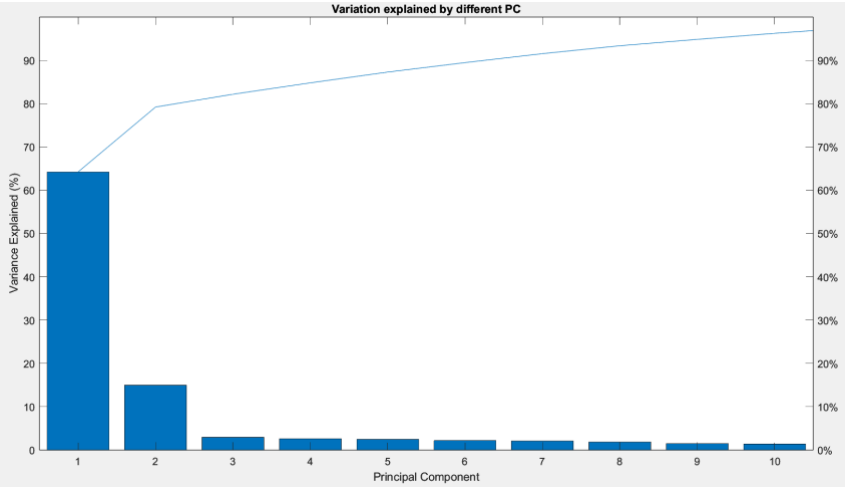

Figure 3: Pareto Graph explaining the variance contribution from the different PCs.

The results of our PCA analysis are visually represented in a Pareto Graph. This graph allows us to understand how much variance each Principal Component explains within the dataset. Notably, PC1 and PC2 together explain a substantial 79% of the data's variance, while PC1 to PC3 collectively account for 82%. This insight is invaluable, as it guides us in focusing on the most influential components for our subsequent modelling and analysis.

Contribution of variance by different predictors in PC 1 and 2

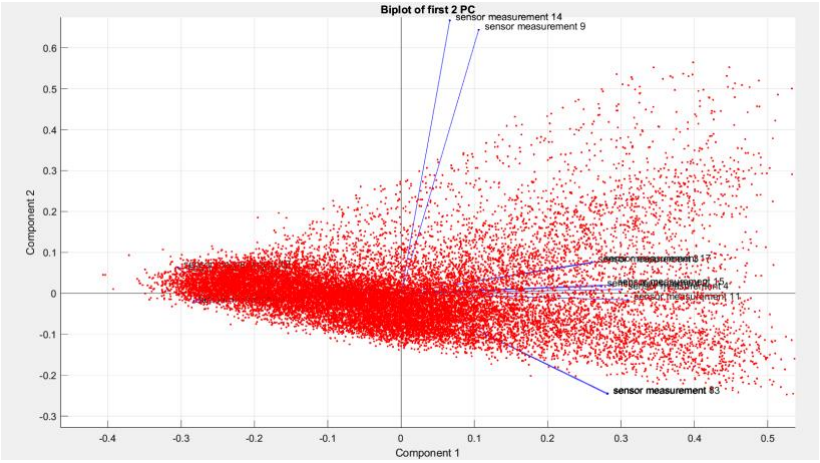

Figure 4: Bi plot of PC 1 and 2

In the context of PC2, Sensor Measurement 14 and 9 emerge as prominent contributors, signifying their substantial influence on the variation captured by this component. Conversely, in PC1, Sensor Measurement 11, 4, and 3 are the primary drivers of variance, underlining their pivotal roles within the dataset.

Furthermore, the Biplot allows us to understand correlations between predictors. For instance, a negative correlation is evident between Sensor Measurement 4 and 7, as well as between Sensor Measurement 11 and 12.

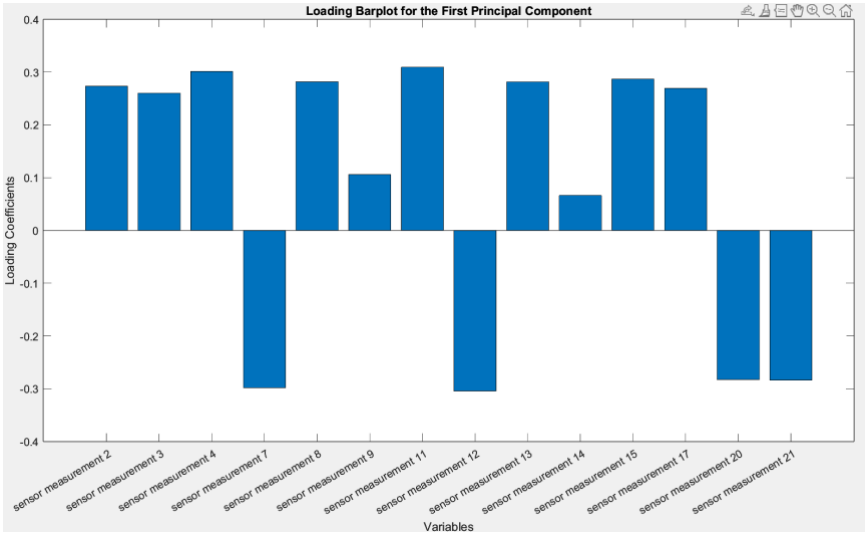

Figure 5: Contribution of the variance by the predictors in PC1.

Outlier detection

In our task of outlier detection, we employed the T2 graph, a tool used to detect and uncover anomalies within datasets. These anomalies may signify data points that diverge markedly from the expected norms or manifest uncommon patterns.

To effectively detect these outliers, we employed a control limit set at 3 standard deviations. This threshold acts as a benchmark, helping us distinguish between data points that fall within the expected range and those that warrant closer scrutiny. When a data point surpasses this limit, it serves as an alert. We thus removed data points that fell way above our threshold.

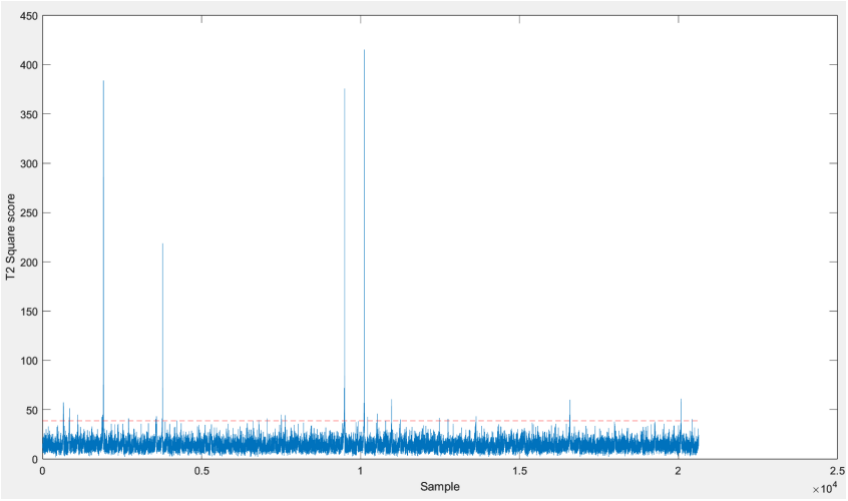

Figure 6:T2 Graph

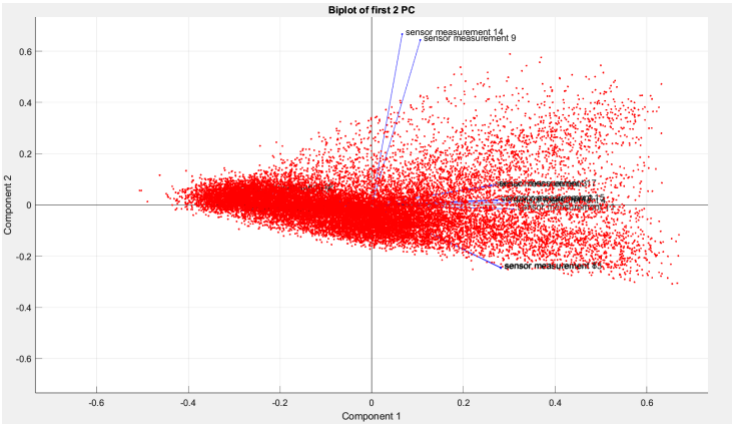

Figure 7: Bi plot after Outlier Reduction.