Data Analysis and Model Development Report

- Yash Mayur

7. Data Collection

There were 3 types of dataset available for predictive maintenance; out of that ai4i2020 dataset was used as it has 10,000 product information. So it was good for training and testing purposes. Silicon wafer data was also collected but electronic maintenance was somehow getting diverted from actual 3 types of dataset what we already collected. Hence Silicon wafer data is kept as reference purpose.
Below is the feature description of the dataset used for analysis and model development.

a. Feature Description:

1) **Product ID**: consisting of a letter L, M, or H for low (50% of all products), medium (30%) and high (20%) as product quality variants and a variant-specific serial number.

2) **Type**: just the product type L, M or H from column 2.

3) **Air Temperature [K]**: generated using a random walk process later normalized to a standard deviation of 2 K around 300 K.

4) **Process Temperature [K]:** generated using a random walk process normalized to a standard deviation of 1 K, added to the air temperature plus 10 K.

5) **Rotational Speed [rpm]:** calculated from a power of 2860 W, overlaid with a normally distributed noise.

6) **Torque [Nm]:** torque values are normally distributed around 40 Nm with a SD = 10 Nm and no negative values.

7) **Tool Wear [min]:** (breakdown and gradual failure of a cutting tool due to regular operation) The quality variants H/M/L add 5/3/2 minutes of tool wear to the used tool in the process.

8) **Machine failure:** Machine Failure label that indicates, whether the machine has failed in this particular datapoint for any of the following failure modes are true. The machine failure consists of five independent failure modes as follows:

a) **Tool wear failure (TWF):** the tool will be replaced of fail at a randomly selected tool wear time between 200 - 240 mins (120 times in our dataset). At this point in time, the tool is replaced 69 times, and fails 51 times (randomly assigned).

b) **Heat dissipation failure (HDF):** heat dissipation causes a process failure, if the difference between air and process temperature is below 8.6 K and the tools rotational speed is below 1380 rpm. This is the case for 115 data points.¶

c) **Power failure (PWF):** the product of torque and rotational speed (in rad/s) equals the power required for the process. If this power is below 3500 W or above 9000 W, the process fails, which is the case 95 times in our dataset.

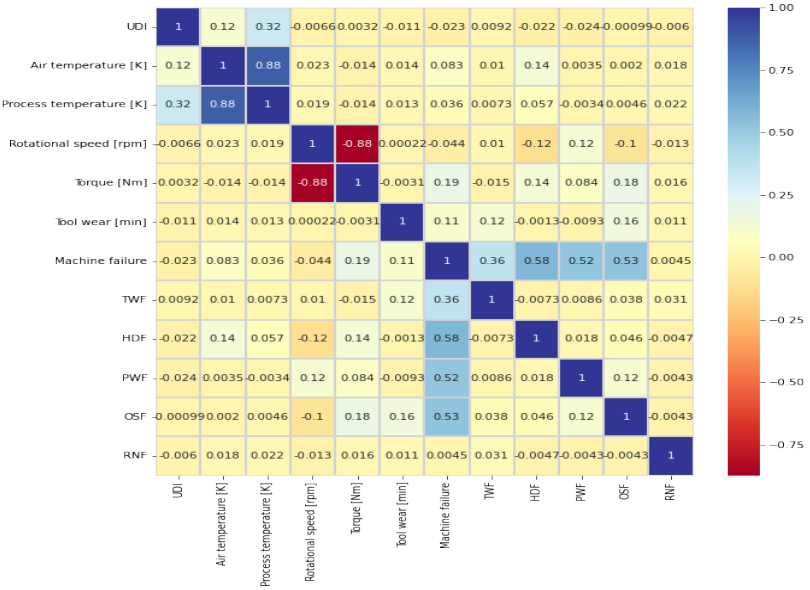
d) **Overstrain failure (OSF):** if the product of tool wear and torque exceeds 11,000 minNm for the L product variant (12,000 M, 13,000 H), the process fails due to overstrain. This is true for 98 datapoints.

e) **Random failures (RNF):** each process has a chance of 0,1 % to fail regardless of its process parameters. This is the case for only 5 datapoints, less than could be expected for 10,000 datapoints in our dataset.

Note that if at least one of the above failure modes is true, the process fails and the 'machine failure' label is set to 1. It is therefore not transparent to the machine learning method, which of the failure modes has caused the process to fail.

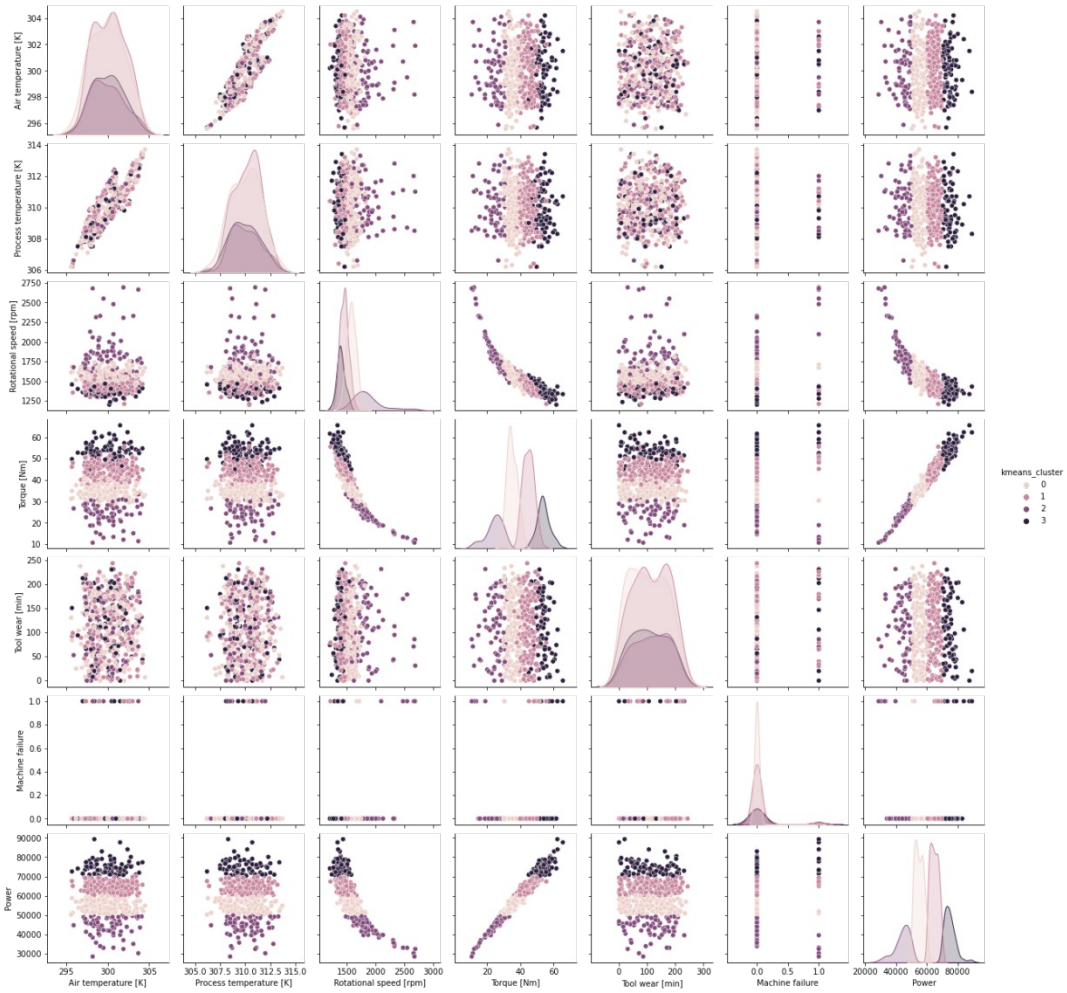
8. Exploratory Data Analysis

a. Correlation:



We can see that there are strongly correlated features namely process temperature and air temperature. Torque and rotational speed are also negatively correlated. We can drop one of the temperatures, but the torque to rotational speed difference might be a indication of a failure, so we'll keep both.

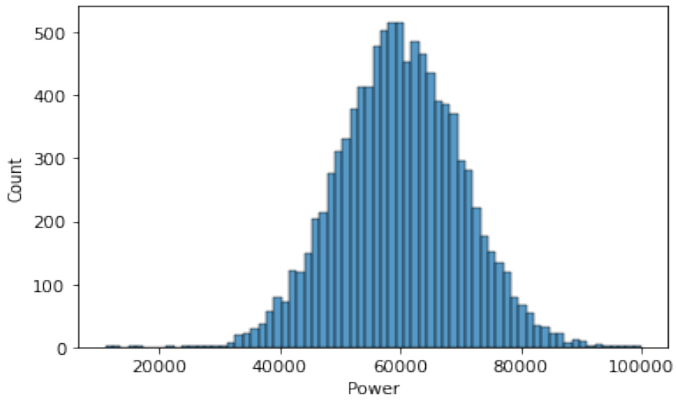
b. Linearly dependable:



Process Temperature and Air Temperature have Linear Relationship with each other and both the data has gaussian normal distribution

c. Power Attribute:

A new feature has been added to dataset



Power = Torque × Rotational speed

d. Silhouette Score:

We can say that the clusters are well apart from each other as the silhouette score is closer to 1 and greater than 0.5

9. Feature Engineering (Data Preprocessing):

a. Encoding of Categorical Feature:

Use any one from the below encoding technique

i. **Label Encoding:**
ii. **One Hot Encoding:**

b. Imbalance Dataset Handling:

As analysed and Mentioned earlier that the dataset is imbalanced so we have to use SMOTE technique on training data inorder to make balance data

c. Missing Values

There are no missing values present so the dataset is clean

10. Model Development and Evaluation:

- As this is binary classification problem, we used many types of classification algorithms like Random forest, Decision tree, KNN classifier etc (all details available in Jupiter notebook.
- Catboost classifier comes to be the best performing classifier with highest accuracy.
- Model.pkl file is created for catboost classifier for the deployment purpose.
- Classifier model is deployed on Google server (GCP) and it is attached to webapp as well.