Welcome to the Digital M&S Data Science Challenge!

In this challenge, you will construct models related to manufacturing yield that can help Sanofi process engineers answer their questions.

Good luck, and let's get started!

**Background**

Manufacturing yield is a critical indicator used in the manufacturing to gauge production processes efficiency. It represents the percentage of high-quality output produced from raw materials.

Manufacturing yield significantly impacts a company’s profitability by affecting production costs and sales revenue. A low yield can result in increased production costs and reduced profits. As a result, manufacturers prioritize monitoring and enhancing yield.

The drug manufacturing process is complex and involves hundreds of parameters, making it difficult for process engineers to identify which parameters most significantly impact yield.

This is where AI/ML comes into play. By leveraging machine learning algorithms, we can analyze vast amounts of data and identify critical parameters that impact yield. The algorithm can recommend thresholds to optimize yield while adhering to regulatory standards. Process engineers can make informed decisions based on these algorithmic outputs.

Now it is time to put your skills into good use and see if you can help the Sanofi process engineers to understand yield. Your task is to construct machine learning models that answer the following questions from Sanofi process engineers

* What are the key parameters influencing manufacturing yield?
* What recommendations would you give to improve the manufacturing yield?

**The Data**

To answer these questions, refer to the provided file "Datainput.xlsx". It contains yield metrics and key data sources that may affect yield. The following sheets are included:

1. Sheet **“Combined\_Yields”**

Target metric “YIELD\_VALUE”:

* 1. This tab gives you the “output batches” coming out from the manufacturing process and their yields. You can use the “activity\_coefficient” (column YIELD\_NAME) of these output batches as the target in your model
  2. [Bonus point] if you have more time and want to explore more, you can also try fitting another model using “overall\_activity\_yield” (column YIELD\_NAME) as the target variable

1. Sheet **“Campaign\_info”**

Campaign information:

1. The manufacturing process is organized into separate campaigns with multiple batches produced within each campaign. The sheet provides the **CAMPAIGN\_ID** and **BATCH\_POSITION** within the campaign to indicate the sequence of each batch produced
2. Engineers know that later batches in a campaign may show artificially higher yield due to accumulated remains in the tank. This is a known factor and may appear in your model, but process engineers cannot learn much from it
3. ***How to use it:*** Join the campaign details to the **output batches** using BATCH\_ID and MATERIAL\_ID
4. Sheet **“IoT”**

Sensor data:

1. During the manufacturing process, sensors record batch attributes like temperature and chemical levels. Process engineers focus on these variables and often seek to figure out the optimal settings for these values to achieve the best yield
2. ***How to use it:*** Link the sensor information to the final batches by the join key BATCH\_ID and MATERIAL\_ID
3. Sheet “**QC\_Data”**

Quality control (QC) is a crucial gatekeeper at every step of the manufacturing process. This sheet captures all the QC tests that have been conducted on the batches:

* 1. QC tests are specific to the material and batch stages, so not all batches will have the same tests. This may result in empty cells on the sheet. Determine how to manage the empty cells in this scenario
  2. ***How to use it:*** As mentioned in the next section (Batch Genealogy), there are multiple ways to add QC data to the model. The process engineers are interested in seeing two things:
     1. The QC attributes of the output batches
     2. The 'weighted QC feature' from the parent batches of the output batches. For more information on what constitutes a parent batch, please refer to the Batch Genealogy section below

1. Sheet **“Batch Genealogy”** Important!

Table

Description automatically generated Table

Description automatically generated

Figure 1: Batch genealogy for a hypothetical final “Batch D”.

1. The manufacturing process often involves multiple steps, and the batch ID changes at each stage to differentiate it from earlier stages. Refer to Figure 1 where Batch “A” becomes Batch “D” at a later stage
2. You can trace the output batch (e.g., Batch D) back to earlier batches that served as its parents, such as Batch A and Batch B. This family tree is called batch genealogy and is often depicted as a graph or tree
3. To identify parent batches for an output batch.
   * 1. Locate the output batches by selecting **movement type “101”** in the sheet “Batch\_Genealogy”
     2. The parent batches for each output batch will have the same **“PROCESS\_ORDER\_NUMBER” and a movement type as “261”** (Refer to Figure 1)
     3. Movement type “531” can be disregarded in this exercise

|  |  |
| --- | --- |
|  | Movement type |
| Parent batch | 261 |
| Output batch | 101 |

1. ***How to use it:*** Understanding batch genealogy is crucial in analyzing manufacturing yield:
   * 1. Output batches, which have movement type “261” in the Batch\_Genealogy sheet, are the ones used in yield calculation – they should have the target yield numbers in sheet Combined\_Yields
     2. However, process parameters affecting yield need to be analyzed in both output and parent batches. To incorporate the QC variable mentioned in the previous section, you need to look at the QC test associated with **both the output batch and the parent batches** (“Batch A, B” in Figure 1)
     3. To account for QC tests in parent batches, a weighted feature can be constructed using the input quantities (“QUANTITY” in Batch\_Genealogy). The weighted QC attribute for a batch can be obtained by summing up the QC variables of its parent batches weighted by their quantities, and divided by the sum of their quantities.
     4. Weighted QC attributes =

|  |  |  |
| --- | --- | --- |
| Batch | Quantities (Qi) (“QUANTITY” from sheet “Batch\_Genealogy”) | QC attribute value (QCi)  (“VALUE” from sheet “QC\_Data”) |
| Batch A | 10 | 2.0 |
| Batch B | 20 | 3.0 |
| Weighted QC attribute value to be used in the model for Batch D |  | 2.66 |

Table 1: An example showing calculation of weighted QC attributes.

Table

Description automatically generated

Table 2: An example from this dataset

**Additional comments**

1. To add some final information, the QC features listed below were used to determine the target yield definitions, which means they are inherently related to the yield target

|  |  |
| --- | --- |
| **Yield Definitions** | Features Used |
| **Activity Coefficient** | caog\_dry, xa\_dry |
| **Overall activity yield** | lod, coag\_dry, xa\_dry |
|  |  |

1. Make sure to consider the BATCH\_ID and MATERIAL\_ID for all sheets as they define a production unit in the process. All metrics and raw data inputs are aggregated to the BATCH\_ID+MATERIAL\_ID level

With this information, you should be able to get started. Don't hesitate to reach out if you have any questions. We'll see you at the finish line!