Proposal:

* Running Correlation on Sub-unit level, instead of Plant Level
* Finding form the past 10k Live Models.
  + Inference tested on 110 LIVE Models

Evidence:



Experimentation:

* Identify the sub-unit level tags and running correlation on those set of tags can save computation cost up to 90%.
* Ran the methodology to test out how many features belong to same unit for the above tags which have LIVE models.
  + On an Average **87%** of the features belong to the same subunit/unit as of the Target feature.

Table

Description automatically generated

* Using this as expansion ideology, ran the same methodology on set of 100 LIVE TAGS.
  + *78 LIVE Models have 100% of features belonging to the same Sub-unit to which the Target Tag also belongs to.*
  + *Only 22 Models do not reach 100% features.*
    - *Out of 22, only 20 models do not have up to 90% of features.*

Chart

Description automatically generated

Table

Description automatically generated

'''

1. Let user decide if they want to run the extension of correlation on the whole plant?

2. Introduce a new technique SelectKBest with f\_regression at this moment

- Default: Subunit

- Extension: On whole plant

'''



Comments from TAO:

* Santhip:
  + Holds TRUE for 80 to 90% of Assets.
  + May not work for assets where the UNIT NUMBER is not provided.
    - In such cases, the team would like to perform correlation on PLANT LEVEL.
* Amit:
  + The Models in the Rheinland asset have features from other units, but their feature importance is either 0 or too low to consider?
    - Coz, during bulk modelling those features got picked up, but the heaviest feature importance value is provided to features from same unit.
  + Jurong and Rheinland do follow the deduced nomenclature.
* Guru:
  + The Nomenclature seems perfectly fine for Pernis Asset.
  + The 3rd part which is termed as equipment in the assumption is known as type of measurement the equipment is installed for.
* Vijay:
  + The Nomenclature stand solid for all the downstream assets but the same does not applies to Upstream assets.
  + Also, there is no proper way to differentiate between Downstream and Upstream assets
    - Due to low Upstream assets currently present the correlation logic on UNIT will them identify better features.
* Thriambak:
  + Geismar follows the 3rd Nomenclature.
  + For Norco, there is no specific nomenclature, and they do not provide UNIT number in the tag.
* Sahl:
  + Deerpark follows the same nomenclature.

* For multiple LIVE Models across multiple Assets

Chart, bar chart

Description automatically generated

* ***The other method is that we have hierarchy in C3/while loading data of Unit/Sub-unit which we can reload and make it more feasible instead of running regex and finding the sub-unit.***