Machine Learning Chemical Manufacturing

This document will contain the areas researched for using ML in Honeywell mfg. The data is time series of continious batch processing sensor values of multiple phases from one line and product.

Of the multiple sensor data provided by client, only pi tags with .pv suffixed will be considered for analysis as the remainging may be irrelevant.

The output or dependent variables is the results of a lab test eg.ABD, the RCA files point towards the PCS and ABD caused by incorrect concentration of Al .Sol. Further analysis will be conducted on the sensor that captured the volume of Al and volume of its solution with the aim of reducing offspec batches.

New Pi tags data named "L3 Tags 2018 a" will be used for analysis henceforth. As the sensor data to be modeled with has been narrowed down to .pv temp and pressure data it is possible to conduct simple tests that will act as framework for detecting variance of pi tags across batches.

Upon building the framework for vector calculation, vectors will be calculated using statistical methods for a single batch and then compared to another randonmly selected batch. After which the single batch vector of choice will be compared with the batch that was offspec and test the

Null hypothesis: vectors for on and off spec batches can be different.

Extending that idea several offspec batches vectors will compared to an equal number of onspec batches, this method is commonly known as Clinical Research Trials or A/B testing. If the hypothesis holds true then all the offspec batches of the entire data provided will be mined out and compared to randomly selected on spec batches.

To find more differnece between SOME onspec and ALL offspec batches, the randomly selected onspec batches will be continiously rebagged or resampled. Then the conditions under which offspec batch can be detected will be clear.

In the event that the NULL hypothesis is accepted viz. there is no difference in the vectors between on and off spec batches. The A/B test will have to be weighted in favor of offspec batches.

Meaning an imblanced dataset is created opposite to the original where 98% of the batches are offspec and only 2% are onspec. The task here will be to identify an algorithim which could be Oversampling,distance,XGboost,SMOTE,ROSE or SVM which will clearly segregate the 2% onspec from the imbalanced dataset with minimal false positive rate.

Obviously there are several parameters that can be controlled such raising or lowering the threshold that classifies a batch failing or passing a parameter test and even the aggregated features upon which vectors are built and upon which statistical methods are applied could also be tinkered with.

The objctive is to find an algorithim that can either classify SOME offspec batches from ALL onspec batches or classify SOME onspec batches from ALL offspec batches.

By tinkering with

spec limits spec type sensor selection statistical treatment/vector creation variance measurement rebagging of control group Oversampling rare cases Undersampling majority cases OR Synthetic Minority Oversampling Technique

The hypothesis can eventually be proven that vectors of offspec batches are different from vectors of onspec batches consistently or that vectors of offspec batches are similar to vectors of other offspec batches.