Background materials for interview

In this case we outline a simplified chemical process, with accompanying data as an excel file. This is followed by some questions related to this dummy data. The goal is to discuss the outcome of these questions during the interview in more detail.

There is no fixed format or required platform for this analysis. It is fully up to you how to present these results, for example, but not limited to: guide us through the results, show us the reasoning steps, guide us through your data processing steps and code, show potential alternatives, potential problems, etc. However, please be advised that any code explained should be understandable by someone that does not understand the specific coding language.

General process introduction

In this case we are dealing with a (fictious) chemical process. In this process our aim is to create product B, starting from product A. Product A is brought into the vessel in which it spontaneously reacts to product B, with some (unwanted) remaining product A still inside.

The product A is continuously brought into the vessel and product B is continuously taken out of the vessel. In this way the level of the vessel does not change. This simplified process is shown in the figure below.

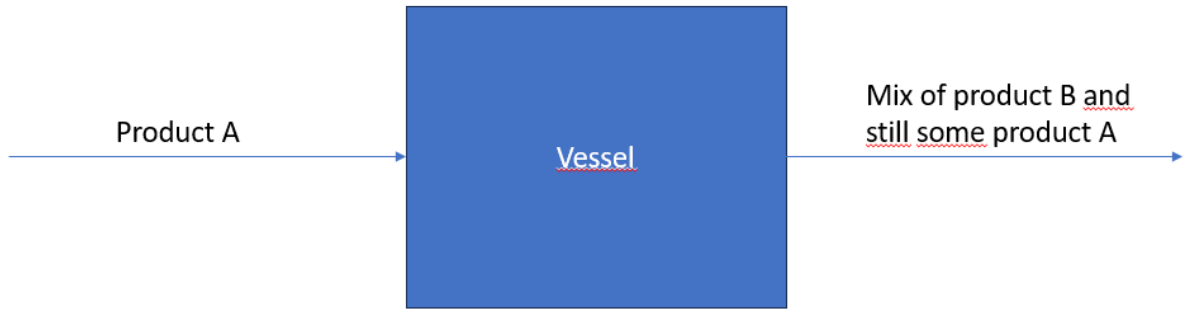

Figure 1: Simplified chemical process

Control of the vessel

In order to control the chemical process, different parameters are continuously monitored and shown to the operator (the person who is responsible for ensuring that the process runs safely and smoothly). The following table gives an overview of the parameters as they are measured and the basic meaning.

Table 1: Overview of the parameters measured and their meaning.

Location measurement	Parameter name	Meaning
Inflow	conti_Ca_in	Concentration of the incoming product A (~purity of incoming product)
Inflow	conti_m_in	Flow of the incoming product (e.g., kg/h)
Outflow	conti_Ca	Concentration of the outgoing product A, remember, this is an unwanted product
Outflow	conti_Cb	Concentration of the outgoing product B
Vessel	conti_Tsp	Temperature measured inside the vessel
Vessel	conti_mfree_A_liq	Total weight of product A in the liquid in the vessel

Vessel	conti_mfree_B_liq	Total weight of product B in the liquid in the vessel
Vessel	conti_mfree_A_gas	Total weight of product A in the gas in the vessel
Vessel	conti_mfree_B_gas	Total weight of product B in the gas in the vessel
Vessel	conti_Pinert	The pressure of the nitrogen gas inside the vessel
Vessel	conti_P_A	The pressure of product A in the vessel
Vessel	conti_P_B	The pressure of product B in the vessel
Vessel	conti_Ptot	The pressure measured inside the vessel

The data for each of these parameters measured across ~5 days is provided in the excel attached. Every row is exactly 15 seconds apart from the next row.

Questions

1. Currently the data is sampled at 15 second intervals.
    a. Comment on potential benefits of reducing the time interval and potential downsides of doing so.
    b. For reasons of complementarity with other data sets, we need to reduce the time interval to 5 seconds. Unfortunately we cannot remeasure. Create a secondary dataset with an interval of 5 seconds.
    c. Comment on the potential issues with taking your approach of the previous question.

2. When inspecting the data, you will see that data is missing. We agreed with the customer that any gap larger smaller than 9 minutes can be filled without comment. Gaps equal or larger than 9 minutes should also be filled, but we should also provide the customer with a list of timestamps when this event started.
    a. Fill the data in a way that seems appropriate to you.
    b. Provide a list of timestamps (=indices) where the gap was equal or larger than 9 minutes.
    c. Comment on potential issues when the gap becomes larger.

3. Currently, our feeling is that we might have too many variables (or parameters) to work with.
    a. Please provide an overview of the variables that are highly correlated (threshold can be chosen by you) and explain how this correlation was found.
    b. Create a new dataset where the number of parameters are reduced in such a way that a minimum of information is lost. Again explain how this was determined.

4. For the operator in charge of the process, it is important to understand how the concentrations of the products (conti_Ca & conti_Cb) in the outgoing flows evolve as a function of all other parameters.
    a. Is it possible to create a model to predict the evolution of these parameters for many steps in the future? How would such a model look like? Explain potential alternative models (the question is not about implementing these models) and their benefits and downsides.
    b. How would you assess the accuracy of this model over time (so not only 1 timestep ahead, but many steps into the future).

5. The operator sees the following values for conti_Ca appear on the screen (as shown in the table below). This trend seems familiar, but the operator does not immediately recall when this trend showed up almost the same in the past.

a. Could you retrieve the index in the database where this trend also appeared? Please take into account that you might want to look at the x best matches.
b. How similar is the current trend with the trend in the database?

Table 2: Trend seen by the operator.

conti_Ca
3.6348
3.6302
3.5898
3.5495
3.6168
3.5765
3.6079
3.5321
3.5634
3.5235
3.5903
3.5151
3.5850
3.5518
3.5202
3.5245
3.5633
3.6001
3.5982
3.5923
3.5464