Data Scientist Technical Challenge

This challenge is meant to test your technical skills on a problem similar to the ones faced at Pani Energy.

Background - Reverse Osmosis

Reverse osmosis is the process of creating fresh water from seawater. This is done by bringing the water into contact with a semipermeable membrane that allows water to pass much more easily than salt. When hydraulic pressure applied by a pump overcomes the chemical desire of the liquid to stay mixed, water permeates through the membrane. In practice, membranes are spirally-wound into membrane elements. Many membrane elements are placed in series inside a pressure vessel; these pressure vessels are placed in parallel to form the final configuration.

A RO Train is a group of pressure vessels in parallel which operates independently. A system is all RO trains in a given water treatment plant.

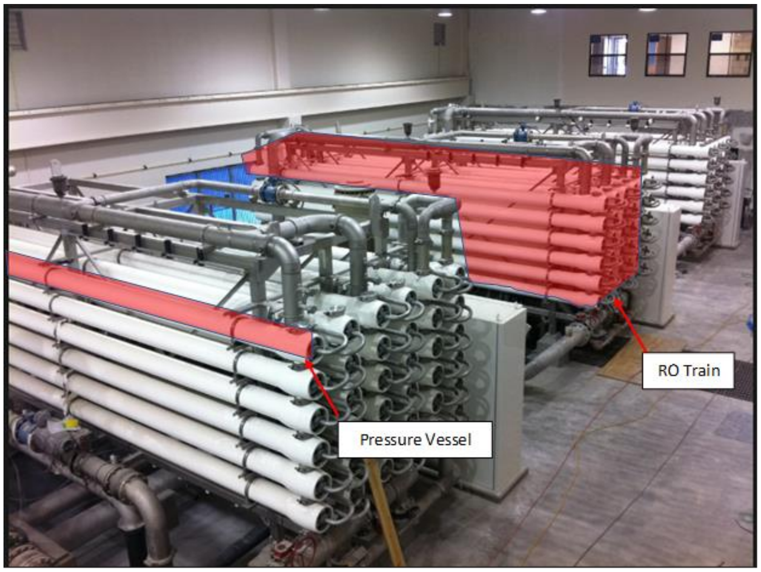

Figure 1: Assembly of membrane trains. High-pressure feed and waste streams are carried by the large stainless steel pipes, while product water at low pressure is carried by the smaller pipes

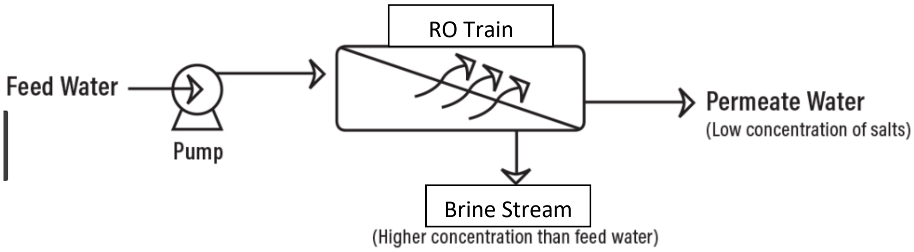
RO Train
Feed Water
Pump
Permeate Water
(Low concentration of salts)
Brine Stream
(Higher concentration than feed water)

Figure 2: Common representation of membrane train in drawings. Input feed water: salty water. Reject brine water: very salty water. This is the water that has all the salts rejected by the membrane. Permeate water or product water: water with very low quantity of salt.

Problem

Membranes degrade over time. Managing, controlling, and understanding this degradation is critically important for desalination plants. Chemical cleaning procedures are performed to restore performance, and eventually the membranes need to be replaced. Often, the cleaning and replacement costs are significantly more over the lifetime of the membrane than the power costs of the pump.

Goal

Fundamentally, we want a model to predict membrane current performance given past performance. This technical challenge is focused on this problem. More specifically:

Create a model of the target column that is accurate for as long a period in the test data as possible. Choose as a target column either A or stage_1_feed_pressure (see details section). This is timeseries data and with a basic algorithm the later you get in the test data the worse the performance you will get. You will also experience different performance in different test data. The challenge is to produce a model that is accurate for as long as possible in as many different regions of the data as possible.

Please submit:

- All code written to solve the problem
- Summary of your modelling process

Solving this problem in practice faces challenges. Significant among them are a lack of data, both in features and time. It is possible that the dataset does not contain enough features to completely describe the physical process; it is also possible that it does not contain enough rows. However, a key part of the job is working with real data and doing what you can with what you have. We often must present results with caveats, or sometimes no results at all. High importance is given to the problem-solving process and modelling fundamentals you display.

Details

The system given as an example has this specific configuration:

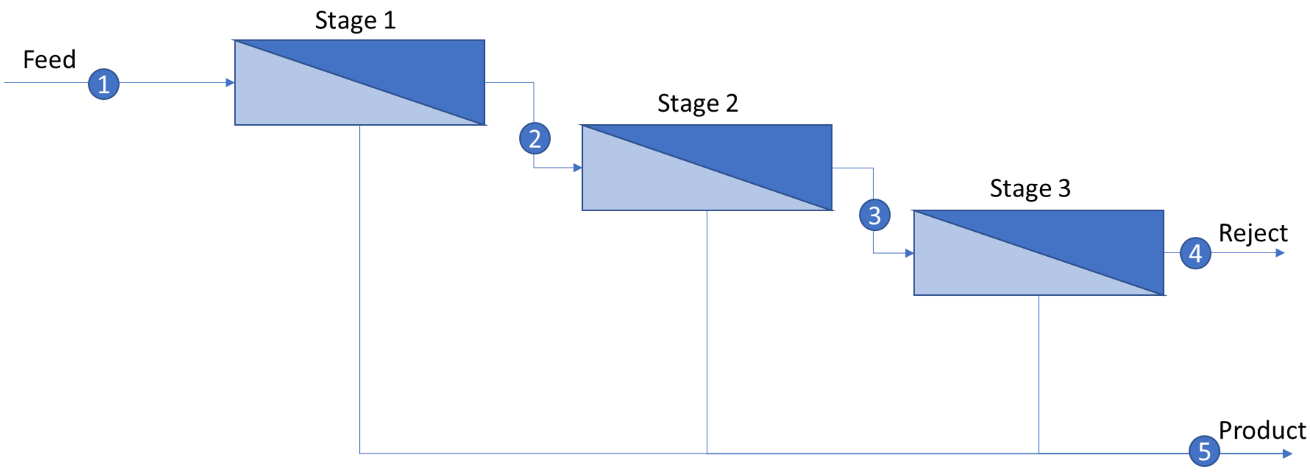

Figure 3: Three stage RO membrane train. The reject outlet of one stage is sent to the feed input of the next. The train does not contain full intermediate sensors due to the cost of installing sensors at high pressure.

This is presented to fully explain the physical system that the data represents. However, looking at the above system and the lack of instrumentation, it is only of limited use to consider the system like this. We must treat parts of it as Fig (2), a higher-level black box.

Features

The specific configuration and sensors available are given below with reference to locations in the figure above and with reference to columns in the file 'df_released.csv'.

Input Sensor Variables (features)

Name	**Type**	**Notes**	**Location**	**Unit**
feed_flow	Sensor		1	m3/h
product_flow	Sensor		5	m3/h
feed_concentration	Sensor		1	mg/L
product_pressure	Sensor		5	bar
feed_temperature	Sensor		1	Celsius
feed_pH	Sensor	In theory, could affect degradation rate but not instantaneous operation. Sparser than other sensors	1	unitless
feed_turbidity	Sensor	In theory, could affect degradation rate but not instantaneous operation. Sparser than other sensors	1	NTU (Nephelometric Turbidity Unit)
Membrane Cleaning	Event	Not explicitly in dataset. Information contained in running_time_since_last_clean and salt_loading_since_last_clean	everywhere	N/A
Membrane Replacement	Event	Not explicitly in dataset. Information contained in running_time_since_last_replace and salt_loading_since_last_replace	everywhere	N/A

Input Calculated Variables (features or possible features)

Name	**Type**	**Notes**	**Location**	
running_time_since_clean	Calculated	Time since last cleaning. Cleaning events occur over a non-negligible period.	n/a	hours
running_time_since_replace	Calculated	Time since last replacement. Replacement events occur over a non-negligible period. No replacement events occur in the dataset	n/a	hours
salt_loading	Calculated	feed_flow * feed_concentration Better representation of a filter's wear than running hours	1	kg/h
salt_loading_since_clean	Calculated	Integrated salt loading since last cleaning. Cleaning	1	kg

		events occur over a non-negligible period.		
salt_loading_since_replace	Calculated	Integrated since last replacement. Replacement events occur over a non-negligible period. No replacement events occur in the dataset	1	kg
recovery	Calculated	product_flow / feed_flow Meaningful variable to engineers. If this crosses a critical threshold, degradation will in theory happen faster.	n/a	unitless ratio between 0 and 1

Output Sensors (targets)

To simplify the technical challenge, only model the bolded variable.

Name	**Type**	**Notes**	**Location**	
stage_1_feed_pressure	**Sensor**		1	bar
stage_2_feed_pressure	Sensor	Contains stage_1_feed_pressure. stage_2_feed_pressure = stage_1_feed_pressure – stage_1_pressure_drop	2	bar
stage_3_feed_pressure	Sensor	Contains stage_2_feed_pressure. stage_3_feed_pressure = stage_2_feed_pressure – stage_2_pressure_drop	3	bar
reject_pressure	Sensor	Contains stage_3_feed_pressure. reject_pressure = stage_3_feed_pressure – stage_3_pressure_drop	4	bar
product_concentration	Sensor		5	mg/L

Output Calculated Variables (alternate targets)

Pani Energy has used flow and mass transport models to calculate coefficients representing the instantaneous state of the membrane from both the inputs and the outputs. These parameters cannot be used as features for the outputs above because they require the outputs to be calculated and the problem becomes circular. For this same reason, outputs are not valid features for these variables. However, they provide an alternative option to the outputs. They may prove an easier target to model than the output sensors; if that is the case, the outputs can be calculated from these coefficients and the inputs using the previously mentioned physical models.

To simplify the technical challenge, only model the bolded variable.

Name	Type	Notes	Location
A	**Calculated**	**Representative of average membrane water flux**	Average of stage 1, 2, and 3
B	Calculated	Representative of average membrane salt flux	Average of stage 1, 2, and 3
k1	Calculated	Representative of average stage 1 pressure drop	Stage 1 (between 1 and 2)
k2	Calculated	Representative of average stage 1 pressure drop	Stage 2 (between 2 and 3)
k3	Calculated	Representative of average stage 1 pressure drop	Stage 3 (between 3 and 4)

Data

There are 14 months in this dataset. 12 months of data is contained within the attached csv. You are free to split into train/test/validation as you wish. The performance on the csv you have will be evaluated, as well as the data we have retained.

The dataset is raw and contains noise. If you choose to remove noise or otherwise aggregate the data, please include the details in your modelling summary.

The dataset contains raw sensors. It also contains pre-calculated engineering features and information which process experts have identified as being useful in theory, but that usefulness has not been confirmed in this dataset. Please contain details of feature selection as well as any potential additional engineered features in your dataset.