Data Mining to Predict Operational Outcome

Sembcorp Industries Ltd

40.000 Group 6: Basil Yap Ji Tsing, Hilda Thian Qing Wei, Liu Yawen, Sim Yan Jie, Yuan Nan

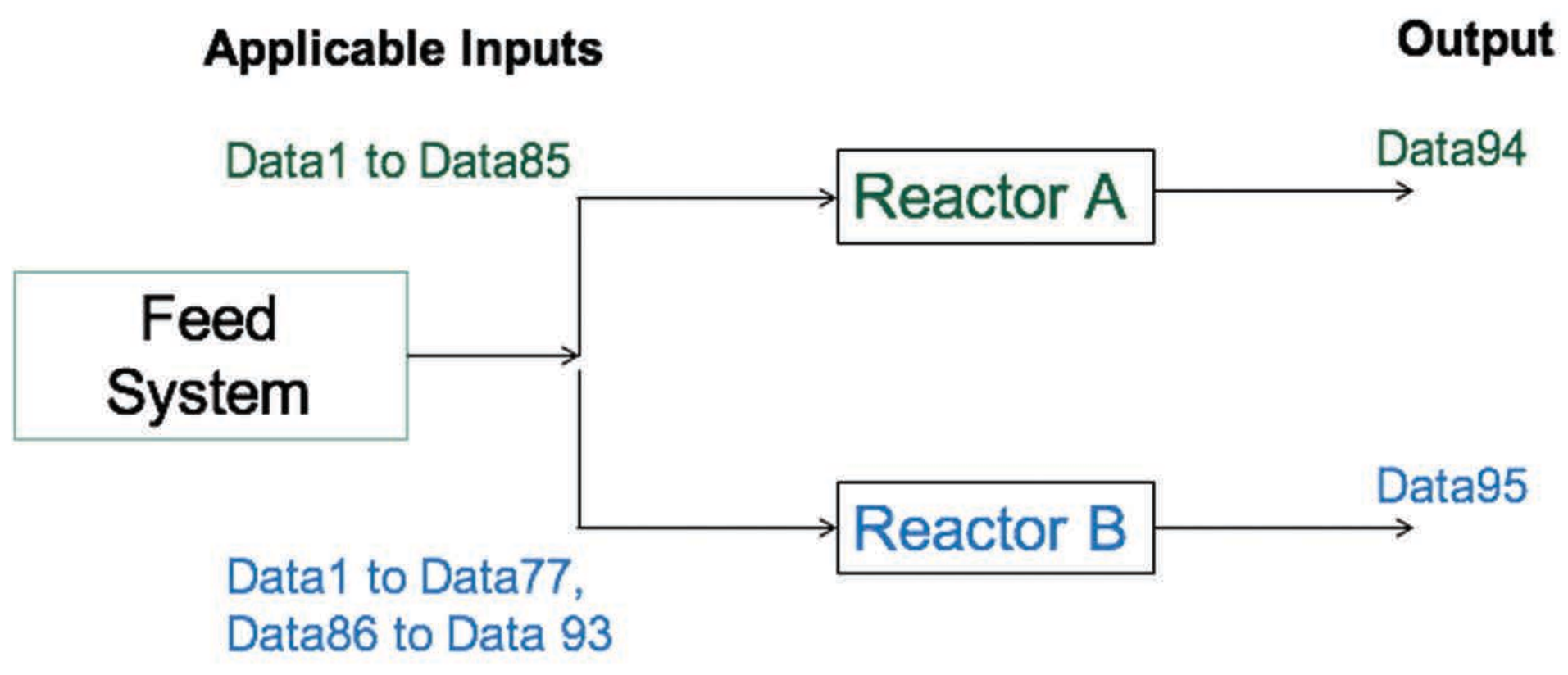
ESD ENGINEERING SYSTEMS & DESIGN

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN
Established in collaboration with MIT

Introduction

Microorganisms are used to decompose the contaminants in the biological wastewater treatment processes. However, they are sensitive to various factors such as wastewater composition and reactor conditions.

This project aims to identify the key predictors affecting the treatment plants' performances via data mining and analysis so as to help Sembcorp to improve their operational outcome.



3 years worth of data are provided for 2 reactors, A and B. Each data set consists of observations for 93 predictors and 2 ouput variables.

Challenges

1. Data provided was not in same format as Sembcorp's definition of a day.
2. Different predictors had different number of observations recorded a day.
3. Anomalous values are indistinguishable because data provided was purely numerical without physical meaning.
4. Several measures available in accessing predictive performance of models.

Methodology

1. Data Cleaning

Each predictor's time value is reformatted according to Sembcorp's definition of a day. A time-series plot of each predictor is then plotted to observe their behaviors. The problem of numerous 0 values in some predictors was resolved through several hypotheses with varying percentage tolerance to the 0 values. Aggregation of data for predictors with multiple entries a day was done by averaging all observations of the predictor into a single daily value.

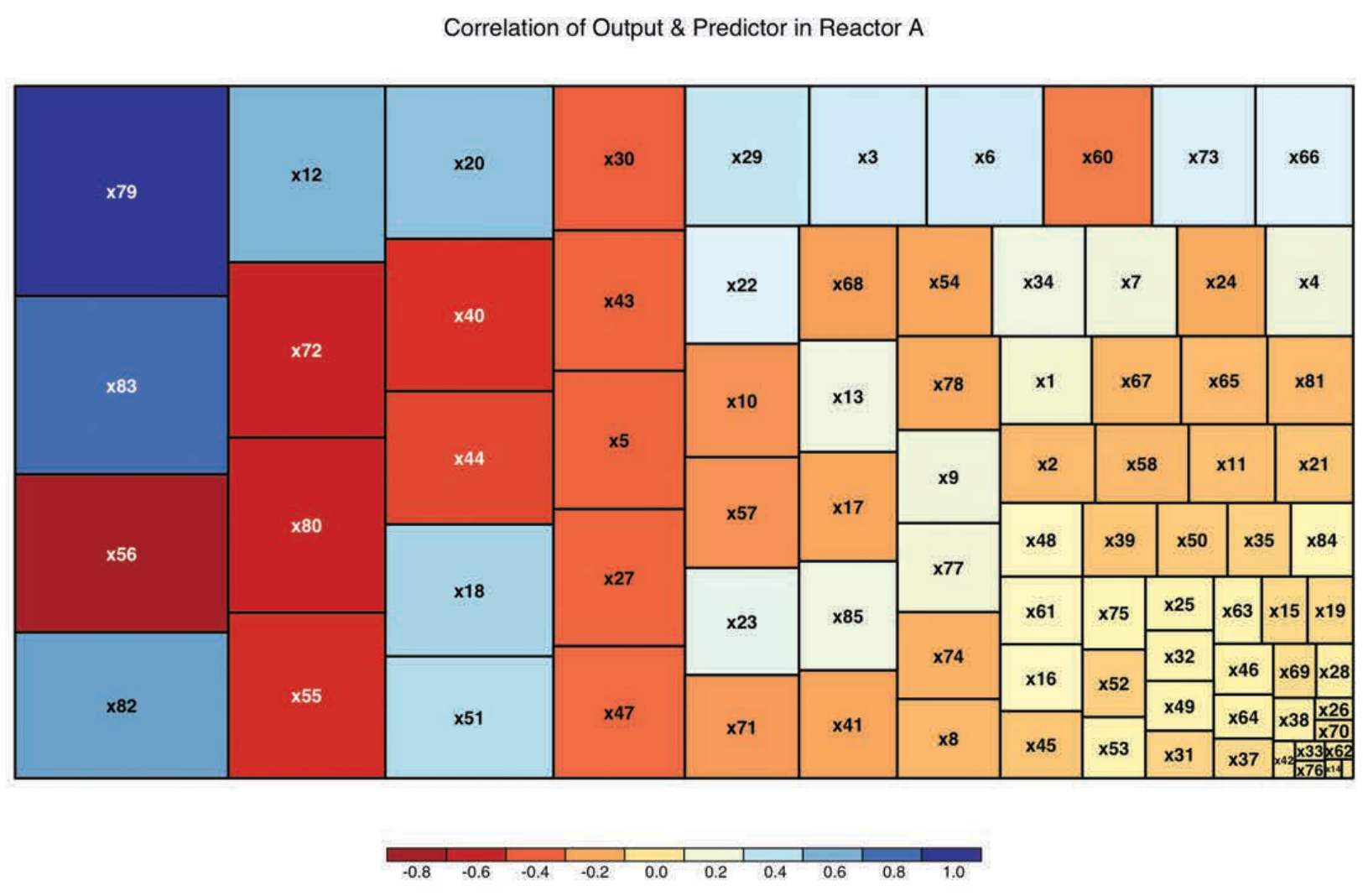
2. Matrix Construction

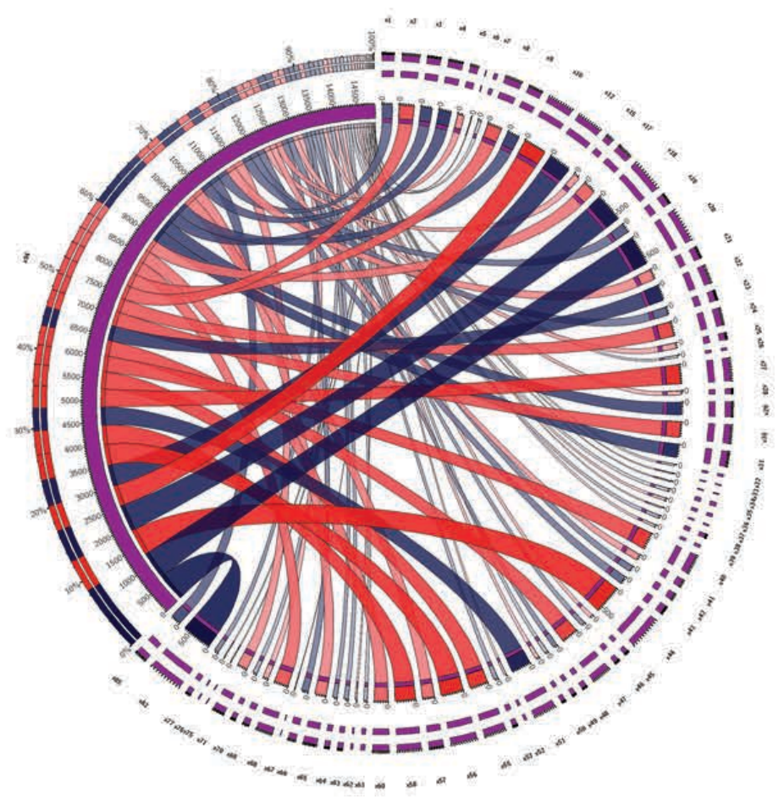
2 matrices containing the relevant predictors and output are constructed for each reactor. The data of every predictor is matched with the output observation according to day that output was generated. Date, output variable and predictors make up the column of the matrix. For the initial few hypotheses, with regard to the handling of null values, only complete entries fulfilling all columns of the matrix were used.

3. Correlation Study

To reduce the number of the predictors before constructing the model, the correlations between variables were studied. To start with, a correlation matrix was constructed in Excel, showing the correlation value between every pair of variables. To better visualise the relationship of variables, different softwares were used:

(1) A tree map in R illustrating the correlation value of each predictor with respect to the output variable. The size of the block represents the absolute correlation value while the color shows if it is positively or negatively correlated.





(2) A Circos chart presents the correlation between each variable and the output in a circle with bands connecting from each predictor to the output variable. The wider the band, the higher the correlation.

4. Model Calibration

To generate a more robust model, multiple models of different combinations of predictors were formulated and then compared based on their predictive performance. There were 3 approaches in formulation:

To generate a more robust model, multiple models of different combinations of predictors were formulated and then compared based on their predictive performance. There were 3 approaches in formulation:

(1) Only predictors with at least 280 observation were used to generate a model. The number of predictors used in modelling is then reduced based on their significance level, keeping only those with more than 95% confidence.

(2) Only predictors with pairwise correlation with absolute value of more than 0.3 were used in modelling. This range is further reduced by increasing the threshold correlation value to 0.35 and 0.4.

(3) Variables were selected using the stepAIC() in R.

5. Testing and Improvement

To validate the models, a portion of the data was set aside. A random selection of 6 out of 36 months worth of data was used as testing data, leaving the remaining 30 months as training data. 5 different sets of training and testing data was generated for each model.

Akaike information criterion (AIC), Root Mean Squared Error (RMSE), Adjusted R Squared (Adj R^2), Average Relative Error (ARE) values were generated to compare the predictive performance of each model. Cross-validation is conducted on the training data for every model to check their performances against each other.

$$\text{AIC} = 2k - 2\ln(L)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2{}_{adj} = R^2 - (\frac{p}{n-p-1})(1 - R^2)$$

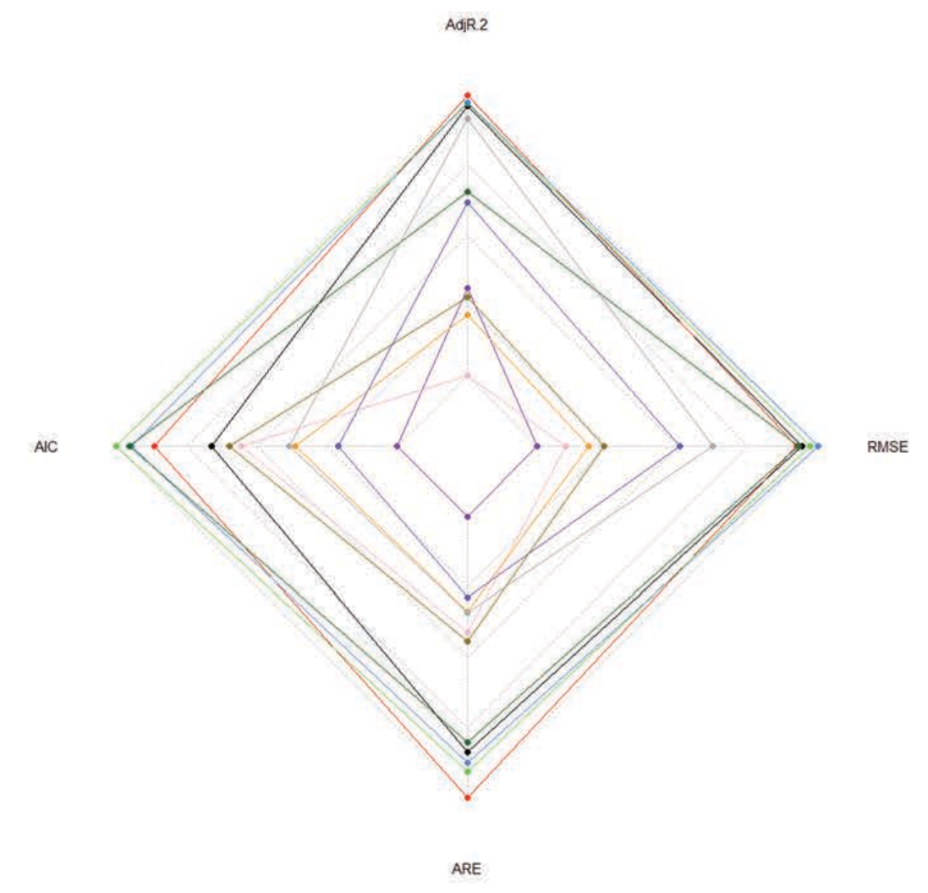
$$\boldsymbol{ARE = \sum_i^n (abs(\frac{y_i - \hat{y}_i}{y_i}))/n}$$

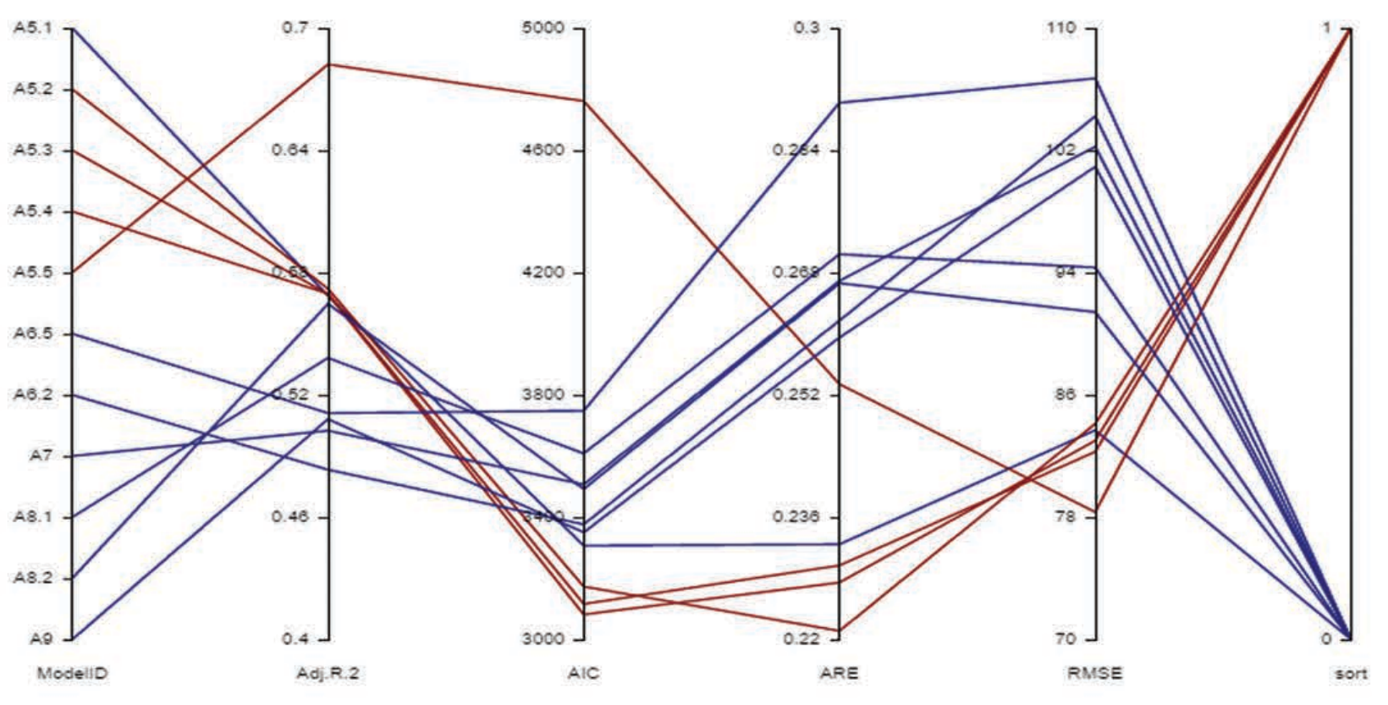
A time-series algorithm is then formulated as it is more appropriate for modelling a dynamic process.

Results and Analysis

The best performing models should give the minimum AIC, RMSE, ARE and maximum Adj R^2 values. Across the 5 different sets of training and testing data for each model, the average values of AIC, RMSE, Adj R^2 and ARE were calculated, which produced a multi-objective optimization problem. Pareto sorting is used to obtain the best performing models. The number of times each predictor appears in the pareto-efficient models was then determined to find common predictors which Sembcorp might be interested to look further into.

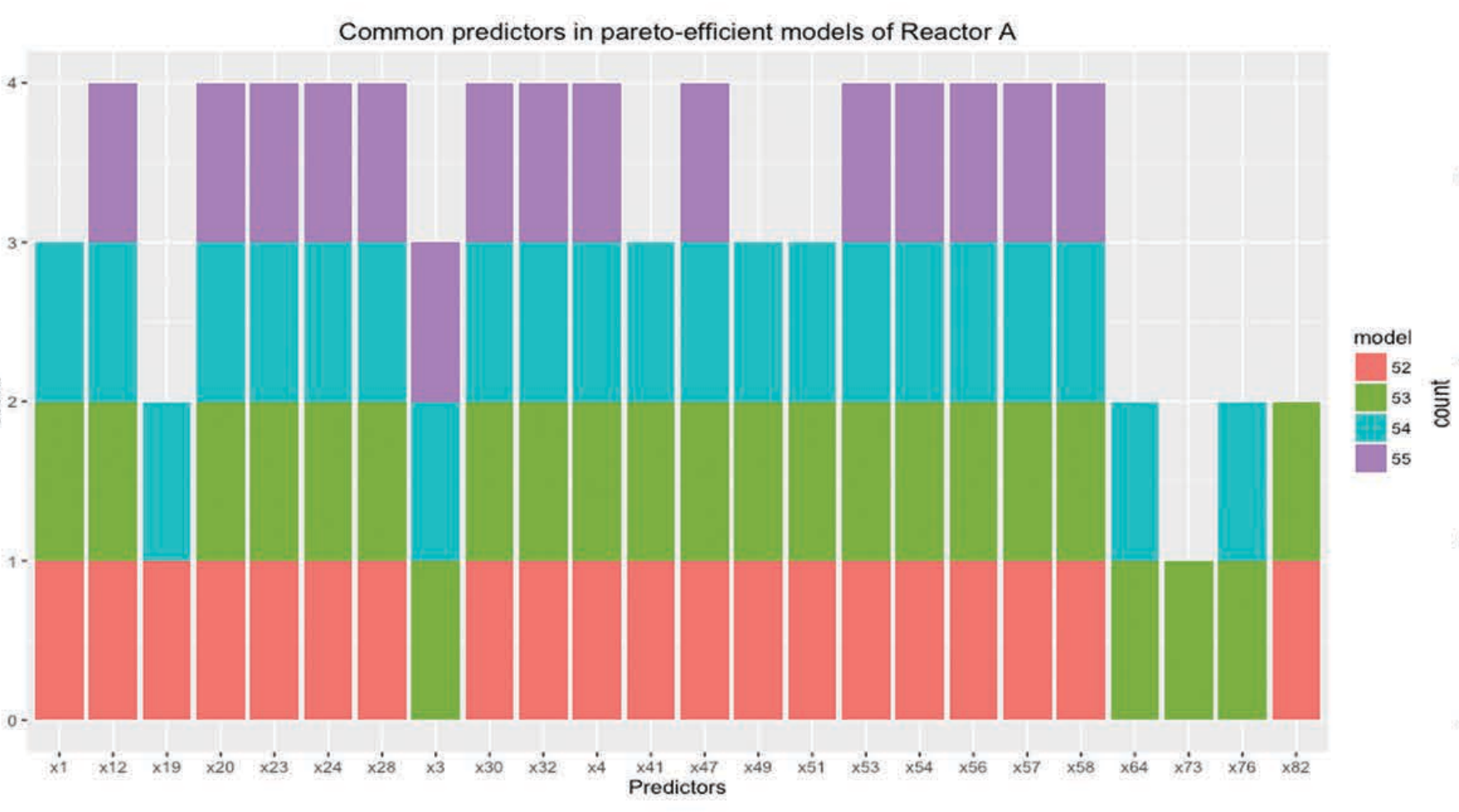
In this radar chart plotted in R , with ModelA55 removed, the better performing models are those which have points lying closer to the corners of the diamond.

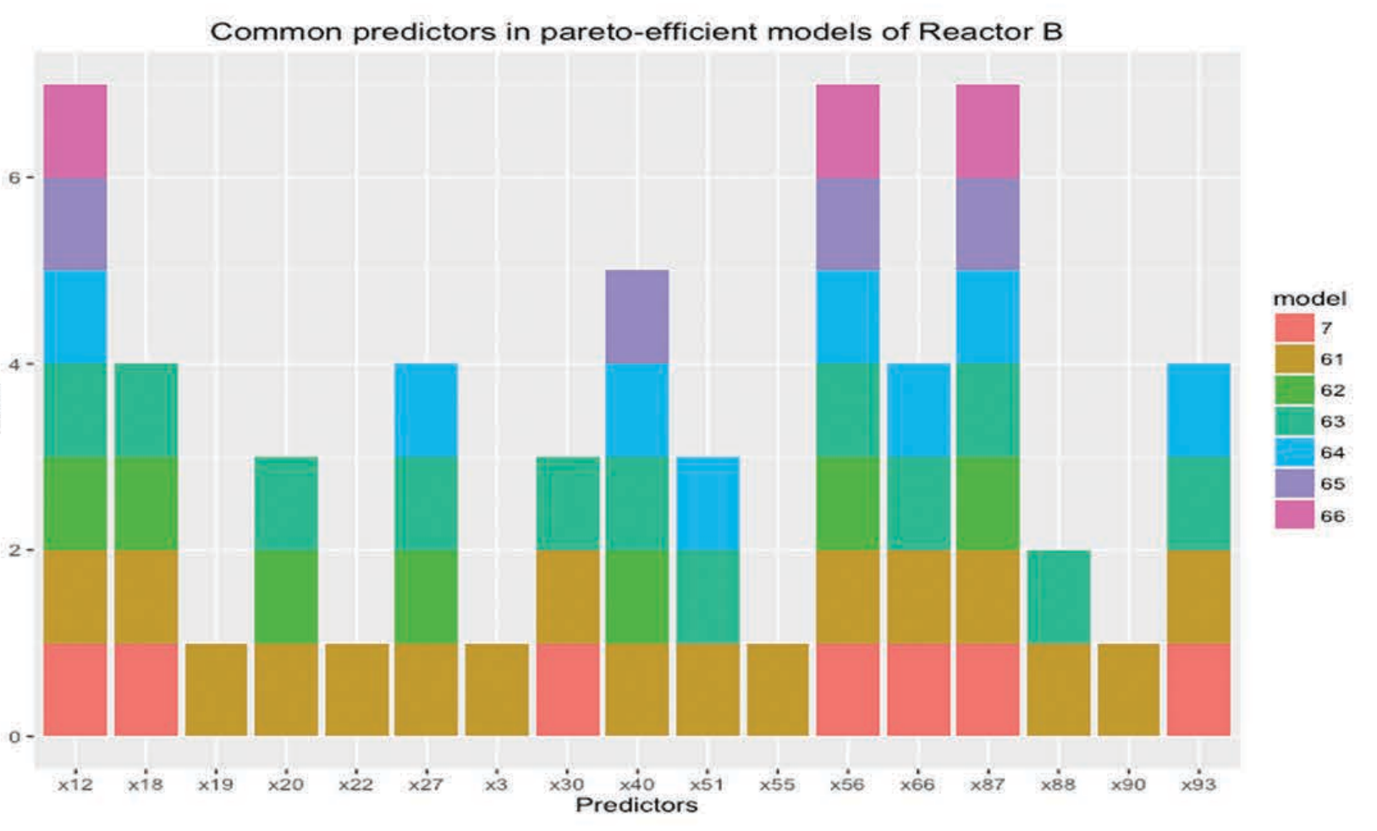




In this parallel coordinate plot from DiscoveryDV, the pareto-efficient models are colored red.

In the two histogram below, the number of times each predictor appears in our pareto-efficient models is counted.





Conclusion and Future Consideration

Sembcorp is the Developer, Owner and Operator of Water and Wastewater Treatment Facilities with capacity of 8.8 million m3/day and Sembcorp Utilities across the globe has a turnover of 4,200 million SGD in 2015 [1]. This shows the importance of managing the process efficiently as the scale of the impact is very large, not just in Singapore, but internationally.

Therefore, having a statistical model allows Sembcorp to predict and analyse the performance of their processes beforehand, which means they can intervene and make adjustments in advance and reduce the occurrence of process abnormalities. This also helps to improve the operation of Sembcorp's processes and cuts down on unnecessary expenses and downtime relating to such abnormalities.

Softwares Used

[1] http://www.sembcorp.com/en/media/469912/sembcorp-industries_facts-figures-2015.pdf

Acknowledgement: Dr. Stefano Galelli, Ms. Macauley Kenney, Dr. Peter Jackson, Dr. Ni Wangdong, Mr. Aaron Yuen