Chemical Plant Prediction and Control

Soham Inamdar(210100149)

Anannay Jain(210110021)

Description of the Problem

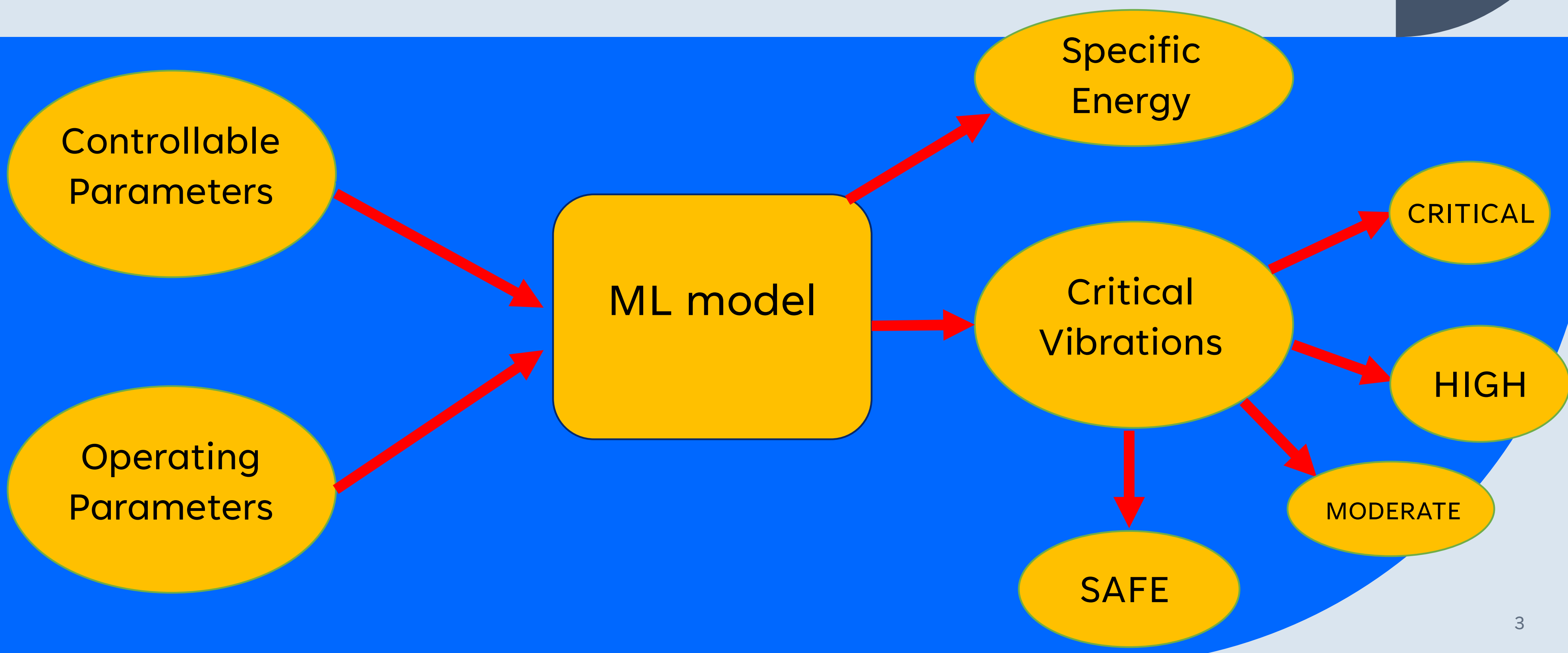
We were given the daily averaged values of observations logged by a data acquisition system at a chemical processing plant. There were three primary objectives of this problem:

•Create separate ML models to predict vibration levels of the **critical parameters(c51, c52, c53, c54)** using only the **controllable parameters** and using all parameters

•Identification of levels which are defined as **Safe**(<5), **Moderate**(5-10), **High**(10-20) and **Critical**(>20).

•Create a prediction model to study which parameters significantly affect the **free energy(c241).**

Description of the Problem



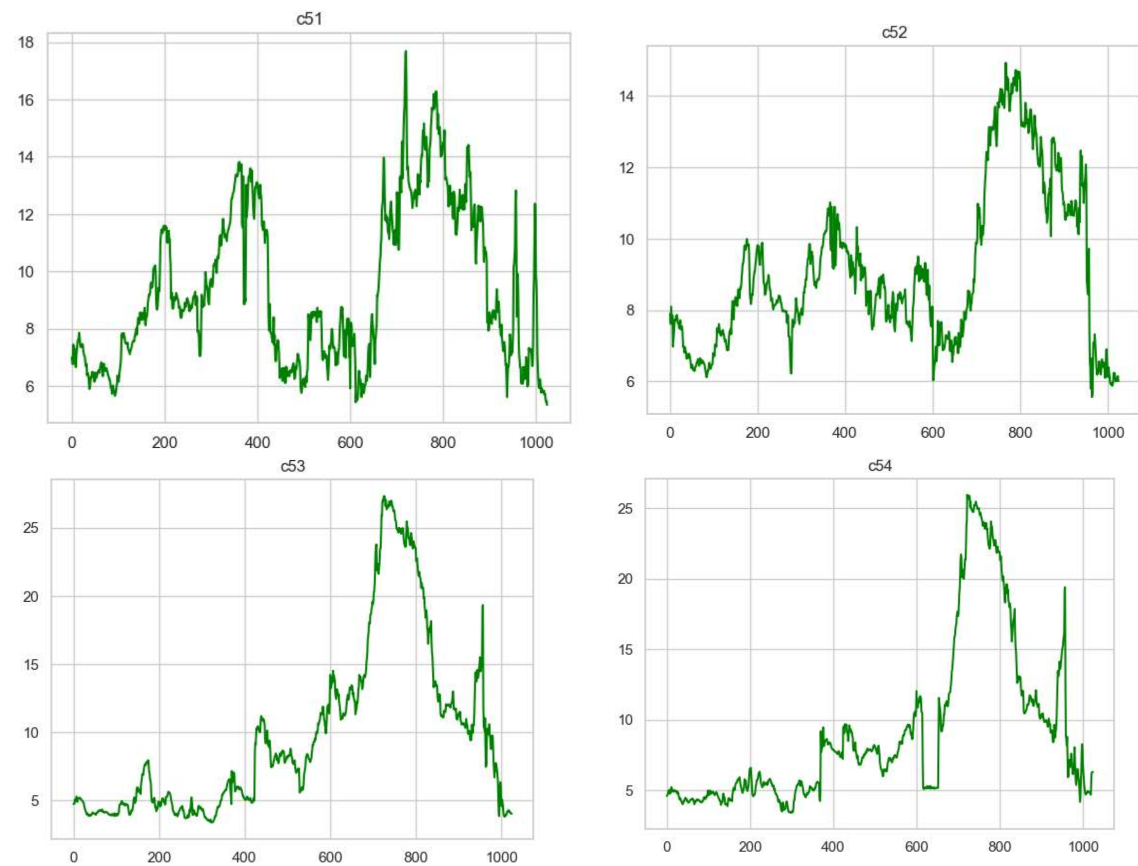
Steps Involved

Analyse Data → Handling Missing Data → Removing obsolete columns → Replace Outliers → Create and Improve ML Model

Data Analysis

- Due to the large number of columns, it was difficult to use visual inspection, hence we have used statistical methods.
- We visualized the target columns to get an idea of the correctness/validity of the data given to us and the model we need to use.

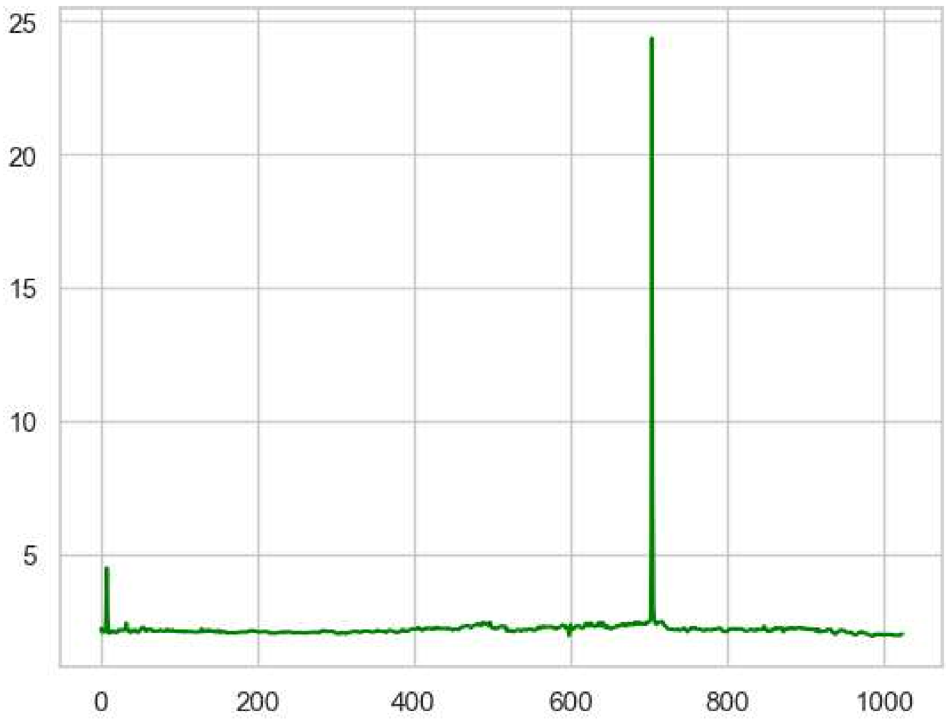
Plots of target columns



Insights from these plots:

- Non-linear variation of target columns.
- Presence of Outliers

Plot of Specific Energy c241



Insights:

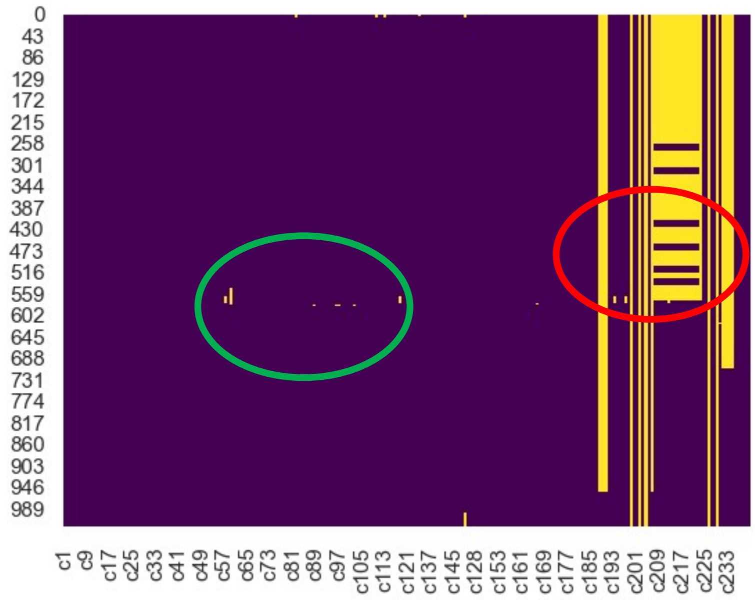
- Free energy data is more or less saturated around a single value.
- Data consists of very few outliers.

Problems in the Data

- Arbitrary characters such as “#REF”, “#VALUE” were present in the original data, we replaced these values with NaN for the time being.
- Missing values in the dataset
- Obsolete columns which have a constant value throughout
- Outliers
- Multi-collinearity between various columns

Missing Values

The following plot depicts the distribution of missing values across the dataset:

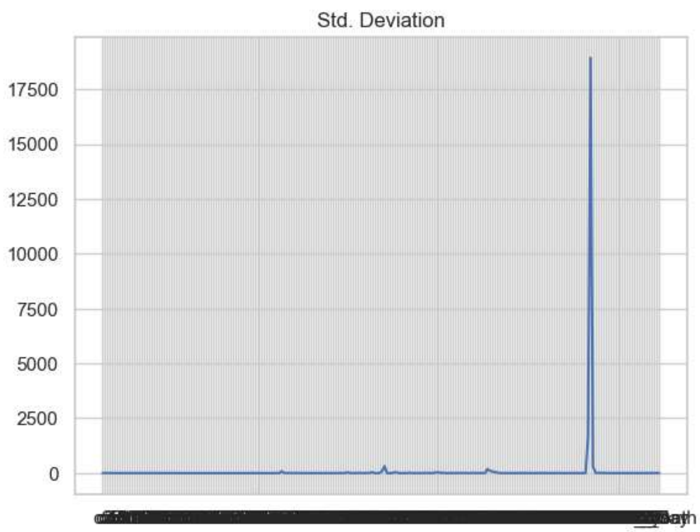


Insights from this plot:

- The circled(red) columns are those which are mostly empty and hence, useless to us and hence, we dropped them.
- The green circle indicates columns where only a few values are missing. We filled in these values using **linear interpolation**.

Obsolete Columns

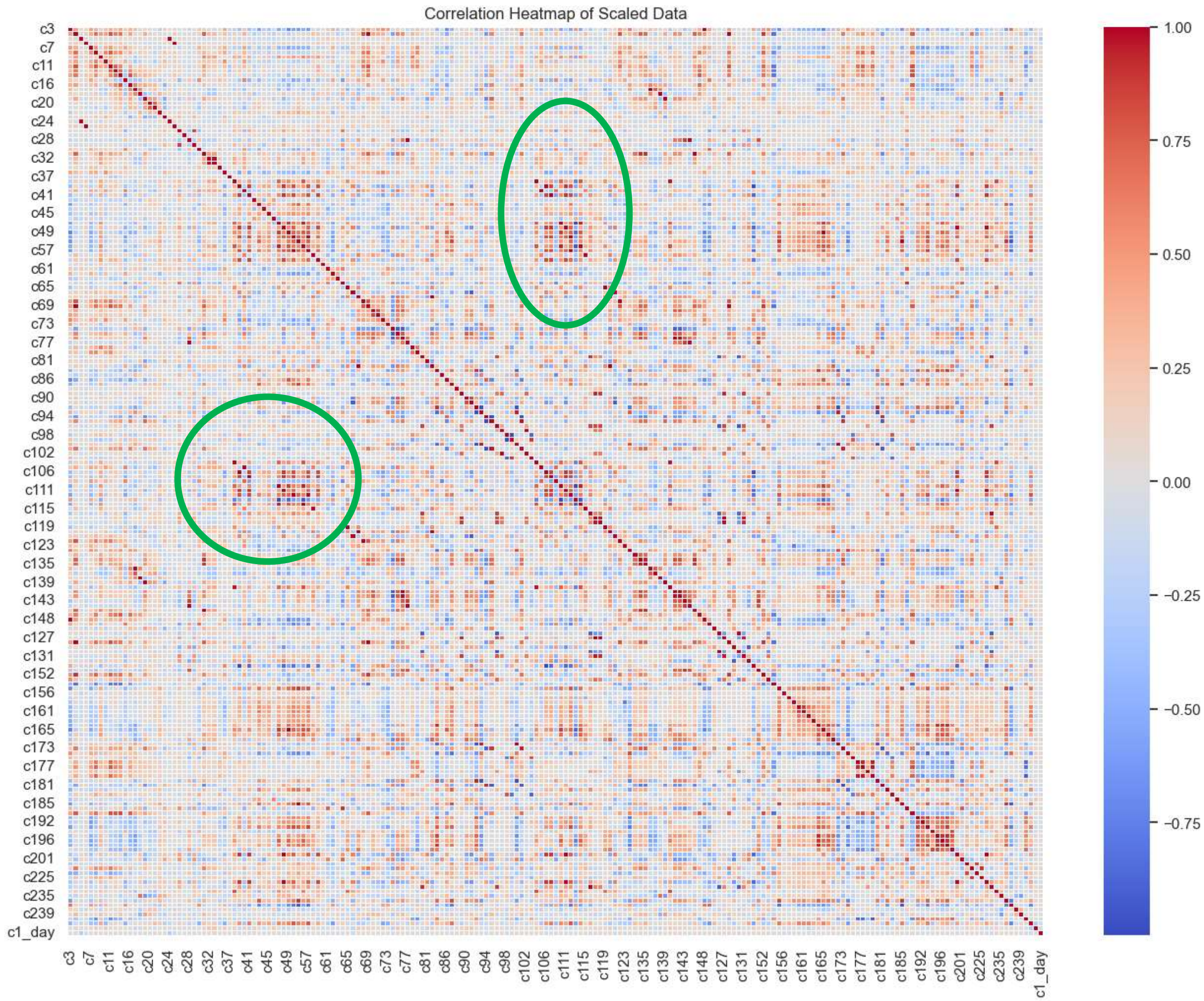
- Some columns in the dataset are constant throughout, hence these columns are of no use to us. We identified these columns by calculating the standard deviation:



We dropped the columns with a standard deviation equal to zero to counter this issue.

Dealing with Multicollinearity

- Multicollinearity means multiple columns are strongly correlated. To visualize the correlation, we plotted the **correlation matrix**.
- However, before computing the correlation matrix, we normalized the data in the range [0,1] using min-max scaling.



- The circled columns are two examples of highly correlated columns.
- We need to eliminate such columns.
- However, we cannot do it by visual inspection.
- Hence, we need to do it using a mask that filters and eliminates columns having correlation greater than 0.95.

Solving the Problem

We decided to build 3 ML models for the following three parts:

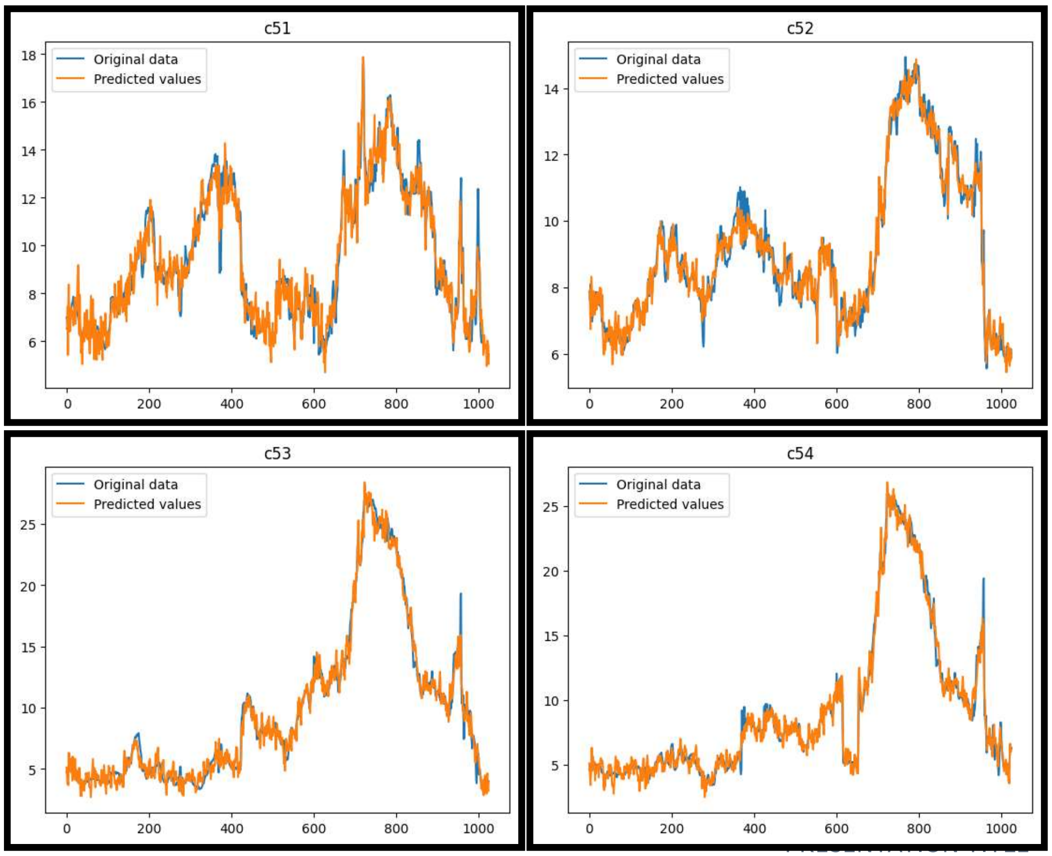
- Part 1: Prediction of Critical Vibrations using all parameters
- Part 2: Prediction of Critical Vibrations using controllable parameters
- Part 3: Analysis of Specific Energy Data(c241)

PART 1: Prediction of critical vibrations using all parameters

ML Model

- We have used multiple linear regression(MLR) to predict the values of the critical parameters and that of the free energy.
- We have used an iterative procedure to drop columns that have a **p-value less than 0.05 for** feature engineering one at a time.
- The results obtained from this model can be seen in the upcoming slides.

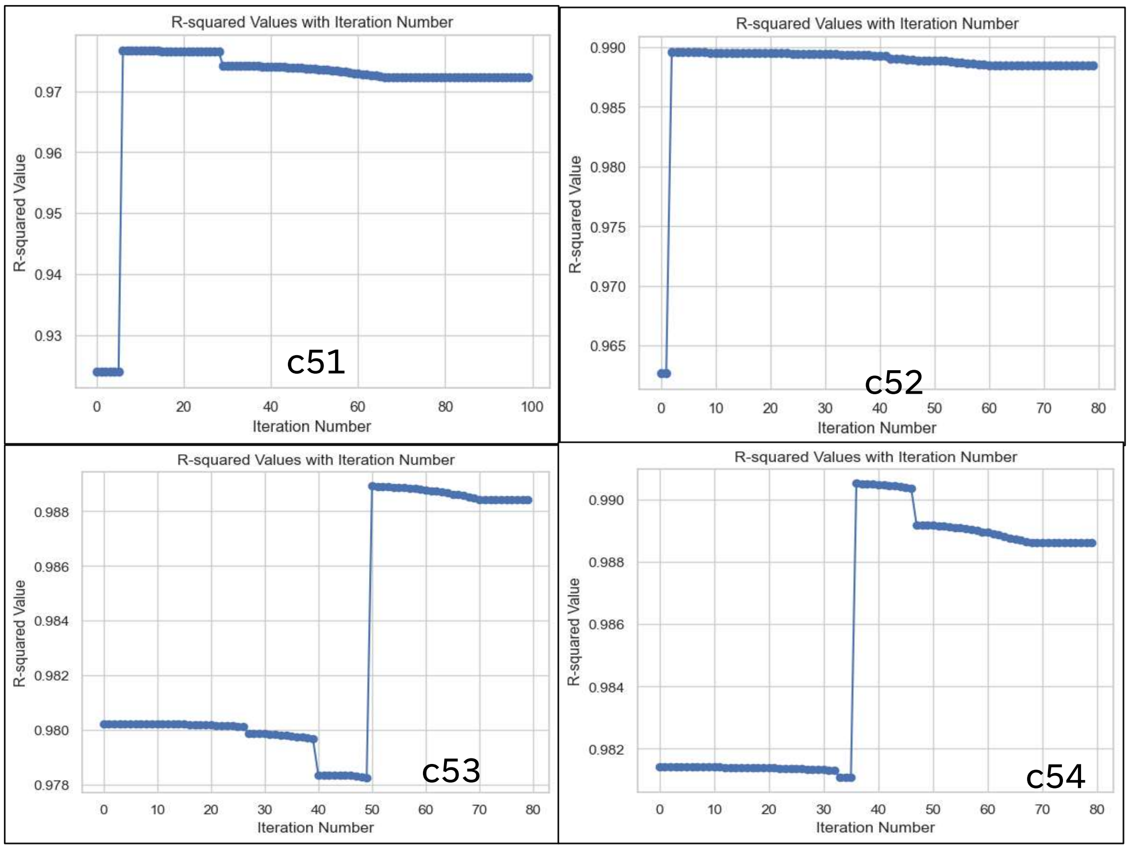
Plots of predicted data and original data



After looking at the plots of data predicted, we can conclude the following things:

- The model does a very good job at replacing outliers from the training data.

Part 1: Plots of R-square value vs number of iterations



Plots for Iterative Feature Elimination

It can be seen from these plots that **R-square** drastically increases after a few iterations which is clearly a good thing

Part 1: Prediction of critical vibrations using all parameters

- Following is a tabulated summary of the results of part 1:

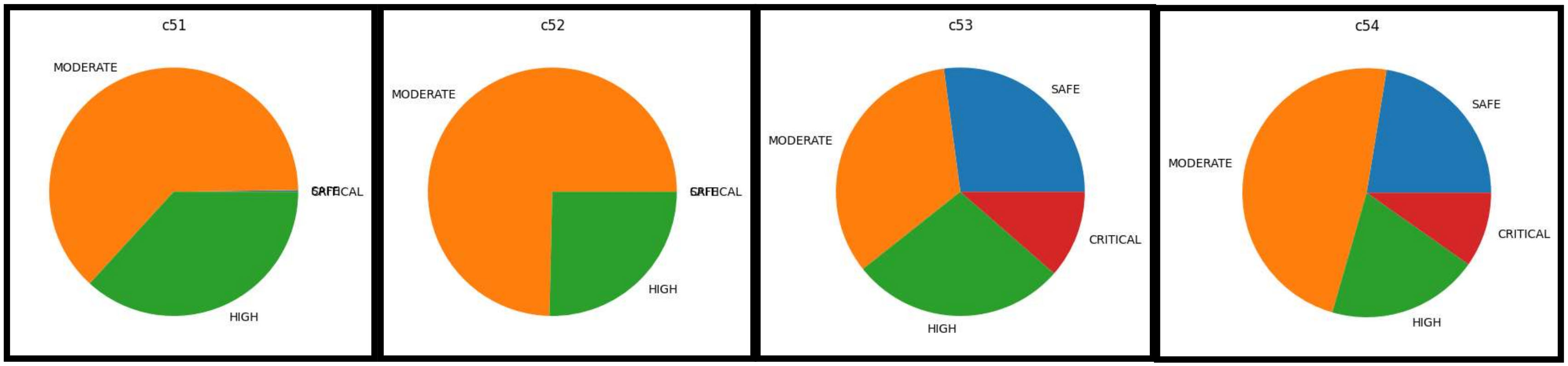
Parameter	R-Square	Train-MSE	Test-MSE
c51	**0.972**	1.24	3.58
c52	**0.988**	1.205	1.898
c53	**0.988**	0.61	1.38
c54	**0.988**	0.461	1.258

Part 1: Prediction of critical vibrations using all parameters

- Following is a table depicting the classification of “vibration values”:

Parameter	SAFE	MODERATE	HIGH	CRITICAL
c51	2	646	377	0
c52	0	765	260	0
c53	278	344	286	117
c54	230	493	202	100

Part 1: Prediction of critical vibrations using all parameters



The distribution of SAFE, MODERATE, HIGH, CRITICAL values can be seen from the above pie charts.

PART 2: Prediction of critical vibrations using controllable parameters

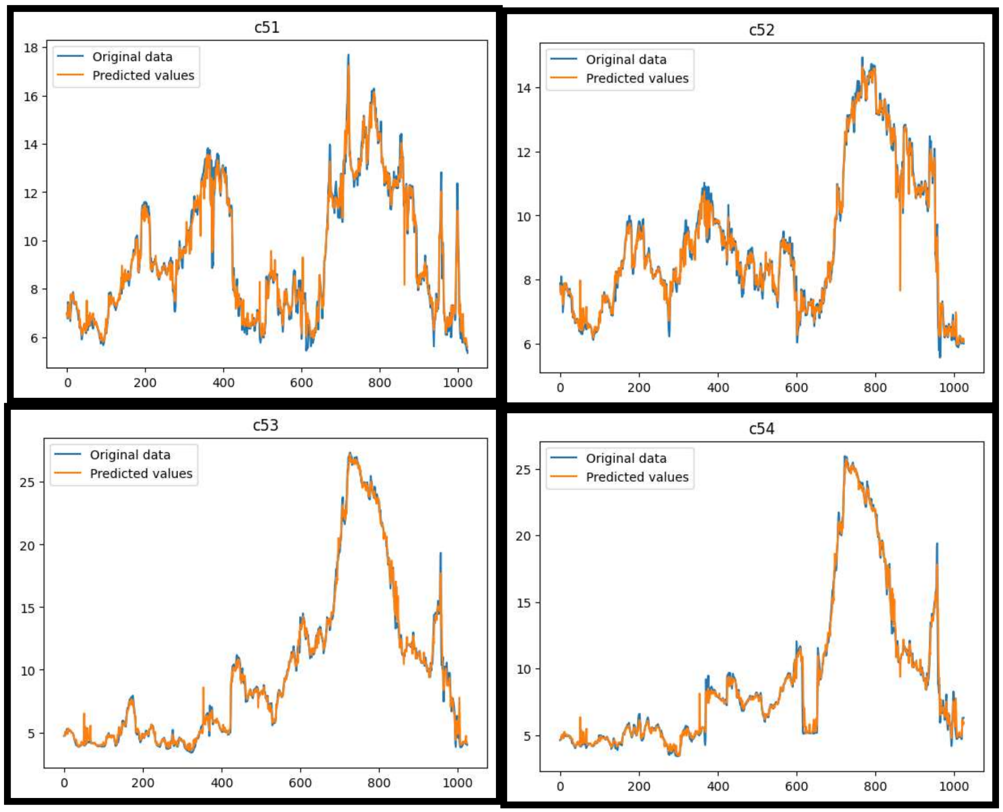
ML Model

- We have used a **random forest regressor** to solve this part of the problem.
- Reasons for choosing this model:
  1. Relatively Low number of features which reduces chances of overfitting when using Random Forest
  2. High prediction accuracy of Random Forest
  3. Random Forest provides us with **feature importance**

  Following are the results obtained using the ML model:

Train MSE	0.0003
Test MSE	0.002

Plots of predicted values and original values



After looking at the plots of data predicted, we can conclude the following things:

- The model performs better than that in Part 1.
- Reducing features is an important factor in improving model accuracy.

Most Important Parameters

- Following are the most important controllable parameters to reduce the vibrations:

Parameter	Importance
c155	0.653
c161	0.086
c143	0.061
c39	0.034
c158	0.033

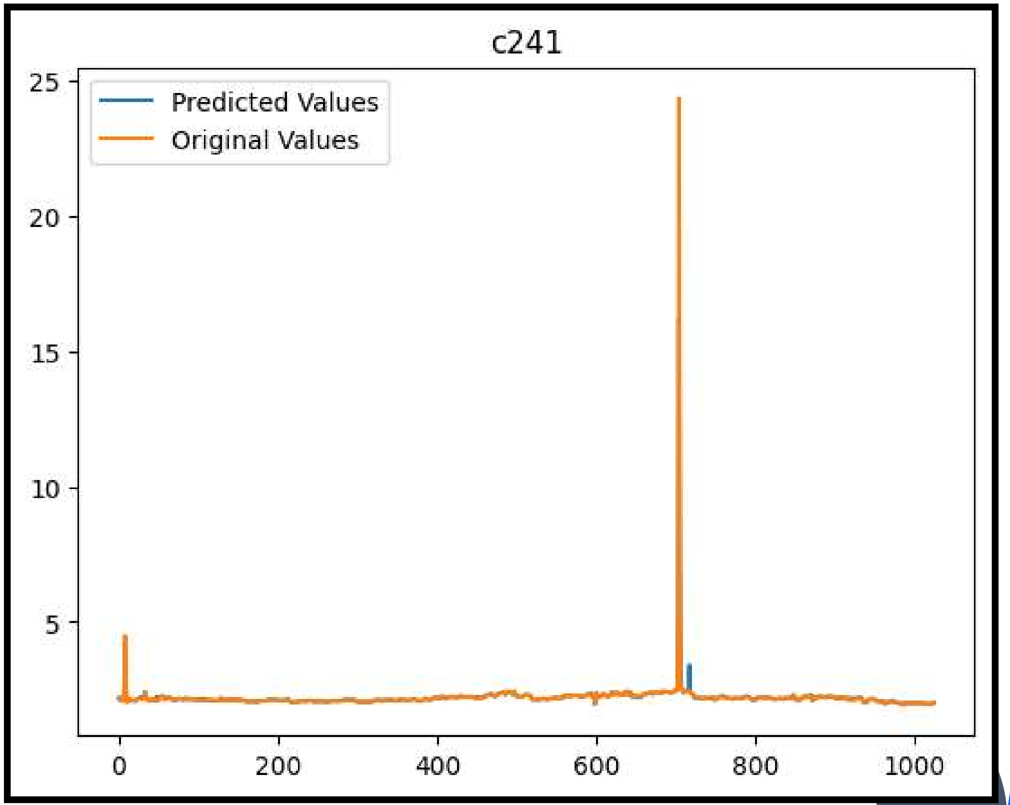
PART 3: Prediction and Analysis of Specific Energy

ML Model

- We have used a **random forest regressor** to solve this part of the problem.
- Reasons for choosing this model:
  1. High prediction accuracy of Random Forest
  2. Random Forest provides us with **feature importance**
- As we can see, the results obtained by using all the parameters are very good.

Train MSE	**1e-5**
Test MSE	**1e-6**

Plots of predicted and original specific energy data



Most Important Parameters

Following are the top 5 **most important** parameters to predict free energy: **(Model was trained using all parameters)**

Parameter	Importance
c193	0.107
c192	0.097
c151	0.088
c99	0.058
c103	0.056

Most Important Parameters

- Now, to find the minimum number of parameters, we started by using the **top 5 most important parameters** in our model.
- Following are the results obtained:

Train MSE	**1e-4**
Test MSE	**1e-5**

- Hence, we can see that using only the top 5 parameters provides us with acceptable results.

Why our approach is a good one?

- High R-square values(close to 0.98) and small difference between train and test MSE.
- We have removed columns where more than 50% of the data is missing, it would have been impossible to accurately interpolate data for these columns.
- We have removed constant columns(0 standard deviation) which results in further **dimensionality reduction.**
- We are using an iterative process for feature elimination rather than eliminating all the seemingly irrelevant features in one go.

Insights

- The data predicted by the model can be applied in training **alerting systems** to prevent damage to the chemical plant.
- The vibrations can be controlled using the top 5 parameters listed above.
- Energy conservation can be achieved by tuning the parameters specified above. (for specific energy)
- One can also use **online learning** to monitor and control the chemical plant as the problem here is quite similar to the **prediction and control problem.**

Challenges Faced

- The data provided had arbitrary string values such as #REF!, #VALUE!. We solved this problem by inspecting the Excel sheet and changing such values to NULL for the time being.
- We used Variance Inflation Factor(VIF) to detect multicollinearity, but it gave bad results. Hence, we did not use it in our final model.
- Initially, we used MLR for Part 2(using only controllable parameters), but it did not yield satisfactory results. So, we used a different model, Random Forest Regressor instead.

Key Achievements

- Reduced the number of features from 240(given) to 60-70(in ML model) using various Data Analysis and Interpolation techniques.
- Fully automated iterative feature-elimination in Multiple Linear Regression(MLR).
- Used different models(MLR and Random Forest) based on the difference in nature of the problem.

Thank you