E-11
DS-203

TEAM

EXPLOREDATA

Our Team

Yash Mehta	22B1504
Nishlesh Goel	22B0017
Tarshit Sehgal	22B0077

EXPLOREDATA

Problem Statement

1.Creating a **MLR model** to predict vibrations for a **chemical operating plant.** These vibrations need to be closely monitored to keep them under control. The MLR model thus created would be used to raise alerts and alarms if/when they reach HIGH and CRITICAL levels respectively

2.Identify the controllable parameters which contribute maximum towards those vibrations, so that these can can be altered to reduce the vibrations

3.Create a model to predict specific energy on all the given parameters

Before we jump into soving the problem, we need to come up with an planned approach to proceed with the solving this problem

Our plan of action

- Data Cleaning
- MLR Model creation
- Backward Elimination
- Comparing the predicted safety levels of vibrations with the actual ones
- Identifying the controllable parameters which contribute maximum towars the vibration

EXPLOREDATA

1. Data Cleaning

We began with making three new columns by splitting the date into day, month and year so that we can use these additional parameters in the regression; just in case the output depends on any of the these then these will be included, or else these will be removed in the backward elimination performed later

S	HT	HU	HV
	day_column	month_column	year_column
4083	1	9	2018
3879	2	9	2018
8296	3	9	2018
8927	4	9	2018
6676	5	9	2018
2196	6	9	2018
7302	7	9	2018
2217	8	9	2018
9179	9	9	2018
0244	10	9	2018
4788	11	9	2018
4667	12	9	2018
6899	13	9	2018
9532	14	9	2018
1304	15	9	2018
4098	16	9	2018
9755	17	9	2018
4907	18	9	2018
1752	19	9	2018
8679	20	9	2018
6208	21	9	2018
3681	22	9	2018
1589	23	9	2018
6438	24	9	2018
4196	25	9	2018
8417	26	9	2018
4852	27	9	2018
2114	28	9	2018
3677	29		

Continued...

All those cells with #REF, #VALUE! and #NA we replaced by empty cells.

Made a new csv after these alterations named Chemical_plant.csv

DE	DF	DG	DH	DI	
)	c110	c111	c112	c113	c1
0.072	#REF!	301.0413	792.6062	#REF!	3
.6493	#REF!	301.3578	792.4817	#REF!	3
.7706	#REF!	304.1794	790.1263	#REF!	
.7196	#REF!	297.9304	794.0051	#REF!	3
.0775	#REF!	294.428	793.9269	#REF!	
.2505	#REF!	296.2236	795.5415	#REF!	3
.0933	#REF!	295.7032	795.5313	#REF!	3
.0319	298	302.2139	795.6536	-0.00011	3

DE	DF	DG	DH	DI	DJ	
9	c110	c111	c112	c113	c114	c1
00.072		301.0413	792.6062		303.6975	3
9.6493		301.3578	792.4817		303.4686	3
9.7706		304.1794	790.1263		303.41	3
7.7196		297.9304	794.0051		302.9316	3
7.0775		294.428	793.9269		302.718	3
8.2505		296.2236	795.5415		302.6212	3
8.0933		295.7032	795.5313		302.3302	3
8.0319	298	302.2139	795.6536	-0.00011	302.1605	3
8.6795	298	303.7463	791.4832	-0.00238	302.2078	
9.1156	298	304.4502	787.941	-0.00393	302.6943	3

EXPLOREDATA

Continued...

On plotting each column as a scatter plot, we see the rest of the values of that column are almost similar (with negligible variance)

Therefore, we filled all the empty cells of each column with the mean of the rest of the data in that column.

c109	c110	c111	c112	c113	c114
[illegible]	298	[illegible]	[illegible]	-0.01256	[illegible]
[illegible]	298	[illegible]	[illegible]	-0.01256	[illegible]
[illegible]	298	[illegible]	[illegible]	-0.01256	[illegible]
[illegible]	298	[illegible]	[illegible]	-0.01256	[illegible]
[illegible]	298	[illegible]	[illegible]	-0.01256	[illegible]
[illegible]	298	[illegible]	[illegible]	-0.01256	[illegible]
[illegible]	298	[illegible]	[illegible]	-0.01256	[illegible]
[illegible]	298	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	298	[illegible]	791.4832	[illegible]	[illegible]
[illegible]	[illegible]	304.4502	787.941	-0.00393	[illegible]
298.2915	[illegible]	299.9357	796.9341	-0.001	[illegible]

Continued...

Some columns were completely empty (like c229, c226, c199, c202, c204) and were removed

Other columns like c188, c189 and c190 had ‘#REF!’ error for more than 60% of data so they were also removed

Continued...

Columns in range 207, 222 initially had a lot of cells with N/A error.. These were replaced with empty cells first, followed by replacing with the mean.

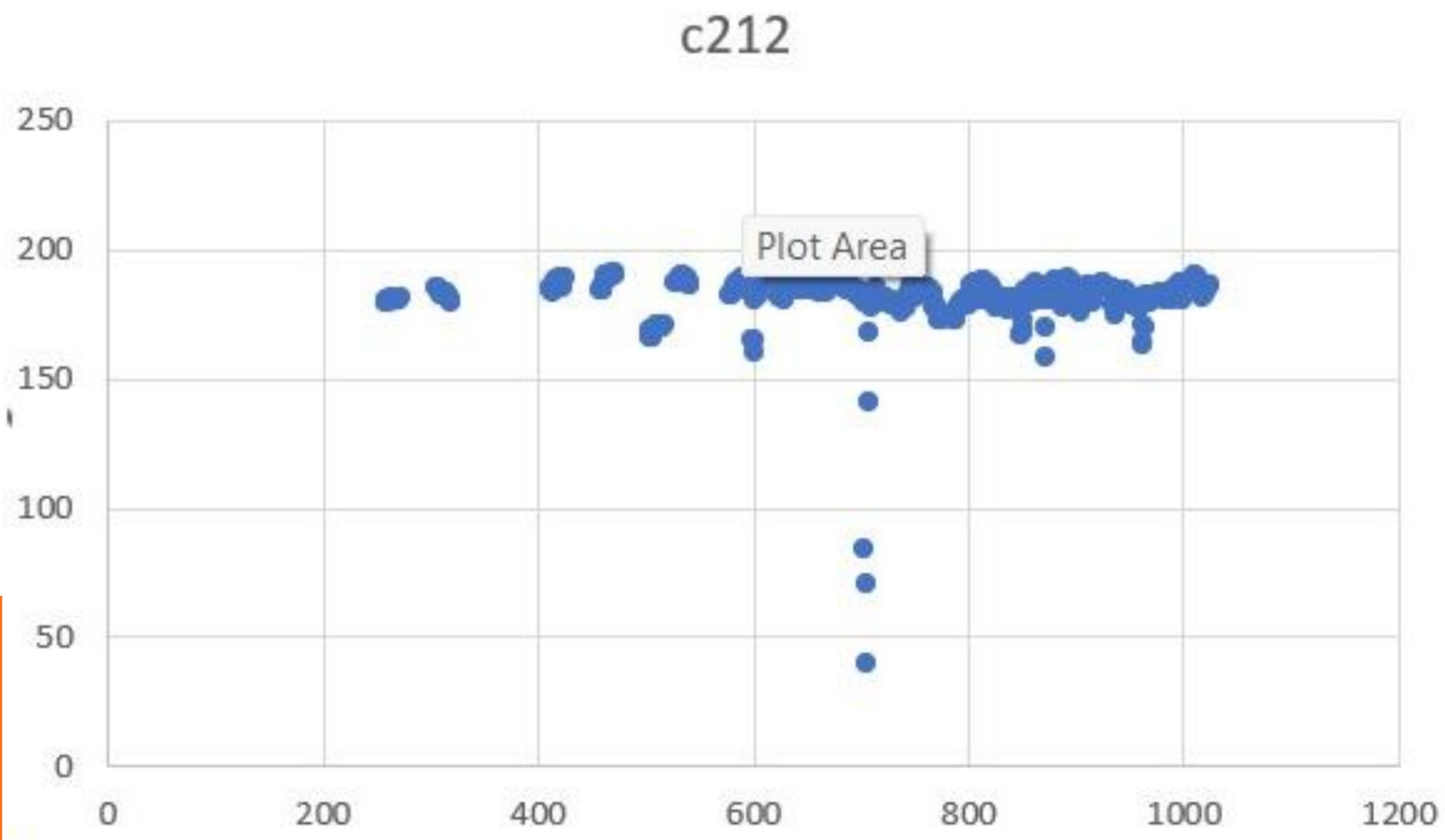
:205	c206	c207	c208	c209	c210	c211	c212	c213	c214	c215	c216	c217	c218	c219
22.125	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
21.6621	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
20.4655	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
21.9522	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
22.5881	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
22.9129	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
22.4222	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
18.2498	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
16.2837	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
17.0183	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
21.0298	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
21.0298	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
22.5605	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
22.6113	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
23.3058	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
23.363	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
23.6938	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
23.7424	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
23.6282	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
23.8147	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
23.9792	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
23.628	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
23.5903	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
23.965	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
23.5758	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
23.5758	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
23.8588	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
23.347	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
23.7504	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
23.6419	#REF!	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A

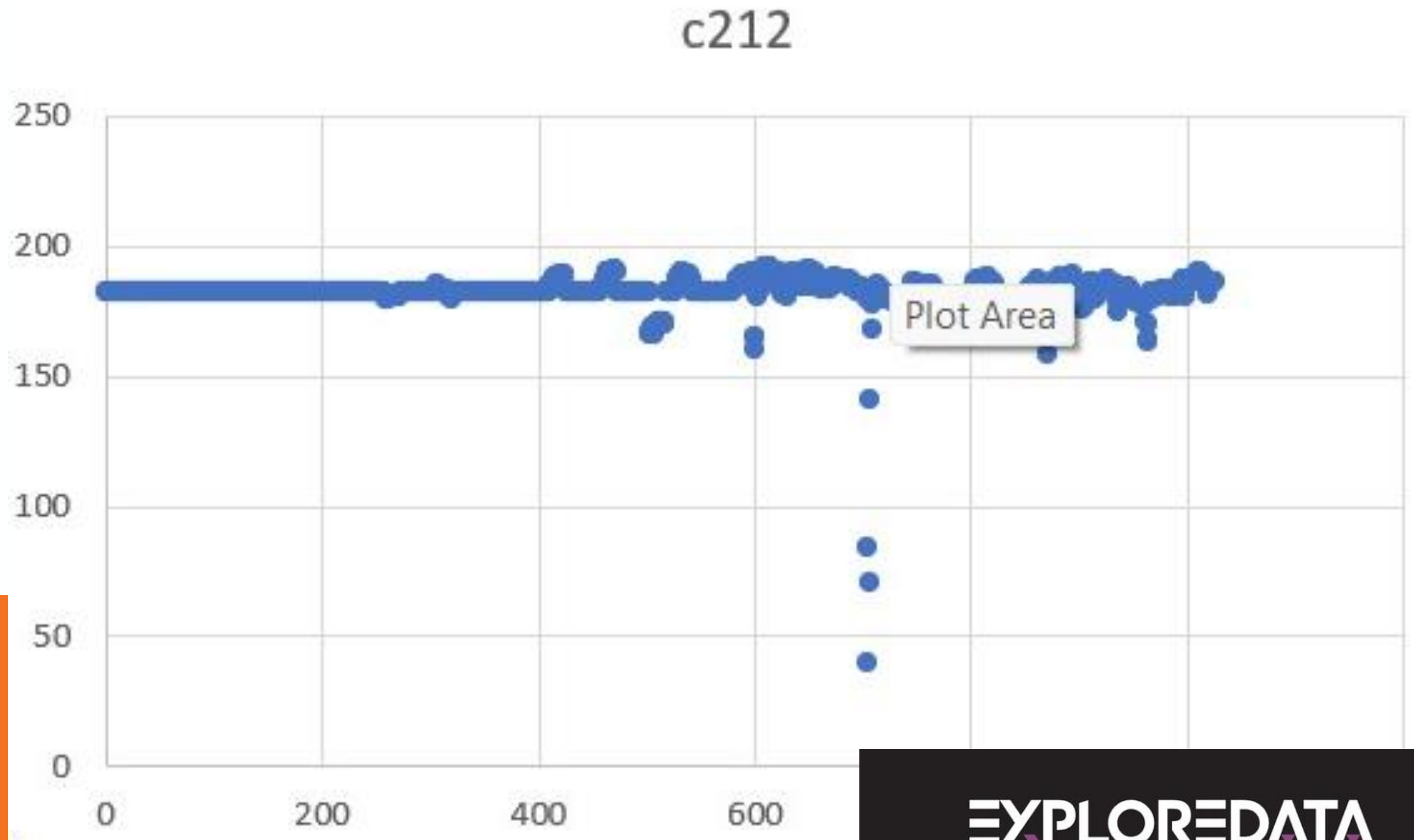
GX	GY	GZ	HA	HB	HC	HD	HE	HF	HG	HH	HI	HJ	HK	HL	HM	HN	HO	HP
06	c207	c208	c209	c210	c211	c212	c213	c214	c215	c216	c217	c218	c219	c220	c221	c222	c223	c224
																		0.16414
																		0.16629
																		0.18067
																		0.16267
																		0.15824
																		0.1539
																		0.15368
																		0.17167
																		0.17529
																		0.18116
																		0.16152
																		0.15308
																		0.14853
																		0.12738
																		0.12321
																		0.11571
																		0.12681
																		0.13227
																		0.13289
																		0.13427
																		0.11394
																		0.11077
																		0.10985
																		0.12119
																		0.12966
																		0.13142
																		0.1336
																		0.13881
																		0.13687
																		0.1407
																		0.14035
																		0.14041
																		0.14009

Continued...

How do we deal with these empty cells?

Here also once again, plotting on the graph yielded values with very less variance before replacing empty cells (as we can see in the left figure for c212). Hence, the empty cells were replaced with the mean of the remaining data of that column





Continued...

After performing all these steps, there were still a few cells with some error values in them and a few more cells which were already empty

Solution to this problem:
We replaced all this empty cells with zero values.

That’s all with Data Cleaning!!

We now proceed to create MLR models ahead by creating a new csv with this cleaned data, named Modified_plant.csv

Creating MLR Model

- The columns c51, c52, c53, c54 are the vibration columns

- We named these 4 columns as y1, y2, y3 and y4 and thus 4 different MLR models were created for each y

- This was followed by **Back Elimination** to remove unwanted extra variables, by eliminating the variable in the **descending order of their p-value**

- This gave us our final 4 MLR models

EXPLOREDATA

Creating MLR Model

Let's take a look at the each of the Models:

Model 1

OLS Regression Results

Dep. Variable:	c51	R-squared (uncentered):	0.994
Model:	OLS	Adj. R-squared (uncentered):	0.994
Method:	Least Squares	F-statistic:	1352.
Date:	Sun, 12 Nov 2023	Prob (F-statistic):	0.00
Time:	17:27:40	Log-Likelihood:	-1129.2
No. Observations:	1025	AIC:	2498.
Df Residuals:	905	BIC:	3090.
Df Model:	120		
Covariance Type:	nonrobust		

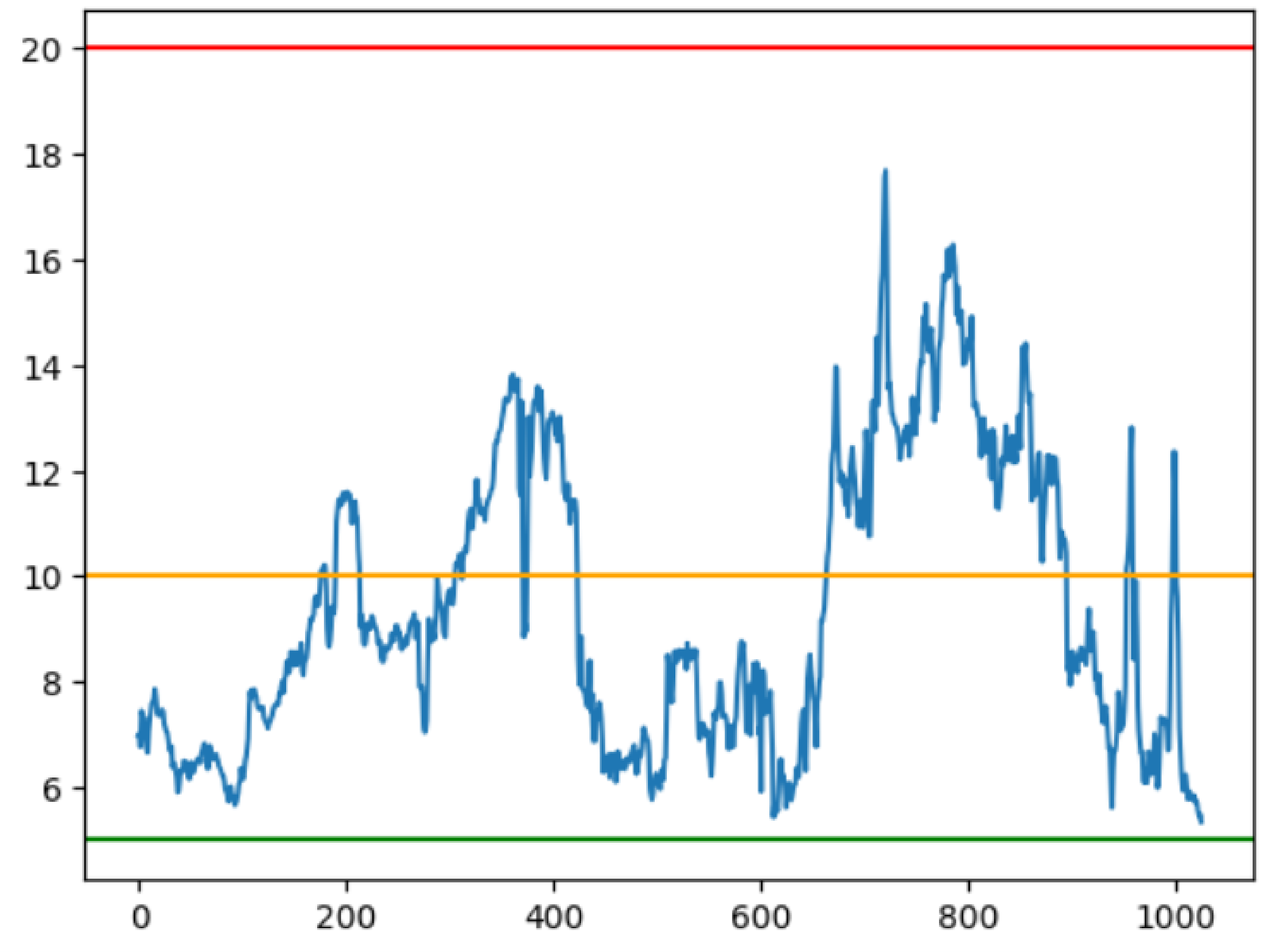
Overview of some regression parameters of our model

Model 1

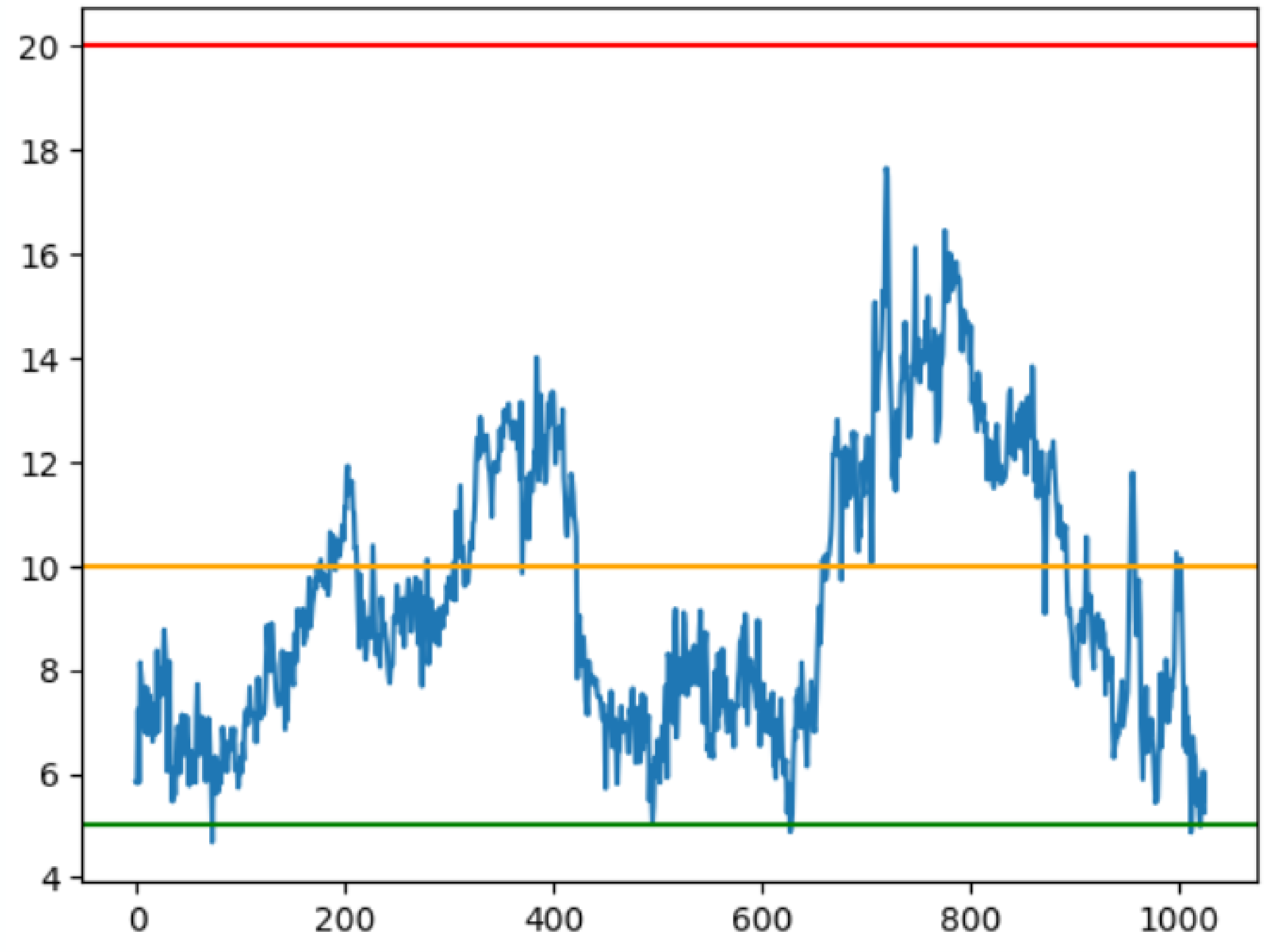
	coef	std err	t	P>|t|	[0.025	0.975]
c3	-0.0724	0.017	-4.337	0.000	-0.105	-0.040
c6	2.983e+06	9.21e+05	3.238	0.001	1.18e+06	4.79e+06
c7	0.9118	0.463	1.970	0.049	0.003	1.820
c13	-0.1877	0.059	-3.184	0.002	-0.303	-0.072
c17	-70.2048	20.272	-3.463	0.001	-109.991	-30.419
c18	-1238.1589	447.635	-2.766	0.006	-2116.683	-359.635
c19	-1.148e+04	4323.996	-2.654	0.008	-2e+04	-2991.576
c22	-0.1202	0.032	-3.706	0.000	-0.184	-0.057
c25	-2.983e+06	9.21e+05	-3.238	0.001	-4.79e+06	-1.18e+06
c31	3708.8857	1218.665	3.043	0.002	1317.147	6100.625
c34	3.8796	1.124	3.452	0.001	1.674	6.085
c35	-3.3499	1.194	-2.806	0.005	-5.693	-1.007
c38	-37.4471	10.506	-3.564	0.000	-58.067	-16.827
c42	-1.0167	0.168	-6.057	0.000	-1.346	-0.687

Glimpse of the independent variable in the model and their coefficient value alongside it

Model 1:
Actual and Predicted Plots for Reference



Actual
c51



Predicted
c51

EXPLOREDATA

Creating MLR Model

Model 2

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    c52   R-squared:                       0.966
Model:                            OLS   Adj. R-squared:                  0.962
Method:                 Least Squares   F-statistic:                     228.2
Date:                Sun, 12 Nov 2023   Prob (F-statistic):               0.00
Time:                        17:40:32   Log-Likelihood:                -540.90
No. Observations:                1025   AIC:                             1310.
Df Residuals:                     911   BIC:                             1872.
Df Model:                         113
Covariance Type:            nonrobust
```

Overview of some regression parameters of our model

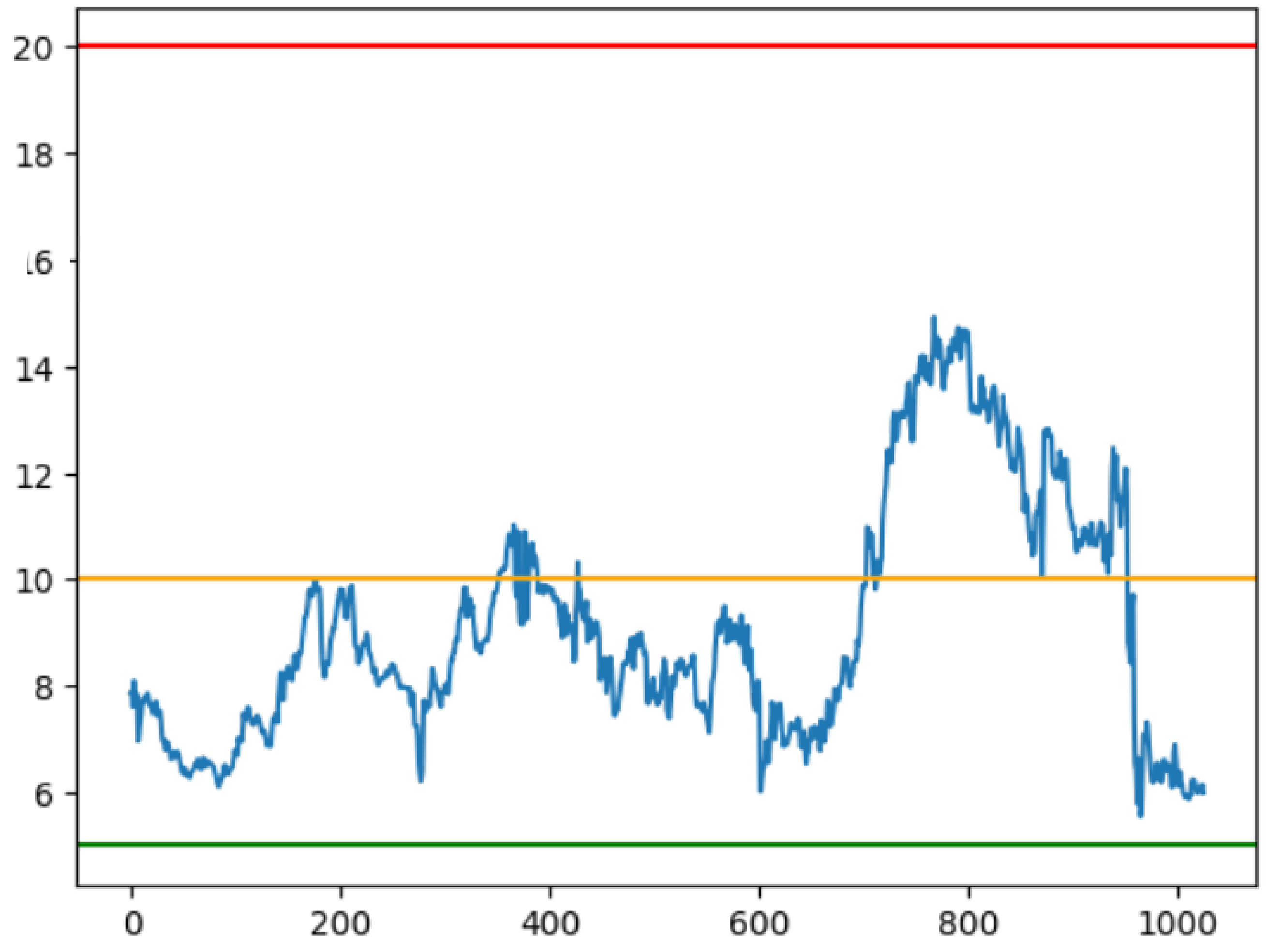
EXPLOREDATA

Model 2

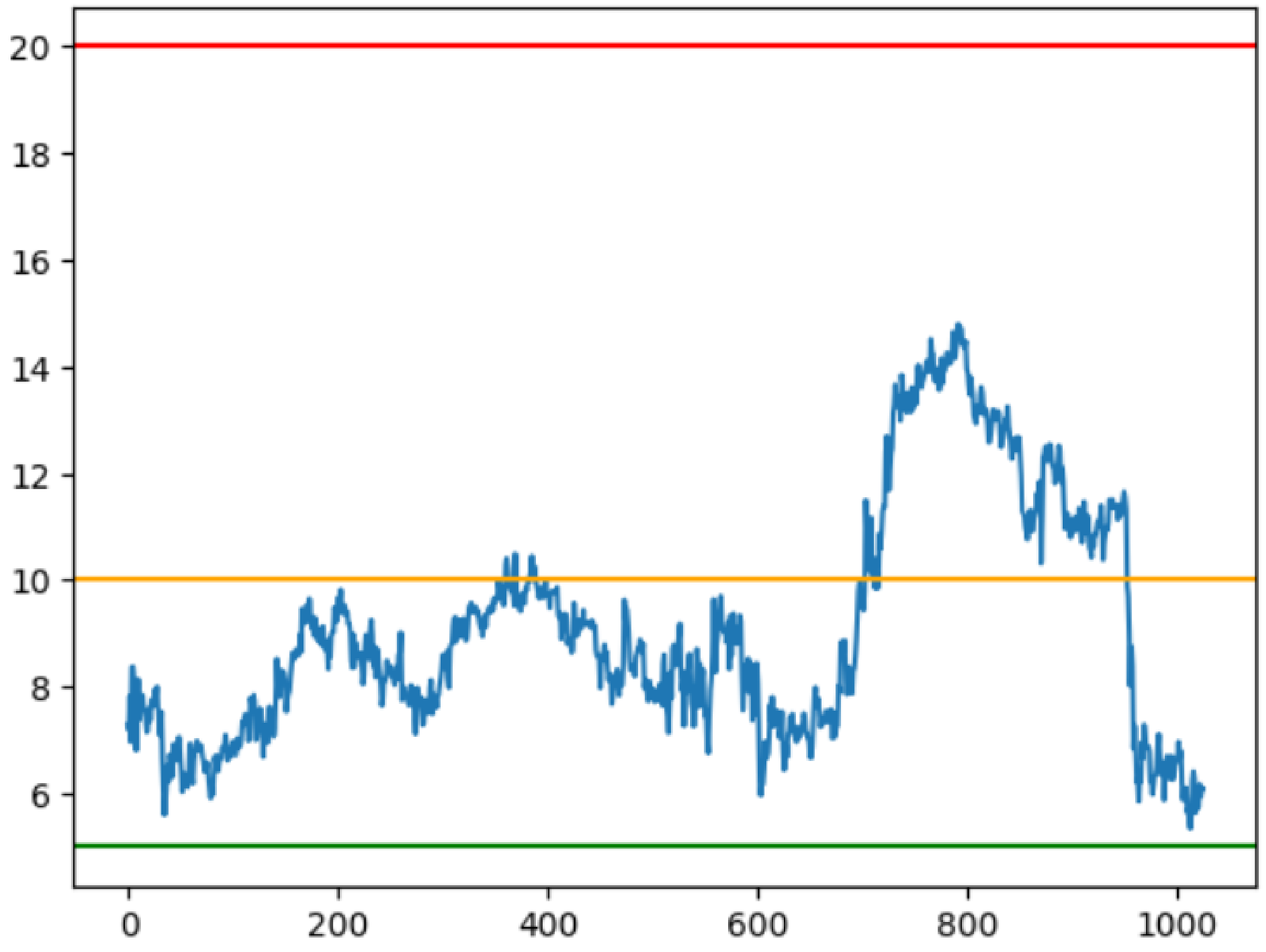
c3	-0.3250	0.076	-4.276	0.000	-0.474	-0.176
c4	0.9193	0.181	5.074	0.000	0.564	1.275
c5	-5.493e+05	2.69e+05	-2.044	0.041	-1.08e+06	-2.2e+04
c6	-4.076e+06	1.5e+06	-2.718	0.007	-7.02e+06	-1.13e+06
c7	2.0683	0.293	7.063	0.000	1.494	2.643
c11	-0.1490	0.041	-3.610	0.000	-0.230	-0.068
c17	87.0615	32.995	2.639	0.008	22.306	151.817
c18	2012.1944	694.936	2.896	0.004	648.332	3376.057
c19	1.923e+04	6591.012	2.917	0.004	6291.826	3.22e+04
c22	-0.0363	0.017	-2.142	0.032	-0.070	-0.003
c24	5.493e+05	2.69e+05	2.044	0.041	2.2e+04	1.08e+06
c25	4.076e+06	1.5e+06	2.718	0.007	1.13e+06	7.02e+06
c28	1.5125	0.262	5.778	0.000	0.999	2.026
c31	-5354.8345	1923.185	-2.784	0.005	-9129.222	-1580.447
c34	1.2867	0.643	2.000	0.046	0.024	2.550

Glimpse of the independent variable in the model and their coefficient value alongside it

Model 2:
Actual and Predicted Plots for Reference



Actual c52



Predicted
c52

EXPLOREDATA

Creating MLR Model

Let's take a look at the each of the Models:

Model 3

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    c53   R-squared:                       0.985
Model:                            OLS   Adj. R-squared:                  0.984
Method:                 Least Squares   F-statistic:                     587.0
Date:                Sun, 12 Nov 2023   Prob (F-statistic):               0.00
Time:                        17:44:44   Log-Likelihood:                -1194.3
No. Observations:                1025   AIC:                             2601.
Df Residuals:                     919   BIC:                             3124.
Df Model:                         105
Covariance Type:            nonrobust
==============================================================================
```

Overview of some regression parameters of our model

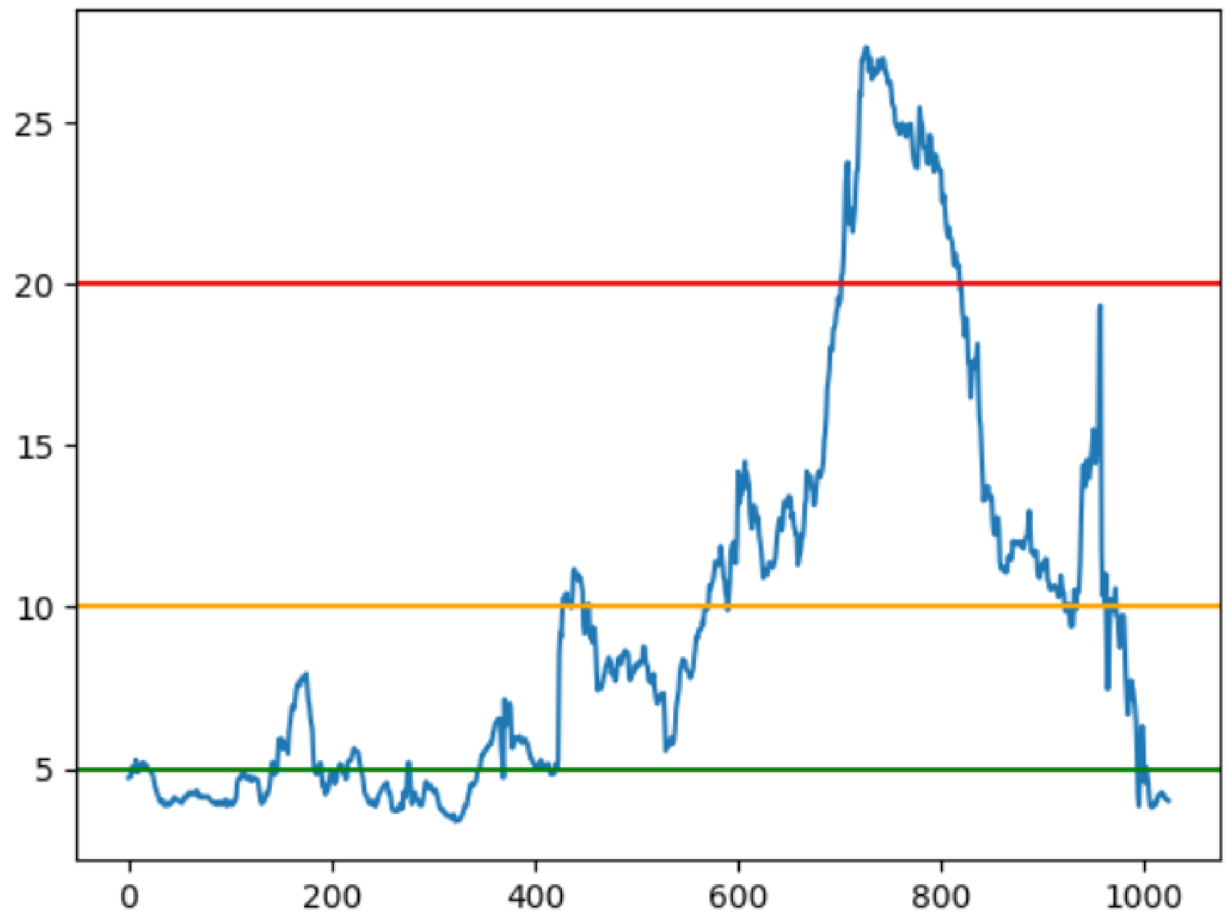
EXPLOREDATA

Model 3

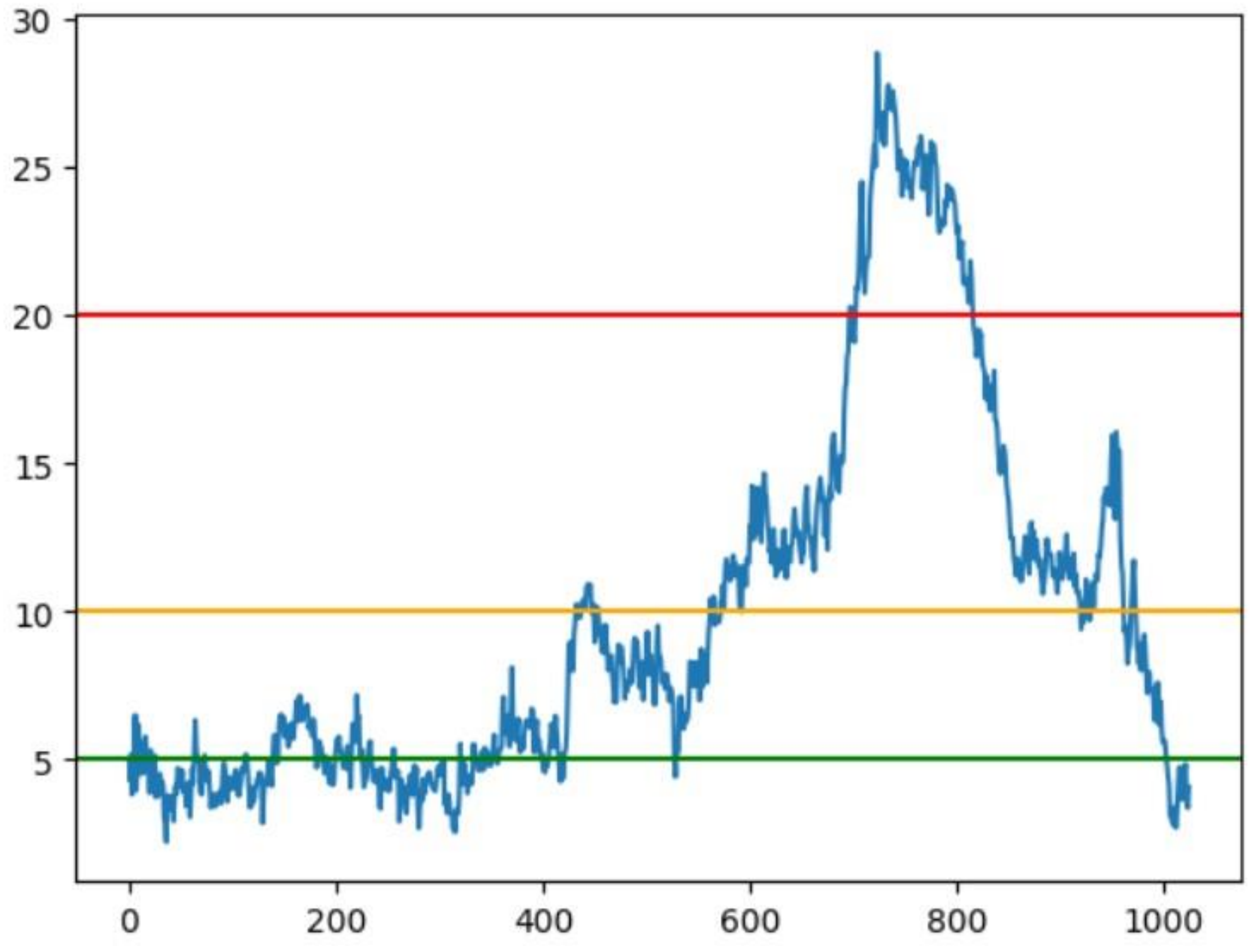
	coef	std err	t	P>|t|	[0.025	0.975]
c2	1.349e+04	5717.368	2.359	0.019	2264.704	2.47e+04
c3	-0.0909	0.017	-5.411	0.000	-0.124	-0.058
c6	-2.706e+05	7.86e+04	-3.445	0.001	-4.25e+05	-1.16e+05
c9	-0.5633	0.060	-9.312	0.000	-0.682	-0.445
c17	8.7448	2.371	3.688	0.000	4.091	13.399
c22	-0.1237	0.034	-3.643	0.000	-0.190	-0.057
c25	2.706e+05	7.86e+04	3.445	0.001	1.16e+05	4.25e+05
c27	-0.8986	0.263	-3.421	0.001	-1.414	-0.383
c28	6.1256	1.618	3.786	0.000	2.950	9.301
c29	0.2740	0.070	3.907	0.000	0.136	0.412
c30	-6.9417	2.335	-2.972	0.003	-11.525	-2.358
c31	-166.1326	74.286	-2.236	0.026	-311.923	-20.343
c34	4.1465	1.189	3.486	0.001	1.812	6.481
c40	2.184e+06	9.3e+05	2.348	0.019	3.59e+05	4.01e+06
c42	0.4534	0.178	2.546	0.011	0.104	0.803

Glimpse of the independent variable in the model and their coefficient value alongside it

Model 3:
Actual and Predicted Plots for Reference



Actual
c53



Predicted
c53

EXPLOREDATA

Creating MLR Model

Let’s take a look at the each of the Models:

Model 4

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    c54   R-squared:                       0.984
Model:                            OLS   Adj. R-squared:                  0.982
Method:                 Least Squares   F-statistic:                     546.9
Date:                Sun, 12 Nov 2023   Prob (F-statistic):               0.00
Time:                        17:47:59   Log-Likelihood:                -1153.0
No. Observations:                1025   AIC:                             2510.
Df Residuals:                     923   BIC:                             3013.
Df Model:                         101
Covariance Type:            nonrobust
```

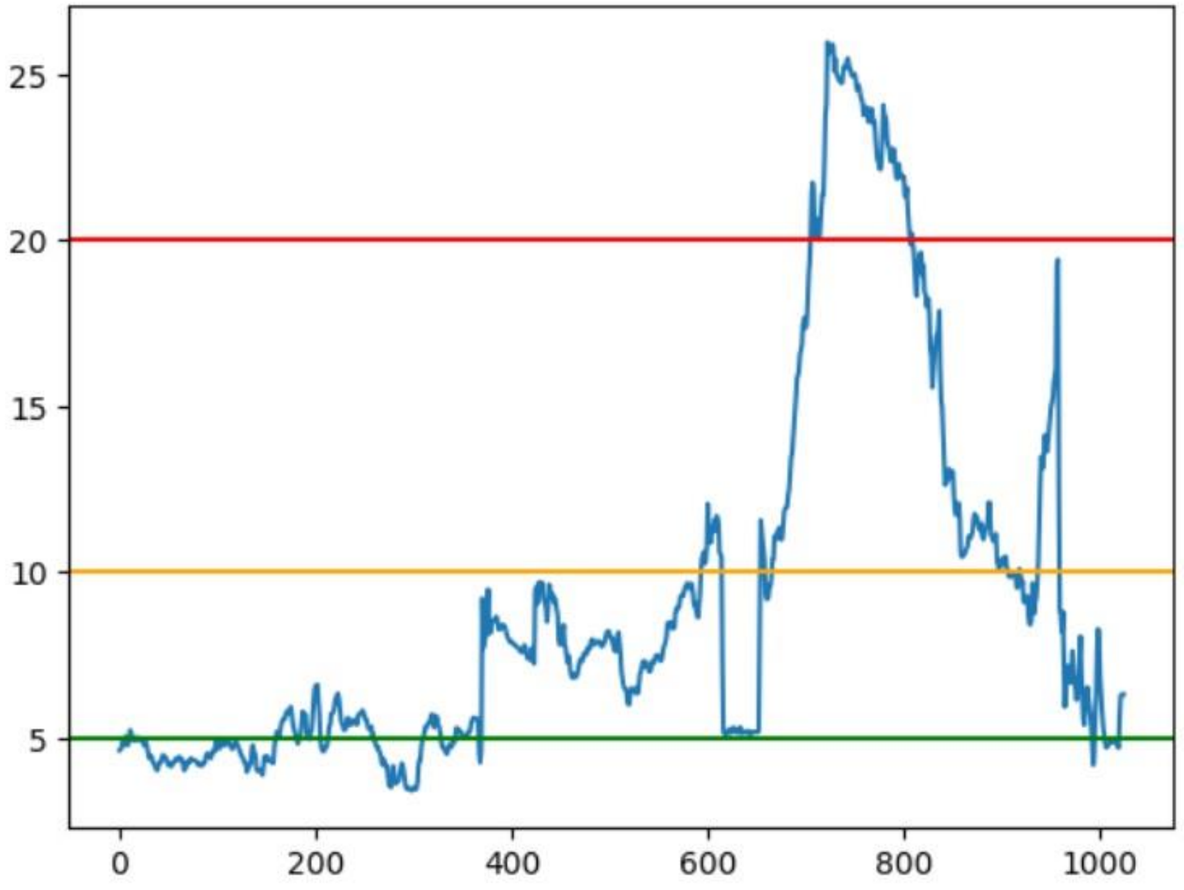
Overview of some regression parameters of our model

Model 4

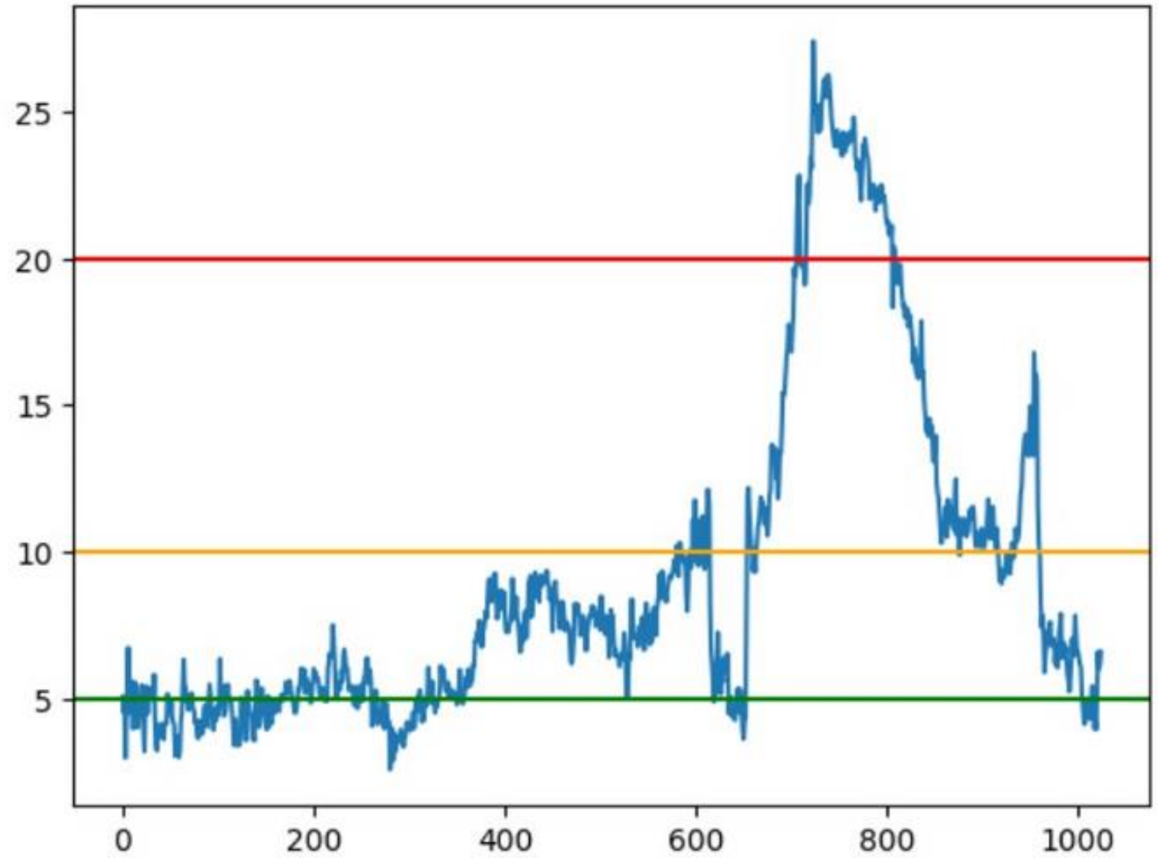
c2	1.203e+04	5471.043	2.200	0.028	1297.409	2.28e+04
c3	-0.1637	0.025	-6.474	0.000	-0.213	-0.114
c7	2.9237	0.446	6.555	0.000	2.048	3.799
c21	-0.0999	0.042	-2.355	0.019	-0.183	-0.017
c22	-0.2186	0.035	-6.240	0.000	-0.287	-0.150
c26	0.6433	0.100	6.442	0.000	0.447	0.839
c27	-1.0251	0.258	-3.969	0.000	-1.532	-0.518
c33	-0.9136	0.158	-5.779	0.000	-1.224	-0.603
c34	5.3310	1.136	4.693	0.000	3.102	7.560
c40	2.004e+06	8.89e+05	2.255	0.024	2.6e+05	3.75e+06
c44	-0.3019	0.071	-4.254	0.000	-0.441	-0.163
c47	-24.1999	4.408	-5.491	0.000	-32.850	-15.550
c49	21.8659	3.790	5.770	0.000	14.429	29.303
c55	90.9540	39.345	2.312	0.021	13.737	168.171

Glimpse of the independent variable in the model and their coefficient value alongside it

Model 4



Actual
c54



Predicted
c54

Use of these MLR Models Created

These MLR models are used to predict whether the vibrations are ‘critical’ or ‘high’ for a given set of parameters.

On comparing the predictions with the actual levels for each of the vibrations, we get that the predictions were accurate with a percentage of -

95.4 for y1
97.8 for y2
88.78 for y3 and
84.68 for y4

	HW	HX	HY	HZ
ur	Alert1	Alert2	Alert3	Alert4
.8	MODERATE	MODERATE	SAFE	MODERATE
.8	MODERATE	MODERATE	SAFE	SAFE
.8	MODERATE	MODERATE	MODERATE	SAFE
.8	MODERATE	MODERATE	SAFE	SAFE
.8	MODERATE	MODERATE	SAFE	SAFE
.8	MODERATE	MODERATE	MODERATE	MODERATE
.8	MODERATE	MODERATE	MODERATE	MODERATE
.8	MODERATE	MODERATE	SAFE	SAFE
.8	MODERATE	MODERATE	MODERATE	SAFE
.8	MODERATE	MODERATE	MODERATE	MODERATE
.8	MODERATE	MODERATE	MODERATE	MODERATE
.8	MODERATE	MODERATE	SAFE	SAFE
.8	MODERATE	MODERATE	SAFE	MODERATE
.8	MODERATE	MODERATE	MODERATE	MODERATE
.8	MODERATE	MODERATE	SAFE	SAFE
.8	MODERATE	MODERATE	MODERATE	SAFE
.8	MODERATE	MODERATE	MODERATE	SAFE
.8	MODERATE	MODERATE	MODERATE	MODERATE
.8	MODERATE	MODERATE	SAFE	SAFE
.8	MODERATE	MODERATE	SAFE	MODERATE
.8	MODERATE	MODERATE	SAFE	MODERATE
.8	MODERATE	MODERATE	MODERATE	SAFE
.8	MODERATE	MODERATE	SAFE	SAFE
.8	MODERATE	MODERATE	SAFE	SAFE

Predicted levels for each of the four vibrations for given values

EXPLOREDATA

Classifying the Vibrations based on MLR Models

We now create 4 new MLR models with independent variables as y1, y2, y3 and y4 each. But this time the dependent variables are only the controllable parameters. We again perform Back Elimination and get the most appropriate controllable parameters for each of the y's.

This is done to find the importance of each of the controllable parameter in contributing towards each of the vibration. This will help us in reducing CRITICAL and HIGH y's by altering most important parameters first.

EXPLOREDATA

Finding Dependence of Vibrations on Controllable Parameters

- Here we take a look at contribution of each of the variables by multiplying the coefficient of each of the variable with the mean of values of that variable. Lets call this sum as **si** for each **i**

- This value is further divided by summation of **si** to get a normalized value. This normalized values gives us the relative contribution of each variable in the total.

- The magnitude of these value are arranged in descending order which gives us their importance order too..

EXPLOREDATA

Finding Dependence of Vibrations on Controllable Parameters

The ‘|’ represents change of sign :

To reduce vibrations, reducing the value of the left most parameter is as significant as increasing the value of right most parameter

What we get is -

For y1

'c33', 'c32', 'c31', 'c28', 'c39', 'c161' | 'c143', 'c158', 'c155', 'c156', 'c139', 'c157', 'c160', 'c142', 'c27'

For y2

'c26', 'c31', 'c30', 'c143', 'c28', 'c39', 'c161', 'c158', 'c155' | 'c157', 'c160', 'c139', 'c32', 'c142', 'c27', 'c33', 'c29'

For y3

'c26', 'c31', 'c143', 'c30', 'c28', 'c155', 'c163' | 'c156', 'c162', 'c160', 'c157', 'c39', 'c139', 'c33', 'c27', 'c142'

For y4

'c26', 'c31', 'c30', 'c143', 'c155', 'c158', 'c163', 'c161' | 'c156', 'c162', 'c160', 'c157', 'c139', 'c39', 'c33', 'c142', 'c27'

Logic to Find Important parametrs to reduce vibrations

- Here we take a look at contribution of each of the variables by multiplying the coefficient of each of the variable with the mean of values of that variable. Lets call this sum as **si** for each **i**

- This value is further divided by summation of **si** and to get a normalized value. This normalized values gives us the relative contribution of each variable in the total.

- The magnitude of these value are arranged in descending order which gives us their importance order too..

EXPLOREDATA

PART 2

SPECIFIC ENERGY PREDICTION BY USING ML MODEL

The column c241 represents the specific energy. The task here is to create an ML prediction model which will be used to understand which parameters (operating + controllable) significantly contribute to the 'specific energy', so that energy reduction research and efforts can be focused on them

EXPLOREDATA

SPECIFIC ENERGY PREDICTION BY USING ML MODEL

OLS Regression Results

Dep. Variable:	c241	R-squared:	0.999
Model:	OLS	Adj. R-squared:	0.999
Method:	Least Squares	F-statistic:	1.016e+04
Date:	Sun, 12 Nov 2023	Prob (F-statistic):	0.00
Time:	16:03:10	Log-Likelihood:	2477.3
No. Observations:	1025	AIC:	-4745.
Df Residuals:	920	BIC:	-4227.
Df Model:	104		
Covariance Type:	nonrobust		

Regression parameters of the created ML model (for reference)

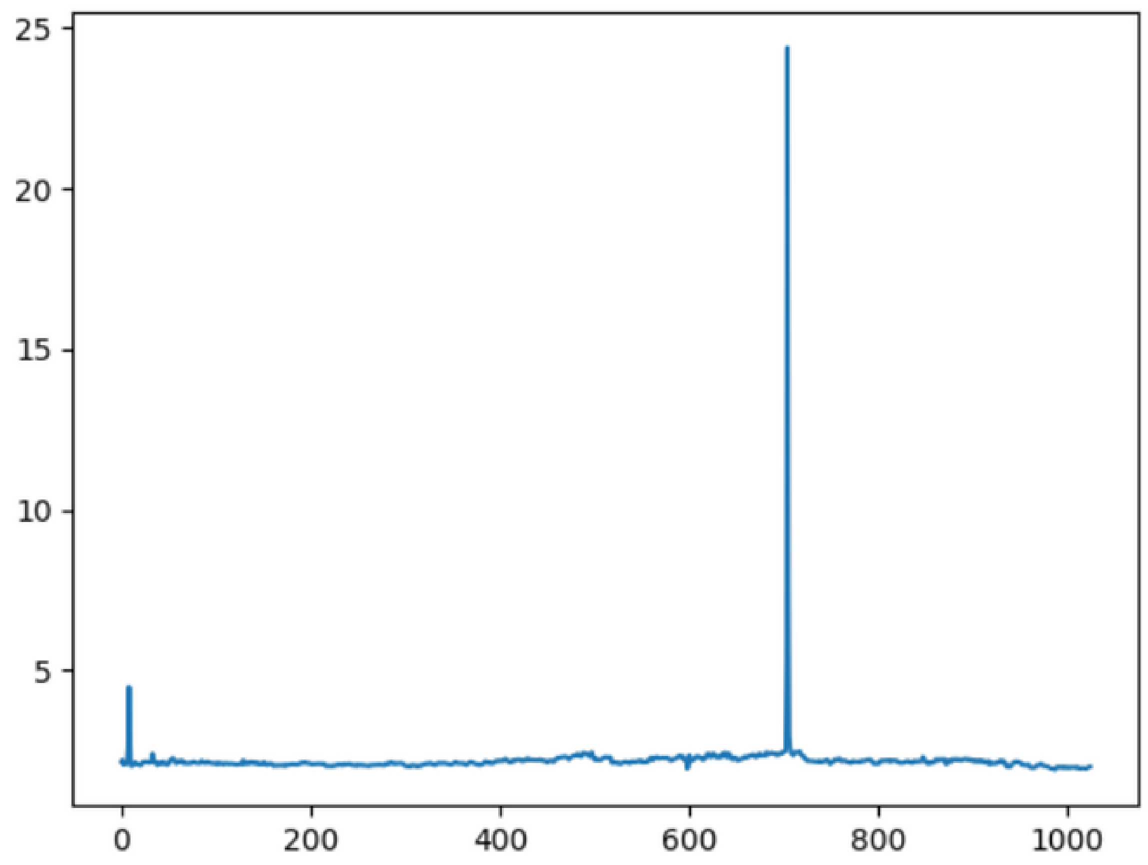
SPECIFIC ENERGY PREDICTION BY USING ML MODEL

	coef	std err	t	P>|t|	[0.025	0.975]
c5	0.0202	0.005	3.935	0.000	0.010	0.030
c6	0.0214	0.002	9.119	0.000	0.017	0.026
c7	0.1024	0.021	4.901	0.000	0.061	0.143
c8	-0.0536	0.004	-12.437	0.000	-0.062	-0.045
c9	-0.0340	0.005	-7.275	0.000	-0.043	-0.025
c11	-0.1085	0.003	-39.887	0.000	-0.114	-0.103
c13	-0.0118	0.002	-7.035	0.000	-0.015	-0.008
c16	-0.0136	0.002	-6.347	0.000	-0.018	-0.009
c17	-0.0068	0.002	-4.465	0.000	-0.010	-0.004
c26	-0.0255	0.006	-4.596	0.000	-0.036	-0.015
c28	0.1035	0.042	2.477	0.013	0.022	0.186
c29	0.0092	0.002	4.370	0.000	0.005	0.013
c31	0.0903	0.041	2.228	0.026	0.011	0.170
c32	0.0085	0.003	2.855	0.004	0.003	0.014

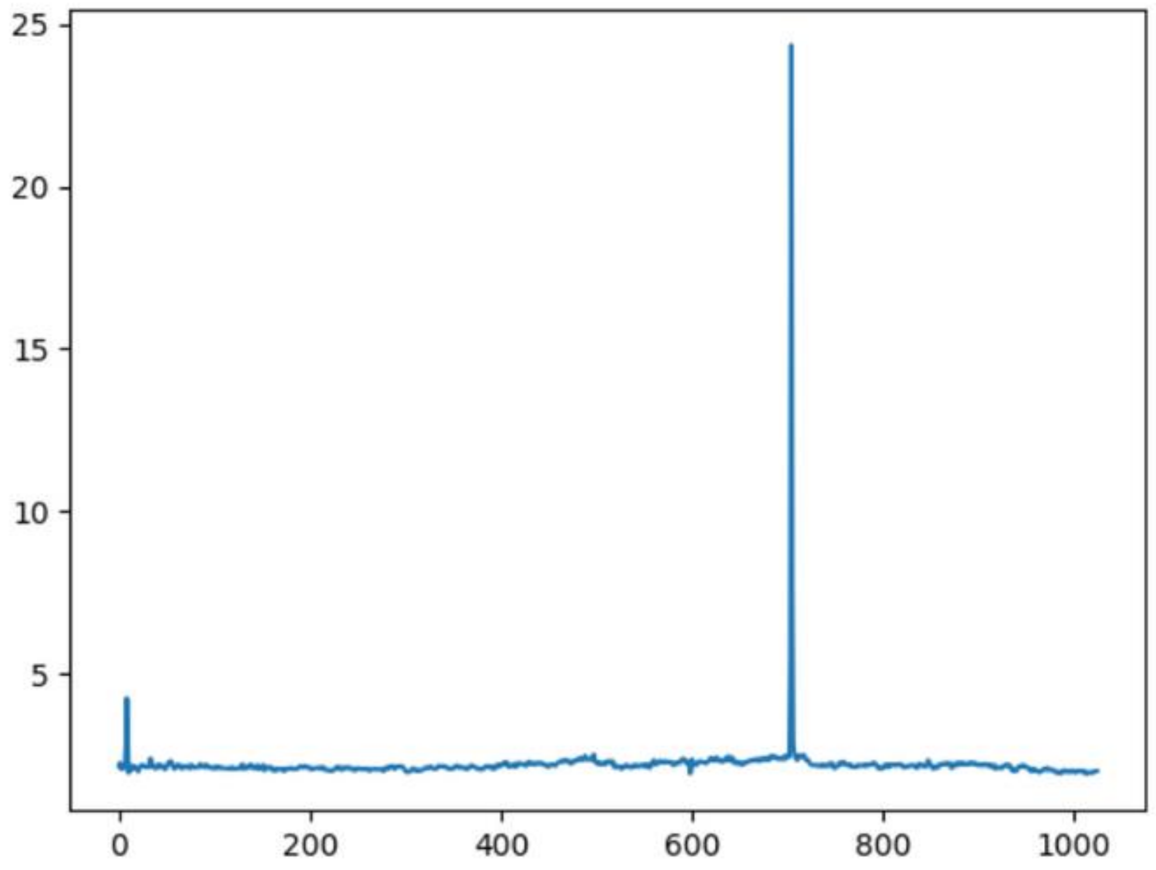
Coefficients of independent variable alongside it

EXPLOREDATA

SPECIFIC ENERGY PREDICTION BY USING ML MODEL



Actual
c241



Predicted
c241

SPECIFIC ENERGY PREDICTION BY USING ML MODEL

- Now we need to filter out the predicted parameters form the pool of predicted + controlled parameters
- We accomplish this by considering the controllable parameters as the one in which the values are close to each other with less variance
- This is because if we can control some inputs of the system, the change in its value is usually not observed, even if is there, it is quite systematic and does not change randomly

EXPLOREDATA

SPECIFIC ENERGY PREDICTION BY USING ML MODEL

This method was employed by taking controlling parameters as those in which the standard deviation of values is less than 0.065 (i.e 6.5%) of the mean of values of that variable; rest being the predicted one.

We came to the conclusion of choosing 0.065 after taking various values and plotting graphs for each of the parameter. We came to 0.065 when we were satisfied that the visual depiction of the data of each variable matched with its classification.. (i.e those graphs which had very low variance visually, were mostly correctly classified as controlling.

EXPLOREDATA

SPECIFIC ENERGY PREDICTION BY USING ML MODEL

These are the filtered predicting parameters

```
['c5', 'c6', 'c7', 'c9', 'c17', 'c28', 'c37', 'c39', 'c42', 'c43', 'c45', 'c46', 'c52', 'c53', 'c55', 'c70', 'c89', 'c96', 'c9
7', 'c102', 'c113', 'c118', 'c137', 'c144', 'c149', 'c127', 'c130', 'c150', 'c153', 'c159', 'c160', 'c161', 'c162', 'c181', 'c1
84', 'c191', 'c197', 'c214', 'c221']
```

These are the controlling parameters

```
['c8', 'c11', 'c13', 'c16', 'c26', 'c29', 'c31', 'c32', 'c38', 'c41', 'c48', 'c50', 'c58', 'c62', 'c63', 'c67', 'c69', 'c75',
'c78', 'c79', 'c80', 'c81', 'c86', 'c87', 'c88', 'c90', 'c91', 'c92', 'c93', 'c95', 'c99', 'c100', 'c101', 'c103', 'c105', 'c10
7', 'c108', 'c112', 'c119', 'c123', 'c135', 'c140', 'c141', 'c148', 'c152', 'c176', 'c178', 'c179', 'c187', 'c192', 'c194', 'c1
95', 'c196', 'c198', 'c201', 'c203', 'c205', 'c206', 'c212', 'c213', 'c215', 'c216', 'c222', 'c224', 'c225', 'c231', 'c232']
```

We can verify this by comparing this with the controllable parameters given to us and can see there are very few given controllable parameters in the predicting parameters predicted by us

EXPLOREDATA

SPECIFIC ENERGY PREDICTION BY USING ML MODEL

The Regression Results of MLR using Prediction Parameters

```
                            OLS Regression Results
=======================================================================================
Dep. Variable:                   c241   R-squared (uncentered):                   0.992
Model:                            OLS   Adj. R-squared (uncentered):              0.992
Method:                 Least Squares   F-statistic:                              4585.
Date:                Sun, 12 Nov 2023   Prob (F-statistic):                        0.00
Time:                        23:26:25   Log-Likelihood:                          171.61
No. Observations:                1025   AIC:                                     -287.2
Df Residuals:                     997   BIC:                                     -149.1
Df Model:                          28
Covariance Type:            nonrobust
==============================================================================
```