RUL Prediction Model

1. Data Preprocessing:

Datasets were in the form of space-separated text files (without any labels) instead of usual csv file. We used read_csv function with separator set to regex operator (\s+) and set the column names by using names parameter. There were no null values and string labels.

Now RUL column was not present in the provided training datasets.
It was calculated by taking the difference between maximum time instance for that particular unit and its current time in cycles because it is given that the data in training set is till the system failure.

Feature Reduction
Correlation heatmap was plotted and a new dataframe was created to eliminate the columns which have very low correlation with RUL column. However, on training the Random Forest Regressor with this new dataset, the R squared score was only 0.57(as tested on splitted data) whereas the same model gave R squared score of 0.94 (when including unit column).

2. Model Building
Other models such as SVR gave R2 score of 0.74 on splitted data and similar results were obtained when the model was hyper-tuned. XGB also gave the same R squared score.

3. Predicting RUL values

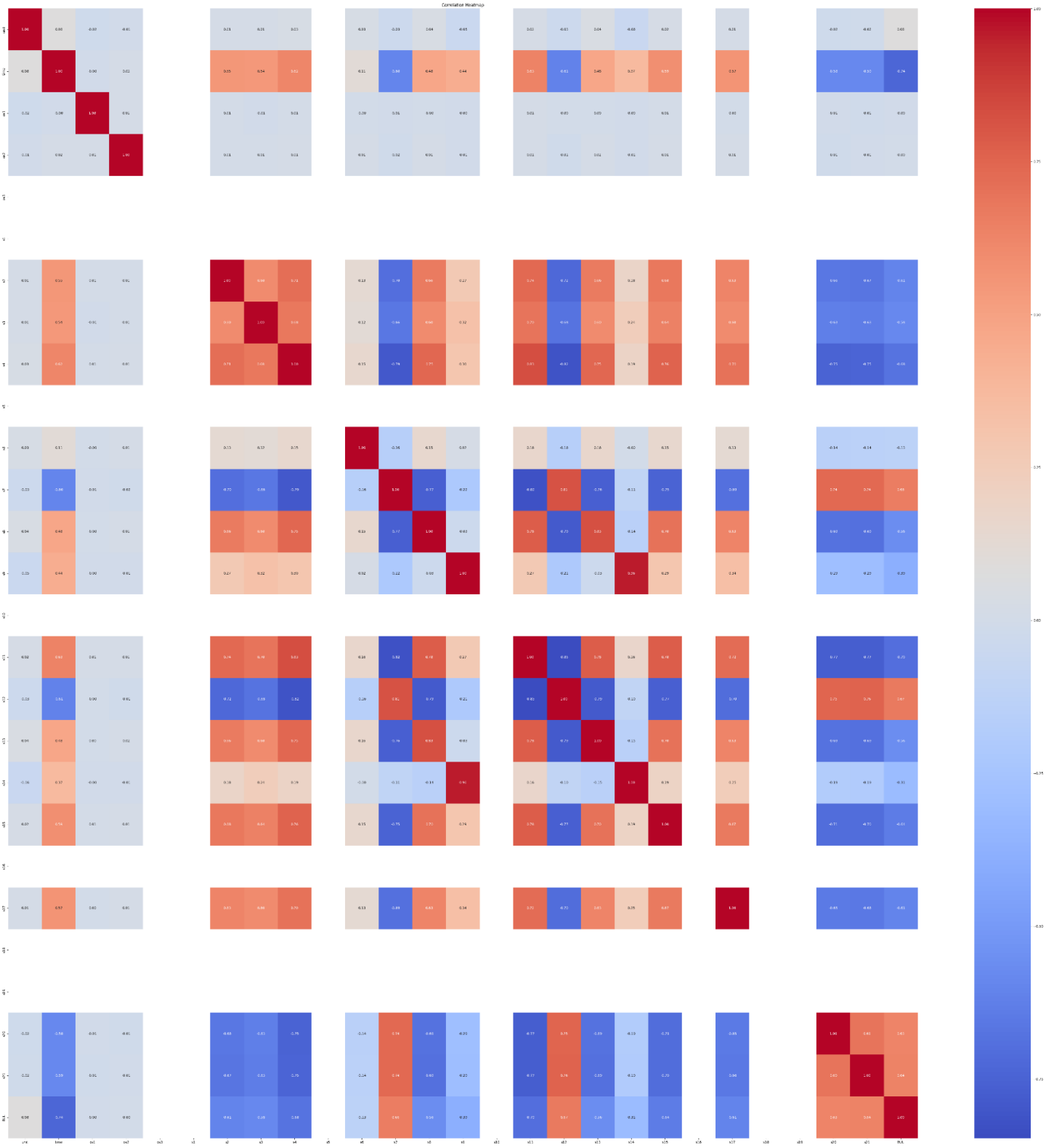
Finally, we chose Random Forest Regressor for predicting the test dataset values. The RUL values were predicted with an average R squared score of 0.62.

Observation: Model performed better on split training dataset as compared to given test dataset.

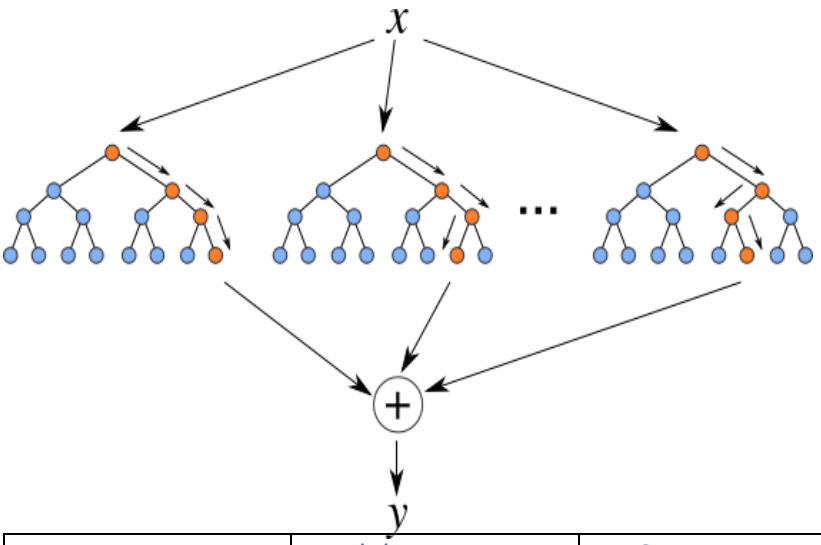
```
dft1.describe()
```

	unit	time	os1	os2	os3	s1	s2	s3	s4	s5	...	s13	s14	s15	
count	20631.000000	20631.000000	20631.000000	20631.000000	20631.0	2.063100e+04	20631.000000	20631.000000	20631.000000	2.063100e+04	...	20631.000000	20631.000000	20631.000000	2.0631
mean	51.506568	108.807862	-0.000009	0.000002	100.0	5.186700e+02	642.680934	1590.523119	1408.933782	1.462000e+01	...	2388.096152	8143.752722	8.442146	3.000
std	29.227633	68.880990	0.002187	0.000293	0.0	6.537152e-11	0.500053	6.131150	9.000605	3.394700e-12	...	0.071919	19.076176	0.037505	1.556
min	1.000000	1.000000	-0.008700	-0.000600	100.0	5.186700e+02	641.210000	1571.040000	1382.250000	1.462000e+01	...	2387.880000	8099.940000	8.324900	3.000
25%	26.000000	52.000000	-0.001500	-0.000200	100.0	5.186700e+02	642.325000	1586.260000	1402.360000	1.462000e+01	...	2388.040000	8133.245000	8.414900	3.000
50%	52.000000	104.000000	0.000000	0.000000	100.0	5.186700e+02	642.640000	1590.100000	1408.040000	1.462000e+01	...	2388.090000	8140.540000	8.438900	3.000
75%	77.000000	156.000000	0.001500	0.000300	100.0	5.186700e+02	643.000000	1594.380000	1414.555000	1.462000e+01	...	2388.140000	8148.310000	8.465600	3.000
max	100.000000	362.000000	0.008700	0.000600	100.0	5.186700e+02	644.530000	1616.910000	1441.490000	1.462000e+01	...	2388.560000	8293.720000	8.584800	3.000

8 rows × 27 columns


Correlation Heatmap

RANDOM FOREST REGRESSOR



Dataset	Model	RMSE on Test set:	RMSE on Training set:	R^2 Score on Test set
FD001	Random Forest Regressor (without unit column)	36.81	13.55	0.73
FD001	Random Forest Regressor (with unit column)	16.55	6.09	0.94
FD001(columns reduced)	Random Forest Regressor (without unit column)	45.734	17.022	0.58
FD001(columns reduced)	Random Forest Regressor (with unit column)	38.125	14.08	0.70
FD001	SVR	36.16	35.52	0.71
FD001	XGB Regressor	35.59	32.10	0.72
FD002	Random Forest Regressor	37.20	13.91	0.71
FD003	Random Forest Regressor	49.06	18.71	0.76
FD004	Random Forest Regressor	48.60	18.59	0.70