KL AIoT HACKATHON

- Submitted by Amit Pawar

Problem Statement

- A billion dollar chemical manufacturing unit creates a Chemical Product with the help of very costly raw materials. The current economic crisis and impeding fear of war has put them in a place where it is crucial for them to reduce expense while not compromising on the target.

1. Raw Materials & Catalyst:
    1. Raw Materials are used to create the product. As this is a chemical reaction, using more than what is required does not yield more product. [Stoichiometry - 101]. Find the optimum amount of raw material to be fed for highest conversion.

2. Reactor Temperature:
    1. Steam has to be supplied in order to maintain the reactor temperature. Does lowering steam reduce the amount of product the plant manufactures? Are we supplying more steam than necessary?

3. Conversion Rate:
    1. The ultimate KPI that tells us how efficiently is the reaction churning out the product. The higher this value is, the better it is.

- Analyse the dataset to help the plant reduce expense while not compromising on the conversion rate.
- Help the plant achieve highest conversion rate everyday!

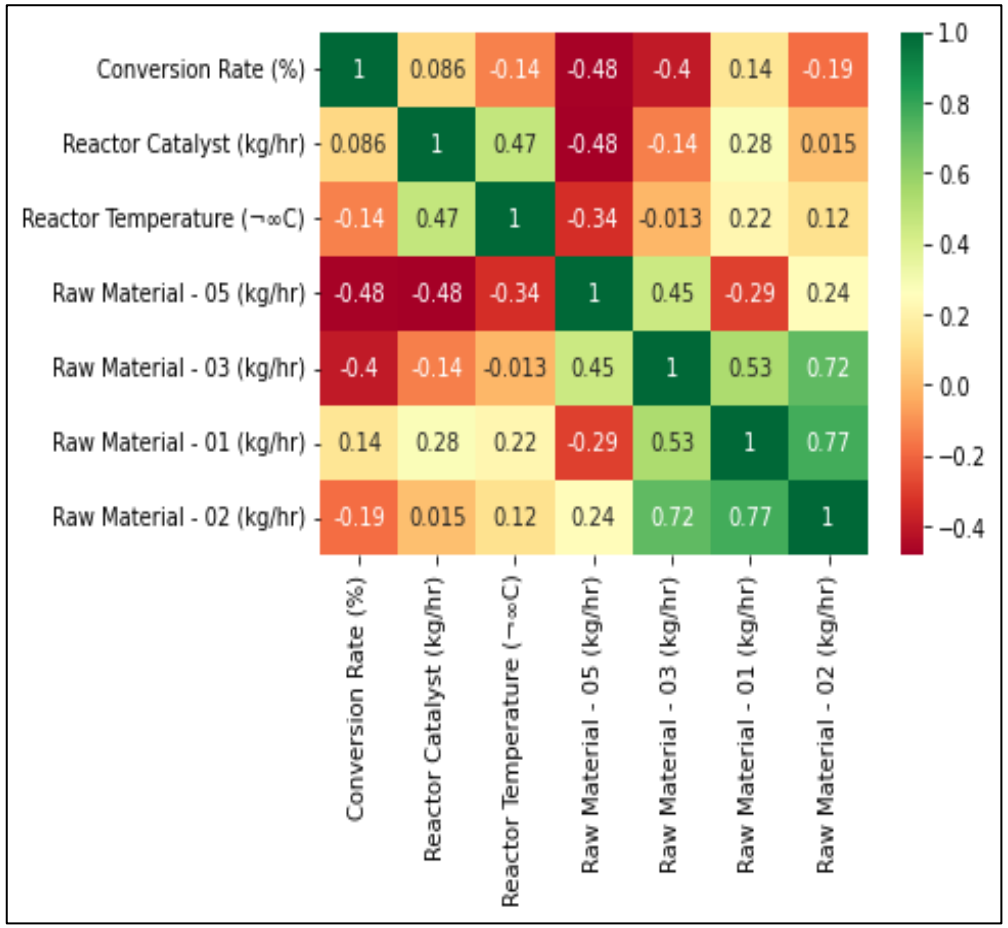
Important parameters in the dataset

- **Conversion rate (%)** : tells us how efficiently is the reaction churning out the product.
- **Reactor Catalyst** : to speed up the reaction
- **Reactor Temperature (¬∞C)** : required temperature for the reaction. Steam is needed to maintain temperature.
- **Raw Materials** : raw materials needed for reaction (Raw Material - 01 ,02 ,03 ,04 ,05).

So let's begin with our analysis…

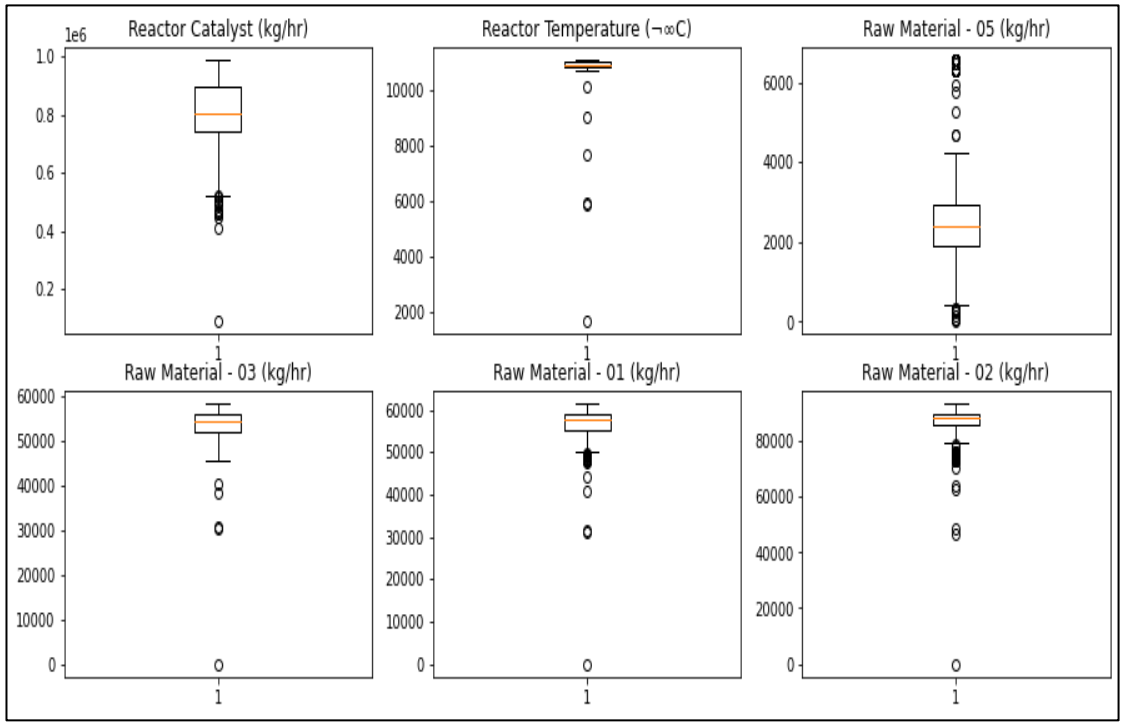
Data Cleaning and Preparation

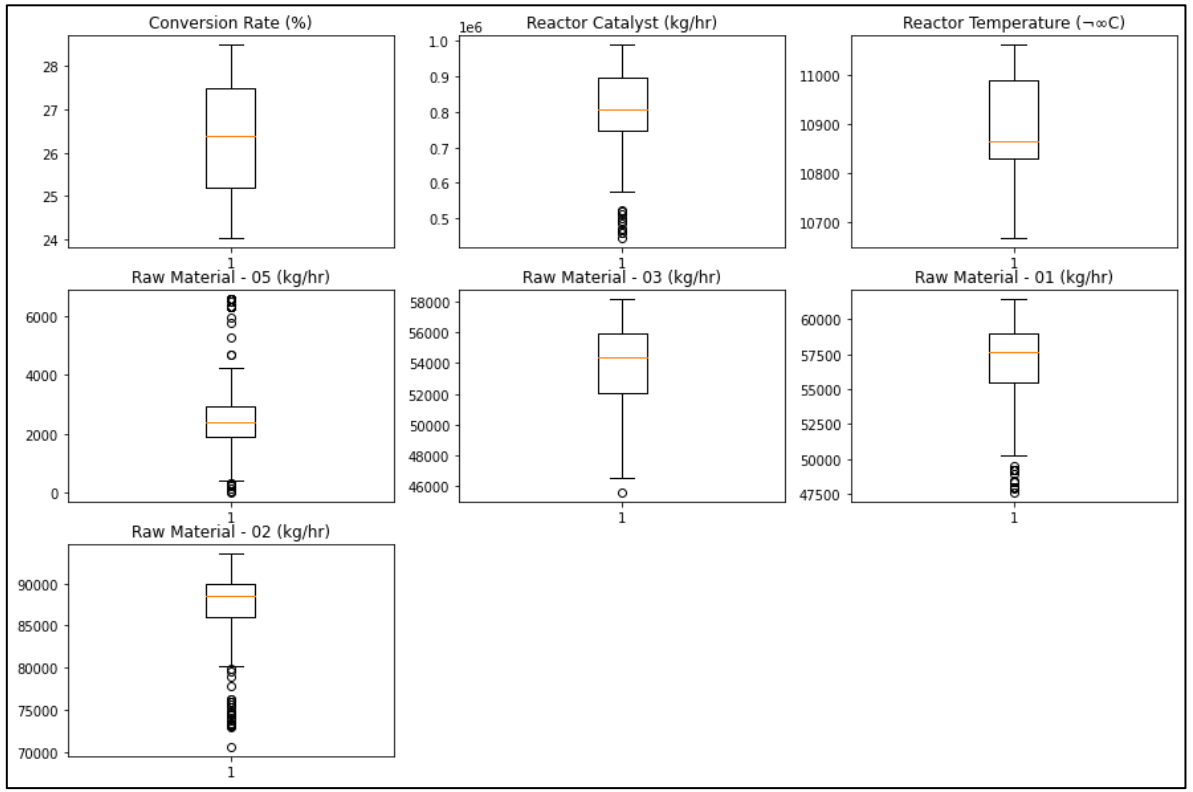
- Handling missing values : Dropped the missing values as they were very small in number(1.8%) and will not have much effect on our analysis
- Changing 'timestamp' column datatype to datetime from object.
- Dropped the **Raw material 04** column as it had all 0 values and thus it was not helping in our analysis.
- Imputed negative values of **Raw material 05** columns with mean value of the column(as quantity of raw material cannot be a negative value).
- Checked correlations. Very low correlation found with the target variable.

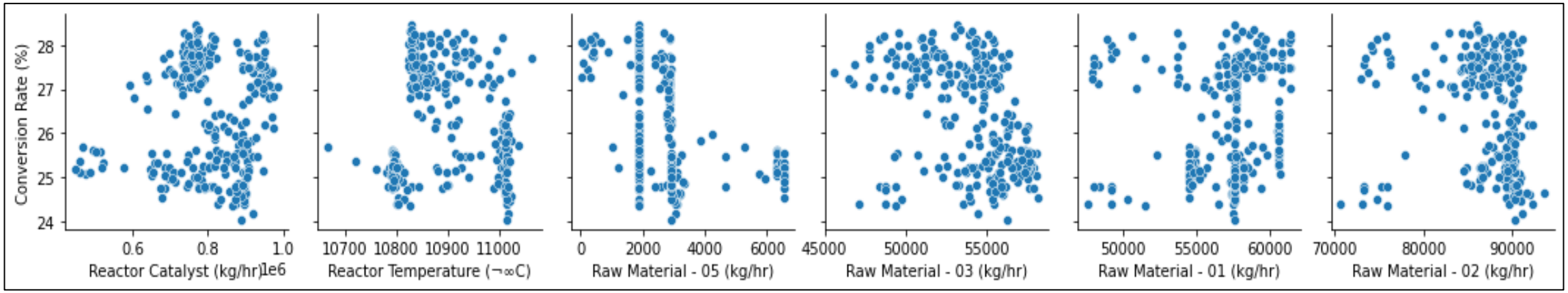


Outlier treatment

Removed outliers from the data set using Inter-quartile range







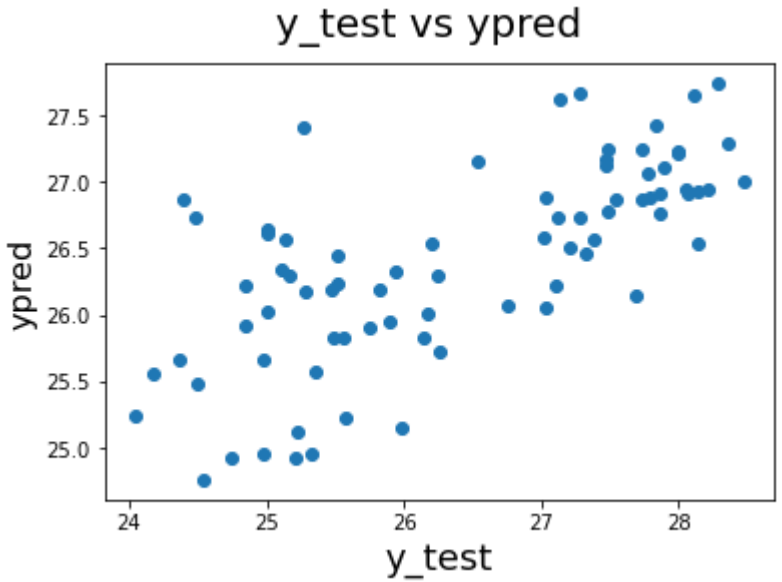
Checking relationship with our target variable

- There are not many patterns visible in the relationships affecting Conversion rate
- Few noticeable features are:
  1. Conversion rate goes down with increase in quantity of Raw material 05
  2. Reactor temperature in the range 10800 to 11000 C has higher conversion rate . Anything above or below leads to drop in the conversion rate.
  3. Reactor catalyst has some linear relationship with Conversion rate, although not very strong.

Machine learning solution

Linear Regression

- Split our data into train and test sets .
- Then , built a Linear regression model . Trained the model on Train set and made predictions on test set. But we got a very low R2 score of 0.47, which means our model was able to explain only 47% variance in the data.



- This was expected after looking at the correlations with target variable. Using Feature elimination technique we could have improved model performance , but we need to know impact of all the features here.

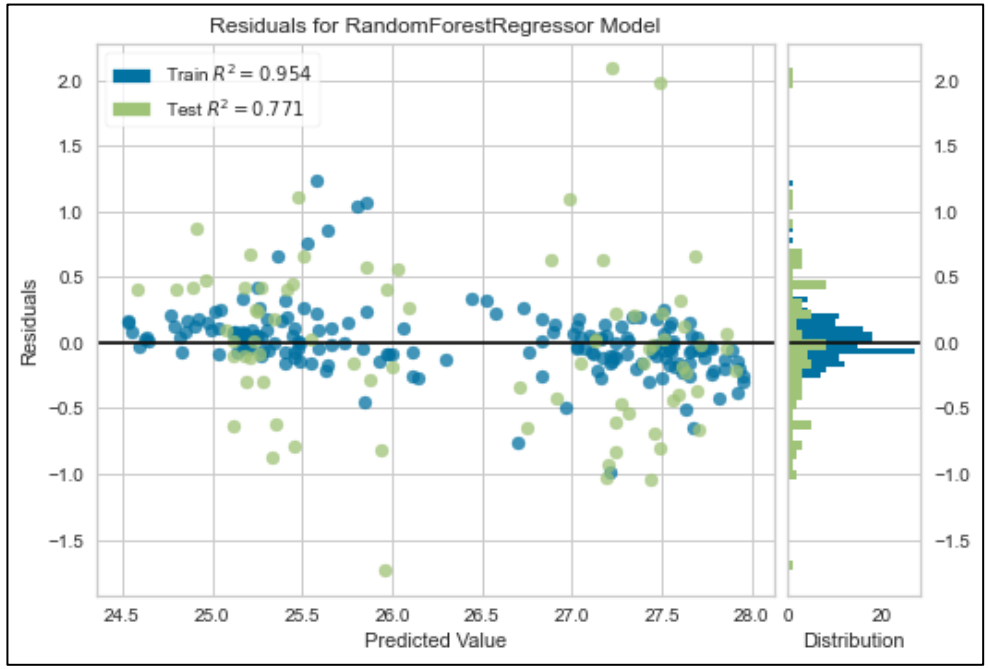
Random forest Regressor model

- Then , built a Random forest Regressor model . Trained the model on Train set and made predictions on test set.

```
print (f'Train Accuracy - : {rfm.score(X_train,y_train):.3f}')
print (f'Test Accuracy - : {rfm.score(X_test,y_test):.3f}')
```

```
Train Accuracy - : 0.954
Test Accuracy - : 0.771
```

We got train accuracy of 95%, which is very good as compared to our previous model but on test set it dropped down to 77% , which led to overfitting situation on training set. To reduce the gap, we need to peform **hyperparameter tuning** on our regressor model.



Hyper parameter tuning

- Gridsearch CV method is used to determine best parameters for our regressor model.
- We have used cross validation parameter as 10 for our model, we got our best fit parameter as :

```
1  rf_Grid.best_params_
```

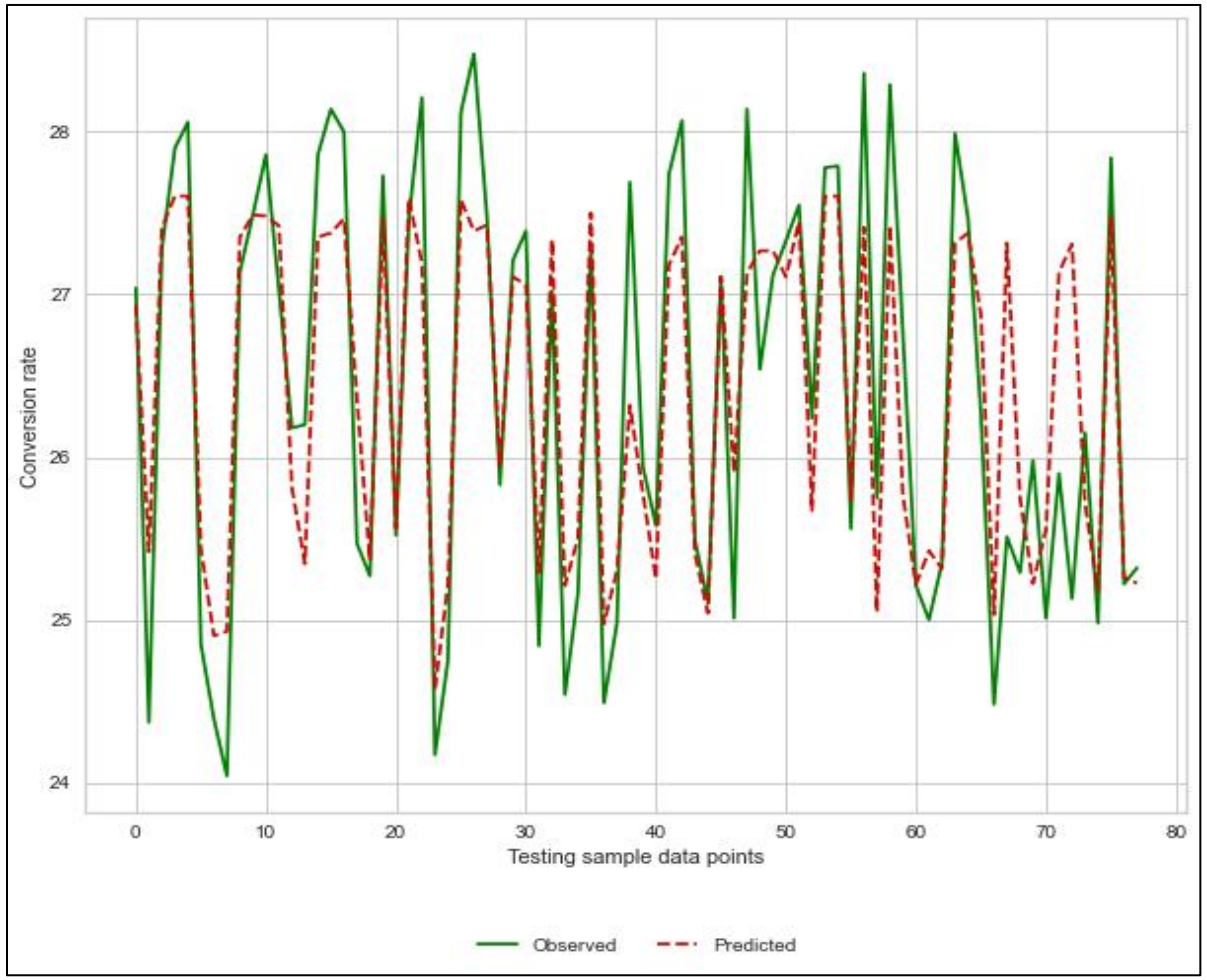
```
{'bootstrap': True,
 'max_depth': 4,
 'max_features': 'auto',
 'min_samples_leaf': 2,
 'min_samples_split': 2,
 'n_estimators': 25}
```

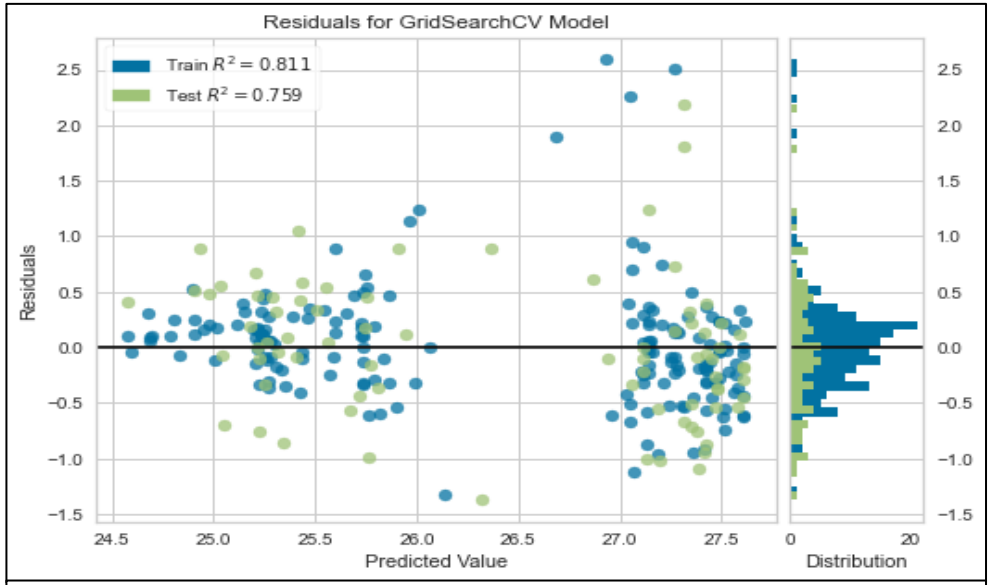
- After training the model, the accuracy on train set is 81% and on test set is 75%. The accuracy is reduced as compared to our base RF model, but we have managed to solve our problem of overfitting. We can tune the parameters more to increase the accuracy or we need to add more data to get more accuracy. However, let's now stick with this model.

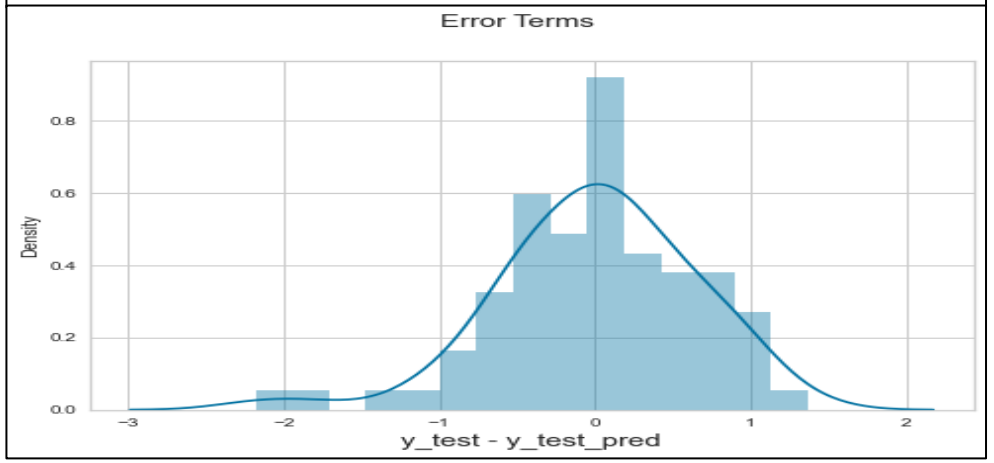
```
1  print (f'Train Accuracy - : {rf_Grid.score(X_train,y_train):.3f}')
2  print (f'Test Accuracy - : {rf_Grid.score(X_test,y_test):.3f}')
```

```
Train Accuracy - : 0.811
Test Accuracy - : 0.759
```

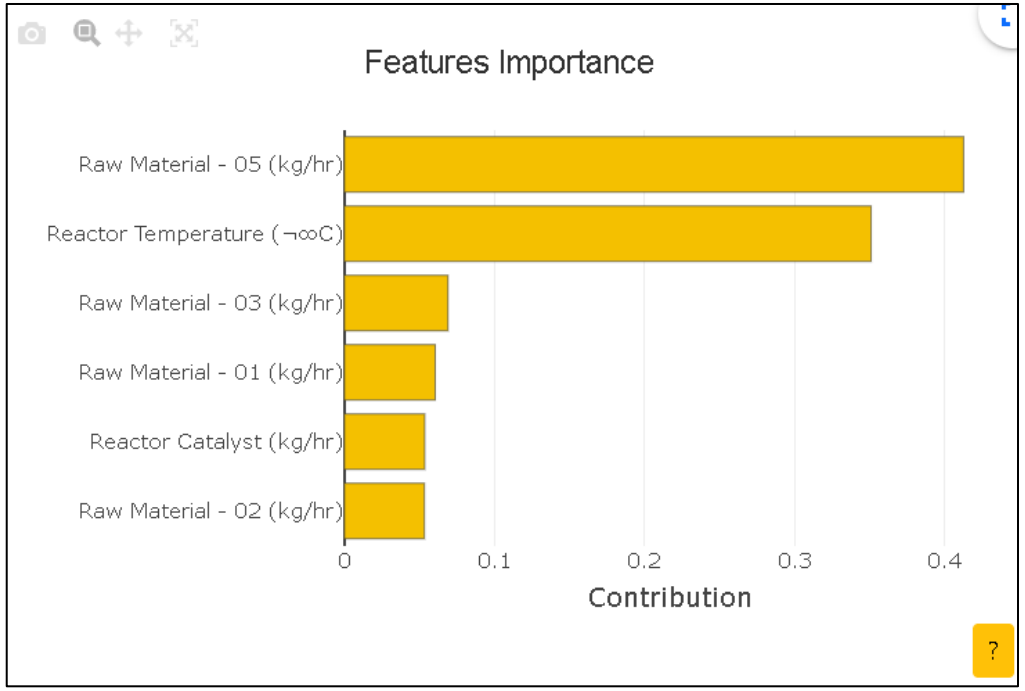
- Figure on left shows Actual vs Prediction plot on Test set. Our model has done a decent enough job on predicting Conversion rate.
- On the right, the residuals are normally distributed with mean at 0, and there are no visible patterns in the error terms. So, this model fit looks good.

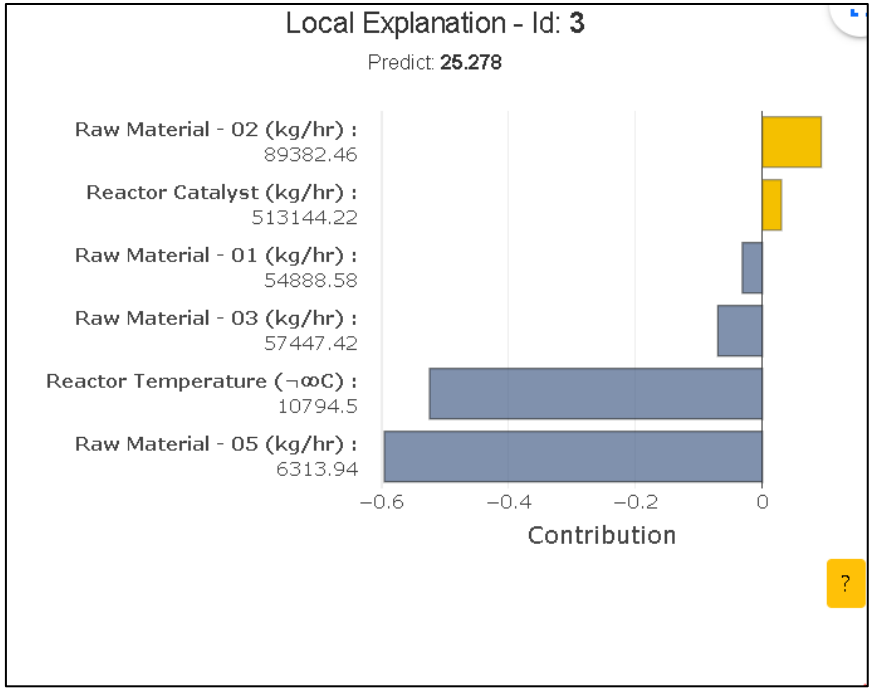






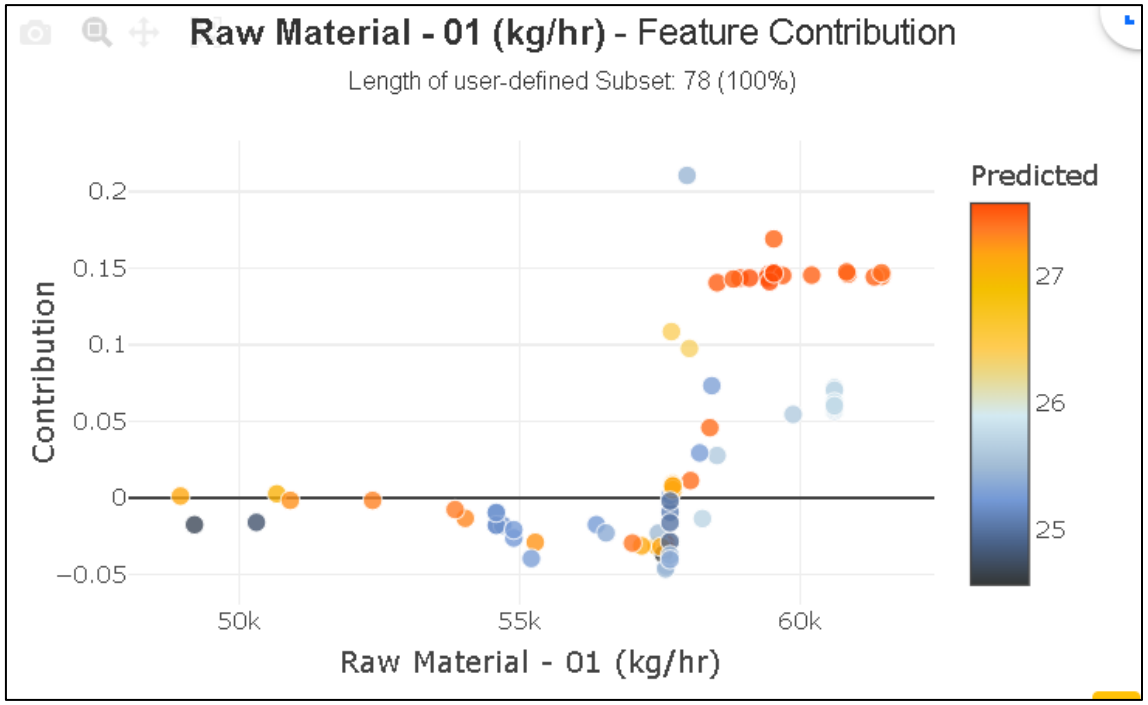
Some more Visualizations obtained from Shapash





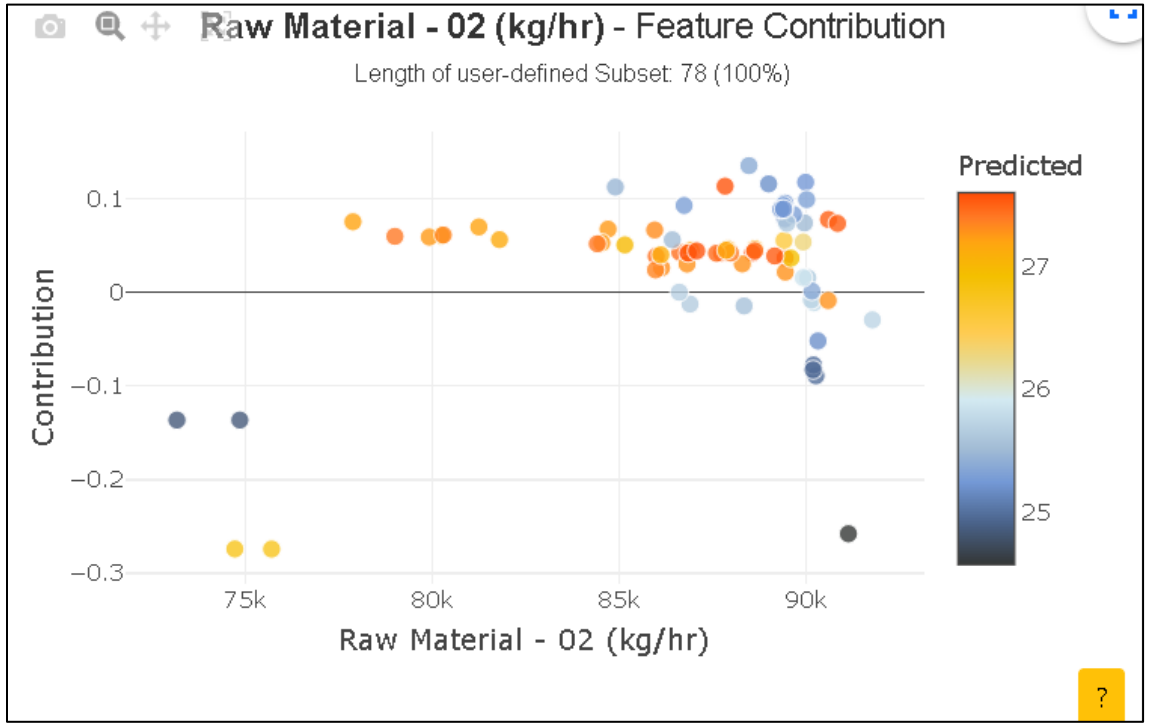
Raw material – 05 and Reactor temperature looks like the most important features. However from the second image it can be seen that they contribute negatively towards the Conversion rate. So using more is a clear NOO!

Raw Material -01 contribution



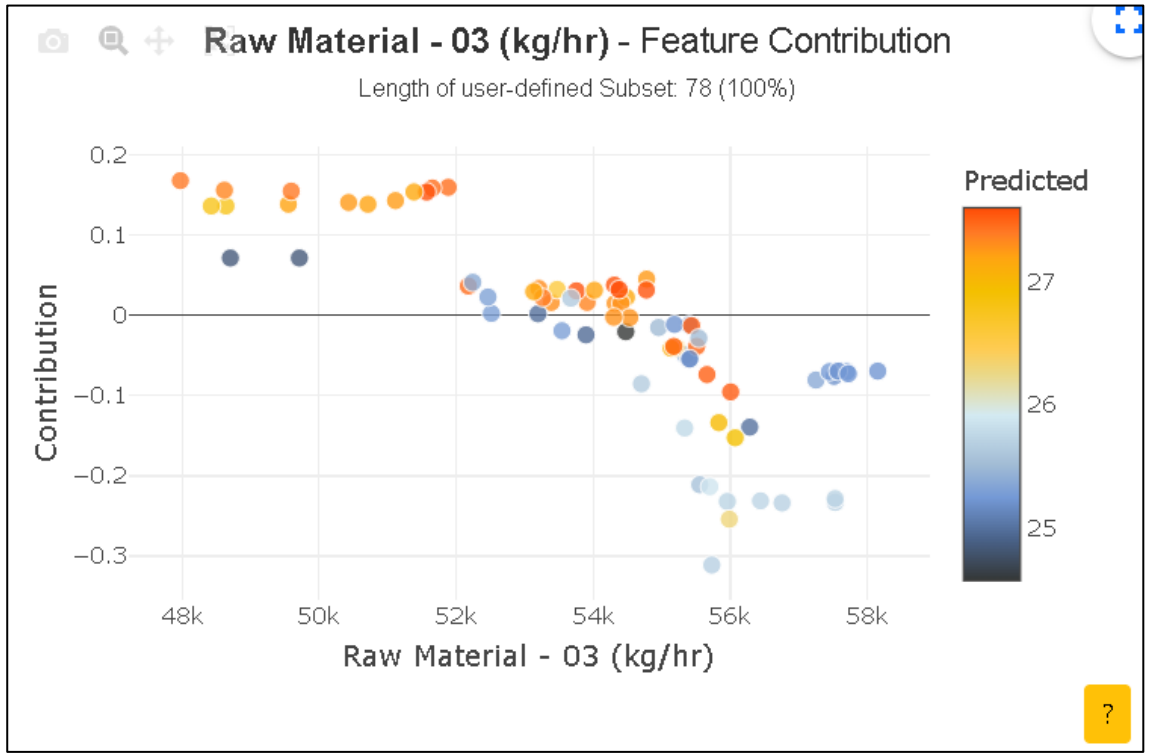
- Conversion rate increases with increase in quantity of Raw material 01.
- Raw material 01 should be used more than 57000 kg/hr for better conversion rate.

Raw Material -02 contribution



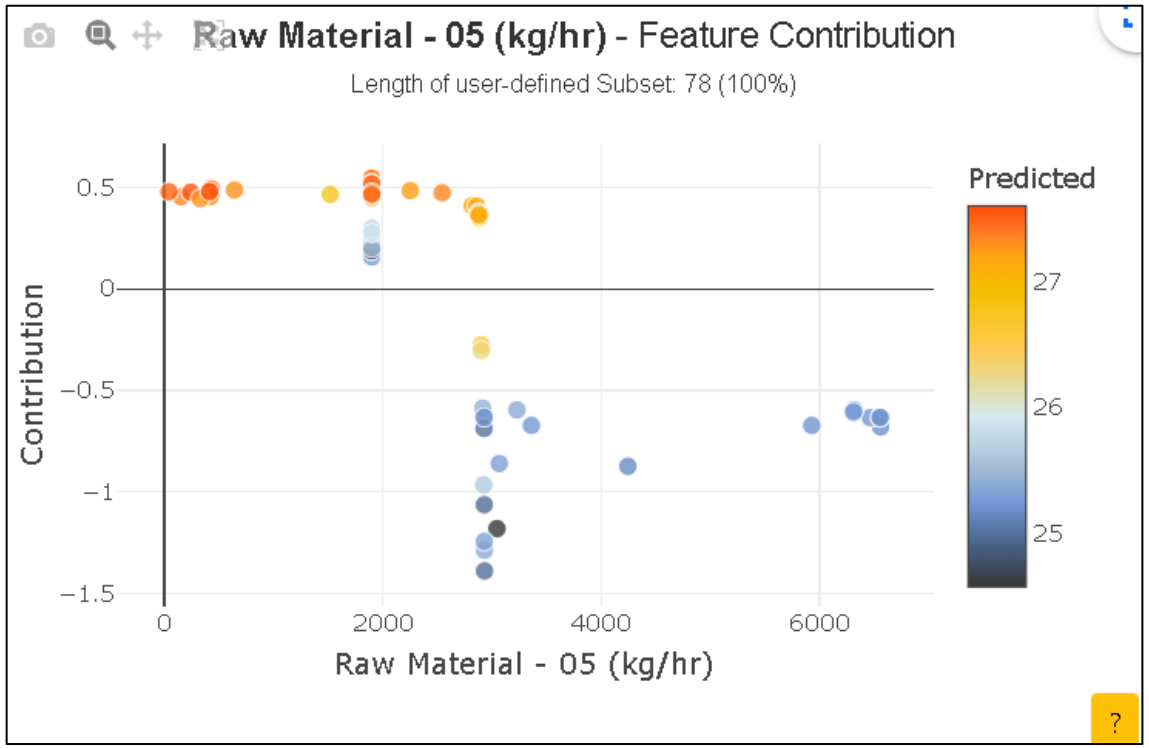
- Restrict the usage of Raw material -02 between 77000 kg/hr to 90000 kg/hr for better conversion rate

Raw Material -03 contribution



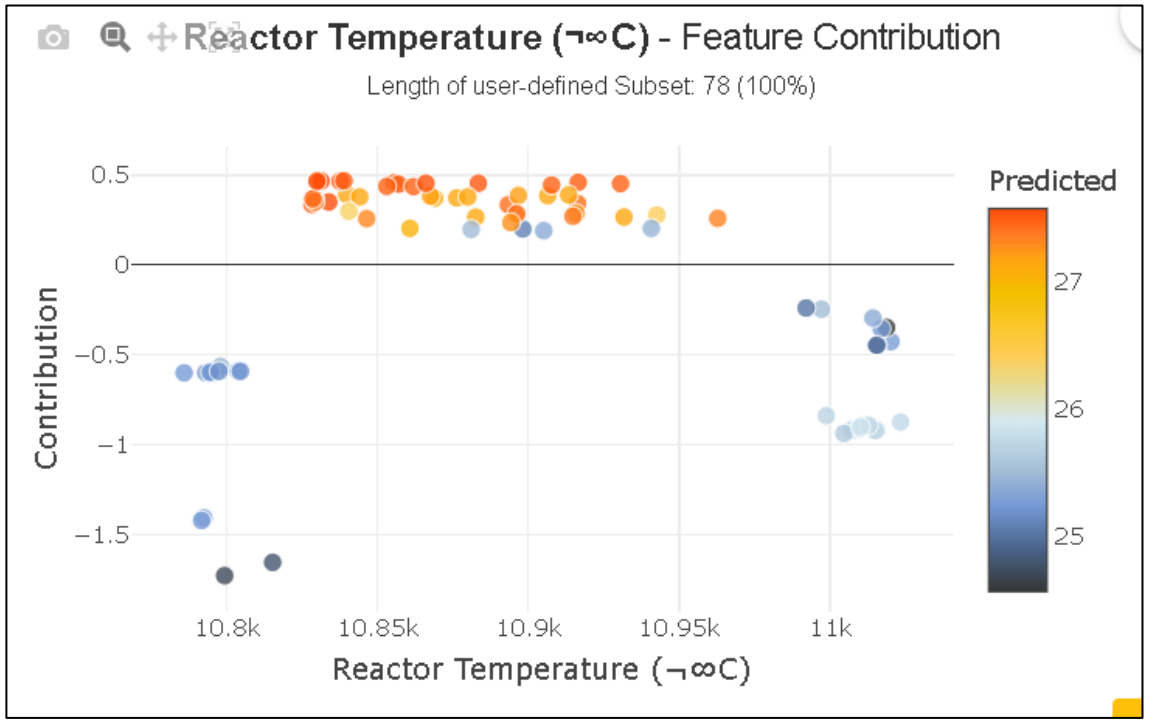
- Similarly for Raw Material 03, quantity above 52177 kg/hr leads to drop in Conversion rate.
- So we should limit our usage for Raw material -03 at this level

Raw Material -05 contribution



- We have seen earlier raw material contributes most to the conversion rate.
- However it affects negatively to the conversion rate. So adding more will not help here
- It can be seen that after quantity higher than approx. 2800 kg/hr , the conversion rate has decreased.
- So we can use this as a bar for optimum quantity of Raw material- 05 to be used.

Reactor Temperature contribution



- As we have seen from feature importance bar plot, Reactor temperature is an important factor in determining conversion rate.
- So maintaining optimum temperature is crucial for the conversion rate as well to control the amount of steam provided to plant to get higher production
- Reactor temperature should be maintained between 10.8k to 11k (¬∞C), so steam should be provided accordingly .

Summary

- Raw material 01 should be used more than 57000 kg/hr for better conversion rate.
- Restrict the usage of Raw material -02 between 77000 kg/hr to 90000 kg/hr .
- Raw Material 03's quantity above 52177 kg/hr leads to drop in Conversion rate.
- Raw Material 04 had all 0 values. Probably it was due to error in data or it was not used throughout the period. If it is used in specific conditions only, then it should be stored as per process needs.
- Raw material 05 quantity should be provided at rate of 2800 kg/hr or lesser for higher conversion of product.
- Reactor temperature should be maintained between 10.8k to 11k (¬∞C), so steam should be provided accordingly

THANK YOU !!