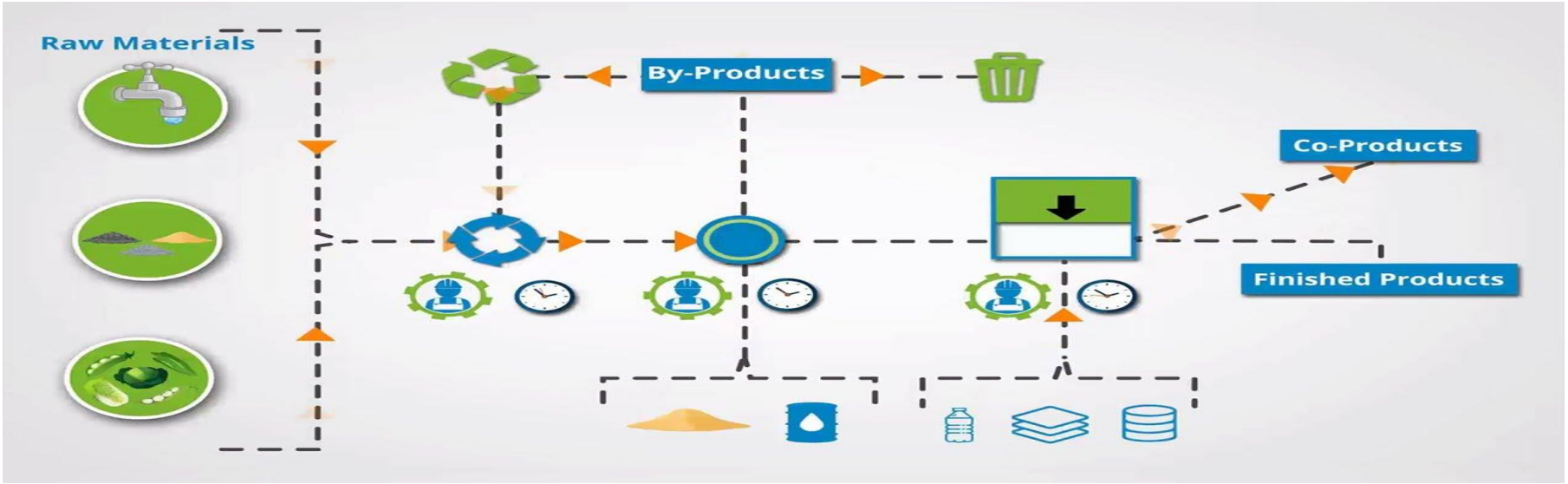


Predicting Batch Performance of Machines

Team Neoteric:

Parakrant , Namrata , Abhishek ,Somnath

Problem Statement

Data is collected from various condition monitoring sensors in a manufacturing plant.

- The **main goal** of the problem is to *find the relevant sensors which may be impacting the quality of final product.*

Stakeholders

- CEO/ CTO (Decision maker)
- Operation Head
- Operation Team
- Production Manager

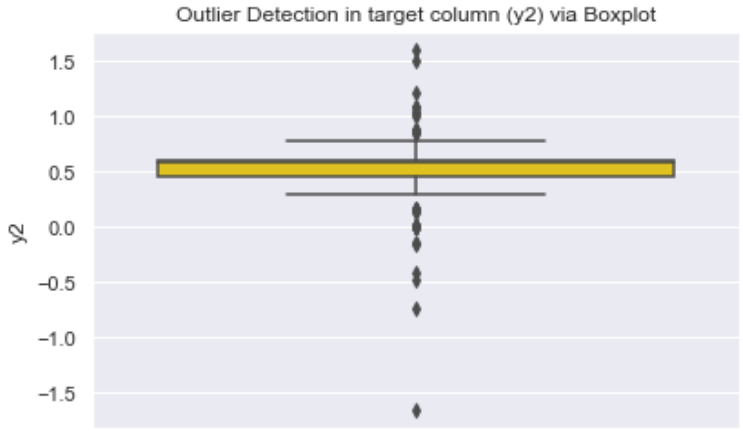
Solving the problem will help the stakeholders in getting the following business insights:

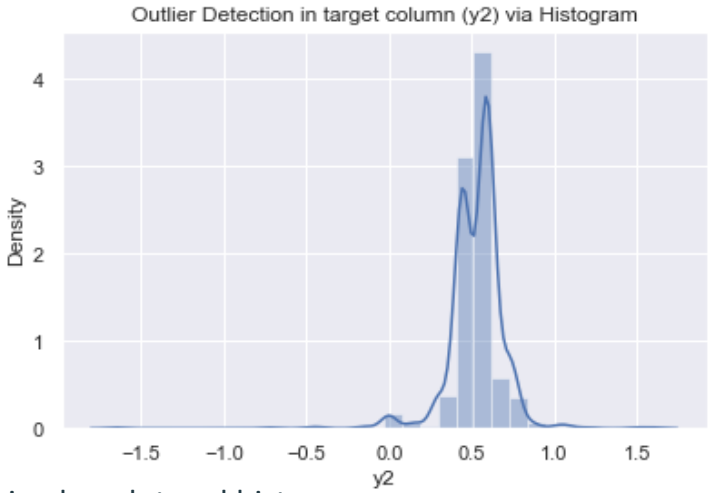
- Possibility of sensors failures.
- Maintainance of sensors.
- Quality control of product.
- Reduce the downtime of process assembly.
- Analyzing various sensors types to be used.
- Increase production line operation capability

Understanding the Data

- 54 sensors named as x0,x1…..x54.
- Each sensor records a reading at given timestamps: t0,t1,t2,t3,t4,t5 and t6.
- Output parameter y2 describes the batch performance.

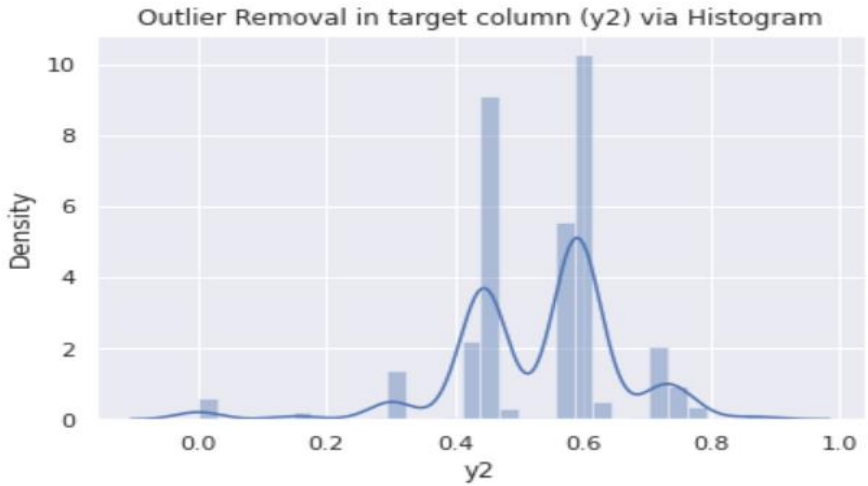
Exploratory Data Analysis (EDA)





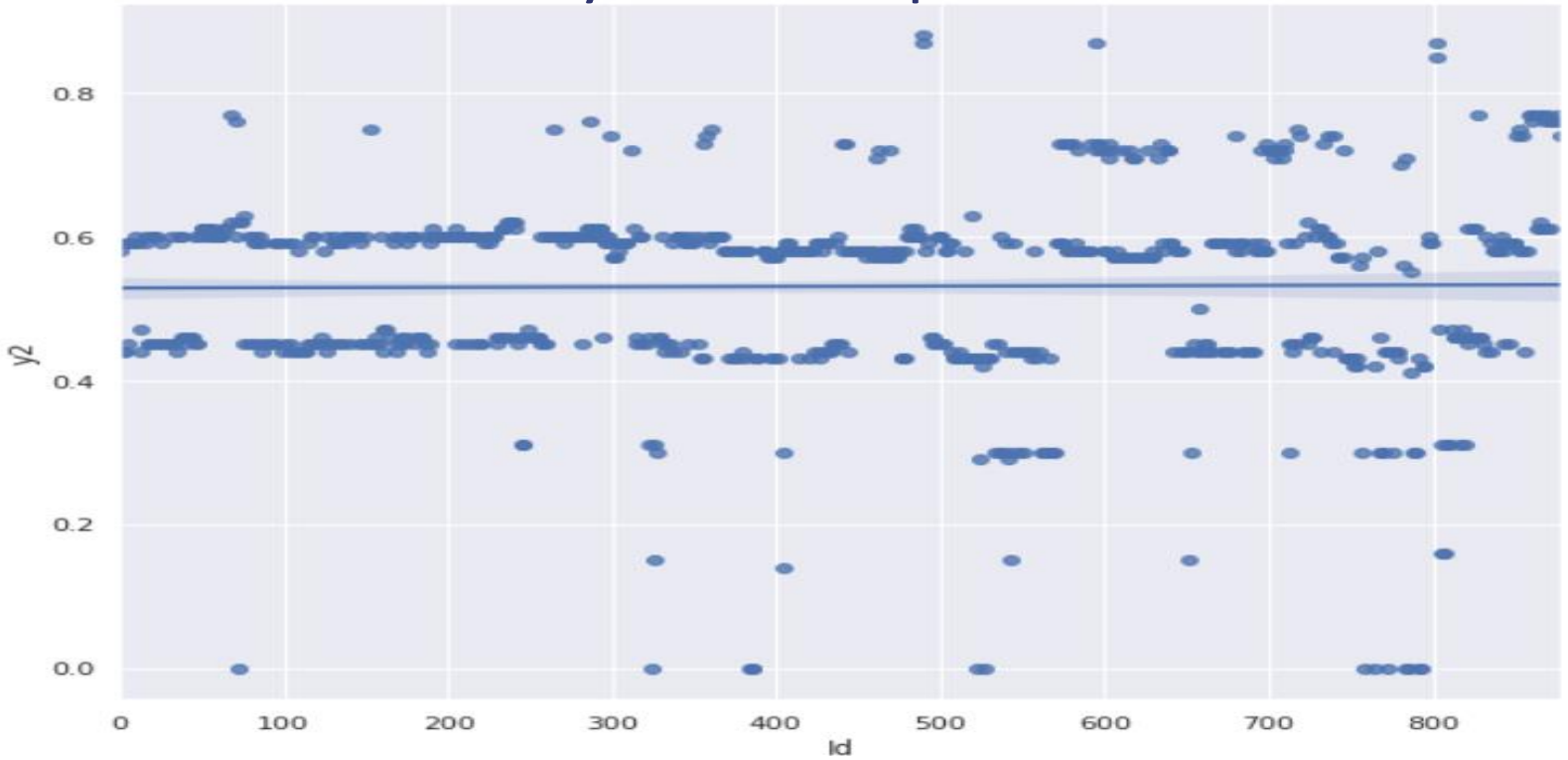
- Understanding the nature of the outliers in y2 using boxplot and histogram.

Exploratory Data Analysis (EDA)



- Considering the values in between 0.0 to 1.0.

Distribution of y2 with respect to Batch ID



Feature Engnieering:

- We considered all the features and scaled it using Robust Scalar.
- Removed the columns that showed zero variance. Further we removed the duplicate columns.
- We calculated the *VIF values* for all the features.
- Feature selection is carried out by thresholding various *VIF* values such as 10, 8, 6, 5, 4 etc.
- VIF value greater than 10 shows high correlation with the target value. Hence we removed it.
- We got best results for the features having VIF values in between **8** and **4**. (4<=VIF=<8)

Experimentation:

- Machine Learning models for regression analysis.
  - Linear Regression
  - Lasso
  - Ridge
  - Decision Trees
  - Random Forest
  - XGBoost
  - SVM

Results

R2 Scores	Baseline	Dropping Zero Variance columns	Dropping VIF > 10	All Sensors at timestamp t0	All Sensors at timestamp t2	All Sensors at timestamp t6
Models						
Linear Regression	-1319.907915	-0.712915	-5.5216	0.06	-0.06	-0.90
Lasso	-0.1050881123	-0.016240	-0.032	-0.02	-0.02	-0.09
Ridge	-2.31495	-0.428700	-0.4093	0.06	-0.05	-0.85
Decision Tree Reg	-0.01893	-0.448500	-0.9139	-0.86	-2.08	-0.09
Random Forest Reg	0.34864	0.365620	0.3906	0.16	0.12	0.03
Xgboost	0.14893	0.067700	0.24952	0.14	0.08	-0.01
SVM	-0.02951	0.056220	-0.024	0.07	-0.02	-0.12

Conclusion and Future Work

- We got best results for the features having VIF values in between 8 and 4. (4<=VIF=<8) for Linear Regression with R2 score of -**0.59**
- We tried with PCA analysis but the results were not satisfactory
- In Future work, we can look into the time series analysis of the sensor data collected at various time stamps.
- The models used in our experiments can be further tuned to achieve better accuracy.