DAASA HACKATHON (11-04-23)

TEAM ML MAVERICKS

PROBLEM - I (BAD CYCLE PREDICTION)

TEAM MEMBERS:

1. BHUVVAAN CHANDRA (LEAD)
2. TRINAY GANGISETTY
3. SIDDHARTH KALYANASUNDARAM
4. ANUDEEP NAYAK

PROBLEM STATEMENT:

Objective: Develop a predictive maintenance system for aluminum manufacturing.

Issue: Poor maintenance timing leads to bad furnace cycles, causing costly downtime and low-quality metal.

Solution: Create a model to predict bad cycles and schedule maintenance.

Benefits: Reducing downtime, minimizing maintenance costs, and improving metal quality.

Evaluation: Measure success using AUC-ROC score and Accuracy of the Model.

SAMPLE DATA:

Period Code	Cycle ID	B_2	B_3	B_4	B_5	B_9	B_10	B_14	B_15	B_16	B_17	B_18	B_19	B_20	B_21	B_22	B_23	B_24	B_25	Good/Bad	timestamp
1	1	-0.0007	-0.0004	100	518.67	14.62	21.61	1.3	47.47	521.66	2388.02	8138.62	8.4195	0.03	392	2388	100	39.06	23.419	0	3/1/2020 0:00
1	2	0.0019	-0.0003	100	518.67	14.62	21.61	1.3	47.49	522.28	2388.07	8131.49	8.4318	0.03	392	2388	100	39	23.4236	0	3/1/2020 0:05
1	3	-0.0043	0.0003	100	518.67	14.62	21.61	1.3	47.27	522.42	2388.03	8133.23	8.4178	0.03	390	2388	100	38.95	23.3442	0	3/1/2020 0:10
1	4	0.0007	0	100	518.67	14.62	21.61	1.3	47.13	522.86	2388.08	8133.83	8.3682	0.03	392	2388	100	38.88	23.3739	0	3/1/2020 0:15
1	5	-0.0019	-0.0002	100	518.67	14.62	21.61	1.3	47.28	522.19	2388.04	8133.8	8.4294	0.03	393	2388	100	38.9	23.4044	0	3/1/2020 0:20
1	6	-0.0043	-0.0001	100	518.67	14.62	21.61	1.3	47.16	521.68	2388.03	8132.85	8.4108	0.03	391	2388	100	38.98	23.3669	0	3/1/2020 0:25
1	7	0.001	0.0001	100	518.67	14.62	21.61	1.3	47.36	522.32	2388.03	8132.32	8.3974	0.03	392	2388	100	39.1	23.3774	0	3/1/2020 0:30
1	8	-0.0034	0.0003	100	518.67	14.62	21.61	1.3	47.24	522.47	2388.03	8131.07	8.4076	0.03	391	2388	100	38.97	23.3106	0	3/1/2020 0:35
1	9	0.0008	0.0001	100	518.67	14.62	21.61	1.3	47.29	521.79	2388.05	8125.69	8.3728	0.03	392	2388	100	39.05	23.4066	0	3/1/2020 0:40
1	10	-0.0033	0.0001	100	518.67	14.62	21.61	1.3	47.03	521.79	2388.06	8129.38	8.4286	0.03	393	2388	100	38.95	23.4694	0	3/1/2020 0:45
1	11	0.0018	-0.0003	100	518.67	14.62	21.61	1.3	47.15	521.4	2388.01	8140.58	8.434	0.03	392	2388	100	38.94	23.4787	0	3/1/2020 0:50
1	12	0.0016	0.0002	100	518.67	14.62	21.61	1.3	47.18	521.8	2388.02	8134.25	8.3938	0.03	391	2388	100	39.06	23.366	0	3/1/2020 0:55
1	13	-0.0019	0.0004	100	518.67	14.62	21.61	1.3	47.38	521.85	2388.08	8128.1	8.4152	0.03	393	2388	100	38.93	23.2757	0	3/1/2020 1:00
1	14	0.0009	0	100	518.67	no response	21.61	1.3	47.44	521.67	2388	8134.43	8.3964	0.03	393	2388	100	39.18	23.3826	0	3/1/2020 1:05
1	15	-0.0018	-0.0003	I/O	518.67	14.62	21.61	1.3	47.3	start	2388.08	8127.56	8.4199	0.03	Missing	2388	100	38.99	23.35	0	3/1/2020 1:10

DATA EXPLORATION (STEP - 1):

1. The first step and the most important step is understanding the columns.
2. We have explored a few insights from the data and they are as follows:
    1. We have identified the target column, quantitative, qualitative variables
    2. We have identified all the columns which had non-numerical text data
    3. We looked into the data set for any missing values
    4. We looked into the statistical analysis for all the numerical columns
    5. We further looked if the data in the columns is normally distributed or if the data is skewed
    6. We have identified that Period and Cycle goes together and cycle starts newly for each period again
    7. We have identified that duration between each cycle is 5 mins
    8. We have identified that there is no overlapping between any two periods and a new period starts only after the previous period is over and not simultaneously.

DATA CLEANING (STEP - 2):

1. **The next step is data cleaning**

   a. Non - numerical text data - > NaN

   b. NaN -> imputed the mean / median values with help of (Kolmogorov - Smirnov) test to check for the "Normality of the data".

   c. If P <= 0.05 -> Median else Mean

   d. We have converted the data types of the numerical columns from object -> float

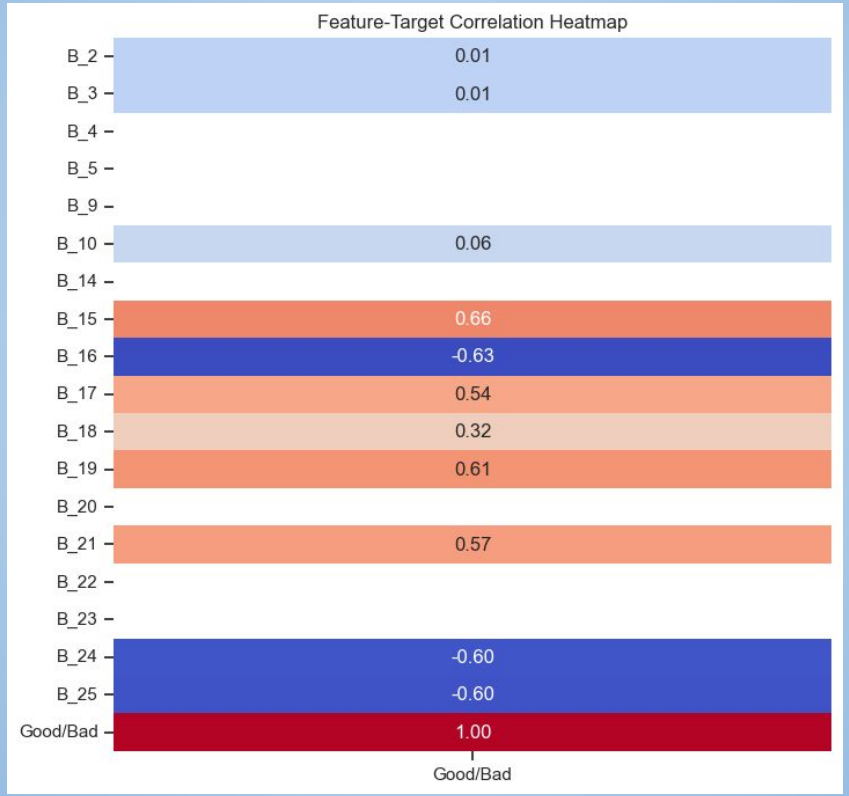
   e. Grouped the values by period -> added +1 to the previous value of missing cycle value and replaced it.

FEATURE ENGINEERING:

We decided to use only a few sensors that are highly correlated and have high variation, and drop the rest of the features. So we explored the same using several methods

1. **RFE - Recursive Feature Elimination**
2. **VIF - Variation Influence Factor**
3. **ANOVA (Analysis of Variance) F - Test**

Correlation Matrix



MODELLING:

ALGORITHM	ACCURACY
RANDOM FOREST CLASSIFIER	91.12%
XGBOOST	95.60%

Why XGBOOST?

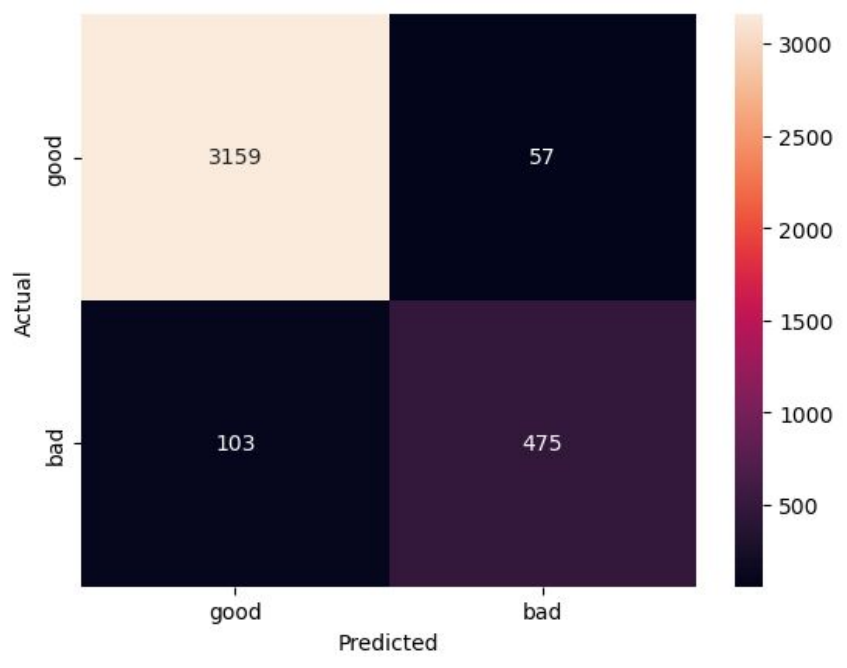
1. **XGBoost is a great algorithm for classification and regression problems.**
2. **It has inbuilt feature selection capability**
3. **Really good for imbalanced datasets such as ours.**
4. **Lots of hyperparameter tuning possibilities.**

Hypertuning:

1. Objective - Logistic
2. Parameter Grid Search
   a. Learning rate - step size learning - 0.02
   b. N_estimators - no of base learners - [150, 300, 450]
   c. Max_depth - height of DT- [3,4,5]

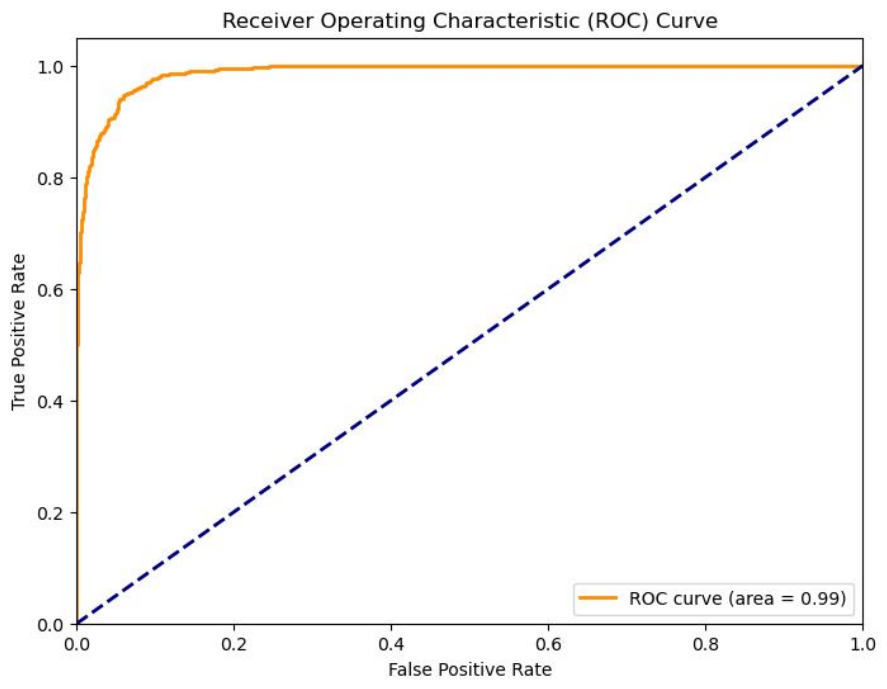
Total 9 possible combinations, 300 and 5 being the best

Confusion Matrix



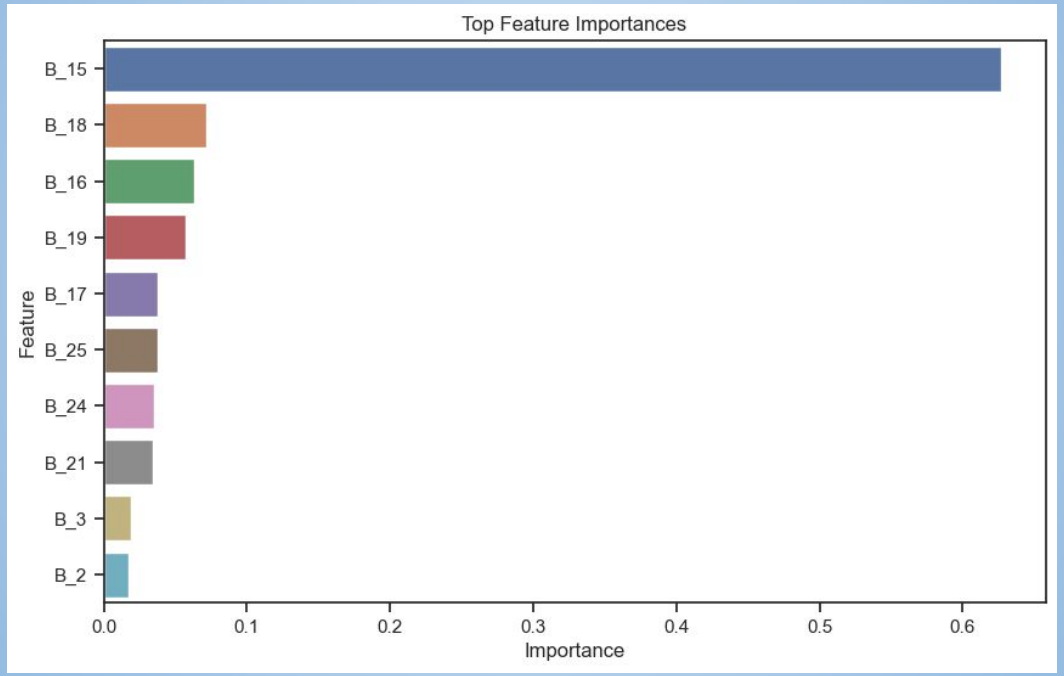
- The train data set is further split into test and train data sets to estimate the accuracy of the model.
- 75% of the data set is used for training and 25% of the data set is used for testing.
- This leads to a test data set of ~12000.

ROC Curve



- The train data set is further split into test and train data sets to estimate the accuracy of the model.
- 75% of the data set is used for training and 25% of the data set is used for testing.
- This leads to a test data set of ~12000.

Feature Extraction



THANK YOU! :)