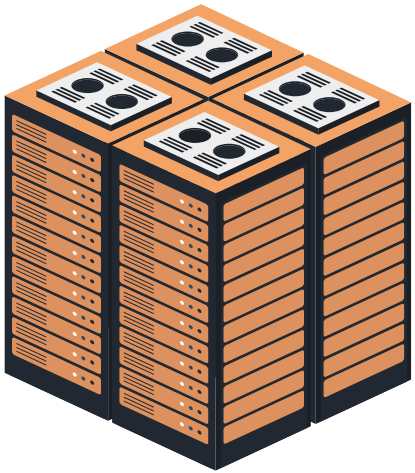
CLASSIFICATION TECHNIQUES

CU Boulder, DaSSA Hackathon 2

Team Members:
Sasi Bonu, Aman Rathi, Divya Nallawar, Taahaa Dawe, Ghizlane Rehioui

Table of Contents

About the Data



Data is about Aluminum decoating.

Timely alarms of poor quality decoating cycles are critical and potentially **avoid business losses**.

Train data

.csv file with 22 columns and a total of 15,175 rows

Test data

.csv file with 21 columns and a total of 5059 rows

Content

timestamp, Period Code, Cycle ID, Censor B_index (18 censors), and Good/Bad.

Objectives

1. Determine maintenance time based on the model predictions.
2. Build a classification model with good accuracy.
3. Better predictions help avoid business losses.

Data Pre-processing & Cleaning

Step 1

Sort the dataframe in ascending order by timestamp

Step 2

Correcting misplaced values of Period Code using order of time

Step 3

Correcting misplaced values of Cycle ID using sequential order

Step 4

Replacing nulls and misplaced by average and removing columns with repeated rows

Data Pre-processing & Cleaning

100% of the 15,175 rows retained

14 out of 21 columns retained

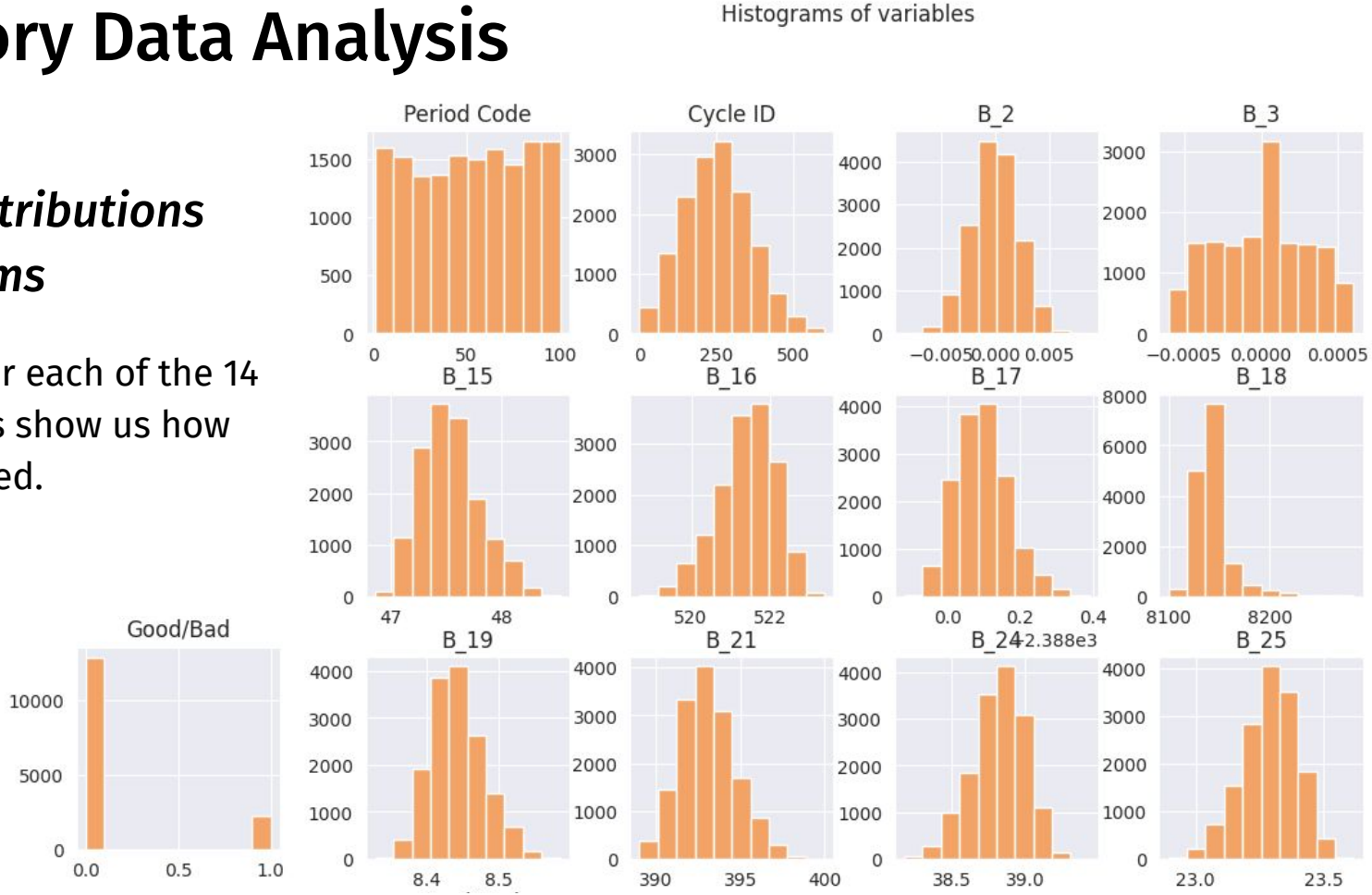
No missing/null values & correctly formatted columns

	Period Code	Cycle ID	B_2	B_3	B_15	B_16	B_17	B_18	B_19	B_21	B_24	B_25	timestamp	Good/Bad
0	1	1	-0.0007	-0.0004	47.47	521.66	2388.02	8138.62	8.4195	392.0	39.06	23.4190	1/3/2020 00:00	0
1	1	3	-0.0043	0.0003	47.27	522.42	2388.03	8133.23	8.4178	390.0	38.95	23.3442	1/3/2020 00:10	0
2	1	4	0.0007	0.0000	47.13	522.86	2388.08	8133.83	8.3682	392.0	38.88	23.3739	1/3/2020 00:15	0

Exploratory Data Analysis

1. Numeric Distributions with Histograms

The histograms for each of the 14 retained variables show us how they are distributed.



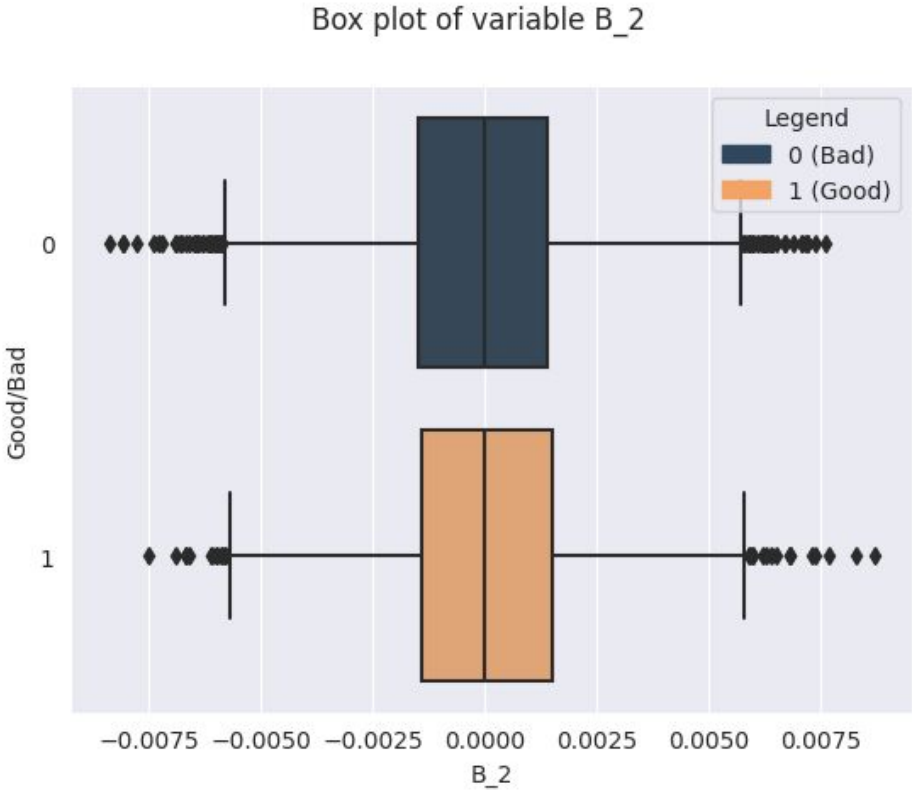
Exploratory Data Analysis

2. Numeric Distributions with Boxplots

Box plots allow us to visualize python's *describe* function, the characteristics of data, and locate outliers.

```
df.describe()
```

	Period Code	Cycle ID	B_2	B_3
count	15174.000000	15174.000000	15174.000000	15174.000000
mean	51.287663	253.754382	-0.000029	0.000002
std	29.287305	109.467632	0.002168	0.000291
min	1.000000	1.000000	-0.008400	-0.000600



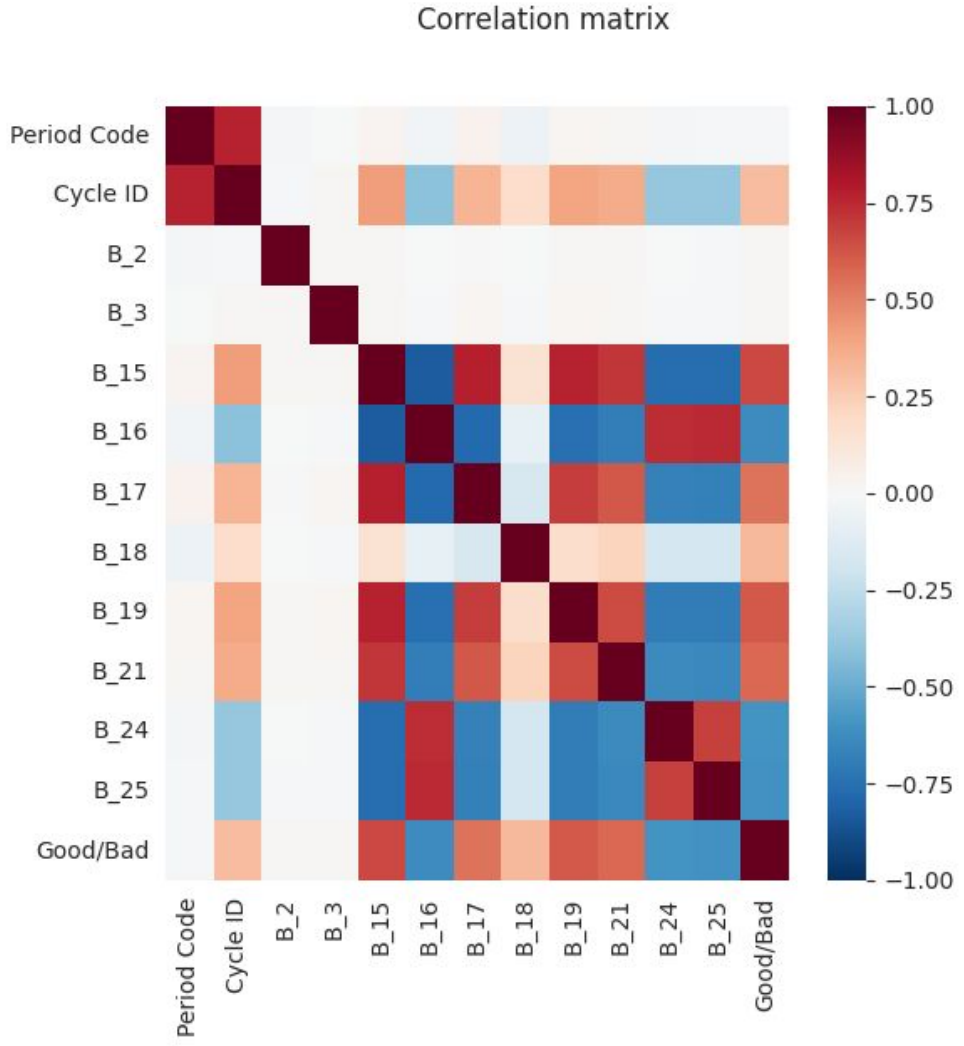
3. Interactions between numeric features with Correlation Matrix

Highly positively correlated

B15 - B16
B16 - B17
B24 - B15
B25 - B15

Highly negatively correlated

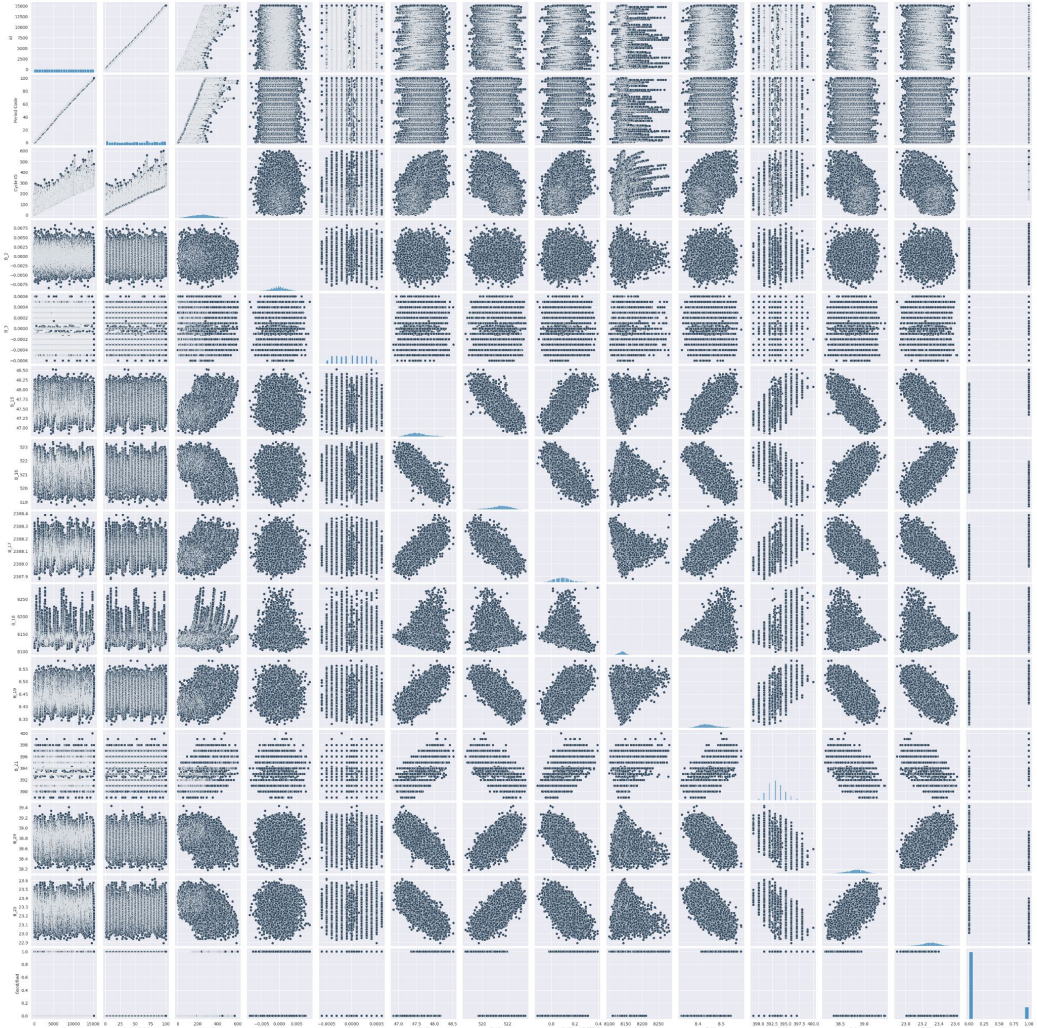
B15 - B17
B15 - B19



4. Pair plots

Pair plots visualize the relationship between variables.

Looking at this plot, we can identify which variables might be correlated and in which direction.



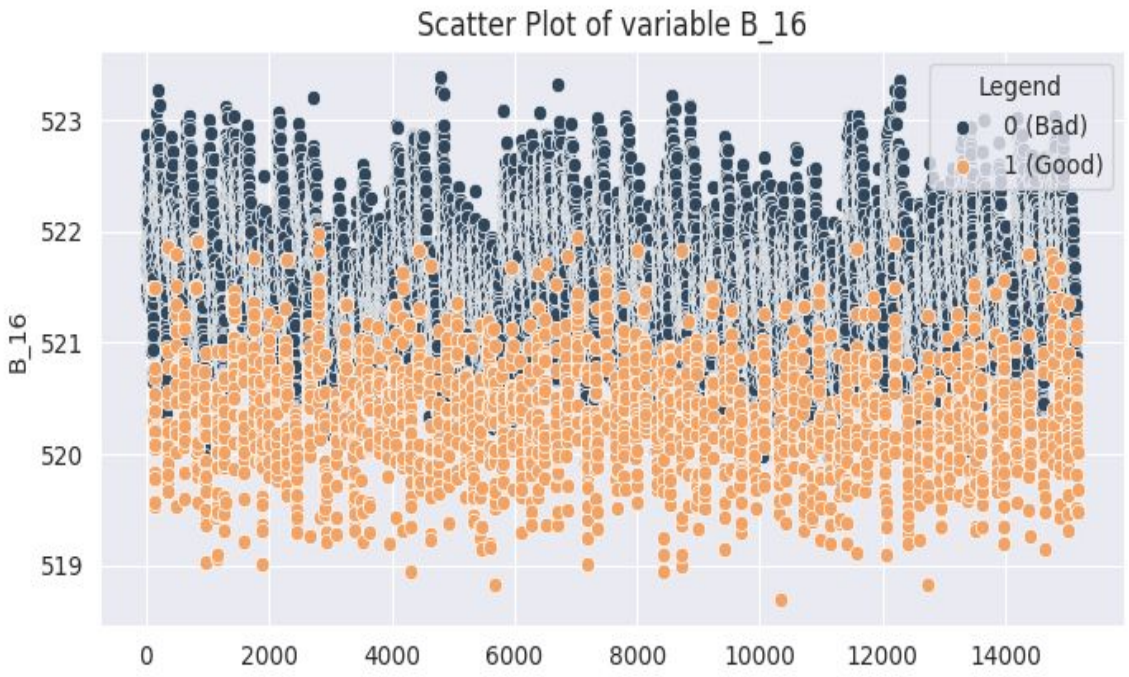
5. Scatter plots

This plot allows us to compare how high the values of B_16 are for the case of 0 (Bad) and 1 (Good).

We can also say that Good occurs less frequently (less points) than Bad for B_16.

In fact:

Bad	**Good**
12,887	2287



Modeling

Binary Time Series Classification

Random Forest

Binary Time Series Classification

Definition

We have labeled time series data and we want to classify data into different categories or states.

Why?

We have labeled time series data and a binary variable *Good/Bad*.

Results

```
DTC = DecisionTreeClassifier()
DTC.fit(features_filtered_direct, y_train)
```

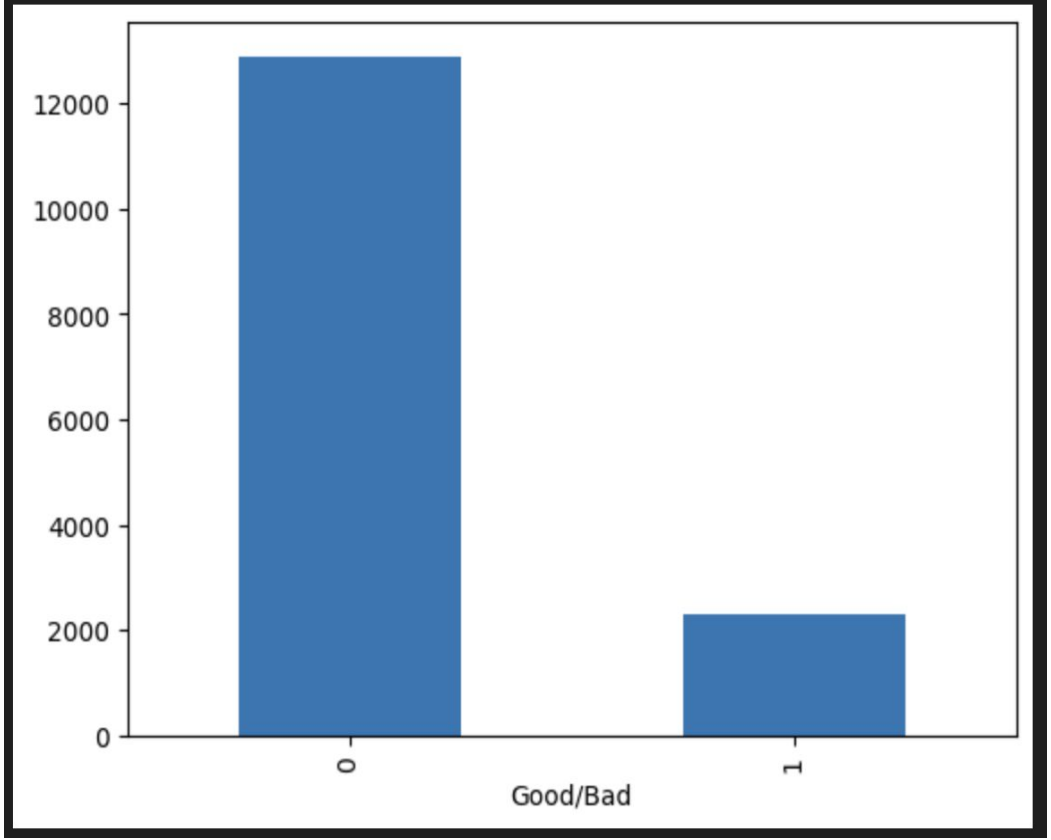
▾ DecisionTreeClassifier
DecisionTreeClassifier()

```
KNN = KNeighborsClassifier(n_neighbors=5)

KNN.fit(features_filtered_direct,y_train)
```

▾ KNeighborsClassifier
KNeighborsClassifier()

```
target["Good/Bad"].value_counts().plot(kind="bar")
plt.show()
```



Random Forest

Definition

It is an ML algorithm used in time series analysis for forecasting and anomaly detection.

Why?

We have a mix of different types of features (temporal and non-temporal).

Results

Accuracy: 0.9447312481020346

Classification Report:

	precision	recall	f1-score	support
0	0.96	0.98	0.97	2757
1	0.87	0.84	0.86	482
Bad	0.00	0.00	0.00	6
I/O	0.00	0.00	0.00	9
Missing	0.00	0.00	0.00	9
No Data	0.00	0.00	0.00	6
no connection	0.00	0.00	0.00	7
no response	0.00	0.00	0.00	4
start	0.00	0.00	0.00	7
stop	0.00	0.00	0.00	6
accuracy			0.94	3293
macro avg	0.18	0.18	0.18	3293
weighted avg	0.93	0.94	0.94	3293

Thank you!

CU Boulder, DaSSA Hackathon 2

Team Members:
Sasi Bonu, Aman Rathi, Divya Nallawar, Taahaa Dawe, Ghizlane Rehioui