

# Scania Engine Fault Prediction

## Introduction:

The Air Pressure System (APS) is an essential part of a heavy duty vehicle, where compressed air causes the piston to apply pressure on the brake pads to slow down the vehicle. The advantages of having an APS instead of a hydraulic setup is the easy availability and sustainability of air from nature.

The dataset consists of data collected from heavy Scania trucks in everyday usage. These are failure cases of the trucks during operation, and our task is to predict whether a given failure is caused due to a specific component of the Air Pressure System. This may help in avoiding failure during truck operation and thereby reducing maintenance cost.

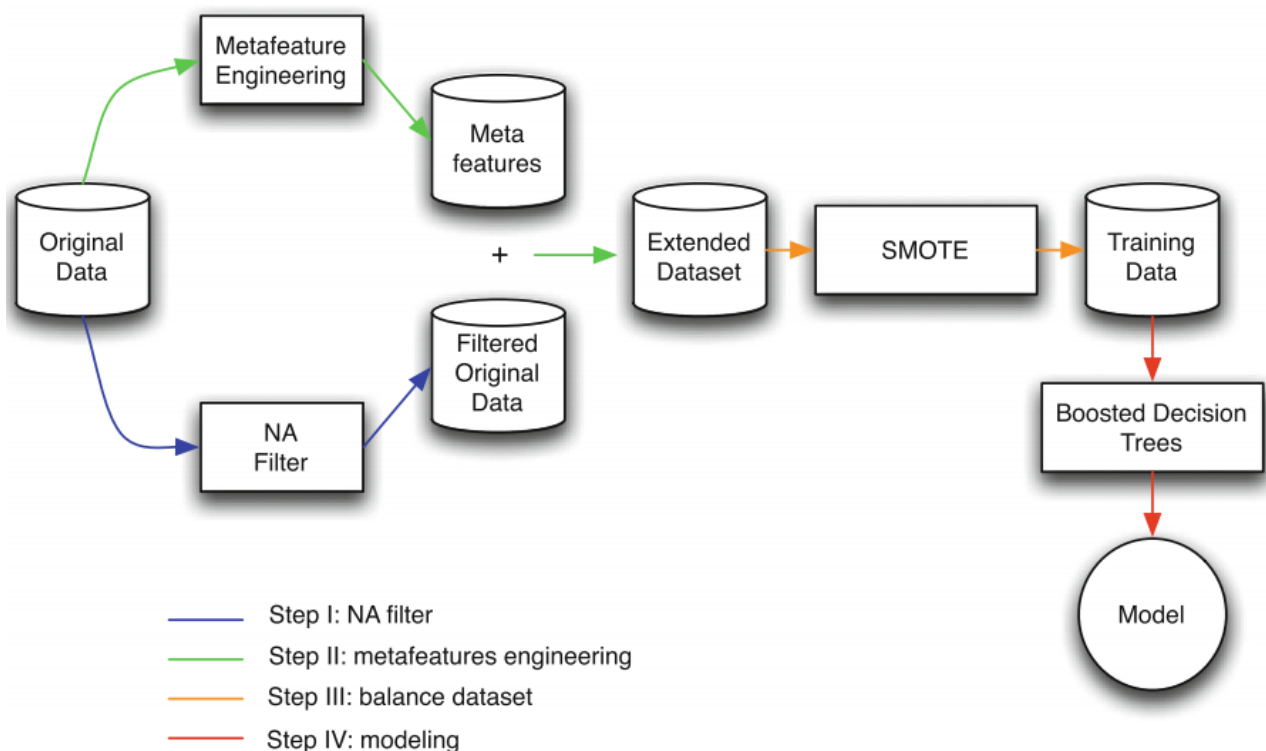
## Objective:

This is a Binary Classification problem where the positive class tells us that the failure was due to a specific component of the APS, whereas, the negative class tells us that the failure has nothing to do with that component. Therefore, given a new data point (sensor information), we can build an ML model that would tell us if the failure was due to the truck's APS or not.

## Benefits:

- ◆ Latency must be fairly low, to detect a failure in the APS and avoid increase in maintenance cost.
- ◆ Cost of misclassification is very high since an undetected failure of the APS component can lead to failure of the truck during operation and therefore an increase in maintenance cost.

## Methodology:



## Data Sharing Agreement:

- Sample file name (ex APS\_20162021\_112)
- Length of data stamp(8 digits)
- Length of time stamp(6 digits)
- Number of Columns
- Column names
- Column datat type

## Data Description:

- The dataset consists of data collected from heavy Scania trucks in everyday usage. The system in focus is the Air Pressure system (APS) which generates pressurised air that are utilized in various functions in a truck, such as braking and gear changes. The datasets' positive class consists of component failures for a specific component of the APS system. The negative class consists of trucks with failures for components not related to the APS. The data consists of a subset of all available data, selected by experts.
- The attribute names of the data have been anonymized for proprietary reasons.
- Apart from training files, we also require a "schema" file from the client, which contains all the relevant information about the training files such as:
- Name of the files, Length of Date value in FileName, Length of Time value in FileName, Number of Columns, Name of the Columns, and their datatype.

## Data Validation and Data Transformation:

### ➤ **Name Validation:**

We validate the name of the files based on the given name in the schema file. We have created a regex pattern as per the name given in the schema file to use for validation.

If all the values are as per requirement, we move such files to "Good\_Data\_Folder" else we move such files to "Bad\_Data\_Folder."

### ➤ **Number of columns:**

We validate the number of columns present in the files, and if it doesn't match with the value given in the schema file, then the file is moved to "Bad\_Data\_Folder."

### ➤ **Name of columns:**

The name of the columns is validated and should be the same as given in the schema file. If not, then the file is moved to "Bad\_Data\_Folder".

### ➤ **The data type of columns:**

The data type of columns is given in the schema file. It is validated when we insert the files into Database. If the data type is wrong, then the file is moved to "Bad\_Data\_Folder"

### ➤ **Null values in columns:**

The data type of columns is given in the schema file. It is validated when we insert the files into Database. If the data type is wrong, then the file is moved to "Bad\_Data\_Folder"

## **Data Insertion into Database:**

### ➤ **Database creation and connection:**

Create a database with the given name passed. If the database has already been created, open a connection to the database.

### ➤ **Table creation in the database:**

Table with name - "Good\_Data", is created in the database for inserting the files in the "Good\_Data\_Folder" based on given column names and datatype in the schema file.

### ➤ **Insertion of files in the table:**

The files in the "Good\_Data\_Folder" are inserted in the above-created table. If any file has invalid data type in any of the columns, the file is not loaded in the table and is moved to "Bad\_Data\_Folder".

## **Model Training:**

### **Data Export from db:**

The data in a stored database is exported as a CSV file to be used for model training.

### **Data Preprocessing:**

- a) Drop the columns not required for prediction.
- b) Remove the unwanted spaces in data.
- c) For this dataset, the null values were replaced with '?' in the client data. Those '?' have been replaced with NaN values.
- d) Check for null values in the columns. If present, impute the null values using the categorical imputer.
- e) Replace and encode the categorical values with numeric values.
- f) Scale the numeric values using the standard scaler.
- g) Handle the imbalanced dataset using oversampling.

## Clustering:

Kmeans algorithm is used to create clusters in the preprocessed data. The optimum number of clusters is selected by plotting the elbow plot, and for the dynamic selection of the number of clusters, we are using KneeLocator function. The idea behind clustering is to implement different algorithms.

## Model selection:

- Performing EDA to get insight of data like identifying distribution, outlier, trend, among data etc.
- Check for null values in the columns. If present impute the null values.
- Encode the categorical values with numeric values.
- Perform standard scaler to scale down the values.

## Prediction:

- The Client will send the data in multiple sets of files in batches at a given location. Data will contain the annual income of various persons.
- Apart from prediction files, we also require a "schema" file from the client, which contains all the relevant information about the training files such as
- Name of the files, Length of Date value in FileName, Length of Time value in FileName, Number of Columns, Name of the Columns and their datatype.

## Q & A :

Q1). What is the source of data?

The data for training provided by the client in multiple batches and each batch contain multiple files.

Q2). What was the type of data?

The data was the combination of numerical and categorical values.

Q3). What is the complete flow you followed in this Project?

Refer slide 3 for better understanding

Q4). After the file validation what you do with incomplete file or files which didn't pass the validation?

Files like these are moved to the archive folder and list of the files has been shared with the client and we removed the bad data folder.

Q5). How logs are managed?

We are using different logs as per the steps that we follow in validation and modeling like file validation log, Data insertion, Model Training log, Prediction log.