Analysis of Scania Truck APS System Failures

Introduction

The primary goal of this analysis is to understand the APS (Air Pressure System) failures in heavy Scania trucks, which can help the company in deciding their investment strategy for the upcoming year. We will be using Python for all data wrangling, analysis, and visualizations. The steps are designed to meet the minimum requirements set by the company's CTO, including data characterization, data preparation and evaluation, and Principal Component Analysis (PCA) for dimensionality reduction.  
Characterisation of the data

First, we import the Pandas framework and assign it the alias 'pd' for ease of code management, allowing us to avoid lengthy references. We proceed to read the file by creating a variable with the same name as the file and utilize the 'pd.read\_csv("filename.csv")' method. The 'read\_csv' function is a component of the Pandas library designed for file reading, and we specify the file's name within the brackets.

In the next step, I make sure that the correct file has been loaded. I suing the code .heas(),but there are many ways to visualize the table like put the name of the variable.

Implementing the code ().shape for can se the size of the data set, so the size of our data set is the 60,000 x 171 that means is 60,000 Observation and 171 attributes

What we can see is that we have a lot rows and is not easy to manage it,the strategy is reduce,so we can reduce the columns?is not specified what is the mean of the columns,we don’t have enof information for can remove the column or reduce,because can be a component of the scania truck,so we cannot even rename it,like for example “aa\_000, ab\_000” = “gear” for example,so what we can do is reduce the number of rows

Implementing .describe for analyse the data, we can se that it shows only one columns, that could be that no has number for can calculate the mean, median…,it show only one column

Let’s check if there is a missing values,for the moment it show that there are no is missing values,that is a little bir raro  
So let check with .types for analyse the columns, and yes the problem is that the other columns are object and not integer or float.

So lets describe the table,with the object value included,and what we ca see is that the value na is repited many time in the clase ab\_000 and ad\_000 there are more value na,and it complicated our analysis so let change it

Now we change the values of the data set,we check if there is a missing values,and surprise,the are many missing values.  
The good thing is when we describe the data,it show all the column,but we se that are many value 0 that is not good for our calculation  
so I I’m going to make 2 type of variable

One variable,a create a new columns based on the result of the media of the other columns based in the information of prediction,but without know of the name the column what can represent,I agroupate the names of the columns based on the letters  
  
  
the curse of dimesionality

Understandably, navigating the complexities of high-dimensional data spaces is much like the challenge of finding a specific landmark in a vast and varied landscape. Each characteristic or feature of the data introduces a new dimension, just as additional geographic features—such as elevation, temperature, or rainfall—would add layers of complexity to a map.

As the number of features (dimensions) in a dataset increases, the difficulty in finding patterns, or in the case of our analogy, landmarks (data points), escalates. This phenomenon is commonly referred to as the "Curse of Dimensionality" (Bellman, 1961). Essentially, in a high-dimensional space, the data becomes so spread out that any new data point—akin to our landmark—can appear to be far from any existing points. This vastness makes it incredibly challenging to perform analyses or predictions because traditional statistical techniques lose their effectiveness as the space expands (Bishop, 2006).

However, it's crucial to strike a balance. While dimensionality reduction techniques such as Principal Component Analysis (PCA) can help by removing redundant or irrelevant features, thereby simplifying the landscape and making it easier to find our landmarks, there is a caveat. If we simplify too much, akin to removing too many details from our map, we risk "underfitting" our model. Underfitting occurs when our model is overly simplified and lacks the complexity needed to capture the underlying patterns and structures in the data, much like a map that lacks the necessary details to navigate the terrain effectively (James et al., 2013).

In practice, the key is to find the sweet spot where the model is complex enough to accurately capture the important patterns (landmarks) in the data without being overwhelmed by the sheer number of dimensions. Techniques for balancing this include cross-validation to test model performance and regularization methods to penalize complexity, ensuring that our model remains robust and generalizable (Hastie et al., 2009).

References:

Bellman, R. (1961). Adaptive Control Processes: A Guided Tour. Princeton University Press.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.

After we insert the code of .info() that is give to us a idea of our data set

Dataset Overview

Data Source: "aps\_failure\_set.csv"

Data Structure:

Number of Rows (Observations): 60,000

Number of Columns (Attributes): 171

1 column of int64 data type (target variable)

170 columns of object data type