CA1 - Integrated

Ariel Goldman

ID sbs23073

*Higher Diploma in Science in Data Analytics for Business*

*Data Preparation – James Garza*

*Statistical Techniques for D.A. – Mariana Iantoro*

Table of Contents

[Introduction 1](#_Toc130887682)

[Dataset summary 1](#_Toc130887683)

[Data Preparation 1](#_Toc130887684)

[Data Frame 2](#_Toc130887685)

[Exploratory Data Analysis 2](#_Toc130887686)

[Figure 1. Head of the dataset. 2](#_Toc130887687)

[Figure 2. Tail of the dataset 2](#_Toc130887688)

[Figure 3. Info of the dataset 3](#_Toc130887689)

[Figure 4. Null/Na values of the dataset. 3](#_Toc130887690)

[Figure 5. Replacing ‘na’ values for ‘NaN’. 3](#_Toc130887691)

[Figure 6. Retrieving all columns. 4](#_Toc130887692)

[Figure 6. Filtering columns +70% missing values 4](#_Toc130887693)

[Figure 7. Dropping columns +70% missing values 4](#_Toc130887694)

[Figure 8. Percentage positive and negative classes 5](#_Toc130887695)

[Principal Component Analysis 6](#_Toc130887696)

[Figure 1. PCA Elbow Plot 7](#_Toc130887697)

[Curse of Dimensionality 8](#_Toc130887698)

[Conclusions and Findings 8](#_Toc130887699)

[Descriptive Statistical Analysis 9](#_Toc130887700)

[Figure 1. Central Tendency Metrics. 9](#_Toc130887701)

[Figure 2. Variance and Standard Deviation in both classes. 10](#_Toc130887702)

[Figure 3. Histogram. 11](#_Toc130887703)

[Figure 4. Box Plot. 12](#_Toc130887704)

[Normal Distribution 13](#_Toc130887705)

[Combined Analysis 13](#_Toc130887706)

[References 14](#_Toc130887707)

# Introduction

The data presented in this dataset was obtained from Scania Trucks that are frequently used for heavy-duty transportation. Specifically, the dataset focuses on the performance of the Air Pressure System (APS), which generates pressurized air used for various truck functions like braking and gear changes.   
The dataset provides two classes: the positive class, which consists of component failures in a specific APS component, and the negative class, which includes trucks with failures that are not related to the APS system.  
  
Through this report, I will present an in-depth analysis of the data and make recommendations for the company’s investment strategy based on my findings.

## Dataset summary

The dataset from Scania Trucks contains 60,000 observations and 171 features, with features indicating the negative class and another indicating the positive class.  
Specifically, the negative class contains 59,000 data points while the positive class contains the remaining 1,000.

# Data Preparation

Different libraries have been used for the purpose of performing the analysis of the dataset called ‘aps\_failure\_set (1).csv’, which is being implemented in Jupyter Notebook.

The following libraries are crucial for data analysis.

* Pandas
* Numpy

Visualisation libraries

* Seaborn
* Matplotlib

Machine Learning library

* Sklearn

Statistical Library

* Scipy.stats

The function import warnings will suppress the errors that would normally be displayed.

## Data Frame

‘df\_failure’ is the name given for this DataFrame that stores the data from the CSV file. This DataFrame can be used to manipulate and analyse the data in a variety of ways, such as filtering, dropping, and visualizing the data.

# Exploratory Data Analysis

Using Exploratory Analysis (EDA), I will be summarizing the main characteristics of Scania Trucks Dataset.

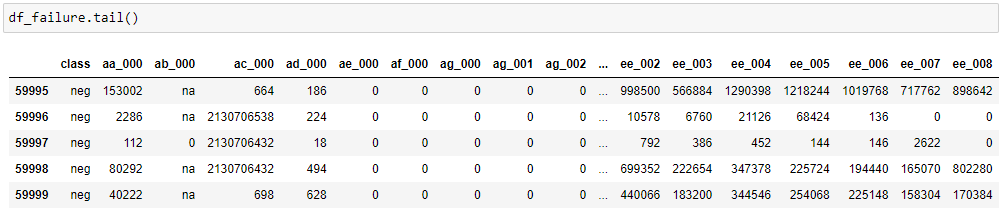
Executing the first 10 rows, it can be observed that there is missing data and the dataset contains 10 rows and 171 columns.

Table

Description automatically generated

## Figure 1. Head of the dataset.

Executing the last 5 rows, the total number of observations are 60.000 (0 to 59.999)



## Figure 2. Tail of the dataset

Checking the summarised information, the total number of observations are 60,000 and total number of attributes 171 that the data set includes for the analysis.

Text, letter

Description automatically generated

## Figure 3. Info of the dataset

Checking missing/null values.

There are no missing or null values in the dataset.

Graphical user interface, text, application, Word

Description automatically generated

## Figure 4. Null/Na values of the dataset.

There is a significant amount of undefined data. Therefore, I replaced the ‘na’ values for ‘NaN’.

A picture containing text

Description automatically generated

## Figure 5. Replacing ‘na’ values for ‘NaN’.

Retrieving all the column names of the dataset and converting them into a list can make it easier to work with the columns.

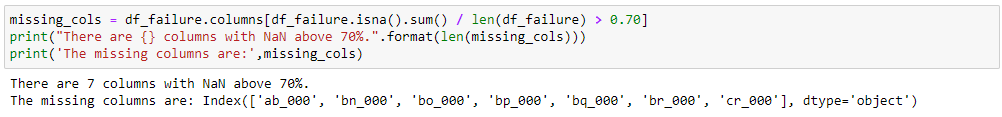
Chart, scatter chart

Description automatically generated

## Figure 6. Retrieving all columns.

I have decided to filter the variables where the percentage of missing values are greater than 70%, given the high number of missing values in the dataset.

During the Exploratory Data Analysis, it was detected that some features had a high percentage of missing values. In particular, 7 out of 171 features had more than 70% of their values missing. As a result, those features were dropped.



## Figure 6. Filtering columns +70% missing values

Text

Description automatically generated with medium confidence

## Figure 7. Dropping columns +70% missing values

The dataset is highly imbalanced, with much larger numbers of negative class points than positive ones. This imbalance can cause the training model to spend most of its time on negative examples and not learn enough from the positive one, which may lead to poor performance on the positive class. (https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data)

Text

Description automatically generated

## Figure 8. Percentage positive and negative classes

# Principal Component Analysis

To perform Principal Component Analysis (PCA), I replaced categorical data with numerical values. ‘class’ negative and positive were replaced with 0 and 1, respectively, and NaN values were filled with mean values.

Basically, we are trying to capture the variance of the dataset by capturing the most important information, while ignoring any noise or less important data points in order to allow us to reduce the dimensions to make it simpler to analyse and visualise. By performing this analysis, we can create a more efficient model to make a more accurate decision.

Minimum number of features needed for retaining 95% variance in the data.

Chart, line chart

Description automatically generated

## Figure 1. PCA Elbow Plot

Based on the plot, it can be observed that the initial Principal Components describe a significant amount of variance in the data, while the amount of variance explained by each successive Principal Component decreases rapidly. In this scenario, it seems that with 18 Principal Components, a substantial portion of the variance present in the data can be accounted for, as the cumulative explained variance at around 0.95. This means that we can retain 95% of the variance in the data with 18 principal components.  
By specifying ‘n=18’ in the PCA model, we are capturing a minimum of 95% of the total variance in the data.

## Curse of Dimensionality

The Curse of Dimensionality is a problem when we have loads of variables to consider at once. From my point of view, if we have a lot of variables, data points or features in a dataset, it becomes difficult to find consistency or understand the patterns between them. Basically, this whole process of collecting data can be tough, challenging or impossible to collect.

# Conclusions and Findings

Based on the analysis and findings, it is recommended that in the upcoming year Scania Trucks invest in a predictive maintenance system that can monitor the APS system’s pressure levels and detect potential failures in advance. This approach can improve the system’s reliability, reduce downtime and save the company significant costs. (https://www.tibco.com/reference-center/what-is-predictive-maintenance#:~:text=Predictive%20maintenance%2C%20also%20referred%20to,predictive%20maintenance%20in%20the%20nineties.)

In conclusion, Scania Track dataset emphasizes the significance of dealing with class imbalance and selecting appropriate features when constructing machine learning models.

The findings can provide valuable insights for Scania Trucks to optimize their maintenance strategy and improve the reliability of their vehicles.

# Descriptive Statistical Analysis

Central Tendency Metrics of the APS component failures in both classes

Text

Description automatically generated

## Figure 1. Central Tendency Metrics.

On one hand, the mean and the median show that the positive class has higher APS component failures compared to the negative class, which can be seen from the mean and median values being much larger in the positive class compared to the negative class.

On the other hand, zero is the value that appears most frequently in the dataset (except for the positive class column which has a value of 1) for both the positive and negative classes.

Variance and Standard Deviation of the APS failures in both classes.

*Compare and give interpretation to these results.*

Text

Description automatically generated

## Figure 2. Variance and Standard Deviation in both classes.

The variance and standard deviation are measures which explain how spread the dataset is.

Having said that, the variance of APS component failures in the positive class, is relatively low compared to the variance in the negative class (2.2307). Additionally, the standard deviation in the positive class is also smaller (around 40 million) compared to the standard deviation in the negative class (around 61 million), implying that there is less disperse in APS failures in the positive class.

To sum up, the above results suggest that the APS component failures have a wider range and higher variability in the negative class compared to the positive class.

Histogram to visualise the distribution of APS in positive class.

Chart

Description automatically generated

## Figure 3. Histogram.

The Histogram displays a symmetrical distribution cantered around the positive class value. This suggests that the positive class has relatively consistent level of APS component failures as there are no outliers or extreme values. Additionally, it appears that there is a spike in the frequency of APS component failures at around 160,000.

Box Plot to visualise the distribution of APS in positive class.

Chart

Description automatically generated

## Figure 4. Box Plot.

The box plot represents the distribution of feature values for the positive class. After applying the Interquartile Range method (IQR) there are 10 observations which are represented by dots after removing the outliers. Any values outside this range are considered outliers and are not included in the plot. It means that 59,990 observations were identified as outliers and removed from the dataset.

# Normal Distribution

Two insightful probabilities based on Normal Distribution:

1. Using Normal Distribution, we can calculate the probability of observing a specific number of APS component failures. Assuming that the mean of APS failures is 3 per day with standard deviation of 1, we can determine the likelihood of observing exactly 5 APS failures.

It is possible to recommend maintenance or repairs to the company, as we can anticipate the number of APS failures that are likely to occur and allocate resources accordingly.

1. The Normal Distribution can also be used to observe a range of APS component failures. Assuming that the mean of APS failures is 3 per day with standard deviation of 1, we can determine the probability of observing between 2 and 4 APS failures.

In this particular range, we could suggest setting targets and establish a range of acceptable failures to improve performance within that range.

# Combined Analysis

Based on the analysis of the dataset and the results above, the use of statistical techniques can help the haulage company make informed decisions about their investment strategy by providing a better understanding of the data related to component failures, maintenance costs, repairs and performance objectives. By using this information, we can plan accordingly to optimize the outcomes for the upcoming year.

# 

# References

sklearn.preprocessing.LabelBinarizer.examples. “Sklearn.preprocessing.LabelBinarizer.”

*Scikit-Learn.org*, 18 Mar. 2023, scikit

https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data. “Imbalanced Data | Machine Learning.” Google Developers, July 18AD, developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data#:~:text=A%20classification%20data%20set%20with.

Educative. “What Is a Subplots in Matplotlib?” *Educative: Interactive Courses for Software Developers*, 26 Mar. 2023, www.educative.io/answers/what-is-a-subplots-in-matplotlib.

https://www.tibco.com/reference-center/what-is-predictive-maintenance#:~:text=Predictive%20maintenance%2C%20also%20referred%20to,predictive%20maintenance%20in%20the%20nineties. “What Is Predictive Maintenance?” *TIBCO Software*, 26 Mar. 2023, www.tibco.com/reference-center/what-is-predictive-maintenance#:~:text=Predictive%20maintenance%2C%20also%20referred%20to.

https://www.pluralsight.com/guides/cleaning-up-data-from-outliers. “Cleaning up Data from Outliers | Pluralsight.” Www.pluralsight.com, 22 Oct. 2019, www.pluralsight.com/guides/cleaning-up-data-from-outliers.