Optimizing APS Failure Prediction: An Evaluation of Techniques and Algorithms

1st Yusuf Sahin
Computer Engineering Department
Istanbul Technical University
Istanbul, Turkey
sahiny20@itu.edu.tr

2nd Mustafa Can Caliskan
Computer Engineering Department
Istanbul Technical University
Istanbul, Turkey
caliskanmu20@itu.edu.tr

***Abstract*—This study focuses on the APS Failure Prediction context hosted on Kaggle, aiming to predict failures in the APS system. Preprocessing steps involved handling missing data, outliers, and feature engineering. Specific machine learning algorithms (such as PCA, LDA, Factor Analysis) were implemented and comprehensively evaluated for their performance. Assessment based on metrics like accuracy, precision, recall, and F1-score highlighted the effectiveness of these specific techniques or algorithms in accurately predicting APS failures.**

I. Introduction

The objective of this study is to determine the condition of the Air Pressure System (APS) in heavy Scania trucks using a binary classification approach. The dataset, sourced from the daily operations of these trucks, poses a challenge of distinguishing between APS component failures (positive class) and failures unrelated to the APS (negative class). Our training dataset contains 60,000 instances—comprising 59,000 negatives and 1,000 positives—while the test set holds 16,000 examples. With a total of 171 attributes, including 7 histogram variables, the dataset provides ample opportunity for feature engineering and model refinement. Through our predictive modeling efforts, the primary aim is to develop a robust algorithm capable of accurately identifying APS-related failures, thereby improving maintenance strategies and operational efficiency within heavy Scania trucks. This paper outlines our approach, methodologies, and findings in addressing this critical challenge within the transportation domain.

II. Data Exploration

When examining the dataset, we observed numerous 'NaN' values and generally noted its imbalance, particularly an overabundance in the negative class. Figure 1 illustrates the distribution of 'NaN' values across the features. To assess the relationship between features, we applied correlation analysis. Figure 2 shows the correlation matrix created by the features. Additionally, we visualized the features to detect outliers. As an example, the visualization we conducted to examine the outliers of the 'aa_000' feature can be seen in Figure 3.

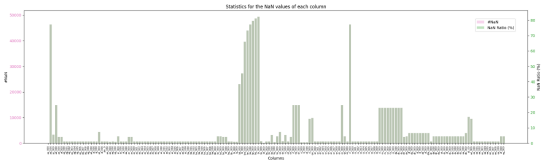

Fig. 1. Statistics of 'NaN' Values

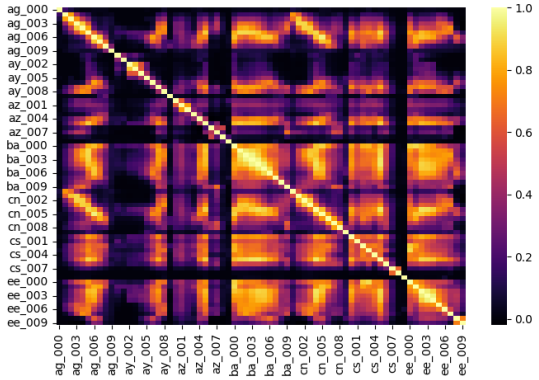

Fig. 2. Correlation Matrix

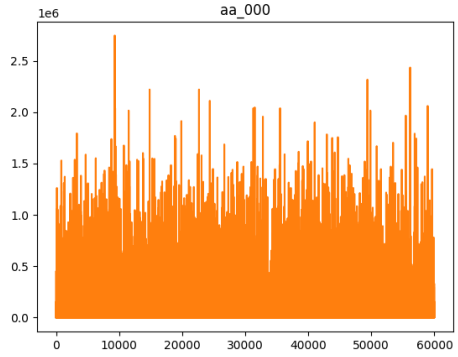

Fig. 3. Visualization of 'aa_000' Feature

III. Methodology

Firstly, we've established the methodology for preprocessing. In the initial phase, we attempted several well-known dimensionality reduction techniques including PCA, LDA, and Factor Analysis, but we did not achieve the desired results. Subsequently, Figure I and II depict the outcomes of PCA and LDA respectively. After, we continue with different approach. If the 'NaN' ratio within columns exceeds 67%, we delete that column. Afterwards, we conduct 'undersampling' based on the majority class. Then, we fill the 'NaN' values based on the median using the SimpleImputer class from the scikit-learn library.

TABLE I
CLASSIFICATION REPORT USING PCA

	precision	recall	f1-score	support
neg	0.99	0.90	0.94	182
pos	0.92	0.99	0.96	218
accuracy			0.95	400
macro avg	0.96	0.95	0.95	400
weighted avg	0.95	0.95	0.95	400

TABLE II
CLASSIFICATION REPORT USING LDA

	precision	recall	f1-score	support
neg	0.90	0.87	0.89	182
pos	0.89	0.92	0.91	218
accuracy			0.90	400
macro avg	0.90	0.90	0.90	400
weighted avg	0.90	0.90	0.90	400

We decided on the random forest model for model selection, as it yields efficient results with imbalanced datasets and demonstrates flexibility. Finally, we trained the model using the dataset.

IV. Results

The classification report of the model can be observed in Table III.

TABLE III
CLASSIFICATION REPORT

	precision	recall	f1-score	support
neg	0.98	0.97	0.97	182
pos	0.97	0.98	0.98	218
accuracy			0.97	400
macro avg	0.98	0.97	0.97	400
weighted avg	0.98	0.97	0.97	400

The classification report outlines the model's performance in categorizing data into "neg" and "pos" classes. Its high precision scores indicate the model's capability to accurately label instances within both classes. However, the slightly lower recall for the "neg" class suggests that it misses a small portion of actual "neg" instances compared to "pos" instances. Nevertheless, the F1-scores, harmonizing precision and recall, demonstrate a strong balance between identifying relevant instances and minimizing false positives. Overall, with a solid accuracy of 97%, this model showcases commendable proficiency in distinguishing between the two classes, indicating its reliability in this classification task.

V. Conclusion

The model developed in this study exhibits impressive performance, boasting a 97% accuracy in categorizing APS failures and non-failures within heavy Scania trucks. Despite a slightly lower recall for the non-failure class, the model maintains high precision and F1-scores, indicating a strong ability to accurately identify APS-related failures while minimizing false positives. This proficiency holds significant promise for enhancing maintenance strategies and operational efficiency in these trucks. The methodology, encompassing preprocessing techniques and the application of a random forest model, has proven effective in handling imbalanced data and achieving robust predictive outcomes.