Predicting of APS failure of Scania truck

CSE445.02 Phase 2 Report

Md. Masudur Rahman
#1631189042
Electric and Computer Engineering
North South University
Dhaka, Bangladesh

Saraf Sumaita Hasan
#1631258042
Electric and Computer Engineering
North South University
Dhaka, Bangladesh

Md Rifat Hasan
#1620259042
Electric and Computer Engineering
North South University
Dhaka, Bangladesh

Rehnuma Sharmin
#1620739042
Electric and Computer Engineering
North South University
Dhaka, Bangladesh

***Abstract*— This report emphasizes on describing the steps and approaches taken in order to keep in check the maintenance costs of the Air Pressure System (APS) of Scania trucks. It can be achieved with the help of accurate predictions. The Logistic Regression model has been used in this case. The results achieved were moderately successful and hence can be considered to be potentially useful for the industrial sector.**

I. Introduction

This problem primarily deals with the well-being of the Scania trucks, the particular unit of interest being the air pressure system (APS).

These heavy-load trucks are highly prone to malfunctions which can prove to be very costly. Therefore this problem was proposed by the IDA Challenge 2016 with the aim of minimizing the maintenance costs as a whole. The given goal can be achieved by predicting beforehand any possible failures and taking necessary precautions. If solved successfully, it can be very cost-effective. The Logistic Regression model has been used as the first approach.

The upcoming sections of the report are:

Background, Dataset description and preprocessing, Methodology, Results, Conclusion and finally a brief take on our Future Plans concerning the problem.

II. Background

Some of the previous works related to this topic are:

A combination of data cleaning, feature engineering, and feature selection was used for data preparation. The model used was Random Forest, with which pretty good results were achieved. [1]

This paper explores and compares the performances given by the various classification trees using different entropies (the measure of information present in a dataset) applied to the Scania trucks dataset. A C5.0 model is the best performing tree in this case. Other entropy classification trees such as Renyi and Tsallis have also proven to be useful. [2]

III. Dataset

The dataset was published by Scania CV AB on the UCI Machine Learning Repository. It contains data for heavy-load Scania trucks which are used on a regular basis and is specifically centered on the Air Pressure System (APS). For any given truck, the target is to find out the cause of its possible breakdown. It can either result from the failure of a component in the APS or outside of it. In short, it gives rise to a classification problem.

The two classes in the dataset are the positive class and the negative class. Positive class implies that the failing component belongs to the APS and the negative class implies otherwise.

Calculating Total cost for miss-classification (Cost Metric):

- False Positive: Given a failure belongs to the negative class but the predicted class is positive, the resulting cost will be cost_1 (cost_1 = 10 which is the cost of unnecessary checking/servicing)

-False Negative: Given a failure belongs to the positive class but the predicted class is negative, the resulting cost will be cost_2 (cost_2 = 500 which is the cost of overlooking a faulty truck and the possible damage resulting from it)

Total cost calculation of a prediction model:

*Total cost = Cost_1 * Number Instances (False positive) + Cost_2 * Number Instances (False negative)*

The sum of (cost_1 times the number of instances of the positive class failure) and (cost_2 times the number of instances of the negative class failure).

The data comes with a training set and a testing set.

The summary of the datasets are given below:

- Training Set :
  - Number of Features: 171
  - Records : 60,000
  - Features: Anonymized due to proprietary reasons
  - Type: Labeled (positive or negative) – Imbalanced (59000 – Negative, 1000-Positive)

- Testing Set :
  - Number of Features: 171
  - Records : 16,000

- Features: Anonymized due to proprietary reasons
- Type: Labeled

Class	aa_000	ab_000	ac_000	ad_000	ae_000
neg	**76698**	**0.713189**	**2.13E+09**	**280**	**0**
neg	**33058**	**0.713189**	**0**	**1.90620.64**	**0**
neg	**41040**	**0.713789**	**228**	**100**	**0**
neg	**12**	**0.000000**	**70**	**66**	**0**
neg	**60874**	**0.713180**	**1368**	**458**	**0**

Data Preprocessing:
Upon inspecting the dataset, initially, a large amount of missing values was observed. This can be illustrated with the help of a heatmap:

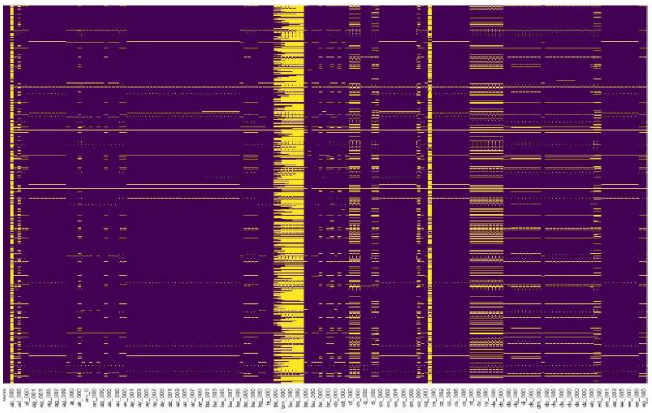

Fig1: Heatmap to illustrate the missing values (Yellow lines indicate the missing values)

The following approaches were taken in handling the missing values:
1st Approach: We dropped columns having more than 80% missing values.
2nd Approach: We kept all features and replaced all the missing values with the mean value. [3]

Both the approaches left us with a clean heatmap with zero missing values:

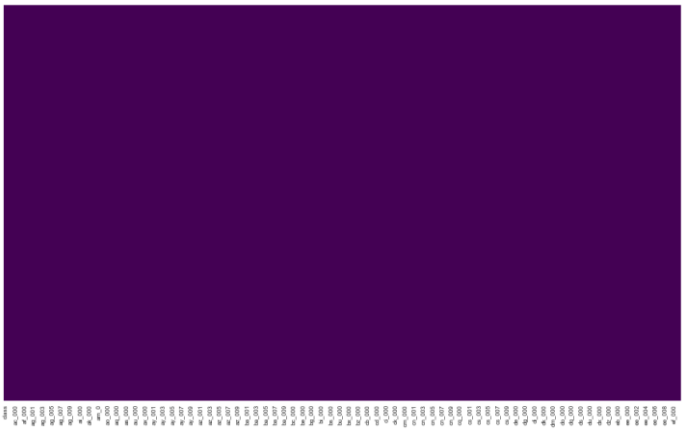

Fig2: Heatmap to illustrate the missing values (Yellow lines indicate the missing values)

The next thing to deal with was our response variable which had string values ("Pos" or "Neg"). In order to turn the values into a machine readable form, a Dummy variable was used. Since this is a binary classification problem, the values 1 and 0 were easily used as replacements (1 = Positive class and 0= Negative class). Finally the dataset was ready to be fed to a model.

The testing dataset was preprocessed in the same way.

IV. METHODOLOGY

Since this is classification problem with only two classes, the Linear Regression model could be an option. It would have done quite well as a classifier (Classify as Positive if Y >^ 0.5). The only problem would be the possibility of it producing values less than 0 or bigger than 1).
Thus the **Logistic Regression** model became our solution.

If we have $p(X) = Pr(Y = 1|X)$
Logistic Regression has the form:

$$p(X) = e\ \beta 0+\beta 1X\ /1 + e\ \beta 0+\beta 1X\ .$$

(e ≈ 2.71828 is a mathematical constant [Euler's number.])

$p(X)$ will always have values between 0 and 1, no matter what values β0, β1 or X take.

Algorithm:

1. The CSV file is read
2. Data Preprocessing is started
3. Missing values management (2 Approaches)
4. A Dummy variable is used
5. Data Preprocessing is completed
6. Model evaluation using Cross Validation (CV=5 and CV=10)
7. Threshold tuning using testing data
8. Model implementation with testing dataset
9. Results are obtained

Algorithm Flowchart:

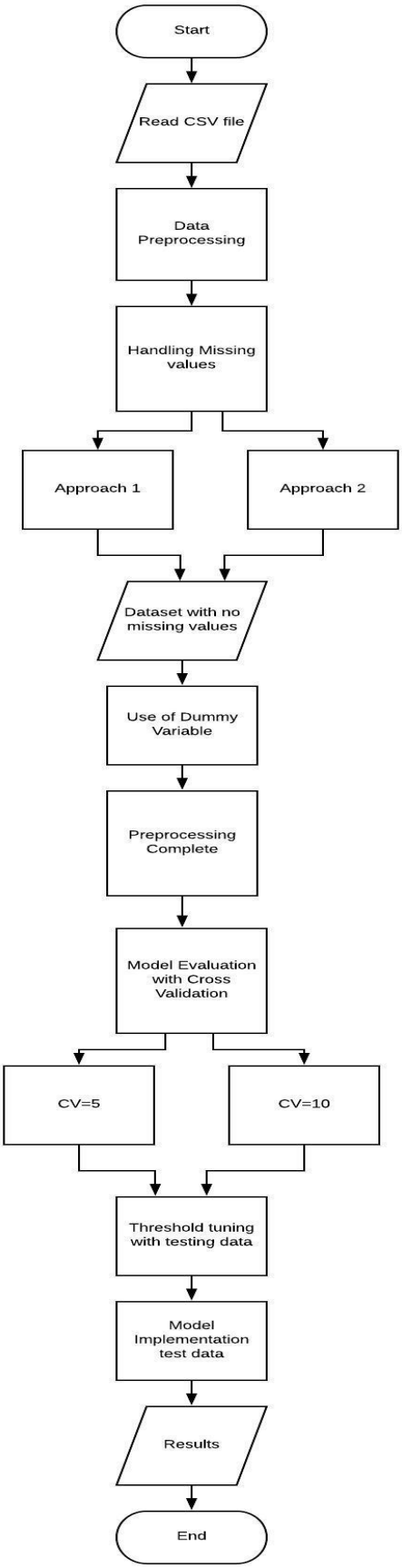

Fig 3: Algorithm flowchart

V. RESULT

So, throughout the Result analysis, it is observed that the best outcome actually resulted from Approach 2, with a modified threshold which is 0.253

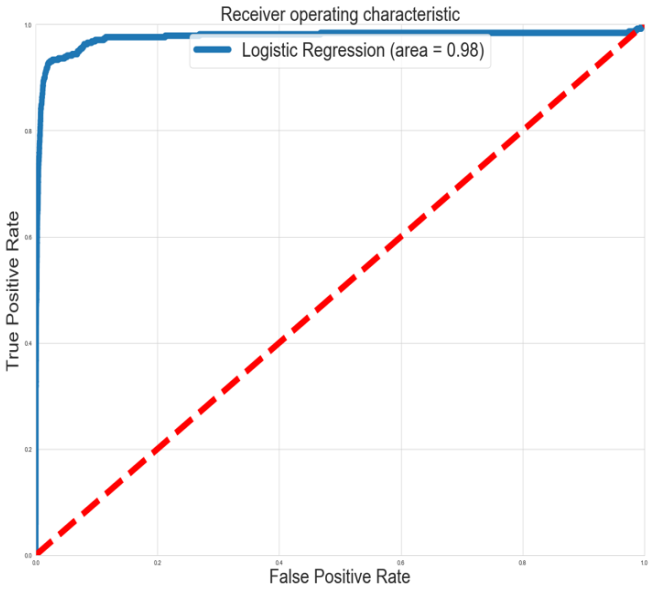

Fig 3: Roc curve for approach 2

Where the accuracy score of 98.9%, R^2 score of 50.8%, F1 score of 75.7%, recall score of 74.9% and lastly the MSE is 1.125%. Also it has AUC of 98% which is nearly perfect. With these settings, we got the total cost of 47860 which is our final score.

Our main target was to minimize this cost as much as possible. But with this Logistic Regression model, this is the best minimization we could arrive to.

VI. CONCLUSION

In conclusion, it can be said that the result achieved with logistic regression is decent, but there is room for more optimization.

VII. FUTURE WORKS

Our expected result was something around 15000~20000. So in that sense, there is a gap of ~27000, which is possible to achieve using some more advanced algorithms such as Random forest or Support Vector Machine.

a) References

[1] Christopher Gondek, Daniel Hafner, and Oliver R. Sampson, "Prediction of Failures in the Air Pressure System of Scania Trucks using a Random Forest and Feature Engineering"

[2] Eleonora Peruff, "Improving predictive maintenance classifiers of industrial sensors' data using entropy. A case stud"

[3] Hyun Kang, "The prevention and handling of the missing data", Department of Anesthesiology and Pain Medicine, Chung-Ang University College of Medicine, Seoul, Korea