# Predicting of APS failure of Scania truck

CSE445.02 Phase 3 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, Support Vector Classifier and Random Forest Classifier models have been used in this case. The results achieved were moderately successful and hence, can be considered to be potentially useful for the industrial sector.**

***Keywords—LR, SVM, SVC, RF, Scaling, Parameter tuning, Feature selection, Dimension reduction***

## 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, Support Vector Classifier and the Random Forest Classifier [1] models have been used for this purpose.

The upcoming sections going to describe the background, dataset description, preprocessing, methodologies, result analysis, model analysis and conclusion which will eventually determine the success of these implemented models.

## II. Background

### A. Classification Algorithm

Logistic Regression (LR) [2] is one of the basic and commonly used binary classification where it classifies between two classes having the values, 0 and 1. It generally calculates the probability of each observation of having any of the upper class and then with the help of the threshold, it determines the classification. The default threshold is generally 0.5. It determines the classes with the help of a decision boundary, which can also be work as a disadvantage of this algorithm.

Standard Support Vector Machine (SVM) [4] can also be described as a binary classifier based on its origin. Apart from the LR, it is not probabilistic. It determines a hyperplane by which the classification is determined. This hyperplane will be optimal if it has the largest distance from the nearest data points from any class' observation. On the other hand, "kernel trick" allows [3, 4] this to perform nonlinear classification. It is suitable for high dimensional data space.

Random Forest (RF) [5] is used for both regression and classification problem. Its default parameter returns a good result without any modification, but for the sensitive result, adjustment is needed. Its runtime is fast. On the other hand, it works well with the unbalanced data as well as the datasets that have a large amount of missing data

### B. Handling Missing Data

There are several ways for handling missing data such as single imputation [6], multiple imputations, pairwise deletion and so on [7]. Among the missing value handling by the imputation with mean/median values is a very common and sometimes a very efficient method.

## 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 a positive class and a negative class. Positive class implies that the failing component belongs to the APS and the negative class implies otherwise.

False Positive defines a situation when 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 defines the situation of 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 is the following

$$Cost = Cost\_1 \times FP + Cost\_2 \times FN \qquad (1)$$

"(1)" 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 main aim is to reduce the cost as much as possible.

This dataset contains a training and a testing datasets. The details of the datasets are given below,

Table 1: Training Dataset

| | |
|---|---|
| Number of features | 171 |
| Records | 60000 |
| Features | Anonymized |
| Type | Labeled |
| Balanced or imbalanced | Imbalanced |

Table 2: Testing Dataset

| | |
|---|---|
| Number of features | 171 |
| Records | 16000 |
| Features | Anonymized |
| Type | Labeled |
| Balanced or imbalanced | Imbalanced |

As it is seen that this dataset is a high dimensional dataset. A sample of the whole dataset is given below,

Table 3: Sample Dataset

| Class | aa_000 | ab_000 | ad_000 | ae_000 |
|---|---|---|---|---|
| neg | 76698 | 0.713189 | 280 | 0 |
| neg | 33058 | 0.713189 | 1.90620.64 | 0 |
| neg | 41040 | 0.713789 | 100 | 0 |
| neg | 12 | 0.000000 | 66 | 0 |
| neg | 60874 | 0.713180 | 458 | 0 |

### A. Data Preprocessing

Upon inspecting the dataset, initially, a large number of missing values was observed. This dataset contains up to 80% missing data.

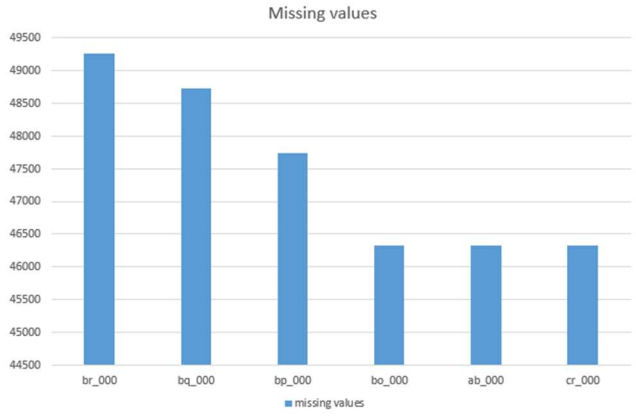


Fig 1: Top 6 column having more than 80% missing values from the dataset

Fig 1 is an illustration of the top 6 columns that have more than 80% missing values from the dataset.

For handling these missing data, all the observation with the missing data are replaced with the mean value* of their corresponding feature.

The next thing to deal with was our response variable which had string values ("Pos" or "Neg"). 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 a classification problem with only two classes, three models will be implemented for the prediction purpose. Binary classifier of Logistic Regression, Support Vector Machine and Random Forest will be implemented.

### A. Logistic Regression

Logistic Regression is a predictive analysis. It is used to describe the data and explain the relationship. It can be said that it is a method for binary classification problems. For the binary classification problem, it is well suited and can perform excellently in some domain.

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

$$p(X) = \frac{e^{\beta 0+\beta 1X}}{1+e^{\beta 0+\beta 1X}} \quad (2)$$

($e \approx 2.71828$ is a mathematical constant)

$p(X)$ from "(2)" will always have values between 0 and 1, no matter what values $\beta 0, \beta 1$ or $X$ is taken.

### B. Support Vector Machine

The Support Vector Machine is also a binary classifier as well as can be used for regression, but the difference between this and Logistic Regression is that SVM is not probabilistic. Rather than, it separates the classes using a hyperplane.

If $f(X) = \beta 0 + \beta 1X1 + \ldots. + \beta pXp$, then $f(X) > 0$ points on one side of the plane and $f(X) < 0$ points on the other side.

This hyperplane will be optimal if it has the biggest gap or margin between the nearest points of the data observations of the classes.

### C. Random Forest

Like its name, it is consist of a large number of individual trees like an ensemble. Each tree in the Random Forest gives a class prediction and the class with the most voted is decided as a result or sometimes mean is chosen.

### D. Algorithm

Step 1. The CSV file is read
Step 2. Data Preprocessing is started
Step 3. Data Standardization
Step 4. Parameter/ Threshold tuning with cross-validation or grid-search
Step 5. Model implementation

Step 6. Primary model evaluation with only train data

Step 7. Implementing test data on the fitted model on Step 6

Step 8. Results are obtained

### E. *Flowchart*

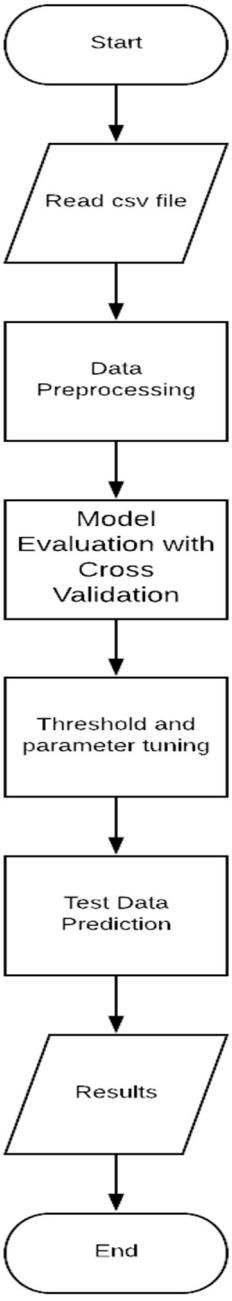


Fig 2: Algorithm Flowchart

## V. Result Analysis

### A. *Logistic Regression*

For this model implementation, the parameters needed to be optimally chosen as well as the threshold. Cross-validation is used to determine the parameter "C".

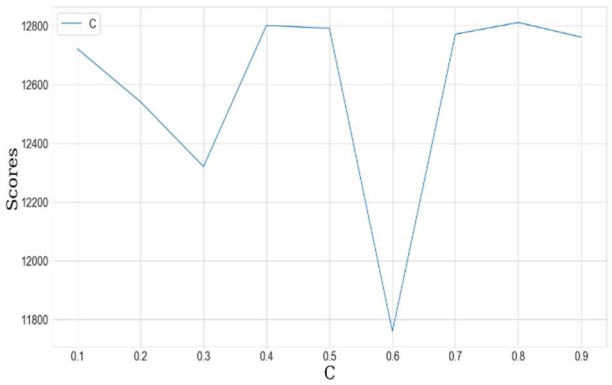


Fig 3: Selecting the optimal value of "C" with CV

(Based on Fig 3) optimal value of C = 0.6 is chosen. For the threshold, ROC curve is used.

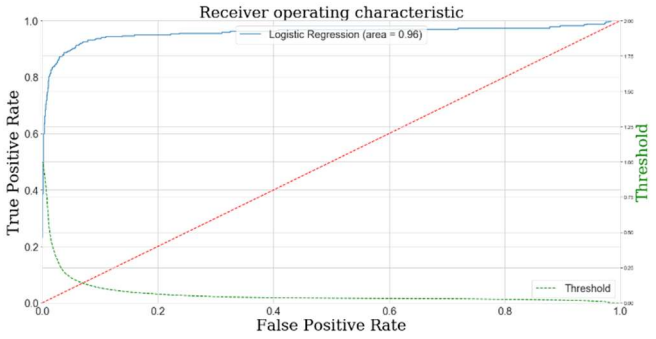


Fig 4: Roc curve for finding the optimal threshold

(Based on Fig 4) Optimal threshold of 0.42 is chosen. Now with these values, if the model is implemented, then the result came,

Table 4: Cost and Misclassification

| Cost | Type 1 fault | Type 2 fault |
|---|---|---|
| 15040 | 554 | 19 |

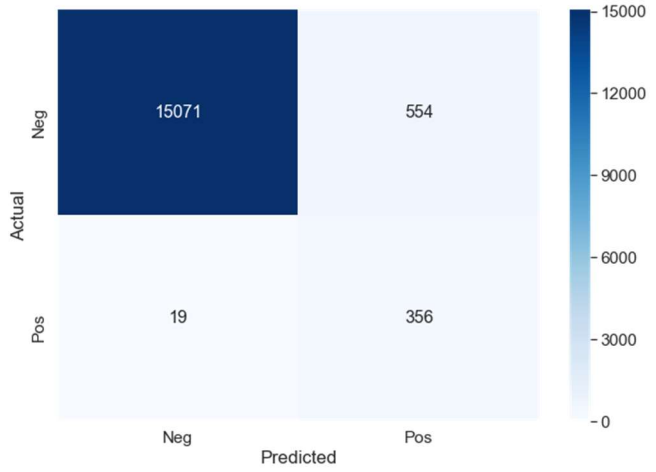


Fig 5: Confusion matrix for Logistic Regression

(Based on Table 4 and Fig 5) the total cost is 15040.

### B. *Support Vector Machine*

For this algorithm, the standardization of data was need. After the normalization, Principal Component Analysis was done along with the classifier and put into a pipeline for grid-

search and to find the optimal values for PCA n components as well as the classifier's parameters.

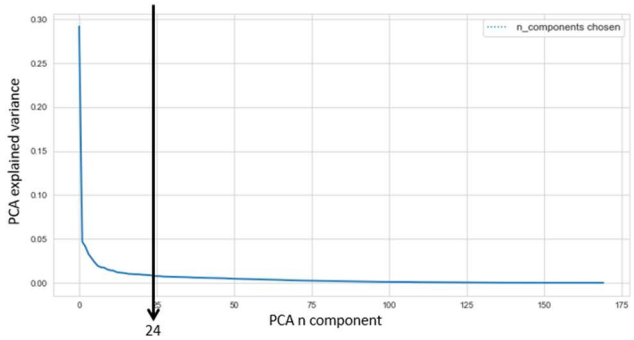


Fig 6: Choosing the optimal n component for PCA from elbow

With the value of SVC model gamma = "0.01" and with PCA n component = 24 (Based on Fig 6) the model is implemented and after the testing data evaluation the result obtained,

Table 5: Cost and Misclassification

| Cost | Type 1 fault | Type 2 fault |
|---|---|---|
| 15220 | 522 | 20 |

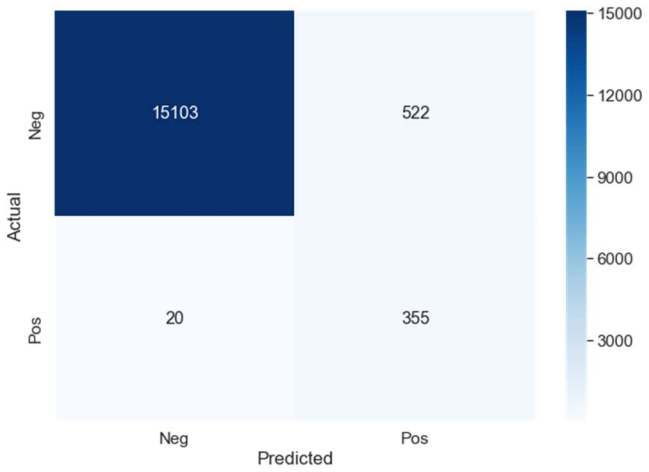


Fig 7: Confusion matrix for SVC

(Based on table 5 and fig 7) The total cost is 15220.

## C. *Random Forest*

For this algorithm, the best result was found after performing feature selection. From 170 features, the feature selection was performed based on their importance.

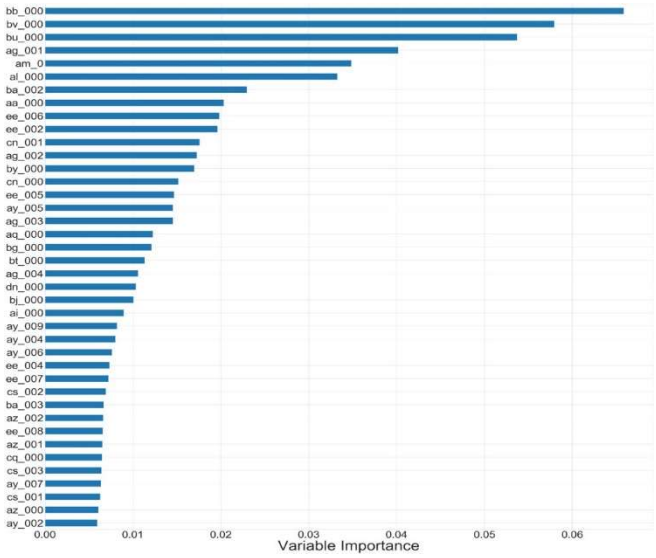


Fig 8: Variable importance of top 40 features

(Based on Fig 8) from 170 features, top 40 is selected and then with the help of observation of the curve [8] the number of trees is selected.

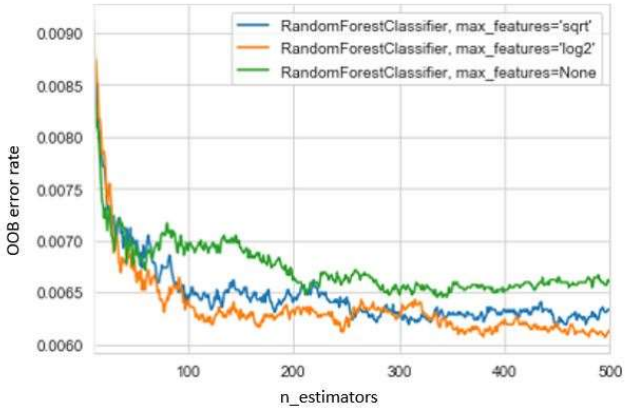


Fig 9: OOB curve

The optimal number of trees = 380 (based on Fig 9) is chosen and max feature =$log2$. After model implementation and applying test data, the result obtained,

Table 6: Cost and Misclassification

| Score | Type 1 fault | Type 2 fault |
|---|---|---|
| 10810 | 681 | 8 |

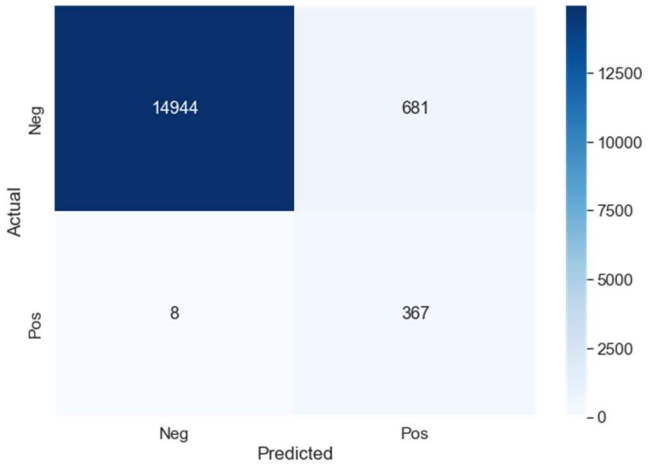

Fig 10: Confusion matrix for Random Forest

(Based on Table 6 and Fig 10) the total cost is 10810

## VI. Model Comparison

To solve the APS problem, three models were implemented, among them Support Vector Machine and Logistic Regression performs almost in the same way, but Random Forest shows the best result. The evaluation scores of these three model obtained,

Table 7: Evaluation Scores

| | Accuracy Score | F1 Score | MSE |
|---|---|---|---|
| Logistic Regression | 96.4 | 55.4 | 3.5 |
| SVM | 96.6 | 56.7 | 3.4 |
| Random Forest | 95.7 | 51.5 | 4.3 |

(Based on Table 7) The evaluation score is almost the same for these three models. Now, for the type misclassification and cost for these three classes,

Table 8: Cost and Misclassification Analysis

| | Cost | Type I Fault | Type II fault |
|---|---|---|---|
| Logistic Regression | 15040 | 554 | 19 |
| SVM | 15520 | 522 | 20 |
| Random Forest | 10810 | 681 | 8 |

(Based on Table 8) It is quite visible that if an individual criterion is chosen, then Random forest will be behind in the game in terms of type I fault, which is the false positive. But according to the problem domain, here false positive is still considerable and has a penalty of 10 whereas type II fault, which is the false-negative has a penalty of 500 and in terms of this criteria Random forest is a clear winner with a cost of 10810.

## VII. Conclusion

For predicting the APS failure system on Scania trucks, three models were implemented and the main aim to reduce the cost as much as possible. The targeted value was 9020 which was the contest winning score, which was almost reached through the Random forest classifier. Although the rest of the two models show very close approximation, but as false-negative has a higher weight, Random forest could make the combination and as a result gives the best result.

## References

[1] Breiman, L.: Random Forests. In: Machine Learning. Vol. 45, No. 1, pp. 5-32. (2001)

[2] Cox, D.R.: The regression analysis of binary sequences. J. R. Stat. Soc. Ser. B 20(2), 215–242 (1958).

[3] Aizerman, pM.A., Braverman, E.A., Rozonoer, L.: Theoretical foundations of the potential function method in pattern recognition learning. In: Automation and Remote Control, pp. 821–837 (1964)

[4] Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the 5th COLT, pp. 144–152 (1992)

[5] Ho, T.K.: Random decision forests. In: Proceedings of the 3rd IJDAR, pp. 278–282 (1995)

[6] Zhang Z.: Missing data imputation: focusing on single imputation. Ann Transl Med. 2016;4(1):9. DOI:10.3978/j.issn.2305-5839.2015.12.38

[7] Breiman, L.: Random Forests. In: Machine Learning. Vol. 45, No. 1, pp. 5-32. (2001)

[8] Mayumi Oshiro, Thais & Santoro Perez, Pedro & Baranauskas, José. (2012). How Many Trees in a Random Forest?. Lecture notes in computer science. 7376. 10.1007/978-3-642-31537-4_13.