Prediction of Failures in the Air Pressure System of Scania Trucks using a Random Forest and Feature Engineering

Christopher Gondek, Daniel Hafner, and Oliver R. Sampson

University of Konstanz, Germany
{christopher.gondek,daniel.hafner,oliver.sampson}@uni-konstanz.de
https://www.uni-konstanz.de

Abstract. This paper demonstrates an approach in data analysis to minimize overall maintenance costs for the air pressure system of Scania trucks. Feature creation on histograms was used. Randomly chosen subsets of attributes were then evaluated to generate an order and a final subset of features. Finally, a Random Forest was applied and fine-tuned. The results clearly show that data analysis in the field is beneficial and improves upon the naive approaches of checking every truck or no truck until failure.

Keywords: Data Mining, Feature Extraction, Dimension Reduction, Random Forest

1 Introduction

Given a high dimensional dataset by this year's Industrial Challenge, we chose a combination of feature engineering and feature reduction whilst constantly evaluating the results using a Random Forest.[1] Our work is structured closely to the KDD process, the model we used throughout the challenge.

2 Project Understanding

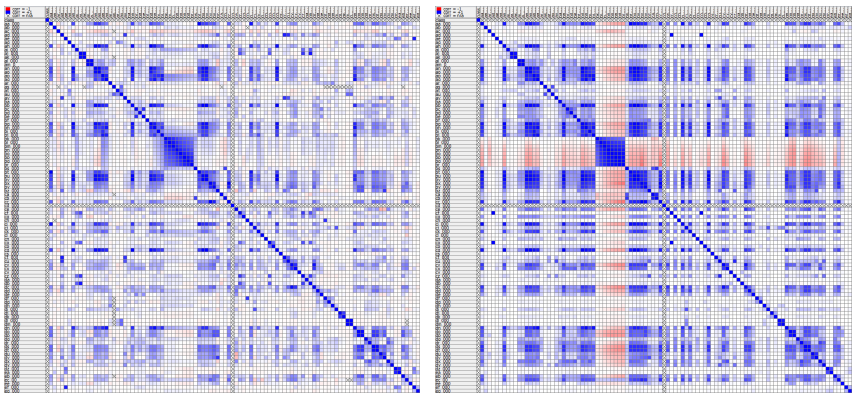
The goal of the task, as presented by the Industrial Challenge for IDA 2016, was to minimize maintenance costs of the air pressure system (APS) of Scania trucks. Therefore, failures should be predicted before they occur. Falsely predicting a failure has a cost of 10, missing a failure a cost of 500. This leads to the need of cost minimization.

3 Data Understanding

The data given to us contains a training set and a test set. The training set contains 60,000 rows, of which 1,000 belong to the positive class and 171 columns, of which one is the class column. All attributes are numeric. 70 of these attributes

belong to 7 histograms with ten bins each. Based on visual inspection we guessed, that the sum across each histogram indicates the age of the APS. Also, most failures could be predicted by using one or two features. It appeared that the hard part is to correctly predict failures for records that are actually very close to the non-failure class. Some visual inspection methods we used were:

- Box plots to get an overview of the variance of the values.
- Correlation matrices for identifying features that correlate. (see Figure 1)
- Scatter plots to see how the classes are spread.
- Radar charts to recognize outliers.



(a) (b)

Fig. 1: Correlation matrices of the attributes for the (a) positive class and the (b) negative. A blue color means a positive correlation, red a negative one. The saturation indicates the strength.

4 Data Preperation

4.1 Data Cleaning

The dataset contains up to 82% missing values per attribute. Furthermore, many of the attributes contain outliers. Therefore, we chose to replace the missing values by the median.

4.2 Normalization

After evaluating several classifiers including Naive Bayes, Multilayer Perceptron, and Support Vector Machines, we determined that a Random Forest would perform best. Hence, a normalization was not necessary.

4.3 Feature Engineering

In our first models we only considered the sum of each histogram and excluded the single bins. This lead to good results but disregarded the distribution of the values. Therefore, we then calculated 16 different features for each histogram. All of these are distances to other distributions using two different distance functions.

The two distance functions we used are the χ^2-distance and the Earth Mover's Distance. The χ^2-distance, proposed by Pearson in the early 1900's [2], is a bin-wise comparison of the observed value to the expected one. The Earth Mover's Distance, introduced by Rubner et al. in 1998 [3], finds the cheapest way to transform one histogram into another one. For this purpose it takes the distances of two bins to each other into account.

With these functions we calculated the distances to the following four different distributions:

1. Mean distribution of the positive examples. It is calculated by filtering the data points with the positive class and computing the mean value for each bin. Based on these, the distance from the computed to the measured histogram is calculated.
2. Mean distribution of the negative examples. The calculation was done in a similar manner as the first one but considers the negative class instead of the positive.
3. Normal distribution with the parameters $\mu = 5,\ \sigma = 1.5$.
4. Mirrored normal distribution. It was achieved by mirroring the normal distribution along the x-axis and shifting it on the y-axis into the positive.

All the above mentioned distances are highly correlated to the sum of the bins. To resolve this dependency the histograms are normalized by their sum. With these additional histograms the same distances as before are calculated.

4.4 Feature selection

The features given and calculated combined resulted in 282 dimensions, excluding the class column. As stated in Section 3, many of them were correlated and hence probably not needed. Consequently, feature selection was introduced. It was done in two steps: Ranking the features by their expressiveness and testing the performance of the feature sets varying in size.

Feature Expressiveness Our approach to rank the dimensions according to their expressiveness was the following:

1. Take 200 random features out of the 282.
2. Learn a RANDOM FOREST and predict the class.
3. Store the precision of the results together with the features used.
4. Repeat steps 1-3 2,000 times.
5. Calculate the mean precision of each dimension and rank them in decreasing order.

Testing the Feature Sets Using the ranked features from Section 4.4, we could compute the costs of the prediction model with a different number of dimensions.

This was done by training a Random Forest and predicting the class using a 10-fold cross-validation and calculating the average costs starting with the feature set containing only the most expressive feature, i.e., `am_0`. Afterwards, the set got expanded by the second best feature and the prediction was repeated. This was done until all dimensions were included.

The analysis showed that the average costs per data record as determined by the given cost function of only one dimension are about 3.1 and decrease rapidly. Between 10 to all 282 features, the costs fluctuates between 0.85 and 0.6. This led us to the conclusion that we do not need all dimensions and can reduce them.

5 Modeling

The Random Forest algorithm always tries to minimize the prediction error. It assumes that all wrong predicted classes are equally expensive. But that is not the case for the IDA Challenge. In fact, the cost of a false negative is 50 times higher than a false positive. We tried to overcome this problem by correcting the predicted class based on the confidence of our classifier. For that we slightly adjusted the procedure described in Section 4.4. For every feature subset we set a threshold for the prediction confidence and changed it in steps of one percent. Whenever the confidence was below or equal to the threshold, the predicted class was set to "pos."

An analysis of the results showed that in most cases the best threshold was 95%. Using that, we got the costs that are shown in Figure 2. With at least 10 dimensions the costs fluctuates between about 0.75 and 0.57. To get the best prediction possible, we used the global minimum with 210 features.

6 Evaluation

The naive approach to solve the challenge would be to label all records as negative which has a mean cost of $500 \cdot 1000/60000 = 8.33$ or to label all records as positive which results in a mean cost of 9.83. Therefore, our approach with average costs of around 0.6 is able to reduce the mean cost by the factor of 13.9.

To get reproducible results with the stratified sampling in the 10-fold cross-validation a fixed seed was used. Since this may lead to an adaption of the subsets of features to said seed, we did repeat the cross-validation with several others. Overall, the costs per truck stayed approximately the same. Therefore, we assumed that overfitting is not a major problem and the expected costs are in the neighborhood of around 0.6.

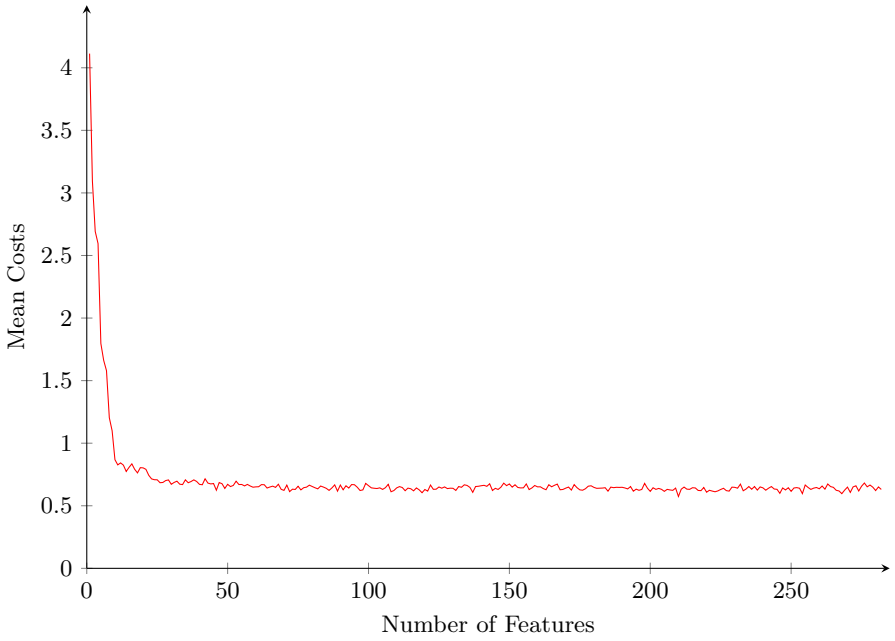


Fig. 2: Curve of the mean costs evaluated using a confidence threshold of 95% and 10-fold cross-validation with a fixed seed.

7 Conclusion

An early detection of a failure in an Air Pressure System in trucks can save the company a lot money. The prediction of a fault can be performed even if the meaning of the measured values is unknown or only histograms are available. We demonstrated how meaningful features of histograms can be computed to improve the prediction. Also, we showed how the forecasts can be adapted to a cost function using a threshold on the confidence of a Random Forest. Finally, a significantly lowered main cost compared to the naive approaches was achieved.

References

1. Breiman, L.: Random Forests. In: Machine Learning. Vol. 45, No. 1, pp. 5-32. 2001
2. Pearson, K.: On the criterion that a given system of derivations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In: The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. Vol. 50, No. 5, pp. 157-175. 1900
3. Rubner, Y., Tomasi, C., Guibas, L.J.: A Metric for Distributions with Applications to Image Databases. In: Proceedings of the Sixth International Conference on Computer Vision. pp. 59–. IEEE Computer Society, Washington, DC, USA (1998)