AIR PRESSURE SYSTEM FAILURE

Cost Reduction Model

Satish Agrawal
sagrawal@my365.bellevue.edu

Overview

In the trucking industry, maintenance, breakdown, and safety are a few of the major concerns. There are preventive, routine maintenance to confirm the current state of equipment in the truck and avoid any breakdowns. And there are expenses after the breakdown (due to a missed routine maintenance) to get the truck back on the road. Unnecessary preventive checks cost money and add burden to the company, and on the other hand, a missed maintenance may cause a breakdown and cost the company even more. There is a need to predict the maintenance timeline to minimize both types of expenses. This project aims at developing the best possible model to predict the need for preventive care and avoid the breakdown while keeping the maintenance cost as low as possible.

Approach and Modelling Methodology

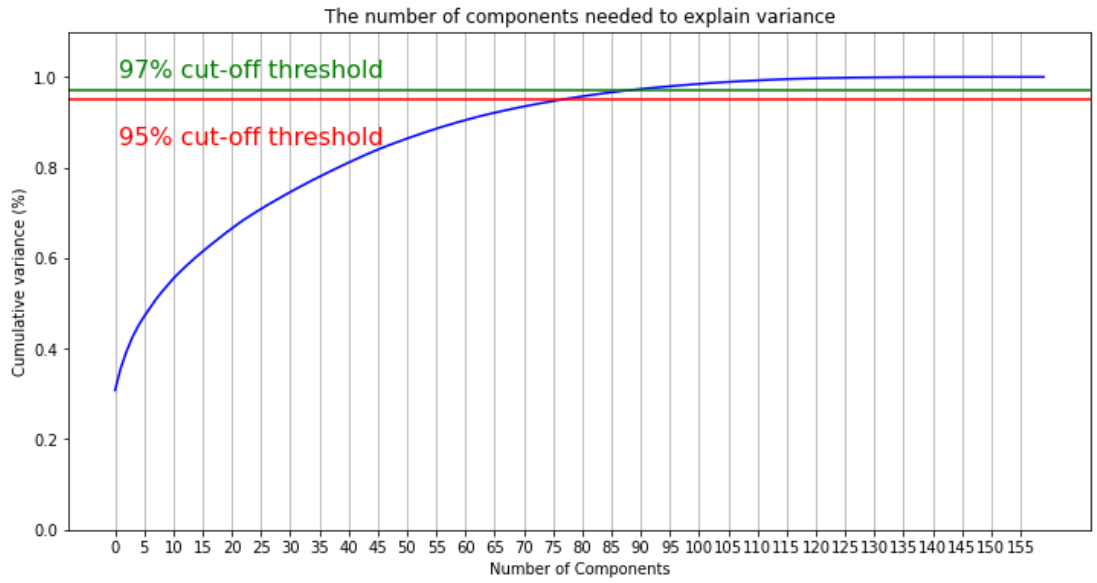
Dataset available comes in two files one for a train set and another for a test set. The dataset was released by Scania CV AB [1] on the UCI Machine Learning Repository [2]. The target is to predict the failure of the Scania Air Pressure System (APS) in trucks to enable preventive maintenance and thereby reduce the maintenance costs. The dataset is anonymized due to proprietary reasons and should not be a problem for the usage of this project.

The target variable in the dataset is a class that can take the values of either positive or negative. The positive value indicates that the maintenance was due to the APS failure and a negative value indicates that the maintenance was due to some other non-APS failure. It would have been ideal if explanatory variables are independent identically distributed which is not the case for the dataset at hand. There are some correlations found between the explanatory variables but it not strong and does not impact the project. Dataset had a lot of missing values and imputation techniques were used to impute missing values using the most frequent value.

Also, it was found that many of the attributes have strong correlations with the target variable. The data set has 171 attributes which were cleaned down to 161 attributes, removing attributes with excessive missing values and then reduced to 108 attributes using Principle Component Analysis (PCA). PCA graph shows that 95% of the cumulative variance is explained by only 75 components and 97% of the cumulative variance is explained by 85 components. I ran PCA for 99% of the variance and successfully reduced the components down to 108.

For the evaluation of the model, we have come up with a metric to come up with a cost equation. Negative class maintenance was assigned $10 and a positive class maintenance event is assigned $500. In other words, broken equipment costs much more to fix (or replace) and then get the truck back to operating mode as opposed to routine maintenance to avoid the breakdown. However, too many preventive maintenances also cost money and need to be

minimized. The dollar amount here merely depicts the ratio of expenditure which is 1:50 for negative versus positive cases.



A series of models including Logistics regression, Random forest classifier, and Xgboost classifier were trained using k fold cross-validation and hyperparameters are tuned to pick the best performing model. Finally, Random Forest Classifier came out as the most effective model given the minimum cost prediction using the given test set. The cost equation as explained above is used to evaluate the model's performance in the test data (and in the validation set). The lower the cost incurred by the model, the better the model is.

Explanation of Results

Table 1 depicts the result in the form of a confusion matrix when the best model (based on the custom scoring metrics) was run through the test data set. The table shows that the model was able to predict accurately to true when the actual target variable was true 15609 times out of 16000 observations. It also shows the model predicted false when the actual value was false 249 times which sums up to 15858 times the model predicted correctly. Now looking at the other two values, the model predicted true 126 times the actual value was false, and the model predicted false 16 times when the actual value was true.

The equation for the cost is
Cost = False Positive X 500 + False Negative X 10

Putting values to the equation the total values comes to

126 X 500 + 16 X 10 = 63160

	Predicted True	Predicted False
Actual True	15609	16
Actual False	126	249

Table 1. Predicted versus actual results

Conclusion and next steps

While this project concludes that the Random forest classifier gets the minimal cost, there could be future enhancements to the model to minimize the total cost. Random forest works well because it is simply a collection of decision trees whose results are aggregated into one final result. Random Forests has the ability to limit overfitting without substantially increasing error due to bias. One way Random Forests reduce variance is by training on different samples of the data. We can try CatBoost classifiers to see how it performs in terms of getting the least cost.

A good next step could be to work with adjusting the probability threshold of the fitted model. Looking closely at the cost equation, reducing the false-negative even slightly will reduce the cost by big amounts. The best probability threshold is decided by using cross-validation which gives the least cost. In this process, we will end up finding the optimal values for FN and FP.

The dataset is highly imbalanced, and it has the majority of the observations for negative and very few of the positive outcomes. It would be beneficial to do some class balancing by down sampling the negative class and/or oversampling the positive class. This means one possible next step would be to use the smote technique for class balancing.

References:

1. *More information on Scania trucks and services can be found at - https://www.scania.com/group/en/home/products-and-services.html*
2. *Dataset - https://archive.ics.uci.edu/ml/datasets/APS+Failure+at+Scania+Trucks*